ESAIM: PROCEEDINGS AND SURVEYS, November 2023, Vol. 74, p. 119-136
Editors: Marguerite Zani, Clément Marteau - Journées MAS 2020, Fédération Siméon Poisson (Orléans-Tours),

TOPICS IN ROBUST STATISTICAL LEARNING *

Claire Brecheteau[1], Edouard Genetay[2], Timothee Mathieu[3] and Adrien Saumard[4]

Abstract. Some recent contributions to robust inference are presented. Firstly, the classical problem of robust M-estimation of a location parameter is revisited using an optimal transport approach - with specifically designed Wasserstein-type distances - that reduces robustness to a continuity property. Secondly, a procedure of estimation of the distance function to a compact set is described, using union of balls. This methodology originates in the field of topological inference and offers as a byproduct a robust clustering method. Thirdly, a robust Lloyd-type algorithm for clustering is constructed, using a bootstrap variant of the median-of-means strategy. This algorithm comes with a robust initialization.

Résumé. Quelques contributions récentes à l'inférence robuste sont présentées. Premièrement, le problème classique de la M-estimation robuste d'un paramètre de localisation est revisité en utilisant une approche de transport optimal, avec des distances de type Wasserstein spécifiquement conçues, qui réduit la robustesse à une propriété de continuité. La deuxième contribution décrit une procédure d'estimation de la fonction de distance à un ensemble compact, en utilisant une union de boules. Cette méthodologie trouve son origine dans le domaine de l'inférence topologique et offre comme sous-produit une méthode de clustering robuste. Enfin, un algorithme robuste de type Lloyd pour le clustering est présenté, en utilisant une variante bootstrap de la stratégie "median-of-means". Cet algorithme s'accompagne notamment d'une initialisation robuste.

Introduction

This article presents some recent studies on the topic of robust statistical learning. The results were presented at the session Robust Statistical Learning of the Journées MAS 2020, organized by Adrien Saumard. The three sections of the article are based on the talks given by Claire Brécheteau, Edouard Genetay and Timothée Mathieu. A fourth talk was given by Jules Depersin, but it will be only briefly mentioned in this introduction. Adrien Saumard has written the introduction and coordinated the present article.

After the seminal work of Catoni [10] on sub-Gaussian estimation of the mean and variance, robust estimation has recently seen a resurgence of theoretical interest [30]. On the practical side, robustness is a property more sought after than ever, since the data scientist often faces massive and complex datasets that may contain a variety of outliers. Designing efficient robust procedures that allow to get rid of most of the usual, time consuming, preprocessing of data has thus become an attractive direction of research in the machine learning

* *Timothée Mathieu is thankful to Matthieu Lerasle and Guillaume Lecué for their support and their precious advices when writing this work.*

[1] Univ. Rennes 2, Rennes, France. claire.brecheteau@univ-rennes2.fr

[2] CREST, ENSAI, Univ. Rennes, LumenAI, Tours, France. egenetay@lumenai.fr

[3] INRIA, Scool team. Univ. Lille, CRIStAL, CNRS, France. timothee.mathieu@inria.fr

[4] CREST, ENSAI, Univ. Rennes, Bruz, France. adrien.saumard@ensai.fr

Article published online by EDP Sciences and available at https://www.esaim-proc.org or https://doi.org/10.1051/proc/202374119

community. This led to several polynomial time robust learning algorithms [18], the research being still very active in this area.

Before introducing the different sections of this article, let us first describe in a few lines the presentation of Jules Depersin, which is based on his preprint written in collaboration with Guillaume Lecué [16].

Median-of-means (MOM) versions of the median absolute deviation (MAD) and Stahel-Donoho outlyingness (SDO) functions are studied, aiming at constructing estimators that are robust to contaminated or heavy-tailed data. Recall that the MAD estimator in a direction $v \in \mathbb{R}^d$, based on a sample $(X_1, \dots, X_n)$ of vectors in $\mathbb{R}^d$, is defined by,

$$\mathrm{MAD}(v) = \mathrm{Med}(|\langle X_i, v\rangle - \mathrm{Med}(\langle X_j, v\rangle)|).$$

In addition, the SDO in a direction $\mu \in \mathbb{R}^d$ writes,

$$\mathrm{SDO}(\mu) = \sup_{v \in \mathbb{R}^d} \frac{|\langle \mu, v\rangle - \mathrm{Med}(\langle X_j, v\rangle)|}{\mathrm{MAD}(v)}$$

and the so-called "Stahel-Donoho median" $\hat{\mu}^{SD}$ is then givn by,

$$\hat{\mu}^{SD} \in \arg\min_{\mu \in \mathbb{R}^d} \mathrm{SDO}(\mu).$$

To define MOM versions of the latter quantities, it suffices to consider a partition $\{B_1, \dots, B_K\}$ of the set of indices $\{1, \dots, n\}$ - with each block B_k having the same number of elements, up to one element if necessary - and to apply $\mathrm{MAD}, \mathrm{SDO}$ and $\hat{\mu}^{SD}$ to the collection of vectors $(\bar{X}_1, \dots, \bar{X}_K)$ rather than the initial sample, where we set $\bar{X}_k = 1/\mathrm{Card}(B_k) \sum_{i \in B_k} X_i$.

The first non-asymptotic bounds of the "Stahel-Donoho median" and its MOM version are established by Depersin and Lecué (see Theorems 1, 2, 3, 4 in [16]). These bounds are naturally obtained in the "prediction norm" - that is the $\|\Sigma^{-1/2} \cdot\|_2$-norm where Σ stands for the covariance matrix of the non-corrupted data vectors - and not only in the quadratic norm, due to almost isometric properties of MAD and SDO with respect to the prediction norm. It is also shown that the MOM version of MAD can be used to construct an estimator of the covariance matrix under a single assumption of finite second moment or of a scaling parameter if a second moment does not exist.

In the first section of this article, Timothée Mathieu presents the results obtained during his PhD thesis [33], pertaining to robustness properties of M-estimators of location parameters. Hampel's continuity approach to robustness is revisited using specifically designed Wasserstein-type distance and new stability results for M-estimators are obtained.

Claire Brécheteau has written the second section, that deals with robust estimation of the distance function to a compact set, a theme arising from topological inference and having applications in robust clustering. The method is based on a carefully chosen union of balls or ellipsoids recovering data points, a detailed algorithm is provided and rates of convergence are obtained. The approach is detailed in [4, 6], where the second article is written in collaboration with Clément Levrard.

The last section is written by Edouard Genetay. It presents the methodology and some of the results and experiments obtained in [8], an article written in collaboration with Camille Brunet-Saumard and Adrien Saumard. A robust quantization algorithm is proposed, using in parallel Lloyd iterations on blocks of data that are generated according to a bootstrap sampling process. This algorithm comes with a robust initialization. A theoretical guarantee in terms of a probabilistic breakdown point is presented and some experimental results for the initialization show evidence of robustness to the presence of a small enough proportion of outliers.

1. Robust study of consistent M-estimators via an optimal transport distance

1.1. Location estimators and corrupted distributions

We study geometric M-estimators computed as minimizers of an empirical loss: let $X_1, \dots, X_n$ be an i.i.d sample in a Hilbert space $\mathcal{H}$ and define the statistic

$$T(\widehat{P}_n) \in \arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \rho(\|X_i - \theta\|) \tag{1}$$

where ρ is some convex loss function (convex, even, positive and with $\rho(0) = 0$), $\widehat{P}_n$ is the empirical distribution of $X_1, \dots, X_n$ and $\|\cdot\|$ is the norm associated to the inner product $\langle \cdot, \cdot \rangle$ in the Hilbert space $\mathcal{H}$. The goal is to estimate some location parameter of the law P that spawned the data. Typically we want to estimate the mean. One of our purposes is to study the effect of the choice of ρ on the robustness of $T(\widehat{P}_n)$, in particular the asymptotic properties of $T(\widehat{P}_n)$ for different choices of ρ. Our second purpose is to define and use a family of distances between distributions using the theory of optimal transport. These distances between distributions are well adapted to working with robust estimators of the mean.

We consider robustness because we aim at dealing with a dataset that is not ideal. Instead, the data may be corrupted by some anomalous data called outliers (one can think of outliers as "data [...] which do not fit the pattern suggested by the majority of the data" [22]) and the inliers (points that are not outliers) may come from a heavy-tailed distribution.

Let us define a model of corruption that can be used to represent such a corrupted dataset.

Definition 1.1 (Huber's contamination model)**.** Let $\varepsilon \in [0, 1/2)$ be the corruption proportion, let P be a probability distribution with at least two finite moments and H an arbitrary probability distribution. We consider $X_1, \dots, X_n$ drawn i.i.d. from $(1 - \varepsilon)P + \varepsilon H$.

