Robust clustering

J.D. Gonzalez[1] V.J. Yohai[2] R.H. Zamar[3]

Universidad de Buenos Aires and University of British Columbia

Clustering

- Cluster analysis is a useful tool for unsupervised data analysis.
- The goal is to divide the population in K homogeneous groups $C_1, ..., C_k$ called clusters.
- One way to define the clusters is by giving the centers of the clusters

$$\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$$

- Then

$$C_k = \{\mathbf{x}_i : ||\mathbf{x}_i - \boldsymbol{\mu}_k|| = \min_{1 \leq j \leq K} ||\mathbf{x}_i - \boldsymbol{\mu}_j||\}. \tag{1}$$

- In this talk we will discuss how to choose the centers.

K-means

- One of the most popular procedures to choose the cluster centers is called *K-means*.
- Introduced by Steinhaus (1956) and popularized MacQueen (1967).

Set

$$D(\mathbf{x}_i, \boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K) = \min_{1 \leq k \leq K} ||\mathbf{x}_i - \boldsymbol{\mu}_k||$$

Then

$$(\widehat{\boldsymbol{\mu}}_1, \ldots \widehat{\boldsymbol{\mu}}_K) = \arg \min_{\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K} \frac{1}{n} \sum_{i=1}^{n} D^2(\mathbf{x}_i, \boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K)$$

Robust-clustering

- Cuesta-Albertos et al. (1997) noted that K-means is not robust against outliers and proposed a robust alternative, called trimmed K-means.
- We propose another robust alternative called K-TAU.
- We use the concepts of scale estimators.

Scale estimators

Given real numbers $u_1, ..., u_n$, an scale estimator $s(u_1, ..., u_n)$ measures the absolute size of the sample.

An scale estimator has the following properties:

P1. Only depend on $|u_1|, ..., |u_n|$

P2. $s(0, ..., 0) = 0$

P3. If

$$|v_1| \geq |u_1|, ..., |v_n| \geq |u_n|$$

then

$$s(v_1, ..., v_n) \geq s(u_1, ..., u_n)$$

P4. $s(\lambda u_1, ..., \lambda u_n) = |\lambda| s(u_1, ..., u_n)$

Examples of scale estimators

- L_2 scale

$$s_2(u_1, \ldots, u_n) = \left(1/n \sum_{i=1}^{n} u_i^2\right)^{1/2}$$

- L_1 scale

$$s_1(u_1, \ldots, u_n) = 1/n \sum_{i=1}^{n} |u|_i$$

Examples of robust scale estimators

- Median scale

$$s_{\text{med}}(u_1, ..., u_n) = median(|u_1|, .., |u_n|)$$

- M-scale $\quad s_{\text{M}}(u_1, ..., u_n)$ solves

$$\frac{1}{n}\sum \rho\left(\frac{|u_i|}{s}\right) = b$$

$\rho : R \to [0, \infty)$ is even, non decreasing, $\rho(\infty) = 1$

$b = 1/2$ for breakdown point equal to $1/2$.

The Tau Scale

Yohai and Zamar (1988) introduce a new family of scale estimators, τ-estimators to achieve high Gaussian efficiency and breakdown point $1/2$ simultaneously.

To define a τ scale estimator we use two ρ functions

$$\rho_1 \geq \rho_2$$

We first compute an M-scale $s = s(u_1, ..., u_n)$ using ρ_1

The (squared) τ-scale is then defined as

$$s_\tau^2(u_1, ..., u_n) = s^2 \frac{1}{n} \sum_{i=1}^{n} \rho_2(\frac{u_i}{s})$$

