STAT 593
Robust statistics:
Equivariance and breakdown point

Joseph Salmon
http://josephsalmon.eu

Télécom Paristech, Institut Mines-Télécom
&
University of Washington, Department of Statistics
(Visiting Assistant Professor)

Outline

Statistical invariance / equivariance

Breakdown point

Table of Contents

Dataset / point clouds and statistics

In this part we follow the concepts introduced by Donoho[1,2]: we right $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$ for the "cloud" of points representing n points in the space $\mathbb{R}^p$.

A **statistic** T is a (measurable) function from $\mathbb{R}^{p \times n}$ to $\mathbb{R}^{p'}$.
We write $T^{(n)}$ when the dependence on n is needed; we also use the notation $T(x_1, \ldots, x_n) = T(X)$ whenever needed.

Rem: often $p' = p$

Rem: notation different from standard design matrix (transposed)

[1] D. L. Donoho. "Breakdown properties of multivariate location estimators". PhD thesis. Harvard University, 1982.

[2] D. L. Donoho and M. Gasko. "Breakdown properties of location estimates based on halfspace depth and projected outlyingness". In: *Ann. Statist.* 20.4 (1992), pp. 1803–1827.

Transformations / invariance

For a permutation $\pi \in \mathfrak{S}_n$ we write:

$$\textbf{relabeling} : \pi(X) = [x_{\pi(1)}, \dots, x_{\pi(n)}]$$

Targeted property: **Permutation invariance**

$$\forall \pi \in \mathfrak{S}_n, T(\pi(X)) = T(X)$$

Interpretation: labeling should not matter to summarize a dataset

- Examples: mean, median, trimmed means, etc.
- Counter-example: *e.g.*, the first/last point (x_1 or x_n)

Translation

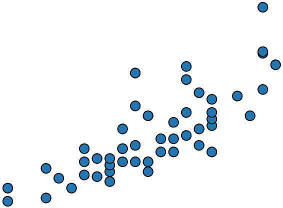
For a vector $\mu \in \mathbb{R}^p$ and a dataset X we write:

$$\textbf{Translation}: \; X + \mu = [x_1 + \mu, \ldots, x_n + \mu]$$

Translation

For a vector $\mu \in \mathbb{R}^p$ and a dataset X we write:

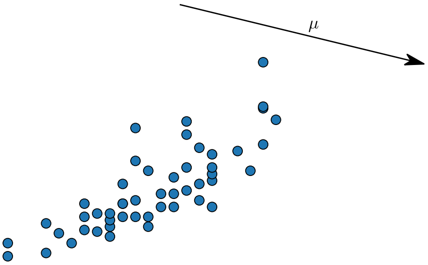
$$\textbf{Translation} : X + \mu = [x_1 + \mu, \ldots, x_n + \mu]$$



Translation

For a vector $\mu \in \mathbb{R}^p$ and a dataset X we write:

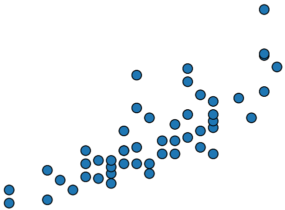
$$\textbf{Translation}: \; X + \mu = [x_1 + \mu, \ldots, x_n + \mu]$$



Translation

For a vector $\mu \in \mathbb{R}^p$ and a dataset X we write:

$$\textbf{Translation} : X + \mu = [x_1 + \mu, \ldots, x_n + \mu]$$



Translation equivariance

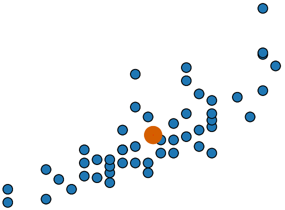
A statistic T is said **translation equivariant** if it satisfies:
for any vector $\mu \in \mathbb{R}^p$, and any dataset X the following holds

$$T(X + \mu) = T(X) + \mu$$

Translation equivariance

A statistic T is said **translation equivariant** if it satisfies:
for any vector $\mu \in \mathbb{R}^p$, and any dataset X the following holds

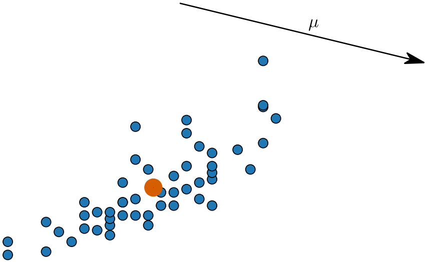
$$T(X + \mu) = T(X) + \mu$$



Translation equivariance

A statistic T is said **translation equivariant** if it satisfies:
for any vector $\mu \in \mathbb{R}^p$, and any dataset X the following holds

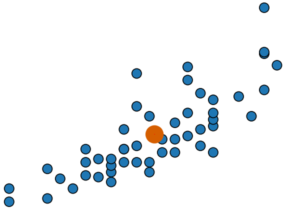
$$T(X + \mu) = T(X) + \mu$$



Translation equivariance

A statistic T is said **translation equivariant** if it satisfies:
for any vector $\mu \in \mathbb{R}^p$, and any dataset X the following holds

$$T(X+\mu) = T(X) + \mu$$



Translation equivariance (bis)

- Examples: mean, median, trimmed means, etc.
- Counter-example: **shrinkage** estimators, *e.g.*, James-Stein estimator $(n = 1, p > 2)$

$$\widehat{\boldsymbol{\mu}}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|x_1\|^2}\right) x_1, \text{or} \left(1 - \frac{(p-2)\sigma^2}{\|x_1\|^2}\right)_+ x_1$$

or extension with n observations:

$$\widehat{\boldsymbol{\mu}}_{JS} = \left(1 - \frac{(p-2)\frac{\sigma^2}{n}}{\|\overline{x}_n\|^2}\right) \overline{x}_n \text{ or } \left(1 - \frac{(p-2)\frac{\sigma^2}{n}}{\|\overline{x}_n\|^2}\right)_+ \overline{x}_n$$

