THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY
OF NORTH CAROLINA

LOCAL POLYNOMIAL ESTIMATION OF REGRESSION FUNCTIONS FOR MIXING PROCESSES

by

E. Masry and J. Fan

October 1993

Mimeo Series #2311

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

E.Masry & J. Fan

MIMEO SERIES #2311

LOCAL POLYNOMIAL ESTIMATION OF REGRESSION FUNCTIONS FOR MIXING PROCESSES

NAME	DATE

Local Polynomial Estimation of Regression Functions for Mixing Processes

Elias Masry†
Department of Electrical and Computer Engineering
University of California
La Jolla, CA 92093-0407

Jianqing Fan‡
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

Abstract

Local polynomial fitting has many exciting statistical applications. Yet, the results are rarely known for dependent data. However, the desire for nonlinear time series modeling, constructing predictive intervals, understanding divergence of nonlinear time series requires the theory and applications of local polynomial fitting for dependent data. In this paper, we study the problem of estimating conditional mean functions and their derivatives via a local polynomial fit. The functions include conditional moments, conditional distribution as well as conditional density functions. Joint asymptotic normality for derivatives estimation is established for both strongly mixing and ρ-mixing processes.

1 Introduction

Local polynomial fitting, systematically studied by Stone (1977), Cleveland (1979), Tsybakov (1986) and Fan (1992, 93), has many exciting statistical applications, in particular to statistical function estimation. It reduces the bias of the Nadaraya-Watson estimators

† Supported by the Office of Naval Research under Grant N00014-90-J-1175

‡ Partially supported by an UNC Research Council grant, NSF grant DMS-9203135 and an NSF postdoctoral fellowship.

Abbreviated title. Nonparametric regression.

AMS 1991 Subject Classification. Primary 62G07. Secondary 62H10, 60F05.

Key words and phrases. Asymptotic normality, local polynomial fitting, mixing processes.

and the variance of the Gasser-Müller (1979) estimator (see Chu and Marron (1991) and Fan (1992)). It adapts automatically to the boundary of design points (see Fan and Gijbels (1992), Hastie and Loader (1993), Ruppert and Wand (1993)) — No boundary modification is required. The design adaptation and the advantages of local polynomial fitting was made clear in Fan (1992). Since then, many interesting statistical properties have been discovered. See Ruppert and Wand (1993) and Fan, Gasser, Gijbels, Brockman and Engle (1993). In particular, Ruppert and Wand (1993), Fan and Gijbels (1993) emphasize the superiority of using local polynomial as a device for derivatives estimation. Since local polynomial fitting is basically a regression problem, the choice of the local neighborhood is very easy. Fan and Gijbels (1993) propose two procedures for bandwidth selection. See also Ruppert, Sheater and Wand (1993) on the development of this subject. While the bandwidth selection for the local polynomial fitting is simple and useful, it does not apply to the Nadaraya-Watson estimator due to its inefficiency (See Fan and Gijbels (1993)). Interesting applications to robust regression can be found in Tsybakov (1986) who discovered many exciting minimaxity and asymptotic normality results. The local polynomial fitting as useful statistical graphical tools can be found in Cleveland (1993).

While the theory and applications of local polynomial regression are very well studied, the focus of all of the above mentioned papers is on i.i.d. observations. Yet, the statistical properties of local polynomial regression for dependent data have not been studied previously. The desire for nonlinear time series modeling, estimating time trend, constructing predictive intervals, understanding divergence of nonlinear time series compels us to consider the dependent data case. See Tong (1990) and Yao and Tong (1993 a,b) for these problems and the use of the local polynomial fitting as a device to these important issues.

We now abstract our problem as follows: Suppose that we observe a stationary sequence $(X_1, Y_1), \cdots, (X_n, Y_n)$. Of interest is the estimation of the conditional mean function

$$m(x) = E(Y|X = x) \tag{1.1}$$

and its derivative function $m^{(\nu)}(x)$ by using local polynomial regression. If the $(p+1)^{th}$

derivative of $m(z)$ at the point x exists, we approximate $m(z)$ locally by a polynomial of order p:

$$m(z) \approx m(x) + \cdots + m^{(p)}(x)(z-x)^p/p! \equiv \beta_0 + \cdots + \beta_p(z-x)^p. \tag{1.2}$$

Thinking locally about the unknown function modeled by (1.1), one then carries a local polynomial regression by minimizing

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p}\beta_j(X_i-x)^j\right)^2 K\left(\frac{X_i-x}{h}\right), \tag{1.3}$$

where $K(\cdot)$ denotes a nonnegative weight function and h — a smoothing parameter — determines the size of the neighborhood of x. If $\{\hat{\beta}_\nu\}$ denotes the solution to the above weighted least squares problem, then it is clear from (1.1) that $\nu!\hat{\beta}_\nu(x)$ estimates $m^{(\nu)}(x), \nu = 0, \cdots, p$. We remark that since we fit (1.3) locally, we do not need to know whether $\mathrm{var}(Y|X=x)$ remains constant or not, because it is approximately the same in a local neighborhood. This is another advantage of the local polynomial fitting.

The above setup is broad enough to include estimating functions of form $m_\psi(x) = E(\psi(Y)|X=x)$ by using the new data set $\{(X_1, Z_1), \cdots, (X_n, Z_n)\}$ with $Z_i = \psi(Y_i)$. These functions include the conditional moment functions, conditional distribution functions as well as conditional density functions, and their derivatives with respect to x. For simplicity of notation, we do not explicitly treat these functions. The above setup can also be applied to time series; in that case, by letting $Y_i = X_{i+d}$, we have the case of d-step prediction and we are able to estimate d-step conditional moments and distribution functions.

