Latent Block Models

Mohamed Nadif

LIPADE, Université Paris Descartes, France

Plan

1. **Co-clustering**
2. **Latent block model**
   - Bernoulli LBM
   - Gaussian LBM
3. **CML and ML approaches**
   - CML approach
   - Links between BCEM and classical co-clustering algorithms
   - Gaussian latent block models
4. **Contingency Table**
   - Criteria and Algorithms
5. **Latent block model**
   - Poisson Latent Block Model
   - Algorithms
6. **Conclusion**

Outline

1. **Co-clustering**
2. **Latent block model**
   - Bernoulli LBM
   - Gaussian LBM
3. **CML and ML approaches**
   - CML approach
   - Links between BCEM and classical co-clustering algorithms
   - Gaussian latent block models
4. **Contingency Table**
   - Criteria and Algorithms
5. **Latent block model**
   - Poisson Latent Block Model
   - Algorithms
6. **Conclusion**

Simultaneous clustering on both dimensions

- First works in J.A. Hartigan, Direct Clustering of a Data Matrix, JASA, 1972.
- Referred in the literature as bi-clustering, co-clustering, double clustering, direct clustering, coupled clustering
  - no-overlapping co-clustering
  - overlapping co-clustering
- Different approaches are proposed: they differ in the pattern they seek and the types of data on which they apply
- All proposed methods aim to organize the data matrix into homogeneous blocks
- The co-clustering methods have attracted much attention to cluster the sets of objects and features simultaneously
  - Text mining: documents, terms
  - Bioinformatics: genes, experiments

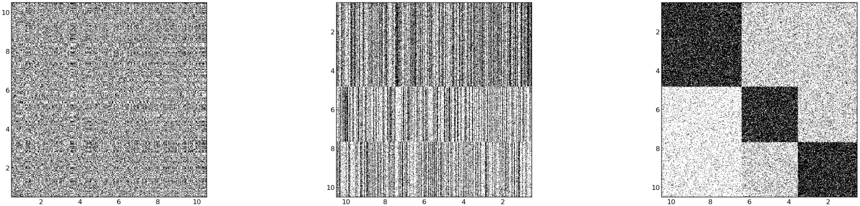


Figure: Left : Original data. Middle : data reorganized according to row clusters. Right : data reorganized according to row and column clusters.

Interests

- Extracting relevant clusters and co-clusters
- Generating compact representation
- Enriching visualization methods (for instance with CA)
- Reducing running time

Approaches

- Metric
- Matrix Factorization
- Spectral
- Probabilistic

Notations

Data

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

Partition of I in g clusters

- $\mathbf{z} = (z_1, \ldots, z_i, \ldots, z_n)$ where $z_i \in \{1, \ldots, g\}$
- $\mathbf{Z} = (z_{ik})$ where $z_{ik} = 1$ if $i \in k$th cluster and $z_{ik} = 0$ otherwise

z		Z	
3	0	0	1
2	0	1	0
3	0	0	1
2	0	1	0
1	1	0	0

Partition of J in s clusters

- $\mathbf{w} = (w_1, \ldots, w_j, \ldots, w_d)$ where $w_j \in \{1, \ldots, s\}$
- $\mathbf{W} = (w_{ij})$ where $w_{j\ell} = 1$ if $j \in \ell$th cluster and $w_{j\ell} = 0$ otherwise

From Z and W

- Block (k, ℓ) is defined by the x_{ij}'s with $z_{ik} w_{j\ell} = 1$

Co-clustering algorithms (1)

Four algorithms (Govaert, 1977, 1983)

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

Optimization of criterion $\mathcal{C}(\mathbf{Z}, \mathbf{W}, \mathbf{A})$

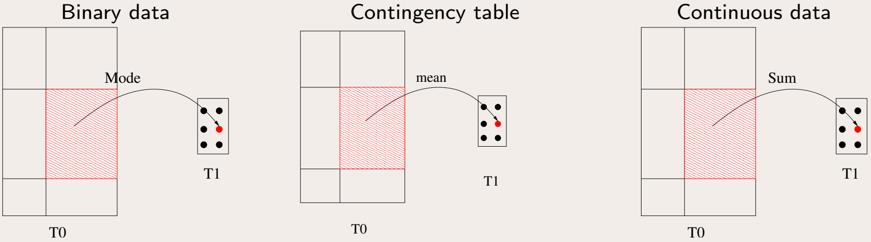
- $\mathbf{Z}$ and $\mathbf{W}$ partitions of I and J
- $\mathbf{A} = (a_{k\ell})$ summary matrix of dimensions $g \times s$ having the same structure that the initial data matrix
- $\mathcal{C}$ depends on the type of data.

