Anomaly Detection Robustness

Paul Irofti
Cristian Rusu
Andrei Pătraşcu

Computer Science Department
University of Bucharest

Outline

- Simple regression methods and scalar robustness
- Multidimensional regression and trimming
- Clustering, K-Means and Trimmed K-Means

We assume that some one-dimensional measurements data is provided:

$$y_1, y_2, \cdots, y_m \in \mathbb{R}.$$

Small example:

3.6548	2.8729	1.5856	2.4937	1.2595	2.0692

We assume that some one-dimensional measurements data is provided:

$$y_1, y_2, \cdots, y_m \in \mathbb{R}.$$

Therefore, we desire to explain data through a single variable model:

$$y_i = \theta + \epsilon_i \qquad \forall 1 \leq i \leq m,$$

where

- θ is the unknown 1D parameter

We assume that some one-dimensional measurements data is provided:

$$y_1, y_2, \cdots, y_m \in \mathbb{R}.$$

Therefore, we desire to explain data through a single variable model:

$$y_i = \theta + \epsilon_i \qquad \forall 1 \leq i \leq m,$$

where

- θ is the unknown 1D parameter
- ϵ_i normally distributed noise over $\mathcal{N}(0, \sigma)$

We assume that some one-dimensional measurements data is provided:

$$y_1, y_2, \cdots, y_m \in \mathbb{R}.$$

Therefore, we desire to explain data through a single variable model:

$$y_i = \theta + \epsilon_i \qquad \forall 1 \leq i \leq m,$$

where

- θ is the unknown 1D parameter
- ϵ_i normally distributed noise over $\mathcal{N}(0, \sigma)$
- How to compute a good (location) estimator of $\mathbb{E}[y]$?

We assume that some one-dimensional measurements data is provided:

$$y_1, y_2, \cdots, y_m \in \mathbb{R}.$$

Therefore, we desire to explain data through a single variable model:

$$y_i = \theta + \epsilon_i \qquad \forall 1 \leq i \leq m,$$

where

- θ is the unknown 1D parameter
- ϵ_i normally distributed noise over $\mathcal{N}(0, \sigma)$
- How to compute a good (location) estimator of $\mathbb{E}[y]$?
- How to compute a good (scale) estimator of $\sigma(y)$?

$y_1,$	$y_2,$	y_3	$\cdots$	y_{m-1}	y_m
3.6548	2.8729	1.5856	2.4937	1.2595	2.0692

Mean (Least-Squares) estimator:

$$\arg\min_{\theta} \sum_{i=1}^{m} (y_i - \theta)^2.$$

Solution:

$$\theta_{mean} = \frac{1}{m} \sum_{i=1}^{m} y_i \quad (= 2.5)$$

y_1,	y_2,	y_3	$\cdots$	y_{m-1}	y_m
3.6548	2.8729	1.5856	2.4937	1.2595	2.0692

Scale (variance) estimator:

$$\hat{\sigma} = \sqrt{1/m \sum_{i=1}^{m} (y_i - \theta)^2} \quad (= 0.7689)\,.$$

Now assume that we have a measurement error:

y_1,	y_2,	y_3	$\cdots$	y_{m-1}	y_m
3.6548	2.8729	1.5856	2.4937	125.95	2.0692

$$\theta_{mean} = \frac{1}{m}\sum_{i=1}^{m} y_i \quad (= 23.104)$$

$$\hat{\sigma} = \sqrt{1/m\sum_{i=1}^{m}(y_i - \theta)^2} \quad (= 2539)\,.$$

Now assume that we have a measurement error:

y_1,	y_2,	y_3	$\cdots$	y_{m-1}	y_m
3.6548	2.8729	1.5856	2.4937	125.95	2.0692

$$\theta_{mean} = \frac{1}{m}\sum_{i=1}^{m} y_i \quad (= 23.104)$$

$$\hat{\sigma} = \sqrt{1/m \sum_{i=1}^{m} (y_i - \theta_{mean})^2} \quad (= 2539)\,.$$

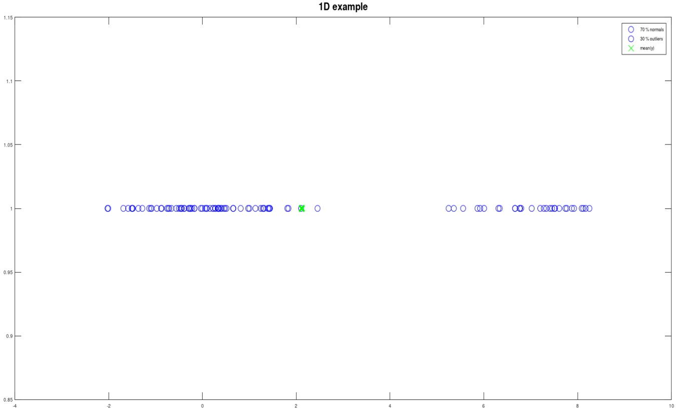
- The error manifests strongly in the mean/scale estimator even for a single outlier!
- We say that the mean estimator has a *breakdown value* of $\frac{1}{m}$ (or 0%)

Definition

The breakdown value *bdv* of a given estimator is given by the smallest proportion of the dataset that need to be replaced in order to carry the estimation arbitrary far away.

