Nonparametric Regression With Nearly Integrated Regressors Under Long Run Dependence*

Zongwu CAI\textsuperscript{a,b}, Bing-Yi JING\textsuperscript{c}, Xin-Bing KONG\textsuperscript{c}, Zhi LIU\textsuperscript{c}

\textsuperscript{a}Department of Mathematics & Statistics and Department of Economics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA.
\textsuperscript{b}The Wang Yanan Institute for Studies in Economics, Xiamen University, Fujian, China.
\textsuperscript{c}Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China.

E-mails: zcai@uncc.edu, majing@ust.hk, kongxin@ust.hk, and liuzhi@ust.hk.

November 24, 2010

We study nonparametric estimation of regression function with nonstationary (integrated or nearly integrated) covariates and the error series of the regressor process following a fractional ARIMA model. A local linear estimation method is developed to estimate the unknown regression function. The asymptotic results of the resulting estimator at both interior points and boundaries are obtained. The asymptotic distribution is mixed normal, associated with the local time of an Ornstein-Uhlenbeck (O-U) fractional Brownian motion. Furthermore, we study the Nadaraya-Watson estimator and examine its asymptotic results. As a result, it shares exactly the same asymptotic results as those for the local linear estimator for the zero energy situation. But for the non-zero energy case, the local linear estimator is superior over the Nadaraya-Watson estimator in terms of optimal convergence rate. Moreover, a comparison of our results with the conventional results for stationary covariates is presented. Finally, a Monte Carlo simulation is conducted to illustrate the finite sample performance of the proposed estimator.

Keywords: Asymptotics; kernel smoothing; local time of an Ornstein-Uhlenbeck fractional Brownian motion; nonlinearity; nonstationary covariates; unit root.

\textsuperscript{*}Cai’s research was supported, in part, by the National Science Foundation of China grants \#70871003 and \#70971113, and funds provided by the Chang Jiang Scholars Program from Chinese Ministry of Education, the Minjiang Scholars Program from Fujian Province, China and Xiamen University. Jing’s research is supported, in part, by Hong Kong RGC grants HKUST6011/07P and HKUST6015/08P.
1 Introduction

Nonparametric estimation techniques have become cornerstone research topics in statistics and econometrics for the last three decades due to their numerous advantages relative to parametric techniques such as more flexibility and robustness to functional form misspecification, and have been embraced by applied researchers in many fields; see the books by Härdle (1990), Fan and Gijbels (1996), Fan and Yao (2003), and Li and Racine (2007), and the survey papers by Cai and Hong (2009) and Cai, Gu and Li (2009) for nonparametric methods with applications in finance and economics. Asymptotic theory underlying various nonparametric estimators and test statistics for many commonly used models have been well established for independent and identically distributed (iid) data and some weak and strong dependent stationary time series. The only nonparametric asymptotic analysis when covariates are integrated (unit root, denoted by I(1)) or nearly integrated (nearly unit root or local-to-unity, denoted by NI(1)) time series that we are aware of includes the work by Phillips and Park (1998), Park and Hahn (1999), Chang and Park (2003), Juhl (2005), Cai, Li and Park (2009), Wang and Phillips (2009a, 2009b, 2009c) and Xiao (2009). It is worth pointing out that for local-to-unity or nearly integrated regressors, the main focus in the literature is on a linear regression model; see, for example, Elliott and Stock (1994), Cavanagh, Elliott, and Stock (1995), Torous, Valkanov, and Yan (2004), Campbell and Yogo (2006), Polk, Thompson, and Vuolteenaho (2006), Rossi (2007), Cai and Wang (2009) among others. However, a nonparametric regression model with nearly integrated variables has not been formally addressed.

In this paper, for the observed data \( \{(y_t, x_t)\} \) for \( t = 1, \ldots, n \), we study a nonparametric regression function with nonstationary covariate as follows,

\[
y_t = f(x_t) + u_t, \quad 1 \leq t \leq n, \tag{1.1}
\]

where \( \{u_t\} \) is stationary (denoted by I(0)) and \( f(\cdot) \) is an unknown regression function. Here, \( x_t \) is an integrated or nearly integrated process satisfying

\[
x_t = \beta x_{t-1} + \epsilon_t, \quad 1 \leq t \leq n,
\]

where \( \beta = 1 + c/n \) for \( c \leq 0 \), and \( \{\epsilon_t\} \) is a stationary sequence with a possible long run
dependence as a fractional integrated autoregressive moving average (FARIMA) process

\[(1 - B)^d \epsilon_t = \eta_t \equiv \sum_{j=0}^{\infty} \psi_j \xi_{t-j}, \quad (1.2)\]

where \(B\) is the backward operator, \(\psi_j, j > 0\) are some constants, \(\xi_j, j > 0\) are independent and identically distributed (i.i.d.) random variables (r.v.s) with zero mean and finite variance, and \(|d| < 1/2\). The fractional power \((1 - B)^d\) is defined as \(\sum_{k=0}^{\infty} c_{k,d} x^k\), where \(c_{k,d} = \Gamma(-d + k)/\Gamma(-d)\Gamma(k + 1)\) and \(\Gamma(\cdot)\) denotes the \(\Gamma\)-function. It is easy to see that \(c_{k,d} \sim k^{-d-1}/\Gamma(-d)\) as \(k \to \infty\). Clearly, when \(d = 0\), \(\{\epsilon_t\}\) in (1.2) becomes a linear process and it is a stationary and \(\alpha\)-mixing time series if it satisfies a mild condition such as Assumption 1 (later).

Indeed, model (1.1) is not new in the literature but its asymptotics is novel when \(x_t\) is persistent and nonstationary. For example, if \(x_t\) is stationary, model (1.1) has been studied extensively in the literature; see Härdle (1990), Fan and Gijbels (1996), Fan and Yao (2003), and Li and Racine (2007) for details, while it was investigated by Karlsen and Tjøstheim (2001) for \(x_t\) being null recurrent time series and Karlsen, Myklebust and Tjøstheim (2007) for the \(\phi\)-irreducible Markov chain time series. A functional coefficient type model and nonlinear cointegration were investigated by Cai, Li and Park (2009) and Xiao (2009) for both I(0) and I(1) covariates and by Cai and Wang (2009) for NI(1) covariates, while Wang and Phillips (2009a, 2009b) considered a nonparametric regression and structure regression when \(x_t\) is I(1). For simplicity of notation, we consider only one-dimensional case since extension to multivariate \(x_t\) involves fundamentally no new ideas but complicated notations.

Model (1.1) might have a great potential in many applications. For example, in macroeconomics, a particular parametric form of (1.1) can be used for forecasting inflation rate based on some persistent and nonstationary covariates such as velocity; see Bachmeier, Leelahanon and Li (2006), which showed that the velocity is an I(1) process. Also, it can be employed for testing the predictability and stability of stock returns using various lagged financial variables, such as the dividend yield, term and default premia, the dividend-price ratio, the earning-price ratio, the book-to-market ratio, and interest rates; see Elliott and Stock (1994), Cavanagh, Elliott, and Stock (1995), Torous, Valkanov, and Yan (2004), Campbell and Yogo
(2006), Polk, Thompson, and Vuolteenho (2006), Rossi (2007), and Cai and Wang (2009), among others. In fact, Campbell and Yogo (2006) showed that the 95% confidence intervals for the coefficient $\beta$ in an AR(1) are $[0.957, 1.007]$ and $[0.939, 1.000]$ for the log dividend-price ratio and the log earnings-price ratio, respectively; see Panel A in Table 4 of Campbell and Yogo (2006). As advocated by Campbell and Yogo (2006), Bachmeier, Leelahanon and Li (2006), and Cai, Li and Park (2009), the predictive power of using integrated or nearly integrated (highly persistent) covariates in a regression model can be improved significantly due to less noise. Therefore, our motivation is based on the aforementioned real examples.

