Robust Nonlinear Regression Estimation in Null Recurrent Time Series

Francesco Bravo[*], Degui Li[†], Dag Tjøstheim[‡]

Version: March 23, 2020

Abstract

In this article, we study parametric robust estimation in nonlinear regression models with regressors generated by a class of non-stationary and null recurrent Markov process. The nonlinear regression functions can be either integrable or asymptotically homogeneous, covering many commonly-used functional forms in parametric nonlinear regression. Under regularity conditions, we derive both the consistency and limit distribution results for the developed general robust estimators (including the nonlinear least squares, least absolute deviation and Huber's M-estimators). The convergence rates of the estimation depend on not only the functional form of nonlinear regression, but also on the recurrence rate of the Markov process. Some Monte-Carlo simulation studies are conducted to examine the numerical performance of the proposed estimators and verify the established asymptotic properties in finite samples. Finally two empirical applications illustrate the usefulness of the proposed robust estimation method.

Keywords: Asymptotically homogeneous functions, β-null recurrence, Integrable functions, Nonlinear regression, Outliers, Robust estimation.

[*]Department of Economics and Related Studies, University of York, UK.
[†]Department of Mathematics, University of York, UK. Corresponding author: degui.li@york.ac.uk
[‡]Department of Mathematics, University of Bergen, Norway.

1 Introduction

Suppose that $\{Y_t\}$ and $\{X_t\}$ are two sequences of time series processes which are either stationary or non-stationary. A flexible framework to study the relationship between Y_t and X_t is via the following nonlinear regression model:

$$Y_t = m(X_t, \boldsymbol{\gamma}_0) + W_t, \quad t = 1, \cdots, n, \tag{1.1}$$

where $\boldsymbol{\gamma}_0 = (\gamma_{01}, \cdots, \gamma_{0d})^\tau$ is a column vector of unknown parameters, $m(\cdot, \cdot)$ is a pre-specified nonlinear regression function, $\{W_t\}$ is a sequence of model errors, and n is the sample size. Without loss of generality, we assume that $\boldsymbol{\gamma}_0$ lies in the interior of $\boldsymbol{\Upsilon}$, a convex and compact parameter set. Such a parametric nonlinear regression model has been extensively studied in the statistical and econometric literature when the observations are either stationary (c.f., Jennrich, 1969; Malinvaud, 1970; Wu, 1981; Severini and Wong, 1992; Lai, 1994; Skouras, 2000) or non-stationary (c.f., Park and Phillips, 2001; Chan and Wang, 2015; Li, Tjøstheim and Gao, 2016; Tjøstheim, 2020). The main interest of this paper lies on estimation of the unknown parameter vector $\boldsymbol{\gamma}_0$. Most of the aforementioned literature constructs consistent parameter estimation via the so-called nonlinear least squares approach, which has some nice asymptotic properties and performs well in practice when the model errors W_t follow (or approximate) the centred normal distribution. However, it is well-known that the least squares based estimation is not robust and performs poorly when the observations contain outliers or the model error distribution is heavy-tailed. The latter features are particularly common for the data collected from economics and finance. Hence, it becomes imperative to develop a robust methodology to estimate model (1.1).

When the observations are stationary, there has been extensive literature on studying various parametric robust estimation approaches and developing their asymptotic properties (c.f., Huber, 1964, 1981; Bai, Rao and Wu, 1992; He and Shao, 1996; Knight, 1998; Sinha, Field and Smith, 2003; Koenker, 2005; Zou and Yuan, 2008). Extension of robust estimation to parametric non-stationary linear regression can be found in Knight (1991), Phillips (1995) and Maddala and Kim (1998). In this paper, we consider a general setting where the observations are collected from a null recurrent Markov process which includes the classic random walk process. In recent years, there have been increasing interests on parametric and nonparametric estimation of null recurrent Markov chains (c.f., Karlsen and Tjøstheim, 2001; Karlsen, Myklebust and Tjøstheim, 2007, 2010; Schienle, 2011; Chen, Gao and Li, 2012; Chan and Wang, 2015; Gao *et al*, 2015; Li, Tjøstheim and Gao, 2016; Tjøstheim, 2020). Meanwhile, nonparametric extensions of the classic M-type estimation and composite quantile estimation for recurrent time series have been studied in Lin, Li and Chen (2009) and Li and Li (2016), respectively. However, as we are aware of, there is virtually no work on robust estimation for parametric nonlinear models like (1.1) with X_t being nonstationary. This

paper aims to fill this gap.

In the robust nonparametric estimation, both Lin, Li and Chen (2009) and Li and Li (2016) use the kernel-based approach and assume the kernel function is bounded, making it feasible to use some existing limit theory developed for bounded functions of Harris recurrent Markov processes. However, it is often the case that the nonlinear function $m(\cdot,\cdot)$ or its derivatives may be unbounded, leading to difficulties in developing the relevant asymptotic properties when we use classic parametric robust estimation methods such as the least absolute deviation and Huber's M-estimation. To address this issue, in this paper we introduce a trimmed version of robust estimation, disregarding the observations when $|X_t|$ exceeds ζ_n which is a tuning parameter diverging to infinity at certain rate of n. Through an appropriate choice of ζ_n, we remove some aberrant observations in the robust estimation procedure and keep sample information loss controlled.

The seminal paper by Park and Phillips (2001) shows that, when the regressor X_t follows a non-stationary unit root process, the asymptotic behavior for the nonlinear least squares estimator of γ_0 relies on the property of the nonlinear regression function $m(\cdot,\cdot)$ and its derivatives, i.e., the convergence rate is slower than the well-known root-n rate in stationary parametric regression estimation when the regression function is integrable, whereas a super-fast convergence rate can be achieved when the regression function is asymptotically homogeneous. Such a dichotomy in the asymptotic theory is also studied in Chan and Wang (2015) and Li, Tjøstheim and Gao (2016). In this paper, we make a further extension of the theory and develop both the consistency and limit distribution results for the general robust estimation when X_t is β-null recurrent. The restriction on the convex loss function in the developed estimation procedure is mild, covering a few commonly-used parametric estimation methods. As an application of our main theoretical results, we derive the weak consistency and limit distribution results for the so-called modified nonlinear least squares estimation similar to those in Li, Tjøstheim and Gao (2016), but the conditions imposed in the present paper are much weaker. In addition, we also obtain the asymptotic results for the modified least absolute deviation and Huber's M-estimators. As in Li, Tjøstheim and Gao (2016), the convergence rates of the proposed estimators in this paper depend on not only the functional form of nonlinear regression, but also the recurrence rate of the Markov process. Some Monte-Carlo simulation studies are given to examine the numerical performance of the proposed parametric estimators in finite samples, and two real data examples illustrate their applicability.

The rest of the paper is organised as follows. Section 2 introduces the robust estimation methodology and some basic definitions and results for null recurrent Markov processes. Section 3 gives the main asymptotic theory and discusses some possible extensions. Section 4 provides the simulation studies and Section 5 contains the empirical applications. Section 6 concludes the paper. Proofs of the main theoretical results are given in Appendix A. Appendices B and C, which contain, respectively, proofs of some relevant technical lemmas and some additional simulation

studies, are available in a supplemental document.

2 Estimation methodology and recurrent Markov processes

In this section we first introduce the robust method to estimate the parameter vector $\boldsymbol{\gamma}_0$ in model (1.1) and then review some basic results for the recurrent Markov processes which will be used in the following sections.

2.1 Robust estimation methodology

Suppose that model (1.1) holds and we have the observations (Y_t, X_t), $t = 1, \cdots, n$. To estimate the unknown parameter vector $\boldsymbol{\gamma}_0$, we may define the following objective function:

$$\mathcal{L}_n(\boldsymbol{\gamma}) = \sum_{t=1}^{n} \rho\big(Y_t - m(X_t, \boldsymbol{\gamma})\big), \tag{2.1}$$

where $\rho(\cdot)$ is a convex loss function satisfying some conditions given in Section 3. Some commonly-used loss functions include $\rho(u) = u^2$ (corresponding to the nonlinear least squares), $\rho(u) = |u|$ (corresponding to the nonlinear least absolute deviation), and Huber (1964)'s loss function:

$$\rho_\delta(u) = \begin{cases} u^2/2, & |u| \leqslant \delta, \\ \delta\left(|u| - \delta/2\right), & |u| > \delta, \end{cases}$$

where δ is a pre-specified number. The estimator of $\boldsymbol{\gamma}_0$ can be obtained by minimising the loss function $\mathcal{L}_n(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ and we denote the resulting estimate by $\overline{\boldsymbol{\gamma}}_n$.

When the observations are stationary, it can be proved that $\overline{\boldsymbol{\gamma}}_n$ has some nice asymptotic properties such as the root-n consistency and asymptotic normality under some mild conditions. However, when X_t are non-stationary and generated from a null recurrent Markov chain, the study of limit theory for the parametric estimation of $\boldsymbol{\gamma}_0$ becomes much more involved. For the nonlinear least squares estimator with $\rho(u) = u^2$, as shown by Li, Tjøstheim and Gao (2016), the asymptotic behavior of $\overline{\boldsymbol{\gamma}}_n$ relies on the functional properties of the regression function $m(\cdot, \cdot)$ and its derivatives, which may be not bounded and integrable. In the latter case, we cannot directly make use of some classic limit results for the recurrent Markov process (such as the ergodic theorem for bounded functions of the recurrent Markov processes), making it very challenging to develop sensible asymptotic theory for the parametric estimation.

In order to address the above concern, we need to modify the objective function $\mathcal{L}_n(\boldsymbol{\gamma})$ using

the truncation technique as in Ling (2007) and Li, Tjøstheim and Gao (2016). Let ζ_n be a positive tuning parameter satisfying $\zeta_n \to \infty$ as $n \to \infty$. Some conditions on ζ_n will be given in Section 3 below. Consider a new objective function:

$$\mathcal{L}_{n,\zeta_n}(\boldsymbol{\gamma}) = \sum_{t=1}^{n} \rho\big(Y_t - m(X_t, \boldsymbol{\gamma})\big) I\left(|X_t| \leqslant \zeta_n\right), \tag{2.2}$$

where $I(\cdot)$ is an indicator function. With such a truncation technique, we can remove the influence of some aberrant observations of X_t, some of which may not be automatically deleted through the use of the outlier-resistant loss function $\rho(\cdot)$, and could therefore affect the estimation of $\boldsymbol{\gamma}_0$ (see Section 4 for some Monte-Carlo evidence). A similar truncation idea can also be found in the recent papers by Chen and Christensen (2015) and Hansen (2015), which consider series estimation in general nonparametric regression setting. The modified robust estimator of $\boldsymbol{\gamma}_0$ is defined as

$$\widetilde{\boldsymbol{\gamma}}_n = \arg\min_{\boldsymbol{\gamma} \in \boldsymbol{\Upsilon}} \mathcal{L}_{n,\zeta_n}(\boldsymbol{\gamma}). \tag{2.3}$$

If $\rho(u) = u^2$, $\widetilde{\boldsymbol{\gamma}}_n$ becomes the modified nonlinear least squares estimator (Li, Tjøstheim and Gao, 2016) denoted by $\widetilde{\boldsymbol{\gamma}}_{LS}$. If $\rho(u)$ is chosen as $|u|$ or Huber's loss function $\rho_\delta(u)$, $\widetilde{\boldsymbol{\gamma}}_n$ would be the nonlinear least absolute deviation estimator denoted by $\widetilde{\boldsymbol{\gamma}}_{LAD}$ or Huber's M-estimator denoted by $\widetilde{\boldsymbol{\gamma}}_M$, both of which have been studied by Phillips (1995) for the parametric linear regression setting with X_t generated by a unit root $I(1)$ process. Section 3 below will give the asymptotic properties for $\widetilde{\boldsymbol{\gamma}}_n$, $\widetilde{\boldsymbol{\gamma}}_{LS}$, $\widetilde{\boldsymbol{\gamma}}_{LAD}$ and $\widetilde{\boldsymbol{\gamma}}_M$.

2.2 Basic Markov theory

To make our paper self-contained, we next review some basic definitions and results for a recurrent Markov process. The notation we will use is similar to that in some existing papers (c.f. Karlsen and Tjøstheim, 2001; Karlsen, Myklebust and Tjøstheim, 2007; Li, Tjøstheim and Gao, 2016), and further details are available in the two books: Nummelin (1984) and Meyn and Tweedie (2009).

Throughout the paper, we let $\{X_t, t \geqslant 0\}$ be a ϕ-irreducible Markov chain on the state space $(\mathbb{E}, \mathcal{E})$ with transition probability P, meaning that for any set $\mathbb{A} \in \mathcal{E}^+$, we have $\sum_{t=1}^{\infty} \mathsf{P}^t(x, \mathbb{A}) > 0$ for $x \in \mathbb{E}$, where $\mathcal{E}^+$ denotes the class of nonnegative measurable functions with ϕ-positive support and a set $\mathbb{A} \in \mathcal{E}^+$ if the indicator function $I(\mathbb{A}) \in \mathcal{E}^+$. The transition probability P is assumed to satisfy the following minorization inequality:

$$\mathsf{P} \geqslant s \otimes \nu, \tag{2.4}$$

where $s(\cdot) \in \mathcal{E}^+$ is a small function, $\nu(\cdot)$ is a small measure satisfying $\nu(\mathbb{E}) = 1$ and $s \otimes \nu(x, \mathbb{A}) = s(x)\nu(\mathbb{A})$ for any $(x, \mathbb{A}) \in (\mathbb{E}, \mathcal{E})$. The definitions of the small function and small measure can be found in Karlsen and Tjøstheim (2001), see also Nummelin (1984) and Meyn and Tweedie (2009). Example 3.1 in Karlsen and Tjøstheim (2001) shows that the inequality (2.4) holds for the nonlinear AR(1) process defined by

$$X_t = g(X_{t-1}) + x_t, \tag{2.5}$$

where the nonlinear autoregressive function $g(\cdot)$ is bounded on any compact set and $\{x_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean and with a density function $f(\cdot)$ on $\mathbb{E} = \mathbb{R}$. Karlsen and Tjøstheim (2001) further assume that $\inf_{x \in \mathbb{A}} f(x)$ is strictly positive for any compact set $\mathbb{A}$ and obtain (2.4) by choosing the small function and the small measure as

$$s(x) = \chi(f) \cdot I(x \in \mathbb{A}) \text{ and } \nu(dy) = [\chi(f)]^{-1} \cdot f_0(y)dy,$$

where $\chi(f) = \int_{\mathbb{R}} f_0(z)dz$ and $f_0(z) = \inf_{x \in \mathbb{A}} f(z - g(x))$. In particular, letting $g(x) = x$ in (2.5), we obtain the minorisation inequality for the random walk process.

In this paper, we further assume that the ϕ-irreducible Markov chain is Harris recurrent, i.e., for any set $\mathbb{A} \in \mathcal{E}^+$ and given $X_0 = x$ for all $x \in \mathbb{E}$, the Markov process $\{X_t\}$ returns to the set $\mathbb{A}$ infinitely often with probability one. The Harris recurrence is a key assumption to make the kernel-based local method applicable to the nonparametric estimation of recurrent time series (c.f., Karlsen and Tjøstheim, 2001; Karlsen, Myklebust and Tjøstheim, 2007) and it also allows one to construct a split chain (Nummelin, 1984). The split chain technique can lead to decomposition of the partial sum of functions of the Harris recurrent process into blocks of *i.i.d.* parts (which are asymptotically dominant) and two asymptotically negligible remaining parts. As in Karlsen and Tjøstheim (2001), we let τ_k be the recurrence times of the split chain, and define $T(n)$ as

$$T(n) = \max_k \{k : \tau_k \leqslant n\} \vee 0.$$

For the process $\{G(X_t) : t \geqslant 0\}$ with $G(\cdot)$ being a real function, we define

$$U_k(G) = \begin{cases} \sum_{t=0}^{\tau_0} G(X_t), & k = 0, \\ \sum_{t=\tau_{k-1}+1}^{\tau_k} G(X_t), & k = 1, \cdots T(n), \\ \sum_{t=\tau_{T(n)}+1}^{n} G(X_t), & k = T(n) + 1, \end{cases}$$

and consequently have

$$S_n(G) = \sum_{t=0}^{n} G(X_t) = U_0(G) + \sum_{k=1}^{T(n)} U_k(G) + U_{T(n)+1}(G). \tag{2.6}$$

From Nummelin (1984), we know that $\{U_k(G), k \geqslant 1\}$ is a sequence of *i.i.d.* random variables, and the first term $U_0(G)$ and the third term $U_{T(n)+1}(G)$ on the right side of (2.6) can be shown to be bounded with probability approaching one.

