stats

MDPI

Article

Multi-Step-Ahead Prediction Intervals for Nonparametric Autoregressions via Bootstrap: Consistency, Debiasing, and Pertinence

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

[1] Department of Mathematics and Halicioğlu Data Science Institute, University of California, San Diego, CA 92093, USA
[2] Department of Mathematics, University of California, San Diego, CA 92093, USA; kwu@ucsd.edu
* Correspondence: dpolitis@ucsd.edu

Abstract: To address the difficult problem of the multi-step-ahead prediction of nonparametric autoregressions, we consider a forward bootstrap approach. Employing a local constant estimator, we can analyze a general type of nonparametric time-series model and show that the proposed point predictions are consistent with the true optimal predictor. We construct a quantile prediction interval that is asymptotically valid. Moreover, using a debiasing technique, we can asymptotically approximate the distribution of multi-step-ahead nonparametric estimation by the bootstrap. As a result, we can build bootstrap prediction intervals that are pertinent, i.e., can capture the model estimation variability, thus improving the standard quantile prediction intervals. Simulation studies are presented to illustrate the performance of our point predictions and pertinent prediction intervals for finite samples.

Keywords: bootstrap; non-linear time-series prediction; nonparametric estimation

check for updates

Citation: Politis, D.N.; Wu, K. Multi-Step-Ahead Prediction Intervals for Nonparametric Autoregressions via Bootstrap: Consistency, Debiasing, and Pertinence. *Stats* **2023**, *6*, 839–867. https://doi.org/10.3390/stats6030053

Academic Editor: Wei Zhu

Received: 19 July 2023
Revised: 5 August 2023
Accepted: 7 August 2023
Published: 11 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Since the 1980s, non-linear time-series models have attracted attention for modeling asymmetry in financial returns, the volatility of stock markets, switching regimes, etc. Compared to linear time-series models, non-linear models are more capable of depicting the underlying data-generating mechanism; see the review in [1], for example. However, unlike linear models, where the one-step-ahead predictor can be iterated, the multi-step-ahead prediction of non-linear models is cumbersome, since the innovation severely influences the forecasting value.

In this paper, by combining the forward bootstrap in [2] with nonparametric estimation, we develop multi-step-ahead (conditional) predictive inference for the general model:

$$X_t = m(X_{t-1}, \ldots, X_{t-p}) + \sigma(X_{t-1}, \ldots, X_{t-q})\epsilon_t; \tag{1}$$

where the ϵ_t values are assumed to be independent and identically distributed (i.i.d.) with mean 0 and variance 1, and $m(\cdot)$ and $\sigma(\cdot)$ are some functions that satisfy some smoothness conditions. We will also assume that the time series satisfying Equation (1) is geometrically ergodic and causal, i.e., that for any t, ϵ_t is independent of $\{X_s, s < t\}$.

In Equation (1), we have the trend/regression function $m(\cdot)$ depending on the last p data points, while the standard deviation/volatility function $\sigma(\cdot)$ depends on the last q data points; in many situations, p and q are taken to be equal for simplicity. Some special cases deserve mention: e.g., if $\sigma(X_{t-1}, \ldots, X_{t-q}) \equiv \sigma$ (constant), Equation (1) yields a non-linear/nonparametric autoregressive model with homoscedastic innovations. The well-known ARCH/GARCH models are a special case of Equation (1) with $m(X_{t-1}, \ldots, X_{t-p}) \equiv 0$.

Although the L_2-optimal one-step-ahead prediction of Equation (1) is trivial when we know the regression function $m(\cdot)$ or have a consistent estimator of it, the multi-step-ahead prediction is not easy to obtain. In addition, it is nontrivial to find the L_1-optimal prediction, even for one-step-ahead forecasting. In several applied areas, e.g., econometrics, climate modeling, and water resources management, data might not possess a finite second moment, in which case, optimizing L_2 loss is vacuous. 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 discussion in Ch. 10 of [2]. Later, we will show that our method is compatible with both L_2- and L_1-optimal multi-step-ahead predictions.

Efforts to overcome the difficulty of forecasting non-linear time series can be traced back to the work of [3], where a numerical approach was proposed to explore the exact conditional k-step-ahead L_2-optimal prediction of X_{T+k} for the homoscedastic Equation (1). However, this method is computationally intractable with long-horizon prediction and requires knowledge of the distribution of innovations and the regression function, which is not realistic in practice.

Consequently, practitioners started to investigate some suboptimal methods to perform multi-step-ahead prediction. Generally speaking, these methods take one of two avenues: (1) direct prediction or (2) iterative prediction. The first idea involves working with a different ("direct") model, specific to k-step-ahead prediction, namely:

$$X_t = m_k(X_{t-k}, \ldots, X_{t-k-p+1}) + \sigma_k(X_{t-k}, \ldots, X_{t-k-q+1})\xi_t. \tag{2}$$

Even though $m_k(\cdot)$ and $\sigma_k(\cdot)$ are unknown to us, we can construct nonparametric estimators, $\widehat{m}_k$ and $\widehat{\sigma}_k$, and plug them into Equation (2) to perform k-step-ahead prediction. Ref. [4] gives a review of this approach. However, as pointed out by [5], a drawback of this approach is that information from intermediate observations $\{X_t, \ldots, X_{t-k+1}\}$ is disregarded. Furthermore, if ϵ_t in Equation (1) is i.i.d., then ξ_t in Equation (2) cannot be i.i.d. In other words, a practitioner must employ the (estimated) dependence structure of ξ_t in Equation (2) in order to perform the prediction in an optimal fashion.

The second idea is "iterative prediction", which employs one-step-ahead predictors in a sequential way to perform a multi-step-ahead forecast. For example, consider a two-step-ahead prediction using Model Equation (1); first, note that the L_2-optimal predictor of X_{T+1} is $\hat{X}_{T+1} = m(X_T, \ldots, X_{T+1-p})$. The L_2-optimal predictor of $X_{T+2} = m(X_{T+1}, X_T, \ldots, X_{T+2-p})$, but since X_{T+1} is unknown, it is tempting to plug in $\hat{X}_{T+1}$ in its place. This plug-in idea can be extended to multi-step-ahead forecasts, but it does *not* lead to the L_2-optimal predictor, except in the special case where the function $m(\cdot)$ is linear, e.g., in the case of a linear autoregressive (LAR) model.

Remark 1. *Since neither of the above two approaches is satisfactory, we propose to approximate the distribution of the future value via a particular type of simulation when the model is known or, more generally, by the bootstrap. To describe this approach, we rewrite Equation* (1) *as*

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

where $\boldsymbol{X}_{t-1}$ *is a vector that represents* $\{X_{t-1}, \ldots, X_{t-max(p,q)}\}$*, and* $G(\cdot, \cdot)$ *is some appropriate function. Then, when the model and the innovation information are known to us, we can create a pseudo-value* X^*_{T+k}*. Taking a three-step-ahead prediction as an example, the pseudo-value* X^*_{T+3} *can be defined as follows:*

$$X^*_{T+3} = G(G(G(\boldsymbol{X_T}, \epsilon^*_{T+1}), \epsilon^*_{T+2}), \epsilon^*_{T+3}); \tag{3}$$

where $\{\epsilon^*_i\}_{i=T+1}^{T+3}$ *is simulated as i.i.d. from* F_ϵ*. Repeating this process to M pseudo-*X^*_{T+3}*, the* L_2*-optimal prediction of* X_{T+3} *can be estimated from the mean of* $\{X^{*(m)}_{T+3}\}_{m=1}^{M}$*. As already discussed, constructing the* L_1*-optimal predictor may also be required since sometimes* L_2 *loss is not well defined; in our simulation framework, we can construct the optimal* L_1 *prediction by taking the*

median of $\{X_{T+k}^{*(m)}\}_{m=1}^{M}$. *Moreover, we can even build a prediction interval (PI) to measure the forecasting accuracy based on the quantile values of simulated pseudo-values. The extension of this algorithm to longer-step-ahead prediction is illustrated in Section 2.*

Realistically, practitioners will not know F_ϵ, $m(\cdot)$, or $\sigma(\cdot)$. In this situation, the first step is to estimate these quantities and plug them into the above simulation, which then turns into a bootstrap method. The bootstrap idea was introduced by [6] to carry out statistical inference for independent data. After that, many variants of the bootstrap were developed to handle time-series data. Prominent examples include the sieve bootstrap and the block bootstrap in its many variations, e.g., the circular bootstrap of [7] and the stationary bootstrap of [8]; see [9] for a review. Once some model structure of the data is assumed, participators can rely on model-based bootstrap methods, e.g., the residual and/or wild bootstrap; see [10] for a book-length treatment. The bootstrap technique can also be applied to a recently popular model, namely, a neural network. In particular, ref. [11] applied the bootstrap for the estimation inference of neural networks' parameters, while [12] utilized the bootstrap to estimate the performance of neural networks.

In the spirit of the idea of the bootstrap, ref. [13] proposed a *backward bootstrap* trick to predict an $AR(p)$ model. The advantage of the backward method is that each bootstrap prediction is naturally conditional on the latest p observations, which coincide with the conditional prediction in the real world. However, this method cannot handle non-linear time series, whose backward representation may not exist. Later, ref. [14] proposed a strategy to generate forward bootstrap $AR(p)$ series. To resolve the conditional prediction issues, they fixed the last p bootstrap values to be the true observations and computed predictions iteratively in the bootstrap world starting from there. They then extended this procedure to forecast the GARCH model in [15].

Sharing a similar idea, ref. [16] defined the *forward bootstrap* to perform prediction, but they proposed a different PI format that empirically has better performance, according to the coverage rate (CVR) and the length (LEN), compared to the PI of [14]. Although ref. [16] covered the forecasting of a non-linear and/or nonparametric time-series model, only one-step-ahead prediction was considered. The case of the multi-step-ahead prediction of non-linear (but parametric) time-series models was recently addressed in [17]. In the paper at hand, we address the case of the multi-step-ahead prediction of nonparametric time-series models, as in Equation (1). Beyond discussing optimal L_1 and L_2 point predictions, we consider two types of PI—quantile PI (QPI) and pertinent PI (PPI). As already mentioned, the former can be approximated by taking the quantile values of the future value's distribution in the bootstrap world. The PPI requires a more complicated and computationally heavy procedure to be built, as it attempts to capture the variability in parameter estimation. This additional effort results in improved finite-sample coverage as compared to the QPI.

As in most nonparametric estimation problems, the issue of bias becomes important. We will show that debiasing on the inherent bias-type terms of local constant estimation is necessary to guarantee the pertinence of a PI when multi-step-ahead predictions are required. Although the QPI and PPI are asymptotically equivalent, the PPI renders a better CVR in finite-sample cases; see the formal definition of PPI in the work of [2,16]. Analogously to the successful construction of PIs in the work of [18], we can employ predictive—as opposed to fitted—residuals in the bootstrap process to further alleviate the finite-sample undercoverage of bootstrap PIs in practice. There are several other nonparametric approaches to carry out the prediction inference of future values; e.g., see the work of [5,19] for variants of kernel-based methods; see the work of [20–22] for prediction with a neural network using the sieve bootstrap or various ensemble strategies; finally, see the work of [23–25] for a novel transformation-based approach for model-free prediction. The comparison of these various nonparametric techniques could be an independent study.

This paper is organized as follows. In Section 2, forward bootstrap prediction algorithms with local constant estimators will be given. The asymptotic properties of point predictions and PIs will be discussed in Section 3. Simulations are given in Section 4 to substantiate the finite-sample performance of our methods. Conclusions are given in Section 5. All proofs can be found in Appendix A. Discussions on the debiasing and pertinence related to building PIs are presented in Appendices B–D.

2. Nonparametric Forward Bootstrap Prediction

As discussed in the remark in Section 1, we can apply the *simulation* or *bootstrap* technique to approximate the distribution of future values. In general, this idea works for any geometrically ergodic autoregressive model, regardless of whether it is in a linear or non-linear format. For example, if we have a known general model $X_t = G(\boldsymbol{X}_{t-1}, \epsilon_t)$ at hand, we can perform k-step-ahead predictions according to the same logic of the three-step-ahead prediction example in Section 1.

To elaborate, we need to simulate $\{\epsilon_i^*\}_{i=T+1}^{T+k}$ as i.i.d. from F_ϵ and then compute the pseudo-value X_{T+k}^* iteratively with simulated innovations as follows:

$$X_{T+k}^* = G(\cdots G(G(G(\boldsymbol{X_T}, \epsilon_{T+1}^*), \epsilon_{T+2}^*), \epsilon_{T+3}^*), \ldots, \epsilon_{T+k}^*). \tag{4}$$

Repeating this procedure M times, we can make a prediction inference with the empirical distribution of $\{X_{T+k}^{*(m)}\}_{m=1}^M$. Similarly, if the model and innovation distribution are unknown to us, we can perform the estimation first to obtain $\widehat{G}(\cdot,\cdot)$ and $\widehat{F}_\epsilon$. Then, the above simulation-based algorithm turns out to be a bootstrap-based algorithm. More specifically, we bootstrap $\{\hat{\epsilon}_i^*\}_{i=T+1}^{T+k}$ from $\widehat{F}_\epsilon$ and calculate the pseudo-value $\widehat{X}_{T+k}^*$ iteratively with $\widehat{G}(\cdot,\cdot)$. The prediction inference can also be conducted with the empirical distribution of $\{\widehat{X}_{T+k}^{*(m)}\}_{m=1}^M$.

This simulation/bootstrap idea was recently implemented by [17] in the case where the model G is either known or parametrically specified. In what follows, we will focus on the case of the nonparametric model in Equation (1) and will analyze the asymptotic properties of the point predictor and prediction interval. For the sake of simplicity, we consider only the case in which $p = q = 1$; the general case can be handled similarly, but the notation is much more cumbersome. Assume that we observe $T+1$ data points and that we denote them by $\{X_0, \ldots, X_T\}$; our goal is the prediction inference of X_{T+k} for some $k \geq 1$. If we know $m(\cdot)$, $\sigma(\cdot)$, and F_ϵ, we can take a simulation approach to develop the prediction inference, as we explained in Section 1. When $m(\cdot)$, $\sigma(\cdot)$, and F_ϵ are unknown, we start by estimating $m(\cdot)$ and $\sigma(\cdot)$; we then estimate F_ϵ based on the empirical distribution of residuals. Subsequently, we can deploy a bootstrap-based method to approximate the distribution of future values. Several algorithms are given for this purpose later in the paper.

2.1. Bootstrap Algorithm for Point Prediction and QPI

For concreteness, we focus on local constant estimators, i.e., kernel-smoothed estimators of the Nadaraya–Watson type; other estimators can be applied similarly. The local constant estimators of $m(\cdot)$ and $\sigma(\cdot)$ are, respectively, defined as:

$$\widetilde{m}_h(x) = \frac{\sum_{t=1}^T K(\frac{x-X_{t-1}}{h}) X_t}{\sum_{t=1}^T K(\frac{x-X_{t-1}}{h})} \text{ and } \widetilde{\sigma}_h(x) = \frac{\sum_{t=1}^T K(\frac{x-X_{t-1}}{h})(X_t - \widetilde{m}_h(X_{t-1}))^2}{\sum_{t=1}^T K(\frac{x-X_{t-1}}{h})}; \tag{5}$$

where K is a non-negative kernel function that satisfies some regularity assumptions; see Section 3 for details. We use h to represent the bandwidth of kernel functions, but h may take a different value for mean and variance estimators. Due to theoretical and practical issues, we need to truncate the above local constant estimators as follows:

$$\widehat{m}_h(x) = \begin{cases} -C_m & \text{if } \widetilde{m}_h(x) < -C_m \\ \widetilde{m}_h(x) & \text{if } |\widetilde{m}_h(x)| \le C_m \\ C_m & \text{if } \widetilde{m}_h(x) > C_m \end{cases}; \widehat{\sigma}(x) = \begin{cases} c_\sigma & \text{if } \widetilde{\sigma}_h(x) < c_\sigma \\ \widetilde{\sigma}_h(x) & \text{if } c_\sigma \le \widetilde{\sigma}_h(x) \le C_\sigma \\ C_\sigma & \text{if } \widetilde{\sigma}_h(x) > C_\sigma \end{cases}; \tag{6}$$

where C_m and C_σ are large enough, and c_σ is small enough.

