Numerical cost for time series prediction via aggregation

Andrés Sánchez Pérez, François Roueff

Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI
Télécom ParisTech, 37 rue Dareau, 75014 Paris, France
andres.sanchez-perez@telecom-paristech.fr, francois.roueff@telecom-paristech.fr

Résumé – Dans ce travail, on étudie le problème de la prédiction pour les processus de Bernoulli Causaux et Décalés (CBS). La technique de l'agrégation permet de définir un prédicteur avec des propriétés théoriques remarquables. Le calcul numérique de cet estimateur repose sur une méthode de chaîne de Markov Monte Carlo dont il s'agit d'évaluer les performances. En particulier, il est important de borner le nombre de simulations nécessaires pour atteindre une précision comparable à celle de l'erreur de prédiction. Nous présentons un résultat général et son application à un modèle autorégressif. Des expériences numériques confirment les résultats attendus.

Abstract – In this work, we study the problem of forecasting a time series for a Causal Bernoulli Shifts (CBS) model. The aggregation technique provides an estimator with well established and excellent theoretical properties. However the numerical computation of this estimator relies on a Markov chain Monte Carlo method whose performances should be evaluated. In particular, it is crucial to bound the number of simulations needed to achieve a numerical precision of the same order as the prediction error. We present a fairly general result and its application to the autoregressive model. Some numerical experiments are carried out to support our results.

1 Introduction

An aggregation method consists in building a new estimator or a new predictor from a collection of different ones (typically via an integration), which is nearly as good as the best among them, given a risk criterion. The problem has been treated in different scenarios, with a few contributions in the dependent context, see [1], on which we shall rely in this work. The aggregate estimator is usually computed via a numerical procedure which raises an implementation issue. The most common application of Markov chain Monte Carlo methods is to deal with this kind of problems : numerically calculating multi-dimensional integrals.
We establish an oracle inequality in a quite general context : the Causal Bernoulli Shifts. For the practical aspect, a result of Łatuszyński [3], jointly with other properties of the basic MCMC algorithms that we used, allows us to control the error that we make by approximating the mean of a random variable by the empirical estimate obtained via MCMC. We show an oracle inequality that applies to the numerical approximation, instead of the theoretical aggregate estimator. Finally we treat the autoregressive process as an example and we present some numerical results.

2 Gibbs estimator

2.1 Statement of the problem and notation

Let us observe $(X_1, \dots, X_n)$ from a stationary time series $X = (X_t)_{t\in\mathbb{Z}}$ valued in $\mathbb{R}^r$ for some $r \geq 1$. Consider a family of predictors $\{f_\theta, \theta \in \Theta\}$. There exists $d \in \mathbb{N}^*$ such that for any $\theta \in \Theta$, $f_\theta : (\mathbb{R}^r)^d \to (\mathbb{R}^r)$ is a function from which we obtain :

$$\hat{X}_t^\theta \quad = \quad f_\theta\left(X_{t-1}, \dots, X_{t-d}\right) \ , \tag{1}$$

a possible forecasting of X_t according to θ. Let ℓ be a loss function ; we define the prediction risk as $R(\theta) = \mathbb{E}\left[\ell\left(\hat{X}_t^\theta, X_t\right)\right]$. It has a key role in the evaluation of the performance of any θ that we consider. As in general we do not know the distribution of the process, it is convenient to take into account the empirical version of the risk :

$$r_n(\theta; X_1, \dots, X_n) \quad = \quad \frac{1}{n-d}\sum_{t=d+1}^{n} \ell\left(\hat{X}_t^\theta, X_t\right) \ .$$

