Influence function of the error rate of classification based on clustering

Joint work with G. Haesbroeck

Ch. Ruwet

Department of Mathematics - University of Liège

19 May 2009

cruwet@ulg.ac.be

Outline

Outline

Classification setting

Suppose

$X \sim F$ arises from G_1 and G_2 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 density $f = \pi_1(F)f_1 + \pi_2(F)f_2$.

Additional assumption : one dimension !

The generalized 2-means clustering method

- Aim of clustering : Find estimations $C_1(F)$ and $C_2(F)$ (called clusters) of the two underlying groups.
- The clusters' centers $(T_1(F), T_2(F))$ are solutions of

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

for a suitable nondecreasing penalty function $\Omega : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$.

- Classical penalty functions :

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

Classification rule

- The classification rule is

$$R_F(x) = C_j(F) \Leftrightarrow \Omega(\|x - T_j(F)\|) = \min_{1 \leq i \leq 2} \Omega(\|x - T_i(F)\|)$$

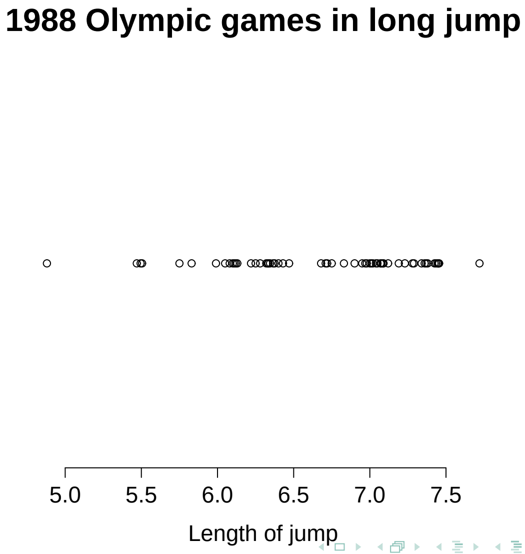
- In one dimension, the clusters are simply :

$$C_1(F) =]-\infty, C(F)[$$

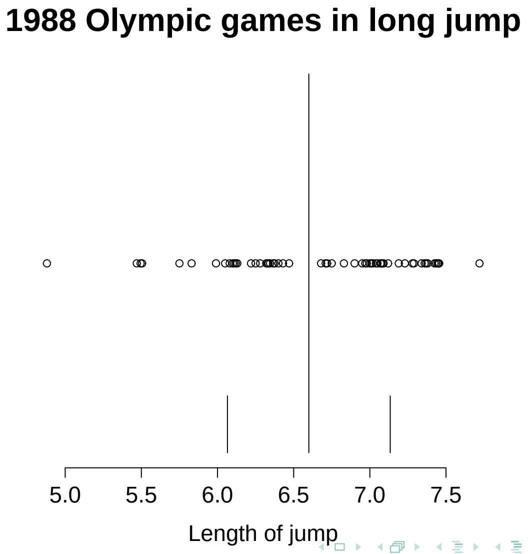
$$C_2(F) =]C(F), +\infty[$$

where $C(F) = \frac{T_1(F) + T_2(F)}{2}$ is the cut-off point.

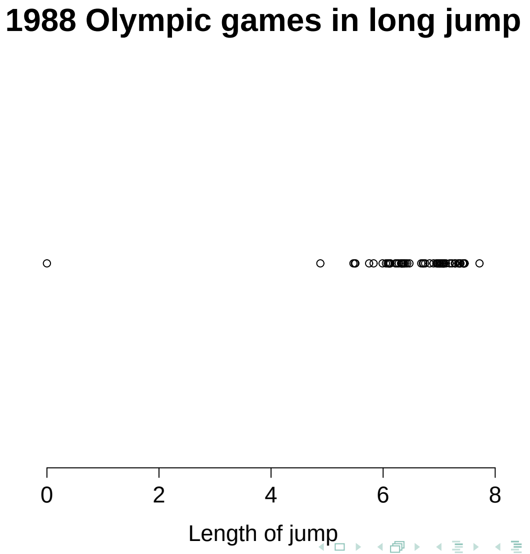
Example (from Hand et al., 1991)



