1993

Essays On Recursive Nonparametric Kernel Estimation Of Regression

Miang Hong Ngerng

Essays On Recursive Nonparametric Kernel Estimation of Regression

by

Ngerng, Miang Hong

Department of Economics

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London Ontario
October 1992

ISBN 0-315-81302-4

Abstract

In the practice of economics it is common that the data are observed not as a sample of fixed size, but rather as an ongoing sequence of a time series. It could be computationally advantageous if the estimate of the unknown function could be updated for each newly arriving data point. On some occasions, there is also a need to update the existing estimate with the newly realized observations. Kalman filter and Bayesian estimation are the commonly encountered techniques to handle these problems in the paradigm of linear parametric estimation. However, few procedures are available for nonlinear models, especially in the nonparametric setting. This thesis attempts to formulate such an estimator using the recursive version of the Nadaraya-Watson estimator.

The recursive estimator for the conditional mean of a nonparametric regression model with independent observations was thoroughly explored in the late 1970's and early 1980's by authors such as Greblicki and Pawlak (1987). The first chapter of this thesis summarizes the constructs and methods of analysis developed in connection with such estimators for independent observations and briefly demonstrates some of their asymptotic properties under the chosen conditions. However, economic time series are generated as economic agents engage in intertemporal optimization and are usually heterogeneous, correlated and unlikely to be linear. There is an incentive for us to extend the study of this recursive nonparametric regression estimator to the case where the observations are correlated. This investigation forms the content of

Chapter Two. In Chapter Three, we propose a recursive version of nonparametric kernel estimator of the derivative of a regression function and establish the conditions to ensure that it is consistent and has an asymptotically normal distribution. In Chapter Four, we show an implementation of the recursive estimator and examine its finite sample properties.

Acknowledgements

I would like to thank the members of my Thesis Committee, John Knight, R.A.L. Carter and Kim Balls for their continuous encouragement, intellectual stimulation, time and interest which they gave me during the course of writing this thesis. I would also like to thank Aman Ullah for his guidance and assistance. Above all I would like to thank them for teaching me econometrics.

I owe my successes to instructors from the Department of Statistical and Actuarial Science of University of Western Ontario and University of Waterloo, from whom I have learnt statistics.

Table of Contents

List of Figures

Introduction

Economic theory can very often tell us only that there exists a relationship between a dependent variable, y, and a vector of exogenous variables, x. The relationship could be represented by a model $y = m(x) + \epsilon$ with $E(\epsilon|x) = 0$ for example. However, the exact form of the function $m(x)$ and the distribution of ϵ is often unknown. Traditionally, $m(x)$ is taken to be a linear function, partly due to expediency and partly because most functions can be approximated by their linear first order Taylor expansions. At any rate, this approximation is a crude one and it does not give us any clue regarding the statistical behavior of this linearized approximation. The nonparametric technique improves upon the situation by providing an approximation scheme for the unknown $m(x)$ that improves as the sample size increases. What is more important is that the asymptotic properties of the estimators can easily be established.

In the 1970s, some authors of empirical economic research, Christensen, Jorgenson and Lau (1973), Diewert (1971), *inter alia* used families of functions like Cobb-Douglas and C.E.S. functions in the estimation of production functions. Later, this development led to the use of flexible functional forms such as the trans-log and generalized Leontief. Lau (1986) gives a comprehensive survey on the use of nonlinear functional forms in economics. The major setback with this approach is that it essentially provides a local truncated Taylor series approximation to the unknown underlying function and has very poor statistical properties. For example the estimate of the

OLS covariance matrix is usually inconsistent. In the econometric literature, Gallant (1981) pioneered the use of the Flexible Fourier form to produce an approximation to $m(x)$ with an accuracy that can be precisely described. To achieve improved accuracy with increased sample size, the number of parameters to be estimated in this method expands with sample size. It is now commonly known as the series expansion approach.

The technique used in this thesis is an alternate approximation scheme commonly called kernel estimation. A kernel estimator produces a local approximation centered at a fixed point, x, by taking the weighted average of the y_i corresponding to all observations of x_i that are close to x. The weight is small for an observation x_i that is far from x. It is determined by a kernel function, $W(.)$, and a parameter h_n called the window width. The same window width, h_n is applied to all the observations in a given sample, $\{ x_i, y_i \}_1^n$. The estimated conditional mean is $m_n(x) = \frac{\sum_1^n W(x_i, h_n, x) y_i}{\sum_1^n W(x_i, h_n, x)}$. The improving approximation as the sample size gets larger requires that the window width h_n converges to zero as the sample size, n, grows.

In economics, it is sometime the case that the data are observed not as a sample of fixed size, but as an ongoing sequence of time series data. This is the most common of many situations that require the updating of prior calculated estimates with newly arrived observations through some form of recursive estimation. The techniques that are commonly encountered in the econometrics and statistics literature are recursive least squares, Kalman filtering (see for example, Harvey (1989), for some details) and Bayesian techniques (for example West (1990)).

Consider the estimation of a linear model, $y = x\beta + \epsilon$, for example, the ordinary least squares estimator can be of the form: $\beta_n = M_n^{-1} C_n$ where $M_n = \frac{1}{n} \sum_1^n x_i x_i'$, $C_n = \frac{1}{n} \sum_1^n x_i y_i$.

The recursive updating with the $(n+1)$st observation can be easily implemented as follows: $M_{n+1} = \frac{n}{n+1}M_n + \frac{1}{n+1}x_{n+1}x_{n+1}^t$, $C_{n+1} = \frac{n}{n+1}C_n + \frac{1}{n+1}x_{n+1}y_{n+1}$ and $\beta_{n+1} = M_{n+1}^{-1}C_{n+1}$.

The recursive estimator here is the same as the nonrecursive estimator. However, the technique is meant for linear parametric models. So far few techniques are available for nonlinear models, especially in the nonparametric setting.

Economic time series are generated as economic agents engage in intertemporal optimization and usually are correlated and heterogeneous stochastic processes. As the environmental factors facing the agents change from time to time, the dynamic structures that generate the data are unlikely to be linear. The aim of this thesis is to handle these features by sequentially updating the estimate in nonparametric settings with correlated and heterogeneous exogenous variables.

In nonparametric estimation, the window width usually decreases as the sample size increases. In order to incorporate recursive updating, we have to use a different window width for each observation, $m_n(x) = \frac{\sum_1^n W(x_i\ ;h_i,x)y_i}{\sum_1^n W(x_i\ ;h_i,x)}$. The recursive estimator in a nonparametric setting therefore cannot be an exact equivalent to its fixed width counterpart. One of the main features in this thesis is the development of asymptotic results that parallel those of the normal nonrecursive nonparametric estimator. The other issue is how the use of nonuniform width in recursive estimation affects the finite sample behavior.

In the past ten to twenty years, extensive research has been produced in the area of recursive nonparametric kernel estimation with independent observations. The first section of Chapter One surveys the literature in the area of recursive nonparametric estimation with independent observations. Section 2 then defines the assumptions which are used in this thesis as well as in the version of recursive nonparametric

estimators of the conditional mean that was first proposed by Ahmad and Lin (1976). The next section then goes on to reconstruct versions of asymptotic results that have been established by Ahmad and Lin (1976) which then are used as framework for further extension in the later chapters. The last section derives the approximate finite sample bias and asymptotic variance of the estimator.

In Chapter Two, we extend recursive estimation to α-mixing observations and derive the asymptotic properties of the estimator. Section 1, surveys the literature relating to nonparametric kernel estimation with α-mixing observations. Sections 2 and 3 define and develop the constructs that are needed for the derivation of asymptotic results in section 4. This proof uses a version of the central limit theorem due to Davidson (1990), the unpublished of which is reconstructed as an appendix to Chapter Two.

Section 1 of Chapter Three surveys the literature in the area of nonparametric kernel estimation of derivatives. In section 2, we propose a recursive version of the nonparametric estimator of the first (partial) derivatives of the regression. The asymptotic properties of the estimator are derived in section 3.

Chapter Four has two objectives, one is to provide an implementation of the recursive estimator and the other is to analyse the finite sample properties of the recursive estimator through a Monte Carlo study. After the introductory section, section 2, explains the issues pertaining to the implementation of kernel estimators, and the rationale behind our implementation of recursive estimators and their algorithm. Section 3, then describes the Monte Carlo experiment and the implementation of two other fixed window width estimators. Section 4, analyses the outcomes of the experiment.

The last chapter summarizes the major contributions of the thesis and indicates directions for further research.

Chapter 1

Review on the Recursive Nonparametric Kernel Estimation for Independent Observations

1.1 Introduction

As in most regression models, we assume that there exists a Borel measurable function $m(x)$ such that $Y = m(X) + \epsilon$ and $E(\epsilon|X) = 0$ that is $m(x) = E(Y|X = x)$.

The error ϵ needs not be from a Gaussian process, and, unlike parametric regression, $m(x)$ is not restricted to any specific functional form. All that we require is the smoothness of $m(x)$ in the sense that it has derivatives at least up to the 2nd order. We can write $m(x)$ as a ratio of two functions:

$$\begin{aligned} m(x) &= E(Y|x) \\ &= \int y f_{y|x}(y)dy \quad f_{y|x} \text{ is the conditional density of y given x,} \\ &= \int \frac{yf(x,y)}{f_x(x)}dy \quad f(x,y) \text{ is the joint density of (x, y) and} \\ &\qquad\qquad\qquad\qquad f_x \text{ is the marginal density of x} \\ &= \frac{r(x)}{f(x)} \end{aligned}$$

where $r(x) = \int yf(x,y)dy$ and we simply denote the marginal density of x by $f(x)$.

Both Watson (1964) and Nadaraya (1964) have, independently, used the method

of kernel density estimation, first proposed by Rosenblatt (1956), to separately estimate $r(x)$ and $f(x)$ in sequence. They then use the ratio of these two estimates as an estimate of $m(x)$. This version of the kernel estimator for regression function is commonly known as the Nadaraya-Watson (NW) Estimator. Our research is based on a recursive version of this NW estimator first used by Ahmad and Lin in nonparametric regression estimation for independent observations. In this chapter, we reconstruct our version of proofs to be used as framework for further extension in the later chapters. We will extend the existing work to the case where the regressors are generated by strong mixing processes. We also propose a version of the recursive kernel estimator for the first partial derivative of the regression function and we study its large sample properties.

1.2 Review on Existing Works on Kernel Regression Estimation

1.2.1 Survey of Recursive Kernel Regression Estimator and Related Works

The idea of recursive nonparametric estimation was first mentioned in the paper by Wolverton and Wagner (1969). Later Yamato (1971) independently introduced the recursive estimator as a variant of the Parzen's (1962) kernel density estimator.

Density estimator f_n

The estimator uses a window width that varies with the number of observations instead of a constant window width for the entire sample. This slight modification certainly makes the analysis of the estimator more clumsy. However, with it one could easily update the estimate with new information using a recursive algorithm, thereby avoid the need to re-compute a new estimate from the whole data. Yamato (1971) derived the weak consistency conditions and established the asymptotic normality of the recursive density estimator for a sample of independent observations under fairly general regularity conditions. Deheuvels (1974), and later also Devroye (1979), gave conditions for strong consistency of this estimator. Singh and Ullah (1986) studied the multivariate marginal density and conditional density estimators and gave some examples of applying these estimators in econometrics. Some variants and generalisations of the recursive kernel density estimator for independent observations were proposed and investigated by several authors in the past, including Davies and Wegman (1975,1979), Carroll (1976) and others.

Regression estimator m_n

Ahmad and Lin (1976) were the first in applying the recursive kernel estimator to estimate a regression function. They obtained the conditions for both weak and

strong consistency. They also proved the asymptotic normality of the regression estimate. Greblicki, Pawlak and Krzyzak (1984) and Greblicki and Pawlak (1987) give a set of necessary and sufficient consistency conditions of such an estimator when observations are independent. Another interesting development was put forth by Rutkowski (1985). He proposed an algorithm for independent observations based on the kernel estimator to identify a time-varying non-stationary system which is modeled by the regression relationship $Y_i = m_i(X_i) + \epsilon_i$ with a slowly varying mean function $m_i(x)$. He investigated the convergence conditions of this estimator.

1.2.2 The Recursive Estimator of the Mean Function $m(x)$

The nonparametric kernel estimators are typically defined to deal with the entire sample as a whole, thus if data were to come sequentially then for each newly arriving observation it would require recomputation of the estimate with the entire sample. It could therefore be of computational advantage if the regression estimate based on (n+1) points can be evaluated from the (n+1)th observation (X_{n+1}, Y_{n+1}) and the estimate based on the first n points without recalling the previous data points from computer memory and recomputing with the entire sample of observations.

Here we will look at a recursive version of the Nadaraya-Watson Kernel estimator, $m_n = r_n/f_n$, where

$$r_n = \frac{1}{n}\sum_{1}^{n} K_i Y_i$$

$$f_n = \frac{1}{n}\sum_{1}^{n} K_i \quad \text{and} \quad K_i = \frac{1}{h_i^d} K(\frac{x - X_i}{h_i}).$$

With this estimator, the function $m_n(x)$ can be recursively estimated by the algorithm

$$r_n = \frac{n-1}{n} r_{n-1} + \frac{1}{n h_n^d} Y_n K(\frac{x - X_n}{h_n}) .$$

$$f_n = \frac{n-1}{n} f_{n-1} + \frac{1}{nh_n^d} K(\frac{x - X_n}{h_n}) .$$

This modification enables dynamic updating of the estimates. Such estimation is analogous to the recursive least square of the traditional OLS paradigm, or the Kalman filter for the linear time series model in state-space form, see for example, Harvey (1989) for some details, or even the Bayesian estimation of a dynamic model see for example, West and Harrison(1990). The recursive kernel estimator provides an alternative to these techniques for dynamic updating which are required in some occasions for applications of econometrics and statistics.

1.3 Preliminary Notions and Some General Assumptions

In the literature on kernel estimation, similar estimators might not share similar properties due to differences in the conditions imposed on the kernel function, underlying density functions and the regression function. At this juncture, we will make plain the assumptions employed in the thesis. Assumptions like heterogeneity and correlated data generating processes for examples, are chosen with the intention of meeting the needs of common econometric applications.

1.3.1 The Data Generating Process

Let $(X_{n,i}, Y_{n,i})_{i=1}^n$ be a random sample of size n such that X is an $\Re^d$ valued random vector and Y is an $\Re$ valued random variable defined on probability space $(\Omega, \mathcal{F}, \mathcal{P})$. We assume that:

(A.0) The joint and marginal probability density of (X_n, Y_n) exist and have 2nd order derivatives. The marginal density of X in particular is independent of n and is bounded away from zero on its support.

(A.1) $(X_{n,i}, Y_{n,i})_{i=1}^{n}$ is a sequence of independent observations sampled from (X_n, Y_n). In all cases, the second order derivative of marginal densities of X's exist.

(A.2) The existence of the absolute moments of Y, that is $\exists \ \bar{Y}$ finite, such that $E(|Y_i|^k) \leq \bar{Y}$ for k > 1, $\forall i$, and

(A.3) the conditional variance of Y given x is bounded throughout the support of x.

1.3.2 Kernel Function

The estimator involves a kernel function, denoted as K, which is a scalar function of $\Re^d$. We impose the following restrictions on the selection of kernel function.

(K.1) $\|x\|^d K(x) \to 0$ as $\|x\| \to \infty$ and $\exists \bar{K}$ finite such that $|K(x)| \leq \bar{K}$ $\forall x$ in $\Re^d$ and $\int |K| < \infty$.

(K.2) $\int K = 1$, $\int x_j \ K = 0$ for all j and $\int xx^T K < \infty$.

1.3.3 Window Width

Besides the kernel function, the kernel estimator is also controlled by another parameter commonly referred to as window width or band width in some literature. The following conditions are normally imposed on the selection of the window width in order to achieve asymptotic unbiasedness and for the variance of the estimate to diminish with increasing sample size. Let h_n represents the window width for the nth observation.

(H.1) $h_n \to 0$ as $n \to \infty$ and

(H.2) $nh_n^d \to \infty$ (or $\sum_1^n h_i^d \to \infty$) as $n \to \infty$.

(H.3) $\lim_{n\to\infty} \frac{1}{n} \sum_1^n (\frac{h_n}{h_i})^\gamma < \infty$, for some $\gamma \in [d, 2d]$.

Proposition 1.3.1 *Suppose $h_n = cn^{-\alpha}$, that is, the window width for each observation decays as the sample size increases exponentially at rate α such that $\alpha \in (0, \frac{1}{\gamma})$ with $\gamma \in (d, 2d]$ and c is some finite constant, then the following conditions are satisfied.*

H.1 $h_n \to 0$ *as* $n \to \infty$.

H.2 $nh_n^d \to \infty$ *as* $n \to \infty$.

H.3 $\lim_{n\to\infty} \frac{1}{n} \sum_1^n (\frac{h_n}{h_i})^\gamma < \infty$.

Proof As $\alpha > 0$, $h_n = cn^{-\alpha} \xrightarrow{n} 0$ and hence H.1 is satisfied.

As $d\alpha < 1$, $nh_n^d = c^d n^{1-d\alpha} \to \infty$ as $n \to \infty$ thus H.2 is satisfied.

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_1^n (\frac{h_n}{h_i})^\gamma &= \lim_{n\to\infty} \frac{1}{n} \sum_1^n (\frac{i}{n})^{\alpha\gamma} \\
&= \int_0^1 x^{\alpha\gamma} dx \\
&= \frac{x^{\alpha\gamma+1}}{\alpha\gamma+1}|_0^1 \\
&= \frac{1}{\alpha\gamma+1} < 1 \quad (finite) \qquad \diamond
\end{aligned}$$

1.4 Large Sample Properties of the Estimator when Regressors are Independent Sequences

1.4.1 Decomposition of the Deviation between the Estimand and its Estimate

By algebraic manipulation we can break $m_n - m$ down into four terms to facilitate our proof.

$$\begin{aligned} m_n - m &= \frac{r_n}{f_n} - \frac{r}{f} \\ &= f_n^{-1}[r_n - \frac{f_n}{f}r] \\ &= f_n^{-1}[r_n - r - (\frac{f_n}{f}r - r)] \\ &= f_n^{-1}[r_n - r \cdot\cdot \frac{r}{f}(f_n - f)] \\ &= f_n^{-1}[r_n - r - m(f_n - f)] \\ &= f_n^{-1}[r_n - Er_n + Er_n - r - m(f_n - Ef_n + Ef_n - f)] \\ &= f_n^{-1}[A_1 + A_2 - m(A_3 + A_4)] \end{aligned}$$

Where: the stochastic component of $r_n - r$

is $A_1 = r_n - Er_n$ and EA_1^2 is the variance of r_n;

the deterministic component of $r_n - r$

is $A_2 = Er_n - r$ and A_2 is the bias of r_n;

the stochastic component of $f_n - f$

is $A_3 = f_n - Ef_n$ and EA_3^2 is the variance of f_n;

the deterministic component of $f_n - f$

is $A_4 = Ef_n - f$ and A_4 is the bias of f_n.

In terms of this setup, the estimate is asymptotically unbiased if $A_i \to 0$ as $n \to \infty$ for i=2 and 4. In section 4.2 below we will establish asymptotic unbiasedness

under the condition of the existence of second order derivatives of the data generating process (A.0) and the decay of the window width as sample size increases (H.1). A_1 and A_2 are respectively the deterministic component of $f_n - f$ and $r_n - r$ and are independent of the underlying random mechanism. Thus the results can be extended to the later chapters where the regressor is not independent.

In section 4.3, with bounded moments, (A.2) and slow enough rate of decay of window width, (H.2) we are able to show that $A_i \rightarrow 0$ in probability for i=1 and 3. The weak consistency of m_n for m follows naturally. In addition, if the rate of decay of window width satisfies the condition $lim_{n\rightarrow\infty} \sum_1^n (i^2 h_i^{-d})^{-1} < \infty$ then $A_i \rightarrow 0$ almost surely for i=1 and 3. It follows that m_n is strongly consistent for m. With the same conditions as that for weak consistency, the asymptotic distribution of the estimate is Normal and we establish the result in section 4.4.

1.4.2 Asymptotic Unbiasedness of the Kernel Estimate

By the choice of a kernel function that satisfies K.1 and K.2, and a window width that satisfies H.1 we can show that the expectation of the kernel estimate can approximate the true value up to the order of h_n^2. At this juncture we introduce some lemmas that are to be used in the proof below.

Lemma 1.4.1 *Suppose the kernel function K satisfies K.1 and K.2 with a window width satisfies H.1, and for any function $g(x)$ on $\Re^d$ that has bounded second order derivative, then the convolution of K_i and $g(x)$ approximates $g(x)$ up to the order of h_i^2 that is*

$\int_{\Re^d} h_i^{-d} K(\frac{x-u}{h_i}) g(u) du = g(x) + O(h_i^2)$

Proof

$$\int_{\Re^d} h_i^{-d} K(\frac{x-u}{h_i}) g(u) du = \int_{\Re^d} K(s) g(x - s h_i) ds$$

by transformation of variable

$$= \int_{\Re^d} K(s)[g(x) - (sh_i)g'(x)^T + sh_i g^{(2)}(x^*)(sh_i)^T]ds$$

where $x^* \in (x - sh_i, x)$,

by Taylor Expansion $= g(x) + h_i^2 \int sg^{(2)}(x^*)s^T\ K(s)\ ds$ by K.2,

up to the 2nd order, $= g(x) + O(h_i^2)$ by K.1 $\diamond$

Remarks

1. For the case where g is univariate, Parzen (1962) was the first in establishing similar results with weaker conditions that only required the continuity of g. The result was later extended by Cacoullos (1966) to multivariate g.

   Their proof is essentially based on a theorem that was first put forth by Bochner (1955). The idea of the proof is to break down the following expression into three terms,

$$\begin{aligned}\int_{\Re^d} h_i^{-d} K(\frac{x-u}{h_i})g(u)du - g(x) &= \int_{\Re^d} K(s)\{g(x - sh_i) - g(x)\}ds \\ &= \int_{|x-s|\le\xi} K(s)\{g(x - sh_i) - g(x)\}ds \\ &\quad + \int_{\{|x-s|>\xi\}} \frac{1}{h_i^d} g(s)K(\frac{x-s}{h_i})ds \\ &\quad + g(x)\int_{\{|s|>\frac{\xi}{h_i}\}} K(s)ds.\end{aligned}$$

   The first term with $|x - s| \le \xi$ for some arbitrarily small ξ; then by continuity of g the term can be made arbitrarily small.

   By K.1 the second term with $\int_{\{|x-s|>\xi\}} \frac{1}{h_i^d} g(s)K(\frac{x-s}{h_i})ds$, also goes to zero as h_i goes to zero.

   By K.2 the third term, $g(x)\int_{\{|s|>\frac{\xi}{h_i}\}} K(s)ds$, goes to zero as h_i goes to zero.

2. For the case where g might have an isolated discontinuity, then the Lebesgue Density theorem is employed instead. Some authors like Greblicki Krzyzak and

Pawlak (1984) have followed such an approach.

3. Here we have imposed a stronger condition that the existence of a 2nd order derivative of g so that we can specify the order of convergence as well.

4. Singh (1974) advanced the idea of an optimal k order kernel; for example in univariate case, the optimal k order kernel is a polynomial over the interval [-1,1] that minimizes the asymptotic IMSE and satisfies:

$$\int_{-1}^{1} K(u)u^j du = \begin{cases} 1 & \text{j=0} \\ 0 & \text{j=1,..,k-1} \\ k! & \text{j=k.} \end{cases}$$

The k order kernel is able to reduce the bias term to a higher order, $O(h^k)$. However, the bias reduction is accompanied by a larger constant in the asymptotic variance term, $\int k^2$ and the asymptotic variance is increased. Furthermore, the distribution of the weight is not positive throughout, hence the estimate could be negative for an all non-negative data set.

Unlike the case of uniform width, where the above lemma alone will be sufficient to establish asymptotic unbiasedness, we need to show as well that asymptotically, the estimate with different width for each observation will average out to yield results equivalent to that of a constant window width throughout the whole sample. Here we state without proof another lemma that gives a result analogous to the Toeplitz lemma on the convergence of a sequence of functions.

Lemma 1.4.2 *If a sequence of functions $g_n(x)$ converges to a function $g(x)$ at a point x as $n \to \infty$, then $n^{-1}\sum_{j=1}^{n} g_j(x) \to g(x)$ as $n \to \infty$.*

Now we are in position to prove the asymptotic unbiasedness of the estimators r_n and f_n. Theorem 1.4.1 was first proved by Yamato (1971) and Theorem 1.4.2 was first introduced by Ahmad and Lin (1976).

1.4.3 Consistency

Weak Consistency

Theorem 1.4.3 *Chebyshev's WLLN*

Let $E(X_i) = \mu_i$ *and* $Var(X_i) = \sigma_i^2$ *and* $Cov(X_i, X_j) = 0 \quad i \neq j$. *Then*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_1^n \sigma_i^2 = 0 \Longrightarrow \quad \bar{X}_n - \bar{\mu}_n \xrightarrow{p} \quad 0.$$

Theorem 1.4.4 *Suppose K.1, K.2, A.1, A.2 and H.2 hold, and in addition H.3 holds with* $\gamma = d$, *then* $f_n - Ef_n \xrightarrow{p} \quad 0$.

Proof Here $f_n = \frac{1}{n}\sum_1^n K_i$ and $K_i = h_i^{-d} K(\frac{x-u}{h_i})$,

so by Chebyshev's WLLN stated above it would be sufficient for us to show that

$$\begin{aligned} Var(f_n) &= Var(\frac{1}{n}\sum_1^n K_i) \\ &= \frac{1}{n^2}\sum_1^n Var(K_i) \quad \to \quad 0 \quad as\ n \to \infty. \end{aligned}$$

By Proposition 2.3.1, $\exists$ finite C_i such that $Var(K_i) = h_i^{-d} C_i$ for all i.

Let $sup_i\ h_i^d\ Var(K_i) = C < \infty$, and

$$\begin{aligned} \frac{1}{n^2}\sum_1^n Var(K_i) &\leq \frac{1}{nh_n^d}(\frac{1}{n}\sum_1^n (\frac{h_n}{h_i})^d)\ C \\ &\leq \frac{1}{nh_n^d} C' \quad \text{where} \quad C' \text{ is finite by the hypothesis above.} \\ &\to \quad 0 \quad \text{as} \quad n \to \infty. \text{ by H.2} \quad \diamond \end{aligned}$$

Theorem 1.4.5 *Suppose K.1, K.2, A.1, A.2 and H.2 hold, and in addition H.3 holds with* $\gamma = d$, *then* $r_n - Er_n \xrightarrow{p} \quad 0$.

Proof By Proposition 2.3.2, $\exists$ a finite C_i such that $Var(Y_i\ K_i) = h_i^{-d} C_i$.

Then the rest of the proof proceeds as in the proof of the above theorem. $\diamond$

Combining the results of the two theorems here together with the Theorems 1.4.1 and 1.4.2, we can conclude that m_n is at least weakly consistent for m under the stated conditions.

Theorem 1.4.6 *Suppose K.1, K.2, A.0, A.1, A.2, H.1 and H.2 hold and H.3, with* $\gamma = d$ *holds, then* $m_n - m \xrightarrow{p} 0$.

Strong Consistency

For the proof of strong consistency we will use a LLN due to Kolmogorov as stated below.

Theorem 1.4.7 *Kolmogorov's SLLN*

Let $\{X_i\}$ *be a sequence of independent random variables such that* $E(X_i) = \mu_i$ *and* $Var(X_i) = \sigma_i^2$. *Then*

$$\lim_{n\to\infty} \sum_1^n \frac{1}{i^2}\sigma_i^2 < \infty \Longrightarrow \quad \bar{X}_n - \bar{\mu}_n \xrightarrow{a.s.} 0.$$

Theorem 1.4.8 *Suppose H.2, K.1, K.2, A.1 and A.0 hold, in addition*

$$\lim_{n\to\infty} \sum_1^n (i^2 h_i^{-d})^{-1} < \infty,$$

then $f_n - Ef_n \xrightarrow{a.s.} 0$.

Proof As before, by the result of Proposition 2.3.1,

$\exists$ finite C_i such that $Var(K_i) = h_i^{-d} C_i$ for all i.

Put $C = sup_i C_i$ and obtain

$$\sum_1^n \frac{1}{i^2} Var(K_i) \leq \sum_1^n (\frac{1}{i^2 h_i^d})\, C \quad \text{finite by the hypothesis above}$$

Thus the result follows by Kolmogorov's SLLN $\diamond$

Theorem 1.4.9 *Suppose K.1, K.2, H.2, A.1 and A.2 hold, and in addition*

$$\lim_{n\to\infty}\sum_{1}^{n}(i^2h_i^{-d})^{-1} < \infty,$$

then $r_n - Er_n \xrightarrow{a.s.} 0$.

Proof Follows the same argument as in Theorem 1.4.8. ◇

Combine Theorems 1.4.5 and 1.4.6 with Theorems 1.4.1 and 1.4.2 and we can draw the following conclusion.

Theorem 1.4.10 *Suppose K.1, K.2, A.0, A.1, A.2, H.1 and H.2 hold, and furthermore*

$$\lim_{n\to\infty}\sum_{1}^{n}(i^2h_i^{-d})^{-1} < \infty,$$

then $m_n - m \xrightarrow{a.s.} 0$.

1.4.4 Asymptotic Normality

In this section we establish the limiting distribution of $T = \frac{m_n - m}{\sqrt{Var(m_n)}}$ as $n \to \infty$, which can be shown to be normal.

In section 4.1 we decompose $m_n - m$ into four components, namely

$$A_1 = r_n - Er_n, \quad A_2 = Er_n - r, \quad A_3 = f_n - Ef_n \text{ and } A_4 = Ef_n - f.$$

By Theorem 1.4.1 and 1.4.2, it is shown that A_2 and $A_4 \to 0$ as $n \to \infty$, and Theorem 1.4.5 will ensure that under the stated conditions $A_3 \xrightarrow{p} 0$.

Hence if we can establish the asymptotic normality of A_1, that of T will follow by Slutsky's theorem (Slutsky 1925).

As usual we will prove the asymptotic normality by one of the standard conditions, namely Lindeberg's or Liapunov's condition. It would be easier to establish the proof with Liapunov's condition. Anyways, if the required conditions for consistency are met they will render the two conditions equivalent. If we could show that under the

same conditions as that of Theorem 1.4.6 the following criterion ρ_n, converges to zero as $n \rightarrow \infty$, then the asymptotic normality of T_n will follow by Liapunov's theorem.

$$\rho_n = \frac{(\sum_1^n E|V_i - EV_i|^3)^{\frac{1}{3}}}{(\sum_1^n Var(V_i))^{\frac{1}{2}}}$$

where $V_i = K_i Y_i$.

