Weighted Nadaraya-Watson Regression Estimation

ZONGWU CAI
Department of Mathematics
University of North Carolina
Charlotte, NC 28223, USA
E-mail: zcai@uncc.edu

Abstract

In this article we study nonparametric estimation of regression function by using the weighted Nadaraya-Watson approach. We establish the asymptotic normality and weak consistency of the resulting estimator for α-mixing time series at both boundary and interior points, and we show that the estimator preserves the bias, variance, and more importantly, automatic good boundary behavior properties of local linear estimator. Also, the asymptotic minimax efficiency is discussed. Finally, comparisons between weighted Nadaraya-Watson approach and local linear fitting are given.

Key words and phrases: α-mixing; asymptotic properties; forecasting; local linear smoothers; minimax efficiency; Nadaraya-Watson estimator; nonparametric regression function; prediction interval; time series analysis.

AMS 1991 subject classification. Primary 62G07, 62M10; secondary 62G20, 62F35.

1 Introduction

In various statistical problems, regression techniques are commonly used for modeling the relationship between response and covariates for both independent and time series data. The main purpose of this article is to study an easily implemented smoothing method for exploring the association between response and covariates when the observed values of response variable are not possibly independent (see later for the kind of dependence stipulated). Although we only focus on the univariate case, we would like to mention that the basic ideas of our methodology hold for the case of multivariate situations. One of key applications of regression estimation, this time involving dependent data, is construction of prediction intervals for the next value in stationary time series $\{Y_1, \ldots, Y_n\}$. If the time series is Markovian, then we may solve easily the prediction problem by estimating the expectation of Y_{n+1}, conditional on $Y_n = x$.

There is a vast literature on modeling the relationship between two variables. There are many linear smoothers proposed to estimate nonparametric regression functions: kernel, spline, local polynomial, orthogonal series methods, and others. For the available methods and results on both theory and applications, we refer to the books by Eubank (1988), Müller (1988), Härdle (1990), Wahba (1990), Hastie and Tibshirani (1990), Green and Silverman (1994), Wand and Jones (1995), Simonoff (1996), Fan and Gijbels (1996), Bosq (1998), among others. Among the aforementioned linear smoothers, local polynomial method has become popular in recent years due to its attractive mathematical efficiency, bias reduction and adaptation of edge effects; see the book by Fan and Gijbels (1996) for the detailed discussions on methods and results. In the time series context, the nonparametric estimates of regression function have been investigated by many authors in the case where the observations exhibit some kind of dependence such as Markovian chains, ergodic processes, mixing sequences, long-range memory processes, association random variables, and so on. Here, we mention just few references, such as Robinson (1983), Roussas (1990) and Troung and Stone (1992) for employing kernel methods such as Nadaraya-Watson and Masry and Fan (1997) for using local polynomial approaches to estimate regression curves.

It can be seen easily that the implementation of the Nadaraya-Watson estimator of regression function is much easier than the local linear estimator, and the estimated values of regression function are always within the range of the response variables. However, it is well-known that the Nadaraya-Watson estimator is inferior to the local linear estimator due to the limitations such as larger bias, non-adaptation and boundary effects. To take the advantages of both the Nadaraya-Watson and the local linear estimate, a weighted version of Nadaraya-Watson estimate is prosed. The basic idea is to use the weighted local constant fitting. In particular, this method is useful for estimating conditional distribution and regression quantile. This approach was proposed first by Hall and Presnell (1999) for estimating regression function under the independent samples and they derived the mean squared error of the estimator. For the time series data, it was used by Hall, Wolff and Yao (1998) and Cai (1999) for estimating conditional distribution and Cai (1999) for estimating conditional quantile. But there has not been previously advocated in the general nonparametric regression setup for time series data. Therefore, the main purpose of this article is to explore the weighted Nadaraya-Watson method in detail and to obtain the asymptotic properties such as consistency, normality and efficiency at both interior and boundary points, and to made a comparison between weighted Nadaraya-Watson and local linear.

Although our interest in nonparametric estimation of regression function is motivated by the forecasting from time series data, we introduce our methods in a more general setting (α-mixing) which includes time series modeling as a special case. Our theoretical results are derived under α-mixing assumption.

For reference convenience, we first introduce the mixing coefficient. Let $\mathcal{F}_a^b$ be the σ-algebra generated by $\{(X_t, Y_t)\}_{t=a}^b$. Define

$$\alpha(t) = \sup\{|P(A\,B) - P(A)\,P(B)| : \ A \in \mathcal{F}_{-\infty}^0, \ B \in \mathcal{F}_t^\infty\}.$$

It is called the strong mixing coefficient of the stationary process $\{(X_t, Y_t)\}_{-\infty}^\infty$. If $\alpha(t) \to 0$ as $t \to \infty$, the process is called strongly mixing or α-mixing.

Among various mixing conditions used in literature, α-mixing is reasonably weak, and is known to be fulfilled for many time series models. Gorodetskii (1977) and Withers (1981) derived the conditions under which a linear process is α-mixing. In fact, under very mild assumptions linear autoregressive and more generally bilinear time series models are α-mixing with mixing coefficients decaying exponentially. Auestad and Tjøstheim (1990) provided illuminating discussions on the role of α-mixing (including geometric ergodicity) for model identification in nonlinear time series analysis. Chen and Tsay (1993) showed that the functional autoregressive process is geometrically ergodic under certain conditions. Furthermore, Masry and Tjøstheim (1995, 1997) demonstrated that under some mild conditions, both ARCH processes and nonlinear additive autoregressive models with exogenous variables, which are particularly popular in finance and econometrics, are stationary and α-mixing.

