Accepted Manuscript

Multivariate and functional robust fusion methods for structured Big Data

Catherine Aaron, Alejandro Cholaquidis, Ricardo Fraiman, Badih Ghattas

PII: S0047-259X(17)30689-9
DOI: https://doi.org/10.1016/j.jmva.2018.06.012
Reference: YJMVA 4384

To appear in: *Journal of Multivariate Analysis*

Received date : 20 November 2017

Please cite this article as: C. Aaron, A. Cholaquidis, R. Fraiman, B. Ghattas, Multivariate and functional robust fusion methods for structured Big Data, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.06.012

Multivariate and functional robust fusion methods for structured Big Data

Catherine Aaron,[a,1], Alejandro Cholaquidis[b,2], Ricardo Fraiman[b,c,3], Badih Ghattas[d,4]

[a] *Université Clermont Auvergne, Campus universitaire des Cézeaux, France*
[b] *CABIDA and Centro de Matemática, Universidad de la República, Uruguay*
[c] *Instituto Pasteur de Montevideo, Uruguay*
[d] *Aix Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, 13453, Marseille, France*

Abstract

We address one of the important problems in Big Data, namely how to combine estimators from different subsamples by robust fusion procedures, when we are unable to deal with the whole sample. We propose a general framework based on the classic idea of 'divide and conquer'. In particular we address in some detail the case of a multivariate location and scatter matrix, the covariance operator for functional data, and clustering problems.

Key words: Big Data, Clustering, Functional Data, Robustness

1. Introduction

Big Data has emerged in recent years in various contexts such as social networks, biochemistry, health care systems, politics, and retail, among many others. New developments are necessary to address many of the related issues. Typically, classical statistical approaches that perform reasonably well for small data sets fail when dealing with huge data sets. To handle these challenges, new mathematical and computational methods are needed.

The challenges posed by Big Data cover a wide range of problems, and have been recently considered in abundant literature; see, e.g., [1, 22, 23] and references therein. We address one of these problems, namely, how to combine, using robust techniques, estimators obtained from different subsamples in the case where we are computationally unable to deal with the whole sample. In what follows, we will refer to such approaches as robust fusion methods (RFM). Specifically, this paper proposes a general algorithm which, in spirit, is related with the well-known idea of divide-and-combine. We consider the case where the data belong to finite- and infinite-dimensional spaces.

Functional Data Analysis (FDA) has become a central area of statistics in recent years, having gained much momentum from the work of Ramsay and his collaborators in the early 2000s. Since then, both the quantity and quality of its results have enjoyed a marked growth, while addressing a great diversity of problems. FDA faces several specific challenges, most of them associated with the infinite-dimensional nature of the data. Essential recent references for FDA include [3, 11, 12, 15], as well as the recent surveys [9] and [21].

Divide-and-combine is a well-known technique for dealing with huge data sets; see, e.g., [2]. In the FDA setting, it has also been considered recently in [20] for linear regression problems with the Lasso, a problem that is not addressed here. In the present paper, we focus on a general robust procedure for different problems. The consistency and robustness of our method is studied in the general setting of FDA, and we apply our proposed algorithm to some statistical problems in finite- and infinite-dimensional settings, namely, the location and scatter matrix, clustering, and impartial trimmed k-means. Also, a new robust estimator of the covariance operator is proposed.

We start by describing one of the simplest problems in this area as a toy example. Suppose we are interested in the median of a huge set of iid random variables $X_1, \ldots, X_n$ with common density f_X, and we split the sample into m

[1] catherine.aaron@math.univ-bpclermont.fr
[2] acholaquidis@cmat.edu.uy.
[3] rfraiman@cmat.edu.uy.
[4] badihghattas@gmail.com

subsamples of size ℓ, so that $n = m\ell$. We compute the median of each subsample, resulting in m random variables $Y_1, \ldots, Y_m$. Then we take the median of the set $Y_1, \ldots, Y_m$, i.e., we consider the well-known median of medians, which, in this case, will be our RFM estimator. This estimator clearly does not coincide with the median of the whole original sample $X_1, \ldots, X_n$, but it will be close. What can be said regarding its efficiency and robustness of this estimator?

In this particular case, the RFM estimator is nothing but the median of m iid random variables, but now with a different distribution, given by the distribution of the median of ℓ random variables with density f_X. Suppose for simplicity that $\ell = 2k + 1$. Then, the density of each random variable $Y_1, \ldots, Y_m$ is given by

$$g_Y(t) = \frac{(2k+1)!}{(k!)^2} F_X(t)^k \{1 - F_X(t)\}^k f_X(t).$$

On the one hand, if $f_X\{F_X^{-1}(0.5)\} \neq 0$, the empirical median $\hat{\theta} = \text{med}(X_1, \ldots, X_n)$ has, asymptotically, a normal distribution centered at the true median θ with variance $\text{var}(\hat{\theta}) = 1/\{4nf_X(\theta)^2\}$. On the other hand, the distribution of $\tilde{\theta}^{RFM}$, the median of medians, is asymptotically normal and centered at θ with variance $\text{var}(\tilde{\theta}^{RFM}) = 1/\{4mg_Y(\theta)^2\}$, where $g_Y(\theta) = (1/2)^{2k}(2k+1)!/(k!^2)f_X(\theta) \sim \sqrt{2k/\pi}$. So in this case we can explicitly compute the asymptotic relative loss of efficiency, i.e., $\lim_{n\to\infty} \text{var}(\hat{\theta})/\text{var}(\tilde{\theta}^{RFM}) = 2/\pi \approx 0.6366$.

In Section 2 we generalize this RFM idea and study its consistency, robustness, breakdown point, and efficiency. In Section 3, we show how the RFM can be applied to multivariate location and scatter matrix estimation, covariance operator estimation for functional data, and robust clustering. Section 4 reports simulation results for these problems.

2. A general setup for RFM

We start by introducing a general framework for RFM. The idea is quite simple: given a sample $X_1, \ldots, X_n$ of iid random elements in a metric space E (e.g., $E = \mathbb{R}^d$) and a statistical problem (such as multivariate location, covariance operators, linear regression, or principal components, among many others), we split the sample into m subsamples of equal size. For each subsample we compute a robust solution for the statistical problem considered. The solution given by RFM corresponds to the deepest point among the m solutions (in terms of the appropriate norm associated to the problem) obtained from the subsamples.

In order to introduce the notion of depth, we will use throughout this paper the following notation. Let X be a random variable taking values in some Banach space $(E, \|\cdot\|)$, with probability distribution P_X, and let $x \in E$. The depth of x with respect to P_X is defined by

$$D(x, P_X) = 1 - \left\| \text{E}_{P_X}\left(\frac{X - x}{\|X - x\|} \right) \right\|. \tag{1}$$

It was introduced in [6], formulated (in a different way) in [21], and extended to a very general setup in [5].

Given a sample $X_1, \ldots, X_n$, let us write P_n for the empirical measure. The empirical version of (1) is

$$D(x, P_n) = 1 - \left\| \text{E}_{P_n}\left(\frac{X - x}{\|X - x\|} \right) \right\| = 1 - \frac{1}{n} \left\| \sum_{i=1}^{n} \frac{X_i - x}{\|X_i - x\|} \right\|. \tag{2}$$

Although we suggest using the depth function, we point out that this is not always suitable, e.g., in clustering. In such cases, the deepest point may be replaced by other robust estimators, as we will show in Section 3.3. We summarize our approach in Table 1 for a general framework of parameter estimation. This may be easily applied to any situation where robust estimators exist or can be designed.

We will address the consistency, efficiency, robustness, and computational time of the RFM proposals.

2.1. Consistency, robustness and breakdown point of the RFM

We start by proving that, given a sample $X_1, \ldots, X_n$ of a random element X, its deepest point (i.e., the value that maximizes (2)) converges almost surely to the value that maximizes (1). Although similar results have already been obtained, e.g., in [5], we will need this fact when P_n is not necessarily the empirical measure associated to a sample, but any measure converging weakly to a probability distribution P. We will need the following assumption.