For the sake of simplicity we will identify
$r_n(\theta) \equiv r_n(\theta; X_1, \dots, X_n)$ but without forgetting that it is a random variable which depends on n observations of the series.
A probability measure π over Θ is labelled as the prior. It will serve to control the complexity of predictors in Θ and to construct one in particular, as detailed in the following.
For a measure ν and a measurable function h (called "energy function") such that $\nu[\exp(h)] = \int \exp(h) \ \mathrm{d}\nu < +\infty$, we denote by $\nu\{h\}$ the measure defined by :

$$\nu\{h\}(\mathrm{d}\theta) \quad = \quad \frac{\exp(h(\theta))}{\nu[\exp(h)]}\nu(\mathrm{d}\theta) \ . \tag{2}$$

It is a particular Gibbs measure where the inverse temperature is equal to -1.
Given a $\lambda > 0$, another temperature parameter, we define the Gibbs estimator as the expectation of a random variable drawn

under $\pi\{-\lambda r_n\}$:

$$\hat{\boldsymbol{\theta}}_{\lambda,n} = \pi\{-\lambda r_n\}[\mathrm{Id}] = \int_{\Theta} \boldsymbol{\theta}\pi\{-\lambda r_n(\cdot)\}(\mathrm{d}\boldsymbol{\theta}) \ . \tag{3}$$

So far we have presented a quite general framework : a time series that we aim to predict using a parameter $\boldsymbol{\theta}$, and a proposition of estimation for this $\boldsymbol{\theta}$. Let us introduce now the context in which this proposition will be studied.
A time series is defined as *Causal Bernoulli Shifts* (CBS) if it satisfies the representation :

$$X_t \quad = \quad H(\xi_t, \xi_{t-1}, \xi_{t-2}, \ldots), \forall t \in \mathbb{Z} \ , \tag{4}$$

where (ξ_s) is an i.i.d. sequence of $\mathbb{R}^{r'}$-valued r.v.s, for some $r' \geq 1$ and $H : \left(\mathbb{R}^{r'}\right)^{\mathbb{N}} \to \mathbb{R}^r$ is a function satisfying :

$$\|H(v) - H(v')\| \quad \leq \quad \sum_{j=0}^{\infty} a_j(H)\|v_j - v'_j\| \ , \tag{5}$$

for any $v = \left(v_j\right)_{j\in\mathbb{N}}, v' = \left(v'_j\right)_{j\in\mathbb{N}} \in \mathbb{R}^{r'}$,
where $\sum_{j=0}^{\infty} j a_j(H) < +\infty$.

2.2 Oracle inequality

In the context of CBS, adapting [1], some generic Oracle inequalities can be established on the aggregate predictor. For instance, for a bounded $\Theta \subset \mathbb{R}^p$, a uniform prior π yields that there exists a constant $\mathcal{E}$, such that for all $\epsilon > 0$, with probability at least $1-\epsilon$,

$$R\left(\hat{\boldsymbol{\theta}}_{\sqrt{n},n}\right) \quad \leq \quad \inf_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}) + \mathcal{E}\frac{\log^2(n)}{\sqrt{n}} + \frac{2}{\sqrt{n}}\log\left(\frac{1}{\epsilon}\right) \ . \tag{6}$$

The proof uses a Hoeffding type inequality for dependent sequences [6] and a lemma about the Legendre transform of the Kullback divergence function that we can find in [2].
Such inequality can be extended to more difficult situations which are not detailed here for brevity. Here however we shall focus on the fact that this inequality does not take into account the complexity to compute $\hat{\boldsymbol{\theta}}_{\sqrt{n},n}$.

3 MCMC approximation

We use the Metropolis - Hastings algorithm in order to compute the mean of a target probability whose density ρ, possibly unnormalised, is relatively easy to calculate. We will work over $\Theta \subseteq \mathbb{R}^p$ equipped with $\mathcal{T}$, the Borel σ- algebra. We will consider probability measures which are absolutely continuous, and have a known density with respect to the Lebesgue measure.
The Metropolis-Hastings algorithm generates a Markov chain $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_i\}_{i\geq 0}$ with the target distribution as a unique invariant measure, based on another Markov chain which serves as a proposal. We shall consider the two following classical setups for the proposal :

- The independent Hastings algorithm where the proposal is i.i.d. with density q such that $\frac{q(y)}{\rho(y)} \geq \beta, \forall y \in \Theta$ for some $\beta > 0$.
- The Metropolis-Hastings algorithm where the proposal is a Markov chain with conditional density kernel q on $\bar{\Theta}\times\bar{\Theta}$ such that $\beta = \inf_{x\in\bar{\Theta},y\in\bar{\Theta}} \frac{\rho(y)}{\rho(x)} \inf_{x\in\bar{\Theta},y\in\bar{\Theta}} q(x,y) > 0$.