Using $\widehat{m}_h(\cdot)$ and $\widehat{\sigma}_h(\cdot)$ on Equation (1), we can obtain the fitted residuals $\{\hat{\epsilon}_t\}_{t=1}^T$, which are defined as:

$$\hat{\epsilon}_t = \frac{X_t - \widehat{m}_h(X_{t-1})}{\widehat{\sigma}_h(X_{t-1})}, \text{ for } t = 1, \ldots, T. \tag{7}$$

Later, in Section 3, we will show that the innovation distribution F_ϵ can be consistently estimated from the centered empirical distribution of $\{\hat{\epsilon}_t\}_{t=1}^T$, i.e., $\widehat{F}_\epsilon$, under some standard assumptions. We now have all the ingredients to perform the bootstrap-based Algorithm 1 to yield the point prediction and QPI of X_{T+k}.

Algorithm 1 Bootstrap prediction of X_{T+k} with fitted residuals

Step 1	With data $\{X_0, \ldots, X_T\}$, construct the estimators $\widehat{m}_h(x)$ and $\widehat{\sigma}_h(x)$ with Equation (6).
Step 2	Compute fitted residuals based on Equation (7), and let $\bar{\epsilon} = \frac{1}{T}\sum_{i=1}^T \hat{\epsilon}_i$. Let $\widehat{F}_\epsilon$ denote the empirical distribution of the centered residuals $\hat{\epsilon}_t - \bar{\epsilon}$ for $t = 1, \ldots, T$.
Step 3	Generate $\{\hat{\epsilon}_i^*\}_{i=T+1}^{T+k}$ i.i.d. from $\widehat{F}_\epsilon$. Then, construct bootstrap pseudo-values X_{T+1}^*, $\cdots$, X_{T+k}^* iteratively, i.e., $$X_{T+i}^* = \widehat{m}_h(X_{T+i-1}^*) + \widehat{\sigma}_h(X_{T+i-1}^*)\hat{\epsilon}_{T+i}^*, \text{ for } i = 1, \ldots, k. \tag{8}$$ For example, $X_{T+1}^* = \widehat{m}_h(X_T^*) + \widehat{\sigma}_h(X_T^*)\hat{\epsilon}_{T+1}^*$, and $X_{T+2}^* = \widehat{m}_h(\widehat{m}_h(X_T) + \sigma_h(X_T)\hat{\epsilon}_{T+1}^*) + \widehat{\sigma}_h(\widehat{m}(X_T) + \widehat{\sigma}_h(X_T)\hat{\epsilon}_{T+1}^*)\hat{\epsilon}_{T+2}^*$.
Step 4	Repeating Step 3 M times, we obtain pseudo-value replicates of X_{T+k}^* that we denote by $\{X_{T+k}^{(1)}, \ldots, X_{T+k}^{(M)}\}$. Then, L_2- and L_1-optimal predictors can be approximated by $\frac{1}{M}\sum_{i=1}^M X_{T+k}^{(i)}$ and the median of $\{X_{T+k}^{(1)}, \ldots, X_{T+k}^{(M)}\}$, respectively. Furthermore, a $(1-\alpha)100\%$ QPI can be built as (L, U), where L and U denote the $\alpha/2$ and $1-\alpha/2$ sample quantiles of M values $\{X_{T+k}^{(1)}, \ldots, X_{T+k}^{(M)}\}$.

Remark 2. *To construct the QPI of Algorithm 1, we can employ the optimal bandwidth rate, i.e., $h = O(T^{-1/5})$. However, in practice with small sample size, the QPI has a better empirical CVR for multi-step-ahead predictions by adopting an under-smoothing bandwidth; see Appendix B for a related discussion, and see Section 4 for simulation comparisons between applying optimal and under-smoothing bandwidths to the QPI.*

In the next section, we will show the conditional asymptotic consistency of our optimal point predictions and the QPI. In particular, we will verify that our point predictions converge to oracle optimal point predictors in probability—conditional on X_T. In addition, we will look for an asymptotically valid PI with a $(1-\alpha)100\%$ CVR to measure the prediction accuracy conditional on the latest observed data, which is defined as:

$$\mathbb{P}(L \le X_{T+k} \le U) \to 1-\alpha, \text{ as } T \to \infty, \tag{9}$$

where L and U are lower and higher PI bounds, respectively. Although not explicitly denoted, the probability $\mathbb{P}$ should be understood as the conditional probability given X_T. Later, based on a sequence of sets that contains the observed sample with a probability tending to 1, we will show how to build a prediction interval that is asymptotically valid by the bootstrap technique, even if the model information is unknown.

Although asymptotically correct, in finite samples, the QPI typically suffers from undercoverage; see the discussion in [2,16]. To improve the CVR in practice, we consider taking the predictive residuals to boost the bootstrap process. To derive such predictive residuals, we need to estimate the model based on the delete-X_t dataset, i.e., the available data for the scatter plot of X_i vs. $\{X_{i-1}\}$ for $i = 1, \ldots, t-1, t+1, \ldots, T$, i.e., excluding the single point at $i = t$. More specifically, we define the delete-X_t local constant estimators as:

$$\widetilde{m}_h^t(x) = \frac{\sum_{i=1, i\neq t}^{T} K(\frac{|x-X_{i-1}|}{h}) X_i}{\sum_{i=1, i\neq t}^{T} K(\frac{|x-X_{i-1}|}{h})} \text{ and } \widetilde{\sigma}_h^t(x) = \frac{\sum_{i=1, i\neq t}^{T} K(\frac{|x-X_{i-1}|}{h}) (X_i - \widetilde{m}_h^t(X_{i-1}))^2}{\sum_{i=1, i\neq t}^{T} K(\frac{|x-X_{i-1}|}{h})}. \tag{10}$$

Similarly, the truncated delete-X_t local estimators $\widehat{m}^t(x)$ and $\widehat{\sigma}_h^t(x)$ can be defined according to Equation (6). We now construct the so-called predictive residuals as:

$$\hat{\epsilon}_t^p = \frac{X_t - \widehat{m}_h^t(X_{t-1})}{\widehat{\sigma}_h^t(X_{t-1})}, \text{ for } t = 1, \ldots, T. \tag{11}$$

The k-step-ahead prediction of X_{T+k} with predictive residuals is depicted in Algorithm 2. Although Algorithms 1 and 2 are asymptotically equivalent, Algorithm 2 gives a QPI with a better CVR for finite samples; see the simulation comparisons of these two approaches in Section 4.

Algorithm 2 Bootstrap prediction of X_{T+k} with predictive residuals

Step 1	The same as Step 1 of Algorithm 1.
Step 2	Compute predictive residuals based on Equation (11). Let $\widehat{F}_\epsilon^p$ denote the empirical distribution of the centered predictive residuals $\hat{\epsilon}_t^p - \frac{1}{T}\sum_{i=1}^{T} \hat{\epsilon}_i^p, t = 1, \ldots, T$.
Steps 3–4	Replace $\widehat{F}_\epsilon$ by $\widehat{F}_\epsilon^p$ in Algorithm 1. All the rest are the same.

2.2. Bootstrap Algorithm for PPI

To improve the CVR of a PI, we can try to take the variability in the model estimation into account when we build the PI; i.e., we need to mimic the estimation process in the bootstrap world. Employing this idea results in a pertinent PI (PPI), as discussed in Section 1; see also [26].

Algorithm 3 outlines the procedure to build a PPI. Although this algorithm is more computationally heavy, the advantage is that the PPI gives a better CVR compared to the QPI in practice, i.e., with finite samples; see the examples in Section 4.

Remark 3 (Bandwidth choices)**.** *In Step 3 (b) of Algorithm 3, we can use an optimal bandwidth h and an over-smoothing bandwidth g to generate bootstrap time series so that we can capture the asymptotically non-random bias-type term of nonparametric estimation by the forward bootstrap; see the application in [27]. We can also apply an under-smoothing bandwidth h (and then use $g = h$) to render the bias term negligible. It turns out that both approaches work well for one-step-ahead prediction, although applying the over-smoothing bandwidth may be slightly better. However, taking under-smoothing bandwidth(s) is notably better for multi-step-ahead prediction. The reason for this is that the bias term cannot be captured appropriately for multi-step-ahead estimation with an over-smoothing bandwidth. On the other hand, with an under-smoothing bandwidth, the bias term is negligible; see Section 3.2 for further discussion; also, see [28] for a related discussion. The simulation studies in Appendix C explore the differences between these two bandwidth strategies.*

Algorithm 3 Bootstrap PPI of X_{T+k} with fitted residuals

Step 1	With data $\{X_0, \ldots, X_T\}$, construct the estimators $\widehat{m}_h(x)$ and $\widehat{\sigma}_h(x)$ by using Equation (6). Furthermore, compute fitted residuals based on Equation (7). Denote the empirical distribution of centered residuals by $\hat{\epsilon}_t - \frac{1}{T}\sum_{i=1}^{T}\hat{\epsilon}_i, t = 1, \ldots, T$ by $\widehat{F}_\epsilon$.
Step 2	Construct the L_1 or L_2 prediction $\widehat{X}_{T+k}$ using Algorithm 1.
Step 3	(a) Resample (with replacement) the residuals from $\widehat{F}_\epsilon$ to create pseudo-errors $\{\hat{\epsilon}_i^*\}_{i=1}^{T}$ and $\{\hat{\epsilon}_i^*\}_{i=T+1}^{T+k}$. (b) Let $X_0^* = X_I$, where I is generated as a discrete random variable uniformly distributed on the values $0, \ldots, T$. Then, create bootstrap pseudo-data $\{X_t^*\}_{t=1}^{T}$ in a recursive manner from the formula $$X_i^* = \widehat{m}_g(X_{i-1}^*) + \widehat{\sigma}_g(X_{i-1}^*)\hat{\epsilon}_i^*, \text{ for } i = 1, \ldots, T. \quad (12)$$ (c) Based on the bootstrap data $\{X_t^*\}_{t=0}^{T}$, re-estimate the regression and variance functions according to Equation (6) and obtain $\widehat{m}_h^*(x)$ and $\widehat{\sigma}_h^*(x)$; we use the same bandwidth h as the original estimator $\widehat{m}_h(x)$. (d) Guided by the idea of the forward bootstrap, re-define the latest value of X_T^* to match the original, i.e., re-define $X_T^* = X_T$. (e) With the estimators $\widehat{m}_g(x)$ and $\widehat{\sigma}_g(x)$, the bootstrap data $\{X_t^*\}_{t=0}^{T}$, and the pseudo-errors $\{\hat{\epsilon}_t^*\}_{t=T+1}^{T+k}$, use Equation (12) to recursively generate the future bootstrap data $X_{T+1}^*, \ldots, X_{T+k}^*$. (f) With bootstrap data $\{X_t^*\}_{t=0}^{T}$ and the estimators $\widehat{m}_h^*(x)$ and $\widehat{\sigma}_h^*(x)$, utilize Algorithm 1 to compute the optimal bootstrap prediction, which is denoted by $\widehat{X}_{T+h}^*$; to generate bootstrap innovations, we still use $\widehat{F}_\epsilon$. (g) Determine the bootstrap predictive root: $X_{T+k}^* - \widehat{X}_{T+k}^*$.
Step 4	Repeat Step 3 B times; the B bootstrap root replicates are collected in the form of an empirical distribution whose β-quantile is denoted by $q(\beta)$. The $(1-\alpha)100\%$ equal-tailed prediction interval for X_{T+k} centered at $\widehat{X}_{T+k}$ is then estimated by $[\widehat{X}_{T+k} + q(\alpha/2), \widehat{X}_{T+k} + q(1-\alpha/2)]$.

As Algorithm 2 is a version of Algorithm 1 using predictive (as opposed to fitted) residuals, we now propose Algorithm 4, which constructs a PPI with predictive residuals.

Algorithm 4 Bootstrap PPI of X_{T+k} with predictive residuals

Step 1	With data $\{X_0, \ldots, X_T\}$, construct the estimators $\widehat{m}_h(x)$ and $\widehat{\sigma}_h(x)$ by using Equation (6). Furthermore, compute predictive residuals based on Equation (11). Denote the empirical distribution of centered residuals $\hat{\epsilon}_t^p - \frac{1}{T}\sum_{i=1}^{T}\hat{\epsilon}_i^p, t = 1, \ldots, T$ by $\widehat{F}_\epsilon^p$.
Steps 2–4	The same as in Algorithm 3, but change the residual distribution from $\widehat{F}_\epsilon$ to $\widehat{F}_\epsilon^p$, and change the application of Algorithm 1 to Algorithm 2.

3. Asymptotic Properties

In this section, we provide the theoretical substantiation of our nonparametric bootstrap prediction methods—Algorithms 1–4. We start by analyzing optimal point predictions and the QPI based on Algorithms 1 and 2.

Remark 4. *Since the effect of leaving out one data pair X_t vs. $\{X_{t-1}\}$ is asymptotically negligible for large T, the delete-X_t estimators $\widehat{m}^t(x)$ and $\widehat{\sigma}^t(x)$ are asymptotically equal to $\widehat{m}(x)$ and $\widehat{\sigma}(x)$, respectively. Then, the predictive residual $\hat{\epsilon}_t^p$ is asymptotically the same as the fitted residual $\hat{\epsilon}_t$; see Lemma 5.5 of [16] for a formal comparison of these two types of estimators and residuals. Thus, we just give theorems to guarantee the asymptotic properties of point predictions and PIs with fitted residuals. The asymptotic properties for variants with predictive residuals also hold true.*

3.1. On Point Prediction and QPI

First, to conduct statistical inference for time series, we need to quantify the degree of the asymptotic dependence of the time series. In this paper, we consider that the time series is geometrically ergodic, which is equivalent to the β-mixing condition with an exponentially fast mixing rate; see [29] for a detailed introduction to different mixing conditions and ergodicity. To simplify the proof, we make the following assumptions:

A1 $|m(x)| + \sigma(x)\mathbb{E}|\epsilon_1| \leq c_1 + c_2|x|$ for all $x \in \mathbb{R}$ and some $c_1 < \infty, c_2 < 1$;
A2 $\sigma(x) \geq c_3 > 0$ for all $x \in \mathbb{R}$ and some $c_3 > 0$;
A3 $f_\epsilon(x)$ is positive everywhere.

A1–A3 can guarantee that the time-series process is geometrically ergodic; see Theorem 1 of [30] for proof and see the work of [31] for a discussion on the sufficient conditions of higher-order time series.

Since we need to build consistent properties of the nonparametric estimation, we further assume that:

A4 The regression function $m(x)$ is twice continuously differentiable with bounded derivatives, and we denote its Lipschitz continuous constant as L_m;
A5 The volatility function $\sigma(x)$ is twice continuously differentiable with bounded derivatives, and we denote its Lipschitz continuous constant as L_σ. Moreover, for all $M < \infty$, there is $c_M < \infty$ with $\mathbb{E}|\sigma(X_0)\epsilon_1|^M \leq c_M$, where X_0 is the initial point of the time series;
A6 For L_m and L_σ, $L_m + L_\sigma\mathbb{E}|\epsilon_1| < 1$;
A7 For the innovation distribution, f_ϵ is twice continuously differentiable; f_ϵ, $f_\epsilon^{'}$, and $f_\epsilon^{''}$ are bounded; and $\sup_{x\in\mathbb{R}} |xf_\epsilon^{'}(x)| < \infty$;
A8 The kernel function $K(x)$ is a compactly supported and symmetric probability density on $\mathbb{R}$ and has a bounded derivative.

