K-bMOM: a robust Lloyd-type clustering algorithm based on bootstrap Median-of-Means

Camille Brunet-Saumard[a], Edouard Genetay[b,c], Adrien Saumard[b,*]

[a] *twice.ai, Rennes, France*
[b] *Université de Rennes, Ensai, CREST-UMR 9194, Rennes F-35000, France*
[c] *LumenAI, Pau, France*

Abstract

The median-of-means is an estimator of the mean of a random variable that has emerged as an efficient and flexible tool to design robust learning algorithms with optimal theoretical guarantees. However, its use for the clustering task suggests dividing the dataset into blocks, which may provoke the disappearance of some clusters in some blocks and lead to bad performances. To overcome this difficulty, a procedure termed "bootstrap median-of-means" is proposed, where the blocks are generated with a replacement in the dataset. Considering the estimation of the mean of a random variable, the bootstrap median-of-means has a better breakdown point than the median-of-means if enough blocks are generated. A clustering algorithm called K-bMOM is designed, by performing Lloyd-type iterations together with the use of the bootstrap median-of-means. Good performances are obtained on simulated and real-world datasets for color quantization and an emphasis is put on the benefits of our robust intialization procedure. On the theoretical side, K-bMOM is proven to achieve a non-trivial breakdown point for well-clusterizable situations. Finally, by considering an idealized version of the estimator, robustness is also tackled by deriving rates of convergence for the K-means distortion in the adversarial contamination setting. It is the first result of this kind for the K-means distortion.

Keywords: bootstrap, breakdown point, color quantization, contamination, median-of-means, robust clustering

2010 MSC: 62-07, 62H30, 62P99

*Corresponding author
Email addresses: csaumard@twice.ai (Camille Brunet-Saumard), edouard.genetay@ensai.fr (Edouard Genetay), adrien.saumard@ensai.fr (Adrien Saumard)

1. Introduction

Massive and complex datasets are often corrupted by outliers. Classical data mining procedures such as K-means or more general EM algorithms for instance, are however, sensitive to the presence of outliers, which can induce time consuming data pre-processing.

In this context, robust versions of data mining procedures are particularly relevant and we investigate a way to produce a Lloyd-type algorithm for hard clustering that is robust with respect to the presence of outliers. We propose more precisely using a variant of median-of-means (MOM) statistics, that we call "bootstrap median-of-means" (bMOM). The MOM principle has been the object of recent intensive research in mean estimation, regression, high-dimensional framework and supervised classification and machine learning ([1, 2, 3, 4, 5, 6, 7, 8]). Other approaches to robustness for K-means also exist in the literature, such as for instance, K-median or trimmed K-means [9, 10] to name but a few. The design of robust estimators with a control of the algorithmic complexity has also been investigated [11].

Given a dataset, bMOM consists of first generating a (large) bootstrap sample and then performing a classical median-of-means on this bootstrap sample. This can be seen also as a so-called subragging procedure - for "sub-sample robust aggregating" - in the terminology of Bühlmann [12]. We prove in Section 3.1.1 that if enough blocks are generated from the bootstrap sampling, then for a fixed block size, bMOM has a higher breakdown point than MOM. In other words, bMOM is more robust to contamination than the classical MOM. Note that one strength of bMOM, that will be very useful in the context of clustering, is that sampling is done *with replacement* when constructing the blocks. Hence, the number of blocks for a fixed length is not limited by the amount of initial data, unlike MOM or its variant by sampling without replacement ([13]).

We propose a robust-to-outliers version of K-means, that we call K-bMOM, and that performs Lloyd-type iterations through the use of bMOM estimates of the K-means distortion, as further explained in Section 2. In that section, a robust variant to the traditional K-means++ initialization strategy by applying the MOM principle is also presented.

Theoretical results are summarized in Section 3. We prove in particular in Section 3.1.2, that the K-bMOM algorithm is robust in a sense of a probabilistic version of a breakdown point if the initial data is in a well-clusterizable situation. This is very much in line with the results on the trimmed K-means for example ([14]). We provide in Section 3.2 some deviation bounds for the

performance in terms of K-means distortion of an idealized version of the estimator produced by our algorithm. We consider indeed a minimizer of the median-of-means of the K-means distortion loss along possible codebooks and call it K-MOM. We prove that K-MOM is robust to adversarial contamination of the dataset if the number of outliers is sufficiently small compared to the number of blocks in the MOM statistics. We also prove in Section 3.1.2, that K-MOM has a non-trivial breakdown point in any clustering configuration, which is a strong result, but note that it does directly affect the practice, because the computation of K-MOM is NP-hard like for the K-means.

In Section 4, the scope of application of K-bMOM is illustrated and practical considerations and guidelines are provided for choosing the number and size of the blocks. In Section 5, the proposed initialization procedure and the K-bMOM approach are tested in several simulation settings of outliers. It is also compared to existing robust K-means based clustering approaches in Section 6. And finally, this algorithm is applied to the well-known problem of color quantization in the image processing field.

Our framework is close to the recent work [15] that investigates the use of median-of-means statistics to produce a robust K-means type clustering. However, the latter work is theoretical only and the authors study probabilistic performance bounds for the minimizer of the median-of-means of the K-means distortion loss under a finite second moment assumption. In particular the authors do not discuss the use of median-of-means through Lloyd-type iterations nor a practical way to compute the estimator. Neither do they discuss the possibility of generating blocks with replacements in the dataset.

2. K-bMOM algorithm

We recall first in Section 2.1, the Median-of-Means procedure and introduce a variant, called bootstrap Median-of-Means (bMOM), for the estimation of the mean in dimension one. We then use the latter procedure in a robust iterative clustering algorithm presented in Section 2.2. Our algorithm applies to multi-dimensional data, while performing bMOM estimates of the K-means risk, that is real valued. Moreover, since the resulting partition of most of clustering approaches depends on the starting centers, in Section 2.3 we propose a bMOM-based initialization procedure.

2.1. Median-of-Means and bootstrap Median-of-Means

The median-of-means (MOM) estimator of the mean in dimension one consists of taking a median of some arithmetic means computed on a collection of disjoint blocks $(x_i)_{i\in b_j}$, where $\{b_j : j \in \{1,...,B\}\}$ form a partition of the set of indices {1,...,n} of a real valued sample $x_1^n = (x_1,...,x_n)$. The number of blocks is thus equal to B. The lengths of the blocks are generally taken to be equal, possibly up to one data. By denoting b_1^B the collection of blocks, we can thus write

$$\mathrm{MOM}(x_1^n, b_1^B) = \mathrm{med}\left\{ \frac{1}{\#b_j} \sum_{i\in b_j} x_i : j \in \{1,...,B\} \right\}.$$

where $\#b_j$ denotes the cardinal of b_j and med is a median, that is $\#\{j \in \{1,...,B\}\,; a_j \leq \mathrm{med}\{a_l\}\} \geq B/2$ and $\#\{j \in \{1,...,B\}\,; a_j \geq \mathrm{med}\{a_l\}\} \geq B/2$. In the following, when the set of possible medians is not a singleton, we always consider its middle point as being our choice of median, that is thus uniquely defined.

We may consider that the blocks are generated according to a random drawing process, that proceeds without replacements (disjoint blocks) and according to the uniform distribution over the remaining data at each step. This formulation naturally leads to consider more general random block generating processes.

For any positive integers n_B and B, denote $q = Bn_B$ and generate a bootstrap sample $y_1^q = (y_1,...,y_q)$ from the dataset x_1^n. More precisely, each y_i is taken uniformly at random from the values $(x_1,...,x_n)$ and independently from $(y_j)_{j\neq i}$. The boostrap median-of-means (bMOM) of the dataset x_1^n with parameters n_B and B is then the (classical) MOM estimator over the boostrap sample y_1^q with blocks $b_j = (n_B(j-1)+1,...,n_Bj)$ for $j \in \{1,...,B\}$,

$$\mathrm{bMOM}(x_1^n, n_B, B) = \mathrm{MOM}(y_1^q, b_1^B).$$

Note that the bMOM is a randomized estimator. Also, for any fixed sample size n, we can choose any block size n_B and number of blocks B to define a bMOM estimator, unlike the classical MOM, where the product of the block size with the number of blocks should be equal to the sample size. This will turn out to be extremely useful in the clustering context, where we do not want too small block sizes in order to avoid the disappearance of some clusters in the blocks.

2.2. A robust Lloyd-type algorithm

In this section we propose an estimation procedure based on bMOM statistics for clustering unlabeled data.

Let us introduce the following notations. Let $x_1, \ldots, x_n \in \mathbb{R}^p$ denote a dataset of n observations that we want to cluster into K homogeneous groups. Then $b \in \{1, \ldots, B\}$ stands for the index of a block b and $B \in \mathbb{N}^*$ the number of blocks, containing at least $n_B > K$ datapoints. We define the empirical risk of the block b as:

$$R_b(\mathbf{c}) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k^{(b)}} \left\| x_i^{(b)} - c_k^{(b)} \right\|^2$$

where $x_i^{(b)}$ stands for the ith datapoint contained in the block b, $\mathcal{C}_k^{(b)}$ stands for the set of datapoints belonging to cluster k in the block b and $\|.\|$ is the Euclidean norm. Furthermore, $c_k^{(b)}$ stands for the mean vector of the cluster k in block b. Finally, we denote by $\mathcal{P}(\mathbf{c})$, the Voronoï partition obtained from the set of centroids $\mathbf{c}$.

The K-bMOM algorithm

Due to the nature of the bMOM statistics and the clustering goal, the algorithm that we propose alternates three main steps. At iteration t, and given the centers fitted in the median block of the previous iteration, B blocks of n_B data are built by uniform sampling with replacement. Then, a partition per block is computed by assigning each data point to its closest centroids fitted on the median block at iteration $(t-1)$. The centroids of each block are updated according to their block partition and the empirical risk $\hat{R}_b(\mathbf{c})$ is returned. The block with the median empirical risk is selected and the fitted centers of this median block become the current ones. These steps are repeated several times. The final partition over the whole dataset is obtained by assigning each data point to its nearest closest centroid $\left(\hat{c}_1^{(bmed)}, \ldots, \hat{c}_K^{(bmed)}\right)$ of the current median block. A pseudo algorithm of this procedure is detailed in Algorithm 1.

Our algorithm shares some similarity with the techniques of so-called consensus/ensemble clustering ([16]), since it amounts at each step to producing a robust clustering, given by a codebook, from a collection of candidates computed on bootstrap sub-samples. However, there are also essential differences, since we select one of the candidates by a simple median criterion for dimension one statistics, whereas consensus clusterings aggregate the candidates in a more complicated fash-

ion, using some similarity measures between clusterings. Interestingly, so-called bagged clustering ([17, 18, 19]) proposes performing clusterings on bootstrap samples and to aggregate them using a hierarchical clustering on the collection of obtained centroids. But the size of the bootstrap samples are equal to the original sample size, whereas in our approach the sub-sampling is crucial and directly related to the allowed proportion of outliers (see Section 3.1.2). A robust trimmed clustering approach for probabilities in Wasserstein space is developed in [20] and used to advantage to robustly aggregate model-based clusterings on multivariate data - each clustering being seen as a probability - that are previously learned on sub-samples of the original data. This approach, however, concerns the robust aggregation of model-based clusterings, whereas our focus is on robust hard clustering in the context of the K-means problem.

Algorithm 1: Iteration phase structure

Input: $\{x_1, \ldots, x_n\}$, B the number of blocks and n_B size of blocks ($n_B > K$)

initialization: Let $(c_1, \ldots, c_K)$, K initial centroids.

Set: $q = 0$.

Main Loop: while $q < \bar{q}_{max}$:

1. Create B blocks of the data of size n_B randomly and uniformly with replacement
2. In each block b:
   - Assign each data point to its closest centroid.
   - If $n_k^{(b)} > 1$, $\forall k \in \{1, \ldots, K\}$:
     - for $k \in \{1, \ldots, K\}$: $c_k^{(b)} \leftarrow 1/n_k^{(b)} \sum_{i \in \mathcal{C}_k} x_i^{(b)}$
     - $\hat{R}_b(\mathbf{c}) \leftarrow \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k^{(b)}} \left\| x_i^{(b)} - \hat{c}_k^{(b)} \right\|^2$
   - Else, skip the block.
3. Get the median empirical risk $\hat{R}_{bmed}(\mathbf{c})$ and the associated quantities of the median block : b_{med}, $\left(\widehat{c}_1^{(bmed)}, \ldots, \widehat{c}_K^{(bmed)}\right)$.
4. $q \leftarrow q + 1$

Output: $\left(\bar{c}_1^{(bmed)}, \ldots, \bar{c}_K^{(bmed)}\right)$ and $\mathcal{P}(\bar{\mathbf{c}}^{(bmed)})$

Stopping criterion

In practice, the algorithm is run a given number of maximum iterations ($\bar{t} = 25$ by default). In order to obtain a more precise estimation of centroids at the end of the maximum number of iterations, instead of retrieving the centroids of the median block computed in the last iteration, centroids of the last 10 iterations are aggregated $\bar{\mathbf{c}} = (\bar{c}_1^{(bmed)}, \ldots, \bar{c}_K^{(bmed)})$ such as $\bar{c}_k^{(bmed)} = 1/10 \sum_{i=0}^{9} \widehat{c}_{k,\bar{t}-i}^{(bmed)}$ where $\widehat{c}_{k,t}^{(bmed)}$ stands for the centroid of the cluster k of the median block at iteration t.

Model selection

In model-based clustering, it is frequent to consider several models in order to find the most appropriate one for the considered data. In particular, for most clustering algorithms, the model is specified by its number of clusters K. There are lots of ad-hoc approaches in the literature to select the number of components K and we can therefore think of the Gap statistics from [21], the Silhouette criterion and so one. However, since the K-means algorithm can be seen as a hard version of an EM-like algorithm which tries to estimate a mixture of K Gaussians with isotropic covariance matrices, we can therefore apply classical tools for model selection including BIC, ICL criteria and the heuristic slope [22] for example. We use such criteria on the proposed robust version of the K-means by processing the K-bMOM on several values of K, computing the BIC criterion for each model and select the model defined by its number of components which maximizes it.

2.3. A robust initialization

It is well-known that since the clustering problem is non-convex, the initialization step is a keystone for the resulting partition. We propose therefore, a robust variant of traditional initialization strategies by applying the MOM principle. To do so, the idea is to build uniformly and with replacement, B blocks of n_B datapoints where the number of points is strictly greater than the number of groups. A traditional K-means++ initialization [23] is operated in each block. Such an approach proceeds in an iterative way: it starts with a centroid picked at random among the data points. Iteratively and until the number of groups K is reached, a new centroid is then chosen from the data points with a probability which increases exponentially with the distance $D^2(x, c)$ to the already chosen closest centers. In each block, the empirical risk is therefore, computed and the centers linked to the median empirical risk, called the median block, are selected as the initial

centers. This algorithm is summarized in Algorithm 2. Let us note that the robustness of this initialization scheme is evaluated in Section 5.2.

Algorithm 2: initialization strategy

Input: the dataset $\{x_1, \ldots, x_n\}$, B the number of blocks and $n_B > K$ size of blocks

1. Iterate from 1 until B blocks:
   (a) Select at random, uniformly and with n_B replacement data points
   (b) Proceed a K-means++ initialization
   (c) Compute the empirical risk $\hat{R}_b(\mathbf{c})$ of block b
2. Select the centers from the block having the median empirical risk and get: $\left(\widehat{c}_1^{(bmed)}, \ldots, \widehat{c}_K^{(bmed)}\right)$.

Output: $\left(\widehat{c}_1^{(bmed)}, \ldots, \widehat{c}_K^{(bmed)}\right)$

3. Theoretical analysis

3.1. Breakdown points

We prove in Section 3.1.1 below that bMOM enables us to perform a more robust mean estimation than MOM, if enough blocks are generated, in the sense that the breakdown point of the bMOM is higher. This indeed has an interest for us, since we use the bMOM statistics in K-bMOM to provide a robust estimate of the risk, which is a real-valued quantity corresponding to the mean of the K-means loss (see Section 2). This result still has some limitations however in the perspective of clustering, since K-bMOM does not correspond to the bMOM estimator for $K = 1$. In Section 3.1.2 therefore, we added a study of the breakdown points of some K-MOM estimator and of our K-bMOM algorithm.

3.1.1. Breakdown points for mean estimation

The breakdown point is a classical concept of robust statistics ([24, 25]), that gives the maximal proportion of outliers that is allowed so that the deviations of the estimator stay bounded compared

to the no-corruption setting.

Assume that we are given a sample $x_1^n = (x_1, ..., x_n)$ of real valued random variables.

Definition 1 (Deterministic breakdown point) *The (deterministic) breakdown point $\delta_n(T_n, x_1^n)$ of a real-valued estimator T_n given the sample x_1^n, is the maximum proportion of outliers that leaves the value of the estimator bounded.*

$$\delta_n(T_n, x_1^n) = \frac{1}{n} \max \left\{ m; \max_{i_1,...,i_m} \sup_{y_1,...,y_m} |T_n(z_1, ..., z_n)| < +\infty \right\},$$

where the sample $(z_1, ..., z_n)$ is obtained by replacing the m data points $x_{i_1}, ..., x_{i_m}$ of the sample x_1^n by arbitrary values $y_1, ..., y_m$.

Definition 1 corresponds to a worst case analysis, the outliers potentially appearing at the worst places for the estimator T_n. If the estimator T_n is randomized - rather denoted T_n^ω in this case -, then its breakdown point is a random variable.

For the median $\text{med}\{x_1^n\}$, it holds that $\delta_n(\text{med}\{x_1^n\}, x_1^n) = \lfloor (n-1)/2 \rfloor / n$ and for the empirical mean, $\bar{x}_n = 1/n \sum_{i=1}^n x_i$, $\delta_n(\bar{x}_n, x_1^n) = 1/n$.

Proposition 1 *The breakdown point of the median-of-means estimator is*

$$\delta_n\left(\text{MOM}(x_1^n, b_1^B), x_1^n\right) = \frac{\left\lfloor \frac{B-1}{2} \right\rfloor}{n}.$$

The proof of Proposition 1 is direct since MOM diverges if and only if there is at least one outlier in a majority of blocks. Note that the same breakdown point is achieved for a class of a much more general estimator of a multi-dimensional mean called the median-of-means tournament ([26]).

