The breakdown behavior of the TCLUST procedure

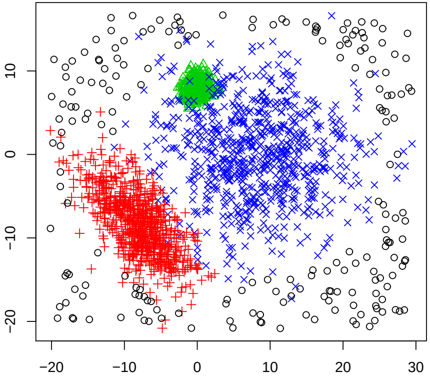
Joint work with L.A. García-Escudero, A. Gordaliza and A. Mayo-Iscar from the University of Valladolid (Spain)

Ch. Ruwet

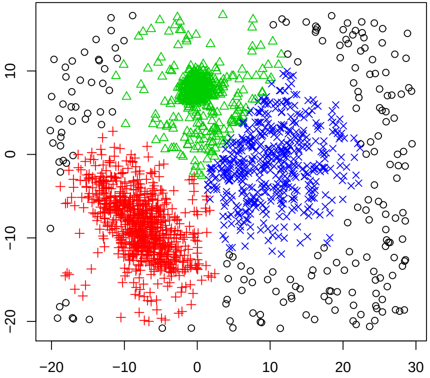
University of Liège

Namur - May 18th 2011

Introduction - Simulated dataset



Introduction - Estimation by trimmed k-means



Outline

- Definition of the TCLUST procedure
- Choice of the different parameters
- A real example
- Breakdown behavior
- Conclusions

Outline

- Definition of the TCLUST procedure
- Choice of the different parameters
- A real example
- Breakdown behavior
- Conclusions

Notations

- k the fixed number of clusters;
- $\alpha \in [0,1[$ the trimming size;

(PR) $X_n = \{x_1, \ldots, x_n\} \in \mathbb{R}^p$ a dataset that is not concentrated on k points after removing a mass equal to α;

- $R_0, R_1, \ldots, R_k$ a partition of $\{1, \ldots, n\}$ with $|R_0| = \lfloor n\alpha \rfloor$;
- $\varphi(\cdot; \mu, \Sigma)$ the probability density function (pdf) of the p-variate normal distribution with mean μ and covariance matrix Σ.

Clustering procedures based on trimming (1)

- The trimmed k-means: k centers $T_1, \ldots, T_k$ that minimize

$$\sum_{j=1}^{k} \sum_{i \in R_j} \|x_i - T_j\|^2$$

(Cuesta-Albertos *et al.*, 1997)

- The trimmed determinant criterion: k centers $T_1, \ldots, T_k$ and a $p \times p$ scatter matrix S that maximize

$$\sum_{j=1}^{k} \sum_{i \in R_j} \log \varphi\left(x_i; T_j, S\right)$$

(Gallegos and Ritter, 2005)

Clustering procedures based on trimming (2)

- Heterogeneous clustering: k centers $T_1, \ldots, T_k$ and k $p \times p$ scatter matrices $S_1, \ldots, S_k$ that maximize

$$\sum_{j=1}^{k} \sum_{i \in R_j} \log \varphi\left(x_i; T_j, S_j\right)$$

under the constraint $\det(S_1) = \ldots = \det(S_k)$ (Gallegos, 2002)

TCLUST procedure $E_{c,k,\alpha}$

García-Escudero *et al.*, 2008

- k centers $T_1, \ldots, T_k$, k $p \times p$ scatter matrices $S_1, \ldots, S_k$ and k weights $p_j \in [0,1], j = 1, \ldots, k$ with $\sum_{j=1}^k p_j = 1$ that maximize

$$\sum_{j=1}^{k} \sum_{i \in R_j} \log \left(p_j \varphi\left(x_i; T_j, S_j\right)\right)$$

- Eigenvalues-ratio restriction **(ER)**:

$$\frac{M_n}{m_n} = \frac{\max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l(S_j)}{\min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l(S_j)} \leq c$$

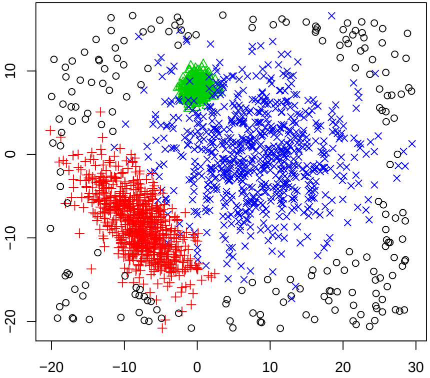
  for a constant $c \geq 1$ and where $\lambda_l(S_j)$ are the eigenvalues of S_j, $l = 1, \ldots, p$ and $j = 1, \ldots, k$.

