Introduction to robust statistics

Pavlo Mozharovskyi*

*LTCI, Telecom Paris, Institut Polytechnique de Paris

Tail events analysis:
Robustness, outliers and models for extreme values

Palaiseau, February 10, 2020

Outline of the course

Format: 6 $\times$ 3.5 hours + exam

- Class 1: Introduction to robust statistics
- Class 2: Lab session I
- Class 3: Data depth
- Weeks 4: Extreme value statistics
- Week 5 : Multi-dimensional setting
- Week 6: Lab session II

Programming language: R

Grading: Exam

Today

Contents

Observations in the tail of a distribution

Given observations in the tail of a distribution,
there are two statistical points of view:

- The observations are contaminating the data and should be ignored: outliers.

Robust statistics

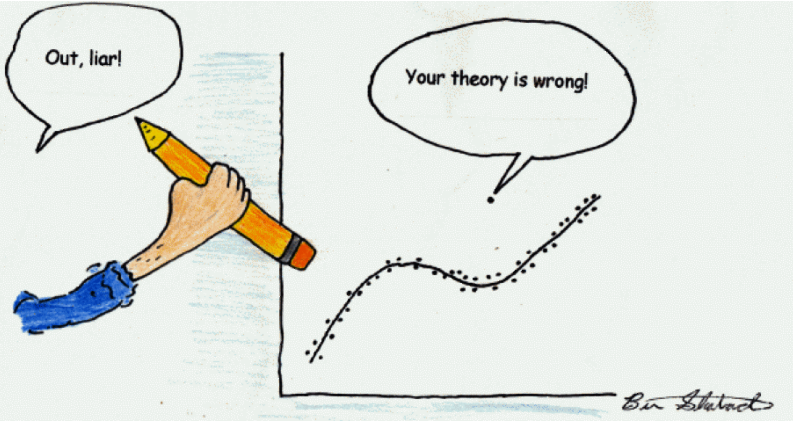
- The observations are (even more) of interest (than the "normal" data itself) and thus their modeling should be studied in detail: extreme values.

Extreme value theory

What is an outlier?

Definition

An outlier is an observation that deviates from the (model fit suggested by the) majority of the observations.



What is robust statistics?

- Often, real data contain outliers. Results of most statistical methods are (highly) influenced by these outliers.

- Robust statistical methods try to fit the model imposed by the majority of the data. They aim to find a *robust* fit, which is possibly close to the fit one would have found without outliers.

- This further allows outlier detection: flagging those observations deviating from the robust fit.

Assumptions

- One often assumes that the majority of observations follow a specific (parametric) model and one is interested in estimating parameters of this model.

$$
\begin{aligned}
E.g.: x_i &\sim \mathcal{N}(\mu, \sigma^2) \\
\boldsymbol{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \,\text{with}\, \varepsilon_i \sim \mathcal{N}(0, \sigma^2)\,.
\end{aligned}
$$

- Further, one assumes that some observations might not follow this specified model.
- !!! But, the model of outlier(s) generating process(es) is unknown.
- !!! Also, the portion of outliers is unknown.
- An example is the Huber contamination model:

$$X \sim (1 - p_{outliers}) F_{normal} + p_{outliers} F_{outliers}\;,\text{where}$$

F_{normal} is the probability distribution of "normal" observations,
$F_{outliers}$ is the probability distribution of the outlying observations,
$p_{outliers}$ is the prior probability of outliers.

A simple example

Consider the 10 most recent observations from the data set on *Nuclear power plant accidents* with available and positive accident cost. The logarithm of the total accident cost is presented in the table below:

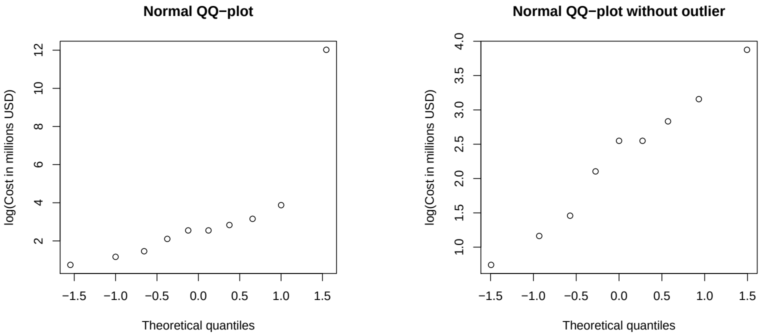
Date	Power plant	log(Cost)
2011-03-11	Fukushima Prefecture, Japan	12.02
2011-08-23	Mineral, Virginia, US	3.875
2011-09-12	Marcoule, France	2.549
2012-01-30	Rock River, Illinois, US	0.742
2012-03-12	Wanli, Taiwan	1.163
2012-04-05	Dieppe, France	2.549
2013-06-21	Wanli, Taiwan	1.459
2013-07-15	Shimen, Taiwan	3.157
2014-02-14	Waste Isolation Pilot Plant, New Mexico, US	2.104
2014-08-11	Lancashire, UK	2.833

Assume the Gaussian model for "normal" data:

$$x_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for} \quad i = 1, \dots, 10.$$

A simple example: QQ-plot

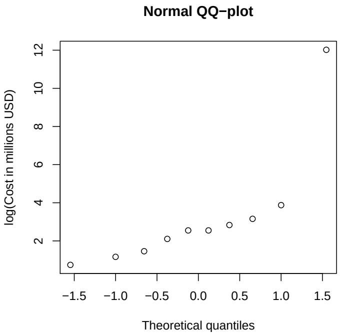
A normal QQ-plot is a plot of the observations versus theoretical quantiles of the Gaussian distribution: ideal fit should give a straight line.

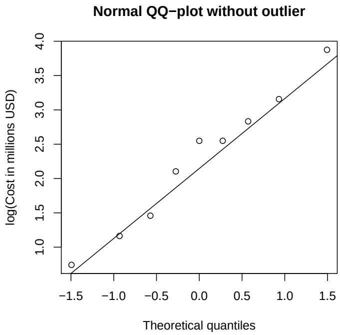


For the 9 observations (*i.e.* except for the Fukushima accident) the Gaussianity cannot be rejected.

A simple example: QQ-plot

A normal QQ-plot is a plot of the observations versus theoretical quantiles of the Gaussian distribution: ideal fit should give a straight line.





For the 9 observations (*i.e.* except for the Fukushima accident) the Gaussianity cannot be rejected.

Classical versus robust estimators: location

Classic estimator: arithmetic mean.

$$\hat{\mu} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \,.$$

Value for the given sample: $\bar{X}_n = 3.245\,.$

Robust estimator: sample median.

$$\hat{\mu} = \mathrm{med}(X_n) = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd}\,, \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even}\,. \end{cases}$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$ are the ordered observations.

Value for the given sample: $\mathrm{med}(X_n) = 2.549\,.$

Classical versus robust estimators: scale

Classic estimator: standard deviation.

$$\hat{\sigma} = \mathrm{sd}(X_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}_n)^2}\,.$$

Value for the given sample: $\mathrm{sd}(X_n) = 3.226$.

Robust estimator: interquantile range.

$$\hat{\sigma} = \mathrm{IQR}(X_n) = x_{\{0.75\}} - x_{\{0.25\}}\,,$$

where $x_{\{q\}}$ is the q-th empirical quantile for $q \in [0, 1]$.

Value for the given sample: $\mathrm{IQR}(X_n) = 1.456$.

Classical versus robust estimators: comparison

Compare the estimates excluding (only 9 “normal” observations) and including (all 10 observations) the Fukushima accident.

