The Econometrics Journal

Econometrics Journal (2017), volume **20**, pp. 118–138.
doi: 10.1111/ectj.12082

Nonparametric regression with nearly integrated regressors under long-run dependence

ZONGWU CAI[†], BINGYI JING[‡], XINBING KONG[§,1] AND ZHI LIU[‖]

[†]*Department of Economics, University of Kansas, Lawrence, KS 66045, USA; Wang Yanan Institute for Studies in Economics and the Fujian Provincial Key Laboratory of Statistical Science, Xiamen University, Fujian 361005, China.*
E-mail: caiz@ku.edu

[‡]*Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong, China.*
E-mail: majing@ust.hk

[§]*School of Science, Nanjing Audit University, Nanjing, China.*
E-mail: kongxblqh@gmail.com

[‖]*Department of Mathematics, University of Macau, Avenida da Universidade, Taipa, Macau, China; UMacau Zhuhai Research Institute, No.1 Software Road, Zhuhai Hi-tech Zone, Guangdong 519080, China.*
E-mail: liuzhi@umac.mo

First version received: October 2013; final version accepted: January 2017

Summary We study the nonparametric estimation of a regression function with nonstationary (integrated or nearly integrated) covariates and the error series of the regressor process following a fractional integrated autoregressive moving average 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 fractional Brownian motion. Furthermore, we study the Nadaraya–Watson estimator and we 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. However, for the non-zero energy case, the local linear estimator is superior to the Nadaraya–Watson estimator in terms of optimal convergence rate. We also present a comparison of our results with the conventional results for stationary covariates. Finally, we conduct a Monte Carlo simulation to illustrate the finite sample performance of the proposed estimator.

Keywords: *Local time, Ornstein–Uhlenbeck fractional Brownian motion, Unit root.*

1. INTRODUCTION

Nonparametric estimation techniques have become cornerstone research topics in statistics and econometrics for the last three decades because of their numerous advantages relative

[1]Corresponding author.

to parametric techniques, such as more flexibility and robustness to functional form misspecification. These techniques 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 et al. (2009a) 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 (i.i.d.) 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 but is not limited to the work by Phillips and Park (1998), Park and Hahn (1999), Chang and Martinez-Chombo (2003), Chang and Park (2003), Juhl (2005), Cai et al. (2009b, 2015), Wang and Phillips (2009a, 2011, 2012), Xiao (2009), Cai (2011), Cai and Wu (2013) and Sun et al. (2013, 2016). 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, e.g. Elliott and Stock (1994), Cavanagh et al. (1995), Torous et al. (2004), Campbell and Yogo (2006), Polk et al. (2006), Rossi (2007), Cai and Wang (2014) and Zhu et al. (2014), among others.

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

$$y_t = f(x_t) + u_t, \quad 1 \le t \le n. \tag{1.1}$$

Here, $f(x_t) = E[y_t|x_t]$ is an unknown regression function, $\{u_t\}$ is some stationary sequence and x_t is an integrated or nearly integrated process satisfying

$$x_t = \beta\, x_{t-1} + \epsilon_t, \quad 1 \le t \le n,$$

where $\beta = 1 - c/n$ for $c \ge 0$ and $\{\epsilon_t\}$ is assumed to be a stationary sequence with a possible long-run dependence as a fractional integrated autoregressive moving average (FARIMA) process, which can be expressed as

$$(1 - B)^d \epsilon_t = \eta_t \equiv \sum_{j=0}^{\infty} \psi_j \xi_{t-j}. \tag{1.2}$$

Here, B is the backward operator, ψ_j, $j < \infty$ are some constants, ξ_j, $j > 0$ are i.i.d. random variables with zero mean and finite variance, and $|d| < 1/2$. The fractional power $(1 - B)^d$ is defined as $\sum_{k=0}^{\infty} c_{k,d} B^k$, where $c_{k,d} = \Gamma(-d + k)/\Gamma(-d)\Gamma(k + 1)$ and $\Gamma(\cdot)$ denotes the Γ-function. It is easy to see that $c_{k,d} \sim k^{-d-1}/\Gamma(-d)$ as $k \to \infty$. Therefore, $\{\epsilon_t\}$ in (1.2) can be expressed as a linear process.

Note that the nonparametric regression model in (1.1) is not new in the literature. For example, if x_t is i.i.d. or stationary, model (1.1) has been studied extensively in the literature; see the books by Härdle (1990), Fan and Gijbels (1996), Fan and Yao (2003) and Li and Racine (2007) for details. It was investigated by Karlsen and Tjøstheim (2001) for x_t being a null recurrent time series, by Karlsen et al. (2007) for the ϕ-irreducible Markov chain time series and by Bandi (2002), Cai (2011) and Cai and Wu (2013) for both integrated and nearly integrated time series. Wang and Phillips (2009a, 2009b, 2011, 2012) have considered a nonparametric regression and structure regression when x_t is $I(1)$ and is possibly correlated with u_t, by assuming that $\{\epsilon_t\}$ is either i.i.d. or a linear process. A functional coefficient type model and nonlinear cointegration were investigated by Cai et al. (2009b), Xiao (2009) and Sun et al.

(2013, 2016), for both $I(0)$ and $I(1)$ covariates, and by Cai et al. (2015) for $NI(1)$ covariates. Finally, Miller and Park (2010) investigated the probability properties of model (1.1) by assuming that x_t is an $I(1)$ process and that ϵ_t has a heavy-tailed distribution. However, to identify a nonparametric regression, as pointed out by Miller and Park (2010), the common approaches to three nonstandard modelling approaches (nonlinearity, nonstationarity and long memory) are either separated or in conjunction, to account for the nonstandard features observed in many time series data in economics and finance, such as discontinuous sample paths, excessive volatility or leptokurtosis. Therefore, in this paper, we consider a more general setting under which the aforementioned three nonstandard modelling approaches are combined. It turns out that all our results depend explicitly or implicitly on d.

Model (1.1) might have great potential in many applications. For example, in macroeconomics, a nonparametric form of (1.1) can be used for forecasting the inflation rate based on some persistent and nonstationary covariates, such as velocity; see Bachmeier et al. (2006), which shows that velocity is an $I(1)$ process. Also, it can be used to model a nonlinear cointegration relationship in the purchasing power parity (PPT) hypothesis and to test whether or not the PPT theory holds between two countries; see Hong and Phillips (2010), Sun et al. (2013) and Li et al. (2015) for details on empirical examples. In finance, 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 et al. (1995), Bandi (2002), Torous et al. (2004), Campbell and Yogo (2006), Polk et al. (2006), Rossi (2007), Cai and Wu (2013), Cai and Wang (2014), Zhu et al. (2014) and Cai et al. (2015), among others. In fact, Campbell and Yogo (2006) have shown that the the log dividend–price ratio and the log earnings–price ratio are indeed nonstationary; see Panel A in Table 4 of Campbell and Yogo (2006). Therefore, motivated by the aforementioned empirical examples, we need to study our model.

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 when the regressor x_t is either $I(1+d)$ or $NI(1+d)$ with long-run dependence errors. 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 the stationary case – see Fan and Yao (2003) – and when $0 \leq d < 1/2$, it is slower than the rate $n^{1/5}$ for the case where ϵ_t is short-dependence ($d = 0$); see Cai et al. (2009b). 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 $NI(1+d)$ or $I(1+d)$ 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 Ornstein–Uhlenbeck (O–U) fractional Brownian motion in which the nearly unit root series can be embedded. Furthermore, the nearly integrated covariate requires 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 of the order $O(n^{-1/5})$. Clearly, the use of the conventional bandwidth has the theoretical potential of undersmoothing in the presence of $NI(1+d)$ or $I(1+d)$ covariates. Finally, similar to Cai (2011) and Wang and Philips (2009b), the interesting new finding is that both local linear and local constant estimators share exactly the 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. In Section 2, we present the nonparametric kernel estimators of $f(\cdot)$ using both local linear and Nadaraya–Watson (local constant) estimation methods and their asymptotic behaviours 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. We present concluding remarks in Section 4. Finally, the mathematical proofs of the main results of the paper are relegated to the Appendix.