<u>Rem</u>: James-Stein useful when estimating the mean of *i.i.d.* Gaussian with variance σ^2

Location estimator

Definition: location estimator

A statistics T is a **location estimator** if it is both

- permutation invariant
- translation equivariant

Example :

- the empirical mean $T(X) = T(x_1, \ldots, x_n) = \overline{x}_n$
- more generally if T is linear, it is translation equivariant
- we will see that any M-estimator is translation equivariant

Affine transformation

For a vector $\mu \in \mathbb{R}^p$ and a nonsingular matrix $\Sigma \in \mathbb{R}^{p \times p}$ and a dataset X we write:

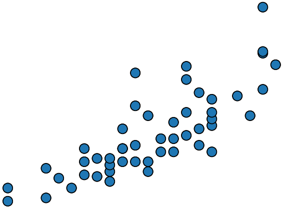
Affine transformation[3] : $\Sigma X + \mu = [\Sigma x_1 + \mu, \ldots, \Sigma x_n + \mu]$

[3] there is an abuse of notation as the matrix size do not match...

Affine transformation

For a vector $\mu \in \mathbb{R}^p$ and a nonsingular matrix $\Sigma \in \mathbb{R}^{p \times p}$ and a dataset X we write:

Affine transformation[3] $: \Sigma X + \mu = [\Sigma x_1 + \mu, \ldots, \Sigma x_n + \mu]$

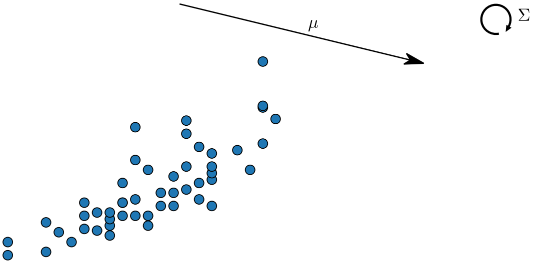


[3] there is an abuse of notation as the matrix size do not match...

Affine transformation

For a vector $\mu \in \mathbb{R}^p$ and a nonsingular matrix $\Sigma \in \mathbb{R}^{p \times p}$ and a dataset X we write:

Affine transformation[3] : $\Sigma X + \mu = [\Sigma x_1 + \mu, \ldots, \Sigma x_n + \mu]$

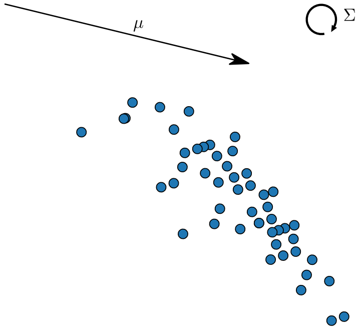


[3] there is an abuse of notation as the matrix size do not match...

Affine transformation

For a vector $\mu \in \mathbb{R}^p$ and a nonsingular matrix $\Sigma \in \mathbb{R}^{p \times p}$ and a dataset X we write:

Affine transformation[3] $: \Sigma X + \mu = [\Sigma x_1 + \mu, \ldots, \Sigma x_n + \mu]$



[3] there is an abuse of notation as the matrix size do not match...

Affine equivariance

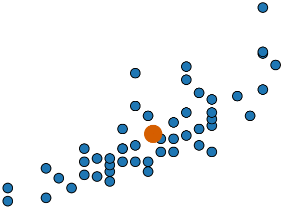
A statistic T is said **affine equivariant** if it satisfies:
For any nonsingular matrix $\Sigma \in \mathbb{R}^{p \times p}$, for any vector $\mu \in \mathbb{R}^p$ and for any dataset X the following holds:

$$T(\Sigma X + \mu) = \Sigma T(X) + \mu$$

Affine equivariance

A statistic T is said **affine equivariant** if it satisfies:
For any nonsingular matrix $\Sigma \in \mathbb{R}^{p \times p}$, for any vector $\mu \in \mathbb{R}^p$ and for any dataset X the following holds:

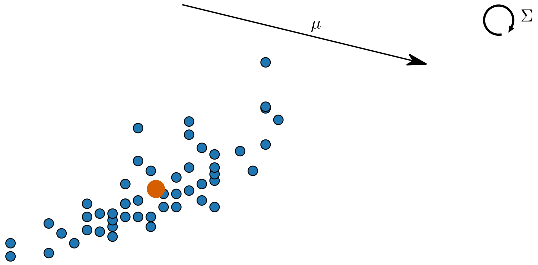
$$T(\Sigma X + \mu) = \Sigma T(X) + \mu$$



Affine equivariance

A statistic T is said **affine equivariant** if it satisfies:
For any nonsingular matrix $\Sigma \in \mathbb{R}^{p \times p}$, for any vector $\mu \in \mathbb{R}^p$ and for any dataset X the following holds:

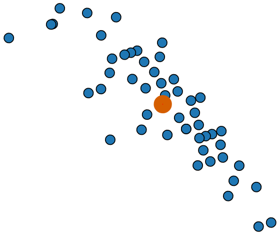
$$T(\Sigma X + \mu) = \Sigma T(X) + \mu$$



