Estimation for Double-Nonlinear Cointegration

Yingqian Lin[a], Yundong Tu[a,*] and Qiwei Yao[b]
[a]Guanghua School of Management and
Center for Statistical Science, Peking University, China
[b]Department of Statistics,
London School of Economics, U.K.

June 5, 2019

Abstract

In recent years statistical inference for nonlinear cointegration has attracted attention from both academics and practitioners. This paper proposes a new type of cointegration in the sense that two univariate time series y_t and x_t are cointegrated via two (unknown) smooth nonlinear transformations, further generalizing the notion of cointegration initiated revealed by Box and Tiao (1977), and more systematically studied by Engle and Granger (1987). More precisely, it holds that $G(y_t, \beta) = g(x_t) + u_t$, where $G(\cdot, \beta)$ is strictly increasing and known upto an unknown parameter β, $g(\cdot)$ is unknown and smooth, x_t is $I(1)$, and u_t is the stationary disturbance. This setting nests the nonlinear cointegration model of Wang and Phillips (2009b) as a special case with $G(y, \beta) = y$. It extends the model of Linton et al. (2008) to the cases with a unit-root nonstationary regressor. Sieve approximations to the smooth nonparametric function g are applied, leading to an extremum estimator for β and a plugging-in estimator for $g(\cdot)$. Asymptotic properties of the estimators are established, revealing that both the convergence rates and the limiting distributions depend intimately on the properties of the two nonlinear transformation functions. Simulation studies demonstrate that the estimators perform well even with small samples. A real data example on the environmental Kuznets curve portraying the nonlinear impact of per-capita GDP on air-pollution illustrates the practical relevance of the proposed double-nonlinear cointegration.

JEL classification: C14, C22, Q53.

Keywords: Box-cox transformation; Nonlinear cointegration; Semiparametrics; Sieve method; Transformation models.

*Corresponding author. Address: Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, 100871, China. E-mail: yundong.tu@gsm.pku.edu.cn.

1 Introduction

The phenomenon that there exist stable linear relationships among nonstationary time series was illustrated first by Box and Tiao (1977), and was later coined as "cointegration" after the seminal work of Granger (1981) and Engle and Granger (1987). This concept has proved to be important in both econometric theory and economic application, and has been one of the most active research areas in the past 30 years. For earlier development of cointegration, see Johansen (1995) for an excellent survey. Recently, Liao and Phillips (2015) considered automated estimation of vector error correction models using adaptive shrinkage. Tu and Yi (2017) considered model averaging estimation in cointegrated vector autoregressive systems. Zhang et al. (2019) dealt with the cointegration of the processes with different integration orders in a high dimensional setting. Tu et al. (2019) studied the error correction factor models in high dimensional cointegration models. For further information on linear cointegration models, see the above papers and the references therein.

Nonlinear cointegration started to attract the research attention since Granger (1991). Building on the framework of Park and Phillips (1999) that developed an asymptotic theory for stochastic processes generated from nonlinear transformations of integrated time series, Park and Phillips (2000) studied the binary choice models with integrated regressors, and Park and Phillips (2001) considered parametric nonlinear regression with integrated processes. Furthermore, Chang and Park (2003) studied index models with integrated processes, which include as special cases the simple neural network models and the smooth transition regressions.

Besides the above effort in building parametric nonlinear cointegration, the recent literature has witnessed a surge of interest in developing nonparametric and semiparametric cointegration models. Karlsen et al. (2007), Wang and Phillips (2009b,a, 2016) and Linton and Wang (2016) considered kernel estimation of nonparametric cointegration models. Cai et al. (2009), Xiao (2009), Gao and Phillips (2013b), Hirukawa and Sakudo (2018) and Tu and Wang (2019) considered functional coefficient cointegration models. Gao and Phillips (2013a) studied the semiparametric estimation in triangular system equations with nonstationarity. Dong et al. (2016b) considered a semiparametric single index model with integrated regressors. Phillips et al. (2017) studied estimation of smooth structural change in cointegration models. Dong and Linton (2018) studied non-

parametric regression with time variable, nonstationary and stationary variables. Gao et al. (2009), Wang and Phillips (2012), Wang et al. (2018) and Dong and Gao (2018) considered specification test in nonlinear cointegration models. Kasparis and Phillips (2012) considered dynamic misspecification test in nonparametric cointegration models. Kasparis et al. (2015) studied inferences in nonparametric predictive regressions. Phillips (2009) studied spurious regression in nonparametric regression with integrated processes, Tu and Wang (2018) considered spurious regression in functional coefficient regressions with integrated processes and provide a robust solution for spurious detection.

This paper aims to complement the growing literature on cointegration by considering a double-nonlinear cointegration model, in which the dependent variable and the integrated regressor are cointegrated after possible double nonlinear transformations. Our setting is quite general and it nests the models considered in Park and Phillips (2001) and Wang and Phillips (2009b) as special cases with transformation function $G(y, \beta) = y$. It also extends Linton et al. (2008) in which semiparametric transformations are analyzed for random samples, to the case that incorporates a nonstationary integrated regressor. The motivation for such a development is to extend further the notion of cointegration that there exist stable relationships among nonstationary variables (Box and Tiao, 1977; Engle and Granger, 1987). In the current setup, this relationship is described through the nonlinear transformations of both the dependent variable and the regressor, unlike the nonlinear cointegration of earlier studies where transformation is only applied to the regressor.

Related to this paper is a large literature on transformation models. Bickel and Doksum (1981) provided asymptotic properties of the maximum likelihood estimators in the linear regression model where the dependent variable is subject to a Box-Cox transformation (Box and Cox, 1964). Han (1987) provided an improved nonparametric estimator of this model based on the rank correlation. Carroll and Ruppert (1984) proposed parametric transformations of both sides of the regression, which is later generalized by Ramsay (1988) and Wang and Ruppert (1995, 1996) to the case where the transformation of the dependent variable is nonparametric, and is further relaxed to nonparametric transformations of both sides of the regression by Breiman and Friedman (1985) and Tibshirani (1988). Chen (2002), Horowitz (1996) and Ye and Duan (1997) proposed $\sqrt{n}$-consistent semiparametric estimators for a linear regression model where the dependent variable is transformed by an unknown monotonic function. Abrevaya (1999) considered a rank esti-

mator of the transformation model with observed truncation. Fan and Fine (2013) considered linear transformation models with parametric covariate transformation. Chiappori et al. (2015) studied identification and estimation of nonparametric transformations and Lewbel et al. (2015) provided a specification test for such a nonparametric transformation model. More recently, Florens and Sokullu (2017), Vanhems and Van Keilegom (2018), and Lin and Tu (2019) studied semiparametric transformation models in the presence of endogeneity. For more references on this literature, see Lin and Tu (2019) and references therein.

This paper contributes to the literature in several aspects. First, we propose an estimation strategy for the double-nonlinear cointegration model. To begin with, we propose to approximate the unknown transformation on the integrated process using Hermite polynomials. For a given parameter value in the transformation of the dependent variable, we can estimate the unknown coefficients in the Hermite expansion using the least squares method. Then, we estimate the parameter in the transformation of the dependent variable using a semiparametric least squares criterion, which is similar to the least squares objective function used by Breiman and Friedman (1985). The loss measures the relative variation of the regression residual compared to the variation in the transformed dependent variable. In this sense, the parametric estimator is to maximize the goodness-of-fit in the relationship described by the transformations.

Secondly, we establish the consistency and asymptotic distribution for the parametric estimator and the sieve estimator for the unknown transformation function. The parametric estimator is super consistent, with the rate of convergence depending on the property of the unknown transformations. The derivations build on Park and Phillips (2001) and Chan and Wang (2015), which laid down the foundation for nonlinear regressions with integrated series. The sieve estimator is shown to be asymptotically standard normal, after self-normalization. This result complements those derived in Dong et al. (2016b) for semiparametric regressions with integrated processes.

Finally, numerical studies illustrate the merit of our proposed estimators. We carry out simulation experiments to examine the finite sample performance of the proposed estimators. Results show that the biases of the proposed estimators are small, and their variances decay to zero fast as the sample size increases. These findings confirm that our estimators are super consistent. A real data example on environmental Kutznets curve is also included to demonstrate the practical value of our proposed model.

The rest of this article is organized as follows. Section 2 presents our model and the estimation procedure. Section 3 establishes the large sample properties of the proposed estimators. Section 4 illustrates the finite sample performance of the estimators using Monte Carlo experiments, and provides also a real data example. Section 5 concludes the paper with remarks on future studies. The proofs for the main results are relegated to the Appendix, with the technical details contained in an online supplementary document.

Notations. $\mathbb{R}$ denotes the real line and $\mathbb{R}_+$ its positive part. Convergence in probability and convergence in distribution are signified, respectively, as $\xrightarrow{p}$ and $\Rightarrow$.

2 Model and Estimation

We consider a semiparametric transformation model

$$G(y_t, \beta_0) = g(x_t) + u_t, \tag{2.1}$$

where the dependent variable y_t, after a strictly increasing transformation specified by the parametric family $\{G(y, \beta) : \beta \in \Theta\}$, is related to the univariate unit root regressor x_t via an unknown link function $g : \mathbb{R} \to \mathbb{R}$, the true parameter β_0 is assumed to be an interior point of a compact set $\Theta \subset \mathbb{R}$, and the innovation u_t is a stationary sequence.

The model in (2.1), which is referred to as the double-nonlinear cointegration model, is quite general. It nests many popularly studied models in the literature as special cases. For example, when $G(y, \beta) = y$, this model reduces to the nonparametric cointegration models of Wang and Phillips (2009a,b), and it also includes the nonlinear nonstationary regression models of Park and Phillips (2001), Chan and Wang (2015) and Uematsu (2017) when the g function is parametrically specified. In addition, when g is a linear parametric function, this model reduces to the linear cointegration model of Engle and Granger (1987). Furthermore, this model extends the semiparametric transformation model of Linton et al. (2008) to the case where x_t is unit root nonstationary. In the case where β_0 is known, the model may be analyzed either using the kernel method as in Wang and Phillips (2009b, 2016), or the sieve approach as in Dong et al. (2016b), Dong and Linton (2018). Here we take the latter approach, because the analysis in the sieve framework is much simpler when β_0 is unknown.

We assume that the link function $g(\cdot)$ belongs to a Hilbert space, $L^2(\mathbb{R}, e^{-x^2/2}) = \{g(x) : \int_{\mathbb{R}} g^2(x) e^{-x^2/2} dx < \infty\}$, with inner product given by $\langle f_1, f_2 \rangle = \int f_1(x) f_2(x) e^{-x^2/2} dx$

and the induced norm $\|f\|^2 = \langle f, f\rangle$. Note that the Hilbert space $L^2(\mathbb{R}, e^{-x^2/2})$ covers all polynomials, all power functions and all bounded functions on $\mathbb{R}$, to name a few. Note that Hermite orthogonal polynomial sequence $\{H_j(x)\}$ is a complete orthogonal basis in $L^2(\mathbb{R}, e^{-x^2/2})$. Recall that the Hermite polynomials $\{H_j(x)\}$ are defined by

$$H_j(x) = (-1)^j \exp\Big(\frac{x^2}{2}\Big)\frac{d^j}{dx^j}\exp\Big(-\frac{x^2}{2}\Big), \quad j = 0, 1, 2, \cdots,$$

and $\langle H_i(x), H_j(x)\rangle = \sqrt{2\pi} j! \delta_{ij}$, where δ_{ij} is the Kronecker delta. Let

$$h_j(x) = (j!)^{-1/2} H_j(x), \quad j \geq 0.$$

Then for any continuous function $g(x) \in L^2(\mathbb{R}, e^{-x^2/2})$, it holds that

$$g(x) = \sum_{j=0}^{\infty} c_j h_j(x), \qquad c_j = \frac{1}{\sqrt{2\pi}}\langle g, h_j\rangle. \tag{2.2}$$

For any integer $k \geq 1$, let $g_k(x) = \sum_{j=0}^{k-1} c_j h_j(x)$ be the truncated expansion, and $\gamma_k(x) = g(x) - g_k(x)$ be the residue after truncation.

By virtue of (2.2), model (2.1) can be written as

$$G(y_t, \beta_0) = Z_k(x_t)^T c + \gamma_k(x_t) + u_t, \tag{2.3}$$

where $Z_k(\cdot) = (h_0(\cdot), \cdots, h_{k-1}(\cdot))^T$, $c = (c_0, \cdots, c_{k-1})^T$ and k is the truncation parameter. With observations $\{x_t, y_t\}_{t=1}^n$, let $\boldsymbol{G}(\beta) = (G(y_1, \beta), \cdots, G(y_n, \beta))^T$, $Z = (Z_k(x_1), \cdots, Z_k(x_n))^T$. Hence, the ordinary least squares (OLS) estimator for c is,

$$\widetilde{c} = \widetilde{c}(\beta_0) = (Z^T Z)^{-1} Z^T \boldsymbol{G}(\beta_0), \tag{2.4}$$

which depends on β_0. As a result, the sieve estimator for the link function g is $\widetilde{g}(x) = Z_k^T(x)\widetilde{c}$.

However, the above OLS estimator is infeasible as the transformation parameter β_0 is unknown. Put

$$L_n(\beta) = \frac{\sum_{t=1}^n [G(y_t, \beta) - Z_k^T(x_t)\widetilde{c}(\beta)]^2}{\sum_{t=1}^n G(y_t, \beta)^2}, \tag{2.5}$$

where $\widetilde{c}(\cdot)$ is defined as in (2.4). An estimator for β_0 is then defined as

$$\widehat{\beta}_n = \arg\min_{\beta \in \Theta} L_n(\beta).$$

Consequently a plug-in estimator for g is defined as $\widehat{g}(x) = Z_k^T(x)\widehat{c}$, where $\widehat{c} = \widetilde{c}(\widehat{\beta}_n)$.

The loss function in (2.5) has a normalizing denominator, and is different from that used for the standard nonlinear regressions (e.g. Park and Phillips (2001), and Chan and Wang (2015)). Since $g(\cdot)$ is completely unspecified, the direct least squares estimation (i.e. without the normalizing denominator) for model (2.1) tends to choose β such that $G(\cdot, \beta)$ is flat with little variation. Furthermore, the normalization excludes the trivial specification $G(y, \beta) = \beta y$ or y/β, under which the loss function in (2.5) is invariant to β. Nevertheless, such a model may be simply analyzed without the transformation on y, as originally considered by Wang and Phillips (2009b). Finally, the minimization of (2.5) is effectively to choose $G(\cdot, \beta)$ and $g(\cdot)$ such that the squared regression correlation coefficient is maximized. See Breiman and Friedman (1985) for a similar objective function used in estimating nonparametric transformation models.

3 Asymptotic Theory

3.1 Functional classes

Let $\mathcal{F}_{LB}^0$ be the class of locally bounded functions that are exponentially bounded, i.e., f fulfills condition $f(x) = O(e^{c|x|})$ as $|x| \to \infty$ for some $c \in \mathbb{R}_+$. Let $\mathcal{F}_B^0$ denote the class of functions that are bounded and vanish at infinity in the sense that $f(x) \to 0$ as $|x| \to \infty$.

Definition 3.1 *A function $g(x)$ is called* H-regular *if it satisfies the following conditions.*

(a) $g(\lambda x) = \kappa(\lambda)h(x) + R(x, \lambda)$ *with $h(x)$ being a continuous function,*

and either

(b.i) $|R(x, \lambda)| \leq a(\lambda)P(x)$ *with* $\limsup_{\lambda\to\infty} |\kappa(\lambda)^{-1}a(\lambda)| = 0$, *or*

(b.ii) $|R(x, \lambda)| \leq b(\lambda)P(x)Q(\lambda x)$ *with* $\limsup_{\lambda\to\infty} |\kappa(\lambda)^{-1}b(\lambda)| < \infty$,

where

(c) $P(x) \in \mathcal{F}_{LB}^0$ *and* $Q(x) \in \mathcal{F}_B^0$.

The continuity requirement on h in Definition 3.1 (a) is somewhat stronger than that imposed in the H-regular class defined in Definition 4.2 of Park and Phillips (1999).

However, this condition is satisfied by most functions used in practical nonlinear time series analyses, including polynomial functions, logarithmic functions, etc.

Definition 3.2 *A function h is called* regular *on Θ if it satisfies the following conditions.*

(a) For all $\beta \in \Theta$, $h(\cdot, \beta)$ is continuous and twice differentiable;

(b) For all $x \in \mathbb{R}$, $h(x, \cdot)$ and $h'(x, \cdot)$ are equicontinuous in a neighborhood of x, where $h'(x, \cdot) = \partial h(x, \cdot)/\partial x$.

