The impact of high dimension on clustering

Gilles Celeux

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

Cluster Analysis

Cluster analysis aims to discover homogeneous clusters in a data set.

Data sets

- (Dis)Similarity table : matrix D with dimension (n, n)
- Objects-variables table: matrix X with dimension (n, d)

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

Large dimensions

- We are concerned with large n and d objects-variables tables
- We restrict attention to partitions

Outline of the talk

First, three families of methods are discussed

- Standard geometrical k-means-like methods (data analysis community)
- Model-based clustering methods (statistics community)
- Spectral clustering (machine learning community)

Second, the Latent Block Model which is a specific model for summarizing large tables will be considered.

Partitions: k-means type algorithm

- Within-cluster type inertia criterion :

$$W(C, L) = \sum_k \sum_{i \in C_k} ||\boldsymbol{x}_i - \lambda_k||^2$$

  where $L = (\lambda_1, \ldots, \lambda_g)$ 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 $W(C, L)$
- L can take many forms (points, axes, points and distances, densities, ...) to lead to many algorithms.
- For the standard k-means algorithm, λ_k is the center of cluster C_k

Features of the k-means algorithm

k-means is simple

- The k-means algorithms converges (rapidly) in a finite number of iterations.
- Cluster summary is parsimonious.
- It is the most popular clustering method.

k-means is not versatile

- The standard k-means algorithm has a tendency to provide spherical clusters, with equal sizes and volumes.
- Many local optimal solutions.
- Variable selection procedures for k-means are unrealistic and poor: a variable has to be relevant or independent of the clustering.

Model-based clustering

Finite Mixture Model

The general form of a mixture model with g components is

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

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

Each mixture component is associated to a cluster

- The parametrisation of the cluster densities depends of the nature of the data. Typically:
- quantitative data: multivariate Gaussian mixture,
- qualitative data: multinomial latent class model.

Quantitative data: multivariate Gaussian Mixture (MGM)

Multidimensional observations $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ 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 $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \alpha_1^{11}, \ldots, \alpha_g^{dm_d})$ 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.

The advantages of model-based clustering

Model-based clustering provides a solid ground to answer to the cluster analysis problems.

- Many efficient algorithms to estimate the model parameters.
- Choosing the number of clusters can be achieved with relevant penalized information criteria (BIC, ICL)
- Those criteria are also helpful to choose a relevant model with a fixed number of clusters.
- Defining the possible roles of the variables can be achieved properly (relevant, redundant and independent variables).
- Efficient softwares: http://www.mixmod.org
- Specific situations can be dealt with efficiently. Examples:
  - taking missing data into account
  - robust analysis
  - hidden Markov models for depending data

EM algorithm (maximum likelihood estimation)

Algorithm

- **Initial Step** : initial solution θ^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

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 behaviour
- 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 maximising the completed likelihood

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

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.

Model-based clustering via EM

Relevant clustering can be deduced from EM

- Estimating the mixture parameters with EM
- 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 maximising t_{ik} (MAP : Maximum a posteriori)

This strategy could be preferred since CEM provides biased estimates of the mixture parameters.
But CEM is doing the job for well-separated mixture components.

Penalized Likelihood Selection Criteria

The BIC criterion

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

BIC works well to choose a model in a density estimation context

The ICL criterion

$$\mathrm{ICL}(m) = \mathrm{BIC}(m) - \sum_{k,i} t_{ik}^m \log t_{ik}^m.$$

ICL is focussing on the clustering purpose and favoring mixtures with well separated components.

Drawbacks of Model-based Cluster Analysis

Model-based clustering is not tailored to deal with large data sets.

