Statistica Sinica Preprint No: SS-2024-0414	
Title	GROS: A General Robust Aggregation Strategy
Manuscript ID	SS-2024-0414
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0414
Complete List of Authors	Alejandro Cholaquidis, Emilien Joly and Leonardo Moreno
Corresponding Authors	Leonardo Moreno
E-mails	leonardo.moreno@fcea.edu.uy

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

GROS:

A General Robust Aggregation Strategy

Alejandro Cholaquidis [1], Emilien Joly[2], Leonardo Moreno [3]

[1] Centro de Matemáticas, Facultad de Ciencias, UDELAR, Uruguay.

[2] Centro de Investigación en Matemáticas, CIMAT, Guanajuato, México

[3] Instituto de Estadística, Departamento de Métodos Cuantitativos, Facultad de Ciencias Económicas y de Administración, UDELAR, Uruguay.

Abstract

A new, very general, robust procedure for combining estimators in metric spaces is introduced (GROS). The method is reminiscent of the well-known median of means, as described in Devroye, Lerasle, Lugosi and Oliveira (2016). Initially, the sample is divided into K groups. Subsequently, an estimator is computed for each group. Finally, these K estimators are combined using a robust procedure. We prove that this estimator is sub-Gaussian and we get its breakdown point, in the sense of Donoho. The robust procedure involves a minimization problem on a general metric space, but we show that the same (up to a constant) sub-Gaussianity is obtained if the mini-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

mization is taken over the sample, making GROS feasible in practice. The performance of GROS is evaluated through five simulation studies: the first one focuses on classification using k-means, the second one on the multi-armed bandit problem, the third one on the regression problem. The fourth one is the set estimation problem under a noisy model. We apply GROS to get a robust persistent diagram. Lastly, an application of robust estimation techniques to determine the home-range of Canis dingo in Australia is implemented.

Keywords: Bandits, Median of means, Robustness, Sub-Gaussian estimator, Topological data analysis.

MSC code: 62G05, 62G20,62G35.

1. Introduction

The problem of combining estimators has been extensively studied in statistics, with foundational contributions tracing back to the seminal work of (James and Stein, 1961), who showed that combining biased estimators could outperform unbiased ones under quadratic loss. More recently, methods such as stacking Wolpert (1992); Breiman (1996) and ensemble learning approaches like boosting and bagging Freund and Schapire (1997); Breiman (2001) have provided robust frameworks for combining estimators

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

in both parametric and non-parametric settings. There are recent proposals that merge regression estimators (see, for instance, Biau, Fischer, Guedj and Malley (2016)), classifiers (Cholaquidis, Fraiman, Kalemkerian and Llop (2016)), and density estimators (Cholaquidis, Fraiman, Ghattas and Kalemkerian (2021)), among others. In these scenarios, the aim is to merge the estimators to generate one that, at least asymptotically, surpasses the best of the group. In other instances, the aim is to derive a robust estimator.

Robust estimation techniques aim to produce reliable statistical inference even in the presence of deviations from idealized assumptions, such as outliers or heavy-tailed distributions. The use of these estimators has proven to be valuable in varied statistical scenarios, such as in machine learning, see Lecué and Lerasle (2020). In these contexts, it is advisable to consider estimators that, without removing outliers, do not reduce their precision. Robust statistics point in this direction, see Maronna, Martin, Yohai and Salibián-Barrera (2019). Over the past decades, two main generations of robust estimators have emerged. Among the classical approaches to robust estimation, M-estimators are a prominent example, see Huber (1964). These estimators retain consistency and asymptotic normality under mild conditions and offer improved resistance to outliers compared to

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

least squares. However, their performance can degrade significantly when the contamination is not sparse or is adversarially structured. Moreover, they often require tuning parameters and careful implementation to balance robustness and efficiency.

In contrast, Median-of-Means (MOM) estimators, initially proposed in the context of mean estimation under heavy-tailed noise, provide finite-sample guarantees and are particularly well-suited to modern statistical challenges such as high-dimensionality, online learning, and adversarial contamination.

The MOM estimator is a robust statistical technique for estimating the mean of a distribution, particularly useful when the data may contain outliers or come from heavy-tailed distributions. The method traces back to the foundational work of Nemirovsky and Yudin Nemirovsky and Yudin (1983), who introduced it in the context of stochastic optimization. Their motivation was to develop estimators that remain stable under uncertainty and variability, providing strong guarantees even in the presence of noise. Later, the estimator was further explored from a statistical standpoint by Devroye, Györfi, and Lugosi Devroye, Györfi and Lugosi (1996), who presented MOM as an essential tool for robust estimation in the context of pattern recognition. In recent years, the work of Lugosi and Mendelson

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

Lugosi and Mendelson (2019) revitalized interest in MOM estimators, especially in high-dimensional and adversarial settings.

In the MOM the data, $\aleph_n$ (an i.i.d. sample of a random variable X), is first randomly partitioned into K groups. Subsequently, the mean of each group is computed. The MOM estimator is then the median of these K means. If the variance of the data is assumed to be finite, this estimator is sub-Gaussian. For further details, we refer to Devroye, Lerasle, Lugosi and Oliveira (2016), Joly, Lugosi and Oliveira (2017), and the references therein. For the case of random vectors, the so-called median of means tournament is introduced in Lugosi and Mendelson (2019), where is proved to be sub-Gaussian. In Rodriguez and Valdora (2019) it is proved that the median of means tournament has break-down point $\lfloor (K-1)/2 \rfloor / n$, where $\lfloor x \rfloor$ denotes the floor of x and n is the sample size.

Following this idea of dividing into K groups, calculating the estimator in each group, and then combining them, we will introduce a new way to combine the estimators in order to obtain, in a very general framework, a new one that, under the assumption that the estimators by group are independent, turns out to be sub-Gaussian (see Theorem 3 below). This new strategy, in what follows: GROS, has breakdown point $\lceil K/2 \rceil / n$, where $\lceil x \rceil$ denotes the ceiling of x, see Section 2. The only assumption we make is that

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

the original sample comes from a random variable with finite variance and that the space where the group estimators take their values is a separable and complete metric space.

While the combined estimator requires solving a minimization problem in a metric space, we prove that if it is minimized on the sample of the group estimators, an appropriate candidate is obtained. We also determine how much is lost by this choice, see Section 3.

Due to the immense generality of GROS, it can be applied to various areas of statistics where robustness plays a key role. Furthermore, the space in which the estimators reside doesn't need to be a metric space: a pseudometric suffices. This permits the consideration of estimators in the space of bounded subsets of $\mathbb{R}^d$ equipped with the Hausdorff distance or the measure distance, which is the case when the object to be estimated is, for instance, a set. This space will be used in our fourth simulation example, see subsection 4.2.

We have chosen to present five problems to demonstrate its performance, comparing it with techniques explicitly crafted for these specific issues. Some of these techniques were already designed to yield robust estimators. Specifically, we treat:

- The traditional clustering problem.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

- The multi-armed bandit problem with heavy-tailed rewards; included in the sumplementary material.
- Regression in the presence of noisy data; included in the sumplementary material.
- The estimation of a convex set when dealing with a noisy sample.
- An application to topological data analysis.

It is worth noting that the fourth problem cannot be successfully treated using conventional methods such as convex hulls or r-convex hulls, as we will see.

As expected, in all cases the performance of GROS is noticeably better than the proposals that do not consider the presence of outliers, see Section 4. Moreover, the performance is good compared to methods that do consider the presence of outliers, even surpassing some in certain cases.

The structure of the paper is outlined as follows: Section 2 introduces GROS within a broad context and examines its robustness properties. Section 3 presents a modification of GROS to simplify its computational aspects. In Section 4, the application of GROS across five problems is discussed. The codes were developed in R. They are available at `https://github.com/mrleomr/GROS`. An example using real data is pro-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

vided in Section 5. The paper concludes with Section 6, where the findings and implications of the study are elaborated.

2. Robust aggregation of weakly convergent estimators

In this section, we define and study a new proposal of a robust estimator based on the aggregation of estimators. We assume given a sample $\aleph_n = \{X_1, \ldots, X_n\}$ of i.i.d. random elements with common distribution P. Let μ be a certain characteristic of P. We assume that μ belongs to a complete and separable metric space $\mathcal{M}$ endowed with a metric d. All the results in this paper remains true if d is a pseudo metric.

