PAC-Bayesian Online Clustering

Le LI, Benjamin GUEDJ, Sébastien LOUSTAU

Université d'Angers & Inria Lille & iAdvize

April 20, 2016

Table of contents

Online learning

A **blackbox** reveals: $z_t \in \mathcal{Z}, t = 1, 2, \ldots, T$ (denoted by $(z_t)_{1:T}$).
A **forecaster** predicts: $\hat{z}_t$, based on $(z_s)_{1:s}, s < t$ and other available information.
Sequence z_t: not realization of stochastic process.

1. **Prediction with experts' advice**:
   $\{f_{e,t} \in \mathcal{D}, e \in \mathcal{E}\}$, where $f_{e,t}$ prediction of expert e at time t, $\mathcal{E}$ finite.
   Task: build $\hat{z}_t$, based on $\{f_{e,s} \in \mathcal{D}, e \in \mathcal{E}\}_{1:t}$, and $(z_s)_{1:t-1}$ such that, uniformly in $(z_t)_{1:T}$,

$$\sum_{t=1}^{T} \ell(\hat{z}_t, z_t) - \min_{e \in \mathcal{E}} \left\{ \sum_{t=1}^{T} \ell\left(f_{e,t}, z_t\right) \right\} \leq \Delta_T(\mathcal{E}),$$

   where ℓ is the loss function and $\Delta_T(\mathcal{E})$ remainder term. $\Delta_T(\mathcal{E})$ order $\sqrt{\ln|\mathcal{E}|T}$ if $|\mathcal{E}| < \infty$, ℓ bounded, convex in its first argument (CBL 06).

Online learning

2. **Online regression**:
   $z_t = (x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$. Predict $\hat{y}_t$ based on x_t and past observations.
   Task: build $\hat{y}_t$ such that, uniformly in $(x_t, y_t)_{1:T}$,

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \ell(\langle \theta, x_t \rangle, y_t) \right\} \leq \Delta_T(d),$$

   where $\langle x, y \rangle$ dot product in $\mathbb{R}^d$, $\Delta_T(d)$ remainder term. $\Delta_T(d)$ of order $d \log T$ (Vovk 2001). In high-dimension, growing logarithmically with d and T (Gerchinovitz 2011).

Online Learning

3 Online clustering

Let $(x_t)_{1:T}, x_t \in \mathbb{R}^d$, online dataset.

Task: Learn a partition $\hat{\mathbf{c}}_t = (\hat{c}_{t,1}, \hat{c}_{t,2}, \ldots, \hat{c}_{t,K_t}) \in \mathscr{C}$, where K_t time-dependent; $\mathscr{C} = \cup_{k=1}^{p} \mathscr{C}(k,R)$, partition space, $\mathscr{C}(k,R) \in \mathbb{R}^{dk}$, $R > 0$.

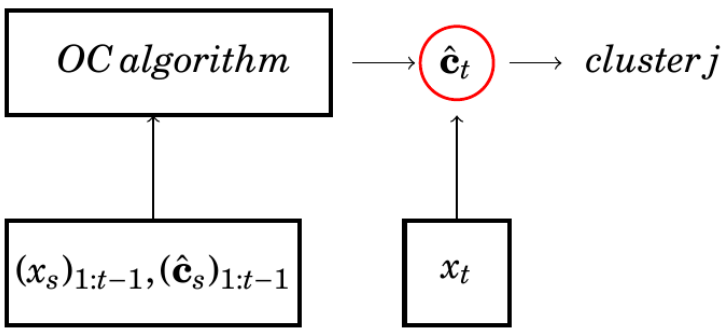

Figure: General structure of OC

- for any $\hat{\mathbf{c}}_t \in \mathscr{C}$, the loss $\ell(\hat{\mathbf{c}}_t, x_t) = \min_{1 \le k \le K_t} ||x_t - \hat{c}_{t,k}||_2^2$.
- cluster index j: $j = \arg\min_{1 \le k \le K_t} ||x_t - \hat{c}_{t,k}||_2^2$.

Pseudo-code (Audibert 2009, Loustau 2014)

Parameters p, $\lambda > 0$, $R > 0$, a well-defined prior π on $\mathscr{C}$.

1. At the beginning, draw $\hat{\mathbf{c}}_1 \sim \mathrm{d}\hat{\rho}_1 = \mathrm{d}\pi$.
2. For $t = 1, \ldots, T-1$, get x_t and compute recursively for $\mathbf{c} \in \mathscr{C}$

$$S_t(\mathbf{c}) = S_{t-1}(\mathbf{c}) + \ell(\mathbf{c}, x_t) + \frac{\lambda}{2}\left[\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)\right]^2,$$

   where $S_0(\mathbf{c}) = 0$ for any $\mathbf{c} \in \mathscr{C}$.
3. Update Gibbs quasi-posterior

$$d\hat{\rho}_{t+1}(\mathbf{c}) = \frac{\exp(-\lambda S_t(\mathbf{c}))}{\int \exp(-\lambda S_t(\mathbf{c}))\mathrm{d}\pi(\mathbf{c})}\mathrm{d}\pi(\mathbf{c})$$