- MBC makes use of versatile models which are too complex for large dimensions
- Algorithmic difficulties increase dramatically with the dimension
- Since all models are wrong, penalised likelihood criteria as BIC become inefficient for large sample sizes.
- Choosing a model cannot be independent of the modelling purpose

Solutions exist to attenuate these problems:

- restrict attention to parsimonious models
- prefer CEM to EM algorithm
- prefer ICL to BIC to select a model

An antinomic approach: Spectral Clustering

Spectral Clustering is based on non directed similarity graph $G = (V, E)$, (s_{ij}) such that

- The vertices V are the objects.
- There is an edge between two objects i and j if $s_{ij} > 0$.
- A weighted adjacency matrix W w_{ij} is associated to s_{ij}.

We define

- the degree of edge i as $d_i = \sum_j w_{ij}$,
- D is the diagonal matrix $(d_i, i = 1, \ldots, n)$,
- for $A \in V$, $|A| = \text{card}A$, $\text{vol}(A) = \sum_{i \in A} d_i$.

The connected components of G define a partition of V.

Which similarities ?

- all points whose pairwise distances are smaller threshold ε are connected, then $w_{ij} = 1$.
- the connected points are k nearest neighbor symmetrised, then $w_{ij} = s_{ij}$.
- the connected points are mutual k nearest neighbor, then$w_{ij} = s_{ij}$.
- a Gaussian similarity

$$s_{ij} = \exp[-\frac{||x_i - x_j||^2}{2\sigma^2}].$$

  is chosen.

The tuning parameters ε, k or σ are sensitive. . . as the choice of g.

Laplacian graphs

- Non normalised Laplacian graph :

$$L = D - W$$

- Symetrised Laplacian graph

$$L_s = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$

- Random walk Laplacian graph

$$L_r = D^{-1} L = I - D^{-1} W$$

Spectral clustering Algorithm

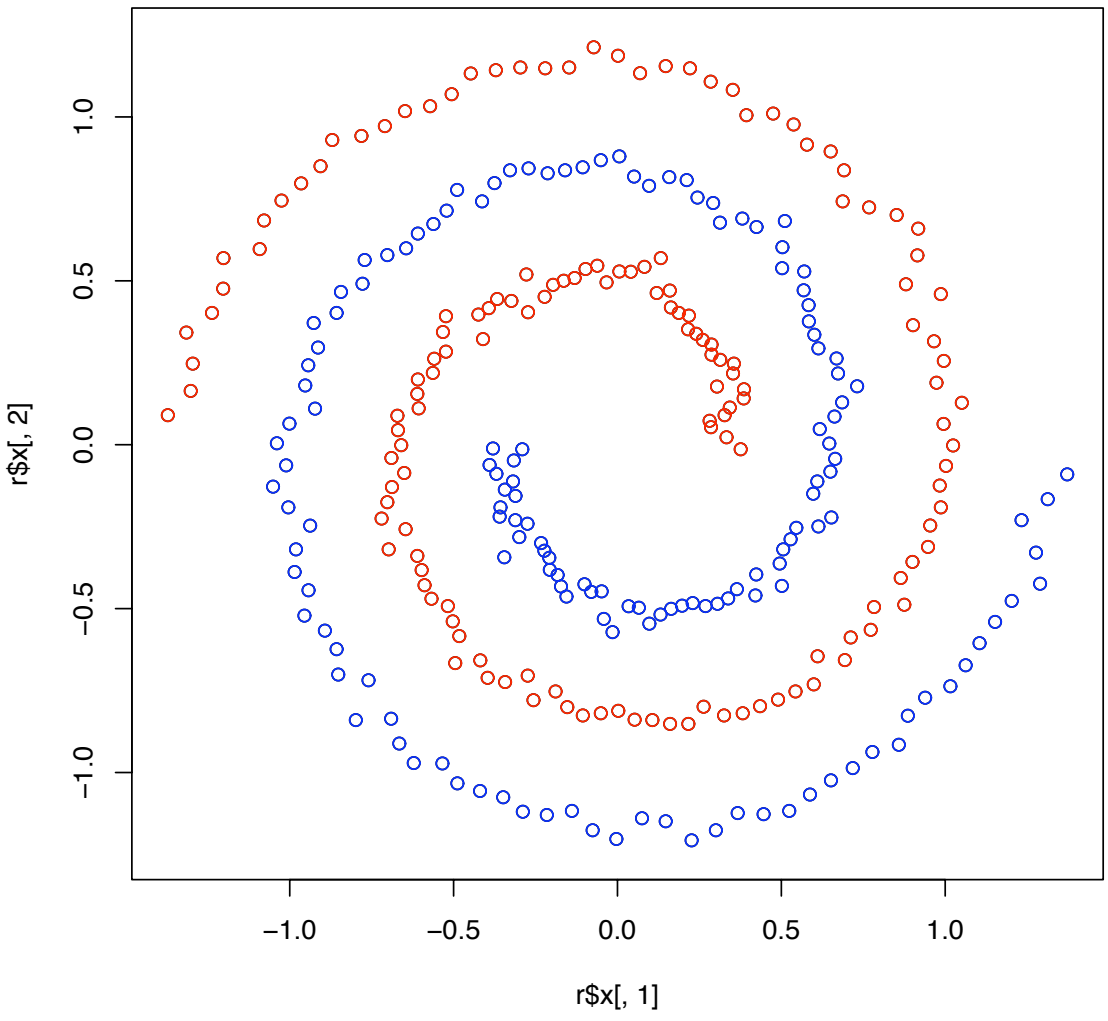
Input The similarity matrix S and the number of clusters g.