	9 “normal” observations	all 10 observations
$\bar{X}_n$	2.27	3.245
med(X_n)	2.549	2.549
sd(X_n)	1.005	3.226
IQR	1.375	1.456

- The classic estimators are highly influenced by the outlier.
- The robust estimators are less influenced by the outlier.
- The robust estimates computed from the 9 “normal” observations only are comparable with the estimates obtained using all 10 observations.

Classical versus robust estimators

- Robustness: Being less influenced by outliers.
- Efficiency: Being precise on uncontaminated data.

One requires from robust estimators being both:

robust and *efficient*.

Outlier detection

Usual rule: *an outlier has high z-score* (standardized residual).

Using classic estimates:

$$r_i = \frac{x_i - \bar{X}_n}{\mathrm{sd}(X_n)} = 2.72\,.$$

One flags an observation as outlier if $|r_i| > 3$.

For the Fukushima accidnet: $|r_1| = 2.72$; conclusion: ?

Using robust estimates:

$$r_i = \frac{x_i - \mathrm{med}(X_n)}{\mathrm{IQR}(X_n)}\,.$$

For the Fukushima accidnet: $|r_1| = 6.504$; conclusion: *an outlier*.

Contents

Contents

Breakdown value

Definition

Given an estimator T a data set X_n consisting of n observations. Let m be an integer such that:

- the estimator T stays in a fixed bounded set if $m - 1$ observations are replaced by *any* outliers;
- this does not hold anymore if m observations are replaced by *any* outliers.

The breakdown value of the estimator T at the data set X_n is $\frac{m}{n}$.

- Notation:
$$\varepsilon_n^*(T_n, X_n) = \frac{m}{n}.$$
- Typically, the breakdown value does not depend (much) on the data set.
- Often, it is a fixed constant as long as the (original) data set satisfies certain weak condition(s), *e.g.* the absence of ties.

Breakdown value: arithmetic mean

Example (Arithmetic mean)

Given:

- A univariate data set $X_n = \{x_1, ..., x_n\}$.
- The estimator $T(X_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- Replace one (arbitrary) observation from X_n by *any* value x^*, yielding a new data set X_n^*.
- If $x^* = +\infty$, then $T(X_n^*) = +\infty$ as well.
- Thus, the breakdown value of T_n being the arithmetic mean at X_n is:

$$\varepsilon_n^*(T, X_n) = \frac{1}{n} \cdot 1 = \frac{1}{n}\,.$$

- The limit — if $n \to \infty$ — of the finite sample breakdown value is called the asymptotic breakdown value:

$$\lim_{n\to\infty} \varepsilon_n^*(T, X_n) = \lim_{n\to\infty} \frac{1}{n} = 0\,.$$

Contents

Sensitivity curve

- Now, we study the behavior of the estimator when adding one observation to the sample.

Definition

Given an estimator T and a data set $X_{n-1} = \{x_1, ..., x_{n-1}\}$ consisting of $n-1$ observations. For $x \in \mathbb{R}$, let $X_n = \{x_1, ..., x_{n-1}, x\}$ be the completed data set. Then, the sensitivity curve is defined as:

$$SC(x, T, X_{n-1}) = \frac{T(X_n) - T(X_{n-1})}{\frac{1}{n}}.$$

Remarks:

- The sensitivity curve measures the effect of a single outlier on the estimator.
- The sensitivity curve depends strongly on the data set.

Sensitivity curve: arithmetic mean

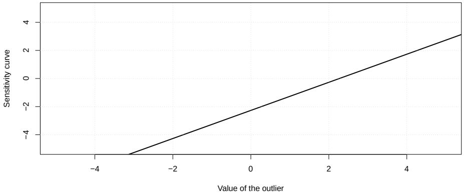
Example (Arithmetic mean)

Given:

- A univariate data set of "normal" observations X_9.
- The estimator $T(X_n)$.
- For the arithmetic mean $T(X_{n-1}) = \sum_{i=1}^{n-1} x_i$ (using notation from above) we obtain:

$$SC(x, T, X_{n-1}) = \frac{T(X_n) - T(X_{n-1})}{\frac{1}{n}} = \frac{\frac{1}{n}\left(\sum_{i=1}^{n-1} x_i + x\right) - \frac{1}{n-1}\sum_{i=1}^{n-1} x_i}{\frac{1}{n}}$$

$$= \frac{\frac{n-1}{n}\bar{X}_{n-1} + \frac{1}{n}x - \bar{X}_{n-1}}{\frac{1}{n}} = \frac{\frac{1}{n}x - \frac{1}{n}\bar{X}_{n-1}}{\frac{1}{n}} = x - \bar{X}_{n-1}\,.$$



Contents

Influence function

- The influence function can be seen as the *asymptotic version of* the *influence curve*.
- It is computed given an estimator T and a distribution F.
- The influence function measures how $T(F)$ changes with contamination added in one point x.

Definition

Given an estimator T and a distribution F. For $x \in \mathbb{R}$, let the contaminated distribution be defined as:

$$F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$$

for $\varepsilon > 0$, where Δ_x is the Dirac distribution at x.
Then, the influence function is defined as:

$$IF(x,T,F) = \lim_{\varepsilon \to 0} \frac{T(F_{\varepsilon,x}) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_{\varepsilon,x})\,|_{\varepsilon=0}\,.$$

Influence function: arithmetic mean

Example (Arithmetic mean)

Given:

- A distribution: $\mathcal{N}(0, \sigma^2)$.
- The estimator $T(X_n)$.

- For this purpose, the estimator should be written as a function of distribution F.
- For the sample mean we obtain $T(F) = \mathbb{E}_F[X]$.
- For the standard normal distribution we obtain:

$$
\begin{aligned}
IF(x, T, F) &= \frac{\partial}{\partial \varepsilon} \mathbb{E}_F[(1-\varepsilon)F + \varepsilon \Delta_x] \,|_{\varepsilon=0} \\
&= \frac{\partial}{\partial \varepsilon} (1-\varepsilon)\mathbb{E}_F[F] + \varepsilon \mathbb{E}_F[\Delta_x] \,|_{\varepsilon=0} \\
&= \mathbb{E}_F[\Delta_x] - \mathbb{E}_F[F] = x - \mathbb{E}_F[F] = x.
\end{aligned}
$$

- One prefers estimators with a bounded influence function.

Contents

Contents

Sample median

Definition (Sample median)

For a univariate data set $X_n = \{x_1, ..., x_n\}$ the sample median is defined as follows:

$$\hat{\mu} = \text{med}(X_n) = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd}\,, \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even}\,. \end{cases}$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$ are the ordered observations,
$\lfloor x \rfloor$ is the "floor" function $\lfloor x \rfloor = \max\{y \,:\, y \in \mathbb{Z},\, y \leq x\}$, and
$\lceil x \rceil$ is the "ceiling" function $\lceil x \rceil = \min\{y \,:\, y \in \mathbb{Z},\, y \geq x\}$.

For $\text{med}(X_n)$, let us study:

- (asymptotic) breakdown value,
- sensitivity curve (for the "normal" nuclear accident sample X_9),
- influence function (for the standard normal distribution Φ).

Sample median: breakdown value

Assume n is odd, then $T(X_n) = x_{(\frac{n+1}{2})}$.

- Replace $\frac{n-1}{2}$ observations from X_n by any values, which yields a data set X_n^*.
- Then, $T(X_n^*)$ belongs to the interval $[x_{(1)}, x_{(n)}]$, hence $T(X_n^*)$ is bounded.
- Replace $\frac{n+1}{2}$ observations by ∞.
- Then, $T(X_n^*){=}\infty$.