Model

- $\mathbf{X} = \mathbf{Z}\mathbf{A}\mathbf{W}^T + \mathbf{R}$

Co-clustering algorithms (2)

General principle



Criteria

Data	Summary	Criterion $\mathcal{C}$
Binary	Mode	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} \lvert x_{ij} - a_{k\ell} \rvert$
Continuous	Mean	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - \mu_{k\ell})^2 = \lvert\lvert \mathbf{X} - \mathbf{ZAW}^T \rvert\rvert^2$
Contingency	Sum	$\chi^2(\mathbf{z}, \mathbf{w}) = N \sum_{k,\ell} \frac{(p_{k\ell} - p_{k.} p_{.\ell})^2}{p_{k.} p_{.\ell}}$

Binary data

Illustration

	v1	v2	v3	v4	v5	z
a	1	0	0	0	0	1
b	0	1	0	1	1	1
c	0	0	1	1	0	2
d	1	0	0	0	1	1
e	1	0	1	1	1	2
f	0	0	1	1	1	2
w	1	1	2	2	1	

- $\mathbf{z} = (1,1,1,2,1,2,2)$
  - 1th cluster $= \{a, b, d\}$, 2th cluster $= \{c, e, f\}$
- $\mathbf{w} = (1,1,2,2,1)$
  - 1th cluster $= \{v1, v2, v5\}$, 2th cluster $= \{v3, v4\}$

w fixed $\Rightarrow \mathbf{X}^{\mathbf{w}} = (x^{\mathbf{w}}_{i\ell})$ where $x^{\mathbf{w}}_{i\ell} = \sum_j w_{j\ell} x_{ij}$

	v1	v2	v5	v4	v3
a	1	0	0	0	0
b	0	1	1	1	0
d	1	0	1	0	0
c	0	0	0	1	1
e	1	0	1	1	1
f	0	0	1	1	1

$\mathbf{X}^{\mathbf{w}}$	$\{v1, v2, v5\}$	$\{v3, v4\}$
a	1	0
b	2	1
d	2	0
c	0	2
e	2	2
f	1	2

z fixed $\Rightarrow \mathbf{X}^{\mathbf{z}} = (x^{\mathbf{w}}_{kj})$ where $x^{\mathbf{z}}_{kj} = \sum_i z_{ik} x_{ij}$

	v1	v2	v5	v4	v3
a	1	0	0	0	0
b	0	1	1	1	0
d	1	0	1	0	0
c	0	0	0	1	1
e	1	0	1	1	1
f	0	0	1	1	1

$\mathbf{X}^{\mathbf{z}}$	v1	v2	v5	v4	v3
$\{a, b, d\}$	2	1	2	1	0
$\{c, e\}$	1	0	2	3	3

Binary data: CROBIN

Algorithm

Alternated minimization of $\mathcal{C}(\mathbf{Z},\mathbf{W},\mathbf{A}) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|$

- $\text{argmin}_{\mathbf{Z},\mathbf{A}}\, \mathcal{C}(\mathbf{Z},\mathbf{A}|\mathbf{W}) = \sum_{i,k,\ell} z_{ik} |x^{\mathbf{w}}_{i\ell} - w_{.\ell} a_{k\ell}|$ where $w_{.\ell} = \sum_j w_{j\ell}$
  - *nuées dynamiques* on $\mathbf{X}^{\mathbf{w}}$ of size $n \times s$
- $\text{argmin}_{\mathbf{W},\mathbf{A}}\, \mathcal{C}(\mathbf{W},\mathbf{A}|\mathbf{Z}) = \sum_{j,k,\ell} w_{j\ell} |x^{\mathbf{z}}_{kj} - z_{.k} a_{k\ell}|$ where $z_{.k} = \sum_i z_{ik}$
  - *nuées dynamiques* on $\mathbf{X}^{\mathbf{z}}$ of size $g \times d$

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 1 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 (A)

0	1
1	0

Homogeneity

0.8	0.9
0.9	0.8

Heterogenity (ε)

0.2	0.1
0.1	0.2

Continuous Data

Minimization of $\mathcal{C}(\mathbf{Z},\mathbf{W},\mathbf{A}) = ||\mathbf{X} - \mathbf{Z}\mathbf{A}\mathbf{W}^T||^2$

Algorithm

- Choose initial **Z** and **W**
- Repeat the following steps
  - update **A**
  - update **Z**
  - update **W**

Two-mode k-means

- Choose initial **Z** and **W**
- Repeat the following steps
  - update **A**
  - update **Z**
  - update **A**
  - update **W**

The Croeuc Algorithm

(a) $\text{argmin}_{\mathbf{Z},\mathbf{A}} \mathcal{C}(\mathbf{Z},\mathbf{A}|\mathbf{W}) = \sum_{i,k,\ell} z_{ik}(x^{\mathbf{w}}_{i\ell} - a_{k\ell})^2$ where $x^{\mathbf{w}}_{i\ell} = \sum_j w_{j\ell}x_{ij}/w_{.\ell}$

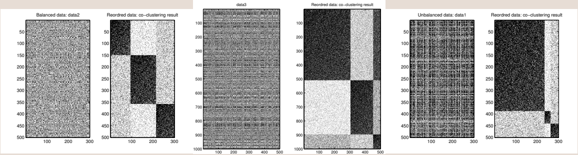
(a.1) *k-means* on $\mathbf{X}^{\mathbf{w}}$

(b) $\text{argmin}_{\mathbf{W},\mathbf{A}} \mathcal{C}(\mathbf{W},\mathbf{A}|\mathbf{Z}) = \sum_{j,k,\ell} w_{j\ell}(v^{\mathbf{z}}_{kj} - a_{k\ell})^2$ where $v^{\mathbf{z}}_{kj} = \sum_i z_{ik}x_{ij}/z_{.k}$

