Robustness of classification based on clustering

Ch. Ruwet

DEPARTMENT OF MATHEMATICS - UNIVERSITY OF LIÈGE

September 2nd 2011
Robustness of classification based on clustering

Ch. Ruwet

Introduction

Statistical functionals

Error Rate

Influence functions

Asymptotic Loss

Breakdown point

Some improvements
1 Introduction

2 Statistical functionals

3 Error Rate

4 Influence functions

5 Asymptotic Loss

6 Breakdown point

7 Some improvements
Robustness of classification based on clustering

Ch. Ruwet

Introduction

Statistical functionals

Error Rate

Influence functions

Asymptotic Loss

Breakdown point

Some improvements

Classification

![Classification Diagram](image)

- **Succès**
- **Echec**

Legend: Success and Failure scatter plot.
Robustness of classification based on clustering

Ch. Ruwet

Introduction

Statistical functionals

Error Rate

Influence functions

Asymptotic Loss

Breakdown point

Some improvements

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:
    $$x \in C_j^n \iff \|x - T_j^n\| = \min_{1 \leq i \leq k} \|x - T_i^n\|;$$
- 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:
    \[
    x \in C_j^n \iff \|x - T_j^n\| = \min_{1 \leq i \leq k} \|x - T_i^n\|;
    \]
- 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:
    \[
    x \in C_j^n \iff \| x - T_j^n \| = \min_{1 \leq i \leq k} \| x - T_i^n \|;
    \]
- Let us focus on \( k = 2 \) groups:
  \[
  C_1^n = \left\{ x \in \mathbb{R}^p : (T_1^n - T_2^n)^t x - \frac{1}{2} (\| T_1^n \|^2 - \| T_2^n \|^2 > 0) \right\}.
  \]
Robustness of classification based on clustering
Ch. Ruwet

Introduction
Statistical functionals
Error Rate
Influence functions
Asymptotic Loss
Breakdown point
Some improvements

Result of 2-means
Introduction of contamination

Algèbre

Analyse

Robustness of classification based on clustering

Ch. Ruwet

Introduction

Statistical functionals

Error Rate

Influence functions

Asymptotic Loss

Breakdown point

Some improvements
Robustness of classification based on clustering

Ch. Ruwet

Introduction
Statistical functionals
Error Rate
Influence functions
Asymptotic Loss
Breakdown point
Some improvements

Result of 2-means with contamination
The generalized 2-means clustering method

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

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

for an increasing penalty function \(\Omega : \mathbb{R}^+ \rightarrow \mathbb{R}^+\) such that \(\Omega(0) = 0\).

- Classical penalty functions:
  \[
  \Omega(x) = x^2 \rightarrow \text{2-means method}
  \]
  \[
  \Omega(x) = x \rightarrow \text{2-medoids method}
  \]

- The classification rule:
  \[
  x \in C_1^n \iff \Omega(\|x - T_1^n\|) \leq \Omega(\|x - T_2^n\|)
  \]
  \[
  \iff \|x - T_1^n\| \leq \|x - T_2^n\|.
  \]
Classical penalty functions

- Robustness of classification based on clustering
- Ch. Ruwet

Introduction
- Statistical functionals
- Error Rate
- Influence functions
- Asymptotic Loss
- Breakdown point
- Some improvements

Graph showing the classical penalty functions for 2-means and 2-medoids.
Robustness of classification based on clustering

Ch. Ruwet

Introduction
Statistical functionals
Error Rate
Influence functions
Asymptotic Loss
Breakdown point
Some improvements

Result of 2-medoids
Result of 2-medoids with contamination

Robustness of classification based on clustering
Ch. Ruwet

Introduction
Statistical functionals
Error Rate
Influence functions
Asymptotic Loss
Breakdown point
Some improvements
1. Introduction

2. Statistical functionals

3. Error Rate

4. Influence functions

5. Asymptotic Loss

6. Breakdown point

7. Some improvements
Statistical functionals

- The empirical distribution $F_n$ is replaced by a cumulative distribution $F \in \mathcal{F}$;
- A statistical functional $T : \mathcal{F} \rightarrow \mathbb{R}^l : F \mapsto T(F)$ such that $T(F_n) = T^n$. 

![Graph showing empirical distribution $F_n$ versus theoretical distribution $F$]
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

\[ F \text{ is a mixture of two distributions} \]

\[ F = \pi_1(F)F_1 + \pi_2(F)F_2 \]

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 \( f_2 \).
Robustness of classification based on clustering

Ch. Ruwet

Introduction

Statistical functionals

Error Rate

Influence functions

Asymptotic Loss

Breakdown point

Some improvements

Univariate mixture distribution

![Graph showing two normal distributions with means 0.4 and 0.6]
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 \(\Omega\);

- The classification rule is

\[
R_F : x \mapsto j \iff \|x - T_j(F)\| = \min_{1 \leq i \leq 2} \|x - T_i(F)\|
\]