Note that [2, Section 4.2] proposes automatically selecting the number of blocks of the MOM estimator by a Lepskii-type procedure. This consists of choosing the smallest number of blocks, such that the intersection of some confidence intervals constructed for MOM with greater numbers of blocks, is empty. The resulting estimator will inherit from the value of the breakdown point corresponding to the highest number of blocks in the considered collection. If the highest number of blocks is equal to the sample size n, thus corresponding to a median, then the method of intersection of confidence intervals, gives an optimal value of the breakdown point, corresponding to $\lfloor n/2 \rfloor / n$.

Computing such selection procedure is however, time consuming and as we want to make an iterative use of (bootstrap) MOM estimates, this method seems to be out of scope for us. Instead, we

show below that the use of replacements while constructing the blocks, already gives an improvement of the breakdown point if enough blocks are considered, compared to the use of disjoint blocks when applied to MOM statistics.

Proposition 2 *Assume first that the sample size satisfies $n = Bn_B$. We then have*

$$\delta_n\left(\mathrm{bMOM}(x_1^n, n_B, B), x_1^n\right) \leq \delta_n\left(\mathrm{MOM}(x_1^n, b_1^B), x_1^n\right) \ a.s.$$

Secondly, fix the block size n_B and the sample size n and let the number of blocks taken in the bMOM, tend to infinity. It holds that

$$\lim_{B\to+\infty} \delta_n\left(\mathrm{bMOM}(x_1^n, n_B, B), x_1^n\right) = 1 - \frac{1}{2^{1/n_B}} > \frac{1}{2n_B} \ a.s.$$

Note that $1 - \frac{1}{2^{1/n_B}} \sim_{n_B\to+\infty} \frac{\log 2}{n_B} \simeq \frac{0.69}{n_B}$.

Proof. The first display follows from the fact that the blocks taken in bMOM, are not necessarily disjoint, which could cause damage if repeated data are outliers. For the second display, assume that the sample is corrupted by m outliers. Denote S_i, the indicator that the block B_i is not corrupted. Then S_i is a Bernoulli random variable of mean $(1 - m/n)^{n_B}$. Then $\sup_{y_1,\ldots,y_m} |\mathrm{bMOM}(x_1^n, n_B, B)|$ is finite if the proportion of corrupted blocks is less than $1/2$. This corresponds to the condition $\sum_{i=1}^B S_i/B > 1/2$. By the strong law of large numbers, the latter inequality is almost surely realized asymptotically if $(1 - m/n)^{n_B} > 1/2$, hence the result. ■

On the one hand, the first display in Proposition 2 states that when the number of blocks in bMOM is equal to the number of blocks in MOM, bMOM has a breakdown point that is less than or equal to the breakdown point of MOM. On the other hand, the second display in Proposition 2 states that for a fixed block size, when the number of blocks in bMOM tends to infinity, its breakdown point tends to a value that is strictly greater than the breakdown point of MOM taken with the same block size.

Considering that the contaminated sample is given (fixed), it is interesting to evaluate the probability that a randomized estimator does not diverge when the outliers go to infinity. It can indeed happen that the indices of the outliers are not the worst with respect to the block drawing process. This leads to the following definition.

Definition 2 (Probabilistic breakdown point) *The probabilistic breakdown point of a random-*

ized estimator T_n^ω *given the sample* x_1^n *is*

$$p_n\left(T_n^\omega, x_1^n, (i_1, ..., i_m)\right) = \mathbb{P}\left(\left\{\omega : \sup_{y_1,...,y_m} |T_n^\omega(z_1, ..., z_n)| < +\infty\right\}\right)$$

where the sample $(z_1, ..., z_n)$ *is obtained by replacing the* m *data points* $x_{i_1}, ..., x_{i_m}$*, for some fixed indices* $(i_1, ..., i_m)$*, by the arbitrary values* $y_1, ..., y_m$*.*

As $p_n\left(\text{bMOM}(x_1^n, n_B, B), x_1^n, (i_1, ..., i_m)\right)$ only depends on m, but not on the values of $(i_1, ..., i_m)$, we will rather denote it $p_n\left(\text{bMOM}(x_1^n, n_B, B), m\right)$. We have the following bound.

Proposition 3 *Assume that the block length* n_B *in bMOM and the proportion of outliers* m/n *are such that* $(1 - m/n)^{n_B} > 1/2$*. Then it holds that*

$$p_n\left(\text{bMOM}(x_1^n, n_B, B), m\right) \geq 1 - \exp\left(-2B\left((1 - m/n)^{n_B} - 1/2\right)^2\right).$$

Proof. As in the proof of Proposition 2, denote S_i, the indicator that the block B_i is not corrupted. As $(1 - m/n)^{n_B} > 1/2$, we have by Hoeffding's inequality ([27, Theorem 2.27]),

$$\begin{aligned}
\mathbb{P}\left(\left\{\omega : \sup_{y_1,...,y_m} |\text{bMOM}(x_1^n, n_B, B)| = +\infty\right\}\right) &= \mathbb{P}\left(\sum_{i=1}^{B}(1 - S_i) > B/2\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{B}(1 - S_i) - \mathbb{E}[1 - S_i]) > B(1 - m/n)^B - B/2\right) \\
&\leq \exp\left(-2B\left((1 - m/n)^{n_B} - 1/2\right)^2\right).
\end{aligned}$$

■

If the number of outliers m and the sample size n are fixed, then the block length n_B should be such that $(1 - m/n)^{n_B} > 1/2$, i.e. $n_B < \log(2)/\log(1/(1 - m/n))$ (Figure 1). Hence, in case of a large proportion of outliers m/n, the block length should not be taken too large. Furthermore, by denoting $D = (1 - m/n)^{n_B} - 1/2 > 0$, we have that $p_n\left(\text{bMOM}(x_1^n, n_B, B), m\right) \geq 1 - R$ if $B > \log(1/R)/\left(2D^2\right)$. We illustrate the behavior of the latter lower bound on the block size in Figure 2. This implies in particular that if the block size n_B is rightly chosen (not too large according to the proportion of outliers), then the probability that the bMOM remains stable under the adversarial contamination tends to 1 when the number of blocks B tends to infinity.

Another direction for discussing robustness properties of bMOM for mean estimation would be to investigate sub-Gaussian deviation bounds, that were obtained in [2] for MOM. We leave it as an interesting open problem.

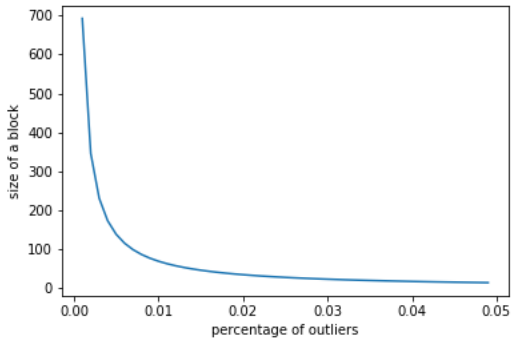


Figure 1: Maximum admissible block size n_B for bMOM and K-bMOM as a function of the proportion of outliers $p = m/n$.

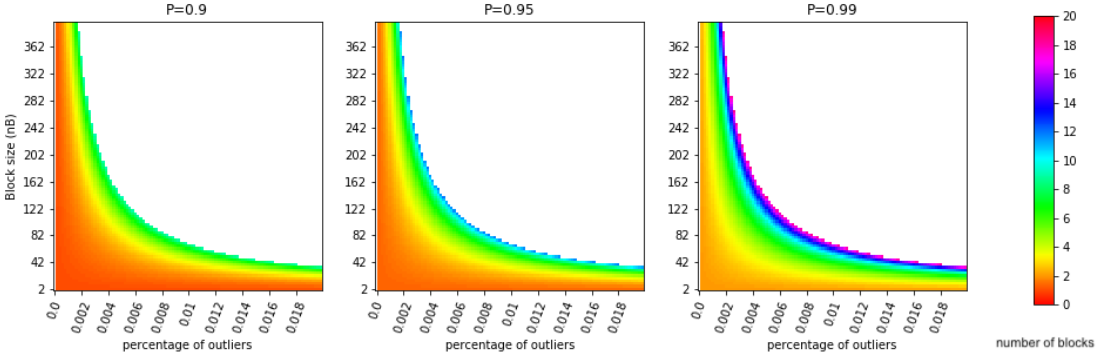


Figure 2: Evolution of the lower bound on the number of blocks (colorbar) as a function of the proportion of outliers and the size of the blocks in bMOM for different levels of confidence.

3.1.2. Some breakdown points for robust K-means estimation

In this section we consider that the data at hand $x_1^n = (x_1, ..., x_n)$ take values in a separable Hilbert space $(\mathcal{X}, \|\cdot\|)$. Typically $\mathcal{X} = \mathbb{R}^d$, as in our experiments, but we do not need to specify further the set $\mathcal{X}$ for now.

We discuss the values of some breakdown points for some robust estimates in the K-means problem. The task is thus to estimate a codebook $\mathbf{c}_*$ in the set C_* of optimal codebooks, given by

$$C_* = \arg \min_{\mathbf{c}=\{c_1,...,c_K\} \in \mathcal{X}^K} \left\{ \mathbb{E}\left[\min_{j=1,...,K} \|X - c_j\|^2 \right] \right\},$$

where the random variable X follows the true distribution of the data and K stands for the number of classes. Let us first consider the estimators studied in [15] and in Section 3.2 below. These estimators can be formulated as

$$\hat{\mathbf{c}}_n \in \arg \min_{\mathbf{c} \in \mathcal{S}^K} \left\{ \mathrm{MOM}\left(\ell_{\mathbf{c}}\right) \right\}, \tag{1}$$

where $\ell_{\mathbf{c}}$ is the standard K-means loss (see Section 3.2) and $\mathcal{S}$ is a subset of the space $\mathcal{X}$. In the minimization problem (1), the blocks are chosen once and for all and the quantities $\mathrm{MOM}(\ell_{\mathbf{c}})$ are computed using the same blocks. For simplicity, we assume that the blocks have the same length n_B, so that $n = Bn_B$, where B is the number of blocks taken in the MOM estimates. Different choices can be made for $\mathcal{S}$ ([15]), but in essence its value does not affect the breakdown point study. For convenience, we take $\mathcal{S} = \mathcal{X}$ in the following of this section.

Definition 3 (Breakdown point for K-MOM) *The breakdown point $\delta_n(\hat{\mathbf{c}}_n, x_1^n)$ of the codebook $\hat{\mathbf{c}}_n$, defined in (1) with $\mathcal{S} = \mathcal{X}$ and given the sample x_1^n, is the maximum proportion of outliers that leaves the norms of the centroids bounded,*

$$\delta_n(\hat{\mathbf{c}}_n, x_1^n) = \frac{1}{n} \max \left\{ m; \max_{i_1,...,i_m} \sup_{y_1,...,y_m} \max_{c \in \hat{\mathbf{c}}_n(z_1,...,z_n)} \|c\| < +\infty \right\},$$

where the sample $(z_1, ..., z_n)$ is obtained by replacing the m data points $x_{i_1}, ..., x_{i_m}$ of the sample x_1^n by arbitrary values $y_1, ..., y_m$ and $\hat{\mathbf{c}}_n(z_1, ..., z_n)$ is the codebook defined in (1) and learned from the sample $(z_1, ..., z_n)$.

Theorem 1 *Let the K-MOM estimator $\hat{\mathbf{c}}_n$ be defined by (1) with $\mathcal{S} = \mathcal{X}$ and B blocks for the MOM estimates. For any dataset x_1^n, its breakdown point is given by*

$$\delta_n(\hat{\mathbf{c}}_n, x_1^n) = \frac{\left\lfloor \frac{B-1}{2} \right\rfloor}{n}. \tag{2}$$

Proof. Let us prove first that the lower bound: $\delta_n(\hat{\mathbf{c}}_n, x_1^n) \geq \lfloor (B-1)/2 \rfloor / n$. Set $r > 0$ such that all regular data are in a ball $B(0,r)$ centered at the origin and of radius r, that is: $\max_{i=1,...,n} ||x_i|| \leq r$. If the number of outliers is less than or equal to $\lfloor (B-1)/2 \rfloor$, then a majority of blocks do not contain any outlier, so the median values that give the quantities $\mathrm{MOM}(\ell_{\mathbf{c}})$ are sandwiched in the risk values along these regular blocks. Now, consider a centroid c_i, $i \in \{1, ..., K\}$ such that $||c_i|| > r$. The projection $\bar{c}_i$ of c_i onto the ball $B(0,r)$ is then at a lesser distance to any data point than c_i. Hence, c_i cannot be a centroid of a codebook solution of (1) for the contaminated sample $(z_1, ..., z_n)$, since replacing it with $\bar{c}_i$ would decrease the risk on each regular block and so, would also decrease the median of the risks along the blocks. For any $c \in \hat{\mathbf{c}}_n(z_1, ..., z_n)$, we thus get $||c|| \leq r$, which proves the lower bound.

Let us prove now the upper bound: $\delta_n(\hat{\mathbf{c}}_n, x_1^n) \leq \lfloor (B-1)/2 \rfloor / n$. Consider that the number of outliers is strictly larger than $\lfloor (B-1)/2 \rfloor$ and that a majority of blocks contains at least one outlier. Take all the outliers to be equal to a vector y and set a constant $r > 0$ such that $\max_{i=1,...,n} ||x_i|| \leq r$. We will show that the procedure breaks down when y diverges to infinity. We proceed by *reductio ad absurdum*. Let us assume that there exists a constant $t > 0$ such that for any value y of the outliers, there exists a codebook $\mathbf{c}(y) = \{c_1(y), ..., c_K(y)\}$ solution of (1) for the contaminated sample $(z_1, ..., z_n)$ and depending on y, such that $\max_{i=1,...,K} ||c_i(y)|| \leq t$. Then if $||y|| > t$, the risk on each contaminated block is larger than the value $(||y|| - t)^2/n_B$ due to the contribution of y. As y is contained in a majority of blocks, we also have $\mathrm{MOM}(\ell_{\mathbf{c}(y)}) \geq (||y|| - t)^2/n_B$, for the MOM computed with the contaminated sample $(z_1, ..., z_n)$. Now, take other codebooks $\tilde{\mathbf{c}}(y) = (\tilde{c}_1(y), ..., \tilde{c}_K(y))$ where $\max_{i=1,...,K-1} ||\tilde{c}_i(y)|| \leq r$ and $\tilde{c}_K(y)) = y$. We get $\mathrm{MOM}(\ell_{\tilde{\mathbf{c}}(y)}) \leq 4r^2 < (||y|| - t)^2/n_B$, for $||y||$ sufficiently large and for the MOM computed with the contaminated sample $(z_1, ..., z_n)$, which gives a contradiction with $\mathbf{c}(y)$ being a solution of (1), for the contaminated sample $(z_1, ..., z_n)$. The upper bound is thus proven, which concludes the proof. ■

Theorem 1 is based on the fact that for a smaller amount of outliers than the value of the breakdown point in (2), a majority of blocks does not contain any outlier. Hence the medians along the blocks stay bounded for bounded centroids, whatever the values of the outliers. As noted in Section 3.1.1 above, the same breakdown point is attained for the median-of-mean tournament in multi-dimensional mean estimation ([26]).

The K-MOM procedure (1) has the strong feature of achieving a non-trivial breakdown point *in any clustering configuration*, as opposed to other robust K-means procedures such as for instance

the trimmed K-means, that requires a so-called well-clusterizable configuration where clusters are sufficiently "compact" and sufficiently "separated" ([14] ; see also [28] and references therein).

Theorem 1 complements the results available for K-MOM in [15] where it is shown that it achieves sub-Gaussian bounds when the distribution of data has only a finite second moment and in Section 3.2 below where we give rates of convergence in the presence of outliers.

K-MOM as defined in (1) has however essentially a theoretical interest, since the minimization problem is intractable in general. Thus, we turn now to the study of the breakdown point of the K-bMOM algorithm presented in Section 2.2 above. As it produces a randomized estimator, we denote it $\bar{\mathbf{c}}^{\omega}$ rather than $\bar{\mathbf{c}}$, the notation ω accounting for randomization. We discuss the following notion of probabilistic breakdown point.

Definition 4 (Probabilistic breakdown point for K-bMOM) *The probabilistic breakdown point $p_n\left(\bar{\mathbf{c}}^{\omega}, x_1^n, (i_1, ..., i_m)\right)$ of the randomized codebook $\bar{\mathbf{c}}^{\omega} = \bar{\mathbf{c}}$, defined in Section 2.2 as the output of the algorithm K-bMOM, is given by*

$$p_n\left(\bar{\mathbf{c}}^{\omega}, x_1^n, (i_1, ..., i_m)\right) = \mathbb{P}\left(\left\{\omega : \sup_{y_1,...,y_m} \max_{c \in \bar{\mathbf{c}}^{\omega}(z_1,...,z_n)} \|c\| < +\infty\right\}\right) ,$$

where the sample $(z_1, ..., z_n)$ is obtained by replacing the m data points $x_{i_1}, ..., x_{i_m}$, for some fixed indices $(i_1, ..., i_m)$, by the arbitrary values $y_1, ..., y_m$ and $\bar{\mathbf{c}}^{\omega}(z_1, ..., z_n)$ is the output of K-bMOM learned from $(z_1, ..., z_n)$.

The K-bMOM algorithm will be proven to be robust in terms of a probabilistic breakdown point in the case of a "well-clusterizable" clustering configuration, that is a classical assumption for obtaining robustness in clustering ([9, 28]). Roughly speaking, a well-clusterizable configuration is made of "compact" clusters that are well "separated". We give the following formal definition, suitable for our needs.

Definition 5 *A dataset x_1^n is said to be in a well-clusterizable configuration, with compactness parameter r and separation parameter R satisfying $R > 2r > 0$, if the points x_1^n lie in a union of K disjoint balls $B(a_i, r)$, $i = 1, ..., K$, of radius r with centers a_i separated from each other by at least a distance $R : \min_{i \neq j} ||a_i - a_j|| \geq R$. Each ball $B(a_i, r)$ is assumed to contain exactly one cluster.*

Theorem 2 *Let $\bar{\mathbf{c}}^{\omega}$ be the K-bMOM codebook, computed using bMOM statistics with block size n_B and number of blocks B. Assume that the block length n_B and the proportion of outliers m/n are*

such that $(1-m/n)^{n_B} > 1/2$. *Assume furthermore that the regular data points* x_1^n *are in a well-clusterizable situation, with compactness and separation parameters denoted respectively* r *and* R, *such that* $R^2 > 16 n_B r^2$. *Finally, assume that at the beginning of the last* 10 *iterations, the algorithm has identified the right partition of the regular data, meaning that one cluster is associated to one centroid. Then it holds* $p_n(\bar{\mathbf{c}}^\omega, x_1^n, (i_1, ..., i_m)) \geq \max\{p_1 - p_2, 0\}$ *with*

$$p_1 = \left(1 - \sum_{i=1}^{K} \left(1 - \frac{n_i^r}{n}\right)^{n_B}\right)^{10B} \tag{3}$$

and

$$p_2 = 10 \exp\left(-2B\left(\left(1 - \frac{m}{n}\right)^{n_B} - \frac{1}{2}\right)^2\right), \tag{4}$$

where the quantity n_i^r *in display (3) stands for the number of regular data belonging to cluster* i *in the sample* $(z_1, ..., z_n)$.

