Detection of influential observations on the error rate based on the generalized k-means clustering procedure

Joint work with G. Haesbroeck

Ch. Ruwet

Department of Mathematics - University of Liège

14 October 2009

cruwet@ulg.ac.be

Outline

Outline

Statistical clustering

Let

$X \sim F$ 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 $f = \pi_1(F)f_1 + \pi_2(F)f_2$.

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

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

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

- Classical penalty functions :

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

$$\Omega(x) = x \rightarrow \text{ 2-medoids method}$$

Classification rule

- The classification rule is

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

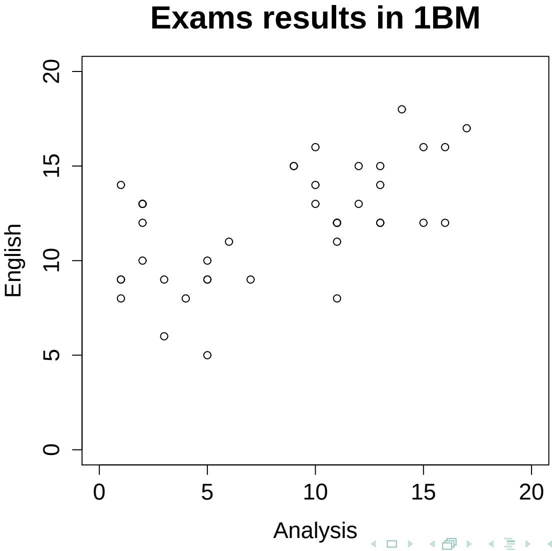
- The clusters are half spaces delimited by the hyperplane :

$$\mathcal{C} = \left\{ x \in \mathbb{R}^p : A(F)^T x + b(F) = 0 \right\}$$

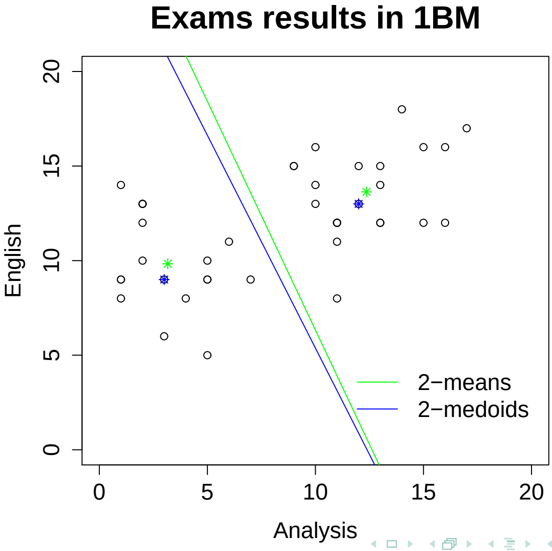
with

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

Example: Exams results in 1BM ($n = 40$)

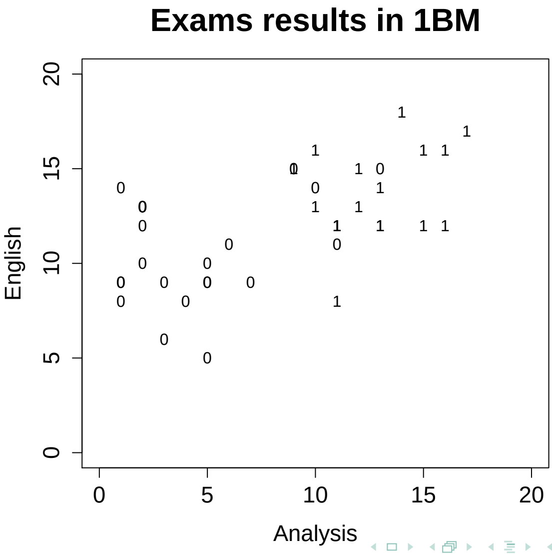


Example: Exams results in 1BM ($n = 40$)