- The worst: 0% (the case of the mean estimator)
- The best: 50% (the case of the robust trimmed estimators)

A larger example



70 % Normals : $N(0,1)$ [30% Outliers : $N(7,1)$] $Mean(y) = 2.1161e + 00$

Remark (The idea of trimming estimators)

Trim the both tail sides of the data and evaluate a classical non-robust estimator on the remained data sector.

Remark (The idea of trimming estimators)

Trim the both tail sides of the data and evaluate a classical non-robust estimator on the remained data sector.

- Median
- Least Trimmed Squares
- Least Median Squares
- $\alpha-$ trimmed
- Their generalization to multidimensional context

Remark (The idea of trimming estimators)

Trim the both tail sides of the data and evaluate a classical non-robust estimator on the remained data sector.

Most used in one dimension:

$$\sigma = MAD(y) = \text{med}_i \ (|y_i - \text{med}(y)|)$$

MAD = Median of the Absolute Deviations from the median.

$y_{[1]}$	$\leq y_{[2]}$	$\leq y_{[3]}$	$\cdots$	$\leq y_{[m]}$	$\leq y_{[m]}$
1.5856	2.0692	2.4937	2.8729	3.6548	125.95

Median estimator:

$$\arg\min_{\theta} \sum_{i=1}^{m} |y_i - \theta|.$$

$y_{[1]}$	$\leq y_{[2]}$	$\leq y_{[3]}$	$\cdots$	$\leq y_{[m]}$	$\leq y_{[m]}$
1.5856	2.0692	2.4937	2.8729	3.6548	125.95

Median estimator:

$$\arg\min_{\theta} \sum_{i=1}^{m} |y_i - \theta|.$$

Solution:

$$\theta_{med} = \begin{cases} y_{[m+1/2]} & \text{if n odd} \\ (y_{[m/2]} + y_{[m/2+1]})/2 & \text{if n even.} \end{cases} \quad (= 2.6833)\,.$$

$y_{[1]}$	$\leq y_{[2]}$	$\leq y_{[3]}$	$\cdots$	$\leq y_{[m]}$	$\leq y_{[m]}$
1.5856	2.0692	2.4937	2.8729	3.6548	125.95

Median estimator:

$$\arg\min_{\theta} \sum_{i=1}^{m} |y_i - \theta|.$$

Solution:

$$\theta_{med} = \begin{cases} y_{[m+1/2]} & \text{if n odd} \\ (y_{[m/2]} + y_{[m/2+1]})/2 & \text{if n even.} \end{cases} \quad (= 2.6833)\,.$$

- The median = trim 50% each side of the sorted data

$$\begin{array}{cccccc} y_{[1]} & \leq y_{[2]} & \leq y_{[3]} & \cdots & \leq y_{[m]} & \leq y_{[m]} \\ 1.5856 & 2.0692 & 2.4937 & 2.8729 & 3.6548 & 125.95 \end{array}$$

Median estimator:

$$\arg\min_{\theta} \sum_{i=1}^{m} |y_i - \theta|.$$

Solution:

$$\theta_{med} = \begin{cases} y_{[m+1/2]} & \text{if n odd} \\ (y_{[m/2]} + y_{[m/2+1]})/2 & \text{if n even.} \end{cases} \quad (= 2.6833)\,.$$

- The median = trim 50% each side of the sorted data
- It is robust to up to half of data outliers (the median breakdown value of 50%)

$y_{[1]}$	$\leq y_{[2]}$	$\leq y_{[3]}$	$\cdots$	$\leq y_{[m]}$	$\leq y_{[m]}$
1.5856	2.0692	2.4937	2.8729	3.6548	125.95