Proof. As described in Section 2.2, the output of the K-bMOM algorithm $\bar{\mathbf{c}}^\omega$ is given by the average of the last 10 codebooks computed through the iterations. Let us first prove the following property (P): if all the K clusters are represented in each of the blocks generated during the last 10 iterations and if in each of these iterations, a majority of blocks are made of regular data only, the procedure does not break down.

Indeed, consider the first of the last 10 iterations, when each cluster is represented in all the B blocks, a majority of which consisting of regular data only. For such regular blocks, as the algorithm has found the right partition, the new centroids are barycenters of data in one cluster, that belong to a ball of radius r and so the risk of the new codebooks in regular blocks is less than $4r^2$.

Now, consider a block that contains at least one outlier. Set $A > 0$ such that the regular data is contained in a ball centered at the origin and of radius A. After updating the centroids in this block, two cases are possible. Either one of the centroids is a barycenter between some regular data and some outliers, among which an outlier denoted y that has the greatest norm in the dataset z_1^n. The risk in this configuration is thus greater than $(||y|| - A)^2/(4n_B)$, where $||y||$ is assumed to be greater than A without loss of generality.

Or, the second case could be that the outlier y with the greatest norm lies in a cluster that does not contain any regular data. Hence, at most $K-1$ centroids are assigned to the regular data. Consequently, one centroid is a barycenter between regular data points of at least two clusters and the risk in this configuration is greater than $R^2/(4n_B)$. Finally, we see that if $R^2 > 16 n_B r^2$ and

$(||y|| - A)^2 > 16 n_B r^2$, then the risk of any regular block is less than the risk of any block containing some outliers. As a majority of blocks are regular, selecting the set of centroids achieving the median of the risks along the blocks, will give a codebook corresponding to a regular block. As we already noted, these centroids induce the right partition in the sense that they are associated with data in one cluster only. Hence, the reasoning extends to the next iterations and property (P) is proven.

To obtain a lower bound on the probabilistic breakdown point $p_n(\bar{\mathbf{c}}^\omega, x_1^n, (i_1, ..., i_m))$, it suffices now to provide a lower bound on the probability of the event described in property (P). Denote E_1 by the event where the K clusters are represented in each block generated during the last 10 iterations and E_2 the event where a majority of blocks are contaminated by some outliers in at least one of the last 10 iterations. By the first part of the proof, it holds

$$p_n(\bar{\mathbf{c}}^\omega, x_1^n, (i_1, ..., i_m)) \geq \mathbb{P}\left(E_1 \bigcap E_2^c\right) \geq \mathbb{P}(E_1) - \mathbb{P}(E_2)\,.$$

By denoting F_1 by the event where one block, considered to be fixed, contains representatives of the K clusters, we get $\mathbb{P}(E_1) = (\mathbb{P}(F_1))^{10B}$ by independence of the block generating process. Now, considering events where the data in the considered block come from outliers and all but one cluster, we obtain $\mathbb{P}(F_1) \geq 1 - \sum_{i=1}^K (1 - n_i^r/n)^{n_B}$ where n_i^r is the number of regular data in cluster i, which gives

$$\mathbb{P}(E_1) \geq \left(1 - \sum_{i=1}^K \left(1 - \frac{n_i^r}{n}\right)^{n_B}\right)^{10B} = p_1\,.$$

To conclude, it remains to bound $\mathbb{P}(E_2)$ from above by the quantity p_2. Denote S_i the indicator that the block B_i is not corrupted, for some generation of B blocks of length n_B. A simple union bound along the 10 iterations gives

$$\mathbb{P}\left(E_2\right) \leq 10 \times \mathbb{P}\left(\sum_{i=1}^B (1 - S_i) > \frac{B}{2}\right)\,.$$

Since $(1 - m/n)^{n_B} > 1/2$, we apply by Hoeffding's inequality as in the proof of Proposition 3 and get

$$\mathbb{P}\left(\sum_{i=1}^B (1 - S_i) > B/2\right) \leq \exp\left(-2B\left(\left(1 - \frac{m}{n}\right)^{n_B} - \frac{1}{2}\right)^2\right)\,.$$

Putting the two latter inequalities together gives the desired upper bound on $\mathbb{P}(E_2)$ and concludes the proof. ■

To conclude this section, let us make some comments on Theorem 2. We assume in Theorem 2 that the K-bMOM algorithm is not too far from the solution - given by the set of optimal codebooks C_* - by postulating that it has found the right partition at the beginning of the last 10 iterations. This assumption seems legitimate, since analyzing the behavior of clustering algorithms in a neighborhood of the optimal solutions, by assuming a "warm start" for instance, is very classical. We could also assume a warm start by requiring that the initialization procedure has found the right partition, at the price of considering the total number of iterations of K-bMOM instead of the last 10 iterations. But as we only retain the last 10 codebooks selected by the algorithm to produce its output, we find it more realistic to focus on the last 10 iterations only or equivalently, controlling the behavior of all the iterations seems too restrictive.

We also assume that the data is in a well-clusterizable configuration with compactness and separation parameters r and R that satisfy $R^2 > 16 n_B r^2$. Again, the necessity of this relation is surely pessimistic compared to practice, but it allows us to deduce a non-trivial lower bound for the probabilistic breakdown point with a reasoning that is kept rather simple. The rationale to keep in mind for practice is that if r and R can be well defined, then when the ratio R/r increases, the algorithm is more likely to be robust with respect to the presence of some outliers.

We see from Theorem 2 that the K-bMOM algorithm has a high probability to be robust if the quantity p_1 defined in (3) is close to 1 and the quantity p_2 given by (4) is close to 0. To analyze p_1, note that if the clusters are well-balanced and if the proportion of outliers is not too large, then the quantities n_i^r/n can be approximated by $1/K$. In this case p_1 simplifies to $p_1 \simeq (1 - K\,(1 - 1/K)^{n_B})^{10B}$. Taking a block size $n_B = \gamma K$ for a parameter $\gamma = 5, 6, 7, 8, 9, 10$ and taking various number of blocks B and number of clusters K, we give in Figure 3, the values of the approximation of p_1. We see from these calculations that p_1 can indeed be close to 1. For instance, if $K = 3$, $n_B = 10K = 30$ and $B = 100$, the value of p_1 is greater than 0.98.

Furthermore, as seen from the proof of Theorem 2, the quantity p_1 is actually a lower bound on the probability that all the blocks generated during the last 10 iterations contain representatives of the K clusters. In practice, what counts most is that the K clusters are represented in the blocks that contain outliers. This would indeed ensure a high risk in these blocks. The quantity p_1 thus consists of a simplification of this latter case, that we can manage in our theoretical reasoning. When using the bound of Theorem 2 however, on the probabilistic breakdown point as a guide to set the hyper-parameters of K-bMOM, we prefer to simplify the value of p_1 to 1, as we think that

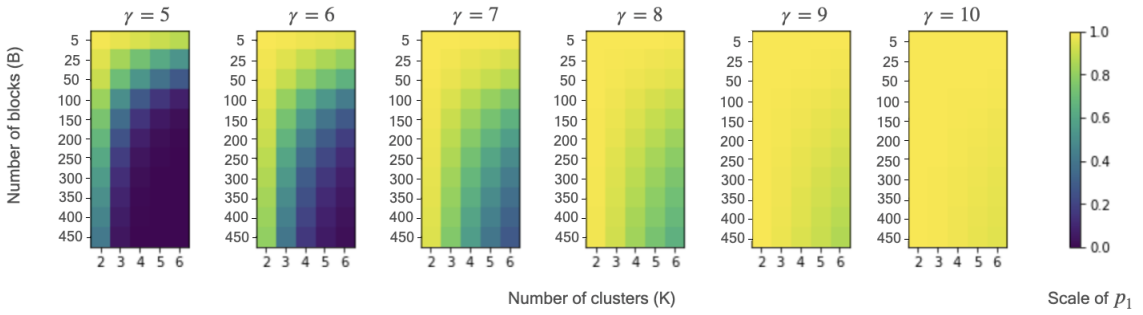


Figure 3: Evolution of the quantity p_1 defined in Theorem 2 with respect to the number of clusters K, the block size $n_B = \gamma K$ and the number of blocks B.

it is too pessimistic for practice.

When choosing the hyper-parameters of K-bMOM, we thus a priori fix the proportion of outliers m/n and take a block length n_B such that $(1 - m/n)^{n_B} = 0.55 > 1/2$. We then fix a confidence parameter H (typically equal to 0.05) such that $H = 1 - p_2 = 10 \exp(-2B\left((1 - m/n)^{n_B} - 1/2\right)^2)$. By denoting $D = (1 - m/n)^{n_B} - 1/2 (= 0.55 - 0.5 = 0.05)$, we get $B = \lceil \log(10/R)/(2D^2) \rceil$. See Section ?? for more details, that discuss the use of the bound $1 - p_2$ for practice.

3.2. Convergence rates for an idealized robust estimator

In this section, we give probabilistic performance bounds for an idealized version of the estimator produced by our algorithm presented in Section 2 above.

We first need to describe our setting. We study the *robustness against adversarial contamination*. Since we are in a probabilistic framework, we denote the sample $(X_1, ..., X_n)$, rather than $(x_1, ..., x_n)$ in the previous sections. We assume that the dataset is made of two disjoint components, indexed by $\mathcal{I}$ and $\mathcal{O}$ with $\mathcal{I} \bigcup \mathcal{O} = \{1, ..., n\}$ and $\mathcal{I} \bigcap \mathcal{O} = \emptyset$: the set of regular data $(X_i)_{i \in \mathcal{I}}$, corresponding to data that provide information and are not corrupted, and the set of outliers $(X_j)_{j \in \mathcal{O}}$, that may be completely misleading for the clustering task. The random variables X_i, $i = 1, ..., n$, take values in a separable Hilbert space $(\mathcal{X}, \|\cdot\|)$ and the regular data $(X_i)_{i \in \mathcal{I}}$ are independent and identically distributed random variables. No assumption is made on the behavior of the outliers $(X_j)_{j \in \mathcal{O}}$, that may have infinite moments.

We also set a generic random variable X, independent of the sample and of the same distribution P as X_i, for any index $i \in \mathcal{I}$.

For any codebook $\mathbf{c} = \{c_1, ..., c_K\}$, we denote by $\ell_{\mathbf{c}}$ a loss function on $\mathcal{X}$ such that $\ell_{\mathbf{c}}(x) = \min_{j=1,...,K} \left\{-2\langle x, c_j\rangle + \|c_j\|^2\right\}$, where $\langle\cdot,\cdot\rangle$ is the scalar product associated to the Hilbertian norm $\|\cdot\|$ on $\mathcal{X}$. Notice that $\|x - c_j\|^2 = \|x\|^2 - 2\langle x, c_j\rangle + \|c_j\|^2$. The loss $\ell_{\mathbf{c}}$ is classically associated with the K-means procedure (see for instance [29]).

For any function f, denote $Pf := \mathbb{E}[f(X)]$. For the K-means problem to make sense, we assume that $P\|X\|^2 < +\infty$. Our goal is to find from the sample $(X_1, ..., X_n)$ a collection of centroids that is close to the following set of optimal codebooks,

$$\begin{aligned} C_* &= \arg\min_{\mathbf{c}\in\mathcal{X}^K} \{P\ell_{\mathbf{c}}\} \\ &= \arg\min_{\mathbf{c}=\{c_1,...,c_K\}\in\mathcal{X}^K} \left\{\mathbb{E}\left[\min_{j=1,...,K} \|X - c_j\|^2\right]\right\}. \end{aligned}$$

Also denote $\ell_* = \ell_{\mathbf{c}_*}$ for any $\mathbf{c}_* \in C_*$, the optimal distorsion risk. Note that an optimal codebook always exists but may not be unique, $C_* \neq \emptyset$ (see for instance [30]).

Furthermore, we assume that the magnitude of an optimal codebook is known. This means that there exists a constant $M_* > 0$ such that there exists $\mathbf{c}_* = (c_{*,1}, ..., c_{*,k}) \in C_*$ with $\max_{i=1,...,k} \|c_{*,i}\| \leq M_*$ and we may restrict our search within codebooks $\mathbf{c}$ satisfying $\max_{c\in\mathbf{c}} \|c\| \leq M_*$. This assumption is rather natural (in practice in general we do not want to have centroids too far away) and is also considered in [15, Section 2].

Hence, we set

$$\hat{C} = \arg\min_{\mathbf{c}\in\mathcal{X}_{M_*}^K} \{\mathrm{MOM}(\ell_{\mathbf{c}})\}, \tag{5}$$

the set of codebooks minimizing the median-of-means of the loss along the data, where $\mathcal{X}_{M_*} = \{x \in \mathcal{X}; \|x\| \leq M_*\}$ is the ball of radius M_* in $\mathcal{X}$ and we recall that

$$\mathrm{MOM}(\ell_{\mathbf{c}}) = \mathrm{med}\left\{\frac{1}{\#b_i}\sum_{j\in b_i} \ell_{\mathbf{c}}(X_j) : i \in \{1, ..., B\}\right\}.$$

We consider that our algorithm, presented in Section 2 above, is an approximation to the minimization task defined in (5). Indeed, our algorithm iteratively computes codebooks in a Lloyd-type fashion in each block of data and then at each step, chooses to keep the codebook that achieves the median of the K-means distortion in each block.

Note also that in (5) we consider the "classical" MOM, instead of the bootstrap MOM. Considering a bMOM however, with the same block length and number of blocks as a MOM should

give rather similar performances. The point in using the MOM statistics is that its mathematical analysis is simpler than for bMOM, since the blocks of MOM are disjoint and so, independent. Consequently, empirical process techniques are available (see the proof in Section 8).

The estimators given by (5) have been recently studied in [15, Section 2], where they are proved to achieve sub-Gaussian performance bounds just using a two finite moments assumption for the random variable X. In our result below, we take a different route by studying robustness against adversarial contamination, under the hypothesis that regular data are uniformly bounded. In the framework of supervised learning, Lecué et al. [31] also studied estimators of the form of (5) - but with different losses -, both in the cases of data with finite second moment and data contamination.

Let O denote the set of indexes of blocks that contain at least one outlier and I denote the set of indexes of blocks that are not corrupted, i.e. that do not contain any outlier. We thus have $|O| \leq n_o$, where n_o is the number of outliers , and $|I| \geq B - n_o$.

Denote also $R(\mathbf{c}) = P\ell_{\mathbf{c}}$, the risk of a codebook $\mathbf{c}$ and $R_* = P\ell_*$ the best possible risk, that is the minimum value of the risk over all possible codebooks $\mathbf{c} \in \mathcal{X}_{M_*}^K$. For any estimator $\hat{\mathbf{c}}_n \in \hat{C}$, we give probabilistic bounds on the quantity $R(\hat{\mathbf{c}}_n) - R_*$, also known as the excess K-means distortion risk.

Theorem 3 *If there exists $M_I > 0$ such that $\|X\| \leq M_I$ a.s. and if the number of outliers n_o satisfies $n_o \leq B/4$, then two numerical constants $l_1, l_2 > 0$ exist such that it holds, with probability greater than $1 - 2\exp(-l_1 B)$,*

$$R(\hat{\mathbf{c}}_n) - R_* \leq l_2 \max\left\{ M\sqrt{\frac{B\mathbb{E}\left[\|X\|^2\right]}{n}}, \frac{K\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]} + M^2/2\right]}{\sqrt{n}} \right\}, \tag{6}$$

where $M = \max\{M_*, M_I\}$. It can be seen from the proof that $l_1 = 3/64$ and $l_2 = 512$ work.

The proof of Theorem 3 can be found in Section 8.

Note that in Theorem 3 we assume that the regular data are defined in a bounded domain of the Hilbert space $\mathcal{X}$ and robustness is considered through the fact that there may be outliers in the dataset. If the number of outliers is small enough compared to the number of blocks ($n_o \leq B/4$), the upper bound given in (6) for the excess K-means distortion risk is composed of two terms. The second term in the maximum appearing at the right-hand side of (6) corresponds to the classical convergence rate of the K-means for a sample that is bounded in a separable Hilbert space that

do not contain any outlier, see [29]. Note that it has been proved that in the no-contamination, bounded setting, the right dependence in K in the convergence rate is of the order $\sqrt{K}$ - up to logarithmic factors - rather than K, see [32] and [33]. This brings many technicalities however, and we leave the optimal dependence in K in our setting as an interesting open question.

The first term in the maximum appearing on the right-hand side of (6) reflects the price to pay for the presence of outliers. In particular, it does not change the rate of convergence of the no-contamination setting if B is on the order of K^2. In such a case, estimators given by (5) are robust to adversarial contamination in the sense that their estimation rates are not impacted by the presence of outliers.

4. Scope of K-bMOM and practical considerations

4.1. Block length, number of clusters and proportion of outliers

The purpose of this section is to clarify the scope of the K-bMOM algorithm according to the proportion of outliers, the number of clusters and the block size.

The reader is reminded that Proposition 3 holds if the probability that a block is not contaminated by outliers ie $(1 - m/n)^{n_B}$ is strictly superior to $1/2$. The block length n_B is therefore dependent of the proportion of outliers m/n and it can be determined for a given probability that a block is not contaminated by outliers. Taking a block size $n_B = \gamma K$ with $\gamma \in \mathbb{N}^*$ (it can be seen roughly as the number of data points per cluster), the maximum proportion of outliers for which our proposed approach is robust can be evaluated for a given probability that a block is not contaminated by outliers.

