Block Clustering models and algorithms

Mohamed Nadif

Université Paris Descartes, France

Outline

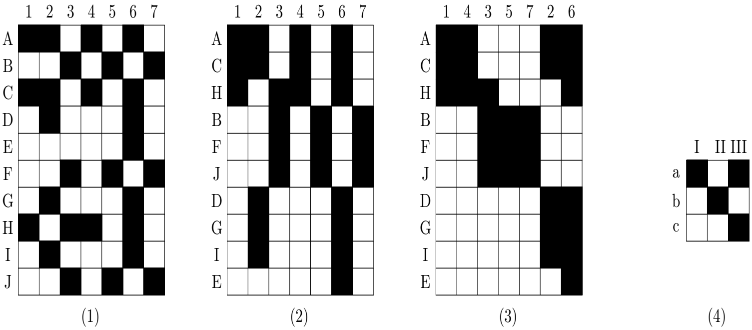
Simultaneous clustering on both dimensions

- They have attracted much attention in recent years
- The problem of block clustering had an increasing influence in applied mathematics (Jennings, 1968)
- Referred in the literature as bi-clustering, co-clustering, direct clustering,...
  - no-overlapping co-clustering
  - overlapping co-clustering
- First works in J.A. Hartigan, Direct Clustering of a Data Matrix, J. Am. Statistical Assoc. (JASA), vol. 67, no. 337, pp. 123-129, 1972.
- Different approaches are proposed: they differ in the pattern they seek and the types of data they apply to
- Organization of the data matrix into homogeneous blocks

Aim

- To cluster the sets of rows and columns simultaneously
- To permutate the rows and the columns in order to obtain homogeneous blocks

Example of block clustering



- (1) : Initial data matrix
- (2) : Data matrix reorganized according a partition of rows
- (3) : Data matrix reorganized according partitions of rows and columns
- (4) : Summary of this matrix

Notations

Data

- matrix $\mathbf{x} = (x_{ij})$
- $i \in I$ set of n rows
- $j \in J$ set of d columns

Partition $\mathbf{z}$ of I in g clusters

- $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) = (z_{ik})$
  - $\mathbf{z}_i$ cluster number of i
  - $z_{ik} = 1$ if $i \in k$ and $z_{ik} = 0$ otherwise

3	0	0	1
2	0	1	0
3	0	0	1
2	0	1	0
1	1	0	0

Partition $\mathbf{w}$ of J in m clusters

- $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_p) = (w_{j\ell})$
  - $\mathbf{w}_j$ cluster number of j
  - $w_{j\ell} = 1$ if $j \in \ell$ and $w_{j\ell} = 0$ otherwise

From $\mathbf{z}$ and $\mathbf{w}$

- block $k\ell$ is defined by the x_{ij}'s with $z_{ik}w_{j\ell} = 1$

Block clustering algorithms (1)

Four algorithms (Govaert, 1977, 1983)

- CROBIN: binary data
- CROKI2: contingency data
- CROEUC: continuous data
- CROMUL: categorical data

Optimization of criterion $W(\mathbf{z}, \mathbf{w}, \mathbf{a})$

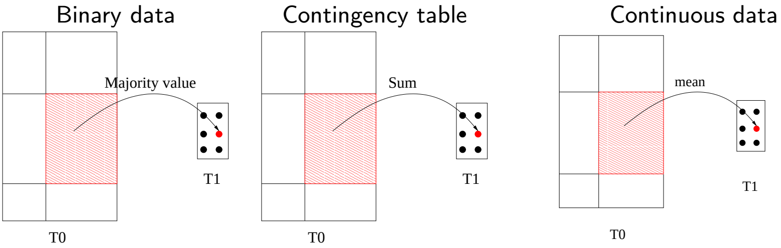
- $\mathbf{z}$ and $\mathbf{w}$ partitions of I and J
- $\mathbf{a} = (a_{k\ell})$ summary matrix of dimensions $K \times M$ having the same structure that the initial data matrix
- W depends on the type of data.

Additive model

- $\mathbf{x} = \mathbf{z}\mathbf{a}\mathbf{w}^T + \mathbf{e}$

Block clustering algorithms (2)

General principle



Criteria

Data	$a_{k\ell}$	Criterion W
Binary	Mode	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} \lvert x_{ij} - a_{k\ell} \rvert$
Contingency	Sum	$\chi^2(\mathbf{z},\mathbf{w}) = N \sum_{k,\ell} \frac{(f_{k\ell} - f_{k.} f_{.\ell})^2}{f_{k.} f_{.\ell}}$
Continuous	Mean	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - a_{k\ell})^2 = \lvert\lvert \mathbf{x} - \mathbf{zaw}^T \rvert\rvert^2$

