A robust clustering approach to fraud detection

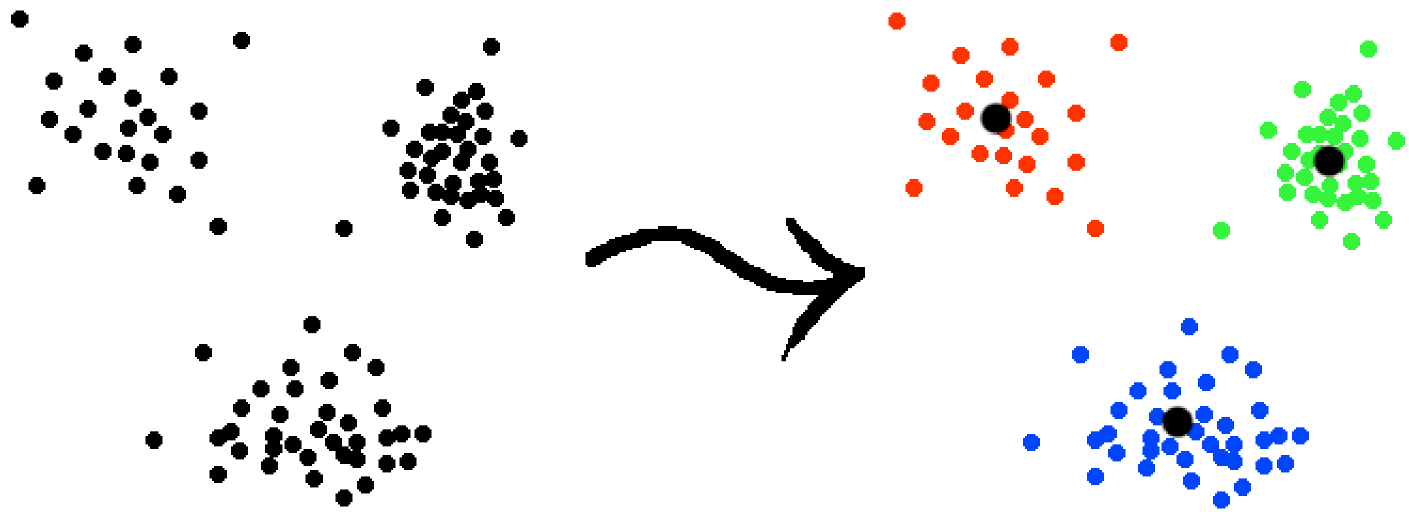
Luis Angel García-Escudero

Dpto. de Estadística e I.O. and IMUVA - Universidad de Valladolid

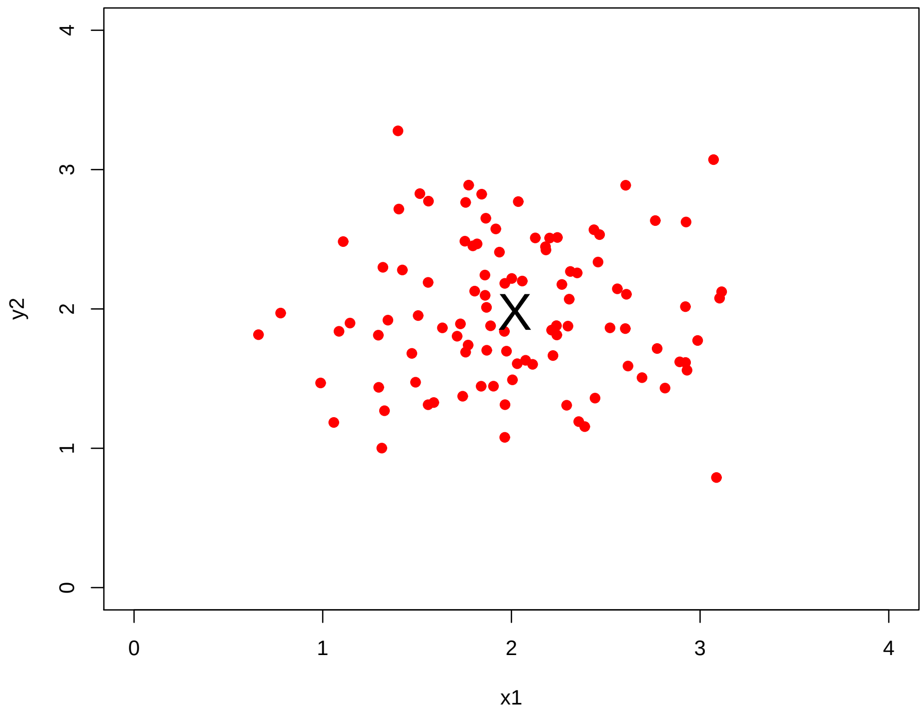
joint (and on-going...) work with A. Mayo-Iscar, A. Gordaliza, C. Matrán (U. Valladolid) and colleagues from M. Riani, A. Cerioli (U. Parma) and D. Perrotta, F. Torti (JRC-Ispra)

1. CLUSTERING AND ROBUSTNESS

- **Clustering** is the task of **grouping** a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other clusters:



- **Sample mean:**
  - ◇ $m = \frac{1}{n}\sum_{i=1}^{n} x_i$ minimizes $\sum_{i=1}^{n} \|x_i - m\|^2$
  - ◇ m may be seen as the "center" of a data-cloud:



- k **clusters** $\Rightarrow$ k "data-clouds" $\Rightarrow$ k**-means**

- **k-means:** Search for

  - $\diamond$ k centers $m_1, ..., m_k$
  - $\diamond$ a partition $\{R_1, ..., R_k\}$ of $\{1, 2, ..., n\}$

  minimizing

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

- **Cluster** j:

$$R_j = \{i : \|x_i - m_j\| \leq \|x_i - m_l\| \text{ for every } l = 1, ..., k\}$$

  (...assignment to the closest center...)

- **Robustness:** Many statistical procedures are strongly affected by even few outlying observations:

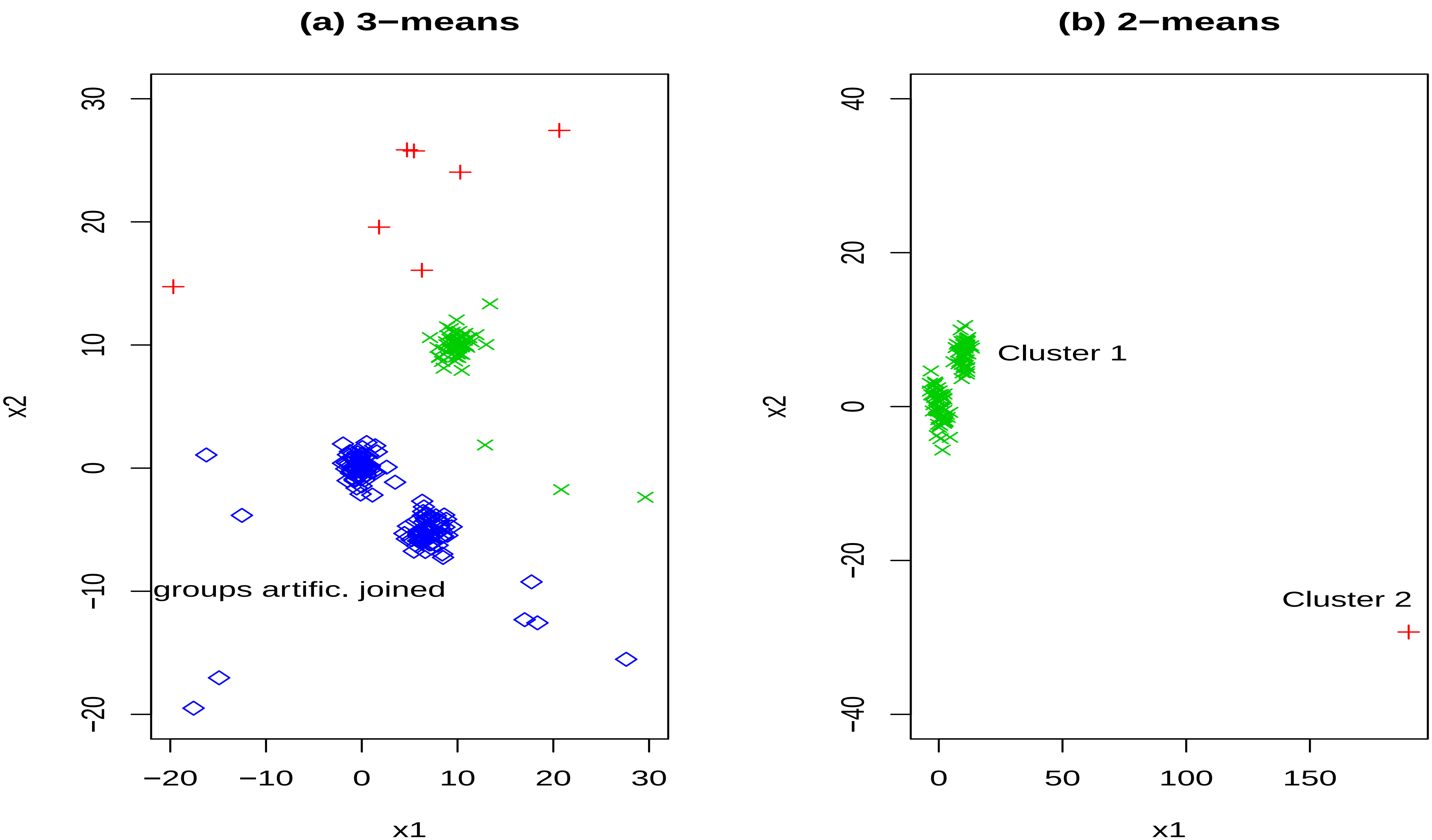
  ◇ The mean is not robust:

$$\overline{x} = \frac{1.72 + 1.67 + 1.80 + 1.70 + 1.82 + 1.73 + 1.78}{7} = 1.745$$

$$\overline{x} = \frac{1.72 + 1.67 + 1.80 + 1.70 + 182 + 1.73 + 1.78}{7} = 27.485$$

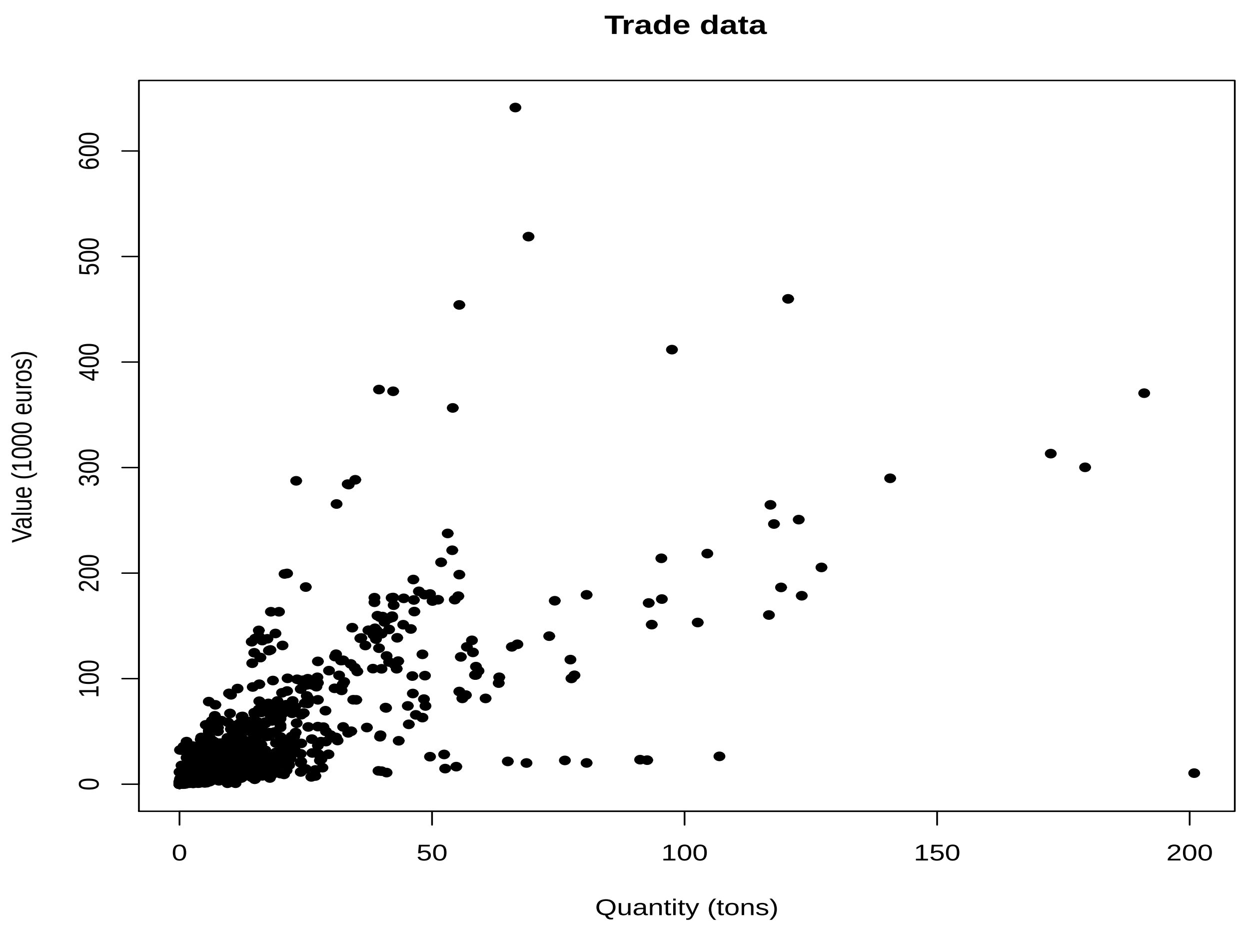
  ◇ k-**means** inherits that **lack of robustness from the mean**

- **Lack of robustness of k-means:**



- **Outliers** can be seen as **"clusters by themselves"**
- **So, why not increasing the number of clusters...?**
  - **But:**
    - Due to (physical, economical,...) reasons we could have an initial idea of k without being aware of the existence of outliers
    - "Radial/background" noise requires large k's
- **Moreover, the detection of outliers may be the goal itself!!!**

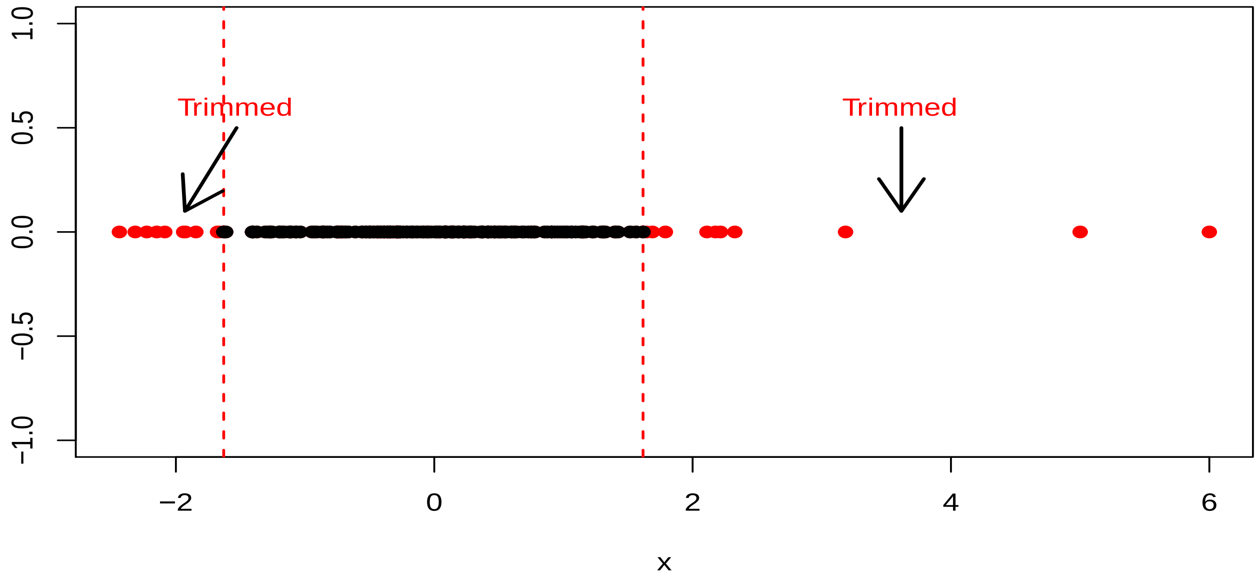
- **Outliers in trade data can be associated to “frauds”:**
  - ◇ Heterogeneous sources of data (clustering) + Few outliers (frauds??)