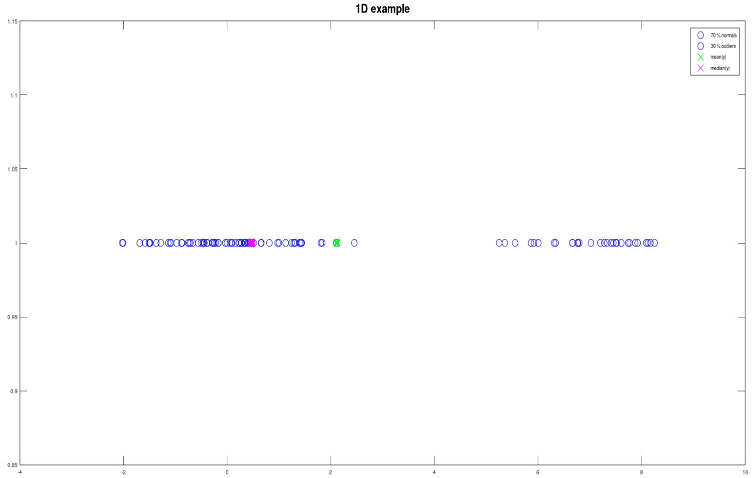
Median estimator:

$$\arg\min_{\theta} \sum_{i=1}^{m} |y_i - \theta|.$$

Solution:

$$\theta_{med} = \begin{cases} y_{[m+1/2]} & \text{if n odd} \\ (y_{[m/2]} + y_{[m/2+1]})/2 & \text{if n even.} \end{cases} \quad (= 2.6833)\,.$$

- The median = trim 50% each side of the sorted data
- It is robust to up to half of data outliers (the median breakdown value of 50%)
- Slightly more costly to compute (than the mean) in the scalar case.



$$70\ \%\ \text{Normals} : N(0,1) \quad [30\%\ \text{Outliers} : N(7,1)]$$

$$Mean(y) = 2.1161e+00,\ \ Med(y) = 4.6965e-01$$

Median

- We saw that the median cuts half of data left and right

- We saw that the median cuts half of data left and right
- This trimming proportion of 50% is fixed ”by the user” and equal on both sides

Median

- We saw that the median cuts half of data left and right
- This trimming proportion of 50% is fixed ”by the user” and equal on both sides
- α-trimming: trimming proportion α, equal on both sides

Median

- We saw that the median cuts half of data left and right
- This trimming proportion of 50% is fixed ”by the user” and equal on both sides
- α-trimming: trimming proportion α, equal on both sides
- May be ”too robust”: α is given a priori (in no connection with the data)

Median

- We saw that the median cuts half of data left and right
- This trimming proportion of 50% is fixed ”by the user” and equal on both sides
- α-trimming: trimming proportion α, equal on both sides
- May be ”too robust”: α is given a priori (in no connection with the data)
- What if the trimming is driven by the data?

Least Median Squares (LMS) estimator:

$$\theta_{LMS} := \arg\min_{\theta} \ \text{med}_i \ (y_i - \theta)^2.$$

Least Median Squares (LMS) estimator:

$$\theta_{LMS} := \arg\min_{\theta} \ \mathrm{med}_i \ (y_i - \theta)^2.$$

- Determines the center of the region where the ”normal” samples stay close together.

Least Median Squares (LMS) estimator:

$$\theta_{LMS} := \arg\min_{\theta} \; \mathrm{med}_i \, (y_i - \theta)^2.$$

- Determines the center of the region where the ”normal” samples stay close together.
- Unlike the median, lets the data to decide the estimated mean

Least Median Squares (LMS) estimator:

$$\theta_{LMS} := \arg\min_{\theta} \ \text{med}_i \ (y_i - \theta)^2.$$

- Determines the center of the region where the ”normal” samples stay close together.
- Unlike the median, lets the data to decide the estimated mean
- To compute LMS, one has to compute the ”shortest” half of data: take $h = \lfloor m/2 \rfloor + 1$

$$\min\{y_{[h]} - y_{[1]}, y_{[h+1]} - y_{[2]}, \cdots, y_{[n]} - y_{[n-h+1]}\}$$

Least Median Squares (LMS) estimator:

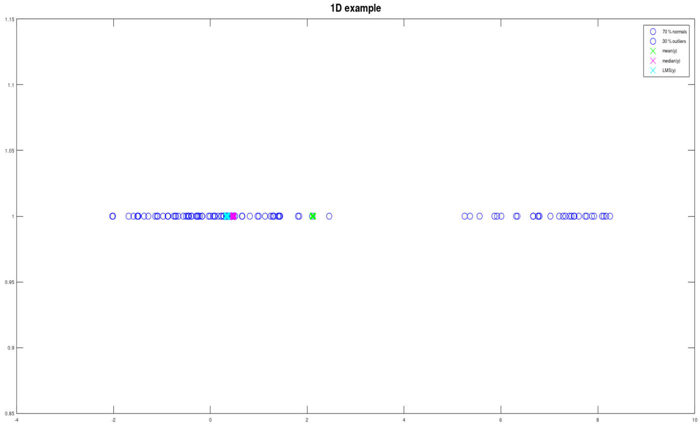
$$\theta_{LMS} := \arg\min_{\theta} \ \mathrm{med}_i \ (y_i - \theta)^2.$$

- Determines the center of the region where the "normal" samples stay close together.
- Unlike the median, lets the data to decide the estimated mean
- To compute LMS, one has to compute the "shortest" half of data: take $h = \lfloor m/2 \rfloor + 1$

$$\min\{y_{[h]} - y_{[1]}, y_{[h+1]} - y_{[2]}, \cdots, y_{[n]} - y_{[n-h+1]}\}$$

- Then, θ_{LMS} is the midpoint of this shortest interval.