Binary data: CROBIN

Algorithm

Alternated minimization of the criterion $W(\mathbf{z}, \mathbf{w}, \mathbf{a})$

- minimization of $W(\mathbf{z}, \mathbf{a}|\mathbf{w}) = \sum_{i,k,\ell} z_{ik}|u_{i\ell} - \#w_\ell a_{k\ell}|$ where $u_{i\ell} = \sum_j w_{j\ell}x_{ij}$
  - *nuées dynamiques* on **u**
- minimization of $W(\mathbf{w}, \mathbf{a}|\mathbf{z}) = \sum_{j,k,\ell} w_{j\ell}|v_{j\ell} - \#z_k a_{k\ell}|$ where $v_{kj} = \sum_i z_{ik}x_{ij}$
  - *nuées dynamiques* on **v**

Data

	abcdefghij
y_1	1010001101
y_2	0101110011
y_3	1000001100
y_4	1010001100
y_5	0111001100
y_6	0101110101
y_7	0111110111
y_8	1100111011
y_9	0100110000
y_{10}	1010101101
y_{11}	1010001100
y_{12}	1010000100
y_{13}	1010001101
y_{14}	0010011100
y_{15}	0010010100
y_{16}	1111001100
y_{17}	0101110011
y_{18}	1010011101
y_{19}	1010001000
y_{20}	1100101100

Reorganized matrix

	a c g h	*b d e f i j*
y_2	0 0 0 0	1 1 1 1 1 1
y_6	0 0 0 1	1 1 1 1 0 1
y_7	0 1 0 1	1 1 1 1 1 1
y_8	1 0 1 0	1 0 1 1 1 1
y_9	0 0 0 0	1 0 1 1 0 0
y_{17}	0 0 0 0	1 1 1 1 1 1
y_1	1 1 1 1	0 0 0 0 0 1
y_3	1 0 1 1	0 0 0 0 0 0
y_4	1 1 1 1	0 0 0 0 0 0
y_5	0 1 1 1	1 1 0 0 0 0
y_{10}	1 1 1 1	0 0 1 0 0 1
y_{11}	1 1 1 1	0 0 0 0 0 0
y_{12}	1 1 0 1	0 0 0 0 0 0
y_{13}	1 1 1 1	0 0 0 0 0 1
y_{14}	0 1 1 1	0 0 0 1 0 0
y_{15}	0 1 0 1	0 0 0 1 0 0
y_{16}	1 1 1 1	1 1 0 0 0 0
y_{18}	1 1 1 1	0 0 0 1 0 1
y_{19}	1 1 1 0	0 0 0 0 0 0
y_{20}	1 0 1 1	1 0 1 0 0 0

Summary

0	1
1	0

Homogeneity

0.80	0.87
0.86	0.84

Continuous Data

Minimization of the criterion $W(\mathbf{z},\mathbf{w},\mathbf{a}) = ||\mathbf{x} - \mathbf{zaw}^T||^2$

Two-mode k-means

- Choose initial **z** and **w**
- repeat the following steps
  - update **a**, $a_{k\ell} = \sum_{i,j} z_{ik} w_{j\ell} x_{ij} / \sum_{i,j} z_{ik} w_{j\ell}$
  - update **z**, $z_{ik} = 1$ if $c_{ik} = min_{1 \leq k \leq g} c_{ik}$ where $c_{ik} = \sum_{j,\ell} w_{j\ell}(x_{ij} - a_{k\ell})^2$
  - update **a**
  - update **w**, $w_{j\ell} = 1$ if $d_{j\ell} = min_{1 \leq \ell \leq m} d_{j\ell}$ where $d_{j\ell} = \sum_{i,k} z_{ik}(x_{ij} - a_{k\ell})^2$

Alternating Exchanges : Gaul and Schader (1996)

1. For each transfer row object i to row cluster k, we re-calculate **a**
2. For each transfer column object j to column cluster ℓ, we re-calculate **a**

The Croeuc Algorithm

(a) minimization of $W(\mathbf{z},\mathbf{a}|\mathbf{w}) = \sum_{i,k,\ell} z_{ik}(u_{i\ell} - \#w_\ell a_{k\ell})^2$ where $u_{i\ell} = \sum_j w_{j\ell} x_{ij} / \#w_\ell$

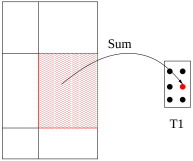
(a.1) *k-means* on **u** and we obtain **z**

(b) minimization of $W(\mathbf{w},\mathbf{a}|\mathbf{z}) = \sum_{j,k,\ell} w_{j\ell}(v_{j\ell} - \#z_k a_{k\ell})^2$ where $v_{kj} = \sum_i z_{ik} x_{ij} / \#z_k$

(b.1) *k-means* on **v** and we obtain **w**

Contingency table

- Summary of T_0 can be obtained by



- T_1 and T_0 have the same structure $\chi^2(T_0) \geq \chi^2(T_1)$
- Problem: find partitions $\mathbf{z}$ and $\mathbf{w}$ maximizing $\chi^2(\mathbf{z}, \mathbf{w})$.
- Solution: Alternated maximization of $\chi^2(\mathbf{z}, J)$ and $\chi^2(I, \mathbf{w})$
- Croki2: Alternated application of $k$means with the χ^2 metric on intermediate reduced matrices of size $(K \times p)$ and $(n \times M)$

Interests

Complementary methods to factor analysis methods

- PCA, Correspondence analysis, etc.

