PAC-Bayesian bounds for dependent observations

Joint work with Pierre Alquier & Olivier Wintenberger

Xiaoyin LI
Xiaoyin.li@u-cergy.fr

University of Cergy-Pontoise

TEST February 2013

Outline

A problem of statistical inference

Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary time series in $\mathbb{R}^p$. We want to learn to forecast the series from the observations $X_1, \dots, X_n$.

We have a family of predictors

$$\mathcal{F} = \left\{ f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p \text{ mesurable}, \theta \in \Theta \right\}$$

We consider a model-selection type approach:

$$\Theta = \bigcup_{j=1}^{m} \Theta_j.$$

Objective: find a $\theta \in \Theta$ such that $f_\theta(X_{t-1}, \dots, X_{t-k})$ is a good prediction of X_t.

Familes of classical predictor

Definition

For any $\theta \in \Theta$,

$$\hat{X}_t^{\theta} = f_\theta(X_{t-1}, \ldots, X_{t-k}).$$

Linear auto-regressive class of predictors:

$$f_\theta(X_{t-1}, \ldots, X_{t-k}) = \theta_0 + \sum_{j=1}^{k} \theta_j X_{t-j}.$$

Non-parametric auto-regression predictors:

$$f_\theta(X_{t-1}, \ldots, X_{t-k}) = \sum_{i=1}^{j} \theta_i \varphi_i(X_{t-1}, \ldots, X_{t-k}).$$

Measure of risk

Let ℓ be a loss function: $\ell(\hat{X}_t^\theta, X_t) \geq 0$ measures the forecasting error of predictor θ at time t.

The prevision risk is defined as:

$$R(\theta) = \mathbb{E}\left[\ell\left(\hat{X}_t^\theta, X_t\right)\right] \textit{ is unknown.}$$

On the other hand, we observe the empirical risk:

$$r_n(\theta) = \frac{1}{n-k}\sum_{i=k+1}^{n} \ell\left(\hat{X}_i^\theta, X_i\right).$$

with $\mathbb{E}\left[r_n(\theta)\right] = R(\theta)$.

Objective

Build a parameter θ on the basis of observations $X_1, \ldots, X_n$ such that :

$$R(\theta) \simeq \inf_{\theta \in \Theta} R(\theta).$$

More precisely,

$$R(\theta) \leq \left\{ \inf_{\theta \in \Theta} R(\theta) + \Delta(n, \Theta) \right\}$$

Prior and Estimators

Let $\mathcal{M}_{+}^{1}(\Theta)$ denote the set of all probability measures on $(\Theta, \mathcal{T})$. Let us take $\pi \in \mathcal{M}_{+}^{1}(\Theta)$, the prior.

Definition (Gibbs Estimators)

We put, for any $\lambda > 0$,

$$\hat{\theta}_{\lambda} = \int_{\Theta} \theta \hat{\rho}_{\lambda}(\mathrm{d}\theta)$$

where

$$\hat{\rho}_{\lambda}(\mathrm{d}\theta) = \frac{e^{-\lambda r_n(\theta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda r_n(\theta')}\pi(\mathrm{d}\theta')}.$$

Outline

Overview of the results

The idea is that the risk of the Gibbs estimator will be close to $\inf_\theta R(\theta)$ up to a small remainder term Δ called the rate of convergence.

For the sake of simplicity, let $\overline{\theta} \in \Theta$ be such that

$$R(\overline{\theta}) = \inf_\theta R(\theta).$$

We want to prove that our estimators satisfy, for any $\varepsilon > 0$,

$$\mathbb{P}\left(R\left(\hat{\theta}\right) \leq R(\overline{\theta}) + \Delta(n, \lambda, \pi, \varepsilon)\right) \geq 1 - \varepsilon$$

where $\Delta(n, \lambda, \pi, \varepsilon) \to 0$ as $n \to \infty$ for some $\lambda = \lambda(n)$.

