

Robust estimators based on generalization of trimmed mean

Lukáš Adam & Přemysl Bejda

To cite this article: Lukáš Adam & Přemysl Bejda (2017): Robust estimators based on generalization of trimmed mean, Communications in Statistics - Simulation and Computation, DOI: [10.1080/03610918.2017.1337136](https://doi.org/10.1080/03610918.2017.1337136)

To link to this article: <http://dx.doi.org/10.1080/03610918.2017.1337136>



Accepted author version posted online: 06 Jun 2017.
Published online: 06 Jun 2017.



Submit your article to this journal [↗](#)



Article views: 12




View related articles [↗](#)



View Crossmark data [↗](#)



Robust estimators based on generalization of trimmed mean

Lukáš Adam ^a and Přemysl Bejda^a

^aInstitute of Information Theory and Automation, Czech Academy of Sciences, Prague, Czech Republic;

^bFaculty of Mathematics and Physics, Charles University, Prague, Czech Republic

ABSTRACT

In this article, we propose new estimators of location. These estimators select a robust set around the geometric median, enlarge it, and compute the (iterative) weighted mean from it. By doing so, we obtain a robust estimator in the sense of the breakdown point, which uses more observations than standard estimators. We apply our approach on the concepts of boxplot and bagplot. We work in a general normed vector space and allow multi-valued estimators.

ARTICLE HISTORY

Received 15 February 2016

Accepted 25 May 2017

KEYWORDS

Breakdown point; Estimators; Geometric median; Location; Trimmed mean

MATHEMATICS SUBJECT

CLASSIFICATION

62G35; 62G05

1. Introduction

Robust statistical methods try to weaken quite restrictive assumptions of classical methods. For instance, in many statistical models it is assumed that residuals are independent and identically distributed. Moreover, the assumption of normality of residuals is often added and the Euclidean norm is usually employed. But for real data these assumptions are often violated and classical approaches fail. This leads to a necessity to consider robust statistical methods. As classical examples of robust methods we mention replacing the l_2 norm by the l_1 norm or replacing the mean by the median. Such development was also enabled by the progress in computer technologies as the robust methods are usually more complicated and computationally demanding than the classical methods.

In this article, we are interested in robust estimators for the parameter of location. One of the first attempts to deal with such estimators are M-estimators, see Huber (1964) or Maronna (1976). They are computationally simple but suffer from low breakdown point, see Hampel (1971) or Donoho and Huber (1983). This, loosely speaking, expresses the data fraction, which can be “arbitrarily modified” without affecting the finiteness of the estimator. Its value is at most $\frac{1}{d+1}$, see Maronna et al. (2006), where d is the dimension of observations. Later, multiple estimators got proposed, among others we mention:

- minimum volume ellipsoid estimators (MVE, Rousseeuw, 1985) whose name stems from the fact that among all “proper” ellipsoids containing at least half of observations, the one given by MVE has minimal volume. However, their efficiency is rather poor.
- S-estimators (Davies, 1987) have been suggested to overcome the low efficiency of MVE. They combine approaches of MVE and M-estimators.
- τ -estimators (Lopuhaä, 1991) also employ the idea of M-estimators but they do not require preliminary scale estimator.

- Stahel–Donoho estimator (Donoho, 1982; Stahel, 1981) is based on the idea that any outlier in multivariate case should be an outlier in some univariate projection.

The advantage of these estimators is their high breakdown point, which is the highest which a shift equivariant estimator can attain. However, their computation usually requires heavy effort. Therefore, Maronna and Zamarb (2002) suggested a way of reducing the computational complexity while sustaining the high breakdown point. As a price to pay, one is no longer able to estimate the covariance structure.

In this article, we follow the goal of finding easily computable robust estimators with high breakdown point. From a sample, we first select a set A of observations, which are robust in the sense of breakdown point. Then we utilize these observations to construct further estimators, for example, by enlarging A and computing the (iterative) weighted mean of observations from the enlarged set. Set A is based on the geometric median (also called the spatial median), which is a direct generalization of the real median proposed by Haldane (1948). Since the geometric median has breakdown point of $\frac{1}{2}$, our estimators will be able to keep this property as well.

Our estimators enjoy the following nice properties:

- Instead of considering the \mathbb{R}^d space, we work with a general normed vector space X . This opens a natural way to tackle time series by our approach.
- Our estimators have high breakdown point and are simple to compute.
- We partially consider the covariance structure.
- We are able to work with set-valued estimators instead of single-valued estimators.

Note that estimators usually only satisfy several of the above properties, for example, either they have high breakdown point or they do not take into account the covariance structure at all.

This article is organized as follows: in the first part of [Section 2](#) we define basic concepts of breakdown point and geometric median. Even though geometric median may be a set and not a point in general, most authors do not handle this fact. Because of this, we have decided to work with estimators, which are multifunctions (also known as set-valued maps). The second part of [Section 2](#) contains new results. We propose new estimators, discuss their breakdown point, and provide a comparison between our algorithms and M-estimators.

Our notation is as follows: by $(X, \|\cdot\|)$ we understand a normed vector space. For a set $A \subset X$, we define $\|A\| := \sup_{x \in A} \|x\|$. Often we will use the bold notation for $\mathbf{x} = (x_1, \dots, x_n) \in X^n$. By the lower index we understand a component of a vector while by the upper index, we mean an iteration number. A multifunction $R : X \rightrightarrows Y$ is a generalization of a function, where $R(x)$ does not have to be one point but may be a (nonempty) subset of Y . We say that R is bounded on bounded sets if $\cup_{x \in A} R(x)$ is a bounded set for all bounded sets $A \subset X$. Since we consider multi-valued estimators, some of the definitions are slightly generalized.