The Harris recurrence provides a very general framework for time series analysis and includes both the positive recurrent (stationary) and null recurrent (non-stationary) processes. The main interest of this paper lies on the null recurrent case as the robust estimation in the stationary and positive recurrent time series has been systematically studied in the literature. To obtain a specific rate of the random number $T(n)$ for the null recurrent process, we further impose some restrictions on the tail behavior of the distribution of the recurrence times of the Markov process. *A Markov chain $\{X_t\}$ is β-null recurrent if there exist a small nonnegative function $f(\cdot)$, an initial measure λ, a constant $\beta \in (0,1)$, and a slowly varying function $L_f(\cdot)$ (which may depend on the function f) such that*

$$E_\lambda \left[\sum_{t=1}^{n} f(X_t) \right] \sim \frac{1}{\Gamma(1+\beta)} n^\beta L_f(n), \tag{2.7}$$

where E_λ denotes the expectation with initial distribution λ and $\Gamma(\cdot)$ is the Gamma function.

Choosing $f(x) = I(x \in \mathbb{C})$ for a small set $\mathbb{C} \in \mathcal{E}^+$ and using (2.7), we find that assuming β-null recurrence restricts the number of X_t returning to the set $\mathbb{C}$ in the time interval $[0, n]$ to be a regularly varying function. In practice, the set $\mathbb{C}$ can be chosen as a compact set with ϕ-positive support, which is a small set under some mild conditions. Let $\pi_s(\cdot)$ be an invariant measure of the Markov chain $\{X_t\}$ which is defined as

$$\pi_s = \sum_{l=0}^{\infty} \nu(P - s \otimes \nu)^l, \tag{2.8}$$

where $s(\cdot)$ and $\nu(\cdot)$ are defined as in the inequality (2.4). For all small functions f, we can find a slowly varying function $L_s(\cdot)$ such that (2.7) holds for the β-null recurrent Markov chain with $L_f(\cdot) = \pi_s(f) L_s(\cdot)$, where $\pi_s(f) = \int f(x) \pi_s(dx)$. From Lemma 3.4 in Karlsen and Tjøstheim (2001), we have

$$n^{\beta-\eta} \ll T(n) \ll n^{\beta+\eta} \tag{2.9}$$

for any $\eta > 0$ almost surely (*a.s.*). Furthermore, Theorem 3.2 in Karlsen and Tjøstheim (2001) shows that $T(n)$ has the following asymptotic distribution:

$$\frac{T(n)}{n^\beta L_s(n)} \xrightarrow{d} M_\beta(1), \tag{2.10}$$

where $\{M_\beta(t), t \geqslant 0\}$ is the Mittag-Leffler process with parameter β (c.f., Kasahara, 1984).

Examples of β-null recurrent Markov processes can be found in Schienle (2011), Myklebust,

Karlsen and Tjøstheim (2012) and Li, Tjøstheim and Gao (2016). The classic random walk process is β-null recurrent with $\beta = 1/2$ and with $\pi_s(\cdot)$ as the Lebesgue measure. As $\beta < 1$, we have $T(n) < n$ *a.s.*, leading to slower rates of convergence for the nonparametric kernel estimation in null recurrent time series (c.f., Karlsen, Myklebust and Tjøstheim, 2007; Gao *et al*, 2015). However, Li, Tjøstheim and Gao (2016) show that this is not the case for the parametric estimation. In particular, they prove that the rate of convergence for the nonlinear least squares estimator $\widetilde{\gamma}_{LS}$ in the null recurrent case can be faster than that for the stationary time series case when the regression function $m(\cdot,\cdot)$ in model (1.1) is asymptotically homogeneous. In Section 3 below, we will show that similar properties hold for the more general M-estimation defined in (2.3).

3 Large sample theory

In this section, we first give the main limit theorems for the robust estimation $\widetilde{\gamma}_n$ when the regression function $m(\cdot,\cdot)$ is either integrable on the set Υ (in Section 3.1) or asymptotically homogeneous on Υ (in Section 3.2), and then derive the results for $\widetilde{\gamma}_{LS}$, $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_M$. In Section 3.3, we discuss how to relax the independence restriction between $\{X_t\}$ and $\{W_t\}$, and a possible extension of the model structure to semiparametric single-index models.

3.1 The case of integrable regression function

A real function $g(x,\gamma)$ is said to be integrable on the set Υ if for each $\gamma \in \Upsilon$, $g(\cdot,\gamma)$ is π_s-integrable in the sense that

$$\int_{\mathbb{R}} |g(x,\gamma)|\,\pi_s(x)\,dx < \infty,$$

and there exist a neighborhood $\mathbb{N}_\gamma$ and a function $B : \mathbb{R} \to \mathbb{R}$, positive, bounded and π_s-integrable such that $|g(x,\gamma_1) - g(x,\gamma)| \leqslant \|\gamma_1 - \gamma\| B(x)$ for any $\gamma_1 \in \mathbb{N}_\gamma$, where $\|\cdot\|$ denotes the Euclidean norm. Extension of the above definition to the case of a vector (or matrix) of integrable functions on Υ is straightforward. Such a definition is similar to the definition of I-regular functions in Park and Phillips (2001). Examples of integrable function include $g(x,\gamma) = \gamma g(x)$ with $g(x)$ being a bounded and π_s-integrable function, and $g(x,\gamma) = \exp\{-\gamma x^2\}$ with $\gamma \in \mathbb{R}^+$ (the set of positive real numbers). Throughout this subsection, we assume that the nonlinear regression function $m(x,\boldsymbol{\gamma})$ and its first and second partial derivatives with respect to $\boldsymbol{\gamma}$:

$$\dot{m}(x,\boldsymbol{\gamma}) = \left(\frac{\partial m(x,\boldsymbol{\gamma})}{\partial \gamma_j}\right)_{d\times 1}, \quad \ddot{m}(x,\boldsymbol{\gamma}) = \left(\frac{\partial^2 m(x,\boldsymbol{\gamma})}{\partial \gamma_i \partial \gamma_j}\right)_{d\times d},$$

are integrable on $\boldsymbol{\Upsilon}$. The following regularity conditions are used to derive the limit theorems for the robust estimation $\widetilde{\boldsymbol{\gamma}}_n$.

Assumption 1. *Let $\{W_t\}$ be a sequence of i.i.d. random variables and $\{X_t\}$ be a β-null recurrent Markov process. In addition, W_t is independent of $\{X_s, s \leqslant t\}$.*

Assumption 2. *(i) Let $\psi(\cdot)$ be any choice of the sub-gradient of the convex loss function $\rho(\cdot)$. The function $\psi(\cdot)$ is either Lipschitz continuous or bounded, satisfying*

$$\mathsf{E}\left[\psi(W_t + \varepsilon)\right] = \phi_1 \varepsilon + o(|\varepsilon|), \quad \varepsilon \to 0, \tag{3.1}$$

where $\phi_1 > 0$ is a constant.

(ii) There exists a positive constant ϕ_2 such that $\phi_2 = \mathsf{E}\left[\psi^2(W_t)\right]$. For any positive constant M, there exists a function $Q(\cdot)$ such that for $|w| < M$

$$\mathsf{E}\left\{\left[\psi(W_t + w) - \psi(W_t)\right]^2\right\} \leqslant Q(w), \tag{3.2}$$

where $Q(w)$ is bounded over $|w| < M$ and continuous at $w = 0$, and $Q(0) = 0$.

(iii) The function $\psi(\cdot)$ satisfies that

$$\mathsf{E}\left[|\psi(W_t + w_1) - \psi(W_t + w_2)|\right] = O(|w_1 - w_2|), \tag{3.3}$$

uniformly over w_1 and w_2 in a small neighborhood of 0.

Assumption 3. *(i) The regression function $m(x, \boldsymbol{\gamma})$ and its partial derivatives $\dot{m}(x, \boldsymbol{\gamma})$ and $\ddot{m}(x, \boldsymbol{\gamma})$ are integrable on the set $\boldsymbol{\Upsilon}$.*

(ii) For any $\boldsymbol{\gamma} \in \boldsymbol{\Upsilon}$ and $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}_0$, $\int_{\mathbb{R}} \left[m(x, \boldsymbol{\gamma}) - m(x, \boldsymbol{\gamma}_0)\right]^2 \pi_s(dx) > 0$. Furthermore, the $d \times d$ matrix $\boldsymbol{\Delta}_m(\boldsymbol{\gamma}_0) := \int_{\mathbb{R}} \dot{m}(x, \boldsymbol{\gamma}_0) \dot{m}^\intercal(x, \boldsymbol{\gamma}_0) \pi_s(dx)$ is positive definite.

Remark 3.1. As in Karlsen, Myklebust and Tjøstheim (2007), we may replace the i.i.d. condition on W_t by some stationary and mixing dependence condition at the cost of more lengthy arguments in mathematical proofs. The independence assumption between W_t and $\{X_s, s \leqslant t\}$ facilitates the construction of martingale differences in the proofs of some key technical lemmas (say, Lemma A.1 in Appendix A) for the estimation consistency. However, to derive the limit distribution theory, we need to strengthen the condition to mutual independence between $\{W_t\}$ and $\{X_t\}$ (e.g., Karlsen, Myklebust and Tjøstheim, 2007; Gao *et al*, 2015; Li, Tjøstheim and Gao, 2016), which ensures that the β-null recurrence can be retained for the compound Markov process $\{(X_t, W_t)\}$. Section 3.3 below will discuss how to relax such an independence restriction. Assumption 2 imposes some

mild conditions on the loss function and its derivative, some of which are similar to those in Bai, Rao and Wu (1992). Assumption 3(i) restricts the nonlinear regression function and its derivatives to be integrable and Assumption 3(ii) guarantees that the parameters are identifiable (e.g., Park and Phillips, 2001; Chan and Wang, 2015; Li, Tjøstheim and Gao, 2016).

We next state the asymptotic theorem for the robust estimation $\widetilde{\gamma}_n$ when the nonlinear regression function and its derivatives are integrable.

Theorem 3.1. *Suppose that Assumptions 1–3 are satisfied and that $\zeta_n \to \infty$ as $n \to \infty$.*

(i) The robust estimation $\widetilde{\gamma}_n$ defined in (2.3) is weakly consistent.

(ii) If, in addition, $\{W_t\}$ is independent of $\{X_t\}$, the following limit distribution theory holds:

$$T_{\mathbb{C}}^{1/2}(n)\,(\widetilde{\gamma}_n - \gamma_0) \xrightarrow{d} N\left(\mathbf{0}, \pi_s(\mathbb{C})(\phi_2/\phi_1^2)\Delta_m^{-1}(\gamma_0)\right), \tag{3.4}$$

where $T_{\mathbb{C}}(n) = \sum_{t=1}^{n} I(X_t \in \mathbb{C})$ and $\mathbb{C} \in \mathcal{E}^+$ is a small set.

Remark 3.2. (i) Following the proof of Theorem 3.1 in Appendix A, we may show that $\overline{\gamma}_n$ which minimises the untruncated loss function $\mathcal{L}_n(\cdot)$ in (2.1), is also weakly consistent and has the same limit distribution as that in (3.4). In fact, for an integrable function $g(\cdot, \gamma)$, by the ergodic theorem for the Harris recurrent Markov process, we may show that

$$\frac{1}{T(n)}\sum_{t=1}^{n} g(X_t, \gamma) I(|X_t| \leqslant \zeta_n) \xrightarrow{P} \int_{|x|\leqslant\zeta_n} g(x, \gamma)\pi_s(dx), \tag{3.5}$$

and

$$\frac{1}{T(n)}\sum_{t=1}^{n} g(X_t, \gamma) \xrightarrow{P} \int_{\mathbb{R}} g(x, \gamma)\pi_s(dx). \tag{3.6}$$

As $g(\cdot, \gamma)$ is π_s-integrable, letting $\zeta_n \to \infty$, we have $\int_{|x|>\zeta_n} g(x, \gamma)\pi_s(dx) \to 0$, which indicates that the limits in (3.5) and (3.6) would be the same. Furthermore, we may prove that

$$\frac{1}{T_{\mathbb{C}}^{1/2}(n)}\sum_{t=1}^{n} g(X_t, \gamma)\psi(W_t) \text{ and } \frac{1}{T_{\mathbb{C}}^{1/2}(n)}\sum_{t=1}^{n} g(X_t, \gamma)\psi(W_t) I(|X_t| \leqslant \zeta_n)$$

have the same limit distribution when $\zeta_n \to \infty$. Applying the above arguments, we may show that $T_{\mathbb{C}}^{1/2}(n)\,(\widetilde{\gamma}_n - \overline{\gamma}_n) = o_P(1)$.

(ii) By Lemma 3.2 in Karlsen and Tjøstheim (2001), we have $\frac{T_{\mathbb{C}}(n)}{T(n)} \to \pi_s(\mathbb{C})$ a.s. Hence, an alternative limit distribution result can be written as

$$T^{1/2}(n)\,(\widetilde{\gamma}_n - \gamma_0) \xrightarrow{d} N\left(\mathbf{0}, (\phi_2/\phi_1^2)\Delta_m^{-1}(\gamma_0)\right). \tag{3.7}$$

In practice, it is often preferred to use $T_{\mathbb{C}}(n)$ as the random normalisation in the asymptotic normal distribution as it is observable, facilitating statistical inference of the unknown parameter γ_0.

Using Theorem 3.1 and verifying Assumption 2, we can derive the asymptotic results for $\widetilde{\gamma}_{LS}$, $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_{M}$. When $\rho(u) = u^2$, assuming that $\mathsf{E}[W_t] = 0$, it is easy to show that Assumption 2 is satisfied with $\phi_1 = 2$ and $\phi_2 = 4 \cdot \mathsf{E}[W_t^2] =: 4\sigma_W^2$. When $\rho(u) = |u|$, assuming that W_t has a symmetric density function $f_W(\cdot)$ continuous at point 0, we may show that $\psi(u) = \mathsf{sign}(u)$ and Assumption 2 is satisfied with $\phi_1 = 2f_W(0)$ and $\phi_2 \equiv 1$. When $\rho(u)$ is chosen as Huber's loss function $\rho_\delta(u)$, we have

$$\psi_\delta(u) = \begin{cases} u, & |u| \leqslant \delta, \\ \delta \cdot \mathsf{sign}(u), & |u| > \delta, \end{cases}$$

and Assumption 2 is satisfied with $\phi_1 = \mathsf{P}(|W_t| \leqslant \delta)$ and $\phi_2 = \mathsf{E}\left[\psi_\delta^2(W_t)\right]$ if $\mathsf{E}[\psi_\delta(W_t)] = 0$. Combining the above arguments and Theorem 3.1, we readily have the following corollary.