The main purpose of the present paper is to estimate nonparametric regression $f(\cdot)$ by using local linear (polynomial) and local constant (Nadaraya-Watson) fitting schemes. For simplicity, the main results can be summarized as follows. First, the optimal rate of convergence is $n^{(1-2d)/5}$ with $|d| < 1/2$ slower than the usual rate $n^{2/5}$ for stationary case, see Fan and Yao (2003), and slower than the rate $n^{1/5}$ for the case where $\{\epsilon_t\}$ is short-dependence ($d = 0$); see Cai, Li and Park (2009). Consequently, the order of the asymptotic mean-squared error (AMSE) is $n^{(2-4d)/5}$ rather than the standard rate $n^{-4/5}$. The intuitive explanation to this phenomenon is that an NI(1) or I(1) time series (under long run dependence) takes longer to revisit levels in its range. Second, the asymptotic bias term, similar to the stationary case, is independent of the distributions of regressors and is only due to the linear approximation, which is typical for a local linear fitting scheme. Third, the limiting distribution is mixed-normal (conditional normal) in that the asymptotic variance depends inversely on the local time of an O-U fractional Brownian motion in which the nearly unit root series can be embedded. Furthermore, the nearly integrated covariate requires the larger bandwidths. Indeed, the optimal (in the AMSE sense) bandwidth is $O_p(n^{-(1-2d)/10})$ implying a larger optimal bandwidth than in conventional kernel regressions with stationary regressors where the optimal bandwidth is known to be $O(n^{-1/5})$. Clearly, the use of conventional bandwidth has the theoretical potential of under-smoothing in the presence of NI(1) or I(1) covariates. Finally, it is very interesting that both local linear and local constant estimators share exactly same asymptotic properties at both interior and boundary points for the zero energy case. However, for the non-zero energy case, the local linear estimator is superior over the Nadaraya-Watson estimator in terms of optimal convergence rate.
The remainder of the paper is organized as follows. Section 2 is devoted to presenting nonparametric kernel estimators of $f(\cdot)$ using both local linear and Nadaraya-Watson (local constant) estimation methods and their asymptotic behaviors for both interior and boundary points, together with assumptions and remarks on comparisons of our results with conventional findings. In Section 3, we illustrate the finite sample performance of the estimators with a Monte Carlo experiment. A concluding remark is presented in Section 4. Finally, the proofs of the main results of the paper are relegated to Section 5.

2 Statistical Properties

2.1 Local Linear Estimation

We estimate $f(\cdot)$ using local linear fitting from observations $\{ (y_t, x_t) \}_{t=1}^n$. Our motivation of using local linear fitting is its high statistical efficiency in an asymptotic minimax sense, design adaptation and automatic correction for edge effects, as discussed in Fan and Gijbels (1996). Although a general local polynomial technique is applicable as well, it is well known that the local linear fitting will suffice for many applications; see Fan and Gijbels (1996) for a very comprehensive discussion, and that the theory developed for the local linear estimator continues to hold for the local polynomial estimator with only slight modification.

Another virtue of using local polynomials is that both the unknown functions as well as their derivatives can be estimated simultaneously. For simplicity, we only focus on local linear estimation and leave the generalization for additional research.

We assume throughout the paper that $f(\cdot)$ is twice continuously differentiable, so that at any given $x$, we use a local approximation: $f(x_t) \simeq f(x) + f'(x) (x_t - x)$, when $x_t$ is in the neighborhood of $x$, where $\simeq$ denotes the first order Taylor approximation and $f'(x)$ is the first derivative of $f(x)$. Hence (1.1) is approximated by

$$y_t \simeq \theta_0 + (x_t - x) \theta_1 + u_t,$$

and it becomes a local linear model. Therefore, the locally weighted sum of squares is

$$\sum_{t=1}^n [y_t - \theta_0 - (x_t - x) \theta_1]^2 K_h(x_t - x), \quad (2.3)$$

where $K_h(x) = K(x/h)/h$, $K(\cdot)$ is the kernel function, and $h = h_n > 0$ is the bandwidth satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$, which controls the amount of smoothing used.
in the estimation. By minimizing (2.3) with respect to \( \theta_0 \) and \( \theta_1 \), we obtain the local linear estimate of \( f(x) \), denoted by \( \hat{f}(x) \), and the local linear estimator of the derivative of \( f(x) \), denoted by \( \hat{f}'(x) \). It is easy to show that the minimizer of (2.3) is given by

\[
\begin{bmatrix}
\hat{f}(x) \\
\hat{f}'(x)
\end{bmatrix} = \left[ \sum_{t=1}^{n} \left( \frac{1}{x_t - x} \right)^2 K_h(x_t - x) \right]^{-1} \sum_{t=1}^{n} \left( \frac{1}{x_t - x} \right) y_t K_h(x_t - x),
\]

where \( A \otimes 2 = AA^T \) (\( A \otimes 1 = A \)) for a vector or matrix \( A \).

### 2.2 Notations and Assumptions

We first introduce some notation before making the model assumptions. Denote by \( W_d(\cdot) \) the fractional Brownian motion with \(|d| < 1/2\). Then \( W_d(t) \) for \( t > 0 \) admits the following representation in the sense of Itô integral,

\[
W_d(t) = \frac{1}{A(d)} \int_{-\infty}^{0} ((t - s)^d - (-s)^d) dW(s) + \int_{0}^{t} (t - s)^d dW(s),
\]

where \( W(\cdot) \) is a standard Brownian motion and

\[
A^2(d) = \frac{1}{2d + 1} + \int_{0}^{\infty} ((1 + s)^d - s^d)^2 ds.
\]

It is well known that for \( d \in (0, 1/2) \), \( W_d(\cdot) \) inherits long run dependence in its increments; that is

\[
\sum_{m=1}^{\infty} \text{Cov}(W_d(m) - W_d(m - 1), W_d(1)) = \infty.
\]

For detailed properties of a fractional Brownian motion, we refer to the paper by Mandelbrot and Van Ness (1968). A stochastic process \( W_{c,d}(\cdot) \) is called an O-U fractional Brownian motion with parameters \((c, d)\) if it admits the following expression,

\[
W_{c,d}(s) = W_d(s) - c \int_{0}^{s} e^{-c(s-u)} W_d(u) du,
\]

where \( W_d(\cdot) \) is a fractional Brownian motion and \( c \) and \( d \) are two parameters satisfying \( c \geq 0 \). Clearly, when \( c = 0 \), \( W_{c,d}(\cdot) \) reduces to \( W_d(\cdot) \), when \( d = 0 \), an O-U fractional Brownian motion becomes an O-U process driven by a standard Brownian motion, and when \( c = 0 \) and \( d = 0 \), \( W_{c,d}(\cdot) \) is simply a standard Brownian motion. Therefore, \( W_{c,d}(\cdot) \) provides a flexible way in approximating a normalized non-stationary series. A more general definition of an O-U process can be found in Buchmann and Chan (2007).

We now list some assumptions to be used later. Let \( c, \sigma, \rho \) and \( q \) be some constants.
ASSUMPTION 1. \( n(1-\beta) \to c \), \( d \in [0,1/2) \), \( \sum_{j=0}^{\infty} |\psi_j| < \infty \) and \( b_0 \equiv \sum_{j=0}^{\infty} \psi_j \neq 0 \), and \( E\xi_0^2 < \infty \).

ASSUMPTION 2. \( F_t = \sigma\{u_i, \xi_j, 1 \leq i \leq t, -\infty < j \leq t+1\} \) is the smallest \( \sigma \)-field generated by \( (u_i, \xi_j), 1 \leq i \leq t, -\infty < j \leq t+1 \). Assume that \( E((u_t, \xi_{t+1})|F_{t-1}) = 0 \), \( E(u_t \xi_{t+1}) \to \rho \) and \( E(u_t^2|F_{t-1}) \to \sigma_u^2 > 0 \) a.s. as \( t \to \infty \). Also, \( \sup_{1 \leq t \leq n} E|u_t|^q < \infty \) for some \( q > 2 \).

Assumption 1 is commonly used in the literature; see Wang et al. (2003b) and the references therein. The condition \( n(1-\beta) \to c \) includes the special setting \( \beta = 1-c/n \). Although we assume that \( E\xi_0^2 < \infty \), it is possible to consider the more general setting, where the \( \xi_j \)'s belong to the domain of attraction of some stable law. But this is out of the scope of the present paper. The definition of the filtration (or called information flow) in Assumption 2 implies that \( x_t \in F_{t-1} \) while \( u_t \in F_t \). The second condition implies that \((u_t, \xi_{t+1})\) is a two dimensional martingale difference with respect to \( F_t \) which is slightly more demanding than just saying that \( E(\xi_t|\sigma\{\xi_j, -\infty < j \leq t-1\}) = 0 \). \( F_t \) is more informative than \( \sigma\{\xi_j, -\infty < j \leq t+1\} \) while the former implies further that \( E(u_t \xi_{t+l}) = 0 \) for \( l > 1 \). Assumption 2 allows for heteroskedasticity of model (1.1) for finite samples. The last condition in Assumption 2 guarantees the Linderberg condition in the martingale central limit theorem. The most important implication of Assumption 2 is that \( E(x_t u_t) = 0 \). Wang and Phillips (2009b, 2009c) considered a structural model for which the aforementioned orthogonality may not hold.