Our goal is to establish the joint asymptotic normality of the vector $\hat{\boldsymbol{\beta}}(x) = (\hat{\beta}_1(x), \cdots, \hat{\beta}_p(x))^T$. The major technical difficulty of the local polynomial fitting for dependent data is that the conditional arguments of Fan (1992) and Ruppert and Wand (1993) are no longer applicable. Upon conditioning on the design vector $(X_1, \cdots, X_n)$, the values of $(Y_1, \cdots, Y_n)$ could be fixed, in particular in the time series context where $Y_i = X_{i+d}$ — d-step prediction. Thus, the bias and variance of the estimator $\hat{\beta}_\nu(x)$ can not easily be obtained as in Fan and Gijbels (1993).

In the case of a local constant fit, i.e. $p = 0$, minimizing (1.3) reduces to the Nadaraya-Watson estimator. This estimator has been extensively studied in the literature by, for example, Mack and Silverman (1982) and Härdle (1990) and references therein for i.i.d. observations, and Rosenblatt (1969), Robinson (1983, 1986), Collomb and Härdle (1986), Roussas (1990), Truong (1991) and Roussas and Tran (1991), among others for dependent observations. Our technical devices are analogous to those, but are also quite different. The 'classical' arguments rely strongly on the simple form of the Nadaraya-Watson estimator. However, for general polynomial fitting (minimizing (1.3) for arbitary p) the derivation of the asymptotic distributions of the resulting estimators is considerably more involved.

The outline of the paper is as follows. Section 2 deals with the mean-square convergence of the hat matrix of the regression problem (1.3) and of other related quantities. These studies serve as a building block to our main result. The joint asymptotic normality and its implications are presented in Section 3.

2 Mean Square Convergence

As indicated in the Introduction many statistical problems, such as the estimation of conditional moment functions and conditional distributions, involve the same design matrix. Thus, it is worthwhile to study its convergence properties. The study here also serves as a building block to our main results and has applications elsewhere.

2.1 Preliminaries

We first introduce some notation. Denote $K_h(t) = K(t/h)/h$. Let

$$\boldsymbol{X} = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^p \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{W} = \mathrm{diag}\Big(K_h(X_i - x)\Big).$$

Then, the solution to the problem (1.3) is $\hat{\beta}(x) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}$. The matrix $(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})$ is positive definite as long as there are at least $p + 1$ local effective design points. This assumption is granted since we always assume that $nh \to \infty$. Denote by

$$s_{n,j} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - x}{h}\right)^j K_h(X_i - x), \quad t_{n,j} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - x}{h}\right)^j K_h(X_i - x)Y_i. \tag{2.1}$$

Putting

$$\boldsymbol{S}_n = \begin{pmatrix} s_{n,0} & \cdots & s_{n,p} \\ \vdots & \ddots & \vdots \\ s_{n,p} & \cdots & s_{n,2p} \end{pmatrix}, \quad \boldsymbol{t}_n = \begin{pmatrix} t_{n,0} \\ \vdots \\ t_{n,p} \end{pmatrix}, \tag{2.2}$$

the solution to (1.3) can be expressed as (see Fan and Gijbels (1993))

$$\hat{\boldsymbol{\beta}}(x) = \text{diag}(1, h^{-1}, \cdots, h^{-p})\boldsymbol{S}_n^{-1}\boldsymbol{t}_n. \tag{2.3}$$

For the convenience of notation, we denote

$$\mu_j = \int_{-\infty}^{+\infty} u^j K(u)du, \quad \nu_j = \int_{-\infty}^{+\infty} u^j K^2(u)du.$$

and

$$\boldsymbol{S} = \begin{pmatrix} \mu_0 & \cdots & \mu_p \\ \vdots & \ddots & \vdots \\ \mu_p & \cdots & \mu_{2p} \end{pmatrix}, \quad \tilde{\boldsymbol{S}} = \begin{pmatrix} \nu_0 & \cdots & \nu_p \\ \vdots & \ddots & \vdots \\ \nu_p & \cdots & \nu_{2p} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_p \\ \vdots \\ \mu_{2p+1} \end{pmatrix}. \tag{2.4}$$

In the i.i.d. case, it can be easily seen via conditioning on $(X_1, \cdots, X_n)$ that $\hat{\boldsymbol{\beta}}(x)$ is estimating the vector

$$\hat{\boldsymbol{\beta}}^*(x) \equiv (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(m(X_1), \cdots, m(X_n))^T. \tag{2.5}$$

Since the regression is conducted in the neighborhood of $|X_i - x| \le h$, by Taylor's expansion,

$$\begin{aligned} \boldsymbol{m} &\equiv (m(X_1), \cdots, m(X_n))^T \\ &= \boldsymbol{X}\boldsymbol{\beta}(x) + \frac{m^{(p+1)}(x)}{(p+1)!}((X_1 - x)^{p+1}, \cdots, (X_n - x)^{p+1})^T + o_P(h^{p+1}), \end{aligned} \tag{2.6}$$

where $\boldsymbol{\beta}(x) = (m(x), \cdots, m^{(p)}(x)/p!)^T$. Substituting this into (2.5), using the fact that $s_{n,j} \xrightarrow{P} s_j$, we have

$$\begin{aligned} \hat{\boldsymbol{\beta}}^*(x) &= \boldsymbol{\beta}(x) + \text{diag}(h^{p+1}, \cdots, h)\boldsymbol{S}_n^{-1}\left\{\frac{m^{(p+1)}(x)}{(p+1)!}\Big(s_{n,p+1}(x), \cdots, s_{n,2p+1}(x)\Big) + o_P(1)\right\} \\ &= \boldsymbol{\beta}(x) + \frac{m^{(p+1)}(x)}{(p+1)!}\text{diag}(h^{p+1}, \cdots, h)\left\{\boldsymbol{S}^{-1}\boldsymbol{\mu} + o_P(1)\right\}. \end{aligned} \tag{2.7}$$

Therefore, $\hat{\beta}(x)$ is asymptotically unbiased estimator to $\beta(x)$. The order of bias is also indicated in (2.7). For dependent data, we shall use the bias expressions (2.5) and (2.7) to center our estimators even though they can not be obtained by conditioning arguments on $X_1, \cdots, X_n$ as in the i.i.d. case.