- $\Theta_c = \left\{\theta = \left(\{p_j\}_{j=1}^k, \{T_j\}_{j=1}^k, \{S_j\}_{j=1}^k\right) : \textbf{(ER)} \text{ is OK}\right\}$

An R package

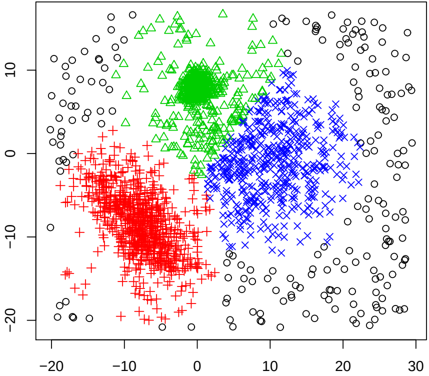

- ```
  > library(tclust)
  > tclust(data, k = 3 , alpha = 0.05,
  restr = "eigen", restr.fact = 12,
  equal.weights = FALSE)
  ```
- `restr` is the type of restriction to be applied: "eigen" (default), "deter" and "sigma"
- `restr.fact` is the constant *c* that constrains the allowed differences among group scatters
- `equal.weights` leads to a model without estimation of the weights
  ```

Example - Simulated dataset



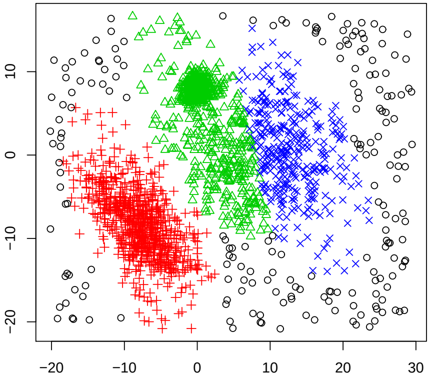
Example - Trimmed *k*-means


```
restr = "eigen", restr.fact = 1,
equal.weights = TRUE
```



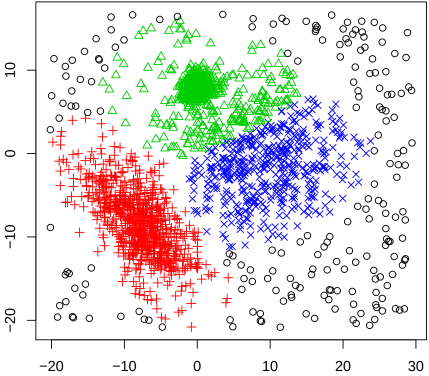
Example - Trimmed determinant criterion

```
restr = "sigma", restr.fact = 1,
equal.weights = TRUE
```



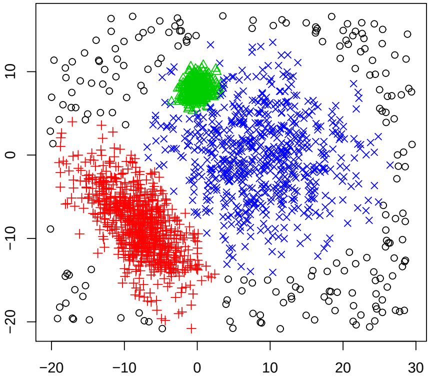
Example - Heterogeneous clustering


```
restr = "deter", restr.fact = 1,
equal.weights = TRUE
```



Example - TCLUST