2. ECONOMETRIC MODELLING

2.1. Local linear estimation

We estimate $f(\cdot)$ using local linear fitting from observations $\{(y_t,\, x_t)\}_{t=1}^n$. Our motivation for 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 we 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 neighbourhood 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

$$\sum_{t=1}^{n} \big(y_t - \theta_0 - (x_t - x)\,\theta_1\big)^2\, K_h(x_t - x), \tag{2.1}$$

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 $n\,h \to \infty$ as $n \to \infty$, which controls the amount of smoothing used in the estimation. By minimizing (2.1) with respect to θ_0 and θ_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 (2.1) is given by

$$(\hat{f}(x),\, \hat{f}'(x))^{\mathrm{T}} = \Big(\sum_{t=1}^{n} \begin{pmatrix} 1 \\ x_t - x \end{pmatrix}^{\otimes 2} K_h(x_t - x) \Big)^{-1} \sum_{t=1}^{n} (1,\, x_t - x)^{\mathrm{T}}\, y_t\, K_h(x_t - x), \tag{2.2}$$

where $A^{\otimes 2} = A\,A^{\mathrm{T}}$ ($A^{\otimes 1} = A$) 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 Type I fractional Brownian motion with $|d| < 1/2$. That is, $W_d(\cdot)$ is a time-continuous Gaussian process with the following covariance structure,

$$E[W_d(t)W_d(s)] = \frac{1}{2}(t^{2d+1} + s^{2d+1} - |t-s|^{2d+1}),$$

where $d + 1/2$ is the so-called Hurst parameter in the literature; see Mandelbrot and Van Ness (1968). Then, $W_d(t)$ for $t > 0$ admits the following representation of the Weyl 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

$$\sum_{m=1}^{\infty} \mathrm{Cov}(W_d(m) - W_d(m-1), W_d(1)) = \infty.$$

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,

$$W_{c,d}(s) = W_d(s) - c \int_0^s e^{-c(s-u)} W_d(u) du, \tag{2.3}$$

where $W_d(\cdot)$ is a fractional Brownian motion and c and d are two parameters satisfying $c \geq 0$ and $|d| < 1/2$. Clearly, when $c = 0$, $W_{c,d}(\cdot)$ reduces to $W_d(\cdot)$, and when $d = 0$, an O–U fractional Brownian motion becomes an O–U process driven by a standard Brownian motion. Further, when both c and d are zero, $W_{c,d}(\cdot)$ is simply a standard Brownian motion. Therefore, $W_{c,d}(\cdot)$ provides a flexible way in approximating a normalized nonstationary 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, σ, ρ and q be some constants.

ASSUMPTION 2.1. *$n(1-\beta) \to c$, $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and $b_\psi \equiv \sum_{j=0}^{\infty} \psi_j \neq 0$. For $d \in [0, 1/2)$, we assume $E[\xi_0^2] < \infty$, and for $d \in (-1/2, 0)$, we assume $E[|\xi_0|^{(2+\delta)/(1+2d)}] < \infty$ for some $\delta > 0$.*

ASSUMPTION 2.2. *$\mathcal{F}_t = \sigma\{u_i, \xi_j, 1 \leq i \leq t, -\infty < j \leq t+1\}$ is the smallest σ-field generated by (u_i, ξ_j), $1 \leq i \leq t$, $-\infty < j \leq t+1$. Assume that $E[(u_t, \xi_{t+1}]|\mathcal{F}_{t-1}) = 0$, $E[u_t \xi_{t+1}] \equiv \rho_{t+1} \to \rho$ and $E[u_t^2|\mathcal{F}_{t-1}] \to \sigma_u^2 > 0$ almost surely (a.s.) as $t \to \infty$. Also, $\sup_{1 \leq t \leq n} E|u_t|^q < \infty$ for some $q > 2$.*

REMARK 2.1. Assumption 2.1 is commonly used in the literature; see Wang et al. (2003b) and 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

$\{\xi_j\}$ belong to the domain of attraction of some stable law. However, this is out of the scope of the present paper. The definition of the filtration (or called information flow) in Assumption 2.2 implies that $x_t \in \mathcal{F}_{t-1}$ while $u_t \in \mathcal{F}_t$. The second condition implies that (u_t, ξ_{t+1}) is a two-dimensional martingale difference with respect to $\mathcal{F}_t$, which is slightly more demanding than just saying that $E[\xi_t|\sigma\{\xi_j, -\infty < j \leq t-1\}] = 0$. Here, $\mathcal{F}_t$ 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 2.2 allows for heteroscedasticity of model (1.1) and the last condition in Assumption 2.2 guarantees the Linderberg condition in the martingale central limit theorem.

REMARK 2.2. The most important implication of Assumption 2.2 is that $E[(x_t\, u_t)] = 0$, which implies that $\{x_t\}$ is a strictly exogenous series. Wang and Phillips (2009b, 2012) and Chang and Park (2010) considered a structural model for which the aforementioned orthogonality may not hold. We expect that, based on the results of Wang (2014) and Chang and Park (2010), relaxing the exogeneity assumption would yield qualitatively similar results but with lengthier mathematical proofs. To this end, we conduct a small simulation at the end of Section 3 to demonstrate this conjecture. As pointed by Miller and Park (2010), for the I-regular class of functions defined in Miller and Park (2010), such robustness immediately follows from the limiting mixed normality that is obtained, even under reasonable allowances for endogeneity. Finally, it is possible to relax the martingale difference assumption on u_t but the proof would be much lengthier.

Let $\Rightarrow$ denote the weak convergence in the Skorohod space $D[0, 1]$; see Billingsley (1999). Denote the convergence in probability and in distribution by $\xrightarrow{p}$ and $\xrightarrow{d}$, respectively. The notations $x_n = o_P(y_n)$, $x_n = o(y_n)$, $x_n = O_P(y_n)$ and $x_n = O(y_n)$ used later denote, respectively, the convergence in probability to 0, the convergence a.s. to zero, tightness and boundedness in limit, of the quotient x_n/y_n. Define

$$U_n(s) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[ns]} u_t \qquad \text{and} \qquad V_n(s) = \frac{1}{\gamma_n} \sum_{t=1}^{[ns]} \epsilon_t, \tag{2.4}$$

where $\gamma_n = k(d)n^{1/2+d}$ with $k^2(d) = b_\psi^2 E\xi_0^2 \Gamma(1-2d)/[(1+2d)\Gamma(1+d)\Gamma(1-d)]$. Then, we have the following two lemmata.

LEMMA 2.1. *Under Assumptions 2.1 and 2.2,* $(U_n, V_n) \Rightarrow (U, W_d)$*, where*

$$U(s) = \sigma_u\big(\rho' W(s) + \sqrt{1-\rho'^2}\, W^{\perp}(s)\big),$$

with $\rho' = \rho b_\psi/\sigma_u k(d)\Gamma(1+d)$*,* ρ *given in Assumption 2.2, and* $W^{\perp}(\cdot)$ *denoting a standard Brownian motion orthogonal to* $W(\cdot)$*.*

LEMMA 2.2. *Under Assumption 2.1, 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 nonstationary kernel density estimation, we need the following two assumptions.

