Time-varying autoregressive models forecasting

Amina Bouchafaa Maxime Godin

March 2018

Contents

Introduction

This project is about time series prediction. We are interested in a larger class of models than the usual weakly stationary framework. We study more specifically the case of the time-varying autoregressive model which is a generalisation of the autoregressive model. Actually, this model can be seen as a specialisation of the locally stationary assumption. Roughly, locally stationary processes can be approximated by weakly stationary processes when one looks on short sub-sequences of the series.

We study in particular three predictors: the first one, presented in [RS16] is based on the Yule-Walker equation and relies on covariance estimation. The second one is a recursive identification method which is described in [MPR05]. Finally, we take a glance at an exponential aggregation technique of the previous predictors based on [GRS15].

We start by recalling the stationary framework and then introduce the TVAR model. We show some stability results for TVAR and explain how they can be seen as locally stationary processes.

Then we introduce the concept of prediction for time series and explain some theoretical results concerning the three predictors in the case of TVAR processes.

Finally, we numerically illustrate these results. We developed a *Shiny* application in order to explore and assess interactively the predictors.

1. Time varying auto-regressive models

In this section, we aim to introduce a model for time series which goes beyond the usual stationary framework. We first introduce the stationarity assumption and its main properties. In this framework, we recall the classic autoregressive model. We then naturally extend the autoregressive model to a time-varying autoregressive model and state some stability conditions. We finally introduce the local stationarity property and show it encompasses the time-varying autoregressive model provided some smoothness conditions.

1.1 Introducing TVAR models

1.1.1 Weak stationarity and autoregressive processes

We now introduce random processes, weak stationarity and autoregressive processes. The definitions, the results and some remarks are taken from Chapters 2 and 3 of [Rou17].

We consider a probability space $(\Omega, \mathbb{F}, \mathbb{P})$, an index set T and a measurable space $(\mathbb{X}, \chi)$, called the observation space.

Definition 1.1.1. (Random process). A random process defined on $(\Omega, \mathbb{F}, \mathbb{P})$, indexed on T and valued in $(\mathbb{X}, \chi)$ is a collection $(X_t)_{t\in T}$ of random variables defined on $(\Omega, \mathbb{F}, \mathbb{P})$ and taking their values in $(\mathbb{X}, \chi)$.

The idea behind random processes is to enable complex dependence structures which go far beyond the independent and identically distributed assumption.

In what follows, we consider $T = \mathbb{N}$ or $\mathbb{Z}$ and $\mathbb{X} = \mathbb{R}$ or $\mathbb{C}$ embedded with their Euclidian structure and their Borel σ-field. This narrows the framework of this study to univariate random processes as referred to the dimension of the observation space. Such a random process is called a univariate time series. Time series are a general framework which can be used to model temporal data.

Definition 1.1.2. (Time series). A random process indexed on $T = \mathbb{N}$ or $\mathbb{Z}$ and with values in $\mathbb{X} = \mathbb{R}$ or $\mathbb{C}$ is called a univariate time series.

One very common assumption when one models time series is the weakly stationary assumption. Roughly, a weakly stationary process has its first and second moment constant overtime.

Definition 1.1.3. (Weakly stationary time series)

Let $\mu \in \mathbb{C}$ and $\gamma : \mathbb{Z} \longmapsto \mathbb{C}$. A process $(X_t)_{t\in\mathbb{Z}}$ with values in $\mathbb{C}$ is said to be weakly stationary with mean μ and autocovariance function γ if all the following assertions hold:

1. X is an L^2 process, i.e. $\mathbb{E}[|X_t|^2] < \infty$,
2. for all $t \in \mathbb{Z}$, $\mathbb{E}[X_t] = \mu$,
3. for all $(s,t) \in \mathbb{Z} \times \mathbb{Z}, \mathrm{Cov}(X_s, X_t) = \gamma(s-t)$.

Remark. There are some widely-used techniques to transform non-stationary random processes into stationary ones. We do not focus on them as this work aims to go beyond the weak stationarity assumption.

In what follows, we assume that the weakly stationary processes we consider are centered, that is $\mu = 0$. In a statistical context, this means that one must first estimate the mean of a weakly stationary process before going any further in modelling it.

Let us now state the first properties of weakly stationary processes.

Proposition 1.1.1. The autocovariance function $\gamma : \mathbb{Z} \longmapsto \mathbb{C}$ of a complex valued weakly stationary process satisfies the following properties:

1. Hermitian symmetry: for all $s \in Z, \gamma(-s) = \overline{\gamma(s)}$
2. Nonnegative definiteness: for all integer $n \geq 1$ and $a_1, \ldots, a_n \in \mathbb{C}$,

$$\sum_{s=1}^{n}\sum_{t=1}^{n} \overline{a_s}\gamma(s-t)a_t \geq 0$$

The autocovariance matrix Γ_n of n consecutive samples $X_1, \ldots, X_n$ of the time series has a particular structure, namely it is constant on its diagonals, $(\Gamma_n)_{i,j} = \gamma(i-j)$, $\Gamma_n = \mathrm{Cov}([X_1, \ldots, X_n]^T)$

$$\Gamma_n = \begin{pmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(1-n) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(2-n) \\ \vdots & & & \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{pmatrix}.$$

Let $\mathbb{T}$ denotes any interval congruent to $[0, 2\pi)$. We denote by $\mathcal{B}(\mathbb{T})$ the associated Borel σ-field. The Herglotz's theorem shows that the autocovariance function of a weakly stationary process X is entirely determined by a finite nonnegative measure on $(\mathbb{T}, \mathcal{B}(\mathbb{T}))$. This measure is called the spectral measure of X.

Theorem 1.1.1. (Herglotz). A sequence $(\gamma(h))_{h\in Z}$ is a nonnegative definite hermitian sequence in the sense of Proposition 1.1.1 if and only if there exists a finite nonnegative measure ν on$(\mathbb{T}, \mathcal{B}(\mathbb{T}))$ such that:

$$\gamma(h) = \int_{\mathbb{T}} e^{ih\lambda}\nu(\mathrm{d}\lambda)$$

This relation defines ν uniquely.

Remark. By Proposition 1.1.1, Theorem 1.1.1 applies to all γ which is an autocovariance function of a weakly stationary process X. In this case ν is called the spectral measure of X. If ν admits a density f, it is called the spectral density function. The Herglotz's theorem shows that the spectral measure completely characterises the autocovariance function of a weakly stationary processs. This representation may be more convenient in some applications and further extensions to locally stationary processes.

Definition 1.1.4. (Innovation process). Let $(X_t)_{t\in\mathbb{Z}}$ be a centered weakly stationary process. We call innovation process the process $(\epsilon_t)_{t\in Z}$ defined by:

$$\epsilon_t := X_t - \text{proj}(X_t|\mathcal{H}_{t-1}^X)$$

where $\mathcal{H}_t^X := \overline{\text{Span}}(X_s, s \leq t)$ and proj denotes the L^2 orthogonal projection.

We now introduce the auto-regressive model in the stationary context.

Definition 1.1.5. (Autoregressive process). An autoregressive process of order p (AR(p)) with coefficients $\theta_1, \cdots, \theta_p$ is a random process $X = (X_t)_{t\in\mathbb{Z}}$ that satisfies the AR(p) equation which is:

$$X_t = \sum_{k=1}^{p} \theta_k X_{t-k} + \epsilon_t \tag{1.1}$$

where ϵ_t is a centered weak noise with variance σ^2.

The process dependence in the p past values creates a regression model where the regression coefficients are given by $(\theta_k)_{k=1\cdots p}$. This justifies the term auto-regressive. This dependence makes the modelling coherent and interesting to study.
Let us note that in the case of AR(p) the regression coefficients $(\theta_k)_{k=1\cdots p}$ do not depend on time. The AR(p) equation holds for all $t \in \mathbb{Z}$.

Theorem 1.1.2. (Existence and uniqueness of a weakly stationary solution of the AR(p) equation). For ϵ a centered weak noise with variance $\sigma^2 > 0$, and $\theta_1 \cdots \theta_p \in \mathbb{C}$, we set Θ to be the polynomial:

$$\Theta(z) := 1 - \sum_{k=1}^{p} \theta_k z^k \tag{1.2}$$

Then, the AR(p) equation admits a unique weakly stationary solution if and only if Θ has no roots on the unit circle Γ_1.
Furthermore, the unique solution X is given by $X = \sum_{k\in\mathbb{Z}} \phi_k \epsilon^k$ where $\phi \in l^1$ is uniquely defined by the equation:

$$\sum_{k\in\mathbb{Z}} \phi_k z^k = \frac{1}{\Theta(z)}, \quad \forall z \in \Gamma_1.$$

Remark. Let us point out that the $(\epsilon_t)_{t\in Z}$ in equation (1.1) are not always the innovations of the unique stationary solution of the AR(p) equation. This is only the case when the polynomial Θ defined by equation (1.2) has no roots in the closed unit disk (see [Rou17], Theorem 3.4.1).

1.1.2 TVAR equation

We are interested in a specific extension of the AR processes where the regression coefficients do depend on time. They are called the time-varying autoregressive (TVAR) processes. Let us now introduce the time-varying autoregressive equation which describes the model we will consider for prediction.

Definition 1.1.6. (TVAR equation). Let $(X_{k,n})_{0\leq k\leq n}$ be the doubly indexed random process with values in $\mathbb{R}$ uniquely defined by the equation:

$$X_{k+1,n} = \boldsymbol{\theta}\left(\frac{k}{n}\right) \cdot \boldsymbol{X}_{k,n} + \sigma\left(\frac{k+1}{n}\right) \epsilon_{k+1,n} \tag{1.3}$$

for $0 \leq k < n$ where $\boldsymbol{X}_{k,n} = (X_{k,n}, \ldots, X_{k-d+1,n})^T$ with the appropriate padding convention when $k \leq d$ and:

- $d \geq 1$ is the order of the equation,
- $(\boldsymbol{X}_{0,n})_{n\geq 0}$ is a collection of random variables in $\mathbb{R}^d$ called the initial conditions,
- $(\epsilon_{k,n})_{1\leq k\leq n}$ is a triangular array of real-valued random variables referred to as the (normalised) innovations,
- $\boldsymbol{\theta}(t) := [\theta_1(t), \ldots, \theta_d(t)]^T, t \in [0,1]$, is a d-dimensional vector referred to as the local autoregression vector,
- $\sigma(t)$, $t \in [0,1]$, is a nonnegative number referred to as the local innovation standard deviation.

With a fixed $n \geq 1$, the time series $(X_{k,n})_{1\leq k\leq n}$ is referred to as a TVAR time series.

We can interpret equation (1.3) with a fixed n as an autoregression with coefficients varying over time.

Remark. The fact that the random process is doubly indexed is useful to develop an asymptotic estimation theory. Indeed, in a stationary context, the asymptotic estimation theory is built on an increasing number of observations of the time series $(X_k)_{k\in\mathbb{N}}$. This cannot be the case for a model based on the time varying autoregressive equation as the coefficients are varying over time, as we can see in Figure 1.1. Observing data for long times cannot obviously be helpful to improve estimation at the beginning of the process. Instead, the asymptotic theory is based on a increasing number of observations in a given time interval, here $[0,1]$. This means that when n tends to $+\infty$, we increase the sampling frequency of the data. As we will elaborate in section 1.2, n can be seen as the sharpness of the local approximation of the TVAR time series $(X_{k,n})_{1\leq k\leq n}$ with n fixed by a weakly stationary process.

Figure 1.2 represents a sample of a TVAR process of length 2^{10}. It follows a 1-order TVAR equation where the only component of the local autoregression vector is represented on Figure 1.1. This process has normal innovations, a constant standard deviation equal to 1 and normal initial conditions.