The (finite-sample) breakdown value ε_n^* of $T(X_n^*)$ is

$$\varepsilon_n^*(T, X_n) = \frac{1}{n}\left\lfloor \frac{n+1}{2} \right\rfloor \approx 0.5\,.$$

The asymptotic breakdown value is:

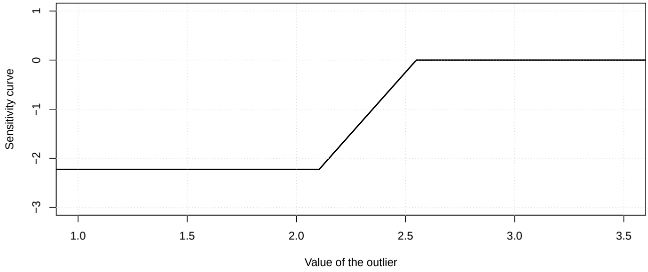
$$\lim_{n\to\infty} \varepsilon_n^*(T, X_n) = \frac{1}{n}\left\lfloor \frac{n+1}{2} \right\rfloor = 0.5\,(= 50\%)\,.$$

Sample median: sensitivity curve

For X_{n-1} $(= X_9)$,
assume $n-1$ is odd, then $T(X_{n-1}) = x_{(\frac{n}{2})}$.

$$SC(x, T, X_{n-1}) = \begin{cases} n\left(\frac{x_{(\frac{n}{2}-1)}+x_{(\frac{n}{2})}}{2} - x_{\frac{n}{2}}\right) & \text{if } x < x_{(\frac{n}{2}-1)}\,, \\ n\left(\frac{x+x_{(\frac{n}{2})}}{2} - x_{\frac{n}{2}}\right) & \text{if } x_{(\frac{n}{2}-1)} \leq x \leq x_{(\frac{n}{2}+1)}\,, \\ n\left(\frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2} - x_{\frac{n}{2}}\right) & \text{if } x > x_{(\frac{n}{2}+1)}\,. \end{cases}$$

For the nuclear accidents data we obtain:



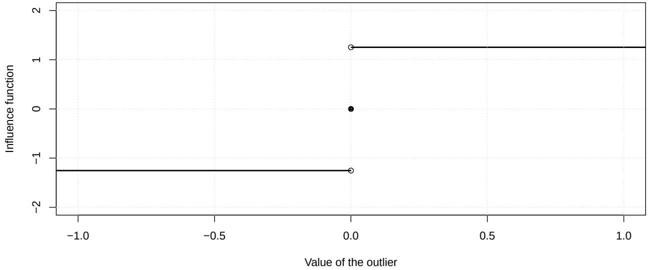
Sample median: influence function

For some F, assume that $f_X(x) > 0 \,\forall\, x \in \mathbb{R}$ and is continuous at x_q.

$$IF(x, T, F) = \begin{cases} \frac{q-1}{f_X(x_q)} & \text{if } x < x_q\,, \\ 0, & \text{if } x = x_q\,, \\ \frac{q}{f_X(x_q)} & \text{if } x > x_q\,, \end{cases}$$

where x_q is the qth quantile of F: $x_q = \inf\{x \,:\, F(x) \geq q\}$.

For the median $q = 0.5$, and with F being c.d.f. of $\mathcal{N}(0, \sigma^2)$, we obtain:



Contents

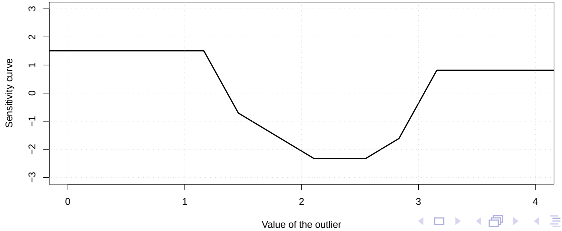
Interquantile range

Definition (Interquantile range)

For a univariate data set $X_n = \{x_1, \ldots, x_n\}$ the q-interquantile range is defined as follows:

$$\hat{\sigma} = \mathrm{IQR}_q(X_n) = x_{\{1-q\}} - x_{\{q\}} \,.$$

- A special case in common use is the 0.25-interquantile range, so that $\hat{\sigma}$ is the difference between the 0.75 and the 0.25 quantiles.
- Using similar considerations as those for the median, its asymptotic breakdown point is $\lim_{n\to\infty} \varepsilon_n^*(T, X_n) = 0.25$.
- For the "normal" part of the nuclear accidents data set the sensitivity curve looks as follows:



Interquantile range: influence function

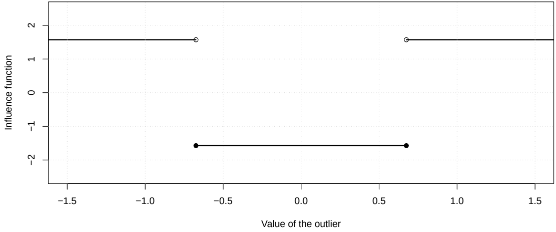
For F, assume that $f_X(x) > 0 \, \forall \, x \in \mathbb{R}$ and is continuous at x_q and x_{1-q}.

$$IF(x, T, F) = \begin{cases} \frac{1}{f_X(x_q)} - C & \text{if } x < x_q \,, \\ -C, & \text{if } x_q \leq x \leq x_{1-q} \,, \\ \frac{1}{f_X(x_q)} - C & \text{if } x > x_{1-q} \,, \end{cases}$$

where

$$C = q\left(\frac{1}{f_X(x_q)} + \frac{1}{f_X(x_{1-q})} \right) .$$

For the median $q = 0.25$, and with F being c.d.f. of $\mathcal{N}(0, \sigma^2)$, we obtain:

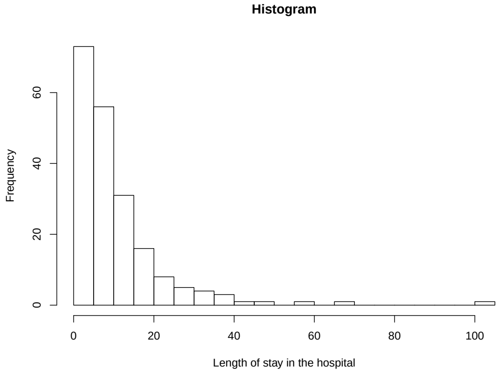


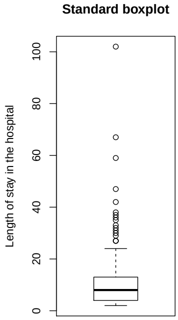
Contents

IQR and boxplot

- The boxplot is a useful tool for exploratory data analysis.
- Among others, it flags the *outliers* as the observations beyond the "whiskers".

Regard a data set of the length of stay (in days) for 201 patients at the University Hospital of Lausanne during the year 2000; see [RPM00] and R-package `robustbase` [MRC$^+$19, TF09] for a reference.





Medcouple

Definition (Medcouple)

For a univariate data set $X_n = \{x_1, ..., x_n\}$ the medcouple is defined as follows:

$$\hat{\gamma} = MC(X_n) = \text{med}\big(\{h(x_i, x_j) \,:\, x_i < Q_2 < x_j\}\big)\,,$$

where

$$h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}$$

and $Q_2 = \text{med}(X_n)$.

- *Medcouple* is sensitive to asymmetry, and thus is well suited for measuring deviations of the data from symmetry in practice.
- It has asymptotic breakdown value 0.25.