Remark 5. *Assumption A6 is originally used to show that the expected value of X_t^* is $O_p(1)$ in the bootstrap world for all t. In practice, this assumption is not strict; see examples in Section 4. For assumption A8, we can apply a kernel with a support on the whole real line as long as the part outside a large enough compact set is asymptotically negligible.*

Under A1–A8, [27] shows that truncated the local constant estimators in Equation (6) are uniformly consistent with the true functions in an expanding region. We summarize this result in the lemma below:

Lemma 1. *Under A1–A8 and observed data $\{X_0, \ldots, X_T\}$, for local constant estimation as in Equation (6), we have:*

$$\sup_{|x|\leq c_T} |\widehat{m}_h(x) - m(x)| \xrightarrow{p} 0 \text{ and } \sup_{|x|\leq c_T} |\widehat{\sigma}_h(x) - \sigma(x)| \xrightarrow{p} 0. \tag{13}$$

where c_T is an appropriate sequence that converges to infinity as $T \to \infty$.

In addition, for the centered empirical distribution of $\hat{\epsilon}$, we can derive Lemma 2 to describe its consistency property.

Lemma 2. *Under A1–A8 and observed data $\{X_0, \ldots, X_T\}$, for the centered empirical distribution $\widehat{F}_\epsilon$, we have:*

$$\sup_{x\in\mathbb{R}} |\widehat{F}_\epsilon(x) - F_\epsilon(x)| \xrightarrow{p} 0. \tag{14}$$

See Theorem 5 of [27] for the proof of Lemmas 1 and 2. Combining all the pieces, we present Theorem 1 to show that the optimal point prediction and QPI returned by

Algorithm 1 or Algorithm 2 are consistent and asymptotically valid, respectively, conditionally on the latest observations.

Theorem 1. *Under assumptions A1–A8 and observed data* $\{X_0, \ldots, X_T\}$*, we have:*

$$\sup_{|x| \leq c_T} \left| F_{X^*_{T+k}|X_T,\ldots,X_0}(x) - F_{X_{T+k}|X_T}(x) \right| \xrightarrow{p} 0, \text{ for } k \geq 1, \tag{15}$$

where X^*_{T+k} *is a future value in the bootstrap world that can be determined iteratively by applying the expression* $X^*_{T+i} = \widehat{m}_h(X^*_{T+i-1}) + \widehat{\sigma}_h(X^*_{T+i-1})\hat{\epsilon}^*_{T+i}$ *for* $i = 1, \ldots k$*;* $\{\hat{\epsilon}^*_{T+i}\}_{i=1}^k$ *is i.i.d., with its distribution given by the empirical distribution of fitted (or predictive) residuals;* $F_{X^*_{T+k}|X_T,\ldots,X_0}(x)$ *represents the distribution* $\mathbb{P}^*(X^*_{T+k} \leq x|X_T, \ldots, X_0)$*, and here, we take* $\mathbb{P}^*$ *to represent the probability measure conditional on the sample of data; and* $F_{X_{T+k}|X_T}(x)$ *represents the (conditional) distribution of* X_{T+k} *in the real world, i.e.,* $\mathbb{P}(X_{T+k} \leq x|X_T)$*.*

3.2. On PPI with Homoscedastic Errors

With more complicated prediction procedures, such as Algorithms 3 and 4, we expect to find a more accurate PI, i.e., a PPI. The superiority of such PIs is that the estimation variability can be captured when we use the distribution of the predictive root in the bootstrap world to approximate its variant in the real world. We consider models with homoscedastic errors throughout this section; the model with heteroscedastic errors will be analyzed later.

Firstly, let us consider the one-step-ahead predictive root centered at the optimal L_2 point prediction in the real and bootstrap worlds, as given below:

$$\begin{aligned} X_{T+1} - \widehat{X}_{T+1} &= m(X_T) + \epsilon_{T+1} - \frac{1}{M}\sum_{i=1}^{M}(\widehat{m}_h(X_T) + \hat{\epsilon}_{i,T+1}); \\ X^*_{T+1} - \widehat{X}^*_{T+1} &= \widehat{m}_g(X_T) + \hat{\epsilon}_{T+1} - \frac{1}{M}\sum_{i=1}^{M}\left(\widehat{m}^*_h(X_T) + \hat{\epsilon}^*_{i,T+1}\right), \end{aligned} \tag{16}$$

where M is the number of bootstrap replications that we employ to approximate the optimal L_2 point prediction. Since we have centered the residuals to a mean of zero, Equation (16) degenerates to the following simple form asymptotically as $M \to \infty$:

$$\begin{aligned} X_{T+1} - \widehat{X}_{T+1} &= m(X_T) + \epsilon_{T+1} - \widehat{m}_h(X_T); \\ X^*_{T+1} - \widehat{X}^*_{T+1} &= \widehat{m}_g(X_T) + \hat{\epsilon}_{T+1} - \widehat{m}^*_h(X_T). \end{aligned} \tag{17}$$

To acquire a pertinent PI according to Definition 2.4 of [16], in addition to Equation (14), we also need asymptotically valid confidence intervals for local constant estimation in the bootstrap world; i.e., we should be able to estimate the distribution of the nonparametric estimator in the bootstrap world. For one-step-ahead prediction, this condition can be formulated as follows:

$$\sup_x |\mathbb{P}(a_T A_m \leq x) - \mathbb{P}^*(a_T A^*_m \leq x)| \xrightarrow{p} 0, \tag{18}$$

where

$$A_m = m(X_T) - \widehat{m}_h(X_T)\ ;\ A^*_m = \widehat{m}_g(X_T) - \widehat{m}^*_h(X_T), \tag{19}$$

and a_T is an appropriate sequence such that $\mathbb{P}(a_T A_m \leq x)$ has a nontrivial limit as $T \to \infty$. In [16], it was assumed that the nontrivial limit of $\mathbb{P}(a_T A_m \leq x)$ is continuous. In this case, the uniform convergence in Equation (18) follows from the pointwise convergence of all x.

Remark 6. *As we have discussed in Remark 3, the bootstrap procedure cannot capture the bias term of nonparametric estimation exactly unless delicate manipulations are made. Ref. [16] adopts two strategies to solve this issue: (B1) let $g = h$, and take a bandwidth rate satisfying $hT^{1/5} \to 0$, i.e., under-smoothing in function estimation; (B2) use the optimal smoothing rate with h proportional to $T^{-1/5}$, but generate time series in the bootstrap world with over-smoothing estimators, i.e., $g \neq h$ and $g/h \to \infty$. No matter which approach we take, Equation (18) can be shown; see details from Theorem 1 of [27] and Theorem 5.4 of [16].*

The following corollary is immediate:

Corollary 1. *Under assumptions A1—A3 and observed data $\{X_0, \ldots, X_T\}$, the one-step-ahead PI returned by Algorithms 3 and 4 with fitted or predictive residuals is asymptotically pertinent, respectively.*

However, for multi-step-ahead predictions, the analysis becomes more complicated, and the under-smoothing strategy turns out to work better. For example, considering the two-step-ahead prediction, the two predictive roots can be written as follows:

$$\begin{aligned} X_{T+2} - \widehat{X}_{T+2} &= m(X_{T+1}) + \epsilon_{T+2} - \frac{1}{M}\sum_{i=1}^{M}(\widehat{m}_h(\widehat{m}_h(X_T) + \hat{\epsilon}_{i,T+1}) + \hat{\epsilon}_{i,T+2}) \\ &\approx m(m(X_T) + \epsilon_{T+1}) + \epsilon_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\widehat{m}_h(\widehat{m}_h(X_T) + \hat{\epsilon}_{i,T+1}). \end{aligned} \tag{20}$$

Correspondingly, the predictive root in the bootstrap world is:

$$\begin{aligned} X^*_{T+2} - \widehat{X}^*_{T+2} &= \widehat{m}_g(X^*_{T+1}) + \hat{\epsilon}^*_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\left(\widehat{m}^*_h\left(\widehat{m}^*_h(X_T) + \hat{\epsilon}^*_{i,T+1}\right) + \hat{\epsilon}^*_{i,T+2}\right) \\ &\approx \widehat{m}_g(\widehat{m}_g(X_T) + \hat{\epsilon}^*_{T+1}) + \hat{\epsilon}^*_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\widehat{m}^*_h\left(\widehat{m}^*_h(X_T) + \hat{\epsilon}^*_{i,T+1}\right), \end{aligned} \tag{21}$$

where the approximated equality is due to the application of the LLN on the sample mean of the centered residuals.

Remark 7. *We should note that the over-smoothing approach may work better for finite samples. The reason is that applying the optimal bandwidth rate is superior when the bias-type term of the nonparametric estimation can be captured by the bootstrap. However, we will soon show that applying an under-smoothing bandwidth strategy is more accurate for multi-step-ahead predictions since it can solve the bias issue and render a PPI. Thus, in practice, we recommend adopting strategy (B2) to perform one-step-ahead predictions and adopting strategy (B1) to perform multi-step-ahead predictions. For a time series with heteroscedastic errors, the optimal bandwidth strategy is slightly different; see Section 3.3 for reference.*

Based on Equations (20) and (21), as we prove that the future distribution of X^*_{T+k} converges uniformly to the future distribution of X_{T+k} in probability, we can show that the distribution of the predictive root $X^*_{T+2} - \widehat{X}^*_{T+2}$ in the bootstrap world also converges uniformly in probability to the distribution of the predictive root $X_{T+2} - \widehat{X}_{T+2}$ in the real world. This result guarantees the asymptotic validity of the PPI. We summarize this conclusion in Theorem 2.

Theorem 2. *Under assumptions A1–A8 and observed data $\{X_0, \ldots, X_T\}$, we have:*

$$\sup_{|x| \leq c_T} \left| F_{X^*_{T+k} - \widehat{X}^*_{T+k} | X_T, \ldots, X_0}(x) - F_{X_{T+k} - \widehat{X}_{T+k} | X_T, \ldots, X_0}(x) \right| \xrightarrow{p} 0, \text{ for } k \geq 1, \tag{22}$$

where $X^*_{T+k} - \widehat{X}^*_{T+k}$ *is the k-step-ahead predictive root in the bootstrap world, and* $F_{X^*_{T+k}-\widehat{X}^*_{T+k}|X_T,\ldots,X_0}(x)$ *represents its distribution at point x;* $X_{T+k} - \widehat{X}_{T+k}$ *is the k-step-ahead predictive root in the real world, and* $F_{X_{T+k}-\widehat{X}_{T+k}|X_T,\ldots,X_0}(x)$ *represents its (conditional) distribution at point x. This theorem holds for both bandwidth selection strategies.*

However, since we apply a more complicated procedure to capture estimation variability, we anticipate that this results in a PPI. To see this, we first apply the Taylor expansion on the r.h.s. of Equations (20) and (21); the two predictive roots can be decomposed into several parts:

$$\begin{aligned} X_{T+2} - \widehat{X}_{T+2} &= m(m(X_T)) - \widehat{m}_h(\widehat{m}_h(X_T)) + m^{(1)}(\hat{x})\epsilon_{T+1} + \epsilon_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\widehat{m}_h^{(1)}(\hat{\hat{x}}_i)\hat{\epsilon}_{i,T+1}; \\ X^*_{T+2} - \widehat{X}^*_{T+2} &= \widehat{m}_g(\widehat{m}_g(X_T)) - \widehat{m}^*_h(\widehat{m}^*_h(X_T)) + \widehat{m}_g^{(1)}(\hat{x}^*)\hat{\epsilon}^*_{T+1} + \hat{\epsilon}^*_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\widehat{m}_h^{*(1)}(\hat{\hat{x}}^*_i)\hat{\epsilon}^*_{i,T+1}, \end{aligned} \tag{23}$$

where $\hat{x}$ and $\hat{x}^*$ are some points between $m(X_T)$ and $m(X_T) + \epsilon_{T+1}$ and between $\widehat{m}_g(X_T)$ and $\widehat{m}_g(X_T) + \hat{\epsilon}^*_{T+1}$, respectively; $\hat{\hat{x}}_i$ and $\hat{\hat{x}}^*_i$ are some points between $\widehat{m}_h(X_T)$ and $\widehat{m}_h(X_T) + \hat{\epsilon}_{i,T+1}$ and between $\widehat{m}^*_h(X_T)$ and $\widehat{m}^*_h(X_T) + \hat{\epsilon}^*_{i,T+1}$, respectively. The k-step-ahead predictive root can be expressed similarly when $k > 2$. We can consider the r.h.s of Equation (23) to be made up of two components in both the real and bootstrap worlds: (1) the two-step-ahead estimation variability component, $m(m(X_T)) - \widehat{m}_h(\widehat{m}_h(X_T))$ and $\widehat{m}_g(\widehat{m}_g(X_T)) - \widehat{m}^*_h(\widehat{m}^*_h(X_T))$; (2) the rest of the terms, which are related to future innovations. For the second component, the bootstrap can mimic the real-world situation well.

We expect that the first component, i.e., the variability in local constant estimation of the mean function $m(m(X_T)) - \widehat{m}_h(\widehat{m}_h(X_T))$, can be well approximated by its variant $\widehat{m}_g(\widehat{m}_g(X_T)) - \widehat{m}^*_h(\widehat{m}^*_h(X_T))$ in the bootstrap world. Although PPIs with either of the two bandwidth selection approaches are both asymptotically valid, the PPI with the bandwidth strategy (B2) is only "almost" pertinent for multi-step-ahead predictions since the variability in local constant estimation is not well estimated in finite samples; see also the simulation results in Section 4 and Appendix C. On the other hand, the PPI with the bandwidth strategy (B1) meets our goal. We summarize this finding in Theorem 3.

Theorem 3. *Under assumptions A1–A8 and with observed data* $\{X_0, \ldots, X_T\} \in \Omega_T$*, where* $\mathbb{P}((X_0, \ldots, X_T) \in \Omega_T) = 1 - o(1)$ *as* $T \to \infty$*, by taking the bandwidth strategy (B1), we can build a confidence bound for the local constant estimation at step k:*

$$\begin{aligned} \sup_{|x|\le c_T} \Big| &\mathbb{P}\Big(a_T\Big(\mathcal{M}_k(X_T) - \widehat{\mathcal{M}}_{h,k}(X_T)\Big) \le x\Big) - \\ &\mathbb{P}\Big(a_T\Big(\mathcal{M}^*_{h,k}(X_T) - \widehat{\mathcal{M}}^*_{h,k}(X_T)\Big) \le x\Big)\Big| \xrightarrow{p} 0, \text{ for } k \ge 1; \end{aligned} \tag{24}$$

$\mathcal{M}_k(X_T)$ *can be expressed by iteratively computing* $X_{T+i} = m(X_{T+i-1})$ *for* $i = 1, \ldots, k$*; i.e., it has the form below:*

$$\mathcal{M}_k(X_T) = m(m(\cdots(m(m(X_T))))); \tag{25}$$

$\widehat{\mathcal{M}}_{h,k}(X_T)$ *can be expressed by iteratively computing* $X_{T+i} = \widehat{m}_h(X_{T+i-1})$ *for* $i = 1, \ldots, k$*; i.e., it has the form below:*

$$\widehat{\mathcal{M}}_{h,k}(X_T) = \widehat{m}_h(\widehat{m}_h(\cdots \widehat{m}_h(\widehat{m}_h(X_T))\cdots)); \tag{26}$$

$\mathcal{M}^*_{h,k}(X_T)$ *and* $\widehat{\mathcal{M}}^*_{h,k}(X_T)$ *can be expressed similarly.*

The direct implication of Theorem 3 is that the PPI generated by Algorithms 3 and 4 should have a better CVR for small sample sizes than the QPI since the estimation variability is included in the PI with *high probability*; see the simulation examples in Section 4.