The plan of the paper is as follows. In Section 2, we introduce the weighted Nadaraya-Watson estimation, and in Section 3, we explore the asymptotic normality and week consistency of the estimator and the asymptotic minimax efficiency at both boundary and interior points. Also, the detailed comparisons between weighted Nadaraya-Watson and local linear are presented. All technical proofs are given in the Appendix.

2 Weighted Nadaraya-Watson Estimation

Assume that the process $\{(X_t, Y_t)\}_{t=-\infty}^\infty$ is stationary. Of interest is the nonparametric estimation of the regression function $m(x) = E(\phi(Y_t)\,|\,X_t = x)$, where $\phi(\cdot)$ is an arbitrary measurable function on the real line and it is assumed that $E|\phi(Y_t)| < \infty$. The introduction of $\phi(\cdot)$ allows us to estimate not only regression function but also conditional distribution ($\phi(Y_t) = I(Y_t \le y)$ for any fixed y) and conditional moment ($\phi(Y_t) = Y_t^r$, $r > 0$). Typically, the (locally) weighted least-squared approach is used often in the literature. For the given data $\{(X_t, Y_t)\}_{t=1}^n$, at the grid point x,

$$\sum_{t=1}^n \left[\phi(Y_t) - \sum_{j=0}^p \beta_j\,(X_t - x)^j\right]^2 c_{n,t}(x)\,K_h(x - X_t) \tag{2.1}$$

is minimized with respect to $\beta_0, \ldots, \beta_p$, where $K(\cdot)$ is a kernel function, $K_h(\cdot) = K(\cdot/h)/h$, $c_{n,t}(x)$ is a weight function, which might depend on the data $X_1, \ldots, X_n$, and $h = h_n$ is the smoothing parameter. The estimator of $m(x)$ is $\widehat{\beta}_0$. When $p = 0$ and $c_{n,t}(x) = 1$, the estimator of $m(x)$ becomes

$$\widehat{m}_{NW}(x) = \frac{\sum_{t=1}^n K_h(x - X_t)\, \phi(Y_t)}{\sum_{t=1}^n K_h(x - X_t)},$$

which is the well-known Nadaraya-Watson estimate and when $p = 1$ and $c_{n,t}(x) = 1$, it is the local linear estimate

$$\widehat{m}_{LL}(x) = \sum_{t=1}^n w_t\, \phi(Y_t), \quad w_t = \frac{K_h(x - X_t)\, \{S_{n,2} - (X_t - x)\, S_{n,1}\}}{S_{n,0}\, S_{n,2} - S_{n,1}^2},$$

where $S_{n,j} = \sum_{t=1}^n K_h(x - X_t)\, (X_t - x)^j$. It is easy to show (see Fan and Gijbels, 1996, p.63) that $\{w_t\}$ satisfy the following discrete moment conditions

$$\sum_{t=1}^n w_t = 1, \qquad \text{and} \qquad \sum_{t=1}^n (X_t - x)\, w_t = 0. \tag{2.2}$$

A direct consequence of this relation is that the finite sample bias is zero, but the Nadaraya-Watson estimator does not have this nice feature.

Let $w_t(x)$, for $1 \le t \le n$, denote the weight functions of the data $X_1, \ldots, X_n$ and the design point x with the property that each $w_t(x) \ge 0$, $\sum_{t=1}^n w_t(x) = 1$, and

$$\sum_{t=1}^n (X_t - x)\, w_t(x)\, K_h(x - X_t) = 0. \tag{2.3}$$

Motivated by the property of local linear estimator, the constraint (2.3) can be regarded as a discrete moment condition (2.2). Of course, $\{w_t(x)\}$ satisfying these conditions are not uniquely defined, and we specify them by maximizing $\prod_{t=1}^n w_t(x)$ subject to the constraints. The weighted version of Nadaraya-Watson estimator of conditional mean $m(x)$ of Y_t given $X_t = x$ is defined by

$$\widehat{m}(x) = \frac{\sum_{t=1}^n w_t(x)\, K_h(x - X_t)\, \phi(Y_t)}{\sum_{t=1}^n w_t(x)\, K_h(x - X_t)}, \tag{2.4}$$

which minimizes (2.1) with $c_{n,t}(x) = w_t(x)$ and $p = 0$. Clearly, unlike the local linear estimator, $\min_i\{\phi(Y_i)\} \le \widehat{m}(x) \le \max_i\{\phi(Y_i)\}$ for any x. Note that when $\phi(Y_t) = I(Y_t \le y)$, the weighted Nadaraya-Watson estimator $\widehat{m}(x)$ becomes the estimation of the conditional distribution of Y_t given $X_t = x$, which was studied by Hall, Wolff and Yao (1998) and Cai (1999) for time series data. We show in Theorem 1 (below) that $\widehat{m}(x)$ is first-order equivalent to a local linear estimator, and more importantly, in Theorem 2 (below) that $\widehat{m}(x)$ has automatic good behavior at boundaries.