Definition 3.3 *A function g is called H-*regular *on Θ if it satisfies the following conditions.*

(a) $g(\lambda x, \beta) = \kappa(\lambda, \beta)h(x, \beta) + R(x, \lambda, \beta)$ with h being regular on Θ, and $R(x, \lambda, \beta)$ is differentiable with respect to x. $R(x, \lambda, \beta)$ and its first-order partial derivative $R_1(x, \lambda, \beta) \equiv \partial R(x, \lambda, \beta)/\partial x$ satisfy either

(b.i) $|R(x, \lambda, \beta)| \leq a(\lambda, \beta)P(x, \beta)$ with $\limsup_{\lambda\to\infty} \sup_{\beta\in\Theta} |\kappa^{-1}(\lambda, \beta)a(\lambda, \beta)| = 0$, $|R_1(x, \lambda, \beta)| \leq a_1(\lambda, \beta)P_1(x, \beta)$ with $\limsup_{\lambda\to\infty} \sup_{\beta\in\Theta} |\kappa^{-1}(\lambda, \beta)a_1(\lambda, \beta)| = 0$, or

(b.ii) $|R(x, \lambda, \beta)| \leq b(\lambda, \beta)P(x, \beta)Q(\lambda x, \beta)$ with $\limsup_{\lambda\to\infty} \sup_{\beta\in\Theta} |\kappa^{-1}(\lambda, \beta)b(\lambda, \beta)| < \infty$, $|R_1(x, \lambda, \beta)| \leq b_1(\lambda, \beta)P_1(x, \beta)Q_1(\lambda x, \beta)$ with $\limsup_{\lambda\to\infty} \sup_{\beta\in\Theta} |\kappa^{-1}(\lambda, \beta)b_1(\lambda, \beta)| < \infty$, where

(c) $\sup_{\beta\in\Theta} P(\cdot, \beta), \sup_{\beta\in\Theta} P_1(\cdot, \beta) \in \mathcal{F}^0_{LB}$ and $\sup_{\beta\in\Theta} Q(\cdot, \beta), \sup_{\beta\in\Theta} Q_1(\cdot, \beta) \in \mathcal{F}^0_B$.

We call κ the asymptotic order, h the limit homogeneous function and R the remainder function of g. Roughly speaking, the class of H-regular functions consists of functions that are asymptotically equivalent to their (uniquely defined) limit homogeneous functions. Condition (b) and (c) allow us to establish this asymptotic equivalence. The regularity requirement for the limit homogeneous function h in the condition (a) is necessary to ensure that h has well defined asymptotics. The regularity conditions in Definition 3.2 and Definition 3.3 are stronger than the corresponding conditions introduced in Park and Phillips (2001, Definition 3.2, Definition 3.5) and Uematsu (2017, Definition 7.1, Definition 7.2). In particular, smooth conditions on the regular function h and the remainder function R are imposed for technical convenience.

3.2 Assumptions

The following assumptions are needed for the theoretical development.

Assumption 1

(a) There exists a filtration $\{\mathcal{F}_{nt}\}$ *such that* $\{u_t, \mathcal{F}_{nt}\}$ *is a martingale difference sequence with* $E(u_t^2|\mathcal{F}_{n,t-1}) = \sigma_u^2$, $E(u_t^4|\mathcal{F}_{n,t-1}) = \mu_4$ *almost surely for* $t = 1, 2, \cdots, n$, *and* $\sup_{1\le t\le n} E(|u_t|^q|\mathcal{F}_{n,t-1}) < \infty$ *for some* $q > 4$.

(b) $x_t = x_{t-1} + v_t$ *for* $t \ge 1$ *and* $x_0 = O_p(1)$.

(c) x_t *is adapted to* $\mathcal{F}_{n,t-1}$, $t = 1, 2, \cdots$.

(d) Let $U_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t$ *and* $V_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} v_t$. *Suppose that* $(U_n(r), V_n(r)) \to (U(r), V(r))$ *as* $n \to \infty$. *Here,* $(U(r), V(r))$ *is a vector of Brownian motion.*

Assumption 2 Θ *is a convex and compact set and in* $\mathbb{R}$ *and* β_0 *is an interior point of* Θ.

Remark 3.1 *Conditions in Assumption 1 are commonly used in the literature on non-stationary processes. Assumption 1 (a) assumes that the error is a martingale difference sequence. Like linear cointegrating regression theory, serial correlation in the errors is allowed in our model. For instance, for an MA(1) process* $u_t = \varepsilon_t + \rho_1 \varepsilon_{t-1}$, *where* $\{\varepsilon_t\}$ *is a sequence of independent white noises, the MDS assumption can be made satisfied with the choice of* $\mathcal{F}_{nt} = (\varepsilon_{t-2}, \varepsilon_{t-3}, \cdots)$. *Note that the correlations do not affect the consistency of the estimator, but generally affect the limiting distribution theory. Assumption 1 (b) stipulates that the regressor* x_t *is an integrated process. See Park and Phillips (2001), Uematsu (2017), Tu and Wang (2019) for similar settings. Under Assumption 1 (c),* x_t *becomes predetermined. This condition can be simply satisfied with* $\mathcal{F}_{nt} = \{x_0, u_1, \cdots, u_t, v_1, \cdots, v_{t+1}\}$. *Assumption 2 is commonly used for the parametric space.*

Remark 3.2 *In Assumption 1, the requirement that the partial sum of* v_t *converges to a continuous Brownian Motion is quite weak and permits a variety of innovations that may have serial correlation. For example, for a linear process,* $v_t = \sum_{i=0}^{\infty} \phi_i \epsilon_{t-i}$, *where* $\{\epsilon_i, -\infty < i < \infty\}$ *is a sequence of i.i.d. random variables with* $E\epsilon_0 = 0$, $E\epsilon_0^2 = 1$,

$\sum_{i=0}^{\infty} i|\phi_i| < \infty$ *and* $\phi \equiv \sum_{i=0}^{\infty} \phi_i \neq 0$*, we have* $V_n(r) \Rightarrow \phi V_0(r) \equiv V(r)$*, where* $V_0(r)$ *is a standard Brownian Motion. Thus, Assumption 1* (d) *is easily satisfied.*

Furthermore, the unit root assumption on x_t *can be further relaxed to a general non-stationary process as stated in Assumption 3.3 of Chan and Wang (2015). The subsequent limiting theory will continue to hold with some modifications in the proof. However, the case when* x_t *has a drift or time trend would lead to different asymptotics, the results of which will be reported separately elsewhere.*

Remark 3.3 *The stochastic vector process* $(U_n(r), V_n(r))$ *takes values in* $D[0,1]^2$*, where* $D[0,1]$ *denotes the space of* cadlag *functions defined on the interval* $[0,1]$*. It follows from Skorohod representation theorem (e.g., Pollard (1984), pp.71-72) that there exists* $(U_n^0(r), V_n^0(r))$ *in a richer probability space such that* $(U_n(r), V_n(r)) \stackrel{d}{=} (U_n^0(r), V_n^0(r))$*, where* $\stackrel{d}{=}$ *signifies equivalence in distribution and for which* $(U_n^0(r), V_n^0(r)) \stackrel{a.s.}{\to} (U_n(r), V_n(r))$ *uniformly on* $[0,1]^2$*. For our purpose, it causes no loss of generality to assume* $(U_n(r), V_n(r)) = (U_n^0(r), V_n^0(r))$ *instead of* $(U_n(r), V_n(r)) \stackrel{d}{=} (U_n^0(r), V_n^0(r))$*. This convention will be made throughout the paper. It helps us to avoid repetitious embedding of* $(U_n(r), V_n(r))$ *in the probability space where* $(U_n^0(r), V_n^0(r))$ *is defined. For more details, see the discussions in Park and Phillips (1999), Park and Phillips (2001) and Dong et al. (2016b).*

Assumption 3

(a) *The link function* $g(x) \in L^2(\mathbb{R}, e^{-x^2/2})$ *is differentiable up to* m*-th order on* $\mathbb{R}$ *and* $g^{(m)}(x) \in L^2(\mathbb{R}, e^{-x^2/2})$*.*

(b) $g(x)$ *is an H-regular function with asymptotic order* $\kappa_g(\cdot)$ *and limiting homogeneous function* $h_g(x)$*.*

(c) *There exists* $m_0 > 0$ *such that* $\max_{1 \le t \le n} E[\gamma_k^2(x_t)] = O(k^{-m_0})$*.*

(d) *The sieve order* k *diverges with* n *such that* $k/n \to 0$*,* $n/k^{m_0} \to 0$*,* $n/k^{m+1} \to 0$*, as* $n \to \infty$*, where* m *and* m_0 *are defined as in Assumption 3* (a) *and* (c)*, respectively.*

Remark 3.4 *The smoothness condition of* $g(\cdot)$ *in Assumption 3* (a) *is standard in the literature and ensures the negligibility of the truncation residuals. See Dong et al. (2015, 2016a) for similar treatment. Assumption 3* (b) *requires that* $g(x)$ *is an H-regular function defined in Definition 3.1. Assumption 3* (c) *and Assumption 3* (d) *ensure not only that*

the residual term in the sieve approximation for g is sufficiently small, but also that it can be smoothed out when we establish the asymptotic normality. This is the so-called "under-smoothing" condition in the literature. See Dong et al. (2015, 2016a) for similar conditions.

For the ease of presentation, we define $\xi(x,\beta) \equiv \dot{G}(G^{-1}(x,\beta_0),\beta)$, $\dot{\xi}(x,\beta) \equiv \ddot{G}(G^{-1}(x,\beta_0),\beta)$ $\ddot{\xi}(x,\beta) \equiv \dddot{G}(G^{-1}(x,\beta_0),\beta)$, where $G^{-1}(x,\beta)$ is the inverse of $G(x,\beta)$ with respect to x, $\dot{G}(x,\beta) = \partial G(x,\beta)/\partial\beta$, $\ddot{G}(x,\beta) = \partial^2 G(x,\beta)/\partial\beta^2$, and $\dddot{G}(x,\beta) = \partial^3 G(x,\beta)/\partial\beta^3$.

Assumption 4

(a) $\{G(\cdot,\beta) : \beta \in \Theta\}$ *is a parametric family of strictly increasing functions and* $G(y,\beta)$ *is supposed to be three times continuously differentiable with respect to* β.

(b) ξ *is an H-regular function with asymptotic order* $\kappa_\xi(x,\beta)$ *and limiting homogeneous function* $h_\xi(x,\beta)$*, Moreover, for* $f = h_\xi, h'_\xi, P_\xi$*, we have* $\sup_{\beta\in\Theta}|f(h_g(x),\beta)| \in \mathcal{F}^0_{LB}$*, where* h_g *is defined in Assumption 3 (b) and* h_ξ, h'_ξ, P_ξ *are defined in Definition 3.3.*

(c) Let $\varpi_n = \min\{\sqrt{n}, \kappa_{ng}\}$*, define a neighborhood of* β_0 *by* $N(\varepsilon,\lambda) = \{\beta : |\kappa_\xi(\lambda,\beta_0)(\beta - \beta_0)| \le \varpi_n^{-1+\varepsilon}\}$ *for given* $\varepsilon > 0$*. For any given* $\bar{s} > 0$*, there exists* $\varepsilon > 0$ *such that as* $\lambda \to \infty$,

$$\kappa_\xi(\lambda,\beta_0)^{-2} \sup_{|s|\le\bar{s}} |\dot{\xi}(\lambda s,\beta_0)| \to 0, \tag{3.1}$$

$$\varpi_n^{-1+\varepsilon}\kappa_\xi(\lambda,\beta_0)^{-1} \sup_{|s|\le\bar{s}} \sup_{\beta\in N(\varepsilon,\lambda)} |\xi(\lambda s,\beta)| \to 0, \tag{3.2}$$

$$\varpi_n^{-1+\varepsilon}\kappa_\xi(\lambda,\beta_0)^{-2} \sup_{|s|\le\bar{s}} \sup_{\beta\in N(\varepsilon,\lambda)} |\dot{\xi}(\lambda s,\beta)| \to 0, \tag{3.3}$$

$$\varpi_n^{-1+\varepsilon}\kappa_\xi(\lambda,\beta_0)^{-3} \sup_{|s|\le\bar{s}} \sup_{\beta\in N(\varepsilon,\lambda)} |\ddot{\xi}(\lambda s,\beta)| \to 0, \tag{3.4}$$

where $\kappa_{ng} = \kappa_g(\sqrt{n})$*, and* $\kappa_g(\cdot)$ *is defined in Assumption 3 (b).*

(d) $n/(k\kappa_{ng}^2\kappa_{n\xi,\beta_0}^2) \to 0$.

Remark 3.5 *The strictly increasing property of* $G(y,\beta)$ *in Assumption 4 (a) is commonly imposed for identification. Assumption 4 (b) further stipulates that the composite function* ξ *is H-regular (see Definition 3.3). Note that* κ_ξ, h_ξ, *in Assumption 4 (b) may depend on* β_0*. Assumption 4 (c) is similar to Assumption (b) of Theorem 5.3 in Park and Phillips*

(2001), and is required to prove a uniform convergence result. It holds for many H-regular functions used in nonlinear analysis. Assumption 4 (d) is imposed to remove the estimation effect of $\widehat{\beta}_n$ on $\widehat{g}$, and is easily satisfied for g being H-regular functions and G being the Box-Cox transformation (with appropriate choice of k).

Example 3.1 *Consider the power transformation $G(y,\beta_0) = y^\beta$ $(y>0, \beta>0)$. It is obvious that Assumption 4 (a) holds. In this case $\xi(y,\beta) = \frac{1}{\beta_0} y^{\beta/\beta_0}\ln(y)$. To see that Assumption 4 (b) is satisfied, write*

$$\xi(\lambda y,\beta) = \lambda^{\beta/\beta_0}\ln(\lambda)\cdot\frac{1}{\beta_0}y^{\beta/\beta_0} + \frac{1}{\beta_0}(\lambda y)^{\beta/\beta_0}\ln(y).$$

Define $\kappa_\xi(\lambda,\beta) = \lambda^{\beta/\beta_0}\ln(\lambda)$ (depending on β_0), $h_\xi(y,\beta) = y^{\beta/\beta_0}/\beta_0$ and $R(y,\lambda,\beta) = (\lambda y)^{\beta/\beta_0}\ln(y)/\beta_0$. Obviously, h_ξ is regular on Θ and

$$|R(y,\lambda,\beta)| = \left|\frac{1}{\beta_0}(\lambda y)^{\beta/\beta_0}\ln(y)\right| \le \lambda^{\beta/\beta_0}\cdot\frac{1}{\beta_0}\left|y^{\beta/\beta_0}\ln(y)\right|,$$
$$|R_1(y,\lambda,\beta)| = \left|\frac{1}{\beta_0}\lambda^{\beta/\beta_0}y^{\beta/\beta_0-1}\left[\frac{\beta}{\beta_0}\ln(y)+1\right]\right| \le \lambda^{\beta/\beta_0}\left|\frac{1}{\beta_0}y^{\beta/\beta_0-1}\left[\frac{\beta}{\beta_0}\ln(y)+1\right]\right|.$$

Let $a(\lambda,\beta) = a_1(\lambda,\beta) = \lambda^{\beta/\beta_0}$, then

$$\limsup_{\lambda\to\infty}\sup_{\beta\in\Theta}|\kappa_\xi^{-1}(\lambda,\beta)a(\lambda,\beta)| = \limsup_{\lambda\to\infty}\sup_{\beta\in\Theta}\left|\frac{1}{\ln(\lambda)}\right| = 0,$$

showing that R and R_1 satisfy Definition 3.3 (b.i). Let $P(x,\beta) = y^{\beta/\beta_0}|\ln(y)|/\beta_0$ and $P_1(x,\beta) = y^{\beta/\beta_0}\left|\frac{\beta}{\beta_0}\ln(y)+1\right|/\beta_0$. It is apparent that $\sup_{\beta\in\Theta}P(x,\beta), \sup_{\beta\in\Theta}P_1(x,\beta) \in \mathcal{F}^0_{LB}$. Thus, $\xi(y,\beta)$ is H-regular with asymptotic order $\kappa_\xi(\lambda,\beta) = \lambda^{\beta/\beta_0}\ln(\lambda)$ and limit homogeneous function $h_\xi(y,\beta) = y^{\beta/\beta_0}/\beta_0$. Hence, Assumption 4 (b) is satisfied.

Moreover, by simple calculation, we have $\dot{\xi}(y,\beta) = y^{\beta/\beta_0}\ln^2(y)/\beta_0^2$, and $\ddot{\xi}(y,\beta) = y^{\beta/\beta_0}\ln^3(y)/\beta_0^3$. It is easy to show

$$\kappa_\xi(\lambda,\beta_0)^{-2}\sup_{|s|\le\bar{s}}|\dot{\xi}(\lambda s,\beta_0)| = \frac{1}{\beta_0^2\lambda^2\ln^2(\lambda)}\sup_{|s|\le\bar{s}}|\lambda s\ln^2(\lambda s)| \to 0,$$

as $\lambda\to\infty$. This establishes (3.1). Similarly, one can verify that (3.2)-(3.4) of Assumption 4(c) hold.

