PAC-Bayesian Online Clustering

Le LI[*], Benjamin GUEDJ[†] and Sébastien LOUSTAU[‡]

January 28, 2016

Abstract

This paper addresses the online clustering problem. When faced with high frequency streams of data, clustering raises theoretical and algorithmic pitfalls. Working under a sparsity assumption, a new online clustering algorithm is introduced. Our procedure relies on the PAC-Bayesian approach, allowing for a dynamic (*i.e.*, time-dependent) estimation of the number of clusters. Its theoretical merits are supported by sparsity regret bounds, and an RJMCMC-flavored implementation called PACO is proposed along with numerical experiments to assess its potential.

Keywords: Online clustering, PAC-Bayesian theory, Reversible Jump Markov Chain Monte Carlo, Sparsity regret bounds.

1 Introduction

Online learning has been extensively studied these last decades in game theory and statistics (see Cesa-Bianchi and Lugosi, 2006, and references therein). The problem could be described as a sequential game: a blackbox reveals at each time t some $z_t \in \mathcal{Z}$. Then, the forecaster predicts the next value based on the past observations and possibly other available information. The difference with the classical statistical framework lies in the fact that the sequence (z_t) is not assumed to be a realization of some stochastic process. Research efforts in online learning began in the framework of prediction with experts advices. In this setting, the forecaster has access to a set $\{f_{e,t} \in \mathcal{D} : e \in \mathcal{E}\}$ of experts' predictions, where $\mathcal{E}$ is a finite set of

[*]Université d'Angers & iAdvize, le@iadvize.com.

[†]Modal project-team, Inria, benjamin.guedj@inria.fr.

[‡]Université d'Angers, loustau@math.univ-angers.fr.

experts (such as deterministic physical models, or stochastic decisions). Predictions made by the forecaster and experts are assessed with a loss function $\ell : \mathscr{D} \times \mathscr{Z} \longrightarrow \mathbb{R}_+$. The goal is to build a sequence $\hat{z}_1, \ldots, \hat{z}_T$ (denoted in the sequel $(\hat{z}_t)_{1:T}$) of predictions which are nearly as good as the best expert's predictions in the first T time rounds, *i.e.*, satisfying uniformly over any sequence (z_t) the following regret bound:

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

where $\Delta_T(\mathscr{E})$ is a remainder term. This term should be as small as possible and in particular sublinear in T. When $\mathscr{E}$ is finite, and the loss is bounded in $[0,1]$ and convex in its first argument, an explicit $\Delta_T(\mathscr{E}) = \sqrt{(T/2)\log|\mathscr{E}|}$ is given by Cesa-Bianchi and Lugosi (2006). The optimal forecaster is then obtained by forming the exponentially weighted average of all experts. For similar results, we refer the reader to Littlestone and Warmuth (1994), Cesa-Bianchi et al. (1997) and Cesa-Bianchi and Lugosi (2006).

Online learning techniques have also been applied to the regression framework. In particular, sequential ridge regression has been studied by Vovk (2001). For any $t = 1, \ldots, T$, we now assume that $z_t = (x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$. At each time t, the forecaster gives a prediction $\hat{y}_t$ of y_t, using only newly revealed side information x_t and past observations $(x_s, y_s)_{1:(t-1)}$. Let $\langle \cdot, \cdot \rangle$ denote the scalar product in $\mathbb{R}^d$. A possible goal is to build a forecaster whose performance is nearly as good as the best linear forecaster $f_\theta : x \mapsto \langle \theta, x \rangle$, *i.e.*, such that uniformly over all sequences $(x_t, y_t)_{1:T}$,

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

where $\Delta_T(d)$ is a remainder term. This setting has been addressed by numerous contributions to the literature. In particular, Azoury and Warmuth (2001) and Vovk (2001) each provide an algorithm close to the ridge regression with a remainder term $\Delta_T(d) = \mathscr{O}(d \log T)$. Other contributions have investigated the Gradient-Descent algorithm (Cesa-Bianchi et al., 1996; Kivinen and Warmuth, 1997) and the Exponentiated Gradient Forecasters (Kivinen and Warmuth, 1997; Cesa-Bianchi, 1999). Gerchinovitz (2011) extends the notation $\langle u, x_t \rangle$ in (1) to $\langle u, \varphi(x_t) \rangle = \sum_{j=1}^{d} u_j \varphi_j(x_t)$, where $\varphi = (\varphi_1, \ldots, \varphi_d)$ is a dictionary of base forecasters. In the so-called high dimensional setting ($d \gg T$), a sparsity regret bound with a remainder term $\Delta_T(d)$ growing logarithmically with d and T is proved by Gerchinovitz (2011, Proposition 3.1).

The ambition of the present work is to transpose the aforecited framework to the clustering problem. Online clustering has attracted some attention from the machine learning and streaming communities. As an example, Guha et al. (2003), Barbakh and Fyfe (2008) and Liberty et al. (2014) study the so-called data streaming clustering problem. Its amounts to cluster online data to a fixed number of groups in a single pass, or a small number of passes, while using little memory. From a machine learning perspective, Choromanska and Monteleoni (2012) aggregate online clustering algorithms, with a fixed number K of centers. To the best of our knowledge, Loustau (2014) is the first attempt to perform online clustering with an unfixed K, which serves as a starting point to the present article.

Following Gerchinovitz (2011), we rely on the so-called PAC-Bayesian theory to derive sparsity regret bounds. The PAC-Bayesian theory originates in the machine learning community in the late 1990s, in the seminal works of Shawe-Taylor and Williamson (1997) and McAllester (1999a,b), see also Seeger (2002) and Seeger (2003). In the statistical learning community, the PAC-Bayesian approach has been extensively developed by Catoni (2004), Audibert (2004), Alquier (2006), Catoni (2007) and Dalalyan and Tsybakov (2008), and later on adapted to the high dimensional setting (Alquier and Lounici, 2011; Dalalyan and Tsybakov, 2012; Alquier and Biau, 2013; Guedj and Alquier, 2013; Guedj and Robbiano, 2015; Alquier and Guedj, 2016, among others). In a parallel effort, the online learning community has contributed to the PAC-Bayesian theory in the online regression setting (Kivinen and Warmuth, 1999), and Audibert (2009) has been the first attempt to merge both lines of research. Gerchinovitz (2011) proved similar results, completed by Loustau (2014) for the clustering problem.

The paper is organized as follows. Section 2 introduces our notation and online clustering procedure. Section 3 contains our mathematical claims, consisting in sparsity regret bounds for our PAC-Bayesian online clustering algorithm. In particular, remainder terms which are sublinear in T are obtained for a sparsity-inducing quasi-prior. We then discuss in Section 4 the practical implementation of our method, which relies on an adaptation of the Reversible Jump Monte Carlo Markov Chain (RJMCMC) algorithm to our setting. The performance of the resulting algorithm, called PACO, is evaluated in Section 5. For the sake of clarity, proofs are postponed to Section 6, and Appendix A contains an extension of our work to the case of a multivariate Student quasi-prior.

2 The PAC-Bayesian online clustering algorithm

Let $(x_t)_{1:T}$ be an online dataset, where $x_t \in \mathbb{R}^d$. Our goal is to learn a time-dependent parameter K_t and a partition of the observed points into K_t cells, for any $t = 1,\dots,T$. To this aim, the output of our algorithm at time t is a vector $\hat{\mathbf{c}}_t = (\hat{c}_{t,1}, \hat{c}_{t,2},\dots,\hat{c}_{t,K_t}) \in \mathbb{R}^{dK_t}$, depending on the past information $(x_s)_{1:(t-1)}$ and $(\hat{\mathbf{c}}_s)_{1:(t-1)}$. A partition is then fulfilled by assigning $(x_s)_{1:(t-1)}$ to its closest center. When x_t is newly revealed, the instantaneous loss is computed as

$$\ell(\hat{\mathbf{c}}_t, x_t) = \min_{1\le k\le K_t} |\hat{c}_{t,k} - x_t|_2^2, \tag{2}$$

where $|\cdot|_2$ is the ℓ_2-norm in $\mathbb{R}^d$. In what follows, we investigate sparsity regret bounds for cumulative losses. Given a measurable space Θ (embedded with its Borel σ-algebra), we let $\mathscr{P}(\Theta)$ denote the set of probability distributions on Θ, and for some reference measure ν, we let $\mathscr{P}_\nu(\Theta)$ be the set of probability distributions absolutely continuous with respect to ν. For any $\rho, \pi \in \mathscr{P}(\Theta)$, the Kullback-Leibler divergence $\mathscr{K}(\rho,\pi)$ is defined as

$$\mathscr{K}(\rho,\pi) = \begin{cases} \int_\Theta \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\pi & \text{when } \rho \in \mathscr{P}_\pi(\Theta), \\ +\infty & \text{otherwise.} \end{cases}$$

Note that for any bounded measurable function $h : \Theta \to \mathbb{R}$ and any probability distribution $\rho \in \mathscr{P}(\Theta)$ such that $\mathscr{K}(\rho,\pi) < +\infty$,

$$-\log \int_\Theta \exp(-h)\mathrm{d}\pi = \inf_{\rho\in\mathscr{P}(\Theta)} \left\{ \int_\Theta h\mathrm{d}\rho + \mathscr{K}(\rho,\pi) \right\}. \tag{3}$$

This result, which may be found in Csiszar (1975) and Catoni (2004, Equation 5.2.1), is critical in our scheme of proofs. Further, the infimum is achieved at the so-called Gibbs quasi-posterior $\hat{\rho}$, defined by

$$\mathrm{d}\hat{\rho} = \frac{\exp(-h)}{\int \exp(-h)\mathrm{d}\pi}\mathrm{d}\pi.$$

We now introduce the notation to our online clustering setting. Let $\mathscr{C} = \cup_{k=1}^p \mathbb{R}^{dk}$ for some integer $p \ge 1$. We denote by q a discrete probability distribution on the set $[\![1,p]\!] := \{1,\dots,p\}$. For any $k \in [\![1,p]\!]$, let π_k denote a probability distribution on $\mathbb{R}^{dk}$. For any $\mathbf{c} \in \mathscr{C}$, we define $\pi(\mathbf{c})$ as

$$\pi(\mathbf{c}) = \sum_{k\in[\![1,p]\!]} q(k)\mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}\pi_k(\mathbf{c}). \tag{4}$$

Note that (4) may be seen as a distribution over the set of partitions of $\mathbb{R}^d$: any $\mathbf{c} \in \mathscr{C}$ corresponds to a partition of $\mathbb{R}^d$ with at most p cells. In the sequel, we denote by $\mathbf{c} \in \mathscr{C}$ a partition of $\mathbb{R}^d$ and by $\pi \in \mathscr{P}(\mathscr{C})$ a quasi-prior over this set. Let $\lambda > 0$ be some (inverse temperature) parameter. At each time t, we observe x_t and a random partition $\hat{\mathbf{c}}_{t+1} \in \mathscr{C}$ is sampled from the Gibbs quasi-posterior

$$\mathrm{d}\hat{\rho}_{t+1}(\mathbf{c}) \propto \exp\big(-\lambda S_t(\mathbf{c})\big)\mathrm{d}\pi(\mathbf{c}). \tag{5}$$

This quasi-posterior is based on the quasi-prior π defined in (4) and the following cumulative loss

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

where the latter term is a consequence of the non-convexity of the loss ℓ (see Audibert, 2009). Note that since $(x_t)_{1:T}$ is deterministic, no likelihood is attached to our approach, hence the term "quasi-posterior" for $\hat{\rho}_{t+1}$. The resulting estimate is a realization of $\hat{\rho}_{t+1}$ with a random number K_t of cells. This scheme is described in Algorithm 1.

Algorithm 1 The PAC-Bayesian online clustering algorithm

1: **Input parameters**: $p > 0, \pi \in \mathscr{P}(\mathscr{C})$, $\lambda > 0$ and $S_0 \equiv 0$
2: **Initialization**: Draw $\hat{\mathbf{c}}_1 \sim \pi$
3: **For** $t \in [\![1, T-1]\!]$
4: Get the data x_t
5: Draw $\hat{\mathbf{c}}_{t+1} \sim \hat{\rho}_{t+1}(\mathbf{c})$ where $\mathrm{d}\hat{\rho}_{t+1} \propto \exp\big(-\lambda S_t(\mathbf{c})\big)\mathrm{d}\pi(\mathbf{c})$, and

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

6: **End for**

---

3 Sparsity Regret Bounds

Let $\mathbb{E}_{\mathbf{c}\sim\nu}$ stands for the expectation with respect to the distribution ν of $\mathbf{c}$ (abbreviated as $\mathbb{E}_\nu$ where no confusion is possible).