Firstly and for illustration purposes, let's take this probability equal to 0.55, such that $(1 - m/n)^{n_B} = 0.55 > 1/2$. According to the previous condition, we get:

$$\frac{m}{n} = 1 - (0.55)^{\frac{1}{\gamma K}}.$$

Figure 4 stands for the maximum level of contamination according to the number of clusters for different values of $\gamma \in [1, 10]$. As can be observed, the case $\gamma = 1$ enables us to deal with a contamination level up to 10% for a number of clusters $K < 7$. However, if by chance each cluster is represented in a block, the estimation of centroids is based on one data point only which can lead to inaccurate estimations. On the other hand, by taking $\gamma = 5$ (roughly 5 data points per

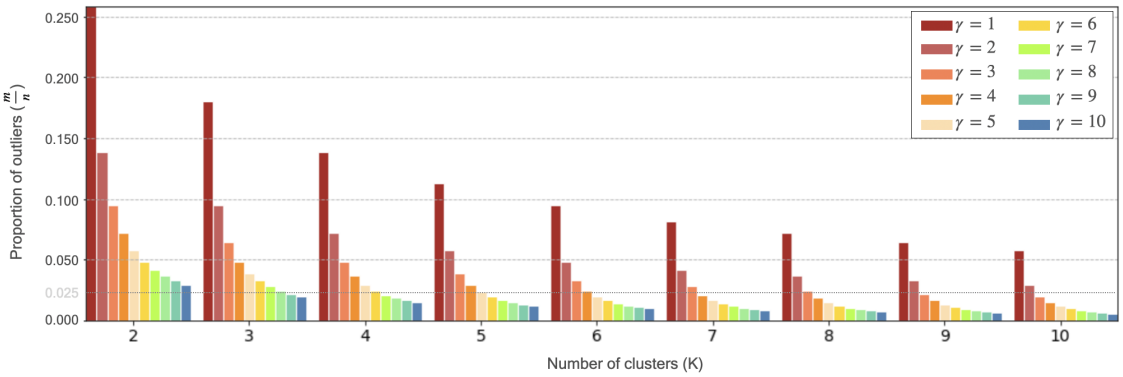


Figure 4: Evolution of the upper level of outliers for robustness according to the number of clusters and the coefficient γ.

cluster), the level of outliers for which the K-bMOM algorithm is robust, has to be low (under 5% if $K > 2$) but we can expect having an accurate estimation of centroids. Therefore, there is a trade-off in practice between the number of clusters, the level of outliers and the accuracy of the centroid estimation.

The second illustration proposed in Table 1 evaluates the sample size n_B, according to a range of probabilities that a data block is healthy, and a range of proportion of outliers. It can be noted that the block size n_B dramatically decreases when the percentage of outliers increases. A proportion of outliers up to 0.04 leads to a majority of block sizes containing less than 10 data points. When the number of clusters remains small ($K \in 2, 3$) with a uniform presence in the block ($\gamma \in 3, 4, 5$), the K-bMOM algorithm should behave correctly insuring that the probability that a data block is healthy is greater than $1/2$ for a proportion of outliers up to 0.4 (see bold values located in the upper left diagonal in Table 1). However, if the number of groups becomes quite high, e.g. $K = 10$, the scope of K-bMOM narrows to a proportion of outliers of 0.01 and below for $\gamma = 5$ as illustrated by the blue bold values.

In conclusion, when the number of clusters is small ($K \leq 5$) the K-bMOM algorithm should be robust with respect to a proportion of outliers up to $m/n = 0.03$ with a limited block size ($n_B \simeq 25$ and $\gamma \geq 5$). For a higher number of groups, the K-bMOM algorithm should remain accurate but for smaller percentage of outliers (below 1%). In practice, this situation should not be too restrictive. Indeed, since an outlier is a data point that differs considerably from all or most other data in a dataset, we do not expect having a big proportion of them in a dataset (in contrast to noisy data).

		Probability that a data block is healthy									
		0.51	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
	0.001	**673**	**597**	**510**	**430**	**356**	**287**	**223**	**162**	**105**	**51**
	0.005	**134**	**119**	**101**	**85**	**71**	**57**	**44**	**32**	**21**	10
Proportion	0.01	**66**	**59**	**50**	**42**	**35**	**28**	**22**	**16**	10	5
of outliers	0.02	**33**	**29**	**25**	**21**	**17**	14	11	8	5	2
m/n	0.03	**22**	**19**	**16**	14	11	9	7	5	3	1
	0.04	**16**	14	12	10	8	7	5	3	2	1
	0.05	13	11	9	8	6	5	4	3	2	1
	0.1	6	5	4	4	3	2	2	1	1	0

Table 1: Lookup table of the block size n_B evaluated according to a range of proportion of outliers m/n and a range of probabilities that a data block is healthy. In bold, the ranges of block sizes for K=3 with $\gamma = 5$. In bold blue, the possible ranges of n_B for $K = 10$ with $\gamma = 5$.

Section 5.3 evaluates the performance of the K-bMOM algorithm in different simulations contexts.

4.2. Influence of the number of blocks

In the previous section, we showed the strong influence of the block size on the proportion of maximum outliers to guarantee a percentage of healthy blocks. In particular, the smaller the size, the more robust the algorithm is to a large proportion of outliers. In this section, we focus on the influence of the second hyper-parameter of the K-bMOM algorithm which is the number of blocks. To do so, we consider a 2-dimensional Gaussian mixture model of $K = 3$ components with equal size $n_1 = n_2 = n_3 = 300$. The mean vectors are set to $\mu_1 = [3, 12], \mu_2 = [6, 3]$ and $\mu_3 = [-6, 9]$ and the variance parameter is set to $\sigma^2 = 0.6$. Twenty outliers are randomly selected from the data and their coordinates are multiplied by 10. The block size is set to $5 * K = 15$ to be robust to any level of outliers (see Table 1) with a sufficient number of elements per group. The number of blocks varies between 1 block until 1000 blocks and the process is iterated 100 times. The violinplots of accuracies, distortions and number of clusters computed on the regular data are illustrated in Figure 5. The median is shown by a bold black dash. As it can be observed, the more the number of blocks increases, the more robust the algorithm is. Moreover, the performance of the algorithm is equivalent from a number of blocks $B \geq 50$. These results confirm the choice of a small size of a

block (see Section 4.1 and Section 3.1.1) and a high number of blocks ($B \geq 50$) to have a procedure that is likely to be robust.

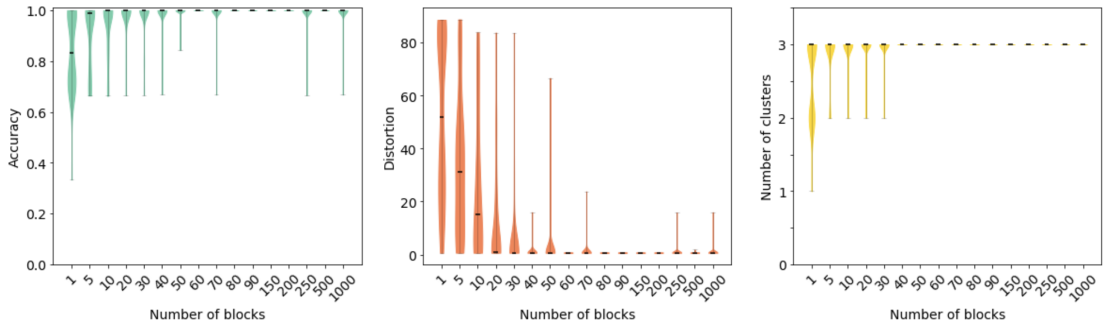


Figure 5: Violinplots of accuracy (left), distortion (middle), number of clusters (right) computed on the partition of regular data obtained by the K-bMOM algorithm according to the number of blocks.

5. Experimental simulations

This Section aims at evaluating the scope of performances of the K-bMOM as an initialization strategy and as a clustering algorithm according to a taxonomy of different types of outliers on one hand and according to several specifications such as sample size, dimension and number of groups on the other hand.

5.1. Experimental contexts and practical considerations

The same experimental context will be addressed to evaluate the proposed robust initialization and the K-bMOM algorithm. In particular, the different situations considered will depend on these different aspects:

1. the outliers typology. Three types of outliers are considered: isolated multi-directional outliers, isolated oriented outliers and a cluster of outliers.
2. the level m/n of outliers contamination,
3. the sample size n of data,

4. the dimension p of data. Let us note that our strategy is not designed to deal with high-dimensional data depending on the form of the optimisation function. Therefore, the number of dimensions tested will remain small (to avoid the curse of dimensionality introduced by Bellman) and discriminant since we do not deal with the problem of noisy dimensions.

5. the separability of clusters by varying the level of the scaling parameter σ^2 of the identity covariance matrix.

6. the number K of clusters.

7. the distribution of data: data points are generated according to a K-dimensional Gaussian Mixture model with equal size and isotropic variance, following the design of the K-means based approach.

Regular data and outliers generation procedures

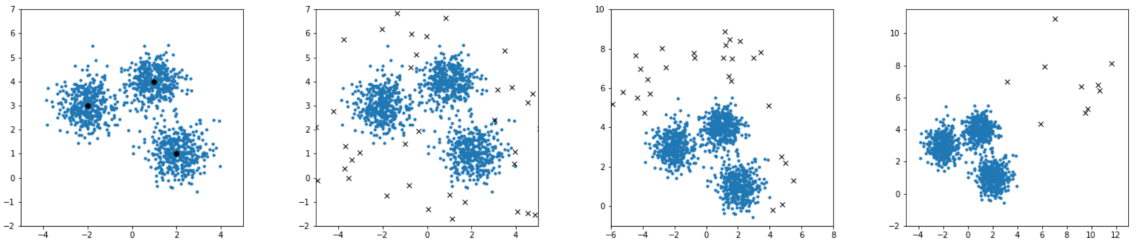
n- data are generated from K multivariate Gaussian distributions of dimension p with equal size n/K, isotropic variance $\Sigma = \sigma^2 \mathbf{I}_p$ and average vectors μ_k with $k \in \{1, \ldots, K\}$. Figure 6.a illustrates one realisation of the simulated context in the case $K = 3$ groups.

Three typologies of outliers are considered and are generated as expressed below:

- isolated outliers: they are generated uniformly in a parallelogram defined by the coordinate-wise ranges of the regular data points. Data points having squared Mahalanobis distances from the centers greater than $\chi^2_{0.975}$ are retained and only m of them are going to replace the same number of randomly selected regular data. This case is illustrated in Figure 6.b.

- isolated oriented outliers: from these n regular data points, we randomly select m of regular data points as potential outliers and their coordinates are multiplied by a constant term β which quantifies how far these outliers are from their own distribution. Figure 6.c illustrates such a type of outliers.

- Cluster of outliers: m regular data points are randomly replaced by a cluster of outliers of size m generated according to a 2-dimensional Gaussian distribution with average $\mu_{outlier} = \beta[1, 1]$ and $\sigma^2_{outlier}$ as a scaling parameter of the covariance matrix. This situation is depicted in Figure 6.d.



(a) Regular data (b) Punctual outliers (T1) (c) Punctual oriented outliers (T2) (d) Cluster of outliers (T3)

Figure 6: Illustrations of simulated regular data generated according to a Gaussian Mixture Model with isotropic variance (blue points) and different types of outliers (black crosses).

Experimental values

The experimental values taken for each of these scenarios are detailed below:

	detailed case	**values**
degree of outliers		$\beta \in \{9, 27\}$
outlier typology	T1: isolated	$m/n \in \{0, 0.001, 0.005, 0.01, ..., 0.04\}$
	T2: isolated oriented	$\beta \in \{9, 27\}$, $m/n \in \{0, 0.001, 0.005, 0.01, ..., 0.04\}$
	T3: clustered	$\mu_{outlier} = \beta\,[1, 1]$, $\sigma^2_{outlier} = 2$
		$m/n \in \{0, 0.001, 0.005, 0.01, ..., 0.04\}$
dimension		$p \in \{2, 5\}$
sample size		$n \in \{120, 1200, 12000\}$
separability of clusters	(high, medium, low)	$\sigma^2 \in \{0.4, 0.6, 0.8\}$
number of clusters		$K \in \{3, 5, 10\}$
average vector	$(K = 3, p = 2)$	$\mu_1 = [1, 4], \mu_2 = [2, 1], \mu_3 = [-2, 3]$
	$(K = 5, p = 2)$	$\mu_4 = [0, -1], \mu_5 = [1, -3]$
	$(K = 10, p = 2)$	$\mu_6 = [0, 7], \mu_7 = [3, 6], \mu_8 = [5, 1]$
		$\mu_9 = [7, 0], \mu_{10} = [8, 4]$
	$(K = 3, p = 5)$	$\mu_1 = [1, 4, b, a, a], \mu_2 = [2, 1, a, b, a]$
		$\mu_3 = [-2, 3, a, a, b]$
	$(K = 5, p = 5)$	$\mu_4 = [0, -1, b, b, b], \mu_5 = [1, -3, a, a, a]$
	$(K = 10, p = 5)$	$\mu_6 = [0, 7, a, b, b], \mu_7 = [3, 6, b, a, a],$
		$\mu_8 = [5, 1, a, b, a], \mu_9 = [7, 0, b, a, a],$
		$\mu_{10} = [8, 4, a, b, b]$

with $a = 0$ and $b = -1$. Let us note that for all the experiments, the number of clusters (and the level of outliers) is supposed to be known and fixed to its true value K (resp. m/n).

Performance criteria

In order to compare the different starting strategies in terms of performance, we compute three criteria:

- the Root Mean Square Error (RMSE) in order to evaluate the robustness of fitted centers once the initialization step is performed. This criterion is calculated between the centers proposed

by the initialization process and those used to simulate the data, given by:

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{K} \|\hat{c}_k - \mu_k\|^2}{K}},$$

where $\hat{c}_k$ stands for the initial center the most probable for the class k and μ_k the average parameter of the kth component.

- the accuracy computed between the partition obtained by the nearest initial centers and computed on the regular data.

- the empirical distortion obtained at the end of the initialization step and computed over the $(1 - m/n)n$ regular data:

$$\hat{R}(\hat{\mathbf{c}}) = \frac{1}{(1 - m/n)n} \sum_{k=1}^{K} \sum_{x_i \in \mathcal{C}_k} \|x_i - \hat{c}_k\|^2 .$$

For each simulation context, the experiment is repeated 1000 times. These criteria are averaged and standard deviations have been computed for each initialization or clustering approach.

5.2. Comparing initialization strategies for the clustering task

The experimental context of this subsection aims at evaluating the scope of performance of the initialization process based on the bMOM principle. In this section we propose to apply the MOM principle to the most widely used initialization methods among which K-means++ and K-medians++ as expressed in Algorithm 2 in Section 2.2. We consider the following 3 traditional initialization strategies: Random initialization, K-means++ [23], K-medians++ and also a robust initialization strategy developed by [34] named ROBIN. The implementations that we used in this study for the above approaches come from SCIKIT-LEARN library which is a free software machine learning library for the Python programming language and is publicly available in [35].

Global comments and results

First of all, it is important to notice that the behavior of 6 initialization procedures is really stable and comparable in terms of accuracy when the data are not polluted by any kind of outliers. Indeed as illustrated in Table 2, the average and the standard deviation of accuracy are equivalent between approaches (except for the uniform one). As expected, the accuracy decreases when the cluster

separability becomes weaker (ie when the scaling parameter σ^2 increases). Besides, an interesting point is the level of accuracy obtained by the different approaches to estimate the centers of each cluster. It appears that, on regular data, this is the K-bMOM-km++ and K-bMOM-kmed which fits better the centers of clusters than the other methods as it can be observed in Table 3.

	average accuracy (and standard deviation) of initialization processes:					
σ^2	uniform	K-medians++	K-means++	**ROBIN**	K-bMOM-km	K-bMOM-kmed
0.4	0.588 (0.028)	0.995 (0.009)	0.995 (0.009)	0.979 (0.028)	0.997 (0.015)	0.970 (0.027)
0.6	0.531 (0.033)	0.883 (0.033)	0.895 (0.035)	0.898 (0.039)	0.902 (0.042)	0.886 (0.048)
0.8	0.476 (0.047)	0.712 (0.047)	0.711 (0.050)	0.735 (0.053)	0.716 (0.053)	0.702 (0.056)

Table 2: Accuracy averaged performance and the standard deviations (in brackets) of initialization procedures on regular data according to the separability of clusters.

	average RMSE (and standard deviation) of initialization processes:					
σ^2	**uniform**	K-medians++	K-means++	ROBIN	K-bMOM-km	K-bMOM-kmed
0.4	1.333 (0.205)	0.612 (0.164)	0.614 (0.179)	0.822 (0.175)	**0.398 (0.093)**	**0.422 (0.147)**
0.6	1.528 (0.242)	0.955 (0.265)	0.943 (0.254)	1.160 (0.259)	**0.737 (0.185)**	**0.739 (0.186)**
0.8	1.860 (0.275)	1.267 (0.351)	1.258 (0.350)	1.521 (0.362)	**1.096 (0.242)**	**1.105 (0.251)**

Table 3: RMSE averaged performance and the standard deviations (in brackets) of initialization procedures on regular data according to the separability of clusters.

Tables 4, 5 and 6 highlight aggregated performances on the different situations (dimension, separability, number of clusters, etc) of 6 initialization strategies according to the typology of outliers considered. Median and standard deviations (in brackets) of three metrics are represented. They are computed on all simulated contexts with outliers. First of all, it can be noted in Table 4 that in the case of isolated outliers (T1), all methods (except for the uniform case) perform quite well on average: their average accuracy remains above 0.83. However, our proposed strategies K-bMOM-km++ and K-bMOM-kmed seem to be more robust than the other approaches even though their standard deviations remain high. This last point will be explained later but is mainly due to the different levels of simulation complexity we address.

On the other hand, the traditional approaches are being impacted by oriented isolated outliers

(T2) and clustered outliers (T3). Indeed, in the case T2 in Table 5, the average accuracy of uniform initialization, K-means++, K-medians++ and ROBIN, performs 10% (as K-medians++) to 40% (K-means++) worse than the MOM-based approaches. Such an impact is also visible on the average distortion computed on the regular data and on the RMSE. The K-bMOM-km++ ones remain smaller compared to the rest of the approaches. In the case T3 illustrated in Table 6, the performance difference between the proposed initialization approaches and the traditional ones is narrower than in case T2 – especially for K-means++ – but still exists.

type of outlier		initialization	RMSE	distortion	accuracy
T1	isolated	uniform	1.643 (0.370)	4.307 (1.700)	0.538 (0.069)
		K-medians++	0.934 (0.389)	1.887 (1.656)	0.833 (0.137)
		K-means++	0.979 (0.405)	1.752 (1.668)	0.857 (0.137)
		ROBIN	1.351 (1.122)	2.674 (2.273)	0.847 (0.196)
		K-bMOM-km++	**0.702 (0.534)**	**1.421 (1.363)**	**0.894 (0.136)**
		K-bMOM-kmed	**0.727 (0.412)**	**1.491 (1.355)**	**0.871 (0.134)**

Table 4: Aggregated performances on the case of isolated outliers (T1).

type of outlier		initialization	RMSE	distortion	accuracy
T2	oriented & isolated	uniform	**4.155 (5.653)**	4.652 (1.699)	0.708 (0.046)
		K-medians++	39.53 (39.09)	7.936 (5.574)	0.412 (0.189)
		K-means++	23.38 (33.49)	3.458 (2.339)	0.770 (0.146)
		ROBIN	15.95 (50.41)	7.646 (89.79)	0.635 (0.346)
		K-bMOM-km++	6.552 (9.142)	**1.828 (1.491)**	**0.874 (0.085)**
		K-bMOM-kmed	7.420 (8.819)	**1.972 (1.505)**	**0.849 (0.081)**

Table 5: Aggregated performances on the case of isolated oriented outliers (T2).