Example 3.2 *Consider the Box-Cox transformation* $G(y,\beta) = (y^\beta - 1)/\beta$ $(y > 0, \beta > 0)$. *It follows from simple calculations that*

$$\xi(y,\beta) = (\beta\beta_0)^{-1}(\beta_0 y + 1)^{\beta/\beta_0} \ln(\beta_0 y + 1) - \beta^{-2}(\beta_0 y + 1)^{\beta/\beta_0} + \beta^{-2},$$

$$\dot{\xi}(y,\beta) = (\beta_0 y + 1)^{\beta/\beta_0}\Big[\frac{1}{\beta\beta_0^2}\ln^2(\beta_0 y + 1) - \frac{2}{\beta^2\beta_0}\ln(\beta_0 y + 1) + \frac{2}{\beta^3}\Big] - \frac{2}{\beta^3},$$

$$\ddot{\xi}(y,\beta) = (\beta_0 y + 1)^{\beta/\beta_0}\Big[\frac{1}{\beta\beta_0^3}\ln^3(\beta_0 y + 1) - \frac{3}{\beta^2\beta_0^2}\ln^2(\beta_0 y + 1) + \frac{6}{\beta^3\beta_0}\ln(\beta_0 y + 1) - \frac{6}{\beta^4}\Big] + \frac{6}{\beta^4}.$$

We can show that ξ *is H-regular with asymptotic order* $\kappa_\xi(\lambda,\beta) = \lambda^{\beta/\beta_0}\ln(\lambda)$ *and homogeneous function* $h_\xi(y,\beta) = \beta^{-1}\beta_0^{-1}(\beta_0 y)^{\beta/\beta_0}$. *In addition, Assumption 4(c) can be easily verified as well. More details are provided in the online supplementary document.*

3.3 Distribution theory

The following theorem presents the asymptotic distribution of $\widehat{\beta}_n$, which depends on the asymptotic order (κ_g) of the regression function g and that (κ_ξ) of the transformation function G.

Theorem 3.1 *Let Assumptions 1-4 hold. Assume that for all* $\delta > 0$,

$$\int_{|s|<\delta} h_\xi^2(h_g(s),\beta_0)ds > 0 \text{ and } \int_{|s|<\delta} h_g^2(s,\beta_0)ds > 0. \tag{3.5}$$

Then the following assertions hold as $n \to \infty$.

(a) If $\sqrt{n}/\kappa_{ng} \to 0$,

$$\sqrt{n}\kappa_{n\xi,\beta_0}(\widehat{\beta}_n - \beta_0) \Rightarrow -\bigg(\int_0^1 h_\xi^2\Big[h_g\Big(V(r)\Big),\beta_0\Big]dr\bigg)^{-1}\int_0^1 h_\xi\Big[h_g\Big(V(r)\Big),\beta_0\Big]dU(r).$$

(b) If $\sqrt{n}/\kappa_{ng} \to \alpha \in \mathbb{R}_+$,

$$\begin{aligned}
&\sqrt{n}\kappa_{n\xi,\beta_0}(\widehat{\beta}_n - \beta_0)\\
&\Rightarrow -\bigg(\int_0^1 h_\xi^2\Big[h_g\Big(V(r)\Big),\beta_0\Big]dr\bigg)^{-1}\bigg\{\int_0^1 h_\xi\big[h_g\big(V(r)\big),\beta_0\big]dU(r)\\
&+\alpha\sigma_u^2\int_0^1 h_\xi'\big[h_g\big(V(r)\big),\beta_0\big]dr\bigg\}\\
&+\alpha\sigma_u^2\bigg\{\int_0^1 h_\xi^2\Big[h_g\Big(V(r)\Big),\beta_0\Big]dr\int_0^1 h_g^2(V(r))dr\bigg\}^{-1}\int_0^1 h_g\Big(V(r)\Big)h_\xi\Big[h_g\Big(V(r)\Big),\beta_0\Big]dr.
\end{aligned}$$

(c) If $\kappa_{ng}/\sqrt{n} \to 0$ *and* $\kappa_{ng} \to \infty$, *as* $n \to \infty$,

$$\begin{aligned}
&\kappa_{ng}\kappa_{n\xi,\beta_0}(\widehat{\beta}_n - \beta_0) \Rightarrow \\
&-\sigma_u^2\Big(\int_0^1 h_\xi^2\Big[h_g\Big(V(r)\Big),\beta_0\Big]dr\Big)^{-1}\int_0^1 h_\xi'\big[h_g\big(V(r)\big),\beta_0\big]dr \\
&+\sigma_u^2\Big\{\int_0^1 h_\xi^2\Big[h_g\Big(V(r)\Big),\beta_0\Big]dr\int_0^1 h_g^2\big(V(r)\big)dr\Big\}^{-1}\int_0^1 h_g\Big(V(r)\Big)h_\xi\Big[h_g\Big(V(r)\Big),\beta_0\Big]dr.
\end{aligned}$$

In the above expressions, $h_\xi'(x,\beta) = \partial h_\xi(x,\beta)/\partial x$, *and* $U(r), V(r)$ *are the Brownian motions defined in Assumption 1* (d).

The identifiability condition in (3.5) is similar to that of Park and Phillips (2001, Theorem 5.2) and that of Uematsu (2017, Theorem 3.1). Note that the result in (b) degenerates to that in (a) when $\alpha = 0$. The limiting distributions in all three cases are nonstandard. When U and V are independent, the limiting distribution in (a) is mixed normal. In general, all the limiting results are not mixed normal and $\widehat{\beta}_n$ is likely to have an asymptotic bias. However, how to construct the bias-corrected estimator in this setup remains a challenging issue, and is beyond the scope of this study. See Park and Phillips (2001) and Chan and Wang (2015) for similar issues.

We now illustrate the above asymptotic results with an example.

Example 3.3 *For the functions discussed in previous examples, the asymptotic results presented in Theorem 3.1 can be easily simplified. Consider the Box-Cox transformation function* G *given in Example 3.2 and the link function given by* $g(x) = ax^b + d$ $(b > 0)$. *It is straightforward to see that* g *is H-regular with asymptotic order* $\kappa_g(\lambda) = \lambda^b$ *and homogeneous function* $h_g(x) = ax^b$. *By Example 3.2, we have* $\kappa_\xi(\lambda,\beta_0) = \lambda\ln(\lambda)$ *and* $h_\xi(y,\beta_0) = \beta_0^{-1}y$.

(a). In the case that $b = 2$, $\kappa_{ng} = n$ *and* $\sqrt{n}/\kappa_{ng} = 1/\sqrt{n} \to 0$. *Meanwhile,* $\kappa_{n\xi,\beta_0} = n\ln(n)$ *and* $h_\xi(h_g(x),\beta_0) = \beta_0^{-1}ax^2$. *Then Theorem 3.1* (a) *reduces to*

$$n^{3/2}\log(n)(\widehat{\beta}_n/\beta_0 - 1) \Rightarrow -\left(\int_0^1 V^4(r)dr\right)^{-1}\int_0^1 V^2(r)dU(r).$$

(b). When $b = 1$, we have $\kappa_{ng} = \sqrt{n}$, satisfying the condition in Theorem 3.1 (b) with $\alpha = 1$. And $\kappa_{n\xi,\beta_0} = \sqrt{n}\ln(\sqrt{n})$, $h_\xi(h_g(x), \beta_0) = ax/\beta_0$ and $h'_\xi(h_g(x), \beta_0) = 1/\beta_0$. Thus, by the limiting result given in Theorem 3.1 (b),

$$n\log(\sqrt{n})(\widehat{\beta}_n/\beta_0 - 1) \Rightarrow -\left(a\int_0^1 V^2(r)dr\right)^{-1}\int_0^1 V(r)dU(r).$$

(c). When $b = 1/2$, $\kappa_{ng} = \sqrt[4]{n}$ diverges slower than $\sqrt{n}$. Also, $\kappa_{n\xi,\beta_0} = \sqrt[4]{n}\ln(\sqrt[4]{n})$, $h_\xi(h_g(x), \beta_0) = a\sqrt{x}/\beta_0$ and $h'_\xi(h_g(x), \beta_0) = 1/\beta_0$. By the limiting result in Theorem 3.1 (c),

$$\sqrt{n}\log(\sqrt[4]{n})(\widehat{\beta}_n/\beta_0 - 1) \xrightarrow{p} 0.$$

The theorem bellow presents the asymptotic property of the plug-in estimator $\widehat{g}(x) = Z_k^T(x)\widehat{c}$ defined in Section 2.

Theorem 3.2 *Let the conditions in Theorem 3.1 hold. As $n \to \infty$,*

$$\sigma_u^{-1}\Delta_z(x)^{-1/2}[\widehat{g}(x) - g(x)] \Rightarrow N(0, 1),$$

where $\Delta_z(x) = Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x)$, $C_k = diag(n, n^2, \cdots, n^k)$ and $D_k = \left(\int_0^1 \frac{1}{\sqrt{(i-1)!(j-1)!}}V^{i+j-2}(r)dr\right)_{1\le i,j\le k}$.

Remark 3.6 *(i) The order involved in the normality is $O_p(\sqrt{n/k})$, due to the fact that*

$$\begin{aligned}\Delta_z(x) &\le \lambda_{\min}^{-1}(D_k)\lambda_{\max}(C_k^{-1}) \cdot Z_k^T(x)Z_k(x)\\ &= O_p(1) \cdot O_p(n^{-1}) \cdot O_p(k) = O(k/n),\end{aligned}$$

in view of $\|Z_k(x)\|^2 = O_p(k)$. (ii) It can be shown that a consistent estimator for the error variance σ_u^2 is $\widehat{\sigma}_u^2 = \frac{1}{n}\sum_{t=1}^n [G(y_t, \widehat{\beta}_n) - Z_k^T\widetilde{c}(\widehat{\beta}_n)]^2$, and that $\widehat{D}_k - D_k \xrightarrow{p} 0$, where $\widehat{D}_k = C_k^{-1/2}Z^TZC_k^{-1/2}$ (Lemma A.1). As a result, the point-wise confidence interval of g can be easily constructed based on the above limiting distribution.

4 Numerical Results

4.1 Simulations

We conduct simulation studies to investigate the finite sample performance of the proposed estimator for the transformation index β_0 and that for the link function g. Let x_t be

generated according to

$$x_t = x_{t-1} + v_t,$$

for $t = 1, \cdots, n$, $x_0 = 0$, and $v_t \sim N(0, 0.8^2)$. The regression error u_t is independent of v_t and follows

$$u_t = \varepsilon_t + \rho_1 \varepsilon_{t-1},$$

in which $\rho_1 = 0, 0.2, 0.5, 0.8$, and $\varepsilon_t \sim N(0, 0.5^2)$. We consider the Box-Cox transformation $G(y, \beta_0) = (y^{\beta_0} - 1)/\beta_0$ for $\beta_0 = 0.5, 1, 1.5$, and $G(y, \beta_0) = \ln y$ for $\beta_0 = 0$. Three types of link functions in line with Example 3.3 are entertained:

(1) $g_1(x) = x^2 + 2$;

(2) $g_2(x) = a_2 x + 2$, with $a_2 = 1, 10$;

(3) $g_3(x) = a_3\sqrt{x} + 2$, with $a_3 = 1, 10$.

To save space, some selected results for sample size $n = 100, 200, 400, 800, 1200$ are reported with $M = 1000$ replications. Other results are similar and are available upon request.

The estimators presented in Section 2 depend on the sieve order k that is used to approximate the unknown link function. In practice, we choose the value k which minimizes the penalized in-sample mean squared forecast error:

$$\widehat{k} = \arg\min_{k \in K_n} \Big(MSE(k) + \frac{2k}{n} \Big), \tag{4.1}$$

where $MSE(k) = \frac{1}{n}\sum_{t=1}^{n} (\widehat{y}_t(k) - y_t)^2$, $\widehat{y}_t(k) = \frac{1}{n}\sum_{i=1}^{n} G^{-1}(\widehat{g}_k(x_t) + \widehat{e}_i(k), \widehat{\beta}_n(k))$ is the smearing estimate (Duan, 1983) that is to remove the prediction bias, with $\widehat{e}_t(k) = G(y_t, \widehat{\beta}_n(k)) - \widehat{g}_k(x_t)$, $\widehat{g}_k(x_t) = Z_k^T(x_t)\widehat{c}_k$, and $K_n = \{2, 3, 4, 5\}$. Note that the estimators $\widehat{\beta}_n(k)$ and $\widehat{c}_k$ are as defined in Section 2 with the sieve order being k. The optimal $\widehat{k}$ selected are 3, 2 and 5 for the above specifications of g, respectively. Consequently, we shall report the results for these corresponding $\widehat{k}$'s only, to save space. We shall also suppress the dependence of the notations on $\widehat{k}$ for brevity.

To evaluate estimation accuracy of the estimator $\widehat{\beta}_n$, we calculate the bias, standard deviation (SD) and root mean squared error (RMSE):

$$Bias = \overline{\beta}_n - \beta_0, \quad SD = \Big(\frac{1}{M} \sum\nolimits_{l=1}^{M} (\widehat{\beta}_{n,l} - \overline{\beta}_n)^2 \Big)^{1/2}, \quad RMSE = \sqrt{Bias^2 + SD^2},$$

where $\overline{\beta}_n = \frac{1}{M}\sum_{l=1}^{M}\widehat{\beta}_{n,l}$, with $\widehat{\beta}_{n,l}$ standing for the estimate of β_0 in the l-th replication. The above measures for $\widehat{\beta}_n$ (multiplied by 10^3) are reported in Table 1.

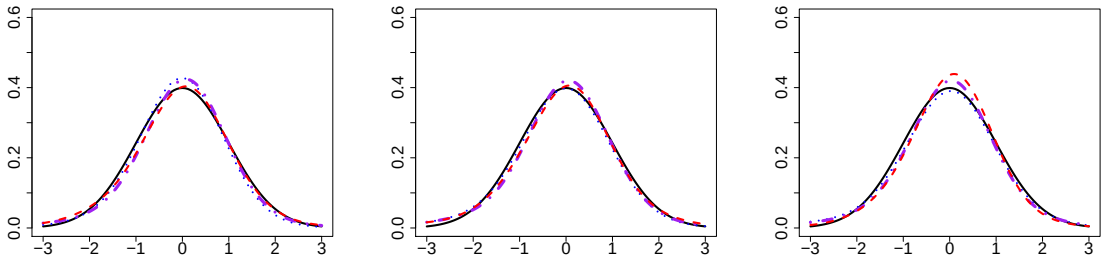
It is observed from Table 1 that $\widehat{\beta}_n$ performs well for all the specifications of g and choices of β_0 and ρ_1. Even for a sample of size $n = 100$, $\widehat{\beta}_n$ has small bias and RMSE. In addition, all three measures for $\widehat{\beta}_n$ decrease fast as sample size n increases. Noticeably, the rate that the variance and/or RMSE of $\widehat{\beta}_n$ converges depends on the specification of the link function g, but is not affected by the dependence parameter ρ_1 in u_t. In particular, we see that the RMSE of $\widehat{\beta}_n$ diminishes much faster for g_1 than that for g_2, and that of the latter much faster than that for g_3. This is consistent with the asymptotic theory derived in Example 3.3.

We next evaluate the finite sample distribution of $\widehat{\beta}_n$ under the three choices of g. As the limiting distributions given in Theorem 3.1 or Example 3.3 are nonstandard, we shall compare the estimated density of normalized $\widehat{\beta}_n$ with the corresponding limiting distribution. First, consider the case with $g_1(x) = x^2 + 2$. Example 3.3 shows that $\beta_1 \equiv n^{3/2}\ln(n)(\widehat{\beta}_n/\beta_0 - 1)$ is asymptotically mixed normal, which leads to the t-ratio $\widetilde{\beta}_1 \equiv \beta_1/sd(\beta_1)$ being standard normal. Second, for the linear function $g_2(x) = 10x + 2$, we obtain from Example 3.3 that $\beta_2 \equiv n\ln(\sqrt{n})(\widehat{\beta}_n/\beta_0 - 1)$ is also asymptotically mixed normal, indicating that the t-ratio $\widetilde{\beta}_2 = \beta_2/sd(\beta_2)$ is standard normal. Therefore, an intuitive way is to compare the kernel density of $\widetilde{\beta}_1$ and that of $\widetilde{\beta}_2$ with the standard normal density. The kernel density estimates of $\widetilde{\beta}_1$, and those of $\widetilde{\beta}_2$ are plotted in Figure 4.1 and Figure 4.2, respectively, with the second-order Gaussian kernel and bandwidth $h = 2n^{-1/5}$, along with the standard normal density. Comparing the curves in Figures 4.1-4.2 reveals that the asymptotic distributions provide good approximations in finite samples. This finding is robust with respect to different choices of β_0, ρ_1, specifications of g and sample sizes n. As for $g_3(x) = 10\sqrt{x}+2$, Example 3.3 shows that $\beta_3 \equiv \sqrt{n}\ln(n^{1/4})(\widehat{\beta}_n/\beta_0-1) \xrightarrow{p} 0$. The mean squared errors of β_3 are reported is Table 2. This shows that β_3 is converging to zero in probability, although the rate of convergence seems slow in finite samples.