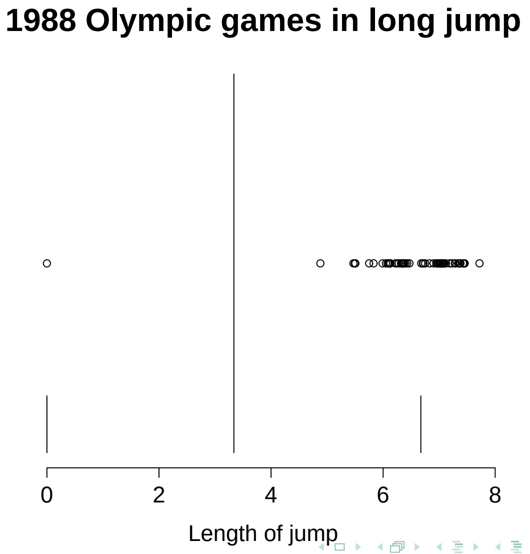
Example


1988 Olympic games in long jump
5.0
5.5
6.0
6.5
7.0
7.5
Length of jump

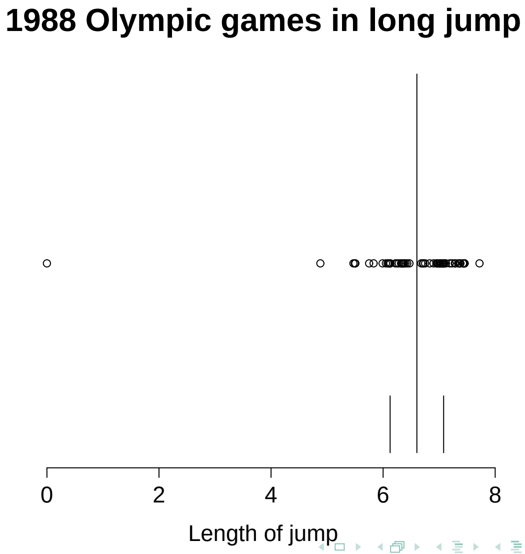
Contaminated example



Contaminated example with the 2-means

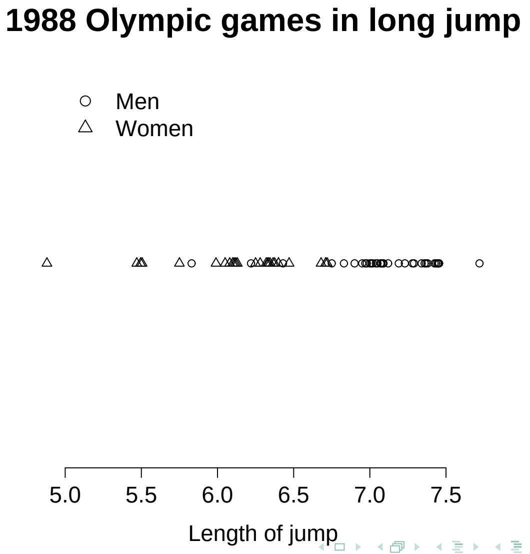


Contaminated example with the 2-medoids

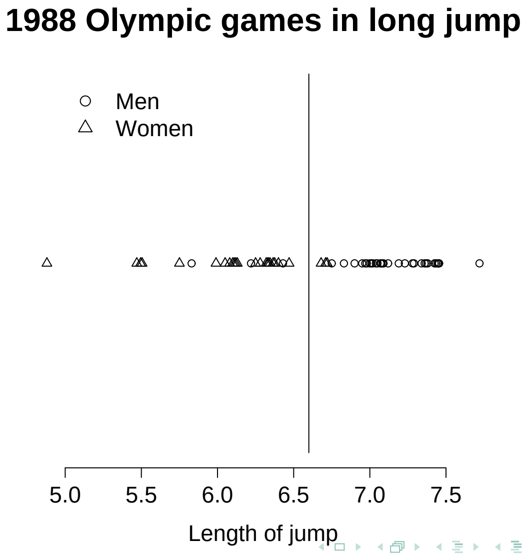


Outline

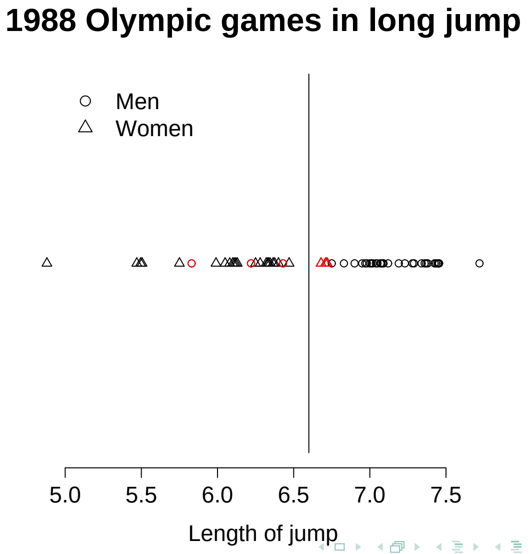
Example


1988 Olympic games in long jump
Men
Women
5.0
5.5
6.0
6.5
7.0
7.5
Length of jump

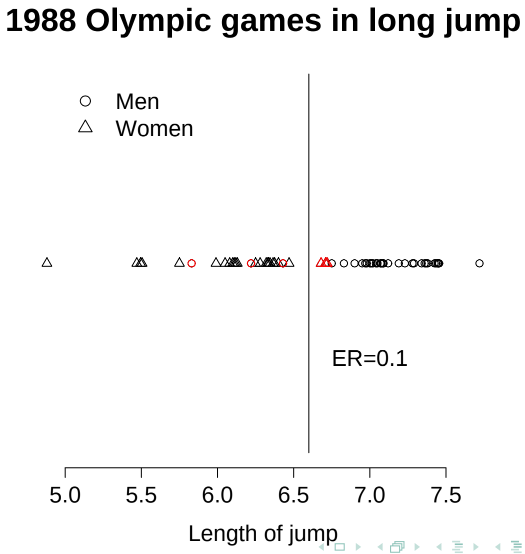
Example with the 2-means



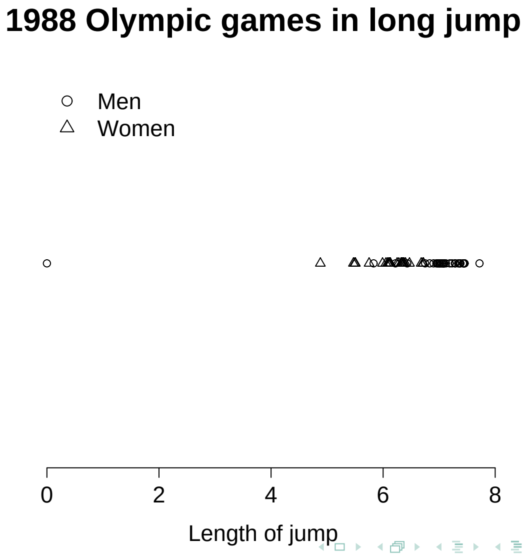
Example with the 2-means



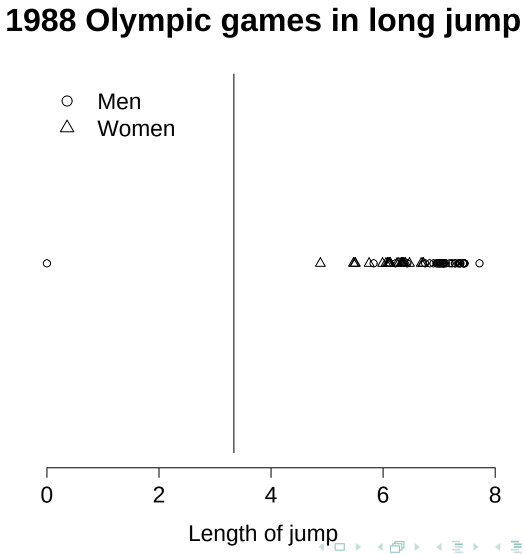
Example with the 2-means



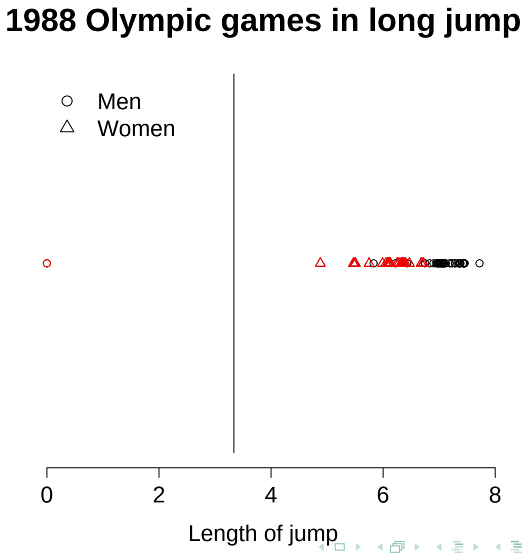
Contaminated example