Corollary 3.1. *Suppose that Assumptions 1 and 3 are satisfied and that $\zeta_n \to \infty$ as $n \to \infty$.*

(i) If $\rho(u) = u^2$, assuming that $\mathsf{E}[W_t] = 0$, the modified nonlinear least squares estimator $\widetilde{\gamma}_{LS}$ is weakly consistent. Furthermore, assuming that $\{W_t\}$ is independent of $\{X_t\}$, we have the following limit distribution:

$$T_{\mathbb{C}}^{1/2}(n)\left(\widetilde{\gamma}_{LS} - \gamma_0\right) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \pi_s(\mathbb{C})\sigma_W^2 \boldsymbol{\Delta}_m^{-1}(\gamma_0)\right). \tag{3.8}$$

(ii) Suppose that W_t has a symmetric density function $f_W(\cdot)$ continuous at point 0 and $f_W(0) > 0$. If $\rho(u) = |u|$, the modified nonlinear least absolute deviation estimator $\widetilde{\gamma}_{LAD}$ is weakly consistent. Furthermore, assuming that $\{W_t\}$ is independent of $\{X_t\}$, we have the following limit distribution:

$$T_{\mathbb{C}}^{1/2}(n)\left(\widetilde{\gamma}_{LAD} - \gamma_0\right) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \pi_s(\mathbb{C})[2f_W(0)]^{-2} \boldsymbol{\Delta}_m^{-1}(\gamma_0)\right). \tag{3.9}$$

(iii) If $\rho(u)$ is the Huber's loss function $\rho_\delta(u)$, assuming that $\mathsf{E}[\psi_\delta(W_t)] = 0$, the modified Huber's M-estimator $\widetilde{\gamma}_{M}$ is weakly consistent. Furthermore, assuming that $\{W_t\}$ is independent of $\{X_t\}$, we have the following limit distribution:

$$T_{\mathbb{C}}^{1/2}(n)\left(\widetilde{\gamma}_{M} - \gamma_0\right) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \pi_s(\mathbb{C})\phi(\delta) \boldsymbol{\Delta}_m^{-1}(\gamma_0)\right), \tag{3.10}$$

where

$$\phi(\delta) = \mathsf{E}\left[\psi_\delta^2(W_t)\right] / \mathsf{P}^2(|W_t| \leqslant \delta).$$

Remark 3.3. (i) Corollary 3.1 above covers and extends some existing results developed in the literature. For example, Corollary 3.1(i) is the same as Theorem 3.1 and Corollary 3.1 in Li,

Tjøstheim and Gao (2016); Corollary 3.1(ii) and (iii) complements the limiting results in Bai, Rao and Wu (1992) and Phillips (1995) which study the robust estimation in linear regression for stationary and non-stationary time series, respectively. The main difference among the limit distribution results in (3.8)–(3.10) is the asymptotic variance. Treating $\widetilde{\gamma}_{LS}$ as the benchmark, we may calculate the following two ratios:

$$R(\widetilde{\gamma}_{LAD}, \widetilde{\gamma}_{LS}) = \sigma_W^{-2}[2f_W(0)]^{-2} \text{ and } R(\widetilde{\gamma}_M, \widetilde{\gamma}_{LS}) = \phi(\delta)\sigma_W^{-2},$$

which can be used to examine the estimation efficiency. For the special case of standard normally distributed W_t, we readily have that $\sigma_W^2 = 1$, $f_W(0) = 1/\sqrt{2\pi}$ and $\phi(\delta) \to 1$ when δ is sufficiently large. Consequently, we can show that $R(\widetilde{\gamma}_{LAD}, \widetilde{\gamma}_{LS}) = 2\pi/4 > 1$ (indicating that $\widetilde{\gamma}_{LS}$ is asymptotically more efficient than $\widetilde{\gamma}_{LAD}$) and $R(\widetilde{\gamma}_M, \widetilde{\gamma}_{LS}) \to 1$ as $\delta \to \infty$ (indicating that $\widetilde{\gamma}_M$ would be as efficient as $\widetilde{\gamma}_{LS}$ if δ is chosen to be large enough).

(ii) By (2.10) and $\frac{T_{\mathbb{C}}(n)}{T(n)} \to \pi_s(\mathbb{C}) > 0$ a.s., the random number $T_{\mathbb{C}}(n)$ has the asymptotic divergence rate of $n^{\beta}L_s(n)$. Then, from Theorem 3.1 and Corollary 3.1 above, the convergence rate for the developed robust estimators (including $\widetilde{\gamma}_{LS}$, $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_M$) becomes $\sqrt{n^{\beta}L_s(n)}$, slower than the $\sqrt{n}$-rate in stationary parametric regression estimation as $\beta < 1$ and $L_s(\cdot)$ is a slowly varying function. In particular, if X_t follows the random walk process, we have $\beta = 1/2$ and $L_s(n) \equiv 1$. The convergence rate in the above limit theory becomes $O_P(n^{-1/4})$.

3.2 The case of asymptotically homogeneous function

A real function $G(x, \boldsymbol{\gamma})$ *is said to be asymptotically homogeneous on the set* $\boldsymbol{\Upsilon}$ *if*

$$G(wx, \boldsymbol{\gamma}) = v(w)H(x, \boldsymbol{\gamma}) + R(x, w, \boldsymbol{\gamma}), \tag{3.11}$$

where $v(\cdot)$ *is a positive real function,* $H(x, \boldsymbol{\gamma})$ *and* $R(x, w, \boldsymbol{\gamma})$ *satisfy the following two conditions: (i)* $H(x, \boldsymbol{\gamma})$ *is locally bounded uniformly over* $\boldsymbol{\gamma} \in \boldsymbol{\Upsilon}$ *and continuous with respect to* $\boldsymbol{\gamma}$*; (ii)* $R(x, w, \boldsymbol{\gamma})$ *is of order smaller than* $v(w)$ *as* $w \to \infty$ *on* $\boldsymbol{\Upsilon}$. Throughout the paper, we call $v(\cdot)$ the asymptotic order and $H(\cdot, \cdot)$ the limit homogeneous function of $G(\cdot, \cdot)$, and assume that $v(\cdot)$ does not depend on $\boldsymbol{\gamma}$. A vector (or matrix) of asymptotically homogeneous functions on $\boldsymbol{\Upsilon}$ is defined similarly by assuming each component function has the decomposition as in (3.11). Examples of asymptotically homogeneous functions include $G(x, \gamma) = \gamma G(x)$ with $G(x)$ being an asymptotically homogeneous function independent of γ, i.e., (3.11) is simplified to $G(wx) = v(w)H(x) + R(x, w)$, where the conditions imposed on $H(\cdot)$ and $R(\cdot, \cdot)$ are similar to those on $H(\cdot, \cdot)$ and $R(\cdot, \cdot, \cdot)$ in (3.11). A special case is to take $G(x) = x$ and thus $G(x, \gamma) = \gamma x$. Throughout this subsection, we assume that the

nonlinear regression function $m(x, \boldsymbol{\gamma})$ and its first and second partial derivatives with respect to $\boldsymbol{\gamma}$:

$$\dot{m}(x, \boldsymbol{\gamma}) = \left(\frac{\partial m(x, \boldsymbol{\gamma})}{\partial \gamma_j} \right)_{d \times 1} =: [\dot{m}_1(x, \boldsymbol{\gamma}), \cdots, \dot{m}_d(x, \boldsymbol{\gamma})]^\tau$$

$$\ddot{m}(x, \boldsymbol{\gamma}) = \left(\frac{\partial^2 m(x, \boldsymbol{\gamma})}{\partial \gamma_i \partial \gamma_j} \right)_{d \times d} =: [\ddot{m}_{ij}(x, \boldsymbol{\gamma})]_{d \times d},$$

are asymptotically homogeneous on $\boldsymbol{\Upsilon}$. The asymptotic orders of $m(x, \boldsymbol{\gamma})$, $\dot{m}_i(x, \boldsymbol{\gamma})$ and $\ddot{m}_{ij}(x, \boldsymbol{\gamma})$ are denoted by $\kappa(\cdot)$, $\dot{\kappa}_i(\cdot)$ and $\ddot{\kappa}_{ij}(\cdot)$, respectively; and their limit homogeneous functions are denoted by $h(\cdot, \cdot)$, $\dot{h}_i(\cdot, \cdot)$ and $\ddot{h}_{ij}(\cdot, \cdot)$, respectively.

To derive the asymptotic properties of the robust estimator $\widetilde{\boldsymbol{\gamma}}_n$ when the nonlinear regression function and its derivatives are asymptotically homogeneous, we use some additional assumptions.

Assumption 4. *(i) There exists a positive constant* $\underline{\psi}$ *such that*

$$M \cdot \mathsf{E}\left[\psi(W_t + M)) - \psi(W_t)\right] \geqslant \underline{\psi} \cdot M^2$$

for any real number M.

(ii) Letting $\pi_s(\zeta_n) = \int_{-\zeta_n}^{\zeta_n} \pi_s(dx)$*, for any* $\zeta_n^\star = o(\zeta_n)$*,* $\pi_s(\zeta_n^\star) = o\left(\pi_s(\zeta_n)\right)$*. The invariant density function* $p_s(\cdot)$ *exists, and is bounded satisfying*

$$\inf_{|x| \leqslant \zeta_n} p_s(x) \geqslant n^{4\xi_0 - \beta},$$

where $0 < \xi_0 < \beta/4$*. In addition,* ζ_n *diverges to infinity at a polynomial rate of* n*, and for any* $x \in [-\zeta_n, \zeta_n]$*,* $\mathbb{N}_x(1)$*, a neighborhood of* x *with radius* 1*, is a small set.*

Assumption 5. *(i) The limit homogeneous functions* $h(x, \boldsymbol{\gamma})$*,* $\dot{h}_i(x, \boldsymbol{\gamma})$ *and* $\ddot{h}_{ij}(x, \boldsymbol{\gamma})$ *are Lipschitz continuous over* $|x| \leqslant 1$ *and* $\boldsymbol{\gamma} \in \boldsymbol{\Upsilon}$*. In addition, for any* $\boldsymbol{\gamma} \in \boldsymbol{\Upsilon}$ *and* $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}_0$*,*

$$\lim_{n \to \infty} \frac{1}{\pi_s(\zeta_n)} \int_{-\zeta_n}^{\zeta_n} \left[h\left(x/\zeta_n, \boldsymbol{\gamma}\right) - h\left(x/\zeta_n, \boldsymbol{\gamma}_0\right)\right]^2 \pi_s(dx) > 0 \ \text{ as } \ n \to \infty. \tag{3.12}$$

(ii) The asymptotic orders satisfy that $\kappa(\zeta_n) \to \infty$ *and* $\sup_{1 \leqslant i,j \leqslant d} \frac{\kappa(\zeta_n)\ddot{\kappa}_{ij}(\zeta_n)}{\dot{\kappa}_i(\zeta_n)\dot{\kappa}_j(\zeta_n)} < \infty$.

(iii) The $d \times d$ *matrix* $\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)$ *defined as*

$$\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0) = \lim_{n \to \infty} \frac{1}{\pi_s(\zeta_n)} \int_{-\zeta_n}^{\zeta_n} \dot{h}(x/\zeta_n, \boldsymbol{\gamma}_0)\dot{h}^\tau(x/\zeta_n, \boldsymbol{\gamma}_0)\pi_s(dx) \tag{3.13}$$

with $\dot{h}(\cdot,\cdot) = \left[\dot{h}_1(\cdot,\cdot), \cdots, \dot{h}_d(\cdot,\cdot)\right]^\tau$, *is positive definite.*

Remark 3.4. Assumption 4(i) implies some additional smoothness and moment conditions on $\psi(\cdot)$ and W_t. The convexity of $\rho(\cdot)$ ensures that $M \cdot \mathsf{E}\left[\psi(W_t + M)) - \psi(W_t)\right]$ is non-negative for any real number M. It includes the case of $\rho(u) = u^2$ (corresponding to the nonlinear least squares estimation) but excludes the cases of L_1 and Huber (1964)'s loss functions, which will be dealt with separately. Assumption 4(ii) is on some restrictions on the invariant measure $\pi_s(\cdot)$ and its derivative $p_s(\cdot)$. In particular, we need to assume that there exists an order for the lower bound of $p_s(x)$ over $|x| \leqslant \zeta_n$, which is allowed to be divergent to zero. Such a condition is automatically satisfied if we consider the random walk process ($p_s(x) \equiv 1$). The conditions on the asymptotic orders and limit homogeneous functions in Assumption 5 are comparable to those in Park and Phillips (2001) and Chan and Wang (2015). In particular, if X_t is a random walk process, the identification condition (3.12) becomes

$$\int_{-1}^{1} \left[h\left(x, \gamma\right) - h\left(x, \gamma_0\right)\right]^2 dx > 0,$$

and the definition of $\Delta_h(\gamma_0)$ in (3.13) can be simplified to

$$\Delta_h(\gamma_0) = \frac{1}{2} \lim_{n\to\infty} \frac{1}{\zeta_n} \int_{-\zeta_n}^{\zeta_n} \dot{h}(x/\zeta_n, \gamma_0)\dot{h}^\tau(x/\zeta_n, \gamma_0)dx = \frac{1}{2}\int_{-1}^{1} \dot{h}(x, \gamma_0)\dot{h}^\tau(x, \gamma_0)dx.$$

We next give the asymptotic theory of $\widetilde{\gamma}_n$ for the case of asymptotically homogeneous nonlinear regression function.

Theorem 3.2. *Suppose that Assumptions 1, 2, 4 and 5 are satisfied.*

(i) The robust estimation $\widetilde{\gamma}_n$ *defined in (2.3) is weakly consistent.*

(ii) If, in addition, $\{W_t\}$ *is independent of* $\{X_t\}$*, the following limit distribution theory holds:*

$$\mathbf{D}(\zeta_n, \dot{\kappa})(\widetilde{\gamma}_n - \gamma_0) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, (\phi_2/\phi_1^2)\Delta_h^{-1}(\gamma_0)\right), \tag{3.14}$$

where $\mathbf{D}(\zeta_n, \dot{\kappa})$ *is a* $d \times d$ *diagonal matrix with the i-th diagonal number being* $T^{1/2}(n)\dot{\kappa}_i(\zeta_n)\pi_s^{1/2}(\zeta_n)$.

Remark 3.5. Note that we allow the diagonal elements of $\mathbf{D}(\zeta_n, \dot{\kappa})$ to have different orders, which is mainly caused by the varying asymptotic orders $\dot{\kappa}_i(\cdot)$. As a result, different components of $\widetilde{\gamma}_n$ could have different convergence rates. Furthermore, by choosing $\zeta_n \sim n^{1-\beta}L_s^{-1}(n)$ as in Li, Tjøstheim and Gao (2016), we may find that the convergence rates also rely on the value of β.

From Theorem 3.2 above, we can easily derive the asymptotic property of the modified nonlinear least squares estimator $\widetilde{\gamma}_{LS}$ by choosing the convex loss function as $\rho(u) = u^2$, which satisfies

Assumption 4(i) by letting $\underline{\psi} = 2$.

Corollary 3.2. *Suppose that Assumptions 1, 4(ii) and 5 are satisfied and* $\mathsf{E}[W_t] = 0$. *If* $\rho(u) = u^2$, *the modified nonlinear least squares estimator* $\widetilde{\gamma}_{LS}$ *is weakly consistent. Furthermore, assuming that* $\{W_t\}$ *is independent of* $\{X_t\}$, *we have the following limit distribution:*

$$\mathbf{D}(\zeta_n, \dot{\kappa})(\widetilde{\gamma}_{LS} - \gamma_0) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \sigma_W^2 \Delta_h^{-1}(\gamma_0)\right), \tag{3.15}$$

where $\sigma_W^2 = \mathsf{E}\left[W_t^2\right]$ *as defined in Corollary 3.1(i).*

Remark 3.6. Corollary 3.2 generalises Theorem 3.2 in Li, Tjøstheim and Gao (2016) which implicitly assumes that the asymptotic orders $\dot{\kappa}_i(\cdot)$ remain the same over $i = 1, \cdots, d$. As discussed in Remark 3.5, we allow that different components of $\widetilde{\gamma}_{LS}$ could have different random convergence rates $T^{1/2}(n)\dot{\kappa}_i(\zeta_n)\pi_s^{1/2}(\zeta_n)$. In addition, the technical assumptions in Assumption 5 are substantially simpler than those in Li, Tjøstheim and Gao (2016).

As pointed out in Remark 3.4, Assumption 4(i) cannot cover the convex loss functions used in the least absolute deviation estimator and Huber's M-estimator. The following theorem establishes the asymptotic theory for $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_M$.