K-Tau clustering procedure

The K-tau cluster centers are given by:

$$(\widetilde{\boldsymbol{\mu}}_1, ...\widetilde{\boldsymbol{\mu}}_K) = \arg \min_{\boldsymbol{\mu}_1,\dots\boldsymbol{\mu}_K} s_\tau(D(\mathbf{x}_1, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K), ..., D(\mathbf{x}_n, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K))$$

and the clusters $C_1, C_2, ..., C_K$ are defined

$$C_k = \{\mathbf{x}_i : \min_{1\leq j\leq K} ||\mathbf{x}_i - \widetilde{\boldsymbol{\mu}}_j|| = ||\mathbf{x}_i - \widetilde{\boldsymbol{\mu}}_k||\}, \qquad (2)$$

as in the case of $K-$means

Estimating equations

The cluster centers corresponding to the $K-$TAU clustering procedure can not be given explicitly, but they satisfy the following fixed point equations

$$\boldsymbol{\mu}_k = \frac{\sum_{i \in C_k} w\left(\frac{\|\mathbf{x}_i - \boldsymbol{\mu}_k\|}{s}\right)\mathbf{x}_i}{\sum_{i \in C_k} w\left(\frac{\|\mathbf{x}_i - \boldsymbol{\mu}_k\|}{s}\right)} \quad 1 \leq k \leq K, \tag{3}$$

with

$$w(t) = A_n \frac{\psi_1(t)}{t} + B_n \frac{\psi_2(t)}{t}. \tag{4}$$

Estimating equations

$$A_n = \sum_{i=1}^{n} \left[2\rho_2 \left(\frac{D(\mathbf{x}_i, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K)}{s} \right) - \psi_2 \left(\frac{D(\mathbf{x}_i, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K)}{s} \right) \right]$$

$$B_n = \sum_{i=1}^{n} \psi_1 \left(\frac{D(\mathbf{x}_i, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K)}{s} \right)$$

and the M-scale s satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{D(\mathbf{x}_i, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K)}{s} \right) = b$$

Estimating equations

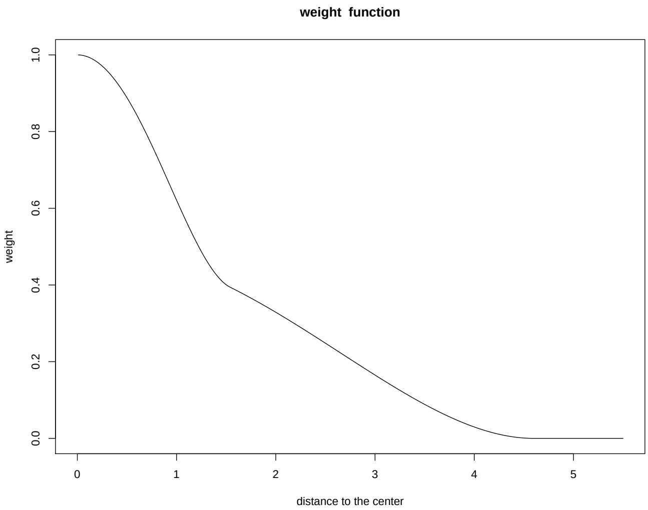


Figure: Weights based on ρ_1 and ρ_2 bisquare functions

K-TAU computing algorithm

The following steps are repeated H times.

Initialization. K initials centers $\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}$ are obtained to start the algorithm.

Update Centers and Groups. Let $\boldsymbol{\mu}_1^{(l)}, \ldots, \boldsymbol{\mu}_K^{(l)}$, $C_1^{(l)}, \ldots, C_K^{(l)}$, and $s^{(l)}$ be the current values.