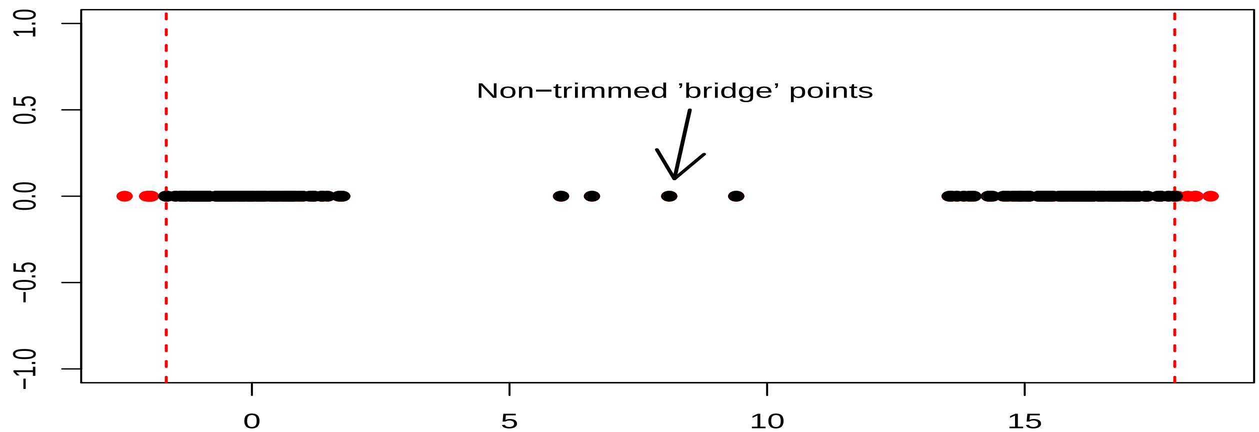
2.- TRIMMED k-MEANS

- **Trimming** is the **oldest** and most widely **used** way to achieve robustness.

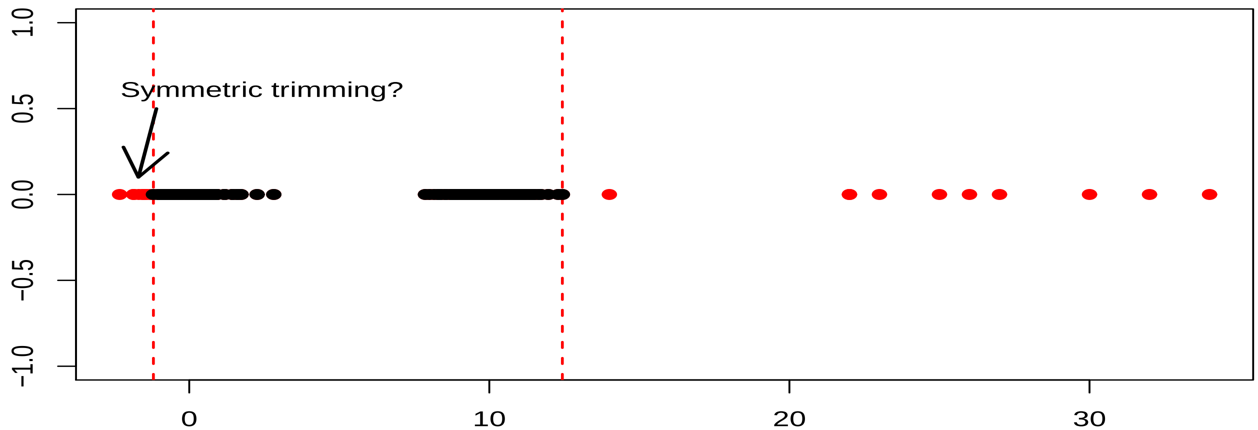
- **Trimmed mean:** The proportion $\alpha/2$ smallest and $\alpha/2$ largest observations are discarded before computing the mean:



- **But,... how to trim in clustering?**
  - ◇ Why not trimming outlying "bridge" points?



  - ◇ Why a symmetric trimming?



  - ◇ How to trim in *multivariate* clustering problems?

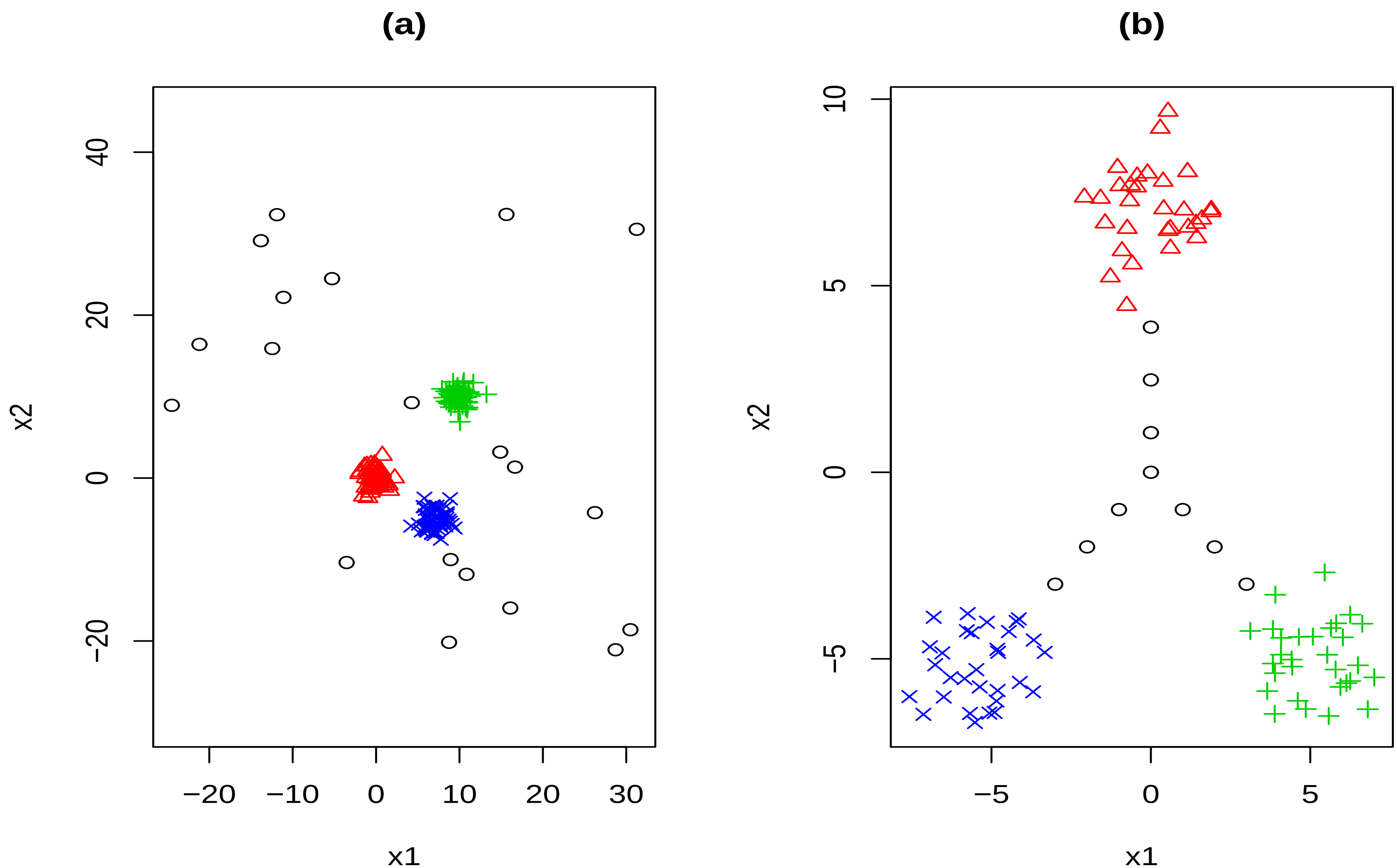
- **Idea: Data itself tell us which are the most outlying observations!!**
  - ◇ Data-driven, adaptive, impartial,... trimming!
- **Trimmed k-means:** we search for
  - ◇ k centers $m_1, ..., m_k$ and
  - ◇ a partition $\{R_0, R_1, ..., R_k\}$ of $\{1, 2, ..., n\}$ with $\boxed{\#R_0 = [n\alpha]}$

minimizing

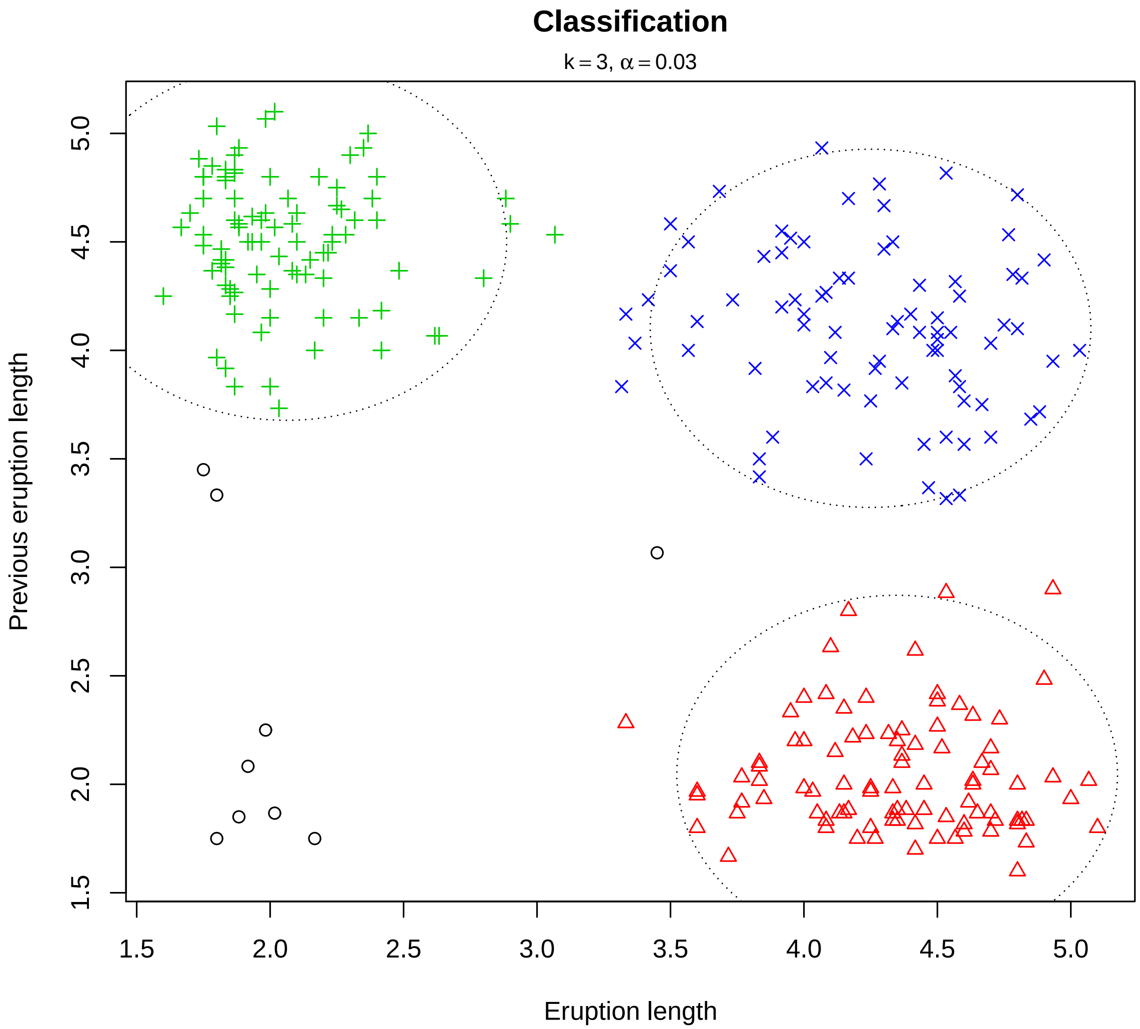
$$\sum_{j=1}^{k} \sum_{i \in R_j} \|x_i - m_j\|^2.$$

[A fraction α of data is not taken into account $\rightsquigarrow$ Trimmed]

- Black circles: *trimmed points* $(k = 3$ and $\alpha = 0.05)$:



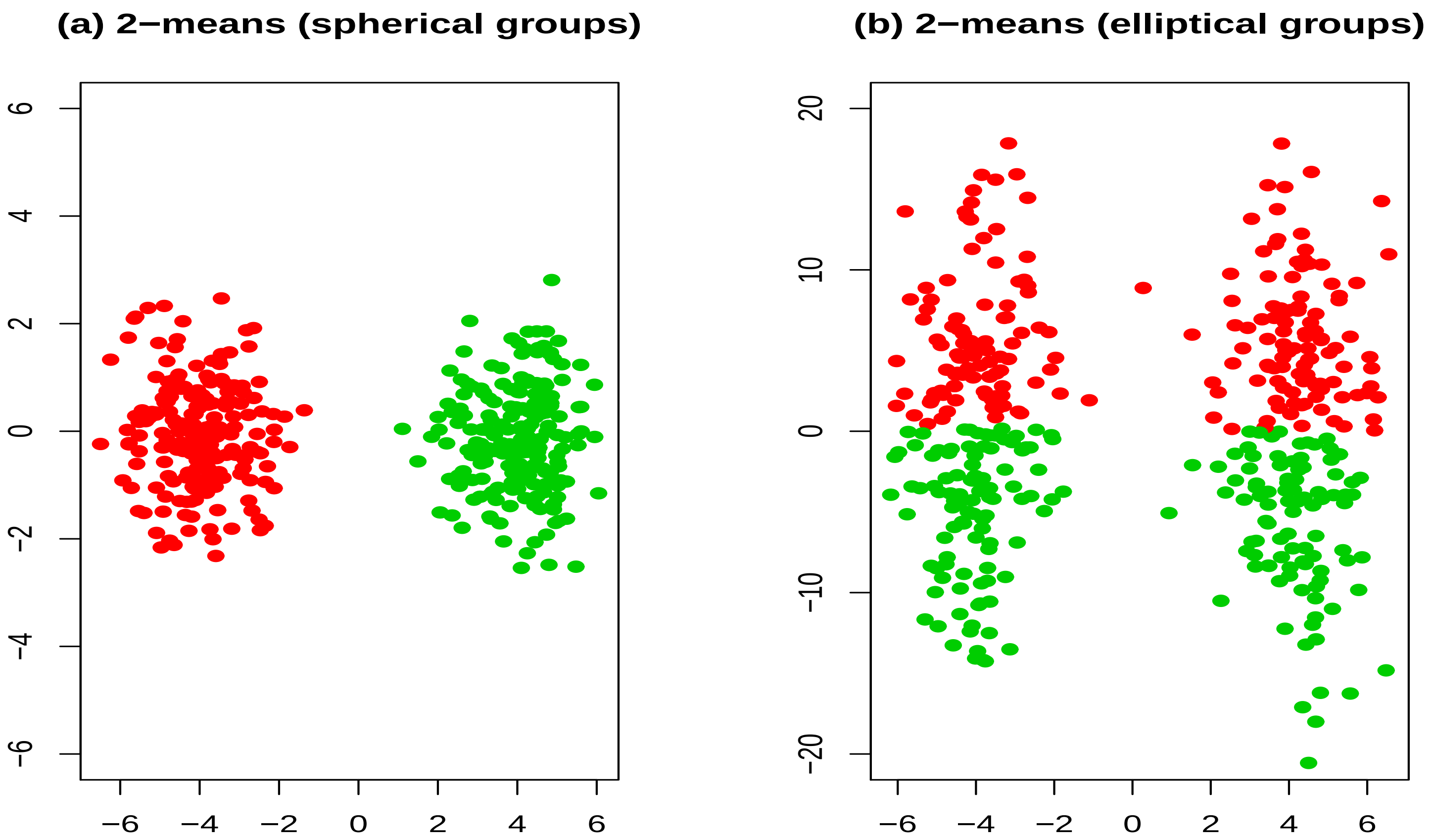
- **Old Faithful Geyser data**: $x_1 =$ "Eruption length", $x_2 =$ "Previous eruption length" and $n = 271$



◇ $k = 3$ and $\alpha = 0.03$ ($0.03 \cdot 271 \simeq 9$ **trimmed obs.**): 6 rare "short-followed-by-short" eruptions trimmed, 3 bridge points...