Theorem 3.3. *Suppose that Assumptions 1, 4(ii) and 5 are satisfied with (3.12) replaced by*

$$\lim_{n\to\infty} \frac{1}{\pi_s(\zeta_n)} \int_{-\zeta_n}^{\zeta_n} |h(x/\zeta_n, \gamma) - h(x/\zeta_n, \gamma_0)| \, \pi_s(dx) > 0 \ \text{ as } \ n \to \infty. \tag{3.16}$$

(i) Suppose that W_t *has a symmetric density function* $f_W(\cdot)$ *continuous at point* 0 *and* $f_W(0) > 0$. *If* $\rho(u) = |u|$, *the modified nonlinear least absolute deviation estimator* $\widetilde{\gamma}_{LAD}$ *is weakly consistent. Furthermore, assuming that* $\{W_t\}$ *is independent of* $\{X_t\}$, *we have the following limit distribution:*

$$\mathbf{D}(\zeta_n, \dot{\kappa})(\widetilde{\gamma}_{LAD} - \gamma_0) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, [2f_W(0)]^{-2} \Delta_h^{-1}(\gamma_0)\right). \tag{3.17}$$

(ii) Suppose that $|m(x, \gamma_0) - m(x, \gamma)| \to \infty$ *as* $|x| \to \infty$ *for* $\gamma \neq \gamma_0$. *If* $\rho(u)$ *is the Huber's loss function* $\rho_\delta(u)$ *and* $\mathsf{E}[\psi_\delta(W_t)] = 0$, *the modified Huber's M-estimator* $\widetilde{\gamma}_M$ *is weakly consistent. Furthermore, assuming that* $\{W_t\}$ *is independent of* $\{X_t\}$, *we have the following limit distribution:*

$$\mathbf{D}(\zeta_n, \dot{\kappa})(\widetilde{\gamma}_M - \gamma_0) \xrightarrow{d} \mathsf{N}\left(\mathbf{0}, \phi(\delta) \Delta_h^{-1}(\gamma_0)\right), \tag{3.18}$$

where $\phi(\delta)$ *is defined as in Corollary 3.1(iii).*

Remark 3.7. For the case of linear regression model with $m(x, \gamma) = x\gamma$, if the non-stationary

regressor X_t is generated from a random walk process, we would find that the limit distribution results (3.17) and (3.18) are comparable to those in Theorems 4.1 and 5.1 in Phillips (1995). In particular, choosing $\zeta_n = c_\zeta n^{1/2}$ with c_ζ being a positive constant, we obtain super-consistency for $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_M$ with n-convergence rate.

3.3 Extensions

In Sections 3.1 and 3.2 above, to derive the weak consistency for the proposed robust estimators, we assume that W_t is independent of the β-null recurrent regressors X_s with $s \leqslant t$; and to establish the limit distribution theory, we further assume that $\{W_t\}$ and $\{X_t\}$ are mutually independent. Although the latter is commonly used in parametric and nonparametric regression models with null recurrent regressors (c.f., Karlsen, Myklebust and Tjøstheim, 2007; Li, Tjøstheim and Gao, 2016), it can be rather restrictive as it excludes some interesting cases in practical applications, such as, for example, a nonlinear autoregressive time series model with nonstationarity. In the context of parametric nonlinear least squares estimation (c.f., Park and Phillips, 2001; Chan and Wang, 2015), the following assumption is often imposed on W_t: there exists a filtration $\mathcal{F}_t$ such that X_t is adapted to $\mathcal{F}_{t-1}$ and $\{(W_t, \mathcal{F}_t)\}$ is a martingale difference sequence (*m.d.s.*) with $\mathsf{E}(W_t|\mathcal{F}_{t-1}) = 0$ a.s. To further derive the limit distribution theory when X_t is null recurrent (see Theorem 3.3 in Chan and Wang, 2015), $\mathcal{F}_t$ is replaced by $\mathcal{F}_{nt} := \sigma(\mathcal{F}_t, X_1, X_2, \cdots, X_n)$, which is slightly weaker than the mutual independence assumption between $\{W_t\}$ and $\{X_t\}$. However, in the context of general robust estimation, the above m.d.s. assumption would be insufficient as $\mathsf{E}(\psi(W_t)|\mathcal{F}_{t-1})$ does not equal to zero in general unless $\psi(u) = 2u$, corresponding to the nonlinear least squares estimation. A possible alternative assumption is to impose the m.d.s. condition directly on $\psi(W_t)$. We conjecture that the limiting properties in Sections 3.1 and 3.2 may still hold, but substantial modifications are needed in the mathematical proofs.

As discussed in Tjøstheim (2020), it is feasible to extend the methodology and theory developed in this paper to a parametric nonlinear autoregression setting defined by

$$X_t = m^\star(X_{t-1}, \boldsymbol{\gamma}_0^\star) + W_t, \tag{3.19}$$

where $m^\star(\cdot,\cdot)$ is a pre-specified nonlinear autoregression function, and $\boldsymbol{\gamma}_0^\star$ is an unknown vector of parameters. For the model (3.19), it is unreasonable to assume that $\{W_t\}$ is independent of $\{X_t\}$ as X_t clearly depends on W_t. However, under some structural assumptions on $m^\star(\cdot,\cdot)$ and W_t (c.f., Gao, Tjøstheim and Yin, 2013; Cai, Gao and Tjøstheim, 2017), we may prove that $\{X_t\}$ is β-null

recurrent and rewrite

$$\sum_{t=1}^{n} \dot{m}(X_t, \boldsymbol{\gamma}_0)\psi(W_t)\mathbb{I}(|X_t| \leqslant \zeta_n) = \sum_{t=1}^{n} \dot{m}(X_t, \boldsymbol{\gamma}_0)\psi(X_t - m^\star(X_{t-1}, \boldsymbol{\gamma}_0^\star))\mathbb{I}(|X_t| \leqslant \zeta_n)$$
$$=: \sum_{t=1}^{n} M(X_t, X_{t-1})$$

in Lemma A.4, a key result for proving the limit distribution. Then, using the technique in Section 4 of Karlsen and Tjøstheim (2001), we can prove the central limit theorem for $\sum_{t=1}^{n} M(X_t, X_{t-1})$ with appropriate random normalisation.

A second possible extension is to assume that the compound process $\{(X_t, W_t)\}$ is β-null recurrent in the nonlinear regression model setting (1.1). As discussed in Section 4 of Karlsen, Myklebust and Tjøstheim (2007), we may use the following assumption: $\{X_t\}$ and $\{W_t\}$ are asymptotically independent in the sense that the invariant measure for the compound process can be factorised into a product of two measures corresponding to $\{X_t\}$ and $\{W_t\}$, respectively. Such an assumption allows for dependence between X_t and W_t for fixed t (see Example 4.1 in Karlsen, Myklebust and Tjøstheim, 2007), but the stationary process $\{W_t\}$ has little influence on the nonstationary and null recurrent process $\{X_t\}$ in the asymptotic sense. However, Karlsen, Myklebust and Tjøstheim (2007) need some additional conditions to establish the limit distribution theory. For example, they have to assume that the errors are bounded, excluding the case of heavy-tailed model errors.

The developed methodology and theory heavily rely on the assumption that the nonlinear regression function $m(\cdot, \cdot)$ is pre-specified. An interesting extension is to relax this restriction and consider a semiparametric regression model with an unknown link function. In particular, the following single-index model structure [1] has been extensively studied in the literature:

$$Y_t = m\left(\mathbf{X}_t^\tau \boldsymbol{\gamma}\right) + W_t, \tag{3.20}$$

where the link function $m(\cdot)$ is unspecified and the parameter vector $\boldsymbol{\gamma}$ is unknown. When the observations are stationary, various semiparametric estimation methods have been proposed to estimate both the parametric and nonparametric components in model (3.20) (c.f., Härdle, Hall and Ichimura, 1993; Xia *et al*, 2002). A recent paper by Dong, Gao and Tjøstheim (2016) considers estimating model (3.20) when $\mathbf{X}_t$ satisfies a nonstationary unit root framework. We next briefly discuss how to construct a robust version of Dong, Gao and Tjøstheim (2016)'s estimation method. Dong, Gao and Tjøstheim (2016) suggest approximating the nonparametric function $m(\cdot)$ (for given $\boldsymbol{\gamma}$) via an orthogonal series, obtain its nonparametric estimation as a function of $\boldsymbol{\gamma}$, and then estimate the unknown parameter vector $\boldsymbol{\gamma}$ through ordinary least squares. It seems sensible to

[1] We thank the Associate Editor for suggesting this model specification.

replace their loss function by a general convex loss function as described in Section 2, and obtain a robust semiparametric estimation method for γ in model (3.20). Appendix C in the supplemental material will report a simulation study for the robust single-index model estimation. Note that the theory developed in Sections 3.1 and 3.2 is limited to the univariate null recurrent regressor setting. As pointed out by Park and Phillips (2001) and Li, Tjøstheim and Gao (2016), extension to multivariate nonstationary regressors is often challenging. For example, a three-dimensional vector containing three independent random walks is transient (Myklebust, Karlsen and Tjøstheim, 2012), indicating that the theory developed in Sections 3.1 and 3.2 would be not applicable. The relevant theoretical properties for the single-index model estimation will be considered in our future studies.

4 Simulation studies

In this section we conduct Monte Carlo simulation studies to illustrate the finite-sample numerical performance of the proposed robust estimators. The two simulated examples are similar to those considered by Li, Tjøstheim and Gao (2016), but we allow presence of outliers in the data and heavier tails in the error distribution, a typical setting for us to verify the robustness property of $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_{M}$ in finite samples. We start with a brief discussion on how the proposed robust estimators $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_{M}$ (which lack closed-form solutions) can be computed in practice. Computation of the LAD estimator $\widetilde{\gamma}_{LAD}$ can be carried out using standard derivative free minimising algorithms such as those used by the function `optimize` of the statistical programming language R (used in Example 4.1) or the function `fminsearch` of MATLAB. Alternatively, it can be computed using the `nlr` function with $\tau = 0.5$ available in the R package `quantreg`. Computation of the Huber estimator $\widetilde{\gamma}_{M}$ can be carried out using the iterated reweighted least squares algorithm which is used in the function `nlrob` available in the R package `Robustbase` (used in Example 4.1). For the linear regression in Example 4.2, computation of the proposed estimators can be carried out using standard quantile and robust regression packages available in MATLAB, R and Stata.

Example 4.1. Consider the following nonlinear regression model:

$$Y_t = \exp\left(-\gamma_0 X_t^2\right) + W_t, \;\; \gamma_0 = 1, \tag{4.1}$$

where X_t is either a random walk process $X_t = X_{t-1} + \varepsilon_t$ or a threshold autoregressive (TAR) process $X_t = 0.5I\left(|X_{t-1}| \leqslant 1\right) + X_{t-1}I\left(|X_{t-1}| > 1\right) + \varepsilon_t$, $X_0 = 0$, and both ε_t and W_t are i.i.d. random variables independent of each other. The innovations ε_t follow the standard normal distribution, whereas the unobservable errors W_t are generated from either $N\left(0, 0.5^2\right)$ or a mixed normal (MN) distribution $0.95N\left(0,1\right) + 0.05N\left(0, 10^2\right)$. The outliers are added to X_t by randomly replacing 5%

of their values with those randomly drawn from a uniform distribution between 20 and 25. We consider three estimation methods for γ_0: $\widetilde{\gamma}_{LS}$, $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_{M}$ defined in Section 2. Note that Li, Tjøstheim and Gao (2016) use $\widetilde{\gamma}_{LS}$ in their simulation studies, but do not consider any robust estimation. In the simulations, we consider sample sizes: $n = 500, 1000, 5000$, and 1000 replications. The tuning parameter ζ_n is chosen as $2.58n^{1/2}$, the same as that in Li, Tjøstheim and Gao (2016).

Tables 1 and 2 report the finite-sample mean absolute bias (upper position) and standard deviations (lower position) of the three estimators for γ_0 with the two different generating schemes of X_t. In addition, we also report the number of times in which the estimation procedure fails to converge (after 500 iterations, in which case the estimate obtained in the last iteration is used). From the tables, in the absence of outliers and with normal errors, the robust estimators, i.e., $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_{M}$, have finite-sample mean absolute biases and standard deviations that are comparable to those of the least squares based estimator $\widetilde{\gamma}_{LS}$. With mixed normal errors, the numerical performance of $\widetilde{\gamma}_{LS}$ is negatively affected by the heavy tails, resulting in larger values of the mean absolute bias and standard deviations. In contrast, $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_{M}$ have robust performance with mean absolute bias and standard deviations similar to those in the light-tailed case. The finite-sample performance of $\widetilde{\gamma}_{LS}$ further deteriorates when outliers are randomly added to X_t. In addition, outliers and/or mixed normal errors affect the numerical stability of the nonlinear least squares estimator, as the number of convergence failures clearly indicates. Noting that the nonlinear regression function in (4.1) is integrable, from the asymptotic theory in Section 3.1, the convergence rates of the three estimators is rather slow. This can be seen in finite samples by comparing the results between Tables 1–2 and Tables 3–4 (in Example 4.2 below).

Example 4.2. Consider the following linear regression model:

$$Y_t = \gamma_0 X_t + W_t, \quad \gamma_0 = 0.5, \tag{4.2}$$

where X_t, W_t and the outliers are defined as those in Example 4.1 above. We again compare three estimators for γ_0: $\widetilde{\gamma}_{LS}$, $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_{M}$, and consider sample sizes: $n = 500, 1000, 5000$ and 1000 replications. Note that the linear regression function in (4.2) is asymptotically homogeneous.

Tables 3 and 4 report the finite-sample mean absolute bias and standard deviations of the three estimators for the parameter γ_0 in model (4.2). As in Tables 1 and 2, the robust estimators $\widetilde{\gamma}_{LAD}$ and $\widetilde{\gamma}_{M}$ have reliable numerical performance even when the outliers are present and the model errors have heavy tails, following the mixed normal distribution. In contrast, the performance of the least squares based estimator $\widetilde{\gamma}_{LS}$ becomes significantly worse in the presence of outliers in X_t and/or heavy tails in model errors. In addition, the values of the mean absolute bias and standard deviations in Tables 3 and 4 are much smaller than those in Tables 1 and 2, supporting the super fast consistency theory developed in Section 3.2.