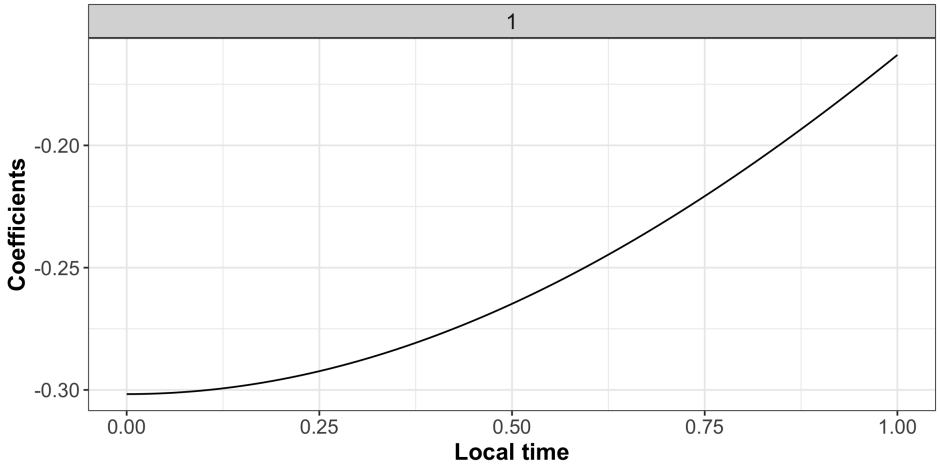

Figure 1.1: Example of TVAR(1) coefficients over time

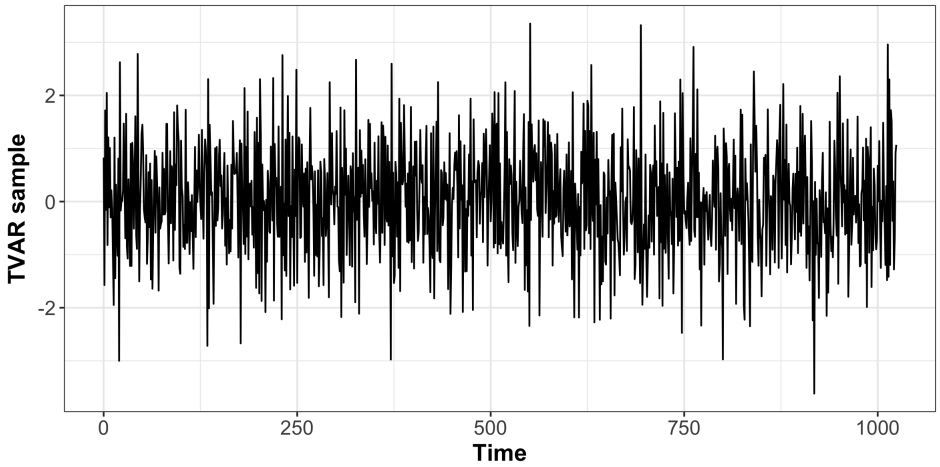

Figure 1.2: TVAR(1) sample of length 2^{10} with normal innovations

1.1.3 Stability conditions

The recurrence equation 1.3 which rules a TVAR process has a nice interpretation. However, when one uses auto-regression to model a real phenomenon, they may wonder whether the model will remain physically realistic in the sense that it does not grow indefinitely. This is a question of stability of the model. Stability conditions for stationary autoregressive processes are well-known: they require that the polynomial associated to the recurrence equation does not vanish on the unit disk . They can be extended in a non-stationary context under further assumptions concerning the roots of the local polynomial and some smoothness conditions on the coefficients.

Let us first introduce some definitions and notations :

For $\rho > 0$, we denote :

$$S(\rho) := \{\boldsymbol{\theta} \colon [0,1] \mapsto \mathbb{R}^d, 1 - \sum_{j=1}^{d} \boldsymbol{\theta}_j(t) z^j \neq 0, \forall |z| < \rho^{-1} \text{ and } t \in [0,1]\} \tag{1.4}$$

Definition 1.1.7. (β-Lipschitz ball). For any $\beta \in (0,1]$, denote the β-Lipschitz semi-norm of a mapping $\boldsymbol{f} : [0,1] \mapsto \mathbb{R}^d$ by :

$$|\boldsymbol{f}|_{\Lambda,\beta} := \sup_{t \neq s} \frac{|\boldsymbol{f}(t) - \boldsymbol{f}(s)|}{|t-s|^\beta}.$$

Define for $0 < L < \infty$ the β-Lipschitz ball :

$$\Lambda_d(\beta, L) := \left\{ \boldsymbol{f} \colon [0,1] \mapsto \mathbb{R}^d, |\boldsymbol{f}|_{\Lambda,\beta} \leq L, \sup_{t \in [0,1]} |\boldsymbol{f}(t)| \leq L \right\}.$$

For all $d \in \mathbb{N}^*$, $\beta > 1$, let $k \in \mathbb{N}$ and $\alpha \in (0,1]$ be uniquely defined by $\beta = k + \alpha$. Then we define:

$$\Lambda_d(\beta, L) := \left\{ \boldsymbol{f} \colon [0,1] \mapsto \mathbb{R}^d, |\boldsymbol{f}^{(k)}|_{\Lambda,\alpha} \leq L, \sup_{t \in [0,1]} |\boldsymbol{f}(t)| \leq L \right\}$$

where $\boldsymbol{f}^{(k)}$ denotes the derivative of order k and $|\boldsymbol{f}^{(k)}|_{\Lambda,\alpha} = \infty$ if $\boldsymbol{f}$ is not k times differentiable.

For all $d \in \mathbb{N}^*$, $\beta > 0$, $L > 0$, $0 < \rho < 1$ and $0 < \sigma_- \leq \sigma_+$, we define:

$$\mathcal{C}_d^+(\beta, L, \rho, \sigma_-, \sigma_+) := \{(\boldsymbol{\theta}, \sigma) \colon \Lambda_d(\beta, L) \cap S(\rho), \sigma \colon [0,1] \mapsto [\sigma_-, \sigma_+]\} \tag{1.5}$$

Let us now state a L^q uniform boundedness result:

Theorem 1.1.3. (L^q uniform boundedness stability conditions). Let $d \in \mathbb{N}^*$, $\beta \in (0,1]$, $L > 0$, $0 < \rho < \tau < 1$, $0 < \sigma_- \leq \sigma_+$ and $\mathcal{C} = \mathcal{C}_d^+(\beta, L, \rho, \sigma_-, \sigma_+)$.

Suppose the random variables $\{\epsilon_{k,n}\}$ are independent, have zero mean and unit variance and are independent of the initial conditions $\mathbf{X}_{0,n}$.

Further suppose that $\sup_{n \geq 0} ||\mathbf{X}_{0,n}||_q < \infty$ and $\epsilon_q^* = \sup_{1 \leq k \leq n} ||\epsilon_{k,n}||_q < \infty$ for some $q \geq 2$.

Then we have the following L^q uniform boundedness result:

$$\sup_{(\boldsymbol{\theta},\sigma) \in \mathcal{C}} \sup_{0 \leq k \leq n} ||\mathbf{X}_{k,n}||_{q,\boldsymbol{\theta},\sigma} < +\infty$$

where $(X_{k,n})_{0 \leq k \leq n}$ denotes the TVAR process of order d defined by the local autoregression vector $\boldsymbol{\theta}$, the local innovation standard deviation σ, the innovations $\{\epsilon_{k,n}\}_{0 \leq k \leq n}$ and the initial conditions $\{\mathbf{X}_{0,n}\} \in \mathbb{R}^d$.

Proof. Let first rewrite the recurrence equation with matrices:

$$\mathbf{X}_{k+1,n} = \Theta(l/n, \boldsymbol{\theta}) \mathbf{X}_{k,n} + \boldsymbol{\sigma}(k+1/n) \epsilon_{k+1,n}$$

where

$$\boldsymbol{X}_{k,n} = [X_{k,n}, X_{k-1,n} \cdots, X_{k-d+1,n}]^T$$
$$\boldsymbol{\sigma}(k/n) = [\sigma(k/n), 0, \cdots, 0]^T$$

and $\Theta(u, \boldsymbol{\theta})$ for $u \in [0,1]$ is the local companion matrix of the recurrence equation defined by:

$$\Theta(u, \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\theta}_1(u) & \boldsymbol{\theta}_2(u) & \cdots & \boldsymbol{\theta}_d(u) \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}.$$

We therefore can rewrite the process as a function of the initial conditions, the innovations, the local companion matrix and the local innovation standard deviation:

$$\boldsymbol{X}_{k,n} = \prod_{l=1}^{k} \Theta(l/n, \boldsymbol{\theta}) \boldsymbol{X}_{0,n} + \sum_{l=1}^{k} \prod_{j=l+1}^{k} \Theta(j/n, \boldsymbol{\theta}) \boldsymbol{\sigma}(l/n) \epsilon_{l,n} \tag{1.6}$$

We now recall a useful lemma to control the operator norm of companion matrices products derived from Proposition 13 of [MPR05].

Lemma 1.1.1. Let $\beta \in (0,1]$, $L > 0$, $0 < \rho < \tau < 1$. Then there exists a constant $M > 0$ such that, for all $\boldsymbol{\theta} \in \Lambda_d(\beta, L) \cap S(\rho)$ and $0 \leq k \leq k + m \leq n$ we have:

$$\left| \prod_{l=k+1}^{k+m} \Theta(l/n, \boldsymbol{\theta}) \right| \leq M \tau^m.$$

where $|\cdot|$ denotes the operator norm associated with the Euclidian norm.

Using the Minkowski's inequality, we can control $||\mathbf{X}_{k,n}||_{q,\boldsymbol{\theta},\sigma}$:

$$||\mathbf{X}_{k,n}||_{q,\boldsymbol{\theta},\sigma} \leq || \prod_{l=1}^{k} \Theta(l/n, \boldsymbol{\theta}) \boldsymbol{X}_{0,n}||_{q,\boldsymbol{\theta},\sigma} + \sum_{l=1}^{k} || \prod_{j=l+1}^{k} \Theta(j/n, \boldsymbol{\theta}) \boldsymbol{\sigma}(l/n) \epsilon_{l,n}||_{q,\boldsymbol{\theta},\sigma}$$

Applying Lemma 1.1.1 we further have:

$$| \prod_{l=1}^{k} \Theta(l/n, \boldsymbol{\theta}) \boldsymbol{X}_{0,n}| \leq M \tau^k |\boldsymbol{X}_{0,n}| \leq M |\boldsymbol{X}_{0,n}|$$
$$| \prod_{j=l+1}^{k} \Theta(j/n, \boldsymbol{\theta}) \boldsymbol{\sigma}(l/n) \epsilon_{l,n}| \leq M \tau^{k-l} \sigma_+ |\epsilon_{l,n}|$$

as $\tau \leq 1$ and $\sigma \leq \sigma_+$.

Raising to power q, taking the expectation and raising to power $1/q$, we then have:

$$|| \prod_{l=1}^{k} \Theta(l/n, \boldsymbol{\theta}) \boldsymbol{X}_{0,n}||_{q,\boldsymbol{\theta},\sigma} \leq M ||\boldsymbol{X}_{0,n}||_{q,\boldsymbol{\theta},\sigma} \leq M \sup_{n \leq 0} ||\boldsymbol{X}_{0,n}||_{q,\boldsymbol{\theta},\sigma}$$

$$\sum_{l=1}^{k} || \prod_{j=l+1}^{k} \Theta(j/n, \boldsymbol{\theta})\boldsymbol{\sigma}(l/n)\epsilon_{l,n}||_{q,\boldsymbol{\theta},\sigma} \leq \sum_{l=1}^{k} M\tau^{k-l}\sigma_+\epsilon_q^* \leq M\sigma_+\epsilon_q^*$$

given the boundedness assumptions $\sup_{n\geq 0} ||\mathbf{X}_{0,n}||_q < \infty$ and $\epsilon_q^* = \sup_{1\leq k\leq n} ||\epsilon_{k,n}||_q < \infty$.