```
restr = "eigen", restr.fact = 50,
equal.weights = FALSE
```



Outline

- Definition of the TCLUST procedure
- Choice of the different parameters
- A real example
- Breakdown behavior
- Conclusions

Choice of c

- The choice of c should depend on prior knowledge of type of clusters we are searching for;
- Large values of c lead to rather unrestricted solutions;
- Small values of c yield similarly structured clusters;
- This constant can be viewed as a "robustness" constant.

For fixed $c \geq 1$,

- For $k \geq 1$ and $\alpha \in [0, 1[$,

$$\mathcal{L}_c(\alpha, k) := \max_{\{R_j\}_{j=0}^k, \theta \in \Theta_c} \sum_{j=1}^{k} \sum_{i \in R_j} \log\left(p_j \varphi\left(x_i; T_j, S_j\right)\right)$$

- $\Delta_c(\alpha, k) = \mathcal{L}_c(\alpha, k+1) - \mathcal{L}_c(\alpha, k) \geq 0$ is the "gain" achieved by increasing the number of clusters from k to $k+1$
- k^* should be the smallest value of k such that $\Delta_c(\alpha, k) \approx 0$, except for small values of α
- α^* should be the smallest value of α such that $\Delta_c(\alpha, k^*) \approx 0$ for all $\alpha \geq \alpha^*$

Classification trimmed likelihood function

García-Escudero *et al.*, 201x

For fixed $c \geq 1$,

- For $k \geq 1$ and $\alpha \in [0, 1[$,

$$\mathcal{L}_c(\alpha, k) := \max_{\{R_j\}_{j=0}^k, \theta \in \Theta_c} \sum_{j=1}^{k} \sum_{i \in R_j} \log \left(p_j \varphi \left(x_i ; T_j, S_j \right) \right)$$

- $\Delta_c(\alpha, k) = \mathcal{L}_c(\alpha, k+1) - \mathcal{L}_c(\alpha, k) \geq 0$ is the "gain" achieved by increasing the number of clusters from k to $k + 1$
- k^* should be the smallest value of k such that $\Delta_c(\alpha, k) \approx 0$, except for small values of α
- α^* should be the smallest value of α such that $\Delta_c(\alpha, k^*) \approx 0$ for all $\alpha \geq \alpha^*$

For fixed $c \geq 1$,

- For $k \geq 1$ and $\alpha \in [0, 1[$,

$$\mathcal{L}_c(\alpha, k) := \max_{\{R_j\}_{j=0}^k, \theta \in \Theta_c} \sum_{j=1}^{k} \sum_{i \in R_j} \log\left(p_j \varphi\left(x_i; T_j, S_j\right)\right)$$

- $\Delta_c(\alpha, k) = \mathcal{L}_c(\alpha, k+1) - \mathcal{L}_c(\alpha, k) \geq 0$ is the "gain" achieved by increasing the number of clusters from k to $k+1$
- k^* should be the smallest value of k such that $\Delta_c(\alpha, k) \approx 0$, except for small values of α
- α^* should be the smallest value of α such that $\Delta_c(\alpha, k^*) \approx 0$ for all $\alpha \geq \alpha^*$

For fixed $c \geq 1$,

- For $k \geq 1$ and $\alpha \in [0, 1[$,

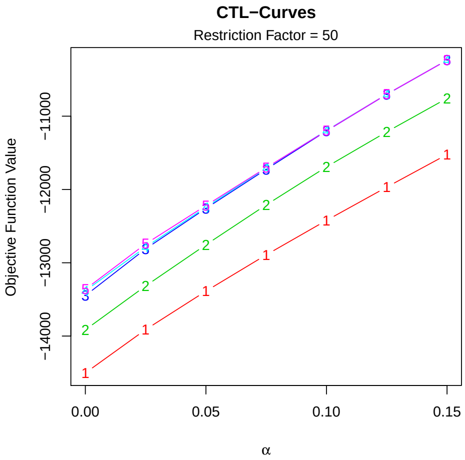
$$\mathcal{L}_c(\alpha, k) := \max_{\{R_j\}_{j=0}^k, \theta \in \Theta_c} \sum_{j=1}^{k} \sum_{i \in R_j} \log \left(p_j \varphi\left(x_i; T_j, S_j\right)\right)$$

- $\Delta_c(\alpha, k) = \mathcal{L}_c(\alpha, k+1) - \mathcal{L}_c(\alpha, k) \geq 0$ is the "gain" achieved by increasing the number of clusters from k to $k+1$
- k^* should be the smallest value of k such that $\Delta_c(\alpha, k) \approx 0$, except for small values of α
- α^* should be the smallest value of α such that $\Delta_c(\alpha, k^*) \approx 0$ for all $\alpha \geq \alpha^*$

The R package