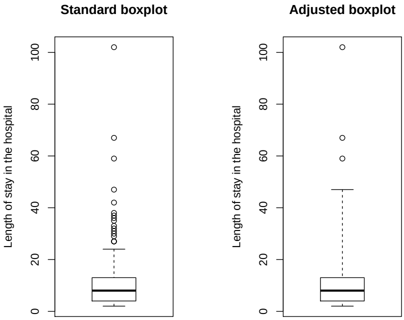
Adjusted boxplot

- Using *medcouple* we can define a boxplot adjusted to asymmetry.
- For this, one can define "whiskers" as:

$$[Q_1 - 1.5\, e^{-4\, MC(X_n)} IQR(X_n),\ Q_3 + 1.5\, e^{3\, MC(X_n)} IQR(X_n)]\,,$$

where $Q_1 = x_{0.25}$ and $Q_3 = x_{0.75}$ are 1st and 3rd quartiles of X_n.

For the length of stay data one can compare:



Contents

Multivariate data

- In most cases, data are multivariate, *i.e.* $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ where the observations $\boldsymbol{x}_i$ for $i = 1, ..., n$ are d-variate (column) vectors.
- Their coordinates can be summarized as a $n \times d$ matrix:

$$\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^\top = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix} .$$

- The model for the observations is the multivariate normal distribution:

$$X \sim \mathcal{N}_d(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X),$$

  where $\boldsymbol{\mu}_X \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_X$ is a positive semi-definite $d \times d$ matrix.
- More generally, one can assume that the data are generated from an elliptical distribution; the contours of an elliptical distribution are d-variate ellipsoids as well.

Affine equivariance

- Being unknown, in practice one evaluates $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ as *estimators* of location ($\hat{\boldsymbol{\mu}}$) and scatter ($\hat{\boldsymbol{\Sigma}}$).
- We often require from estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ affine equivariance.

Definition (Affine equivariance)

Location and scatter estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are affine equivariant if the satisfy:

$$\hat{\boldsymbol{\mu}}(\{\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}\}) = \boldsymbol{A}\hat{\boldsymbol{\mu}}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}) + \boldsymbol{b}\,,$$
$$\hat{\boldsymbol{\Sigma}}(\{\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}\}) = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\})\boldsymbol{A}^\top\,,$$

for any non-singular $d \times d$ matrix $\boldsymbol{A}$ and any vector $\boldsymbol{b} \in \mathbb{R}^d$.

- Affine invariance implies that the estimator "follows" any linear non-singular transformation/reparametrization of $\mathbb{R}^d$.
- The data can thus be translated, rotated or rescaled (*e.g.* due to the change of the measurement unit) without changing the *order statistics*, and thus without influencing the *outlier detection* diagnostics.

Breakdown value

- A *location estimator* $\hat{\boldsymbol{\mu}}$ "breaks down" if it can be contained beyond any bounded set.
- The breakdown value of a *scatter estimator* $\hat{\boldsymbol{\Sigma}}$ is defined as the smallest of the *explosion* and *implosion* breakdown values.
  - Explosion of a scatter estimator $\hat{\boldsymbol{\Sigma}}$ occurs when its largest eigenvalue becomes arbitrary large.
  - Implosion of a scatter estimator $\hat{\boldsymbol{\Sigma}}$ occurs when its smallest eigenvalue becomes arbitrary small.

Definition (General position)

A data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ is in general position if at most p observations from $\boldsymbol{X}$ lie in any affine subspace of dimension $p - 1$ for $p = 1, ..., d$.

Breakdown value

- Any affine equivariant location estimator $\hat{\boldsymbol{\mu}}$ satisfies:

$$\varepsilon_n^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X}) \leq \frac{1}{n} \left\lfloor \frac{n+1}{2} \right\rfloor .$$

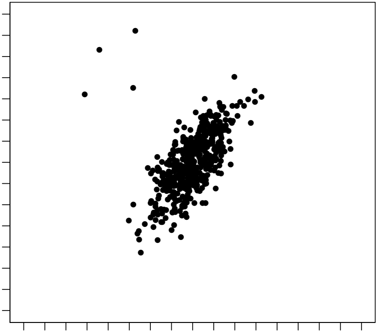
- If $\boldsymbol{X}$ is in *general position*, then any affine equivariant scatter estimator $\hat{\boldsymbol{\Sigma}}$ satisfies:

$$\varepsilon_n^*(\hat{\boldsymbol{\Sigma}}, \boldsymbol{X}) \leq \frac{1}{n} \left\lfloor \frac{n-d+1}{2} \right\rfloor .$$

Contents

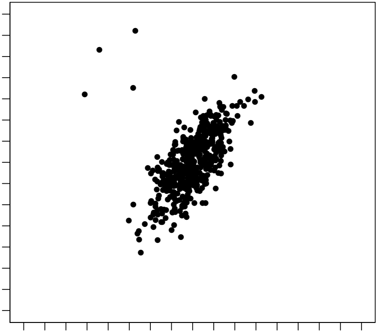
Detection of multivariate outliers

Regard two measurements during a test:



Detection of multivariate outliers

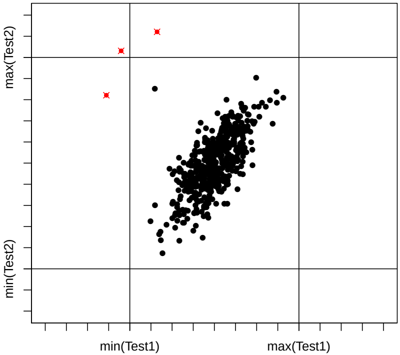
Regard two measurements during a test:



- Checking for minimum and maximum in each test result.

Detection of multivariate outliers

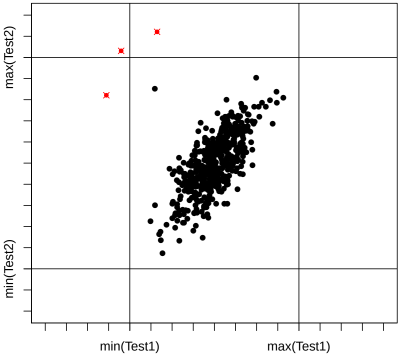
Regard two measurements during a test:



- Checking for minimum and maximum in each test result.

Detection of multivariate outliers

Regard two measurements during a test:

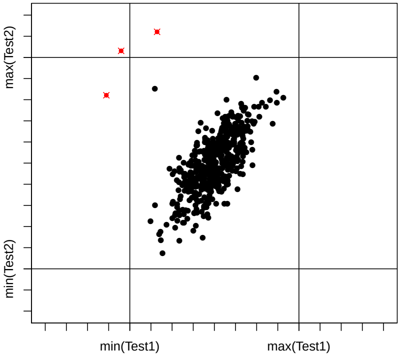


- Checking for minimum and maximum in each test result.
- Label observation $\boldsymbol{x}$ as outlier if:

$$\boldsymbol{x} \notin [\min(\text{Test1}), \max(\text{Test1})] \times [\min(\text{Test2}), \max(\text{Test2})] \,.$$

Detection of multivariate anomalies

Regard two measurements during a test:



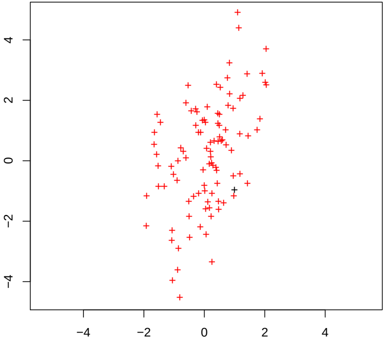
- Checking for minimum and maximum in each test result.
- Label observation $\boldsymbol{x}$ as outlier if:

$$\boldsymbol{x} \notin [\min(\text{Test1}), \max(\text{Test1})] \times [\min(\text{Test2}), \max(\text{Test2})] \,.$$

- !!! Not all anomalies can be detected.