Let “\( \Rightarrow \)” denote the weak convergence in the Skorohod space \( D[0,1] \). Denote the convergence in probability and in distribution by \( \to^P \) and \( \to^d \), respectively. The notation \( x_n = o_P(y_n), x_n = o(y_n), x_n = O_P(y_n) \) and \( x_n = O(y_n) \) used later stand respectively for the convergence in probability to 0, the convergence almost surely (a.s.) to zero, tightness and boundedness in limit, of the quotient \( x_n/y_n \). Define

\[
U_n(s) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor ns \rfloor} u_t, \quad \text{and} \quad V_n(s) = \frac{1}{\gamma_n} \sum_{t=1}^{\lfloor ns \rfloor} \epsilon_t, \quad (2.6)
\]

where \( \gamma_n = k(d)n^{1/2+d} \) with \( k^2(d) = b_0^2 E\xi_0^2 \Gamma(1-2d)/[(1+2d)\Gamma(1+d)\Gamma(1-d)] \). Then, we have the following two lemmas with their proofs given in Section 5.
**Lemma 1.** Under Assumptions 1-2, \((U_n, V_n) \Rightarrow (U, W_d)\), where
\[
U(s) = \sigma_u \left( \rho' W(s) + \sqrt{1 - \rho'^2} W^\perp(s) \right)
\]
with \(\rho' = \rho \psi / \sigma_u k(d) \Gamma(1 + d)\) and \(W^\perp(\cdot)\) denoting a standard Brownian motion orthogonal to \(W(\cdot)\).

**Lemma 2.** Under Assumption 1, we have \(x_{[ns]} / \gamma_n \Rightarrow W_{c,d}\), where \(W_{c,d}(\cdot)\) is an O-U fractional Brownian motion.

To obtain the local time approximation of the non-stationary kernel density estimation, we need the following two assumptions.

**Assumption 3.** \(K(\cdot)\) is a continuous kernel function with a compact support.

**Assumption 4.** \(\xi_1\) satisfies the Cramér condition: \(\lim \sup_{|u| \to \infty} |\psi(u)| < 1\), where \(\psi(u)\) is the characteristic function of \(\xi_1\).

Assumption 3 is commonly used in the kernel estimation literature. Assumption 4 is easily fulfilled. For example, any r.v. with the distribution function having a non-zero absolutely continuous component will be strongly non-lattice which amounts to the Cramér condition. Recall that a measurable process \(\{L_{W_{c,d}}(t, x) ; t \geq 0, x \in \mathbb{R}\}\) is called the local time of \(W_{c,d}(\cdot)\) at state \(x\) up to time \(t\) if for each \(t \geq 0\),
\[
\int_0^t I_A(W_{c,d}(s))ds = \int_R I_A(x)L_{W_{c,d}}(t, x)dx,
\]
for all Borel subset \(A \in \mathbb{R}\),
\[
(2.7)
\]
where \(I_A(\cdot)\) is an indicator function of the event \(A\). For ease of notation, we drop \(W_{c,d}\) from \(L_{W_{c,d}}(t, x)\) so that \(L_{W_{c,d}}(t, x)\) becomes \(L(t, x)\). Finally, let \(S = \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix}\) with \(\mu_j = \int_R u^j K(u)du\) for \(j = 0, 1, 2\), \(S^* = \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix}\) with \(\nu_j = \int_R u^j K^2(u)du\) for \(j = 0, 1, 2\) and \(c_2 = \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix}\). Denote by \(e\) the unit vector \(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\).

**2.3 Asymptotic Results**

We now state our main result below and the proof is relegated to Section 5. The notation \(A^T\) and \(A^{-1}\) stands for the transpose and the inverse of a matrix \(A\), respectively.
**Theorem 1.** Under Assumptions 1–4, if $nh^7/\gamma_n \to 0$, we have

$$
\lambda_n \left[ \hat{f}(x) - f(x) - h^2 B(x) \right] \xrightarrow{d} MN(\sigma_f^2),
$$

where $\lambda_n = \sqrt{n h / \gamma_n}$, $B(x) = f''(x)/2e^T S^{-1} c_x$ and $MN(\sigma_f^2)$ is a mixed normal distribution with mean zero and conditional variance $\sigma_f^2 = \sigma_u^2 e^T S^{-1} S^* S^{-1} e/L(1,0)$.

**Remark 1.** The asymptotic properties for $\hat{f}'(x)$ can be obtained as the same fashion as those in Theorem 1 and omitted. By comparing the results in Theorem 1 and conventional findings in Fan and Gijbels (1996) and Fan and Yao (2003) for the stationary covariates, our new results can be summarized as follows. Clearly, $B(x)$ serves as the asymptotic bias, which is the same as that for stationary case when one uses a local linear estimation method; see Fan and Yao (2003). If we choose $K(u)$ as a probability density function with zero mean, the bias $B(x)$ and the variance $\sigma_f^2$ become respectively $f''(x)/2$ and $\sigma^2 \nu_0/L(1,0)$ which are the same as those for the Nadaraya-Watson estimator (see Theorem 3 later). This is consistent with the fact that the asymptotic bias term comes mainly from the local linear approximation. However, the convergence rate is of order $\lambda_n$ much slower than that for stationary covariates. Also, the stochastic asymptotic variance is independent of the grid point $x$. Indeed, one can show that the results in Theorem 1 hold true as long as any $x = x_n$ satisfies $x_n/\gamma_n \to 0$ and $\lambda_n h^2 f''(x_n) = O(1)$; see Theorem 2 later. Furthermore, from the asymptotic bias and variance presented in Theorem 1, the stochastic AMSE is given by

$$
\text{AMSE} = \text{Var} + \text{bias}^2 = \sigma_f^2 \lambda_n^{-2} + \frac{h^4}{4} \mu_2^2(K) [f''(x)]^2.
$$

The minimization of the AMSE with respect to $h$ yields the optimal bandwidth

$$
h_{opt} = \left[ \frac{4 \sigma_f^2 \gamma_n}{\mu_2^2(f''(x))^2 n} \right]^{1/5} = O_p\left( (n^{d-1/2})^{1/5} \right), \tag{2.8}
$$

which is stochastic and much larger than the conventional optimal bandwidth $h_{opt,s} = O(n^{-1/5})$ for the stationary case; see Fan and Yao (2003). Therefore, if $h_{opt,s}$ is used in estimating $f(\cdot)$ in (1.1), the nonparametric estimator given in (2.4) is undersmoothing. Hence, it is of own interest of investigating theoretically and empirically the data-driven (optimal) bandwidth selection and it can be a interesting future research topic.
To make Theorem 1 applicable in statistical inference, a consistent estimator of the stochastic variance has to be given. Let $\hat{\sigma}^2_f = \sigma^2_0 e^T S^{-1} S^* S^{-1} e / \sum_{t=1}^n K(\frac{x_t-x}{h})$. In virtue of Proposition 1 in Section 5, we have the following

**Corollary 1.** Under the conditions in Theorem 1,
\[
\frac{1}{\sqrt{\hat{\sigma}^2_f}} \left[ \hat{f}(x) - f(x) - h^2 B(x) \right] \xrightarrow{d} N(0, 1),
\]
where $N(0, 1)$ stands for standard normal random variable.

Now, we embark on investigating the asymptotic behaviors at boundaries. When $x_t$ is NI(1), it follows from Lemma 2 that when $x = a \gamma_n$ ($a \neq 0$) and $r = t/n$,
\[
P(x_t \geq x) = P(x_t \geq a \gamma_n) \rightarrow P(W_{c,d}(r) \geq a) > 0.
\]
This means that there is a great chance that $|x_t|$ can take large values. In other words, an NI(1) time series takes longer to revisit levels in its range. Now the question is how the asymptotic behaviors of the estimator look like when $x$ is large like $x = a \gamma_n$ for any fixed $a$. To this end, we obtain the following asymptotic results at the boundary $x = a \gamma_n$ for any fixed $a$. However, we do not provide the detailed proofs since they follow closely the same arguments as those used in the proof of Theorem 1.