K-TAU computing algorithm

The new centers are given by

$$\boldsymbol{\mu}_k^{(l+1)} = \frac{\sum_{i \in C_k} w\left(\frac{\|\mathbf{x}_i - \boldsymbol{\mu}_k^{(l)}\|}{s^{(l)}}\right)\mathbf{x}_i}{\sum_{i \in C_k} w\left(\frac{\|\mathbf{x}_i - \boldsymbol{\mu}_k^{(l)}\|}{s^{(l)}}\right)}, 1 \leq k \leq K,$$

and $s^{(l+1)}$ satisfies:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{D(\mathbf{x}_i, \boldsymbol{\mu}_1^{(l)}, \ldots, \boldsymbol{\mu}_K^{(l)})}{s}\right) = b$$

K-TAU computing algorithm

Stopping Rule. Given a tolerance value ε, we stop when

$$\frac{\left\|\boldsymbol{\mu}_k^{(l+1)} - \boldsymbol{\mu}_k^{(l)}\right\|}{\left\|\boldsymbol{\mu}_k^{(l)}\right\|} \leq \varepsilon, \ 1 \leq i \leq K.$$

Accounting for Elliptical Shapes

For each of the clusters we compute robust S-estimators of multivariate location and covariance $\widetilde{\boldsymbol{\mu}}_k$ and $\widetilde{\Sigma}_k$.

The new clusters are defined by

$$C_k = \{\mathbf{x}_i : M(\mathbf{x}_i, \widetilde{\mu}_k, \widetilde{\Sigma}_k\} = \min_{1 \le j \le K} M(\mathbf{x}_i, \widetilde{\mu}_j, \widetilde{\Sigma}_j\}$$

with

$$M((\mathbf{x}, \mu, \boldsymbol{\Sigma}) = [(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)]^{\mathbf{1/2}}$$

Identifying Outliers

Once we have the new clusters, we identify those observations that are outlying:

$$M^2(\mathbf{x}_i, \mu_k, \mathbf{\Sigma}_k) \geq \chi^2_{p,0.975}$$

Here $\chi^2_{p,\alpha}$ is the α quantile of the χ^2 distribution with p degrees of freedom

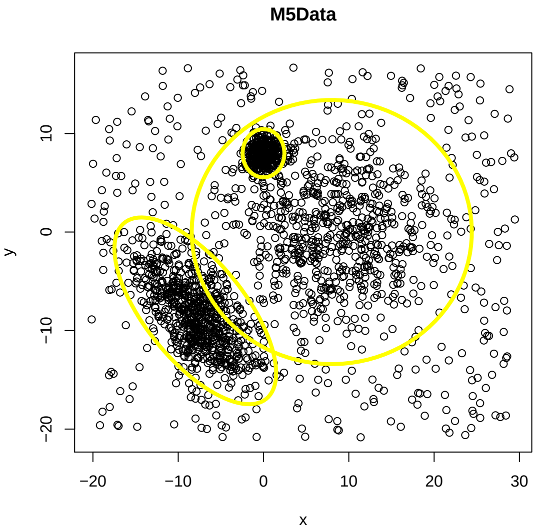
Example

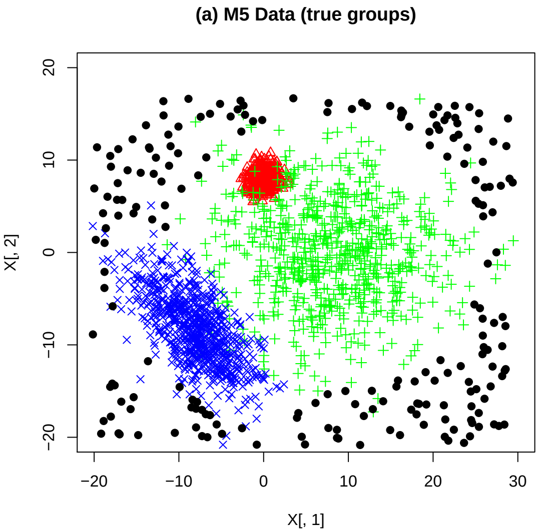
The following example is the set M5 of artificial data generated by García Escudero et al. (2008)
The data set consist of 2000 points from three bivariate normal populations $N(\mu_i, \Sigma_i)$, where