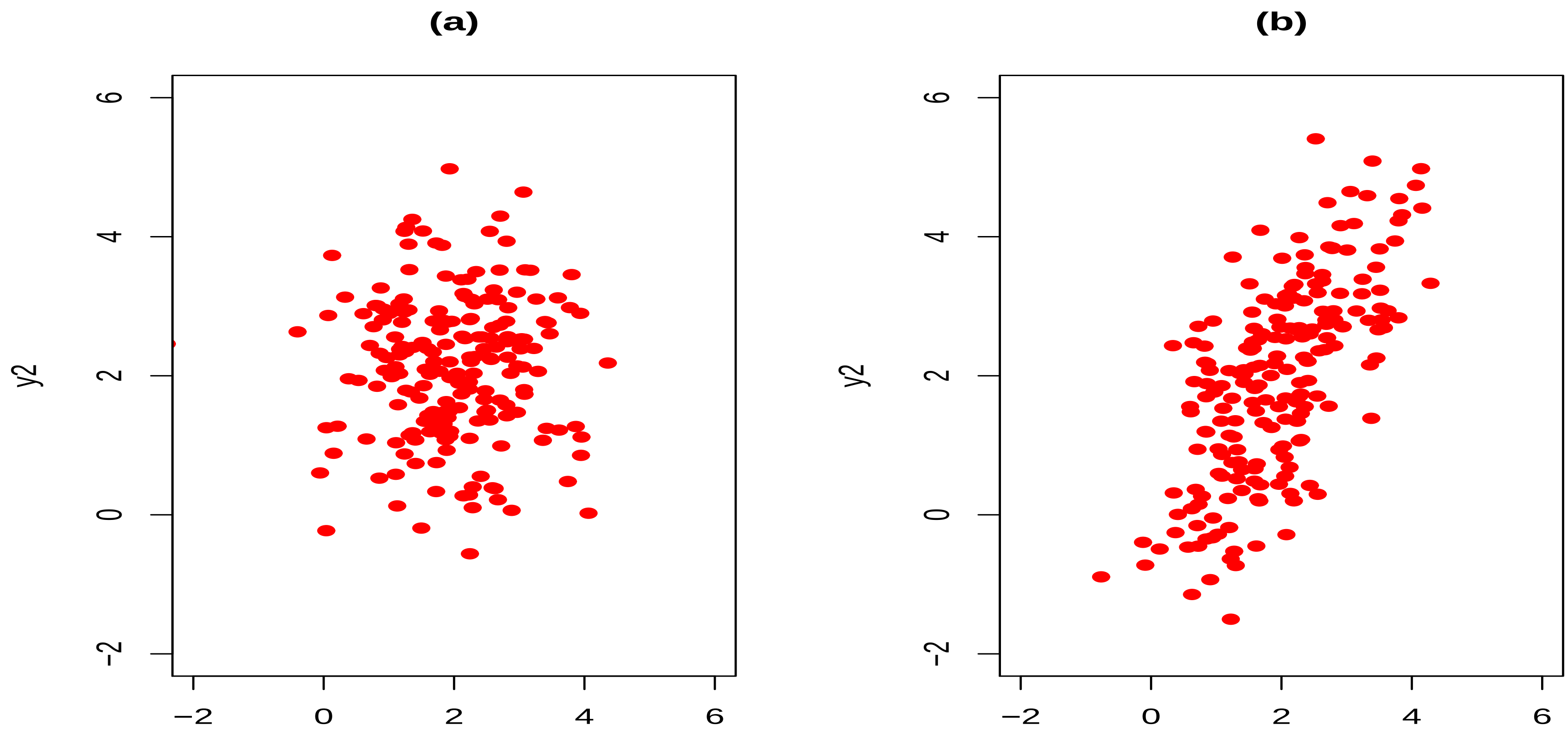
3.- ROBUST MODEL-BASED CLUSTERING

- k-means and trimmed k-means prefer **spherical clusters:**



- Elliptically contoured clusters?

- **Multivariate normal** distributions with densities $\phi(\cdot; \mu, \Sigma)$:
  - ◇ $\mu = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ [spherical] in (a)
  - ◇ $\mu = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ [non-spherical] in (b)



$$\phi(x; \mu, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\big(- (x - \mu)' \Sigma^{-1} (x - \mu)/2 \big)$$

- **Trimmed likelihoods:** Search for
  - ◇ k centers $m_1, ..., m_k$,
  - ◇ k scatter matrices $S_1, ..., S_k$, and,
  - ◇ a partition $\{R_0, R_1, ..., R_k\}$ of $\{1, 2, ..., n\}$ with $\# R_0 = [n\alpha]$

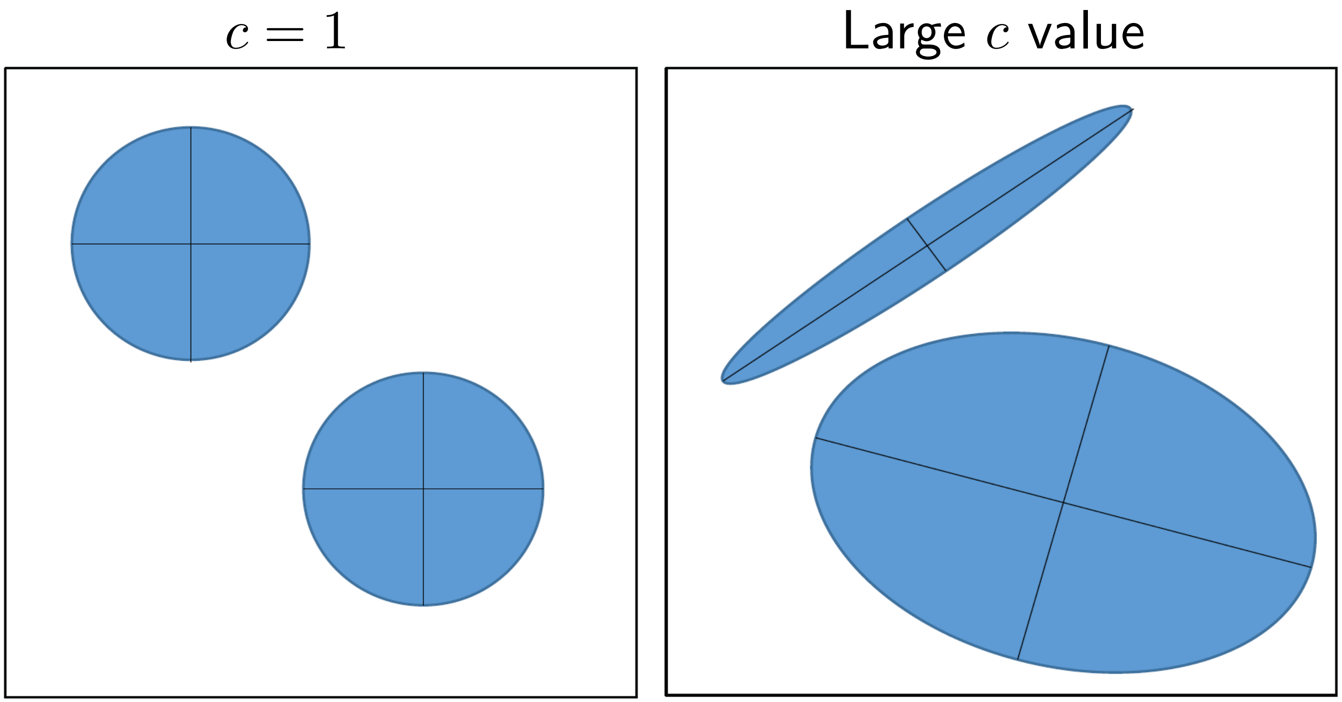
  maximizing

$$\sum_{j=1}^{k} \sum_{x_i \in R_j} \log \phi(x_i; m_j, S_j) \qquad \text{(obs. in } R_0 \text{ not taken into account)}$$

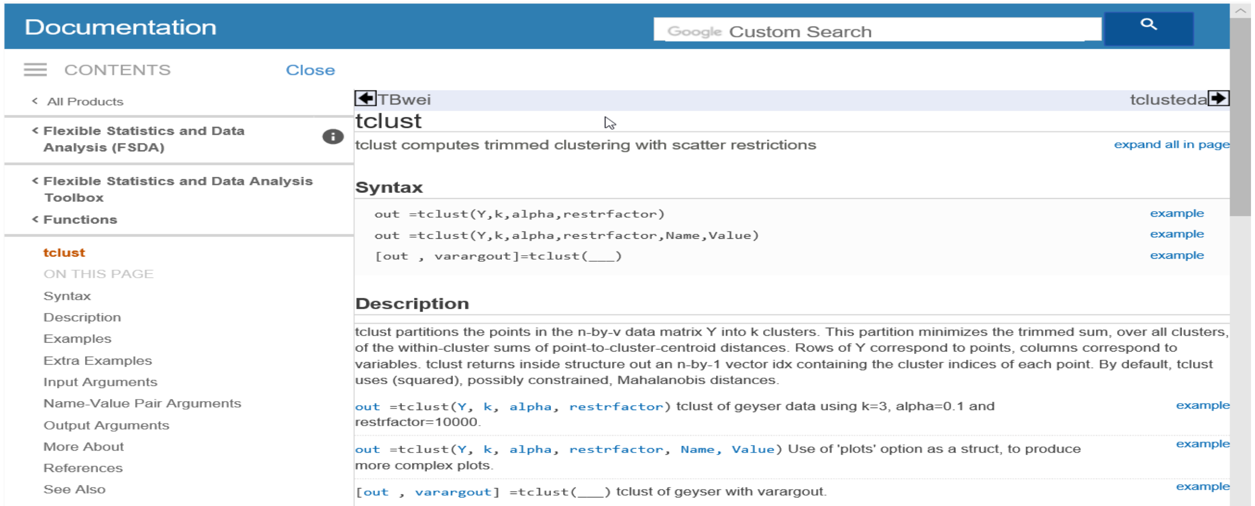
García-Escudero et al 2008, Neykov et al 2007, Gallegos and Ritter 2005,...

- **Constraints** on the S_j scatter matrices **needed**:
  - ◇ Unbounded target likelihood functions
  - ◇ Avoid detecting (non-interesting) “spurious” clusters

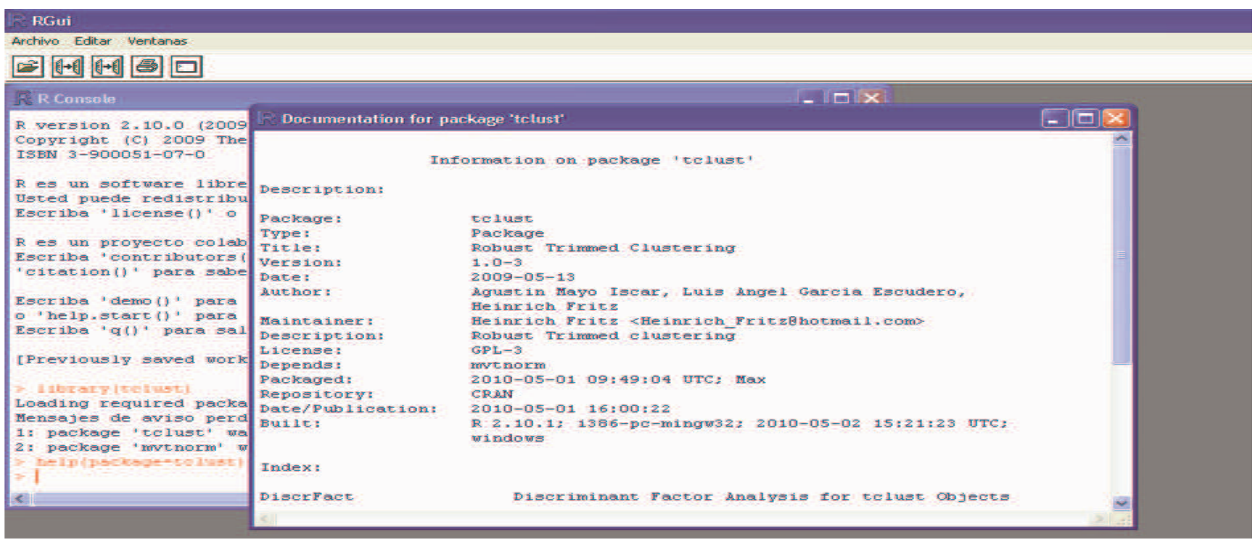
- Control relative axes’ lengths (eigenvalues constraints):



- The FSDA Matlab toolbox:



- The R package tclust at CRAN repository:



- The R package `tclust`:

  `> library(tclust)`

- `tkmeans(data, k, alpha)`

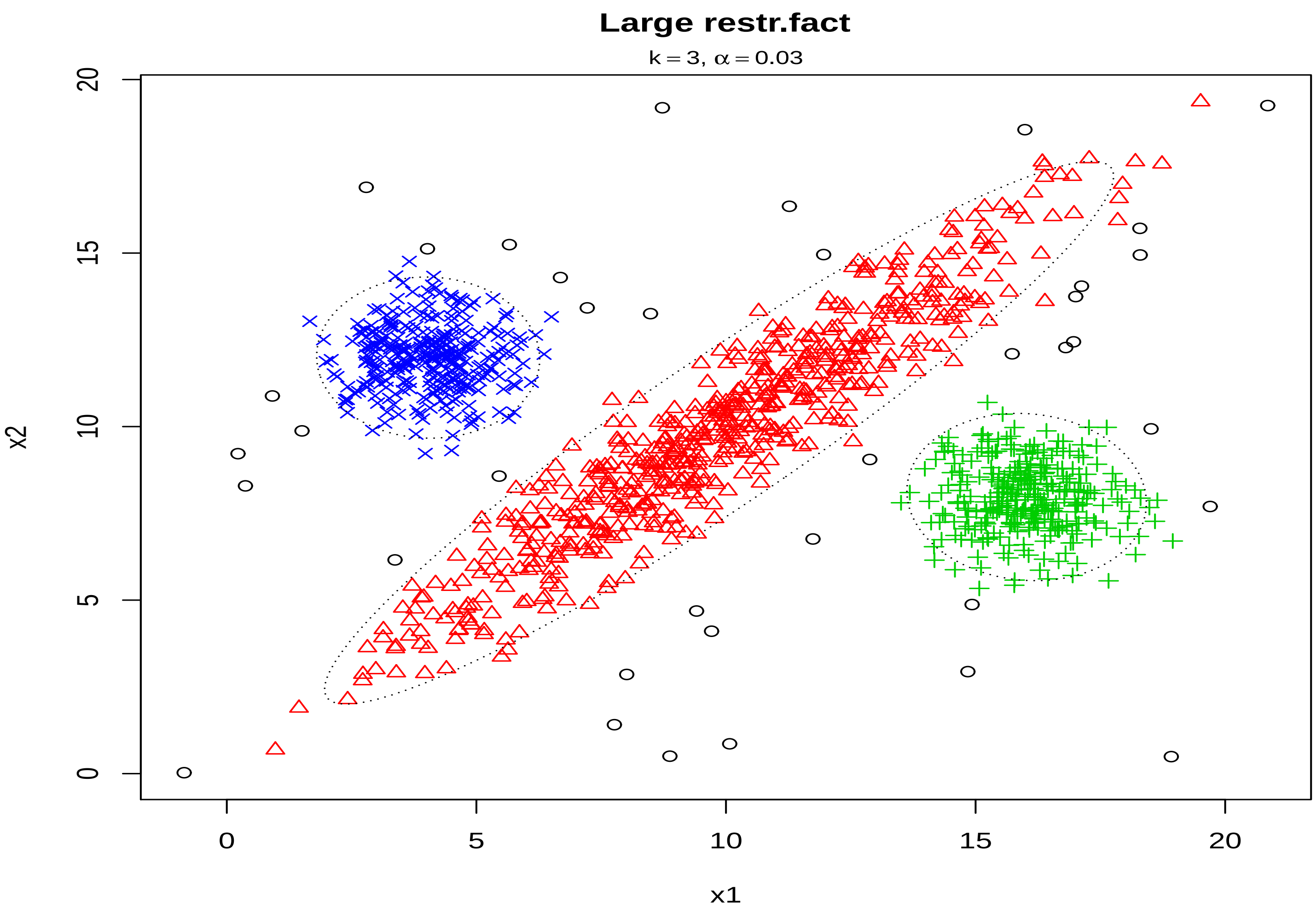
  ◇ `k` = "number of groups"

