Online Supplement for "Bootstrap Prediction Inference of Non-linear Autoregressive Models"

Kejin Wu[1] and Dimitris N. Politis [*,2]

[1]Department of Mathematics, University of California, San Diego
[2]Department of Mathematics and Halicioğlu Data Science Institute, University of California, San Diego

The supplement contains three parts. One set of sufficient conditions for geometric ergodicity of

$$X_t = \phi(\boldsymbol{X_{t-1}}) + \sigma(\boldsymbol{X_{t-1}})\epsilon_t \tag{1}$$

are discussed in Appendix A. The application of the forward bootstrap prediction algorithm on the general NLAR model

$$X_t = G(\boldsymbol{X_{t-1}}, \epsilon_t)$$

is presented in Appendix B. The proofs for the lemmas and theorems in the main text are given in Appendix C.

Appendix A Sufficient conditions for geometric ergodicity

Checking geometric ergodicity for a LAR model is simple; it is well known that the LAR model is stationary and geometrically ergodic as long as the corresponding characteristic polynomial does not have zero roots inside or on the unit circle. However, this check criterion depends on the linearity assumption, and can not be extended to serve for NLAR model directly An and Huang (1996). Thus, practitioners rely on Markov chain techniques to explore conditions under which the NLAR model is geometrically ergodic. The motivation is that the NLAR model can be described as a Markov chain in a general state space; the extensive discussions and literature related to Markov chains can guide the development of criteria to check the ergodicity of NLAR models.

One of the earliest criteria developed to guarantee the ergodicity of a Markov chain is Doeblin's condition given by Doob (1953). Later, Tweedie (1975) proposed a more generalized condition, the so-called *Drift criterion*. This criterion gives a sufficient condition for an aperiodic and irreducible Markov chain to be geometrically ergodic. For completeness, we present this criterion using the version applied by An and Huang (1996):

*Correspondence to: Dimitris N. Politis. Email: dpolitis@ucsd.edu.

Lemma A.1 (Drift criterion)**.** *Let $\{X_t\}$ be an aperiodic and irreducible Markov chain. Suppose that there exists a small set C, a non-negative measurable function k, positive constants c_1, c_2 and $\rho < 1$ such that:*

$$\begin{aligned} &\mathbb{E}(k(\boldsymbol{X_{t+1}})|\boldsymbol{X_t} = \boldsymbol{x}) \leq \rho k(\boldsymbol{x}) - c_1, \text{ for any } \boldsymbol{x} \notin C, \\ &\mathbb{E}(k(\boldsymbol{X_{t+1}})|\boldsymbol{X_t} = \boldsymbol{x}) \leq c_2, \text{ for any } \boldsymbol{x} \in C. \end{aligned} \tag{2}$$

Remark. *The function $k(\cdot)$ is called the test function in the literature. For the formal definition of a 'small set'; see Tjøstheim (1990) or Tong (1990). We will soon see that the small set can be taken as a compact set in some situations.*

Thus, to ensure the ergodicity of an NLAR model, people can check if $\{X_t\}$ is aperiodic and irreducible and Lemma A.1 holds. Along with this idea, An and Huang (1996) give several kinds of sufficient conditions for Eq. (1) with $\sigma(\boldsymbol{X_{t-1}}) \equiv 1$ (homoscedastic errors case) to be geometrically ergodic. Note that the test function $k(\cdot)$ and the specific small set will change according to which condition is based. Then, Min and Hongzhi (1999) extend these results to the region of NLAR models with heteroscedastic errors. Based on this body of work, if we assume:

A4 The probability density function of innovation $f_\epsilon(\cdot)$ is continuous and everywhere positive.

A5 The conditional mean and volatility functions satisfy the inequalities:

$$\sup_{||\boldsymbol{x}||_2 \leq K} |\phi(\boldsymbol{x})| < \infty \; ; \quad \sup_{||\boldsymbol{x}||_2 \leq K} |\sigma(\boldsymbol{x})| < \infty, \text{ for each } K > 0, \tag{3}$$

where $\boldsymbol{x} \in \mathbb{R}^p$, and $||\cdot||_2$ is the Euclidean norm.

we can obtain a useful Lemma:

Lemma A.2. *Let $\{X_t\}$ satisfy the model Eq. (1). Suppose A4 and A5 are fulfilled. Then $\{X_t\}$ is aperiodic and irreducible with respect to μ which is the Lebesgue measure. Moreover, the μ-non-null compact sets are small sets.*

Remark A.1. *The proof of Lemma A.2 can be found in Min and Hongzhi (1999). The original proof only requires the density function to be lower semi-continuous. In this paper, since the estimation of the innovation distribution and the L_1 optimal predictor will be discussed, we require a stronger condition for the density function. Besides, we should notice that we can check the classical properties defined by the Markov chain to verify the aperiodicity and irreducibility. Nevertheless, it is more natural to apply Lemma A.2 for analyzing NLAR models. For example, when we consider a homoscedastic NLAR model with $p = 1$, the drift criterion and the boundedness of the conditional mean function on a suitable compact set can be satisfied by a simple condition: $|\phi(x)| \leq C_2|x| + C_1$, for all x and some $C_1 < \infty, C_2 < 1$; this is exactly Assumption 3 (i) of Franke, Kreiss, et al. (2002). In their work, they mentioned that the everywhere-positive assumption (A4) on the density function is unnecessarily restrictive. Thus, they replace it with their Assumption 3 (iii). However, for simplifying the proofs of our paper, we still require the everywhere-positive property. Furthermore, we will apply a convolutional technique to acquire an everywhere-positive "innovation distribution" in the bootstrap world to satisfy this strong assumption.*

The condition which ensures the Drift criterion is satisfied for a homoscedastic NLAR models can be drawn from An and Huang (1996). If we assume:

A6 There exists a positive number $\lambda < 1$ and a constant C such that the conditional mean function satisfies:

$$|\phi(\boldsymbol{x})| \leq \lambda \max\{|x_1|, \ldots, |x_p|\} + C, \tag{4}$$

then we can get the below lemma:

Lemma A.3 (Theorem 3.2 of An and Huang (1996)). *Let $\{X_t\}$ satisfy the homoscedastic version of the model Eq. (1). If A4 and A6 are fulfilled, then this NLAR model is geometrically ergodic.*

Similarly, we can obtain the condition for heteroscedastic NLAR models to be geometrically ergodic by imposing an additional assumption on the variance function:

A7 The conditional variance function satisfies:

$$\lim_{||\boldsymbol{x}||_2 \to \infty} \frac{\sigma(\boldsymbol{x})}{||\boldsymbol{x}||_2} = 0. \tag{5}$$

Then the Lemma holds:

Lemma A.4 (Theorem 3.5 of Min and Hongzhi (1999)). *Let $\{X_t\}$ satisfy the model Eq. (1). Suppose the conditional variance function satisfies A5 and A7. In addition, A6 holds true for the conditional mean function, and A4 holds true for the probability density function. Then, this heteroscedastic NLAR model is geometrically ergodic.*

Appendix B Forward bootstrap prediction of general NLAR models

Similar to the setup in the main context, we suppose that we observe $T + p$ number of real-valued samples $\{X_{-p+1}, X_{-p+2}, \ldots, X_T\}$, but these samples were generated from the below general ergodic NLAR model:

$$X_t = G(\boldsymbol{X_{t-1}}, \epsilon_t); \tag{6}$$

here $\{\epsilon_t\}$ is still assumed to be *i.i.d.* with mean zero, and $\boldsymbol{X_{t-1}}$ represents vector $\{X_{t-1}, \ldots, X_{t-p}\}$; $G(\cdot, \cdot)$ can be any continuous (possibly non-linear) function that makes the variance and mean of $\{X_t\}$ finite. The problem is how can we make multi-step ahead prediction inferences with such a complicated model. As different from the location-scale model in the main context, it may be hard to estimate F_ϵ by residuals explicitly in practice. Thus, we present forward bootstrap algorithms for two ideal cases: (a) $G(\cdot, \cdot)$ and F_ϵ are known; (b) $G(\cdot, \cdot)$ and F_ϵ are unknown, but they can be estimated consistently.