ASSUMPTION 2.3. $K(\cdot)$ *is a continuous kernel function with a compact support.*

ASSUMPTION 2.4. $\int |\psi(u)|du < 1$*, where* $\psi(u)$ *is the characteristic function of* ξ_1*.*

Assumption 2.3 is commonly used in the kernel estimation literature and Assumption 2.4 is easily fulfilled. For example, any random variable 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 $\{L_{W_{c,d}}(t,x); t \geq 0, x \in R\}$ is called the local time of $W_{c,d}(\cdot)$ at state x up to time t for each $t \geq 0$, defined by

$$L_{W_{c,d}}(t,x) = \lim_{\eta \to 0} \frac{1}{2\eta} \int_0^t I_{\{x-\eta < W_{c,d}(s) < x-\eta\}} ds,$$

where $I_A(\cdot)$ is an indicator function of the event A. Then,

$$\int_0^t I_A(W_{c,d}(s))ds = \int_R I_A(x) L_{W_{c,d}}(t,x)dx, \quad \text{for all Borel subset } A \in R. \tag{2.5}$$

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 $(1, 0)^{\mathrm{T}}$. Clearly, $\mu_1 = \nu_1 = 0$ if $K(\cdot)$ is symmetric.

2.3. Asymptotic results

We now state our main result. The notations A^{T} and A^{-1} denote the transpose and the inverse of a matrix A, respectively.

THEOREM 2.1. *Under Assumptions 2.1–2.4, if* $nh^7/\gamma_n \to 0$, $f \in C^2$ *and* $\gamma_n/(nh) \to 0$, *we have*

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

where $\lambda_n = \sqrt{nh/\gamma_n}$, $B(x) = f''(x)/2 e^{\mathrm{T}} S^{-1} c_2$ *and* $MN(\sigma_f^2)$ *is a mixed normal distribution with mean zero and conditional variance* $\sigma_f^2 = \sigma_u^2 e^{\mathrm{T}} S^{-1} S^* S^{-1} e / L(1,0)$.

REMARK 2.3. The asymptotic properties for $\hat{f}'(x)$ can be obtained in the same way as those in Theorem 2.1 and omitted. By comparing the results in Theorem 2.1 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 the 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 σ_f^2 become $f''(x)/2$ and $\sigma^2 \nu_0 / L(1,0)$, respectively, which are the same as those for the Nadaraya–Watson estimator (see Theorem 2.3). 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 λ_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 2.1 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 2.2. Furthermore, from the asymptotic bias and variance presented in Theorem 2.1, the stochastic AMSE is given by

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

where $\mu_2(K) = e^{\mathrm{T}} S^{-1} c_2$. The minimization of the AMSE with respect to h yields the optimal bandwidth

$$h_{\text{opt}} = \Big(\frac{4\sigma_f^2 \gamma_n}{\mu_2^2 (f''(x))^2 n} \Big)^{1/5} = O_p((n^{d-1/2})^{1/5}), \tag{2.6}$$

which is stochastic and much larger than the conventional optimal bandwidth $h_{\text{opt,s}} = O(n^{-1/5})$ for the stationary case; see Fan and Yao (2003). Therefore, if $h_{\text{opt,s}}$ is used in estimating $f(\cdot)$ in (1.1), the nonparametric estimator given in (2.2) is undersmoothing. Hence, it is of interest to investigate theoretically and empirically the data-driven (optimal) bandwidth selection and it might be an interesting future research topic.

To make Theorem 2.1 applicable in statistical inference, an estimator of the stochastic variance has to be given. Let $\hat{\sigma}_f^2 = \hat{\sigma}_u^2 \, e^{\mathrm{T}} S^{-1} S^* S^{-1} e \mu_0 / \sum_{t=1}^n K((x_t - x)/h)$, where $\hat{\sigma}_u^2$ is some consistent estimator of σ_u^2, i.e. $\hat{\sigma}_u^2 \xrightarrow{p} \sigma_u^2$. By virtue of Proposition A.1 in the Appendix, we find that $\hat{\sigma}_f^2/\lambda_n$ converges to σ_f^2 in distribution. Generally, using the classic martingale central limit theorem, one needs convergence in probability so that the studentized sequence (normalize the left-hand side of the equation in Theorem 2.1 by $\hat{\sigma}_f^2$) should converge to the standard normal random variable in distribution. However, by a slight modification of the proof of Theorem 2.1 in Wang (2014), we can easily show that $\hat{\sigma}_f^2/\lambda_n$ converges to σ_f^2 in distribution jointly with $\hat{f}(x) - f(x) - h^2 B(x)$. Hence, we have the following result.

COROLLARY 2.1. *Under the conditions in Theorem 2.1,*

$$\frac{1}{\sqrt{\hat{\sigma}_f^2}} \big(\hat{f}(x) - f(x) - h^2 B(x) \big) \xrightarrow{d} N(0, 1),$$

where $N(0, 1)$ *stands for the standard normal random variable.*

REMARK 2.4. In Corollary 2.1, we assume that there exists a consistent estimator of σ_u^2. Practically, one could use the sample variance of the residuals for x_t in a compact set. Let Ω^* be a compact set on R. Then, a possible estimator of σ_u^2 is

$$\hat{\sigma}_u^2 \equiv \frac{\sum_{t=1}^n (y_t - \bar{f}(x_t))^2 I(x_t \in \Omega^*)}{\sum_{t=1}^n I(x_t \in \Omega^*)},$$

where $\bar{f}(x)$ is either the local linear estimator $\hat{f}(x)$ or the local constant estimator $\tilde{f}(x)$ given later. A key condition for $\hat{\sigma}_u^2$ to be consistent to the target is the following uniform convergence,

$$\sup_{x \in \Omega^*} |\bar{f}(x) - f(x)| = o_p(1).$$

The uniform convergence of the estimator of the nonparametric regression function for nonstationary covariates has become increasingly interesting recently; see Wang and Wang

(2013), Wang and Chan (2014), Gao et al. (2011), and references therein. When $d = 0$ in the definition of ϵ_t in (1.2), the uniform convergence of $\tilde{f}(x)$ to $f(x)$ in probability in a compact set has recently been established by Wang and Wang (2013). However, for $d \neq 0$, the uniform convergence result is still not theoretically underpinned and definitely nontrivial to be extended from Wang and Wang (2013). We leave the derivation of the required uniform convergence in compact sets, as well as the efficient estimation of σ_u^2, for our future research work.

Now, we embark on investigating the asymptotic behaviours at boundaries. When x_t is $NI(1)$, it follows from Lemma 2.2 that when $x = a\,\gamma_n$ ($a \neq 0$) and $r = t/n$,

$$P(x_t \geq x) = P(x_t \geq a\gamma_n) \;\to\; P(W_{c,d}(r) \geq a) > 0.$$

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 what the asymptotic behaviours of the estimator look like when x is large, such as $x = a\,\gamma_n$ 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 detailed proofs because they follow closely the same arguments as those used in the proof of Theorem 2.1.

THEOREM 2.2. *If Assumptions 2.1–2.4 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 $\sigma_a^2 = \sigma_u^2 e^{\mathrm{T}} S^{-1} S^* S^{-1} e / L(1, a)$.

REMARK 2.5. Comparing Theorem 2.2 with Theorem 2.1, 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 finding is different from its i.i.d. and stationary counterparts; see Fan and Gijbels (1996) for the i.i.d. case and see Fan and Yao (2003) 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

$$\tilde{f}(x) = \sum_{t=1}^{n} y_t\, K_h(x_t - x) / \sum_{t=1}^{n} K_h(x_t - x). \tag{2.7}$$

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

THEOREM 2.3. *Under the assumptions of Theorem 2.1, if further $K(\cdot)$ is a symmetric density function and $nh^7/\gamma_n \to 0$, then 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 2.2 holds for $\tilde{f}(x)$.