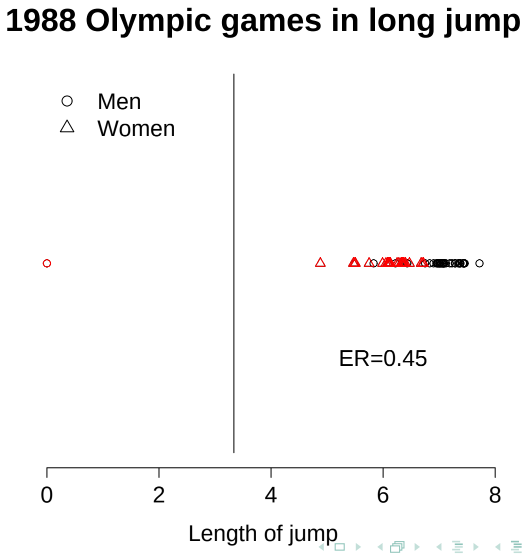
Contaminated example with the 2-means



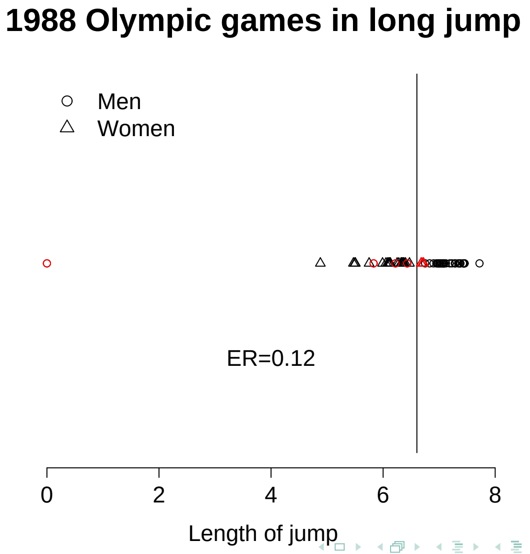
Contaminated example with the 2-means



Contaminated example with the 2-means



Contaminated example with the 2-medoids



Error rate

- Training sample according to F : estimation of the rule
- Test sample according to F_m : evaluation of the rule
- In ideal circumstances : $F = F_m$

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

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)

- The 2-means procedure is optimal under the model

$$F_N = 0.5\, N(\mu_1, \sigma^2) + 0.5\, N(\mu_2, \sigma^2) \text{ with } \mu_1 < \mu_2$$

(Qiu and Tamhane, 2007)

Outline

Outline

Definition and properties of the first order influence function

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

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

Definition and properties of the first order influence function

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

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

First order influence function of the error rate

Now, the training sample is distributed as F_ε which is a contaminated mixture.

$$\begin{aligned} \mathrm{ER}(F_\varepsilon, F_m) &= \sum_{j=1}^{2} \pi_j(F_m) \mathbb{P}_{F_m}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) \middle| G_j\right] \\ &= \pi_1(F_m)\{1 - F_{m,1}(C(F_\varepsilon))\} + \pi_2(F_m) F_{m,2}(C(F_\varepsilon)) \end{aligned}$$

First order influence function of the error rate

Now, the training sample is distributed as F_ε which is a contaminated mixture.