Theorem 1. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, for any quasi-prior distribution $\pi \in \mathscr{P}(\mathscr{C})$ and any $\lambda > 0$, the procedure described in Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{\rho \in \mathscr{P}_\pi(\mathscr{C})} \Bigg\{ \mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T} \ell(\mathbf{c}, x_t)\right] + \frac{\mathscr{K}(\rho, \pi)}{\lambda} + \frac{\lambda}{2} \mathbb{E}_{(\hat{\rho}_1, \ldots, \hat{\rho}_T)} \mathbb{E}_{\mathbf{c}\sim\rho} \sum_{t=1}^{T} [\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2 \Bigg\}.$$

Theorem 1 is a straightforward consequence of Audibert (2009, Theorem 4.6) applied to the loss function defined in (2), the set of partition $\mathscr{C}$, and any quasi-prior $\pi \in \mathscr{P}(\mathscr{C})$.

3.1 Preliminary sparsity regret bounds

In the following, we refine the regret bound introduced in Theorem 1. Distribution q in (4) is chosen as the following discrete distribution on the set $[\![1,p]\!]$

$$q(k) = \frac{\exp(-\eta k)}{\sum_{i=1}^{p} \exp(-\eta i)}, \quad \eta \geq 0. \tag{6}$$

When $\eta > 0$, the larger the number of cells k, the smaller the probability mass. Further, π_k in (4) is chosen as a product of k independent uniform distributions on ℓ_2-balls in $\mathbb{R}^d$:

$$\mathrm{d}\pi_k(\mathbf{c}, R) = \left(\frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{\frac{d}{2}}} \right)^{k} \frac{1}{(2R)^{dk}} \left\{ \prod_{j=1}^{k} \mathbb{1}_{\{B_d(2R)\}}(c_j) \right\} \mathrm{d}\mathbf{c}, \tag{7}$$

where $R > 0$, Γ is the Gamma function and $B_d(r) = \{x \in \mathbb{R}^d, |x|_2 \leq r\}$ is an ℓ_2-ball in $\mathbb{R}^d$, centered in $0 \in \mathbb{R}^d$ with radius $r > 0$. Finally, for any $k \in [\![1,p]\!]$ and any $R > 0$, let

$$\mathscr{C}(k,R) = \left\{ \mathbf{c} = (c_j)_{j=1,\ldots,k} \in \mathbb{R}^{dk}, \text{ such that } |c_j|_2 \leq R \quad \forall j \right\}.$$

Corollary 1. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$ and any $p \geq 1$, consider π defined by (4), (6) and (7) with $\eta \geq 0$ and $R \geq \max_{t=1,\ldots,T} |x_t|_2$. If $\lambda \geq (d+2)/(2TR^2)$, the procedure described in Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1, \hat\rho_2, \ldots, \hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{k \in [\![1,p]\!]} \left\{ \inf_{\mathbf{c} \in \mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{dk}{2\lambda} \log\left(\frac{8R^2 \lambda T}{d+2} \right) + \frac{\eta}{\lambda} k \right\} + \left(\frac{\log p}{\lambda} + \frac{d}{2\lambda} + \frac{\lambda T C_1^2}{2} \right),$$

where $C_1 = (2R + \max_{t=1,\ldots,T} |x_t|_2)^2$.

Note that $\inf_{\mathbf{c} \in \mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t)$ is a non-increasing function of the number k of cells while the penalty is linearly increasing with k. Small values for λ (or equivalently, large values for R) lead to small k. The price to pay for the complexity of $\mathscr{C} = \bigcup_{k=1,\ldots,p} \mathbb{R}^{dk}$ is $\log p$. The calibration $\lambda = (d+2)/(2\sqrt{T}R^2)$ yields a sublinear remainder term in the following corollary.

Corollary 2. *Under the previous notation with $\lambda = (d+2)/(2\sqrt{T}R^2)$ and $R \geq \max_{t=1,\dots,T} |x_t|_2$, the procedure in Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\mathbf{c}\in\mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + k\frac{dR^2}{d+2}\sqrt{T}\log\left(4\sqrt{T}\right) \right.$$
$$\left. k\frac{2R^2\eta}{d+2}\sqrt{T} \right\} + \left(\frac{2R^2\log p}{d+2} + \frac{dR^2}{d+2} + \frac{(d+2)C_1^2}{4R^2} \right)\sqrt{T}.$$

Let us stress that if there exist $k^\star \in [\![1,p]\!]$ and $\mathbf{c}^\star \in \mathscr{C}(k^\star, R)$ which achieve the infimum on the right-hand side, then

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) - \sum_{t=1}^{T} \ell(\mathbf{c}^\star, x_t) \leq J\, k^\star \sqrt{T}\log T, \tag{8}$$

where J is a constant depending on d, R, $\log p$ and C_1^2. In (8) the regret of the expected cumulative loss is sublinear in T. This nice behavior is illustrated in Section 5. However, whenever $k^\star > p$, the term $k\sqrt{T}\log(4\sqrt{T})$ emerges and the bound in Corollary 2 is deteriorated.

3.2 Adaptive sparsity regret bounds

The time horizon T is usually unknown, prompting us to choose a time-dependent inverse temperature parameter $\lambda = \lambda_t$. We thus propose a slight adaptation of Algorithm 1, described in Algorithm 2. This adaptation allows for the following refined result.

Algorithm 2 The adaptive PAC-Bayesian online clustering algorithm

1: **Input parameters**: $p > 0, \pi \in \mathscr{P}(\mathscr{C})$, $(\lambda_t)_{0:T} > 0$ and $S_0 \equiv 0$
2: **Initialization**: Draw $\hat{\mathbf{c}}_1 \sim \pi$
3: **For** $t \in [\![1, T-1]\!]$
4: Get the data x_t
5: Draw $\hat{\mathbf{c}}_{t+1} \sim \hat\rho_{t+1}(\mathbf{c})$ where $\mathrm{d}\hat\rho_{t+1} \propto \exp\left(-\lambda_t(\mathbf{c})\right)\mathrm{d}\pi(\mathbf{c})$, and

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

6: **End for**

Theorem 2. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, any quasi-prior distribution π on $\mathscr{C}$, if $(\lambda_t)_{0:T}$ is a non-increasing sequence of positive numbers, then the procedure described in Algorithm 2 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \le \inf_{\rho\in\mathscr{P}_\pi(\mathscr{C})} \left\{ \mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T} \ell(\mathbf{c}, x_t)\right] + \frac{\mathscr{K}(\rho,\pi)}{\lambda_T} \right.$$
$$\left. + \mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)} \mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T} \frac{\lambda_{t-1}}{2}[\ell(\mathbf{c},x_t) - \ell(\hat{\mathbf{c}}_t,x_t)]^2\right]\right\}.$$

If λ is chosen in Theorem 1 as $\lambda = \lambda_T$, the only difference between Theorem 1 and Theorem 2 lies on the last term of the regret bound. This term will be larger in the adaptive setting than in the simpler non-adaptive setting since $(\lambda_t)_{0:T}$ is non-increasing. In other words, here is the price to pay for the adaptivity of our algorithm. However, a suitable choice of λ_t allows, again, for a refined result.

Corollary 3. *For any deterministic sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, if q and π_k in (4) are taken respectively as in (6) and (7) with $\eta \ge 0$ and $R \ge \max_{t=1,\dots,T} |x_t|_2$, if $\lambda_t = (d+2)/\left(2\sqrt{t}R^2\right)$ for any $t \in [\![1, T]\!]$ and $\lambda_0 = 1$, then the procedure described in Algorithm 2 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \le \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\mathbf{c}\in\mathscr{C}(k,R)} \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \frac{dkR^2}{d+2}\sqrt{T}\log\left(4\sqrt{T}\right) \right.$$
$$\left. + k\frac{2R^2\eta}{d+2}\sqrt{T} \right\} + \left(\frac{2R^2\log p}{d+2} + \frac{dR^2}{d+2} + \frac{(d+2)C_1^2}{2R^2} \right)\sqrt{T}.$$

This result is consistent with the general results stated by Audibert (2009) and the clustering counterpart developed by Loustau (2014). The adaptive algorithm described in Algorithm 2 is supported by a sparsity regret bound with rate $\sqrt{T}\log T$.

4 The Paco algorithm

Since direct sampling from the Gibbs quasi-posterior is usually not possible, we focus on a stochastic approximation in this section, called Paco. Both implementation and convergence (towards the Gibbs quasi-posterior) of this scheme are discussed.

4.1 Structure and links with RJMCMC

In Algorithm 1 and Algorithm 2, it is required to sample at each t from the Gibbs quasi-posterior $\hat\rho_t$. Since $\hat\rho_t$ is defined on the massive and complex-structured space $\mathscr{C}$ (let us recall that $\mathscr{C}$ is a union of heterogeneous spaces), direct sampling from $\hat{\rho}_t$ is not an option and is much rather an algorithmic challenge. Our approach consists in approximating $\hat{\rho}_t$ through Markov Chain Monte Carlo (MCMC) under the constraint of favoring local moves of the Markov chain. To do so, we rely on the popular RJMCMC devised by Green (1995), adapted with ideas from the Subspace Carlin and Chib algorithm proposed by Dellaportas et al. (2002) and Petralias and Dellaportas (2013). An adaptation of this approach to the batch sparse regression appears in Guedj and Alquier (2013) and is further developed in Guedj and Robbiano (2015). Since sampling from $\hat{\rho}_t$ is similar for any $t = 1, \dots, T$, the time index t is now omitted for the sake of brevity.

Let $(k^{(n)}, \mathbf{c}^{(n)})_{0 \le n \le N}$, $N \ge 1$ be the states of the Markov Chain of interest of length N, where $k^{(n)} \in [\![1, p]\!]$ and $\mathbf{c}^{(n)} \in \mathbb{R}^{dk^{(n)}}$. At each RJMCMC iteration, only local moves are possible from the current state $(k^{(n)}, \mathbf{c}^{(n)})$ to a proposal state $(k', \mathbf{c}')$, in the sense that the proposal state should only differ from the current state by at most one covariate. Hence, $\mathbf{c}^{(n)} \in \mathbb{R}^{dk^{(n)}}$ and $\mathbf{c}' \in \mathbb{R}^{dk'}$ may be in different spaces (*i.e.*, $k' \neq k^{(n)}$). Two auxiliary vectors $v_1 \in \mathbb{R}^{d_1}$ and $v_2 \in \mathbb{R}^{d_2}$ $(d_1, d_2 \ge 1)$ are needed to compensate for this dimensional difference, *i.e.*, satisfying the dimension matching condition introduced by Green (1995)

$$dk^{(n)} + d_1 = dk' + d_2,$$

such that the pairs $(v_1, \mathbf{c}^{(n)})$ and $(v_2, \mathbf{c}')$ are of analogous dimension. This condition is a preliminary to the detailed balance condition that ensures that the Gibbs quasi-posterior $\hat{\rho}_t$ is the invariant distribution of the Markov chain. The structure of PACO is presented in Figure 1.