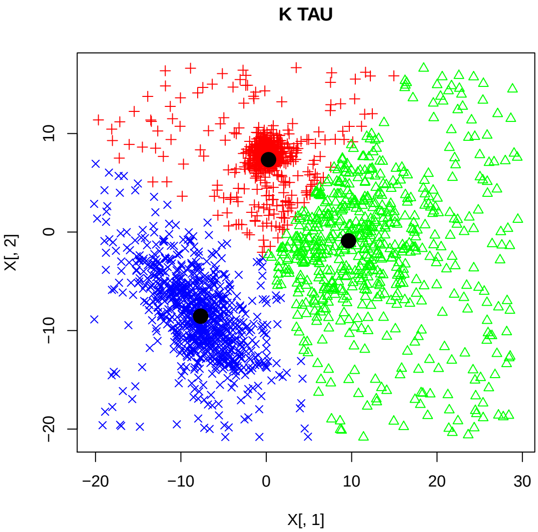
$$\mu_1 = (0, 8), \quad \mu_2 = (8, 0), \quad \mu_3 = (-8, 8)$$

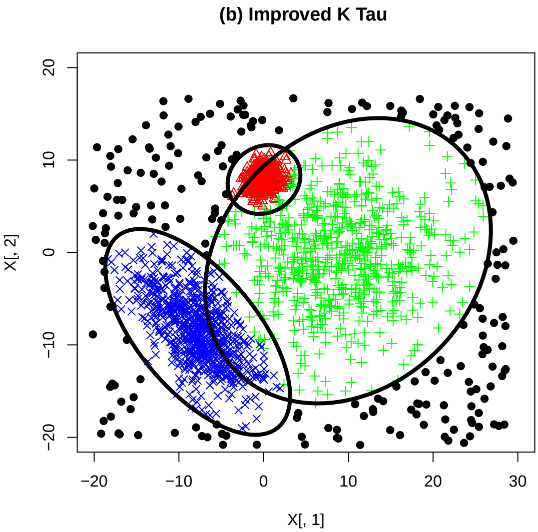
$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 45 & 0 \\ 0 & 30 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 15 & -10 \\ -10 & 15 \end{pmatrix}$$

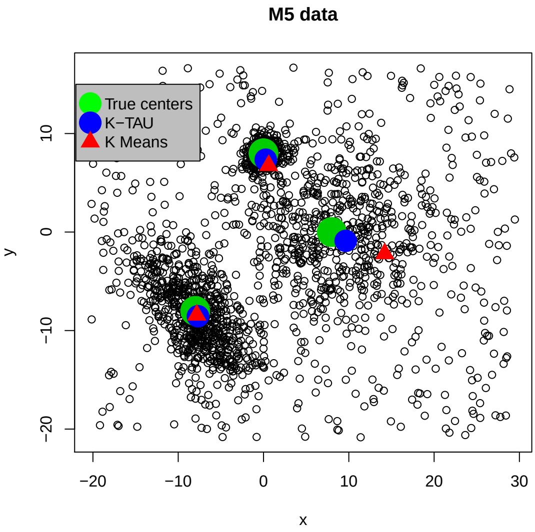
There are 1800 regular points and 200 outliers generated with a uniform distribution in a box around the three populations
From the 1800 regular points, 20% correspond to population 1, 40% to population 2 and 40% to population 3


