Electronic Journal of Statistics
Vol. 13 (2019) 3710–3774
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1606

Predictive, finite-sample model choice for time series under stationarity and non-stationarity

Tobias Kley[*]

University of Bristol
School of Mathematics, Faculty of Science, University Walk, Bristol BS8 1TW, UK
e-mail: tobias.kley@bristol.ac.uk

Philip Preuß[†]

Ruhr-Universität Bochum
Department of Mathematics, Institute of Statistics, 44780 Bochum, Germany
e-mail: philip.preuss@rub.de

and

Piotr Fryzlewicz[*]

London School of Economics and Political Science
Department of Statistics, Columbia House, Houghton Street, London, WC2A 2AE, UK
e-mail: p.fryzlewicz@lse.ac.uk

Abstract: In statistical research there usually exists a choice between structurally simpler or more complex models. We argue that, even if a more complex, locally stationary time series model were true, then a simple, stationary time series model may be advantageous to work with under parameter uncertainty. We present a new model choice methodology, where one of two competing approaches is chosen based on its empirical, finite-sample performance with respect to prediction, in a manner that ensures interpretability. A rigorous, theoretical analysis of the procedure is provided. As an important side result we prove, for possibly diverging model order, that the localised Yule-Walker estimator is strongly, uniformly consistent under local stationarity. An R package, **forecastSNSTS**, is provided and used to apply the methodology to financial and meteorological data in empirical examples. We further provide an extensive simulation study and discuss when it is preferable to base forecasts on the more volatile time-varying estimates and when it is advantageous to forecast as if the data were from a stationary process, even though they might not be.

MSC 2010 subject classifications: Primary 62M20; secondary 62M10.
Keywords and phrases: Forecasting, Yule-Walker estimate, local stationarity, covariance stationarity.

Received January 2019.

[*]Supported by the Engineering and Physical Sciences Research Council grant no. EP/L014246/1.
[†]Supported by the Sonderforschungsbereich "Statistical modelling of nonlinear dynamic processes" (SFB 823, Teilprojekt C1) of the Deutsche Forschungsgemeinschaft.

Contents

1. Introduction

A well-trodden path in applied statistical research is to propose a model believed to be a good approximation to the data-generating process, and then to

estimate the model parameters with a view to performing a specific task, for example, prediction. However, even if the analyst were 'lucky' and chose the right model family, thereby reducing modelling bias, the resulting parameter estimators could be so variable that the selected model might well be sub-optimal from the point of view of the task in question. Choosing a slightly wrong model but with less variable parameter estimates may well lead to superior performance in, for example, prediction. This effect is usually referred to as the bias-variance trade-off and it has frequently been discussed in the literature. In this paper we explore how this unsurprising but interesting phenomenon could and should affect model choice in the analysis of non-stationary time series.

Choosing between stationary and non-stationary modelling is, typically, an important step in the analysis of time series data. Stationarity, which assumes that certain probabilistic properties of the time series model do not evolve over time, is a key assumption in time series analysis, and several excellent monographs focus on stationary modelling; see, e.g., [10], [11] or [47]. However, in practice, many time series are deemed to be better-suited for non-stationary modelling; this judgement can be based on diverse factors, such as, for example, visual inspection, formal tests against stationarity, or the observation that the data have been collected in a time-evolving environment and therefore are unlikely to have come from a stationary model.

Early contributions to the literature of non-stationary time series are [56], where the tvAR model was introduced, and [27], who defined the tvARMA model. A general non-stationary time series framework was provided by [46], who defined the evolutionary spectrum. A now particularly popular framework for the rigorous description of non-stationary time series models is that of local stationarity, in which the data are modelled locally as approximately stationary [18, 19]. We now illustrate the main idea of the paper using a simple example of a locally stationary time series model, the time-varying autoregressive model (of order 1)

$$X_{t,T} = a(t/T)X_{t-1,T} + Z_t, \quad t = 1, ..., T,$$

with T denoting the sample size, $a : [0, 1] \to (-1, 1)$ being some suitable function and Z_t being an i.i.d. sequence with mean zero and variance one. Typically, to forecast future observations, one would require an estimate of $a(1)$, see e.g. [14]. Before constructing a suitable estimator, some analysts would wish to test if a was indeed time-varying, and there exist a vast amount of techniques to validate the assumption of a constant second-order structure in this framework; see [59], [40], [23], [41], [22], [38], [45] or [58]. If the process was found to be non-stationary, it would be tempting to estimate $a(1)$ by a localised estimate based on the most recent observations of $X_{t,T}$. This localisation would most likely reduce the bias of the estimator if the true dependency structure was indeed time-varying, but also increase its variance. However, if, for example, the function a was varying only slowly over time, this estimation procedure might result in sub-optimal estimation from the point of view of the mean squared prediction error, yielding inferior forecasts compared to the classical stationary AR(1) model. This would be particularly likely if the test of stationarity em-

ployed at the start was not constructed with the same performance measure in mind (i.e., mean squared prediction error) and was therefore 'detached' from the task in question (i.e., prediction). One of the findings of this paper is that even if the function a varied over time, one should in some cases treat it as constant in order to obtain smaller prediction errors, or in other words, 'prefer the wrong model' from the point of view of prediction.

The main aim of this paper is to propose an alternative model choice methodology in time series analysis that avoids the pitfalls of the above-mentioned process of testing followed by model choice. More precisely, our work has the following objectives:

- To propose a generic procedure for finite-sample model choice which avoids the path of hypothesis testing but instead chooses the model that offers better empirical finite-sample performance in terms of prediction on a validation set, with associated performance guarantees for the test set of yet unobserved data. Although the procedure is proposed and analysed theoretically in the framework of choice between stationarity and local stationarity and in the context of prediction, the procedure is applicable more generally whenever a decision needs to be made between two competing approaches, and can therefore be viewed as model- and problem-free. At the end of Section 3.2, we provide two examples of other situations in which the general principle of our procedure can be applied.
- To suggest 'rules of thumb' indicating when the (wrong) stationary model may be preferred in a time-varying, locally stationary situation from the point of view of forecasting; and when a time-varying model should be preferred.

Our procedure validates and puts on a solid footing the possibly counter-intuitive observation that it is sometimes beneficial to choose the 'wrong' (but possibly simpler) model in time series analysis, if that model relies on more reliable estimators of its parameters than the right (but possibly more complex) model. While we stop short of conveying the message that simplicity in time series should always be preferred, part of our aim is to draw time series analysts' attention to the fact that particularly complex time series models may well appear attractive on first glance as they have the potential to capture features of the data well, but on the other hand can be so hard to estimate that this makes them inferior to simple and easy-to-estimate alternative models, even if the latter are wrong.

We now briefly describe related recent literature. The work of [60], who, while discussing time series prediction, select the model based on the minimisation of up to m-step ahead prediction errors (rather than the usual 1-step ahead ones) also appears to carry the general message that different models may be preferred for the same dataset depending on the task in question, or, in the language of the authors, on the 'features to be matched'. Besides similarities in this general outlook, our model-fitting methodology and the context in which it is proposed are entirely different. Forecasting in the presence of structural changes is a widely studied topic in the econometrics literature, see e.g. the

comprehensive review by [52] and the references therein. In particular, [26] also use the minimisation of the 1-step ahead prediction error as a basis for model choice under non-stationarity, but, unlike us, do not consider the question of how this may lead to the preference for the 'wrong' model in finite samples. [21] apply the model-free prediction principle of [43] in the context of locally stationary time series and construct 1-step-ahead point and interval predictors.

Instead of pursuing the cross-validation approach, [35] evaluate the upper bound on the generalisation error in time series forecasting, and use its heuristically estimated version to guide model choice. We note, however, that this approach requires the estimation of some possibly difficult to estimate parameters, unlike cross-validation-based approaches. The empirical mean squared prediction error (MSPE) which we will employ in our method is closely related to the population MSPE under parameter uncertainty. The strand of literature discussing this population quantity includes [3] and [49], where approximating expressions were derived for stationary VAR time series. For locally stationary tvMA(∞) processes, [39] discuss optimal h-step ahead forecasting, in terms of the true model characteristics. Yet, they do not take parameter uncertainty into account.

While the main question we are concerned with is whether a stationary or a time-varying autoregressive model should be used for prediction, a nested question is what order the stationary or non-stationary model should have. Traditionally, order selection is done via minimisation of an information criterion, see, e.g., [11], p.301. [62] develop an adaptive criterion for model selection based on predictive risk. [1] introduced the Final Prediction Error (FPE) as a figure of merit for a potential predictor and adopts a decision theoretic approach, called the minimum FPE procedure, where the predictor with the best FPE is chosen. In practice, the decision is then based on an estimate of the FPE. In [2] a theoretical basis of the procedure is provided. [42] derive and compare MSPE for univariate and multivariate predictors when the parameters are known. They then define and estimate a criterion (a measure of predictability) to choose between these two prediction options. Their approach is similar to ours in spirit, but, firstly, it chooses between univariate and multivariate models while we consider stationary and non-stationary models and, secondly, their methodology works with the population MSPE (which moves the focus away from the observed data to the postulated model), while we work with the corresponding empirical quantity directly. This difference in approaching the problem also holds for another, more general class of special-purpose-criteria: the focused information criteria (FIC), which were introduced in [16]. The FIC methodology with the focus on choosing the model best suited for prediction was then applied in the field of time series analysis in [15], where the best AR(p) model for prediction is chosen, in [51], where the best ARMA(p, q) model for this purpose is chosen, and in [12], where models for volatility forecasting are chosen. The idea of the FIC is that the model which minimises the asymptotic MSPE is the best one and the FIC is then based on an estimator of that asymptotic MSPE. Contrary to this, our approach is based on the empirical MSPE directly, which we believe to be the more relevant quantity in many applications. Contrary to

the FIC which is based on the large-sample theory of the estimators involved, we provide finite-sample exponential bounds that imply a performance guarantee for our method. This approach can be advantageous, when it is preferred that the model choice also depends on the size of the sample, which in our view should be a natural requirement.

Our paper is organised as follows. In Section 2 we provide a simple motivating example. In Sections 3.1 and 3.2 we introduce and comment on our new time series model choice methodology. The statistical properties of our procedure are discussed in Section 3.3, where also the performance guarantee (Theorem 3.1) is provided. The results of a simulation study and the analysis of three empirical examples can be found in Sections 4 and 5. In Section 6 we discuss statistical properties of the local Yule-Walker estimator and prove its strong uniform consistency under local stationarity (Corollary 6.2). We conclude with a summary in Section 7. Proofs, technical details, additional tables and figures from the simulations section are gathered in Appendices A–J. Note that Appendices F–J are only available in the arXiv'ed version of the manuscript [32].

2. Motivating example

We consider the time-varying autoregressive (tvAR) model of order 2:

$$X_{t,T} = a_1(t/T)X_{t-1,T} + a_2(t/T)X_{t-2,T} + Z_t, \quad t = 1, ..., T,$$

where $a_1(u) := 0.15 + 0.15u$, $a_2(u) := 0.25 - 0.15u$, and Z_t is Gaussian white noise. $X_{t,T}$ is a non-stationary process which lies in the locally stationary class of [18]. We will now compare different forecasting procedures for $X_{0.9T,T}$, where $T \in \{50, 500, 5000\}$. The predictor that minimises the mean squared prediction error is given by

$$\hat{X}^{\text{oracle}}_{0.9T,T} = 0.285X_{0.9T-1,T} + 0.115X_{0.9T-2,T}.$$

Yet, since in practice the underlying model is unknown, the analyst needs to

(1) make assumptions regarding the model, and
(2) estimate the assumed model's parameters.

For the purpose of this illustration, we discuss four possible models. In the first two models we falsely assume that the data were stationary and model $X_{t,T}$ to satisfy a traditional, autoregressive (AR) equation.

- In the first of the two cases we assume an AR(1) model and
- in the second case we assume the model to be an AR(2) model.

We further, discuss cases 3–4, where the correct class of models (tvAR) is assumed. Yet,

- in case three, we falsely assume a tvAR(1) model, before
- in case four, we correctly assume the model to be a tvAR(2) model.

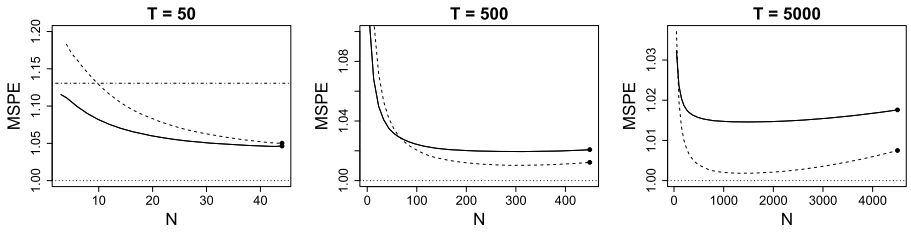

FIG 1. *Mean squared prediction errors (MSPEs) for forecasting $X_{0.9T,T}$ with predictors $\hat{X}^{1,N}_{0.9T,T}$ and $\hat{X}^{2,N}_{0.9T,T}$ associated with tvAR(1) and tvAR(2) modelling of the data, where N varies. Left, middle and right column correspond to $T = 50$, $T = 500$ or $T = 5000$, respectively. The solid lines corresponds to $\mathbb{E}(\hat{X}^{1,N}_{0.9T,T} - X_{0.9T,T})^2$, the dashed line corresponds to $\mathbb{E}(\hat{X}^{2,N}_{0.9T,T} - X_{0.9T,T})^2$. The endpoints of each line indicate the MSPEs of the predictor associated with the stationary AR(1) and AR(2) models. The dotted horizontal lines (at level 1.00) indicate the MSPE of the oracle predictor. The dashed-dotted line (approximately at level 1.13) indicates the variance of $X_{0.9T,T}$.*

Note that the true model, the tvAR(2) model, is the most complex one of the four choices. In each of the models we estimate the parameters by solving the empirical Yule-Walker equations. In the case of the tvAR models we localise by using the segment $X_{0.9T-N,T}, \dots X_{0.9T-1,T}$. In the case of the traditional, stationary AR models we use all available observations $X_{1,T}, \dots X_{0.9T-1,T}$. Details on the estimation are deferred to Section 3.1.

Denoting the localised Yule-Walker estimates of order 1 by $\hat{a}^{(1)}_{N,T}(0.9T-1)$ and the ones of order 2 by $\hat{a}^{(2)}_{1;N,T}(0.9T-1)$ and $\hat{a}^{(2)}_{2;N,T}(0.9T-1)$ we obtain the predictors

$$\hat{X}^{1,N}_{0.9T,T} := \hat{a}^{(1)}_{N,T}(0.9T-1)X_{0.9T-1,T},$$
$$\hat{X}^{2,N}_{0.9T,T} := \hat{a}^{(2)}_{1;N,T}(0.9T-1)X_{0.9T-1,T} + \hat{a}^{(2)}_{2;N,T}(0.9T-1)X_{0.9T-2,T},$$

where $\hat{X}^{1,N}_{0.9T,T}$ corresponds to the models of order 1 and $\hat{X}^{2,N}_{0.9T,T}$ corresponds to the models of order 2. The segment length N will be chosen as $0.9T-1$ in the AR models and strictly smaller than this in the tvAR models.

In Figure 1, we observe that the predictors associated with the simpler, stationary AR model perform better than or similarly well as the predictors associated with the more complex, locally stationary tvAR model if $T = 50$ or $T = 500$. If $T = 5000$ the predictor associated with the locally stationary tvAR model performs visibly better in terms of its MSPE when the segment size N is chosen appropriately. In conclusion, this example illustrates how it can be advantageous to assume a wrong, but structurally simpler model when only a short time series is available. In particular, the model chosen should depend on the task at hand (here: prediction) and on the amount of data available. For $T = 50$ the best result is obtained by assuming the AR(1) model which is the simplest of the four candidates. When $T = 500$ the more complex AR(2) model becomes advantageous. Note that this model is more complex than the AR(1) model and

thus provides a better approximation to the true tvAR(2) mechanism, but is still simplifying, because it does not take the time-varying characteristics into account at all. Only when even more data (here: $T = 5000$) are available, then the variability of the parameter estimates of the tvAR(2) model is small enough not to overshadow the modelling bias, which in this example is rather small.

Obviously, the bias-variance trade-off is at work here, which is well-known but interestingly, to our knowledge, has previously been unexplored in the important context of stationary versus non-stationary modelling for prediction. The observation to be made here, thus, is that finding the 'right' model may not always be a suitable way of proceeding when it comes to the prediction of future observations. We point out that this observation was made in other contexts of time series analysis. For example, basic exponential smoothing is a widely used forecasting and trend extrapolation technique, and although it is well-known that it corresponds to standard Box-Jenkins forecasting in the ARIMA$(0, 1, 1)$ model, it is also frequently used for data that does not follow it.

This paper investigates the question of what is the best model in terms of forecasting performance in the context of the choice between stationarity and non-stationarity. To ask this question explicitly instead of applying a test for stationarity is important since the smallest sample size T needed to reject the null hypothesis of stationarity may be smaller than the sample size needed to obtain improvement in terms of our task of interest, namely forecasting. In the following section, we will elaborate more on this question. Further, in Section 4, we see, as results of a simulation study, under which conditions using the true model is advantageous and when it can become disadvantageous.

3. When (not) to use locally stationary models under local stationarity: the new model choice methodology

3.1. Precise description of the procedure

We work in the framework of general locally stationary time series (a rigorous definition is deferred to Section 3.3), in which the available data is a finite stretch $X_{1,T}, \ldots, X_{T,T}$ from an array $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ of random variables with mean zero and finite variances. Our aim is to determine a linear predictor for the unobserved $X_{T+h,T}$ from the observed $X_{1,T}, \ldots, X_{T,T}$.

Our proposal is to compare candidate h-step ahead predictors in terms of their empirical mean squared prediction error and choose the predictor with the best forecasting performance. To this end, we proceed as follows:

Step 1. Separate the final $2m$ observations from the T available observations. The observations with indices $M_0 := \{1, \ldots, T-2m\}$, $M_1 := \{T-2m+1, \ldots, T-m\}$ and $M_2 := \{T-m+1, \ldots, T\}$ will be referred to as the training set, first validation set and second validation set, respectively. The set of unobserved data with the indices $M_3 := \{T+1, \ldots, T+m\}$ will be referred to as the test set. The size m of the separated sets will be small in comparison to the sample size T

(and hence also to the training set). Comments on why we require two distinct validation set are deferred to Section 3.2.

Step 2. Compute the linear 1-step ahead prediction coefficients

$$\hat{a}^{(p)}_{N,T}(t) := \big(\hat{\Gamma}^{(p)}_{N,T}(t)\big)^{-1}\hat{\gamma}^{(p)}_{N,T}(t) = \big(\hat{a}^{(p)}_{1;N,T}(t),\dots,\hat{a}^{(p)}_{p;N,T}(t)\big)', \tag{1}$$

(a' denotes the transposed vector a) for $t+h \in M_1 \cup M_2$, $p = 1,\dots,\max\mathcal{P}$, and $N \in \mathcal{N}$,

$$\hat{\Gamma}^{(p)}_{N,T}(t) := \big[\hat{\gamma}_{i-j;N,T}(t)\big]_{i,j=1,\dots,p}, \quad \hat{\gamma}^{(p)}_{N,T}(t) := \big(\hat{\gamma}_{1;N,T}(t),\dots,\hat{\gamma}_{p;N,T}(t)\big)' \tag{2}$$

and

$$\hat{\gamma}_{k;N,T}(t) := \frac{1}{N}\sum_{\ell=t-N+|k|+1}^{t} X_{\ell-|k|,T}X_{\ell,T}, \quad k = 0,\dots,\max\mathcal{P}. \tag{3}$$

The set of possible model orders $\mathcal{P} \subset \{0,1,\dots,\min\mathcal{N}-1\}$, with $\mathcal{P} \neq \emptyset$ and $\max\mathcal{P} \geq 1$, and the set of possible segment lengths $\mathcal{N} \subset \{\max\mathcal{P}+1,\dots,T-2m-h+1\}$, with $\mathcal{N} \neq \emptyset$, are parameters to be specified by the user. Further comments on how they are to be chosen are deferred to Section 3.2.

Step 3. Compute the linear h-step ahead prediction coefficients

$$\begin{aligned}\big(\hat{v}^{(p,h)}_{N,T}(t)\big)' &:= \big(\hat{v}^{(p,h)}_{1;N,T}(t),\cdots,\hat{v}^{(p,h)}_{p;N,T}(t)\big) := e_1'\big(\hat{A}^{(p)}_{N,T}(t)\big)^h \\ &:= e_1'\big(e_1\big(\hat{a}^{(p)}_{N,T}(t)\big)' + H\big)^h,\end{aligned} \tag{4}$$

where $\hat{a}^{(p)}_{N,T}(t)$ is defined in (1), e_1 denotes the first canonical unity vector of dimension p and H denotes a $p \times p$ Jordan block with all eigenvalues equal to zero; cf. equation (39), in the appendix. Comments on an equivalent, recursive definition are provided in Section 3.2. Next, define $f^{\text{loc.}}_{t,h;0,N} := 0$, $f^{\text{stat.}}_{t,h;0} := 0$ and, for $p \in \mathcal{P} \setminus \{0\}$ and $N \in \mathcal{N}$, compute

$$f^{\text{loc.}}_{t,h;p,N} := e_1'\big(\hat{A}^{(p)}_{N,T}(t)\big)^h(X_t, X_{t-1},\dots,X_{t-p+1})' := \sum_{i=1}^{p}\hat{v}^{(p,h)}_{i;N,T}(t)X_{t-i+1,T}, \tag{5}$$

$$f^{\text{stat.}}_{t,h;p} := e_1'\big(\hat{A}^{(p)}_{t,T}(t)\big)^h(X_t, X_{t-1},\dots,X_{t-p+1})' := \sum_{i=1}^{p}\hat{v}^{(p,h)}_{i;t,T}(t)X_{t-i+1,T} \tag{6}$$

In Figure 2, a time line is shown that illustrates the relation of the sets M_j, $j = 0,1,2,3$ and the quantities t, p, and N.

Step 4. Amongst predictors (5) select $f^{\text{loc.}}_{t,h} := f^{\text{loc.}}_{t,h;\hat{p}_{\text{loc.}},\hat{N}_{\text{loc.}}}$, with

$$(\hat{p}_{\text{loc.}}, \hat{N}_{\text{loc.}}) := \arg\min_{\substack{p\in\mathcal{P}\\ N\in\mathcal{N}}} \sum_{t+h\in M_1}\big(X_{t+h,T} - f^{\text{loc.}}_{t,h;p,N}\big)^2,$$

and, amongst predictors (6) select $f^{\text{stat.}}_{t,h} := f^{\text{stat.}}_{t,h;\hat{p}_{\text{stat.}}}$, with

$$\hat{p}_{\text{stat.}} := \arg\min_{p\in\mathcal{P}} \sum_{t+h\in M_1}\big(X_{t+h,T} - f^{\text{stat.}}_{t,h;p}\big)^2.$$

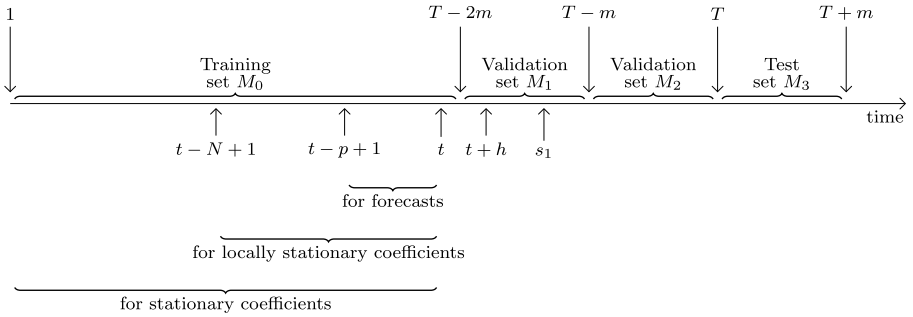

FIG 2. *Time line to illustrate the sets M_j, $j = 0, 1, 2, 3$ and relations of t, p, h, m and N. Downward pointing arrows indicate first or last indices of the four sets. The three upward pointing arrows from the left and braces indicate the indices of the observations used to compute the forecasting coefficients and the observations that are weighted by the coefficients to constitute the forecasts. The upward pointing arrow second from the right indicates the index of an observation for which the forecast is computed. The rightmost upward pointing arrow indicates $s_1 := T - m - h$, the observation up to which the MSPEs can be evaluated; cf. eq. (24).*

Note that $f^{\text{loc.}}_{t,h}$ and $f^{\text{stat.}}_{t,h}$ are the forecasts of type (5) and (6) that minimise the empirical MSPE (on M_1) within the classes of tvAR and AR models of orders $p \in \mathcal{P}$, respectively.

Step 5. Use $f^{\text{loc.}}_{t,h}$ as h-step ahead forecast of X_{t+h}, with $t + h > T$, if

$$\hat{R}_{T,j}(h) := \frac{\text{MSPE}^{\text{stat.}}_{T,j}(h)}{\text{MSPE}^{\text{loc.}}_{T,j}(h)} \geq 1 + \delta \tag{7}$$

holds for $j = 2$, and $f^{\text{stat.}}_{t,h}$ otherwise, where

$$\text{MSPE}^{*}_{T,j}(h) := \frac{1}{m} \sum_{t+h \in M_j} (X_{t+h,T} - f^{*}_{t,h})^2, \tag{8}$$

with $*$ indicating the corresponding model (we write 'loc.' for the locally stationary approach and 'stat.' for the stationary model) and $\delta \geq 0$ is a parameter by which the user of the procedure specifies which degree of superiority of the more complex procedure is required before it is preferred over the simpler alternative (cf. the end of Section 3.2).

By Theorem 3.1 we have that, with an appropriately chosen δ, the decision rule of type (7) will, with high probability, prefer the same models for forecasting observations from the second validation set ($j = 2$) and the test set ($j = 3$).

3.2. Remarks on the procedure

Some further explanations regarding the procedure are in order now. Our comments are organised according to the steps of the previous section.

Step 1. While it is common practice to separate one validation set when tuning the model parameters to avoid over-fitting, we require two such sets. This is necessary, because we would otherwise compare candidates in an unbalanced situation where $|\mathcal{P}|$ stationary predictors compete with $|\mathcal{N}| \times |\mathcal{P}|$ locally stationary ones. In our procedure, where we first choose the hyper-parameters by minimising the mean squared error on the first validation set and then choose between the two model classes by minimisation of the mean squared error on the second validation set, we achieve a fairer competition of the two model classes.

Step 2. The coefficients (1) are estimates for the coefficient functions $a_1(t/T), \ldots, a_p(t/T)$ if the data follows the tvAR(p) model

$$X_{t,T} = \sum_{j=1}^{p} a_j(t/T) X_{t-j,T} + \sigma(t/T) Z_t, \quad t = 1, ..., T, \tag{9}$$

(see, for example, [20]). Recall that Z_t is usually assumed to be white noise and that $X_{t,T}$ is non-stationary if at least one of the functions a_j, $j = 1, \ldots, p$, or σ is non-constant. A recursive algorithm to estimate the parameters was described and analysed in [37].

We are interested in linear forecasts that will perform well for time series possessing a general dependency structure. The tvAR(p) model (9) is a natural choice to approximate the linear dynamics of the observed, non-stationary time series, because in this model the coefficient functions at time t/T coincide with the 1-step ahead prediction coefficients (of order p) which define the best linear predictor. In Section 6, we show that $\hat{a}_{N,T}^{(p)}(t)$ from Step 2 can be used as estimates for the 1-step ahead linear prediction coefficients

$$\tilde{a}_T^{(p)}(t) := \arg \min_{a=(a_1,\ldots,a_p)' \in \mathbb{R}^p} \mathbb{E}\Big[\Big(X_{t,T} - \sum_{j=1}^{p} a_j X_{t-j,T}\Big)^2\Big],$$

also when the observations do not satisfy (9). A forecasting procedure derived within the tvAR(p) model can therefore be expected to behave reasonably, irrespective of whether the tvAR(p) model is true or just an approximation to the truth. Note that we use the tvAR(p) model to approximate the dynamic structure of the data in Section 3.2 and most of our examples in Section 4 are of this kind, but we do not assume that the data actually satisfies it.