In the context of robust estimation, one goal is to obtain sub-Gaussian type inequalities for the deviation of an estimator $\widehat{\mu}$ from μ. A common way of defining a robust, distribution-free estimator is to make K disjoint groups out of $\aleph_n$, hence, to create a collection of K independent estimators $\mu_1, \ldots, \mu_K \in \mathcal{M}$. We define the GROS of $\mu_1, \ldots, \mu_K$ by

$$\mu^* = \underset{\nu \in \mathcal{M}}{\operatorname{argmin}} \min_{I : |I| > \frac{K}{2}} \max_{j \in I} d(\mu_j, \nu). \tag{2.1}$$

The minimization is taken over all the possible subsets I of $\{1, \ldots, K\}$ that contain at least $\lfloor K/2 \rfloor + 1$ indices. However, it is enough to minimize over all possible subsets I whose cardinality, $|I|$, fulfills $|I| = \lfloor K/2 \rfloor + 1$. Indeed, for any set I of cardinality strictly bigger than $\lfloor K/2 \rfloor + 1$ we can

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

find a set I_0 such that $|I_0| = \lfloor K/2 \rfloor + 1$ and for which $\max_{j \in I_0} d(\mu_j, \nu) \leq \max_{j \in I} d(\mu_j, \nu)$. Observe that, for any $\nu \in \mathcal{M}$, $\min_{I:|I|>\frac{K}{2}} \max_{j \in I} d(\mu_j, \nu) =: d(\nu, \nu_{(\lfloor K/2 \rfloor+1)-\mathrm{NN}})$, where $\nu_{(\lfloor K/2 \rfloor+1)-\mathrm{NN}}$ denotes the $(\lfloor K/2 \rfloor + 1)$-nearest neighbor of ν in $\mu_1, \ldots, \mu_K$. This last quantity is a measure of the depth of ν inside the set $\mu_1, \ldots, \mu_K$. Then, μ^* is the point with the least depth from all the candidates $\nu \in \mathcal{M}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

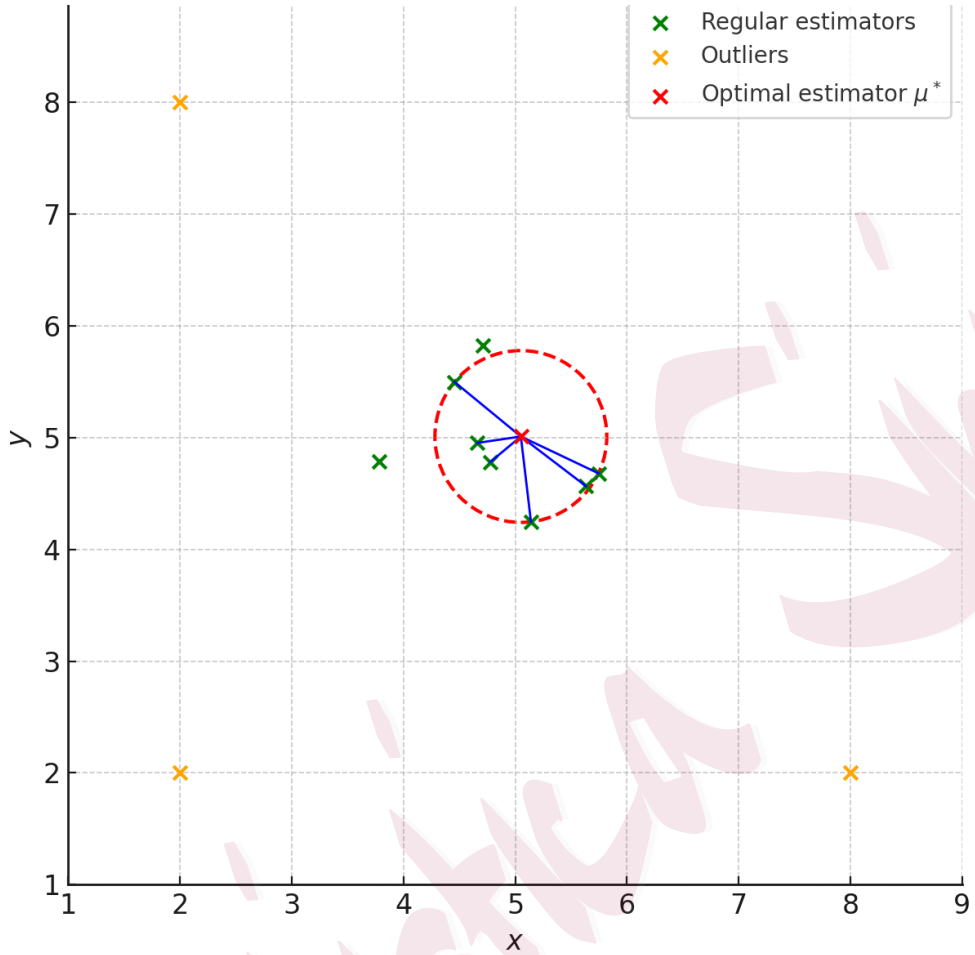


Figure 1: Illustrating the definition of the robust estimator μ^* for 11 estimators μ_j. In this case, 3 of the estimators μ_j have an erratic behavior and are mark as outliers. The blue lines indicate the 6 estimators taken into account in the minimization (2.1).

In full generality, the set of minimizers of (2.1) may not be unique. In

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

that case, we still denote by μ^* one of the minimizers arbitrarily chosen.

Lemma 1. Assume that $\mathcal{M}$ such that $B(\mu, r)$, the closed ball of center μ and radious $r > 0$, is compact, for all $\mu \in \mathcal{M}$ and r, then the set μ^* given by (2.1) is non-empty.

Remark 1. A natural generalization of μ^* is to define, for $q \in [1/2, 1)$, $\mu_q^* = \text{argmin}_{\nu \in \mathcal{M}} \min_{I:|I|>Kq} \max_{j \in I} d(\mu_j, \nu)$. All the results we present are for $q = 1/2$ but they remain true for any $q \in [1/2, 1)$.

As we said in the Introduction, we aim to combine the estimators $\mu_1, \ldots, \mu_K$ in a robust way. More precisely, let us recall the definition of finite-sample breakdown point introduced by Donoho (see Donoho (1982)).

Definition 1. Let $\mathbf{x} = \{x_1, \ldots, x_n\} \subset \mathcal{M}$ be a dataset, θ an unknown parameter lying in a metric space (Θ, ρ), and $\widehat{\theta}_n = \widehat{\theta}_n(\mathbf{x})$ an estimator based on $\mathbf{x}$. Let $\mathcal{X}_p$ be the set of all datasets $\mathbf{y}$ of size n having $n - p$ elements in common with $\mathbf{x}$: $\mathcal{X}_p = \{\mathbf{y} : |\mathbf{y}| = n \text{ and } |\mathbf{x} \cap \mathbf{y}| = n - p\}$. Then, the breakdown point of $\widehat{\theta}_n$ at $\mathbf{x}$ is $\epsilon_n^*(\widehat{\theta}_n, \mathbf{x}) = p^*/n$, where

$$p^* = \max\{p \geq 0 : \forall \mathbf{y} \in \mathcal{X}_p, \widehat{\theta}_n(\mathbf{y}) \text{ is bounded and } \exists \epsilon > 0 : \rho(\widehat{\theta}_n(\mathbf{y}), \partial\Theta) > \epsilon \quad \text{if } \partial\Theta \neq \emptyset\}.$$

From (2.1) it follows easily that the finite-sample breakdown point of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

μ^* is $\lceil K/2 \rceil / n$, which is the same order obtained in Rodriguez and Valdora (2019) for the MOM aggregation strategy mentioned in the Introduction.

The following lemma states that if there exists an η for which at least $K/2$ of the μ_i are at a distance at most t from η, then any minimum in (2.1) is at a distance at most $2t$ from η. This, as we will see, implies the robustness and sub-Gaussianity of the estimator (2.1). Let us write $[K] = \{1, \dots, K\}$.

Lemma 2. Let $t > 0$. Assume that there exist an $\eta \in \mathcal{M}$ and an $I \subset [K]$ with $|I| > K/2$ such that for all $j \in I$, $d(\mu_j, \eta) \leq t$. Then, $d(\mu^*, \eta) \leq 2t$.