Reduction of the size of data

- They distil the initial data matrix into a simpler one having the same structure
- High dimensionality

Methods able to handle large data sets

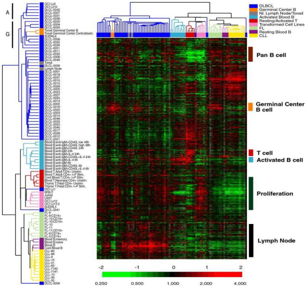
- Less computation required than for processing the two sets separately

n	p	K	M	separately	simultaneously
100	100	5	5	5×10^5	1.25×10^5
1000	1000	10	5	7.5×10^6	1.375×10^6
1000	1000	10	10	100×10^6	5×10^6

- Using $(n \times M)$ and $(K \times p)$ reduced matrices (good tool in data mining)
- To treat sparse data

Applications

- Text mining: clustering of documents and words simultaneously is better than
  - clustering of documents on basis of words
  - clustering of words on basis of documents
- Bioinformatics: clustering of genes and tissus simultaneously



Defects of algorithms cited

- Choice of the criterion not often easily
- Implicit hypotheses unknown
- Crobin not able to propose a solution when the clusters are not well-separated and
  - proportions of clusters dramatically different
  - degrees of homogeneity of blocks dramatically different

$$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|$$

- Croki2 not depending on the proportions of clusters

$$\chi^2(\mathbf{z},\mathbf{w}) = N \sum_{k,\ell} \frac{(f_{k\ell} - f_{k.}f_{.\ell})^2}{f_{k.}f_{.\ell}}$$

Aim

Propose a general framework able to formalize the hypotheses of block clustering algorithms: latent block model

- to overcome the defects of criteria and therefore to propose other criteria
- to develop other efficient algorithms

Algorithm of Block clustering

Algorithm of Block clustering

- Consists to permutate the rows and the columns in order to obtain homogeneous blocks

Optimisation of criterion $W(\mathbf{z}, \mathbf{w}, \mathbf{a})$

- **z** and **w** partitions of I and J
- $\boldsymbol{\alpha} = (\alpha_{k\ell})$ is a $K \times M$ data matrix having the same structure that the initial data matrix $n \times p$
- The criterion W depends on the type of data

Why to consider a probabilistic model ?

- We have seen the limits of a numerical criterion, interpretation not often easy, depend only the data and the centers
- Solution = "Block Mixture Model"

Outline

New formulation of the classical mixture model

Traditional formulation

$$f(\mathbf{x};\theta) = \prod_i \sum_k \pi_k \varphi(\mathbf{x}_i;\alpha_k)$$

- φ a statistical distribution with parameter $\boldsymbol{\alpha}_k$
- π_k the proportion of the kth component

Alternative formulation

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}\in\mathcal{Z}} P(\mathbf{z}) f(\mathbf{x}|\mathbf{z};\boldsymbol{\alpha})$$

- $P(\mathbf{z}) = \prod_i \pi_{\mathbf{z_i}}$
- $f(\mathbf{x}|\mathbf{z};\alpha) = \prod_i \varphi(\mathbf{x}_i;\alpha_{\mathbf{z_i}})$
- $\mathcal{Z}$ set of all the partitions of I

Proof

$$
\begin{aligned}
f(\mathbf{x},\boldsymbol{\theta}) &= \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)\\
&= \prod_{i=1}^{n}\sum_{\mathbf{z_i}\in\{1,\ldots,K\}} p_{\mathbf{z_i}}\varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_{\mathbf{z_i}})\\
&= \sum_{\mathbf{z}\in\mathcal{Z}}\prod_{i=1}^{n} p_{\mathbf{z_i}}\varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_{\mathbf{z_i}})\\
&= \sum_{\mathbf{z}\in\mathcal{Z}}\prod_{i=1}^{n} p_{\mathbf{z_i}}\prod_{i=1}^{n}\varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_{\mathbf{z_i}})\\
&= \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{z})f(\mathbf{x}|\mathbf{z};\alpha)
\end{aligned}
$$

where

- $P(\mathbf{z}) = \prod_i \pi_{\mathbf{z_i}}$
- $f(\mathbf{x}|\mathbf{z};\alpha) = \prod_i \varphi(\mathbf{x}_i;\alpha_{\mathbf{z_i}})$

Latent block model

Generalization on $I \times J$, (Govaert and Nadif, 2003)