- The clusters are

\[
C_1(F) = \{ x \in \mathbb{R}^p : A(F)^t x + b(F) > 0 \}
\]

\[
C_2(F) = \mathbb{R}^p \setminus C_1(F)
\]

with \(A(F) = T_1(F) - T_2(F)\)

and \(b(F) = -\frac{1}{2} (\|T_1(F)\|^2 - \|T_2(F)\|^2)\).
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: \( g(r) = \exp\left( -\frac{r}{2} \right) \)
- 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):

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

with $\mu = \mu_1 e_1$ and $\mu_1 > 0$. 
Robustness of classification based on clustering

Ch. Ruwet

Introduction

Statistical functionals

Error Rate

Influence functions

Asymptotic Loss

Breakdown point

Some improvements

Representation
Robustness of classification based on clustering

Ch. Ruwet

Introduction
Statistical functionals
Error Rate
Influence functions
Asymptotic Loss
Breakdown point
Some improvements

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.
Robustness of classification based on clustering

Ch. Ruwet

Introduction

Statistical functionals

Error Rate

Influence functions

Asymptotic Loss

Breakdown point

Some improvements
Robustness of classification based on clustering

Ch. Ruwet

Introduction
Statistical functionals
Error Rate
Influence functions
Asymptotic Loss
Breakdown point
Some improvements

Error rate of the 2-means

![Graph showing error rate of the 2-means with points representing success and failure.](Image)
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 $F_m$:

$$ER(F, F_m) = \pi_1(F_m)P_{F_m}[R_F(X) \neq 1 | G_1]$$
$$+ \pi_2(F_m)P_{F_m}[R_F(X) \neq 2 | G_2]$$
$$= \sum_{j=1}^{2} \pi_j(F_m)P_{F_m}[R_F(X) \neq j | G_j]$$
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 $F_m$:

$$
\text{ER}(F, F_m) = \pi_1(F_m) \mathbb{P}_{F_m} [R_F(X) \neq 1 \mid G_1]
+ \pi_2(F_m) \mathbb{P}_{F_m} [R_F(X) \neq 2 \mid G_2]
= \sum_{j=1}^{2} \pi_j(F_m) \mathbb{P}_{F_m} [R_F(X) \neq j \mid G_j]
$$
A classification rule is optimal if the corresponding error rate is minimal;
The optimal classification rule is the Bayes rule:

\[ x \in C_1(F) \iff \pi_1(F)f_1(x) > \pi_2(F)f_2(x) \]

Optimality in classification

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

\[ x \in C_1(F) \iff \pi_1(F)f_1(x) > \pi_2(F)f_2(x) \]


**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 \( F_O \).
A contaminated distribution is defined by

\[ F_\varepsilon \]

\[ 1 - \varepsilon : F \]

\[ \varepsilon : G \]

where \( G \) is a arbitrary distribution function.
Contaminated distribution

A contaminated distribution is defined by

\[ F_\varepsilon \]

1 \( - \varepsilon \) : \( F \)

\[ \varepsilon \) : \( G \]

where \( G \) is an 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

\[ F_{\varepsilon, x} \]

\[ \begin{align*}
(1-0.05) & \quad 0.4 \\
0.05 & \\
0.05 & \\
(1-0.05) & \quad 0.6
\end{align*} \]
Error rate under contamination

- Contaminated training sample according to $F_\epsilon$: estimation of the rule
- Test sample according to $F_m$: evaluation of the rule

$$\text{ER}(F_\epsilon, F_m) = \sum_{j=1}^{2} \pi_j(F_m) \mathbb{P}_{F_m} [R_{F_\epsilon}(X) \neq j \mid G_j]$$
Robustness of classification based on clustering

Ch. Ruwet

Outline

1. Introduction

2. Statistical functionals

3. Error Rate

4. Influence functions

5. Asymptotic Loss

6. Breakdown point

7. Some improvements
Hampel et al. (1986) : For any statistical functional \( T \) and any distribution \( F \),

\[
\text{IF}(x; T, F) = \lim_{\varepsilon \to 0} \frac{T((1 - \varepsilon)F + \varepsilon \Delta_x) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T((1 - \varepsilon)F + \varepsilon \Delta_x) \bigg|_{\varepsilon=0}
\]

(under condition of existence);

\[
E_F[\text{IF}(X; T, F)] = 0;
\]

First order Taylor expansion of \( T \) at \( F \):

\[
T(F_{\varepsilon,x}) \approx T(F) + \varepsilon \text{IF}(x; T, F)
\]

for \( \varepsilon \) small enough.
Definition and properties of the first order influence function

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

$$\text{IF}(x; T, F) = \lim_{\varepsilon \to 0} \frac{T((1 - \varepsilon)F + \varepsilon \Delta x) - T(F)}{\varepsilon}$$

$$= \left. \frac{\partial}{\partial \varepsilon} T((1 - \varepsilon)F + \varepsilon \Delta x) \right|_{\varepsilon=0}$$

(under condition of existence);

- $E_F[\text{IF}(X; T, F)] = 0$;

- First order Taylor expansion of $T$ at $F$:

$$T(F_{\varepsilon,x}) \approx T(F) + \varepsilon \text{IF}(x; T, F)$$

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

If the model distribution is $F_O$,

- $\text{ER}(F_{\varepsilon,x}, F_O) \approx \text{ER}(F_O, F_O) + \varepsilon \text{IF}(x; \text{ER}, F_O)$
- $\text{ER}(F_{\varepsilon,x}, F_O) \geq \text{ER}(F_O, F_O)$
- $E_{F_O}[\text{IF}(X; \text{ER}, F_O)] = 0$
First order influence function of the error rate

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

If the model distribution is $F_O$,

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

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

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 $F$,

$$\text{IF2}(x; T, F_O) = \left. \frac{\partial^2}{\partial \varepsilon^2} T((1 - \varepsilon)F_O + \varepsilon \Delta x) \right|_{\varepsilon=0}$$

(under condition of existence).
For any statistical functional $T$ and any distribution $F$,

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

(under condition of existence).