Let $\rho_{k'}(\cdot, \mathfrak{c}_{k'}, \tau_{k'})$ denote the multivariate Student distribution on $\mathbb{R}^{dk'}$

$$\rho_{k'}(\mathbf{c}, \mathfrak{c}_{k'}, \tau_{k'}) = \prod_{j=1}^{k'} \left\{ C_{\tau_{k'}}^{-1} \left(1 + \frac{|c_j - \mathfrak{c}_{k',j}|_2^2}{6\tau_{k'}^2} \right)^{-\frac{3+d}{2}} \right\} \mathrm{d}\mathbf{c}, \tag{9}$$

where $C_{\tau_{k'}}^{-1}$ is the normalizing constant. Let us now detail the proposal mechanism. First, a local move from $k^{(n)}$ to k' is proposed by choosing $k' \in [\![k^{(n)} - 1, k^{(n)} + 1]\!]$ with probability $q(k^{(n)}, \cdot)$. Next, choosing $d_1 = dk'$, $d_2 = dk^{(n)}$, we sample v_1 from $\rho_{k'}$ in (9). Finally, the pair $(v_2, \mathbf{c}')$ is obtained by

$$(v_2, \mathbf{c}') = g\left(v_1, \mathbf{c}^{(n)}\right),$$

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 one to one, first order derivative mapping. The resulting RJMCMC acceptance probability is

$$\alpha\left[\left(k^{(n)}, \mathbf{c}^{(n)}\right), \left(k', \mathbf{c}'\right)\right] = \min\left\{1, \frac{\hat{\rho}_t(\mathbf{c}') q(k', k^{(n)}) \rho_{k^{(n)}}(v_2)}{\hat{\rho}_t(\mathbf{c}^{(n)}) q(k^{(n)}, k') \rho_{k'}(v_1)} \left| \frac{\partial g\left(v_1, \mathbf{c}^{(n)}\right)}{\partial v_1 \partial \mathbf{c}^{(n)}} \right| \right\}, \tag{10}$$

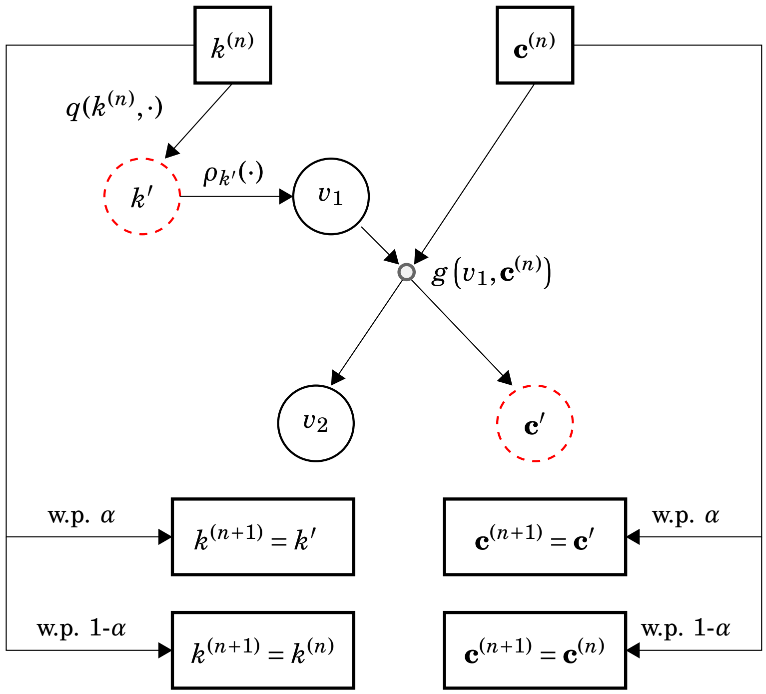


Figure 1: General structure of PACO.

where the final term is the determinant of the Jacobian matrix of g. The resulting PACO algorithm is described in Algorithm 3. Note that $\mathfrak{c}'$ in line 6 is generated from a standard k-means algorithm trained on $(x_s)_{1:(t-1)}$, and that the second equality in line 10 is due to the fact that the determinant of the Jacobian matrix of g is 1.

4.2 Convergence of PACO towards the Gibbs quasi-posterior

We prove in this subsection that Algorithm 3 builds a Markov chain whose invariant distribution is precisely the Gibbs quasi-posterior as N goes to $+\infty$. To do so, we need to prove that the chain is $\hat{\rho}_t$-irreducible, aperiodic and Harris recurrent, see Robert and Casella (2004, Theorem 6.51) and Roberts and Rosenthal (2006, Theorem 20).

Recall that at each RJMCMC iteration in Algorithm 3, the chain is said to propose a "between model move" if $k' \neq k^{(n)}$ and a "within model move" if $k' = k^{(n)}$ and $\mathbf{c}' \neq \mathbf{c}^{(n)}$. The following result gives a sufficient condition for the chain to be Harris recurrent.

Algorithm 3 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 $n \in [\![1, N-1]\!]$
    Sample $k' \in [\![k^{(n)}-1, k^{(n)}+1]\!]$ from $q(k^{(n)}, \cdot) = \frac{1}{3}$.
    Let $\mathfrak{c}' \leftarrow$ standard k-means output.
    Let $\tau' = 1/\sqrt{pt}$.
    Sample $v_1 \sim \rho_{k'}(\cdot, \mathfrak{c}_{k'}, \tau_{k'})$, .
    Let $(v_2, \mathbf{c}') = g(v_1, \mathbf{c}^{(n)})$.
    Accept the move $(k^{(n)}, \mathbf{c}^{(n)}) = (k', \mathbf{c}')$ with probability

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

    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

Lemma 1. *Let D be the event that no "within-model move" is ever accepted and $\mathscr{E}$ be the support of $\hat{\rho}_t$. Then the chain generated by Algorithm 3 satisfies*

$$\mathbb{P}\left[D | \left(k^{(0)}, \mathbf{c}^{(0)}\right) = (k, \mathbf{c})\right] = 0,$$

for any $k \in [\![1, p]\!]$ and $\mathbf{c} \in \mathbb{R}^{dk} \cap \mathscr{E}$.

Lemma 1 states that the chain must eventually accept a "within-model move". It remains true for other choice of $q(k^{(n)}, \cdot)$ in Algorithm 3, provided that it preserves the stationarity of $\hat{\rho}_t$.

Theorem 3. *Let $\mathscr{E}$ denote the support of $\hat{\rho}_t$. Then for any $\mathbf{c}^{(0)} \in \mathscr{E}$, the chain $\left(\mathbf{c}^{(n)}\right)_{1:N}$ generated by Algorithm 3 is $\hat{\rho}_t$-irreducible, aperiodic and Harris recurrent.*

Theorem 3 legitimates our approximation PACO to perform online clustering, since it asymptotically mimics the behavior of the computationally unavailable $\hat{\rho}_t$. To the best of our knowledge, this kind of guarantee is original in the PAC-Bayesian literature.

5 Numerical studies

This section is devoted to the illustration of the potential of our PAC-Bayesian approach. Several synthetic models are considered.

From a theoretical perspective, the implementation of PACO requires a fine calibration of R and λ (or λ_t in the online setting). A too small value for R will result in the sticking of the algorithm, preventing the chain to move from the initial state. Indeed, most of the proposals will be discarded since the acceptance ratio will likely be 0 (exactly or numerically). On the other hand, a too large value for R will induce a poor upper bound in Corollary 1 and Corollary 2. In addition, λ needs to be calibrated as well since the number of cells is very sensitive to its value. Recall that large values for λ enforce the Gibbs quasi-posterior to account more for past data, whereas small values make the quasi-posterior rather similar to the quasi-prior. In our numerical experiments, in accordance with the order of magnitude suggested by our theoretical results, we set $\lambda = 0.6 \times (d+2)/(2\sqrt{n})$ (resp. $\lambda_t = 0.6 \times (d+2)/(2\sqrt{t})$) in batch setting (resp. online setting), where d is the dimension of the observations. Further, we set R to be the maximum ℓ_2-norm of the observations.

5.1 Batch clustering setting

A large variety of methods have been proposed in the literature for selecting the number of clusters k in batch clustering (see Milligan and Cooper, 1985; Gordon, 1999, for a survey). These methods may be of local or global nature. For local methods, at each step, each cluster is either merged with another one, split in two or remain unaltered. Global methods evaluate the empirical distortion of any clustering as a function of the number k of cells over the whole dataset, and selects the minimizer of this distortion.

In particular, the rule of Hartigan (1975) is a well-known representative of local methods. Popular global methods include the works of Calinski and Harabasz (1974), Krzanowski and Lai (1988) and Kaufman and Rousseeuw (1990), where functions based on the empirical distortion or on the average of within-cluster dispersion of each point are constructed and the optimal number of clusters is the maximizer of these functions. In addition, the Gap Statistic (Tibshirani et al., 2001) compares the change in within-cluster dispersion with the one expected under an appropriate reference null distribution. More recently, CAPUSHE (CAlibrating Penalty Using Slope Heuristics) introduced by Fischer (2011) and Baudry et al. (2012) addresses the problem from the penalized model selection perspective.

In this subsection, we compare PACO to the aforecited methods in a batch setting with $n = 200$ observations simulated from the following 5 models.

Model 1 (1 group in dimension 5)**.** *Observations are sampled from a uniform distribution on the unit hypercube in* $\mathbb{R}^5$.

Model 2 (4 weakly separated Gaussian groups in dimension 2)**.** *Observations are sampled from 4 bivariate Gaussian distributions with identity covariance matrix, whose mean vectors are respectively* $(0,0),(-2,-1),(0,4),(3,1)$*. Each observation is uniformly drawn from one of the four groups.*

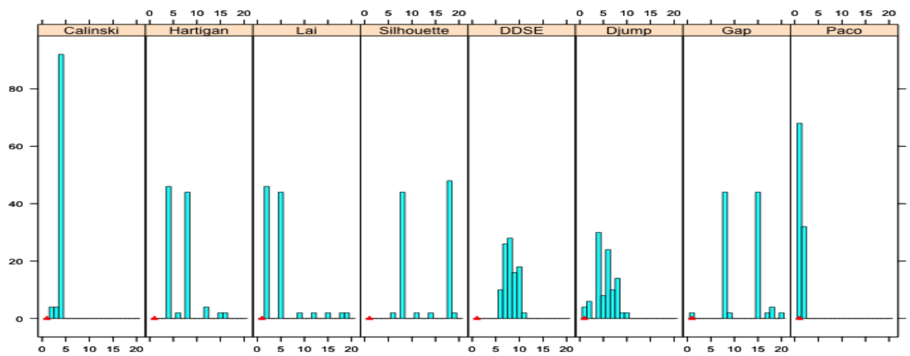
Model 3 (4 strongly separated Gaussian groups in dimension 2)**.** *Same as the previous model but with more separated mean vectors:* $(0,0),(-4,-1),(0,7),(5,2)$*.*

Model 4 (7 Gaussian groups in dimension 50)**.** *Observations are sampled from 7 multivariate Gaussian distributions in* $\mathbb{R}^{50}$ *with identity covariance matrix, whose mean vectors are chosen randomly according to a uniform distribution on* $[-10,10]^{50}$*. Each observation is uniformly drawn from one of the seven groups.*

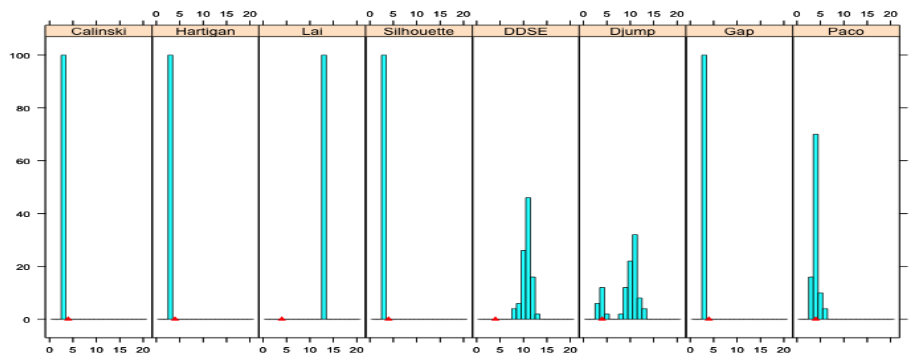
Model 5 (3 lognormal groups in dimension 3)**.** *Observations are sampled from 3 multivariate lognormal distributions in* $\mathbb{R}^3$ *with identity covariance matrix, whose mean vectors are respectively* $(1,1,1),(6,5,7),(10,9,11)$*. Each observation is uniformly drawn from one of the three groups.*

R packages implementing concurrent methods are used with their default parameters in our simulations.
Figure 2 and Figure 3 present the percentage of the estimated number of cells k on 50 realizations of the 5 aforementioned models, for 8 methods including PACO. In each graph, the red dot indicates the real number of groups. The methods used for selecting k are presented on the top of each panel, where DDSE (Data-Driven Slope Estimation) and Djump (Dimension jump) are the two methods introduced in CAPUSHE (Baudry et al., 2012). The maximum number of cells is set to 20.

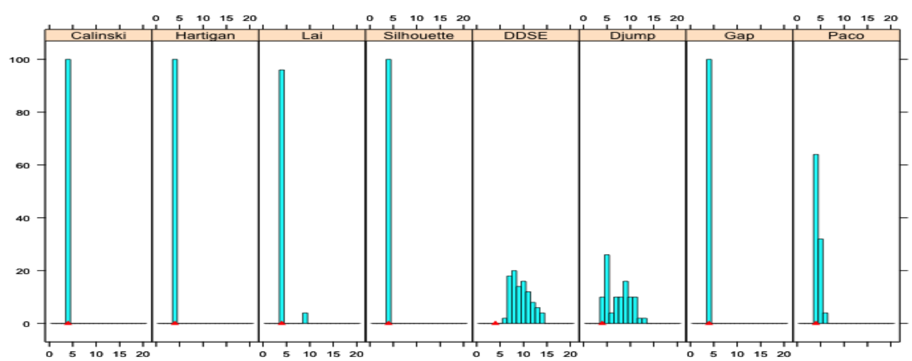
For Model 1, PACO outperforms all competitors, since it selects the correct number of cells in almost 70% of our simulations, when all other methods barely find it (Figure 2a).
For Model 2 where 4 groups are weakly separated, Calinski, Hartigan, Silhouette and Gap underestimate the number of cells by identifying 3 groups.