Proposition 1.4.1 $\lim_{n\rightarrow\infty} \rho_n = 0$.

Proof

$$\begin{aligned} E|V_i - EV_i|^3 &\leq 2E|V_i|^3 \\ &\leq 2\int h_i^{-3d}|y|^3|K(\frac{x-u}{h_i})|^3 f(su,y)dydu \\ &\leq 2\int h_i^{-2d} E\{|y|^3|x-sh_i\}|K(s)|^3 f(x-sh_i)ds \\ &\qquad \text{by transformation of variable and } \exists \text{ finite } D_i. \\ &\leq h_i^{-2d} D_i. \end{aligned}$$

Also, by Proposition 2.3.2, $Var(V_i) = h_i^{-d} C_i$ for some finite C_i.

Therefore, ρ_n is bounded by C
$\frac{(\sum_1^n h_i^{-2d})^{\frac{1}{3}}}{(\sum_1^n h_i^{-d})^{\frac{1}{2}}}$ for some finite C. And

$$\begin{aligned} \frac{(\sum_1^n h_i^{-2d})^{\frac{1}{3}}}{(\sum_1^n h_i^{-d})^{\frac{1}{2}}} &= \frac{(\frac{1}{n}\sum_1^n(\frac{h_n}{h_i})^{2d})^{\frac{1}{3}}}{(\frac{1}{n}\sum_1^n(\frac{h_n}{h_i})^{d})^{\frac{1}{2}}} \frac{h_n^{-\frac{2}{3}d}}{h_n^{-\frac{1}{2}d}} \frac{n^{\frac{1}{3}}}{n^{\frac{1}{2}}} = O((nh_n)^{-\frac{1}{6}}) \\ &\rightarrow 0 \text{ by H.2, } nh_n \rightarrow \infty \text{ as } n \rightarrow \infty \end{aligned}$$

And the asymptotic normality of T_n will follow. We sum up these arguments in the proof of the theorem below.

Theorem 1.4.11 *Suppose the recursive kernel regression estimator satisfies the conditions of Theorem 1.4.6, then* $S_n = \frac{m_n - m}{\sqrt{Var(m_n)}} \xrightarrow{D} N(B,1)$, *where* $B = O(\sqrt{nh_n^{d+4}})$.

Proof Here we will show that $\rho_n = O(nh)^{-\frac{1}{2}}$ hence $\lim_{n\to\infty} \rho_n = 0$, as by H.2, $nh \to \infty$ as $n \to \infty$.

Furthermore, the asymptotic normality of m_n follows by Liapunov's theorem.

Here the asymptotic bias term, $m_n - m$, is of the order $O(h_n^2)$; and the asymptotic variance is of the order of $O((nh_n^d)^{-1})$. Hence, $\frac{Em_n - m}{\sqrt{Var(m_n)}}$ is of order $O(\sqrt{nh_n^{d+4}})$. $\diamond$

Remarks:

1. According to the Berry-Esséen theorem, the rate of convergence to the normal distribution in this case is dominated by ρ_n^3 and therefore is $O((nh_n)^{-\frac{1}{2}})$. In comparison with conventional parametric cases, where the rate of convergence is of order $O(n^{-\frac{1}{2}})$, the rate of convergence of the nonparametric estimator is slowed down by the factor h_n which is typical of nonparametric estimation.

2. If we were to choose width that decays at the rate $n^{-\alpha}$, that would require that $\alpha > \frac{1}{d+4}$, in order to ensure that the Bias term, B goes asymptotically to zero. However, even if B does go to zero its rate of convergence which is given by the difference between the rate of decays in bias term, $E(m_n) - m$ and that of the square root of variance will be very slow for most practical purposes.

   For most commonly encountered finite samples we cannot completely ignore the bias term altogether. Also the rate of window width decay that attains the "optimal" convergence rate in MSE is at $\alpha = \frac{1}{d+4}$ Stone (1982). Therefore estimating the bias term and incorporating its effects in our inference with the asymptotic normality theorem above could be a better approach. We are, therefore, required to take a step beyond the asymptotic bias and variance and analyse their first order approximation.

1.5 Finite Sample Properties and Rate of Weak Convergence

In this section, we will examine the approximation to the asymptotic variance and bias components of the mean square error. In sequence, we gave the optimal rate of decay for the selection of window width and the resulting rate of convergence in mean square error.

1.5.1 Approximate Asymptotic Bias

Proposition 1.5.1 *The bias of m_n can be asymptotically approximated by* $bias(m_n) = \frac{h_n^2}{2}\beta \int u\ u^T K(u)du\{r^{(2)}(x) - m(x)\ f^{(2)}(x)\}/f(x)$

where[1] $\beta = \frac{1}{n}\sum(\frac{h_i}{h_n})^2$.

Proof: Here $m_n - m = \frac{f}{f_n}\{\frac{r_n}{f} - \frac{f_n}{f}m\}$, and $f/f_n \xrightarrow{P} 1$, therefore asymptotically,

$$Bias(m_n) = \frac{1}{f}\{Bias(r_n) - n.\ Bias(f_n)\}.$$

By Lemma 1.4.1, the bias of each $h_i^{-d}K(\frac{x-X_i}{h_i})$ is approximated by $\frac{h_i^2}{2}\ f^{(2)}(x)\int u\ u^T K(u)du$. Hence

$$\begin{aligned} Bias(f_n) &= \frac{1}{n}\sum\ Bias(h_i^{-d}K(\frac{x-X_i}{h_i})) \\ &= \frac{1}{n}\sum\frac{h_i^2 f^{(2)}(x)}{2}\int u\ u^T K(u)du + o(h_n^2) \\ &= \frac{h_n^2}{2}\beta f^{(2)}(x)\int u\ u^T K(u)du \quad \text{where } \beta = \frac{1}{n}\sum(\frac{h_i}{h_n})^2. \end{aligned}$$

By the same token, the approximate bias of r_n is given by

$$\frac{h_n^2}{2}\beta r^{(2)}(x)\int u\ u^T K(u)du.$$

Hence the approximate bias of m_n is,

$$bias(m_n) = \frac{h_n^2}{2}\beta\int u\ u^T K(u)du\{r^{(2)}(x) - m(x)\ f^{(2)}(x)\}/f(x). \qquad \diamond$$

[1] Here, β is finite, provided the width decays at rate $O(n^{-\alpha})$ and α lies in the interval (0, 0.5).

1.5.2 Approximate Asymptotic Variance

Proposition 1.5.2 *The asymptotic variance of m_n can be approximated by*

$$Var(m_n) = \frac{1}{nh_n^d}\tau\frac{Var(Y|x)\int K^2}{f(x)} \qquad \text{where } \tau = \frac{1}{n}\sum(\frac{h_n}{h_i})^d.$$

Proof:

Let $V_i = Y_i K(\frac{x-X_i}{h_i})$, then

$$\begin{aligned} Var(V_i) &= E(V_i^2) - E(V_i)^2 \\ &= \int \frac{1}{h_i^{2d}} K^2(\frac{x-s}{h_i}) y^2 f(s,y)dyds \; - \; E(V_i)^2 \\ &= \int \frac{1}{h_i^d} K^2(u) y^2 f(x-uh_i, y)dydu + \; E(V_i)^2 \\ &= \frac{1}{h_i^d} Var(Y|x) f(x) \int K^2 + o(\frac{1}{h_i^d}) \qquad \diamond \end{aligned}$$

Here $r_n = \frac{1}{n}\sum V_i$, hence

$$\begin{aligned} Var(r_n) &= \frac{1}{n\; h_n^d}\,(\frac{1}{n}\sum(\frac{h_n}{h_i})^d) Var(Y|x) f(x) \int K^2 + o(\frac{1}{n\; h_n^d}) \\ &= \frac{1}{n\; h_n^d}\,\tau Var(Y|x) f(x) \int K^2 + o(\frac{1}{n\; h_n^d}) \end{aligned}$$

By $f/f_n \xrightarrow{p} 1$, the asymptotic variance of m_n can be obtained as $\frac{1}{f^2} Var(r_n)$.

Hence the approximate asymptotic variance of m_n is, $\frac{1}{n\; h_n^d}\,\tau\frac{Var(Y|x)\int K^2}{f(x)}$ $\diamond$.

From the two propositions above, we can obtain the approximate asymptotic mean square error as a function of the h's:

$$\begin{aligned} MSE \;=\; & \frac{1}{n\; h_n^d}\,\tau\frac{Var(Y|x)\int K^2}{f(x)} \\ & + \frac{h_n^4}{4}\left[\beta \int u\; u^T K(u)du\{r^{(2)}(x) - m(x)\; f^{(2)}(x)\}/f(x)\right]^2. \end{aligned}$$

1. The minimizer of the approximate MSE is,

$$h_n = \left[\frac{\tau \; d \; Var(Y|x) \int K^2}{n \; (\beta\{r^{(2)}(x) - m(x) \; f^{(2)}(x)\} \int u \; u^T K(u) du)^2} \right]^{\frac{1}{d+4}}.$$

This expression is not explicitly in term of h_n and in fact unlike the case of fixed window width there is no explicit solution for the minimizer of the approximate MSE. However, what we can learn from this expression is that the approximate rate of decay for the 'optimal' width $h_n \sim n^{-\frac{1}{d+4}}$ and the resulting rate of decay for the MSE is of the order $O(n^{-\frac{4}{d+4}})$. Hence the rate of weak convergence or rather convergence in the mean square, is of the order $O(n^{-\frac{4}{d+4}})$ which is the same as that of the fixed width estimator

2. One possible consistent estimator of the variance of m_n is

$$\hat{V}ar(m_n) = \frac{1}{n^2} \sum_1^n (\frac{1}{h_i^d}) \frac{\hat{V}ar(Y|x) \int K^2}{f_n(x)},$$

where f_n is the usual recursive estimator of density. Let the recursive kernel estimator for the $E(Y^2|x)$ be $w_n(x) = \sum K(\frac{x - X_i}{h_i}) Y_i^2$, then the $\hat{V}ar(Y|x)$ can be obtained as $w_n(x) - m_n(x)^2$. Here the consistency of the w_n can be established in the same way as the m_n except that the moment condition will be more stringent, that is $E|Y|^\gamma$ for $\gamma > 4$. With such estimator we can have an adaptive estimator of S_n and make use of the central limit theorem, at least in providing the confidence interval for an estimate.

Chapter 2

Recursive Kernel Estimation of the Regression Function when Regressors are Mixing Sequences

The object of this chapter is to extend the technique of recursive kernel . parametric regression estimation to cases in which the observations on the regressors are correlated.

The main focus here is the conditions required to establish the recursive estimator's asymptotic properties. Here we are concerned with local pointwise consistency of the estimator in the setting that we had explained in Chapter One. Re ently, Roussas (1990) has established the conditions for the recursive estimator applied to mixing sequences to be uniformly strongly consistent. Roussas (1992) went on to develop results for the estimator's exact rates of almost sure convergence using an approach resembling the law of iterated logarithms. Our results can be considered as supplementary to his. The central limit theorem that is established here for the recursive estimator is closely comparable to that of Robinson (1983) Theorem 5.3 (p.193) for the fixed width estimator applied to time series. [1]

[1] In the thesis, we consider only a real valued dependent variable. However, the analysis could be easily extend to a real valued function defined on R^q for some $q > 1$, as is the case in Roussas' paper. Whereas in Robinson's paper the dependent variable is a real valued functional of the time series X_t at some future t-value given the past observations.

We give a brief review of the literature on the estimation of the density of a strong mixing DGP in the next section. Section 2 describes the way we characterize dependency, gives inequalities associated with dependency and presents Laws of Large Number and Central Limit Theorems that are used as the basis for the proofs of asymptotic properties in Section 3. Section 4 addresses some issues pertaining to the application of asymptotic normality for inference purposes and to the rate of convergence of the estimator. As Davidson's CLT used in this chapter has yet to make its way to publication, we attach its proof as an appendix.

2.1 Survey of Recursive Kernel Regression Estimator for Dependent Observations and Related Works

Nonparametric Estimator for Correlated Observations

Robinson (1983) established the asymptotic normality of the Nadaraya-Watson type of estimator applied to a time series model, that is a model of the form: $X_{i+l} = m(X_i) + \epsilon_i$. Singh and Ullah (1985) obtained the conditions which ensure the asymptotic properties, such as consistency and asymptotic normality of the Nadaraya-Watson estimator for a stationary model: $Y_i = m(X_i) + \epsilon_i$, with $\{X_i, Y_i\}$ being a sequence of strong mixing observations. Using Georgiev (1988) version of a general nonparametric estimator for the fixed design regression, $Y_i = m(\frac{i}{n}) + \epsilon_i$, Fan (1990) proved that the estimator is consistent and has an asymptotically normal distribution for the case where $\{\epsilon_i\}$ is a mixingale sequence.

Recursive Density Estimator

For a density estimate of a sample of dependent observations, Györfi (1981) examined the behavior of the recursive estimator of univariate density where the underlying DGP satisfies the asymptotic independence condition. Marsy (1986) proved mean square convergence and asymptotic normality of recursive kernel estimators of a univariate density which is strong mixing and asymptotically uncorrelated processes. Marsy and Györfi (1987), and Marsy (1987) derived the strong consistency results for recursive univariate density estimates of strong mixing and of asymptotically uncorrelated processes.

Recursive Regression Estimator

We briefly surveyed the development of the recursive kernel regression estimator for independent observations in Chapter One. However, little attention was given to

the case where the exogenous variables are correlated and dependent which is of more relevant to econometric applications.

Recently there were two papers published by Roussas (1990, 1992) that establish the conditions for uniform strong consistency of the recursive estimator and the exact rate of almost sure convergence. Here we supplement what he did by developing results pertaining to pointwise consistency and asymptotic normality of the recursive kernel regression estimator for the situation where regressors are sampled from mixing processes.

The aim of this chapter is to address these issues with the aid of some results on CLT and LLN for mixing sequences recently published in the econometrics or statistics literature.

2.2 Definitions and Asymptotic Laws Pertaining to Mixing Sequence

In this chapter, we would like to relax the assumption (A.1) as follows:

(A.1') $(X_{n,i}, Y_{n,i})_{i=1}^{n}$ is a random sample from a strong mixing process which can be heterogenous.

Let X_i be sequence of random variables defined on probability space $(\Omega, \mathcal{F}, P)$ and $\mathcal{F}_a^b$ be the sub-sigma algebra generated by the sub-sequence$(X_j)_{j=a}^{b}$ for some integers a,b $(a \leq b)$.

Definition 2.2.1 *Let* $\alpha_m = sup_k[|P(A \cap B) - P(A)P(B)|], \quad A \in \mathcal{F}_1^k$ *and* $B \in \mathcal{F}_{k+m}^{\infty}$. *And if* α_m *goes to zero with* $m \to \infty$, *then the process is said to be strong mixing.*

Here we define a notion of *size* to describe the rate of decay of a sequence which is a simplified version of the notion put forth by McLeish (1975).

Definition 2.2.2 *We will call a sequence* $\{a_m\}$ *of size -p if* $a_m = O(m^\lambda)$ *for some* $\lambda < -p$.

In order to put such characterization of the dependency structure to work we need some useful inequalities associated with such a process.

Lemma 2.2.1 *(Ioannides and Roussas 1985) Suppose* X *and* Y *are* $\mathcal{F}_1^k$*-measureable and* $\mathcal{F}_{k+m}^\infty$*-measureable respectively, and* $E|X|^p$, $E|Y|^q < \infty$ *for some p,q such that* $\frac{1}{q} + \frac{1}{p} < 1$. *Then under strong mixing,*

$$|EXY - EXEY| \leq [\alpha_m]^{1-\frac{1}{q}-\frac{1}{p}} ||X||_p ||Y||_q.$$

Where $||X||_p = (E|X|^p)^{\frac{1}{p}}$.

Comments

1. The measure of dependency between two sub-σ -algebras of $\mathcal{F}$ by the strong mixing coefficient mentioned here was introduced by M. Rosenblatt (1956), it is also referred to as α-mixing by some authors. It is one of the commonly encountered mixing concepts in the literature. Another notion of mixing, uniform mixing was introduced by Ibragimov (1962) and Billingsley (1968) section 20 who gave a comprehensive exposition of its properties. The concept of maximal correlation was investigated by Kolmogorov and Rozanov (1960) and Castellana and Leadbetter (1986) who studied the properties of their kernel density estimator on correlated data that are characterized by such notion.

2. These three measures of dependency are equivalent when the process is Gaussian. However, in general, strong mixing is the weakest condition among the three and it usually requires more stringent moment conditions to establish similar results.

3. Any Borel measurable function of α-mixing sequences is also itself α-mixing. This is the one property of strong mixing that make proofs of functions of α-mixing sequences very convenient.

4. There is no simple link between the mixing process and the popular linear ARMA process. Withers (1981) and Athreya and Pantula (1986) gave some conditions under which an ARMA process is also mixing. From time to time there are counter-examples put forth by some authors for examples, Andrew (1984), Gorodetskii (1977), Chanda (1974) and Ibragimov and Linnik (1971) to name a few.

The notion of a mixingale was introduced by McLeish (1975) and extended by Andrews (1988).

Definition 2.2.3 *(McLeish,1975 and Andrews 1988) The sequence* $\{X_i, \mathcal{F}_i\}$ *is an* L^p*-mixingale if there exist nonnegative constants* $\{c_i : i \geq 1\}$ *and* $\{\psi_m : m \geq 0\}$ *such that* $\psi_m \to 0$ *as* $m \to \infty$ *, and we have*

$$(a) \qquad ||E(X_i|\mathcal{F}_{i-m})||_p \leq c_i\psi_m,$$

$$(b) \qquad ||X_i - E(X_i|\mathcal{F}_{i+m})||_p \leq c_i\psi_{m+1}.$$

In order to put such a characterization of dependency to use we need some useful inequalities associated with such process.

Lemma 2.2.2 *(McLeish 1975) Suppose* X *is a random variable measurable with respect to* $\mathcal{A}$*, and* $1 \leq p \leq r \leq \infty$*. Then*

$||E(X|\mathcal{F}) - EX||_p \leq 2(2^{\frac{1}{p}} + 1)\{\alpha(\mathcal{F}, \mathcal{A})\}^{\frac{1}{p}-\frac{1}{r}}||X||_r.$

Comments

1. The construct $||E(X_i|\mathcal{F}_{i-m})||_p$ can be intuitively interpreted as the forecast of the random variable X_i based on the information $\mathcal{F}_{i-m}$ at lag m and it can be expressed as a sum of a martingale difference sequence. It is the link between this construct and the martingale sequence that some authors like McLeish and Andrew exploit so the development of proofs about mixingale sequences can utilitize the established results on martingale difference sequences.

2. The Lemma 2.2.2 is important because that it links an α- mixing sequence to a mixingale sequence and consequently permits one to make use of the established results on mixingale sequence for the development of proofs about an α-mixing sequence.

One aim of this thesis is to investigate the properties of the Ahmad and Lin (1976) recursive estimator in the case where the underlying DGP is strong mixing. Our strategy of proof is to make use of the recently published results pertaining to CLT and LLN's of a strong mixing process and make the proof look similar to those for the independent cases.

Theorem 2.2.1 *(McLeish 1975) Suppose* $\{Z_i,\mathcal{F}_i\}$ *is a* L^2*-mixingale with* $\{\psi_m\}$ *of size* $-\frac{1}{2}$*, and constant* $\{c_i\}$ *such that*
$\sum_1^\infty \frac{c_i^2}{i^2}$ *is bounded then* $\frac{1}{n}\sum_1^n Z_i \to 0$ *almost surely.*

For our purpose in establishing the asymptotic normality of $r_n - r$ we make use of a CLT for heterogeneous, strong mixing sequences according to Davidson (1990). This theorem is an extension of McLeish (1974)'s CLT for martingale to CLT for strong mixing sequences by the Bernstein sums method. It relaxes the second order stationary condition that was imposed by most authors, including Billingsley (1968) and Serfling (1968). Here we will state the theorem below for reference in our proof.

Theorem 2.2.2 *(Davidson, 1990) Suppose a sequence of random variables* $\{Z_i\}$ *on a probability space* $(\Omega, \mathcal{F}, \mathcal{P})$ *satisfies the following conditions:*

(a) $E(Z_i) = 0$ *for all* i,

(b) there exists a sequence of positive constants $\{c_i\}$ *such that* $\{\tilde{z}_i\}$ *is uniformly* L_r *bounded*[2] *for* $r > 2$, *where* $\tilde{z}_i = \frac{Z_i}{c_i}$;

(c) the sequence is strong mixing with α_m *of size* $-\frac{r}{r-2}$;

(d) $\sup_n \frac{nM_n^2}{s_n^2} < \infty$, *where* $M_n = \max_{1 \le i \le n} c_i$ *where* $s_n^2 = E(\sum_1^n Z_i)^2$ *and* $S_n = \sum_1^n \frac{Z_i}{s_n}$.

Then $S_n \xrightarrow{\mathcal{D}} N(0,1)$.

In order obtain results on the asymptotic properties of our estimator, some intermediate results, not formally shown elsewhere, will be proved in section below.

2.3 The Bound of the Moments of the Kernel Functions

In the following two propositions we give the bound of the moments of K_i and V_i, where $V_i = Y_i K_i$ and $K_i = K(\frac{x - X_i}{h_i})$, which will be required for the proofs in the next section.

Proposition 2.3.1 *Under the conditions that K.1, K.2, A.0, and H.1 hold, then*

(i) $E(K_i) = f(x) + O(h_i^2) = O(1)$

(ii) $Var(K_i) = O(h_i^{-d})$, *and*

(iii) $||K_i||_r = O(h_i^{-(1-\frac{1}{r})d})$.

Proof (i) Under the given conditions the assumptions of Lemma 1.4.1 are satisfied and hence by Lemma 1.4.1, $EK_i = f(x) + O(h_i^2)$.

[2]Here, Z is L_r bounded means $\{E|Z|^r\}^{\frac{1}{r}}$ is bounded.

By H.1 as n $\to \infty$ $h_i \to 0$, so the claim of the proposition follows. $\diamond$

(ii) $\quad Var(K_i) = E(K_i^2) - [E(K_i)]^2$

$$\begin{aligned} Var(K_i) &= \int h_i^{-2d} K(\frac{x-u}{h_i})^2 f(u)du \ + \ O(1) \\ &= \int h_i^{-d} K(s)^2 f(x - sh_i)ds \ + \ O(1) \\ &= h_i^{-d}[f(x)\int K^2 + O(h_i^2)] \ + \ O(1) \\ &= O(h_i^{-d}) \quad \diamond \end{aligned}$$

$$\begin{aligned} \text{(iii)} \quad E|K_i|^r &= \frac{1}{h_i^{d(r-1)}} \int h_i^{-d} |K(\frac{x-u}{h_i})|^r f(u)du \\ &= h_i^{-d(r-1)} \int |K(s)|^r f(x - sh_i)ds \\ &= h_i^{-d(r-1)}[f(x)\int |K|^r + O(h_i^2)] \\ &= O(h_i^{-d(r-1)}) \end{aligned}$$

$$\begin{aligned} \text{Hence} \quad ||K_i||_r &= (E|K_i|^r)^{\frac{1}{r}} \\ &= O(h_i^{-d(1-\frac{1}{r})}) \quad \diamond \end{aligned}$$

Corollary 2.3.1 *Under the same conditions as in the above proposition, if in addition A.1' is applied, then*

$Cov(K_i,\ K_j) \ \leq \ O(\alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})-1}\ h_j^{-d(1-\frac{1}{r})-1}\ )$.

Proof

$$\begin{aligned} \text{By Lemma 2.2.1,} \quad Cov(K_i,\ K_j) &\leq |E(K_i\ K_j) \ - \ E(K_i)E(K-j)| \\ &\leq \alpha_{|i-j|}^{1-\frac{2}{r}}\ ||K_i||_r ||K_j||_r \\ &\leq O(\alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})-1}\ h_j^{-d(1-\frac{1}{r})-1}\ ) \quad \diamond \end{aligned}$$

Proposition 2.3.2 *Under the conditions that K.1, K.2, A.0, A.2 and H.1 hold, then*

(i) $\quad E(V_i) = r(x) + O(h_i^2) = O(1)$

(ii) $Var(V_i) = O(h_i^{-d})$; *and*

(iii) $||V_i||_r = O(h_i^{-(1-\frac{1}{r})d})$.

Proof (i) Under the given conditions the assumptions of Lemma 1.4.1 are satisfied and hence by Lemma 1.4.1, $EV_i = r(x) + O(h_i^2)$.

By H.1 as n $\to \infty$ $h_i \to 0$, so the claim of the proposition follows. $\diamond$

(ii) $Var(V_i) = E((Y_i\ K_i)^2) - [E(Y_i\ K_i)]^2$

$$\begin{aligned} Var(Y_i\ K_i) &= \int\int h_i^{-2d} y^2 K(\frac{x-u}{h_i})^2 f(u,y)dydu \ + \ O(1) \\ &= \int h_i^{-d} K(s)^2 E(Y^2|x - sh_i) f(x - sh_i)ds \ + \ O(1) \\ &= h_i^{-d}[f(x)E(Y^2|x)\int K^2 + O(h_i^2)] \ + \ O(1) \\ &= O(h_i^{-d}) \qquad \diamond \end{aligned}$$

$$\begin{aligned} \text{(iii)} \qquad E|V_i|^r &= \frac{1}{h_i^{d(r-1)}} \int\int h_i^{-d} |K(\frac{x-u}{h_i})|^r |y|^r f(u,y)dydu \\ &= h_i^{-d(r-1)} \int |K(s)|^r E(|Y|^r|x - sh_i) f(x - sh_i)ds \\ &= h_i^{-d(r-1)}[f(x)E(|Y|^r|x)\int |K|^r + O(h_i^2)] \\ &= O(h_i^{-d(r-1)}) \end{aligned}$$

$$\begin{aligned} \text{Hence} \quad ||V_i||_r &= (E|V_i|^r)^{\frac{1}{r}} \\ &= O(h_i^{-d(1-\frac{1}{r})}) \qquad \diamond \end{aligned}$$

Corollary 2.3.2 *Under the same conditions as in the above proposition, if in addition A.1' is applied, then*

$Cov(V_i,\ V_j) \ \leq \ O(\alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})-1}\ h_j^{-d(1-\frac{1}{r})-1}\ )$.

Proof Same as in the corollary above. $\diamond$

2.4 Asymptotic Properties of the Recursive Regression Estimator for Strong Mixing Observations

2.4.1 Weak Consistency

Theorem 2.4.1 *Suppose the recursive kernel density estimator f_n with kernel function satisfying K.1 and K.2; and a window width satisfying H.1, H.2 and H.3 with $\gamma=2d(1-\frac{1}{r})$, $r > 2$; and that the underlying DGP satisfies A.0 and A.1' such that α_m is of size $-\frac{r}{r-2}$ $r > 2$, then $f_n - f \to 0$ in probability.*

Proof As the assumptions K.1, K.2 and H.1 are in force the asymptotic unbiasedness follows by Theorem 1.4.1.

It is, therefore, sufficient for us to demonstrate that the variance of the estimator converges to zero as n $\to \infty$ and it follows that the estimator is consistent in mean square.

In sequence the result follows by Chebyshev's inequality.

Together with A.0 and the assumptions on window width, it has been shown in Proposition 2.3.1 that

$Var(K_i) = O(h_i^{-d})$ and in general $||K_i||_r = O(h_i^{-d(1-\frac{1}{r})})$.

The fact that α_m is of size $-\frac{r}{r-2}$ $r > 2$ means $\sum_1^\infty \alpha_n^{1-\frac{2}{r}}$ is bounded.

Hence together with Lemma 2.2.2,

$$\begin{aligned} Var(f_n) &= Var(\frac{1}{n}\sum_1^n K_i) \\ &= \frac{1}{n^2}[\sum_i Var(K_i) + \sum\sum_{i\neq j} Cov(K_i, K_j)] \\ &\leq \frac{1}{n^2}[\sum_i C_i\, h_i^{-d} + \sum\sum_{i\neq j} D_{ij}\ \alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})}\ h_j^{-d(1-\frac{1}{r})}] \end{aligned}$$

Where C_i and D_{ij} are constants independent of n.

Rearrange the terms to obtain

$$\begin{aligned} Var(f_n) &\leq \frac{1}{n^2}[\sum_i C_i h_i^{-d} + 2\sum_{i=1}^{n} h_i^{-2d(1-\frac{1}{r})} \sum_{j<i} D_{ij}\ \alpha_{|i-j|}^{1-\frac{2}{r}}\ (\frac{h_j}{h_i})^{-d(1-\frac{1}{r})}] \\ &= O(\frac{1}{n^2}\sum_i h_i^{-\gamma}) \\ &= O\left(\frac{1}{nh_n^{\gamma}}(\frac{1}{n}\sum_i(\frac{h_n}{h_i})^{\gamma})\right) \rightarrow 0 \end{aligned}$$

where $\gamma = 2d(1-\frac{1}{r})$ for $r > 2$ and by H.2 $nh_n \rightarrow \infty$ as $n \rightarrow \infty$.

Hence the result follows. $\diamond$

The condition imposed on the window width is slightly more stringent than in the case of independent oberservations. It amounts to $h_n = Cn^{-\frac{1}{\gamma}}$ where C is a constant and γ lies between d to 2d. That is, the window width is decaying at a slower rate due to the correlation between observations which necessarily means that the bias is larger than that of the independent case.

As before, the moment of K_iY_i grows with h_i at the same order as K_i under the assumptions A.2 and A.0. Therefore the same arguments will serve as a proof of weak convergency of r_n and, in sequence, that of m_n as well. We will make the statement more formally below.

Theorem 2.4.2 *Suppose the recursive kernel estimator r_n satisfies the conditions in Theorem 2.4.1 in addition A.2 is satisfied, then $r_n - r \rightarrow 0$ in probability.*

Proof As mentioned before, it is sufficient to show that the variance of the estimator converges to zero as n $\rightarrow \infty$.

Put $V_i = K_iY_i$. It is shown in the appendix that $Var(V_i) = O(h_i^{-d})$ and $||V_i||_r = O(h_i^{-d(1-\frac{1}{r})})$.