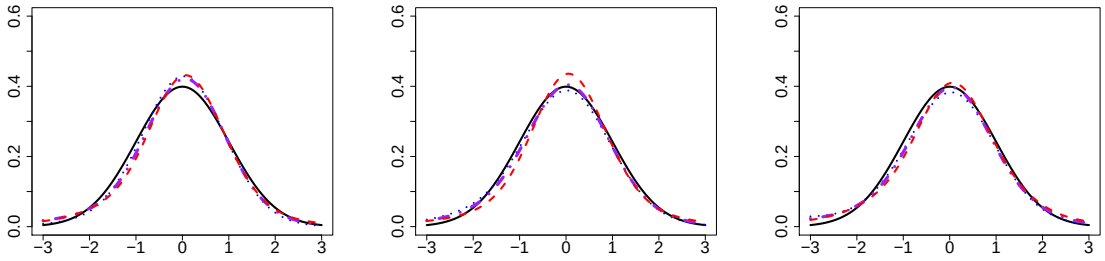
We then turn to evaluate the performance of the link function estimator, with the results reported in Figures 4.3-4.5 for sample size $n = 200$. Figure 4.3 shows the average of the estimates of $g_1(x) = x^2 + 2$ for $\beta_0 = 0, 1, 1.5$ and $\rho_1 = 0, 0.5, 0.8$, together with the 95% point-wise simulated confidence band and the 95% point-wise asymptotic confidence band. Figure 4.4 and 4.5 report those for $g_2(x) = x + 2$ and $g_3(x) = \sqrt{x} + 2$, respectively. Figures 4.3-4.5 indicate that the estimator performs very well, with very small bias and

Table 1: Bias, SD and RMSE ($\times 10^3$) for $\widehat{\beta}_n$.

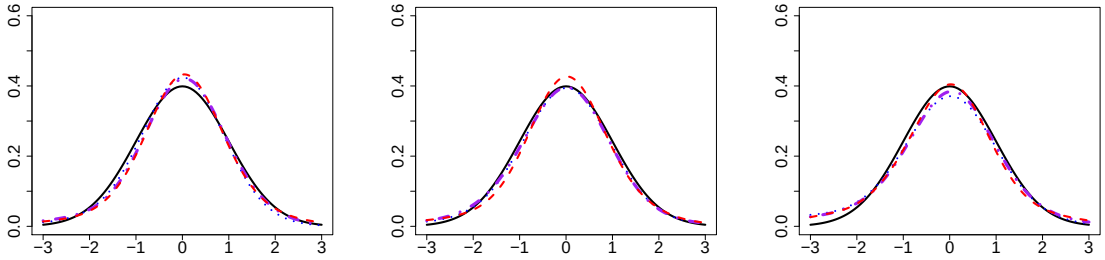
	β_0	0			1			1.5		
	ρ_1	0	0.5	0.8	0	0.5	0.8	0	0.5	0.8
n		$g(x) = x^2 + 2$								
	Bias	0.21	0.25	0.27	-0.33	0.24	-0.34	-0.29	-7.49	-12.0
100	SD	0.45	0.53	0.51	7.08	7.38	7.62	7.48	51.8	59.1
	RMSE	0.49	0.58	0.58	7.09	7.39	7.63	7.49	52.3	60.3
	Bias	0.09	0.11	0.12	0.16	0.09	0.03	0.03	-1.33	-0.77
200	SD	0.07	0.13	0.14	2.20	2.31	2.33	2.32	17.3	20.2
	RMSE	0.11	0.17	0.19	2.20	2.31	2.33	2.32	17.4	20.2
	Bias	0.06	0.07	0.08	0.00	0.01	0.06	0.04	-0.09	0.29
400	SD	0.02	0.04	0.04	0.69	0.71	0.72	0.71	6.28	6.99
	RMSE	0.06	0.08	0.09	0.69	0.71	0.72	0.71	6.28	7.00
		$g(x) = 10x + 2$								
	Bias	0.08	0.10	0.11	-1.79	-3.02	-3.74	-1.93	-5.45	-6.56
100	SD	0.07	0.10	0.13	5.25	10.4	12.3	6.06	15.3	17.4
	RMSE	0.11	0.14	0.17	5.55	10.8	12.8	6.36	16.2	18.6
	Bias	0.06	0.07	0.07	-0.80	-1.88	-1.95	-0.99	-2.44	-3.64
200	SD	0.02	0.03	0.05	2.31	5.53	6.35	2.71	8.26	9.87
	RMSE	0.07	0.08	0.09	2.45	5.84	6.64	2.88	8.61	10.5
	Bias	0.07	0.06	0.06	-0.37	-0.91	-1.23	-0.41	-1.35	-1.96
400	SD	0.01	0.02	0.02	1.08	2.58	3.44	1.24	3.95	4.90
	RMSE	0.07	0.06	0.06	1.14	2.74	3.66	1.30	4.17	5.28
		$g(x) = 10\sqrt{x} + 2$								
	Bias	1.08	0.64	0.22	-4.02	-5.44	-6.78	-3.91	-5.68	-6.54
100	SD	3.06	2.25	1.24	7.58	6.64	5.32	7.65	6.49	5.61
	RMSE	3.24	2.34	1.26	8.58	8.58	8.62	8.59	8.63	8.61
	Bias	0.65	0.35	0.15	-2.13	-3.62	-4.30	-2.68	-3.65	-4.60
200	SD	1.99	1.39	0.87	4.82	3.84	3.07	4.54	3.83	2.58
	RMSE	2.09	1.44	0.88	5.27	5.28	5.28	5.27	5.29	5.28
	Bias	0.31	0.16	0.10	-1.93	-2.41	-2.91	-1.86	-2.60	-2.91
400	SD	1.03	0.72	0.44	2.66	2.22	1.53	2.71	2.01	1.53
	RMSE	1.07	0.74	0.45	3.28	3.28	3.29	3.28	3.29	3.29



(a) $\beta_0 = 0.5$, $\rho_1 = 0$ (b) $\beta_0 = 1$, $\rho_1 = 0$ (c) $\beta_0 = 1.5$, $\rho_1 = 0$

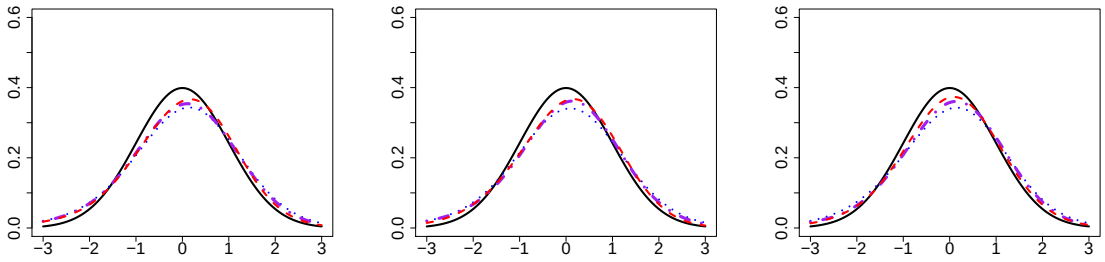


(d) $\beta_0 = 0.5$, $\rho_1 = 0.5$ (e) $\beta_0 = 1$, $\rho_1 = 0.5$ (f) $\beta_0 = 1.5$, $\rho_1 = 0.5$

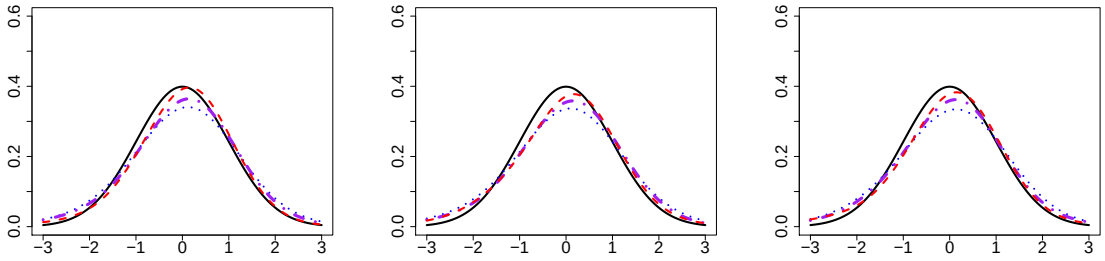


(g) $\beta_0 = 0.5$, $\rho_1 = 0.8$ (h) $\beta_0 = 1$, $\rho_1 = 0.8$ (i) $\beta_0 = 1.5$, $\rho_1 = 0.8$

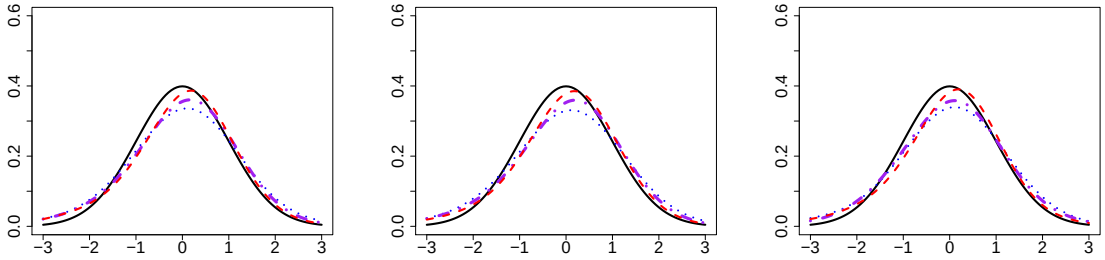
Figure 4.1: Plot of standard normal density (black solid line) and kernel density of $\widetilde{\beta}_1$ (the normalized $\widehat{\beta}_n$) for $n = 100$ (blue dotted), $n = 200$ (purple dot-dashed), $n = 400$ (red dashed) with $g_1(x) = x^2 + 2$.



(a) $\beta_0 = 0.5$, $\rho_1 = 0$ (b) $\beta_0 = 1$, $\rho_1 = 0$ (c) $\beta_0 = 1.5$, $\rho_1 = 0$



(d) $\beta_0 = 0.5$, $\rho_1 = 0.5$ (e) $\beta_0 = 1$, $\rho_1 = 0.5$ (f) $\beta_0 = 1.5$, $\rho_1 = 0.5$



(g) $\beta_0 = 0.5$, $\rho_1 = 0.8$ (h) $\beta_0 = 1$, $\rho_1 = 0.8$ (i) $\beta_0 = 1.5$, $\rho_1 = 0.8$

Figure 4.2: Plot of standard normal density (black solid line) and kernel density of $\widetilde{\beta}_2$ (the normalized $\widehat{\beta}_n$) for $n = 100$ (blue dotted), $n = 200$ (purple dot-dashed), $n = 400$ (red dashed) with $g_2(x) = 10x + 2$.

Table 2: Mean squared errors (MSEs) ($\times 10^3$) for β_3 (the normalized $\widehat{\beta}_n$)

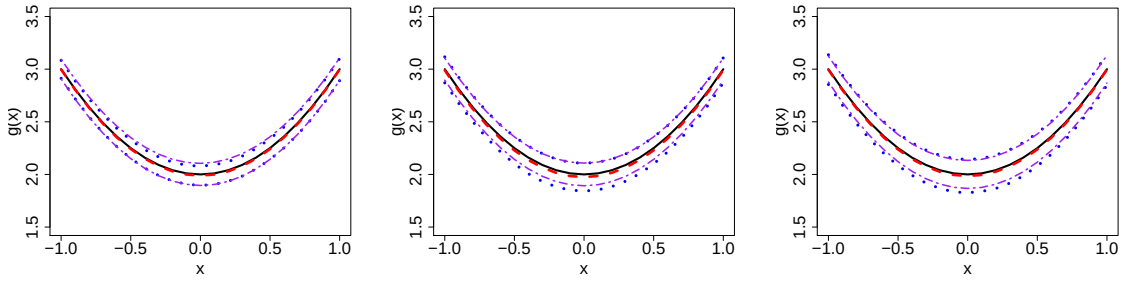
β_0	ρ_1	$n = 100$	200	400	800	1200
	0	18.70	17.90	16.58	14.40	10.56
0.5	0.5	18.69	17.92	16.60	14.40	10.56
	0.8	18.71	17.92	16.60	14.40	10.56
	0	9.348	8.957	8.300	7.201	5.279
1	0.5	9.342	8.959	8.294	7.203	5.279
	0.8	9.350	8.949	8.301	7.203	5.279
	0	6.232	5.971	5.534	4.802	3.519
1.5	0.5	6.236	5.973	5.534	4.802	3.519
	0.8	6.232	5.973	5.534	4.802	3.519

variance for the most part of the domain of x, despite the fact that the estimator is based on merely $n = 200$ observations. We also plot these curves for $n = 100, 200, 400$ with $\beta_0 = 1$ and $\rho_1 = 0$ in Figure 4.6, which reveals that the confidence bands are shrinking as n increases. The simulated and asymptotic confidence bands almost overlap, indicating that the asymptotic distribution given in Theorem 3.2 provides a good approximation in finite samples.

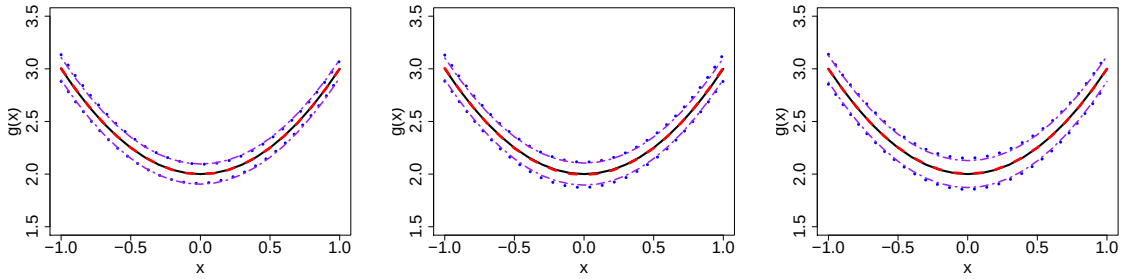
4.2 An empirical study

We now provide a real data example to illustrate the practical merit of our proposed model. In particular, we consider the environmental Kuznets curve for $PM_{2.5}$ observations in one city and two provinces in northen China, i.e., Beijing (BJ), Hebei (HB) and Shandong (SD).

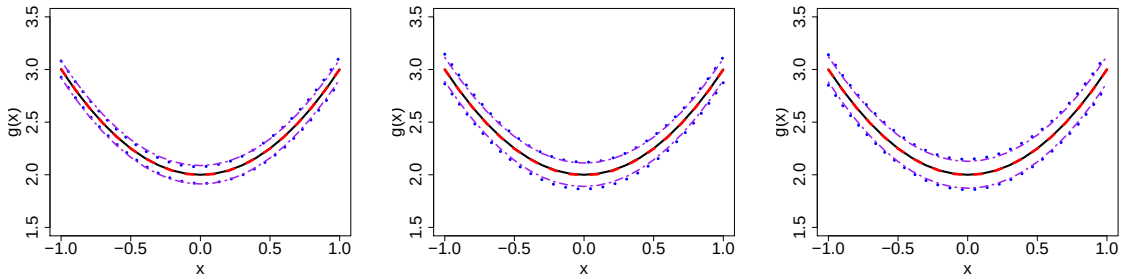
The environmental Kuznets curve. The concept of environmental Kuznets curve (EKC) emerged in the 1990's, when Grossman and Krueger (1991) revealed that the air pollution measures increased with income at first but decreased once per-capita GDP passes a certain threshold, hence hypothesized an inverted-U shaped relationship between indicators of environmental degradation and income per capita. It was named after Kuznets (1955), who first hypothesized a similar relationship between income inequality and economic development. Together with a period of extraordinary economic



(a) $\beta_0 = 0$, $\rho_1 = 0$ (b) $\beta_0 = 0$, $\rho_1 = 0.5$ (c) $\beta_0 = 0$, $\rho_1 = 0.8$

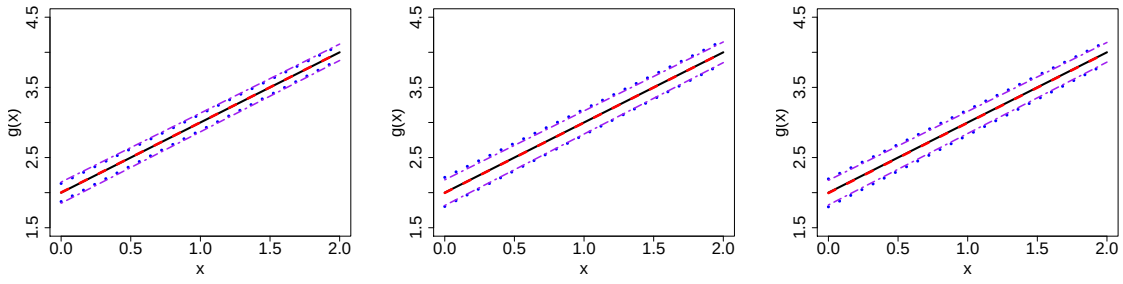


(d) $\beta_0 = 1$, $\rho_1 = 0$ (e) $\beta_0 = 1$, $\rho_1 = 0.5$ (f) $\beta_0 = 1$, $\rho_1 = 0.8$

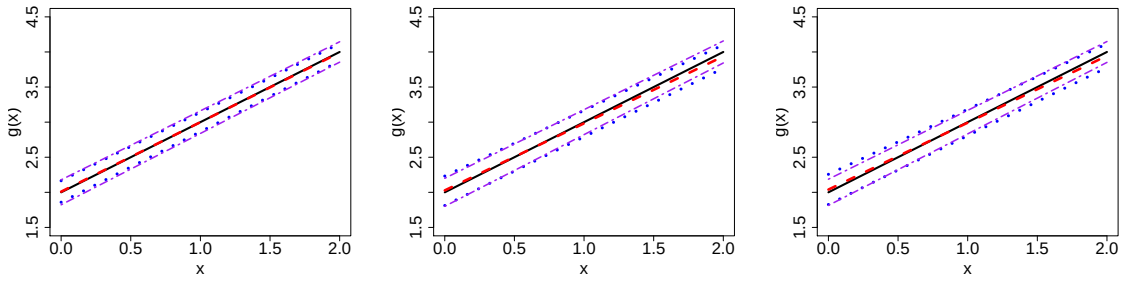


(g) $\beta_0 = 1.5$, $\rho_1 = 0$ (h) $\beta_0 = 1.5$, $\rho_1 = 0.5$ (i) $\beta_0 = 1.5$, $\rho_1 = 0.8$

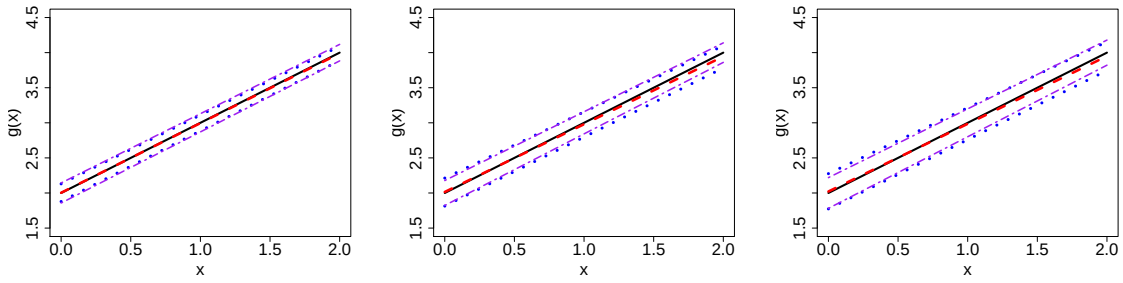
Figure 4.3: Plots of $g_1(x) = x^2 + 2$ (black solid line), the averaged estimate of $g_1(x)$ (red dashed), the simulated 95% confidence bands (blue dotted) and the asymptotic 95% confidence bands (purple dot-dashed) for $n = 200$.