(a) Model 1.

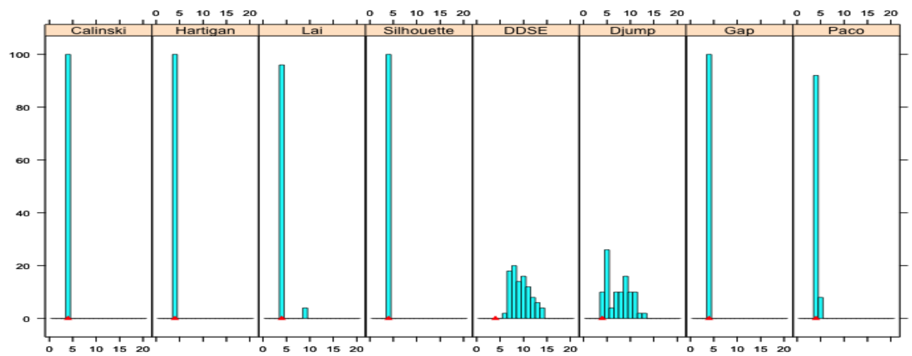


(b) Model 2.

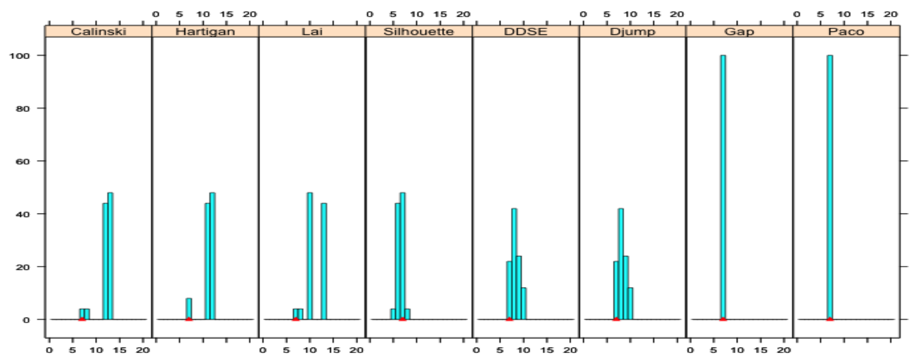


(c) Model 3.

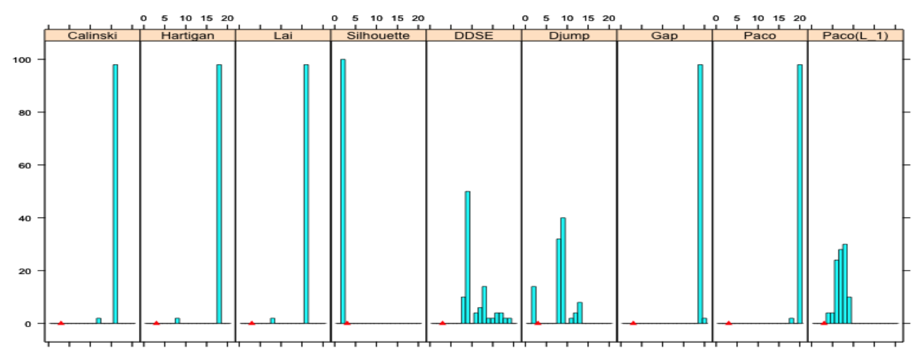
Figure 2: Histograms of the estimated number of cells on 50 realizations. The red mark indicates the true number of cells.



(a) Model 3 with a smaller λ.



(b) Model 4.



(c) Model 5.

Figure 3: Histograms of the estimated number of cells on 50 realizations. The red mark indicates the true number of cells.

Djump finds the true value $k = 4$ less than 10%. PACO identifies 4 groups in 60% of our runs (Figure 2b).
For Model 3, PACO does not perform as well as others (see Figure 2c), illustrating the need for a fine tuning of λ. Indeed, dividing λ by 3 yields far better results in Figure 3a.
For Model 4 PACO is one of the two best methods, together with Gap (Figure 3b).
For Model 5 where 3 groups of observations are generated from a heavy-tailed distribution, we consider a variant of PACO with the ℓ_1-norm in $\mathbb{R}^d$, *i.e.*, we replace the loss in (2) by $\ell(\hat{\mathbf{c}}_t, x_t) = \min_{1\le k\le K_t} |\hat{c}_{t,k} - x_t|_1$. Figure 3c shows that most methods perform poorly, to the notable exception of this PACO(ℓ_1).

5.2 Online clustering setting

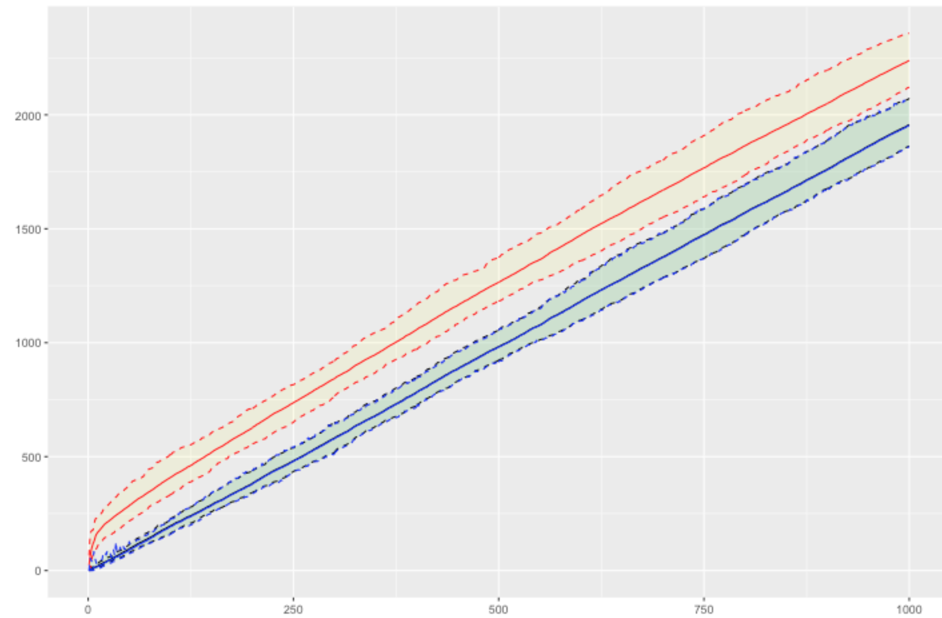
Recall from Corollary 3 that the difference between the expected cumulative loss (ECL) of Algorithm 2 and the oracle cumulative loss (OCL) is sublinear in T. In the next series of experiments, on the basis of 20 realizations with $T = 1000$ observations sampled from Model 3, we compute at each time $t \in [\![1, T]\!]$ the aforementioned two cumulative losses and compare them to validate the performance of our method in the online setting. The time-dependent choice of λ_t is fixed to $\lambda_t = 0.6 \times (d+2)/(2\sqrt{t})$ as above.

Figure 4a shows the means of ECL (orange line) and OCL (blue line) as functions of t with their 95% confidence intervals (shaded lines). The gap between ECL and OCL fills as t grows. This behavior is also pictured in Figure 4b, where the difference (normalized by t so it converges to 0) and the ratio (which converges to 1) between the mean of ECL and that of OCL are given.

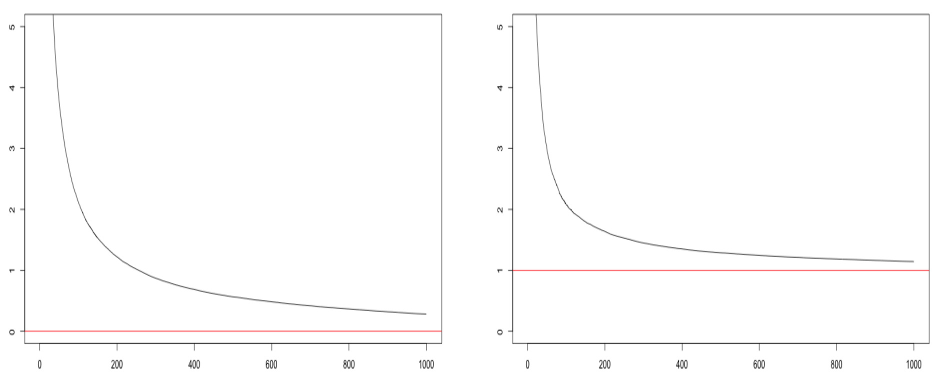
In our last experiment, we apply PACO to a more complicated online dataset from the following model.

Model 6 (10 mixed groups in dimension 2)**.** *Observations $(x_t)_{t=1,\ldots,T=200}$ are simulated in the following way: define firstly for each $t \in [\![1, T]\!]$ a pair $(c_{x,t}, c_{y,t}) \in \mathbb{R}^2$, where $c_{x,t} = \frac{20}{9}\lfloor\frac{t-1}{20}\rfloor$ and $c_{y,t} = 10\sin(c_{x,t}\pi/10)$. Then for $t \in [\![1, 100]\!]$, x_t is sampled from a uniform distribution on the unit cube in $\mathbb{R}^2$, centered at $(c_{x,t}, c_{y,t})$. For $t \in [\![101, 200]\!]$, x_t is generated by a bivariate Gaussian distribution, centered at $(c_{x,t}, c_{y,t})$ with identity covariance matrix.*

Under this model, the true number of groups will augment of 1 unit every 20 time steps to reach 10 at the end. Figure 5a shows the number of cells as a function of t, given respectively by PACO and Gap, one of the best among

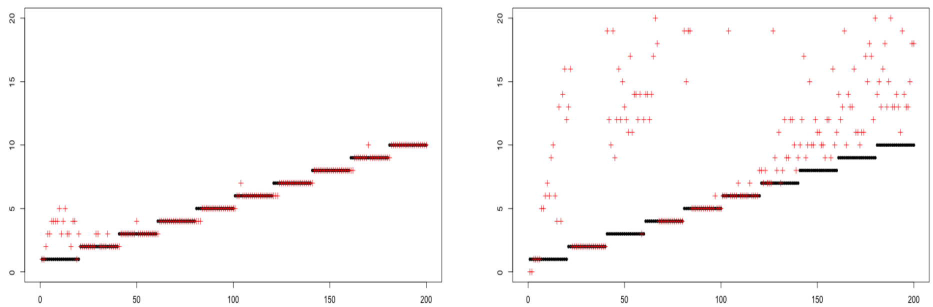


(a) ECL (orange line) and OCL (blue line) as functions of t, with 95% confidence intervals (dashed lines).

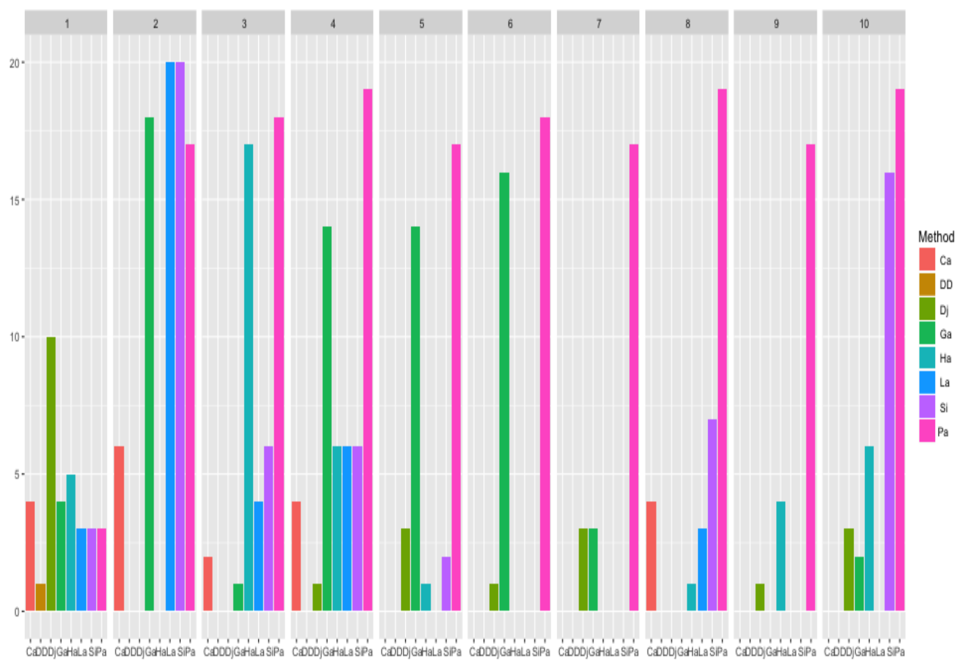


(b) Difference (left) and ratio (right) between ECL and OCL as functions of t. Convergence towards 0 for the difference and 1 for the ratio are indicated in solid red.