(b.1) *k-means* on $\mathbf{X}^{\mathbf{z}}$

Limits of classical co-clustering methods

- $\sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|$, $\sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - \mu_{k\ell})^2$,
- Choice of the criterion not often easily, implicit hypotheses unknown
- Algorithms not able to propose a solution when
  - the clusters are not well-separated
  - degrees of homogeneity of blocks dramatically different
  - proportions of clusters dramatically different



Aim

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

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

Outline

1. Co-clustering
2. **Latent block model**
   - Bernoulli LBM
   - Gaussian LBM
3. CML and ML approaches
   - CML approach
   - Links between BCEM and classical co-clustering algorithms
   - Gaussian latent block models
4. Contingency Table
   - Criteria and Algorithms
5. Latent block model
   - Poisson Latent Block Model
   - Algorithms
6. Conclusion

Latent block model

Definition (Govaert and Nadif, 2003)

The pdf of **X**:

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

where $\Theta = (\pi_1,\ldots,\pi_g;\rho_1,\ldots,\rho_s;\boldsymbol{\alpha}_{11},\ldots,\boldsymbol{\alpha}_{gs})$

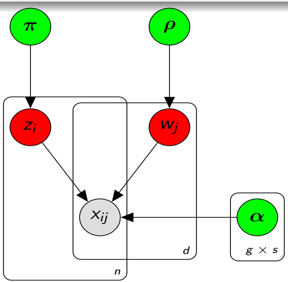
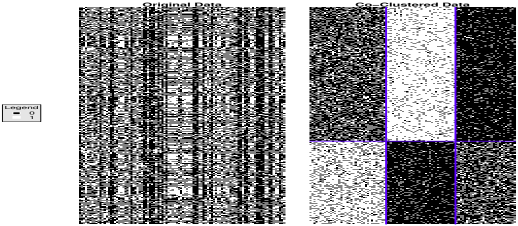


Figure: LBM as a graphical model

Advantages (see for instance; Govaert and Nadif, 2013)

- Parsimonious models giving probabilistic interpretations of classical criteria

```
# Simple example with simulated binary data
#load data
data(binarydata)
#usage of cocluster function in its most simplest form
library(blockcluster)
out<-cocluster(binarydata,datatype="binary",nbcocluster=c(2,3))
#Summarize the output results
summary(out)
#Plot the original and Co-clustered data
plot(out)
```



Binary data: Classical Bernoulli Mixture model

- We have $f(\mathbf{x}_i;\Theta)=\sum_k \pi_k \prod_j \alpha_{kj}^{x_{ij}}(1-\alpha_{kj})^{(1-x_{ij})}$, $\boldsymbol{\alpha}_k$ can be replaced by the two parameters a_k and ε_k : $f(\mathbf{x}_i;\Theta)=\sum_k \pi_k \prod_j \varepsilon_{kj}^{|x_{ij}-a_{kj}|}(1-\varepsilon_{kj})^{1-|x_{ij}-a_{kj}|}$ where

$$\begin{cases} a_{kj}=0, \varepsilon_{kj}=\alpha_{kj} & \text{if } \alpha_{kj}\leq 0.5 \\ a_{kj}=1, \varepsilon_{kj}=1-\alpha_{kj} & \text{if } \alpha_{kj}> 0.5 \end{cases}$$

  - $p(x_{ij}=1|a_{kj}=0)=p(x_{ij}=0|a_{kj}=1)=\varepsilon_{kj}$
  - $p(x_{ij}=0|a_{kj}=0)=p(x_{ij}=1|a_{kj}=1)=1-\varepsilon_{kj}$

Bernoulli Latent block model: $\mathcal{B}(\alpha_{k\ell})$

$$\varphi(x_{ij};\boldsymbol{\alpha}_{k\ell})=\alpha_{k\ell}^{x_{ij}}(1-\alpha_{k\ell})^{(1-x_{ij})}$$

$$\alpha_{k\ell}\Rightarrow(a_{k\ell},\boldsymbol{\varepsilon}_{k\ell})\begin{cases} a_{k\ell}=0, \varepsilon_{k\ell}=\alpha_{k\ell} & \text{if } \alpha_{k\ell}\leq 0.5 \\ a_{k\ell}=1, \varepsilon_{k\ell}=1-\alpha_{k\ell} & \text{if } \alpha_{k\ell}> 0.5 \end{cases} \quad a_{k\ell}\in\{0,1\} \text{ and } \varepsilon_{k\ell}\in]0,1/2[$$

Number of parameters

$\Theta = (\pi_1, \ldots, \pi_g; \rho_1, \ldots, \rho_s; \alpha_{11}, \ldots, \alpha_{gs})$

- Number of parameters: $(g-1) + (s-1) + g \times s$
  - $n = 1000$, $d = 500$, $g = 4$, $s = 3$, $\pi_k = 1/g$, $\rho_\ell = 1/s$
  - Bernoulli latent block model: $4 \times 3 = 12$ parameters
  - Two mixture models: $(4 \times 500 + 3 \times 1000) = 5000$ parameters