Mahalanobis distance

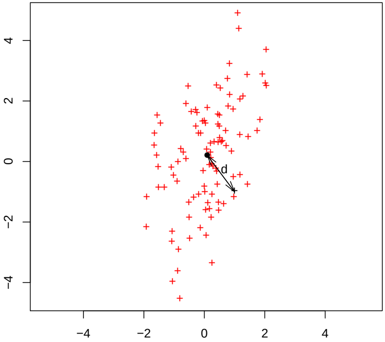
- Regard a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, , ..., \boldsymbol{x}_n\} \subset \mathbb{R}^d$ and a point $\boldsymbol{x} \in \mathbb{R}^d$.



- How central (or representative) is $\boldsymbol{x}$ with respect to $\boldsymbol{X}$?

Mahalanobis distance

- Regard a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^d$ and a point $\boldsymbol{x} \in \mathbb{R}^d$.



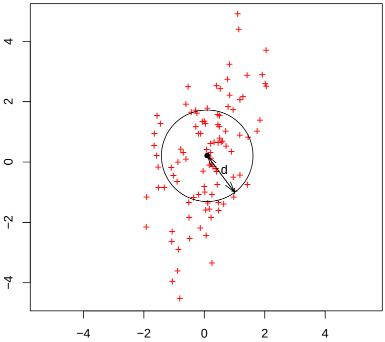
- Euclidean distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}_{\boldsymbol{X}}$:

$$d^2_{Eucl}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{X}}) = (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}) .$$

- Sample mean: $\boldsymbol{\mu}_{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$.

Mahalanobis distance

- Regard a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^d$ and a point $\boldsymbol{x} \in \mathbb{R}^d$.



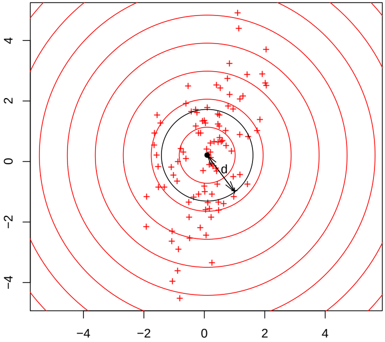
- Euclidean distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}_{\boldsymbol{X}}$:

$$d^2_{Eucl}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{X}}) = (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}) \,.$$

- Sample mean: $\boldsymbol{\mu}_{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$.

Mahalanobis distance

- Regard a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$ and a point $\boldsymbol{x} \in \mathbb{R}^d$.



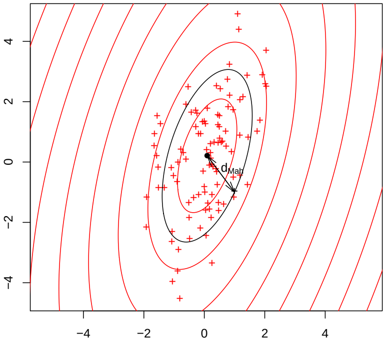
- Euclidean distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}_{\boldsymbol{X}}$:

$$d^2_{Eucl}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{X}}) = (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}).$$

- Sample mean: $\boldsymbol{\mu}_{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$.

Mahalanobis distance

- Regard a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^d$ and a point $\boldsymbol{x} \in \mathbb{R}^d$.

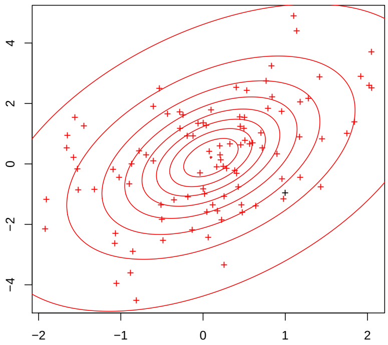


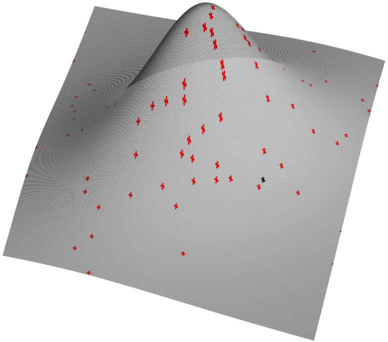
- Mahalanobis distance: $d^2_{Mah}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{X}}; \boldsymbol{\Sigma}_{\boldsymbol{X}}) = (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})$.

- Sample mean: $\boldsymbol{\mu}_{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i$.
- Sample covariance matrix: $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \frac{1}{n-1} \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}})^\top$.

Mahalanobis depth (Mahalanobis, 1936)

- Regard a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$ and a point $\boldsymbol{x} \in \mathbb{R}^d$.



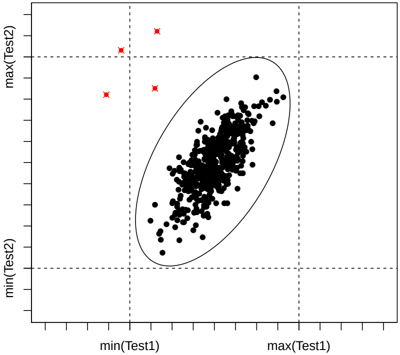


- Mahalanobis depth of $\boldsymbol{x}$ = a *centrality measure*:

$$D^{Mah(n)}(\boldsymbol{x}|\boldsymbol{X}) = \frac{1}{1 + d^2_{Mah}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{X}}; \boldsymbol{\Sigma}_{\boldsymbol{X}})} = \frac{1}{1 + (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})^\top \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}})}$$

Mahalanobis distance: detection of multivariate outliers

- Regard a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^d$ and a point $\boldsymbol{x} \in \mathbb{R}^d$.



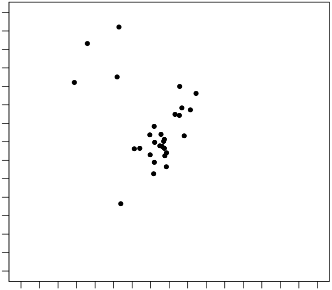
- Label $\boldsymbol{x}$ as outlier $d_{Mah}(\boldsymbol{x}|\boldsymbol{X}) > \max(d_{Mah})$.
- A reasonable (and often acceptable) choice is to take $\max(d_{Mah}$ to be a quantile of the χ^2 distribution, *e.g.* $\max(d_{Mah}) = \sqrt{\chi^2_{d,0.975}}$.
- This is called classical tolerance allipsoid.

Mahalanobis distance: robustness

- Since $\boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ are both affine equivariant estimators, the Mahalanobis distance is *affine invariant*, *i.e.*:

$$d_{Mah}(\boldsymbol{x}|\{\boldsymbol{A}\boldsymbol{x}_1+\boldsymbol{b}, ..., \boldsymbol{A}\boldsymbol{x}_n+\boldsymbol{b}\}) = d_{Mah}(\boldsymbol{x}|\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}) = d_{Mah}(\boldsymbol{x}|\boldsymbol{X}) .$$

- Nevertheless, Mahalanobis distance d_{Mah} is not robust, neither are estimators $\boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}$:
  - their breakdown value is 0;
  - their influence function is not bounded.
- With less available data, *e.g.* at the beginning of the production process, when abnormal behavior is in addition more likely:

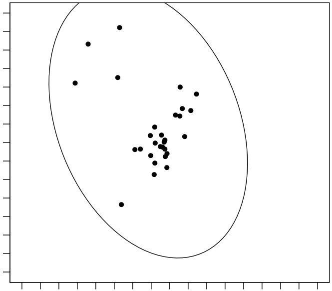


Mahalanobis distance: robustness

- Since $\boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ are both affine equivariant estimators, the Mahalanobis distance is *affine invariant*, *i.e.*:

  $$d_{Mah}(\boldsymbol{x}|\{\boldsymbol{A}\boldsymbol{x}_1+\boldsymbol{b}, ..., \boldsymbol{A}\boldsymbol{x}_n+\boldsymbol{b}\}) = d_{Mah}(\boldsymbol{x}|\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}) = d_{Mah}(\boldsymbol{x}|\boldsymbol{X}) .$$

- Nevertheless, Mahalanobis distance d_{Mah} is not robust, neither are estimators $\boldsymbol{\mu}_{\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}$:
  - their breakdown value is 0;
  - their influence function is not bounded.
- With less available data, *e.g.* at the beginning of the production process, when abnormal behavior is in addition more likely:



Contents

Stahel-Donoho estimator: idea behind

- The Stahel-Donoho estimator is the first affine-equivariante estimator of location and scatter with 50% asymptotic breakdown value [Sta81, Don82].
- It is based on the projection pursuit principle:
  "A multivariate outlier should be outlier in at least one direction, but not necessarily the direction(s) of the coordinate axes".