It follows that

$$Var(r_n) = Var(\frac{1}{n}\sum_1^n V_i)$$

$$
\begin{aligned}
&= \frac{1}{n^2}[\sum_i Var(V_i) + \sum\sum_{i\neq j} Cov(V_i, V_j)] \\
&\leq \frac{1}{n^2}[\sum_i C_i h_i^{-d} + 2\sum_{i=1}^{n} h_i^{-2d(1-\frac{1}{r})} \sum_{j<i} D_{ij}\ \alpha_{|i-j|}^{1-\frac{2}{r}}\ (\frac{h_j}{h_i})^{-d(1-\frac{1}{r})}] \\
&= O(\frac{1}{n^2}\sum_i h_i^{-\gamma}) \\
&= O(\frac{1}{nh_n^{\gamma}}(\frac{1}{n}\sum_i (\frac{h_n}{h_i})^{\gamma})) \\
&\to 0 \quad as\ n \to \infty \quad \diamond
\end{aligned}
$$

Where C_i, D_{ij} are constants independent of n and $\gamma = 2d(1-\frac{1}{r})$ for $r > 2$.

Theorem 2.4.3 *Suppose the recursive kernel regression estimator m_n satisfies the conditions in Theorem 2.4.2, then $m_n - m \to 0$ in probability.*

Proof As discussed in Section 1.4.1 the combined result of Theorem 2.4.1 and 2.4.2 above will be sufficient to establish the claim. $\diamond$

2.4.2 Strong consistency

As in the case of independent regressors, we will use SLLN, Theorem 1.2.1 in our proof of strong consistency of recursive kernel regression estimator on dependent observations.

Theorem 2.4.4 *Suppose the recursive kernel density estimator f_n with kernel function satisfies K.1 and K.2; and that the window width satisfies H.1 and H.2 such that $\lim_{n\to\infty} \sum_1^n (ih_i^{d(1-\frac{1}{r})})^{-2}$ is bounded, where $r > 2$; and also that the underlying DGP satisfies A.0 and A.1' such that α_m is of size $-\frac{r}{r-2}$ $r > 2$,*
then $f_n - f \to 0$ almost surely.

Proof Let $Z_i = K_i - E(K_i)$ and
$\mathcal{F}_i = \sigma(\{X_j\}_{j=1}^{i})$ for i$\geq$ 1 and $\mathcal{F}_i = \{\emptyset, \Omega\}$ for i < 1.

Then we have $||EZ_i|\mathcal{F}_{i+m}||_2 = 0$ and, by Lemma 2.2.2 above,

$$||EZ_i|\mathcal{F}_{i-m}||_2 \leq 2(2^{\frac{1}{2}}+1)\alpha_m^{\frac{1}{2}-\frac{1}{r}}||EZ_i||_r \quad \text{for} \quad r>2$$

Hence $\{Z_i,\mathcal{F}_i\}$ is a L^2-mixingale with $c_i=||Z_i||_r$ and $\psi_m=5\alpha^{\frac{1}{2}-\frac{1}{r}}$.

As α_m is of size $-\frac{r}{r-2}$ $r>2$ it follows that ψ_m is of size $-\frac{1}{2}$.

Furthermore, $||Z_i||_r \leq 2||K_i||_r$ and by assumption A.0 it is shown in Proposition 2.3.1 that

$||K_i||_r = O(h_i^{-2d(1-\frac{1}{r})})$.

So $\exists$ a finite C such that the constant of the mixingale

$c_i = ||Z_i||_r \leq C\ h_i^{-2d(1-\frac{1}{r})}$ for all i and

$\sum_1^n \frac{c_i^2}{i^2} \leq C\sum_1^n i^{-2}h_i^{-2d(1-\frac{1}{r})}$ is bounded as n $\to \infty$ by the assumption stated above.

Therefore the conditions of Theorem 2.2.1 are fulfilled and we have $\frac{1}{n}\sum_1^n Z_i = f_n - Ef_n$ $\to 0$ almost surely.

Also the conditions imposed here include that of Theorem 1.4.1. Therefore the result of Theorem 1.4.1 is in force.

Hence the conclusion the of theorem follows. $\diamond$

By the same token we can demonstrate the strong consistency of the $r(x)$ estimator and in sequence that of the regression estimator.

Theorem 2.4.5 *Suppose the recursive kernel estimator r_n satisfies the conditions of Theorem 2.4.4 and Assumption A.2' then $r_n - r \to 0$ almost surely.*

Proof Let $Z_i = K_iY_i - EK_iY_i$ and

$\mathcal{F}_i = \sigma(\{X_j,Y_j\}_{j=1}^i)$ for i$\geq$ 1 and $\mathcal{F}_i = \{\emptyset,\Omega\}$ for i $<$ 1

Then $||EZ_i|\mathcal{F}_{i+m}||_2 = |EZ_i| = 0$ and by Lemma 2.2.2 above

$||EZ_i|\mathcal{F}_{i-m}||_2 \leq 2(2^{\frac{1}{2}}+1)\alpha_m^{\frac{1}{2}-\frac{1}{r}}||EZ_i||_r$ for r $>$ 2.

Hence $Z_i,\mathcal{F}_i$ is a mixingale sequence with ψ_m of size $-\frac{1}{2}$ and $c_i = ||Z_i||_r$.

By A.0 and A.2, it is shown in Proposition 2.3.1 that $||Z_i||_r = O(h_i^{-d(1-\frac{1}{r})})$.
Therefore, $\sum_1^\infty \frac{c_i^2}{i^2}$ is bounded. It follows that $\frac{1}{n}\sum_1^n Z_i = r_n - r \to 0$ *a.s.* by Theorem 2.2.1 $\diamond$

These two theorems allows us to draw the following conclusion about the strong consistency of the recursive kernel regression estimator applies to samples with regressors generated from mixing processes.

Theorem 2.4.6 *Supvose the recursive kernel regression estimator satisfies the conditions in Th:orem 2.4.5,*
then $m_n - m \to 0$ *almost surely.*

Asymptotic Normality

As we have discussed in the previous section, our main aim is to establish the asymptotic normality of $r_n - r$; then that of $m_n - m$ will follow by Slutsky's theorem.

Theorem 2.4.7 *Suppose the recursive kernel estimator of* $r(x)$ *with a kernel function that satisfies K.1 and K.2; and a window width that satisfies H.1, H.2 and H.3 for some* γ *between d to 2d; and that the underlying DGP satisfies A.0, A.1' and A.2 such that* α_m *is of size* $-\frac{r}{r-2}$ $r > 2$,
then $S_n = \frac{r_n - Er_n}{\sqrt{Var(r_n)}} \xrightarrow{\mathcal{D}} N(0,1)$.

Proof Let $Z_i = K_iY_i - EK_iY_i$ and
$\mathcal{F}_i = \sigma(\{X_j, Y_j\}_{j=1}^i)$ for i$\geq$ 1 and $\mathcal{F}_i = \{\emptyset, \Omega\}$ for i < 1
Then $EZ_i = 0$ and part (a) of the above theorem is trivially satisfied.
Z_i is a measurable function of strong mixing sequences, and therefore, it is a strong mixing sequence of the same size. Hence part (b) of Davidson's theorem is satisfied.
Let $c_i = h_i^{-d(1-\frac{1}{r})}$, then

$$\tilde{z}_i = \frac{Z_i}{c_i}$$

$$= h_i^d(Y_iK_i - EK_iY_i)$$
$$= Y_iK(\frac{x - X_i}{h_i}) - E(Y_iK(\frac{x - X_i}{h_i}))$$

Clearly $E|\tilde{z}_i|^r \leq 2E|Y_iK(\frac{x-X_i}{h_i})|^r$, and by assumptions A.2 and K.1, $||Y_iK(\frac{x-X_i}{h_i})||_r^2 \leq ||Y_i||_{2r}||K(\frac{x-X_i}{h_i})||_{2r}$ is always bounded. Therefore condition (c) is satisfied.

$$s_n^2 = E(\sum_1^n Z_i)^2$$
$$= \sum_1^n EZ_i^2 + 2\sum_{i=1}^n\sum_{j=1}^{i-1} EZ_iZ_j$$
by Proposition 2.3.2, and V_i and $C_{i,i-j}$ are some finite constants
$$\sim \sum_1^n h_i^{-d}V_i + 2\sum_1^n\sum_1^{i-1}(h_ih_{i-j})^{-d(1-\frac{1}{r})}\alpha_{i,i-j}^{1-\frac{2}{r}}C_{i,i-j} \quad for\ r > 2$$
$$= O(\sum_1^n h_i^{-\gamma})$$

where $\gamma = 2d(1 - \frac{1}{r})$.

Here we make use of the fact that $\{\alpha_m\}$ is of size $-\frac{r}{r-2}$ which means $\sum_1^\infty \alpha_m^{\frac{r-2}{r}}$ is bounded; and $\frac{h_i}{h_{i-j}} \leq 1$ for all i and j.

Also $M_n = \max_{i\leq n} c_i = O(h_n^{-d(1-\frac{1}{r})})$, therefore

$$\frac{nM_n^2}{s_n^2} \sim \frac{nh_n^{-d(1-\frac{1}{r})}}{\sum_1^n h_i^{-\gamma}}$$
$$= h_n^{\gamma-d(1-\frac{1}{r})}[\frac{1}{n}\sum_1^n(\frac{h_n}{h_i})^\gamma]^{-1}$$
$$< \infty.$$

Now, part (d) of the above theorem is also satisfied.

Therefore, the conclusion of the theorem follows by Davidson's CLT.

Following the discussion in Section 4.5 of last chapter, we are going to extend the result to the asymptotic normality of $\frac{m_n - m}{\sqrt{Var(m_n)}}$, the standardized recursive kernel regression estimator.

Theorem 2.4.8 *Suppose the recursive kernel regression estimator satisfies the conditions in Theorem 2.4.7,*

then $T_n = \frac{m_n - m}{\sqrt{Var(m_n)}} \xrightarrow{\mathcal{D}} N(B,1)$, *for* $B = O(\sqrt{nh_n^{\gamma+4}})$, *where* $\gamma = 2d(1-\frac{1}{r})$ *for* $r > 2$.

Proof Under the given conditions, we have

Theorem 2.4.1 $f_n - f \xrightarrow{P} 0$. Hence,

$$\begin{aligned} T_n &= \frac{(r_n - Er_n)\frac{1}{f}}{\sqrt{\frac{1}{f^2}Var(r_n)}} + \frac{Em_n - m}{\sqrt{Var(m_n)}} \\ &\xrightarrow{\mathcal{D}} S_n + B \\ &\xrightarrow{\mathcal{D}} N(B,1) \text{ by Theorem 2.4.7} \end{aligned}$$

The Bias term is not affected by the stochastic process of the regresssor and it will be exactly the same as in the case of independent regressors. Here the asymptotic bias term, $m_n - m$, is of the order $O(h_n^2)$; and the asymptotic variance is of the order of $O(nh_n^{\gamma})^{-1})$. Hence, $\frac{Em_n - m}{\sqrt{Var(m_n)}}$ is of the order of $O(\sqrt{nh_n^{\gamma+4}})$. $\diamond$

2.4.3 Rate of Weak Convergence and Selection of Optimal Window Width

In order to derive the rate of convergence we need to obtain an approximation to the local mean square error (MSE). However, as we have demonstrated in Chapter 1, we can only obtain an expression for the first order approximation up to a constant multiplicative factor and we are therefore only concerned with the rate of convergence.

Theorem 2.4.9 *The local MSE=*$O(h_n^4 + \frac{1}{nh_n^{\gamma}})$ *where* γ*=d for the case in which regressors are sampled from independent processes, and*

γ*=*$2d(1-\frac{1}{r})$ *when regressors are sampled from* α*-mixing processes of size* $-\frac{r}{r-2}$, $r > 2$.

Proof Here $Var(m_n) = O((nh_n^{\gamma})^{-1})$,

for regressors sampled from independent processes, $\gamma = d$ as shown in the proof of

Theorem 1.4.3 and 1.4.4.

For regressors which are from an α-mixing sequence, $\gamma = 2d(1 - \frac{1}{r})$ as shown in the proof of Theorem 2.4.1 and 2.4.2.

By Theorems 1.4.1 and 1.4.2, $Bias(m_n) = O(h_n^2)$.

This implies that

$$\begin{aligned} MSE &= Variance + Bias^2 \\ &= O((nh_n^\gamma)^{-1} + h_n^4). \quad \diamond \end{aligned}$$

With the above results we can make the following claim.

Proposition 2.4.1 *The optimal rate of decay of window width for recursive regression estimator m_n is given by $h_n = cn^{-\frac{1}{4+\gamma}}$ where $\gamma = d$ for the independent case and $\gamma = 2d(1 - \frac{1}{r})$ for the α-mixing case.*

The corresponding rate of weak convergence is $O(n^{-\frac{4}{4+\gamma}})$.

Proof In the approximate MSE, the bias term decreases as h_n decreases whereas the variance increases as h_n decreases for constant n.

Thus the approximate MSE is minimized when the two terms decrease at the same rate as n increases.

$$\begin{aligned} (nh_n^\gamma)^{-1} &\sim h_n^4 \\ h_n &\sim n^{-\frac{1}{4+\gamma}} \\ MSE &= O(n^{-\frac{4}{4+\gamma}}). \quad \diamond \end{aligned}$$

For the case of regressors generated by independent processes the rate has achieved the optimal rate given by Stone's (1982) theorem.

Appendix

In our proof of the estimator's asymptotic normality we invoke a theorem due to Davidson which is not yet published, so we reproduce the proof of the theorem below. Davidson's CLT is an extension of McLeish (1974)'s CLT:

Theorem 2.4.10 *(McLeish 1974) Suppose* $\{Z_{n,i}, i = 1, .., r_n\}$ *is an array of random variables on probability space* $(\Omega, \mathcal{F}, P)$. *For each* n, r_n *define a random variable* $T_n = \prod_1^{r_n}(1 + it\ Z_{n,i})$, *for all real* t,

(a) $\max_{j \le r_n} |Z_{n,j}| \xrightarrow{p} 0$;

(b) $\sum_{j=1}^{r_n} Z_{n,j}^2 \xrightarrow{p} 1$;

(c) $\{T_n\}$ *is uniformly integrable;*

(d) $E(T_n) \to 1$.

Then $S_n = \sum_i^{r_n} Z_{n,j} \xrightarrow{\mathcal{D}} N(0,1)$.

In the same paper, he (McLeish) also established that the conditions (a) and (c) above are the consequence of the Lindeberg condition:

$$\sum Z_{n,i}^2 I(|Z_{n,i}| > \epsilon) \xrightarrow{p} 0.$$

Proof:

(1) By $P\{\max_i |Z_{n,i}| > \epsilon\} = P\{\sum_i Z_{n,i}^2 I(|Z_{n,i}| > \epsilon) > \epsilon^2\}$, (a) is satisfied.

(2) The above condition implies that $\max_{i \le r_n} |Z_{n,i}|$ is uniformly bounded in L_2 norm, since for all $\epsilon > 0$, $\max_i Z_{n,i}^2 \le \epsilon^2 + \sum_1^{r_n} Z_{n,i}^2 I(|Z_{n,i}| > \epsilon)$, and the second term converges in probability to zero.

The boundedness of $Z_{n,i}$ will imply the uniform integrability of $\{T_n\}$ with the proof as follow;

Define an array $\tilde{z}_{n,i} = Z_{n,i} I(\sum_1^{i-1} Z_{n,i}^2 \le 2)$,

then $P\{\ \tilde{z}_{n,i} \neq Z_{n,i}\ for\ some\ i \le r_n\} \le P\{\ \sum Z_{n,i}^2 > 2\} \to 0$ as $n \to \infty$.

$$J_n = \begin{cases} min\{j \le r_n; \sum_1^j Z_{n,i}^2 > 2\} & \sum Z_{n,i}^2 > 2 \\ r_n & \text{otherwise} \end{cases}$$

$$\begin{aligned} E|T_n|^2 &= E\prod_1^{r_n}(1 + t^2 \tilde{z}_{n,j}^2) \\ &\le E\ exp(t^2 \sum_1^{J_n - 1} Z_{n,j}^2)(1 + t^2 Z_{n,J_n}^2) \\ &\le e^{2t^2}(1 + t^2 E Z_{n,J_n}^2). \end{aligned}$$

Which is uniformly bounded by because $Z_{n,j}^2$ is.◇

The proof of the theorem below will then establish the links between its assumptions with that of the above, that is the assumptions A.1 to A.4 below satisfy the Lindeberg condition, conditions (b) and (d) above.

Theorem 2.4.11 *(Davidson, 1990) Suppose a sequence of random variables $\{X_i\}$ on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ satisfies the following conditions:*

(A.1) $E(X_i) = 0$ for all i,

(A.2) there exists a sequence of positive constants $\{c_i\}$ such that $\{\tilde{X}_i\}$ is uniformly L_r bounded for $r > 2$, where $\tilde{X}_i = \frac{X_i}{c_i}$;

(A.3) the sequence is strong mixing with α_m of size $-\frac{r}{r-2}$;

(A.4) $\sup_n \frac{nM_n^2}{s_n^2} < \infty$, *where* $M_n = \max_{1\le i\le n} c_i$, $s_n^2 = E(\sum_1^n X_t)^2$ *and* $S_n = \sum_1^n \frac{X_t}{s_n}$.

Then $S_n \xrightarrow{\mathcal{D}} N(0,1)$.

Here we define the constructs typical of the Bernstein approach, $Z_{n,i} = s_n^{-1} \sum_{b_n(i-1)+1}^{ib_n} X_t$, for $i = 1, .., r_n$, where $b_n = [n^{1-\alpha}]$ for $\alpha \in (0,1)$.

So $\sum_1^n X_t = s_n \sum_1^{r_n} Z_{n,i}$+ remainder terms, and the remainder term is asymptotically negligible as $b_n\, r_n/n \to 1$.

Proposition 2.4.2 *Under assumptions A.1-A.4,* $\{Z_{n,i}\}$ *satisfies the Lindeberg condition.*

Proof: Let $W_{n,i} = b_n^{-\frac{1}{2}} \sum_{b_n(i-1)+1}^{ib_n} \tilde{X}_t$, for some $B \ge 0$, the event

$E_i(B) = \{\omega : |W_{n,i}(\omega)| > B\}$.

By the definition of $W_{n,i}$,

$$\begin{aligned} E\{I\{E_i(B)\}W_{n,i}^2\} &\le \frac{1}{b_n} \sum_{b_n(i-1)+1}^{ib_n} (E(I\{E_i(B)\}\tilde{X}_t^2 \\ &+ 2 \sum_{m=1}^{t-(i-1)b_n-1} |E(I\{E_i(B)\}\tilde{X}_t\tilde{X}_{t-m})|) \\ &\le \max_{b_n(i-1)+1\le t\le b_n i} (E(I\{E_i(B)\}\tilde{X}_t^2 \\ &+ 2 \sum_{m=1}^{t-(i-1)b_n-1} |E(I\{E_i(B)\}\tilde{X}_t\tilde{X}_{t-m})|) \end{aligned}$$

Here, by A.2 the $\{\tilde{X}_t^2\}$ is uniformly integrable and also by A.3 and the inequality associated with strong mixing due to McLeish, $b_n^{-1}E(\sum_{b_n(i-1)+1}^{ib_n} \tilde{X}_t)^2$ is bounded uniformly in n and i. It follows that

$$\sup_n \max_{1\le i\le r_n} (\sup_t E(I\{E_i(B)\}\tilde{X}_t^2)) \to 0 \quad \text{as} \quad B \to \infty.$$

Similarly, by modulus inequality,

$$\begin{aligned} |E(I\{E_i(B)\}\tilde{X}_t\tilde{X}_{t-m})| &\leq E|I\{E_i(B)\}\tilde{X}_t\tilde{X}_{t-m}| \\ &\leq ||I\{E_i(B)\}\tilde{X}_t||_2||I\{E_i(B)\}\tilde{X}_{t-m}||_2 \end{aligned}$$

we can also establish that the second term in the above expression goes to zero uniformly in i, n as $B \to \infty$.

Hence the uniformly integrability of $\{W_{n,i}^2\}$ is established, by the fact that,

$$\begin{aligned} Z_{n,i} &= s_n^{-1} \sum_{b_n(i-1)+1}^{ib_n} X_t \\ &= s_n^{-1} \sum_{b_n(i-1)+1}^{ib_n} c_t\tilde{X}_t \\ &\leq s_n^{-1}M_n \sum_{b_n(i-1)+1}^{ib_n} \tilde{X}_t \\ &\leq s_n^{-1}M_n b_n^{\frac{1}{2}} W_{n,i}. \end{aligned}$$

Hence, $Z_n^2 \leq s_n^{-2}M_n^2 b_n W_{n,i}^2 \sim \frac{W_{n,i}^2}{r_n}$ and the uniform integrability of $\{Z_n^2\}$ follows from that of $\{W_{n,i}^2\}$ $\diamond$

Proposition 2.4.3 *Under the conditions A.1 to A.4, the condition (d) is satisfied, that is*

$$E(T_n) \to 1.$$

Proof: Here

$$\begin{aligned} T_n &= \prod_1^{r_n}(1 + i\lambda Z_{n,i}) \\ &= 1 + i\lambda \sum_{i=1}^{r_n} T_{i-1}Z_{n,i}. \end{aligned}$$

T_i is a $\mathcal{F}_{ib_n}$-measurable random variable. Hence

$$E(T_n) = 1 + i\lambda \sum_{i=1}^{r_n} E(T_{i-1}Z_{n,i})$$

$$= \quad 1 + i\lambda \sum_{i=1}^{r_n} E(T_{i-1}\ E(Z_{n,i}|\mathcal{F}_{(i-1)b_n}))$$

Here, $\sum_{i=1}^{r_n} E(T_{i-1}\ E(Z_{n,i}|\mathcal{F}_{(i-1)b_n})) \leq \sum_{1}^{r_n} ||T_{i-1}^2||_2 ||E(Z_{n,i}|\mathcal{F}_{(i-1)b_n})||_2$, and the uniformly boundedness of the $||T_{i-1}||_2$ follows the Lindeberg condition as we had shown in the above.

So $E(T_n) \to 1$ if $\sum_{1}^{r_n} ||E(Z_{n,i}|\mathcal{F}_{(i-1)b_n})||_2 \to 0$.

By assumption A.3, we can establish that the sequence $\{X_t\}$ is a mixingale of size $\frac{1}{2}$; that is

$$||E(X_t|\mathcal{F}_{(i-1)b_n})||_2 \leq c_t\zeta_m \quad \text{where } \max_{1\leq t\leq n} c_t \leq 5KM_n$$

$K = sup\tilde{X}$, M_n is as defined in A.4. Also $\zeta_m = O(m^{-(\frac{1}{2}+\mu)})$ for $\mu > 0$.

$$\begin{aligned}
||E(Z_{n,i} \quad | \quad \mathcal{F}_{(i-1)b_n})||_2 &= s_n^{-1}\{E[\sum_{b_n(i-1)+1}^{ib_n} E(X_t|\mathcal{F}_{(i-1)b_n})]^2\}^{\frac{1}{2}} \\
&\leq \quad s_n^{-1}\{\sum_{b_n(i-1)+1}^{ib_n} E[E(X_t|\mathcal{F}_{(i-1)b_n})]^2 \\
&+ \quad 2\sum_{b_n(i-1)+2}^{ib_n} \sum_{1}^{t-(i-1)b_n-1} |E[E(X_t|\mathcal{F}_{(i-1)b_n})E(X_{t-m}|\mathcal{F}_{(i-1)b_n})]|\}^{\frac{1}{2}} \\
&\leq \quad 5s_n^{-1}M_nK\{\sum_{1}^{b_n}\zeta_t^2 + 2\sum_{2}^{b_n}\zeta_t\sum_{1}^{t-1}\zeta_m\}^{\frac{1}{2}}.
\end{aligned}$$

By the fact that $\{x_t\}$ is a mixingale sequence of size half, the first term is of order $\sum_{1}^{b_n} O(t^{-1-2\mu}) = O(b_n^{-2\mu})$ and the second term is of order

$$\sum_{2}^{b_n} O(t^{-\frac{1}{2}-\mu}) \sum_{1}^{t-1} O(m^{-\frac{1}{2}-\mu}) = O(b_n^{1-2\mu}).$$

By A.4, $\frac{M_n}{s_n} = O(n^{-\frac{1}{2}})$ and given that $b_n = [n^{1-\alpha}]$ and $r_n = O(n^{\alpha})$.

$$\begin{aligned}
\sum_{i=1}^{r_n} E(T_{i-1}\ E(Z_{n,i}|\mathcal{F}_{(i-1)b_n})) &= O(n^{-\frac{1}{2}} r_n b_n^{\frac{1-2\mu}{2}}) \\
&= O(n^{\alpha(\frac{1}{2}+\mu)-\mu}).
\end{aligned}$$

Choosing $\alpha < \frac{2\mu}{2\mu+1}$ will be sufficient to ensure that $\sum_{i=1}^{r_n} E(T_{i-1}\ E(Z_{n,i}|\mathcal{F}_{(i-1)b_n})) \to 0$ as $n \to \infty$ $\diamond$

It remains to be shown that $\sum_{1}^{r_n} Z_{n,i}^2 \xrightarrow{P} 1$. Here,

$$\begin{aligned}\sum_{1}^{r_n} Z_{n,i}^2 - 1 &= s_n^{-2}[\sum_{1}^{r_n}(\sum_{(i-1)b_n+1}^{ib_n} X_t)^2 - s_n^2] \\ &= A_n - B_n.\end{aligned}$$

Where $A_n = \sum_{1}^{r_n}(Z_{n,i}^2 - E(Z_{n,i}^2))$, and

$B_n = 2s_n^{-2}\sum_{2}^{r_n}\{\sum_{t=(i-1)b_n+1}^{ib_n}[\sum_{(i-1)b_n+1}^{t-1}\sigma_{t,t-m}]\}$, where $\sigma_{t,t-m} = E(X_tX_{t-m})$.

Proposition 2.4.4 *(1)* $A_n \xrightarrow{P} 0$.

(2) $B_n \to 0,\ n \to \infty$.

Proof:

(1) Put $A_n = A_{n,0} + 2\sum_{m=1}^{b_n-1} A_{n,m}$, where $A_{n,m} = s_n^{-2}\sum_{i=1}^{r_n}\sum_{(i-1)b_n+1}^{ib_n}(X_tX_{t-m} - \sigma_{t,t-m})$ for $m = 0,..,b_n - 1$ and $\sigma_{t,s} = E(X_tX_s)$.

By A.2 $\{\tilde{X}_t\}$ is L_r bounded, let $K = sup||\tilde{X}_t||_r < \infty$. Put $E(\tilde{X}_t\tilde{X}_{t-m}) = \tau_{t,t-m}$ then

$$\begin{aligned}|\tilde{X}_t\tilde{X}_{t-m} - \tau_{t,t-m}| &\leq 2\ |\tau_{t,t-m}| \\ &\leq 2\ E|\tilde{X}_{t-m}E(\tilde{X}_t|\mathcal{F}_{t-m})| \\ &\leq 2\ ||\tilde{X}_{t-m}||_r||E(\tilde{X}_t|\mathcal{F}_{t-m})||_{\frac{r}{r-1}} \\ &\leq 12\ ||\tilde{X}_{t-m}||_r\ \alpha_m^{\frac{r-2}{r}}\ ||\tilde{X}_t||_r \quad \text{by the } \alpha\text{-mixing inequality,} \\ &\leq 12\ K^2\ \zeta_m\end{aligned}$$

where $\zeta_m = O(m^{-1-\delta})$.

Now, $s_n^{-2}\ E|X_tX_{t-m} - \sigma_{t,t-m}| \leq 12K^2M_n^2s_n^{-2}\ \zeta_m$, and $M_n^2s_n^{-2} \sim n^{-1}$.

$E|A_{n,m}|$ contains r_nb_n of such term, as $r_nb_n = O(n)$ hence $E|A_{n,m}| = O(\zeta_m)$ for each

m. That is $E|A_{n,m}| = O(m^{-1-\delta})$. Therefore

$$\begin{aligned} E|A_n| &\leq E|A_{n,0}| + 2\sum_{1}^{b_n-1} E|A_{n,m}| \\ &= \sum_{1}^{b_n} O(m^{-1-\delta}) \\ &= O(b_n^{-\delta}) \\ &= O(n^{-a\delta}) \end{aligned}$$

Hence $E|A_n|$ converges to zero as $n \to \infty$ which is sufficient to establish the desired result. $\diamond$

$$\begin{aligned} (2)\ B_n &\leq 2s_n^{-2}\sum_{2}^{r_n}\{\sum_{t=(i-1)b_n+1}^{ib_n}[\sum_{(i-1)b_n+1}^{t-1}|\sigma_{t,t-m}|]\} \\ &\leq 2s_n^{-2}\, M_n^2K^2\sum_{2}^{r_n}\{\sum_{t=(i-1)b_n+1}^{ib_n}[\sum_{(i-1)b_n+1}^{t-1}\zeta_m]\} \\ &\leq 2s_n^{-2}\, M_n^2K^2\sum_{2}^{r_n}\sum_{j=1}^{b_n}\sum_{m=j}^{(i-1)b_n+j-1}\zeta_m \\ &\leq 2s_n^{-2}\, M_n^2K^2b_n\sum_{2}^{r_n}\sum_{m=1}^{ib_n}\zeta_m \\ &\leq 2s_n^{-2}\, M_n^2K^2b_n\sum_{2}^{r_n}\sum_{m=1}^{ib_n}O(m^{-1-\delta}) \\ &= O(n^{-1}b_{n^{1-\delta}r_n^{1-\delta}}) \\ &= O(n^{-\delta}) \to 0 \text{ as } n \to 0. \ \diamond \end{aligned}$$

With the three propositions above, the Davidson's CLT can easily be established on the basis of McLeish's CLT. $\diamond$

Chapter 3

Estimation of the First Derivative

In economics measures such as elasticities, marginal productivity and marginal cost are of considerable importance. These measures are related to the first derivative of one economic variable with respect to its exogenous determinant(s). In this chapter, we shall estimate the response coefficient of a regression function with respect to changes in an exogenous variable. This is the first (partial) derivative of $m(x)$ with respect to x_k, and we denote the estimand $\frac{\partial m(x)}{\partial x_k}$ by $\beta^k(x)$. Authors like McMillan, Ullah and Vinod (1988) and Rilstone (1987), for example, have studied the estimation of the derivative of a regression function by the standard fixed window width Watson-Nadaraya type of estimator. However, less attention has been devoted to the study of a recursive version of the Watson-Nadaraya type of derivative estimator. The objective of this chapter is to extend the theory of recursive kernel estimation in the previous chapter to the estimation of the first derivative of the regression function.

In the next section, we give a brief survey of some related work on kernel estimation of the derivatives of density and regression functions. In section 2, we propose a recursive estimator of the first derivative and in section 3 we derive its asymptotic properties for the cases of independent as well as α-mixing exogenous variables.