Affine equivariance

A statistic T is said **affine equivariant** if it satisfies:
For any nonsingular matrix $\Sigma \in \mathbb{R}^{p \times p}$, for any vector $\mu \in \mathbb{R}^p$ and for any dataset X the following holds:

$$T(\Sigma X + \mu) = \Sigma T(X) + \mu$$



Affine equivariance (bis)

A case of interest is the case: $\mu = 0$ and Σ is diagonal with with positive elements:

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{pmatrix}$$

This corresponds to scale equivariance, *i.e.*, the statistics should be equivariant w.r.t. change of unit (*e.g.*, kilometers vs miles)

Table of Contents

Breakpoint: history

A geometrical concept, though

- introduced by Hampel[4] in a probabilist framework
- the proposed formulation was provided by Donoho[5];
- another variant is provided in Maronna *et al.* (2006)

Donoho: "Imagine contaminating your dataset; how extensively must you contaminate it in order to make your estimator misbehave arbitrarily"

[4] F. R. Hampel. "Contributions to the theory of robust estimation". PhD thesis. University of California, Berkeley, 1968.

[5] D. L. Donoho. "Breakdown properties of multivariate location estimators". PhD thesis. Harvard University, 1982.

Merge dataset

Notation:

- X is a dataset of size n, $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$
- Y is a dataset of size m, $Y = [y_1, \ldots, y_m] \in \mathbb{R}^{p \times m}$

The **merged** dataset, of size $n + m$ is written $X \cup Y$ and is the concatenation of X and Y:

$$X \cup Y = [x_1, \ldots, x_n, y_1, \ldots, y_m] \in \mathbb{R}^{(n+m) \times p}$$

Breakdown point: Donoho's definition

Definition: Breakdown point

For a dataset X of size n, the **breakdown point** of a statistic T is:

$$\varepsilon^* = \varepsilon^*(T, X) = \frac{m^*}{n + m^*}$$

where

$$m^* = \min \left\{ m : \sup_{\#Y = m} \| T(X \cup Y) - T(X) \| = +\infty \right\}$$

Rem: coined ε-contamination in Huber and Ronchetti (2009)

Rem: ε-replacement variant, *cf.* Maronna *et al.* (2006),Huber and Ronchetti (2009) consists in arbitrary corrupting some points from the dataset (not adding some more)

Remarks and first properties

$$\varepsilon^* = \frac{m^*}{n+m^*}, m^* = \min\left\{m : \sup_{\#Y=m} \|T(X \cup Y) - T(X)\| = +\infty\right\}$$

- $\varepsilon^* = \varepsilon^*(T, X)$: depends both on the statistic T and on the dataset X (but not so much on the later)
- m^*, ε^* do not depend on the norm chosen (proof: equivalence of norm in Euclidean spaces)
- $\forall \mu \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p\times p}$(nonsingular)$, \varepsilon^*(T, \Sigma X + \mu) = \varepsilon^*(T, X)$ when T is affine equivariant (blackboard)

Lower bound

Theorem

$$\varepsilon^*(T, X) \geq \frac{1}{n+1},$$

moreover this value is attained for the empirical mean

Lower bound

Theorem

$$\varepsilon^*(T, X) \geq \frac{1}{n+1},$$

moreover this value is attained for the empirical mean

Proof: Let $T(x_1, \ldots, x_n) = \overline{x}_n$. Hence,

Lower bound

Theorem

$$\varepsilon^*(T, X) \geq \frac{1}{n+1},$$

moreover this value is attained for the empirical mean

Proof: Let $T(x_1, \ldots, x_n) = \overline{x}_n$. Hence,

$$T(x_1, \ldots, x_n, y_1) - T(x_1, \ldots, x_n) = \frac{y_1 + n\overline{x}_n}{n+1} - \overline{x}_n$$

Lower bound

Theorem

$$\varepsilon^*(T, X) \geq \frac{1}{n+1},$$

moreover this value is attained for the empirical mean

Proof: Let $T(x_1, \ldots, x_n) = \overline{x}_n$. Hence,

$$\begin{aligned} T(x_1, \ldots, x_n, y_1) - T(x_1, \ldots, x_n) &= \frac{y_1 + n\overline{x}_n}{n+1} - \overline{x}_n \\ &= \frac{y_1}{n+1} + \frac{n}{n+1}\overline{x}_n - \overline{x}_n \end{aligned}$$

Lower bound

Theorem

$$\varepsilon^*(T, X) \geq \frac{1}{n+1},$$

moreover this value is attained for the empirical mean

Proof: Let $T(x_1, \dots, x_n) = \overline{x}_n$. Hence,

$$\begin{aligned}
T(x_1, \dots, x_n, y_1) - T(x_1, \dots, x_n) &= \frac{y_1 + n\overline{x}_n}{n+1} - \overline{x}_n \\
&= \frac{y_1}{n+1} + \frac{n}{n+1}\overline{x}_n - \overline{x}_n \\
&= \frac{y_1}{n+1} - \frac{1}{n+1}\overline{x}_n
\end{aligned}$$

Lower bound

Theorem

$$\varepsilon^*(T, X) \geq \frac{1}{n+1},$$

moreover this value is attained for the empirical mean

Proof: Let $T(x_1, \ldots, x_n) = \overline{x}_n$. Hence,

$$\begin{aligned} T(x_1, \ldots, x_n, y_1) - T(x_1, \ldots, x_n) &= \frac{y_1 + n\overline{x}_n}{n+1} - \overline{x}_n \\ &= \frac{y_1}{n+1} + \frac{n}{n+1}\overline{x}_n - \overline{x}_n \\ &= \frac{y_1}{n+1} - \frac{1}{n+1}\overline{x}_n \end{aligned}$$