We now introduce the mixing coefficients. Let $\mathcal{F}_i^k$ be the σ-algebra of events generated by the random variables $\{X_j, Y_j, i \le j \le k\}$ and $L_2(\mathcal{F}_i^k)$ denote the collection of all second-order random variables which are $\mathcal{F}_i^k$-measurable. The stationary processes $\{X_j, Y_j\}$ are called strongly mixing (Rosenblatt, 1956) if

$$\sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_k^\infty} |P(AB) - P(A)P(B)| = \alpha(k) \to 0 \quad \text{as} \quad k \to \infty$$

and are called ρ-mixing (Kolmogorov and Rozanov, 1960) if

$$\sup_{U\in L_2(\mathcal{F}_{-\infty}^0), V\in L_2(\mathcal{F}_k^\infty)} \frac{|\mathrm{cov}(U, V)|}{\mathrm{var}^{1/2}(U)\mathrm{var}^{1/2}(V)} = \rho(k) \to 0 \quad \text{as} \quad k \to \infty.$$

It is well known that these mixing coefficients satisfy

$$\alpha(k) \le \frac{1}{4}\rho(k).$$

2.2 Results

We make the following assumptions on the kernel function and the mixing processes:

Condition 1:

i) The kernel function $K \in L_1$ is bounded and $u^{2p+1}K(u) \to 0$, as $|u| \to \infty$.

ii) $|f(u, v; \ell) - f(u)f(v)| \le M < \infty, \forall \ell \ge 1$, where $f(u)$ and $f(u, v; \ell)$ denote respectively the density of X_0 and (X_0, X_ℓ).

iii) Either the processes $\{X_j, Y_j\}$ are ρ-mixing with $\sum \rho(\ell) < \infty$; or are strongly mixing with $\sum \ell^a[\alpha(\ell)]^{1-2/\delta} < \infty$ for some $\delta > 2$ and $a > 1 - 2/\delta$. In the latter case, we assume further $u^{2\delta p+2}K(u) \to 0$ as $|u| \to \infty$.

Theorem 1. *Under Condition 1 and the assumption that* $h_n \to 0, \quad nh_n \to \infty$, *as* $n \to \infty$, *we have at every continuity point of* f,

$$Es_{n,j} \to f(x)\mu_j, \qquad nh_n \operatorname{var}(s_{n,j}) \to f(x)\nu_{2j},$$

for each $0 \le j \le 2p$ *and*

$$\boldsymbol{S}_n \xrightarrow{m.s.} f(x)\boldsymbol{S}$$

in the sense that each element converges in mean square.

To study the joint asymptotic normality of $\hat{\boldsymbol{\beta}}(x)$, we need to center the vector $\boldsymbol{t}_n$ as follows: Let

$$t^*_{n,j} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - x}{h}\right)^j K_h(X_i - x)(Y_i - m(X_i)), \quad \boldsymbol{t}^*_n = (t_{n,0}, \cdots, t_{n,p})^T.$$

Then consider the arbitrary linear combination of $t^*_{n,j}$.

$$Q_n = \sum_{j=0}^{p} c_j t^*_{n,j} = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{2.8}$$

where with $C(u) = \sum_{j=0}^{p} c_j u^j K(u)$ and $C_h(u) = C(u/h)/h$,

$$Z_i = (Y_i - m(X_i))C_h(X_i - x). \tag{2.9}$$

Once the joint asymptotic normality of $t^*_{n,j}$ is established, by using (2.3) – (2.7), we can easily obtain the joint asymptotic normality of $\hat{\boldsymbol{\beta}}(x)$. We need the following conditions.

Condition 2.

i) The kernel K is bounded with a bounded support.

ii) Assume that

$$f_{X_0,X_\ell|Y_0,Y_\ell}(x_1, x_\ell|y_1, y_\ell) \le A_1 < \infty, \forall \ell \ge 1. \tag{2.10}$$

iii) For ρ-mixing processes we assume that

$$\sum \rho(\ell) < \infty, \quad EY_0^2 < \infty;$$

for strongly mixing processes, we assume that for some $\delta > 2$ and $a > 1 - 2/\delta$,

$$\sum \ell^a[\alpha(\ell)]^{1-2/\delta} < \infty, \quad E|Y_0|^\delta < \infty, \quad f_{X_0|Y_0}(x|y) \le A_2 < \infty. \tag{2.11}$$

Put $\sigma^2(x) = \text{var}(Y|X = x)$.

Theorem 2. *Under Condition 2, if $h_n \to 0$ and $nh_n \to \infty$, we have the following convergence at every continuity point of $\sigma^2 f$:*

a) $h_n \text{var}(Z_1) \to \sigma^2(x)f(x)\int_{-\infty}^{+\infty} C^2(u)du$.

b) $h_n \sum_{\ell=1}^{n-1} |\text{cov}(Z_1, Z_{\ell+1}| = o(1)$;

c) $nh_n \text{ var}(Q_n) \to \sigma^2(x)f(x)\int_{-\infty}^{+\infty} C^2(u)du$;

d) $nh_n \text{ cov}(\mathbf{t}_n^*) \to f(x)\sigma^2(x)\tilde{S}$, *where* $\text{cov}(\mathbf{t}_n^*)$ *denotes the variance-convariance matrix of* $\mathbf{t}_n^*$.