- Construct the similarity graph and the related matrix W.
- Compute the chosen Laplacian matrix L.
- Compute the first g eigenvectors of L, $Lu = \lambda u$.
- Let U the matrix of the g first eigenvectors.
- For $i = 1, \ldots, n$, let $y_i \in R^g$ be the vector corresponding to the row i of U.
- Cluster the y_is with k-means in g clusters $C_1, \ldots, C_g$.

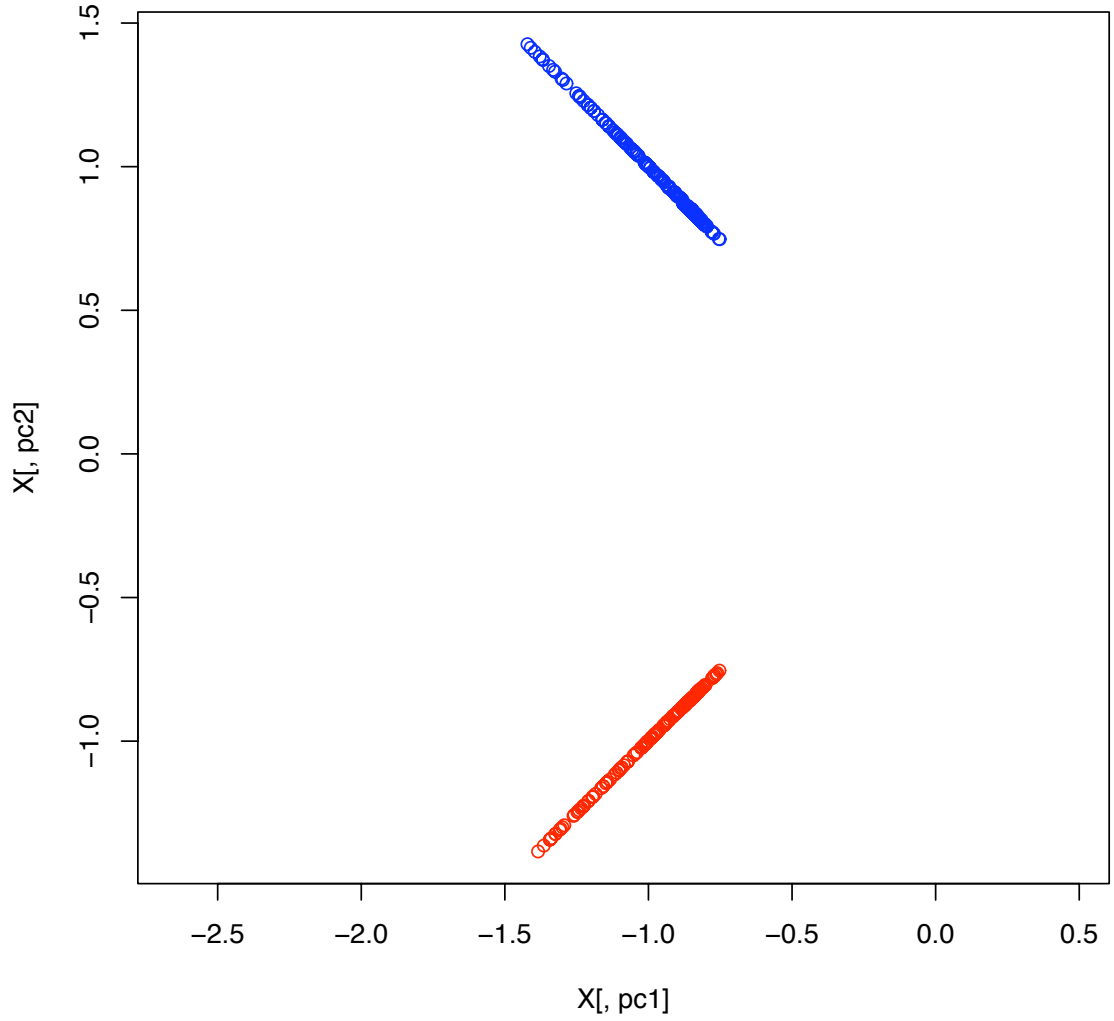
Output The g clusters $C_1, \ldots, C_g$ with $C_\ell = \{i \ / \ y_i \in C_\ell\}$.

