Journal of Econometrics 222 (2021) 909–932

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A weighted sieve estimator for nonparametric time series models with nonstationary variables

Chaohua Dong [a,*], Oliver Linton [b], Bin Peng [c]

[a] Zhongnan University of Economics and Law, China
[b] University of Cambridge, UK
[c] Monash University, Australia

ARTICLE INFO

Article history:
Received 10 May 2019
Received in revised form 3 March 2020
Accepted 11 March 2020
Available online 17 September 2020

JEL classification:
C12
C22
C32

Keywords:
Nonparametric regression
Nonstationary variable
Sieve estimation
Stationary variable
Time trend
Unbounded support
Weighted least squares

ABSTRACT

We study a class of nonparametric regression models that includes deterministic time trends and both stationary and nonstationary stochastic processes (whose shocks are allowed to be mutually correlated). We propose a unified approach to estimation based on the weighted sieve method to tackle the issue of unbounded support of the covariates. This approach improves on the existing technology in terms of some key regularity conditions such as moment conditions and the α-mixing coefficients for the stationary process. We establish self-normalized central limit theorems for the sieve estimator and other related quantities. Monte Carlo simulation confirms the theoretical results. We use our methodology to study the effect of CO_2 and solar irradiance on global sea level rise.

1. Introduction

Nonparametric methods are widely used in various areas of economics and finance due to their flexibility and generality. New methodology such as machine learning is in many cases building on the tools and devices developed over the last half century by work on nonparametrics. See Härdle and Linton (1994), Chen (2007) and Li and Racine (2007) for surveys on the theoretical tools and practical applications. For stationary weakly dependent time series the theoretical properties of kernel and sieve methods are well understood. For nonstationary or strongly dependent time series the theory is still incomplete, despite significant works in this direction. Nonstationarity leads to slower rates of convergence, unlike the stationary case (Stone, 1980), although limiting distributions often remain normal or mixed normal allowing standard inference techniques. What remains relatively unstudied is the multiple covariate case where some of the variables may be strongly dependent or even nonstationary and others are stationary or are deterministic trends. This paper aims to address this issue.

In many applications, the outcome variable may be affected by multiple types of variables. For example, consumers' consumption may be determined by their income and the interest rate; a stock price may be affected by the prices

* Correspondence to: School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China.
E-mail address: chaohuadong@outlook.com (C. Dong).

https://doi.org/10.1016/j.jeconom.2020.03.024
0304-4076/

of related stocks and the trading volumes for them; global sea level rise is primarily attributed to the rise of global temperature caused by too much carbon dioxide (chemical formula CO_2) and solar irradiance (we shall return to this question in the empirical study section). Some of these variables are stationary while others are nonstationary. Besides, all of them may contain deterministic components that change over time. These observations motivate us to incorporate different types of variables into a single regression model. We also expect in many cases, for example climate modeling, that the effects of interest are nonlinear, and that linear models are neither scientifically grounded nor adequate approximations.

In this paper, we consider a class of nonparametric regression models that incorporate all kinds of variables aforementioned. Suppose that

$$y_t = m_0(\tau_t, z_t, x_t) + e_t, \qquad t = 1, \ldots, n, \tag{1.1}$$

where m_0 is an unknown function, $\tau_t = t/n$, z_t a stationary process, x_t an integrated process, while e_t is an error term that satisfies at least $\mathbb{E}[e_t|z_t, x_t] = 0$. Here, $m_0(\cdot)$ is defined on $[0, 1] \times V_z \times \mathbb{R}$, where V_z is the support of z_1. The integrated process x_t definitely has an unbounded support due to the divergent nature of the I(1) process whereas the stationary z_t may reasonably be supposed to have either bounded or unbounded support depending on the application. Here, we allow z_t to assume values in $(-\infty, \infty)$ or $(0, \infty)$. By contrast, researchers normally require the support of the regressor to be compact when the sieve method is used. See, e.g. Assumption 8 of Newey (1997, p.156) and Assumption 3.1 of Ai and Chen (2003, p.1803). This restriction excludes the frequently encountered and important normal random variable, as well as prima facie unit root processes.

We propose a weighted least squares sieve method to deal with the issues of unbounded support of variables and nonstationarity. The kernel methodology developed in Wang and Phillips (2009) faces some difficulty in this case, as we discuss below. Our methodology is novel and simpler than earlier approaches. We make use of the density of the Hilbert space that includes the regression function in a weighted least squares sieve method. By contrast, both Chen and Christensen (2015) and Hansen (2015) consider weighted (least squares) estimators in stationary settings but their weighting is more akin to trimming, i.e., restricting attention to expanding compact support sets.

Our weighting scheme facilitates the establishment of our asymptotic theory in the presence of globally nonstationary variables that may have unbounded support. Additionally, as explained below Assumption B, the weighting scheme also allows us to very much weaken the condition on the α-mixing coefficients for the stationary process compared with the literature. As a result, the model can deal with a much broader range of stationary variables. Furthermore, we allow the shocks of the stationary and integrated variables to be correlated. We establish pointwise self-normalized central limit theorems for the estimated regression functions and various functionals thereof and provide feasible inference procedures. The rates of convergence obtained are generally slower than in the purely stationary or deterministic case (Stone, 1980, 1985), and for the regression function itself the rates are determined by the slowest component, the nonstationary part. However, certain marginal effects may converge at rates corresponding to the stationary case. In practice both z_t and x_t might be vectors, so the setting of model (1.1) needs to be reformulated to deal with these situations. We give a brief discussion on this issue in the conclusion section.

Our simulation evidence shows that our estimation procedures work satisfactorily in finite sample situations. There is much recent work by econometrics on climate modeling, see Atak et al. (2011) for example and the special issue of the Journal of Econometrics (No.1, Vol.214, 2020). We apply our methodology to an important question in climate econometrics. Specifically, we study the effect of CO_2 and solar irradiance on global sea level rise (hereafter, SLR) using annual data from 1880 to 2005. Visser et al. (2015) review the work on modeling of SLR; their Table 1 reveals that a wide range of methods have been used for the trend, including: kernel methods, wavelets, MARS regression, neural networks, and spline methods, but in most cases covariate effects have been treated linearly. We allow both trend and covariates to affect SLR nonlinearly, which is more consistent with the type of nonlinear differential equation models with possibly chaotic dynamic that are favored by meteorological offices concerned with numerical weather prediction, see Lynch (2006). We find evidence of nonlinearity and indeed some interaction effects between the main variables.

Literature Review. The class of nonstationary processes is extremely broad, and different types of nonstationarity can generate quite different behavior and require quite different analytical techniques. There are two main approaches to depicting the structure of nonstationary data. One is the unit root theory for integrated time series (or similar techniques for fractional integrated time series that covers unit root process as a special case). This theory and associated techniques are studied and developed by Park and Phillips (1999, 2001), Marinucci and Robinson (2001), Wang and Phillips (2009), Hualde and Robinson (2011), Wang (2015) and Dong et al. (2016), among others. A second setting to describe nonstationarity is based on the class of null recurrent Markov processes, that was mainly developed by Karlsen and Tjøstheim (2001), Karlsen et al. (2007), Mykelbust et al. (2012) and Li et al. (2016b). The theory of linear models including all these types of variables is well understood, but the analysis of nonlinear models is at an earlier stage and presents certain challenges. In an unpublished paper, Schienle (2008) considers an additive nonparametric regression model with multiple nonstationary variables based on the kernel backfitting methodology. There are some papers that consider diverse types of variables in one model. Although they have accommodated all the three types of regressors, Chang et al. (2001) study a nonlinear parametric model where all functions are supposed to be known up to a finite dimensional vector of parameters; Park and Hahn (1999) investigate linear regression with an I(1) regressor and time varying coefficients depending on a fixed design; Xiao (2009) studies a functional-coefficient cointegration

regression where the coefficients depend on a stationary variable and the regressor is an I(1) vector; Cai et al. (2009) study a similar model with more flexibility; Wang (2015) considers the estimation of a nonparametric regression model with both stationary and nonstationary variables via the kernelmethod; Li et al. (2016a) investigate the convergence of sample covariances that have an I(1) process and a variable that can be a fixed design or a random design (but not both); more recently, Dong and Linton (2018) consider an additive nonparametric model with the three types of variables considered here and use standard least squares sieve estimation. Additivity is a strong assumption ruling out interaction effects and can be violated by some datasets. In recent years some effort has been devoted to relax the compact support assumption in nonparametric estimation. Chen and Christensen (2015) establish the uniform consistency with optimal rate for sieve estimators with weakly dependent data. They propose a sequence of expanding compact sets to approximate the unbounded support.

Throughout the paper, $\|u\|$ is Euclidean norm for any vector and $\|A\| = \sqrt{\mathrm{tr}(A^{\intercal}A)}$ is entry-wise norm for any matrix; $\int f(x)dx$ is an integral on the entire $\mathbb{R}$; $A_n \asymp B_n$ means that A_n/B_n is bounded from below and above uniformly in n; C can be different constant at each appearance.

The rest of the paper is organized as follows. Section 2 gives assumptions and the estimation procedure; Section 3 presents the asymptotic theory for the estimator proposed in the preceding section; Section 4 shows the results of numerical experiments followed by the empirical study in Section 5; and Section 6 concludes. All technical lemmas are given in Appendix A and the main results are proven in Appendix B. The proofs of lemmas, some secondary experiment results and a discussion of heteroscedasticity setting are shown in the supplementary material file.

2. Assumptions and estimation procedure

2.1. Assumptions

We first give the structure of the integrated regressor x_t.

Assumption A.

A.1 Let $\{\epsilon_j, -\infty < j < \infty\}$ be a scalar sequence of independent and identically distributed random variables having an absolutely continuous distribution with respect to the Lebesgue measure and satisfying $\mathbb{E}[\epsilon_1] = 0, \mathbb{E}[\epsilon_1^2] = 1, \mathbb{E}|\epsilon_1|^{q_1} < \infty$ for some $q_1 \geq 4$. The characteristic function of ϵ_1 satisfies that $\int |\lambda||\mathbb{E}\exp(i\lambda\epsilon_1)|d\lambda < \infty$.

A.2 Let $w_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$, where $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ and $\psi := \sum_{j=0}^{\infty} \psi_j \neq 0$.

A.3 For $t \geq 1$, $x_t = x_{t-1} + w_t$, and $x_0 = O_P(1)$.

The conditions of Assumption A are commonly used in the literature on nonstationary unit root time series (see, e.g. Park and Phillips, 1999, 2001, Wang and Phillips, 2009, Dong et al., 2016). The innovation variables $\{\epsilon_j\}$ are building blocks for the linear process w_t from which the regressor is integrated. All properties for x_t given in Lemma A.1 that are crucial for our theoretical development are derived from the $I(1)$ structure and Condition A.1 postulated for the innovation.

From the structure of x_t, we have $d_t^2 := \mathbb{E}(x_t^2) = \psi^2 t(1 + o(1))$ when $t \to \infty$ simply by virtue of the Beveridge–Nelson decomposition for w_t (Phillips and Solo, 1992, p. 972). More importantly, because of $x_t = O_P(\sqrt{t})$, the third argument of $m_0(\cdot, \cdot, \cdot)$ has to have $\mathbb{R}$ as its support.

Assumption B.

B.1 Suppose that $z_t = \rho(\epsilon_t, \ldots, \epsilon_{t-d+1}; \eta_t)$ with fixed nonnegative integer d and measurable function $\rho : \mathbb{R}^{d+1} \mapsto \mathbb{R}$, where the sequence $\{\eta_t\}$ is independent of $\{\epsilon_j\}$, and z_t has finite second moment; moreover, suppose that $\{\eta_t\}$ is a strictly stationary α-mixing process with mixing coefficients $\alpha(i)$ such that $\sum_{i=1}^{\infty} \alpha(i) < \infty$.

B.2 Suppose that f is a density function on V_z, and if V_z is a bounded interval it is continuous in the interior of V_z and satisfies additionally $f(z) \geq c > 0$ for some constant c. Suppose also that there exists an orthonormal function sequence $\{p_i(\cdot)\}$ in the space $L^2(V_z, f(z))$ such that $\sup_{z \in V_z} \sup_{i \geq 0} |p_i(z) f^{1/2}(z)| < \infty$.

B.3 Suppose that e_t and the filtration $\mathcal{F}_{t,n} = \sigma(z_{j+1}, e_j, j \leq t; x_1, \ldots, x_n)$ form a martingale difference sequence such that almost surely $\mathbb{E}(e_t^2|\mathcal{F}_{t-1,n}) = \sigma_e^2$ and $\max_{1 \leq t \leq n} \mathbb{E}(|e_t|^{q_2}|\mathcal{F}_{t-1,n}) \leq C < \infty$ for some $q_2 \geq 4$.

Condition B.1 allows z_t to be correlated with x_t by sharing the same shocks $\epsilon_t, \ldots, \epsilon_{t-d+1}$, but in the special case $d = 0$, where $\rho(\epsilon_t, \ldots, \epsilon_{t-d+1}; \eta_t) \equiv \tilde{\rho}(\eta_t)$, they could be mutually independent. By construction, z_t itself is strictly stationary and α-mixing and its mixing coefficients satisfy the same property as $\{\eta_t\}$. Here, the condition $\sum_{i=1}^{\infty} \alpha(i) < \infty$ is very much weaker than the common requirement in the literature where researchers usually impose that $\sum_{i=1}^{\infty} \alpha^{\delta/(2+\delta)}(i) < \infty$ for some $\delta > 0$, which implies a much quicker decay rate of $\alpha(i)$ when i increases. See, for example, Assumption A6 in Cai et al. (2009, p.103) and Assumption 1 in Dong et al. (2015, p.303). It will be made clear that this is due to the use of the weighted least squares in our estimation procedure. Precisely, the use of the weight makes all variables bounded so that Billingsley's inequality, $|\mathrm{cov}(X, Y)| \leq 4\|X\|_{\infty}\|Y\|_{\infty}\alpha$, is applicable. See Bosq (1996, p.20). Hence, the summability of $\alpha(i)$ is sufficient.

The density f introduced in B.2 is not necessarily the actual Lebesgue density of z_t, and the condition $f(z) \geq c > 0$ is satisfied by several orthonormal sequences when its support is compact. For example, $f(z) = (1-z^2)^{-1/2} \geq 1$ on $[-1, 1]$ for Chebyshev polynomials of the first kind; $f(z) \equiv 1$ for the sequence $\{\varphi_j(r), j \geq 1\}$ defined in the next subsection. It is a user-chosen density defined on V ($\equiv V_z$, the subscript is suppressed henceforth), but which one is chosen depends on how large the Hilbert space $L^2(V, f(z))$ is required or expected. For example, if $V = \mathbb{R}$, then $L^2(V, 1/(1+z^2))$ is much smaller than $L^2(V, e^{-z^2})$. Normally, the thinner the tail of the density, the larger the Hilbert space. Condition B.2 also stipulates an orthogonal sequence $\{p_i(z), i \geq 0\}$ with respect to $f(z)$. Most orthogonal sequences used in the literature are polynomial sequences and very much depend on the specification of V. If $V = \mathbb{R}$, the sequence may be Hermite polynomial sequence orthogonal with density $f(z) = e^{-z^2}$; if $V = [0, \infty)$, the sequence may be Laguerre polynomials orthogonal with density $f(z) = e^{-z}$; if $V = [a, b]$ is bounded interval, one may use trigonometric functions or Chebyshev polynomials possibly with a linear mapping such that they are orthogonal on $[a, b]$.

Notice also that the uniform boundedness in Condition B.2 is fulfilled for all bases mentioned above after normalization (such that they become orthonormal). In fact, (1) if $p_i(z), i \geq 0$, are Laguerre polynomials with density $f(z) = e^{-z}$ and $V = [0, \infty)$, then $\sup_{z\in V} \sup_{i\geq 0} |p_i(z)f^{1/2}(z)| < 1$; (2) if $p_i(z), i \geq 0$, are Hermite polynomials with density $f(z) = e^{-z^2}$ and $V = \mathbb{R}$, then $\sup_{z\in V} \sup_{i\geq 0} |p_i(z)f^{1/2}(z)| \leq \pi^{-1/4}$; (3) orthogonal trigonometric function sequence satisfies the uniform boundedness automatically when the support is a bounded interval. In this case $f(z) \equiv 1$, and again $\{\varphi_j(r), j \geq 1\}$ defined in the next subsection is an example. See pages 205 and 208 of Erdelyi et al. (1953), Indritz (1961), Todd (1963) and Gautschi (2004) for more details.

The martingale difference structure for the error term in Condition B.3 is extensively used in the literature such as Park and Phillips (1999, 2001) and Wang and Phillips (2009) among others. However, the inclusion of $\{x_t, t \leq n\}$ is a bit strong. In defence of this assumption, some papers impose the independence between the unit root process and the error term (Wang and Phillips, 2009, Theorem 3.1), which is even more stringent than our assumption, and some papers use the strong approximation for integrated process to Brownian motion (Kasparis et al., 2015, Assumption2.2(b)), which also has a theoretical drawback pointed out by Wang (2014). To circumvent the drawback, Wang (2014, 2015) and Wang and Phillips (2016) establish weak asymptotic theory for a kernel estimator that instead uses the information $\{x_s, s \leq t+1\}$ in $\mathcal{F}_{t,n}$. To do so, these papers take advantage of the form of the kernel estimator and establish the joint weak convergence for the numerator and denominator.

It is worth noting that with the inclusion of $\{x_t, t \leq n\}$ in the information filtration, we establish self-normalized normality for our sieve estimator; we point out in the next section that there are some situations where the inclusion can be relaxed.

2.2. Estimation procedure

The sieve estimation method is used to estimate the unknown function in model (1.1). This gives rise to the questions of: which function space does the unknown function belong to, and which basis should be used to represent it? The function space should be sufficiently large to include a wide range of reasonable choices for the regression function and to accommodate a broad class of processes for the regressors.