2.3 Proofs

The proof of Theorem 1 is similar to that of Theorem 2. Since the proof of Theorem 2 is more involved, we only prove Theorem 2. In the following, we denote by D a generic constant, which may take different values at different places. We suppress the dependence of h_n on n.

First of all, we have by conditioning on X_1,

$$\begin{aligned} \text{var}(Z_1) &= E\sigma^2(X_1)C_h^2(X_1 - x) \\ &= \frac{1}{h}\left(\sigma^2(x)f(x)\int_{-\infty}^{+\infty} C^2(x)dx + o(1)\right), \end{aligned} \tag{2.12}$$

at a continuity point of $\sigma^2 f$. The result c) follows directly from a) and b) along with

$$\text{var}(Q_n) = \frac{1}{n}\text{var}(Z_1) + \frac{2}{n}\sum_{\ell=1}^{n-1}(1 - \ell/n)\text{cov}(Z_1, Z_{\ell+1}). \tag{2.13}$$

Conclusion d) follows directly from conclusion c) and some simple algebra. So, it remains to prove part b).

Let $d_n \to \infty$ be a sequence of integers such that $d_n h_n \to 0$. Define

$$J_1 = \sum_{\ell=1}^{d_n-1} |\text{cov}(Z_1, Z_{\ell+1})|, \quad J_2 = \sum_{\ell=d_n}^{n-1} |\text{cov}(Z_1, Z_{\ell+1})|.$$

It remains to show that $J_1 = o(1/h)$ and $J_2 = o(1/h)$.

We remark that since K has a bounded support, $m(X_j)$ is bounded in the neighborhood of $X_j \in x \pm h$. Let $B = \sup_{X \in x \pm h} |m(X)|$. By conditioning on (Y_1, Y_ℓ) and using (2.10), we have for $\ell > 1$,

$$\begin{aligned} |\mathrm{cov}(Z_1, Z_\ell)| &= |E(Y_1 - m(X_1))(Y_\ell - m(X_\ell))C_h(X_1 - x)C_h(X_\ell - x)| \\ &\le A_1 E(|Y_1| + B)(|Y_\ell| + B)\left(\int_{-\infty}^{+\infty} |C_h(u - x)|du\right)^2 \\ &\le D. \end{aligned} \tag{2.14}$$

It follows that

$$J_1 \le d_n D = o(1/h_n),$$

by the choice of d_n.

Next, we consider the contribution of J_2. For ρ-mixing process,

$$|\mathrm{cov}(Z_1, Z_{\ell+1})| \le \rho(\ell)\mathrm{var}(Z_1).$$

By using (2.12), we have

$$J_2 \le \mathrm{var}(Z_1) \sum_{j=d_n}^{\infty} \rho(j) = o(1/h).$$

For strongly mixing process, we have by using Davydov's lemma [see Hall and Heyde (1980), Corollary A2]

$$|\mathrm{cov}(Z_1, Z_{\ell+1})| \le 8[\alpha(\ell)]^{1-2/\delta}[E|Z_1|^\delta]^{2/\delta}. \tag{2.15}$$

By conditioning on Y_1 and using (2.11), we have

$$E|Z_1|^\delta \le A_2 E(|Y_1| + B)^\delta \int_{-\infty}^{+\infty} |C_h(x - u)|^\delta \le D h^{-\delta+1}. \tag{2.16}$$

The combination of (2.15) and (2.16) leads to

$$\begin{aligned} J_2 &\le D h^{2/\delta - 2} \sum_{\ell=d_n}^{\infty} [\alpha(\ell)]^{1-2/\delta} \\ &\le D h^{2/\delta - 2} d_n^{-a} \sum_{\ell=d_n}^{\infty} \ell^a [\alpha(\ell)]^{1-2/\delta} \\ &= o(1/h) \end{aligned}$$

by taking $h^{1-2/\delta} d_n^a = 1$ so that the requirement that $d_n h_n \to 0$ is satisfied.

3 Joint asymptotic normality for mixing processes

3.1 Main results

We need the following conditions in order to establish the joint asymptotic normality of $\{m^{(\nu)}(x)\}_{\nu=0}^{p}$.

Condition 3. Assume that $h_n \to 0$ and $nh_n \to \infty$. For ρ-mixing and strongly mixing processes, we assume respectively that there exists a sequence of positive integers satisfying $s_n \to \infty$ and $s_n = o((nh_n)^{1/2})$ such that

$$(n/h_n)^{1/2}\rho(s_n) \to 0 \quad \text{and} \quad (n/h_n)^{1/2}\alpha(s_n) \to 0, \text{ as } n \to \infty.$$

Theorem 3. *Under Conditions 1 - 3, we have the following asymptotic normality as $n \to \infty$:*

$$\sqrt{nh_n}Q_n \xrightarrow{\mathcal{L}} N(0, \theta^2(x)),$$

where Q_n is defined by (2.8) and $\theta^2(x) = f(x)\sigma^2(x)\int_{-\infty}^{+\infty} C^2(u)du$.

The proof of Theorem 3 is given in Section 3.2. Before we give the implications of Theorem 3, we make a few remarks on the mixing conditions and the applicability of the theorem.

Remark 1. It can be shown that if $h_n = dn^{-\gamma}, (0 < \gamma < 1, d > 0)$, a sufficient condition for Condition 3 is respectively $\alpha(n) = O(n^{-\gamma'})$ and $\rho(n) = O(n^{-\gamma'})$ with $\gamma' > \frac{1+\gamma}{1-\gamma}$ for strongly mixing and ρ-mixing processes [with $s_n = (nh_n)^{1/2}/\log n)$]. In particular, if $\gamma = 1/5$, then $\gamma' > 1.5$. A sufficient condition for (2.11) is $\alpha(n) = O(n^{-\gamma''}), \gamma'' > \frac{2\delta-2}{\delta-2}$. Therefore, if $h_n \sim n^{-1/5}$ and $EY_0^4 < \infty$, a sufficient condition for the mixing coefficients to satisfy Conditions 1 - 3 is $\alpha(n) = O(n^{-\gamma^*})$ with $\gamma^* > 3$, $\rho(n) = O(n^{-\gamma^*})$ with $\gamma^* > 1.5$.