  ◇ `alpha` = "trimming proportion"

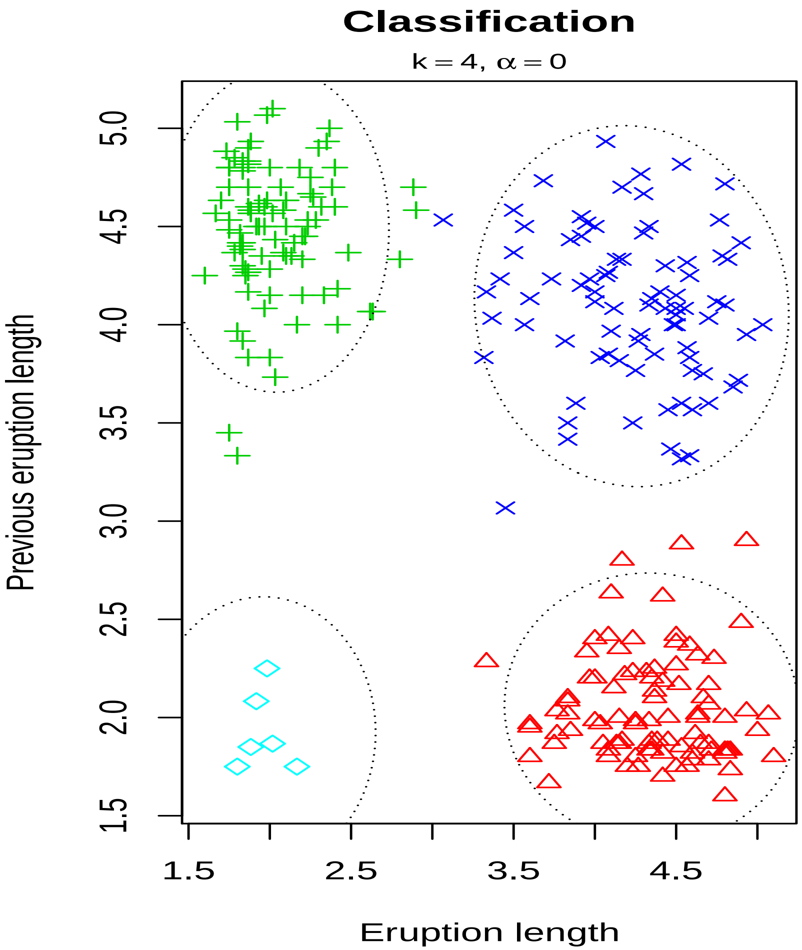
- `tclust(data, k, alpha, restr.fact, ...)`

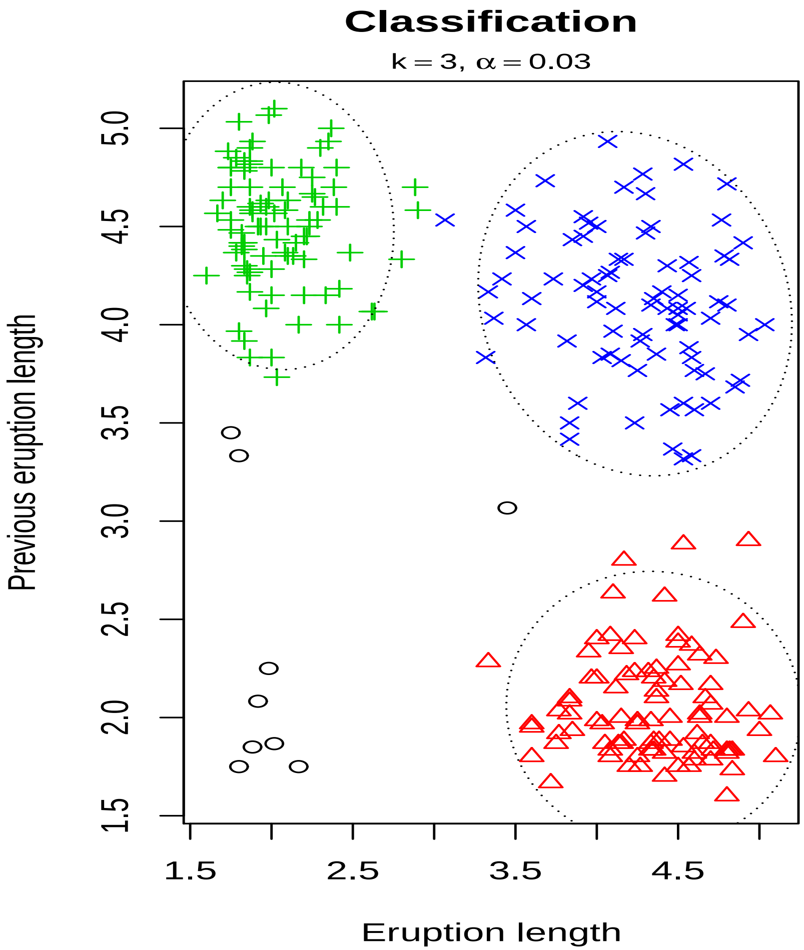
  ◇ `restr.fact` = "Strength of the constraints"

- tclust(X,k=3,alpha=0.03,restr.fact=50)



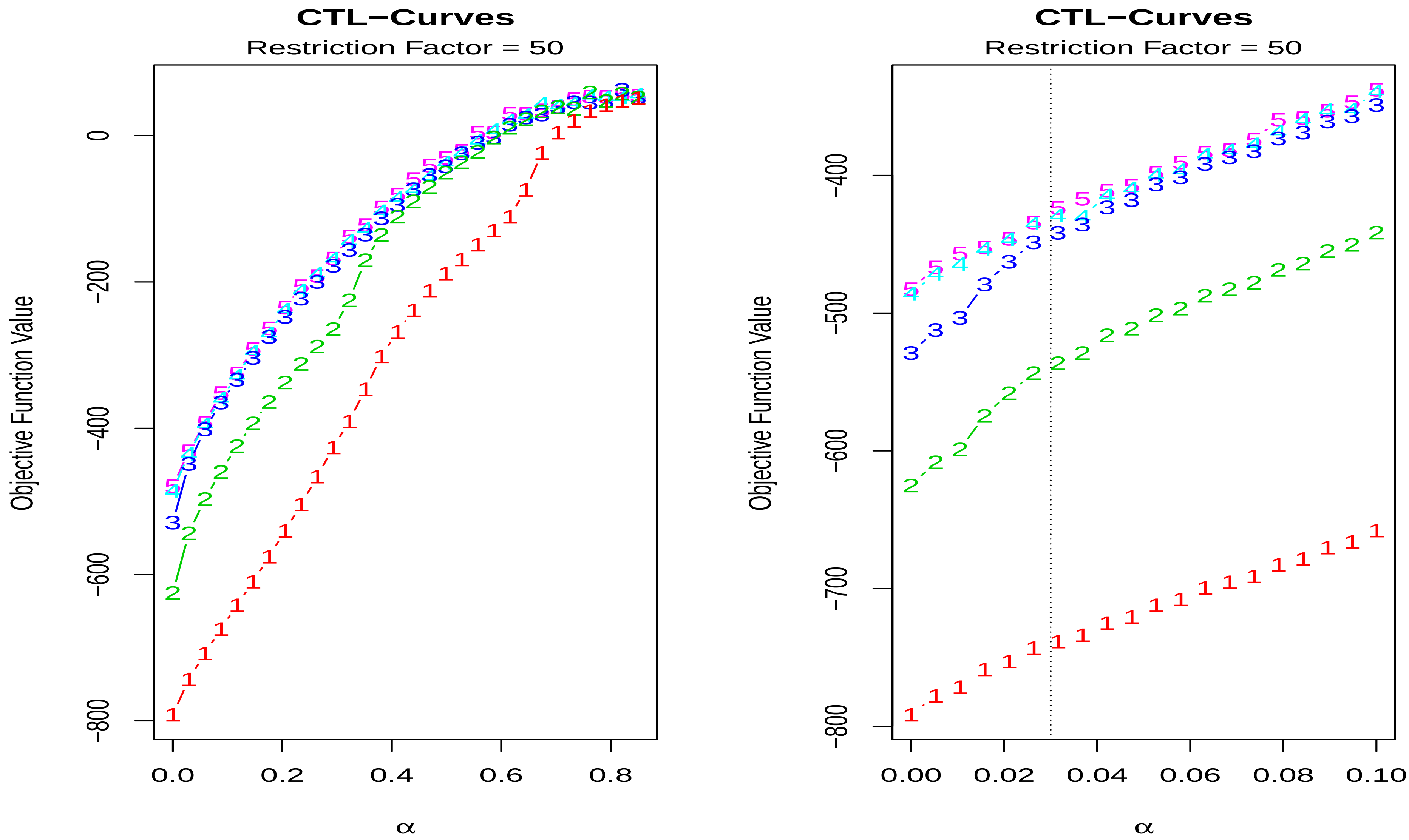
- **Old Faithful Geyser data again:**





- Why $k = 3$ and $\alpha = 0.03$ was a sensible solution?

- **Applying** `ctlcurves` to the **Old Faithful Geyser** data:



4.- ROBUST CLUSTERING AROUND LINEAR SUBSPACES

- **Robust linear grouping:** Higher p dimensions, but assuming that our data "live" in k low-dimensional (affine) subspaces...

  ◇ We search for

  · k linear subspaces $h_1, ..., h_k$ in $\mathbb{R}^p$

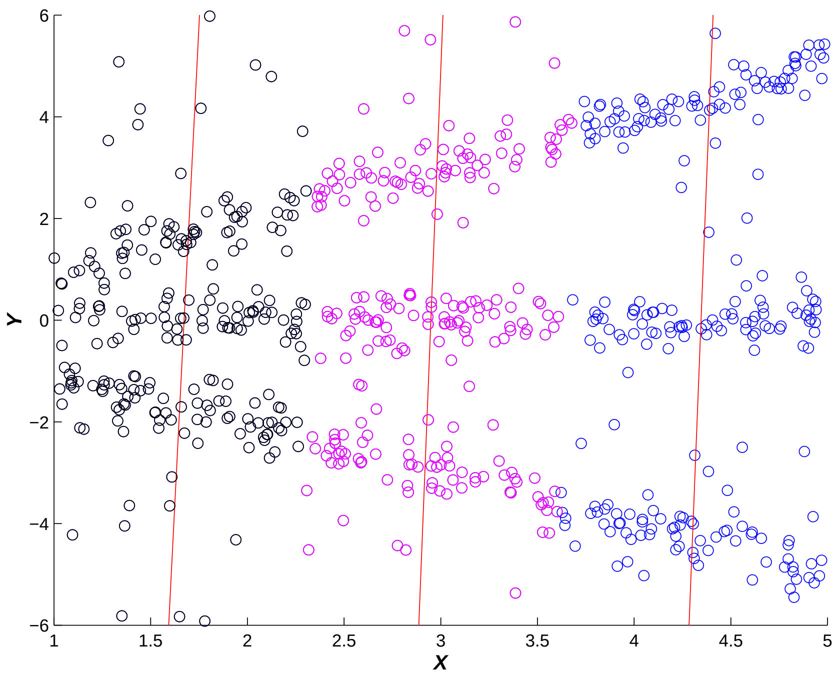
  · a partition $\{R_0, R_1, ..., R_k\}$ of $\{1, 2, ..., n\}$ with $\boxed{\#R_0 = [n\alpha]}$

  minimizing

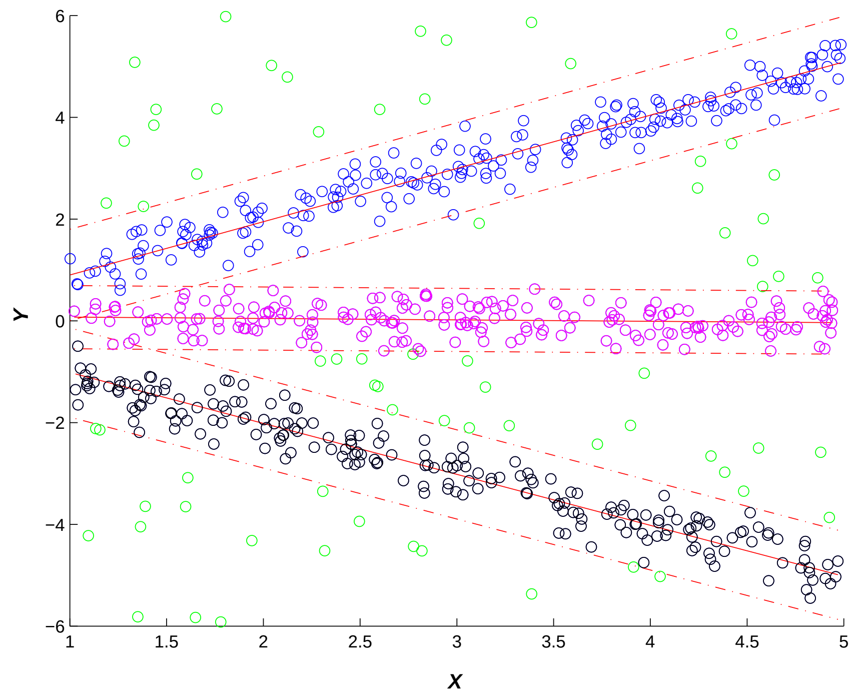
$$\sum_{j=1}^{k} \sum_{i \in R_j} \|x_i - \mathrm{Pr}_{h_j}(x_i)\|^2.$$

  ◇ $\mathrm{Pr}_h(\cdot)$ denotes the "orthogonal" projection onto the linear subspace h

- **Example:** Three linear structures in presence of noise:



(a) $\alpha = 0$



(b) $\alpha = 0.1$ (○ = "Trimmed")

Trimmed "mixtures of regressions" can also be applied...

