Robustness of classification based on clustering

Ch. Ruwet

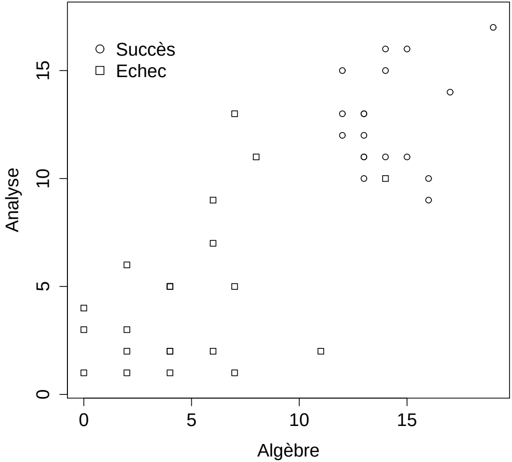
DEPARTMENT OF MATHEMATICS - UNIVERSITY OF LIÈGE

September 2nd 2011

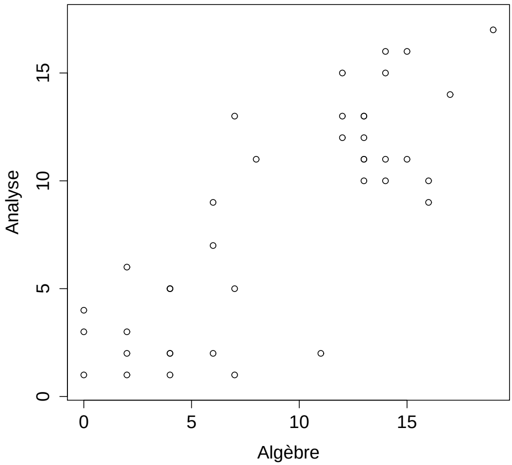
Outline

Outline

Classification



Clustering



The k-means clustering method

- $X_n = \{x_1, \ldots, x_n\}$ a dataset in p dimensions;
- Aim of clustering : Group similar observations in k clusters $C_1, \ldots, C_k$;
- The k-means algorithm constructs clusters in order to minimize the within cluster sum of squared distances
  - The clusters centers $(T_1^n, \ldots, T_k^n)$ are solutions of

$$\min_{\{t_1,\ldots,t_k\}\subset\mathbb{R}^p} \sum_{i=1}^{n} \left(\inf_{1\leq j\leq k} \|x_i - t_j\| \right)^2 ;$$

  - The classification rule:

$$x \in C_j^n \Leftrightarrow \|x - T_j^n\| = \min_{1\leq i\leq k} \|x - T_i^n\|;$$

- Let us focus on $k = 2$ groups:

$$C_1^n = \left\{ x \in \mathbb{R}^p : (T_1^n - T_2^n)^t x - \frac{1}{2} \left(\|T_1^n\|^2 - \|T_2^n\|^2 > 0 \right) \right\}.$$

The k-means clustering method

- $X_n = \{x_1, \ldots, x_n\}$ a dataset in p dimensions;
- Aim of clustering : Group similar observations in k clusters $C_1, \ldots, C_k$;
- The k-means algorithm constructs clusters in order to minimize the within cluster sum of squared distances
  - The clusters centers $(T_1^n, \ldots, T_k^n)$ are solutions of

$$\min_{\{t_1,\ldots,t_k\}\subset\mathbb{R}^p} \sum_{i=1}^{n} \left(\inf_{1\leq j\leq k} \|x_i - t_j\| \right)^2 ;$$

  - The classification rule:

$$x \in C_j^n \Leftrightarrow \|x - T_j^n\| = \min_{1\leq i\leq k} \|x - T_i^n\|;$$

- Let us focus on $k = 2$ groups:

$$C_1^n = \left\{ x \in \mathbb{R}^p : (T_1^n - T_2^n)^t x - \frac{1}{2}\left(\|T_1^n\|^2 - \|T_2^n\|^2 > 0 \right) \right\}.$$

The k-means clustering method

- $X_n = \{x_1, \ldots, x_n\}$ a dataset in p dimensions;
- Aim of clustering : Group similar observations in k clusters $C_1, \ldots, C_k$;
- The k-means algorithm constructs clusters in order to minimize the within cluster sum of squared distances
  - The clusters centers $(T_1^n, \ldots, T_k^n)$ are solutions of

$$\min_{\{t_1,\ldots,t_k\}\subset\mathbb{R}^p} \sum_{i=1}^{n} \left(\inf_{1\leq j\leq k} \|x_i - t_j\| \right)^2 ;$$

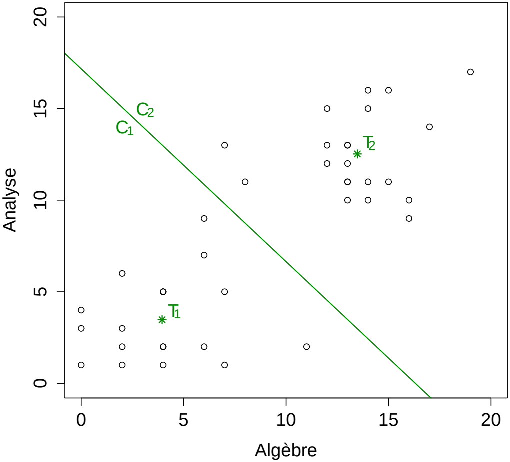
  - The classification rule:

$$x \in C_j^n \Leftrightarrow \|x - T_j^n\| = \min_{1\leq i\leq k} \|x - T_i^n\|;$$

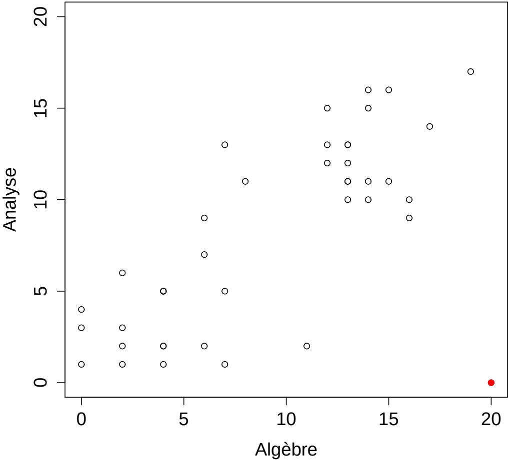
- Let us focus on $k = 2$ groups:

$$C_1^n = \left\{ x \in \mathbb{R}^p : (T_1^n - T_2^n)^t x - \frac{1}{2}\left(\|T_1^n\|^2 - \|T_2^n\|^2 > 0 \right) \right\}.$$

