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

Abstract

The non-linear autoregressive (NLAR) model plays an important role in modeling and predicting time series. One-step ahead prediction is straightforward using the NLAR model, but the multi-step ahead prediction is cumbersome. For instance, iterating the one-step ahead predictor is a convenient strategy for linear autoregressive (LAR) models, but it is suboptimal under NLAR. In this paper, we first propose a simulation and/or bootstrap algorithm to construct optimal point predictors under an L_1 or L_2 loss criterion. In addition, we construct bootstrap prediction intervals in the multi-step ahead prediction problem; in particular, we develop an asymptotically valid quantile prediction interval as well as a pertinent prediction interval for future values. In order to correct the undercoverage of prediction intervals with finite samples, we further employ predictive—as opposed to fitted—residuals in the bootstrap process. Simulation and empirical studies are also given to substantiate the finite sample performance of our methods.

Keywords: Bootstrap, NLAR forecasting, Pertinence prediction.
MOS subject classification: 62M10, 62F40.

1 Introduction

In the domain of time series analysis, accurate forecasting based on observed data is an important topic. Such single- or multi-step ahead predictions play an important role in forecasting crop yields, stock prices, traffic volume, etc. For Linear Autoregressive (LAR) models with finite order and independent, identically distributed (*i.i.d.*) or martingale difference innovations, it is easy to construct the optimal (with respect to L_2 risk) multi-step ahead predictor by iterating the one-step ahead predictor. However, the LAR model may not be enough to analyze complicated data in the real world. As pointed out by the work of De Gooijer and Kumar (1992) and Tjøstheim (1994), there are various occasions when prior knowledge indicates the data-generating process is in a non-linear form; see the review of Politis (2009) for example. Furthermore, there are several ways to test the hypothesis of linearity of the data at hand; see the work of Berg, McMurry, and Politis (2012) for traditional and bootstrap/subsampling approaches.

The analysis of Non-linear Autoregressive (NLAR) models can be traced back to the work of Jones and Cox (1978). Although the one-step ahead optimal (with respect to L_2 risk) prediction of (causal) NLAR models is usually easy to obtain, the optimal (no matter in L_2 or L_1 loss) multi-step ahead prediction can not be obtained

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

by the iterative procedure we employ for LAR models, even when the NLAR model parameters are known. To resolve this issue, Pemberton (1987) proposed a numerical integration approach to get the exact solution. However, his approach assumes that the distribution of innovation is known, which is usually not realistic in the real world. Besides, this numerical approach can be very computationally heavy for long-horizon predictions. Instead, some suboptimal ideas were proposed, such as estimating a model by minimizing multi-step ahead mean squared errors and then making predictions directly; this multi-step estimation criterion can improve the long-horizon forecasting accuracy compared to the standard 1-step ahead error minimization strategy when models are misspecified for the data generation process; see the work of Zhang, Patuwo, and Hu (1998), Clements and Hendry (1996) and Lee and Billings (2003) for a discussion.

The work of Guo, Bai, and An (1999) further shed some light on the multi-step ahead prediction of NLAR models. Taking advantage of the true innovation distribution or the empirical residual distribution, they proposed an analytic predictor that asymptotically converges to the optimal predictor. Nevertheless, their analyses are limited to the L_2 optimal point prediction and are lacking details when the model is unknown. In several applied areas, e.g. econometrics, climate modeling, water resources management, etc., data might not possess a finite 2nd moment in which case optimizing L_2 loss is vacuous. For example, financial returns typically do not possess a finite 4th moment; hence, to predict their volatility which is usually mimicked by its 2nd moment, it is not appropriate to rely on L_2 optimal prediction since the MSE of predicting squared returns is essentially a fourth moment. For all such cases—but also of independent interest—prediction that is optimal with respect to L_1 loss should receive more attention in practice; see detailed discussions from Ch. 10 of Politis (2015).

Unfortunately, the aforementioned numerical integration and analytic methods for NLAR prediction can not be extended to L_1 optimal prediction directly. In addition, even for linear autoregressions, the multi-step ahead L_1 optimal predictor is elusive since iterating the one-step ahead predictor does not work in the L_1 loss setup. Beyond the point prediction, we should also be concerned about the accuracy of our point predictions. In analogy to the construction of Confidence Intervals (CI) in estimation problems, we may attempt to measure the accuracy of point predictions by constructing Prediction Intervals (PI); see the formal definition of such measures in Section 2. In the paper at hand, we provide an algorithm to make prediction inferences for a popular type of NLAR model with a specific structural form that contains separate parametric mean and volatility/variance functions. We also indicate the potential extension of our algorithm to a more general class of NLAR models. When the model and innovation distribution are known, we can deploy Monte Carlo (MC) *simulation* to achieve consistent forecasting. When the model is unknown—which is typically the case—we need to fit the model to get estimated parameters and innovation distribution. Throughout, we assume that the order of the parametric non-linear time series model p is known; when we say the model is unknown, we mean that the corresponding parameter values of this model are unknown. Performing MC simulation using the fitted model and estimated innovation distribution effectively becomes a *bootstrap* method; see the book Kreiss and Paparoditis (2023) for discussions of the bootstrap technique.

For a meaningful prediction in the time series domain, all future predictions must be conditional on the latest p observed data where p is the order of the NLAR model. Thus, to construct a reasonable predictor in the bootstrap world, we need to make sure predictors of the bootstrap series are also conditional on the exact same p data; this is the idea of the forward bootstrap proposed by Politis (2015) and Pan and Politis (2016). This forward bootstrap method is similar to the density forecast of future values—see Chen, Yang, and Hafner (2004), Manzan and Zerom (2008), and Pascual, Romo, and Ruiz (2001) who applied various approaches to do density forecast. To quantify the point prediction accuracy, the straightforward Quantile Prediction Interval (QPI) based on quantile values of future value distribution is typically characterized by finite-sample undercoverage because it does not take the variability of the model estimation into account; see Wang and Politis (2021) made a related discussion. Recently, Politis (2015) introduced the notion of a so-called Pertinent PI (PPI) that has a better empirical Coverage Rate (CVR) in finite-sample cases; see more explanations about the necessity of capturing estimation variability in Section 3. To implement the PPI, we need to impose more requirements on

the bootstrap series, i.e., we require that the estimated model in the bootstrap world is also consistent with the true one. To check the consistency, the bootstrap series should possess some mixing or weak dependence properties. As developed in the work of Franke, Kreiss, Mammen, and Neumann (2002), it is possible to get a self-ergodic bootstrap series that also approximates the true series via a non-linear autoregressive residual bootstrap (AR bootstrap) approach. In the paper at hand, we focus on the forward-bootstrap prediction of parametric non-linear models; see Politis and Wu (2023) for prediction based on non-parametric models. To further boost the empirical CVR of our bootstrap-based PPIs, we may use predictive (instead of fitted) residuals analogously to the successful construction of PI for regression and autoregression with predictive residuals in work of Politis (2013) and Pan and Politis (2016); the formal definition of predictive residuals is presented in Section 2.2.

The paper is organized as follows. In Section 2, we introduce the forward bootstrap methods to predict a specific class of NLAR model for two situations in which the model and innovation information are known or unknown. Under standard assumptions, we show the consistency of optimal point prediction and asymptotic validity of QPI. In Section 3, we present the algorithm to build the PPI and check its asymptotic pertinence. In Section 4, some simulation results will be presented. Empirical studies are deplored in Section 5. Conclusions are given in Section 6. The proofs of theorems from Sections 2 and 3 are given in Appendix C of the additional Supporting Information.

2 Prediction inference for a specific NLAR of interest

In this paper, we suppose that we observe $T+p$ number of real-valued sample $\{X_{-p+1}, X_{-p+2}, \ldots, X_T\}$. Here and in all that follows, we are exclusively interested in analyzing the NLAR model of the specific form:

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

where $\{\epsilon_t\}$ are *i.i.d.* innovations with mean zero, and $\boldsymbol{X_{t-1}}$ represents vector $\{X_{t-1}, \ldots, X_{t-p}\}$. Model (1) possesses two components. One is $\phi(\cdot)$ —that represents the conditional mean— plus the second one which is the variance function multiplying the innovations ϵ_t. Under Eq. (1), the innovations are explicitly defined and thus easily estimable as residuals after model fitting. If $\sigma(\boldsymbol{X_{t-1}}) \equiv 1$, then we have an NLAR with homoscedastic errors. For simplifying the notation, we consider a common order p for mean and variance functions. We first impose some standard assumptions and suppose they are met throughout this paper:

- A1 $\phi(\cdot)$ and $\sigma(\cdot)$ are continuous functions from $\mathbb{R}^p$ to $\mathbb{R}$, and $\sigma(\cdot)$ is positive and bounded. Moreover, for quantifying the boundness of X_t in probability to serve the proof, we assume that there are $C_M < \infty$ with $\mathbb{E}(|\sigma(X_0)\epsilon_1|^M) \leq C_M$ for all $M < \infty$, where X_0 is the starting point of the time series.
- A2 $\{\epsilon_t\}$ are *i.i.d.* with distribution F_ϵ, satisfying $\mathbb{E}(\epsilon_t) = 0$ and $E(\epsilon_t^2) = 1$. However, if $\sigma(\boldsymbol{X_{t-1}}) \equiv 1$ (homoscedastic errors case), then $E(\epsilon_t^2)$ is not restricted to equal one, but needs to be finite.
- A3 For all t, ϵ_t are *i.i.d.* innovations and are independent of the initialization $X_{-p+1}, \ldots, X_0$.

We attempt to propose a method to make prediction inferences with Eq. (1), especially for multi-step ahead predictions. As known to us, for a stochastic process $\{X_t\}_{t=-\infty}^{T}$, the L_2 optimal predictor of X_{T+h}, $h \geq 1$, given the (infinite) past is:

$$\mathbb{E}[X_{T+h}|X_s, s \leq T], \tag{2}$$

when it exists. As pointed out by Pemberton (1987), this result does not require the stochastic process to be stationary. Since we assume the order of the NLAR model p is finite, Eq. (2) can be simplified to:

$$\mathbb{E}[X_{T+h}|X_T, \ldots, X_{T-p+1}]. \tag{3}$$

Similarly, the L_1 optimal predictor of X_{T+h} given past history is the conditional median:

$$Q_{X_{T+h}|X_T,\ldots,X_{T-p+1}}(1/2), \tag{4}$$

where $Q_{X_{T+h}|X_T,\ldots,X_{T-p+1}}(\cdot)$ is the conditional quantile function of X_{T+h}.

We will call Eq. (3) and Eq. (4) the exactly optimal point predictors based on L_1 or L_2 loss. However, it is hard to compute them directly. Subsequently, we will propose the simulation or bootstrap-based method to find an approximation of the exactly optimal prediction. Moreover, we also consider the PI of future values; an asymptotically valid PI of X_{T+h} with $(1-\alpha)100\%$ CVR given past history can be defined as:

$$\mathbb{P}(L \leq X_{T+h} \leq U) \xrightarrow{p} 1-\alpha, \text{ as } T \to \infty, \tag{5}$$

where L and U are lower and higher PI bounds, respectively. Implicitly, the probability $\mathbb{P}$ should be understood as the conditional probability given the latest p observations. We typically construct a PI that is centered at some meaningful point predictor $\widehat{X}_{T+h}$. An asymptotically valid centered PI with $(1-\alpha)100\%$ CVR given past history can then be defined as:

$$\mathbb{P}(\widehat{X}_{T+h} + R_{\alpha/2} \leq X_{T+h} \leq \widehat{X}_{T+h} + R_{1-\alpha/2}) \xrightarrow{p} 1-\alpha, \text{ as } n \to \infty, \tag{6}$$

where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ denote the lower $\alpha/2$ and $1-\alpha/2$ quantiles with respect to the conditional distribution of the so-called *predictive root*: $X_{T+h} - \widehat{X}_{T+h}$. We should notice that the distribution of the *predictive root* may not be symmetric. Thus, the PI defined by Eq. (6) may not be symmetric, but it is equal-tailed and centered around some meaningful points, e.g., optimal L_1 or L_2 predictions. Notice that it is easy to create symmetric PIs, but to create shortest PIs is not feasible computationally; with symmetric *predictive root* distribution, equal-tailed property implies the shortest property.

Remark. *Beyond providing the prediction inference for the location-scale model (1), our forward bootstrap prediction algorithm could be extended to work for a general NLAR model of the type* $X_t = G(\boldsymbol{X_{t-1}}, \epsilon_t)$ *for some function* $G(\cdot,\cdot)$. *We present the corresponding algorithm in Appendix B of the additional Supporting Information. However, the application of the forward bootstrap prediction algorithm on the general NLAR model hinges on the ability to estimate the function* $G(\cdot,\cdot)$, *and the distribution of the errors* ϵ_t.