Lemma 2 can be applied when $\eta = \mu$, in which case if a group of more than $K/2$ estimators is reasonably close to the objective μ, then μ^* itself is reasonably close. Such an estimator is robust to outliers since this effect will not be altered by the bad behavior of up to $K/2 - 1$ estimators. This lemma is a technical fact that will allow us to use the so called binomial argument. Indeed, assume that μ is such that for any $0 < p < 1/2$, there exists $t = t(n, K)$ such that for all $k \in 1, \dots, K$, $\mathbb{P}\left(d(\mu_k, \mu) > t\right) \leq p$. Since the estimators $\mu_1, \dots, \mu_K$ are independent and identically distributed,

Negating Lemma 2 leads to the following fact: if, for a certain $t > 0$, $d(\mu^*, \eta) > 2t$ then, for all I containing at least half of the points ($|I| \geq \lceil K/2 \rceil$) there exists $j \in I$ such that $d(\mu_j, \eta) > t$. Equivalently, there exists

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

I with $|I| \geq \lfloor K/2 \rfloor$ (by taking I to be the set of indexes i such that $d(\mu_i, \mu) > t$) such that $\forall i \in I, d(\mu_i, \mu) > t$. Using this reformulation, we get

$$\begin{aligned}
\mathbb{P}\left(d(\mu^*, \mu) > 2t\right) &\leq \mathbb{P}(\exists I : |I| \geq \lfloor K/2 \rfloor \text{ and } \forall i \in I, d(\mu_i, \mu) > t) \\
&= \mathbb{P}\left(\sum_{k=1}^{K} \mathbb{I}_{\{d(\mu_k, \mu) > t\}} \geq \lfloor K/2 \rfloor\right) \\
&\leq \mathbb{P}\left(B_{K,p} \geq \lfloor K/2 \rfloor\right) \\
&\leq e^{\frac{-2(\lfloor K/2 \rfloor - Kp)^2}{K}}, \qquad (2.2)
\end{aligned}$$

where $B_{K,p}$ denotes a random variable with binomial distribution, with parameters K and p. Let us assume that the μ_i are identically distributed such that $\mathbb{E}d^2(\mu_i, \mu)$ is finite. Then, if we choose $p = 1/4$ and $K = \lceil 8 \log(\delta^{-1}) \rceil$, using the fact that $\lfloor \lceil x \rceil / 2 \rfloor \leq \lceil x \rceil / 2$, we get from (2.2) together with Markov's inequality

$$\mathbb{P}\left(d(\mu^*, \mu) > 4\sqrt{\mathbb{E}d^2(\mu_1, \mu)}\right) \leq \delta.$$

Lastly, we have proved the following theorem:

Theorem 3. Let $\aleph_n = \{X_1, \ldots, X_n\}$ be an i.i.d. sample from a distribution P. Let $\mu \in \mathcal{M}$ a certain characteristic of P where $(\mathcal{M}, d)$ is a metric space. Let $\delta \in (0, 1)$ and $K = \lceil 8 \log(\delta^{-1}) \rceil$. We split $\aleph_n$ into K disjoint groups (assume that n guarantee that $n/K = \ell \in \mathbb{N}$), and create K inde-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

pendent and identically distributed estimators $\mu_1, \ldots, \mu_K$ of μ. Let μ^* be the aggregation defined by (2.1). Then,

$$\mathbb{P}\left(d(\mu^*, \mu) > 4\sqrt{\mathbb{E}d^2(\mu_1, \mu)}\right) \leq \delta. \tag{2.3}$$

As Theorem 3 is written, the role of the hyperparameter K can be unclear at first sight. It is set to $\lceil 8\log(\delta^{-1})\rceil$ to balance two effects. The first aspect is that if K is large, the upper bound in Equation 2.2 is small and so the aggregation step is fairly effective. In this regime n/K is small and so the number of data points inside each group that form the samples used for the calculation of each individual estimators μ_i is small. This tends to deteriorate the performance of each of the individual estimators hence the value of $\mathbb{E}d^2(\mu_1, \mu)$. On the contrary, if K is small, the mean squared error $\mathbb{E}d^2(\mu_1, \mu)$ becomes small but the bound of Equation 2.2 looses in usefulness so that it may not be possible to upper bound it by the level δ. The parameter K has been chosen as small as possible in the regime where the upper bound of Equation 2.2 is less than δ.

Remark 2. In Theorem 3, the restriction that n guarantee $n/K = \ell \in \mathbb{N}$ is purely technical, to ensure that the $\mu_1, \ldots, \mu_K$ are identically distributed and then to get the clean expression (2.3). If this is not the case, that is, the groups are unbalanced, the estimators $\mu_1, \ldots, \mu_K$ are independent but

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

not necessarily identically distributed, and the obtained bound is

$$\mathbb{P}\left(d(\mu^*,\mu) > 4\sqrt{\max_{i=1,\dots,K}\mathbb{E}d^2(\mu_i,\mu)}\right) \leq \delta. \tag{2.4}$$

Remark 3. Under the slightly weaker hypothesis $\mathbb{E}\left[d(\mu_1,\mu)^{1+\epsilon}\right] < \infty$, and using that $\mathbb{P}(d(\mu_k,\mu) > t) \leq \frac{1}{t^{1+\epsilon}}\mathbb{E}(d^{1+\epsilon}(\mu_k,\mu))$, it follows from (2.2), taking $p = 1/4$ that, $\mathbb{P}\left(d(\mu^*,\mu) > \left[4\mathbb{E}d^{1+\epsilon}(\mu_1,\mu)\right]^{\frac{1}{1+\epsilon}}\right) \leq \delta$.

This assumption in quite classical in the robust estimation community.

2.1 Mis-specification of the set $\mathcal{M}$

Through this section we assume that $\mathcal{M}$ is a subset of a metric space $(\mathcal{T}, d)$ and $\widetilde{\mathcal{M}} \subset \mathcal{T}$ possibly disjoint with $\mathcal{M}$. Lemma 2 uses the fact that η is inside the set $\mathcal{M}$. Now, assume that $\widetilde{\mathcal{M}}$ is mis-specified in the sense that $d(\eta,\widetilde{\mathcal{M}}) = \inf_{\nu\in\widetilde{\mathcal{M}}} d(\eta,\nu) = \epsilon > 0$. The true parameter of interest does not belong to the set of features $\widetilde{\mathcal{M}}$. The minimization is then given by

$$\widetilde{\mu} = \operatorname*{argmin}_{\nu\in\widetilde{\mathcal{M}}} \min_{I:|I|>\frac{K}{2}} \max_{j\in I} d(\mu_j,\nu). \tag{2.5}$$

Lemma 4. Assume that there exist an $\eta \in \mathcal{M}$ and an $I \subset [K]$ with $|I| > K/2$ such that for all $j \in I$, $d(\mu_j,\eta) \leq t$. Then, $d(\widetilde{\mu},\eta) \leq 2t + \epsilon$.

The following corollary is a direct consequence of Lemma 4.

Corollary 1. *Let $\aleph_n = \{X_1,\dots,X_n\}$ be an i.i.d. sample from a distribution P. Let $\mu \in \mathcal{M}$ be a certain characteristic of P, where $(\mathcal{M},d)$ is a metric*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

space. We assume given a set $\widetilde{\mathcal{M}}$ *(possibly random) such that* $d(\eta, \widetilde{\mathcal{M}}) = \epsilon > 0$. *Let* $\delta > 0$ *and* $K = \lceil 8 \log(\delta^{-1}) \rceil$. *We construct the* K *disjoint groups and* K *estimators* $\mu_1, \ldots, \mu_K$ *of* μ *as in Theorem 3. Then,*

$$\mathbb{P}\left(d(\widetilde{\mu}, \eta) > 4\sqrt{\mathbb{E}d^2(\mu_1, \mu)} + \epsilon \right) \leq \delta. \tag{2.6}$$

3. Computational aspects

Equation (2.1) supposes that one is able to find minimizers of a complex functional on the metric space $\mathcal{M}$, which is often an unfeasible problem. To simplify that task, one can restrict the minimization to the set of estimators $\mu_1, \ldots, \mu_K$. That is, we find the index j^* such that

$$j^* = \underset{j=1,\ldots,K}{\operatorname{argmin}} \min_{I:|I|>K/2} \max_{i \in I} d(\mu_i, \mu_j). \tag{3.1}$$

The next lemma and theorem state that μ_{j^*} has the same sub-Gaussian type bound (up to a constant) as μ^*.

