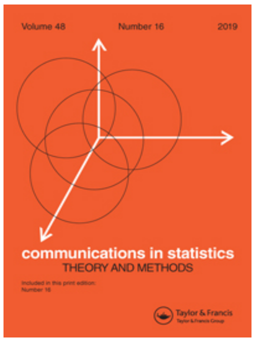

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Adaptive estimation for varying coefficient models with nonstationary covariates

Zhiyong Zhou & Jun Yu

To cite this article: Zhiyong Zhou & Jun Yu (2019) Adaptive estimation for varying coefficient models with nonstationary covariates, Communications in Statistics - Theory and Methods, 48:16, 4034-4050, DOI: 10.1080/03610926.2018.1484483

To link to this article: https://doi.org/10.1080/03610926.2018.1484483

© 2018 Informa UK Limited, trading as Taylor & Francis Group

Published online: 29 Dec 2018.

Submit your article to this journal

Article views: 491

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lsta20

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2019, VOL. 48, NO. 16, 4034–4050
https://doi.org/10.1080/03610926.2018.1484483

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Adaptive estimation for varying coefficient models with non stationary covariates

Zhiyong Zhou and Jun Yu

Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden

ABSTRACT
In this paper, the adaptive estimation for varying coefficient models proposed by Chen, Wang, and Yao (2015) is extended to allowing for non stationary covariates. The asymptotic properties of the estimator are obtained, showing different convergence rates for the integrated covariates and stationary covariates. The nonparametric estimator of the functional coefficient with integrated covariates has a faster convergence rate than the estimator with stationary covariates, and its asymptotic distribution is mixed normal. Moreover, the adaptive estimation is more efficient than the least square estimation for non normal errors. A simulation study is conducted to illustrate our theoretical results.

ARTICLE HISTORY
Received 13 April 2017
Accepted 28 May 2018

KEYWORDS
Varying coefficient model; adaptive estimation; local linear fitting; non stationary covariates

1. Introduction

Nonstationarity is a very important empirical feature in many economic and financial time series. Over the past decade, there has been great interests in nonparametric and semiparametric models with non stationary covariates, existing literature includes Cai, Li, and Park (2009), Chan and Wang (2015), Chen, Fang, and Li (2015), Chen, Gao, and Li (2012), Dong, Gao, and Tjøstheim (2016), Gu and Liang (2014), Gao and Phillips (2013), Juhl and Xiao (2005), Karlsen, Myklebust, and Tjøstheim (2007), Karlsen and Tjostheim (2001), Liang, Lin, and Hsiao (2015), Li et al. (2017), Sun, Cai, and Li (2013), Sun and Li (2011), Wang (2014), Wang (2015), Wang and Phillips (2009a), Wang and Phillips (2009b), Wang and Phillips (2016), Xiao (2009), Zhou and Lin (2018). As we know, compared with nonparametric regression model, semiparametric regression models have the advantage of attenuating the problem of "curse of dimensionality." Among them, varying coefficient models proposed by Hastie and Tibshirani (1993) have gained considerable attention due to their flexibility and good interpretability. By allowing the functional coefficients to vary over a index variable, it is an useful extension of the classical linear model. Thus, to study the varying coefficient models with non stationary covariates is very meaningful from both theoretical and practical aspects.

In this paper, we focus on the varying coefficient models with the form:

$$Y_t = X_t^T\beta(U_t) + \varepsilon_t = X_{t,1}^T\beta_1(U_t) + X_{t,2}^T\beta_2(U_t) + \varepsilon_t, \quad t = 1, 2, \dots, n \tag{1}$$

where $X_t = (X_{t,1}^T, X_{t,2}^T)^T$, $X_{t,i}$ is a $d_i \times 1$ vector, $i = 1, 2$, $d_1 + d_2 = d$, $X_{t,1}$ is stationary, $X_{t,2}$ is non stationary, specifically an integrated of order one time series (denoted by I(1), see

CONTACT Zhiyong Zhou zhiyong.zhou@umu.se Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, 901 87, Sweden.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.
© 2018 The Author(s), Published by Taylor & Francis Group, LLC This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Hamilton (1994) for details), $\beta(\cdot) = (\beta_1(\cdot)^T, \beta_2(\cdot)^T)^T$ is a d-dimensional vector of unknown smooth functions, U_t is a scalar stationary variable, and ε_t is a stationary error term. In what follows, we assume $E(\varepsilon_t|X_t, U_t) = 0$, that is X_t and U_t are uncorrelated with ε_t. Note that Y_t is allowed to be stationary or non stationary in this model.

Cai, Li, and Park (2009) and Xiao (2009) have considered this model by using local liner estimation method, which is based on the least squares type criteria. For ordinary least squares based estimators, a well-known drawback is that the efficiency will be reduced when non-Gaussian errors are present. Motivated by efficiency considerations, we adopt the adaptive estimation procedure so that the new estimator can be adapted to different error distributions. This approach has been introduced in nonparametric models (Linton and Xiao 2007; Jin, Su, and Xiao 2015) and varying coefficient models (Chen, Wang, and Yao 2015). However, only the cases where covariates are independent or stationary have been considered. In this paper, we will extend the adaptive estimation procedure to the varying coefficient model Equation (1) which involves both integrated (non stationary) covariates and stationary covariates. We establish the asymptotic distributions of the proposed adaptive estimators and show different convergence rates for the non stationary covariates and stationary covariates. The efficient EM algorithm proposed by Chen, Wang, and Yao (2015) is applied to implement our adaptive estimation procedure. Our simulation study illustrates that the adaptive estimator is more efficient than the ordinary least squares estimator when the error distribution deviates from normal.

The paper is organized as follows. Section 2 introduces adaptive estimation method and the EM algorithm. The asymptotic results for the estimators are given in Section 3. A simulation study is presented in Section 4. Section 5 is devoted to the conclusion. Finally, the proofs are postponed to Section 6.

2. Adaptive estimation

Suppose that $\{X_t, U_t, Y_t, t = 1, 2, \ldots, n\}$ are the observations. Assuming that $\beta(\cdot)$ has second order derivative, we have the Taylor expansion:

$$\beta(u) \approx \beta(u_0) + \beta'(u_0)(u - u_0)$$

for u in a small neighborhood of u_0. By adopting the local linear fitting method (Fan and Gijbels 1996), the traditional least squares based local linear estimation of $\Psi(u_0) = (\beta(u_0)^T, h\beta'(u_0)^T)^T$ is to minimize the weighted loss function (with respect to Θ):

$$\sum_{t=1}^{n} \{Y_t - X_t(u_0)^T \Theta\}^2 K_h(U_t - u_0) \tag{2}$$

where Θ is a $2d$-dimensional vector, $X_t(u_0) = \begin{pmatrix} X_t \\ h^{-1}(U_t - u_0)X_t \end{pmatrix}$, and $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ being the kernel function and h being the bandwidth. We use a Gaussian kernel for $K(\cdot)$ throughout the paper for convenience.