So, $\|T(x_1, \ldots, x_n, y_1) - T(x_1, \ldots, x_n)\| \geq \dfrac{\|y_1\|}{n+1} - \dfrac{\|\overline{x}_n\|}{n+1}$

Lower bound

Theorem

$$\varepsilon^*(T, X) \geq \frac{1}{n+1},$$

moreover this value is attained for the empirical mean

Proof: Let $T(x_1, \ldots, x_n) = \overline{x}_n$. Hence,

$$\begin{aligned} T(x_1, \ldots, x_n, y_1) - T(x_1, \ldots, x_n) &= \frac{y_1 + n\overline{x}_n}{n+1} - \overline{x}_n \\ &= \frac{y_1}{n+1} + \frac{n}{n+1}\overline{x}_n - \overline{x}_n \\ &= \frac{y_1}{n+1} - \frac{1}{n+1}\overline{x}_n \end{aligned}$$

So, $\|T(x_1, \ldots, x_n, y_1) - T(x_1, \ldots, x_n)\| \geq \frac{\|y_1\|}{n+1} - \frac{\|\overline{x}_n\|}{n+1}$

Taking the sup over all $y_1 \in \mathbb{R}^p$ leads to the conclusion. □

Upper bound

Theorem

$$\varepsilon^*(T, X) \leq 1,$$

moreover this value is attained for constant estimators, say $T = 0$

Upper bound

Theorem

$$\varepsilon^*(T, X) \leq 1,$$

moreover this value is attained for constant estimators, say $T = 0$

Proof: Let $T(x_1, \ldots, x_n) = 0$.

Upper bound

Theorem

$$\varepsilon^*(T, X) \leq 1,$$

moreover this value is attained for constant estimators, say $T = 0$

Proof: Let $T(x_1, \ldots, x_n) = 0$.
Hence,

$$T(x_1, \ldots, x_n, y_1, \ldots, y_m) - T(x_1, \ldots, x_n) = 0, \forall m$$

Upper bound

Theorem

$$\varepsilon^*(T, X) \leq 1,$$

moreover this value is attained for constant estimators, say $T = 0$

Proof: Let $T(x_1, \ldots, x_n) = 0$.
Hence,

$$T(x_1, \ldots, x_n, y_1, \ldots, y_m) - T(x_1, \ldots, x_n) = 0, \forall m$$

So $m^* = +\infty$ and $\varepsilon^*(T, X) = 1$. □

Refined upper bound: translation invariance

Theorem

Whenever T is translation equivariant the following holds:

$$\varepsilon^*(T, X) \leq \frac{1}{2}$$

Interpretation 1: if one adds more contaminated points than the number of points already present, the estimator should break down

Interpretation 2: if more than half a dataset if phony, the "good" data must look like outliers contaminating the phony data!

Proof adapted from Donoho (1982)

Assume that the following holds:

$$\sup_{\#Y=n} \|T(X \cup Y) - T(X)\| = \infty \qquad (\star)$$

Then,

$$m^* := \min \left\{ m : \sup_{\#Y=m} \|T(X \cup Y) - T(X)\| = +\infty \right\} \leq n.$$

Next,

$$\varepsilon^* = \frac{m^*}{m^* + n} \leq \frac{n}{n+n} = \frac{1}{2}$$

holds true as $x \to \frac{x}{x+n}$ is a non-decreasing function.

Proof adapted from Donoho (1982) (bis)

ab absurdum: if $(\star)$ does not hold, there exists B such that

$$\sup_{\#Y=n} \|T(X \cup Y) - T(X)\| < B$$

1

2

3

4

Proof adapted from Donoho (1982) (bis)

ab absurdum: if $(\star)$ does not hold, there exists B such that

$$\sup_{\#Y=n} \|T(X \cup Y) - T(X)\| < B$$

Let $\mu \in \mathbb{R}^p$ such that $\|\mu\| = 3B$, then

$$\|T([X+\mu] \cup X) - T(X+\mu)\| \overset{1}{=} \|T(X \cup [X-\mu]) - T(X)\|$$

---

[1] T is translation equivariant

[2]

[3]

[4]

Proof adapted from Donoho (1982) (bis)

ab absurdum: if $(\star)$ does not hold, there exists B such that

$$\sup_{\#Y=n} \|T(X \cup Y) - T(X)\| < B$$

Let $\mu \in \mathbb{R}^p$ such that $\|\mu\| = 3B$, then

$$\|T([X+\mu] \cup X) - T(X+\mu)\| \overset{1}{=} \|T(X \cup [X-\mu]) - T(X)\| \overset{2}{\leq} B.$$

[1] T is translation equivariant

[2] use $\#[X-\mu] = n$ and *ab absurdum* hypothesis

[3]

[4]

Proof adapted from Donoho (1982) (bis)

ab absurdum: if $(\star)$ does not hold, there exists B such that

$$\sup_{\#Y=n} \|T(X \cup Y) - T(X)\| < B$$

Let $\mu \in \mathbb{R}^p$ such that $\|\mu\| = 3B$, then

$$\|T([X+\mu] \cup X) - T(X+\mu)\| \overset{1}{=} \|T(X \cup [X-\mu]) - T(X)\| \overset{2}{\leq} B.$$

Moreover,

$$\|T(X \cup [X+\mu]) - T(X)\| \overset{3}{\geq} \|T([X+\mu]) - T(X)\| \\ - \|T([X+\mu] \cup X) - T(X+\mu)\|$$

[1] T is translation equivariant

[2] use $\#[X-\mu] = n$ and *ab absurdum* hypothesis

[3] triangle inequality

[4]