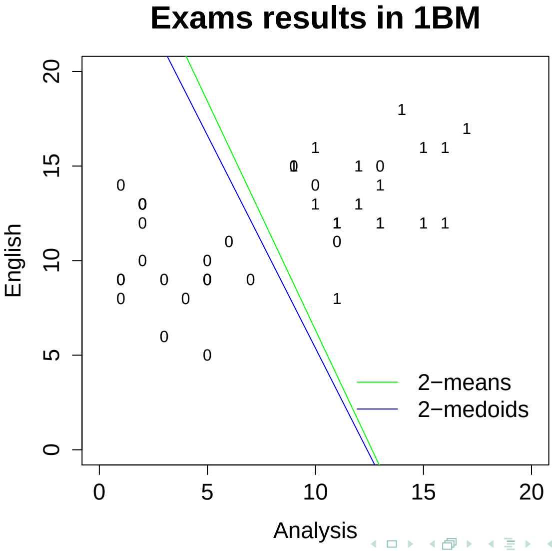


Outline

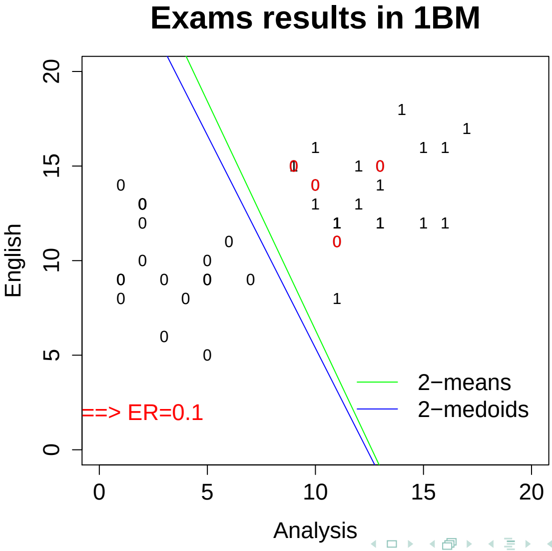
Example: Exams results in 1BM ($n = 40$)



Example: Exams results in 1BM ($n = 40$)



Example: Exams results in 1BM ($n = 40$)



Error rate (ER)

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

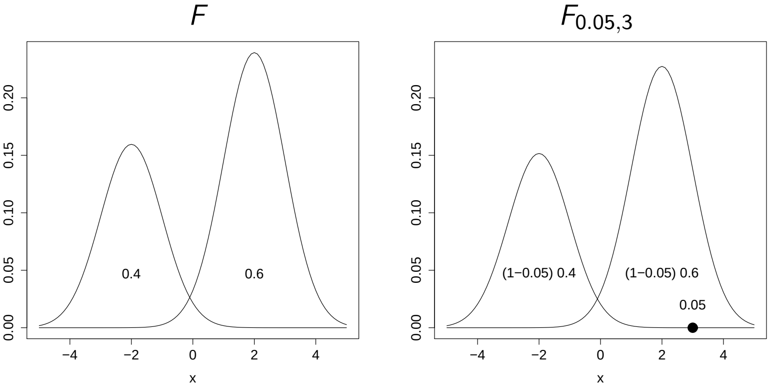
Error rate (ER)

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

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

Outline

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



Definition and properties of the IF

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 :
  $$\mathrm{T}((1-\varepsilon)F+\varepsilon\Delta_x)\approx \mathrm{T}(F)+\varepsilon\mathrm{IF}(x;\mathrm{T},F)$$
  for ε small enough.

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 :

  $$\mathrm{T}((1-\varepsilon)F+\varepsilon\Delta_x) \approx \mathrm{T}(F) + \varepsilon\mathrm{IF}(x;\mathrm{T},F)$$

  for ε small enough.

Influence function of the error rate (1)

Now, the training sample is distributed as $F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$ which is a contaminated mixture.

$$\begin{aligned}
\text{ER}(F_{\varepsilon,x}, F_m) &= \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_{F_{\varepsilon,x}}(X) \neq C_j(F_{\varepsilon,x}) \middle| G_j\right] \\
&= \pi_1(F_m)\mathbb{P}_{F_{m,1}}[X^T A(F_{\varepsilon,x}) + b(F_{\varepsilon,x}) < 0] \\
&\quad + \pi_2(F_m)\mathbb{P}_{F_{m,2}}[X^T A(F_{\varepsilon,x}) + b(F_{\varepsilon,x}) > 0]
\end{aligned}$$

Influence function of the error rate (1)