2. New estimators based on generalization of trimmed mean

In this section, we first recall the geometric median and on its basis derive other estimators. The basic idea is to find first the geometric median, then restrict ourselves to a set of neighboring observations and construct an estimator based only on this restricted set. If this set is chosen in a proper way, the estimator will have a breakdown point of $\frac{1}{2}$. The breakdown point is nowadays one of the standard measures of robustness and expresses the minimal proportion of the data, which can be corrupted (made arbitrarily distant) before the estimator becomes unbounded.

Definition 2.1. Consider a normed vector space X and an estimator $T_n : X^n \rightrightarrows X$ of some functional T . For $\mathbf{x} = (x_1, \dots, x_n) \in X^n$ and $m = 1, \dots, n$ define

$$A_{m,n}(\mathbf{x}) := \{\tilde{\mathbf{x}} \in X^n \mid \tilde{\mathbf{x}} \text{ and } \mathbf{x} \text{ have at most } m \text{ different coordinates}\},$$

$$m_n^*(T_n, \mathbf{x}) := \max_{m \in \{1, \dots, n\}} \left\{ m \mid \sup_{\tilde{\mathbf{x}} \in A_{m,n}(\mathbf{x})} \inf_{z \in T_n(\tilde{\mathbf{x}}), z \in T_n(\mathbf{x})} \|\tilde{z} - z\| < \infty \right\}.$$

Then we say that T_n has the breakdown point

$$\varepsilon_n^*(T_n, \mathbf{x}) := \frac{1}{n} m_n^*(T_n, \mathbf{x}).$$

Finally, for family of estimators $\{T_n : X^n \rightrightarrows X\}$ we define the asymptotic breakdown point as

$$\varepsilon^* := \liminf_{n \rightarrow \infty} \inf_{\mathbf{x} \in X} \varepsilon_n^*(T_n, \mathbf{x}).$$

We continue with the definition of geometric median.

Definition 2.2. We define the geometric median as a multifunction $\hat{T}_n : X^n \rightrightarrows X$ satisfying

$$\hat{T}_n(x_1, \dots, x_n) = \operatorname{argmin}_{a \in X} \sum_{j=1}^n \|a - x_j\|. \quad (1)$$

We present now two examples. The first one shows that the choice of the norm can change the geometric median in a significant way and that the geometric median may indeed be multi-valued. The second one depicts a simple situation where we are able to compute the geometric median. Moreover, it will be used later in some proofs.

Example 2.1. Consider $X = \mathbb{R}^2$ and points $x_1 = (1, 0)$, $x_2 = (-1, 0)$, and $x_3 = x_4 = (0, 1)$. Then it is not difficult to verify that for the following norms, we have

$$\begin{aligned} (\mathbb{R}^2, \|\cdot\|_1) &\Rightarrow \hat{T}_4(x_1, \dots, x_4) = \operatorname{conv}\{(0, 0), (0, 1)\}, \\ (\mathbb{R}^2, \|\cdot\|_2) &\Rightarrow \hat{T}_4(x_1, \dots, x_4) = \{(0, 1)\}, \\ (\mathbb{R}^2, \|\cdot\|_\infty) &\Rightarrow \hat{T}_4(x_1, \dots, x_4) = \{(0, 1)\}, \end{aligned}$$

where conv stands for the convex hull. We see that for $\|\cdot\|_2$ and $\|\cdot\|_\infty$ the geometric median is determined in a unique way. This does not hold any more for $\|\cdot\|_1$.

Example 2.2. Consider $\bar{x} \in X$ and $\mathbf{x} = (x_1, \dots, x_n)$, where $x_1 = \dots = x_m = \bar{x}$ for some $m \geq \frac{n}{2}$. Fix any $y \in X$. Then we have

$$\begin{aligned} \sum_{i=1}^n \|\bar{x} - x_i\| &= \sum_{i=m+1}^n \|\bar{x} - x_i\| \leq \sum_{i=m+1}^n \|\bar{x} - y\| + \sum_{i=m+1}^n \|y - x_i\| \\ &= \sum_{i=m+1}^n \|\bar{x} - y\| + \sum_{i=m+1}^n \|y - x_i\| + \sum_{i=1}^m \|y - x_i\| - \sum_{i=1}^m \|y - \bar{x}\| \\ &= \sum_{i=1}^n \|y - x_i\| + (n - 2m)\|y - \bar{x}\| \leq \sum_{i=1}^n \|y - x_i\| \end{aligned}$$

due to the $m \geq \frac{n}{2}$. But this means that $\bar{x} \in \hat{T}_n(\mathbf{x})$.

Recall that a shift equivariant estimator T_n satisfies $T_n(x_1 + y, \dots, x_n + y) = T_n(x_1, \dots, x_n) + y$ for all $x_1, \dots, x_n \in X$ and $y \in X$. For single-valued estimators, it has been

shown in Maronna et al. (2006, formula (3.25)) that a shift equivariant estimator satisfies

$$\varepsilon_n^*(T_n, \mathbf{x}) \leq \frac{1}{n} \left\lfloor \frac{n-1}{2} \right\rfloor. \quad (2)$$

Since the geometric median possesses this property, it is not surprising, that we obtain formula (2) as well. Moreover, we obtain even equality in this estimate.

Lemma 2.1. *For any $\mathbf{x} = (x_1, \dots, x_n) \in X^n$, we have for the geometric median*

$$\varepsilon_n^*(\hat{T}_n, \mathbf{x}) = \frac{1}{n} \left\lfloor \frac{n-1}{2} \right\rfloor,$$

and thus for the asymptotic breakdown point we have $\varepsilon^* = \frac{1}{2}$.

Proof. The second statement is an immediate consequence of the first one. Note that the first statement is equivalent to $m_n^*(\hat{T}_n, \mathbf{x}) = n_0$ with $n_0 := \lfloor \frac{n-1}{2} \rfloor$. From Example 2.2, we see that $m_n^*(\hat{T}_n, \mathbf{x}) < \frac{n}{2}$, which further implies $m_n^*(\hat{T}_n, \mathbf{x}) \leq \frac{n}{2} - 1 \leq n_0$. To finish the proof, it is sufficient to show that $m_n^*(\hat{T}_n, \mathbf{x}) \geq n_0$.