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{u} \in U} P(\mathbf{u}) f(\mathbf{x}|\mathbf{u}; \boldsymbol{\alpha})$$

where U is the set of all the partitions of $I \times J$

Hypotheses

- $\mathbf{u} = \mathbf{z} \times \mathbf{w}$
- Hypothesis : $f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\alpha}) = \prod_{i,j} \varphi(x_{ij}; \boldsymbol{\alpha}_{z_i, w_j})$ where $\varphi(., \boldsymbol{\alpha})$ are pdf on $\mathbb{R}$

Latent block model

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} \prod_{i} \pi_{z_i} \prod_{j} \rho_{w_j} \prod_{i,j} \varphi(x_{ij}; \boldsymbol{\alpha}_{z_i w_j})$$

where $\theta = (\pi_1, \ldots, \pi_K, \rho_1, \ldots, \rho_M, \boldsymbol{\alpha}_{11}, \ldots, \boldsymbol{\alpha}_{gm})$

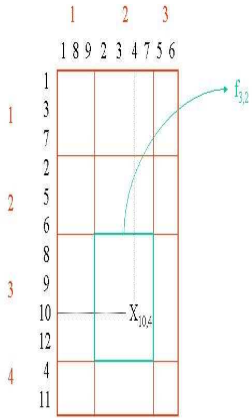
Interpretation

Given

- the proportions $\pi_1, \ldots, \pi_K$, $\rho_1, \ldots, \rho_M$
- the pdf of each pair of clusters,

the randomized data generation process can be described as follows:

- Generate the partition $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ according to the multinomial distribution $(\pi_1, \ldots, \pi_K)$
- Generate the partition $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_p)$ according to the multinomial distribution $(\rho_1, \ldots, \rho_M)$
- Generate for $i = 1, ..., n$ and $j = 1, ..., p$ a real value x_{ij} according to the distribution $\varphi(.; \boldsymbol{\alpha}_{z_i w_j})$



Types of data

Bernoulli latent block model

- Binary data
- φ Bernoulli distribution $\mathcal{B}(\alpha_{k\ell})$

More parsimonious than using classical mixture model on I and J

- Binary data
- $n = 1000$, $p = 500$, $K = 4$, $M = 3$, $\pi_k = 1/K$, $\rho_\ell = 1/M$
- Bernoulli latent block model : $4 \times 3 = 12$ parameters
- Two mixture models : $(4 \times 500 + 3 \times 1000) = 5000$ parameters

Many versatile or parsimonious models available

As for classical mixture models, it is possible to impose various constraints

- Fixed proportions
- Bernoulli latent model : $\alpha_{k\ell} \rightarrow (a_{k\ell}, \varepsilon_{k\ell})$ where $a_{k\ell} \in \{0, 1\}$ and $\varepsilon \in]0, 1/2[$
- Different models with ε, ε_k, ε_ℓ, $\varepsilon_{k\ell}$

Poisson latent block model

Poisson latent block model

- Contingency table
- φ Poisson distribution $\mathcal{P}(\mu_i \nu_j \alpha_{k\ell})$
  - μ_i and ν_j the effects of the row i and the column j
  - $\alpha_{k\ell}$ the effect of the block $k\ell$.
- Constraints for identifiability of the model : $\mu_i = (\mu_1, \ldots, \mu_n)$ and $\nu_j = (\nu_1, \ldots, \nu_p)$ are assumed to be known

Example

- Text mining
- I: set of documents
- J: set of words
- x_{ij} frequency of word j in document i
- Model : if i is in cluster k and j is in cluster ℓ, then

$$x_{ij} \sim \mathcal{P}(\mu_i \nu_j \alpha_{k\ell})$$

Outline

Clustering: find optimal $(\mathbf{z}^*, \mathbf{w}^*)$

Maximum Likelihood (ML) approach

- Estimation of θ by maximizing the likelihood of data
- MAP to propose optimal $(\mathbf{z}^*, \mathbf{w}^*)$
- Some problems for the block clustering
- BEM algorithm

Classification Maximum Likelihood (CML) approach

- Maximization of the complete data likelihood
- No problems to propose $(\mathbf{z}^*, \mathbf{w}^*)$
- BCEM

Remarks about CML approach

- To find the classical criteria and to propose the news
- To find the algorithms used and to propose other variants