P is called the law of "inliers" and H the law of "outliers".

Said differently, in Huber's contamination model, we draw the data from P the inlier distribution with probability $(1 - \varepsilon)$ and otherwise we draw according to the outlier distribution. This models a naturally-occuring error that happens with small probability and is independent of the inliers. We will consider other corruption neighborhoods in the sequel but Huber's contamination model will play the role of benchmark.

1.2. M-estimators and robustness as a continuity

M-estimators have traditionally been studied in two contexts, the parametric context and the robustness context. In a parametric context, M-estimators are dictated by the model and ρ is fixed to the negative log-likelihood. The associated M-estimator is efficient at the hypothesized model but it can be highly sensitive to a small proportion of outlier data. On the other hand, in the robustness context, ρ is supposed to be Lipshitz and the associated M-estimator is then not sensitive to a small portion of outliers but it is not clear what is the most efficient M-estimator in a given context. The choice of the optimal loss function ρ for a given problem is hard in general except for some simple specific neighborhoods of parametric model. Indeed, in Huber contamination model with a fixed (and known) inlier probability P, a very ingenious analysis by Huber [23] gives us a minimax M-estimator, but for more general neighborhoods and when the model is not parametric, we don't know which ρ is better. The goal of this article is to give some pointers on how to choose ρ for more general models than Huber's contamination model.

The definition of robustness we use here is related to Hampel's definition of robustness in [20]. Let $T(P)$ be the asymptotic value of $T(\widehat{P}_n)$ when n goes to infinity, if there is a small change in the probability then it should not cause a large change in the value of $T(P)$. And if this is the case, $T(P)$ will be said to be robust. More formally, let d be a distance between probabilities (Hampel used Levy, Prokhorov or Kolmogorov distance

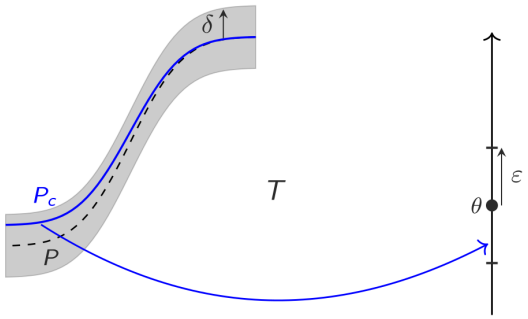


FIGURE 1. Illustration of robustness as a continuity of a functional T in dimension 1 for Kolmogorov distance

for example) then T is robust at P if we have

$$\|T(P) - T(Q)\| \xrightarrow[d(P,Q)\to 0]{} 0.$$

This is a continuity property for the functional T and robustness theory shows that this type of robust property of T implies robustness properties for the associated M-estimator $T(\widehat{P}_n)$, see [21, 24], see Figure 1 for an illustration. Continuity is linked to the notion of neighborhood and after defining robustness as a continuity, it is natural to define the notion of corrupted neighborhood. An ε-corruption neighborhood of P is defined as all the distributions Q such that $d(P,Q) \leq \varepsilon$ where d is some distance between distributions. For instance if $d = TV$ is the total variation distance, then Huber's contamination model of P is included in the ε-corrupted neighborhood of P because $TV((1-\varepsilon)P + \varepsilon H, P) \leq \varepsilon$ for any H. On the other hand, there is no distance d for which Huber's contamination neighborhood is equal to the corruption neighborhood for d.

To study the outlier-resistance properties of $T(\widehat{P}_n)$, let us define $\psi = \rho'$ (when it exists) so that $T(\widehat{P}_n)$ may be defined alternatively by

$$\sum_{i=1}^{n} \frac{X_i - T(\widehat{P}_n)}{\|X_i - T(\widehat{P}_n)\|} \psi(\|X_i - T(\widehat{P}_n)\|) = 0.$$

This equation is true almost surely as a consequence of the derivability of the norm away from 0 in an Hilbert space, and taking by convention $\frac{x}{\|x\|}\psi(\|x\|) = 0$ for $x = 0$.

We will see that the properties of $T(\widehat{P}_n)$ are easy to state using properties of ψ. ψ is supposed to be non-decreasing but unlike many robust statistics articles, we don't suppose that ψ is bounded and instead we will see that different M-estimators constructed with unbounded ψ function will verify a weaker definition of robustness. For example,

- If $\psi(x) = x$, we get $T(\widehat{P}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$,
- If $\psi(x) = 1$, then $T(\widehat{P}_n)$ is the empirical geometric median.

The empirical median can be seen as the most robust estimator in some sense (see [24, Section 4.2]) but it is not very efficient to estimate the mean. We will use other ψ functions to compute M-estimators that will be more robust than the empirical mean and more efficient than the empirical median to estimate the mean. Some examples of trade-off between the mean and the median are obtain for ψ function such as

Huber's score function: Let $\beta > 0$. For all $x \geq 0$, let

$$\psi_H(x) = x\, \mathbb{1}\{x \leq \beta\} + \beta\, \mathbb{1}\{x > \beta\}. \tag{2}$$

In dimension 1, the M-estimator constructed from this score function is called Huber's estimator [23].

Catoni's score function: Let $\beta > 0$. For all $x \geq 0$, let

$$\psi_C(x) = \beta \log\left(1 + \frac{x}{\beta} + \frac{1}{2}\left(\frac{x}{\beta}\right)^2\right). \tag{3}$$

The associated M-estimator is one of the estimators considered by Catoni in [11]. We call the resulting M-estimator Catoni's estimator.

Polynomial score function: Let $p \in \mathbb{N}^*$, $\beta > 0$. For all $x \geq 0$, let

$$\psi_P(x) = \frac{x}{1 + \left(\frac{x}{\beta}\right)^{1-1/p}}. \tag{4}$$

We call Polynomial estimator the M-estimator obtained using this score function.

In the rest of the section, we aim at showing some continuity and asymptotic results for geometric M-estimators, extending the classical robustness theory using optimal transport distances. Our main contributions are given in Theorem 1.3 and Theorem 1.4 where we prove the robustness of some family of distances between distributions and some consequence of the continuity of M-estimators with respect to such distances.

1.3. **A family of Wasserstein distances adapted to robust statistics**

An important new tool of independent interest that we use here is a family of Wasserstein-type robust distances defined by

$$W_\psi(P, Q) = \sup_{h \preceq \psi} \left\{ \int h(x) dP(x) - \int h(x) dQ(x) \right\},$$

where $h \preceq \psi$ if and only if for any $x, y \in \mathcal{H}$, $h(x) - h(y) \leq \psi(\|x - y\|)$. Two special cases of W_ψ are the total variation distance (for $\psi(x) = \text{sign}(x)$ with $\psi(0) = 0$) and the Wasserstein-1 distance (for $\psi(x) = x$). The distance W_ψ can also be efficiently computed using for instance techniques from [36]. From the theory of optimal transport we can prove the following theorem.

Theorem 1.2. *[From Theorem 6.9 in [39]] Suppose that ψ is increasing on $\mathbb{R}_+$, concave, $\psi(0) = 0$ and ψ is not constant equal to 0. Then, W_ψ metrizes the weak convergence in $\mathcal{P}_\psi = \{P : \mathbb{E}_P[\psi(\|X\|)] < \infty\}$. In other words, if $(P_k)_{k\in\mathbb{N}}$ is a sequence of probability measures in $\mathcal{P}_\psi$ and $P \in \mathcal{P}$ where $\mathcal{P}$ is the set of probability distributions on $\mathcal{H}$. Then the following statements are equivalent:*

$$P_k \xrightarrow[k\to\infty]{law} P \quad \textit{and} \quad W_\psi(P_k, P) \xrightarrow[k\to\infty]{} 0.$$

In particular, from Glivenko-Cantelli's theorem, $W_\psi(\widehat{P}_n, P) \xrightarrow[n\to\infty]{} 0$ *a.s.*. To illustrate the difference between W_ψ and the Wasserstein-1 distance, let us consider the transport problem of Figure 2. In this pair of distribution, clearly the red distribution contains some outliers. We represent in Figure 3 the optimal transport map for the Wasserstein-1 distance and the optimal transport map for W_ψ distance where ψ is Huber's score function for $\beta = 1$. There is a lot less disruption due to outliers in the transport map with W_ψ than in the transport map with Wasserstein-1 distance.

We prove a theoretical guarantee linked to the illustration in Figures 2 and 3 expressed in the following stability result for W_ψ.