We assume that

$$m_0(r, z, x) \in L^2([0, 1] \times V \times \mathbb{R}, \phi(z, x)), \tag{2.1}$$

that is, $m_0(r, z, x)$ satisfies $\iiint_{[0,1]\times V\times\mathbb{R}} m_0^2(r, z, x)\phi(z, x)drdzdx < \infty$, where ϕ is a density function defined on $V \times \mathbb{R}$. Here, we take $\phi(z, x) = f(z)\exp(-x^2)$, where $f(z)$ is given by Assumption B. Henceforth, denote $L^2_\phi = L^2([0, 1] \times V \times \mathbb{R}, \phi(z, x))$ for convenience. For L^2_ϕ, we construct a basis as the tensor product of the bases chosen from $L^2[0, 1] = \{u(r) : \int_0^1 u^2(r)dr < \infty\}$, $L^2(V, f(z)) = \{p(z) : \int_V p^2(z)f(z)dz < \infty\}$ and $L^2(\mathbb{R}, e^{-x^2}) = \{g(x) : \int g^2(x)e^{-x^2}dx < \infty\}$, respectively. We stipulate these as follows.

Firstly, let $\varphi_0(r) \equiv 1$, and for $j \geq 1$, $\varphi_j(r) = \sqrt{2}\cos(\pi jr)$. Then, $\{\varphi_j(r)\}$ is an orthonormal basis in the Hilbert space $L^2[0, 1]$. Here, the inner product is given by $\langle u_1, u_2\rangle = \int_0^1 u_1(r)u_2(r)dr$ for any $u_1(\cdot), u_2(\cdot) \in L^2[0, 1]$ with the induced norm $\|u\|^2 = \langle u, u\rangle$ for any $u(\cdot) \in L^2[0, 1]$. It follows that $\langle\varphi_i(r), \varphi_j(r)\rangle = \delta_{ij}$ the Kronecker delta. Note that $\{\varphi_j(r)\}$ can be replaced by any other orthonormal basis in $L^2[0, 1]$, as shown in Chen and Shen (1998), Gao et al. (2001) and Phillips (2005) among others. However, with this specific basis other than a general one, we do not need any assumption on it. All quantities related to the basis are easily and directly calculated.

Secondly, the orthogonal function sequence $\{p_i(z), i \geq 0\}$ stipulated in Assumption B is chosen as an orthonormal basis of $L^2(V, f(z))$.

Thirdly, we choose from the space $L^2(\mathbb{R}, e^{-x^2})$ the Hermite orthogonal polynomial sequence $\{H_j(x)\}$ as the basis where the inner product is given by $\langle f_1, f_2\rangle = \int f_1(x)f_2(x)e^{-x^2}dx$ with the induced norm $\|f\|^2 = \langle f, f\rangle$. Recall that Hermite polynomials $\{H_j(x)\}$ are defined by

$$H_j(x) = (-1)^j \exp(x^2)\frac{d^j}{dx^j}\exp(-x^2), \qquad j \geq 0, \tag{2.2}$$

and satisfy $\int H_i(x)H_j(x)\exp(-x^2)dx = \sqrt{\pi}2^j j!\delta_{ij}$, meaning that they are orthogonal with respect to the density $\exp(-x^2)$. It is well-known that $\{H_j(x)\}$ is a complete orthogonal polynomial sequence and hence $h_j(x) := (\sqrt{\pi}2^j j!)^{-1/2}H_j(x)$ is an orthonormal polynomial basis in $L^2(\mathbb{R}, e^{-x^2})$. Unlike in the space $L^2[0,1]$, it seems impossible to have an orthonormal polynomial basis other than Hermite polynomials $\{H_j(x)\}$ in $L^2(\mathbb{R}, e^{-x^2})$. This is because in general an orthogonal polynomial sequence is uniquely determined by the support and the density up to a constant.

Finally, the tensor product $\{\varphi_i(r)\} \otimes \{p_j(z)\} \otimes \{h_\ell(x)\}$ is an orthonormal basis in L^2_ϕ. For better exposition, denote $\mathcal{B}_{ij\ell}(r,z,x) := \varphi_i(r)p_j(z)h_\ell(x)$, which is used to represent the expansion of the unknown regression function m_0 into an orthogonal series, i.e.

$$m_0(r,z,x) = \sum_{i,j,\ell=0}^{\infty} c_{ij\ell}\mathcal{B}_{ij\ell}(r,z,x), \quad \text{where } c_{ij\ell} = \iiint_{[0,1]\times V\times\mathbb{R}} m_0(r,z,x)\mathcal{B}_{ij\ell}(r,z,x)\phi(z,x)drdzdx. \tag{2.3}$$

Let k_i, $i = 1,2,3$, be positive integers and $K = k_1k_2k_3$. Define the truncated series with truncation parameter $k = (k_1,k_2,k_3)$,

$$m_k(r,z,x) = \sum_{i=0}^{k_1-1}\sum_{j=0}^{k_2-1}\sum_{\ell=0}^{k_3-1} c_{ij\ell}\mathcal{B}_{ij\ell}(r,z,x) := Z_k(r,z,x)^\intercal c, \tag{2.4}$$

where $Z_k(r,z,x)^\intercal := (\mathcal{B}_{000}(r,z,x),\ldots,\mathcal{B}_{k_1-1,k_2-1,k_3-1}(r,z,x))$ is the K-dimensional vector of the basis functions used to approximate the regression function in which $\mathcal{B}_{ij\ell}(r,z,x)$ is organized in a certain ordering and $c^\intercal := (c_{000},\ldots,c_{k_1-1,k_2-1,k_3-1})$ is in the same ordering.

The residual after truncation, denoted by $\gamma_k(r,z,x)$, is the series that consists of all terms for which $(i,j,\ell) \notin \mathcal{K} := \{0,\ldots,k_1-1\}\times\{0,\ldots,k_2-1\}\times\{0,\ldots,k_3-1\}$. That is,

$$\gamma_k(r,z,x) = \sum_{(i,j,\ell)\notin\mathcal{K}} c_{ij\ell}\mathcal{B}_{ij\ell}(r,z,x), \tag{2.5}$$

and it can be spelt out as a sum of seven terms including the following three:

$$\gamma_{1k}(r,z,x) := \sum_{i=k_1,j=0,\ell=0}^{\infty} c_{ij\ell}\mathcal{B}_{ij\ell}(r,z,x), \quad \gamma_{2k}(r,z,x) := \sum_{i=0,j=k_2,\ell=0}^{\infty} c_{ij\ell}\mathcal{B}_{ij\ell}(r,z,x),$$
$$\gamma_{3k}(r,z,x) := \sum_{i=0,j=0,\ell=k_3}^{\infty} c_{ij\ell}\mathcal{B}_{ij\ell}(r,z,x),$$

which constitute the leading terms in $\gamma_k(r,z,x)$ (they are slower in convergence than the other terms.[1]) In view of the expansion (2.3), the truncation series (2.4) and the residual (2.5), model (1.1) can be written as

$$y_t = Z_k(\tau_t,z_t,x_t)^\intercal c + \gamma_k(\tau_t,z_t,x_t) + e_t, \tag{2.6}$$

for $t = 1,\ldots,n$. To write all equations in (2.6) into a matrix form, let: $y = (y_1,\ldots,y_n)^\intercal$, $Z_{nK} = (Z_k(\tau_1,z_1,x_1),\ldots,Z_k(\tau_n,z_n,x_n))^\intercal$ an $n\times K$ matrix, $\gamma = (\gamma_k(\tau_1,z_1,x_1),\ldots,\gamma_k(\tau_n,z_n,x_n))^\intercal$, and $e = (e_1,\ldots,e_n)^\intercal$. Hence, we have

$$y = Z_{nK}c + \gamma + e. \tag{2.7}$$

We are now ready to define our estimator. Let $W_n = \mathrm{diag}(\phi(z_1,x_1),\ldots,\phi(z_n,x_n))$. The estimate of the coefficients is derived from a weighted least squares (WLS)

$$\widehat{c} = \arg\min_{c\in\mathbb{R}^K}(y - Z_{nK}c)^\intercal W_n(y - Z_{nK}c), \tag{2.8}$$

which yields the closed form solution $\widehat{c} = (Z_{nK}^\intercal W_n Z_{nK})^{-1}Z_{nK}^\intercal W_n y$. Then, let for any $r\in[0,1]$, $z\in V$ and $x\in\mathbb{R}$,

$$\widehat{m}_n(r,z,x) = Z_k(r,z,x)^\intercal\widehat{c}, \tag{2.9}$$

be an estimator of the unknown $m_0(r,z,x)$.

The WLS method facilitates the derivation of our large sample theory below and makes the estimation procedure robust. This is because the density involved in the estimation would automatically "draw back" outlier observations.

[1] To see this, take $\gamma_{1k}(r,z,x)$ as an example. Rewrite $\gamma_{1k}(r,z,x) = \sum_{i=k_1}^{\infty} c_i(z,x)\varphi_i(r)$ where $c_i(z,x) = \sum_{j=0,\ell=0}^{\infty} c_{ij\ell}p_j(z)h_\ell(x) = \int_0^1 m_0(r,z,x)\varphi_i(r)dr$ can be viewed as the coefficients in the expansion of $m_0(r,z,x)$, for given (z,x), in terms of $\{\varphi_i(r)\}$. Thus the decay rate of $\gamma_{1k}(r,z,x)$ is only related to the series of univariate function expansion, and such rate can be found in Lemma A.5. Similar interpretation goes to $\gamma_{2k}(r,z,x)$ and $\gamma_{3k}(r,z,x)$. All other residual terms have at least two truncations so they converge quicker than these three.

Moreover, without the WLS approach, we have to make stronger condition on the α-mixing coefficients in Assumption B and higher moment condition on the orthogonal functions $p_i(\cdot)$, or we have to truncate the support of the argument in m_0. In addition, although Wang (2015, p.209) considers a nonparametric model including stationary and nonstationary variables using the kernel method, the methodology therein may not be usable for our model (1.1). This is because the kernel estimator suffers from a degeneracy issue found in Phillips et al. (2017) when the deterministic trend is involved, while as illustrated by Park and Hahn (1999) and our paper below, the weighted sieve estimator does not have such an issue. Hence, dealing with model (1.1), the sieve method has certain advantages over the kernel method.

Nonetheless, there remains the question of how we should choose the density function. As we mentioned previously, the choice of the density determines the function space where the regression function resides. If the support V of z_t is bounded, one may simply use $f(z) \equiv 1$ without any loss of generality. In the unbounded support case, there is a trade-off to be made: a density f with very thin tails will allow a very large class of regression functions that can grow rapidly in the tails, on the other hand, the larger the space allowed, the higher the asymptotic variance.

Before showing the asymptotic theory of the estimator, we briefly discuss the quantity of $\|Z_k(r,z,x)\|$ that is crucial in our theory because it determines partially the convergence rate of our estimator. By definition,

$$\|Z_k(r,z,x)\|^2 = \sum_{i=0}^{k_1-1} \varphi_i^2(r) \sum_{j=0}^{k_2-1} p_j^2(z) \sum_{\ell=0}^{k_3-1} h_\ell^2(x).$$

A straightforward calculation similar to Lemma A.4 in Dong and Linton (2018) yields

$$\frac{1}{k_1} \sum_{i=0}^{k_1-1} \varphi_i^2(r) = 1 + O(k_1^{-1}), \tag{2.10}$$

when $k_1 \to \infty$ for $r \in (0,1)$. Moreover, in the mathematical literature the reciprocal of $\sum_{j=0}^{k_2-1} p_j^2(z)$ is called the Christoffel function for a general orthonormal polynomial sequence. See (3.3) of Nevai (1986, p.6), (33) and (38) of Máté et al. (1991, p.445) and Levin and Lubinsky (2001, p.18). When $V = [-1,1]$ and for the density $f(z)$ in Assumption B.2, Corollary 1.3 of Lubinsky (2009, p.917) gives

$$\lim_{k_2\to\infty} \frac{1}{k_2} \sum_{j=0}^{k_2-1} p_j^2(z) = [\pi f(z)\sqrt{1-z^2}]^{-1}, \tag{2.11}$$

for every $z \in (-1,1)$; for any bounded interval $[a,b]$ the diverging order of $\sum_{j=0}^{k_2-1} p_j^2(z)$ is the same as $O(k_2)$ for any orthonormal polynomial system on the interval with density satisfying Assumption B.2 and any $z \in (a,b)$, because the linear mapping $z = 2(u-a)/(b-a) - 1$ transforms $[a,b]$ into $[-1,1]$. When the support is $\mathbb{R}$ and the density is $\exp(-x^2)$, as a special case, Theorem 1.1 of Levin and Lubinsky (1992) shows that

$$\frac{1}{\sqrt{k_3}} \sum_{\ell=0}^{k_3-1} h_\ell^2(x) \asymp \exp(x^2) \left(\max\left\{ k_3^{-2/3}, 1 - \frac{|x|}{\sqrt{2k_3}} \right\} \right)^{1/2}, \tag{2.12}$$

uniformly for $k_3 \geq 1$ and $x \in \{u : |u| \leq \sqrt{2k_3}(1 + Lk_3^{-2/3})\}$ where $L > 0$ is a constant. Here, the relationship $a \asymp b$ in the above means that there exist positive absolute constants c_1, c_2 such that $c_1 < a/b < c_2$. Thus, if $1 - |x|/\sqrt{2k_3} > k_3^{-2/3}$, we have $\sum_{\ell=0}^{k_3-1} h_\ell^2(x) \asymp \sqrt{k_3}$. Notice that the constant $\exp(x^2)$ in the equivalence relationship is also important at least in practice, though it is innocuous in theory. This order is also applicable to $\{p_j(z)\}$ when $V = \mathbb{R}$ and $f(z) = \exp(-z^2)$. One may be curious about why the order is $\sqrt{k_3}$ other than k_3. Corollary 1.4 in Levin and Lubinsky (1992) shows that $\sup_{x\in\mathbb{R}} |h_\ell(x)| e^{-x^2/2} \asymp \ell^{-1/12}$ for $\ell \to \infty$, which implies $\sum_{\ell=0}^{k_3-1} h_\ell^2(x) \leq Ce^{x^2} \sum_{\ell=1}^{k_3-1} \ell^{-1/6} = O(k_3^{5/6})$. This partially gives the answer.

Accordingly, we may conclude for fixed $r \in [0,1], z \in V, x \in \mathbb{R}$, the order of $\|Z_k(r,z,x)\|^2$ is $O(k_1 k_2 \sqrt{k_3})$ when V is bounded or $O(k_1\sqrt{k_2 k_3})$ when $V = \mathbb{R}$. Here we assume the condition on x and/or z is fulfilled automatically, i.e. $1 - |x|/\sqrt{2k_3} > k_3^{-2/3}$, because we allow the truncation parameters to diverge. Notice also that the dependence of $\|Z_k(r,z,x)\|^2$ on (r,z,x) asymptotically boils down to

$$\Phi(z,x) = \begin{cases} \left[\pi f(z)\sqrt{1-z^2}\right]^{-1} \exp(x^2) & \text{if } V = [-1,1], \\ \exp(z^2 + x^2) & \text{if } V = \mathbb{R}, \end{cases}$$

an explicit and fixed function. Moreover, from the above analysis one can easily find the lower and upper bounds for $\|Z_k(r,z,x)\|^2$ when $\min(k_1,k_2,k_3) \to \infty$.

3. Asymptotic theory

3.1. Main result

Notice that for any (r, z, x) in the domain of m_0, the estimation error $\widehat{m}_n(r, z, x) - m_0(r, z, x)$ has leading stochastic term

$$Z_k(r, z, x)^\intercal (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_{nK}^\intercal W_n e$$

under Assumptions A, B, and C given below. From this, a self-normalized limit theory will be established. It follows from Assumption B that the normalizer should be the square root of

$$\Sigma_n(r, z, x)^2 := \sigma_e^2 Z_k(r, z, x)^\intercal (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_{nK}^\intercal W_n^2 Z_{nK} (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_k(r, z, x), \tag{3.1}$$

which includes all observations as well as the conditional variance of the error term. In Lemma A.4, we show approximations for $Z_{nK}^\intercal W_n Z_{nK}$ and $Z_{nK}^\intercal W_n^2 Z_{nK}$, namely, there exist two matrices Ψ_K and Ξ_K defined in Appendix A such that $\|\frac{d_n}{n} Z_{nK}^\intercal W_n Z_{nK} - \Psi_K\| = o_P(1)$ and $\|\frac{d_n}{n} Z_{nK}^\intercal W_n^2 Z_{nK} - \Xi_K\| = o_P(1)$. Denote the minimum and maximum eigenvalues of Ψ_K and Ξ_K by $\lambda_{\min}^{\Psi}$, $\lambda_{\max}^{\Psi}$, $\lambda_{\min}^{\Xi}$ and $\lambda_{\max}^{\Xi}$, respectively. We now present Assumption C.

Assumption C.

- C.1 Suppose that $m_0(r, z, x) \in L_\phi^2$ and $m_0(r, z, x)$ is differentiable with respect to r, z and x, respectively, up to the orders of s_1, s_2 and s_3.
- C.2 Suppose that k_1, k_2 and k_3 are divergent as $n \to \infty$, and that $\lambda_{\min}^{\Xi} \asymp K^{-\iota_1}$, $\lambda_{\max}^{\Xi} \asymp K^{\iota_2}$, $\lambda_{\min}^{\Psi} \asymp K^{-\varsigma_1}$ and $\lambda_{\max}^{\Psi} \asymp K^{\varsigma_2}$ in probability uniformly in n, $\iota_i \geq 0$ and $\varsigma_i \geq 0$, $i = 1, 2$, such that:
  - (a) $K^{4+\iota_1+2(\varsigma_1+\varsigma_2)} k_3 = o(n)$.
  - (b) $K^{\iota_1+\varsigma_2} \sqrt{n} \max(k_1^{-2s_1} \log^2(k_1), k_2^{-s_2}, k_3^{-s_3}) = o(1)$.
  - (c) $K^{2\varsigma_2+\iota_1} \sqrt{n} \max(k_1^{-2s_1} \log^2(k_1), k_2^{-s_2+1}, k_3^{-s_3+1}) = o(1)$.