Figure 4: Comparison between ECL and OCL.



(a) Estimated number of cells by PACO (left) and Gap (right) as a function of t. Black lines represent the true number and red dots the estimates.



(b) Counts of correct estimations of the true number of clusters (top line) by 8 different methods including PACO (pink) , out of 20 trials.

Figure 5: Performance of PACO on Model 6.

the pool of competitors we consider. The black line indicates the real number of groups with respect to t. Both methods fail to identify a single group at the beginning: a probable miss due to the very small amount of observed data. Later on, PACO almost perfectly identifies the true number of groups, whereas it is a bit more of a struggle for Gap.
Finally, Figure 5b presents the number of correct estimation out of 20 trials for each of the ten time intervals (corresponding to the ten cells appearance). Each of the eight methods is represented by a different color. PACO outperforms all other methods, most of the time spectacularly. More graphics are presented in section Appendix A (Figure 6).

6 Proofs

This section contains the proofs to all original results claimed in Section 3 and Section 4.

6.1 Proof of Corollary 1

Let us first introduce some notation. For any $k \in [\![1,p]\!]$ and $R > 0$, let

$$\mathscr{C}(k,R) = \left\{\mathbf{c} = (c_j)_{j=1,\ldots,k} \in \mathbb{R}^{dk} : |c_j|_2 \le R, \forall j\right\},$$
$$\Xi(k,R) = \left\{\xi = (\xi_j)_{j=1,\ldots,k} \in \mathbb{R}^{k} : 0 < \xi_j \le R, \forall j\right\}.$$

We denote by $\rho_k(\mathbf{c},\mathfrak{c},\xi)$ the density consisting in the product of k independent uniform distributions on ℓ_2-balls in $\mathbb{R}^d$, namely,

$$\mathrm{d}\rho_k(\mathbf{c},\mathfrak{c},\xi) = \prod_{j=1}^{k}\left\{\frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}}}\left(\frac{1}{\xi_j}\right)^d \mathbb{1}_{\{B_d(\mathfrak{c}_j,\xi_j)\}}(c_j)\right\}\mathrm{d}\mathbf{c},$$

where $\mathfrak{c} \in \mathscr{C}(k,R)$, $\xi \in \Xi(k,R)$ and $B_d(\mathfrak{c}_j,\xi_j)$ is an ℓ_2-ball in $\mathbb{R}^d$, centered in $\mathfrak{c}_j$ with radius ξ_j. In the following, we will shorten $\rho_k(\mathbf{c},\mathfrak{c},\xi)$ to ρ_k when no confusion can arise.

The proof relies on choosing a specific ρ in Theorem 1. For any $k \in [\![1,p]\!]$, $\mathfrak{c} \in \mathscr{C}(k,R)$ and $\xi \in \Xi(k,R)$, let $\rho = \rho_k \mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}$. Then ρ is a well-defined distribution on $\mathscr{C}$ and belongs to $\mathscr{P}_\pi(\mathscr{C})$. Theorem 1 yields

$$\sum_{t=1}^{T}\mathbb{E}_{(\hat\rho_1,\hat\rho_2,\ldots,\hat\rho_t)}\ell(\hat{\mathbf{c}}_t,x_t) \le \inf_{k\in[\![1,p]\!]}\inf_{\substack{\rho\in\mathscr{P}_\pi(\mathscr{C})\\ \rho=\rho_k\mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}}}\left\{\mathbb{E}_{\mathbf{c}\sim\rho}\sum_{t=1}^{T}[\ell(\mathbf{c},x_t)] + \frac{\mathscr{K}(\rho,\pi)}{\lambda}\right.$$

$$
+\frac{\lambda}{2}\mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_T)}\mathbb{E}_{\mathbf{c}\sim\rho}\sum_{t=1}^{T}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2\Bigg\}. \tag{11}
$$

For any $\rho = \rho_k \mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}$, the first term on the right-hand side of (11) satisfies

$$
\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}_{\mathbf{c}\sim\rho}[\ell(\mathbf{c},x_t)] &= \sum_{t=1}^{T}\mathbb{E}_{\mathbf{c}\sim\rho_k}[\ell(\mathbf{c},x_t)] \\
&\le \sum_{t=1}^{T}\min_{j=1,\ldots,k}\left\{\mathbb{E}_{\mathbf{c}\sim\rho_k}\left[|c_j-\mathfrak{c}_j|_2^2\right]+|\mathfrak{c}_j-x_t|_2^2\right\} \\
&= \sum_{t=1}^{T}\min_{j=1,\ldots,k}\left\{\frac{d}{d+2}\xi_j^2+|\mathfrak{c}_j-x_t|_2^2\right\} \\
&\le \frac{dT}{d+2}\max_{j=1,\ldots,k}\xi_j^2+\sum_{t=1}^{T}\ell(\mathfrak{c},x_t),
\end{aligned} \tag{12}
$$

Let us now compute the second term on the right-hand side of (11).

$$
\begin{aligned}
\mathscr{K}(\rho,\pi) &= \int_{\mathscr{C}}\log\frac{\rho(\mathbf{c})}{\pi(\mathbf{c})}\rho(\mathbf{c})\mathrm{d}\mathbf{c} \\
&= \int_{\mathbb{R}^{dk}}\left(\log\frac{\rho_k(\mathbf{c})}{\pi_k(\mathbf{c})}+\log\frac{\pi_k(\mathbf{c})}{\pi(\mathbf{c})}\right)\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} \\
&= \mathscr{K}(\rho_k,\pi_k)+\log\frac{1}{q(k)} \\
&=: A+B,
\end{aligned}
$$

where

$$
A = \int_{\mathbb{R}^{dk}}\log\prod_{j=1}^{k}\frac{\left(\frac{1}{\xi_j}\right)^d}{\left(\frac{1}{2R}\right)^d}\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} = d\sum_{j=1}^{k}\log\left(\frac{2R}{\xi_j}\right).
$$

and since the function $x \mapsto (1-e^{-\eta x})/x$ is non-increasing for $x>0$ and $\eta>0$, we have

$$
\begin{aligned}
B &= \log\left(\frac{e^{-\eta}(1-e^{-\eta p})}{1-e^{-\eta}}e^{\eta k}\right) \\
&\le \log\left(pe^{\eta(k-1)}\right) \\
&= \eta(k-1)+\log p.
\end{aligned} \tag{13}
$$

When $\eta = 0$, q is a uniform distribution on $[\![1,p]\!]$, and the above inequality holds as well. Then, $\mathscr{K}(\rho,\pi)/\lambda$ in (11) may be upper bounded.

$$
\frac{\mathscr{K}(\rho,\pi)}{\lambda} \le \frac{d}{\lambda}\sum_{j=1}^{k}\log\left(\frac{2R}{\xi_j}\right)+\frac{\eta(k-1)}{\lambda}+\frac{\log p}{\lambda}. \tag{14}
$$

Finally

$$\begin{aligned}|\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)| &= \left|\min_{j=1,\dots,k}|c_j-x_t|_2^2-\min_{j=1,\dots,K_t}|\hat{c}_{t,j}-x_t|_2^2\right| \\ &\le \min_{j=1,\dots,k}|c_j-x_t|_2^2 \vee \min_{j=1,\dots,K_t}|\hat{c}_{t,j}-x_t|_2^2 \\ &\le \left(2R+\max_{t=1,\dots,T}|x_t|_2\right)^2 \\ &=: C_1.\end{aligned}$$

Then, the third term of the right-hand side in (11) is controlled as follows:

$$\frac{\lambda}{2}\mathbb{E}_{(\hat{\rho}_1,\dots,\hat{\rho}_T)}\mathbb{E}_{\mathbf{c}\sim\rho_k}\sum_{t=1}^{T}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2 \le \frac{\lambda T}{2}C_1^2. \tag{15}$$

Combining inequalities (12), (14) and (15) gives, for any $\xi\in\Xi(k,R)$,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\dots,\hat{\rho}_t)}\ell(\hat{\mathbf{c}}_t,x_t) \le \inf_{k\in[\![1,p]\!]}\inf_{\mathfrak{c}\in\mathscr{C}(k,R)}\Bigg\{&\sum_{t=1}^{T}\ell(\mathfrak{c},x_t)+\frac{dT}{d+2}\max_{j=1,\dots,k}\xi_j^2 \\ &+\frac{d}{\lambda}\sum_{j=1}^{k}\log\left(\frac{2R}{\xi_j}\right)+\frac{\eta}{\lambda}(k-1)\Bigg\}+\frac{\lambda T}{2}C_1^2+\frac{\log p}{\lambda}.\end{aligned}$$

Under the assumption that $\lambda>(d+2)/(2TR^2)$, the global minimizer of the function

$$(\xi_1,\dots,\xi_k)\mapsto\frac{Td}{d+2}\max_{j=1,\dots,k}\xi_j^2+\frac{d}{\lambda}\sum_{j=1}^{k}\log\left(\frac{2R}{\xi_j}\right) \tag{16}$$

does not necessarily belong to $\Xi(k,R)$. A possible choice of $(\xi_j)_{1:k}\in\Xi(k,R)$ is given by

$$\xi_1^\star=\xi_2^\star=\cdots=\xi_k^\star=\sqrt{\frac{d+2}{2\lambda T}}.$$

Then (16) amounts to

$$\frac{d}{2\lambda}+\frac{dk}{2\lambda}\log\left(\frac{8R^2\lambda T}{d+2}\right).$$

Hence,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\dots,\hat{\rho}_t)}\ell(\hat{\mathbf{c}}_t,x_t) \le \inf_{k\in[\![1,p]\!]}\inf_{\mathfrak{c}\in\mathscr{C}(k,R)}\Bigg\{&\sum_{t=1}^{T}\ell(\mathfrak{c},x_t)+\frac{dk}{2\lambda}\log\left(\frac{8R^2\lambda T}{(d+2)k}\right) \\ &+\frac{\eta}{\lambda}k\Bigg\}+\left(\frac{\log p}{\lambda}+\frac{d}{2\lambda}+\frac{\lambda T}{2}C_1^2\right).\end{aligned}$$

6.2 Proof of Theorem 2

The proof is mainly based on the online variance inequality described in Audibert (2009), *i.e.*, for any $\lambda > 0$, any $\hat{\rho} \in \mathscr{P}_\pi(\mathscr{C})$ and any $x \in \mathbb{R}^d$,

$$\mathbb{E}_{\mathbf{c}' \sim \hat{\rho}}[\ell(\mathbf{c}', x)] \leq -\frac{1}{\lambda} \mathbb{E}_{\mathbf{c}' \sim \hat{\rho}} \log \mathbb{E}_{\mathbf{c} \sim \hat{\rho}} \left[e^{-\lambda \left[(\ell(\mathbf{c}, x) + \frac{\lambda}{2} (\ell(\mathbf{c}, x) - \ell(\mathbf{c}', x))^2 \right]} \right]. \tag{17}$$

By (17), we have

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) = \sum_{t=1}^T \mathbb{E}_{(\hat{\rho}_1, \dots, \hat{\rho}_{t-1})} \mathbb{E}_{\hat{\rho}_t} [\ell(\hat{\mathbf{c}}_t, x_t) \mid \hat{\mathbf{c}}_1, \dots, \hat{\mathbf{c}}_{t-1}] \\
\leq &\sum_{t=1}^T \mathbb{E}_{(\hat{\rho}_1, \dots, \hat{\rho}_{t-1})} \left[-\frac{1}{\lambda_{t-1}} \mathbb{E}_{\hat{\mathbf{c}}_t \sim \hat{\rho}_t} \log \mathbb{E}_{\mathbf{c} \sim \hat{\rho}_t} \left(e^{-\lambda_{t-1} [\ell(\mathbf{c}, x_t) + \frac{\lambda_{t-1}}{2} (\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t))^2]} \right) \right] \\
\leq &\mathbb{E}_{(\hat{\rho}_1, \dots, \hat{\rho}_T)} \left[\sum_{t=1}^T -\frac{1}{\lambda_{t-1}} \log \frac{\int e^{-\lambda_{t-1} S_t(\mathbf{c})} \mathrm{d}\pi(\mathbf{c})}{\int e^{-\lambda_{t-1} S_{t-1}(\mathbf{c})} \mathrm{d}\pi(\mathbf{c})} \right] \\
= &\mathbb{E}_{(\hat{\rho}_1, \dots, \hat{\rho}_T)} \left[\sum_{t=1}^T -\frac{1}{\lambda_{t-1}} \log \frac{V_t}{W_{t-1}} \right] \\
= &\mathbb{E}_{(\hat{\rho}_1, \dots, \hat{\rho}_T)} \left[\sum_{t=1}^T \left[\frac{1}{\lambda_{t-1}} \log W_{t-1} - \frac{1}{\lambda_{t-1}} \log V_t \right] \right].
\end{aligned} \tag{18}$$

