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:

  ![Diagram showing a sample of data points with a marked mean.]

• $k$ **clusters** $\Rightarrow$ $k$ "data-clouds" $\Rightarrow$ $k$-**means**
- **k-means:** Search for
  - $k$ centers $m_1, ..., m_k$
  - 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:

\[
\bar{x} = \frac{1.72 + 1.67 + 1.80 + 1.70 + 1.82 + 1.73 + 1.78}{7} = 1.745
\]

\[
\bar{x} = \frac{1.72 + 1.67 + 1.80 + 1.70 + 182 + 1.73 + 1.78}{7} = 27.485
\]

◊ *k*-means inherits that **lack of robustness from the mean**
• **Lack of robustness of $k$-means:**

![Graphs showing lack of robustness of $k$-means](image)

(a) 3-means
(b) 2-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 $m_1, \ldots, m_k$ and
  ◦ a partition $\{R_0, R_1, \ldots, R_k\}$ of $\{1, 2, \ldots, n\}$ with $\#R_0 = \lfloor n\alpha \rfloor$

minimizing

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

[A fraction $\alpha$ of data is not taken into account $\leadsto$ Trimmed]
• Black circles: *trimmed points* ($k = 3$ and $\alpha = 0.05$):
Old Faithful Geyser data: $x_1 = \text{“Eruption length”}$, $x_2 = \text{“Previous eruption length”}$ and $n = 271$

$k = 3$ and $\alpha = 0.03$ ($0.03 \cdot 271 \sim 9$ 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)

\[
\phi(x; \mu, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left( - (x - \mu)' \Sigma^{-1} (x - \mu) / 2 \right)
\]
• **Trimmed likelihoods:** Search for

- $k$ centers $m_1, \ldots, m_k$,
- $k$ scatter matrices $S_1, \ldots, S_k$, and,
- a partition \{R_0, R_1, \ldots, R_k\} of \{1, 2, \ldots, n\} with $\#R_0 = \lfloor n\alpha \rfloor$

maximizing

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

---

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

- Control relative axes’ lengths (eigenvalues constraints):

  \[ c = 1 \]

  ![Diagram showing different scales of ellipses with labels](image)

  - Large $c$ value
- The **FSDA** Matlab toolbox:

- The **R** package **tclust** at CRAN repository:
• The R package \texttt{tclust}:
  \begin{itemize}
  \item > library(tclust)
  \item \texttt{tkmeans(data, k, alpha)}
    \begin{itemize}
    \item k = “number of groups”
    \item alpha = “trimming proportion”
    \end{itemize}
  \item \texttt{tclust(data, k, alpha, restr.fact, ...)}
    \begin{itemize}
    \item restr.fact = “Strength of the constraints”
    \end{itemize}
  \end{itemize}
\textbullet \ t\text{clust}(X, k=3, \text{alpha}=0.03, \text{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:

![ CTL-Curves](image)

- **Objective Function Value**
- **Restriction Factor = 50**
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 $h_1, \ldots, h_k$ in $\mathbb{R}^p$
    - a partition $\{R_0, R_1, \ldots, R_k\}$ of $\{1, 2, \ldots, n\}$ with $\#R_0 = \lfloor n\alpha \rfloor$

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

  - $\text{Pr}_h(\cdot)$ denotes the “orthogonal” projection onto the linear subspace $h$
• **Example:** Three linear structures in presence of noise:

Trimmed “mixtures of regressions” can also be applied...
• \( k = 1 \) case \( \Rightarrow \) Robust "Principal Components Analysis (PCA)":

\[\text{PCA provides a } q\text{-dimensional } (q << p) \text{ representation of data by }\]

\[
\min_{B_q, A_q, m} \sum_{i=1}^{n} ||x_i - \hat{x}_i||^2 \text{ for } \\
\hat{x}_i = \Pr_h(x_i) = \hat{x}_i(B_q, A_q, m) = m + B_qa_i
\]

\[\cdot \quad A_q = \begin{pmatrix}
-a_1 \\
\vdots \\
-a_i \\
\vdots \\
-a_n
\end{pmatrix}
\]

is the scores matrix \((n \times q)\)

\[\cdot \quad B_q = \begin{pmatrix}
-b_1 \\
\vdots \\
-b_j \\
\vdots \\
-b_p
\end{pmatrix}
\]