**Theorem 2.** If Assumptions 1 – 4 hold and $\lambda_n h^2 f''(a \gamma_n) = O(1)$ for any $a$, then, we have
\[
\lambda_n \left[ \hat{f}(a \gamma_n) - f(a \gamma_n) - h^2 B(a \gamma_n) \right] \xrightarrow{d} MN(\sigma_a^2),
\]
where $MN(\sigma_a^2)$ is a mixed normal distribution with mean zero and conditional variance $\sigma_a^2 = \sigma_0^2 e^T S^{-1} S^* S^{-1} e / L(1, a)$.

**Remark 2.** Comparing Theorem 2 with Theorem 1, we observe that the magnitude of the asymptotic variance of $\hat{f}(\cdot)$ at the boundary points ($x = O(\gamma_n)$) differs from that for the interior points ($x = o(\gamma_n)$). This is different from its stationary counterparts; see Fan and Gijbels (1996) for the stationary case.
2.4 Nadaraya-Watson Estimation

Now we turn to the asymptotic properties for the local constant estimator of $f(\cdot)$. It is well documented that the Nadaraya-Watson estimator is given by

$$\tilde{f}(x) = \frac{\sum_{t=1}^{n} y_t K_h(x_t - x)}{\sum_{t=1}^{n} K_h(x_t - x)}. \tag{2.9}$$

For $\tilde{f}(x)$, we have the following theorem.

**Theorem 3.** Under the assumptions of Theorem 1, if further $\int_R K(u)du = 1$, $\int_R uK(u)du = 0$, $\gamma_n/nh \to 0$ and $nh^2/\gamma_n \to 0$, both $\tilde{f}(x)$ and $\hat{f}(x)$ share the exact same asymptotic properties. That is, we have

$$\lambda_n \left[ \tilde{f}(x) - f(x) - h^2 B(x) \right] \xrightarrow{d} MN(\sigma_f^2),$$

where $B(x) = \mu_2 f''(x)/2$ and $MN(\sigma_f^2)$ is a mixed normal distribution with mean zero and conditional covariance $\sigma_f^2 = \sigma_u^2 \nu_0/L(1,0)$. Further, Theorem 2 holds for $\tilde{f}(x)$.

**Remark 3.** It is clear that $h^2 \mu_2 f''(x)/2$ serves as the asymptotic bias, which is the same as the case when one uses a local linear estimation method (see Theorem 1). However, for the stationary $x_t$ case with a local constant estimation method, there is an additional leading bias term which has the form of $h^2 \mu_2 f''_x(x)f'(x)/2f_x(x)$ where $f_x(\cdot)$ is the stationary density of $x_t$ when $x_t$ is stationary; see Fan and Gijbels (1996). Theorem 3 shows that for non-stationary $x_t$, the local constant estimator has the same leading bias as that of a local linear method. This is an interesting new finding that is not shared by a local constant estimator if $x_t$ is stationary. It can be shown that with non-stationary $x_t$, the bias term associated with $f_{t,x}(x)f'(x)$, where $f_{t,x}(x)$ is the density of $(x_t - x)/\sqrt{t}$, has an order of $h\sqrt{\gamma_n h/n}$, which is smaller than $h^2$; see (5.41) in Section 5. Therefore, the leading bias contains only one term associated with $f''(x)$ with the order $h^2$. Interestingly, as in the case of standard local polynomial methods, the Nadaraya-Watson estimator is design-adaptive too in the sense of Fan and Gijbels (1996). Clearly, this property should be interpreted as follows. The clustered designs are not expected to occur in the presence of integrated or nearly integrated (highly persistent) processes. Therefore, the theoretical relevance of the design-adaptation property and the theoretical appeal of local polynomial methods over the standard Nadaraya-Watson kernel estimates seem to vanish.

10
3 A Monte Carlo Simulation Study

In this section we report a Monte Carlo simulation to examine the finite sample property of the proposed estimator. In our computation, the Epanechnikov kernel $K(u) = 0.75 (1 - u^2) I(|u| \leq 1)$ is used. To assess the performance of finite samples, we compute the mean absolute deviation errors (MADE) for $\hat{f} (\cdot)$, which is defined as

$$\text{MADE} = m^{-1} \sum_{k=1}^{m} \left| \hat{f}(v_k) - f(v_k) \right|,$$

where $\hat{f}(\cdot)$ is the local linear estimate of $f(\cdot)$. We take \(\{v_k = -1 + 0.1k, k = 1, \cdots, 20\}\) for $f_A$ and \(\{v_k = 0.05k, k = 1, \cdots, 20\}\) for $f_B$.

We consider the following data generating process

$$y_t = f(x_t) + \epsilon_t, \quad t = 1, \ldots, n,$$

where $x_t$ is generated from the integrated or nearly integrated model $x_t = \rho x_{t-1} + u_t$ with $\rho = 1 + c/n$ and $u_t \sim \text{FARIMA}(0, d, 0)$, and $\epsilon_t \sim N(0, 1)$. In the simulations, we consider two functions: $f_A(z) = z^3$ and $f_B(z) = \sum_{i=1}^{4} (-1)^i \sin(j\pi z) / j!$. The Monte Carlo simulation is repeated 500 times for each sample size $n = 200, 500, \text{and} 1000$. Here, we take $c$ to be 0, $-5$ and $-25$ and $d$ to be 1/4 and 0.45 for simplicity. Theoretically, the optimal bandwidth given in (2.8) is $h_{\text{opt}} = A_0 \times n^{-(1-2d)/10}$, where $A_0$ (depending on unknown parameters and functions) can often be estimated in practice by some data-driven methods such as the cross validation method, and $d$ can be replaced by its estimate. In our simulations, we would like to see how the MADE values change with different choices of $A_0$.

The simulation results are presented in Tables 1 and 2 (the median and the standard deviation (in parentheses) of 500 MADE-values). From Tables 1 and 2 first, we can see that the median and the standard deviation (in parentheses) of 500 MADE-values for both $\hat{f}_A$ and $\hat{f}_B$ decrease for all settings when the sample size increases. This is consistent with the asymptotic theory. Secondly, it can be also seen that as the value of $A_0$ increases, the MADE values for both $\hat{f}_A$ and $\hat{f}_B$ start to decrease first, reach the minimum and then increase for all settings. This pattern is invariant for different sample sizes. We note that the MADE values for different sample sizes achieve the minimum when $A_0 = 0.8$ for $\hat{f}_A$ with $d = 1/4$.
Table 1: The median and the standard deviation (in parentheses) of 500 MADE-values for $\hat{f}_A$ and $\hat{f}_B$ with different sample sizes and different values of $c$, $h = A_0 n^{-1/20}$ and $d = 0.25$.