The natural question arises regarding how to choose the weights. The idea of finding the best weight functions is from the empirical likelihood method (Owen, 1998). Namely, for fixed point x, by maximizing $\sum_{t=1}^n \log\{w_t(x)\}$ subject to the constraints $\sum_{t=1}^n w_t(x) = 1$ and (2.3) through the Lagrange multiplier, the $\{w_t(x)\}$ are simplified to

$$w_t(x) = n^{-1}\, \{1 + \lambda\, (X_t - x)\, K_h(x - X_t)\}^{-1}, \tag{2.5}$$

where λ, a function of data and x, is uniquely defined by (2.3), which ensues that $\sum_{t=1}^{n} w_t(x) = 1$. Equivalently, λ is chosen to maximize

$$L_n(\lambda) = \frac{1}{n\,h} \sum_{t=1}^{n} \log\left\{1 + \lambda\,(X_t - x)\,K_h(x - X_t)\right\}. \tag{2.6}$$

In implementation, the New-Raphson iteration scheme is recommended to find the root of equation $L_n'(\lambda) = 0$.

3 Asymptotic Theory

In this section, we establish the weak consistency with a rate and the asymptotic normality for the weighted Nadaraya-Watson estimator $\widehat{m}(x)$ under α-mixing. We first introduce some notation. Let $g(\cdot)$ denote the marginal density of X_t. Define $\mu_j = \int u^j\,K(u)\,du$, $\nu_j = \int u^j\,K^2(u)\,du$, and $\sigma^2(x) = \mathrm{Var}(Y_t \mid X_t = x)$. To obtain the asymptotic properties, the following regularity conditions are needed.

- **B1.** For fixed x, $g(x) > 0$, $g(\cdot)$ and $\sigma^2(\cdot)$ are continuous at x, and $m(\cdot)$ has continuous second order derivative in a neighborhood of x.
- **B2.** The kernel $K(\cdot)$ is a symmetric and bounded density with a bounded support $[-1,\,1]$.
- **B3.** $|g_{1,t}(y_1,\,y_2\,|x_1,\,x_2) \le M_1 < \infty$ for all $t \ge 2$, where $g_{1,t}(y_1,\,y_2\,|\,x_1,\,x_2)$ be the conditional density of Y_1 and Y_t given $X_1 = x_1$ and $X_t = x_2$.
- **B4.** Assume that $E\left(|\phi(Y_t)|^{\delta}\,|\,X_t = u\right) \le M_3 < \infty$ for some $\delta > 2$, in a neighborhood of x.
- **B5.** The mixing coefficient of the α-mixing process $\{(X_t,\,Y_t)\}_{t=-\infty}^{\infty}$ satisfies $\sum t^a\,[\alpha(t)]^{1-2/\delta} < \infty$ for some $a > 1 - 2/\delta$.
- **B6.** Assume that there exists a sequence of integers $s_n > 0$ such that $s_n \to \infty$, $s_n = o((n\,h_n)^{1/2})$, and $(n/h_n)^{1/2}\alpha(s_n) \to 0$, as $n \to \infty$.
- **B7.** There exists $\delta^* > \delta$ such that $E\left(|\phi(Y_t)|^{\delta^*}\,|\,X_t = u\right) \le M_4 < \infty$ in a neighborhood of x, $\alpha(t) = O(t^{-\theta^*})$, where $\theta^* \ge \delta^*\,\delta/\{2(\delta^* - \delta)\}$, and $n^{1/2-\delta/4}\,h^{\delta/\delta^*-1/2-\delta/4} = O(1)$.

Note that the conditions B1-B6 are used commonly in the literature of time series data, see Masry and Fan (1997). Since the common technique – truncation approach for time series data is not applicable to our setting, the purpose of the condition B7 is to use the moment inequality.

Theorem 1. Suppose that Conditions B1-B5 hold. Then, as $n \to \infty$,

$$\widehat{m}(x) - m(x) = \frac{1}{2}\,h^2\,\mu_2\,m''(x) + o_p(h^2) + O_p\left((n\,h)^{-1/2}\right). \tag{3.1}$$

In addition, if Conditions B6 and B7 are satisfied, then

$$\sqrt{n\,h}\,\left[\widehat{m}(x) - m(x) - \frac{1}{2}\,h^2\,\mu_2\,m''(x) + o_p(h^2)\right] \xrightarrow{D} N\left(0,\,\nu_0\,\sigma^2(x)/g(x)\right). \tag{3.2}$$

It may be seen from the theorem that first, the weighted Nadaraya-Watson estimator $\widehat{m}(x) \to m(x)$ in probability with a rate, which, of course, implies that $\widehat{m}(x)$ is consistent. Also, it is easy to see that the asymptotic mean squared error (MSE) is

$$\mathrm{MSE} = \frac{1}{4}\, h^4\, \mu_2^2\, \{m''(x)\}^2 + \frac{\nu_0}{n\,h}\, \frac{\sigma^2(x)}{g(x)}.$$

This implies that to the first order, the weighted Nadaraya-Watson method enjoys the same convergence rates as those of local linear procedure. Some algebra yields the optimal bandwidth minimizing the MSE,

$$h_{opt} = \left(\frac{\nu_0\, \sigma^2(x)}{\mu_2^2\, g(x)\, \{m''(x)\}^2} \right)^{1/5} n^{-1/5},$$

and the optimal asymptotic MSE

$$\mathrm{MSE}_{opt} = \frac{5}{4} \left(\frac{\sqrt{\mu_2\, |m''(x)|}\, \nu_0\, \sigma^2(x)}{n\, g(x)} \right)^{4/5}. \tag{3.3}$$