Assumption C.1 summarizes the properties of the regression function where the differentiability guarantees the convergence of the orthogonal series expansion of the function m_0 with a certain rate. Assumption C.2 allows that the smallest eigenvalues of Ψ_K and Ξ_K decay to zero whereas the largest ones diverge to infinity with a certain rate. We emphasize that assuming the dependence of such eigenvalues on sample size is not new in the literature. See, for example, Assumption 5, Lemmas 2.1–2.4 and Theorem 3.2 in Chen and Christensen (2015) and Corollary 1 in Chang et al. (2015). Certainly, when $\iota_i = 0$ and $\varsigma_i = 0$, $i = 1, 2$, the requirement implies that all eigenvalues are bounded below from zero and above from infinity, a quite common assumption. See also Theorems 1–4 and Corollaries 2–3 in Chang et al. (2015), and Condition A.2 in Belloni et al. (2015).

Moreover, the unit root process plays a different role from the other two variables in our analysis, which is reflected in Assumption C.2(a) where an extra k_3 is involved. It can be seen from the proof of Lemma A.4 that, when $\iota_i = 0$ and $\varsigma_i = 0$, $i = 1, 2$, and if $m_0(r, z, x) \equiv m_0(z)$, the condition becomes $k_2^2 = o(n)$, which coincides with the literature such as Newey (1997) and Ai and Chen (2003), while if $m_0(r, z, x) \equiv m_0(x)$ the condition becomes $k_3^5 = o(n)$, which is the same as in Dong et al. (2016) and Dong and Gao (2018). Meanwhile, Assumptions C.2(b)–(c) are undersmoothing conditions that guarantee the negligibility of the truncation error terms; note that the truncation parameters in different directions play different roles depending on whether their supports are bounded (r) or unbounded (x and potentially z).

Theorem 3.1. *For fixed $r \in [0, 1]$, $z \in V$ and $x \in \mathbb{R}$, under Assumptions A–C, we have*

$$\Sigma_n^{-1}(r, z, x)[\widehat{m}_n(r, z, x) - m_0(r, z, x)] \to_D N(0, 1) \tag{3.2}$$

as $n \to \infty$ where Σ_n is given by (3.1).

Furthermore, let for $(r, z, x) \neq (r', z', x')$,

$$\Lambda_n^2(r, z, x; r', z', x') := \sigma_e^2 \begin{pmatrix} Z_k(r, z, x)^\intercal \\ Z_k(r', z', x')^\intercal \end{pmatrix} \Omega_{nK} \begin{pmatrix} Z_k(r, z, x)^\intercal \\ Z_k(r', z', x')^\intercal \end{pmatrix}^\intercal,$$

a 2×2 matrix. Then, we have

$$\Lambda_n^{-1}(r, z, x; r', z', x') \begin{pmatrix} \widehat{m}_n(r, z, x) - m_0(r, z, x) \\ \widehat{m}_n(r', z', x') - m_0(r', z', x') \end{pmatrix} \to_D N(0, I_2)$$

as $n \to \infty$ provided that Λ_n is invertible, where I_2 is the identity matrix of order 2.

The proof is given in Appendix B. The result in (3.2) is a self-normalized version of normality. Note that $\frac{d_n}{n} K^{\iota_1 - 2\varsigma_2} \|Z_k(r, z, x)\|^2 \leq \Sigma_n^2(r, z, x) \leq \frac{d_n}{n} K^{\iota_2 + 2\varsigma_1} \|Z_k(r, z, x)\|^2$ in probability under Assumption C. Hence, the convergence rate of the estimator $\widehat{m}_n(r, z, x)$ depends on both $\|Z_k(r, z, x)\|^2$ and the related eigenvalues. As explained in the preceding section,

when V is bounded, the order of $\|Z_k(r,z,x)\|^2$ is $O(k_1k_2\sqrt{k_3})$ provided that the point z is an interior point of V; when $V=\mathbb{R}$, the order of $\|Z_k(r,z,x)\|^2$ is $O(k_1\sqrt{k_2k_3})$. While the quantity $\Phi(z,x)$ induced by (2.11) and (2.12) affects the efficiency, it does not affect the convergence rate of the estimator under our assumptions.

If $m_0(r,z,x)$ reduces to $m_0(x)$ and $\iota_i=0$ and $\varsigma_i=0$, $i=1,2$, the rate becomes $\sqrt[4]{k_3/n}$. This is comparable with Theorem 3.1 of Wang and Phillips (2009) where the unknown function in a cointegrating regression is estimated by the kernel method and the estimator has convergence rate $1/\sqrt{\sqrt{n}h}$ in which h is bandwidth. See Remark 3.3 of Wang and Phillips (2009, p. 722).

Note that $\|\widehat{m}_n(r,z,x)-m_0(r,z,x)\|^2_{L^2_\phi}=\|\widehat{c}-c\|^2+\|\gamma_k\|^2_{L^2_\phi}$ by the orthogonality of the basis functions. This can give some clue about the choice of $k=(k_1,k_2,k_3)$. Basically, the optimal k should balance the variance $\mathbb{E}\|\widehat{c}-c\|^2$ and the squared bias $\|\gamma_k\|^2_{L^2_\phi}$. In the simulation study we investigate a cross-validation method for selection of the tuning parameters that is adapted to our framework, but we have no theoretical result on the choice of k.[2]

Notice further that the 2×2 matrix Λ_n^2 has elements on the diagonal proportional to $\|Z_k(r,z,x)\|^2$ and $\|Z_k(r',z',x')\|^2$ and elements off-diagonal $Z_k(r,z,x)^\top Z_k(r',z',x')$, apart from a factor d_n/n and the impact of related matrices. Hence, the convergence rate in the second assertion is the same as the first result of Theorem 3.1. Normally, $\widehat{m}_n(r,z,x)-m_0(r,z,x)$ and $\widehat{m}_n(r',z',x')-m_0(r',z',x')$ do not have independent limits as they share the same $I(1)$ process that has a random limit. However, they are asymptotically conditionally independent, and this is significant because it allows averaging over (r,z,x), as occurs inside semiparametric or partial mean functionals (see the next section), to improve the rates of convergence.

To make statistical inference, the nuisance parameter σ_e in Σ_n should be replaced by a consistent estimator. Note that the estimator of σ_e is different from the usual one, due to the use of WLS and the involvement of the unit root process. Here, $\widehat{\sigma}_e^2$ is defined from $\widehat{e}_t$ with weight depending on the observations rather than the usual equal weighting.

Corollary 3.1. *Let $\widehat{e}_t=y_t-\widehat{m}_n(\tau_t,z_t,x_t)$ for $t=1,\dots,n$. Define*

$$\widehat{\sigma}_e^2=\left(\sum_{t=1}^n\phi(z_t,x_t)\right)^{-1}\sum_{t=1}^n\widehat{e}_t^2\,\phi(z_t,x_t).$$

Then, under Assumptions A–C, $\widehat{\sigma}_e\to_P\sigma_e$ as $n\to\infty$.

The proof is given in Appendix B. This result facilitates the construction of consistent pointwise confidence intervals.

We conclude with a discussion regarding the inclusion of $\{x_t,t\le n\}$ in $\mathcal{F}_{t,n}$ stipulated in Assumption B. We would like to point out that such an inclusion is not necessary in the following two situations. First, if we are only interested in the order of $\widehat{m}_n-m_0$ in either norm or point-wise sense, we can use a usual filtration $\mathcal{F}^*_{t,n}=\sigma(e_j,z_{j+1},x_{j+1},j\le t)$ in the martingale structure in Assumption B. Indeed, for any (r,z,x) in the domain of m_0, we have

$$|\widehat{m}_n(r,z,x)-m_0(r,z,x)|\le\frac{d_n}{n}\|Z_k(r,z,x)\|(\lambda^{\Psi}_{\min})^{-1}\|Z_{nK}^\top W_n(e+\gamma)\|+|\gamma_k(r,z,x)|.$$

Hence, one only needs $\mathcal{F}^*_{t,n}$ to calculate the conditional variance of $\|Z_{nK}^\top W_n(e+\gamma)\|$. Second, the inclusion can be relaxed if a heteroscedasticity structure is imposed such as $e_t=\sigma(\tau_t,z_t,x_t)\varepsilon_t$ where ε_t is independent of $\{z_t,x_t\}$ with some condition on $\sigma(\cdot)$. Then, the result of Theorem 3.1 still holds if $\mathcal{F}_{t,n}$ is replaced by $\mathcal{F}^*_{t,n}$ and $\Sigma_n(r,z,x)^2$ is substituted by

$$\Sigma_n^*(r,z,x)^2=Z_k(r,z,x)^\top(Z_{nK}^\top W_nZ_{nK})^{-1}Z_{nK}^\top W_n\Omega_nW_nZ_{nK}(Z_{nK}^\top W_nZ_{nK})^{-1}Z_k(r,z,x),\tag{3.3}$$

where $\Omega_n=\mathrm{diag}(\sigma(\tau_1,z_1,x_1),\dots,\sigma(\tau_n,z_n,x_n))$. Moreover, if we denote $\hat{\Omega}_n=\mathrm{diag}(\hat{e}_1^2,\dots,\hat{e}_n^2)$, and let $\hat{\Sigma}_n^*(r,z,x)^2$ be obtained from $\Sigma_n^*(r,z,x)^2$ with replacement of Ω_n by $\hat{\Omega}_n$, under certain conditions we may show $\hat{\Sigma}_n^*(r,z,x)^2/\Sigma_n^*(r,z,x)^2\to_P 1$ as $n\to\infty$. See Section E in the supplementary material of the paper.

3.2. Marginal effects

We next consider certain linear functionals of $m_0(r,z,x)$, which are often of interest in applications where they may represent average marginal effects. Here, we discuss the estimates of

$$\delta_1=\int_0^1\int_{V\times\mathbb{R}}\frac{\partial m_0(r,z,x)}{\partial x}d\omega(r,z,x),\quad \delta_2=\int_0^1\int_{V\times\mathbb{R}}\frac{\partial m_0(r,z,x)}{\partial r}d\omega(r,z,x),$$
$$\delta_3=\int_0^1\int_{V\times\mathbb{R}}\frac{\partial m_0(r,z,x)}{\partial z}d\omega(r,z,x),\quad \delta_{12}=\int_0^1\int_{V\times\mathbb{R}}\frac{\partial^2 m_0(r,z,x)}{\partial r\partial z}d\omega(r,z,x),$$

[2] Our central limit theorems reflect the undersmoothing condition that downplays the bias terms, which leads to "suboptimal" convergence rates albeit the convergence rates we achieve are arbitrarily close to these "optimal rates", where it should be acknowledged that we know of no formal theory of optimal estimation in the current context. It is widespread practice to employ undersmoothing to facilitate simple inference methods, although there is some recent work that tries to take account of bias terms in inference. We shall leave this for future research.

$$\delta_{13} = \int_0^1 \int_{V\times\mathbb{R}} \frac{\partial^2 m_0(r,z,x)}{\partial r \partial x} d\omega(r,z,x), \quad \delta_{23} = \int_0^1 \int_{V\times\mathbb{R}} \frac{\partial^2 m_0(r,z,x)}{\partial z \partial x} d\omega(r,z,x),$$

where $d\omega(r,z,x) = \phi(z,x)drdzdx$. We take this specific weighting for simplicity and because it reflects our choice of parameter space. One could choose other weighting schemes, especially regarding the r,z coordinates, but our theory requires downweighting of the x coordinate comparable to e^{-x^2} and so any weighting scheme should satisfy $c\phi(z,x) \leq w(r,z,x) \leq C\phi(z,x)$ for positive finite constants c, C. Many papers consider density weighted (which is comparable with our weighting scheme) average derivatives for the same technical reasons, see for example Powell et al. (1989).

Because of the orthogonality of the basis, all these quantities can be expressed by the coefficients in the orthogonal expansion of m_0. We can naturally define the estimators of the δ's simply by replacing m_0 with $\widehat{m}_n$, e.g.

$$\widehat{\delta}_1 := \int_0^1 \int_{V\times\mathbb{R}} \frac{\partial \widehat{m}_n(r,z,x)}{\partial x} d\omega(r,z,x).$$

Note that the partial derivative will eventually be taken on the basis functions for $\widehat{m}_n(r,z,x) = Z_k(r,z,x)^\intercal \widehat{c}$, which is of linear form. In the next result we state the asymptotic properties of the $\widehat{\delta}$'s.

Theorem 3.2. *Under the same conditions as Theorem 3.1, we have*

$$B_{1n}^{-1}(\widehat{\delta}_1 - \delta_1) \to_D N(0, 8\sqrt{\pi}), \qquad B_{2n}^{-1}(\widehat{\delta}_2 - \delta_2) \to_D N(0, 8),$$
$$B_{3n}^{-1}(\widehat{\delta}_3 - \delta_3) \to_D N(0, 8\sqrt{\pi}), \qquad B_{4n}^{-1}(\widehat{\delta}_{12} - \delta_{12}) \to_D N(0, 64\sqrt{\pi}),$$
$$B_{5n}^{-1}(\widehat{\delta}_{13} - \delta_{13}) \to_D N(0, 64\sqrt{\pi}), \qquad B_{6n}^{-1}(\widehat{\delta}_{23} - \delta_{23}) \to_D N(0, 64\pi),$$

as $n \to \infty$, where denoting $\Omega_{nK} = (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_{nK}^\intercal W_n^2 Z_{nK} (Z_{nK}^\intercal W_n Z_{nK})^{-1}$,

$$B_{1n}^2 := \sigma_e^2\, \ell_3^\intercal \Omega_{nK} \ell_3, \quad B_{2n}^2 := \sigma_e^2\, L_1^\intercal \Omega_{nK} L_1, \quad B_{3n}^2 := \sigma_e^2\, \ell_2^\intercal \Omega_{nK} \ell_2,$$
$$B_{4n}^2 := \sigma_e^2\, L_2^\intercal \Omega_{nK} L_2, \quad B_{5n}^2 := \sigma_e^2\, L_3^\intercal \Omega_{nK} L_3, \quad B_{6n}^2 := \sigma_e^2\, \ell_{11}^\intercal \Omega_{nK} \ell_{11},$$

and $\ell_3 = (0,0,1,0,\ldots,0)^\intercal$, $\ell_2 = (0,1,0,\ldots,0)^\intercal$, $\ell_{11} = (0,\ldots,0,1,0,\ldots,0)^\intercal$ where in the last vector 1 is located at the same place as c_{011} in the c, L_1 is defined to be a sparse column vector (viz. most elements are zero) where 1's conformably are in the same place as c_{i00} (odd i only and $i \leq k_1 - 1$) in c, L_2 is defined similarly but 1's conformably are in the same place as c_{i10} (odd i only and $i \leq k_1 - 1$) in c, and sparse L_3 with 1 at the same place as c_{i01} (odd i only and $i \leq k_1 - 1$) in c; all of them are of dimension K.

All the quantities in the theorem have a faster convergence rate than the function estimator in Theorem 3.1. To see this, note that $\Omega_{nK} = \frac{d_n}{n} \Psi_K^{-1} \Xi_K \Psi_K^{-1}(1 + o_P(1))$ by Lemma A.4. To make the comparison simpler, we temporarily suppose that all eigenvalues of Ψ_K and Ξ_K are bounded away from zero and above from infinity, i.e. $\iota_i = 0$ and $\varsigma_i = 0$ for $i = 1, 2$ in Assumption C. Therefore, $B_{1n}^2 \asymp \frac{d_n}{n}$. Hence, $\widehat{\delta}_1$ has rate $n^{-1/4}$, whereas $\widehat{m}_n$ has rate $\|Z_k(r,z,x)\| n^{-1/4}$. The comparison of convergence rates between any one of all the other $\widehat{\delta}$'s and $\widehat{m}_n$ is clear because B_{3n}^2 and B_{6n}^2 are all proportional to $\frac{d_n}{n}$, and B_{2n}^2, B_{4n}^2 and B_{5n}^2 are all proportional to $\frac{d_n}{n}\sqrt{k_1}$. One thing we have to mention is that in the derivations for δ_3, δ_{12} and δ_{23} we specify the sequence $\{p_j(z)\}$ to be Hermite polynomials, otherwise we have to make a great number of assumptions on $\{p_j(z)\}$, which we eschew for space limitation reasons.

3.3. Additive and multiplicative separability

The above result is valid without any further functional form restrictions. We next consider the case where we are willing to impose a separability assumption on the regressions surface, either additive separability $m_0(r,z,x) = \beta_0(r) + g_0(z) + q_0(x)$ or multiplicative separability $m_0(r,z,x) = \beta_1(r) g_1(z) q_1(x)$. Similar to Linton and Nielsen (1995), we shall consider the following contrasts

$$\chi_0 = \int_0^1 \int_{V\times\mathbb{R}} m_0(r,z,x) f(z) w(x) drdzdx, \qquad \chi_1(r) = \int_{V\times\mathbb{R}} m_0(r,z,x) f(z) w(x) dzdx,$$
$$\chi_2(z) = \int_{[0,1]\times\mathbb{R}} m_0(r,z,x) w(x) drdx, \qquad \chi_3(x) = \int_{[0,1]\times V} m_0(r,z,x) f(z) drdz,$$

where $w(x) = \pi^{-1/2} e^{-x^2}$ and $f(z)$ is given in Assumption B. Notice that actually $\chi_0 = c_{000}$ the first coefficient in the expansion (2.3).