From a simulated sequence $\Phi_1, \ldots, \Phi_m$, a numerical estimate of $\int x\rho(x)\,\mathrm{d}x$ is obtained by setting $\bar{\boldsymbol{\theta}}_m = \frac{1}{m}\sum_{i=0}^{m-1} \boldsymbol{\Phi}_i$. We have the following result.
Theorem 1. *Define*

$$M(\alpha,\gamma,\epsilon) = \frac{(2-\gamma)\,\mathrm{diam}(\Theta)}{2\alpha^2\epsilon\gamma} + \frac{1}{2}\sqrt{\left(\frac{(2-\gamma)\,\mathrm{diam}(\Theta)}{\alpha^2\epsilon\gamma}\right)^2 + \frac{4\mathrm{diam}(\Theta)}{\alpha^2\epsilon\gamma}} \ , \tag{7}$$

where $\mathrm{diam}(\Theta) = \sup_{x,y\in\Theta}\|x-y\|$. *Then, for any of two considered setups, for all* $m \geq M(\alpha,\beta,\epsilon)$, *with probability at least* $1-\epsilon$,

$$\left|\bar{\boldsymbol{\theta}}_m - \int x\rho(x)\,\mathrm{d}x\right| \quad \leq \quad \alpha \ . \tag{8}$$

By setting α appropriately, this result says how many iterations of the MCMC method are required in order to be reach a precisions of the same order as the prediction error enjoyed by the target Gibbs estimator.

4 Application to the AR(d) process with bounded innovations

4.1 Theoretical facts

We study the autoregressive model of order d or simply the AR(d), defined as the stationary solution of :

$$X_t \quad = \quad \sum_{j=1}^{d} \theta_j X_{t-j} + \sigma\xi_t \ , \tag{9}$$

where the ξ_t are i.i.d. with $\mathbb{E}\xi_t = 0$. We denote $s_d(\rho) = \left\{(\theta_1,\ldots,\theta_d) \ : \ 1 - \sum_k \theta_k z^k \neq 0 \text{ for } |z| < \rho^{-1}\right\}$ the set of $\boldsymbol{\theta}$s for which the autoregressive polynomial $\boldsymbol{\theta}(z) = 1 - \sum_k \theta_k z^k$ has all its roots outside the circle of radius ρ^{-1}. In this context, the CBS assumption implies that the true parameter $\bar{\boldsymbol{\theta}} = (\theta_1,\ldots,\theta_d) \in s_d(1)$. In the following we moreover assume that the innovations (ξ_t) have compact support (as in [1]) and denote by $\mathcal{B}$ a constant such that $X_t \in [-\mathcal{B},\mathcal{B}]$ for all t.
Since $s_d(1) \subseteq B_d\left(2^d-1\right)$ (see [5]), the prior π can be defined on $\Theta = s_d(1)$ or $B_d\left(2^d-1\right)$. These two possible priors are combined with two different proposals in the Metropolis-Hasting algorithm.
Uniform prior on $B_d\left(2^d-1\right)$
Suppose that π is the Lebesgue measure in $\Theta = B_d\left(2^d-1\right)$. As proposal chain we will use the uniform distribution over

the entire ball (independent of current state) and the truncated Gaussian one.