Lemma 5. Assume that there exists an $I \subset [K]$ such that $|I| > K/2$, and for all $j \in I$, $d(\mu, \mu_j) \leq t$. Then $d(\mu_{j^*}, \mu) \leq 3t$.

By means of Lemma 5, it is possible to give a practical version of Theorem 3.

Theorem 6. Assume the hypotheses of Theorem 3. Let $\mu_{j^*} \in \{\mu_1, \ldots, \mu_K\}$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

defined by the optimization (3.1). Then,

$$\mathbb{P}\left(d(\mu_{j^*}, \mu) > 6\sqrt{\mathbb{E}d^2(\mu_1, \mu)}\right) \leq \delta. \tag{3.2}$$

Note that Equation (3.2) is the same as Equation (2.3) except that it has the constant 6 in the right-hand side. This shows that the practical version of the estimator, μ_{j^*}, has essentially the same rate of convergence as μ^*, but it can be deteriorated by a constant factor.

4. Some applications of GROS

4.1 Clustering by k-means

One of the most popular procedures for determining clusters in a dataset is the k-means method. Although the precursors of this algorithm were MacQueen in 1967, see McQueen (1967), and Hartigan in 1978, see Hartigan (1978). Pollard in 1981 Pollard (1981) proved the strong consistency of the method and in 1982, in Pollard (1982), determined its asymptotic distribution.

Given k, the k-means clustering procedure partitions a set $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ into k groups as follows: first k cluster centres a_j are chosen, in such a way as to minimise $W_n = \frac{1}{n}\sum_{l=1}^{n} \min_{1\leq j\leq k}\|x_l - a_j\|^2$. Then it assigns each x_l to its nearest cluster centre. In this way, each centre acquires a subset

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

C_l as its associated cluster. The mean of the points in C_l must equal a_l, otherwise W_n could be decreased by replacing a_l with the cluster mean in the first instance, and then reassigning some of the x's to their new centres. This criterion is then equivalent to that of minimising the sum of squares between clusters.

The standard k-means algorithm starts from a set of initial centres $a_1^{(1)}, \ldots, a_k^{(1)}$, and alternates, up to a stopping criterion, two steps: *Assignment step:* Assigns each observation x_j to the cluster whose centre is the closest one. *Update step:* Recalculate means (centroids) for the observations assigned to each cluster, by averaging the observations in each cluster.

These k-centres are not robust to the presence of outliers, nor to the distribution's possession of heavy tails. There are several proposals in the literature that seek to make the k-means algorithm more robust. An example is the k-medoids algorithm (called PAM), see Kaufman (1990); Kaufman and Rousseeuw (2009). In this algorithm, in the second step, one chooses that point in the cluster which minimises the sum of the distances to the remaining observations. These points are called the medoids.

Another proposal for a robust version of k-means, TClust, is developed in Cuesta-Albertos, Gordaliza and Matrán (1997). It is based on an $\alpha > 0$ trimming of the data. This trimming is self-determined by the data and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

aims to mitigate the impact of extreme data.

We propose a simple modification to the second step of the k-means algorithm, which will be referred to as *RobustkM*. The idea is very simple: instead of calculating the centroids by taking the arithmetic mean in each group, the centroids are determined using (3.1).

4.1.1 Simulations

To evaluate the performance of this proposal, we run a small simulation study. In all cases, the data consist of an i.i.d. sample $X_1, \ldots, X_n$, whose common distribution, F_X, is given by the mixture of three bi-variate Student distributions. More precisely,

$$F_X(x) = 0.45T\left(x, \mu_1, \nu, \Sigma_1\right) + 0.45T\left(x, \mu_2, \nu, \Sigma_2\right) + 0.1T\left(x, \mu_3, \nu, \Sigma_3\right), \tag{4.1}$$

where $T\left(x, \mu, \nu, \Sigma\right)$ denotes the bi-variate cumulative Student distribution function with mean $\mu \in \mathbb{R}^2$, variance and covariance matrix Σ, and $\nu = 2$ degrees of freedom.

In the examples, we chose $\mu_1 = (6, 0)$, $\mu_2 = (-6, 0)$, $\mu_3 = (0, 6)$, $\Sigma_1 = \Sigma_2 = \left(\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right)$ and $\Sigma_3 = \left(\begin{smallmatrix} 4 & 1 \\ 1 & 9 \end{smallmatrix}\right)$.

Figure 2 shows a simulation with $n = 1000$ points. It can be seen that the dispersion of the third group makes the clustering problem more

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

difficult.

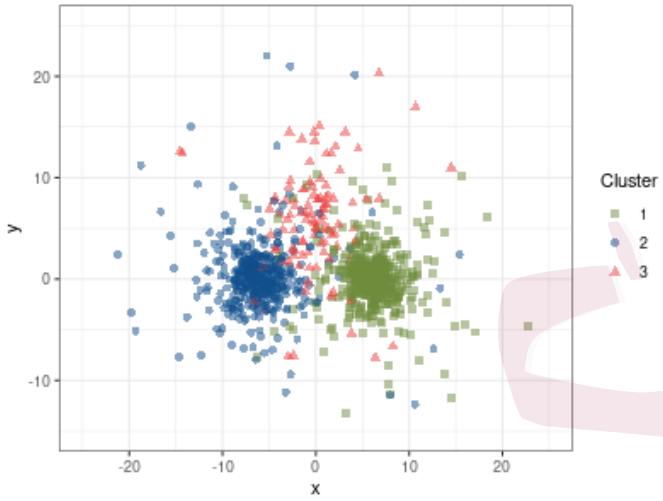


Figure 2: Simulation of 1000 observations of the multivariate Student mixture (4.1). Observations are colored according to the component of the mixture which the data comes from.

Let π be a permutation of the set of labels $\{1, 2, 3\}$. Denote by $C(x) \in \{1, 2, 3\}$ the true (unknown) label, and $\widehat{C}(x) \in \{1, 2, 3\}$ the label assigned by the algorithm to observation x. Then the classification error is given by

$$\min_{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\left\{C(x_i) \neq \pi(\widehat{C}(x_i))\right\}}. \tag{4.2}$$

Figure 3 shows the performance of RobustkM (with $K = 10$), k-means, PAM, and TClust (with $\alpha = 0.01$), over 1000 replications. In TClust, the trimmed data (at the end of the algorithm) are assigned to the nearest cen-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

tres. This toy example shows that the proposed algorithm is a competitive alternative to other methods that "robustify" k-means.

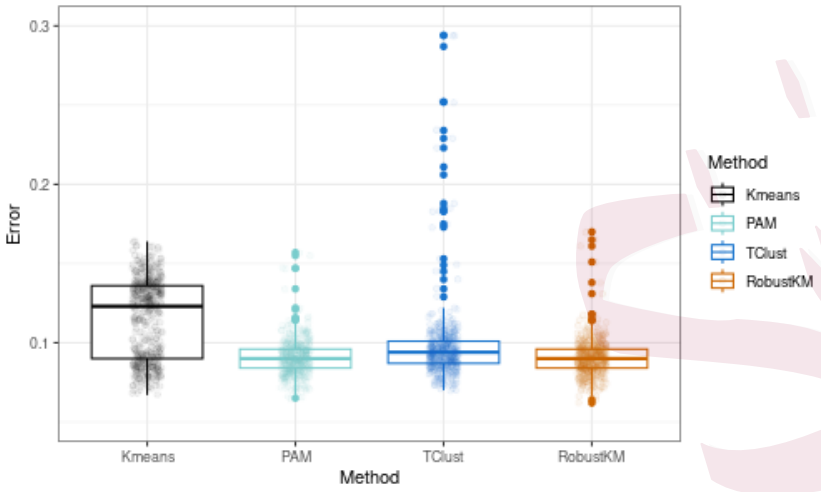


Figure 3: Box plot of classification errors, according to (4.2), of K-means, TClust, PAM and RobustKM over 1000 replicates.