Theorem 1.3 (Theorem 21 in [33])**.** *Suppose that ψ is increasing, concave, $\psi(0) = 0$ and ψ is not constant equal to 0. Then for $H : [0, 1] \to \mathcal{P}_\psi$ if*

$$t\mathbb{E}_{H(t)}[\psi(\|X\|)] \xrightarrow[t\to 0]{} 0$$

then,

$$W_\psi\left((1-t)P + tH(t), P\right) \xrightarrow[t\to 0]{} 0.$$

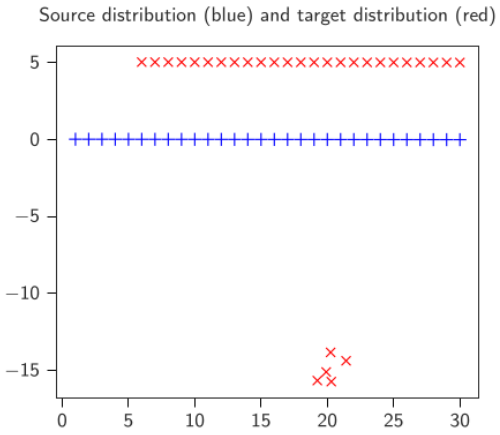


FIGURE 2. Corrupted sources distributions in an optimal transport problem.

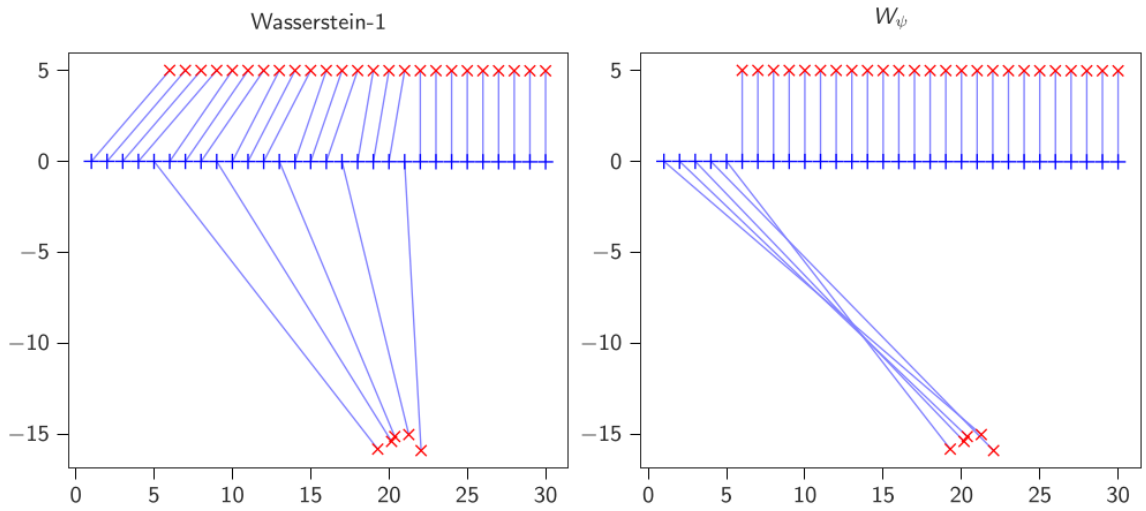


FIGURE 3. Transport map for Wasserstein-1 distance and for W_ψ distance.

Theorem 1.3 gives the condition that an outlier distribution H must verify in order to have $(1-t)\,P+tH(t)$ that converges to P for W_ψ. It is proved in [33] and we feel that this new result justify the use of W_ψ in robust statistics. This is a property of stability of the W_ψ neighborhoods which shows that for some score function ψ, W_ψ neighborhoods are more general than Huber's contamination neighborhoods. A consequence of Theorem 1.3 is that any estimator that is continuous with respect to W_ψ will show some robustness properties, we make that clear in the next section.

1.4. **Application to the robustness of M-estimators**

Wasserstein-type distances are better suited than Kolmogorov or total-variation distances for studying of empirical distribution in a robust context and we show this through consistency results in the corrupted setting.

One of the reason why Wasserstein distances are better suited for empirical work than total variation distance is that the total variation does not take the ambient distance into account: for any $x \neq y$, we have $TV(\delta_x, \delta_y) = 1$ whereas $W_\psi(\delta_x, \delta_y) = \psi(\|x - y\|)$ where δ_x denotes the Dirac mass in x. Hence W_ψ will be better suited for working with empirical densities. We also consider unbounded score functions ψ. Unbounded score functions have been studied extensively in parametric setting (see [37]) and more recently using concentration inequalities in the non-parametric setting [11, 12, 34]. An important first result is that the functional T constructed from some score function ψ is continuous in $\mathcal{P}_\psi$ with respect to the distance W_ψ.

Using Theorem 1.3 and the continuity of T with respect to W_ψ (Theorem 22 in [33]), we can prove the following result.

Theorem 1.4 (from Corollary 8 in [33])**.** *If $X'_1, \ldots, X'_n$ are i.i.d data with law P and $X_1, \ldots, X_n$ is a copy of $X'_1, \ldots, X'_n$ except for one point that has been contaminated and has value M_n, i.e. $X_i = X'_i$ for all $i \neq i_0$ and $X_{i_0} = M_n \neq X'_{i_0}$. If $\psi(\|M_n\|)/n \xrightarrow[n\to\infty]{} 0$, then*

$$\lim_{n\to\infty} T\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\right) = \lim_{n\to\infty} T\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{X'_i}\right).$$

This result shows that if we have mild assumptions on the outlier (for instance if ψ is logarithmic at infinity, M_n must be $o(e^n)$) then the M-estimator ignores the outlier asymptotically. To show this result, we prove an extension of Hampel's Theorem [24, Section 2.6] to the case of unbounded score function ψ using heavily the properties of the family of distribution W_ψ.

2. Dataset approximation with unions of ellipsoids and clustering

Let P be a probability distribution supported on a compact subset $\mathcal{K}$ of $\mathbb{R}^d$, equipped with the Euclidean norm $\|\cdot\|$. Let $\mathbb{X} = \{X_1, X_2, \ldots, X_n\}$ be an i.i.d n-sample from a noisy version Q of P. In this section, we introduce methods to build a family of k balls or ellipsoids to represent the data $\mathbb{X}$. These methods aim at recovering the compact set $\mathcal{K}$, or more precisely, the distance function $\mathrm{d}_\mathcal{K} : x \mapsto \inf_{y\in\mathcal{K}} \|x - y\|$ to this compact set. Our estimators of $\mathrm{d}_\mathcal{K}$ may be used to cluster the data $\mathbb{X}$, accordingly to their intrinsic shape. Therefore, they are robust alternatives to spectral clustering methods [40] or to Topological Mode Analysis Tool (ToMATo algorithm) [14].

The construction of the balls and ellipsoids is based on a risk minimisation task. The centers of the balls $c_1, c_2, \ldots, c_k \in \mathbb{R}^d$ and, for the directions of the ellipsoids, the covariance matrices $\Sigma_1, \Sigma_2, \ldots, \Sigma_k \in \mathrm{Cov}_d$ are both obtained by minimising some k-means - type empirical risk [29]

$$R_n : (\mathbf{t}, \mathbb{X}) \in (\mathbb{R}^d \times \mathrm{Cov}_d)^k \times \mathbb{R}^{d\times n} \mapsto \frac{1}{n}\sum_{i=1}^{n} \gamma(\mathbf{t}, X_i),$$

based on the criterion

$$\gamma : (\mathbf{t} = (t_1, t_2, \ldots, t_k), x) \in (\mathbb{R}^d \times \mathrm{Cov}_d)^k \times \mathbb{R}^d \mapsto \min_{l\in[\![1,k]\!]} \delta(t_l, x). \tag{5}$$

Here, $\delta(t_l, x)$ defines a divergence between a pair $t_l = (c_l, \Sigma_l) \in \mathbb{R}^d \times \mathrm{Cov}_d$ and a point $x \in \mathbb{R}^d$, and Cov_d denotes the set of symmetric positive semi-definite real $d \times d$-matrices. Replacing δ with the squared Euclidean distance in γ leads to the k-means criterion [29]. The estimators of $\mathrm{d}_\mathcal{K}$ that we introduce in this section are of the form $\sqrt{\gamma(\hat{\mathbf{t}}, \cdot)}$, with $\hat{\mathbf{t}}$, a minimiser of the empirical risk R_n. The compact set $\mathcal{K}$ may be approximated with sublevel sets of $\gamma(\hat{\mathbf{t}}, \cdot)$, that is, with unions of k balls or k ellipsoids. The performance of our distance estimators is assessed through the $L_1(P)$-distance between the squared estimator $\gamma(\hat{\mathbf{t}}, \cdot)$ and $\mathrm{d}^2_\mathcal{K}$. We derive non-parametric rates of convergence.