Consider thus any $\tilde{\mathbf{x}} \in A_{n_0, n}(\mathbf{x})$ and denote by I the index set of coordinates where \mathbf{x} and $\tilde{\mathbf{x}}$ differ and by J its complement. Denote by n_1 the cardinality of I and observe that $n_1 \leq n_0$. Denoting further $R := \max_{i=1, \dots, n} \|x_i\|$, we have $\|\tilde{x}_j\| = \|x_j\| \leq R$ for all $j \in J$. Taking any $j \in J$ and $y \in X$, we obtain the following estimate:

$$\begin{aligned} \sum_{l=1}^n \|\tilde{x}_j - \tilde{x}_l\| &\leq \sum_{l \in I} \|\tilde{x}_j - \tilde{x}_l\| + 2(n - n_1)R \leq \sum_{l \in I} \|\tilde{x}_j - y\| + \sum_{l \in I} \|y - \tilde{x}_l\| + 2(n - n_1)R \\ &= \sum_{l=1}^n \|y - \tilde{x}_l\| - \sum_{l \in J} \|y - \tilde{x}_l\| + n_1 \|\tilde{x}_j - y\| + 2(n - n_1)R \\ &\leq \sum_{l=1}^n \|y - \tilde{x}_l\| - \sum_{l \in J} \|y - \tilde{x}_l\| + \sum_{l \in J} \|\tilde{x}_j - \tilde{x}_l\| + n_1 \|\tilde{x}_j - y\| + 2(n - n_1)R \\ &\leq \sum_{l=1}^n \|y - \tilde{x}_l\| + (2n_1 - n) \|y - \tilde{x}_j\| + 4(n - n_1)R. \end{aligned}$$

Since $2n_1 - n \leq 2n_0 - n < 0$, we obtain that there is $\tilde{R}_I > 0$ such that for all $\|y\| \geq \tilde{R}_I$ we have

$$\sum_{l=1}^n \|\tilde{x}_j - \tilde{x}_l\| < \sum_{l=1}^n \|y - \tilde{x}_l\|.$$

But this means that the geometric median lies in a ball with radius \tilde{R}_I . Since there is only finite number of possible subsets I , we have finished the proof. \square

The next lemma allows us to compute the breakdown point of an estimator.

Lemma 2.2. *Consider multifunctions $\Phi_1 : X^n \rightrightarrows X^m$ and $\Phi_2 : X^n \times X^m \rightrightarrows X$. Assume that the following assumptions are satisfied:*

1. *All components of Φ_1 have breakdown point at least p .*
2. *There exists $\Phi_3 : X^m \rightrightarrows X$ which is bounded on bounded sets such that $\|\Phi_2(\mathbf{x}, \mathbf{y})\| \leq \|\Phi_3(\mathbf{y})\|$ for all $\mathbf{x} \in X^n$ and $\mathbf{y} \in X^m$.*

Then estimator T_n defined as

$$T_n(\mathbf{x}) := \bigcup_{\mathbf{y} \in \Phi_1(\mathbf{x})} \Phi_2(\mathbf{x}, \mathbf{y})$$

has breakdown point at least p .

Proof. Due to the first assumption, there exists some $R > 0$ such that for $m := np$, all $\tilde{\mathbf{x}} \in A_{m,n}(\mathbf{x})$ and for all $\mathbf{y} \in \Phi_1(\tilde{\mathbf{x}})$ we have $\|\mathbf{y}\| \leq R$. But then we have $\|\Phi_2(\tilde{\mathbf{x}}, \mathbf{y})\| \leq \|\Phi_3(\mathbf{y})\|$, which is uniformly bounded due to the second assumption. Thus, the statement has been proved. \square

We come now to new estimators. For a set $S \subset X$ and a point $\mathbf{x} = (x_1, \dots, x_n) \in X^n$, we define

$$\mathcal{L}(S, \mathbf{x}) := \bigcup_{y \in S} \left\{ \mathbf{x}_I \in X^{\lfloor \frac{n-1}{2} \rfloor} \mid \exists I \subset \{1, \dots, n\} : \max_{i \in I} \|x_i - y\| \leq \min_{i \in \{1, \dots, n\} \setminus I} \|x_i - y\| \right\},$$

where \mathbf{x}_I denotes the restriction of \mathbf{x} to components I . The interpretation of this set goes as follows: we select some $y \in S$ and choose \mathbf{x}_I to be the $\lfloor \frac{n-1}{2} \rfloor$ observations closest to y . Then $\mathcal{L}(S, \mathbf{x})$ is the union of all such subsets with respect to all choices of $y \in S$. Since every such \mathbf{x}_I contains less than $\frac{n}{2}$ components of \mathbf{x} , this set is stable with respect to perturbations of \mathbf{x} whenever less than one half of observations is contaminated.

We will use $\mathcal{L}(S, \mathbf{x})$ to define further estimators. The next theorem says that if we start with the geometric median $S = \hat{T}_n(\mathbf{x})$ and a multifunction R with certain boundedness properties, we obtain an estimator with the same breakdown point as the geometric median.