$$\begin{aligned}\mathrm{ER}(F_\varepsilon, F_m) &= \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) | \, G_j\right] \\ &= \pi_1(F_m)\{1 - F_{m,1}(C(F_\varepsilon))\} + \pi_2(F_m)F_{m,2}(C(F_\varepsilon))\end{aligned}$$

- $\mathrm{ER}(F_\varepsilon, F_N) \approx \mathrm{ER}(F_N, F_N) + \varepsilon \mathrm{IF}(x; \mathrm{ER}, F_N)$
- $\mathrm{ER}(F_\varepsilon, F_N) \geq \mathrm{ER}(F_N, F_N)$

First order influence function of the error rate

Now, the training sample is distributed as F_ε which is a contaminated mixture.

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

$$= \pi_1(F_m)\{1 - F_{m,1}(C(F_\varepsilon))\} + \pi_2(F_m)F_{m,2}(C(F_\varepsilon))$$

- $\mathrm{ER}(F_\varepsilon, F_N) \approx \mathrm{ER}(F_N, F_N) + \varepsilon \mathrm{IF}(x; \mathrm{ER}, F_N)$
- $\mathrm{ER}(F_\varepsilon, F_N) \geq \mathrm{ER}(F_N, F_N)$

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

First order influence function of the error rate

Proposition

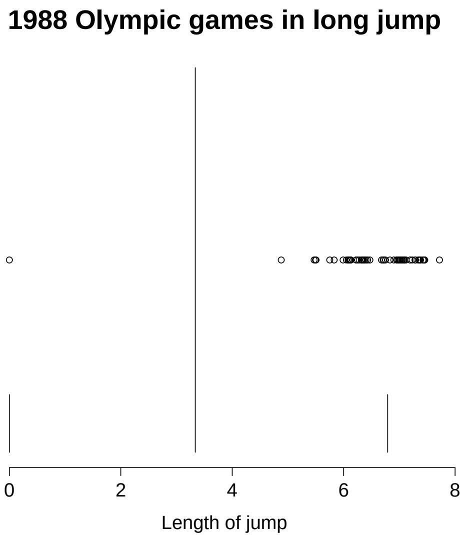
The influence function of the error rate of the generalized 2-means classification procedure is given by

$$\mathrm{IF}(x; \mathrm{ER}, F) = \frac{1}{2}\{\mathrm{IF}(x; T_1, F) + \mathrm{IF}(x; T_2, F)\}$$
$$\{\pi_2(F) f_2(C(F)) - \pi_1(F) f_1(C(F))\}$$

for all $x \neq C(F)$.

Expressions of $\mathrm{IF}(x; T_1, F)$ and $\mathrm{IF}(x; T_2, F)$ were computed by García-Escudero and Gordaliza (1999).

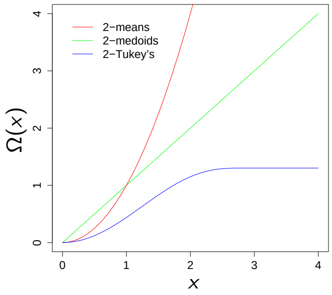
New contaminated example with the 2-medoids



Penalty functions

- 2-means : $\Omega(x) = x^2$
- 2-medoids : $\Omega(x) = x$
- 2-Tukey's : $\Omega(x) = \frac{b^2}{6} \begin{cases} 1 - \left[1 - \left(\frac{x}{b}\right)^2\right]^3 & \text{if } |x| \leq b \\ 1 & \text{if } |x| > b \end{cases}$