(a) $\beta_0 = 0$, $\rho_1 = 0$ (b) $\beta_0 = 0$, $\rho_1 = 0.5$ (c) $\beta_0 = 0$, $\rho_1 = 0.8$

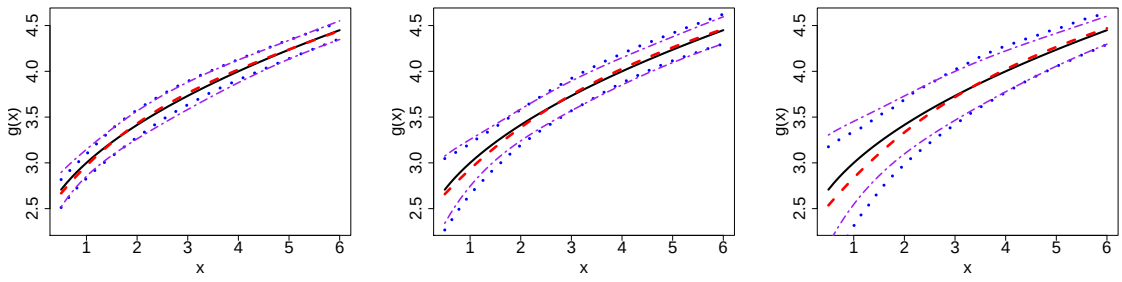


(d) $\beta_0 = 1$, $\rho_1 = 0$ (e) $\beta_0 = 1$, $\rho_1 = 0.5$ (f) $\beta_0 = 1$, $\rho_1 = 0.8$

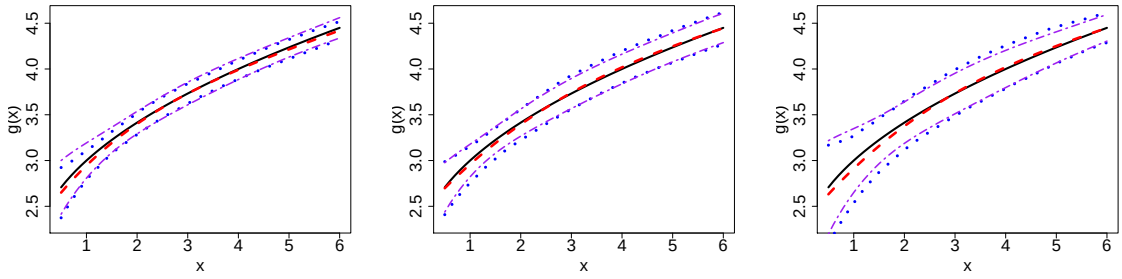


(g) $\beta_0 = 1.5$, $\rho_1 = 0$ (h) $\beta_0 = 1.5$, $\rho_1 = 0.5$ (i) $\beta_0 = 1.5$, $\rho_1 = 0.8$

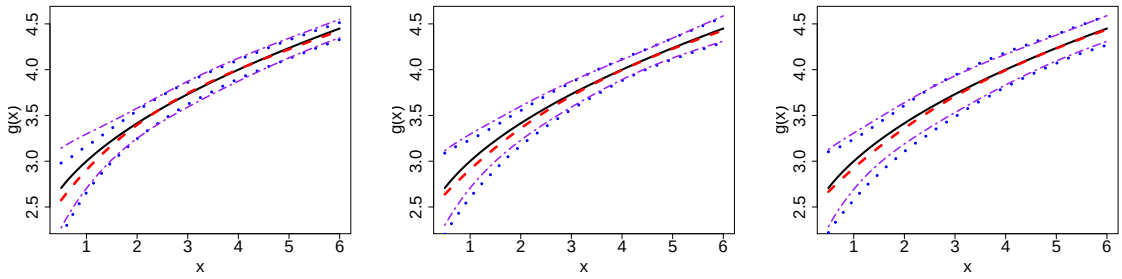
Figure 4.4: Plots of $g_2(x) = x + 2$ (black solid line), the averaged estimate of $g_2(x)$ (red dashed), the simulated 95% confidence bands (blue dotted) and the asymptotic 95% confidence bands (purple dot-dashed) for $n = 200$.



(a) $\beta_0 = 0$, $\rho_1 = 0$ (b) $\beta_0 = 0$, $\rho_1 = 0.5$ (c) $\beta_0 = 0$, $\rho_1 = 0.8$

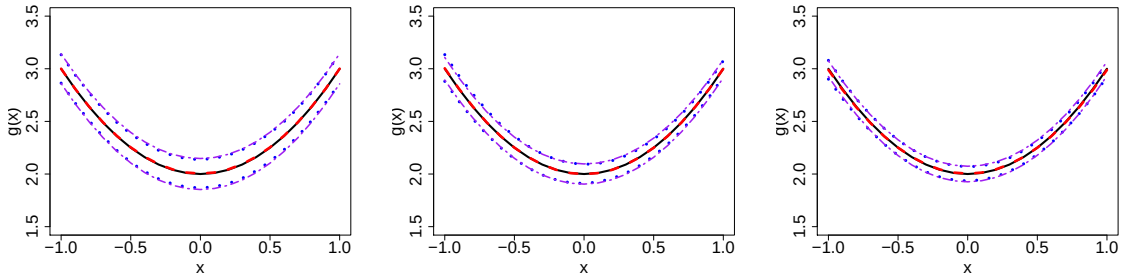


(d) $\beta_0 = 1$, $\rho_1 = 0$ (e) $\beta_0 = 1$, $\rho_1 = 0.5$ (f) $\beta_0 = 1$, $\rho_1 = 0.8$

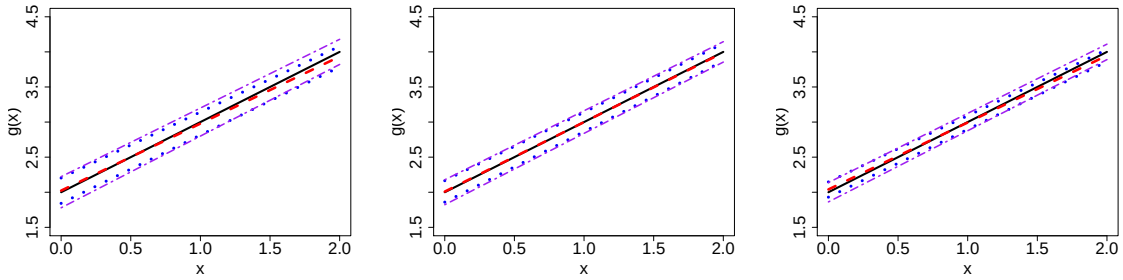


(g) $\beta_0 = 1.5$, $\rho_1 = 0$ (h) $\beta_0 = 1.5$, $\rho_1 = 0.5$ (i) $\beta_0 = 1.5$, $\rho_1 = 0.8$

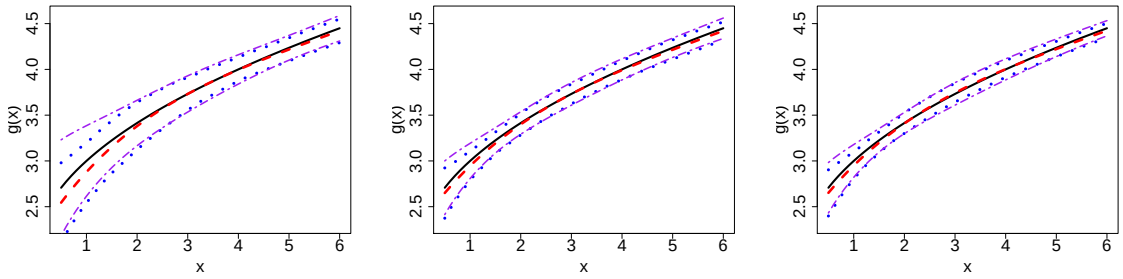
Figure 4.5: Plots of $g_3(x) = \sqrt{x} + 2$ (black solid line), the averaged estimate of $g_3(x)$ (red dashed), the simulated 95% confidence bands (blue dotted) and the asymptotic 95% confidence bands (purple dot-dashed) for $n = 200$.



(a) $g_1, n = 100$ (b) $g_1, n = 200$ (c) $g_1, n = 400$



(d) $g_2, n = 100$ (e) $g_2, n = 200$ (f) $g_2, n = 400$



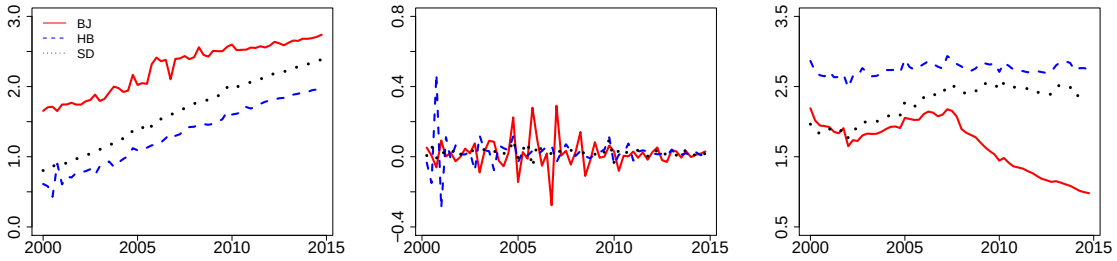
(g) $g_3, n = 100$ (h) $g_3, n = 200$ (i) $g_3, n = 400$

Figure 4.6: Plots of $g = g_1, g_2, g_3$ (black solid line), the averaged estimate of $g(x)$ (red dashed), the simulated 95% confidence bands (blue dotted) and the asymptotic 95% confidence bands (purple dot-dashed) for $\beta_0 = 1$ and $\rho_1 = 0$, with $g_1(x) = x^2 + 2, g_2(x) = x + 2, g_3(x) = \sqrt{x} + 2$.

growth during the past 40 years, the increase of pollution emission in China has led to serious environmental problems, as well as heated debate on how economic growth affects environmental quality and whether China's high growth is sustainable once more environment-friendly policies are implemented. Central to this debate, is the study of the EKC, which is to be analyzed using our proposed model. Here we focus on the EKC for particulate matter that has a diameter of less than 2.5 micrometers, i.e. $PM_{2.5}$, which is of serious concern in China nowadays.

The data. The $PM_{2.5}$ emission data are obtained from the research group at Department of Environmental Science and Laboratory for Earth Surface Processes, Peking University, or from the website http://inventory.pku.edu.cn/. The original $PM_{2.5}$ emission data are available at monthly frequency, with 0.1 by 0.1 degree spatial resolution measured by g/km^2. To obtain the quarterly emissions, we first add up the emissions of the three months in each quarter, and then multiply it by the area of the city/province (the sum of area of the grid squares falling into the city/province). The quarterly emissions are then divided by the population to obtain the per capita $PM_{2.5}$ emissions (y_t), which are measured by 10^2kg per capita. Following the literature, we use the natural logarithm of real GDP per capita (x_t) to measure economic development. The nominal GDP and population data are downloaded from the Wind database, originally published by the National Bureau of Statistics (NBS) of China. Real GDP is then calculated with the GDP deflator obtained from NBS. Divided by the population, we then obtain the real per capita GDP, measured by one thousand CNY of the year 2000. The quarterly data ranges from 2000Q1 to 2014Q4 with a total of $n = 60$ observations for each city, with the starting point determined by the availability of GDP and the ending point determined by the availability of the $PM_{2.5}$ emissions. All the series are seasonally adjusted using X-13-ARIMA developed by the United States Census Bureau.

The natural logarithm of real GDP per capita x_t and the per capita $PM_{2.5}$ emissions y_t are plotted in Figure 4.7 (a) and (c). It is observed that in the three cases x_t has stochastic trend and is likely to be a unit root process, while y_t looks more stable over time. To be precise, we apply the augmented Dickey-Fuller (ADF) unit root test to the two series, with the results reported in Table 3. The test fails to reject the null that x_t is unit root in all three cases. It rejects the same null for y_t in Beijing with a p-value 0.054, while fails to reject the null for y_t in Hebei and Shandong with p-values 0.589 and 0.917, respectively. The differenced series Δx_t, plotted in Figure 4.7 (b), looks stationary, which



(a) lnGDP. (b) Δ lnGDP. (c) $PM_{2.5}$.

Figure 4.7: The log of real GDP per capita and its difference, and $PM_{2.5}$ per capita for Beijing, Hebei and Shandong. Data range: 2000Q1-2014Q4.

is also supported by the ADF test.

Table 3: p-values for the Augmented Dickey-Fuller (ADF) Test

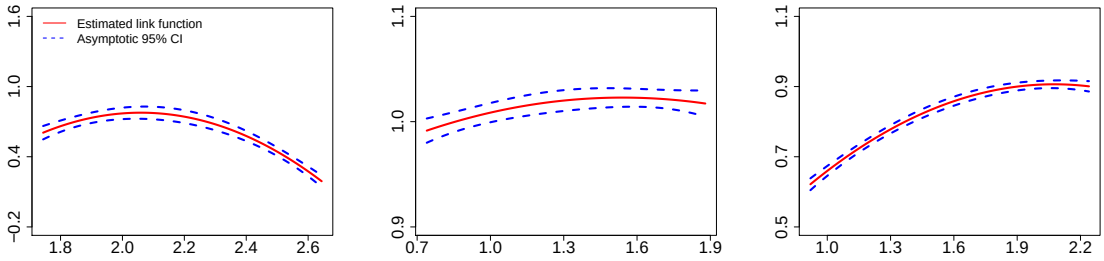
	BJ	HB	SD
lnGDP	0.985	0.990	0.990
$PM_{2.5}$	0.054	0.589	0.917
Δ lnGDP	0.010	0.010	0.010

The model and results. We first consider our proposed model, i.e.,

$$M1: \quad G(y_t, \beta) = g(x_t) + u_t,$$

where we use the Box-Cox transformation $G(y, \beta) = (y^\beta - 1)/\beta$, for $\beta > 0$, and $G(y, 0) = \ln y$, an H-regular function $g(\cdot)$ left unspecified, and u_t is the error term.

The truncation parameter k in the sieve approximation is selected according to (4.1), for $K_n = \{3, 4, 5\}$. It is found that $k = 3$ is selected for all three cases under this criterion. We shall report results for this optimal choice of k. For Beijing, we obtain $\widehat{\beta}_n = 0$ and $\widehat{c} = (-8.22, 7.07, -2.43)^T$; for Hebei, we obtain $\widehat{\beta}_n = 0$ and $\widehat{c} = (0.86, 0.15, -0.07)^T$; for Shandong, we obtain $\widehat{\beta}_n = 0$ and $\widehat{c} = (-0.23, 0.89, -0.30)^T$. The Box-Cox transformation parameters are estimated to be zero in all three cases, suggesting that a natural log transformation would be suitable for the data. That is, we shall consider $G(y, \widehat{\beta}_n) = \ln(y)$. Figure 4.8 plots the estimated curve for the link function $\widehat{g}(x)$, with the 95% point-wise



(a) Beijing. (b) Hebei. (c) Shandong.

Figure 4.8: The estimated link function $\widehat{g}(\cdot)$ and its asymptotic 95% confidence interval.

confidence interval based on the asymptotic distribution in Theorem 3.1 (a). Therefore, the estimated relationship between $PM_{2.5}$ and GDP per capita is inverted-U shaped, supporting the central hypothesis of EKC.