Theorem 2.1. Consider any multifunction $R : X^{\lfloor \frac{n-1}{2} \rfloor} \rightrightarrows X$, which is bounded on bounded sets and for which there exists z^k such that $\|R(z^k, \dots, z^k)\| \rightarrow \infty$. Then for estimator $T_n : X^n \rightrightarrows X$ defined as

$$T_n^1(\mathbf{x}) := \bigcup_{\mathbf{y} \in \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})} R(\mathbf{y}) \quad (3)$$

and for every $\mathbf{x} = (x_1, \dots, x_n) \in X^n$, we have the following relation

$$\varepsilon_n^*(T_n^1, \mathbf{x}) = \varepsilon_n^*(\hat{T}_n, \mathbf{x}) = \frac{1}{n} \left\lfloor \frac{n-1}{2} \right\rfloor.$$

Proof. From Lemma 2.2 with $m = \lfloor \frac{n-1}{2} \rfloor$, $\Phi_1(\mathbf{x}) = \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})$, and $\Phi_2(\mathbf{x}, \mathbf{y}) = R(\mathbf{y})$ and Lemma 2.1 we obtain

$$\varepsilon_n^*(T_n^1, \mathbf{x}) \geq \varepsilon_n^*(\hat{T}_n, \mathbf{x}) = \frac{1}{n} \left\lfloor \frac{n-1}{2} \right\rfloor.$$

To show the opposite inequality, realize that the statement is equivalent to $m_n^*(T_n^1, \mathbf{x}) \leq \lfloor \frac{n-1}{2} \rfloor$. For contradiction assume that $m_n^*(T_n^1, \mathbf{x}) \geq \lfloor \frac{n-1}{2} \rfloor + 1 \geq \lfloor \frac{n}{2} \rfloor$. We change the first $\lfloor \frac{n}{2} \rfloor$ coordinates of \mathbf{x} to z^k and denote the perturbed point by $\tilde{\mathbf{x}}^k$. Then Example 2.2 tells us that $z^k \in \hat{T}_n(\tilde{\mathbf{x}}^k)$. Due to the definition of \mathcal{L} , we see that $(z^k, \dots, z^k) \in \mathcal{L}(\hat{T}_n(\tilde{\mathbf{x}}^k), \tilde{\mathbf{x}}^k)$ and the imposed assumption of R implies a contradiction. \square

If both R and $\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})$ are single-valued functions, expression (3) reduces to

$$T_n^1(\mathbf{x}) = R(\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})).$$

Moreover, in such a case $\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})$ denotes one half of observations, which are closest to the geometric median. There are several natural choices for R : for example, mean, weighted mean, or geometric median.

One of possible drawbacks of estimator (3) is that it utilizes only half of the original data. We want to make use of as many observations as possible while maintaining the high breakdown point. To this aim, we first consider a general set $\mathcal{S} \subset X^m$ for some $m \in \mathbb{N}$, for example, we may consider mean of \mathbf{x} as a subset of X or $\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})$ as a subset of $X^{\lfloor \frac{n-1}{2} \rfloor}$. Then we consider some $b : X \times X^n \rightarrow [0, \infty)$ and enlarge \mathcal{S} by defining

$$\mathcal{E}_b(\mathcal{S}, \mathbf{x}) := \bigcup_{\mathbf{y} \in \mathcal{S}} \left\{ \mathbf{x}_I \mid I = \{i \mid x_i \in \bigcup_{j=1}^m \mathbb{B}(y_j, b(y_j, \mathbf{x}))\} \right\}. \quad (4)$$

Here, $\mathbb{B}(y_j, b(y_j, \mathbf{x}))$ stands for a ball around y_j with radius $b(y_j, \mathbf{x})$. The interpretation goes as follows: from \mathcal{S} we select \mathbf{y} , make balls around all of its components, and select all components of \mathbf{x} , which lie in the union of this balls.

Example 2.3. Consider the case of $X = \mathbb{R}$, $n = 5$, and $\mathbf{x} = (-3, -2, 0, 2, 4)$. Then the geometric median equals to $\hat{T}_n(\mathbf{x}) = 0$ and since $n_0 = \lfloor \frac{n-1}{2} \rfloor = 2$, we also have

$$\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}) = \{(-2, 0), (0, 2)\} \subset \mathbb{R}^2.$$

If we consider $b \equiv 1$, then

$$\mathcal{E}_b(\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}), \mathbf{x}) = \{(-3, -2, 0), (0, 2)\}.$$

Note that both elements of $\mathcal{E}_b(\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}), \mathbf{x})$ are of a different dimension.

We obtain the following variant of [Theorem 2.1](#), for which we omit its identical proof.

Theorem 2.2. Consider $b : X \times X^n \rightarrow [0, \infty)$ bounded on bounded sets in the first variable, uniformly in the second one, any family of multifunctions $R_s : X^s \rightrightarrows X$ for $s = 1, \dots, n$, which are all bounded on bounded sets and for which there exists z^k such that $\|R_s(z^k, \dots, z^k)\| \rightarrow \infty$. Then for estimators $T_n^2 : X^n \rightrightarrows X$ and $T_n^3 : X^n \rightrightarrows X$ defined as

$$T_n^2(\mathbf{x}) := \bigcup_{\mathbf{y} \in \mathcal{E}_b(\hat{T}_n(\mathbf{x}), \mathbf{x})} R_{\dim \mathbf{y}}(\mathbf{y}), \quad (5a)$$

$$T_n^3(\mathbf{x}) := \bigcup_{\mathbf{y} \in \mathcal{E}_b(\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}), \mathbf{x})} R_{\dim \mathbf{y}}(\mathbf{y}) \quad (5b)$$

and for every $\mathbf{x} = (x_1, \dots, x_n) \in X^n$, we have the following relation for breakdown points

$$\varepsilon_n^*(T_n^2, \mathbf{x}) = \varepsilon_n^*(T_n^3, \mathbf{x}) = \varepsilon_n^*(\hat{T}_n, \mathbf{x}) = \frac{1}{n} \left\lfloor \frac{n-1}{2} \right\rfloor.$$

Function b should neither have too large values (which corresponds to big enlargement of the set in question) because outliers may be close to the non-contaminated data, nor too small values because some information could be missed. We suggest a possible choice in the [Appendix](#).