Hence:

$$||\mathbf{X}_{k,n}||_{q,\boldsymbol{\theta},\sigma} \leq M\left(\sup_{n\geq 0}||\boldsymbol{X}_{0,n}||_{q,\boldsymbol{\theta},\sigma} + \sigma_+\epsilon_q^*\right).$$

□

Remark. Let us recall from Theorem 1.1.2 that a stationary solution of the AR(p) equation exists as long as the polynomial associated to the recurrence equation does not vanish on the unit circle. Furthermore, the AR(p) equation is the canonical representation of the process (that is the $\{\epsilon_{k,n}\}$ are the innovation process) if and only if the polynomial does not vanish on the closed unit disk (see Theorem 3.4.1 from [Rou17]). Remark that the condition $\boldsymbol{\theta} \in \Lambda_d(\beta, L) \cap S(\rho)$ is stronger than this usual condition in the sense that we require the roots of all the local polynomials to stay away from the unit disk.

1.2 From weak stationarity to local stationarity

1.2.1 Local stationarity

The following definitions are adapted from [RS16] to match our setting.

We consider a doubly indexed time series $(X_{k,n})_{0\leq k\leq n}$. Here k refers to a discrete time sample index and n is an additional index indicating the sharpness of the local approximation of the time series $(X_{k,n})_{1\leq k\leq n}$ with a fixed n by a stationary one. $(X_{k,n})_{0\leq k\leq n}$ is considered to be locally stationary if, for n large, given a set S_n of sample indices such that $\frac{k}{n} \approx u$ for $k \in S_n$, the sample $(X_{k,n})_{k\in S_n}$ can be somehow be viewed as the sample of a weakly stationary process with a re-scaled location u.

Definition 1.2.1. (Time varying covariance function). Let $(X_{k,n})_{0\leq k\leq n}$ be a triangular array of random variables with finite variances. The local time varying covariance function γ^* is defined for all $n \in \mathbb{N}^*$, $1 \leq k \leq n$ and $k - n \leq l \leq k - 1$ by:

$$\gamma^*(k, n, l) = \mathrm{Cov}(X_{k,n}, X_{k-l,n})$$

Definition 1.2.2. (Local covariance function and local spectral density). A local spectral density f is a $[0, 1] \times \mathbb{R} \mapsto \mathbb{R}^+$ function, (2π)-periodic and locally integrable with respect to the second variable. The local covariance function γ associated with the local spectral density f is defined on $[0, 1] \times \mathbb{Z}$ by:

$$\gamma(u, l) = \int_{-\pi}^{\pi} e^{il\lambda} f(u, \lambda)\mathrm{d}\lambda$$

Definition 1.2.3. (Weakly locally stationary processes). Let $(X_{k,n})_{0\leq k\leq n, n\geq n_0}$ be an array of random variables with finite variances, $n_0 \geq 1$, $\beta, R > 0$ and $(k_n^0)_{n\geq n_0}$ such that for all n, $k_n^0 \geq 1$ and $\frac{k_n^0}{n}$ tends to zero. We say that $(X_{k,n})_{1\leq k\leq n, n\geq n_0}$ is (β, R)-weakly locally stationary with local spectral density f if:

1. for all $\lambda \in \mathbb{R}$, we have $f(\cdot, \lambda) \in \Lambda_1(\beta, R)$,
2. the time varying covariance function γ^* of $(X_{k,n})_{1\leq k\leq n, n\geq n_0}$ and the local covariance function γ associated with f satisfy, for all $n \geq n_0$, for all $k \geq k_n^0$ and for all $k - n \leq l \leq k - 1$:

$$\left|\gamma^*(k, n, l) - \gamma\left(\frac{k}{n}, l\right)\right| \leq Rn^{-\min(1,\beta)}.$$

Remark. The previous definition definition can be interpreted as follows: a part from a fraction of the sample which is going to zero as n goes to $+\infty$, the time-varying covariance function can be approximated by the autocovariance function of a stationary process with an appropriate rescaling of the time index.

1.2.2 TVAR models as locally stationary processes

We now state a result showing that under some assumptions, a TVAR process is a locally stationary process.

Theorem 1.2.1. Let $(X_{k,n})_{0\leq k\leq n}$ be a TVAR process of order d with initial conditions $(\boldsymbol{X}_{0,n})_{n\in\mathbb{N}}$, innovations $(\epsilon_{k,n})_{1\leq k\leq n}$, local autoregression vector $\boldsymbol{\theta}$ and local standard deviation σ.

Assume there exist $\beta \in (0, 1]$, $L > 0$, $0 < \rho < \tau < 1$, $0 < \sigma_- \leq \sigma_+$ such that $(\boldsymbol{\theta}, \sigma) \in \mathcal{C} = \mathcal{C}_d^+(\beta, L, \rho, \sigma_-, \sigma_+)$.

Suppose also that the random variables $\{\epsilon_{k,n}\}$ are independent, have zero mean and unit variance and are independent of the initial conditions $\mathbf{X}_{0,n}$.

Further suppose that the initial conditions $\mathbf{X}_{0,n}$ are centered and that $\sup_{n\geq 0} ||\mathbf{X}_{0,n}||_q < \infty$ and $\epsilon_q^* = \sup_{1\leq k\leq n} ||\epsilon_{k,n}||_q < \infty$ for $q = 2$.

Then there exists $n_0 \in \mathbb{N}^*$, $(k_n^0)_{n\geq n_0}$, $R > 0$ such that $(X_{k,n})_{0\leq k\leq n, n\geq n_0}$ is (β, R)-weakly locally stationary with local spectral density f defined by:

$$f(\lambda, u) = \frac{\sigma(u)^2}{2\pi}|\theta(e^{i\lambda}, u)|^{-2} \tag{1.7}$$

for $(\lambda, u) \in \mathbb{R}^2$ where $\theta(\cdot, u)$ is the local polynomial defined by:

$$\theta(z, u) = 1 - \sum_{l=1}^{d} \boldsymbol{\theta}_l(u) z^l$$

for $z \in \mathbb{C}$.

Proof. First remark that by Theorem 1.1.3, $(X_{k,n})_{0\leq k\leq n}$ is well-defined and as finite variances.

Let us now state a useful lemma to control $|\theta(e^{i\lambda}, u)|$.

Lemma 1.2.1. Let $\boldsymbol{\theta} \in S(\rho)$ for some $\rho < 1$. Then we have for all $u \in [0,1]$, for all $z \in \mathbb{C}$ with $|z| = 1$:

$$|\theta(z,u)| \geq (\rho^{-1} - 1)^d.$$

Proof. First remark that if $\boldsymbol{\theta} \in S(\rho)$, then all the roots of all the polynomials $\theta(\cdot, u)$, $u \in [0,1]$ are outside the closed disk of radius $\rho^{-1} > 1$. We then rewrite for a fixed $u \in [0,1]$:

$$\theta(z,u) = \prod_{l=1}^{d} (z - \nu_l(u)).$$

where $\nu_l(u)$ are the roots of the polynomial $\theta(\cdot, u)$. Using that $|z - \nu_l(u)| \geq \rho^{-1} - 1$, the proof follows. □

Now we show that there exists R_1 such that for all $\lambda \in \mathbb{R}$, we have $f(\cdot, \lambda) \in \Lambda_1(\beta, R_1)$ where f is defined by equation 1.7.

Applying the previous lemma, f is obviously bounded by $\frac{\sigma_+^2}{2\pi}(\rho^{-1} - 1)^{-2d}$.

Fixing λ and let s and t be in $[0,1]$, we have:

$$\frac{|f(s,\lambda) - f(t,\lambda)|}{|t-s|^\beta} \leq \frac{1}{2\pi |t-s|^\beta} |\sigma(t)^2 |\theta(e^{i\lambda}, t)|^{-2} - \sigma(t)^2 |\theta(e^{i\lambda}, s)|^{-2} + \sigma(t)^2 |\theta(e^{i\lambda}, s)|^{-2} - \sigma(s)^2 |\theta(e^{i\lambda}, s)|^{-2}|$$

We will now control the two terms separately.

Applying the previous lemma, using $\sigma \in \Lambda_1(\beta, L)$ and given that $x \colon \mapsto x^2$ is $2\sigma_+$-Lipschitz on $[\sigma_-, \sigma_+]$:

$$\frac{\left|\sigma(t)^2 |\theta(e^{i\lambda}, s)|^{-2} - \sigma(s)^2 |\theta(e^{i\lambda}, s)|^{-2}\right|}{|t-s|^\beta} \leq (\rho^{-1} - 1)^{-2d} \frac{|\sigma(t)^2 - \sigma^2(s)|}{|t-s|^\beta} \leq 2(\rho^{-1} - 1)^{-2d} \sigma_+ \frac{|\sigma(t) - \sigma(s)|}{|t-s|^\beta} \leq 2(\rho^{-1} - 1)^{-2d} \sigma_+ L$$

Applying the previous lemma, using $\sigma < \sigma_+$, $\theta \in \Lambda_d(\beta, L)$ and given that $z \mapsto |z|^{-2}$ is C-Lipschitz for some $C > 0$ on $\{z \in \mathbb{C}, |z| \geq (\rho^{-1} - 1)^d\}$, there exists $C' > 0$ such that:

$$\frac{\left|\sigma(t)^2 |\theta(e^{i\lambda}, t)|^{-2} - \sigma(t)^2 |\theta(e^{i\lambda}, s)|^{-2}\right|}{|t-s|^\beta} \leq \sigma_+^2 C \frac{\left|\theta(e^{i\lambda}, t) - \theta(e^{i\lambda}, s)\right|}{|t-s|^\beta} \leq \sigma_+^2 C \frac{\sum_{l=1}^{d} \left|e^{il\lambda}\right| |\boldsymbol{\theta}_l(t) - \boldsymbol{\theta}_l(s)|}{|t-s|^\beta} \leq \sigma_+^2 C' L$$

By picking the correct R_1', we then have that for all $\lambda \in \mathbb{R}$ $f(\cdot, \lambda)$ is in $\Lambda_1(\beta, R_1')$.

Concerning the inequality:

$$\left|\gamma^*(k,n,l) - \gamma\left(\frac{k}{n}, l\right)\right| \leq Rn^{-\beta}$$

for all $n \geq 1$ and for all $k - n \leq l \leq k - 1$, we will not go into details.

Fix n, k and l and consider $u = \frac{k}{n}$.

We will focus on the case $l \geq 0$, the demonstration can be adapted to the case $l < 0$.

Let us compute the cross-covariance matrix of $\boldsymbol{X}_{k,n}$ and $\boldsymbol{X}_{k-l,n}$, that is $\mathbb{E}\left[\boldsymbol{X}_{k,n}\boldsymbol{X}_{k-l,n}^T\right]$ using representation (1.6).

We easily get that :

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{X}_{k,n}\boldsymbol{X}_{k-l,n}^T\right] = &\left(\prod_{j=1}^{k}\Theta(j/n,\boldsymbol{\theta})\right)\mathrm{Cov}(\boldsymbol{X}_{0,n},\boldsymbol{X}_{0,n})\left(\prod_{j=1}^{k-l}\Theta(j/n,\boldsymbol{\theta})\right)^T \\ &+\sum_{j=1}^{k-l}\left(\prod_{r=j+1}^{k}\Theta(r/n,\boldsymbol{\theta})\right)\left(\boldsymbol{\sigma}(k/n)\boldsymbol{\sigma}(k/n)^T\right)\left(\prod_{r=j+1}^{k-l}\Theta(r/n,\boldsymbol{\theta})\right)^T.\end{aligned}$$

Remark that as for $X \in \mathbb{R}^d$, $|XX^T| = |X|^2$ and $\sup_{n\geq 0}||\mathbf{X}_{0,n}||_q < \infty$, there exists K^0_{Cov} such that :

$$|\mathrm{Cov}(\boldsymbol{X}_{0,n},\boldsymbol{X}_{0,n})| \leq K^0_{\mathrm{Cov}}.$$

Now consider a stationary AR(d) process $(Y_p)_{p\in mathbbZ}$ with auto-regression coefficients $\boldsymbol{\theta}(u)$ and standard deviation $\sigma(u)$ whose spectral density is exactly $f(\cdot,u)$. Let us not $\boldsymbol{Y}_p = (Y_p,\ldots,Y_{p-d+1})$.