In the additive case, we have:

$$\chi_0 = \int_0^1 \beta_0(r) dr + \int_V g_0(z) f(z) dz + \int q_0(x) w(x) dx,$$
$$\chi_1(r) = \beta_0(r) + \theta_1, \quad \theta_1 := \int_V g_0(z) f(z) dz + \int q_0(x) w(x) dx,$$

$$\chi_2(z) = g_0(z) + \theta_2, \quad \theta_2 := \int_0^1 \beta_0(r)dr + \int q_0(x)w(x)dx,$$
$$\chi_3(x) = q_0(x) + \theta_3, \quad \theta_3 := \int_0^1 \beta_0(r)dr + \int_V g_0(z)f(z)dz,$$

and accordingly $\chi_1(r) + \chi_2(z) + \chi_3(x) - 2\chi_0 = m_0(r, z, x)$. In the multiplicative situation, we have:

$$\chi_0 = \int_0^1 \beta_1(r)dr \int_V g_1(z)f(z)dz \int q_1(x)w(x)dx,$$
$$\chi_1(r) = \pi_1\beta_1(r), \quad \pi_1 := \int_V g_1(z)f(z)dz \int q_1(x)w(x)dx,$$
$$\chi_2(z) = \pi_2 g_1(z), \quad \pi_2 := \int_0^1 \beta_1(r)dr \int q_1(x)w(x)dx,$$
$$\chi_3(x) = \pi_3 q_1(x), \quad \pi_3 := \int_0^1 \beta_1(r)dr \int_V g_1(z)f(z)dz,$$

and consequently $\chi_1(r)\chi_2(z)\chi_3(x)/\chi_0^2 = m_0(r, z, x)$ provided that $\chi_0 \neq 0$.

Naturally, we estimate $\chi_i(\cdot)$ by

$$\begin{aligned} \widehat{\chi}_0 &= \widehat{c}_{000}, \qquad \widehat{\chi}_1(r) = \int_{V\times\mathbb{R}} \widehat{m}_n(r, z, x)f(z)w(x)dzdx, \\ \widehat{\chi}_2(z) &= \int_{[0,1]\times\mathbb{R}} \widehat{m}_n(r, z, x)w(x)drdx, \quad \widehat{\chi}_3(x) = \int_{[0,1]\times V} \widehat{m}_n(r, z, x)f(z)drdz. \end{aligned} \tag{3.4}$$

Define four K-dimensional vectors:

$$\ell_1^\intercal := (1, 0, \ldots, 0), \quad P_{k_2}(z)^\intercal := (0, \ldots, 0, p_0(z), \ldots, p_{k_2-1}(z), 0, \ldots, 0),$$
$$\Phi_{k_1}(r)^\intercal := (\varphi_0(r), \ldots, \varphi_{k_1-1}(r), 0, \ldots, 0), \quad U_{k_3}(x)^\intercal := (0, \ldots, 0, h_0(x), \ldots, h_{k_3-1}(x)),$$

and define four quantities related to conditional variances below:

$$A_{0n}^2 := \sigma_e^2 \ell_1^\intercal \Omega_{nK} \ell_1, \quad A_{1n}^2(r) := \sigma_e^2 \Phi_{k_1}(r)^\intercal \Omega_{nK} \Phi_{k_1}(r),$$
$$A_{2n}^2(z) := \sigma_e^2 P_{k_2}(z)^\intercal \Omega_{nK} P_{k_2}(z), \quad A_{3n}^2(x) := \sigma_e^2 U_{k_3}(x)^\intercal \Omega_{nK} U_{k_3}(x),$$

where Ω_{nK} is defined in Theorem 3.2.

Theorem 3.3. *Under the same conditions as Theorem 3.1, we have*

$$A_{0n}^{-1}(\widehat{\chi}_0 - \chi_0) \to_D N(0, 1), \quad A_{1n}(r)^{-1}(\widehat{\chi}_1(r) - \chi_1(r)) \to_D N(0, 1),$$
$$A_{2n}(z)^{-1}(\widehat{\chi}_2(z) - \chi_2(z)) \to_D N(0, 1), \; A_{3n}(x)^{-1}(\widehat{\chi}_3(x) - \chi_3(x)) \to_D N(0, 1),$$

as $n \to \infty$ for any r, z and x.

Similar to the comment for Theorem 3.2, apart from the affect of the eigenvalues of Ψ_K and Ξ_K, A_{0n}^2 is proportional to d_n/n, while $A_{1n}^2 \asymp (d_n/n)\sum_{i=0}^{k_1-1} \varphi_i^2(r)$, $A_{2n}^2 \asymp (d_n/n)\sum_{j=0}^{k_2-1} p_j^2(z)$ and $A_{3n}^2 \asymp (d_n/n)\sum_{\ell=0}^{k_3-1} h_\ell^2(x)$, and these three sums involved can be spelt out from the analysis of $\|Z_k(r, z, x)\|$ in Section 2. As a result, the convergence rates in Theorem 3.3 are all faster than that in Theorem 3.1. This is because the integrations shorten the vector of basis functions by orthogonality. We provide a simulation study in the supplement of the paper that illustrates how the results of this section can allow one to discriminate between additive and multiplicative structures.

4. Monte Carlo simulations

We conduct Monte Carlo simulations in order to validate the relevance of our theoretical results for finite sample situations. Consider model (1.1) with the following data generation procedure: $z_t = |\xi_t|$ with $\xi_t \sim$ i.i.d. $N(0, 1)$; the unit root regressor x_t is integrated by an AR(1) process w_t, i.e., $x_t = x_{t-1} + w_t$, where $w_t = \rho_w w_{t-1} + \epsilon_t$, $\rho_w = 0.2$, $\epsilon_t \sim$ i.i.d. $N(0, 0.2^2)$, $w_0 \sim N(0, 1/(1-\rho_w^2))$ and $x_0 \sim N(0, 0.2^2)$; finally, $e_t \sim$ i.i.d. $N(0, 1)$. Note that both z_t and x_t have unbounded support.

Table 1
Estimation results for $\widehat{m}$ and $\widehat{\sigma}_e^2$.

	n	400	600	800
Case 1	$\text{RMSE}_{\widehat{m}}$	0.104	0.088	0.077
	$\text{RMSE}_{\widehat{\sigma}_e^2}$	0.265	0.247	0.215
Case 2	$\text{RMSE}_{\widehat{m}}$	0.088	0.074	0.065
	$\text{RMSE}_{\widehat{\sigma}_e^2}$	0.083	0.059	0.059
Case 3	$\text{RMSE}_{\widehat{m}}$	0.068	0.055	0.048
	$\text{RMSE}_{\widehat{\sigma}_e^2}$	0.207	0.183	0.169

For the regression function m_0 in (1.1), we consider the following three cases:

Case 1. $m_0(\tau, z, x) = \exp(\tau + x/6) + z$;

Case 2. $m_0(\tau, z) = \tau(z + z^2) + 2$;

Case 3. $m_0(\tau, x) = \exp((\tau + x)/6)$.

Notice that m_0 in Cases 2 and 3 are bivariate functions, hence they are special cases of the general model. They are used to plot 3-dimensional graphs for the estimated functions.

In view of the data process, we adopt $\{\varphi_i(r)\}$ and $\{h_\ell(x)\}$ for time series $\{\tau_t\}$ and $\{x_t\}$ respectively as defined in the beginning of Section 2.2, while for $\{z_t\}$, we let $\{p_j(z)\}$ be the Laguerre polynomials. This means that the basis used to expand m_0 is the tensor product $\{\varphi_i(r)\} \otimes \{p_j(z)\} \otimes \{h_\ell(x)\}$ in Case 1, the tensor product $\{\varphi_i(r)\} \otimes \{p_j(z)\}$ in Case 2 and the tensor product $\{\varphi_i(r)\} \otimes \{h_\ell(x)\}$ in Case 3. More possible choices of basis functions under different scenarios can be found in Chen (2007).

For the choice of truncation parameters, we follow Gao et al. (2002) to consider the minimization of a generalized cross validation (GCV) function. Take Case 1 as an example. The GCV function is defined as

$$(\widehat{k}_1, \widehat{k}_2, \widehat{k}_3) = \underset{k_1,k_2,k_3}{\text{argmin}} \frac{(y - Z_{nK}\widehat{c})^\intercal W_n (y - Z_{nK}\widehat{c})}{n(1 - \frac{k_1 k_2 k_3}{n})^2}. \tag{4.1}$$

For the other cases, we drop k_2 or k_3 in the above definition according to the data generating process.

According to our estimation procedure in Section 2, we construct the weight matrix W_n in each case as follows:

Case 1. $W_n = \text{diag}\left(\exp(-x_1^2 - z_1), \ldots, \exp(-x_n^2 - z_n)\right)$;

Case 2. $W_n = \text{diag}\left(\exp(-z_1), \ldots, \exp(-z_n)\right)$;

Case 3. $W_n = \text{diag}\left(\exp(-x_1^2), \ldots, \exp(-x_n^2)\right)$.

For each generated dataset, with the proposed estimation procedure we compute two values

$$\varpi_1 = \frac{1}{n}(\Upsilon - Z_{nK}\widehat{c})^\intercal W_n (\Upsilon - Z_{nK}\widehat{c}) \quad \text{and} \quad \varpi_2 = \widehat{\sigma}_e^2,$$

where $\widehat{c}$ is defined in (2.8), $\Upsilon = (m_0(\tau_1, z_1, x_1), \ldots, m_0(\tau_n, z_n, x_n))^\intercal$ (this is simplified in Case 2 and Case 3 according to the function forms), and $\widehat{\sigma}_e^2$ is defined in Corollary 3.1. Note that ϖ_1 is essentially equivalent to the mean squared errors under parametric setting. After M replications of the Monte Carlo simulation, we calculate three quantities by

$$\text{RMSE}_{\widehat{m}} = \sqrt{\frac{1}{M}\sum_{j=1}^{M} \varpi_1(j)}, \quad \text{and} \quad \text{RMSE}_{\widehat{\sigma}_e^2} = \sqrt{\frac{1}{M}\sum_{j=1}^{M} [\varpi_2(j) - \sigma_e^2]^2},$$

where $\varpi_1(j)$ and $\varpi_2(j)$ respectively stand for the values of ϖ_1 and ϖ_2 in the jth replication, and $\sigma_e^2 = 1$ by our DGP. Moreover, we expect that both $\text{RMSE}_{\widehat{m}}$ and $\text{RMSE}_{\widehat{\sigma}_e^2}$ to be sufficiently small.

We summarize the results in Table 1. It is obvious that the values of $\text{RMSE}_{\widehat{m}}$ and $\text{RMSE}_{\widehat{\sigma}_e^2}$ decrease to 0 as n goes up. It seems that the convergence of the RMSE for $\widehat{\sigma}_e^2$ is slower than that of $\widehat{m}$, in particular in the first and the third cases where the unit root process is involved. This is because the later is weighted RMSE. Note that these simulations verify our theoretical results in Theorem 3.1 and Corollary 3.1. (Due to space limitations, some extra simulation results regarding our discussion associated to additive and multiplicative forms of regression are provided in the supplementary material of this paper).

In order to visualize our simulation results, we plot $\phi^{1/2}\widehat{m}$ (with $n = 800$) and $\phi^{1/2}m_0$ of Cases 2 and 3 in Fig. 1, where ϕ is the density of the function space as defined in the beginning of Section 2.2. Fig. 1 includes the true $\phi^{1/2}m_0$ and the 90% confidence interval of $\phi^{1/2}\widehat{m}$ for each given point based on 1000 replications. As can be seen, the three layers in each subplot are close to each other though the interval for Case 3 is a bit wider than Case 2 that we understand is due to the involvement of the unit root process. This further confirms our proposed estimation procedure and the theoretical development via finite sample simulation.

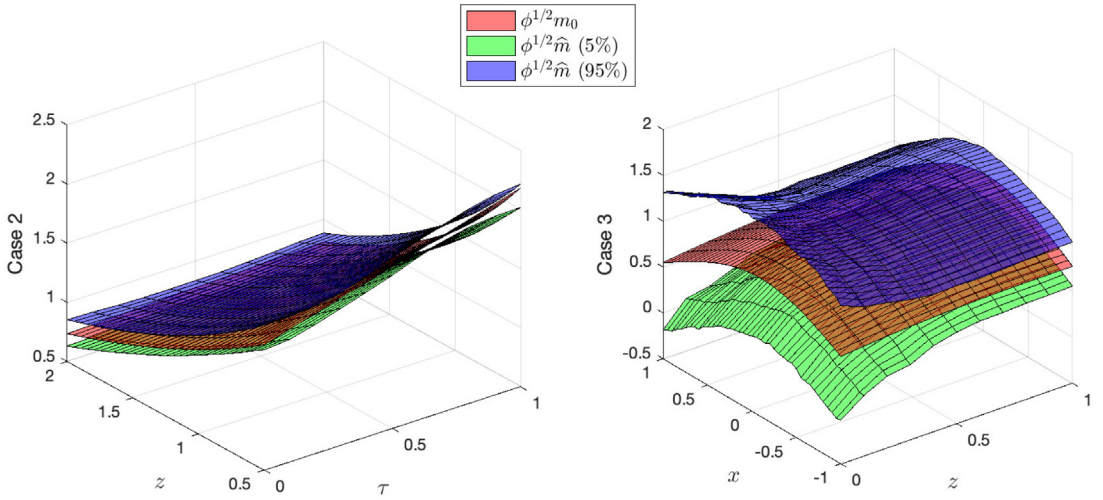

Fig. 1. $\widehat{m}$ and m_0 of Cases 2 and 3.

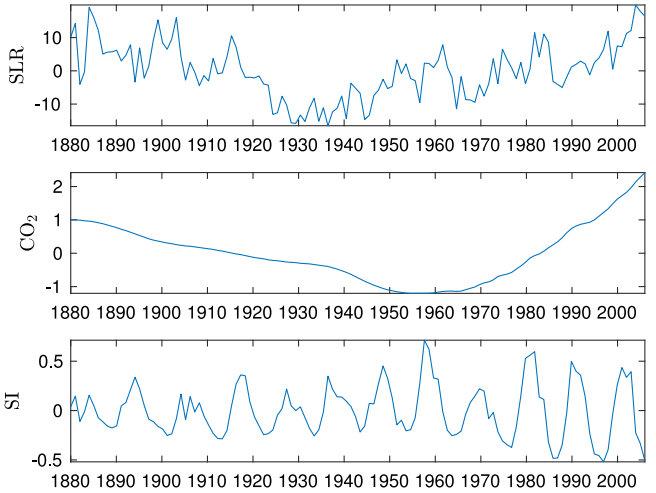

Fig. 2. Three detrended time series.

5. Empirical study

It is well understood that global sea level rise is primarily a problem of the rise of global temperature caused by too much carbon dioxide (chemical formula CO_2) and solar irradiance (referred to as SI hereafter). This section studies how CO_2 and SI affect SLR using the proposed nonparametric model.

Annual data for SLR, CO_2 and SI from 1880 to 2005 are used in our study. The data are collected from the websites: SLR (in mm) from CSIRO,[3] CO_2 (in parts per million) from NASA[4] and SI (in W/m^2) from University of Colorado Boulder.[5] Note that the range of original data of CO_2 is from 291.2 to 379.2 parts per million. As we shall show below, this series is an $I(1)$ process. Therefore, due to the numerical limitation of our computer, if we stick to the current unit, the weight $\exp(-w^2)$ will wipe out almost all information when implementing the estimation procedure. To deal with this computational issue, we change the units of CO_2 from parts per million to parts per one hundred thousand in this study (i.e., dividing all original data of CO_2 by 10). After this, we firstly remove a linear time trend from the three time series in order to get rid of strong time effects; we plot detrended time series in Fig. 2. All the following calculations are implemented based on the three detrended series.

[3] http://www.cmar.csiro.au/sealevel/index.html.

[4] https://climate.nasa.gov/vital-signs/carbon-dioxide.

[5] http://lasp.colorado.edu/home/sorce/data/tsi-data.

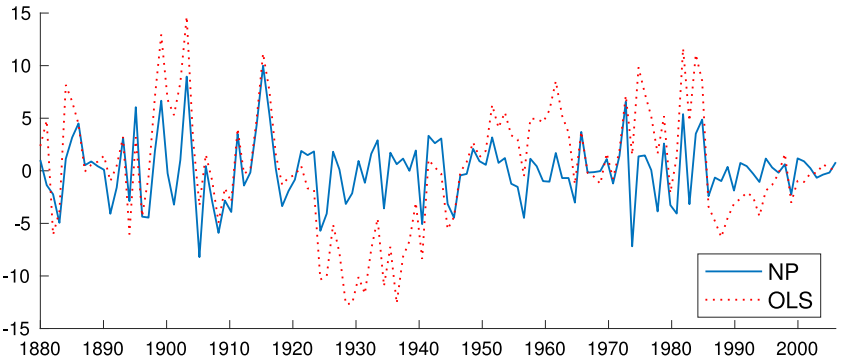

Fig. 3. Estimation residuals.

Table 2
Forecasting errors.

$\widetilde{n}$	1	2	3	4	5
FError	1.7333	1.4400	2.5567	2.5532	2.3137

In order to ensure our model (1.1) fits this study well, we implement the Augmented Dickey–Fuller (ADF) test to examine the stationarity of CO_2 and SI, and report the relevant p-values 0.9990 and 0.0131, respectively. It is clear that, at the 5% significant level, we fail to reject the null for CO_2 (i.e., having a unit root), and reject the null for SI (i.e., no unit root). Therefore, $\{z_t \mid t = 1, \ldots, n\}$ take the data of SI and $\{x_t \mid t = 1, \ldots, n\}$ take the data of CO_2 in model (1.1), respectively. Moreover, since SI moves between -1 and 1, we adopt the cosine sequence after a linear transformation from $[-1, 1]$ to [0,1]. Same as the simulation study, the truncation parameters $(\widehat{k}_1, \widehat{k}_2, \widehat{k}_3) = (4, 2, 4)$ are chosen by minimizing the GCV function. For the sake of space limitation, the detailed results associated to the GCV function are reported in the supplementary file of this paper.