To improve the behavior of the estimators, we implement an iterative procedure. We start with geometric median $z^0 = \hat{T}_n(\mathbf{x})$ and in every iteration k compute a new estimate z^k . To do so, we employ (5b) with R being the weighted mean, where the (non-normalized)

weights satisfy

$$w_i(z^{k-1}, y_i) = \begin{cases} 1 & \text{if } y_i \in \mathcal{L}(\{z^{k-1}\}, \mathbf{x}), \\ \max_{y_j \in \mathcal{L}(\{z^{k-1}\}, \mathbf{x})} \left(1 - \frac{\|y_i - y_j\|}{b^k(y_j, \mathbf{x})}\right) & \text{otherwise} \end{cases} \quad (6)$$

for some b^k based on $\mathcal{L}(\{z^{k-1}\}, \mathbf{x})$. This choice of weights makes use of the possibly division of components of $\mathbf{y} \in \mathcal{E}_b(\mathcal{L}(\{z^{k-1}\}, \mathbf{x}), \mathbf{x})$ into two parts: those who belong to $\mathcal{L}(\{z^{k-1}\}, \mathbf{x})$ and those who were added by enlarging this set. For the first part, we choose the (non-normalized) weight equal to one, the weight for observations from the second part decreases with the increasing distance to $\mathcal{L}(\{z^{k-1}\}, \mathbf{x})$. We summarize this approach in Algorithm 2.1. Considering the termination criterion, any standard criterion may be used, for example, if the (relative) change in z^k is small.

Algorithm 2.1 An estimator based on iterative weighting

Input: observations $\mathbf{x} = (x_1, \dots, x_n)$

- 1: $k \leftarrow 0, z^0 \leftarrow \hat{T}_n(\mathbf{x})$
 - 2: **while not terminate do**
 - 3: $k \leftarrow k + 1$
 - 4: determine b^k based on $\mathcal{L}(\{z^{k-1}\}, \mathbf{x})$
 - 5: pick any $\mathbf{y}^k \in \mathcal{E}_{b^k}(\mathcal{L}(\{z^{k-1}\}, \mathbf{x}), \mathbf{x})$
 - 6: compute w^k according to formula (6) and renorm them such that their sum equals to 1
 - 7: $z^k \leftarrow \sum_{i=1}^{\dim \mathbf{y}} w_i^k y_i^k$
 - 8: **end while**
 - 9: **return** estimate $\hat{x} \leftarrow z^k$
-

Finally, we would like to point out that every update step in Algorithm 2.1 keeps the stability result, which we already mentioned several times in the previous text. For $k = 1$, we may write

$$z^k = \bigcup_{\mathbf{y} \in \Phi_1(\mathbf{x})} \Phi_2(\mathbf{x}, \mathbf{y}),$$

where $\mathbf{y} = \mathbf{y}^1$, $\Phi_1(\mathbf{x}) := \mathcal{E}_b(\mathcal{L}(\{z^0\}, \mathbf{x}), \mathbf{x})$, and $\Phi_2(\mathbf{x}, \mathbf{y}) := \sum_{i=1}^{\dim \mathbf{y}^1} w_i(\mathbf{x}, \mathbf{y}^1) y_i^1$. Then Φ_1 has breakdown point $\frac{1}{n} \lfloor \frac{n-1}{2} \rfloor$ and since $\|\Phi_2(\mathbf{x}, \mathbf{y})\| \leq n \max_{y_i \in \mathbf{y}} \|y_i\|$ holds true, thanks to [Lemma 2.2](#) we obtain that z^1 has the same breakdown point. By applying the same procedure to subsequent iterations, we obtain the same result for all z^k .

In the previous text, we highlighted some benefits of our estimators. Note that naturally there are also situations where it is better to use standard estimators. Consider, for example, the one-dimensional double exponential distribution with density $\frac{1}{2} \exp(-|x - \mu|)$. Then the (geometric) median coincides with the maximum likelihood estimator of μ , see sec. 6.3 in [Lehmann and Casella \(2006\)](#), and therefore the median is the most efficient estimator. Thus, by employing additional observations apart from the median, we only worsen the quality of an estimator. However, to benefit from such situation, we would have to know the true distribution and know that there is no contamination.

Remark 2.1. There is some connection between our estimators with M-estimators. In Algorithm 2.2, we summarize the algorithm from [Maronna et al. \(2006, sec. 2.7.3\)](#).

Algorithm 2.2 M-estimator from**Require:** observations $\mathbf{x} = (x_1, \dots, x_n)$, weighting functions W_1 and W_2

- 1: $k \leftarrow 0, z^0 \leftarrow \hat{T}_n(\mathbf{x})$, dispersion estimate σ^0
- 2: **while not terminate do**
- 3: $k \leftarrow k + 1$
- 4: $r_i^k \leftarrow \frac{x_i - z^{k-1}}{\sigma^{k-1}}$
- 5: $w_{1,i}^k \leftarrow W_1(r_i^k), w_{2,i}^k \leftarrow W_2(r_i^k)$ and norm the weights to sum to one
- 6: $z^k \leftarrow \sum_{i=1}^n w_{1,i}^k x_i, \sigma^k \leftarrow \frac{1}{n} \sum_{i=1}^n w_{2,i}^k (x_i - z^{k-1})^2$
- 7: **end while**
- 8: **return** estimate $\hat{x} \leftarrow z^k$

Both approaches (iteratively) compute a weighted mean of observations. While M-estimators are based on the maximum likelihood estimate, ours are based on geometric intuition and trimmed mean. If $X = \mathbb{R}$ and if W_1 in Algorithm 2.2 has finite support, we can say that our estimators belong to the very wide class of M-estimators. This changes for \mathbb{R}^d though. Under the standard assumption that W_1 is symmetric around zero and non-increasing on rays emanating from zero, the weight of an observation depends only on the distance from z^{k-1} . Thus, two observations have the same weight if and only if their distance to z^{k-1} is identical. On the other hand, in our approach all points in $\mathcal{L}(\{z^{k-1}\}, \mathbf{x})$ have the same weight and this set is enlarged farther for distant observations. Thus, even though none of the algorithms estimates the covariance structure, our algorithm makes at least an attempt to consider it.