B.1 Simulation-based prediction

First, consider an idealized situation where the model $G(\boldsymbol{X_{t-1}}, \epsilon_t)$ and the distribution of innovations F_ϵ are known. In this case, we propose to approximate the h-step ahead prediction by simulating innovations from F_ϵ and plugging them into the NLAR model. To describe the idea, focus on the

2-step ahead prediction, and note that the distribution of the future value X_{T+2} (conditional on the observed $\boldsymbol{X_T}$) is identical to the distribution of

$$G(G(\boldsymbol{X_T}, \epsilon^*_{T+1}), \epsilon^*_{T+2}),$$

where ϵ^*_{T+1} and ϵ^*_{T+2} are $i.i.d.\sim F_\epsilon$. Going to the h-step ahead prediction, the distribution of the future value X_{T+h} (conditional on the observed $\boldsymbol{X_T}$) is identical to the distribution of the quantity

$$G(\cdots G(G(G(\boldsymbol{X_T}, \epsilon^*_{T+1}), \epsilon^*_{T+2}), \epsilon^*_{T+3}), \ldots, \epsilon^*_{T+h}), \tag{7}$$

where $\epsilon^*_{T+1}, \ldots, \epsilon^*_{T+h}$ are $i.i.d.\sim F_\epsilon$.

Of course, in order to obtain the L_2 or L_1 optimal point predictor, we would need to approximate the mean or median of the quantity (7). We can do this by Monte Carlo (MC) simulation; the simulation will be based on M replicates of the quantity (7); these are denoted $\{X^{(m)}_{T+h}\}_{m=1}^M$. Then, the L_2 or L_1 optimal predictor of X_{T+h} can be approximated by the mean or median, respectively, of $\{X^{(1)}_{T+h}, \ldots, X^{(M)}_{T+h}\}$; the empirical distribution of the values $\{X^{(1)}_{T+h}, \ldots, X^{(M)}_{T+h}\}$ can also be used to approximate the distribution of the future value X_{T+h} (conditional on the observed $\boldsymbol{X_T}$), leading to the construction of asymptotically valid PIs as $M \to \infty$.

B.2 Bootstrap-based prediction

In the more realistic scenario, both $G(\cdot,\cdot)$ and F_ϵ are unknown but we assume that they can be estimated from the data at hand; denote their estimators by $\widehat{G}(\cdot,\cdot)$ and the empirical distribution of residuals by $\widehat{F}_\epsilon$, respectively, and assume they are consistent. Conducting a simulation as described in the previous subsection using $\widehat{G}(\cdot,\cdot)$ and $\widehat{F}_\epsilon$ in place of the unknown $G(\cdot,\cdot)$ and F_ϵ turns the MC simulation into a *bootstrap* procedure. The bootstrap version of the quantity (7) is now given by

$$\widehat{G}(\cdots \widehat{G}(\widehat{G}(\widehat{G}(\boldsymbol{X_T}, \hat{\epsilon}^*_{T+1}), \hat{\epsilon}^*_{T+2}), \hat{\epsilon}^*_{T+3}), \ldots, \hat{\epsilon}^*_{T+h}), \tag{8}$$

where $\{\hat{\epsilon}^*_t\}_{t=T+1}^{T+h}$ are $i.i.d.\sim \widehat{F}_\epsilon$; $\widehat{G}(\cdot,\cdot)$ is an estimator to the true model. Next, we can take a similar approach as the simulation-based method previously described to approximate the L_1 or L_2 optimal predictor of X_{T+h}. Also, the asymptotically valid PI can be determined as $T \to \infty$ and $M \to \infty$. However, such a QPI will suffer from finite-sample undercoverage since the variability of estimating the model is not taken into account. We will prefer constructing a PPI instead that possesses a stronger property than the asymptotic validity. The setup of PPI for general NLAR models is similar to the procedure explained by Algorithm 2 in the main text.

Appendix C Proofs

Proof of Theorem 2.1. Denote the exactly L_2 optimal predictor of X_{T+h} by $X^{L_2}_{T+h}$. We know it is a conditional mean given observed data:

$$X^{L_2}_{T+h} = \mathbb{E}(X_{T+h} | X_T, \ldots, X_{T-p+1}). \tag{9}$$

Due to the strong stationary property, we can rewrite Eq.(10) in the main text as $X_{T+h} = f(\boldsymbol{Y}, \boldsymbol{\epsilon_h})$, where $\boldsymbol{Y}$ and $\boldsymbol{\epsilon_h}$ represent $\{X_t\}_{t=1}^p$ and $\{\epsilon_t\}_{t=T+1}^{T+h}$, respectively. Moreover, by the causality assumption,

we get $\boldsymbol{Y}$ and $\boldsymbol{\epsilon_h}$ are independent. In addition, $\{\boldsymbol{\epsilon_h}^{(i)}\}_{i=1}^{M}$ are also *i.i.d.*. Thus, $\{X_{T+h}^{(i)}\}_{i=1}^{M}$ are conditionally *i.i.d.* given $\boldsymbol{Y}$. Based on Theorem 4.2 of Majerek, Nowak, and Zieba (2005), the conditional version of the strong law of large numbers implies that:

$$\widehat{X_{T+h}^{L_2}} = \frac{1}{M}\sum_{i=1}^{M} X_{T+h}^{(i)} \xrightarrow{a.s.} X_{T+h}^{L_2}, \text{assuming that } X_{T+h}^{L_2} \text{ exits.} \tag{10}$$

The existence of $X_{T+h}^{L_2}$ is guaranteed by assumptions A2 and A5 - A7. This proof can be directly extended to the NLAR model with heteroscedastic errors, since the relationship $X_{T+h} = f(\boldsymbol{Y}, \boldsymbol{\epsilon_h})$ is also satisfied, so $\{X_{T+h}^{(i)}\}_{i=1}^{M}$ are still conditionally *i.i.d.* given $\boldsymbol{Y}$. Therefore, without changing other parts of this proof, we can show the analogous theorem for NLAR models with heteroscedastic error cases. □

Proof of Theorem 2.2. We actually want to show that under the ergodic property:

$$\widehat{X_{T+h}^{L_1}} \xrightarrow{p} X_{T+h}^{L_1}, \text{as } M \text{ converges to infinity.} \tag{11}$$

We can write $\widehat{X_{T+h}^{L_1}}$ and $X_{T+h}^{L_1}$ as:

$$\widehat{X_{T+h}^{L_1}} = H_M(\boldsymbol{Y}, \mathscr{E}_{\boldsymbol{h}}) = \text{Median}(f(\boldsymbol{Y}, \boldsymbol{\epsilon_h}^{(1)}), \ldots, f(\boldsymbol{Y}, \boldsymbol{\epsilon_h}^{(M)})) \ ; \ X_{T+h}^{L_1} = H(\boldsymbol{Y}, \boldsymbol{\epsilon_h}) = Q_{X_{T+h}|\boldsymbol{X}_T}(1/2), \tag{12}$$

where $\boldsymbol{Y}$ and $\boldsymbol{\epsilon_h}$ represent $\{X_t\}_{t=1}^{p}$ and $\{\epsilon_t\}_{t=T+1}^{T+h}$, respectively; $\mathscr{E}_{\boldsymbol{h}}$ represents the whole set $\{\boldsymbol{\epsilon_h}^{(i)}\}_{i=1}^{M}$. $Q_{X_{T+h}|\boldsymbol{X}_T}$ is the conditional quantile function of X_{T+h}. By assumption A8, $f(\cdot,\cdot)$ is also uniformly continuous in $\boldsymbol{x}$ since it is a composition of uniformly continuous functions. Thus, for a given $\eta > 0$, there exists a constant $\delta > 0$ such that:

$$|f(\boldsymbol{y_1}, \boldsymbol{\epsilon_h}) - f(\boldsymbol{y_2}, \boldsymbol{\epsilon_h})| < \eta, \text{ when } ||\boldsymbol{y_1} - \boldsymbol{y_2}|| \leq \delta, \tag{13}$$

where $||\cdot||$ is any norm equivalent to the Euclidean norm.

