Model-Based Cluster Analysis: a Survey

Gilles Celeux

Inria Saclay-Île-de-France, Université Paris-Sud

The data sets

(Dis)Similarity table : matrix D with dimension (n, n)

- Distances
- Dissimilarities
- Similarities

Objects-variables table: matrix X with dimension (n, p)

- p variables measured on n objects
  - quantitative variables : n points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in $\mathbb{R}^p$
  - qualitative variables

Clustering structures

Structures

- Partitions
- Overlapping Classes
- Density classes
- Fuzzy Partitions

$$\mathbf{z} = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \\ z_{41} & z_{42} & z_{43} \\ z_{51} & z_{52} & z_{53} \end{pmatrix}$$

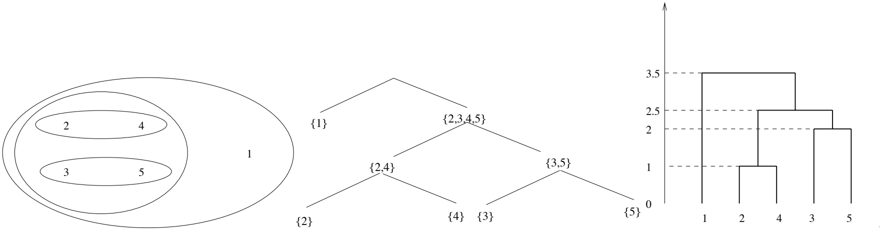
Problem

x				**z**		
3.5	2.3	0.3	4.2	?	?	?
2.2	1.4	2.9	1.3	?	?	?
4.2	1.7	2.2	1.1	?	?	?
2.5	2.3	0.3	4.2	?	?	?
9.2	2.4	2.9	1.3	?	?	?
6.2	1.2	2.2	1.1	?	?	?

Estimation of the clustering matrix **z** from the data **x**

Hierarchy

Sequence of embedded partitions



CAH algorithm

Defining a distance between clusters : agregation criterion

- Single link : $D(A,B) = \min\{d(i,i'), i \in A \text{ et } i' \in B\}$
- Complete link : $D(A,B) = \max\{d(i,i'), i \in A \text{ et } i' \in B\}$
- Mean link : $D(A,B) = \frac{\sum_{i \in A} \sum_{i' \in B} d(i,i')}{n_A . n_B}$

Algorithm

- Initialisation :
  - Stating from the singletons
  - Distances between singletons
- For a number of clusters greater than > 1
  - Merging of the nearest clusters
  - Updating of the distance table

Hierarchy and ultrametric distance

- CAH : algorithmic approach
- Opimal properties ?
- Equivalence between indexed hierarchy ultrametric distance
- Problem : find the ultrametric δ minimising $\Delta(\delta, d)$ where Δ is a dissimilarity measure between distances
- Partial solutions
  - CAH D_{min} : ultrametric smaller than d optimal for any Δ
  - CAH D_{moy} : ultrametric near optimum for Δ such that

$$\Delta(d, \delta) = \sum_{i,i' \in \Omega} (d(i,i') - \delta(i,i'))^2$$

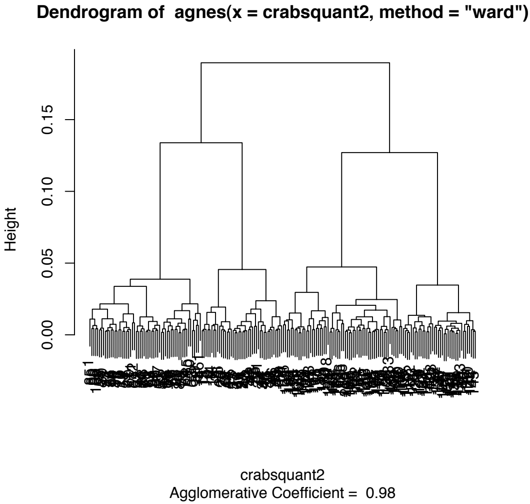
Ward method

- Data : quantitative object-variables and d Euclidian distance
- Algorithm : CAH with Ward criterion

$$D(A,B) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B)$$

- Local optimality : merging the 2 clusters minimising the within-cluster intertia
- Often used with a PCA and the k-means algorithm

An example: crabs data set



Partitions: *k*-means type algorithm

- Within-cluster type inertia criterion :

$$W(P,L) = \sum_{k} \sum_{i \in P_k} ||\boldsymbol{x}_i - \lambda_k||^2$$

  where $L = (\lambda_1, \dots, \lambda_g)$ with $\lambda_k \in \mathbb{R}^p$ (in the standard situation.
- Algorithm: alterned minimisation of W'
- It leads to a stationary sequence of partitions decreasing in $W(P,L)$
- L can take many forms (points, axes, points and distances, densities, ...) to lead to many algorithms.