From the criterion $\gamma(\hat{\mathbf{t}}, \cdot)$, we derive two types of clustering methods for $\mathbb{X}$. A first one is based on the space decomposition induced by the criterion, when k is small and corresponds to the expected number of clusters. A second one is based on a hierarchical clustering-type procedure based on the sublevel sets of $\gamma(\hat{\mathbf{t}}, \cdot)$. For this scheme, k may be much larger than the number of clusters, since it represents the number of balls or ellipsoids used to fit the data possibly sampled from a complex geometric structure.

The methods are proven robust. The estimator $\gamma(\hat{\mathbf{t}}, \cdot)$ is robust to additive noise. The continuous version of the estimator, $\gamma(\mathbf{t}_Q^*, \cdot)$, based on the noisy version Q of P, with $\mathbf{t}_Q^*$, a minimiser of the continuous criterion $\mathbf{t} \mapsto Q\gamma(\mathbf{t}, \cdot)$, is robust to Wasserstein noise and to additive noise. By robust, we mean that additive noise does not substantially modifies the $L_1(P)$- distance of the squared distance estimator to $\mathrm{d}^2_{\mathcal{K}}$. Moreover, by trimming our criteria [15], our methods can be used as procedures for outliers detection. Clustering procedures, based on the trimmed criterion with a Bregman divergence δ, do have theoretical guaranties in terms of breakdown point, [5].

The work exposed in this section has been published in two papers [4, 6]. In the following four parts, we define our divergence δ for the approximation of $\mathcal{K}$ with balls, for the approximation of $\mathcal{K}$ with ellipsoids, we give the algorithm for our trimmed distance estimators and finally we explain how to make use of these distance estimators for data clustering.

2.1. **Approximating data with a family of k balls**

As aforementionned, the k-means criterion is obtained by replacing $\delta : (c, x) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \|c - x\|^2$, in the expression of γ, in (5). Note that δ does not have any covariance matrix as an argument. Therefore, in the resulting expression, the k-uplets $\mathbf{t} = ((c_1, \Sigma_1), (c_2, \Sigma_2) \ldots, (c_k, \Sigma_k)) \in \left(\mathbb{R}^d \times \mathrm{Cov}_d\right)^k$ are identified to k-uplets $\mathbf{t} = (c_1, c_2, \ldots, c_k) \in \left(\mathbb{R}^d\right)^k$. This criterion lacks of robustness with respect to outliers and additive noise in the data. Indeed, a single element t of $\mathbf{t}$, far from $\mathcal{K}$, makes the L_∞-distance to $\mathrm{d}_{\mathcal{K}}$ large, since $\|\mathrm{d}_{\mathbf{t}} - \mathrm{d}_{\mathcal{K}}\|_\infty \geq \inf_{x \in \mathcal{K}} \|x - t\|$. Therefore, we introduce a smoothed version of this criterion using a parameter $q \in [\![1, n]\!]$, that corresponds to a number of nearest-neighbours. Equivalently, we set $h = \frac{q}{n} \in (0, 1]$, the proportion of neighbours to consider in order to smooth the distance estimator.

For $c \in \mathbb{R}^d$, consider $\tilde{\mathbb{X}}(c, h)$ the subset of $\mathbb{X}$ composed with the $q = nh$ nearest neighbours of c in $\mathbb{X}$, for the Euclidean norm. Set $m_{c,h} = \frac{1}{q} \sum_{\tilde{x} \in \tilde{\mathbb{X}}(c,h)} \tilde{x}$, the mean of these neighbours, and $v_{c,h} = \frac{1}{q} \sum_{\tilde{x} \in \tilde{\mathbb{X}}(c,h)} \|m_{c,h} - \tilde{x}\|^2$ their inertia or variance. We consider the divergence $\delta_{\mathbb{X},h}$, defined by :

$$\delta_{\mathbb{X},h} : (c, x) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \|x - m_{c,h}\|^2 + v_{c,h} = \frac{1}{q} \sum_{\tilde{x} \in \tilde{\mathbb{X}}(c,h)} \|x - \tilde{x}\|^2.$$

The function $x \mapsto \sqrt{\delta_{\mathbb{X},h}(x, x)} = \sqrt{\inf_{c \in \mathbb{R}^d} \delta_{\mathbb{X},h}(c, x)}$ coincides with the empirical distance-to-measure function $\mathrm{d}_{\mathbb{X},h}$ [13], a robust approximation of the distance function $\mathrm{d}_{\mathcal{K}}$, widely used in the robust geometric data analysis area.

The continuous analogue of $\delta_{\mathbb{X},h}$ is defined for the probability distribution P by $\delta_{P,h} : (c, x) \mapsto \tilde{P}_{c,h} \|x - \cdot\|^2$. It corresponds to the expectation of the squared Euclidean distance of x to a random variable X of distribution $\tilde{P}_{c,h}$, the restriction of P to the Euclidean ball $B_{P,c,h}$ centered at c with P-mass $P(B_{P,c,h}) = h$. The distance to the measure P is defined by $\mathrm{d}_{P,h} : x \in \mathbb{R}^d \mapsto \sqrt{\delta_{P,h}(x, x)} = \sqrt{\inf_{c \in \mathbb{R}^d} \delta_{P,h}(c, x)}$ [13].

The k-power distance-to-measure function is defined by

$$\mathrm{d}_{P,h,k} : x \mapsto \sqrt{\gamma_{P,h}(\mathbf{t}^*, x)} \coloneqq \sqrt{\min_{l \in [\![1,k]\!]} \delta_{P,h}(t_l^*, x)}, \tag{6}$$

with

$$\mathbf{t}^* \in \arg\min_{\mathbf{t}\in(\mathbb{R}^d)^k} \left\{P\gamma_{P,h}(\mathbf{t},\cdot)\right\}. \tag{7}$$

The empirical k-power distance-to-measure function is defined by

$$\mathrm{d}_{\mathbb{X},h,k} : x \mapsto \sqrt{\gamma_{\mathbb{X},h}(\hat{\mathbf{t}},x)} \coloneqq \sqrt{\min_{l\in[\![1,k]\!]} \delta_{\mathbb{X},h}(\hat{t}_l,x)}, \tag{8}$$

with

$$\hat{\mathbf{t}} \in \arg\min_{\mathbf{t}\in(\mathbb{R}^d)^k} \left\{ \frac{1}{n}\sum_{i=1}^{n} \gamma_{\mathbb{X},h}(\mathbf{t},X_i) \right\}. \tag{9}$$

Note that for every $x \in \mathbb{R}^d$, $\mathrm{d}_{\mathbb{X},h,k}(x) \geq \mathrm{d}_{\mathbb{X},h}(x)$ and $\mathrm{d}_{P,h,k}(x) \geq \mathrm{d}_{P,h}(x)$.

In order to assess the performance of our estimator $\mathrm{d}_{\mathbb{X},h,k}$ of $\mathrm{d}_{\mathcal{K}}$, we compare $\mathrm{d}_{\mathbb{X},h,k}$ to the distance-to-measure function $\mathrm{d}_{P,h}$, which is proven to be close to $\mathrm{d}_{\mathcal{K}}$ for the infinity norm, under regularity assumptions [13].

The performance is measured in terms of the $L_1(P)$-norm between the squares of the two distance functions. To this aim, we make use of the following bias-variance decomposition :

$$P\left|\mathrm{d}^2_{\mathbb{X},h,k}(\cdot) - \mathrm{d}^2_{P,h}(\cdot)\right| \leq P\left|\mathrm{d}^2_{\mathbb{X},h,k}(\cdot) - \mathrm{d}^2_{Q,h,k}(\cdot)\right| + P\left|\mathrm{d}^2_{Q,h,k}(\cdot) - \mathrm{d}^2_{P,h}(\cdot)\right|.$$

The bias term depends on the regularity of the distribution P and of the compact set $\mathcal{K}$.