3.3. On PPI with Heteroscedastic Errors

For time-series models with heteroscedastic errors, i.e., where the variance function $\sigma(x)$ represents the heteroscedasticity of innovations, we do not need to care about the bias term in the nonparametric estimation of the variance function. In other words, we use neither under-smoothing nor over-smoothing bandwidth tricks on the variance function to generate the bootstrap series for covering the bias term; we can just use the bandwidth with the optimal rate to estimate the variance function from real and bootstrap series.

To see this, let us consider the two-step-ahead predictive root with heteroscedastic errors. In the real world, we have:

$$\begin{aligned} & X_{T+2} - \widehat{X}_{T+2} \\ & = m(X_{T+1}) + \sigma(X_{T+1})\epsilon_{T+2} - \frac{1}{M}\sum_{i=1}^{M}(\widehat{m}_h(\widehat{m}_h(X_T) + \widehat{\sigma}_h(X_T)\hat{\epsilon}_{i,T+1}) + \widehat{\sigma}_h(X_{T+1})\hat{\epsilon}_{i,T+2}) \\ & \approx m(m(X_T) + \sigma(X_T)\epsilon_{T+1}) + \sigma(X_{T+1})\epsilon_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\widehat{m}_h(\widehat{m}_h(X_T) + \widehat{\sigma}_h(X_T)\hat{\epsilon}_{i,T+1}). \end{aligned} \tag{27}$$

Correspondingly, the predictive root in the bootstrap world is:

$$\begin{aligned} & X^*_{T+2} - \widehat{X}^*_{T+2} \\ & = \widehat{m}_g(X^*_{T+1}) + \widehat{\sigma}_g(X^*_{T+1})\epsilon^*_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\left(\widehat{m}^*_h\left(\widehat{m}^*_h(X_T) + \widehat{\sigma}^*_h(X_T)\hat{\epsilon}^*_{i,T+1}\right) + \widehat{\sigma}^*_h(X^*_{T+1})\hat{\epsilon}^*_{i,T+2}\right) \\ & \approx \widehat{m}_g(\widehat{m}_g(X_T) + \widehat{\sigma}_g(X^*_T)\hat{\epsilon}^*_{T+1}) + \widehat{\sigma}_g(X^*_{T+1})\hat{\epsilon}^*_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\widehat{m}^*_h\left(\widehat{m}^*_h(X_T) + \widehat{\sigma}^*_h(X_T)\hat{\epsilon}^*_{i,T+1}\right). \end{aligned} \tag{28}$$

Through Taylor expansion, we can obtain:

$$\begin{aligned} X_{T+2} - \widehat{X}_{T+2} &\approx m(m(X_T)) - \widehat{m}_h(\widehat{m}_h(X_T)) \\ &\quad + m^{(1)}(\hat{x})\sigma(X_T)\epsilon_{T+1} + \sigma(X_{T+1})\epsilon_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\widehat{m}^{(1)}_h(\hat{\hat{x}}_i)\widehat{\sigma}_h(X_T)\hat{\epsilon}_{i,T+1}; \\ X^*_{T+2} - \widehat{X}^*_{T+2} &\approx \widehat{m}_g(\widehat{m}_g(X_T)) - \widehat{m}^*_h(\widehat{m}^*_h(X_T)) \\ &\quad + \widehat{m}^{(1)}_g(\hat{x}^*)\widehat{\sigma}_g(X^*_T)\hat{\epsilon}^*_{T+1} + \widehat{\sigma}_g(X^*_{T+1})\hat{\epsilon}^*_{T+2} - \frac{1}{M}\sum_{i=1}^{M}\widehat{m}^{*(1)}_h(\hat{\hat{x}}^*_i)\widehat{\sigma}^*_h(X_T)\hat{\epsilon}^*_{i,T+1}. \end{aligned} \tag{29}$$

We can still consider the r.h.s. of Equation (29) to contain two components. Once we use the under-smoothing technique to cover the estimation variability for the mean function, since the residual distribution is determined by the estimated mean and variance functions, the convergence rate of the residual distribution to the true innovation distribution is dominated by the convergence rate of $\widehat{m}_h(x)$ to $m(x)$. In addition, all estimators of the variance function in Equation (29) are tied with future estimated innovations; so, we are free to use the bandwidth $g = h$ with the optimal smoothing rate to estimate the variance function, and the overall convergence rate will not change. To show this benefit, we ran some simulations, which are presented in Appendix D, to compare the performance of PIs when applying under-smoothing or the optimal bandwidth in estimating the variance function. In Section 4, we will demonstrate the use of the optimal bandwidth to estimate the variance function if the time series is heteroscedastic.

To analyze the pertinence of the PPI for time series with heteroscedastic errors, from Equation (29), it is apparent that the distribution of $m(m(X_T)) - \widehat{m}_h(\widehat{m}_h(X_T))$ can still be approximated by $\widehat{m}_g(\widehat{m}_g(X_T)) - \widehat{m}^*_h(\widehat{m}^*_h(X_T))$. For the rest of the terms, the bootstrap can still mimic the real-world situation.

4. Simulations

In this section, we describe the simulations that we deployed to check the performance of five-step-ahead point predictions and the corresponding PIs of our algorithms in the *R* platform with finite samples. To obtain the optimal bandwidth h_{op} for our local constant estimators, we relied on the function *npregbw* from the *R* package *np*. The under-smoothing and over-smoothing bandwidths were taken as $0.5 \cdot h_{op}$ and $2 \cdot h_{op}$, respectively.

4.1. Optimal Point Prediction

We first consider a simple non-linear model:

$$X_t = \log(X_{t-1}^2 + 1) + \epsilon_t, \tag{30}$$

where $\{\epsilon_t\}$ is assumed to have a standard normal distribution. The geometric ergodicity of Equation (30) can be easily checked.

We apply the "oracle" prediction as the benchmark. The oracle prediction is returned by employing a simulation approach, assuming that we know the true model and the error distribution, i.e., the simulation-based prediction, as we discussed in Section 1; see Section 3.2 of [17] for more details and the theoretical validation of this approach. Since this oracle prediction should have the best performance, we would like to challenge our nonparametric bootstrap-based methods by comparing them with the oracle prediction. We also pretend that the true model and innovation distribution are unknown when we perform the nonparametric bootstrap-based prediction. For point predictions, we just utilize fitted residuals. The application of predictive residuals will play a role in building PIs later.

In a single experiment, we take $X_0 \sim \text{Uniform}(-1, 1)$ and then iteratively generate a series with size $C + T + 1$ according to Equation (30). Here, C is taken as 200 to remove the effects of the initial distribution of X_0. To perform oracle predictions, we take $M = 1000$ to obtain a satisfying approximation. For a fair comparison, we also apply a 1000 times bootstrap in Algorithms 1 and 2 to obtain bootstrap-based predictions.

Referring to the simulation studies in [16], we take $T = 100,\ 200$, $k = 1, \ldots, 5$, and employ the Mean-Squared Prediction Error (MSPE) to compare oracle and bootstrap predictions. The metric MSPE can be approximated based on the formula below:

$$\text{MSPE of the } k\text{-th ahead prediction} = \frac{1}{N}\sum_{n=1}^{N}(X_{n,k} - P_{n,k})^2, \text{ for } k = 1, \ldots, 5, \tag{31}$$

where $P_{n,k}$ represents the k-th step-ahead optimal L_1 or L_2 point predictions implied by the bootstrap or simulation approach, and $X_{n,k}$ stands for the true future value in the n-th replication. We take $N = 5000$ and record all MSPEs in Table 1.

Table 1. The MSPEs of different predictions using Model Equation (30) with a standard normal innovation.

Model:	$X_t = \log(X_{t-1}^2 + 1) + \epsilon_t, \epsilon_t \sim N(0,1)$					
$T = 100$	Prediction step	1	2	3	4	5
L_2-Bootstrap		1.1088	1.5223	1.6088	1.5886	1.6282
L_1-Bootstrap		1.1123	1.5290	1.6212	1.6011	1.6385
L_2-Oracle		1.0181	1.4521	1.5529	1.5273	1.5731
L_1-Oracle		1.0198	1.4540	1.5554	1.5305	1.5734
$T = 200$						
L_2-Bootstrap		1.0142	1.4006	1.5380	1.5956	1.6102
L_1-Bootstrap		1.0134	1.4041	1.5426	1.6024	1.6171
L_2-Oracle		0.9790	1.3671	1.4982	1.5556	1.5791
L_1-Oracle		0.9793	1.3681	1.4999	1.5568	1.5791

From Table 1, we can find that the MSPEs of oracle- and bootstrap-based L_1- or L_2-optimal predictions are very close to each other, respectively. The MSPEs of oracle optimal predictions are always smaller than the corresponding bootstrap predictions. This phenomenon is in line with our expectation since the bootstrap prediction is obtained with an estimated model and innovation distribution.

Rather than applying the standard normal distribution, we consider a skewed innovation, i.e., $\epsilon_t \sim \chi^2(3) - 3$. Repeating the above process, we present the MSPEs in Table 2.

Table 2. The MSPEs of different predictions using Model Equation (30) with $\chi(3) - 3$ innovation.

Model:		$X_t = \log(X_{t-1}^2 + 1) + \epsilon_t, \epsilon_t \sim \chi(3) - 3$				
$T = 100$	Prediction step	1	2	3	4	5
L_2-Bootstrap		6.7286	7.6087	7.8202	7.3395	7.6966
L_1-Bootstrap		7.1093	7.9908	8.2598	7.6761	7.9988
L_2-Oracle		6.2972	7.3608	7.6953	7.1766	7.5157
L_1-Oracle		6.6937	7.6540	8.0064	7.3889	7.7174
$T = 200$						
L_2-Bootstrap		6.2457	7.1662	7.5042	7.6227	7.1980
L_1-Bootstrap		6.6355	7.4942	7.7964	7.9285	7.5006
L_2-Oracle		5.9531	7.0244	7.3823	7.4382	7.0738
L_1-Oracle		6.3519	7.2785	7.5810	7.6443	7.2600

The performance of bootstrap-based predictions is also competitive with oracle predictions. Another notable phenomenon indicated by Table 2 is that the MSPE of L_2-optimal predictions is always less than its corresponding value in L_1-optimal predictions. The reason for this is that the L_2-optimal prediction coincides with the L_2 loss used in MSPE. However, this phenomenon is not remarkable for the results in Table 1 since the innovation distribution is symmetric in that case.

For the non-linear model with heteroscedastic errors, we consider the following model:

$$X_t = \sin(X_{t-1}) + \epsilon_t \sqrt{0.5 + 0.25 X_{t-1}^2}. \tag{32}$$

Model Equation (32) is in a GARCH form, except that the regression function is non-linear. This model was also considered by [16]. We present the MSPEs of different predictions in Table 3. It reveals that our bootstrap-based optimal point prediction methods can work for the non-linear time-series model with heteroscedastic errors, and its performance is still competitive with oracle predictions.

Table 3. The MSPEs of different predictions using Model Equation (32) with standard normal innovation.

Model:		$X_t = \sin(X_{t-1}) + \epsilon_t \sqrt{0.5 + 0.25 X_{t-1}^2}, \epsilon_t \sim N(0,1)$				
$T = 100$	Prediction step	1	2	3	4	5
L_2-Bootstrap		0.9447	1.1306	1.2373	1.2091	1.2714
L_1-Bootstrap		0.9461	1.1374	1.2396	1.2127	1.2731
L_2-Oracle		0.8454	1.0726	1.1832	1.1722	1.2186
L_1-Oracle		0.8457	1.0730	1.1841	1.1737	1.2183
$T = 200$						
L_2-Bootstrap		0.8798	1.1539	1.2600	1.2901	1.2717
L_1-Bootstrap		0.8833	1.1600	1.2649	1.2949	1.2749
L_2-Oracle		0.8103	1.0991	1.2227	1.2680	1.2509
L_1-Oracle		0.8107	1.1000	1.2239	1.2684	1.2511

Remark 8. *In practice, we should mention that both local constant estimators* $\widehat{m}(x)$ *and* $\widehat{\sigma}(x)$ *will only be accurate when x falls in the area where data are dense. Estimations in the sparse area will return large fitted residuals. These large residuals will spoil the multi-step-ahead prediction process in the bootstrap procedure. Thus, depending on which optimal prediction we are pursuing, we replace all inappropriate or numerical NaN values with the sample mean or sample median of observed data. In addition, during the simulation studies, we truncate* $\widetilde{m}(x)_h$*; i.e., we take* C_m *as* $5 \cdot \max\{|x_0|, \ldots, |x_T|\}$*. For the mean function estimator* $\widetilde{m}(x)_h^*$ *in the bootstrap world, we take* C_m^* *as* $\min\{2 \cdot C_m, 5 \cdot \max\{|x_0^*|, \ldots, |x_T^*|\}\}$ *since we want to allow more variability for the bootstrap series. For the local constant estimator of the variance function, we take* c_σ *and* c_σ^* *as* 0.01*. We take* C_σ *and* C_σ^* *as* $2 \cdot \hat{\sigma}$ *and* $\min\{4 \cdot \hat{\sigma}, 2 \cdot \hat{\sigma}^*\}$*, respectively;* $\hat{\sigma}$ *and* $\hat{\sigma}^*$ *are the sample standard deviations of the observed series in the real world and bootstrap world, respectively. These truncating constants work well for the above two models. In practice, a cross-validation approach could be taken to find the optimal truncating constants.*

4.2. QPI and PPI

In this subsection, we try to evaluate the CVR of the QPI and PPI based on the nonparametric forward bootstrap prediction method. Similarly, we take the oracle prediction interval as the benchmark, which is computed by the QPI with a known model and innovation distribution; see the discussion in Section 1 and Section 3.2 of [17] for references on this approach.

Due to the time complexity of the double bootstrap in the bootstrap world, we only take $B = 500$ and $M = 100$ in Algorithms 3 and 4 to derive the PPI. Correspondingly, we take $M = 500$ to compute the QPI. In practice, people can increase the values of B and M. To make the result as consistent as possible, we still repeat the simulation process 5000 times.

The empirical CVR of the bootstrap-based QPI and PPI for $k = 1, \ldots, 5$-step-ahead predictions is determined with the formula below:

$$\text{CVR of the } k\text{-th ahead prediction} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{X_{n,k} \in [L_{n,k}, U_{n,k}]}, \text{for } k = 1, \ldots, 5, \tag{33}$$

where $[L_{n,k}, U_{n,k}]$ and $X_{n,k}$ represent the k-th step-ahead prediction interval and the true future value in the n-th replication, respectively. In addition to the CVR, we are also concerned about the empirical LEN of different PIs. The empirical LEN of a PI is defined as follows:

$$\text{LEN of the } k\text{-th ahead PI} = \frac{1}{N}\sum_{n=1}^{N} (U_{n,k} - L_{n,k}), \text{for } k = 1, \ldots, 5. \tag{34}$$

Recall that the PPI can be centered at the L_1- or L_2-optimal point predictor, and the QPI can be found with the optimal bandwidth and the under-smoothing bandwidth; thus, we have four types of PIs based on the bootstrap. In particular, each type of PI can be obtained with fitted or predictive residuals. In total, we have eight bootstrap-type PIs and one oracle PI. In addition, to observe the effects of introducing the predictive residuals and the superiority of the PPI, we consider three sample sizes, 50, 100, and 200. All CVRs and LENs for different PIs for predicting Equations (30) and (32) are presented in Tables 4 and 5, respectively.

From there, we can observe that the SPI (oracle PI) is the best one, according to the most accurate CVR and relatively small LEN. For the QPI with fitted residuals, it severely under-covers the true future value, especially for data with a small sample size. With predictive residuals, although the LEN of the PI gets amplified, the CVR of the QPI improves significantly. After applying the under-smoothing bandwidth with the QPI, the CVR is further improved for multi-step-ahead (i.e., $k \geq 2$) predictions, regardless of whether fitted or predictive residuals are used. The PPI with fitted residuals outperforms the QPI with fitted residuals. The PPI with predictive residuals can achieve the most

accurate CVR among various bootstrap-based PIs, especially when the data are short, though the price is that its LEN is the largest compared to other PIs. We should note that the QPI with predictive residuals and the under-smoothing bandwidth can achieve a great CVR with 200 samples for these two models. However, we may not know the sufficiently large sample size to guarantee that the QPI can work well. Thus, we recommend taking the PPI with predictive residuals as the first choice.

