Model-based methods to identify multiple cluster structures in a data set

Giuliano Galimberti

joint work with Gabriele Soffritti

Dipartimento di Scienze Statistiche
Alma Mater Studiorum - Università di Bologna

Outline

1. Model-based clustering: basic notions
2. Definition of multiple cluster structures
3. Detecting multiple cluster structure: a model-based perspective
4. Some results
5. Concluding remarks
1. Model-based clustering: basic notions

1.1) Cluster analysis

Basic assumption:

The population under study is composed by \( K \) distinct sub-populations or "clusters"

\[ \Rightarrow \quad \text{Cluster membership for units and } K \ (\geq 1) \text{ are not known and need to be discovered} \]
1.2) Model-based cluster analysis

\( \mathbf{X} = (X_1, \ldots, X_m) \) \( m \) observed variables

Basic idea:
Each sub-population is characterized by a different probability distribution for \( \mathbf{X} \)
Conditional probability distribution of \( \mathbf{X} \) given cluster \( k \)

\[ f_k(\mathbf{X} | \boldsymbol{\theta}_k) \quad \text{different parameter values AND/OR} \]
\[ \text{different parametric families} \]

1.3) Finite mixture models for cluster analysis

\( Y \) nominal latent (non observable) variable describing cluster membership

\[ P(Y = k) = \pi_k \quad \text{prior probability of belonging to cluster } k \]
\( (k=1,\ldots,K) \)

Unconditional probability distribution of \( \mathbf{X} \):

(Standard) mixture model:
\[ f(\mathbf{X} | \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{X} | \boldsymbol{\theta}_k) \]
1.4) Posterior probability assignments

\{ \mathbf{x}_i = (x_{i1}, \ldots, x_{im}), \ i = 1, \ldots, n \} \quad \text{random sample}

Model parameter estimation

Maximum likelihood with EM algorithm

Posterior probability estimation:

\[ \hat{\pi}_{ik} = \frac{\hat{\pi}_k f_i(x_i | \hat{\theta}_k)}{\sum_{k=1}^{K} \hat{\pi}_k f_i(x_i | \hat{\theta}_k)} \quad k = 1, \ldots, K \]

Probability of belonging to cluster \( k \) for a unit with values for the observed variables equal to \( \mathbf{x}_i \)

1.5) soft vs. hard clustering

Each sample unit can belong to each cluster, but with different posterior probabilities

Bayes classification rule: unit \( i \) is assigned to cluster \( k^* \) if

\[ \arg\max_k \left( \hat{\pi}_{ik} f_i(x_i | \hat{\theta}_k) \right) = k^* \]

Maximum a posteriori probability classification
1.6) Choice of the number of clusters

Model selection criteria:

- BIC (Bayesian Information Criterion)
- AIC (Akaike Information Criterion)
- CAIC (Consistent Akaike Information Criterion)
- ICL (Integrated Classification Likelihood)

1.7) BIC

Model 1

\[ f_{M1}(\mathbf{X} | \theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{X} | \phi_k) \quad \text{loglikelihood} \quad \log L_{M1} = \sum_{i=1}^{n} f_{M1}(x_i | \theta) \]

Model 2

\[ f_{M2}(\mathbf{X} | \theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{X} | \phi_k) \quad \log L_{M2} = \sum_{i=1}^{n} f_{M2}(x_i | \theta) \]

\[ \text{BIC}_{M1} = 2\log L_{M1} - n \text{par}_{M1} \log n \quad \text{BIC}_{M2} = 2\log L_{M2} - n \text{par}_{M2} \log n \]

\[ \text{Fraley e Raftery 2002} \]

The model with the largest BIC represents the better trade-off between goodness of fit and complexity.
1.8) An example on simulated data (n=250, m=2)

![Simulated data scatter plot]

1.9) Gaussian mixture models

\[ f_k(x | \theta_k) = \frac{\exp \left\{ -\frac{1}{2} (x - \mu_k)^\top \Sigma_k^{-1} (x - \mu_k) \right\} }{(2\pi)^{m/2} \det(\Sigma_k)^{1/2}} \]

- \( \mu_k \): mean vector for the \( k \)-th sub-population
- \( \Sigma_k \): variance-covariance matrix for the \( k \)-th sub-population

*Same parametric family for each component*

*Restrictions on \( \Sigma_k \) can be introduced to reduce the total number of parameters*
1.10) An example on simulated data (n=250, m=2)

Choice of the number of clusters

![Graph showing choice of number of clusters]

1.11) An example on simulated data (n=250, m=2)

Estimated joint density function

![Graphs showing estimated joint density function]
1.10) An example on simulated data (n=250, m=2)