  with $b = 2.795$



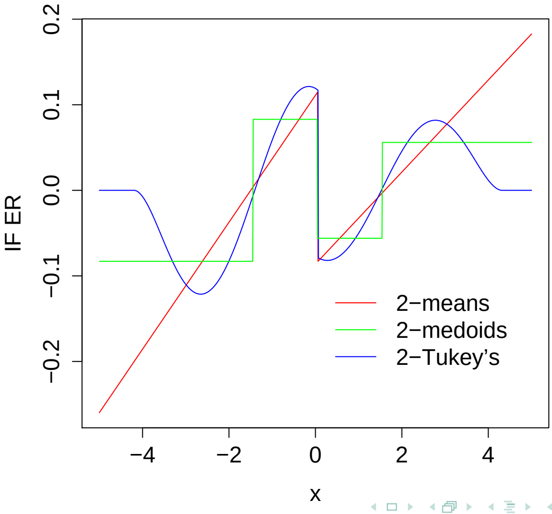
Model distributions

$$F = \pi_1 \, N(-\Delta/2, 1) + (1 - \pi_1) \, N(\Delta/2, 1)$$

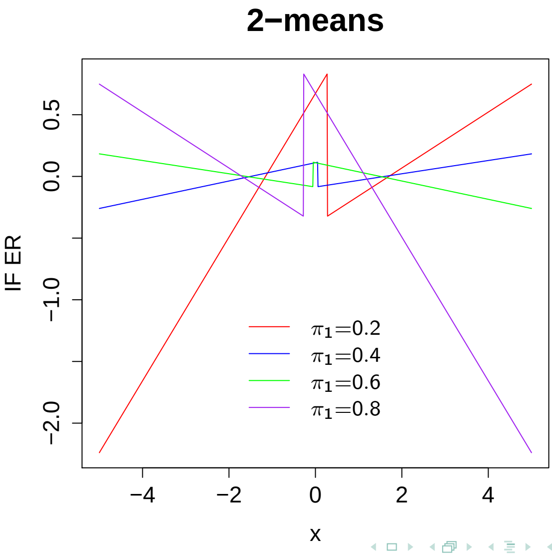
with

- $\pi_1 = 0.4$ and $\Delta = 3$
- π_1 is varying and $\Delta = 3$
- $\pi_1 = 0.4$ and Δ is varying

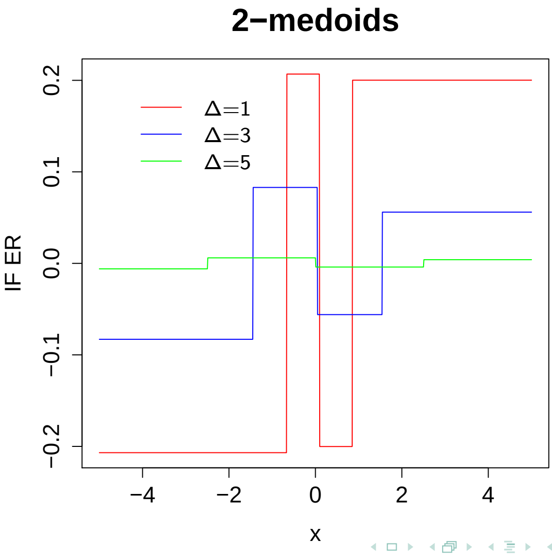
Graph of IF(x; ER, F)



Graph of IF(x; ER, F)



Graph of IF(x; ER, F)



Outline

Influence function of the error rate of classification based on clustering

Ch. Ruwet

Definition of the second order influence function

Under the model F_N, $\text{IF}(x; \text{ER}, F_N) \equiv 0$

One needs to go a step further !

Definition of the second order influence function

Under the model F_N, $\text{IF}(x; \text{ER}, F_N) \equiv 0$

One needs to go a step further !

For any statistical functional T and any distribution F,

$$\text{IF2}(x; \text{T}, F) = \left.\frac{\partial^2}{\partial \varepsilon^2} \text{T}(F_\varepsilon)\right|_{\varepsilon=0}$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon \Delta_x$ (under condition of existence).

Definition of the second order influence function

Under the model F_N, $\text{IF}(x; \text{ER}, F_N) \equiv 0$

One needs to go a step further !

For any statistical functional T and any distribution F,

$$\text{IF2}(x; \text{T}, F) = \left.\frac{\partial^2}{\partial \varepsilon^2} \text{T}(F_\varepsilon)\right|_{\varepsilon=0}$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$ (under condition of existence).
Second order von Mises expansion of ER at F_N :

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

for ε small enough.