Parsimonious models available

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

- Constraints on the proportions: $\pi_1 = \ldots = \pi_g$ and $\rho_1 = \ldots = \rho_s$
- Constraints on $\varepsilon_{k\ell}$: ε_k, ε_ℓ, ε
- Constraints on $a_{k\ell}$

Gaussian Latent block model

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

- Fixed proportions: $\pi_1 = \ldots = \pi_g$ and $\rho_1 = \ldots = \rho_s$
- Gaussian LBM : $\alpha_{k\ell} \rightarrow (\mu_{k\ell}, \sigma_{k\ell})$
- Constraints on $\sigma_{k\ell}$: σ_k, σ_ℓ, σ
- Constraints on $\mu_{k\ell}$

Outline

1. Co-clustering

2. Latent block model
   - Bernoulli LBM
   - Gaussian LBM

3. CML and ML approaches
   - CML approach
   - Links between BCEM and classical co-clustering algorithms
   - Gaussian latent block models

4. Contingency Table
   - Criteria and Algorithms

5. Latent block model
   - Poisson Latent Block Model
   - Algorithms

6. Conclusion

Clustering: find optimal (Z^*, 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
- VEM (Variational Expectation-Maximization) algorithm

Classification Maximum Likelihood (CML) approach

- Maximization of the complete data likelihood
- No problems to propose $(\mathbf{Z}^*, \mathbf{W}^*)$
- BCEM (Block Classification EM) algorithm

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 data (or classification) log-likelihood

$$L_C(\Theta, \mathbf{Z}, \mathbf{W}) = \sum_k z_{.k} \log \pi_k + \sum_\ell w_{.\ell} \log \rho_\ell + \sum_{i,j,k,\ell} z_{ik} w_{j\ell} \log \varphi(x_{ij}; \boldsymbol{\alpha}_{k\ell})$$

- Constraints on $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$: $L_{CR}(\Theta, \mathbf{Z}, \mathbf{W}) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} \log \varphi(x_{ij}; \boldsymbol{\alpha}_{k\ell})$

Various alternated maximization of L_C using from an initial position $(\mathbf{Z}, \mathbf{W}, \Theta)$, the three steps:

$$a) : \operatorname*{argmax}_{\mathbf{Z}} L_C(\Theta, \mathbf{Z}, \mathbf{W}) \quad b) : \operatorname*{argmax}_{\mathbf{W}} L_C(\Theta, \mathbf{Z}, \mathbf{W}) \quad c) : \operatorname*{argmax}_{\Theta} L_C(\Theta, \mathbf{Z}, \mathbf{W})$$

A version among others

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

- Maximization of L_C by an alternated maximization of
  - Step 1: maximization of $L_C(\Theta, \mathbf{Z}|\mathbf{W})$
  - Step 2: maximization of $L_C(\Theta, \mathbf{W}|\mathbf{Z})$

  - $L_C(\Theta, \mathbf{Z}|\mathbf{W})$ associated to a classical mixture model on $\mathbf{X}^{\mathbf{z}}$ a $(n \times s)$ data matrix
  - $L_C(\Theta, \mathbf{W}|\mathbf{Z})$ associated to a classical mixture model on $\mathbf{X}^{\mathbf{w}}$ a $(g \times d)$ data matrix

  - Classical Classification EM on $\mathbf{X}^{\mathbf{w}}$
  - Classical Classification EM on $\mathbf{X}^{\mathbf{z}}$
- BCEM is an alternated application of the CEM algorithm on $\mathbf{X}^{\mathbf{w}}$ and $\mathbf{X}^{\mathbf{z}}$

For Bernoulli and Poisson latent block models

- $L_C(\Theta, \mathbf{Z}|\mathbf{W})$ and $L_C(\Theta, \mathbf{W}|\mathbf{Z})$ associated to a mixture of Binomial distributions
- $L_C(\Theta, \mathbf{Z}|\mathbf{W})$ and $L_C(\Theta, \mathbf{W}|\mathbf{Z})$ associated to a mixture of multinomial distributions

Link between BCEM and Crobin

Bernoulli LBM

- Constraints: $\pi_1 = \ldots = \pi_g$, $\rho_1 = \ldots = \rho_s$ and $\varepsilon_{k\ell} = \varepsilon \quad \forall k, \ell$

$$\underset{\mathbf{Z},\Theta,\mathbf{W}}{\operatorname{argmax}} L_{RC}(\Theta, \mathbf{Z}, \mathbf{W}) \equiv \underset{\mathbf{Z},\mathbf{A},\mathbf{W}}{\operatorname{argmin}} \sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|$$

- BCEM=Crobin
- Example of Constraints: ε and $a_{kk} = 1$ and $a_{k\ell} = 0$ for all k $\neq \ell$ (Laclau and Nadif, 16)

$$\underset{\mathbf{Z},\Theta,\mathbf{W}}{\operatorname{argmax}} L_{RC}(\Theta, \mathbf{Z}, \mathbf{W}) \equiv \underset{\mathbf{Z},\mathbf{A},\mathbf{W}}{\operatorname{argmin}} \sum_{i,j,k} z_{ik} w_{j\ell} |x_{ij} - 1| + \sum_{i,j,k,\ell \neq k} z_{ik} w_{j\ell} x_{ij}$$