As the efficiency for the resulting least squares based estimator in Equation (2) may be reduced when the error is apart from normal, an estimation procedure which is adaptive to error distributions (Chen, Wang, and Yao 2015; Linton and Xiao 2007; Jin, Su, and Xiao 2015) is adopted here. Specifically, with $f(\cdot)$ being the probability density function of the error term and assuming to be known, the parameter vector $\Psi(u_0)$ is estimated by maximizing the following local log-likelihood function (with respect to Θ):

$$\sum_{t=1}^{n} \log f\{Y_t - X_t(u_0)^T\Theta\}K_h(U_t - u_0) \tag{3}$$

Notice that the error probability density function is generally unknown in practice and has to be estimated beforehand. One can use a kernel density estimator, among others, based on some initial error estimators. For example, applying least squares or median regression based local linear estimator to obtain initial parameter estimator $\tilde{\beta}(\cdot)$, we can get an error estimator $\tilde{\varepsilon}_s = Y_s - \tilde{\beta}(U_s)^T X_s, s = 1, 2, \ldots, n$. These residuals can then be used to attain the kernel density estimator of $f(\cdot)$ as follows:

$$\tilde{f}(\varepsilon_t) = \frac{1}{n}\sum_{s\neq t}^{n} K_{h_0}(\varepsilon_t - \tilde{\varepsilon}_s), \quad \text{for } t, s = 1, 2, \ldots, n \tag{4}$$

where h_0 is an appropriately chosen bandwidth. Here, the leave-one-out kernel density estimator is used to wash off the estimation bias. Then the adaptive local linear estimator for $\Psi(u_0)$ is

$$\hat{\Psi}(u_0) = \arg\max_{\Theta} L_n(\Theta, u_0) \tag{5}$$

where $L_n(\Theta, u_0)$ is the estimated local log-likelihood function:

$$\begin{aligned} L_n(\Theta, u_0) &= \sum_{t=1}^{n} \log \tilde{f}\{Y_t - X_t(u_0)^T\Theta\}K_h(U_t - u_0) \\ &= \sum_{t=1}^{n} \log\left(\frac{1}{n}\sum_{s\neq t}^{n} K_{h_0}\{Y_t - X_t(u_0)^T\Theta - \tilde{\varepsilon}_s\}\right) K_h(U_t - u_0) \end{aligned} \tag{6}$$

Algorithm: At $(k+1)$th step, the following E and M steps are operated:
E-step: Calculation of the classification probabilities $p_{ts}^{(k+1)}$ with

$$p_{ts}^{(k+1)} = \frac{K_{h_0}\{Y_t - X_t(u_0)^T\hat{\Psi}^{(k)}(u_0) - \tilde{\varepsilon}_s\}}{\sum_{s\neq t}^{n} K_{h_0}\{Y_t - X_t(u_0)^T\hat{\Psi}^{(k)}(u_0) - \tilde{\varepsilon}_s\}}, \quad 1 \le s \neq t \le n$$

M-step: Update of the estimator $\hat{\Psi}^{(k+1)}(u_0)$ with

$$\begin{aligned} \hat{\Psi}^{(k+1)}(u_0) &= \arg\max_{\Theta} \sum_{t=1}^{n}\sum_{s\neq t}^{n}\{p_{st}^{(k+1)} K_{h_0}(U_t - u_0)\log\left(K_{h_0}\{Y_t - X_t(u_0)^T\Theta - \tilde{\varepsilon}_s\}\right)\} \\ &= \arg\min_{\Theta} \sum_{t=1}^{n}\sum_{s\neq t}^{n}\{p_{st}^{(k+1)} K_{h_0}(U_t - u_0)[Y_t - X_t(u_0)^T\Theta - \tilde{\varepsilon}_s]^2\} \\ &= \left(\sum_{t=1}^{n}\sum_{s\neq t}^{n} p_{st}^{(k+1)} K_{h_0}(U_t - u_0)X_t(u_0)X_t(u_0)^T\right)^{-1} \\ &\quad \sum_{t=1}^{n}\sum_{s\neq t}^{n} p_{st}^{(k+1)} K_{h_0}(U_t - u_0)(Y_t - \tilde{\varepsilon}_s)X_t(u_0) \end{aligned}$$

Compared to the estimation Equation (2) which has a solution with closed form (see Cai, Li, and Park 2009), the estimation Equation (5) does not have an explicit solution. A modified EM algorithm proposed in Chen, Wang, and Yao (2015) is applied here. That is, let $\hat{\Psi}^{(0)}(u_0)$ be an initial estimator of parameter $\Psi(u_0)$, then we can update the parameter estimator according to the Algorithm 2.1 in Chen, Wang, and Yao (2015). In order to be self-contained, we list the algorithm as above.

Then, according to the Proposition 2.1 in Chen, Wang, and Yao (2015) (it still holds in our settings as its proof procedure involves no conditions of dependence), the above EM algorithm monotonically increases the estimated local log-likelihood $L_n(\Psi, u_0)$ after each iteration, i.e., $L_n(\hat{\Psi}^{(k+1)}(u_0), u_0) \geq L_n(\hat{\Psi}^{(k)}(u_0), u_0)$. Thus, $\hat{\Psi}^{(k+1)}(u_0)$ converges to $\hat{\Psi}(u_0)$ for large k.

3. Asymptotic results

Before presenting the asymptotic results, we give some notations and assumptions. For $j \geq 0$, we define $\mu_j = \int_{-\infty}^{\infty} r^j K(r)\mathrm{d}r$ and $\nu_j = \int_{-\infty}^{\infty} r^j K^2(r)\mathrm{d}r$. Assume the I(1) vector $X_{t,2} = X_{t-1,2} + \eta_t$ $(1 \leq t \leq n)$, where η_t is a strictly stationary α-mixing random vector process with mean zero satisfying, for some $p_0 > 0$,

$$E|\eta_t|^{2+p_0} < \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} k^{(2+p_0)/p_0}\alpha(k) < \infty$$

where $\alpha(\cdot)$ is the mixing coefficient. Here we adopt the condition (2.5) in Cai, Li, and Park (2009) to ensure that

$$X_{[nr],2}/\sqrt{n} \Rightarrow B(r)$$

where $r \in [0,1]$, $[\cdot]$ denotes the bracket function and $B(r)$ is a d_2-dimensional Brownian motion on $[0,1]$ with covariance matrix $\Omega_\eta = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2}\sum_{t=1}^{n}\eta_t)$, see Cai, Li, and Park (2009); Merlevède, Peligrad, and Utev (2006) for more details. Then, for $k = 1, 2$, we have

$$\frac{1}{n}\sum_{t=1}^{n}(X_{t,2}/\sqrt{n})^{\otimes k} \xrightarrow{d} \int_0^1 B(r)^{\otimes k}\mathrm{d}r =: \Gamma_k \quad \text{as } n \to \infty$$

where $A^{\otimes 2} = AA^T(A^{\otimes 1} = A)$ for a vector or matrix A, (see Berkes and Horváth (2006); Billingsley (1999) for details). Define $M_\ell(u_0) = E[X_{t,1}^{\otimes \ell}|U_t = u_0]$ for $\ell = 1, 2$. And, let

$$S(u_0) = \begin{pmatrix} M_2(u_0) & M_1(u_0)\Gamma_1^T \\ \Gamma_1 M_1(u_0)^T & \Gamma_2 \end{pmatrix}$$

Then, throughout the paper, we make the following assumptions:

- A1. $M_2(u_0)$ is positive-definite. Moreover, $M_1(u)$ and $M_2(u)$ are continuous in a neighborhood of u_0.
- A2. Let $\omega_t = (X_{t,1}^T, \eta_t^T, \varepsilon_t)^T$. Assume that $\{(\omega_t^T, U_t)\}$ is strictly stationary α-mixing process with the p_1-th moment $(p_1 > 2)$. $E[|\varepsilon_t X_{t,1}^2|^{p_2}|U_t = u] \leq C_1 < \infty$ with $p_2 > p_1$ and $\alpha(t) = O(t^{-p_3})$ for some $p_3 > \min\{p_2p_1/(p_2 - p_1), p_5, 2p_6/(2 - p_6)\}$, where $p_5 = p_4p_1/(p_4p_1 - p_1 - p_4)$ for some p_4 satisfying $p_1/(p_1 - 1) < p_4 < 2$. Also, $\|\eta_t\|_{q_0} = [E|\eta_t|^{q_0}]^{1/q_0} < \infty$ with $q_0 = p_4p_6/(p_4 - p_6)$ for some $1 < p_6 < p_4$. Further, $\sup_k E[\eta_1^2\varepsilon_{k+1}^2|U_{k+1} = u] \leq C_2 < \infty$.