Remark 2. Theorem 3 remains valid when one estimates a function of form $m_\psi(x) = E(\psi(Y)|X = x)$. The only changes are that the moment conditions should now be either $E|\psi(Y_0)|^2 < \infty$ (for ρ-mixing processes) or $E|\psi(Y_0)|^\delta < \infty$ for some $\delta > 2$ (for strongly mixing processes). Under Conditions 1 – 3, Theorem 3 now reads as

$$\sqrt{h_n/n}\sum_{i=1}^{n}(\psi(Y_i) - m_\psi(X_i))C_h(X_i - x) \xrightarrow{\mathcal{L}} N\left(0, \sigma_\psi^2(x)f(x)\int_{-\infty}^{+\infty} C^2(u)du\right), \qquad (3.1)$$

where $\sigma_\psi^2(x) = \text{var}(\psi(Y)|X = x)$.

We now give some applications of Theorem 3. Since Theorem 3 holds for all linear combinations of $t_{n,j}^*$, we have the joint asymptotic normality (see also part (d) of Theorem 2:)

$$\sqrt{nh_n}\boldsymbol{t}_n^* = \sqrt{nh_n}(t_{n,0}^*, \cdots, t_{n,p}^*)^T \xrightarrow{\mathcal{L}} N(0, \sigma^2(x)f(x)\tilde{\boldsymbol{S}}),$$

where $\tilde{\boldsymbol{S}}$ is given by (2.4). Thus, by Theorem 1,

$$\sqrt{nh_n}\boldsymbol{S}_n^{-1}\boldsymbol{t}_n^* \xrightarrow{\mathcal{L}} N(0, \sigma^2(x)\boldsymbol{S}^{-1}\tilde{\boldsymbol{S}}\boldsymbol{S}^{-1}/f(x)) \qquad (3.2)$$

at continuity points of $\sigma^2 f$ whenever $f(x) > 0$. By (2.3), (2.5) and (2.7), we have

$$\begin{aligned}\boldsymbol{S}_n^{-1}\boldsymbol{t}_n^* &= \text{diag}(1, \cdots, h_n^p)[\hat{\boldsymbol{\beta}}(x) - \hat{\boldsymbol{\beta}}^*(x)] \\ &= \text{diag}(1, \cdots, h_n^p)[\hat{\boldsymbol{\beta}}(x) - \boldsymbol{\beta}(x)] - \frac{h_n^{p+1}m^{(p+1)}(x)}{(p+1)!}\boldsymbol{S}^{-1}\boldsymbol{\mu} + o_P(h_n^{p+1}).\end{aligned}$$

This and (3.2) give

Theorem 4. *Under Conditions 1 – 3, if $h_n = O(n^{1/(2p+1)})$, then, as $n \to \infty$,*

$$\begin{aligned}&\sqrt{nh_n}\left(\text{diag}(1, \cdots, h_n^p)[\hat{\boldsymbol{\beta}}(x) - \boldsymbol{\beta}(x)] - \frac{h_n^{p+1}m^{(p+1)}(x)}{(p+1)!}\boldsymbol{S}^{-1}\boldsymbol{\mu}\right) \\ &\xrightarrow{\mathcal{L}} N(0, \sigma^2(x)\boldsymbol{S}^{-1}\tilde{\boldsymbol{S}}\boldsymbol{S}^{-1}/f(x))\end{aligned}$$

at continuity points of $\sigma^2 f$ whenever $f(x) > 0$.

Theorem 4 gives the joint asymptotic normality for the estimators $\hat{\beta}_\nu(x) = \hat{m}^{(\nu)}(x)/\nu!$. In particular, we have the following convergence for the individual components.

Theorem 5. *Under Conditions 1 – 3, if $h_n = O(n^{1/(2p+1)})$, then, as $n \to \infty$,*

$$\sqrt{nh_n^{2\nu+1}}\left(\hat{m}^{(\nu)}(x) - m^{(\nu)}(x) - \frac{m^{(p+1)}(x)\nu!B_\nu}{(p+1)!}h_n^{p+1-\nu}\right) \xrightarrow{\mathcal{L}} N\left(0, \frac{(\nu!)^2V_\nu\sigma^2(x)}{f(x)}\right),$$

at continuity points of $\sigma^2 f$ with $f(x) > 0$, where B_ν and V_ν are, respectively, the ν^{th} element of $\boldsymbol{S}^{-1}\boldsymbol{\mu}$ and the ν^{th} diagonal element of $\boldsymbol{S}^{-1}\tilde{\boldsymbol{S}}\boldsymbol{S}^{-1}$.

Theorem 5 indicates that the local polynomial fit has the asymptotic bias and variance respectively as:

$$\text{bias of } \hat{m}^{(\nu)}(x) = \frac{m^{(p+1)}(x)\nu! B_\nu}{(p+1)!} h_n^{p+1-\nu}, \qquad \text{"variance" of } \hat{m}^{(\nu)}(x) = \frac{(\nu!)^2 V_\nu \sigma^2(x)}{n h_n^{2\nu+1} f(x)}. \tag{3.3}$$

The optimal bandwidth for estimating the ν^{th} derivative can be defined to be the one which minimizes the squared bias plus variance. Assume that $p - \nu$ is odd (see Fan and Gijbels (1993) for the reason why this assumption is needed and why this assumption is natural). Then the optimal bandwidth is given by

$$h_{\nu,opt} = \left(\frac{[(p+1)!]^2 V_\nu \sigma^2(x)/f(x)}{2(p+1-\nu)[m^{(p+1)}(x)]^2 B_\nu^2} \right)^{1/(2p+3)}.$$

By (3.1), we have the following generalized results.