This, together with Theorem 3.5 in Fan and Gijbels (1996, p.85), implies that the asymptotic minimax efficiency (AME), defined in Fan and Gijbels (1996, pp.85), for the weighted Nadaraya-Watson estimator is determined by

$$0.268\, (\mu_2\, \nu_0^2)^{-1/2},$$

which is nearly 100% with the Epanechnikov kernel and the optimal bandwidth h_{opt}, in a minimax sense, among estimators that are linear in the response variable; see Fan and Gijbels (1996, pp.84–94) for the detailed discussions.

As for the boundary behavior of the weighted Nadaraya-Watson estimator, we offer Theorem 2 below. Without loss of generality, we consider the left boundary point $x = c\, h$, $0 < c < 1$. Following Fan, Hu and Troung (1994), we take $K(\cdot)$ to have support $[-1,\ 1]$ and $g(\cdot)$ to have support $[0,\ 1]$. First, we introduce the following notation. Let

$$L_c(\lambda) = \int_{-1}^{c} \frac{u\, K(u)}{1 - \lambda\, u\, K(u)}\, du, \tag{3.4}$$

and λ_c be the root of equation $L_c(\lambda) = 0$, namely, $L_c(\lambda_c) = 0$. With the Epanechnikov kernel, Figure 1(a) plots λ_c versus c.

Theorem 2. Suppose that the conditions of Theorem 1 hold. Then

$$\sqrt{n\, h}\, \left[\widehat{m}(c\, h) - m(c\, h) - B_c(y) + o_p(h^2) \right] \ \xrightarrow{D}\ N\left(0,\ \sigma_c^2(y)\right), \tag{3.5}$$

where with

$$\beta_0(c) = \int_{-1}^{c} \frac{u^2\, K(u)}{1 - \lambda_c\, u\, K(u)}\, du, \quad \text{and} \quad \beta_j(c) = \int_{-1}^{c} \frac{K^j(u)}{\{1 - \lambda_c\, u\, K(u)\}^j}\, du\ (1 \le j \le 2),$$

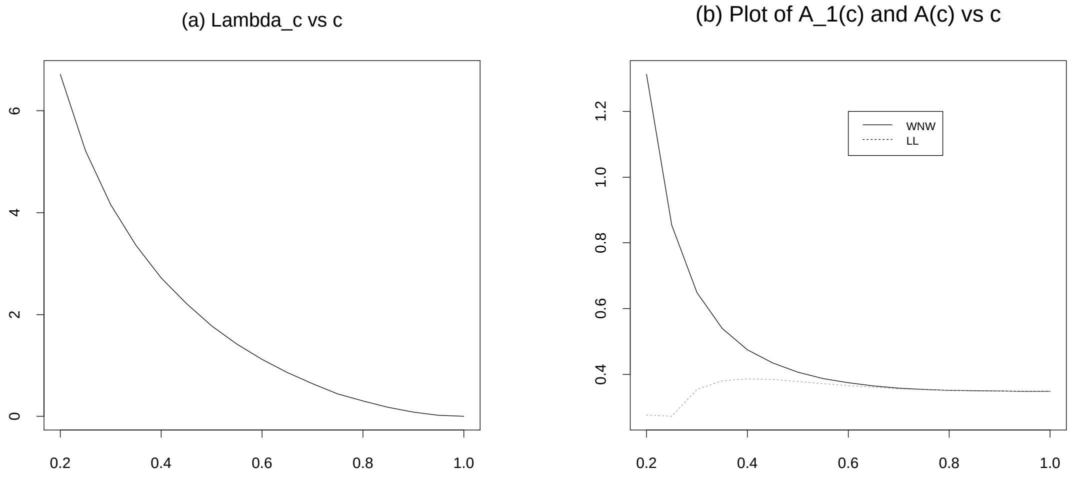

Figure 1: *(a) Plot of λ_c versus c. (b) Plot of $A_1(c)$ for WNW (solid line) and $A(c)$ for local linear (dot line).*

the bias and variance are given, respectively, by

$$B_c(y) = \frac{h^2\,\beta_0(c)\,m''(0)}{2\,\beta_1(c)}, \quad \text{and} \quad \sigma_c^2(y) = \frac{\beta_2(c)\,\sigma^2(0)}{\beta_1^2(c)\,g(0)}.$$

These theorems reflect two of the major advantages of the weighted Nadaraya-Watson estimator: (a) no dependence of the asymptotic bias on the design density $g(\cdot)$, and indeed its dependence on the simple regression function curvature $m''(\cdot)$; and (b) automatic good behavior at boundaries, at least with regard to orders of magnitude, without the need of boundary correction. Also, we remark that a similar result (3.5) holds for the right boundary point $x = 1 - c\,h$. If the point 0 were an interior point, then the expression (3.5) would hold with $c = 1$ and $\lambda_c = 0$. Furthermore, note that since the proofs of theorems 1 and 2 are similar, we present only the detailed proof of Theorem 1 and give the brief proof of Theorem 2 in the Appendix.