Model diagnostics. The residuals can be estimated as $\widehat{u}_t = \ln(y_t) - \widehat{g}(x_t)$. The augmented Dickey-Fuller test suggests that the estimated residuals in Beijing, Hebei and Shandong are stationary, with p-values 0.011, 0.010 and 0.010, respectively.

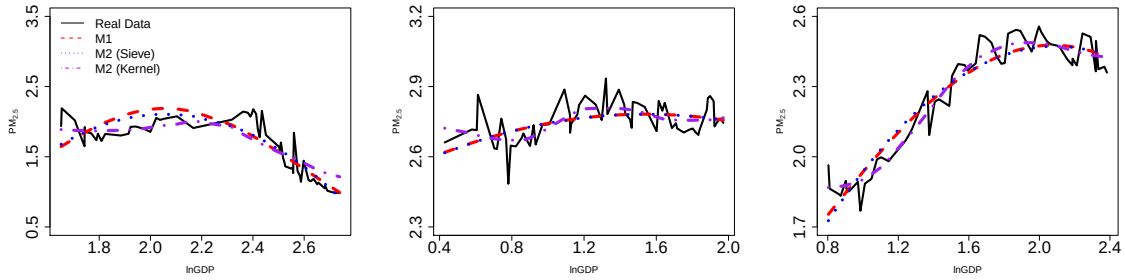
To shed further light on the serial dependence of the fitted residuals, we perform the portmanteau test of Ljung and Box (1978) for no serial correlation. For Beijing, the Ljung-Box test shows that $\widehat{u}_t$ is not white noise with p-value smaller than 0.001. Therefore, we consider an autoregressive moving average process for the residuals, with the order selected using the extended autocorrelation function (EACF) proposed by Tsay and Tiao (1984). This results an MA(6) model, which is

$$\widehat{u}_t = e_t + 0.71e_{t-1} + 0.36e_{t-2} + 0.34e_{t-3} + 0.44e_{t-4} + 0.4e_{t-5} + 0.007e_{t-6},$$

where e_t is white noise (p-value of Ljung-box test is 0.75) with $E(e_t^2) = 0.06$. For Hebei, the Ljung-Box test suggests that $\widehat{u}_t$ is white noise with p-value 0.068. For Shandong, the Ljung-Box test suggests that $\widehat{u}_t$ is not white noise with p-value 0.002. The EACF suggests that an MA(2) model is suitable, and we obtain,

$$\widehat{u}_t = e_t + 0.62e_{t-1} + 0.22e_{t-2},$$

where e_t is white noise (p-value of the Ljung-Box test is 0.28) with $E(e_t^2) = 0.0007$. The above results indicate that the estimated residuals in the three cases all can be



(a) Beijing. (b) Hebei. (c) Shandong.

Figure 4.9: The real data (black solid line) and fitted curves for M1 (red dashed), M2 with sieve estimator (blue dotted) and M2 with kernel estimator (purple dot-dashed).

well approximated by finite order MA processes. This finding is consistent with our Assumption 1 on the residuals.

Forecasting comparison. To evaluate the performance of the proposed model, we compare it with the nonparametric cointegration model considered by Wang and Phillips (2009b),

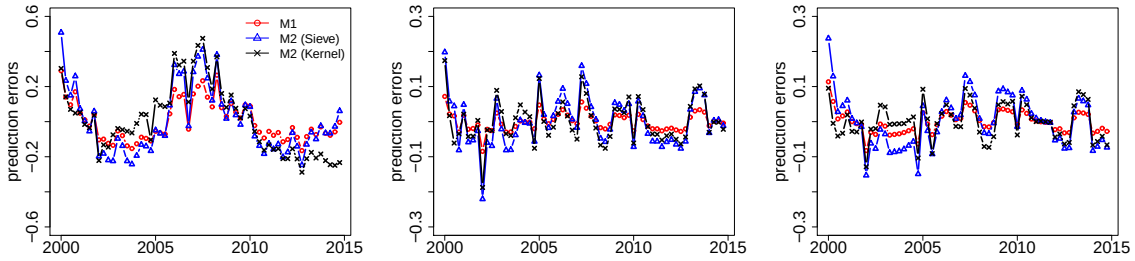
$$M2: \quad y_t = f(x_t) + u_{2t}, \tag{4.2}$$

where $f(\cdot)$ is an unknown link function, u_{2t} is the error term. Here, we use both kernel and sieve methods to estimate the unknown function $f(\cdot)$. For the kernel estimator, the second order Epanechnikov kernel is used with bandwidth $h = n^{-1/3}$, following Wang and Phillips (2009b). For the sieve estimator, we expand f by hermite polynomials like (2.2) and use least squares method to estimate the unknown coefficient c with the truncation parameter k selected by (4.1). It is found that $k = 3$ is selected in all three cases. The fitted curves for $PM_{2.5}$ versus lnGDP from both M1 and M2 are plotted in Figure 4.9.

We next evaluate the performance of the above models using two criteria, i.e., the in-sample and out-of-sample mean square forecast errors.

(i) In-sample mean squared forecast error (MSE_{is}). First, all unknown quantities in the two models are estimated based on the whole sample (x_t, y_t), $t = 1, 2, \cdots, n$. Then, we calculate the predictions $\widehat{y}_t^{\ell}$ with $\ell = 1, 2$,

$$\widehat{y}_t^1 = \frac{1}{n}\sum\nolimits_{i=1}^{n} \exp(\widehat{g}(x_t) + \widehat{e}_i),$$
$$\widehat{y}_t^2 = \widehat{f}(x_t),$$



(a) Beijing. (b) Hebei. (c) Shandong.

Figure 4.10: The in-sample prediction errors from M1, M2 with sieve estimator and M2 with kernel estimator.

where $\widehat{g}(x) = Z_k^T(x)\widehat{c}$, $\widehat{e}_i = \ln(y_i) - \widehat{g}(x_i)$. Note that $\widehat{y}_t^1$ is the smearing estimate of Duan (1983). The in-sample prediction errors produced from the two models are plotted in Figure 4.10. Finally, the in-sample mean squared forecast errors are calculated, for $\ell = 1, 2$, by

$$MSE_{is}(\ell) = \frac{1}{n}\sum\nolimits_{t=1}^{n}(y_t - \widehat{y}_t^{\ell})^2. \tag{4.3}$$

Meanwhile, to check the robustness of the results to the choice of sieve order k and bandwidth h, the MSE_{is} for $k = 4, 5$ or $h = 0.8n^{-1/3}, 1.2n^{-1/3}$ are calculated as well.

(ii) Out-of-sample mean squared forecast error (MSE_{oos}). We use the earlier part of the observations to fit the models, based on which we make predictions for the rest of the observations using a rolling window scheme. More precisely, for $j = 1, \cdots, 5$, we use the observations $\{(y_t, x_t)\} : j \le t \le 54 + j$ to obtain the estimates of unknown parameters and functions, with selected $k = 3$ for M1 and M2 with sieve estimator, and $h = n^{-1/3}$ for M2 with kernel estimator. Then forecasts are produced for y_{55+j}, respectively,

$$\begin{aligned}
\widehat{y}_{55+j}^1 &= \frac{1}{55}\sum\nolimits_{i=j}^{55+j} G^{-1}(\widehat{g}_j(x_{55+j}) + \widehat{e}_{j,i}, \widehat{\beta}_{j,n}),\\
\widehat{y}_{55+j}^2 &= \widehat{f}(x_{55+j}),
\end{aligned}$$

where $\widehat{e}_{j,i} = G(y_i, \widehat{\beta}_{j,n}) - \widehat{g}_j(x_i)$. The MSE_{oos} is calculated, for $\ell = 1, 2$, by

$$MSE_{oos}(\ell) = \frac{1}{5}\sum\nolimits_{j=1}^{5}(y_{55+j} - \widehat{y}_{55+j}^{\ell})^2. \tag{4.4}$$

In addition, to assess the choice of k and h, the MSE_{oos} for $k = 4, 5$ and that for $h =$

$0.8n^{-1/3}, 1.2n^{-1/3}$ are computed as well. All the MSE_{is} and MSE_{oos} are reported in Table 4.

Table 4: The MSEs ($\times 10^3$) for M1 and M2.

		M1			M2					
					Sieve			Kernel		
	k/h	3	4	5	3	4	5	$0.8n^{-1/3}$	$n^{-1/3}$	$1.2n^{-1/3}$
BJ	MSE_{is}	41.07	18.16	15.49	34.56	22.83	15.38	26.18	37.78	50.52
	MSE_{oos}	4.892	18.75	4.952	6.396	82.47	13.05	55.57	78.46	117.0
HB	MSE_{is}	4.858	4.702	4.063	4.869	4.700	4.071	3.362	3.651	3.775
	MSE_{oos}	1.629	3.145	4.869	1.644	3.145	4.771	1.765	1.731	1.653
SD	MSE_{is}	4.644	3.355	2.501	5.317	3.228	2.460	2.500	2.671	2.841
	MSE_{oos}	4.011	4.774	4.758	4.987	4.416	4.247	5.363	5.149	5.313

It is seen from Table 4 that M1 performs the best when $k = 5$ in terms of in-sample fit as one might expect. However, its performance achieves the best when $k = 3$ in terms of out-sample fit. This suggests that the criterion (4.1) to select k works well for out-of-sample prediction. Furthermore, our model M1 with the selected $k = 3$ has the smallest out-of-sample MSE in all three cases, compared to M2 with either sieve or kernel estimator. To be specific, for example, the out-of-sample MSE of M1 for Beijing is 4.892×10^{-3}, and is smaller than those of M2 with $k = 3$ (6.396×10^{-3}) and M2 with $h = n^{-1/3}$ (78.46×10^{-3}). Finally, the M2 with kernel estimator seems to provide better in-sample fits for Hebei and Shandong, but tends to perform worse than the sieve estimator (with the selected $\widehat{k} = 3$) in out-of-sample predictions. To sum up, the results suggest that, to analyze the EKC using Chinese data, the dependent variable $\text{PM}_{2.5}$ should be transformed by a logarithmic function, and our double nonlinear cointegration model opens spaces to improve the out-of-sample predictions.

5 Concluding Remarks

This paper studies a double nonlinear cointegration model, where the dependent variable, after a transformation by a strictly increasing parametric function, is related to a unit root nonstationary regressor with an unknown smooth link function. Estimation of the

unknown quantities is investigated. The asymptotic properties of the proposed estimators are established. Numerical studies reveal the nice performance of the estimators.

There are several possibilities to extend this work further. First, the univariate unit root regressor may be extended to a multiple case via an index structure. Second, the current setting assumes that the regressor is a unit root process, which may be generalized to nonstationary processes with long memory. Third, incorporating the information in the residual dependence might lead to more efficient estimation as in Linton and Wang (2016). Fourth, how to test the existence of the double-nonlinear cointegration remains a challenging issue. Some residual-based tests may be worth considering. Finally, the current model assumes a parametric transformation on y_t, which may be relaxed to be a nonparametric function as in Chiappori et al. (2015). Such extensions are technically demanding and the results will be reported elsewhere.

Acknowledgements

The authors thank the Co-editor, two anonymous referees, Chaohua Dong, and Ying Wang for helpful suggestions and comments. Tu would like to thank support from China's National Key Research Special Program Grants 2016YFC0207705 and National Natural Science Foundation of China (Grant 71532001, 71671002), the Center for Statistical Science at Peking University, and Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education.

Appendix

This appendix contains two parts. Part A presents some useful lemmas to facilitate the proofs for the main theorems of the paper in Part B.

Notations. For a matrix (vector) $A = (a_{ij})_{n\times m}$, $\|A\| = (\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2)^{1/2}$. Throughout this appendix, we denote a generic constant by C, which may be different at each appearance. And we simply denote $G(y_t, \beta)$ as $G_t(\beta)$.

A Useful Lemmas

Lemma A.1 *If $\{x_t\}_{t=1}^n$ satisfy Assumption 1, define $C_k = diag(n, n^2, \cdots, n^k)$, we have*

$$C_k^{-1/2} Z^T Z C_k^{-1/2} - D_k = o_p(1),$$

where $D_k = \Big(\int_0^1 \frac{1}{\sqrt{(i-1)!(j-1)!}} V^{i+j-2}(r) dr \Big)_{1 \le i,j \le k}$ is a $k \times k$ matrix.

Lemma A.2 *Denote that*

$$A_{n1}(\beta) = \sum_{t=1}^n G_t^2(\beta), \qquad A_{n5}(\beta) = \sum_{t=1}^n \dot{G}_t^2(\beta),$$
$$A_{n2}(\beta) = \sum_{t=1}^n [G_t(\beta) - g(x_t)]^2, \qquad A_{n6}(\beta) = \sum_{t=1}^n [G_t(\beta) - g(x_t)] \ddot{G}_t(\beta),$$
$$A_{n3}(\beta) = \sum_{t=1}^n G_t(\beta) \dot{G}_t(\beta), \qquad A_{n7}(\beta) = \sum_{t=1}^n G_t(\beta) \ddot{G}_t(\beta).$$
$$A_{n4}(\beta) = \sum_{t=1}^n [G_t(\beta) - g(x_t)] \dot{G}_t(\beta),$$

Let Assumptions 1-4 hold. As $n \to \infty$, it holds that

(a) $\frac{1}{n\kappa_{ng}^2} A_{n1}(\beta_0) \Rightarrow \int_0^1 h_g^2\big(V(r)\big) dr$;

(b) $\frac{1}{n} A_{n2}(\beta_0) \xrightarrow{p} \sigma_u^2$;

(c) $\frac{1}{n\kappa_{ng}\kappa_{n\xi,\beta_0}} A_{n3}(\beta_0) \Rightarrow \int_0^1 h_g\big(V(r)\big) h_\xi\big[h_g\big(V(r)\big), \beta_0\big] dr$;

(d) (1) if $\sqrt{n}/\kappa_{ng} \to 0$, $\frac{1}{\sqrt{n}\kappa_{n\xi,\beta_0}} A_{n4}(\beta_0) \Rightarrow \int_0^1 h_\xi\big[h_g\big(V(r)\big), \beta_0\big] dU(r)$,

(2) if $\sqrt{n}/\kappa_{ng} \to \alpha$, $\frac{1}{\sqrt{n}\kappa_{n\xi,\beta_0}} A_{n4}(\beta_0) \Rightarrow \int_0^1 h_\xi\big[h_g\big(V(r)\big), \beta_0\big] dU(r) + \alpha\sigma_u^2 \int_0^1 h'_\xi\big[h_g\big(V(r)\big), \beta_0\big] dr$,

(3) if $\kappa_{ng}/\sqrt{n} \to 0$ and $\kappa_{ng} \to \infty$, $\frac{\kappa_{ng}}{n\kappa_{n\xi,\beta_0}} A_{n4}(\beta_0) \Rightarrow \sigma_u^2 \int_0^1 h'_\xi\big[h_g\big(V(r)\big), \beta_0\big] dr$;

(e) $\frac{1}{n\kappa_{n\xi,\beta_0}^2} A_{n5}(\beta_0) \Rightarrow \int_0^1 h_\xi^2\Big[h_g\Big(V(r)\Big), \beta_0\Big] dr$;

(f) $A_{n6}(\beta_0) = o_p(n\kappa_{n\xi,\beta_0}^2)$;

(g) $A_{n7}(\beta_0) = o_p(n\kappa_{ng}\kappa_{n\xi,\beta_0}^2)$,

where α is a constant, $\kappa_{ng} = \kappa_g(\sqrt{n})$, $\kappa_{n\xi,\beta} = \kappa_\xi(\kappa_{ng}, \beta)$ and $h'_\xi(x,\beta) = \frac{\partial h_\xi(x,\beta)}{\partial x}$.

Lemma A.3 *Define* $C_n = n^{1/2-\rho}\kappa_{n\xi,\beta_0}$, *where* $0 < \rho < \varepsilon/6$ *with* ε *defined in Assumption 4 (c) and* $N_n = \{\beta : |C_n(\beta - \beta_0)| \leq 1\}$. *Let Assumptions 1-4 hold. If* $\sqrt{n}/\kappa_{ng} \to \alpha \geq 0$, *as* $n \to \infty$, *we have*

(a) $\sup_{\beta\in N_n} |A_{n1}(\beta) - A_{n1}(\beta_0)| = o_p(n^{1-2\rho}\kappa_{ng}^2)$;

(b) $\sup_{\beta\in N_n} |A_{n2}(\beta) - A_{n2}(\beta_0)| = o_p(n^{1-2\rho})$,

(c) $\sup_{\beta\in N_n} |A_{n3}(\beta) - A_{n3}(\beta_0)| = o_p(n^{1-2\rho}\kappa_{ng}\kappa_{n\xi,\beta_0})$;

(d) $\sup_{\beta\in N_n} |A_{n4}(\beta) - A_{n4}(\beta_0)| = o_p(n^{1-2\rho}\kappa_{n\xi,\beta_0})$;

(e) $\sup_{\beta\in N_n} |A_{n5}(\beta) - A_{n5}(\beta_0)| = o_p(n^{1-2\rho}\kappa_{n\xi,\beta_0}^2)$;

(f) $\sup_{\beta\in N_n} |A_{n6}(\beta) - A_{n6}(\beta_0)| = o_p(n^{1-2\rho}\kappa_{n\xi,\beta_0}^2)$;

(g) $\sup_{\beta\in N_n} |A_{n7}(\beta) - A_{n7}(\beta_0)| = o_p(n^{1-2\rho}\kappa_{ng}\kappa_{n\xi,\beta_0}^2)$.