Table 1: Absolute estimation bias and standard deviation for model (4.1) with random walk X_t

	$W_t \sim N(0, 0.5^2)$						$W_t \sim MN$					
n	NLS		LAD		Huber		NLS		LAD		Huber	
	No outlier						No outlier					
500	.033 .215	$(0)^a$	.034 .223	$(0)^a$	.035 .220	$(0)^a$	.143 .517	$(12)^a$	.038 .229	$(0)^a$	.035 .232	$(0)^a$
1000	.021 .186	$(0)^a$	.020 .182	$(0)^a$	.024 .185	$(0)^a$	.121 .449	$(9)^a$	.030 .186	$(0)^a$	.023 .189	$(0)^a$
5000	.010 .126	$(0)^a$	.012 .128	$(0)^a$	.012 .130	$(0)^a$	.064 .316	$(5)^a$	.016 .134	$(0)^a$	.015 .136	$(0)^a$
	With outliers						With outliers					
500	.204 .347	$(130)^a$	.035 .230	$(0)^a$	.039 .228	$(0)^a$	.299 .613	$(148)^a$	.039 .235	$(0)^a$	.041 .239	$(0)^a$
1000	.175 .312	$(101)^a$	.027 .186	$(0)^a$	.030 .189	$(0)^a$	.224 .558	$(123)^a$	.030 .191	$(0)^a$	.031 .193	$(0)^a$
5000	.110 .272	$(78)^a$	.016 .132	$(0)^a$	.015 .136	$(0)^a$	.145 .483	$(95)^a$	.020 .141	$(0)^a$	.021 .145	$(0)^a$

a number of convergence failures

Table 2: Absolute estimation bias and standard deviation for model (4.1) with TAR X_t

	$W_t \sim N(0, 0.5^2)$						$W_t \sim MN$					
n	NLS		LAD		Huber		NLS		LAD		Huber	
	No outlier						No outlier					
500	.030 .200	$(0)^a$	.031 .205	$(0)^a$	.034 .209	$(0)^a$	.139 .455	$(8)^a$	.029 .204	$(0)^a$	.027 .207	$(0)^a$
1000	.020 .176	$(0)^a$	.018 .178	$(0)^a$	.020 .179	$(0)^a$	.102 .401	$(5)^a$	.020 .181	$(0)^a$	.022 .180	$(0)^a$
5000	.009 .115	$(0)^a$	.013 .122	$(0)^a$	.016 .124	$(0)^a$	.060 .312	$(3)^a$	.013 .128	$(0)^a$	.012 .130	$(0)^a$
	With outliers						With outliers					
500	.189 .321	$(115)^a$	.030 .213	$(0)^a$	.032 .217	$(0)^a$	.255 .583	$(125)^a$	.035 .215	$(0)^a$	.038 .220	$(0)^a$
1000	.162 .299	$(93)^a$	.023 .179	$(0)^a$	.023 .176	$(0)^a$	.198 .538	$(113)^a$	.028 .186	$(0)^a$	.027 .184	$(0)^a$
5000	.104 .259	$(72)^a$	.018 .129	$(0)^a$	.020 .132	$(0)^a$	.136 .471	$(84)^a$	.015 .134	$(0)^a$	.018 .136	$(0)^a$

a number of convergence failures

Table 3: Mean absolute bias and standard deviation for model (4.2) with random walk X_t

	$W_t \sim N(0, 0.5^2)$			$W_t \sim MN$		
n	NLS	LAD	Huber	NLS	LAD	Huber
	No outliers			No outliers		
	.00005	.00006	.00006	.00012	.00008	.00009
500	.0026	.0025	.0024	.0056	.0029	.0030
	.00001	.00002	.00002	.00008	.00004	.00004
1000	.0012	.0014	.0012	.0027	.0016	.0017
	.00000	.00000	.00001	.00006	.00002	.00001
5000	.0002	.0003	.0002	.0007	.0003	.0003
	With outliers			With outliers		
	.00012	.00008	.00008	.00612	.00012	.00014
500	.0164	.0031	.0032	.0213	.0037	.0039
	.00009	.00004	.00005	.00515	.00007	.00008
1000	.0101	.0017	.0018	.0126	.0018	.0019
	.00005	.00002	.00002	.00315	.00004	.00005
5000	.0038	.0004	.0004	.0068	.0005	.0005

Table 4: Mean absolute bias and standard deviation for model (4.2) with TAR X_t

	$W_t \sim N(0, 0.5^2)$			$W_t \sim MN$		
n	NLS	LAD	Huber	NLS	LAD	Huber
	No outliers			No outliers		
	.00004	.00005	.00006	.00011	.00007	.00007
500	.0023	.0024	.0023	.0046	.0031	.0033
	.00009	.00009	.00010	.00007	.00004	.00005
1000	.0012	.0011	.0010	.0024	.0016	.0017
	.00000	.00000	.00000	.0006	.00001	.00001
5000	.0002	.0002	.0002	.0006	.0003	.0003
	With outliers			With outliers		
	.00341	.00008	.00008	.00586	.00019	.00020
500	.0154	.0029	.0030	.0192	.0034	.0034
	.00282	.00072	.00070	.00485	.00007	.00006
1000	.0102	.0015	.0016	.0112	.0017	.0018
	.00212	.00038	.00036	.00299	.0004	.0004
5000	.0049	.0003	.0003	.0054	.0005	.0005

Tables 1-4 above show the usefulness of the proposed robust estimators when outliers and/or heavy tails in model errors are present. However, they do not show the usefulness of the truncating mechanism in the robust estimation procedure. To do so we increase the magnitude of outliers in the sample by using random draws from a uniform distribution between 50 and 80 and the percentage of outliers from 5 to 10. Tables 5 and 6 report the finite-sample mean absolute bias and standard deviation of the Huber's M and LAD estimators with and without truncation, i.e., comparison between the two robust estimators defined by minimising the objective functions in (2.1) and (2.2), respectively. Tables 5 and 6 show that without involvement of truncation, the robust estimators are influenced by the presence of very large (aberrant) outliers in the simulated sample. On the other hand, the trimmed robust estimators have mean absolute biases and standard deviations that are directly comparable to those of Tables 1-4.

Table 5: Comparison between trimmed and untrimmed robust estimators for model (4.1)

	Random walk X_t							
	$W_t \sim N(0, 0.5^2)$		$W_t \sim MN$		$W_t \sim N(0, 0.5^2)$		$W_t \sim MN$	
	LAD	Huber	LAD	Huber	LAD	Huber	LAD	Huber
n	untrimmed estimators				trimmed estimators			
	.040	.046	.048	.047	.036	.041	.042	.042
500	.241	.243	.246	.249	.231	.235	.238	.242
	.035	.030	.036	.035	.028	.026	.032	.032
1000	.198	.204	.205	.203	.190	.199	.195	.195
	.019	.019	.025	.024	.016	.015	.022	.022
5000	.138	.141	.151	.150	.133	.136	.147	.146
	TAR X_t							
	$W_t \sim N(0, 0.5^2)$		$W_t \sim MN$		$W_t \sim N(0, 0.5^2)$		$W_t \sim MN$	
	LAD	Huber	LAD	Huber	LAD	Huber	LAD	Huber
n	untrimmed estimators				trimmed estimators			
	.038	.040	.043	.042	.031	.033	.037	.038
500	.227	.232	.229	.233	.214	.220	.217	.222
	.030	.030	.030	.032	.025	.030	.030	.034
1000	.189	.190	.193	.195	.181	.190	.187	.185
	.021	.023	.018	.021	.018	.020	.014	.018
5000	.135	.137	.140	.143	.128	.131	.134	.135

Table 6: Comparison between trimmed and untrimmed robust estimators for model (4.2)

	Random walk X_t							
	$W_t \sim N(0, 0.5^2)$		$W_t \sim MN$		$W_t \sim N(0, 0.5^2)$		$W_t \sim MN$	
	LAD	Huber	LAD	Huber	LAD	Huber	LAD	Huber
n	untrimmed estimators				trimmed estimators			
	.00014	.00015	.00017	.00018	.00009	.00010	.00012	.00013
500	.0051	.0048	.0052	.0053	.0030	.0032	.0038	.0040
	.00009	.00010	.00013	.00014	.00005	.00005	.00008	.00008
1000	.0024	.0026	.0025	.0026	.0017	.0019	.0018	.0019
	.00004	.00005	.00007	.00008	.00002	.00002	.00004	.00005
5000	.0011	.0012	.0011	.0012	.0004	.0003	.0006	.0005
	TAR X_t							
	$W_t \sim N(0, 0.5^2)$		$W_t \sim MN$		$W_t \sim N(0, 0.5^2)$		$W_t \sim MN$	
	LAD	Huber	LAD	Huber	LAD	Huber	LAD	Huber
n	untrimmed estimators				trimmed estimators			
	.00019	.00018	.00024	.00027	.0009	.00008	.00019	.00021
500	.0051	.0052	.0049	.0051	.0030	.0032	.0035	.0036
	.00014	.00015	.00018	.00020	.00008	.00005	.00008	.00008
1000	.0027	.0028	.0028	.0030	.0016	.0017	.0018	.0019
	.0008	.00008	.00010	.00009	.0003	.0004	.00004	.00005
5000	.0010	.0009	.0011	.0010	.0005	.0006	.0006	.0006

5 Empirical applications

In this section we provide two empirical examples to illustrate usefulness of the proposed robust estimation methodology. The first example uses the same data set as in Li, Tjøstheim and Gao (2016) to study the relationship between UK imports from (or exports to) USA and real exchange rates, and the second example examines the so-called Environmental Kuznets Curve (EKC) hypothesis by estimating the relationship between per capita CO_2 emission and per capita GDP in Denmark.

Example 5.1. Consider the data used by Li, Tjøstheim and Gao (2016) to investigate the relationship between UK imports from (or exports to) USA and the real exchange rates between the two nations. The data set[2] consists of 212 monthly observations of the nominal UK-USA exchange rate (E_t), the UK and USA consumer price indices (P_t^{UK} and P_t^{US}), the UK exports to the USA (EXP_t) and the UK imports from the USA (IMP_t), collected from January 1996 to August 2013. Let Y_t denote the log of either EXP_t or IMP_t and $X_t = \log(E_t) + \log(P_t^{UK}) - \log(P_t^{USA})$ denote the real UK-USA

[2]The data are available at `https://www.uktradeinfo.com`

exchange rate. The empirical analysis in Li, Tjøstheim and Gao (2016) suggests that X_t may follow a TAR process, a 1/2 recurrent Markov process, see also Gao, Tjøstheim and Yin (2013). To evaluate the usefulness of the proposed robust estimators, we consider the same polynomial specification between Y_t and X_t as that considered by Li, Tjøstheim and Gao (2016),

$$Y_t = \gamma_1 + \gamma_2 X_t + \gamma_3 X_t^2 + \gamma_4 X_t^3 + W_t, \tag{5.1}$$

where the unknown parameters $\gamma_1, \gamma_2, \gamma_3$ and γ_4 are estimated by one of the following three methods: modified LS, LAD and Huber's M estimation defined as in Section 2.

To construct the out-of-sample prediction, we split the sample into two parts: the training set consisting of the first 169 observations (corresponding to 80% of the sample), and the validation set consisting of the remaining 43 observations. The mean squared error (MSE) for out-of-sample forecasting is computed as

$$\text{MSE} = \frac{1}{43} \sum_{t=170}^{212} \left(Y_t - \widehat{Y}_t \right)^2,$$

where $\widehat{Y}_t$ is the fitted value of Y_t using the modified LS, LAD or Huber's M estimates of the unknown parameters in (5.1). To evaluate the estimation robustness, we consider two scenarios: one based on the original data, and the other one with 5 outliers (ranging from 1.5 times to twice the original value) randomly replacing the responses Y_t in the training set. To eliminate possible location effect due to the random positioning of outliers, we generate 100 samples of responses with randomly positioned outliers.

Table 7: MSE of out-of-sample forecasting

Exports	Original data	Data with outliers	Imports	Original data	Data with outliers
LS	.1336	.1787	LS	.0434	.3162
Huber	.1388	.1255	Huber	.0423	.0403
LAD	.1386	.1388	LAD	.0412	.0396

Table 7 reports the MSE for the case of original data and the average MSE (over the 100 replications) for the case with randomly positioned outliers. For the original data, the MSE values (based on all the three estimators) are very similar. When there are random outliers, the MSE values based on the LAD and Huber's M estimates are close to those obtained in the original data, whereas the MSE values based on the LS estimates are seriously affected by the outliers present in the training set. Figure 1 plots the MSE values over the 100 replications for the three estimators when outliers are present. This figure clearly shows that the MSE values of the out-of-sample forecasting based on the two robust estimators are very stable, as opposed to those based on the

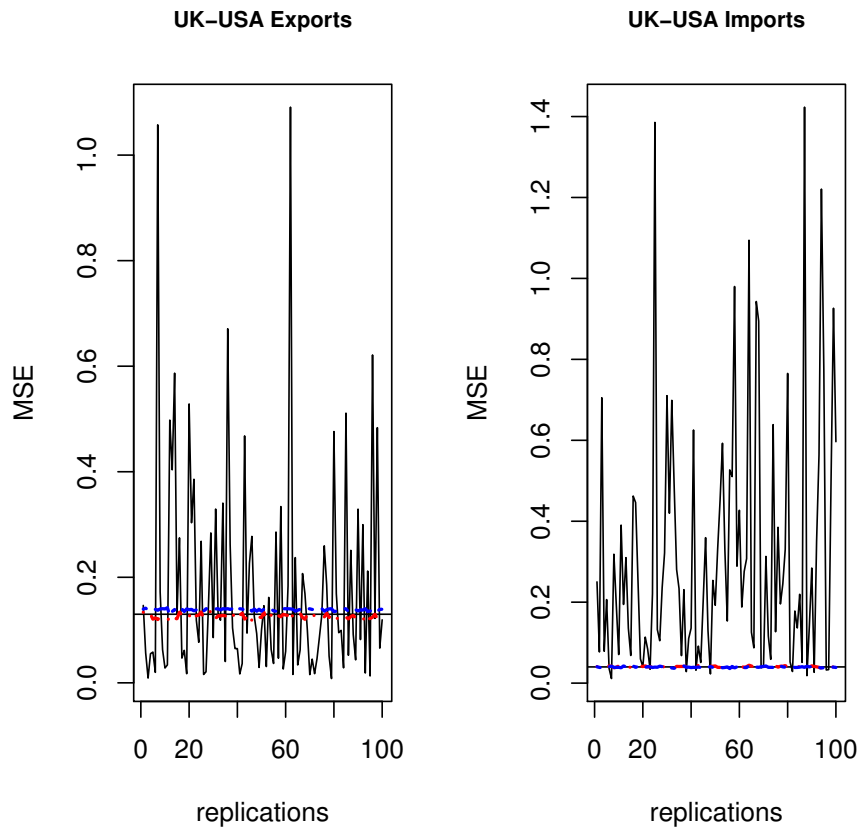

Figure 1: MSE of out-of-sample forecasting based on LS estimate (solid line), Huber-M estimate (dashed line) and LAD estimate (short dashed line) when the training set contains outliers

modified LS estimator.

Example 5.2. The EKC hypothesis suggests the existence of an inverted-U shaped relationship between enviromental degradation and income level of a given country, see, for example, Dinda (2004) for a review. Following Piaggio and Padilla (2012) and Chan and Wang (2015), the EKC is specified as

$$Y_t = \gamma_{10} + \gamma_{20} X_t + \gamma_{30} X_t^2 + W_t, \tag{5.2}$$

where Y_t and X_t denote, respectively, the log of per capita CO_2 emission and the log of per capita GDP of a given country. Both Piaggio and Padilla (2012) and Chan and Wang (2015) estimate (5.2) for 16 countries, after testing that both Y_t and X_t are nonstationary. Wang, Wu and Zhu (2018) further test the assumption made by Chan and Wang (2015) that the regressor X_t in (5.2) is endogenous, by examining the correct specification of (5.2), and conclude that out of the 16 countries, only three (Denmark, India and Ireland) are correctly specified - implying that the regressor X_t in (5.2) is exogenous. In the following empirical analysis we only consider the

Denmark case. We use the same annual data as in Piaggio and Padilla (2012) and Chan and Wang (2015) from 1950 to 2008.[3] To evaluate the usefulness of the proposed robust estimator, we replace one original response Y_t with an artificial outlier equaling to four times the original value. As with the previous empirical example, to eliminate possibility of the location effect, we generate 50 samples of responses with randomly positioned outliers. Table 8 reports the estimates of the unknown parameters of the original data and the average estimates (over the 50 replications) of the data with the artificial outliers. Table 8 also reports the standard deviations of the estimators (in parentheses) over the 50 replications when the data set contains an artificial outlier.

Table 8: Parameter estimates of the EKC with Danish data

	Original data			Data with outliers		
	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$
LS	-109.78	22.87	-1.17	-103.76 (18.62)	21.63 (3.89)	-1.11 (0.20)
Huber	-110.22	22.98	-1.18	-110.03 (1.58)	22.94 (0.33)	-1.18 (0.02)
LAD	-109.56	22.87	-1.18	-109.85 (1.07)	22.93 (0.22)	-1.18 (0.01)

Table 8 shows that both the LAD and Huber-M (average) estimates are virtually unaffected by the presence of a single outlier, whereas the LS estimates are severely affected, as the standard deviations clearly indicate. Figure 2 illustrates the regression fit of the EKC for the case with original data and one of the cases with an artificial outlier.