Then, we can split the p-dimensional ball $B(D) = \{\boldsymbol{y}, ||\boldsymbol{y}|| \leq D\}$ into some disjoint subsets S_j, $j = 1, \ldots, k$. Let $||\boldsymbol{y} - \boldsymbol{s_j}|| < \delta$ for $\boldsymbol{s_j} \in S_j$ and any point $\boldsymbol{y} \in B(D)$ such that:

$$|f(\boldsymbol{y}, \boldsymbol{\epsilon_h}) - f(\boldsymbol{s_i}, \boldsymbol{\epsilon_h})| < \eta, \forall \eta > 0. \tag{14}$$

Thus, $|f(\boldsymbol{y}, \boldsymbol{\epsilon_h}^{(i)}) - f(\boldsymbol{s_j}, \boldsymbol{\epsilon_h}^{(i)})| < \eta$, $\forall i \in 1, \ldots, M$. It is possible to fix a small enough η to make sure that the order of $\{f(\boldsymbol{y}, \boldsymbol{\epsilon_h}^{(1)}), \ldots, f(\boldsymbol{y}, \boldsymbol{\epsilon_h}^{(M)})\}$ is same with the order of $\{f(\boldsymbol{s_j}, \boldsymbol{\epsilon_h}^{(1)}), \ldots, f(\boldsymbol{s_j}, \boldsymbol{\epsilon_h}^{(M)})\}$. In other words, we have:

$$\sum_{j=1}^{k} I_j |H_M(\boldsymbol{Y}, \mathscr{E}_{\boldsymbol{h}}) - H_M(\boldsymbol{s_j}, \mathscr{E}_{\boldsymbol{h}})| < \eta, \tag{15}$$

where I_j represents the indicator function $I(\boldsymbol{Y} \in S_j)$, $j = 1, \ldots, k$. In addition, we also have:

$$\sum_{j=1}^{k} I_j |H(\boldsymbol{Y}, \boldsymbol{\epsilon_h}) - H(\boldsymbol{s_j}, \boldsymbol{\epsilon_h})| < \eta. \tag{16}$$

Therefore, define $I_0 := I(\boldsymbol{Y} \notin B(D))$, by combining Eqs. (15) and (16), we can get:

$$
\begin{aligned}
&|H_M(\boldsymbol{Y}, \mathscr{E}_{\boldsymbol{h}}) - H(\boldsymbol{Y}, \boldsymbol{\epsilon}_{\boldsymbol{h}})| \\
&= \sum_{j=1}^{k} I_j \, |H_M(\boldsymbol{Y}, \mathscr{E}_{\boldsymbol{h}}) - H_M(\boldsymbol{s_j}, \mathscr{E}_{\boldsymbol{h}}) + H_M(\boldsymbol{s_j}, \mathscr{E}_{\boldsymbol{h}}) - H(\boldsymbol{s_j}, \boldsymbol{\epsilon}_{\boldsymbol{h}}) + H(\boldsymbol{s_j}, \boldsymbol{\epsilon}_{\boldsymbol{h}}) - H(\boldsymbol{Y}, \boldsymbol{\epsilon}_{\boldsymbol{h}})| + \\
&I_0 |H_M(\boldsymbol{Y}, \mathscr{E}_{\boldsymbol{h}}) - H_M(\boldsymbol{s_0}, \mathscr{E}_{\boldsymbol{h}}) + H_M(\boldsymbol{s_0}, \mathscr{E}_{\boldsymbol{h}}) - H(\boldsymbol{s_0}, \boldsymbol{\epsilon}_{\boldsymbol{h}}) + H(\boldsymbol{s_0}, \boldsymbol{\epsilon}_{\boldsymbol{h}}) - H(\boldsymbol{Y}, \boldsymbol{\epsilon}_{\boldsymbol{h}})| \\
&\leq \sum_{j=1}^{k} I_j \left(|H_M(\boldsymbol{Y}, \mathscr{E}_{\boldsymbol{h}}) - H_M(\boldsymbol{s_j}, \mathscr{E}_{\boldsymbol{h}})| + |H_M(\boldsymbol{s_j}, \mathscr{E}_{\boldsymbol{h}}) - H(\boldsymbol{s_j}, \boldsymbol{\epsilon}_{\boldsymbol{h}})| + |H(\boldsymbol{s_j}, \boldsymbol{\epsilon}_{\boldsymbol{h}}) - H(\boldsymbol{Y}, \boldsymbol{\epsilon}_{\boldsymbol{h}})| \right) + I_0 \cdot C \\
&\leq 2\eta + I_0 \cdot C + \sum_{j=1}^{k} I_j |H_M(\boldsymbol{s_j}, \mathscr{E}_{\boldsymbol{h}}) - H(\boldsymbol{s_j}, \boldsymbol{\epsilon}_{\boldsymbol{h}})|.
\end{aligned}
\tag{17}
$$

Comparing $H_M(\boldsymbol{s_j}, \mathscr{E}_{\boldsymbol{h}})$ and $H(\boldsymbol{s_j}, \boldsymbol{\epsilon}_{\boldsymbol{h}})$, where $\boldsymbol{s_j}$ is a fixed point, by applying the CLT on the sample median for the ergodic series, we can get $H_M(\boldsymbol{s_j}, \mathscr{E}_{\boldsymbol{h}})$ converges to $H(\boldsymbol{s_j}, \boldsymbol{\epsilon}_{\boldsymbol{h}})$ in probability. The non-zero property of the probability density of X_{n+h} at the median is guaranteed by the everywhere positive density function of innovation. Thus, we have:

$$
\mathbb{P}(|H_M(\boldsymbol{Y}, \mathscr{E}_{\boldsymbol{h}}) - H(\boldsymbol{Y}, \boldsymbol{\epsilon}_{\boldsymbol{h}})| \leq 2\eta + I_0 \cdot C) \to 1. \tag{18}
$$

Besides, D can be arbitrarily large. Also, η can be arbitrarily small. Finally, we get:

$$
\mathbb{P}(|H_M(\boldsymbol{Y}, \mathscr{E}_{\boldsymbol{h}}) - H(\boldsymbol{Y}, \boldsymbol{\epsilon}_{\boldsymbol{h}})| \leq \varepsilon) \to 1. \tag{19}
$$

The above proof can be extended to other quantile estimators. Thus, we can build asymptotically valid QPI with any CVR. For extending such proof to NLAR models with heteroscedastic errors, we need the variance function is also uniformly continuous. Then, $f(\cdot, \cdot)$ is still uniformly continuous in $\boldsymbol{x}$. Therefore, without changing other parts of this proof, we can show the analogous theorem for NLAR models with heteroscedastic error cases. □

Proof of Lemma 2.1. To simplify the notation, we just consider the NLAR model in Eq.(11) of the main text with order 1 and homoscedastic errors. The case with higher order and heteroscedastic errors can be proved similarly.