2. Definition of multiple cluster structures
2.1) An introductory example

\[ \{ \mathbf{x}_i = (x_{i1}, \ldots, x_{im}), \ i = 1, \ldots, n \}, \ n = 60, \ m = 4 \]

2.2) Comparison between the partitions in the two subspace (case A)

| Cluster in \((X_{12}, X_{23})\) | 1  | 2  | 3
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \Rightarrow \text{A single cluster structure in} \ \mathbf{X} \]
2.3) Results from standard Gaussian mixture models (case A)

Choice of the number of clusters

2.4) Introductory example: case B

\( \{x_i = (x_{i1}, ..., x_{im}), i = 1, ..., n\}, n = 60, m = 4 \)
### 2.5) Comparison between the partitions in the two subspace (case B)

<table>
<thead>
<tr>
<th>Cluster in ((X_1, X_2))</th>
<th>Cluster in ((X_3, X_4))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 3 \quad \text{p-value} = 0.5578 \]

\[ \Rightarrow \text{two INDEPENDENT cluster structures in } X \]

### 2.6) Results from standard Gaussian mixture models (case B)

#### Choice of the number of clusters

![Graph showing the choice of the number of clusters](image)
2.7) Note on results from standard cluster analysis (case B)

The 9 detected clusters are the results of all possible intersections of the clusters in the two subspaces (this happens when the two partitions are independent).

It can be very hard to disentangle these intersections by simply looking at the final results.

These clusters provide only a partial representation of the relevant features in $\mathbf{X}$ (the two cluster structures).

3. Detecting multiple cluster structures: a model-based perspective
3.1) The proposed solution
(Galimberti e Soffritti, 2007)

Main features:
use of model-based clustering

Basic idea:
recast the multiple cluster structures detection problem as a model selection problem

3.2) Introductory example: a model-based perspective – mixtures of bivariate Gaussian distributions
3.3) A standard mixture model (case A)

Cluster in \((X_1, X_2)\) | Cluster in \((X_3, X_4)\)
--- | --- | ---
1 | \(\pi_1\) | 0 | 0 | \(\pi_1\)
2 | 0 | \(\pi_2\) | 0 | \(\pi_2\)
3 | 0 | 0 | \(\pi_3\) | \(\pi_3\)

\[f(\mathbf{X} | \Theta) = f(X_1, X_2, X_3, X_4 | \Theta) = \sum_{k=1}^{3} \pi_k f_k(\mathbf{X} | \Theta_k)\]

3.4) Mixture models for case B

Cluster in \(A = (X_1, X_2)\) | Cluster in \(B = (X_3, X_4)\)
--- | --- | --- | ---
1 | \(\pi_{1A} \pi_{1B}\) | \(\pi_{1A} \pi_{2B}\) | \(\pi_{1A} \pi_{3B}\) | \(\pi_{1A}\)
2 | \(\pi_{2A} \pi_{1B}\) | \(\pi_{2A} \pi_{2B}\) | \(\pi_{2A} \pi_{3B}\) | \(\pi_{2A}\)
3 | \(\pi_{3A} \pi_{1B}\) | \(\pi_{3A} \pi_{2B}\) | \(\pi_{3A} \pi_{3B}\) | \(\pi_{3A}\)

\[f(X_1, X_2 | \Theta_A) = \sum_{i=1}^{3} \pi_{iA} f(X_2, X_2 | \Theta_i)\] mixture for \((X_1, X_2)\)

\[f(X_3, X_4 | \Theta_B) = \sum_{h=1}^{3} \pi_{hB} f_3(X_3, X_4 | \Theta_h)\] mixture for \((X_3, X_4)\)
3.5) A joint mixture model with two independent cluster structures

Joint model for \((X_1, X_2, X_3, X_4)\)

\[
f(X \mid \theta) = f(X_1, X_2 \mid \theta_A) f(X_3, X_4 \mid \theta_B)
\]

\[
= \left[ \sum_{l=1}^{3} \pi_{lA} f_l(X_1, X_2 \mid \theta_l) \right] \left[ \sum_{h=1}^{3} \pi_{hB} f_h(X_3, X_4 \mid \theta_h) \right]
\]

\[
= \sum_{k=1}^{9} \pi_k f_k(X_1, X_2, X_3, X_4 \mid \theta_k)
\]

model UM

The following constraints hold:

- \(\hat{\theta}_k = \hat{\theta}_1 \cup \hat{\theta}_h\)
- \(\pi_k = \pi_{lA} \pi_{hB}\) for \(l = 1, 2, 3, h = 1, 2, 3\) and \(k = 1, \ldots, 9\)
- \(f_k(X_1, X_2, X_3, X_4 \mid \theta_k) = f_l(X_1, X_2 \mid \theta_l) f_h(X_3, X_4 \mid \theta_h)\)