The algorithm of the Stahel-Donoho estimator is the following:

1. Data $\boldsymbol{X}$ are projected on a direction $\boldsymbol{u} \in \mathbb{S}^{d-1}$, with $\mathbb{S}^{d-1} = \{\boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}^d, \|\boldsymbol{y}\| = 1\}$ being the unit hypersphere.
2. For each data point, its robustly standardized distance to the median is computed of its projection $\boldsymbol{x}_i^\top \boldsymbol{u}$.
3. For each data point, the largest distance over all directions is retained. This distance is called outlyingness of $\boldsymbol{x}_i$.
4. The Stahel-Donoho estimator of location and scatter is the weighted mean and covariance matrix, where the weight function $W(t)$ is a strictly positive and weakly decreasing function of the outlyingness of $\boldsymbol{x}_i$.

Stahel-Donoho estimator: definition

Definition

For a multivariate data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}$, the Stahel-Donoho outlyingness of a point $\boldsymbol{x}_i$ is given by:

$$O_{SD}(\boldsymbol{x}_i) = \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \frac{|\boldsymbol{x}_i^\top \boldsymbol{u} - \text{med}(\boldsymbol{x}_1^\top \boldsymbol{u}, ..., \boldsymbol{x}_n^\top \boldsymbol{u})|}{\text{MAD}(\boldsymbol{x}_1^\top \boldsymbol{u}, ..., \boldsymbol{x}_n^\top \boldsymbol{u}))},$$

where

$$\text{MAD}(X_n) = \text{med}(|x_1 - \text{med}(X_n)|, ..., |x_n - \text{med}(X_n)|)$$

is the absolute median deviation *from the median* — a robust univariate measure of scale.

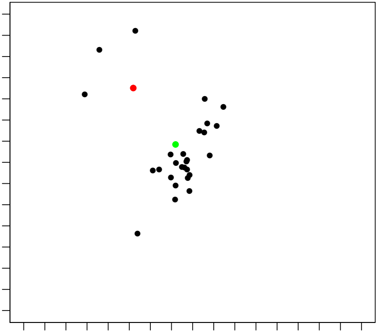
A typical weight function is

$$W(t) = \min\left(1, \frac{\chi^2_{d,0.95}}{t^2}\right).$$

Then, the estimator itself is defined as the weighted mean or weighted covariance matrix of the data with weights $w_i = W\big(O_{SD}(\boldsymbol{x}_i)\big)$.

Stahel-Donoho estimator: illustration

Regard again the two measurements during a test:



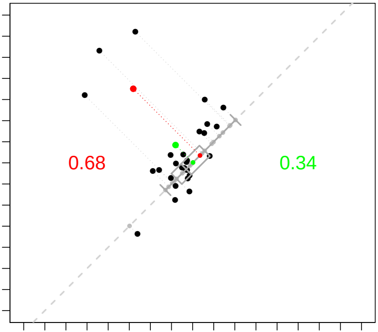
- *Stahel-Donoho outlyingness* of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top\boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}\,.$$

where 'med' and 'MAD' are median and median absolute deviation from it.

Stahel-Donoho estimator: illustration

Regard again the two measurements during a test:



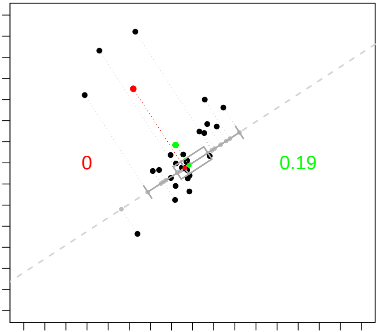
- *Stahel-Donoho outlyingness* of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top\boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}.$$

where 'med' and 'MAD' are median and median absolute deviation from it.

Stahel-Donoho estimator: illustration

Regard again the two measurements during a test:



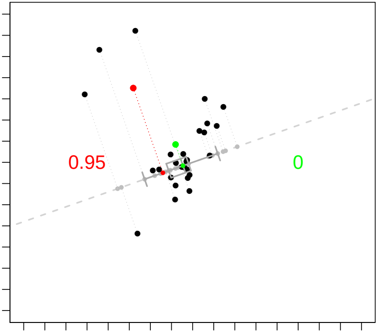
- *Stahel-Donoho outlyingness* of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top\boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}\,.$$

where 'med' and 'MAD' are median and median absolute deviation from it.

Stahel-Donoho estimator: illustration

Regard again the two measurements during a test:



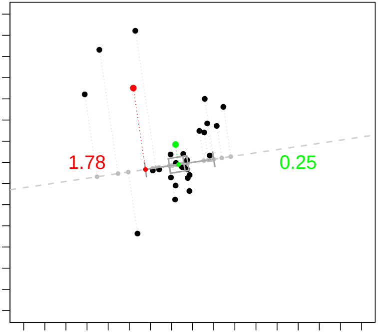
- *Stahel-Donoho outlyingness* of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}\,.$$

where 'med' and 'MAD' are median and median absolute deviation from it.

Stahel-Donoho estimator: illustration

Regard again the two measurements during a test:



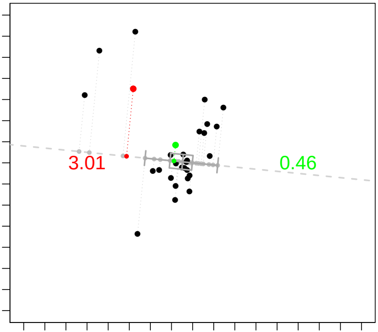
- *Stahel-Donoho outlyingness* of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top\boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}\,.$$

where 'med' and 'MAD' are median and median absolute deviation from it.

Stahel-Donoho estimator: illustration

Regard again the two measurements during a test:



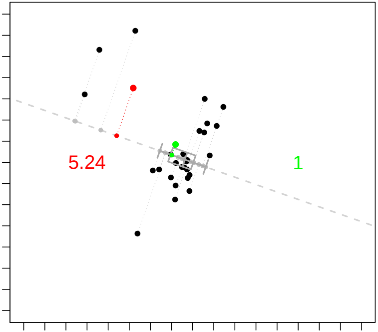
- *Stahel-Donoho outlyingness* of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}\,.$$

where 'med' and 'MAD' are median and median absolute deviation from it.