The proof is inspired by the work of Boldin (1983). For connecting $F_\epsilon(x)$ and $\widehat{F}_\epsilon(x)$, we should notice that the empirical distribution $F_T(x)$ based on $\{\epsilon_i\}_{i=1}^T$ can be a bridge, i.e., we have:

$$
\begin{aligned}
\sup_x |\widehat{F}_\epsilon(x) - F_\epsilon(x)| &= \sup_x |\widehat{F}_\epsilon(x) - F_T(x) + F_T(x) - F_\epsilon(x)| \\
&\leq \sup_x |\widehat{F}_\epsilon(x) - F_T(x)| + \sup_x |F_T(x) - F_\epsilon(x)|.
\end{aligned}
\tag{20}
$$

From the Glivenko–Cantelli theorem, we know $\sup_x |F_T(x) - F_\epsilon(x)|$ converges to 0 a.s.. Thus, we only need to show:

$$
\sup_x |\widehat{F}_\epsilon(x) - F_T(x)| \xrightarrow{p} 0. \tag{21}
$$

First, we know

$$
\text{if } \Delta_i(x) = \begin{cases} 1, & \epsilon_i \leq x \\ 0, & \epsilon_i > x \end{cases} \text{ then, } F_T(x) = \frac{1}{T} \sum_{i=1}^{T} \Delta_i(x). \tag{22}
$$

Thus, we can get:

$$\widehat{F}_\epsilon(x) = \frac{1}{T}\sum_{i=1}^{T}\Delta_i(x + \phi(X_{i-1}, \widehat{\theta}_1) - \phi(X_{i-1}, \theta_1)), \tag{23}$$

since $\hat{\epsilon}_i = \phi(X_{i-1}, \theta_1) - \phi(X_{i-1}, \widehat{\theta}_1) + \epsilon_i$. For handling the randomness of $\phi(X_{i-1}, \widehat{\theta}_1) - \phi(X_{i-1}, \theta_1)$ inside $\Delta_i(\cdot)$ of Eq. (23), we use nonrandom $\eta_T, T = 1, 2, 3, \ldots,$ to replace $\phi(X_{i-1}, \widehat{\theta}_1) - \phi(X_{i-1}, \theta_1)$. Then, we can consider the process:

$$z_T(x, \eta_T) = \widehat{F}_\epsilon(x) - F_T(x) = \frac{1}{T}\sum_{i=1}^{T}\left(\Delta_i(x + \eta_T) - \Delta_i(x)\right). \tag{24}$$

Indeed, we have:

$$\mathbb{P}(\sup_x |\widehat{F}_\epsilon(x) - F_T(x)| > \epsilon) \leq \mathbb{P}(\sup_x \sup_{|\eta_T| \leq T^{-\lambda}} |z_T(x, \eta_T)| > \epsilon) + \mathbb{P}(|\phi(X_{i-1}, \widehat{\theta}_1) - \phi(X_{i-1}, \theta_1)| > T^{-\lambda}). \tag{25}$$

Without loss of generality, we select an appropriate λ to make sure the second term on the right-hand side of the above inequality converges to 0 under A9 and A10.

Then, we shall show $\mathbb{P}(\sup_x \sup_{|\eta_T| \leq T^{-\lambda}} |z_T(x, \eta_T)| > \epsilon)$ also converges to 0. Since this term depends on the continuum of values of x, we can partition the real axis into $N_T \sim T^{1/2}$ parts by points:

$$-\infty = x_0 < x_1 < \cdots x_k < \cdots < x_{N_T - 1} < x_{N_T} = \infty, \text{ where } F_\epsilon(x_k) = kN_T^{-1}. \tag{26}$$

Hence, for x_r and x_{r+1} such that $x_r \leq x \leq x_{r+1}$, we have:

$$x_r + \eta_T \leq x + \eta_T \leq x_{r+1} + \eta_T. \tag{27}$$

In addition, since $\Delta_i(x)$ is monotonic, we obtain:

$$\begin{aligned} z_T(x, \eta_T) &\geq z_T(x_r, \eta_T) + \frac{1}{T}\sum_{i=1}^{T}\Delta_i(x_r) - \frac{1}{T}\sum_{i=1}^{T}\Delta_i(x_{r+1}); \\ z_T(x, \eta_T) &\leq z_T(x_{r+1}, \eta_T) + \frac{1}{T}\sum_{i=1}^{T}\Delta_i(x_{r+1}) - \frac{1}{T}\sum_{i=1}^{T}\Delta_i(x_r). \end{aligned} \tag{28}$$

Therefore, we have:

$$\begin{aligned} &\sup_x \sup_{|\eta_T| \leq T^{-\lambda}} |z_T(x, \eta_T)| \\ &\qquad \leq \sup_{k \leq N_T - 1} \sup_{|\eta_T| \leq T^{-\lambda}} |z_T(x_{k+1}, \eta_T)| \\ &\qquad + \sup_{k \leq N_T} \sup_{|\eta_T| \leq T^{-\lambda}} |z_T(x_k, \eta_T)| \\ &\qquad + \sup_{|t_1 - t_2| \leq N_T^{-1}} \frac{1}{T}\left|\sum_{i=1}^{T}\left(\Delta_i(F_\epsilon^{-1}(t_1)) - \Delta_i(F_\epsilon^{-1}(t_2))\right)\right|. \end{aligned} \tag{29}$$

For the last term on the r.h.s. of Eq. (29):

$$
\begin{aligned}
&\sup_{|t_1-t_2|\le N_T^{-1}} \frac{1}{T}\left|\sum_{i=1}^{T}\left(\Delta_i(F_\epsilon^{-1}(t_1))-\Delta_i(F_\epsilon^{-1}(t_2))\right)\right| \\
&= \sup_{|t_1-t_2|\le N_T^{-1}} \frac{1}{T}\left|\sum_{i=1}^{T}\left(\Delta_i(F_\epsilon^{-1}(t_1))-t_1+t_1-t_2+t_2-\Delta_i(F_\epsilon^{-1}(t_2))\right)\right| \\
&\le \sup_{t_1,s.t.|t_1-t_2|\le N_T^{-1}}\left|\frac{1}{T}\sum_{i=1}^{T}\Delta_i(F_\epsilon^{-1}(t_1))-t_1\right| + \sup_{|t_1-t_2|\le N_T^{-1}}|t_1-t_2| + \sup_{t_2,s.t.|t_1-t_2|\le N_T^{-1}}\left|t_2-\frac{1}{T}\sum_{i=1}^{T}\Delta_i(F_\epsilon^{-1}(t_2))\right|.
\end{aligned}
\tag{30}
$$

By the Glivenko–Cantelli theorem, it is obvious that $\sup_{|t_1-t_2|\le N_T^{-1}} \frac{1}{T}\left|\sum_{i=1}^{T}\left(\Delta_i(F_\epsilon^{-1}(t_1))-\Delta_i(F_\epsilon^{-1}(t_2))\right)\right|$ is $o_p(1)$. Next, we consider the second term of the r.h.s of Eq. (29):

$$
\begin{aligned}
&\sup_{k\le N_T}\sup_{|\eta_T|\le T^{-\lambda}}|z_T(x_k,\eta_T)| \\
&= \sup_{k\le N_T}\sup_{|\eta_T|\le T^{-\lambda}}\left|\frac{1}{T}\sum_{i=1}^{T}\left(\Delta_i(x_k+\eta_T)-\Delta_i(x_k)\right)\right| \\
&= \sup_{k\le N_T}\sup_{|\eta_T|\le T^{-\lambda}}\left|\frac{1}{T}\sum_{i=1}^{T}\Delta_i(x_k+\eta_T)-F_\epsilon(x_k+\eta_T)+F_\epsilon(x_k+\eta_T)-F_\epsilon(x_k)+F_\epsilon(x_k)-\frac{1}{T}\sum_{i=1}^{T}\Delta_i(x_k)\right| \\
&\le \sup_{k\le N_T}\sup_{|\eta_T|\le T^{-\lambda}}\left\{\left|\frac{1}{T}\sum_{i=1}^{T}\Delta_i(x_k+\eta_T)-F_\epsilon(x_k+\eta_T)\right|+|F_\epsilon(x_k+\eta_T)-F_\epsilon(x_k)|+\left|F_\epsilon(x_k)-\frac{1}{T}\sum_{i=1}^{T}\Delta_i(x_k)\right|\right\}.
\end{aligned}
\tag{31}
$$

Applying the Glivenko–Cantelli theorem again, we can find the first and third term in the r.h.s. of Eq. (31) converges to 0 a.s.. For the middle term:

$$
\begin{aligned}
&\sup_{k\le N_T}\sup_{|\eta_T|\le T^{-\lambda}}|F_\epsilon(x_k+\eta_T)-F_\epsilon(x_k)| \\
&= \sup_{k\le N_T}\sup_{|\eta_T|\le T^{-\lambda}}\left|F_\epsilon(x_k)+F_\epsilon'(\tilde{x})\eta_T-F_\epsilon(x_k)\right| \\
&\le \sup_{x}\sup_{|\eta_T|\le T^{-\lambda}}\left|F_\epsilon'(\tilde{x})\eta_T\right| \to 0, \text{ Under A11.}
\end{aligned}
\tag{32}
$$

We can do a similar analysis for the first term of the r.h.s of Eq. (29). Combining all parts together, we prove Lemma 2.1. The proof with heteroscedastic errors can be written similarly. □

Proof of Theorem 2.3. For simplifying the notations, we consider the NLAR model with order 1 and homoscedastic errors. For NLAR models with higher order and heteroscedastic errors, the proof can be written similarly. We show the proof of $h = 2$ as an example. The proof of higher steps or one-step prediction can be written similarly.