Now, the training sample is distributed as $F_{\varepsilon,x} = (1-\varepsilon)F + \varepsilon\Delta_x$ which is a contaminated mixture.

$$\begin{aligned}
\mathrm{ER}(F_{\varepsilon,x}, F_m) &= \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_{F_{\varepsilon,x}}(X) \neq C_j(F_{\varepsilon,x}) \middle| G_j\right] \\
&= \pi_1(F_m)\mathbb{P}_{F_{m,1}}[X^T A(F_{\varepsilon,x}) + b(F_{\varepsilon,x}) < 0] \\
&\quad + \pi_2(F_m)\mathbb{P}_{F_{m,2}}[X^T A(F_{\varepsilon,x}) + b(F_{\varepsilon,x}) > 0]
\end{aligned}$$

Taylor expansion : for ε small enough,

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

- $\mathrm{IF}(x; \mathrm{ER}, F_m) \geq 0 \Leftrightarrow \mathrm{ER}(F_{\varepsilon,x}, F_m) \geq \mathrm{ER}(F_m, F_m)$
- $\mathrm{IF}(x; \mathrm{ER}, F_m) \leq 0 \Leftrightarrow \mathrm{ER}(F_{\varepsilon,x}, F_m) \leq \mathrm{ER}(F_m, F_m)$

Influence function of the error rate (2)

Decomposition	$y = (y_1, y_2^T)^T$ with $y_1 \in \mathbb{R}$;
Notations	$T_i(F_m) = \tau_i$, $A(F_m) = (\alpha_1, \alpha_2^T)^T$ and $b(F_m) = \beta$;
Hypothesis	w.l.o.g., $-\tau_1 = \tau_2 = (t, 0, \ldots, 0)^T$ with $t > 0$;

Influence function of the error rate (3)

Proposition

Under these hypotheses and with these notations, the IF of the ER of the generalized 2-means classification procedure is given by

$$\text{IF}(x; \text{ER}, F_m) = \int \left(\frac{\text{IF}(x; b, F_m)}{\alpha_1} + y_2^T \frac{\text{IF}(x; A_2, F_m)}{\alpha_1} \right) [\pi_1 f_{m,1}(0, y_2) - \pi_2 f_{m,2}(0, y_2)] \, dy_2$$

with

$$\text{IF}(x; b, F_m) = t[\text{IF}(x; T_{21}, F_m) + \text{IF}(x; T_{11}, F_m)]$$
$$\text{IF}(x; A_2, F_m) = \text{IF}(x; T_{12}, F_m) - \text{IF}(x; T_{22}, F_m)$$

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

Influence functions of T_1 and T_2

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

where $\omega_i(x) = -\operatorname{grad}_y \Omega(\|y\|)\big|_{y=x-T_i(F)} \mathrm{I}(x \in C_i(F))$ and with M independent of x.

2-means method $\quad \omega_i(x) = -2(x - T_i(F))\mathrm{I}(x \in C_i(F));$

2-medoids method $\quad \omega_i(x) = -\dfrac{x - T_i(F)}{\|x - T_i(F)\|}\mathrm{I}(x \in C_i(F)).$