- A3. Let $\rho(\cdot) = \log f(\cdot)$. Assume $E[\rho'(\varepsilon_t)] = 0$, $E[\rho''(\varepsilon_t)] = \delta_1 < 0$, $E[\rho'(\varepsilon_t)^2] < \infty$ and $\rho'''(\cdot)$ is bounded.
- A4. U_t has a compact support Ξ. Its probability density function $g(u)$ is positive and bounded away from 0 and infinity, and has second order continuous derivative when u is in Ξ. Furthermore, the conditional density function of (U_0, U_s) given $(X_{0,1} = x_0, X_{s,1} = x_s)$, $g(u_0, u_s|x_0, x_s, s)$, is bounded for all $s \geq 1$.
- A5. The function $\beta(u)$ has second order continuous derivatives when u is in the compact support Ξ.
- A6. As $n \to \infty$, we have $h \to 0$ and $nh \to \infty$. Furthermore, $n^{1/2-p_1/4}h^{p_1/p_2-1/2-p_1/4} = O(1)$.

These regularity conditions are adopted from Cai, Li, and Park (2009); Chen, Wang, and Yao (2015) to facilitate the proofs of our main theoretical results. As is discussed in Cai, Li, and Park (2009), the assumption A2 is fulfilled with some standard moment conditions if $\alpha(\cdot)$ decays geometrically and the assumption A6 is satisfied for the optimal bandwidths selection ($h = cn^{-\gamma}$ for $0 < \gamma < 1$ and some $c > 0$) under minor conditions, see Cai, Li, and Park (2009) for details. These two technical conditions are assumed to ensure that our mixed normal limiting results Equation (18) hold by adopting the existing results in Cai, Li, and Park (2009). To obtain the mixed normal limiting results, these conditions might be relaxed by using the weak convergence to stochastic integrals tool given in Liang et al. (2016). But we do not pursue this extension here. With these notations and assumptions, we have obtained the following asymptotic properties of the proposed adaptive estimator for parameter vector $\Psi(u_0)$.

Theorem 3.1. *Suppose that the assumptions A1–A6 are satisfied. Then with probability approaching 1, there exists a consistent local maximizer $\hat{\Psi}(u_0)$ of Equation (5) such that*

$$\hat{\Psi}(u_0) - \Psi(u_0) = O_p\left(\left(\sqrt{nh}\right)^{-1} + h^2\right) \tag{7}$$

Theorem 3.1 gives us the consistency of the adaptive estimator $\hat{\Psi}(u_0)$. To present the asymptotic distribution of $\hat{\Psi}(u_0)$, define $D_n = \mathrm{diag}\{\mathrm{I}_{d_1}, \sqrt{n}\mathrm{I}_{d_2}\}$, $\Delta_n = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes D_n$, where $\otimes$ denotes the Kronecker product.

Theorem 3.2. *Suppose that the assumptions A1–A6 hold. Then $\hat{\Psi}(u_0)$, given in Theorem 3.1, has the following asymptotic distribution*

$$\sqrt{nh}\Delta_n\left[\hat{\Psi}(u_0) - \Psi(u_0) - \frac{h^2}{2}R(u_0)^{-1}\begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \otimes \left(S(u_0)\beta''(u_0)\right)(1 + o_p(1))\right] \xrightarrow{d} MN(\Sigma_\Psi(u_0)) \tag{8}$$

where $R(u_0) = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \otimes S(u_0)$, $MN(\Sigma_\Psi(u_0))$ is a mixed normal distribution with zero mean and random covariance matrix $\Sigma_\Psi(u_0) = [E\{\rho'(\varepsilon_t)^2\}]^{-1}g(u_0)^{-1}R(u_0)^{-1}\Lambda(u_0)R(u_0)^{-1}$ where $\Lambda(u_0) = \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes S(u_0)$

Remark 1. When there is no integrated covariates ($d_2 = 0$), this theorem reduces to the Theorem 2.2 in Chen, Wang, and Yao (2015). As is presented, when non stationary covariates are incorporated, the convergence rate involves an additional item Δ_n compared to the convergence rate $\sqrt{nh}$ in the stationary case. This indicates that the convergence rate of the estimator for non stationary part is $n\sqrt{h}$, which is faster than that of the estimator for the stationary part. Moreover, it shows that the asymptotic distribution of the non stationary part is mixed normal while that of the stationary part is normal.

Based on Theorem 3.2, we can immediately obtain the asymptotic distribution of the parameter estimator $\hat{\beta}(u_0) = (\mathrm{I}_d, 0_d)\hat{\Psi}(u_0)$ as presented in Theorem 3.3.

Theorem 3.3. *When the assumptions of Theorem 3.2 hold, $\hat{\beta}(u_0)$ has the following asymptotic distribution*

$$\sqrt{nh}D_n\left[\hat{\beta}(u_0) - \beta(u_0) - \frac{h^2}{2}\mu_2\beta''(u_0)(1+o_p(1))\right] \xrightarrow{d} MN(\Sigma_\beta(u_0)) \tag{9}$$

where $MN(\Sigma_\beta(u_0))$ is a mixed normal distribution with zero mean and random covariance matrix $\Sigma_\beta(u_0) = [E\{\rho'(\varepsilon_t)^2\}]^{-1}\nu_0 S(u_0)^{-1}/g(u_0)$.

Remark 2. Note that the proposed adaptive estimator and the traditional least squares based local linear estimator by minimizing Equation (2) (see Theorem 2.1 in Cai, Li, and Park (2009)) have the same asymptotic bias, but have slight different asymptotic variance (the asymptotic variance for traditional local linear estimator is obtained just by replacing $[E\{\rho'(\varepsilon_t)^2\}]^{-1}$ by $E(\varepsilon_t^2)$ in $\Sigma_\beta(u_0)$). In fact, the adaptive estimator has always smaller asymptotic variance than the traditional local linear estimator for non-Gaussian errors. This can be easily verified by using Cauchy-Schwarz inequality, since it holds that $E(\varepsilon_t^2)E\{\rho'(\varepsilon_t)^2\} \geq [E\{\varepsilon_t\rho'(\varepsilon_t)\}]^2 = 1$ and the equality holds if and only if $f(\cdot)$ is a normal density. Therefore, $[E\{\rho'(\varepsilon_t)^2\}]^{-1} \leq E(\varepsilon_t^2)$. Thus, the adaptive estimator is more efficient than the least squares based estimator when the error is not normal.

4. Simulation

In this section, a simulation study is carried out to compare the performance between the adaptive estimator (adapt) and the least squares based local linear estimator (LS), for varying coefficient models with non stationary covariates. In order to facilitate the comparison, we first consider the independent and identically distributed (i.i.d) error with the following five distributions in the simulation experiment which were also considered in Chen, Wang, and Yao (2015):

(a) $N(0,1)$
(b) t_3
(c) $0.5N(-1,0.5^2) + 0.5N(1,0.5^2)$
(d) $0.3N(-1.4,1) + 0.7N(0.6,0.4^2)$
(e) $0.9N(0,1) + 0.1N(0,10^2)$

Among these distributions, the standard normal distribution $N(0,1)$ is the baseline for the comparison. t_3 denotes the standard t-distribution with degree of freedom 3. The three remaining distributions are all mixed normal, but own different characteristics: (c) is bimodal, (d) is left skewed, while (e) is contaminated by some outliers.

Table 1. Comparison of AMSE and its standard error in brackets for Case (i).

	$n=200$		$n=400$	
ε	LS	Adapt	LS	Adapt
a	0.0362(0.0196)	0.0355(0.0182)	0.0183(0.0089)	0.0180(0.0086)
b	0.0893(0.0620)	0.0708(0.0380)	0.0501(0.0311)	0.0437(0.0296)
c	0.0453(0.0256)	0.0394(0.0249)	0.0232(0.0108)	0.0180(0.0095)
d	0.0425(0.0249)	0.0339(0.0230)	0.0229(0.0114)	0.0160(0.0077)
e	0.3281(0.2881)	0.1572(0.1478)	0.1870(0.1363)	0.0867(0.1037)

Table 2. Comparison of AMSE and its standard error in brackets for Case (ii).