- $k=1$ **case** $\Rightarrow$ **Robust "Principal Components Analysis (PCA)":**
  - ◇ PCA provides a q-dimensional ($q << p$) representation of data by

$$\min_{\boldsymbol{B}_q, \boldsymbol{A}_q, m} \sum_{i=1}^{n} ||\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i||^2 \text{ for}$$

$$\widehat{\boldsymbol{x}}_i = \mathsf{Pr}_h(\boldsymbol{x}_i) = \widehat{\boldsymbol{x}}_i(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m}) = \boldsymbol{m} + \boldsymbol{B}_q \boldsymbol{a}_i$$

· $\boldsymbol{A}_q = \begin{pmatrix} -\boldsymbol{a}_1- \\ \cdots \\ -\boldsymbol{a}_i- \\ \cdots \\ -\boldsymbol{a}_n- \end{pmatrix}$ is the **scores** matrix ($n \times q$)

· $\boldsymbol{B}_q = \begin{pmatrix} -\boldsymbol{b}_1- \\ \cdots \\ -\boldsymbol{b}_j- \\ \cdots \\ -\boldsymbol{b}_p- \end{pmatrix}$ is a matrix ($p \times q$) whose columns generate a q-dimensional **approximating subspace** h

- **Principal Components Analysis is highly non-robust!!!**

- **Least Trimmed Squares PCA** (Maronna 2005): Minimize

$$\sum_{i=1}^{n} w_i \|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i\|^2 = \sum_{i=1}^{n} w_i \|\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i(\boldsymbol{B}_q, \boldsymbol{A}_q, \boldsymbol{m})\|^2,$$

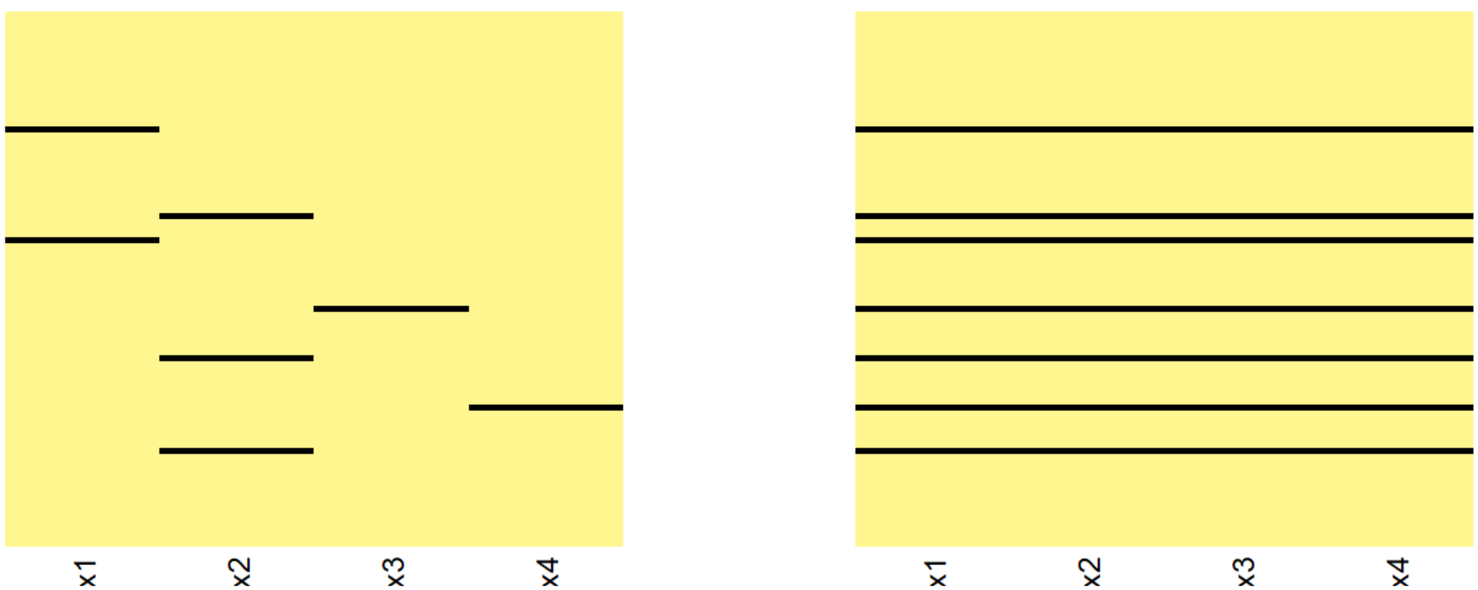
with $\{w_i\}_{i=1}^{n}$ being "0-1 **weights**" such that

$$\sum_{i=1}^{n} w_i = [n(1 - \alpha)]$$

◇ Weights: $w_i = \begin{cases} 1 & \text{If } \boldsymbol{x}_i \text{ is not trimmed} \\ 0 & \text{If } \boldsymbol{x}_i \text{ is trimmed} \end{cases}.$

- **Cases** $\to \boldsymbol{x}_i = (x_{i1}, ..., x_{ip})' \in \mathbb{R}^p$ and **Cells** $\to x_{ij} \in \mathbb{R}$
  - ◇ i denotes a country (or a trader; company;...) for $i = 1, ..., n$
  - ◇ x_{ij} is the "quantity-value ratio" for country i in the j-th month (or the j-th year; the j-th product;...) for $j = 1, ..., p$
- **Casewise trimming:** Trim $\boldsymbol{x}_i$ cases with (at least one) outlying x_{ij}

$n = 100 \times p = 4$ data matrix with 2% outlying cells:

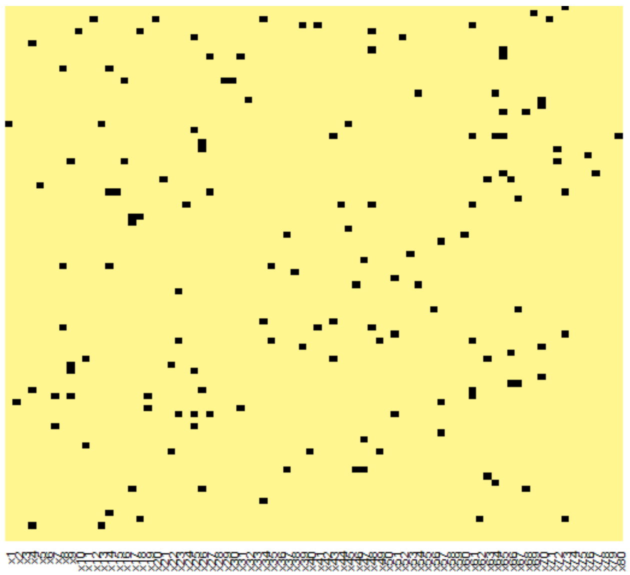


Outlying x_{ij} cells

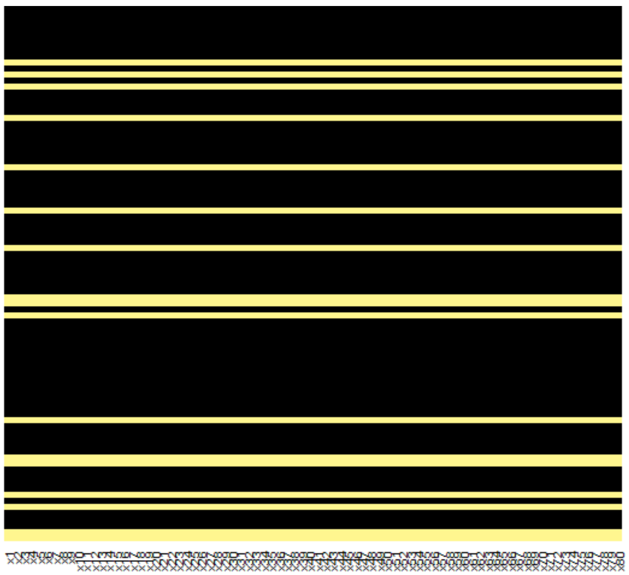
Trimmed $\boldsymbol{x}_i$ cases (black lines)

- But when the dimension p increases... we do not expect many $\boldsymbol{x}_i$ completely free of outlying x_{ij} cells:

  $n = 100 \times p = 80$ data matrix with 2% outlying cells:



Outlying x_{ij} cells



Trimmed $\boldsymbol{x}_i$ cases (black lines)

- **Cellwise trimming:**

  ◇ Only trimming outlying cells... ($\Rightarrow$ “Particular” frauds identified...??)

- PCA approximation $\widehat{\boldsymbol{x}}_i = \boldsymbol{m} + \boldsymbol{B}_q \boldsymbol{a}_i = (\widehat{x}_{i1}, ..., \widehat{x}_{ip})^T$ re-written as

$$\widehat{x}_{ij} = m_j + \boldsymbol{a}_i^T \boldsymbol{b}_j.$$

- **Cellwise LTS** (Cevallos-Valdiviezo 2016): Minimize

$$\sum_{i=1}^{n} w_{ij}(x_{ij} - m_j - \boldsymbol{a}_i^T \boldsymbol{b}_j)^2$$

  ◇ $w_{ij} = 0$ if cell x_{ij} is trimmed and $w_{ij} = 1$ if not with

$$\sum_{i=1}^{n} w_{ij} = [n(1-\alpha)], \text{ for } j = 1, ..., p.$$

- Different patterns/structures in data $\Rightarrow$ G **subspace approximations:**

$$\widehat{\boldsymbol{x}}_i^g(\boldsymbol{B}_{q_g}^g, \boldsymbol{A}_{q_g}^g, \boldsymbol{m}^g) = \boldsymbol{m}^g + \boldsymbol{B}_{q_g}^g \boldsymbol{a}_i^g \quad \text{or} \quad \widehat{x}_{ij}^g = m_j^g + (\boldsymbol{a}_i^g)^T \boldsymbol{b}_j^g,$$

for $g = 1, ..., G$

- Minimize

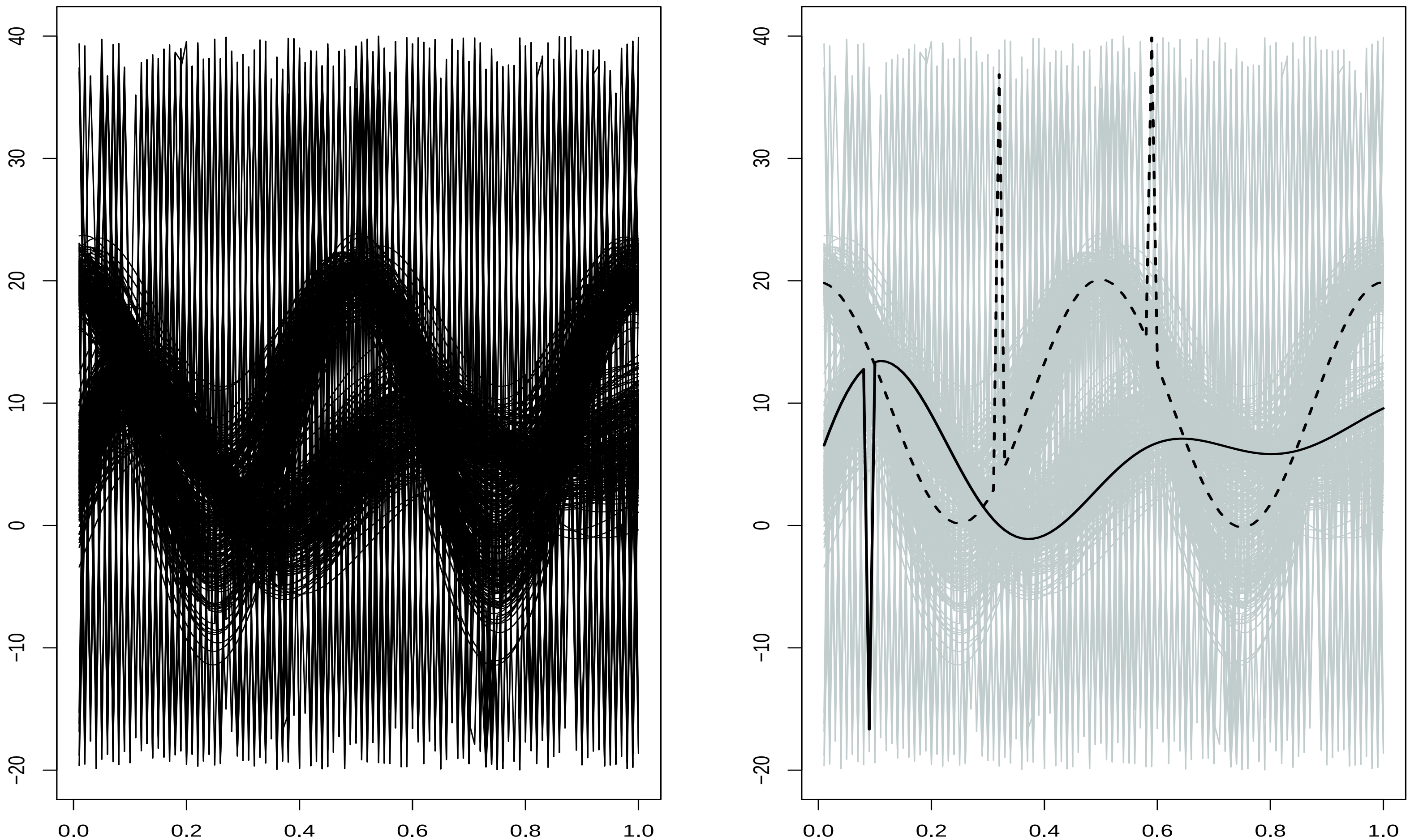
$$\min_{w_{ij}^g, B_q^g, A_q^g, \boldsymbol{m}^g} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{g=1}^{G} w_{ij}^g (x_{ij} - \widehat{x}_{ij}^g)^2.$$

  - ◇ $w_{ij}^g = 1$ if cell x_{ij} is assigned to cluster g and non-trimmed and 0 otherwise

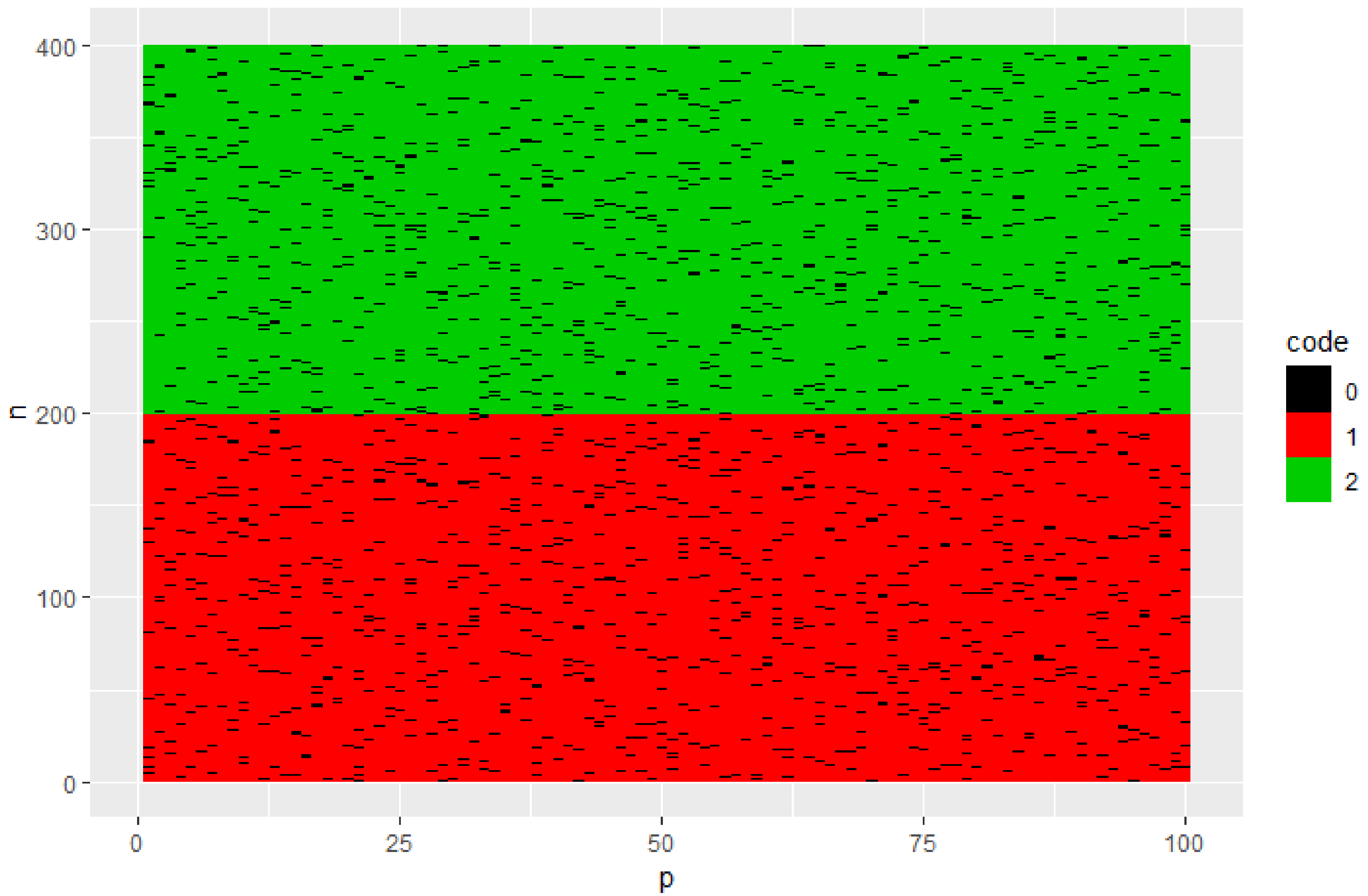
  - ◇ Appropriate constraints on the w_{ij}^g

$q_1, \ldots, q_G$ are intrinsic dimensions...