Remark 9. *We should clarify that the CVR computed using Equation* (33) *is the unconditional coverage rate of X_{T+k} since it is an average of the conditional coverage of X_{T+k} for all replications.*

Table 4. The CVRs and LENs of PIs for Equation (30).

Model 1:	$X_t = \log(X_{t-1}^2 + 1) + \epsilon_t, \epsilon_t \sim N(0,1)$									
	CVR for each step					LEN for each step				
$T = 200$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.936	0.935	0.931	0.928	0.925	3.80	4.38	4.52	4.55	4.57
QPI-p	0.943	0.944	0.939	0.935	0.937	3.94	4.54	4.69	4.73	4.74
QPI-f-u	0.936	0.941	0.940	0.937	0.937	3.80	4.51	4.69	4.76	4.77
QPI-p-u	0.942	0.949	0.949	0.945	0.949	3.95	4.68	4.86	4.92	4.94
L_2-PPI-f-u	0.940	0.944	0.944	0.940	0.939	3.94	4.59	4.76	4.81	4.83
L_2-PPI-p-u	0.947	0.954	0.951	0.947	0.947	4.09	4.75	4.92	4.98	5.00
L_1-PPI-f-u	0.942	0.945	0.944	0.940	0.941	3.95	4.61	4.77	4.83	4.84
L_1-PPI-p-u	0.948	0.954	0.952	0.948	0.949	4.10	4.77	4.94	4.99	5.01
SPI	0.951	0.948	0.950	0.944	0.946	3.88	4.58	4.77	4.82	4.84
$T = 100$										
QPI-f	0.921	0.918	0.912	0.913	0.909	3.74	4.28	4.40	4.44	4.45
QPI-p	0.940	0.935	0.931	0.931	0.928	3.99	4.54	4.67	4.71	4.72
QPI-f-u	0.916	0.928	0.931	0.930	0.927	3.74	4.46	4.63	4.69	4.71
QPI-p-u	0.937	0.943	0.943	0.944	0.943	3.99	4.72	4.89	4.95	4.97
L_2-PPI-f-u	0.931	0.934	0.935	0.934	0.931	3.97	4.58	4.73	4.78	4.80
L_2-PPI-p-u	0.949	0.948	0.947	0.944	0.947	4.22	4.84	4.99	5.04	5.07
L_1-PPI-f-u	0.931	0.936	0.934	0.933	0.934	3.98	4.60	4.75	4.79	4.82
L_1-PPI-p-u	0.949	0.948	0.949	0.944	0.948	4.23	4.86	5.01	5.06	5.09
SPI	0.951	0.941	0.946	0.942	0.944	3.89	4.58	4.76	4.82	4.84
$T = 50$										
QPI-f	0.891	0.898	0.899	0.890	0.887	3.64	4.14	4.25	4.29	4.30
QPI-p	0.923	0.926	0.931	0.924	0.917	4.04	4.56	4.67	4.71	4.72
QPI-f-u	0.884	0.916	0.921	0.918	0.907	3.64	4.37	4.54	4.60	4.62
QPI-p-u	0.914	0.939	0.940	0.939	0.934	4.03	4.79	4.95	5.00	5.02
L_2-PPI-f-u	0.906	0.924	0.924	0.927	0.919	3.99	4.56	4.69	4.74	4.76
L_2-PPI-p-u	0.936	0.951	0.948	0.944	0.943	4.41	4.97	5.10	5.15	5.16
L_1-PPI-f-u	0.907	0.925	0.924	0.927	0.920	4.00	4.58	4.72	4.76	4.79
L_1-PPI-p-u	0.939	0.952	0.948	0.945	0.941	4.43	5.00	5.12	5.17	5.18
SPI	0.947	0.949	0.944	0.947	0.942	3.88	4.58	4.76	4.81	4.84

Note: With no other specifications, throughout all simulations, QPI-f and QPI-p represent QPIs based on optimal bandwidth with fitted and predictive residuals, respectively; QPI-f-u and QPI-p-u represent QPIs based on under-smoothing bandwidth with fitted and predictive residuals, respectively; L_2-PPI-f-u and L_2-PPI-p-u represent PPIs centered at L_2-optimal point prediction with fitted and predictive residuals, respectively; L_1-PPI-f-u and L_1-PPI-p-u represent PPIs centered at L_1-optimal point prediction with fitted and predictive residuals, respectively; all PPIs with the "-u" symbol are based on applying the under-smoothing bandwidth to estimate the model; SPI represents the oracle PI.

Table 5. The CVRs and LENs of PIs for Equation (32).

Model 2:	$X_t = \sin(X_{t-1}) + \epsilon_t\sqrt{0.5 + 0.25X_{t-1}^2}, \epsilon_t \sim N(0,1)$									
	CVR for each step					LEN for each step				
$T = 200$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.913	0.918	0.916	0.924	0.924	3.30	3.93	4.07	4.11	4.12
QPI-p	0.935	0.936	0.933	0.941	0.940	3.62	4.29	4.46	4.49	4.51
QPI-f-u	0.904	0.934	0.935	0.943	0.944	3.34	4.25	4.50	4.55	4.57
QPI-p-u	0.926	0.949	0.951	0.958	0.955	3.65	4.62	4.89	4.95	4.97
L_2-PPI-f-opv	0.909	0.938	0.937	0.948	0.946	3.51	4.38	4.60	4.65	4.67
L_2-PPI-p-opv	0.932	0.952	0.951	0.961	0.959	3.87	4.80	5.03	5.08	5.10
L_1-PPI-f-opv	0.912	0.939	0.937	0.949	0.946	3.53	4.38	4.59	4.64	4.66
L_1-PPI-p-opv	0.933	0.951	0.950	0.960	0.960	3.88	4.79	5.02	5.07	5.08
SPI	0.948	0.948	0.940	0.950	0.946	3.37	4.11	4.32	4.38	4.40
$T = 100$										
QPI-f	0.901	0.907	0.912	0.909	0.906	3.28	3.85	3.97	4.01	4.01
QPI-p	0.933	0.931	0.938	0.933	0.938	3.82	4.41	4.55	4.58	4.59
QPI-f-u	0.901	0.923	0.931	0.929	0.932	3.28	4.07	4.29	4.35	4.37
QPI-p-u	0.931	0.943	0.950	0.950	0.947	3.82	4.64	4.85	4.90	4.93
L_2-PPI-f-opv	0.915	0.925	0.935	0.936	0.935	3.52	4.25	4.43	4.48	4.50
L_2-PPI-p-opv	0.941	0.948	0.954	0.955	0.954	4.17	4.90	5.07	5.11	5.13
L_1-PPI-f-opv	0.916	0.926	0.935	0.936	0.936	3.53	4.25	4.43	4.48	4.50
L_1-PPI-p-opv	0.941	0.947	0.954	0.952	0.955	4.17	4.90	5.07	5.12	5.13
SPI	0.951	0.947	0.947	0.946	0.942	3.41	4.13	4.33	4.39	4.40
$T = 50$										
QPI-f	0.844	0.874	0.884	0.883	0.888	3.09	3.68	3.83	3.87	3.89
QPI-p	0.903	0.921	0.929	0.929	0.934	4.01	4.74	4.85	4.93	4.95
QPI-f-u	0.845	0.892	0.907	0.910	0.910	3.09	3.93	4.15	4.23	4.26
QPI-p-u	0.905	0.929	0.934	0.940	0.946	4.03	4.91	5.17	5.23	5.24
L_2-PPI-f-opv	0.871	0.905	0.917	0.918	0.922	3.45	4.19	4.38	4.46	4.47
L_2-PPI-p-opv	0.934	0.941	0.948	0.950	0.954	4.71	5.48	5.60	5.67	5.68
L_1-PPI-f-opv	0.873	0.907	0.920	0.919	0.923	3.46	4.20	4.40	4.47	4.48
L_1-PPI-p-opv	0.934	0.942	0.948	0.950	0.954	4.69	5.44	5.57	5.64	5.64
SPI	0.942	0.946	0.948	0.939	0.950	3.39	4.11	4.33	4.38	4.40

Note: All PPIs with the "-opv" symbol are based on applying under-smoothing and optimal bandwidths to estimate mean and variance functions, respectively.

4.3. Simulation Results for Appendices

We carried out a simulation study to show that the QPIs with the optimal bandwidth and under-smoothing bandwidth are asymptotically equivalent; see the results in Table 6, and see the formal analysis in Appendix B.

Table 6. The CVRs and LENs of QPIs with 1000 samples using Equation (30).

Model 1:	$X_t = \log(X_{t-1}^2 + 1) + \epsilon_t, \epsilon_t \sim N(0,1)$									
	CVR for each step					LEN for each step				
$T = 1000$	1	2	3	4	5	1	2	3	4	5
QPI-f	0.950	0.940	0.948	0.947	0.939	3.86	4.50	4.66	4.70	4.71
QPI-f-u	0.947	0.943	0.952	0.954	0.946	3.86	4.56	4.74	4.79	4.81
QPI-p	0.949	0.938	0.951	0.951	0.943	3.91	4.54	4.71	4.75	4.76
QPI-p-u	0.951	0.947	0.954	0.956	0.950	3.90	4.62	4.80	4.84	4.86

We also deployed simulations to check the effects of applying under-smoothing or over-smoothing tricks on the performance of the PPI. We took the sample size $T + 1$ to be 50 or 500 and performed simulations 5000 times on the first model; see Table 7 and Appendix C for the results and analysis, respectively.

Table 7. The CVRs and LENs of PPIs with under-smoothing or over-smoothing bandwidth strategies using Equation (30).

Model 1:	$X_t = \log(X_{t-1}^2 + 1) + \epsilon_t, \epsilon_t \sim N(0,1)$									
	CVR for each step					LEN for each step				
$T = 500$	1	2	3	4	5	1	2	3	4	5
L_2-PPI-f-u	0.943	0.940	0.945	0.943	0.948	3.88	4.54	4.71	4.77	4.78
L_1-PPI-f-u	0.942	0.941	0.946	0.947	0.949	3.89	4.55	4.72	4.78	4.80
L_2-PPI-p-u	0.946	0.949	0.947	0.952	0.954	3.96	4.63	4.79	4.85	4.8
L_1-PPI-p-u	0.946	0.950	0.947	0.951	0.954	3.97	4.64	4.81	4.86	4.88
L_2-PPI-f-o	0.942	0.926	0.916	0.915	0.923	3.86	4.26	4.33	4.34	4.35
L_1-PPI-f-o	0.943	0.925	0.921	0.918	0.922	3.87	4.27	4.34	4.36	4.36
L_2-PPI-p-o	0.948	0.929	0.927	0.927	0.925	3.94	4.34	4.42	4.43	4.43
L_1-PPI-p-o	0.949	0.931	0.928	0.925	0.924	3.95	4.35	4.43	4.44	4.44
SPI	0.946	0.947	0.948	0.950	0.956	3.89	4.57	4.76	4.82	4.84
$T = 50$										
L_2-PPI-f-u	0.912	0.919	0.919	0.925	0.931	3.95	4.53	4.67	4.72	4.74
L_1-PPI-f-u	0.913	0.921	0.919	0.928	0.931	3.96	4.55	4.69	4.74	4.76
L_2-PPI-p-u	0.943	0.945	0.942	0.946	0.950	4.38	4.95	5.08	5.12	5.14
L_1-PPI-p-u	0.944	0.946	0.943	0.948	0.950	4.39	4.98	5.10	5.15	5.16
L_2-PPI-f-o	0.911	0.880	0.869	0.869	0.873	3.78	3.93	3.96	3.97	3.97
L_1-PPI-f-o	0.912	0.882	0.868	0.868	0.871	3.79	3.95	3.98	3.98	3.98
L_2-PPI-p-o	0.940	0.918	0.903	0.908	0.910	4.20	4.37	4.40	4.41	4.42
L_1-PPI-p-o	0.941	0.919	0.902	0.909	0.909	4.22	4.39	4.42	4.43	4.43
SPI	0.950	0.947	0.946	0.947	0.950	3.89	4.58	4.76	4.82	4.84

Note: "-o" indicates that the corresponding PPI is built with an over-smoothing bandwidth in generating bootstrap series.

For a model with heteroscedastic errors, as we mentioned in Section 3.3, we can rely on the optimal bandwidth to estimate the variance functions. To check this claim, we consider two strategies for the bandwidth of the estimator for the variance function: (1) take the under-smoothing bandwidth as we do for the mean function estimator; (2) take the bandwidth with the optimal rate. To estimate the mean function in the bootstrap world, we continue using the under-smoothing bandwidth strategy. The simulation results based on Equation (32) with a small sample size are shown in Table 8; see the corresponding discussion in Appendix D.

Table 8. The CVRs and LENs of PPIs with two strategies for estimating the variance function.

Model 1:	$X_t = \sin(X_{t-1}) + \epsilon_t \sqrt{0.5 + 0.25X_{t-1}^2}, \epsilon_t \sim N(0,1)$									
	CVR for each step					LEN for each step				
$T = 50, \text{Rep} = 5000$										
L_2-PPI-f-u	0.871	0.896	0.921	0.915	0.923	3.50	4.24	4.41	4.48	4.52
L_1-PPI-f-u	0.877	0.901	0.919	0.918	0.925	3.52	4.24	4.42	4.49	4.53
L_2-PPI-p-u	0.925	0.939	0.946	0.946	0.946	4.82	5.47	5.63	5.71	5.81
L_1-PPI-p-u	0.927	0.935	0.945	0.949	0.949	4.80	5.39	5.51	5.65	5.75
L_2-PPI-f-opv	0.885	0.891	0.923	0.920	0.918	3.45	4.12	4.34	4.39	4.43
L_1-PPI-f-opv	0.885	0.893	0.927	0.919	0.917	3.47	4.14	4.36	4.41	4.45
L_2-PPI-p-opv	0.934	0.939	0.947	0.950	0.947	4.75	5.28	5.49	5.56	5.60
L_1-PPI-p-opv	0.940	0.940	0.946	0.951	0.943	4.72	5.21	5.40	5.45	5.55
SPI	0.943	0.939	0.958	0.945	0.945	3.38	4.11	4.33	4.38	4.40

Note: "-opv" indicates the corresponding PPI is built by optimal bandwidth for the variance function estimator.

5. Conclusions

In this paper, we propose some forward bootstrap prediction algorithms based on the local constant estimation of the model. With theoretical and practical validations, we show that our bootstrap-based point predictions work well, and their MSPEs are very close to those of the oracle predictions. By debiasing the nonparametric estimation with the under-smoothing bandwidth, we show that the confidence bounds for the multi-step-ahead estimator can be approximated by the bootstrap. As a result, we can obtain a pertinence prediction interval by using a specifically designed algorithm. Empirically, we further take the predictive residuals to make predictions that can alleviate the under-coverage of the PI for a small sample size. Among different bootstrap-based PIs, as revealed by simulation studies, the PPI with predictive residuals is the best one, and it is competitive with the oracle PI.

Author Contributions: All authors D.N.P. and K.W. contributed equally to this project. All authors have read and agreed to the published version of the manuscript.