For the purpose of comparison, we also consider a linear parametric model as follows:

$$y_t = \beta_0 + \beta_1 \tau_t + \beta_2 z_t + \beta_3 x_t + e_t. \tag{5.1}$$

The RMSE (i.e., $\sqrt{\frac{1}{n}(y-\widehat{y})^{\tau} W_n (y-\widehat{y})}$, where $\widehat{y}$ stands for the estimate of y) for each model is calculated. Here, to put the two models on an equal footing, the same W_n is used when estimating the parametric linear model (5.1). The RMSE for the nonparametric model (1.1) is 3.6587, while for the parametric linear model (5.1) it is 5.5823, indicating that nonparametric model is favorable to the parametric model. Furthermore, we plot the scaled version of the estimation residuals (i.e., $W_n^{1/2}(y-\widehat{y})$) of the two models in Fig. 3. Clearly, the estimated residuals of both models fluctuate around zero, but the nonparametric model (NP) yields a better fit in general than the parametric model (OLS). The full estimation results of linear model are provided in the online supplementary file.

In the following, we focus on reporting the results associated with the nonparametric model. Firstly, we take a look at the forecasting ability of our nonparametric model (1.1). Specifically, the forecasting errors are calculated as follows:

$$\text{FError} = \sqrt{\frac{1}{\widetilde{n}} \sum_{s=n-\widetilde{n}+1}^{n} \exp(-x_s^2) \cdot (y_s - \widehat{y}_s)^2},$$

where $\widehat{y}_s = Z_k(1, z_s, x_s)^{\tau} \widehat{c}_s$ and $\widehat{c}_s$ is calculated by using the sample $\{(\frac{t}{s}, z_t, x_t) \mid 1 \le t \le s-1\}$. One can also consider the above expression as a rolling out-of-sample weighted mean squared error. The results are summarized in Table 2. As can be seen, the forecasting errors are quite stable in terms of the choice of $\widetilde{n}$ and even smaller than the RMSE reported above.

Second, we plot the marginal effects of the weak time trend τ, SI and CO_2 respectively. Specifically, they are calculated by

$$\text{ME}(\tau) = \iint_{V \times \mathbb{R}} \frac{\partial \widehat{m}_n(\tau, z, x)}{\partial \tau} dz dx, \quad \text{ME}(z) = \iint_{[0,1] \times \mathbb{R}} \frac{\partial \widehat{m}_n(\tau, z, x)}{\partial z} d\tau dx$$

$$\text{ME}(x) = \iint_{[0,1] \times V} \frac{\partial \widehat{m}_n(\tau, z, x)}{\partial x} d\tau dz.$$

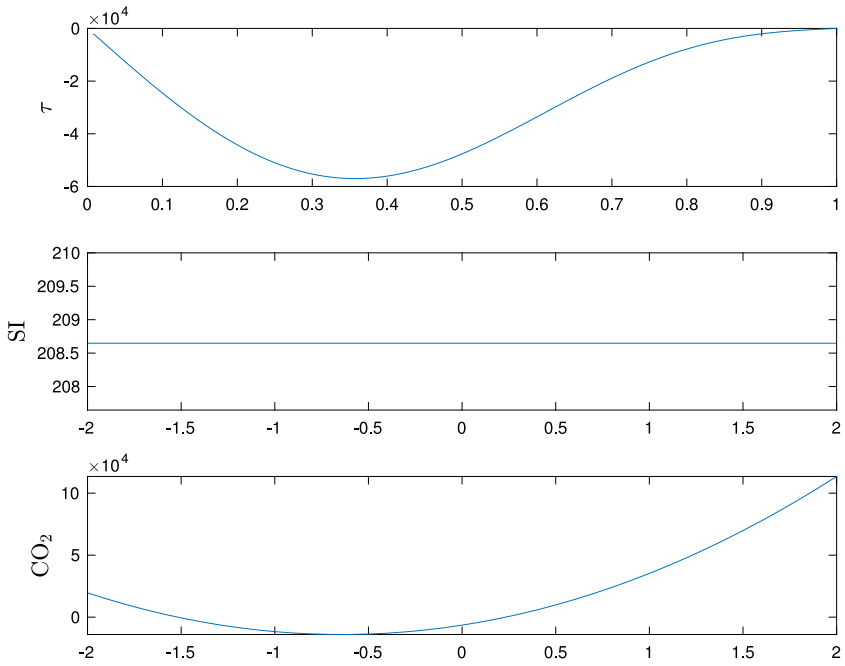

Fig. 4. Marginal effects.

It is not surprising that the marginal effects of SI is a horizontal line, given that the truncation parameter $k_2 = 2$. While the quantity of CO_2 increases, the marginal effects of CO_2 tend to increase, which quantifies the well known concern regarding the "Greenhouse Effect".

To further examine the interactions of different variables, we also plot the following three quantities in Fig. 5.

$$\mathrm{ME}(\tau, z) = \int_{\mathbb{R}} \frac{\partial^2 \widehat{m}_n(\tau, z, x)}{\partial \tau \partial z} dx, \quad \mathrm{ME}(\tau, x) = \int_V \frac{\partial^2 \widehat{m}_n(\tau, z, x)}{\partial \tau \partial x} dz$$

$$\mathrm{ME}(z, x) = \int_{[0,1]} \frac{\partial^2 \widehat{m}_n(\tau, z, x)}{\partial z \partial x} d\tau.$$

Although Fig. 4 shows that the marginal effects of CO_2 increase as CO_2 increases, the second sub-figure of Fig. 5 shows that the impacts of CO_2 become weak as time flies. It seems to suggest that the ecosystem can heal itself eventually. Of course, it depends on how we protect this ecosystem in the future.

Finally, we take a look at the interaction between CO_2 and SI in some selected time periods, so that we plot $\widehat{m}(0.25, \cdot, \cdot)$, $\widehat{m}(0.5, \cdot, \cdot)$, $\widehat{m}(0.75, \cdot, \cdot)$ and $\widehat{m}(1, \cdot, \cdot)$ in Fig. 6. These 3-dimensional plots in different periods show the change of the relationship among the dependent and independent variables. More interesting results can be drawn by looking at other plots at additional time periods.

6. Conclusion

The methodology we have proposed is relatively simple and works well in theory and practice. It may be adapted to allow for and exploit parametric short run weak dependence in error terms by adapting the objective function along the lines of Linton and Xiao (2019). We have not addressed uniform convergence issues as considered by Chen and Christensen (2015), and this remains an open problem for this model setting; we expect that weighted uniform convergence can be obtained with suitable rates. Some other extensions of the paper are possible. The scalar variable z_t might be replaced by a vector but, if the dimension is large, semiparametric models like single-index model or additive model are recommended. On the other hand, unit root vector can be involved in single-index structure like Dong et al. (2016). We may take these issues up in future studies. In terms of the application, we have found some evidence of nonlinearity and interaction effects in the relationship between time, CO_2, SI, and the output variable SLR. We will not overclaim the statistical significance of our findings, as Amreheim et al. (2019), but the marginal effect curves seem broadly consistent with evidence presented elsewhere. Other areas of application where our methods may prove useful are predictive regression for stock returns, where it is common to consider very persistent possibly nonstationary predictors as well as less persistent possible stationary predictors (Phillips and Lee, 2013; Andersen and Varneskov, 2018; Cheng et al., 2019).

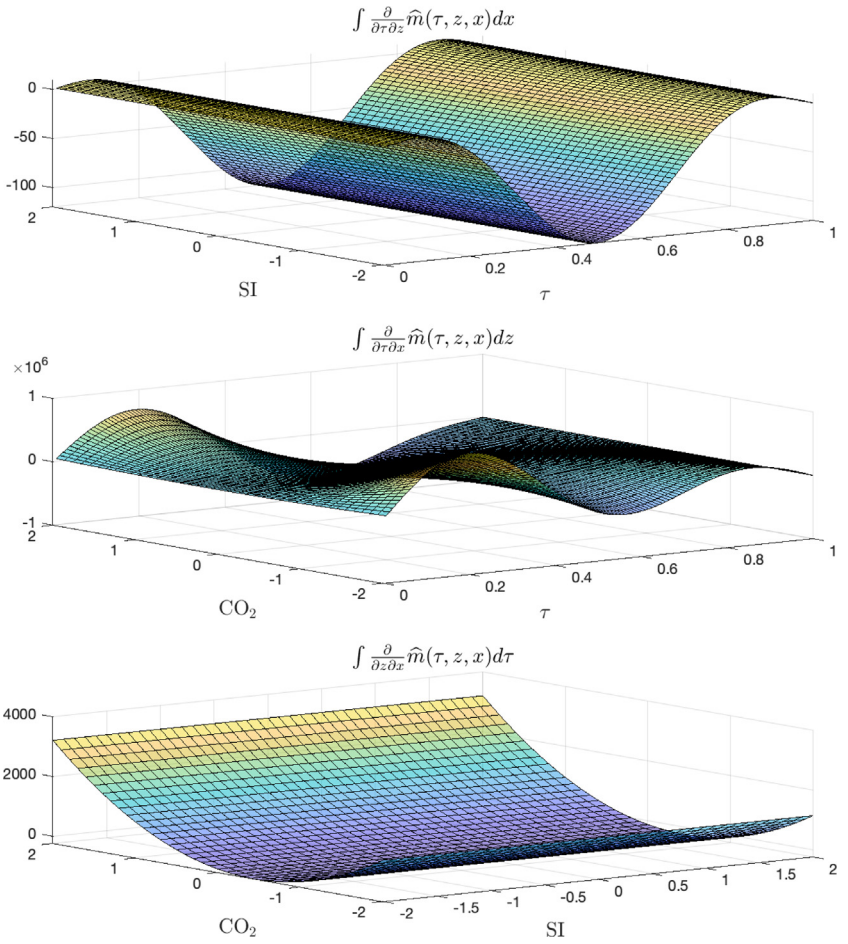

Fig. 5. Marginal interaction effects.

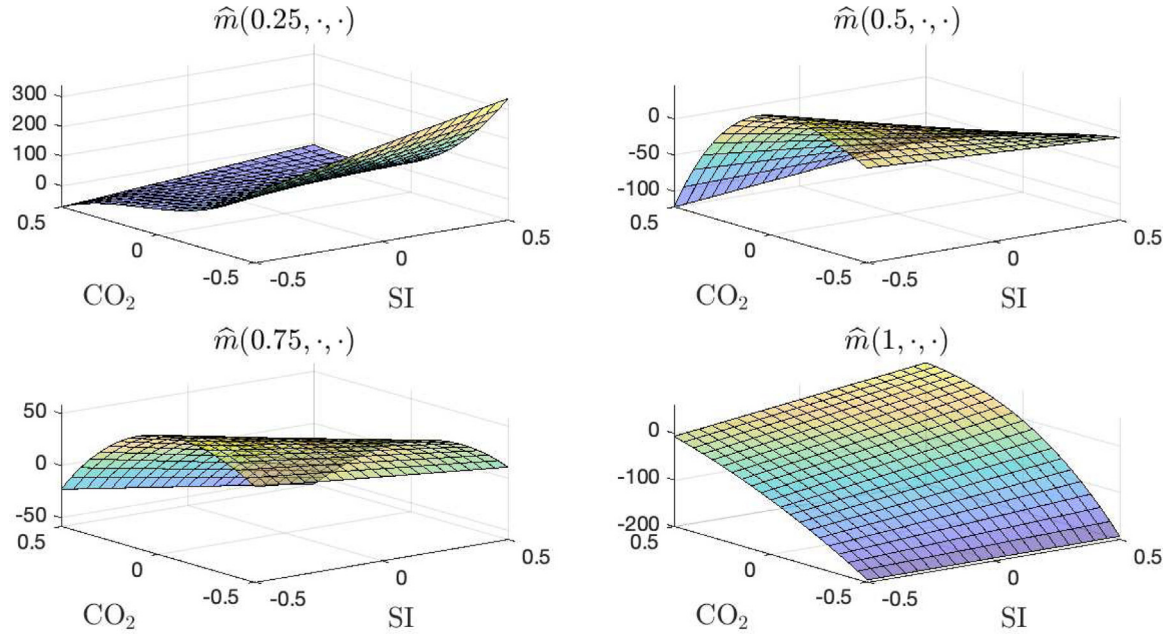

Fig. 6. Estimated $\widehat{m}$ at $\tau = 0.25, 0.5, 0.75$ and 1.

Acknowledgments

We thank the co-editor, AE and two anonymous referees for their insightful suggestions and comments that improve our paper considerably. The first author thanks the financial support from National Natural Science Foundation of China under Grant No. 71671143.

Appendix A. Lemmas

Without loss of generality, in what follows let $x_0 = 0$ almost surely. It follows that

$$x_t = \sum_{\ell=1}^{t} w_\ell = \sum_{\ell=1}^{t} \sum_{i=-\infty}^{\ell} \psi_{\ell-i}\epsilon_i = \sum_{i=-\infty}^{t} \left(\sum_{\ell=\max(1,i)}^{t} \psi_{\ell-i} \right) \epsilon_i =: \sum_{i=-\infty}^{t} b_{t,i}\epsilon_i. \tag{A.1}$$

Additionally, letting $1 \le s < t$, x_t also has the following decomposition:

$$x_t = x_s^* + x_{ts}, \tag{A.2}$$

where $x_s^* = x_s + \bar{x}_s$ with $\bar{x}_s = \sum_{i=s+1}^{t} \sum_{a=-\infty}^{s} \psi_{i-a}\epsilon_a$ containing all information available up to s and $x_{ts} = \sum_{i=s+1}^{t} b_{t,i}\epsilon_i$ which captures all information containing in x_t on the time periods $(s, t]$. Let $d_{ts} := (\mathbb{E}[x_{ts}^2])^{1/2}$ for later use. Moreover, $\bar{x}_s = O_P(1)$ by virtue of Assumption A.

Lemma A.1. *Suppose that Assumption A holds. For $t \to \infty$ or $t - s \to \infty$,*

(1) $d_t^{-1}x_t$ have uniformly bounded densities $f_t(x)$ over all t and x satisfying a uniform Lipschitz condition $\sup_x |f_t(x+y) - f_t(x)| \le C|y|$ for any y and some constant $C > 0$. In addition, $\sup_x |f_t(x) - \phi(x)| \to 0$ as $t \to \infty$ where $\phi(x)$ is the standard normal density function.

(2) Let $1 \le s < t$. $d_{ts}^{-1}x_{ts}$ have uniformly bounded densities $f_{ts}(x)$ over all (t, s) and x satisfying the above uniform Lipschitz condition as well.

This lemma is exactly Lemma A.2 in Dong and Gao (2018) whose proof is omitted here.

Taking into account that in Assumption B.1, z_t maybe contains $\epsilon_t, \ldots, \epsilon_{t-d+1}$, we decompose, for $t > d$,

$$x_t = \sum_{i=t-d+1}^{t} b_{t,i}\epsilon_i + \sum_{i=-\infty}^{t-d} b_{t,i}\epsilon_i := x_t^{(d)} + x_t^{(t-d)}. \tag{A.3}$$

Thus, $x_t^{(d)}$ and $x_t^{(t-d)}$ are mutually independent, and $x_t^{(d)}$ is stationary since it is a combination of $\epsilon_t, \ldots, \epsilon_{t-d+1}$ with fixed coefficients $\psi_0, \ldots, \sum_{\ell=0}^{d-1} \psi_\ell$ (i.e., a MA(d) process), while $x_t^{(t-d)}$ is still nonstationary since obviously its variance varies with t. More importantly, $x_t^{(t-d)}$ is independent of z_t.

Additionally, since z_t and z_s maybe share the same ϵ's, we decompose

$$x_t = x_t^{(d)} + x_{ts}^{(d)} + x_s^{(d*)} + x_s^{(s-d*)}, \tag{A.4}$$

$$\text{where } x_t^{(d)} = \sum_{i=t-d+1}^{t} b_{t,i}\epsilon_i, \quad x_{ts}^{(d)} = \sum_{i=s+1}^{t-d} b_{t,i}\epsilon_i$$

$$x_s^{(d*)} = x_s^{(d)} + \bar{x}_s^{(d)}, \quad x_s^{(s-d*)} = x_s^{(s-d)} + \bar{x}_s^{(s-d)},$$

recalling that $x_s^{(d)}$ and $\bar{x}_s^{(d)}$ are the sums of the first d terms of x_s and $\bar{x}_s$, respectively, whereas $x_s^{(s-d)}$ and $\bar{x}_s^{(s-d)}$ are the rests of them in x_s and $\bar{x}_s$, respectively. Obviously, all four components in (A.4) are mutually independent.

Lemma A.2. *Suppose that Assumption A holds.*

(1) Let $\tilde{d}_t^2 = \mathbb{E}[(x_t^{(t-d)})^2]$. When $t \to \infty$, $\tilde{d}_t^{-1}x_t^{(t-d)}$ have uniformly bounded densities $f_{t/d}(x)$ over all t and x satisfying a uniform Lipschitz condition $\sup_x |f_{t/d}(x+y) - f_{t/d}(x)| \le C|y|$ for any y and some constant $C > 0$. In addition, $\sup_x |f_{t/d}(x) - \phi(x)| \to 0$ as $t \to \infty$ where $\phi(x)$ is the standard normal density function.

(2) For $1 \le s < t$ and $t - s > d$, let $\tilde{d}_{ts}^2 = \mathbb{E}[(x_{ts}^{(t-d)})^2]$. When $t - s \to \infty$, $\tilde{d}_{ts}^{-1}x_{ts}^{(t-d)}$ have uniformly bounded densities $f_{ts/d}(x)$ over all (t, s) and x satisfying the above uniform Lipschitz condition as well.

