Online adaptive forecasting of a locally stationary time varying autoregressive process

Andrés SÁNCHEZ PÉREZ and François ROUEFF

Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI

May 23, 2013

Abstract

In this work, we study the problem of online adaptive forecasting for locally stationary Time Varying Autoregressive processes (TVAR). The Normalized Mean Least Squares algorithm (NMLS) is an online stochastic gradient method which has been shown to perform efficiently, provided that the gradient step size is well chosen. This choice highly depends on the smoothness exponent of the evolving parameters. In this contribution, we show that a sequential aggregation of several NLMS estimators at various gradient step sizes is able to adapt to an unknown smoothness, resulting in an online adaptive predictor.

1 Introduction

In many applications where high frequency data is collected, one wishes to predict the next values of an observed time series through an online predictor learning algorithm, allowing one to process a large amount of data. However, as a counterpart, the usual stationarity assumption has to be weakened to take into account some smooth evolution of the environment. An interesting approach to cope with this non-stationarity issue is to rely on a *local stationarity* assumption. We refer to [2] and the references therein for a recent general view about statistical inference for locally stationary processes. Here we focus on a particular model, which is obtain by apply this approach to a time-varying autoregressive process.

Definition 1 (Time-varying autoregressive process (TVAR)) *The T-sample $X_{1,T}, \ldots, X_{T,T}$ of a TVAR process or order d satisfies,*

$$X_{t,T} \quad = \quad \sum_{j=1}^{d} \theta_j \left(\frac{t-1}{T} \right) X_{t-j,T} + \sigma \left(\frac{t}{T} \right) \xi_t \;, \tag{1}$$

where the ξ_t are i.i.d. with $\mathbb{E}\xi_t = 0$, $\mathbb{E}|\xi_t| < \infty$, and θ_j are the time-varying autoregressive coefficients rescaled on the interval $[0, 1]$.

Some initial conditions can be added but we omit them here for brevity. It is usually assumed that ξ_t is independent of the past of $X_{s,T}$ up to $s = t - 1$ so that the best predictor of $X_{t,T}$ given this past is $\boldsymbol{\theta}' \left(\frac{t-1}{T} \right) \underline{X}_{t-1,T}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ and $\underline{X}_{t-1,T} = (X_{t-1,T}, \ldots, X_{t-d,T})'$. Here and in the following, we let A' denote the transpose of the matrix A.

The local stationary time-varying autoregressive polynomial for a TVAR is defined as $\boldsymbol{\theta}(z; t) = 1 - \sum\limits_{j=1}^{d} \theta_j(u) z^j$.

For $\zeta > 0$ we denote by $s_d(\zeta) = \left\{ \boldsymbol{\theta} : [0, 1] \to \mathbb{R}^d, \boldsymbol{\theta}(z; u) \neq 0, \forall |z| < \zeta^{-1}, u \in [0, 1] \right\}$.

Following [3] and [4], a TVAR process is *locally stationary* if $\boldsymbol{\theta}$ and σ satisfy some smoothness conditions and $\boldsymbol{\theta} \in s_d(\delta)$ for some $\delta \in (0, 1)$. These conditions imply that there exists a solution with representation :

$$X_{t,T} \quad = \quad \sum_{j=0}^{\infty} a_{t,T}(j) \xi_{t-j} \;, \tag{2}$$

and that there exist $\bar{K} > 0$ and $\rho \in [0, 1)$ such that $\sup\limits_{t,T} |a_{t,T}(j)| \leq \bar{K} \rho^j$.