Funding: The research of the first author was partially supported by NSF grant DMS 19-14556. The research of the second author was partially supported by the Richard Libby Graduate Research Award.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Proof of Theorem 1. To show that Equation (15) is satisfied for $k \geq 1$, we can just show the case with $k = 2$. Cases with $k = 1$ and $k > 2$ can be handled similarly. $F_{X_{T+2}|X_T}(x)$ is equivalent to:

$$\begin{aligned} F_{X_{T+2}|X_T}(x) &= \mathbb{P}(X_{T+2} \leq x | X_T) \\ &= \mathbb{P}(m(X_{T+1}) + \sigma(X_{T+1})\epsilon_{T+2} \leq x | X_T) \\ &= \mathbb{P}\left(\epsilon_{T+2} \leq \frac{x - m(m(X_T) + \sigma(X_t)\epsilon_{T+1})}{\sigma(m(X_T) + \sigma(X_t)\epsilon_{T+1})} \middle| X_T\right) \\ &= \mathbb{E}\left[\mathbb{P}\left(\epsilon_{T+2} \leq \frac{x - m(m(X_T) + \sigma(X_T)\epsilon_{T+1})}{\sigma(m(X_T) + \sigma(X_T)\epsilon_{T+1})} \middle| \epsilon_{T+1}, X_T\right) \middle| X_T\right] \\ &= \mathbb{E}\left[F_\epsilon\left(\frac{x - m(m(X_T) + \sigma(X_T)\epsilon_{T+1})}{\sigma(m(X_T) + \sigma(X_T)\epsilon_{T+1})}\right) \middle| X_T\right] \\ &= \mathbb{E}\left[F_\epsilon(\mathcal{G}(x, X_T, \epsilon_{T+1})) \middle| X_T\right]; \end{aligned} \tag{A1}$$

we use $\mathcal{G}(x, X_T, \epsilon_{T+1})$ to represent $\frac{x - m(m(X_T) + \sigma(X_T)\epsilon_{T+1})}{\sigma(m(X_T) + \sigma(X_T)\epsilon_{T+1})}$ to simplify notations. Similarly, we can analyze $F_{X^*_{T+2}}(x)$, and it has the following equivalent expressions:

$$\begin{aligned} F_{X^*_{T+2}|X_T,\ldots,X_0}(x) &= \mathbb{P}(X^*_{T+2} \leq x | X_T, \ldots, X_0) \\ &= \mathbb{E}\left[\mathbb{P}\left(\hat{\epsilon}^*_{T+2} \leq \widehat{\mathcal{G}}(x, X_T, \hat{\epsilon}^*_{T+1}) \middle| \hat{\epsilon}^*_{T+1}, X_T, \ldots, X_0\right) \middle| X_T, \ldots, X_0\right] \\ &= \mathbb{E}^*\left[\widehat{F}_\epsilon\left(\widehat{\mathcal{G}}(x, X_T, \hat{\epsilon}^*_{T+1})\right)\right], \end{aligned} \tag{A2}$$

where $\widehat{\mathcal{G}}(x, X_T, \hat{\epsilon}^*_{T+1})$ represents $\frac{x - \widehat{m}_h(\widehat{m}_h(X_T) + \widehat{\sigma}_h(X_T)\hat{\epsilon}^*_{T+1})}{\widehat{\sigma}_h(\widehat{m}_h(X_T) + \widehat{\sigma}_h(X_T)\hat{\epsilon}^*_{T+1})}$, 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:

$$\sup_{|x|\leq c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,X_T,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}\left[F_\epsilon(\mathcal{G}(x,X_T,\epsilon_{T+1}))\middle|X_T\right]\right|\xrightarrow{p}0. \tag{A3}$$

However, it is hard to analyze Equation (A3) since there is a random variable X_T inside $\mathbb{E}^*(\cdot)$ and $\mathbb{E}(\cdot)$. Thus, we consider two regions of X_T, i.e., (1) $|X_T|>\gamma_T$ and (2) $|X_T|\leq\gamma_T$, where γ_T is an appropriate sequence that converges to infinity. Under A1, A2, and A5, by Lemma 1 of [30], we have:

$$\mathbb{P}(|X_T|>\gamma_T)\to 0. \tag{A4}$$

In addition, we have the relationship:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{|x|\leq c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,X_T,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}\left[F_\epsilon(\mathcal{G}(x,X_T,\epsilon_{T+1}))\middle|X_T\right]\right|>\varepsilon\right)\\
&\leq\mathbb{P}((|X_T|>\gamma_T))\\
&+\mathbb{P}\left((|X_T|\leq\gamma_T)\bigcap\left(\sup_{|x|\leq c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,X_T,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}\left[F_\epsilon(\mathcal{G}(x,X_T,\epsilon_{T+1}))\middle|X_T\right]\right|>\varepsilon\right)\right).
\end{aligned} \tag{A5}$$

Thus, to verify Equation (A3), we just need to show that the second term on the r.h.s. of Equation (A5) converges to 0. We can take the sequences 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 following expression:

$$\sup_{|x|\leq c_T,|y|\leq c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}[F_\epsilon(\mathcal{G}(x,y,\epsilon_{T+1}))]\right|>\varepsilon. \tag{A6}$$

We decompose the l.h.s. of Equation (A6) into:

$$\begin{aligned}
&\sup_{|x|\leq c_T,|y|\leq c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}[F_\epsilon(\mathcal{G}(x,y,\epsilon_{T+1}))]\right|\\
&=\sup_{|x|\leq c_T,|y|\leq c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}^*\left[F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]\right.\\
&\left.+\mathbb{E}^*\left[F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}[F_\epsilon(\mathcal{G}(x,y,\epsilon_{T+1}))]\right|\\
&\leq\sup_{|x|\leq c_T,|y|\leq c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}^*\left[F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]\right|\\
&+\sup_{|x|\leq c_T,|y|\leq c_T}\left|\mathbb{E}^*\left[F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}[F_\epsilon(\mathcal{G}(x,y,\epsilon_{T+1}))]\right|.
\end{aligned} \tag{A7}$$

Then, we analyze the two terms on the r.h.s. of Equation (A7) separately. For the first term, we have:

$$\begin{aligned}
&\sup_{|x|\leq c_T,|y|\leq c_T}\left|\mathbb{E}^*\left[\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}^*\left[F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]\right|\\
&\leq\sup_{|x|\leq c_T,|y|\leq c_T}\mathbb{E}^*\left|\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))-F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right|\\
&\leq\sup_{|x|\leq c_T,|y|\leq c_T,z}\left|\widehat{F}_\epsilon(\widehat{\mathcal{G}}(x,y,z))-F_\epsilon(\widehat{\mathcal{G}}(x,y,z))\right|\xrightarrow{p}0,\ \text{under Equation (14)}.
\end{aligned} \tag{A8}$$

For the second term on the r.h.s. of Equation (A7), we have:

$$
\begin{aligned}
&\sup_{|x|\le c_T,|y|\le c_T}\left|\mathbb{E}^*\left[F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}^*_{T+1}))\right]-\mathbb{E}[F_\epsilon(\mathcal{G}(x,y,\epsilon_{T+1}))]\right| \\
&=\sup_{|x|\le c_T,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_i))-\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathcal{G}(x,y,\epsilon_i))\right. \\
&\left.+\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathcal{G}(x,y,\epsilon_i))-\mathbb{E}[F_\epsilon(\mathcal{G}(x,y,\epsilon_{T+1}))]\right| \qquad \text{(A9)} \\
&\le\sup_{|x|\le c_T,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_i))-\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathcal{G}(x,y,\epsilon_i))\right| \\
&+\sup_{|x|\le c_T,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathcal{G}(x,y,\epsilon_i))-\mathbb{E}[F_\epsilon(\mathcal{G}(x,y,\epsilon_{T+1}))]\right|,
\end{aligned}
$$

where $\{\epsilon_i\}_{i=1}^T$ is taken as $(X_i - m(X_{i-1}))/\sigma(X_{i-1})$ for $i = 1,\dots,T$. And $\{\hat{\epsilon}_i\}_{i=1}^T$ is computed by $(X_i - \widehat{m}(X_{i-1}))/\widehat{\sigma}(X_{i-1})$ for $i = 1,\dots,T$. We can show:

$$
\begin{aligned}
&\mathbb{P}\left(\max_{i=1,\dots,T}\left|\epsilon_i-\hat{\epsilon}_i\right|>\varepsilon\right) \\
&=\mathbb{P}\left(\max_{i=1,\dots,T}\left|\frac{X_i-m(X_{i-1})}{\sigma(X_{i-1})}-\frac{X_i-\widehat{m}(X_{i-1})}{\widehat{\sigma}(X_{i-1})}\right|>\varepsilon\right) \\
&\le\mathbb{P}\left(\left(\max_{i=1,\dots,T}|X_i|>c_T\right)\bigcup\left(\max_{i=1,\dots,T}|X_{i-1}|>c_T\right)\right) \\
&+\mathbb{P}\left(\left(\max_{i=1,\dots,T}|X_i|<c_T\right)\bigcap\left(\max_{i=1,\dots,T}|X_{i-1}|<c_T\right)\right. \qquad \text{(A10)} \\
&\left.\bigcap\left(\max_{i=1,\dots,T}\left|\frac{X_i-m(X_{i-1})}{\sigma(X_{i-1})}-\frac{X_i-\widehat{m}(X_{i-1})}{\widehat{\sigma}(X_{i-1})}\right|>\varepsilon\right)\right) \\
&\le o(1)+\mathbb{P}\left(\sup_{|x|,|y|\le c_T}\left|\frac{x-m(y)}{\sigma(y)}-\frac{x-\widehat{m}(y)}{\widehat{\sigma}(y)}\right|>\varepsilon\right) \\
&\to 0.
\end{aligned}
$$

We further consider the two terms on the r.h.s. of Equation (A9) separately. For the first term, by applying Taylor expansion, we have:

$$
\begin{aligned}
&\sup_{|x|\le c_T,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_i))-\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathcal{G}(x,y,\epsilon_i))\right| \\
&=\sup_{|x|\le c_T,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}\Big(F_\epsilon(\mathcal{G}(x,y,\epsilon_i))+f_\epsilon(o_i)(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_i)-\mathcal{G}(x,y,\epsilon_i))\Big)-\frac{1}{T}\sum_{i=1}^{T}F_\epsilon(\mathcal{G}(x,y,\epsilon_i))\right| \\
&=\sup_{|x|\le c_T,|y|\le c_T}\left|\frac{1}{T}\sum_{i=1}^{T}f_\epsilon(o_i)(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_i)-\mathcal{G}(x,y,\epsilon_i))\right| \\
&\le\sup_{|x|\le c_T,|y|\le c_T}\frac{1}{T}\sum_{i=1}^{T}\left|f_\epsilon(o_i)(\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_i)-\mathcal{G}(x,y,\epsilon_i))\right| \qquad \text{(A11)} \\
&\le\sup_{|x|\le c_T,|y|\le c_T}\sup_z|f_\epsilon(z)|\cdot\frac{1}{T}\sum_{i=1}^{T}\left|\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_i)-\mathcal{G}(x,y,\epsilon_i)\right| \\
&\le\sup_{|x|\le c_T,|y|\le c_T}C\cdot\frac{1}{T}\sum_{i=1}^{T}\left|\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_i)-\mathcal{G}(x,y,\epsilon_i)\right| \text{ (under A7)} \\
&\le\sup_{|x|\le c_T,|y|\le c_T,j\in\{1,\dots,T\}}C\cdot\left|\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_j)-\mathcal{G}(x,y,\epsilon_j)\right|.
\end{aligned}
$$

From Equation (A10) and Lemma 1, we verify that Equation (A11) converges to 0 in probability. For the second term on the r.h.s. of Equation (A9), by the uniform law of large numbers, we have:

$$\sup_{|x|\leq c_T, |y|\leq c_T} \left| \frac{1}{T}\sum_{i=1}^{T} F_\epsilon(\mathcal{G}(x,y,\epsilon_i)) - \mathbb{E}[F_\epsilon(\mathcal{G}(x,y,\epsilon_{T+1}))] \right| \xrightarrow{p} 0. \tag{A12}$$

Combining all the pieces, Equation (A6) converges to 0 in probability, which implies Theorem 1. □

Proof of Theorem 2. We want to show:

$$\sup_{|x|\leq c_T} \left| F_{X^*_{T+k}-\widehat{X}^*_{T+k}|X_T,\ldots,X_0}(x) - F_{X_{T+k}-\widehat{X}_{T+k}|X_T,\ldots,X_0}(x) \right| \xrightarrow{p} 0, \text{ for } k \geq 1, \tag{A13}$$

where $X^*_{T+k} - \widehat{X}^*_{T+k}$ and $X_{T+k} - \widehat{X}_{T+k}$ are predictive roots. We still present the proof for the two-step prediction. The proof for higher-step predictions can be shown similarly. When we are dealing with two-step-ahead predictions, predictive roots have the same expression as in Equations (20) and (21). Thus, we want to measure the asymptotic distance between the following two quantities:

$$\begin{aligned} &\mathbb{P}\left(m(m(X_T) + \epsilon_{T+1}) + \epsilon_{T+2} - \frac{1}{M}\sum_{j=1}^{M} \widehat{m}_h(\widehat{m}_h(X_T) + \hat{\epsilon}_{j,T+1}) \leq x \middle| X_T, \ldots, X_0 \right); \\ &\mathbb{P}\left(\widehat{m}_h(\widehat{m}_h(X_T) + \hat{\epsilon}^*_{T+1}) + \hat{\epsilon}^*_{T+2} - \frac{1}{M}\sum_{j=1}^{M} \widehat{m}^*_h\left(\widehat{m}^*_h(X_T) + \hat{\epsilon}^*_{j,T+1}\right) \leq x \middle| X_T, \ldots, X_0 \right). \end{aligned} \tag{A14}$$

Compared to Equations (A1) and (A2), Equations (20) and (21) just have two more terms, $\frac{1}{M}\sum_{j=1}^{M} \widehat{m}_h(\widehat{m}_h(X_T) + \hat{\epsilon}_{j,T+1})$ and $\frac{1}{M}\sum_{j=1}^{M} \widehat{m}^*_h\left(\widehat{m}^*_h(X_T) + \hat{\epsilon}^*_{j,T+1}\right)$, in the predictive root in the real and bootstrap worlds, respectively. By the LLN, these two terms converge to their corresponding means in the real or bootstrap world. Based on the consistency between $\widehat{m}_h(\cdot)$ and $\widehat{m}^*_h(\cdot)$, we can show Theorem 2 similarly with the procedure used to prove Theorem 1. □

Proof of Theorem 3. The proof is based on $\{X_0, \ldots, X_T\} \in \Omega_T$. We need to verify Equation (24); i.e., we can build confidence bounds for the k-step-ahead estimation by the bootstrap. Still, we focus on the two-step-ahead prediction; i.e., we want to show:

$$\begin{aligned} \sup_{|x|\leq c_T} \Big| &\mathbb{P}\left(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(X_T)) - m(m(X_T)) \leq x \right) - \\ &\mathbb{P}\left(\sqrt{Th}(\widehat{m}^*_h(\widehat{m}^*_h(X_T)) - \widehat{m}_h(\widehat{m}_h(X_T))) \leq x \right) \Big| \xrightarrow{p} 0. \end{aligned} \tag{A15}$$

Applying the property $\mathbb{P}(|X_T| > c_T) \to 0$ again, it is enough to show:

$$\begin{aligned} \sup_{|x|,|y|\leq c_T} \Big| &\mathbb{P}\left(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y)) - m(m(y))) \leq x \right) - \\ &\mathbb{P}\left(\sqrt{Th}(\widehat{m}^*_h(\widehat{m}^*_h(y)) - \widehat{m}_h(\widehat{m}_h(y))) \leq x \right) \Big| \xrightarrow{p} 0. \end{aligned} \tag{A16}$$