Proof adapted from Donoho (1982) (bis)

ab absurdum: if ($\star$) does not hold, there exists B such that

$$\sup_{\#Y=n} \|T(X \cup Y) - T(X)\| < B$$

Let $\mu \in \mathbb{R}^p$ such that $\|\mu\| = 3B$, then

$$\|T([X+\mu] \cup X) - T(X+\mu)\| \overset{1}{=} \|T(X \cup [X-\mu]) - T(X)\| \overset{2}{\leq} B.$$

Moreover,

$$\begin{aligned} \|T(X \cup [X+\mu]) - T(X)\| &\overset{3}{\geq} \|T([X+\mu]) - T(X)\| \\ &\quad - \|T([X+\mu] \cup X) - T(X+\mu)\| \\ &\geq \|T([X+\mu]) - T(X)\| - B \\ &\overset{4}{=} \|\mu\| - B = 2B \end{aligned}$$

[1] T is translation equivariant

[2] use $\#[X-\mu] = n$ and *ab absurdum* hypothesis

[3] triangle inequality

[4]

Proof adapted from Donoho (1982) (bis)

ab absurdum: if $(\star)$ does not hold, there exists B such that

$$\sup_{\#Y=n} \|T(X \cup Y) - T(X)\| < B$$

Let $\mu \in \mathbb{R}^p$ such that $\|\mu\| = 3B$, then

$$\|T([X+\mu] \cup X) - T(X+\mu)\| \overset{1}{=} \|T(X \cup [X-\mu]) - T(X)\| \overset{2}{\leq} B.$$

Moreover,

$$\begin{aligned}
\|T(X \cup [X+\mu]) - T(X)\| &\overset{3}{\geq} \|T([X+\mu]) - T(X)\| \\
&\quad - \|T([X+\mu] \cup X) - T(X+\mu)\| \\
&\geq \|T([X+\mu]) - T(X)\| - B \\
&\overset{4}{=} \|\mu\| - B = 2B \\
&> B \quad \text{(contradiction)} \quad \square
\end{aligned}$$

[1] T is translation equivariant

[2] use $\#[X-\mu] = n$ and *ab absurdum* hypothesis

[3] triangle inequality

[4] T is translation equivariant

Median in dimension 1 ($p = 1$)

Theorem

The (1D) median $T(X) = \mathrm{Med}_n(X)$ achieves the best possible breakdown point value for a location parameter :

$$\varepsilon^*(T, X) = \frac{1}{2}$$

Reminder: the definition of “a” median is

$$\mathrm{Med}_n(X) \in \arg\min_{\delta \in \mathbb{R}} \sum_{i=1}^{n} |\delta - x_i|$$

Median properties

Property (I)

Any median $\mathrm{Med}_n(X)$ satisfies:

$$\#\{i \in [n] : x_i < \mathrm{Med}_n(X)\} \leq \#\{i \in [n] : x_i \geq \mathrm{Med}_n(X)\}$$
$$\#\{i \in [n] : x_i > \mathrm{Med}_n(X)\} \leq \#\{i \in [n] : x_i \leq \mathrm{Med}_n(X)\}$$

Proof: will be given in the "sub-gradient" lesson

Rem: beware that

$$\#\{i \in [n] : x_i \leq \mathrm{Med}_n(X)\} \neq \#\{i \in [n] : x_i \geq \mathrm{Med}_n(X)\}$$

Take for instance $X = (1, 2, 2, 3, 3)$, so that $\mathrm{Med}_n(X) = 2$ and

$$\#\{i \in [n] : x_i \leq \mathrm{Med}_n(X)\} = 3 < \#\{i \in [n] : x_i \geq \mathrm{Med}_n(X)\} = 4$$

Median properties (II)

Corrollary

Any median $\mathrm{Med}_n(X)$ satisfies:

$$\#\{i \in [n] : x_i < \mathrm{Med}_n(X)\} \leq \frac{n}{2}$$
$$\#\{i \in [n] : x_i > \mathrm{Med}_n(X)\} \leq \frac{n}{2}$$

Proof: simply remark the two following points

$$\#\{i \in [n] : x_i < \mathrm{Med}_n(X)\} + \#\{i \in [n] : x_i \geq \mathrm{Med}_n(X)\} = n$$
$$\#\{i \in [n] : x_i > \mathrm{Med}_n(X)\} + \#\{i \in [n] : x_i \leq \mathrm{Med}_n(X)\} = n$$

□

Proof (Median optimality)

Fact 1: $\mathrm{Med}_n(X)$ is translation equivariant so $\varepsilon^* \leq \frac{1}{2}$.

Proof (Median optimality)

Fact 1: $\mathrm{Med}_n(X)$ is translation equivariant so $\varepsilon^* \leq \frac{1}{2}$.

Proof: Let $\mu \in \mathbb{R}$ and $X + \mu = [x_1 + \mu, \dots, x_n + \mu]$. Then,

$$\mathrm{Med}_n(X + \mu) \in \arg\min_{\delta \in \mathbb{R}} \sum_{i=1}^{n} |\delta - (x_i + \mu)|$$

Noticing that for any function f:

$$\arg\min_{\nu \in \mathbb{R}} f(\nu) + \mu = \arg\min_{\delta \in \mathbb{R}} f(\delta - \mu)$$

we get that $\mathrm{Med}_n(X + \mu) = \mathrm{Med}_n(X) + \mu$ $\square$

Proof (Median optimality)

Fact 1: $\mathrm{Med}_n(X)$ is translation equivariant so $\varepsilon^* \leq \frac{1}{2}$.