Applying Jensen's inequality, for any $1 \leq t \leq T$,

$$\begin{aligned}
\frac{1}{\lambda_{t-1}} \log V_t &= \frac{1}{\lambda_{t-1}} \log \mathbb{E}_{\mathbf{c} \sim \pi} \left[\left(e^{-\lambda_t S_t(\mathbf{c})} \right)^{\frac{\lambda_{t-1}}{\lambda_t}} \right] \\
&\geq \frac{1}{\lambda_{t-1}} \log \left(\mathbb{E}_{\mathbf{c} \sim \pi} \left[e^{-\lambda_t S_t(\mathbf{c})} \right] \right)^{\frac{\lambda_{t-1}}{\lambda_t}} \\
&= \frac{1}{\lambda_t} \log W_t.
\end{aligned}$$

Therefore, since $W_0 = 1$,

$$\sum_{t=1}^T \left[\frac{1}{\lambda_{t-1}} \log W_{t-1} - \frac{1}{\lambda_{t-1}} \log V_t \right] \leq -\frac{1}{\lambda_T} \log W_T, \tag{19}$$

and by (18), (19) and the duality formula (3), we have

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t) \\
\leq &\mathbb{E}_{(\hat{\rho}_1, \dots, \hat{\rho}_T)} \left[-\frac{1}{\lambda_T} \log \mathbb{E}_{\mathbf{c} \sim \pi} \left[e^{-\lambda_T S_T(\mathbf{c})} \right] \right]
\end{aligned}$$

$$
\begin{aligned}
&\leq -\frac{1}{\lambda_T}\log \mathbb{E}_{\mathbf{c}\sim\pi}\left[e^{-\lambda_T \mathbb{E}_{(\hat\rho_1,\ldots,\hat\rho_T)} S_T(\mathbf{c})}\right] \qquad \text{(by Audibert, 2009, Lemma 3.2)}\\
&= \inf_{\rho\in\mathscr{P}_\pi(\mathscr{C})}\left\{\mathbb{E}_{\mathbf{c}\sim\rho}\left[\sum_{t=1}^{T}\ell(\mathbf{c},x_t)\right] + \mathbb{E}_{\mathbf{c}\sim\rho}\mathbb{E}_{(\hat\rho_1,\ldots,\hat\rho_T)}\left[\sum_{t=1}^{T}\frac{\lambda_{t-1}}{2}(\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t))^2\right]\right.\\
&\quad \left. + \frac{\mathscr{K}(\rho,\pi)}{\lambda_T}\right\},
\end{aligned}
$$

which achieves the proof.

The proof of Corollary 3 is similar to the proof of Corollary 1, the only difference lies in the fact that (15) is replaced by

$$
\begin{aligned}
\mathbb{E}_{(\hat\rho_1,\ldots,\hat\rho_T)}\mathbb{E}_{\mathbf{c}\sim\rho_k}\sum_{t=1}^{T}\frac{\lambda_{t-1}}{2}[\ell(\mathbf{c},x_t)-\ell(\hat{\mathbf{c}}_t,x_t)]^2 &\leq \frac{(d+2)C_1^2}{4R^2}\left(1+\sum_{t=2}^{T}\frac{1}{\sqrt{t-1}}\right)\\
&\leq \frac{(d+2)C_1^2}{2R^2}\sqrt{T}.
\end{aligned}
$$

6.3 Proof of Lemma 1

Let D_n denote the event that no "within-model move" is ever accepted in the first n moves. Then $D_1 = D_1^{\text{within}} \cup D_1^{\text{between}}$, where D_1^{within} stands for the event that a "within-model move" is proposed but rejected in one step and D_1^{between} that a "between-model move" is proposed in one step. Then we have

$$
\begin{aligned}
\mathbb{P}\left[D_1|(k^{(0)},\mathbf{c}^{(0)})=(k,\mathbf{c})\right] &= \mathbb{P}\left[k'\neq k|(k,\mathbf{c})\right] + \mathbb{P}\left[k'=k, \text{but rejected}|(k,\mathbf{c})\right]\\
&= \frac{2}{3}+\frac{1}{3}\left[1-\int_{\mathbb{R}^{dk}}\alpha\left[(k,\mathbf{c}),(k,\mathbf{c}')\right]\rho_k\left(\mathbf{c}',\mathfrak{c}_k,\tau_k\right)\mathrm{d}\mathbf{c}'\right],
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha\left[(k,\mathbf{c}),(k,\mathbf{c}')\right] &= \min\left\{1, \frac{\hat\rho_t(\mathbf{c}')\rho_k(\mathbf{c},\mathfrak{c}_k,\tau_k)}{\hat\rho_t(\mathbf{c})\rho_k(\mathbf{c}',\mathfrak{c}_k,\tau_k)}\right\}\\
&= \min\left\{1, h_t\left(\mathbf{c}'|(k,\mathbf{c})\right)\right\}.
\end{aligned}
$$

Under the assumption of $k'=k$, we have that $\mathbf{c}',\mathbf{c}\in\mathbb{R}^{dk}$, therefore the restriction of $\hat\rho_t$ to $\mathbb{R}^{dk}$ is well defined. Moreover, by the definition of π_k in (7), the support of the restriction of $\hat\rho_t$ to $\mathbb{R}^{dk}$ is $\mathbb{R}^{dk}\cap\mathscr{E}=(B_d(2R))^k$. Hence the function $(\mathbf{c}',\mathbf{c})\mapsto h_t\left(\mathbf{c}'|(k,\mathbf{c})\right)$ is strictly positive and continuous on the compact set $(B_d(2R))^k\times(B_d(2R))^k$. As a consequence, the minimum of $h_t\left(\mathbf{c}'|(k,\mathbf{c})\right)$ on $(B_d(2R))^k\times(B_d(2R))^k$ is achieved and we denote it by m_k, *i.e.*,

$$
m_k = \inf_{\mathbf{c}',\mathbf{c}\in(B_d(2R))^k} h_t\left(\mathbf{c}'|(k,\mathbf{c})\right) > 0.
$$

In addition, due to the continuity and positivity of ρ_k on $\mathbb{R}^{dk}$, it is clear that for any $k \in [\![1,p]\!]$

$$z_k = \int_{(B_d(2R))^k} \rho_k\left(\mathbf{c}', \mathfrak{c}_k, \tau_k\right) \mathrm{d}\mathbf{c}' > 0.$$

Therefore, for any k,

$$\begin{aligned}\int_{\mathbb{R}^{dk}} \alpha\left[(k,\mathbf{c}),(k,\mathbf{c}')\right] \rho_k\left(\mathbf{c}', \mathfrak{c}_k, \tau_k\right) \mathrm{d}\mathbf{c}' &\geq \inf_{k\in[\![1,p]\!]} (m_k z_k) \\ &=: m^\star > 0.\end{aligned}$$

Hence, uniformly on $k \in [\![1,p]\!]$ and $\mathbf{c} \in \mathbb{R}^{dk} \cap \mathscr{E}$, we have,

$$\mathbb{P}\left[D_1 | (k,\mathbf{c})\right] \leq \left[\frac{2}{3} + \frac{1}{3}(1-m^\star)\right] < 1.$$

To conclude,

$$\mathbb{P}\left[D | (k,\mathbf{c})\right] = \lim_{n\longrightarrow\infty} \mathbb{P}\left[D_n | (k,\mathbf{c})\right] \leq \lim_{n\longrightarrow\infty} \left[\frac{2}{3} + \frac{1}{3}(1-m^\star)\right]^n = 0.$$

6.4 Proof of Theorem 3

For any $\mathbf{c} \in \mathscr{E}$, there exists some $k \in [\![1,p]\!]$ such that $\mathbf{c} \in (B_d(2R))^k \subset \mathscr{E}$. For any $k' \in [\![k-1,k+1]\!]$ and for any $A \in \mathscr{B}\left(\mathbb{R}^{dk'}\right)$ such that $\hat{\rho}_t(A) > 0$, the transition kernel H of the chain is given by

$$H\big(\mathbf{c}, \mathbf{c}' \in A\big) = \int \mathbb{1}_{\{v_1 \in A\}} \alpha\left[(k,\mathbf{c}), \left(k', v_1\right)\right] q(k,k') \rho_{k'}(v_1, \mathfrak{c}_{k'}, \tau_{k'}) \mathrm{d}v_1 + r(\mathbf{c}) \delta_{\mathbf{c}}(A), \tag{20}$$

where $\rho_{k'}(\cdot, \mathfrak{c}_{k'}, \tau_{k'})$ is the multivariate Student distribution in (9) and

$$r(\mathbf{c}) = \sum_{k' \in [\![k-1,k+1]\!]} q(k,k') \int \left(1 - \alpha\left[(k,\mathbf{c}), \left(k', v_1\right)\right]\right) \rho_{k'}(v_1, \mathfrak{c}_{k'}, \tau_{k'}) \mathrm{d}v_1$$

is the probability of rejection when starting at state $\mathbf{c}$, and $\delta_{\mathbf{c}}(\cdot)$ is a Dirac measure in $\mathbf{c}$. One can easily note that $H(\mathbf{c}, \mathbf{c}' \in A)$ in (20) is strictly positive, indicating that the chain, when starting from $\mathbf{c}$, has a positive chance to move. Therefore, for any $A \in \mathscr{B}(\mathscr{C})$ such that $\hat{\rho}_t(A) > 0$, we can prove with the Chapman-Kolmogorov equation that there exists some $m \in \mathbb{N}^*$ such that

$$H^m(\mathbf{c}, A) > 0,$$

where $H^m(\mathbf{c}, A) = \int H^{m-1}(y, A) H(\mathbf{c}, \mathrm{d}y)$ is the m-step transition kernel. In other words, the chain is $\hat{\rho}_t$-irreducible.

Next, a sufficient condition for the chain to be aperiodic is that Algorithm 3 allows transitions such as $\left\{\left(k^{(n+1)}, \mathbf{c}^{(n+1)}\right) = \left(k^{(n)}, \mathbf{c}^{(n)}\right)\right\}$, *i.e.*,

$$\begin{aligned}
&\mathbb{P}\left(\alpha\left[(k^{(n)}, \mathbf{c}^{(n)}), (k', \mathbf{c}')\right] < 1\right) \\
&= \mathbb{P}\left(\frac{\hat{\rho}_t(\mathbf{c}')q(k', k^{(n)})\rho_{k^{(n)}}(\mathbf{c}^{(n)}, \mathfrak{c}_{k^{(n)}}, \tau_{k^{(n)}})}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)}, k')\rho_{k'}(\mathbf{c}', \mathfrak{c}_{k'}, \tau_{k'})} < 1\right) > 0. \qquad (21)
\end{aligned}$$

Since for any $\mathbf{c}' \in A \subset \mathscr{B}\left(\mathbb{R}^{dk'}\right) \cap \mathscr{E}^c$ such that

$$\mathbb{P}\left(\mathbf{c}' \in A\right) = \int_A \rho_{k'}(\mathbf{c}', \mathfrak{c}_{k'}, \tau_{k'})\mathrm{d}\mathbf{c}' > 0,$$

we have $\hat{\rho}_t(\mathbf{c}') = 0$, (21) holds. Therefore,

$$\mathbb{P}\left(\frac{\hat{\rho}_t(\mathbf{c}')q(k', k^{(n)})\rho_{k^{(n)}}(\mathbf{c}^{(n)}, \mathfrak{c}_{k^{(n)}}, \tau_{k^{(n)}})}{\hat{\rho}_t(\mathbf{c}^{(n)})q(k^{(n)}, k')\rho_{k'}(\mathbf{c}', \mathfrak{c}_{k'}, \tau_{k'})} < 1\right) \geq \mathbb{P}\left(\mathbf{c}' \in A\right) > 0.$$

The chain is therefore aperiodic.

Finally, the Harris recurrence of the chain is a consequence of Lemma 1 (based on Roberts and Rosenthal, 2006, Theorem 20). As a conclusion, the chain converges to the target distribution $\hat{\rho}_t$.

Acknowledgements

The authors gratefully acknowledge financial support from iAdvize and ANRT (CIFRE grant 2014-00757).