To handle the uniform convergence on y, we make a ε-covering of X_T. Let the ε-covering number of $[-c_T, c_T]$ be $C_N = N(\varepsilon; [-c_T, c_T]; |\cdot|)$, which means that for every $y \in [-c_T, c_T]$, $\exists\, i \in \{1, 2, \ldots, C_N\}$ s.t. $|y - y^i| \leq \varepsilon$ for $\forall \varepsilon > 0$. Defining $y_0 \in \{y^1, \ldots, y^{C_N}\}$, we can consider:

$$
\begin{aligned}
&\sup_{|x|,|y|\le c_T}\Big|\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y))-m(m(y))\le x\Big)-\\
&\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h^*(\widehat{m}_h^*(y))-\widehat{m}_h(\widehat{m}_h(y)))\le x\Big)\Big|\le\\
&\sup_{|x|,|y|\le c_T}\Big|\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y))-m(m(y)))\le x\Big)-\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y_0))-m(m(y_0)))\le x\Big)\Big|\\
&+\sup_{\substack{|x|\le c_T,\\ y_0\in\{y^1,\ldots,y^{C_N}\}}}\Big|\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y_0))-m(m(y_0)))\le x\Big)\\
&-\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h^*(\widehat{m}_h^*(y_0))-\widehat{m}_h(\widehat{m}_h(y_0)))\le x\Big)\Big|\\
&+\sup_{|x|,|y|\le c_T}\Big|\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h^*(\widehat{m}_h^*(y_0))-\widehat{m}_h(\widehat{m}_h(y_0)))\le x\Big)-\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h^*(\widehat{m}_h^*(y))-\widehat{m}_h(\widehat{m}_h(y)))\le x\Big)\Big|.
\end{aligned}
\tag{A17}
$$

For the first term on the r.h.s. of Equation (A17), we have:

$$
\begin{aligned}
&\sup_{|x|,|y|\le c_T}\Big|\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y))-m(m(y)))\le x\Big)-\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y_0))-m(m(y_0)))\le x\Big)\Big|\\
&=\sup_{|x|,|y|\le c_T}\Big|\mathbb{P}(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y_0))-m(m(y_0)))+\sqrt{Th}(C_1(\widehat{m}_h(y)-\widehat{m}_h(y_0))\\
&+C_2(m(y_0)-m(y)))\le x)-\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y_0))-m(m(y_0)))\le x\Big)\Big|.
\end{aligned}
\tag{A18}
$$

where C_1 and C_2 are some finite constants since the derivatives of $\widehat{m}_h(\cdot)$ and $m(\cdot)$ are bounded. Considering the first term inside the absolute bracket on the r.h.s. of Equation (A18), we can think of this as a convolution of two random variables:

$$
\begin{aligned}
&\mathbb{P}\Big(\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y_0))-m(m(y_0)))+\sqrt{Th}(C_1(\widehat{m}_h(y)-\widehat{m}_h(y_0))+C_2(m(y_0)-m(y)))\le x\Big)\\
&=\mathbb{P}(X+Z\le x).
\end{aligned}
\tag{A19}
$$

Further, based on the smoothing property of $\widehat{m}_h(\cdot)$ and $m(\cdot)$, again, we can take ε to be small enough to make the random variable Z close to being degenerated, i.e., $\mathbb{P}(Z=0)=1-\mathbb{P}(Z\in A)=1-o(1)$; A is a small set around 0 without containing 0. Thus, Equation (A18) can be written as:

$$
\begin{aligned}
&\sup_{|x|,|y|\le c_T}|\mathbb{P}(X+Z\le x)-\mathbb{P}(X\le x)|\\
&=\sup_{|x|,|y|\le c_T}|\mathbb{P}(X+0\le x,Z=0)+\mathbb{P}(X+Z\le x,Z\in A)-\mathbb{P}(X\le x)|\\
&\le\sup_{|x|,|y|\le c_T}|\mathbb{P}(X\le x)+o(1)+o(1)-\mathbb{P}(X\le x)|\\
&=o(1).
\end{aligned}
\tag{A20}
$$

Similarly, the last term on the r.h.s. of Equation (A17) can also be made to converge to 0. We can then focus on analyzing the middle term. In other words, it is enough to analyze the pointwise convergence property between distributions in the real and bootstrap worlds. According to the idea for estimating the distribution of nonparametric estimation by the bootstrap in the work of [27], we decompose $\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y_0))-m(m(y_0)))$ into bias-type and variance terms:

$$\begin{aligned}
&\sqrt{Th}(\widehat{m}_h(\widehat{m}_h(y_0)) - m(m(y_0))) \\
&= \sqrt{Th}\left(\frac{\sum_{t=0}^{T-1} K_h(\widehat{m}_h(y_0) - X_t)X_{t+1}}{T\hat{f}_h(\widehat{m}_h(y_0))} - \frac{\sum_{t=0}^{T-1} K_h(\widehat{m}_h(y_0) - X_t)\cdot m(m(y_0))}{T\hat{f}_h(\widehat{m}_h(y_0))}\right) \\
&= \sqrt{Th}\left(\frac{\hat{r}_{V,h}(\widehat{m}_h(y_0))}{\hat{f}_h(\widehat{m}_h(y_0))} + \frac{\hat{r}_{B,h}(\widehat{m}_h(y_0))}{\hat{f}_h(\widehat{m}_h(y_0))}\right),
\end{aligned} \tag{A21}$$

where

$$\begin{aligned}
\hat{r}_{V,h}(\widehat{m}_h(y_0)) &= \frac{1}{T}\sum_{t=0}^{T-1} K_h(\widehat{m}_h(y_0) - X_t)\epsilon_{t+1}; \\
\hat{r}_{B,h}(\widehat{m}_h(y_0)) &= \frac{1}{T}\sum_{t=0}^{T-1} K_h(\widehat{m}_h(y_0) - X_t)(m(X_t) - m(m(y_0))),
\end{aligned} \tag{A22}$$

where $K_h(\cdot)$ represents the function in the form $\frac{1}{h}K(\cdot/h)$. By also carrying this process out for $\sqrt{Th}(\widehat{m}_h^*(\widehat{m}_h^*(y_0)) - \widehat{m}_h(\widehat{m}_h(y_0)))$, we can obtain:

$$\sqrt{Th}(\widehat{m}_h^*(\widehat{m}_h^*(y_0)) - \widehat{m}_h(\widehat{m}_h(y_0))) = \sqrt{Th}\left(\frac{\hat{r}_{V,h}^*(\widehat{m}_h^*(y_0))}{\hat{f}_h^*(\widehat{m}_h^*(y_0))} + \frac{\hat{r}_{B,h}^*(\widehat{m}_h^*(y_0))}{\hat{f}_h^*(\widehat{m}_h^*(y_0))}\right), \tag{A23}$$

where

$$\begin{aligned}
\hat{r}_{V,h}^*(\widehat{m}_h^*(y_0)) &= \frac{1}{T}\sum_{t=0}^{T-1} K_h(\widehat{m}_h^*(y_0) - X_t^*)\hat{\epsilon}_{t+1}^*; \\
\hat{r}_{B,h}^*(\widehat{m}_h^*(y_0)) &= \frac{1}{T}\sum_{t=0}^{T-1} K_h(\widehat{m}_h^*(y_0) - X_t^*)(\widehat{m}_h(X_t^*) - \widehat{m}_h(\widehat{m}_h(y_0))).
\end{aligned} \tag{A24}$$

For the variance term, by Lemma 4.4 of [27], we have:

$$\begin{aligned}
\sup_x \left|\mathbb{P}(\sqrt{Th}\hat{r}_{V,h}(x_0) \le x) - \mathbb{P}(Z(x_0) \le x)\right| &= o(1); \\
\sup_x \left|\mathbb{P}(\sqrt{Th}\hat{r}_{V,h}^*(x_0) \le x) - \mathbb{P}(Z(x_0) \le x)\right| &= op(1),
\end{aligned} \tag{A25}$$

where $Z(x_0)$ has the distribution $N(0, \tau^2(x_0))$; $\tau^2(x_0) = f_X(x_0)\int K^2(v)dv$; $x_0 \in \mathbb{R}$. Since $\widehat{m}_h(y_0)$ and $\widehat{m}_h^*(y_0)$ both converge to $m(y_0)$ in probability and the target distribution is continuous, by the continuous mapping theorem, we can obtain the uniform convergence between the distributions of $\sqrt{Th}\hat{r}_{V,h}(m(y_0))$ and $\sqrt{Th}\hat{r}_{V,h}(\widehat{m}_h(y_0))$, i.e.:

$$\sup_x \left|\mathbb{P}(\sqrt{Th}\hat{r}_{V,h}(\widehat{m}_h(y_0)) \le x) - \mathbb{P}(\sqrt{Th}\hat{r}_{V,h}(m(y_0)) \le x)\right| = o(1). \tag{A26}$$

To show the uniform convergence relationship between $\sqrt{Th}\hat{r}_{V,h}^*(m(y_0))$ and $\sqrt{Th}\hat{r}_{V,h}^*(\widehat{m}_h^*(y_0))$, we need the continuous $\sqrt{Th}\hat{r}_{V,h}^*(m(y_0))$, which is a convolution of *i.i.d.* random variables, $\{\hat{\epsilon}_i^*\}_{i=1}^T \sim \widehat{F}_\epsilon$. Unfortunately, $\widehat{F}_\epsilon$ is the empirical distribution of residuals and is discrete. To make the analysis more convenient, we take a convolution approach to smooth the distribution of empirical residuals; i.e., we define another random variable, which is the sum of $\hat{\epsilon}$ and a standard normal random variable ξ:

$$\tilde{\epsilon} = \hat{\epsilon} + \xi, \tag{A27}$$

where $\xi \sim N(0, \mathcal{L}(T))$ and $\mathcal{L}(T) \to 0$ at an appropriate rate. It is easy to show that the distribution of $\tilde{\epsilon}$, $\widetilde{F}_\epsilon$ is asymptotically equivalent to $\widehat{F}_\epsilon$; i.e., Equation (14) is also satisfied for $\widetilde{F}_\epsilon$. In practice, we can take $\mathcal{L}(T)$ to be small enough, and then we still bootstrap time

series based on $\widehat{F}_\epsilon$ in practice. However, from a theoretical view, we would like to take $\widetilde{F}_\epsilon$. To simplify the notation, we use $\widehat{F}_\epsilon$ throughout this paper, and its representation changes according to the context.

Combining all the pieces, we can obtain:

$$\begin{aligned}\sup_x\left|\mathbb{P}(\sqrt{Th}\hat{r}_{V,h}(\widehat{m}_h(y_0)) \le x) - \mathbb{P}(Z(m(y_0)) \le x)\right| &= op(1);\\ \sup_x\left|\mathbb{P}(\sqrt{Th}\hat{r}^*_{V,h}(\widehat{m}^*_h(y_0)) \le x) - \mathbb{P}(Z(m(y_0)) \le x)\right| &= op(1).\end{aligned} \tag{A28}$$

Then, the bias-type term in the real and bootstrap worlds remains to be analyzed. We first consider the bias-type term $\hat{r}_{B,h}(\widehat{m}_h(y_0))$:

$$\begin{aligned}&\sqrt{Th}\hat{r}_{B,h}(\widehat{m}_h(y_0))\\ &= \sqrt{\frac{h}{T}}\sum_{t=0}^{T-1} K_h(\widehat{m}_h(y_0) - X_t)\cdot(m(X_t) - m(m(y_0)))\\ &= \sqrt{\frac{h}{T}}\sum_{t=0}^{T-1}\left[K_h(m(y_0) - X_t) + K_h^{(1)}(\hat{x})\cdot(\widehat{m}_h(y_0) - m(y_0))\right]\cdot(m(X_t) - m(m(y_0)))\\ &= \sqrt{\frac{h}{T}}\sum_{t=0}^{T-1} K_h(m(y_0) - X_t)\cdot(m(X_t) - m(m(y_0)))\\ &+ \sqrt{\frac{h}{T}}\sum_{t=0}^{T-1} K_h^{(1)}(\hat{x})\cdot(\widehat{m}_h(y_0) - m(y_0))\cdot(m(X_t) - m(m(y_0))).\end{aligned} \tag{A29}$$

For the first term on the r.h.s. of Equation (A29), based on the ergodicity of $\{X_t\}$ series, we can find that the mean of this term is:

$$\begin{aligned}&\mathbb{E}\left[\sqrt{\frac{h}{T}}\sum_{t=0}^{T-1} K_h(m(y_0) - X_t)\cdot(m(X_t) - m(m(y_0)))\right]\\ &= \mathbb{E}\left[\sqrt{Th}\mathbb{E}[K_h(m(y_0) - X_1)\cdot(m(X_1) - m(m(y_0)))|X_0]\right]\\ &= \mathbb{E}\left[\sqrt{Th}\int K(v)\cdot(m(vh + m(y_0)) - m(m(y_0)))\cdot f_\epsilon(vh + m(y_0) - m(X_0))dv\right]\\ &= \mathbb{E}\left[\sqrt{Th}\int K(v)\cdot\left(m^{(1)}(m(y_0))vh + m^{(2)}(\hat{y})\cdot v^2h^2\right)\cdot\left(f_\epsilon(m(y_0) - m(X_0)) + f_\epsilon^{(1)}(\hat{x})\cdot vh\right)dv\right].\end{aligned} \tag{A30}$$

If we take the bandwidth satisfying $Th^5 \to 0$, Equation (A30) converges to 0. Then, we consider the mean of the second term on the r.h.s. of Equation (A29):

$$\begin{aligned}&\mathbb{E}\left[\sqrt{\frac{h}{T}}\sum_{t=0}^{T-1} K_h^{(1)}(\hat{x})\cdot(\widehat{m}_h(y_0) - m(y_0))\cdot(m(X_t) - m(m(y_0)))\right]\\ &= \sqrt{\frac{h}{T}}\sum_{t=0}^{T-1}\mathbb{E}\left[K_h^{(1)}(\hat{x})\cdot(\widehat{m}_h(y_0) - m(y_0))\cdot(m(X_t) - m(m(y_0)))\right]\\ &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathbb{E}\left[\sqrt{Th}\cdot K_h^{(1)}(\hat{x})\cdot(\widehat{m}_h(y_0) - m(y_0))\cdot(m(X_t) - m(m(y_0)))\middle| X_t\right]\right].\end{aligned} \tag{A31}$$

Since $\mathbb{E}(\sqrt{Th}\cdot(\widehat{m}_h(y_0) - m(y_0))$ is $O(\sqrt{Th^5})$ (see Lemma 4.6 of [27] for the proof), under the assumption that $K(\cdot)$ has a bounded derivative and $m(\cdot)$ is bounded in a compact set, we have $\mathbb{E}(\mathbb{E}(\sqrt{Th}\cdot K_h^{(1)}(\hat{x})\cdot(\widehat{m}_h(y_0) - m(y_0))\cdot(m(X_t) - m(m(y_0)))|X_t))$ equal to $O(\sqrt{Th^5})$; once we select the under-smoothing bandwidth that satisfies $Th^5 \to 0$, Equation (A31) converges to 0. Then, we need to analyze the variance of $\sqrt{Th}\hat{r}_{B,h}(\widehat{m}_h(y_0))$. Similarly, we can show that it is $op(1)$. All in all, $\sqrt{Th}\hat{r}_{B,h}(\widehat{m}_h(y_0))$ converges to 0 in probability.