Similar to (3.3), the optimal bandwidth and MSE at boundary are given, respectively, by

$$h_{opt}^0 = \left(\frac{v_1(c)}{b_1^2(c)}\,\frac{\sigma^2(0)}{g(0)\,\{m''(0)\}^2}\right)^{1/5} n^{-1/5}, \quad \text{and} \quad \text{MSE}_{opt}^0 = \frac{5}{4}\,A_1(c)\,\left(\frac{\sqrt{|m''(0)|}\,\sigma^2(0)}{n\,g(0)}\right)^{4/5}, \tag{3.6}$$

where

$$b_1^2(c) = \frac{\beta_0^2(c)}{\beta_1^2(c)}, \qquad v_1(c) = \frac{\beta_2(c)}{\beta_1^2(c)}, \quad \text{and} \quad A_1(c) = \left\{b_1^2(c)\right\}^{1/5}\,v_1^{4/5}(c). \tag{3.7}$$

Clearly,

$$\lim_{c\to 1} b_1^2(c) = \mu_2^2 \qquad \text{and} \qquad \lim_{c\to 1} v_1(c) = \nu_0.$$

By a comparison of (3.6) with Theorem 3.3 in Fan and Gijbels (1996, p.72), it is easy to see that both the weighted Nadaraya-Watson and local linear methods have the same order of convergent rate, but there is a constant factor. To see how the factor changes with c, we define

$$A(c) = \left\{b^2(c)\right\}^{1/5}\,v^{4/5}(c)$$

for the local linear estimator, where with $\mu_{j,c} = \int_{-1}^{c} u^j\, K(u) du$ and $\nu_{j,c} = \int_{-1}^{c} u^j\, K^j(u)\, du$,

$$b^2(c) = \left(\frac{\mu_{2,c}^2 - \mu_{1,c}\, \mu_{3,c}}{\mu_{2,c}\, \mu_{0,c} - \mu_{1,c}^2} \right)^2, \qquad \text{and} \qquad v(c) = \frac{\mu_{2,c}^2\, \nu_{0,c} - 2\, \mu_{2,c} \mu_{1,c}\, \nu_{1,c} + \mu_{1,c}^2\, \nu_{2,c}}{\left(\mu_{2,c}\, \mu_{0,c} - \mu_{1,c}^2 \right)^2};$$

see Fan and Gijbels (1996, p.74) for details. We plot $A_1(c)$ and $A(c)$ versus c in Figure 1(b) with the Epanechnikov kernel, which shows that when $c \geq 0.5$, the two methods almost have the same performance, however, the local linear estimator performs better when $c < 0.5$. Finally, it follows from (3.6) and Theorem 3.8 in Fan and Gijbels (1996, p.89) that the AME for $\widehat{m}(c\, h)$ is bounded from above and below by

$$1.51786 \leq \sqrt{\frac{\beta_0(c)\, \beta_2^2(c)}{\beta_1^5(c)}}\ \text{AME} \leq 2.07844.$$

Appendix: Proofs

Note that we still use the same notation as in Sections 2 and 3. Let $\varepsilon_t = \phi(Y_t) - m(X_t)$, and

$$b_t(x) = \left[1 - \frac{h\, \mu_2\, g'(x)}{2\, \nu_2\, g(x)}\, (X_t - x)\, K_h(x - X_t) \right]^{-1}. \tag{A.1}$$

Set $\zeta_t = \sqrt{h}\, b_t(x)\, \varepsilon_t\, K_h(x - X_t)$, and

$$J_1 = \sqrt{\frac{h}{n}} \sum_{t=1}^{n} b_t(x)\, \varepsilon_t\, K_h(x - X_t) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \zeta_t. \tag{A.2}$$

Let C denote a positive constant which might take a different value at the different place.

Lemma 1. Under the assumptions B1-B5, we have

$$\text{Var}(J_1) \ \rightarrow\ \nu_0\, \sigma^2(x)\, g(x) \equiv \theta^2(x).$$

Proof. It is easy to see that $E[\zeta_t] = 0$ since $E[\varepsilon_t \,|\, X_t] = 0$, and

$$\text{Var}(J_1) = E\left[\zeta_t^2 \right] + \sum_{t=2}^{n} \left(1 - \frac{t-1}{n} \right) \text{Cov}(\zeta_1,\, \zeta_t). \tag{A.3}$$

A straightforward manipulation yields

$$E\left[\zeta_t^2 \right] = h\, E\left[b_t^2(x)\, \varepsilon_t^2\, K_h^2(x - X_t) \right] = \theta^2(x) + o(1).$$

Choose the positive integers $\{d_n\}$ such that $d_n\, h_n \rightarrow 0$, and decompose the second term on the right hand side of (A.3) into two terms as follows

$$\sum_{t=2}^{n} = \sum_{t=2}^{d_n} + \sum_{t=d_n+1}^{n} \equiv J_{11} + J_{12}.$$

For J_{11}, by conditioning on (X_1, X_t) and using Condition B3, it follows that $|\mathrm{Cov}(\zeta_1, \zeta_t)| \le C\,h$, so that $J_{11} = O(d_n\,h) = o(1)$. For J_{12}, by applying Corollary A.2 (the Davydov's inequality) in Hall and Heyde (1980), one has

$$|\mathrm{Cov}(\zeta_1, \zeta_t)| \le C\,[\alpha(t-1)]^{1-2/\delta} \left(E|\zeta_1|^\delta\right)^{2/\delta}. \tag{A.4}$$