Second order Taylor expansion of $ER$ at $F_O$ :

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

for $\varepsilon$ 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

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

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

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

where $\omega_i(x) = -\text{grad}_y\Omega(\|y\|)|_{y=x-T_i(F_m)} I(x \in C_i(F_m))$ 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 $\Omega$ 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

$$\text{IF}_2(x; \text{ER}, F_O) = -\frac{1}{4} f'_{-\mu,\sigma^2}(0) \left( \text{IF}(x; T_1, F_O) + \text{IF}(x; T_2, F_O) \right)^2$$

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 $\text{IF}(x; \text{ER}, F_M)$
Robustness of classification based on clustering
Ch. Ruwet

Introduction
Statistical functionals
Error Rate
Influence functions
Asymptotic Loss
Breakdown point
Some improvements

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

2–means

\[ \pi_1 = 0.2 \]
\[ \pi_1 = 0.4 \]
\[ \pi_1 = 0.6 \]
\[ \pi_1 = 0.8 \]
Robustness of classification based on clustering

Ch. Ruwet

Introduction
Statistical functionals
Error Rate
Influence functions
Asymptotic Loss
Breakdown point
Some improvements

Graph of $\text{IF}(x; \text{ER}, F_M)$
Robustness of classification based on clustering

Ch. Ruwet

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

- $2$–means
- $2$–medoids
Robustness of classification based on clustering

Ch. Ruwet

Outline

1. Introduction
2. Statistical functionals
3. Error Rate
4. Influence functions
5. Asymptotic Loss
6. Breakdown point
7. Some improvements
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

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

As in Croux et al. (2008):

Proposition

*Under some regularity conditions of the clusters centers estimators,*

\[
A\text{-Loss} = \frac{1}{2} E_{F_O}[IF2(X; ER, F_O)]
\]
Proposition (Ruwet and Haesbroeck, 2011)

Under an optimal mixture of normal distributions, \( F_N \), with \( \mu = \frac{\Delta}{2} e_1 \), the asymptotic loss of the generalized 2-means procedure is given by

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

where \( \text{ASV} \) and \( \text{ASC} \) 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

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

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

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

\[
\text{ARCE}(\text{Method 1, Method 2}) = \frac{\text{A-Loss(Method 2)}}{\text{A-Loss(Method 1)}}.
\]
ARCE of 2-medoids w.r.t. 2-means
ARCE of classification procedures w.r.t. 2-means
Robustness of classification based on clustering

Ch. Ruwet

Introduction

1 Introduction

2 Statistical functionals

3 Error Rate

4 Influence functions

5 Asymptotic Loss

6 Breakdown point

7 Some improvements
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);
- ...
Let the observation $x$ of the training sample $(F_\varepsilon, x)$ tend to infinity

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

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

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

⇒ $\text{ER}$ has broken down in the sense of Genton and Lucas (2003);

⇒ 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;

$\Rightarrow$ $\text{ER}$ 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

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

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

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

⇒ ER has broken down in the sense of Genton and Lucas (2003);

⇒ The BDP of the ER is $1/n$ which tends to zero as $n \to \infty$. 
Let the observation $x$ of the training sample $(F_\varepsilon, x)$ tend to infinity

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

$\Rightarrow$ $\text{ER}$ 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;

$\Rightarrow$ $\text{ER}$ 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 $n \to \infty$. 
Robustness of classification based on clustering

Ch. Ruwet

Outline

1 Introduction
2 Statistical functionals
3 Error Rate
4 Influence functions
5 Asymptotic Loss
6 Breakdown point
7 Some improvements
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 $\alpha$;
- Optimization over partitions $\mathcal{R} = \{R_1, \ldots, R_k\}$ of $\{1, \ldots, n\}$ with $\lceil n(1 - \alpha) \rceil$ observations;
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:

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

$x \in \mathcal{R}$
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:

$$
x \in C_j^n \iff \begin{cases} 
\|x - T_j^n\| = \min_{1 \leq i \leq k} \|x - T_i^n\| \\
x \in \mathcal{R}
\end{cases}
$$
García-Escudero and Gordaliza, 1999

- Bounded IF whatever $\Omega$;
- 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

Simulated dataset
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

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

where $\varphi$ 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


Bibliography


Bibliography


- Ruwet C., and Haesbroeck G. (201x), Classification performance resulting from a 2-means, *Submitted* (under revision).