Of course, there are M-estimators, which along with the location also properly estimate the covariance structure. But this raises the computational complexity and reduces the breakdown point to $\frac{1}{d+1}$. To summarize: we can say that our algorithms try to pick the best properties of M-estimators, on one hand they have high breakdown point and are simple to compute; on the other hand, they at least partially consider the covariance structure.

Another advantage of our estimator over M-estimators is a simpler theoretical analysis. To show that our estimator has the limiting breakdown point $\frac{1}{2}$, it is sufficient to apply [Lemma 2.2](#), which itself directly follows from the definition of the breakdown point. To the best of our knowledge, such direct application of the definition is not possible for M-estimators, for example, one has to take care of properties of weighting function due to the division by dispersion.

3. Numerical results

In this section, we first show numerical performance of our estimators and then show how they comply with the well-known concepts of boxplot and bagplot.

3.1. Numerical performance

We consider $X = \mathbb{R}^d$ with $d \in \{1, 15\}$ and compare our algorithms with known estimators; for reader's convenience we summarize the used algorithms in [Table 1](#).

To generate the samples, we first generate z_i from $N(0, 1)$, then contaminate them by some distribution with probability p and finally modify them via a covariance structure. This modification is performed in the following way: we randomly generate a correlation matrix C and diagonal matrix Σ^2 with diagonal elements having distribution $U[0.5, 10]$. Then we compute covariance matrix $V = \Sigma C \Sigma$, its Cholesky decomposition $V = S^\top S$, and finally set $y_i = S z_i + \mu$, where $\mu := (0, \dots, d-1)$. Together we consider $N = 10,000$ samples with

Table 1. Summary of algorithms. The horizontal line divides them in known and our algorithms.

Mean	mean
Med	median or geometric median
Trun	α -truncated mean with $\alpha = 0.2$
Winsor	α -winsorized mean with $\alpha = 0.2$
M1	Huber M-estimator, see Huber (1981)
M2	Algorithm 2.2 from Maronna et al. (2006)
SD	Stahel–Donoho estimator, see Stahel (1981) and Donoho (1982)
GM1	formula (3), where R is the mean
GM2	formula (5a), where R is the mean
GM3	Algorithm 2.1

Table 2. Value of loss function (7) for contaminations of $N(0, 1)$ by some other distribution with probability p for $X = \mathbb{R}$.

Distribution	p	Mean	Med	Trun	Winsor	M1	M2	GM1	GM2	GM3
$N(0, 100)$	0	0.080	0.099	0.085	<i>0.083</i>	<i>0.081</i>	<i>0.081</i>	0.116	0.085	<i>0.083</i>
Cauchy	0.05	0.190	0.103	0.089	0.089	0.084	<i>0.088</i>	0.119	<i>0.087</i>	<i>0.086</i>
$U(-10, 10)$	0.05	0.657	0.099	<i>0.086</i>	<i>0.084</i>	0.082	<i>0.083</i>	0.116	<i>0.086</i>	<i>0.084</i>
$N(0, 100)$	0.1	0.129	0.103	0.090	<i>0.089</i>	0.084	<i>0.088</i>	0.120	<i>0.087</i>	<i>0.086</i>
Cauchy	0.1	0.259	0.110	0.096	0.097	0.089	0.098	0.125	<i>0.090</i>	<i>0.089</i>
$U(-10, 10)$	0.1	0.642	0.101	<i>0.088</i>	<i>0.087</i>	0.084	<i>0.086</i>	0.116	<i>0.088</i>	<i>0.086</i>
$U(-20, 10)$	0.1	0.162	0.108	0.095	0.096	0.090	0.096	0.123	<i>0.090</i>	<i>0.090</i>
	0.4	1.992	0.288	0.565	1.098	0.192	0.596	0.198	<i>0.161</i>	0.153

$n = 100$ observations. The loss function equals to

$$\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^d |\hat{x}_{i,j} - \mu_j|, \tag{7}$$

where $\hat{x}_{i,j}$ denotes an estimate for sample i and coordinate $j = 1, \dots, d$.

We present the results in Table 2 for $X = \mathbb{R}$ and in Table 3 for $X = \mathbb{R}^{15}$. The values show the loss function (7) for contaminating distribution (first column) with given probability (second column). Bold numbers are the best values among all estimators and numbers in italic are those within 5% of the best loss function value. For $X = \mathbb{R}$, the results between M-estimators and our estimators are comparable with M-estimators in a slight lead, which is not surprising due to Remark 2.1. For $X = \mathbb{R}^{15}$, the performance of both estimators turn around and now our estimators perform better than the examined M-estimators. This is again expected as our estimators try to take into account the covariance structure as explained in Remark 2.1. For the second case, our estimators manage to beat the Stahel–Donoho estimators almost in all cases.

Table 3. Value of loss function (7) for contaminations of $N(0, I)$ by some other distribution with probability p for $X = \mathbb{R}^{15}$.

Distribution	p	Mean	Med	M1	SD	GM1	GM2	GM3
$N(0, 100)$	0	0.151	<i>0.154</i>	<i>0.151</i>	0.159	0.213	<i>0.151</i>	<i>0.152</i>
Cauchy	0.05	0.360	<i>0.162</i>	0.155	<i>0.160</i>	0.215	<i>0.155</i>	<i>0.156</i>
$U(-10, 10)$	0.05	2.424	<i>0.160</i>	<i>0.160</i>	<i>0.160</i>	0.214	0.155	<i>0.158</i>
$N(0, 100)$	0.1	0.241	<i>0.161</i>	<i>0.160</i>	<i>0.160</i>	0.215	0.155	<i>0.157</i>
Cauchy	0.1	0.491	0.169	<i>0.160</i>	<i>0.161</i>	0.217	0.159	<i>0.160</i>
$U(-10, 10)$	0.1	2.095	<i>0.167</i>	0.171	<i>0.162</i>	0.216	0.160	<i>0.165</i>
$U(-20, 10)$	0.1	0.308	0.168	0.171	<i>0.162</i>	0.217	0.159	<i>0.162</i>
	0.4	3.586	0.610	0.371	0.194	0.234	0.215	0.214