3.1 Survey of Related Works

Derivative of a Density

Bhattacharya (1967) proposed an estimator of a derivative of a density, $\hat{f}^{(p)}(x) = \frac{1}{nh^{2+p}} \sum K^{(p)}(\frac{X_i - x}{h})$, where the Kernel function satisfies

$$\int_{-\infty}^{\infty} |u|K(u)du < \infty, \quad lim_{|y|\to\infty} K^{(r)}(y) = 0 \text{ for } r = 0, 1, .., p-1,$$

and $|K^{(p+1)}(y)| < \infty$. He showed that under suitable conditions, $||E\hat{f}^{(p)} - f^{(p)}|| = O(h)$.

Schuster (1969) further showed that under certain conditions,
with $h = Cn^{-\frac{1}{2p+3}}$, $||\hat{f}^{(p)} - f^{(p)}|| = o(n^{-c})$ almost surely, where $0 < c < \frac{1}{2p+1}$.

Singh (1977,1979) used a kernel function that satisfies the orthogonality conditions:

$$\begin{aligned} |\int y^{p+1}K(y)dy| &\to 0 \quad as \quad |y| \to \infty \\ \int |y^p K(y)dy| &< \infty \end{aligned}$$

$$\int y^j K(y)dy = \begin{cases} j! & \text{if } j = p \\ 0 & \text{if } j = 0, 1, .., p-1, p+1, .., r > p. \end{cases}$$

and suggested an estimator for derivative of density function of the form,
$\hat{f}^{(p)}(x) = \frac{1}{nh^{2+1}} \sum K(\frac{X_i - x}{h})$.
Under suitable conditions, he then obtained an approximation for the bias and variance of $\hat{f}^{(p)}$.

$$\begin{aligned} Bias(\hat{f}^{(p)}) &\sim \frac{h^{(r-p)}f^{(r)}(x)}{r!} \int y^r K(y)dy \\ Var(\hat{f}^{(p)}) &\sim \frac{f(x)}{nh^{2p+1}} \int K^2. \end{aligned}$$

Singh (1981) extended the results to multivariate density and established the asymptotic normality of $\hat{f}^{(p)}$.

Menon, Prasad and Singh (1984) proposed a recursive version of the Singh's estimator of derivative of density and established conditions under which the estimator is consistent. Their estimator is of the form

$$\hat{f}^{(p)}(x) = \frac{1}{n}\sum h_j^{-(d+1)} K(\frac{X_i - x}{h_j}).$$

Derivative of Regression function

Schuster and Yakowitz (1979) proposed estimates of $m^{(p)}(x)$ as

$$m_n^{(p)}(x) = \sum_{k=0}^{p} \binom{p}{k} w_n^{(k)}(x)[g_n^{-1}(x)]^{(p-k)}$$

where $w_n^{(k)} = \frac{1}{nh_n}\sum Y_i K^{(k)}(\frac{x-X_i}{h_n})$ and $g_n(x) = \frac{1}{nh_n}\sum K(\frac{x-X_i}{h_n})$. This estimator attempts to estimate the object $\frac{\partial^k}{\partial x^k}(\frac{w(x)}{g(x)})$, where $g(x)$ is the marginal density of x and $w(x) = \int y f(x,y)dy$. Ullah and Vinod (1987,1988) have discussed an estimator in the same vein and gave examples of its applications to some econometric problems.

Rilstone (1987) proposed a first derivative estimator that is based on differencing two conditional mean estimates. Singh, Ullah and Carter (1987), and Ahmad and Ullah (1988), provide other results using this approach. In particular, Ahmad and Ullah (1988) propose the following estimator for pth order derivative, $m^{(p)}(x)$,

$$m_n^{(p)}(x) = (\frac{1}{2h})^p \sum_{k=0}^{p} \binom{p}{k} m_n(x + (p-2k)h).$$

In this case the conditions imposed on $K(.)$ are the same as those required for the estimation of $m(x)$. Hence the consistency and asymptotic normality of the estimator can be proved on the basis of the asymptotic properties of m_n via devices such as Cramer-Wold theorem.

In this chapter, we propose a recursive version of the derivative estimator for the case p=1. We then derive its asymptotic properties with independent observations as well as α-mixing observations.

3.2 Recursive Kernel Estimator of the 1st Derivative of Regression

Proposed Recursive Estimator

The recursive estimator of the first derivative of the density function:

$$f_n^k = \frac{n-1}{n} f_{n-1}^k + \frac{1}{n} h_n^{-d} \frac{1}{2h_n} \left[K(\frac{x_n^+ - X_n}{h_n}) - K(\frac{x_n^- - X_n}{h_n}) \right]$$

The recursive estimator of the first derivative of $r(x)$:

$$r_n^k = \frac{n-1}{n} r_{n-1}^k + \frac{1}{n} Y_n h_n^{-d} \frac{1}{2h_n} \left[K(\frac{x_n^+ - X_n}{h_n}) - K(\frac{x_n^+ - X_n}{h_n}) \right]$$

Combined with the estimator of the conditional mean and the density estimator, we obtain the estimator of the first derivative of the regression:

$$b_n^k(x) = \frac{r_n^k(x) - m_n(x) f_n^k(x)}{f_n(x)}$$

where $x_n^+ = x + h_n$ and $x_n^- = x - h_n$, $f_n(x) = \frac{1}{n} \sum_{i=1}^n h_i^{-d} K(\frac{X_i - x}{h_i})$,

and $m_n(x) = f_n(x)^{-1} \frac{1}{n} \sum_{i=1}^n h_i^{-d} Y_i K(\frac{X_i - x}{h_i})$.

Put $K_i^- = \frac{1}{2h_i} \left[K(\frac{x_i^+ - X_i}{h_i}) - K(\frac{x_i^- - X_i}{h_i}) \right]$ and hence

$$b_n^k = \frac{1}{f_n(x)} \frac{1}{n} \sum_1^n Y_i h_i^{-d} K_i^- - \frac{m_n(x)}{f_n(x)} \frac{1}{n} \sum_1^n h_i^{-d} K_i^-$$

correspondingly, $f_n^k(x) = \frac{1}{n} \sum_1^n h_i^{-d} K_i^-$ and $r_n^k(x) = \frac{1}{n} \sum_1^n Y_i h_i^{-d} K_i^-$.

Remark

Here, another way to estimate the derivative is to use higher order kernel function in place of K_i^-. The resulting estimator is able to deliver an estimate of the same object. However, it requires stronger smoothness assumption on the underlying regression function. This approach was first proposed by Singh (1974) for the estimation of the derivative of a density function. Beside stronger smoothness assumption, the proofs to establish this estimator's consistency and asymptotic normality are similar to those in section 3 below. Hence, we will not discuss them in this thesis.

3.3 Asymptotic properties

In the previous chapter, we established the conditions for the weak and strong consistency of $f_n(x)$ and $m_n(x)$. It is sufficient for us to specify the additional conditions required to establish the weak and strong consistency of f_n^k and r_n^k for $\frac{\partial}{\partial x_k} f(x)$ and $\frac{\partial}{\partial x_k} m(x)$ respectively. The consistency of b_n^k for $\frac{\partial}{\partial x_k} m(x)$ follows from the arguments below.

As the conditions required to establish the consistency of the proposed estimator will also satisfy that of Theorem 1.4.6 thus it is true that $f_n(x) \xrightarrow{p} (or \xrightarrow{a.s.}) f(x)$ and $m_n(x) \xrightarrow{p} (or \xrightarrow{a.s.}) m(x)$. Therefore, by the properties of stochastic convergence we can conclude that

$$\begin{aligned} b_n^k(x) &= \frac{r_n^k(x)}{f_n(x)} - \frac{m_n(x) f_n^k(x)}{f_n(x)} \\ &\xrightarrow{p} (\xrightarrow{a.s.}) \frac{f(x)\frac{\partial}{\partial x_k} m(x) + m(x)\frac{\partial}{\partial x_k} f(x)}{f(x)} - \frac{m(x)\frac{\partial}{\partial x_k} f(x)}{f(x)} \\ &\xrightarrow{p} (\xrightarrow{a.s.}) \frac{\partial}{\partial x_k} m(x). \end{aligned}$$

3.3.1 Asymptotic Unbiasedness

Theorem 3.3.1 *Suppose K.1, K.2, H.1 and A.0 hold then* $Ef_n^k \to \frac{\partial}{\partial x_k} f$ *as* $n \to \infty$. *In addition, if A.2 holds then* $Er_n^k \to \frac{\partial}{\partial x_k} r$ *as* $n \to \infty$.

Proof Under the given conditions Lemma 1.4.1 applies and hence

$$\begin{aligned} E(h_i^{-d} K_i^*) &= \frac{1}{2h_i}[f(x + h_i) - f(x - h_i) + O(h_i^2)] \\ &= \frac{\partial}{\partial x_k} f(x) + O(h_i) \end{aligned}$$

By Lemma 1.4.2, $E(f_n^k) \to \frac{\partial}{\partial x_k} f + O(h_n)$.

Similarly,

$$E(h_i^{-d}\, Y_i\, K_i^*) = \frac{1}{2\, h_i}[r(x + h_i) - r(x - h_i) + O(h_i^2)]$$

$$= \frac{1}{2\,h_i}[m(x+h_i)f(x+h_i) - m(x-h_i)f(x-h_i)] + O(h_i)$$
$$= f(x)\frac{\partial}{\partial x_k}m(x) + m(x)\frac{\partial}{\partial x_k}f(x) + O(h_i)$$

By Lemma 1.4.2, $E(r_n^k) \to f(x)\frac{\partial}{\partial x_k}m(x) + m(x)\frac{\partial}{\partial x_k}f(x) + O(h_n)$.

3.3.2 The Bound of the Moments of the Kernel Functions

In the following two propositions we give the bound of the moments of U_i and V_i where $U_i = h_i^{-d}K_i^*$ and $V_i = h_i^{-d}Y_i\,K_i^*$ which are required for the proofs in the next section.

Proposition 3.3.1 *Under the conditions that K.1, K.2, A.0, and H.1 hold, then*

(i) $E(U_i) = O(1)$

(ii) $Var(U_i) = O(h_i^{-(d+2)})$, *and*

(iii) $||U_i||_r = O(h_i^{-(1-\frac{1}{r})d-1})$.

Proof (i) Under the given conditions, the assumptions of Theorem 3.3.1 are satisfied and

hence by Theorem 3.3.1, $EU_i = f(x) + O(h_i)$.

By H.1 as n $\to \infty$ $h_i \to 0$, so the claim of the proposition follows. $\diamond$

(ii) $\quad Var(U_i) = E(U_i^2) - [E(U_i)]^2$

$$Var(U_i) = \int h_i^{-2d}[\frac{1}{2h_i}\left(K(\frac{x+h_i-u}{h_i}) - K(\frac{x-h_i-u}{h_i})\right)]^2 f(u)du \;+\; O(1)$$
$$= \int h_i^{-d}[\frac{1}{2h_i}(K(s+1) - K(s-1))]^2 f(x-sh_i)ds \;+\; O(1)$$
$$= h_i^{-(d+2)}[f(x)\int K'^2 ds] \;+\; O(1)$$
$$= O(h_i^{-(d+2)}) \quad \diamond$$

(iii) $$E|U_i|^r = h_i^{-dr}\int h_i^{-d}|K^*|^r f(u)du$$
$$= h_i^{-d(r-1)}h_i^{-r}\int |\frac{1}{2h_i}(K(s+1) - K(s-1))|^r f(x-sh_i)ds$$

$$= h_i^{-d(r-1)-r}[f(x)\int |K'|^r]$$
$$= O(h_i^{-d(r-1)-r})$$

Hence $||U_i||_r = (E|U_i|^r)^{\frac{1}{r}}$
$$= O(h_i^{-d(1-\frac{1}{r})-1}) \quad \diamond$$

Corollary 3.3.1 *Under the same conditions as in the above proposition, and in addition A.1' holds, then*

$Cov(U_i,\ U_j) \leq O(\alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})-1}\ h_j^{-d(1-\frac{1}{r})-1}\ )$.

Proof

By Lemma 2.2.1, $Cov(U_i,\ U_j) \leq |E(U_i\ U_j) - E(U_i)E(U-j)|$
$$\leq \alpha_{|i-j|}^{1-\frac{2}{r}}\ ||U_i||_r||U_j||_r$$
$$\leq O(\alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})-1}\ h_j^{-d(1-\frac{1}{r})-1}\ ) \quad \diamond$$

Proposition 3.3.2 *Under the conditions that K.1, K.2, A.0, A.2 and H.1 hold, then*

(i) $E(V_i) = O(1)$

(ii) $Var(V_i) = O(h_i^{-(d+2)})$; *and*

(iii) $||V_i||_r = O(h_i^{-(1-\frac{1}{r})d-1})$.

Proof (i) Under the given conditions, the assumptions of Theorem 3.3.1 are satisfied and

hence by Theorem 3.3.1, $EV_i = f(x)\frac{\partial}{\partial x_k}m(x) + m(x)\frac{\partial}{\partial x_k}f(x) + O(h_i)$.

By H.1 as n $\rightarrow \infty$ $h_i \rightarrow 0$, so the claim of the proposition follows. $\diamond$

(ii) $Var(V_i) = E((h_i^{-d}\ Y_i\ K_i^*)^2) - [E(h_i^{-d}\ Y_i\ K_i^*)]^2$

$$Var\ (\frac{h_i^d}{Y_i\ K_i^*})$$
$$= \int\int h_i^{-2d}y^2[\frac{1}{2h_i}\left(K(\frac{x+h_i-u}{h_i}) - K(\frac{x-h_i-u}{h_i})\right)]^2 f(u,y)dudy + O(1)$$

$$= \int h_i^{-d}[\frac{1}{2h_i}(K(s+1)-K(s-1))]^2 E(Y^2|x-sh_i)f(s)ds \ + \ O(1)$$
$$= h_i^{-d}h_i^{-2}[f(x)E(Y^2|x)\int K'^2] \ + \ O(1)$$
$$= O(h_i^{-(d+2)}) \qquad \diamond$$

(iii) $E|\ V_i\ |^r$

$$= h_i^{-dr}\int\int |\frac{1}{2h_i}\left(K(\frac{x+h_i-u}{h_i})-K(\frac{x-h_i-u}{h_i})\right)|^r |y|^r f(u,y)dydu$$
$$= h_i^{-d(r-1)}\int |\frac{1}{2h_i}(K(s+1)-K(s-1))|^r E(|Y|^r|x-sh_i)f(s)ds$$
$$= h_i^{-d(r-1)-r}[f(x)E(|Y|^r|x)\int |K'|^r + O(h_i^2)]$$
$$= O(h_i^{-d(r-1)-r})$$

Hence $||V_i||_r = (E|V_i|^r)^{\frac{1}{r}}$

$$= O(h_i^{-d(1-\frac{1}{r})-1}) \qquad \diamond$$

Corollary 3.3.2 *Under the same conditions as in the above proposition, and in addition A.1' holds, then* $Cov(V_i,\ V_j) \ \leq \ O(\alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})-1}\ h_j^{-d(1-\frac{1}{r})-1}\ )$.

Proof Same in the corollary above. $\diamond$

3.3.3 Consistency of the Proposed Estimator

Weak Consistency for Independent Regressors

Theorem 3.3.2 *Suppose K.1, K.2, A.0, A.1 A.2, H.1 hold, and*

(H.2') $n\ h_n^{d+2} \to \infty$;

(H.3) $\lim_{n\to\infty}\frac{1}{n}\sum_1^n(\frac{h_{in}}{h_n})^\gamma < \infty$, *with* $\gamma = d+2$

are satisfied then $b_n^k(x) \xrightarrow{p} \frac{\partial}{\partial x_k}m(x)$.

Proof

$$Var(f_n^k(x)) \ = \ Var(\frac{1}{n}\sum U_i)$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum C_i h_i^{-(d+2)} \quad \text{by Proposition 3.3.1} \\
&\leq \frac{1}{n^2}\sum C h_i^{-(d+2)} \quad \text{where by A.0 } \exists C = sup_i C_i < \infty \\
&= \frac{C}{nh_n^{d+2}}\frac{1}{n}\sum(\frac{h_n}{h_i})^{d+2} \\
&= O((nh_n^{d+2})^{-1}) \\
&\to 0 \quad \text{as} \quad n \to \infty \text{ by H.2'}.
\end{aligned}
$$

Hence $f_n^k(x) \xrightarrow{p} \frac{\partial}{\partial x_k} f(x)$ by Chebyshev's WLLN and Theorem 3.3.1.

Similarly,

$$
\begin{aligned}
Var(r_n^k(x)) &= Var(\frac{1}{n}\sum V_i) \\
&\leq \frac{1}{n^2}\sum C h_i^{-(d+2)} \\
&= \frac{C}{nh_n^{d+2}}\frac{1}{n}\sum(\frac{h_n}{h_i})^{d+2} \\
&= O((nh_n^{d+2})^{-1}) \\
&\to 0 \quad \text{as} \quad n \to \infty \text{ by H.2'}.
\end{aligned}
$$

And $r_n^k(x) \xrightarrow{p} f(x)\frac{\partial}{\partial x_k}m(x) + m(x)\frac{\partial}{\partial x_k}f(x)$.

Hence the weak convergence of $b_n^k(x)$ follows. ◇

3.3.4 Strong Consistency for Independent Regressors

Theorem 3.3.3 *Suppose K.1, K.2, A.0, A.1 A.2, H.1 hold, and*

(H.2") $\sum_i \frac{1}{i^2 h_i^{d+2}} < \infty$;

is satisfied then $b_n^k(x) \xrightarrow{a.s.} \frac{\partial}{\partial x_k}m(x)$.

Proof

$$
\begin{aligned}
\sum \frac{1}{i^2} Var(U_i) &= \sum \frac{1}{i^2} h_i^{-(d+2)} C_i \quad \text{by Proposition 3.3.1} \\
&\leq C\sum \frac{1}{i^2 h_i^{d+2}} \quad \text{by A.0 } \exists C = sup_i C_i < \infty \\
&< \infty \quad \text{by H.2"}
\end{aligned}
$$

By Kolmogorov's SLLN and Theorem 3.3.1 $f_n^k(x) = \frac{1}{n}\sum U_i \xrightarrow{a.s.} \frac{\partial}{\partial x_k} f(x)$.

Similarly,

$$\begin{aligned}\sum \frac{1}{i^2} Var(V_i) &= \sum \frac{1}{i^2} h_i^{-(d+2)} C_i \quad \text{by Proposition 3.3.1} \\ &\leq C \sum \frac{1}{i^2 h_i^{d+2}} \quad \text{by A.2 } \exists C = sup_i C_i < \infty \\ &< \infty \quad \text{by H.2"}\end{aligned}$$

By Kolmogorov's SLLN and Theorem 3.3.1, $r_n^k(x) = \frac{1}{n}\sum V_i \xrightarrow{a.s.} f(x)\frac{\partial}{\partial x_k} m(x) + m(x)\frac{\partial}{\partial x_k} f(x)$.

Therefore, $b_n^k(x) \xrightarrow{a.s.} \frac{\partial}{\partial x_k} m(x)$. $\diamond$

Remark

If we set $h_i = O(i^{-\alpha})$ then the condition (H.2') of Theorem 3.3.2 will be equivalent to the condition that $1 - \alpha(d+2) > 0$ or $\alpha < \frac{1}{d+2}$. H.2" of Theorem 3.3.3 will mean that $2 - \alpha(d+2) > 1$ which also leads to the same condition $\alpha < \frac{1}{d+2}$. Hence under the settings stated in the Theorems 3.3.2 and 3.3.3 the requirements for the strong and weak consistency of the proposed estimator are in fact equivalent.

Weak (also Mean Square) Consistency α-mixing Sequences

Theorem 3.3.4 *Suppose the derivative estimator $b_n^k(x)$ with*
kernel function satisfies K.1 and K.2; and
window width satisfies H.1 and H.2 $nh_n^\gamma \to \infty$ as $n \to \infty$ such that*
H.3 $\lim_{n\to\infty} \frac{1}{n}\sum_1^n (\frac{h_n}{h_i})^\gamma$ converges,
where $\gamma = 2d(1-\frac{1}{r})+2$, $r > 2$; and the underlying DGP satisfies A.0 and A.1' such that
α_m is of size $-\frac{r}{r-2}$ $r > 2$,
then $b_n^k(x) - \frac{\partial}{\partial x_k} m(x) \to 0$ in probability.

Proof By Proposition 3.3.1, $Var(U_i)$ and $Var(V_i) = O(h_i^{-d-2})$ and in general

$||U_i||_r$ and $||V_i||_r = O(h_i^{-d(1-\frac{1}{r})-1})$.

The fact that α_m is of size$-\frac{r}{r-2}$, $r > 2$, means $\sum_1^\infty \alpha_n^{1-\frac{2}{r}}$ converges.

$$\begin{aligned}
Var\ (\ f_n^k) &= Var(\frac{1}{n}\sum_1^n U_i) \\
&= \frac{1}{n^2}[\sum_i Var(U_i) + \sum\sum_{i\neq j} Cov(U_i, U_j)] \\
&\leq \frac{1}{n^2}[\sum_i C_i\ h_i^{-d-1} + \sum\sum_{i\neq j} C_i\ C_j\ \alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})-1}\ h_j^{-d(1-\frac{1}{r})-1}] \\
&\leq \frac{1}{n^2}[\sum_i C_i h_i^{-d-1} + 2\sum_{i=1}^n h_i^{-2d(1-\frac{1}{r})-2}\sum_{j<i} C^2\ \alpha_{|i-j|}^{1-\frac{2}{r}}\ (\frac{h_j}{h_i})^{-d(1-\frac{1}{r})-1}] \\
&= O(\frac{1}{n^2}\sum_i h_i^{-\gamma}) \\
&= O\left(\frac{1}{nh_n^\gamma}(\frac{1}{n}\sum_i(\frac{h_n}{h_i})^\gamma)\right) \ \rightarrow\ 0
\end{aligned}$$

where $\gamma = 2d(1-\frac{1}{r}) + 2$ for $r > 2$, by H.3 $\frac{1}{n}\sum(\frac{h_n}{h_i})^\gamma < \infty$, and by H.2* $nh_n^\gamma \rightarrow \infty$ as $n \rightarrow \infty$.

Together with Theorem 3.3.1, we have f_n^k converges to $\frac{\partial}{\partial x_k}f(x)$ in mean square and by Chebyshev's inequality $f_n^k \xrightarrow{p} \frac{\partial}{\partial x_k}f(x)$.

Similarly,

$$\begin{aligned}
Var(r_n^k) &= Var(\frac{1}{n}\sum_1^n V_i) \\
&= \frac{1}{n^2}[\sum_i Var(V_i) + \sum\sum_{i\neq j} Cov(V_i, V_j)] \\
&\leq \frac{1}{n^2}[\sum_i C_i\ h_i^{-d-1} + \sum\sum_{i\neq j} C_i\ C_j\ \alpha_{|i-j|}^{1-\frac{2}{r}}\ h_i^{-d(1-\frac{1}{r})-1}\ h_j^{-d(1-\frac{1}{r})-1}] \\
&\leq \frac{1}{n^2}[\sum_i C_i h_i^{-d-1} + 2\sum_{i=1}^n h_i^{-2d(1-\frac{1}{r})-2}\sum_{j<i} C^2\ \alpha_{|i-j|}^{1-\frac{2}{r}}\ (\frac{h_j}{h_i})^{-d(1-\frac{1}{r})-1}] \\
&= O(\frac{1}{n^2}\sum_i h_i^{-\gamma}) \\
&= O\left(\frac{1}{nh_n^\gamma}(\frac{1}{n}\sum_i(\frac{h_n}{h_i})^\gamma)\right) \ \rightarrow\ 0
\end{aligned}$$

By Theorem 3.3.1, we have r_n^k converges to $f(x)\frac{\partial}{\partial x_k}m(x) + m(x)\frac{\partial}{\partial x_k}f(x)$ in mean square and by Chebyshev's inequality $r_n^k \xrightarrow{p} f(x)\frac{\partial}{\partial x_k}m(x) + m(x)\frac{\partial}{\partial x_k}f(x)$.

Hence $b_n^k(x) \xrightarrow{p} \frac{\partial}{\partial x_k}m(x)$. $\diamond$

3.3.5 Strong Consistency for α-mixing Sequences

Theorem 3.3.5 *Suppose the kernel derivative estimator $b_n^k(x)$ with kernel function satisfies K.1 and K.2; and the window width satisfies H.1 and H.2 such that $\lim_{n\to\infty}\sum_1^n (ih_i^{d(1-\frac{1}{r})+1})^{-2}$ converges, where $r > 2$; and the underlying DGP satisfies A.0, A.2 and A.1' such that α_m is of size $-\frac{r}{r-2}$, $r > 2$, then $b_n^k(x) - \frac{\partial}{\partial x_k}m(x) \to 0$ almost surely.*

Proof Let $Z_i = U_i - E(U_i)$ and

$\mathcal{F}_i = \sigma(\{X_j\}_{j=1}^i)$ for i$\geq$ 1 and $\mathcal{F}_i = \{\emptyset, \Omega\}$ for i $<$ 1.

Then we have $||EZ_i|\mathcal{F}_{i+m}||_2 = 0$ and using the inequality associated with α-mixing sequences,

$$||EZ_i|\mathcal{F}_{i-m}||_2 \leq 2(2^{\frac{1}{2}}+1)\alpha_m^{\frac{1}{2}-\frac{1}{r}}||EZ_i||_r \quad \text{for} \quad r > 2$$

Hence $\{Z_i,\mathcal{F}_i\}$ is a L^2-mixingale with c_i=$||Z_i||_r$ and ψ_m=$5\alpha^{\frac{1}{2}-\frac{1}{r}}$.

As α_m is of size $-\frac{r}{r-2}$, $r > 2$ it follows that ψ_m is of size $-\frac{1}{2}$.

Furthermore, $||Z_i||_r \leq 2||U_i||_r$ and by assumption A.0 it is shown in Proposition 3.3.1, $||U_i||_r = O(h_i^{-2d(1-\frac{1}{r})-2})$.

So $\exists$ a finite C such that the constant of the mixingale

$c_i = ||Z_i||_r \leq C\ h_i^{-2d(1-\frac{1}{r})-2}$ for all i and

$\sum_1^n \frac{c_i^2}{i^2} \leq C\sum_1^n i^{-2}h_i^{-2d(1-\frac{1}{r})-2}$ converges as n $\to \infty$ by the assumption stated above.

Therefore the conditions of McLeish's (1975) SLLN are satisfied and we have

$\frac{1}{n}\sum_{1}^{n} Z_i = f_n^k - Ef_n^k \rightarrow 0$ almost surely.

Hence together with Theorem 3.3.1, $f_n^k \xrightarrow{a.s.} \frac{\partial}{\partial x_k} f(x)$.

Let $Z_i = h_i^{-d} K_i^* Y_i - E h_i^{-d} K_i^* Y_i$ and

$\mathcal{F}_i = \sigma(\{X_j, Y_j\}_{j=1}^{i})$ for i$\geq$ 1 and $\mathcal{F}_i = \{\emptyset, \Omega\}$ for i < 1

Then $||EZ_i|\mathcal{F}_{i+m}||_2 = |EZ_i| = 0$ and by the inequality associated with α-mixing sequences,

$||EZ_i|\mathcal{F}_{i-m}||_2 \leq 2(2^{\frac{1}{2}}+1)\alpha_m^{\frac{1}{2}-\frac{1}{r}}||EZ_i||_r$ for r > 2.

Hence $Z_i, \mathcal{F}_i$ is a mixingale sequence with ψ_m of size $-\frac{1}{2}$ and $c_i = ||Z_i||_r$.

By A.0 and A.2, it is shown in Proposition 3.3.1 that $||Z_i||_r = O(h_i^{-d(1-\frac{1}{r})-1})$. Therefore, $\sum_{1}^{\infty} \frac{c_i^2}{i^2}$ converges.

It follows that $\frac{1}{n}\sum_{1}^{n} Z_i = r_n^k - E(r_n^k) \xrightarrow{a.s.} 0$ by McLeish's (1975) SLLN and together with Theorem 3.3.1, $r_n^k \xrightarrow{a.s.} f(x)\frac{\partial}{\partial x_k} m(x) + m(x)\frac{\partial}{\partial x_k} f(x)$

The conclusion of the above theorem follows. $\diamond$

3.3.6 Asymptotic Normality

It is sufficient for us to show that under appropriate conditions to be specified,

$$\frac{r_n^k(x) - E(r_n^k(x))}{\sqrt{Var(r_n^k)}} \xrightarrow{D} N(0,1).$$

Under the conditions of Theorem 3.3.1, we can extend this result to

$$\frac{r_n^k(x) - \frac{\partial}{\partial x_k} r(x)}{\sqrt{Var(r_n^k)}} \xrightarrow{D} N(B,1) \quad \text{where B is the bias as explained before.}$$

Under appropriate conditions as specified in the previous theorem pertaining to weak convergence of estimators, we have shown that

$$f_n^k(x) \xrightarrow{p} \frac{\partial}{\partial x_k} f(x),$$

$$m_n(x) \xrightarrow{p} m(x),$$

$f_n(x) \xrightarrow{P} f(x)$.

Hence the asymptotic Normality of $b_n^k(x)$ will follow by Slutsky's theorem.

Regresssors are Independent Processes

Theorem 3.3.6 *Suppose the recursive kernel estimator of the 1st derivative of regression satisfies the conditions of Theorem 3.3.2,*
then $T_n = \frac{r_n^k - \frac{\partial}{\partial x_k} r}{\sqrt{Var(r_n^k)}} \xrightarrow{\mathcal{D}} N(B,1)$ *where* $B = O(\sqrt{nh^{4+d}})$.

Proof Here, if we could show that under the same conditions as that of Theorem 3.3.4 the following criterion ρ_n, converges to zero as $n \to \infty$, then the asymptotic normality of T_n will follow by Liapunov's theorem.

$$\rho_n = \frac{(\sum_1^n E|V_i - EV_i|^3)^{\frac{1}{3}}}{(\sum_1^n Var(V_i))^{\frac{1}{2}}}$$

where $V_i = h_i^{-d} K_i^* Y_i$.

By Proposition 3.3.1, with r=3 $||V_i||_3 = O(h_i^{-d(1-\frac{1}{3})-1})$,

$$E|V_i - E_i|^3 \leq 2\ E|V_i|^3 = O(h_i^{-2d-3}),$$

and $Var(V_i) = O(h_i^{-d-2})$.

Therefore, ρ_n is bounded by $C\ \frac{(\sum_1^n h_i^{-3d-3})^{\frac{1}{3}}}{(\sum_1^n h_i^{-d-2})^{\frac{1}{2}}}$ for some finite C.