LMS



$$70\ \%\ \text{Normals} : N(0,1) \quad [30\%\ \text{Outliers} : N(7,1)]$$

$$Mean(y) = 2.1161e+00,\ Med(y) = 4.6965e-01, LMS(y) = 3.4312e-01$$

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - \theta)^2$ are the residuals.

- A mean estimator over the $h \in [n/2, n]$ "normal" samples.

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - \theta)^2$ are the residuals.

- A mean estimator over the $h \in [n/2, n]$ "normal" samples.
- A small h means a high breakdown vs. a large h better approximates the true mean.

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - \theta)^2$ are the residuals.

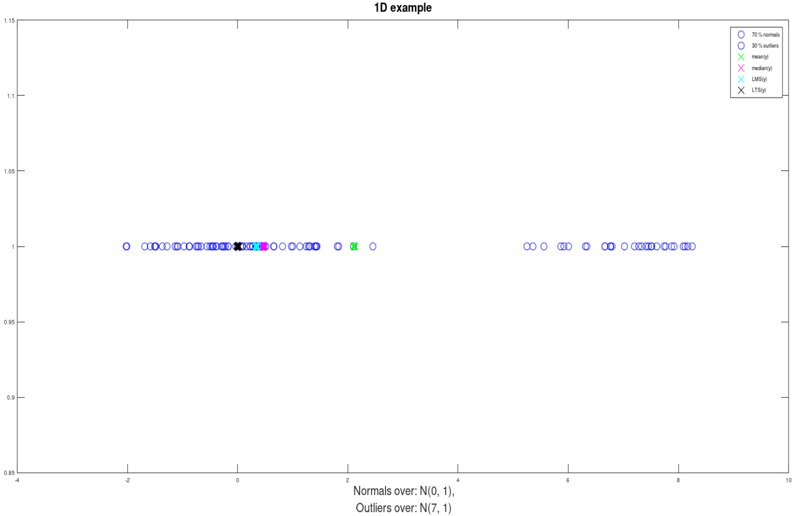
- A mean estimator over the $h \in [n/2, n]$ ”normal” samples.
- A small h means a high breakdown vs. a large h better approximates the true mean.
- Basically, assumes that an outlier is ”far off” the mean of the normal samples.

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - \theta)^2$ are the residuals.

- A mean estimator over the $h \in [n/2, n]$ ”normal” samples.
- A small h means a high breakdown vs. a large h better approximates the true mean.
- Basically, assumes that an outlier is ”far off” the mean of the normal samples.
- As in the previous case, mainly lets the data to decide the estimated mean.



$$Mean(y) = 2.1161e+00,\ Med(y) = 4.6965e-01, LMS(y) = 3.4312e-01,$$
$$LTS(y) = 4.1184e-03$$

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - \theta)^2$ are the residuals.

- To compute LTS, look at subsamples:

$$\{y_{[1]}, \cdots, y_{[h]}\}, \{y_{[2]}, \cdots, y_{[h+1]}\}, \cdots, \{y_{[n-h+1]}, \cdots, y_{[n]}\}$$

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - \theta)^2$ are the residuals.

- To compute LTS, look at subsamples:

$$\{y_{[1]}, \cdots, y_{[h]}\}, \{y_{[2]}, \cdots, y_{[h+1]}\}, \cdots, \{y_{[n-h+1]}, \cdots, y_{[n]}\}$$

- Compute means: $\hat{y}^{(i)} = \frac{1}{h} \sum_{j=i}^{i+h-1} y_{[j]}$

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - \theta)^2$ are the residuals.

- To compute LTS, look at subsamples:

$$\{y_{[1]}, \cdots, y_{[h]}\}, \{y_{[2]}, \cdots, y_{[h+1]}\}, \cdots, \{y_{[n-h+1]}, \cdots, y_{[n]}\}$$

- Compute means: $\hat{y}^{(i)} = \frac{1}{h} \sum_{j=i}^{i+h-1} y_{[j]}$
- Compute sum of squares: $s^{(i)} = \sum_{j=i}^{i+h-1} (y_{[j]} - \hat{y}^{(j)})^2$

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - \theta)^2$ are the residuals.