Table 1: Parameter estimation using RFM

$X_1, \ldots, X_n$: iid random elements in a Banach space E; θ_0: a parameter to estimate
a) Split the sample into m subsamples with $n = m\ell$, viz. $\{X_1, \ldots, X_\ell\}, \{X_{\ell+1}, \ldots, X_{2\ell}\}, \ldots, \{X_{(m-1)\ell+1}, \ldots, X_{\ell m}\}$.
b) Compute a robust estimate of θ_0 on each subsample, yielding $\hat{\theta}_1, \ldots, \hat{\theta}_m$.
c) Compute the final estimate $\tilde{\theta}^{RFM}$ by RFM combining $\hat{\theta}_1, \ldots, \hat{\theta}_m$ by a robust approach.
For instance, $\tilde{\theta}^{RFM}$ can be the deepest point, or the average of 40% of the deepest points among the $\hat{\theta}_1, \ldots, \hat{\theta}_m$.

Assumption H1: A probability measure P defined on a separable Hilbert space $\mathcal{H}$ fulfils H1 if $P\{\partial B(y, r)\} = 0$ for all $r > 0$ and $y \in \mathcal{H}$, where ∂A stands for the boundary of a set $A \subset \mathcal{H}$.

Observe that H1 is fulfilled if the random variable $\|X - y\|$ is absolutely continuous for every $y \in \mathcal{H}$, where X is a random variable with distribution P.

Theorem 1. *Let $X_1, X_2, \ldots$ be a sequence of random elements with common distribution P_n, defined in a separable Hilbert space $(\mathcal{H}, \|\cdot\|)$. Let P be a probability distribution fulfilling H1. Assume that $P_n \to P$ weakly, and $\|\mathrm{E}_P\{(X - x)/\|X - x\|\}\|$ has a unique minimum. Then, as $n \to \infty$,*

$$\arg\max_x \ D(x, P_n) \to \arg\max_x \ D(x, P) \quad a.s. \tag{3}$$

In order to prove (3) we will use the following fundamental result proved by Billingsley and Topsøe [4] and which still holds when $\mathcal{H}$ is a separable Banach space; see Theorem 1 and Example 3.

Theorem 2. *Suppose $S \subset \mathcal{H}$ and let $\mathcal{B}(S, \mathcal{H})$ be the class of all bounded measurable functions mapping S into $\mathcal{H}$. Suppose $\mathcal{F} \subset \mathcal{B}(S, \mathcal{H})$ is a subclass of functions. Then*

$$\sup_{f \in \mathcal{F}} \left\| \int f dP_n - \int f dP \right\| \to 0,$$

for every sequence P_n that converges weakly to P if, and only if,

$$\sup\{\|f(z) - f(t)\| : f \in \mathcal{F}, z, t \in S\} < \infty,$$

and for all $\epsilon > 0$,

$$\lim_{\delta \to 0} \sup_{f \in \mathcal{F}} P[\{x : \omega_f\{B(x, \delta)\} \geq \epsilon\}] = 0, \tag{4}$$

where $\omega_f(A) = \sup\{|f(x) - f(y)| : x, y \in A\}$ and $B(x, \delta)$ is the open ball of radius $\delta > 0$.

Proof of Theorem 1. Consider $S = \mathcal{H}$ and $\mathcal{F}$ the subclass $\{f_y : y \in \mathcal{H}\}$ of functions such that $f_y(z) = (z - y)/\|z - y\|$. Then $\sup\{\|f_y(z) - f_y(t)\| : y, x, z \in \mathcal{H}\} \leq 2$. Let $2\sqrt{\delta} < \epsilon$. Then, for all y,

$$\{x : \omega_{f_y}\{B(x, \delta)\} > \epsilon\} = \{x \in B(y, \sqrt{\delta}) : \omega_{f_y}\{B(x, \delta)\} > \epsilon\} \cup \{x \notin B(y, \sqrt{\delta}) : \omega_{f_y}\{B(x, \delta)\} > \epsilon\}.$$

Observe that $\omega_{f_y}\{B(x, \delta)\} = 2\delta/\|x - y\|$ if $\|x - y\| > \delta$. Thus if $x \notin B(y, \sqrt{\delta})$, then $\omega_{f_y}\{B(x, \delta)\} \leq 2\sqrt{\delta} < \epsilon$, and so $\{x \notin B(y, \sqrt{\delta}) : \omega_{f_y}\{B(x, \delta)\} > \epsilon\} = \emptyset$. Finally, we get that for all y,

$$\{x : \omega_{f_y}\{B(x, \delta)\} > \epsilon\} = \{x \in B(y, \sqrt{\delta}) : \omega_{f_y}\{B(x, \delta)\} > \epsilon\} \subset B(y, \sqrt{\delta}).$$

Now, since $P\{\partial B(y, \sqrt{\delta})\} = 0$ we have that $\mathbf{1}_{B(y_k, \sqrt{\delta})}(x) \to \mathbf{1}_{B(y, \sqrt{\delta})}(x)$ a.s. with respect to P, whenever $y_k \to y$ for every y. Lebesgue's Dominated Convergence Theorem then implies that

$$P\{B(y_k, \sqrt{\delta})\} \to P\{B(y, \sqrt{\delta})\}.$$

This entails that $P\{B(y, \sqrt{\delta})\}$ is a continuous function of y, so its maximum in a compact set, is attained.

Let $\epsilon > 0$ and K_ϵ be a compact set such that $P[\{K_\epsilon \ominus B(0,1)\}^{\complement}] < \epsilon$ where $K_\epsilon \ominus B(0,1) = \{z \in K_\epsilon : d(z, K_\epsilon^{\complement}) > 1\}$, where $K_\epsilon^{\complement}$ is the complement of the set K_ϵ. Let $y_{\epsilon,\delta} = \arg\max_{y \in K_\epsilon} P\{B(y, \sqrt{\delta})\}$. We will prove that for any fixed $\epsilon > 0$, $P\{B(y_{\epsilon,\delta}, \sqrt{\delta})\} \to 0$ as $\delta \to 0$. If this is not the case there exists $\eta > 0$, $y_n \in K_\epsilon$ and $\delta_n \to 0$ such that $P\{B(y_n, \sqrt{\delta_n})\} > \eta$ for all $n \in \mathbb{N}$. Since K_ϵ is compact we can assume that $y_n \to y$ for some $y \in K_\epsilon$ by considering a subsequence. From $P\{\partial B(x,r)\} = 0$ for all x, it follows that $P(\{y\}) = 0$; indeed, just consider x and $r > 0$ such that $y \in \partial B(x,r)$.

Define $\rho_n = \max_{j \geq n}(\sqrt{\delta_j} + \|y - y_j\|)$ and $B_n = B(y, \rho_n)$. Then $P(B_n) \geq \eta$ and $B_1 \supseteq \cdots \supseteq B_n \cdots$. As a result, $0 = P(\{y\}) = \lim P(B_n)$, which contradicts the fact that $P\{B(y_n, \sqrt{\delta_n})\} > \eta$. Now for all $\delta < 1$,

$$\sup_y P\{B(y, \sqrt{\delta})\} \leq \max\left[\sup_{y \in K_\epsilon} P\{B(y, \sqrt{\delta})\}, P[\{K_\epsilon \ominus B(0,1)\}^{\complement}]\right].$$

Therefore $\sup_y P\{B(y, \sqrt{\delta})\} \leq \max[P\{B(y_{\epsilon,\delta}, \sqrt{\delta})\}, \epsilon] < \epsilon$ for δ small enough, showing that (4) holds. Finally (3) is a consequence of the uniform convergence of $D(x, P_n)$ to $D(x, P)$ and the argmax argument. □

The following corollary states the consistency of the RFM explained in Table 1 when the sample $X_1, \ldots, X_n$ is distributed as a random variable X with distribution P_0 fulfilling H1.

Corollary 1. *Assume that P_0 fulfils H1 and that there exists a unique θ_0 such that, for all ℓ,*

$$\mathrm{E}_{P_0}\left(\frac{\hat{\theta}_1 - \theta_0}{\|\hat{\theta}_1 - \theta_0\|}\right) = 0.$$

Then, under P_0, $\tilde{\theta}^{RFM} \to \theta_0$ a.s., as $m \to \infty$.