Step 3. Linear h-step ahead predictors can either be obtained by iterating model equation (9) or by using a separate model for each h in which the indices of the sum on the right hand side run from $j = h, \ldots, p+h-1$. These approaches have been referred to as the plug-in method and the direct method, respectively. A comparison of the two approaches can, for example, be found in [8], where results for a class of linear, stationary processes were derived. We employ the plug-in method.

The coefficients $\hat{v}_{N,T}^{(p,h)}(t)$ defined in (4) can be computed efficiently via the recursion:

$$\hat{v}_{i;N,T}^{(p,1)}(t) := \hat{a}_{i;N,T}^{(p)}(t), \quad i = 1, \ldots, p,$$
$$\hat{v}_{i;N,T}^{(p,\eta)}(t) := \hat{a}_{i;N,T}^{(p)}(t) \hat{v}_{1;N,T}^{(p,\eta-1)}(t) + \hat{v}_{i+1;N,T}^{(p,\eta-1)}(t) I\{i \leq p-1\}, \quad \eta = 2, 3, \ldots, h.$$

From the previous comments it can be seen how the predictors $f^{\text{loc.}}_{t,h;p,N}$ and $f^{\text{stat.}}_{t,h;p}$ relate to the choice of modelling the time series' dynamics by a tvAR(p) or AR(p) model, respectively. In each of these model classes, increasing the order p will give a better approximation of the dynamics, but increase the complexity of the model, and make it more difficult to deal with under parameter uncertainty.

The parameters $\mathcal{P}$ and $\mathcal{N}$ are sets of integers to be chosen by the user. The choice should depend on T. $p \in \mathcal{P}$ determines the order of the tvAR(p) model that is used to approximate the dynamics. $N \in \mathcal{N}$ determines the degree of locality in the estimation of the coefficients. The parameters $p \in \mathcal{P}$ and $N \in \mathcal{N}$ will influence the degree of bias and variance of the predictor. Our selection mechanism will balance them implicitly.

Traditional choice of N. It is obvious that the variance of the estimator can decrease when a larger segment is used, but that the non-stationarity will potentially inflict an additional bias that increases with N. Under the condition that $N/T + T/N^2 = o(1)$, [20] derive asymptotic expansions for the local Yule-Walker estimator's bias and variance for a centred sample. It follows from their results, that for the one-sided sample we require for forecasting, N should be chosen at the order of $T^{2/3}$, with the constant depending on the second derivatives of the true model quantities, which are unknown and difficult to estimate. The choice of $\mathcal{N}$ should thus, ideally, be such that $N \asymp T^{2/3}$, for all $N \in \mathcal{N}$. In practice, since the true model parameters are unknown, this rate provides very little guidance to the user of the method. We recommend, though, to adhere to two facts: the upper and lower bound of $\mathcal{N}$ should be bounded away from 0 and T, respectively. In other words, we recommend to choose $\mathcal{N}$ with $\min \mathcal{N}$ large enough, for the performance guarantee to be valid (cf. Theorem 3.1) and $\max \mathcal{N}$ being substantially smaller than T, to ensure that there is a clear boundary between the locally stationary and the stationary approach. [50] propose to adaptively choose a bandwidth for local M-estimators by minimising a cross-validation functional.

Traditional choice of p. As described in the beginning of this section we use the tvAR(p) model to approximate the dynamic structure of the data. Intuitively, we have that the larger the order p the better the approximation to the true dynamic structure. In opposition to the previously discussed question of how to choose the segment length N, we here have that a smaller p will inflict a modelling bias, while a larger p will typically be accompanied by an inflation of the variance of the estimation, because it implies that more parameters need to be estimated. Traditionally, the model order is chosen by minimizing information criteria as for example AIC or BIC. [15] propose to use a version of the focused information criterion (FIC, see [16]) to select the model order of a stationary AR(p) model optimal with respect to forecasting when the true model is known to be AR(∞). However, as mentioned in the introduction, the FIC-based methods employs an estimator of the asymptotic MSPE, while our approach is based on the empirical MSPE, which facilitates our focus on the finite sample performance.

Step 4 and 5. Our procedure performs two stages of selections. Firstly (in Step 4), it selects the model order p and, for the locally stationary approach, the segment length N by comparing predictors within each class of models under consideration (i.e., time-varying or non-time-varying autoregressive models). The parameters p and N are chosen such that the empirical MSPE (predicting observations from M_1) is minimised. Secondly (in Step 5), a final competition of the winners is performed to select among the two classes of models. The procedure that minimises the empirical MSPE (predicting M_2) is selected and used for forecasting of the test set (M_3). In our theoretical analysis of the next section (see, in particular, Theorem 3.1) we show that the proposed procedure will, with high probability, choose the same class of models on the validation as on the test set, implying that the procedure with the best empirical performance will be selected.

The parameter δ. By introduction of the parameter $\delta \geq 0$ the user is given additional control over which model the procedure prefers. In the simplest case, $\delta = 0$, this reduces to a straight choice between the two model classes, whereby the time-varying model is chosen if it performs better or equally well. Choosing $\delta > 0$ introduces penalisation against the choice of more complex models. In this case, the predictor derived from the more complex, locally stationary model is only chosen if it performs at least $\delta \cdot 100\%$ better than the one derived from the simpler, stationary model.

Generalisations. Besides linear predictions for stationary or locally stationary time series models, the general principle of our method can also be applied in many other situation. To illustrate this, we outline two examples below.

Non-linear predictions with neural networks. In this scenario we either choose a neural network trained with the N most recent observations (i.e., loc.) or with all available observations (i.e., stat.). To this end proceed as follows: with the available data partitioned as described in Step 1, consider a range of candidate networks (with different network topologies) suitable for forecasting the observations from the first validation set. Train them either with the N most recent observations (loc.) or with all available observations (stat.). After first choosing the network for which we see the smallest MSPE on the first validation set within each class (loc. or stat.), we then choose that class for which the winner from the previous step obtains the best performance on the second validation set.

Predictors obtained from locally stationary or long-memory time series models. Long-range dependence and non-stationarity can lead to the same stylised facts in financial time series; cf. [36]. Choosing a test, for example from [44], to distinguish between the two model classes seems to be a sensible approach, but this might not lead to the best choice if the aim is to choose a model for the purpose of prediction. With the model choice methodology in this paper, one proceeds as follows: first, fit a set of long-range dependence models and a range of locally stationary models, use the implied predictors to forecast the observations from the first validation set and choose the model with the best forecasting performance within each model class. Then, choose between the two model classes by comparing the winning models within each class with respect

to their empirical performance in predicting the observations from the second validation set.

3.3. Performance guarantee: theoretical result for the general case

In this section, we establish theoretical results that will facilitate our analysis of the model choice suggested by decision rule (7). We show that the probability of choosing different models on the validation and the test set decays to zero at an exponential rate, which can be viewed as a performance guarantee of our model choice methodology.

To rigorously prove the results, some definitions and assumptions are in order. Throughout this paper, we work with the doubly indexed process $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$. The first index (i.e., t) refers to the time. The second index (i.e., T) indicates how well the covariance structure of $(X_{t,T})_{t\in\mathbb{Z}}$ can be approximated locally by the autocovariance function of a stationary process. We will assume that, for large T, segments of observations $X_{t,T}$ with their indices $t \approx uT$ are approximately weakly stationary. The parameter u is continuous and often referred to as the rescaled time. If the index T coincides with the number of observations in a time series, then $u \in [0,1]$ (cf. [17]). This restriction is not necessary and in fact, because we will consider m unobservables (to be forecast) in addition to the T observations (available at the time when the forecasting is done) it is more convenient to allow $u > 1$, as was also done by [54].

The following definitions from [54] are required for our assumptions. For an array $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ with finite variances, the time-varying covariance function is defined for all $t \in \mathbb{Z}, T \in \mathbb{N}^*$ and $k \in \mathbb{Z}$ as

$$\tilde{\gamma}_{k,T}(t) = \mathrm{Cov}\left(X_{t,T}, X_{t-k,T}\right). \tag{10}$$

A local spectral density f is a $\mathbb{R}^2 \to \mathbb{R}_+$ function, (2π)-periodic and locally integrable with respect to the second variable. The local covariance function γ associated with the time-varying spectral density f is defined for $(k,u) \in \mathbb{Z}\times\mathbb{R}$ by

$$\gamma_k(u) = \int_{-\pi}^{\pi} \exp\left(\mathrm{i}k\lambda\right) f(u,\lambda)\mathrm{d}\lambda. \tag{11}$$

The first five assumptions are specific to the kind of data we may apply our result to.

Assumption 1 (Local stationarity, [54])**.** *Let the array of random variables $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ have finite variances. We say that $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ is locally stationary with local spectral density f if the time-varying covariance function of $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ and the local covariance function associated with f satisfy*

$$\left|\tilde{\gamma}_{k,T}\left(t\right) - \gamma_k\left(\frac{t}{T}\right)\right| \leq \frac{C}{T}, \tag{12}$$

where $C \geq 0$ is a constant.

Assumption 2 (Geometrically α-mixing)**.** *There exist constants $K > 0$ and $\rho > 1$ such that, for every $n \in \mathbb{N}$,*

$$\alpha(n) := \sup_{T\in\mathbb{N}^*} \sup_{t\in\mathbb{Z}} \sup_{A\in\sigma(X_{s,T}:s\le t)} \sup_{B\in\sigma(X_{s,T}:s\ge t+n)} \left|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\right| \le K\rho^{-n}. \tag{13}$$

Assumption 3. *The local spectral density f is bounded from above and below:*

$$0 < m_f \le f(u,\lambda) \le M_f. \tag{14}$$

Assumption 4. *The local spectral density f is continuously differentiable with respect to the first argument and the partial derivative is uniformly bounded. More precisely, assume the existence of $M'_f \ge 0$ such that*

$$\left|\frac{\partial}{\partial u} f(u,\lambda)\right| \le M'_f. \tag{15}$$

Assumption 5 (Bernstein-type condition)**.** *There exist $c > 0$ and $d \ge 1/2$, such that*

$$\mathbb{E}|X_{t,T}|^k \le c^{k-2}(k!)^d \mathrm{Var}(X_{t,T}) \quad t\in\mathbb{Z};\ k = 2,3,\dots$$

The assumptions are reasonable and non-restrictive in the sense that many popular and widely used time series models (e.g., a wide range of tvARMA models) satisfy the full set of assumptions. The notion of local stationarity we impose (Assumption 1) goes beyond that of locally stationary linear processes and, in particular, we do not require the data to be tvAR. Assumption 1 is satisfied for second order stationary process (then we have $C = 0$), the general (linear) locally stationary process introduced by [17], but also non-linear processes as elaborated by [54]. Assumption 2 is satisfied for a broad class of (linear and non-linear) locally stationary time series models; see, for example, [25] or [57]. Assumptions 3 and 4 can be verified by considering the local spectral density when it is given explicitly. For example, in the tvAR model that we used to motivate our prediction approach in Section 3.2, see (9), and as examples in Section 4, the local spectral density and local covariances are naturally those of the stationary AR process when the parameter u of the coefficient functions is chosen as t/T. We will refer to these AR processes as the tangent processes of the tvAR process. Similar assumptions with respect to the local spectral density are common in the literature; cf. [18]. Processes with sub-Gaussian marginal distributions satisfy Assumption 5; cf. Lemma E.8 in the appendix. We recall, from Section 3.2, that the tvAR(p) model is used to approximate the linear dynamic structure of the data, but that we do not assume that the data actually satisfies it. Thus our results apply in a more general context. We require Assumptions 2 and 5 to prove that the probabilities in our results decay at an exponential rate.

As a consequence of Assumptions 1 and 3, we have

$$\pi m_f \le \sigma^2_{t,T} := \mathrm{Var}(X_{t,T}) \le 3\pi M_f, \quad \text{for all } T \ge \frac{C}{\pi m_f}. \tag{16}$$

Further, by Assumption 4 and Leibniz's integral rule, we have that $\gamma_k'(u)$ exists and has the following form

$$\gamma_k'(u) := \frac{\partial}{\partial u}\gamma_k(u) = \frac{\partial}{\partial u}\int_{-\pi}^{\pi} \exp(\mathrm{i}\ell\lambda)\, f(u,\lambda)\mathrm{d}\lambda = \int_{-\pi}^{\pi} \exp(\mathrm{i}\ell\lambda)\, \frac{\partial}{\partial u} f(u,\lambda)\mathrm{d}\lambda, \tag{17}$$

which, in particular, implies that $|\gamma_k'(u)| \leq 2\pi M_f'$.

Assumptions 6 and 7, which we state below, are more specific to our procedure. They concern minimum requirements for the size m of the validation sets and the minimum segment size $\min \mathcal{N}$ which are used to compute the forecast as well as the number T of observations required to be available at the time the forecasts are to be determined. To precisely state the final two assumptions, we will define $q(\delta)$ that quantifies the difference between the two approaches in terms of their expected empirical mean square prediction error forecasting performance.

To make the definition of $q(\delta)$ precise in an accessible manner we now present it from the inside outwards. At the core we have the local covariance function defined in (11) and averaged versions

$$\begin{aligned}
\gamma_\Delta^{(p)}(u) &:= \int_0^1 \gamma^{(p)}(u + \Delta(x-1))\mathrm{d}x, \quad \gamma^{(p)}(u) := [\gamma_1(u) \ \dots \ \gamma_p(u)]', \\
\Gamma_\Delta^{(p)}(u) &:= \int_0^1 \Gamma^{(p)}(u + \Delta(x-1))\mathrm{d}x, \quad \Gamma^{(p)}(u) := (\gamma_{i-j}(u);\, i,j = 1,\dots,p).
\end{aligned} \tag{18}$$

If $\Delta := (N-|k|)/T$ or N/T and $u = t/T$, then the entries $\int_0^1 \gamma_k(u+\Delta(x-1))\mathrm{d}x$ in $\gamma_\Delta^{(p)}(u)$ and $\Gamma_\Delta^{(p)}(u)$ are approximations for the expectation $\mathbb{E}\hat{\gamma}_{k;N,T}(t)$ of the lag k autocovariance estimate $\hat{\gamma}_{k;N,T}(t)$ computed from $X_{t-N+1,T},\dots,X_{t,T}$; cf. Lemma D.1. This seemingly complicated construction is necessary, because we do not require that N/T is negligible. By allowing $\Delta > 0$ we can capture the evolving second moments of the processes. Further note that, for every $u \in \mathbb{R}$ and $\Delta \geq 0$, the averaged local autocovariances form the autocovariance function of a stationary process that can be seen as an average of the stationary approximations $X_t(\cdot)$ over the local times in $[u-\Delta, u]$. Solving the Yule-Walker equations for this average process yields

$$a_\Delta^{(p)}(u) := \big(a_{1,\Delta}^{(p)}(u),\dots,a_{p,\Delta}^{(p)}(u)\big)' := \Gamma_\Delta^{(p)}(u)^{-1}\gamma_\Delta^{(p)}(u). \tag{19}$$

As can be seen from Theorem 6.1 and Lemma B.2, $a_\Delta^{(p)}(u)$ is an approximation to the limit of the Yule-Walker estimate obtained from $X_{t-N+1,T},\dots,X_{t,T}$. It further is related to the 1-step ahead linear forecasting coefficients, as can be seen from Lemma B.1. The h-step ahead counterpart of $a_\Delta^{(p)}(u)$ is defined as

$$\begin{aligned}
\big(v_\Delta^{(p,h)}(u)\big)' &:= \big(v_{1;\Delta}^{(p,h)}(u), v_{2;\Delta}^{(p,h)}(u), \cdots, v_{p;\Delta}^{(p,h)}(u)\big) \\
&:= e_1'\big(A_\Delta^{(p)}(u)\big)^h := e_1'\big(e_1\big(a_\Delta^{(p)}(t)\big)' + H\big)^h,
\end{aligned} \tag{20}$$

where e_1 and H are the same as in (4). Then, for $u \in \mathbb{R}$, $\Delta_1, \Delta_2 \geq 0$, the functions $\mathrm{MSPE}^{(p,h)}_{\Delta_1,\Delta_2}(u)$ are defined as

$$\mathrm{MSPE}^{(p,h)}_{\Delta_1,\Delta_2}(u) := \int_0^1 g^{(p,h)}_{\Delta_1}\Big(u + \Delta_2(1-x)\Big)\mathrm{d}x, \tag{21}$$

where $g^{(0,h)}_{\Delta}(u) := \gamma_0(u)$ and, for $p \in \mathbb{N}^*$, with $\gamma^{(p,h)}_0(u) := \big(\gamma_h(u), \ldots, \gamma_{h+p-1}(u)\big)'$,

$$g^{(p,h)}_{\Delta}(u) := \gamma_0(u) - 2\big(v^{(p,h)}_{\Delta}(u)\big)'\gamma^{(p,h)}_0(u) + \big(v^{(p,h)}_{\Delta}(u)\big)'\Gamma^{(p)}_0(u)v^{(p,h)}_{\Delta}(u). \tag{22}$$

From Lemmas A.1 and A.3, it can be seen that

$$\mathrm{MSPE}^{(p,h)}_{s,m,N,T} := \frac{1}{m}\sum_{t=s+1}^{s+m}\Big(X_{t+h,T} - \sum_{i=1}^{p}\hat{v}^{(p,h)}_{i;N,T}(t)X_{t-i+1,T}\Big)^2,$$

concentrates around $\mathrm{MSPE}^{(p,h)}_{\Delta_1,\Delta_2}(u)$ with $\Delta_1 = N/T$, $\Delta_2 = m/T$ and $u = s/T$. Note that two arguments Δ_1 and Δ_2 are required to allow for the averaging of possible effects due to non-stationarity originating from (a) either the computation of the forecasting coefficients or (b) the computation of the mean squared prediction errors. The quantity $g^{(p,h)}_{N/T}(t/T)$ approximates the MSPE of $f^{\mathrm{loc.}}_{t,h;p,N}$ defined in (5). In the case of 1-step ahead forecasts we can simplify the expression in (22) to

$$g^{(p,1)}_{\Delta}(u) = \mathbb{E}[(\hat{X}^{(p)}_t(u) - X_t(u))^2] + \big\|a^{(p)}_{\Delta}(u) - a^{(p)}_0(u)\big\|^2_{\Gamma^{(p)}_0(u)}, \tag{23}$$

where $\hat{X}^{(p)}_t(u) := \sum_{j=1}^{p} a^{(p)}_{j,0}(u)X_{t-j}(u)$ is the best linear 1-step ahead forecast for $X_t(u)$ and $\|x\|^2_{\Gamma} := x'\Gamma x$ denotes the quadratic form associated with Γ. Decomposition (23) is into two non-negative quantities. The first term only depends on the characteristics of the stationary tangent process $X_t(u)$ and will be a decreasing sequence with index p for any u. The second term is the squared weighted difference of the forecasting coefficients obtained from the stationary approximation at time u and the forecasting coefficients obtained from the non-stationary data; more precisely from the stationary approximations $X_t(\cdot)$ "averaged" over $[u - \Delta, u]$.

The final two assumptions require that the size m of the validation sets, the smallest segment size $\min\mathcal{N}$ from which locally stationary forecasting coefficients are computed, and the number of available observations T are large enough in relation to the maximum model order $\max\mathcal{P}$, the forecasting horizon h, and the minimum possible difference of performance of stationary and locally stationary forecasts in terms of MSPE, which we measure by

$$q(\delta) := \min_{\substack{p_1,p_2\in\mathcal{P}\\ N\in\mathcal{N}}}\Big|\mathrm{MSPE}^{(p_1,h)}_{s_1/T,m/T}\big(\frac{s_1}{T}\big) - (1+\delta)\cdot\mathrm{MSPE}^{(p_2,h)}_{N/T,m/T}\big(\frac{s_1}{T}\big)\Big|, \tag{24}$$

where $s_1 := T - m - h$.

Assumption 6 requires the size m of the validation sets and the smallest possible segment lengths $N \in \mathcal{N}$ from which to estimate the forecasting coefficients to be 'large enough'.

Assumption 6 (Minimum size for m and $\min\mathcal{N}$)**.** *Let* $K_0 := 4C_0(2C_0+1)$*. For* $\delta \geq 0$*,* $m, T, h \in \mathbb{N}^*$*,* $\mathcal{P} \subset \{0, 1, \ldots, \min\mathcal{N} - 1\}$*, such that* $\mathcal{P} \neq \emptyset$*,* $\max\mathcal{P} \geq 1$*, and* $\emptyset \neq \mathcal{N} \subset \{\max\mathcal{P} + 1, \ldots, T - 2m - h + 1\}$*, assume that*

$$\min\mathcal{N} \geq 8h2^h\big(C_0\big)^{2h+1}(\max\mathcal{P})^2 \max\left\{\frac{20(1+\delta)}{q(\delta)}, 1\right\}\Big[6(2\pi M_f' + C) + 1\Big]$$

and

$$\max\left\{\Big(\frac{h+\max\mathcal{P}}{m}\Big)^{\frac{1+4d}{3+8d}} K_0^h(\max\mathcal{P})^2, \Big(\frac{\max\mathcal{P}}{\min\mathcal{N} - \max\mathcal{P}}\Big)^{\frac{1+2d}{3+4d}} K_0^h(\max\mathcal{P})^3 h\right\} < \frac{q(\delta)}{20(1+\delta)}. \quad (25)$$

Assumption 7 requires the sample size T to be 'large enough'.

Assumption 7 (Minimum sample size T)**.** *With* C_0 *and* C_1 *defined in* (36)*, in the appendix, and* C *and* M_f' *from Assumptions 1 and 4, respectively,*

$$T \geq \max\left\{6h2^hC_1(\max\mathcal{P})^2, 4m\big(2h+1\big)\big(C_0\big)^{2h+1}M_f'\frac{20(1+\delta)}{q(\delta)}\right\}.$$

The intuition behind the final two assumptions is that if two forecasts exist, one stationary and one locally stationary, that behave similarly well in terms of approximations to their expected empirical mean squared errors, then m and $\min\mathcal{N}$ need to be large enough (in relation to $q(\delta), h$, and $\max\mathcal{P}$). Further, we require that T exceeds a specified level (depending on $q(\delta), h, \max\mathcal{P}$, and m) to be able to provide bounds of the error of approximation of the local stationary process with the tangent process. The specific form of Assumptions 6 and 7 are due to technical reasons in our proof and, in fact, our simulation results in Section 4 suggest that the probability bounded in Theorem 3.1 will also be large for T smaller than the threshold, as long as δ is chosen appropriately. The quantity $q(\delta)$ is constructed to measure the difference between the MSPEs of the stationary predictors for different p_1 and the MSPEs of the locally stationary predictors for different (p_2, N) scaled by a factor of $1+\delta$. Assumptions 6 and 7 are slightly stronger than necessary, as we do not only require only those procedures to perform differently for the p_1 and (p_2, N) that yield the best result, but we require it for any combination. This is due to our method of proof. On the other hand, it is obvious that some condition like this is required for consistency of the procedure, because if there is no difference in performance either approach may equally likely be chosen. It is important to note that in the situation where both approaches perform equally well we do not need the selection to be consistent.

The quantity $q(\delta)$ depends on the model under consideration and, as $|\mathcal{P}|$ and $|\mathcal{N}|$ get larger, may potentially tend to zero. Thus, to employ Theorem 3.1 in practice, one has to analyse $q(\delta)$ to determine the right bounds stated in Assumptions 6 and 7. In Section 3.4 we show how this can be done in the special case where $\mathcal{P} = \{1\}$ and $h = 1$. There we show that if δ is chosen large

enough or, in the case where the true model is non-stationary, if δ is chosen small enough, then q is bounded away from 0. If $q(\delta) > \varepsilon_0 > 0$, then, even in an asymptotic framework where h and $\max \mathcal{P}$ do not need to be bounded and $m, \min \mathcal{N} \to \infty$ as $T \to \infty$, then condition (25) will hold for T large enough, if

$$(h + \max \mathcal{P})\big(K_0^h (\max \mathcal{P})^2\big)^{\frac{3+8d}{1+4d}} = o(m), \quad \text{and}$$
$$(\max \mathcal{P})^{\frac{10+14d}{1+2d}} \big(K_0^h h\big)^{\frac{3+4d}{1+2d}} = o(\min \mathcal{N}).$$

Note that, $(\max \mathcal{P})^{1+2\frac{3+8d}{1+4d}} \leq (\max \mathcal{P})^{17/3}$ and $(\max \mathcal{P})^{\frac{10+14d}{1+2d}} \leq (\max \mathcal{P})^{17/2}$. Therefore, if $h = O(1)$, we have that condition (25) will hold for T large enough, if $\max \mathcal{P} = O(m^{3/17})$ and $\max \mathcal{P} = O\big((\min \mathcal{N})^{2/17}\big)$.

For the finite sample case, the quantity $q(\delta)$ can easily be computed for any tvAR(p) model. A function performing the necessary calculations is provided in our R package **forecastSNSTS**. Numerical illustrations are provided in Section 4.

We are now ready to state the main result that guarantees that our procedure will, with high probability, choose the predictor that achieves the best empirical performance on the test set.

Theorem 3.1. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1–5 and $\mathbb{E}X_{t,T} = 0$. Further, let $\delta, m, T, h, \mathcal{P}$, and $\mathcal{N}$ be such that Assumptions 6–7 are satisfied. Then, with $\hat{R}_{T,j}(h)$, $j = 2, 3$, defined in (7), we have*

$$\begin{aligned}
&\mathbb{P}\Big((\hat{R}_{T,2}(h) \geq 1+\delta \text{ and } \hat{R}_{T,3}(h) \geq 1+\delta) \text{ or } (\hat{R}_{T,2}(h) < 1+\delta \\
&\quad \text{and } \hat{R}_{T,3}(h) < 1+\delta)\Big) \\
&\geq 1 - 6D_1|\mathcal{P}|^2|\mathcal{N}|\Bigg[(\max \mathcal{P})^2 \exp\left(-D_2\Big(\frac{m}{h+\max \mathcal{P}}\Big)^{1/(3+8d)}\right) \\
&\qquad + m(\max \mathcal{P})^3 \exp\left(-D_3\Big(\frac{\min \mathcal{N} - \max \mathcal{P}}{\max \mathcal{P}}\Big)^{1/(3+4d)}\right)\Bigg],
\end{aligned}$$

where D_1, D_2 and D_3 are constants, defined in (41), in the appendix, that only depend on K and ρ, m_f and M_f, and c and d the constants from Assumption 2, 3 and 5, respectively.

The proof of Theorem 3.1 is long and technical and therefore deferred to Section A. The probability in Theorem 3.1 tends to one if $m \gg (h + \max \mathcal{P}) \times [\log(|\mathcal{P}|^2|\mathcal{N}|(\max \mathcal{P})^2)]^{3+8d}$ and $\min \mathcal{N} \gg \max \mathcal{P}[\log(|\mathcal{P}|^2|\mathcal{N}|m(\max \mathcal{P})^3)]^{3+4d}$, where we have used the notation $a_T \gg b_T$ for $a_T/b_T \to \infty$, as $T \to \infty$. Thus, Theorem 3.1 provides a "performance guarantee" of our model choice methodology in the sense that it asserts that, with high probability, the method which we have observed to perform better empirically in forecasting the observations from the second validation set will also perform better empirically in forecasting the future, not yet observed values of the test set.

3.4. Theoretical results for a simple, special cases

To illustrate the usefulness of Theorem 3.1 we now discuss the special case in which the model order is pre-determined to be 1, for both locally stationary and stationary forecasts, and the forecasting horizon is 1-step ahead; i.e., $\mathcal{P} := \{1\}$ and $h = 1$. Though this special case is usually not of practical interest, restricting ourselves will allow to illustrate how the general conditions simplify and can more easily be understood. For the simplification we proceed by finding lower bounds for $q(\delta)$ (uniformly in T and $N \in \mathcal{N}$) which in turn allows us to state more explicit conditions that imply Assumptions 6 and 7.