<table>
<thead>
<tr>
<th>$c$</th>
<th>$A_0$</th>
<th>$n = 200$</th>
<th>$n = 500$</th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.2374 (0.8654)</td>
<td>0.5169 (0.4119)</td>
<td>0.4075 (0.1688)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1822 (0.6543)</td>
<td>0.2973 (0.0840)</td>
<td>0.2412 (0.0612)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1562 (0.2977)</td>
<td>0.2247 (0.0505)</td>
<td>0.1987 (0.0421)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.1663 (0.2646)</td>
<td>0.2075 (0.0470)</td>
<td>0.1811 (0.0332)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.2228 (0.3265)</td>
<td>0.2136 (0.0758)</td>
<td>0.1945 (0.0387)</td>
</tr>
<tr>
<td>$\hat{f}_A$</td>
<td>$-5$</td>
<td>0.2278 (0.8246)</td>
<td>0.5169 (0.4119)</td>
<td>0.4075 (0.1688)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1715 (0.2602)</td>
<td>0.2973 (0.0840)</td>
<td>0.2412 (0.0612)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1288 (0.0921)</td>
<td>0.2247 (0.0505)</td>
<td>0.1987 (0.0421)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.1422 (0.0857)</td>
<td>0.2075 (0.0470)</td>
<td>0.1811 (0.0332)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.2010 (0.0704)</td>
<td>0.2136 (0.0758)</td>
<td>0.1945 (0.0387)</td>
</tr>
<tr>
<td></td>
<td>$-20$</td>
<td>0.2278 (0.8246)</td>
<td>0.5169 (0.4119)</td>
<td>0.4075 (0.1688)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1715 (0.2602)</td>
<td>0.2973 (0.0840)</td>
<td>0.2412 (0.0612)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1288 (0.0921)</td>
<td>0.2247 (0.0505)</td>
<td>0.1987 (0.0421)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.1422 (0.0857)</td>
<td>0.2075 (0.0470)</td>
<td>0.1811 (0.0332)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.2010 (0.0704)</td>
<td>0.2136 (0.0758)</td>
<td>0.1945 (0.0387)</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.2487 (0.3106)</td>
<td>0.2090 (0.2919)</td>
<td>0.2044 (0.1658)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.2228 (0.1542)</td>
<td>0.1565 (0.1340)</td>
<td>0.1531 (0.1212)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.2243 (0.1456)</td>
<td>0.1785 (0.1205)</td>
<td>0.1682 (0.1165)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.2893 (0.1232)</td>
<td>0.2336 (0.1170)</td>
<td>0.2354 (0.1032)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.3807 (0.1333)</td>
<td>0.3479 (0.0958)</td>
<td>0.3240 (0.0878)</td>
</tr>
<tr>
<td>$\hat{f}_B$</td>
<td>$-5$</td>
<td>0.2360 (0.3556)</td>
<td>0.1718 (0.3229)</td>
<td>0.1387 (0.1188)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1625 (0.2295)</td>
<td>0.1300 (0.0738)</td>
<td>0.1134 (0.0919)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1815 (0.0990)</td>
<td>0.1601 (0.0718)</td>
<td>0.1511 (0.0660)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.2637 (0.1265)</td>
<td>0.2304 (0.0750)</td>
<td>0.2161 (0.0620)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.3525 (0.1161)</td>
<td>0.2274 (0.0827)</td>
<td>0.3061 (0.0698)</td>
</tr>
<tr>
<td></td>
<td>$-20$</td>
<td>0.1580 (0.1018)</td>
<td>0.1176 (0.0501)</td>
<td>0.0996 (0.0360)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1250 (0.0563)</td>
<td>0.1050 (0.0404)</td>
<td>0.0959 (0.0361)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1621 (0.0637)</td>
<td>0.1545 (0.0498)</td>
<td>0.1403 (0.0420)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.2469 (0.0698)</td>
<td>0.2252 (0.0516)</td>
<td>0.2162 (0.0469)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.3456 (0.0733)</td>
<td>0.3152 (0.0580)</td>
<td>0.2943 (0.0493)</td>
</tr>
</tbody>
</table>
Table 2: The median and the standard deviation (in parentheses) of 500 MADE-values for $\hat{f}_A$ and $\hat{f}_B$ with different sample sizes, $h = A_0 n^{-1/100}$ and $d = 0.45$.

<table>
<thead>
<tr>
<th>c</th>
<th>$A_0$</th>
<th>$n = 200$</th>
<th>$n = 500$</th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$A_0$</td>
<td>$A_0$</td>
<td>$A_0$</td>
</tr>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.2649 (0.5432)</td>
<td>0.2196 (0.4132)</td>
<td>0.1988 (0.2881)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1913 (0.2987)</td>
<td>0.1511 (0.2751)</td>
<td>0.1270 (0.1632)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1566 (0.1865)</td>
<td>0.1295 (0.1765)</td>
<td>0.1167 (0.1325)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.1670 (0.1443)</td>
<td>0.1501 (0.1349)</td>
<td>0.1336 (0.1124)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.2120 (0.1646)</td>
<td>0.1941 (0.1086)</td>
<td>0.1773 (0.0897)</td>
</tr>
<tr>
<td>$\hat{f}_A$</td>
<td>−5</td>
<td>0.2</td>
<td>0.2563 (0.6392)</td>
<td>0.2221 (0.6098)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1581 (0.4453)</td>
<td>0.1378 (0.4087)</td>
<td>0.0968 (0.0504)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1345 (0.2514)</td>
<td>0.1143 (0.2301)</td>
<td>0.0903 (0.0420)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.1434 (0.1207)</td>
<td>0.1367 (0.1153)</td>
<td>0.1120 (0.0425)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.2032 (0.0824)</td>
<td>0.1820 (0.0876)</td>
<td>0.1622 (0.0414)</td>
</tr>
<tr>
<td></td>
<td>−20</td>
<td>0.2</td>
<td>0.1598 (0.1295)</td>
<td>0.1258 (0.0455)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1044 (0.0366)</td>
<td>0.0858 (0.0287)</td>
<td>0.0722 (0.0249)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1014 (0.0395)</td>
<td>0.0854 (0.0317)</td>
<td>0.0745 (0.0279)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.1294 (0.0446)</td>
<td>0.1157 (0.0355)</td>
<td>0.1086 (0.0297)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.1899 (0.0439)</td>
<td>0.1686 (0.0343)</td>
<td>0.1560 (0.0303)</td>
</tr>
<tr>
<td>$\hat{f}_B$</td>
<td>−5</td>
<td>0.2</td>
<td>0.2422 (0.4126)</td>
<td>0.1827 (0.3761)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.2082 (0.2546)</td>
<td>0.1618 (0.2231)</td>
<td>0.1434 (0.0764)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.2249 (0.2237)</td>
<td>0.1793 (0.1925)</td>
<td>0.1588 (0.0432)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.2765 (0.1811)</td>
<td>0.2370 (0.1471)</td>
<td>0.2323 (0.0343)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.3673 (0.1867)</td>
<td>0.3205 (0.1228)</td>
<td>0.3127 (0.0675)</td>
</tr>
<tr>
<td></td>
<td>−20</td>
<td>0.2</td>
<td>0.2299 (0.3806)</td>
<td>0.2252 (0.3139)</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.1551 (0.2005)</td>
<td>0.1492 (0.1738)</td>
<td>0.1201 (0.0737)</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.1902 (0.1151)</td>
<td>0.1749 (0.1087)</td>
<td>0.1560 (0.0450)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.2422 (0.0971)</td>
<td>0.2426 (0.0901)</td>
<td>0.2242 (0.0325)</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.3543 (0.1075)</td>
<td>0.3434 (0.0933)</td>
<td>0.3117 (0.0532)</td>
</tr>
</tbody>
</table>

13
and $A_0 = 0.6$ for $\hat{f}_A$ with $d = 0.45$, and when $A_0 = 0.4$ for $\hat{f}_B$ with both $d = 1/4$ and $d = 0.45$. This is in line with the fact that $A_0$ only depends on the population parameter and functionals, but not on the sample size. The results in Tables 1 and 2 also show that the MADE values are not too sensitive to the choice of the bandwidth if $A_0$ is in $(0.4, 1)$. This is a good thing in practice since one does not need to worry too much about getting a rough estimate of the bandwidth.

4 Discussion

In this paper, we studied a nonparametric regression model for nearly integrated time series data with a possible long range dependence. We suggested using the local polynomial and local constant fitting schemes to estimate the nonparametric function and derived the asymptotic properties of the proposed estimators. Our theoretical results show that the asymptotic bias is of the same order as that for stationary covariates. But, the convergence rate for the nonstationary covariates is slower than that for the stationary covariates by a factor of $n^{-1/4}$. Further, the asymptotic distribution is not normal any more but just a mixed normal associated with the local time of an O-U fractional Brownian motion. Moreover, we showed that the asymptotic properties for both the local linear and local constant estimators are exactly same. We would like to mention some interesting future research topics related to this paper. First, it would be very useful and important to discuss how to select the data-driven (optimal) bandwidth empirically. Secondly, the model may include both stationary and nonstationary covariates. Finally, it is worth considering some extensions to other types of nonstationary models such as semiparametric models, additive models, index models and varying coefficient models.

5 Proofs

Throughout this section, we denote by $C$ or $\tilde{C}$ a generic positive constant, which may take different values at different places. In the sequel we drop the dependence on $d$ of $c_{k,-d}$ and write for simplicity $c_k$ for $c_{k,-d}$.