Theorem 6. *Let $\hat{m}_\psi(x)$ be the estimator of $m_\psi(x) = E(\psi(Y)|X = x)$ by using a local polynomial fit. Under Conditions 1 - 3 with suitable modification as indicated in Remark 2, if $h_n = O(n^{1/(2p+1)})$, then as $n \to \infty$*

$$\sqrt{n h_n^{2\nu+1}} \left(\hat{m}_\psi^{(\nu)}(x) - m_\psi^{(\nu)}(x) - \frac{m_\psi^{(p+1)}(x)\nu! B_\nu}{(p+1)!} h^{p+1-\nu} \right) \xrightarrow{\mathcal{L}} N\left(0, (\nu!)^2 V_\nu \sigma_\psi^2(x)/f(x)\right)$$

at continuity points of $\sigma^2 f$ whenever $f(x) > 0$, where B_ν and V_ν are given in Theorem 5, and $\sigma_\psi^2(x) = \mathrm{var}(\psi(Y)|X = x)$.

Example 1. a) When $\psi(Y) = I\{Y \le y\}$, the problem corresponds to estimating the conditional distribution $m_\psi(x) = P(Y \le y|X = x)$ and its derivative with respect to x. In this case,

$$\sigma_\psi^2(x) = m_\psi(x)(1 - m_\psi(x)).$$

b) If $\psi(Y) = Y^2$, then the problem corresponds to estimating conditional second moment. In this case,

$$\sigma_\psi^2(x) = E\left(Y^4|X = x\right) - \left[E\left(Y^2|X = x\right)\right]^2.$$

3.2 Proof of Theorem 3

We employ the small-block and large-block argument. Partition the set $\{1, \cdots, n\}$ into $2k+1$ subsets with large blocks of size $r = r_n$ and small block of size $s = s_n$. Put

$$k = k_n = \lfloor \frac{n}{r_n + s_n} \rfloor \tag{3.4}$$

In the following derivation, we suppress the dependence of h_n on n. Let $Z_{n,i} = \sqrt{h} Z_{i+1}, i = 0, \cdots, n-1$. Then,

$$\sqrt{nh} Q_n = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} Z_{n,i}$$

and by Theorem 2

$$\operatorname{var}(Z_{n,0}) = \theta^2(x)(1 + o(1)), \quad \sum_{\ell=1}^{n-1} |\operatorname{cov}(Z_{n,0}, Z_{n,\ell})| = o(1). \tag{3.5}$$

Define the random variables

$$\eta_j = \sum_{i=j(r+s)}^{j(r+s)+r-1} Z_{n,i}, \quad 0 \le j \le k-1,$$

$$\xi_j = \sum_{i=j(r+s)+r}^{(j+1)(r+s)-1} Z_{n,i}, \quad 0 \le j \le k-1,$$

and

$$\zeta_k = \sum_{i=k(r+s)}^{n-1} Z_{n,i}.$$

Then,

$$\begin{aligned} \sqrt{nh} Q_n &= \frac{1}{\sqrt{n}} \left\{ \sum_{j=0}^{k-1} \eta_j + \sum_{j=0}^{k-1} \xi_j + \zeta_k \right\} \\ &\equiv \frac{1}{\sqrt{n}} \left\{ Q_n' + Q_n'' + Q_n''' \right\}. \end{aligned} \tag{3.6}$$

We will show that as $n \to \infty$,

$$\frac{1}{n} E(Q_n'')^2 \to 0, \quad \frac{1}{n} E(Q_n''')^2 \to 0 \tag{3.7}$$

$$\left| E\left[\exp(itQ_n')\right] - \prod_{j=0}^{k-1} E\left[\exp(it\eta_j)\right] \right| \to 0 \tag{3.8}$$

$$\frac{1}{n}\sum_{j=0}^{k-1} E\left(\eta_j^2\right) \to \theta^2(x) \tag{3.9}$$

$$\frac{1}{n}\sum_{j=0}^{k-1} E\left(\eta_j^2 I\{|\eta_j| \geq \varepsilon\theta(x)\sqrt{n}\}\right) \to 0 \tag{3.10}$$

for every $\varepsilon > 0$. (3.7) implies that Q_n'' and Q_n''' are asymptotically negligible, (3.8) implies that the summands $\{\eta_j\}$ in Q_n' are asymptotically independent, and (3.9) and (3.10) are the standard Lindeberg-Feller conditions for asymptotic normality of Q_n' under independence. Expressions (3.7) – (3.10) entail the following asymptotic normality:

$$\sqrt{nh}Q_n \xrightarrow{\mathcal{L}} N\left(0, \theta^2(x)\right). \tag{3.11}$$

We now establish (3.7) – (3.10). The proof concentrates on the strongly mixing case. We remark on the difference for ρ-mixing processes.

We first choose the block sizes. Condition 3 implies that there exist constants $q_n \to \infty$ such that

$$q_n s_n = o(\sqrt{nh_n}); \quad q_n(n/h_n)^{1/2}\alpha(s_n) \to 0. \tag{3.12}$$

[For ρ-mixing process, $q_n(n/h_n)^{1/2}\rho(s_n) \to 0$]. Define the large block size r_n by

$$r_n = \lfloor (nh_n)^{1/2}/q_n \rfloor.$$

Then, it can easily be shown that, as $n \to \infty$,

$$s_n/r_n \to 0, \quad r_n/n \to 0, \quad r_n/(nh_n)^{1/2} \to 0, \tag{3.13}$$

and

$$\frac{n}{r_n}\alpha(s_n) \to 0. \tag{3.14}$$

[For ρ-mixing processes, (3.14) is proved via the inequality $\alpha(s_n) \leq \rho(s_n)/4$.]