To apply Theorem 3.1, we require that the MSPE of the stationary predictors are not to close to $1+\delta$ times the MSPE of the locally stationary predictors (cf. Assumption 6). Therefore, we now consider the following two cases:

(a) The parameter δ is chosen large enough.
(b) The parameter δ is chosen small enough and the true model is non-stationary.

To make the requirements precise, we define

$$\rho := \sup_{1-m/T \le u \le 1} \left| \frac{\gamma_1(u)}{\gamma_0(u)} \right|, \tag{26}$$

$$D_{\sup} := \sup_{1-m/T \le u \le 1} \left| \frac{\int_0^1 \gamma_1\big(u + \frac{s_1}{T}(x-1)\big)\mathrm{d}x}{\int_0^1 \gamma_0\big(u + \frac{s_1}{T}(x-1)\big)\mathrm{d}x} - \frac{\gamma_1(u)}{\gamma_0(u)} \right|, \tag{27}$$

and

$$D_{\inf} := \inf_{1-m/T \le u \le 1} \left| \frac{\int_0^1 \gamma_1\big(u + \frac{s_1}{T}(x-1)\big)\mathrm{d}x}{\int_0^1 \gamma_0\big(u + \frac{s_1}{T}(x-1)\big)\mathrm{d}x} - \frac{\gamma_1(u)}{\gamma_0(u)} \right|, \tag{28}$$

where $\gamma_0(u)$ and $\gamma_1(u)$ are the local autocovariances from Assumption 1. The suprema and the infimum are with respect to points u of the second validation set. Averaging of autocovariances in the first terms of $D_{\inf}$ and $D_{\sup}$ is across the training set and first validation set. Note that $D_{\inf} \le D_{\sup} \le 2$ and that $D_{\inf}$ is a measure for the non-stationarity of the training set. In particular, it will vanish if the data stems from a stationary process. Further, note that ρ is a measure for the strength of serial dependence.

The simplified conditions that imply Assumptions 6 and 7 for the special case, will be stated in terms of ρ, $D_{\sup}$ and $D_{\inf}$. Note that, also in the case where $\mathcal{P} = \{1\}$, the quantity $q(\delta)$ in Assumptions 6 and 7 depends on $\mathcal{N}$, but the $D_{\inf}$, $D_{\sup}$ and ρ only depend on m, T, $\gamma_1(\cdot)$ and $\gamma_0(\cdot)$. Therefore, the conditions in Lemmas 3.2 and 3.3 are indeed simpler than Assumptions 6 and 7. Further note that the local autocovariances $\gamma_k(\cdot)$ can be determined easily for many time series models. If, for example the data stems from a tvAR(1) process with coefficient function a, then we have $\gamma_k(u) = a(u)^{|k|}/(1-a(u)^2)$, $k \in \mathbb{Z}$.

We now state two results about the special case of the procedure for 1-step ahead forecasting. The first result illustrates that the modified procedure will be consistent if δ is chosen large enough:

Lemma 3.2. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1–5, and $\mathbb{E}X_{t,T} = 0$. Assume that $\rho < 1$ and $\delta \geq 2D_{\sup}^2/(1-\rho^2)$, where ρ and $D_{\sup}$ are defined in (26) and (27). Then, $q(\delta) \geq \delta\pi m_f(1-\rho^2)$, where m_f is from Assumption 3. In particular, this implies that constants K_1, K_2 and K_3, defined in the proof, exist such that, if $m > K_1$ and $\min\mathcal{N} > K_2$ then Assumption 6 holds. Further, if $T \geq K_3 m$, then Assumption 7 holds.*

Further more, we have as a second result that if the true model is non-stationary in the sense that the quantity $D_{\inf}$ is large compared to N/T for all $N \in \mathcal{N}$, then we also have consistency for δ's that are small enough:

Lemma 3.3. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1–5, $\mathbb{E}X_{t,T} = 0$, and*

$$D_{\inf}^2 \geq 2\Big(\frac{M_f'}{m_f}\frac{\max\mathcal{N}}{T}\Big)^2, \tag{29}$$

with $D_{\inf}$ defined in (28). Assume that $\delta \leq \frac{1}{8}D_{\inf}^2$. Then, $q(\delta) \geq \pi D_{\inf}^2 m_f/2$, where m_f is from Assumption 3. In particular, this implies that constants K_4, K_5 and K_6, defined in the proof, exist such that, if $m > K_4$ and $\min\mathcal{N} > K_5$ then Assumption 6 holds. Further, if $T \geq K_6 m$ then Assumption 7 holds.

By Lemma 3.2 we have that, in the case where $\mathcal{P} = \{1\}$, $h = 1$ and $\delta \geq 0$ have been fixed, Assumption 6 will hold if m and $\min\mathcal{N}$ are chosen larger than some constant. This requirement is not restrictive, in the sense that we would typically consider m and $\min\mathcal{N}$ to diverge as T diverges, such that by Theorem 3.1 the probability for consistent model choice will tend to one. In Lemma 3.3 the restrictions on m and $\min\mathcal{N}$ are even weaker, as in a typical application $\max\mathcal{N}/T$ will tend to 0. In both Lemmas 3.2 and 3.3 the condition that implies Assumption 7 to hold is that T is chosen larger than a multiple of m, which is eventually satisfied if m/T tends to zero.

Remark 3.4. *In Lemmas 3.2 and 3.3 a lower bound of the form*

$$\frac{q(\delta)}{20(1+\delta)} \geq \varepsilon_0 \tag{30}$$

is proven, for the special case where $\mathcal{P} = \{1\}$ and $h = 1$. This lower bound implies that Assumption 6 holds, but it is in fact stronger, as Assumption 6 allows for $q(\delta)$ tending to 0, as $|\mathcal{N}|$ and $|\mathcal{P}|$ increase, as long as m and $\min\mathcal{N}$ are increasing fast enough. Under condition (30) and the conditions of Theorem 3.1 we have the following, stronger result:

$$\begin{aligned}
&\mathbb{P}\Big((\hat{R}_{T,2}(h) \geq 1+\delta \text{ and } \hat{R}_{T,3}(h) \geq 1+\delta) \text{ or } (\hat{R}_{T,2}(h) < 1+\delta \\
&\quad \text{and } \hat{R}_{T,3}(h) < 1+\delta)\Big) \\
&\geq 1 - 6D_1|\mathcal{P}|^2|\mathcal{N}|\Bigg[(\max\mathcal{P})^2\exp\left(-D_2\varepsilon_0^{\frac{1}{2+4d}}\Big(\frac{m}{K_0^h(\max\mathcal{P})^2(h+\max\mathcal{P})}\Big)^{\frac{1}{2+4d}}\right) \\
&\qquad + m(\max\mathcal{P})^3\exp\left(-D_3\varepsilon_0^{\frac{1}{2+2d}}\Big(\frac{\min\mathcal{N}-\max\mathcal{P}}{hK_0^h(\max\mathcal{P})^4}\Big)^{\frac{1}{2+2d}}\right)\Bigg],
\end{aligned}$$

which can be proved along the same lines of the proof of Theorem 3.1, together with inequality (70) *from the proof of Lemma A.2, which is available in the arXiv'ed version of the manuscript [32].*

In particular, when the parameters $\max \mathcal{P} = (\max \mathcal{P})(T)$, $h = h(T)$ *and* $\varepsilon_0 = \varepsilon_0(T)$ *are bounded sequences* (ε_0 *also bounded away from zero), we get the following bound:*

$$\begin{aligned}&\mathbb{P}\Big((\hat{R}_{T,2}(h) \geq 1+\delta \text{ and } \hat{R}_{T,3}(h) \geq 1+\delta) \text{ or } (\hat{R}_{T,2}(h) < 1+\delta \\ &\quad \text{and } \hat{R}_{T,3}(h) < 1+\delta)\Big) \\ &\geq 1 - \kappa_1 |\mathcal{N}| \Big(\exp(-m^{1/(2+4d)}\kappa_2) + m \exp(-N^{1/(2+2d)}\kappa_3)\Big) \end{aligned} \tag{31}$$

where $\kappa_1, \kappa_2, \kappa_3$ *are constants that do not depend on* m *or* N *and* d *is the constant from Assumption 5 (e.g., for sub-Gaussian processes:* $d = 1/2$*).*

4. Simulations

In this section we discuss finite sample properties of the estimates $\hat{R}_{T,i}(h)$, defined in (7), and their population counterparts $R_{T,j}(h) := (\mathbb{E}(\mathrm{MSPE}^{\mathrm{stat.}}_{T,j}(h)))/(\mathbb{E}(\mathrm{MSPE}^{\mathrm{loc.}}_{T,j}(h)))$. The simulation study was conducted with the R package `forecastSNSTS` [48, 31], available from The Comprehensive R Archive Network (CRAN). In particular, we investigate the performance of decision rule (7). To this end, we have considered 15 different tvAR models. Three of the models are stationary, the other 12 are non-stationary. Amongst the non-stationary processes we have some where the covariance structure changes quickly and some where the covariances change slowly. Further, we will have examples where the processes given by the parameters at some local time u are almost unit root and some where they are not.

For each of the models we proceed as follows. We simulate sequences of length $T + m = n \in \{100, 200, 500, 1000, 2000, 4000, 6000, 8000, 10000\}$. The $T + m$ observations, with T and m as in Section 3, contain the training, validation and test set. We separate the test and validation sets of length $m := \lfloor n^{.85}/4 \rfloor$. Thus, $n_i := n - (3-i)\lfloor n^{.85}/4 \rfloor$, $i = 0, ..., 3$, mark the end indices of the training set, the validation sets and the test set, respectively. We have chosen m as a function of n in such a way that $m = o(n)$ and $m \to \infty$, as $n \to \infty$. The sizes of the three sets therefore are $12, 22, 49, 88, 159, 288, 406, 519$, and 627 for the different sequence sizes, respectively.

As described in Section 3.1 we then, for any $h = 1, \ldots, H := 10$, determine linear h-step ahead predictions for $X_{t+h,T}$ with $t + h \in \{n_0 + 1, \ldots, n_1\}$. We determine the 'stationary predictions', with coefficients estimated for a given $p = 0, \ldots, p_{\max} := 7$, from $X_{1,T}, \ldots, X_{t,T}$ by $\hat{v}^{(p,h)}_{t,T}(t)$ from Step 3 of the procedure. For simplicity, we have chosen the same $p_{\max}$ for every T. We further determine 'locally stationary predictions' where the coefficients $\hat{v}^{(p,h)}_{N,T}(t)$ are used for $p = 0, \ldots, p_{\max}$ and

$$\mathcal{N} = \{N := N_{\min} + i\lceil (N_{\max} - N_{\min})/25 \rceil \ : \ i \in \mathbb{N}, N \leq N_{\max}\},$$

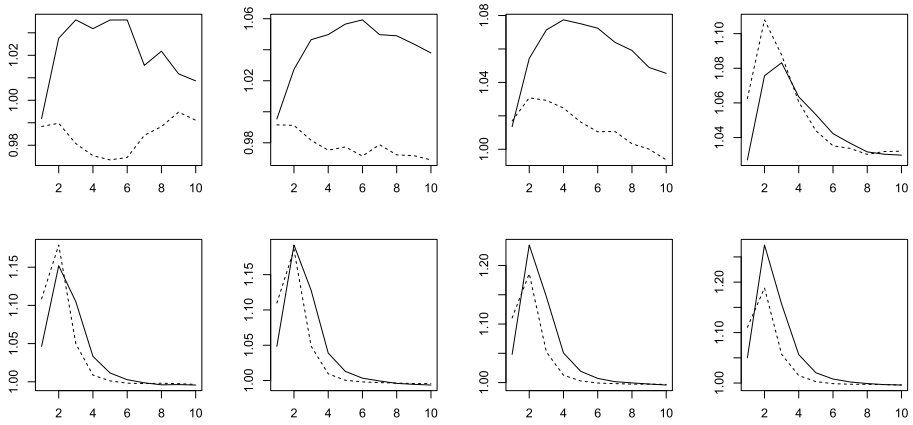

FIG 3. *Plot of $h \mapsto R_{T,i}(h)$ for model (32) and different values of n (from left to right: n=100, n=200, n=500, n=1000 [first row], n=4000, n=6000, n=8000, n=10000 [second row]). Solid line: $i = 3$ (test set), dashed line: $i = 2$ (second validation set).*

where $N_{\min} := \lfloor (n/2)^{4/5} \rfloor$ and $N_{\max} := \lfloor n^{4/5} \rfloor$. Instead of considering every integer between $N_{\min}$ and $N_{\max}$ as a possible segment size, we restrict the number $\#\mathcal{N}$ of possible values for N to a maximum of 25 elements to reduce computation time. The results did not change significantly when a larger number of elements was used. We then compare the predictors with respect to their empirical mean squared prediction error (MSPE) on the first validation set and, according to Step 4 of the procedure, choose the stationary predictor with $\hat{p}_{\text{stat.}}$ that minimises the MSPE on M_1 amongst all stationary predictors and the locally stationary predictor with $(\hat{p}_{\text{loc.}}, \hat{N}_{\text{loc.}})$ which minimizes the MSPE on M_1 amongst all the locally stationary predictors.

For those two predictors we then determine the empirical mean squared prediction errors $\text{MSPE}^*_{T,2}$ and $\text{MSPE}^*_{T,3}$, defined in (8), on the validation and test set, respectively. We record seven pieces of information: $\hat{p}_{\text{stat.}}$, $\hat{p}_{\text{loc.}}, \hat{N}_{\text{loc.}}$, $\text{MSPE}^{\text{stat.}}_{T,2}$, $\text{MSPE}^{\text{loc.}}_{T,2}$, $\text{MSPE}^{\text{stat.}}_{T,3}$, and $\text{MSPE}^{\text{loc.}}_{T,3}$. We replicate the experiment 10000 times.

Now we define the first two models. Both are tvAR(1) models defined by two periodic coefficient functions, namely the models are

$$X_{t,T} = (0.8 + 0.19 \sin(4\pi \frac{t}{T}))X_{t-1,T} + Z_t, \tag{32}$$

$$X_{t,T} = (0.3 + 0.19 \sin(4\pi \frac{t}{T}))X_{t-1,T} + Z_t. \tag{33}$$

The innovations Z_t are i.i.d Gaussian white noise. In this section we discuss the above two models in detail. The remaining processes are defined in Appendix J [32], where to also the corresponding tables and figures for them are being deferred.

In Figure 3, note that, since in the numerator we have the MSPE for the

TABLE 1
Proportions of the individual events in (34) for the process (32) and selected combinations of n and δ.

n		$R^{s,ls}_{T,2}(1)$ ≥ 1.01	$R^{s,ls}_{T,2}(1)$ < 1.01		$R^{s,ls}_{T,2}(5)$ ≥ 1.01	$R^{s,ls}_{T,2}(5)$ < 1.01
100	$R^{s,ls}_{T,3}(1) \geq 1.01$	0.1825	0.2777	$R^{s,ls}_{T,3}(5) \geq 1.01$	0.1747	0.2479
	$R^{s,ls}_{T,3}(1) < 1.01$	0.1888	0.351	$R^{s,ls}_{T,3}(5) < 1.01$	0.1424	0.435
n		$R^{s,ls}_{T,2}(1)$ ≥ 1.2	$R^{s,ls}_{T,2}(1)$ < 1.2		$R^{s,ls}_{T,2}(5)$ ≥ 1.2	$R^{s,ls}_{T,2}(5)$ < 1.2
1000	$R^{s,ls}_{T,3}(1) \geq 1.2$	5e-04	0.0055	$R^{s,ls}_{T,3}(5) \geq 1.2$	0.0758	0.0636
	$R^{s,ls}_{T,3}(1) < 1.2$	0.0063	0.9877	$R^{s,ls}_{T,3}(5) < 1.2$	0.0699	0.7907
n		$R^{s,ls}_{T,2}(1)$ ≥ 1	$R^{s,ls}_{T,2}(1)$ < 1		$R^{s,ls}_{T,2}(5)$ ≥ 1	$R^{s,ls}_{T,2}(5)$ < 1
10000	$R^{s,ls}_{T,3}(1) \geq 1$	0.9916	0	$R^{s,ls}_{T,3}(5) \geq 1$	0.7567	0.2054
	$R^{s,ls}_{T,3}(1) < 1$	0.0084	0	$R^{s,ls}_{T,3}(5) < 1$	0.0251	0.0128
n		$R^{s,ls}_{T,2}(1)$ ≥ 1.05	$R^{s,ls}_{T,2}(1)$ < 1.05		$R^{s,ls}_{T,2}(5)$ ≥ 1.05	$R^{s,ls}_{T,2}(5)$ < 1.05
10000	$R^{s,ls}_{T,3}(1) \geq 1.05$	0.4917	4e-04	$R^{s,ls}_{T,3}(5) \geq 1.05$	0.0019	0.1698
	$R^{s,ls}_{T,3}(1) < 1.05$	0.5077	2e-04	$R^{s,ls}_{T,3}(5) < 1.05$	0.0025	0.8258
n		$R^{s,ls}_{T,2}(1)$ ≥ 1.1	$R^{s,ls}_{T,2}(1)$ < 1.1		$R^{s,ls}_{T,2}(5)$ ≥ 1.1	$R^{s,ls}_{T,2}(5)$ < 1.1
10000	$R^{s,ls}_{T,3}(1) \geq 1.1$	0.0033	9e-04	$R^{s,ls}_{T,3}(5) \geq 1.1$	1e-04	0.0119
	$R^{s,ls}_{T,3}(1) < 1.1$	0.7025	0.2933	$R^{s,ls}_{T,3}(5) < 1.1$	1e-04	0.9879
n		$R^{s,ls}_{T,2}(1)$ ≥ 1.15	$R^{s,ls}_{T,2}(1)$ < 1.15		$R^{s,ls}_{T,2}(5)$ ≥ 1.15	$R^{s,ls}_{T,2}(5)$ < 1.15
10000	$R^{s,ls}_{T,3}(1) \geq 1.15$	0	0	$R^{s,ls}_{T,3}(5) \geq 1.15$	0	7e-04
	$R^{s,ls}_{T,3}(1) < 1.15$	0.0188	0.9812	$R^{s,ls}_{T,3}(5) < 1.15$	2e-04	0.9991

best stationary predictor and in the denominator the MSPE for the best locally stationary predictor, a ratio above 1 corresponds to the situation where the best locally stationary predictor outperforms the best stationary predictor. It can be seen whether this happens on average, while in Table 2 we can see the proportion of simulated cases in which this has happens. In Figure 3, we thus observe that, for $n = 100$, the stationary approach performs better on average across all values of h on both the test and the second validation set. For $n = 200$ the locally stationary approach performs better for $3 \leq h \leq 6$ on the test set, while the stationary approach still excels for all h on the second validation set. For $n \geq 500$ the locally stationary approach is better across all values of h on the test set and for $2 \leq h \leq 4$ it outperforms the stationary approach on the second validation set. For $n \geq 1000$ the locally stationary approach is always as least as good as the stationary approach for all h. It is striking that, for this particular model and for the larger n's we see that as h gets larger the two approaches (stationary and locally stationary) perform almost equally well on average, which can be seen from the lines in Figure 3 being close to one. Another important observation is that, as n gets larger and m/n gets smaller, we see the lines for the validation and test set converging, which is in line with what Theorem 3.1 suggests should happen.

We now, briefly, compare the outcome of model (32) to that of model (33); details are shown in Appendix J [32]. Note that in model (32) the coefficient

TABLE 2
Proportion of (7) being fulfilled for the process (32) and different values of h, δ and n.

	(34) holds for $h = 1$, $i = 2$								
δ	n=100	n=200	n=500	n=1000	n=2000	n=4000	n=6000	n=8000	n=10000
0	0.4182	0.4414	0.5873	0.8844	0.9962	0.9999	1	1	1
0.01	0.3713	0.3821	0.5152	0.8422	0.993	0.9999	1	1	1
0.05	0.2431	0.2172	0.27	0.5986	0.9106	0.9855	0.9958	0.9985	0.9994
0.1	0.1518	0.1029	0.0893	0.2347	0.4748	0.6212	0.67	0.6995	0.7058
0.15	0.0983	0.0483	0.024	0.0475	0.071	0.0617	0.0449	0.0305	0.0188
0.2	0.0622	0.0233	0.0072	0.0068	0.0046	6e-04	1e-04	0	0
0.4	0.0138	0.0026	8e-04	2e-04	5e-04	1e-04	0	0	0
0.6	0.0059	0.0015	4e-04	2e-04	4e-04	0	0	0	0
	(34) holds for $h = 1$, $i = 3$								
δ	n=100	n=200	n=500	n=1000	n=2000	n=4000	n=6000	n=8000	n=10000
0	0.4968	0.5118	0.5989	0.6929	0.8099	0.9279	0.9681	0.9829	0.9916
0.01	0.4602	0.4671	0.5436	0.6357	0.7485	0.8725	0.9276	0.9559	0.9751
0.05	0.3357	0.3103	0.3372	0.3905	0.416	0.4577	0.4776	0.4732	0.4921
0.1	0.2292	0.1745	0.1481	0.1427	0.0954	0.0458	0.0195	0.0113	0.0042
0.15	0.1487	0.0942	0.0536	0.0313	0.0098	3e-04	2e-04	0	0
0.2	0.0983	0.0491	0.0194	0.006	7e-04	1e-04	0	0	0
0.4	0.0209	0.006	0.0013	3e-04	5e-04	1e-04	0	0	0
0.6	0.0078	0.0023	7e-04	2e-04	5e-04	1e-04	0	0	0
	(34) holds for $h = 5$, $i = 2$								
δ	n=100	n=200	n=500	n=1000	n=2000	n=4000	n=6000	n=8000	n=10000
0	0.3365	0.3525	0.3646	0.323	0.1831	0.1944	0.2508	0.3045	0.359
0.01	0.3171	0.3246	0.3416	0.3137	0.1648	0.1409	0.1642	0.1853	0.2101
0.05	0.2535	0.236	0.2609	0.2798	0.1216	0.0307	0.0121	0.0073	0.0044
0.1	0.1968	0.159	0.18	0.2378	0.0985	0.0132	0.0013	3e-04	2e-04
0.15	0.1541	0.1101	0.1235	0.1929	0.0844	0.012	0.0011	1e-04	2e-04
0.2	0.1244	0.0757	0.0843	0.1457	0.0696	0.0101	9e-04	1e-04	2e-04
0.4	0.0575	0.0207	0.0164	0.0263	0.0101	0.0013	1e-04	0	0
0.6	0.0321	0.0068	0.0032	0.0033	4e-04	0	0	0	0
	(34) holds for $h = 5$, $i = 3$								
δ	n=100	n=200	n=500	n=1000	n=2000	n=4000	n=6000	n=8000	n=10000
0	0.4423	0.5487	0.5253	0.3178	0.2155	0.2979	0.3955	0.4736	0.5393
0.01	0.4226	0.5269	0.5054	0.3086	0.1943	0.2561	0.3474	0.4249	0.4931
0.05	0.3563	0.4417	0.4281	0.2705	0.1312	0.0999	0.1168	0.1477	0.1717
0.1	0.2823	0.3492	0.3378	0.2245	0.0859	0.0243	0.0153	0.0119	0.012
0.15	0.2291	0.2779	0.2602	0.1806	0.0625	0.0098	0.0023	8e-04	7e-04
0.2	0.1864	0.2192	0.1923	0.1394	0.0451	0.0065	8e-04	1e-04	0
0.4	0.0815	0.0823	0.0561	0.0398	0.0081	5e-04	0	0	0
0.6	0.0406	0.0303	0.0164	0.0071	4e-04	0	0	0	0

function ranges from 0.61 to 0.99, placing some of its tangent processes close to the unit root. In model (33) the coefficient function ranges from 0.11 to 0.49. Thus, the two models have the same variation of the coefficient function, but in model (33) the tangent processes are further away from the unit root. In Figure 15, it can be seen that the stationary approach is preferred over the locally stationary approach for sequences up to length $n = 1000$. Further, we observe that the advantage of using the locally stationary approach for sequences of length $n \geq 4000$ is minuscule and visible only for 1-step ahead forecasting. For the other models we can make similar observations:

Rules of Thumb. The locally stationary approach outperforms the stationary approach only if either the sequence is long, or the coefficient function exhibits considerable variation, or the tangent processes (cf. the comment after Assumptions 1–5) are close to the unit root. In any other case the stationary approach can be chosen without (a large) loss.

Our observation that the locally stationary forecast performs better when the stationary approximations are near unit root may possibly be explained by the fact that the coefficient of a near unit root AR(1) process can be estimated at a better rate than in the classical case where the rate is $T^{-1/2}$; cf. [13, 24]. A rigorous analysis of the issue is beyond the scope of this paper and left for future research.

TABLE 3
Proportion of (34) being fulfilled for the process (32) and different values of h, δ and n.

	(34) holds for $h = 1$								
δ	n=100	n=200	n=500	n=1000	n=2000	n=4000	n=6000	n=8000	n=10000
0	0.5254	0.5176	0.5088	0.6377	0.8071	0.9278	0.9681	0.9829	0.9916
0.01	0.5335	0.5308	0.4984	0.5877	0.7441	0.8724	0.9276	0.9559	0.9751
0.05	0.606	0.6265	0.5928	0.4899	0.4346	0.46	0.4776	0.4737	0.4919
0.1	0.7114	0.7714	0.7972	0.7034	0.53	0.395	0.3381	0.3054	0.2966
0.15	0.8028	0.8743	0.9284	0.9256	0.9214	0.9384	0.9549	0.9695	0.9812
0.2	0.8667	0.9358	0.9756	0.9882	0.9957	0.9995	0.9999	1	1
0.4	0.9727	0.9936	0.9995	0.9999	1	1	1	1	1
0.6	0.9903	0.9982	0.9993	1	0.9999	0.9999	1	1	1
	(34) holds for $h = 5$								
δ	n=100	n=200	n=500	n=1000	n=2000	n=4000	n=6000	n=8000	n=10000
0	0.6054	0.5684	0.6755	0.8888	0.879	0.7975	0.7729	0.7699	0.7695
0.01	0.6097	0.5711	0.673	0.8851	0.8819	0.783	0.7324	0.6858	0.6538
0.05	0.633	0.5991	0.6752	0.8705	0.9246	0.9008	0.8801	0.8492	0.8277
0.1	0.6817	0.6582	0.7082	0.8601	0.9546	0.9825	0.9854	0.988	0.988
0.15	0.7252	0.715	0.7547	0.8583	0.9551	0.9938	0.998	0.9993	0.9991
0.2	0.7664	0.7677	0.807	0.8665	0.9553	0.9944	0.9991	1	0.9998
0.4	0.8816	0.9046	0.9373	0.9463	0.9872	0.9984	0.9999	1	1
0.6	0.9337	0.9647	0.9814	0.99	0.9992	1	1	1	1

TABLE 4
Values of $q(\delta)$, defined in (25), for the process (32) and different values of h, δ and n.