To conduct statistical inference for non-linear time series data in the following sections, we need to find a tool to quantify the degree of asymptotic independence of time series. Popular choices are various mixing conditions. For simplifying proofs and relying on existing results, in this paper, we focus on time series with geometrically ergodic property which is equivalent to β-mixing condition with at least exponentially fast mixing rate; see Bradley (2005) made a detailed introduction of different mixing conditions and ergodicity. We further 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:

$$\sup_{||\boldsymbol{x}||_2 \leq K} |\phi(\boldsymbol{x})| < \infty \; ; \quad \sup_{||\boldsymbol{x}||_2 \leq K} |\sigma(\boldsymbol{x})| < \infty, \text{ for each } K > 0, \tag{7}$$

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

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

$$|\phi(\boldsymbol{x})| \leq \lambda \max\{|x_1|, \ldots, |x_p|\} + C. \tag{8}$$

A7 The conditional variance function satisfies:

$$\lim_{||\boldsymbol{x}||_2 \to \infty} \frac{\sigma(\boldsymbol{x})}{||\boldsymbol{x}||_2} = 0. \tag{9}$$

The deduction to get the geometrically ergodic property under the above sufficient conditions is presented in Appendix A of the additional Supporting Information. We refer readers to the work of Stockis, Franke, and Kamgaing (2010) for other conditions to guarantee the geometric ergodicity.

2.1 Prediction inference for NLAR models with known form

We start with a relatively simple case, i.e., prediction inference for known NLAR models. To simplify the notation, we only consider the homoscedastic NLAR model in the main text, the analogous algorithms and theorems that serve for NLAR models with heteroscedastic errors can be shown similarly. In short, we deploy the Monte Carlo simulation to approximate the exact optimal point prediction or PI conditional on the past history. The procedure is summarized in Algorithm 1.

Algorithm 1 h-step ahead prediction of X_{T+h} under homoscedastic Eq. (1) of known form

Step 1 Write homoscedastic Eq. (1) as $X_{T+1} = \phi(X_T, \ldots, X_{T+1-p}) + \epsilon_{T+1}$; Iterate this equation to find the expression of X_{T+h}:

$$X_{T+h} = \mathscr{G}(X_T, \ldots, X_{T-p+1}; \epsilon_{T+1}, \ldots, \epsilon_{T+h}), \tag{10}$$

where we used the notation $\mathscr{G}(X_T, \ldots, X_{T-p+1}; \epsilon_{T+1}, \ldots, \epsilon_{T+h})$ to specify that X_{T+h} depends on $X_T, \ldots, X_{T-p+1}$ and $\{\epsilon_i\}_{i=T+1}^{T+h}$.

Step 2 Simulate $\{\epsilon^*_{T+1}, \ldots, \epsilon^*_{T+h}\}$ *i.i.d.* from F_ϵ.

Step 3 Plug the $\{\epsilon^*_t\}_{t=T+1}^{T+h}$ from Step 2 and $\{X_{T-p+1}, \ldots, X_T\}$ into Eq. (10) to obtain a pseudo-value of X_{T+h} given by $\mathscr{G}(X_T, \ldots, X_{T-p+1}; \epsilon^*_{T+1}, \ldots, \epsilon^*_{T+h})$.

Step 4 Repeat Steps 2 and 3 M times to get M pseudo values $\{X^{(1)}_{T+h}, \ldots, X^{(M)}_{T+h}\}$. The L_2 and L_1 optimal predictor can be approximated by $\frac{1}{M}\sum_{i=1}^{M} X^{(i)}_{T+h}$ and Median$(X^{(1)}_{T+h}, \ldots, X^{(M)}_{T+h})$, respectively. Furthermore, a QPI can be built by taking corresponding quantiles of the empirical distribution of $\{X^{(1)}_{T+h}, \ldots, X^{(M)}_{T+h}\}$.

We first show that the mean of pseudo values derived from Algorithm 1 can be consistent with the exactly L_2 optimal predictor. This is formalized in Theorem 2.1.

Theorem 2.1. *Under assumptions A1-A6, the point predictor of homoscedastic Eq. (1) as* $\widehat{X^{L_2}_{T+h}} = \frac{1}{M}\sum_{i=1}^{M} X^{(i)}_{T+h}$ *converges to the exactly L_2 optimal predictor almost sure as M tends to infinity. Here,* $X^{(i)}_{T+h} = \mathscr{G}(X_T, \ldots, X_{T-p+1}; \epsilon^{(i)}_{T+1}, \ldots, \epsilon^{(i)}_{T+h})$; $\{\epsilon^{(i)}_{T+1}, \ldots, \epsilon^{(i)}_{T+h}\}$ *are i.i.d. with common distribution F_ϵ for all $i = 1, \ldots, M$.*

Next, inspired by the proof of Guo, Bai, and An (1999), we can show the median of pseudo values in Algorithm 1 is also consistent with the exactly L_1 optimal predictor. We can also build a PI that is asymptotically valid with any arbitrary CVR. To achieve this goal, we need additional one mild assumption:

A8 The mean function $\phi(\cdot)$ is uniformly continuous.

This will lead to Theorem 2.2 below.

Theorem 2.2. *Under assumptions A1-A6, if we take the point predictor as* $\widehat{X^{L_1}_{T+h}} = Median\,(\{X^{(1)}_{T+h}, \ldots, X^{(M)}_{T+h}\})$, *it is consistent to the exactly L_1 optimal predictor when M converges to infinity. Here, $X^{(i)}_{T+h}$ and $\{\epsilon^{(i)}_{T+1}, \ldots, \epsilon^{(i)}_{T+h}\}$ have the same definitions with Theorem 2.1. We can further show that the QPI is asymptotically valid with any arbitrary CVR.*

2.2 Prediction inference for NLAR models with unknown parameters

We further consider the case that the NLAR model (1) has a known parametric specification but the parameter values are unknown. After estimating the model by the Least Square (LS) technique, we show that the prediction

inference can also be built with standard assumptions. We assume that the NLAR model (1) has the parametric specification

$$X_t = \phi(\boldsymbol{X}_{t-1}, \theta_1) + \sigma(\boldsymbol{X}_{t-1}, \theta_2)\epsilon_t, \tag{11}$$

where the functional form of $\phi(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ is known but the real-valued parameters θ_1 and θ_2 are unknown.

For carrying out prediction inference, we assume:

A9 The parameter estimator $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are consistent to θ_1 and θ_2 respectively.

A10 For all $\boldsymbol{x}$ in the support $\mathscr{X}$ of $\boldsymbol{X}_t$, the non-linear functions $\phi(\boldsymbol{x},\cdot)$ and $\sigma(\boldsymbol{x},\cdot)$ are both Lipschitz continuous with respect to their 2nd argument. The Lipschitz constants could be different.

A11 The probability density of innovation $f_\epsilon(x)$ satisfies $\sup_x f_\epsilon(x) < \infty$ and $\int |f_\epsilon(x) - f_\epsilon(x+c)|dx = O(c)$ for finite c.

The reason for choosing the LS method and the procedure of estimation will be discussed in Section 3.2. First, we want to show the Cumulative Distribution Function (CDF) of innovations can be approximated by the empirical CDF of residuals. Their relationship can be summarized in Lemma 2.1.

Lemma 2.1. *Under A1–A7, A9–A11, the CDF of innovation $F_\epsilon(x)$ can be approximated by the empirical CDF of residuals $\widehat{F}_\epsilon(x)$ in a way:*

$$\sup_x |\widehat{F}_\epsilon(x) - F_\epsilon(x)| \xrightarrow{p} 0, \tag{12}$$

where $\widehat{F}_\epsilon(x) := \frac{1}{T}\sum_{i=1}^{T} \mathbb{1}_{\hat{\epsilon}_i \le x}$; $\mathbb{1}(\cdot)$ is the indicator function, and we define the residual $\hat{\epsilon}_i = \frac{(X_i - \phi(\boldsymbol{X}_{i-1}, \widehat{\theta}_1))}{\sigma(\boldsymbol{X}_{i-1}, \widehat{\theta}_2)}$ for $i = 1, \ldots, T$.

With Lemma 2.1, we can build a QPI or find approximations of optimal L_1 and L_2 predictors. Of course, for this case, we need to apply the forward bootstrap prediction method. The algorithm is similar to the Algorithm 1. The difference is that we replace the true models by estimators $\phi(\boldsymbol{X}_{i-1}, \widehat{\theta}_1)$ and $\sigma(\boldsymbol{X}_{i-1}, \widehat{\theta}_2)$, respectively; we also use the corresponding residual distribution $\widehat{F}_\epsilon$ to approximate F_ϵ. The asymptotic validity of QPI and consistency of optimal L_1 or L_2 point prediction are guaranteed by Theorem 2.3 under the additional assumption of mean and volatility functions given below:

A12 For all parameter values in their respective domains, the non-linear functions $\phi(\cdot,\cdot)$ and $h(\cdot,\cdot)$ are continuous w.r.t their first argument.

Theorem 2.3. *Under A1-A7, A9–A12, let $\{X_t\}$ satisfy Eq. (11). For $h \ge 1$ we have:*

$$\sup_{|x| \le c_T} \left| F_{X^*_{T+h}|X_T,\ldots,X_{-p+1}}(x) - F_{X_{T+h}|X_T,\ldots,X_{T-p+1}}(x) \right| \xrightarrow{p} 0, \tag{13}$$

where $X^*_{T+h} = \mathscr{G}(X_T, \ldots, X_{T-p+1}; \hat{\epsilon}^*_{T+1}, \ldots, \hat{\epsilon}^*_{T+h}; \widehat{\theta})$; this is computed as $X^*_{T+k} = \phi(\boldsymbol{X}^*_{T+k-1}, \widehat{\theta}_1) + \sigma(\boldsymbol{X}^*_{T+k-1}, \widehat{\theta}_2)\hat{\epsilon}^*_{T+k}$ iteratively for $k = 1, \ldots, h$; $\{\hat{\epsilon}^*_i\}_{i=T+1}^{T+h}$ are i.i.d. $\sim \widehat{F}_\epsilon$; c_T is an appropriate sequence converges to infinity as T converges to infinity; we can take $c_T = T^\delta$ for some small $\delta < 1/2$. $F_{X^*_{T+h}|X_T,\ldots,X_{-p+1}}(x)$ is the distribution of h-step ahead future value in the bootstrap world, i.e., conditional on all observed data.

As we discussed in the Section 1, instead of adopting the fitted (traditional) residuals, we can apply the predictive residuals to compute QPI. To acquire such predictive residuals which are denoted $\hat{\epsilon}^p_t$ hereafter, we need to estimate models based on delete-X_t data, i.e., the available data for the scatter plot of X_i vs. $\{X_{i-p}, \ldots, X_{i-1}\}$ excludes the single point at $i = t$. Evaluate and collect the estimation residual at this point and repeat it for $t = 1, \ldots, T$, we obtain all predictive residuals $\{\hat{\epsilon}^p_t\}_{t=1}^T$. When T tends to infinity, the effects of leaving out one data pair X_t vs. $\{X_{t-1}, \ldots, X_{t-p}\}$ is negligible. Hence, for large T, the predictive residual $\hat{\epsilon}^p_t$ is approximately the same as the fitted residual $\hat{\epsilon}_t$. Therefore, Lemma 2.1 and Theorem 2.3 are still true with predictive residuals.

Remark. *Beyond the prediction based on optimal L_1 or L_2 loss criterion, our bootstrap prediction procedure can be extended to predict any quantile value of the future distribution of X_{T+h} when the sample size is large enough. To get such optimal prediction in practice, we can take the desired quantile of M pseudo values $\{X_{T+h}^{(1)}, \ldots, X_{T+h}^{(M)}\}$ generated by the forward bootstrap prediction method mentioned above. This type of prediction can serve the needs of predictions under asymmetric loss functions.*

3 Pertinent PIs

As shown in Theorem 2.3, it is straightforward to build a QPI for X_{T+h}. Although this type of prediction interval is asymptotically valid, it can not capture the variability arising from the model estimation. We can illustrate the necessity of including the model estimation variability by a simple example. Suppose we want to do a 1-step ahead prediction with data generated by the underlying model: $X_t = g(X_{t-1}) + \epsilon_t$; here we assume that innovations ϵ_t are *i.i.d.* and the non-linear function $g(\cdot)$ makes the series $\{X_t\}$ geometrically ergodic. If we want to build a PI for X_{T+1} based on $\{X_T, \ldots, X_1\}$. We can rely on the distribution of predictive root $\widehat{X}_{T+1} - X_{T+1}$; here we assume $\widehat{X}_{T+1}$ is the L_2-optimal prediction, i.e., $\hat{g}(X_T)$ in this context; $\hat{g}()$ is the estimation of $g()$. Thus,

$$\widehat{X}_{T+1} - X_{T+1} = \hat{g}(X_T) - g(X_T) - \epsilon_{T+1}.$$

In a large-sample case, $\hat{g}(X_T) - g(X_T)$ could be negligible. The standard QPI treats it as exactly zero and tries to explain the variability of the root as the variability of the innovation ϵ. However, for the finite sample case, the variability inherent in $\hat{g}(X_T) - g(X_T)$ needs to be captured in which the bootstrap plays a vital role. Due to this reason, we propose a new bootstrap procedure to build the so-called PPI in Algorithm 2. Analogously to the development of QPI with predictive residuals, we can also build PPI with predictive residuals.