REMARK 2.6. 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 2.1). 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) and Fan and Yao (2003). Theorem 2.3 shows that for nonstationary x_t, the local constant estimator has the same leading bias as that of a local linear method. This is an interesting phenomenon that is not shared by a local constant estimator if x_t is stationary. It can be shown that with nonstationary 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 (A.25) in the Appendix. 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. Finally, the interesting finding is that for the non-zero energy case, the local linear estimator is superior to the Nadaraya–Watson estimator in terms of the optimal convergence rate; see (A.25) in the Appendix.

3. MONTE CARLO SIMULATION STUDIES

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.

We consider the following data-generating process

$$y_t = f(x_t) + u_t, \qquad t = 1, \ldots, n,$$

where x_t is generated from the integrated or nearly integrated model $x_t = \beta\, x_{t-1} + \epsilon_t$ with $\beta = 1 - c/n$ with $c \geq 0$ and $\epsilon_t \sim$ FARIMA(0, d, 0), and $u_t \sim N(0, 1)$. In the simulations, we consider two functions: $f_A(z) = z^3$ and $f_B(z) = \sum_{i=1}^{4}(-1)^j \sin(j\pi z)/j!$. To assess the performance of finite samples, we compute the mean absolute deviation errors (MADE) for $\hat{f}(\cdot)$, which is defined as

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

where $\hat{f}(\cdot)$ is the local linear estimate of $f(\cdot)$. We take $\{v_k = -1 + 0.1k, k = 1, \ldots, 20\}$ for f_A and $\{v_k = 0.05k, k = 1, \ldots, 20\}$ for f_B. The Monte Carlo simulation is repeated 500 times for each sample size $n = 200$, 500 and 1,000. Here, we take c to be 0, -5 and -20 and d to be $1/4$ and 0.45 for simplicity. Theoretically, the optimal bandwidth given in (2.6) 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.

Table 1. MADE median and standard deviation: $d = 0.25$.

c	A_0	$n = 200$		$n = 500$		$n = 1{,}000$	
				$\hat{f}_A$			
0	0.2	0.2374	(0.8654)	0.5169	(0.4119)	0.4075	(0.1688)
	0.4	0.1822	(0.6543)	0.2973	(0.0840)	0.2412	(0.0612)
	0.6	0.1562	(0.2977)	0.2247	(0.0505)	0.1987	(0.0421)
	0.8	0.1663	(0.2646)	0.2075	(0.0470)	0.1811	(0.0332)
	1.0	0.2228	(0.3265)	0.2136	(0.0758)	0.1945	(0.0387)
−5	0.2	0.2278	(0.8246)	0.5169	(0.4119)	0.4075	(0.1688)
	0.4	0.1415	(0.2602)	0.2973	(0.0840)	0.2412	(0.0612)
	0.6	0.1288	(0.0921)	0.2247	(0.0505)	0.1987	(0.0421)
	0.8	0.1422	(0.0857)	0.2075	(0.0470)	0.1811	(0.0332)
	1.0	0.2010	(0.0704)	0.2136	(0.0758)	0.1945	(0.0387)
−20	0.2	0.1585	(0.7470)	0.5169	(0.4119)	0.4075	(0.1688)
	0.4	0.1050	(0.0373)	0.2973	(0.0840)	0.2412	(0.0612)
	0.6	0.0996	(0.0399)	0.2247	(0.0505)	0.1987	(0.0421)
	0.8	0.1320	(0.0442)	0.2075	(0.0470)	0.1811	(0.0332)
	1.0	0.1896	(0.0425)	0.2136	(0.0758)	0.1945	(0.0387)
				$\hat{f}_B$			
0	0.2	0.2487	(0.3106)	0.2090	(0.2919)	0.2044	(0.1658)
	0.4	0.2228	(0.1542)	0.1565	(0.1340)	0.1531	(0.1212)
	0.6	0.2243	(0.1456)	0.1785	(0.1205)	0.1682	(0.1165)
	0.8	0.2893	(0.1232)	0.2336	(0.1170)	0.2354	(0.1032)
	1.0	0.3807	(0.1333)	0.3479	(0.0958)	0.3240	(0.0878)
−5	0.2	0.2360	(0.3556)	0.1718	(0.3229)	0.1387	(0.1188)
	0.4	0.1625	(0.2295)	0.1300	(0.0738)	0.1134	(0.0919)
	0.6	0.1815	(0.0990)	0.1601	(0.0718)	0.1511	(0.0660)
	0.8	0.2637	(0.1265)	0.2304	(0.0750)	0.2161	(0.0620)
	1.0	0.3525	(0.1161)	0.2274	(0.0827)	0.3061	(0.0698)
−20	0.2	0.1580	(0.1018)	0.1176	(0.0501)	0.0996	(0.0360)
	0.4	0.1250	(0.0563)	0.1050	(0.0404)	0.0959	(0.0361)
	0.6	0.1621	(0.0637)	0.1545	(0.0498)	0.1403	(0.0420)
	0.8	0.2469	(0.0698)	0.2252	(0.0516)	0.2162	(0.0469)
	1.0	0.3456	(0.0733)	0.3152	(0.0580)	0.2943	(0.0493)

Note: MADE values are shown for $\hat{f}_A$ and $\hat{f}_B$ with different sample sizes and different values of c and $h = A_0 n^{-1/20}$.

The simulation results are presented for $d = 0.25$ and $d = 0.45$ in Tables 1 and 2 (the median and the standard deviation (in parentheses) of 500 MADE values), respectively. First, we can see from Tables 1 and 2 that the mean squared error of 500 MADE values for both $\hat{f}_A$ and $\hat{f}_B$ decrease overall when the sample size increases. This is consistent with the asymptotic theory. Secondly, it can also be 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 $d = 0.25$ 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 $d = 0.25$ and $d = 0.45$. This is in line with the fact that

Table 2. MADE median and standard deviation: $d = 0.45$.

c	A_0	$n = 200$		$n = 500$		$n = 1{,}000$	
				$\hat{f}_A$			
0	0.2	0.2649	(0.5432)	0.2196	(0.4132)	0.1988	(0.2881)
	0.4	0.1913	(0.2987)	0.1511	(0.2751)	0.1270	(0.1632)
	0.6	0.1566	(0.1865)	0.1295	(0.1765)	0.1167	(0.1325)
	0.8	0.1670	(0.1443)	0.1501	(0.1349)	0.1336	(0.1124)
	1.0	0.2120	(0.1646)	0.1941	(0.1086)	0.1773	(0.0897)
−5	0.2	0.2563	(0.6392)	0.2221	(0.6098)	0.1394	(0.1099)
	0.4	0.1581	(0.4453)	0.1378	(0.4087)	0.0968	(0.0504)
	0.6	0.1345	(0.2514)	0.1143	(0.2301)	0.0903	(0.0420)
	0.8	0.1434	(0.1207)	0.1367	(0.1153)	0.1120	(0.0425)
	1.0	0.2032	(0.0824)	0.1820	(0.0876)	0.1622	(0.0414)
−20	0.2	0.1598	(0.1295)	0.1258	(0.0455)	0.1012	(0.0277)
	0.4	0.1044	(0.0366)	0.0858	(0.0287)	0.0722	(0.0249)
	0.6	0.1014	(0.0395)	0.0854	(0.0317)	0.0745	(0.0279)
	0.8	0.1294	(0.0446)	0.1157	(0.0355)	0.1086	(0.0297)
	1.0	0.1899	(0.0439)	0.1686	(0.0343)	0.1560	(0.0303)
				$\hat{f}_B$			
0	0.2	0.2422	(0.4126)	0.1827	(0.3761)	0.1796	(0.0987)
	0.4	0.2082	(0.2546)	0.1618	(0.2231)	0.1434	(0.0764)
	0.6	0.2249	(0.2237)	0.1793	(0.1925)	0.1588	(0.0432)
	0.8	0.2765	(0.1811)	0.2370	(0.1471)	0.2323	(0.0343)
	1.0	0.3673	(0.1867)	0.3205	(0.1228)	0.3127	(0.0675)
−5	0.2	0.2299	(0.3806)	0.2252	(0.3139)	0.1564	(0.0963)
	0.4	0.1551	(0.2005)	0.1492	(0.1738)	0.1201	(0.0737)
	0.6	0.1902	(0.1151)	0.1749	(0.1087)	0.1560	(0.0450)
	0.8	0.2422	(0.0971)	0.2426	(0.0901)	0.2242	(0.0325)
	1.0	0.3543	(0.1075)	0.3434	(0.0933)	0.3117	(0.0532)
−20	0.2	0.1657	(0.0909)	0.1305	(0.0546)	0.1061	(0.0379)
	0.4	0.1284	(0.0533)	0.1106	(0.0512)	0.1001	(0.0389)
	0.6	0.1689	(0.0648)	0.1534	(0.0560)	0.1426	(0.0399)
	0.8	0.2502	(0.0691)	0.2252	(0.0616)	0.2118	(0.0463)
	1.0	0.3531	(0.0759)	0.3211	(0.0619)	0.2980	(0.0483)