	Value of $q(\delta)$, defined in (25), for $h = 1$								
δ	n=100	n=200	n=500	n=1000	n=2000	n=4000	n=6000	n=8000	n=10000
0	0	0	0	0	0	0	0	0	0
0.01	6.2e-05	1.1e-05	7.6e-05	0.02	0.02	0.021	0.021	0.021	0.021
0.05	0.00018	1.7e-05	2e-05	6.4e-05	0.012	0.0087	0.0054	0.0032	0.0014
0.1	0.023	6.4e-05	2.6e-05	3.7e-06	0.0056	0.011	0.015	0.017	0.018
0.15	0.076	0.051	0.044	0.049	0.056	0.062	0.066	0.067	0.069
0.2	0.13	0.1	0.094	0.1	0.11	0.11	0.12	0.12	0.12
0.4	0.2	0.28	0.3	0.3	0.31	0.31	0.32	0.32	0.32
0.6	3e-04	0.054	0.19	0.33	0.42	0.47	0.49	0.51	0.52
	Value of $q(\delta)$, defined in (25), for $h = 5$								
δ	n=100	n=200	n=500	n=1000	n=2000	n=4000	n=6000	n=8000	n=10000
0	0	0	0	0	0	0	0	0	0
0.01	3.3e-06	0.00029	0.00041	6.5e-05	0.02	0.021	0.021	0.021	0.021
0.05	0.00016	4.9e-06	0.001	0.042	0.034	0.026	0.021	0.018	0.015
0.1	0.005	0.0015	3.9e-05	0.14	0.13	0.13	0.12	0.12	0.12
0.15	0.094	5.5e-05	0.00022	0.055	0.16	0.13	0.12	0.11	0.1
0.2	0.2	0.00088	0.00018	0.00091	0.058	0.031	0.015	0.0046	0.0035
0.4	4e-04	0.05	0.099	0.14	0.17	0.2	0.22	0.23	0.24
0.6	0.33	0.4	0.46	0.52	0.56	0.6	0.62	0.64	0.65

The proportions shown in Table 3 provide information on the consistency of the procedure, as we see the proportion of cases in which the same procedure (stationary or locally stationary) is chosen on both the test set and the second validation set. This validates Theorem 3.1 for the example. It is interesting to compare the observed proportions with the corresponding value of $q(\delta)$, which we provide in Table 4. We see that a larger proportion typically goes along with a larger value of $q(\delta)$ indicating the relevance of condition (25). To make it more precise: the tables are concerned with the proportion for which the decision rule (7) yields the same result no matter if we take $i = 2$ or $i = 3$, i.e. we count what proportion of runs satisfies

$$(\hat{R}_{T,2}(h) \geq 1+\delta \text{ and } \hat{R}_{T,3}(h) \geq 1+\delta) \text{ or } (\hat{R}_{T,2}(h) < 1+\delta \text{ and } \hat{R}_{T,3}(h) < 1+\delta). \tag{34}$$

We see that if δ is chosen large enough then the probability for the event (34) approaches 1, as T and m increase. More precisely, this is the case, if δ is chosen smaller than the ratio of MSPEs depicted in Figure 3 on both the validation and test set or larger than both those ratios. This is as expected from Corollary 3.2 and 3.3. A more detailed analysis is possible, employing the information provided

in Table 1. In the third row of tables we see, for example, that for $n = 10000$ and $\delta = 0$ the procedure will consistently choose the locally stationary approach on both the test set and the second validation set for 1-step ahead forecasting. For $n = 10000$ and $\delta = 0.05$, on the other hand, we see that the procedure almost consistently chooses the locally stationary approach on the validation set while it is rather undecided (50%-50%) on the test set. For $\delta = 0.1$ the procedure almost consistently chooses the stationary approach on the test set and is to some degree undecided (70%-30%) on the second validation set. Finally, if $\delta = 0.15$, we see that the stationary approach gets chosen almost consistently on both validation and test sets. This is just what we would expect, as a smaller δ must lead to the locally stationary approach being preferred, as the more complex locally stationary approach only gets selected if the empirical MSPE of the stationary approach is at least $(1 + \delta)$-times of the empirical MSPE of the locally stationary approach.

The remaining part of the simulation studied is deferred to Section J [32].

5. Data examples

5.1. London housing prices

We analyse average housing prices from the UK House Price Index (HPI). The HPI is updated monthly with data from the Land Registry, the Registers of Scotland, and the Land and Property Services Northern Ireland. The data is combined by the Office of National Statistics using hedonic regression; cf. [34]. The sequence we used for the analysis contains 264 monthly index values from 1995 to 2016. It was obtained as follows: In the 'customise your search' part of the 'search the UK house price index' form we have selected the 'English region' London, the period from 01-1995 to 12-2016, and then obtained the 'average price' for 'all property types'. The data is depicted in the left panel of Figure 4. For the analysis we consider $T + m = 263$ monthly changes (in percent). The prices are centred by subtracting the arithmetic mean prior to the analysis. We clearly see autocorrelation at lags less or equal than 4 and at lag 12 in the right panel of Figure 4.

We then compute the 1-step to 6-step prediction coefficients, defined in (4), with which we can predict an observation X_{t+h} from $X_t, \ldots, X_{t-p+1}$, where X_{t+h} is an observation made either in 2014, 2015 or 2016, respectively. We choose $p = 0, 1, \ldots, 18$, where $p = 0$ shall mean that we are predicting with 0. Note that the maximum p was chosen larger than 12, as we are dealing with monthly data and dependence at lag 12 can be seen from the autocorrelation function. We consider the stationary predictors as well as locally stationary predictors with $N = 50, 51, \ldots, 87 = \lceil 263^{4/5} \rceil$.

Interestingly, in Figure 5, we observe that the MSPE of the locally stationary forecasts are typically larger than corresponding ones of the stationary forecasts.

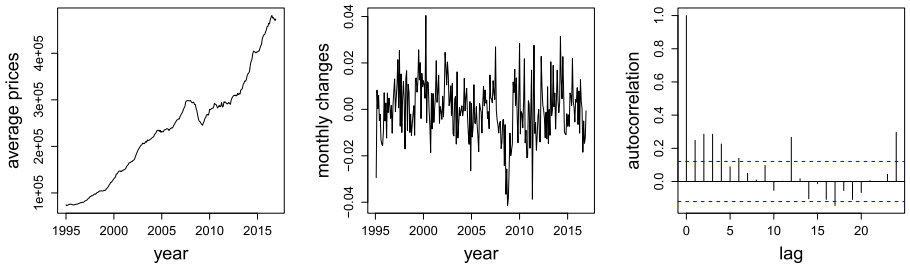

FIG 4. *Data for London from the UK House Price Index. Left: average monthly housing prices. Middle: monthly changes of average housing prices in percent, demeaned by subtracting arithmetic mean. Right: autocorrelation function, computed from the sequence in the middle.*

TABLE 5

Minimum empirical mean squared prediction errors (MSPEs) for h-step ahead prediction, $h = 1, \ldots, 6$, of the house price data. Top table shows values computed on the first validation set. Bottom table shows values computed on the second validation set and on the test set.

h	$\hat{p}_{\text{stat.}}$	$\text{MSPE}^{\text{stat.}}_{T,1}(h)$	$\hat{p}_{\text{loc.}}$	$\hat{N}_{\text{loc.}}$	$\text{MSPE}^{\text{loc.}}_{T,1}(h)$
1	18	8.033024e-05	18	73	7.701586e-05
2	18	8.547987e-05	18	72	9.027318e-05
3	18	9.362087e-05	18	71	9.512262e-05
4	18	1.079008e-04	18	71	1.039368e-04
5	18	1.164369e-04	18	87	1.291897e-04
6	18	1.097551e-04	18	86	1.160201e-04

h	$\text{MSPE}^{\text{stat.}}_{T,2}(h)$	$\text{MSPE}^{\text{loc.}}_{T,2}(h)$	$\hat{R}_{T,2}(h)$	$\text{MSPE}^{\text{stat.}}_{T,3}(h)$	$\text{MSPE}^{\text{loc.}}_{T,3}(h)$	$\hat{R}_{T,3}(h)$
1	3.473298e-05	3.501655e-05	0.992	9.740925e-05	0.0001385059	0.703
2	3.560845e-05	4.308688e-05	0.826	9.547598e-05	0.0001351634	0.706
3	4.31916e-05	4.21518e-05	1.025	0.0001052688	0.0001309526	0.804
4	4.57004e-05	4.429208e-05	1.032	0.0001053983	0.0001421635	0.741
5	5.970928e-05	4.943228e-05	1.208	0.0001210628	0.0001195622	1.012
6	6.412237e-05	5.234349e-05	1.225	0.0001152908	0.0001146555	1.006

As described in our procedure we now determine the $\hat{p}_{\text{stat.}}$, $\hat{p}_{\text{loc.}}$, and $\hat{N}$ that minimise the MSPE within each class of predictors. For 1-step ahead prediction we find $\hat{p}_{\text{stat.}} = 18$, $\hat{p}_{\text{loc.}} = 18$, and $\hat{N} = 73$. For 6-step ahead prediction we find $\hat{p}_{\text{stat.}} = 18$, $\hat{p}_{\text{loc.}} = 18$, and $\hat{N} = 86$. The numbers are summarised in Table 5.

We then determine the MSPE for forecasting the observations from the second validation set (here: the year 2015) using these predictors. For 1-step ahead prediction we find that $\text{MSPE}^{\text{stat.}}_{251,2}(1) = 3.47 \cdot 10^{-5}$ and $\text{MSPE}^{\text{loc.}}_{251,2}(1) = 3.50 \cdot 10^{-5}$, with $\text{MSPE}^{*}_{T,j}(h)$ defined in (8). For 6-step ahead prediction we find that $\text{MSPE}^{\text{stat.}}_{251,2}(6) = 6.41 \cdot 10^{-5}$ and $\text{MSPE}^{\text{loc.}}_{251,2}(6) = 5.23 \cdot 10^{-5}$. Consequently, we decide to use the stationary approach for 1-step and the locally stationary approach for 6-step ahead forecasting of the observations made in 2016.

The MSPEs computed from 1-step ahead forecasting the observations from the test set (here: the year 2016) are $\text{MSPE}^{\text{stat.}}_{251,3}(1) = 9.74 \cdot 10^{-5}$ and $\text{MSPE}^{\text{loc.}}_{251,3}(1) = 1.39 \cdot 10^{-4}$. The MSPEs computed from 6-step ahead forecasting the observations from 2016 are $\text{MSPE}^{\text{stat.}}_{251,3}(6) = 1.153 \cdot 10^{-4}$ and $\text{MSPE}^{\text{loc.}}_{251,3}(6) = 1.147 \cdot 10^{-4}$. We have thus chosen the better performing procedure for 1-step and 6-step ahead forecasting.

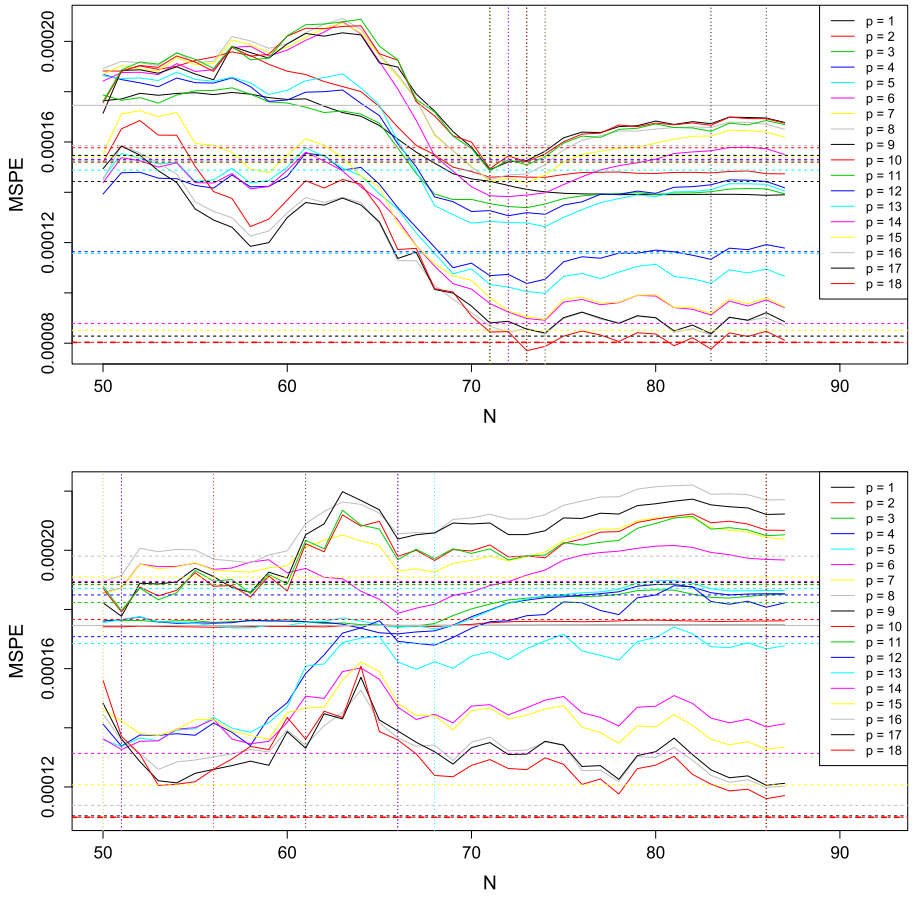

FIG 5. *Empirical mean squared prediction errors (MSPEs) computed on the frist validation set (predicting the 12 observations from 2014). Top panel shows MSPEs for 1-step ahead prediction. Bottom panel shows MSPEs for 6-step ahead prediction. The colours indicate which p was used. The colour code is described in the plot's legend. The solid lines correspond to the MSPEs for different N when the locally stationary approach is used. The dashed lines show the MSPE when the stationary approach is used. The horizontal grey line indicates the MSPE for the trivial forecasts* ($f^{\text{loc.}}_{t,h;0,N}$ *and* $f^{\text{stat.}}_{t,h;0}$). *The MSPE in this case is* 0.000175.

In conclusion, our analysis has revealed that, from the point of view of 1-month ahead prediction of the 2016 observations, treating the data as stationary does not have a negative effect. We were able to see that using the estimates from the stationary AR(18) model gave us better predictions than using the (locally stationary) estimates of segments of 73 month (roughly 6 years). For the 6-month ahead prediction the local estimates are better, but only by a small margin. Contrary to what one might naively expect, the impact of, for example, the 2008-2009 financial crisis on the stationary estimates is not profound enough to substantially worsen the predictors' performance.

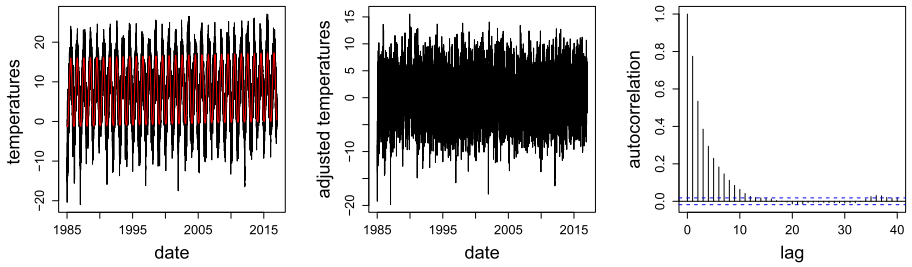

FIG 6. *Temperature data Hohenpeißenberg. Left: daily temperatures and fitted harmonic regression model. Middle: adjusted data (demeaned and detrended). Right: autocorrelation function, computed from the sequence in the middle.*

5.2. Temperatures Hohenpeißenberg

In this example, we analyse seasonally adjusted, daily temperature data collected at the meteorological observatory in Hohenpeißenberg (Germany). More precisely, we use $n = T + m = 11680$ observations of daily mean temperatures that were recorded between 1985 and 2016.[1] The data are shown in the left panel of Figure 6. To eliminate the clearly visible trend and seasonality, we have fitted a harmonic linear regression model of the form

$$y_t = c + \alpha t + \sum_{i=1}^{4} \Big(\beta_i \sin(2\pi t i/365) + \gamma_i \cos(2\pi t i/365) \Big),$$

to capture the trend and annual variation. The red curve in the left panel of Figure 6 is the prediction of the fitted model. We then consider the residuals of this model which are shown in the middle panel of Figure 6. The right panel of Figure 6 shows the autocorrelation function, which clearly indicates that serial dependence is present. [9] analyse the same data set and fit a stationary ARMA(3,1) model to capture the serial dependence.

In Figure 7, the MSPE are presented in the same manner as in Section 5.1. In this example we have chosen $p_{\max} = 10$ and $\mathcal{N} := \{365, 366 \ldots, \lceil n^{4/5} \rceil\} = \{365, 366, \ldots, 1794, 1795\}$ and $m := 365$. The MSPE corresponding to $p = 0$ is 110.2 in this example and therefore not visible in the plot.

By minimising the empirical MSPE on the first validation set the procedure chooses, for the stationary approach $\hat{p}_{\text{stat.}} = 2$ for $h = 1, 2$. For the locally stationary approach the procedure chooses $(\hat{p}_{\text{loc.}}, \hat{N}_{\text{loc.}}) = (3, 910)$ and $(\hat{p}_{\text{loc.}}, \hat{N}_{\text{loc.}}) = (2, 985)$ for $h = 1$ and $h = 2$, respectively. Empirical MSPEs for other values of p and N are shown in Figure 7. The numbers are summarised in Table 6.

For 1-step ahead forecasting and on validation 2 set this yields, given the $\hat{p}_{\text{stat.}}$ chosen by the procedure, that $\text{MSPE}^{\text{stat.}}_{11315,2}(1) = 8.11$ and, given the $\hat{p}_{\text{loc.}}$

[1] The data was obtained from http://www.dwd.de/DE/klimaumwelt/cdc/cdc_node.html.

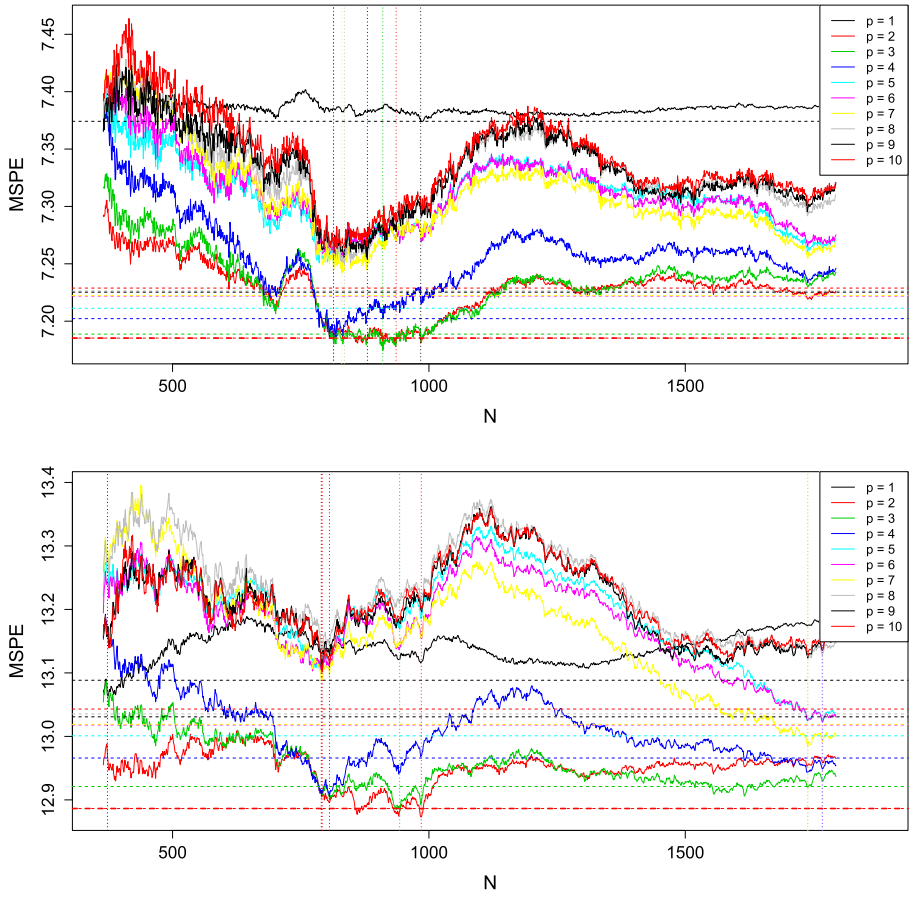

FIG 7. *Empirical mean squared prediction errors (MSPEs) computed on the first validation set (predicting the 365 observations from 2014) of the temperature data. Top panel shows MSPEs for 1-step ahead prediction. Bottom panel shows MSPEs for 2-step ahead prediction. The colours indicate which p was used. The colour code is described in the plot's legend. The solid lines correspond to the MSPEs for different N when the locally stationary approach is used. The dashed lines show the MSPE when the stationary approach is used.*

and N chosen by the procedure, that $\mathrm{MSPE}^{\mathrm{loc.}}_{11315,2}(1) = 8.06$. Similarly, for 2-step ahead forecasting, we have $\mathrm{MSPE}^{\mathrm{stat.}}_{11315,2}(2) = 14.87$ and $\mathrm{MSPE}^{\mathrm{loc.}}_{11315,2}(2) = 14.94$. The respective ratios are both very close to 1. The procedure thus chooses the stationary approach over the locally stationary approach if $\delta = 0.01$ is chosen and, obviously, this superiority will continue to hold if δ is chosen larger than that. On the test set we have $\mathrm{MSPE}^{\mathrm{stat.}}_{11315,3}(1) = 8.09$ and $\mathrm{MSPE}^{\mathrm{loc.}}_{11315,3}(1) = 7.97$ for 1-step ahead forecasting. Likewise, for 2-step ahead forecasting, we have $\mathrm{MSPE}^{\mathrm{stat.}}_{11315,3}(2) = 15.43$ and $\mathrm{MSPE}^{\mathrm{loc.}}_{11315,3}(2) = 15.41$. Thus, again, both approaches for 1-step and 2-step ahead forecasting behave almost equally well and we see that had we chosen $\delta > 0.015$ our procedure chose the stationary

TABLE 6

Minimum empirical mean squared prediction errors (MSPEs) for h-step ahead prediction, $h = 1, 2, 3, 4, 5$, of the temperature data Hohenpeißenberg. Top table shows values computed on the first validation set. Bottom table shows values computed on the second validation set and on the test set.

h	$\hat{p}_{\text{stat.}}$	$\text{MSPE}^{\text{stat.}}_{T,1}(h)$	$\hat{p}_{\text{loc.}}$	$\hat{N}_{\text{loc.}}$	$\text{MSPE}^{\text{loc.}}_{T,1}(h)$
1	2	7.185208	3	910	7.173272
2	2	12.886257	2	985	12.870544
3	2	15.397509	2	870	15.343298
4	2	16.605640	2	800	16.504915
5	2	17.226943	2	800	17.093823

h	$\text{MSPE}^{\text{stat.}}_{T,2}(h)$	$\text{MSPE}^{\text{loc.}}_{T,2}(h)$	$\hat{R}_{T,2}(h)$	$\text{MSPE}^{\text{stat.}}_{T,3}(h)$	$\text{MSPE}^{\text{loc.}}_{T,3}(h)$	$\hat{R}_{T,3}(h)$
1	8.10974	8.058095	1.006	8.08899	7.967895	1.015
2	14.86848	14.94354	0.995	15.42535	15.39907	1.001
3	17.72551	17.92775	0.989	17.4254	17.3617	1.004
4	19.63724	19.8143	0.991	17.68487	17.60241	1.005
5	20.97236	21.0989	0.994	17.92979	17.83498	1.005

approach, which performs almost equally well as the more complicated locally stationary approach.

In conclusion, in this example, we have provided clear evidence that the temperature data, after adjusting for trend and seasonality, collected in the Hohenpeißenberg observatory, from the point of view of prediction, can be treated as if they were stationary. We see that using the estimates related to a AR(2) [or AR(3)] model yielded forecasts that in all cases perform almost equally well as the estimates localised to the segment suggested by the procedure (using the past 2.2–2.7 years; 800–910 days). This observation is remarkable, in the sense that, in 30 years of data an analyst might typically expect non-stationarity (e.g., changes due to global warming) to worsen the predictions. Our conclusion indicates that the variation of covariance structure might be less substantial than the change in mean. Note that our procedure did not consistently chose the approach with the better performance on the test set, but that both approaches perform almost equally well on either set. It is thus legitimate to use the simpler, stationary approach.

5.3. Volatility around the time of the EU referendum in the UK, 2016

This example is about forecasting volatility of the FTSE 100 stock index. More precisely, we consider a sequence of $n = T + m = 607$ (daily) opening prices p_{open} and closing prices p_{close}, dated from 2 January 2015 to 26 May 2017.[2] The analysis is then based on the sequence $((p_{\text{close}} - p_{\text{open}})/p_{\text{close}})^2$, centred by subtracting the arithmetic mean of this sequence. The data are shown in Figure 8.

We separate the final 60 observations of the data as test set, first validation set and second validation set (used for determining the model orders and segment sizes). Each set is of size $m := 20$. Visual inspection of these 60 observations suggested that some returns are unusually small or large. Indeed, the returns

[2]The data was obtained from http://www.finanzen.net/index/FTSE_100/Historisch.

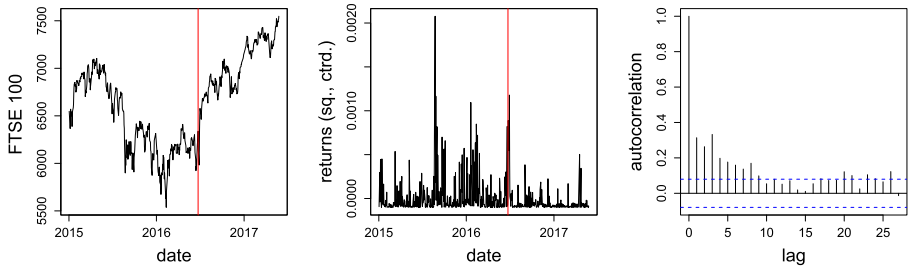

FIG 8. *Volatility of the FTSE 100 Index, for 2 January 2015 to 26 May 2017. Left: FTSE 100 closing price. Middle: squared and centred returns. Right: autocorrelation function, computed from the sequence in the middle. Red vertical line in the left and middle plot marks 23/06/2016, the day of the EU referendum in the UK.*

of 1 March, 18 April, and 24 April 2017 are either more than 1.5 times the interquartile range (IQR) smaller than the lower quartile or 1.5 times the IQR larger than the upper quartile. By Tukey's criterion they can thus be classified as outliers. To better deal with the outliers, we use a robustified measure of accuracy to compare the forecasts in this example. More precisely, instead of the MSPE in Steps 4 and 5 of our procedure, we now use an empirical trimmed mean of absolute prediction errors (trMAPE), where we trim the largest 25%, averaging only the remaining 15 out of 20 absolute errors. We have further chosen $p_{\max} = 8$ and $\mathcal{N} := \{40, 41, \ldots, 250\}$.

First, we consider the trMAPEs of forecasting the 20 observations from the first validation set to determine the optimal $\bar{p}_{\text{stat.}}$, $\bar{p}_{\text{loc.}}$ and $\bar{N}_{\text{loc.}}$. We use a bar instead of the hat to indicate that the trMAPEs were used instead of the MSPEs. In Figure 9 we can see, for the 1-step, 2-step and 3-step ahead forecasts, that the lines depicting the trMAPEs have a characteristic shape: as N increases the trMAPEs slightly decreases (for each p at a different level) until it starts increasing around $N \approx 60$. After this follows another phase of slight decreasing and increasing with the new minimum higher than the minimum of the previous phase. We further observe that the overall level is typically lower than that of the trMAPEs of the stationary approach. The last such minimum in our plots is obtained when N is around 170–180.