And

$$\frac{(\sum_1^n h_i^{-2d-3})^{\frac{1}{3}}}{(\sum_1^n h_i^{-d-2})^{\frac{1}{2}}} = \frac{(\frac{1}{n}\sum_1^n(\frac{h_n}{h_i})^{2d+3})^{\frac{1}{3}}}{(\frac{1}{n}\sum_1^n(\frac{h_n}{h_i})^{d+2})^{\frac{1}{2}}} \frac{h_n^{-\frac{2}{3}d}}{h_n^{-\frac{1}{2}d}} \frac{n^{\frac{1}{3}}}{n^{\frac{1}{2}}} = O((nh_n)^{-\frac{1}{6}})$$
$$\to 0 \quad \text{by H.2, } nh_n \to \infty \text{ as } n \to \infty$$

Therefore the asymptotic normality of T_n is proved.

By Theorem 3.3.2, $Var(r_n^k) = O((nh^{2+d})^{-1})$, and by Theorem 3.3.1, $Bias(r_n^k) = O(h_n)$. Thus $B = O(\sqrt{nh^{4+d}})$. $\diamond$

As explained at the beginning of this section we can claim the following corollary.

Corollary 3.3.3 *If the conditions of Theorem 3.3.2 hold, then* $\frac{b_n^k - \frac{\partial}{\partial x_k} m}{\sqrt{Var(b_n^k)}} \xrightarrow{\mathcal{D}} N(B,1)$ *where* $B = O(\sqrt{nh^{4+d}})$.

Regressors are α-mixing Processes

Theorem 3.3.7 *Suppose the recursive kernel 1st derivative estimator of* $r(x)$ *satisfies the conditions of Theorem 3.3.4,*
then $S_n = \frac{r_n^k - r^k}{\sqrt{Var(r_n^k)}} \xrightarrow{\mathcal{D}} N(B,1)$ *where* $B = O(\sqrt{nh^{4+2d(1-\frac{1}{r})}})$.

Proof We will apply Davidson's (1990) CLT for mixingales to prove the required result.

Let $Z_i = h_i^{-d} K_i^{-} Y_i - E h_i^{-d} K_i^{-} Y_i$ and

$\mathcal{F}_i = \sigma(\{X_j, Y_j\}_{j=1}^{i})$ for i$\geq$ 1 and $\mathcal{F}_i = \{\emptyset, \Omega\}$ for i $<$ 1

Then $EZ_i = 0$ and part (a) of the Davidson's theorem is trivially satisfied.

Z_i is a measurable function of strong mixing sequences, and therefore, it is a strong mixing sequence of the same size. Hence part (b) of Davidson's theorem is satisfied.

Let $c_i = h_i^{-d(1-\frac{1}{r})-1}$ then

$E|z_i|^k \leq 2h_i^{dk(1-\frac{1}{r})+k} E|h_i^{-d} Y_i K_i^{-}|^k$, and by Proposition 3.3.1,

$$\begin{aligned} E|z_i|^k &= O(h_i^{dk(1-\frac{1}{r})+k}) \, ||V_i||_k^k \\ &= O(h_i^{dk(1-\frac{1}{r})+k}) O(h_i^{-d(1-\frac{1}{r})+k}) \\ &= O(1) \end{aligned}$$

Therefore condition (c) is satisfied.

$$\begin{aligned} s_n^2 &= E(\sum_1^n Z_i)^2 \\ &= \sum_1^n EZ_i^2 + 2\sum_{i=1}^n \sum_{j=1}^{i-1} EZ_iZ_j \\ &\sim \sum_1^n h_i^{-d-2} V_i + 2\sum_1^n \sum_1^{i-1} (h_i h_{i-j})^{-d(1-\frac{1}{r})-2} \alpha_{i,i-j}^{1-\frac{2}{r}} C_{i,i-j} \quad for\ r > 2 \end{aligned}$$

$$= O(\sum_{1}^{n} h_i^{-\gamma})$$

where γ lies between d+2 to 2d+2.

Here we make use of the fact that $\{\alpha_m\}$ is of size $-\frac{r}{r-2}$ which means $\sum_1^\infty \alpha_m^{\frac{r-2}{r}}$ converges to some finite value; and $\frac{h_i}{h_{i-j}} \leq 1$ for all i and j.

And $M_n = \max_{i \leq n} c_i = O(h_n^{-d(1-\frac{1}{r})-1})$, therefore

$$\begin{aligned} \frac{nM_n^2}{s_n^2} &\sim \frac{nh_n^{-d(1-\frac{1}{r})-1}}{\sum_1^n h_i^{-\gamma}} \\ &= O(h_n^{\gamma-d(1-\frac{1}{r})-1}) \\ &= O(1) \\ &< \infty. \end{aligned}$$

Now, part (d) of the above theorem is also satisfied.

Therefore, the conclusion of the theorem follows by Davidson's (1990) CLT.

By Theorem 3.3.4, $Var(r_n^k) = O((nh^\gamma)^{-1})$, where $\gamma = 2d(1-\frac{1}{r})+2$, and by Theorem 3.3.1, $Bias(r_n^k) = O(h_n)$. Thus $B = O(\sqrt{nh^{4+2d(1-\frac{1}{r})}})$. $\diamond$

Following the discussion at the beginning of the section we give the corollary below,

Corollary 3.3.4 *If the conditions of Theorem 3.3.4 hold, then* $\frac{b_n^k - \frac{\partial}{\partial x_k} m}{\sqrt{Var(b_n^k)}} \xrightarrow{D} N(B,1)$ *where* $B = O(\sqrt{nh^{4+2d(1-\frac{1}{r})}})$.

3.3.7 Optimal Window Width and Optimal Rate of Convergence

In order to obtain the rate of convergence we need to derive an approximation of the local mean square error (MSE).

Theorem 3.3.8 *The local MSE=*$O(h_n^2 + \frac{1}{nh_n^\gamma})$

where γ*=d+2 for case where observations are independent, or*

γ=$2d(1-\frac{1}{r})+2$ *where observations are from an* α*-mixing process of size* $-\frac{r}{r-2}$, $r > 2$.

Proof Here $Var(m_n) = O((nh_n^\gamma)^{-1})$,

for observations from an independent process, $\gamma = d + 2$ as shown in the proof of Theorem 3.3.2.

And for α-mixing sequence, $\gamma = 2d(1-\frac{1}{r})+2$ as shown in the proof of Theorem 3.3.4.

By Theorem 3.3.1, $Bias(m_n) = O(h_n)$.

It implies that

$$\begin{aligned} MSE &= Variance + Bias^2 \\ &= O((nh_n^\gamma)^{-1} + h_n^2). \quad \diamond \end{aligned}$$

With the above results we can make the following claim.

Proposition 3.3.3 *The optimal rate of decay of window width for the recursive regression estimator* m_n *is given by*

$h_n = cn^{-\frac{1}{2+\gamma}}$ *and* $\gamma = d + 2$, *for the independent case, and* $\gamma = 2d(1-\frac{1}{r}) + 2$ *for the* α*-mixing case.*

The corresponding rate of weak convergence is $O(n^{-\frac{2}{2+\gamma}})$.

Proof In the approximate MSE, the bias term decreases as h_n decreases whereas the variance increases as h_n decreases for fixed sample size n.

So the approximate MSE is minimized when the two terms are decreasing at the same rate with n increasing.

$$\begin{aligned} (nh_n^\gamma)^{-1} &\sim h_n^2 \\ h_n &\sim n^{-\frac{1}{2+\gamma}} \\ MSE &= O(n^{-\frac{2}{2+\gamma}}). \quad \diamond \end{aligned}$$

Remark

The optimal rate of decay for the window width is slower than that of for the regression estimate. The MSE here converges at a slower rate.

Chapter 4

Recursive Estimation: its Implementation and Finite Sample Properties

4.1 Introduction

The asymptotic theory we established in earlier chapters was concerned only with the rate of decay of the window width in order to attain consistency and asymptotic normality. Therefore, it does not provide us with enough information to implement a recursive estimator. Hence, there is a need to show how the recursive estimator can be implemented. This requires a discussion on window width selection and the algorithm in general which we will deal with in section 2.

In the first three chapters, we were concerned only with the asymptotic properties of the recursive estimators. Consistency of the estimator is important; since if the estimator is not able to correctly estimate the target with indefinitely large data sets then perhaps it is just not worthy of any further attention. On the other hand, given that an estimator does have the required asymptotic properties under certain regularity conditions, it is only natural to investigate its finite sample properties or at least demonstrate instances where the asymptotic behavior can be a close approximation to its finite sample counterpart

Usually, the finite sample behavior of a nonparametric kernel estimator does not have a simple analytical form that can be treated generally. It is common practice to illustrate the estimator's finite sample behavior with some controlled experiments through a simulation study. In addition to the usual problems faced by the kernel nonparametric estimator, the recursive estimator's window width is not always at its optimum. Does this handicap matter to an extent that defeats the purpose of attempting recursive nonparametric estimation altogether? This is the question that we try to address in this chapter. The other objective of this chapter is to provide a Monte Carlo study of the recursive estimator in particular and a comparison of its performance with other commonly used estimators.

In section 3, we construct a model for Monte Carlo simulation for the purpose of studying the finite sample properties of the recursive estimator under different data generating processes for the regressor at different sample sizes. We will compare the performance of the recursive estimator with the two other kernel regression estimators with controlled experiments. Section 4 reports the outcomes and their implications. A summary and some concluding comments are provided in the last section.

4.2 Implementation of the Recursive Nonparametric Estimator

4.2.1 Introduction

Implementation of a nonparametric kernel estimator requires us to deal with the problems of selecting window width and kernel function. Nonparametric kernel estimation in general can be regarded as a local weighted averaging scheme while the role of the window width is to determine the span of a sampling of points. The kernel function determines the distribution of weight inside the chosen window. It appears that the selection of the kernel generally plays a minor role in nonparametric ker-

nel estimation. Its impact on the efficiency of the estimator is marginal. This was first reported by Rosenblatt (1956) in the density estimation setting and his findings can easily be transferred to a regression estimation case. Singh (1974) advanced the ideas of bias reduction by higher order kernels which might relax the selection of window width. However, such an approach might end up with a larger variance and assign negative weight to its local averaging scheme which might not be acceptable at times. In this chapter, we limit our choice of kernel to a second order kernel function which is usually some form of symmetric density function.[1] The main focus of the implementation of the recursive estimation, however, is the algorithm for selecting the window width. In order to explain the rationale behind the our algorithm, the next subsection will describe the two main strategies in window width selection, the cross-validation and plug-in methods, and assess their strengths and weaknesses as window width selection methods for recursive estimation.

4.2.2 Techniques of Window Width Selection

Heuristically it is not hard for us to see that a wider window width would mean the estimate is an average over a larger number of points and, hence, would tend to have a smaller variance. On the other hand, with larger window width, several other features of the regression might be averaged out and the estimate would tend to have increased bias. The choice of window width involves this delicate task of balancing the two components; the variance on one hand, and the bias in the mean square error on the other. One direct and intuitive way is to sample plots of estimates with different window widths and choose an appropriate one by inspection. However, such an intuitive approach will not appeal to any serious researcher who intends to perform rigorous inference with the estimates. An automatic data-driven procedure

[1] We might still use higher order kernels in our pilot estimation of the derivative of density and regression functions.

that is able to remove this element of subjective choice from the procedure is definitely desirable.

The usual objective for the window width selection is to minimize some error criteria such as integrated mean square error (IMSE), mean square error (MSE), etc. However, we normally do not have knowledge of the underlying process that generates the data and these error criteria are not available to us in practice. One possible means of circumventing the problem is to replace the error criterion, IMSE say, by its consistent estimator and select a window width to minimize the sample estimate of this estimator. Given the fact that the window width, so chosen, minimized a consistent estimate of the IMSE, it is not difficult to establish that under proper regularity conditions the selected window width is optimal in the sense that it converges in probability to the actual minimizer of the IMSE. This approach gives rise to a paradigm that is known as Generalized Cross-Validation and we will describe a version of it, Least-Square Cross-Validation (LS-CV), in the next section. Later, the window width selection procedure is used as a benchmark in gauging the performance of a recursive estimator in our simulation study.

Another way to resolve the problem of unavailability of IMSE in practice is to substitute for it an approximation which is normally obtained from a Taylor series expansion This approximation is a function of the window width and, hence, an explicit solution for the optimal window width can be found. The main drawback of this approach is that it requires us to plug-in estimates of unknown parameters in order to obtain the optimal window width. Studies by simulation and asymptotic analysis alike have indicated that this type of estimator is more efficient and has more reliable small sample behavior than the cross-validation type.[2] However, there might be room for improvement.

[2]The study done by Gasser, Kneip and Kehler (1991) is an example.

One problem with the window width selection based on global error criterion such as IMSE is that the estimate tends to have increased bias near the peak of the regression function. This fact motivates a researcher to use local error criteria such as approximate mean square error (AMSE). The minimizer of this criterion is a window width that is dependent on the local value of some unknown functions which have to be consistently estimated through pilot estimation. In this sense, the estimator is adaptive.[3] We will further elaborate on these points and explain our rationale in choosing strategies for the implementation of the recursive estimator below.

Cross-validation Approach

The selection of window width has the objective of minimizing the error of the estimate in some sense. One common measure of error is the integrated mean square error, IMSE, that is $M(h) = E\{\int[m(x) - m_h(x)]^2 dx\}$, where the integration is over the interval of estimation. This error criterion is a global one which means that the estimator that minimizes it will represent the best fit to the underlying function over the interval of estimation. In practice, we do not have knowledge of the underlying regression $m(x)$ and such an objective is an ideal that is not attainable.[4] One way to get around the problem is to replace $M(h)$ by its estimate $\hat{M}(h)$ and one natural choice would be its sample counterpart, $\hat{M}(h) = \frac{1}{n}\sum_1^n (Y_i - m_h(X_i))^2$ which happens to be an inconsistent estimator for $M(h)$. The natural minimization solution for this criterion will be zero, that is, the best estimate for $m_h(X_i)$ is Y_i itself. Rudemo (1982) initiated the idea of using a leave-out estimator in the win-

[3]The term adaptive kernel nonparametric estimator might embrace a wider class of estimator which selects window width based on the local value of some functions. Some examples and a detailed asymptotic study can be found in the paper by Mack and Müller (1987). However, throughout this paper we reserved the term adaptive estimator for the type of window width selector that attempts to minimize the AMSE.

[4]The same scenario would apply to other criteria such as ISE, MSE etc.

dow width selection for kernel nonparametric estimation of density. The idea can be easily adapted to the regression setting. In this case the leave-one-out estimator can be $m_{h,-i}(X_i) = \frac{1}{n-1}\sum_{j\neq i} K(\frac{X_j - X_i}{h})Y_j$ which is a consistent estimator of $m(X_i)$ with Y_i deleted from the sample. The minimizer of such a criterion is one type of window width selector that is commonly referred to as Least-Square Cross-Validation (LS-CV).

Cross-validation score is defined as:

$$CV(h) = \frac{1}{n}\sum(Y_i - m_i(X_i))^2$$

where $m_i(X_i)$ is the leave-out estimator which is simply the kernel estimator of $E(Y|x = X_i)$ with Y_i taken out of the sample. The cross-validation window width selector is given by:

$$h_{cv} = arg \min_{h\in N_h} CV(h).$$

Usually the CV-score is not a smooth function of h and very often has multiple local minima. The usual strategy is to search for the minimizer of the CV-score over a grid of algebraically many h's.

In some ways, this approach can be viewed as a special case of the penalizing principle in which the original criterion is weighed by a penalizing function that penalizes enormously for zero window width. The idea is simply to make a correction for the bias toward zero through the working of the penalizing function. Such an estimator is referred to as Generalized Cross-Validation and there are several choices for the penalizing function (e.g Härdle (1990)). We will focus our attention on the LS-CV which shares similar properties with many other of the GCV variants.

The CV-type of window width selector is well explored. Its optimality, as we mentioned in the introduction, is established by Härdle and Marron (1986) and Härdle, Hall and Marron (1988). Although it displays asymptotic normality, it converges at

very slow rate, $O_p(n^{-\frac{1}{10}})$. Furthermore, normality is not a close approximation to its distribution when the sample size is moderate or small. Chiu (1991) shows that χ^2 of appropriate degree of freedom is a closer approximation. Besides our extensive knowledge of its behavior, the popularity of this type of estimator is also due to its directness and intuitive approach. It requires minimum knowledge about the underlying regression and assumes the least restriction on the smoothness of the regression function. This type of estimation is a data-driven procedure that needs no estimation of unknown parameters. It is automatically adapted to the data on hand. Recent studies by Hart and Vieu (1990), in the density estimation setting, and Härdle and Vieu (1990), for regression estimation, have shown that this type of estimator works well even on correlated data. The idea of CV can be extended easily to the problem of window width selection for derivatives whose value is not observable directly. Rice (1986) uses the differencing operation to provide a discrete approximation to the unobservable derivative's value and to show that the resulting cross-validation criterion works well just as in the case of ordinary cross-validation. Müller, Stadtmüller and Schmitt (1987) use higher order kernels and, combined with the idea of common factors, they manage to reduce the problem to one that selects the window width for regression. While it seems that the cross-validation technique has many virtues, it is not without flaws.

One main problem with the cross-validation technique is that it takes enormous computing efforts to obtain an estimate. One evaluation of the cross- validation function itself at a given window width will take computing operations by the order of n^2 (Here n being the sample size).[5] The matter is further complicated by the fact that the CV function itself has many local minima and an ordinary efficient

[5]That is, estimation with 20 observations is four times as expensive as estimation with 10 observations.

algorithm for optimatization does not work properly in this case. One would have to resort to a grid search over a net of algebraically many points, (that is $O(n^\gamma)$ points where γ lies in the interval $(0,1)$), just to ensure that optimality of the estimate is attained. Therefore, the computation cost for an estimate is of the order of $O(n^{2+\gamma})$. This problem has motivated researchers to design close substitutes for direct cross-validation, for example see Scott (1985) for averaged shifted histogram, see Härdle (1990) for the weighted averaging of rounded points, see Jon Breslaw (1992) for fast Fourier transformation leave-out estimator.

The wide sample variability is another main criticism of the CV-type of estimator. In fact the CV-function itself deviates randomly from $M(h)$ because it is just a consistent estimate of $M(h)$. In addition, the possibility of having multiple local minima could increase the variability. The CV-type selector, therefore, has a tendency to produce a poor estimate especially when the sample size is small. Scott and Terrell (1987), in the density estimation setting, propose including an additional term to penalize the choice of window width that has higher value for its second derivative. In the context of regression, this approach forces the selector to select a window width that yields an estimate with lower curvature. Although such an attempt has alleviated the problem, the criterion has strayed from the original idea of cross-validation. We will turn to the other approach on window width selection for a solution.

Plug-In Approach

Another approach to circumvent the problem of unavailability of the IMSE in practice is to replace it by its approximation. In chapter I, we showed through asymptotic analysis with a 2nd order kernel that the asymptotic bias and variance of the regression estimator are, respectively, given by

$$B(x) = \{h^2[r^{(2)}(x) - f^{(2)}(x)m(x)\} \int x^2 K(x)dx]f(x) + o(h^2)$$

$$V(x) = \frac{\sigma^2(x) \int K^2}{nhf(x)} + o(\frac{1}{nh}).$$

We can approximate the IMSE and obtain the minimizer of this AIMSE as below:

$$h_{opt} = \left[\frac{\int \sigma^2(x)f(x)dx \int K^2}{n \int [r^{(2)}(x) - f^{(2)}(x)m(x)]^2 dx [\int x^2 K(x)dx]^2}\right]^{\frac{1}{5}}$$

This method is generally referred to as the plug-in method since the unknown parameters such as $\int \sigma^2(x)dx$, $\int f(x)dx$, $\int r^{(2)}(x)dx$ etc. have to be estimated from the data and plugged into the above formulae for estimating h_{opt}.

The algorithm for the plug-in based estimator is not as straight forward as the CV's. It takes three stages, namely the pilot estimation of the parameters, followed by the window width estimation, then the regression estimation itself. Regression estimation uses Nadaraya- Watson fixed kernel estimation and the estimation of window width is by the formula described above where we now replace the parameters by their estimates.

The estimator has also been shown to be asymptotically optimal. The rationale behind this is that if the parameters are consistently estimated then the h_{opt} above optimizes a consistent estimator of an approximate IMSE that converges to the actual one asymptotically. Thus the h_{opt} so determined will converge asymptotically to the actual minimizer of the IMSE in probability. Gasser and Kohler (1990) developed an iterative plug in algorithm for a fixed design regression estimation. They demonstrate through asymptotic analysis that the relative rate of convergence for the plug-in and CV are the same, $O_p(n^{-\frac{1}{10}})$. However, the plug-in is superior in that it has smaller asymptotic variance. In kernel density estimation setting, Marron and Park (1990) have made a comprehensive comparison between the plug-in and CV type of window width selection. Their study, which includes asymptotic analysis and simulation,

indicates that on the whole the plug-in rule performs better especially for smaller sample sizes. They attribute the gain in efficiency of the plug-in method to the fact that it requires a stronger assumption on the underlying smoothness. In conclusion, they state their belief that the plug-in is the most practical method currently available while still feeling that there is room for improvement. We can easily bring forward these findings to the regression estimation setting.

Simple analysis[6] would lead one to the result that the integral of a consistently minimized function should be smaller than the minimum of the integral of the function, i.e.:

$$\int \left[\inf_h E\{[m(x) - m_h(x)]^2\}\right] dx \leq \inf_h \int E\left\{[m(x) - m_h(x)]^2\right\} dx.$$

Therefore, there may be gains in trying to minimize local error criterion like MSE instead of the IMSE. Following the plug-in approach as before, we arrive at the expression for the window width that minimizes the local MSE.

$$h_{opt}(x) = \left[\frac{\sigma^2(x) f(x) \int K^2}{n\{r^{(2)}(x) - f^{(2)}(x)m(x)\} \int x^2 K(x)dx]^2}\right]^{\frac{1}{5}}$$

Now, the optimal window width is a function of x with unknown parameters, $r^{(2)}$, σ^2, $m(x)$, $f(x)$ and $f^{(2)}(x)$. In the actual estimation these unknown parameters are replaced by their pilot estimates which gives rise to the question of adaptability. That is, can the estimator be as efficient as one where full knowledge of the underlying functions were available? The analysis that has been done by Mack and Müller (1987) has provided us with an affirmative answer. However, the task of adaptive nonparametric regression estimation is more involved than the usual plug-in type estimation described above. Here we are dealing with a local estimate of the derivative which should have larger sample variability than that of the integral of the

[6]Suppose $g(x,h) = E\{[m(x) - m_h(x)]^2\}$ and $g^*(x) = \inf_h g(x,h)$, then $g^*(x) \leq g(x,h)$ for all h. Hence, $\int g^*(x)\,dx \leq \int g(x,h)\,dx$ for all h which includes the required result.

derivative as in the case of the plug-in method. Moreover, numerical differentiation itself is inherently unstable, and even more so with noisy data. The irregularity in the estimation of the derivative can seriously mask the performance of the adaptive estimator. However, in the fixed design regression setting, Müller and Stadtmüller (1987) have shown, by asymptotic analysis and simulation, that so long as the parameters are consistently estimated the local adaptive window width selection scheme is always better than the global one. We will also implement a version of the plug-in based adaptive estimator in our simulation as a benchmark in gauging the performance of the recursive estimation.

4.2.3 Implementation of the Recursive Estimator

After describing the two window width selectors that we will use for comparison in our simulation study, we will examine the alternative for window width selection strategy in our recursive estimation. Since cross-validation always requires a complete data set, is computationally inefficient, and has relatively poor small sample performance, it is not a good choice as window width selection for our recursive estimation.

The plug-in based adaptive scheme is a better choice for a window width selection strategy in recursive estimation. It requires less computation for an estimate as it requires only the local estimates instead of the estimate of the integral of the underlying function, and it exploits all potential gains to their fullest. Of course this occurs at the expense of requiring more stringent underlying smoothness.[7] The simulation study done by Müller and Stadtmüller (1987) indicates that the local window width selector's small sample behavior is more reliable than the global one which is an advantage that is needed by the recursive scheme due to its dependency

[7]For a kernel estimation with 2nd order kernel this scheme will require estimation of the 2nd derivative which means the regression itself must be differentiable at least up to the 4th order just to ensure that the local 2nd order differentiate itself is estimatiable.

on the success of the initial estimate. Hence, we will adopt the adaptive based scheme as the window width selection strategy in the recursive estimation and we will explain the detail of the algorithm below.

Start-up

For any sample of size smaller than 50, the computation effort with the whole sample is minimal. Moreover, with a small sample we might try to use more expensive algorithms to compensate for the small sample. Hence, the recursive approach would not be justifiable. We will start with a sample size of 50 and perform the adaptive estimation which can be the same as the one described in next section. The outcome here is just an adaptive estimate of 50 observations.

Recursive Updating

The recursive algorithm is designed to incorporate information brought in by the newly arrived observation, the (n+1)th observation $\{X_{n+1}, Y_{n+1}\}$, into the existing estimates, m_n. The recursive estimator for the regression function is given by the ratio of two estimators, $m_n = \frac{r_n}{f_n}$, that can be recursively estimated:

$$r_n = r_{n-1} + K(\frac{x - X_n}{h_n})\, Y_n.$$

$$f_n = f_{n-1} + K(\frac{x - X_n}{h_n}).$$

Where the window width is determined by the past observations. The density estimator at x, $\hat{f}(x)$ is given by: $\frac{1}{nh} f_n$. And the local variance, S_n, estimator of $E(y^2|x) - (E(y|x))^2$, is recursively estimated by:

$$YY_n = YY_{n-1} + K(\frac{x - X_n}{h_n})\, Y_n^2$$

$$S_n = \frac{YY_n}{f_n} - m_n^2.$$

To simplify the computation of the estimates of the derivative, we treated the sequence of estimates $\{x_i, m_n(x_i)\}, i = 1,..,125$, as equidistant design points and estimated the second derivative by:

$$\hat{r}_n^{(2)}(x) = \frac{1}{n_x h_x^3} \sum K_{(2,6)}(\frac{x - x_i}{h_x}) m_n(x_i)$$

where $K_{(2,6)}$ is the optimal kernel of order (2,6). [8]

The summation is over all points of estimation, $\{x_i\}$, where h_x is fixed at $0.95ss_x$ and x is a point at which we carry out out the regression estimation. The sequence of x's, $\{x_i\}$ is a deterministic sequence with constant standard deviation, ss_x. In this algorithm, it is only $\hat{r}_n^{(2)}$ that changes with the observations through the sequence of estimates $\{m_n(x_i)\}_i$.

With all these parameters estimated we proceed to select the 'optimal' window width for the next observation:

$$\hat{h}_{opt}(x) = \left[\frac{\hat{\sigma}^2(x) \int K^2}{n\hat{f}(x)[\hat{r}^{(2)}(x) \int x^2 K(x)dx]^2}\right]^{\frac{1}{5}}.$$

The procedure is repeated each time a new observation is realized.

The whole procedure can be summarized by the algorithm below:

Step 1 Initial estimation

An adaptive estimation of size 50 with the first 50 observations.

---

[8]The 2 indicates the order of derivative and 6 indicates the degree of smoothness of the kernel function. It is a polynomial over the interval [-1,1] satisfying

$$\int_{-1}^{1} K(u)u^j du = \begin{cases} 0 & j=1,..,3 \\ \frac{1}{30} \int K^2 & j=4. \end{cases}$$

$$\int_{-1}^{1} K^{(2)}(u)u^j du = \begin{cases} 0 & j=0,1,3,..,5 \\ 2 & j=2 \\ \int K^2 & j=4 \end{cases}$$

and minimizes the approximate asymptotic mean square error. See for example, Müller (1979) for details.

Step 2 Recursive estimation

DO LOOP (with the rest of the observations)

(1) Sampled one observation from the sequence and perform the recursive updating of r_n, f_n, YY_n, S_n and m_n.

(2) Estimation of $r_n^{(2)}$.

(3) Estimation of optimal h for the next observation.

ENDLOOP.

4.3 Model and Estimation methods in the Monte Carlo Simulation

The purpose of the simulation is of twofold, one is to make comparison among the three methods. The other is provide a detailed analysis of the properties of the finite sample behavior of the recursive estimator. In this section we outline the design of our model and explain the methods of kernel estimation in our simulation study. The recursive estimator described in this section is one way of choosing the window width. We gauge the recursive estimator's performance by comparing the distance of the estimates it generates from the true curve with that of the other two window width selectors described in last section.

4.3.1 Model

Regression Curve

We assume that the underlying nonparametric regression model is of the form:

$$y = m(x) + \epsilon$$

where $\epsilon \sim N(0, 0.1)$, $Var(Y|X = x) = \sigma^2(x) = \sigma_\epsilon^2$=0.1 and

$$m(x) = x sin(2\pi x) exp(\frac{1}{2}(2x - 1)^2).$$

This regression curve has one mild turning around x=0.35, a sharper turning at around $x = 0.8$ and a linear portion around x=0.5. It is designed to show the performance of the estimators in capturing these features. We can assess the impact of the regression curve on the estimation by examining the estimates at different regions of the regression curve for x between 0 and 1.

Data Generating Process (DGP) for the regressor, X's

We examine three scenarios. In each case the regressor $\{X_i\}$ is sampled from one of the DGP's described below:

Independent Uniform $X \sim U[0, 1]$. The situation here is equivalent to unequidistant, fixed design regression. As several published results are cast in a fixed design setting, we would like to generate results in a comparable setting.

Independent Normal $X \sim N(0.5, 0.5^2)$. The Gaussian data generating process is standard for most simulation studies. Here we choose a variance so that roughly one third of the points are outside the interval of estimation, $[0, 1]$. It represents the situation where the distribution of the regressor is not evenly spread over the interval of estimation.

Correlated Normal X follows a Gaussian AR(1) process. That is $X_i = \rho X_{i-1} + \sqrt{1 - \rho^2} Z_i$, for $i > 1$ and $X_1 = Z_1$. Where Z_i's are independent $N(0.5, 0.5^2)$ with $\rho = 0.25$.

Since we have established the asymptotic properties of the recursive estimator with a strong mixing regressor in Chapter Two we would like to choose a process that has its link with the strong mixing process. To this end, we borrow a result from

creased to 1000, the region is further narrowed to intervals: $(0.12, 0.2)$, $(0.56, 0.61)$ and $(0.9, 0.96)$ for cross-validation. For the adaptive, the region consists of intervals, $(0.12, 0.2)$, $(0.56, 0.61)$ and $(0.72, 0.92)$.

In general, the recursive estimator agrees well with the other two benchmark estimators in the vicinity of a inflexion point, that is the point at which the second derivative is zero, where the absolute bias converges to zero at a rate smaller than h_n^2. The picture becomes even clearer when a normal regressor is used.