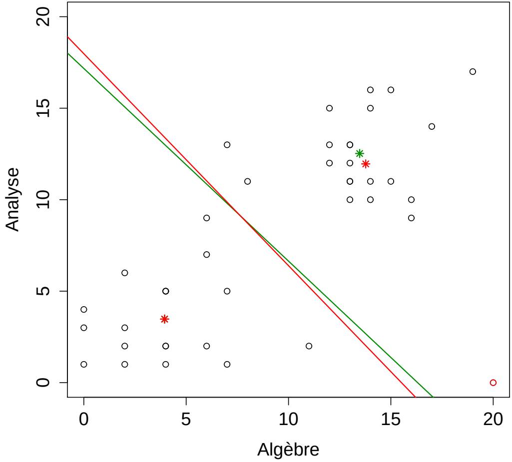
Result of 2-means



Introduction of contamination



Result of 2-means with contamination



The generalized 2-means clustering method

- The clusters centers (T_1^n, T_2^n) are solution of

$$\min_{\{t_1,t_2\}\subset\mathbb{R}^p} \sum_{i=1}^{n} \Omega\left(\inf_{1\le j\le 2} \|x_i - t_j\|\right)$$

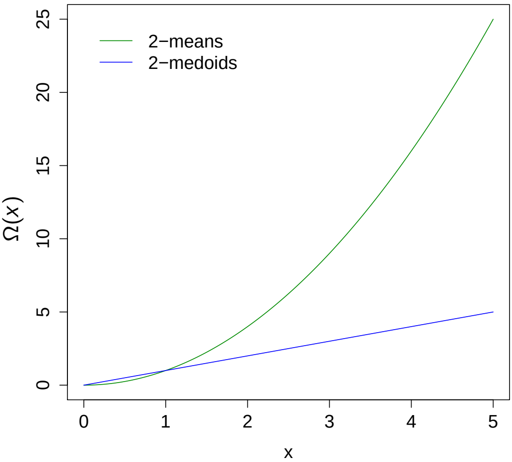
  for an increasing penalty function $\Omega : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\Omega(0) = 0$.
- Classical penalty functions:

$$\Omega(x) = x^2 \to \text{ 2-means method}$$
$$\Omega(x) = x \to \text{ 2-medoids method}$$

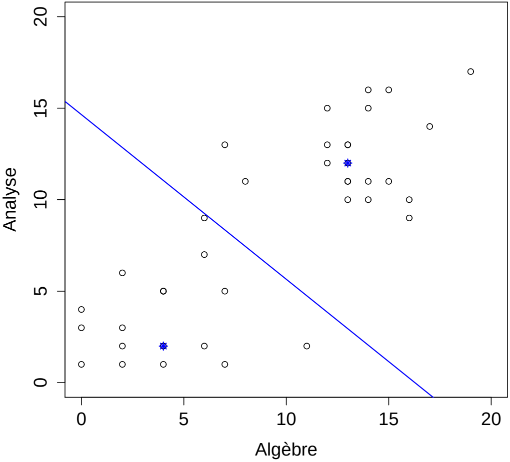
- The classification rule:

$$x \in C_1^n \Leftrightarrow \Omega(\|x - T_1^n\|) \le \Omega(\|x - T_2^n\|)$$
$$\Leftrightarrow \|x - T_1^n\| \le \|x - T_2^n\|.$$

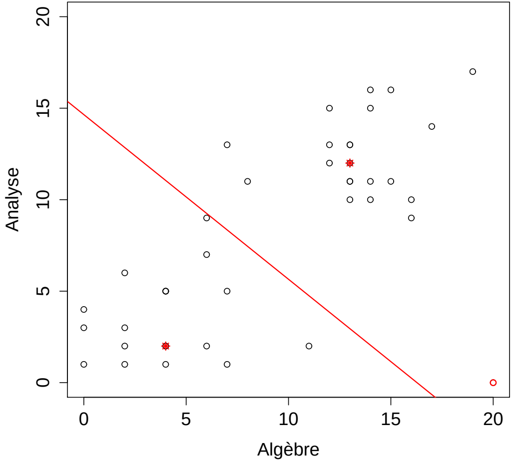
Classical penalty functions



Result of 2-medoids



Result of 2-medoids with contamination



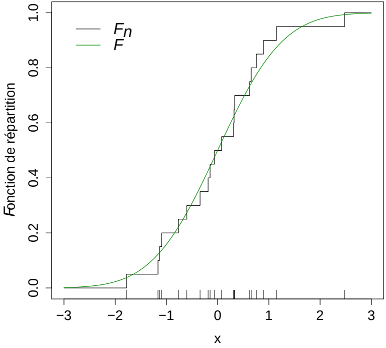
Outline

Statistical functionals

- The empirical distribution F_n is replaced by a cumulative distribution $F \in \mathcal{F}$;
- A statistical functional

$$T : \mathcal{F} \to \mathbb{R}^l : F \mapsto T(F)$$

such that $T(F_n) = T^n$.



Classification setting

Suppose

$$X \sim F \text{ arises from } G_1 \text{ and } G_2 \text{ with } \pi_i(F) = \mathbb{P}_F[X \in G_i]$$

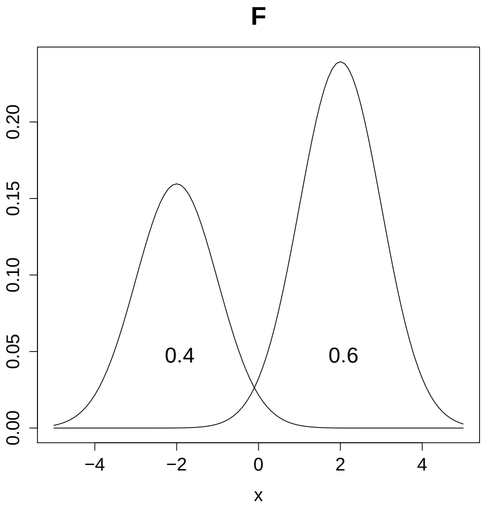
then

F is a mixture of two distributions

$$F = \pi_1(F)F_1 + \pi_2(F)F_2$$

with $\pi_1 + \pi_2 = 1$ and where F_1 and F_2 are the conditional distributions under G_1 and G_2 with densities f_1 and f_2.