First, by the tower property, we can show that $F_{X_{T+2}|X_T}(x)$ is equivalent to:

$$\begin{aligned} F_{X_{T+2}|X_T}(x) &= \mathbb{P}(X_{T+2} \leq x | X_T) \\ &= \mathbb{P}(\phi(X_{T+1}, \theta_1) + \epsilon_{T+2} \leq x | X_T) \\ &= \mathbb{P}\left(\epsilon_{T+2} \leq x - \phi(\phi(X_T, \theta_1) + \epsilon_{T+1}, \theta_1) \middle| X_T\right) \\ &= \mathbb{E}\left[\mathbb{P}\left(\epsilon_{T+2} \leq x - \phi(\phi(X_T, \theta_1) + \epsilon_{T+1}, \theta_1) \middle| \epsilon_{T+1}, X_T\right) \middle| X_T\right] \\ &= \mathbb{E}\left[F_\epsilon\left(x - \phi(\phi(X_T, \theta_1) + \epsilon_{T+1}, \theta_1)\right) \middle| X_T\right] \\ &= \mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x, X_T, \epsilon_{T+1})\right) \middle| X_T\right]; \end{aligned} \tag{33}$$

we use $\mathscr{L}(x, X_T, \epsilon_{T+1})$ to represent $x - \phi(\phi(X_T, \theta_1) + \epsilon_{T+1}, \theta_1)$ to simplify notations. Similarly, we can analyze $F_{X^*_{T+2}|X_T,\ldots,X_0}(x)$, it has below equivalent expressions:

$$\begin{aligned} F_{X^*_{T+2}|X_T,\ldots,X_0}(x) &= \mathbb{P}(X^*_{T+2} \leq x | X_T, \ldots, X_0) \\ &= \mathbb{E}\left[\mathbb{P}\left(\hat{\epsilon}^*_{T+2} \leq \widehat{\mathscr{L}}(x, X_T, \hat{\epsilon}^*_{T+1}) \middle| \hat{\epsilon}^*_{T+1}, X_T, \ldots, X_0\right) \middle| X_T, \ldots, X_0\right] \\ &= \mathbb{E}^*\left[\widehat{F}_\epsilon\left(\widehat{\mathscr{L}}(x, X_T, \hat{\epsilon}^*_{T+1})\right)\right], \end{aligned} \tag{34}$$

where $\widehat{\mathscr{L}}(x, X_T, \hat{\epsilon}^*_{T+1})$ represents $x - \phi(\phi(X_T, \widehat{\theta}_1) + \hat{\epsilon}^*_{T+1}, \widehat{\theta}_1)$ and $\mathbb{E}^*(\cdot)$ represents the expectation in the bootstrap world, i.e., $\mathbb{E}(\cdot | X_T, \ldots, X_0)$. Thus, we hope to show:

$$\sup_{|x| \leq c_T} \left| \mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x, X_T, \hat{\epsilon}^*_{T+1}))\right] - \mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x, X_T, \epsilon_{T+1})\right) \middle| X_T\right] \right| \xrightarrow{p} 0. \tag{35}$$

From here, we first bound the region of X_t by Lemma 1 of Franke, Neumann, and Stockis (2004) under A1–A7:

$$\mathbb{P}(|X_T| > \gamma_T) \to 0, \tag{36}$$

where $\{\gamma_T\}$ is a sequence of sets, such that $\gamma_1 \subseteq \cdots \subseteq \gamma_T \subseteq \gamma_{T+1} \subseteq \cdots$ with the form $\gamma_T = [-T^c, T^c]$; c is some appropriate constant. For deriving such a result for a time series model with heteroscedastic errors, we need the additional assumption of variance function in A1; then the proof is referred to Lemma 1 of Franke, Neumann, and Stockis (2004). In addition, we have a relationship:

$$\begin{aligned} &\mathbb{P}\left(\sup_{|x| \leq c_T} \left| \mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x, X_T, \hat{\epsilon}^*_{T+1}))\right] - \mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x, X_T, \epsilon_{T+1})\right) \middle| X_T\right] \right| > \varepsilon\right) \\ &\leq \mathbb{P}(|X_T| > \gamma_T) + \mathbb{P}\left((|X_T| \leq \gamma_T) \bigcap \left(\sup_{|x| \leq c_T} \left| \mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x, X_T, \hat{\epsilon}^*_{T+1}))\right] - \mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x, X_T, \epsilon_{T+1})\right) \middle| X_T\right] \right| > \varepsilon\right)\right). \end{aligned} \tag{37}$$

Thus, to verify Eq. (35), we just need to show that the second term of the r.h.s. of Eq. (37) converges to 0. We can take the sequence c_T and γ_T to be the same sequence which converges to infinity slowly enough. Then, it is enough for us to analyze the asymptotic probability of the below expression:

$$\sup_{|x|,|y| \leq c_T} \left| \mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x, y, \hat{\epsilon}^*_{T+1}))\right] - \mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x, y, \epsilon_{T+1})\right)\right] \right| > \varepsilon. \tag{38}$$

Decompose the l.h.s. of Eq. (38) as:

$$\begin{aligned}
&\sup_{|x|,|y|\le c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x,y,\epsilon_{T+1})\right)\right]\right| \\
&\le \sup_{|x|,|y|\le c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}^*\left[F_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))\right]\right| \\
&+ \sup_{|x|,|y|\le c_T}\left|\mathbb{E}^*\left[F_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x,y,\epsilon_{T+1})\right)\right]\right|.
\end{aligned} \tag{39}$$

Then, we analyze two terms on the r.h.s. of Eq. (39) separately. For the first term, we have:

$$\begin{aligned}
&\sup_{|x|,|y|\le c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}^*\left[F_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))\right]\right| \\
&\le \sup_{|x|,|y|\le c_T}\mathbb{E}^*\left|\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))-F_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))\right| \\
&\le \sup_{|x|,|y|\le c_T,z}\left|\widehat{F}_\epsilon(\widehat{\mathscr{L}}(x,y,z))-F_\epsilon(\widehat{\mathscr{L}}(x,y,z))\right| \xrightarrow{p} 0, \text{ under Lemma 4.1.}
\end{aligned} \tag{40}$$