Recall that a sequence of estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots$ is qualitatively robust at a probability distribution P if for all $\epsilon > 0$ there exists $\delta > 0$, for all probability distribution Q, $\Pi(P,Q) < \delta \Rightarrow \Pi\{\mathcal{L}_P(\hat{\theta}_n), \mathcal{L}_Q(\hat{\theta}_n)\} < \epsilon$ (see [14]), where Π denotes the Prokhorov distance and $\mathcal{L}_F(\hat{\theta}_n)$ denotes the probability distribution of $\hat{\theta}_n$ under F. As Π metrizes weak convergence, we have the following corollary.

Corollary 2. *Robustness of RFM estimators. Under the hypotheses of Corollary 1, $\tilde{\theta}^{RFM}$ is qualitatively robust.*

Remark 1. Qualitative robustness ensures the good behavior of the estimator in a neighborhood of P_0. However, there are some estimators that still converge to θ_0 even if P is far from P_0. For instance "the shorth", defined as the average of the observations lying in the shortest interval containing half of the data, has this property. Indeed, consider the case where $P_0 = U(-1,1)$, $P_1 = U(3,4)$, and $P = (1-\alpha)P_0 + \alpha P_1$, for any $\alpha < 0.5$. This is also the case for the impartial trimmed estimators, the minimum volume ellipsoid, and the redescendent (with compact support) M-estimators; see Section 2.2. If the estimators corresponding to each subsample have this property, the RFM estimator will inherit it.

2.2. Efficiency of the fusion of M-estimators

In this section we obtain the asymptotic variance of the RFM method, for the special case of M-estimators. Recall that an M-estimator T can be defined (see Section 3.2 in [17]) by the implicit functional equation $\int \psi\{x; T(F)\}F(dx) = 0$, where $\psi(x;\theta) = (\partial/\partial\theta)\rho(x;\theta)$ and F stands for the true underlying common distribution of the observations. For instance, the Maximum Likelihood estimator is obtained with $\rho(x;\theta) = -\ln\{f(x,\theta)\}$. The estimator T_n is given by the empirical version of T, based on a sample $X_1, \ldots, X_n$. It is well known that $\sqrt{n}\{T_n - T(F)\}$ is asymptotically normal with mean 0 and variance $A(F,T)$ given by the integral of the square of the influence curve, i.e., $A(F,T) = \int IC(x;F,T)^2 F(dx)$, where the influence curve, IC, is

$$IC(x;F,T) = \frac{\psi\{x;T(F)\}}{-\int(\partial/\partial\theta)\psi\{x;T(F)\}F(dx)}.$$

For the location problem (i.e., $\int \psi(x-\mu_0)F(dx) = 0$), we get $IC(x,F,T) = -\psi(x-\mu_0)/\int \psi'(x-\mu_0)F(dx)$. The asymptotic efficiency of T_n is defined as $\mathrm{Eff}(T_n) = \sigma^2_{ML}/A(F,T)$, where σ^2_{ML} is the asymptotic variance of the

maximum likelihood estimator. Then the asymptotic variance of an M-estimator built from a sample $T_n^1, \ldots, T_n^m$ of m M-estimators of T can be computed easily. The strong consistency of the M-estimators under the model (see [16]) entails that $\tilde{\theta}^{RFM}$ built from M-estimators is consistent (see Corollary 1) whenever the empirical version of the implicit functional equation has an unique solution.

The choice of m and ℓ has an impact on the robustness of the estimator and on the computation time. Indeed, if the computation time of each $\hat{\theta}_i = O(\ell^a)$ and the computation time of the fusion step is $O(m^b)$, then the optimal choice is $\ell = O\{n^{(b-1)/(a+b-1)}\}$ if $b > 1$.

2.3. Breakdown point for the RFM

Following [10] we consider the finite-sample breakdown point. Intuitively the breakdown point corresponds to the maximum percentage of outliers (located at the worst possible positions) we can have in a sample before the estimate breaks in the sense that it can be arbitrarily large (or close to the boundary of the parameter space).

Definition 1. *Let* $\mathbf{x} = \{x_1, \ldots, x_n\}$ *be a data set,* θ *an unknown parameter lying in a metric space* Θ*, and* $\hat{\theta}_n = \hat{\theta}_n(\mathbf{x})$ *an estimate based on* $\mathbf{x}$*. Let* $\mathcal{X}_p$ *be the set of all data sets* $\mathbf{y}$ *of size* n *having* $n - p$ *elements in common with* $\mathbf{x}$*, viz.*

$$\mathcal{X}_p = \{\mathbf{y} : \text{card}(\mathbf{y}) = n,\ \text{card}(\mathbf{x} \cap \mathbf{y}) = n - p\}.$$

Then the breakdown point of $\hat{\theta}_n$ *at* $\mathbf{x}$ *is* $\epsilon_n^*(\hat{\theta}_n, \mathbf{x}) = p^*/n$*, where* $p^* = \max\{p \geq 0 : \forall_{\mathbf{y} \in \mathcal{X}_p}\ \hat{\theta}_n(\mathbf{y})$ *is bounded and also bounded away from the boundary* $\partial\Theta$*, if* $\partial\Theta \neq \emptyset\}$.

To analyze the breakdown point of the RFM, we consider the case where the breakdown point of the robust estimators is 0.5 (high breakdown point estimators). For each observation X_i from the sample, let $B_i = 1$ if X_i is an outlier and 0 otherwise. Assume that the variables B_i are iid following a Bernoulli distribution with parameter p and let $S_j = \sum_{s=1}^{\ell} B_{(j-1)\ell+s}$ be the number of outliers in the subsample j, for $j \in \{1, \ldots, m\}$. The RFM estimator will break down if and only if there are more than $m/2$ cases where S_j is greater than k (recall that $\ell = 2k + 1$).

To take a glance of the behavior of the breakdown point, we performed 5000 replications where we generated $n = 30{,}000$ binomial random variables with parameter p. We split each of the samples of size 30,000 randomly into m subsamples. Next we computed the number of its subsamples which contained more than k values of 1 (outliers). In Table 2 we report the average number of times (over the 5000 replications) that this number was greater than $m/2$, for different values of p and m. The best result is obtained when $m = 5$.

3. Some applications of RFM

In this section we will show how RFM may be used to tackle three classic statistical problems for large samples: estimating the multivariate location and scatter matrix, estimating the covariance operator for functional data, and clustering. For each problem we show how to apply our approach, given in Table 1. Solutions for many other problems may be derived from these cases, e.g., Principal Components, both for non-functional and functional data.

Table 2: Average (over 5000 replications) of estimator breakdowns for different values of m and p and fixed $n = 30{,}000$; p is the proportion of outliers.

m	$p = 0.45$	$p = 0.49$	$p = 0.495$	$p = 0.499$
5	0	0.0020	0.0820	0.3892
10	0	0.0088	0.1564	0.5352
30	0	0.0052	0.1426	0.5186
50	0	0.0080	0.1598	0.5412
100	0	0.0192	0.2162	0.6084
150	0	0.0278	0.2728	0.6780

3.1. Robust fusion for location and scatter matrix in finite-dimensional spaces

Given an iid random sample $X_1, \ldots, X_n$ in $\mathbb{R}^d$, we consider the location and scatter matrix estimation problem. To perform RFM we only need to make explicit the estimators used for each of the m subsamples, and the depth function in the fusion stage. For the location parameters, we propose to use simple robust estimates, denoted by $\hat{\theta}_1, \ldots, \hat{\theta}_m$; see, e.g., [19].