Motivations of Model-based cluster analysis

Classical clustering

Use of more or less empirical methods being based on metric criteria : k-means, hierarchical algorithm of Ward, ...

Difficulties

- Choice of metric and criterion
- Selection of the method and the number of classes

One solution

Embed clustering in the framework of probabilistic clustering models:

- Objects to be classified: sample of a random vector
- Clustering obtained by analyzing the density of this vector

Different approaches of model-based cluster analysis

Non-parametric

- Multimodality
- Hight-density clusters

Parametric

- Mixture model
- Poisson cluster process in spatial statistics

Finite mixture models: principle

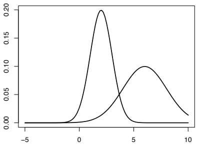
- Mixture model clustering consists of assuming that the data come from a source with several subpopulations.
- Each subpopulation is modeled separately.
- The overall population is a mixture of these subpopulations.
- The resulting model is a finite mixture model.

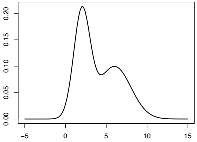
Finite mixture models: definition

The general form of a mixture model with *g* groups is

$$f(\boldsymbol{x}) = \sum_k \pi_k f_k(\boldsymbol{x})$$

- π_k : mixing proportions
- $f_k(.)$: densities of components





The parameterisation of the group densities depends of the nature (continuous or discrete) of the observed data.

Finite mixture models: an hidden structure model

- The mixture model is an incomplete data structure model
- The complete data are

$$\mathbf{y} = (\mathbf{x}, \mathbf{z}) = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = ((\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n))$$

  where the missing data are $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = (z_{ik})$
  - $\boldsymbol{z}_i$ = component of i
  - $z_{ik} = 1$ if i arises from group k and 0 otherwise

$\mathbf{z}$ defines a partition $P = (P_1, \ldots, P_g)$ of the observed data $\mathbf{x}$ with $P_k = \{i \mid z_{ik} = 1\}$.

Finite mixture models: generative model

Knowing

- the proportions $\pi_1, \ldots, \pi_g$ and
- the component distributions f_k ,

data are drawn according to the following scheme

- $z_i \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$ (multinomial distribution)
- $\boldsymbol{x}_i \sim$ distribution of density f_{z_i}

Mixture model for cluster analysis

Two approaches:

Estimation method

- Estimating the mixture parameters
- Computing of t_{ik}, conditional probability that observation $\boldsymbol{x}_i$ comes from cluster k using the estimated parameters.
- Assigning each observation to the cluster maximizing t_{ik} (MAP : Maximum a posteriori)

Clustering approach

Simultaneous estimation of both the mixture parameter and the underlying partition

Quantitative data: multivariate Gaussian Mixture (MGM)

Multidimensional observations $\mathbf{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ in $\mathbb{R}^d$ are assumed to be a sample from a probability distribution with density

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_k \pi_k \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k)$$

where

- π_k : mixing proportions
- $\phi(.|\boldsymbol{\mu}_k, \Sigma_k)$: Gaussian density with mean $\boldsymbol{\mu}_k$ and variance matrix Σ_k.

This is the most popular model for clustering of quantitative data.

Qualitative Data: latent class model (LCM)

- Observations to be classified are described with d qualitative variables.
- Each variable j has m_j response levels.

Data $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are defined by

$$\boldsymbol{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$$

with

$$\begin{cases} x_i^{jh} = 1 & \text{if } i \text{ has response level } h \text{ for variable } j \\ x_i^{jh} = 0 & \text{otherwise.} \end{cases}$$

The standard latent class model (LCM)

Data are supposed to arise from a mixture of g multivariate multinomial distributions with pdf

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_k \pi_k m_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k) = \sum_k \pi_k \prod_{j,h} (\alpha_k^{jh})^{x_i^{jh}}$$

where $\boldsymbol{\theta} = (\pi_1, \dots, \pi_g, \alpha_1^{11}, \dots, \alpha_g^{dm_d})$ is the parameter of the latent class model to be estimated :

- α_k^{jh} : probability that variable j has level h in cluster k,
- π_k : mixing proportions

Latent class model is assuming that the variables are conditionnally independent knowing the latent clusters.