Assumptions

We assume:

- $(X_t)_{t\in\mathbb{Z}}$ is bounded and θ-dependent, *ie* a.s. $\|X_0\|_\infty \le \mathcal{B} < \infty$ and $\theta_{\infty,n}(1) \le \mathcal{C} < \infty$.
- $\ell(x,x') = g(x-x')$ for some convex function g and g is K-Lipschitz.
- for any f,

$$\left\|f_\theta\left(x_1,\ldots,x_k\right) - f_\theta\left(y_1,\ldots,y_k\right)\right\| \le \sum_{j=1}^{k} a_j\left(\theta\right)\left\|x_j - y_j\right\|.$$

$$L := \sup_{\theta\in\Theta}\sum_{j=1}^{k} a_j\left(\theta\right)$$

- $k = k(n) \le n/2$.

A basic PAC-Bayesian Bound

Theorem

for any $\lambda > 0$, *for any* $\varepsilon > 0$,

$$\mathbb{P}\left\{\int R(\hat{\theta}_\lambda) \leq \inf_\rho \left[\int R \mathrm{d}\rho + \frac{2\lambda\kappa_n^2}{n} + 2\frac{\mathcal{K}(\rho,\pi) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right]\right\} \geq 1 - \varepsilon$$

where $\kappa_n := \sqrt{2}K(1+L)(\mathcal{B} + \theta_{\infty,n}(1))$.

Example of loss function

1. Absolute loss $\ell(x, x') = \|x - x'\|$ with $K = 1$ (Alquier and Wintenberger).
2. Quadratic loss $\ell(x, x') = \|x - x'\|^2$ with $K = 4\mathcal{B}$ (Meir).
3. Quantile loss

$$\ell_\tau(x, y) = \begin{cases} \tau\,(x - y)\,, & \text{if } x - y > 0 \\ -\,(1 - \tau)\,(x - y)\,, & \text{otherwise} \end{cases}.$$

with $K = \max(\tau, 1 - \tau) \leq 1$(Alquier and Li).

θ weak dependent coefficient

Introduced by Doukhan and Louhichi (SPA, 1999). θ coefficient is as follow:

$$\theta_\infty(\mathfrak{S}, Z) = \sup_{f \in \Lambda_1^q} \left\| \mathbb{E}\left[f(Z)|\mathfrak{S}\right] - \mathbb{E}\left[f(Z)\right] \right\|_\infty$$

where

$$\Lambda_1^q = \left\{ f : (\mathbb{R}^p)^q \to \mathbb{R}, \quad \frac{|f(z_1, \ldots, z_q) - f(z_1', \ldots, z_q')|}{\sum_{j=1}^q \|z_j - z_j'\|} \le 1 \right\},$$

and that

$$\theta_{\infty,k}(1) := \sup_{p < j_1 < \ldots < j_\ell, 1 \le \ell \le k} \left\{ \theta_\infty(\sigma(X_t, t \le p), (X_{j_1}, \ldots, X_{j_\ell})) \right\}.$$

Example of θ-weakly dependent series

For any series

$$X_t = F(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots)$$

with ε_i iid, upper bounded by b, and

$$\|F(x_1, x_2, \ldots) - F(x'_1, x'_2, \ldots)\| \leq \sum_{j=1}^{\infty} a_j \left\|x_j - x'_j\right\|$$

we have

$$\theta_{\infty,n}(1) \leq 2b \sum_{j=1}^{\infty} j a_j.$$

Examples of θ-weakly dependent series

Let us remind the ϕ-mixing coefficient:

$$\phi(r) = \sup_{A \in \sigma(X_t, t \leq 0), B \in \sigma(X_t, t \geq r)} |P(B|A) - P(B)|$$

Then, for (X_t) upper bounded by $\mathcal{B}$, we have

$$\theta_{\infty,n}(1) \leq 2\mathcal{B} \sum_{r=1}^{n} \phi(r).$$

Reminder

Theorem

for any $\lambda > 0$*, for any* $\varepsilon > 0$*,*

$$\mathbb{P}\left\{\int R(\hat{\theta}_\lambda) \leq \inf_\rho \left[\int R\mathrm{d}\rho + \frac{2\lambda\kappa_n^2}{n} + \frac{2\mathcal{K}(\rho,\pi) + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right]\right\} \geq 1 - \varepsilon$$

where $\kappa_n := \sqrt{2}K(1+L)(\mathcal{B} + \theta_{\infty,n}(1))$.