Remark 3.1. *To build the PPI for Eq. (11) with heteroscedastic error, we need to normalize the variance of fitted/predictive residuals to 1 since we assume the innovation ϵ_1 has variance 1 if $\sigma(\cdot) \not\equiv 1$. This additional manipulation for Eq. (11) with heteroscedastic error simplifies the theoretical proof. Moreover, from a practical issue, the length of PPI will decrease with this additional step.*

3.1 Asymptotic validity of the PPIs

The idea that underlies Algorithm 2 is approximating the distribution of predictive root $X_{T+h} - \widehat{X}_{T+h}$ by its bootstrap version $X_{T+h}^* - \widehat{X}_{T+h}^*$. From a theoretical view, as we have clarified, applying fitted residuals or predictive residuals is asymptotically equivalent. Thus, in what follows, we analyze the asymptotic performance of a PPI with fitted residuals. However, the PI with predictive residuals invariably has a larger finite-sample CVR. All in all, we want to compare two predictive roots $X_{T+h} - \widehat{X}_{T+h}$ and $X_{T+h}^* - \widehat{X}_{T+h}^*$. If we can show

$$\sup_{|x| \le c_T} \left| \mathbb{P}\left(X_{T+h}^* - \widehat{X}_{T+h}^* \le x | X_T, \ldots, X_{-p+1} \right) - \mathbb{P}\left(X_{T+h} - \widehat{X}_{T+h} \le x | X_T, \ldots, X_{T-p+1} \right) \right| \xrightarrow{p} 0, \tag{14}$$

then we can utilize the distribution of $X_{T+h}^* - \widehat{X}_{T+h}^*$ to consistently estimate the distribution of $X_{T+h} - \widehat{X}_{T+h}$; as a result, the PPI has asymptotic validity within c_T. Eq. (14) can be shown based on Theorem 2.3 with one additional condition, namely that $\widehat{\theta}^* = (\widehat{\theta}_1^*, \widehat{\theta}_2^*)$ is consistent to $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)$. Before going into detail about this property, we first discuss the conditions under which $\widehat{\theta}$ is consistent to $\theta = (\theta_1, \theta_2)$.

Algorithm 2 h-steps ahead PPI of X_{T+h} for unknown homoscedastic Eq. (11) with fitted residuals

Step 1 Fit the homoscedastic model (11) based on $\{X_{-p+1}, \ldots, X_T\}$ to get parameter estimation $\widehat{\theta}_1$ which satisfies A9. Furthermore, compute and record $\hat{\epsilon}_t - T^{-1}\sum_{i=1}^{T}\hat{\epsilon}_i$ for $t = 1, \ldots, T$ to get $\widehat{F}_\epsilon$.

Step 2 Find the prediction $\widehat{X}_{T+h}$ based on the forward bootstrap method.

Step 3 (a) Resample (with replacement) the residuals from $\widehat{F}_\epsilon$ to create pseudo-errors $\{\hat{\epsilon}_t^*\}_{t=p+1}^{T}$ and $\{\hat{\epsilon}_t^*\}_{t=T+1}^{T+h}$.

(b) Let $(X_{-p+1}^*, \ldots, X_0^*)' = (X_{0+I}, \cdots, X_{p+I-1})'$ where I is generated as a discrete random variable uniformly on the values $-p+1, \ldots, T-p+1$. Then, use the fitted homoscedastic NLAR model of Step 1 and the generated $\{\hat{\epsilon}_t^*\}_{t=p+1}^{T}$ in Step 3 (a) to create bootstrap pseudo-data $\{X_t^*\}_{t=1}^{T}$ in a recursive manner, i.e., compute $X_k^* = \phi(\boldsymbol{X}_{k-1}^*, \widehat{\theta}_1) + \hat{\epsilon}_1^*$ for $k = 1, \ldots, T$.

(c) Based on the bootstrap data $\{X_t^*\}_{t=-p+1}^{T}$, re-estimate the homoscedastic NLAR model to get $\widehat{\theta}_1^*$.

(d) Guided by the idea of forward bootstrap, re-define the last p values of the bootstrap data to match the original, i.e., re-define $X_t^* = X_t$ for $t = T-p+1, \ldots, T$.

(e) With parameter estimation $\widehat{\theta}_1$, the bootstrap data $\{X_t^*\}_{t=-p+1}^{T}$, and the pseudo-errors $\{\hat{\epsilon}_t^*\}_{t=T+1}^{T+h}$ to generate recursively the future bootstrap data $X_{T+1}^*, \ldots, X_{T+h}^*$.

(f) With bootstrap data $\{X_t^*\}_{t=-p+1}^{T}$ and parameter estimation $\widehat{\theta}_1^*$, utilize the forward bootstrap method to compute the bootstrap predictor which is denoted by $\widehat{X}_{T+h}^*$. For generating innovations, we still use $\widehat{F}_\epsilon$.

(g) Determine the bootstrap root: $X_{T+h}^* - \widehat{X}_{T+h}^*$.

Step 4 Repeat Step 3 K times; the K bootstrap root replicates are collected in the form of an empirical distribution whose β-quantile is denoted $q(\beta)$. The $(1-\alpha)100\%$ equal-tailed prediction interval for X_{T+h} centered at $\widehat{X}_{T+h}$ is then estimated by $[\widehat{X}_{T+h} + q(\alpha/2), \widehat{X}_{T+h} + q(1-\alpha/2)]$.

3.2 The consistency of $\widehat{\theta}$ to θ

In this paper, we adopt the non-linear LS (NLS) technique to perform parameter estimation; the reason is that NLS is based entirely on the scatter plot of X_t vs. $(X_{t-1}, \ldots, X_{t-p})$ so that predictive residuals can be easily defined. First, we consider a homoscedastic version of Eq. (11). The heteroscedastic version will be handled by a two-step estimation approach later. To simplify notation in the proofs, we consider an NLAR model with order one, and we attempt to minimize a quadratic empirical loss function as given below:

$$\widehat{\theta}_1 = \arg\min_{\vartheta \in \Theta_1} L_T(\vartheta) \quad \text{where} \quad L_T(\vartheta) = \frac{1}{T}\sum_{t=1}^{T}(X_t - \phi(X_{t-1}, \vartheta))^2, \tag{15}$$

where Θ_1 is the domain of possible values of θ_1. With a correctly specified model, the true parameter θ_1 satisfies that:

$$\theta_1 = \arg\min_{\vartheta \in \Theta_1} L(\vartheta) \quad \text{where} \quad L(\vartheta) = \mathbb{E}(X_t - \phi(X_{t-1}, \vartheta))^2. \tag{16}$$

The consistency of the non-linear least squares estimator $\widehat{\theta}_1$ to θ_1 can be guaranteed with below additional assumptions:

A13 Θ_1 is bounded, closed and with finite dimension.

A14 θ_1 uniquely minimizes $L(\vartheta)$ over $\vartheta \in \Theta_1$.

If we can not correctly specify the model, we call θ_1 the optimal parameter in the sense of minimizing $L(\vartheta)$. We can still build the consistency relationship between $\widehat{\theta}_1$ and θ_1 if we assume A14. As we have clarified at the beginning of Section 1, we focus on the case where we can correctly specify the model. The model misspecification case can be analyzed similarly.

Lemma 3.1. *Under assumptions A1-A7 and A10, A13-A14, if $\{X_t\}$ satisfies a homoscedastic Eq. (11), the non-linear least squares estimation $\widehat{\theta}_1$ converges to the true parameter θ_1 in probability, i.e., for any $\epsilon > 0$,*

$$\mathbb{P}(|\widehat{\theta}_1 - \theta_1| > \epsilon) \to 0. \tag{17}$$

To handle the heteroscedastic Eq. (11), we still estimate θ_1 by minimizing the empirical risk Eq. (15). The corresponding true risk with respect to θ_1 is:

$$\mathbb{E}(X_t - \phi(X_{t-1}, \vartheta))^2 = \mathbb{E}(\phi(X_{t-1}, \theta_1) - \phi(X_{t-1}, \vartheta))^2 + \mathbb{E}(\sigma(X_{t-1}, \theta_2)^2), \tag{18}$$

which implies that the minimizer of risk Eq. (18) is equal to the true θ_1. Thus, the minimizer of empirical risk will still converge to the true parameter in probability. After securing this consistent estimation of θ_1, we proceed to estimate the θ_2 by minimizing the below empirical risk:

$$\widehat{\theta}_2 = \arg\min_{\vartheta \in \Theta_2} K_T(\vartheta, \widehat{\theta}_1) = \arg\min_{\vartheta \in \Theta_2} \left| \frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t - \phi(X_{t-1}, \widehat{\theta}_1)}{h(X_{t-1}, \vartheta)}\right)^2 - 1 \right|. \tag{19}$$

The corresponding true risk should be:

$$\begin{aligned} \vartheta_2 = \arg\min_{\vartheta \in \Theta_2} K(\vartheta, \theta_1) &= \arg\min_{\vartheta \in \Theta_2} \left| \mathbb{E}\left(\frac{X_t - \phi(X_{t-1}, \theta_1)}{h(X_{t-1}, \vartheta)}\right)^2 - 1 \right| \\ &= \arg\min_{\vartheta \in \Theta_2} \left| \mathbb{E}\left(\frac{\phi(X_{t-1}, \theta_1) + h(X_{t-1}, \theta_2)\epsilon_t - \phi(X_{t-1}, \theta_1)}{h(X_{t-1}, \vartheta)}\right)^2 - 1 \right| \\ &= \arg\min_{\vartheta \in \Theta_2} \left| \mathbb{E}\left(\frac{h(X_{t-1}, \theta_2)}{h(X_{t-1}, \vartheta)}\right)^2 - 1 \right|, \end{aligned} \tag{20}$$

which implies $\vartheta_2 = \theta_2$. Under the additional assumptions:

A15 Θ_2 is bounded, closed and with finite dimension.

A16 θ_2 uniquely minimizes $K(\vartheta, \theta_1)$ over $\vartheta \in \Theta_2$,

we can derive the lemma below to ensure the consistency of $\widehat{\theta}_2$ to θ_2:

Lemma 3.2. *Under assumptions A1-A7 and A10, A13-A16, if $\{X_t\}$ satisfies a heteroscedastic model (11), the NLS estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ converge respectively to the true parameters θ_1 and θ_2 in probability.*

3.3 The consistency of $\widehat{\theta}^*$ to $\widehat{\theta}$ in the bootstrap world

From the last subsection, we have seen the non-linear least squares can return a satisfied estimation but this is still not enough for us to derive the asymptotic validity of the PPI. As we discussed in Section 3.1, the necessary component is the consistency between $(\widehat{\theta}_1^*, \widehat{\theta}_2^*)$ and $(\widehat{\theta}_1, \widehat{\theta}_2)$. We first investigate the relationship between $\widehat{\theta}_1^*$ and θ_1. Once this relationship is determined, the consistency between $\widehat{\theta}_1^*$ and $\widehat{\theta}_1$ is trivial. In the work of Franke and Neumann (2000), a similar problem is considered for the regression case. In short, this consistency can be proved by showing that analogous $L_T^*(\vartheta)$ converges uniformly to $L(\vartheta)$. In our case, $L_T^*(\vartheta)$ has the form as below:

$$L_T^*(\vartheta) = \frac{1}{T}\sum_{t=1}^{T}(X_t^* - \phi(X_{t-1}^*, \vartheta))^2, \tag{21}$$

here $\{X_t^*\}$ is the bootstrap series, which mimics the property of the true series and $\widehat{\theta}_1^*$ satisfies that:

$$\widehat{\theta}_1^* = \arg\min_{\vartheta\in\Theta_1} L_T^*(\vartheta) = \arg\min_{\vartheta\in\Theta_1} \frac{1}{T}\sum_{t=1}^{T}(X_t^* - \phi(X_{t-1}^*, \vartheta))^2. \tag{22}$$

As discussed in Section 3.2, it is necessary that we have the additional condition: the bootstrap series is also geometrically ergodic. Then, with close enough empirical residual distribution and true innovation distribution, we may show the uniform convergence of $L_T^*(\vartheta)$ and $L(\vartheta)$. Then, the consistency of $\widehat{\theta}_1^*$ to θ_1 is easily found.