Classification likelihood

The criterion

- Complete data: $(\mathbf{x}, \mathbf{z}, \mathbf{w})$
- Complete (or classification) log-likelihood

$$
\begin{aligned}
L_C(\theta, \mathbf{z}, \mathbf{w}) &= L(\theta; \mathbf{x}, \mathbf{z}, \mathbf{w}) = \log \left(\prod_i \pi_{\mathbf{z}_i} \prod_j \rho_{\mathbf{w}_j} \prod_{i,j} \varphi(x_{ij}; \boldsymbol{\alpha}_{\mathbf{z}_i \mathbf{w}_j}) \right) \\
&= \sum_i \log \pi_{\mathbf{z}_i} + \sum_j \log \rho_{\mathbf{w}_j} + \sum_{i,j} \log \varphi(x_{ij}; \boldsymbol{\alpha}_{\mathbf{z}_i \mathbf{w}_j}) \\
&= \sum_k n_k \log \pi_k + \sum_\ell d_\ell \log \rho_\ell + \sum_{i,j,k,\ell} z_{ik} w_{j\ell} \log \varphi(x_{ij}; \boldsymbol{\alpha}_{k\ell})
\end{aligned}
$$

- Find the partitions $\mathbf{z}$ and $\mathbf{w}$ and the parameter $\boldsymbol{\theta}$ maximizing L_C

Block CEM algorithm (BCEM)

Various alternated maximization of L_C using from an initial position $(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta})$, the three steps:

$$a) : \underset{\mathbf{z}}{\operatorname{argmax}} L_C(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) \quad b) : \underset{\mathbf{w}}{\operatorname{argmax}} L_C(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) \quad c) : \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L_C(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w})$$

Version 1

Repeat the two following steps until convergence

1. Repeat steps a) and b) until convergence
2. Step c)

Version 2

Repeat the two following steps until convergence

1. Repeat steps a) and c) until convergence
2. Repeat steps b) and c) until convergence

Some remarks on BCEM

Version 2

- Maximization of L_C by an alternated maximization of
  - Step 1: maximization of $L_C(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w})$
  - Step 2: maximization of $L_C(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z})$

  - $L_C(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w})$ associated to a classical mixture model on **u** a $(n \times M)$ data matrix
  - $L_C(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z})$ associated to a classical mixture model on **v** a $(K \times p)$ data matrix

  - Classical CEM on **u**
  - Classical CEM on **v**
- BCEM is an alternated application of the CEM algorithm on **u** and **v**

For Bernoulli and Poisson latent block models

- $L_C(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w})$ and $L_C(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z})$ associated to a mixture of Binomial distributions
- $L_C(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w})$ and $L_C(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z})$ associated to a mixture of multinomial distributions

Different computes for BCEM: Bernoulli latent block model

Notations

$$n_k = \sum_i z_{ik} \qquad d_\ell = \sum_j w_{j\ell}$$
$$v_{kj} = \sum_i z_{ik} x_{ij} \qquad u_{i\ell} = \sum_j w_{j\ell} x_{ij}$$

E-step (1,2): computation of s and t

$$s_{ik} \propto \pi_k \prod_\ell \alpha_{k\ell}^{u_{i\ell}} (1 - \alpha_{k\ell})^{d_\ell - u_{i\ell}}$$

$$t_{j\ell} \propto \rho_\ell \prod_k \alpha_{k\ell}^{v_{kj}} (1 - \alpha_{k\ell})^{n_k - v_{kj}}$$

C-step (1,2): computation of classification matrices z and w

$$z_{ik} = 1 \text{ if } k = \operatorname*{argmax}_{k'=1,\ldots,K} s_{ik'} \text{ and } w_{j\ell} = 1 \text{ if } \ell = \operatorname*{argmax}_{\ell'=1,\ldots,M} t_{j\ell'}$$

M-step (1,2): computation of θ

$$\pi_k = \frac{n_k}{n} \qquad \rho_\ell = \frac{d_\ell}{d} \qquad \alpha_{k\ell} = \frac{\sum_{ij} z_{ik} w_{j\ell} x_{ij}}{\sum_{ij} z_{ik} w_{j\ell}}$$

Links between BCEM and Crobin or Croki2

Crobin

- Constraints on the $(\alpha_{k\ell})$'s and the proportions
  - $\alpha_{k\ell} = (a_{k\ell}, \varepsilon)$ where $a_{k\ell} \in \{0,1\}$ and $\varepsilon \in]0, 1/2[$
  - Assumption : $\pi_1 = \ldots = \pi_K$ and $\rho_1 = \ldots = \rho_M$

$$L_c = \log(\frac{\varepsilon}{1-\varepsilon})W(\mathbf{z},\mathbf{w},\mathbf{a}) + cst$$

- Maximization of L_C equivalent to minimization of $W(\mathbf{z},\mathbf{w},\mathbf{a})$
- $L_C(\boldsymbol{\theta},\mathbf{z}|\mathbf{w})$ and $L_C(\boldsymbol{\theta},\mathbf{w}|\mathbf{z})$ correspond to $W(\mathbf{z},\mathbf{a}|\mathbf{w})$ and $W(\mathbf{w},\mathbf{a}|\mathbf{z})$