```
R > ctl <- ctlcurves (data, k = 1:4, alpha =
seq (0, 0.2, length = 6),restr.fact = 50)
R > plot(ctl)
```

Example - Simulated dataset



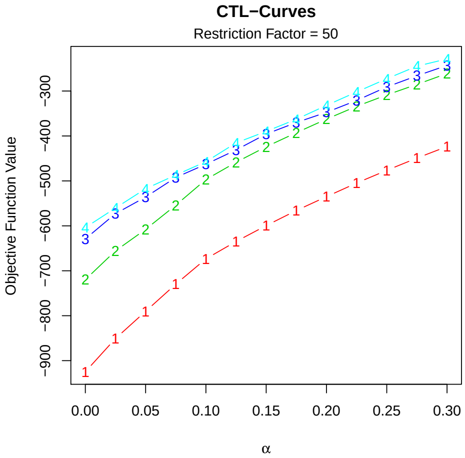
Outline

Swiss bank notes data (1)

- Flury and Riedwyl, 1988
- 6 variables (measurements on the bank notes)
- 200 observations divided in 2 groups: 100 genuine and 100 forged old Swiss 1000-franc bank notes

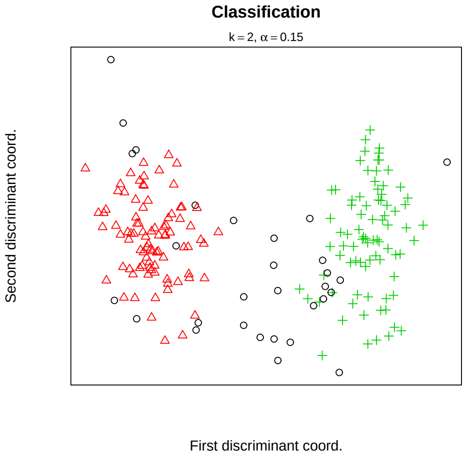
Swiss bank notes data (2)


```
R > plot(ctlcurves(Swiss, k = 1:4 , alpha =
seq(0,0.3,by=0.025))
```



Swiss bank notes data (3)