First Interest of MBC

Many versatile or parsimonious models available

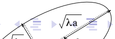
The variance matrix eigenvalue decomposition (1)

Decomposition $\Sigma_k = \lambda_k D_k^t A_k D_k$

- $\lambda_k = |\Sigma_k)|^{1/d}$: component volume
- $D_k =$ matrix of eigenvectors of Σ_k : component orientation
- $A_k =$ diagonal matrix of normalised eigenvalues : component shape

Example in R^2

- D_k rotation matrix defined by θ
- A_k diagonal matrix defined by a and $1/a$
- Equidensity ellipse



28 different models (1)

- For each component : proportion, volume, shape, orientation
- By allowing some of these quantities to vary between components, we get different and easily interpreted models
  - The general family: Assuming equal or free proportions, volumes orientations and shapes leads to 16 models.

    $$[\pi_k \lambda_k D_k A_k], \quad \dots \quad [\pi \lambda D A]$$

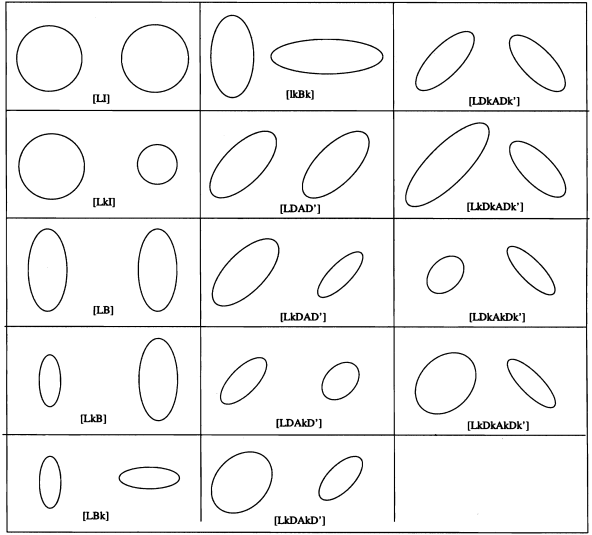
  - The diagonal family: Assuming that the component variances matrices are diagonal leads to 8 models.

    $$[\pi_k \lambda_k B_k], \quad \dots \quad [\pi \lambda B]$$

  - The spherical family: Assuming that the variance matrices are proportional to the identity matrix leads to 4 models.

    $$[\pi_k \lambda_k I] \quad [\pi \lambda_k I] \quad [\pi_k \lambda I] \quad [\pi \lambda I]$$

28 different models(2)



LMC: a Reparameterization

For each cluster k and each variable j :

$$(\alpha_k^{j1}, \ldots, \alpha_k^{jm_j}) \longrightarrow (a_k^{j1}, \ldots, a_k^{jm_j}, \varepsilon_k^{j1}, \ldots, \varepsilon_k^{jm_j})$$

where binary vector $a_k^{j1}, \ldots, a_k^{jm_j}$ provides the mode levels in cluster k for variable j

$$a_k^{jh} = \begin{cases} 1 & \text{if } h = \arg\max_h \alpha_k^{jh} \\ 0 & \text{otherwise} \end{cases}$$

and the ε_k^{jh} can be regarded as scattering values :

$$\varepsilon^{jh} = \begin{cases} 1 - \alpha_k^{jh} & \text{if } a_k^{jh} = 1 \\ \alpha_k^{jh} & \text{if } a_k^{jh} = 0. \end{cases}$$

Example: $(0.7, 0.2, 0.1) \longrightarrow (1, 0, 0,\ \ 0.3, 0.2, 0.1)$.

Five latent class models

Using this form, it is possible to impose various constraints to the scattering parameters ε_k^{jh}.

The models

- $[\varepsilon_k^{jh}]$ (standard latent class model): the scattering is depending upon clusters, variables and levels.
- $[\varepsilon_k^{j}]$: the scattering is depending upon clusters and variables but not upon levels.
- $[\varepsilon_k]$: the scattering is depending upon clusters, but not upon variables.
- $[\varepsilon^{j}]$: the scattering is depending upon variables, but not upon clusters.
- $[\varepsilon]$: the scattering is constant over variables and clusters.