	$n=200$		$n=400$	
ε	LS	Adapt	LS	Adapt
a	0.0068(0.0034)	0.0069(0.0035)	0.0021(0.0011)	0.0021(0.0011)
b	0.0151(0.0305)	0.0111(0.0068)	0.0048(0.0063)	0.0037(0.0020)
c	0.0082(0.0049)	0.0074(0.0041)	0.0025(0.0015)	0.0022(0.0012)
d	0.0081(0.0053)	0.0067(0.0039)	0.0024(0.0016)	0.0019(0.0009)
e	0.0431(0.0474)	0.0226(0.0267)	0.0113(0.0079)	0.0051(0.0038)

For each of the above error distributions, we generate our data as follows:

$$Y_t = X_t^T\beta(U_t) + \varepsilon_t = X_{t,1}\beta_1(U_t) + X_{t,2}\beta_2(U_t) + \varepsilon_t, \quad t = 1, 2, \ldots, n \tag{10}$$

where $\beta_1(u) = 2u(1-u)$, $\beta_2(u) = \sin(2\pi u)$, and $U_t \overset{\text{i.i.d}}{\sim} U[0,1]$. For $X_t = (X_{t,1}, X_{t,2})^T$, we consider the following two cases:

(i) $\{X_t\}$ is generated by $X_t = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix} X_{t-1} + \eta_t$

(ii) $\{X_t\}$ is generated by $X_t = X_{t-1} + \eta_t$

where $\eta_t = (\eta_{t,1}, \eta_{t,2})^T \overset{\text{i.i.d}}{\sim} N((0,0)^T, \text{diag}(1,1))$.

Case (i) assumes that X_t is a mixture of stationary and non stationary covariates, while Case (ii) considers a non stationary X_t. In our simulation, we draw samples of sizes $n = 200, 400$ with $N = 100$ replications. Choose $h = 4s_U \times n^{-2/5}$ and $h_0 = 2s_{\tilde{\varepsilon}} \times n^{-1/5}$ for both Cases (i) and (ii), where s_U and $s_{\tilde{\varepsilon}}$ are the standard errors of $\{U_t\}$ and $\{\tilde{\varepsilon}_t\}$, respectively. Based on the theoretical result in Sun and Li (2011), the CV-selected bandwidth via the local linear method is $O_p(n^{-2/5})$ when X_t contains I(1) components. But instead of using this data-driven optimal bandwidth, we adopted this optimal order with a fixed constant $4s_U$. And for both cases, we use the same bandwidth in order to facilitate the comparison. The performance of estimates $\hat{\beta}(\cdot)$ is assessed via the averaged mean squared errors:

$$\text{AMSE} = \frac{1}{N}\sum_{j=1}^{N} \text{MSE}_\text{j}, \quad \text{MSE}_\text{j} = \frac{1}{100}\sum_{k=1}^{100}\sum_{i=1}^{2}\left[\hat{\beta}_i(u_k, j) - \beta_i(u_k)\right]^2$$

where $\{u_k, k = 1, 2, \ldots, 100\}$ are the grid points on the interval $[0, 1]$.

Tables 1 and 2 summarize the results from our simulation experiments. It can be seen that the AMSE and standard error of the adaptive estimator are smaller for all cases when the error is i.i.d and non-Gaussian. The efficiency gain is substantial even for moderate sample sizes. The corresponding results for normal errors are very similar for these two methods. It has also been shown in Figure 1 that the biases of these two estimators are almost the same, but the adaptive estimator owns narrower standard error curves than the least squares based ones. These findings are largely consistent with our theoretical results. In addition, when the

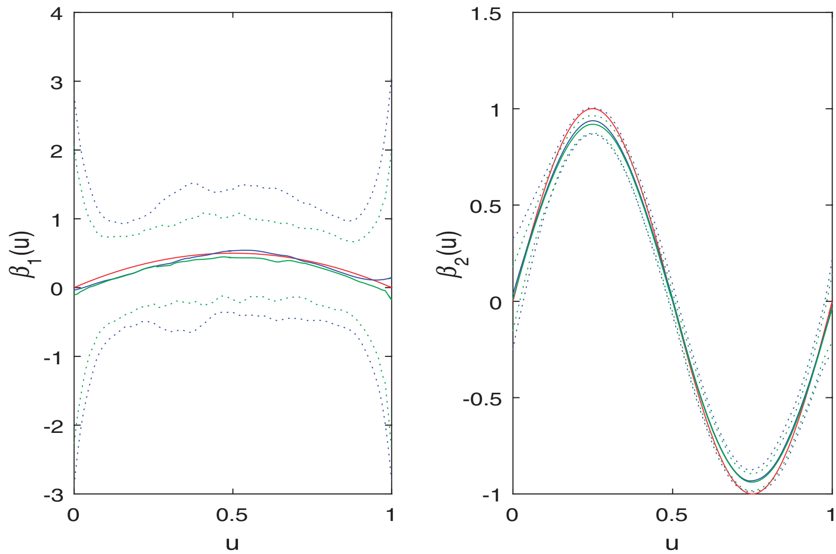

Figure 1. The averaged nonparametric estimators of $\beta_1(u)$ and $\beta_2(u)$ for Case (i) with $n = 200, N = 100$ and the error is the contaminated normal mixture (e). The red line is the true function. The blue line is LS estimator, while the green line is our adaptive estimator. The dashed lines are the corresponding standard error curves.

covariates vary from stationary to non stationary, our AMSE decreases. This largely illustrates that the estimator with non stationary covariates enjoys faster convergence rate than that with stationary covariates.

Next, in order to make comparisons, we conduct simulations for the cases that the error term ε_t is weakly dependent, and that η_t is weakly dependent. Specially, we consider the case that the error is AR(1) sequence: $\varepsilon_t = 0.5\varepsilon_{t-1} + s_t$ with $s_t \overset{\text{i.i.d}}{\sim} N(0, 0.75)$, and the case that the error is MA(1) sequence: $\varepsilon_t = s_t + \frac{\sqrt{3}}{3}s_{t-1}$ with $s_t \overset{\text{i.i.d}}{\sim} N(0, 0.75)$. Both cases ensure that the errors are weakly dependent and have a $N(0, 1)$ stationary distribution. Comparing to Tables 1 and 2, it is as expected that the week dependences of ε_t (AR(1) or MA(1) processes) don't have much impact on the performance of the estimates from Table 3. And the corresponding results for normal errors are also quiet similar for these two methods in the dependent cases. Besides, we consider the cases that η_t is weakly dependent, AR(1) or MA(1) processes, while assuming that the error is i.i.d with the previous five different distributions. It is shown that the weak dependences of η_t don't have much impact on the performance of the estimates for both cases either from Tables 4 and 5. The observed simulation results are largely consistent with our theoretical results.

5. Discussion

In this paper, we extended the adaptive estimation method to varying coefficient models with non stationary covariates. We derived the asymptotic properties of the proposed estimators. The proposed estimation procedure can be applied to situations when the errors departure from normal. It provides a more efficient estimator than the least squares based estimation procedure. Simulation studies confirmed our theoretical results.

Table 3. Performance of the estimates when the error term ε_t is weakly dependent.