The observations 1 through to 373 were recorded from 2 January 2015 to 23 June 2016 (the day of the EU referendum) and observations 374 through to 607 were recorded from 24 June 2016 to 26 May 2017. This implies that the final 234 observations were recorded after the EU referendum, meaning that there are 175 observations between the EU referendum and the observations to be forecast in the first step. Thus, the last minimum of the lines, when N is roughly about 170, corresponds to the time of the referendum. The sudden increase of the trMAPE indicates the change in bias of the Yule Walker estimator due to non-stationarity when pre-referendum data is starting to be used for the estimation of the prediction coefficients. Another important observation is that also the post-referendum part of the diagram ($40 \leq N \leq 175$) shows signs of

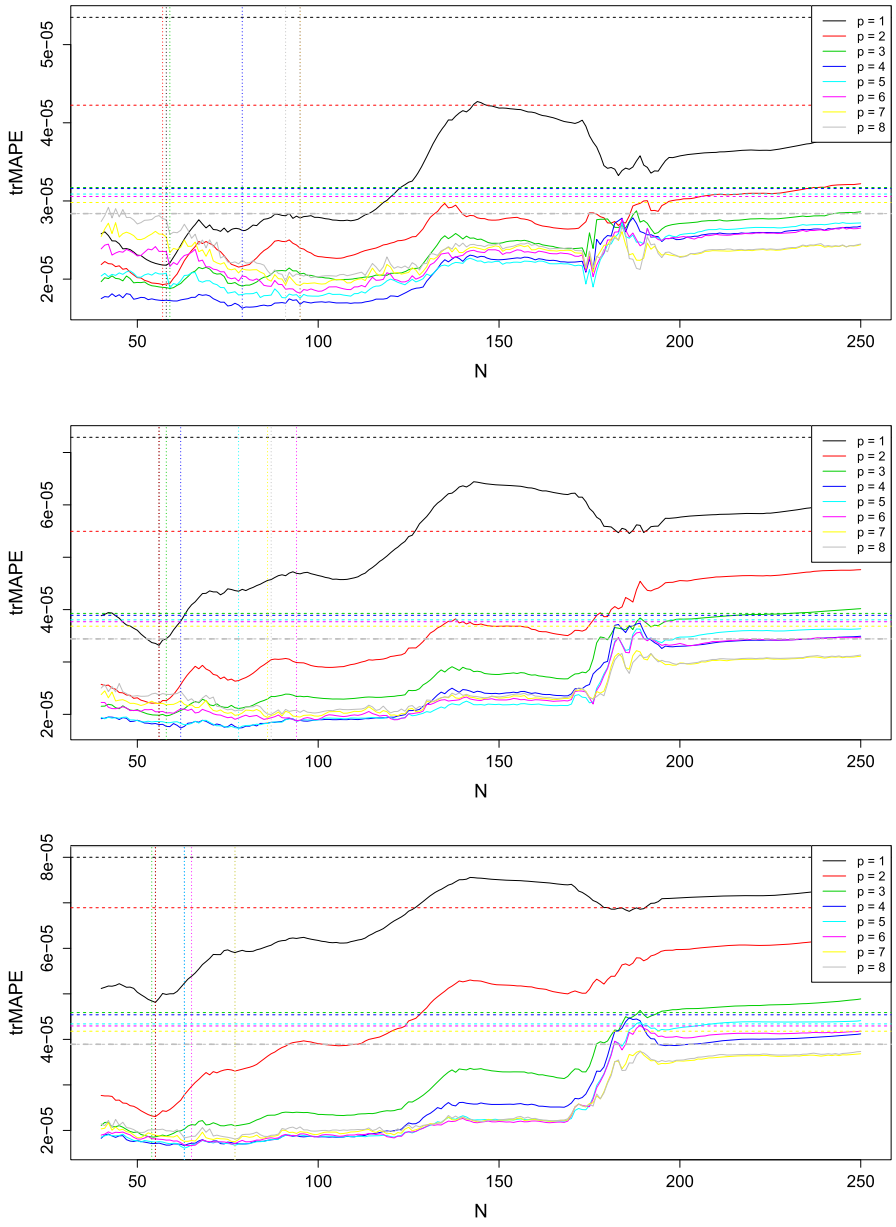

FIG 9. *Empirical trimmed mean absolute prediction errors (trMAPE) computed on the first validation set (predicting the observations 548 to 567) of the squared and centred FTSE returns. Top, middle and bottom panel show the trMAPE for the 1, 2 and 3-step ahead predictions, respectively. The colours indicate which p was used. The colour code is described in the plot's legend. The solid lines correspond to the trMAPE for different N when the locally stationary approach is used. The dashed lines show the trMAPE when the stationary approach is used. The horizontal grey line indicates the trMAPE for the trivial forecasts ($f^{\text{loc.}}_{t,h;0,N}$ and $f^{\text{stat.}}_{t,h;0}$). The trMAPE in this case is* 8.2×10^{-5}.

TABLE 7

Minimum empirical trimmed mean absolute prediction errors (trMAPE) for h-step ahead prediction, $h = 1, 2, 3, 4, 5$, of the squared and centred FTSE 100 data. Analysis performed with $m := 20$ and $p_{\max} = 8$. Top table shows values computed on the first validation set. Bottom table shows values computed on the second validation set and on the test set.

h	$\bar{p}_{\text{stat.}}$	$\text{trMAPE}^{\text{stat.}}_{T,1}(h)$	$\bar{p}_{\text{loc.}}$	$\bar{N}_{\text{loc.}}$	$\text{trMAPE}^{\text{loc.}}_{T,1}(h)$
1	8	2.838118e-05	4	79	1.628012e-05
2	8	3.440985e-05	5	78	1.736197e-05
3	8	3.892572e-05	5	63	1.610966e-05
4	8	4.786001e-05	6	65	1.731724e-05
5	8	5.161963e-05	5	53	2.209272e-05

h	$\text{trMAPE}^{\text{stat.}}_{T,2}(h)$	$\text{trMAPE}^{\text{loc.}}_{T,2}(h)$	$\bar{R}_{T,2}(h)$	$\text{trMAPE}^{\text{stat.}}_{T,3}(h)$	$\text{trMAPE}^{\text{loc.}}_{T,3}(h)$	$\bar{R}_{T,3}(h)$
1	2.838118e-05	1.628012e-05	1.743	3.33206e-05	2.395498e-05	1.391
2	3.440985e-05	1.736197e-05	1.982	3.817851e-05	2.505271e-05	1.524
3	3.892572e-05	1.610966e-05	2.416	4.369565e-05	2.750278e-05	1.589
4	4.786001e-05	1.731724e-05	2.764	4.974355e-05	2.81844e-05	1.765
5	5.161963e-05	2.209272e-05	2.336	5.390384e-05	4.560496e-05	1.182

non-stationarity. More specifically, each phase of up-movement indicate that the variance is reduced less than the squared bias increases. The increase from the first (and global) minimum at around $N \approx 60$ onwards corresponds to taking data from the end of November 2016 and earlier into account and might correspond to changes due to effects of the election of the US president. The minimum trMAPE for forecasting the data from the end of the estimation set are summarised in Table 7. We observe that for $h = 1, 2, 3, 4$ the optimum segment size is roughly 60 such that no observations prior to November 2016 are used for estimation. For $h = 5$ the optimum segment size is 41 and thus even smaller. This implies that no observation prior to the presidential election in the US are used for estimation of the forecasting coefficients.

Using these predictors to forecast the 20 observations from the second validation set we see, in Table 7, that the trMAPE of the stationary approach are typically (2.1 to 3.7 times) larger than the trMAPE of the locally stationary approach. We thus choose to work with the locally stationary approach. In Table 7 we denote the ratios of the trMAPE of the stationary approach over the trMAPE of the locally stationary approach by $\bar{R}_{T,j}(h)$, where the bar indicates that the trMAPE and not the MSPE is used. Forecasting the 20 observations from the test set we see that the trMAPEs of the stationary approach are again larger than those of the locally stationary approach, but not quite as much as on the second validation set. Still, following our procedure, we chose the better performing approach (the locally stationary one).

For this example, we further conducted a sensitivity analysis, by varying the parameters m and $p_{\max}$. Selected results, in which we see the results are mostly stable when changing the parameters are shown in Appendix I [32].

6. Analysis of the localised Yule-Walker estimator under general conditions and local stationarity

In this section we discuss the probabilistic properties of the localised Yule-Walker estimator $\hat{a}^{(p)}_{N,T}(t)$ defined in (1). We believe the results to be of inde-

pendent interest and therefore present them in this separate section. They are also key results for the proofs of the result in Section 3.3. Our results will hold under Assumptions 1–5 (cf. Section 3.3). The assumptions are not restrictive and, in particular, the concentration result in this section will hold for a broad class of locally stationary processes and, in particular, does not require that the data come from a tvAR(p) model. Further, we allow for any $1+p \leq N \leq T$ and, in particular, allow for a diverging model order p, as $T \to \infty$. We do not, as do for example [20], require that $N = o(T)$.

The main result of this section (Theorem 6.1) provides a non-asymptotic bound for the Euclidean distance of $\hat{a}^{(p)}_{N,T}(t)$ to the following population quantity:

$$\bar{a}^{(p)}_{N,T}(t) := \big(\mathbb{E}\hat{\Gamma}^{(p)}_{N,T}(t)\big)^{-1}\big(\mathbb{E}\hat{\gamma}^{(p)}_{N,T}(t)\big) = \big(\bar{a}^{(p)}_{1,N,T}(t), \ldots, \bar{a}^{(p)}_{p,N,T}(t)\big)'. \tag{35}$$

The Yule-Walker estimator is widely used in practice and $\hat{a}^{(p)}_{N,T}(t)$ and its properties have been studied in detail under various conditions. [6, 5] and [4] derive large deviation principles for Gaussian AR processes when the model order is 1. A simple exponential inequality, also for model order 1, is given in Section 5.2 of [7]. [61] prove a large deviation principle for general, but fixed, model order. [29, 30] derives simultaneous confidence bands. The cited results all require that the underlying process is stationary. [20] analyse the bias and variance of the localised Yule-Walker estimator in the framework of local stationarity. They do not, however, provide an exponential inequality, and, as far as we are aware, no result as the one we provide below is available at present. The exponential inequality in Theorem 6.1, which we now state, is explicit in terms of all parameters and constants. We make use of the explicitness to derive Corollary 6.2, by which the localised Yule-Walker estimator is strongly, uniformly consistent, even when the model order is diverging as the sample size grows.

Theorem 6.1. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1–5 and $\mathbb{E}X_{t,T} = 0$. Then, for every $T \geq 2C_1p^2$, $N \geq 1+p \geq 2$ and $\varepsilon > 0$, we have:*

$$\begin{aligned}
&\mathbb{P}\big(\|\hat{a}^{(p)}_{N,T}(t) - \bar{a}^{(p)}_{N,T}(t)\| > \varepsilon\big) \\
&\leq 3p\exp\left(-\frac{\left(\frac{m_f}{4p}\min\left\{1, \varepsilon\frac{1}{8C_0}\right\}\right)^2}{2\big(C_{1,1}\frac{p}{N-p} + \big(\frac{m_f}{4p}\min\left\{1,\varepsilon\frac{1}{8C_0}\right\}\big)^{(3+4d)/(2+2d)}\big(C_{2,1}\frac{p}{N-p}\big)^{1/(2+2d)}\big)}\right) \\
&\leq \begin{cases}
3p\exp\left(-\frac{m_f^2}{32C_{1,1}\frac{p^3}{N-p} + m_f^{(3+4d)/(2+2d)}\left(32C_{2,1}\frac{p^2}{N-p}\right)^{1/(2+2d)}}\right) & \varepsilon \geq 1/(8C_0) \\
3p\exp\left(-\varepsilon^2\frac{m_f^2}{2^{12}C_{1,1}}\left(C_0^2\frac{p^3}{N-p}\right)^{-1}\right) & \varepsilon \leq \min\{U_{p,N}, \frac{1}{8C_0}\} \\
3p\exp\left(-\varepsilon^{1/(2+2d)}\left(\frac{m_f}{2^{9+4d}C_{2,1}}\right)^{1/(2+2d)}\left(C_0\frac{p^2}{N-p}\right)^{-1/(2+2d)}\right) & \\
\qquad\qquad \frac{1}{8C_0} > \varepsilon \geq \min\{U_{p,N}, \frac{1}{8C_0}\}
\end{cases}
\end{aligned}$$

where $\hat{a}_{N,T}^{(p)}(t)$ *is defined in* (1), $\bar{a}_{N,T}^{(p)}(t)$ *is defined in* (35),

$$U_{p,N} := \frac{32C_0}{m_f}\Big(\frac{C_{1,1}^{2+2d}}{C_{2,1}}\Big)^{1/(3+4d)}\Big(\frac{p^{(4+6d)/(3+4d)}}{(N-p)^{(1+2d)/(3+4d)}}\Big),$$

and C_0, C_1 *and* $C_{1,1}$, $C_{2,1}$ *are defined in* (36) *and* (42), *respectively.*

The proof of Theorem 6.1 is deferred to Section A of the appendix.

Theorem 6.1 is a key ingredient to the proof of Lemma A.1 which is essential to the proof of the performance-guarantee-result (Theorem 3.1) of our procedure. Further, it implies

Corollary 6.2. *Let* $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ *satisfy Assumptions 1–5,* $\mathbb{E}X_{t,T} = 0$ *and let* $P = P_T$ *and* $N = N_T$ *be sequences of integers that satisfy* $2 \le 1 + P \le N \le T$. *Assume that* $P = o(N^{(1+2d)/(4+6d)})$ *and* $N \to \infty$, *as* $T \to \infty$. *Further, assume that there exists a sequence* R_T *with* $0 \le R_T \to \infty$ *and* $R_T \log(T) = o\big((N/P)^{1/(3+4d)}\big)$, *as* $T \to \infty$, *where* d *is the constant from Assumption 5. Then, we have*

$$\sup_{p=1,\dots,P}\sup_{t=N,\dots,T}\|\hat{a}_{N,T}^{(p)}(t) - \bar{a}_{N,T}^{(p)}(t)\| = O\left(P^{3/2}\Big(\frac{\log(T)}{N}\Big)^{1/2}\right),$$

almost surely, as $T \to \infty$.

Remark 6.3. *For any stationary AR(p) model we have that* $\bar{a}_{N,T}^{(p)}(u)$ *corresponds to the vector of coefficients. This can be seen from Lemma B.2 and the fact that* $C_1 = 0$ *if the model is stationary. Thus, choosing* $N_T = T$ *and* $P_T = p$, *our result yields the same rate as Theorem 1 in [33], by which the (least squares) estimator is strongly consistent with rate* $(\log(T)/T)^{1/2}$. *An early consistency result for the Yule-Walker estimate with diverging model order is Theorem 6 in [28]. Under the assumption that* $P = O(\log(T)^a)$, $a > 1$ *or* $P = C\log T$, C *sufficiently large, they prove that the rate of convergence is* $O\big((\log\log T/T)^{1/2}\big)$.

7. Conclusion

In this paper, we have presented a method to choose between different forecasting procedures, based on the empirical mean squared prediction errors the procedures achieve. Using the empirical rather than the asymptotic mean squared prediction error, our procedure automatically takes into account that different models should be preferred depending on the amount of data available, which is an important difference to the Focused Information Criterion by [16]. Working in the general framework of locally stationary time series we choose from two classes of forecasts that were motivated by approximating the serial dependence of the time series by time-varying or traditional autoregressive models. The procedure implicitly balances the modelling bias (which is lower if the model is more complex) and the variance of estimation (which increases for more complex

models). Our two step procedure automatically chooses the number of forecasting coefficients to be used and the segment size from which the forecasting coefficients are estimated.

In a comprehensive simulation study we have illustrated that it is often advisable to use a forecasting procedure derived from a simpler model when not a vast amount of data is available. In particular, in the tvAR models of our simulations, if the variation over time is not very pronounced and when the tangent processes are not close to being unit root it is advisable to work with the simpler stationary model, even when the data are non-stationary.

As an important side result of our rigorous theoretical analysis of the method, we have shown that the localised Yule-Walker estimator is strongly, uniformly consistent under local stationarity.

Appendix

In Section A we provide proofs of the results in the main text. In Section A.3, we provide a proof for Theorem 3.1, the performance guarantee of our model selection procedure. The proof relies on properties of the empirical mean squared prediction errors for fixed model order and segment (Lemmas A.1–A.3) which we state in Section A.2. Theorem 6.1 which is about concentration properties of the localised Yule-Walker estimate under local stationarity, is proved in Section A.4. Corollary 6.2 which is about the strong consistency of the localised Yule-Walker estimate is proved in Section A.5. Lemmas 3.2 and 3.3, which fascilitate our discussion of the special case of our procedure are proved in Section A.6.

In Sections B–D we provide technical results about the properties of quantities related to the second order moments. In Section B we state results about the vector $\bar{a}^{(p)}_{N,T}(u)$, defined in (19), around which the localised Yule-Walker estimator concentrates. We also discuss how it is related to the mean square minimising 1-step ahead forecasting coefficients. In Section C we discuss properties of $\bar{v}^{(p,h)}_{N,T}(u)$, the h-step ahead version of $\bar{a}^{(p)}_{N,T}(u)$. Further, we establish properties of $g^{(p,h)}_{\Delta}(u)$ and $\mathrm{MSPE}^{(p,h)}_{\Delta_1,\Delta_2}(u)$ from the definition of $q(\delta)$ that is important for Assumptions 6 and 7. In Section D.2, we provide approximation results for expectations of Toepliz matrices of empirical localised autocovariances $\hat{\gamma}_{k;N,T}(t)$, defined in (3) and in Section D.3 we establish concentration results. In Section E we state a number of technical lemmas that we use in the proofs of our results. We state these results in a separate section, because we believe that they are useful for proving similar results in the future.

Sections F–J that are only available in the extended, arXiv'ed version of the manuscript, cf. [32], contain supplementary material. In Section F we provide the proofs of Lemmas A.1–A.3. In Section H we cite two results from [55] which we use for our proof in Section G.4. In Sections I–J we provide additional material for our simulation and empirical study.

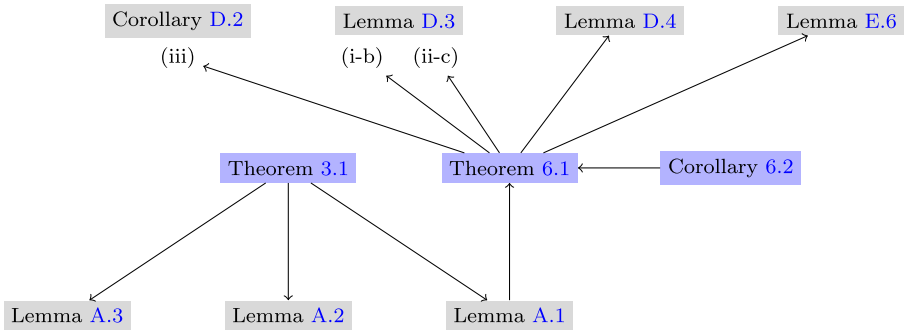

FIG 10. *Map of the results proved in Section A.*

Appendix A: Proofs of Theorems 3.1 and 6.1, of Corollary 6.2, and of Lemmas 3.2 and 3.3

A.1. Outlook

In this section we provide the proofs of the results from Sections 3.3 and 6. In Section A.2 we state and discuss three auxiliary results (Lemmas A.1–A.3) which facilitate the proof of our main result (Theorem 3.1). The auxiliary results are about the empirical mean squared prediction error. Their proofs are deferred to Section F [32]. The proof of Theorem 3.1, by which our model selection procedure chooses models consistently with high probability, is then stated in Section A.3. Because the proof of Lemma A.1 heavily relies on our result about the Yule-Walkers estimators (Theorem 6.1), our proof of Theorem 3.1, implicitly, also depends on it. The proof of Theorem 6.1 and its corollary (Corollary 6.2), by which the localised Yule-Walker estimator is uniformly, strongly consistent, are stated in Sections A.4 and A.5, respectively. For the proof of Theorem 6.1 we employ some of our results about the localised empirical autocovariance estimate from Section D and a technical result from Section E. For the readers convenience, we include Figure 10 in which the dependence of the various results is illustrated graphically.

A.2. Three technical lemmas for the proof

We now introduce two quantities that combine constants from the assumptions. Stating the results in terms of these constants will help to better interpret the bounds and significantly shorten otherwise complicated expressions. To this end, we define

$$C_0 := (2\pi)^{1/2} M_f / m_f, \quad \text{and} \quad C_1 := (2\pi M_f' + C) m_f^{-1}. \tag{36}$$

The constant C_0 can be interpreted in terms of the strength of serial correlation. Note that C_0 will be smaller if there is little variation (uniform in local

time) of the spectral density with respect to frequency. In particular, it will be minimal if the spectral density is constant. This would corresponds to the case of white noise. The constant C_1 can be interpreted as divergence from stationarity. In particular, note the meaning of the two summands of the first factor. The constant M_f' corresponds to the rapidity of changes in stationarity and will vanish in case of stationarity. The constant C corresponds to the quality of locally approximating the correlation structure with a stationary processes correlation structure. It, also, vanishes if the underlying process is stationary.

The aim of the auxiliary results is to approximate general mean squared prediction errors of the form

$$\mathrm{MSPE}_{s,m,N,T}^{(p,h)} := \frac{1}{m}\sum_{t=s+1}^{s+m}\big(X_{t+h,T} - \sum_{i=1}^{p}\hat{v}_{i;N,T}^{(p,h)}(t)X_{t-i+1,T}\big)^2, \tag{37}$$

with $\hat{v}_{i;N,T}^{(p,h)}(t)$ defined in (4) and $\hat{v}_{i;0,T}^{(p,h)}(t) := \hat{v}_{i;t,T}^{(p,h)}(t)$.

The first auxiliary result (Lemma A.1) entails that the quantity defined in (37) is, with high probability, close to

$$\overline{\mathrm{MSPE}}_{s,m,N,T}^{(p,h)} := \frac{1}{m}\sum_{t=s+1}^{s+m}\mathbb{E}\big(X_{t+h,T} - \sum_{i=1}^{p}\bar{v}_{i,N,T}^{(p,h)}(t)X_{t-i+1,T}\big)^2, \tag{38}$$

with

$$\begin{aligned}
&\big(\bar{v}_{N,T}^{(p,h)}(t)\big)'\\
&:= \big(\bar{v}_{1;N,T}^{(p,h)}(t), \bar{v}_{2;N,T}^{(p,h)}(t), \cdots, \bar{v}_{p;N,T}^{(p,h)}(t)\big)\\
&:= e_1'\big(\bar{A}_{N,T}^{(p)}(t)\big)^h\\
&:= e_1'\big(e_1\big(\bar{a}_{N,T}^{(p)}(t)\big)' + H\big)^h,\\
&\qquad e_1 := \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix},\ H := \begin{pmatrix}0 & 0 & \cdots & 0 & 0\\1 & 0 & \cdots & 0 & 0\\0 & 1 & \cdots & 0 & 0\\\vdots & \ddots & \cdots & 0 & 0\\0 & 0 & \cdots & 1 & 0\end{pmatrix},
\end{aligned} \tag{39}$$

where $\bar{a}_{N,T}^{(p)}(t)$ is defined in (35), e_1 denotes the first canonical unity vector of dimension p and H denotes a $p \times p$ Jordan block with all eigenvalues equal to zero. The second auxiliary result (Lemma A.2) provides a simplified probability bound for the result in Lemma A.1 that can be applied in an especially relevant case.

By our third auxiliary result (Lemma A.3) we have that $\overline{\mathrm{MSPE}}_{s,m,N,T}^{(p,h)}$ in turn can be approximated by $\mathrm{MSPE}_{N/T,m/T}^{(p,h)}(s/T)$, where $\mathrm{MSPE}_{\Delta_1,\Delta_2}^{(p,h)}(u)$ is the quantity defined in (21), with continuous time indices Δ_1 and Δ_2. Note that this quantity also appears in $q(\delta)$ defined in (24) which is a relevant component of Assumptions 6 and 7.

Some comparison of $\overline{\mathrm{MSPE}}^{(p,h)}_{s,m,N,T}$, defined in (38), and $\mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s/T)$, as defined in (21) are in order: Note that $\overline{\mathrm{MSPE}}^{(p,h)}_{s,m,N,T}$ is defined as the expectation of a modified version of $\mathrm{MSPE}^{(p,h)}_{s,m,N,T}$, the modification being that $\hat{v}^{(p,h)}_{N,T}(t)$ is exchanged by $\bar{v}^{(p,h)}_{N,T}(t)$. As before, we will denote $\bar{v}^{(p,h)}_{0,T}(t) := \bar{v}^{(p,h)}_{t,T}(t)$.

We have that $g^{(p,h)}_{N/T}(t/T)$ approximates $\mathbb{E}[(X_{t+h,T} - f^{\mathrm{loc.}}_{t,h;p,N})^2]$, with $f^{\mathrm{loc.}}_{t,h;p,N}$ defined in (5). Therefore, the expectation of the empirical mean squared prediction error (37) we are considering is naturally an average of these quantities:

$$\mathbb{E}[\mathrm{MSPE}^{(p,h)}_{s,m,N,T}] = \frac{1}{m}\sum_{t=s+1}^{s+m} \mathbb{E}[(X_{t+h,T} - f^{\mathrm{loc.}}_{t,h;p,N})^2].$$

We now state the results that the quantities defined in (37) and (38) are close, with high probability.

Lemma A.1. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1–5 and $\mathbb{E}X_{t,T} = 0$. Then, for every $m, h \in \mathbb{N}^*$, $p \in \mathbb{N}$, $N \geq 6C_0p^2$, $\varepsilon > 0$ and $T \geq 10C_1p^2$, with $\mathrm{MSPE}^{(p,h)}_{s,m,N,T}$ defined in (37) and $\overline{\mathrm{MSPE}}^{(p,h)}_{s,m,N,T}$ defined in (38), we have that*

$$\mathbb{P}\Big(\Big|\mathrm{MSPE}^{(p,h)}_{s,m,N,T} - \overline{\mathrm{MSPE}}^{(p,h)}_{s,m,N,T}\Big| > \varepsilon\Big) \leq P^{(p,h)}_{m,N}(\varepsilon)$$

and

$$\mathbb{P}\Big(\Big|\mathrm{MSPE}^{(p,h)}_{s,m,0,T} - \overline{\mathrm{MSPE}}^{(p,h)}_{s,m,0,T}\Big| > \varepsilon\Big) \leq P^{(p,h)}_{m,s}(\varepsilon)$$

with

$$P^{(p,h)}_{m,N}(\varepsilon) := (1 + 4p + 2p^2)$$
$$\cdot\exp\Bigg(-\frac{\frac{\varepsilon^2}{(p+1)^4}}{8((2C_0+1)^{4h}\frac{C_{1,2}(h+p-1)}{m} + (\frac{\varepsilon}{2(p+1)^2})^{(3+8d)/(2+4d)}((2C_0+1)^{2h}\frac{C_{2,2}(h+p-1)}{m})^{1/(2+4d)})}\Bigg)$$
$$+ 6mp^2(p+1)\exp\left(-\frac{\eta^2}{2\big(C_{1,1}\frac{p}{N-p} + \eta^{(3+4d)/(2+2d)}(C_{2,1}\frac{p}{N-p})^{1/(2+2d)}\big)}\right),$$

where

$$\eta := \frac{m_f}{4p}\min\Big\{1, \bar{\mu}/(8C_0)\Big\}, \qquad \bar{\mu} := 2^{1-h}\frac{\mu}{\mu + h(2C_0)^{h-1}},$$
$$\mu := \frac{\bar{\varepsilon}}{2\Big(\big(2C_0+1\big)^{2h} + \bar{\varepsilon}\Big)^{1/2}}, \qquad \bar{\varepsilon} := \frac{\varepsilon/(p+1)^2}{2\big((6\pi M_f c^2 24^d)^2 + \varepsilon^2/(p+1)^4\big)^{1/4}},$$

and the constants C_0, C_1, and $C_{1,1}$, $C_{1,2}$, $C_{2,1}$, $C_{2,2}$, and m_f, M_f, and c, d are defined in (36), (42), and Assumptions 3 and 5, respectively.

In a typical application the bound $P^{(p,h)}_{m,N}(\varepsilon)$ will be small. More precisely, the following, more accessible bound for $P^{(p,h)}_{m,N}(\varepsilon)$, proved in Section F [32], will be useful

Lemma A.2. *There exist constants $D_1, D_2, D_3 > 0$ and $K_0 > 1$, defined in the proof, such that for any*

$$\max\left\{\left(\frac{h+p}{m}\right)^{\frac{1+4d}{3+8d}} K_0^h p^2, \left(\frac{p}{N-p}\right)^{\frac{1+2d}{3+4d}} K_0^h p^3 h\right\} < \varepsilon \le \min\{6\pi M_f c^2 24^d, 1\}(p+1)^2, \tag{40}$$

we have

$$P_{m,N}^{(p,h)}(\varepsilon) \le D_1\left[p^2 \exp\left(-D_2\left(\frac{m}{h+p}\right)^{1/(3+8d)}\right) + mp^3 \exp\left(-D_3\left(\frac{N-p}{p}\right)^{1/(3+4d)}\right)\right].$$

Note that we are interested in the scenario where $\varepsilon > 0$ may be small. Therefore, if we allow that p and h may be large, we have to require m and N to be of a minimum size.