For the second term on the r.h.s. of Eq. (39), we have:

$$\begin{aligned}
&\sup_{|x|,|y|\le c_T}\left|\mathbb{E}^*\left[F_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x,y,\epsilon_{T+1})\right)\right]\right| \\
&\le \sup_{|x|,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}_i))-\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathscr{L}(x,y,\epsilon_i))\right| \\
&+ \sup_{|x|,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathscr{L}(x,y,\epsilon_i))-\mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x,y,\epsilon_{T+1})\right)\right]\right|,
\end{aligned} \tag{41}$$

where $\{\epsilon_j\}_{j=1}^T$ are taken as $X_i-\phi(X_{i-1},\theta_1)$ for $i=1,\dots,T$. We can show:

$$\begin{aligned}
&\mathbb{P}\left(\max_{i=1,\dots,T}\left|\epsilon_i-\hat{\epsilon}_i\right|>\varepsilon\right) \\
&=\mathbb{P}\left(\max_{i=1,\dots,T}\left|X_i-\phi(X_{i-1},\theta_1)-X_i+\phi(X_{i-1},\widehat{\theta}_1)\right|>\varepsilon\right) \\
&\le \mathbb{P}\left(\left(\max_{i=1,\dots,T}|X_{i-1}|>c_T\right)\right) \\
&+\mathbb{P}\left(\left(\max_{i=1,\dots,T}|X_{i-1}|<c_T\right)\bigcap\left(\max_{i=1,\dots,T}\left|\phi(X_{i-1},\widehat{\theta}_1)-\phi(X_{i-1},\theta_1)\right|>\varepsilon\right)\right) \\
&\le o(1)+\mathbb{P}\left(\sup_{|x|\le c_T}\left|\phi(x,\widehat{\theta}_1)-\phi(x,\theta_1)\right|>\varepsilon\right) \\
&\to 0, \text{ under A9 and A10.}
\end{aligned} \tag{42}$$

We further consider two terms on the r.h.s. of Eq. (41) separately. For the first term, by Taylor

expansion, we have:

$$
\begin{aligned}
&\sup_{|x|,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}_i))-\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathscr{L}(x,y,\epsilon_i))\right| \\
&=\sup_{|x|,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}\Big(F_\epsilon(\mathscr{L}(x,y,\epsilon_i))+f_\epsilon(o_i)(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}_i)-\mathscr{L}(x,y,\epsilon_i))\Big)-\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathscr{L}(x,y,\epsilon_i))\right| \\
&=\sup_{|x|,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}f_\epsilon(o_i)(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}_i)-\mathscr{L}(x,y,\epsilon_i))\right| \\
&\le\sup_{|x|,|y|\le c_T}\frac{1}{T}\sum_{i=1}^{T}\left|f_\epsilon(o_i)(\widehat{\mathscr{L}}(x,y,\hat{\epsilon}_i)-\mathscr{L}(x,y,\epsilon_i))\right| \\
&\le\sup_{|x|,|y|\le c_T}\sup_z|f_\epsilon(z)|\cdot\frac{1}{T}\sum_{i=1}^{T}\left|\widehat{\mathscr{L}}(x,y,\hat{\epsilon}_i)-\mathscr{L}(x,y,\epsilon_i)\right| \\
&\le\sup_{|x|,|y|\le c_T}C\cdot\frac{1}{T}\sum_{i=1}^{T}\left|\widehat{\mathscr{L}}(x,y,\hat{\epsilon}_i)-\mathscr{L}(x,y,\epsilon_i)\right|\ \text{(under A11)} \\
&\le\sup_{|x|,|y|\le c_T,j\in\{1,\dots,T\}}C\cdot\left|\widehat{\mathscr{L}}(x,y,\hat{\epsilon}_j)-\mathscr{L}(x,y,\epsilon_j)\right|.
\end{aligned}
\tag{43}
$$

From Eq. (42) and A9–A12, we have Eq. (43) converges to 0 in probability. For the second term on the r.h.s. of Eq. (41), by the uniform law of large numbers, we have:

$$
\sup_{|x|,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathscr{L}(x,y,\epsilon_i))-\mathbb{E}\left[F_\epsilon\left(\mathscr{L}(x,y,\epsilon_{T+1})\right)\right]\right|\xrightarrow{p}0. \tag{44}
$$

Combine all pieces, Eq. (38) converges to 0 in probability, which implies Theorem 2.3. □

Proof of Lemma 3.1. Under A4-A7, the time series $\{X_t\}$ is ergodic. Besides, under A2 and A10, we can show $L(\vartheta)$ is uniformly finite for $\vartheta\in\Theta_1$. With A13, by the uniform law of larger numbers for the ergodic series, see Theorem 6 of Kirch and Kamgaing (2012) for a reference, we have:

$$
\sup_{\vartheta\in\Theta_1}|L_T(\vartheta)-L(\vartheta)|\xrightarrow{p}0. \tag{45}
$$

Under A13 and A14, we can easily show that:

$$
\inf_{|\vartheta-\theta_1|>\varepsilon}L(\vartheta)>L(\theta_1),\text{for }\forall\epsilon, \tag{46}
$$

Eq. (46) implies that given $\forall\varepsilon>0$, $\exists C>0$ such that $|\vartheta-\theta_1|>\epsilon\Rightarrow L(\vartheta)-L(\theta_1)\ge C>0$, thus we have:

$$
\begin{aligned}
\mathbb{P}(|\widehat{\theta}_1-\theta_1|>\epsilon)&\le\mathbb{P}(L(\widehat{\theta}_1)-L(\theta_1)\ge C) \\
&=\mathbb{P}(L(\widehat{\theta}_1)-L_T(\widehat{\theta}_1)+L_T(\widehat{\theta}_1)-L(\theta_1)\ge C) \\
&\le\mathbb{P}(L(\widehat{\theta}_1)-L_T(\widehat{\theta}_1)+L_T(\theta_1)-L(\theta_1)\ge C) \\
&\le\mathbb{P}(2\sup_{\vartheta\in\Theta_1}|L_T(\vartheta)-L(\vartheta)|>C)\to0.
\end{aligned}
\tag{47}
$$

The last inequality of Eq. (47) is guaranteed by Eq. (45). □

Proof of Lemma 3.2. The proof of the consistency of $\widehat{\theta}_1$ to θ_1 is the same as the proof of Lemma 4.2. Similar to the proof of Lemma 4.2, by the ergodic property of the series $\{X_t\}$, we know:

$$\sup_{\vartheta\in\Theta_2}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t-\phi(X_{t-1},\theta_1)}{h(X_{t-1},\vartheta)}\right)^2-\mathbb{E}\left(\frac{X_t-\phi(X_{t-1},\theta_1)}{h(X_{t-1},\vartheta)}\right)^2\right|\xrightarrow{p}0. \tag{48}$$

Actually, after the first step estimation, we have got the $\widehat{\theta}_1$ which is consistent to the true parameter θ_1. Thus, we need to find the below convergence relationship:

$$\sup_{\vartheta\in\Theta_2}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t-\phi(X_{t-1},\widehat{\theta}_1)}{h(X_{t-1},\vartheta)}\right)^2-\mathbb{E}\left(\frac{X_t-\phi(X_{t-1},\theta_1)}{h(X_{t-1},\vartheta)}\right)^2\right|\xrightarrow{p}0. \tag{49}$$

Since $\widehat{\theta}_1$ converges to θ_1 in probability, it is easily to find:

$$\sup_{\vartheta\in\Theta_2}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t-\phi(X_{t-1},\widehat{\theta}_1)}{h(X_{t-1},\vartheta)}\right)^2-\frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t-\phi(X_{t-1},\theta_1)}{h(X_{t-1},\vartheta)}\right)^2\right|\xrightarrow{p}0, \tag{50}$$

which implies Eq. (49) for a compact set Θ_2 in conjunction with Eq. (48). Then, since $|\cdot-a|$ is a uniformly continuous function for a constant a, by applying the uniform continuous mapping theorem which is the Theorem 1 of Kasy (2019) on Eq. (49), we can get:

$$\sup_{\vartheta\in\Theta_2}|K_T(\vartheta,\widehat{\theta}_1)-K(\vartheta,\theta_1)|\xrightarrow{p}0, \tag{51}$$

where $K(\vartheta,\theta_1)$ and $K_T(\vartheta,\widehat{\theta}_1)$ are:

$$K(\vartheta,\theta_1)=\left|\mathbb{E}\left(\frac{X_t-\phi(X_{t-1},\theta_1)}{h(X_{t-1},\vartheta)}\right)^2-1\right|\ ;\ K_T(\vartheta,\widehat{\theta}_1)=\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t-\phi(X_{t-1},\widehat{\theta}_1)}{h(X_{t-1},\vartheta)}\right)^2-1\right|. \tag{52}$$