Example of a coefficient of the matrix M in 2-D

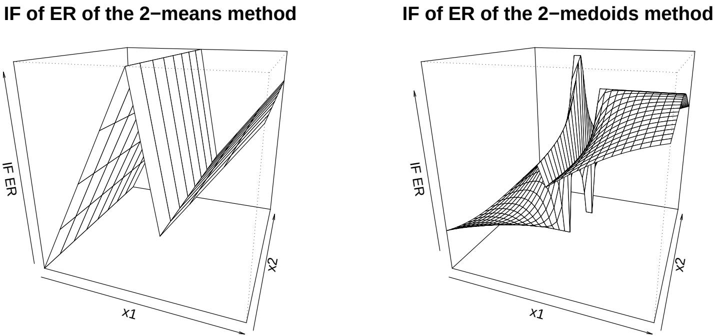
For the 2-means procedure, $M_{11} =$

$$\frac{2(T_{21}(F)-T_{11}(F))}{\|T_2(F)-T_1(F)\|^4}\Big(\int_{\{v\in\mathbb{R}^2:v_1<0\}}[(T_{21}(F)-T_{11}(F))v_1+(T_{12}(F)-T_{22}(F))v_2]$$
$$[f(c^v)+(c_1^v-T_{11}(F))f_1'(c^v)]dv$$
$$+\int_{\{v\in\mathbb{R}^2:v_1<0\}}[(T_{22}(F)-T_{12}(F))v_1+(T_{21}(F)-T_{11}(F))v_2](c_1^v-T_{11}(F))f_2'(c^v)dv\Big)$$
$$-\frac{1}{\|T_2(F)-T_1(F)\|^2}\int_{\{v\in\mathbb{R}^2:v_1<0\}}(v_1+1/2)[f(c^v)+(c_1^v-T_{11}(F))f_1'(c^v)]dv$$
$$-\int_{\{v\in\mathbb{R}^2:v_1<0\}} v_2(c_1^v-T_{11}(F))f_2'(c^v)dv-\int_{\{v\in\mathbb{R}^2:v_1<0\}} f(c^v)dv$$

with

$$c^v=\frac{1}{\|T_2(F)-T_1(F)\|^2}\begin{pmatrix} T_{21}(F)-T_{11}(F) & T_{12}(F)-T_{22}(F) \\ T_{22}(F)-T_{12}(F) & T_{21}(F)-T_{11}(F)\end{pmatrix}\begin{pmatrix} v_1 \\ v_2\end{pmatrix} + \frac{T_1(F)+T_2(F)}{2}.$$

$F_m = 0.4\, N_2((-2,0), I_2) + 0.6\, N_2((2,0), I_2)$



Some comments on the IF of the ER

- The IF of the ER is bounded as soon as the gradient of the penalty function is bounded and the first moment of the model distribution exists;
- Outliers have a bigger influence in the smallest group;
- The closer the two groups are, the bigger the influence of some contamination is.

IF$(x; ER, F_N) \equiv 0$

- For some models F, IF$(x; ER, F) \equiv 0$;
- It is the case for

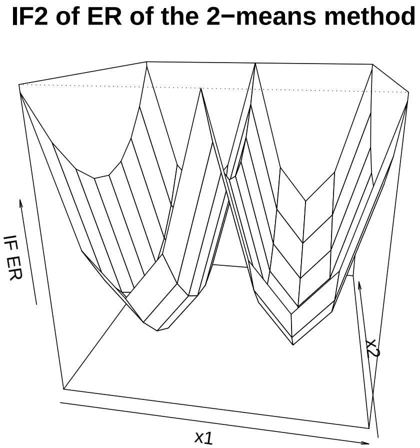
$$\textbf{(N)} \qquad F_N = 0.5\, N_p(-\mu, I_p) + 0.5\, N_p(\mu, I_p)$$

with $\mu = (\mu_1, 0, \ldots, 0)^T$;

- One needs to go one step further in the Taylor expansion;

$\Rightarrow$ Second order influence function

$F_N = 0.5\, N_2((-2,0), I_2) + 0.5\, N_2((2,0), I_2)$



Outline

Hair-plot in 1-D (Genton and Ruiz-Gazen, 2009)

- Let T be an estimator of interest;
- Let $S = (x_1, \ldots, x_n)$ be a n-vector of observations;
- Define n functions of $\xi \in \mathbb{R}$:

$$S[i, \xi] = S + \xi\, e_i$$

  where e_i is the i-th unit vector $\in \mathbb{R}^n$;
- The hair-plot is the representation of n curves

$$\mathbb{R} \to \mathbb{R} : \xi \mapsto T(S[i, \xi])$$

  for $i = 1, \ldots, n$.

Empirical influence function (EIF)

- Let T be an estimator of interest;
- Let $S = (x_1, \dots, x_n)$ be a dataset of size n;
- The EIF with replacement is defined by

$$\mathbb{R} \to \mathbb{R} : \xi \mapsto \frac{T(x_1, \dots, x_{n-1}, \xi) - T(S)}{1/n}.$$

Diagnostic plot in 1-D (1)

- Let T be an estimator of interest;
- Let $S = (x_1, \ldots, x_n)$ be a n-vector of observations;
- Define n functions of $\xi \in \mathbb{R}$:

$$S[i,\xi] = S + \xi\, e_i$$

where e_i is the i^eth unit vector in $\mathbb{R}^n$;
- The diagnostic plot is the representation of n EIF's

$$\mathbb{R} \to \mathbb{R} : \xi \mapsto \frac{T(S[i,\xi]) - T(S)}{1/n}$$

for $i = 1, \ldots, n$.