6 Conclusions

In this paper we consider parametric robust estimation in nonlinear regression models with a univariate regressor generated by a class of non-stationary and null recurrent Markov process and the regression function being either integrable or asymptotically homogeneous. We show that the proposed estimators are weakly consistent and asymptotically normal with convergence rates that depend on the functional form of nonlinear regression and the recurrence rate of the Markov process (i.e., the value of β). The developed asymptotic results substantially generalise some classic asymptotic theory in the existing literature (e.g., Phillips, 1995; Park and Phillips, 2001; Li, Tjøstheim and Gao, 2016). Monte-Carlo simulation studies as well as two real data applications show that the proposed robust estimators are characterised by good finite-sample properties and are stable for data contaminated by outliers.

[3]The CO_2 emission data is published by the Carbon Dioxide Information Analysis Center (Boden, Marland and Andres, 2009) at `https://cdiac.ess-dive.lbl.gov/frequent_data_products.html`, and the per capita GDP data is available at the Maddison Project database `http://www.ggdc.net/maddison`.

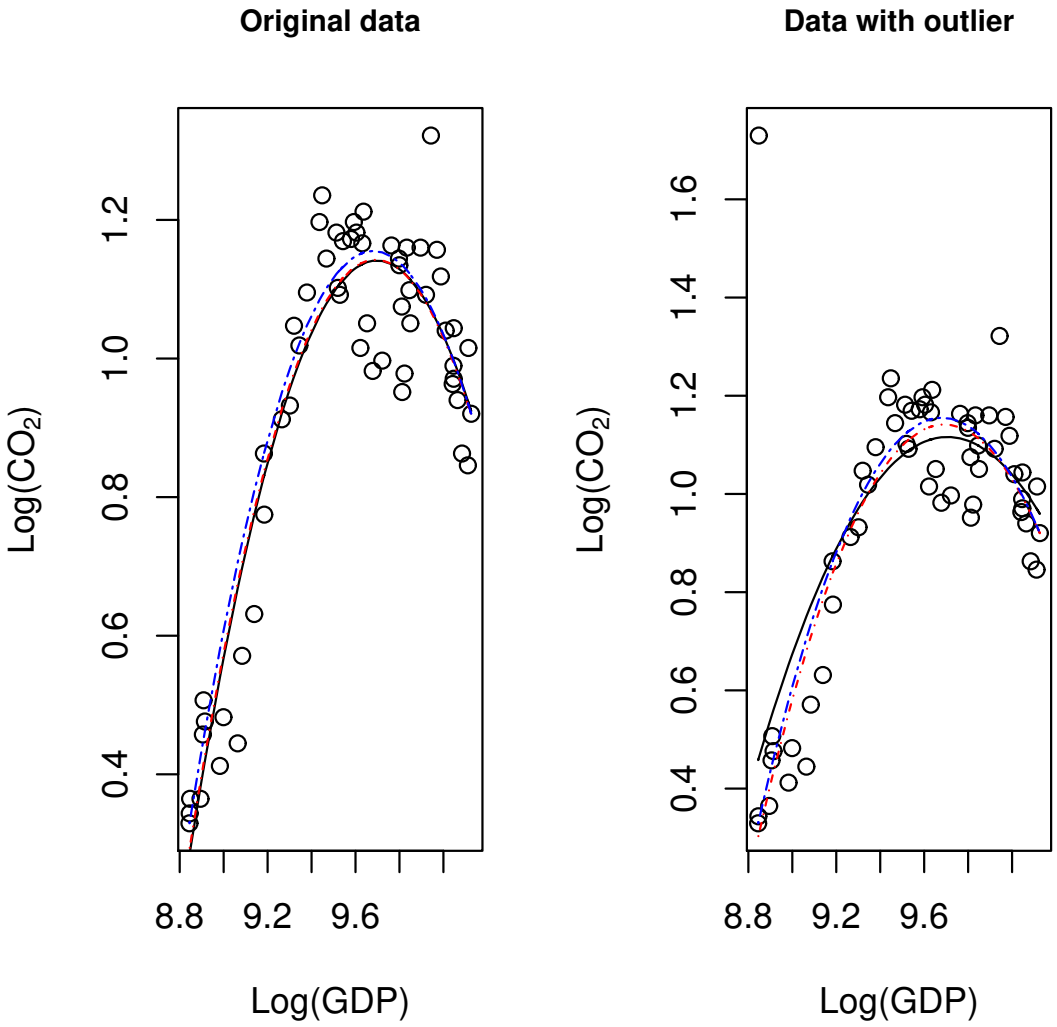

Figure 2: EKC for the Danish data based on LS estimate (solid line), Huber-M estimate (dashed line) and LAD estimate (short dashed line)

Acknowledgement

The authors thank a Co-Editor, an Associate Editor and two reviewers for helpful comments, which have substantially improved an earlier version of the paper.

Supplement

The supplementary document contains proofs of some technical lemmas as well as additional simulation results.

Appendix A: Proofs of the main results

In this appendix, we provide the detailed proofs of the main asymptotic results stated in Section 3 as well as some technical lemmas. The proofs of the technical lemmas are given in Appendix B which is available in the supplemental document. Define

$$V(X_t, W_t, \boldsymbol{\gamma}) = \rho\big(Y_t - m(X_t, \boldsymbol{\gamma})\big) I_t - \rho(W_t) I_t + [m(X_t, \boldsymbol{\gamma}) - m(X_t, \boldsymbol{\gamma}_0)]\, \psi(W_t) I_t,$$

where $I_t = I(|X_t| \leqslant \zeta_n)$. We start with two technical lemmas which play an important role in the proof of robust estimation consistency.

Lemma A.1. *Suppose that Assumptions 1, 2(ii) and 3(i) are satisfied. For any positive constant* M*, we have*

$$\sup_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \leqslant M} \left| \sum_{t=1}^{n} V(X_t, W_t, \boldsymbol{\gamma}) - \sum_{t=1}^{n} E\left[V(X_t, W_t, \boldsymbol{\gamma}) | \mathcal{F}_{t-1}\right] \right| = O_P\left(n^{\beta-\eta}\right), \tag{A.1}$$

where $\mathcal{F}_t$ *is a* σ*-field generated by* X_s *with* $s \leqslant t+1$ *and* W_s *with* $s \leqslant t$*,* $\|\cdot\|$ *denotes the Euclidean norm, and* $0 < \eta < \beta/(3+2d)$ *with* β *defined in (2.7).*

Remark A.1. The asymptotic order in (A.1) is not a sharp rate, but is sufficient to the proofs of our main results. In fact, it is easy to show that the rate $O_P\left(n^{\beta-\eta}\right)$ can be replaced by $o_P\left(n^{\beta-\eta}\right)$.

Lemma A.2. *Suppose that Assumptions 1, 2(i) and 3(i) are satisfied. For any sufficiently small* $\iota > 0$*, we have*

$$\sup_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \leqslant \iota} \left| \frac{1}{T_{\mathbb{C}}(n)} \sum_{t=1}^{n} E\left[V(X_t, W_t, \boldsymbol{\gamma}) | \mathcal{F}_{t-1}\right] - \frac{\phi_1}{2\pi_s(\mathbb{C})} \int_{\mathbb{R}} [m(x, \boldsymbol{\gamma}) - m(x, \boldsymbol{\gamma}_0)]^2 \pi_s(dx) \right| = o_P(1), \tag{A.2}$$

where $T_{\mathbb{C}}(n)$ *is defined in Theorem 3.1,* ϕ_1 *is defined in Assumption 2(i) and* $\pi_s(\cdot)$ *is the invariant measure for the* β*-null recurrent Markov process* $\{X_t\}$*.*

We next narrow the range of γ from those in Lemmas A.1 and A.2 to $T_{\mathbb{C}}^{1/2}(n)\|\gamma-\gamma_0\| \leqslant M$, and give the following lemma which is crucial to the proof of Theorem 3.1(ii).

Lemma A.3. *Suppose that Assumptions 1–3 are satisfied. For any positive constant* M*, we have*

$$\sup_{T_{\mathbb{C}}^{1/2}(n)\|\gamma-\gamma_0\|\leqslant M}\left|\sum_{t=1}^{n} V(X_t,W_t,\gamma)-\sum_{t=1}^{n}\mathsf{E}\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right]\right| = o_P(1), \tag{A.3}$$

and

$$\sup_{T_{\mathbb{C}}^{1/2}(n)\|\gamma-\gamma_0\|\leqslant M}\left|\sum_{t=1}^{n}\mathsf{E}\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right]-\frac{\phi_1 T_{\mathbb{C}}(n)}{2\pi_s(\mathbb{C})}(\gamma-\gamma_0)^{\intercal}\boldsymbol{\Delta}_m(\gamma_0)(\gamma-\gamma_0)\right| = o_P(1), \tag{A.4}$$

where $\boldsymbol{\Delta}_m(\gamma_0)$ *is defined in Assumption 3(ii).*

Lemma A.4. *Suppose that Assumptions 1, 2(ii) and 3 are satisfied and that* $\{W_t\}$ *is independent of* $\{X_t\}$*. Then, we have*

$$\frac{1}{T_{\mathbb{C}}^{1/2}(n)}\sum_{t=1}^{n}\dot{m}(X_t,\gamma_0)\psi(W_t)I_t \xrightarrow{d} \mathsf{N}\left(\mathbf{0},\phi_2\boldsymbol{\Delta}_m(\gamma_0)/\pi_s(\mathbb{C})\right), \tag{A.5}$$

where ϕ_2 *is defined in Assumption 2(ii).*

Proof of Theorem 3.1(i). Let $\mathbb{S}_{\gamma_0}(\varepsilon)$ be a circle centered at γ_0 with radius ε, where $\varepsilon>0$ is sufficiently small. It is sufficient to prove that

$$\mathsf{P}\left(\inf_{\gamma\in\mathbb{S}_{\gamma_0}(\varepsilon)}\mathcal{L}_{n,\zeta_n}(\gamma) > \mathcal{L}_{n,\zeta_n}(\gamma_0)\right)\to 1, \tag{A.6}$$

where $\mathcal{L}_{n,\zeta_n}(\cdot)$ is defined in (2.2). In fact, due to the convexity of the loss function $\rho(\cdot)$, (A.6) implies that $\inf_{\|\gamma-\gamma_0\|\geqslant\varepsilon}\mathcal{L}_{n,\zeta_n}(\gamma) > \mathcal{L}_{n,\zeta_n}(\gamma_0)$ with probability approaching one. Therefore, $\mathcal{L}_{n,\zeta_n}(\gamma)$ must have a local minimum in the interior of $\mathbb{S}_{\gamma_0}(\varepsilon)$, and consequently,

$$\mathsf{P}\left(\|\widetilde{\gamma}_n-\gamma_0\|>\varepsilon\right)\to 0, \tag{A.7}$$

completing the proof of Theorem 3.1(i).

By (2.9) and Lemmas A.1 and A.2 above, we have that uniformly over $\|\gamma-\gamma_0\|\leqslant\varepsilon$,

$$\begin{aligned}
&\frac{1}{T_{\mathbb{C}}(n)}\sum_{t=1}^{n}\rho\big(Y_t-m(X_t,\gamma)\big)I_t-\frac{1}{T_C(n)}\sum_{t=1}^{n}\rho(W_t)I_t\\
=\;&-\frac{1}{T_{\mathbb{C}}(n)}\sum_{t=1}^{n}\left[m(X_t,\gamma)-m(X_t,\gamma_0)\right]\psi(W_t)I_t+\\
&\frac{\phi_1}{2\pi_s(\mathbb{C})}\int_{\mathbb{R}}\left[m(x,\gamma)-m(x,\gamma_0)\right]^2\pi_s(dx)+o_P(1).
\end{aligned} \tag{A.8}$$

Following the proof of Lemma A.1 in the supplemental document and using (2.9) and the fact of $\frac{T_{\mathbb{C}}(n)}{T(n)}\to$

$\pi_s(\mathbb{C})$ a.s., it is easy to show that

$$\sup_{\|\gamma-\gamma_0\|\leqslant\varepsilon}\left|\sum_{t=1}^{n}[m(X_t,\gamma)-m(X_t,\gamma_0)]\,\psi(W_t)I_t\right| = O_P(n^{\beta-\eta}) = o_P(T_{\mathbb{C}}(n)). \tag{A.9}$$

By Assumption 3(ii), we have

$$\inf_{\gamma\in\mathbb{S}_{\gamma_0}(\varepsilon)}\int_{\mathbb{R}}[m(x,\gamma)-m(x,\gamma_0)]^2\,\pi_s(dx) > 0. \tag{A.10}$$

In view of (A.8)–(A.10), we can prove (A.6). ■

Proof of Theorem 3.1(ii). The proof is analogous to the proof of Theorem 2.4 in Bai, Rao and Wu (1992). Define

$$\widehat{\gamma}_n = \gamma_0 + \left[\frac{\phi_1}{\pi_s(\mathbb{C})}\boldsymbol{\Delta}_m(\gamma_0)\right]^{-1}\frac{1}{T_{\mathbb{C}}(n)}\sum_{t=1}^{n}\dot{m}(X_t,\gamma_0)\psi(W_t)I_t.$$

By Lemma A.4, we readily have

$$\begin{aligned} T_{\mathbb{C}}^{1/2}(n)(\widehat{\gamma}_n-\gamma_0) &= \left[\frac{\phi_1}{\pi_s(\mathbb{C})}\boldsymbol{\Delta}_m(\gamma_0)\right]^{-1}\frac{1}{T_{\mathbb{C}}^{1/2}(n)}\sum_{t=1}^{n}\dot{m}(X_t,\gamma_0)\psi(W_t)I_t \\ &\xrightarrow{d} N\left(\mathbf{0},\pi_s(\mathbb{C})(\phi_2/\phi_1^2)\boldsymbol{\Delta}_m^{-1}(\gamma_0)\right). \end{aligned} \tag{A.11}$$

By (A.11), we only need to prove

$$P\left(T_{\mathbb{C}}^{1/2}(n)\|\widetilde{\gamma}_n-\widehat{\gamma}_n\| > \varepsilon\right) \to 0 \tag{A.12}$$

for any small $\varepsilon > 0$.