References

P. Alquier. *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. PhD thesis, Université Paris 6, 2006. 3

P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013. 3

P. Alquier and B. Guedj. Bayesian Non-Negative Matrix Factorization. arXiv preprint, 2016. URL http://arxiv.org/abs/1601.01345. 3

P. Alquier and K. Lounici. PAC-Bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011. 3

J.-Y. Audibert. *Une approche PAC-bayésienne de la théorie statistique de l'apprentissage*. PhD thesis, Université Paris 6, 2004. 3

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009. 3, 5, 6, 8, 22, 23

K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001. 2

W. Barbakh and C. Fyfe. Online clustering algorithm. *International Journal of Neural Systems*, 18(3):185–194, 2008. 3

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012. 13

R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974. 12

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. École d'Été de Probabilités de Saint-Flour 2001. Springer, 2004. 3, 4

O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture notes – Monograph Series*. Institute of Mathematical Statistics, 2007. 3

N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59(3):392–411, 1999. 2

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, New York, 2006. 1, 2

N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, 7(3):604–619, 1996. 2

N. Cesa-Bianchi, D. Helmbold, N. Freund, Y. Haussler, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997. 2

A. Choromanska and C. Monteleoni. Online clustering with experts. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 227–235, 2012. 3

I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975. 4

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008. 3

A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012. 3

P. Dellaportas, J. J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36, 2002. 9

A. Fischer. On the number of groups in clustering. *Statistics and probability letters*, 81:1771–1781, 2011. 13

S. Gerchinovitz. *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, 2011. 2, 3

A. D. Gordon. *Classification*, volume 82 of *Monographs on Statistics and Applied Probability*. Chapman Hall/CRC, Boca Raton, 1999. 12

P. J. Green. Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. 9

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013. 3, 9

B. Guedj and S. Robbiano. PAC-Bayesian High Dimensional Bipartite Ranking. arXiv preprint, 2015. URL `http://arxiv.org/abs/1511.02729`. 3, 9

S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):511–528, 2003. 3

J. A. Hartigan. *Clustering Algorithm*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1975. 12

L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, Hoboken, 1990. 12

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997. 2

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Computational Learning Theory: 4th European Conference (EuroCOLT '99)*, pages 153–167. Springer, 1999. 3

W. J. Krzanowski and Y. T. Lai. A criterion for determination the number of clusters in a data set. *Biometrics*, 44:23–34, 1988. 12

E. Liberty, R. Sriharsha, and M. Sviridenko. An algorithm for online k-means clustering. Preprint, 2014. URL `http://arxiv.org/abs/1412.5721`. 3

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–216, 1994. 2

S. Loustau. Online clustering of individual sequences. Preprint, 2014. URL `https://hal.archives-ouvertes.fr/hal-00943384`. 3, 8

D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3): 355–363, 1999a. 3

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999b. 3

G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985. 12

A. Petralias and P. Dellaportas. An MCMC model search algorithm for regression problems. *Journal of Statistical Computation and Simulation*, 83 (9):1722–1740, 2013. 9

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2004. 10

G. O. Roberts and J. S. Rosenthal. Harris Recurrence of Metropolis-Within-Gibbs and Trans-Dimensional Markov Chains. *The Annals of Applied Probability*, 16(4):2123–2139, 2006. 10, 25

M. Seeger. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002. 3

M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, 2003. 3

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997. doi: 10.1145/267460.267466. 3

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63:411– 423, 2001. 12

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69 (2):213–248, 2001. 2

A Extension to a different quasi-prior

For the sake of completion, this appendix presents additional sparsity regret bounds for a different heavy-tailed quasi-prior. Doing so, we stress that the PAC-Bayesian approach is flexible in the sense that it allows for sparsity regret bounds for a large variety of quasi-priors. The appendix ends with additional numerical results.

Let us consider π_k as a product of k independent truncated multivariate Student distributions with 3 degrees of freedom in $\mathbb{R}^d$, namely, for any $\mathbf{c} \in \mathbb{R}^{dk} \subset \mathscr{C}$,

$$\mathrm{d}\pi_k(\mathbf{c}, \tau_0, 2R) = \prod_{j=1}^{k} \left\{ C_{2R,\tau_0}^{-1} \left(1 + \frac{|c_j|_2^2}{6\tau_0^2}\right)^{-\frac{3+d}{2}} \mathbb{1}_{\{|c_j|_2 \leq 2R\}} \right\} \mathrm{d}\mathbf{c}, \tag{22}$$

where $\tau_0 > 0$ and $R > 0$ are respectively the scale and truncation parameters, and C_{2R,τ_0} is the normalizing constant accounting for the truncation. When $R = +\infty$, $\pi_k(\mathbf{c}, \tau_0, 2R)$ amounts to a distribution without truncation. In the following, we shorten $\pi_k(\mathbf{c}, \tau_0, 2R)$ to π_k where no confusion is possible.

Denote by ν the multivariate Student distribution in $\mathbb{R}^d$, with mean vector $0 \in \mathbb{R}^d$, scale parameter 1, and 3 degrees of freedom. Fix $k \in [\![1, p]\!]$, $R > 0$ and $\mathfrak{c} \in \mathscr{C}(k, R)$, and recall that $\Xi(k, R)$ denotes the hypercube in $\mathbb{R}^k$ defined by

$$\Xi(k,R) := \left\{ \xi = (\xi_j)_{j=1,\ldots,k} \in \mathbb{R}^k : 0 < \xi_j \leq R, \forall j \right\}.$$

Moreover, define for any $k \in [\![1,p]\!]$, $\mathbf{c} \in \mathbb{R}^{dk} \subset \mathscr{C}$, $\mathfrak{c} \in \mathscr{C}(k,R)$, $\xi \in \Xi(k,R)$, $0 < \tau^2 \leq \sqrt{3}R^2/(6\sqrt{d})$ and $R > 0$ a probability distribution ρ_k on $\mathbb{R}^{dk}$ by

$$\rho_k(\mathbf{c},\mathfrak{c},\tau,\xi) = \prod_{j=1}^{k} \left\{ C_{\xi_j,\tau}^{-1} \left(1 + \frac{|c_j - \mathfrak{c}_j|_2^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \mathbb{1}_{\{|c_j - \mathfrak{c}_j|_2 \leq \xi_j\}} \right\}, \tag{23}$$

where $C_{\xi_j,\tau}$ are normalizing constants defined as $C_{\xi_j,\tau} = \mathbb{P}\left(|\nu|_2 \leq \xi_j/\sqrt{2}\tau\right)/A_{d,\tau}$, where $A_{d,\tau}$ is the constant in the density of ν. Moreover, when $(\xi_j)_{j=1,\ldots,k} = +\infty$, we let $\rho_k(\mathbf{c},\mathfrak{c},\tau,\xi)$ denote the multivariate Student distribution without truncation. In the sequel, we will shorten $\rho_k(\mathbf{c},\mathfrak{c},\tau,\xi)$ to ρ_k where no confusion is possible.

Lemma 2. *Assume that q and π_k in (4) are defined respectively as in (6) and (22), and that ρ_k is defined as (23) for each $k \in [\![1,p]\!]$. For the probability distribution $\rho(\mathbf{c},\mathfrak{c},\tau,\xi) = \mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}\rho_k(\mathbf{c},\mathfrak{c},\tau,\xi)$ defined on $\mathscr{C}$, if $R \geq \max_{t=1,\ldots,T}|x_t|_2$, then*

$$\begin{aligned}\mathscr{K}(\rho,\pi) \leq &\sum_{j=1}^{k}\left[\frac{3+d}{2}\log\left(1+\frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2}\log\xi_j^2\right] - k\log c_d \\ &+ (3+d)k\log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^{k}|\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right) + kd\log\tau_0 + \log p + \eta(k-1).\end{aligned}$$

Proof. By the definition of Kullback-Leibler divergence, we have

$$\mathscr{K}(\rho,\pi) = \mathscr{K}(\rho_k,\pi_k) + \log\frac{1}{q(k)} =: A + B, \tag{24}$$

where

$$\begin{aligned}A &= \int_{\mathbb{R}^{dk}} \log\left[\prod_{j=1}^{k}\frac{C_{2R,\tau_0}}{C_{\xi_j,\tau}}\left(\frac{\tau_0^2}{\tau^2}\frac{6\tau^2+|c_j-\mathfrak{c}_j|_2^2}{6\tau_0^2+|c_j|_2^2}\right)^{-\frac{3+d}{2}}\right]\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} \\ &= \sum_{j=1}^{k}\log\frac{C_{2R,\tau_0}}{C_{\xi_j,\tau}} + \frac{3+d}{2}\int_{\mathbb{R}^{dk}}\sum_{j=1}^{k}\log\left(\frac{\tau^2}{\tau_0^2}\frac{6\tau_0^2+|c_j|_2^2}{6\tau^2+|c_j-\mathfrak{c}_j|_2^2}\right)\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} \\ &= \sum_{j=1}^{k}\log\frac{\mathbb{P}\left(|\nu|_2 \leq \frac{2R}{\sqrt{2}\tau_0}\right)}{\mathbb{P}\left(|\nu)|_2 \leq \frac{\xi_j}{\sqrt{2}\tau}\right)} + kd\log\frac{\tau_0}{\tau} \\ &+ \frac{3+d}{2}\int_{\mathbb{R}^{dk}}\sum_{j=1}^{k}\log\left(\frac{\tau^2}{\tau_0^2}\frac{6\tau_0^2+|c_j|_2^2}{6\tau^2+|c_j-\mathfrak{c}_j|_2^2}\right)\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} =: A_1 + A_2 + A_3.\end{aligned} \tag{25}$$

By the definition of the multivariate Student distribution ν

$$\begin{aligned}\mathbb{P}\left(|\nu|_2 \le \frac{\xi_j}{\sqrt{2}\tau}\right) &= \int_{|\nu|_2 \le \frac{\xi_j}{\sqrt{2}\tau}} \frac{\Gamma(\frac{3+d}{2})}{\Gamma(\frac{3}{2})(3\pi)^{\frac{d}{2}}} \left(1+\frac{|\nu|_2^2}{3}\right)^{-\frac{3+d}{2}} \mathrm{d}\nu \\ &\ge \left(1+\frac{\xi_j^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \frac{\Gamma(\frac{3+d}{2})}{\Gamma(\frac{3}{2})(3\pi)^{\frac{d}{2}}} \int_{|\nu|_2 \le \frac{\xi_j}{\sqrt{2}\tau}} \mathrm{d}\nu \\ &= c_d \tau^{-d} \left(1+\frac{\xi_j^2}{6\tau^2}\right)^{-\frac{3+d}{2}} \xi_j^d,\end{aligned}$$

where $\Gamma(\cdot)$ is the Gamma function and

$$c_d = \frac{\Gamma\left(\frac{3+d}{2}\right)}{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{d}{2}+1\right)6^{\frac{d}{2}}}$$

Hence, the term A_1 in (25) verifies

$$\begin{aligned}A_1 &= k\log\mathbb{P}\left(|\nu|_2 \le \frac{2R}{\sqrt{2}\tau_0}\right) - \sum_{j=1}^{k}\log\mathbb{P}\left(|\nu)|_2 \le \frac{\xi_j}{\sqrt{2}\tau}\right) \\ &\le -\sum_{j=1}^{k}\log\mathbb{P}\left(|\nu|_2 \le \frac{\xi_j}{\sqrt{2}\tau}\right) \\ &\le \sum_{j=1}^{k}\left[\frac{3+d}{2}\log\left(1+\frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2}\log\xi_j^2\right] + kd\log\tau - k\log c_d. \qquad (26)\end{aligned}$$

In addition, we have

$$\begin{aligned}\frac{6\tau_0^2 + |c_j|_2^2}{6\tau^2 + |c_j - \mathfrak{c}_j|_2^2} &\le 1 + \frac{2|\mathfrak{c}_j|_2}{2\sqrt{6}\tau}\frac{2\sqrt{6}\tau|c_j - \mathfrak{c}_j|_2}{6\tau^2 + |c_j - \mathfrak{c}_j|_2^2} + \frac{|\mathfrak{c}_j|_2^2}{6\tau^2 + |c_j - \mathfrak{c}_j|_2^2} + \frac{\tau_0^2}{\tau^2} \\ &= 1 + \frac{|\mathfrak{c}_j|_2}{\sqrt{6}\tau} + \frac{|\mathfrak{c}_j|_2^2}{6\tau^2} + \frac{\tau_0^2}{\tau^2} \le \left(1 + \frac{|\mathfrak{c}_j|_2}{\sqrt{6}\tau} + \frac{\tau_0}{\tau}\right)^2,\end{aligned}$$

where we used the Cauchy–Schwarz inequality. Due to the above inequality, the term A_3 in (25) satisfies

$$\begin{aligned}A_3 &\le (3+d)\int \sum_{j=1}^{k}\log\left(1+\frac{\tau}{\tau_0}+\frac{|\mathfrak{c}_j|_2}{\sqrt{6}\tau_0}\right)\rho_k(\mathbf{c})\mathrm{d}\mathbf{c} \\ &\le (3+d)k\int\log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^{k}|\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right)\rho_k(\mathbf{c})\mathrm{d}\mathbf{c}\end{aligned}$$

$$= (3+d)k \log\left(1 + \frac{\tau}{\tau_0} + \frac{\sum_{j=1}^k |c_j|_2}{\sqrt{6}k\tau_0}\right). \tag{27}$$