By conditioning on X_1 and using Condition B4, we obtain

$$E\left[|\zeta_1|^\delta\right] \le C\,h^{1-\delta/2}, \tag{A.5}$$

which, in conjunction with (A.4) and Condition B5, implies that

$$J_{12} \le C\,h^{2/\delta-1} \sum_{t \ge d_n} [\alpha(t)]^{1-2/\delta} \le C \sum_{t \ge d_n} t^a\,[\alpha(t)]^{1-2/\delta} = o(1)$$

by choosing d_n such that $h^{1-2/\delta}\,d_n = O(1)$, so that the requirement that $d_n\,h_n \to 0$ is satisfied. This completes the proof of the lemma. □

Lemma 2. Under the assumptions B1-B5, we have

$$\lambda = -\frac{h\,\mu_2\,g'(x)}{2\,\nu_2\,g(x)}\,\{1 + o_p(1)\},$$

so that $w_t(x) = b_t(x)\,\{1 + o_p(1)\}$.

Proof. Define, for $j \ge 1$,

$$A_j = \frac{1}{n}\sum_{t=1}^{n}(X_t - x)^j\,K_h^j(x - X_t).$$

Using the same arguments as those employed in the proof of Lemma 1, we have

$$A_1 = -\frac{1}{2}\mu_2\,h^2\,g'(x) + o_p(h^2), \quad A_2 = h\,\mu_2\,g(x) + o_p(h^2), \quad \text{and} \quad A_3 = O_p(h^2). \tag{A.6}$$

By (6.4) in Chen and Hall (1993),

$$|\lambda| \le \frac{|A_1|}{A_2 - C_1\,|A_1|} = O_p(h).$$

By a Taylor expansion,

$$0 = A_1 - \lambda\,A_2 + \lambda^2\,A_3 - \lambda^3\,A_4 + \cdots,$$

so that

$$\lambda = \frac{A_1}{A_2} + \lambda^2\,\frac{A_3}{A_2} - \lambda^3\,\frac{A_4}{A_2} + \cdots.$$

Therefore, substituting (A.6) into the above equation, we prove the lemma. □

Proof of Theorem 1. It follows from Lemma 2 that

$$\begin{aligned}\widehat{m}(x) - m(x) &= \frac{\sum_{t=1}^{n}[\phi(Y_t) - m(x)]\,w_t(x)\,K_h(x - X_t)}{\sum_{t=1}^{n} w_t(x)\,K_h(x - X_t)} \\ &\equiv \left\{(n\,h)^{-1/2}\,J_1 + J_2\right\}\,J_3^{-1}\,\{1 + o_p(1)\},\end{aligned} \tag{A.7}$$

where

$$J_2 = \frac{1}{n}\sum_{t=1}^{n}[m(X_t) - m(x)]\, w_t(x)\, K_h(x - X_t), \quad \text{and} \quad J_3 = \frac{1}{n}\sum_{t=1}^{n} b_t(x)\, K_h(x - X_t).$$

By Condition B1 and (2.3) as well as the Taylor expansion, we have

$$J_2 = \frac{1}{2\,n}\sum_{t=1}^{n} m''(x)\,(X_t - x)^2\, b_t(x)\, K_h(x - X_t) + o_p(h^2) = \frac{h^2}{2}\,\mu_2\, m''(x)\, g(x) + o_p(h^2)$$

by following the same line as in the proof of Lemma 1. Similarly,

$$J_3 = g(x) + o_p(1). \tag{A.8}$$

Therefore,

$$\sqrt{n\,h}\left[\widehat{m}(x) - m(x) - \frac{h^2}{2}\,\mu_2\, m''(x) + o_p(h^2)\right] = g^{-1}(x)\, J_1 + o_p(1). \tag{A.9}$$

This, in conjunction with Lemma 1, implies (3.1). To prove (3.2), it suffices to establish the asymptotic normality of J_1 by (A.9). To this end, we employ the Doob's small-block and large-block technique. Namely, partition $\{1, \ldots, n\}$ into $2\,q_n + 1$ subsets with large-block of size $r = r_n$ and small-block of size $s = s_n$. Set

$$q = q_n = \left\lfloor \frac{n}{r_n + s_n} \right\rfloor. \tag{A.10}$$

Define the random variables, for $0 \le j \le q - 1$,

$$\eta_j = \sum_{i=j(r+s)}^{j(r+s)+r-1} \zeta_i, \qquad \xi_j = \sum_{i=j(r+s)+r}^{(j+1)(r+s)} \zeta_i, \qquad \text{and} \qquad \eta_q = \sum_{i=q(r+s)}^{n-1} \zeta_i.$$

Then,

$$J_1 = \frac{1}{\sqrt{n}}\left\{\sum_{j=0}^{q-1}\eta_j + \sum_{j=0}^{q-1}\xi_j + \eta_q\right\} \equiv \frac{1}{\sqrt{n}}\left\{Q_{n,1} + Q_{n,2} + Q_{n,3}\right\}.$$