Spectral clustering is attractive for large tables because efficient programs are available to find the eigenvectors of large sparse matrices.

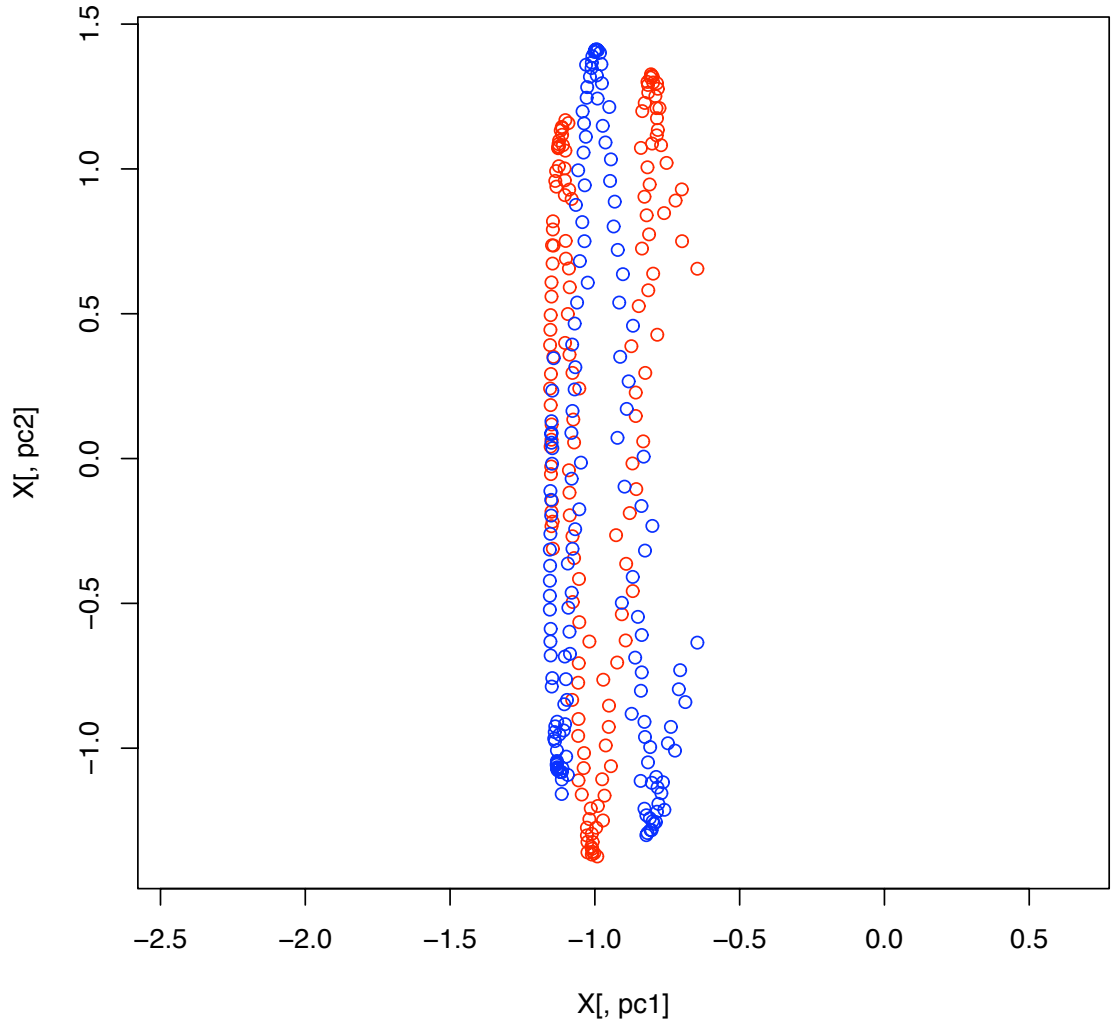
An example: the data



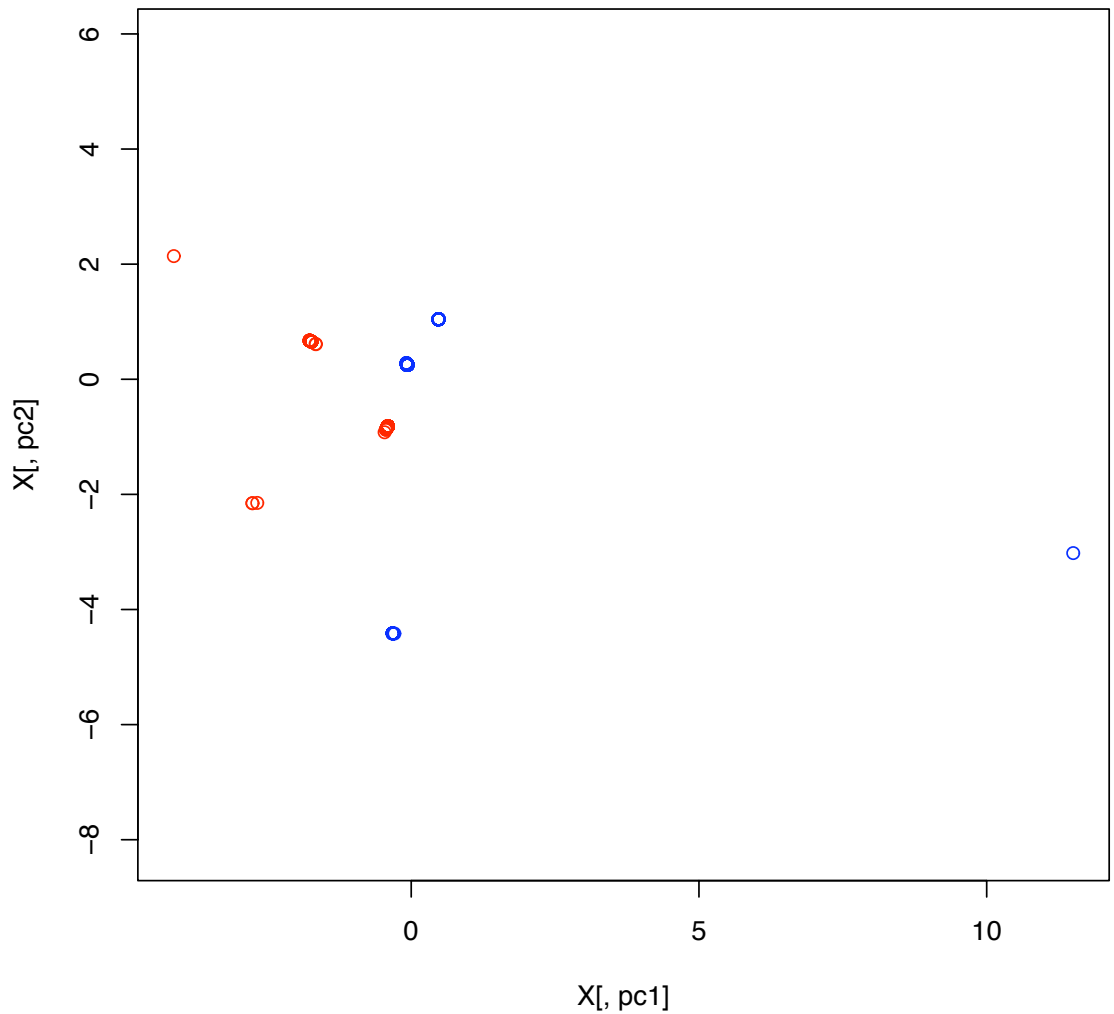
Representation before k means with $\sigma = 300$