Analogously, we derive the cross-covariance of this process :

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{Y}_k\boldsymbol{Y}_{k-l}^T\right] = &\left(\Theta(u,\boldsymbol{\theta})^k\right)\mathrm{Cov}(\boldsymbol{Y}_0,\boldsymbol{Y}_0)\left(\Theta(u,\boldsymbol{\theta})^{k-l}\right)^T \\ &+\sum_{j=1}^{k-l}\left(\Theta(u),\boldsymbol{\theta})^{k-j}\right)\left(\boldsymbol{\sigma}(u)\boldsymbol{\sigma}(u)^T\right)\left(\Theta(u,\boldsymbol{\theta})^{k-j-l}\right)^T.\end{aligned}$$

Remark that by the Herglotz's theorem, the boundedness assumption on σ and Lemma 1.2.1, there exists $K_0^{stat} > 0$ independent of u such that:

$$|\mathrm{Cov}(\boldsymbol{Y}_0,\boldsymbol{Y}_0)| \leq K^{stat}_{\mathrm{Cov}}.$$

Let us now look split in three terms the operator norm of the difference of these cross-covariance matrices:

$$\begin{aligned}\left|\mathbb{E}\left[\boldsymbol{X}_{k,n}\boldsymbol{X}_{k-l,n}^T\right]-\mathbb{E}\left[\boldsymbol{Y}_k\boldsymbol{Y}_{k-l}^T\right]\right| \leq & \left|\left(\prod_{j=1}^{k}\Theta(j/n,\boldsymbol{\theta})\right)\mathrm{Cov}(\boldsymbol{X}_{0,n},\boldsymbol{X}_{0,n})\left(\prod_{j=1}^{k-l}\Theta(j/n,\boldsymbol{\theta})\right)^T\right| \\ & + \left|\left(\Theta(u,\boldsymbol{\theta})^k\right)\mathrm{Cov}(\boldsymbol{Y}_0,\boldsymbol{Y}_0)\left(\Theta(u,\boldsymbol{\theta})^{k-l}\right)^T\right| \\ & + \left|\sum_{j=1}^{k-l}\left(\prod_{r=j+1}^{k}\Theta(r/n,\boldsymbol{\theta})\right)\left(\boldsymbol{\sigma}(k/n)\boldsymbol{\sigma}(k/n)^T\right)\right. \\ & \qquad \left.\times\left(\prod_{r=j+1}^{k-l}\Theta(r/n,\boldsymbol{\theta})\right)^T\right. \\ & - \left.\left(\Theta(u),\boldsymbol{\theta})^{k-j}\right)\left(\boldsymbol{\sigma}(u)\boldsymbol{\sigma}(u)^T\right)\left(\Theta(u,\boldsymbol{\theta})^{k-j-l}\right)^T\right|.\end{aligned}$$

We will first show that provided that $k \geq k_n^0 := \lceil \frac{\beta\log(n)}{\log(\tau^{-1})} \rceil$, the two first terms, which are due to the inclusion of the initial conditions in the model, are decaying faster than $n^{-\beta}$. Note that we have $\frac{k_n^0}{n}$ going to zero with n. Actually, it is a direct consequence of Lemma 1.1.1 and of the previous bounds on the covariance matrices of the initial conditions.

Let us recall from equation (81) in [MPR05] that given any matrices $A_1,\ldots,A_r$ and $B_1,\ldots,B_r$ with compatibles sizes, we have:

$$\prod_{i=1}^{r}A_i-\prod_{i=1}^{r}B_i=\sum_{j=1}^{r}\left(\prod_{k=1}^{j-1}A_k\right)(A_j-B_j)\left(\prod_{k=j+1}^{r}B_k\right) \tag{1.8}$$

Applying this identity and Lemma 1.1.1, we have that there exists M' such that:

$$\begin{aligned}&\left|\left(\prod_{r=j+1}^{k}\Theta(r/n,\boldsymbol{\theta})\right)\left(\boldsymbol{\sigma}(k/n)\boldsymbol{\sigma}(k/n)^T\right)\left(\prod_{r=j+1}^{k-l}\Theta(r/n,\boldsymbol{\theta})\right)^T-\right. \\ &\left.\left(\Theta(u),\boldsymbol{\theta})^{k-j}\right)\left(\boldsymbol{\sigma}(u)\boldsymbol{\sigma}(u)^T\right)\left(\Theta(u,\boldsymbol{\theta})^{k-j-l}\right)^T\right| \leq \\ &\sum_{r=j+1}^{k}\sigma_+^2M'\tau^{k-j-1+k-j-l}|\Theta(r/n,\boldsymbol{\theta})-\Theta(u,\boldsymbol{\theta})|+ \\ &M'\tau^{k-j+k-j-l}|\boldsymbol{\sigma}(k/n)\boldsymbol{\sigma}(k/n)^T-\boldsymbol{\sigma}(u)\boldsymbol{\sigma}(u)^T|+ \\ &\sum_{r=j+1}^{k-l}\sigma_+^2M'\tau^{k-j+k-j-l-1}|\Theta(r/n,\boldsymbol{\theta})^T-\Theta(u,\boldsymbol{\theta})^T|.\end{aligned}$$

Putting this with the assumption $(\boldsymbol{\theta},\sigma)\in\mathcal{C}$, we finally have that there exists some $R_2>0$ such that:

$$\left|\gamma^*(k,n,l)-\gamma\left(\frac{k}{n},l\right)\right| \leq R_2n^{-\beta}.$$

□

Remark. Note that the above conditions imply that the local spectral density is uniformly bounded from below by some $f_- > 0$. This remark implies that under the assumptions of the previous theorem, assumption (M-1) in [RS16] holds.

2. Prediction for TVAR models

In this section, we first introduce the notion of prediction for time series using conditional expectation. We also explain the idea of linear prediction in the stationary framework and then extend it in the case of TVAR processes. We then introduce three ways to build predictors for TVAR processes and discuss some of their theoretical properties.

2.1 What do we mean by time series prediction ?

Let $(X_t)_{t\in T}$ be a centered L^2 process. The main idea of time series prediction is to find ways to approximate the best predictor, that is the conditional expectation of X_t knowing its past. In the L^2 case, this conditional expectation resumes to computing the projection of X_t on the vector subspace of L^2 associated to the σ-field spanned by $\{X_h, h \in T, h < t\}$. As this projection cannot be computed in general, we have to find ways to approximate it.

A solution to this issue can be to compute instead the linear projection of X_t on its past, as we introduced in Definition 1.1.4. More realistically, we can limit this projection to a recent past. We clarify this statement in the weakly stationary case with the following definition derived from 2.5.2 in [Rou17].

Definition 2.1.1. (Prediction coefficients and partical innovation process). Let $(X_t)_{t\in T}$ be a centered weakly stationary process. We call predictors of order p the random variables $\text{proj}(X_t|\mathcal{H}^X_{t-1,p})$ and the partial innovation process of order p the process $(\epsilon^+_{t,p})_{t\in\mathbb{Z}}$ defined by:

$$\epsilon^+_{t,p} := X_t - \text{proj}(X|\mathcal{H}^X_{t-1,p})$$

where $\mathcal{H}^X_{t,p} := \text{Span}(X_s, t-p \leq s \leq t)$.

The prediction coefficients are any coefficients $(\phi^+_{k,p})_{k=1,\cdots,p}$ which satisfy, for all $t \in \mathbb{Z}$,

$$\text{proj}(X_t|\mathcal{H}^X_{t-1,p}) = \sum_{k=1}^{p} \phi^+_{k,p} X_{t-k}$$

which is equivalent to

$$\Gamma^+_p \phi^+_p = \gamma^+_p \tag{2.1}$$

where $\gamma^+_p = [\gamma(1), \gamma(2), \ldots, \gamma(p)]^T$ and

$$\Gamma^+_p = \text{Cov}([X_{t-1} \cdots X_{t-p}]^T)^T$$

$$= \begin{pmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(1-p) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(2-p) \\ \vdots & & & \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{pmatrix}$$

The prediction error is equal to:

$$\sigma_p^2 = \gamma(0) - (\phi_p^+)^H \gamma_p^+, \tag{2.2}$$

Remark. Equations (2.1) and (2.2) are called the Yule-Walker equations. They can be obtained by writing the orthogonality conditions which define the L^2 projection. We will see again equation (2.1) in the locally stationary framework.

Remark. In the case of the stationary AR(d) processes introduced in Definition 1.1.5, the predictors of order p are equal to the best predictor, that is the conditional expectation, as soon as $p \geq d$. This means that the error due to the approximation of the conditional expectation by a linear projection vanishes. This remark motivates the use of linear prediction for time series in more general settings. We will see is the following section whether this approximation error vanishes for TVAR processes.

2.2 Linear prediction for TVAR processes

As we noted in the previous remark, when we use linear prediction for AR(d), there is no approximation error. In fact, this result is also true for TVAR processes.

Proposition 2.2.1. (Linear prediction). Let $(X_{k,n})_{0 \leq k \leq n}$ be a centered TVAR process of order d for which the stability conditions of Theorem 1.1.3 are satisfied for some $q \geq 2$. Then the best predictor of $X_{k+1,n}$, defined by the conditional expectation:

$$\widehat{X}^*_{k+1,n} := \mathbb{E}\left[X_{k+1,n} | X_{s,n},\ s \leq k\right] \tag{2.3}$$

and the best linear predictor of order p defined by:

$$\widehat{X}^*_{p,k+1,n} := \boldsymbol{\theta}^*_{p,k,n} \cdot \boldsymbol{X}^p_{k,n} \tag{2.4}$$

where $\boldsymbol{X}^p_{k,n} = (X_{k,n}, \ldots, X_{k-p+1,n})^T$ with appropriate padding and

$$\boldsymbol{\theta}^*_{p,k,n} := \underset{\boldsymbol{\theta} \in \mathbf{R}^p}{\arg\min}\, \mathbb{E}\left[\left(X_{k+1,n} - \boldsymbol{\theta} \cdot \boldsymbol{X}^p_{k,n}\right)^2\right] \tag{2.5}$$

correspond as soon as $p \geq d$.
Moreover, for $p = d$:

$$\boldsymbol{\theta}^*_{d,k,n} = \boldsymbol{\theta}(k/n).$$

The following proposition is based on section 2.3 of [RS16]

Proposition 2.2.2. (Prediction error development). Let $(X_{k,n})_{0\leq k\leq n}$ be a centered TVAR process of order d for which the stability conditions of Theorem 1.1.3 are satisfied for some $q \geq 4$. Let $\widetilde{\boldsymbol{\theta}}_{d,k,n}$ be an estimator of $\boldsymbol{\theta}^*_{d,k,n}$ which only depends on the observations for $s \leq k$. Then, we have the following inequality:

$$\left(\mathbb{E}\left[\left|\widetilde{\boldsymbol{\theta}}_{d,k,n}\cdot \boldsymbol{X}^d_{k,n} - \widehat{X}^*_{d,k+1,n}\right|^2\right]\right) \leq \mathbb{E}\left[\|\boldsymbol{X}^d_{k,n}\|^4\right]^{1/4}\mathbb{E}\left[\|\widetilde{\boldsymbol{\theta}}_{d,k,n} - \boldsymbol{\theta}^*_{d,k,n}\|^4\right]^{1/4}.$$

Remark. Note that we can control $\mathbb{E}\left[\|\boldsymbol{X}^d_{k,n}\|^4\right]^{1/4}$ using Theorem 1.1.3.