Univariate mixture distribution



The generalized 2-means statistical functionals

- The clusters centers $(T_1(F), T_2(F))$ are solution of

$$\min_{\{t_1,t_2\}\subset\mathbb{R}^p} \int \Omega\left(\inf_{1\leq j\leq 2} \|x - t_j\|\right) dF(x)$$

for a suitable increasing penalty function Ω;

- The classification rule is

$$R_F : x \mapsto j \Leftrightarrow \|x - T_j(F)\| = \min_{1\leq i\leq 2} \|x - T_i(F)\|;$$

- The clusters are

$$C_1(F) = \left\{x \in \mathbb{R}^p : A(F)^t x + b(F) > 0\right\}$$
$$C_2(F) = \mathbb{R}^p \backslash C_1(F)$$

with $A(F) = T_1(F) - T_2(F)$
and $b(F) = -\frac{1}{2}\left(\|T_1(F)\|^2 - \|T_2(F)\|^2\right)$.

Mixture of spherically symmetric distributions

$X \sim F_{\mu,\sigma^2}$ if its density is

$$f_{\mu,\sigma^2}(x) = \frac{K}{\sigma^p}\, g\left(\frac{(x-\mu)^t(x-\mu)}{\sigma^2}\right)$$

where g is a non-increasing generator function and with K a constant such that the honesty condition holds.

Examples:

- Multivariate Normal distribution: $g(r) = \exp(-\frac{r}{2})$
- Multivariate Student distribution: $g(r) = \left(1 + \frac{r}{\nu}\right)^{-\frac{\nu+p}{2}}$

Mixture of spherically symmetric distributions

$X \sim F_{\mu,\sigma^2}$ if its density is

$$f_{\mu,\sigma^2}(x) = \frac{K}{\sigma^p}\, g\left(\frac{(x-\mu)^t(x-\mu)}{\sigma^2}\right)$$

where g is a non-increasing generator function and with K a constant such that the honesty condition holds.

Examples:

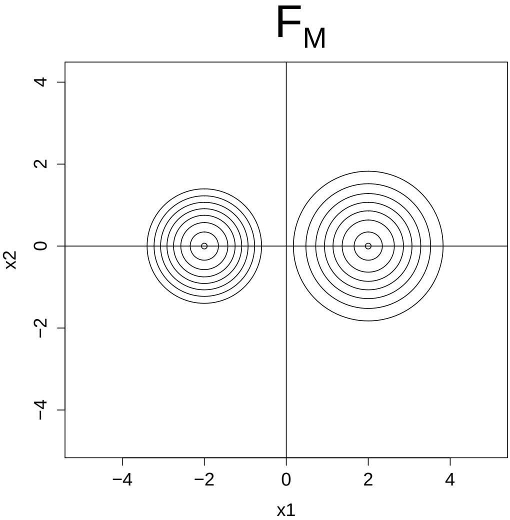
- Multivariate Normal distribution: $g(r) = \exp(-\frac{r}{2})$
- Multivariate Student distribution: $g(r) = \left(1 + \frac{r}{\nu}\right)^{-\frac{\nu+p}{2}}$

Model **(M)**:

$$\textbf{(M)} \qquad F_M = \pi_1 F_{-\mu,\sigma^2} + \pi_2 F_{\mu,\sigma^2}$$

with $\mu = \mu_1 e_1$ and $\mu_1 > 0$.

Representation



Position of $T_1(F_M)$ and $T_2(F_M)$

- 2-means:

Proposition (Kurata and Qiu, 2011)

*Under the model distribution **(M)**, the 2-means centers are on the first axis.*

- Generalized 2-means:

Conjecture (Ruwet and Haesbroeck, 2011)

*Under the model distribution **(M)**, the generalized 2-means centers are on the first axis.*

Position of $T_1(F_M)$ and $T_2(F_M)$

- 2-means:

Proposition (Kurata and Qiu, 2011)

Under the model distribution ***(M)****, the 2-means centers are on the first axis.*

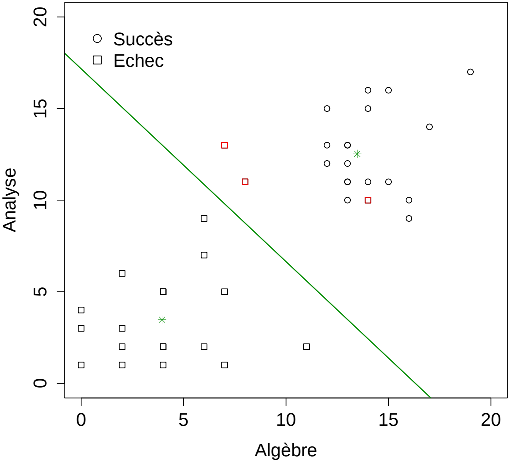
- Generalized 2-means:

Conjecture (Ruwet and Haesbroeck, 2011)

Under the model distribution ***(M)****, the generalized 2-means centers are on the first axis.*

Outline

Error rate of the 2-means



Error rate

- Training sample according to F : estimation of the rule
- Test sample according to F_m : evaluation of the rule
- In ideal circumstances : $F = F_m$
- Probability to misclassify data coming from F_m:

$$\begin{aligned} \mathrm{ER}(F, F_m) &= \pi_1(F_m)\mathbb{P}_{F_m}\left[R_F(X) \neq 1 \,|\, G_1\right] \\ &\quad + \pi_2(F_m)\mathbb{P}_{F_m}\left[R_F(X) \neq 2 \,|\, G_2\right] \\ &= \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_F(X) \neq j \,|\, G_j\right] \end{aligned}$$