For the bias-type term $\hat{r}^*_{B,h}(\widehat{m}^*_h(y_0))$ in the bootstrap world, we can perform a similar decomposition to the one applied in Equation (A29), and then we can obtain:

$$\begin{aligned}
&\sqrt{Th}\hat{r}^*_{B,h}(\widehat{m}^*_h(y_0)) \\
&= \sqrt{\frac{h}{T}}\sum_{t=0}^{T-1} K_h(\widehat{m}_h(y_0) - X^*_t)\cdot(\widehat{m}_h(X^*_t) - \widehat{m}_h(\widehat{m}_h(y_0))) \\
&+ \sqrt{\frac{h}{T}}\sum_{t=0}^{T-1} K^{(1)}_h(\hat{x})\cdot(\widehat{m}^*_h(y_0) - \widehat{m}_h(y_0))\cdot(\widehat{m}_h(X^*_t) - \widehat{m}_h(\widehat{m}_h(y_0))).
\end{aligned} \tag{A32}$$

We first rely on the fact that $\mathbb{E}^*(\mathbb{E}^*(\sqrt{Th}\cdot K^{(1)}_h(\hat{x})\cdot(\widehat{m}^*_h(y_0) - \widehat{m}_h(y_0))\cdot(\widehat{m}_h(X^*_t) - \widehat{m}_h(\widehat{m}_h(y_0)))\,|X^*_t))$ is also $O(\sqrt{Th^5})$; see Lemma 4.6 of [27] for more details. Thus, using the under-smoothing bandwidth strategy, the second term on the r.h.s. of Equation (A32) also converges to 0. For the first term, we can rely on the fact that the bootstrap series is also ergodic with high probability; see Theorem 2 of [30,32] for a time-series model with homoscedastic or heteroscedastic errors, respectively. Thus, with a similar analysis of the variant in the real world, we can see that the bias-type term in the bootstrap world also converges to 0 in probability. Given the consistent relationship between $\hat{f}_h(\widehat{m}_h(y_0))$ and $\hat{f}^*_h(\widehat{m}_h(y_0))$, which is implied by Lemma 4.5 of [27], Equation (A15) follows from the analysis of variance and bias-type terms in the real and bootstrap worlds. □

Appendix B. The Advantage of Applying Under-Smoothing Bandwidth for QPI with Finite Sample

The proof of Theorem 1 provides the *big picture* of the asymptotic validity of the QPI. Although the choice of the bandwidth does not influence the asymptotic validity of the QPI, we can find that the QPI with the under-smoothing bandwidth has a better CVR for multi-step-ahead predictions from the simulation results. We attempt to analyze this phenomenon informally. Starting from the convergence result, we want to show:

$$\sup_{|x|\le c_T}\left|F_{X^*_{T+k}|X_T,\ldots,X_0}(x) - F_{X_{T+k}|X_T}(x)\right| \xrightarrow{p} 0, \text{ for } k \ge 1. \tag{A33}$$

We still take the case with $k = 2$ as an example. From the analyses in the proof of Theorem 1, we can obtain:

$$\sup_{|x|\le c_T}\left|F_{X^*_{T+2}|X_T,\ldots,X_0}(x) - F_{X_{T+2}|X_T}(x)\right| \le op(1) + \sup_{|x|\le c_T, |y|\le c_T, j\in\{1,\ldots,T\}} C\cdot\left|\widehat{\mathcal{G}}(x,y,\hat{\epsilon}_j) - \mathcal{G}(x,y,\epsilon_j)\right|. \tag{A34}$$

Recall that $\mathcal{G}(x, X_T, \epsilon_{T+1})$ represents $\frac{x - m(m(X_T)+\sigma(X_T)\epsilon_{T+1})}{\sigma(m(X_T)+\sigma(X_T)\epsilon_{T+1})}$, and $\widehat{\mathcal{G}}(x, X_T, \hat{\epsilon}^*_{T+1})$ represents $\frac{x - \widehat{m}_h(\widehat{m}_h(X_T)+\widehat{\sigma}_h(X_T)\hat{\epsilon}^*_{T+1})}{\widehat{\sigma}_h(\widehat{m}_h(X_T)+\widehat{\sigma}_h(X_T)\hat{\epsilon}^*_{T+1})}$. To simplify the notation, we consider the model when $\sigma(x) \equiv 1$. Then, Equation (A34) becomes:

$$\begin{aligned}
&\sup_{|x|\le c_T}\left|F_{X^*_{T+k}|X_T,\ldots,X_0}(x) - F_{X_{T+k}|X_T}(x)\right| \le op(1) \\
&+ \sup_{|y|\le c_T, j\in\{1,\ldots,T\}} C\cdot\left|\widehat{m}_h(\widehat{m}_h(y) + \hat{\epsilon}^*_j) - m(m(y) + \epsilon_j)\right|.
\end{aligned} \tag{A35}$$

Then, we can focus on analyzing $\widehat{m}_h(\widehat{m}_h(X_T) + \hat{\epsilon}^*_j) - m(m(X_T) + \epsilon_j)$. By applying Taylor expansion, we can obtain:

$$\begin{aligned}
&\widehat{m}_h(\widehat{m}_h(y)+\hat{\epsilon}_j^*)-m(m(y)+\epsilon_j)\\
&=\widehat{m}_h(m(y)+\epsilon_j)-m(m(y)+\epsilon_j)\\
&+\widehat{m}_h^{(1)}(\hat{\hat{x}})(\widehat{m}_h(y)+\hat{\epsilon}_j^*-m(y)-\epsilon_j).
\end{aligned} \tag{A36}$$

For the first term on the r.h.s. of Equation (A36), based on the ergodicity, asymptotically, we have:

$$\begin{aligned}
&\widehat{m}_h(m(y)+\epsilon_j)=\frac{\frac{1}{Th}\sum_{i=0}^{T-1}K\Big(\frac{m(y)+\epsilon_j-X_i}{h}\Big)X_{i+1}}{\widehat{f}_h(m(y)+\epsilon_j)}\\
&\frac{\frac{1}{Th}\sum_{i=0}^{T-1}K\Big(\frac{m(y)+\epsilon_j-X_i}{h}\Big)(m(X_i)+\epsilon_{i+1})}{\widehat{f}_h(m(y)+\epsilon_j)}\\
&=\frac{1}{\widehat{f}_h(m(y)+\epsilon_j)}\left(\frac{1}{h}\mathbb{E}(K\Big(\frac{m(y)+\epsilon_j-X_1}{h}\Big)m(X_1))+\frac{1}{h}\mathbb{E}(K\Big(\frac{m(y)+\epsilon_j-X_1}{h}\Big)\epsilon_1)\right)\\
&=\frac{1}{\widehat{f}_h(m(y)+\epsilon_j)}\left(\frac{1}{h}\int K\Big(\frac{u-m(y)-\epsilon_j}{h}\Big)m(u)f_X(u)du+0\right)\\
&=\frac{1}{\widehat{f}_h(m(y)+\epsilon_j)}\left(\int K(v)m(vh+m(y)+\epsilon_j)f_X(vh+m(y)+\epsilon_j)dv+0\right)\\
&=\frac{1}{\widehat{f}_h(m(y)+\epsilon_j)}\Big(\int K(v)[m(m(y)+\epsilon_j)+vhm^{(1)}(m(y)+\epsilon_j)+v^2h^2m^{(2)}(\hat{\hat{y}})]\cdot\\
&[f_X(m(y)+\epsilon_j)+vhf_X^{(1)}(m(y)+\epsilon_j)+v^2h^2f_X^{(2)}(\hat{\hat{z}})]dv\Big)\\
&=\frac{1}{\widehat{f}_h(m(y)+\epsilon_j)}\Big(m(m(y)+\epsilon_j)f_X(m(y)+\epsilon_j)+O(h^2)\Big).
\end{aligned} \tag{A37}$$

The convergence of $\widehat{f}_h(m(y)+\epsilon_j)$ to $f_X(m(y)+\epsilon_j)$ guarantees the consistency relationship of Equation (A37) and $m(m(y)+\epsilon_j)$. Similarly, for the third term on the r.h.s. of Equation (A36), we can conduct a similar analysis to find the convergence to 0 in probability. Moreover, the convergence speed is related to $O(h^2)$. When multiple-step-ahead predictions are required, we will obtain more and more such $O(h^2)$ terms. If we have large enough data, it is "safe" to focus on the bandwidth with the optimal rate to estimate the model. However, for the finite-sample cases, it is better to take an under-smoothing h, though the corresponding LEN of the prediction interval will get larger due to the mean–variance trade-off. This conclusion coincides with the results shown in Tables 4 and 5. From there, we can observe that the one-step-ahead QPI with the optimal bandwidth has a better CVR compared to the version with the under-smoothing bandwidth. In the meantime, the LEN of the PI with the optimal bandwidth is also slightly smaller. When the prediction horizon is larger than 1, although the QPI with the under-smoothing bandwidth has a slightly larger LEN, its CVR is notably better than the QPI with the optimal bandwidth. Here, we conducted more simulation studies to show that the QPIs with the optimal bandwidth and under-smoothing bandwidth are asymptotically equivalent. We performed simulations with Equation (30) and took $T+1$ to be 1000. The CVR and LEN of different QPIs are tabulated in Table 6. From the simulation results, although the LEN of the QPI with the optimal bandwidth is always less than the variant with the under-smoothing bandwidth, the difference is marginal. In addition, these two types of QPIs have indistinguishable performance according to the CVR, which implies the asymptotic equivalence of applying the optimal bandwidth or under-smoothing bandwidth. It also implies that adopting fitted or predictive residuals is also asymptotically equivalent.

Appendix C. The Effects of Applying Under-Smoothing or Over-Smoothing Bandwidth on PPI

To see the effects of applying under-smoothing or over-smoothing tricks on the performance of the PPI, we took the sample size $T + 1$ to be 50 or 500 and performed simulations 5000 times on the first model. The simulation results are shown in Table 7. These results coincide with Corollary 1: i.e., both bandwidth strategies can give one-step-ahead PPIs with satisfactory CVR even for a small sample size with predictive residuals. The implication of Theorem 3 is also verified: i.e., taking the under-smoothing bandwidth can keep the CVR at a high level for multi-step-ahead predictions when the sample size is small. In addition, as the sample size increases, the CVR of the PPI with the over-smoothing bandwidth also increases. This phenomenon is guaranteed by the asymptotically valid property of the PPI regardless of whether the over-smoothing or under-smoothing bandwidth is used; see Theorem 2.

Appendix D. The Comparison of Applying Under-Smoothing and Optimal Bandwidths on Estimating the Variance Function for Building PPI

In Appendix C, we have seen the advantage of applying the under-smoothing bandwidth to estimate the model in the real and bootstrap worlds when the model has homoscedastic errors. For the model with heteroscedastic errors, as we have mentioned in Section 3.3, we can rely on the optimal bandwidth to estimate the variance functions.

To check this claim, we consider two strategies for the bandwidth of the estimator for the variance function: (1) take the under-smoothing bandwidth as we do for the mean function estimator; (2) take the bandwidth with the optimal rate. To estimate the mean function in the bootstrap world, we keep using the under-smoothing bandwidth strategy. The simulation results based on Equation (32) with a small sample size are shown in Table 8. We can see that the LEN of the PPI with the optimal bandwidth when estimating the variance function is always smaller than that of the corresponding PPI with the under-smoothing bandwidth. At the same time, the CVRs of both types of PPI are indistinguishable for $k > 1$. For the one-step-ahead prediction, the former PPI is notably better than the latter PPI. This phenomenon is implied by Remark 7: i.e., the best strategy for the one-step-ahead PPI is choosing bandwidths with the optimal rate for both estimators of mean and variance functions.

References

1. Politis, D.N. Financial time series. *Wiley Interdiscip. Rev. Comput. Stat.* **2009**, *1*, 157–166. [CrossRef]
2. Politis, D.N. Pertinent Prediction Intervals. In *Model-Free Prediction and Regression*; Springer: New York, NY, USA, 2015; pp. 43–45.
3. Pemberton, J. Exact least squares multi-step prediction from nonlinear autoregressive models. *J. Time Ser. Anal.* **1987**, *8*, 443–448. [CrossRef]
4. Lee, K.; Billings, S. A new direct approach of computing multi-step ahead predictions for non-linear models. *Int. J. Control* **2003**, *76*, 810–822. [CrossRef]
5. Chen, R.; Yang, L.; Hafner, C. Nonparametric multistep-ahead prediction in time series analysis. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2004**, *66*, 669–686. [CrossRef]
6. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
7. Politis, D.N.; Romano, J.P. A Circular Block-Resampling Procedure for Stationary Data. In *Exploring the Limits of Bootstrap*; LePage, R., Billard, L., Eds.; Wiley: New York, NY, USA, 1992; pp. 263–270.
8. Politis, D.N.; Romano, J.P. The stationary bootstrap. *J. Am. Stat. Assoc.* **1994**, *89*, 1303–1313. [CrossRef]
9. Politis, D.N. The Impact of Bootstrap Methods on Time Series Analysis. *Stat. Sci.* **2003**, *18*, 219–230. [CrossRef]
10. Kreiss, J.P.; Paparoditis, E. *Bootstrap for Time Series: Theory and Methods*; Springer: Heidelberg, Germany, 2023.
11. Franke, J.; Neumann, M.H. Bootstrapping neural networks. *Neural Comput.* **2000**, *12*, 1929–1949. [CrossRef]
12. Michelucci, U.; Venturini, F. Estimating neural network's performance with bootstrap: A tutorial. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 357–373. [CrossRef]
13. Thombs, L.A.; Schucany, W.R. Bootstrap prediction intervals for autoregression. *J. Am. Stat. Assoc.* **1990**, *85*, 486–492. [CrossRef]
14. Pascual, L.; Romo, J.; Ruiz, E. Bootstrap predictive inference for ARIMA processes. *J. Time Ser. Anal.* **2004**, *25*, 449–465. [CrossRef]
15. Pascual, L.; Romo, J.; Ruiz, E. Bootstrap prediction for returns and volatilities in GARCH models. *Comput. Stat. Data Anal.* **2006**, *50*, 2293–2312. [CrossRef]

16. Pan, L.; Politis, D.N. Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. *J. Stat. Plan. Inference* **2016**, *177*, 1–27. [CrossRef]
17. Wu, K.; Politis, D.N. Bootstrap Prediction Inference of Non-linear Autoregressive Models. *arXiv* **2023**, arXiv:2306.04126.
18. Politis, D.N. Model-free model-fitting and predictive distributions. *Test* **2013**, *22*, 183–221. [CrossRef]
19. Manzan, S.; Zerom, D. A bootstrap-based non-parametric forecast density. *Int. J. Forecast.* **2008**, *24*, 535–550. [CrossRef]
20. Giordano, F.; La Rocca, M.; Perna, C. Forecasting nonlinear time series with neural network sieve bootstrap. *Comput. Stat. Data Anal.* **2007**, *51*, 3871–3884. [CrossRef]
21. Khosravi, A.; Nahavandi, S.; Creighton, D.; Atiya, A.F. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Trans. Neural Netw.* **2011**, *22*, 1341–1356. [CrossRef]
22. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* **2017**, 6402–6413 .
23. Chen, J.; Politis, D.N. Optimal multi-step-ahead prediction of ARCH/GARCH models and NoVaS transformation. *Econometrics* **2019**, *7*, 34. [CrossRef]
24. Wu, K.; Karmakar, S. Model-free time-aggregated predictions for econometric datasets. *Forecasting* **2021**, *3*, 920–933. [CrossRef]
25. Wu, K.; Karmakar, S. A model-free approach to do long-term volatility forecasting and its variants. *Financ. Innov.* **2023**, *9*, 59. [CrossRef]
26. Wang, Y.; Politis, D.N. Model-free Bootstrap and Conformal Prediction in Regression: Conditionality, Conjecture Testing, and Pertinent Prediction Intervals. *arXiv* **2021**, arXiv:2109.12156.
27. Franke, J.; Kreiss, J.P.; Mammen, E. Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli* **2002**, *8*, 1–37.
28. Politis, D.N. Studentization vs. Variance Stabilization: A Simple Way Out of an Old Dilemma. 2022. Available online: https://mathweb.ucsd.edu/~politis/PAPER/DGP_Aug_11.pdf (accessed on 18 July 2023).
29. Bradley, R.C. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2005**, *2*, 107–144. [CrossRef]
30. Franke, J.; Neumann, M.H.; Stockis, J.P. Bootstrapping nonparametric estimators of the volatility function. *J. Econom.* **2004**, *118*, 189–218. [CrossRef]
31. Min, C.; Hongzhi, A. The probabilistic properties of the nonlinear autoregressive model with conditional heteroskedasticity. *Acta Math. Appl. Sin.* **1999**, *15*, 9–17. [CrossRef]
32. Franke, J.; Kreiss, J.P.; Mammen, E.; Neumann, M.H. Properties of the nonparametric autoregressive bootstrap. *J. Time Ser. Anal.* **2002**, *23*, 555–585. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.