The first problem we face is that the probability density of the fitted residual $\hat{\epsilon}$ is not continuous and positive everywhere which means the basic assumption A4 needed to show the ergodicity of the bootstrap series is not met. In a similar situation, Franke, Kreiss, Mammen, and Neumann (2002) apply the kernel technique to build a probability density of $\hat{\epsilon}$. Here, we take a convolution approach to make the density function of empirical residual continuous and positive everywhere, i.e., we define another random variable $\widetilde{\epsilon}_i$ which is the sum of empirical fitted residual $\hat{\epsilon}_i$ and an independent normal random variable with mean 0 and suitable variance $\xi(T)$, i.e., let:

$$\widetilde{\epsilon}_i = \hat{\epsilon}_i + z_i, \text{ for } i = 1, \ldots, T, \tag{23}$$

where $z_i \sim N(0, \xi(T))$, where $\xi(T)$ converges to 0 as $T \to \infty$ with a suitable rate. Then, we create bootstrap residuals by drawing *i.i.d.* from $\widetilde{F}_\epsilon$, the CDF of $\{\widetilde{\epsilon}_i\}$, in order to build a bootstrap series $\{\widetilde{X}_t^*\}$ as we did in Algorithm 2. Subsequently, we re-estimate the parameter of NLAR based on the bootstrap series $\{\widetilde{X}_t^*\}$. Since the convergence in mean squares implies the convergence in probability, we can easily see that Lemma 2.1 still stands true for $\widetilde{F}_\epsilon$.

Remark. *Here, we take a convolution approach to create residuals that possess a continuous probability density function. We should notice that this approach is equivalent to the kernel density estimator taken by Franke, Kreiss, Mammen, and Neumann (2002) and Franke, Kreiss, and Mammen (2002) with a Gaussian kernel. More specifically, the variance $\xi(T)$ plays the same role as the bandwidth h in the Gaussian kernel. Thus, if we take $\xi(T) = O(T^{-\delta'})$ for some constant $\delta' > 0$, we can show that the probability density of $\widetilde{\epsilon}$, $\widetilde{f}_\epsilon$ converges uniformly to f_ϵ,i.e., $||\widetilde{f}_\epsilon - f_\epsilon||_\infty = op(1)$, see the proof of Lemma 4 from Franke, Kreiss, Mammen, and Neumann (2002) for a reference. In addition, this convolution technique can also be applied to predictive residuals. As we have discussed in Section 3.1, the predictive residual is equivalent to the fitted residual asymptotically.*

Although we will use $\widetilde{F}_\epsilon$ and the corresponding density function $\widetilde{f}_\epsilon$ to develop subsequent theorems, in practice we still apply $\widehat{F}_\epsilon$ since effects stemming from z_i are negligible when we sample innovations from the empirical residual distribution. For simplifying notations, we keep using $\widehat{F}_\epsilon$ and $\widehat{f}_\epsilon$, though their representation may change according to the context. Similarly with the deduction of Lemma 3.1, we can get:

$$\mathbb{P}(|\widehat{\theta}_1^* - \theta_1| > \epsilon) \leq \mathbb{P}(2 \sup_{\vartheta \in \Theta_1} |L_T^*(\vartheta) - L(\vartheta)| > C), \tag{24}$$

for some constant $C > 0$. Focusing on analyzing $\sup_{\vartheta \in \Theta_1} |L_T^*(\vartheta) - L(\vartheta)|$, we can partition the parameter space into different balls, i.e., make a ε-covering of Θ_1. Let the ε-covering number of Θ be $C_N = N(\varepsilon; \Theta_1; ||\cdot||)$ which means for every $\vartheta \in \Theta_1$, $\exists\ i \in \{1, 2, \ldots, C_N\}$ s.t. $||\vartheta - \theta^i|| \leq \varepsilon$ for $\forall \varepsilon > 0$. Define $\Xi_\theta \in \{\theta^1, \ldots, \theta^{C_N}\}$, we can consider:

$$\sup_{\vartheta \in \Theta_1} |L_T^*(\vartheta) - L(\vartheta)| \leq \max_{\Xi_\theta \in \{\vartheta^1, \ldots, \vartheta^{C_N}\}} |L_T^*(\Xi_\theta) - L(\Xi_\theta)| + \sup_{\vartheta \in \Theta_1} |L_T^*(\vartheta) - L_T^*(\Xi_\theta)| + \sup_{\vartheta \in \Theta_1} |L(\vartheta) - L(\Xi_\theta)|. \tag{25}$$

Consider the second term of the r.h.s. of the above inequality. From Lipschitz continuous assumption of $\phi(\cdot, \cdot)$ with respect to ϑ, we can get:

$$\sup_{\vartheta \in \Theta_1} |L_T^*(\vartheta) - L_T^*(\Xi_\theta)| \leq \sup_{\vartheta \in \Theta_1} L||\vartheta - \Xi_\theta|| \leq L \cdot \varepsilon \to 0. \tag{26}$$

Similarly, we can find the $\sup_{\vartheta \in \Theta} |L(\vartheta) - L(\Xi_\theta)| \to 0$. For the first term of the r.h.s of Eq. (25), if we can show the bootstrap series is also ergodic when the parameter is fixed, then we actually have $L_T^*(\Xi_\vartheta) \xrightarrow{p} \mathbb{E}^*[X_1^* - \phi(X_0^*, \Xi_\vartheta)]$, such a similar result is also implied by Theorem 5 of Franke, Kreiss, Mammen, and Neumann (2002), here $\mathbb{E}^*(\cdot)$ stands for the conditional expectation in the bootstrap world. Therefore, for getting the uniform convergence of $L_T^*(\vartheta)$ to $L(\vartheta)$ in probability, it is enough to show:

$$\mathbb{E}^*[X_1^* - \phi(\boldsymbol{X}_0^*, \Xi_\theta)]^2 \xrightarrow{p} \mathbb{E}[X_1 - \phi(\boldsymbol{X}_0, \Xi_\theta)]^2, \text{ for each } \Xi_\theta. \tag{27}$$

For notational simplicity, we consider an NLAR(1) model; then, the l.h.s. of Eq. (27) equals:

$$\int_{\mathbb{R}^2} (x_1 - \phi(x_0, \Xi_\theta))^2 \widehat{f}_\epsilon(x_1 - \phi(x_0, \widehat{\theta}_1)) \pi^*(x_0) dx_1 dx_0, \tag{28}$$

where $\pi^*(\cdot)$ stands for the marginal stationary density function of the bootstrap series. As we can see, the uniform convergence of Eq. (25) in probability depends on the ergodic property of the bootstrap series and the closeness of $\pi^*(\cdot)$ and $\pi(\cdot)$ which is the true stationary density function of the real series. In other words, the ergodic property of the bootstrap series is not enough to get our desired result. We also require the stationary distribution of the bootstrap series and the real series should be close enough to show Eq. (27). Here, we develop a theorem to illustrate the required conditions.

Theorem 3.1. *Suppose that the data generating process obeys Eq. (11) and the bootstrap time series* $\{X_t^*\}_{t=-p+1}^T$ *are generated by our forward bootstrap methods. Under A1-A7, A9-A12, we have:*

$$\mathbb{E}^*[X_1^* - \phi(X_0^*, \Xi_\theta)]^2 \xrightarrow{p} \mathbb{E}[X_1 - \phi(X_0, \Xi_\theta)]^2. \tag{29}$$

Then, the following Corollary 3.1 is trivial:

Corollary 3.1. *Under assumptions A1-A7, A9-A12,* $\widehat{\theta}^*$ *is consistent to* $\widehat{\theta}$ *with both fitted and predictive residuals, which substantiates Eq. (14). Thus, the conditional distribution of* $X_{T+h} - \widehat{X}_{T+h}$ *can be asymptotically approximated by the conditional distribution of* $X_{T+h}^* - \widehat{X}_{T+h}^*$ *which guarantees the validity of PPI.*

3.4 The estimation inference of $\widehat{\theta}$ and $\widehat{\theta}^*$

With a more complicated prediction procedure in Algorithm 2, we expect to get a stronger property, i.e., pertinence. The crucial part behind the pertinence is that we can approximate the distribution of $\widehat{\theta}$ by the distribution of $\widehat{\theta}^*$. In other words, the estimation variability can be captured by the bootstrap-based PI. To derive the estimation inference, we need stronger assumptions than A10 on the mean and variance function. We assume:

A17 For all $\boldsymbol{x}$ in the support $\mathscr{X}$ of $\boldsymbol{X}_t$, $\phi(\boldsymbol{x},\cdot)$ and $\sigma(\boldsymbol{x},\cdot)$ are twice differentiable w.r.t. parameters in some neighborhood of true parameters.

A18 If we write $L_T(\vartheta)$ as $\frac{1}{T}\sum_{t=1}^{T} q_t(\vartheta)$, we need $\mathbb{E}\nabla^2\ q_1(\theta_1)$ is non-singular; If we write $K_T(\vartheta,\theta_1)$ as $\frac{1}{T}\sum_{t=1}^{T} q_t(\vartheta,\theta_1)$, we need $\mathbb{E}\nabla^2\ q_1(\theta_2,\theta_1)$ is non-singular.

A19 The first order condition of minimizing the empirical risk function satisfies that $\nabla L_T(\widehat{\theta}_1) = op(T^{-1/2})$ and $\nabla K_T(\widehat{\theta}_2,\widehat{\theta}_1) = op(T^{-1/2})$. Similarly, we assume we can achieve such accuracy for optimization in the bootstrap world.

A19 implies that the first-order conditions for minimizing criterion functions hold approximately since it may be hard to find exactly $\widehat{\theta}_1$ or $\widehat{\theta}_2$. Then, we first develop the estimation inference of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ in the Theorem 3.2.

Theorem 3.2. *Under A1-A7, A13-A19, with consistent parameter estimations $\widehat{\theta}_1$ and $\widehat{\theta}_2$, we have:*

$$\sqrt{T}(\widehat{\theta}_1 - \theta_1) \xrightarrow{d} N(0, B_1^{-1}\Omega_1 B_1^{-1}), \tag{30}$$

where $\Omega_1 = 4\cdot\mathbb{E}(\sigma(X_0,\theta_2)R_1\sigma(X_0,\theta_2))$; $B_1 = 2\cdot\mathbb{E}\left(\nabla\phi(X_0,\theta_1)(\nabla\phi(X_0,\theta_1))^\top\right)$; $R_1 = \nabla\phi(X_0,\theta_1)\ \nabla\phi(X_0,\theta_1)^\top$; ∇ is the gradient operator w.r.t. θ_1. Similarly, we can analyze the distribution of parameter estimation $\widehat{\theta}_2$:

$$\sqrt{T}(\widehat{\theta}_2 - \theta_2) \xrightarrow{d} N(0, B_2^{-1}\Omega_2 B_2^{-1}), \tag{31}$$

where $\Omega_2 = 4\cdot\mathbb{E}(B_3R_2B_3^\top)$; $B_3 = \mathbb{E}(\nabla g(X_1,X_0,\theta_2,\theta_1))$; $R_2 = (g(X_1,X_0,\theta_2,\theta_1)-1)^2$; $B_2 = 2\cdot(\mathbb{E}(\nabla g(X_1,X_0,\theta_2,\theta_1))\cdot(\mathbb{E}(\nabla g(X_1,X_0,\theta_2,\theta_1))^\top$; $g(X_1,X_0,\theta_2,\theta_1) = \left(\frac{X_1-\phi(X_0,\theta_1)}{\sigma(X_0,\theta_2)}\right)^2$; here ∇ is the gradient operator w.r.t. θ_2.