Croki2

- Assumption : $\pi_1 = \ldots = \pi_K$ and $\rho_1 = \ldots = \rho_M$

$$L_c = N \underbrace{\sum_{k,\ell} f_{k\ell} \log \frac{f_{k\ell}}{f_{k.} f_{.\ell}}}_{I(\mathbf{z},\mathbf{w})/\chi^2(\mathbf{z},\mathbf{w})/Croki2} + cst$$

Maximization of likelihood

- EM algorithm
- Complete data : $(\mathbf{x}, \mathbf{z}, \mathbf{w})$
- Iterative maximization of the conditional expectation of $L_C(\theta, \mathbf{z}, \mathbf{w})$
  - given the data $\mathbf{x}$ and using the current fit θ' for the parameter :

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{ik} s_{ik} \log \pi_k + \sum_{j\ell} t_{j\ell} \log \rho_\ell + \sum_{ijk\ell} e_{ijk\ell} \log \varphi(x_{ij}; \boldsymbol{\alpha}_{k\ell})$$

  - $s_{ik} = P(z_{ik} = 1|\mathbf{x}, \boldsymbol{\theta}')$, $t_{j\ell} = P(w_{j\ell} = 1|\mathbf{x}, \boldsymbol{\theta}')$
  - $e_{ijk\ell} = P(z_{ik} w_{j\ell} = 1|\mathbf{x}, \boldsymbol{\theta}')$

Difficulties

- Dependence structure among the variables x_{ij}
- Determination of $e_{ikj\ell}$ not tractable

Approximation

- Replace the maximization of the likelihood by the maximization of a new criterion

The Neal and Hinton interpretation of the EM algorithm

Hathaway interpretation of EM : classical mixture model context

- EM = alternated maximization of the fuzzy clustering criterion

$$F_C(\mathbf{s},\boldsymbol{\theta}) = L_C(\mathbf{s};\boldsymbol{\theta}) + H(\mathbf{s})$$

  - $\mathbf{s} = (s_{ik})$: fuzzy partition
  - $L_C(\mathbf{s},\boldsymbol{\theta}) = \sum_{i,k} s_{ik} \log(\pi_k \varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_k))$: fuzzy classification log-likelihood
  - $H(\mathbf{s}) = -\sum_{i,k} s_{ik} \log s_{ik}$: entropy function

Algorithm

- Maximizing F_C w.r. to $\mathbf{s}$ yields the E step
- Maximizing F_C w.r. to $\boldsymbol{\theta}$ yields the M step

Neal and Hinton interpretation of EM: general context

$$F_C(P,\boldsymbol{\theta}) = E_P(L_C(\mathbf{z},\boldsymbol{\theta})) + H(P)$$

- P: distribution over the space of missing data $\mathbf{z}$
- H: entropy function

Fuzzy criterion

By using

- the Neal and Hinton interpretation of the EM algorithm
- the variational mean field approximation: $e_{ikj\ell} = s_{ik} \times t_{j\ell}$

we replace the likelihood criterion by the new criterion (Govaert and Nadif, 2008)

$$G(\boldsymbol{\theta}, \mathbf{s}, \mathbf{t}) = L_C(\boldsymbol{\theta}, \mathbf{s}, \mathbf{t}) + H(\mathbf{s}) + H(\mathbf{t})$$

where $\mathbf{s} = (s_{ik})$, $\mathbf{t} = (t_{j\ell})$ and H is the entropy function.

Various alternated maximization of G using, from an initial position $(\mathbf{s}, \mathbf{t}, \boldsymbol{\theta})$, the three steps:

$$a) : \operatorname*{argmax}_{\mathbf{s}} G(\boldsymbol{\theta}, \mathbf{s}, \mathbf{t}) \quad b) : \operatorname*{argmax}_{\mathbf{t}} G(\boldsymbol{\theta}, \mathbf{s}, \mathbf{t}) \quad c) : \operatorname*{argmax}_{\boldsymbol{\theta}} G(\boldsymbol{\theta}, \mathbf{s}, \mathbf{t})$$

Block EM algorithm: version 1

Repeat the two following steps until convergence

1. Repeat steps a) and b) until convergence
2. Step c)

Block EM algorithm

Version 2

Repeat the two following steps until convergence

1. Repeat steps a) and c) until convergence
2. Repeat steps b) and c) until convergence

Interpretation of Version 2

- Step 1: maximization of $G(\theta, \mathbf{s}|\mathbf{t})$, Hathaway $\rightarrow$ EM
- Step 2: maximization of $G(\theta, \mathbf{t}|\mathbf{s})$, Hathaway $\rightarrow$ EM

Alternated maximization by using reduced matrices u and v

- $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_n)$ where $\mathbf{u}_i = (u_{i1}, \ldots, u_{iM})$
  - $\mathbf{u}_{i\ell} = f(x_{ij}, t_{j\ell})$
- $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_p)$ where $\mathbf{v}_j = (v_{1j}, \ldots, v_{Kj})$
  - $\mathbf{v}_{kj} = f(x_{ij}, s_{ik})$