Remark. We can interpret these two results by saying that finding a good predictor for a TVAR process is somehow deeply related to finding a good estimator of $\boldsymbol{\theta}$.

2.3 Local Yule-Walker predictor

We now study a first way to construct a good estimator of $\boldsymbol{\theta}$ which can be used in prediction of TVAR processes. This estimator is based on the Yule-Walker equation (2.1) adapted to TVAR processes.

Proposition 2.3.1. Suppose the assumptions of Theorem 1.2.1 hold. The solution for the minimisation problem (2.5) which defines the best linear predictor (2.4) of order d for $n \geq d$ is given by:

$$\boldsymbol{\theta}^*_{d,k,n} = {\Gamma^*_{d,k,n}}^{-1}\gamma^*_{d,k,n} \tag{2.6}$$

where $\gamma^*_{d,k,n} = [\gamma^*(k,n,1),\dots,\gamma^*(k,n,d)]$ is the local covariance function and $\Gamma^*_{d,k,n} = (\gamma^*(k-i,n,j-i))_{i,j=1,..,d}$. Note that due to the assumption that σ is bounded from below by a positive constant, $\Gamma^*_{d,k,n}$ is always non-singular.

The idea behind this proposition is to estimate the time-varying covariance function, using its closeness to the local autocovariance function. By plugging this estimator into equation (2.6), we can derive an estimator of $\boldsymbol{\theta}^*_{d,k,n}$.

Definition 2.3.1. Local empirical covariance function. Given a taper function $h \mapsto [0,1]$ and $M \in 2\mathbb{N}$, the local empirical covariance function is given by:

$$\hat{\gamma}_{n,M}(u,l) := \frac{1}{H_M}\sum_{\substack{t_1,t_2=1\\ t_1-t_2=l}}^{M} h\left(\frac{t_1}{M}\right)h\left(\frac{t_2}{M}\right)X_{\lfloor un\rfloor+t_1-M/2,n}X_{\lfloor un\rfloor+t_2-M/2,n}$$

setting $H_M := \sum_{k=1}^M h^2(k/M) \sim M\int_0^1 h^2(x)\mathrm{d}x$.

Remark. Note that in a prediction context, the taper function must vanish for $u > \frac{1}{2}$. Otherwise, the estimator of $\boldsymbol{\theta}^*_{d,k,n}$ depends on $X_{k+1,n}$ and cannot be used to construct a predictor.

Remark. Let us refer to [RS16] Theorem 4.1 (p.10) to say that this is indeed a good estimator in the sense that assumption (C) (p.8) holds provided some conditions on h.

Convergence results
The main result concerning the predictor we derived is Theorem 3.2 (p.8) in [RS16] . We can see that there is a bias-variance tradeoff in the sense that when the bandwidth M is small, the variance of the local empircal covariance function is high, leading to a poor estimation of $\boldsymbol{\theta}^*_{d,k,n}$. We are indeed using a very small sample of observations to estimate the time-varying covariance function. On the contrary, when the bandwidth is long, the variance is reduced but the price to pay is that the bias of the estimator increases. This can be explained by the fact that we are using a too long sample to estimate the time-varying covariance function, preventing the estimator to adapt to the varitions of this very same function. This tradeoff will later be numerically illustrated.

Romberg bias reduction
Theorem 3.4 in [RS16] suggests to combine different predictors using weights for different bandwidths M in order to reduce the bias. This method is adapted from the Romberg method. On one side, we know that this method will reduce the bias. However, we cannot presume the impact of this combination on the variance of the estimation. This specific behaviour will also be numerically illustrated.

2.4 Normalised Least Mean Squares (NLMS)

2.4.1 Definition and intuitive derivation

We now introduce the NLMS predictor using the definition of [MPR05].

We consider real-valued observations $(X_{1,n}, X_{2,n}, ..., X_{n,n})$ from a time-varying autoregressive model (TVAR). They are ruled by the TVAR equation (1.3).

Our aim is to estimate the functions $t \mapsto \boldsymbol{\theta}(t)$ from the observations:

$$\{\boldsymbol{X}_{0,n}, \dots, X_{k,n}, k \geq 1\}.$$

Here we add the initial conditions $\boldsymbol{X}_{0,n}$ in the observations set for convenience.

More precisely, at a given time $t \in (0,1)$, only observations that have been observed before time t are used in the definition of the estimator:

$$\widehat{\boldsymbol{\theta}}_n(t) = \hat{\theta}_n(t, \boldsymbol{X}_{0,n}, X_{1,n}, \dots, X_{[nt],n})$$

where $[x]$ denotes the integer part of x. Hence, we are studying the Normalised Least Squares algorithm (NLMS), which is a recursive identification algorithm, defined as follows:

$$\hat{\boldsymbol{\theta}}_{0,n} = 0 \tag{2.7}$$

$$\hat{\boldsymbol{\theta}}_{k+1,n} = \hat{\boldsymbol{\theta}}_{k,n} + \mu(X_{k+1,n} - \hat{\boldsymbol{\theta}}_{k,n}^T \mathbf{X_{k,n}}) \frac{\mathbf{X_{k,n}}}{1 + \mu|\mathbf{X_{k,n}}|^{\mathbf{2}}}, k = 1 \cdots n-1 \tag{2.8}$$

At each iteration of the algorithm, the parameter estimates are updated by moving in the direction of the gradient of the instantaneous estimate $(X_{k+1,n} - \boldsymbol{\theta}^T \boldsymbol{X}_{k,n})^2$ of the local mean square error $\mathbb{E}[(X_{k+1,n} - \boldsymbol{\theta}^T \boldsymbol{X}_{k,n})^2]$. The normalization $(1 + \mu|\boldsymbol{X}_{k,n}|^2)^{-1}$ is a safeguard again large values of the norm of the regression vector and allows for very mild assumptions on the innovations. Note that equation (2.8) appears somehow related to the definition of the stochastic gradient descent algorithm. We define a pointwise estimate of $t \mapsto \boldsymbol{\theta}(t)$ as a simple interpolation of $\hat{\boldsymbol{\theta}}$.

2.4.2 Convergence results

We now present the main result of [MPR05] concerning the NLMS estimator, namely Theorem 2.

Theorem 2.4.1. Suppose the assumptions of Theorem 1.2.1 hold with $q \geq 4$ and let $p \in [1, \frac{q}{3})$. Let $\beta \in (0,1]$, $L > 0$, $0 < \rho < 1$ and $0 < \sigma^- \leq \sigma^+$. Then there exist $M, \delta > 0$ and $\mu_0 > 0$ such that, for all $\mu \in (0, \mu_0]$, $n \geq 1$, $t \in (0,1]$ and $(\boldsymbol{\theta}, \sigma) \in \mathcal{C}_d^+(\beta, L, \rho, \sigma_-, \sigma_+)$,

$$||\widehat{\boldsymbol{\theta}}_n(t;\mu) - \boldsymbol{\theta}(t)||_{p,\boldsymbol{\theta},\sigma} \leq M(|\boldsymbol{\theta}(0)|(1-\delta\mu)^{tn} + \sqrt{\mu} + (n\mu)^{-\beta}) \tag{2.9}$$

where $||X||_{p,\boldsymbol{\theta},\sigma} := \mathbb{E}_{p,\boldsymbol{\theta},\sigma}[|X|^p]^{1/p}$.

Remark. We have three terms in this upper bound. The first term $|\boldsymbol{\theta}(0)|(1-\delta\mu)^{tn}$ reflects the effect of the initial error (how far we start from the true initial value). We see that this term is decaying exponentially with time. We see that a bigger stepsize increases the rate at which we forget this initial error.

The second term is the fluctuation of the recursive method. It is a statistical error term, in other words a variance term which is increasing with the stepsize.

The third term controls the error involved by time evolution of $\boldsymbol{\theta}(t)$ and mainly relies on the smoothness exponent β. It is a bias term which decreases with the stepsize.

To wrap up this analysis, if we have a small step parameter μ then the first term and the third will converge slowly towards 0. Conversely, with a greater μ the two terms converge quickly but the lag-noise $\sqrt{\mu}$ is greater and counterbalances the two others. This behaviour can be interpreted as a bias-variance tradeoff.

2.4.3 Romberg bias reduction

We can combine recursive estimators to reduce the bias of the estimator. This is inspired by the Romberg method. We define for $\gamma \in (0,1)$

$$\tilde{\boldsymbol{\theta}}_n(t;\mu,\gamma) := \frac{1}{1-\gamma}(\hat{\boldsymbol{\theta}}_n(t;\mu) - \gamma\hat{\boldsymbol{\theta}}_n(t;\gamma\mu)) \tag{2.10}$$

Remark. Once again, this is a bias reduction method. But we have no guarantee on the impact of this combination on the variance of the estimator. This will be numerically illustrated.

2.5 Exponential aggregation of predictors

In this section, we study the aggregation method presented in [GRS15]. The idea is to mix a finite number of predictors such as those presented above in order to improve the prediction. The weighting technique in aggregation has been used and developed in the machine learning community for instance. It presents deep similarities with the so-called Adaboost method.

The new predictor $\hat{X}_t$ is defined as a weighted sum of N predictors $\hat{X}_t^{(i)}$ as follows:

$$\hat{X}_t = \sum_{i=1}^{N} \hat{\alpha}_{i,t}\hat{X}_t^{(i)}, \qquad 1 \le t \le T \tag{2.11}$$

There are two ways to build the weights $(\hat{\alpha}_{i,t})_{i=1,\dots,N}$. This first way uses the gradient of the quadratic loss and gives us the weights:

$$\hat{\alpha}_{i,t} = \frac{\exp(-2\eta\sum_{s=1}^{t-1}(\sum_{j=1}^{N}\hat{\alpha}_{j,s}\hat{X}_s^{(j)} - X_s)\hat{X}_s^{(i)})}{\sum_{k=1}^{N}\exp(-2\eta\sum_{s=1}^{t-1}(\sum_{j=1}^{N}\hat{\alpha}_{j,s}\hat{X}_s^{(j)} - X_s)\hat{X}_s^{(i)})} \tag{2.12}$$

This second way uses the quadratic loss and gives us the weights:

$$\hat{\alpha}_{i,t} = \frac{\exp(\eta \sum_{s=1}^{t-1} (\hat{X}_s^{(j)} - X_s)^2)}{\sum_{k=1}^{N} \exp(\eta \sum_{s=1}^{t-1} (\hat{X}_s^{(j)} - X_s)^2)} \tag{2.13}$$

with the weights initialised for $t = 1$ at $\alpha_t = \frac{1}{N}$.

Remark. The way we build the weights have an immediate interpretation: we penalise the predictors which tend to be wrong and we boost good predictors. Let us remark that when the temperature η is rising, the weights will tend to select exclusively the predictor which has the best performances. On the contrary, when the temperature is decreasing, we will share the weights uniformly on the predictors.

Remark. The NLMS is well adapted to this aggregation method as it can be computed in line. [GRS15] focuses a lot on the use of exponential aggregation of NLMS predictors and shows this has good properties.

3. Numerical study of the predictors

In this section, we represent some numerical experiments. We illustrated numerically some theoretical properties we have presented above. We focus in particular on the bias-variance tradeoff and the bias reduction method based on the Romberg method we explained before.

3.1 Local autoregression vector selection and TVAR simulation

In order to pick $(\boldsymbol{\theta}, \sigma) \in \mathcal{C}$, we use the same methodology as in Section 5 of [RS16]. The method is based on the random sampling of partial autocorrelation coefficients from which we derive autoregression coefficients using the Levinson-Durbin algorithm. σ is set constant for convenience.