Second interest of MBC

Many algorithms to estimate the mixture model
from different points of view

EM: maximum likelihood estimation

Maximisation of the loglikelihood

$$L(\boldsymbol{\theta}) = \ln\left(\prod_i f(\boldsymbol{x}_i; \boldsymbol{\theta})\right) = \sum_i \ln\left(\sum_k \pi_k \varphi_k(\boldsymbol{x}_i; \alpha_k)\right)$$

The EM algorithm is the reference tool to derive the ML estimates in a mixture model.

Algorithme EM

Algorithm

- **Initial Step** : initial solution $\boldsymbol{\theta}^0$
- **E step**: Compute the conditional probabilities t_{ik} that observation i arises from the kth component for the current value of the mixture parameters:

$$t_{ik}^m = \frac{\pi_k^m \varphi_k(\boldsymbol{x}_i; \alpha_k^m)}{\sum_\ell \pi_\ell^m \varphi_\ell(\boldsymbol{x}_i; \alpha_\ell^m)}$$

- **M step**: Update the mixture parameter estimates maximising the expected value of the completed likelihood. It leads to weight the observation i for group k with the conditional probability t_{ik}.
  - $\pi_k^{m+1} = \frac{1}{n} \sum_i t_{ik}^m$
  - α_k^{m+1} : Solving the Likelihood Equations

EM for the multivariate Gaussian mixture

$$\alpha_k = (\mu_k, \Sigma_k)$$

$$\mu_k^{m+1} = \frac{1}{\sum_i t_{ik}^m} \sum_i t_{ik}^m \boldsymbol{x}_i$$

$$\Sigma_k^{m+1} = \frac{1}{\sum_i t_{ik}^m} \sum_i t_{ik}^m (\boldsymbol{x}_i - \mu_k^{m+1})(\boldsymbol{x}_i - \mu_k^{m+1})'$$

Features of EM

- EM is increasing the likelihood at each iteration
- Under regularity conditions, convergence towards the unique consistent solution of likelihood equations
- Easy to program
- Good practical behavior
- Slow convergence situations (especially for mixtures with overlapping components)
- Many local maxima or even saddle points
- Quite popular: see the McLachlan and Krishnan book (1997)

Classification EM

The CEM algorithm, clustering version of EM, estimate both the mixture parameters and the labels by maximizing the completed likelihood

$$L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{k,i} z_{ik} \log \pi_k f(\mathbf{x}_i; \alpha_k)$$

Algorithm

- **E step**: Compute the conditional probabilities t_{ik} that observation i arises from the kth component for the current value of the mixture parameters.
- **C step**: Assign each observation i to the component maximising the conditional probability t_{ik} (MAP principle).
- **M step**: Update the mixture parameter estimates maximising the completed likelihood.

Features of CEM

- CEM aims maximising the complete likelihood where the component label of each sample point is included in the data set.
- Contrary to EM, CEM converges in a finite number of iterations
- CEM provides biased estimates of the mixture parameters.
- CEM is a *K-means*-like algorithm.

CEM and standard clustering algorithms

model	distance	criterion	remarks
$\pi, \lambda I$	$d^2(\boldsymbol{x}_i, \mu_k)$	$\mathrm{trace}(W)$	k-means, (Ward, 1963
$\pi, \lambda_k I$	$\frac{d^2(\boldsymbol{x}_i,\mu_k)}{\lambda_k} + d\ln(\lambda_k)$	$\sum_k n_k \ln \mathrm{tr}(\frac{W_k}{n_k})$	Scott & Symons 1971
$\pi, \lambda B$	$d^2_{B^{-1}}(\boldsymbol{x}_i, \mu_k)$	$\mathrm{diag}(W)$	classification $+$ weight
π, Σ	$d^2_{\Sigma^{-1}}(\boldsymbol{x}_i, \mu_k)$	$\lvert W \rvert$	Friedman & Rubin, 196

$$W = \sum_{i,k} z_{ik}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)'$$

Stochastic EM

- **E step**: Compute the conditional probabilities t_{ik} that object i arises from the kth component for the current value of the mixture parameters.
- **S step**: Assign each object i at random to one of the component according to the distribution defined by $(t_{i1}, \ldots, t_{ig}$.
- **M step**: Update the mixture parameter estimates maximising the completed likelihood.