Different computes for BEM: Bernoulli latent block model

Notations

$$n_k = \sum_i s_{ik} \qquad d_\ell = \sum_j t_{j\ell}$$
$$v_{kj} = \sum_i s_{ik} x_{ij} \qquad u_{i\ell} = \sum_j t_{j\ell} x_{ij}$$

E-step (1,2): computation of s and t

$$s_{ik} \propto \pi_k \prod_\ell \alpha_{k\ell}^{u_{i\ell}} (1 - \alpha_{k\ell})^{d_\ell - u_{i\ell}}$$

$$t_{j\ell} \propto \rho_\ell \prod_k \alpha_{k\ell}^{v_{kj}} (1 - \alpha_{k\ell})^{n_k - v_{kj}}$$

M-step (1,2): computation of θ

$$\pi_k = \frac{n_k}{n} \qquad \rho_\ell = \frac{d_\ell}{d} \qquad \alpha_{k\ell} = \frac{\sum_{ij} s_{ik} t_{j\ell} x_{ij}}{\sum_{ij} s_{ik} t_{j\ell}}$$

Example $n \times r = 200 \times 120$, fairly-separated

θ	True values	Estimations by BEM	Estimations by BCEM
p_1	0.2	0.1979	0.1900
p_2	0.3	0.3140	0.3400
p_3	0.5	0.4881	0.4700
q_1	0.3	0.2929	0.2583
q_2	0.7	0.7071	0.7417
α	$\begin{pmatrix} 0.60 & 0.40 \\ 0.40 & 0.60 \\ 0.60 & 0.65 \end{pmatrix}$	$\begin{pmatrix} 0.6067 & 0.4026 \\ 0.4089 & 0.6041 \\ 0.5989 & 0.6565 \end{pmatrix}$	$\begin{pmatrix} 0.6188 & 0.4063 \\ 0.3861 & 0.6000 \\ 0.6095 & 0.6559 \end{pmatrix}$
$\lVert\theta - \theta^0\rVert$	0	0.0252	0.0824

- Good estimation by BEM

Outline

Some numerical simulations

Parameters

- Characteristics of the data
  - Bernoulli block mixture model
  - $g = 3$ and $m = 2$
- 9 situations:
  - 3 degrees of overlapping:
    - Well-separated (+): 4%
    - Fairly-separated (++): 15%
    - Poorly-separated (+++): 25%
  - 3 sizes of data:
    - Small: $n \times p = 50 \times 30$
    - Medium: $n \times p = 100 \times 60$
    - Large: $n \times p = 200 \times 120$
- For each situation: simulation of 30 samples

Objective

- Comparison of BEM and BCEM by looking at the quality of results and the frequency on 30 that one of the two algorithms outperforms the other
- Clustering (error rate) and estimation contexts ($\|\boldsymbol{\theta} - \boldsymbol{\theta}^0\|$)
- Only Version 2 because it is slightly better and faster

Results with well-separated data (True error rate = 0.03)

Sizes		(50, 30)	(100, 60)	(200, 120)
Error rate	mean for BEM	0.03	0.04	0.02
	mean for BCEM	0.04	0.04	0.03
	#(BEM>BCEM)	1	9	6
	#(BEM=BCEM)	27	18	23
	#(BEM<BCEM)	2	3	1
$\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathbf{0}}\|$	mean for BEM	0.19	0.13	0.08
	mean for BCEM	0.21	0.14	0.08
	#(BEM>BCEM)	15	20	20
	#(BEM=BCEM)	0	0	0
	#(BEM<BCEM)	15	10	10

Results with fairly-separated data (True error rate = 0.15)

Sizes		(50, 30)	(100, 60)	(200, 120)
Error rate	mean for BEM	0.21	0.13	0.13
	mean for BCEM	0.31	0.15	0.20
	#(BEM>BCEM)	17	18	24
	#(BEM=BCEM)	11	8	1
	#(BEM<BCEM)	2	4	5
$\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathbf{0}}\|$	mean for BEM	0.34	0.16	0.10
	mean for BCEM	0.52	0.22	0.21
	#(BEM>BCEM)	27	25	27
	#(BEM=BCEM)	0	0	0
	#(BEM<BCEM)	3	5	3

Results with poorly-separated data (True error rate =0.25)