3.2 Exploring theoretical properties

We picked up a local autoregression coefficient applying the method explained above to simulate 500 independent samples of the TVAR process with length 2^{15}. We used normal innovations, a constant innovation standard deviation equal to 1 and normal initial conditions.

Figure 3.2 represents one sample of the TVAR process. Figure 3.1 represents the local autoregression coefficient we picked up as a function of local time.

In order to assess visually the behaviour of the estimators we study, we estimate the local autoregression vector and predict the next value of the process for each sample. As we know the underlying local autoregression vector, we can directly compare the estimation error for each estimator of the local autoregression vector. We can also compare the best predictions and the predictions computed by the different methods for each sample. The loss measure we used is the mean squared error (MSE).

3.2.1 Local Yule-Walker numerical results

We see on Figure 3.3 that the mean squared error of the estimator (MSE) depends a lot on the bandwidth we use to estimate the local autocovariance function. In this setting, the minimum is reached for a bandwidth equal to 2^{12}. Let us now represent on Figure 3.4 the predictions computed for this optimal bandwidth.

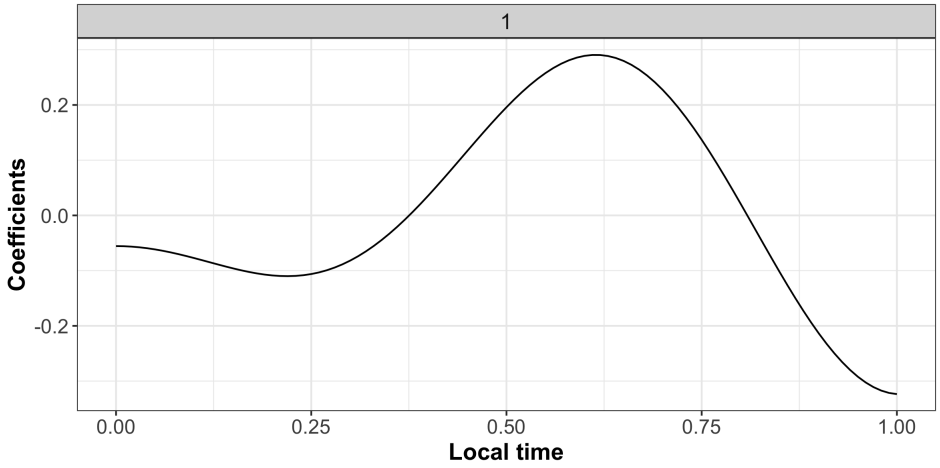

Figure 3.1: Local autoregression coefficient used in the numerical simulations

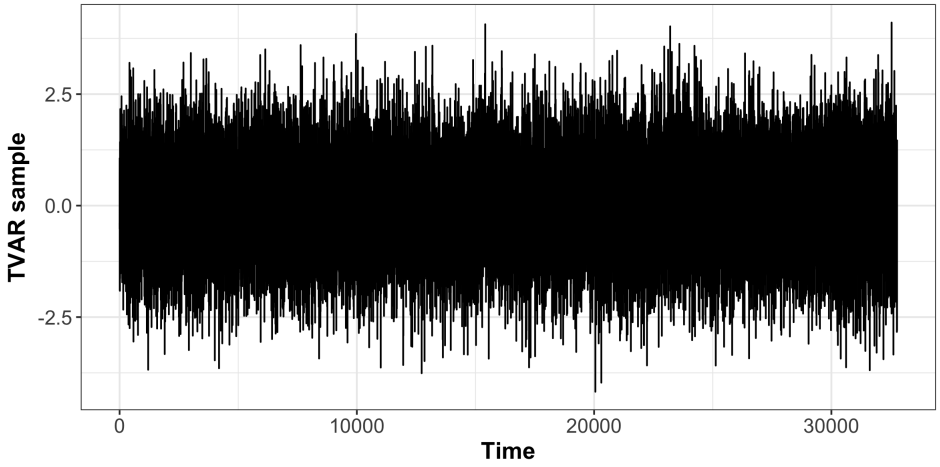

Figure 3.2: One sample of the TVAR process of length 2^{15}

Figure 3.4 represents the best predictions for each sample as a function of the computed predictions. It is a diagnosis plot of the predictor which allows us to identify its specific behaviours. We have added the first bisector to this scatter plot. The closer the scatter plot is from this line, the better the predictor is. We are quite satisfied by this first diagnosis plot.

Bias-variance tradeoff. Let us now illustrate the behaviour of the Yule-Walker predictor for non optimal bandwidths. Figure 3.5 illustrates the bias-variance tradeoff we explained in section 2.3. We have represented the predictions for non optimal bandwidths. We see that a small bandwidth ($M = 2^5$) leads to a high variance of the predictor which is due to the small size of the sample used to estimate the local autocovariance function. Nevertheless, the bias of these predictions is quite satisfying as we see no deviation pattern from the first bisector. On the contrary, a long bandwidth ($M = 2^{15}$) increases the bias of the predictions. This is clearly

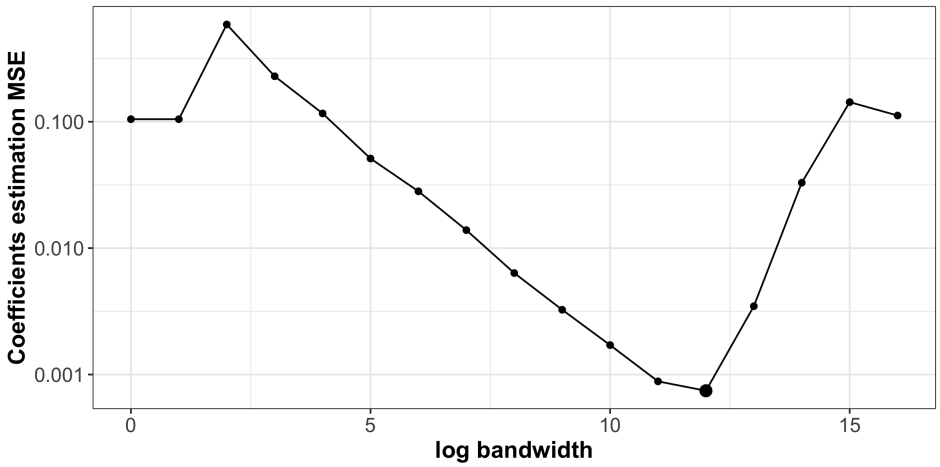

Figure 3.3: Coefficients estimation (Yule-Walker) MSE in function of the bandwdith

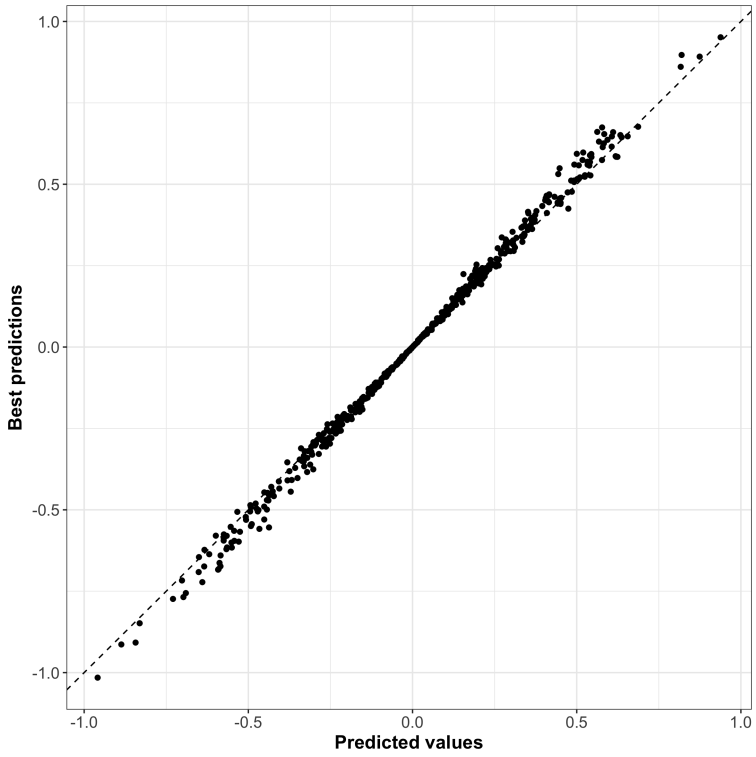

Figure 3.4: Diagnosis plot of the Yule-Walker predictor with optimal bandwidth $M = 2^{12}$

showed by the deviation pattern of the predictions from the first bisector. Note that however the variance is reduced in comparison to the previous bandwidth.

Romberg bias reduction. Figure 3.6 shows a slight gain on the bias of the predictions when one uses the Romberg bias reduction technique. The optimal bandwidth range for the Romberg method was 2^{14}-2^{16}. Figure 3.3 recalls us that

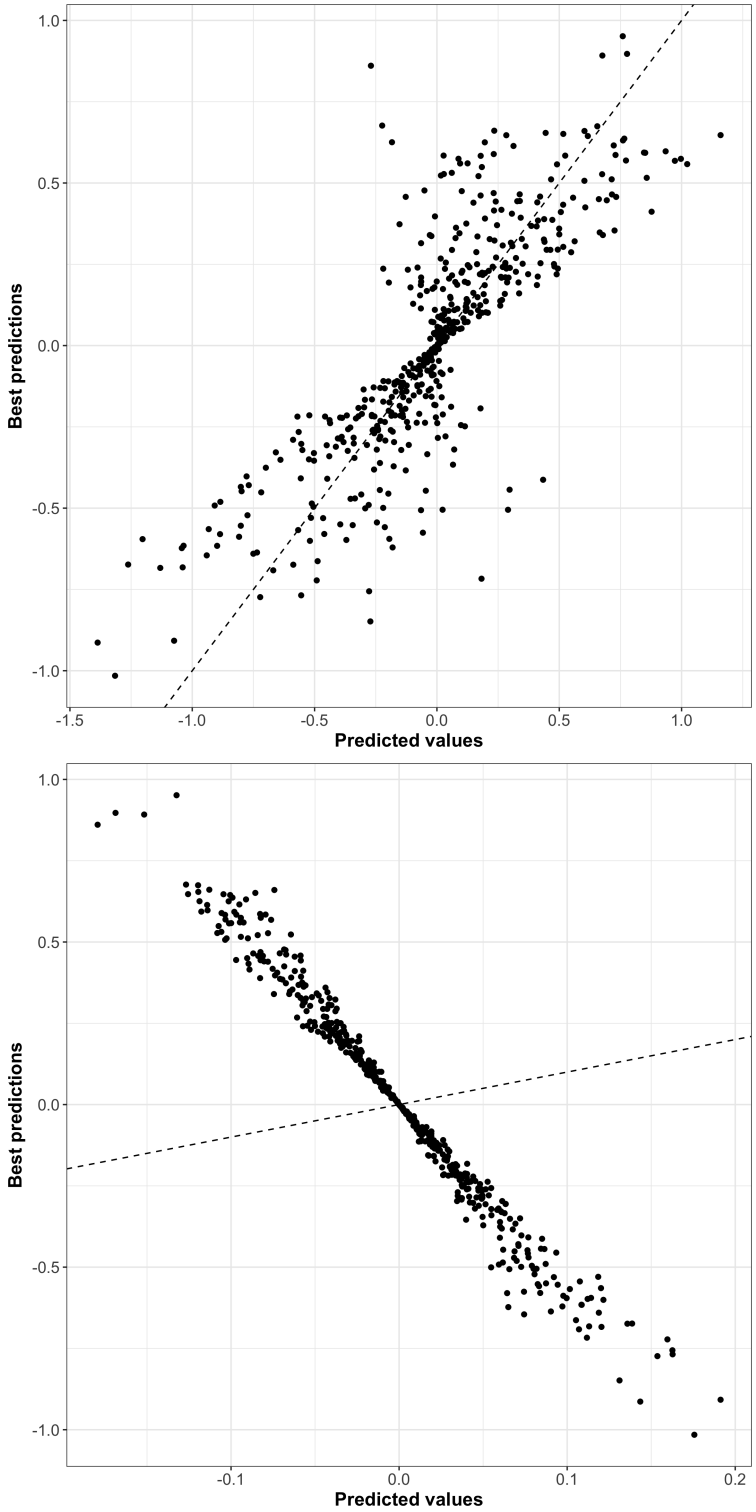

Figure 3.5: Diagnosis plot of the the Yule-Walker predictor with bandwidths $M = 2^5$ and $M = 2^{15}$ respectively

these bandwidth are far from being optimal. However, as we explained, these long bandwidths are not satisfying because of their bias and not because of their variance.