- **Example 1:** $n = 400$ in dimension $p = 100$ with 2 groups and 2% "scattered" outliers:

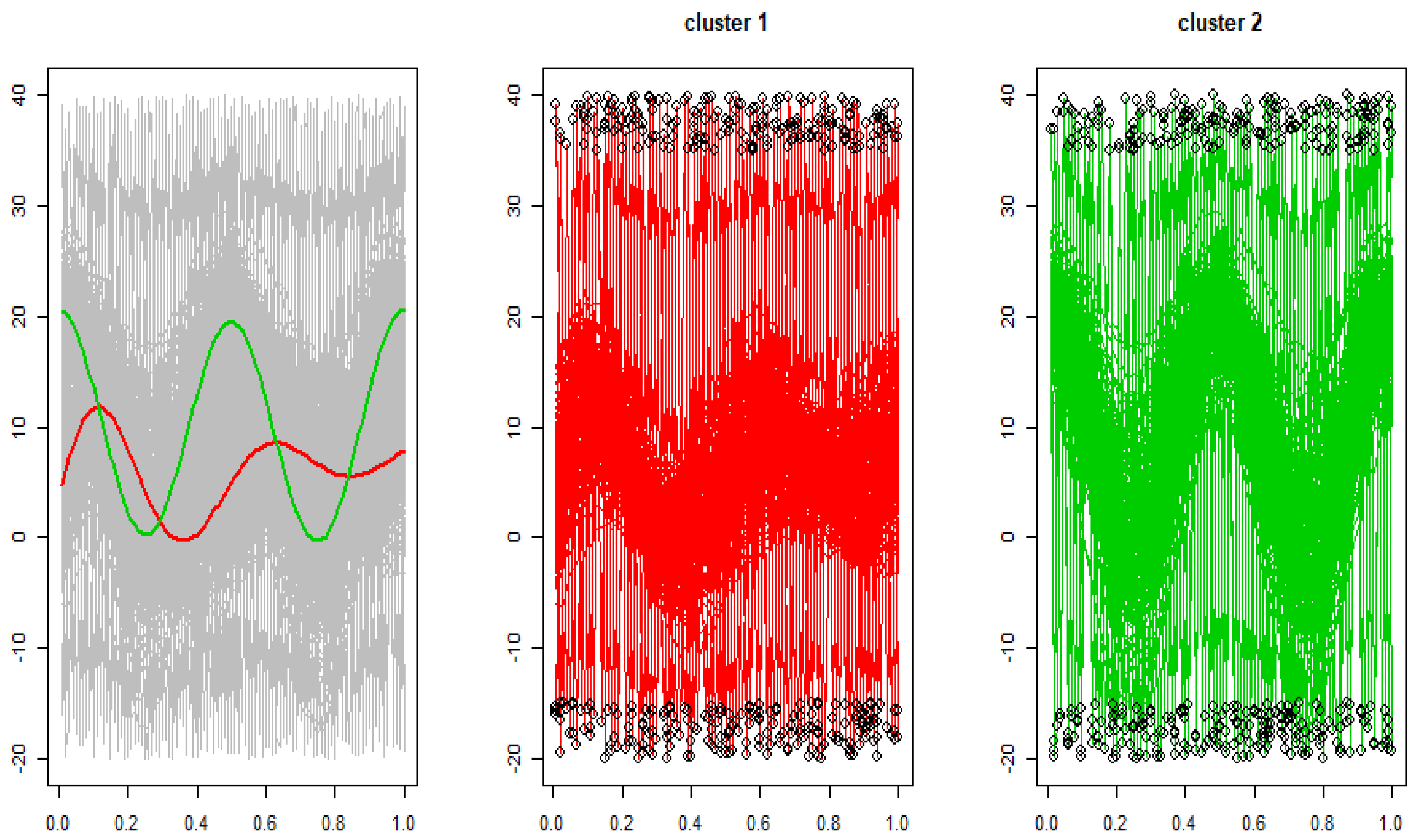


- $k = 2$, $q = 2$ **and** $\alpha = 0.05$:

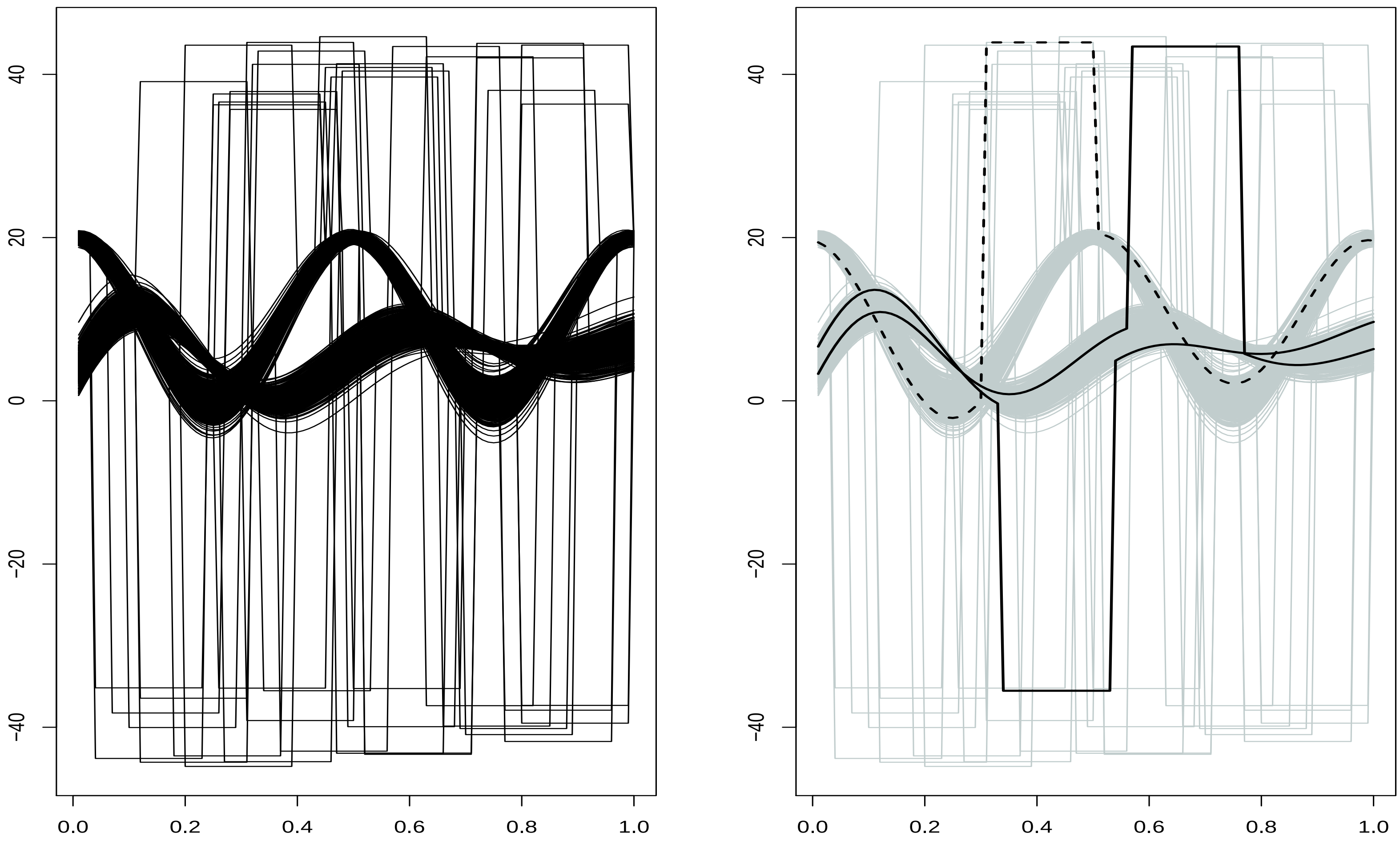


"-" are the trimmed cells

- **Cluster means and trimmed cells (○):**

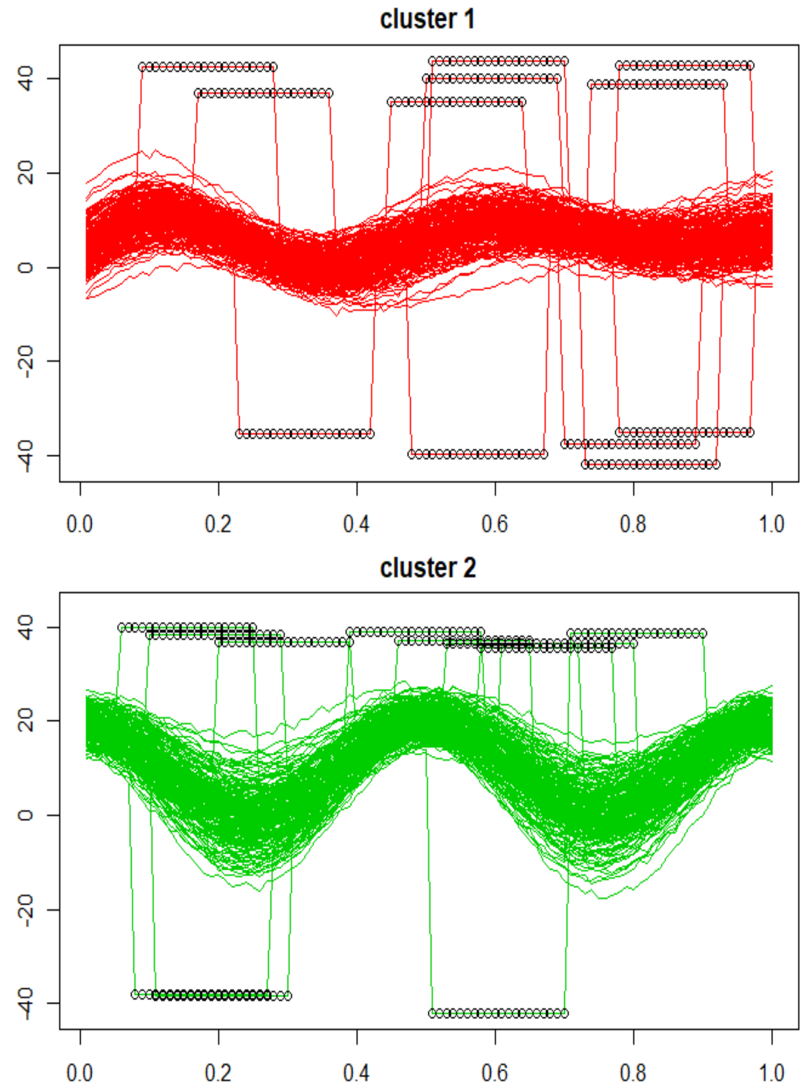


- **Example 2:** $n = 400$ in dimension $p = 100$ with 2 groups and few curves with 20% consecutive cells corrupted:



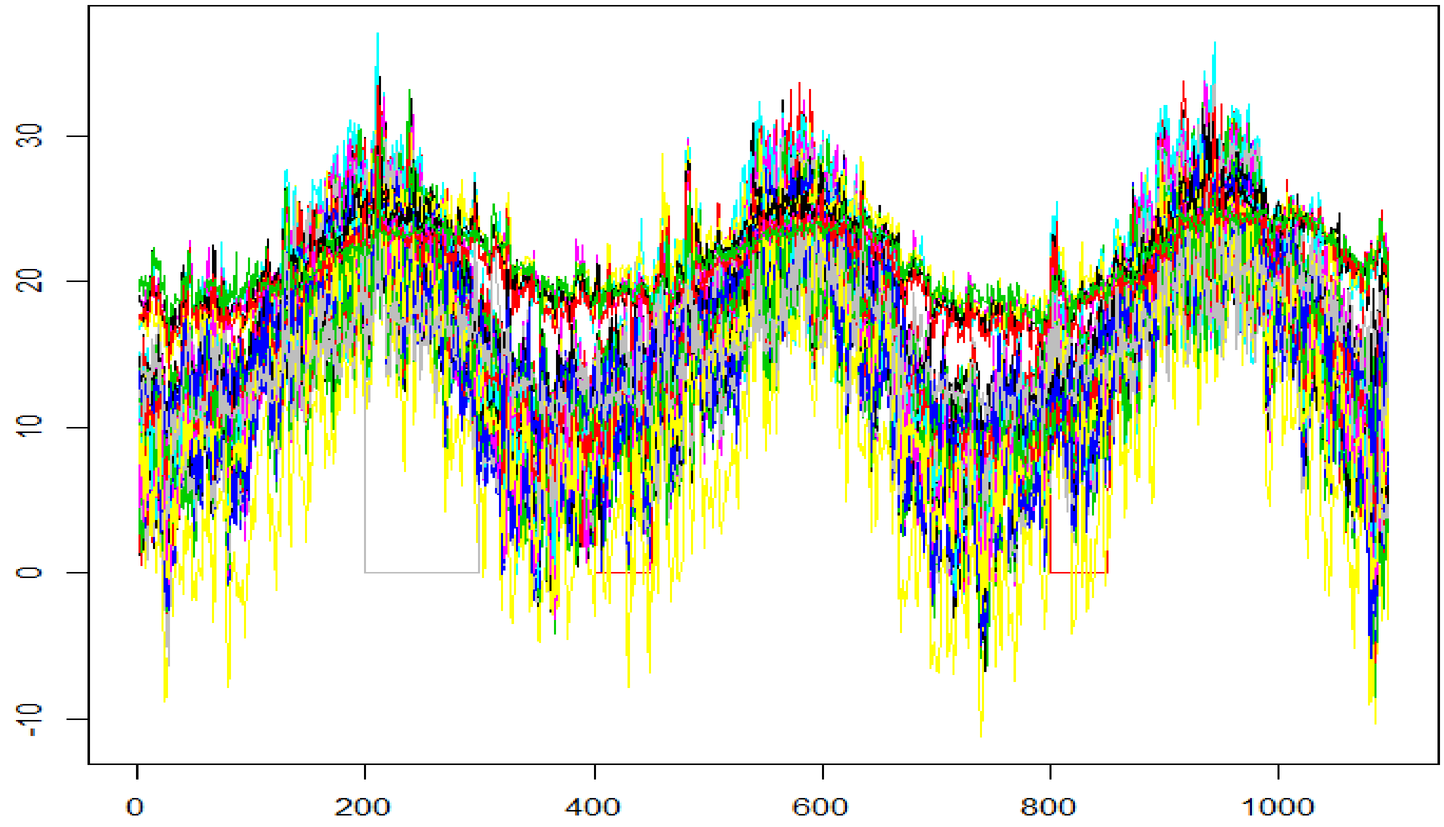
- **Results:**

400
300
200
100
0
n
0
25
50
75
100
p
code
0
1
2

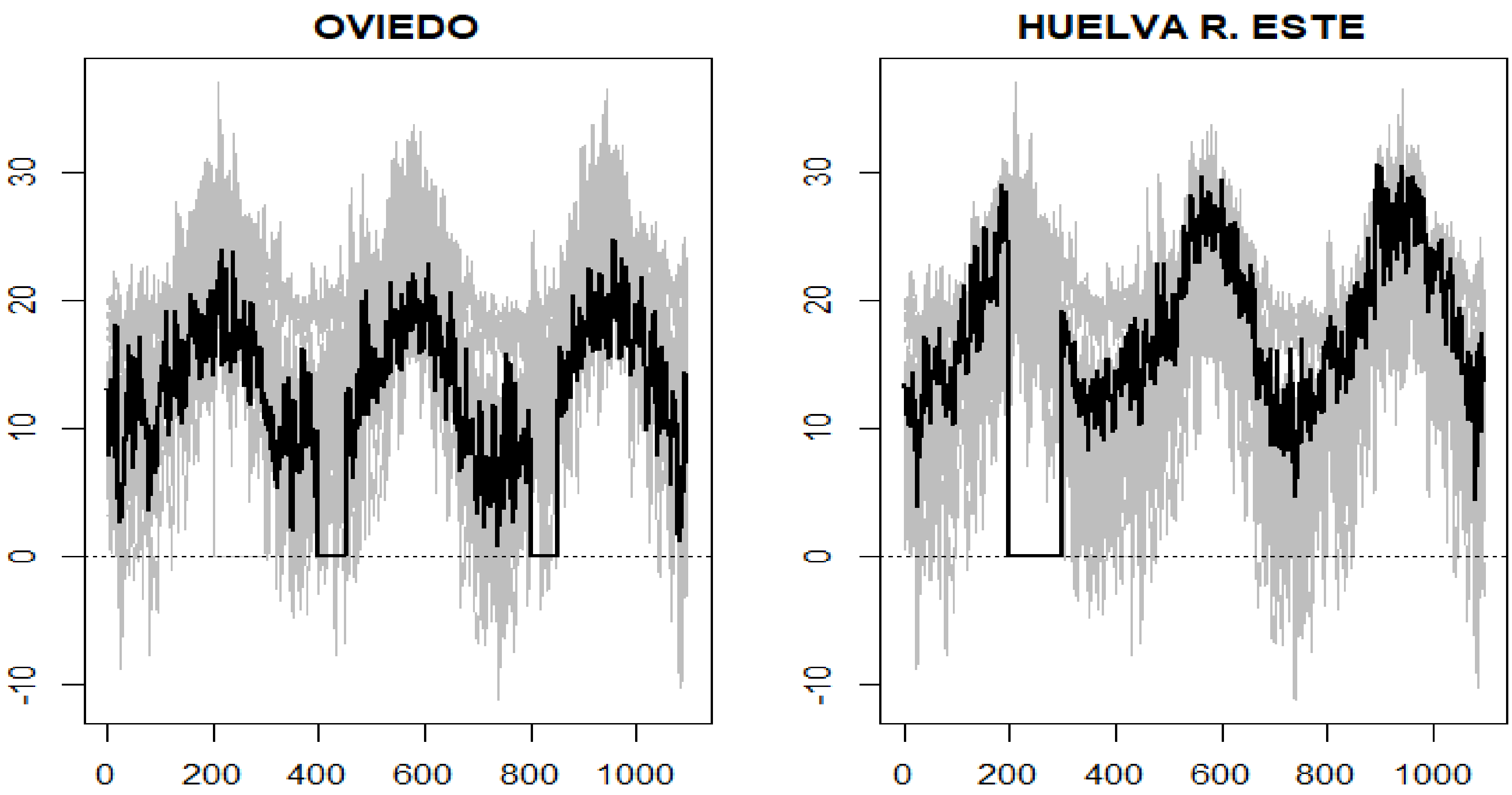


cluster 1
40
20
0
-20
-40
0.0
0.2
0.4
0.6
0.8
1.0
cluster 2
40
20
0
-20
-40
0.0
0.2
0.4
0.6
0.8
1.0

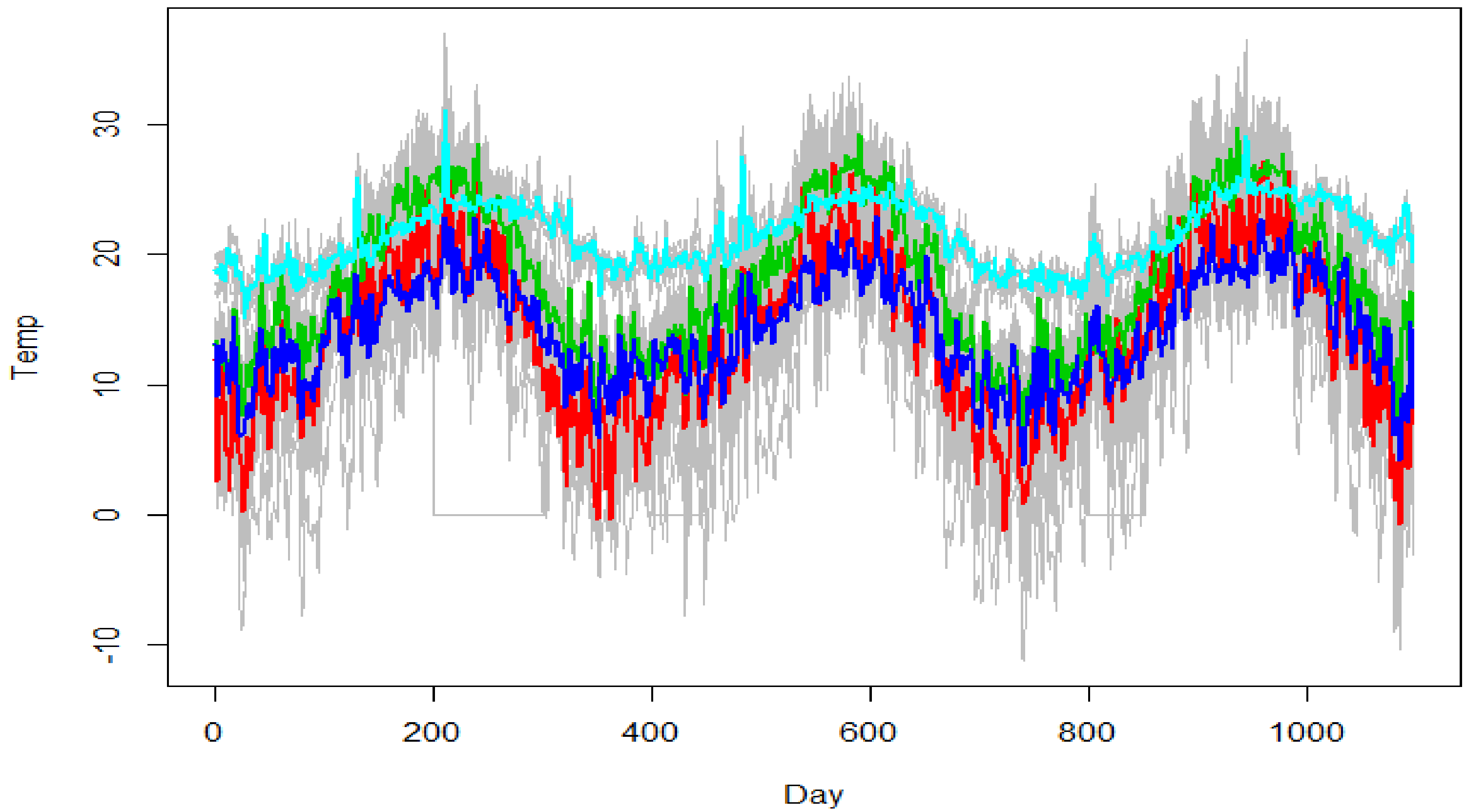
- **Real data example:** Average daily temperatures in 83 Spanish meteorologic stations between 2007-2009 ($n = 83$ and $p = 1096$).



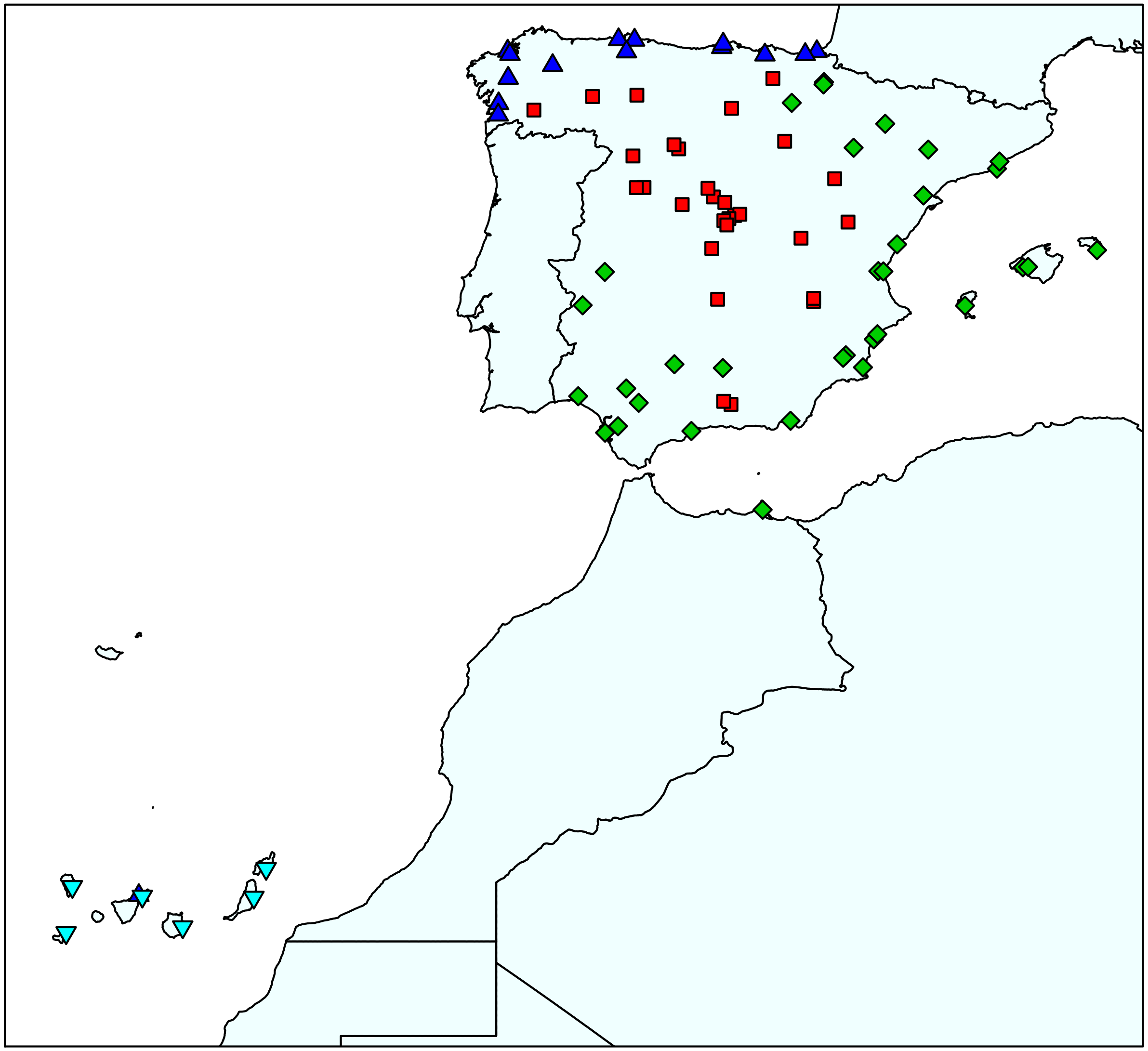
- **Artificial outliers:**
  - ◇ Two periods of 50 consecutive days in *Oviedo* replaced by 0^oC.
  - ◇ 150 consecutive days in *Huelva* temperature replaced by 0^oC.



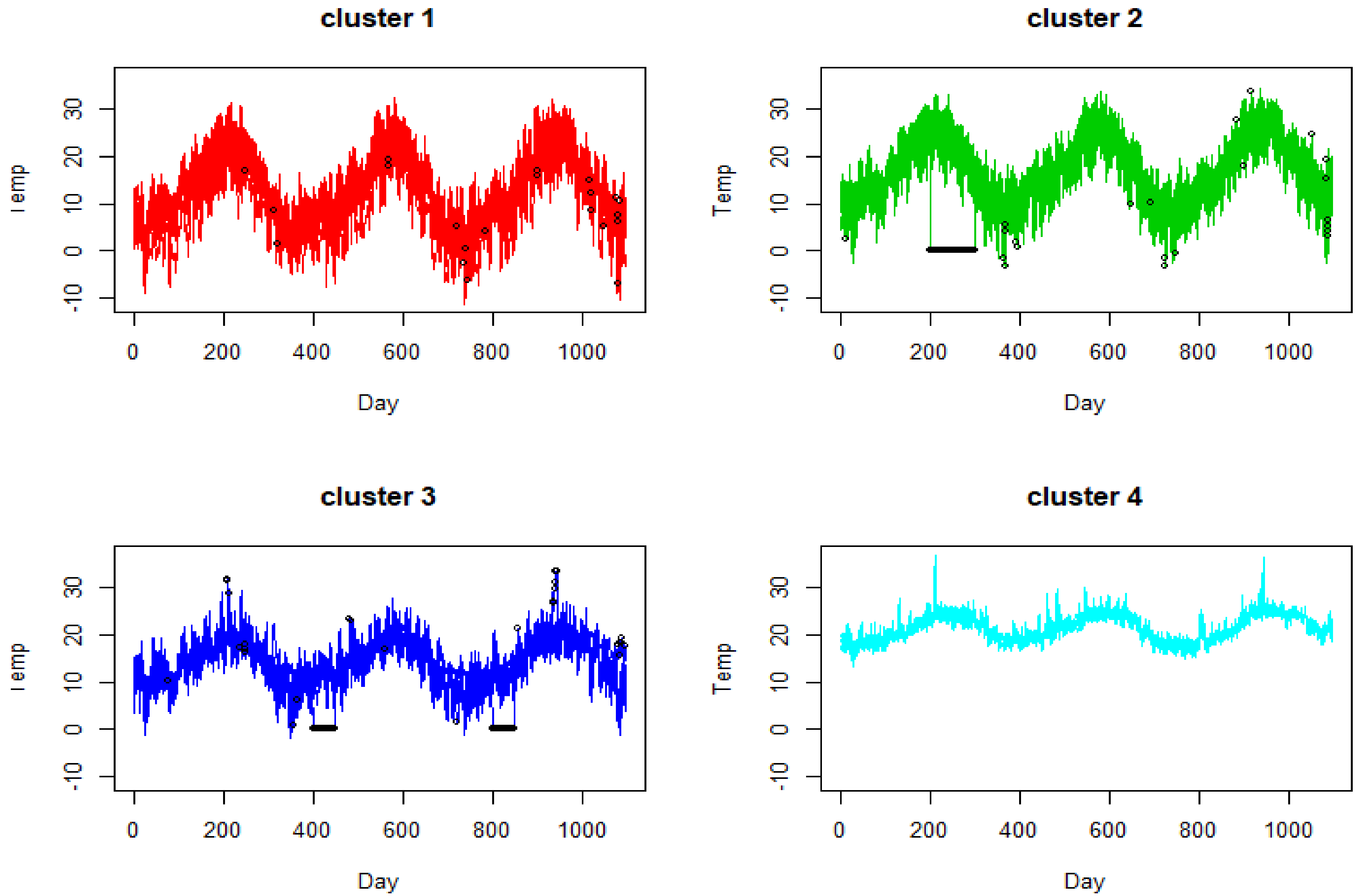
- **Cluster means:**
  - ◇ "Meseta" (Central plateau-Castile): (red line) Mediterranean: (green line)
  - ◇ Cantabrian Coast: (blue line) Canary Islands: (cyan line)