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

\[
\sum_{i=1}^{n} w_i \| \mathbf{x}_i - \hat{\mathbf{x}}_i \|^2 = \sum_{i=1}^{n} w_i \| \mathbf{x}_i - \hat{\mathbf{x}}_i(\mathbf{B}_q, \mathbf{A}_q, m) \|^2,
\]

with \( \{ w_i \}_{i=1}^{n} \) being “0-1 weights” such that

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

◊ **Weights:**

\[
\begin{cases} 
1 & \text{If } \mathbf{x}_i \text{ is not trimmed} \\
0 & \text{If } \mathbf{x}_i \text{ is trimmed}
\end{cases}
\]
• **Cases** → $\mathbf{x}_i = (x_{i1}, ..., x_{ip})' \in \mathbb{R}^p$ and **Cells** → $x_{ij} \in \mathbb{R}$
  
  ◦ $i$ 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 $\mathbf{x}_i$ cases with (at least one) outlying $x_{ij}$

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

  ![Outlying $x_{ij}$ cells](image1)

  ![Trimmed $\mathbf{x}_i$ cases (black lines)](image2)
• But when the dimension $p$ increases... we do not expect many $x_i$ completely free of outlying $x_{ij}$ cells:

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

![Outlying $x_{ij}$ cells](image1)

![Trimmed $x_i$ cases (black lines)](image2)

• **Cellwise trimming:**

  ◊ Only trimming outlying cells... (⇒ “Particular” frauds identified...??)
• PCA approximation $\hat{x}_i = m + B_q a_i = (\hat{x}_{i1}, \ldots, \hat{x}_{ip})^T$ re-written as

$$\hat{x}_{ij} = m_j + a_i^T b_j.$$ 

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

$$\sum_{i=1}^{n} w_{ij} (x_{ij} - m_j - a_i^T b_j)^2$$

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

$$\sum_{i=1}^{n} w_{ij} = [n(1 - \alpha)], \text{ for } j = 1, \ldots, p.$$
• Different patterns/structures in data ⇒ $G$ subspace approximations:

$$\hat{x}_i^g(B_{qg}^g, A_{qg}^g, m^g) = m^g + B_{qg}^g a_i^g \quad \text{or} \quad \hat{x}_{ij}^g = m_j^g + (a_i^g)^T b_j^g,$$

for $g = 1, \ldots, G$

• Minimize

$$\min_{w_{ij}^g, B_{qg}^g, A_{qg}^g, m^g} \sum_{i=1}^n \sum_{j=1}^p \sum_{g=1}^G w_{ij}^g (x_{ij} - \hat{x}_{ij}^g)^2.$$

◊ $w_{ij}^g = 1$ 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:
• **Real data example:** Average daily temperatures in 83 Spanish meteorologic stations between 2007-2009 \((n = 83 \text{ and } p = 1096)\).
• Artificial outliers:
  ◦ Two periods of 50 consecutive days in Oviedo replaced by 0°C.
  ◦ 150 consecutive days in Huelva temperature replaced by 0°C.
• Cluster means:
  ◇ “Meseta” (Central plateau-Castile): red  Mediterranean: green
  ◇ Cantabrian Coast: blue  Canary Islands: cyan
• Clustered stations:
• Clusters found and trimmed cells:
• **Reconstructed curves** “—” and true real data “—” in Oviedo:
• Conclusions:

- Different patterns/structures in data ⇒ Cluster Analysis

- Robust clustering aimed at (jointly) detecting main clusters (bulk of data) and outliers ⇒ Potential “frauds”...

- Higher dimensional problems: Assume clusters “living” in low-dimensional subspaces

- “Casewise” and “cellwise” trimming
Some References:


Thanks for your attention!!!