Proof: Let $\mu \in \mathbb{R}$ and $X + \mu = [x_1 + \mu, \ldots, x_n + \mu]$. Then,

$$\mathrm{Med}_n(X + \mu) \in \arg\min_{\delta \in \mathbb{R}} \sum_{i=1}^{n} |\delta - (x_i + \mu)|$$

Noticing that for any function f:

$$\arg\min_{\nu \in \mathbb{R}} f(\nu) + \mu = \arg\min_{\delta \in \mathbb{R}} f(\delta - \mu)$$

we get that $\mathrm{Med}_n(X + \mu) = \mathrm{Med}_n(X) + \mu$ $\square$

Partial conclusion: we only need to show $\varepsilon^* \geq \frac{1}{2}$, *i.e.*, $m^* \geq n$

Proof (II)

<u>Fact 2</u>: To show that $m^* \geq n$, it is sufficient to have

$$\sup_{\#Y=n-1} |\operatorname{Med}_{2n-1}(X \cup Y) - \operatorname{Med}_n(X)| < \infty.$$

Proof: simply remind that

$$\varepsilon^* = \frac{m^*}{n+m^*}, m^* = \min\left\{m : \sup_{\#Y=m} \|T(X \cup Y) - T(X)\| = +\infty\right\}$$

$\square$

We will now prove that:

$$\sup_{\#Y=n-1} |\operatorname{Med}_{2n-1}(X \cup Y) - \operatorname{Med}_n(X)| \leq x_{(n)} - x_{(1)} < +\infty$$

where the dataset X has been ordered s.t. $x_{(1)} \leq \cdots \leq x_{(n)}$

Proof (III)

Fact 3:
Let Y be arbitrary s.t. $\#Y = n-1$, $Z := X \cup Y = [z_1, \ldots, z_{2n-1}]$ for any $t \in \mathbb{R}$,

$$\#\{i \in [2n-1] : z_i \geq t\} \geq n \Rightarrow \mathrm{Med}_{2n-1}(Z) \geq t$$
$$\#\{i \in [2n-1] : z_i \leq t\} \geq n \Rightarrow \mathrm{Med}_{2n-1}(Z) \leq t$$

Proof (ab absurdum): we show only the first point, the second is proved similarly. If $M < t$ then one has

$n \leq \#\{i \in [2n-1] : z_i \geq t\}$

1

2

Proof (III)

Fact 3:
Let Y be arbitrary s.t. $\#Y = n-1$, $Z := X \cup Y = [z_1, \dots, z_{2n-1}]$ for any $t \in \mathbb{R}$,

$$\#\{i \in [2n-1] : z_i \geq t\} \geq n \Rightarrow \mathrm{Med}_{2n-1}(Z) \geq t$$
$$\#\{i \in [2n-1] : z_i \leq t\} \geq n \Rightarrow \mathrm{Med}_{2n-1}(Z) \leq t$$

Proof (ab absurdum): we show only the first point, the second is proved similarly. If $M < t$ then one has

$$n \leq \#\{i \in [2n-1] : z_i \geq t\} \overset{1}{\leq} \#\{i \in [2n-1] : z_i > M\}$$

[1] use $M < t$

[2]

Proof (III)

Fact 3:
Let Y be arbitrary s.t. $\#Y = n-1$, $Z := X \cup Y = [z_1, \dots, z_{2n-1}]$ for any $t \in \mathbb{R}$,

$$\#\{i \in [2n-1] : z_i \geq t\} \geq n \Rightarrow \mathrm{Med}_{2n-1}(Z) \geq t$$
$$\#\{i \in [2n-1] : z_i \leq t\} \geq n \Rightarrow \mathrm{Med}_{2n-1}(Z) \leq t$$

Proof (ab absurdum): we show only the first point, the second is proved similarly. If $M < t$ then one has

$$n \leq \#\{i \in [2n-1] : z_i \geq t\} \overset{1}{\leq} \#\{i \in [2n-1] : z_i > M\} \overset{2}{\leq} \frac{2n-1}{2}$$

[1] use $M < t$

[2] apply last corollary to the z_i's

Proof (IV)

<u>Fact 4</u>: Let us order X so that $x_{(1)} \leq \cdots \leq x_{(n)}$, then

$$\mathrm{Med}_{2n-1}(Z) \in [x_{(1)}, x_{(n)}]$$

Proof (IV)

<u>Fact 4</u>: Let us order X so that $x_{(1)} \leq \cdots \leq x_{(n)}$, then

$$\mathrm{Med}_{2n-1}(Z) \in [x_{(1)}, x_{(n)}]$$

Proof: one can check that

$$\{x_{(1)}, \ldots, x_{(n)}\} \subset \{z_i : z_i \geq x_{(1)}\}$$

hence

$$\#\{i \in [2n-1] : z_i \geq x_{(1)}\} \geq n.$$

We can apply Fact 3 so that:

$$\mathrm{Med}_{2n-1}(X \cup Y) = \mathrm{Med}_{2n-1}(Z) \geq x_{(1)}$$

$$\mathrm{Med}_{2n-1}(X \cup Y) = \mathrm{Med}_{2n-1}(Z) \leq x_{(n)} \quad \square$$

Proof (IV)

Fact 4: Let us order X so that $x_{(1)} \leq \cdots \leq x_{(n)}$, then

$$\mathrm{Med}_{2n-1}(Z) \in [x_{(1)}, x_{(n)}]$$

Proof: one can check that

$$\{x_{(1)}, \ldots, x_{(n)}\} \subset \{z_i : z_i \geq x_{(1)}\}$$

hence

$$\#\{i \in [2n-1] : z_i \geq x_{(1)}\} \geq n.$$

We can apply Fact 3 so that:

$$\mathrm{Med}_{2n-1}(X \cup Y) = \mathrm{Med}_{2n-1}(Z) \geq x_{(1)}$$

$$\mathrm{Med}_{2n-1}(X \cup Y) = \mathrm{Med}_{2n-1}(Z) \leq x_{(n)} \quad \square$$

Finally,

$$\sup_{\#Y=n-1} |\mathrm{Med}_{2n-1}(X \cup Y) - \mathrm{Med}_n(X)| \leq x_{(n)} - x_{(1)} < +\infty$$

and this conclude the proof using Fact 2.