Theorem 2.1 (Bias term [6, Corollary 16, Proposition 17])**.** *If* $\mathrm{supp}(P) = \mathcal{K} \subset B(0,K)$ *for some* $K > 0$*, and if* Q *has a finite first moment, then,*

$$P\left|\mathrm{d}^2_{Q,h,k}(\cdot) - \mathrm{d}^2_{P,h}(\cdot)\right| \leq 3\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\|_{\infty,B(0,K)} + P\left(\mathrm{d}^2_{P,h,k}(.) - \mathrm{d}^2_{P,h}(.)\right) + 4W_1(P,Q)\sup_{c\in\mathbb{R}^d}\|m_{P,c,h}\|, \tag{10}$$

with $m_{P,c,h}$*, the expectation of* $\tilde{P}_{c,h}$*, the restriction of* P *to the ball centered at* c *with* P*-mass* h*.*

Moreover, if $\mathcal{K}$ *is a compact* d'*-dimensional* $\mathcal{C}^2$*-submanifold, if* P *has a density* $0 < f_{min} \leq f \leq f_{max}$ *with respect to the volume measure on* $\mathcal{K}$ *and if it satisfies, for every* $x \in \mathcal{K}$ *and* $r > 0$*, the following inequality*

$$P(B(x,r)) \geq c f_{min} r^{d'} \wedge 1,$$

then, for $k \geq c_{N,f_{min}}$ *and* $h \leq C_{N,f_{min}}$*, we have*

$$0 \leq P(d^2_{P,h,k} - d^2_{P,h}) \leq C_{N,f_{min},f_{max}} k^{-\frac{2}{d'}}, \tag{11}$$

where $c_{N,f_{min}}$*,* $C_{N,f_{min}}$ *and* $C_{N,f_{min},f_{max}}$ *are three constants, depending on* N*,* f_{min} *and* f_{max}*.*

Note that the first term $\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\|_{\infty,B(0,K)}$ in the right hand side of (10) may be upper bounded by the L_2-Wasserstein distance $W_2(P,Q)$, up to a constant term, due to the stability of the distance-to-measure function with respect to the L_2-Wasserstein distance [13].

A bound for $\|\mathrm{d}_{Q,h,k} - \mathrm{d}_{\mathcal{K}}\|_\infty$ is available in [6, Proposition 18]. This bound is stated in terms of the Wasserstein distance between P and Q, and of the smoothing parameter h. In particular, for h large enough, the distance estimator may be close to $\mathrm{d}_{\mathcal{K}}$, unlike the distance estimator based on k-means.

Parametric rates of convergence are obtained for the variance term, under additive noise assumption.

Theorem 2.2 (Variance term [6, Theorem 19])**.** *Let* P *be supported on* $\mathcal{K} \subset B(0,K)$*. Assume that we observe* $\mathbb{X} = \{X_1,\ldots,X_n\}$ *such that* $X_i = Y_i + Z_i$*, where the* Y_i*'s and* Z_i*'s are all independent,* Y_i *is sampled from* P *and* Z_i *is sub-Gaussian with variance* σ^2*, with* $\sigma \leq K$*. That is,* $\|Z_i\| \leq t$ *with probability larger than* $1 - \exp\left(-\frac{t^2}{2\sigma^2}\right)$*, for every* $t \geq \sigma$*.*

Let $p > 0$, with probability larger than $1 - 10n^{-p}$, we have

$$\left|P\left(\mathrm{d}^2_{Q_n,h,k}(.) - \mathrm{d}^2_{Q,h,k}(.)\right)\right| \leq C\sqrt{k\log(k)d}\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{K\sigma}{\sqrt{h}}. \tag{12}$$

Consequently, under the additive noise assumption, optimising in k the sum of the two upper bounds from (10), (11) and (12) consists in optimising in k the quantity

$$\frac{C\sqrt{k\log(k)}K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C_{N,f_{min},f_{max}}k^{-\frac{2}{d'}}.$$

Therefore, an optimal choice would consist in taking k of order $n^{\frac{d'}{d'+4}}$, where d' is the intrinsic dimensionality of $\mathcal{K}$. Therefore, there is no need of taking k of order n, the sample size, to have the best distance to $\mathcal{K}$ approximation, from a noisy sample.

2.2. **Approximating data with a family of k ellipsoids**

If $\mathcal{K}$ is embedded into a d'-dimensional submanifold with $d' < d$, the ambiant dimension, summarizing data with ellipsoids instead of balls may be more appropriate: notably, to have a sparser descriptor of the data, that is, a descriptor with a smaller number of ellipsoids.

For $h \in (0, 1]$, the function $c \mapsto \delta_{P,h}(c, c) = \tilde{P}_{c,h}\|c - \cdot\|^2$ actually corresponds, up to some absolute constants, to the negative h-trimmed log-likelihood [38] for the normal isotropic model $(\mathcal{N}(c, I_d))_{c\in\mathbb{R}^d}$. Indeed, the ball $B_{P,c,h}$ defined in the previous section is the upper-level set of the normal isotropic distribution $\mathcal{N}(c, I_d)$, with P-mass h, where I_d stands for the identity matrix on $\mathbb{R}^d$.

By enriching the normal distributions model with covariance matrices Σ in Cov_d, we extend the function $\delta_{P,h}$ to the whole space $\mathbb{R}^d \times \mathrm{Cov}_d \times \mathbb{R}^d$. It is defined by $\delta_{P,h} : (t = (c, \Sigma), x) \mapsto \tilde{P}_{(c,\Sigma),h}\left(\|x - \cdot\|^2_{\Sigma^{-1}} + \log(\det(\Sigma))\right)$, where, for $x \in \mathbb{R}^d$, $\|x\|_{\Sigma^{-1}} = \sqrt{x^T\Sigma^{-1}x}$ denotes the Σ-Mahalanobis norm of x. In particular, $t = (c, \Sigma) \mapsto \delta_{P,h}(t, c)$ coincides with the h-trimmed negative loglikelihood, for $(\mathcal{N}(c, \Sigma))_{(c,\Sigma)\in\mathbb{R}^d\times\mathrm{Cov}_d}$, the normal anisotropic model. Note that $\tilde{P}_{(c,\Sigma),h}$ is the restriction of P to the upper-level set of $\mathcal{N}(c, \Sigma)$ of P-mass h. This is an ellipsoid centered at c, directed by the matrix Σ.

Let $\tilde{\mathbb{X}}((c, \Sigma), h)$ be the subset of $\mathbb{X}$ composed with the q $\|\cdot\|_{\Sigma^{-1}}$-nearest neighbours of c in $\mathbb{X}$. The discrete version of the criterion is defined from $\tilde{\mathbb{X}}((c, \Sigma), h)$, by :

$$\delta_{\mathbb{X},h} : ((c, \Sigma), x) \in \mathbb{R}^d\times\mathrm{Cov}_d\times\mathbb{R}^d \mapsto \|x - m_{(c,\Sigma),h}\|^2_{\Sigma^{-1}} + v_{(c,\Sigma),h} + \log(\det(\Sigma)) = \frac{1}{q}\sum_{\tilde{x}\in\tilde{\mathbb{X}}((c,\Sigma),h)}\|x - \tilde{x}\|^2_{\Sigma^{-1}} + \log(\det(\Sigma)),$$

with $m_{(c,\Sigma),h} = \frac{1}{q}\sum_{\tilde{x}\in\tilde{\mathbb{X}}((c,\Sigma),h)}\tilde{x}$, and $v_{(c,\Sigma),h} = \frac{1}{q}\sum_{\tilde{x}\in\tilde{\mathbb{X}}((c,\Sigma),h)}\|m_{(c,\Sigma),h} - \tilde{x}\|^2_{\Sigma^{-1}}$.

The k-power likelihood-to-measure function and the empirical k-power likelihood-to-measure function are estimators of $\mathrm{d}_{\mathcal{K}}$ defined from $\delta_{\mathbb{X},h}$ and $\delta_{P,h}$, by the equations (6), (7), (8) and (9).

2.3. **Algorithms**

An adaptation of the Lloyd's algorithm for k-means is possible and boils down to compute a local minimiser of the empirical risk R_n, for both the empirical k-power distance-to-measure and the empirical loglikelihood-to-measure. An adaptation of the trimmed k-means algorithm [15] is also possible.

For a given $\alpha \in [0, 1]$, the α-trimmed empirical risk is defined by :

$$\tilde{R}_{n,\alpha} : (\mathbf{t}, \mathbb{X}) \in (\mathbb{R}^d \times \mathrm{Cov}_d)^k \times \mathbb{R}^{d\times n} \mapsto \inf_{\tilde{\mathbb{X}}\subset\mathbb{X},|\tilde{\mathbb{X}}|=\lfloor\alpha n\rfloor} R_n(\mathbf{t}, \tilde{\mathbb{X}}).$$

Here, $|\tilde{\mathbb{X}}|$ denotes the cardinality of $\tilde{\mathbb{X}}$ and $\lfloor \alpha n \rfloor$ denotes the floor of αn. Minimising the trimmed risk $\tilde{R}_\alpha$ consists in selecting the subset of $\lfloor \alpha n \rfloor$ points of $\mathbb{X}$ for which the optimal risk $\inf_{\mathbf{t}} R(\mathbf{t}, \tilde{\mathbb{X}})$ is minimal. To a certain extent, it corresponds to the subset of the data that can be best approximated with a family of k points, according to our given criterion γ.