Error rate

- Training sample according to F : estimation of the rule
- Test sample according to F_m : evaluation of the rule
- In ideal circumstances : $F = F_m$
- Probability to misclassify data coming from F_m:

$$
\begin{aligned}
\mathrm{ER}(F, F_m) &= \pi_1(F_m)\mathbb{P}_{F_m}\left[R_F(X) \neq 1 | G_1\right] \\
&\quad + \pi_2(F_m)\mathbb{P}_{F_m}\left[R_F(X) \neq 2 | G_2\right] \\
&= \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_F(X) \neq j | G_j\right]
\end{aligned}
$$

Optimality in classification

- A classification rule is optimal if the corresponding error rate is minimal;
- The optimal classification rule is the Bayes rule :

$$x \in C_1(F) \Leftrightarrow \pi_1(F)f_1(x) > \pi_2(F)f_2(x)$$

(Anderson, 1958).

Optimality in classification

- A classification rule is optimal if the corresponding error rate is minimal;
- The optimal classification rule is the Bayes rule :

$$x \in C_1(F) \Leftrightarrow \pi_1(F) f_1(x) > \pi_2(F) f_2(x)$$

(Anderson, 1958).

Proposition (Ruwet and Haesbroeck, 2011)

The 2-means procedure is optimal under the model

$$F_O = 0.5\, F_{-\mu,\sigma^2} + 0.5\, F_{\mu,\sigma^2} \text{ with } \mu = \mu_1\, e_1 \text{ and } \mu_1 > 0.$$

With the Conjecture, the generalized 2-means procedures are also optimal under F_O.

Contaminated distribution

A contaminated distribution is defined by

$$\begin{tikzcd} & F_{\varepsilon} \arrow[dl] \arrow[dr] & \\ 1-\varepsilon : F & & \varepsilon : G \end{tikzcd}$$

where G is a arbitrary distribution function.

Contaminated distribution

A contaminated distribution is defined by

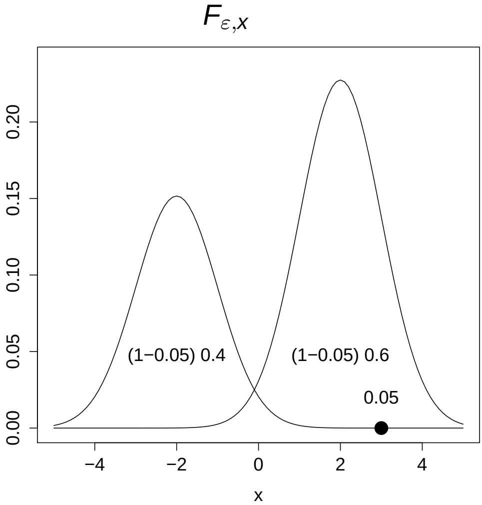
$$\begin{tikzcd} & F_\varepsilon \arrow[dl] \arrow[dr] & \\ 1-\varepsilon : F & & \varepsilon : G \end{tikzcd}$$

where G is a arbitrary distribution function.

To see the influence of one singular point x, $G = \Delta_x$ leading to

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

Contaminated mixture



Error rate under contamination

- Contaminated training sample according to F_ε : estimation of the rule
- Test sample according to F_m : evaluation of the rule

$$\mathrm{ER}(F_\varepsilon, F_m) = \sum_{j=1}^{2} \pi_j(F_m) \mathbb{P}_{F_m}\left[R_{F_\varepsilon}(X) \neq j \,|\, G_j\right]$$

Outline

Definition and properties of the first order influence function

Hampel *et al.* (1986) : For any statistical functional T and any distribution F,

- $$\mathrm{IF}(x;\mathrm{T},F) = \lim_{\varepsilon\to 0}\frac{\mathrm{T}((1-\varepsilon)F+\varepsilon\Delta_x)-\mathrm{T}(F)}{\varepsilon} = \frac{\partial}{\partial\varepsilon}\mathrm{T}((1-\varepsilon)F+\varepsilon\Delta_x)\Big|_{\varepsilon=0}$$
  (under condition of existence);
- $E_F[\mathrm{IF}(X;\mathrm{T},F)] = 0$;
- First order Taylor expansion of T at F:
  $$\mathrm{T}(F_{\varepsilon,x}) \approx \mathrm{T}(F) + \varepsilon\mathrm{IF}(x;\mathrm{T},F)$$
  for ε small enough.

Definition and properties of the first order influence function

Hampel *et al.* (1986) : For any statistical functional T and any distribution F,

- $$\begin{aligned}\mathrm{IF}(x;\mathrm{T},F) &= \lim_{\varepsilon\to 0}\frac{\mathrm{T}((1-\varepsilon)F+\varepsilon\Delta_x)-\mathrm{T}(F)}{\varepsilon}\\ &= \frac{\partial}{\partial\varepsilon}\mathrm{T}((1-\varepsilon)F+\varepsilon\Delta_x)\Big|_{\varepsilon=0}\end{aligned}$$
  (under condition of existence);
- $E_F[\mathrm{IF}(X;\mathrm{T},F)]=0$;
- First order Taylor expansion of T at F:
  $$\mathrm{T}(F_{\varepsilon,x})\approx\mathrm{T}(F)+\varepsilon\mathrm{IF}(x;\mathrm{T},F)$$
  for ε small enough.

First order influence function of the error rate

Now, the training sample is distributed as $F_{\varepsilon,x}$.

First order influence function of the error rate

Now, the training sample is distributed as $F_{\varepsilon,x}$.

If the model distribution is F_O,

- $\mathrm{ER}(F_{\varepsilon,x}, F_O) \approx \mathrm{ER}(F_O, F_O) + \varepsilon \mathrm{IF}(x; \mathrm{ER}, F_O)$
- $\mathrm{ER}(F_{\varepsilon,x}, F_O) \geq \mathrm{ER}(F_O, F_O)$
- $E_{F_O}[\mathrm{IF}(X; \mathrm{ER}, F_O)] = 0$

First order influence function of the error rate

Now, the training sample is distributed as $F_{\varepsilon,x}$.