Features of SEM

- SEM generates a Markov chain whose stationary distribution is (more or less) concentrated around the ML parameter estimator.
- Thus a natural parameter estimate from a SEM sequence is the mean of the iterates values obtain after a burn-in period (SEMmean).
- An alternative estimate is to consider the parameter value leading to the largest likelihood in a SEM sequence (SEMmax).
- Different variants (Monte Carlo EM, Simulated Annealing EM) are possible.

Third interest of MBC

Finite mixture models can be compared
and assessed in an objective way

Model selection

- Choosing a parsimonious model in a collection of models.
- The problem is to solve the bias-variance dilemma.
  - A too simple model leads to a large approximation error.
  - A too complex model leads to a large estimation error.
- Standard criteria of model selection are AIC and BIC criteria.
- Both criteria are penalized likelihood criteria

The AIC criterion

AIC is approximating the deviance of a model m with ν_m free parameters

$$d(\mathbf{x}) = 2[\log \mathbf{p}(\mathbf{x}) - \log \mathbf{p}(\mathbf{x}|m, \hat{\theta}_m)]$$

on a single test observation X. The penalization is an estimation of $nD(X) - E(d(\mathbf{x}))$ where

$$D(X) = 2E[\log p(X) - \log p(X|m, \hat{\theta}_m)]$$

is the expected deviance on X.

Assuming that the data arose from a distribution belonging to the collection of models in competition, AIC is

$$\text{AIC}(m) = 2 \log \mathbf{p}(\mathbf{x}|m, \hat{\theta}_m) - 2\nu_m.$$

The BIC criterion

BIC is a pseudo-Bayesian criterion. It is approximating the integrated likelihood of the data

$$\mathbf{p}(\mathbf{x}|m) = \int \mathbf{p}(\mathbf{x}|m,\theta_m)\pi(\theta_m)d\theta_m,$$

$\pi(\theta_m)$ being a prior distribution for parameter θ_m.

BIC is

$$\mathrm{BIC}(m) = \log \mathbf{p}(\mathbf{x}|m,\hat{\theta}_m) - \frac{\nu_m}{2}\log(n).$$

This approximation is appropriate when one and only one of the competing models is true.

Choosing a mixture model in a density estimation context

Despite theoretical difficulties in the mixture context

- Simulation experiments (see Roeder & Wasserman 1997) show that BIC works well at a practical level to choose a sensible Gaussian mixture model,
- See also the good performances of a cross-validated likelihood criterion proposed by Smyth (2000).

Choosing a clustering model

Since BIC does not take into account the clustering purpose for assessing m, BIC has a tendency to overestimate g regardless of the separation of the clusters.

The ICL criterion

ICL is a BIC-like approximation of the integrated completed likelihood

$$\mathbf{p}(\mathbf{x}, \mathbf{z} \mid m) = \int_{\Theta_m} \mathbf{p}(\mathbf{x}, \mathbf{z} \mid m, \theta)\pi(\theta \mid m) d\theta,$$

$$\text{ICL}(m) = \log \mathbf{p}(\mathbf{x}, \hat{\mathbf{z}} | m, \hat{\theta}) - \frac{\nu_m}{2} \log n,$$

where the missing data have been replaced by their most probable value for parameter estimate $\hat{\theta}$.

Roughly speaking criterion ICL is the criterion BIC penalized by the estimated mean entropy

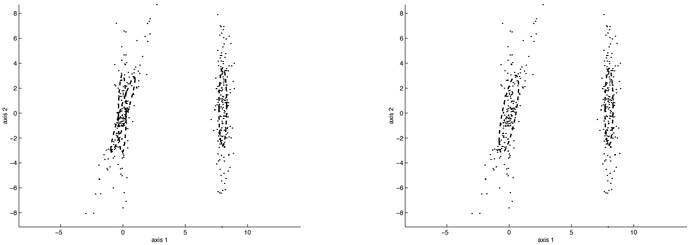
$$E(m) = -\sum_{k,i} t_{ik}^m \log t_{ik}^m \geq 0.$$

Behavior of the ICL criterion

Because of this additional entropy term, ICL favors model giving rise to partitioning the data with the greatest evidence.