We now state the result that the quantities defined in (38) and (21) are close. The quality of the approximation depends on the parameters T, p and h, but is uniform with respect to s, m and N:

Lemma A.3. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1–5 and $\mathbb{E}X_{t,T} = 0$. Then, for every $m, h \in \mathbb{N}^*$, $p \in \mathbb{N}$, $T \ge 6h2^h C_1 p^2$, and $N \ge 4h2^h C_0 p^2$, with $\overline{\mathrm{MSPE}}_{s,m,N,T}^{(p,h)}$ defined in (38) and $\mathrm{MSPE}_{\Delta_1,\Delta_2}^{(p,h)}(u)$ defined in (21), we have*

$$\left|\overline{\mathrm{MSPE}}_{s,m,N,T}^{(p,h)} - \mathrm{MSPE}_{N/T,m/T}^{(p,h)}(s/T)\right| \le 8h2^h\big(C_0\big)^{2h+1}\left[6(2\pi M_f' + C)\frac{p^2}{T} + \frac{p^2}{N}\right]$$

and

$$\left|\overline{\mathrm{MSPE}}_{s,m,0,T}^{(p,h)} - \mathrm{MSPE}_{s/T,m/T}^{(p,h)}(s/T)\right| \le 8h2^h\big(C_0\big)^{2h+1}\left[6(2\pi M_f' + C)\frac{p^2}{T} + \frac{p^2}{N}\right].$$

The proofs of the three lemmas are long and technical. We therefore defer them to Section F [32].

A few comments about Lemma A.3 are in order. Note that the approximation error is zero in case of a stationary time series, as then $2\pi M_f' + C = 0$. Note further, that the approximation will be better, if h and p are small compared to T. More precisely, if $h(2C_0^2)^h p^2 = o(T)$, then the difference will vanish asymptotically. In particular, if $h = O(1)$, then it would suffice to assume that $p = o(T^{1/2})$, for the approximation error to vanish asymptotically.

A.3. Proof of Theorem 3.1

The constants D_1, D_2 and D_3 are defined as

$$\begin{aligned} D_1 &:= 12, \\ D_2 &:= \big(2^8 \max\{C_{1,2}, C_{2,2}^{1/(2+4d)}\}\big)^{-1}, \text{ and} \\ D_3 &:= K_1^2/\big(2^{12}\max\{C_{1,1}, (K_1^{3+4d}C_{2,1})^{1/(2+2d)}\}\big), \end{aligned} \tag{41}$$

where $K_1 := m_f/(32\min\{(6\pi M_f c^2 24^d)^{1/2}, 1\})$ and

$$\begin{aligned} C_{1,\alpha} &:= 12\cdot 2^{10\alpha d+7}\alpha^{4\alpha d}\big(\max\{c^2, 3\pi M_f, 1\}\big)^{2\alpha}\mathrm{e}\Big(1+\frac{1}{\log\rho}\Big)\Big(1+K^{1/2}\Big), \\ C_{2,\alpha} &:= 12\cdot 2^{4\alpha d+3}\alpha^{2\alpha d}\big(\max\{c^2, 3\pi M_f, 1\}\big)^{\alpha}\mathrm{e}\Big(1+\frac{1}{\log\rho}\Big), \end{aligned} \tag{42}$$

with $\alpha \in \{1, 2\}$. In the definitions, we have K and ρ the constants from Assumption 2, M_f and m_f the constants from Assumption 3, and c and d the constants from Assumption 5.

To compact notation, we denote $s_2 := T - h$, $\mathrm{MSPE}^{(p_1,h)}_{s_i,m,N,T}$ by X_i and $\mathrm{MSPE}^{(p_2,h)}_{s_i,m,0,T}$ by Y_i. Further, denote $\mathrm{MSPE}^{(p_1,h)}_{N/T,m/T}(\frac{s_i}{T})$ and $\mathrm{MSPE}^{(p_2,h)}_{s_1/T,m/T}(\frac{s_i}{T})$ by $\bar{Y}_i$ and $\bar{X}_i$, respectively. Further, we abbreviate $A := Y_1 - X_1(1+\delta)$ and $B := Y_1 - Y_2 + (X_2 - X_1)(1+\delta)$.

First note that Assumptions 6 and 7 imply that

$$T \geq \max\big\{10C_1(\max\mathcal{P})^2, 6h2^hC_1(\max\mathcal{P})^2\big\}, \quad \min\mathcal{N} \geq 4h2^hC_0(\max\mathcal{P})^2$$

Therefore, the conditions of Lemmas A.1 and A.3 are satisfied. Further, note that since

$$\min\mathcal{N} \geq 8h2^h\big(C_0\big)^{2h+1}\big(\max\mathcal{P}\big)^2\Big[6(2\pi M_f' + C) + 1\Big]\Big(20(1+\delta)/q(\delta)\Big)$$

and because $N \leq T$ for all $N \in \mathcal{N}$, we have that the bound from Lemma A.3 can again be bounded

$$8h2^h\big(C_0\big)^{2h+1}\big(\max\mathcal{P}\big)^2\Big[6(2\pi M_f' + C)\frac{1}{T} + \frac{1}{N}\Big] \leq \frac{q(\delta)}{20(1+\delta)} =: \varepsilon \tag{43}$$

Finally, note that by Assumption 7, we have

$$T \geq 4m\big(2h+1\big)\big(C_0\big)^{2h+1}M_f'\frac{20(1+\delta)}{q(\delta)}$$

which implies that (a quantity related to the bound from Lemma C.4(iv)) can be bounded

$$4\big(2h+1\big)\big(C_0\big)^{2h+1}M_f'\Big|\frac{s_1 - s_2}{T}\Big| \leq \varepsilon. \tag{44}$$

Now, for the proof of the Theorem, note that

$$\begin{aligned} &\mathbb{P}\Big((\hat{R}_{T,2}(h) \geq 1+\delta \text{ and } \hat{R}_{T,3}(h) \geq 1+\delta) \text{ or } (\hat{R}_{T,2}(h) < 1+\delta \\ &\quad \text{and } \hat{R}_{T,3}(h) < 1+\delta)\Big) \\ &\quad \geq 1 - \sum_{p_1,p_2\in\mathcal{P}}\sum_{N\in\mathcal{N}}\Big(\mathbb{P}\big(|A| \leq q(\delta)/2\big) + \mathbb{P}\big(|B| > q(\delta)/2\big)\Big), \end{aligned} \tag{45}$$

which we prove in Section G.1 [32].

We now bound the part of the right hand side of (45) that involves the quantity A. Using the fact that

$$\begin{aligned}|\bar{Y}_1 - \bar{X}_1(1+\delta)| &= |Y_1 + \bar{Y}_1 - Y_1 - X_1(1+\delta) + (X_1 - \bar{X}_1)(1+\delta)| \\ &\le |Y_1 - X_1(1+\delta)| + |\bar{Y}_1 - Y_1| + |X_1 - \bar{X}_1|(1+\delta),\end{aligned}$$

we have the first inequality of

$$\begin{aligned}&\mathbb{P}\Big(|A| \le q(\delta)/2\Big) = \mathbb{P}\Big(|Y_1 - X_1(1+\delta)| \le q(\delta)/2\Big) \\ &\le \mathbb{P}\Big(|Y_1 - \bar{Y}_1| + |X_1 - \bar{X}_1|(1+\delta) \ge |\bar{Y}_1 - \bar{X}_1(1+\delta)| - q(\delta)/2\Big) \\ &\le \mathbb{P}\Big(|Y_1 - \bar{Y}_1| \ge \frac{1}{2}(|\bar{Y}_1 - \bar{X}_1(1+\delta)| - q(\delta)/2)\Big) \\ &\qquad + \mathbb{P}\Big(|X_1 - \bar{X}_1| \ge \frac{1}{2(1+\delta)}(|\bar{Y}_1 - \bar{X}_1(1+\delta)| - q(\delta)/2)\Big) \\ &\le \mathbb{P}\Big(|Y_1 - \bar{Y}_1| > q(\delta)/10)\Big) + \mathbb{P}\Big(|X_1 - \bar{X}_1| > \frac{q(\delta)}{10(1+\delta)})\Big) \qquad (46) \\ &\le P^{(p_2,h)}_{m,T-m}\Big(\frac{q(\delta)}{20}\Big) + P^{(p_1,h)}_{m,N}\Big(\frac{q(\delta)}{20(1+\delta)}\Big) \le 2P^{(p_{\max},h)}_{m,N_{\min}}\Big(\frac{q(\delta)}{20(1+\delta)}\Big), \qquad (47)\end{aligned}$$

where $p_{\max} := \max \mathcal{P}$ and $N_{\min} := \min \mathcal{N}$. For the inequality in (46) we have used the definition of $q(\delta)$ and $1/4 > 1/10$. For the first inequality in (47) we have used Lemmas A.1 and A.3 and (43) to obtain

$$\mathbb{P}\Big(\Big|\mathrm{MSPE}^{(p,h)}_{s,m,N,T} - \mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s/T)\Big| > 2\varepsilon\Big) \le P^{(p,h)}_{m,N}(\varepsilon) \qquad (48)$$

and

$$\mathbb{P}\Big(\Big|\mathrm{MSPE}^{(p,h)}_{s,m,0,T} - \mathrm{MSPE}^{(p,h)}_{s/T,m/T}(s/T)\Big| > 2\varepsilon\Big) \le P^{(p,h)}_{m,s}(\varepsilon). \qquad (49)$$

For the second inequality in (47) we have used that

$$p_1 \le p_2 \Rightarrow P^{(p_1,h)}_{m,N}(\varepsilon) \le P^{(p_2,h)}_{m,N}(\varepsilon), \quad N_1 \le N_2 \Rightarrow P^{(p,h)}_{m,N_1}(\varepsilon) \ge P^{(p,h)}_{m,N_2}(\varepsilon),$$

and $\varepsilon_1 \le \varepsilon_2 \Rightarrow P^{(p,h)}_{m,N}(\varepsilon_1) \ge P^{(p,h)}_{m,N}(\varepsilon_2)$.

We now bound the part of the right hand side of (45) that involves the quantity B. We have

$$\begin{aligned}&\mathbb{P}\Big(|B| > q(\delta)/2\Big) = \mathbb{P}\Big(|Y_1 - Y_2 + (X_2 - X_1)(1+\delta)| > q(\delta)/2\Big) \\ &\le \mathbb{P}\Big(|Y_1 - Y_2| > q(\delta)/4\Big) + \mathbb{P}\Big(|X_2 - X_1| > \frac{q(\delta)}{4(1+\delta)}\Big) \\ &\le 2P^{(p_2,h)}_{m,T-m}\Big(\frac{q(\delta)}{20}\Big) + 2P^{(p_1,h)}_{m,N}\Big(\frac{q(\delta)}{20(1+\delta)}\Big) \le 4P^{(p_{\max},h)}_{m,N_{\min}}\Big(\frac{q(\delta)}{20(1+\delta)}\Big). \qquad (50)\end{aligned}$$

Note that we have

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\mathrm{MSPE}^{(p,h)}_{s_1,m,N,T} - \mathrm{MSPE}^{(p,h)}_{s_2,m,N,T} > 5\varepsilon\Big)\\
&\le \mathbb{P}\Big(\Big|\mathrm{MSPE}^{(p,h)}_{s_1,m,N,T} - \mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s_1/T)\Big| > 2\varepsilon\Big)\\
&\quad + \mathbb{P}\Big(\Big|\mathrm{MSPE}^{(p,h)}_{s_2,m,N,T} - \mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s_2/T)\Big| > 2\varepsilon\Big)\\
&\quad + I\Big\{\Big|\mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s_1/T) - \mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s_2/T)\Big| > \varepsilon\Big)\Big\},
\end{aligned}$$

where the first two terms can be bound by an application of (48) and the indicator function vanishes for all T satisfying the condition of the Theorem, because

$$\begin{aligned}
&\Big|\mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s_1/T) - \mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s_2/T)\Big|\\
&\quad \le 4\big(2h+1\big)\big(C_0\big)^{2h+1} M'_f \Big|\frac{s_1 - s_2}{T}\Big|,
\end{aligned}$$

where Lemma C.4(iv) was employed to obtain (44) for the last inequality.

Thus, combining (45), (47) and (50), we have shown that

$$\begin{aligned}
&\mathbb{P}\Big((\hat{R}_{T,2}(h) \ge 1+\delta \text{ and } \hat{R}_{T,3}(h) \ge 1+\delta) \text{ or } (\hat{R}_{T,2}(h) < 1+\delta\\
&\quad \text{and } \hat{R}_{T,3}(h) < 1+\delta)\Big)\\
&\qquad \ge 1 - 6|\mathcal{P}|^2|\mathcal{N}|P^{(p_{\max},h)}_{m,N_{\min}}\Big(\frac{q(\delta)}{20(1+\delta)}\Big).
\end{aligned}$$

An application of Lemma A.2 finishes the proof of the theorem.

Remark A.4. *Equations* (48)–(49)*, which are immediate consequences of Lemmas A.1 and A.3, can be used to derive the almost sure convergence of*

$$\Big|\mathrm{MSPE}^{(p,h)}_{s,m,N,T} - \mathrm{MSPE}^{(p,h)}_{N/T,m/T}(s/T)\Big| \text{ and } \Big|\mathrm{MSPE}^{(p,h)}_{s,m,0,T} - \mathrm{MSPE}^{(p,h)}_{s/T,m/T}(s/T)\Big|,$$

under appropriate conditions, using a classical Borel-Cantelli argument.

This asymptotic view of $\mathrm{MSPE}^{(p,h)}_{s,m,N,T}$ *and* $\mathrm{MSPE}^{(p,h)}_{s,m,0,T}$*, in particular, implies that we may interpret* $\mathrm{MSPE}^{(p,h)}_{\Delta_1,\Delta_2}(u)$ *as an approximation of the expectation of the empirical MSPE for an* h*-step ahead linear forecast of order* p*, where observations up to (local) time* u *have been made. The* Δ_1 *and* Δ_2 *are (localised) length which are related to the segment length of observations used for the estimation of the forecasting coefficients and the segment from which the observations* $X_{t+h,T}$ *that are being forecasted are taken, respectively.*

We now proceed with the proofs of the results from Section 6.

A.4. Proof of Theorem 6.1

Let $M := \hat{\Gamma}^{(p)}_{N,T}(t)$, $M_0 := \mathbb{E}M$, $v := \hat{\gamma}^{(p)}_{N,T}(t)$, and $v_0 := \mathbb{E}v$. By Lemma D.3(ii-c) we deduce that M_0 is invertible for $T \geq 2p^2C_1$, because it is positive definite with smallest eigenvalue larger or equal to $m_f/2$. An application of Lemma E.6, with the spectral norm as the matrix norm and the Euclidean norm as the vector norm yields

$$\begin{aligned}
&\mathbb{P}\big(\|\hat{a}^{(p)}_{N,T}(t) - \bar{a}^{(p)}_{N,T}(t)\| > \varepsilon\big) = \mathbb{P}\big(\|M^{-1}v - M_0^{-1}v_0\| > \varepsilon\big)\\
&\leq \mathbb{P}\Big(\|M - M_0\| > \frac{1}{2\|M_0^{-1}\|}\Big) + \mathbb{P}\Big(\|v - v_0\| > \frac{\varepsilon}{4}\frac{1}{\|M_0^{-1}\|}\Big)\\
&\qquad + \mathbb{P}\Big(\|M - M_0\| > \frac{\varepsilon}{4}\frac{1}{(\|M_0^{-1}\|)^2\,\|v_0\|}\Big) I\{\|v_0\| \neq 0\}\\
&\leq \mathbb{P}\Big(\max_{k=0,\ldots,p-1} |\hat{\gamma}_{k;N,T}(t) - \mathbb{E}\hat{\gamma}_{k;N,T}(t)| > \frac{1}{4p}m_f\Big)\\
&\qquad + \mathbb{P}\Big(\max_{k=1,\ldots,p} |\hat{\gamma}_{k;N,T}(t) - \mathbb{E}\hat{\gamma}_{k;N,T}(t)| > \frac{\varepsilon}{8p^{1/2}}m_f\Big)\\
&\qquad + \mathbb{P}\Big(\max_{k=0,\ldots,p-1} |\hat{\gamma}_{k;N,T}(t) - \mathbb{E}\hat{\gamma}_{k;N,T}(t)| > \frac{\varepsilon}{32(2\pi)^{1/2}M_f p}m_f^2\Big)\\
&\leq 3p \max_{k=0,\ldots,p} \mathbb{P}\Big(|\hat{\gamma}_{k;N,T}(t) - \mathbb{E}\hat{\gamma}_{k;N,T}(t)| > \frac{m_f}{4p}\min\Big\{1, \frac{\varepsilon p^{1/2}}{2}, \frac{\varepsilon}{8C_0}\Big\}\Big),\\
&= 3p \max_{k=0,\ldots,p} \mathbb{P}\Big(|\hat{\gamma}_{k;N,T}(t) - \mathbb{E}\hat{\gamma}_{k;N,T}(t)| > \frac{m_f}{4p}\min\Big\{1, \frac{\varepsilon}{8C_0}\Big\}\Big),
\end{aligned}$$

where we have use Lemma D.3(ii-c) again to bound $1/\|M_0^{-1}\|$. In the last step we employed that $\frac{p^{1/2}}{2} \geq \frac{1}{8C_0}$. Further, we have used that $M - M_0$ satisfies

$$\begin{aligned}
\|M - M_0\|_1 = \|M - M_0\|_\infty &= \max_{1\leq\ell\leq p} \sum_{h=1}^{p} |\hat{\gamma}_{h-\ell;N,T}(t) - \mathbb{E}\hat{\gamma}_{h-\ell;N,T}(t)|\\
&\leq p \max_{k=0,\ldots,p-1} |\hat{\gamma}_{k;N,T}(t) - \mathbb{E}\hat{\gamma}_{k;N,T}(t)|.
\end{aligned}$$

Thus, by Hölder's inequality

$$\|M - M_0\| \leq \Big(\|M - M_0\|_1 \|M - M_0\|_\infty\Big)^{1/2} \leq p \max_{k=0,\ldots,p-1} |\hat{\gamma}_{k;N,T}(t) - \mathbb{E}\hat{\gamma}_{k;N,T}(t)|.$$

For the Euclidean norm we have used

$$\|v - v_0\| \leq p^{1/2}\|v - v_0\|_\infty = p^{1/2} \max_{k=1,\ldots,p} |\hat{\gamma}_{k;N,T}(t) - \mathbb{E}\hat{\gamma}_{k;N,T}(t)|.$$

Finally, by Corollary D.2(iii) and Lemma D.3(i-b), we have

$$\begin{aligned}
\|v_0\| = \|\mathbb{E}\hat{\gamma}^{(p)}_{N,T}(t/T)\| &\leq \|f_{p,N} \circ \gamma^{(p)}_{N/T}(t/T)\| + \|\mathbb{E}\hat{\gamma}^{(p)}_{N,T}(t/T) - f_{p,N} \circ \gamma^{(p)}_{N/T}(t/T)\|\\
&\leq (2\pi)^{1/2}M_f + 2T^{-1}p^{3/2}C_1 m_f \leq 2(2\pi)^{1/2}M_f,
\end{aligned}$$

where the second inequality holds for $T \geq 2\frac{p^{3/2}C_1 m_f}{(2\pi)^{1/2}M_f} = 2C_1 p^{3/2}/C_0$, which is the case, as $T \geq 2C_1 p^2$ is assumed. Here we also have used that $\|f_{p,N} \circ x\| \leq \|x\|$, as all entries of $f_{p,N}$ are between 0 and 1. Applying Lemma D.4 yields the assertion, because

$$\begin{aligned}
&\mathbb{P}\big(\|\hat{a}^{(p)}_{N,T}(t) - a^{(p)}_{N,T}(t)\| > \varepsilon\big) \\
&\leq 3p \max_{h=0,\dots,p} \exp\left(- \frac{\eta^2}{2\Big(C_{1,1}\frac{h_*}{N-|h|} + \eta^{(3+4d)/(2+2d)}\Big(C_{2,1}\frac{h_*}{N-|h|}\Big)^{1/(2+2d)}\Big)}\right) \\
&= 3p \exp\left(- \frac{\eta^2}{2\Big(C_{1,1}\frac{p}{N-p} + \eta^{(3+4d)/(2+2d)}\Big(C_{2,1}\frac{p}{N-p}\Big)^{1/(2+2d)}\Big)}\right)
\end{aligned}$$

where $\eta := \frac{m_f}{4p}\min\Big\{1, \varepsilon\frac{1}{8C_0}\Big\}$, and the third line follows from the fact, for any two integers N and p with $N \geq 1 + p \geq 2$ we have that $(\frac{h_*}{N-|h|})_{h=0,1,\dots,p}$ is an increasing sequence. This is easy to see: $\frac{1}{N-0} \leq \frac{1}{N-1} \leq \dots \leq \frac{p-1}{N-p+1} \leq \frac{p}{N-p}$. Note that $T \geq 2pC_1 \geq C/(\pi m_f)$, such that this condition of Lemma D.4 is met. □

A.5. *Proof of Corollary 6.2*

Note the fact that, if $R_n \geq 0$ is a sequence with $R_n \to \infty$, as $n \to \infty$, then

$$b_n = O(1) \Leftrightarrow b_n = o(r_n), \forall\, 0 \leq r_n \leq R_n, \text{ with } r_n \to \infty, \text{ as } n \to \infty.$$

Thus, employing the Borel-Cantelli lemma, it suffices to show that, for any given $\varepsilon > 0$ and sequence $0 \leq r_T \leq R_T^{1/2}$ with $r_T \to \infty$, we have

$$\sum_{T=1}^{\infty} \mathbb{P}\Big(\sup_{p=1,\dots,P}\sup_{t=N,\dots,T} \|\hat{a}^{(p)}_{N,T}(t) - \bar{a}^{(p)}_{N,T}(t)\| > \varepsilon P^{3/2}\Big(\frac{\log(T)}{N}\Big)^{1/2} r_T\Big) < \infty.$$

This follows, since we have

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{p=1,\dots,P}\sup_{t=N,\dots,T} \|\hat{a}^{(p)}_{N,T}(t) - a^{(p)}_{N,T}(t)\| > \varepsilon P^{3/2}\Big(\frac{\log(T)}{N}\Big)^{1/2} r_T\Big) \\
&\leq P \cdot T \cdot \sup_{p=1,\dots,P}\sup_{t=N,\dots,T} \mathbb{P}\Big(\|\hat{a}^{(p)}_{N,T}(t) - a^{(p)}_{N,T}(t)\| > \varepsilon P^{3/2}\Big(\frac{\log(T)}{N}\Big)^{1/2} r_T\Big) \\
&\leq P \cdot T \cdot \sup_{p=1,\dots,P}\sup_{t=N,\dots,T} \mathbb{P}\Big(\|\hat{a}^{(p)}_{N,T}(t) - a^{(p)}_{N,T}(t)\| > \varepsilon p^{3/2}\tilde{C}^{1/2}\Big(\frac{\log(T)}{N-p}\Big)^{1/2} r_T\Big) \\
&\leq P \cdot T \cdot \sup_{p=1,\dots,P}\sup_{t=N,\dots,T} 3p \exp\left(- \varepsilon^2 \frac{p^3 \log(T)}{N-p}\tilde{C} r_T^2 \frac{m_f^2}{2^{12}C_{1,1}}\Big(C_0^2\frac{p^3}{N-p}\Big)^{-1}\right)\Big) \\
&= 3T^3 \exp\left(- \varepsilon^2 \log(T)\tilde{C} r_T^2 \frac{m_f^2}{2^{12}C_{1,1}} C_0^{-2}\right)\Big) \leq 3T^{-2},
\end{aligned}$$

for T large enough. In the second inequality we have used the fact that, due to $P = o(N)$, there exists a $\tilde{C} > 0$ such that $1/N \geq \tilde{C}/(N-P)$, for T large enough. Note that we have $P = o(T^{1/2})$, from $N \leq T$, $P = o(N^{(1+2d)/(4+6d)})$ and $d \geq 1/2$, such that, in the third inequality, Theorem 6.1 can be applied, where we have also used the fact that, under the assumptions made

$$p^{3/2}\Big(\frac{\log(T)}{N-p}\Big)^{1/2} R_T^{1/2} = o\Big(\frac{P^{(4+6d)(3+4d)}}{N^{(1+2d)/(3+4d)}}\Big),$$

implying that, for T large enough, we have

$$\varepsilon p^{3/2}\Big(\frac{\log(T)}{N-p}\Big)^{1/2} r_T \leq \min\{U_{p,N}, 1/(8C_0)\} = U_{p,N}.$$

This completes the proof. □

A.6. Proofs of Lemmas 3.2 and 3.3

For the proof of Lemma 3.2 it suffices to show that

$$q(\delta) := \min_{N\in\mathcal{N}} \Big|\mathrm{MSPE}^{(1,1)}_{s_1/T,m/T}\big(\frac{s_1}{T}\big) - (1+\delta)\cdot \mathrm{MSPE}^{(1,1)}_{N/T,m/T}\big(\frac{s_1}{T}\big)\Big| \geq \delta\pi m_f\big(1-\rho^2\big).$$

Likewise, to show Lemma 3.3, we bound $q(\delta)$ with $\pi m_f D^2_{\inf}/2$ on the right hand side.