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

$$\frac{1}{n}\, E\,[Q_{n,2}]^2 \to 0, \qquad \frac{1}{n}\, E\,[Q_{n,3}]^2 \to 0, \tag{A.11}$$

$$\left| E\,[\exp(i\,t\,Q_{n,1})] - \prod_{j=0}^{q-1} E\,[\exp(i\,t\,\eta_j)] \right| \to 0, \tag{A.12}$$

$$\frac{1}{n}\sum_{j=0}^{q-1} E\left(\eta_j^2\right) \to \theta^2(x), \tag{A.13}$$

where $\theta^2(x)$ is defined in Lemma 1, and

$$\frac{1}{n}\sum_{j=0}^{q-1} E\left[\eta_j^2 I\left\{|\eta_j| \ge \varepsilon\,\theta(x)\,\sqrt{n}\right\}\right] \to 0 \tag{A.14}$$

for every $\varepsilon > 0$. (A.11) implies that $Q_{n,2}$ and $Q_{n,3}$ are asymptotically negligible in probability; (A.12) shows that the summands η_j in $Q_{n,1}$ are asymptotically independent; and (A.13) and (A.14) are the standard Lindeberg-Feller conditions for asymptotic normality of $Q_{n,1}$ for the independent setup.

Let us first establish (A.11). To this effect, Condition B6 implies that there is a sequence of positive numbers $\gamma_n \to \infty$ such that

$$\gamma_n\, s_n / \sqrt{n\,h} \to 0, \qquad \text{and} \qquad \gamma_n\, (n/h)^{1/2} \alpha(s_n) \to 0. \tag{A.15}$$

Define the large-block size r_n by $r_n = \lfloor (n\,h_n)^{1/2}/\gamma_n \rfloor$ and the small-block size s_n. Then, as $n \to \infty$,

$$s_n/r_n \to 0, \qquad r_n/n \to 0, \qquad r_n\,(n\,h)^{-1/2} \to 0, \quad \text{and} \quad (n/r_n)\,\alpha(s_n) \to 0. \tag{A.16}$$

Observe that

$$E\,[Q_{n,2}]^2 = \sum_{j=0}^{q-1} \mathrm{Var}(\xi_j) + 2 \sum_{0 \le i < j \le q-1} \mathrm{Cov}(\xi_i,\, \xi_j) \equiv F_1 + F_2. \tag{A.17}$$

It follows from stationarity and Lemma 1 that

$$F_1 = q_n\, \mathrm{Var}(\xi_1) = q_n\, \mathrm{Var}\left(\sum_{j=1}^{s_n} \zeta_j \right) = q_n\, s_n\, [\theta^2(x) + o(1)]. \tag{A.18}$$

Next consider the second term F_2 on the right-hand side of (A.17). Let $r_j^* = j(r_n + s_n)$, then $r_j^* - r_i^* \ge r_n$ for all $j > i$, we therefore have

$$\begin{aligned} |F_2| \quad &\le \quad 2 \sum_{0 \le i < j \le q-1} \sum_{j_1=1}^{s_n} \sum_{j_2=1}^{s_n} |\mathrm{Cov}(\zeta_{r_i^*+r_n+j_1},\, \zeta_{r_j^*+r_n+j_2})| \\ &\le \quad 2 \sum_{j_1=1}^{n-r_n} \sum_{j_2=j_1+r_n}^{n} |\mathrm{Cov}(\zeta_{j_1},\, \zeta_{j_2})|. \end{aligned}$$

By stationarity and Lemma 1, one obtains

$$|F_2| \le 2n \sum_{j=r_n+1}^{n} |\mathrm{Cov}(\zeta_1,\, \zeta_j)| = o(n). \tag{A.19}$$

Hence, by (A.16)-(A.19), we have

$$\frac{1}{n}\, E[Q_{n,2}]^2 = O\left(q_n\ s_n\ n^{-1} \right) + o(1) = o(1). \tag{A.20}$$

It follows from stationarity, (A.16) and Lemma 1 that

$$\mathrm{Var}\,[Q_{n,3}] = \mathrm{Var}\left(\sum_{j=1}^{n-q_n(r_n+s_n)} \zeta_j \right) = O(n - q_n(r_n + s_n)) = o(n). \tag{A.21}$$

Combining (A.16), (A.20) and (A.21), we establish (A.11). As for (A.13), by stationarity, (A.16) and Lemma 1, it is easily seen that

$$\frac{1}{n} \sum_{j=0}^{q_n-1} E\left(\eta_j^2 \right) = \frac{q_n}{n} E\left(\eta_1^2 \right) = \frac{q_n\, r_n}{n} \cdot \frac{1}{r_n} \mathrm{Var}\left(\sum_{j=1}^{r_n} \zeta_j \right) \to \theta^2(x).$$

In order to establish (A.12), we make use of Lemma 1.1 in Volkonskii and Rozanov (1959) to obtain

$$\left| E\left[\exp(i\,t\,Q_{n,1})\right] - \prod_{j=0}^{q_n-1} E\left[\exp(i\,t\,\eta_j)\right] \right| \le 16\,(n/r_n)\,\alpha(s_n)$$

tending to zero by (A.16).