Gaussian LBM

- Constraints: $\pi_1 = \ldots = \pi_g$, $\rho_1 = \ldots = \rho_s$ and $\sigma_{k\ell} = \sigma \quad \forall k, \ell$

$$\underset{\mathbf{Z},\Theta,\mathbf{W}}{\operatorname{argmax}} L_{RC}(\Theta, \mathbf{Z}, \mathbf{W}) \equiv \underset{\mathbf{Z},\mathbf{A},\mathbf{W}}{\operatorname{argmin}} \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - a_{k\ell})^2 = \underset{\mathbf{Z},\mathbf{A},\mathbf{W}}{\operatorname{argmin}} ||\mathbf{X} - \mathbf{Z}\mathbf{A}\mathbf{W}^T||^2$$

- BCEM=Croeuc

Continuous data

We assume that for each block $k\ell$ the values x_{ij} are distributed according to a Gaussian distribution

$$\mathcal{G}(\mu_{k\ell}, \sigma^2_{k\ell}) \qquad \text{with} \quad \mu_{k\ell} \in \mathbb{R} \quad \text{and} \quad \sigma^2_{k\ell} \in \mathbb{R}^+,$$

we obtain the Gaussian latent block model with the following pdf $f(\mathbf{X}; \Theta)$ taking this form

$$\sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \left(\frac{1}{\sqrt{2\pi\sigma^2_{k\ell}}} \exp - \left\{ \frac{1}{2\sigma^2_{k\ell}} (x_{ij} - \mu_{k\ell})^2 \right\} \right)^{z_{ik} w_{j\ell}} \tag{1}$$

With this model, the complete-data log-likelihood is, up to the constant $-\frac{nd}{2} \log 2\pi$, given by

$$\begin{aligned} L_C(\Theta, \mathbf{Z}, \mathbf{W}) &= \sum_{k,\ell} z_{ik} \log \pi_k + \sum_{j,\ell} w_{j\ell} \log \rho_\ell \\ &- \frac{1}{2} \sum_{k,\ell} \left(z_{.k} w_{.\ell} \log \sigma^2_{k\ell} + \frac{1}{\sigma^2_{k\ell}} \sum_{i,j} z_{ik} w_{j\ell} (x_{ij} - \mu_{k\ell})^2 \right) \end{aligned}$$

Link between LBCEM and Croeuc

Criterion

Parsimonious model can be defined by imposing constraints on the variances: we obtain the $[\sigma], [\sigma_k], [\sigma^j], \ldots$
In the simplest case, the $[\sigma]$ model, given identical proportions ($\pi_k = 1/g, \rho_\ell = 1/s$)

$$L_C(\mathbf{Z}, \mathbf{W}, \Theta) = -\frac{nd}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i,j,k,\ell} z_{ik} w_{j\ell}(x_{ij} - \mu_{k\ell})^2 - n \log g - d \log s$$

and it is easy to see that maximizing L_C is equivalent to minimizing $W(\mathbf{Z}, \mathbf{W})$ where

$$W(\mathbf{Z}, \mathbf{W}) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell}(x_{ij} - \mu_{k\ell})^2 \text{ minimized by Croeuc}$$

Assignation steps

It suffices to remark that in step 1 of LBCEM we have

$$z_i = \operatorname*{argmax}_k \log \pi_k - \frac{1}{2} \sum_\ell w_{.\ell} \left(\log \sigma^2_{k\ell} + \frac{u^{\mathbf{w}}_{i\ell} - 2\mu_{k\ell} x^{\mathbf{w}}_{i\ell} + \mu^2_{k\ell}}{\sigma^2_{k\ell}} \right).$$

For the $[\sigma]$ model, this leads to $z_i = \operatorname{argmin}_k \sum_\ell w_{.\ell}(x^{\mathbf{w}}_{i\ell} - \mu_{k\ell})^2$. In the same way we can prove that in step 3 of LBCEM we have $w_j = \operatorname{argmin}_\ell \sum_k z_{.k}(x^{\mathbf{z}}_{kj} - \mu_{k\ell})^2$

Outline

1 Co-clustering

2 Latent block model
- Bernoulli LBM
- Gaussian LBM

3 CML and ML approaches
- CML approach
- Links between BCEM and classical co-clustering algorithms
- Gaussian latent block models

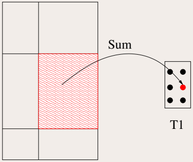
4 Contingency Table
- Criteria and Algorithms

5 Latent block model
- Poisson Latent Block Model
- Algorithms

6 Conclusion

Contingency table

- Summary of $\mathbf{X}$ can be obtained by



- $\mathbf{X}$ and $\mathbf{Y}$ have the same structure $\mathcal{A}(\mathbf{X}) \geq \mathcal{A}(\mathbf{Y})$
- Problem: Find partitions $\mathbf{z}$ and $\mathbf{w}$ maximizing $\mathcal{A}(\mathbf{z}, \mathbf{w})$. The $(\mathbf{z}, \mathbf{w})$ is obtained in making the sums of values per block