Remark. *From here, we can see the distribution of $\widehat{\theta}_1$ depends on the time series structure. If we do not assume that we can specify the correct model format, the covariance matrix of the parameter's asymptotic distribution will depend on the whole structure of the time series. This is the reason why we need the forward bootstrap to generate time series and do estimation in the bootstrap world, otherwise, we can not approximate the covariance term well.*

In the bootstrap world, we can perform a similar parameter estimation procedure as we did in Section 3.3. As we have seen in the proof of Theorem 3.1, for $(X_{-p+1},\ldots,X_T)\in\Omega_T$, where $\Omega_T\subseteq\mathbb{R}^{T+p}$ and $\mathbb{P}((X_{-p+1},\ldots,X_T)\notin\Omega_T) = o(1)$, under the consistency of parameter estimation in the real world, the bootstrap series is also ergodic in the sense of β-mixing. Thus, we can have below consistency results in the bootstrap world:

$$\begin{aligned}
&\nabla^2 L_T^*(\widetilde{\theta}_1^*) = \frac{1}{T}\nabla^2\sum_{t=1}^{T}\left(X_t^* - \phi(X_{t-1}^*,\widetilde{\theta}_1^*)\right)^2 \xrightarrow{p} 2\cdot\mathbb{E}^*\left(\nabla\phi(X_0^*,\widehat{\theta}_1)(\nabla\phi(X_0^*,\widehat{\theta}_1))^\top\right) = B_1^*;\\
&\nabla^2 K_T^*(\widetilde{\theta}_2^*,\widehat{\theta}_1^*) = 2\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\nabla g^*(X_t^*,X_{t-1}^*,\widetilde{\theta}_2^*,\widehat{\theta}_1^*)\right)\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\nabla g^*(X_t^*,X_{t-1}^*,\widetilde{\theta}_2^*,\widehat{\theta}_1^*)\right)^\top\\
&+2\cdot\left(\frac{1}{T}\sum_{t=1}^{T} g^*(X_t^*,X_{t-1}^*,\widetilde{\theta}_2^*,\widehat{\theta}_1^*) - 1\right)\cdot\left(\frac{1}{T}\sum_{t=1}^{T}\nabla^2 g^*(X_t^*,X_{t-1}^*,\widetilde{\theta}_2^*,\widehat{\theta}_1^*)\right)\\
&\xrightarrow{p} 2\cdot\mathbb{E}^*\left(\nabla g^*(X_1^*,X_0^*,\widehat{\theta}_2,\widehat{\theta}_1)\right)\mathbb{E}^*\left(\nabla g^*(X_1^*,X_0^*,\widehat{\theta}_2,\widehat{\theta}_1)\right)^\top = B_2^*,
\end{aligned} \tag{32}$$

where $\widetilde{\theta}_1^*$ is between $\widehat{\theta}_1^*$ and $\widehat{\theta}_1$; $\widetilde{\theta}_2^*$ is between $\widehat{\theta}_2^*$ and $\widehat{\theta}_2$; hence $\widetilde{\theta}_1^*$ and $\widetilde{\theta}_2^*$ also converge to θ_1 and $\theta_2$in probability, respectively. It is easily to find $B_1^* \to B_1$ and $B_2^* \to B_2$ for $(X_{-p+1}, \ldots, X_T) \in \Omega_T$. To simplify the result about $\nabla^2 K_T^*(\widetilde{\theta}_2^*, \widehat{\theta}_1^*)$, we need the variance of ϵ_1^* to be one to remove the second term. This is guaranteed since we normalize the variance of the residuals to 1 when we perform the bootstrap prediction algorithms for models with heteroscedastic errors; see Remark 3.1. Another advantage of this manipulation comes from analyzing $\nabla K_T^*(\widehat{\theta}_2, \widehat{\theta}_1) = 2 \cdot \left(\frac{1}{T}\sum_{t=1}^T g^*(X_t^*, X_{t-1}^*, \widehat{\theta}_2, \widehat{\theta}_1) - 1\right) \cdot \left(\frac{1}{T}\sum_{t=1}^T \nabla g^*(X_t^*, X_{t-1}^*, \widehat{\theta}_2, \widehat{\theta}_1)\right)$. With this additional manipulation, $\mathbb{E}^*(g^*(X_t^*, X_{t-1}^*, \widehat{\theta}_2, \widehat{\theta}_1) - 1)$ is 0, which implies that the asymptotic distribution of $\nabla K_T^*(\widehat{\theta}_2, \widehat{\theta}_1^*)$ has mean 0. By the CLT for a triangular array of strongly mixing series given in Politis, Romano, and Wolf (1997), we can further show:

$$\begin{aligned} &\sqrt{T}\nabla L_T^*(\widehat{\theta}_1) \xrightarrow{d} N(0, \Omega_1); \\ &\sqrt{T}\nabla K_T^*(\widehat{\theta}_2, \widehat{\theta}_1^*) \xrightarrow{d} N(0, \Omega_2). \end{aligned} \tag{33}$$

The required assumptions can be checked in the same way shown in Theorem 5 of Franke, Kreiss, Mammen, and Neumann (2002). All in all, we can develop estimation inference for parameter estimation in the bootstrap world, i.e., Corollary 3.2 as below:

Corollary 3.2. *If we restrict on observed data $\{X_{-p+1}, \ldots, X_T\} \in \Omega_T$, where $\mathbb{P}((X_{-p+1}, \ldots, X_T) \notin \Omega_T) = o(1)$ as $T \to \infty$, under assumptions of Theorem 3.2, we can further build the estimation inference of parameter estimations in the bootstrap world, i.e., we have:*

$$\begin{aligned} &\sqrt{T}(\widehat{\theta}_1^* - \widehat{\theta}_1) \xrightarrow{d} N(0, B_1^{-1}\Omega_1 B_1^{-1}); \\ &\sqrt{T}(\widehat{\theta}_2^* - \widehat{\theta}_2) \xrightarrow{d} N(0, B_2^{-1}\Omega_2 B_2^{-1}). \end{aligned} \tag{34}$$

Theorem 3.2 and Corollary 3.2 together guarantee the pertinence of PPI returned by Algorithm 2 with *high probability*. The notable advantage of this type of PI will be illustrated in Sections 4 and 5.

4 Simulations

In this section, we deploy simulations to check the performance of our bootstrap point predictions and the performance of various PIs in R platform for a finite sample size. We first consider a simple case: NLAR model with order one and homoscedastic errors. We present the model below:

$$X_t = a + \log(b + |X_{t-1}|) + \epsilon_t, b > 0, \tag{35}$$

where ϵ_t satisfies A2. Assuming that we have observed series $\{X_1, \ldots, X_T\}$, we want to predict the value of X_{T+h}. As pointed out before, the exactly L_2 optimal predictor is the conditional mean of X_{T+h}:

$$\mathbb{E}(X_{t+h}|X_1, \ldots, X_T) = \int \cdots \int \mathscr{G}(X_T, \epsilon_{T+1}, \ldots, \epsilon_{T+h}) dF_{\epsilon_{T+1}} \cdots dF_{\epsilon_{T+h}}, \tag{36}$$

where $\mathscr{G}(X_T, \epsilon_{T+1}, \ldots, \epsilon_{T+h})$ represents the analytic formula of X_{T+h} which can be obtained by computing $X_{T+k} = a + \log(b + |X_{T+k-1}|) + \epsilon_{T+k}$ for $k = 1, \ldots, h$ iteratively. When we know the NLAR model and the innovation distribution, Eq. (36) can be computed by multiple-integration directly. However, to avoid the computational difficulty, we take the simulation repeating number $M = 1000$ to get a satisfying approximation. According to the forward bootstrap prediction method, we also do 1000 times bootstrap to get the prediction when the model and innovation are unknown. Starting from a simple example, we consider $a = 0.2, b = 0.5$ and $\{\epsilon_i\} \sim N(0, 1)$. For the prediction horizon, we consider $h = 1, 2, \ldots, 5$. To generate the data of Eq. (35), we take $X_0 \sim \text{Uniform}(-1, 1)$, then generate a series with size $B + T$; B is a large enough burn-in number to remove the effects of the initial distribution of X_0.

To see the crucial difference between the prediction of LAR and NLAR models, we apply two naive prediction methods which predict X_{T+h} of Eq. (35) by computing $X_{T+k} = a + \log(b + |X_{T+k-1}|)$ or $X_{T+k} = \widehat{a} + \log(\widehat{b} + |X_{T+k-1}|)$ repeatedly for $k = 1, \ldots, h$; $\widehat{a}$ and $\widehat{b}$ are estimators of a and b, respectively. In total, we compare four methods to make predictions. We call them (1) Simulation, with a known model and innovation; (2) Bootstrap, with an unknown model and innovation; (3) True Naive Prediction—naive prediction with the known model; (4) Estimated Naive Prediction—naive prediction with the estimated model. The simulation (1) method is "oracle" since we assume that model and innovation information are known to us. We set the burn-in number $B = 1000$ and $T = 400$. To get a comprehensive comparison, we repeat the above experiment $N = 5000$ times and compute the MSPE of various predictions based on the below formula.

$$\text{MSPE of the } h\text{-th ahead prediction} = \frac{1}{N}\sum_{n=1}^{N}(X_{n,h} - P_{n,h})^2, \text{ for } h = 1, \ldots, 5, \tag{37}$$

where $P_{n,h}$ represents h-th step ahead predictions implied by four approaches and $X_{n,h}$ stands for the true future value in the n-th replication. All MSPE values are presented in Table 1.

Table 1: The MSPE of all prediction methods with $N(0, 1)$ innovation under Model Eq. (35)

Prediction Horizon	1	2	3	4	5
L_2-Simulation	0.9595	1.2357	1.2101	1.1905	1.2153
L_1-Simulation	0.9594	1.2360	1.2107	1.1901	1.2156
L_2-Bootstrap	0.9639	1.2390	1.2144	1.1958	1.2181
L_1-Bootstrap	0.9640	1.2406	1.2158	1.1960	1.2193
True Naive	0.9596	1.3748	1.4894	1.5581	1.6309
Estimated naive	0.9641	1.3826	1.4910	1.5518	1.6084

We can find that the MSPE of simulation- and bootstrap-based L_1 or L_2 optimal predictions are very close, respectively. Since the bootstrap optimal prediction is obtained with an estimated model and innovation distribution, it is not surprising that the MSPE is slightly larger than the simulation-based (oracle) optimal prediction, no matter if L_2 or L_1 is the loss criterion. In our expectation, the MSPE of simulation- and bootstrap-based prediction are all smaller than the MSPE of two naive predictions. The importance of including the innovation effect in NLAR prediction is highlighted. Rather than applying the standard normal innovation distribution, we also researched the MSPE of different methods with a skewed innovation distribution, e.g., $\epsilon_t \sim \chi^2(3) - 3$, the relative performance of different prediction methods is consistent with results implied by Table 1.

Beyond analyzing the performance of point prediction, we are also interested in measuring prediction accuracy by building PIs. As discussed before, we can build two types of bootstrap-based prediction intervals: (1) Quantile PI; (2) Pertinent PI. The advantage of pertinent PI is that it can be centered at the optimal L_2 or L_1 predictors. Moreover, it includes the estimation error of parameters into consideration, which means a superior empirical CVR, especially in short data size situations. We take $K = 1000$ in Algorithm 2 to derive pertinent PI. We repeat experiment $N = 5000$ times and set significance level $\alpha = 0.05$. Then, we compute empirical CVR of bootstrap-based QPI and PPI for $h = 1, \ldots, 5$ step ahead predictions with the below formula:

$$\text{CVR of the } h\text{-th ahead prediction} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{X_{n,h}\in[L_{n,h},U_{n,h}]}, \text{for } h = 1, \ldots, 5, \tag{38}$$

where $[L_{n,h}, U_{n,h}]$ and $X_{n,h}$ represent h-th step ahead prediction intervals and the true future value in the n-th replication, respectively. We denote all considered PIs by (1) QPI-f, QPI with fitted residuals; (2) QPI-p, QPI

with predictive residuals; (3) L_2-PPI-f, PPI centered at L_2 optimal predictor with fitted residuals; (4) L_2-PPI-p, PPI centered at L_2 optimal predictor with predictive residuals; (5) L_1-PPI-f, PPI centered at L_1 optimal predictor with fitted residuals; (6) L_1-PPI-p, PPI centered at L_1 optimal predictor with predictive residuals; (7) SPI, which is QPI based on simulations. In addition, since building a valid PI is more challenging work, we take seven different models to check the feasibility of our methods:

- Model 1: $X_t = (0.1 \cdot X_{t-1})I(X_{t-1} \leq 0) + (0.8 \cdot X_{t-1})I(X_{t-1} > 0) + \epsilon_t$.
- Model 2: $X_t = (0.5 \cdot X_{t-1} + 0.2 \cdot X_{t-2} + 0.1 \cdot X_{t-3})I(X_{t-1} \leq 0) + (0.8X_{t-1})I(X_{t-1} > 0) + \epsilon_t$.
- Model 3: $X_t = (0.1 \cdot X_{t-1} + 0.5 \cdot e^{-X_{t-1}^2}\epsilon_t)I(X_{t-1} \leq 0) + (0.8 \cdot X_{t-1} + 0.5 \cdot e^{-X_{t-1}^2}\epsilon_t)I(X_{t-1} > 0)$.
- Model 4: $X_t = 0.2 + \log(0.5 + |X_{t-1}|) + \epsilon_t$.
- Model 5: $X_t = 2 \cdot \log(X_{t-1}^2) + \epsilon_t$.
- Model 6: $X_t = \log(10 + 5 \cdot e^{0.9 \cdot X_{t-1}}) + \epsilon_t$.
- Model 7: $X_t = \log(4 \cdot e^{0.9 \cdot X_{t-2}} + 5 \cdot e^{0.9 \cdot X_{t-1}} + 6 \cdot e^{0.9 \cdot X_{t-3}}) + \epsilon_t$,

where $\epsilon_t \sim N(0, 1)$ and $I(X_{t-1} \leq 0)$ is the indicator function which equals to 0 when $X_{t-1} \leq 0$ and 1 otherwise. Throughout the simulation studies, we pretend that all coefficients except threshold values of these 7 models are unknown to build PIs based on bootstrap methods. Besides the analyses of CVR, we are also concerned about the average length of PIs of different methods. In practice, a wide PI is less useful even though it has great coverage probability. We define the average length (LEN) of PIs as below:

$$\text{LEN of the } h\text{-th ahead PI} = \frac{1}{N}\sum_{n=1}^{N}(U_{n,h} - L_{n,h}), \text{for } h = 1, \ldots, 5, \tag{39}$$

where $U_{n,h}$ and $L_{n,h}$ are higher and lower bounds of the h-th step ahead PI in the n-th replication, respectively. Accordingly, we present LEN of different PIs along with CVR in Tables 2 to 8.