If the model distribution is F_O,

- $\mathrm{ER}(F_{\varepsilon,x}, F_O) \approx \mathrm{ER}(F_O, F_O) + \varepsilon \mathrm{IF}(x; \mathrm{ER}, F_O)$
- $\mathrm{ER}(F_{\varepsilon,x}, F_O) \geq \mathrm{ER}(F_O, F_O)$
- $E_{F_O}[\mathrm{IF}(X; \mathrm{ER}, F_O)] = 0$

$$\Rightarrow \mathrm{IF}(x; \mathrm{ER}, F_O) \equiv 0$$

A second order term is necessary in the Taylor expansion !

Definition of the second order influence function

For any statistical functional T and any distribution F,

$$\mathrm{IF2}(x; \mathrm{T}, F_O) = \frac{\partial^2}{\partial \varepsilon^2} \mathrm{T}((1-\varepsilon)F_O + \varepsilon \Delta_x)\Big|_{\varepsilon=0}$$

(under condition of existence).

Definition of the second order influence function

For any statistical functional T and any distribution F,

$$\mathrm{IF2}(x;\mathrm{T},F_O)=\frac{\partial^2}{\partial\varepsilon^2}\mathrm{T}((1-\varepsilon)F_O+\varepsilon\Delta_x)\Big|_{\varepsilon=0}$$

(under condition of existence).

Second order Taylor expansion of ER at F_O :

$$\mathrm{ER}(F_{\varepsilon,x},F_O)\approx\mathrm{ER}(F_O,F_O)+\frac{\varepsilon^2}{2}\mathrm{IF2}(x;\mathrm{ER},F_O)$$

for ε small enough.

First order influence function of the error rate under F_M

Proposition (Ruwet and Haesbroeck, 2011)

Under F_M, the first order influence function of the error rate of the generalized 2-means classification procedure is given by

$$\mathrm{IF}(x;\mathrm{ER},F_M) = \frac{\pi_2 - \pi_1}{2} f_{\mu,\sigma^2}(0)\big(\mathrm{IF}(x;T_1,F_M)+\mathrm{IF}(x;T_2,F_M)\big)^t e_1$$

for all x such that $A(F_M)^t x + b(F_M) \neq 0$.

This influence function is bounded as soon as the influence functions of the generalized 2-means centers (see next slide) are bounded.

The influence function is also available for any model distribution F.

First order influence function of the generalized 2-means centers

Proposition (García-Escudero and Gordaliza, 1999)

The influence function of the generalized 2-means centers T_1 and T_2 is given by

$$\begin{pmatrix} \mathrm{IF}(x; T_1, F_m) \\ \mathrm{IF}(x; T_2, F_m) \end{pmatrix} = M^{-1} \begin{pmatrix} \omega_1(x) \\ \omega_2(x) \end{pmatrix}$$

where $\omega_i(x) = -\, grad_y \Omega(\|y\|)\Big|_{y = x - T_i(F_m)} \mathrm{I}(x \in C_i(F_m))$ *and where the matrix* M *depends only on the distribution* F_m.

This influence function is bounded as soon as M^{-1} exists and as soon as the gradient of Ω is bounded.

Univariate second order influence function of the error rate under F_O

Proposition (Ruwet and Haesbroeck, 2011)

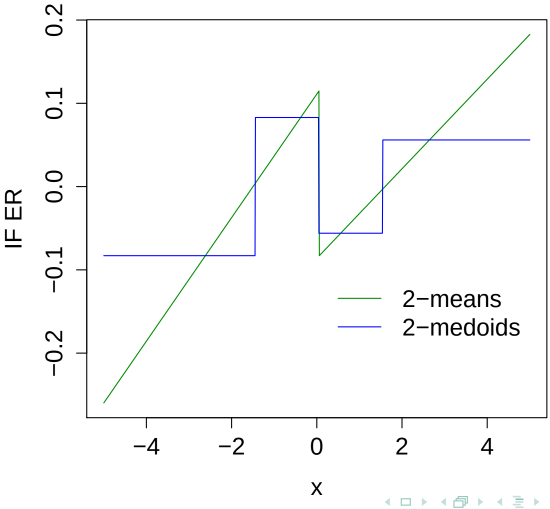
Under F_O, the univariate second order influence function of the error rate of the generalized 2-means classification procedure is given by

$$\mathrm{IF2}(x; \mathrm{ER}, F_O) = -\frac{1}{4} f'_{-\mu,\sigma^2}(0) \big(\mathrm{IF}(x; T_1, F_O) + \mathrm{IF}(x; T_2, F_O)\big)^2$$

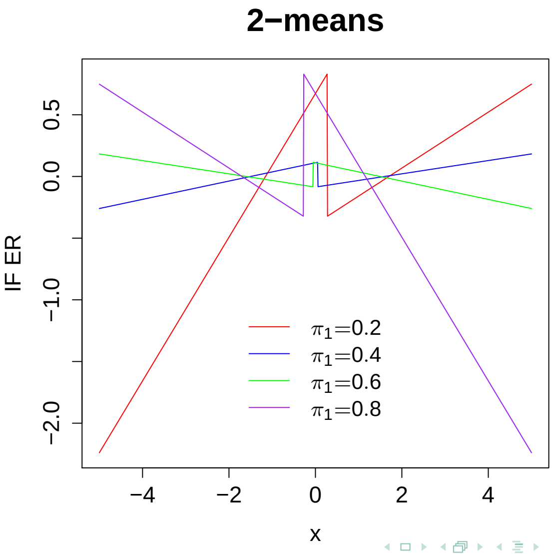
for all x such that $A(F_O)\,x + b(F_O) \neq 0$.

The influence function is also available for multivariate distributions.