	$v1$	$v2$	$v3$	$v4$	$v5$	Z
a	5	0	0	0	0	1
b	0	2	0	1	1	1
c	0	0	1	4	0	2
d	1	0	0	0	1	1
e	2	0	1	3	1	2
f	0	4	1	1	1	2
W	1	1	2	2	1	

	$v1$	$v2$	$v5$	$v3$	$v4$
a	5	0	0	0	0
b	0	2	2	0	1
d	1	0	1	0	0
c	0	0	0	1	4
e	1	0	1	1	3
f	0	0	1	1	1

11	1
3	11

- Solution: Alternated maximization of $\mathcal{A}(\mathbf{z}, J)$ and $\mathcal{A}(I, \mathbf{w})$
- Idea: Alternated application of k-means (nuées dynamiques, Diday 1971) with an appropriate metric on intermediate reduced matrices of size $(g \times d)$ and $(n \times s)$

Connections between these approaches

There exists a large variety of co-clustering methods for contingency tables (Govaert (1983), Bock (1992, 2003)) which can be applied in document clustering context.

Let $\mathbf{X}$ be a two-way contingency table associated to two categorical random variables that take values in sets $I = \{1, \ldots, i, \ldots, n\}$ and $J = \{1, \ldots, j, \ldots, d\}$. The entries x_{ij} are co-occurrences of row and column categories, each of them counts the number of entities that fall simultaneously in the corresponding row and column categories.

Let $P_{IJ} = (p_{ij})$ denote the sample joint probability distribution. It is a matrix of size $n \times d$ defined by $p_{ij} = \frac{x_{ij}}{N}$ where $N = \sum_{ij} x_{ij}$. The sample marginal probability distributions are defined by $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$.

	$\mathbf{1}$	$\ldots$	j	$\ldots$	d	
$\mathbf{1}$	$x_{\mathbf{1}j}$	$\ldots$	$x_{\mathbf{1}j}$	$\ldots$	$x_{\mathbf{1}d}$	$x_{\mathbf{1}.}$
			$\vdots$	$\ddots$		
i	$x_{i\mathbf{1}}$	$\ldots$	x_{ij}	$\ldots$	x_{id}	$x_{i.}$
			$\vdots$	$\ddots$		
n	$x_{n\mathbf{1}}$	$\ldots$	x_{nj}	$\ldots$	x_{nd}	$x_{n.}$
	$x_{.\mathbf{1}}$	$\ldots$	$x_{.j}$	$\ldots$	$x_{.d}$	N

	$\mathbf{1}$	$\ldots$	j	$\ldots$	d	
$\mathbf{1}$	$p_{\mathbf{1}j}$	$\ldots$	$p_{\mathbf{1}j}$	$\ldots$	$p_{\mathbf{1}d}$	$p_{\mathbf{1}.}$
			$\vdots$	$\ddots$		
i	$p_{i\mathbf{1}}$	$\ldots$	p_{ij}	$\ldots$	p_{id}	$p_{i.}$
			$\vdots$	$\ddots$		
n	$p_{n\mathbf{1}}$	$\ldots$	p_{nj}	$\ldots$	p_{nd}	$p_{n.}$
	$p_{.\mathbf{1}}$	$\ldots$	$p_{.j}$	$\ldots$	$p_{.d}$	$\mathbf{1}$

Measures of association

Introduction

The contingency table characterizes the dependency links between the two sets, and measuring the strength of this association is a long tradition in statistics, going back to at least Pearson (1900).

Phi-squared: $\Phi^2(P_{IJ}) = \frac{\chi^2(\mathbf{X})}{N} = \sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \sum_{i,j} \frac{p_{ij}^2}{p_{i.}p_{.j}} - 1$

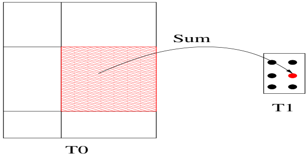
This coefficient can be seen as an estimation of the deviation between the probabilities $\xi_{i.}\xi_{.j}$, that we would have if the two categorical random variables were independent, and the probabilities ξ_{ij}

Mutual Information: $\mathrm{I}(P_{IJ}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}}$

This measure of association is defined by $\mathrm{I}(P_{IJ}) = H(P_I) + H(P_J) - H(P_{IJ})$ where $H(P_I)$, $H(P_J)$ are the marginal entropies, $H(P_{IJ})$ is the joint entropy of I and J.

Contingency table associated to co-clustering $(\mathbf{z}, \mathbf{w})$

A two-way contingency table $\mathbf{Y}^{\mathbf{zw}} = (y^{\mathbf{zw}}_{k\ell})$ associated to two categorical random variables that take values in sets $K = \{1, \ldots, \ldots, g\}$ and $L = \{1, \ldots, \ldots, s\}$. It can be obtained from the initial table in computing the sum of the rows and columns according to the partitions $\mathbf{z}$ and $\mathbf{w}$:
$y^{\mathbf{zw}}_{k\ell} = \sum_{i,j} z_{ik} w_{j\ell} x_{ij} \qquad \forall k \in K \quad \text{and} \quad \forall \ell \in L$