In order to have a general view of the sensitivity of initialization procedures according to experimental dimensions, an analysis of variance explaining the average accuracy has been done for each procedure. Table 7 summarizes the effect of parameters on any type of outliers. A star indicated that the p-value was under the threshold 0.05 on at least one type of outlier and the modality of the parameter is filled when its negative effect has an impact on the accuracy. It appears that

type of outlier		initialization	RMSE	distortion	accuracy
		uniform	1.505 (0.360)	4.157 (1.597)	0.544 (0.066)
		K-medians++	0.842 (0.358)	1.872 (1.667)	0.810 (0.152)
T3	cluster of	K-means++	0.880 (0.360)	2.472 (1.755)	0.756 (0.158)
	outliers	ROBIN	1.256 (0.817)	3.847 (4.067)	0.694 (0.330)
		K-bMOM-km++	**0.637 (0.429)**	**1.630 (1.523)**	**0.851 (0.153)**
		K-bMOM-kmed	0.697 (0.421)	1.718 (1.522)	0.800 (0.152)

Table 6: Aggregated performances on the case of clustered outliers (T3).

all methods are impacted by the proportion of outliers m/n and the class separability σ^2. While ROBIN remains insensitive to the number of components K, this is the only approach which is very sensitive to the sample size. Moreover, the outlier degree β has a negative impact on the traditional initialization approaches such as K-means++ and K-medians++.

initialization	n	σ^2	m/n	β	K	p
uniform		*	*		*	
K-medians++		*	*	*	*	
K-means++		*	*	*	*	
ROBIN	$*(n = 12000)$	*	*			
K-bMOM-km++		*	$(m/n > 0.03)$		*	
K-bMOM-kmed		*	$(m/n > 0.03)$		*	

Table 7: Summary of the influence of parameters on the accuracy of each initialization method. Note that when a p-value is less than 0.05, a star is indicated and also involved modalities.

More particularly, the quantitative effects of cluster separability is detailed in Table 8 where the average accuracies of initialization methods according to the modalities of σ^2 are indicated. As expected, the less the groups are separated, the more difficult it is to find a correct data partition with or without the presence of outliers for any kind of outliers. The accuracy decreases from 15% to 30% between the high separability case ($\sigma = 0.4$) and the lowest one ($\sigma = 0.4$).

Similarly, since the initialization methods behaves equally in average when the cluster separability decreases, we are going to focus on the case $\sigma^2 = 0.4$ in order to highlight the benefits

	isolated outliers (T1)			oriented outliers (T2)			cluster of outliers (T3)		
	0.4	0.6	0.8	0.4	0.6	0.8	0.4	0.6	0.8
uniform	0.591	0.545	0.469	0.586	0.539	0.477	0.587	0.540	0.472
K-medians++	0.969	0.853	0.671	0.454	0.450	0.383	0.912	0.811	0.660
K-means++	0.987	0.870	0.680	0.825	0.731	0.595	0.817	0.747	0.617
ROBIN	0.905	0.829	0.661	0.639	0.799	0.664	0.666	0.729	0.606
K-bMOM-km++	0.989	0.887	0.711	0.953	0.867	0.697	0.958	0.840	0.673
K-bMOM-kmed	0.965	0.872	0.706	0.909	0.840	0.672	0.920	0.816	0.652

Table 8: Average accuracies for each initialization approaches according to the type of outliers and the level of separability.

and limitations of the proposed initialization approaches, depending on the number of clusters, the percentage and the degree of outliers. Moreover, it has to be noted that the case of isolated outliers (T1) impacts none of the initialization approaches. Hence, the specific comments will focus on types T2 and T3 of outliers.

Specific comments: sensibility to the type of outliers

The three plots shown in Figure 7 stand for the average accuracies of 6 initialization approaches for different proportion of type T3 outliers depending on the number of simulated clusters with K=3 (left side), K=5 (middle) and K=10 (right side) groups respectively. In the easiest configuration (K=3), it is worth noting that K-bMOM-km++ and K-bMOM-kmed++ are on average, the most robust approaches to noise. This is mainly due to their independence to the distance of outliers from regular data and also to the sample size. K-means resists quite well until a proportion of noise equals 0.01 before dropping out. ROBIN has the same kind of behavior and its decrease is mainly linked to the level of outliers. When the percentage of noise remains low (less than 0.01), ROBIN remains really competitive with initialization K-bMOM-km++ and exceeds it when K=10.

In the case of isolated outliers (T2), K-bMOM-km++ has the same kind of behavior that previously as illustrated in Figure 8: it resists very well to the increasing level of outliers when the number of clusters is relatively small (K= 3 or K=5). However, when the number of groups is high (K=10), even though it is remains relatively stable in regards to the increasing level of outliers, it

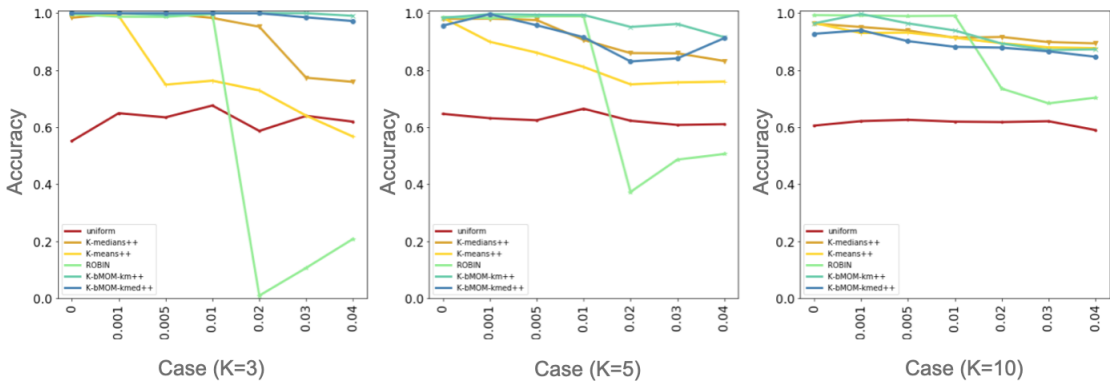


Figure 7: Case: cluster of outliers (T3). Comparison of the average accuracies per initialization approach depending on the proportion of outliers and the number of clusters.

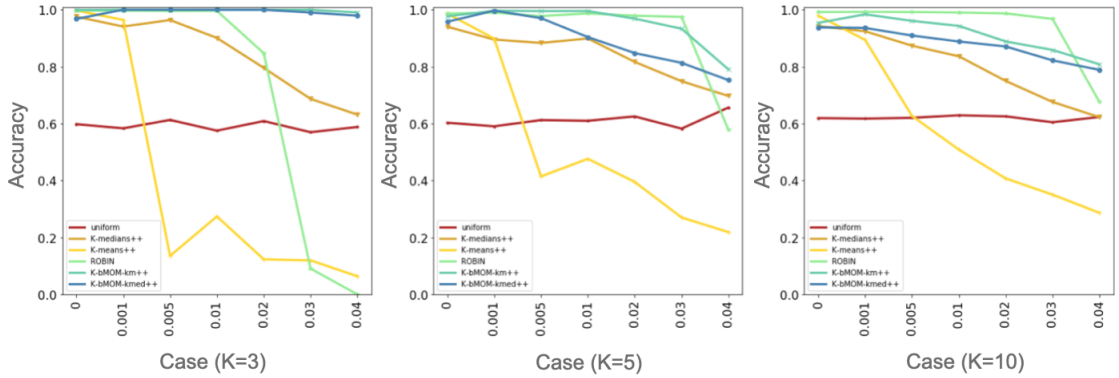


Figure 8: Case: isolated oriented outliers (T2). Comparison of the average accuracies per initialization approach depending on the proportion of outliers, the number of clusters

appears that ROBIN has better performances until 3% of polluted data. Regarding to this kind of outliers, the accuracy of K-means++ and K-medians++ drop out even with a low percentage of outliers and keep going down at each additional percentage of outliers.

5.3. Evaluation of the behaviour of K-bMOM algorithm

The aim of this subsection is to evaluate how K-bMOM behaves according to several considerations such as the type of outliers, the proportion of contamination, the sample size, the dimension of data and the number of groups.

type of outlier		m/n	RMSE	distortion	accuracy
	regular data	0	0.203 (0.084)	0.921 (0.555)	**0.993 (0.010)**
		0.001	0.076 (0.097)	0.921 (0.576)	**0.988 (0.025)**
		0.005	0.150 (0.117)	0.911 (0.554)	**0.992 (0.015)**
		0.01	0.149 (0.099)	0.904 (0.548)	**0.992 (0.015)**
T1	isolated	0.02	0.175 (0.090)	0.912 (0.552)	**0.993 (0.011)**
		0.03	0.200 (0.095)	0.922 (0.557)	**0.991 (0.016)**
		0.04	0.215 (0.099)	0.930 (0.560)	**0.990 (0.020)**

Table 9: Aggregated performance of K-bMOM depending on the percentage of outliers for type T1 (isolated) outliers.

Tables 9, 10 and 11 show the mean and the standard deviation (in parenthesis) of RMSE, distortion and accuracy for each type of outlier (T1, T2, T3) with several levels of outliers. These metrics have been averaged over the different combinations linked to the number of clusters, the dimension of data and the level of cluster separability.

First of all, it can be observed that isolated outliers have almost no influence on the clustering results. On average, the accuracy remains around 0.99, really close to that obtained on regular data irrespective of the percentage of outliers. The same stability is observed on the RMSE and distortion values. This is mainly explained by the distribution of outliers which is well-dispersed around regular data from the isotropic Gaussian mixture model and therefore, it does not affect the clustering task.

Secondly, concerning the T2 and T3 type outliers, it can be noted that the accuracy rate remains over 0.90 for a percentage of noise under 0.2. This behavior is totally explained by the theoretical limitations seen in Section 4.1. Moreover, the K-bMOM algorithm seems to be more robust in the configuration T3 of a cluster of outliers than in the configuration T2 of isolated and oriented outliers. Indeed, for a percentage of outliers less than 0.2, the accuracy rates decrease by 3% at maximum compared to the performances obtained on regular data while distortions increase of 0.05, in the case of T3. In the case T2, the accuracy rate can decrease by 8% and the average distortion rises of 0.2.

By analysing the importance of each factors of simulation settings, in cases verifying the theoretical aspects (ie $m/n \leq 0.2$), it can be observed that the sample size and the dimension of

type of outlier		m/n	RMSE	distortion	accuracy
	regular data	0	0.203 (0.084)	0.921 (0.555)	**0.993 (0.010)**
		0.001	0.053 (0.028)	0.903 (0.542)	**0.994 (0.007)**
		0.005	0.191 (0.352)	1.234 (2.779)	**0.983 (0.063)**
		0.01	0.383 (0.515)	1.552 (2.460)	**0.937 (0.107)**
T2	oriented & isolated	0.02	0.465 (0.575)	1.808 (2.921)	**0.911 (0.131)**
		0.03	0.674 (0.709)	2.309 (3.283)	0.864 (0.146)
		0.04	0.885 (0.713)	2.518 (2.798)	0.814 (0.146)

Table 10: Aggregated performance of K-bMOM depending on the percentage of outliers for type T2 (isolated oriented) outliers.

type of outlier		m/n	RMSE	distortion	accuracy
	regular data	0	0.203 (0.084)	0.921 (0.555)	**0.993 (0.010)**
		0.001	0.067 (0.082)	0.915 (0.565)	**0.990 (0.021)**
		0.005	0.213 (0.218)	0.978 (0.571)	**0.984 (0.036)**
		0.01	0.201 (0.151)	0.959 (0.561)	**0.976 (0.039)**
T3	cluster of outliers	0.02	0.260 (0.177)	1.016 (0.609)	**0.966 (0.055)**
		0.03	0.416 (0.221)	1.247 (0.694)	0.898 (0.091)
		0.04	0.480 (0.247)	1.330 (0.732)	0.882 (0.089)

Table 11: Aggregated performance of K-bMOM depending on the percentage of outliers for type T3 (clustered) outliers.

data (as soon as the dimensions remain discriminant and their number weak) do not influence the performance of the proposed procedure. The main factors which impact the K-bMOM algorithm in the oriented and isolated outliers case (T2) are the percentage of outliers (m/n), the outlier degree (β) and the number of clusters. Such remarks can be observed in Table 12 that shows the p-values of the variance analysis for each parameter. In particular, the outlier degree plays a role when the number of clusters becomes quite high ($K = 10$). This confirms the link between the number of clusters, the size of the block and the level of outliers for which K-bMOM is robust and is highlighted in Figure 4 in Section 4.1. The same kind of behavior is obtained for outliers of type T3 whose performances worsen when the number of clusters is high ($K = 10$). For the type T1, the main parameters which influence the performances are the number of classes (K) and the separability criterion (σ^2).

	influence parameters					
cases	K	σ^2	n	$\frac{m}{n}$	β	p
T1 isolated	*	*	0.62	0.59	-	0.45
T2 isolated oriented	*	0.32	1.00	*	*	0.90
T3 clustered	*	0.12	0.15	*	*	0.99

Table 12: p-values linked to F-statistics of analysis of variances for each influence parameter. Note that a p-value less than 0.05 is denoted by a star * and highlights the influence of this parameter on the performance of the K-bMOM procedure.

The following paragraph focuses on cases $\sigma = 0.4$ to visualize the evolution of RMSE and accuracy of the K-bMOM algorithm according to the number of clusters, the proportion and the type of outliers.

The case of isolated oriented outliers (T2) is displayed in Figures 9, 10 and 11. They show violinplots of accuracies, RMSE and number of unique clusters fitted by the K-bMOM algorithm for a number of clusters equal to K=3, K=5 and K=10 respectively and for several levels of outliers. As can be observed, for a low number of groups ($K \in 3, 5$), the K-bMOM algorithm is robust to a level of outlier up to 3%: the median accuracy on regular data is equal to 1 with centroids well fitted (RMSE remain around 0.2). However, as expected, when the number of clusters is high (K=10), the procedure remains robust for a lower proportion of outliers ($m/n \leq 0.005$).

The case of a cluster of outliers (T3) is displayed in Figures 12, 13 and 14. We can note the

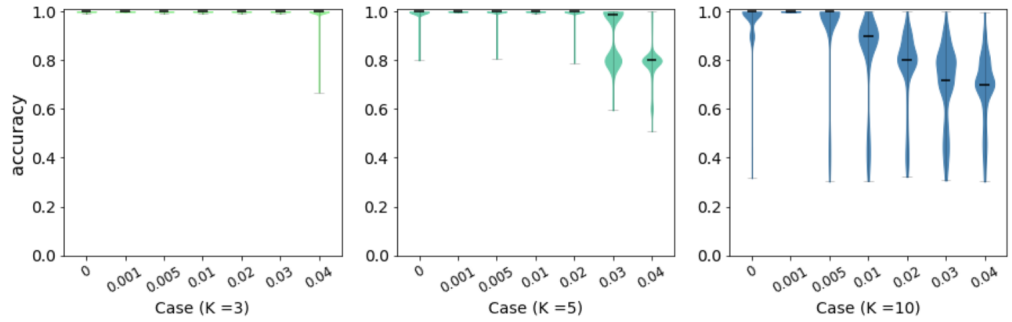


Figure 9: Violinplots of K-bMOM accuracies obtained for different level of T2 outliers, among different sample sizes, outlier degrees and dimensions for K=3 clusters (left), K=5 groups (middle) and K=10 groups (right).

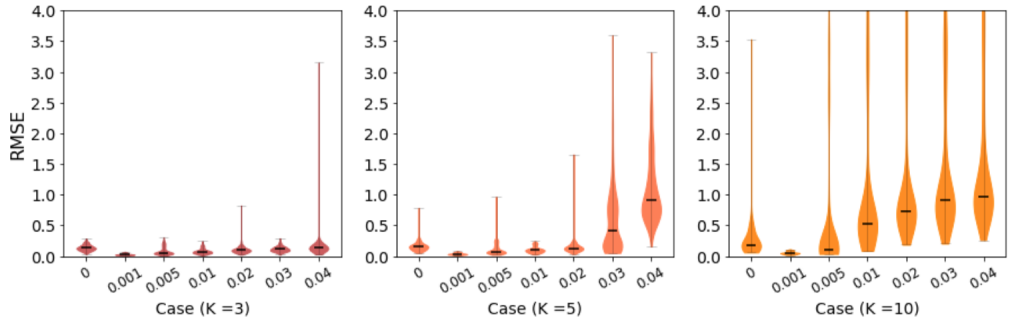


Figure 10: Violinplots of K-bMOM RMSE obtained for different level of T2 outliers, among different sample sizes, outlier degrees and dimensions for K=3 clusters (left), K=5 groups (middle) and K=10 groups (right).

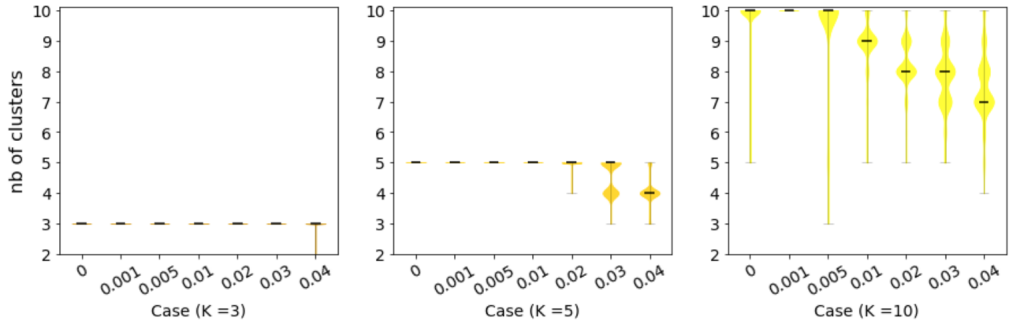


Figure 11: Violinplots of number of clusters obtained for different level of T2 outliers, among different sample sizes, outlier degrees and dimensions for K=3 clusters (left), K=5 groups (middle) and K=10 groups (right).

same kind of behavior as that for the case of isolated outliers: for a low level of outliers, the procedure is robust and this robustness decreases with the number of components of the mixture model. Moreover, the drop-out of metrics occurs at the same time: $m/n = 0.03$ for K=5 and $m/n = 0.01$ when $K = 10$. However, the impact of the proportion of outliers on the level of metrics is weaker when outliers are clustered (T3) than isolated and oriented (T2). Indeed, especially when K=10, the accuracies stay high at about 0.9 whatever the level of pollution and the RMSE remain below 0.5 in case T3.