M5Data
10
0
−10
−20
y
−20
−10
0
10
20
30
x

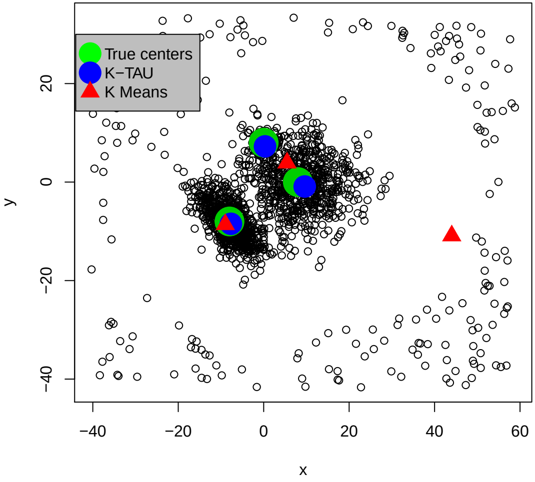

(a) M5 Data (true groups)
X[, 2]
20
10
0
−10
−20
X[, 1]
−20
−10
0
10
20
30


K TAU
X[, 2]
X[, 1]
10
0
−10
−20
−20
−10
0
10
20
30

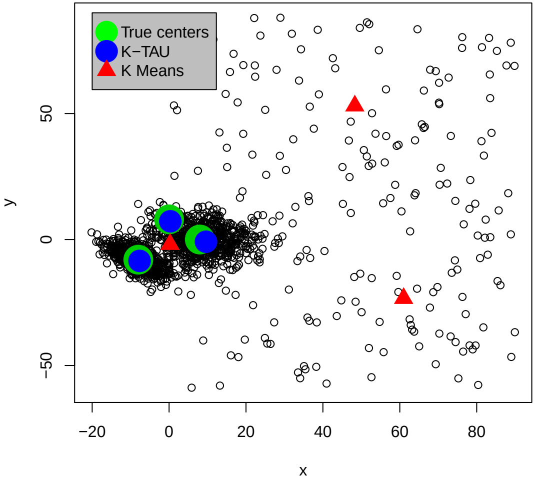

(b) Improved K Tau
X[, 1]
X[, 2]
−20
−10
0
10
20
30


M5 data
True centers
K−TAU
K Means
10
0
−10
−20
y
−20
−10
0
10
20
30
x

M5 data Outliers multiplied by 2



M5 data outliers in [0,90] x [−60,90]

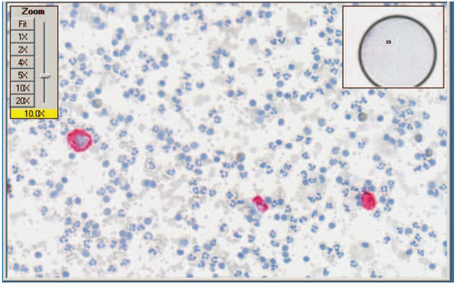


Automatic Processing of Digitized Images

Potential fields of application for robust cluster analysis

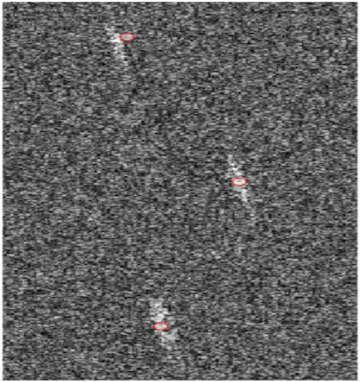
- Computer vision
- Anomaly detection
- Inspection of industrial items
- Search for tumors in microscopic images
- Search for a lost vessel in satellite images

- Micrometastasis of tumor cells in circulating blood
- An algorithm tuned to find the "rare events" can process an entire slide quickly and reliably without fatigue



Satellite Imagery Used by Frontex to Detect and Rescue Migrant Boats

While the use by Frontex of satellite imagery is not new, Frontex released a copy of a satellite image used last week to detect and rescue 370 people on board three inflatable boats off the Libyan coast. (It is unclear whether the image made available by Frontex shows the actual spatial resolution available to Frontex.)