Note: MADE values are shown for $\hat{f}_A$ and $\hat{f}_B$ with different sample sizes and different values of c and $h = A_0 n^{-1/20}$.

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 as it is not necessary to worry too much about obtaining a rough estimate of the bandwidth.

Finally, to find out the effect of the appearance of a non-zero correlation between ϵ_t and u_t on the local linear estimator, we conduct the following simulation study. We choose the function $f_B(\cdot)$ and let $d = 0.25$ and $A_0 = 0.8$. We obtain the correlation pairs of (ϵ_t, u_t) by letting $u_t = \rho(\epsilon_t / sd(\epsilon_t)) + \sqrt{1-\rho^2}N(0,1)$, where ρ varies from 0.05 to 0.9 with step size 0.05. For each fixed ρ, we repeat the simulation 500 times, and the MADE values are calculated for each

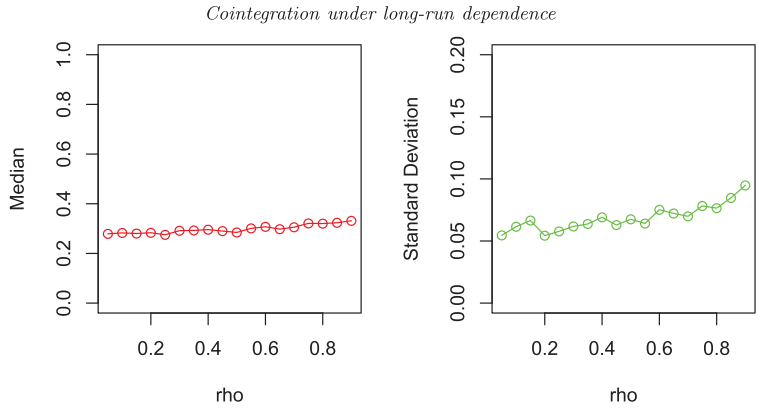

Figure 1. MADE median (left) and standard deviation (right). [Colour figure can be viewed at wileyonlinelibrary.com]

simulation. We display the median and standard deviation of these 500 MADE values against ρ in Figure 1. From the figure, we can see that the increasing correlation between ϵ_t and u_t does not change the performance of the local linear estimator too much in view of the flat median curve, while the standard deviation increases slightly but not too much. This demonstrates that our results might hold for a certain cross-correlation between u_t and ϵ_t.

4. DISCUSSION

In this paper, we have studied a nonparametric regression model for nearly integrated time series data with a possible long-range dependence. We have suggested using the local polynomial and local constant fitting schemes to estimate the nonparametric function and we have 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. However, 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 no longer normal but just a mixed normal associated with the local time of an O–U fractional Brownian motion. Moreover, we have shown that the asymptotic properties for both the local linear and local constant estimators are exactly the 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. Thirdly, an extension to allow for a certain cross-correlation between u_t and ϵ_t is warranted. 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.

ACKNOWLEDGEMENTS

Z. Cai's research was supported, in part, by the National Nature Science Foundation of China (NSFC) grants 71131008 (Key Project) and 71631004 (Key Project). B. Jing's research was supported, in part, by the Hong Kong Research Grants Council, grants HKUST6019/10P and

HKUST6019/12P. X. Kong's research was supported, in part, by NSFC 11571250, HSSYF of the Chinese Ministry of Education (12YJC910003) and PAPD of the Jiangsu Higher Education Institution. Z. Liu's work was partially supported by the Macau Science and Technology Development Fund (No. 078/2013/A3) and NSFC (No. 11401607).

REFERENCES

Bachmeier, L., S. Leelahanon and Q. Li (2006). Money growth and inflation in the United States. *Macroeconomic Dynamics 11*, 113–27.

Bandi, F. M. (2002). On persistence and nonparametric estimation (with an application to stock return predictability). Working paper, Graduate School of Business, University of Chicago.

Billingsley, P. (1999). *Convergence of Probability Measures* (2nd ed.). New York, NY: Wiley.

Buchmann, B. and N. H. Chan (2007). Asymptotic theory of least squares estimation for nearly unstable processes under strong dependence. *Annals of Statistics 35*, 2001–17.

Cai, Z. (2011). Nonparametric regression models with integrated covariates. In J. Jiang, G. G. Roussas and F. J. Samaniego (Eds.), *Nonparametric Statistical Methods and Related Topics: A Festschrift in Honor of Professor P. K. Bhattacharya on his 80th Birthday*, 257–75. Singapore: World Scientific.

Cai, Z. and Y. Hong (2009). Some recent developments in nonparametric finance. *Advances in Econometrics 25*, 379–432.

Cai, Z. and Y. Wang (2014). Testing predictive regression models with nonstationary regressors. *Journal of Econometrics 178*, 4–14.

Cai, Z. and L. Wu (2013). A consistent nonparametric test on nonlinear regression models with nearly integrated covariates. Working paper, University of North Carolina at Charlotte.

Cai, Z., J. Gu and Q. Li (2009a). Recent developments in nonparametric econometrics. *Advances in Econometrics 25*, 495–549.

Cai, Z., Q. Li and J. Park (2009b). Functional-coefficient models for nonstationary time series data. *Journal of Econometrics 148*, 101–13.

Cai, Z., Y. Wang and Y. Wang (2015). Testing instability in a predictive regression model with nonstationary regressors. *Econometric Theory 31*, 953–80.

Campbell, J. Y. and M. Yogo (2006). Efficient tests of stock return predictability. *Journal of Financial Economics 81*, 27–60.

Cavanagh, C. L., G. Elliott and J. H. Stock (1995). Inference in models with nearly integrated regressors. *Econometric Theory 11*, 1131–47.

Chang, Y. and E. Martinez-Chombo (2003). Electricity demand analysis using cointegration and error-correction models with time varying parameters: The Mexican case. Working paper, Department of Economics, Indiana University.

Chang, Y. and J. Park (2003). Index models with integrated time series. *Journal of Econometrics 114*, 73–106.

Chang, Y. and J. Park (2010). Endogeneity in nonlinear regressions with integrated time series. *Econometric Reviews 30*, 51–87.

Elliott, G. and J. H. Stock (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric Theory 10*, 672–700.

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

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Berlin: Springer.

Gao, J., D. Li and D. Tjøstheim (2011). Uniform consistency for nonparametric estimates in null recurrent time series. Working paper 0085, School of Economics, University of Adelaide.

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

Hong, S. H. and P. C. B. Phillips (2010). Testing linearity in cointegrating relations with an application to purchasing power parity. *Journal of Business and Economic Statistics 28*, 96–114.

Juhl, T. (2005). Functional coefficient models under unit root behavior. *Econometrics Journal 8*, 197–213.