It is noteworthy that $\tilde{d}_t = O(\sqrt{t})$, the same order as d_t, and $\tilde{d}_{ts} = O(\sqrt{t-s})$, the same order as d_{ts}, noting by that d is fixed. This fact will be used frequently in the following derivation which, for simplicity, will not be mentioned repeatedly. The proof of the lemma is much similar to Lemma A.1 so that it is omitted too.

Lemma A.3. *Suppose that Assumptions A and B.1(b) hold.*

(1) Let $p(\cdot)$ be a function such that $\mathbb{E}|p(z_t)| < \infty$, $h(\cdot)$ be such that $\int |h(x)|dx < \infty$. Then, for $t \to \infty$, $|\mathbb{E}[p(z_t)h(x_t)]| < C\tilde{d}_t^{-1}\mathbb{E}|p(z_t)| \int |h(x)|dx(1 + O(\tilde{d}_t^{-1}))$.

(2) *Let $p_1(\cdot)$ and $p_2(\cdot)$ satisfy the above condition for $p(\cdot)$; and $h_1(\cdot)$ is integrable and $h_2(\cdot)$ is such that $\int |xh_2(x)|dx < \infty$. For $1 \le s < t$ and $t - s > d$,*

$$|\mathbb{E}[p_1(z_t)p_2(z_s)h_1(x_t)h_2(x_s)]| \le C\tilde{d}_{ts}^{-1}\tilde{d}_s^{-1}\mathbb{E}|p_1(z_t)|\mathbb{E}|p_2(z_s)| \left|\int h_1(x)dx\right| \int |h_2(x)|dx + C_1\tilde{d}_{ts}^{-2}\tilde{d}_s^{-1}\mathbb{E}|p_1(z_t)|\mathbb{E}|p_2(z_s)| \int |h_1(x)|dx \int |xh_2(x)|dx,$$

and if $\int h_1(x)dx = 0$,

$$|\mathbb{E}[p_1(z_t)p_2(z_s)h_1(x_t)h_2(x_s)]| \le C_1\tilde{d}_{ts}^{-2}\tilde{d}_s^{-1}\mathbb{E}|p_1(z_t)|\mathbb{E}|p_2(z_s)| \int |h_1(x)|dx \int |xh_2(x)|dx.$$

The proof is relegated to the supplementary material of the paper. Next, we shall consider the matrices $Z_{nK}^\intercal W_n Z_{nK}$ and $Z_{nK}^\intercal W_n^2 Z_{nK}$ in the normalizer $\Sigma_n(r, z, x)$ defined in Section 3. We establish a *one-step approximation* for the matrices that facilitates the proof of our main result. Specifically, denote $\mathscr{P}_j(z) = p_j(z)f^{1/2}(z)$ and $\mathscr{H}_\ell(x) = h_\ell(x)\exp(-x^2/2)$ and further denote $\Phi_{ii'}(\cdot) = \varphi_i(\cdot)\varphi_{i'}(\cdot)$, $\mathcal{P}_{jj'}(\cdot) = \mathscr{P}_j(\cdot)\mathscr{P}_{j'}(\cdot)$ and $\mathcal{H}_{\ell\ell'}(\cdot) = \mathscr{H}_\ell(\cdot)\mathscr{H}_{\ell'}(\cdot)$ for brevity. Then, the $K \times K$ matrix $Z_{nK}^\intercal W_n Z_{nK}$ has elements $\sum_{t=1}^n \Phi_{ii'}(t/n)\mathcal{P}_{jj'}(z_t)\mathcal{H}_{\ell\ell'}(x_t)$ with all $(i, j, \ell), (i', j', \ell') \in \mathcal{K}$. At element level, the one-step approximation is

$$\frac{d_n}{n}\sum_{t=1}^n \Phi_{ii'}(t/n)\mathcal{P}_{jj'}(z_t)\mathcal{H}_{\ell\ell'}(x_t) = \mathbb{E}[\mathcal{P}_{jj'}(z_t)]\frac{d_n}{n}\sum_{t=1}^n \Phi_{ii'}(t/n)\mathcal{H}_{\ell\ell'}(x_t) + o_P(1)$$

uniformly over all indices under Assumptions A and B. Let Ψ_K be a square matrix with dimension K and elements $\mathbb{E}[\mathcal{P}_{jj'}(z_t)]\frac{d_n}{n}\sum_{t=1}^n \Phi_{ii'}(t/n)\mathcal{H}_{\ell\ell'}(x_t)$ in concert with the same ordering as the elements in $Z_{nK}^\intercal W_n Z_{nK}$. Thus, $\|\frac{d_n}{n}Z_{nK}^\intercal W_n Z_{nK} - \Psi_K\| = o_P(1)$ under Assumption C when $n \to \infty$, as shown rigorously in Lemma A.4. Similarly, $\|\frac{d_n}{n}Z_{nK}^\intercal W_n^2 Z_{nK} - \Xi_K\| = o_P(1)$ where Ξ_K has elements $\mathbb{E}[\mathcal{P}_{jj'}(z_t)f(z_t)]\frac{d_n}{n}\sum_{t=1}^n \Phi_{ii'}(t/n)\mathcal{H}_{\ell\ell'}(x_t)\exp(-x_t^2)$. As a result, all functions of z_t are replaced by their expectation while all x_t are remained in both Ψ_K and Ξ_K. Hence, the one-step approximation mitigates the condition on z_t to establish the limit theory for the estimator, but no further approximation can be made in the original probability space without using the strong convergence for the $I(1)$ process x_t in an expanded space. This is why the $\mathcal{F}_{t,n}$ includes all $x_s, s \le n$, but only z_s up to $t + 1$ in Assumption B. By contrast, for kernel estimator Wang (2014) takes advantage of the ratio form to establish a joint weak convergence for the numerator and denominator that implies the normality of the estimator. In this regard, the condition on the integrated variable for the kernel estimator is a bit weaker than for the sieve estimator.

Lemma A.4. *Under Assumptions A–C, as $n \to \infty$, we have $\|\frac{d_n}{n}Z_{nK}^\intercal W_n Z_{nK} - \Psi_K\| = o_P(1)$ and $\|\frac{d_n}{n}Z_{nK}^\intercal W_n^2 Z_{nK} - \Xi_K\| = o_P(1)$.*

The proof of Lemma A.4 is relegated to the supplementary material of the paper.

Now, let $[a, b]$ be a bounded interval and assume $g(x) \in C[a, b]$, all continuous functions on $[a, b]$. Let P_k be the set of all polynomials with order no higher than k and define $E_k(g) = \inf_{p_k \in P_k}\max_{x\in[a,b]}|g(x) - p_k(x)|$. For each k, as shown in page 26 of Chapter 3 and proven in page 79 of Chapter 8 of Todd (1963), there exists a $p_k^* \in P_k$ such that $E_k(g) = \max_{x\in[a,b]}|g(x) - p_k^*(x)|$. In the literature p_k^* is called the best approximation polynomial.

On the other hand, let $\{\varphi_j(x), j \ge 0\}$ be an orthonormal polynomial sequence in $L^2([a, b], \rho(x))$ where without loss of generality, $\int_a^b \rho(x)dx = 1$. Denote by $\|\cdot\|_{L^2}$ the norm in the space. Then, $g(x) = \sum_{j=0}^\infty c_j\varphi_j(x)$ where $c_j = \int_a^b g(u)\varphi_j(u)\rho(u)du$. Let $S_k(g; x) = \sum_{j=0}^k c_j\varphi_j(x)$, which can be viewed as a project operator from any function of $L^2([a, b], \rho(x))$ to a kth polynomial. For better exposition, denote $\Phi_k(x) = (\varphi_0(x), \ldots, \varphi_k(x))^\intercal$, $c = (c_0, \ldots, c_k)^\intercal$ and $\gamma_k(g; x) = \sum_{j=k+1}^\infty c_j\varphi_j(x)$. Thence, $g(x) = S_k(g; x) + \gamma_k(g; x)$ and $S_k(g; x) = \Phi_k(x)^\intercal c$.

Lemma A.5. *(1). For $g(x) \in C[a, b]$, we have (a) $\|\gamma_k(g; x)\|_{L^2} \le E_k(g)$; (b) For any $x \in [a, b]$, $|\gamma_k(g; x)| \le (1 + \|\Phi_k(x)\|)E_k(g)$; (c) In particular, for the first kind of Chebyshev polynomial sequence on $[-1, 1]$, $\max_{x\in[-1,1]}|\gamma_k(g; x)| \le (1 + \sqrt{k})E_k(g)$.*

(2) If g is continuously differentiable on $[a, b]$ up to sth order, then $E_k(g) \le Ck^{-s}\omega^{(s)}(k^{-1})$ for $k > s$, where $\omega^{(s)}(\cdot)$ is the modulus of continuity for $g^{(s)}$.

(3) If $h(r) \in L^2[0, 1]$ and the orthogonal sequence, $\varphi_0(r) \equiv 1$ and $\varphi_j(r) = \sqrt{2}\cos(\pi jr)$, $j \ge 1$, is used to expand the function. Define similarly $\gamma_k(h; r)$. If h is continuously differentiable on $[0, 1]$ up to sth order, then $\max_{r\in[0,1]}|\gamma_k(h; r)| \le Ck^{-s}\log(k)$.

(4) Suppose that $x^{s-\ell}g^{(\ell)}(x) \in L^2(\mathbb{R}, e^{-x^2})$ for $\ell = 0, \ldots, s$ for some integer $s > 0$. When $s = 1$, $g(x) = \sum_{i=0}^\infty c_i h_i(x)$ converges absolutely at any point on $\mathbb{R}$, where $h_i(x)$ are Hermite polynomials defined in Section 2; when $s \ge 1$, $|\gamma_k(x)| = |g(x) - \sum_{i=0}^{k-1} c_i h_i(x)| = o(k^{-(s-1)/2-1/12})$ at any point on $\mathbb{R}$, and $\int |\gamma_k(x)|^2 e^{-x^2}dx = o(k^{-s})$ when $k \to \infty$.

The proof is given in the supplementary material of the paper. The lemma mainly gives the point convergence rate of orthogonal series expansion in two situations, that is, on bounded interval and unbounded interval, respectively. In particular, the assertions of (1) and (3) connect the uniform rate on the finite interval with the error of the best polynomial approximation in line with Belloni et al. (2015). See Jackson (1930) and Schultz (1969) for the best polynomial approximation. We notice that the upper bound in (1b) may not be sharp, but it is sufficient for the paper. In this regard, see Belloni et al. (2015) for detailed discussion.

Appendix B. Proofs of the main results

Proof of Theorem 3.1. (1) By Lemma A.4, for any (r,z,x), we define by $\Delta_n^2(r,z,x)$ the one-step approximation to $\Sigma_n^2(r,z,x)$, that is, $\Sigma_n^2(r,z,x)=\Delta_n^2(r,z,x)(1+o_P(1))$ where $\Sigma_n^2(r,z,x)$ is given by Eq. (3.1) and

$$\Delta_n^2(r,z,x) := \sigma_e^2\frac{d_n}{n}Z_k(r,z,x)^\intercal\Psi_K^{-1}\Xi_K\Psi_K^{-1}Z_k(r,z,x).$$

Observe further that,

$$\begin{aligned}
&\Sigma_n^{-1}(r,z,x)[\widehat{m}_n(r,z,x)-m(r,z,x)]\\
=&\Sigma_n^{-1}(r,z,x)Z_k(r,z,x)^\intercal(\widehat{c}-c)-\Sigma_n^{-1}(r,z,x)\gamma_k(r,z,x)\\
=&\Sigma_n^{-1}(r,z,x)Z_k(r,z,x)^\intercal(Z_{nK}^\intercal W_nZ_{nK})^{-1}Z_{nK}^\intercal W_n(\gamma+e)-\Sigma_n^{-1}(r,z,x)\gamma_k(r,z,x)\\
=&\frac{d_n}{n}\Delta_n^{-1}(r,z,x)Z_k(r,z,x)^\intercal\Psi_K^{-1}Z_{nK}^\intercal W_n(\gamma+e)(1+o_P(1))-\Sigma_n^{-1}(r,z,x)\gamma_k(r,z,x),
\end{aligned}$$

where the leading term is

$$\begin{aligned}
&\frac{d_n}{n}\Delta_n^{-1}(r,z,x)Z_k(r,z,x)^\intercal\Psi_K^{-1}Z_{nK}^\intercal W_ne\\
=&\frac{d_n}{n}\Delta_n^{-1}(r,z,x)Z_k(r,z,x)^\intercal\Psi_K^{-1}\sum_{t=1}^n Z_k(\tau_t,z_t,x_t)\phi(z_t,x_t)e_t
\end{aligned}$$

from which the normality will be derived.

Let $\xi_{nt}=\frac{d_n}{n}\Delta_n(r,z,x)^{-1}Z_k(r,z,x)^\intercal\Psi_K^{-1}Z_k(\tau_t,z_t,x_t)\phi(z_t,x_t)e_t$, so that the leading term is written as $\xi_n:=\sum_{t=1}^n\xi_{nt}$. Moreover, noting by the elements of Ψ_K and Ξ_K and Assumption B, $(\xi_{nt},\mathcal{F}_{nt})$ form a martingale difference sequence and hence ξ_n is a martingale. Its conditional covariance is

$$\begin{aligned}
&\sigma_e^2\frac{d_n}{n}\Delta_n^{-2}(r,z,x)Z_k(r,z,x)^\intercal\Psi_K^{-1}\left(\frac{d_n}{n}\sum_{t=1}^n Z_k(\tau_t,z_t,x_t)Z_k(\tau_t,z_t,x_t)^\intercal\phi(z_t,x_t)^2\right)\Psi_K^{-1}Z_k(r,z,x)\\
=&\sigma_e^2\frac{d_n}{n}\Delta_n^{-2}(r,z,x)Z_k(r,z,x)^\intercal\Psi_K^{-1}\Xi_K\Psi_K^{-1}Z_k(r,z,x)(1+o_P(1))\equiv 1+o_P(1),
\end{aligned}$$

in view of the expression of $\Delta_n^2(r,z,x)$.

Moreover, for any $\eta>0$,

$$\begin{aligned}
&\sum_{t=1}^n\mathbb{E}[\xi_{nt}^2I(|\xi_{nt}|>\eta)|\mathcal{F}_{n,t-1}]\le\frac{1}{\eta^2}\sum_{t=1}^n\mathbb{E}[\xi_{nt}^4|\mathcal{F}_{n,t-1}]\\
\le&\frac{1}{\eta^2}\mu_4\Delta_n^{-4}(r,z,x)\frac{d_n^4}{n^4}\sum_{t=1}^n[Z_k(r,z,x)^\intercal\Psi_K^{-1}Z_k(\tau_t,z_t,x_t)\phi(z_t,x_t)]^4,
\end{aligned}$$

where $\mu_4=\max_{1\le t\le n}\mathbb{E}(e_t^4|\mathcal{F}_{t-1})<\infty$ stipulated in Assumption B.

Notice that

$$[Z_k(r,z,x)^\intercal\Psi_K^{-1}Z_k(\tau_t,z_t,x_t)\phi(z_t,x_t)]^2\le(\lambda_{\min}^\Psi)^{-2}K\|Z_k(r,z,x)\|^2$$

where $\|Z_k(\tau_t,z_t,x_t)\phi(z_t,x_t)\|^2\le K$ because all elements of $Z_k(\tau_t,z_t,x_t)\phi(z_t,x_t)$ are uniformly bounded. Meanwhile, $\Delta_n^2(r,z,x)\ge\frac{d_n}{n}\lambda_{\min}^\Xi(\lambda_{\max}^\Psi)^{-2}\|Z_k(r,z,x)\|^2$ where the constant σ_e^2 is ignored that does not affect the following derivation. It follows that

$$\begin{aligned}
&\sum_{t=1}^n\mathbb{E}[\xi_{nt}^2I(|\xi_{nt}|>\eta)|\mathcal{F}_{n,t-1}]\\
\le&C(\lambda_{\min}^\Xi)^{-1}(\lambda_{\max}^\Psi/\lambda_{\min}^\Psi)^2K\Delta_n^{-2}(r,z,x)\frac{d_n^3}{n^3}\sum_{t=1}^n[Z_k(r,z,x)^\intercal\Psi_K^{-1}Z_k(\tau_t,z_t,x_t)\phi(z_t,x_t)]^2\\
=&C(\lambda_{\min}^\Xi)^{-1}(\lambda_{\max}^\Psi/\lambda_{\min}^\Psi)^2K\Delta_n^{-2}(r,z,x)\frac{d_n^2}{n^2}Z_k(r,z,x)^\intercal\Psi_K^{-1}\left(\frac{d_n}{n}Z_{nK}^\intercal W_n^2Z_{nK}\right)\Psi_K^{-1}Z_k(r,z,x)\\
=&C(\lambda_{\min}^\Xi)^{-1}(\lambda_{\max}^\Psi/\lambda_{\min}^\Psi)^2K\Delta_n^{-2}(r,z,x)\frac{d_n^2}{n^2}Z_k(r,z,x)^\intercal\Psi_K^{-1}\Xi_K\Psi_K^{-1}Z_k(r,z,x)(1+o_P(1))\\
=&C(\lambda_{\min}^\Xi)^{-1}(\lambda_{\max}^\Psi/\lambda_{\min}^\Psi)^2K\frac{d_n}{n}(1+o_P(1))=o_P(1)
\end{aligned}$$

by the definition of $\Delta_n(r,z,x)$ and Assumption C. This verifies the Lindeberg condition for ξ_n. Hence, we have shown that $\xi_n \to_D N(0,1)$ as $n \to \infty$ by Corollary 3.1 in Hall and Heyde (1980).