4.2 Robust set estimation

Set estimation consists in determining a set, or a characteristic parameter of that set, based on a random sample of points. This set could represent various things, such as the support of a probability distribution (see for instance Rodríguez-Casal (2007)), its boundary (see, for instance Cuevas and Rodríguez-Casal (2004), the surface area of the boundary (see, for instance Aaron, Cholaquidis and Fraiman (2022)), or in applied contexts, the home-range of a specie (see, for instance Baíllo and Chacón (2021)), to

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

name a few.

Within this framework, shape constraints are usually imposed, convexity and r-convexity being some of the most used. Convexity can be restrictive for some applications, such as, for instance, if the set is the home-range of a species (see Cholaquidis, Fraiman, Mordecki and Papalardo (2021), Cholaquidis, Hernández and Fraiman (2023) and references therein), see also Section 5. Usually, the available data is an i.i.d. sample of a random vector whose support is the unknown set. For classic estimators such as the convex hull, r-convex hull, or cone-convex hull, any noise in the sample, no matter how small, drastically changes the estimators. This behavior is especifically pronounced in the convex hull case. In Section 5 we apply our robust aggregation strategy to tackle a home-range estimation problem, when the r-convex hull is employed.

4.2.1 Simulations

We show the performance of our robust proposal under a noisy model, where the aim is to estimate a convex set. More precisely, let $D(r, R)$ be the uniform distribution on the ring in $\mathbb{R}^2$ with inner radius r and outer radius R. We simulated 2000 i.i.d. observations of the mixture $(1-\lambda)D(0,1) + \lambda D(1, 1.25)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

The aim is to estimate $D(0,1)$ from this sample. We have chosen $\lambda = 0.01$ as the proportion of noise. The estimator, referred to as RChull, is the one proposed in Section 3, and we considered the Hausdorff distance between compact sets to measure the discrepancy between $D(0,1)$ and the estimator. To build our estimator, we first split the original sample at random into $K = 20$ disjoint groups of points of size 100, and compute the convex hull on each group. Then, we select from the K hulls $H_1, \ldots, H_{20}$, the hull H_{j^*}, with j^* as in (3.1).

To gain an insight into the improvement resulting from using this robust procedure, we show on the left of Figure 4 the set to be estimated, the classical estimator (the convex hull (Chull) of the whole sample) and the 20 hulls of the subsamples. The right panel shows the convex hull of the whole sample, together with H_{j^*}.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

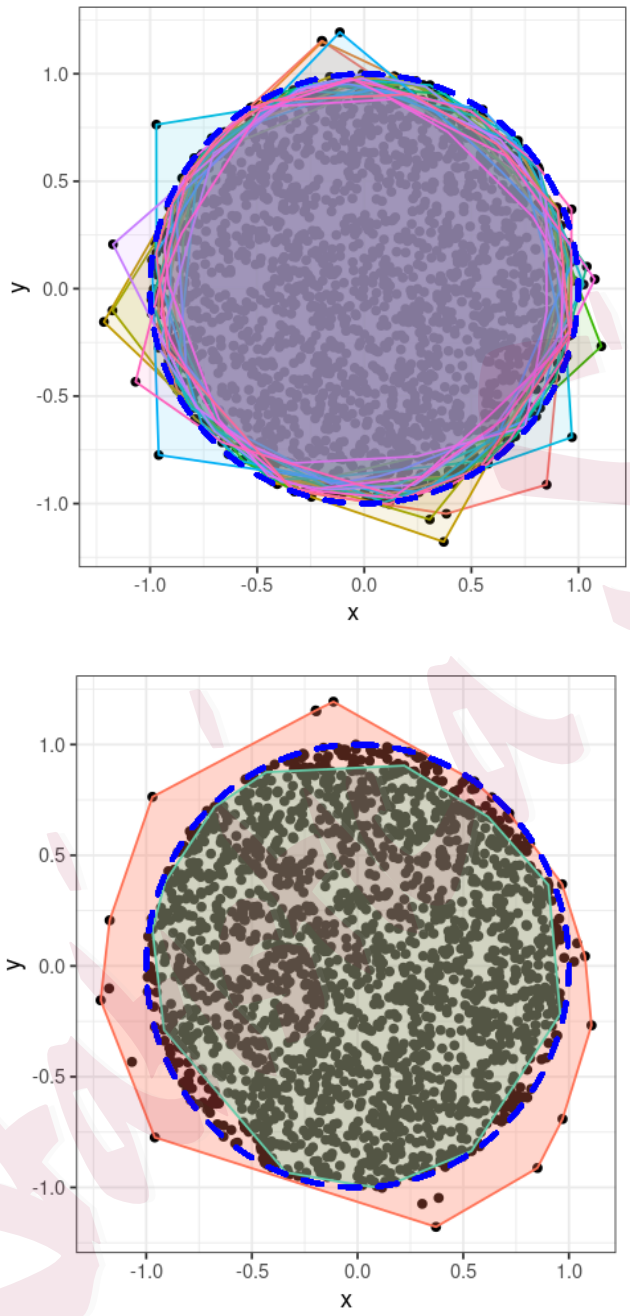


Figure 4: In the blue dotted line the boundary of the ball of radius 1. The sample is shown as solid black points. Outside this ball the sample generated from $D(1, 1.25)$. On the top panel, we show the 20 convex hulls of the selected subsamples (of size 100). On the bottom panel, we show the convex hull of the whole sample (Chull, coral color) and the robust estimator based on the 20 convex hulls (RChull, green color).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

To evaluate the performance of GROS , we run 100 replicates and estimate the set by both methods (the classical convex hull and our robust proposal) in each replicate. We calculate the Hausdorff distance of the estimated sets RChull and Chull from the set D(0,1). In Figure 5 box plots of these distances are shown. The RChull estimator outperforms the classical Chull estimator, but as can be seen, it has larger variability.

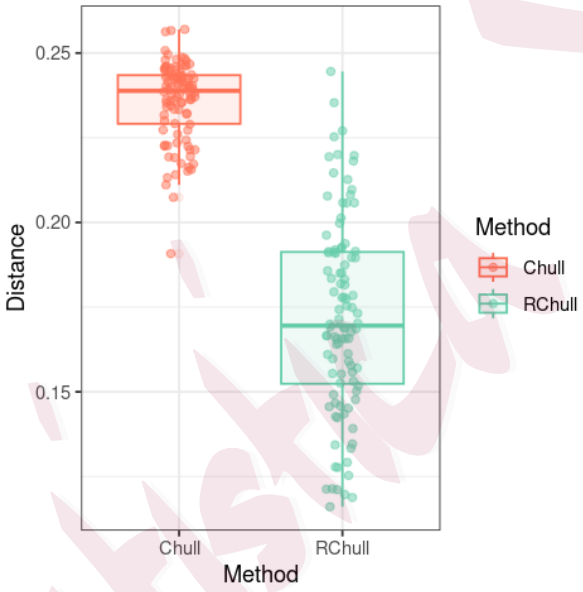


Figure 5

4.3 Robust persistent diagram

The robustness of persistent homology to perturbations in data measured by the Hausdorff distance is well established. However, it has a high sensitivity

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

to outliers, as discussed in Vishwanath, Fukumizu, Kuriki and Sriperumbudur (2020); Vishwanath, Sriperumbudur, Fukumizu and Kuriki (2022).

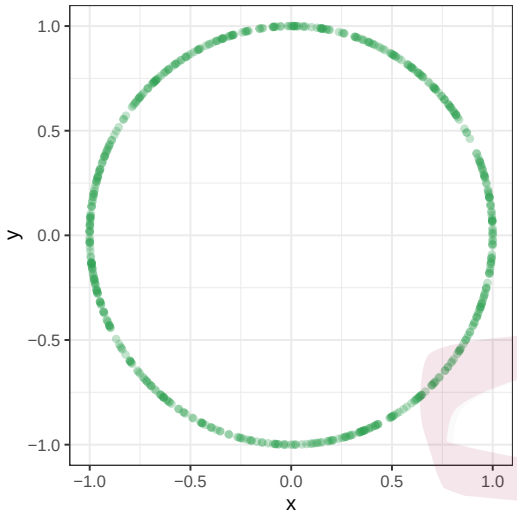
In this section, we introduce a robust persistence diagram, which we call the *Robust Wasserstein Estimator*, using (3.1).