Karlsen, H. and D. Tjøstheim (2001). Nonparametric estimation in null recurrent time series. *Annals of Statistics 29*, 372–416.

Karlsen, H. A., T. Myklebust and D. Tjøstheim (2007). Nonparametric estimation in a nonlinear cointegration type model. *Annals of Statistics 35*, 252–99.

Li, H., Z. Lin and C. Hsiao (2015). Testing purchasing power parity hypothesis: a semiparametric varying coefficient approach. *Empirical Economics 48*, 427–38.

Li, Q. and J. Racine (2007). *Nonparametric Econometrics: Theory and Applications*. Princeton, NJ: Princeton University Press.

Mandelbrot, B. B. and J. M. Van Ness (1968). Fractional Brownian motion, fractional noises and applications. *SIAM Review 10*, 422–37.

Miller, J. I. and J. Park (2010). Nonlinearity, nonstationarity, and thick tails: how they interact to generate persistence in memory. *Journal of Econometrics 155*, 83–89.

Park, J. and S. B. Hahn (1999). Cointegrating regressions with time varying coefficients. *Econometric Theory 15*, 664–703.

Phillips, P. C. B. and J. Park (1998). Nonstationary density and kernel autoregression. Cowles Foundation Discussion Paper 1181, Yale University.

Polk, C., S. Thompson and T. Vuolteenaho (2006). Cross-sectional forecasts of the equity premium. *Journal of Financial Economics 81*, 101–141.

Rossi, B. (2007). Expectation hypothesis tests and predictive regressions at long horizons. *Econometrics Journal 10*, 1–26.

Sun, Y., Z. Cai and Q. Li (2013). Semiparametric functional coefficient models with integrated covariates. *Econometric Theory 29*, 659–72.

Sun, Y., Z. Cai and Q. Li (2016). A consistent nonparametric test on semiparametric smooth coefficient models with integrated time series. *Econometric Theory 32*, 988–1022.

Torous, W., R. Valkanov and S. Yan (2004). On predicting stock returns with nearly integrated explanatory variables. *Journal of Business 77*, 937–66.

Wang, Q. (2014). Martingale limit theorem revisited and nonlinear cointegrating regression. *Econometric Theory 30*, 509–35.

Wang, Q. and N. Chan (2014). Uniform convergence rates for a class of martingales with application in nonlinear cointegrating regression. *Bernoulli 20*, 207–30.

Wang, Q. and P. C. B. Phillips (2009a). Asymptotic theory for local time density estimation and nonparametric cointegrating regression. *Econometric Theory 25*, 710–38.

Wang, Q. and P. C. B. Phillips (2009b). Structural nonparametric cointegrating regression. *Econometrica 77*, 1901–48.

Wang, Q. and P. C. B. Phillips (2011). Asymptotic theory for zero energy density estimation with nonparametric regression applications. *Econometric Theory 27*, 235–59.

Wang, Q. and P. C. B. Phillips (2012). A specification test for nonlinear nonstationary models. *Annals of Statistics 40*, 727–58.

Wang, Q. and Y. Wang (2013). Nonparametric cointegrating regression with NNH errors. *Econometric Theory 29*, 1–27.

Wang, Q., Y-X. Lin and C. M. Gulati (2003a). Strong approximation for long memory processes with applications. *Journal of Theoretical Probability 16*, 377–89.

Wang, Q., Y-X. Lin and C. M. Gulati (2003b). Asymptotics for general fractionally integrated processes with applications to unit root tests. *Econometric Theory 69*, 117–61.

Xiao, Z. (2009). Functional-coefficient cointegration models. *Journal of Econometrics 152*, 81–92.

Zhu, F., Z. Cai and L. Peng (2014). Perceptive regressions for macroeconomic data. *Annals of Applied Statistics 8*, 577–94.

APPENDIX

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

Proof of Lemma 2.1: Let $c_k = \Gamma(d+k)/[\Gamma(d)\Gamma(k+1)]$. Then $c_k \sim k^{d-1}/\Gamma(d)$, as $k \to \infty$. We make a convention that $c_k = 0$ for $k < 0$. By Theorem 2.1 in Wang et al. (2003b),

$$V_n(s) \Rightarrow W_d(s), \quad 0 \le s \le 1. \tag{A.1}$$

However, by the functional central limit theorem, we have

$$U_n(s) \Rightarrow \sigma_u \tilde{W}_s, \tag{A.2}$$

where $\tilde{W}$ is a standard Brownian motion. Because two marginal sequences of processes are tight, the joint sequences of processes must also be tight. Then, it suffices to prove the convergence in distribution of finite vectors at different time points. Because the limiting joint distribution is multivariate normal and (U_n, V_n) has the form of sums, it is enough to prove the pairwise convergence of their covariances. Therefore, it is enough to show the following

$$E[U_n(s_1)V_n(s_2)] \to \sigma_u E[\tilde{W}_{s_1} W_d(s_2)] = \frac{\rho b_\psi}{k(d)\Gamma(2+d)}(s_2^{d+1} - ((s_2 - s_1) \vee 0)^{d+1}). \tag{A.3}$$

In fact, suppose $s_1 < s_2$

$$\begin{aligned}
E[U_n(s_1)V_n(s_2)] &= \frac{1}{k(d)n^{1+d}} \sum_{t=1}^{[ns_1]} \sum_{l=1}^{[ns_2]} E[u_t \epsilon_l] \\
&= \frac{1}{k(d)n^{1+d}} \sum_{t=1}^{[ns_1]} \sum_{l=t+1}^{[ns_2]} E[u_t \sum_{j=0}^{\infty} \psi_j \sum_{k=0}^{\infty} c_k \xi_{l-j-k}] \\
&= \frac{1}{k(d)n^{1+d}} \sum_{t=1}^{[ns_1]} \sum_{l=t+1}^{[ns_2]} \sum_{j=0}^{\infty} \psi_j c_{l-j-(t+1)} E[u_t \xi_{t+1}] \\
&\equiv \frac{1}{k(d)n^{1+d}} \sum_{t=1}^{[ns_1]} \sum_{l=t+1}^{[ns_2]} \sum_{j=0}^{\infty} \psi_j c_{l-j-(t+1)} \rho_{t+1} \\
&= \frac{1}{k(d)n^{1+d}} \sum_{j=0}^{\infty} \psi_j \sum_{t=1}^{[ns_1]} \sum_{l=t+1}^{[ns_2]} c_{nl/n-n(t+1)/n-j} \rho_{n(t+1)/n} \\
&= \frac{1}{k(d)n^{1+d}} \sum_{j=0}^{\infty} \psi_j n^2 \int_{2/n}^{s_1} \int_{[nv]/n}^{s_2} c_{[nu]-[nv]-j} \rho_{[nv]} du dv
\end{aligned}$$