Uniform proposal : $q\left(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2\right) \propto 1_{B(2^p-1)}\left(\tilde{\boldsymbol{\theta}}_2\right)$. It can be shown in this case that the posterior distribution (which depends on λ and $X_1, \ldots, X_n$) satisfies the assumptions of Theorem 1 with

$$\beta_{\lambda,n} = \exp\left(-\lambda \mathcal{B}^2 \left(1 + \sqrt{d}\left(2^d - 1\right)\right)^2\right) .$$

Constrained random walk with Gaussian increment : $q\left(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2\right) \propto \exp\left(-\frac{n}{2}\|\tilde{\boldsymbol{\theta}}_2 - \tilde{\boldsymbol{\theta}}_1\|^2\right) 1_{\{\tilde{\boldsymbol{\theta}}_2 \in B(2^p-1)\}}$. Here we chose the variance of the increments so that the corresponding coefficient β can be guaranteed to be at least

$$\beta_{\lambda,n} = \left(\frac{n}{2\pi}\right)^{\frac{d}{2}} \exp\left(-2\left(2^d - 1\right)\left(\lambda 2^{d+1}\mathcal{B}^2 + \left(2^d - 1\right)n\right)\right) .$$

Pushforward measure on $s_d(1)$

It is more natural to choose $\Theta_d = s_d(1)$ than $\Theta_d = B_d\left(2^d - 1\right)$. There are several ways to define a prior on $s_d(1)$. We propose to use the function which maps the reciprocal roots of $\boldsymbol{\theta}(z)$ to the coefficients $\theta_1, \ldots \theta_d$.

Given $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$, $\boldsymbol{\theta}$ is then obtained by the transformation $\boldsymbol{\theta} = T(\boldsymbol{\lambda})$, defined as :

$$\theta_k = (-1)^{k-1} \sum_{1 \le i_1 < \ldots < i_k \le d} \lambda_{i_1} \ldots \lambda_{i_k} . \qquad (10)$$

However not any $\boldsymbol{\lambda} \in S^d$ (with $S = \{z \in \mathbb{C}, |z| > 1\}$) can be picked up, since $\boldsymbol{\theta}(z)$ is restricted to real polynomials. It is easy to deal with this picking couples of complex conjugates. A possible procedure is (and this is how the probability distribution is set) :

- $C \in \left[0, \left\lfloor \frac{d}{2} \right\rfloor + 1\right)$. $\lfloor C \rfloor \in \left\{0, 1, \ldots, \left\lfloor \frac{d}{2} \right\rfloor\right\}$ represents the number of couples of non-real roots.
- $U_i \in [-1, 1]$ is the value of λ_i when it is real, $i \in [1, d]$.
- $V_i \in (0, 1]$ is the modulus of λ_{2i-1} and λ_{2i} when they are not real, $i \in \left[1, \left\lfloor \frac{d}{2} \right\rfloor\right]$.
- $W_i \in (0, \pi)$ is the argument of λ_{2i-1} and $\bar{\lambda}_{2i}$ when they are not real, $i \in \left[1, \left\lfloor \frac{d}{2} \right\rfloor\right]$.

Transformation $\boldsymbol{\lambda} = L(C, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ is detailed below

$$\begin{aligned} \lambda_{2i-1} &= V_i \left[\cos(W_i) + i \sin(W_i)\right] \cdot 1_{\{C \ge i\}} + U_{2i-1} \cdot 1_{\{C \le i-1\}} \\ \lambda_{2i} &= V_i \left[\cos(W_i) - i \sin(W_i)\right] \cdot 1_{\{C \ge i\}} + U_{2i} \cdot 1_{\{C \le i-1\}} \end{aligned}$$

Let $\Omega = \left[0, \left\lfloor \frac{d}{2} \right\rfloor + 1\right) \times [-1, 1]^d \times (0, 1]^{\lfloor \frac{d}{2} \rfloor} \times (0, \pi)^{\lfloor \frac{d}{2} \rfloor}$ the space to which variables $(C, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ belong. The transformation $F(\omega) = T \circ L(\omega)$ covers the whole set $s_d(1)$. With this map, it is possible to define a prior measure on $s_d(1)$ from a measure on Ω (precisely the pushforward measure).