To finish the proof, we need to show the negligibility of the remained terms, A_{1n} and A_{2n}, say, that is,

$$A_{1n} = \Delta_n^{-1}(r,z,x)\frac{d_n}{n} Z_k(r,z,x)^\intercal \Psi_K^{-1} Z_{nK}^\intercal W_n \gamma = o_P(1), \tag{B.1}$$

$$A_{2n} = \Delta_n^{-1}(r,z,x)\gamma_k(r,z,x) = o_P(1). \tag{B.2}$$

To show (B.1), note that

$$\begin{aligned}
|A_{1n}|^2 =& \Delta_n^{-2}(r,z,x)\frac{d_n^2}{n^2}|Z_k(r,z,x)^\intercal \Psi_K^{-1} Z_{nK}^\intercal W_n \gamma|^2 \\
\le& \Delta_n^{-2}(r,z,x)\frac{d_n^2}{n^2}\|Z_k(r,z,x)^\intercal \Psi_K^{-1} Z_{nK}^\intercal W_n^{1/2}\|^2 \|W_n^{1/2}\gamma\|^2 \\
=& \Delta_n^{-2}(r,z,x)\frac{d_n}{n} Z_k(r,z,x)^\intercal \Psi_K^{-1}\left(\frac{d_n}{n} Z_{nK}^\intercal W_n Z_{nK}\right)\Psi_K^{-1} Z_k(r,z,x) \times \gamma^\intercal W_n \gamma \\
=& \frac{d_n}{n}\Delta_n^{-2}(r,z,x) Z_k(r,z,x)^\intercal \Psi_K^{-1} Z_k(r,z,x)(1+o_P(1)) \times \gamma^\intercal W_n \gamma \\
\le& (\lambda_{\min}^{\Xi})^{-1}\lambda_{\max}^{\Psi}\sum_{t=1}^n \phi(z_t,x_t)\gamma_k^2(\tau_t,z_t,x_t),
\end{aligned}$$

where we use $\Delta_n^2(r,z,x) \ge \lambda_{\min}^{\Xi}(\lambda_{\max}^{\Psi})^{-1}\frac{d_n}{n} Z_k(r,z,x)^\intercal \Psi_K^{-1} Z_k(r,z,x)$. It follows from Lemma A.3 that

$$\begin{aligned}
\sum_{t=1}^n \mathbb{E}\phi(z_t,x_t)\gamma_k(\tau_t,z_t,x_t)^2 \le& \sum_{t=1}^n \frac{1}{\tilde{d}_t}\mathbb{E}\int \phi(z_1,x)\gamma_k(\tau_t,z_1,x)^2 dx \\
\le& \sqrt{n}\max_{r\in[0,1]}\mathbb{E}\int \phi(z_1,x)\gamma_k(r,z_1,x)^2 dx,
\end{aligned}$$

where $\tilde{d}_t \sim \sqrt{t}$. Moreover, $\gamma_k^2(r,z,x)$ has leading terms $\gamma_{1k}^2(r,z,x)$, $\gamma_{2k}^2(r,z,x)$ and $\gamma_{3k}^2(r,z,x)$ as explained in Section 2, and they relate with the truncation residues of the three orthogonal expansions of $m_0(r,z,x)$ in terms of $\{\varphi_i(r)\}$, $\{p_j(z)\}$ and $\{h_\ell(x)\}$, respectively. Thus, in the sequel we mainly consider $\mathbb{E}\int \phi(z_1,x)\gamma_{ik}(\tau_t,z_1,x)^2 dx$ for $i = 1,2,3$.

Now that $\gamma_{1k}(r,z,x)$ is the residue for $m_0(r,z,x)$ about r, $\max_{r\in[0,1]}|\gamma_{1k}(r,z,x)| = o(k_1^{-s_1}\log(k_1))$ and hence $\mathbb{E}\int \phi(z_1,x)\gamma_{1k}(\tau_t,z_1,x)^2 dx = o(k_1^{-2s_1}\log^2(k_1))$. See Lemma A.5 for this uniform convergence rate. Notice that $\mathbb{E}\int \phi(z_1,x)\gamma_{ik}(\tau_t,z_1,x)^2 dx$ are square norms of $\gamma_{ik}(r,z,x)$ for $i = 2,3$, so they have decay rates $o(k_1^{-s_i})$, respectively, $i = 2,3$, by Lemma A.5 again, where innocuously we suppose V is unbounded. Consequently, we have

$$\sum_{t=1}^n \mathbb{E}\phi(z_t,x_t)\gamma_k(\tau_t,z_t,x_t)^2 = C\sqrt{n}\max(k_1^{-2s_1}\log^2(k_1), k_2^{-s_2}, k_3^{-s_3}).$$

Thus, $A_{1n} = o_P(1)$ by Assumption C.

To show (B.2), note that

$$\begin{aligned}
&|A_{2n}| = \Delta_n^{-1}(r,z,x)|\gamma_k(r,z,x)| \\
\le& \Delta_n^{-1}(r,z,x)(|\gamma_{1k}(r,z,x)| + |\gamma_{2k}(r,z,x)| + |\gamma_{3k}(r,z,x)|) \\
=& (\lambda_{\min}^{\Xi})^{-1/2}\lambda_{\max}^{\Psi}\|Z_k(r,z,x)\|^{-1} n^{1/4} o(\max(k_1^{-s_1}\log(k_1), k_2^{-(s_2-1)/2}, k_3^{-(s_3-1)/2})) = o_P(1)
\end{aligned}$$

in view of Assumption C and $\Delta_n^2(r,z,x) \ge \lambda_{\min}^{\Xi}(\lambda_{\max}^{\Psi})^{-2}\frac{d_n}{n}\|Z_k(r,z,x)\|^2$, where again the convergence rate for $|\gamma_{ik}(r,z,x)|$, $i = 1,2,3$, can be found in Lemma A.5. Definitely, the rates are in point-wise sense as the convergence relates to the point (z,x). The first result holds.

(2) Observe that

$$\begin{aligned}
\widehat{m}_n(r,z,x) - m_0(r,z,x) =& Z_k(r,z,x)^\intercal (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_{nK}^\intercal W_n e \\
&+ Z_k(r,z,x)^\intercal (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_{nK}^\intercal W_n \gamma - \gamma_k(r,z,x),
\end{aligned}$$

and thus

$$\begin{aligned}
&\begin{pmatrix} \widehat{m}_n(r,z,x) - m_0(r,z,x) \\ \widehat{m}_n(r',z',x') - m_0(r',z',x') \end{pmatrix} \\
&= \sum_{t=1}^{n} \begin{pmatrix} Z_k(r,z,x)^\intercal \\ Z_k(r',z',x')^\intercal \end{pmatrix} (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_k(\tau_t, z_t, x_t)\phi(z_t, x_t) e_t \\
&\quad + \begin{pmatrix} Z_k(r,z,x)^\intercal \\ Z_k(r',z',x')^\intercal \end{pmatrix} (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_{nK}^\intercal W_n \gamma - \begin{pmatrix} \gamma_k(r,z,x) \\ \gamma_k(r',z',x') \end{pmatrix}.
\end{aligned}$$

While the last two terms are negligible similar to the proof of the first part, the normality is derived from the first term for which we call the leading term. Indeed, by the one-step approximation in Lemma A.4, the normalizer $\Lambda_n^2(r,z,x;r',z',x') = \lambda_n^2(r,z,x;r',z',x')(1+o_P(1))$ where

$$\lambda_n^2(r,z,x;r',z',x') = \sigma_e^2 \frac{d_n}{n} \begin{pmatrix} Z_k(r,z,x)^\intercal \\ Z_k(r',z',x')^\intercal \end{pmatrix} \Psi_K^{-1} \Xi_K \Psi_K^{-1} \begin{pmatrix} Z_k(r,z,x)^\intercal \\ Z_k(r',z',x')^\intercal \end{pmatrix}^\intercal.$$

Thence, after normalization the leading term is

$$\begin{aligned}
&\lambda_n^{-1}(r,z,x;r'z'x') \sum_{t=1}^{n} \begin{pmatrix} Z_k(r,z,x)^\intercal \\ Z_k(r',z',x')^\intercal \end{pmatrix} (Z_{nK}^\intercal W_n Z_{nK})^{-1} Z_k(\tau_t, z_t, x_t)\phi(z_t, x_t) e_t \\
&= \frac{d_n}{n} \lambda_n^{-1}(r,z,x;r'z'x') \begin{pmatrix} Z_k(r,z,x)^\intercal \\ Z_k(r',z',x')^\intercal \end{pmatrix} \sum_{t=1}^{n} \Psi_K^{-1} Z_k(\tau_t, z_t, x_t)\phi(z_t, x_t) e_t,
\end{aligned}$$

which by Assumption B is a two-dimensional martingale sequence and clearly has conditional variance I_2. Hence, using Cramér–Wold device and similar to the first part, it follows that

$$\Lambda_n^{-1}(r,z,x;r'z'x') \begin{pmatrix} \widehat{m}_n(r,z,x) - m_0(r,z,x) \\ \widehat{m}_n(r',z',x') - m_0(r',z',x') \end{pmatrix} \to_D N(0, I_2)$$

as $n \to \infty$. We omit the details for the similarity and the proof then is complete. □

Proof of Corollary 3.1. Note that

$$\begin{aligned}
\widehat{\sigma}_e^2 &= \left(\sum_{t=1}^{n} \phi(z_t, x_t) \right)^{-1} \sum_{t=1}^{n} \widehat{e}_t^2 \phi(z_t, x_t) \\
&= \left(\sum_{t=1}^{n} \phi(z_t, x_t) \right)^{-1} \sum_{t=1}^{n} [e_t + m(\tau_t, z_t, x_t) - \widehat{m}_n(\tau_t, z_t, x_t)]^2 \phi(z_t, x_t) \\
&= \left(\sum_{t=1}^{n} \phi(z_t, x_t) \right)^{-1} \sum_{t=1}^{n} e_t^2 \phi(z_t, x_t) \\
&\quad + \left(\sum_{t=1}^{n} \phi(z_t, x_t) \right)^{-1} \sum_{t=1}^{n} [m(\tau_t, z_t, x_t) - \widehat{m}_n(\tau_t, z_t, x_t)]^2 \phi(z_t, x_t) \\
&\quad + 2 \left(\sum_{t=1}^{n} \phi(z_t, x_t) \right)^{-1} \sum_{t=1}^{n} e_t [m(\tau_t, z_t, x_t) - \widehat{m}_n(\tau_t, z_t, x_t)] \phi(z_t, x_t).
\end{aligned}$$

It suffices to show that

$$\left(\sum_{t=1}^{n} \phi(z_t, x_t) \right)^{-1} \sum_{t=1}^{n} e_t^2 \phi(z_t, x_t) \to_P \sigma_e^2, \tag{B.3}$$

and

$$\left(\sum_{t=1}^{n} \phi(z_t, x_t) \right)^{-1} \sum_{t=1}^{n} [m(\tau_t, z_t, x_t) - \widehat{m}_n(\tau_t, z_t, x_t)]^2 \phi(z_t, x_t) = o_P(1) \tag{B.4}$$

as $n \to \infty$, since the last term is implied to be $o_P(1)$ by Cauchy–Schwarz and the convergence of the first two.

Note that

$$\left(\sum_{t=1}^{n}\phi(z_t,x_t)\right)^{-1}\sum_{t=1}^{n}e_t^2\phi(z_t,x_t)-\sigma_e^2=\left(\frac{d_n}{n}\sum_{t=1}^{n}\phi(z_t,x_t)\right)^{-1}\frac{d_n}{n}\sum_{t=1}^{n}(e_t^2-\sigma_e^2)\phi(z_t,x_t),$$

and similarly to Lemma A.4 we may show that $\frac{d_n}{n}\sum_{t=1}^{n}\phi(z_t,x_t)=\mathbb{E}[f(z_t)]\frac{d_n}{n}\sum_{t=1}^{n}w(x_t)(1+o_P(1))=O_P(1)$ by Lemma A.1 and $\frac{d_n}{n}\sum_{t=1}^{n}(e_t^2-\sigma_e^2)\phi(z_t,x_t)=o_P(1)$ by virtue of the martingale difference structure. Thus, the assertion (B.3) holds.

To show (B.4), observe that

$$\begin{aligned}
&\left(\sum_{t=1}^{n}\phi(z_t,x_t)\right)^{-1}\sum_{t=1}^{n}[m(\tau_t,z_t,x_t)-\widehat{m}_n(\tau_t,z_t,x_t)]^2\phi(z_t,x_t)\\
=&O_P\left(\frac{d_n}{n}\right)\sum_{t=1}^{n}[Z_k(\tau_t,z_t,x_t)^{\intercal}(c-\widehat{c})+\gamma_k(\tau_t,z_t,x_t)]^2\phi(z_t,x_t)\\
\le&O_P\left(\frac{d_n}{n}\right)\sum_{t=1}^{n}[Z_k(\tau_t,z_t,x_t)^{\intercal}(c-\widehat{c})]^2\phi(z_t,x_t)+O_P\left(\frac{d_n}{n}\right)\sum_{t=1}^{n}\gamma_k^2(\tau_t,z_t,x_t)\phi(z_t,x_t)\\
=&O_P(1)(c-\widehat{c})^{\intercal}\left(\frac{d_n}{n}Z_{nK}^{\intercal}W_nZ_{nK}\right)(c-\widehat{c})+O_P(1)\frac{d_n}{n}\sum_{t=1}^{n}\gamma_k^2(\tau_t,z_t,x_t)\phi(z_t,x_t)\\
=&O_P(1)(c-\widehat{c})^{\intercal}\Psi_K(c-\widehat{c})+O_P(1)\frac{d_n}{n}\sum_{t=1}^{n}\gamma_k^2(\tau_t,z_t,x_t)\phi(z_t,x_t)\\
=&O_P(1)\frac{d_n^2}{n^2}e^{\intercal}W_nZ_{nK}\Psi_K^{-1}Z_{nK}^{\intercal}W_ne+O_P(1)\frac{d_n}{n}\sum_{t=1}^{n}\gamma_k^2(\tau_t,z_t,x_t)\phi(z_t,x_t)\\
\le&O_P(1)\frac{d_n}{n}K\lambda_{\max}^{\Xi}(\lambda_{\min}^{\Psi})^{-1}+O_P(1)\frac{d_n}{n}\sum_{t=1}^{n}\gamma_k^2(\tau_t,z_t,x_t)\phi(z_t,x_t)=o_P(1),
\end{aligned}$$

where the first term is $o_P(1)$ derived from $\widehat{c}-c=(Z_{nK}^{\intercal}W_nZ_{nK})^{-1}Z_{nK}^{\intercal}W_ne$ and the martingale difference structure, while in the second term $\frac{d_n}{n}\sum_{t=1}^{n}\gamma_k^2(\tau_t,z_t,x_t)\phi(z_t,x_t)=o_P(1)$ as shown in the proof of Theorem 3.1. Therefore the corollary is complete. □

Proof of Theorem 3.2. Given Assumption C, the function m_0 is sufficiently smooth such that its series expansion, combined with the density in the integrals in the definition of δ's, is uniformly convergent. Hence, its derivatives can be calculated term by term and again due to the presence of the density the integrals can be computed termwise in what follows. Without loss of generality, assume $p_0(z)\equiv 1$.

(a) Notice that $H_i'(x)=2iH_{i-1}(x)$. By (2.3) we have

$$\begin{aligned}
\delta_1&=\int_0^1\int_{V\times\mathbb{R}}\frac{\partial m_0(r,z,x)}{\partial x}d\omega(r,z,x)=\sum_{i,j,\ell=0}^{\infty}c_{ij\ell}\int_0^1\int_{V\times\mathbb{R}}\frac{\partial\mathcal{B}_{ij\ell}(r,z,x)}{\partial x}d\omega(r,z,x)\\
&=\sum_{i,j,\ell=0}^{\infty}c_{ij\ell}\int_0^1\varphi_i(r)dr\int_V p_j(z)f(z)dz\int_{\mathbb{R}}\frac{\partial h_\ell(x)}{\partial x}e^{-x^2}dx\\
&=\sum_{\ell=1}^{\infty}\frac{c_{00\ell}}{(\sqrt{\pi}\,\ell!2^{\ell})^{-1/2}}\int_{\mathbb{R}}(2\ell)H_{\ell-1}(x)e^{-x^2}dx=c_{001}(\sqrt{\pi}2)^{-1/2}4\sqrt{\pi}=2\sqrt{2}\sqrt[4]{\pi}c_{001}.
\end{aligned}$$

Thus, $\widehat{\delta}_1-\delta_1=2\sqrt{2}\sqrt[4]{\pi}(\widehat{c}_{001}-c_{001})$.

Let $\ell_3=(0,0,1,0,\ldots,0)$ of dimension K. Then,

$$\widehat{c}_{001}-c_{001}=\ell_3^{\intercal}(\widehat{c}-c)=\ell_3^{\intercal}(Z_{nK}^{\intercal}W_nZ_{nK})^{-1}Z_{nK}^{\intercal}W_n(\gamma+e)$$

which is similar to the first two terms of $\widehat{m}_n(r,z,x)-m(r,z,x)$ in Theorem 3.1 but with $Z_k(r,z,x)$ being replaced by ℓ_3. Therefore, under the conditions of Theorem 3.1 and defining $B_{1n}^2:=\sigma_e^2\ell_3^{\intercal}\Omega_{nK}\ell_3$, we may show in the same fashion as Theorem 3.1 that

$$B_{1n}^{-1}(\widehat{c}_{001}-c_{001})\to_D N(0,1),$$

as $n\to\infty$ and hence the assertion holds.