6. Benchmark K-means type robust clustering algorithm

The objective of this section is to compare the performance of the K-bMOM strategy in its scope of application with the robust clustering algorithms based on K-means approaches on a specific framework with isolated oriented outliers.

Benchmark algorithms

We consider 6 different algorithms: our proposed robust clustering algorithm named K-bMOM, the traditional K-means for comparison and also 4 well-known robust versions of the K-means. These methods are described below:

K-bMOM algorithm introduced in Section 2. The time complexity of the K-bMOM for each iteration is $\mathcal{O}(Kn_BBp)$.

K-medoids aims at finding K data points as centers such as the within inertia is minimized. The Partition Around Medoids algorithm named PAM [36] aims to achieve this in two steps : an assignment step where each data point is assigned to its closest medoid; a refinement step which looks for better medoids than the current ones. The search each time is exhaustive in the data PAM has a complexity dominated by $\mathcal{O}\left(K(n-K)^2p\right)$ per iteration. Faster versions have been proposed in [37].

K-medians is a robust variant of the K-means algorithm [38] : in the aggregation step, instead of computing the barycenter of each group as in the K-means procedure, the K-medians computes

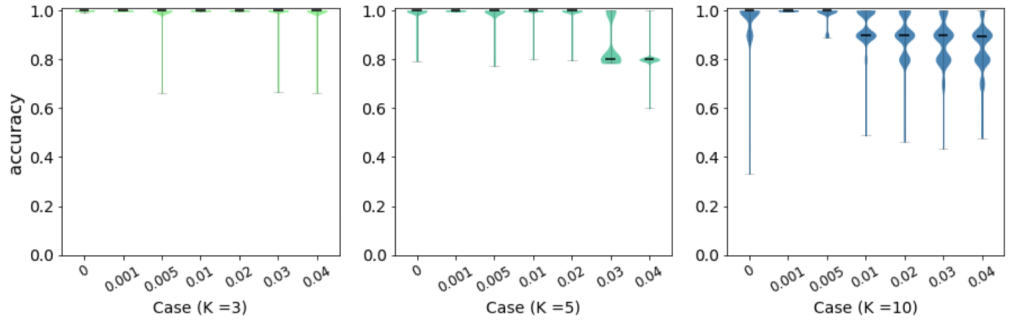


Figure 12: Violinplots of K-bMOM accuracies obtained for different level of T3 outliers, among different sample sizes, outlier degrees and dimensions for K=3 clusters (left), K=5 groups (middle) and K=10 groups (right).

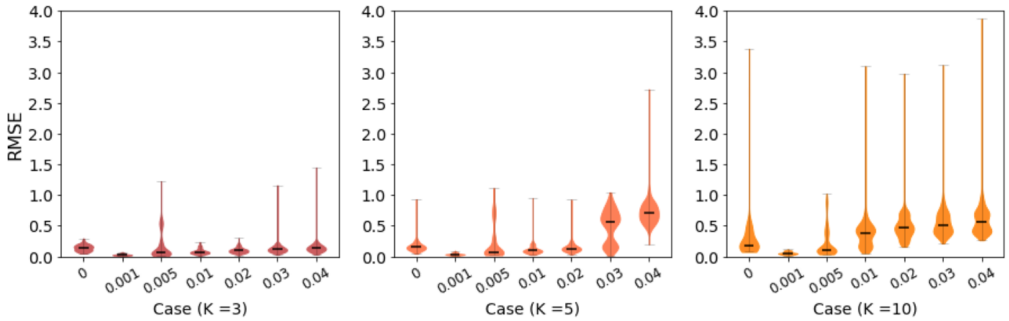


Figure 13: Violinplots of K-bMOM RMSE obtained for different level of T3 outliers, among different sample sizes, outlier degrees and dimensions for K=3 clusters (left), K=5 groups (middle) and K=10 groups (right).

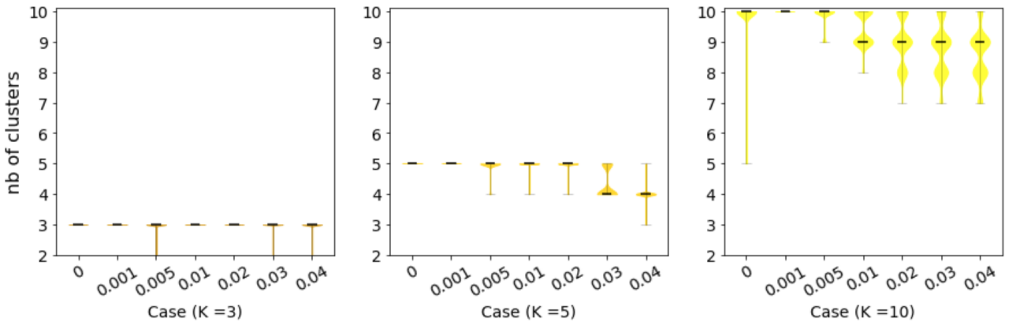


Figure 14: Violinplots of number of clusters obtained for different level of T3 outliers, among different sample sizes, outlier degrees and dimensions for K=3 clusters (left), K=5 groups (middle) and K=10 groups (right).

in each single dimension, the median in the Manhattan-distance formulation. This makes the algorithm more reliable for extreme values. The complexity of such a procedure is dominated by $\mathcal{O}(nKp)$ per iteration as for Loyd's algorithm [39].

trimmed-K-means (trimK-means) implementation is an EM-like algorithm introduced by Cuesta et al. [40] in the late 90s. It is derived from the K-means and benefits robustness properties from the trimming action during the maximisation step where only a proportion $1 - m/n$ of the closest data point from their assigned centroid, is taken into account. Since the trimming needs to sort the data points according to their distance to centroid, it leads therefore to an overall complexity of $\mathcal{O}(Knp + n.\log n)$ at each iteration. Moreover, note that in practice, the user needs to choose a value m/n for the proportion of data points to be discarded and no practical information is given to calibrate such an hyper-parameter. In the simulations, m/n is set to the true value of the number of outliers ie m/n.

K-PDTM is a robust quantization algorithm introduced by Brecheteau et al. [10] that aims to infer the manifold from which the data points are drawn. This inference is done by means of K centroids that should be on the manifold if the algorithm runs well. It is based also on a Lloyd-type algorithm where in the updating step, the centroid is computed as the barycenter of the q nearest neighbours of the barycenter of the cluster. In the assignment step, the data point is assigned according to a Bregman divergence. This algorithm has two hyper-parameters: q, the number of neighbors used to compute the centroid and the number of clusters K. This leads to the following complexity for one iteration: $\mathcal{O}(Kqp + n.\log n)$.

The implementations used for the clustering approaches to compare the MOM-based ones in this experiment are publicly available. Table 13 details the programming languages and associated libraries used as well as selected hyper-parameters. Let us note that the trimming approach from the TRIMCLUSTER[41] R package has been implemented in Python language in order to be able to play easily with the initialization conditions.

Simulation context

We dispose of $N = 1500$ points of dimension $p = 3$ which are generated according to a mixture of $K \in \{3, 4, 5\}$ multivariate Gaussian density functions with isotropic covariance matrix. The average

Algorithm	Language	hyper-parameters
K-means	Python [35]	K, $init$ =given*,$n\_init = 1$
K-medoids	Python [42]	$initial\_index\_medoids$ = given*
K-medians	Python [42]	$initial\_centers$ =given*
trimK-means	R [40, 41]	K, $trim = m/n$, $runs = n$, $points$ =given*,$maxit = 300$
K-pdtm	Python [10, 43]	K, $query\_pts$ =given*,$q = K$, $k = K$, $sig = N - m$,
		$iter\_max = 300$, $nstart = 1$,$leaf\_size = 30$
K-bMOM	Python [44]	K, $n_B = 5K$, $B = 500$, $iter\_max = 25$, $initial\_centers$ =given*

*given: same centers obtained either with a random initialization or K-bMOM-km++

Table 13: Implementations and hyper-parameters

vectors for the K components are respectively $\mu_1 = [0, 1, 4]$, $\mu_2 = [2, 1, 0]$, $\mu_3 = [0, -2, 3]$, for $K = 3$, $\mu_4 = [0, 5, -5]$ is added in the case $K = 4$, then $\mu_5 = [-1, -2, 0]$ for $K = 5$. Isolated outliers have been generated by randomly taken 30 datapoints from the regular dataset and their coordinates have been multiplied by a factor of +/-10. The proportion of outliers is therefore equal to 0.02, which imposes some restrictions on the K-bMOM algorithm, notably on the block size which needs to be smaller than 22 (cf. lookup Table 1 in Section 4.1). Besides, all the algorithms have been initialized with the exact same conditions: a random initialization, a K-means++ strategy and a K-bMOM-km++ strategy presented in Section 5.2. These conditions have been repeated 1000 times and in order to compare the performances of these algorithms, the RMSE, the distortion and the accuracy have been computed based on the true parameters of data distribution and their label membership. Moreover, the average number of clusters found among the regular data have also been computed.

General Results and Analysis

First of all, let us note that the clustering algorithms have been executed on regular data with the same random initializations. As it can be observed on Table 14, all the methods perform equally well in average and whatever the number of clusters, when there is no outlier.

The average accuracies calculated for the different number of groups are summarized in Table 15. As it can be observed, the initialization scheme affects most of the clustering approaches and the

	Average performances on regular data ($m/n = 0$)					
	K-means	**K-pdtm**	**trimK-means**	**K-medians**	**K-medoids**	**K-bMOM**
RMSE	0.071 (0.367)	0.068 (0.384)	0.071 (0.366)	0.726 (0.379)	0.173 (0.579)	0.200 (0.03)
distortion	1.096 (0.671)	1.093 (0.911)	1.144 (0.821)	1.700 (1.241)	1.120 (1.505)	1.111 (0.024)
accuracy	0.995 (0.111)	0.995 (0.127)	0.995 (0.111)	0.989 (0.088)	0.995 (0.121)	0.996 (0.001)

Table 14: Average performances and standard deviations (in brackets) of the K-means and the 5 robust clustering approaches on regular data with the same random initializations, among $K \in \{3, 4, 5\}$.

K	init	**K-means**	**K-pdtm**	**trimK-means**	**K-medians**	**K-medoids**	**K-bMOM**
	1	0.819 (0.162)	0.938 (0.151)	0.938 (0.156)	0.915 (0.082)	0.864 (0.161)	0.943 (0.123)
3	2	0.613 (0.124)	0.931 (0.131)	0.833 (0.124)	0.612 (0.123)	0.613 (0.124)	0.810 (0.163)
	3	0.994 (0.006)	0.942 (0.123)	0.997 (0.001)	0.980 (0.059)	0.995 (0.003)	0.997 (0.002)
	1	0.869 (0.124)	0.889 (0.142)	0.869 (0.124)	0.903 (0.116)	0.839 (0.120)	0.879 (0.124)
4	2	0.502 (0.001)	0.916 (0.116)	0.915 (0.056)	0.502 (0.001)	0.502 (0.001)	0.958 (0.001)
	3	0.988 (0.049)	0.989 (0.021)	0.988 (0.049)	0.984 (0.0498)	0.988 (0.049)	0.988 (0.049)
	1	0.820 (0.110)	0.860 (0.155)	0.900 (0.114)	0.896 (0.104)	0.854 (0.113)	0.961 (0.075)
5	2	0.534 (0.093)	0.894 (0.141)	0.777 (0.134)	0.520 (0.097)	0.534 (0.093)	0.996 (0.002)
	3	0.965 (0.069)	0.898 (0.132)	0.996 (0.002)	0.971 (0.058)	0.984 (0.042)	0.996 (0.002)

Table 15: Average accuracies and standard deviations of clustering approaches depending on the number of clusters and the initialization scheme with type (1) random initialization, type (2) K-means++ and type (3) K-bMOM-km++.

more sensitive ones are the K-means, K-medians and K-medoids procedures with a K-means++ initialization. The average accuracies of K-bMOM and trimK-means decreases a little bit when initialized by a K-means++ procedure. On the contrary, K-pdtm seems to be quite insensitive to the tested initialization scheme.

Specific Results and Analysis

We are going to focus on the more complex case ie $K = 5$ and present all the performance criteria on this configuration. The results of the simulated context with different initialization schemes presented above are summarized in Tables 16, 17, 18 where averages and standard deviations of the

RMSE, the distortion, the accuracy and the number of clusters are displayed. Let us note that the detailed contents of performances for the cases $K = 3$ and $K = 4$ are available in the appendix.

Besides, the whole distribution of 1000 repetitions for each metric and tested algorithm are illustrated according to violinplots in Figures 15, 16, 17 and 18 where the median of each distribution is depicted by a bold black dash and the average by a thin black dash.

It is interesting to observe that the initialization conditions have non negligible impacts on the robust K-means algorithms of the literature. Random initialization gives convincing results for all the robust algorithms. The K-means algorithm on the other hand systematically positions a cluster on an outlier as can be seen in Table 16 and as illustrated in Figure 18 a. When the initialization is of type K-means++, the traditional robust algorithms (K-medians and K-medoids) do not find the right clusters and the associated performances are weak (see Table 17 and figures 15b, 16b, 17b and 18b). On the other hand, K-pdtm, trimK-means and K-bMOM resist rather well. Finally, when a robust initialization instantiates all the algorithms, their performances are very good and equivalent. It is still interesting to note the real impact of the initial conditions on the clustering task especially on data with outliers whatever the iterative procedure.

7. Color quantization in image processing

We have seen in Section 5.2 that using a robust initialization even in the context of few outliers, is the best strategy in terms of resulting partition stability and accuracy. Given that spirit, in this last experimental section, the K-bMOM procedure is applied to the problem of color quantization addressed in image processing and computer graphics.

Color quantization (CQ) is a procedure commonly used for color analysis, image compression, segmentation, non photorealistic rendering, etc. It is a process which aims to reduce the number of colors used in an image with the goal of keeping the same quality of visualisation as the original. CQ is a challenging problem since most of real-world images contain tens of thousands of colors. CQ can be viewed as a 3-dimensional clustering problem according to the Red, Green, Blue channels of pixels of an image. A wide literature is devoted to this problem and it appears that the K-means algorithm is not used so often mainly because of its sensitivity to the initialization. We propose therefore to use the K-bMOM procedure as a robust CQ process providing confident and high-quality quantization on a noisy image.

context	methods	RMSE	distortion	accuracy	nb K
random initialization	K-means	1.076 (0.264)	2.560 (0.680)	0.820 (0.110)	4.1 (0.5)
	K-pdtm	0.508 (0.491)	1.993 (1.199)	0.860 (0.155)	4.9 (0.3)
	trimK-means	0.389 (0.385)	1.577 (0.662)	0.900 (0.114)	4.8 (0.3)
	K-medians	0.993 (0.559)	2.498 (1.569)	0.896 (0.104)	4.7 (0.4)
	K-medoids	1.068 (0.32)	2.535 (1.042)	0.854 (0.113)	4.7 (0.4)
	K-bMOM	0.29 (0.236)	1.284 (0.375)	0.961 (0.075)	4.9 (0.2)

Table 16: Average and standard deviations of accuracies of K-means and 5 robust algorithms for different settings.

initialization	methods	RMSE	distortion	accuracy	nb K
K-means++ initialization	K-means	1.669 (0.271)	4.881 (1.09)	0.534 (0.093)	2.6 (0.5)
	K-pdtm	0.385 (0.393)	1.742 (0.977)	0.884 (0.141)	5.0 (0.0)
	trimK-means	0.624 (0.452)	1.836 (1.437)	0.857 (0.134)	4.9 (0.3)
	K-medians	1.828 (0.785)	7.650 (2.454)	0.520 (0.097)	2.6 (0.4)
	K-medoids	1.686 (0.449)	5.128 (1.203)	0.534 (0.093)	2.6 (0.4)
	K-bMOM	0.186 (0.034)	1.105 (0.025)	0.996 (0.002)	5.0 (0.0)

Table 17: Average and standard deviations of the performances 5 robust algorithms on polluted data with K-means++ initializations.

initialization	methods	RMSE	distortion	accuracy	nb K
K-bMOM-km++ initialization	K-means	0.882 (0.122)	1.891 (0.286)	0.965 (0.069)	4.8 (0.3)
	K-pdtm	0.427 (0.429)	1.743 (0.871)	0.888 (0.132)	4.9 (0.2)
	trimK-means	0.063 (0.011)	1.067 (0.019)	0.996 (0.002)	5.0 (0.0)
	K-medians	0.738 (0.228)	1.638 (0.365)	0.971 (0.058)	4.9 (0.3)
	K-medoids	0.819 (0.154)	1.721 (0.205)	0.984 (0.042)	5.0 (0.0)
	K-bMOM	0.185 (0.025)	1.096 (0.022)	0.996 (0.002)	5.0 (0.0)

Table 18: Average and standard deviations of the performances 5 robust algorithms on polluted data with a robust initialization (K-bMOM-km++).

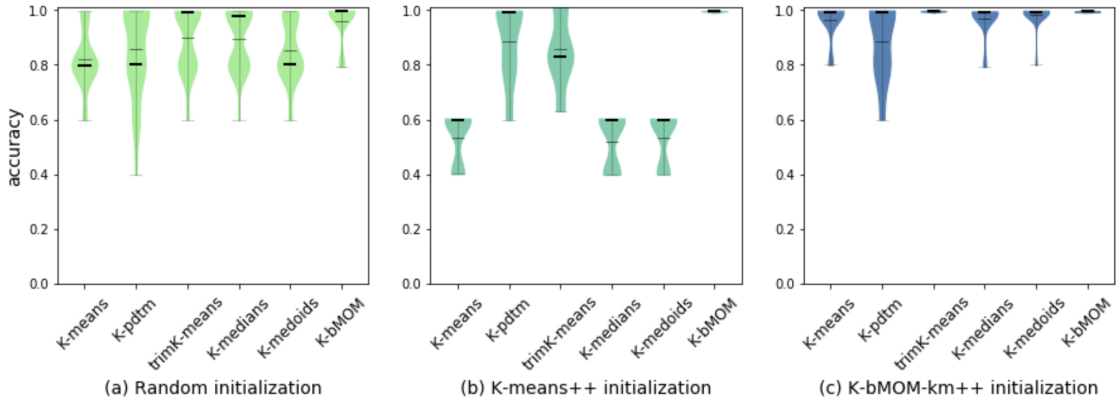


Figure 15: Violinplots of accuracies according to different initialization strategies.