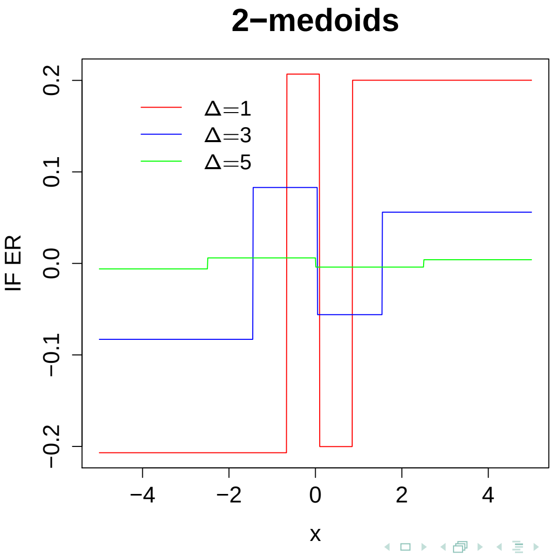
Graph of $\mathrm{IF}(x; \mathrm{ER}, F_M)$



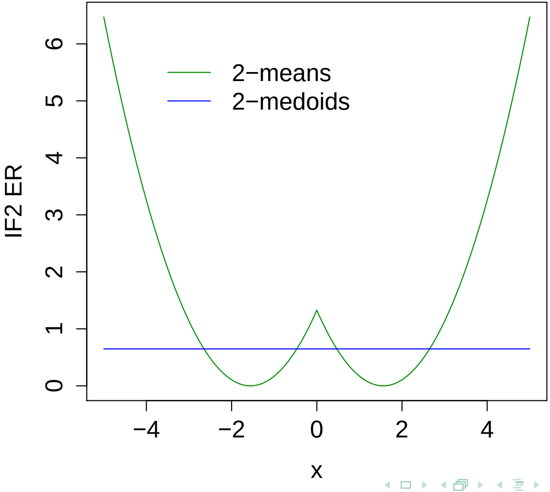
Graph of IF$(x; \text{ER}, F_M)$



Graph of IF$(x; \mathrm{ER}, F_M)$



Graph of IF2$(x; \mathrm{ER}, F_O)$



Outline

Asymptotic loss

Under optimality (F_O), a measure of the expected increase in error rate when estimating the optimal clustering rule from a finite sample with empirical cdf F_n is

$$\text{A-Loss} = \lim_{n\to\infty} n\, E_{F_O}[\mathrm{ER}(F_n, F_O) - \mathrm{ER}(F_O, F_O)].$$

As in Croux *et al.* (2008) :

Proposition

Under some regularity conditions of the clusters centers estimators,

$$A\text{-}Loss = \frac{1}{2} E_{F_O}[IF2(X; ER, F_O)]$$

Asymptotic loss

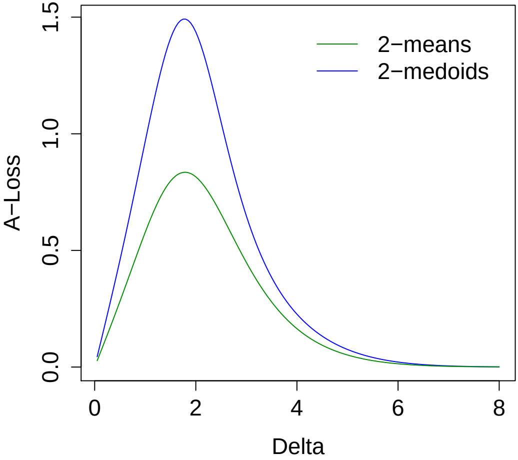
Proposition (Ruwet and Haesbroeck, 2011)

Under an optimal mixture of normal distributions, F_N, with $\mu = \Delta/2\, e_1$, the asymptotic loss of the generalized 2-means procedure is given by

$$\text{A-Loss} = \frac{\Delta}{16\sigma^3\tau^2} f_{0,1}\left(\frac{\Delta}{2\sigma}\right)\Big(\tau^2[\text{ASV}(T_{21}) + \text{ASV}(T_{11}) + 2\text{ASC}(T_{11}, T_{21})] + \sigma^2[\text{ASV}(T_{12}) + \text{ASV}(T_{22}) - 2\text{ASC}(T_{12}, T_{22})]\Big)$$

where ASV *and* ASC *stand for the asymptotic variance and covariance of their component (at the model distribution).*

Graph of the asymptotic loss



Asymptotic relative classification efficiencies

A measure of the price one needs to pay in error rate for protection against the outliers when using a robust procedure instead of the classical one is

$$\text{ARCE(Robust,Classical)} = \frac{\text{A-Loss(Classical)}}{\text{A-Loss(Robust)}}.$$

Asymptotic relative classification efficiencies

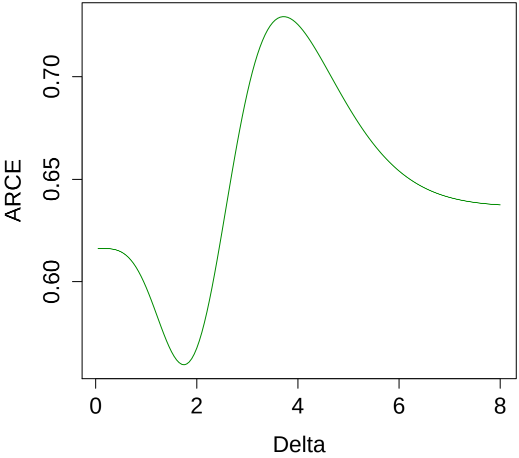
A measure of the price one needs to pay in error rate for protection against the outliers when using a robust procedure instead of the classical one is

$$\text{ARCE(Robust,Classical)} = \frac{\text{A-Loss(Classical)}}{\text{A-Loss(Robust)}}.$$

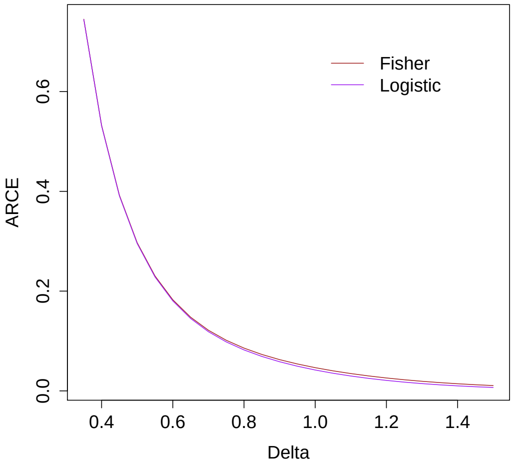
More generally, the ARCE of a method (Method 1) w.r.t. another one (Method 2) is given by

$$\text{ARCE(Method 1,Method 2)} = \frac{\text{A-Loss(Method 2)}}{\text{A-Loss(Method 1)}}.$$

ARCE of 2-medoids w.r.t. 2-means



ARCE of classification procedures w.r.t. 2-means



Outline

Intuitive definition

The breakdown point (BDP) is the minimal fraction of outliers that needs to be added (addition BDP) or replaced (replacement BDP) in order to destroy completely the estimator, i.e. to get an estimation

- at infinity (Hampel, 1971);
- at the bounds of the support of the estimator (He and Simpson, 1992);
- which is restricted to a finite set while it could lie in an infinite set without contamination (Genton and Lucas, 2003);
- ...