For the depth function we propose to use the empirical version of (1), replacing P_X by the empirical distribution P_m of $\{\hat{\theta}_1, \ldots, \hat{\theta}_m\}$,

$$D(\theta, P_m) = 1 - \left\| \frac{1}{m} \sum_{j=1}^{m} \frac{\hat{\theta}_j - \theta}{\|\hat{\theta}_j - \theta\|} \right\|, \tag{5}$$

where $\theta \in \mathbb{R}^d$, and $\|\cdot\|$ is the Euclidean distance. Equivalently, for the scatter matrix we use the depth function

$$D(\Sigma, P_m) = 1 - \left\| \frac{1}{m} \sum_{j=1}^{m} \frac{\hat{\Sigma}_j - \Sigma}{\|\hat{\Sigma}_j - \Sigma\|} \right\|, \tag{6}$$

where $\hat{\Sigma}_1, \ldots, \hat{\Sigma}_m$ are robust estimators of the scatter matrix, the norm is $\|\Sigma\| = \max_{i \in \{1,\ldots,d\}} \sum_{j=1}^{d} |\Sigma_{ij}|$. P_m denotes the empirical distribution of $\hat{\Sigma}_1, \ldots, \hat{\Sigma}_m$. A simulation study is presented in Section 4.

3.2. Robust fusion for the covariance operator

The estimation of the covariance operator of a stochastic process is a very important topic in FDA, which helps to understand the fluctuations of a random element, as well as to derive the principal functional components from its spectrum. Several robust and non-robust estimators have been proposed; see, e.g., [5] and references therein. In order to perform RFM, we introduce a new robust estimator to use for each of the m subsamples, which can be implemented using parallel computing. It is based on the notion of impartial trimming in the Hilbert–Schmidt space where the covariance operators are defined. It was introduced in [13] and has been shown to be a very successful tool in robust estimation. Next, the RFM estimator is defined as the deepest point among the m estimators ('impartial trimmed means') corresponding to each subsample.

To better understand the construction of our new estimator, we first recall the general framework used for the estimation of covariance operators.

3.2.1. A general framework for the estimation of covariance operators

Let $E = L^2(I)$, where I is a finite interval in $\mathbb{R}$, and $X, X_1, X_2, \ldots$ be iid random elements taking values in E. Assume that $\mathrm{E}\{X(t)^2\} < \infty$ for all $t \in I$, and $\int_I \int_I \rho^2(s,t) ds dt < \infty$, so that the covariance function, given by $\rho(s,t) = \mathrm{E}[\{X(t) - \mu(t)\}\{X(s) - \mu(s)\}]$, where $\mathrm{E}\{X(t)\} = \mu(t)$, is well defined. For notational simplicity we assume that $\mu(t) = 0$ for all $t \in I$. Under these conditions, the covariance operator, given for all $f \in E$ by

$$\Gamma_0(f)(t) = \mathrm{E}\{\langle X, f \rangle X(t)\} = \int_I \rho(s,t) f(s) ds,$$

is diagonalizable, with non-negative eigenvalues λ_i such that $\sum_i \lambda_i^2 < \infty$. Moreover Γ_0 belongs to the Hilbert–Schmidt space $HS(E)$ of linear operators with norm and inner product given by

$$\|\Gamma\|_{HS}^2 = \sum_{k=1}^{\infty} \|\Gamma(e_k)\|^2 < \infty, \quad \langle \Gamma_1, \Gamma_2 \rangle_{HS} = \sum_{k=1}^{\infty} \langle \Gamma_1(e_k), \Gamma_2(e_k) \rangle,$$

respectively, where $\{e_1, e_2, \ldots\}$ is any orthonormal basis of E, and $\Gamma, \Gamma_1, \Gamma_2 \in HS(E)$. In particular, $\|\Gamma_0\|_{HS}^2 = \lambda_1^2 + \lambda_2^2 + \cdots$. Given an iid sample $X_1, \ldots, X_n$, we define the Hilbert–Schmidt operators of rank 1 by setting, for each $i \in \{1, \ldots, n\}$, $W_i : E \to E$, as $W_i(f) = \langle X_i, f \rangle X_i(\cdot)$. Let $\phi_i = X_i / \|X_i\|$, then $W_i(\phi_i) = \|X_i\|^2 \phi_i \equiv \eta_i \phi_i$.

The standard estimator of Γ_0 is just the average of these operators, i.e., $\hat{\Gamma}_n = (W_1 + \cdots + W_n)/n$, which is a consistent estimator of Γ_0 by the Law of Large Numbers in the space $HS(E)$. We replace this average by a trimmed version in the space $HS(E)$.

3.2.2. A new robust estimator for the covariance operator

Our proposal is to consider an impartial trimmed estimator as a resistant estimator. The notion of impartial trimming was introduced in [13], and the functional data setting was considered in [7], from which one can obtain the asymptotic theory for our setting. The construction of our estimator needs an explicit expression of the distances $\|W_i - W_j\|$, computed for all $i, j \in \{1, \ldots, \ell\}$ with $i < j$, which we will derive using the following lemma.

Lemma 1. *We have, for all $i, j \in \{1, \ldots, n\}$ with $i < j$,*

$$d_{ij}^2 \equiv \|W_i - W_j\|_{HS}^2 = \|X_i\|^4 + \|X_j\|^4 - 2\langle X_i, X_j\rangle^2. \tag{7}$$

Proof. Let us write

$$\begin{aligned}\langle W_i - W_j, W_i - W_j\rangle_{HS} &= \langle W_i, W_i\rangle_{HS} + \langle W_j, W_j\rangle_{HS} - 2\langle W_i, W_j\rangle_{HS} = \eta_i^2 + \eta_j^2 - 2\sum_{k=1}^{\infty}\langle W_i(e_k), W_j(e_k)\rangle \\ &= \eta_i^2 + \eta_j^2 - 2\sum_{k=1}^{\infty}\langle\langle X_i, e_k\rangle X_i, \langle X_j, e_k\rangle X_j\rangle = \eta_i^2 + \eta_j^2 - 2\langle X_i, X_j\rangle\sum_{k=1}^{\infty}\langle X_i, e_k\rangle\langle X_k, e_k\rangle.\end{aligned}$$

Now Eq. (7) follows from the identity

$$\sum_{k=1}^{\infty}\langle X_i, e_k\rangle\langle X_k, e_k\rangle = \langle X_i, X_j\rangle.$$

This concludes the argument. □

Given the sample, which we took to have mean zero for notational simplicity, and $\alpha \in (0, 1)$, we provide a simple algorithm to compute an approximate impartial trimmed mean estimator of the covariance operator which is strongly consistent.

Step 1: Compute $d_{ij} = \|W_i - W_j\|_{HS}$, for all $i, j \in \{1, \ldots, n\}$ with $i < j$, using Lemma 1.

Step 2: Let $r = \lfloor(1-\alpha)n\rfloor + 1$. For each $i \in \{1, \ldots, n\}$, consider the set of indices $I_i \subset \{1, \ldots, n\}$ corresponding to the r nearest neighbors of W_i among $W_1, \ldots, W_n$, and the order statistic of the vector $(d_{i1}, \ldots, d_{in})$, $d_i^{(1)} \leq \cdots \leq d_i^{(n)}$.

Step 3: Let $\gamma = \operatorname{argmin}\{d_1^{(r)}, \ldots, d_n^{(r)}\}$.

Step 4: The impartial trimmed mean estimator of Γ_0 is given by the average of the r nearest neighbors of W_γ among $W_1, \ldots, W_n$, i.e., the average of the rank-1 operators W_i such that $i \in I_\gamma$. The covariance function is then estimated by $\hat{\rho}(s,t) = \sum_{j\in I_\gamma} X_j(s)X_j(t)/r$. Observe that Steps 1 and 2 of the algorithm can be performed using parallel computing.

The final estimator given by the RFM may be obtained by taking the deepest point (or the average of the 40% deepest points) among the m estimators obtained from the algorithm above. The norm used for the depth function in this case is the functional analogue of (6).

3.3. Robust fusion for cluster analysis

In this section we describe a robust fusion method for clustering. Our approach is based on the use of impartial trimmed k-means (ITkM, see [8]) in two steps. In the first one we apply ITkM with a given trimming level α_1 to each of the m subsamples, and obtain m sets of k centres $\hat{\mathcal{M}}_1, \ldots, \hat{\mathcal{M}}_m$. In the second step we apply ITkM with a trimming level α_2 to the set $\hat{\mathcal{M}}_1 + \cdots + \hat{\mathcal{M}}_m$, as suggested in Section 5.1 of [8]. We start by describing briefly ITkM.