- To compute LTS, look at subsamples:

$$\{y_{[1]}, \cdots, y_{[h]}\}, \{y_{[2]}, \cdots, y_{[h+1]}\}, \cdots, \{y_{[n-h+1]}, \cdots, y_{[n]}\}$$

- Compute means: $\hat{y}^{(i)} = \frac{1}{h} \sum_{j=i}^{i+h-1} y_{[j]}$
- Compute sum of squares: $s^{(i)} = \sum_{j=i}^{i+h-1} (y_{[j]} - \hat{y}^{(j)})^2$
- Solution: $\hat{y}^{(i)}$ associated with the smallest $s^{(i)}$.

Outline

- Simple regression methods and scalar robustness
- **Multidimensional regression and trimming**
- Clustering, K-Means and Trimmed K-Means

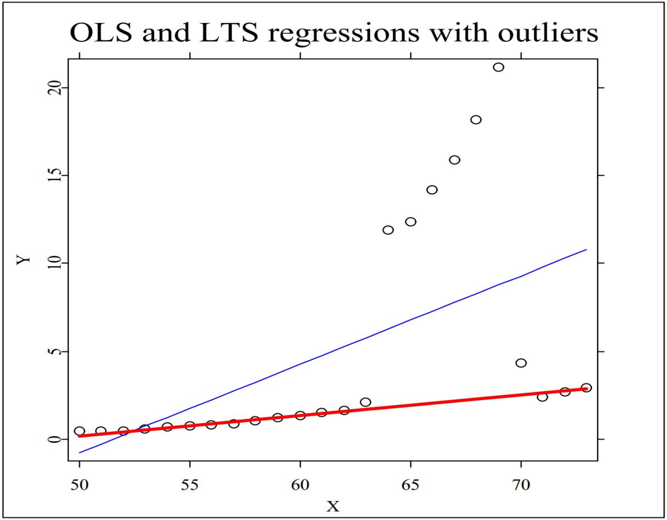
Multidimensional regression

The LR problem:

$$\min_{\theta} \sum_{i=1}^{m} (y_i - a_i^T \theta)^2.$$

- Traditional non-robust regression method with explicit solution.
- It breaks even for a single outlier.
- How to extends the previous models to this problem?

Multidimensional regression



Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - a_i^T\theta)^2$.

- The algorithm from the scalar case does not help!

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - a_i^T\theta)^2$.

$$\begin{aligned}\theta_{LTS} &= \arg\min_{\theta} \min_{\omega\geq 0, \sum_i \omega_i = h} \sum_{i=1}^{n} \omega_i (y_i - a_i^T\theta)^2 \\ &= \arg\min_{\theta,\omega\in\Delta_h} \sum_{i=1}^{n} \omega_i (y_i - a_i^T\theta)^2\end{aligned}$$

Good news: nonconvex in joint variable (θ, ω), but convex over variable θ and ω, separately.

Least Trimmed Squares (LTS) estimator:

$$\theta_{LTS} := \arg\min_{\theta} \sum_{i=1}^{h} r_{[i]}^2(\theta),$$

where $r_i(\theta) := (y_i - a_i^T\theta)^2$.
In other words:

$$\theta(\omega) = \arg\min_{\theta} \sum_{i=1}^{n} \omega_i(y_i - a_i^T\theta)^2 \text{(just a LS solution)}$$

$$\omega(\theta) = \arg\min_{\omega \in \Delta_h} \sum_{i=1}^{n} \omega_i(y_i - a_i^T\theta)^2 \text{(bottom h residuals)}$$

Alternating Minimization algorithm

Initialize $\theta^0 \in \mathbb{R}^n, \omega^0 \in \Delta_h$ and iterate

$$\theta^{k+1} := \arg\min_{\theta} \sum_{i=1}^{n} \omega_i^k (y_i - a_i^T \theta)^2$$

$$\omega^{k+1} = \arg\min_{\omega \in \Delta_h} \sum_{i=1}^{n} \omega_i (y_i - a_i^T \theta^{k+1})^2$$

Initialize $\theta^0 \in \mathbb{R}^n, \omega^0 \in \Delta_h$ and iterate

$$\theta^{k+1} := \arg\min_{\theta} \sum_{i=1}^{n} \omega_i^k (y_i - a_i^T \theta)^2$$

The iteration θ^k is an usual LR estimator over the $h-$subset of data featuring the h smallest residuals.

Initialize $\theta^0 \in \mathbb{R}^n, \omega^0 \in \Delta_h$ and iterate

$$\omega^{k+1} = \arg\min_{\omega \in \Delta_h} \sum_{i=1}^{n} \omega_i (y_i - a_i^T \theta^{k+1})^2$$

For the new θ^{k+1}, the weights ω^{k+1} are the 1 for the new smallest h residual and 0 otherwise.