```
R > plot(tclust(Swiss, k = 2 , alpha = 0.15,
restr.fact = 50)
```



Outline

- Definition of the TCLUST procedure
- Choice of the different parameters
- A real example
- Breakdown behavior
- Conclusions

Intuition about Breakdown

- Breakdown point (BDP): the fraction of outliers needed to bring the estimator to its bounds
  - Replacement BDP (RBDP): observations are replaced by outliers
  - Addition BDP (ABDP): outliers are added
- Explosion of the centers
- $\hat{p}_j = 0$ (sign of a badly chosen k)
- Implosion or explosion of the scatter matrices
  - Some of them : impossible due to **(ER)**
  - All of them : impossible due to existence under **(PR)** (García-Escudero *et al.*, 2008)

Replacement breakdown point

Proposition

The replacement breakdown point of the TCLUST procedure satisfies the optimistic relation

$$RBDP \leq \min\left\{\frac{\lfloor n\alpha\rfloor+1}{n}, \min_{j=1,\dots,k}\frac{|C_j|}{n}\right\}.$$

- Data dependent
- Same upper bound as the trimmed k-means (García-Escudero and Gordaliza, 1999) even if we expect a smaller RBDP for the TCLUST (estimation of weights and scatters)

Ideal model of "well-clustered" data sets
Hennig, 2004

- $n_1 < \ldots < n_k$ and $A_m^j = \{x_{(n_{j-1}+1),m}, \ldots, x_{n_j,m}\}$, $j = 1, \ldots, k$;
- $X_m = \bigcup_{j=1}^k A_m^j$ is said to be "well k-clustered" if $\exists\, b < \infty$ s.t., $\forall\, m \in \mathbb{N}$,
  - (1) $\max_{1 \le j \le k} \max_{x_{i,m}, x_{l,m} \in A_m^j} \|x_{i,m} - x_{l,m}\| < b$
  - (2) $\lim_{m \to \infty} \min_{x_{i,m} \in A_m^h, x_{l,m} \in A_m^j, j \ne h} \|x_{i,m} - x_{l,m}\| = \infty$;
- Addition of r outliers $y_{1,m}, \ldots, y_{r,m}$:
  - (3) $\lim_{m \to \infty} \min \|y_{i,m} - x_{l,m}\| = \infty$
  - (4) $\lim_{m \to \infty} \min_{i \ne l} \|y_{i,m} - y_{l,m}\| = \infty$.

Addition breakdown point

Proposition

Let $X_m, m \in \mathbb{N}$, be an ideal sequence of data sets in $\mathbb{R}^p$ that are "well k-clustered" in clusters $A_m^1, \ldots, A_m^k$ verifying conditions (1) and (2). The addition of $r \leq \lfloor n\alpha \rfloor$ outliers verifying conditions (3) and (4) does not break down the TCLUST procedure with trimming size α:

$$ABDP \geq \frac{\lfloor n\alpha \rfloor}{n + \lfloor n\alpha \rfloor}.$$