AR(1): $\varepsilon_t = 0.5\varepsilon_{t-1} + s_t$ with $s_t \overset{\text{i.i.d}}{\sim} N(0, 0.75)$				
	$n = 200$		$n = 400$	
Case	LS	Adapt	LS	Adapt
Case (i)	0.0392(0.0203)	0.0390(0.0204)	0.0192(0.0094)	0.0192(0.0106)
Case (ii)	0.0077(0.0050)	0.0079(0.0050)	0.0026(0.0016)	0.0028(0.0028)
MA(1): $\varepsilon_t = s_t + \frac{\sqrt{3}}{3}s_{t-1}$ with $s_t \overset{\text{i.i.d}}{\sim} N(0, 0.75)$				
	$n = 200$		$n = 400$	
Case	LS	Adapt	LS	Adapt
Case (i)	0.0407(0.0212)	0.0395(0.0201)	0.0204(0.0094)	0.0200(0.0091)
Case (ii)	0.0069(0.0040)	0.0070(0.0041)	0.0020(0.0008)	0.0020(0.0008)

Table 4. Performance of the estimates when η_t is weakly dependent for Case (i).

AR(1): $\eta_t = 0.2\eta_{t-1} + r_t$ with $r_t = (r_{t,1}, r_{t,2})^T \overset{\text{i.i.d}}{\sim} N((0,0)^T, \text{diag}(0.96, 0.96))$				
	$n = 200$		$n = 400$	
ε	LS	Adapt	LS	Adapt
a	0.0314(0.0155)	0.0315(0.0161)	0.0159(0.0071)	0.0161(0.0087)
b	0.0745(0.0563)	0.0611(0.0345)	0.0432(0.0278)	0.0398(0.0297)
c	0.0385(0.0214)	0.0348(0.0213)	0.0198(0.0090)	0.0162(0.0082)
d	0.0369(0.0217)	0.0310(0.0212)	0.0197(0.0094)	0.0171(0.0273)
e	0.2888(0.2789)	0.1412(0.1533)	0.1563(0.1118)	0.0713(0.0839)
MA(1): $\eta_t = r_t + \frac{2}{\sqrt{96}}r_{t-1}$ with $r_t = (r_{t,1}, r_{t,2})^T \overset{\text{i.i.d}}{\sim} N((0,0)^T, \text{diag}(0.96, 0.96))$				
	$n = 200$		$n = 400$	
ε	LS	Adapt	LS	Adapt
a	0.0331(0.0162)	0.0329(0.0160)	0.0178(0.0100)	0.0176(0.0100)
b	0.0762(0.0565)	0.0663(0.0422)	0.0429(0.0282)	0.0366(0.0244)
c	0.0399(0.0198)	0.0364(0.0201)	0.0198(0.0079)	0.0159(0.0075)
d	0.0435(0.0241)	0.0369(0.0236)	0.0181(0.0080)	0.0134(0.0073)
e	0.2852(0.3169)	0.1223(0.0956)	0.1309(0.0694)	0.0636(0.0448)

However, there are still some interesting future research topics left. Firstly, it may be possible to generate our model to allow U_t is non stationary or both U_t and X_t are non stationary. Secondly, it would be important to investigate the bandwidth selection method. In addition, it will be interesting to consider the hypothesis tests for the functional coefficients with the adaptive estimators. Finally, the idea of the adaptive estimation might also be extended to many other semiparametric models with non stationary regressors, such as semi-varying coefficient models (Li et al. 2017), single-index and partially linear single-index integrated models (Dong, Gao, and Tjøstheim 2016), and varying coefficient partially non linear models (Zhou and Lin 2018).

6. Proofs

By the adaptive nonparametric regression result of Linton and Xiao (2007), we conjecture that the asymptotic results of $\hat{\Psi}(u_0)$ are the same whether the true density function $f(\cdot)$ is

Table 5. Performance of the estimates when η_t is weakly dependent for Case (ii).

AR(1): $\eta_t = 0.2\eta_{t-1} + r_t$ with $r_t = (r_{t,1}, r_{t,2})^T \overset{\text{i.i.d}}{\sim} N((0,0)^T, \text{diag}(0.96, 0.96))$				
	$n = 200$		$n = 400$	
ε	LS	Adapt	LS	Adapt
a	0.0057(0.0026)	0.0061(0.0033)	0.0017(0.0008)	0.0019(0.0020)
b	0.0115(0.0215)	0.0088(0.0051)	0.0036(0.0044)	0.0029(0.0015)
c	0.0065(0.0035)	0.0062(0.0030)	0.0020(0.0011)	0.0022(0.0035)
d	0.0066(0.0039)	0.0058(0.0030)	0.0019(0.0011)	0.0018(0.0021)
e	0.0309(0.0338)	0.0168(0.0195)	0.0081(0.0056)	0.0038(0.0026)
MA(1): $\eta_t = r_t + \frac{2}{\sqrt{96}} r_{t-1}$ with $r_t = (r_{t,1}, r_{t,2})^T \overset{\text{i.i.d}}{\sim} N((0,0)^T, \text{diag}(0.96, 0.96))$				
	$n = 200$		$n = 400$	
ε	LS	Adapt	LS	Adapt
a	0.0057(0.0030)	0.0059(0.0032)	0.0017(0.0006)	0.0017(0.0006)
b	0.0100(0.0068)	0.0098(0.0067)	0.0033(0.0022)	0.0032(0.0023)
c	0.0067(0.0040)	0.0065(0.0039)	0.0019(0.0009)	0.0021(0.0025)
d	0.0066(0.0034)	0.0059(0.0030)	0.0019(0.0010)	0.0031(0.0086)
e	0.0337(0.0398)	0.0237(0.0351)	0.0096(0.0080)	0.0047(0.0037)

used or not. The rigorous proofs are left for future work. Therefore, we will mainly show the consistence and asymptotic distribution of $\hat{\Psi}(u_0)$ assuming $f(\cdot)$ is known.

Adopting the same arguments given in Proof of Theorem 3.1 in Wang and Phillips (2009a), it holds that under a suitable probability space $\{\Omega, \mathcal{F}, P\}$, there exists an equivalent process $X^*_{t,2}$ of $X_{t,2}$ (i.e., $X^*_{t,2} \overset{d}{=} X_{t,2}, 1 \le t \le n$) such that

$$\sup_{0 \le r \le 1} \|X^*_{[nr],2} - B(r)\| = o_p(1) \tag{11}$$

by using the fact that $X_{[nr],2} \Rightarrow B(r)$ and the Skorohod–Dudley–Wichura representation theorem. Therefore, because of the consistency result in Theorem 3.1 and the asymptotic distribution result in Theorem 3.2 to be proved involves only weak convergence, without loss of generality we assume that $X_{t,2}$ satisfies Equation (11), and X_t, U_t and ε_t, $1 \le t \le n$ are defined on the same probability space $\{\Omega, \mathcal{F}, P\}$.

Proof of Theorem 3.1. Denote $K_t = K_h(U_t - u_0)$, $r(U_t, X_t) = X_t^T \beta(U_t) - X_t(u_0)^T \Psi(u_0) = X_t^T[\beta(U_t) - \beta(u_0) - \beta'(u_0)(U_t - u_0)]$ and $a_n = (\sqrt{nh})^{-1} + h^2$. We assume the objective function is

$$Q(\Theta) = \frac{1}{n} \sum_{t=1}^{n} K_t \rho(Y_t - X_t^T(u_0)\Theta) \tag{12}$$