Moreover,

(i) $\sup_{\beta\in N_n} |A_{n1}(\beta)| = O_p(n\kappa_{ng}^2)$;

(ii) $\sup_{\beta\in N_n} |A_{n2}(\beta)| = O_p(n)$,

(iii) $\sup_{\beta\in N_n} |A_{n3}(\beta)| = O_p(n\kappa_{ng}\kappa_{n\xi,\beta_0})$;

(iv) $\sup_{\beta\in N_n} |A_{n5}(\beta)| = O_p(n\kappa_{n\xi,\beta_0}^2)$;

(v) $\sup_{\beta\in N_n} |A_{n7}(\beta)| = O_p(n\kappa_{ng}\kappa_{n\xi,\beta_0}^2)$.

Lemma A.4 *Let Assumptions 1-4 hold. For each* β *satisfying* $|D_n(\beta-\beta_0)| \leq C$, *we have*

$$L_n(\beta) = \frac{\sum_{t=1}^n [G_t(\beta) - g(x_t)]^2}{\sum_{t=1}^n G_t^2(\beta)}(1 + o_p(1)), \tag{A.1}$$

where

$$D_n = \begin{cases} \sqrt{n}\kappa_{n\xi,\beta_0}, & \text{if } \sqrt{n}/\kappa_{ng} \to \alpha \geq 0; \\ \kappa_{ng}\kappa_{n\xi,\beta_0}, & \text{if } \kappa_{ng}/\sqrt{n} \to 0,\ \kappa_{ng} \to \infty. \end{cases}$$

B Proof of Main Theorems

Proof of Theorem 3.1. We consider three cases, i.e., Case I, if $\sqrt{n}/\kappa_{ng} \to 0$; Case II, if $\sqrt{n}/\kappa_{ng} \to \alpha$, where $\alpha \in \mathbb{R}_+$; and Case III, if $\kappa_{ng}/\sqrt{n} \to 0$, $\kappa_{ng} \to \infty$ as $n \to \infty$. Because the arguments used to prove the three cases are similar, we present the proofs for Case I and II below, and leave the detailed proofs for Case III to the online supplement.

Our proof contains four parts. Part (a) gives the score and hessian, and part (b) and (c) establish their asymptotics. Part (d) includes a detailed proof for the limit of $\widehat{\beta}_n - \beta_0$.

(a) The loss function. Since by Lemma A.4, we have

$$L_n(\beta) = \frac{\sum_{t=1}^n [G_t(\beta) - g(x_t)]^2}{\sum_{t=1}^n G_t^2(\beta)}(1 + o_p(1)),$$

and $n\kappa_{ng}^2$ does not rely on β, then minimizing $L_n(\beta)$ with respect to β is equivalent to minimizing $\tilde{L}_n(\beta) \equiv \frac{n\kappa_{ng}^2}{2}\frac{\sum_{t=1}^n [G_t(\beta)-g(x_t)]^2}{\sum_{t=1}^n G_t^2(\beta)}$. Therefore, the score function is

$$S_n(\beta) = n\kappa_{ng}^2 \frac{A_{n1}(\beta)A_{n4}(\beta) - A_{n2}(\beta)A_{n3}(\beta)}{A_{n1}^2(\beta)},$$

and the hessian is

$$\begin{aligned} J_n(\beta) = n\kappa_{ng}^2 A_{n1}^{-3}(\beta)\Big[&A_{n1}^2(\beta)\big(A_{n5}(\beta) + A_{n6}(\beta)\big) - A_{n1}(\beta)A_{n2}(\beta)\big(A_{n5}(\beta) + A_{n7}(\beta)\big) \\ &- 4A_{n1}(\beta)A_{n3}(\beta)A_{n4}(\beta) + 4A_{n2}(\beta)A_{n3}^2(\beta)\Big] \equiv n\kappa_{ng}^2 \frac{J_{n2}(\beta)}{J_{n1}(\beta)}. \end{aligned}$$

(b) The score. In view of Lemma A.2, we have
Case I: if $\sqrt{n}/\kappa_{ng} \to 0$,

$$\begin{aligned} (\sqrt{n}\kappa_{n\xi,\beta_0})^{-1} S_n(\beta_0) &= \frac{(\sqrt{n}\kappa_{n\xi,\beta_0})^{-1} A_{n4}(\beta_0)}{(n\kappa_{ng}^2)^{-1} A_{n1}(\beta_0)} + o(1). \\ &\Rightarrow \Big(\int_0^1 h_g^2\big(V(r)\big)dr\Big)^{-1} \int_0^1 h_\xi\Big[h_g\Big(V(r)\Big), \beta_0\Big] dU(r). \end{aligned}$$

Case II: if $\sqrt{n}/\kappa_{ng} \to \alpha$, where $\alpha \in \mathbb{R}_+$,

$$\begin{aligned} (\sqrt{n}\kappa_{n\xi,\beta_0})^{-1} S_n(\beta_0) &= \frac{(\sqrt{n}\kappa_{n\xi,\beta_0})^{-1} A_{n4}(\beta_0)}{(n\kappa_{ng}^2)^{-1} A_{n1}(\beta_0)} - \alpha \frac{n^{-1}A_{n2}(\beta_0)\cdot(n\kappa_{ng}\kappa_{n\xi,\beta_0})^{-1}A_{n3}(\beta_0)}{(n\kappa_{ng}^2)^{-2}A_{n1}^2(\beta_0)}. \\ &\Rightarrow \Big(\int_0^1 h_g^2\big(V(r)\big)dr\Big)^{-1}\Big(\int_0^1 h_\xi\big[h_g\big(V(r)\big), \beta_0\big]dU(r) + \alpha\sigma_u^2 \int_0^1 h_\xi'\big[h_g\big(V(r)\big), \beta_0\big]dr\Big) \\ &\quad - \alpha\sigma_u^2\Big(\int_0^1 h_g^2\big(V(r)\big)dr\Big)^{-2}\int_0^1 h_g\Big(V(r)\Big)h_\xi\Big[h_g\Big(V(r)\Big), \beta_0\Big]dr. \end{aligned}$$

(c) The hessian. Similarly, by Lemma A.2,

$$(\sqrt{n}\kappa_{n\xi,\beta_0})^{-2}J_n(\beta_0) = \frac{(\sqrt{n}\kappa_{n\xi,\beta_0})^{-2}A_{n5}(\beta_0)}{(n\kappa_{ng}^2)^{-1}A_{n1}(\beta_0)} + o_p(1)$$
$$\Rightarrow \Big(\int_0^1 h_g^2\big(V(r)\big)dr\Big)^{-1}\int_0^1 h_\xi^2\Big[h_g\Big(V(r)\Big),\beta_0\Big]dr.$$

(d) Detailed proof for the limit of $\widehat{\beta}_n - \beta_0$. Notice that

$$0 = S_n(\widehat{\beta}_n) = S_n(\beta_0) + J_n(\beta_n)(\widehat{\beta}_n - \beta_0), \tag{B.1}$$

where β_n is between $\widehat{\beta}_n$ and β_0. Define $D_n = \sqrt{n}\kappa_{n\xi,\beta_0}$. To obtain the limiting theorem, it is sufficient to show that

$$D_n(\widehat{\beta}_n - \beta_0) = -[D_n^{-1}J_n(\beta_0)D_n^{-1}]^{-1}\cdot D_n^{-1}S_n(\beta_0) + o_p(1). \tag{B.2}$$

We shall use Theorem 10.1 of Wooldridge (1994) to complete our proof, four conditions of which will be verified subsequently. Note that the first two conditions are trivially satisfied due to Assumptions 2-4. To verify the third condition, rewrite (B.1) as

$$S_n(\beta_0) + J_n(\beta_0)(\widehat{\beta}_n - \beta_0) + [J_n(\beta_n) - J_n(\beta_0)](\widehat{\beta}_n - \beta_0) = 0,$$

where β_n is between $\widehat{\beta}_n$ and β_0, $S_n(\beta_0)$ and $J_n(\beta_0)$ are the score and hessian at β_0, respectively, and $J_n(\beta_n)$ is the hessian at β_n. Let $C_n = n^{-\rho}D_n$ for some $0 < \rho < \varepsilon/6$ such that $C_nD_n^{-1} = o_p(1)$, where ε is defined in Assumption 4 (c). It follows that

$$\begin{aligned}0 &= D_n^{-1}S_n(\beta_0) + D_n^{-1}J_n(\beta_0)D_n^{-1}D_n(\widehat{\beta}_n - \beta_0) + D_n^{-1}[J_n(\beta_n) - J_n(\beta_0)]D_n^{-1}D_n(\widehat{\beta}_n - \beta_0)\\ &= D_n^{-1}S_n(\beta_0) + D_n^{-1}J_n(\beta_0)D_n^{-1}D_n(\widehat{\beta}_n - \beta_0) + n^{-2\rho}C_n^{-1}[J_n(\beta_n) - J_n(\beta_0)]C_n^{-1}D_n(\widehat{\beta}_n - \beta_0).\end{aligned}$$

As a result, the condition (iii) of Theorem 10.1 in Wooldridge (1994) will be satisfied if we can show

$$\sup_{\beta\in N_n}|C_n^{-1}[J_n(\beta_n) - J_n(\beta_0)]C_n^{-1}| = o_p(1), \tag{B.3}$$

where $N_n = \{\beta : |C_n(\beta - \beta_0)| \le 1\}$. Towards this end, write

$$\begin{aligned} J_n(\beta) - J_n(\beta_0) &= n\kappa_{ng}^2 \Bigg\{ \frac{1}{J_{n1}(\beta_0)}(J_{n2}(\beta) - J_{n2}(\beta_0)) - \frac{J_{n2}(\beta_0)}{J_{n1}^2(\beta_0)}(J_{n1}(\beta) - J_{n1}(\beta_0)) \\ &\quad + \frac{J_{n1}(\beta) - J_{n1}(\beta_0)}{J_{n1}(\beta)J_{n1}(\beta_0)} \Big[J_{n2}(\beta) - J_{n2}(\beta_0) - \frac{J_{n2}(\beta_0)(J_{n1}(\beta) - J_{n1}(\beta_0))}{J_{n1}(\beta_0)} \Big] \Bigg\} \\ &= n\kappa_{ng}^2 \Bigg\{ \frac{1}{J_{n1}(\beta_0)}(J_{n2}(\beta) - J_{n2}(\beta_0)) - \frac{J_{n2}(\beta_0)}{J_{n1}^2(\beta_0)}(J_{n1}(\beta) - J_{n1}(\beta_0)) \Bigg\} \frac{J_{n1}(\beta_0)}{J_{n1}(\beta)}. \end{aligned} \tag{B.4}$$

By Part (c) and Lemma A.2, we have

$$J_{n1}(\beta_0) = A_{n1}^3(\beta_0) = O_p((n\kappa_{ng}^2)^3),$$

and

$$J_{n2}(\beta_0) = A_{n1}^2(\beta_0)A_{n5}(\beta_0)(1 + o_p(1)) = O_p(n^3\kappa_{ng}^4\kappa_{n\xi,\beta_0}^2).$$

Thus, to show (B.3), it is suffices to show

$$\sup_{\beta \in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}(J_{n2}(\beta) - J_{n2}(\beta_0))| = o_p(1), \tag{B.5}$$

$$\sup_{\beta \in N_n} |n^{2\rho-3}\kappa_{ng}^{-6}(J_{n1}(\beta) - J_{n1}(\beta_0))| = o_p(1), \tag{B.6}$$

$$\sup_{\beta \in N_n} \Big| \frac{J_{n1}(\beta_0)}{J_{n1}(\beta)} \Big| = O_p(1). \tag{B.7}$$

First consider (B.5). Write

$$\begin{aligned} J_{n2}(\beta) - J_{n2}(\beta_0) &= \Big[A_{n1}^2(\beta)\big(A_{n5}(\beta) + A_{n6}(\beta)\big) - A_{n1}^2(\beta_0)\big(A_{n5}(\beta_0) + A_{n6}(\beta_0)\big) \Big] \\ &\quad - \Big[A_{n1}(\beta)A_{n2}(\beta)\big(A_{n5}(\beta) + A_{n7}(\beta)\big) - A_{n1}(\beta_0)A_{n2}(\beta_0)\big(A_{n5}(\beta_0) + A_{n7}(\beta_0)\big) \Big] \\ &\quad - 4\Big[A_{n1}(\beta)A_{n3}(\beta)A_{n4}(\beta) - A_{n1}(\beta_0)A_{n3}(\beta_0)A_{n4}(\beta_0) \Big] \\ &\quad + 4\Big[A_{n2}(\beta)A_{n3}^2(\beta) - A_{n2}(\beta_0)A_{n3}^2(\beta_0) \Big] \\ &\equiv \Upsilon_{n2,1} - \Upsilon_{n2,2} - 2\Upsilon_{n2,3} + 4\Upsilon_{n2,4}, \end{aligned} \tag{B.8}$$

where the definitions of $\Upsilon_{n2,1}$-$\Upsilon_{n2,4}$ should be obvious. For $\Upsilon_{n2,1}$, we have

$$
\begin{aligned}
&\sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\Upsilon_{n2,1}(\beta)| \\
&\leq n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\sup_{\beta\in N_n} |A_{n1}^2(\beta)[A_{n5}(\beta)-A_{n5}(\beta_0)]| \\
&+ n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\sup_{\beta\in N_n} |A_{n1}^2(\beta)[A_{n6}(\beta)-A_{n6}(\beta_0)]| \\
&+ n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\sup_{\beta\in N_n} |A_{n1}^2(\beta)-A_{n1}^2(\beta_0)|[A_{n5}(\beta_0)+A_{n6}(\beta_0)] \\
&\leq n^{-2}\kappa_{ng}^{-4}\sup_{\beta\in N_n} |A_{n1}(\beta)|^2\cdot n^{2\rho-1}\kappa_{n\xi,\beta_0}^{-2}\sup_{\beta\in N_n} |A_{n5}(\beta)-A_{n5}(\beta_0)| \\
&+ n^{-2}\kappa_{ng}^{-4}\sup_{\beta\in N_n} |A_{n1}(\beta)|^2\cdot n^{2\rho-1}\kappa_{n\xi,\beta_0}^{-2}\sup_{\beta\in N_n} |A_{n6}(\beta)-A_{n6}(\beta_0)| \\
&+ n^{2\rho-1}\kappa_{ng}^{-2}\sup_{\beta\in N_n} |A_{n1}(\beta)-A_{n1}(\beta_0)|\cdot n^{-1}\kappa_{ng}^{-2}\sup_{\beta\in N_n} |A_{n1}(\beta)+A_{n1}(\beta_0)|\cdot n^{-1}\kappa_{n\xi,\beta_0}^{-2}[A_{n5}(\beta_0)+A_{n6}(\beta_0)] \\
&= o_p(1),
\end{aligned}
$$

by Lemma A.3. Similarly, we can show

$$
\begin{aligned}
\sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\Upsilon_{n2,2}(\beta)| &= o_p(1), \\
\sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\Upsilon_{n2,3}(\beta)| &= o_p(1), \\
\sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\Upsilon_{n2,4}(\beta)| &= o_p(1).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}(J_{n2}(\beta)-J_{n2}(\beta_0))| \\
&\leq \sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\Upsilon_{n2,1}(\beta)| + \sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\Upsilon_{n2,2}(\beta)| \\
&+ 2\sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\Upsilon_{n2,3}(\beta)| + 4\sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-4}\kappa_{n\xi,\beta_0}^{-2}\Upsilon_{n2,4}(\beta)| \\
&= o_p(1),
\end{aligned}
$$

proving (B.5). Next, we show (B.6). Write that

$$
\begin{aligned}
&\sup_{\beta\in N_n} |n^{2\rho-3}\kappa_{ng}^{-6}(J_{n1}(\beta)-J_{n1}(\beta_0))| \\
&\leq n^{2\rho-1}\kappa_{ng}^{-2}\sup_{\beta\in N_n} |A_{n1}(\beta)-A_{n1}(\beta_0)|\cdot n^{-2}\kappa_{ng}^{-4}\sup_{\beta\in N_n} |A_{n1}^2(\beta)+A_{n1}(\beta)A_{n1}(\beta_0)+A_{n1}^2(\beta_0)| \\
&= o_p(1),
\end{aligned}
$$

by Lemma A.3. Regarding of (B.7), using arguments similar to those of Theorem 2.2 in Chan and Wang (2014), we have

$$\sup_{\beta\in N_n}\left|\frac{J_{n1}(\beta_0)}{J_{n1}(\beta)}\right| = O_p((n\kappa_{ng}^2)^3)\cdot O_p((n\kappa_{ng}^2)^{-3}) = O_p(1).$$

This completes the proof of (B.3). Finally, the convergence of $S_n(\beta_0)$ and $J_n(\beta_0)$ in Part (b) and (c) indicates that the condition (iv) in Wooldridge's theorem holds. Consequently, there exists a sequence of estimator $\widehat{\beta}_n$ of β_0 such that $D_n(\widehat{\beta}_n - \beta_0) = O_p(1)$ and hence the limiting distribution follows.