- ICL appears to provide a stable and reliable estimate of g for real data sets and also for simulated data sets from mixtures when the components are not too much overlapping.
- But ICL, which is not aiming to discover the true number of mixture components, can underestimate the number of components for simulated data arising from mixture with poorly separated components.

Contrasting BIC and ICL



Typical solutions proposed by BIC (left) (92%) and ICL (right) (88%) with the following features: Gaussian mixture with free variance matrices, $n = 400$.
The criteria select g and the form of the variance matrices from their eigenvalue decomposition.

- BIC outperforms ICL from the density estimation point of view...

Fourth interest of MBC

Special questions can be tackled in a proper way in the MBC context

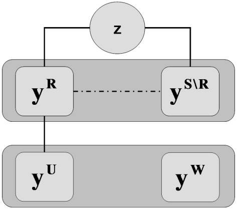
Variables selection

- Model (Maugis, Celeux and Martin-Magniette 2009):

$$\mathbf{x} \in \mathbb{R}^Q \mapsto f_{\text{clust}}(\mathbf{x}^S|g, m, \alpha)\; f_{\text{reg}}(\mathbf{x}^U|r, a+\mathbf{x}^R\beta, \Omega)\; f_{\text{ind}}(\mathbf{x}^W|\ell, \gamma, \tau)$$

  - clustering variables (S): Gaussian Mixture density
  - redundant variables (U): linear regression of $\mathbf{x}^U$ with $\mathbf{x}^R$
  - independent variables (W): Gaussian density

- Graphical representation:



Robust Cluster Analysis

- Using multivariate Student distributions instead of Multivariate Gaussian distributions lead to attenuate the influence of outliers (McLachlan & Peel 2000).
- Including in the mixture a group from a uniform distribution allows to take into account noisy data (DasGupta & Raftery 1998).
- Shrinking the group variance matrix in a proper way (see for instance Ciuperca, Idier & Ridolfi 2002)

Some other special questions solved with MBC

- Imposing specific constraints as groups with known distributions.
- Dealing with missing data at random in a proper way (Hunt & Basford 1999, 2001).
- Mixture of Factor Analyses could be efficient to deal with high dimensional data sets.
- Simple and efficient models in semi-supervised Classification (see for instance Ganesalingam & McLachlan 1978, . . .).
- Assuming the variables are conditionally independent knowing the groups makes valid, in a simple way, the treatment of continuous and discrete data with the same MBC.

Softwares

- Many free softwares for finite mixture analysis are available.
- Some of them (EMMIX, MCLUST) are more devoted to a multidimensional context in a cluster analysis or classification purpose.

MIXMOD software

- Since 2001, C++ library including most of the features described in this talk, Matlab interface
- Website: www.mixmod.org
- In 2012, mixmodGUI: a Graphical User Interface for MIXMOD
- In 2012, Rmixmod package: a set of functions to use MIXMOD in R environment

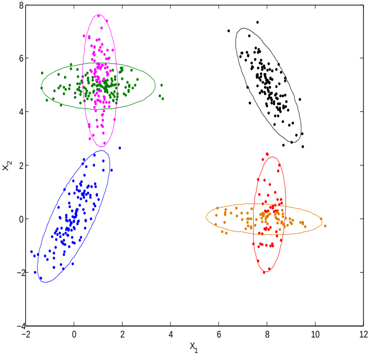
Combining components for clustering: mixtures of mixtures

Limits of ICL

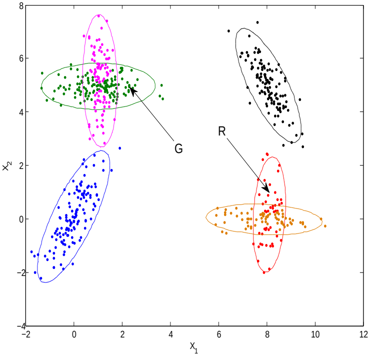
- ICL criterion takes into account the clustering purpose, but still lies on the principle '1 cluster = 1 mixture component'.
- ICL provides a more relevant clustering, but the fit is degraded in comparison with BIC.