Remark. *We should clarify that the CVR computed by Eq. (38) is the unconditional coverage rate of X_{T+h} since it is an average of the conditional coverage of X_{T+h} for all replications. Also, when the sample size is small, we may get parameter estimations that make the time series close to being unstationary, especially for estimating different regions of a threshold model where the sample size further decreases. This will destroy our prediction process when multi-step ahead predictions are required. Thus, we redo the simulation once we find such abnormal larger or smaller predictions.*

From these simulations, the first thing we can notice is that all CVR for SPI is great and close to the nominal coverage level even for short data, which implies the simulation-based approach works well once we know the true model and innovation distribution. For $T = 400$, we can find all PPIs work well and are even competitive compared to the SPI. On the other hand, the QPI with fitted residuals is the worst one, especially for complicated Models 6 and 7. By applying predictive residuals, the CVR gets improved for QPI. For $T = 100$, no matter with fitted or predictive residuals, PPIs dominate QPIs. For $T = 50$, the gap between QPI and PPI also gets amplified. For the LEN of different PIs, we can find that the LENs of SPIs are barely changed for a specific model with various sample sizes. For PPI, although its LEN tends to be slightly larger than the LEN of SPI and QPI, it is the best bootstrap-type PI according to the CVR. Based on these simulation results, we summarize some important conclusions below:

- If we know the parameters of the model and innovation distribution, SPI can work well and give accurate CVR even for short data, but it is usually unrealistic in practice.
- If we do not have model information and the data is short, PPI with predictive residuals is the best method, which can give competitive performance compared to SPI. On the other hand, the QPI can not cover future values well and its CVR is severely lower than the nominal level.

- If we do not have model information and the data in hand is large enough, both QPI and PPI work well.
- Since in practice we can not judge whether the data in hand is large enough for the problem at hand, using the PPI (with predictive residuals) is recommendable.

Remark. *To perform bootstrap-based prediction, we ran simulations in a parallel fashion using 30 Xeon(R) E5-2630 CPUs. Besides, we should notice that the constant parameter inside the* log *function of Model 6 is the hardest one to estimate, since the low change rate of the partial derivative. This may be the reason for the relatively poor performance of bootstrap-based prediction methods on Model 6.*

Table 2: The CVR and LEN of PIs for Model 1

Model 1:	$X_t = (0.1 \cdot X_{t-1})I(X_{t-1} \leq 0) + (0.8 \cdot X_{t-1})I(X_{t-1} > 0) + \epsilon_t$									
	CVR for each step					LEN for each step				
$T = 400$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.9456	0.9472	0.9478	0.9496	0.9484	3.89	4.60	4.87	5.01	5.07
QPI-p	0.9470	0.9468	0.9478	0.9502	0.9500	3.91	4.62	4.90	5.03	5.09
L_2-PPI-f	0.9474	0.9454	0.9486	0.9500	0.9514	3.90	4.61	4.89	5.03	5.09
L_2-PPI-p	0.9474	0.9480	0.9480	0.9510	0.9526	3.92	4.63	4.92	5.05	5.12
L_1-PPI-f	0.9468	0.9456	0.9494	0.9494	0.9520	3.90	4.61	4.89	5.03	5.09
L_1-PPI-p	0.9464	0.9468	0.9484	0.9512	0.9530	3.92	4.63	4.92	5.05	5.12
SPI	0.9484	0.9474	0.9500	0.9502	0.9546	3.90	4.62	4.91	5.04	5.10
$T = 100$										
QPI-f	0.9388	0.9408	0.9362	0.9328	0.9330	3.86	4.52	4.78	4.91	4.97
QPI-p	0.9438	0.9446	0.9394	0.9366	0.9352	3.94	4.61	4.88	5.00	5.07
L_2-PPI-f	0.9416	0.9424	0.9382	0.9350	0.9358	3.91	4.58	4.85	4.98	5.05
L_2-PPI-p	0.9478	0.9478	0.9442	0.9396	0.9402	3.99	4.67	4.94	5.08	5.15
L_1-PPI-f	0.9428	0.9430	0.9376	0.9346	0.9358	3.91	4.58	4.85	4.97	5.04
L_1-PPI-p	0.9476	0.9482	0.9442	0.9402	0.9404	3.99	4.67	4.94	5.07	5.14
SPI	0.9502	0.9482	0.9464	0.9468	0.9460	3.90	4.61	4.89	5.03	5.09
$T = 50$										
QPI-f	0.9168	0.9248	0.9204	0.9106	0.9218	3.74	4.44	4.69	4.81	4.87
QPI-p	0.9296	0.9360	0.9334	0.9238	0.9324	3.91	4.63	4.90	5.02	5.09
L_2-PPI-f	0.9306	0.9318	0.9268	0.9176	0.9306	3.91	4.57	4.83	4.96	5.04
L_2-PPI-p	0.9402	0.9438	0.9392	0.9292	0.9390	4.07	4.76	5.04	5.18	5.26
L_1-PPI-f	0.9302	0.9314	0.9264	0.9170	0.9300	3.91	4.56	4.82	4.95	5.02
L_1-PPI-p	0.9390	0.9438	0.9366	0.9290	0.9364	4.08	4.75	5.03	5.16	5.24
SPI	0.9486	0.9492	0.9508	0.9452	0.9464	3.90	4.61	4.90	5.03	5.09

Table 3: The CVR and LEN of PIs for Model 2

Model 2:	$X_t = (0.5 \cdot X_{t-1} + 0.2 \cdot X_{t-2} + 0.1 \cdot X_{t-3})I(X_{t-1} \leq 0) + (0.8 \cdot X_{t-1})I(X_{t-1} > 0) + \epsilon_t$									
	CVR for each step					LEN for each step				
$T = 400$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.9420	0.9506	0.9468	0.9444	0.9372	3.88	4.68	5.11	5.40	5.58
QPI-p	0.9462	0.9512	0.9502	0.9474	0.9428	3.92	4.72	5.16	5.45	5.64
L_2-PPI-f	0.9446	0.9510	0.9486	0.9470	0.9408	3.90	4.71	5.15	5.44	5.63
L_2-PPI-p	0.9466	0.9542	0.9516	0.9494	0.9434	3.94	4.75	5.20	5.49	5.69
L_1-PPI-f	0.9448	0.9518	0.9478	0.9468	0.9402	3.90	4.71	5.15	5.44	5.62
L_1-PPI-p	0.9470	0.9544	0.9500	0.9486	0.9436	3.94	4.75	5.20	5.49	5.68
SPI	0.9446	0.9534	0.9508	0.9510	0.9454	3.90	4.71	5.16	5.46	5.65
$T = 100$										
QPI-f	0.9270	0.9304	0.9294	0.9272	0.9250	3.81	4.57	4.98	5.23	5.40
QPI-p	0.9370	0.9412	0.9368	0.9372	0.9372	3.98	4.76	5.19	5.46	5.63
L_2-PPI-f	0.9358	0.9352	0.9338	0.9314	0.9298	3.95	4.71	5.13	5.40	5.59
L_2-PPI-p	0.9454	0.9454	0.9444	0.9430	0.9418	4.10	4.90	5.34	5.63	5.83
L_1-PPI-f	0.9364	0.9360	0.9336	0.9310	0.9304	3.95	4.71	5.13	5.39	5.58
L_1-PPI-p	0.9450	0.9456	0.9432	0.9422	0.9412	4.11	4.90	5.33	5.62	5.81
SPI	0.9446	0.9472	0.9498	0.9474	0.9478	3.90	4.71	5.16	5.46	5.65
$T = 50$										
QPI-f	0.8980	0.9054	0.9018	0.8950	0.8926	3.66	4.47	4.87	5.14	5.38
QPI-p	0.9260	0.9314	0.9272	0.9218	0.9212	4.05	4.97	5.42	5.74	5.99
L_2-PPI-f	0.9340	0.9268	0.9214	0.9164	0.9152	4.22	5.10	5.86	6.89	8.97
L_2-PPI-p	0.9522	0.9478	0.9404	0.9400	0.9376	4.60	5.57	6.36	7.33	9.03
L_1-PPI-f	0.9338	0.9268	0.9194	0.9144	0.9130	4.23	5.09	5.82	6.79	8.71
L_1-PPI-p	0.9522	0.9482	0.9384	0.9378	0.9356	4.61	5.55	6.30	7.20	8.71
SPI	0.9494	0.9448	0.9464	0.9458	0.9462	3.90	4.71	5.16	5.46	5.65

Table 4: The CVR and LEN of PIs for Model 3

Model 3:	$X_t = (0.1 \cdot X_{t-1} + 0.5 \cdot e^{-X_{t-1}^2} \cdot \epsilon_t) I(X_{t-1} \leq 0) + (0.8 \cdot X_{t-1} + 0.5 \cdot e^{-X_{t-1}^2} \cdot \epsilon_t) I(X_{t-1} > 0)$									
	CVR for each step					LEN for each step				
$T = 400$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.9478	0.9442	0.9526	0.9444	0.9418	1.47	1.74	1.82	1.84	1.85
QPI-p	0.9474	0.9486	0.9504	0.9444	0.9432	1.47	1.74	1.82	1.84	1.85
L_2-PPI-f	0.9520	0.9488	0.9542	0.9436	0.9434	1.59	2.22	2.24	2.30	2.29
L_2-PPI-p	0.9510	0.9486	0.9522	0.9454	0.9446	1.64	2.37	2.37	2.44	2.42
L_1-PPI-f	0.9514	0.9480	0.9540	0.9448	0.9440	1.63	1.88	2.10	2.17	2.18
L_1-PPI-p	0.9480	0.9514	0.9530	0.9474	0.9448	1.68	1.92	2.19	2.27	2.28
SPI	0.9500	0.9500	0.9516	0.9444	0.9442	1.47	1.74	1.82	1.84	1.85
$T = 100$										
QPI-f	0.9344	0.9388	0.9420	0.9390	0.9372	1.47	1.73	1.81	1.84	1.85
QPI-p	0.9318	0.9348	0.9404	0.9392	0.9378	1.47	1.73	1.82	1.84	1.86
L_2-PPI-f	0.9406	0.9418	0.9452	0.9418	0.9434	1.55	2.08	2.11	2.13	2.11
L_2-PPI-p	0.9424	0.9422	0.9452	0.9410	0.9452	1.64	2.40	2.38	2.40	2.36
L_1-PPI-f	0.9400	0.9426	0.9464	0.9430	0.9440	1.60	1.91	2.01	2.02	2.03
L_1-PPI-p	0.9398	0.9440	0.9466	0.9412	0.9454	1.72	2.04	2.16	2.17	2.18
SPI	0.9482	0.9474	0.9506	0.9518	0.9456	1.47	1.74	1.82	1.84	1.85
$T = 50$										
QPI-f	0.9060	0.9268	0.9266	0.9222	0.9302	1.43	1.71	1.80	1.83	1.84
QPI-p	0.9030	0.9286	0.9262	0.9206	0.9312	1.44	1.73	1.82	1.85	1.87
L_2-PPI-f	0.9300	0.9394	0.9350	0.9338	0.9408	1.55	3.37	3.34	3.23	3.11
L_2-PPI-p	0.9302	0.9414	0.9358	0.9372	0.9398	1.64	3.85	3.74	3.60	3.44
L_1-PPI-f	0.9302	0.9410	0.9358	0.9356	0.9412	1.61	2.39	2.58	2.54	2.50
L_1-PPI-p	0.9308	0.9422	0.9364	0.9384	0.9412	1.79	2.79	2.81	2.74	2.68
SPI	0.9486	0.9520	0.9486	0.9444	0.9518	1.47	1.74	1.81	1.84	1.85