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

$$\sim \frac{\rho b_\psi}{k(d)\Gamma(2+d)} (s_2^{d+1} - (s_2 - s_1)^{d+1}), \tag{A.4}$$

which verifies (A.3). The case for $s_1 \geq s_2$ can be done similarly. □

Proof of Lemma 2.2: Following the ideas as in Buchmann and Chan (2007), we have

$$\begin{aligned}
\frac{1}{\gamma_n} x_{[ns]} &= \frac{\beta^{[ns]}}{\gamma_n} x_0 + \frac{1}{\gamma_n} \sum_{k=1}^{[ns]} \beta^{[ns]-k} \epsilon_k \\
&= \frac{\beta^{[ns]}}{\gamma_n} x_0 + \sum_{k=1}^{[ns]} \beta^{[ns]-k} \left(V_n\left(\frac{k}{n}\right) - V_n\left(\frac{k-1}{n}\right) \right) \\
&= \frac{\beta^{[ns]}}{\gamma_n} x_0 + V_n\left(\frac{[ns]}{n}\right) - \beta^{[ns]-1} V_n(0) - \beta^{[ns]} \sum_{k=1}^{[ns]-1} (\beta^{-(k+1)} - \beta^{-k}) V_n\left(\frac{k}{n}\right) \\
&= \frac{\beta^{[ns]}}{\gamma_n} x_0 + V_n\left(\frac{[ns]}{n}\right) + \beta^{[ns]} n \log(\beta) \sum_{k=1}^{[ns]-1} \int_{k/n}^{(k+1)/n} V_n(u) \beta^{-nu} du \\
&\equiv I_s^n - II_s^n + III_s^n + IV_s^n + V_s^n,
\end{aligned} \tag{A.5}$$

where

$$I_s^n = \frac{1}{\gamma_n} \beta^{[ns]} x_0, \quad II_s^n = \left(\int_{[ns]/n}^{s} + \int_0^{1/n} \right) (n \log(\beta)) \beta^{[ns]-nu} V_n(u) du;$$

$$III_s^n = V_n\left(\frac{[ns]}{n}\right) - V_n(s), \quad IV_s^n = \int_0^s [(n \log(\beta)) \beta^{[ns]-nu} + c e^{-c(s-u)}] V_n(u) du;$$

$$V_s^n = V_n(s) - c \int_0^s e^{-c(s-u)} V_n(u) du.$$

By the condition in Lemma 2.2, $I_s^n \xrightarrow{p} 0$. By Theorem 2.1 of Wang et al. (2003a), $V_n(\cdot) \Rightarrow W_d(\cdot)$ in $D[0, 1]$. Then, by the Skorohod representation theorem, there exists V_n^* and W_d^* such that $V_n^* \stackrel{d}{=} V_n$ and $W_d^* \stackrel{d}{=} W_d$, and $\sup_{0 \leq s \leq 1} |V_n^*(s) - W_d^*(s)| \to 0$. To simplify notation, let $V_n = V_n^*$ and $W_d = W_d^*$. Then

$$V_s^n - W_{c,d}(s) = (V_n(s) - W_d(s)) - c \int_0^s e^{-c(s-u)} (V_n(u) - W_d(u)) du. \tag{A.6}$$

This shows that $\sup_{0 \leq s \leq 1} |V_s^n - W_{c,d}(s)| \to 0$, which implies that $V_s^n \Rightarrow W_{c,d}$ 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 $[0, 1]$, $II_s^n \xrightarrow{p} 0$, and $III_s^n \xrightarrow{p} 0$. By the definition $\beta = 1 - c/n$, $\beta^{[ns]-nu}$ is uniformly close to $e^{-c(s-u)}$ for $0 \leq s \leq 1$, $0 \leq u \leq s$, and $n \log \beta$ converges to $-c$. Hence, the product of $\beta^{[ns]-nu}$ and $n \log \beta$ converges to $-ce^{-c(s-u)}$ uniformly in $0 \leq s \leq 1$, $0 \leq u \leq s$. However, for large enough n, on the enlarged probability space, $|V_n(u) - W_d(u)| \leq \epsilon$ for any $\epsilon > 0$. This implies $IV_s^n \to 0$ uniformly in s to 0. Combining the above arguments, Lemma 2.2 is proved. □

Before we prove the main results of this paper, we first give a useful proposition, the proof of which relies on checking the conditions of Theorem 2.1 of Wang and Phillips (2009a). Note that the proof of Corollary 2.2 in their paper is not applicable as there d is assumed to be equal to 0.

PROPOSITION A.1. *Let $F(\cdot)$ be a continuous function with compact support. Under Assumptions 2.1, 2.3 and 2.4, if $\gamma_n/(nh) \to 0$, then for any $0 \le s \le 1$, we have (a) if $\int_R F(u)du \neq 0$,*

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

(b) 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: We prove Proposition A.1(a) by checking the conditions given in the very general Theorem 2.1 in Wang and Phillips (2009a), which cannot be obtained with several equations. We define $d_{l,k,n} = c_{l-k}/c_n$ for any $0 < k < l \le n$. Then it suffices to prove Assumption 2.3(a) and (b) of Wang and Phillips (2009a). Their Assumption 2.3(a) can be verified directly by using the definition of $d_{l,k,n}$ and the fact that $c_n \sim n^{d-1}/\Gamma(d)$. To prove their Assumption 2.3(b), let $g(j) = \sum_{i=1}^{j} c_i \beta^{j-i}$. Then

$$x_l - \beta^{l-k}x_k = \sum_{j=k+1}^{l}\left(\sum_{m=0}^{l-j}\psi_m g(l-j-m)\xi_j\right) + \tilde{x}_k, \tag{A.7}$$

where $\tilde{x}_k$ is some random variable adapted to $\mathcal{F}_{k-1}$. Now we need to prove that

$$\frac{1}{\gamma_{l-k}}\sum_{j=k+1}^{l} a(l-j)\xi_j$$

has integrable characteristic function where $a(j) := \sum_{m=0}^{j}\psi_m g(j-m)$, i.e., $\int |f_{l,k}(u)|du < \infty$. To this end, we need the following two facts.

(a) For large enough $l-k$, there exist $0 < c_1^* < c_2^* < \infty$ such that $c_1^*/\sqrt{l-k} < a(l-j)/\gamma_{l-k} < c_2^*/\sqrt{l-k}$ when $((l-k)/2) \le l-j \le l-k-1$.
(b) For some $\delta_0 > 0$, there exists a $0 < \eta < 1$ such that

$$|\phi(t)| = |Ee^{\sqrt{-1}t\xi_0}| \le \begin{cases} e^{-t^2/4}, & \text{for } |t| \le \delta_0 \\ \eta, & |t| \ge \delta_0. \end{cases}$$

Fact (b) is due to the existence of second moment of ξ_0. We will prove fact (a) later. In view of facts (a) and (b), for some $\delta > 0$, we have

$$\begin{aligned}\int |f_{l,k}(\theta)|d\theta &\le \int \prod_{j=k+1}^{(l+k)/2} |Ee^{\sqrt{-1}\theta(a(l-j)/(\gamma_{l-k}))\xi_j}|d\theta \\ &= \left(\int_{|\theta|\le\delta\sqrt{l-k}} + \int_{|\theta|>\delta\sqrt{l-k}}\right)\prod_{j=k+1}^{(l+k)/2} |Ee^{\sqrt{-1}\theta(a(l-j)/(\gamma_{l-k}))\xi_j}|d\theta \\ &\le \int e^{-\theta^2/8}d\theta + C\eta^{(l-k)/2-1}\int |Ee^{\sqrt{-1}\theta\xi_0}|d\theta < \infty, \end{aligned} \tag{A.8}$$

in view of (a) and (b) and Assumption 2.4. Therefore, conditional on $\mathcal{F}_{k-1}$, $(x_l - x_k/\gamma_n)/d_{l,k,n}$ has a density function $h_{l,k,n}(x)$ that is uniformly bounded by a constant. Hence the first claim of Theorem 2.1(b) of Wang and Phillips (2009a) is verified. The second claim can be proved using the same lines as in the proof of Corollary 2.2 of the same paper.