Independent Normal At n=100, the recursive estimator agrees well with the other two generally in the neighbourhood of x=0.16 and x=0.56. To be exact, intervals $(0.04, 0.26)$, and $(0.48, 0.65)$ for the adaptive and intervals $(0.08, 0.24)$ and $(0.52, 0.60)$ for the cross-validation. The interval in which the recursive agrees with the other estimators narrowed as the sample size increased. At n=350, the intervals become $(0.1, 0.26)$ and $(0.53, 0.64)$ for the adaptive and $(0.13, 0.19)$ and $(0.54, 0.58)$ for the cross-validation. For the adaptive, the intervals narrowed to $(0.13, 0.24)$ and $(0.54, 0.58)$ when the sample size increased to 1000. For the cross-validation, at n=1000, the intervals also narrowed to $(0.14, 0.18)$ and $(0.56, 0.58)$, however, the recursive becomes marginally better off in the interval $(0.7, 0.91)$.

We would expect to lower the rate of convergence for all the estimators with correlated regressors. Hence the departure of the recursive from the the benchmark estimators would be more gradual.

Dependent Normal At n=100, the recursive is not different or is marginally better than the other two over a relatively wide region, $(0, 0.68)$ excluding interval $(0.4, 0.5)$ for the adaptive, and $(0, 0.62)$ excluding interval $(0.4, 0.51)$ for

Godoretskii (1977) in our design of correlated regressor's DGP. He has shown that if an independent Gaussian process $\{\nu_i\}$ links to a process $\{X_i\}$ by the following relationship, for $|\rho| < 1$,

$$X_i = \rho X_{i-1} + \nu_i,$$

then the process $\{X\}$ is an α-mixing process with mixing coefficient $\alpha(m) = O(m|\rho|^{-m})$.

In our correlated normal case, each X_i is a $N(0.5, 0.5^2)$ and $\rho(X_i, X_{i-1}) = 0.25$ for all i, then, according to the result stated above, the process $\{X_i\}$ forms an α-mixing sequence with mixing coefficient $\alpha(m) = O(m|4|^{-m})$.

Implementation of the experiment

The whole experiment is implemented in FORTRAN and run on a VAX 6430 computer with VMS as operating system. The three DGP's described above are coded in three subroutines namely IU(X,Y,N), IN(X,Y,N) and DN(X,Y,N) [9] for generating observations with the regressor sampled from independent uniform process, independent normal and dependent normal processes respectively. The uniform and standard normal variates are respectively simulated by subroutines DRNUN and DRNNOR from the IMSL/STAT library. For the independent uniform regressor, the X values are generated directly by the subroutine DRNUN. Independent normal X values are generated by changing the mean and variance of the variable generated by DRNNOR. For the AR1 regressor, we generate N independent normal variates by DRNNOR as before. Then we use a do loop to generate the AR1 regressor, $(X(J) = 0.25\ X(J-1) + \sqrt{1 - 0.25^2}\ Z(J))$. The Y is then generated from the corresponding functional value of the X plus an error term. The error term is a normal variate that is generated by another call to the subroutine DRNNOR.

In addition to the three variations in DGP for the regressor, we also use three different

[9] N is the length of the observations specified by the user and X,Y are output variables that stored the simulated observations.

sample sizes, namely 100, 350 and 1000 to form 9 controlled experiments. In each experiment, we generate a thousand replicates.[10] For each replicate we obtain the estimate of the nonparametric regression of 125 distinct points in the $(0, 1]$ interval by three methods that we will describe below. We write the one thousand estimates at each x to a file. For each estimator, we generate 125 files, one at each x point in the grid of 125 equally spaced points in (0,1]. We also report the minimum, median, maximum, mean and variance of the thousand estimates at each estimation point in a separate file. There are altogether 125 entries of the five statistics of the thousand estimates for each of the three methods of estimation in each contingency.

As the amount of disk storage required for a thousand different samples can be substantial, the required data is generated directly in the program that processes it. In order to control the generation of random sequences such that each time the same sequence is used for each of the three estimators, we use RNSET to set the seed to 123457 for the random number generation. We then use RNUND to generate one thousand different integers which in turn are used as seeds to generate the required thousand sequences of random numbers for the thousand replications in the simulation study. We subject a given sequence generated by a method described above to the three mentioned estimators, namely, the cross-validation, the adaptive and the recursive.

[10] A replicate is a pseudo random sequence generated by a call to the respective subroutines described above.

4.3.2 Implementation of the Estimators in the Simulation Study

We apply the kernel estimator to estimate the regression at a sequence of 125 equally spaced points over an interval [0,1] and we denote this sequence as $\{x_i\}_{i=1}^{125}$. We have described the algorithm for the implementation of the recursive estimator in the last section. In this subsection we describe the algorithm for the implementation of the two benchmark estimators and specify some details of the implementations.

We use the Gaussian kernel function for our regression estimation.

Cross-Validation

The direct evaluation of the CV score at a given h is computationally expensive and we implement Jon Breslaw's (1992) version of the Fast Fourier Transform (FFT) leave-out estimator in our computation for the CV-score. In our implementation of the cross-validation selector, we probe over 300 points in an interval $[0.3\sigma_x n^{-0.2}, 1.5\sigma_x n^{-0.2}]$, where σ_x is the standard deviation of the realization of the process for the regressor, $\{X_i\}_{i=1}^{n}$.

The procedure is described in the following algorithm:

Step 1 Discretize the data into 1000 points over an interval, [0,1] for the independent uniform regressor and [-1,2] for the normal regressor. The discretized data is then padded with 23 zero's and one unit weight placed at the end to makeup a sequence of 1024 (2^{10}) points.

Step 2 Transform the discretized data by a FFT subroutine.

Step 3 DO LOOP { over a grid of h values }

(i) Compute the convolution of the transformed data and the kernel sequence at each h.

(ii) Invert the convoluted data and use the last point as K(0) to obtain the leave-one-out estimates.

(iii) Compute the CV score and identify the minimum score and the width that generated it.

ENDLOOP

Step 4 Use the minimizer of the CV score as window width to obtain the usual kernel regression estimate.

Adaptive

We estimated five parameters in the pilot estimation which requires some explanation. In the pilot estimation, we start off by using some heuristic rules such as setting window width $h = \sigma_x n^{-0.2}$ for the pilot estimation of the regression $\hat{m}(x)$, local density $\hat{f}(x)$, and the local variance $\hat{\sigma}(x)^2 = \frac{\sum K(.)Y^2}{\sum K(.)} - \hat{m}^2$. The derivatives $\hat{f}^{(2)}(x)$ and $\hat{r}^{(2)}(x)$ are estimated by an optimal kernel of order (2,6), $K_{(2,6)}(x) = \frac{3465}{512}(-117\ x^8 + x^6 - 270\ x^4 = 84\ X^2 - 5)$.[11]

Algorithm:

Step 1 Pilot estimation

Use rule $h = \sigma_x n^{-\frac{1}{5}}$ to estimate $\hat{m}(x)$, $\hat{f}(x)$ and $\hat{\sigma}^2(x)$.

use rule $h = \sigma_x n^{-\frac{1}{13}}$ to estimate $\hat{f}^{(2)}(x)$ and $\hat{r}^{(2)}(x)$.

Step 2 Plug-in estimation of window width.

$$\hat{h}_{opt}(x) = \left[\frac{\hat{f}(x)\hat{\sigma}^2(x)\int K^2}{n[(\hat{r}^{(2)}(x) - \hat{m}(x)\hat{f}^{(2)}(x))\int x^2 K(x)dx]^2}\right]^{\frac{1}{5}},$$

where $\int K^2 = \frac{1}{2\sqrt{\pi}}$ and $\int x^2 K(x)dx = 1$ for the Gaussian kernel function.

[11] Müller (1988),p.69.

Step 3 Re-estimation

use the optimal h obtained in 2 to estimate $\hat{m}(x)$, $\hat{f}(x)$ and $\hat{\sigma}^2(x)$ again.

Step 4 Repeat step 2 and 3 one more time to get better estimates of the parameters of the optimal h and in turn a closer optimal h estimate. The resulting $\hat{m}(x)$ is then the adaptive estimate of the regression at x.

Recursive Estimation

In the simulation study, we implement the recursive estimator as described in the last section and with Gaussian kernel and optimal (2,6) kernel as specified above.

4.4 Analysis of Results from the Monte Carlo Study

The recursive algorithm is specifically designed to perform continuous estimation with sequential data flows. In this simulation study, we are particularly concerned with the effect of non-uniform window width on the performance of the estimator with a finite sample and the recursive estimator's finite sample behavior in relation to the behavior predicted by the asymptotic analysis. In the subsections below, we describe our methods for comparison and we report the outcomes and draw inferences pertaining to the properties of the estimators. We also examine more closely the finite sample behavior of the recursive estimator in relation to the asymptotic results obtained in earlier chapters.

4.4.1 Comparison of the Recursive With the Other Two Nonrecursive Estimators

In the simulation we subject the same computer generated pseudo random sample to the three estimation methods described earlier: the recursive estimator, adaptive estimator and the cross-validation estimator. For each of the controlled experiments mentioned in the last section, we replicate each of the three estimations at the 125 points over (0,1] a thousand times. We attempt to draw inferences about the recursive estimator's optimality through a comparison of its performance with the two representative optimal fixed window width kernel estimators described in the earlier section. Quantitatively, a better estimator is one with smaller average mean square error, significantly smaller absolute average bias over a large portion of the estimation interval, and which generated estimates with smaller dispersion.

In the simulation study, we have the true regression function and thus we can determine the actual absolute bias of each estimate. The absolute bias is defined

as $|m_n(x_i) - m(x_i)|$ where $m_n(x_i)$ is the estimate and $m(x_i)$ is the true value of the regression at x_i. We judge if the distribution of the absolute bias of the estimates at each of the 125 estimation points by the recursive estimator is significantly different from that of each of the two benchmark estimators through a Wilcoxon matched-paired signed-ranks test. Then we visually examine the curve of the average bias and the standard deviation of the estimates by each of the three estimators. Finally, we compare the overall mean square error of the three estimators. The mean square error of the estimator is approximated by its sample counterpart:

$$\frac{1}{1000 \text{ x } 125} \sum_{r=1}^{1000} \sum_{i=1}^{125} \{m_n^r(x_i) - m(x_i)\}^2$$

where r is the replicate number and x_i is one of the 125 estimation points.

We also plot the curves of the maximum, minimum, median and mean in a graph to make comparisons across estimators. The subsections below report the outcomes and draw inferences about the finite sample properties of the recursive estimator with respect to the two benchmark estimators.

The comparison of the three estimators at each estimation point

We attempt to find out whether the error in the estimate generated by the recursive estimator is significantly different from that of the other two benchmark estimators. This amounts to testing the equality of central tendency of the absolute bias of the estimate by the recursive method with that of the other two estimators. For this purpose, we apply the Wilcoxon test to test the equality of the median of the absolute bias of the estimate by the three estimators two at a time, that is, recursive vs adaptive and recursive vs cross-validation. We conduct the Wilcoxon test for each of the two pairs at each of the 125 estimation points over (0,1] for all the nine experiments considered. The Wilcoxon test is employed for its robustness against the violation of

the normality assumption. That is, it is less efficient than the standard paired-t test under the normality assumption, however, its efficiency does not deteriorate even if the assumption is violated.

The outcome of the test is, for most cases, either the recursive is significantly worse than the other, (adaptive or cross-validation), or there is no significant difference. In some experiments, the recursive is slightly better over some small regions. We marked on the estimation interval, (0,1], regions where the test fails to reject the null hypothesis of no difference and the region where there is significant difference between the two methods with the sign, that is (+) when the other method is better and (-) when the recursive is better. For each of the nine experiments, the results on the 125 tests for the recursive vs each of the two benchmark methods are summarized by a line. There are two lines for each experiment and eighteen in total for the nine experiments. All these lines together with the true regression curve are plotted on a graph, Figure 4.1. The paragraphs below highlight some features we observed from the results.

Independent Uniform At n=100, except at the two ends, $x < 0.08$ and $x > 0.93$, the absolute bias of the recursive estimator is not significantly different from that of the adaptive estimator over a large portion of the (0,1] interval. The cross-validation estimator is significantly better than the recursive estimator over the interval $(0.3, 0.56)$ and one of the end portions $x > 0.94$.

As sample size increased, the portion where the recursive estimator is not significantly different with the other estimators was reduced. At n=350, the recursive is not different from the cross-validation only at intervals, $(0.12, 0.264)$, $(0.53, 0.64)$ and $(0.8, 0.92)$. For the adaptive, the region of no difference is slightly more extensive, $(0.12, 0.256)$ and $(0.53, 0.92)$. When the sample size in-

creased to 1000, the region is further narrowed to intervals: $(0.12, 0.2)$, $(0.56, 0.61)$ and $(0.9, 0.96)$ for cross-validation. For the adaptive, the region consists of intervals, $(0.12, 0.2)$, $(0.56, 0.61)$ and $(0.72, 0.92)$.

In general, the recursive estimator agrees well with the other two benchmark estimators in the vicinity of a inflexion point, that is the point at which the second derivative is zero, where the absolute bias converges to zero at a rate smaller than h_n^2. The picture becomes even clearer when a normal regressor is used.

Independent Normal At n=100, the recursive estimator agrees well with the other two generally in the neighbourhood of x=0.16 and x=0.56. To be exact, intervals $(0.04, 0.26)$, and $(0.48, 0.65)$ for the adaptive and intervals $(0.08, 0.24)$ and $(0.52, 0.60)$ for the cross-validation. The interval in which the recursive agrees with the other estimators narrowed as the sample size increased. At n=350, the intervals become $(0.1, 0.26)$ and $(0.53, 0.64)$ for the adaptive and $(0.13, 0.19)$ and $(0.54, 0.58)$ for the cross-validation. For the adaptive, the intervals narrowed to $(0.13, 0.24)$ and $(0.54, 0.58)$ when the sample size increased to 1000. For the cross-validation, at n=1000, the intervals also narrowed to $(0.14, 0.18)$ and $(0.56, 0.58)$, however, the recursive becomes marginally better off in the interval $(0.7, 0.91)$.

We would expect to lower the rate of convergence for all the estimators with correlated regressors. Hence the departure of the recursive from the the benchmark estimators would be more gradual.

Dependent Normal At n=100, the recursive is not different or is marginally better than the other two over a relatively wide region, $(0, 0.68)$ excluding interval $(0.4, 0.5)$ for the adaptive, and $(0, 0.62)$ excluding interval $(0.4, 0.51)$ for

the cross-validation. At n=350, for the cross-validation, the region is narrowed to intervals $(0.11, 0.224)$ and $(0.56, 0.60)$. For the adaptive, the region is reduced to intervals $(0.06, 0.28)$ and $(0.56, 0.7)$. With sample size increased to 1000, the region further shrank to intervals, $(0.14, 0.18)$, $(0.56, 0.6)$ and $(0.71, 0.74)$ for the cross-validation and $(0.11, 0.19)$ and $(0.56, 0.66)$ for the adaptive.

The curves of Average bias and the finite sample standard deviation of the three estimators

Instead of looking at the thousand estimates at each estimation point, we now turn to the summary statistics of the thousand estimates such as mean and standard deviation for comparison. We plot the curves of the average bias and standard deviation of the estimates by the three methods in a graph for each of the nine experiments considered. In this subsection, we attempt to draw inferences on the properties of the recursive estimator in relation to the two benchmark estimators through visual examination on the these curves.

Average Bias

The general features of the average bias curves for the experiments considered are essentially the same. We only exhibited the case of independent normal regressor with sample size n=100 and n=1000 as Figure 4.2. All the three curves corresponding to the three different estimators in the graphs can be roughly partitioned into three portions. The interval, $x < 0.16$, the average bias of each of the three estimators is positive. In between $x = 0.16$ and $x = 0.58$, the average bias of each of the three is negative, and for the rest of the $(0, 1]$ the average is positive. For the normal regressor, however, the humps corresponding to the vicinity of the turning points of the regression are bigger than that of the

uniform regressor. For the uniform regressor, there is an irregularity near the end, x=1.

The cross-validation curves in general reduce fairly uniformly over the whole interval of estimation, whereas the adaptive curve reduces more quickly at the vicinity of the turning points. The three curves meet at the inflexion points, x=0.16 and x=0.58. At n=100, the three curves are of marginal difference and the recursive is closer to the adaptive. However, at n=1000, the adaptive curve moves quickly to flatten its humps. For the cross-validation estimator, its average bias curve at n=1000, becomes closer to the recursive. The difference between the recursive curve with the other two also becomes more obvious than at n=100. Perhaps, the reduction in the bias with the recursive estimator is just not as fast as the other two.

Standard Deviation

As we observed the curves for different regressors generally have similar features and we showed only the case of independent normal regressor with sample size n=100 and 1000 as Figure 4.3. The standard deviation curve of the cross-validation in general is flat in between x=0.2 and x=0.6 and gradually sloping upward at the two end portions. For most cases considered, the three curves meet around x=0.2 and x=0.6. The adaptive's standard deviation fluctuates around the cross-validation's and the standard deviation of the recursive is just as wavy as that of the adaptive.

At n=100, the standard deviation of the recursive comes closer to the that of the adaptive which has bigger humps than the cross-validation and generally lies above the cross-validation as well. However, as sample size increases, the adaptive curve moves quickly to flatten its humps and moves below the cross-

validation at around x=0.8. The reduction in the cross-validation curve with increased sample size is more gradual and uniform. However, its standard deviation is essentially smaller than the other two even at n=1000. The standard deviation of the recursive is usually smaller than the other two around the points of inflexion, x=0.16 and x=0.58. However, it is always larger elsewhere.

The results indicate that the recursive estimator does not converge to zero as fast as the other two. As the sample size increases, the recursive estimator applies a smaller window width to the most recent observation and discounts the information of the most recent observation more than all the past observations. This feature is required to ensure consistency of the estimator. However, it is also responsible for its slow convergence and hence the reduction in efficiency of the estimator especially for the case where the initial estimator is far from ideal. For example, for the normal regressor at x=0.95, the estimated bias of the recursive estimator is 0.16 at n=100, 0.14 at n=350 and 0.12 at n=1000.

Mean Square Error of the Estimations

The mean square error (MSE) of the three methods in general decreases with increased sample size. There is significant increase in the MSE when the regressor is switched from independent uniform variate to independent normal variate. However, the increase in the MSE was only marginal when we introduced correlation in the normal regressor. The difference due to different methods of estimation is also marginal in absolute terms. However, the relative difference in the mean square error of the recursive estimator to that of the other two estimators became significant as sample size increased. We compile the estimated average MSE for various controlled experiments in Table 1.3a. In table 1.3b, we report the relative efficiency of the re-

cursive estimator with respect to each of the two benchmark estimators. The relative efficiency means the ratio of the mean square error of one of the other two estimators to that of the recursive under the same controlled situation, that is, $\frac{M.S.E._o}{M.S.E._{rec}}$, where o can be adaptive or cross-validation.

Table 1.3a Average Mean Square Error for all 125 Estimates

DGP	n	100	350	1000
Independent Uniform	*Adaptive*	0.011699	0.001966	0.001781
	Cross – Validation	0.01183	0.004490	0.002125
	Recursive	0.01265	0.006615	0.002866
Independent Normal	*Adaptive*	0.018278	0.00449	0.002125
	Cross – validation	0.014795	0.005850	0.002546
	Recursive	0.0214481	0.008684	0.003977
Correlated Normal	*Adaptive*	0.018014	0.005378	0.001966
	Cross – validation	0.015841	0.006764	0.002358
	Recursive	0.02372	0.01020	0.004796

Table 1.3b Relative Efficiency of the recursive with respective to the other two estimators

DGP	n	100	350	1000
Independent Uniform	*Adaptive*	0.925	0.297	0.621
	Cross – Validation	0.935	0.679	0.741
Independent Normal	*Adaptive*	0.852	0.517	0.534
	Cross – validation	0.690	0.674	0.640
Correlated Normal	*Adaptive*	0.759	0.527	0.410
	Cross – validation	0.668	0.663	0.492

The recursive estimator is strictly dominated by both cross-validation and adaptive estimators. The gaps widen as the DGP for the regressor goes from independent uniform variate to independent normal variate. The correlation in the regressor increases the gap ever further. The gaps narrowed with increased sample size which means the recursive estimation converges to the same limit as the other two but at a slower rate. However, in terms of relative efficiency, the proportion in reduction of the mean square error can be significant. For the case of correlated normal with n=1000, the mean square error of the adaptive and cross-validation is respectively 0.410 and 0.492 of that of the recursive estimator, which indicates a significant loss

of efficiency in the recursive estimation. The results are consistent with what we had pointed out in the earlier subsection. The recursive estimator usually applies larger discounts to the information contained in the later observations and consequently converges at a slower rate to the true value. Hence, we could not apply the recursive scheme indefinitely as the loss in efficiency of the estimation can be substantial.

The curves of mean, median, maximum and minimum of the estimates

We plotted the curves of the mean, median, maximum and minimum of the estimates generated by different methods and compared them visually. The true curve is always above the minimum curve and below the maximum curve. This is normally expected of an estimator that does not consistently over or under estimate the regression function. At n=100, the maximum and minimum curves of the recursive estimator resemble that of the adaptive, Figure 4.4, however, the reduction in the range of the recursive is more gradual with sample size increases, Figure 4.4 and 4.6. At n=1000 Figure 4.6, the range of the recursive estimates is generally closer to the cross-validation's, however, they are as spiked as the adaptive's when n=100.

Generally, the mean curve of the adaptive estimates is closer to the true curve than the cross-validation estimates, and the recursive estimates are normally worse than the other two methods.

At n=100, all the three estimators display a visible difference between their median and mean curves. For the adaptive, the difference diminished at n=350, Figure 4.5. The reduction in the gap between median and mean is slow for the case of cross-validation estimator, still its median curve almost coincides with its mean curve at n=1000, Figure 4.6. For recursive estimator the gap between the median curve and the mean curve does reduce with increased sample size, but it is not eliminated at n=1000.

These observations reflect the fact that the recursive estimate improves very gradually with sample size increases and its behavior is very much dependent on the initial estimation. This point could also explain the fact that even for a sample size of 1000, the empirical finite sample distribution of the recursive estimator does not well approximate a normal distribution.

4.4.2 Finite Sample Properties of the Recursive Estimator

In the last subsection, we were concerned with the finite sample performance of the recursive estimator in comparison with the other two non-recursive estimators. In this subsection, we study the finite sample properties of the recursive estimator, in particular the behavior of the estimator's finite sample bias, variance, and normality. The issue we focus on is how well these finite sample properties accord with the predicted properties of the asymptotic analysis.

Finite sample average bias and standard deviation of the recursive estimator

We plot statistics (average absolute bias or standard deviation) for the three sample size (100, 350, and 1000) in a graph for visual examination. All the graphs except the average absolute bias for the independent uniform regressor show clearly gradual reduction in standard deviation, and bias with increased sample size, graph Figure 4.7 for example. As for the case of independent uniform regressor, Figure 4.8, the curves can roughly be divided into two portions with the first portion, $x < 0.6$ showing no significant differences among the three curves and for the remaining portion, $x > 0.6$ the average bias of the estimators do clearly converge to zero with increased sample size. It means the rate of convergence for independent uniform regressor is slower than expected.

To gather some idea about the rate of convergence we use the summary statistics to gauge the rate of convergence:

the average absolute bias, $b_n = \frac{1}{1000 \times 125} \sum_{r=1}^{1000} \sum_{i=1}^{125} |m_n^r(x_i) - m(x_i)|$, and

the average standard deviation, $s_n = \frac{1}{125} \sum_{i=1} 125 s(x_i)$, where $\{x_i\}$ is the sequence of estimation points and $s(x_i)$ being the standard deviation of the thousand estimates at x_i.

We further assume that $b_n = c_b n^{-\gamma_b}$, $s_n = c_s n^{-\gamma_s}$, for c_b and c_s being constants independent of n and the γ is the rate of convergence.

We then regress $ln\ b_n$, $ln\ n$ on n to obtain the respective rate of convergence, γ and the results are shown as table 2.1.

Table 2.1 **Estimated rate of convergence of the average bias and standard deviation**

DGP	γ_b	γ_s	c_b	c_s
Independent Uniform	0.078	0.428	0.055	0.669
Independent Normal	0.253	0.433	0.246	0.832
Correlated Normal	0.253	0.421	0.230	0.841

The results in table 2.1, are just rough indication of the way the standard deviation and bias are converging. The optimal rate of convergence in mean square error for the case of independent univariate regressor is 0.8 and lower for a correlated regressor. The corresponding rate of convergence for the standard deviation and absolute bias is 0.4 and the rates in the table 2.1 seem to indicate that the standard deviation of the cases considered converges faster than 0.4. However, the rate of convergence for the bias is well below the optimal, 0.4. These results indicate that the window width used in the recursive estimation is often larger than the optimal window width on the average. This may well be that in the recursive estimation the optimal window width applies only to the most recent observation and the window widths applied to the past observations are usually larger than the optimal.

Asymptotic variance and the variance of the finite sample estimator

Usually, we can have only one sample and we approximate the variance of the finite sample estimator by the estimate of its asymptotic variance. In this subsection, we try to find out how well the asymptotic variance agrees with the sample variance estimated from the thousand replicates that we obtained through the experiment. As we have derived in chapter 1, $Var(m_n(x)) = \frac{1}{nh_n(x)}\tau\frac{Var(Y|x)\int K^2}{f(x)}$ where $\tau = \frac{1}{n}\sum\frac{h_{ni}}{h_i}$.[12] the expression for the asymptotic variance of the recursive estimator contains a multiplicative factor τ. This factor is the result of using different window widths with different observation and it is dependent of the data sampled, which cannot be determined analytically. However, the factor is asymptotically bounded by some finite constant which is 1 in this case, and so the asymptotic variance for the recursive estimator should be roughly equal to the fixed window width estimator. In the simulation, we used the sample variance of the thousand estimates generated by the thousand replicates to approximate the variance of the finite sample estimation at the 125 estimation points. Then we attempted to find out how closely the finite sample variance of the recursive estimator agreed with the asymptotic variance of a fixed window width estimator. For comparison between the two variances, we plotted the asymptotic standard deviation evaluated according to the procedure above and the estimated standard deviation of the finite sample estimator on the same graph, Figure 4.9.

Generally, the asymptotic standard deviation evaluated analytically has dips corresponding to the points of inflexion of the regression curve, roughly x=0.15 and x=0.56. It also has humps at the turning points of the regression curve, x=0.35 and x=0.81. The curve for the sample standard deviation of the recursive estimates has

[12] For univariate regression function, d=1, with an independent process for the regressor.

a similar shape on the whole. However, it lies above its asymptotic counterpart most of the times. The portion of curve corresponding to the turning point of the regression curve has significant vertical noise and becomes more pronounced around the sharper turning point. For the case where the regressor follows a Gaussian process, independent and AR1 alike, the noisy portion around the sharper turning point of the regression , x=0.8, turns into a pointed peak at around x=0.915 when sample size increases. This phenomenon can be attributed to the instability in the estimation of derivatives of the regression and density functions. Normally, the estimation of derivative can be unstable even for deterministic sequences not to mention the noisy random sequence in our case. Therefore, the observed variance of the finite sample estimates is larger than the asymptotic variance. For the case of the uniformly distributed regressor, we observed pronounced a 'boundary effect' at the two end points, x=0 and x=1. This effect is often reported by studies on kernel nonparametric regression estimation in fixed design setting.

Other than the peaked portions around the turning points of the regression and the boundary points of the support of the marginal density of the regressor, the discrepancy between the two curves is well within a margin of error of 0.02. In practice, we usually estimate the asymptotic variance through estimation of its parameters. Alternatively, it is possible for us to estimate finite sample variance directly through resampling techniques. In any case, the estimation of the variance itself would incur an error that is likely to exceed the discrepancy between variance predicted by the asymptotic theory and the actual variance of the finite sample estimator.

Normality of the finite sample recursive estimation

Here, we focus on the distribution of the estimates generated by the recursive estimator. We attempt to verify the hypothesis that the estimate is sampled from a normal

population. To this end we employ the Kolmogorov-Smirnov goodness of fit test. The test is generally more powerful than the usual chi-square goodness of fit test, however, it requires a simple null hypothesis so that the hypothesized distribution has to be completely specified. We choose to compare the distribution of the estimates with a normal distribution with the same mean and variance as the corresponding sample mean and variance rather than asymptotic normality,[13] since if the test rejected the null hypothesis of asymptotic normality then it is not altogether clear that the rejection is the result of the difference in the mean and/or in the variance or due to the departure from normality. Whereas if the test rejected normality with the corresponding sample mean and variance then there is no question of asymptotic normality. Furthermore, with a large sample, the power of the test is sufficiently high and we can afford to lower the level of significance.[14]

We choose the point x=0.12, where the sample variance agrees well with the analytically obtained asymptotic variance in most of the experiments considered. Beside, we also conducted the same test at points x=0.312, x=0.552 and x=0.808 and observed similar outcomes as the one we report. Hence we do not repeat them here. The results of the Kolmogorov-Smirnov test are compiled in table 2.3.

Table 2.3 Kolmogorov-Smirnov Test on the normality of the recursive

[13] A normal distribution with mean and variance predicted by the asymptotic theory.

[14] Lilliefors (1967) observed that the standard Kolmogorov-Smirnov test with composite null, that is, the parameters of the hypothesized distribution are estimated from the sample, tends to accept the null more often than it should.

estimates

DGP	n	Abs Diff	K-S score	p-value
Independent Uniform	100	0.03410	1.078	.195
	350	0.04164	1.317	.062
	1000	0.03747	1.185	.121
Independent Normal	100	0.03734	1.181	.123
	350	0.04130	1.306	.066
	1000	0.04198	1.327	.059
Correlated Normal	100	0.04073	1.288	.073
	350	0.03086	0.976	.297
	1000	0.04965	1.570	.014

Of all the experiments considered, the null hypothesis of normality can not be rejected, at 1% level of significance. However, according to Lilliefors, the adjusted critical value for the K-S score at 0.1% level is 1.22 which means the observed sampling distribution of the estimates exhibit significant departure from normality even at 0.1% level of significance in some cases. The cases where the null hypothesis is rejected are for independent uniform regressor, n=350 , for independent normal regressor, n=350 and 1000 and for dependent normal regressor, n=100, and 1000. These results show that the evidence for normality of the estimate's sampling distribution is weak. We conclude from these results that the asymptotic normal approximation is not a close fit.

We plotted the empirical distribution of the standardized estimates [15] alongs with the distribution of the standard normal ($N(0,1)$) in a graph for each of the nine experiments. These graphs reveal that the extent of the departure of the empirical distribution from the standard normal distribution varies among the experiments. However, there is one feature that is common among all the nine graphs plotted, the empirical distribution curve lies above the standard normal distribution close to both ends and below the standard normal distribution near the center. Among the nine graphs plotted, the graph shown as Figure 4.10 is the most obvious one the others

[15]The estimate is standardized by the corresponding sample mean and standard deviation, that is $\hat{z} = \frac{(z-\bar{z})}{s_z}$, where $\bar{z}$ and s_z are respectively the sample mean and standard deviation.

agree with the characterization to different extent.