Proof of Theorem 3.2. Note that

$$\begin{aligned}\widehat{c} &= \widetilde{c}(\widehat{\beta}_n) = (Z^TZ)^{-1}Z^T\boldsymbol{G}(\widehat{\beta}_n)\\ &= (Z^TZ)^{-1}Z^T\boldsymbol{G}(\beta_0) + (Z^TZ)^{-1}Z^T(\boldsymbol{G}(\widehat{\beta}_n) - \boldsymbol{G}(\beta_0))\\ &= c + (Z^TZ)^{-1}Z^T(\gamma+u) + (Z^TZ)^{-1}Z^T(\boldsymbol{G}(\widehat{\beta}_n) - \boldsymbol{G}(\beta_0)).\end{aligned}$$

Then, write

$$\begin{aligned}\widehat{g}(x) - g(x) &= Z_k^T(x)\widehat{c} - g(x) = Z_k^T(x)(\widehat{c}-c) - \gamma_k(x)\\ &= Z_k^T(x)(Z^TZ)^{-1}Z^T(\gamma+u) + Z_k^T(x)(Z^TZ)^{-1}Z^T(\boldsymbol{G}(\widehat{\beta}_n) - \boldsymbol{G}(\beta_0)) - \gamma_k(x).\end{aligned}$$

Let $\Delta_z(x) = Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x)$. To fulfill the normality, we aim to show

$$\begin{aligned}&(i).\quad \sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)(Z^TZ)^{-1}Z^Tu \Rightarrow N(0,1);\\ &(ii).\quad \sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)(Z^TZ)^{-1}Z^T\gamma = o_p(1);\\ &(iii).\quad \sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)(Z^TZ)^{-1}Z^T(\boldsymbol{G}(\widehat{\beta}_n) - \boldsymbol{G}(\beta_0)) = o_p(1);\\ &(iv).\quad \sigma_u^{-1}\Delta_z^{-1/2}(x)\gamma_k(x) = o_p(1).\end{aligned}$$

We first show (i). It follows from Lemma A.1 that

$$\begin{aligned}&\sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)(Z^TZ)^{-1}Z^Tu\\ &\quad= \sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z^Tu(1+o_p(1))\\ &\quad= \sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}\sum_{t=1}^{n}Z_k(x_t)u_t(1+o_p(1)),\end{aligned}$$

which is a martingale array in view of Assumption 1. We shall use the martingale central limit theorem (Pollard (1984), Theorem VIII.1) to show the normality in (i). Define $\zeta_{nt} = \sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x_t)u_t$. The conditional variance process is

$$\begin{aligned}
\sum_{t=1}^n E_{\mathcal{F}_{n,t-1}}(\zeta_{nt}^2) &= \sigma_u^{-2}\Delta_z^{-1}(x)\sum_{t=1}^n [Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x_t)]^2 E_{\mathcal{F}_{n,t-1}}(u_t^2) \\
&= \Delta_z^{-1}(x)\cdot Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}\sum_{t=1}^n [Z_k(x_t)Z_k^T(x_t)]C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x) \\
&= \Delta_z(x)\cdot Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z^TZC_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x) \\
&= \Delta_z^{-1}(x)\cdot\Delta_z(x)(1+o_p(1)) \\
&= 1+o_p(1).
\end{aligned}$$

Next, we show that the Lindeberg's condition is satisfied, i.e.,

$$\forall\ \ \eta>0, \quad \sum_{t=1}^n E_{\mathcal{F}_{n,t-1}}(\zeta_{nt}^2 1\{|\zeta_{nt}|>\eta\}) = o_p(1). \tag{B.9}$$

Write

$$\begin{aligned}
&\sum_{t=1}^n E_{\mathcal{F}_{n,t-1}}(\zeta_{nt}^2 1\{|\zeta_{nt}|>\eta\}) \\
&= \sigma_u^{-2}\Delta_z^{-1}(x)\sum_{t=1}^n [Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x_t)]^2 E_{\mathcal{F}_{n,t-1}}(u_t^2 1\{|\zeta_{nt}|>\eta\}) \\
&\le \max_{1\le t\le n} E_{\mathcal{F}_{n,t-1}}(u_t^2 1\{|\zeta_{nt}|>\eta\})\cdot\sigma_u^{-2}\Delta_z^{-1}(x)\sum_{t=1}^n [Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x_t)]^2 \\
&= \max_{1\le t\le n} E_{\mathcal{F}_{n,t-1}}(u_t^2 1\{|\zeta_{nt}|>\eta\})\cdot\sigma_u^{-2}(1+o_p(1)).
\end{aligned}$$

Therefore, (B.9) will follow from

$$\max_{1\le t\le n} E_{\mathcal{F}_{n,t-1}}(u_t^2 1\{|\zeta_{nt}|>\eta\}) = o_p(1). \tag{B.10}$$

Applying the Hölder and Chebyshev inequalities, for some $0<\delta\le 2$, we have

$$\begin{aligned}
E_{\mathcal{F}_{n,t-1}}(u_t^2 1\{|\zeta_{nt}|>\eta\}) &\le E_{\mathcal{F}_{n,t-1}}^{\frac{2}{2+\delta}}(u_t^{2+\delta})\cdot P_{\mathcal{F}_{n,t-1}}^{\frac{\delta}{2+\delta}}(|\zeta_{nt}|>\eta) \\
&\le E_{\mathcal{F}_{n,t-1}}^{\frac{2}{2+\delta}}(u_t^{2+\delta})\cdot\left(\frac{E_{\mathcal{F}_{n,t-1}}[Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x_t)u_t]^2}{\eta^2\sigma_u^2\Delta_z(x)}\right)^{\frac{\delta}{2+\delta}}.
\end{aligned}$$

Since $E_{\mathcal{F}_{n,t-1}}(u_t^{2+\delta}) < \infty$ by Assumption 1, the condition

$$\max_{1\le t\le n} \frac{E_{\mathcal{F}_{n,t-1}}[Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x_t)u_t]^2}{\sigma_u^2\Delta_z(x)} = o_p(1), \tag{B.11}$$

is sufficient for (B.10). Note that

$$\begin{aligned}&\frac{E_{\mathcal{F}_{n,t-1}}[Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x_t)u_t]^2}{\sigma_u^2\Delta_z(x)}\\ &= \frac{[Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z_k(x_t)]^2}{\Delta_z(x)} \le \frac{\lambda_{\min}^{-2}(D_k)\|C_k^{-1/2}Z_k(x_t)\|^2}{\lambda_{\max}^{-1}(D_k)}.\end{aligned}$$

Since for $1 \le t \le n$, we have $\|C_{kz}^{-1/2}Z_k(x_t)\|^2 = \sum_{j=0}^{k-1}\frac{1}{n^{j+1}}h_j^2(x_t) = O_p(k/n) = o_p(1)$. This proves (B.11). Therefore, the asymptotic normality follows from the martingale central limit theorem of Pollard (1984).

Assertion (ii) holds since

$$\begin{aligned}&\sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)(Z^TZ)^{-1}Z^T\gamma\\ &\quad= \sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z^T\gamma(1+o_p(1))\\ &\quad\le \sigma_u^{-1}\Delta_z^{-1/2}(x)\|Z_k^T(x)C_k^{-1/2}D_k^{-1}C_k^{-1/2}Z^T\|\|\gamma\|(1+o_p(1))\\ &\quad= O_p(\|\gamma\|) = o_p(1),\end{aligned}$$

by Assumption 3 (c) and (d).

To prove (iii), using the mean value theorem, we have

$$\begin{aligned}&\sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)(Z^TZ)^{-1}Z^T(\boldsymbol{G}(\widehat{\beta}_n) - \boldsymbol{G}(\beta_0))\\ &\quad= \sigma_u^{-1}\Delta_z^{-1/2}(x)Z_k^T(x)(Z^TZ)^{-1}Z^T\dot{\boldsymbol{G}}(\beta_n)(\widehat{\beta}_n - \beta_0)\\ &\quad= \sigma_u^{-1}\Delta_z^{-1/2}(x)E(\dot{G}_t(\beta_n)|x_t)(\widehat{\beta}_n - \beta_0)(1+o_p(1))\\ &\quad= O_p(\sqrt{n/k})\cdot O_p(\widehat{\beta}_n - \beta_0) = o_p(1),\end{aligned}$$

by Theorem 3.1 and Assumption 4 (d).

It is readily seen that assertion (iv) holds, since by Lemma A.1 in Dong et al. (2016a) and Assumption 3 (d), $\sigma_u^{-1}\Delta_z^{-1/2}(x)\gamma_k(x) = O_p(\sqrt{n/k})\cdot O_p(k^{-m/2}) = o_p(1)$.

References

Abrevaya, J. (1999). Rank estimation of a transformation model with observed truncation. *The Econometrics Journal*, 2(2):292–305.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252.

Box, G. E. and Tiao, G. C. (1977). A canonical analysis of multiple time series. *Biometrika*, 64(2):355–365.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598.

Cai, Z., Li, Q., and Park, J. Y. (2009). Functional-coefficient models for nonstationary time series data. *Journal of Econometrics*, 148(2):101–113.

Carroll, R. J. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79(386):321–328.

Chan, N. and Wang, Q. (2014). Uniform convergence for nonparametric estimators with nonstationary data. *Econometric Theory*, 30(5):1110–1133.

Chan, N. and Wang, Q. (2015). Nonlinear regressions with nonstationary time series. *Journal of Econometrics*, 185(1):182–195.

Chang, Y. and Park, J. Y. (2003). Index models with integrated time series. *Journal of Econometrics*, 114(1):73–106.

Chen, S. (2002). Rank estimation of transformation models. *Econometrica*, 70(4):1683–1697.

Chiappori, P.-A., Komunjer, I., and Kristensen, D. (2015). Nonparametric identification and estimation of transformation models. *Journal of Econometrics*, 188(1):22–39.

Dong, C. and Gao, J. (2018). Specification testing in structural nonparametric cointegration. *Econometric Theory*, forthcoming.

Dong, C., Gao, J., and Peng, B. (2015). Semiparametric single-index panel data models with cross-sectional dependence. *Journal of Econometrics*, 188(1):301–312.

Dong, C., Gao, J., and Peng, B. (2016a). Another look at single-index models based on series estimation. Technical report, Monash University, Department of Econometrics and Business Statistics.

Dong, C., Gao, J., and Tjøstheim, D. (2016b). Estimation for single-index and partially linear single-index integrated models. *The Annals of Statistics*, 44(1):425–453.

Dong, C. and Linton, O. (2018). Additive nonparametric models with time variable and both stationary and nonstationary regressors. *Journal of Econometrics*, 207(1):212–236.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383):605–610.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, 55:251–276.

Fan, C. and Fine, J. P. (2013). Linear transformation model with parametric covariate transformations. *Journal of the American Statistical Association*, 108(502):701–712.

Florens, J.-P. and Sokullu, S. (2017). Nonparametric estimation of semiparametric transformation models. *Econometric Theory*, 33(4):839–873.

Gao, J., King, M., Lu, Z., and Tjøstheim, D. (2009). Nonparametric specification testing for nonlinear time series with nonstationarity. *Econometric Theory*, 25(6):1869–1892.

Gao, J. and Phillips, P. C. (2013a). Semiparametric estimation in triangular system equations with nonstationarity. *Journal of Econometrics*, 176(1):59–79.

Gao, J. and Phillips, P. C. B. (2013b). Functional coefficient nonstationary regression with non-and semi-parametric cointegration. *Working Paper*, Monash University.

Granger, C. (1991). Some recent generalisations of cointegration and the analysis of longrun relationships. In Engle, R. and Granger, C., editors, *Long-Run Economic Relationships*, pages 277–287. Oxford University Press.

Granger, C. W. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1):121–130.

Grossman, G. M. and Krueger, A. B. (1991). Environmental impacts of a north american free trade agreement. *NBER Working Paper*, No.3914.

Han, A. K. (1987). A non-parametric analysis of transformations. *Journal of Econometrics*, 35:191–209.

Hirukawa, M. and Sakudo, M. (2018). Functional-coefficient cointegration models in the presence of deterministic trends. *Econometric Reviews*, 37(5):507–533.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, pages 103–137.

Johansen, S. (1995). *Likelihood-Based inference in Cointegrated Vector in Gaussian Vector Autoregressive Model*. Oxford University Press, Oxford.

Karlsen, H. A., Myklebust, T., and Tjøstheim, D. (2007). Nonparametric estimation in a nonlinear cointegration type model. *The Annals of Statistics*, 35(1):252–299.

Kasparis, I., Andreou, E., and Phillips, P. C. (2015). Nonparametric predictive regression. *Journal of Econometrics*, 185(2):468–494.

Kasparis, I. and Phillips, P. C. (2012). Dynamic misspecification in nonparametric cointegrating regression. *Journal of Econometrics*, 168(2):270–284.

Kuznets, S. (1955). Economic growth and income inequality. *American Economic Review*, 45(1):1–28.

Lewbel, A., Lu, X., and Su, L. (2015). Specification testing for transformation models with an application to generalized accelerated failure-time models. *Journal of Econometrics*, 184(1):81–96.

Liao, Z. and Phillips, P. C. (2015). Automated estimation of vector error correction models. *Econometric Theory*, 31(3):581–646.

Lin, Y. and Tu, Y. (2019). Identification and estimation of a semiparametric single index transformation model. *Working Paper*, Peking University.

Linton, O., Sperlich, S., and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *The Annals of Statistics*, 36(2):686–718.

Linton, O. and Wang, Q. (2016). Nonparametric transformation regression with nonstationary data. *Econometric Theory*, 32(1):1–29.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Park, J. Y. and Phillips, P. C. B. (1999). Asymptotics for nonlinear transformations of integrated time series. *Econometric Theory*, 15(3):269–298.

Park, J. Y. and Phillips, P. C. B. (2000). Nonstationary binary choice. *Econometrica*, 68(5):1249–1280.

Park, J. Y. and Phillips, P. C. B. (2001). Nonlinear regressions with integrated time series. *Econometrica*, 69(1):117–161.

Phillips, P. C. B. (2009). Local limit theory and spurious nonparametric regression. *Econometric Theory*, 25(6):1466–1497.

Phillips, P. C. B., Li, D., and Gao, J. (2017). Estimating smooth structural change in cointegration models. *Journal of Econometrics*, 196:180–195.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441.

Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83(402):394–405.

Tsay, R. and Tiao, G. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary arma models. *Journal of the American Statistical Association*, 79(385):84–96.

Tu, Y. and Wang, Y. (2018). Spurious functional-coefficient regression models and robust inference with marginal integration. *Working Paper*, Peking University.

Tu, Y. and Wang, Y. (2019). Functional coefficient cointegration models subject to time-varying volatility with an application to the purchasing power parity. *Oxford Bulletin of Economics and Statistics*, forthcoming.

Tu, Y., Yao, Q., and Zhang, R. (2019). Error-correction factor models for high-dimensional cointegrated time series. *Statistica Sinica*, forthcoming.

Tu, Y. and Yi, Y. (2017). Forecasting cointegrated nonstationary time series with time-varying variance. *Journal of Econometrics*, 196(1):83–98.

Uematsu, Y. (2017). Nonstationary nonlinear quantile regression. *Econometric Reviews*, forthcoming.

Vanhems, A. and Van Keilegom, I. (2018). Estimation of a semiparametric transformation model in the presence of endogeneity. *Econometric Theory*, forthcoming.

Wang, N. and Ruppert, D. (1995). Nonparametric estimation of the transformation in the transform-both-sides regression model. *Journal of the American Statistical Association*, 90(430):522–534.

Wang, N. and Ruppert, D. (1996). Estimation of regression parameters in a semiparametric transformation model. *Journal of Statistical Planning and Inference*, 52(3):331–351.

Wang, Q. and Phillips, P. C. B. (2009a). Asymptotic theory for local time density estimation and nonparametric cointegrating regression. *Econometric Theory*, 25(3):710–738.

Wang, Q. and Phillips, P. C. B. (2009b). Structural nonparametric cointegrating regression. *Econometrica*, 77(6):1901–1948.

Wang, Q. and Phillips, P. C. B. (2012). A specification test for nonlinear nonstationary models. *The Annals of Statistics*, 40(2):727–758.

Wang, Q. and Phillips, P. C. B. (2016). Nonparametric cointegrating regression with endogeneity and long memory. *Econometric Theory*, 32(2):359–401.

Wang, Q., Wu, D., and Zhu, K. (2018). Model checks for nonlinear cointegrating regression. *Journal of Econometrics*, 207(2):261–284.

Wooldridge, J. M. (1994). Estimation and inference for dependent processes. In R.F., E. and D., M., editors, *Handbook of Econometrics*, volume IV, chapter 45, pages 2639–2738. North-Holland, Amsterdam.

Xiao, Z. (2009). Functional-coefficient cointegration models. *Journal of Econometrics*, 152(2):81–92.

Ye, J. and Duan, N. (1997). Nonparametric $n^{-1/2}$-consistent estimation for the general transformation models. *The Annals of Statistics*, 25(6):2682–2717.

Zhang, R., Robinson, P., and Yao, Q. (2019). Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association*, forthcoming.