Denoting

$$\gamma_k(u,\Delta) := \int_0^1 \gamma_k(u+\Delta(x-1))\mathrm{d}x = \Delta^{-1}\int_{u-\Delta}^{u} \gamma_k(v)\mathrm{d}v$$

we have, by definition (21), that

$$\begin{aligned}\mathrm{MSPE}^{(1,1)}_{\Delta_1,\Delta_2}(u) &= \int_0^1 g^{(1,1)}_{\Delta_1}\big(u+\Delta_2(1-x)\big)\mathrm{d}x \\ &= \Delta_2^{-1}\int_u^{u+\Delta_2}\Big(\gamma_0(w) - 2\frac{\gamma_1\big(w;\Delta_1\big)}{\gamma_0\big(w;\Delta_1\big)}\gamma_1(w) + \Big(\frac{\gamma_1\big(w;\Delta_1\big)}{\gamma_0\big(w;\Delta_1\big)}\Big)^2\gamma_0(w)\Big)\mathrm{d}w.\end{aligned}$$

To find the lower bound we want, it therefore suffices to proof lower bounds, for every $w \in [s_1/T, (s_1+m)/T]$, of the following difference

$$\begin{aligned}&\Big(\Big(\gamma_0(w) - 2\frac{\gamma_1\big(w;s_1/T\big)}{\gamma_0\big(w;s_1/T\big)}\gamma_1(w) + \Big(\frac{\gamma_1\big(w;s_1/T\big)}{\gamma_0\big(w;s_1/T\big)}\Big)^2\gamma_0(w)\Big) \\ &- (1+\delta)\Big(\gamma_0(w) - 2\frac{\gamma_1\big(w;N/T\big)}{\gamma_0\big(w;N/T\big)}\gamma_1(w) + \Big(\frac{\gamma_1\big(w;N/T\big)}{\gamma_0\big(w;N/T\big)}\Big)^2\gamma_0(w)\Big)\Big).\end{aligned} \tag{51}$$

For Lemma 3.3 we will bound $-1\times$ (51). For notational convenience we omit the w's and denote

$$E := \frac{\gamma_1(w,N/T)}{\gamma_0(w,N/T)}, \text{ and } F := \frac{\gamma_1(w,s_1/T)}{\gamma_0(w,s_1/T)}.$$

By elementary considerations it can be shown that

$$(51) = \gamma_0 \left(\left(F - \frac{\gamma_1}{\gamma_0} \right)^2 - \left(\frac{\gamma_1}{\gamma_0} - E \right)^2 - \delta \left(1 - \left(\frac{\gamma_1}{\gamma_0} \right)^2 + \left(\frac{\gamma_1}{\gamma_0} - E \right)^2 \right) \right). \quad (52)$$

By (28), we have $|F - \frac{\gamma_1}{\gamma_0}| \geq D_{\text{inf}}$ and by (27), we have $|F - \frac{\gamma_1}{\gamma_0}| \leq D_{\text{sup}}$. Further, we have that $|\frac{\gamma_1}{\gamma_0} - E| \leq M_f' N/(m_f T)$, uniformly with respect to ω, which can be seen as follows: first, note that

$$\begin{aligned} \left| \gamma_k(w, N/T) - \gamma_k(w, 0) \right| &\leq \int_0^1 \left| \gamma_k(w) - \gamma_k(w - \frac{N}{T}(1-x)) \right| \mathrm{d}x \\ &\leq 2\pi M_f' \int_0^1 \frac{N}{T}(1-x)\mathrm{d}x = \pi M_f' \frac{N}{T} \end{aligned}$$

Further, note that we have $\frac{x}{y} - \frac{x_0}{y_0} = \frac{1}{y_0}\left(\frac{x}{y}(y_0 - y) + (x - x_0) \right)$ and thus

$$\left| \frac{\gamma_1}{\gamma_0} - E \right| \leq \frac{1}{\gamma_0(w; N/T)} \left(\frac{|\gamma_1|}{\gamma_0} + 1 \right) \pi M_f' \frac{N}{T} \leq \frac{M_f'}{m_f} \frac{N}{T}, \quad (53)$$

where we have used that $2\pi m_f \leq \gamma_0(w; \Delta) := \int_0^1 \gamma_0(w + \Delta(x-1))\mathrm{d}x$ and $|\gamma_1|/\gamma_0 \leq 1$. Employing (52), we have now brought the tools together to prove Lemma 3.2:

$$\begin{aligned} -1 \times (51) &= \gamma_0 \left(\delta \left(1 - \left(\frac{\gamma_1}{\gamma_0} \right)^2 + \left(\frac{\gamma_1}{\gamma_0} - E \right)^2 \right) - \left(F - \frac{\gamma_1}{\gamma_0} \right)^2 + \left(\frac{\gamma_1}{\gamma_0} - E \right)^2 \right) \\ &\geq 2\pi m_f \left(1 - \rho^2 \right) \left(\delta/2 + \delta/2 - D_{\text{sup}}^2/(1-\rho^2) \right) \big) \big) \geq \pi m_f \delta (1 - \rho^2). \end{aligned}$$

For the first inequality we have used the fact that $(\gamma_1/\gamma_0 - E)^2 \geq 0$ and the definitions of ρ and D_{sup}. For the second inequality we have used the condition imposed on δ.

Finally, employing (52) again, we prove Lemma 3.3:

$$\begin{aligned} (51) &= \gamma_0 \left(\left(F - \frac{\gamma_1}{\gamma_0} \right)^2 - \left(\frac{\gamma_1}{\gamma_0} - E \right)^2 - \delta \left(1 - \left(\frac{\gamma_1}{\gamma_0} \right)^2 + \left(\frac{\gamma_1}{\gamma_0} - E \right)^2 \right) \right) \\ &\geq 2\pi m_f \left(\left(F - \frac{\gamma_1}{\gamma_0} \right)^2 - \left(\frac{\gamma_1}{\gamma_0} - E \right)^2 - 2\delta \right) \\ &\geq 2\pi m_f \left(D_{\text{inf}}^2 - \left(\frac{M_f'}{m_f} \frac{N}{T} \right)^2 - 2\delta \right) \geq 2\pi m_f \left(D_{\text{inf}}^2/2 - 2\delta \right) \geq \pi m_f D_{\text{inf}}^2/2, \end{aligned}$$

where in the first inequality we have used

$$\left(\frac{\gamma_1}{\gamma_0} - E \right)^2 \leq \left(\frac{M_f'}{m_f} \frac{N}{T} \right)^2 \leq 1,$$

as we have $D_{\text{inf}} \leq 2$ and thus $\max \mathcal{N} \leq (m_f/M_f')T$ follows from condition (29). For the second inequality we have used the definition of D_{inf} and again condition (29), by which we have $D_{\text{inf}}^2/2 \geq \big(M_f' N/(m_f T)\big)^2$. Finally, for the third inequality we have used that by assumption in the Corollary $2\delta \leq D_{\text{inf}}^2/4$.

The first bound, $q(\delta) \geq \delta \pi m_f (1-\rho^2)$, implies that if

$$m > 2\Big(\frac{\pi m_f(1-\rho^2)}{20K_0}\frac{\delta}{1+\delta}\Big)^{\frac{3+8d}{1+4d}} \quad \text{and} \quad \min\mathcal{N} > \Big(\frac{\pi m_f(1-\rho^2)}{20K_0}\frac{\delta}{1+\delta}\Big)^{\frac{3+4d}{1+2d}} + 1$$

and

$$\min\mathcal{N} \geq 16\big(C_0\big)^3 \max\Big\{\frac{20(1+\delta)}{\delta\pi m_f(1-\rho^2)}, 1\Big\}\Big[6(2\pi M_f' + C) + 1\Big]$$

then Assumption 6 holds, and if

$$T \geq \max\Big\{12C_1, 12m\big(C_0\big)^3 M_f' \frac{20}{\pi m_f(1-\rho^2)}\frac{1+\delta}{\delta}\Big\},$$

then Assumption 7 holds. Hence, we have proven Lemma 3.2 where the constants can be chosen as

$$K_1 := 2\Big(\frac{\pi m_f(1-\rho^2)}{20K_0}\Big)^{\frac{3+8d}{1+4d}},$$

$$K_2 := \max\Big\{\Big(\frac{\pi m_f(1-\rho^2)}{20K_0}\Big)^{\frac{3+4d}{1+2d}} + 1,\\ 16\big(C_0\big)^3 \max\Big\{\frac{20(1+(1-\rho^2)/(2D_{\text{sup}}^2))}{\pi m_f(1-\rho^2)}, 1\Big\}\Big[6(2\pi M_f' + C) + 1\Big]\Big\}$$

and

$$K_3 := \max\Big\{12C_1, 12\big(C_0\big)^3 M_f' \frac{20}{\pi m_f(1-\rho^2)}\Big(1 + \frac{1-\rho^2}{2D_{\text{sup}}^2}\Big)\Big\}.$$

The second bound, $q(\delta) \geq \pi D_{\text{inf}}^2 m_f/2$, implies that if

$$m > 2\Big(\frac{\pi(M_f')^2}{20K_0 m_f(1+\delta)}\Big(\frac{\max\mathcal{N}}{T}\Big)^2\Big)^{\frac{3+8d}{1+4d}}$$

and

$$\min\mathcal{N} > \max\Big\{\Big(\frac{\pi(M_f')^2}{20K_0 m_f(1+\delta)}\Big(\frac{\max\mathcal{N}}{T}\Big)^2\Big)^{\frac{3+4d}{1+2d}} + 1,\\ 16\big(C_0\big)^3 \max\Big\{\frac{20(1+\delta)}{\pi D_{\text{inf}}^2 m_f/2}, 1\Big\}\Big[6(2\pi M_f' + C) + 1\Big]\Big\}$$

then Assumption 6 holds, and if

$$T \geq \max\Big\{12C_1, 12m\big(C_0\big)^3 M_f' \frac{20(1+\delta)}{\pi D_{\text{inf}}^2 m_f/2}\Big\}$$

then Assumption 7 holds. Hence, we have proven Lemma 3.3 where the constants can be chosen as

$$K_4 := 2\Big(\frac{\pi(M_f')^2}{20K_0 m_f}\Big)^{\frac{3+8d}{1+4d}},$$
$$K_5 := \max\Big\{\Big(\frac{\pi(M_f')^2}{20K_0 m_f}\Big)^{\frac{3+4d}{1+2d}} + 1, \\ 16\big(C_0\big)^3 \max\Big\{\frac{20(1+\frac{1}{8}D_{\text{inf}}^2))}{\pi D_{\text{inf}}^2 m_f/2}, 1\Big\}\Big[6(2\pi M_f' + C) + 1\Big]\Big\}$$

and

$$K_6 := 12\max\Big\{C_1, \big(C_0\big)^3 M_f' \frac{20(1+\frac{1}{8}D_{\text{inf}}^2)}{\pi D_{\text{inf}}^2 m_f/2}\Big\}.$$

This finishes the proof of Lemmas 3.2 and 3.3. □

Appendix B: Lemmas regarding a

B.1. Outlook

In this section we state and discuss results relating quantities that are encountered in connection with the localised Yule-Walker estimator. In Section B.2 we state and discuss three lemmas. In Lemma B.1 we make precise that $a_0^{(p)}(t/T)$ approximates the time-varying 1-step linear prediction coefficients which, for $p \in \mathbb{N}^*$ and $t = 1, \dots, T$, are defined as

$$\tilde{a}_T^{(p)}(t) := \arg\min_{a=(a_1,\dots,a_p)' \in \mathbb{R}^p} \mathbb{E}\Big[\Big(X_{t,T} - \sum_{j=1}^{p} a_j X_{t-j,T}\Big)^2\Big] = \big(\tilde{\Gamma}_T^{(p)}(t)\big)^{-1}\tilde{\gamma}_T^{(p)}(t),$$

where

$$\begin{aligned}\tilde{\gamma}_T^{(p)}(t) &:= \left(\mathrm{Cov}\left(X_{t,T}, X_{t-1,T}\right), \dots, \mathrm{Cov}\left(X_{t,T}, X_{t-p,T}\right)\right)', \\ \tilde{\Gamma}_T^{(p)}(t) &:= \left(\mathrm{Cov}\left(X_{t-i,T}, X_{t-j,T}\right);\, i,j = 1,\dots,d\right).\end{aligned} \tag{54}$$

In Lemma B.2 we make precise that $\bar{a}_{N,T}^{(p)}(t)$, defined in (35), is related to $a_\Delta^{(p)}(u)$, defined in (19), in the sense that $a_0^{(p)}(t/T)$ and $a_{N/T}^{(p)}(t/T)$ approximate $\bar{a}_{N,T}^{(p)}(t)$. In Lemma B.3 a bound for the norm of $a_\Delta^{(p)}(u)$ is provided, which is independent of p, Δ and u.

Proofs of the results in Section B.2 are provided in Section G.2 [32]. The proofs rely on results about expectations of localised autocovariance estimates from Section D and an approximation result for inverses of matrices (Lemma E.2). For the readers convenience, we include Figure 11 in which the dependence of the various results is illustrated graphically.

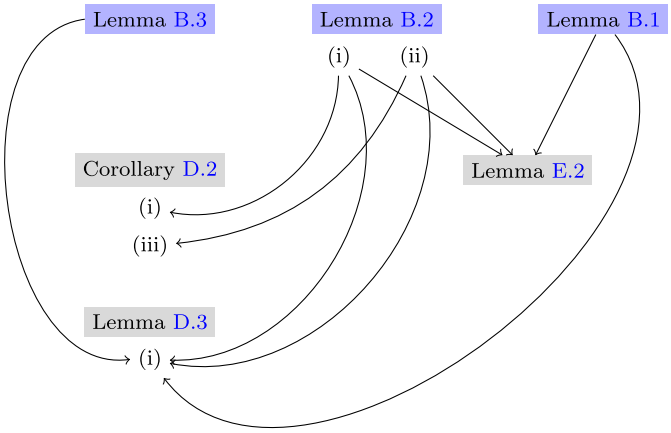

FIG 11. *Map of the lemmas in Section B.*

B.2. Statement of the lemmas

The following two lemmas discuss approximation properties of $\bar{a}_{N,T}^{(p)}(t)$ and $\tilde{a}_T^{(p)}(t)$:

Lemma B.1. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1, 3, 4, and $\mathbb{E}X_{t,T}=0$. Define C_0 and C_1 as in (36). Then, if $T \geq 2p^2C_1$, we have*

$$\|\tilde{a}_T^{(p)}(t) - a_0^{(p)}(t/T)\| \leq \frac{1}{T}\Big(5C_0C_1\,p^2\Big).$$

[53] prove a similar bound (Lemma 3):

$$\|\tilde{a}_T^{(p)}(t) - a_0^{(p)}(t/T)\| \leq \frac{D_1}{T}, \quad D_1 := \frac{Cp^{1/2}(p2^p+1)}{\pi m_f},$$

for $T \geq T_0 := \frac{Cp^{3/2}}{\pi m_f}$. Note that (for larger p) their constant D_1 can be substantially larger than the constant in Lemma B.1, which is largely due to a different representations of $\tilde{a}_T^{(p)}(t) - a_0^{(p)}(t/T)$ in their proof.

It is worth mentioning that in case of a stationary process, where $C_1 = 0$, Lemma B.1 implies that $\tilde{a}_T^{(p)}(t)$ and $a_0^{(p)}(t/T)$ coincide.

Lemma B.2. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1, 3, 4, and $\mathbb{E}X_{t,T}=0$. Define C_0 and C_1 as in (36). Then, if*

(i) $T \geq 8pNC_1$ and $N \geq 4p^2\frac{M_f}{m_f}$, then $\|\bar{a}_{N,T}^{(p)}(t) - a_0^{(p)}(t/T)\| \leq \big(9C_0C_1\big)\frac{p\,N}{T} + \big(3C_0^2\big)\frac{p^2}{N}$.

(ii) $T \geq 4p^2C_1$ and $N \geq 4p^2\frac{M_f}{m_f}$, then $\|\bar{a}_{N,T}^{(p)}(t) - a_{N/T}^{(p)}(t/T)\| \leq \big(5C_0C_1\big)\frac{p^2}{T} + \big(3C_0^2\big)\frac{p^2}{N}$.

Note that, if $p^2 = o(T)$, as $T \to \infty$, then we have, by Lemma B.1, that $\tilde{a}_T^{(p)}(t)$ and $a_0^{(p)}(t/T)$ are asymptotically equivalent in the sense that the Euclidean norm of the difference vanishes asymptotically. For $Np = o(T)$ and $p = o(N^{1/2})$ we have, by Lemma B.2(i), that $\bar{a}_{N,T}^{(p)}(t)$ and $a_0^{(p)}(t/T)$ are asymptotically equivalent, too. Therefore, since $0 \leq p^2 \leq Np$, we have: if $Np = o(T)$ and $p = o(N^{1/2})$, then $\bar{a}_{N,T}^{(p)}(t)$ and $\tilde{a}_T^{(p)}(t)$ are asymptotically equivalent. Note further, that in the case of a tvAR(p) model, the quantity $\tilde{a}_T^{(p)}(t)$ coincides with the vector of coefficients $(a_1(t/T), \ldots, a_p(t/T))$, as is evident from the Yule-Walker equations.

It is worth mentioning that in case of a stationary process, where $C_1 = 0$, the bounds in Lemmas B.1 and B.2 are independent of T.

We will also need the following result that bounds the norm of $a_\Delta^{(p)}(u)$:

Lemma B.3. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1, 3, 4, and $\mathbb{E}X_{t,T} = 0$. Then, for $u \in \mathbb{R}$, $p \in \mathbb{N}^*$ and $\Delta \geq 0$, we have*

$$\|a_\Delta^{(p)}(u)\| \leq (2\pi)^{1/2} M_f / m_f =: C_0.$$

By Lemma 2 in [53] we have $\|a_0^{(p)}(u)\| \leq 2^p$. Their proof adapts arguments from Lemma 4.2 in [20] where $\|\hat{a}_0^{(p)}(u)\| \leq 2^p$ almost surely is proven. We choose to work with the bound from Lemma B.3, because it has the advantage that it does not depend on p. Further, note that neither of the bounds is sharp, as by Cauchy-Schwarz inequality we clearly have $\|a_0^{(1)}(u)\| \leq 1$.

In Lemmas C.1(i) and (ii) we show similar bounds for the approximation of $\bar{v}_{N,T}^{(p,h)}(t)$ with $v_0^{(p,h)}(t/T)$ or $v_{N/T}^{(p,h)}(t/T)$.

Appendix C: Lemmas regarding v, g and MSPE

C.1. Outlook

In this section we state and discuss results relating quantities that are encountered in connection with the h-step ahead forecasting coefficients and the empirical mean squared prediction errors. In particular, this are the quantities $v_\Delta^{(p,h)}(u)$, $g_\Delta^{(p,h)}(u)$ and $\mathrm{MSPE}_{\Delta_1,\Delta_2}^{(p,h)}(u)$. In Section C.2 we state and discuss four lemmas. In Lemma C.1 we make precise that $\bar{v}_{N,T}^{(p,h)}(t)$ can be approximated by $v_0^{(p,h)}(t/T)$ or $v_{N/T}^{(p,h)}(t/T)$, where $\bar{v}_\Delta^{(p,h)}(u)$ was defined in (39). In Lemma C.2 we state bounds for norms of $v_\Delta^{(p,h)}(u)$ and its derivatives with with respect to u or Δ. In Lemma C.3, we state bounds for norms of $\bar{a}_{N,T}^{(p,h)}(t)$. In Lemma C.4(i)–(iii) we state bounds for $g_\Delta^{(p,h)}(u)$ and its derivates with respect to u or Δ. In Lemma C.4(iv)–(vi) we state bounds for the derivatives of $\mathrm{MSPE}_{\Delta_1,\Delta_2}^{(p,h)}(u)$ with respect to u, Δ_1 or Δ_2.

Proofs of the results in Section C.2 are provided in Section G.3 [32]. The proofs rely on some analogous bounds for the quantities related to the Yule-Walker estimator (Section B), on results on expectations of localised autocovariance estimates (Section D) and an approximation result for powers of matrices

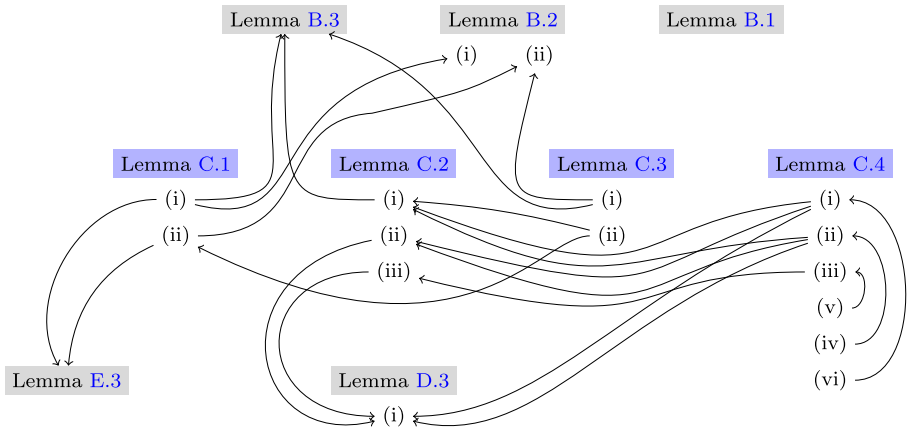

FIG 12. *Map of the lemmas in Section C.*

(Lemma E.3). For the readers convenience, we include Figure 12 in which the dependence of the various results is illustrated graphically.

C.2. Statement of the lemmas

The following lemma is constructed analogously to Lemma B.2, but for the h-step ahead coefficients.

Lemma C.1. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1, 3, 4, and $\mathbb{E}X_{t,T}=0$. Define C_0 and C_1 as in (36). Then, we have, for $\bar{v}_{N,T}^{(p,h)}(t)$ defined in (39),*

(i) if $T \geq 18C_1pN$ and $N \geq 6p^2C_0$, with $v_0^{(p,h)}(t/T)$ defined in (20), that

$$\|\bar{v}_{N,T}^{(p,h)}(t) - v_0^{(p,h)}(t/T)\| \leq h\big(2C_0\big)^h\Big(5C_1\frac{pN}{T} + 2\frac{p^2}{N}C_0\Big).$$

(ii) if $T \geq 10C_1p^2$ and $N \geq 6p^2C_0$, with $v_{N/T}^{(p,h)}(t/T)$ defined in (20), that

$$\|\bar{v}_{N,T}^{(p,h)}(t) - v_{N/T}^{(p,h)}(t/T)\| \leq h\big(2C_0\big)^h\Big(3C_1\frac{p^2}{T} + 2\frac{p^2}{N}C_0\Big).$$

Further, for the norms of $u \mapsto v_\Delta^{(p,h)}(u)$ and $\Delta \mapsto v_\Delta^{(p,h)}(u)$ and the derivatives, we have the following bounds:

Lemma C.2. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1, 3, 4, and $\mathbb{E}X_{t,T}=0$. C_0 as in (36) and m_f, M_f, M_f' from the assumptions. Then, with $v_\Delta^{(p,h)}(u)$ defined in (20), we have*

(i) $\|v_\Delta^{(p,h)}(u)\| \leq \big(C_0\big)^h$,

(ii) $v_\Delta^{(p,h)}(\cdot)$ is continuously differentiable, with

$$\Big\|\frac{\partial}{\partial u}v_\Delta^{(p,h)}(u)\Big\| \leq h\big(C_0\big)^h M_f'(m_f^{-1} + M_f^{-1}),$$

(iii) $\Delta \mapsto v_\Delta^{(p,h)}(u)$, $\Delta > 0$, *is continuously differentiable, with*

$$\left\| \frac{\partial}{\partial \Delta} v_\Delta^{(p,h)}(u) \right\| \leq 2h\big(C_0\big)^h M_f'(m_f^{-1} + M_f^{-1})/\Delta.$$

Lemma C.2 also holds for $h = 1$. Part (i) thus extends Lemma B.3.

Finally, we use Lemmas B.1, B.2 and B.3 to bound the norm of $\bar{a}_{N,T}^{(p)}(t)$ and $\bar{v}_{N,T}^{(p)}(t,h)$.

Lemma C.3. *Let* $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ *satisfy Assumptions 1, 3, 4, and* $\mathbb{E}X_{t,T} = 0$. *Define* C_0 *and* C_1 *as in (36). Then,*

(i) for $T \geq 10C_1p^2$ *and* $N \geq 6C_0p^2$ *we have, for* $\bar{a}_{N,T}^{(p,h)}(t)$ *defined in (35),*

$$\|\bar{a}_{N,T}^{(p)}(t)\| \leq 2C_0, \text{ and } \|\bar{v}_{N,T}^{(p,h)}(t)\|_\infty \leq \Big(2C_0 + 1\Big)^h.$$

Further, (ii) for $T \geq 6h2^hC_1p^2$ *and* $N \geq 4h2^hC_0p^2$ *we have, for* $\bar{v}_{N,T}^{(p,h)}(t)$ *defined in (39),*

$$\|\bar{v}_{N,T}^{(p,h)}(t)\| \leq 2\big(C_0\big)^h.$$

Note that Lemma C.3(i) implies that we have $\|\bar{v}_{N,T}^{(p,h)}(t)\| \leq p^{1/2}\Big(2C_0 + 1\Big)^h$. The bound in Lemma C.3(ii) does not depend on p, but require larger T and N.

An important observation is that, as a function of u, $\mathrm{MSPE}_{N/T,n/T}^{(p,h)}(u)$ is differentiable with bounded derivative

Lemma C.4. *Let* $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ *satisfy Assumptions 1, 3, 4, and* $\mathbb{E}X_{t,T} = 0$. *Define* C_0 *as in (36) and the other constants from the assumptions. Then, the function* $g_\Delta^{(p,h)}$, *defined in (22), is continuously differentiable and the derivatives are bounded. More precisely, we have*

(i)

$$\big|g_\Delta^{(p,h)}(u)\big| \leq 4M_f\big(C_0\big)^{2h},$$

(ii)

$$\left|\frac{\partial}{\partial u} g_\Delta^{(p,h)}(u)\right| \leq 4\big(2h+1\big)\big(C_0\big)^{2h+1} M_f',$$

(iii)

$$\left|\frac{\partial}{\partial \Delta} g_\Delta^{(p,h)}(u)\right| \leq 8\big(2h+1\big)\big(C_0\big)^{2h+1} M_f'/\Delta.$$

In particular, this implies, or $\mathrm{MSPE}_{\Delta_1,\Delta_2}^{(p,h)}(u)$ *defined in (21), that*

(iv)

$$\left|\frac{\partial}{\partial u} \mathrm{MSPE}_{\Delta_1,\Delta_2}^{(p,h)}(u)\right| \leq 4\big(2h+1\big)\big(C_0\big)^{2h+1} M_f'.$$

(v)

$$\left|\frac{\partial}{\partial \Delta_1} \mathrm{MSPE}_{\Delta_1,\Delta_2}^{(p,h)}(u)\right| \leq 8\big(2h+1\big)\big(C_0\big)^{2h+1} M_f'/\Delta_1.$$

(vi)

$$\Big|\frac{\partial}{\partial\Delta_2}\mathrm{MSPE}^{(p,h)}_{\Delta_1,\Delta_2}(u)\Big| \leq 8M_f\big(C_0\big)^{2h}/\Delta_2.$$

Appendix D: Properties of empirical localised autocovariances

D.1. Outlook

In this section we establish properties of the empirical localised autocovariances under local stationarity. In Section D.2 we state three lemmas about the estimators' moments and in Section D.3 we state two lemmas about the concentration of the estimators. More precisely, in Lemma D.1 we approximate the expectation of the empirical autocovariance and state bounds for the approximation error. In Corollary D.2 we employ the approximation results from Lemma D.1 to approximate matrices of such expectations and bound the approximation error (in spectral norm). In Lemma D.3 we establish lower and upper bounds for the eigenvalues of $\Gamma^{(p)}_{\Delta}(u)$ and $\mathbb{E}\hat{\Gamma}^{(p)}_{N,T}(t)$. In Lemma D.4 we establish a concentration result for the localised empirical autocovariance. Lemma D.4 follows as a special case from Lemma D.5 where a concentration result for generalised sums under local stationarity is established.

Proofs of the results are proved in Section G.4 [32]. The proofs rely on technical results to bound the matrix norm of perturbed inverse matrices and approximation of sums by integrals (cf. Section E) as well as on general concentration results from [55] which we cite in Section H [32]. For the readers convenience, we include Figure 13 in which the dependence of the various results is illustrated graphically.

D.2. Approximations for moments

Lemma D.1. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1 and 4, and $\mathbb{E}X_{t,T}=0$. Then, with $\hat{\gamma}_{k;N,T}(t)$ defined in (3), $f(u,\lambda)$ and C from Assumption 1, and M'_f from Assumption 4, we have: (i)*

$$\Big|\mathbb{E}\hat{\gamma}_{k;N,T}(t) - \frac{N-|k|}{N}\int_{-\pi}^{\pi}\Big[\int_0^1 f\big(\frac{t-N+|k|}{T}+\frac{N-|k|}{T}u,\lambda\big)\mathrm{d}u\Big]\mathrm{e}^{\mathrm{i}|k|\lambda}\mathrm{d}\lambda\Big| \leq \frac{2\pi M'_f + C}{T}$$

and (ii)

$$\Big|\mathbb{E}\hat{\gamma}_{k;N,T}(t) - \frac{N-|k|}{N}\int_{-\pi}^{\pi}\Big[\int_0^1 f\big(\frac{t-N}{T}+\frac{N}{T}u,\lambda\big)\mathrm{d}u\Big]\mathrm{e}^{\mathrm{i}|k|\lambda}\mathrm{d}\lambda\Big| \leq \frac{2\pi(|k|+1)M'_f + C}{T}$$

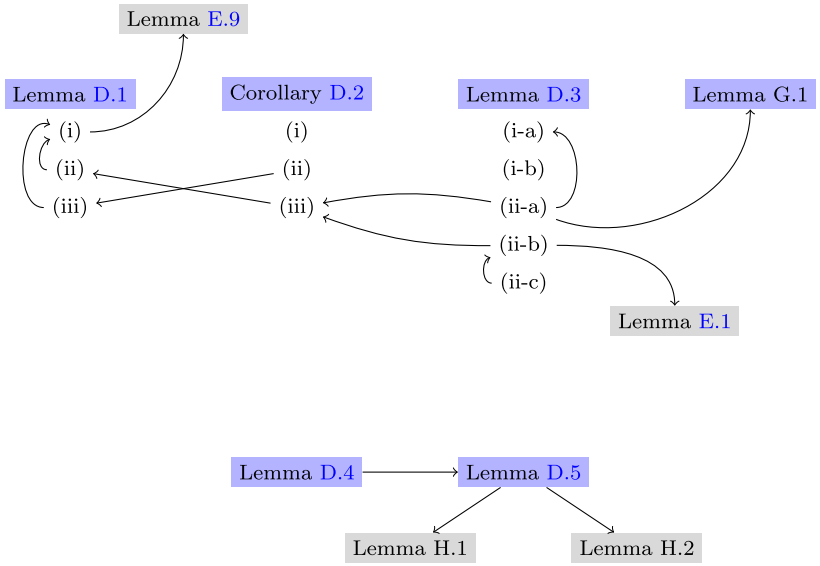

FIG 13. *Map of the lemmas in Section D.*

and (iii)

$$\left|\mathbb{E}\hat{\gamma}_{k;N,T}(t) - \frac{N-|k|}{N}\gamma_k(t/T)\right| \leq \frac{2\pi M_f'(N-|k|+1)+C}{T}.$$

Corollary D.2. *Under the conditions of Lemma D.1, with* $\tilde{\Gamma}_T^{(p)}(t)$ *and* $\tilde{\gamma}_T^{(p)}(t)$ *defined in (54),* $\Gamma_\Delta^{(p)}(u)$ *and* $\gamma_\Delta^{(p)}(u)$ *defined in (18),* $\hat{\Gamma}_{N,T}^{(p)}(t)$ *and* $\hat{\gamma}_{N,T}^{(p)}(t)$ *defined in (2), and* $F_{p,n}$ *and* $f_{p,n}$ *defined for any* $n = 1, 2, \ldots$ *and* $p = 1, \ldots, n$ *as*

$$F_{p,N} := \big(1 - |j-k|/N\big)_{j,k=1,\ldots,p}, \quad \text{and} \quad f_{p,N} := \big(1 - 1/N, \ldots, 1 - p/N\big)',$$

we have: (i)

$$\|\tilde{\Gamma}_T^{(p)}(t) - \Gamma_0^{(p)}(t/T)\| \leq \frac{p^2}{T}(2\pi M_f' + C) \qquad \|\tilde{\gamma}_T^{(p)}(t) - \gamma_0^{(p)}(t/T)\| \leq \frac{p^{1/2}}{T}C$$

and (ii)

$$\|\mathbb{E}\hat{\Gamma}_{N,T}^{(p)}(t) - F_{p,N} \circ \Gamma_0^{(p)}(t/T)\| \leq \frac{p}{T}(2\pi M_f'(N+1)+C)$$
$$\|\mathbb{E}\hat{\gamma}_{N,T}^{(p)}(t) - f_{p,N} \circ \gamma_0^{(p)}(t/T)\| \leq \frac{p^{1/2}}{T}(2\pi M_f' N + C)$$

and (iii)

$$\|\mathbb{E}\hat{\Gamma}_{N,T}^{(p)}(t) - F_{p,N} \circ \Gamma_{N/T}^{(p)}(t/T)\| \leq \frac{p^2}{T}(2\pi M_f' + C)$$

$$\|\mathbb{E}\hat{\gamma}^{(p)}_{N,T}(t) - f_{p,N} \circ \gamma^{(p)}_{N/T}(t/T)\| \leq 2\frac{p^{3/2}}{T}\Big(2\pi M'_f + C\Big)$$

Lemma D.3. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1, 3, and 4, and $\mathbb{E}X_{t,T} = 0$. Then, we have:*

(i-a) the matrices $\Gamma^{(p)}(u)$ and $\Gamma^{(p)}_\Delta(u)$ are positive definite, hence invertible, for $u \in \mathbb{R}$ and $\Delta \geq 0$, with their eigenvalues between m_f and M_f. In other words, the norms of the matrices and their inverses are uniformly bounded:

$$m_f \leq 1/\|\Gamma^{(p)}_\Delta(u)^{-1}\| \leq \|\Gamma^{(p)}_\Delta(u)\| \leq M_f.$$

(i-b) the norms of the respective vectors are uniformly bounded:

$$\|\gamma^{(p)}_\Delta(u)\| \leq (2\pi)^{1/2} M_f.$$

(ii-a) The largest eigenvalue of $\mathbb{E}\hat{\Gamma}^{(p)}_{N,T}(t)$ satisfies the following bound:

$$\|\mathbb{E}\hat{\Gamma}^{(p)}_{N,T}(t)\| \leq M_f + \frac{p^2}{T}(2\pi M'_f + C).$$

(ii-b) if $T > m_f^{-1}p^2(2\pi M'_f + C)$, then the matrix $\mathbb{E}\hat{\Gamma}^{(p)}_{N,T}(t)$ is positive definite, and the smallest eigenvalue satisfies the following bound:

$$m_f - \frac{p^2}{T}(2\pi M'_f + C) \leq \frac{1}{\|\big(\mathbb{E}\hat{\Gamma}^{(p)}_{N,T}(t)\big)^{-1}\|}$$

(ii-c) in particular, if $T \geq 2m_f^{-1}p^2(2\pi M'_f + C)$ we thus have

$$\frac{1}{2}m_f \leq \frac{1}{\|\big(\mathbb{E}\hat{\Gamma}^{(p)}_{N,T}(t)\big)^{-1}\|} \leq \|\mathbb{E}\hat{\Gamma}^{(p)}_{N,T}(t)\| \leq \frac{3}{2}M_f.$$

D.3. Exponential inequalities for empirical covariances

We now state an exponential inequalities for the empirical covariances:

Lemma D.4. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1, 2, 3, and 5 and $\mathbb{E}X_{t,T} = 0$. Then, for $T \geq C/(\pi m_f)$, $n \in \mathbb{N}^*$, $h \in \mathbb{N}$ and $\varepsilon > 0$, we have*

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\hat{\gamma}_{h;N,T}(t) - \mathbb{E}\hat{\gamma}_{h;N,T}(t)\Big| \geq \varepsilon\Big) \\
&\leq \exp\left(-\frac{\varepsilon^2}{2\Big(\frac{C_{1,1}h_*}{N-|h|} + \varepsilon^{(3+4d)/(2+2d)}\Big(\frac{C_{2,1}h_*}{N-|h|}\Big)^{1/(2+2d)}\Big)}\right) \\
&\leq \begin{cases} \exp\Big(-\dfrac{\varepsilon^2}{4}\dfrac{N-|h|}{C_{1,1}h_*}\Big) & \varepsilon \leq \Big(\dfrac{h_*}{N-|h|}\Big)^{(1+2d)/(3+4d)}\Big(\dfrac{C_{1,1}^{2+2d}}{C_{2,1}}\Big)^{1/(3+4d)} \\ \exp\Big(-\dfrac{1}{4}\Big(\varepsilon\dfrac{N-|h|}{C_{2,1}h_*}\Big)^{1/(2+2d)}\Big) & \varepsilon \geq \Big(\dfrac{h_*}{N-|h|}\Big)^{(1+2d)/(3+4d)}\Big(\dfrac{C_{1,1}^{2+2d}}{C_{2,1}}\Big)^{1/(3+4d)} \end{cases},
\end{aligned}$$

where $h_* := |h| + I\{h = 0\}$, $\hat{\gamma}_{h;N,T}(t)$ is defined in (3), and the constants $C_{1,1}$ and $C_{2,1}$ are defined in (42).

Note that the right hand side does not depend on t.

Lemma D.5. *Let $(X_{t,T})_{t\in\mathbb{Z},T\in\mathbb{N}^*}$ satisfy Assumptions 1, 2, 3, and 5 and $\mathbb{E}X_{t,T} = 0$. Let a_t, $t = b,\dots,b+n-1$ be a bounded sequence of numbers; i.e., $|a_t| \le A$. Then, for $\alpha \in \mathbb{N}^*$, $T \ge C/(\pi m_f)$, $n \in \mathbb{N}^*$, $b \in \mathbb{Z}$, $h \in \mathbb{N}$ and $\varepsilon > 0$, we have*

$$\begin{aligned}
&\mathbb{P}\Big(\Big|n^{-1}\sum_{t=b}^{b+n-1} a_t(X^\alpha_{t,T}X^\alpha_{t+h,T} - \mathbb{E}(X^\alpha_{t,T}X^\alpha_{t+h,T}))\Big| > \varepsilon\Big)\\
&\quad\le \exp\left(-\frac{\varepsilon^2}{2\Big(\frac{C_{1,\alpha}A^2h_*}{n} + \varepsilon^{(3+4\alpha d)/(2+2\alpha d)}\Big(\frac{C_{2,\alpha}Ah_*}{n}\Big)^{1/(2+2\alpha d)}\Big)}\right)\\
&\quad\le \begin{cases}
\exp\Big(-\dfrac{(\varepsilon/A)^2}{4}\dfrac{n}{C_{1,\alpha}h_*}\Big)\\
\quad \varepsilon \le A\Big(\dfrac{C_{1,\alpha}^{2+2\alpha d}}{C_{2,\alpha}}\Big)^{1/(3+4\alpha d)}\Big(\dfrac{h_*}{n}\Big)^{(1+2\alpha d)/(3+4\alpha d)}\\
\exp\Big(-\dfrac{1}{4}\Big(\dfrac{\varepsilon}{A}\dfrac{n}{C_{2,\alpha}h_*}\Big)^{1/(2+2\alpha d)}\Big)\\
\quad \varepsilon \ge A\Big(\dfrac{C_{1,\alpha}^{2+2\alpha d}}{C_{2,\alpha}}\Big)^{1/(3+4\alpha d)}\Big(\dfrac{h_*}{n}\Big)^{(1+2\alpha d)/(3+4\alpha d)},
\end{cases}
\end{aligned}$$

where $h_* := |h| + I\{h = 0\}$ and the constants $C_{1,\alpha}$ and $C_{2,\alpha}$ are defined in (42) in the proof [depending only on α, d, C, M_f, ρ, and K].

Note the important fact that the bounds in the inequality do not depend on b.

Appendix E: Technical results

In the previous sections we used the following general results, which are not restricted to locally stationary processes. In some of these technical lemmas we denote by $\|\cdot\|_M$ or $\|\cdot\|_v$ an arbitrary matrix or vector norm, respectively. Special properties we require include submultiplicativity of a matrix norm, and compatibility of a matrix norm with a vector norm. A matrix norm which satisfies $\|AB\|_M \le \|A\|_M\|B\|_M$ for all square matrices ($m = n$), is said to be submultiplicative. A matrix norm $\|\cdot\|_M$ and vector norm $\|\cdot\|_v$ are said to be compatible if $\|Ax\|_v \le \|A\|_M\|x\|_v$ for all square matrices A and vectors x (of sizes that allow for the matrix product).

Lemma E.1. *Let $A \in \mathbb{R}^{p\times p}$ be an invertible matrix and $\Delta \in \mathbb{R}^{p\times p}$ be a matrix with $\|A^{-1}\|_M \cdot \|\Delta\|_M < 1$ for a submultiplicative matrix norm $\|\cdot\|_M$. Then, the matrix $A + \Delta$ is invertible and we have*

$$\|(A+\Delta)^{-1}\|_M \le \frac{\|A^{-1}\|_M}{1 - \|A^{-1}\|_M \cdot \|\Delta\|_M}$$

An important corollary to the above lemma is the following:

Lemma E.2. *Let $A \in \mathbb{R}^{p\times p}$ be an invertible matrix and $\Delta \in \mathbb{R}^{p\times p}$ be a matrix with $\|A^{-1}\|_M \cdot \|\Delta\|_M \le c < 1$ for a submultiplicative matrix norm $\|\cdot\|_M$. Then, the matrix $A+\Delta$ is invertible and we have*

$$\|(A+\Delta)^{-1} - A^{-1}\|_M \le \|\Delta\|_M \frac{\|A^{-1}\|_M^2}{1-c}.$$

Lemma E.3. *Let A and A_0 be two square matrices and $\|\cdot\|_M$ be a submultiplicative matrix norm. Then, for any $h \in \mathbb{N}$,*

$$\|A^h - A_0^h\|_M \le h\|A - A_0\|_M \big(\|A - A_0\|_M + \|A_0\|_M\big)^{h-1}.$$

Lemma E.4. *Let u and v be two real-valued random variables. Further, let u_0 and v_0 be two real numbers. Then, for all $\varepsilon > 0$*

$$\begin{aligned}&\mathbb{P}(|uv - u_0v_0| > \varepsilon)\\ &\quad\le \mathbb{P}\Big(|u-u_0| > \frac{1}{2}\frac{\varepsilon}{(|v_0|^2+\varepsilon)^{1/2}}\Big) + \mathbb{P}\Big(|v-v_0| > \frac{1}{2}\frac{\varepsilon}{(|u_0|^2+\varepsilon)^{1/2}}\Big).\end{aligned}$$

For the proof in the main part we need the following lemma:

Lemma E.5. *Let X_t and $\hat{a}_t$, $t = 1,\dots,n$, be two sequences of random variables, and α_t, $t = 1,\dots,n$ be a sequence of numbers. Assume that there exists a constant $m_2^2 > 0$ such that $\max_{t=1,\dots,n} \mathbb{E}X_t^2 \le m_2^2 < \infty$. Then, for any $\varepsilon > 0$, we have*

$$\begin{aligned}&\mathbb{P}\Big(\Big|\sum_{t=1}^n \big(\hat{a}_tX_t - \alpha_t\mathbb{E}(X_t)\big)\Big| > n\varepsilon\Big)\\ &\le \mathbb{P}\Big(\sup_{t=1,\dots,n}|\hat{a}_t - \alpha_t| > \frac{\varepsilon}{2\big((2m_2^2)^2+\varepsilon^2\big)^{1/4}}\Big)\\ &\qquad + \mathbb{P}\Big(\Big|\sum_{t=1}^n (X_t^2 - \mathbb{E}X_t^2)\Big| > n\varepsilon/2\Big) + \mathbb{P}\Big(\Big|\sum_{t=1}^n \alpha_t(X_t - \mathbb{E}X_t)\Big| > n\varepsilon/2\Big).\end{aligned}$$

We will further use the following lemmas:

Lemma E.6. *Let $M \in \mathbb{R}^{p\times p}$ be a random $p\times p$ matrix with existing expectation $M_0 := \mathbb{E}M$, which is assumed to be invertible. Further, let v be a $\mathbb{R}^p$-valued random vector v with existing expectation $\mathbb{E}v := v_0$. Then, for every submultiplicative matrix norm $\|\cdot\|_M$ that is compatible with the (vector) norm $\|\cdot\|_v$, we have: for every $\varepsilon > 0$*

$$\begin{aligned}&\mathbb{P}\Big(\Big\|M^{-1}v - M_0^{-1}v_0\Big\|_v > \varepsilon\Big)\\ &\le \mathbb{P}\Big(\|M - M_0\|_M > \frac{1}{2\|M_0^{-1}\|_M}\Big) + \mathbb{P}\Big(\|v - v_0\|_v > \frac{\varepsilon}{4}\frac{1}{\|M_0^{-1}\|_M}\Big)\\ &\qquad + \mathbb{P}\Big(\|M - M_0\|_M > \frac{\varepsilon}{4}\frac{1}{(\|M_0^{-1}\|_M)^2\,\|v_0\|_v}\Big) I\{\|v_0\|_v \ne 0\}.\end{aligned}$$

Lemma E.7. *Let $x = (x_1, \dots, x_p)$ be a random vector and $x_0 = (x_{0,1}, \dots, x_{0,p})$ be a deterministic vector. Define two $p \times p$ matrices*

$$A := \begin{pmatrix} x_1 & x_2 & \cdots & x_{p-1} & x_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \quad \text{and } A_0 := \begin{pmatrix} x_{0,1} & x_{0,2} & \cdots & x_{0,p-1} & x_{0,p} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

For any $h = 1, 2, \dots$ define $v := (1, 0, \dots, 0)A^h$ and $v_0 := (1, 0, \dots, 0)A_0^h$. Then, for every $\varepsilon > 0$,

$$\mathbb{P}(\|v - v_0\| > \varepsilon) \le \mathbb{P}\Big(\|x - x_0\| > 2^{1-h} \frac{\varepsilon}{\varepsilon + h(\max\{\|x_0\|, 1\})^{h-1}}\Big).$$

The following lemma ensures that b-sub-Gaussian processes satisfy Assumption 5.

Lemma E.8. *Let $(X_{t,T})_{t\in\mathbb{Z}, T\in\mathbb{N}^*}$ satisfy Assumptions 1 and 3. Assume that $T \ge \frac{C}{\pi m_f}$ and that the standardized variables $X_{t,T}/\sigma_{t,T}$ are b-sub-Gaussian ($b > 0$); i.e.,*

$$\mathbb{E}\Big(\exp\left(\xi X_{t,T}/\sigma_{t,T}\right)\Big) \le \exp\Big(\frac{b^2|\xi|^2}{2}\Big), \quad \xi \in \mathbb{R}.$$

Then, the process $(X_{t,T})_{t\in\mathbb{Z}, T\in\mathbb{N}^*}$ satisfies Assumption 5 with $c := 6\pi b M_f$ and $d := 1/2$.

Lemma E.9. *Let $f : [0,1] \to \mathbb{R}$ be continuous and differentiable on $(0,1)$. Then, for every $A, B = 0, \dots, T$, $T \in \mathbb{N}_*$, $A < B$, we have*

$$\Big|\frac{1}{B-A}\sum_{\ell=A+1}^{B} f(\ell/T) - \int_0^1 f\Big(\frac{A}{T} + \frac{B-A}{T}u\Big)\mathrm{d}u\Big| \le \frac{1}{T}\sup_{A/T<u<B/T}|f'(u)|.$$

References

[1] Akaike, H. (1969). Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics, 21(1):243–247. MR0246476

[2] Akaike, H. (1970). A fundamental relation between predictor identification and power spectrum estimation. Annals of the Institute of Statistical Mathematics, 22(1):219–223. MR0286234

[3] Baillie, R. T. (1979). Asymptotic prediction mean squared error for vector autoregressive models. Biometrika, 66(3):675–8. MR0556749

[4] Bercu, B. (2001). On large deviations in the Gaussian autoregressive process: stable, unstable and explosive cases. Bernoulli, 7(2):299–316. MR1828507

[5] Bercu, B., Gamboa, F., and Lavielle, M. (2000). Sharp large deviations for Gaussian quadratic forms with applications. ESAIM: Probability and Statistics, 4:1–24. MR1749403

[6] Bercu, B., Gamboa, F., and Rouault, A. (1997). Large deviations for quadratic forms of stationary Gaussian processes. Stochastic Processes and their Applications, 71(1):75–90. MR1480640

[7] Bercu, B. and Touati, A. (2008). Exponential inequalities for self-normalized martingales with applications. The Annals of Applied Probability, 18(5):1848–1869. MR2462551

[8] Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. Annals of the Institute of Statistical Mathematics, 48:577–602. MR1424784

[9] Birr, S., Volgushev, S., Kley, T., Dette, H., and Hallin, M. (2017). Quantile spectral analysis for locally stationary time series. Journal of the Royal Statistical Society: Series B, 79(5):1619–1643. MR3731679

[10] Brillinger, D. R. (1975). Time Series: Data Analysis and Theory. Holt, Rinehart and Winston, Inc., New York. MR0443257

[11] Brockwell, P. J. and Davis, R. A. (1991). Time Series: Theory and Methods. Springer, New York. MR2839251

[12] Brownlees, C. T. and Gallo, G. M. (2008). On variable selection for volatility forecasting: The role of focused selection criteria. Journal of Financial Econometrics, 6:513–539.

[13] Chan, N. H. and Wei, C. Z. (1987). Asymptotic inference for nearly nonstationary AR(1) processes. Ann. Statist., 15(3):1050–1063. MR0902245

[14] Chen, Y., Hä rdle, W., and Pigorsch, U. (2010). Localized realized volatility modeling. Journal of the American Statistical Association, 105(492):1376–1393. MR2796557

[15] Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007). Prediction-focused model selection for autoregressive models. Australian & New Zealand Journal of Statistics, 49(4):359–379. MR2413576

[16] Claeskens, G. and Hjort, N. L. (2003). The focused information criterion [with discussion]. Journal of the American Statistical Association, 98:900–916. MR2041482

[17] Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. Stochastic Processes and their Applications, 62(1):139–168. MR1388767

[18] Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. Annals of Statistics, 25(1):1–37. MR1429916

[19] Dahlhaus, R. (2012). Locally stationary processes. In Rao, T. S., Rao, S. S., and Rao, C., editors, Time Series Analysis: Methods and Applications, volume 30 of Handbook of Statistics, pages 351–413. Elsevier. MR3295420

[20] Dahlhaus, R. and Giraitis, L. (1998). On the optimal segment length for parameter estimates for locally stationary time series. Journal of Time Series Analysis, 19(6):629–655. MR1665941

[21] Das, S. and Politis, D. N. (2018). Predictive inference for locally stationary time series with an application to climate data. arXiv:1712.02383.

[22] Dette, H., Preuß, P., and Vetter, M. (2011). A measure of stationarity in locally stationary processes with applications to testing. Journal of the American Statistical Association, 106(495):1113–1124. MR2894768

[23] Dwivedi, J. and Subba Rao, S. (2010). A test for second order stationarity based on the discrete fourier transform. Journal of Time Series Analysis, 32:68–91. MR2790673

[24] Dzhaparidze, K., Kormos, J., van der Meer, T., and van Zuijlen, M. (1994). Parameter estimation for nearly nonstationary AR(1) processes. Mathematical and Computer Modelling, 19(2):29–41. MR1265022

[25] Fryzlewicz, P. and Subba Rao, S. (2011). Mixing properties of ARCH and time-varying ARCH processes. Bernoulli, 17(1):320–346. MR2797994

[26] Giraitis, L., Kapetanios, G., and Price, S. (2013). Adaptive forecasting in the presence of recent and ongoing structural change. Journal of Econometrics, 177(2):153–170. MR3118553

[27] Hallin, M. (1978). Mixed autoregressive-moving average multivariate processes with time-dependent coefficients. Journal of Multivariate Analysis, 8(4):567–572. MR0520964

[28] Hong-Zhi, A., Zhao-Guo, C., and Hannan, E. J. (1982). Autocorrelation, autoregression and autoregressive approximation. Annals of Statistics, 10(3):926–936. MR0663443

[29] Jirak, M. (2012). Simultaneous confidence bands for Yule-Walker estimators and order selection. Annals of Statistics, 40(1):494–528. MR3014315

[30] Jirak, M. (2014). Simultaneous confidence bands for sequential autoregressive fitting. Journal of Multivariate Analysis, 124:130–149. MR3147316

[31] Kley, T., Fryzlewicz, P., and Preuß, P. (2019a). forecastSNSTS: Forecasting for Stationary and Non-Stationary Time Series. R package version 1.3-0.

[32] Kley, T., Preuß, P., and Fryzlewicz, P. (2019b). Predictive, finite-sample model choice for time series under stationarity and non-stationarity. arXiv:1611.04460v3.

[33] Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Annals of Statistics, 10(1):154–166. MR0642726

[34] Landregistry (2019). UK house price index. `http://landregistry.data.gov.uk/app/ukhpi`. Accessed: 2019-01-11.

[35] McDonald, D. J., Shalizi, C. R., and Schervish, M. (2016). Nonparametric risk bounds for time-series forecasting. arXiv:1212.0463. MR3646627

[36] Mikosch, T. and Starica, C. (2004). Non-stationarities in financial time series, the long range dependence and the IGARCH effects. The Review of Economics and Statistics, 86:378–390. MR2327875

[37] Moulines, E., Priouret, P., and Roueff, F. (2005). On recursive estimation for time varying autoregressive processes. Annals of Statistics, 33(6):2610–2654. MR2253097

[38] Nason, G. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. Journal of the Royal Statistical Society: Series B, 75:879–904. MR3124795

[39] Palma, W., Olea, R., and Ferreira, G. (2013). Estimation and forecasting of locally stationary processes. Journal of Forecasting, 32:86–96. MR3017832

[40] Paparoditis, E. (2009). Testing temporal constancy of the spectral structure

of a time series. Bernoulli, 15:1190–1221. MR2597589

[41] Paparoditis, E. (2010). Validating stationarity assumptions in time series analysis by rolling local periodograms. Journal of the American Statistical Association, 105:839–851. MR2724865

[42] Peña, D. and Sánchez, I. (2007). Measuring the advantages of multivariate vs. univariate forecasts. Journal of Time Series Analysis, 28(6):886–909. MR2413412

[43] Politis, D. N. (2015). Model-Free Prediction and Regression: A Transformation-Based Approach to Inference. Springer. MR3442999

[44] Preuß, P. and Vetter, M. (2013). Discriminating between long-range dependence and non-stationarity. Electron. J. Statist., 7:2241–2297. MR3108814

[45] Preuß, P., Vetter, M., and Dette, H. (2013). A test for stationarity based on empirical processes. Bernoulli, 19(5B):2715–2749. MR3160569

[46] Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. Journal of the Royal Statistical Society. Series B (Methodological), 27(2):204–237. MR0199886

[47] Priestley, M. B. (1981). Spectral Analysis and Time Series. Academic Press. MR0628735

[48] R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[49] Reinsel, G. (1980). Asymptotic properties of prediction errors for the multivariate autoregressive model using estimated parameters. Journal of the Royal Statistical Society: Series B, 42(3):328–333. MR0596161

[50] Richter, S. and Dahlhaus, R. (2017). Cross validation for locally stationary processes. arXiv:1705.10046. MR3953447

[51] Rohan, N. and Ramanathan, T. V. (2011). Order selection in arma models using the focused information criterion. Australian & New Zealand Journal of Statistics, 53(2):217–231. MR2851723

[52] Rossi, B. (2013). Advances in forecasting under instability. In Elliott, G. and Timmermann, A., editors, Handbook of Economic Forecasting, volume 2B, pages 1203–1324. Elsevier.

[53] Roueff, F. and Sanchez-Perez, A. (2016). Locally stationary processes prediction by auto-regression. arXiv:1602.01942v1. MR3867204

[54] Roueff, F. and Sanchez-Perez, A. (2018). Prediction of weakly locally stationary processes by auto-regression. arXiv:1602.01942v3. MR3867204

[55] Saulis, L. and Statulevičus, V. A. (1991). Limit Theorems for Large Deviations. Kluwer, Dordrecht. MR1171883

[56] Subba Rao, T. (1970). The fitting of non-stationary time-series models with time-dependent parameters. Journal of the Royal Statistical Society. Series B (Methodological), 32(2):312–322. MR0269065

[57] Vogt, M. (2012). Nonparametric regression for locally stationary time series. Annals of Statistics, 40(5):2601–2633. MR3097614

[58] Vogt, M. and Dette, H. (2015). Detecting smooth changes in locally stationary processes. Annals of Statistics, 43(2):713–740. MR3319141

[59] von Sachs, R. and Neumann, M. H. (2000). A wavelet-based test for stationarity. Journal of Time Series Analysis, 21:597–613. MR1794489

[60] Xia, Y. and Tong, H. (2011). Feature matching in time series modeling. Statistical Science, 26(1):21–46. MR2849904

[61] Yu, M. and Si, S. (2009). Moderate deviation principle for autoregressive processes. Journal of Multivariate Analysis, 100(9):1952–1961. MR2543078

[62] Zhang, Y. and Koreisha, S. (2015). Adaptive order determination for constructing time series forecasting models. Communications in Statistics - Theory and Methods, 44(22):4826–4847. MR3424811