To prove that with probability approaching 1, there exists a consistent local maximizer $\hat{\Psi}(u_0)$ of Equation (12) such that

$$\hat{\Psi}(u_0) - \Psi(u_0) = O_p(a_n)$$

it is sufficient to show that for any given $\gamma > 0$, there exists a large constant c such that

$$P\left(\sup_{\|\mu\|=c} Q(\Psi(u_0) + a_n\mu) < Q(\Psi(u_0))\right) \ge 1 - \gamma$$

where μ has the same dimension as $\Psi(u_0)$, see Proof of Theorem 1 in Fan and Li (2001). By using Taylor expansion, it follows that

$$
\begin{aligned}
&Q(\Psi(u_0)+a_n\mu)-Q(\Psi(u_0))\\
&=\frac{1}{n}\sum_{t=1}^{n}K_t\left[\rho(Y_t-X_t(u_0)^T(\Psi(u_0)+a_n\mu))-\rho(Y_t-X_t(u_0)^T\Psi(u_0))\right]\\
&=\frac{1}{n}\sum_{t=1}^{n}K_t\left[\rho(\varepsilon_t+r(U_t,X_t)-a_nX_t(u_0)^T\mu)-\rho(\varepsilon_t+r(U_t,X_t))\right]\\
&=-\frac{1}{n}\sum_{t=1}^{n}K_t\rho'(\varepsilon_t+r(U_t,X_t))a_nX_t(u_0)^T\mu\\
&\quad+\frac{1}{2n}\sum_{t=1}^{n}K_t\rho''(\varepsilon_t+r(U_t,X_t))a_n^2(X_t(u_0)^T\mu)^2\\
&\quad-\frac{1}{6n}\sum_{t=1}^{n}K_t\rho'''(Z_t)a_n^3(X_t(u_0)^T\mu)^3\\
&=:I_1+I_2+I_3
\end{aligned}
$$

where Z_t lies between $\varepsilon_t+r(U_t,X_t)-a_nX_t(u_0)^T\mu$ and $\varepsilon_t+r(U_t,X_t)$. Since

$$
\rho'(\varepsilon_t+r(U_t,X_t))=\rho'(\varepsilon_t)+\rho''(\varepsilon_t)r(U_t,X_t)+\frac{1}{2}\rho'''(\varepsilon_t)r^2(U_t,X_t)(1+o(1))
$$

then we have

$$
\begin{aligned}
I_1=&-\frac{a_n}{n}\sum_{t=1}^{n}K_t\rho'(\varepsilon_t)X_t(u_0)^T\mu-\frac{a_n}{n}\sum_{t=1}^{n}K_t\rho''(\varepsilon_t)r(U_t,X_t)X_t(u_0)^T\mu\\
&-\frac{a_n}{2n}\sum_{t=1}^{n}K_t\rho'''(\varepsilon_t)r^2(U_t,X_t)X_t(u_0)^T\mu(1+o(1))
\end{aligned}
$$

As for $j=0,1,2$,

$$
\begin{aligned}
\sqrt{\frac{h}{n}}\sum_{t=1}^{n}K_t\rho'(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)^jD_n^{-1}X_t&=\begin{pmatrix}\sqrt{\frac{h}{n}}\sum_{t=1}^{n}K_t\rho'(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)^jX_{t,1}\\ \sqrt{\frac{h}{n}}\sum_{t=1}^{n}K_t\rho'(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)^jX_{t,2}/\sqrt{n}\end{pmatrix}\\
&\xrightarrow{d}MN(E\{\rho'(\varepsilon)^2\}g(u_0)\nu_jS(u_0))=O_p(1)
\end{aligned}
$$

then

$$
\begin{aligned}
-\frac{a_n}{n}\sum_{t=1}^{n}K_t\rho'(\varepsilon_t)X_t(u_0)^T\mu&=-\mu^T\frac{a_n}{n}\sum_{t=1}^{n}K_t\rho'(\varepsilon_t)\begin{pmatrix}1\\ \frac{U_t-u_0}{h}\end{pmatrix}\otimes X_t\\
&=-\mu^TD_n\frac{a_n}{\sqrt{nh}}\cdot\sqrt{\frac{h}{n}}\sum_{t=1}^{n}K_t\rho'(\varepsilon_t)\begin{pmatrix}1\\ \frac{U_t-u_0}{h}\end{pmatrix}\\
&\quad\otimes D_n^{-1}X_t=O_p\left(\frac{ca_n}{\sqrt{h}}\right)
\end{aligned}
$$

Since $r(U_t, X_t) = X_t^T(\beta(U_t) - \beta(u_0) - \beta'(u_0)(U_t - u_0)) = \frac{1}{2}X_t^T\beta''(u_0)(U_t - u_0)^2(1 + o(1))$, then

$$
\begin{aligned}
&-\frac{a_n}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)r(U_t, X_t)X_t(u_0)^T\mu \\
&= -\frac{h^2}{2}\mu^T\frac{a_n}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)\begin{pmatrix}(\frac{U_t-u_0}{h})^2\\(\frac{U_t-u_0}{h})^3\end{pmatrix}\otimes(X_t^T\beta''(u_0)X_t)(1+o(1)) \\
&= -\frac{h^2}{2}\mu^T a_n\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\otimes D_n^2\left[\frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)\begin{pmatrix}(\frac{U_t-u_0}{h})^2\\(\frac{U_t-u_0}{h})^3\end{pmatrix}\otimes(D_n^{-1}X_t^T\beta''(u_0)D_n^{-1}X_t)\right] \\
&\quad (1+o(1)) \\
&= O_p(ca_n nh^2)
\end{aligned}
$$

Similarly, we have $-\frac{a_n}{2n}\sum_{t=1}^{n} K_t\rho'''(\varepsilon_t)r^2(U_t, X_t)X_t(u_0)^T\mu(1 + o(1)) = O_p(ca_n n^{3/2}h^4)$. Therefore, $I_1 = O_p(\frac{ca_n}{\sqrt{h}}) + O_p(ca_n nh^2) + O_p(ca_n n^{3/2}h^4) = O_p(cna_n^2)$. □

In addition,

$$
\begin{aligned}
I_2 &= \frac{1}{2n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t + r(U_t, X_t))a_n^2(X_t(u_0)^T\mu)^2 \\
&= \frac{a_n^2}{2n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)a_n^2(X_t(u_0)^T\mu)^2(1+o(1)) \\
&= \frac{a_n^2}{2}\mu^T\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\otimes D_n^2\left[\frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)\begin{pmatrix}1\\\frac{U_t-u_0}{h}\end{pmatrix}^{\otimes 2}\otimes(D_n^{-1}X_t)^{\otimes 2}\right]\mu \\
&= O_p(na_n^2\delta_1 g(u_0)\mu^T R(u_0)\mu)
\end{aligned}
$$

Similarly, $I_3 = -\frac{1}{6n}\sum_{t=1}^{n} K_t\rho'''(Z_t)a_n^3(X_t(u_0)^T\mu)^3 = O_p(na_n^3)$. As $\delta_1 < 0$, $\|\mu\| = c$ and $R(u_0)$ is a positive-definite matrix, we can choose c large enough such that I_2 dominates both I_1 and I_3 with probability at least $1-\gamma$. Thus $P\left(\sup_{\|\mu\|=c} Q(\Psi(u_0) + a_n\mu) < Q(\Psi(u_0))\right) \geq 1-\gamma$. The proof is completed.

Proof of Theorem 3.2. Now we provide the asymptotic distribution for such consistent estimator $\hat{\Psi}(u_0)$. The proof procedure largely follows from the arguments in Proof of Theorem 2.1 of Cai, Li, and Park (2009).