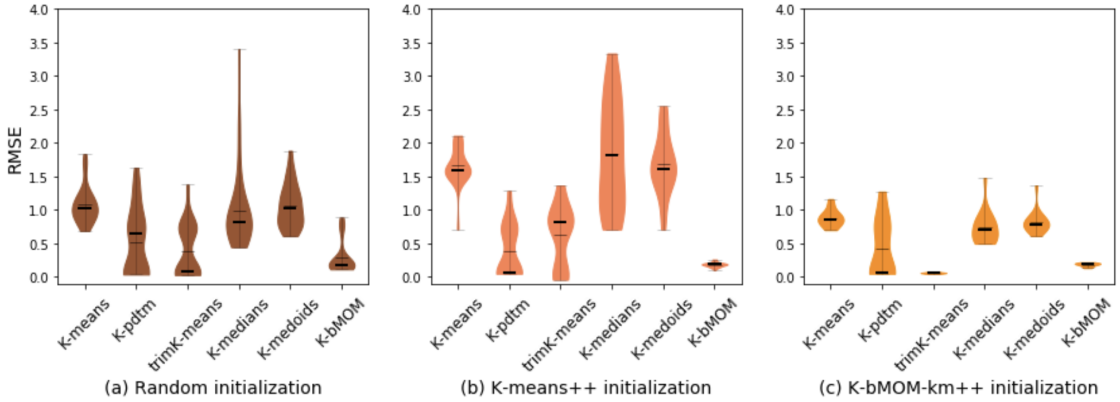


Figure 16: Violinplots of RMSE according to different initialization strategies.

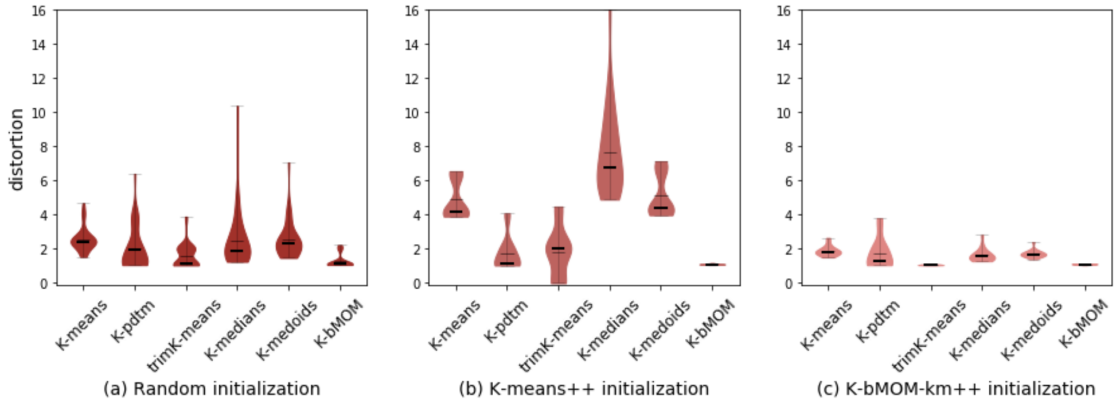


Figure 17: Violinplots of distortions according to different initialization strategies.

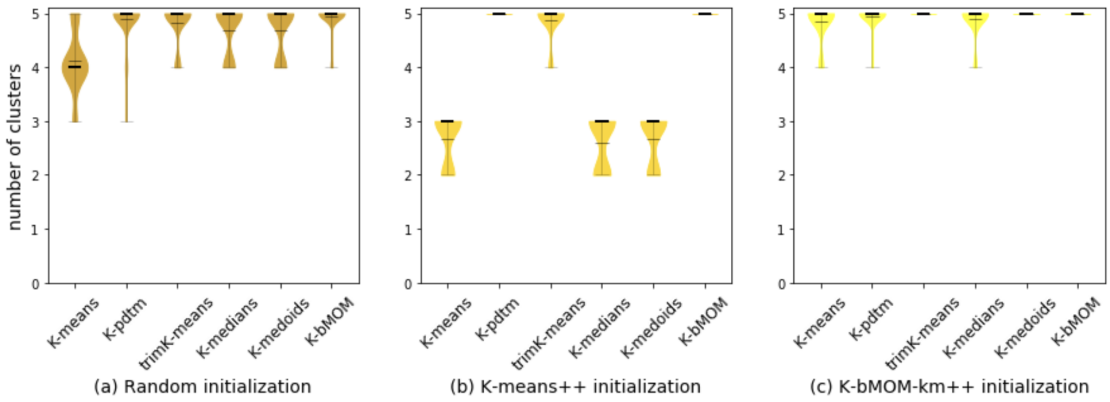


Figure 18: Violinplots of number of clusters according to different initialization strategies.

7.1. Images and experimental setup

The K-bMOM method has been used on a popular 24-bit test image, the Parrots, (768 × 512) coming from the Kodak Lossless True Color Image Suite database and illustrated in Figure 19 a.:

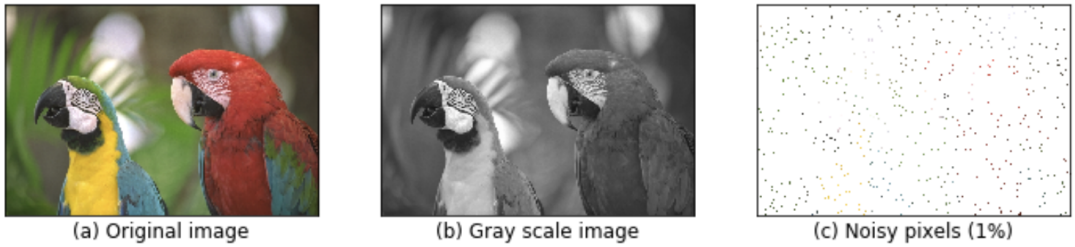


Figure 19: Original Image (left), Gray scale image (middle) and noisy filter applied on the gray scale image (left).

The K-bMOM method has been used on the gray-scale image of parrots as shown in 19 b. from which 1% of randomised pixels have been colorized by the color of the same pixel of the original image. In our experiment, we consider these colored pixels, illustrated in Figure 19 c., as outliers of the gray-scale image. The objective is therefore to extract the main shades of gray of the image. The image of shape 768 × 512 pixels has been shaped into a matrix of 3-dimensions linked to Red Green Blue (RGB) channels. The K-bMOM algorithm has been repeated 50 times for a number K of gray levels (or clusters) equals to 8, 16 and 24 respectively. For these 3 segmentations, the number of blocks has been set to $B = 1000$ and the size of each block set to $n_B = 5 * K$.

7.2. Experimental results

In order to evaluate the quality of the quantization, the empirical distortion has been computed between the pixels $x_i \in \mathbb{R}^3, \forall i \in 1, \ldots, n$ of the original gray-scale image and their segmented version $\hat{c}_k$, their nearest color. It has been averaged among 50 repetitions and the standard deviation has also been computed. Moreover, in order to evaluate the robustness property of the K-bMOM algorithm, the number of fitted gray levels has been computed by comparing the R,G,B values of each centroid. It is expected to have K levels of gray (ie no color among the fitted centroids).

The results are summarized in Table 19. First of all, it can be noted that color quantization processed by the K-bMOM approach seems to be robust. Indeed, the fitted centroids are in the shades of gray: the number of gray levels equals the number of clusters.

	K=8	K=16	K=24
distortion	171.0 (1.69)	56.40 (27.01)	28.70 (3.42)
number of gray levels	8.00 (0.00)	16.00 (0.58)	24.00 (0.00)

Table 19: Median and standard deviation (in parenthesis) of the distortion and the number of gray levels obtained by the K-bMOM procedure for $K = \{8, 16, 24\}$ groups.

Figures 20 illustrates the quantization process on noisy gray-scale Parrot image for $K = 8$, 16 and 24 respectively. Figure 21 show the distortion per pixel in a reverse gray scaled mapping which means that the higher the error, the darker the pixel. It can be seen that the K-bMOM approach performs well in allocating $K-$representative gray levels to the different image regions.

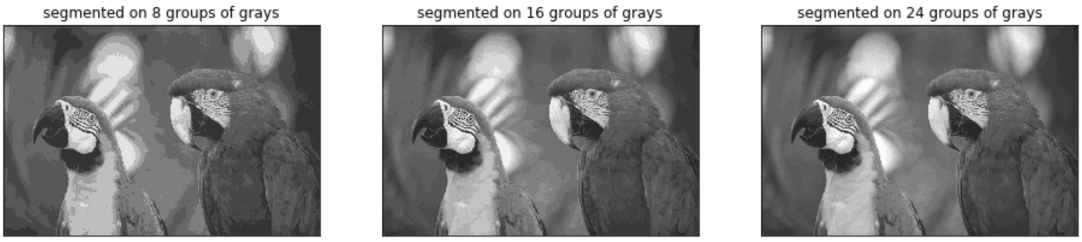


Figure 20: Sample quantization results for $K = 8$, 16 and 32 respectively from left to right on gray-scale noisy Parrot image.

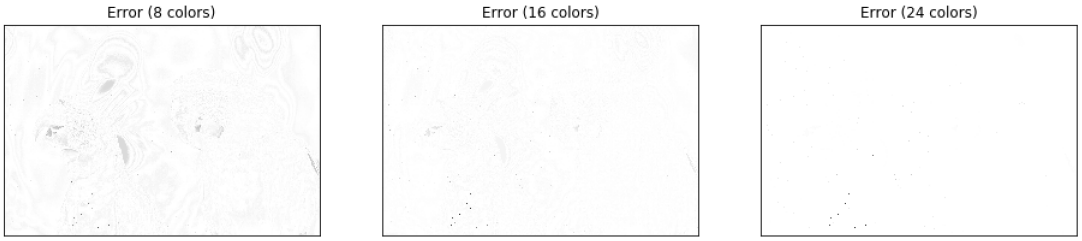


Figure 21: Full scale error images for $K = 8$, 16 and 32 respectively from left to right gray-scale Parrot image.

8. Proof of Theorem 3

Assume without loss of generality, that $B \geq 8$ (otherwise the bound stated in Theorem 3 may occur with probability zero). We have, by definition of $\hat{\mathbf{c}}_n$, for any constant $a > 0$,

$$\begin{aligned}
&\mathbb{P}\left(R\left(\hat{\mathbf{c}}_n\right) - R_* > a\right) \\
\leq &\mathbb{P}\left(\inf_{\mathbf{c}\in\mathcal{F}_{>a}} \mathrm{MOM}\left(\ell_{\mathbf{c}}\right) \leq \inf_{\mathbf{c}\in\mathcal{F}_a} \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right) \\
= &\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{F}_{>a}} \left\{R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right\} \geq \sup_{\mathbf{c}\in\mathcal{F}_a} \left\{R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right\}\right) \\
\leq &\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{F}_{>a}} \left\{R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right\} \geq R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}_*}\right)\right),
\end{aligned}$$

where $\mathcal{F}_a = \left\{\mathbf{c} \in \mathcal{X}_{M_*}^k : R(\mathbf{c}) - R_* \leq a\right\}$ and $\mathcal{F}_{>a} = \left\{\mathbf{c} \in \mathcal{X}_{M_*}^k : R(\mathbf{c}) - R_* > a\right\} = \mathcal{X}_{M_*}^k \setminus \mathcal{F}_a$. Now, on the one hand, for any $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{MOM}\left(\ell_{\mathbf{c}_*}\right) - R_* \geq x\right) \\
= &\mathbb{P}\left(\sum_{j=1}^{B} \mathbf{1}_{\left\{\left(P_{b_j} - P\right)\left(\ell_{\mathbf{c}_*}\right) \geq x\right\}} \geq \frac{B}{2}\right) \\
\leq &\mathbb{P}\left(\sum_{j\in I} \mathbf{1}_{\left\{\left(P_{b_j} - P\right)\left(\ell_{\mathbf{c}_*}\right) \geq x\right\}} \geq \frac{B}{2} - |O|\right) \\
\leq &\sum_{j=\lfloor B/2 - |O|\rfloor}^{B} \binom{B}{j} p^j (1-p)^{B-j} \\
&\leq p^{\lfloor B/2 - |O|\rfloor} 2^B
\end{aligned}$$

where $p = \mathbb{P}\left(\left(P_{b_j} - P\right)\left(\ell_{\mathbf{c}_*}\right) \geq x\right)$. In addition, by the Markov inequality,

$$p \leq \frac{B\mathrm{Var}\left(\ell_{\mathbf{c}_*}\right)}{nx^2}.$$

Hence, by choosing $x = \sqrt{64eB\mathrm{Var}\left(\ell_{\mathbf{c}_*}\right)/n}$, we get

$$\mathbb{P}\left(\mathrm{MOM}\left(\ell_{\mathbf{c}_*}\right) - R_* \geq x\right) \leq 2^B \left(\frac{1}{64e}\right)^{\lfloor B/2 - |O| \rfloor}.$$

Note that since $|O| \leq B/4$ and $B \geq 8$, we have $\lfloor B/2 - |O| \rfloor \geq \lfloor B/4 \rfloor \geq B/8$ and $2^B \leq 16^{\lfloor B/4 \rfloor + 1} \leq 64^{\lfloor B/4 \rfloor}$. This gives

$$\mathbb{P}\left(\mathrm{MOM}\left(\ell_{\mathbf{c}_*}\right) - R_* \geq x\right) \leq \exp\left(-\frac{B}{8}\right).$$

On the other hand,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{F}_{>a}} \left\{R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right\} \geq -x\right) \\
\leq &\mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{F}_{>a}} \left\{\frac{1}{B}\sum_{j=1}^{B} \mathbf{1}_{\left\{R_* - P_{b_j}\left(\ell_{\mathbf{c}}\right) \geq -x\right\}}\right\} \geq \frac{1}{2}\right) \\
\leq &\mathbb{P}\left(\sup_{\mathbf{c} \in \mathcal{F}_{>a}} \left\{\frac{1}{|I|}\sum_{j \in I} \mathbf{1}_{\left\{R_* - P_{b_j}\left(\ell_{\mathbf{c}}\right) \geq -x\right\}}\right\} \geq \frac{B}{2|I|} - \frac{|O|}{|I|}\right)
\end{aligned}$$

Let us denote $\Delta = B/2|I| - |O|/|I|$ and set

$$Z\left(\mathcal{F}_{>a}, x\right) = \sup_{\mathbf{c} \in \mathcal{F}_{>a}} \frac{1}{|I|}\sum_{j \in I} \mathbf{1}_{\left\{R_* - P_{b_j}\left(\ell_{\mathbf{c}}\right) \geq -x\right\}}. \tag{7}$$

By Corollary 1 (see Section 8.1 below), we have

$$\mathbb{P}\left(Z\left(\mathcal{F}_{>a}, x\right) \geq \Delta\right) \leq \exp\left(-\frac{|I|(\Delta - \mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right])^2}{2\mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right] + 2(\Delta - \mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right])/3}\right). \tag{8}$$

It remains to control $\mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right]$. Consider a function $\phi : \mathbb{R} \to \mathbb{R}$, such that $\phi(t) = (t -$

$1)1_{\{1\leq t\leq 2\}} + 1_{\{t\geq 2\}}$. The function ϕ is thus 1-Lipschitz and it holds $\phi(t) \geq 1_{\{t\geq 2\}}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right] = \mathbb{E}&\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\mathbf{1}_{\left\{\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})\geq R(\mathbf{c})-R_*-x\right\}}\right\}\right] \\
\leq&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\mathbf{1}_{\left\{\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})\geq a-x\right\}}\right\}\right] \\
\leq&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\phi\left(\frac{2\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})}{a-x}\right)\right\}\right] \qquad (9)
\end{aligned}$$

Now, for any $i \in I$,

$$\mathbb{E}\left[\phi\left(\frac{2\left(P-P_{b_i}\right)(\ell_{\mathbf{c}})}{a-x}\right)\right] \leq \mathbb{P}\left[\left(P-P_{b_i}\right)(\ell_{\mathbf{c}}) \geq (a-x)/2\right] \leq \frac{BL}{n\left(a-x\right)^2},$$

where the constant L is such that $\sup_c \mathrm{Var}\left(\ell_c\right) \leq L$. More explicitly, we can choose $L = 16M^2\mathbb{E}\left[\|X\|^2\right]$. Hence, by Inequality (9) we get,

$$\mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right] \leq \frac{BL}{n\left(a-x\right)^2} + \mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\phi\left(\frac{2\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})}{a-x}\right) - \mathbb{E}\left[\phi\left(\frac{2\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})}{a-x}\right)\right]\right\}\right].$$

Now, by a standard symmetrisation argument, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\phi\left(\frac{2\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})}{a-x}\right) - \mathbb{E}\left[\phi\left(\frac{2\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})}{a-x}\right)\right]\right\}\right] \\
\leq 2\mathbb{E}&\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\epsilon_j\phi\left(\frac{2\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})}{a-x}\right)\right\}\right],
\end{aligned}$$

where the ϵ_j's are i.i.d. Rademacher variables (i.e. $\mathbb{P}(\epsilon_j = 1) = \mathbb{P}(\epsilon_j = -1) = 1/2$) independent of the sample. Furthermore, as the function ϕ is 1-Lipschitz and $\phi(0) = 0$, we can apply the so-called contraction principle ([45, Section 4]), which gives

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\epsilon_j\phi\left(\frac{2\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})}{a-x}\right)\right\}\right] \\
\leq&\frac{2}{a-x}\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\epsilon_j\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})\right\}\right]
\end{aligned}$$

and by symmetrisation again,

$$\begin{aligned}&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|I|}\sum_{j\in I}\epsilon_j\left(P-P_{b_j}\right)(\ell_{\mathbf{c}})\right\}\right]\\ \leq&\frac{2B}{|I|n}\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\sum_{i\in\mathcal{J}}\epsilon_i\ell_{\mathbf{c}}\left(X_i\right)\right\}\right],\end{aligned}$$

where $\mathcal{J}=\bigcup_{j\in I}b_j$. By Lemma 4.3 in [29],

$$\begin{aligned}\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\sum_{i\in\mathcal{J}}\epsilon_i\ell_{\mathbf{c}}\left(X_i\right)\right\}\right]&\leq 2K\sqrt{|\mathcal{J}|}\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right].\\ &\leq 2K\sqrt{n}\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right]\end{aligned}$$

Putting things together, we obtain

$$\mathbb{E}\left[Z\left(\mathcal{F}_{>a},x\right)\right]\leq\frac{BL}{n\left(a-x\right)^2}+\frac{8B}{\left(a-x\right)|I|\sqrt{n}}2K\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right].$$

Now, by taking

$$a\geq\max\left\{2x,4\sqrt{\frac{BL}{n\Delta}},\frac{128BK\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right]}{\Delta|I|\sqrt{n}}\right\},\tag{10}$$

we get

$$\frac{BL}{n\left(a-x\right)^2}\leq\frac{\Delta}{4}$$

and

$$\frac{8B}{\left(a-x\right)|I|\sqrt{n}}2K\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right]\leq\frac{\Delta}{4}.$$

This gives $\mathbb{E}\left[Z\left(\mathcal{F}_{>a},x\right)\right]\leq\Delta/2$ and so, by using Inequality (12),

$$\mathbb{P}\left(Z\left(\mathcal{F}_{>a},x\right)\geq\Delta\right)\leq\exp\left(-\frac{3|I|\Delta}{16}\right).$$

To conclude, it suffices now to notice that if $n_o\leq B/4$, then $|O|\leq B/4$, $|I|\geq 3B/4$ and $\Delta\geq B/(4|I|)\geq 1/4$. Indeed, in this case, Inequality (10) is achieved by choosing for instance

$$a=\max\left\{8\sqrt{\frac{eBL}{n}},512\frac{K\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right]}{\sqrt{n}}\right\}.$$

8.1. A concentration inequality

We will derive here a concentration inequality for the stochastic process $Z(\mathcal{F}_{>a}, x)$ defined in (7). We proceed by proving that $Z(\mathcal{F}_{>a}, x)$ satisfies the so-called self-bounding condition, we recall now (see also [46, Theorem 6.12]).