By (A.11), for any small $\omega > 0$, there exists a positive constant M_0 such that

$$P\left(T_{\mathbb{C}}^{1/2}(n)\|\widehat{\gamma}_n-\gamma_0\| \leqslant M_0\right) \geqslant 1-\omega. \tag{A.13}$$

For notational simplicity, we let $\boldsymbol{\Omega}_1$ be the event $\{T_{\mathbb{C}}^{1/2}(n)\|\widehat{\gamma}_n-\gamma_0\| \leqslant M_0\}$. Conditional on $\boldsymbol{\Omega}_1$, by (A.11) and Lemma A.3 as well as the definition of $\widehat{\gamma}_n$, we have

$$\begin{aligned} &\sum_{t=1}^{n}\left[\rho\big(Y_t-m(X_t,\widehat{\gamma}_n)\big)-\rho(W_t)\right]I_t \\ =\ & -\sum_{t=1}^{n}\left[m(X_t,\widehat{\gamma}_n)-m(X_t,\gamma_0)\right]\psi(W_t)I_t + \frac{\phi_1 T_{\mathbb{C}}(n)}{2\pi_s(\mathbb{C})}(\widehat{\gamma}_n-\gamma_0)^{\tau}\boldsymbol{\Delta}_m(\gamma_0)(\widehat{\gamma}_n-\gamma_0)+o_P(1) \\ =\ & -\frac{\phi_1 T_{\mathbb{C}}(n)}{\pi_s(\mathbb{C})}(\widehat{\gamma}_n-\gamma_0)^{\tau}\boldsymbol{\Delta}_m(\gamma_0)(\widehat{\gamma}_n-\gamma_0) + \frac{\phi_1 T_{\mathbb{C}}(n)}{2\pi_s(\mathbb{C})}(\widehat{\gamma}_n-\gamma_0)^{\tau}\boldsymbol{\Delta}_m(\gamma_0)(\widehat{\gamma}_n-\gamma_0)+o_P(1) \\ =\ & -\frac{\phi_1 T_{\mathbb{C}}(n)}{2\pi_s(\mathbb{C})}(\widehat{\gamma}_n-\gamma_0)^{\tau}\boldsymbol{\Delta}_m(\gamma_0)(\widehat{\gamma}_n-\gamma_0)+o_P(1). \end{aligned} \tag{A.14}$$

Furthermore, by Lemma A.3 and the definition of $\widehat{\gamma}_n$ again, we have

$$\sup_{T_{\mathbb{C}}^{1/2}(n)\|\gamma-\gamma_0\|\leqslant M_0+\varepsilon}\left|\sum_{t=1}^{n}\rho\big(Y_t-m(X_t,\gamma)\big)I_t-\sum_{t=1}^{n}\rho(W_t)I_t+\frac{\phi_1 T_{\mathbb{C}}(n)}{\pi_s(\mathbb{C})}(\gamma-\gamma_0)^{\intercal}\Delta_m(\gamma_0)(\widehat{\gamma}_n-\gamma_0)\right.$$
$$\left.-\frac{\phi_1 T_{\mathbb{C}}(n)}{2\pi_s(\mathbb{C})}(\gamma-\gamma_0)^{\intercal}\Delta_m(\gamma_0)(\gamma-\gamma_0)\right|=o_P(1). \tag{A.15}$$

A combination of (A.14) and (A.15) leads to

$$\sup_{T_{\mathbb{C}}^{1/2}(n)\|\gamma-\widehat{\gamma}_n\|=\varepsilon}\left|\sum_{t=1}^{n}\rho\big(Y_t-m(X_t,\gamma)\big)I_t-\sum_{t=1}^{n}\rho\big(Y_t-m(X_t,\widehat{\gamma}_n)\big)I_t\right.$$
$$\left.-\frac{\phi_1 T_{\mathbb{C}}(n)}{2\pi_s(\mathbb{C})}(\widehat{\gamma}_n-\gamma)^{\intercal}\Delta_m(\gamma_0)(\widehat{\gamma}_n-\gamma)\right|=o_P(1). \tag{A.16}$$

Note that when $T_{\mathbb{C}}^{1/2}(n)\|\gamma-\widehat{\gamma}_n\|=\varepsilon$, by Assumption 3(ii), there exists a positive constant $m_0^{\star}$ such that

$$\frac{\phi_1 T_{\mathbb{C}}(n)}{2\pi_s(\mathbb{C})}(\widehat{\gamma}_n-\gamma)^{\intercal}\Delta_m(\gamma_0)(\widehat{\gamma}_n-\gamma)\geqslant m_0^{\star}\varepsilon^2. \tag{A.17}$$

Then, by (A.16), (A.17) and the convexity of the function $\rho(\cdot)$, and letting $\omega\to 0$ in (A.13), we can argue that the robust estimation $\widetilde{\gamma}_n$ lies in the interior of the circle centered at $\widehat{\gamma}_n$ with radius $\varepsilon\cdot T_{\mathbb{C}}^{-1/2}(n)$ and thus prove (A.12), completing the proof of Theorem 3.1(ii). ■

We next turn to the proof of Theorem 3.2, and start with some technical lemmas which generalise Lemmas A.1–A.3 from the integrable case to the asymptotically homogeneous case.

Lemma A.5. *Suppose that Assumptions 1, 2(i), 4 and 5(i) are satisfied. For any positive constant* M*, we have*

$$\sup_{\|\gamma-\gamma_0\|\leqslant M}\left|\sum_{t=1}^{n}V(X_t,W_t,\gamma)-\sum_{t=1}^{n}E\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right]\right|=o_P\left(T(n)\kappa^2(\zeta_n)\pi_s(\zeta_n)\right), \tag{A.18}$$

where $\mathcal{F}_t$ *is defined as in Lemma A.1,* $\kappa(\cdot)$ *is the asymptotic order of the nonlinear regression function* $m(\cdot,\cdot)$*, and* $\pi_s(\zeta_n)=\int_{-\zeta_n}^{\zeta_n}\pi_s(dx)$*.*

Lemma A.6. *Suppose that Assumptions 1, 4 and 5(i) are satisfied. For any sufficiently small* $\iota>0$*, we have*

$$\frac{1}{T(n)\kappa^2(\zeta_n)\pi_s(\zeta_n)}\sum_{t=1}^{n}E\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right]\geqslant\frac{\underline{\phi}}{2\pi_s(\zeta_n)}\int_{-\zeta_n}^{\zeta_n}\left[h\left(\frac{x}{\zeta_n},\gamma\right)-h\left(\frac{x}{\zeta_n},\gamma_0\right)\right]^2\pi_s(dx)+o_P(1) \tag{A.19}$$

uniformly over $\|\gamma-\gamma_0\|\leqslant\iota$*, where* $\underline{\phi}$ *is defined in Assumption 4(i) and* $h(\cdot,\cdot)$ *is the limit homogeneous function of the nonlinear regression function* $m(\cdot,\cdot)$*.*

Lemma A.7. *Suppose that Assumptions 1, 2, 4 and 5 are satisfied. For any positive constant* M*, we have*

$$\sup_{\|\mathbf{D}(\zeta_n,\dot{\kappa})(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\|\leqslant M}\left|\sum_{t=1}^{n}V(X_t,W_t,\boldsymbol{\gamma})-\sum_{t=1}^{n}\mathsf{E}\left[V(X_t,W_t,\boldsymbol{\gamma})|\mathcal{F}_{t-1}\right]\right|=o_P(1), \tag{A.20}$$

and

$$\sup_{\|\mathbf{D}(\zeta_n,\dot{\kappa})(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\|\leqslant M}\left|\sum_{t=1}^{n}\mathsf{E}\left[V(X_t,W_t,\boldsymbol{\gamma})|\mathcal{F}_{t-1}\right]-\frac{\phi_1}{2}(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^{\intercal}\mathbf{D}(\zeta_n,\dot{\kappa})\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)\mathbf{D}(\zeta_n,\dot{\kappa})(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)\right|=o_P(1), \tag{A.21}$$

where $\mathbf{D}(\zeta_n,\dot{\kappa})$ *is a* $d\times d$ *diagonal matrix with the* i*-th diagonal number being* $T^{1/2}(n)\dot{\kappa}_i(\zeta_n)\pi_s^{1/2}(\zeta_n)$*, and* $\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)$ *is defined in Assumption 5(iii).*

Proof of Theorem 3.2(i). The proof is very similar to the proof of Theorem 3.1(i). It is sufficient to prove (A.6) for any small $\varepsilon>0$. By (2.9) and Lemmas A.5 and A.6 above, we have that uniformly over $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\leqslant\varepsilon$,

$$\begin{aligned}
&\frac{1}{T(n)\kappa^2(\zeta_n)\pi_s(\zeta_n)}\sum_{t=1}^{n}\rho\big(Y_t-m(X_t,\boldsymbol{\gamma})\big)I_t-\frac{1}{T(n)\kappa^2(\zeta_n)\pi_s(\zeta_n)}\sum_{t=1}^{n}\rho(W_t)I_t\\
\geqslant\;&-\frac{1}{T(n)\kappa^2(\zeta_n)\pi_s(\zeta_n)}\sum_{t=1}^{n}\left[m(X_t,\boldsymbol{\gamma})-m(X_t,\boldsymbol{\gamma}_0)\right]\psi(W_t)I_t+\\
&\frac{\phi}{2\pi_s(\zeta_n)}\int_{-\zeta_n}^{\zeta_n}\left[h\left(\frac{x}{\zeta_n},\boldsymbol{\gamma}\right)-h\left(\frac{x}{\zeta_n},\boldsymbol{\gamma}_0\right)\right]^2\pi_s(dx)+o_P(1).
\end{aligned} \tag{A.22}$$

Following the proof of (A.18) in Appendix B, it is easy to show that

$$\sup_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\leqslant\varepsilon}\left|\frac{1}{T(n)\kappa^2(\zeta_n)\pi_s(\zeta_n)}\sum_{t=1}^{n}\left[m(X_t,\boldsymbol{\gamma})-m(X_t,\boldsymbol{\gamma}_0)\right]\psi(W_t)I_t\right|=o_P(1). \tag{A.23}$$

By (A.22), (A.23) and Assumption 5(i), we can prove (A.6) and then (A.7), completing the proof of Theorem 3.2(i). ■

Proof of Theorem 3.2(ii). The proof is similar to the proof of Theorem 3.1(ii). Let

$$\check{\boldsymbol{\gamma}}_n=\boldsymbol{\gamma}_0+[\phi_1\mathbf{D}(\zeta_n,\dot{\kappa})\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)\mathbf{D}(\zeta_n,\dot{\kappa})]^{-1}\sum_{t=1}^{n}\dot{m}(X_t,\boldsymbol{\gamma}_0)\psi(W_t)I_t.$$

Note that

$$\mathbf{D}(\zeta_n,\dot{\kappa})(\check{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_0)=[\phi_1\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)]^{-1}\mathbf{D}^{-1}(\zeta_n,\dot{\kappa})\sum_{t=1}^{n}\dot{m}(X_t,\boldsymbol{\gamma}_0)\psi(W_t)I_t. \tag{A.24}$$

Similarly to the proof of Lemma A.5 in Appendix B, we may show that

$$\mathbf{D}^{-1}(\zeta_n,\dot{\kappa})\left[\sum_{t=1}^{n}\dot{m}(X_t,\boldsymbol{\gamma}_0)\dot{m}^{\intercal}(X_t,\boldsymbol{\gamma}_0)I_t\right]\mathbf{D}^{-1}(\zeta_n,\dot{\kappa})=\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)+o_P(1),$$

which, together with the arguments in the proof of Lemma A.4, leads to

$$\mathbf{D}^{-1}(\zeta_n, \dot{\kappa}) \sum_{t=1}^{n} \dot{m}(X_t, \boldsymbol{\gamma}_0)\psi(W_t)I_t \xrightarrow{d} N\left(\mathbf{0}, \phi_2 \boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)\right). \tag{A.25}$$

Combining (A.24) and (A.25), we readily have that

$$\mathbf{D}(\zeta_n, \dot{\kappa})(\check{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) \xrightarrow{d} N\left(\mathbf{0}, (\phi_2/\phi_1^2)\boldsymbol{\Delta}_h^{-1}(\boldsymbol{\gamma}_0)\right). \tag{A.26}$$

By (A.26), we only need to prove

$$P\left(\|\mathbf{D}(\zeta_n, \dot{\kappa})\left(\widetilde{\boldsymbol{\gamma}}_n - \check{\boldsymbol{\gamma}}_n\right)\| > \varepsilon\right) \to 0 \tag{A.27}$$

for any small ε.

By (A.26), for any small $\omega > 0$, there exists a positive constant $M_0^\star$ such that

$$P\left(\|\mathbf{D}(\zeta_n, \dot{\kappa})\left(\check{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\right)\| \leqslant M_0^\star\right) \geqslant 1 - \omega. \tag{A.28}$$

For notational simplicity, we let $\boldsymbol{\Omega}_1^\star$ be the event $\{\|\mathbf{D}(\zeta_n, \dot{\kappa})\left(\check{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\right)\| \leqslant M_0^\star\}$. Conditional on $\boldsymbol{\Omega}_1^\star$, using (A.24) and Lemma A.7, we have

$$\begin{aligned}
&\sum_{t=1}^{n} \left[\rho\big(Y_t - m(X_t, \check{\boldsymbol{\gamma}}_n)\big) - \rho(W_t)\right] I_t \\
= \; & -\sum_{t=1}^{n} \left[m(X_t, \check{\boldsymbol{\gamma}}_n) - m(X_t, \boldsymbol{\gamma}_0)\right]\psi(W_t)I_t + \frac{\phi_1}{2}(\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\intercal \mathbf{D}(\zeta_n, \dot{\kappa})\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)\mathbf{D}(\zeta_n, \dot{\kappa})(\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + o_P(1) \\
= \; & -\frac{\phi_1}{2}(\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\intercal \mathbf{D}(\zeta_n, \dot{\kappa})\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)\mathbf{D}(\zeta_n, \dot{\kappa})(\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + o_P(1).
\end{aligned} \tag{A.29}$$

Furthermore, by (A.20) and (A.21) in Lemma A.7 as well as the definition of $\check{\boldsymbol{\gamma}}_n$ again, we have

$$\begin{aligned}
\sup_{\|\mathbf{D}(\zeta_n, \dot{\kappa})(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\| \leqslant M_0^\star + \varepsilon} \Bigg| & \sum_{t=1}^{n} \rho\big(Y_t - m(X_t, \boldsymbol{\gamma})\big) I_t - \sum_{t=1}^{n} \rho(W_t) I_t \\
& + \phi_1(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\intercal \mathbf{D}(\zeta_n, \dot{\kappa})\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)\mathbf{D}(\zeta_n, \dot{\kappa})(\check{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0) \\
& - \frac{\phi_1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\intercal \mathbf{D}(\zeta_n, \dot{\kappa})\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)\mathbf{D}(\zeta_n, \dot{\kappa})(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\Bigg| = o_P(1).
\end{aligned} \tag{A.30}$$

A combination of (A.29) and (A.30) leads to

$$\begin{aligned}
\sup_{\|\mathbf{D}(\zeta_n, \dot{\kappa})(\boldsymbol{\gamma} - \check{\boldsymbol{\gamma}}_n)\| = \varepsilon} \Bigg| & \sum_{t=1}^{n} \rho\big(Y_t - m(X_t, \boldsymbol{\gamma})\big) I_t - \sum_{t=1}^{n} \rho\big(Y_t - m(X_t, \check{\boldsymbol{\gamma}}_n)\big) I_t \\
& - \frac{\phi_1}{2}(\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\intercal \mathbf{D}(\zeta_n, \dot{\kappa})\boldsymbol{\Delta}_h(\boldsymbol{\gamma}_0)\mathbf{D}(\zeta_n, \dot{\kappa})(\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\Bigg| = o_P(1).
\end{aligned} \tag{A.31}$$

By Assumption 5(iii), there exists a positive constant $m_1^\star$ such that

$$\frac{\phi_1}{2}(\check{\gamma}-\gamma)^\intercal \mathbf{D}(\zeta_n,\dot{\kappa})\Delta_h(\gamma_0)\mathbf{D}(\zeta_n,\dot{\kappa})(\check{\gamma}-\gamma) \geqslant m_1^\star \varepsilon^2 \tag{A.32}$$

when $\|\mathbf{D}(\zeta_n,\dot{\kappa})(\gamma-\check{\gamma}_n)\| = \varepsilon$. The remaining proof is the same as that in the proof of Theorem 3.1(ii). ■

Proof of Theorem 3.3(i). We only prove the consistency result as the proof of the limit distribution theory is exactly the same as the proof of Theorem 3.2(ii) above. As in the proof of Theorem 3.1(i), we let $\mathbb{S}_{\gamma_0}(\varepsilon)$ be the circle centered at γ_0 with radius ε, and aim to prove (A.6) where ε is sufficiently small. Note that

$$\begin{aligned}\mathcal{L}_{n,\zeta_n}(\gamma)-\mathcal{L}_{n,\zeta_n}(\gamma_0) &= \sum_{t=1}^{n}\rho\big(Y_t-m(X_t,\gamma)\big)I_t-\sum_{t=1}^{n}\rho(W_t)I_t \\ &= \sum_{t=1}^{n}V(X_t,W_t,\gamma)-\sum_{t=1}^{n}\left[m(X_t,\gamma)-m(X_t,\gamma_0)\right]\psi(W_t)I_t, \end{aligned}\tag{A.33}$$

where $\psi(u)=\mathsf{sign}(u)$ and $V(X_t,W_t,\gamma)$ is defined at the beginning of this appendix. Following the proof of Lemma A.5 with minor modification, we can prove that

$$\sup_{\|\gamma-\gamma_0\|\leqslant\varepsilon}\left|\sum_{t=1}^{n}V(X_t,W_t,\gamma)-\sum_{t=1}^{n}\mathsf{E}\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right]\right| = o_P\left(T(n)\kappa(\zeta_n)\pi_s(\zeta_n)\right), \tag{A.34}$$

and

$$\sup_{\|\gamma-\gamma_0\|\leqslant\varepsilon}\left|\sum_{t=1}^{n}\left[m(X_t,\gamma)-m(X_t,\gamma_0)\right]\psi(W_t)I_t\right| = o_P\left(T(n)\kappa(\zeta_n)\pi_s(\zeta_n)\right). \tag{A.35}$$