Alternating Minimization algorithm

- Easy to show that any limit point satisfy:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \omega_i^* r_i(\theta)^2$$

$$\omega^* = \arg\min_{\omega} \sum_{i=1}^{n} \omega_i r_i(\theta^*)^2.$$

- Easy to show that any limit point satisfy:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \omega_i^* r_i(\theta)^2$$

$$\omega^* = \arg\min_{\omega} \sum_{i=1}^{n} \omega_i r_i(\theta^*)^2.$$

- Usually, the rate convergence is relatively high.

- Easy to show that any limit point satisfy:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \omega_i^* r_i(\theta)^2$$
$$\omega^* = \arg\min_{\omega} \sum_{i=1}^{n} \omega_i r_i(\theta^*)^2.$$

- Usually, the rate convergence is relatively high.
- In simple cases, this kind of stationary point is sufficiently close to optimum and provides reasonable estimation.

- Easy to show that any limit point satisfy:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \omega_i^* r_i(\theta)^2$$

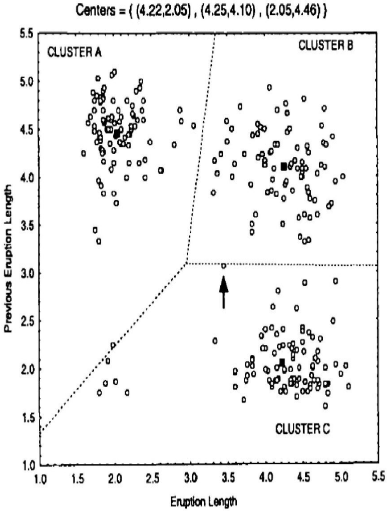
$$\omega^* = \arg\min_{\omega} \sum_{i=1}^{n} \omega_i r_i(\theta^*)^2.$$

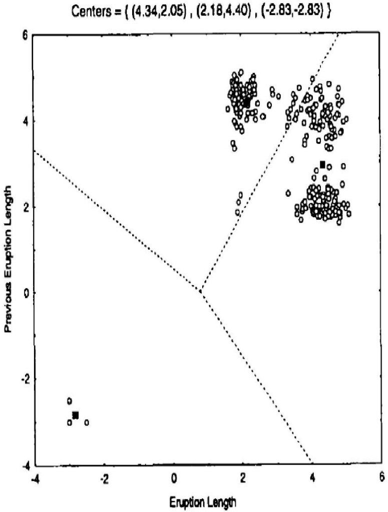
- Usually, the rate convergence is relatively high.
- In simple cases, this kind of stationary point is sufficiently close to optimum and provides reasonable estimation.
- Really hard to improve this class of local minima.

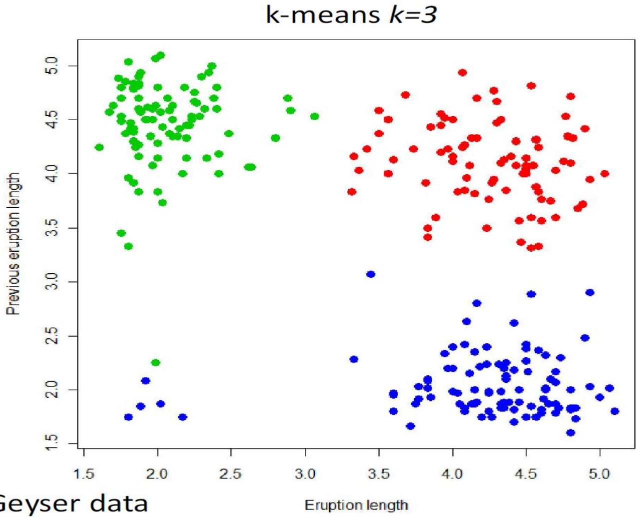
Outline

- Simple regression methods and scalar robustness
- Multidimensional regression and trimming
- **Clustering, K-Means and Trimmed K-Means**

Multidimensional clustering






k-means k=3
Previous eruption length
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Eruption length

Geyser data

The clustering problem:

$$\min_{\theta_1,\cdots,\theta_k} \sum_{i=1}^{m} \min_{1\leq j\leq k} \|y_i - \theta_j\|^2.$$

- Assumes k clusters in data and aims at optimally finding the k centers.
- Highly nonconvex even in 1D.
- K-Means is not robust to outliers: bdp $1/m \to 0\%$ (*when* $m \to \infty$).