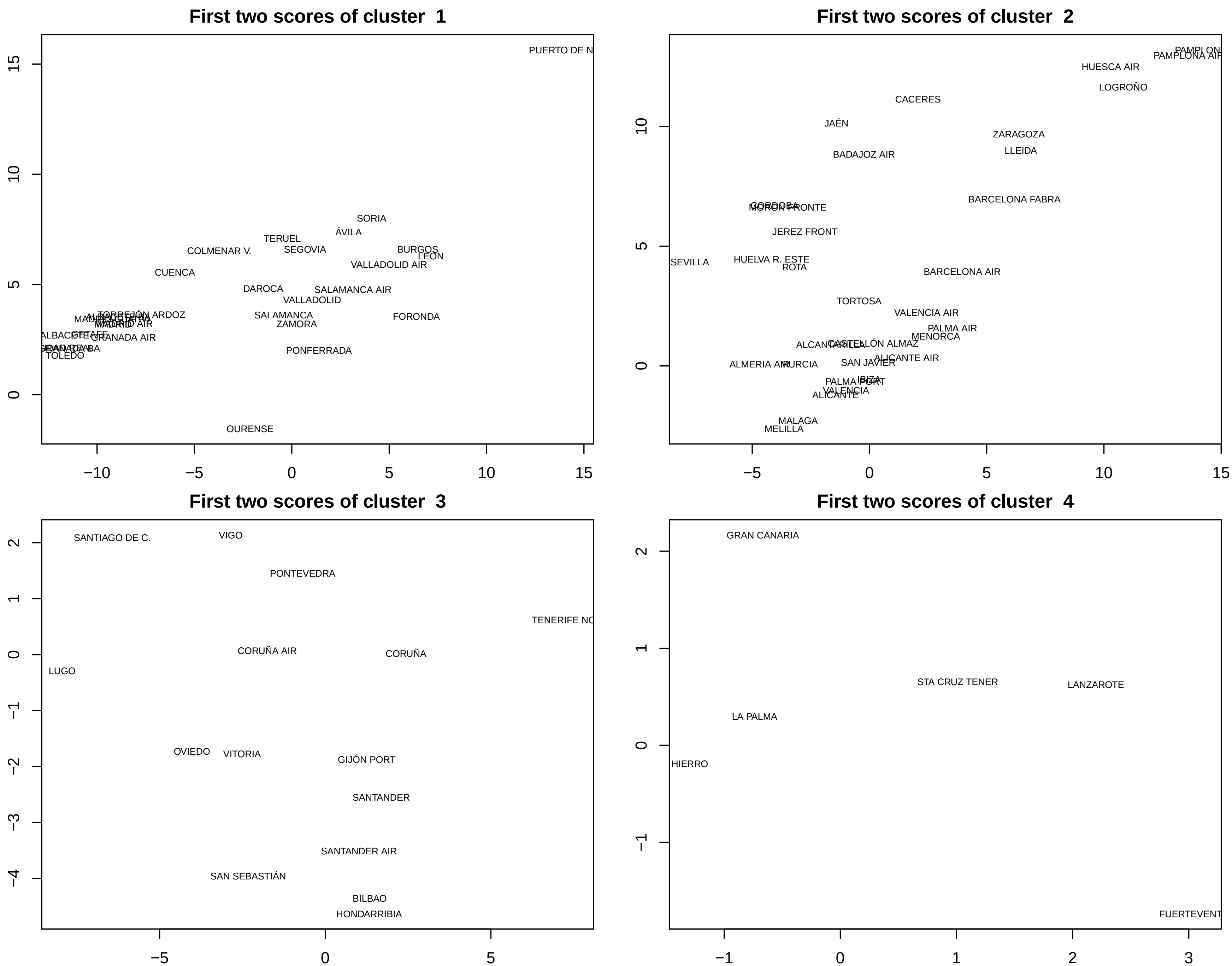


- **Clustered stations:**

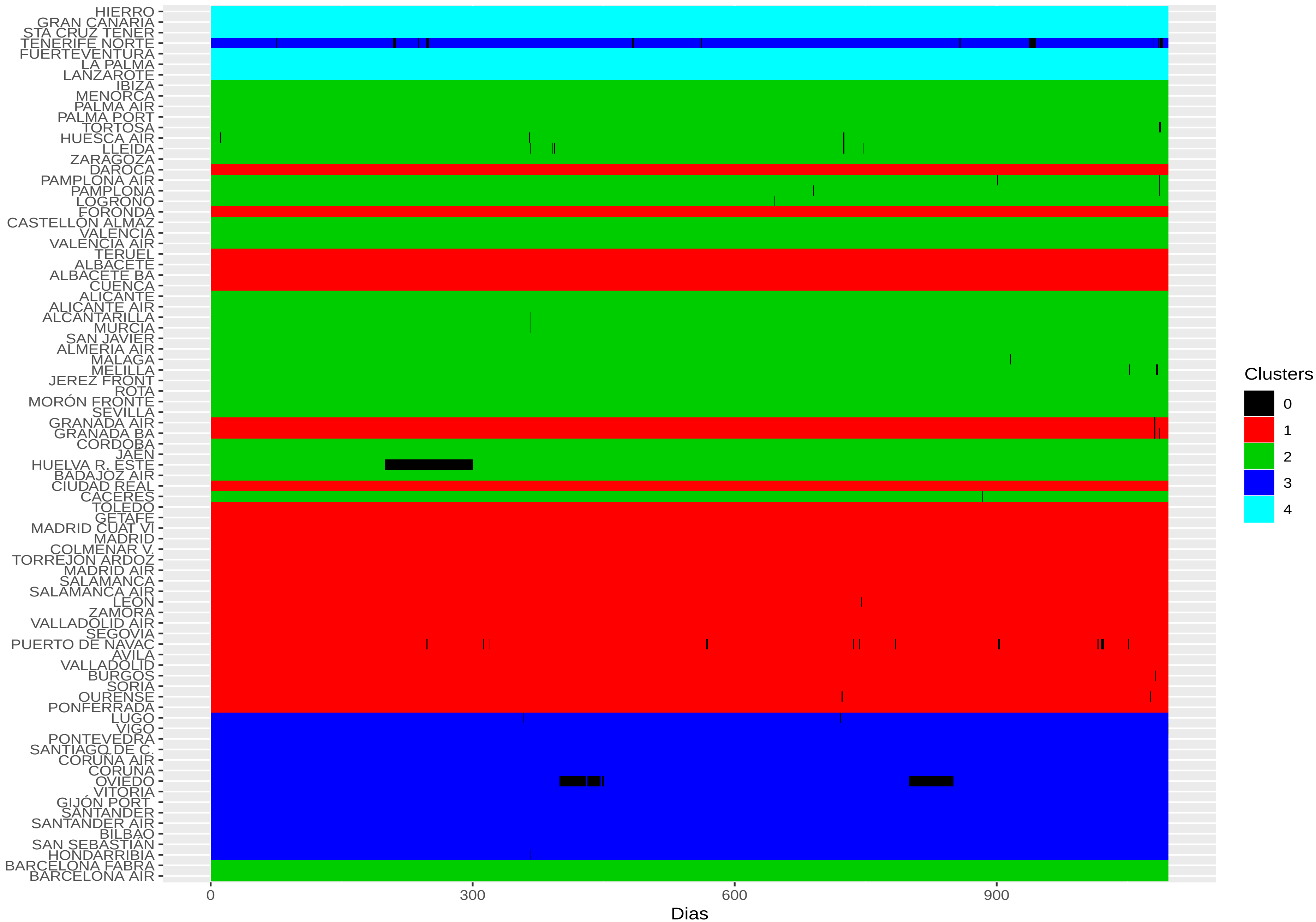


- **Clusters found and trimmed cells:**

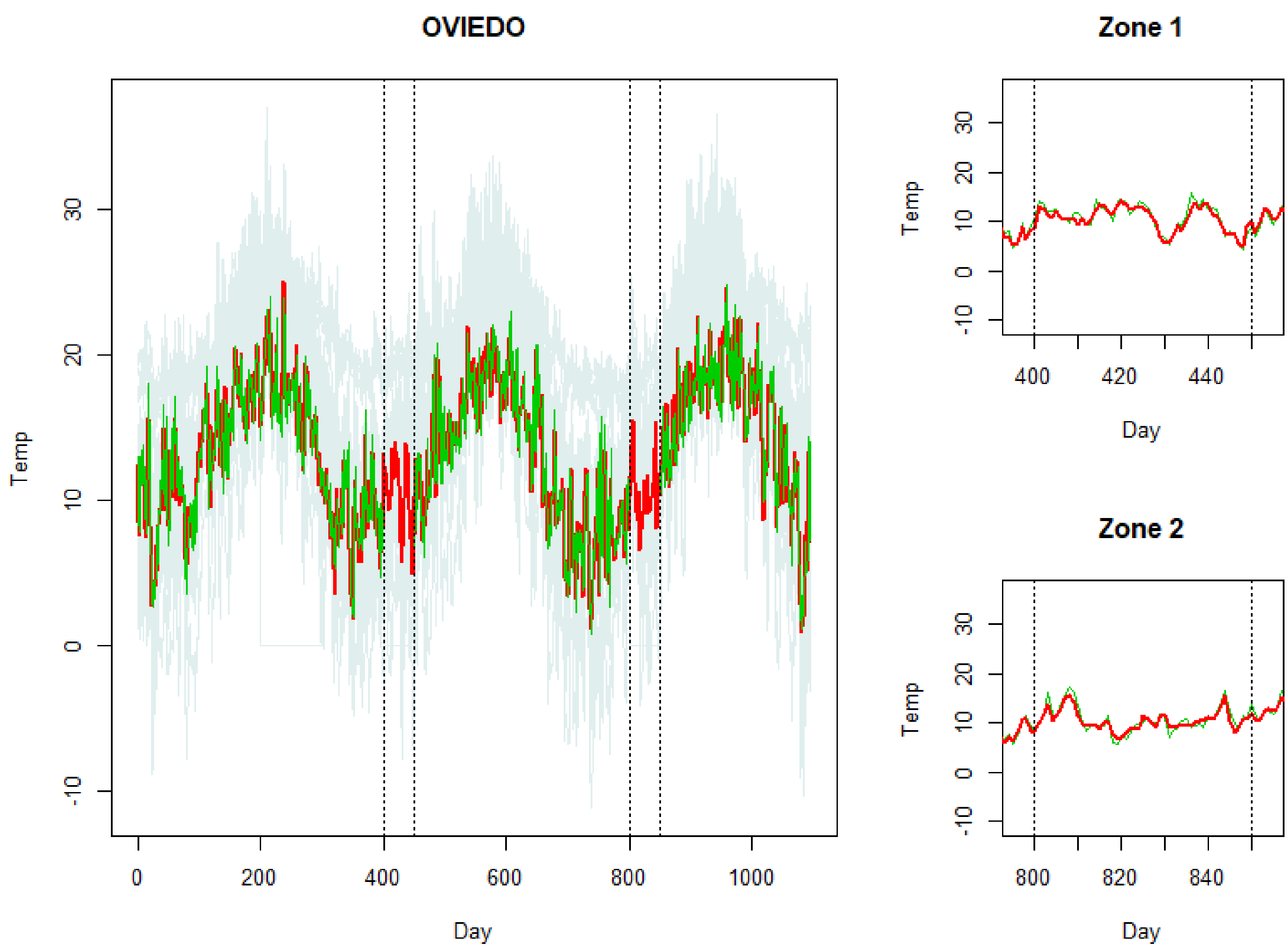




First two scores of cluster 1
PUERTO DE N
SORIA
ÁVILA
TERUEL
SEGOVIA
COLMENAR V.
BURGOS
LEÓN
VALLADOLID AIR
CUENCA
DAROCA
SALAMANCA AIR
VALLADOLID
TORREJÓN ARDOZ
SALAMANCA
FORONDA
ZAMORA
MADRID AIR
MADRID
GETAFE
GRANADA AIR
TOLEDO
PONFERRADA
OURENSE
First two scores of cluster 2
PAMPLONA AIR
HUESCA AIR
LOGROÑO
CACERES
JAÉN
ZARAGOZA
LLEIDA
BADAJOZ AIR
CORDOBA
MORON FRONTE
BARCELONA FABRA
JEREZ FRONT
SEVILLA
HUELVA R. ESTE
ROTA
BARCELONA AIR
TORTOSA
VALENCIA AIR
PALMA AIR
MENORCA
ALCANTARILLA
CASTELLÓN ALMAZ
ALICANTE AIR
SAN JAVIER
ALMERIA AIR
MURCIA
IBIZA
PALMA PORT
VALENCIA
ALICANTE
MALAGA
MELILLA
First two scores of cluster 3
SANTIAGO DE C.
VIGO
PONTEVEDRA
TENERIFE NO
CORUÑA AIR
CORUÑA
LUGO
OVIEDO
VITORIA
GIJÓN PORT
SANTANDER
SANTANDER AIR
SAN SEBASTIÁN
BILBAO
HONDARRIBIA
First two scores of cluster 4
GRAN CANARIA
STA CRUZ TENER
LANZAROTE
LA PALMA
HIERRO
FUERTEVENT


HIERRO
GRAN CANARIA
STA CRUZ TENER
TENERIFE NORTE
FUERTEVENTURA
LA PALMA
LANZAROTE
IBIZA
MENORCA
PALMA AIR
PALMA PORT
TORTOSA
HUESCA AIR
LLEIDA
ZARAGOZA
DAROCA
PAMPLONA AIR
PAMPLONA
LOGROÑO
FORONDA
CASTELLÓN ALMAZ
VALENCIA
VALENCIA AIR
TERUEL
ALBACETE
ALBACETE BA
CUENCA
ALICANTE
ALICANTE AIR
ALCANTARILLA
MURCIA
SAN JAVIER
ALMERIA AIR
MALAGA
MELILLA
JEREZ FRONT
ROTA
MORÓN FRONTE
SEVILLA
GRANADA AIR
GRANADA BA
CORDOBA
JAÉN
HUELVA R. ESTE
BADAJOZ AIR
CIUDAD REAL
CÁCERES
TOLEDO
GETAFE
MADRID CUAT VI
MADRID
COLMENAR V.
TORREJÓN ARDOZ
MADRID AIR
SALAMANCA
SALAMANCA AIR
LEÓN
ZAMORA
VALLADOLID AIR
SEGOVIA
PUERTO DE NAVAC
AVILA
VALLADOLID
BURGOS
SORIA
OURENSE
PONFERRADA
LUGO
VIGO
PONTEVEDRA
SANTIAGO DE C.
CORUÑA AIR
CORUÑA
OVIEDO
VITORIA
GIJÓN PORT
SANTANDER
SANTANDER AIR
BILBAO
SAN SEBASTIÁN
HONDARRIBIA
BARCELONA FABRA
BARCELONA AIR
0
300
600
900
Dias
Clusters
0
1
2
3
4

- **Reconstructed curves** "——" and true real data "——" in Oviedo:



- **Conclusions:**

  - ◇ **Different patterns/structures in data $\Rightarrow$ Cluster Analysis**

  - ◇ **Robust clustering aimed at (jointly) detecting main clusters (bulk of data) and outliers $\Rightarrow$ Potential "frauds"...**

  - ◇ **Higher dimensional problems: Assume clusters "living" in low-dimensional subspaces**

  - ◇ **"Casewise" and "cellwise" trimming**

Some References:

- Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1997), "Trimmed k-means: An attempt to robustify quantizers," *Ann. Statist.*, **25**, 553-576.
- García-Escudero, L.A. and Gordaliza, A. (1999), "Robustness properties of k-means and trimmed k-means," *J. Amer. Statist. Assoc.*, **94**, 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**, 1324-1345.
- García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2010), "A review of robust clustering methods," *Advances in Data Analysis and Classification*, **4**, 89-109.
- Fritz, H.,García-Escudero, L.A. and Mayo-Iscar; A (2012), "`tclust`: An R package for a trimming approach to Cluster Analysis," *Journal of Statistical Software*, **47**, Issue 12.

Thanks for your attention!!!