Proof of Lemma 1: Let $c_k = \Gamma(d+k)/[\Gamma(d)\Gamma(k+1)]$. Then $c_k \sim k^{d-1}/\Gamma(d)$, as $k \to \infty$. By
Theorem 2.1 in Wang et al (2003),

$$V_n(s) \Rightarrow W_d(s), \quad 0 \leq s \leq 1.$$  \hspace{1cm} (5.10)

On the other hand, for fixed time $s$, $U_n(s)$ is the end point of a mean 0 martingale with respect to the filtration $\mathcal{F}_{nt}$, $1 \leq t \leq [ns]$, and we also have

$$\frac{1}{n} \sum_{t=1}^{[ns]} E(u_t^2 | \mathcal{F}_{t-1}) \equiv \frac{1}{n} \sum_{t=1}^{[ns]} \sigma_t^2 = \int_{1/n}^s \sigma_{[nu]}^2 du \to P \sigma_u^2 s.$$  \hspace{1cm}

By the martingale central limit theorem, we deduce that

$$U_n(s) \Rightarrow \sigma_u \tilde{W}_s,$$  \hspace{1cm} (5.11)

where $\tilde{W}$ is a standard Brownian motion. By (5.10) and (5.11) it suffices to prove that

$$E(U_n(s_1)V_n(s_2)) \to \sigma_u E(\tilde{W}_s W_d(s_2)) = \frac{\rho b_\psi}{k(d) \Gamma(2 + d)} (s_2^{d+1} - (s_2 - s_1) \cap 0)^{d+1}).$$  \hspace{1cm} (5.12)

In fact, suppose $s_1 < s_2$

$$E(U_n(s_1)V_n(s_2)) = \frac{1}{k(d)n^{1+d}} \sum_{t=1}^{[ns_1]} \sum_{l=1}^{[ns_2]} E(u_t \epsilon_l)$$

$$= \frac{1}{k(d)n^{1+d}} \sum_{t=1}^{[ns_1]} \sum_{l=t+1}^{[ns_2]} \int_{1/n}^s \sum_{j=0}^{\infty} \psi_j \sum_{k=0}^{\infty} c_k \xi_{t-j-k} E(u_t \xi_{t+1})$$

$$\equiv \frac{1}{k(d)n^{1+d}} \sum_{t=1}^{[ns_1]} \sum_{l=t+1}^{[ns_2]} \int_{1/n}^s \sum_{j=0}^{\infty} \psi_j \xi_{t-j-(t+1)} \rho_{t+1}$$

$$= \frac{1}{k(d)n^{1+d}} \sum_{j=0}^{\infty} \psi_j \sum_{t=1}^{[ns_1]} \sum_{l=t+1}^{[ns_2]} c_{nt-n(t+1)/n-j} \rho_{nt+1/n}$$

$$= \frac{1}{k(d)n^{1+d}} \sum_{j=0}^{\infty} \psi_j n^2 \int_{2/n}^{s_1} \int_{[nv]/n}^{s_2} c_{[nu]-[nv]-j} \rho_{[nu]} dudv$$

$$\sim \frac{1}{k(d)n^{1+d}} \sum_{j=0}^{\infty} \psi_j n^2 \int_{2/n}^{s_1} \int_{[nv]/n}^{s_2} \rho/\Gamma(d)(nu-nv-j)^{d-1} dudv$$

$$\sim \frac{\rho b_\psi}{k(d) \Gamma(2 + d)} (s_2^{d+1} - (s_2 - s_1)^{d+1}),$$  \hspace{1cm} (5.13)
which verifies \((5.12)\). The case for \(s_1 \geq s_2\) can be done similarly.

**Proof of Lemma 2:** Following the ideas as in Buchmann and Chan (2007), we have

\[
\frac{1}{\gamma_n}x_{[ns]} = \frac{\beta^{[ns]}}{\gamma_n}x_0 + \frac{1}{\gamma_n} \sum_{k=1}^{[ns]} \beta^{[ns]-k} \epsilon_k \\
= \frac{\beta^{[ns]}}{\gamma_n}x_0 + \sum_{k=1}^{[ns]} \beta^{[ns]-k} (V_n(\frac{k}{n}) - V_n(\frac{k-1}{n}))
\]

\[
= \frac{\beta^{[ns]}}{\gamma_n}x_0 + V_n(\frac{[ns]}{n}) - \beta^{[ns]-1} V_n(0) - \beta^{[ns]} \sum_{k=1}^{[ns]-1} (\beta^{-(k+1)} - \beta^{-k}) V_n(\frac{k}{n})
\]

\[
= \frac{\beta^{[ns]}}{\gamma_n}x_0 + V_n(\frac{[ns]}{n}) + \beta^{[ns]} n \log(\beta) \sum_{k=1}^{[ns]-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} V_n(u) \beta^{-nu} du
\]

\[
\equiv I - II + III + IV + V,
\]

where

\[
I = \frac{1}{\gamma_n} \beta^{[ns]} x_0, \quad II = \left( \int_{[ns]n}^{\frac{1}{n}} + \int_{0}^{\frac{1}{n}} \right) (n \log(\beta)) \beta^{[ns]-nu} V_n(u) du;
\]

\[
III = V_n(\frac{[ns]}{n}) - V_n(s), \quad IV = \int_{0}^{s} \left[ (n \log(\beta)) \beta^{[ns]-nu} + ce^{-c(s-u)} \right] V_n(u) du;
\]

\[
V = V_n(s) - c \int_{0}^{s} e^{-c(s-u)} V_n(u) du.
\]

By the condition in Lemma 2, \(I \to^P 0\). By Theorem 2.1 of Wang et al (2003), \(V_n(\cdot) \Rightarrow W_d(\cdot)\) in \(D[0,1]\), then by the Skorohod representation Theorem, there exists \(V_n^*\) and \(W^*d\) such that \(V_n^* =^d V_n\) and \(W^*d =^d W_d\), and \(\sup_{0\leq s\leq 1} |V_n^*(s) - W^*_d(s)| \to 0\). Without loss of generality in the context of proving Lemma 2, let \(V_n = V_n^*\) and \(W_d = W^*_d\). Then \(\sup_{0\leq s\leq 1} |V(s) - W_{c,d}(s)| \to 0\), which implies that \(V \Rightarrow W_{c,d}\) in \(D[0,1]\) under uniform topology. Using uniform tightness of \(V_n\) and uniform boundedness of \((n \log(\beta)) \beta^{[ns]-nu}\) in the interval \([0,1]\), \(II \to^P 0\), and \(III \to^P 0\). Since the integrand of \(IV\) goes to 0 uniformly and \(V_n\) is uniformly tight, \(IV \to^P 0\).

Combining the above arguments, Lemma 2 is proved.

Before we prove the main results of this paper, we first give a useful proposition.

**Proposition 1** Let \(F(\cdot)\) be a continuous function with compact support. Under assumption 1, 3 and 4, if \(\gamma_n/(nh) \to 0\) then for any \(0 \leq s \leq 1\), we have
1. If $\int_R F(u)du \neq 0$,

$$\frac{\gamma_n}{nh} \sum_{t=1}^{[ns]} F((x_t - x)/h) \rightarrow^d L_{\gamma_n}(s,0) \int_R F(y)dy.$$ 

2. If $\int_R F(u)du = 0$ and $\int_R F^2(u)du > 0$,

$$\sum_{t=1}^{[ns]} F((x_t - x)/h) = O_P(\sqrt{nh/\gamma_n}).$$

**Proof.** For simplicity, we assume that $\psi_0 = 1$, and $\psi_j = 0$ for $j > 0$ below; the general case can be done similarly. Let $F_n(x) = \gamma_n F(\gamma_n x/h)/h$, then

$$\frac{\gamma_n}{nh} \sum_{t=1}^{[ns]} F((x_t - x)/h) = \frac{1}{n} \sum_{t=1}^{[ns]} F_n(x_t/\gamma_n - x/\gamma_n).$$

We follow the procedure used in Jeganathan (2004) by constructing the following approximation in proving the first part:

$$\frac{1}{n} \sum_{t=1}^{[ns]} F_n(x_t/\gamma_n - x/\gamma_n) - \frac{1}{n} \sum_{t=1}^{[ns]} \int_R F_n(x_t/\gamma_n - x/\gamma_n + z\epsilon)\phi(z)dz \rightarrow L^2 0,$$

(5.15)

uniformly for $0 \leq t \leq 1$ and $x \in R$ by first letting $n \rightarrow \infty$ and then $\epsilon \rightarrow 0$, where $\phi(\cdot)$ denotes the probability density function of the standard random variable. To this end, let $g_n(k) = \sum_{j=0}^{k} c_j \beta^{k-j}$, $0 \leq k \leq n$, and rewrite $\frac{1}{\gamma_t} x_t$, $1 \leq t \leq n$ as

$$\frac{1}{\gamma_t} x_t = \frac{1}{\gamma_t} \left[ \sum_{j=-\infty}^{0} \xi_j g_n(t - j) - g_n(-j) \right] + \sum_{j=1}^{t} \xi_j g_n(t - j) \equiv S_{nt} + S_{nt}^* \equiv 1 \leq t \leq n.$$ 

(5.16)

Follow the arguments of the proof of Proposition 6 of Jeganathan (2004), it suffices to give the following two estimates. Denote the characteristic function of $S_{nt}^*$ by $\widehat{H}_{nt}(u)$.