In this contribution, we shall use $\beta-$ Lipschitz smoothness conditions. For any $\beta \in (0, 1]$, the $\beta-$ Lipschitz semi-norm of a function $f : [0, 1] \to \mathbb{R}^d$ is defined as $|f|_{\Lambda,\beta} = \sup\limits_{s_1 \neq s_2} \dfrac{f(s_1) - f(s_2)}{|s_1 - s_2|^{\beta}}$. For $L \in \mathbb{R}_+^*$ and $\beta > 0$, let $k \in \mathbb{N}$ and $\alpha \in (0, 1]$ be such that $\beta = k + \alpha$. The $\beta-$ Lipschitz ball is :

$$\Lambda_d(\beta, L) \quad = \quad \left\{ f : [0, 1] \to \mathbb{R}^d, \left| f^{(k)} \right|_{\Lambda,\alpha} \leq L, \sup_{s \in [0,1]} |f(s)| \leq L \right\} \;. \tag{3}$$

The quality of the prediction is measured using a loss function $\ell(x, y) = |x - y|^q$ for some $q = 1, 2, 3, \ldots$

2 NLMS estimators

In [5], the normalized least mean squares algorithm (NLMS) estimator of the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ is studied for locally stationary TVAR processes. We will use basically the same estimators but with a slight modification (Eq. (5) below). For a given gradient step size $\mu > 0$, our modified NLMS estimator is defined recursively as follows :

$$\tilde{\boldsymbol{\theta}}_{t,T}(\mu) \quad = \quad \tilde{\boldsymbol{\theta}}_{t-1,T}(\mu) + \mu\left(X_{t,T} - \tilde{\boldsymbol{\theta}}'_{t-1,T}(\mu)\underline{X}_{t-1,T}\right) \frac{\underline{X}_{t-1,T}}{1 + \mu\left|\underline{X}_{t-1,T}\right|_2^2} \ , \tag{4}$$

$$\hat{\boldsymbol{\theta}}_{t,T}(\mu) \quad = \quad \begin{cases} \tilde{\boldsymbol{\theta}}_{t,T}(\mu) & \text{if } \left|\tilde{\boldsymbol{\theta}}_{t,T}(\mu)\right|_2 \leq 2^d - 1 \ , \\ \dfrac{2^d - 1}{\left|\tilde{\boldsymbol{\theta}}_{t,T}(\mu)\right|_2} \tilde{\boldsymbol{\theta}}_{t,T}(\mu) & \text{otherwise.} \end{cases} \tag{5}$$

Here $|\cdot|_2$ stands for the Euclidean norm. The additional step (5) is a projection on the ball of radius $2^d - 1$ which guaranties our estimators to be bounded. The statistic $\hat{\boldsymbol{\theta}}_{t-1,T}(\mu)$ is our estimation for $\boldsymbol{\theta}\left(\frac{t-1}{T}\right)$, from which we obtain the predictor $\hat{\boldsymbol{\theta}}'_{t-1,T}(\mu)\underline{X}_{t-1,T}$ of $X_{t,T}$. One of our objectives is to control the mean risk in prediction using the estimators above. Following [5], it can be shown that, under some mild moment assumptions on the noise (ξ_t), for any $\delta \in (0,1)$, $L, 0 < \sigma_- < \sigma_+$ and $\beta \in (0,1]$, there exists $C > 0$ such that,

$$\sup_{\boldsymbol{\theta} \in s_d(\delta) \cap \Lambda_d(\beta,L), \sigma(u) \in [\sigma_-, \sigma_+], \forall u \in [0,1]} \frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E}\left[\ell\left(\hat{\boldsymbol{\theta}}'_{t-1,T}\mu)\underline{X}_{t-1,T}, X_{t,T}\right)\right] - \sigma^q\left(\frac{t}{T}\right) \mathbb{E}\left[|\xi_0|^q\right]\right) \quad \leq \quad C\left[(T\mu)^{-1} + \left(\sqrt{\mu} + (T\mu)^{-\beta}\right)^q\right] . \tag{6}$$

Observe that, in this bound, the optimal choice of the parameter μ depends on β. This fact leads us to consider an additional aggregation step as explained below.