Stahel-Donoho estimator: illustration

Regard again the two measurements during a test:



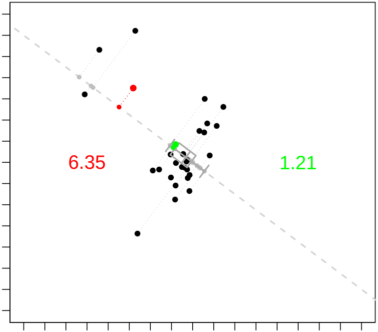
- *Stahel-Donoho outlyingness* of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top\boldsymbol{u} - \mathrm{med}(\boldsymbol{X}\boldsymbol{u})|}{\mathrm{MAD}(\boldsymbol{X}\boldsymbol{u})}\,.$$

  where 'med' and 'MAD' are median and median absolute deviation from it.

Stahel-Donoho estimator: illustration

Regard again the two measurements during a test:



- *Stahel-Donoho outlyingness* of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top\boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}.$$

  where 'med' and 'MAD' are median and median absolute deviation from it.

Contents

The minimum covariance determinant estimator

- The minimum covariance determinant (MCD) estimator [Rou84] is a widely used *high-breakdown* and *affine equivariant* estimator of location and scatter:

Definition (Minimum covariance determinant estimator)

For a multivariate data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \in \mathbb{R}^d$ and for fixed h, with $\frac{\lfloor n+d+1 \rfloor}{2} \leq h \leq n$, let

$$\mathcal{H}_0 \in \underset{\mathcal{H} \subset \{1,...,n\}, \# \mathcal{H} = h}{\arg\min} \det(\boldsymbol{\Sigma}_{\boldsymbol{X}_{\mathcal{H}}})$$

where $\boldsymbol{X}_{\mathcal{H}}$ is the subset of observations from $\boldsymbol{X}$ whose indices are in $\mathcal{H}$. The minimum covariance determinant estimator is then defined as follows:

- $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_{\boldsymbol{X}_{\mathcal{H}_0}}$, *i.e.* it is the mean of the h observations for which the *determinant of the covariance matrix is minimal*;
- $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\boldsymbol{X}_{\mathcal{H}_0}}$, *i.e.* it is the covariance matrix of the h observations for which the *determinant of the covariance matrix is minimal* (multiplied by the consistency factor).

Robustness of the MCD estimator

Properties of the MCD estimator:

- The influence function of MCD is bounded.
- The value h determines the breakdown value.
- For samples in general position

$$\varepsilon_n^* = \min\left(\frac{n-h+1}{n}, \frac{h-d}{n}\right).$$

- The maximal breakdown value is achieved by taking

$$h = \frac{\lfloor n+d+1 \rfloor}{2}.$$

- Usually one speaks about the robustness parameter of the MCD estimator $\alpha = \frac{h}{n} \in [0, 0.5]$.
- Typical choices of $\alpha = 0.5$ or $\alpha = 0.75$, which yields a breakdown value of 50% and 25% respectively.

Computation of the MCD estimator

Exact algorithm:

- Consider all possible $\boldsymbol{X}_{\mathcal{H}}$ with $\mathcal{H} \subset \{1, ..., n\}, \#\mathcal{H} = h$.
- For each of them, compute the mean and the covariance matrix.
- Retain the subset and the values for the mean and the (consistency corrected) covariance matrix with the smallest value of the covariance determinant.

! Infeasible for large n and even moderate d ...

Approximate algorithm:

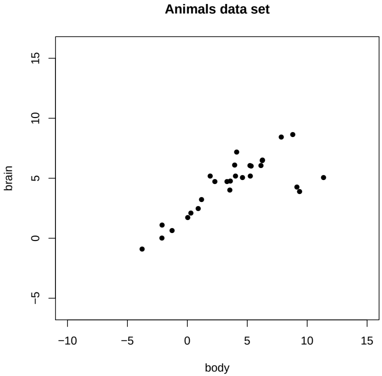
- Consider only a selected set of subsets of cardinality h of $\boldsymbol{X}$, starting from random subsets of size $d + 1$.
- The most used algorithm is FAST-MCD by [RD99].

The FAST-MCD algorithm

1. For $m = 1$ to 500:
   1.1 From $\{1, ..., n\}$, draw a random subset $\mathcal{H}_m$ of size $d+1$ and compute $\boldsymbol{\mu}_{\mathbf{X}_{\mathcal{H}_m}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{H}_m}}$.
   1.2 Apply a C-step:
      1.2.1 For $i = 1, ...n$, compute robust Mahalanobis distances based on $\boldsymbol{\mu}_{\mathbf{X}_{\mathcal{H}_m}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{H}_m}}$:
      $$rd_{Mah}(\mathbf{x}_i, \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{H}_m}}; \boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{H}_m}}) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{H}_m}})^\top \boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{H}_m}}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{H}_m}})}.$$
      1.2.2 Denote $\tilde{\mathcal{H}}$ the subset of $\{1, ..., n\}$ with the h smallest $rd_{Mah}(x_i, \boldsymbol{\mu}_{\mathbf{X}_{\mathcal{H}_m}}; \boldsymbol{\Sigma}_{\mathbf{X}_{\mathcal{H}_m}})$s.
      1.2.3 Compute $\boldsymbol{\mu}_{\mathbf{X}_{\tilde{\mathcal{H}}_m}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}_{\tilde{\mathcal{H}}_m}}$.
   1.3 Apply a second C-step.
2. Retain the 10 subsets with the smallest covariance determinant.
3. Apply C-step on these subsets until convergence.
4. Retain the subset with the smallest covariance determinant.
5. Return the average and the (consistency corrected) covariance matrix for the retained subset.

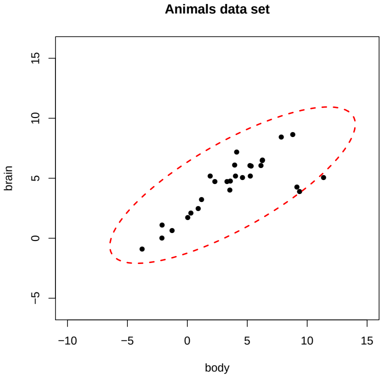
MCD estimator: Animals example

- Regard a data set consisting of the pairs of logarithms of the weight of the brain and of teh body for 28 animal species.



MCD estimator: Animals example

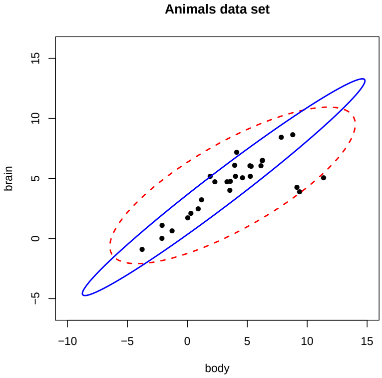
- Regard a data set consisting of the pairs of logarithms of the weight of the brain and of teh body for 28 animal species.



- Tolerance ellipsoid using moment estimates.

MCD estimator: Animals example

- Regard a data set consisting of the pairs of logarithms of the weight of the brain and of teh body for 28 animal species.



- Tolerance ellipsoid using moment estimates.
- Tolerance ellipsoid using MCD estimates.

Contents

Classical PCA

- Consider a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$.
- We assume that the variables are continuous.
- The main objective of principal component analysis (PCA) is to reduce the dimension of the data set without losing too much information.
- One looks for a k-dimensional subspace of $\mathbb{R}^d$ (with $k \ll \min\{n, d\}$) such that the projection of the data on this subspace contains most of the information of the original d-dimensional data.
- We thus search for a center $\boldsymbol{\mu}$ and a loading matrix $\boldsymbol{P}_{d,k}$ (of size $d \times k$) such that the k-dimensional scores t_i

$$t_i = \boldsymbol{P}_{d,k}^\top (\boldsymbol{x}_i - \boldsymbol{\mu})$$

are the most informative.