(b) Observe that

$$\begin{aligned}\delta_2 &= \int_0^1 \int_{V\times\mathbb{R}} \frac{\partial m_0(r,z,x)}{\partial r} d\omega(r,z,x) = \sum_{i,j,\ell=0}^{\infty} c_{ij\ell} \int_0^1 \int_{V\times\mathbb{R}} \frac{\partial \mathcal{B}_{ij\ell}(r,z,x)}{\partial r} d\omega(r,z,x) \\ &= \sum_{i,j,\ell=0}^{\infty} c_{ij\ell} \int_0^1 \frac{d\varphi_i(r)}{dr} dr \int_V p_j(z) f(z) dz \int_{\mathbb{R}} h_\ell(x) e^{-x^2} dx \\ &= \sum_{i=1}^{\infty} c_{i00}[\varphi_i(1) - \varphi_i(0)] = \sqrt{2} \sum_{i=1}^{\infty} c_{i00}[(-1)^i - 1],\end{aligned}$$

by the orthogonality of $\{p_j(z)\}$ and $\{h_\ell(x)\}$. Thus, $\widehat{\delta}_2 = \sqrt{2}\sum_{i=1}^{k_1-1} \widehat{c}_{i00}[(-1)^i - 1]$. We then have

$$\begin{aligned}\widehat{\delta}_2 - \delta_2 =& \sqrt{2} \sum_{i=1}^{k_1-1} [(-1)^i - 1](\widehat{c}_{i00} - c_{i00}) - \sqrt{2} \sum_{i=k_1}^{\infty} c_{i00}[(-1)^i - 1] \\ :=& -2\sqrt{2} L_1^{\intercal}(\widehat{c} - c) + \tilde{\gamma}_{1k_1},\end{aligned}$$

where L_1 is defined to be a sparse column vector with dimension K where 1 conformably is in the place of c_{i00} (odd i) in c. Thus, there are about $[k_1/2]$ places that equal 1 and elsewhere 0; the definition of $\tilde{\gamma}_{1k_1}$ is clear, i.e. the residue of the series. Observe further that

$$\begin{aligned}\widehat{\delta}_2 - \delta_2 =& -2\sqrt{2} L_1^{\intercal}(\widehat{c} - c) + \tilde{\gamma}_{1k_1} \\ =& -2\sqrt{2} L_1^{\intercal}(Z_{nK}^{\intercal} W_n Z_{nK})^{-1} Z_{nK}^{\intercal} W_n(\gamma + e) + \tilde{\gamma}_{1k_1} \\ =& -2\sqrt{2} \frac{d_n}{n} L_1^{\intercal} \Psi_K^{-1} Z_{nK}^{\intercal} W_n(\gamma + e)(1 + o_P(1)) + \tilde{\gamma}_{1k_1},\end{aligned}$$

which has similar structure as $\widehat{m}_n(r,z,x) - m(r,z,x)$ in Theorem 3.1 but with $Z_k(r,z,x)$ being replaced by L_1.

Thus, define $B_{2n}^2 = \sigma_e^2 L_1^{\intercal} \Omega_{nK} L_1$. Following the same derivation as Theorem 3.1, we may show that $B_{2n}^{-1}(\widehat{\delta}_2 - \delta_2) \to_D N(0,8)$.

(c) Here, in order to implement the following calculation, we need to specify the orthogonal sequence $\{p_j(z)\}$, otherwise we have to make a great deal of assumptions on it. Let $\{p_j(z)\}$ be Hermite orthogonal polynomial sequence, i.e. the support of z_t is the entire real line. Similar to (a), we have $\delta_3 = 2\sqrt{2}\sqrt[4]{\pi} c_{010}$ and then $\widehat{\delta}_3 - \delta_3 = 2\sqrt{2}\sqrt[4]{\pi}(\widehat{c}_{010} - c_{010})$. Following exactly the same fashion as (a), we can have the assertion.

(d) Observe that

$$\begin{aligned}\delta_{12} &= \int_0^1 \iint \frac{\partial^2 m_0(r,z,x)}{\partial r \partial z} d\omega(r,z,x) = \sum_{i,j,\ell=0}^{\infty} c_{ij\ell} \int_0^1 \iint \frac{\partial^2 \mathcal{B}_{ij\ell}(r,z,x)}{\partial r \partial z} d\omega(r,z,x) \\ &= \sum_{i,j,\ell=0}^{\infty} c_{ij\ell} \int_0^1 \frac{d\varphi_i(r)}{dr} dr \int \frac{dp_j(z)}{dz} e^{-z^2} dz \int_{\mathbb{R}} h_\ell(x) e^{-x^2} dx \\ &= \sum_{i,j=1}^{\infty} c_{ij0}[\varphi_i(1) - \varphi_i(0)](\sqrt{\pi} j! 2^j)^{-1/2} \int (2j) H_{j-1}(z) e^{-z^2} dz \\ &= 4\sqrt[4]{\pi} \sum_{i=1}^{\infty} c_{i10}[(-1)^i - 1],\end{aligned}$$

similar to (a) and (b). Thus, we define $\widehat{\delta}_{12} = 4\sqrt[4]{\pi}\sum_{i=1}^{k_1-1} \widehat{c}_{i10}[(-1)^i - 1]$ and hence $\widehat{\delta}_{12} - \delta_{12} = 4\sqrt[4]{\pi}\sum_{i=1}^{k_1-1}(\widehat{c}_{i10} - c_{i10})[(-1)^i - 1] - 4\sqrt[4]{\pi}\sum_{i=k_1}^{\infty} c_{i10}[(-1)^i - 1]$. Therefore, the assertion follows in the same fashion as (b).

(e) Similarly, after the same calculation we define $\widehat{\delta}_{13} = 4\sqrt[4]{\pi}\sum_{i=1}^{k_1-1} \widehat{c}_{i01}[(-1)^i - 1]$ and $\widehat{\delta}_{13} - \delta_{13} = 4\sqrt[4]{\pi}\sum_{i=1}^{k_1-1}(\widehat{c}_{i01} - c_{i01})[(-1)^i - 1] - 4\sqrt[4]{\pi}\sum_{i=k_1}^{\infty} c_{i01}[(-1)^i - 1]$. Once again, the assertion follows in the same fashion as (b).

(f) By the similar calculation, $\delta_{23} = 8\sqrt{\pi} c_{011}$ and then define $\widehat{\delta}_{23} = 8\sqrt{\pi}\widehat{c}_{011}$. Then, the assertion follows similarly as (a). This finishes the proof. □

Proof of Theorem 3.3. The assertion about χ_0 follows exactly the same as that for δ_1 in Theorem 3.2, so we omit the proof.

Observe that

$$\widehat{\chi}_1(r) - \chi_1(r) = \int_{V\times\mathbb{R}} [\widehat{m}_n(r,z,x) - m_0(r,z,x)]f(z)w(x)dzdx$$
$$= \int_{V\times\mathbb{R}} Z_k(r,z,x)^{\intercal} f(z)w(x)dzdx(\widehat{c} - c) - \int_{V\times\mathbb{R}} \gamma_k(r,z,x)f(z)w(x)dzdx.$$

Note that $Z_k(r,z,x)$ has elements $\mathcal{B}_{ij\ell}(r,z,x) = \varphi_i(r)p_j(z)h_\ell(x)$ with $i \le k_1 - 1, j \le k_2 - 1$ and $\ell \le k_3 - 1$, and moreover, $\int_V p_j(z)f(z)dz = 0$ unless $j = 0$ and $\int h_\ell(x)w(x)dx = 0$ unless $\ell = 0$. Thus, the vector $\int_{V\times\mathbb{R}} Z_k(r,z,x)^{\intercal} f(z)w(x)dzdx$ reduces to $(\varphi_0(r), \ldots, \varphi_{k_1-1}(r), 0, \ldots, 0)$. Meanwhile, $\int_{V\times\mathbb{R}} \gamma_k(r,z,x)f(z)w(x)dzdx = \sum_{i=k_1}^{\infty} c_{i00}\varphi_i(r)$. This situation is the same as Theorem 3.2 since $\Phi_{k_1}(r)^{\intercal} := (\varphi_0(r), \ldots, \varphi_{k_1-1}(r), 0, \ldots, 0)$ plays the same role as ℓ_2. Let $A_{1n}^2(r) = \sigma_e^2 \Phi_{k_1}(r)^{\intercal} \Omega_{nK} \Phi_{k_1}(r)$. Then, similar to Theorem 3.2, we have

$$\frac{1}{A_{1n}(r)}(\widehat{\chi}_1(r) - \chi_1(r)) \to_D N(0,1)$$

as $n \to \infty$.

Using the orthogonality of the sequence and the similar argument, we may have the normality for $\widehat{\chi}_2(z) - \chi_2(z)$ and $\widehat{\chi}_3(x) - \chi_3(x)$, respectively. Precisely, let $P_{k_2}(z)^{\intercal} := (0, \ldots, 0, p_0(z), \ldots, p_{k_2-1}(z), 0 \cdots, 0)$ and $A_{2n}^2(z) := \sigma_e^2 P_{k_2}(z)^{\intercal} \Omega_{nK} P_{k_2}(z)$; $U_{k_3}(x) := (0, \ldots, 0, h_0(x), \cdots, h_{k_3-1}(x))^{\intercal}$ and $A_{3n}^2(x) := \sigma_e^2 U_{k_3}(x)^{\intercal} \Omega_{nK} U_{k_3}(x)$. Then,

$$\frac{1}{A_{2n}(z)}(\widehat{\chi}_2(z) - \chi_2(z)) \to_D N(0,1),$$
$$\frac{1}{A_{3n}(x)}(\widehat{\chi}_3(x) - \chi_3(x)) \to_D N(0,1),$$

as $n \to \infty$. □

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.03.024.

References

Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, 1795–1843.
Amreheim, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature 567, 305–307.
Andersen, T.G., Varneskov, R.T., 2018. Consistent Inference for Predictive Regressions in Persistent Economic Systems. CREATES Research Paper 2018-9.
Atak, A., Linton, O., Xiao, Z., 2011. A semiparametric panel model for unbalanced data with application to climate change in the United Kingdom. J. Econometrics 164, 92–115.
Belloni, A., Chernozhukov, V., Chetverikov, D., Kato, K., 2015. Some new asymptotic theory for least squares series: Pointwise and uniform results. J. Econometrics 186, 345–366.
Bosq, D., 1996. Nonparametric Statistics for Stochastic Processes: Estimation and Prediction. Springer, New York.
Cai, Z., Li, Q., Park, J., 2009. Functional-coefficient cointegration models for nonstationary time series data. J. Econometrics 148, 101–113.
Chang, J., Chen, S., Chen, X., 2015. High dimensional generalized empirical likelihood for moment restrictions with dependent data. J. Econometrics 185, 283–304.
Chang, Y., Park, J.Y., Phillips, P.C.B., 2001. Nonlinear econometric models with cointegrated and deterministically trending regressors. Econom. J. 4, 1–36.
Chen, X., 2007. In: Engle, R.F., MacFadden, D.L. (Eds.), Large Sample Sieve Estimation of Semi-Parametric Models. In: Handbook of Econometrics, vol. 6B, Elsevier, Amsterdam: North Holland, pp. 5550–5588.
Chen, X., Christensen, T., 2015. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. J. Econometrics 188, 447–465.
Chen, X., Shen, X., 1998. Sieve extremum estimates for weakly dependent data. Econometrica 66, 289–314.
Cheng, T., Gao, J., Linton, O., 2019. Nonparametric Predictive Regressions for Stock Return Prediction. Cambridge Working Papers in Economics 1932, Faculty of Economics, University of Cambridge.
Dong, C., Gao, J., 2018. Specification testing driven by orthogonal series for nonlinear cointegration with endogeneity. Econometric Theory 34, 754–789.
Dong, C., Gao, J., Peng, B., 2015. Semiparametric single-index panel data models with cross-sectional dependence. J. Econometrics 188, 301–312.
Dong, C., Gao, J., Tjøstheim, D., 2016. Estimation for single-index and partially linear single-index integrated models. Ann. Statist. 44, 425–453.
Dong, C., Linton, O., 2018. Additive nonparametric models with time variable and both stationary and nonstationary regressors. J. Econometrics 207, 212–236.
Erdelyi, A., Magnus, W., Oberhettinger, F., Tricomi, F., 1953. Higher Transcendental Functions, Vol. 2. McGraw-Hill, New York.
Gao, J., Tong, H., Wolff, R., 2001. Adaptive orthogonal series estimation in additive stochastic regression models. Statist. Sinica 11, 409–428.
Gao, J., Tong, H., Wolff, R., 2002. Model specification tests in nonparametric stochastic regression models. J. Multivariate Anal. 83, 324–359.
Gautschi, W., 2004. Orthogonal Polynomials: Computation and Approximation. In: Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford.
Hall, P., Heyde, C.C., 1980. Martingale Limit Theory and its Application. Academic Press, New York.
Hansen, B.E., 2015. A Unified Asymptotic Distribution Theory for Parametric and Nonparametric Least Square. Working paper, University of Wisconsin.
Härdle, W., Linton, O., 1994. In: Engle, R.F., MacFadden, D.L. (Eds.), Applied Nonparametric Methods. In: Handbook of Econometrics, vol. 4, Elsevier, pp. 2295–2339.
Hualde, J., Robinson, P.M., 2011. Gaussian pseudo-maximum likelihood estimation of fractional time series models. Ann. Statist. 39, 3152–3181.
Indritz, J., 1961. An inequality for Hermite polynomials. Proc. Amer. Math. Soc. 12, 981–983.

Jackson, D., 1930. The Theory of Approximation, Vol. XI. Amer. Math. Soc. Colloquium Publications, New York.
Karlsen, H.A., Mykelbust, T., Tjøstheim, D., 2007. Nonparametric estimation in a nonlinear cointegration type model. Ann. Statist. 35, 252–299.
Karlsen, H.A., Tjøstheim, D., 2001. Nonparametric estimation in null recurrent time series. Ann. Statist. 29, 372–416.
Kasparis, I., Andreou, E., Phillips, P.C.B., 2015. Nonparametric predictive regression. J. Econometrics 185, 468–494.
Levin, A.L., Lubinsky, D.S., 1992. Christoffel functions, orthogonal polynomails, and nevai's conjecture for freud weights. Constr. Approx. 8, 463–535.
Levin, E., Lubinsky, D.S., 2001. Orthogonal Polynomials for Exponential Weights. Springer-Verlag, New York.
Li, D., Phillips, P.C.B., Gao, J., 2016a. Uniform consistency of nonstationary kernel-weighted sample covariance for nonparametric regressions. Econometric Theory 32, 655–685.
Li, Q., Racine, J.S., 2007. Nonparametric Econometrics Theory and Practice. Princeton University Press.
Li, D., Tjøstheim, D., Gao, J., 2016b. Estimation in nonlinear regression with Harris recurrent Markov chains. Ann. Statist. 44 (5), 1957–1987.
Linton, O., Nielsen, J.P., 1995. A kernel method of estimating nonparametric structured regression based on marginal integration. Biometrika 82, 93–100.
Linton, O., Xiao, Z., 2019. Efficient estimation of nonparametric regression in the presence of dynamic heteroskedasticity. J. Econometrics 213, 608–631.
Lubinsky, D.S., 2009. A new approach to universality limits involving orthogonal polynomials. Ann. of Math. 170, 915–939.
Lynch, P., 2006. The Emergence of Numerical Weather Prediction: Richardson's Dream. Cambridge University Press.
Marinucci, D., Robinson, P.M., 2001. Semiparametric fractional cointegration analysis. J. Econometrics 105, 225–247.
Máté, A., Nevai, P., Totic, V., 1991. Szegö's extremum problem on the unit circle. Ann. of Math. 134, 433–453.
Mykelbust, T., Karlsen, H.A., Tjøstheim, D., 2012. Null recurrent unit root processes. Econometric Theory 28, 1–41.
Nevai, P., 1986. Geza frued, orthogonal polynomials and christoffel functions. a case study. J. Approx. Theory 48, 3–167.
Newey, W.K., 1997. Convergence rates and asymptotic normality for series estimators. J. Econometrics 79, 147–168.
Park, J., Hahn, S., 1999. Cointegrating regressions with time varying coefficients. Econometric Theory 15, 664–703.
Park, J.Y., Phillips, P.C.B., 1999. Asymptotics for nonlinear transformations of integrated time series. Econometric Theory 15, 269–298.
Park, J.Y., Phillips, P.C.B., 2001. Nonlinear regression with integreted time series. Econometrica 69, 117–161.
Phillips, P.C.B., 2005. HAC estimation by automated regression. Econometric Theory 21, 116–142.
Phillips, P.C., Lee, J.H., 2013. Predictive regression under various degrees of persistence and robust long-horizon regression. J. Econometrics 177, 250–264.
Phillips, P.C.B., Li, D., Gao, J., 2017. Estimating smooth structure change in cointegration models. J. Econometrics 196, 180–195.
Phillips, P.C.B., Solo, V., 1992. Asymptotics for linear processes. Ann. Statist. 20, 971–1001.
Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.
Schienle, M., 2008. Nonparametric Nonstationary Regression (Doctoral thesis). Mannheim University, Mannheim, Germany.
Schultz, M., 1969. L^{∞}-multivariate approximation theory. J. Numer. Anal. 6, 161–183.
Stone, C.J., 1980. Optimal rates of convergence for nonparametric estimators. Ann. Statist. 8, 1348–1360.
Stone, C.J., 1985. Additive regression and other nonparametric models. Ann. Statist. 13, 689–705.
Todd, J., 1963. Introduction to the Constructive Theory of Functions. Academic Press, New York.
Visser, H., Dangendorf, S., Petersen, A.C., 2015. A review of trend models applied to sea level data with reference to the "acceleration-deceleration debate". J. Geophys. Res.: Oceans 120, 3873–3895.
Wang, Q., 2014. Martingale limit theorem revisited and nonlinear cointegrating regression. Econometric Theory 30, 509–535.
Wang, Q., 2015. Limit Theorems for Nonlinear Cointegrating Regression. In: Nonlinear Time Series and Chaos, World Scentific, Singarpore.
Wang, Q., Phillips, P.C.B., 2009. Asymptotic theory for local time density estimation and nonparametric cointegreting regression. Econometric Theory 25, 710–738.
Wang, Q., Phillips, P.C.B., 2016. Nonparametric cointegrating regression with endogeneity and long memory. Econometric Theory 32, 359–401.
Xiao, Z., 2009. Functional-coefficient cointegration models. J. Econometrics 152, 81–92.