Since $\hat{\Psi}(u_0)$ maximize $Q(\Theta)$, then we have $Q'(\hat{\Psi}(u_0)) = 0$. By using Taylor expansion, we have

$$
0 = Q'(\hat{\Psi}(u_0)) = Q'(\Psi(u_0)) + Q''(\Psi(u_0))(\hat{\Psi}(u_0) - \Psi(u_0)) + \frac{1}{2}Q'''(\tilde{\Psi}(u_0))(\hat{\Psi}(u_0) - \Psi(u_0))^2
$$

where $\tilde{\Psi}(u_0)$ is a value between $\hat{\Psi}(u_0)$ and $\Psi(u_0)$. Then with Δ_n defined as $\Delta_n = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\otimes D_n$ and using the consistency result of $\hat{\Psi}(u_0)$ in Theorem 3.1, we have

$$\Delta_n(\hat{\Psi}(u_0) - \Psi(u_0)) = \left(\frac{1}{n}\Delta_n^{-1}Q''(\Psi(u_0))\Delta_n^{-1}\right)^{-1}\left(-\frac{1}{n}\Delta_n^{-1}Q'(\Psi(u_0))\right)(1+o_p(1))$$
$$=: S_n(u_0)^{-1}T_n(u_0)(1+o_p(1)) \tag{13}$$

For $S_n(u_0)$, we have

$$\begin{aligned}
S_n(u_0) &= \frac{1}{n}\Delta_n^{-1}Q''(\Psi(u_0))\Delta_n^{-1} \\
&= \frac{1}{n}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \otimes D_n^{-1}\left[\sum_{t=1}^{n} K_t\rho''(Y_t - X_t(u_0)^T\Psi(u_0))X_t(u_0)X_t(u_0)^T\right]\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \otimes D_n^{-1} \\
&= \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(Y_t - X_t(u_0)^T\Psi(u_0))\begin{pmatrix}1\\ \frac{U_t-u_0}{h}\end{pmatrix}^{\otimes 2} \otimes (D_n^{-1}X_t)^{\otimes 2} \\
&=: \begin{pmatrix}S_{n,0}(u_0) & S_{n,1}(u_0)\\ S_{n,1}(u_0) & S_{n,2}(u_0)\end{pmatrix}
\end{aligned}$$

where for $j = 0, 1, 2$,

$$\begin{aligned}
S_{n,j} &= \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(Y_t - X_t(u_0)^T\Psi(u_0))\left(\frac{U_t-u_0}{h}\right)^j (D_n^{-1}X_t)^{\otimes 2} \\
&=: \begin{pmatrix}F_{n,j,0}(u_0) & F_{n,j,1}(u_0)\\ F_{n,j,1}(u_0)^T & F_{n,j,2}(u_0)\end{pmatrix}
\end{aligned}$$

with

$$F_{n,j,0}(u_0) = \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(Y_t - X_t(u_0)^T\Psi(u_0))\left(\frac{U_t-u_0}{h}\right)^j X_{t,1}X_{t,1}^T$$
$$F_{n,j,1}(u_0) = \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(Y_t - X_t(u_0)^T\Psi(u_0))\left(\frac{U_t-u_0}{h}\right)^j X_{t,1}X_{t,2}^T/\sqrt{n}$$

and

$$F_{n,j,2}(u_0) = \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(Y_t - X_t(u_0)^T\Psi(u_0))\left(\frac{U_t-u_0}{h}\right)^j (X_{t,2}/\sqrt{n})^{\otimes 2}$$

We define

$$F^*_{n,j,\ell}(u_0) = \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)^j X_{t,1}^{\otimes \ell}$$

for $\ell = 1, 2$. Then by using the strong stationary property of $\{X_{t,1}^T, U_t, \varepsilon_t\}$ and simple calculation, we have

$$\begin{aligned}
E[F^*_{n,j,\ell}(u_0)] &= E\left[K_t\rho''(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)^j X_{t,1}^{\otimes \ell}\right] \\
&= E[\rho''(\varepsilon_t)]\cdot E\left[K_t\left(\frac{U_t-u_0}{h}\right)^j X_{t,1}^{\otimes \ell}\right] \\
&= \delta_1 g(u_0)\mu_j M_\ell(u_0) + o(1)
\end{aligned}$$

Adopting the similar arguments in Theorem 1 of Cai, Fan, and Yao (2000), one can show that $\mathrm{Var}[F^*_{n,j,\ell}(u_0)] = O(\frac{1}{nh}) = o(1)$. Hence, it holds that, for $\ell = 1, 2$,

$$F^*_{n,j,\ell}(u_0) = E[F^*_{n,j,\ell}(u_0)] + O_p(\sqrt{\mathrm{Var}[F^*_{n,j,\ell}(u_0)]}) = \delta_1 g(u_0)\mu_j M_\ell(u_0) + o_p(1) \tag{14}$$

Therefore, we have

$$\begin{aligned} F_{n,j,0}(u_0) &= \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t + r(U_t, X_t))\left(\frac{U_t - u_0}{h}\right)^j X_{t,1}X_{t,1}^T \\ &= \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)\left(\frac{U_t - u_0}{h}\right)^j X_{t,1}X_{t,1}^T(1 + o_p(1)) \\ &= F^*_{n,j,2}(u_0)(1 + o_p(1)) \\ &= \delta_1 g(u_0)\mu_j M_2(u_0) + o_p(1) \end{aligned}$$

As for $F_{n,j,1}(u_0)$, with the similar procedure to obtain (A.11) in Proof of Theorem 2.1 of Cai, Li, and Park (2009), we denote by $e_t = K_t\rho''(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)^j X_{t,1} - E\left[K_t\rho''(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)^j X_{t,1}\right]$, which gives that $\frac{1}{n}\sum_{t=1}^{n}(X_{t,2}^T/\sqrt{n})e_t = o_p(1)$. As a consequence,

$$\begin{aligned} F_{n,j,1}(u_0) &= \frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)\left(\frac{U_t - u_0}{h}\right)^j X_{t,1}X_{t,2}^T/\sqrt{n}(1 + o_p(1)) \\ &= \left(E\left[K_t\rho''(\varepsilon_t)\left(\frac{U_t - u_0}{h}\right)^j X_{t,1}\right]\cdot\frac{1}{n}\sum_{t=1}^{n} X_{t,2}^T/\sqrt{n} + \frac{1}{n}\sum_{t=1}^{n}(X_{t,2}^T/\sqrt{n})e_t\right)(1 + o_p(1)) \\ &= \delta_1 g(u_0)\mu_j M_1(u_0)\Gamma_1^T + o_p(1) \end{aligned}$$

Similarly,

$$F_{n,j,2}(u_0) = \delta_1 g(u_0)\mu_j\Gamma_2 + o_p(1)$$

Then, we have

$$S_{n,j}(u_0) = \delta_1 g(u_0)\mu_j S(u_0) + o_p(1) \tag{15}$$

By noting that $\mu_0 = 1$ and $\mu_1 = 0$, then it immediately follows that

$$S_n(u_0) = \delta_1 g(u_0)\begin{pmatrix}1 & 0\\ 0 & \mu_2\end{pmatrix}\otimes S(u_0) + o_p(1) = \delta_1 g(u_0)R(u_0) + o_p(1) \tag{16}$$

For $T_n(u_0)$, we can divide it into two parts:

$$\begin{aligned} T_n(u_0) &= -\frac{1}{n}\Delta_n^{-1}Q'(\Psi(u_0)) \\ &= \frac{1}{n}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\otimes D_n^{-1}\left[\sum_{t=1}^{n} K_t\rho'(Y_t - X_t(u_0)^T\Psi(u_0))X_t(u_0)\right] \\ &= \frac{1}{n}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t + r(U_t, X_t))\begin{pmatrix}1\\ \frac{U_t-u_0}{h}\end{pmatrix}\otimes(D_n^{-1}X_t) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t)\begin{pmatrix}1\\ \frac{U_t-u_0}{h}\end{pmatrix}\otimes(D_n^{-1}X_t)\\
&\quad+\left[\frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)r(U_t,X_t)\begin{pmatrix}1\\ \frac{U_t-u_0}{h}\end{pmatrix}\otimes(D_n^{-1}X_t)\right](1+o(1))\\
&=: T_{n,1}(u_0)+T_{n,2}(u_0)(1+o(1))
\end{aligned}
$$