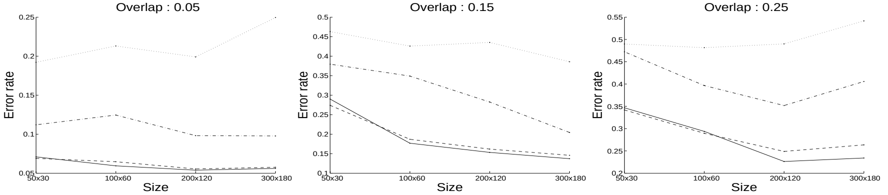
Sizes		(50, 30)	(100, 60)	(200, 120)
Error rate	mean for BEM	0.40	0.28	0.29
	mean for BCEM	0.52	0.53	***
	#(BEM>BCEM)	27	30	30
	#(BEM=BCEM)	0	0	0
	#(BEM<BBCEM)	3	0	0
$\lVert\theta - \theta^0\rVert$	mean for BEM	0.49	0.28	0.17
	mean for BCEM	0.78	0.79	***
	#(BEM>BCEM)	28	30	30
	#(BEM=BCEM)	0	0	0
	#(BEM<BCEM)	2	0	0

Some remarks drawn from these simulations

- BEM outperforms BCEM in most of situations
- Even when the clusters are well separated (favorable situation for BCEM), the performances of both algorithms are not very different
- BEM gives error rates closed to the true value when the size is large enough

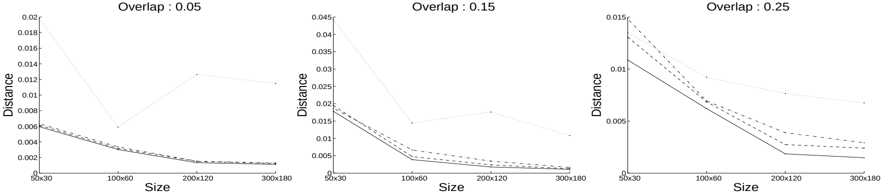
What one can wonder about the performances of 2BEM, 2CEM ?

Clustering



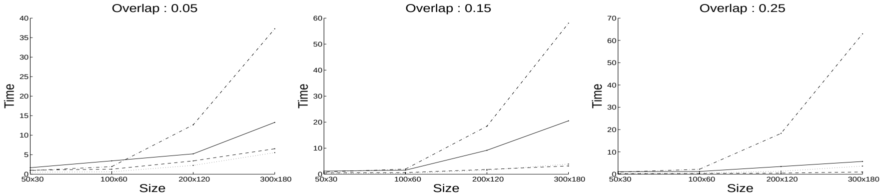
Mean error rates for BEM: solid line, BCEM: dashed line: 2CEM: dotted line and 2EM: dash-dot line

Estimation



Mean distance between true and estimated parameters for the 4 algorithms

Run times

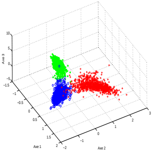


Mean run time (in seconds) according to size and overlap

- BCEM > 2CEM and BEM > 2EM in all situations where the size > 100×60

An illustrative example

- Classic3 data (3893 abstracts, 2000 words) :
  - 1033 abstracts from medical journals,
  - 1460 from **IR** papers,
  - 1400 from aerodynamic systems



Comparison between BEM and BCEM ($g = 3$, $m = 3$)

- Confusion matrices obtained resp. by BEM and BCEM

	Med.	Cis.	Cra.
z_1	1008	4	2
z_2	25	1451	2
z_3	1	16	1383

	Med.	Cis.	Cra.
z_1	1007	3	2
z_2	25	1452	15
z_3	1	6	1382

- BEM > BCEM (52 mis. for BEM and 56 mis. for BCEM)
- 2BEM (54 mis.) and 2CEM (76 mis.)
- BEM is more adapted for clustering even if it is not its aim

Outline

Conclusion

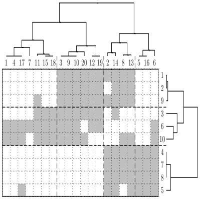
Principal points

- Block clustering methods: BEM and BCEM
- BEM is interesting in clustering and estimation contexts
- Illustrations on binary data and contingency table

Other works related to the latent block model

- Case of continuous data
- number of blocks
- missing data
- speed-up of BEM

- Hierarchical block clustering method



Outline

References

Principal references

- Govaert, G. and Nadif, M., Block clustering with Bernoulli mixture models: Comparison of different approaches, Computational Statistics and Data Analysis , 52, 3233-3245, 2008
- Jollois, F-X. and Nadif, M., Speed up EM algorithm for categorical data, Journal of Global Optimization, 37, 513-525, 2007
- Govaert, G. and Nadif, M., Clustering of contingency table and mixture model, European Journal of Operational Research, 183, 1055-1066, 2007
- Govaert, G. and Nadif, M., An EM algorithm for the Block Mixture Model, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27, 4, pp. 643-647, 2005.
- Govaert, G. and Nadif, M., Fuzzy Clustering to estimate the parameters of block mixture models, Soft Computing, 10, 5, 415-422, 2005
- Govaert, G. and Nadif, M., Clustering with block mixture models, models, Pattern Recognition,36(2), 5, 463-473, 2003
- Govaert, G.,Classification croisée, Thèse d'état, Université Paris Dauphine, 1983