Then by combining them using the appropriate weights, one can kill some bias terms, reducing the bias of the prediction. Hopefully, this combination will not increase the variance of the predictions too much but there is no guarantee for this fact.

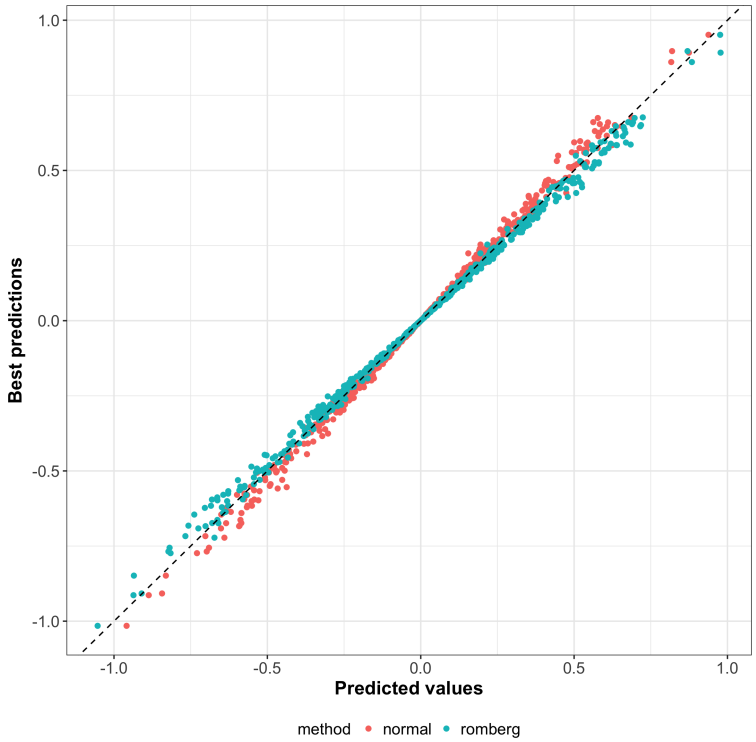

Figure 3.6: Diagnosis plot of the the Yule-Walker predictor with optimal bandwidth and with Romberg bias-reduction for the optimal bandwidth range

3.2.2 Normalised Least Mean Squares

We now present analogue results in the case of the NLMS predictor.

We see on Figure 3.7 that the mean squared error of the estimator (MSE) depends a lot on the stepsize μ we use in the iterative algorithm described by equation (2.8). In this setting, the minimum is reached for a stepsize equal to $10^{-2.8}$. Let us now represent on Figure 3.8 the predictions computed for this optimal bandwidth.

The diagnosis plot on Figure 3.8 is rather satisfying.

Bias-variance tradeoff. Let us now illustrate the behaviour of the NLMS predictor for non optimal stepsizes. Figure 3.9 illustrates the bias-variance tradeoff we explained in section 2.4.2. We have represented the predictions for non optimal stepsizes. We see that a small stepsize ($M = 10^{-3.8}$) leads to a high bias of the predictor. This can be interpreted by the fact that the stochastic gradient descent-like algorithm is stuck because its steps are too small. Following this interpretation, the variance is small as the algorithm cannot move a lot from its initial position. On the contrary, with a larger stepsize, here $\mu = 10^{-1}$, the bias is reduced as we see no

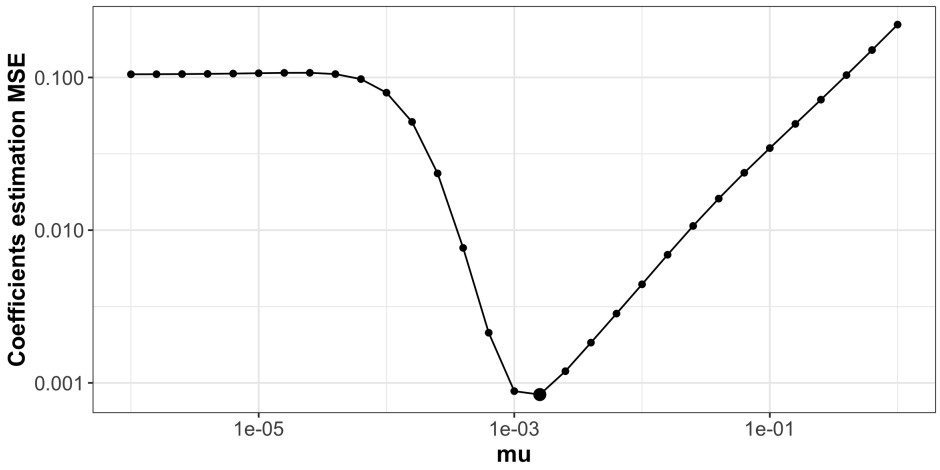

Figure 3.7: Coefficients estimation (NLMS) MSE in function of the stepsize μ

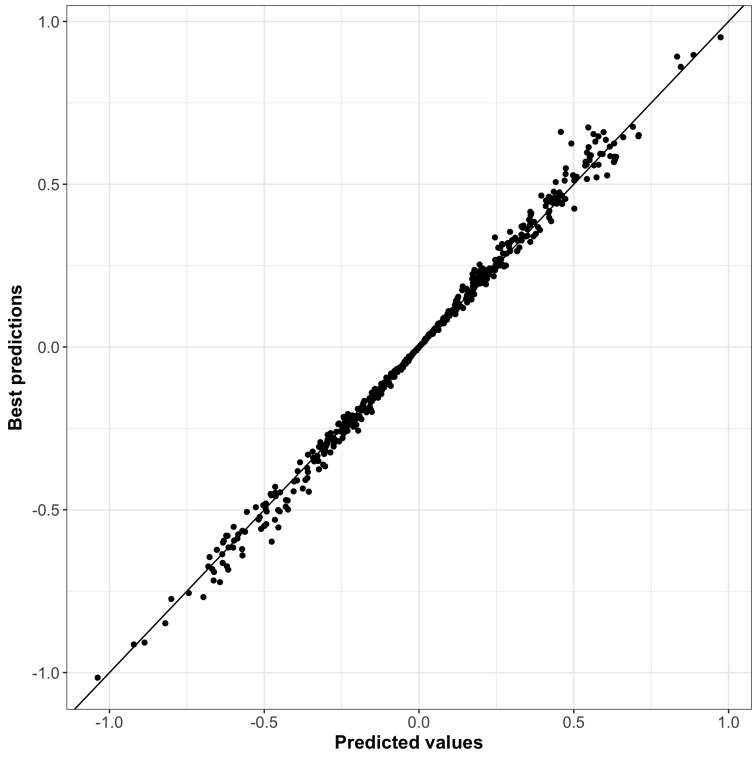

Figure 3.8: Diagnosis plot of the NLMS predictor with optimal step-size $\mu = 10^{-2.8}$

deviation pattern from the first bisector. However, the variance increases. This is to be related to the idea that in stochastic gradient descent, the stepsize needs to be small (even decreasing) to ensure convergence. Note in particular the slope of the MSE for $\mu \geq 10^{-3}$ on Figure 3.7. It is approximately equal to $\frac{1}{2}$ which shows that the dominating term is indeed the variance term $\sqrt{\mu}$ in the error control inequality (2.9) of Theorem 2.4.1.

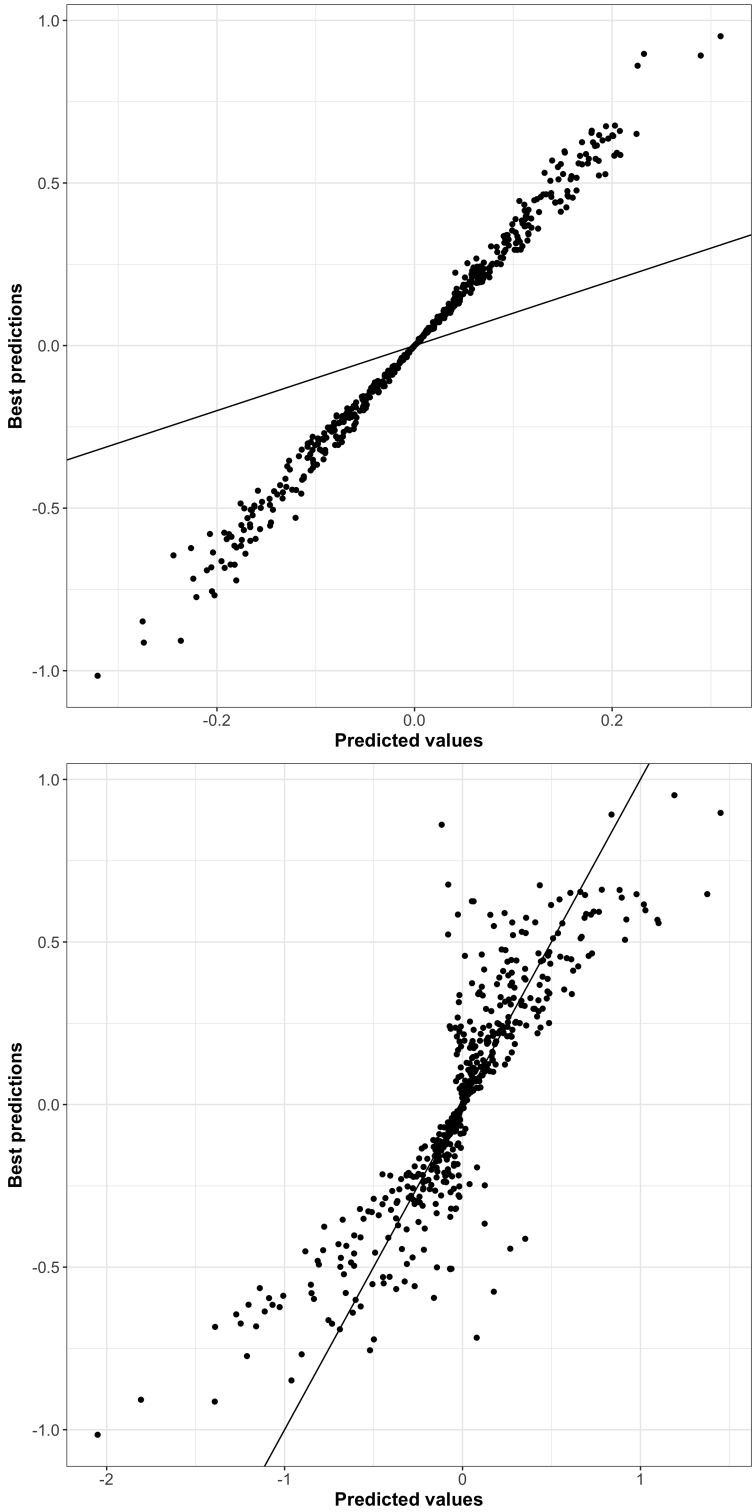

Figure 3.9: Diagnosis plot of the the NLMS predictor with stepsizes $\mu = 10^{-3.8}$ and $\mu = 10^{-1}$ respectively

Romberg bias reduction. Figure 3.10 shows the coefficients estimation MSE for two different NLMS predictors. The first one ($\gamma = 0$) corresponds to the classic

NLMS estimator we introduced in section 2.4.2. The second one ($\gamma = 0.5$) corresponds to the bias reduction technique introduced in section 2.4.3. We see that we can find a stepsize for which the bias-reduced estimator's MSE in better than the optimal MSE for the classic NLMS estimator. Note that the optimal stepsize for this bias-reduced estimator is lower than the optimal stepsize for the classic estimator. We could have expected it as the bias reduction technique allows to take estimators for which the bias is higher as the combination will later reduce it. By picking smaller stepsize, this estimator also benefits from a smaller variance. Finally note (Figure 3.11) the difference of slope when the MSE in decreasing for $10^{-4} \leq \mu \leq 10^{-3}$. We clearly see that the Romberg bias-reduced estimator's MSE is decreasing about twice faster, which is what is expected when one looks at Theorem 8 of [MPR05]. Let us finish by saying that if the variance is increasing too fast for the Romberg predictor, we might not have a lower minimal MSE.