As an attempt to gain a better understanding of the distribution of the recursive estimator, we construct a boxplot for each of the cases considered. The limits of the box are defined by:

$$\textit{maximum} = x_{(75)} + 1.5\ q \tag{4.1}$$

$$\textit{minimum} = x_{(25)} - 1.5\ q. \tag{4.2}$$

Where q is the interquantile range and $x_{(n)}$ is the n^{th} order statistic and all elements which lie outside the limits are indicated by a symbol in each of the cases considered. We plot all the boxplot of all the nine cases in a graph, Figure 4.12. As compared with the boxplot of one thousand observations sampled from a standard normal distribution, Figure 4.11. The boxplots of all cases considered, Figure 4.12, have more symbols at both ends of the box than the boxplot in Figure 4.11. As sample size increases, the symbols of the all the nine boxplots shift toward the center (median) and the number of symbols outside the whisker is also reduced. This means the estimates that are far from the center are adjusting toward the true value through the recursive updating. Some of them are adjusting fast enough to be in the box or whisker, however, even with the sample size of a thousand, many are not in the box or whisker as indicated by the symbol of the boxplot. The sampling distribution has thicker tails than the normal and causes the null hypothesis of normality to be rejected. It has been shown that the estimator has a asymptotically normal distribution. Hence, with a large enough sample size, the empirical cumulative distribution of the estimates obtained through the recursive updating could accord well with the normal distribution. However, this sample size certainly exceeds a thousand from what we have observed here.

4.5 Concluding Remarks

The recursive estimation in general is a quick way of updating the estimates in an ongoing manner. However, unlike its counterpart in the paradigm of parametric estimation, substantial number of recursive updating for the nonparametric estimation can seriously reduce its efficiency even though it is a consistent estimator. The recursive updating bring little improvement to initially bad estimates, that is estimates with large bias. Its ability in correcting a bad estimate is incredibly low. The actual sampling distribution of the recursive estimator also is greatly affected by the way it is estimated before going into the recursive updating. The asymptotic result indicates that regardless of the initial estimation the sampling distribution of the recursive estimator will eventually converge to a normal distribution. However, the actual rate at which it converges to a normal distribution can be so slow that the sampling distribution of the estimator even with a sample of one thousand observations still does not approximate well a normal distribution.

Chapter 5

Conclusion

The conclusion will summarize the main contributions in the thesis and briefly compare the results obtained with those presently existing in the literature. The aim of the dissertation has been to develop techniques for recursive estimation in a nonparametric setting, with particular reference to econometric applications. The thesis has considered the case of heteroscedastic and correlated observations. The recursive techniques commonly encountered in the literature of economics and statistics are recursive least squares, Kalman filtering and Bayesian estimation. As compared to these estimators, the techniques studied in this thesis have some clear advantages and disadvantages. The nonparametric estimator does not require the specification of a structural form, and therefore avoids the problem of misspecification. However, it generally faces the problem of low efficiency. This problem is even more serious with the recursive estimator which is normally sub-optimal to its fixed window width counterpart.

The thesis has extended the techniques of Ahmad and Lin (1976) for recursive estimation of the conditional mean for independent observations to the case in which the exogenous variable is α-mixing. In Chapter Two, we established the conditions which ensure that the estimator is asymptotically unbiased, consistent and has an asymptotically normal distribution. The approach of the proof is to develop simple

results that make possible the direct application of Laws of Large Numbers and Central Limit Theorems derived by McLeish (1975) and Davidson (1990). As a result, this approach has greatly simplified the proofs of theorems in the thesis. Apart from the results obtained in this chapter, Roussas (1990, 1992) has recently provided a more elaborate proof of uniform strong consistency for a similar version of the recursive nonparametric kernel estimator. However, this thesis was concerned mainly with the pointwise consistency of the recursive estimator while Roussas' result can be regarded as a different proof for establishing a stronger property of the estimator. The theorem on the asymptotic normality of the recursive estimator parallels Robinson's (1983) Theorem 5.3 on the asymptotic normality of the fixed window width estimator for the conditional mean in a time series model.

In the third chapter, the techniques were extended to the estimation of the first partial derivative. The chapter proposed a new recursive estimator of the first partial derivative and derived its asymptotic properties for the independent as well as α-mixing regressor. The proof followed the same strategies as in Chapter Two. Consequently they are greatly simplified and made intuitively straight forward. So far no similar result has been established in the literature.

The first half of the fourth chapter illustrated an implementation of the recursive estimator of the conditional mean. The rest of the chapter investigated the estimator's finite sample behaviour through a Monte Carlo study. This chapter considered the practical aspect of the recursive kernel estimation. Up to now, discussions pertaining to the implementation of the recursive estimator and its finite sample behaviour have been lacking in the literature.

The asymptotic results in the earlier chapters indicated that the optimal rate of convergence of the recursive estimator is the same with that of its fixed window width counterpart. A heuristic explanation for these results is that the window width

converges to zero as the sample size increases. For a sufficiently large length of recursive updating the difference in the widths becomes arbitrarily small. However, the performance of the recursive estimation is at best sub-optimal due to the use of non-optimal window widths for the estimation at the initial stage. As the result of the simulations showed, with short lengths of recursive updating (50 or less with initial estimation of size 50) the loss in relative efficiency is roughly between 10 to 20 percent. The loss increases to around 70 percent in some cases as the recursive updating proceeds further. However, the relative efficiency returns to about half for most of the cases considered when the length of recursive updating increased to nearly a thousand. Therefore it may be a good strategy to limit the length of recursive updating to about the same number of observations as in the initial set and then restart the process by reestimating using all the observations with an optimal fixed width estimator from time to time.

The results of the simulation also indicate that the finite sample distribution of the recursive estimator has more elements at the tails than a normal distribution and is not well approximated by a normal distribution. Perhaps the finite sample distribution of the recursive estimator can be better approximated by a distribution with thicker tails such as the t-distribution. It will require more investigation in this direction to justify this conjecture.

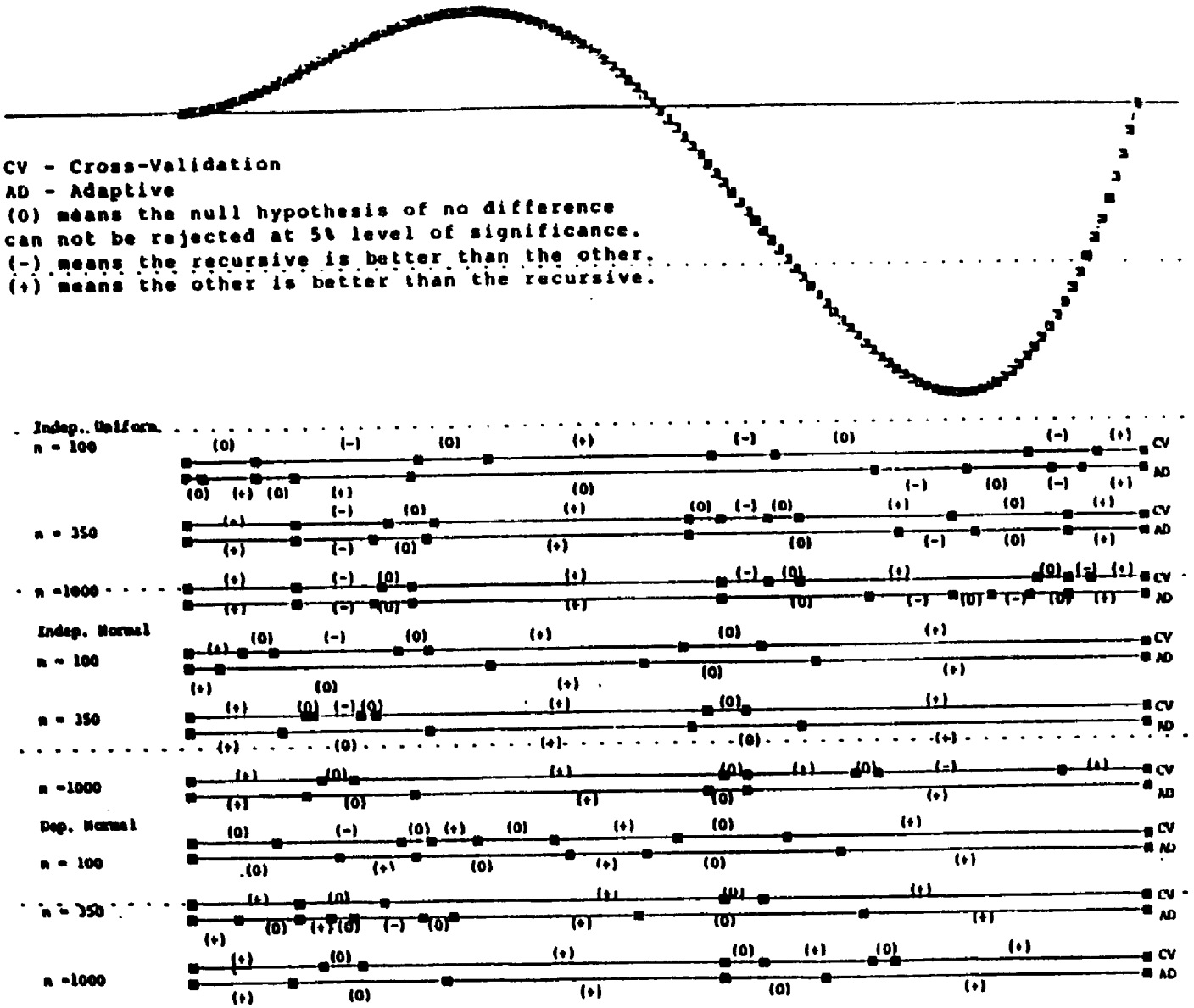

Figure 4.1: Comparison of the Recursive estimator with the other two (Cross-Validation and Adaptive) at each of the 125 estimation points.

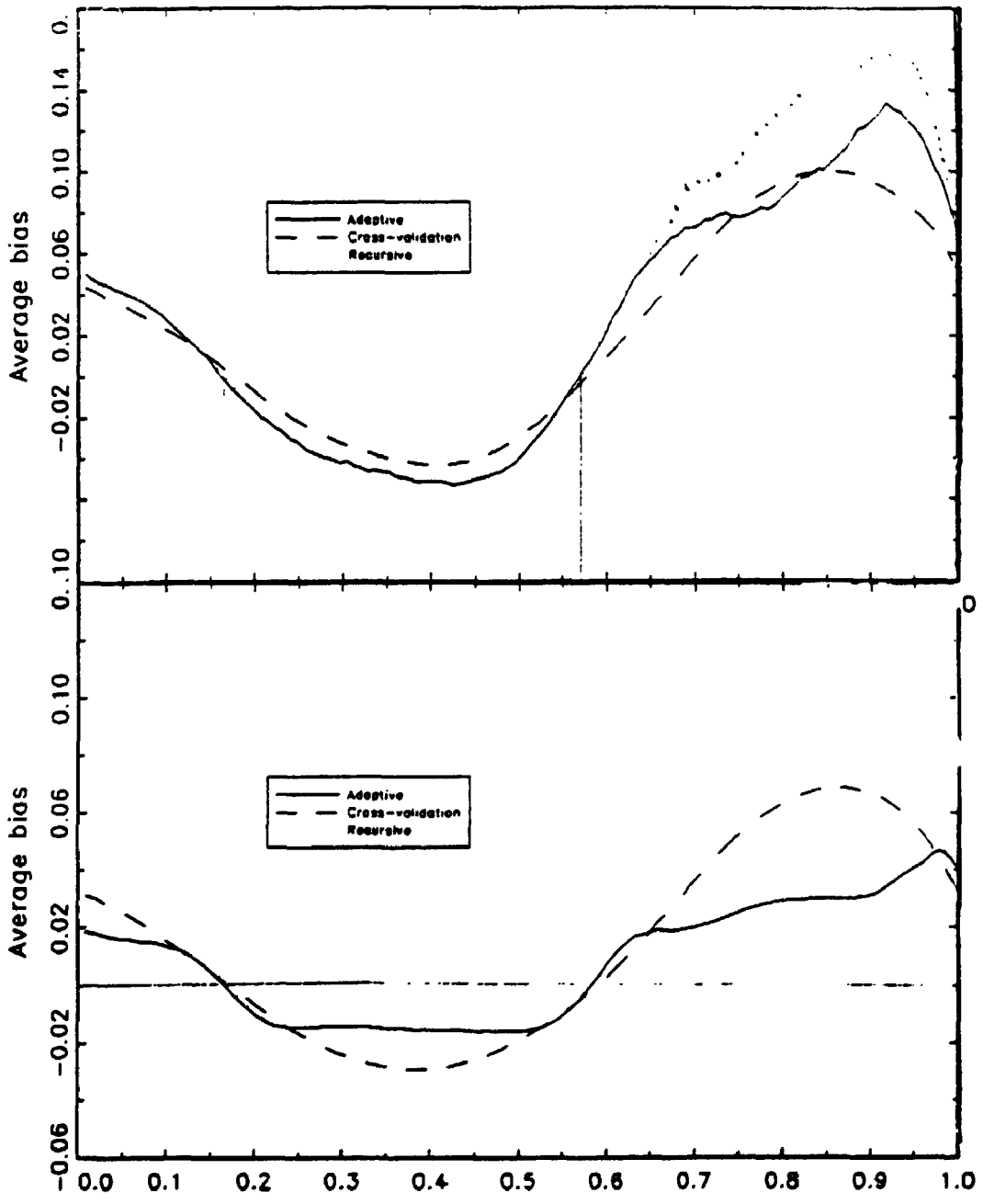

Figure 4.2: The Average Bias Curves of the three estimators for independent Normal Regressor with n=100 (top) and n =1000 (bottom).

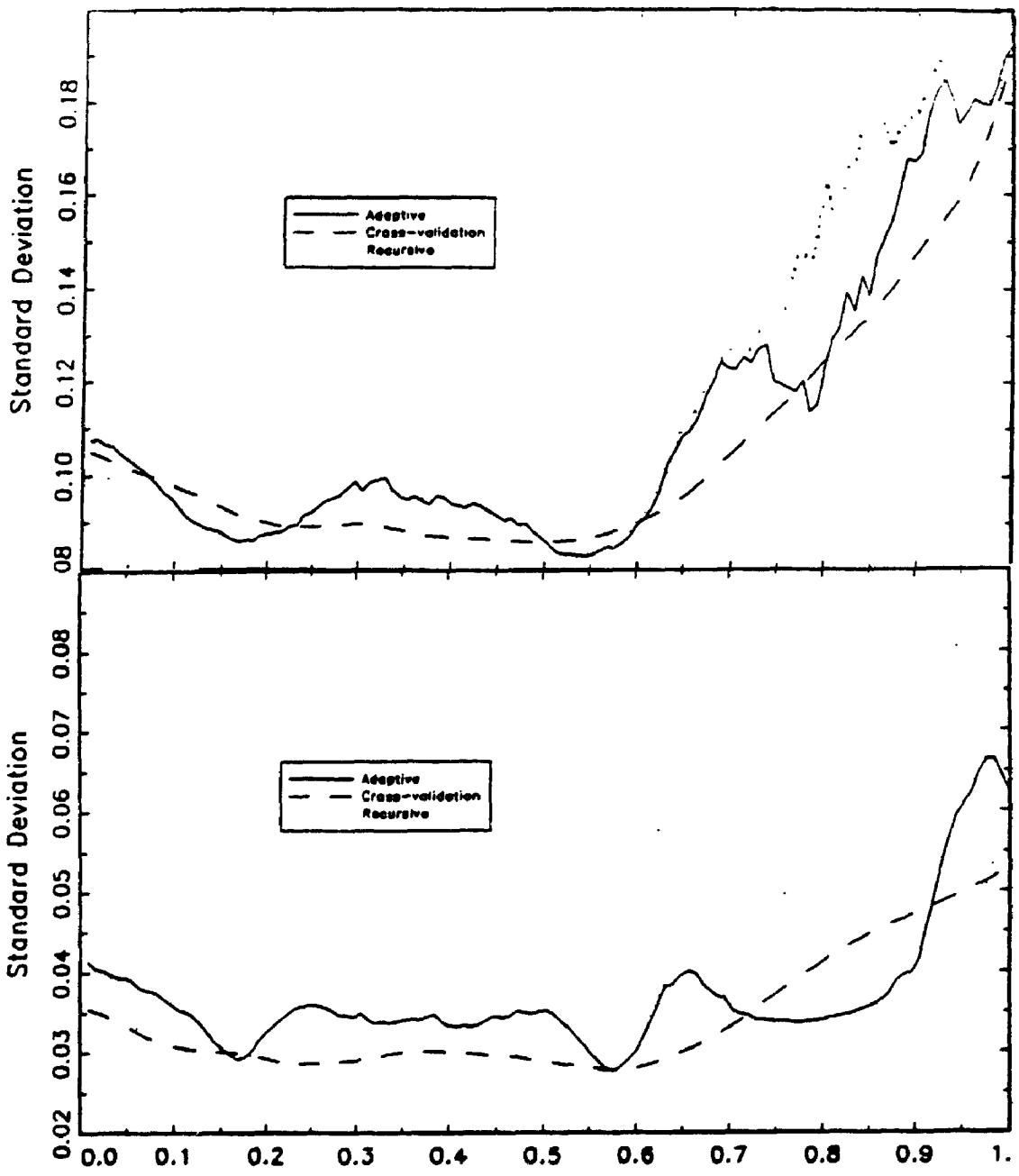

Figure 4.3: The Standard Deviation Curves of the three estimators for independent Normal Regressor with n=100 (top) and n =1000 (bottom).

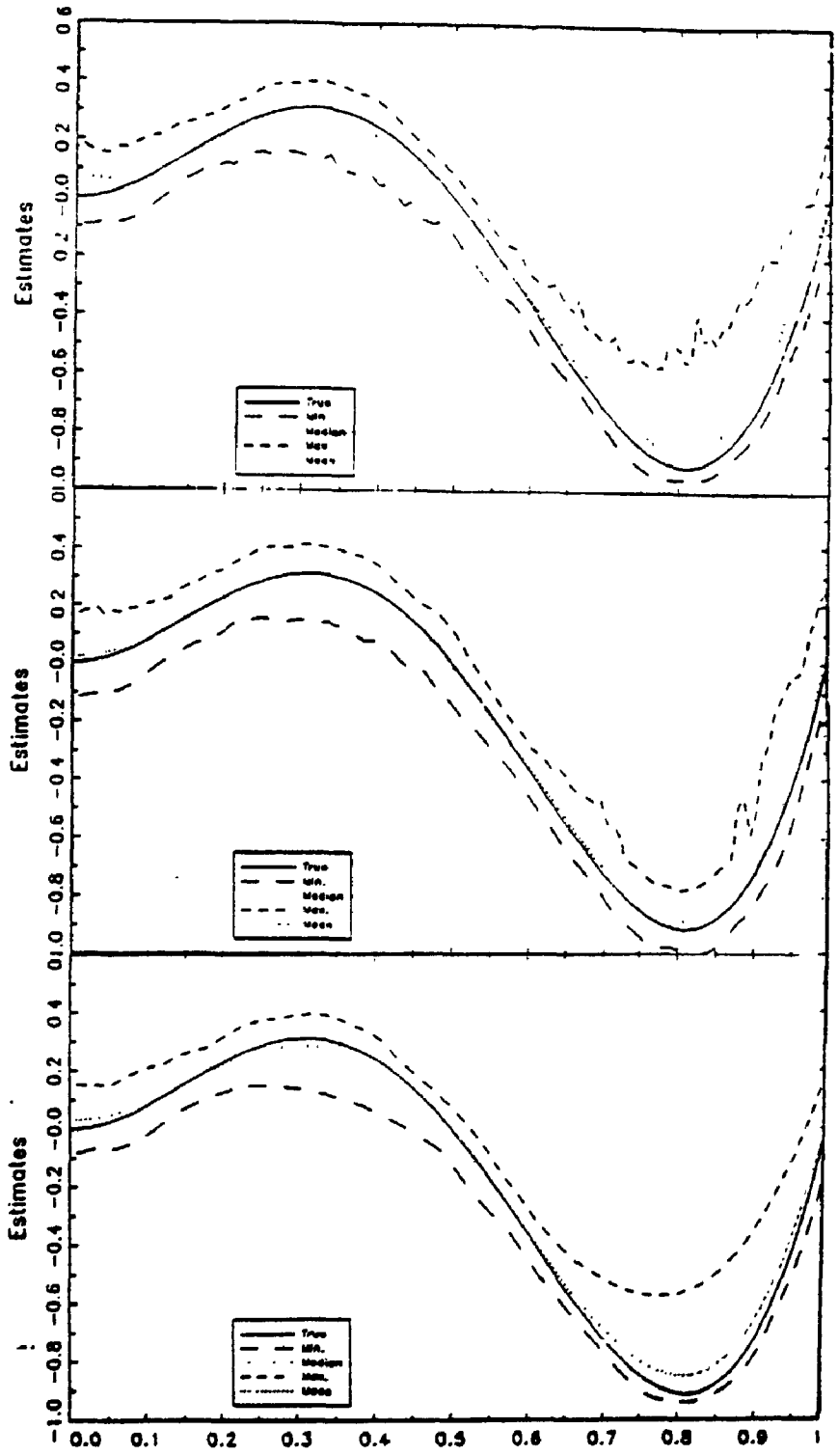

Figure 4.4: The Curves of the Maximum, Minimum, Average and Median of the Estimates from the three methods, Recursive (Top), Adaptive (Middle) and Cross-Validation (Bottom); for independent Normal Regressor with (a.) n=1000

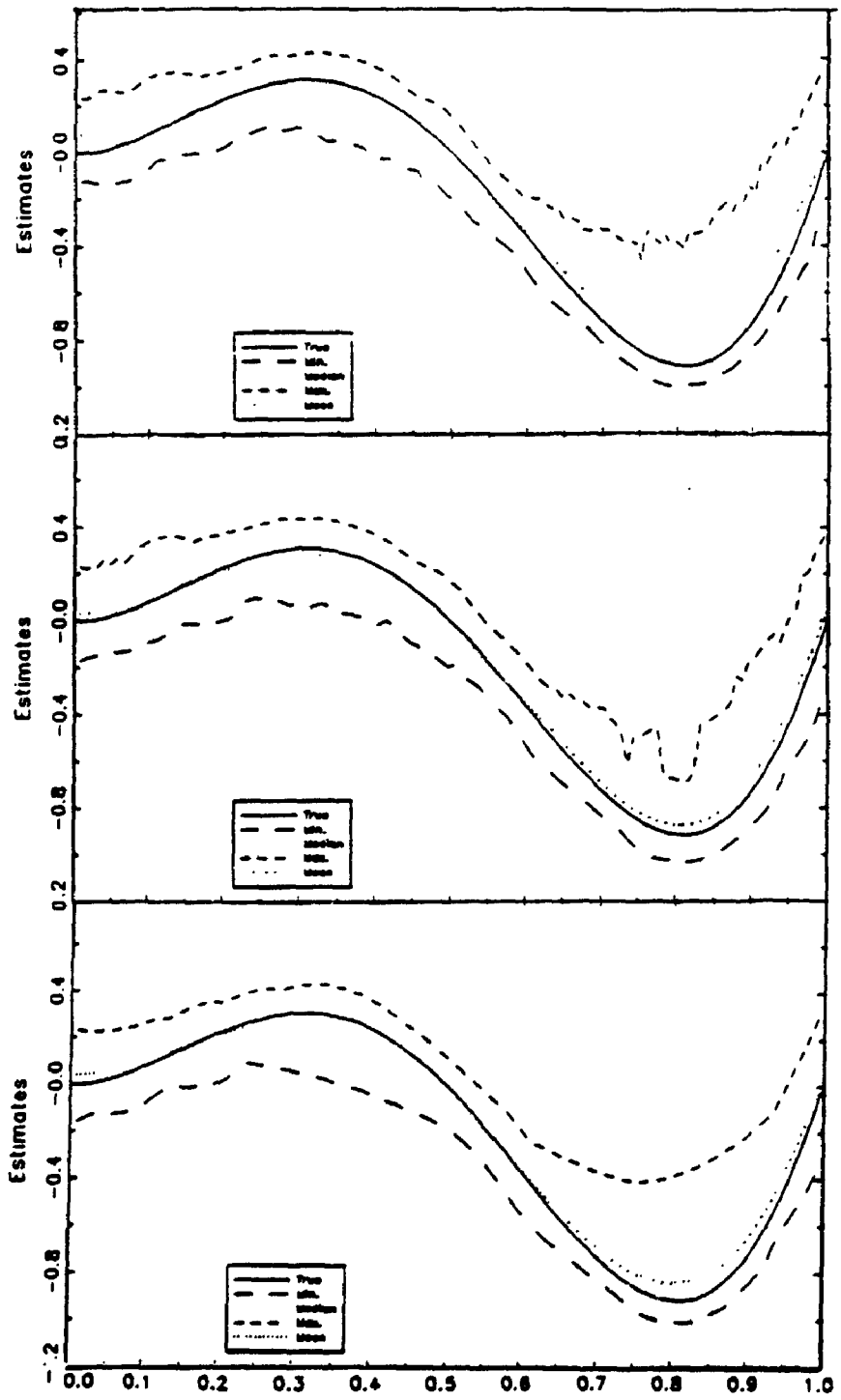

Figure 4.5: The Curves of the Maximum, Minimum, Average and Median of the Estimates from the three methods, Recursive (Top), Adaptive (Middle) and Cross-Validation (Bottom); for independent Normal Regressor with (b.) n=350

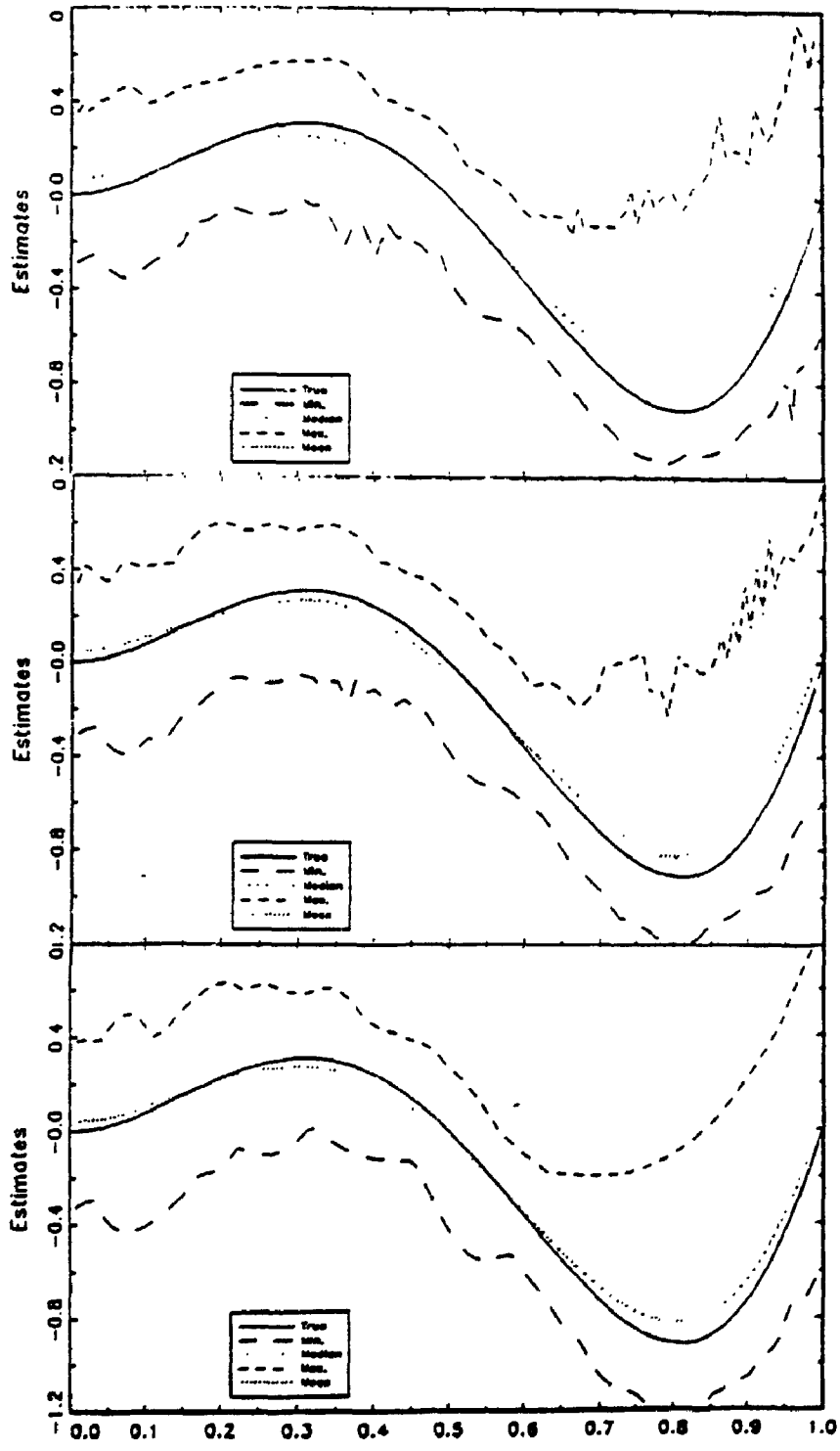

Figure 4.6: The Curves of the Maximum, Minimum, Average and Median of the Estimates from the three methods, Recursive (Top), Adaptive (Middle) and Cross-Validation (Bottom); for independent Normal Regressor with (c.) n =100.