We run the MCMC algorithm precisely on Ω. The convergence properties are inherited by the underlying chain on $s_d(1)$.

Here again, as proposal we will use the uniform distribution over Ω (independent of current state) and the truncated Gaussian one.

Uniform proposal : $\bar{q}(\tilde{\omega}_1, \tilde{\omega}_2) \propto 1_\Omega(\tilde{\omega}_2)$. The posterior distribution satisfies the assumptions of Theorem 1 with

$$\beta_{\lambda,n} = \exp\left(-\lambda \mathcal{B}^2 \left(1 + \sqrt{d}\left(2^d - 1\right)\right)^2\right) .$$

Constrained random walk with Gaussian increment : $\bar{q}(\tilde{\omega}_1, \tilde{\omega}_2) \propto \exp\left(-\frac{n}{2}\|\tilde{\omega}_2 - \tilde{\omega}_1\|^2\right) 1_\Omega(\tilde{\omega}_2)$. We chose the variance of the increments such that β is be guaranteed to be at least

$$\beta_{\lambda,n} = \left(\frac{n}{2\pi}\right)^{1+p+2\lfloor \frac{d}{2} \rfloor} \exp\left(-2\left(\lambda 2^{d+1}\left(2^d - 1\right)\mathcal{B}^2 + \left(2^{1+d+2\lfloor \frac{d}{2} \rfloor} - 1\right)^2 n\right)\right) .$$

Each one of the four above setups give us a $\bar{\boldsymbol{\theta}}_{\lambda,n,m}$ for each m (MCMC iteration number) which can be used for approximating $\hat{\boldsymbol{\theta}}_{\lambda,n}$. In all four cases, the following result is obtained.

Theorem 2. *There exists a constant $\mathcal{F}$ such that for all $m \ge M\left(\frac{\log(n)}{n}, \beta_{\sqrt{n},n}, \epsilon\right)$, with M defined as in Theorem 1,with probability at least $(1-\epsilon)^2$,*

$$R\left(\bar{\boldsymbol{\theta}}_{\sqrt{n},n,m}\right) \le \inf_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}) + \mathcal{F}\frac{\log^2(n)}{\sqrt{n}} + \frac{2}{\sqrt{n}} \log\left(\frac{1}{\epsilon}\right) , (11)$$

where the value of $\beta_{\sqrt{n},n}$ is detailed above depending on the specific scheme.

4.2 Numerical work

We iterate the algorithm with $m = 10000$ times for the four schemes with $d = 8$. It is interesting to note that the prediction error decreases as the number n of observations grows but then stabilizes because the number m of simulations in the MCMC algorithm remains fixed. Indeed, to guaranty the correct error order of magnitude for a fixed n, this number m should be at least $M\left(\frac{\log(n)}{n}, \beta_{\sqrt{n},n}, \epsilon\right)$, which diverges exponentially fast as n increases.

The following graphs resume the behavior of the algorithm for 20 time series in each case.

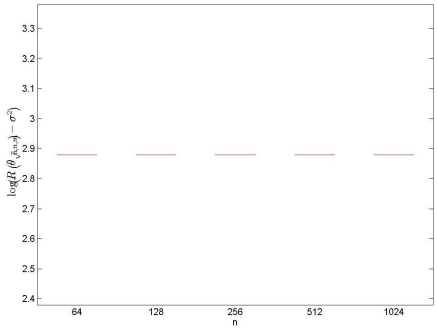

FIG. 1: Uniform proposal, $d = 8$, $\Theta = B_8\left(2^8 - 1\right)$.