The clustering problem:

$$\min_{\theta_1,\cdots,\theta_k} \sum_{i=1}^{m} \min_{1\leq j\leq k} \|y_i - \theta_j\|^2.$$

How we compute the centers? 1) First, introduce slack variables ω

$$\min_{\theta_1,\cdots,\theta_k} \min_{\omega_i\in\Delta_1} \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_i^j \|y_i - \theta_j\|^2.$$

2) Apply alternating minimization scheme over the variables.

Multidimensional clustering

The clustering problem:

$$\min_{\theta_1,\cdots,\theta_k} \sum_{i=1}^{m} \min_{1\leq j\leq k} \|y_i - \theta_j\|^2.$$

Alternating Minimization

While *stopping_criterion = FALSE*:

1. $\theta^i_{k+1} = \arg\min \sum_{j=1}^{m} [\omega^j_k]_i \|y^j - \theta^i\|^2, \qquad \forall 1 \leq i \leq K$
2. $\omega^j_{k+1} = \arg\min \sum_{i=1}^{K} \omega^j_i \|y^j - \theta^i_{k+1}\|^2, \text{ s.l. } \sum_{i=1}^{K} \omega^j_i = 1, \;\; \omega^j \geq 0, \forall 1 \leq j \leq m$
3. $k := k+1$

The clustering problem:

$$\min_{\theta_1,\cdots,\theta_k} \sum_{i=1}^{m} \min_{1\le j\le k} \|y_i - \theta_j\|^2.$$

Alternating Minimization

While *stopping_criterion = FALSE*:

1. $\theta^i_{k+1} = \frac{\sum_{j=1}^{m} [\omega^j_k]_i y^j}{\sum_{j=1}^{m} \omega^j_{i,k}}$
2. $[\omega^j_{k+1}]_i = \begin{cases} 1 & \text{if } i = \arg\min_i \|y^j - \theta^i\|^2 \\ 0 & \text{else} \end{cases}$
3. $k := k+1$

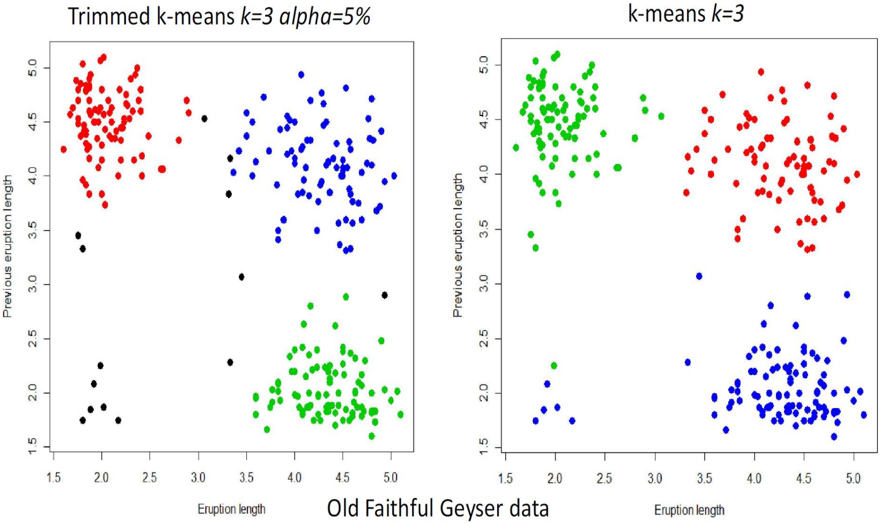
Trimmed-k means:

$$\min_{\theta_1,\cdots,\theta_k} \sum_{i=1}^{h} r_{[i]}(\theta).$$

where $r_i(\theta) = \min_{1 \le j \le k} \|y_i - \theta_j\|^2$.

- We assume k clusters in data and aims at optimally finding the k centers.
- *We trim the points with positions far from any cluster*
- $\alpha = h/m$

Trimmed clustering



Trimmed-k means:

$$\min_{\theta_1,\cdots,\theta_k} \sum_{i=1}^{h} r_{[i]}(\theta) \quad \text{where} \quad r_i(\theta) = \min_{1\le j\le k} \|y_i - \theta_j\|^2.$$

How we compute the centers? 1) Introduce binary slack variables ω, z

$$\min_{\theta_1,\cdots,\theta_k} \min_{\omega_i\in\Delta_1, z\in\Delta_h} \sum_{i=1}^{m} z_i \sum_{j=1}^{k} \omega_i^j \|y_i - \theta_j\|^2.$$

2) Apply alternating minimization scheme over the variables.