Second order influence function of the error rate under optimality

Proposition

Under the optimal model F_N, the second order influence function of the error rate of the generalized 2-means classification procedure is given by

$$\mathrm{IF2}(x; \mathrm{ER}, F_N) = -\frac{1}{4}(f_{N1})' \left(\frac{\mu_1 + \mu_2}{2}\right) \{IF(x; T_1, F_N) + IF(x; T_2, F_N)\}^2$$

for all $x \neq \frac{\mu_1+\mu_2}{2}$. This expression is always positive.

Outline

Asymptotic loss

A measure of the expected increase in error rate when estimating the optimal clustering rule from a finite sample with empirical cdf F_n :

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

As in Croux et al. (2008) :

Proposition

Under some regularity conditions of the clusters'centers estimators,

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

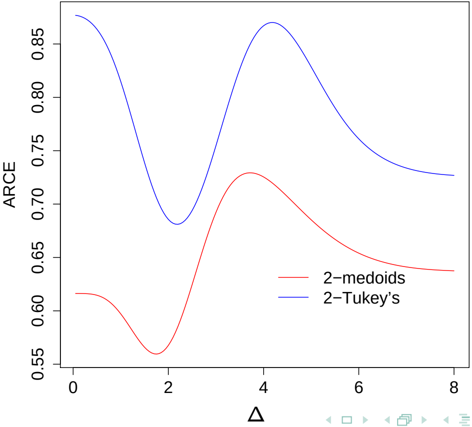
Asymptotic relative classification efficiencies

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 :

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

Graph of ARCE

$$F_N = 0.5\,N(-\Delta, 1) + 0.5\,N(\Delta, 1)$$



Outline

Outline

Conclusion

- The generalized 2-means procedure can give a more robust estimator of the error rate with a good choice of the penalty function;
- The price to pay is a loss in efficiency (depending also on the penalty function).

Outline

Future research

- Generalized trimmed 2-means procedure : for $\alpha \in [0,1]$, $(T_1(F), T_2(F))$ are solutions of

$$\min_{\{A: F(A)=1-\alpha\}} \min_{\{t_1,t_2\}\subset\mathbb{R}} \int_A \Omega\left(\inf_{1\leq j\leq 2} \|x-t_j\|\right) dF(x)$$

(Cuesta-Albertos, Gordaliza, and Matrán, 1997);
- Other robustness properties of the generalized 2-means method defined with nondecreasing penalty function instead of strictly increasing;
- More than 1 dimension and/or more than 2 groups.

Thank you for your attention!

Bibliography

- Anderson T.W., *An Introduction to Multivariate Statistical Analysis*, Wiley, New-York, 1958, pp. 126-133.
- Croux C., Filzmoser P., and Joossens K. (2008), Classification efficiences for robust linear discriminant analysis, *Statistica Sinica* 18, pp. 581-599.
- Croux C., Haesbroeck G., and Joossens K. (2008), Logistic discrimination using robust estimators : an influence function approach, *The Canadian Journal of Statistics*, 36, pp. 157-174.
- Fernholz L. T., On multivariate higher order von Mises expansions in *Metrika* 2001, vol. 53, pp. 123-140.

Bibliography

- García-Escudero L. A., and Gordaliza A., Robustness Properties of k Means and Trimmed k Means, *Journal of the American Statistical Association*, September 1999, Vol. 94, n° 447, pp. 956-969.
- Hampel F.R., Ronchetti E.M., Rousseeuw P.J., and Stahel W.A., Robust Statistics : The Approach Based on Influence Functions, John Wiley and Sons, New-York, 1986.
- Hand D.J., Daly F., Lunn A.D., McConway K.J., and Ostrowski E.,*A Handbook of Small Data Sets*, Chapman and Hall, London, 1991.

Bibliography

- Pollard D., Strong Consistency of k-Means Clustering, *The Annals of Probability*, 1981, Vol.9, nř4, pp.919-926.
- Pollard D., A Central Limit Theorem for k-Means Clustering, *The Annals of Probability*, 1982, Vol.10, nř1, pp.135-140.
- Qiu D. and Tamhane A. C. (2007), A comparative study of the *k*-means algorithm and the normal mixture model for clustering : Univariate case, *Journal of Statistical Planning and Inference*, 137, pp. 3722-3740.