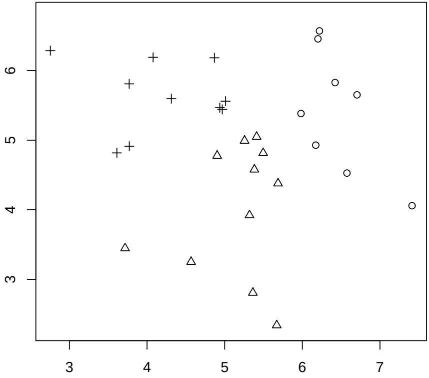
Better than fitting mixtures of t distributions or adding a noise component in normal mixtures (Hennig, 2004).

Dissolution point
Hennig, 2008

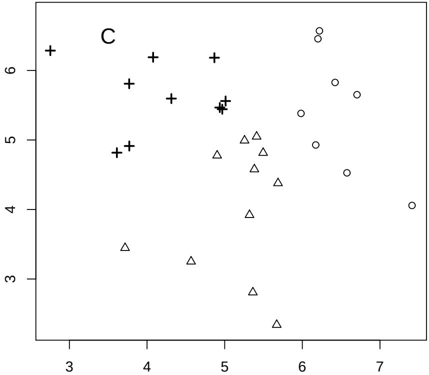
- $E_{c,k,\alpha}(X_n)$ the TCLUST clustering of X_n;
- $E^*_{c,k,\alpha}(X_{n+g})$ the clustering of X_n induced by $E_{c,k,\alpha}(X_{n+g})$;
- $\mathcal{P}$ a partition of X_n;
- For $C \in \mathcal{P}_1$ and $D \in \mathcal{P}_2$, $\gamma(C,D) = \dfrac{|C \cap D|}{|C \cup D|}$;
- A cluster $C \in \mathcal{P}_1$ is dissolved in $\mathcal{P}_2$ if

$$\max_{D \in \mathcal{P}_2} \gamma(C,D) \leq \frac{1}{2}.$$

Example (1)

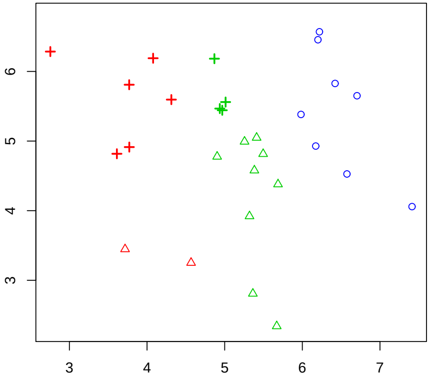


Example (2)



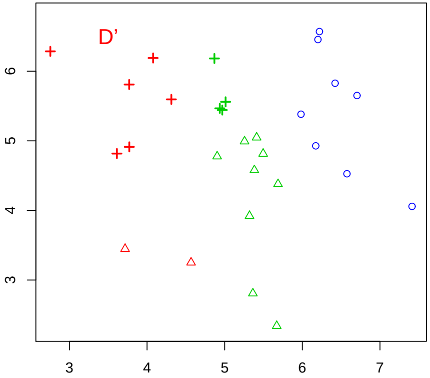
Example (3)

$\mathcal{P}$



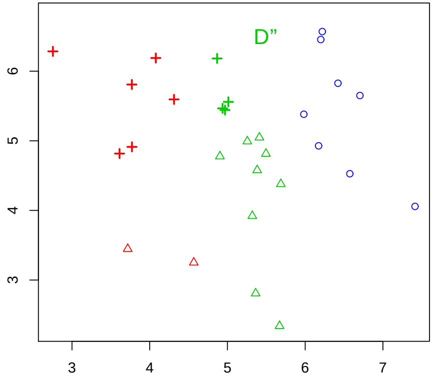
Example (4)

$$|C \cap D'| = 6 \text{ and } |C \cup D'| = 10 + 8$$



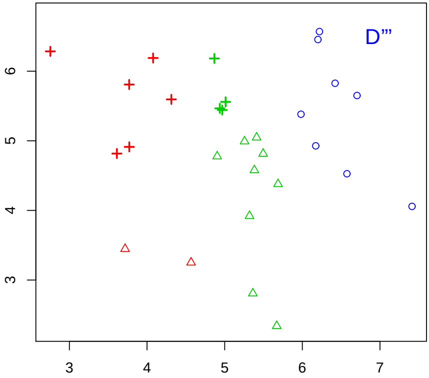
Example (5)

$$|C \cap D''| = 4 \text{ and } |C \cup D''| = 10 + 13$$



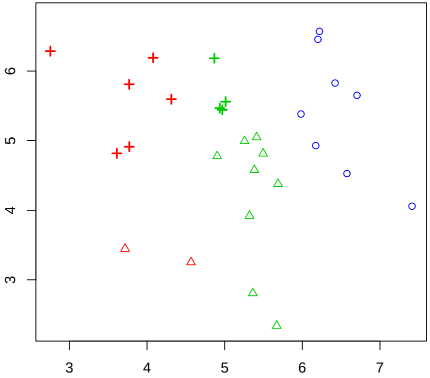
Example (6)

$$|C \cap D'''| = 0 \text{ and } |C \cup D'''| = 10 + 8$$



Example (7)

$$\max_{D \in \mathcal{P}} \gamma(C, D) = 1/3 < 1/2 \rightarrow \text{ C is dissolved in } \mathcal{P}$$



For $C \in E_{c,k,\alpha}(X_n)$, the dissolution point of C is given by

$$\Delta(E_{c,k,\alpha}, X_n, C) = \min_g \left\{ \frac{g}{|C| + g} : \exists x_{n+1}, \ldots, x_{n+g} : \max_{D \in E^*_{c,k,\alpha}(X_{n+g})} \gamma(C, D) \leq 1/2 \right\}.$$

Intuition about the dissolution point theorem

- $g \leq \lfloor n\alpha \rfloor$
- X_n a dataset for which there is no high concentration in X_{n+g} whatever the g added outliers
- $C \in E_{c,k,\alpha}(X_n)$ with $|C| > g$

If there are g points among the trimmed observations that are fitted well enough by the TCLUST clustering, then the cluster C can not be dissolved by the addition of g outliers.

Isolation robustness
Hennig, 2008

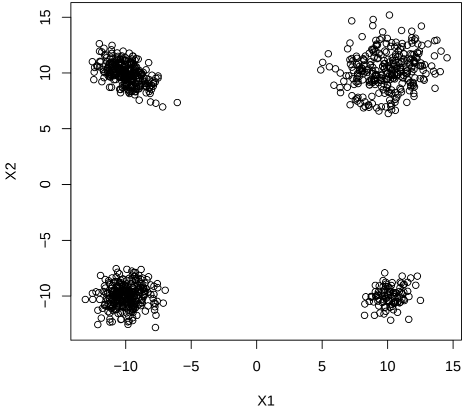
A clustering procedure is said to be isolation robust if for any dataset X_n and for any "well-isolated" cluster C of the partition,

- C is be stable under the addition of points, i.e. for all g, any cluster of the partition of X_{n+g} should not join observations of C and $X_n \backslash C$