Diagnostic plot in 1-D (2)

Our estimator of interest is ER.
To compute ER, information about memberships is necessary:

- It is available ⇒ OK
- It is not available ⇒ Robust estimations

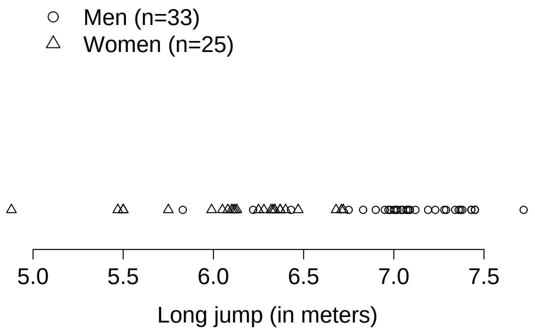
Robust estimations ?

- The 2-medoids method has a bounded IF but its BDP is null;
- It is the same for all generalized 2-means procedures !!! (García-Escudero and Gordaliza, 1999);
- Cuesta-Albertos, Gordaliza and Matrán (1997) introduced the trimmed 2-means method :

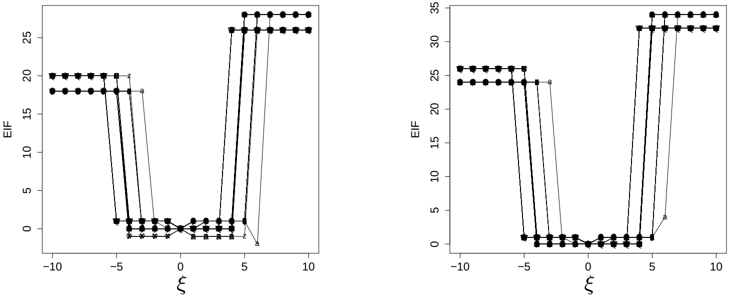
$$\min_{X_\alpha} \min_{\{t_1,\ldots,t_k\}\subset\mathbb{R}^p} \sum_{x_i\in X_\alpha} \Omega\left(\inf_{1\leq j\leq k} \|x_i - t_j\|\right)$$

where X_α ranges on the set of the subsets of $\{x_1,\ldots,x_n\}$ with $\lfloor(1-\alpha)n\rfloor$ data points.

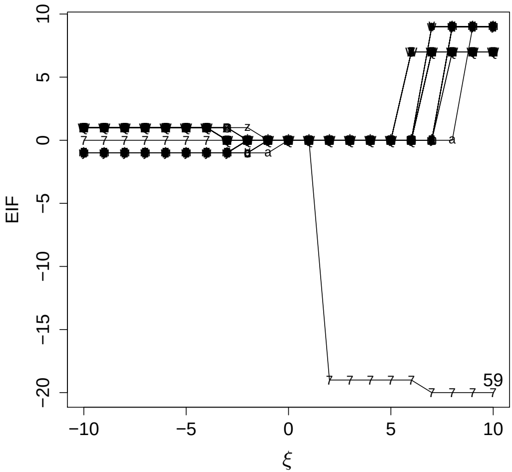
Example: Long jumps in Olympic games (García-Escudero, and Gordaliza, 1999)



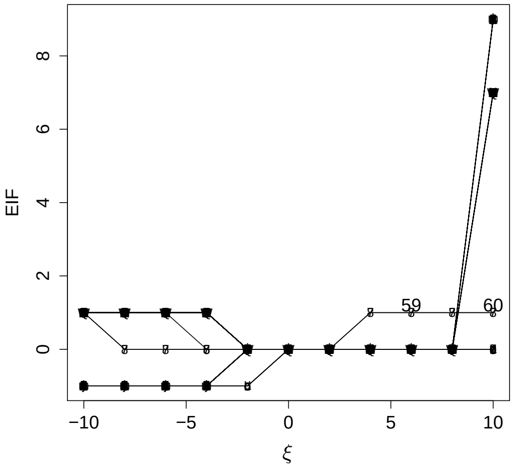
Example: Long jumps in Olympic games



Example: Long jumps in Olympic games with one outliers



Example: Long jumps in Olympic games with two outliers



Adaptation to multiple influential observations

- Let $S = (x_1, \ldots, x_n)$ be a dataset of size n;
- Define C_n^k functions of $\xi \in \mathbb{R}$:

$$S[i, \ldots, j, \xi] = S + \xi\,(e_i + \ldots + e_j)$$

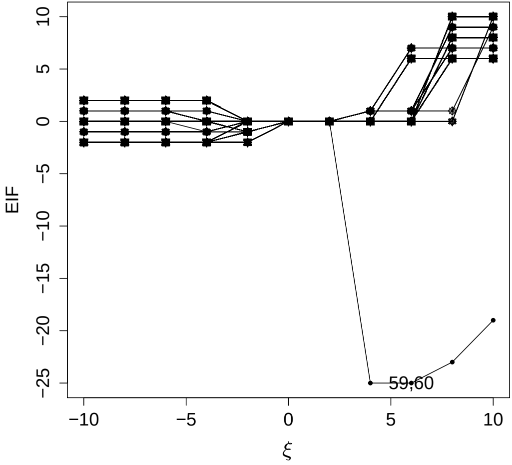
where e_l is the l^eth unit vector in $\mathbb{R}^n$;

- Our diagnostic plot is the representation of C_n^k EIF's

$$\mathbb{R} \to \mathbb{R} : \xi \mapsto \frac{\mathrm{ER}(S[i, \ldots, j, \xi]) - \mathrm{ER}(S)}{1/n}$$

for $i, \ldots, j \in \{1, \ldots, n\}$.

Example: Long jumps in Olympic games with two outliers



Outline

Conclusions

- The generalized 2-means procedure can give a more robust estimator of the error rate with a bounded penalty function;
- But it is not robust w.r.t. all kind of contamination $\Rightarrow$ introduction of the trimmed 2-means method;
- The diagnostic plot presented here can be useful to detect single influential observation or multiple influential observations in one dimension.

Future research

- Adaptation of the diagnostic plot to multivariate cases;
- Error rate of the 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).

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 efficiencies 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
- Cuesta-Albertos J.A., Gordaliza A., and Matrán C. (1997), Trimmed k-means: an attempt to robustify quantizers, *The Annals os Statistics* 25, pp.553-576

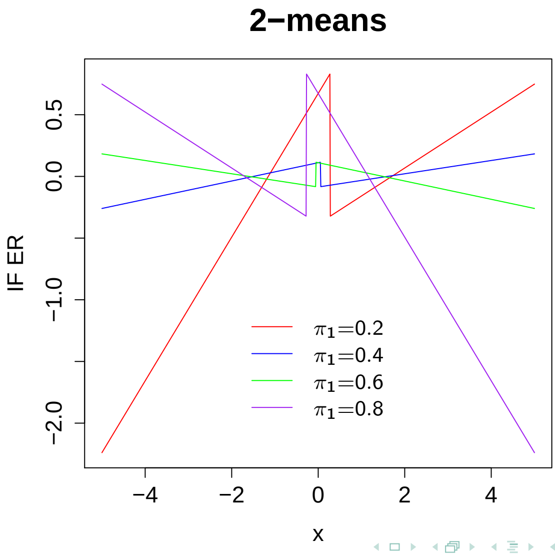
Bibliography

- Fernholz L. T. (2001), On multivariate higher order von Mises expansions, *Metrika* 53, pp. 123-140
- García-Escudero L. A., and Gordaliza A. (1999), Robustness properties of *k*-means and trimmed *k*-means, *Journal of the American Statistical Association* 94, pp. 956-969
- García-Escudero L. A., Gordaliza A., and Matrán C. (2003), Trimming tools in exploratory data analysis, *Journal of Computational and Graphical Statistics* 12, pp. 434-449
- 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

Bibliography

- Pollard D. (1981), Strong Consistency of k-Means Clustering, *The Annals of Probability* 9, pp.919-926
- Pollard D. (1982), A Central Limit Theorem for k-Means Clustering, *The Annals of Probability* 10, 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.

Graph of IF(x; ER, F)



Graph of IF(x; ER, F)