Classical PCA

- Classical principal component analysis (classical PCA, or CPCA) seeks the directions of maximum variability of the data.
- In particular, it computes the loading matrix

$$\boldsymbol{P}_{d,k} = [\boldsymbol{p}_1, \dots, \boldsymbol{p}_k]\,,$$

- where the first column is chosen as

$$\boldsymbol{p}_1 = \underset{\|\boldsymbol{p}\|=1}{\arg\max}\, \mathrm{var}\{\boldsymbol{p}^\top(\boldsymbol{x}_1 - \boldsymbol{\mu}), \boldsymbol{p}^\top(\boldsymbol{x}_2 - \boldsymbol{\mu}), \dots, \boldsymbol{p}^\top(\boldsymbol{x}_n - \boldsymbol{\mu})\}\,,$$

- and all the following columns are chosen sequentially by

$$\boldsymbol{p}_{j+1} = \underset{\|\boldsymbol{p}\|=1, \boldsymbol{p}\perp\boldsymbol{p}_1, \dots, \boldsymbol{p}\perp\boldsymbol{p}_j}{\arg\max}\, \mathrm{var}\{\boldsymbol{p}^\top(\boldsymbol{x}_1 - \boldsymbol{\mu}), \boldsymbol{p}^\top(\boldsymbol{x}_2 - \boldsymbol{\mu}), \dots, \boldsymbol{p}^\top(\boldsymbol{x}_n - \boldsymbol{\mu})\}\,.$$

Classical PCA

- The solution of this maximization problem yields the loading matrix as the matrix containing the k dominant eigenvectors of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ of the data points.
- In particular, the spectral decomposition of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ yields

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top$$

- with $\boldsymbol{P}$ the $d \times d$ orthogonal matrix containing all d eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ and $\boldsymbol{\Lambda}$ the diagonal matrix with the d eigenvalues $l_1, \ldots, l_d$ in decreasing order.
- The classical PCA loading matrix is the matrix $\boldsymbol{P}_{d,k}$ which contains the first k columns of $\boldsymbol{P}$.
- The eigenvalues l_j equal

$$l_j = \text{var}\{\boldsymbol{p}_j^\top(\boldsymbol{x}_1 - \boldsymbol{\mu}), \boldsymbol{p}_j^\top(\boldsymbol{x}_2 - \boldsymbol{\mu}), \ldots, \boldsymbol{p}_j^\top(\boldsymbol{x}_n - \boldsymbol{\mu})\},$$

Robust PCA based on a robust covariance estimator

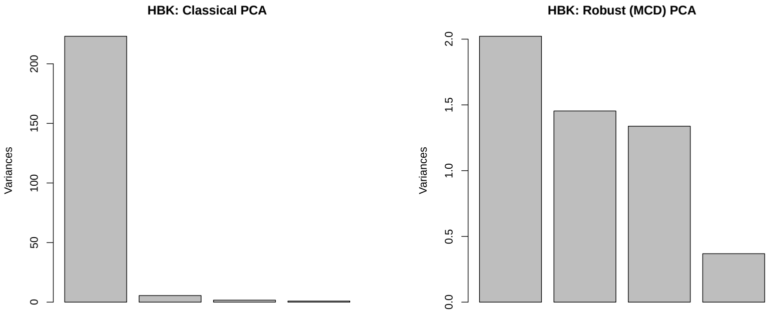
General idea:

- Replace the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ of $\boldsymbol{X}$ by a robust covariance estimate, such as, *e.g.*, MCD. Let us denote it $\boldsymbol{\Sigma}_{\boldsymbol{X},MCD}$.
- The robust center corresponds to the robust location estimate associated with $\boldsymbol{\Sigma}_{\boldsymbol{X},MCD}$.
- The k robust eigenvalues then correspond to the k largest eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{X},MCD}$.
- Take the k corresponding eigenvectors.

This approach can only be used when n is sufficiently larger than d (at least $n > 2d$).

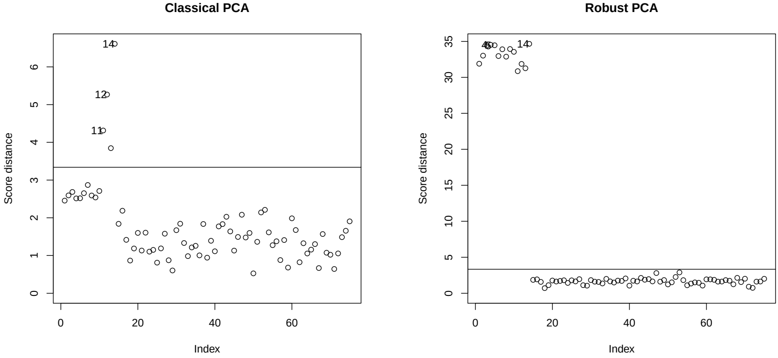
Robust covariance-based PCA: example

- Hawkins-Bradu-Kass data set ($n = 75$, $d = 4$) [HBK84].
- This is an artificial data set with two groups of outliers: observations $1 - 10$ and $11 - 14$.
- We apply classical PCA and robust PCA based on the MCD estimator with $\alpha = 0.5$ (breakdown value).
- This yields the following eigenvalues.



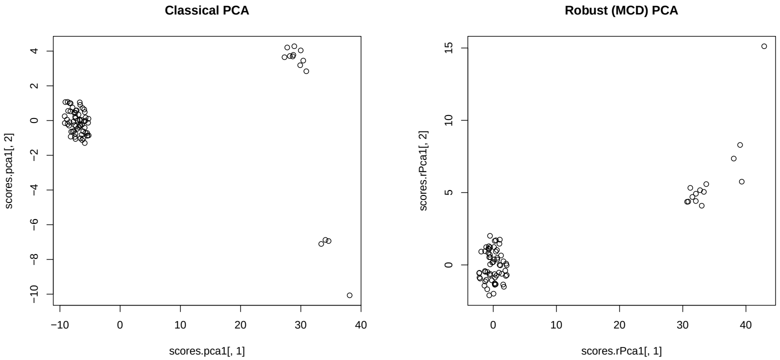
Robust covariance-based PCA: example

- Hawkins-Bradu-Kass data set ($n = 75$, $d = 4$) [HBK84].
- This is an artificial data set with two groups of outliers: observations $1 - 10$ and $11 - 14$.
- We apply classical PCA and robust PCA based on the MCD estimator with $\alpha = 0.5$ (breakdown value).
- This yields the following score distances.



Robust covariance-based PCA: example

- Hawkins-Bradu-Kass data set ($n = 75$, $d = 4$) [HBK84].
- This is an artificial data set with two groups of outliers: observations $1 - 10$ and $11 - 14$.
- We apply classical PCA and robust PCA based on the MCD estimator with $\alpha = 0.5$ (breakdown value).
- This yields the following scores.



Contents

Thank you for your attention! Questions?

[Don82] D.L. Donoho. *Breakdown Properties of Multivariate Location Estimators*. PhD thesis, Harvard University, 1982.

[HBK84] D.M. Hawkins, D. Bradu, and G.V. Kass. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, 26(197–208), 1984.

[MRC+19] M. Maechler, P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, E. L. T. Conceicao, and M. Anna di Palma. *robustbase: Basic Robust Statistics*, 2019. R package version 0.93-5.

[RD99] P. J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[Rou84] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[RPM00] C. Ruffieux, F. Paccaud, and A. Marazzi. Comparing rules for truncating hospital length of stay. *Casemix Quarterly*, 2(1):1422–1424, 2000.

[Sta81] W.A. Stahel. *Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators (In German)*. PhD thesis, Swiss Federal Institute of Technology in Zurich, 1981.

[TF09] V. Todorov and P. Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009.