Toy example: $\mathrm{card}(\Theta) = M < \infty$

We take π as the uniform distribution:

$$\begin{aligned} R(\hat{\theta}_\lambda) &\leq \inf_\rho \left\{ \int R \mathrm{d}\rho + \frac{2\lambda\kappa_n^2}{n} + \frac{2\mathcal{K}(\rho,\pi) + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\} \\ &\leq \inf_\theta \left\{ R(\theta) + \frac{2\lambda\kappa_n^2}{n} + 2\frac{\log(M) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\} \end{aligned}$$

Theorem

Assume that $\mathrm{card}(\Theta) = M$ and let π be the uniform probability distribution on Θ. Then the oracle inequality is satisfied for any $\lambda > 0$, $\varepsilon > 0$ with probability at least $1 - \varepsilon$

$$R(\hat{\theta}_\lambda) \leq \inf_\theta \left\{ R(\theta) + \frac{2\lambda\kappa_n^2}{n} + 2\frac{\log(M) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}$$

Toy example: $\mathrm{card}(\Theta) = M < \infty$

The choice $\lambda = \sqrt{n \log(M)}/\kappa_n$ yields the oracle inequality:

$$R(\hat{\theta}_\lambda) \leq \inf_\theta R + 2\kappa_n \sqrt{\frac{2\log(M)}{n}} + \frac{2\kappa_n \log\left(\frac{2}{\varepsilon}\right)}{n \log(M)}$$

Bad news: the optimal λ depends on $\theta_{\infty,n}(1)$, unknown in practice.

Linear autoregressive predictors

Consider the linear autoregressive model of $\mathrm{AR}(k)$ predictors

$$f_\theta(x_{t-1},\ldots,x_{t-k}) = \sum_{j=1}^{k} \theta_j x_{t-j}$$

with $\theta \in \Theta = \{\theta \in \mathbb{R}^k, \|\theta\| \leq L\}$
We take π uniform and we restrict ρ to the uniform distributions on $\{\theta' : \|\theta - \theta'\| \leq \delta\}$.

$$\begin{aligned} R(\hat{\theta}_\lambda) &\leq \inf_\rho \left\{ \overbrace{\int R \mathrm{d}\rho_{\delta,\theta}}^{\leq R(\theta)+\delta\mathcal{B}} + \frac{2\lambda\kappa_n^2}{n} + \overbrace{\frac{2\mathcal{K}(\rho_{\delta,\theta},\pi) + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda}}^{2\frac{k\log(\frac{L+1}{\delta})+\log\left(\frac{2}{\varepsilon}\right)}{\lambda}} \right\}. \\ &\leq \inf_\theta \left\{ R(\theta) + \delta\mathcal{B} + \frac{2\lambda\kappa_n^2}{n} + 2\frac{k\log(\frac{L+1}{\delta})+\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}. \end{aligned}$$

Linear autoregressive predictors

We can now take $\delta = \frac{k}{\lambda \mathcal{B}}$,

$$R(\hat{\theta}_\lambda) \leq \inf_\theta R + \frac{2\lambda\kappa_n^2}{n} + 2\frac{k \log\left(\frac{K\mathcal{B}(L+1)\sqrt{e}\lambda}{k}\right) + \log(2/\varepsilon)}{\lambda}.$$

The optimal inverse temperature parameter is $\lambda = \frac{\sqrt{nk}}{\kappa_n}$.