3 Aggregation of predictors

From a collection $\left\{\hat{\boldsymbol{\theta}}^{(j)}_{t-1,T}\right\}_{1 \leq j \leq N}$ of N estimators of $\boldsymbol{\theta}$ we obtain the respective predictions of $X_{t,T}$: $f_{j,t} = \hat{\boldsymbol{\theta}}^{(j)\prime}_{t-1,T}\underline{X}_{t-1,T}$, $1 \leq j \leq N$. In particular, each index j may correspond to a NLMS estimator obtained with a given μ_j. In aggregation language, the $f_{j,t}$s are called expert's predictions or forecasts. The strategy used in a different context (bounded observations) by [1] suggest to combine all possible expert's predictions using some weights specified by :

$$\bar{\alpha}_{j,t} \quad = \quad \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{j,s}\right)}{\sum_{i=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{i,s}\right)} \ , \tag{7}$$

where $\tilde{\ell}_{j,t} = \nabla_x \ell\left(\sum_{i=1}^{N} \alpha_{i,t} f_{i,t}, X_{t,T}\right) \cdot f_{j,t}$ and with the convention that a sum over no element is null, i.e. $\bar{\alpha}_{j,1} = \frac{1}{N}$ for all j. The subgradient is taken with respect to the first coordinate on ℓ. The parameter $\eta > 0$ is known as learning rate and will be specified. Based on sequential aggregation techniques (see [1]), and some extensions of the bound (6), we obtain the following result.

Theorem 1 *Suppose that $\mathbb{E}[|\xi_0|^r] < \infty$ for some $r > 3q$. Let $\hat{\theta}^{(j)}_{t,T} = \hat{\boldsymbol{\theta}}_{t,T}\left(T^{-2j/(2j+N)}\right)$ for $j = 1, \ldots, N$. Define the aggregated estimator $\check{\boldsymbol{\theta}}_{t-1,T} = \sum_{j=1}^{N} \bar{\alpha}_{j,t}\hat{\boldsymbol{\theta}}^{(j)}_{t-1,T}$, $1 \leq t \leq T$ with $\eta \propto \sqrt{\frac{\log N}{T}}$. Suppose moreover that $\boldsymbol{\theta} \in s_d(\delta) \cap \Lambda_d(\beta, L)$ for $L > 0$ and $\beta \in (0,1]$, and that σ is bounded between two positive values. Then we have :*

$$\frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E}\left[\ell\left(\check{\boldsymbol{\theta}}'_{t-1,T}\underline{X}_{t-1,T}, X_{t,T}\right)\right] - \sigma^q\left(\frac{t}{T}\right) \mathbb{E}\left[|\xi_0|^q\right]\right) \quad = \quad O\left(T^{-1/(1+2\beta)} + T^{-q\beta/(1+2\beta)}\right) \ , \tag{8}$$

provided that $\frac{\log(T)}{N} = O(1)$ as $T \to \infty$.

Observe that the convergence rate is the same as the one of (6) when μ is optimized according to the given value β for $q\beta \leq 1$. Here, in contrast, the estimator does not require the knowledge of β to achieve this rate.

4 Numerical results

We now provide some numerical experiments to support our theoretical findings. We simulated 100 samples of length $T = 1000$ of a locally stationary TVAR process of order $d = 3$. The definition of the time varying coefficients is omitted here for brevity. We just mention that the stability condition $\boldsymbol{\theta} \in s_d(\delta)$ is satisfied with $\delta = 0.8$. A path of one of these samples is displayed in Figure 1 while the true coefficient θ_1 and the corresponding estimates $\hat{\theta}^{(j)}_{1,t,T}$ for $j = 1, \ldots, N = 6$ and $\check{\theta}_{1,t,T}$ obtained from this sample are displayed in Figure 2. Observe that the behaviour of each expert is as expected : the higher μ_j, the smoother the path of $\hat{\theta}^{(j)}_{1,t,T}$. Finally Figure 3 shows boxplots of the empirical error $R_T(\hat{\boldsymbol{\theta}}) = \frac{1}{T}\sum_{t=1}^{T}\left(\left|\hat{\boldsymbol{\theta}}'_{t-1,T}\underline{X}_{t-1,T} - X_{t,T}\right|^2 - \sigma^2\left(\frac{t}{T}\right)\mathbb{E}\left[|\xi_0|^2\right]\right)$ for $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(j)}_{t,T}$ for $j = 1, \ldots, 6$ and for $\hat{\boldsymbol{\theta}} = \check{\boldsymbol{\theta}}_{t,T}$ in Slot 7. We note that among the NLMS "experts", one indeed seems to have the optimal gradient step (see Slot 2) and that the aggregate predictor achieves comparable performances, as expected.