BDP of the error rate of the generalized 2-means

Let the observation x of the training sample ($F_{\varepsilon,x}$) tend to infinity

$\Rightarrow$ It becomes the center of a cluster with this observation only even if $\Omega(x) = x$ (García-Escudero and Gordaliza, 1999);

$\Rightarrow$ One entire group of the test sample (F) is badly classified while the other is well classified;

$\Rightarrow$ $\text{ER}(F_\varepsilon, F) = \pi_1$ or $\text{ER}(F_\varepsilon, F) = \pi_2$ for any sample;

$\Rightarrow$ ER has broken down in the sense of Genton and Lucas (2003);

$\Rightarrow$ The BDP of the ER is $1/n$ which tends to zero as $n \to \infty$.

BDP of the error rate of the generalized 2-means

Let the observation x of the training sample ($F_{\varepsilon,x}$) tend to infinity

- $\Rightarrow$ It becomes the center of a cluster with this observation only even if $\Omega(x) = x$ (García-Escudero and Gordaliza, 1999);
- $\Rightarrow$ One entire group of the test sample (F) is badly classified while the other is well classified;
- $\Rightarrow$ $\text{ER}(F_\varepsilon, F) = \pi_1$ or $\text{ER}(F_\varepsilon, F) = \pi_2$ for any sample;
- $\Rightarrow$ ER has broken down in the sense of Genton and Lucas (2003);
- $\Rightarrow$ The BDP of the ER is $1/n$ which tends to zero as $n \to \infty$.

BDP of the error rate of the generalized 2-means

Let the observation x of the training sample ($F_{\varepsilon,x}$) tend to infinity

- $\Rightarrow$ It becomes the center of a cluster with this observation only even if $\Omega(x) = x$ (García-Escudero and Gordaliza, 1999);
- $\Rightarrow$ One entire group of the test sample (F) is badly classified while the other is well classified;
- $\Rightarrow$ $\mathrm{ER}(F_\varepsilon, F) = \pi_1$ or $\mathrm{ER}(F_\varepsilon, F) = \pi_2$ for any sample;
- $\Rightarrow$ ER has broken down in the sense of Genton and Lucas (2003);
- $\Rightarrow$ The BDP of the ER is $1/n$ which tends to zero as $n \to \infty$.

BDP of the error rate of the generalized 2-means

Let the observation x of the training sample ($F_{\varepsilon,x}$) tend to infinity

- $\Rightarrow$ It becomes the center of a cluster with this observation only even if $\Omega(x) = x$ (García-Escudero and Gordaliza, 1999);
- $\Rightarrow$ One entire group of the test sample (F) is badly classified while the other is well classified;
- $\Rightarrow$ $\mathrm{ER}(F_\varepsilon, F) = \pi_1$ or $\mathrm{ER}(F_\varepsilon, F) = \pi_2$ for any sample;
- $\Rightarrow$ ER has broken down in the sense of Genton and Lucas (2003);
- $\Rightarrow$ The BDP of the ER is $1/n$ which tends to zero as $n \to \infty$.

BDP of the error rate of the generalized 2-means

Let the observation x of the training sample $(F_{\varepsilon,x})$ tend to infinity

- $\Rightarrow$ It becomes the center of a cluster with this observation only even if $\Omega(x) = x$ (García-Escudero and Gordaliza, 1999);
- $\Rightarrow$ One entire group of the test sample (F) is badly classified while the other is well classified;
- $\Rightarrow$ $\mathrm{ER}(F_\varepsilon, F) = \pi_1$ or $\mathrm{ER}(F_\varepsilon, F) = \pi_2$ for any sample;
- $\Rightarrow$ ER has broken down in the sense of Genton and Lucas (2003);
- $\Rightarrow$ The BDP of the ER is $1/n$ which tends to zero as $n \to \infty$.

BDP of the error rate of the generalized 2-means

Let the observation x of the training sample ($F_{\varepsilon,x}$) tend to infinity

- $\Rightarrow$ It becomes the center of a cluster with this observation only even if $\Omega(x) = x$ (García-Escudero and Gordaliza, 1999);
- $\Rightarrow$ One entire group of the test sample (F) is badly classified while the other is well classified;
- $\Rightarrow$ $\mathrm{ER}(F_\varepsilon, F) = \pi_1$ or $\mathrm{ER}(F_\varepsilon, F) = \pi_2$ for any sample;
- $\Rightarrow$ ER has broken down in the sense of Genton and Lucas (2003);
- $\Rightarrow$ The BDP of the ER is $1/n$ which tends to zero as $n \to \infty$.

Outline

The trimming approach

Idea: Delete extreme observations!

Problem: How can we detect extreme observations?