3.6) BIC for model CM

\[
\log L_{CM} = \sum_{i=1}^{n} \log \left[ f(x_{1i}, x_{2i} \mid \theta_A) f(x_{3i}, x_{4i} \mid \theta_B) \right] =
\]

\[
= \sum_{i=1}^{n} \log f(x_{1i}, x_{2i} \mid \theta_A) + \sum_{i=1}^{n} \log f(x_{3i}, x_{4i} \mid \theta_B) =
\]

\[
= \log L_A + \log L_B
\]

\[
\text{BIC}_{CM} = 2\log L_{CM} - \text{npar}_{CM} \log n
\]

\[
= 2\log L_A + 2\log L_B - \text{npar}_A \log n - \text{npar}_B \log n
\]

\[
= \text{BIC}_A + \text{BIC}_B
\]
3.7) Constrained vs. unconstrained mixture model

Is there a unique cluster structure in in $R^d$ (case A) or two independent cluster structures (case B)?

$\text{BIC}_{CM} > \text{BIC}_{UM} \Rightarrow$ Constrained is better than Unconstrained

$\Rightarrow$ there are two cluster structures in the data

$\text{BIC}_{CM} \leq \text{BIC}_{UM} \Rightarrow$ Unconstrained is better than Constrained

$\Rightarrow$ there is only one cluster structure in the data

3.8) A general formulation of the problem

The best model for the data should be chosen from the set of all possible models:

* 1 model with $m$ independent cluster structures:
  \[ f(X_1, \theta_1) f(X_2, \theta_2) \ldots f(X_m, \theta_m); \]

* \[ \binom{m}{2} \] models with $m-1$ independent cluster structures;

* ...  

* 1 model with a single cluster structure:
  \[ f(X_1, X_2, \ldots, X_m, \theta). \]
3.9) A BIC-based automatic procedure to detect multiple cluster structures

0. \( G_1 = \{X_1\}, ..., G_k = \{X_k\}, ..., G_m = \{X_m\} \quad g=m \) groups of variables
\( \text{BIC}_{G_1}, ..., \text{BIC}_{G_k}, ..., \text{BIC}_{G_m} \quad \text{BIC} \) for each univariate model

1. \( \delta_{j,h} = (\text{BIC}_{G_j} + \text{BIC}_{G_h}) - \text{BIC}_{G_j \cup G_h} \) for \( j, h = 1, ..., g \) (\( j < h \))

2. \( \delta_{A,B} = \min \{ \delta_{j,h} \text{ for } j < h = 1, ..., g \} \)
   * if \( \delta_{A,B} < 0 \) then \( G_A \cup G_B; g \leftarrow g-1; \) return to step 1;
   * otherwise the procedure stops.

3.10) Main properties of the procedure

1. The \( m \) variables are partitioned in \( g \) groups \( G_1, ..., G_g \) (\( 1 \leq g \leq m \)):
   \( G_h \cap G_j = \emptyset \quad \forall \quad h \neq j \quad \text{and} \quad G_1 \cup G_2 \cup ... \cup G_g = \{X_1, ..., X_m\} \).

2. The value of \( g \) gives the number of cluster structures; it is automatically defined by the procedure together with the groups of variables.

3. If \( g > 1 \) then standard methods should be applied not to \( X \) but separately to each group of variables.

4. The stopping rule prevents the analysis from the consideration of irrelevant aggregations.

5. Its computational complexity increases with \( m \) (and \( n \)).
3.11) Further properties of the constrained model

$$\log L_{CM} = \log L_A + \log L_B$$

$$AIC_{CM} = 2\log L_{CM} - 2npar_{CM} = 2(\log L_A + \log L_B) - 2(npar_A + npar_B) = AIC_A + AIC_B$$

$$CAIC_{CM} = 2\log L_{CM} - [\log(n)+1]npar_{CM} = CAIC_A + CAIC_B$$

$$ICL_{CM} = 2BIC_{CM} + 2\sum_{i=1}^{n}\log[\max_{j} \hat{\pi}_{i,j}] = ICL_A + ICL_B$$

3.12) A general automatic procedure to detect multiple cluster structures

0. \(G_1=\{X_1\} \ldots, G_k=\{X_k\}, \ldots, G_m=\{X_m\}\) \(g=m\) groups of variables

\[MSC_{G_1} \ldots, MSC_{G_k} \ldots, MSC_{G_m}\]

where \(MSC = AIC, BIC, CAIC,\) or \(ICL\)