Combining (24), (25), (26), (27) and (13) completes the proof. □

Corollary 4. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, for any $\lambda > 0$, if q and π_k in (4) are taken respectively as in (6) and (22) with parameter $\eta \geq 0$, $\tau_0 > 0$ and $R \geq \max_{t=1,\dots,T} |x_t|_2$, Algorithm 1 satisfies, for any $0 < \tau^2 \leq (\sqrt{3}R^2)/(6\sqrt{d})$,*

$$\sum_{t=1}^T \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{k\in[\![1,p]\!]} \inf_{\mathbf{c}\in\mathscr{C}(k,R)} \left\{ \sum_{t=1}^T \ell(\mathbf{c}, x_t) + \frac{kd}{\lambda}\log\frac{\tau_0}{c_d\tau} + \frac{\eta}{\lambda}k \right.$$
$$\left. + \frac{(3+d)k}{\lambda}\log\left(1 + \frac{\tau}{\tau_0} + \frac{\sum_{j=1}^k |c_j|_2}{\sqrt{6}k\tau_0}\right) + \frac{1}{\lambda}\sqrt{kd(12\tau^2T\lambda + 3k)}\right\} + \frac{\lambda T}{2}C_1^2 + \frac{\log p}{\lambda},$$

where $C_1 = (2R + \max_{t=1,\dots,T} |x_t|_2)^2$ and $c_d = \left(\frac{\Gamma(\frac{3+d}{2})}{\Gamma(\frac{3}{2})\Gamma(\frac{d}{2}+1)}\right)^{1/d}$.

Proof. By Theorem 1,

$$\sum_{t=1}^T \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \leq \inf_{k\in[\![1,p]\!]} \inf_{\substack{\rho\in\mathscr{P}_\pi(\mathscr{C}) \\ \rho=\rho_k \mathbb{1}_{\{\mathbf{c}\in\mathbb{R}^{dk}\}}}} \left\{ \mathbb{E}_{\mathbf{c}\sim\rho} \sum_{t=1}^T [\ell(\mathbf{c}, x_t)] + \frac{\mathscr{K}(\rho,\pi)}{\lambda} \right.$$
$$\left. + \frac{\lambda}{2}\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)}\mathbb{E}_{\mathbf{c}\sim\rho}\sum_{t=1}^T [\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2 \right\} \tag{28}$$

As in (12), the first term on the right-hand side of (28) may be upper bounded.

$$\sum_{t=1}^T \mathbb{E}_{\mathbf{c}\sim\rho}[\ell(\mathbf{c}, x_t)] \leq \sum_{t=1}^T \ell(m, x_t) + T \max_{j=1,\dots,k} \xi_j^2. \tag{29}$$

For the second term in the right-hand side of (28), by Lemma 2,

$$\frac{\mathscr{K}(\rho,\pi)}{\lambda} \leq \frac{(3+d)k}{\lambda}\log\left(1 + \frac{\tau}{\tau_0} + \frac{\sum_{j=1}^k |c_j|_2}{\sqrt{6}k\tau_0}\right) + \frac{1}{\lambda}\sum_{j=1}^k \left[\frac{3+d}{2}\log\left(1 + \frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2}\log\xi_j^2\right]$$
$$+ \frac{kd}{\lambda}\log\tau_0 - \frac{k}{\lambda}\log c_d + \frac{\eta}{\lambda}(k-1) + \frac{\log p}{\lambda}. \tag{30}$$

Likewise to (15), the third term on the right-hand side of (28) is upper bounded by

$$\frac{\lambda}{2}\mathbb{E}_{(\hat\rho_1,\dots,\hat\rho_T)}\mathbb{E}_{\mathbf{c}\sim\rho_k}\sum_{t=1}^T [\ell(\mathbf{c}, x_t) - \ell(\hat{\mathbf{c}}_t, x_t)]^2 \leq \frac{\lambda T}{2}C_1^2. \tag{31}$$

Combining inequalities (29), (30) and (31) yields for $\xi \in \Xi(k,R)$ and $0 < \tau^2 \le \sqrt{3}R^2/(6\sqrt{d})$ that

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t) \\
&\le \inf_{k\in[\![1,p]\!]} \inf_{\mathfrak{c}\in\mathscr{C}(k,R)} \left\{ \sum_{t=1}^{T} \ell(\mathfrak{c}, x_t) + \xi_j^2 + \frac{(3+d)k}{\lambda} \log\left(1 + \frac{\tau}{\tau_0} + \frac{\sum_{j=1}^{k} |\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right) + T \max_{j=1,\dots,k} \xi_j^2 \right. \\
&\left. + \frac{3+d}{2\lambda} \sum_{j=1}^{k} \log\left(1 + \frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2\lambda} \sum_{j=1}^{k} \log \xi_j^2 + \frac{kd}{\lambda} \log \tau_0 - \frac{k}{\lambda} \log c_d + (k-1) \right\} \\
&+ \frac{\lambda T}{2} C_1^2 + \frac{\log p}{\lambda}.
\end{aligned}
$$

Let $\hat\xi_j = \xi_j^2/6\tau^2$ for any $j = 1,\dots,k$, then $0 < \hat\xi_j \le R^2/6\tau^2$ since $\xi = (\xi_j)_{j=1,\dots,k} \in \Xi(k,R)$. This yields

$$
\begin{aligned}
&T \max_{j=1,\dots,k} \xi_j^2 + \frac{3+d}{2\lambda} \sum_{j=1}^{k} \log\left(1 + \frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2\lambda} \sum_{j=1}^{k} \log \xi_j^2 \\
&= 6\tau^2 T \max_{j=1,\dots,k} \hat\xi_j + \frac{3}{2\lambda} \sum_{j=1}^{k} \log\left(1 + \hat\xi_j\right) + \frac{d}{2\lambda} \sum_{j=1}^{k} \log\left(1 + \frac{1}{\hat\xi_j}\right) - \frac{kd}{2\lambda} \log(6\tau^2) \\
&\le 6\tau^2 T \max_{j=1,\dots,k} \hat\xi_j + \frac{3}{2\lambda} \sum_{j} \hat\xi_j + \frac{d}{2\lambda} \sum_{j=1}^{k} \frac{1}{\hat\xi_j} - \frac{kd}{2\lambda} \log(6\tau^2) \\
&\le \left(6\tau^2 T + \frac{3k}{2\lambda}\right) \max_{j=1,\dots,k} \hat\xi_j + \frac{d}{2\lambda} \sum_{j=1}^{k} \frac{1}{\hat\xi_j} - \frac{kd}{2\lambda} \log(6\tau^2). \qquad (32)
\end{aligned}
$$

The minimum of the right-hand side of (32) is reached in

$$
\hat\xi_1 = \cdots = \hat\xi_k = \sqrt{\frac{kd}{12\tau^2 T\lambda + 3k}} \le \frac{R^2}{6\tau^2}, \quad \text{if } 0 < \tau^2 \le \frac{\sqrt{3}R^2}{6\sqrt{d}}.
$$

Therefore, for fixed k, $\mathfrak{c} \in \mathscr{C}(k,R)$ and $0 < \tau^2 \le \frac{\sqrt{3}R^2}{6\sqrt{d}}$,

$$
\begin{aligned}
&\inf_{\xi\in\Xi(k,R)} \left\{ T \max_{j=1,\dots,k} \xi_j^2 + \frac{3+d}{2\lambda} \sum_{j=1}^{k} \log\left(1 + \frac{\xi_j^2}{6\tau^2}\right) - \frac{d}{2\lambda} \sum_{j=1}^{k} \log \xi_j^2 \right\} \\
&\le \frac{1}{\lambda} \sqrt{kd(12\tau^2 T\lambda + 3k)} - \frac{kd}{2\lambda} \log 6\tau^2.
\end{aligned}
$$

Hence

$$
\sum_{t=1}^{T} \mathbb{E}_{(\hat\rho_1,\hat\rho_2,\dots,\hat\rho_t)} \ell(\hat{\mathbf{c}}_t, x_t)
$$

$$\leq \inf_{k\in[\![1,p]\!]} \inf_{\mathfrak{c}\in\mathscr{C}(k,R)} \left\{ \sum_{t=1}^{T} \ell(\mathfrak{c},x_t) + \frac{(3+d)k}{\lambda} \log\left(1+\frac{\tau}{\tau_0}+\frac{\sum_{j=1}^{k}|\mathfrak{c}_j|_2}{\sqrt{6}k\tau_0}\right) \right.$$

$$\left. +\frac{1}{\lambda}\sqrt{kd(12\tau^2T\lambda+3k)} + \frac{kd}{\lambda}\log\frac{\tau_0}{\sqrt{6}\tau c_d^{1/d}} + \frac{\eta}{\lambda}(k-1) \right\} + \frac{\lambda T}{2}C_1^2 + \frac{\log p}{\lambda}.$$

which concludes the proof. □

Tuning parameters λ, τ and η can be chosen to obtain a sublinear regret bound for the cumulative loss of Algorithm 1.

Corollary 5. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, under the assumptions of Corollary 4, if $T \geq 12d\tau_0^4/c_d^2R^4$, Algorithm 1 with $\lambda = 1/\sqrt{T}$, $\tau^2 = \tau_0^2 T^{-1/2}(c_d)^{-2}$ and $\eta \geq 0$ satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t)$$

$$\leq \inf_{k\in[\![1,p]\!]} \inf_{\mathbf{c}\in\mathscr{C}(k,R)} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c},x_t) + (3+d)k\sqrt{T}\log\left(1+\frac{1}{c_dT^{\frac{1}{4}}}+\frac{\sum_{j=1}^{k}|c_j|_2}{\sqrt{6}k\tau_0}\right) \right.$$

$$\left. +\frac{kd}{4}\sqrt{T}\log T + \left(\sqrt{3k^2d+12\tau_0^2(c_d)^{-2}}+\eta k\right)\sqrt{T} \right\} + \left(\log p + \frac{C_1^2}{2}\right)\sqrt{T}.$$

In the adaptive setting (Algorithm 2), applying Theorem 2 to the specific q and π_k in (6) and (22) leads to the following result.

Corollary 6. *For any deterministic sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, under the assumptions of Corollary 4, set $T \geq 12d\tau_0^4/c_d^2R^4$, $\eta \geq 0$, $R \geq \max_{t=1,\ldots,T}|x_t|_2$ and $\lambda_t = 1/\sqrt{t}$ for any $t \in [\![1,T]\!]$ and $\lambda_0 = 1$. Then Algorithm 2 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1,\hat{\rho}_2,\ldots,\hat{\rho}_t)} \ell(\hat{\mathbf{c}}_t, x_t)$$

$$\leq \inf_{k\in[\![1,p]\!]} \inf_{\mathbf{c}\in\mathscr{C}(k,R)} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c},x_t) + (3+d)k\sqrt{T}\log\left(1+\frac{1}{c_dT^{\frac{1}{4}}}+\frac{\sum_{j=1}^{k}|c_j|_2}{\sqrt{6}k\tau_0}\right) \right.$$

$$\left. +\frac{kd}{4}\sqrt{T}\log T + \left(\sqrt{3k^2d+12\tau_0^2(c_d)^{-2}}+\eta k\right)\sqrt{T} \right\} + \left(\log p + C_1^2\right)\sqrt{T}.$$

Proof. The proof is similar to the proof of Corollary 4, the only difference lies in the fact that (31) is replaced by

$$\mathbb{E}_{(\hat{\rho}_1,\ldots,\hat{\rho}_T)}\mathbb{E}_{\mathbf{c}\sim\rho_k} \sum_{t=1}^{T} \frac{\lambda_{t-1}}{2}[\ell(\mathbf{c},x_t) - \ell(\hat{\mathbf{c}}_t,x_t)]^2 \leq C_1^2\sqrt{T}.$$

□

Figure 6: Estimated number of cells as a function of t. Black lines represent the true number and red dots the estimates. From top to bottom and left to right: PACO, Silhouette, Calinski, Hartigan, Djump, DDSE, Lai, Gap.