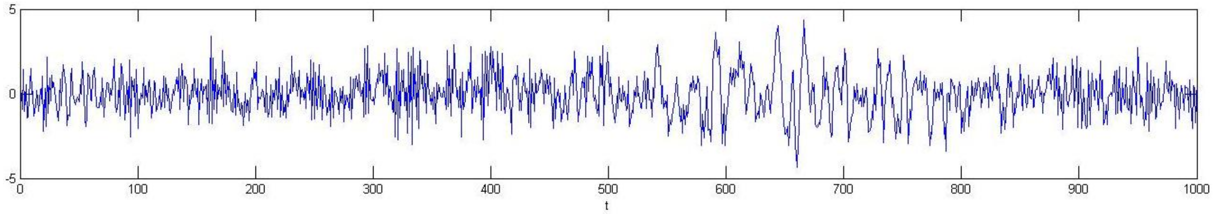

Figure 1: A sample path of the TVAR process

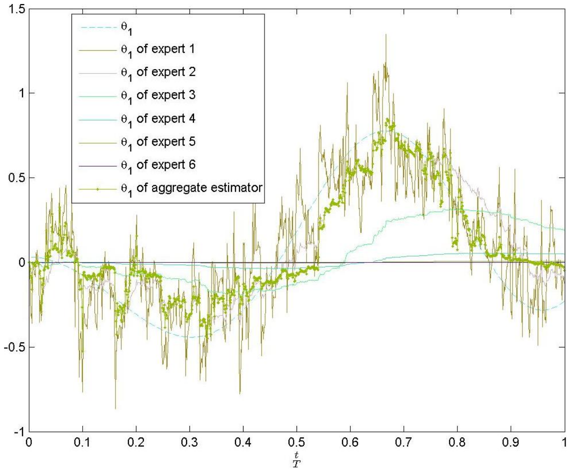

Figure 2: Estimations of θ_1

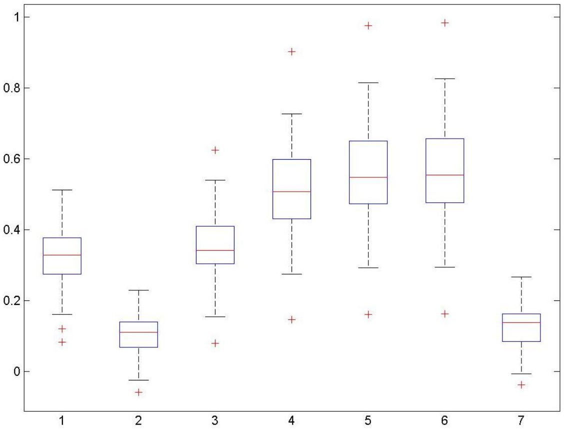

Figure 3: Empirical errors

Acknowledgements

This work has been partially supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île de France (RDM-IdF) for the period 2012 - 2015 and by the Labex LMH (ANR-11-IDEX-003-02). We are grateful to Christophe GIRAUD from Université Paris-Sud for fruitful discussions about this work.

References

[1] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.

[2] Rainer Dahlhaus. Local inference for locally stationary time series based on the empirical spectral measure. *J. Econometrics*, 151(2):101–112, 2009.

[3] Rainer Dahlhaus and Wolfgang Polonik. Empirical spectral processes for locally stationary time series. *Bernoulli*, 15(1):1–39, 2009.

[4] Hans Rudolf Künsch. A note on causal solutions for locally stationary ar-processes. 1995.

[5] Eric Moulines, Pierre Priouret, and François Roueff. On recursive estimation for time varying autoregressive processes. *Ann. Statist.*, 33(6):2610–2654, 2005.