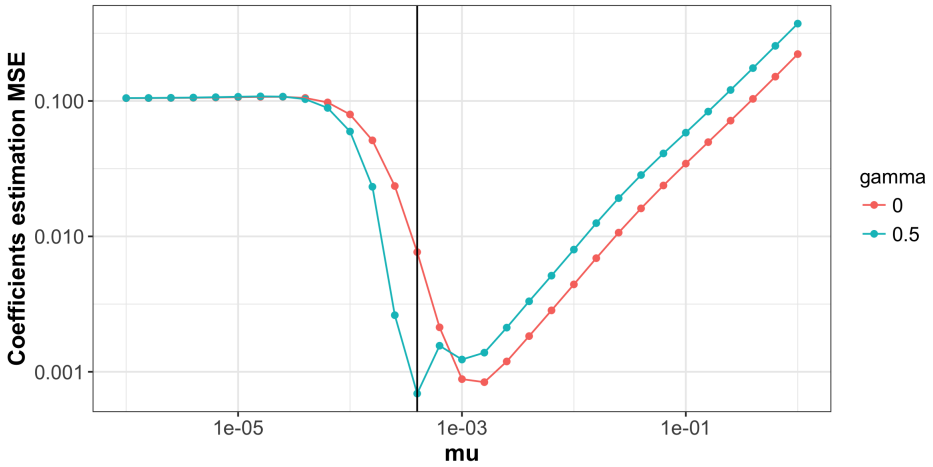

Figure 3.10: Coefficients estimation MSE for the NLMS predictor with $\gamma = 0$ and $\gamma = 0.5$ in function of the stepsize μ

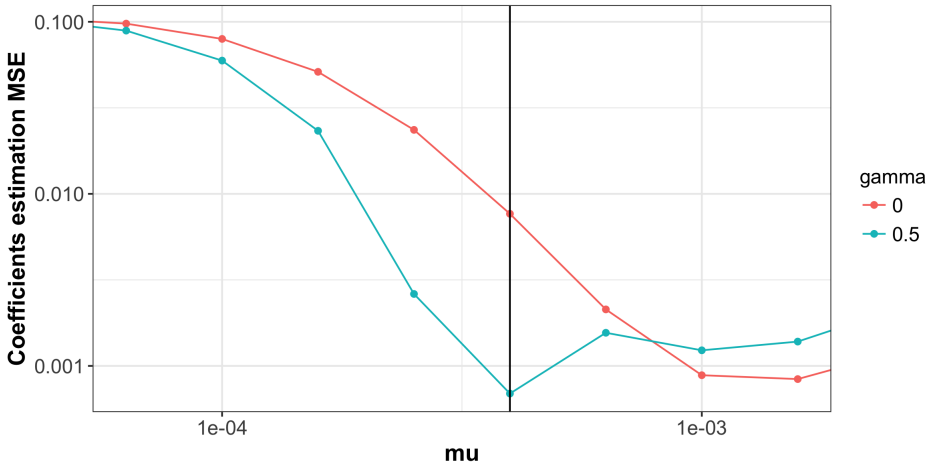

Figure 3.11: Coefficients estimation MSE for the NLMS predictor with $\gamma = 0$ and $\gamma = .5$ in function of the stepsize μ (zoom)

3.2.3 Exponential aggregation of predictors

Let us finally investigate the exponential aggregation technique in the case of the quadratic loss strategy (2.13).

We aggregated 5 NLMS predictors whose stepsizes are given in Table 3.1. We used a temperature $\eta = 0.02$.

We can see on Figure 3.12 that aggregating several predictors can sometimes lead to a better MSE. The horizontal line which represents the MSE of the aggregated predictor is indeed below the MSE of the best predictor used for aggregation.

Predictor	$\log \mu$
a	-4.00
b	-3.25
c	-2.50
d	-1.75
e	-1.00

Table 3.1: Stepsizes of the aggregated NLMS predictors

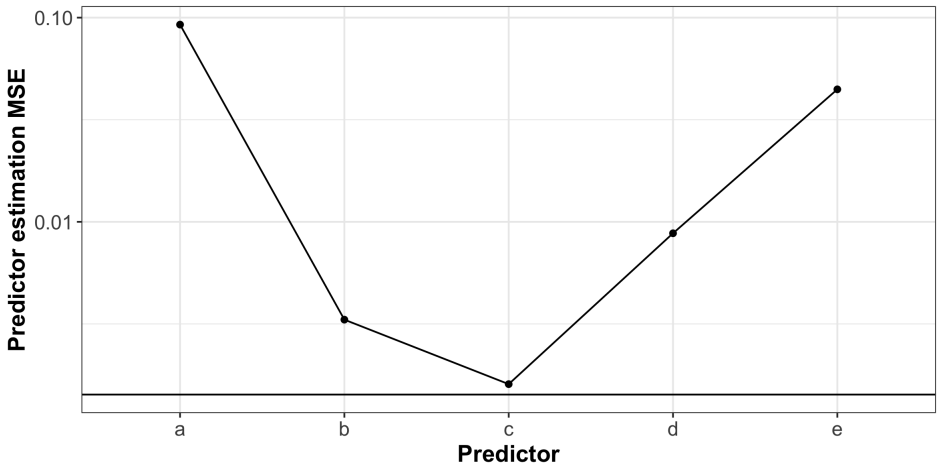

Figure 3.12: Prediction MSEs for the exponential aggregation technique

This improvement can be checked on Figure 3.13 where we see that the aggregated predictor is slightly better than the best predictor (c).

Let us finally note that according to Figure 3.14, the aggregation technique can effectively find the best predictor among a given set of predictors. We indeed see through the boxplots that the best predictor (c) has always the most important weight among the other predictor.

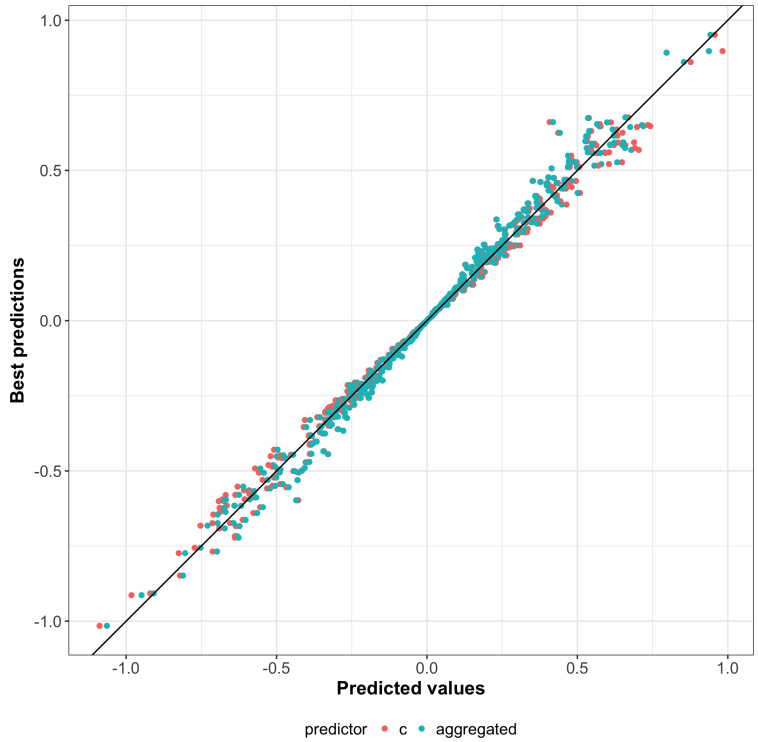

Figure 3.13: Diagnosis plot of the aggregated predictor compared to the best predictor used for aggregation

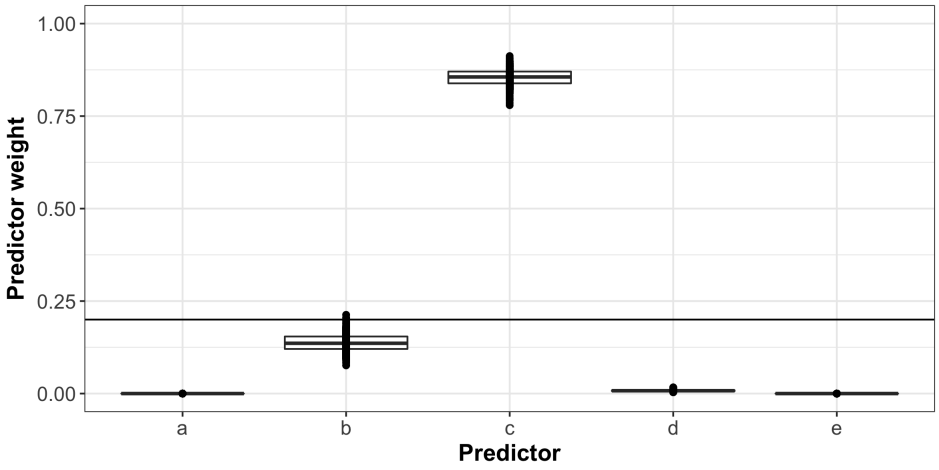

Figure 3.14: Boxplots of the weights used in aggregation for the 500 samples

Conclusion

Throughout this project, we had the opportunity to work on several theoretical concepts about locally stationary processes, more precisely TVAR processes.

The idea to predict time series that are not exactly weakly stationary aims to address a larger statistical problem: predicting series that have only a locally stationary pattern.

We discovered three important methods to predict TVAR processes: local Yule-Walker prediction [RS16], Normalized Mean Least Squares method [MPR05] and finally an exponential aggregation of the previous method [GRS15].

The numerical implementation was very satisfying for our simulated data. We observed the bias-variance compromise and used the bias-reduction Romberg methods to improve our predictions.

Finally, we could ask the question whether these methods remain relevant for the prediction of real life data. We did not explore this question as it was not the core of our project. However, for data that meets the TVAR equation, the prediction methods are very adequate.

Bibliography

[MPR05] Eric Moulines, Pierre Priouret, and François Roueff. "On recursive estimation for time varying autoregressive processes". In: *Ann. Statist.* 33.6 (Dec. 2005), pp. 2610–2654. DOI: 10.1214/009053605000000624. URL: https://doi.org/10.1214/009053605000000624.

[GRS15] Christophe Giraud, François Roueff, and Andres Sanchez-Perez. "Aggregation of predictors for nonstationary sub-linear processes and online adaptive forecasting of time varying autoregressive processes". In: *Ann. Statist.* 43.6 (Dec. 2015), pp. 2412–2450. DOI: 10.1214/15-AOS1345. URL: https://doi.org/10.1214/15-AOS1345.

[RS16] François Roueff and Andres Sanchez-Perez. "Prediction of weakly locally stationary processes by auto-regression". (working paper). Feb. 2016. URL: https://hal.archives-ouvertes.fr/hal-01269137.

[Rou17] François Roueff. *Financial time series in discrete time*. M2 Stat & Finance. Nov. 2017. URL: https://perso.telecom-paristech.fr/roueff/.