Theorem

Let π be the uniform probability distribution on Θ. Then the oracle inequality is satisfied for any $\lambda > 0$, $\varepsilon > 0$ with high probability at least $1 - \varepsilon$

$$R(\hat{\theta}_\lambda) \leq \inf_\theta R + \frac{2\lambda\kappa_n^2}{n} + 2\frac{k \log\left(\frac{K\mathcal{B}(L+1)\sqrt{e}\lambda}{k}\right) + \log(2/\varepsilon)}{\lambda}.$$

General parametric model

We state a general result about finite-dimensional families of predictors.
In general, one can always consider

$$\rho_\delta(\mathrm{d}\theta) \propto \pi(\mathrm{d}\theta)\mathbf{1}(R(\theta) - \inf_\Theta R \leq \delta).$$

Let π be uniform and ρ be the uniform distributions on $\{\theta : R(\theta) - \inf_\Theta R(\theta) \leq \delta\}$.
we assume $\dim(\Theta, \pi) := \sup \frac{-\log \pi\{\theta : R(\theta) - \inf_\Theta R \leq \delta\}}{\log \lambda} = D$

$$R(\hat{\theta}_\lambda) \leq \inf_\rho \left\{ \underbrace{\int R \mathrm{d}\rho_{\delta,\theta}}_{\leq R(\theta)+\delta} + \frac{2\lambda\kappa_n^2}{n} + \underbrace{\frac{2\mathcal{K}(\rho_{\delta,\theta}, \pi) + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda}}_{2\frac{-\log\left(\pi\left\{\theta : R(\theta) - \inf_\Theta R \leq \delta\right\}\right) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}} \right\}.$$

$$\leq \inf_\Theta R + \delta + \frac{2\lambda\kappa_n^2}{n} + 2\frac{d\log(D/\delta) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}$$

The infimum is reached for $\delta = d/\lambda$ and we have:

$$R(\hat{\theta}_\lambda) \leq R(\bar{\theta}) + 2\frac{\lambda\kappa_n^2}{n} + 2\frac{d\log(D\sqrt{e}\lambda/d) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}$$

Theorem

Let π be the uniform probability distribution on Θ. Then the oracle inequality is satisfied for any $\lambda > 0$, $\varepsilon > 0$ with high probability at least $1 - \varepsilon$

$$R\left(\hat{\theta}_\lambda\right) \leq R(\bar{\theta}) + \frac{2\lambda\kappa_n^2}{n} + 2\frac{d\log\left(\frac{D\sqrt{e}\lambda}{d}\right) + \log\left(\frac{2}{\varepsilon}\right)}{\lambda}.$$

Tools used in the proofs

Lemma (Donsker-Varadhan variational formula)

For any $\pi \in \mathcal{M}_+^1(E)$, for any measurable upper-bounded function h, we have:

$$\int \exp(h)\mathrm{d}\pi = \exp\left(\sup_{\rho \in \mathcal{M}_+^1(E)} \left(\int h\mathrm{d}\rho - \mathcal{K}(\rho,\pi)\right)\right)$$

Theorem (Rio,2000)

For any f that is a function 1-Lipshitz

$$\forall t \geq 0, \quad \mathbb{E}[e^{tf(X_1,\ldots,X_n)-t\mathbb{E}[f(X_1,\ldots,X_n)]}] \leq e^{\frac{nt^2(\mathcal{B}+\theta_{\infty,n}(1))^2}{2}}$$

Outline

The context

Objective: at each quarter t, predict the flash estimate of GDP growth: $\Delta\mathrm{GDP}_t$.

Available information:

- $\Delta\mathrm{GDP}_{t'}$, for all $t' < t$
- $I_{t'}$, for all $t' < t$, I_{t-1} is the climate indicator available to the INSEE at time t.
- The observation period is 1988-Q1 (1st quarter of 1988) to 2011-Q3.

Quantile loss function

We define $X_t = (\Delta\mathrm{GDP}_t, I_t)' \in \mathbb{R}^2$. Following Cornec(CIRET conference 2010), we consider predictors of the form:

$$f_\theta(X_{t-1}, X_{t-2}) = \theta_0 + \theta_1 \Delta\mathrm{GDP}_{t-1} + \theta_2 I_{t-1} + \theta_3 (I_{t-1} - I_{t-2})|I_{t-1} - I_{t-2}|$$

These family of predictors allow to obtain a forecasting as precise as the INSEE one.