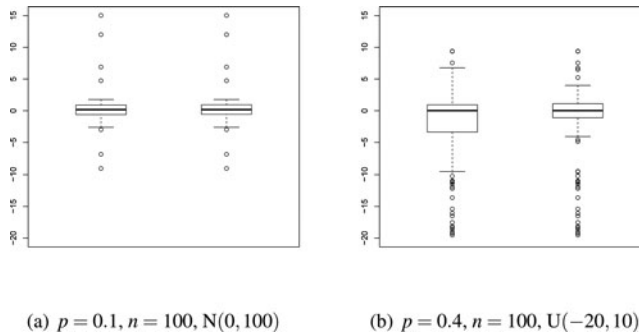


Figure 1. Comparison of the classical boxplot (left-hand side of each figure) and our modification (right-hand side of each figure). In both cases, we contaminate $N(0, 1)$ with a probability p by the distribution described under the figures.

3.2. Relation to boxplot

Boxplot was proposed for the first time in Tukey (1977). It takes the median, then computes the interquartile range (IQR), which is later widened. Observations, which are not present in this widening (known as whiskers) are considered as outliers. We compare boxplot with our method, where instead of considering IQR, we take $\mathbf{y} \in \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})$. Thus, we do not need to take 25% observations with lower value than the median and 25% observations with higher value than the median but we take 50% observations closest to the median. Whiskers are based on $\mathcal{E}_b(\{\mathbf{y}\}, \mathbf{x})$.

We depict this comparison in Fig. 1. We contaminate the standard normal distribution by some other distribution. In each (sub)figure, the left-hand side is the boxplot and the right-hand side is our modification. Since both approaches differ in the way in which they treat non-symmetry, both graphs in the left figure are identical. However, if we assume non-symmetric distributions (right figure), then our modification is able to detect outliers in a better way than the standard boxplot.

3.3. Relation to bagplot

In this subsection, we consider generalization of boxplot into more dimensions. For two dimensions, this is known as bagplot and was studied for the first time in Rousseeuw et al. (1999). For generalization to functional data, see, for example, Sun and Genton (2011).

To construct bagplot, a real number called depth is assigned to every observation, see Zuo and Serfling (2000). Then the observation with highest value of depth is called the depth median and the convex hull of approximately 50% observations with the highest depth is called the bag (this corresponds to IQR for boxplot). Then the bag is enlarged to the so-called fence (which corresponds to whiskers for boxplot). The observations not in the fence are considered as outliers.

We now present our modification of the bagplot, illustrated in Fig. 2. Having a sample \mathbf{x} , we compute first its geometric median $\hat{T}_n(\mathbf{x})$. For simplicity, we assume that it is defined uniquely. Then we choose an arbitrary $\mathbf{y} \in X^{\lfloor \frac{n-1}{2} \rfloor}$ from $\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})$, construct a convex hull containing all coordinates of \mathbf{y} , and call this set the bag. In the last step, we enlarge the bag by considering

$$\text{conv}_i\{\hat{T}_n(\mathbf{x}) + 3(y_i - \hat{T}_n(\mathbf{x}))\},$$

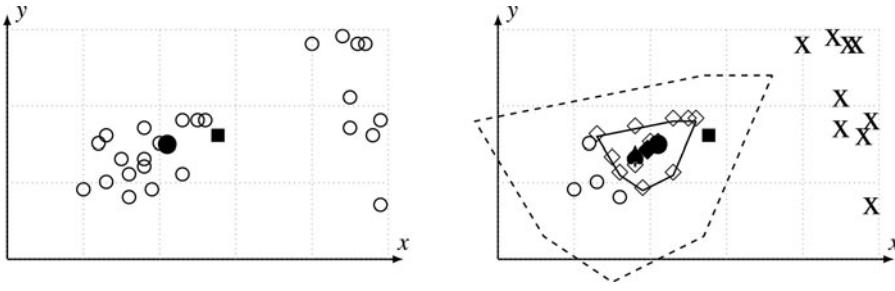


Figure 2. Description of the bagplot based on GM2. \circ : observations, \blacksquare : mean, \bullet : $\hat{T}_n(\mathbf{x})$, full and dashed line: convex envelope of \mathcal{L} and \mathcal{E} , respectively, \blacklozenge : GM1, \blacklozenge : GM2, X: outliers.

see a similar procedure in Rousseeuw et al. (1999). In this figure, we also denote GM1 (as the mean of all observation in the bag) and GM2 (as the mean of all observation in the fence).

Here, the geometric median corresponds to the depth median, the convex hull to the bag, and its enlargement to the fence. Note that the construction of the fence is very similar to the construction of $\mathcal{E}_b(\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}), \mathbf{x})$. Moreover, since the construction of GM2 is based on objects, which have the limiting breakdown point of $\frac{1}{2}$, estimator GM2 has the same property. On the other hand, since bagplot is a generalization of boxplot, which has breakdown point of $\frac{1}{4}$, we cannot expect the bagplot to have higher breakdown point than this. Thus, our version of bagplot deals better with outliers than the original method.

Appendix: Choice of b

In this short section, we derive an estimate for b for algorithms GM2 and GM3 described in Table 1. Note that due to the construction of the algorithm, it is sufficient to define $b_i := b(x_i, \mathbf{x})$ for all observations in $\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})$. Since we want to keep GM2 as simple as possible, we consider constant b and relax this assumption for GM3. Moreover, we derive different value for one- and more-dimensional cases. We start with a technical lemma.