As in Li, Tjøstheim and Gao (2016), we let

$$\mathbb{B}_i^+=[i-1,i),\ \ 1\leqslant i\leqslant\lfloor\zeta_n\rfloor,\ \ \mathbb{B}_{\lfloor\zeta_n\rfloor+1}^+=[\lfloor\zeta_n\rfloor,\zeta_n],$$

and

$$\mathbb{B}_i^-=[-i,-i+1),\ \ 1\leqslant i\leqslant\lfloor\zeta_n\rfloor,\ \ \mathbb{B}_{\lfloor\zeta_n\rfloor+1}^-=[-\zeta_n,\lfloor\zeta_n\rfloor),$$

where $\lfloor\cdot\rfloor$ denotes the floor function. Note that

$$\begin{aligned}\sum_{t=1}^{n}\mathsf{E}\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right] &= \sum_{i=1}^{\lfloor\zeta_n\rfloor+1}\sum_{t=1}^{n}\mathsf{E}\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right]\mathsf{I}\left(X_t\in\mathbb{B}_i^+\right)+ \\ &\quad \sum_{i=1}^{\lfloor\zeta_n\rfloor+1}\sum_{t=1}^{n}\mathsf{E}\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right]\mathsf{I}\left(X_t\in\mathbb{B}_i^-\right).\end{aligned}$$

As $\rho(u)=|u|$, by the triangle inequality, it is easy to show that

$$\mathsf{E}\left[V(X_t,W_t,\gamma)|\mathcal{F}_{t-1}\right] = \mathsf{E}\left\{\left[\rho\big(W_t+m(X_t,\gamma_0)-m(X_t,\gamma)\big)-\rho(W_t)\right]|X_t\right\}$$

$$\geqslant \quad |m(X_t, \gamma_0) - m(X_t, \gamma)| - 2E[|W_t|]. \tag{A.36}$$

As in the proof of (S.29) in Appendix B, we may show that

$$\sum_{i=1}^{\lfloor \zeta_n \rfloor + 1} \sum_{t=1}^{n} E[|W_t|] I\left(X_t \in \mathbb{B}_i^+\right) = O_P\left(T(n)\pi_s(\zeta_n)\right) \tag{A.37}$$

and

$$\begin{aligned} & \frac{1}{T(n)\kappa(\zeta_n)\pi_s(\zeta_n)} \sum_{i=1}^{\lfloor \zeta_n \rfloor + 1} \sum_{t=1}^{n} |m(X_t, \gamma_0) - m(X_t, \gamma)| I\left(X_t \in \mathbb{B}_i^+\right) \\ \xrightarrow{P} \quad & \frac{1}{\pi_s(\zeta_n)} \int_0^{\zeta_n} |h(x/\zeta_n, \gamma) - h(x/\zeta_n, \gamma_0)| \pi_s(dx) \end{aligned} \tag{A.38}$$

uniformly over $\|\gamma - \gamma_0\| \leqslant \varepsilon$. Then, by (A.36)–(A.38) and noting that $\kappa(\zeta_n) \to \infty$, we readily have that, uniformly over $\|\gamma - \gamma_0\| \leqslant \varepsilon$,

$$\frac{1}{T(n)\kappa(\zeta_n)\pi_s(\zeta_n)} \sum_{i=1}^{\lfloor \zeta_n \rfloor + 1} \sum_{t=1}^{n} E\left[V(X_t, W_t, \gamma)|X_t\right] I\left(X_t \in \mathbb{B}_i^+\right) \geqslant \frac{1}{\pi_s(\zeta_n)} \int_0^{\zeta_n} |h(x/\zeta_n, \gamma) - h(x/\zeta_n, \gamma_0)| \pi_s(dx) \tag{A.39}$$

with probability approaching one. Similarly, we also have

$$\frac{1}{T(n)\kappa(\zeta_n)\pi_s(\zeta_n)} \sum_{i=1}^{\lfloor \zeta_n \rfloor + 1} \sum_{t=1}^{n} E\left[V(X_t, W_t, \gamma)|X_t\right] I\left(X_t \in \mathbb{B}_i^-\right) \geqslant \frac{1}{\pi_s(\zeta_n)} \int_{-\zeta_n}^{0} |h(x/\zeta_n, \gamma) - h(x/\zeta_n, \gamma_0)| \pi_s(dx) \tag{A.40}$$

with probability approaching one.

By (A.33)–(A.35), (A.39), (A.40) and using the identification condition (3.16), we can show that (A.6) holds and thus the consistency result can be proved. ■

Proof of Theorem 3.3(ii). The proof of the consistency result is similar to that in the proof of Theorem 3.3(i). Hence, we next only sketch the main difference. Note that when $\rho(u)$ is chosen as Huber's loss function $\rho_\delta(u)$, we have

$$\psi(u) = \psi_\delta(u) = \begin{cases} u, & |u| \leqslant \delta, \\ \delta \cdot \mathrm{sign}(u), & |u| > \delta, \end{cases}$$

which is continuous and bounded. It is straightforward to prove (A.33)–(A.35). By the definition of $\rho_\delta(u)$, we readily have that

$$\rho_\delta(W_t) \leqslant \delta |W_t| \tag{A.41}$$

and

$$\rho_\delta\left(W_t + m(X_t, \gamma_0) - m(X_t, \gamma)\right) \geqslant \delta J_t |m(X_t, \gamma_0) - m(X_t, \gamma)| - \delta |W_t| - \delta^2/2, \tag{A.42}$$

where $J_t = I\left(|W_t + m(X_t, \gamma_0) - m(X_t, \gamma)| \geqslant \delta\right)$. Note that there must exist a positive constant $M_0^\diamond$ such that

$P(|W_t| < M_0^\diamond) > 1/2$. On the other hand, for γ satisfying $\|\gamma - \gamma_0\| = \varepsilon$, there exists a positive integer $\zeta_n^\diamond$ such that $\pi_s(\zeta_n^\diamond) = o(\pi_s(\zeta_n))$ and

$$|m(x, \gamma_0) - m(x, \gamma)| > \delta + M_0^\diamond, \ \ |x| \geqslant \zeta_n^\diamond. \tag{A.43}$$

Combining (A.41)–(A.43), we may show that

$$\begin{aligned} E\left[V(X_t, W_t, \gamma)|\mathcal{F}_{t-1}\right] &= E\left\{\left[\rho_\delta\left(W_t + m(X_t, \gamma_0) - m(X_t, \gamma)\right) - \rho_\delta(W_t)\right]|\mathcal{F}_{t-1}\right\} \\ &\geqslant \frac{\delta}{2}|m(X_t, \gamma_0) - m(X_t, \gamma)| - 2\delta E[|W_t|] - \delta^2 \end{aligned} \tag{A.44}$$

for $X_t \in \mathbb{B}_i^+$ or $\mathbb{B}_i^-$ with $i \geqslant \zeta_n^\diamond + 1$. By (A.37), (A.38), (A.44) and using the fact that $\pi_s(\zeta_n^\diamond) = o(\pi_s(\zeta_n))$ and $h(\cdot, \gamma)$ is locally bounded, we have that, uniformly over $\|\gamma - \gamma_0\| = \varepsilon$,

$$\begin{aligned} &\frac{1}{T(n)\kappa(\zeta_n)\pi_s(\zeta_n)} \sum_{i=1}^{\lfloor \zeta_n \rfloor + 1} \sum_{t=1}^{n} E\left[V(X_t, W_t, \gamma)|\mathcal{F}_{t-1}\right] I\left(X_t \in \mathbb{B}_i^+\right) \\ = &\frac{1}{T(n)\kappa(\zeta_n)\pi_s(\zeta_n)} \left(\sum_{i=1}^{\zeta_n^\diamond} + \sum_{\zeta_n^\diamond + 1}^{\lfloor \zeta_n \rfloor + 1}\right) \sum_{t=1}^{n} E\left[V(X_t, W_t, \gamma)|\mathcal{F}_{t-1}\right] I\left(X_t \in \mathbb{B}_i^+\right) \\ = &\frac{1}{T(n)\kappa(\zeta_n)\pi_s(\zeta_n)} \sum_{\zeta_n^\diamond + 1}^{\lfloor \zeta_n \rfloor + 1} \sum_{t=1}^{n} E\left[V(X_t, W_t, \gamma)|\mathcal{F}_{t-1}\right] I\left(X_t \in \mathbb{B}_i^+\right) + O_P(\pi_s(\zeta_n^\diamond)/\pi_s(\zeta_n)) \\ \geqslant &\frac{\delta}{2\pi_s(\zeta_n)} \int_{\zeta_n^\diamond}^{\zeta_n} |h(x/\zeta_n, \gamma) - h(x/\zeta_n, \gamma_0)|\pi_s(dx) + o_P(1) \\ = &\frac{\delta}{2\pi_s(\zeta_n)} \int_{0}^{\zeta_n} |h(x/\zeta_n, \gamma) - h(x/\zeta_n, \gamma_0)|\pi_s(dx) + o_P(1). \end{aligned} \tag{A.45}$$

Similarly, we also have

$$\frac{1}{T(n)\kappa(\zeta_n)\pi_s(\zeta_n)} \sum_{i=1}^{\lfloor \zeta_n \rfloor + 1} \sum_{t=1}^{n} E\left[V(X_t, W_t, \gamma)|\mathcal{F}_{t-1}\right] I\left(X_t \in \mathbb{B}_i^-\right) \geqslant \frac{\delta}{2\pi_s(\zeta_n)} \int_{-\zeta_n}^{0} |h(x/\zeta_n, \gamma) - h(x/\zeta_n, \gamma_0)|\pi_s(dx) \tag{A.46}$$

with probability approaching one. The remaining proof of the consistency result is the same as that in the proof of Theorem 3.3(i). The proof of the limit distribution theory is the same as the proof of Theorem 3.2(ii), so details are omitted here. ■

References

BAI, Z., RAO, C. R. AND WU, Y. (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica* **2**, 237–254.

BODEN, T. A., MARLAND, G. AND ANDRES, R.J. (2009). Glabal, regional and national fossile fuel CO emissions. Carbon Dioxide Information analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., USA.

CAI, B., GAO, J. AND TJØSTHEIM, D. (2017). A new class of bivariate threshold cointegration models. *Journal of Business and Economics Statistics* **35**, 288–305.

CHAN, N. AND WANG, Q. (2015). Nonlinear cointegrating regressions with nonstationary time series. *Journal of Econometrics* **185**, 182–195.

CHEN, X. AND CHRISTENSEN, T. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics* **188**,447–465.

CHEN, J., GAO, J. AND LI, D. (2012). Estimation in semi-parametric regression with non-stationary regressors. *Bernoulli* **18**, 678–702.

DINDA, S. (2004). Environmental Kuznets curve hypothesis: A survey. *Ecological Economics* **49**, 431–455.

DONG, C., GAO, J. AND TJØSTHEIM, D. (2016). Estimation for single-index and partially linear single-index integrated models. *Annals of Statistics* **44**, 425–453.

GAO, J., KANAYA, S., LI, D. AND TJØSTHEIM, D. (2015). Uniform consistency for nonparametric estimators in null recurrent time series. *Econometric Theory* **31**, 911–952.

GAO, J., TJØSTHEIM, D. AND YIN, J. (2013). Estimation in threshold autoregressive models with a stationary and a unit root regime. *Journal of Econometrics* **172**, 1–13.

HÄRDLE, W., HALL, P. AND ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics* **21**, 157–178.

HANSEN, B. (2015). The integrated mean squared error of series regression and a Rosenthal Hilbert-space inequality. *Econometric Theory* **31**, 337–361.

HE, X. AND SHAO, Q. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Annals of Statistics* **24**, 2608–2630.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73–101.

HUBER, P. J. (1981). *Robust Statistics*. New York: John Wiley.

JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimation. *Annals of Mathematical Statistics* **40**, 633–643.

KASAHARA, Y. (1984). Limit theorems for Lévy processes and Poisson point processes and their applications to Brownian excursions. *J. Math. Kyoto Univ.* **24**, 521–538.

KARLSEN, H. A., MYKLEBUST, T. AND TJØSTHEIM, D. (2007). Nonparametric estimation in a nonlinear cointegration type model. *Annals of Statistics* **35**, 252–299.

KARLSEN, H. A., MYKLEBUST, T. AND TJØSTHEIM D. (2010). Nonparametric regression estimation in a null recurrent time series. *Journal of Statistical Planning and Inference* **140**, 3619–3626.

KARLSEN, H. A. AND TJØSTHEIM, D. (2001). Nonparametric estimation in null recurrent time series. *Annals of Statistics* **29**, 372–416.

KNIGHT, K. (1991). Limit theory for M-estimates in an integrated infinite variance process. *Econometric Theory* **7**, 200-212.

KNIGHT, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Annals of Statistics* **26**, 755–770.

KOENKER, R. (2005). *Quantile Regression*. Cambridge University Press.

LAI, T. L. (1994). Asymptotic properties of nonlinear least squares estimations in stochastic regression models. *Annals of Statistics* **22**, 1917–1930.

LI, D. AND LI, R. (2016). Local composite quantile regression smoothing for Harris recurrent Markov processes. *Journal of Econometrics* **194**, 44–56.

LI, D., TJØSTHEIM, D. AND GAO J. (2016). Estimation in nonlinear regression with Harris recurrent Markov chains. *Annals of Statistics* **44**, 1957–1987.

LIN, Z., LI, D. AND CHEN, J. (2009). Local linear M–estimators in null recurrent time series. *Statistica Sinica* **19**, 1683–1703.

LING, S. (2007). Self-weighted and local quasi-maximum likelihood estimators for ARMA-GARCH/IGARCH models. *Journal of Econometrics* **140**, 849–873.

MADDALA, G. S. AND KIM, I. (1998). *Unit Roots, Cointegration, and Structural Change*. Cambridge University Press.

MALINVAUD, E. (1970). The consistency of nonlinear regressions. *Annals of Mathematical Statistics* **41**, 956–969.

MEYN, S. AND TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability* (2nd Edition). Cambridge University Press.

MYKLEBUST, T., KARLSEN, H. A. AND TJØSTHEIM, D. (2012). Null recurrent unit root processes. *Econometric Theory* **28**, 1–41.

NUMMELIN, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press.

PARK, J. AND PHILLIPS, P. C. B. (2001). Nonlinear regressions with integrated time series. *Econometrica* **69**, 117–162.

PIAGGIO, M. AND PADILLA, E. (2012). CO_2 emissions and economic activity: Heterogeneity across countries and non-stationary series. *Energy Policy* **46**, 370–381.

PHILLIPS, P. C. B. (1995). Robust nonstationary regression. *Econometric Theory* **11**, 912–951.

SCHIENLE, M. (2011). Nonparametric nonstationary regression with many covariates. *Unpublished Manuscript*.

SEVERINI, T. A. AND WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics* **20**, 1768–1802.

SINHA, S. K., FIELD, C. A. AND SMITH, B. (2003). Robust estimation of nonlinear regression with autoregressive errors. *Statistics and Probability Letters* **63**, 49–59.

SKOURAS, K. (2000). Strong consistency in nonlinear stochastic regression models. *Annals of Statistics* **28**, 871–879.

TJØSTHEIM, D. (2020). Some notes on nonlinear cointegration: A partial review with some novel perspectives. Forthcoming in *Econometric Reviews*.

WANG,Q., WU D. AND ZHU, K. (2018). Model checks for nonlinear cointegrating regression. *Journal of Econometrics* **207**, 261–284.

WU, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics* **9**, 501–513.

XIA, Y., TONG, H., LI, W. K. AND ZHU L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B* **64**, 363–410.

ZOU, H. AND YUAN, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics* **36**, 1108–1126.