1. There are constants $A_1 > 0$, $A_2 > 0$, $A_3 > 0$ such that $|\widehat{H}_{nt}(u)| \leq A_1 e^{-A_2 u^2}$ for all $|u| \leq A_3 \sqrt{t}$ as $t$ large enough.

2. For any $B_1 > 0$, there is a $B_2 > 0$ and a $0 < \rho < 1$ such that

$$\sup_{|u| \geq B_1 \sqrt{t}} |\widehat{H}_{nt}(u)| \leq B_2 \rho^t$$

for $t$ large enough.
We first prove item 1. $S^*_n$ can be rewritten as

$$S^*_n = \sum_{j=0}^{t-1} g_n(j)\xi_{t-j} = \sum_{j=0}^{t-1} \left(\sum_{k=0}^{j} c_k\beta^{j-k}\right)\xi_{t-j} \equiv \sum_{j=0}^{t-1} b_n j \xi_{t-j}.$$  

Simple calculation yields as $j \to \infty$

$$C_j \sim b_j \equiv \sum_{k=0}^{j} c_k \beta^{j-k} \leq |b_n| \leq \sum_{k=0}^{j} c_k \sim C j^d \quad (5.17)$$

Since $E(\xi^2) \leq \infty$,

$$\psi(v) = 2 \sum_{k=0}^{2} E\xi_1^k (iv)^k + o(v^2) \quad as \quad v \to 0. \quad (5.18)$$

By definition, for any $0 < \delta < 1$

$$|\hat{H}_{nt}(u)| = |\prod_{j=0}^{t-1} \psi\left(\frac{1}{\gamma_t} b_n j u\right)| = |\exp\left(\sum_{j=0}^{t-1} \log \psi\left(\frac{1}{\gamma_t} b_n j u\right)\right)|$$

$$\leq |\exp\left(\sum_{j=[\delta t]}^{t-1} \log \psi\left(\frac{1}{\gamma_t} b_n j u\right)\right)| \equiv |\tilde{H}_{nt}(u)|. \quad (5.19)$$

Since $|u| \leq A_3\sqrt{t}$ where $A_3$ can be chosen arbitrarily small, by (5.17), $\frac{1}{\gamma_t} b_n j u \to 0$ uniformly for all $0 \leq j \leq t$ by first letting $t \to \infty$ and then $A_3 \to 0$. By this fact,

$$\log \psi\left(\frac{1}{\gamma_t} b_n j u\right) \sim -\frac{E\xi_1^2 (b_n j u)^2}{2}, \quad as \quad t \to \infty \quad first \quad and \quad then \quad A_3 \to 0. \quad (5.20)$$

Combination of (5.20) and the left inequality of (5.17) implies there exist constants $A_1$ and $A_2$, such that when $A_3$ is small enough,

$$|\hat{H}_{nt}(u)| \leq A_1 e^{-A_2 u^2}, \quad for \quad |u| \leq A_3\sqrt{t},$$

which, along with (5.19) proves item 1.

To prove item 2, note that

$$\left|\prod_{j=0}^{t-1} \psi\left(\frac{1}{\gamma_t} b_n j u\right)\right| \leq \left|\prod_{j=[\delta t]}^{t-1} \psi\left(\frac{1}{\gamma_t} b_n j u\right)\right|. \quad (5.21)$$

By (5.17), there is a small enough number $\epsilon$ such that for $j \geq [\delta t]$ and $|u| \geq B_1\sqrt{t}$,

$$\left|\frac{1}{\gamma_t} \sum_{k=0}^{j} c_k \beta^{j-k} u\right| \geq \left|\frac{1}{\gamma_t} (C - \epsilon) u j^d\right| \geq \frac{1}{2} \frac{B_1 (C - \epsilon)}{k(d)} \delta^d, \quad as \quad t \to \infty. \quad (5.22)$$
Item 2 then follows from (5.21) (5.22), and the cramér’s condition of $\psi(\cdot)$.

So far, we have proved the approximation (5.15), now we proceed to finish the proof of proposition 1. Let

$$L_{n,t}(s,x) = \frac{1}{n} \sum_{t=1}^{[ns]} \int_{\mathbb{R}} K_n \left( \frac{1}{\gamma_n} x_t - \frac{x}{\gamma_n} + z\epsilon \right) \phi(z) dz.$$ 

Follow the same lines as in the proof of Lemma 7 of Jeganathan (2004), we have for each $\epsilon > 0$,

$$\sup_{s,x} |L_{n,t}(s,x) - \left( \int_{\mathbb{R}} K(y) dy \right) \frac{1}{n} \sum_{k=1}^{[ns]} \phi(\frac{x_k}{\gamma_n} - \frac{x}{\gamma_n}) | \rightarrow 0, \text{ as } n \rightarrow \infty, \quad (5.23)$$

where $\phi(y) = \frac{1}{\epsilon} \phi(\frac{y}{\epsilon})$. Adopting the notation used in proof of Lemma 2 when we use the Skorohod representation Theorem, we deduce by continuous mapping that for fixed $\epsilon > 0$,

$$\frac{1}{n} \sum_{k=1}^{[ns]} \phi(\frac{x_k}{\gamma_n} - \frac{x}{\gamma_n}) = \int_{s}^{1} \phi(\frac{x[u]}{\gamma_n} - \frac{x}{\gamma_n}) du \rightarrow \int_{0}^{s} \phi(\text{Wc,d}(u)) du, \text{ a.s..} \quad (5.24)$$

If there is a measurable function $L_{W_{c,d}}(s,x)$ with respect to the product $\sigma$ algebra $\mathcal{F} \otimes \mathcal{B}(\mathbb{R}^+) \otimes \mathcal{B}(\mathbb{R})$ satisfying

$$\int_{0}^{s} I_A(W_{c,d}(u)) du = \int_{\mathbb{R}} I_A(x)L_{W_{c,d}}(s,x) dx, \text{ for all Borel set } A \in \mathcal{R}, \quad (5.25)$$

such that

$$\int_{0}^{s} \phi(\text{Wc,d}(u)) du \rightarrow ^{L^2} L_{W_{c,d}}(s,0), \text{ as } \epsilon \downarrow 0, \quad (5.26)$$

then the whole proof is done. Use the Fourier inverse transformation as in the proof of Proposition 11 of Jeganathan (2004) and establish a similar result of Proposition 10 in that reference paper with the fractional Brownian motion replaced by the O-U fractional Brownian motion, (5.25) and (5.26) can be deduced. Now we come to prove the second part of Proposition 1. Let $\Lambda_n$ be the standard deviation of $W_{c,d}(1)$, by Lemma 2, $x_n/\gamma_n \Lambda_n \rightarrow^d N(0,1)$ where $N(0,1)$ stands for a standard normal random variable. Set the notation $d_n$ in the paper by Wang and Phillips (2009c) as $\gamma_n \Lambda_n$, then Lemma 3.1 in the same reference paper hold without changing any notation. Replacing $x_{0,t}$ in that paper by $S_{nt}^*$, since $F(x)$ is continuous with compact support, the estimate (3.8) in that paper holds also. The above two facts imply Proposition 3.4 of that paper which further implies part 2 here by replacing $g(\cdot)$ in that paper by $F(\cdot)$.
Proof of Theorem 1: Let

\[ A_n = \sum_{t=1}^{n} \left( \frac{1}{x_t-x} \frac{x_t-x}{(x_t-x)^2} \right) K_h(x_t-x) \]

and

\[ B_n = \sum_{t=1}^{n} \left( \frac{1}{x_t-x} \right) y_t K_h(x_t-x). \]

Then for some random number \( \xi_t \in (x, x_t) \), and \( H_n = \begin{pmatrix} 1 & 0 \\ 0 & h^{-1} \end{pmatrix}, \)

\[ A_n^{-1}B_n - \left( \frac{f(x)}{f'(x)} \right) = A_n^{-1} \left( \sum_{t=1}^{n} \frac{1}{2} f''(\xi_t)(x_t-x)^2 + u_t \right) \left( \frac{1}{x_t-x} \right) K_h(x_t-x) \]

\[ =: A_n^{-1}(B_n1 + B_n2) = (H_n A_n)^{-1} H_n(B_n1 + B_n2), \quad (5.27) \]

\[ H_n^{-1}(H_n A_n)^{-1} = \sum_{t=1}^{n} \left( \frac{1}{x_t-x} \frac{x_t-x}{(x_t-x)^2} \right) K_h(x_t-x). \quad (5.28) \]