We now establish (3.7). First of all, by stationarity and (3.5),

$$\mathrm{var}(\xi_j) = s\mathrm{var}(Z_{n,0}) + 2s\sum_{j=1}^{s-1}(1 - j/s)\mathrm{cov}(Z_{n,0}, Z_{n,j}) = s\theta^2(x)(1 + o(1)).$$

and

$$E(Q_n'')^2 = \sum_{j=0}^{k-1} \operatorname{var}(\xi_j) + \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ i \neq j}}^{k-1} \operatorname{cov}(\xi_i, \xi_j) \equiv F_1 + F_2. \tag{3.15}$$

By (3.4) and (3.13), $r_n k_n / n \leq s_n/(r_n + s_n) \to 0$ so that $F_1 = O(k_n s_n) = o(n)$. Now, we consider F_2. We first note that with $m_j = j(r+s) + r$,

$$F_2 = \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ i \neq j}}^{k-1} \sum_{\ell_1=0}^{s-1} \sum_{\ell_2=0}^{s-1} \operatorname{cov}(Z_{n,m_i+\ell_1}, Z_{n,m_j+\ell_2}),$$

but since $i \neq j$, $|m_i - m_j + \ell_1 - \ell_2| \geq r$ so that

$$|F_2| \leq 2 \sum_{\ell_1=0}^{n-r-1} \sum_{\ell_2=\ell_1+r}^{n-1} |\operatorname{cov}(Z_{n,\ell_1}, Z_{n,\ell_2})|.$$

By stationarity and (3.5)

$$|F_2| \leq 2n \sum_{j=r}^{n-1} |\operatorname{cov}(Z_{n,0}, Z_{n,j})| = o(n).$$

This together with (3.15) validate the first part of (3.7). For the second part of (3.7), using a similar argument together with (3.5), we obtain that

$$\begin{aligned} \frac{1}{n} E(Q_n''')^2 &\leq \frac{1}{n}(n - k(r+s)) \operatorname{var}(Z_{n,0}) + 2 \sum_{j=1}^{n-1} |\operatorname{cov}(Z_{n,0}, Z_{n,j})| \\ &\leq \frac{r_n + s_n}{n} \theta^2(x) + o(1) \to 0. \end{aligned}$$

Equation (3.8) is now proved as follows. Note that η_a is $\mathcal{F}_{i_a}^{j_a}$-measurable with $i_a = a(r+s)+1$ and $j_a = a(r+s)+r$. Hence, applying Lemma 1 (see the end of the proof) with $V_j = \exp(it\eta_j)$, we have

$$\left| E\exp(itQ') - \prod_{j=0}^{k-1} E[\exp(it\eta_j)] \right| \leq 16k\alpha(s_n+1) \sim 16 \frac{n}{r_n} \alpha(s_n+1),$$

which tends to zero by (3.14).

We now show (3.9). By stationarity and (3.5),

$$\operatorname{var}(\eta_j) = \operatorname{var}(\eta_0) = r_n \theta^2(x)(1 + o(1)).$$

This implies that

$$\frac{1}{n}\sum_{j=0}^{k-1} E(\eta_j^2) = \frac{k_n r_n}{n}\theta^2(x)(1+o(1)) \sim \frac{r_n}{r_n+s_n}\theta^2(x) \to \theta^2(x),$$

since $s_n/r_n \to 0$.

It remains to establish (3.10). We employ a truncation argument as follows. Let

$$a_L(y) = yI\{|y| \le L\},$$

where L is a fixed truncation point. Correspondingly let

$$m_L(x) = E\left(a_L(Y_j)\Big|X_j = x\right),$$

and

$$V_L^2(x) = E\left[(a_L(Y_j) - m_L(X_j))^2\Big|X_j = x\right], \qquad \theta_L^2 = V_L^2(x)f(x)\int_{-\infty}^{+\infty} C^2(u)du.$$

Put

$$Z_j^L = (a_L(Y_j) - m_L(X_j))C_h(X_j - x), \quad \text{and} \quad Z_{n,j-1} = \sqrt{h_n}Z_j^L, \quad j = 1, \cdots, n.$$

and

$$Q_n^L = n^{-1}\sum_{j=1}^{n} Z_j^L, \qquad \tilde{Q}_n^L = n^{-1}\sum_{j=1}^{n}(Z_j - Z_j^L). \tag{3.16}$$

Using the fact that $C(\cdot)$ is bounded (since K is bounded with compact support), we have

$$|Z_{n,j}^L| \le D/h_n^{1/2},$$

for some constant D. This entails that $\max_{0\le j\le k-1}|\eta_j^L|/\sqrt{n} \le Dr_n/\sqrt{nh_n} \to 0$, by (3.13). Hence, when n is large the set $\{|\eta_j^L| \ge \theta_L(x)\varepsilon\sqrt{n}\}$ becomes an empty set, namely (3.10) holds. Consequently, we have the following asymptotic normality:

$$\sqrt{nh_n}Q_n^L \xrightarrow{\mathcal{L}} N(0, \theta_L^2). \tag{3.17}$$

In order to complete the proof, namely to establish (3.11), it suffices to show that as first $n \to \infty$ and then $L \to \infty$ we have

$$nh_n \text{var}\left(\tilde{Q}_n^L\right) \longrightarrow 0. \tag{3.18}$$

Indeed,

$$
\begin{aligned}
& \left| E \exp(it\sqrt{nh_n} Q_n) - \exp(-t^2\theta^2(x)/2) \right| \\
= & \left| E \exp(it\sqrt{nh_n}(Q_n^L + \tilde{Q}_n^L)) - \exp(-t^2\theta_L^2/2) + \exp(-t^2\theta_L^2/2) - \exp(-t^2\theta^2(x)/2) \right| \\
\leq & \left| E \exp(it\sqrt{nh_n} Q_n^L) - \exp(-t^2\theta_L^2/2) \right| + E \left| \exp(it\sqrt{nh_n}\tilde{Q}_n^L) - 1 \right| \\
& \quad + \left| \exp(-t^2\theta_L^2/2) - \exp(-t^2\theta^2/2) \right|.
\end{aligned}
$$