Table 5: The CVR and LEN of PIs for Model 4

Model 4:	$X_t = 0.2 + \log(0.5 + \lvert X_{t-1}\rvert) + \epsilon_t$									
	CVR for each step					LEN for each step				
$T = 400$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.9498	0.9446	0.9482	0.9444	0.9444	3.89	4.30	4.33	4.33	4.33
QPI-p	0.9486	0.9462	0.9512	0.9450	0.9464	3.91	4.33	4.35	4.35	4.35
L_2-PPI-f	0.9492	0.9454	0.9480	0.9440	0.9466	3.90	4.32	4.34	4.34	4.34
L_2-PPI-p	0.9508	0.9442	0.9504	0.9458	0.9466	3.92	4.33	4.35	4.36	4.36
L_1-PPI-f	0.9496	0.9460	0.9486	0.9454	0.9468	3.90	4.32	4.34	4.34	4.34
L_1-PPI-p	0.9510	0.9452	0.9502	0.9462	0.9472	3.93	4.34	4.35	4.36	4.36
SPI	0.9502	0.9456	0.9492	0.9460	0.9504	3.90	4.32	4.34	4.34	4.34
$T = 100$										
QPI-f	0.9350	0.9440	0.9358	0.9412	0.9348	3.85	4.27	4.29	4.29	4.29
QPI-p	0.9412	0.9482	0.9404	0.9456	0.9412	3.93	4.34	4.37	4.37	4.37
L_2-PPI-f	0.9376	0.9442	0.9370	0.9438	0.9362	3.90	4.30	4.32	4.32	4.32
L_2-PPI-p	0.9412	0.9504	0.9406	0.9478	0.9426	3.98	4.38	4.39	4.40	4.40
L_1-PPI-f	0.9386	0.9446	0.9378	0.9448	0.9364	3.90	4.30	4.32	4.32	4.32
L_1-PPI-p	0.9412	0.9502	0.9404	0.9470	0.9418	3.98	4.38	4.40	4.40	4.41
SPI	0.9480	0.9502	0.9426	0.9504	0.9466	3.90	4.32	4.34	4.34	4.34
$T = 50$										
QPI-f	0.9280	0.9288	0.9328	0.9326	0.9312	3.75	4.24	4.27	4.26	4.27
QPI-p	0.9386	0.9396	0.9420	0.9422	0.9428	3.92	4.40	4.43	4.43	4.43
L_2-PPI-f	0.9404	0.9316	0.9372	0.9372	0.9352	3.89	4.30	4.32	4.32	4.33
L_2-PPI-p	0.9496	0.9398	0.9452	0.9448	0.9438	4.06	4.46	4.48	4.49	4.49
L_1-PPI-f	0.9410	0.9308	0.9376	0.9384	0.9350	3.90	4.30	4.33	4.33	4.33
L_1-PPI-p	0.9504	0.9398	0.9452	0.9458	0.9438	4.06	4.47	4.49	4.50	4.49
SPI	0.9530	0.9462	0.9456	0.9444	0.9428	3.90	4.32	4.34	4.34	4.34

Table 6: The CVR and LEN of PIs for Model 5

Model 5:	$X_t = 2 \cdot \log(X_{t-1}^2) + \epsilon_t$									
	CVR for each step					LEN for each step				
$T = 400$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.9432	0.9502	0.9476	0.9478	0.9510	3.90	4.32	4.42	4.44	4.45
QPI-p	0.9440	0.9524	0.9492	0.9480	0.9484	3.91	4.33	4.43	4.46	4.46
L_2-PPI-f	0.9468	0.9502	0.9516	0.9494	0.9518	3.90	4.34	4.43	4.46	4.47
L_2-PPI-p	0.9448	0.9536	0.9500	0.9478	0.9498	3.92	4.35	4.45	4.47	4.48
L_1-PPI-f	0.9448	0.9510	0.9506	0.9496	0.9504	3.91	4.34	4.44	4.46	4.47
L_1-PPI-p	0.9440	0.9532	0.9494	0.9480	0.9498	3.92	4.35	4.45	4.47	4.48
SPI	0.9462	0.9538	0.9510	0.9480	0.9472	3.90	4.33	4.42	4.45	4.45
$T = 100$										
QPI-f	0.9484	0.9406	0.9392	0.9452	0.9418	3.87	4.29	4.39	4.41	4.42
QPI-p	0.9498	0.9450	0.9406	0.9480	0.9450	3.92	4.34	4.44	4.46	4.46
L_2-PPI-f	0.9512	0.9436	0.9418	0.9488	0.9456	3.90	4.33	4.44	4.47	4.48
L_2-PPI-p	0.9526	0.9468	0.9436	0.9480	0.9470	3.94	4.38	4.48	4.52	4.53
L_1-PPI-f	0.9524	0.9440	0.9420	0.9476	0.9454	3.90	4.33	4.44	4.47	4.48
L_1-PPI-p	0.9530	0.9470	0.9438	0.9488	0.9476	3.94	4.38	4.48	4.52	4.53
SPI	0.9562	0.9514	0.9470	0.9512	0.9496	3.90	4.33	4.42	4.45	4.45
$T = 50$										
QPI-f	0.9246	0.9314	0.9326	0.9342	0.9376	3.79	4.27	4.36	4.38	4.39
QPI-p	0.9300	0.9390	0.9390	0.9394	0.9432	3.88	4.36	4.45	4.48	4.48
L_2-PPI-f	0.9336	0.9366	0.9398	0.9404	0.9452	3.89	4.34	4.46	4.50	4.51
L_2-PPI-p	0.9362	0.9418	0.9426	0.9444	0.9492	3.96	4.43	4.55	4.59	4.60
L_1-PPI-f	0.9332	0.9374	0.9392	0.9398	0.9444	3.89	4.35	4.46	4.50	4.51
L_1-PPI-p	0.9364	0.9426	0.9432	0.9436	0.9480	3.97	4.44	4.55	4.59	4.60
SPI	0.9508	0.9498	0.9480	0.9454	0.9516	3.90	4.33	4.43	4.45	4.45

Table 7: The CVR and LEN of PIs for Model 6

Model 6:	$X_t = \log(10 + 5 \cdot e^{0.9 \cdot X_{t-1}}) + \epsilon_t$									
	CVR for each step					LEN for each step				
$T = 400$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.9506	0.9452	0.9440	0.9420	0.9388	3.88	5.18	6.01	6.60	7.03
QPI-p	0.9528	0.9472	0.9454	0.9414	0.9378	3.90	5.22	6.05	6.64	7.07
L_2-PPI-f	0.9532	0.9488	0.9476	0.9434	0.9418	3.90	5.22	6.07	6.67	7.11
L_2-PPI-p	0.9506	0.9500	0.9478	0.9456	0.9412	3.92	5.26	6.12	6.71	7.16
L_1-PPI-f	0.9536	0.9488	0.9470	0.9442	0.9424	3.90	5.22	6.07	6.67	7.12
L_1-PPI-p	0.9514	0.9510	0.9482	0.9456	0.9424	3.93	5.26	6.12	6.72	7.17
SPI	0.9532	0.9508	0.9508	0.9490	0.9498	3.90	5.25	6.13	6.76	7.23
$T = 100$										
QPI-f	0.9350	0.9244	0.9216	0.9108	0.9038	3.83	5.02	5.77	6.28	6.67
QPI-p	0.9404	0.9302	0.9298	0.9176	0.9116	3.94	5.17	5.93	6.46	6.86
L_2-PPI-f	0.9422	0.9332	0.9330	0.9224	0.9136	3.94	5.23	6.04	6.62	7.06
L_2-PPI-p	0.9498	0.9388	0.9392	0.9324	0.9218	4.05	5.37	6.21	6.81	7.25
L_1-PPI-f	0.9424	0.9340	0.9340	0.9224	0.9146	3.95	5.23	6.04	6.63	7.07
L_1-PPI-p	0.9498	0.9400	0.9384	0.9318	0.9214	4.05	5.38	6.21	6.81	7.26
SPI	0.9498	0.9496	0.9504	0.9458	0.9494	3.90	5.25	6.13	6.76	7.23
$T = 50$										
QPI-f	0.9056	0.8930	0.8796	0.8640	0.8526	3.72	4.85	5.50	5.97	6.34
QPI-p	0.9200	0.9102	0.8934	0.8872	0.8716	3.93	5.13	5.83	6.33	6.71
L_2-PPI-f	0.9276	0.9172	0.9032	0.8984	0.8860	4.04	5.32	6.17	6.81	7.36
L_2-PPI-p	0.9412	0.9302	0.9188	0.9160	0.9042	4.27	5.62	6.50	7.16	7.72
L_1-PPI-f	0.9290	0.9170	0.9034	0.8982	0.8856	4.04	5.33	6.17	6.81	7.36
L_1-PPI-p	0.9412	0.9300	0.9192	0.9166	0.9034	4.27	5.62	6.50	7.16	7.72
SPI	0.9508	0.9460	0.9432	0.9484	0.9472	3.90	5.25	6.13	6.75	7.23

Table 8: The CVR and LEN of PIs for Model 7

Model 7:	$X_t = \log(4 \cdot e^{0.9 \cdot X_{t-2}} + 5 \cdot e^{0.9 \cdot X_{t-1}} + 6 \cdot e^{0.9 \cdot X_{t-3}}) + \epsilon_t$									
	CVR for each step					LEN for each step				
$T = 400$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.9450	0.9426	0.9442	0.9384	0.9376	3.88	4.02	4.28	4.70	4.87
QPI-p	0.9484	0.9444	0.9470	0.9410	0.9400	3.92	4.06	4.33	4.75	4.92
L_2-PPI-f	0.9472	0.9434	0.9476	0.9394	0.9406	3.91	4.05	4.31	4.75	4.93
L_2-PPI-p	0.9498	0.9460	0.9498	0.9442	0.9428	3.94	4.09	4.36	4.80	4.98
L_1-PPI-f	0.9462	0.9430	0.9474	0.9408	0.9406	3.91	4.05	4.31	4.75	4.93
L_1-PPI-p	0.9484	0.9466	0.9486	0.9444	0.9434	3.95	4.09	4.36	4.80	4.98
SPI	0.9462	0.9452	0.9480	0.9446	0.9432	3.90	4.04	4.31	4.76	4.94
$T = 100$										
QPI-f	0.9330	0.9404	0.9298	0.9238	0.9278	3.82	3.96	4.19	4.56	4.70
QPI-p	0.9436	0.9496	0.9440	0.9354	0.9390	3.99	4.14	4.38	4.76	4.91
L_2-PPI-f	0.9400	0.9464	0.9404	0.9332	0.9408	3.94	4.09	4.35	4.77	4.95
L_2-PPI-p	0.9498	0.9536	0.9504	0.9430	0.9508	4.10	4.26	4.53	4.98	5.16
L_1-PPI-f	0.9396	0.9466	0.9404	0.9312	0.9406	3.94	4.09	4.35	4.77	4.95
L_1-PPI-p	0.9504	0.9548	0.9508	0.9422	0.9504	4.10	4.26	4.54	4.98	5.16
SPI	0.9502	0.9542	0.9484	0.9468	0.9550	3.90	4.04	4.31	4.76	4.94
$T = 50$										
QPI-f	0.9152	0.9132	0.9202	0.9044	0.9020	3.71	3.89	4.12	4.46	4.60
QPI-p	0.9366	0.9402	0.9428	0.9340	0.9284	4.04	4.25	4.50	4.87	5.03
L_2-PPI-f	0.9344	0.9312	0.9366	0.9254	0.9236	3.97	4.13	4.42	4.88	5.08
L_2-PPI-p	0.9518	0.9506	0.9570	0.9460	0.9432	4.31	4.49	4.82	5.31	5.54
L_1-PPI-f	0.9340	0.9310	0.9360	0.9264	0.9248	3.97	4.13	4.42	4.88	5.08
L_1-PPI-p	0.9528	0.9494	0.9554	0.9450	0.9434	4.31	4.50	4.82	5.32	5.54
SPI	0.9442	0.9464	0.9520	0.9486	0.9508	3.90	4.05	4.31	4.76	4.94

5 Empirical data analyses

In this section, we deploy two real datasets to check the performance of our forward bootstrap methods. These empirical studies could verify the performance of our forward bootstrap prediction algorithm when the model is misspecified. If there is no strong evidence to support the choice of the underlying model in practice, we may apply the forward bootstrap prediction with the non-parametric estimator; see Politis and Wu (2023).

5.1 Flu data

We take the *flu* data from the *R* package *astsa*, which describes the monthly pneumonia and influenza deaths per 10000 people in the United States from 1968 to 1978. Due to the epidemic nature of the flu, the behavior of the series is quite different when the rates go above some threshold value than when it is below the value. Thus, a TAR model is a natural candidate to model the *flu* dataset. Since the slightly downward trend exists in the original series, we consider the first-order differencing and then focus on the prediction of the resulting series, denoted by $\{y_i\}_{i=1}^{131}$. This series is plotted below.