It remains to establish (A.14). To this end, we employ Theorem 4.1 in Shao and Yu (1996) and Condition B7 to obtain,

$$E\left[\eta_1^2\, I\left\{|\eta_1| \ge \varepsilon\,\theta(x)\,\sqrt{n}\right\}\right] \le C\,n^{-\delta/2}\,E\left(|\eta_1|^{\delta}\right) \le C\,n^{-\delta/2}\,r_n^{\delta/2}\,\left\{E\left(|\zeta_1|^{\delta^*}\right)\right\}^{\delta/\delta^*}. \qquad \text{(A.22)}$$

As in (A.5),

$$E\left(|\zeta_1|^{\delta^*}\right) \le C\,h_n^{1-\delta^*/2},$$

which, together with (A.22), gives

$$E\left[\eta_1^2\, I\left\{|\eta_1| \ge \varepsilon\,\theta(x)\,\sqrt{n}\right\}\right] \le C\,n^{1-\delta/2}\,r_n^{\delta/2}\,h_n^{\delta(1/\delta^*-1/2)}.$$

Thus, by (A.10) and the definition of r_n, and using Condition B7, we obtain

$$\frac{1}{n}\sum_{j=0}^{q-1} E\left[\eta_j^2 I\left\{|\eta_j| \ge \varepsilon\,\theta(x)\,\sqrt{n}\right\}\right] \le C\,\gamma_n^{1-\delta/2}\,n^{1/2-\delta/4}\,h_n^{\delta/\delta^*-1/2-\delta/4} \to 0.$$

This completes the proof of the theorem. □

In order to prove Theorem 2, we need the following lemma.

Lemma 3. Under the assumptions B1-B5, we have

$$p_t(c\,h) = n^{-1}\,b_t^c(c\,h)\,\{1 + o_p(1)\},$$

where

$$b_t^c(x) = \left[1 + \lambda_c\,(X_t - x)\,K_h(x - X_t)\right]^{-1}.$$

Proof. Let $\widehat{\lambda} = \operatorname{argmax}_{\lambda} L_n(\lambda)$ so that $L_n'\left(\widehat{\lambda}\right) = 0$, where $L_n(\cdot)$ is defined in (2.6). It suffices to show that $\widehat{\lambda} \to \lambda_c$ in probability. To this end, denote by S_ε the interval $\lambda_c \pm \varepsilon$. We will show that for any sufficiently small ε, the probability

$$\sup_{\lambda \in S_\varepsilon} L_n(\lambda) \le L_n(\lambda_c)$$

tends to one. By the Taylor expansion,

$$L_n(\lambda) - L_n(\lambda_c) = L_n'(\lambda_c)\,(\lambda - \lambda_c) + \frac{1}{2}\,L_n''(\lambda_c)\,(\lambda - \lambda_c)^2 + \frac{1}{6}\,L_n'''(\lambda^*)\,(\lambda - \lambda_c)^3$$

with λ^* lying between λ and λ_c. It is easy to show that

$$L_n'(\lambda_c) = o_p(1), \quad L_n''(\lambda_c) = -\beta_3(c)\, g(0+) + o_p(1), \quad \text{and} \quad L_n'''(\lambda^*) = O_p(1),$$

where

$$\beta_3(c) = \int_{-1}^{c} \frac{u^2\, K^2(u)}{[1 - \lambda_c\, u\, K(u)]^2}\, du.$$

This concludes with probability tending to one that when ε is small enough, for all $\lambda \in S_\varepsilon$,

$$L_n(\lambda) - L_n(\lambda_c) \le 0,$$

which completes the proof of the lemma. □

Proof of Theorem 2. By replacing $b_t(x)$ in ζ_t by $b_t^c(c\,h)$ and following the same arguments as those used in the proof of Theorem 1, we can prove the theorem via Lemma 3. □

References

Auestad, B. and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika*, **77**, 669–687.

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes.* Springer-Verlag, New York.

Cai, Z. (1999). Regression quantiles for time series data. Submitted for publication.

Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, **88**, 298–308.

Chen, S.X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, **21**, 1166–1181.

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

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fan, J., Hu, T.-C., and Troung, Y.K. (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics*, **21**, 433–446.

Gorodetskii, V.V. (1977). On the strong mixing property for linear sequences. *Theory of Probability and Its Applications*, **22**, 411–413.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

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

Hall, P. and Presnell, B. (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society, Series B*, **61**, 143-158.

Hall, P., Wolff, R.C.L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, 154–163.

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

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hurvich, C.M., Simonoff, J.S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, **60**, 271-293.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *The Scandinavian Journal of Statistics* **24**, 165–179.

Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory*, **11**, 258–289.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**, 214–252.

Müller, H.-G. (1988). *Nonparametric Analysis of Longitudinal Data.* Springer-Verlag, Berlin.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Robinson, P.M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**, 185–297.

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

Shao, Q. and Yu, H. (1996). Weak convergence for weighted empirical processes of dependent sequences. *The Annals of Probability* **24**, 2098–2127.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer-Verlag, New York.

Troung, Y.K. and Stone, C.J. (1992). Nonparametric function estimation involving time series. *The Annals of Statistics*, **20**, 77-97.

Volkonskii, V.A. and Rozanov, Yu.A. (1959). Some limit theorems for random functions. I. *Theory of Probability and Its Applications*, **4**, 178–197.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Withers, C.S. (1981). Conditions for linear processes to be strong mixing. *Zeitschrift fur Wahrscheinlichkeitstheorie verwandte Gebiete*, **57**, 477–480.