Representation before k means with $\sigma = 3$



Representation before k means with $\sigma = 3000$



Block clustering setting

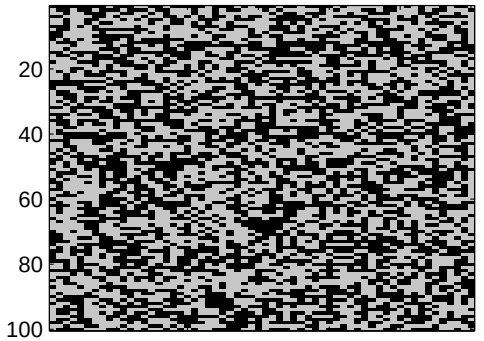
Block clustering of data

- Let $\mathbf{x} = \{(x_{ij}); i \in I, j \in J\}$ be a dimension $n \times d$ matrix, where I is a set n objets and J a set of d variables
- Block clustering of $\mathbf{x}$ is aiming to find a clustering structure on $I \times J$.

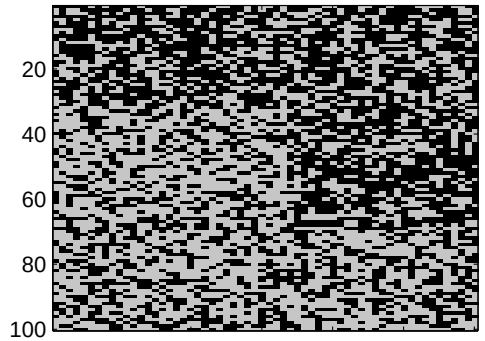
A dramatically parsimonious data representation

- Block clustering is summarizing a data set of nd numbers with gm numbers with $g << n$ and $m << d$.
- This representation is appealing to deal with huge data sets arising in recommendation systems, genomic data analysis, text mining,...

An illustration with binary data



Initial data



Reorganized data


Reorganized and averaged data


Summary

Model-based clustering framework

- Assume that the data are arising from a finite mixture of parametrised densities.
- A cluster is made by observations arising from the same density.
- In a block clustering model, clusters are defined on blocks $\in I \times J$.
- In a block clustering model, data of a block are modelled by the same unidimensional density.