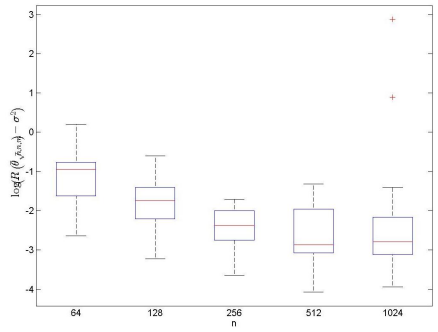

FIG. 2: Gaussian proposal, $d = 8$, $\Theta = B_8\left(2^8 - 1\right)$

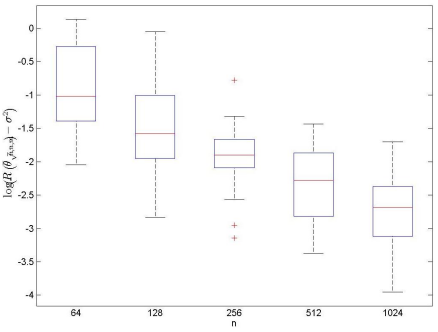

FIG. 3: Uniform proposal, $d = 8$, $\Theta = s_8(1)$

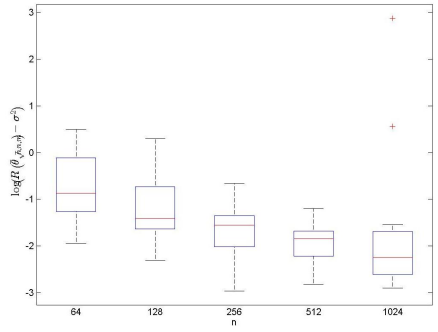

FIG. 4: Gaussian proposal, $d = 8$, $\Theta = s_8(1)$

Figure 1 shows that the uniform proposal in $\Theta = B_d\left(2^d - 1\right)$ leads to unsatisfactory results. Indeed, in this case, the MCMC procedure is unable to get close to the target distribution, whatever the number of simulations is. The main reason is that for $d = 8$ the domain $B_8\left(2^8 - 1\right)$ is too wide so that the domain of interest for the parameter, where the posterior distribution is concentrated, is not explored by the chain. Other situations in Figures 2–4 show good results. However, using (7) and the obtained expressions of β yields to the following equivalence for the minimal number of iterations m guarantying a correct prediction error as a function of the number of observations n :

$$m \geq C_1(d)\,\frac{\log^2 n}{2n^2\epsilon}\exp\left(C_2(d)\,\sqrt{n}\right),$$

where C_1 and C_2 are positive functions.
Hence, for a large number of iterations, the good performance of the Gibbs estimator requires a very costly numerical procedure, making its practical benefit doubtful.

5 Conclusion

The use of aggregate estimators determining a parameter with almost minimal prediction risk has been considered in this work in the context of stationary time series. An approximation of the Gibbs estimator can be computed using the Metropolis Hastings algorithm. This allows us to obtain guaranties on the numerical approximation, that we illustrated by a new oracle inequality. However, this inequality indicates that such an approach is sensible only for a reasonably small number of observations, since the number of iterations of the numerical method needed to achieve the correct prediction error increases exponentially fast.

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).

References

[1] Pierre Alquier and Olivier Wintenberger. *Model selection for weakly dependent time series forecasting*. Bernoulli, 18(3) : 883-913, 2012.

[2] Olivier Catoni and Jean Picard. *Statistical Learning Theory and Stochastic Optimization: Ecole D'été de Probabilités de Saint-Flour XXXI-2001*. Number n 1851 in Ecole d'Été de Probabilités de Saint-Flour. Springer-Verlag, 2004.

[3] Krzysztof Łatuszyński and Wojciech Niemiro. *Rigorous confidence bounds for MCMC under a geometric drift condition*. J. Complexity, 27(1) : 23-38, 2011.

[4] K. L. Mengersen and R. L. Tweedie. *Rates of convergence of the Hastings and Metropolis algorithms*. Ann. Statist., 24(1) : 101-121, 1996.

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

[6] Emmanuel Rio. *Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes*. Comptes Rendus de l'Académie des Sciences Series I Mathematics, 330(10):905908, 2000.