3.3.1. Impartial trimmed k-means

Given a sample $\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$, a trimming level $\alpha \in (0, 1)$, and the number of clusters k, ITkM looks for a set $\{\hat{m}_1, \ldots, \hat{m}_k\} \subset \mathbb{R}^d$ and a partition of the space $C_0, \ldots, C_k$ that minimizes the loss function

$$\frac{1}{n - \lfloor n\alpha \rfloor} \sum_{j=1}^{k} \sum_{X_i \in C_j} \|X_i - \hat{m}_j\|^2.$$

Here, C_0 is the set of trimmed data (with cardinality $\lfloor n\alpha \rfloor$). Let $X \in \mathbb{R}^d$ be a random vector with distribution P_X, the number of clusters k, and a trimming proportion $\alpha \in (0, 1)$.

Step 1: For every k-set $\mathcal{M} = \{m_1, \ldots, m_k\}$, with $m_j \in \mathbb{R}^d$ for all $j \in \{1, \ldots, k\}$, and $x \in \mathbb{R}^d$, define

$$d(x, \mathcal{M}) \equiv \min\{\|x - m_1\|, \ldots, \|x - m_k\|\}.$$

Step 2: The set of trimming functions for P_X at level α is defined by

$$\tau_\alpha(P_X) = \Big\{\tau : \mathbb{R}^d \to [0, 1], \text{ measurable, fulfilling } \int \tau(x) dP_X(x) \geq 1 - \alpha\Big\}.$$

The functions in $\tau_\alpha(P_X)$ are a natural generalization of the indicator functions $\mathbf{1}_A$ with $P_X(A) = 1 - \alpha$.

Step 3: For each pair $(\tau, \mathcal{M})$ such that $\tau \in \tau_\alpha(P_X)$ and $\mathcal{M} \subset \mathbb{R}^d$ with $\text{card}(\mathcal{M}) = k$, consider the function

$$\mathcal{V}(\tau, \mathcal{M}, P_X) = \frac{1}{\int \tau(x) dP_X(x)} \int \tau(x) d^2(x, \mathcal{M}) dP_X(x).$$

Step 4: Finally, define

$$\mathcal{V}(P_X) = \inf_{\tau \in \tau_\alpha(P_X)} \inf_{\substack{\mathcal{M} \subset \mathbb{R}^d \\ \text{card}(\mathcal{M}) = k}} \mathcal{V}(\tau, \mathcal{M}, P_X). \tag{8}$$

Corollary 3.2 in [8] states that there exists a pair $(\tau^*, \mathcal{M}^*)$, not necessarily unique, attaining the value $\mathcal{V}(P_X)$. Moreover, if P_X is absolutely continuous with respect to Lebesgue measure, $\tau^* = \mathbf{1}_{B(\mathcal{M}^*, r^*)}$ with $r^* = r(\alpha, \mathcal{M}^*) = \inf\{r \geq 0 : P_X\{B(\mathcal{M}^*, r)\} \geq 1 - \alpha\}$ and $B(\mathcal{M}^*, r) = \{x \in \mathbb{R}^d : d(x, \mathcal{M}^*) \leq r\}$.

Denote by P_n the empirical distribution based on the sample. Theorem 3.6 in [8] states that if P_X is absolutely continuous with respect to Lebesgue measure and there exists a unique pair $(\tau^*, \mathcal{M}^*)$ solving (8), then $\mathcal{V}(P_n) \to \mathcal{V}(P_X)$ a.s. Moreover, if $\hat{\mathcal{M}}$ is any sequence of empirical trimmed k-means, then $d_H(\hat{\mathcal{M}}, \mathcal{M}^*) \to 0$ a.s., where d_H denotes the Hausdorff distance. It is clear that in this case $\hat{\tau}_n = \mathbf{1}_{B(\hat{\mathcal{M}}, \hat{r})} \to \tau^* \; P_X$ a.s., where $\hat{r} = \inf\{r \geq 0 : P_n\{B(\hat{\mathcal{M}}, r)\} \geq 1 - \alpha\}$.

Now $\mathcal{M}^*$ and $\hat{\mathcal{M}}$ induce partitions of $B(\mathcal{M}^*, r^*)$ and $B(\hat{\mathcal{M}}, \hat{r})$ respectively, into k-clusters, by defining, for $i \in \{1, \ldots, k\}$,

$$\text{Cluster } C_i = \{x \in B(\mathcal{M}^*, r^*) : \|x - m_i^*\| \leq \min_{j \neq i} \|x - m_j^*\|\},$$

$$\text{Cluster } \hat{C}_i = \{x \in B(\hat{\mathcal{M}}, \hat{r}) : \|x - \hat{m}_i\| \leq \min_{j \neq i} \|x - \hat{m}_j\|\}. \tag{9}$$

Points at a boundary between clusters can be assigned arbitrarily. A functional version of ITkM can be found in [7]. With this in hand, the fusion step of the RFM is done by applying ITkM to the set of the $k \times m$ centers. The whole algorithm is summarized in Table 3.

4. Simulation results

We now describe the simulations done with the RFM for the three applications described in the previous sections. As the design of each simulation is specific to its application, we describe them separately. All the simulations were carried out using an 8-core PC, Intel core i7-3770 CPU, 8GB of RAM, 64 bit processor, with the R software package v. 3.3.0 running under Ubuntu.

Table 3: RFM algorithm for clustering

1) Split the sample into m subsamples; recall that $n = m\ell$.
2) To each subsample, apply the empirical version of α-ITkM with $\alpha = \alpha_1$ and obtain $\hat{\mathcal{M}}_1, \ldots, \hat{\mathcal{M}}_m$, each one with k points in $\mathbb{R}^d$.
3) Apply the empirical version of α-ITkM with $\alpha = \alpha_2$ to the set $\hat{\mathcal{M}}_1 \cup \cdots \cup \hat{\mathcal{M}}_m$.
4) Obtain the output of the algorithm $(\hat{\mathcal{M}}_{RFM}, \hat{r}_{RMF})$.
5) Build the clusters by applying (9).

4.1. Location and scatter matrix for finite-dimensional spaces

We use the same simulations to analyze both the location of the parameters and their scatter matrix. For the robust estimator we have applied the function CovMest in the R package `rrcov` with the parameters given by default.

We draw samples from a centered 5-dimensional Gaussian distribution with a covariance matrix whose off-diagonal elements are all equal to 0.2. For the outliers we use a 5-dimensional Cauchy distribution with independent coordinates centered at 50. We test two contamination levels, namely $p = 0.2$ and $p = 0.4$. We vary the sample size n within the set $\{0.1E6, 5E6, 10E6\}$ and the number of subsamples $m \in \{100, 500, 1000, 10{,}000\}$. We replicate each simulation case $K = 5$ times and report the average. The estimators obtained by the RFM are the values which maximize the depth functions given in Eqs. (5)–(6) for the location and the scatter matrix, respectively. In each case, the maximization is done over the set of the m estimates obtained from the subsamples.

The mean squared errors (averaged over five replicates) for the location problem are given in Table 4. The estimators considered are the average of the whole sample (MLE), the average of the robust location estimators (avROB), the average of the 40% deepest robust estimators (RFM1), and the deepest robust estimator (RFM). We can see from Table 4 that the estimator obtained by the RFM behaves very well. Depending on the structure of the outliers, the mean of the robust estimates may behave well or not. Even if only one of the subsamples contains a high proportion of outliers causing the robust estimator to break down, the average of the robust estimators will break down. In contrast, the deepest M-estimator always behaves well. The performances of both estimators decrease in general with m.

Table 4: Location estimators for $p = 0.2$ and $p = 0.4$.