4. Draw $\hat{\mathbf{c}}_{t+1} \sim \mathrm{d}\hat{\rho}_{t+1}$.

Regret bound

Theorem (Loustau 2014, Li 2016)

For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$ and $p \geq 1$, if $R \geq \max_{t=1,\dots,T} |x_t|_2$, $\lambda = (d+2)/(2\sqrt{T}R^2)$, then

$$\sum_{t=1}^{T} \mathbb{E}[\ell(\hat{\mathbf{c}}_t, x_t)] < \inf_{1 \leq k \leq p} \left\{ \inf_{\mathbf{c} \in \mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + C \times k\sqrt{T}\log T \right\}.$$

If there exists $k^\star \in \{1,\dots,p\}$ and $\mathbf{c}^\star \in \mathscr{C}(k,R)$ which achieves the infimum on the right-hand side, then,

$$\sum_{t=1}^{T} \mathbb{E}[\ell(\hat{\mathbf{c}}_t, x_t)] - \sum_{t=1}^{T} \ell(\mathbf{c}^\star, x_t) \leq C \times k^\star \sqrt{T}\log T.$$

where C depends on d, R, $\log p$.

Key facts for the proof

- Duality formula

$$-\ln \int_\Theta \exp(-h)\mathrm{d}\pi = \inf_{\rho \in \mathscr{P}_\pi(\Theta)} \left\{ \int_\Theta h \mathrm{d}\rho + \mathscr{K}(\rho,\pi) \right\}$$

where $\mathscr{K}(\rho,\pi)$ Kullback-Leibler divergence. $h : \Theta \longrightarrow \mathbb{R}$ bounded measurable. Gibbs posterior: $\mathrm{d}\hat{\rho} = \frac{\exp(-h)}{\int_\Theta \exp(-h)\mathrm{d}\pi}\mathrm{d}\pi$.

- Online variance inequality (Audibert 2009)
$\forall \rho \in \mathscr{P}(\Theta), \exists \hat{\rho}, \forall z \in \mathscr{Z}, \lambda > 0,$

$$\mathbb{E}_{g' \sim \hat{\rho}} \log \mathbb{E}_{g \sim \rho} \exp\left(\lambda \left[\ell(g',z) - \ell(g,z) - \frac{\lambda}{2}(\ell(g,z) - \ell(g',z))^2 \right] \right) \leq 0.$$

where ℓ don't need to be convex in its first argument.

Implementation

The target Gibbs $\hat{\rho}_{t+1}$ is defined on a massive and complex space $\mathscr{C} = \cup_{k=1}^{p} \mathscr{C}(k,R)$. Direct sampling is not possible. Hence resorting to,

- Reversible Jump Monte Carlo Markov Chain (RJMCMC) (Green 1995).
- Trans-dimensional MCMC (Guedj and Alquier 2013).

Algorithm

Algorithm 2 PACO

Initialization: (λ_t)
For $t \in [\![1, T]\!]$
Initialization: $\left(k^{(0)}, \mathbf{c}^{(0)}\right) \in [\![1, p]\!] \times \mathbb{R}^{dk^{(0)}}$
For $n \in [\![0, N-1]\!]$
    Given $k^{(n)}$, draw $k' \sim q(k^{(n)}, \cdot)$, where $q(k^{(n)}, \cdot)$ is a conditional distribution on $[\![k^{(n)} - 1, k^{(n)} + 1]\!]$.
    Let $\mathfrak{c}_{k'} \leftarrow$ standard k-means output with k' centers.
    Let $\tau_{k'} = 1/\sqrt{pt}$.
    Sample $v_1 \sim \rho_{k'}(\cdot, \mathfrak{c}_{k'}, \tau_{k'})$, where $\rho_{k'}(\cdot, \mathfrak{c}_{k'}, \tau_{k'})$ is a distribution on $\mathbb{R}^{dk'}$ with parameters $\mathfrak{c}_{k'}$ and $\tau_{k'}$.
    Let $(v_2, \mathbf{c}') = g(v_1, \mathbf{c}^{(n)})$, where $g : (x, y) \in \mathbb{R}^{dk'} \times \mathbb{R}^{dk^{(n)}} \to (y, x) \in \mathbb{R}^{dk^{(n)}} \times \mathbb{R}^{dk'}$ is a bijection with continuous derivative.
    Accept the move $(k^{(n)}, \mathbf{c}^{(n)}) = (k', \mathbf{c}')$ with probability

$$\alpha\left[(k^{(n)}, \mathbf{c}^{(n)}), (k', \mathbf{c}'))\right] = \min\left\{1, \frac{\hat{\rho}_t(\mathbf{c}')q(k', k^{(n)})\rho_{k^{(n)}}(v_2, \mathfrak{c}_{k^{(n)}}, \tau_{k^{(n)}})}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)}, k')\rho_{k'}(v_1, \mathfrak{c}_{k'}, \tau_{k'})} \left|\frac{\partial g(v_1, \mathbf{c}^{(n)})}{\partial v_1 \partial \mathbf{c}^{(n)}}\right|\right\}$$