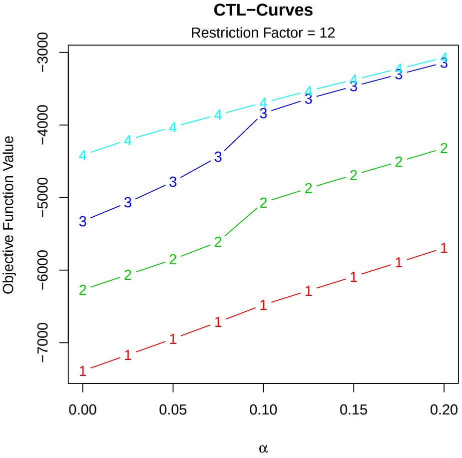
and

- there is at least one cluster in the new partition containing some observations of C.

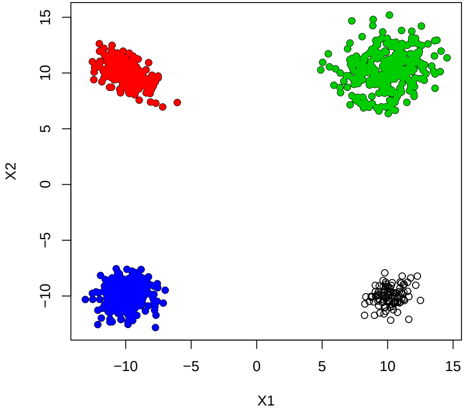
Example



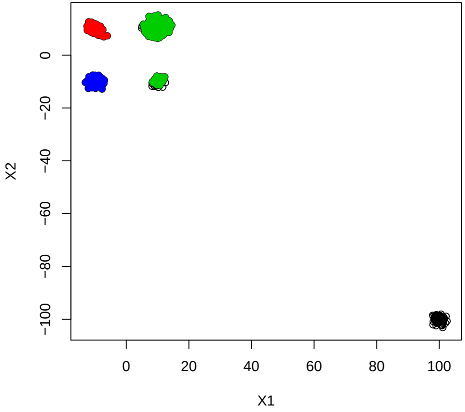
Choice of k and α



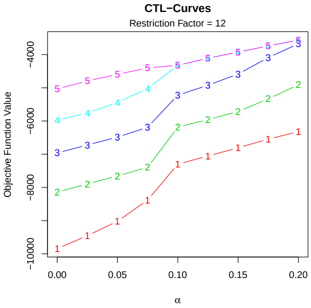
$k = 3, \alpha = 0.1, c = 12$

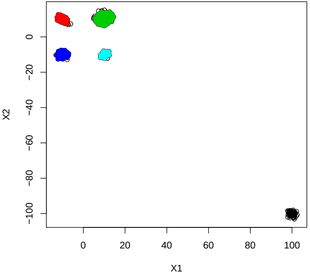


The TCLUST $E_{c,k,\alpha}$ is not isolation robust



The "2-steps" procedure E_c is isolation robust !





Outline

- Definition of the TCLUST procedure
- Choice of the different parameters
- A real example
- Breakdown behavior
- Conclusions

Conclusions

- A flexible clustering procedure;
- A complete R package;
- A graphical tool to chose the parameters;
- Good breakdown behavior under the ideal model of "well-clustered" dataset;
- Isolation robustness of the "2-steps" procedure.

Moreover, the influence functions (not presented here) are bounded.

References (1)

- Cuesta-Albertos J.A., Gordaliza A. and Matrán C. (1997) Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.*, 25(2):553-576
- Gallegos M.T. (2002) Maximum likelihood clustering with outliers. *Classification, clustering, and data analysis (Cracow, 2002)*, Stud. Classification Data Anal. Knowledge Organ., pages 247-255
- Gallegos M.T. and Ritter G. (2005) A robust method for cluster analysis. *Ann. Statist.*, 33(1):347-380
- García-Escudero L.A. and Gordaliza A. (1999) Robustness properties of k-means and trimmed k-means. *J. Amer. Statist. Assoc.*, 94(447):956-969
- García-Escudero L.A., Gordaliza A., Matrán C. and Mayo-Iscar A. (2008) A general trimming approach to robust cluster analysis. *Ann. Statist.*, 36(3):1324-1345

Université de Liège

References (2)

- García-Escudero L.A., Gordaliza A., Matrán C. and Mayo-Iscar A. (201x) Exploring the number of groups in robust model-based clustering. *Stat Comput*
- Flury B. and Riedwyl H. (1988) *Multivariate Statistics. A practical approach.* Chapman and Hall, London
- Hennig C. (2004) Breakdown points for maximum likelihood estimators of location-scale mixtures. *Ann. Statist.*, 32(4):1313-1340
- Hennig C. (2008) Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *J. Multivariate Anal.*, 99(6):1154-1176
- R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.