According to Frontex, the imagery is part of "Frontex's Eurosur Fusion Services ... made possible by the cooperation between experts at Frontex and the European Maritime Safety Agency (EMSA), Italian authorities and EUNAVFORMED. ... The Eurosur [fusion] services already include automated large vessel tracking and detection capabilities, software functionalities allowing complex calculations for predicting positions and detecting suspicious activities of vessels, as well as precise weather and oceanographic forecasts. Fusion Services use optical and radar

TELEFE Picture

TELEFE's "No-Signal" TV Image

$$\mathbf{40 \times 50 = 2000 \ pixels}$$

RGB-Color Coding

Each pixel corresponds to a 3-d vector:

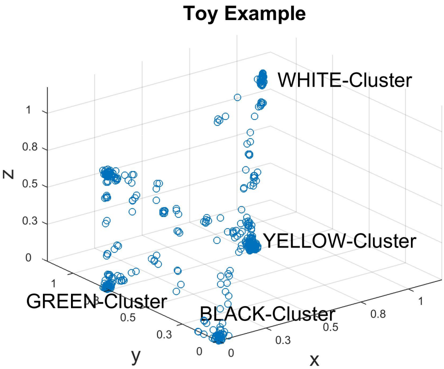
$$(R, \ G, \ B), \quad 0 \le R, G, B \le 1.$$

The vector $(R, \ G, \ B)$ gives the intensity of red, green and blue for the pixel.

For example:

$$\begin{aligned}
(R, G, B) = (1, 0, 0) &= \text{ Red} \\
(R, G, B) = (0, 1, 0) &= \text{ Green} \\
(R, G, B) = (0, 0, 1) &= \text{ Blue} \\
(R, G, B) = (1, 1, 1) &= \text{ White} \\
(R, G, B) = (0, 0, 0) &= \text{ Black} \\
(R, G, B) = (.5, .6, 0) &= \text{ Yellow}
\end{aligned}$$

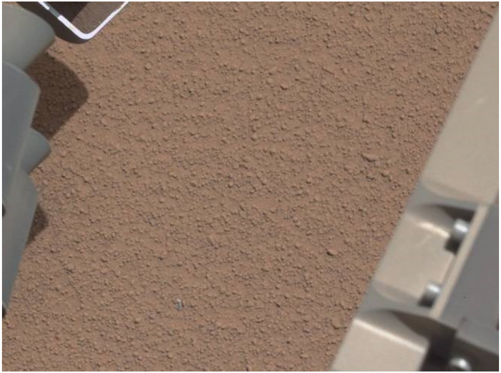
TELEFE Picture Again (left side only)



Mars Rover Curiosity

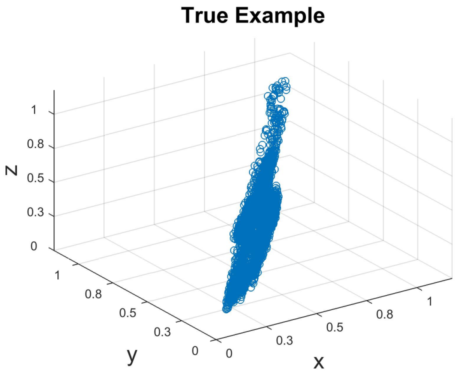
Part of a high resolution NASA's picture