Then, we can repeat the procedure in the proof of Lemma 3.1 to show:

$$\mathbb{P}(|\widehat{\theta}_2-\theta_2|>\varepsilon)\to 0,\ \forall\varepsilon>0. \tag{53}$$

□

Proof of Theorem 3.1. Throughout this proof, we focus on a sequence of sets $\Omega_{T+p}\subseteq\mathbb{R}^{T+p}$, such that $\mathbb{P}\left((X_{-p+1},\ldots,X_T)\notin\Omega_{T+p}\right)=o(1)$.We first explain the truth that the bootstrap series is ergodic for $(X_{-p+1},\ldots,X_T)\in\Omega_{T+p}$. Particularly, we want to check analogous A4-A7 in the bootstrap world. With the consistency property of parameter estimators, i.e., A9 and the continuous density function of residuals after the convolutional manipulation, we can easily find A4-A7 are satisfied in the bootstrap world; see Theorem 2 of Franke, Neumann, and Stockis (2004) for more discussions.

To show the closeness of the stationary distribution of bootstrap and real series. Based on the proof of Theorem 3 of Franke, Kreiss, et al. (2002) and Theorem 3 of Franke, Neumann, and Stockis (2004), we can get:

$$\sup_{B}|\Pi(B)-\Pi^*(B)|=o(1), \tag{54}$$

which holds for all measurable sets B, where $\Pi(B)$ and $\Pi^*(B)$ represent stationary distribution for real series and bootstrap series, respectively. This implies, in conjunction with Eq.(28) in the main text and the condition that $\mathbb{P}(X_t \notin \gamma_T) = o(1)$, we can get:

$$\begin{aligned}
&\int_{\mathbb{R}^2} (x_1 - \phi(x_0, \Xi_\theta))^2 \hat{f}_\epsilon(x_1 - \phi(x_0, \hat{\theta}))\pi^*(x_0)dx_1dx_0 \\
&= \int_{\gamma_T^2} (x_1 - \phi(x_0, \Xi_\theta))^2 \hat{f}_\epsilon(x_1 - \phi(x_0, \hat{\theta}))\pi^*(x_0)dx_1dx_0 + o(1) \\
&\to \mathbb{E}[X_1 - \phi(X_0, \Xi_\theta)]^2,
\end{aligned} \tag{55}$$

which implies Theorem 3.1 directly.

□

Proof of Theorem 3.2. To simplify notation, we consider the case $p = q = 1$. Higher order cases can be handled similarly. We first build the estimation inference for $\widehat{\theta}_1$. Starting from A19, we have:

$$op(T^{-1/2}) = \nabla L_T(\widehat{\theta}_1) = \nabla L_T(\theta_1) + \nabla^2 L_T(\widetilde{\theta}_1)(\widehat{\theta}_1 - \theta_1), \tag{56}$$

where $\widetilde{\theta}_1$ is between $\widehat{\theta}_1$ and θ_1; hence $\widetilde{\theta}_1$ also converges to θ_1 in probability. First, we consider $\nabla^2 L_T(\widetilde{\theta}_1)$, which has a form as below:

$$\nabla^2 L_T(\widetilde{\theta}_1) = \frac{1}{T}\nabla^2 \sum_{t=1}^{T} \left(X_t - \phi(X_{t-1}, \widetilde{\theta}_1)\right)^2 = \frac{1}{T}\sum_{t=1}^{T} \nabla^2 q_t(\widetilde{\theta}_1). \tag{57}$$

Under A18, since it is easy to check $\mathbb{E}\sup_{\vartheta \in \Theta_1^0} ||\nabla^2 q_1(\vartheta)|| < \infty$, by combining the uniform law of larger numbers for the ergodic series, dominated convergence theorem, the consistency between $\widetilde{\theta}_1$ and θ_1 and the continuity of $L_T(\cdot)$ w.r.t. θ_1, we can get:

$$\nabla^2 L_T(\widetilde{\theta}_1) \xrightarrow{p} B_1, \tag{58}$$

where $B_1 = 2 \cdot \mathbb{E}\left(\nabla\phi(X_0, \theta_1)(\nabla\phi(X_0, \theta_1))^\top\right)$. Thus, we can multiply both side of Eq. (56) by $\sqrt{T}$ to get:

$$op(1) = \sqrt{T}\nabla L_T(\theta_1) + (B_1 + op(1))\sqrt{T}(\widehat{\theta}_1 - \theta_1), \tag{59}$$

which further implies that:

$$\sqrt{T}(\widehat{\theta}_1 - \theta_1) = -B_1^{-1}\sqrt{T}\nabla L_T(\theta_1) + op(1), \tag{60}$$

where $\sqrt{T}\nabla L_T(\theta_1)$ has a concrete form as below:

$$-\sqrt{T}\nabla L_T(\theta_1) = \frac{2}{\sqrt{T}}\sum_{t=1}^{T}(X_t - \phi(X_{t-1}, \theta_1))\nabla\phi(X_{t-1}, \theta_1). \tag{61}$$

We need the CLT for strongly mixing processes to show the asymptotic normality of Eq. (61). Based on Theorem 1.7 of Bosq (2012) for univariate case, we can show the normality with the Cramér–Wold theorem. Also, Kirch and Kamgaing (2012) applied the strong invariance principle from Kuelbs and Philipp (1980) to show the asymptotic distribution of multivariate cases directly. Since we need to

analyze the asymptotic distribution of parameters in the bootstrap world, we take the first approach which is more clear. We consider:

$$-\sqrt{T}\boldsymbol{a}^\top \nabla L_T(\theta_1) = \frac{2}{\sqrt{T}} \sum_{t=1}^{T} (X_t - \phi(X_{t-1}, \theta_1)) \boldsymbol{a}^\top \nabla \phi(X_{t-1}, \theta_1), \tag{62}$$

where $\boldsymbol{a}^\top$ is any real vector that has the same dimension as θ_1. Since we assume we can correctly specify the model, we have:

$$-\sqrt{T}\boldsymbol{a}^\top \nabla L_T(\theta_1) \xrightarrow{d} N(0, \tau_1^2), \tag{63}$$

where τ_1^2 has the form as follows:

$$\begin{aligned}
\tau_1^2 &= \sum_{i=-\infty}^{\infty} \mathrm{Cov}\left(2 \cdot (X_1 - \phi(X_0, \theta_1)) \boldsymbol{a}^\top \nabla \phi(X_0, \theta_1), 2 \cdot (X_{i+1} - \phi(X_i, \theta_1)) \boldsymbol{a}^\top \nabla \phi(X_i, \theta_1)\right) \\
&= 4 \cdot \mathbb{E}\left((\boldsymbol{a}^\top \nabla \phi(X_0, \theta_1))(X_1 - \phi(X_0, \theta_1))^2 (\nabla \phi(X_0, \theta_1)^\top \boldsymbol{a})\right) \\
&+ 2 \sum_{i=1}^{\infty} \mathrm{Cov}\left(2 \cdot (X_1 - \phi(X_0, \theta_1)) \boldsymbol{a}^\top \nabla \phi(X_0, \theta_1), 2 \cdot (X_{i+1} - \phi(X_i, \theta_1)) \boldsymbol{a}^\top \nabla \phi(X_i, \theta_1)\right) \\
&= 4 \cdot \mathbb{E}(\boldsymbol{a}^\top \sigma(X_0, \theta_2) \nabla \phi(X_0, \theta_1) \nabla \phi(X_0, \theta_1)^\top \sigma(X_0, \theta_2) \boldsymbol{a}) \\
&+ 2 \sum_{i=1}^{\infty} \mathrm{Cov}\left(2 \cdot \epsilon_1 \boldsymbol{a}^\top \nabla \phi(X_0, \theta_1), 2 \cdot \epsilon_{i+1} \boldsymbol{a}^\top \nabla \phi(X_i, \theta_1)\right) \\
&= 4 \cdot \mathbb{E}(\boldsymbol{a}^\top \sigma(X_0, \theta_2) R_1 \sigma(X_0, \theta_2) \boldsymbol{a}),
\end{aligned} \tag{64}$$

where $R_1 = \nabla \phi(X_0, \theta_1) \nabla \phi(X_0, \theta_1)^\top$. Thus, applying Cramér–Wold theorem, we have:

$$-\sqrt{T} \nabla L_T(\theta_1) \xrightarrow{d} N(0, \Omega_1), \tag{65}$$

where Ω_1 is $4 \cdot \mathbb{E}(\sigma(X_0, \theta_2) R_1 \sigma(X_0, \theta_2))$ Thus, we can conclude from Eq. (60) that:

$$\sqrt{T}(\widehat{\theta}_1 - \theta_1) \xrightarrow{d} N(0, B_1^{-1} \Omega_1 B_1^{-1}). \tag{66}$$