The measure of dissimilarity between two persistence diagrams P_1 and P_2 is quantified by the 1-Wasserstein distance $W_1(P_1, P_2)$. This distance quantifies the cost associated with achieving the optimal alignment of points between the two diagrams, as detailed in (Edelsbrunner and Harer, 2022, p. 202).

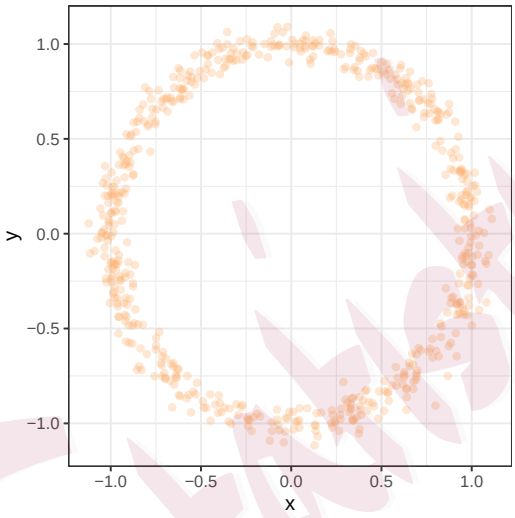
4.3.1 Simulations

The example examines uniformly simulated data on S^1 consisting of 600 points (baseline sample). It explores two potential scenarios of sample distortion, as depicted in Figure 6.

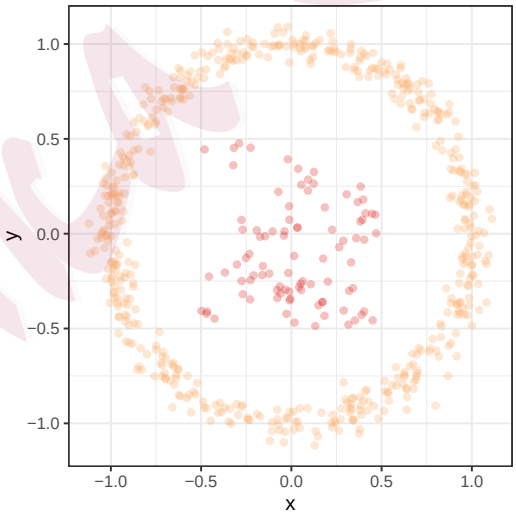
Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414



(a) Baseline sample



(b) Scenario 1: Local perturbation



(c) Scenario 2: Groups of outliers

Figure 6: (a) Baseline sample of 600 points uniformly distributed on S^1. (b) Locally perturbed sample as described in Scenario 1. (c) Sample perturbed in accordance with Scenario 2.

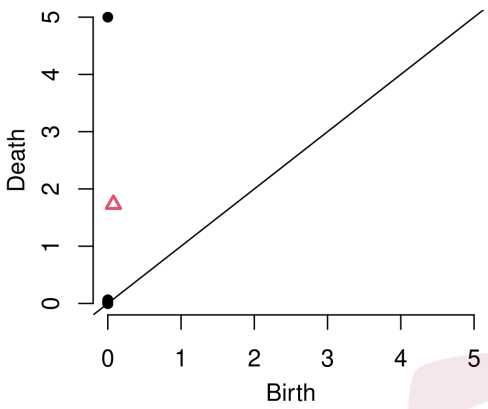
Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

- **Scenario 1:** Local perturbation. The original sample is perturbed using Gaussian noise centered at each sample point, with a standard deviation matrix $0.05 \times Id$.

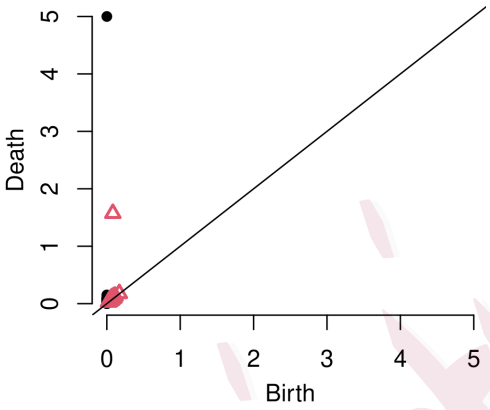
- **Scenario 2:** Group of Outliers: We randomly selected 90% of the perturbed sample as defined in Scenario 1. The remaining 10% are derived from a Matérn cluster process within the square region $[-0.5, 0.5]^2$. This process is characterized by an intensity of 3 for the Poisson process of cluster centers, a scale of 0.25, and an average of 20 points per cluster.

The persistence diagrams for the previously described tree samples were computed. Figure 7 displays these diagrams, where Dgm, Dgm_1, and Dgm_2 represent the persistence diagrams of the baseline sample, the locally perturbed sample (Scenario 1), and the sample with outliers (Scenario 2), respectively. It is evident that the persistence diagram is more distorted in scenarios with groups of outliers than it is with those with local perturbations.

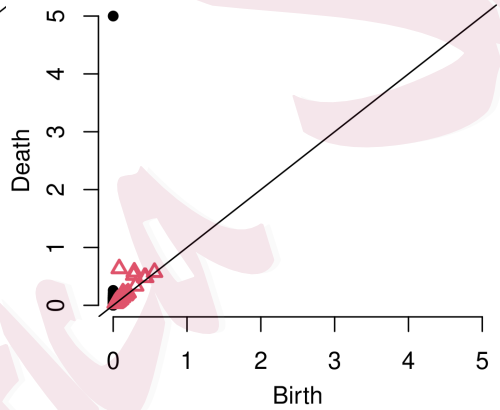
Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414



(a) Dgm



(b) Dgm_1



(c) Dgm_2

Figure 7: (a): Persistence diagram of the baseline sample. (b) and (c): Persistence diagrams for the samples perturbed according to Scenarios 1 and 2, respectively.

To evaluate the performance of our robust proposal (3.1), the sample is divided into $K = 6$ distinct groups. The robust persistence diagrams

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

for Scenarios 1 and 2 are labeled as $RDgm_1$ and $RDgm_2$, respectively. We computed $W_1(Dgm_1, Dgm)$, $W_1(Dgm_2, Dgm)$, $W_1(RDgm_1, Dgm)$, and $W_1(RDgm_2, Dgm)$.

A total of 100 independent iterations of this experiment were conducted, and box plots of the respective distances are shown in Figure 8. The results indicate that in both scenarios, the robust estimates of the persistence diagrams show improved performance, being closer to the baseline diagram in terms of the Wasserstein distance.

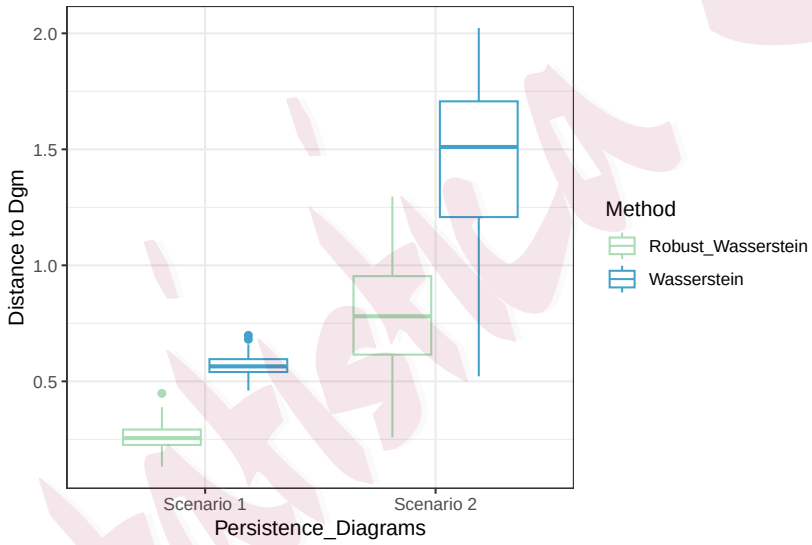


Figure 8: Box plots illustrating the distances between the persistence diagrams of perturbed samples and the baseline sample diagram Dgm, as well as the distances between the robust persistence diagrams and the baseline sample diagram.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

5. A real data study on robust set estimation

The concept of home-range of a specie was initially posed by Burt in 1943, see Burt (1943). It refers to the area that an individual explores while performing essential activities like feeding, mating, and caring for its offspring. Estimating this range is a crucial task in ecology and has been a subject of extensive research. Early methods used the convex hull of observed locations to estimate the home range, but this often led to overestimations. Alternative approaches, such as the r-convex hull and local hulls, have been proposed to address this issue, see Cholaquidis, Fraiman, Mordecki and Papalardo (2021); Cholaquidis, Hernández and Fraiman (2023), and references therein. Baíllo and Chacón (2021) offers a detailed review of various set estimation techniques applied to these problems.