Geometric median

A (Euclidean) **geometric median** is defined by:

$$\mathrm{Med}_n(X) \in \operatorname*{arg\,min}_{\nu \in \mathbb{R}^p} \sum_{i=1}^{n} \|\nu - x_i\|_2$$

Geometric median

A (Euclidean) **geometric median** is defined by:

$$\mathrm{Med}_n(X) \in \operatorname*{arg\,min}_{\nu \in \mathbb{R}^p} \sum_{i=1}^{n} \|\nu - x_i\|_2$$

- Translation equivariant: $T(X+\mu) = T(X) + \mu$, $\forall \mu \in \mathbb{R}^p$

  Hint: use $\operatorname*{arg\,min}_{\nu \in \mathbb{R}} f(\nu) = \operatorname*{arg\,min}_{\nu' \in \mathbb{R}} f(\nu' - \mu) - \mu$

Geometric median

A (Euclidean) **geometric median** is defined by:

$$\mathrm{Med}_n(X) \in \operatorname*{arg\,min}_{\nu \in \mathbb{R}^p} \sum_{i=1}^{n} \|\nu - x_i\|_2$$

- Translation equivariant: $T(X+\mu) = T(X) + \mu$, $\forall \mu \in \mathbb{R}^p$

  Hint: use $\operatorname*{arg\,min}_{\nu \in \mathbb{R}} f(\nu) = \operatorname*{arg\,min}_{\nu' \in \mathbb{R}} f(\nu' - \mu) - \mu$

- Orthogonally equivariant: $T(\Sigma X) = \Sigma T(X)$ for any matrix $\Sigma \in \mathbb{R}^{p \times p}$ such that $\Sigma^\top \Sigma = \mathrm{Id}_p$,

  Hint: use $\operatorname*{arg\,min}_{\nu \in \mathbb{R}} f(\nu) = \Sigma^{-1} \operatorname*{arg\,min}_{\nu' \in \mathbb{R}} f(\Sigma^{-1}\nu')$

Geometric median

A (Euclidean) **geometric median** is defined by:

$$\mathrm{Med}_n(X) \in \arg\min_{\nu \in \mathbb{R}^p} \sum_{i=1}^{n} \|\nu - x_i\|_2$$

- Translation equivariant: $T(X+\mu) = T(X) + \mu$, $\forall \mu \in \mathbb{R}^p$

  Hint: use $\quad \arg\min_{\nu \in \mathbb{R}} f(\nu) = \arg\min_{\nu' \in \mathbb{R}} f(\nu' - \mu) - \mu$

- Orthogonally equivariant: $T(\Sigma X) = \Sigma T(X)$ for any matrix $\Sigma \in \mathbb{R}^{p \times p}$ such that $\Sigma^\top \Sigma = \mathrm{Id}_p$,

  Hint: use $\quad \arg\min_{\nu \in \mathbb{R}} f(\nu) = \Sigma^{-1} \arg\min_{\nu' \in \mathbb{R}} f(\Sigma^{-1}\nu')$

- **But not affine equivariant** (except in 1D):

$$\sum_{i=1}^{n} \|\nu - \Sigma x_i\|_2 = \sum_{i=1}^{n} \sqrt{(\Sigma^{-1}\nu - x_i)^\top \Sigma^\top \Sigma (\Sigma^{-1}\nu - x_i)}$$

$$\mathrm{Med}_n(\Sigma X) = \Sigma \arg\min_{\nu' \in \mathbb{R}^p} \sum_{i=1}^{n} \sqrt{(\nu' - x_i)^\top \Sigma^\top \Sigma (\nu' - x_i)}$$

Breakdown Point of Geometric Median[6]

Theorem

The geometric median $T(X) = \mathrm{Med}_n(X)$ achieves the best possible breakdown point value for a translation equivariant:

$$\varepsilon^*(T, X) = \frac{1}{2}$$

Proof: By translation equivariance, we can assume that $\mathrm{Med}_n(X) = 0$, and writing $Z = [z_1, \ldots, z_{2n-1}] = X \cup Y$ for $\#Y = n - 1$, it is then sufficient to show:

$$\sup_{\#Y = n-1} |\,\mathrm{Med}_{2n-1}(Z)| < \infty.$$

[6] H. P. Lopuhaä and P. J. Rousseeuw. "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices". In: *Ann. Statist.* 19.1 (1991), pp. 229–248.

Proof (I)

Let $M = \max_{i=1,\dots,n} \|x_i\|_2$ and $B(0, 2M)$ be the (Euclidean) ball of center 0 and radius M.

1

Proof (I)

Let $M = \max_{i=1,\ldots,n} \|x_i\|_2$ and $B(0, 2M)$ be the (Euclidean) ball of center 0 and radius M.
Let d be the distance between $\mathrm{Med}_{2n-1}(Z)$ and $B(0, 2M)$, *i.e.*,

$$d := \min_{y \in B(0,2M)} \|y - \mathrm{Med}_{2n-1}(Z)\| = \|y^* - \mathrm{Med}_{2n-1}(Z)\|$$

for some $y^* \in B(0, 2M)$.