Compared to the result in Kirch and Kamgaing (2012), they got another form of Ω_1:

$$\Omega_1 = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\left(\nabla \sum_{t=1}^{T} (X_t - \phi(X_{t-1}, \theta_1))^2\right)\left(\nabla \sum_{t=1}^{T} (X_t - \phi(X_{t-1}, \theta_1))^2\right)^T\right]. \tag{67}$$

With a correctly specified model, it is equivalent to our form.

We can also analyze the distribution of parameter estimation $\widehat{\theta}_2$. By A19, we first have:

$$op(T^{-1/2}) = \nabla K_T(\theta_2, \widehat{\theta}_1) + \nabla^2 K_T(\widetilde{\theta}_2, \widehat{\theta}_1)(\widehat{\theta}_2 - \theta_2). \tag{68}$$

For simplifying the notation, we write:

$$\begin{aligned}
K_T(\widetilde{\theta}_2, \widehat{\theta}_1) &= \left(\frac{1}{T} \sum_{t=1}^{T} \left(\frac{X_t - \phi(X_{t-1}, \widehat{\theta}_1)}{h(X_{t-1}, \widetilde{\theta}_2)}\right)^2 - 1\right)^2 \\
&= \left(\frac{1}{T} \sum_{t=1}^{T} g(X_t, X_{t-1}, \widetilde{\theta}_2, \widehat{\theta}_1) - 1\right)^2.
\end{aligned} \tag{69}$$

We can find $\nabla K_T(\theta_2, \theta_1)$ and $\nabla^2 K_T(\theta_2, \theta_1)$ w.r.t θ_2 have a following forms respectively:

$$\begin{aligned}
\nabla K_T(\theta_2, \theta_1) &= 2 \cdot \left(\frac{1}{T} \sum_{t=1}^{T} g(X_t, X_{t-1}, \theta_2, \theta_1) - 1 \right) \cdot \left(\frac{1}{T} \sum_{t=1}^{T} \nabla g(X_t, X_{t-1}, \theta_2, \theta_1) \right) \\
\nabla^2 K_T(\theta_2, \theta_1) &= 2 \cdot \left(\frac{1}{T} \sum_{t=1}^{T} \nabla g(X_t, X_{t-1}, \theta_2, \theta_1) \right) \cdot \left(\frac{1}{T} \sum_{t=1}^{T} \nabla g(X_t, X_{t-1}, \theta_2, \theta_1) \right)^\top \\
&+ 2 \cdot \left(\frac{1}{T} \sum_{t=1}^{T} g(X_t, X_{t-1}, \theta_2, \theta_1) - 1 \right) \cdot \left(\frac{1}{T} \sum_{t=1}^{T} \nabla^2 g(X_t, X_{t-1}, \theta_2, \theta_1) \right).
\end{aligned} \tag{70}$$

Similarly to analyze $\nabla^2 L_T(\widetilde{\theta}_1)$, under the consistence relationship between $\widetilde{\theta}_2$ and θ_2, $\widehat{\theta}_1$ and θ_1, we can get:

$$\nabla^2 K_T(\widetilde{\theta}_1, \widehat{\theta}_1) \xrightarrow{p} B_2. \tag{71}$$

where $B_2 = 2 \cdot \mathbb{E}(\nabla g(X_1, X_0, \theta_2, \theta_1)) \cdot \mathbb{E}(\nabla g(X_1, X_0, \theta_2, \theta_1)^\top)$. Looking back to Eq. (68), it is left to analyze $\sqrt{T} \nabla K_T(\theta_2, \widehat{\theta}_1)$. Through Slutsky's theorem, it is asymptotically equivalent to analyze $\sqrt{T} \frac{2}{T} \sum_{t=1}^{T} [g(X_t, X_{t-1}, \theta_2, \theta_1) - 1] \cdot B_3$; $B_3 = \mathbb{E}(\nabla g(X_1, X_0, \theta_2, \theta_1))$. Apply the same technique as we analyzed the distribution of $\widehat{\theta}_1$, we can get:

$$\sqrt{T} \boldsymbol{a}^\top \nabla K_T(\theta_2, \widehat{\theta}_1) \xrightarrow{d} N(0, \tau_2^2), \tag{72}$$

where $\tau_2^2 = 4 \cdot \mathbb{E}(\boldsymbol{a}^\top B_3 R_2 B_3^\top \boldsymbol{a})$; $R_2 = (g(X_1, X_0, \theta_2, \theta_1) - 1)^2$. Thus, we have:

$$\sqrt{T} \nabla K_T(\theta_2, \widehat{\theta}_1) \xrightarrow{d} N(0, \Omega_2), \tag{73}$$

where $\Omega_2 = 4 \cdot \mathbb{E}(B_3 R_2 B_3^\top)$. This further implies that:

$$\sqrt{T}(\widehat{\theta}_2 - \theta_2) \xrightarrow{d} N(0, B_2^{-1} \Omega_2 B_2^{-1}). \tag{74}$$

□

References

An H, Huang F. 1996. The geometrical ergodicity of nonlinear autoregressive models. *Statistica Sinica* **6**: 943–956.

Boldin M. 1983. Estimation of the distribution of noise in an autoregression scheme. *Theory of Probability & Its Applications* **27**: 866–871.

Bosq D. 2012. *Nonparametric statistics for stochastic processes: estimation and prediction*. Vol. 110. Springer Science & Business Media.

Doob JL. 1953. *Stochastic processes*. Wiley New York.

Franke J, Kreiss JP, et al. 2002. Properties of the nonparametric autoregressive bootstrap. *Journal of Time Series Analysis* **23**: 555–585.

Franke J, Neumann MH, Stockis JP. 2004. Bootstrapping nonparametric estimators of the volatility function. *Journal of Econometrics* **118**: 189–218.

Kasy M. 2019. Uniformity and the delta method. *Journal of Econometric Methods* **8**: 1–19.

Kirch C, Kamgaing JT. 2012. Testing for parameter stability in nonlinear autoregressive models. *Journal of Time Series Analysis* **33**: 365–385.

Kuelbs J, Philipp W. 1980. Almost sure invariance principles for partial sums of mixing B-valued random variables. *The Annals of Probability*: 1003–1036.

Majerek D, Nowak W, Zieba W. 2005. Conditional strong law of large number. *Int. J. Pure Appl. Math* **20**: 143–156.

Min C, Hongzhi A. 1999. The probabilistic properties of the nonlinear autoregressive model with conditional heteroskedasticity. *Acta Mathematicae Applicatae Sinica* **15**: 9–17.

Tjøstheim D. 1990. Non-linear time series and Markov chains. *Advances in Applied Probability* **22**: 587–611.

Tong H. 1990. *Non-linear Time Series: A Dynamical System Approach*. Oxford university press.

Tweedie RL. 1975. Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. *Stochastic Processes and Their Applications* **3**: 385–403.