In Smith et al. (2019), the home-range of the "native apex predator", the dingo (Canis dingo) is studied: " The research took place within the Matuwa Indigenous Protected Area (IPA) and nearby lands in central Western Australia, approximately 180 km east-northeast of Wiluna (26.23°S, 121.56°E; see Figure 9). Matuwa IPA, a former pastoral lease covering 2,410 km^2, has been co-managed as a conservation reserve by the Wiluna Aboriginal community and the Western Australia Department of Biodiversity, Conservation, and Attractions (formerly known as Parks and Wildlife)

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

since 2000."

The dingoes were captured using traps between June 2013 and June 2014. A total of 26 dingoes, selected based on appropriate weight, were fitted with GPS collars for 70 days, with locations recorded every 2 hours. A more detailed description of the data collection process is provided in Wysong et al. (2020). In total, the dataset contains 51,365 records of dingo locations. The data is open and available at `https://www.movebank.org`.

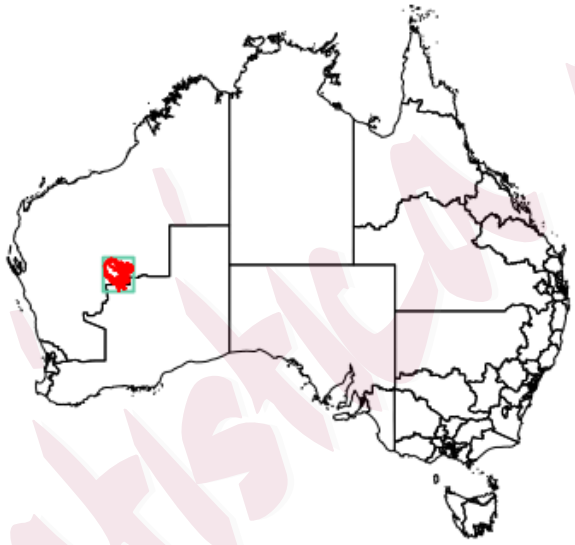


Figure 9: Area where the home-range of the predator Dingo was studied.

The robust estimation was carried out in the same manner as in Section 4.2, but instead of considering convex hulls, r-convex hull estimators were used. In this case, $K = 20$ disjoint groups of points were constructed to

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

perform the robust step. For both the classical and robust r-convex hull estimation, the parameter r was set to 0.05.

The records as well as the home-range estimators by both methods are represented in Figure 10. It can be seen that the classical estimator is influenced by some erratic movements of a few dingoes outside their natural habitat. In contrast, the robust estimation more efficiently delineates the areas where these animals most frequently move, being less affected by these atypical trajectories.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

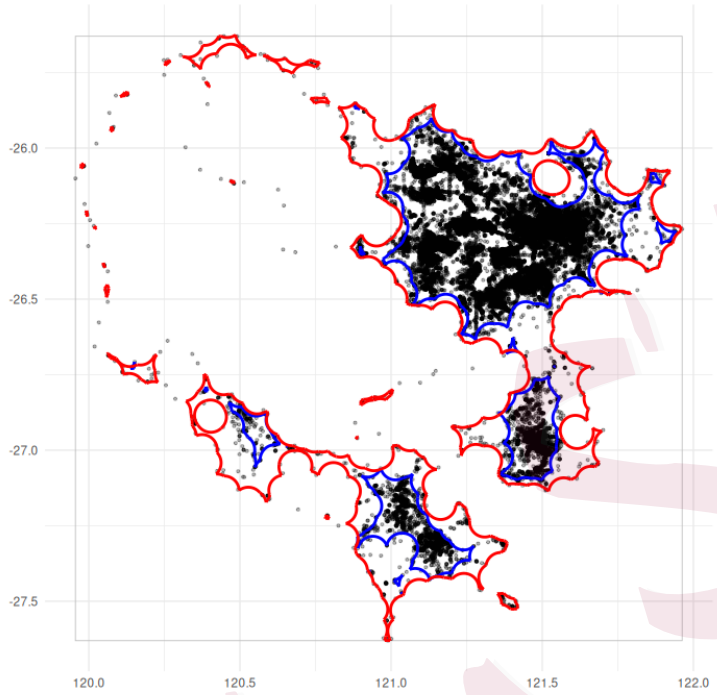


Figure 10: r-convex hulls of the home-range of the predator Dingo. Black points indicate the positions of different animals. The red and blue solid lines are the boundary of the r-convex hull and the robust r-convex hull, respectively.

6. Concluding remarks

We have demonstrated through simulations that GROS , which is applicable to a broad range of problems, significantly improves upon the non-robust, problem-specific solutions for each of the five examples treated. It is

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

also competitive with robust solutions designed for each specific case, even showing some improvements. GROS proposal's flexibility makes it applicable to a wide variety of problems, including those already presented, as well as any other scenario where robustness plays a crucial role. Furthermore, more examples using only pseudo-distances may be of interest for future research.

7. Proofs

Proof of Lemma 1. Let $\nu^l \in \mathcal{M}$ such that $m^l := d(\nu^l, \nu^l_{(\lfloor K/2 \rfloor+1)-\mathrm{NN}}) \downarrow \inf_{\nu\in\mathcal{M}} d(\nu, \nu_{(\lfloor K/2 \rfloor+1)-\mathrm{NN}}) =: L$. Then, there exists $\mu_j \in \{\mu_1, \ldots, \mu_K\}$ such that $\nu^l_{(\lfloor K/2 \rfloor+1)-\mathrm{NN}}) = \mu_j$ for infinitely many values of l. For ease of writing we denote this subsequence as ν^l. Since the closed balls are compact and $\nu^l \in B(\mu_j, L+\epsilon)$ for all l large enough, there exists $\nu \in \mathcal{M}$ such that a subsequence of $\nu^l \to \nu$ (we denote the subsequence by ν^l). $d(\nu, \mu_j) = L$. If $|B(\nu, d(\nu, \mu_j)) \cap \aleph_n| > (\lfloor K/2 \rfloor + 1)$ then $d(\nu, \nu_{(\lfloor K/2 \rfloor+1)}) \leq L$. If it were $d(\nu, \nu_{(\lfloor K/2 \rfloor+1)}) = L$ the lemma is proved. The other possibility is $d(\nu, \nu_{(\lfloor K/2 \rfloor+1)}) < L$ which contradict the definition of L. Let us consider the case $|B(\nu, d(\nu, \mu_j)) \cap \aleph_n| \leq (\lfloor K/2 \rfloor + 1)$. If $|B(\nu, d(\nu, \mu_j)) \cap \aleph_n| = (\lfloor K/2 \rfloor + 1)$ then $\mu_j = \nu_{(\lfloor K/2 \rfloor+1)-\mathrm{NN}}$ and again the lemma is proved. The last case is $|B(\nu, d(\nu, \mu_j)) \cap \aleph_n| < (\lfloor K/2 \rfloor + 1)$. But then $d(\nu, \nu_{\lfloor K/2 \rfloor+1)}) > d(\nu, \mu_j)$. Let

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

$\epsilon < (1/2)(d(\nu, \nu_{\lfloor K/2 \rfloor+1}) - d(\nu, \mu_j))$. Then $B(\nu, d(\nu, \nu_{\lfloor K/2 \rfloor+1}) - \epsilon) \cap \aleph_n| < \lfloor K/2 \rfloor + 1$. But, for l large enough $B(\nu^l, d(\nu^l, \mu_j)) \subset B(\nu, B(\nu, d(\nu, \nu_{\lfloor K/2 \rfloor+1}) - \epsilon))$ and, by definition of μ_j $|B(\nu^l, d(\nu^l, \mu_j)) \cap \aleph_n| = \lfloor K/2 \rfloor + 1$ which contradict $|B(\nu, d(\nu, \mu_j)) \cap \aleph_n| < (\lfloor K/2 \rfloor + 1)$.