Latent block mixture model

Density of the observed data is supposed to be

$$f(\mathbf{x}|g,m,\phi,\alpha) = \sum_{\mathbf{u}\in\mathcal{U}} p(\mathbf{u}|g,m,\phi) f(\mathbf{x}|g,m,\mathbf{u},\alpha)$$

where $\mathbf{u}$ is the indicator block vector.
It is assumed that $u_{ijb} = z_{ik} w_{j\ell}$, $\mathbf{z}$ (resp.$\mathbf{w}$) being the row (resp. column) cluster indicator vector.
Assuming that the $n \times d$ variables Y_{ij} are conditionnally independent knowing $\mathbf{z}$ and $\mathbf{w}$ leads to the model

$$f(\mathbf{x}|g,m,\pi,\rho,\alpha) = \sum_{\mathbf{z},\mathbf{w}\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(y_{ij}|g,m,\alpha_{k\ell})$$

An exemple: Bernoulli latent block model

Mixing proportions

For fixed g, the mixing proportions for the row are $\pi_1, \ldots, \pi_g$.
For fixed m, the mixing proportions for the col. are $\rho_1, \ldots, \rho_m$.

The Bernoulli density per block

$$\varphi(y_{ij}; \alpha_{k\ell}) = (\alpha_{k\ell})^{y_{ij}}(1 - \alpha_{k\ell})^{1-y_{ij}}$$

where $\alpha_{k\ell} \in (0, 1)$.

The mixture density is

$$f(\mathbf{x}|g, m, \pi, \rho, \alpha) = \sum_{\mathbf{z},\mathbf{w} \in \mathcal{Z} \times \mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} (\alpha_{k\ell})^{y_{ij}}(1-\alpha_{k\ell})^{1-y_{ij}}.$$

The $g + m + gm$ parameters to be estimated are the πs, the ρs and the αs.

Maximum likelihood estimation

The EM algorithm is making use of the conditional expectation of the complete loglikelihood

$$Q(\theta|\theta^{(c)}) = \sum_{i,k} s_{ik}^{(c)} \log \pi_k + \sum_{j,\ell} t_{j\ell}^{(c)} \log \rho_\ell + \sum_{i,j,k,\ell} e_{i,j,k,\ell}^{(c)} \log \varphi(x_{ij}; \alpha_{k\ell})$$

where

$$s_{ik}^{(c)} = P(Z_{ik} = 1|\theta^{(c)}, \mathbf{x}), \quad t_{j\ell}^{(c)} = P(W_{j\ell} = 1|\theta^{(c)}, \mathbf{x})$$

and

$$e_{i,j,k,\ell}^{(c)} = P(Z_{ik} W_{j\ell} = 1|\theta^{(c)}, \mathbf{x}).$$

$\hookrightarrow$ Difficulty to compute $e_{i,j,k,\ell}^{(c)}$... Approximations are needed.

Variational approximation of EM (VEM)

It is assumed that

$$e_{i,j,k,\ell}^{(c)} = s_{ik}^{(c)} w_{j\ell}^{(c)}$$

Thus the VEM algorithm is

Govaert and Nadif (2008)

1. E step:
   - 1.1 computing s_{ik} with fixed $w_{jl}^{(c)}$ and $\theta^{(c)}$
   - 1.2 computing w_{jl} with fixed $s_{ik}^{(c+1)}$ and $\theta^{(c)}$
     $\hookrightarrow s^{(c+1)}$ and $w^{(c+1)}$
2. M step: Updating $\theta^{(c+1)}$

Some features of VEM

- The optimised free energy $\mathcal{F}(q_{zw}, \theta)$ is a lower bound of the observed loglikelihood.
- The parameter maximising the free energy could be expected to be a good, if not consistent, approximation of the maximum likelihood estimator.
- Since VEM is minimising $KL(q_{zw}||p(\mathbf{z}, \mathbf{w}|\mathbf{x}; \theta))$ rather than $KL(p(\mathbf{z}, \mathbf{w}|\mathbf{x}; \theta)||q_{zw})$, it is expected to be sensitive to starting values.