Letting $n \to \infty$, the first term goes to zero by (3.17) for every $L > 0$; the second term converges to zero by (3.18) as first $n \to \infty$ and then $L \to \infty$; the third term goes to zero as $L \to \infty$ by the dominated convergence theorem. Therefore, it remains to prove (3.18). Note that by (3.16) $\tilde{Q}_n^L$ has the same structure as Q_n except that the function Y_i is replaced by $Y_i I\{|Y_i| > L\}$. Hence, by part c) of Theorem 2, we have

$$
\lim_{n\to\infty} nh_n \mathrm{var}\left(\tilde{Q}_n^L\right) = \mathrm{var}(YI[|Y| > L]|X = x) f(x) \int_{-\infty}^{+\infty} C^2(u)du.
$$

By dominated convergence, the right hand side converges to 0 as $L \to \infty$. This establishes (3.18) and completes the proof of Theorem 3. □

Lemma (Volkonskii and Rozanov, 1959). *Let $V_1, \cdots, V_L$ be random variables measurable with respect to the σ-algebras $\mathcal{F}_{i_1}^{j_1}, \cdots, \mathcal{F}_{i_L}^{j_L}$ respectively with $1 \leq i_1 < j_1 < i_2 < \cdots < j_L \leq n$, $i_{l+1} - j_l \geq w \geq 1$ and $|V_j| \leq 1$ for $j = 1, \cdots, L$. Then*

$$
\left| E\left(\prod_{j=1}^{L} V_j\right) - \prod_{j=1}^{L} E(V_j) \right| \leq 16(L-1)\alpha(w),
$$

where $\alpha(w)$ is the strongly mixing coefficient.

REFERENCES

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Jour. Amer. Statist. Assoc.*, **74** 829 – 836.

Cleveland, W.S. (1993). Coplots, nonparametric regression, and conditionally parametric fits. *Manuscript.*

Chu, C. K. and Marron, J.S. (1991). Choosing a kernel regression estimator (with discussion). *Statist. Sci.*, **6**, 404-436.

Collomb, G. and Härdle, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stoch. Processes Appl.*, **23**, 77-89.

Fan, J. (1992). Design-adaptive nonparametric regression. *Jour. Amer. Statist. Assoc.*, **87**, 998 - 1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.*, **21**, 196 - 216.

Fan, J. , Gasser, T., Gijbels, I. , Brockman, M. and Engel, J. (1993). Local polynomial fitting: a standard for nonparametric regression. *Manuscript.*

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**, 2008 - 2036.

Fan, J. and Gijbels, I. (1993). Bandwidth and adaptive order selection for local polynomial fitting in function estimation. *Manuscript.*

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Applications.* Academic Press, New York.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Boston.

Hastie, T. and Loader, C. (1993). Local regression: automatic kernel carpentry. (with discussions). *Statistical Science*, **8**, 120-143.

Kolmogorov, A.N. and Rozanov, Yu. A. (1960). On strong mixing conditions for stationary Gaussian processes. *Theory Prob. Appl.*, **52**, 204-207.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie view. Gebiete*, **61**, 405-415.

Nadaraya, E. A. (1964). On estimating regression. *Theor. Prob. Appl.*, **9**, 141-142.

Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Series Anal.*, **4**, 185-297.

Robinson, P. M. (1986). On the consistency and finite sample properties of nonparametric kernel time series regression, auto regression, and density estimators. *Ann. Inst. Statist. Math.*, **38**, 539-549.

Rosenblatt, M. (1956). A central limit theorem and strong mixing conditions. *Proc. Nat. Acad. Sci.*, **4**, 43-47.

Rosenblatt, M. (1969). Conditional probability density and regression estimates. In *Multivariate Analysis II*, ed. Krishnaiah, 25-31.

Roussas, G.G. (1990). Nonparametric regression estimation under mixing conditions. *Stochastic Processes and Their Applications*, **36**, 107-116.

Roussas, G. G. and Tran, L. T. (1991). Asymptotic normality of the recursive kernel regression estimate under dependence conditions. *Ann. Statist.*, to appear.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1993). An effective bandwidth selector for local least squares regression. *Manuscript.*

Ruppert, D. and Wand, M.P. (1993). Multivariate weighted least squares regression. To appear in *Ann. Statist.*

Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach.* Oxford University Press, Oxford.

Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *J. Statistical Planning and Inference*, **28**, 167-183..

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, **22** 133–146.

Volkonskii, V.A. and Rozanov, Yu. A. (1959). Some limit theorems for random functions. *Theory Prob. Appl.*, 4, 178-197.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A.*, **26**, 359-372.

Stone, C.J. (1977). Consistent Nonparametric Regression. *Ann. Statist.*, **5**, 595 – 645.

Yao, Q. and Tong, H. (1993a). Qualifying the influence of initial values on nonlinear prediction. *Jour. Roy. Statist. Assoc.*, to appear.

Yao, Q. and Tong, H. (1993b). Asymmetric least squares regression estimation: a nonparametric approach. *Manuscript.*