Lemma A.1. *Let Y be a random variable with a finite second moment, distribution function G , mean μ , standard deviation σ , and let $q_{\mu,\sigma}$ denote its quantile function. Then for a fixed $\alpha \in [0, 1]$, the following ratio does not depend on the values of μ and σ*

$$K_{G,\alpha} = \frac{q_{\mu,\sigma}(1 - \alpha/2) - q_{\mu,\sigma}(\alpha/2)}{q_{\mu,\sigma}(0.75) - q_{\mu,\sigma}(0.25)}.$$

Proof. This follows from the fact $\sigma q_{0,1}(\alpha) + \mu = q_{\mu,\sigma}(\alpha)$. □

For the case of one dimension, consider a random sample \mathbf{x} from $N(\mu, \sigma^2)$ and denote its median by $\hat{T}_n(\mathbf{x})$. For simplicity assume that $\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}) \subset \mathbb{R}^{\lfloor \frac{n-1}{2} \rfloor}$ is a singleton. Then we may estimate $q_{\mu,\sigma}(0.75) - q_{\mu,\sigma}(0.25)$ by

$$\max \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}) - \min \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}).$$

Then we define b as

$$b := K_{N(0,1),\alpha} \frac{\max \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x}) - \min \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})}{2} \asymp \frac{q_{\mu,\sigma}(1 - \alpha/2) - q_{\mu,\sigma}(\alpha/2)}{2}.$$

This value thus estimates the distance between chosen quantiles divided by two. If we take all observations, which distance from median is b , then we get, under assumption of not contaminated normality, approximately $1 - \alpha$ of our observations.

For the multidimensional case \mathbb{R}^d , we assume that \mathbf{x} follows the $N(\mu, \sigma^2 I)$ distribution. We use $\max_{y \in \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})} \|y - \hat{T}_n(\mathbf{x})\|$ as an approximation of $\sigma \chi_{0.5, d}^2$, where $\chi_{0.5, d}^2$ is the quantile function of χ^2 distribution with d degrees of freedom evaluated at probability 0.5. To include approximately fraction $\alpha \in (0, 1)$ of all observations, we set

$$b := \max_{y \in \mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})} \|y - \hat{T}_n(\mathbf{x})\| \frac{\chi_{\alpha, d}^2}{\chi_{0.5, d}^2}.$$

For GM3, we consider directly \mathbb{R}^d . Assume again that our sample \mathbf{x} has distribution $N(\mu, \sigma^2 I)$, then for x_i from the boundary of $\mathcal{L}(\hat{T}_n(\mathbf{x}), \mathbf{x})$ we have

$$\frac{\|x_i - \hat{T}_n(\mathbf{x})\|}{\sigma} \sim \sqrt{\chi_{0.5, d}^2}.$$

Now fixing given probability level $\alpha \in (0, 1)$ and weight $w \in (0, 1)$, we want to have all x_e with

$$\frac{\|x_e - \hat{T}_n(\mathbf{x})\|}{\sigma} \sim \sqrt{\chi_{\alpha, d}^2}$$

to have weight (6) at least w . But plugging this in the definition of weight results in

$$\begin{aligned} b_i &\geq \frac{\|x_i - x_e\|}{1 - w} \geq \frac{\|x_e - \hat{T}_n(\mathbf{x})\|}{1 - w} - \frac{\|x_i - \hat{T}_n(\mathbf{x})\|}{1 - w} \sim \frac{\sigma}{1 - w} \left(\sqrt{\chi_{\alpha, d}^2} - \sqrt{\chi_{0.5, d}^2} \right) \\ &= \frac{\sigma}{1 - w} \frac{\|x_i - \hat{T}_n(\mathbf{x})\|}{\|x_i - \hat{T}_n(\mathbf{x})\|} \left(\sqrt{\chi_{\alpha, d}^2} - \sqrt{\chi_{0.5, d}^2} \right). \end{aligned}$$

Approximating again the distribution and taking minimum value of b_i , we set

$$b_i := \frac{\sqrt{\chi_{\alpha, d}^2} - \sqrt{\chi_{0.5, d}^2}}{(1 - w)\sqrt{\chi_{0.5, d}^2}} \|x_i - \hat{T}_n(\mathbf{x})\|.$$

In the numerical experiments, we have chosen $\alpha = 0.99$.

Acknowledgment

The authors would like to thank anonymous reviewers for substantially improving this article.

ORCID

Lukáš Adam  <http://orcid.org/0000-0001-8748-4308>

References

- Davies, P. L. (1987). Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics* 15:1269–1292.
- Donoho, D. L. (1982). *Breakdown Properties of Multivariate Location Estimators*. Ph.D. Qualifying Paper, Harvard University.

- Donoho, D. L., Huber, P. J. (1983). The notion of breakdown point. In: Bickel, P. J., Doksum, K. A., and Hodges, J. L., eds. *A Festschrift for Erich Lehman*. California, Berkeley: Wadsworth, pp. 157–184.
- Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika* 35(3–4):414–417.
- Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics* 42(6):1887–1896.
- Huber, P. J. (1964). Robust estimation of location parameter. *The Annals of Mathematical Statistics* 35(1):73–101.
- Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley & Sons.
- Lehmann, E. L., Casella, G. (2006). *Theory of Point Estimation*. New York: Springer Science & Business Media.
- Lopuhaä, H. P. (1991). Multivariate τ -estimators for location and scatter. *Canadian Journal of Statistics* 19:307–321.
- Maronna, R. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics* 4:51–67.
- Maronna, R., Martin, D., Yohai, V. (2006). *Robust Statistics Theory*. New Jersey: John Wiley & Sons.
- Maronna, R. A., Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* 44(4):307–317.
- Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In: *Mathematical Statistics and Applications*. Amsterdam: Elsevier, pp. 101–121.
- Rousseeuw, P. J., Ruts, I., Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *The American Statistician* 53(4):382–387.
- Stahel, W. A. (1981). *Breakdown of Covariance Estimators*. Research Report 31. ETH Zürich: Fachgruppe für Statistik.
- Sun, Y., Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics* 20(2):316–334.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Massachusetts: Addison-Wesley Publishing Co.
- Zuo, Y., Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics* 28(2):461–482.