Definition 6 *A function f is said to have the self-bounding property if, for some functions $f_i : \mathcal{Z}^{n-1} \to \mathbb{R}$, for all $z = (z_1, ..., z_n) \in \mathcal{Z}^n$ and for all $i = 1, ..., n$,*

$$0 \leq f(z) - f_i\left(z^{(i)}\right) \leq 1$$

and

$$\sum_{i=1}^{n} \left(f(z) - f_i\left(z^{(i)}\right)\right) \leq f(z) \ ,$$

where $z^{(i)} = (z_1, ..., z_{i-1}, z_{i+1}, ..., z_n)$.

Lemma 1 *If $\mathcal{A}$ is a class of sets on a measurable space $(\mathcal{Z}, \mathcal{T})$, then the function $h : \mathcal{Z}^p \to \mathbb{R}$ defined by*

$$h(z_1, ..., z_p) = \sup_{A \in \mathcal{A}} \sum_{j=1}^{p} 1_A(z_j) \ ,$$

has the self-bounding property. By consequence, if $(\xi_1, ..., \xi_p) \in \mathcal{X}^p$ is an i.i.d. sample, then by setting $Z = h(\xi_1, ..., \xi_p)$, it holds for any $t > 0$,

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\left(-\frac{t^2}{2\mathbb{E}Z + 2t/3}\right) \ . \tag{11}$$

Proof. Denote $h_i\left(z^{(i)}\right) = \sup_{A \in \mathcal{A}} \sum_{j \neq i} 1_A(z_j)$. Then

$$0 \leq h(z) - h_i\left(z^{(i)}\right) \leq \sup_{A \in \mathcal{A}} 1_A(z_i) \leq 1 \ .$$

Also, assume without loss of generality that $h(z) = \sum_{j=1}^{I} 1_{A_*(z)}(z_j)$ for some $A_*(z) \in \mathcal{A}$, then

$$\begin{aligned} \sum_{i=1}^{I} \left(h(z) - h_i\left(z^{(i)}\right)\right) &\leq \sum_{i=1}^{I} \left(\sum_{j=1}^{I} 1_{A_*(z)}(z_j) - \sum_{j \neq i} 1_{A_*(z)}(z_j)\right) \\ &= \sum_{i=1}^{I} 1_{A_*(z)}(z_i) = h(z) \ . \end{aligned}$$

Hence, h has the self-bounding property. Now, inequality (11) simply follows from [46, Theorem 6.12]. ■

Corollary 1 *The following process*

$$Z\left(\mathcal{F}_{>a},x\right)=\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\frac{1}{|I|}\sum_{j\in I}\mathbf{1}_{\left\{R_{*}-P_{b_{j}}\left(\ell_{\mathbf{c}}\right)\geq-x\right\}}$$

is concentrated around its expected value according to the following inequality,

$$\mathbb{P}\left(Z\left(\mathcal{F}_{>a},x\right)\geq\Delta\right)\leq\exp\left(-\frac{|I|(\Delta-\mathbb{E}\left[Z\left(\mathcal{F}_{>a},x\right)\right])^{2}}{2\mathbb{E}\left[Z\left(\mathcal{F}_{>a},x\right)\right]+2(\Delta-\mathbb{E}\left[Z\left(\mathcal{F}_{>a},x\right)\right])/3}\right)\ . \tag{12}$$

Proof. It suffices to apply Lemma 1 with $p=n_{B}$, $\mathcal{Z}=\mathcal{X}^{n_{B}}$, $\xi_{i}=\left(X_{j}\right)_{j\in b_{i}}$ for $i\in I$ and

$$\mathcal{A}=\left\{\left\{z=(x_{1},...,x_{n_{B}}):\frac{-1}{|n_{B}|}\sum_{j=1}^{n_{B}}\ell_{\mathbf{c}}\left(x_{j}\right)+R_{*}>-x\right\}:\mathbf{c}\in\mathcal{F}_{>a}\right\}\ .$$

■

9. Appendix

The detailed average performances in the cases $K = 3$ and $K = 4$ are summarized in Tables 20, 21, 22, 23, 24 and Table 25 described below:

context	methods	RMSE	distortion	accuracy	nb K
random initialization $(m/n = 0.2)$	K-means	0.949 (0.252)	2.281 (0.705)	0.819 (0.162)	2.4 (0.5)
	K-pdtm	0.371 (0.502)	1.554 (0.742)	0.938 (0.151)	2.9 (0.2)
	trimK-means	0.379 (0.467)	1.612 (0.764)	0.938 (0.156)	2.9 (0.2)
	K-medians	0.711 (0.299)	1.670 (0.479)	0.915 (0.082)	2.9 (0.2)
	K-medoids	0.997 (0.43)	2.197 (0.877)	0.864 (0.161)	2.9 (0.3)
	K-bMOM	0.32 (0.332)	1.367 (0.589)	0.943 (0.123)	3.0 (0.0)

Table 20: Average and standard deviations of the performances 5 robust algorithms on regular data with random initializations in the case K=3

initialization	methods	RMSE	distortion	accuracy	nb K
K-means++ initialization $(m/n = 0.2)$	K-means	1.380 (0.539)	3.489 (1.495)	0.613 (0.124)	1.8 (0.3)
	K-pdtm	0.254 (0.414)	1.403 (0.645)	0.931 (0.131)	3.0 (0.0)
	trimK-means	1.359 (0.564)	3.446 (1.512)	0.833 (0.124)	2.4 (0.3)
	K-medians	1.778 (1.018)	5.036 (2.200)	0.612 (0.123)	1.8 (0.3)
	K-medoids	1.374 (0.621)	3.577 (1.568)	0.613 (0.124)	1.8 (0.3)
	K-bMOM	0.776 (0.536)	2.189 (0.947)	0.810 (0.163)	2.4 (0.4)

Table 21: Average and standard deviations of the performances 5 robust algorithms on polluted data with random initializations in the case K=3.

initialization	methods	RMSE	distortion	accuracy	nb K
K-bMOM-km++ initialization $(m/n = 0.2)$	K-means	0.661 (0.079)	1.499 (0.103)	0.994 (0.006)	3.0 (0.0)
	K-pdtm	0.816 (3.448)	1.343 (0.618)	0.942 (0.123)	3.0 (0.0)
	trimK-means	0.050 (0.012)	1.071 (0.021)	0.997 (0.001)	3.0 (0.0)
	K-medians	0.710 (0.271)	1.683 (0.593)	0.980 (0.059)	2.9 (0.2)
	K-medoids	0.660 (0.071)	1.499 (0.087)	0.995 (0.003)	3.0 (0.0)
	K-bMOM	0.176 (0.045)	1.098 (0.025)	0.997 (0.002)	3.0 (0.0)

Table 22: Average and standard deviations of the performances 5 robust algorithms on polluted data with a robust initialization (K-bMOM-km++) in the case K=3.

context	methods	RMSE	distortion	accuracy	nb K
random initialization $(m/n = 0.2)$	K-means	0.478 (0.404)	1.678 (0.594)	0.869 (0.124)	4.0 (0.0)
	K-pdtm	0.400 (0.431)	1.738 (1.048)	0.889 (0.142)	4.0 (0.0)
	trimK-means	0.478 (0.404)	1.678 (0.593)	0.869 (0.124)	4.0 (0.0)
	K-medians	0.896 (0.326)	2.342 (1.149)	0.903 (0.116)	4.0 (0.0)
	K-medoids	0.759 (0.661)	2.407 (1.889)	0.839 (0.120)	4.0 (0.0)
	K-bMOM	0.510 (0.370)	1.661 (0.599)	0.879 (0.124)	4.0 (0.0)

Table 23: Average and standard deviations of the performances 5 robust algorithms on regular data with random initializations in the case K=4

initialization	methods	RMSE	distortion	accuracy	nb K
K-means++ initialization $(m/n = 0.2)$	K-means	1.735 (0.210)	5.416 (0.048)	0.502 (0.001)	2.0 (0.0)
	K-pdtm	0.323 (0.393)	1.477 (0.571)	0.916 (0.116)	3.9 (0.2)
	trimK-means	1.704 (0.224)	1.251 (0.451)	0.915 (0.056)	3.9 (0.1)
	K-medians	1.990 (0.755)	8.007 (1.622)	0.502 (0.001)	2.0 (0.0)
	K-medoids	1.810 (0.345)	5.726 (0.135)	0.502 (0.001)	2.0 (0.0)
	K-bMOM	0.176 (0.020)	1.101 (0.021)	0.998 (0.001)	4.0 (0.0)

Table 24: Average and standard deviations of the performances 5 robust algorithms on polluted data with random initializations in the case K=4.

initialization	methods	RMSE	distortion	accuracy	nb K
K-bMOM-km++ initialization $(m/n = 0.2)$	K-means	0.084 (0.161)	1.126 (0.240)	0.988 (0.049)	4.0 (0.0)
	K-pdtm	0.378 (0.399)	1.558 (0.592)	0.989 (0.121)	4.0 (0.0)
	trimK-means	0.084 (0.161)	1.126 (0.240)	0.988 (0.049)	4.0 (0.0)
	K-medians	0.774 (0.208)	1.712 (0.350)	0.984 (0.048)	4.0 (0.0)
	K-medoids	0.168 (0.169)	1.142 (0.243)	0.988 (0.049)	4.0 (0.0)
	K-bMOM	0.201 (0.152)	1.151 (0.240)	0.988 (0.049)	4.0 (0.0)

Table 25: Average and standard deviations of the performances 5 robust algorithms on polluted data with a robust initialization (K-bMOM-km++) in the case K=4.

References

[1] M. Lerasle, R. I. Oliveira, Robust empirical mean estimators, arXiv preprint arXiv:1112.3914.

[2] L. Devroye, M. Lerasle, G. Lugosi, R. I. Oliveira, Sub-Gaussian mean estimators, Ann. Statist. 44 (6) (2016) 2695–2725.

[3] G. Lecué, M. Lerasle, Learning from MOM's principles: Le Cam's approach, Stochastic Process. Appl. 129 (11) (2019) 4385–4410.

[4] G. Lecué, M. Lerasle, Robust machine learning by median-of-means: theory and practice, arXiv preprint arXiv:1711.10306.

[5] G. Lugosi, S. Mendelson, Sub-Gaussian estimators of the mean of a random vector, Ann. Statist. 47 (2) (2019) 783–794.

[6] G. Lugosi, S. Mendelson, Mean estimation and regression under heavy-tailed distributions: a survey, Found. Comput. Math. 19 (5) (2019) 1145–1190.

[7] G. Lugosi, S. Mendelson, Risk minimization by median-of-means tournaments, J. Eur. Math. Soc. (JEMS) 22 (3) (2020) 925–965.

[8] S. Minsker, Uniform bounds for robust mean estimators, arXiv preprint arXiv:1812.03523.

[9] L. A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar, A review of robust clustering methods, Adv. Data Anal. Classif. 4 (2-3) (2010) 89–109.

[10] C. Brécheteau, Robust shape inference from a sparse approximation of the Gaussian trimmed loglikelihood, preprint (2018).

[11] I. Diakonikolas, D. M. Kane, Recent Advances in Algorithmic High-Dimensional Robust Statistics (2019). `arXiv:1911.05911`.

[12] P. Bühlmann, Bagging, subagging and bragging for improving some prediction algorithms, in: Recent advances and trends in nonparametric statistics, Elsevier B. V., Amsterdam, 2003, pp. 19–34.

[13] P. Laforgue, S. Clémençon, P. Bertail, On Medians of (Randomized) Pairwise Means, in: 36th International Conference on Machine Learning, Vol. 97, 2019.

[14] L. A. García-Escudero, A. Gordaliza, Robustness properties of k means and trimmed k means, J. Amer. Statist. Assoc. 94 (447) (1999) 956–969.

[15] Y. Klochkov, A. Kroshnin, N. Zhivotovskiy, Robust k-means clustering for distributions with two moments, arXiv preprint arXiv:2002.02339v1.

[16] N. Nguyen, R. Caruana, Consensus clusterings, in: Seventh IEEE international conference on data mining (ICDM 2007), IEEE, 2007, pp. 607–612.

[17] F. Leisch, Bagged Clustering, Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science".
URL `http://epub.wu.ac.at/1272/1/document.pdf`

[18] S. Dolnicar, F. Leisch, Winter tourist segments in Austria: Identifying stable vacation styles using bagged clustering techniques, Journal of Travel Research 41 (3) (2003) 281–292.

[19] P. D'Urso, L. De Giovanni, M. Disegna, R. Massari, Bagged clustering and its application to tourism market segmentation, Expert Systems with Applications 40 (12) (2013) 4944–4956.

[20] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, A. Mayo-Íscar, Robust clustering tools based on optimal transportation, Stat. Comput. 29 (1) (2019) 139–160.

[21] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2) (2001) 411–423.

[22] J.-P. Baudry, C. Maugis, B. Michel, Slope heuristics: overview and implementation, Stat. Comput. 22 (2) (2012) 455–470.

[23] D. Arthur, S. Vassilvitskii, K-means++: The Advantages of Careful Seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, 2007, pp. 1027–1035.

[24] P. J. Huber, E. M. Ronchetti, Robust statistics, 2nd Edition, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, 2009.

[25] R. A. Maronna, R. D. Martin, V. J. Yohai, M. Salibián-Barrera, Robust statistics, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, 2019.

[26] D. Rodriguez, M. Valdora, The breakdown point of the median of means tournament, Statist. Probab. Lett. 153 (2019) 108–112.

[27] B. Bercu, B. Delyon, E. Rio, Concentration inequalities for sums and martingales, SpringerBriefs in Mathematics, Springer, Cham, 2015.

[28] G. Ritter, Robust cluster analysis and variable selection, Vol. 137 of Monographs on Statistics and Applied Probability, CRC Press, Boca Raton, FL, 2015.

[29] G. Biau, L. Devroye, G. Lugosi, On the performance of clustering in Hilbert spaces, IEEE Trans. Inform. Theory 54 (2) (2008) 781–790.

[30] A. Fischer, Quantization and clustering with Bregman divergences, J. Multivariate Anal. 101 (9) (2010) 2207–2221.

[31] G. Lecué, M. Lerasle, T. Mathieu, Robust classification via MOM minimization, Mach. Learn. 109 (8) (2020) 1635–1665.

[32] H. Narayanan, S. Mitter, Sample complexity of testing the manifold hypothesis, in: Advances in neural information processing systems, 2010, pp. 1786–1794.

[33] C. Fefferman, S. Mitter, H. Narayanan, Testing the manifold hypothesis, J. Amer. Math. Soc. 29 (4) (2016) 983–1049.

[34] M. Al Hasan, V. Chaoji, S. Salem, M. J. Zaki, Robust partitional clustering by outlier and density insensitive seeding, Pattern Recognition Letters 30 (11) (2009) 994–1002.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[36] L. Kaufman, P. J. Rousseeuw, Clustering by means of medoids, Statistical Data Analysis Based on the L1 Norm and Related Methods, Amsterdam: North-Holland, 1987, pp. 405–416.

[37] E. Schubert, P. J. Rousseeuw, Faster k-medoids clustering: Improving the pam, clara, and clarans algorithms, arXiv preprint arXiv:1810.05691.

[38] A. K. Jain, R. C. Dubes, Algorithms for clustering data, Englewood Cliffs: Prentice Hall, 1988.

[39] J. Hartigan, M. Wong, Algorithm AS 136: A K-means clustering algorithm, Applied Statistics (1979) 100–108.

[40] J. A. Cuesta-Albertos, A. Gordaliza, C. Matrán, Trimmed k-means: An attempt to robustify quantizers, The Annals of Statistics 25 (2) (1997) 553–576.

[41] C. Hennig, trimcluster : Cluster Analysis with Trimming, R package version 0.1-5 (2021).

[42] A. Novikov, PyClustering: Data Mining Library, Journal of Open Source Software 4 (36) (2019) 1230.

[43] C. Brecheteau, https://www.math.sciences.univ-nantes.fr/ brecheteau/notebooks/ Notebook_kPDTM_kPLM.html, 2020.

[44] C. Brunet-Saumard, E. Genetay, https://github.com/csaumard/kbmom, 2021.

[45] M. Ledoux, M. Talagrand, Probability in Banach spaces, Springer, Berlin, 1991.

[46] S. Boucheron, G. Lugosi, P. Massart, Concentration Inequalities: A Nonasymptotic Theory of Independence, Oxford University Press, Oxford, 2013.