In the sequel proof we strengthen Lemma 1 in an almost sure mode in a suitable probability space by using Skorohod representation Theorem. For simplicity of notation, we still use the same notation and admit that \((U_n, V_n) \rightarrow (U, V), \text{ a.s.}\). We also identify the second component with the notation used in the proof of Lemma 2 when we use the Skorohod representation Theorem there. From (5.15), (5.23), (5.24) and (5.26), we may deduce

\[ \gamma_n \sum_{t=1}^{\lfloor ns \rfloor} F_h(x_t-x) \rightarrow^P L(s, 0) \int R F(y)dy. \quad (5.29) \]

By (5.28) and (5.29),

\[ \frac{n}{\gamma_n} H_n^{-1}(H_n A_n)^{-1} \rightarrow^P L^{-1}(1, 0) \left( \begin{array}{cc} \mu_0 \\ \mu_1 \\ \mu_2 \end{array} \right)^{-1}. \quad (5.30) \]

Similarly we have

\[ \frac{\gamma_n}{nh^2} H_n B_n1 \rightarrow^P \frac{1}{2} f''(x) L(1, 0) \left( \begin{array}{c} \mu_2 \\ \mu_3 \end{array} \right). \quad (5.31) \]

Combination of (5.30) and (5.31) yields

\[ \begin{pmatrix} 1 \\ 0 \end{pmatrix}^T H_n^{-1} A_n^{-1} B_n1 - h^2 B(x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}^T A_n^{-1} B_n1 - h^2 B(x) = o_P(h^2). \quad (5.32) \]
By (5.29) and Assumption 2,
\[
\sum_{t=1}^{n} E \left[ \left( \frac{\gamma_n h}{n} u_t K_h \frac{(x_t - x)}{h} H_n \left( \frac{1}{x_t - x} \right) \right)^{\otimes 2} \bigg| \mathcal{F}_{t-1} \right] \to^P \sigma_u^2 L(1, 0) \begin{pmatrix} \nu_0 & \nu_1 & \nu_2 \end{pmatrix},
\]
So by martingale CLT,
\[
\sqrt{\frac{\gamma_n h}{n}} H_n B_{n2} \to^d \sigma_u \sqrt{L(1, 0)} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},
\tag{5.33}
\]
where \(z_1\) and \(z_2\) are two independent standard normal random variable with the covariance matrix as \(\begin{pmatrix} \nu_0 & \nu_1 & \nu_2 \end{pmatrix}\). In view of (5.27) and (5.33), if \(z_1\) and \(z_2\) are independent of \(W_d\), then the right hand side of (5.33) is a mixed normal random vector and moreover by (5.30)
\[
\sqrt{\frac{nh}{\gamma_n}} \left( \frac{1}{0} \right)^T (H_n A_n)^{-1} H_n B_{n2} = \sqrt{\frac{nh}{\gamma_n}} \left( \frac{1}{0} \right)^T H_n^{-1} (H_n A_n)^{-1} H_n B_{n2} \to^d \frac{\sigma_u S^{-1} (z_1 \ z_2)}{\sqrt{L(1, 0)}}.
\tag{5.34}
\]
Then in virtue of (5.32) and (5.34) Theorem 1 is readily got. Now we return to prove the independence statement. We only prove that \(z_1\) is independent of \(W_d\), since the independence between \(z_2\) and \(W_d\) can be deduced similarly. Without loss of generality as in the proof of Lemma 2.1 in Park and Phillips (1999), we can define
\[
u_t = \sqrt{n}(U(\frac{\tau_{n,t}}{n}) - U(\frac{\tau_{n,t-1}}{n})),
\tag{5.35}
\]
where \(\tau_{n,t}, 0 \leq t \leq 1\), are some stoping times satisfying
\[
\tau_{n,0} = 0, \quad \text{and} \quad \sup_{0 \leq t \leq 1} \left| \frac{\tau_{n,t} - t}{n^\delta} \right| \to 0, \quad \text{as} \quad n \to \infty,
\tag{5.36}
\]
for any \(\delta > \max (1/2, 2/q)\). Define \(M_n(r) = \sqrt{\frac{\tau_n}{nh}} \sum_{t=1}^{[nr]} K(\frac{x_t - x}{h}) u_t\). Then in view of Lemma 1 and (4.28) (4.29), the covariance process between \(M_n\) and \(V\) satisfies
\[
\text{Cov}(M_n, V)_r = \sqrt{\frac{\gamma_n h}{n}} E \sum_{\tau_{n,k} \leq [nr]} K(\frac{x_t - x}{h})(\frac{\tau_{n,k}}{n} - \frac{\tau_{n,k-1}}{n}) + o_P(1)
\]
\[
= \sqrt{\frac{h}{\gamma_n k(d) \Gamma(2+d)}} E \sum_{\tau_{n,k} \leq [nr]} ((r - \frac{\tau_{n,k-1}}{n})^{d+1} - (r - \frac{\tau_{n,k}}{n})^{d+1}) K_n(\frac{x_t - x}{\gamma_n})
\]
21
\[ \sqrt{\frac{h}{\gamma_n k(d)\Gamma(1+d)}} E \sum_{\tau_{n,k} \leq [nr]} \frac{1}{n} \left( \frac{\tau_{n,k} - \frac{n}{\gamma_n}}{n} \right) K_n \left( \frac{x_t - x}{\gamma_n} \right) \leq \sum_{\tau_{n,k} \leq [nr]} \frac{1}{n} \left( \frac{\tau_{n,k} - 1}{n} \right) - \frac{1}{n} \sum_{\tau_{n,k} \leq [nr]} K_n \left( \frac{x_t - x}{\gamma_n} \right) + o_P \left( \sqrt{\frac{h}{\gamma_n}} \right) \rightarrow 0, \] (5.37)

where we have used the estimate \( |(x+h)^{d+1} - x^{d+1}| \leq (d+1)h \) for \( x \in (0, 1) \) and \( 0 < d < 1/2 \), and \( \sqrt{\frac{h}{\gamma_n}} \rightarrow 0 \). (4.30) checks the independence condition. By the condition in Theorem 2 and (4.26), \( E(|VII|\sqrt{\frac{\gamma_n}{n}}) \rightarrow 0 \) which plus (4.26) and the asymptotic normality of \( VI \) complete the proof.

**Proof of Theorem 3:** \( \tilde{f}(x) - f(x) \) can be decomposed as a bias term plus a variance term as follows

\[ \tilde{f}(x) - f(x) = \frac{1}{\sum_{t=1}^{n} K(\frac{x_t - x}{h})} \sum_{t=1}^{n} \left( u_t + f'(x)(x_t - x) + \frac{f''(\xi_t)}{2}(x_t - x)^2 \right) K(\frac{x_t - x}{h}), \] (5.38)

where \( \xi_t \) is some number between \( x \) and \( x_t \). Similar in obtaining (5.34),

\[ \sqrt{\frac{n}{\gamma_n}} \sum_{t=1}^{n} u_t K(\frac{x_t - x}{h}) \Rightarrow \sigma_u \sqrt{\nu_0(K)} \frac{z_3}{\sqrt{L(1,0)}}, \] (5.39)

where \( z_3 \) is independent of \( W_d \), so the right hand side of (5.39) is a mixed normal random variable. Similar in obtaining (5.32),

\[ \frac{1}{\gamma_n} \sum_{t=1}^{n} K(\frac{x_t - x}{h}) \Rightarrow \frac{\sigma_u \sqrt{\nu_0(K)}}{\sqrt{L(1,0)}}, \] (5.39)

where \( z_3 \) is independent of \( W_d \), so the right hand side of (5.39) is a mixed normal random variable. Similar in obtaining (5.32),

\[ \frac{1}{2} \sum_{t=1}^{n} f''(\xi_t)(x_t - x)^2 K(\frac{x_t - x}{h}) \frac{1}{\sum_{t=1}^{n} K(\frac{x_t - x}{h})} - h^2 B(x) = o_P(h^2), \] (5.40)

while

\[ \sum_{t=1}^{n} f'(x)(x_t - x) K(\frac{x_t - x}{h}) = \left\{ \begin{array}{ll} hf'(x)\mu_1 & \mu_1 \neq 0; \\ O_P(h\sqrt{\frac{\gamma_n}{n}}) & \mu_1 = 0. \end{array} \right. \] (5.41)

Combination of (5.39), (5.40) and (5.41) finishes the proof of Theorem 3.

**References**