		MLE	avROB	RFM1	RFM	MLE	avROB	RFM1	RFM
n	m	$p = 0.2$				$p = 0.4$			
0.1	100	31.3	0.0098	0.0124	0.0297	44.2	0.0042	0.0076	0.0288
0.1	500	31.3	0.0097	0.0112	0.0426	44.2	0.1070	0.0081	0.0427
0.1	1000	31.3	0.0097	0.0109	0.0477	44.2	1.3400	0.0231	0.0416
1.0	100	21.4	0.0021	0.0029	0.0074	44.1	0.0038	0.0045	0.0087
1.0	500	21.4	0.0021	0.0037	0.0110	44.1	0.0038	0.0038	0.0164
1.0	1000	21.4	0.0021	0.0030	0.0159	44.1	0.0038	0.0053	0.0186
1.0	10,000	21.4	0.0022	0.0035	0.0261	44.1	1.3900	0.0186	0.0375
5.0	100	22.0	0.0009	0.0014	0.0032	45.9	0.0007	0.0014	0.0044
5.0	500	22.0	0.0009	0.0010	0.0056	45.9	0.0007	0.0014	0.0073
5.0	1000	22.0	0.0009	0.0014	0.0071	45.9	0.0007	0.0014	0.0097
5.0	10,000	22.0	0.0009	0.0015	0.0147	45.9	0.0013	0.0011	0.0159
10.0	100	23.5	0.0009	0.0013	0.0026	47.0	0.0005	0.0010	0.0033
10.0	500	23.5	0.0009	0.0012	0.0038	47.0	0.0005	0.0009	0.0052
10.0	1000	23.5	0.0009	0.0012	0.0047	47.0	0.0005	0.0008	0.0056
10.0	10,000	23.5	0.0009	0.0012	0.0090	47.0	0.0005	0.0009	0.0102

Table 5: Covariance estimators. Using MLE and robust estimates over the entire sample, and aggregating by average, trimmed average or fusion of m subsamples estimators, $p = 0.2$.

n	m	T0	T1	MLE	ROB	avROB	RFM1	RFM
0.1	100	0.460	4.205	23,688,000	0.2594	0.2597	0.2598	0.3722
0.1	500	0.460	14.527	23,688,000	0.2594	0.2675	0.2498	0.4810
0.1	1000	0.460	23.992	23,688,000	0.2594	0.2748	0.2418	0.6130
1.0	100	3.444	6.524	1,617,200	0.2342	0.2345	0.2368	0.2656
1.0	500	3.444	24.028	1,617,200	0.2342	0.2353	0.2371	0.3189
1.0	1000	3.444	45.307	1,617,200	0.2342	0.2360	0.2381	0.3295
1.0	10,000	3.444	945.350	1,617,200	0.2342	0.2464	0.2075	0.4982
5.0	100	20.528	15.984	1,981,900	0.2317	0.2316	0.2340	0.2495
5.0	500	20.528	33.289	1,981,900	0.2317	0.2316	0.2331	0.2687
5.0	1000	20.528	68.342	1,981,900	0.2317	0.2318	0.2336	0.2842
5.0	10,000	20.528	1312.800	1,981,900	0.2317	0.2342	0.2267	0.3810
10.0	100	42.174	29.168	28,135,000	0.2307	0.2306	0.2322	0.2445
10.0	500	42.174	49.992	28,135,000	0.2307	0.2307	0.2322	0.2567
10.0	1000	42.174	73.701	28,135,000	0.2307	0.2308	0.2315	0.2660
10.0	10,000	42.174	1291.000	28,135,000	0.2307	0.2323	0.2290	0.3439

The estimation errors for the covariance are given in Tables 5 and 6, which correspond to the cases $p = 0.2$ and $p = 0.4$, respectively. We compare the MLE estimator (MLE), a robust estimator based on the whole sample (ROB), the average of the robust scatter matrix estimators (avROB), the average of the 40% deepest robust estimators (RFM1), and the deepest robust estimator (RFM). We also report the average time in seconds necessary for both the global estimator (T0, over the whole sample), and T1, the estimator obtained by fusion (including computing the estimators over subsamples and aggregating them by fusion). Since the second step of the algorithm can be parallelized, see point b in Table 1), in practice the computational time T1 can be divided almost by m. The results of RFM are very good for the covariance matrix as well.

Table 6: Covariance estimators. Using MLE and robust estimates over the entire sample, and aggregating by average, trimmed average or fusion the m subsamples estimators, $p = 0.4$.

n	m	T0	T1	MLE	ROB	avROB	RFM1	RFM
0.1	100	0.581	3.416	448,210	0.8247	0.8348	0.8378	1.0111
0.1	500	0.581	13.614	448,210	0.8247	16.3120	0.8057	1.2548
0.1	1000	0.581	22.827	448,210	0.8247	205.3000	0.7772	1.5159
1.0	100	2.631	5.622	6,030,100	0.8081	0.8094	0.8114	0.8790
1.0	500	2.631	21.131	6,030,100	0.8081	0.8143	0.8103	0.9462
1.0	1000	2.631	39.752	6,030,100	0.8081	0.8201	0.8059	0.9690
1.0	10,000	2.631	833.530	6,030,100	0.8081	203.9100	0.7706	1.2760
5.0	100	16.651	14.126	29,809,000	0.8010	0.8012	0.8035	0.8299
5.0	500	16.651	30.762	29,809,000	0.8010	0.8021	0.8025	0.8571
5.0	1000	16.651	60.389	29,809,000	0.8010	0.8032	0.8020	0.8740
5.0	10,000	16.651	1239.300	29,809,000	0.8010	0.9024	0.7877	1.0311
10.0	100	33.922	24.757	93,071,000	0.7988	0.7989	0.8013	0.8185
10.0	500	33.922	43.999	93,071,000	0.7988	0.7993	0.8007	0.8420
10.0	1000	33.922	68.787	93,071,000	0.7988	0.8001	0.8001	0.8555
10.0	10,000	33.922	1486.100	93,071,000	0.7988	0.8117	0.7939	0.9403

4.2. Covariance operator

To generate the data, we used a simplified version of the simulation model used in [18], viz.

$$X(t) = \mu(t) + \sqrt{2}\sum_{k=1}^{10} \lambda_k a_k \sin(2\pi k t) + \sqrt{2}\sum_{k=1}^{10} \nu_k b_k \cos(2\pi k t),$$

where $\nu_k = (1/3)^k$, $\lambda_k = k^{-3}$, and a_k and b_k are random standard Gaussian independent observations, see Figure 1. The central observations were generated using $\mu(t) = 0$ whereas for the outliers we took $\mu(t) = 2 - 8\sin(\pi t)$. For t we used an equally spaced grid of $\mathcal{T} = 20$ points in $[0, 1]$. The covariance operator of this process, given by

$$\text{cov}(s, t) = \sum_{k=1}^{10} A_k(s)A_k(t) + B_k(s)B_k(t),$$

where $A_k(t) = \sqrt{2}\lambda_k \sin(2\pi k t)$ and $A_k(t) = \sqrt{2}\nu_k \cos(2\pi k t)$, was computed for the comparisons.

We varied the sample size n within the set $\{0.1E6, 1E6, 5E6, 10E6\}$ and the number of subsamples $m \in \{100, 500, 1000, 10{,}000\}$. The proportion of outliers was successively fixed to $p = 0.15$ and $p = 0.20$. We replicated each simulation case $K = 5$ times and report the average performance over the replicates. We report also the average time in seconds necessary for both a global estimate T0, over the whole sample, and T1, the estimate obtained by fusion (including computing the estimates over subsamples and aggregating them by fusion). We compare the classical estimator (MLE), the global robust estimate (ROB), the average of the robust estimates from the subsamples (avROB) and the robust fusion estimate (RFM). The results are shown in Tables 7 and 8 for two proportions of outliers, $p = 0.15$ and $p = 0.2$ respectively.

If the proportion of outliers is moderate, $p = 0.15$, the average of the robust estimators still behaves well, better than RFM, but if we increase the proportion of outliers to $p = 0.2$, RFM clearly outperforms all the other estimators.