Trimmed-k means:

$$\min_{\theta_1,\cdots,\theta_k} \sum_{i=1}^{h} r_{[i]}(\theta) \quad \text{where} \quad r_i(\theta) = \min_{1\leq j\leq k} \|y_i - \theta_j\|^2.$$

Alternating Minimization

While *stopping_criterion = FALSE*:

1. $\theta^i_{k+1} = \arg\min \sum_{j=1}^{m} [z_i]_k [\omega^j_k]_i \|y^j - \theta^i\|^2, \qquad \forall 1 \leq i \leq K$
2. $\omega^j_{k+1} = \arg\min \sum_{i=1}^{K} [z_j]_k \omega^j_i \|y^j - \theta^i_{k+1}\|^2$, s.l. $\sum_{i=1}^{K} \omega^j_i = 1, \;\; \omega^j \geq 0, \forall 1 \leq j \leq m$
3. $z_{k+1} = \arg\min \sum_{i=1}^{K} [z_j]_k r_j(\theta_{k+1})$
4. $k := k + 1$

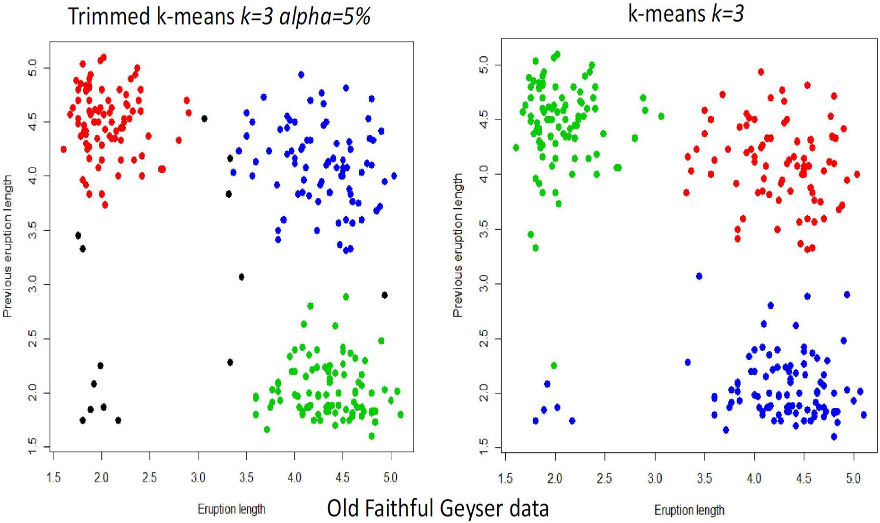
Trimmed-k means:

$$\min_{\theta_1,\cdots,\theta_k} \sum_{i=1}^{h} r_{[i]}(\theta) \quad \text{where} \quad r_i(\theta) = \min_{1\leq j\leq k} \|y_i - \theta_j\|^2.$$

Alternating Minimization

While *stopping_criterion = FALSE*:

1. $\theta^i_{k+1} = \dfrac{\sum_{j=1}^{m} z^k_j [\omega^j_k]_i y^j}{\sum_{j=1}^{m} \omega^j_{i,k}}$
2. $[\omega^j_{k+1}]_i = \begin{cases} 1 & \text{if } i = \arg\min_i \|y^j - \theta^i\|^2 \& rank(y^j) \leq h \\ 0 & \text{else} \end{cases}$
3. $z_{k+1} = hard\_thres(r(\theta))$ % indices of smallest r_i
4. $k := k + 1$


Trimmed k-means k=3 alpha=5%
k-means k=3
Previous eruption length
Eruption length
Old Faithful Geyser data
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

- Robust estimators require tuning parameters in general, e.g. h, α. These parameters encodes our prior knowledge (assumptions) about data. Outlier robustness vs. generalization quality.

- Robust estimators require tuning parameters in general, e.g. h, α. These parameters encodes our prior knowledge (assumptions) about data. Outlier robustness vs. generalization quality.
- Both traditional and robust models are combined in practice; e.g. if their results are highly different, then the data might be contaminated.

- Robust estimators require tuning parameters in general, e.g. h, α. These parameters encodes our prior knowledge (assumptions) about data. Outlier robustness vs. generalization quality.
- Both traditional and robust models are combined in practice; e.g. if their results are highly different, then the data might be contaminated.
- Due to combinatorial nature of robust models, simple algorithms are preffered.

1. García-Escudero, Luis A., and Agustín Mayo-Iscar. ”Robust clustering based on trimming.” Wiley Interdisciplinary Reviews: Computational Statistics 16.4 (2024): e1658.
2. Rousseeuw, Peter J., and Annick M. Leroy. Robust regression and outlier detection. John wiley & sons, 2005.
3. Cizek, Pavel, and J. A. Visek. ”Least trimmed squares.” XPLORE, Application Guide (2000): 49-64.
4. Garcia-Escudero, Luis Angel, and Alfonso Gordaliza. ”Robustness properties of k means and trimmed k means.” Journal of the American Statistical Association 94.447 (1999): 956-969.