Distribution associated to z and w defined on $K \times L$

The distribution $P^{\mathbf{zw}}_{KL} = (p^{\mathbf{zw}}_{k\ell})$ defined on $K \times L$ by

$$p^{\mathbf{zw}}_{k\ell} = \frac{y^{\mathbf{zw}}_{k\ell}}{N} = \sum_{i,j} z_{ik} w_{j\ell}\, p_{ij} \qquad \forall (k, \ell) \in K \times L.$$

the row margins $\sum_\ell p^{\mathbf{zw}}_{k\ell}$ of $P^{\mathbf{zw}}_{KL}$ are equal to $p^{\mathbf{z}}_{k.} = \sum_i z_{ik} p_{i.}$ and then do not depend on the partition $\mathbf{w}$. Similarly, the column margins $\sum_k p^{\mathbf{zw}}_{k\ell}$ are equal to $p^{\mathbf{w}}_{.\ell} = \sum_j w_{j\ell} p_{.j}$.

Distribution associated to z and w defined on $I \times J$

The second distribution $Q_{IJ}^{\mathbf{zw}} = (q_{ij}^{\mathbf{zw}})$ defined on $I \times J$

$$q_{ij}^{\mathbf{zw}} = q_{i.}^{\mathbf{z}} q_{.j}^{\mathbf{w}} \sum_{k,\ell} z_{ik} w_{j\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} \qquad \forall (i,j) \in I \times J$$

We have $\sum_{i,j} q_{ij}^{\mathbf{zw}} = 1, \quad q_{i.}^{\mathbf{z}} = p_{i.} \quad \text{and} \quad q_{.j}^{\mathbf{w}} = p_{.j} \qquad \forall i,j$. We have the same margins as the initial distribution P_{IJ}.

$$q_{ij}^{\mathbf{zw}} = p_{i.} p_{.j} \sum_{k,\ell} z_{ik} w_{j\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} \qquad \forall (i,j) \in I \times J$$

	1	2	3	4	5	
1	0.050	0.040	0.060	0.010	0.000	0.160
2	0.060	0.050	0.040	0.000	0.010	0.160
3	0.010	0.000	0.010	0.070	0.050	0.140
4	0.010	0.010	0.000	0.060	0.050	0.130
5	0.040	0.050	0.030	0.040	0.050	0.210
6	0.050	0.040	0.040	0.030	0.040	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

	1	2	3	4	5	
1	0.056	0.048	0.046	0.005	0.005	0.160
2	0.056	0.048	0.046	0.005	0.005	0.160
3	0.008	0.007	0.006	0.061	0.058	0.140
4	0.007	0.006	0.006	0.057	0.054	0.130
5	0.048	0.041	0.039	0.042	0.040	0.210
6	0.045	0.039	0.037	0.040	0.038	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

Table: Distributions P_{IJ} (left) and $Q_{IJ}^{\mathbf{zw}}$ (right)

Measures of associations associated to z and w

Using the two measures phi-squared and mutual information applied on the two distributions previously defined, we obtain the following measures:

$$\Phi^2(P_{KL}^{\mathbf{zw}}) = \sum_{k,\ell} \frac{(p_{k\ell}^{\mathbf{zw}} - p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}})^2}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} = \Phi^2(Q_{IJ}^{\mathbf{zw}}) = \sum_{i,j} \frac{(q_{ij}^{\mathbf{zw}} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

$$\mathrm{I}(P_{KL}^{\mathbf{zw}}) = \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}} \log \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} = \mathrm{I}(Q_{IJ}^{\mathbf{zw}}) = \sum_{i,j} q_{ij}^{\mathbf{zw}} \log \frac{q_{ij}^{\mathbf{zw}}}{p_{i.}p_{.j}}.$$

Proposition $\Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{\mathbf{zw}}) = \Phi^2(P_{IJ}) - \Phi^2(P_{KL}^{\mathbf{zw}}) = D_{\Phi^2}(P_{IJ}||Q_{IJ}^{\mathbf{zw}})$

where $D_{\Phi^2}(P_{IJ}||Q_{IJ}^{\mathbf{zw}}) = \sum_{i,j} \frac{(p_{ij} - q_{ij}^{\mathbf{zw}})^2}{p_{i.}p_{.j}} = \sum_{i,j} p_{ij} \left(\frac{p_{ij}}{p_{i.}p_{.j}} - \frac{q_{ij}^{\mathbf{zw}}}{p_{i.}p_{.j}} \right)$ *can be viewed as a* Φ^2 *distance between the two distributions* P_{IJ} *and* $Q_{IJ}^{\mathbf{zw}}$

Proposition $\mathrm{I}(P_{IJ}) - \mathrm{I}(Q_{IJ}^{\mathbf{zw}}) = \mathrm{I}(P_{IJ}) - \mathrm{I}(P_{KL}^{\mathbf{zw}}) = \mathrm{KL}(P_{IJ}||Q_{IJ}^{\mathbf{zw}})$

where $\mathrm{KL}(P_{IJ}||Q_{IJ}^{\mathbf{zw}}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}^{\mathbf{zw}}}$ *is the Kullback-Leibler between the two distributions* P_{IJ} *and* $Q_{IJ}^{\mathbf{zw}}$,

$$\mathrm{I}(Q_{IJ}^{\mathbf{zw}}) \leq \mathrm{I}(P_{IJ}) \quad or \quad \mathrm{I}(P_{KL}^{\mathbf{zw}}) \leq \mathrm{I}(P_{IJ}).$$