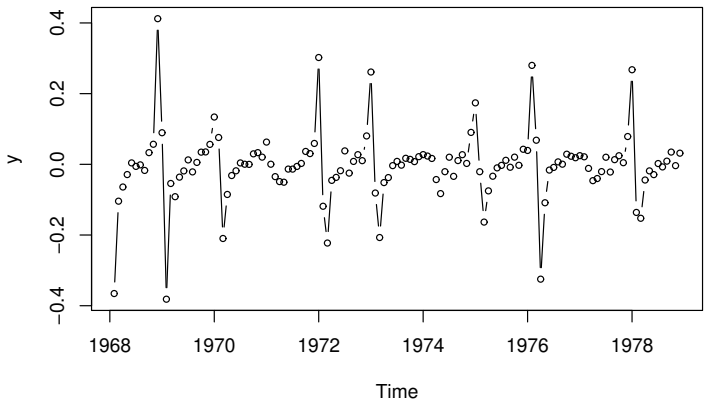

Figure 1: The first-order differencing series of the *flu* data

To perform TAR in practice, we apply the *tsDyn* package built in Stigler (2020). We consider a TAR(2) model in which the threshold value r is determined by the built-in function *setar* automatically. To perform multi-step ahead predictions, the *setar* function can integrate a bootstrap method described in the book of Franses and Van Dijk (2000) to return point prediction and prediction interval denoted by PI-tsDyn. This kind of PI is in a similar spirit to the QPI defined in this paper. As we expect, such PI will suffer from the undercoverage issue in the finite sample case. To compare this PI-tsDyn with our PPI comprehensively with 131 data points, we take a rolling-window pseudo-out-of-sample (rwPOOS) prediction experiment to measure the performance of these two PIs. In short, rwPOOS predictions procedure implies that we use $\{Y_1, \cdots, Y_W\}$ to predict Y_{W+h}; then use $\{Y_2, \cdots, Y_{W+1}\}$ to predict Y_{W+1+h} respectively, and so on till we exhaust all available data; here W is the size of the training window and h is the desired prediction horizon; see Wu and Karmakar (2023) for more description of this prediction setting.

With the *flu* data, to make sure we have enough prediction to compute the average performance, we take $W = 50$ and consider $h = 2, \ldots, 5$. We take the nominal confidence level to be 0.95. To simplify the presentation, we only consider L_2-PPI with fitted or predictive residuals. We take $K = 1000$ and $M = 200$ to build bootstrap PIs. The average empirical coverage rate and length of these two types of PIs are presented in Table 9. We can see the PPI works better than PI-tsDyn. For 1-step to 4-step ahead predictions, the empirical coverage rates

of PI-tsDyn are below 0.9 even though the nominal confidence level is 0.95. On the other hand, our PPIs can give a more accurate coverage rate with a slightly larger length.

Table 9: The CVR and LEN of different PIs on the *flu* dataset under nominal confidence level 0.95

Flu data:	TAR(2)							
	CVR for each step				LEN for each step			
$W = 50$	2	3	4	5	2	3	4	5
PI-tsDyn	0.883	0.896	0.883	0.935	0.252	0.357	0.410	0.458
L_2-PPI-f	0.922	0.935	0.935	0.974	0.399	0.453	0.493	0.492
L_2-PPI-p	0.961	0.974	0.987	0.987	0.544	0.680	0.699	0.695

5.2 Unemployment rates data

We also take the *UnempRate* data from *R* package *astsa*, which records the monthly U.S. unemployment rate from 1948 to 1979. As indicated by the work of Rothman (1998), the unemployment rate increases quickly in recessions but declines relatively slowly during expansions. They suggested a specific Exponential Autoregressive Model (EAR) shown below to model such an asymmetric business cycle:

$$Y_t = c + \exp(-Y_{t-1}^2) \cdot Y_{t-1} + Y_{t-2} + \epsilon_t.$$

After inducing the stationarity in the quarterly unemployment rate series by a so-called log-linear detrending method, they showed that the above EAR(2) model outperformed standard AR(2) and TAR models. To remove the trend in the original series, we took logarithms on the quarterly rate series first and then detrended the series by the *detrend* function in *R* with order to be 5. The final series $\{y_i\}_{i=1}^{124}$ are plotted below.

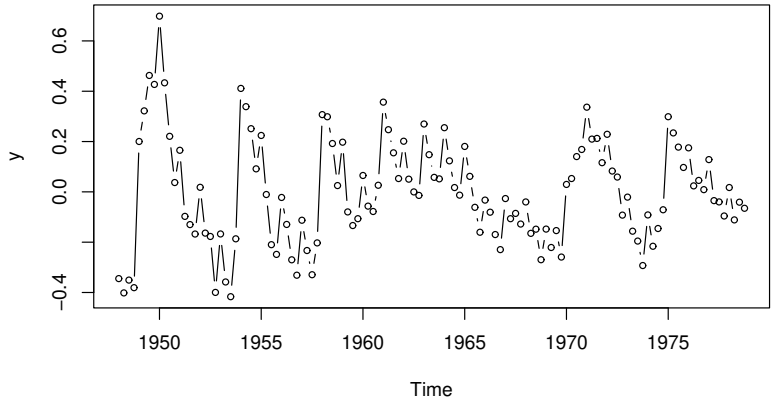

Figure 2: The log-linear detrended series of the quarterly *UnempRate* data

We still focus on data during this period and apply the above EAR(2) model to fit $\{y_i\}_{i=1}^{124}$. We still take the rwPOOS prediction procedure introduced in Section 5.1 to measure the performance of different PIs on this single real data set. As a comparison, we take the iterative regression prediction as the benchmark, and then the naive PI can be built under the assumption of normality of error by *lm* function. We denote the naive PI by PI-Naive. All the prediction settings are the same as the empirical study in Section 5.1. The average empirical

coverage rate and length of these two types of PIs are presented in Table 10. As we can find, the PI-Naive has a worse and worse coverage rate for longer prediction horizons. On the other hand, two PPIs can improve the coverage rate.

Table 10: The CVR and LEN of different PIs on the *UnempRate* dataset under nominal confidence level 0.95

UnempRate data:	EAR(2)							
	CVR for each step				LEN for each step			
$W = 50$	2	3	4	5	2	3	4	5
PI-Naive	0.957	0.929	0.914	0.886	0.580	0.577	0.576	0.576
L_2-PPI-f	0.971	0.943	0.971	0.957	0.658	0.701	0.710	0.718
L_2-PPI-p	0.971	0.957	0.943	0.971	0.690	0.715	0.727	0.736

6 Conclusions and discussions

In the paper at hand, we analyzed prediction inference for a specific form of NLAR model which possesses separate mean and variance functions. When we know the model and innovation information, we show that the simulation-based approach can return consistent predictions. When we only know the form of parametric NLAR models, the bootstrap-based prediction is also shown to be consistent with true optimal future values. Moreover, we can obtain asymptotically valid or pertinent prediction intervals. In addition, we show the possibility that our algorithm could serve for prediction tasks with general NLAR models. Furthermore, we propose the idea of combining predictive residuals with the bootstrap-based NLAR prediction. The simulation and empirical studies verify the superiority of our methods. Constructing pertinent prediction intervals with predictive residuals can improve the empirical CVR, especially for short data.

Notice that Wolf and Wunderli (2015) proposed a so-called Joint Prediction Region (JPR) to cover the whole future path with the desired probability $1 - \alpha$ and allow at most $k - 1$ number of true future predictions to fall outside the JPR. However, they omitted the methodology of constructing the prediction vector and argued that it could be generated by the estimated probability mechanism. Our contribution is solving the difficulty of multi-step ahead prediction of NLAR models. We can expect that the JPR can be combined with our forward bootstrap predictions. Then, a JPR centered at meaningful optimal L_1 or L_2 predictions can be built even for a general non-linear model. Moreover, the JPR also suffers from undercoverage for finite sample cases since there is no procedure to capture the estimation variability in JPR. Thus, our forward bootstrap prediction idea can be applied to JPR to get the pertinent property.

Acknowledgements

The authors are grateful to the Co-Editor and three anonymous referees for their constructive comments that have led to considerable improvements in the revision. The authors are thankful to Professor Yunyi Zhang for his valuable suggestions. The research of the first author was partially supported by the Richard Libby Graduate Research Award. The research of the second author was partially supported by NSF grant DMS 19-14556.

Data availability statement

The data used in the empirical studies is openly available at the R package *astsa*.

Supporting information

Supporting Information can be found online in the supplement section for this article.

References

Berg A, McMurry T, Politis DN. 2012. Testing time series linearity: traditional and bootstrap methods. *Handbook of Statistics* **30**: 27–42.

Bradley RC. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probability surveys* **2**: 107–144.

Chen R, Yang L, Hafner C. 2004. Nonparametric multistep-ahead prediction in time series analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**: 669–686.

Clements MP, Hendry DF. 1996. Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics* **58**: 657–684.

De Gooijer JG, Kumar K. 1992. Some recent developments in non-linear time series modelling, testing, and forecasting. *International Journal of Forecasting* **8**: 135–156.

Franke J, Kreiss JP, Mammen E, Neumann MH. 2002. Properties of the nonparametric autoregressive bootstrap. *Journal of Time Series Analysis* **23**: 555–585.

Franke J, Kreiss JP, Mammen E. 2002. Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli* **8**: 1–37.

Franke J, Neumann MH. 2000. Bootstrapping neural networks. *Neural Computation* **12**: 1929–1949.

Franses PH, Van Dijk D. 2000. *Non-linear time series models in empirical finance*. Cambridge university press.

Guo M, Bai Z, An HZ. 1999. Multi-step prediction for nonlinear autoregressive models based on empirical distributions. *Statistica Sinica* **9**: 559–570.

Jones DA, Cox DR. 1978. Nonlinear autoregressive processes. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **360**: 71–95.

Kreiss JP, Paparoditis E. 2023. *Bootstrap for Time Series: Theory and Methods*. Springer-Verlag, New York.

Lee K, Billings S. 2003. A new direct approach of computing multi-step ahead predictions for non-linear models. *International Journal of Control* **76**: 810–822.

Manzan S, Zerom D. 2008. A bootstrap-based non-parametric forecast density. *International Journal of Forecasting* **24**: 535–550.

Pan L, Politis DN. 2016. Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. *Journal of Statistical Planning and Inference* **177**: 1–27.

Pascual L, Romo J, Ruiz E. 2001. Effects of parameter estimation on prediction densities: a bootstrap approach. *International Journal of Forecasting* **17**: 83–103.

Pemberton J. 1987. Exact least squares multi-step prediction from nonlinear autoregressive models. *Journal of Time Series Analysis* **8**: 443–448.

Politis DN. 2009. Financial time series. *Wiley Interdisciplinary Reviews: Computational Statistics* **1**: 157–166.

— 2013. Model-free model-fitting and predictive distributions. *Test* **22**: 183–221.

— 2015. "Model-free prediction in regression." In: *Model-Free Prediction and Regression*. Springer: pp. 57–80.

Politis DN, Romano JP, Wolf M. 1997. Subsampling for heteroskedastic time series. *Journal of Econometrics* **81**: 281–317.

Politis DN, Wu K. 2023. Multi-Step-Ahead Prediction Intervals for Nonparametric Autoregressions via Bootstrap: Consistency, Debiasing, and Pertinence. *Stats* **6**: 839–867.

Rothman P. 1998. Forecasting asymmetric unemployment rates. *Review of Economics and Statistics* **80**: 164–168.

Stigler M. 2020. "Nonlinear time series in R: Threshold cointegration with tsDyn." In: *Handbook of Statistics*. Vol. 42. Elsevier: pp. 229–264.

Stockis JP, Franke J, Kamgaing JT. 2010. On geometric ergodicity of CHARME models. *Journal of Time Series Analysis* **31**: 141–152.

Tjøstheim D. 1994. Non-linear time series: a selective review. *Scandinavian Journal of Statistics* **21**: 97–130.

Wang Y, Politis DN. 2021. Model-free Bootstrap and Conformal Prediction in Regression: Conditionality, Conjecture Testing, and Pertinent Prediction Intervals. *arXiv preprint arXiv:2109.12156*.

Wolf M, Wunderli D. 2015. Bootstrap joint prediction regions. *Journal of Time Series Analysis* **36**: 352–376.

Wu K, Karmakar S. 2023. A model-free approach to do long-term volatility forecasting and its variants. *Financial Innovation* **9**: 1–38.

Zhang G, Patuwo BE, Hu MY. 1998. Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting* **14**: 35–62.