□

Proof of Lemma 2. By hypothesis, there exists a set I of cardinality greater than $K/2$ such that $\max_{j \in I} d(\mu_j, \eta) \leq t$. Since μ^* is a minimizer of (2.1), there exists a set I_0 (a priori different from I) with $|I_0| > K/2$ such that $\max_{j \in I_0} d(\mu_j, \mu^*) \leq t$. Now, note that $|I| + |I_0| > K$ and so there exists $j \in I \cap I_0$ such that $d(\mu^*, \eta) \leq d(\mu^*, \mu_j) + d(\mu_j, \eta) \leq 2t$, which concludes the proof. □

Proof of Lemma 4. By the triangle inequality $\forall \delta > 0$, there exists an $\eta_\delta \in \widetilde{\mathcal{M}}$ such that $\max_{j \in I} d(\mu_j, \eta_\delta) \leq t + \epsilon + \delta$. So there exists a set I_0 of cardinality $|I_0| > K/2$, such that $\max_{j \in I_0} d(\mu_j, \widetilde{\mu}) \leq t + \epsilon + \delta$. Since $|I| + |I_0| > K$, there exists $j_0 \in I \cap I_0$. Then, $d(\widetilde{\mu}, \eta) \leq d(\widetilde{\mu}, \mu_{j_0}) + d(\mu_{j_0}, \eta) \leq 2t + \epsilon + \delta$. Since this holds for all $\delta > 0$, it follows that $d(\widetilde{\mu}, \eta) < 2t + \epsilon$. □

Proof of Lemma 5. Let $I \subset [K]$ be such that $|I| > K/2$, and for all $j \in I$, $d(\mu, \mu_j) \leq t$, we have $d(\mu_i, \mu_j) \leq d(\mu_i, \mu) + d(\mu, \mu_j) \leq 2t$ for all $i, j \in I$. Then, there exists an I_0 with cardinality greater than $K/2$ such that

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

$d(\mu_{j^*}, \mu_i) \leq 2t$ for all $i \in I_0$. Since $|I|+|I_0|> K$, there exists $j_0 \in I \cap I_0$. Lastly, $d(\mu_{j^*}, \mu) \leq d(\mu_{j^*}, \mu_{j_0}) + d(\mu_{j_0}, \mu) \leq 3t$. □

Supplementary Material

This document provides complementary material to the paper *GROS: A General Robust Aggregation Strategy*. We present two additional application settings illustrating how the GROS principle can be used to build robust procedures under heavy-tailed noise. In Section 1 we adapt the robust aggregated mean estimator (as in Eq. (1) of the main paper) to obtain a robust UCB-type strategy for multi-armed bandits with heavy-tailed rewards, and we include a small simulation study. In Section 2 we apply GROS to nonparametric regression by aggregating kernel estimators computed on independent subsamples, discuss a practical approximation of the L_2-distance used by the method, and compare the resulting estimator with classical and robust alternatives through simulations.

Acknowledgement

We warmly thank the two anonymous referees for their thoughtful comments and valuable suggestions, which have significantly improved the quality of the paper. We also thank the Editors for their careful handling of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

the manuscript and for their constructive guidance throughout the review process. The research of the first and third authors has been partially supported by grant FCE-3-2022-1-172289 from ANII (Uruguay) and grant 22520220100031UD from CSIC (Uruguay).

References

Aaron, C., Cholaquidis, A. and Fraiman, R. (2022). Estimation of surface area. *Electron. J. Statist.* **16**(2), 3751–3788.

Agrawal, R. (1995). Sample mean based index policies by $o(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability* **27**(4), 1054–1078.

Azzalini, A. (2013). *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Baíllo, A. and Chacón, J. E. (2021). Statistical outline of animal home ranges: an application of set estimation. *Handbook of Statistics* **44**, 3–37.

Biau, G., Fischer, A., Guedj, B. and Malley, J. D. (2016). COBRA: A combined regression strategy. *Journal of Multivariate Analysis* **146**, 18–28.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

Boente, G., Martínez, A. and Salibián-Barrera, M. (2017). Robust estimators for additive models using backfitting. *Journal of Nonparametric Statistics* **29**(4), 744–767.

Boursier, E. and Perchet, V. (2022). A survey on multi-player bandits. *arXiv preprint arXiv:2211.16275*.

Breiman, L. (1996). Stacked regressions. *Machine Learning* **24**, 49–64.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory* **59**(11), 7711–7717.

Burtini, G., Loeppky, J. and Lawrence, R. (2015). A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*.

Burt, W. H. (1943). Territoriality and home range concepts as applied to mammals. *Journal of Mammalogy* **24**(3), 346–352.

Cholaquidis, A., Fraiman, R., Ghattas, B. and Kalemkerian, J. (2021). A combined strategy for multivariate density estimation. *Journal of Nonparametric Statistics* **33**(1), 39–59.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

Cholaquidis, A., Fraiman, R., Kalemkerian, J. and Llop, P. (2016). A nonlinear aggregation type classifier. *Journal of Multivariate Analysis* **146**, 269–281.

Cholaquidis, A., Fraiman, R., Mordecki, E. and Papalardo, C. (2021). Level set and drift estimation from a reflected Brownian motion with drift. *Statistica Sinica* **31**, 29–51.

Cholaquidis, A., Hernández, M. and Fraiman, R. (2023). Home range estimation under a restricted sampling scheme. *Journal of Nonparametric Statistics*, to appear.

Cuesta-Albertos, J. A., Gordaliza, A. and Matrán, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.* **25**(2), 553–576.

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Advances in Applied Probability* **36**(2), 340–354.

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

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Harvard University, Boston.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Edelsbrunner, H. and Harer, J. L. (2022). *Computational Topology: An Introduction*. American Mathematical Society.

Fernández, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* **93**(441), 359–371.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**(1), 119–139.

Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-free Theory of Nonparametric Regression*. Springer.

Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria. *Ann. Statist.* **6**, 117–131.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 361–379.

Joly, E., Lugosi, G. and Oliveira, R. I. (2017). On the estimation of the mean of a random vector. *Electron. J. Statist.* **11**(1), 440–451.

Kaufman, L. (1990). Partitioning Around Medoids. In: *Finding Groups in Data*, 344:68–125.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

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

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics* **19**(5), 1145–1190.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

Maronna, R. A., Martin, R. D., Yohai, V. J. and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. Wiley.

McQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, University of California Press, 281–297.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications* **9**(1), 141–142.

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience.

Oh, H.-S., Nychka, D. W. and Lee, T. C. M. (2007). The role of pseudo data for robust smoothing with application to wavelet regression. *Biometrika* **94**(4), 893–904.

Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Probability* **9**(1), 135–140.

Pollard, D. (1982). A central limit theorem for k-means clustering. *Ann. Probability* **10**(4), 919–926.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

Rodriguez, D. and Valdora, M. (2019). The breakdown point of the median of means tournament. *Statistics & Probability Letters* **153**, 108–112.

Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions. *Ann. IHP Probab. Stat.* **43**(6), 763–774.

Salibián-Barrera, M. (2023). Robust nonparametric regression: Review and practical considerations. *Econometrics and Statistics*,in press (corrected proof), available online 25 April 2023. doi: 10.1016/j.ecosta.2023.04.004.

Smith, B. P., Cairns, K. M., Adams, J. W., Newsome, T. M., Fillios, M., Deaux, E. C. *et al.* (2019). Taxonomic status of the Australian dingo: the case for *Canis dingo* Meyer, 1793. *Zootaxa* **4564**(1), 173–197.

Vishwanath, S., Fukumizu, K., Kuriki, S. and Sriperumbudur, B. K. (2020). Robust persistence diagrams using reproducing kernels. *Advances in Neural Information Processing Systems* **33**, 21900–21911.

Vishwanath, S., Sriperumbudur, B. K., Fukumizu, K. and Kuriki, S. (2022). Robust topological inference in the presence of outliers. *arXiv preprint arXiv:2206.01795*.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* **26**(4), 359–372.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0414

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks* **5**(2), 241–259.

Wysong, M. L., Hradsky, B. A., Iacona, G. D., Valentine, L. E., Morris, K. and Ritchie, E. G. (2020). Space use and habitat selection of an invasive mesopredator and sympatric, native apex predator. *Movement Ecology* **8**, 1–115.

Statistica Sinica