4.3. Clustering

We performed a simulation study for large sample sizes, using a model with three clusters with outliers, introduced in [8]. The data were generated using bivariate Gaussian distributions with the following parameters for the clusters and the outliers, respectively:

$$\mu_1 = (0, 0), \quad \mu_2 = (0, 10), \quad \mu_3 = (6, 0), \quad \mu_4 = (2, 10/3),$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = 1.5 \times Id, \quad \Sigma_4 = 20 \times Id$$

where Id is the two-dimensional identity matrix. The outliers were generated with μ_4, Σ_4 and n_4. The sizes of the clusters were fixed at the following values: $n_1 = 15$, $n_2 = 30$, $n_3 = 30$, $n_4 = 40$. As in [8], the outliers lying in the 75% level confidence ellipsoids of the clusters were replaced by others not belonging to that area. The outliers

Table 7: Covariance operator estimator. Using the classical and robust estimators over the entire sample, and aggregating by average or fusion of m subsamples estimators. $p = 0.15$, $\mathcal{T} = 20$.

n	m	T0	T1	MLE	ROB	avROB	RFM
0.05	20	553	18.20	24.3	5.16	5.21	5.52
0.05	50	543	7.81	24.3	5.20	5.24	5.60
0.05	100	528	4.79	24.3	5.20	5.17	5.58
0.05	1000	459	19.40	24.3	5.13	5.54	6.58
0.10	20	2300	69.00	24.2	5.14	5.22	5.43
0.10	50	2300	28.10	24.2	5.04	5.09	5.13
0.10	100	2290	15.20	24.2	5.06	5.15	5.43
0.10	1000	1850	21.60	24.3	5.21	5.35	6.13

Table 8: Covariance operator estimator. Using classical and robust estimators over the entire sample, and aggregating by average or fusion of m subsamples estimators. $p = 0.2, \mathcal{T} = 20$.

n	m	T0	T1	MLE	cvRob	avROB	RFM
0.05	20	572	17.90	30.5	0.879	3.96	1.45
0.05	50	649	7.88	30.5	0.876	7.34	2.10
0.05	100	633	4.61	30.5	0.839	8.86	2.43
0.05	1000	478	19.50	30.5	0.864	13.10	7.08
0.10	20	1970	69.10	30.4	0.914	3.83	1.36
0.10	50	2030	28.10	30.4	0.921	4.32	1.55
0.10	100	2020	15.10	30.4	0.840	8.44	2.35
0.10	1000	1840	21.60	30.4	0.961	12.10	5.20

represent almost 35% of the whole sample. We used this base simulation and varied the whole sample size, multiplying each n_i by a factor "fac" taking the values in $\{10, 100, 1000, 10{,}000\}$. So for the smallest sample, we have $n = 1150$, and the largest, $n = 1{,}150{,}000$.

For each value of n we varied the number of subsamples $m \in \{10, 50, 100, 1000, 10{,}000\}$ with the restriction $m <$ fac. Finally, when applying the trimmed k-means to the samples, we tested three values for the trimming level, $\alpha_1 = 0.2, 0.35, 0.45$, whereas for the fusion we fixed $\alpha_2 = 0.1$.

The left-hand panel of Figure 2 shows an example of the simulated data set for $n = 11{,}500$, the middle panel shows the results obtained by ITkM applied to the whole sample, and the right-hand panel shows the output of the algorithm. The partitions obtained by each approach are compared to the true clusters using the matching error defined by

$$ME = \min_{s \in \mathcal{S}(k+1)} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq s(\hat{y}_i)\}} \tag{10}$$

where $\mathcal{S}(k+1)$ is the set of permutations of $\{0, \ldots, k+1\}$, y_i is the true cluster of observation i and $\hat{y}_i$ is the cluster assigned by the algorithm. The results of the simulation are given in Table 9, where we compare the RFM method,

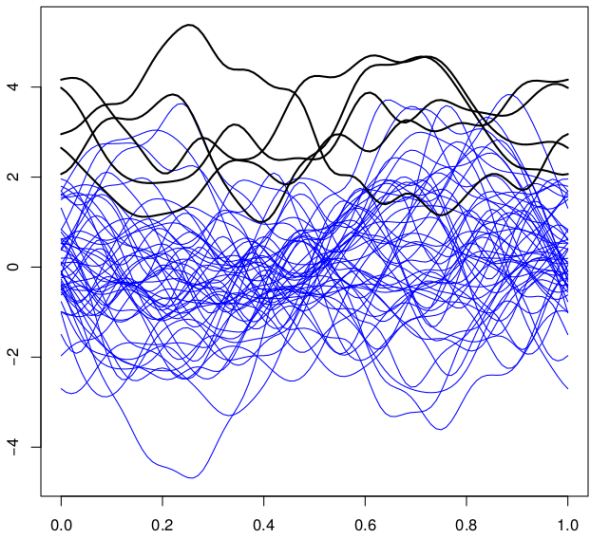

Figure 1: Simulated functions and outliers

with the ITkM calculated with the whole sample. Columns ME1 and ME2 give the matching errors for ITkM applied to the whole sample and for RFM respectively. We also report the average time in seconds necessary for both the global estimator (T0, over the whole sample), and T1, the estimator obtained by fusion (including computing the estimators over subsamples and aggregating them by fusion). Finally T2 is the time using parallel computing.

As expected, the RFM matching errors are often higher than those of ITkM applied to the whole sample. But the loss of performance is very small in general and increases with m. For the smallest values of m with large samples ($n > 10{,}000$), RFM has almost the same performance for all values of α. In contrast, increasing the value of m reduces considerably the computation time of RFM.

4.3.1. A real data example

As an example we have chosen the MNIST data set of handwritten digits (see `https://www.kaggle.com/c/digit-recognizer/data`) to compare the performance of the RFM clustering algorithm with the same clustering procedure without splitting the sample (impartial trimmed k-means). The digits have been size-normalized and centered in a fixed-size image of 28×28 pixels.

The data set consists of a training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of $n = 42{,}000$ data, and a test sample of 10,000 data. As it is explained in the aforementioned link: "this classic data set of handwritten images has served as the basis for benchmarking classification algorithms". However, as we are interested in clustering we will use only the sample $X_1, \ldots, X_n$, searching for $k = 10$ groups. This is a very difficult task: if the labels are chosen at random the probability to get at least half of the 42,000 data well identified is extremely close to zero. We cluster the 42,000 data using both methods.

The design is the same as for the previous simulations. On the one hand we cluster the whole sample using the impartial trimmed k-mean algorithm for $k = 10$. On the other hand we use the RFM clustering method given in Table 3 for $m \in \{10, 100, 500, 1000\}$, with $\alpha_1 = 0.05$ and $\alpha_1 = 0.1$. The labels $Y_1, \ldots, Y_n$ are only used to compute the misclassification error rates ME1 and ME2 defined in (10).

The results are given for $\alpha_1 = 0.05$ and $\alpha_2 = 0.1$ in Table 10 (left), and for $\alpha_1 = 0.1 = \alpha_2$ in Table 10 (right). They show that (a) this clustering problem is very difficult; (b) the relative efficiencies of the RFM clustering procedures for $\alpha_1 = 0.05$ are $5\%, 2\%, 9\%$ and 6% while the computational times fall down drastically to $17\%, 5\%, 3\%$ and 3%, for $m = 10, 100, 500$ and 1000, respectively. For $\alpha_1 = 0.1$ the efficiencies are $8\%, 7\%, 9\%$ and 8%, the computational times fall down to $16\%, 7\%, 5\%$ and 4% for $m = 10, 100, 500$ and 1000, respectively.