The SEM-Gibbs algorithm

SEM

The SEM algorithm: After the E step, a S step is introduced to simulate the missing data according to the distribution $p(\mathbf{z}, \mathbf{w}|\mathbf{x}; \theta^{(c)})$.
A difficulty for the latent block model is to simulate $p(\mathbf{z}, \mathbf{w}| <; \theta)$.

Gibbs sampling

The distribution $p(\mathbf{z}, \mathbf{w}|\mathbf{x}; \theta^{(c)})$ is simulated using a Gibbs sampler. Repeat

Simulate $\mathbf{z}^{(c+1)}$ according to $p(\mathbf{z}|\mathbf{x}, \mathbf{w}^{(c)}; \theta^{(c)})$
Simulate $\mathbf{w}^{(c+1)}$ according to $p(\mathbf{w}|\mathbf{x}, \mathbf{z}^{(c+1)}; \theta^{(c)})$

$\hookrightarrow$ The stationary distribution of the Markov chain is $p(\mathbf{z}, \mathbf{w}|\mathbf{x}; \theta^{(c)})$

SEM features

- SEM is not increasing the loglikelihood at each iteration.
- SEM is generating an irreductible Markov chain with a unique stationary distribution.
- The parameter estimates fluctuate around the ml estimate
  $\hookrightarrow$ A natural estimator of $\theta, \mathbf{z}, \mathbf{w}$ is the mean of $(\theta^{(c)}, \mathbf{z}^{(c)}, \mathbf{w}^{(c)}; c = B, \ldots, B + C)$ get after a burn-in period.

Discussion: VEM vs. SEM

Numerical comparisons lead to the conclusions

- VEM leads rapidly to reasonable parameter estimates when its initial position is near enough the ml estimation.
- VEM is quite sensitive to starting values.
- SEM-Gibbs is (essentially) insensitive to starting values.

$\hookrightarrow$ Coupling SEM and VEM is beneficial to derive sensible ml estimates for the latent block model.

Model selection

Choosing relevant number of clusters in a latent block model is of crucial importance.

Two good news

- The integrated completed likelihood ICL is closed form.
- For n and d large enough, the entropy of the couple of partitions $(\mathbf{z}, \mathbf{w})$ is close to 0. Thus BIC $\approx$ ICL.

Choosing the missing labels

Let $\hat{\theta}$ be the ml estimate of the LBM parameter derived from the SEM-VEM algorithm.

- The missing labels are replaced by

$$(\hat{\mathbf{z}}, \hat{\mathbf{w}}) = \arg\max_{(\mathbf{z},\mathbf{w})} p(\mathbf{z}, \mathbf{w}|\mathbf{x}; g, m, \hat{\theta}).$$

- An alternating optimisation algorithm is used to compute $(\hat{\mathbf{z}}, \hat{\mathbf{w}})$:
  Repeat until convergence
  - $\mathbf{z}^{(c)} = \arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}; g, m, \hat{\theta}, \mathbf{w}^{(c-1)})$
  - $\mathbf{w}^{(c)} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{x}; g, m, \hat{\theta}, \mathbf{z}^{(c)})$.

Discussion on the Latent Block Model

Interest of LBM

- LBM is a parsimonious but crude model.
- To be efficient for large table, g and m have to large.
- The point is to summarize large tables in small tables

Difficulties with the LBM

- In this perspective, empty clusters is a severe issue.
- Bayesian regularisation through Variational Bayes inference could be relevant.
- In a full Bayesian analysis, dealing with the label switching problem for large g and m remains challenging.