Solution: Impartial trimming

- k the fixed number of clusters;
- $\alpha \in [0, 1[$ the trimming size;
- $X_n = \{x_1, \ldots, x_n\} \in \mathbb{R}^p$ a dataset that is not concentrated on k points after removing a mass equal to α;
- Optimization over partitions $\mathcal{R} = \{R_1, \ldots, R_k\}$ of $\{1, \ldots, n\}$ with $\lceil n(1-\alpha) \rceil$ observations;

- The clusters centers $(T_1^n, \ldots, T_k^n)$ are solutions of the double minimization problem

$$\min_{\mathcal{R}} \min_{\{t_1,\ldots,t_k\}\subset\mathbb{R}^p} \sum_{x_i\in\mathcal{R}} \Omega\left(\inf_{1\leq j\leq k} \|x_i - t_j\|\right);$$

- The classification rule:

$$x \in C_j^n \Leftrightarrow \begin{cases} \|x - T_j^n\| = \min_{1\leq i\leq k} \|x - T_i^n\| \\ x \in \mathcal{R} \end{cases}$$

The generalized trimmed k-means method

Cuesta-Albertos *et al.*, 1997

- The clusters centers $(T_1^n, \ldots, T_k^n)$ are solutions of the double minimization problem

$$\min_{\mathcal{R}} \min_{\{t_1,\ldots,t_k\}\subset\mathbb{R}^p} \sum_{x_i\in\mathcal{R}} \Omega\left(\inf_{1\leq j\leq k} \|x_i - t_j\|\right);$$

- The classification rule:

$$x \in C_j^n \Leftrightarrow \begin{cases} \|x - T_j^n\| = \min_{1\leq i\leq k} \|x - T_i^n\| \\ x \in \mathcal{R} \end{cases}$$

Properties of the generalized trimmed 2-means

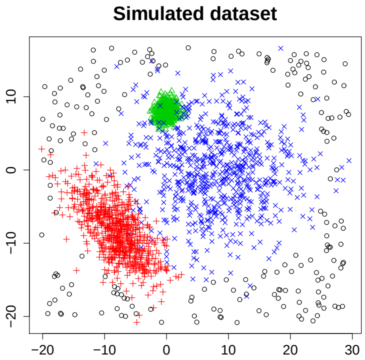
García-Escudero and Gordaliza, 1999

- Bounded IF whatever Ω;
- Better breakdown behavior.

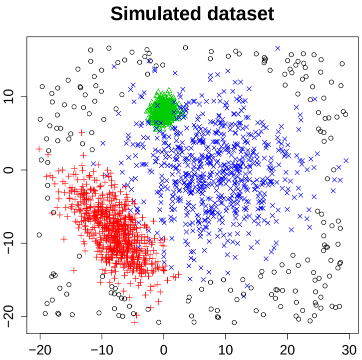
Ruwet and Haesbroeck (unpublished result)

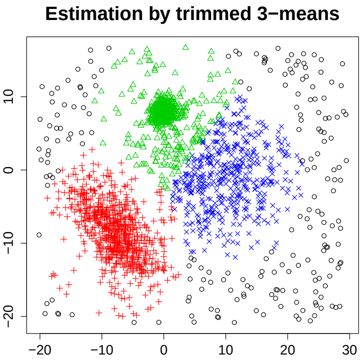
- If the Conjecture also holds for the generalized trimmed 2-means, this procedure is optimal under the model F_O.

Problem of all "*k*-means type" procedures



Problem of all "*k*-means type" procedures





- Optimization also over the scatter matrices S_j^n and the weights p_j^n such that $\sum_{j=1}^k p_j = 1$;
- Maximization of

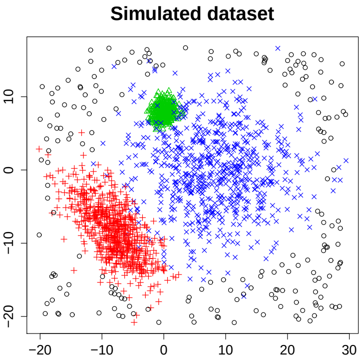
$$\sum_{j=1}^{k} \sum_{i \in R_j} \log \left(p_j \varphi \left(x_i; T_j, S_j\right)\right)$$

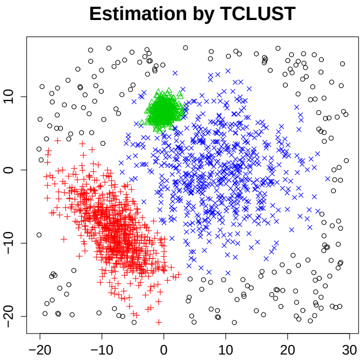
where φ is the pdf of the Gaussian distribution;
- Eigenvalues-ratio restriction:

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

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

Improvement with the TCLUST procedure





Current and future work

Robustness properties of the TCLUST procedure:

- The influence function (Ruwet *et al.*, Submitted)
- The breakdown behavior
- ???

Bibliography

- Anderson T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, Wiley, New-York (pp. 126-133).
- Croux C., Filzmoser P., and Joossens K. (2008), Classification efficiencies for robust linear discriminant analysis, *Statistica Sinica* 18, 581-599.
- Cuesta-Albertos J.A., Gordaliza A., and Matrán C. (1997), Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics* 25, 553-576.
- García-Escudero L.A., and Gordaliza A. (1999), Robustness Properties of k Means and Trimmed k Means, *Journal of the American Statistical Association* 94, 956-969.
- García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar A. (2008), A general trimming approach to robust cluster analysis. *The Annals of Statistics* 36, 1324-1345.

Bibliography

- Genton M.G., and Lucas A. (2003), Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of the Royal Statistical Society Series B* 65, 81-94.
- Hampel F.R. (1971), A general qualitative definition of robustness. *The Annals of Statistics* 42, 1887-1896.
- Hampel F.R., Ronchetti E.M., Rousseeuw P.J., and Stahel W.A. (1986), Robust Statistics : The Approach Based on Influence Functions, John Wiley and Sons, New-York.
- He X., and Simpson D.G. (1992), Robust direction estimation. *The Annals of Statistics* 20, 351-369.
- Kurata H., and Qiu D. (2011), Linear subspace spanned by principal points of a mixture of spherically symmetric distributions, *Communication s in Statistics - Theory and Methods* 40, 2737-2750.

Bibliography

- Ruwet C., and Haesbroeck G. (2011), Impact of contamination on training and test error rates in statistical clustering. *Communications in Statistics - Simulation and Computation*, 40, 394-411.
- Ruwet C., and Haesbroeck G. (201x), Classification performance resulting from a 2-means, *Submitted* (under revision).
- Ruwet C., Gordaliza A., García-Escudero L.A., and Mayo-Iscar A. (201x), The influence function of the TCLUST procedure, *Submitted*.