Now we return to show fact (a). Recall that $((l-k)/2) \le l-j \le l-k-1$ and $l-k$ is large enough. Let $\lambda_{l-k} \le l-k$ satisfying $\lambda_{l-k}/(l-k) \to 0$ and $\lambda_{l-k} \to \infty$. By definition of $a(l-j)$,

$$a(l-j) = \left(\sum_{m=0}^{\lambda_{l-k}} + \sum_{m=\lambda_{l-k}+1}^{l-j}\right)\psi_m \sum_{i=0}^{l-j-m} c_i \beta^{l-j-m-i} := a_1(l-j) + a_2(l-j). \tag{A.9}$$

The second sum in the last equation is bounded by

$$\sum_{m=\lambda_{l-k}}^{l-j} |\psi_m| \sum_{i=0}^{l-j} c_i \le C(l-k)^d \sum_{m=\lambda_{l-k}+1}^{l-k} |\psi_m|.$$

By Assumption 2.1, we have

$$|a_2(l-j)/\gamma_{l-k}| \le C\frac{1}{\sqrt{l-k}} o(1). \tag{A.10}$$

$a_1(l-k)$ can be further decomposed as follows.

$$a_1(l-j) = \sum_{m=0}^{\lambda_{l-k}} \psi_m \left(\sum_{i=0}^{l-j} - \sum_{i=l-j-m+1}^{l-j}\right) c_i \beta^{l-j-m-i} := a_{1,1}(l-j) - a_{1,2}(l-j). \tag{A.11}$$

By definition of λ_{l-k}, we have

$$|a_{1,2}(l-j)| \le C \sum_{m=0}^{\lambda_{l-k}} |\psi_m| c_{l-k} \lambda_{l-k} \le C\lambda_{l-k}(l-k)^{(d-1)}. \tag{A.12}$$

Therefore

$$|a_{1,2}(l-j)/\gamma_{l-k}| \le C\lambda_{l-k}(l-k)^{-3/2}. \tag{A.13}$$

Lastly,

$$a_{1,1}(l-j)/\gamma_{l-k} = \frac{1}{\gamma_{l-k}} \sum_{m=0}^{\lambda_{l-k}} \psi_m \beta^{-m} \sum_{i=0}^{l-j} c_i \beta^{l-j-i} \in \left[\frac{c_1^*}{\sqrt{l-k}}, \frac{c_2^*}{\sqrt{l-k}}\right], \tag{A.14}$$

where we have used Assumption 2.1, $0 < \beta < 1$, and the fact that $\sum_{i=0}^{l-j} c_i \sim C(l-j)^d$. A combination of (A.10)–(A.14) produces fact (a).

Proposition A.1(b) can be proved using (A.10)–(A.14) and the same lines as in the proof of Theorem 2.1 of Wang and Phillips (2011). □

Proof of Theorem 2.1: Let

$$A_n = \sum_{t=1}^{n} \begin{pmatrix} 1 & x_t - x \\ x_t - x & (x_t - x)^2 \end{pmatrix} K_h(x_t - x)$$

and

$$B_n = \sum_{t=1}^{n} \begin{pmatrix} 1 \\ x_t - x \end{pmatrix} y_t K_h(x_t - x).$$

Then, for some random number $\xi_t \in (x, x_t)$, and

$$H_n = \begin{pmatrix} 1 & 0 \\ 0 & h^{-1} \end{pmatrix},$$

$$A_n^{-1}B_n - (f(x),\, f'(x))^{\mathrm{T}} = A_n^{-1}\left(\sum_{t=1}^{n}\left(\frac{1}{2}f''(\xi_t)(x_t-x)^2 + u_t\right)(1,\, x_t-x)^{\mathrm{T}}K_h(x_t-x)\right)$$
$$=: A_n^{-1}(B_{n1}+B_{n2}) = (H_nA_n)^{-1}H_n(B_{n1}+B_{n2}), \tag{A.15}$$

and

$$H_nA_nH_n = \sum_{t=1}^{n}\begin{pmatrix} 1 & \dfrac{x_t-x}{h} \\ \dfrac{x_t-x}{h} & \left(\dfrac{x_t-x}{h}\right)^2 \end{pmatrix} K_h(x_t-x). \tag{A.16}$$

By (A.16) and Proposition A.1,

$$\frac{n}{\gamma_n}H_n^{-1}(H_nA_n)^{-1} \xrightarrow{d} L^{-1}(1,0)\begin{pmatrix}\mu_0 & \mu_1\\ \mu_1 & \mu_2\end{pmatrix}^{-1}. \tag{A.17}$$

Similarly, we have

$$\frac{\gamma_n}{nh^2}H_nB_{n1} \xrightarrow{d} \frac{1}{2}f''(x)L(1,0)(\mu_2,\mu_3)^{\mathrm{T}}. \tag{A.18}$$

Motivated by (A.17) and (A.18), by Proposition A.1 again, we have

$$(1,0)H_n^{-1}A_n^{-1}B_{n1} - h^2B(x) = (1,0)^{\mathrm{T}}A_n^{-1}B_{n1} - h^2B(x) = o_P(h^2). \tag{A.19}$$

Let

$$K_n = \sqrt{\frac{\gamma_n h}{n}}u_tK_h\left(\frac{x_t-x}{h}\right)H_n(1,\, x_t-x)^{\mathrm{T}}.$$

By Proposition A.1, (A.19) and Assumption 2.2, we have

$$\sum_{t=1}^{n}E\left[K_n^{\otimes 2}1_{|\{|K_n>\epsilon|\}}|\mathcal{F}_{t-1}\right] \xrightarrow{p} 0, \qquad \text{for any } \epsilon>0,$$

and

$$\sum_{t=1}^{n}E\left[K_n^{\otimes 2}|\mathcal{F}_{t-1}\right] \xrightarrow{d} \sigma_u^2L(1,0)\begin{pmatrix}\nu_0 & \nu_1\\ \nu_1 & \nu_2\end{pmatrix}. \tag{A.20}$$

Motivated by (A.17) and (A.20), a direct use of Theorem 2.1 of Wang (2014) yields

$$\sqrt{\frac{nh}{\gamma_n}}(1,0)(H_nA_n)^{-1}H_nB_{n2} = \frac{n}{\gamma_n}(1,0)H_n^{-1}(H_nA_n)^{-1}K_n$$
$$\xrightarrow{d} \frac{\sigma_u[(1,0)S^{-1}S^*S^{-1}(1,0)^{\mathrm{T}}]^{1/2}z}{\sqrt{L(1,0)}}, \tag{A.21}$$

where z is a standard normal random variable. Together with (A.19), this completes the proof of Theorem 2.1. □

Proof of Theorem 2.3: $\tilde{f}(x) - f(x)$ can be decomposed as a bias term plus a variance term as follows

$$\tilde{f}(x) - f(x) = \frac{\sum_{t=1}^{n} \left(u_t + f'(x)(x_t - x) + (f''(\xi)/2)(x_t - x)^2\right) K((x_t - x)/h)}{\sum_{t=1}^{n} K((x_t - x)/h)}, \tag{A.22}$$

where ξ_t is some number between x and x_t. Similar to obtaining (A.21),

$$\sqrt{\frac{nh}{\gamma_n}} \frac{\sum_{t=1}^{n} u_t K((x_t - x)/h)}{\sum_{t=1}^{n} K((x_t - x)/h)} \Rightarrow \frac{\sigma_u \sqrt{v_0(K)}}{\sqrt{L(1,0)}} z_3, \tag{A.23}$$

where z_3 is independent of W_d, so the right-hand side of (A.23) is a mixed normal random variable. Similar to obtaining (A.19),

$$\frac{(1/2)\sum_{t=1}^{n} f''(\xi_t)(x_t - x)^2 K((x_t - x)/h)}{\sum_{t=1}^{n} K((x_t - x)/h)} - h^2 B(x) = o_P(h^2), \tag{A.24}$$

while

$$\frac{\sum_{t=1}^{n} f'(x)(x_t - x) K((x_t - x)/h)}{\sum_{t=1}^{n} K((x_t - x)/h)} = \begin{cases} hf'(x)\mu_1 & \mu_1 \neq 0; \\ O_P(h\sqrt{(\gamma_n h)/n}) & \mu_1 = 0. \end{cases} \tag{A.25}$$

A combination of (A.23), (A.24), (A.25) and Theorem 2.1 in Wang (2014) finishes the proof of Theorem 2.3. □

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication files