SFB
823

Prediction in locally stationary time series

Discussion Paper

Holger Dette, Weichi Wu

Nr. 1/2020

SFB
823

Prediction in locally stationary time series

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany

Weichi Wu
Tsinghua University
Center for Statistics
Department of Industrial Engineering
10084 Beijing China

January 7, 2020

Abstract

We develop an estimator for the high-dimensional covariance matrix of a locally stationary process with a smoothly varying trend and use this statistic to derive consistent predictors in non-stationary time series. In contrast to the currently available methods for this problem the predictor developed here does not rely on fitting an autoregressive model and does not require a vanishing trend. The finite sample properties of the new methodology are illustrated by means of a simulation study and a data example.

AMS subject classification: 62M10; 62M20
Keywords and phrases: locally stationary time series, high dimensional auto-covariance, matrices, prediction, local linear regression, prediction

1 Introduction

An important problem in time series analysis is to predict or forecast future observations from a given a stretch of data, say $X_1, \ldots, X_n$, and numerous authors have worked on this problem. Meanwhile there is a well developed theory for prediction under the assumption of stationary processes [see for example Brockwell et al. (2002), Bickel and Gel (2011), McMurry and Politis (2015) among many others]. On the other hand, if data is obtained over a long stretch of time it may be unrealistic to assume that the stochastic structure of

a time series is stable. Moreover, in many shorter time series non-stationarity can also be observed and prediction under the assumption of stationarity might be misleading.
A common approach to deal with this problem of non-stationarity is to assume a location scale model with a smoothly changing trend and variance but a stationary error process, say $X_n = \mu(n) + \sigma(n)\varepsilon_n$ [see, for example, Van Bellegem and Von Sachs (2004), Stărică and Granger (2005), Zhao and Wu (2008), Guillaumin et al. (2017), Das and Politis (2017)]. In this case the trend and variance function can be estimated and prediction can be performed applying methods for stationary data to the standardized residuals. However, there appear also more sophisticated features of non-stationarity in the data, which are not captured by a a simple location scale model, such as time-changing kurtosis or skewness, and the standardized residuals obtained by this procedure may not be stationary.
To address this type of non-stationarity various mathematical concepts modeling a slowly-changing stochastic structure have been developed in the literature [see for example, Priestley (1988), Dahlhaus (1997), Nason et al. (2000), Zhou and Wu (2009) or Vogt (2012)]. The corresponding stochastic processes are usually called *locally stationary* and the problem of predicting future observations in these models is a very challenging one. An early reference is Fryzlewicz et al. (2003) who considered centered *locally stationary wavelet processes*. In this model the sample covariance matrix in the prediction equation is not estimable and the authors proposed an approximation using the (uniquely defined) wavelet spectrum. Van Bellegem and Von Sachs (2004) considered the prediction problem in a location scale model with a smoothly changing variance and stationary error process. More recent work on forecasting in centered locally stationary time series can be found in Roueff and Sanchez-Perez (2018) and Kley et al. (2019). The first named authors investigated a predictor based on auto-regression of a given order, while Kley et al. (2019) considered predictors in stationary and locally stationary models for (possibly) non-stationary data and selected the "better" prediction among the two estimates. A common feature of most of these methods is that they are all based on auto-regressive fitting.
In the present paper we contribute to this literature and propose an alternative method for prediction in physically dependent locally stationary times series, which does not rely on auto-regressive fitting and is therefore more flexible. To be precise we consider the model

$$X_{i,n} = \mu(i/n) + \epsilon_{i,n}, \quad i = 1, \ldots, n \tag{1.1}$$

where μ is a deterministic and smooth mean or trend function on the interval $[0,1]$ and $\{\epsilon_{i,n} : i = 1, \ldots, n\}_{n \in \mathbb{N}}$ is a triangular array modelled by a locally stationary process in the sense of Zhou and Wu (2009) - see Section 2 for mathematical details. We then estimate the regression function μ by local linear smoothing and define a banded estimator for the corresponding auto-covariance matrix

$$\Sigma_n = \big\{\mathrm{Cov}(X_{i,n}, X_{j,n})\big\}_{1\leq i,j\leq n} \tag{1.2}$$

from the residuals of the nonparametric fit, where the width of the band increases with the sample size. Banded estimates of auto-covariance matrices have been considered by Wu and Pourahmadi (2009) and McMurry and Politis (2010) for **centered** and **stationary** processes using the fact that in this case the matrix Σ_n in (1.2) is a Toeplitz matrix. Neither of these results is applicable under the assumption of non-stationarity (even if the locally stationary process $\{X_{i,n}\}_{i=1,\dots,n}$ in (1.1) is centered).

In Section 3 we establish consistency (with respect to the operator norm) of the new covariance operator for locally stationary processes with a time varying mean function. These results are then used in Section 4 to develop new prediction methods, which - in contrast to the currently available literature - do not use autoregressive fitting. In Section 5 we investigate the finite sample properties of the estimator of the covariance matrix and compare the new predictor with the currently available methodology. Finally, all proofs of our main theoretical results and technical details can be found in Section 6.

2 Locally stationary processes

Consider the time series model (1.1) where $\{\epsilon_{i,n} : i = 1, \dots, n\}_{n\in\mathbb{N}}$ is an array of centered random variables, and $\mu : [0, 1] \to \mathbb{R}$ is a smooth mean function. More precisely we assume

(M1) The function μ in model (1.1) has a Lipschitz continuous second order derivative on the interval $[0, 1]$.

In order to model a local stationary error process we use a concept introduced by Zhou and Wu (2009). To be precise, define for an L^q-integrable random variable X its norm by $\|X\|_q = (\mathbb{E}[|X|^q])^{1/q}(q \geq 1)$, let $\{\varepsilon_i : i \in \mathbb{Z}\}$ denote a sequence of independent identically distributed observations and define $\mathcal{F}_i = (\dots, \varepsilon_{i-2}, \varepsilon_{i-1}, \varepsilon_i)$. We assume that there exists a function $G : [0, 1] \times \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ such that

$$\epsilon_{i,n} = G(i/n, \mathcal{F}_i) \tag{2.1}$$

is a well defined random variable. For arbitrary functions G it is not guaranteed that the stochastic structure of $\{\epsilon_{i,n} \colon i \in \mathbb{Z}\}$ varies smoothly, but we can achieve this by the following assumptions.

(L1) For some $q \geq 2$ we have that

$$\sup_{t \in [0,1]} \|G(t, \mathcal{F}_0)\|_q < \infty.$$

(L2) The function G is differentiable with respect to the first coordinate and there exists a constant $M > 0$ such that for all $t, s \in [0,1]$

$$\Big\| \frac{\partial}{\partial t} G(t, \mathcal{F}_0) - \frac{\partial}{\partial t} G(s, \mathcal{F}_0) \Big\|_2 \leq M|t-s|.$$

Next we quantify the dependence structure. For this purpose let $\{\varepsilon_i' : i \in \mathbb{Z}\}$ denote an independent copy of $\{\varepsilon_i : i \in \mathbb{Z}\}$, define $\mathcal{F}_i^* = (\ldots, \varepsilon_{-2}, \varepsilon_{-1}, \varepsilon_0', \varepsilon_1, \ldots, \varepsilon_i)$ and

$$\delta_q(G, i) = \sup_{t \in [0,1]} \|G(t, \mathcal{F}_i) - G(t, \mathcal{F}_i^*)\|_q$$

as a measure of dependence. We assume for the same $q \geq 2$ as in assumption (L1) that

(L3) There exists a constant $\chi \in (0,1)$ such that

$$\delta_q(G, i) = O(\chi^i).$$

Example 2.1. A prominent example of this non-stationary model is a locally stationary $AR(p)$ process where the filter in (2.1) is defined by

$$G(t, \mathcal{F}_i) = \sum_{s=1}^{p} a_s(t) G(t, \mathcal{F}_{i-s}) + \sigma(t)\varepsilon_i \tag{2.2}$$

where $(\varepsilon_i)_{i \in \mathbb{Z}}$ is a sequence of independent identically distributed centered random variables with $\|\varepsilon_1\|_q < \infty$, and $a_1, \ldots, a_p, \sigma : [0,1] \to \mathbb{R}$, are for smooth functions such that for some $\delta_0 > 1$ the polynomial $1 - \sum_{s=1}^{p} a_s(t) z^s$ has no roots in the disc $\{z \in \mathbb{C} \colon |z| \leq \delta_0\}$. If the functions a and σ have bounded derivatives, $G(t, \mathcal{F}_i)$ has a MA representation of the form $G(t, \mathcal{F}_i) = \sigma(t) \sum_{j=0}^{\infty} c_j(t) \epsilon_{i-j}$, where $c_1, c_2, \ldots$ are smooth functions with derivatives satisfying $|c_j'(t)| \leq M\chi^j$ for $j \geq 0$. Therefore assumptions (L1)-(L3) hold for model (2.2). It has been shown in Zhou (2013) that Model (2.2) can approximate the time-varying $AR(p)$ model in Dahlhaus (1997).

Remark 2.1. Note that the definition of a locally stationary error process contains the case that each row of $\{\epsilon_{i,n} \colon i \in \mathbb{Z}\}_{n \in \mathbb{N}}$ is stationary, that is $G(t, \mathcal{F}_i) = H(\mathcal{F}_i)$ for some function $H : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$. In this case the random variables $\epsilon_{i,n} = H(\mathcal{F}_i)$ do not depend on n, Assumption (L2) is obviously satisfied and Assumption (L1) and (L3) reduce to

(S1) For some $q \geq 2$, $\|H(\mathcal{F}_0)\|_q < \infty$.

(S2) There exists a constant $\chi \in (0,1)$ such that

$$\delta_q(H,i) = \|H(\mathcal{F}_i) - H(\mathcal{F}_i^*)\|_q = O(\chi^i) \ .$$

If assumption (L1) holds the covariance matrix $\Sigma_n = (\sigma_{i,j,n})_{1\leq i,j\leq n}$ in (1.2) is well defined, where

$$\sigma_{i,j,n} = \mathrm{Cov}(X_{i,n}, X_{j,n}) = \mathbb{E}(G(i/n, \mathcal{F}_i)G(j/n, \mathcal{F}_j)). \tag{2.3}$$

Throughout this paper we do not reflect the dependence on n in the notation of the entries of a matrix, whenever it is clear from the context. For example we will use $\sigma_{i,j}$ instead of $\sigma_{i,j,n}$ and similarly a simplified notation for corresponding estimates. We also define the (time dependent) auto-covariances

$$\gamma_k(t) = \mathbb{E}(G(t, \mathcal{F}_i)G(t, \mathcal{F}_{i+k})) \quad (k \in \mathbb{Z}) \tag{2.4}$$

of the stationary (for fixed $t \in [0,1]$) process $\{G(t, \mathcal{F}_i)\}_{i\in\mathbb{Z}}$. To estimate the covariances in (2.3) we use a local linear regression estimate of the function γ_k. In order to prove consistency of this estimator we require a smoothness condition on the auto-covariances in (2.4), which is formulated as follows.

(A1) For any $k \in \mathbb{Z}$ the function γ_k in (2.4) is differentiable with derivative $\dot{\gamma}_k(t) = \frac{\partial}{\partial t}\gamma_k(t)$. There exists constants D_k such that for all $t, s \in [0,1]$

$$|\dot{\gamma}_k(t) - \dot{\gamma}_k(s)| \leq D_k|t-s|.$$

An application of the Cauchy-Schwarz inequality and the dominated convergence theorem show that a sufficient condition for assumptions (L2) and (A1), is given by (L1) and

$$\sup_{t\in[0,1]} \Big\| \frac{\partial^2}{\partial t^2} G(t, \mathcal{F}_0) \Big\|_2 < \infty.$$

In the following section we will use the local linear estimates for the function γ_k to define a banded estimate of the covariance matrix Σ_n of a locally stationary process of the form

(1.1) and investigate its asymptotic properties for increasing sample size. We also discuss a corresponding estimator in the stationary case because usually estimators are studied under the assumption of a centered stationary process, that is $\mu \equiv 0$. In the subsequent Section 4 we use these results for prediction in locally stationary processes with a non-vanishing trend.

3 Covariance matrix estimation

The estimation of the covariance matrix has attracted considerable attention in the literature. We refer among many others to the work of Bickel and Levina (2008a), Bickel and Levina (2008b) for high-dimensional independent identically distributed data and Anderson (2003), Wu and Pourahmadi (2009), Chen et al. (2013), Box et al. (2015), and McMurry and Politis (2015) who considered this problem for time series. Most authors consider the case of a vanishing trend, i.e. $\mu \equiv 0$, and assume that the error process $\{\epsilon_{i,n} : i = 1, \ldots, n\}$ is a sequence of independent identical observations or a stationary series. For example, in the case of a stationary centered process Wu and Pourahmadi (2009) proposed the banded estimator

$$\tilde{\Sigma}_n = \{\tilde{\sigma}_{i,j}\mathbf{1}(|i-j| \leq l_n), 1 \leq i, j \leq n\} \tag{3.1}$$

of the matrix Σ_n, where $\mathbf{1}(A)$ denotes the indicator function of the set A and

$$\tilde{\sigma}_{i,j} = \frac{1}{n - |i-j|} \sum_{s=1}^{n-|i-j|} X_{s,n} X_{s+|i-j|,n},$$

is the sample auto-covariance of $\{X_{1,n}, \ldots, X_{n,n}\}$ at lag $|i-j|$ and $l_n \in \mathbb{N}$ denotes a tuning parameter satisfying $l_n \to \infty$, $l_n = o(n)$ as $n \to \infty$. McMurry and Politis (2010) modified this statistic such that the new estimator leaves the band intact, and then gradually down-weighs increasingly distant off-diagonal entries instead of setting them to zero as in the banded matrix case. Both estimators use the fact that for stationary processes the matrix Σ_n is a Toeplitz matrix.

Note that the estimator (3.1) is not consistent for the auto-covariance if the mean function is not constant. As there are many applications where time series have a smoothly changing mean function we begin our discussion analyzing a mean-corrected estimator of the matrix Σ_n for a stationary error process of the form (1.1), which avoids this problem.

Let $\hat{\mu}$ be the local linear estimator defined by

$$(\hat{\mu}(t), \hat{\dot{\mu}}(t))^\top = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{n} \big(X_{i,n} - \beta_0 - \beta_1(i/n - t)\big)^2 K\Big(\frac{i/n - t}{\tau_n}\Big) \tag{3.2}$$

where τ_n denotes the bandwidth. For the kernel K we make the following assumption:

(K) The kernel K is a symmetric, continuously differentiable, bounded density function supported on the interval $[-1, 1]$.

We consider the residuals

$$\hat{\epsilon}_{i,n} = X_{i,n} - \hat{\mu}(i/n) \tag{3.3}$$

obtained from the local linear fit and denote by

$$\hat{\sigma}^\dagger_{i,j} = \frac{1}{n - |i-j|} \sum_{s=1}^{n-|i-j|} \hat{\epsilon}_{s,n}\hat{\epsilon}_{s+|i-j|,n} \quad (i, j = 1, \dots, n)$$

the sample auto-covariance of the residuals $\{\hat{\epsilon}_{1,n}, \dots, \hat{\epsilon}_{n,n}\}$ at lag $|i - j|$. Finally, we define for $l_n \in \mathbb{N}$ the banded matrix

$$\hat{\Sigma}^\dagger_n = \{\hat{\sigma}^\dagger_{i,j}\mathbf{1}(|i - j| \le l_n)\}, \tag{3.4}$$

as an estimator of the matrix Σ_n. It will be shown below that the estimator $\hat{\Sigma}^\dagger_n$ is consistent for Σ_n in the case of a strictly stationary error process. To measure the distance between two matrices (of increasing dimension) we introduce the operator norm

$$\rho(A) = \max_{x\in\mathbb{R}^n:|x|=1} |Ax|$$

of a matrix A, where $|\cdot|$ denotes the Euclidean norm (note that $\rho^2(A)$ is the largest eigenvalue of the matrix $A^\top A$).

Theorem 3.1. *Assume that $n\tau_n^6 = o(1)$, $n\tau_n^3 \to \infty$, $l_n \to \infty$, $\frac{l_n^2}{n} = o(1)$. If conditions (K), (S1), (S2) and (M1) hold, then*

$$\|\rho(\hat{\Sigma}^\dagger_n - \Sigma_n)\|_{q/2} = O(r_n^\diamond),$$

where the sequence $r_n^\diamond$ is defined by

$$r_n^\diamond = l_n(\tau_n^2 + (n\tau_n)^{-1/2}) + \frac{l_n^2}{n} + \chi^{l_n}.$$

Theorem 3.1 establishes consistency of the estimator of the covariance matrix in model (1.1) in the operator norm under the assumption of a stationary error process. However, there also exist many time series exhibiting a non-stationary behaviour in the higher order moments and dependence structure [see Stărică and Granger (2005), Elsner et al. (2008), Guillaumin et al. (2017) among others], and estimation under the assumption of a location model with a stationary error process might be misleading. In this case the estimator $\hat{\Sigma}_n^\dagger$ in (3.4) is not necessarily consistent since the unknown covariance matrix Σ_n is not a Toeplitz matrix. To address this problem we propose an alternative approach which also yields a consistent estimator for non-stationary time series. Roughly speaking, we estimate the elements $\sigma_{i,j}$ in the matrix Σ_n by

$$\hat{\sigma}_{i,j} = \hat{\gamma}_{|i-j|}\Big(\frac{i+j}{2n}\Big), \tag{3.5}$$

where $\hat{\gamma}_k(t)$ is a local linear estimate of the auto-covariance function (2.4) of the process $\{G(t,\mathcal{F}_i)\}_{i\in\mathbb{Z}}$.
To be precise, we distinguish between a lag of odd or even order and define

$$(\hat{\gamma}_k(t), \hat{\gamma}_k'(t))^\top = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^n \big(\hat{\epsilon}_{i-k/2,n}\hat{\epsilon}_{i+k/2,n} - \beta_0 - \beta_1(i/n-t)\big)^2 K\Big(\frac{i/n-t}{b_n}\Big) \tag{3.6}$$

if the lag k is of even order, where b_n is a bandwidth and the residuals $\hat{\epsilon}_{i,n}$ are defined in (3.3). In (3.6) we use the notation $\hat{\epsilon}_{i,n} = 0$ if the index i satisfies $i < 0$ or $i > n$. Similarly, for an odd lag k we define

$$\hat{\gamma}_k(t) = \frac{1}{2}\big(\hat{\gamma}_k^+(t) + \hat{\gamma}_k^-(t)\big), \tag{3.7}$$

where

$$(\hat{\gamma}_k^+(t), (\hat{\gamma}_k^+)'(t))^\top = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^n \big(\hat{\epsilon}_{i-(k-1)/2,n}\hat{\epsilon}_{i+(k+1)/2,n} - \beta_0 - \beta_1(i/n-t)\big)^2 K\Big(\frac{i/n-t}{b_n}\Big),$$

$$(\hat{\gamma}_k^-(t), (\hat{\gamma}_k^-)'(t))^\top = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^n \big(\hat{\epsilon}_{i-(k+1)/2,n}\hat{\epsilon}_{i+(k-1)/2,n} - \beta_0 - \beta_1(i/n-t)\big)^2 K\Big(\frac{i/n-t}{b_n}\Big).$$

The estimator of the element $\sigma_{i,j}$ in Σ_n is finally defined by (3.5) and for the covariance matrix we use again a banded estimator, that is

$$\hat{\Sigma}_n := \Big(\hat{\gamma}_{|i-j|}(\frac{i+j}{2n})\mathbf{1}(|i-j| \leq l_n)\Big)_{1\leq i,j\leq n}. \tag{3.8}$$

Our next result yields the consistency of this estimator in the operator norm.

Theorem 3.2. *Assume that $n\tau_n^3 \to \infty$, $n\tau_n^6 = o(1)$, $\frac{l_n^2}{n} = o(1)$, $l_n b_n^2 = o(1)$,*

$$l_n((nb_n)^{-1/2}b_n^{-2/q} + \tau_n^2 + (n\tau_n)^{-1/2}) = o(1) \quad \textit{and} \quad b_n^2\sum_{k=0}^{l_n} D_k = o(1)$$

If the conditions (K), (L1)–(L3), (A1) and (M1) are satisfied, then we have

$$\|\rho(\hat{\Sigma}_n - \Sigma_n)\|_{q/2} = O(r_n),$$

where the sequence r_n is defined by

$$r_n = l_n((nb_n)^{-1/2}b_n^{-2/q} + \tau_n^2 + (n\tau_n)^{-1/2}) + \frac{l_n^2}{n} + \chi^{l_n} + b_n^2\sum_{k=0}^{l_n} D_k = o(1). \tag{3.9}$$

Remark 3.1.

(a) In the case of a stationary and centered time series it has been demonstrated by McMurry and Politis (2015) that tapering can improve the performance of simply banded estimators of the covariance matrix and similar arguments apply to the covariance estimators (3.4) and (3.8) proposed in this paper for stationary times series with a time varying mean function and for locally stationary times series. To be precise consider the situation in Theorem 3.2 and define the tapering function (other tapers could be used as well) by

$$\kappa(x) = (2 - |x|)\mathbf{1}(1 \leq |x| \leq 2) + \mathbf{1}(|x| < 1)$$

The tapered and banded estimate of the covariance matrix Σ_n is now defined by

$$\hat{\Sigma}_n^{tap} := \Big(\kappa\Big(\frac{|i-j|}{l_n}\Big)\tilde{\gamma}_{|i-j|}(\frac{i+j}{2n})\Big)_{1\leq i,j\leq n}.$$

Using the same arguments as in the proof of Theorem 3.2 it can be shown that

$$\|\rho(\hat{\Sigma}_n^{tap} - \Sigma_n)\|_{q/2} = O(r_n),$$

where the sequence r_n is defined in (3.9).

(b) It is worthwhile to mention that recently Ding and Zhou (2018) proposed an alternative estimate of the the precision matrix Σ_n^{-1} of a **centered** locally stationary series, which is based on a Cholesky decomposition. In contrast the estimator $\hat{\Sigma}_n^{-1}$ considers the inverse of a banded estimator of the covariance matrix of a locally stationary series with a smoothly varying trend.

4 Prediction

In this section we discuss some applications of the proposed estimators in the problem to perform predictions in locally stationary processes. For centered time series this problem has been recently investigated by Roueff and Sanchez-Perez (2018), Kley et al. (2019) who proposed to fit a locally stationary AR model and perform the prediction using an AR approximation. In this section, we suggest an alternative method which is not based on AR fitting. To be precise, assume that we observe a stretch of data $X_{1,n}, \dots, X_{m,n}$ from the model (1.1) and that we are interested in a prediction of the next observation $X_{m+1,n}$. To be precise, our aim is the construction of best linear predictor of $X_{m+1,n}$ based on $X_{1,n}, \dots, X_{m,n}$. For this purpose we define

$$X_{m+1,n}^{\text{Pred}} := a_{m+1,n} + \sum_{s=1}^{m} a_{m+1-s,n} X_{s,n} = \mathbf{a}_m^\top \mathbf{X}_{m,n}, \tag{4.1}$$

where $\mathbf{X}_{m,n} = (1, X_{1,n}, ..., X_{m,n})^\top$ and the prediction vector $\mathbf{a}_m = (a_{m+1,n}, a_{m,n}..., a_{1,n})^\top := (a_{m+1,n}, (\mathbf{a_m}^*)^\top)^\top$ is given by

$$\mathbf{a}_m = (a_{m+1,n}, (\mathbf{a_m}^*)^\top)^\top = \underset{\theta \in \mathbb{R}^{m+1}}{\operatorname{argmin}} \mathbb{E}(X_{m+1,n} - \theta^\top \mathbf{X}_{m,n})^2. \tag{4.2}$$

In order to estimate the vector $\mathbf{a}_m$ we define the local linear estimators from the sample $X_{1,n}, \dots, X_{m,n}$ by

$$(\hat{\mu}^{1:m}(t), \hat{\acute{\mu}}^{1:m}(t))^\top = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{m} (X_{i,n} - \beta_0 - \beta_1(i/n - t))^2 K\left(\frac{i/n - t}{\tau_n}\right), \tag{4.3}$$

and denote by

$$\Sigma_{n,m} = (\sigma_{i,j,n})_{1 \leq i,j \leq m} = \big(\mathrm{Cov}(X_{i,n}, X_{j,n})\big)_{1 \leq i,j \leq m} \tag{4.4}$$

the covariance matrix of the vector $(X_{1,n}, \ldots, X_{m,n})^T$. The residuals (3.3) for estimating the auto-covariances are then replaced by residuals by

$$\hat{\epsilon}_{i,n}^{1:m} = X_{i,n} - \hat{\mu}^{(1:m)}(i/n) \quad (i = 1, \ldots, m)$$

from the nonparametric fit from the data $X_{1,n}, \ldots, X_{m,n}$. Next, we define $\hat{\gamma}_k^{1:m}$ as the analogue of the estimator (3.6) (if the lag k is even) and (3.7) (if the lag is odd), where the residual $\hat{\epsilon}_{\ell,n}$ is replaced by $\hat{\epsilon}_{\ell,n}^{1:m}$. We further define

$$\hat{\Sigma}_{n,m} := \Big(\hat{\gamma}_{|u-v|}^{1:m}\big(\frac{u+v}{2n}\big)\mathbf{1}(|u-v| \leq l_n)\Big)_{1 \leq u,v \leq m} \tag{4.5}$$

as a banded estimator of the covariance matrix $\Sigma_{n,m} := \mathrm{Cov}(X_{i,n}, X_{j,n})_{1 \leq j \leq m}$ in (4.4). It can be shown that, if the assumptions of Theorem 3.2 are satisfied and $m \geq \lfloor cn \rfloor$ for some positive constant c,

$$\|\hat{\Sigma}_{n,m} - \Sigma_{n,m}\|_{q/2} = O(r_n), \tag{4.6}$$

where the sequence r_n is defined in (3.9). We shall construct a predictor based on $\hat{\Sigma}_{n,m}^{-1}$ and for this purpose we show that the consistency of the estimator $\hat{\Sigma}_{n,m}$ in (4.6) can be transferred to its inverse.

Throughout this paper we denote $\lambda_{min}(A)$ the minimum eigenvalue of a symmetric matrix A and make the following assumption.

(E1) There exists a constant $c > 0$ such that

$$\eta = \liminf_{n \to \infty} \inf_{\lfloor cn \rfloor \leq m \leq n} \lambda_{min}(\Sigma_{n,m}) > 0.$$

Corollary 4.1. *Assume that the conditions of Theorem 3.2 and condition (E1) are satisfied. If $n \to \infty$, $\lfloor cn \rfloor \leq m \leq n$ we have*

$$\rho(\hat{\Sigma}_{n,m}^{-1} - \Sigma_{n,m}^{-1}) = O_{\mathbb{P}}(r_n) \tag{4.7}$$

We can now define an estimate $\hat{\mathbf{a}}_m = (\hat{a}_{m+1,n}, (\hat{\mathbf{a}}_m^*)^\top)^\top$ of the vector $\mathbf{a}_m$ in (4.2) by

$$\hat{a}_{m+1,n} = \hat{\mu}^{1:m}(m/n) - \sum_{s=1}^{m} \hat{a}_{m+1-s,n}\hat{\mu}^{1:m}(s/n),$$

and

$$\hat{\mathbf{a}}_m^* \;=\; (\hat{a}_{m,n}, ..., \hat{a}_{1,n})^\top = \hat{\Sigma}_{n,m}^{-1}\hat{\boldsymbol{\gamma}}_n^{1:m}, \tag{4.8}$$

where

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_n^{1:m} &= (\hat{\gamma}_{n,m}^{1:m}, \hat{\gamma}_{n,m-1}^{1:m}, ..., \hat{\gamma}_{n,1}^{1:m})^\top,\\ \hat{\gamma}_{n,s}^{1:m} &= \hat{\gamma}_s^{1:m}\Big(\frac{2m-s+1}{2n}\Big)\mathbf{1}(1 \le s \le l_n).\end{aligned}$$

The final predictor of $X_{m+1,n}$ is defined by

$$\hat{X}_{m+1,n}^{\mathrm{Pred}} := \hat{a}_{m+1,n} + \sum_{s=1}^{m} \hat{a}_{m+1-s,n}X_{s,n}, \tag{4.9}$$

Theorem 4.1. *Assume that the conditions of Theorem 3.2 and assumption (E1) are satisfied,* $\liminf_{n\to 0} \frac{l_n}{\log n} \ge \eta > 0$ *and assume that there exists a constant* $c \in (0,1)$ *such that for* $m \ge cn$, $m \ge \lfloor nb_n \rfloor$.
(a) The vector $\hat{\mathbf{a}}_m = (\hat{a}_{m+1,n}, (\hat{\mathbf{a}}_m^*)^\top)^\top$ *is a consistent estimator of the coefficient vector* $\mathbf{a}_m$ *of the best linear predictor defined in* (4.2), *i.e.,*

$$|\hat{\mathbf{a}}_m^* - \mathbf{a}_m^*| = O_{\mathbb{P}}(r_n), \quad \hat{a}_{m+1,n} - a_{m+1,n} = O_{\mathbb{P}}(r_n^\circ)$$

where r_n *is defined in* (3.9), *and*

$$r_n^\circ = (l_n^{1/2}\log^{1/2} n)r_n + \sqrt{n}\chi^{l_n}. \tag{4.10}$$

(b) Assume that $r_n^\circ = o(1)$. *If the error* $\epsilon_{i,n}$ *is a locally stationary AR(p) process as defined in Example 2.1 and*

(P1) $n^{\frac{1}{q}}r_n = o(1)$.

(P2) $\delta_q(\dot{G}, i) = O(\chi^i)$,

(P3) $\sup_{t\in[0,1]} \|\dot{G}(t, \mathcal{F}_i)\|_q < \infty$,

where $\dot{G}(t,\mathcal{F}_i) = \frac{\partial}{\partial t} G(t,\mathcal{F}_i)$ *denotes the derivative of the filter* G*, we have*

$$\frac{X_{m+1,n} - \hat{X}^{\text{Pred}}_{m+1,n}}{\sigma(\frac{m+1}{n})} \Rightarrow \varepsilon_1 \tag{4.11}$$

where $\Rightarrow$ *denotes the convergence in distribution and* ε_1 *denotes the error in model* (2.2) *.*

The rate r_n° in (4.10) results from convergence rate of the nonparmetric estimate of the time-varying mean and does not appear if the trend is not estimated because it is known to be 0. Conditions (P2) and (P3) can be verified by checking the coefficients of the MA representation of the locally stationary AR process (2.2). They assure that for any i, j, the process $\{\mathbb{E}(G(t,\mathcal{F}_i)G(s,\mathcal{F}_j))\}_{t,s\in[0,1]}$ is sufficiently smooth on $[0,1]\times[0,1]$.

Remark 4.1. Similar arguments as given in the proof of Theorem 3.2 show that the estimator $\hat{\Sigma}_{n,m}$ is positive definite if the sample size is sufficiently large. However, for finite sample sizes the matrix $\hat{\Sigma}_{n,m}$ can be singular. As the prediction in (4.9) requires a non-singular sample covariance matrix we propose in applications to replace the estimator $\hat{\Sigma}_{n,m}$ by a a positive definite estimator, say $\hat{\Sigma}^{pd}_{n,m}$, which is defined as follows. If $\hat{\Sigma}_{n,m} = U_{n,m}V_{n,m}U^\top_{n,m}$ is the spectral decomposition of $\hat{\Sigma}_{n,m}$ and $V_{n,m} = \text{diag}(v_1,\dots,v_m)$ is the diagonal matrix containing the corresponding eigenvalues, we define

$$\hat{\Sigma}^{pd}_{n,m} := U_{n,m}V^+_{n,m}U^\top_{n,m} \tag{4.12}$$

where $V^+_{n,m}$ is a diagonal matrix with its ith diagonal element given by

$$v_i^+ = \max\left\{v_i, \frac{10\int_0^{\frac{m}{n}} \hat{\gamma}_0^{1:m}(t)dt}{m^\beta}\right\}, \ i = 1,\dots .m$$

for some $\beta > 0$. As a rule of thumb, we choose $\beta = 0.5$ because for this choice $\rho(\hat{\Sigma}^{pd}_{n,m} - \hat{\Sigma}_{n,m}) = O(n^{-\beta}) = O(r_n)$. This type of modification has been also advocated by McMurry and Politis (2010) and McMurry and Politis (2015) for stationary time series. Using similar argument as in the proof of Theorem 3.2 of this paper and in the proof of Theorem 3 of McMurry and Politis (2010), it can be shown that $\|\hat{\Sigma}^{pd}_{n,m} - \Sigma_{n,m}\|_{q/2} = O(r_n)$. Now the arguments given in the proof of Corollary 1 of Wu and Pourahmadi (2009) yield an analogue of Corollary 4.1, that is

$$\rho((\hat{\Sigma}^{pd}_{n,m})^{-1} - \Sigma^{-1}_{n,m}) = O_{\mathbb{P}}(r_n).$$

A careful inspection of the proof of Theorem 4.1 finally shows that its assertion remains valid, if $\hat{\Sigma}_{n,m}$ in (4.7) is replaced by $\hat{\Sigma}^{pd}_{n,m}$.

5 Implementation and numerical results

To implement our method we need to choose several tuning parameters: the bandwidths τ_n and b_n for the local linear estimators of the trend μ and auto-covariance function γ_k and the width l_n of the banded estimator of the covariance matrix Σ_n. For choosing τ_n, we recommend the Generalized Cross Validation (GCV) method proposed in Zhou and Wu (2010).
More precisely, let $\hat{\mu}^{1:m}(\cdot,\tau), 1 \leq i \leq m$ be the local linear estimate of the mean trend defined in (4.3) using bandwidth τ, then we choose τ_n as

$$\tau_n = \underset{\tau}{\operatorname{argmin}} \frac{n^{-1}\sum_{i=1}^{m}(X_{i,n} - \hat{\mu}^{1:m}(i/n,\tau))^2}{(1-\sum_{i=1}^{m}(T_{\tau,ii}^{1:m})/n)^2},$$

where $T_{\tau,ii}^{1:m}$ is the i_{th} diagonal entry of the matrix

$$J_0^{1:m}\big((X^{1:m}(i/n))^\top W_\tau^{1:m}(i/n)X^{1:m}(i/n)\big)^{-1}(X^{1:m}(i/n))^\top W_\tau^{1:m}(i/n),$$

$J_0^{1:m}$ and $X^{1:m}(i/n)$ are $m \times 2$ matrices defined by

$$J_0^{1:m} \quad = \quad \begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 \end{pmatrix}^\top \quad , \quad X^{1:m}(i/n) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \frac{1-i}{n} & \frac{2-i}{n} & \dots & \frac{m-i}{n} \end{pmatrix}^\top,$$

respectively, and $W_\tau(x)$ is an $m \times m$ diagonal matrix with elements $\big\{K\big(\frac{x-s/n}{\tau}\big)\big\}_{s=1,\dots m}$.
The bandwidth b_n for the estimation of the auto-covariance function γ_k in (2.4) is defined similarly. For example, if k is even, we choose b_n as

$$b_n = \underset{c}{\operatorname{argmin}} \frac{n^{-1}\sum_{i=1}^{m}(\hat{\epsilon}_{i-k/2,n}^{1:m}\hat{\epsilon}_{i+k/2,n}^{1:m} - \hat{\gamma}_k^{1:m}(i/n,c))^2}{(1-\sum_{i=1}^{m}(T_{c,ii}^{1:m})/n)^2}, \tag{5.1}$$

where $\hat{\gamma}_k^{1:m}(i/n,c)$ is the local linear estimator with bandwidth c defined as in (3.6) using m observations and $T_{c,ii}^{1:m}$ is defined as in the previous paragraph.
To motivate the choice of the width l_n in the banded estimator of the covariance matrix, note that

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{m\wedge(n-k)}\epsilon_{i,n}\epsilon_{i+k,n} - \int_0^{\frac{m}{n}\wedge 1}\gamma_k(t)dt\Big) \Rightarrow \mathcal{N}(0,\tilde{\sigma}_k^2), \tag{5.2}$$

[see Section 4.3 in Zhang and Wu (2012)], where $\tilde{\sigma}_k^2 = \int_0^{\frac{m}{n}\wedge\frac{n-k}{n}} g^2(t)dt$, and the function g^2

is the long-run variance of the locally stationary process $\{\epsilon_{i,n}\epsilon_{i+k,n}\}_{i=1}^{n-k}$. For its estimation we use a statistic proposed by Dette and Wu (2019), which is defined as follows. Consider the partial sum of lag k

$$^{k}S_{r_0,r_1}^{1:m} = \sum_{i=r_0}^{r_1} \hat{\epsilon}_{i,n}^{1:m}\hat{\epsilon}_{i+k,n}^{1:m},$$

where we use the notation $\hat{\epsilon}_{i,n}^{1:m} = 0$ if the index i satisfies $i < 1$ or $i > m$. For an integer $b \geq 2$ we introduce the quantities

$$^{k}\Delta_{j,b}^{1:m} = \frac{^{k}S_{j-b+1,j}^{1:m} - {}^{k}S_{j+1,j+b}^{1:m}}{b}.$$

Finally, we define for $t \in [b/n, (m-b)/n]$

$$\hat{g}^2(t) = \sum_{j=1}^{n} \frac{b(^{k}\Delta_{j,b}^{1:m})^2}{2}\omega(t,j),$$

where

$$\omega(t,i) = K\Big(\frac{i/n-t}{b_n}\Big)\Big/\sum_{i=1}^{n} K\Big(\frac{i/n-t}{b_n}\Big)$$

and the bandwidth b_n is given by (5.1) with $\hat{\epsilon}_{i-k/2,n}^{1:m}\hat{\epsilon}_{i+k/2,n}^{1:m}$ there replaced by $\hat{\epsilon}_{i,n}^{1:m}\hat{\epsilon}_{i+k,n}^{1:m}$. For $t \in [0, b/n)$ and $t \in ((m-b)/n, m/n]$ we define $\hat{g}^2(t) = \hat{g}^2(b/n)$ and $\hat{g}^2(t) = \hat{g}^2((m-b)/n)$, respectively. Finally, we propose

$$l_n = \max\Big\{l \in [l_0, l_1] \;\Big|\; n^{-1/2}|\sum_{i=1}^{n} \hat{\epsilon}_{i,n}^{1:m}\hat{\epsilon}_{i+l,n}^{1:m}| \geq \kappa(0.01)\hat{\sigma}_l\Big\}, \tag{5.3}$$

as a data-driven choice of the width l_n, where $\kappa(\alpha)$ is the $\frac{1+(1-\alpha)^{1/(l_1-l_0+1)}}{2}$-quantile of the standard normal distribution and l_0 and l_1 are constants (if the set $\big\{n^{-1/2}|\sum_{i=1}^{n} \hat{\epsilon}_{i,n}^{1:m}\hat{\epsilon}_{i+l,n}^{1:m}| \geq \kappa(0.01)\hat{\sigma}_l, 1_0 \leq l \leq l_1\big\}$ is empty we define $l_n = l_0 - 1$).

5.1 Covariance estimation

In this section we investigate the finite sample properties of the estimators (3.4) and (3.8) for the covariance matrix Σ_n of a locally stationary process, where we consider

$$\begin{aligned} \mu(t) &= 2\sin 2\pi(t), && (5.4) \\ \mu(t) &= 2 - 8(t-0.5)^2, && (5.5) \\ \mu &= 0\ , && (5.6) \end{aligned}$$

as mean functions. Recalling the notation $\mathcal{F}_i = (\ldots, \varepsilon_{i-1}, \varepsilon_i)$ we investigate four different distributions for the errors in model (1.1):

(a) $\{\epsilon_{i,n} : i = 1, \ldots, n\}$ is a stationary $AR(0.3)$ process with independent standard normal distributed innovations.

(b) $\epsilon_{i,n} = 0.8G(i/n, \mathcal{F}_i)$ where

$$G(t, \mathcal{F}_i) = 0.7\sin(2\pi t)G(t, \mathcal{F}_i) + \varepsilon_i$$

and $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ is a sequence of independent, standardized ($\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = 1$) t-distributed random variables with six degrees of freedom.

(c) $\epsilon_{i,n} = G(i/n, \mathcal{F}_i)$ where

$$G(t, \mathcal{F}_i) = \frac{1}{6}(\exp(4(t-0.5)^2) + 1)\varepsilon_i + 0.6(|\varepsilon_{i-1}| - \mathbb{E}(|\varepsilon_{i-1}|))$$

and $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ is a sequence of independent standard normal distributed random variables.

(d) $\epsilon_{i,n} = G(i/n, \mathcal{F}_i)$ where

$$G(t, \mathcal{F}_i) = \frac{1}{4}(\cos(\pi t) + 2)(\varepsilon_i + 0.9\varepsilon_{i-1} - 0.6\varepsilon_{i-2})$$

and $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ is a sequence of standardized ($\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = 1$) independent chi-square distributed random variables with five degrees of freedom.

Note that model (a) defines a stationary process and model (b) defines a locally stationary AR(1) process. Model (c) defines a nonlinear $tvMA(1)$ process. Since the innovations ε_i in model (c) have a symmetric distribution, the covariance matrix of model (c) is diagonal.

Model (d) defines a $tvMA(2)$ process, where only the entries in the diagonal and the first two off diagonals of the covariance matrix do not vanish.

Table 1: *Simulated mean squared error* $\rho(\hat{\Sigma}_n - \Sigma_n)$ *for the estimators* (3.8) *and* (3.4) *in model* (1.1) *with different mean functions and error processes (a) and (b).*

		Model (a)		Model (b)	
n	μ	(3.8)	(3.4)	(3.8)	(3.4)
250	(5.4)	0.952 (0.0104)	0.637 (0.0105)	5.034 (0.0311)	5.532 (0.0083)
	(5.5)	0.943 (0.0100)	0.632 (0.0102)	5.063 (0.0308)	5.529 (0.0083)
	(5.6)	0.770 (0.098)	0.474 (0.0090)	4.646 (0.0365)	5.388 (0.0103)
500	(5.4)	0.683 (0.0080)	0.410 (0.0051)	4.304 (0.0303)	5.610 (0.0076)
	(5.5)	0.672 (0.0078)	0.421 (0.0053)	4.370 (0.0291)	5.595 (0.0081)
	(5.6)	0.609 (0.0073)	0.346(0.0045)	4.021 (0.0299)	5.490(0.0096)
1000	(5.4)	0.518 (0.0060)	0.329 (0.0043)	3.868 (0.0264)	5.624 (0.0069)
	(5.5)	0.535 (0.0062)	0.322 (0.0043)	3.881 (0.0265)	5.632 (0.0070)
	(5.6)	0.484 (0.0060)	0.282 (0.0042)	3.760 (0.0274)	5.563 (0.0077)

We examine the estimator for covariance matrix Σ_n for sample sizes $n = 250$, 500 and 1000 using 1000 simulation runs. For the estimation of the width l_n of the band in (4.5) we use (5.3) with $l_0 = 1$, $l_1 = 6$. In each simulation run the tuning parameters (τ_n, b_n) are determined as described at the beginning of this section. In Table 1 and 2 we display the simulated mean squared error of the spectral loss $\rho(\hat{\Sigma}_n - \Sigma_n)$ for different estimators $\hat{\Sigma}_n$, where different mean functions and error processes in model (1.1) are considered. In particular we compare the mean corrected estimator (3.8) for non-stationary error processes with the mean corrected estimator (3.4) which assumes a stationary error process. The numbers in brackets show the standard error of the estimates. We observe that in the stationary model (a) the accuracy of both estimators improve with increasing sample size. Moreover, the estimator (3.4) outperforms (3.8) because this estimator is constructed for stationary processes. On the other hand, for the dependence structures (b) - (d) corresponding to locally stationary processes the stationary method in (3.4) is not consistent and the estimator (3.8) shows a substantially superior behaviour.

5.2 Prediction

To illustrate the finite sample properties of the estimator proposed in Section 4 for prediction we examine the mean trend (5.4). As error process we consider a locally stationary AR(6) model defined by

Table 2: *Simulated mean squared error* $\rho(\hat{\Sigma}_n - \Sigma_n)$ *for the estimators* (3.8) *and* (3.4) *in model* (1.1) *with different mean functions and error processes (c) and (d)*

		Model (c)		Model (d)	
n	μ	(3.8)	(3.4)	(3.8)	(3.4)
250	(5.4)	0.647 (0.0114)	1.059 (0.0022)	0.767 (0.0113)	1.024 (0.0071)
	(5.5)	0.623 (0.0116)	1.062 (0.0023)	0.773 (0.011)	1.037 (0.0071)
	(5.6)	0.557 (0.0109)	1.045 (0.0023)	0.745 (0.0109)	1.062 (0.0073)
500	(5.4)	0.482 (0.0094)	1.045 (0.0017)	0.558 (0.010)	0.963 (0.0045)
	(5.5)	0.478 (0.0094)	1.043 (0.0016)	0.569 (0.010)	0.960 (0.0044)
	(5.6)	0.450 (0.0090)	1.037 (0.0016)	0.564 (0.0098)	0.963 (0.0044)
1000	(5.4)	0.357 (0.0069)	1.037 (0.0012)	0.426 (0.0082)	0.964 (0.0030)
	(5.5)	0.374 (0.0071)	1.040 (0.0012)	0.418 (0.0078)	0.959 (0.0031)
	(5.6)	0.360 (0.0074)	1.036(0.0012)	0.405 (0.0079)	0.960 (0.0030)

$$\prod_{s=1}^{6}(1 - a_s(t)\mathcal{B})G(t, \mathcal{F}_i) = \sigma(t)\varepsilon_i, \tag{5.7}$$

where the functions $a_1(t), \ldots, a_6(t)$ are given by

$$a_1(t) = 0.6\sin(2\pi(t - 0.05)),\ a_2(t) = 0.3\cos^2(3\pi t),\ a_3(t) = ((\exp(t - 0.6))^2)/3 - 0.4,$$
$$a_4(t) = -0.4\sin(6\pi t) - 0.1,\ a_5(t) = (t - 0.3)^2 - 0.2,\ a_6(t) = 0.2,$$

$\sigma(t) = (1 + 0.5\sin 2\pi t)^{0.5}$ and $\mathcal{B}$ is the lag operator on the filter $\mathcal{F}_i$, i.e., $\mathcal{B}G(t, \mathcal{F}_i) = G(t, \mathcal{F}_{i-1})$. We consider a standard normal as well as a $\chi^2(6)$ distribution for the errors ε_i (centered and standardized such that $\mathbb{E}[\varepsilon_i] = 0$ $\mathrm{Var}(\varepsilon_i) = 1$) and examine the mean squared error of the prediction for sample sizes $n = 250, n = 500, n = 1000$. We also compare the new predictor with the methods in Roueff and Sanchez-Perez (2018), Kley et al. (2019) and Giraud et al. (2015) which were theoretically investigated for centered data. In a first step we used these methods with the residuals $\hat{\epsilon}_{i,n}^{1:m}$ to obtain a prediction for the de-trended series. In a second step we add to this estimate the value $\hat{\mu}^{1:m}(m/n)$ to obtain the final prediction of $X_{m+1,n}$. Notice that these authors use time-varying AR(d) processes to approximate the time series for prediction without knowing d. Since the error process (5.7) is a locally AR(6) process, we investigate the performance of the methods proposed by Roueff and Sanchez-Perez (2018), Kley et al. (2019) and Giraud et al. (2015) for $d = 3$, $d = 6$ and $d = 9$ (note that in the predictor of Kley et al. (2019) d denotes the maximum lag that their algorithm allows). These cases represent the situation of underestimation, correct-estimation and overestimation of d. Note that in the cited references there are no rules how to select d. Moreover, for the method proposed by Kley et al. (2019) we choose the parameter δ in their procedure as 0.05, as a small parameter δ prefers the choices of a time-varying model to a stationary model.

Table 3: *Simulated mean squared error of different predictors in model* (5.7) *with standard normal distributed* ε_i. *The numbers in brackets show the standard error and the index* $*$ *represents the predictor with the best best performance.*

Method		$t_{pred} = 0.5$			$t_{pred} = 1$		
	lag	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
(4.9)		1.250	1.070*	1.033*	1.283*	1.170	1.077*
	-	(0.0570)	(0.0530)	(0.0464)	(0.0596)	(0.0511)	(0.0464)
R-S	$d = 3$	1.286	1.126	1.057	1.342	1.148*	1.137
		(0.0523)	(0.0499)	(0.0466)	(0.0577)	(0.0589)	(0.0490)
	$d = 6$	1.427	1.250	1.263	1.494	1.288	1.161
		(0.0700)	(0.0510)	(0.0532)	(0.0905)	(0.0536)	(0.0518)
	$d = 9$	1.895	1.297	1.209	32.286	1.779	1.125
		(0.1667)	(0.0566)	(0.0514)	(20.2729)	(0.0630)	(0.0542)
G-R-S	$d = 3$	1.241*	1.244	1.319	2.729	3.262	3.524
		(0.0623)	(0.0607)	(0.0633)	(0.1201)	(0.1676)	(0.2425)
	$d = 6$	1.251	1.241	1.122	2.385	2.868	2.933
		(0.0572)	(0.0537)	(0.0552)	(0.1065)	(0.1280)	(0.1378)
	$d = 9$	1.323	1.166	1.170	2.536	2.461	2.441
		(0.0625)	(0.0548)	0.0500)	(0.1105)	(0.1169)	(0.1311)
K-P-F	$d = 3$	1.314	1.182	1.126	1.346	1.329	1.168
		(0.0628)	(0.0538)	(0.0484)	(0.0674)	(0.0652)	(0.0517)
	$d = 6$	1.336	1.155	1.133	1.448	1.340	1.270
		(0.0565)	(0.0586)	(0.0474)	(0.0726)	(0.0612)	(0.0503)
	$d = 9$	1.343	1.357	1.215	1.459	1.279	1.255
		(0.0598)	(0.0480)	(0.0509)	(0.0588)	(0.0659)	(0.0581)

In Table 3 and 4 we present the simulated mean squared error

$$\mathbb{E}[(\hat{X}^{pred}_{m+1,n} - X_{m+1,n})^2]$$

for the four different prediction methods and different distributions of the innovations. The columns denoted by $t_{pred} = 0.5$ and $t_{pred} = 1$ correspond to a prediction of $X_{\lfloor n/2 \rfloor + 1}$ from on $X_{1,1}, \ldots, X_{\lfloor n/2 \rfloor}$ and a prediction of $X_{n,n}$ from $X_{1,1}, \ldots, X_{n-1,n}$, respectively, where we use $l_0 = \lceil \log(m) \rceil$ and $l_1 = 5 + \lceil \log(m) \rceil$ in (5.3). The first row shows the simulated mean squared error of the prediction (4.9). With increasing sample size this mean squared error approximates 1. This corresponds to our theoretical result in Theorem 4.1, because we have for the model under consideration $\sigma(0.5) = \sigma(1) = 1$. The rows denoted by R-S, G-R-S and K-P-F show the simulated mean squared error for predictors proposed by

Table 4: *Simulated mean squared error of different predictors in model* (5.7) *with (standardized) chi-squared ε_i. The numbers in brackets show the standard error and the index $*$ represents the predictor with the best best performance.*

		$t_{pred} = 0.5$			$t_{pred} = 1$		
Method	lag	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
(4.9)	-	1.201	1.123	1.072*	1.294*	1.116*	1.088
		(0.0577)	(0.0722)	(0.0624)	(0.0871)	(0.0554)	(0.0608)
R-S	$d = 3$	1.276	1.032*	1.100	1.307	1.196	1.061*
		(0.0645)	(0.0757)	(0.0632)	(0.0718)	(0.0794)	(0.0696)
	$d = 6$	4.282	1.263	1.107	1.775	1.298	1.160
		(0.0645)	(0.0787)	(0.0720)	(0.0833)	(0.0868)	(0.0627)
	$d = 9$	1.726	1.347	1.159	50.111	4.181	1.210
		(0.1022)	(0.0573)	(0.0567)	(24.4556)	(0.0804)	(0.0861)
G-R-S	$d = 3$	1.366	1.376	1.346	2.646	3.185	3.162
		(0.0885)	(0.1016)	(0.0784)	(0.1748)	(0.2806)	(0.3451)
	$d = 6$	1.207	1.302	1.274	2.420	2.553	2.844
		(0.0651)	(0.0780)	(0.0632)	(0.1217)	(0.2104)	(0.1783)
	$d = 9$	1.263	1.299	1.182	2.338	2.722	2.721
		(0.0618)	(0.0597)	(0.0683)	(0.1440)	(0.2321)	(0.1664)
K-P-F	$d = 3$	1.120*	1.101	1.176	1.372	1.320	1.061*
		(0.0668)	(0.0508)	(0.0611)	(0.0731)	(0.0753)	(0.0697)
	$d = 6$	1.235	1.163	1.107	1.379	1.195	1.278
		(0.0644)	(0.0621)	(0.0715)	(0.0946)	(0.0589)	(0.0663)
	$d = 9$	1.134	1.283	1.202	1.317	1.293	1.132
		(0.0712)	(0.0710)	(0.0602)	(0.0793)	(0.0801)	(0.0708)

Roueff and Sanchez-Perez (2018), Giraud et al. (2015) and Kley et al. (2019), respectively, with different time lags $d = 3, 6, 9$. In general, the non-stationary predictor (4.9) performs better or similar as the alternative methods with different time lag d in all scenarios. Our simulation results also demonstrate that the performance of R-S, G-R-S and K-P-F predictors depend sensitively on the choice of d. Finally, the large numbers in R-S predictor is due to the singularity of estimated local covariance matrix. We expect that this can be corrected by using an eigenvalue corrected positive definite covariance matrix estimator similar to (4.12).

We also examine the distribution of the prediction error as investigated in Theorem 4.1. For this purpose we show in Figure 1 the QQ plot of prediction errors of the predictors (4.9) for standard normal distributed errors and centered and standardized $\chi^2(6)$-distributed errors in model (5.7), respectively. The model is given by (5.7) and the sample sizes is $n = 1000$. These results confirm the theoretical findings in Theorem 4.1.

Finally, we compare the new predictor (4.9) with the methods proposed by Roueff and Sanchez-Perez (2018), Giraud et al. (2015) and Kley et al. (2019) in a locally stationary

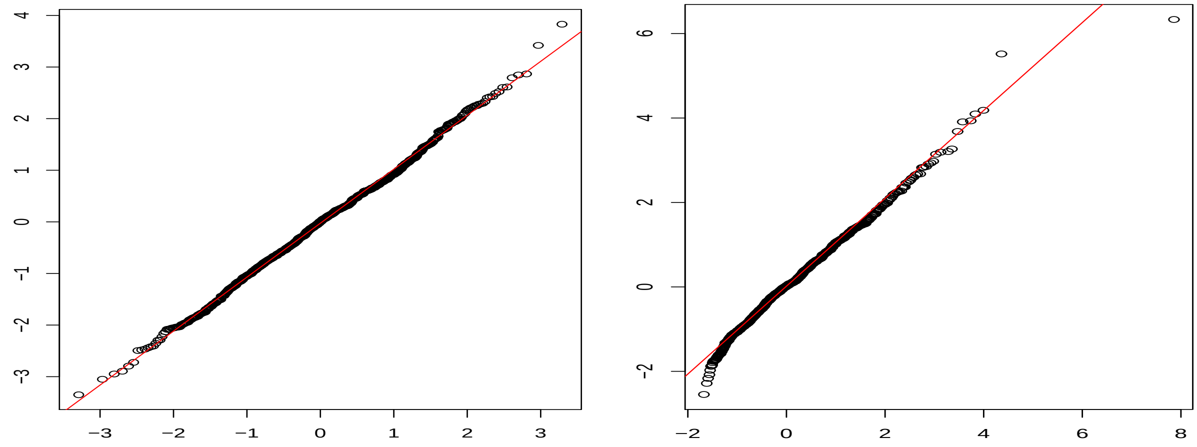

Figure 1: *QQ plots of prediction errors. Left part: standard normal distributed errors. Right part:* $(\mathcal{X}^2(6) - 6)/\sqrt{12}$*-distributed errors.*

MA(6) model defined by

$$G(t, \mathcal{F}_i) = \prod_{s=1}^{6} (1 - a_s(t)\mathcal{B})\sigma(t)\varepsilon_i, \tag{5.8}$$

where the time varying coefficients $a_1, \ldots a_6$ and the function σ are the same as those defined in the locally stationary AR(6) model (5.7), the mean function is given by (5.4) and the random variables ε_i are independent standard normal distributed. The results are presented in Table 5 and we observe similar properties as in the locally stationary AR(6) model (5.7). A detailed discussion is omitted for the sake of brevity.

5.3 Market indices analysis

In this section we apply our method to predict market indices. Let p_t be the adjusted daily closing value at day t, then the log return r_t is defined as

$$r_t = \log p_t - \log p_{t-1}.$$

As pointed out by Stărică and Granger (2005), the sign of r_t is unpredictable. As a result, these authors proposed to model r_t as

$$\log |r_t| = \mu(t) + \sigma(t)\epsilon_t \tag{5.9}$$

Table 5: *Simulated mean squared error of different predictors with MA(6) model* (5.8). *The numbers in brackets show the standard error and the index* $*$ *represents the predictor with the best best performance.*

		$t_{pred} = 0.5$			$t_{pred} = 1$		
Method	lag	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
(4.9)	-	1.187	1.090*	1.083	1.346*	1.234*	1.092*
		(0.0504)	(0.0509)	(0.0470)	(0.0627)	(0.0554)	(0.0511)
R-S	$d = 3$	1.222	1.152	1.144	1.505	1.287	1.102
		(0.0571)	(0.0532)	(0.0503)	(0.0673)	(0.0580)	(0.0475)
	$d = 6$	1.331	1.137	1.228	1.869	1.405	1.266
		(0.0569)	(0.0511)	(0.0519)	(0.0912)	(0.1037)	(0.0511)
	$d = 9$	8.757	1.338	1.138	254.780	2.128	1.247
		(1.213)	(0.0596)	(0.0515)	(175.380)	(0.1643)	(0.0533)
G-R-S	$d = 3$	1.232	1.255	1.296	2.462	2.484	2.042
		(0.0557)	(0.0562)	(0.0642)	(0.1044)	(0.2468)	(0.1060)
	$d = 6$	1.167	1.257	1.035	2.169	1.868	1.793
		(0.0544)	(0.0539)	(0.0492)	(0.0973)	(0.0839)	(0.0849)
	$d = 9$	1.128*	1.178	1.087	1.985	2.064	1.949
		(0.0610)	(0.0604)	(0.0543)	(0.0943)	(0.0925)	(0.0882)
K-P-F	$d = 3$	1.286	1.280	1.051*	1.571	1.404	1.292
		(0.0497)	(0.0599)	(0.0456)	(0.0677)	(0.0595)	(0.0545)
	$d = 6$	1.177	1.179	1.244	1.523	1.321	1.288
		(0.0595)	(0.0538)	(0.0516)	(0.0751)	(0.0669)	(0.0548)
	$d = 9$	1.296	1.238	1.158	1.649	1.449	1.310
		(0.0524)	(0.0511)	(0.0479)	(0.0724)	(0.0640)	(0.0606)

where μ and σ are time varying functions and ϵ_t denotes a zero-mean noise process. Stărică and Granger (2005) used model (5.9) to study the non-stationarity of stock returns. In this section we apply the new method to predict $y_t := \log(|r_t|)$ for the SP500, NASDAQ and Dow Jones Index. We consider data from Dec. 19, 2016 to Dec. 17, 2019. For SP500, NASDAQ and Dow Jones Index, we delete the log return of Jan. 10, 2017, Nov. 13, 2018 and Nov. 12, 2019 respectively due to their negative infinity values. Therefore the lengths of the series are 752. We use the new method to predict the market indices at trading days between April. 8, 2019 and Dec. 17, 2019 for SP500 and NASDAQ and at trading days between April. 5, 2019 and Dec. 17, 2019 for Dow Jones Series, respectively, and calculate the empirical mean squared error for these predictions. For the sake of comparison we also apply the methods of Roueff and Sanchez-Perez (2018) (R-S), Giraud et al. (2015) (G-R-S) and Kley et al. (2019) (K-P-F) to the same series. As in the simulation, for fair comparison we perform those algorithms on non-parametrically de-trended data and use the outcome plus $\hat{\mu}((T-1)/T)$ as the prediction of indices at day T. The corresponding results are listed in Table 6, where we use the different lags $3, 6, 9$ in the procedures based on autoregressive fitting. We observe that the new prediction method (4.9) shows the best performance for all three market indices. For NASDAQ index the method proposed by Kley et al. (2019) with $d = 9$ shows a similar performance. In general the parameter d for the prediction method proposed by Roueff and Sanchez-Perez (2018), Giraud et al. (2015) and Kley et al. (2019) is difficult to select, while it has a complicated impact on the predictions when applying those approaches. In Figure 2 we also plot the prediction error of the different methods for the three market indices. The left panels display $\log |r_t|$, while the right panels show absolute prediction errors of the prediction (4.9) and of the predictors proposed by Roueff and Sanchez-Perez (2018) (R-S), Giraud et al. (2015) (G-R-S) and Kley et al. (2019) (K-P-F) for the corresponding parameter $d \in \{3, 6, 9\}$, which achieves the smallest mean squared error.

Table 6: *Empirical mean squared error of different predictors for SP500, NASDAQ and Dow Jones. The notation $*$ marks the best method.*

Method	lag	SP500	NASDAQ	Dow Jones
(4.9)	-	1.456*	1.119*	1.745*
R-S	d=3	1.535	1.130	1.747
	d=6	1.586	1.142	1.873
	d=9	1.607	1.170	1.860
G-R-S	d=3	1.817	1.826	2.054
	d=6	2.689	1.350	2.361
	d=9	2.225	1.200	2.344
K-P-F	d=3	1.653	1.147	1.883
	d=6	1.707	1.124	1.938
	d=9	1.763	1.119*	1.932

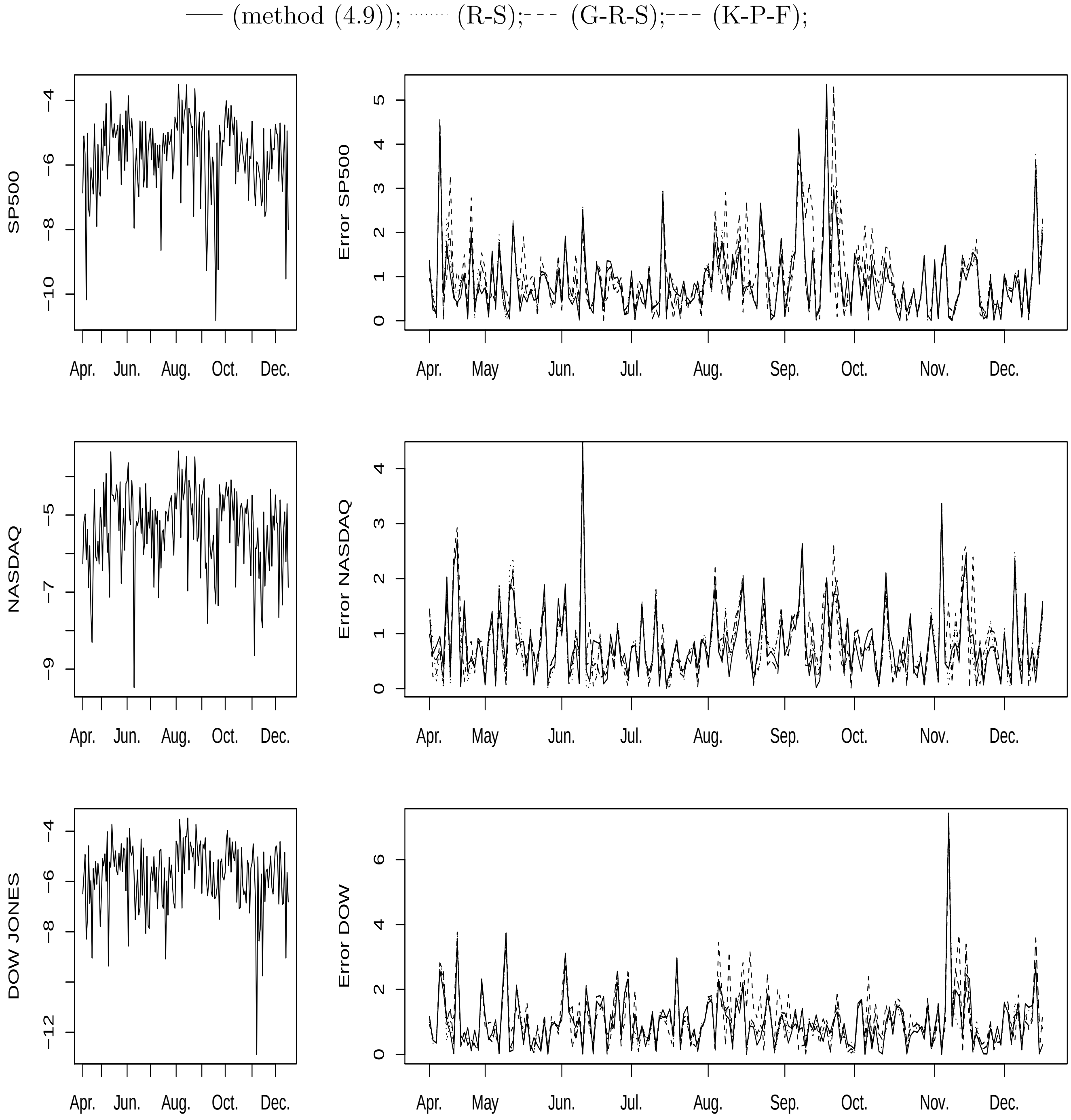

Figure 2: *Prediction of different market indices (left panels). Right Panel: the absolute prediction errors of the different methods*

6 Appendix: Proofs

In the proof, we shall use $\mathcal{P}_i(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_i) - \mathbb{E}(\cdot|\mathcal{F}_{i-1})$ as the projection operator. Let $\epsilon_{i,n} = 0$ and $\hat{\epsilon}_{i,n} = 0$ for $i \leq 0$ or $i > n$ for convenience. For a $p-$dimensional real vector $\mathbf{v} = (v_1, ..., v_p)^\top$, we write $|\mathbf{v}| = (\sum_{i=1}^p v_i^2)^{1/2}$ for its euclidean norm, and write $\|\mathbf{v}\|_q = \mathbb{E}\left(|\mathbf{v}|^q\right)^{1/q}$ if $\mathbf{v}$ is random. Let M denote a sufficiently large constant which varies from line to line. Write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For positive definite matrix A, define $\lambda_{max}(A)$ and $\lambda_{min}(A)$ be its largest and smallest eigenvalues, respectively.

6.1 Some auxiliary results

In this section we provide several auxiliary results, which will be used in the proofs of the main statements. The main result is Proposition 6.3, while Proposition 6.1 and 6.2 are used for a proof of this statement.

Proposition 6.1. *If assumptions (L1)-(L3), (M1) hold, $n\tau_n^3 \to \infty$ and $n\tau_n^6 = o(1)$, and $\lfloor cn \rfloor \leq m \leq n$ for some constant $c, 0 < c < 1$, then the local linear estimate in* (3.2) *satisfies*

$$\sup_{t\in[0,1]} \|\hat{\mu}^{1:m}(t) - \mu(t)\|_q = O(\tau_n^2 + (n\tau_n)^{-1/2}).$$

Proof. Define the quantities $M_k(t)$, $k = 0, 1, 2$ as

$$M_k(t) = \frac{1}{n\tau_n} \sum_{i=1}^m K\Big(\frac{i/n - t}{\tau_n}\Big)\Big(\frac{i/n - t}{\tau_n}\Big)^k.$$

The straightforward but tedious calculations by solving (4.3) we have for $t \in [0, \frac{m}{n}]$ the solution is

$$\hat{\mu}^{1:m}(t) = \frac{1}{n\tau_n} \sum_{i=1}^m \big(\mu(\frac{i}{n}) + \epsilon_{i,n}\big) K^*\Big(\frac{i/n - t}{\tau_n}\Big), \tag{6.1}$$

where

$$K^*\Big(\frac{i/n - t}{\tau_n}\Big) = \frac{M_2(t)K(\frac{i/n-t}{\tau_n}) - M_1(t)K(\frac{i/n-t}{\tau_n})(\frac{i/n-t}{\tau_n})}{M_0(t)M_2(t) - M_1^2(t)},$$

with $0/0 = 0$ for convenience. Observe that K^* is bounded and has a compact support on

$[-1,1]$. Observing the identity

$$\Big\|\sum_{i=1}^{m}\frac{1}{n\tau_n}K^*\Big(\frac{i/n-t}{\tau_n}\Big)\epsilon_{i,n}\Big\|_q=\Big\|\frac{1}{n\tau_n}\sum_{k=0}^{\infty}\sum_{i=1}^{m}\mathcal{P}_{i-k}K^*\Big(\frac{i/n-t}{\tau_n}\Big)\epsilon_{i,n}\Big\|_q, \tag{6.2}$$

and applying Burkholder's inequality to the martingale difference $\sum_{i=1}^m \mathcal{P}_{i-k}K^*\big(\frac{i/n-t}{\tau_n}\big)\epsilon_{i,n}$ shows

$$\Big\|\sum_{i=1}^{m}\mathcal{P}_{i-k}K^*\Big(\frac{i/n-t}{\tau_n}\Big)\epsilon_{i,n}\Big\|_q^2\le C_0q\sum_{i=1}^{m}\Big\|\mathcal{P}_{i-k}K^*\Big(\frac{i/n-t}{\tau_n}\Big)\epsilon_{i,n}\Big\|_q^2\le C_0qn\tau_n\delta_q^2(k) \tag{6.3}$$

for some constant C_0, where we have used the same arguments as given in the proof of Theorem 1 in Wu (2005) for the last inequality, and have used the fact that $m\ge\lfloor cn\rfloor$. Combining (6.2) and (6.3) leads to

$$\Big\|\sum_{i=1}^{m}\frac{1}{n\tau_n}K^*\Big(\frac{i/n-t}{\tau_n}\Big)\epsilon_{i,n}\Big\|_q\le C_0^{1/2}q^{1/2}(n\tau_n)^{-1/2}\sum_{k=0}^{\infty}\delta_q(k). \tag{6.4}$$

Now elementary calculations using condition (M1) with Taylor expansion show that

$$\sup_{t\in[0,1]}\Big|\frac{1}{n\tau_n}\sum_{i=1}^{m}\mu(\frac{i}{n})K^*\Big(\frac{i/n-t}{\tau_n}\Big)-\mu(t)\Big|=O(\tau_n^2). \tag{6.5}$$

Then the the assertion follows from (6.1), (6.4) and (6.5). ♢

Proposition 6.2. *If assumptions (L1)-(L3), (M1) are satisfied, $n\tau_n^3\to\infty$ and $n\tau_n^6=o(1)$, then we have for $1\le k\le n$,*

$$\Big\|\max_{1\le j\le n}|\sum_{i=1}^{j}(\epsilon_{i,n}\epsilon_{i+k,n}-\hat\epsilon_{i,n}\hat\epsilon_{i+k,n})|\Big\|_{q/2}=O(\alpha_n),$$

where $\alpha_n=n\tau_n^3+\tau_n^{-1}+\sqrt{n\tau_n}$.

Proof. Proposition 2 follows using similar arguments as given in the proof of Theorem 3.1 in Dette et al. (2019). ♢

Proposition 6.3. *If the assumptions of Theorem 3.2 are satisfied, and $0\le k\le l_n$, there* *exists a sufficiently large constant M such that*

$$\textit{(i)} \quad \sup_{t\in[0,1]} \|\hat{\gamma}_k(t)-\gamma_k(t)\|_{q/2} \le M\Big((nb_n)^{-1/2}+D_k b_n^2+\frac{k}{n}+\frac{\alpha_n}{nb_n}\Big),$$

$$\textit{(ii)} \quad \Big\|\sup_{t\in[0,1]} |\hat{\gamma}_k(t)-\gamma_k(t)|\Big\|_{q/2} \le M\Big(b_n^{-2/q}(nb_n)^{-1/2}+D_k b_n^2+\frac{k}{n}+\frac{\alpha_n}{nb_n}\Big).$$

Proof. Without loss of generality, we assume that the lag k is even and define $\tilde{\gamma}_k(t)$ as the analogue $\hat{\gamma}_k(t)$ in (3.6), where the residuals $\hat{\epsilon}_{i,n}$ are replaced by the "true" errors $\epsilon_{i,n}$, that is

$$(\tilde{\gamma}_k(t),\tilde{\gamma}_k'(t))^\top = \operatorname*{argmin}_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^n \big(\epsilon_{i-k/2,n}\epsilon_{i+k/2,n}-\beta_0-\beta_1(i/n-t)\big)^2 K\Big(\frac{i/n-t}{b_n}\Big).$$

Elementary calculations show that

$$\tilde{\gamma}_k(t) = \frac{\frac{M_2(t)}{nb_n}\sum_{i=1}^n \epsilon_{i-k/2,n}\epsilon_{i+k/2,n}K(\frac{i/n-t}{b_n}) - \frac{M_1(t)}{nb_n}\sum_{i=1}^n \epsilon_{i-k/2,n}\epsilon_{i+k/2,n}K(\frac{i/n-t}{b_n})(\frac{i/n-t}{b_n})}{M_0(t)M_2(t)-M_1^2(t)}, \tag{6.6}$$

where

$$M_k(t)=\frac{1}{nb_n}\sum_{i=1}^n K\Big(\frac{i/n-t}{b_n}\Big)\Big(\frac{i/n-t}{b_n}\Big)^k, \quad k=0,1,2.$$

Similarly, we have

$$\hat{\gamma}_k(t) = \frac{\frac{M_2(t)}{nb_n}\sum_{i=1}^n \hat{\epsilon}_{i-k/2,n}\hat{\epsilon}_{i+k/2,n}K(\frac{i/n-t}{b_n}) - \frac{M_1(t)}{nb_n}\sum_{i=1}^n \hat{\epsilon}_{i-k/2,n}\hat{\epsilon}_{i+k/2,n}K(\frac{i/n-t}{b_n})(\frac{i/n-t}{b_n})}{M_0(t)M_2(t)-M_1^2(t)}$$

and using the summation by parts formula and Proposition 6.2 it follows that

$$\Big\|\sup_{t\in[0,1]} |\tilde{\gamma}_k(t)-\hat{\gamma}_k(t)|\Big\|_{q/2} = O(\frac{\alpha_n}{nb_n}).$$

uniformly with respect to $1 \le k \le n$ and it remains to show that

$$\text{(a)} \quad \sup_{t\in[0,1]} \Big\| \tilde{\gamma}_k(t) - \gamma_k(t) \Big\|_{q/2} \le M\Big((nb_n)^{-1/2} + D_k b_n^2 + \frac{k}{n} \Big),$$

$$\text{(b)} \quad \Big\| \sup_{t\in[0,1]} |\tilde{\gamma}_k(t) - \gamma_k(t)| \Big\|_{q/2} \le M\Big(b_n^{-2/q} (nb_n)^{-1/2} + D_k b_n^2 + \frac{k}{n} \Big).$$

Let $\eta_{i,k} = \epsilon_{i-k/2,n}\epsilon_{i+k/2,n}$ (note that k is even). By (6.6) we have

$$\tilde{\gamma}(t) = \tilde{M}_1(t)\big(\frac{1}{nb_n}\sum_{i=1}^n \eta_{i,k} K\big(\frac{i/n-t}{b_n}\big)\big) + \tilde{M}_2(t)\big(\frac{1}{nb_n}\sum_{i=1}^n \eta_{i,k} K\big(\frac{i/n-t}{b_n}\big)\big(\frac{i/n-t}{b_n}\big)\big)$$

with $\tilde{M}_1(t) = \frac{M_2(t)}{M_0(t)M_2(t)-M_1^2(t)}$, $\tilde{M}_2(t) = \frac{-M_1(t)}{M_0(t)M_2(t)-M_1^2(t)}$. Notice that

$$\gamma(t) = \tilde{M}_1(t)\big(\frac{1}{nb_n}\sum_{i=1}^n \gamma_k(t) K\big(\frac{i/n-t}{b_n}\big)\big) + \tilde{M}_2(t)\big(\frac{1}{nb_n}\sum_{i=1}^n \gamma_k(t) K\big(\frac{i/n-t}{b_n}\big)\big(\frac{i/n-t}{b_n}\big)\big)$$

As a result, we can decompose $\tilde{\gamma}_k(t) - \gamma_k(t)$ into a random part and a deterministic part, i.e.

$$\tilde{\gamma}_k(t) - \gamma_k(t) = \Xi_k^d(t) + \Xi_k^s(t),$$

where

$$\begin{aligned}
\Xi_k^d(t) &= \tilde{M}_1(t)\frac{1}{nb_n}\sum_{i=1}^n (\mathbb{E}\eta_{i,k} - \gamma_k(t))K\big(\frac{i/n-t}{b_n}\big) \\
&\qquad + \tilde{M}_2(t)\frac{1}{nb_n}\sum_{i=1}^n (\mathbb{E}\eta_{i,k} - \gamma_k(t))K\big(\frac{i/n-t}{b_n}\big)\big(\frac{i/n-t}{b_n}\big), \\
\Xi_k^s(t) &= \tilde{M}_1(t)\frac{1}{nb_n}\sum_{i=1}^n (\eta_{i,k} - \mathbb{E}\eta_{i,k})K\big(\frac{i/n-t}{b_n}\big) \\
&\qquad + \tilde{M}_2(t)\frac{1}{nb_n}\sum_{i=1}^n (\eta_{i,k} - \mathbb{E}\eta_{i,k})K\big(\frac{i/n-t}{b_n}\big)\big(\frac{i/n-t}{b_n}\big).
\end{aligned}$$

To complete the proof we will show that (uniformly for $0 \le k \le l_n$)

$$\sup_{t\in[0,1]} |\Xi^d_k(t)| \le M\left(D_k b_n^2 + \frac{k}{n} + \frac{1}{nb_n}\right), \tag{6.7}$$

$$\sup_{t\in[0,1]} \|\Xi^s_k(t)\|_{q/2} \le M\left((nb_n)^{-1/2}\right), \tag{6.8}$$

$$\| \sup_{t\in[0,1]} |\Xi^s_k(t)| \|_{q/2} \le M\left(b_n^{-2/q}(nb_n)^{-1/2}\right). \tag{6.9}$$

Observe that $\Xi^d_k(t)$ can be further decomposed as

$$\Xi^d_k(t) = \Xi^d_{1,k}(t) + \Xi^d_{2,k}(t),$$

where

$$\begin{aligned}
\Xi^d_{1,k}(t) = \tilde{M}_1(t)\frac{1}{nb_n}&\sum_{i=1}^n (\mathbb{E}\eta_{i,k} - \gamma_k(i/n))K\big(\frac{i/n-t}{b_n}\big) \\
&+ \tilde{M}_2(t)\frac{1}{nb_n}\sum_{i=1}^n (\mathbb{E}\eta_{i,k} - \gamma_k(i/n))K\big(\frac{i/n-t}{b_n}\big)\big(\frac{i/n-t}{b_n}\big), \\
\Xi^d_{2,k}(t) = \tilde{M}_1(t)\frac{1}{nb_n}&\sum_{i=1}^n (\gamma_k(i/n) - \gamma_k(t))K\big(\frac{i/n-t}{b_n}\big) \\
&+ \tilde{M}_2(t)\frac{1}{nb_n}\sum_{i=1}^n (\gamma_k(i/n) - \gamma_k(t))K\big(\frac{i/n-t}{b_n}\big)\big(\frac{i/n-t}{b_n}\big).
\end{aligned}$$

By conditions (L1), (L2) and a Taylor expansion it follows that

$$\begin{aligned}
|\mathbb{E}(\eta_{i,k}) - \gamma_k(i/n)| =& \Big|\mathbb{E}\Big(G\Big(\frac{i-k/2}{n}, \mathcal{F}_0\Big)G\Big(\frac{i+k/2}{n}, \mathcal{F}_k\Big)\Big) - \mathbb{E}\Big(G\Big(\frac{i}{n}, \mathcal{F}_0\Big)G\Big(\frac{i}{n}, \mathcal{F}_k\Big)\Big)\Big| \\
=& O(k/n)
\end{aligned}$$

uniformly with respect to i. A straightforward but tedious calculation now shows that

$$\sup_{t\in[0,1]} |\Xi^d_{1,k}(t)| = O(k/n + \frac{1}{nb_n}) \tag{6.10}$$

as $n \to \infty$, uniformly with respect to $0 \le k \le l_n$. In addition by condition (A1), we obtain that

$$\sup_{t\in[0,1]} |\Xi^d_{2,k}(t)| = O(D_k b_n^2 + \frac{1}{nb_n}) \tag{6.11}$$

(uniformly for $0 \leq k \leq l_n$). As a result, inequality (6.7) follows from (6.10) and (6.11). For $\Xi^s_k(t)$, an application of the Cauchy-Schwartz inequality shows that

$$\|\mathcal{P}_{i+k/2-s}\eta_{i,k}\|_{q/2} \leq M(\delta_q(s) + \mathbf{1}(s \geq k)\delta_q(s-k)),$$

(uniformly with respect to i) and assertion (6.8) now follows using similar arguments as given in the proof of Proposition 6.1. By Assumption (K) and similar arguments as given in the proof of Proposition 6.1 we have

$$\sup_{t\in[0,1]} \Big\|\frac{\partial}{\partial t}\Xi^s_k(t)\Big\|_{q/2} \leq M\left((nb_n)^{-1/2} b_n^{-1}\right) \tag{6.12}$$

(uniformly with respect $0 \leq k \leq l_n$). Finally, inequality (6.9) follows from (6.8), (6.12) and Proposition B.1 in Dette et al. (2019), which completes the proof. $\diamondsuit$

6.2 Proof of Theorem 3.1 and 3.2

For the sake of brevity we restrict ourselves to the proof of Theorem 3.2. Theorem 3.1 can be shown by similar but substantially simpler arguments.

Define the banded matrix $\Sigma_{l_n,n} := (\sigma_{i,j}\mathbf{1}(|i-j| \leq l_n))$, where we use the symbol $\sigma_{i,j}$ for $\sigma_{i,j,n}$ to simplify the notation. Note that $\Sigma_{l_n,n} - \Sigma_n$ is a symmetric matrix and by Gershgorin's circle theorem it follows that

$$\begin{aligned}\rho(\Sigma_{l_n,n} - \Sigma_n) &\leq \max_{1\leq i\leq n} \sum_{j=1}^n |\sigma_{i,j} - \sigma_{i,j}\mathbf{1}(|i-j| \leq l_n)| \\ &\leq \max_{1\leq i\leq n} \Big(\sum_{j=1}^{(i-l_n)\vee 1} |\sigma_{i,j}| + \sum_{j=(i+l_n)\wedge n}^{n} |\sigma_{i,j}|\Big). \end{aligned} \tag{6.13}$$

Using similar arguments as given in the proof of Lemma 5 of Zhou and Wu (2010) it follows that

$$|\sigma_{i,j}| = O\Big(\sum_{s=1}^{\infty} \chi^{2s+|i-j|}\Big) = O(\chi^{|i-j|}) \ , \tag{6.14}$$

for all $i, j \in \mathbb{N}$, and straightforward calculations give

$$\max_{1\le i\le n}\Big(\sum_{j=1}^{(i-l_n)\vee 1}|\sigma_{i,j}|\Big) = O(\chi^{l_n})\ ,\quad \max_{1\le i\le n}\Big(\sum_{j=(i+l_n)\wedge n}^{n}|\sigma_{i,j}|\Big) = O(\chi^{l_n}).$$

Therefore we obtain from (6.13) the estimate

$$\rho(\Sigma_{l_n,n}-\Sigma_n) = O(\chi^{l_n}).$$

Note that, by definition, $\sigma_{i,j} = \mathbb{E}(G(\frac{i}{n},\mathcal{F}_i)G(\frac{j}{n},\mathcal{F}_j))$, $\gamma_{|i-j|}(\frac{i+j}{2n}) = \mathbb{E}(G(\frac{i+j}{2n},\mathcal{F}_i)G(\frac{i+j}{2n},\mathcal{F}_j))$, then using conditions (L1), (L2) we have

$$\max_{|i-j|\le l_n}\Big|\gamma_{|i-j|}(\frac{i+j}{2n}) - \sigma_{i,j}\Big| \le M\frac{l_n}{n} \tag{6.15}$$

for some large constant M.
On the other hand, similarly to (6.13) it follows that

$$\begin{aligned}\rho(\hat{\Sigma}_n - \Sigma_{l_n,n}) &\le \max_{1\le i\le n}\sum_{j=1}^{n}|\big(\sigma_{i,j} - \hat{\gamma}_{|i-j|}(\frac{i+j}{2n})\big)\mathbf{1}(|i-j|\le l_n)| \\ &= \max_{1\le i\le n}\Big(\sum_{j=(i-l_n)\vee 1}^{j=(i+l_n)\wedge n}|\hat{\gamma}_{|i-j|}(\frac{i+j}{2n}) - \sigma_{i,j}|\Big). \end{aligned}\tag{6.16}$$

By Proposition 6.3 it follows that

$$\begin{aligned}&\Big\|\max_{1\le i\le n}\sum_{j=(i-l_n)\vee 1}^{j=(i+l_n)\wedge n}|\hat{\gamma}_{|i-j|}(\frac{i+j}{2n}) - \sigma_{i,j}|\Big\|_{q/2} \\ \le&\Big\|\sum_{j=(i-l_n)\vee 1}^{j=(i+l_n)\wedge n}\max_{1\le i\le n}|\hat{\gamma}_{|i-j|}(\frac{i+j}{2n}) - \gamma_{|i-j|}(\frac{i+j}{2n})|\Big\|_{q/2} + \max_{1\le i\le n}\sum_{j=(i-l_n)\vee 1}^{j=(i+l_n)\wedge n}|\gamma_{|i-j|}(\frac{i+j}{2n}) - \sigma_{i,j}| \\ \le& M\Big(l_n(b_n^{-2/q}(nb_n)^{-1/2} + \frac{\alpha_n}{nb_n}) + \frac{l_n^2}{n} + \sum_{i=0}^{l_n}D_i b_n^2\Big), \end{aligned}\tag{6.17}$$

where the quantities D_k are defined in (A1), for which we have used (6.15) and the estimate

$$\Big\| \sum_{j=(i-l_n)\vee 1}^{j=(i+l_n)\wedge n} \max_{1\le i\le n} |\hat{\gamma}_{|i-j|}(\frac{i+j}{2n}) - \gamma_{|i-j|}(\frac{i+j}{2n})| \Big\|_{q/2}$$

$$= O\Big(\sum_{k=0}^{l_n} \Big(b_n^{-2/q}(nb_n)^{-1/2} + D_k b_n^2 + \frac{k}{n} + \frac{\alpha_n}{nb_n} \Big) \Big)$$

$$= O\Big(l_n \Big(b_n^{-2/q}(nb_n)^{-1/2} + \frac{l_n}{n} + \frac{\alpha_n}{nb_n} \Big) + \sum_{k=0}^{l_n} D_k b_n^2 \Big)$$

Therefore the theorem follows from (6.16) and (6.17).

6.3 Proof of Corollary 4.1

Condition (E1) shows that the quantity

$$W = \Sigma_{n,m}^{-1/2} \hat{\Sigma}_{n,m} \Sigma_{n,m}^{-1/2}.$$

is well defined. By our construction, W is positive definite with probability tending to 1. Then by (4.6) and condition (E1), we have that

$$\|\rho(W - I_{m\times m})\|_{q/2} = O(r_n),$$

where $I_{m\times m}$ is an $m \times m$ diagonal matrix. Now the corollary follows from the argument in the proof of Theorem 2 of McMurry and Politis (2010) and the fact that $\rho(\Sigma_{n,m})$ is bounded which is a consequence of Gershgorin's circle theorem.

6.4 Proof of Theorem 4.1

By the projection theorem, equation (4.2) is equivalent to

$$\mathbb{E}(X_{m+1,n} - X_{m+1,n}^{\text{pred}}) = 0, \quad \mathbb{E}((X_{m+1,n} - X_{m+1,n}^{\text{pred}})X_{j,n}) = 0; \quad j = 1, \ldots, m.$$

Using these equations in (4.1) yields

$$
\begin{aligned}
a_{m+1,n} &= \mu\left(\frac{m+1}{n}\right) - \sum_{s=1}^{m} a_{m+1-s,n}\mu\left(\frac{s}{n}\right)\ , \qquad (6.18)\\
\mathbb{E}\Big[\Big(G\Big(\frac{m+1}{n}, \mathcal{F}_{m+1}\Big) &- \sum_{s=1}^{m} a_{m+1-s,n} G\Big(\frac{s}{n}, \mathcal{F}_s\Big)\Big) G\Big(\frac{j}{n}, \mathcal{F}_j\Big)\Big] = 0
\end{aligned}
$$

$(1 \leq j \leq m)$, which shows that the vector $\mathbf{a}_m^*$ in (4.2) is given by

$$\mathbf{a}_m^* = \Sigma_{n,m}^{-1}\boldsymbol{\gamma}_m, \tag{6.19}$$

where $\boldsymbol{\gamma}_m = (\sigma_{m+1,1}, \ldots, \sigma_{m+1,m})^\top$. Let

$$\boldsymbol{\gamma}_{m,l_n} = (0, \ldots, 0, \sigma_{m+1,m-l_n+1}, \ldots, \sigma_{m+1,m})^\top$$

be the vector with j_{th} entry given by $\sigma_{m+1,j}\mathbf{1}(m+1-j \leq l_n)$. By the representation of $\hat{\mathbf{a}}_m^*$ in (4.8), we have

$$\hat{\mathbf{a}}_m^* - \mathbf{a}_m^* = G_1 + G_2 + G_3,$$

where the terms G_1, G_2 and G_3 are defined by

$$
\begin{aligned}
G_1 &= \hat{\Sigma}_{n,m}^{-1}(\hat{\boldsymbol{\gamma}}_n^{1:m} - \boldsymbol{\gamma}_{m,l_n}),\\
G_2 &= (\hat{\Sigma}_{n,m}^{-1} - \Sigma_{n,m}^{-1})\boldsymbol{\gamma}_{m,l_n},\\
G_3 &= \Sigma_{n,m}^{-1}(\boldsymbol{\gamma}_{m,l_n} - \boldsymbol{\gamma}_m).
\end{aligned}
$$

In the following we shall show that $G_j = O_{\mathbb{P}}(r_n)$ for $j = 1, 2, 3$, which implies

$$|\hat{\mathbf{a}}_m^* - \mathbf{a}_m^*| = O_{\mathbb{P}}(r_n). \tag{6.20}$$

Using similar arguments as given in the derivation of (6.17) we have

$$\|\hat{\boldsymbol{\gamma}}_n^{1:m} - \boldsymbol{\gamma}_{m,l_n}\|_{q/2} = \Big\|\Big(\sum_{s=m+1-l_n}^{m} |\hat{\gamma}_{m+1-s}^{1:m}\big(\frac{m+s}{2n}\big) - \sigma_{m+1,s}|^2\Big)^{1/2}\Big\|_{q/2} = O(r_n).$$

A straightforward calculation using assumption (E1) and Corollary 4.1 show

$$G_1 \leq |\rho(\hat{\Sigma}_{n,m}^{-1})||\hat{\boldsymbol{\gamma}}_n^{1:m} - \boldsymbol{\gamma}_{m,l_n}| = O_{\mathbb{P}}(r_n).$$

By (6.14) $|\boldsymbol{\gamma}_{m,l_n}|$ is bounded. By Corollary 4.1 it also follows $G_2 = O_{\mathbb{P}}(r_n)$. Observing (6.14) we obtain

$$|\boldsymbol{\gamma}_{m,l_n} - \boldsymbol{\gamma}_m| = \Big(\sum_{j=1}^{m-l_n} \sigma_{m,j}^2\Big)^{1/2} \leq M\chi^{l_n} \tag{6.21}$$

which implies $G_3 = O(r_n)$, and hence (6.20) follows. For a proof of part (a), it now remains to show that

$$|\hat{a}_{m+1,n} - a_{m+1,n}| = O_{\mathbb{P}}(r_n^\circ). \tag{6.22}$$

From (6.18) and definition (4) it follows that

$$\begin{aligned}\hat{a}_{m+1,n} - a_{m+1,n} &= \hat{\mu}^{1:m}\big(\frac{m}{n}\big) - \mu\big(\frac{m+1}{n}\big) + \Big(\sum_{s=1}^m a_{m+1-s,n}\mu\big(\frac{s}{n}\big) - \sum_{s=1}^m \hat{a}_{m+1-s,n}\hat{\mu}^{1:m}\big(\frac{s}{n}\big)\Big)\\ &= \Big(\hat{\mu}^{1:m}\big(\frac{m}{n}\big) - \mu\big(\frac{m+1}{n}\big)\Big) + \sum_{s=1}^m a_{m+1-s,n}\Big(\mu\big(\frac{s}{n}\big) - \hat{\mu}^{1:m}\big(\frac{s}{n}\big)\Big)\\ &+ \sum_{s=1}^m \hat{\mu}^{1:m}\big(\frac{s}{n}\big)\big(a_{m+1-s,n} - \hat{a}_{m+1-s,n}\big)\\ &:= H_1 + H_2 + H_3,\end{aligned}$$

where the statistics H_1, H_2 and H_3 are defined in an obvious way. Using assumption (M1) and Proposition 6.1, we have that $\|H_1\|_q = O(\tau_n^2 + (n\tau_n)^{-1/2})$. For an estimate of H_2 we need to determine the order of $\mathbf{a}_m^*$ defined in (4.2). For this purpose we define

$$\Sigma_{n,m,l_n} = (\sigma_{i,j,n}\mathbf{1}(|i-j| \leq l_n))_{1\leq i,j\leq n}, \quad \mathbf{a}_{m,l_n}^* = \Sigma_{n,m,l_n}^{-1}\boldsymbol{\gamma}_{m,l_n}\ ,$$

then using (6.14) and (6.21) we get

$$|\mathbf{a}_{m,l_n}^* - \mathbf{a}_m^*| = O(\chi^{l_n}). \tag{6.23}$$

Denote by $a_{m,l_n,j}$, $\gamma_{m,l_n,j}$ the j_{th} entry of the vector $\mathbf{a}_{m,l_n}^*$ and $\boldsymbol{\gamma}_{m,l_n}$, respectively. Define

$$H_{2,l_n} = \sum_{s=1}^m a_{m,l_n,m+1-s}\Big(\mu\big(\frac{s}{n}\big) - \hat{\mu}^{1:m}\big(\frac{s}{n}\big)\Big),$$

then, by (6.23) and Proposition 6.1, it follows that

$$H_{2,l_n} - H_2 = O_{\mathbb{P}}(\sqrt{n}\chi^{l_n}(\tau_n^2 + (n\tau_n)^{-1/2})). \tag{6.24}$$

Hence it suffices to study the order of H_{2,l_n}. Denote the $(i,j)_{th}$ entry of the matrix Σ^{-1}_{n,m,l_n} by $\Sigma^{-1}_{n,m,l_n}(i,j)$. Since Σ_{n,m,l_n} is l_n-banded, $\lim_{n\to\infty} \|\Sigma_{n,m,l_n}\|_F < \infty$ and condition $(E1)$, we can apply Proposition 2.2 of Demko et al. (1984), and obtain

$$|\Sigma^{-1}_{n,m,l_n}(i,j)| \leq C_n q_n^{\frac{2|i-j|}{l_n}}\ , \tag{6.25}$$

where $q_n = (\sqrt{r_n}-1)/(\sqrt{r_n}+1)$, $r_n = \lambda_{max}(\Sigma_{n,m,l_n})/\lambda_{min}(\Sigma_{n,m,l_n})$, $C_n = \max(\lambda^{-1}_{min}, C_{0n})$, $C_{0n} = (1+r_n^{1/2})^2/(2\lambda_{min}(A)r_n)$. By condition $(E1)$ and (6.14), it follows that there exists a positive constant M and a constant $Q \in (0,1)$ such that

$$C_n \leq M, \quad 0 < q_n \leq Q < 1.$$

Then, if h a is positive constant such that $nQ^{h\log n} l_n^{1/2} = O(\log^{1/2} n)$, we have uniformly for $1 \leq i \leq m - hl_n \log n$

$$\begin{aligned} a_{m,l_n,i} &= \sum_{j=1}^{m} \Sigma^{-1}_{n,m,l_n}(i,j)\gamma_{m,l_n,j} = \sum_{j=m-l_n+1}^{m} \Sigma^{-1}_{n,m,l_n}(i,j)\gamma_{m,l_n,j} \\ &= O(l_n Q^{h\log n}) = O\Big(\frac{l_n^{1/2}\log^{1/2} n}{n}\Big). \end{aligned}$$

On the other hand, observing the fact $|\mathbf{a}^*_m| \leq |\rho(\Sigma^{-1}_{n,m})||\boldsymbol{\gamma}_m| < \infty$ yields

$$\sum_{i\in(m-hl_n\log n, m]} a^2_{m,l_n,i} \leq M' < \infty \tag{6.26}$$

for some constant M'. Thus it follows from Proposition 6.1 and an application of the Cauchy Schwarz inequality that

$$\begin{aligned} |H_{2,l_n}| \leq & \Big| \sum_{s=1}^{m} a_{m,l_n,m+1-s}\Big(\mu(\frac{s}{n}) - \hat{\mu}^{1:m}(\frac{s}{n})\Big)\mathbf{1}(s \geq hl_n \log n + 1)\Big| \\ & + \Big| \sum_{s=1}^{m} a_{m,l_n,m+1-s}\Big(\mu(\frac{s}{n}) - \hat{\mu}^{1:m}(\frac{s}{n})\Big)\mathbf{1}(s < hl_n \log n + 1)\Big| \\ & = O_{\mathbb{P}}(l_n^{1/2}\log^{1/2} n(\tau_n^2 + (n\tau_n)^{-1/2})). \end{aligned} \tag{6.27}$$

Equation (6.24) and (6.27) now show that $H_2 = O_{\mathbb{P}}(r_n^\circ)$, where r_n° is defined in (4.10). Finally, for the estimate of H_3 we define

$$H_{3,l_n} = \sum_{s=1}^{m} \hat{\mu}^{1:m}\big(\frac{s}{n}\big)\big(a_{m,l_n,m+1-s} - \hat{a}_{m+1-s,n}\big).$$

By (6.23) we find $|H_{3,l_n} - H_3| = O_{\mathbb{P}}(\sqrt{n}\chi^{l_n})$. Notice that (6.20) and (6.23) yield that $|\hat{\mathbf{a}}_m^* - \mathbf{a}_{m,l_n}^*| = O_{\mathbb{P}}(r_n)$. Furthermore, similarly to (6.25), using Proposition 2.2 of Demko et al. (1984) it follows that there exist constants $M_0 > 0$ and $Q_0 \in (0,1)$ such that

$$|\hat{\Sigma}_{n,m}^{-1}(i,j)| \leq M_0 Q_0^{\frac{2|i-j|}{l_n}} \ , \ \ 1 \leq i,j \leq m$$

with probability tending to 1. Using this fact and similar arguments as for the derivation of (6.27), we obtain $|H_{3,l_n}| = O_{\mathbb{P}}((l_n^{1/2}\log^{1/2} n) r_n) = O_{\mathbb{P}}(r_n^\circ)$. This proves (6.22) and completes the proof of part (a).

For a proof of part (b), we recall the definition of the filter G in (2.2) and obtain

$$G(\tfrac{m+1}{n}, \mathcal{F}_{m+1}) = \sum_{s=1}^{p} a_s(\tfrac{m+1}{n}) G(\tfrac{m+1}{n}, \mathcal{F}_{m+1-s}) + \sum_{s=p+1}^{m} d_s G(\tfrac{s}{n}, \mathcal{F}_{m+1-s}) + \sigma(\tfrac{m+1}{n})\varepsilon_{m+1},$$

where $d_s = 0$, for $p+1 \leq s \leq m$. Observe that

$$\begin{aligned}\mathbb{E}\big(G(\tfrac{m+1}{n}, \mathcal{F}_{m+1}) G(\tfrac{j}{n}, \mathcal{F}_j)\big) &= \sum_{s=1}^{p} a_s(\tfrac{m+1}{n}) \mathbb{E}\big(G(\tfrac{m+1}{n}, \mathcal{F}_{m+1-s}) G(\tfrac{j}{n}, \mathcal{F}_j)\big) \\ &+ \sum_{s=p+1}^{m} d_s \mathbb{E}\big(G(\tfrac{s}{n}, \mathcal{F}_{m+1-s}) G(\tfrac{j}{n}, \mathcal{F}_j)\big), \qquad (6.28)\end{aligned}$$

$(1 \leq j \leq m-p)$, and

$$\begin{aligned}\mathbb{E}(G\big(\tfrac{m+1}{n}, \mathcal{F}_{m+1}) G(\tfrac{m+1}{n}, \mathcal{F}_j)\big) &= \sum_{s=1}^{p} a_s(\tfrac{m+1}{n}) \mathbb{E}\big(G(\tfrac{m+1}{n}, \mathcal{F}_{m+1-s}) G(\tfrac{m+1}{n}, \mathcal{F}_j)\big) \\ &+ \sum_{s=p+1}^{m} d_s \mathbb{E}\big(G(\tfrac{s}{n}, \mathcal{F}_{m+1-s}) G(\tfrac{m+1}{n}, \mathcal{F}_j)\big)\end{aligned}$$

$(m-p+1 \leq j \leq m)$. Define

$$\begin{aligned}
\tilde{\sigma}_{i,j} &= \mathbb{E}(G(\tfrac{i}{n}, \mathcal{F}_i)G(\tfrac{j}{n}, \mathcal{F}_j)) \ , \quad 1 \leq i \leq m-p \ , \quad 1 \leq j \leq m-p, \\
\tilde{\sigma}_{i,j} &= \mathbb{E}(G(\tfrac{i}{n}, \mathcal{F}_i)G(\tfrac{m+1}{n}, \mathcal{F}_j), \quad 1 \leq i \leq m-p \ , \quad m-p+1 \leq j \leq m+1, \\
\tilde{\sigma}_{i,j} &= \mathbb{E}(G(\tfrac{m+1}{n}, \mathcal{F}_i)G(\tfrac{j}{n}, \mathcal{F}_j)) \ , \quad m-p+1 \leq i \leq m+1 \ , \quad 1 \leq j \leq m-p, \\
\tilde{\sigma}_{i,j} &= \mathbb{E}(G(\tfrac{m+1}{n}, \mathcal{F}_i)G(\tfrac{m+1}{n}, \mathcal{F}_j)) \ , \quad m-p+1 \leq i \leq m+1 \ , \quad m-p+1 \leq j \leq m+1,
\end{aligned}$$

(note that $\tilde{\sigma}_{i,j} = \tilde{\sigma}_{j,i}$). These notations and the equations (6.28) and (6.29) show that the m-dimensional vector $\tilde{\mathbf{a}}_m = \big(0, .., 0, a_p(\frac{m+1}{n}), a_{p-1}(\frac{m+1}{n}), ..., a_1(\frac{m+1}{n})\big)^\top$ satisfies

$$\Sigma_m^{AR} \tilde{\mathbf{a}}_m = \boldsymbol{\gamma}_m^{AR},$$

where the $m \times m$ matrix Σ_m^{AR} and the m-dimensional vector $\boldsymbol{\gamma}_m^{AR}$ are defined by $\Sigma_m^{AR} = (\tilde{\sigma}_{i,j})_{1 \leq i,j \leq m}$ and $\gamma_m^{AR} = (\tilde{\sigma}_{m+1,1}, ..., \tilde{\sigma}_{m+1,m})^\top$, respectively. On the other hand we have

$$\begin{aligned}
\hat{X}_{m+1,n}^{\text{pred}} &= \hat{\mu}^{1:m}(\tfrac{m}{n}) + \sum_{s=1}^{m} \hat{a}_{m+1-s,n}(\mu(\tfrac{s}{n}) - \hat{\mu}^{1:m}(\tfrac{s}{n})) + (\hat{\mathbf{a}}_m^*)^\top \mathbf{Z}, \\
X_{m+1,n} &= \mu(\tfrac{m+1}{n}) + \tilde{\mathbf{a}}_m^\top \tilde{\mathbf{Z}} + \sigma(\tfrac{m+1}{n})\varepsilon_{m+1}.
\end{aligned}$$

where the m-dimensional vectors $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are given by

$$\begin{aligned}
\mathbf{Z} &= \Big(G(\tfrac{1}{n}, \mathcal{F}_1), , ..., , G(\tfrac{m+1-p}{n}, \mathcal{F}_{m+1-p}), G(\tfrac{m+2-p}{n}, \mathcal{F}_{m+2-p}), ..., G(\tfrac{m}{n}, \mathcal{F}_m)\Big)^\top \\
\tilde{\mathbf{Z}} &= \Big(G(\tfrac{1}{n}, \mathcal{F}_1), , ..., G(\tfrac{m-p}{n}, \mathcal{F}_{m-p}), G(\tfrac{m+1}{n}, \mathcal{F}_{m+1-p}), G(\tfrac{m+1}{n}, \mathcal{F}_{m+2-p}), ...G(\tfrac{m+1}{n}, \mathcal{F}_m)\Big)^\top.
\end{aligned}$$

(note that the first $m-p$ elements of the two vectors coincide). Therefore we obtain the following decomposition

$$\hat{X}_{m+1,n}^{\text{Pred}} - X_{m+1,n} = W_1 + W_2 - \sigma(\tfrac{m+1}{n})\varepsilon_{m+1},$$

where

$$\begin{aligned}
W_1 &= \hat{\mu}^{1:m}(\tfrac{m}{n}) - \mu(\tfrac{m+1}{n}) + \sum_{s=1}^{m} \hat{a}_{m+1-s,n}\big(\mu(\tfrac{s}{n}) - \hat{\mu}^{1:m}(\tfrac{s}{n})\big), \\
W_2 &= (\hat{\mathbf{a}}_m^*)^\top \mathbf{Z} - \tilde{\mathbf{a}}_m^\top \tilde{\mathbf{Z}} := W_{2,1} + W_{2,2}, \\
W_{2,1} &= (\hat{\mathbf{a}}_m^*)^\top (\mathbf{Z} - \tilde{\mathbf{Z}}), W_{2,2} = ((\hat{\mathbf{a}}_m^*)^\top - \tilde{\mathbf{a}}_m^\top)\tilde{\mathbf{Z}}.
\end{aligned}$$

It now follows from the proof of (6.22) that $W_1 = O_{\mathbb{P}}(r_n^\circ)$. To derive a similar estimate for the term $W_{2,1}$ we note that by (6.26) and (6.20)

$$|\hat{\mathbf{a}}^*_{\mathbf{m}}| = O_{\mathbb{P}}(1).$$

Straightforward but tedious calculations using condition (L2) yield that

$$|\mathbf{Z} - \tilde{\mathbf{Z}}| = O_{\mathbb{P}}\Big(\frac{p^{3/2}}{n}\Big),$$

which leads to $W_{2,1} = O_{\mathbb{P}}(r_n)$. For estimation of $W_{2,2}$, note that a maximal inequality shows

$$|\tilde{\mathbf{Z}}|_\infty = \max_{i=1}^{m} |\tilde{Z}_i| = O_{\mathbb{P}}(n^{\frac{1}{q}}) \ . \tag{6.29}$$

We will show below that

$$|\tilde{\mathbf{a}}_m - \mathbf{a}^*_m| = O(\frac{p}{n}). \tag{6.30}$$

which yields with (6.20) the estimate $|\hat{\mathbf{a}}^*_m - \tilde{\mathbf{a}}_m| = O_{\mathbb{P}}(r_n)$. Observing (6.29) we have $W_{2,2} = O_{\mathbb{P}}(n^{\frac{1}{q}} r_n)$, which completes the proof of part (b), observing the fact that ε_{m+1} is identically distributed with ε_1.

In order to show (6.30) we use conditions (P2), (P3), will prove that

$$\sup_{1\le i,j\le m+1} |\tilde{\sigma}_{i,j} - \sigma_{i,j}| = O\Big(\frac{2m+2-h_m(i)-h_m(j)}{n}\chi^{|i-j|}\Big), \tag{6.31}$$

where

$$h_m(u) = (m+1)\mathbf{1}(1 \le u \le m-p) + u\mathbf{1}(m-p+1 \le u \le m+1) \ .$$

To see this, we consider exemplarily the case that $i, j \in [m-p+1, m+1]$ - all other cases are treated in the same way. Then

$$\tilde{\sigma}_{i,j} - \sigma_{i,j} = \mathbb{E}\big(G(\tfrac{m+1}{n}, \mathcal{F}_i)G(\tfrac{m+1}{n}, \mathcal{F}_j)\big) - \mathbb{E}\big(G(\tfrac{i}{n}, \mathcal{F}_i)G(\tfrac{j}{n}, \mathcal{F}_j)\big) = K_1 + K_2, \tag{6.32}$$

where K_1 and K_2 are defined by

$$K_1 = \mathbb{E}\Big(G(\tfrac{m+1}{n}, \mathcal{F}_i)\Big(G(\tfrac{m+1}{n}, \mathcal{F}_j) - G(\tfrac{j}{n}, \mathcal{F}_j)\Big)\Big)$$
$$K_2 = \mathbb{E}\Big(G(\tfrac{j}{n}, \mathcal{F}_i)\Big(G(\tfrac{m+1}{n}, \mathcal{F}_i) - G(\tfrac{i}{n}, \mathcal{F}_i)\Big)\Big)$$

For the investigation of K_1, we use the differentiability of the filter to obtain

$$\Big|\mathbb{E}\Big(G(\tfrac{m+1}{n}, \mathcal{F}_i)\big(G(\tfrac{m+1}{n}, \mathcal{F}_j) - G(\tfrac{j}{n}, \mathcal{F}_j)\big)\Big)\Big| \quad \le \int_{\frac{j}{n}}^{\frac{m+1}{n}} \Big|\mathbb{E}\Big(G(\tfrac{m+1}{n}, \mathcal{F}_i)\dot{G}(u, \mathcal{F}_j)\Big)\Big| du \quad (6.33)$$

Observing assumption (P2), (P3) and by the argument of proving (6.14), it follows

$$\big|\mathbb{E}\big(G(\frac{m+1}{n}, \mathcal{F}_i)\dot{G}(u, \mathcal{F}_j)\big)\big| = O(\chi^{|i-j|}) \quad (6.34)$$

(uniformly with respect to $u \in [0, 1]$). Combining the estimates (6.33) and (6.34) yields

$$|K_1| = O\Big(\frac{m+1-j}{n}\chi^{|i-j|}\Big).$$

Similarly it follows that $|K_2| = O(\frac{m+1-i}{n}\chi^{|i-j|})$. These bounds and (6.32) yield

$$\sup_{m-p\le i,j\le m+1} |\tilde{\sigma}_{i,j} - \sigma_{i,j}| = O\Big(\frac{2m+2-i-j}{n}\chi^{|i-j|}\Big),$$

which shows that (6.31) holds uniformly for $i, j \in [m-p, m+1]$. Similar and simpler arguments yield that (6.31) holds uniformly for the other choices of i, j.
Next, observe that Σ_m^{AR} is an $m \times m$ symmetric matrix, and so is $\Sigma_m^{AR} - \Sigma_{n,m}$. By similar arguments as given in the proof of Theorem 3.2, it follows that

$$\rho(\Sigma_m^{AR} - \Sigma_{n,m}) \le \max_{1\le i\le n} \sum_{j=1}^{n} |\sigma_{i,j} - \tilde{\sigma}_{i,j}| = O(\frac{p}{n}),$$

where the last inequality is a consequence from (6.31). This inequality and assumption (E1) imply that Σ_m^{AR} is positive definite if n is sufficiently large. Consequently,

$$\tilde{\mathbf{a}}_m = (\Sigma_m^{AR})^{-1}\boldsymbol{\gamma}_m^{AR}.$$

and by similar arguments as given in the proof of Corollary 4.1 we obtain that

$$\rho((\Sigma_m^{AR})^{-1} - \Sigma_{n,m}^{-1}) = O(\frac{p}{n}), \tag{6.35}$$

$$|\boldsymbol{\gamma}_m^{AR} - \boldsymbol{\gamma}_m| = O(\frac{p}{n}) \tag{6.36}$$

Now (6.30) follows from (6.19) (6.35), (6.36), which completes the proof. □

Acknowledgements This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Project A1, C1) of the German Research Foundation (DFG) and NSFC Young program (No.11901337).

References

Anderson, T. W. (2003). *Multivariate Statistical Analysis.* John Wiley & Sons, New York.

Bickel, P. J. and Gel, Y. R. (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):711–728.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control.* John Wiley & Sons.

Brockwell, P. J., Davis, R. A., and Calder, M. V. (2002). *Introduction to Time Series and Forecasting.* Springer.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25(1):1–37.

Das, S. and Politis, D. N. (2017). Predictive inference for locally stationary time series with an application to climate data. *arXiv preprint arXiv:1712.02383.*

Demko, S., Moss, W. F., and Smith, P. W. (1984). Decay rates for inverses of band matrices. *Mathematics of computation*, 43(168):491–499.

Dette, H. and Wu, W. (2019). Detecting relevant changes in the mean of nonstationary processes - a mass excess approach. *Annals of Statistics*, 47(6):3578–3608.

Dette, H., Wu, W., and Zhou, Z. (2019). Supplement for change point analysis of second order characteristics in non-stationary time series. *Statistica Sinica*, pages 611–643.

Ding, X. and Zhou, Z. (2018). Estimation and inference for precision matrices of non-stationary time series. *arXiv preprint arXiv:1803.01188*.

Elsner, J. B., Kossin, J. P., and Jagger, T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature*, 455(7209):92.

Fryzlewicz, P., Van Bellegem, S., and Von Sachs, R. (2003). Forecasting non-stationary time series by wavelet process modelling. *Annals of the Institute of Statistical Mathematics*, 55(4):737–764.

Giraud, C., Roueff, F., and Sanchez-Perez, A. (2015). Aggregation of predictors for non-stationary sub-linear processes and online adaptive forecasting of time varying autoregressive processes. *The Annals of Statistics*, 43(6):2412–2450.

Guillaumin, A. P., Sykulski, A. M., Olhede, S. C., Early, J. J., and Lilly, J. M. (2017). Analysis of non-stationary modulated time series with applications to oceanographic surface flow measurements. *Journal of Time Series Analysis*, 38(5):668–710.

Kley, T., Preuss, P., and Fryzlewicz, P. (2019). Predictive, finite-sample model choice for time series under stationarity and non-stationarity. *Electronic Journal of Statistics*, 13(2):3710–3774.

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, 31(6):471–482.

McMurry, T. L. and Politis, D. N. (2015). High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics*, 9(1):753–788.

Nason, G. P., Von Sachs, R., and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):271–292.

Priestley, M. B. (1988). Non-linear and non-stationary time series analysis. *London: Academic Press, 1988*.

Roueff, F. and Sanchez-Perez, A. (2018). Prediction of weakly locally stationary processes by auto-regression. *ALEA-Lat. Am. J. Probab. Math. Stat.*, 15:1215–1239.

Stărică, C. and Granger, C. (2005). Nonstationarities in stock returns. *Review of Economics and Statistics*, 87(3):503–522.

Van Bellegem, S. and Von Sachs, R. (2004). Forecasting economic time series with unconditional time-varying variance. *International Journal of Forecasting*, 20(4):611–627.

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

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154.

Wu, W. B. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, 19:1755–1768.

Zhang, T. and Wu, W. B. (2012). Inference of time-varying regression models. *The Annals of Statistics*, 40(3):1376–1402.

Zhao, Z. and Wu, W. B. (2008). Confidence bands in nonparametric time series regression. *The Annals of Statistics*, 36(4):1854–1878.

Zhou, Z. (2013). Inference for non-stationary time-series autoregression. *Journal of Time Series Analysis*, 34(4):508–516.

Zhou, Z. and Wu, W. B. (2009). Local linear quantile estimation for nonstationary time series. *The Annals of Statistics*, 37(5):2696–2729.

Zhou, Z. and Wu, W. B. (2010). Simultaneous inference of linear models with time varying coefficients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):513–531.