For conciseness, we describe the algorithm to compute a local minimiser of the trimmed criterion, for the k-power empirical likelihood-to-measure. The local minimiser for the k-power distance-to-measure function can be derived from this algorithm by replacing all covariance matrices with I_d.

Algorithm 2.3. *Local minimisation of the trimmed risk* $\tilde{R}_{n,\alpha=\frac{a}{n}}$

Input $\mathbb{X}$ *an* n*-sample from* P *;* $q, k, a \in [\![1, n]\!]$.
Sample c_1, c_2,$\dots$ c_k *from* $\mathbb{X}$ *without replacement. Set* $\Sigma_i = I_d$ *for* i *in* $[\![1, k]\!]$.
while *the* $t_i = (c_i, \Sigma_i)$*s vary* ***do***
  for i *in* $[\![1, k]\!]$ ***do***
    Set $\mathcal{C}(t_i) = \{\}$
  end for
  for j *in* $[\![1, n]\!]$ ***do***
    Add X_j *to the cell* $\mathcal{C}(t_i)$ *satisfying*
    $\delta_{\mathbb{X},h}(t_i, X_j) \leq \delta_{\mathbb{X},h}(t_l, X_j) \, \forall l \neq i$.
    Set $t(X) = (c(X), \Sigma(X)) = (c_i, \Sigma_i)$.
  end for
  Sort $(\gamma(X) = \delta_{\mathbb{X},h}(t(X), X))$ *for* $X \in \mathbb{X}$
  Remove the $n - a$ *points* X *associated with the* $n - a$ *largest values of* $\gamma(X)$*, from their cell* $\mathcal{C}(t(X))$
  for i *in* $[\![1, k]\!]$ ***do***
    $c_i = \frac{1}{|\mathcal{C}(t_i)|} \sum_{X \in \mathcal{C}(t_i)} X$; $\Sigma_i = \Sigma(c_i, \Sigma_i, \mathcal{C}(t_i))$
    $t_i = (c_i, \Sigma_i)$
  end for
end while
Output $(t_1, t_2, \dots, t_k)$.

For every $l, m \in [\![1, d]\!]$ and $(c, \Sigma) \in \mathbb{R}^d \times \mathrm{Cov}_d$, $\Sigma(c, \Sigma, \mathcal{C})_{l,m} = \frac{1}{|\mathcal{C}|} \sum_{X \in \mathcal{C}} \frac{1}{|\tilde{\mathbb{X}}((c,\Sigma),h)|} \sum_{\tilde{X} \in \tilde{\mathbb{X}}((c,\Sigma),h)} (X_{(l)} - \tilde{X}_{(l)})(X_{(m)} - \tilde{X}_{(m)})$, with $X_{(m)}$ and $\tilde{X}_{(m)}$ the m-th coordinates of X and $\tilde{X}$.

2.4. **Data clustering**

Just as k-means, our criterion $\gamma_{\mathbb{X},h}(\hat{\mathbf{t}}, \cdot)$ may be used to cluster the data $\mathbb{X}$. For this purpose, a point $x \in \mathbb{X}$ satisfying $\gamma_{\mathbb{X},h}(\hat{\mathbf{t}}, x) = \delta_{\mathbb{X},h}(\hat{t}_l, x)$ for $l \in [\![1, k]\!]$ is given the label l. Moreover, the $n - a$ points $x \in \mathbb{X}$ with largest value $\gamma_{\mathbb{X},h}(\hat{\mathbf{t}}, x)$ are removed from the sample, since they are considered as outliers. An application of the clustering method on a mixture of anisotropic 2-dimensional normal distributions, with outliers uniformly sampled on a subset of $\mathbb{R}^2$, is given in Figure 4 (left). We take $k = 3$ since there are 3 clusters.

For data sampled on a non-connected submanifold, we may use our distance estimator $\mathrm{d}_{\mathbb{X},h,k}$ whose sublevel sets are unions of ellipsoids. Starting from $\hat{\mathbf{t}} = (\hat{t}_1, \hat{t}_2, \dots, \hat{t}_k)$, we may construct a graph filtration (a family of non-decreasing graphs G_τ, indexed with a parameter $\tau \in \mathbb{R}$). In the graph G_τ, vertices correspond to ellipsoids $\delta_{\mathbb{X},h}(\hat{t}_l, \cdot)^{-1}((-\infty, \tau])$ for $l \in [\![1, k]\!]$ and an array between two vertices is in G_τ when the two ellipsoids have a non-empty common intersection. From this graph filtration, it is possible to apply a hierarchical clustering with the dendrogram derived from the graph filtration, or a persistence-based clustering [4] as in [14]. Both procedures are similar. Data points are then clustered according to the label of their $\gamma_{\mathbb{X},h}(\hat{\mathbf{t}}, \cdot)$-associated center. As for the previous procedure, the $n - a$ $\gamma_{\mathbb{X},h}(\hat{\mathbf{t}}, \cdot)$-furthest points to their center are removed. An application of this method is given in Figure 4 (right).

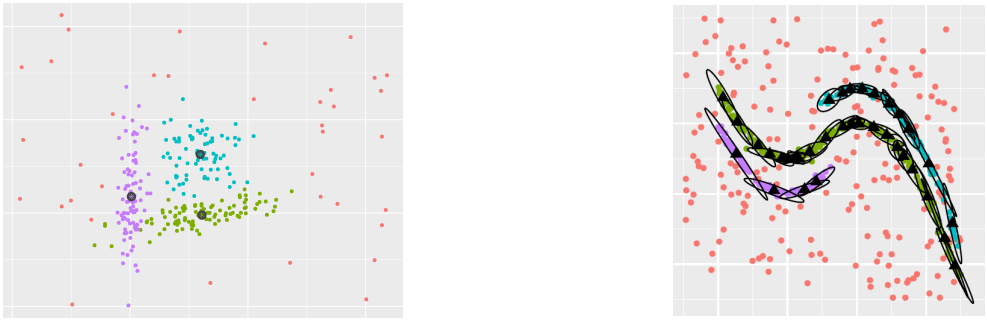


FIGURE 4. Data clustering with a straight use of the criterion $\gamma_{\mathbb{X},h}(\hat{\mathbf{t}},\cdot)$ (left), with a hierarchical type clustering method based on the sublevel sets of $\gamma_{\mathbb{X},h}(\hat{\mathbf{t}},\cdot)$ (right)

2.5. **Conclusion and Opening**

The criteria defined here are useful to cluster noisy data which exhibit an intrinsic geometric shape. Theoretical results for the k-power likelihood-to-measure function are in progress. In order to be compatible with the dimensionality of the data, we also intend to propose a procedure based on a slope heuristic, by considering a subset of the family of covariance matrices, with intrinsic dimensionality smaller than the ambiant dimension. Moreover, the k-power distance-to-measure function's criterion derives from a data-dependent Bregman divergence. Following [5], we expect theoretical guarantees for the trimming procedures.

3. CONSTRUCTION OF A ROBUST CLUSTERING ALGORITHM VIA BOOTSTRAP MOM

3.1. **Introduction**

Classical data mining procedures such as K-means, or EM algorithms for instance, are based on non-robust criteria (least-squares, maximum likelihood) and are thus sensitive to the presence of outliers. A time consuming data pre-processing is thus in general needed before applying such techniques to datasets. To lighten the pre-processing step, robustness is a desirable property for data mining algorithms. Existing approaches for robust K-means consist for instance on considering Huber's losses (K-median) or trimming (trimmed K-means) [3, 19] to name but a few. Trimming comes with practical algorithms and is so far the most theoretically grounded approach, with theoretical garantees such as a breakdown point control. But median-of-means (MOM) strategy has been the object of recent intensive research [17, 26–28, 30–32, 35]. More specifically, Klochkov et al. [25] studied theoretically the MOM strategy applied to K-means on a heavy-tailed data. However, the latter contribution does not give any pratical guidelines.

Although the MOM strategy comes with strong theoretical robust guarantees, it has some flaws from the clustering perspective. Indeed, it consists in making complementary blocks and this is not well suited for clustering because the more blocks, the less chances that all blocks contain one instance of each cluster. We propose therefore to use a variant of the MOM that we call "bootstrap median-of-means" (bMOM), in which the blocks are picked randomly with replacement from the dataset. This procedure can also be viewed as "subragging" [9].