We use the quantile loss function :

$$\ell_\tau((\Delta\mathrm{GDP}_t, I_t), (\Delta'\mathrm{GDP}_t, I'_t)) = \begin{cases} \tau\left(\Delta\mathrm{GDP}_t - \Delta'\mathrm{GDP}_t\right), & \text{if } \Delta\mathrm{GDP}_t - \Delta'\mathrm{GDP}_t > 0 \\ -\left(1-\tau\right)\left(\Delta\mathrm{GDP}_t - \Delta'\mathrm{GDP}_t\right), & \text{otherwise.} \end{cases}$$

Results: GDP forecasting

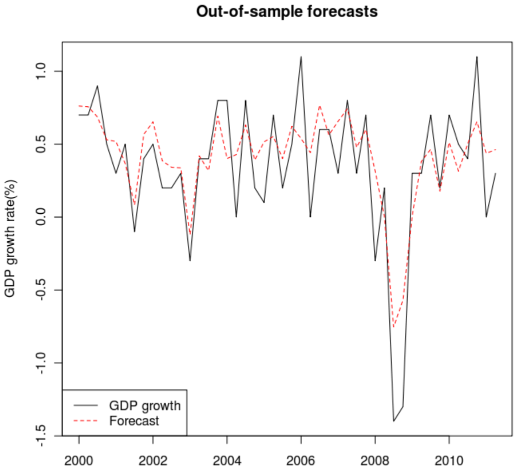

Figure: French GDP online prediction using the quantile loss function with $\tau = 0.5$.

Results: Confidence intervals

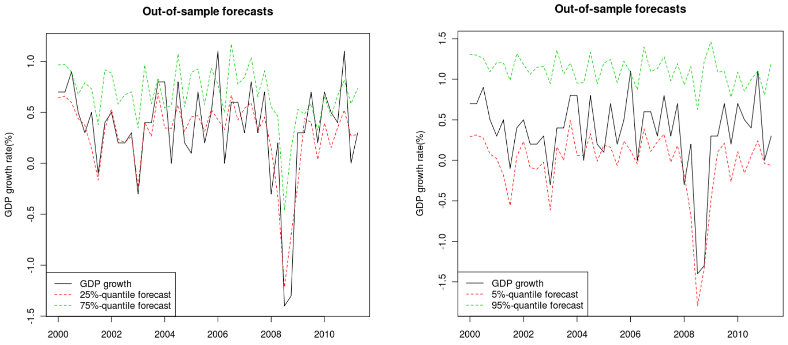

Figure: French GDP online 50%-confidence intervals (left) and 90%-confidence intervals (right).

Reference

- Alquier, P., Li, X. and Wintenberger, O.(2012). *Prediction of time series by statistical learning: general losses and fast rates.Preprint arXiv:1211.1847*
- Alquier, P. and Li, X. (2012). *Prediction of Quantiles by Statistical Learning and Application to GDP Forecasting. in Proceeding of DS'12, Springer LNAI n.7569. pp.22-36*
- Alquier, P. and Wintenberger, O. (2012). *Model selection for weakly dependent time series forecasting. Bernoulli, vol.18, no.3.pp.883-913*
- Rio, E.(2000). *Inégalités de Hoeffding pour les fonctions lipshitziennes de suites dépendantes. CRAS série I, vol 330, pp 905-908*
- Dedecker, J., Doukhan, P., Lang, G.,Léon, J.R., Louhichi, S. and Prieur, C. (2007). *Weak dependence, examples and applications. Springer Lecture Notes in Mathematics n.190*
- Catoni, O.(2004). *Statistical learning theory and stochastic optimization. Saint-Flour 2001 lecture notes. J.Picard Ed. Springer Lecture Notes in Mathematics.*