1. \(\delta_{j,h} = (MSC_{G_j} + MSC_{G_h}) - MSC_{G_j \cup G_h}\) for \(j, h = 1, \ldots, g (j < h)\)

2. \(\delta_{A,B} = \min \{\delta_{j,h} \text{ for } j < h = 1, \ldots, g\}\)

* if \(\delta_{A,B} < 0\) then \(G_A \cup G_B; g \leftarrow g-1;\) return to step 1;

* otherwise the procedure stops.
4. Some results

4.1) A Monte Carlo study

2 cluster structures defined in the 4-dimensional space – data generated by Gaussian mixture models

<table>
<thead>
<tr>
<th>Clusters</th>
<th>{X_1, X_2}</th>
<th>Clusters</th>
<th>{X_3, X_4}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\pi_i</td>
<td>1/3 1/3 1/3</td>
<td>\pi_i</td>
<td>1/3 1/3 1/3</td>
</tr>
<tr>
<td>Means</td>
<td>(0, 0) (\varepsilon, \varepsilon) (2\varepsilon, 2\varepsilon)</td>
<td>Means</td>
<td>(0, 0) (0, \varepsilon) (\varepsilon, \varepsilon)</td>
</tr>
</tbody>
</table>

Var-cov: \[
\begin{bmatrix}
\sigma^2 & 0 \\
0 & \sigma^2
\end{bmatrix}
\] within each cluster
4.2) A simulated data set

4.3) Factors considered in the Monte Carlo study

Factors considered in the simulation experiment:

1. separation between clusters: ε
2. heterogeneity within clusters: σ
3. sample size: n
4. model selection criterion: MSC

100 data matrices have been generated and analysed in R (using MCLUST package) for each combination of factors
4.4) Results (using Gaussian mixture models)

![Table showing AIC, BIC, CAIC, and ICL values for different σ values and ε values.]

Percentage of successes in detecting the two cluster structures

4.5) An application to real data

n = 103 italian provinces (Il Sole 24 Ore, 2002)
m = 7 indicators:

- $X_1$: proportion of total hospital beds used for “day hospital”
- $X_2$: nr. of patients hospitalized outside region out of total patients
- $X_3$: nr. of books bought per 100 inhabitants
- $X_4$: nr. of cultural and artistic associations per 100.000 inhabitants
- $X_5$: nr. of cinema tickets bought per inhabitants
- $X_6$: nr. of gymnasiums per 100.000 inhabitants
- $X_7$: nr. of members of the national sport club per 1.000 inhabitants
4.6) Cluster structures detected using BIC + Gaussian mixture models

Groups of variables: 

- $G_1$: $(X_1, X_2)$ 
- $G_2$: $(X_3, X_4, X_5)$ 
- $G_3$: $(X_6, X_7)$ 

BIC for the selected model: -3703.04

4.7) Comparison among cluster structures

<table>
<thead>
<tr>
<th>Health vs. culture</th>
<th>Health vs. sport</th>
<th>Culture vs. sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_1$</td>
<td>$G_2$</td>
<td>$G_3$</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>32</td>
</tr>
</tbody>
</table>

- $\chi^2 = 1.182$ 
- $\chi^2 = 0.369$ 
- $\chi^2 = 0.150$

- p-value $= 0.277$ 
- p-value $= 0.544$ 
- p-value $= 0.699$

- T $= 0.107$ 
- T $= 0.060$ 
- T $= 0.038$
4.8) Results from standard Gaussian mixture models

4 cluster with 35, 12, 40 and 16 provinces

BIC for the selected model: -3845.92

4.9) Comparison between and global and multiple cluster structures

<table>
<thead>
<tr>
<th>Global vs. health</th>
<th>Global vs. culture</th>
<th>Global vs. sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>73</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 9,522 \]

p-value = 0.023

T = 0.231

\[ \chi^2 = 63,582 \]

p-value = 1.01 \times 10^{-13}

T = 0.597

\[ \chi^2 = 27,970 \]

p-value = 3.68 \times 10^{-6}

T = 0.396
5. Concluding remarks

5.1) Concluding remarks

1. When performing cluster analysis, methods able to detect multiple cluster structures SHOULD always be used
2. There are only few contributions on this topics in the statistical literature
3. Comparative analyses of these methods are still lacking
5.2) Some recent developments on the topic

⇒ Detection of multiple cluster structures in two-level data
   (Galimberti and Soffritti, 2009a)

⇒ Introduction of dependencies among the groups of observed variables
   (Galimberti and Soffritti, 2009b)

5.3) Basic references