We will describe in Section 3.2 an iterative procedure, derived from K-means. In Section 3.3 we introduce a robust initialisation, using parallel K-means++ initialisations. In Section 3.4 we present a theoretical result in terms of a probabilistic breakdown point and finally, we show some experimental experiments in Section 3.5. This short article highlights some parts of our work available in [8] and in its supplementary material [7].

Input: x_1^n, K, B, n_B with $(n_B > K)$ and $(c_{1,0}, \ldots, c_{K,0})$
Let $(c_{1,0}, \ldots, c_{K,0})$ be the K initial centroids and called reference centroids.
Set $t = 1$.
While $t \leq t_{max}$:

(1) Create B blocks $(y_{1,t}^{(b)}, ..., y_{n_B,t}^{(b)})$ for $b \in \{1, ..., B\}$, according to a random sampling process that at each step selects an observation uniformly over the data x_1^n and independently from the other steps.

(2) **For** all $b \in \{1, \ldots, B\}$:

(a) Assign each data point in the block of index b to its closest reference centroid.

(b) Set $n_{k,t}^{(b)}$ the number of data points in the block b belonging to the cluster k.

(c) **if** $n_{k,t}^{(b)} > 1, \forall k \in \{1, \ldots, K\}$:

(i) for all $k \in \{1, \ldots, K\}$:

$$c_{k,t}^{(b)} \leftarrow 1/n_{k,t}^{(b)} \sum_{l=1}^{n_B} y_{l,t}^{(b)} \mathbf{1}\{y_{l,t}^{(b)} \in \mathcal{C}_{k,t}^{(b)}\}.$$

$$R_t^{(b)} \leftarrow \frac{1}{n_B} \sum_{k=1}^{K} \sum_{l=1}^{n_B} \left\| y_{l,t}^{(b)} - c_{k,t}^{(b)} \right\|^2 \mathbf{1}\{y_{l,t}^{(b)} \in \mathcal{C}_{k,t}^{(b)}\}.$$

(d) **otherwise**

(i) Skip the block.

(e) Get the median block *bmed* such that $R_t^{(bmed)} = \text{med}\left\{R_t^{(b)} : b \in \{1, ..., B\}\right\}$ and $\left(\widehat{c}_{1,t}^{(bmed)}, \ldots, \widehat{c}_{K,t}^{(bmed)}\right)$ the centroids assigned to the median block *bmed* at iteration t becoming the reference centroids.

(f) $t \leftarrow t + 1$.

return: $\bar{\mathbf{c}}^{(bmed)} = \left(\bar{c}_1^{(bmed)}, \ldots, \bar{c}_K^{(bmed)}\right)$ such that $\bar{c}_k^{(bmed)} = \frac{1}{10} \sum_{t=t_{max}-10}^{t_{max}} \widehat{c}_{k,t}^{(bmed)}$ for all $k \in \{1, ..., K\}$ and $\mathcal{P}(\bar{\mathbf{c}}^{(bmed)})$.

FIGURE 5. Algorithm of the iteration phase structure

3.2. **K-bMOM algorithm**

In clustering one often wants to get centroids and one evaluates the achieved performances of these centroids with a criterion. In this case, the bMOM strategy can be applied to a non-robust procedure as follows:

- (bootstrap part) make some blocks with datapoints picked uniformly at random with replacement from the original dataset.
- use your non robust procedure on each block
- compute your performance criterion on each block
- (MOM part) select the centroids of the block with the median performance

We apply this strategy to each iteration of Lloyd's algorithm to get the pseudo-code in the algorithm of Figure 5.

3.3. **A robust initialisation: K-bMOM-kmeans++**

In clustering the initialisation is very important. Indeed iterative procedures are often based on non-convex optimisation schemes that are very sensitive to the initial state. To this end, we provide a robust initialisation procedure in the algorithm of Figure 6. It is based on the same routine as in the previous section, applied to the non-robust K-means++ initialisation [2]:

- (bootstrap part) make some blocks with datapoints picked uniformly at random with replacement from the original dataset
- compute kmeans++ on each block

Input: x_1^n, K, B, n_B with $(n_B > K)$

(1) Create B blocks $(y_{1,0}^{(b)}, ..., y_{n_B,0}^{(b)})$ for $b \in \{1, ..., B\}$, according to a random sampling process that at each step selects an observation uniformly over the data x_1^n and independently from the other steps.

(2) **For** all $b \in \{1, \dots, B\}$:

(a) Proceed to a K-means++ initialization based on the sample $(y_{1,0}^{(b)}, ..., y_{n_B,0}^{(b)})$. This gives the centroids $(c_{1,++}^{(b)}, ..., c_{K,++}^{(b)})$.

(b) Compute the empirical risk $R_{++}^{(b)}$ of the block b:
$R_{++}^{(b)} \leftarrow \frac{1}{n_B} \sum_{k=1}^{K} \sum_{l=1}^{n_B} \left\| y_l^{(b)} - c_{k,++}^{(b)} \right\|^2 \mathbf{1}\{y_l^{(b)} \in \mathcal{C}_k^{(b)}\}$.

(3) Select the block *bmed* having the median empirical risk.

Return $\left(\widehat{c}_{1,++}^{(bmed)}, \dots, \widehat{c}_{K,++}^{(bmed)}\right)$ the centroids of the median block *bmed*.

FIGURE 6. Algorithm of the initialization strategy

- compute the performance criterion of the centroids obtained on each block
- (MOM part) select the centroids of the block with the median performance as initialisation centroids.

3.4. **Theoretical garantees**

As claimed in the introduction, we proved that K-bMOM has robust properties in terms of breakdown point as well as in terms of excess risk bounds. We focus in this summary on the result about the breakdown point. For further information, refer to our article [8] and supplementary [7] Sections 3.1 and 3.2. The definition of the probabilistic breakdown point we consider is the following.

Definition 3.1. The probabilistic breakdown point of a randomized estimator $\widehat{T}^{\omega}$ (ω accounts for randomization) and given the sample u_1^n is

$$p\left(\widehat{T}^{\omega}, u_1^n, (i_1, ..., i_m)\right) = \mathbb{P}\left(\left\{\omega : \sup_{e_1,...,e_m} \left|\widehat{T}^{\omega}(s_1, ..., s_n)\right| < \infty\right\}\right). \tag{13}$$

where the sample $(s_1, ..., s_n)$ is obtained by replacing the m data points $u_{i_1}, ..., u_{i_m}$, for some fixed indices $(i_1, ..., i_m)$, by the arbitrary values $e_1, ..., e_m$.

The quantity $p(\widehat{T}, u_1^n, (i_1, ..., i_m))$ defined in (13) can be understood as the probability that the output of the randomized estimator $\widehat{T}$ remains bounded, independently of the values taken by $u_{i_1}, \dots, u_{i_m}$, for fixed indices $i_1, \dots, i_m$. Hence, the only randomness considered on this event is the randomization of the estimator. Let us also highlight that in general, the values of the indices $i_1, ..., i_m$ do not play a significant role, since one either consider the worst case possible for a given m, that is to say, one is interested in: $p(\widehat{T}, u_1^n, m) = \inf_{i_1,...,i_m} p(\widehat{T}, u_1^n, (i_1, ..., i_m))$, or the probabilistic breakdown point does not depend on the values $i_1, ..., i_m$ but on m only, as it is the case in the next theorem.

Under a configuration of the data said "well-clusterisable" we can state a breakdown point for K-bMOM. This configuration is defined as follows.

Definition 3.2. 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 $B(a_k, r)$, $k = 1, ..., K$, of radius r with centers a_k separated from each other by at least a distance R: $\min_{k \neq k'} ||a_k - a_{k'}|| \geq R$. Moreover, each ball $B(a_k, r)$ is assumed to contain exactly one cluster.

In this context, the randomized estimator K-bMOM outputs some centroids $\bar{\mathbf{c}}^{\omega}$ as defined in algorithm 5.

Theorem 3.3. *Let $\omega \mapsto \bar{\mathbf{c}}^{\omega}$ be the K-bMOM randomized output computed iteratively using at each step B blocks of size n_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, satisfying $R^2 > 16 n_B r^2$. Finally, assume that at the beginning of the last 10 iterations, the algorithm has identified the correct partition of the regular data, meaning that one cluster is associated with one centroid. It holds then that $p(\bar{\mathbf{c}}, x_1^n, (i_1, ..., i_m)) \geq \max\{p_1 - p_2, 0\}$ with*

$$p_1 = \left(1 - \sum_{k=1}^{K} \left(1 - \frac{n_k^r}{n}\right)^{n_B}\right)^{10B} \tag{14}$$

and

$$p_2 = 10 \exp\left(-2B\left(\left(1 - \frac{m}{n}\right)^{n_B} - \frac{1}{2}\right)^2\right), \tag{15}$$

where the quantity n_k^r in display (14) stands for the number of regular data (not outliers) belonging to cluster k in the corrupted version of original dataset x_1^n.