    Else $(k^{(n+1)}, \mathbf{c}^{(n+1)}) = (k^{(n)}, \mathbf{c}^{(n)})$.
End for
Let $\hat{\mathbf{c}}_t = \mathbf{c}^{(N)}$.
End for

Numerical assessment

1. Experiment: 4 Gaussian groups in dimension 2 with mean vector $(0,0),(-4,-1),(0,7),(5,2)$ and identity covariance matrix.

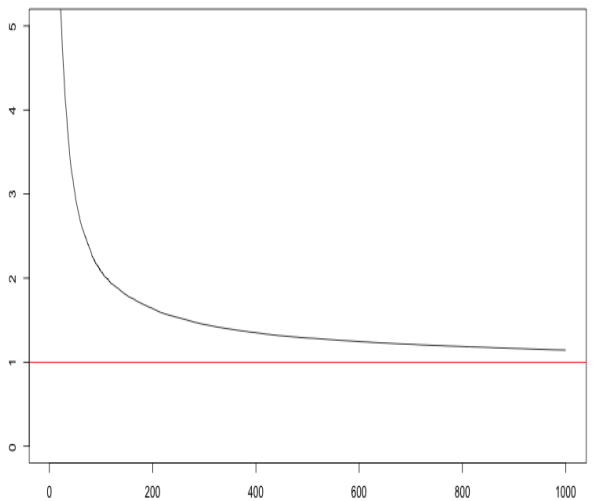

Figure: black line: ECL/OCL; red line: horizontal 1

Numerical assessment

2. Experiment: 10 mixed groups in dimension 2, where the true number of groups will increase of 1 unit every 20 time steps and reach 10 at the end.

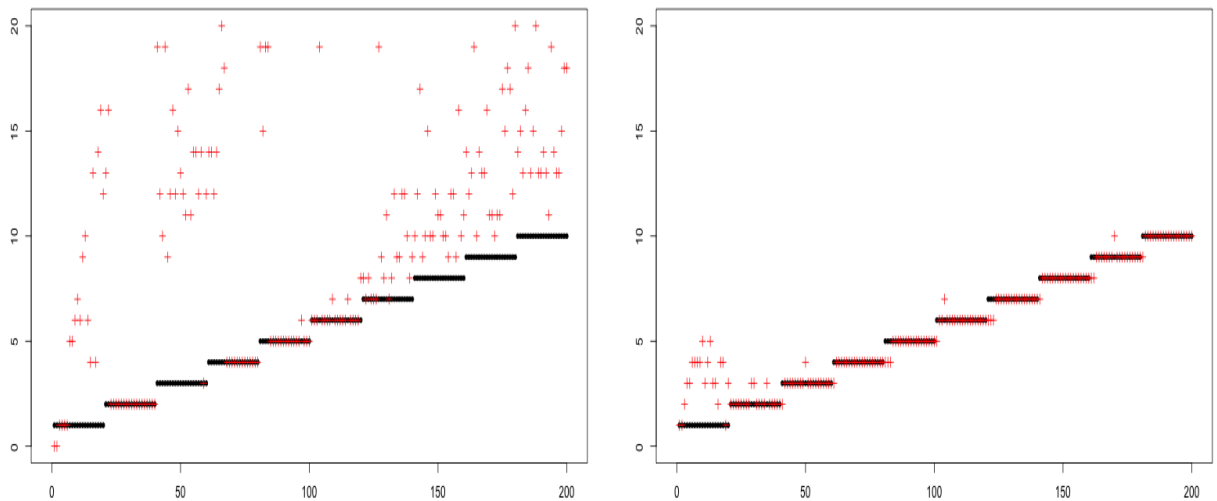

Figure: Red dot: Estimated number of cells by Gap (left) and PACO (right); Black line: true number of cells

[1] N. Cesa-Bianchi and G. Lugosi (2006), *Prediction, learning and Games*, Cambridge University Press, New York.
[2] L. Li, B. Guedj and S. Loustau (2016), *PAC-Bayesian Online Clustering*, preprint: arXiv 1602.00522.
[3] J. Y. Audibert (2009), *Fast learning rates in statistical inference through aggregation*, The Annals of Statistics, 37(4): 1591–1646.
[4] S. Loustau (2014), *Online clustering of individual sequence*, hal-00943384.
[5] S. Gerchinovitz (2011), *Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation*, PhD thesis, Université Paris-Sud.
[6] B. Guedj and P. Alquier (2013), *PAC-Bayesian estimation and prediction in sparse additive models*, Electronic Journal of Statistics.

Thanks for your attention !