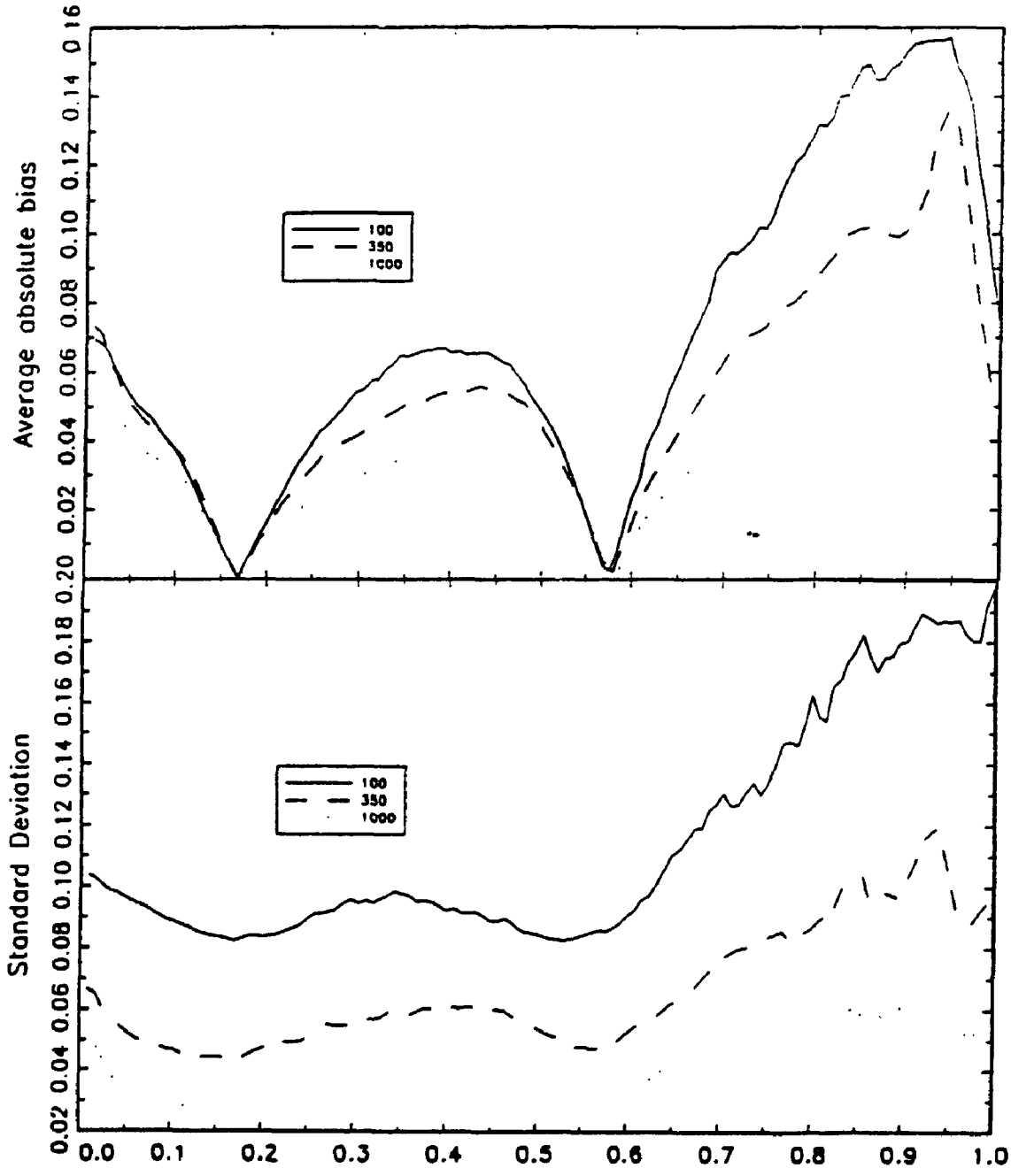

Figure 4.7: Case of Independent Normal, Top: The Average Bias curves of the Recursive estimation with three different sample sizes n=100, n=350 and n=1000. Bottom: The Standard Deviation curves.

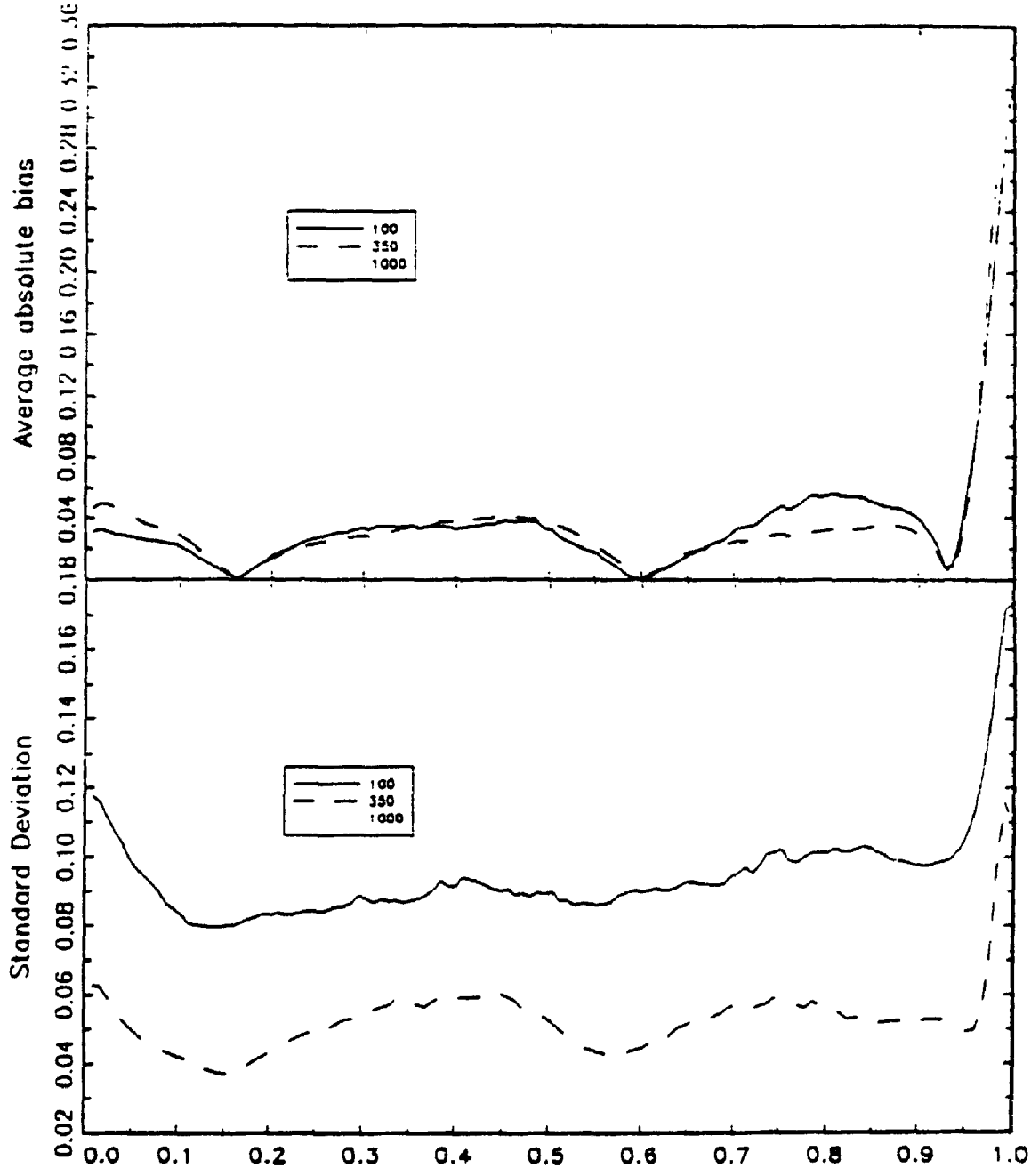

Figure 4.8: Case of Independent Uniform, Top: The Average Bias curves of the Recursive estimation with three different sample sizes n=100, n=350 and n=1000. Bottom: The Standard Deviation curves.

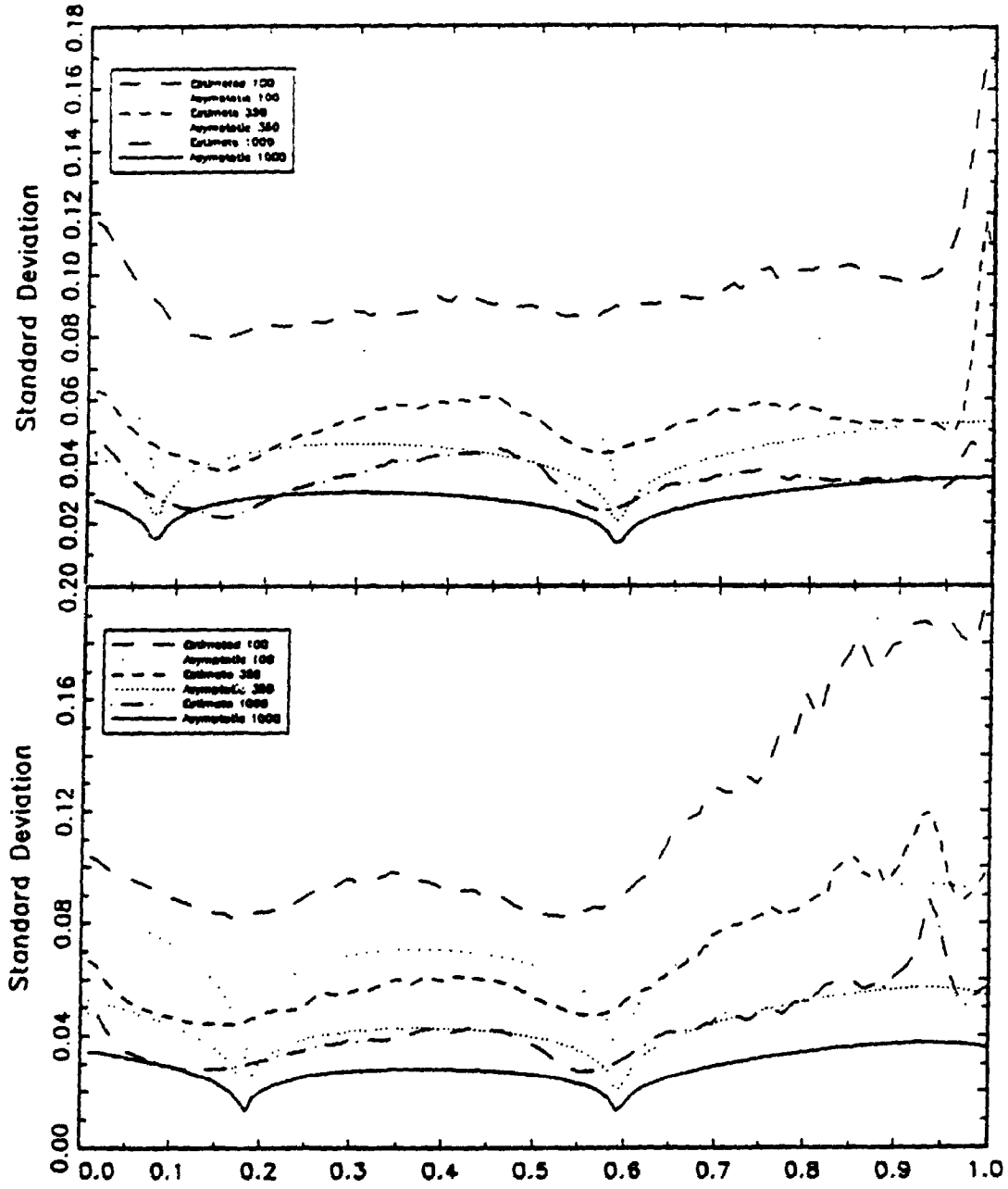

Figure 4.9: Comparison of the Standard Deviation curves of the finite sample estimator with its asymptotic approximates; Top: Independent Uniform regressor, Bottom: Independent Normal regressor.

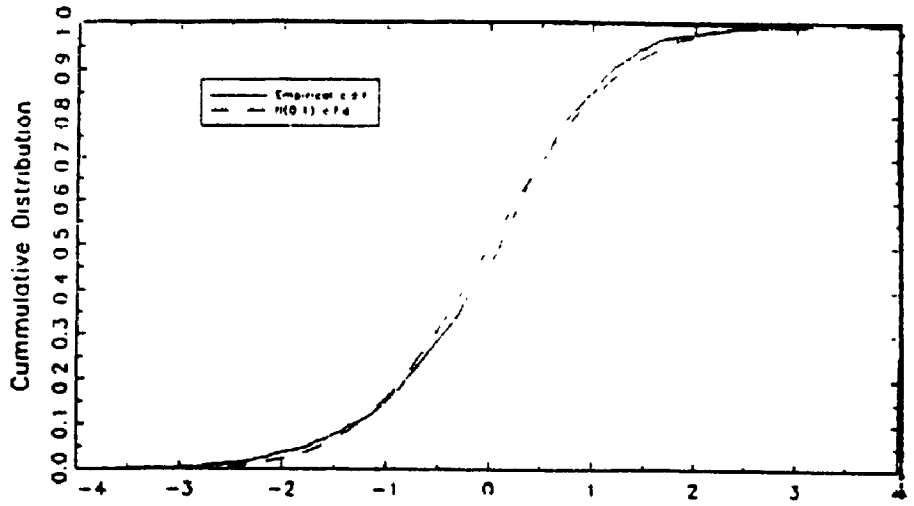

Figure 4.10: The Comparison of the empirical distribution of standardized finite sample estimate with distribution of N(0,1).

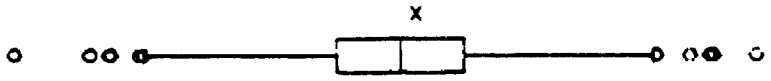

Figure 4.11: The boxplot for 1000 observations from N(0,1).

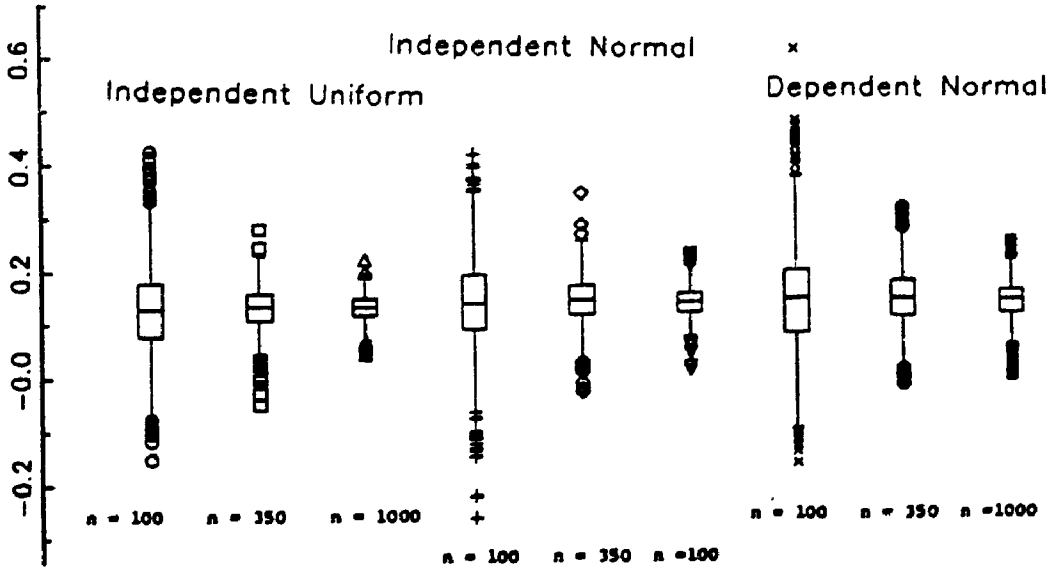

Figure 4.12: The boxplot for the estimates at x=0.12 under each of the nine experiments.

Bibliography

1. Ahmad, I.A. and P.E. Lin (1976) "Nonparametric Sequential Estimation of a Multiple Regression Function", *Bulletin of Mathematical Statistics*, **17**, p.63-75.

2. Ahmad, I.A. and A. Ullah (1988), "Nonparametric Estimation of the p-th Derivative of a Regression Function. Stochastic Regressor Case", *Department of Economics Research Report # 8903, University of Western Ontario.*

3. Andrews, D.W.K. (1984) "Non-strong Mixing Autoregressive Processes", *Journal of Applied Probability*, **21**, p.930-934.

4. Andrews, D.W.K. (1988) "Laws of Large Numbers for Dependent Non- Identically Distributed Random Variables", *Econometric Theory*, **4**, p.458-467.

5. Athreya, K.B. and S.G. Pantula, (1986), "A Note on Strong Mixing of ARMA Processes", *Statistics and Probability Letter*, **4**, p.187-190.

6. Bhattacharya, P.K. (1967), "Estimation of a Probability Density and its Derivatives", *Sankhya Series A* , **29**, p.373-382.

7. Billingsley, P. (1968) *Convergence of Probability Measures* , **John Wiley**, New York, New York, U.S.A.

8. Bochner, S. (1955) *Harmonic Analysis and the Theory of Probability* , **University of California Press**, Berkeley.

9. Breslaw, J.A. (1992) "Kernel Estimation with Cross-Validation Using the Fast Fourier Transform", *Economics Letter*, **38**, p.285-289.

10. Cacoullos, T. (1966) "Estimation of a Multivariate Density", *Annals of the Institute Statistical Mathematics*, **18**, p.179-189.

11. Carroll, R.J. (1976) "On Sequential Density Estimation", *Z.für Wahrscheinlicheitstheorie und Verwandte Gebiet*, **36**, p.136-151.

12. Castellana J.V. and M. R. Leadbetter (1986) "On Smoothed Probability Density Estimation for Stationary Processes", *Stochastic Processes and Their Applications*, **21**, p.179-193.

13. Chanda, K.C. (1974) "Strong Mixing Properties of Linear Stochastic Processes" *Journal of Applied Probability*, **11**, p.401-408.

14. Chiu, S.T. (1990) "On the Asymptotic Distribution of Bandwidth Estimates", *Annals of Statistics*, **18**, p.1696-1711.

15. Christensen, L.R., D.W. Jorgensen and L.J. Lau (1973) "Transcendantal Logarithmic Frontier", *Review of Economics and Statistics*, **55**, p.28-45.

16. Davidson, J. (1990) "Central Limit Theorems for Nonstationary Mixing Processes and Near-Epoch Dependent Functions", *London School Economics EM/90/216*.

17. Davies, H.I. and E.J. Wegman (1975) "Sequential Nonparametric Density Estimation", *IEEE Transactions on Information Theory*, **IT-21**, p.619-628.

18. Davies, H.I. and E.J. Wegman (1979) "Remarks on Some Recursive Estimators of a Probability Density", *Annals of Statistics*, **7**, p.316-327.

19. Deheuvels, P. (1974) "Conditions Nëcessaires et Suffisantes de Convergence Ponctuelle Presque Sûre et Uniforme Presque Sûre des Estimateures de la Densité", *C.R. de l'Acad. des sci. de Paris* , **278**, p.1217-1220.

20. Devroye, L.P. (1979) "On the Pointwise and the Integral Convegence of Recursive Kernel Estimates of Probability Densities", *Utillitas Mathematica*, **15**, p.113-128.

21. Diewert, W.E. (1971) "An Application of the Sheperd Duality Theorem:A Generalized Leontif Production Function", *Journal of Political Economy*, **49**, p.481-507.

22. Fan, Y. (1990) "Seemingly Unrelated Essays in Econometrics - Functions of Mixing Processes, Nonparametric Estimation and Cointegration", *Ph.D. thesis, University of Western Ontario.*

23. Gallant, A.R. (1981), "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form:the Fourier Flexible Form", *Journal of Econometrics*, **15**, p.211-244.

24. Gasser, T., and W. Köhler, (1990), "Selection of Smoothing Parameters in Nonparametric Regression:A Simulation Report", *Unpublished manuscript, Department of Biostatistics, Zentralinstitut für Seelische Gesundheit, 6800 Mannheim, F.R.G.*

25. Gasser, T., A. Kneip, and W.Köhler (1991), "A Flexible and Fast Method for Automatic Smoothing", *Journal of the American Statistical Association*, p.643-652.

26. Gasser, T. and H.G. Müller (1979) "Optimal Convergence Properties of Kernel

Estimates of Derivative of Density Function", T.Gasser and Rosenblatt (eds), *Smoothing Techniques for Curve Estimation*, New York:Springer-Verlag 757, p.23-68.

27. Georgoev. A.A. (1988) "Consistent Nonparametric Multiple Regression:the Fixed Design Case", *Journal of Multivariate Analysis*, **25**, p.100-110.

28. Gorodetskii, V.V. (1977) "On the Strong Mixing Property for Linear Sequences", *Journal of Applied Probability*, **11**, p.411-413.

29. Greblicki, W., A. Krzyzak and M. Pawlak (1984) "Distribution-Free Pointwise Consistency of Kernel Regression Estimate", *Annals of Statistics*, **12**, p.1570-1575.

30. Greblicki, W. and M. Pawlak (1987) "Necessary and Sufficient Consistency Conditions for a Recursive Kernel Regression Estimate", *Journal of Multivariate Analysis*, **23**, p.67-76.

31. Györfi, L. (1981) "Strong Consistent Density Estimate from Ergodic Sample", *Journal of Multivariate Analysis*, **11**, p.81-84.

32. Härdle, W. (1990) *Applied Nonparametric Regression* , **Cambridge University Press**, New York.

33. Härdle, W. and J.S. Marron (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation", *Annals of Statistics*, **13**, p.1465-81.

34. Härdle, W. and P. Vieu (1990),"Nonparametric Regression Smoothing of Time Series", *CORE discussion paper* , **9031**.

35. Hart, J.D. and P. Vieu (1990) "Data-Driven Bandwidth Choice for Density Estimation Based on Dependent Data", *Annals of Statistics* **18**, p.873-890.

36. Harvey, A.C. (1989) *Forecasting, Structural Time Series Models, and the Kalman Filter*, **Cambridge University Press**, New York.

37. Ibragimov, I.A. (1962) "Some Limit Theorems for Stationary Processes", *Theory of Probability and Its Applications*, **7**, p.349-382.

38. Ibragimov, I.A. and V. Y. Linnk (1971) *Independent and Stationary Sequences of Random Variables*, **Wolters-Noordhoff**, Groningen.

39. Ioannides, D. and G.G. Roussas (1985) "Moment Inequalities for Mixing Sequences of Random Variables", *Stochastic Analysis and Applications*, **5**, p.61-120.

40. Kolmogorov, A.N. and Y.A. Rozanov (1960) "On Strong Mixing Conditions for Stationary Gaussian Processes", *Theory of Probability and Its Applications*, **5**, p.204-208.

41. Lau, L.J. (1988) "Functional Forms in Econometric Model Building" *In Handbook of Econometrics* **Vol. 3**, (Ed. Z.Griliches and M.D. Intriligator) Amsterdam: North Holland.

42. Lilliefors, H.W. (1967) "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association* **62**, p.399-402.

43. MacFadden, D. L. (1985) "Specification of Econometric Models", *Presidental Address to the Fifth World Congress of the Econometric Society*, Cambridge, MA, Econometric Department, MIT.

44. Mack, Y.P. and H.G. Müller (1987) "Adaptive Nonparametric Estimation of a

Multivariate Regression Function", *Journal of Multivariate Analysis* **17**, p.163-181.

45. Marron (1988) "Automatic Smoothing Parameter Selection: A Survey", *Empirical Economics*, Vol.13, p.187-208.

46. Marron, J.S. and B.U. Park (1990) "Comparison of Data-driven Bandwidth Selectors", *Journal of the American Statistical Association*, **84**, p.66-72.

47. Marsy, E. (1986) "Recursive Probability Density Estimation for Weakly Dependent Stationary Processes", *IEEE Transactions on Information Theory* **IT-32**, No.2 p.254-267.

48. Marsy, E. (1987) "Almost Sure Convergence of Recursive Density Estimators for Stationary Mixing Processes", *Statistics and Probability Letter*, **5**, p.249-254.

49. Marsy, E. and L. Györfi (1987) "Strong Consistency and Rates for Recursive Probability Density Estimators of Stationary Processes", *Journal of Multivariate Analysis*, **22**, p.79-93.

50. Menon, V.V., V. Prasad and R.S. Singh (1984) "Nonparametric Recursive Estimates of A Probability Density Function and its Derivatives", *Journal of Statistical Planning and Inference* **9**, p.73-82.

51. McLeish, D.L. (1974) "Dependent Central Limit Theorems and Invariance principles", *Annals of Probability*, **3**, p.829-839.

52. McLeish, D.L. (1975) "A Maximal Inequality and Dependent Strong Laws", *Annals of Probability*, **3**, p.829-839.

53. McMillan, J., A. Ullah and H.D. Vinod (1988) "Estimation of the Shape of the Demand Curve by Nonparametric Kernel Methods", **B. Raj** (ed.), *Advances in Econometrics and Modelling*, Holland: Klewer Academic Press.

54. Müller, H.G. (1988) *Nonparametric Regression Analysis of Longitudimal Data* Berlin:Springer-Verlag.

55. Müller, H.G. and U. Stadtmüller (1987) "Variable Bandwidth Kernel Estimator of Regression Analysis", *Annals of Statistics*, **15**, p.610-625.

56. Müller, H.G., U. Stadtmüller and T. Schmitt (1987) "Bandwidth Choice and Confidence Intervals for Derivatives of Noisy Data", *Biometrika*, **74**, p.743-750.

57. Nadaraya. E.A. (1964) "On Estimating Regression", *Theory of Probability and Its Applications* **9**, p.141-142.

58. Pagan, A. R. and M. R. Wickens (1989) "A Survey of Some Recent Econometric Methods", *The Economic Journal*, **99**, p.962-1025.

59. Parzen, E. (1962) "On Estimation of a Probability Density Function and Mode", *Annals of Mathematical Statistics*, **33**, p.1065-1076.

60. Rice., J.A. (1986) "Bandwidth Choice for Differentiation", *Journal of Multivariate Analysis* **19**, p.251-264.

61. Rilstone, P. (1987) "Nonparametric Partial Derivative Estimation" *Ph.D. Thesis University of Western Ontario, Canada.*

62. Rilstone, P. and A. Ullah (1989) "Nonparametric Estimation of Response Coefficients", *Communications in Statistics Part A - Theory and Methods* **18(7)**, p.2615-2627.

63. Robinson, P.M. (1983) "Nonparametric Estimators for Time Series", *Journal of Time Series Analysis* **4**, p.185-207.

64. Rosenblatt, M. (1956) "Remarks on Some Nonparametric Estimates of Density Function", *Annals of Mathematical Statistics* **29**, p.832-837.

65. Rosenblatt, M. (1956) "A Central Limit Theorem and a Strong Mixing Condition", *Proc. Nat. Acad. Sci. U.S.A.* , **42**, p.43-47.

66. Roussas, G.G. (1990) "Nonparametric Regression Estimation Under Mixing Conditions", *Stochastic Processes and Their Applications* **36**, p.107-116.

67. Roussas, G.G. (1992) "Exact Rates of Almost Sure Convergence of a Recursive Kernel Estimate of a Probability Density Function:Application to Regression and Hazard Rate Estimation", *Journal of the Nonparametric Statistics* **1**, p.171-195.

68. Rudemo, M. (1982) "Empirical Choice of Histograms and Kernel Density Estimators", *Scandinavian Journal of Statistics* **9**, p.65-78.

69. Rutkowski, L. (1985) "Real-time Identification of Time-varying System by Nonparametric Algorthms based on Parzen Kernels", *International Journal Systems Science* , **16**, p.1123-1130.

70. Schuster, E.F. (1969) "Estimation of a Probability Density Function and its Derivatives", *Annals of Mathematical Statistics* **40**, p.1187-1195.

71. Schuster, E.F. and S. Yakowitz (1979) "Contributions to the Theory of Nonparametric Regression with Application to System Identification", *Annals of Statistics* **5**, p.394-399.

72. Scott, D. (1985) "Averaged Shifted Histograms:Effective Nonparametric Densition Estimators in Several Dimensions" *Annals of Statistics* **13**, p.1024-1040.

73. Scott, D.W. and G. R. Terrell (1987) "Biased and Unbiased Cross-Validation" *Journal of the American Statistical Association*, **82**, p.1131-1146.

74. Serfling, R.J. (1968) "Contributions to Central Limit Theory for Dependent Variables", *Annals of Mathematical Statistics*, **39**, p.1158-1175.

75. Singh, R.S. (1974) "Estimation of Derivatives of Average of μ-densities and Sequence Compound Estimation Exponential Families," *RM-318, Dept. of Statistics and Probability, Michigan State University, East Lansing, MI.*

76. Singh, R. S. (1977) "Applications of Estimators of a Density and its Derivatives to Certain Statistical Problems," *Journal of the Royal Statistical Society, Ser.B* **39**, p.394-399.

77. Singh, R.S. (1979) "Mean Squared Errors of Estimates of a Density and its Derivatives" *Biometrika* **66**, p.177-180.

78. Singh, R.S. (1981) "Speed of Convergence in Nonparametric Estimation of a Multivariate μ-Density and its Mixed Partial Derivatives," *Journal of Statistical Planning and Inference* **5**, p.287-298.

79. Singh, R.S. and A. Ullah (1985) "Nonparametric Time-Series Estimation of Joint DGP, Conditional DGP, and Vector Autoregression", *Econometric Theory*, **1**, p.27-52.

80. Singh, R.S. and A. Ullah (1986) "Nonparametric Recursive Estimation of a Multivariate Marginal and Conditional DGP with an Application to Specifica-

tion of Econometric Models", *Communications in Statistics Part A - Theory and Methods*, **15**(12), p.3489-3513.

81. Singh, R.S., A. Ullah and R.A.L. Carter (1987) "Nonparametric Inference in Econometric New Applications" *Time-Series and Econometric Modelling* (Ed. I.B.MacNeill and G.J. Umphrey) Kluwer Academic Publishers, Boston, MA, U.S.A., p.253-278.

82. Slutsky, E.E. (1925) "Über Stochastische Asymptoten und Grenzwerte", *Metron*, **5**, p.1-90.

83. Stone, C.J. (1982) "Optimal Global Rates of Convergence for Nonparametric Regression" *Annals of Statistics* **10**, p.1040-53.

84. Watson, G.S. (1964) "Smooth Regression Analysis" *Sankhya Series A* **26**, p.359-72.

85. West, M and J. Harrison (1990) *Bayesian Forecasting and Dynamic Models*, New York, Berlin:Springer Series in Statistics.

86. White, H. (1980) "Using Least Squares to Approximate Unknown Regression Functions", *International Economic Review* **21**, p.149-170.

87. Wolverton and T. J. Wagner (1969) "Asymptotically Optimal Discriminant Functions for Pattern Classification", *IEEE Transactions on Information Theory*, **IT-15**, p.258-265.

88. Withers, C.S. (1981) "Conditions for Linear Processes to be Strong- Mixing", *Z.für Wahrscheinlicheitstheorie und Verwandte Gebiet*, **57**, p.477-480.

89. Ullah, A. and H.D. Vinod (1987) "Flexible Production Function Estimation by Nonparametric Kernel Estimators", *Department of Economics, Technical Report #16 University of Western Ontario.*

90. Ullah, A. and H.D. Vinod (1988) "Nonparametric Kernel Estimation of Econometric Parameters" *Journal of Quantitative Economics* 4, p.81-87.

91. Yamato, H. (1971) "Sequential Estimation of A Continuous Probability Density Function and Mode", *Bulletin of Mathematical Statistics*, 14, p.1-12.