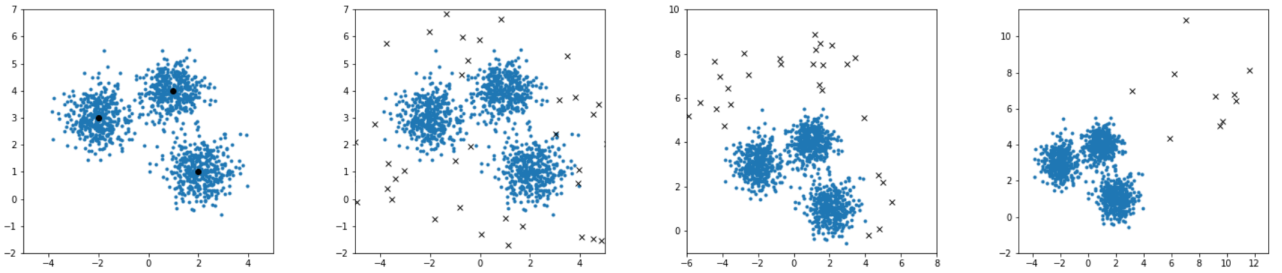
The latter theorem states that when the data are in a so-called "well-clusterisable" configuration and the algorithm benefits from a "warm start", then K-bMOM is able to robustly provide K centroids with positive probability if, among other things, the proportion of outliers m/n is not too large and ranges from 0 to $\left\lfloor n\left(1 - 1/2^{1/n_B}\right)\right\rfloor / n \approx 0.69/n_B$. The 'warm start" assumption corresponds in Theorem 3.3 to postulating that the algorithm has recovered the correct partition before the last ten iterations. Assuming a good enough initialisation is quite common when analyzing iterative algorithms in non-convex situations. In our robust context, we believe that such an assumption is legitimated by the good behavior in practice of our proposed robust initialisation (see Section 3.5 below).

According to Theorem 3.3, one should follow a rationale when choosing the hyper-parameters n_B and B of the K-bMOM algorithm. Indeed, the condition $(1-m/n)^{n_B} > 1/2$ implies that n_B should not be taken too large, but it should also be large enough so that p_1 would be close to one. As for B, the higher its value, the closer p_2 to zero, but the smaller its value, the closer p_1 to one. We refer to [8] for further comments on the behavior of the quantities p_1 and p_2, and also on the practical choice of the hyper-parameters of K-bMOM.

3.5. **Performances of the robust initialisation K-bMOM-kmeans++**

We carried out some experiments to measure the benefits of our approach to robust initialisation. In this short article, we introduce only one of them, and refer to [8] for more experiments. We compare the performance of six different initialisations in three different contexts of outliers. Among the six methods four are well known: random, k-means++ [2], its variant k-medians++ and ROBIN [1]. In addition we compare "k-bmom-kmeans++" and "k-bmom-kmedians++", two methods obtained when one applies bMOM strategy introduced in Section 3.3 to kmeans++ and to kmedians++ respectively. The three types of outliers are said to be punctual (T1), oriented (T2) and clustered (T3). Punctual outliers are located randomly uniformly around the clusters to decrease the contrast between the clusters. Oriented outliers are originally data points whose coordinates get multiplied by a given factor. Finally the third type corresponds to "clustered" outliers because they are all located in the same region of the space, see Figure 7.

The obtained results after 1000 repetitions are shown in Table 1. This table shows overall that our methodology applied on kmeans++ and kmedians++ achieves better performances. In particular, in the case T1, all methods perform well. The performances of the usual methods drop however when other types of outliers are introduced (outlier types T2 and T3).



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

FIGURE 7. Illustrations of simulated regular data (blue points) generated according to a Gaussian Mixture Model with isotropic variance and different types of outliers (black crosses).

type of outlier		initialization	RMSE	distortion	accuracy
T1	isolated	random	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)**
T2	oriented & isolated	random	**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)**
T3	cluster of outliers	random	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)
		K-means++	0.880 (0.360)	2.472 (1.755)	0.756 (0.158)
		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 1. Aggregated performances according to the typology of outliers for different strategies of initialization. RMSE = Root Mean Square Error. The distorsion is the K-means risk computed over the true inliers only. The accuracy is the greatest percentage of well clustered regular data over predicted label permutations.

3.6. **Conclusion**

In this short article, we introduced some important elements of our article [8] and its supplementary material [7]. We explain what we call the "MOM (or bMOM) strategy" and how it can be used in practice. In particular, we introduced the procedures called "K-bMOM" (when applied to the iterations of K-means) and called "K-bmom-kmeans++" (when applied to the initialisation strategy K-means++). Then we stated a theoretical result about a probabilistic breakdown point of the K-bMOM procedure, showing that if the data are in a so-called "well-clusterisable" configuration then K-bMOM is able to provide robust centroids while the proportion of outliers can range from 0 to $\lfloor n\left(1-1/2^{1/n_B}\right)\rfloor/n \approx 0.69/n_B$. Finally, we have shown that the proposed initialisation "K-bmom-kmeans++" matches the definition of a "robust initialisation" because it

keeps good performances even if outliers induce a strong bias/perturbation, while other methods start to break down.

References

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

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

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

[4] C. Brécheteau. Robust anisotropic power-functions-based filtrations for clustering. pages 23:1–23:15, 2020.

[5] C. Brécheteau, A. Fischer, and C. Levrard. Robust Bregman clustering. *Annals of Statistics*, 49(3):1679–1701, 2021.

[6] C. Brécheteau and C. Levrard. A k-points-based distance for robust geometric inference. *Bernoulli*, 26(4):3017–3050, 2020.

[7] C. Brunet-Saumard, E. Genetay, and A. Saumard. Supplement to: "K-bMOM: a robust Lloyd-type clustering algorithm based on bootstrap Median-of-Means". 2020.

[8] C. Brunet-Saumard, E. Genetay, and A. Saumard. K-bMOM: a robust Lloyd-type clustering algorithm based on bootstrap median-of-means. *Comput. Statist. Data Anal.*, 167:Paper No. 107370, 19, 2022.

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

[10] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

[11] O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(4):1148–1185, 11 2012.

[12] O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression, 2017. (`https://arxiv.org/abs/1712.02747`).

[13] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.

[14] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 60(6):A:1 – A:38, 2013.

[15] J.A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed k-means: An attempt to robustify quantizers. *Annals of Statistics*, 25:553–576, 1997.

[16] J. Depersin and G. Lecué. On the robustness to adversarial corruption and to heavy-tailed data of the Stahel-Donoho median of means. *arXiv preprint arXiv:2101.09117*, 2021.

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

[18] I. Diakonikolas and D. M. Kane. Recent Advances in Algorithmic High-Dimensional Robust Statistics, 2019.

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

[20] F. R. Hampel. A general qualitative definition of robustness. *Ann. Math. Statist.*, 42(6):1887–1896, 12 1971.

[21] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley, 1st edition edition, January 1986. missing.

[22] Frank Hampel. Some thoughts about classification. In *Classification, clustering, and data analysis (Cracow, 2002)*, Stud. Classification Data Anal. Knowledge Organ., pages 5–26. Springer, Berlin, 2002.

[23] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[24] P. J. Huber and E. M. Ronchetti. *Robust statistics; 2nd ed.* Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2009.

[25] Y. Klochkov, A. Kroshnin, and N. Zhivotovskiy. Robust k-means clustering for distributions with two moments. *Ann. Statist.*, 49(4):2206–2230, 2021.

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

[27] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *Ann. Statist.*, 48(2):906–931, 2020.

[28] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

[29] S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. on Information Theory*, 28(2):129–136, 1982.

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

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

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

[33] Timothée Mathieu. *M-estimation and Median of Means applied to statistical learning*. Theses, Université Paris-Saclay, January 2021.
[34] Timothée Mathieu. Concentration study of M-estimators using the influence function. *Electronic Journal of Statistics*, 16(1):3695–3750, 2022.
[35] S. Minsker. Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*, 2018.
[36] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
[37] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
[38] D.L. Vandev and N. Neykov. Robust maximum likelihood in the Gaussian case. *New Directions in Data Analysis and Robustness*, 01 1993.
[39] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.
[40] U. von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2017.