Since $r(U_t,X_t) = X_t^T(\beta(U_t)-\beta(u_0)-\beta'(u_0)(U_t-u_0)) = \frac{h^2}{2}X_t^T\beta''(u_0)(\frac{U_t-u_0}{h})^2(1+o(1))$, then

$$
\begin{aligned}
\Delta_n^{-1}T_{n,2}(u_0) &= \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\otimes D_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)r(U_t,X_t)\begin{pmatrix}1\\ \frac{U_t-u_0}{h}\end{pmatrix}\otimes(D_n^{-1}X_t)\right]\\
&= \frac{h^2}{2}\left[\frac{1}{n}\sum_{t=1}^{n} K_t\rho''(\varepsilon_t)\begin{pmatrix}(\frac{U_t-u_0}{h})^2\\ (\frac{U_t-u_0}{h})^3\end{pmatrix}\otimes(D_n^{-1}X_t^T\beta''(u_0)D_n^{-1}X_t)\right](1+o(1))\\
&= \frac{h^2}{2}\delta_1 g(u_0)\begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}\otimes(S(u_0)\beta''(u_0))(1+o_p(1))
\end{aligned}
\tag{17}
$$

where the last equation follows from the proof of Equation (15). Moreover, for $T_{n,1}(u_0)$ which determines the asymptotic distribution, we have

$$
\begin{aligned}
\sqrt{nh}T_{n.1}(u_0) &= \sqrt{\frac{h}{n}}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t)\begin{pmatrix}1\\ \frac{U_t-u_0}{h}\end{pmatrix}\otimes(D_n^{-1}X_t)\\
&= \begin{pmatrix}\sqrt{\frac{h}{n}}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t)D_n^{-1}X_t\\ \sqrt{\frac{h}{n}}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)D_n^{-1}X_t\end{pmatrix}\\
&= \begin{pmatrix}\sqrt{\frac{h}{n}}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t)X_{t,1}\\ \sqrt{\frac{h}{n}}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t)X_{t,2}/\sqrt{n}\\ \sqrt{\frac{h}{n}}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)X_{t,1}\\ \sqrt{\frac{h}{n}}\sum_{t=1}^{n} K_t\rho'(\varepsilon_t)\left(\frac{U_t-u_0}{h}\right)X_{t,2}/\sqrt{n}\end{pmatrix}\xrightarrow{d} MN(\Sigma(u_0))
\end{aligned}
\tag{18}
$$

where $\Sigma(u_0) = E\{\rho'(\varepsilon)^2\}g(u_0)\Lambda(u_0)$. The proof of the last expression concerning the mixed normal limit results follows from the similar arguments for the proof of (A.22), (A.24) and (A.25) in Cai, Li, and Park (2009). The details are omitted here. Therefore, as a consequence of

$$
\begin{aligned}
&\sqrt{nh}\Delta_n\left[\hat{\Psi}(u_0)-\Psi(u_0)-\Delta_n^{-1}S_n(u_0)^{-1}T_{n,2}(u_0)(1+o_p(1))\right]\\
&\quad= \sqrt{nh}S_n(u_0)^{-1}T_{n,1}(u_0)(1+o_p(1))
\end{aligned}
$$

Theorem 3.2 follows immediately. □

Acknowledgments

The authors would like to thank the Editor, the Associate Editor and one anonymous referee for their constructive comments that greatly improved the paper.

Funding

This work is supported by the Swedish Research Council grant (Reg.No. 340-2013-5342).

References

Berkes, I., and L. Horváth. 2006. Convergence of integral functionals of stochastic processes. *Econometric Theory* 22 (2):304–22.
Billingsley, P. 1999. *Convergence of probability measures*. 2nd ed. New York: Wiley.
Cai, Z., J. Fan, and Q. Yao. 2000. Functional-coefficient regression models for non linear time series. *Journal of the American Statistical Association* 95 (451):941–56.
Cai, Z., Q. Li, and J. Y. Park. 2009. Functional-coefficient models for non stationary time series data. *Journal of Econometrics* 148 (2):101–13.
Chan, N., and Q. Wang. 2015. Non linear regressions with non stationary time series. *Journal of Econometrics* 185 (1):182–95.
Chen, H., Y. Fang, and Y. Li. 2015. Estimation and inference for varying-coefficient models with non stationary regressors using penalized splines. *Econometric Theory* 31 (4):753–77.
Chen, J., J. Gao, and D. Li. 2012. Estimation in semi-parametric regression with non stationary regressors. *Bernoulli* 18 (2):678–702.
Chen, Y., Q. Wang, and W. Yao. 2015. Adaptive estimation for varying coefficient models. *Journal of Multivariate Analysis* 137:17–31.
Dong, C., J. Gao, and D. Tjøstheim. 2016. Estimation for single-index and partially linear single-index integrated models. *Annals of Statistics* 44 (1):425–53.
Fan, J., and I. Gijbels. 1996. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66* (Vol. 66). London: CRC Press.
Fan, J., and R. Li. 2001. Variable selection via non concave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 96 (456):1348–60.
Gao, J., and P. C. Phillips. 2013. Functional coefficient non stationary regression. Cowles Foundation Discussion paper 1911. Cowles Foundation for Research in Economics, Yale University. http://cowles.econ.yale.edu/P/cd/d19a/d1911.pdf.
Gu, J., and Z. Liang. 2014. Testing cointegration relationship in a semiparametric varying coefficient model. *Journal of Econometrics* 178:57–70.
Hamilton, J. D. 1994. *Time series analysis* (Vol. 2). Princeton: Princeton university press Princeton.
Hastie, T., and R. Tibshirani. 1993. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)* 757–96.
Jin, S., L. Su, and Z. Xiao. 2015. Adaptive nonparametric regression with conditional heteroskedasticity. *Econometric Theory* 31 (6):1153–91.
Juhl, T., and Z. Xiao. 2005. Partially linear models with unit roots. *Econometric Theory* 21 (5):877–906.
Karlsen, H. A., T. Myklebust, and D. Tjøstheim. 2007. Nonparametric estimation in a non linear cointegration type model. *Annals of Statistics* 35 (1):252–99.
Karlsen, H. A., and D. Tjostheim. 2001. Nonparametric estimation in null recurrent time series. *Annals of Statistics* 29:372–416.
Li, K., D. Li, Z. Liang, and C. Hsiao. 2017. Estimation of semi-varying coefficient models with non stationary regressors. *Econometric Reviews* 36 (1-3):354–69.
Liang, H., P. C. Phillips, H. Wang, and Q. Wang. 2016. Weak convergence to stochastic integrals for econometric applications. *Econometric Theory* 32 (6):1349–75.
Liang, Z., Z. Lin, and C. Hsiao. 2015. Local linear estimation of a nonparametric cointegration model. *Econometric Reviews* 34 (6–10):882–906.

Linton, O., and Z. Xiao. 2007. A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory* 23 (3):371–413.

Merlevède, F., M. Peligrad, and S. Utev. 2006. Recent advances in invariance principles for stationary sequences. *Probability Surveys* 3: 1–36.

Sun, Y., Z. Cai, and Q. Li. 2013. Semiparametric functional coefficient models with integrated covariates. *Econometric Theory* 29 (3):659–72.

Sun, Y., and Q. Li. 2011. Data-driven bandwidth selection for non stationary semiparametric models. *Journal of Business & Economic Statistics* 29 (4):541–51.

Wang, Q. 2014. Martingale limit theorem revisited and non linear cointegrating regression. Singapore: *Econometric Theory* 30 (3):509–35.

Wang, Q. 2015. *Limit theorems for non linear cointegrating regression*. World Scientific.

Wang, Q., and P. C. Phillips. 2009a. Asymptotic theory for local time density estimation and nonparametric cointegrating regression. *Econometric Theory* 25 (3):710–38.

Wang, Q., and P. C. Phillips. 2009b. Structural nonparametric cointegrating regression. *Econometrica* 77 (6):1901–48.

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

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

Zhou, Z., and Z. Lin. 2018. Varying coefficient partially non linear models with non stationary regressors. *Journal of Statistical Planning and Inference* 194:47–64.