An answer: combining components

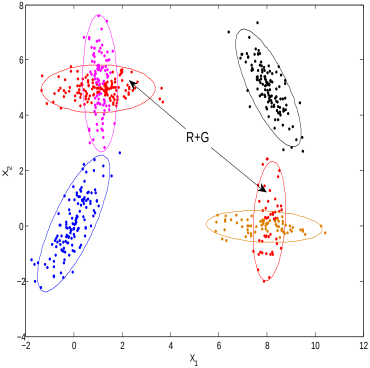
- Start from the BIC solution
- Combine components by merging iteratively the two clusters with the highest entropy.

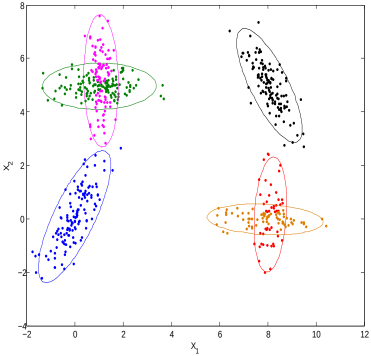


Starting from the BIC solution...

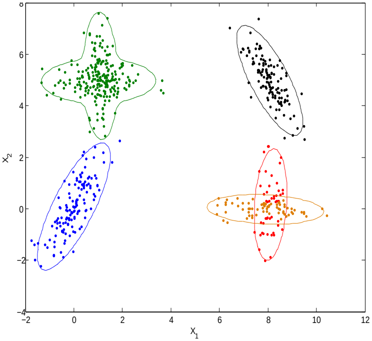


2 clusters are chosen, which one try to combine...


R+G
X_1
X_2
−2
0
2
4
6
8
10
12
−4
−2
0
2
4
6
8

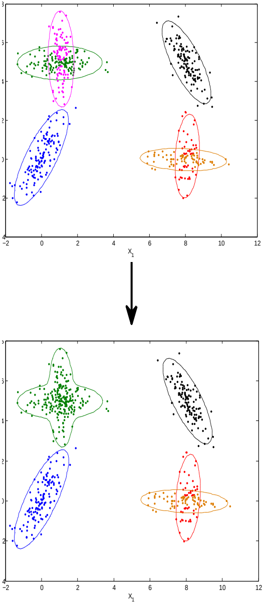


The same is done with each pair of clusters.



The pair of clusters for which the resulting entropy is minimal is actually combined.

This is our 5 clusters solution.



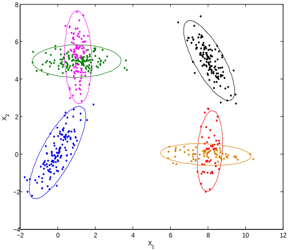
We get a mixture of mixtures:

$$f^{P\cup G}(\cdot) = \sum_{k\neq P,G} \pi_k f_k(\cdot)$$

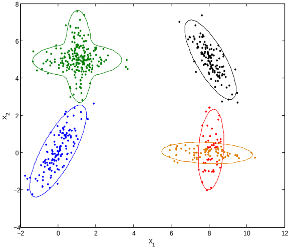
$$+ \pi_{P\cup G}\, f_{P\cup G}(\cdot)$$

$$\text{with} \quad \begin{aligned} & f_k(\cdot) = \phi(\cdot;\hat{\mu}_k,\hat{\Sigma}_k) \quad \forall k, \\ & \pi_{P\cup G} = \pi_P + \pi_G, \\ & f_{P\cup G} = \tfrac{\pi_P}{\pi_P+\pi_G} f_P + \tfrac{\pi_G}{\pi_P+\pi_G} f_G. \end{aligned}$$

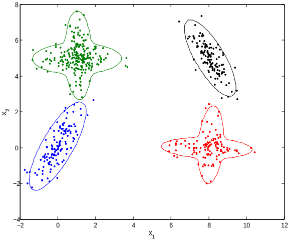
From this new 5-clusters solution, 2 clusters may be combined to get a 4-clusters solution, and so on...



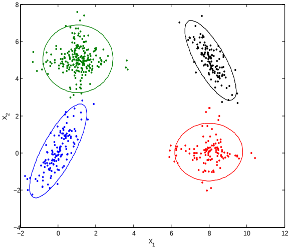
BIC solution: 6-clusters



Combined solution: 5-clusters



Combined solution: 4-clusters



ICL solution: 4-clusters

Solutions obtained by combining:

- provide a good fit, good approximation properties of the Gaussian mixture models;
- allow to choose the number of clusters.

The number of clusters can be chosen from an elbow rule on the entropy criterion.