495 $\times$ 664 = 328,680 pixels



Mars Rover Curiosity

RGB - Representation



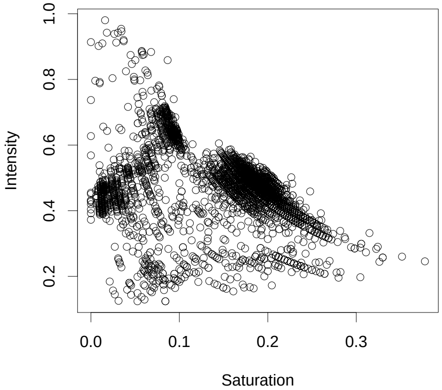
Color picture 2-d representation

$$I = \frac{R + G + B}{3}$$ Intensity

$$S = 1 - \frac{3 \times \min\{R, G, B\}}{R + G + B}$$ Saturation

Mars Rover Curiosity

SI - Representation



Cluster Analysis

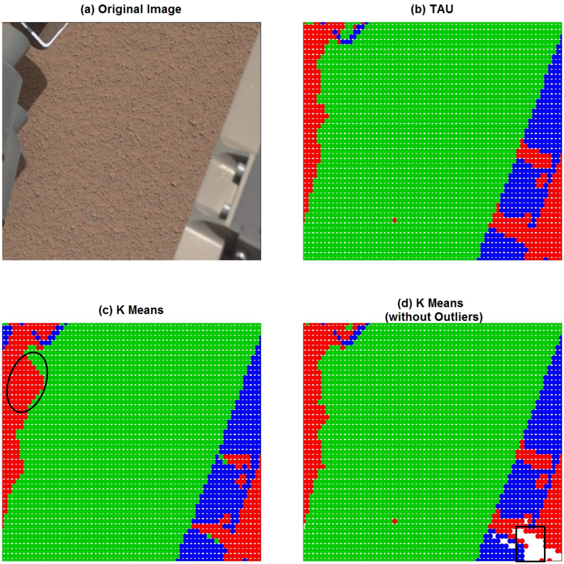
- Image is divided into $n = 3234$ cells with 10×10 pixels
- Each pixel has two values (I, S)
- Each observation represents a 200-d vector

$$(I_1, \ldots, I_{100}, S_1, \ldots, S_{100})$$

- The picture has three components (clusters):
  - *shinning metal (SHM)*
  - *opaque metal (OPM)*
  - *sand (SND)*

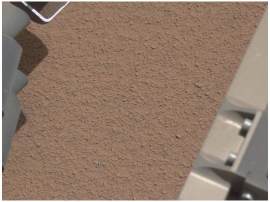
Picture Segmentation

Robust and non-robust clustering

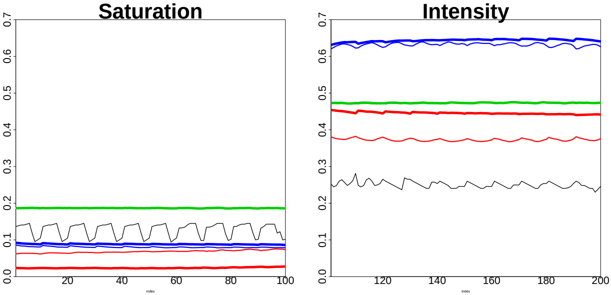


Ouliers

- 15 % of the OPM–cluster cells are very opaque (right lower corner in the picture)
- These cells have unusually low I–levels and high S–levels.
- These outliers upset the K-Means OPM–cluster center.

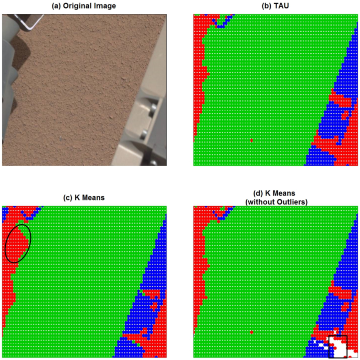


Cluster Centers



Effect of Ouliers

- The shaded sand region is assigned to the OPM–cluster by K-means.
- Recomputing the K-means clusters after removing these outliers validates this reasoning.
- Now the K-means results are consistent with those of the robust clustering procedures.



Automatic detection of missing objects

- The robust clustering sorted the cells as follows:
  - $n_1 = 2500$ SND-cluster cells
  - $n_2 = 405$ OPM-cluster cells
  - $n_3 = 329$ SHM-cluster cells

- The screw is made of a material – opaque metal – that makes up 13% of the image

"Geographic" Step

- Restrict attention to the $n_2 \times 2$ *Geographic Data Matrix* with the position (latitude and longitude) of the OPM–cluster cells
- Perform a second robust cluster analysis on these Geographic data
- Isolated outliers in this second robust clustering are likely to locate the missing piece.

The geographic data