1

Proof (I)

Let $M = \max_{i=1,\dots,n} \|x_i\|_2$ and $B(0, 2M)$ be the (Euclidean) ball of center 0 and radius M.
Let d be the distance between $\mathrm{Med}_{2n-1}(Z)$ and $B(0, 2M)$,*i.e.*,

$$d := \min_{y \in B(0,2M)} \|y - \mathrm{Med}_{2n-1}(Z)\| = \|y^* - \mathrm{Med}_{2n-1}(Z)\|$$

for some $y^* \in B(0, 2M)$. Hence, $d \overset{1}{\geq} \|\mathrm{Med}_{2n-1}(Z)\| - \|y^*\|$, so:

$$\|\mathrm{Med}_{2n-1}(Z)\| \leq \|y^*\| + d \leq 2M + d. \qquad (\star)$$

[1] triangle inequality

Proof (I)

Let $M = \max_{i=1,\dots,n} \|x_i\|_2$ and $B(0, 2M)$ be the (Euclidean) ball of center 0 and radius M.
Let d be the distance between $\mathrm{Med}_{2n-1}(Z)$ and $B(0, 2M)$,*i.e.*,

$$d := \min_{y \in B(0,2M)} \|y - \mathrm{Med}_{2n-1}(Z)\| = \|y^* - \mathrm{Med}_{2n-1}(Z)\|$$

for some $y^* \in B(0, 2M)$. Hence, $d \overset{1}{\geq} \|\mathrm{Med}_{2n-1}(Z)\| - \|y^*\|$, so:

$$\|\mathrm{Med}_{2n-1}(Z)\| \leq \|y^*\| + d \leq 2M + d. \qquad (\star)$$

Now, $\forall i \in [n-1]$, $\|y_i - \mathrm{Med}_{2n-1}(Z)\| \overset{1}{\geq} \|y_i\| - \|\mathrm{Med}_{2n-1}(Z)\|$, so

$$\|y_i - \mathrm{Med}_{2n-1}(Z)\| \geq \|y_i\| - 2M - d \qquad (\star\star)$$

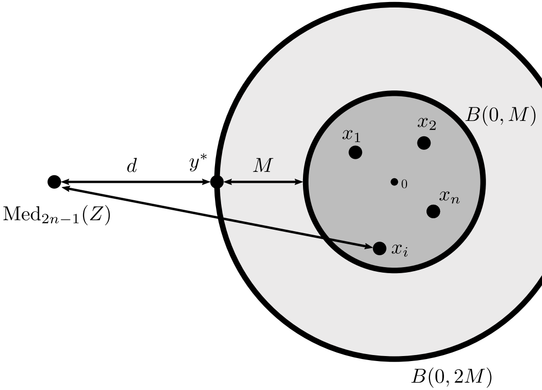
[1] triangle inequality

Proof (II)

Remind that $M = \max_{i=1,\dots,n} \|x_i\|$, so $\forall i \in [n], x_i \in B(0, M)$. Hence, using the figure one can claim that

$$\forall i \in [n], \quad \|x_i - \mathrm{Med}_{2n-1}(Z)\| \geq M + d$$

$$\forall i \in [n], \quad \|x_i - \mathrm{Med}_{2n-1}(Z)\| \geq \|x_i\| + d \qquad (\star\star\star)$$



Proof (III)

$$\forall i \in [n-1], \quad \|y_i - \mathrm{Med}_{2n-1}(Z)\| \geq \|y_i\| - 2M - d \qquad (\star\star)$$
$$\forall i \in [n], \quad \|x_i - \mathrm{Med}_{2n-1}(Z)\| \geq \|x_i\| + d \qquad (\star\star\star)$$

Summing $(\star\star)$ and $(\star\star\star)$

$$\sum_{i=1}^{2n-1} \|z_i - \mathrm{Med}_{2n-1}(Z)\| \geq \sum_{i=1}^{2n-1} \|z_i\| - (2M+d)(n-1) + nd$$
$$= \sum_{i=1}^{2n-1} \|z_i\| + d - 2M(n-1)$$

Now if $d - 2M(n-1) > 0$ then 0 would achieve a smaller objective value than $\mathrm{Med}_{2n-1}(Z)$, leading to a contradiction. Hence, $d \leq 2M(n-1)$ and reminding $(\star)$:

$$\|\mathrm{Med}_{2n-1}(Z)\| \overset{(\star)}{\leq} 2M + d \leq 2nM < \infty$$

$\square$

References I

- Donoho, D. L. "Breakdown properties of multivariate location estimators". PhD thesis. Harvard University, 1982.
- Donoho, D. L. and M. Gasko. "Breakdown properties of location estimates based on halfspace depth and projected outlyingness". In: *Ann. Statist.* 20.4 (1992), pp. 1803–1827.
- Hampel, F. R. "Contributions to the theory of robust estimation". PhD thesis. University of California, Berkeley, 1968.
- Huber, P. J. and E. M. Ronchetti. *Robust statistics*. Second. Wiley series in probability and statistics. Wiley, 2009.
- Lopuhaä, H. P. and P. J. Rousseeuw. "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices". In: *Ann. Statist.* 19.1 (1991), pp. 229–248.
- Maronna, R. A., R. D. Martin, and V. J. Yohai. *Robust statistics: Theory and methods.* Chichester: John Wiley & Sons, 2006.