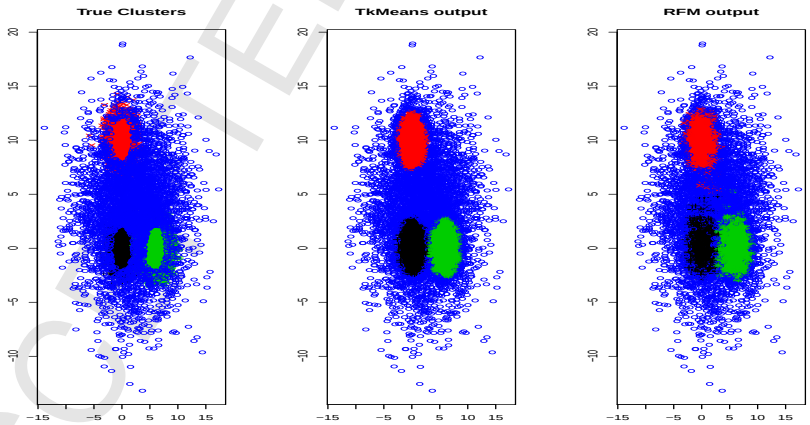

Figure 2: Left panel: the true clusters. Middle panel: Results obtained by ITkM over the whole sample. Right panel: The output obtained by RFM using $m = 100$ subsamples. The outliers are the blue points and $n = 115{,}000$.

Table 9: RFM for clustering using different values of the trimming parameter α_1.

n	m	T0	T1	T2	ME1	ME2
			$\alpha_1 = 0.2$			
1150	10	2.89	1.34	0.55	0.1539	0.1678
11,500	10	21.20	21.69	6.83	0.1594	0.1603
11,500	100	21.20	14.65	4.24	0.1594	0.1693
11,5000	10	274.90	263.80	75.11	0.1585	0.1585
115,000	100	274.90	218.10	56.44	0.1585	0.1591
115,000	1000	274.90	141.50	37.51	0.1585	0.1693
1,150,000	10	3452.00	3149.00	873.40	0.1582	0.1582
1,150,000	100	3452.00	2609.00	680.70	0.1582	0.1583
1,150,000	1000	3452.00	2158.00	546.90	0.1582	0.1590
1,150,000	10,000	3452.00	1434.00	374.70	0.1582	0.1689
			$\alpha_1 = 0.35$			
1150	10	3.45	1.43	0.54	0.1287	0.1310
11,500	10	37.87	33.38	9.89	0.1037	0.1071
11,500	100	37.87	15.30	4.29	0.1037	0.1343
115,000	10	427.70	391.10	109.60	0.1049	0.1050
115,000	100	427.70	307.20	85.70	0.1049	0.1071
115,000	1000	427.70	137.70	38.36	0.1049	0.1331
1,150,000	10	4925.00	4284.00	1166.00	0.1052	0.1053
1,150,000	100	4925.00	3660.00	928.20	0.1052	0.1055
1,150,000	1000	4925.00	3052.00	792.90	0.1052	0.1074
1,150,000	10,000	4925.00	1397.00	372.20	0.1052	0.1336
			$\alpha_1 = 0.45$			
1150	10	2.72	1.27	0.52	0.1330	0.1567
11,500	10	55.58	34.12	9.80	0.1370	0.1403
11,500	100	55.58	13.11	3.65	0.1370	0.1723
115,000	10	698.90	586.60	170.40	0.1325	0.1330
115,000	100	698.90	323.90	86.35	0.1325	0.1355
115,000	1000	698.90	122.50	33.53	0.1325	0.1729
1,150,000	10	7190.00	7087.00	2115.00	0.1327	0.1328
1,150,000	100	7190.00	5654.00	1508.00	0.1327	0.1330
1,150,000	1000	7190.00	3287.00	829.60	0.1327	0.1360
1,150,000	10,000	7190.00	1258.00	328.10	0.1327	0.1726

Table 10: Robust clustering, $\alpha_2 = 0.1$, $\alpha_1 = 0.05$ (left) and $\alpha_1 = 0.1$ (right).

m	T0	T1	ME1	ME2
10	8560	1540	0.477	0.503
100	8560	503	0.477	0.486
500	8560	244	0.477	0.520
1000	8560	253	0.477	0.508

m	T0	T1	ME1	ME2
10	9570	1500	0.492	0.530
100	9570	705	0.492	0.525
500	9570	445	0.492	0.536
1000	9570	417	0.492	0.532

5. Concluding remarks

We have addressed some fundamental statistical problems in the context of Big Data, namely large samples, in the presence of outliers, location and covariance estimation, covariance operator estimation, and clustering. We have proposed a general robust approach, called the robust fusion method (RFM), and shown how it may be applied to these

problems. The simulations gave very good results mainly for the last two problems. Different statistical challenges are associated with these problems. Our approach may be adapted to any other task as soon as a robust efficient estimate is available for the corresponding problem.

Acknowledgments

We are grateful to the Editor-in-Chief, Christian Genest, an Associate Editor, and two referees for their constructive comments. For the last author this work has been partially supported by the ECOS project No. U14E02.

References

References

[1] S.E. Ahmed, Ed., Big and Complex Data Analysis: Methodologies and Applications, Springer, Berlin, 2017.
[2] A. Aho, J.E. Hopcroft, J.D. Ullman, The Design and Analysis of Computer Algorithms, Addison-Wesley, Boston, MA, 1974.
[3] G. Aneiros, E.G. Bongiorno, R. Cao, P. Vieu, Eds., Functional Statistics and Related Fields, Springer, Berlin, 2017.
[4] P. Billingsley, F. Topsøe, Uniformity in weak convergence, Z. Wahrs. und Verw. Gebiete 7 (1967) 1–16.
[5] A. Chakraborty, P. Chaudhuri, The spatial distribution in infinite-dimensional spaces and related quantiles and depths, Ann. Statist. 42 (2014) 1203–1231.
[6] P. Chaudhuri, On a geometric notion of quantiles for multivariate data, J. Amer. Statist. Assoc. 91 (1996) 862–872.
[7] J.A. Cuestas-Albertos, R. Fraiman, Impartial means for functional data, In: R. Liu, R. Serfling, D. Souvaine, Eds. Data Depth: Robust Multivariate Statistical Analysis, Computational Geometry and Applications, Vol. 72 in the DIMACS Series of the American Mathematical Society, 2006, pp. 121–145.
[8] J.A. Cuesta-Albertos, A. Gordaliza, C. Matrán, Trimmed k-means: An attempt to robustify quantizers, Ann. Statist. 25 (1997) 553–576.
[9] A. Cuevas, A partial overview of the theory of statistics with functional data, J. Stat. Plann. Inf. 147 (2014) 1–23.
[10] D.L. Donoho, Breakdown properties of multivariate location estimators, Ph.D. qualifying papers, Dept. of Statistics, Harvard University, Cambridge, MA, 1982.
[11] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer, Berlin, 2006.
[12] A. Goia, P. Vieu, Special issue on statistical models and methods for high or infinite dimensional spaces, J. Multivariate Anal. 146 (2016)1–352.
[13] A. Gordaliza, Best approximations to random variables based on trimming procedures, J. Approx. Theory 64 (1991) 162–180.
[14] F.R. Hampel, A general qualitative definition of robustness, Ann. Math. Statist. 42 (1971) 1887–1896.
[15] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer, Berlin, 2012.
[16] P.J. Huber, The behavior of maximum likelihood estimates under nonstandard conditions, Proc. Fifth Berkeley Symp. on Math. Statist. and Prob. Univ. of Calif. Press, Berkeley, CA, 1 (1967) 221–233
[17] P.J. Huber, E.M. Ronchetti, Robust Statistics, Wiley, Hoboken, NJ, 2009.
[18] D. Kraus, V.M. Panareto, Dispersion operators and resistant second-order functional data analysis, Biometrika 101 (2012) 141–154.
[19] R. Maronna, R. Martin, V. Yohai, Robust Statistics: Theory and Methods, Wiley, Hoboken, NJ, 2006.
[20] L. Tang, L. Zhou, P.X.-K. Song, Method of divide-and-combine in regularised generalised linear models for Big Data, 2016. `https://arxiv.org/abs/1611.06208`.
[21] Y. Vardi, C. Zhang, The multivariate L_1-median and associated data depth, Proc. Nat. Acad. Sci. USA 97 (2000) 1423–1426.
[22] C. Wang, M.-H. Chen, E. Schifano, J. Wu, J. Yan, Statistical methods and computing for Big Data, Statistics and Its Interface 9 (2016) 399–414.
[23] B. Yu, Let us own data science, IMS Bulletin 43(7) (2014) 1 + 13–16.