1. Co-clustering obtained by reducing the size of the contingency table

- Looking for a good co-clustering can be seen as a way to obtain a good summary of the data. The objective of co-clustering can be based on minimizing

$$\Phi^2(P_{IJ}) - \Phi^2(P^{\mathbf{zw}}_{KL}) \quad \text{or} \quad \mathrm{I}(P_{IJ}) - \mathrm{I}(P^{\mathbf{zw}}_{KL}).$$

2. Co-clustering obtained by approximating the original distribution

- The co-clustering problem can be viewed as an approximation of the distribution P_{IJ} by a distribution according to co-clustering termed $Q^{\mathbf{zw}}_{IJ}$ by minimizing the difference between the measures of information of the original distribution and the new distribution:

$$\Phi^2(P_{IJ}) - \Phi^2(Q^{\mathbf{zw}}_{IJ}) \quad \text{or} \quad \mathrm{I}(P_{IJ}) - \mathrm{I}(Q^{\mathbf{zw}}_{IJ}) \text{ criterion optimized by Dhillon et al. (2003)}$$

Objective functions based on measures of association

- Phisquare coefficient Φ^2
- Mutual Information I
- Csizar's Φ divergence: $\sum_{k,\ell} p^{\mathbf{z}}_{k.} p^{\mathbf{w}}_{.\ell} \phi(\frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.} p^{\mathbf{w}}_{.\ell}})$

Phi-squared coefficient

The criterion for the Φ^2 measure of association

$$W_{\Phi^2}(\mathbf{z},\mathbf{w}) = D_{\Phi^2}(P_{IJ}||P^{\mathbf{zw}}_{KL}) = \Phi^2(P_{IJ}) - \Phi^2(Q^{\mathbf{zw}}_{IJ})$$

We introduce a new parameter $\delta = (\delta_{k\ell})$, a matrix of size (g,s) where each $\delta_{k\ell}$ plays the role of centroid of the block $k\ell$ and such that $\delta_{k\ell} > 0 \quad \forall k;\ell \quad \text{and} \quad \sum_{k,\ell} p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}\delta_{k\ell} = 1$.
Using this parameter, a new distribution $R^{\mathbf{zw}\delta}_{IJ} = (r^{\mathbf{zw}\delta}_{ij})$ depending on **z**,**w** and parameter δ can be defined by:

$$r^{\mathbf{zw}\delta}_{ij} = p_{i.}p_{.j}\sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell}.$$

These constraints ensure that $R^{\mathbf{zw}\delta}_{IJ}$ is a distribution. This distribution $R^{\mathbf{zw}\delta}_{IJ}$ has the same column and row margins that the distributions P_{IJ} and $Q^{\mathbf{zw}}_{IJ}$:

$$r^{\mathbf{zw}\delta}_{i.} = p_{i.} \quad \text{and} \quad r^{\mathbf{zw}\delta}_{.j} = p_{.j} \qquad \forall i,j.$$

Using this new distribution, the objective of co-clustering is replaced by minimizing the new criterion for the Φ^2 measure of association

$$\widetilde{W}_{\Phi^2}(\mathbf{z},\mathbf{w},\delta) = \Phi^2(P_{IJ}) - \Phi^2(R^{\mathbf{zw}\delta}_{IJ})$$

Algorithm

$$\widetilde{W}_{\Phi^2}(\mathbf{z},\mathbf{w},\delta)=\sum_{i,j,k;\ell} z_{ik}w_{j\ell}\,p_{i.}p_{.j}\left(\frac{p_{ij}}{p_{i.}p_{.j}}-\delta_{k\ell}\right)^2$$

Algorithm 1 Croki2

input: contingency table $\mathbf{X}$, g and s the desired numbers of row column clusters;
output: partitions $\mathbf{z}$ and $\mathbf{w}$;
initialization:
. start with some initial partitions $\mathbf{z}$, $\mathbf{w}$; $\delta_{k\ell} \leftarrow \frac{p^{\mathbf{zw}}_{k\ell}}{(p^{\mathbf{z}}_{k.})(p^{\mathbf{w}}_{.\ell})}$;
repeat

step 1. $\mathbf{z}$: each i is assigned to the cluster k minimizing $\sum_{j,\ell} w_{j\ell}p_{.j}\left(\frac{p_{ij}}{p_{i.}p_{.j}}-\delta_{k\ell}\right)^2$;

step 2. $\delta_{k\ell} \leftarrow \frac{p^{\mathbf{zw}}_{k\ell}}{(p^{\mathbf{z}}_{k.})(p^{\mathbf{w}}_{.\ell})}$;

step 3. $\mathbf{w}$: each j is assigned to the cluster ℓ minimizing $\sum_{i,k} z_{ik}p_{i.}(\frac{p_{ij}}{p_{i.}p_{.j}}-\delta_{k\ell})^2$;

step 4. $\delta_{k\ell} \leftarrow \frac{p^{\mathbf{zw}}_{k\ell}}{(p^{\mathbf{z}}_{k.})(p^{\mathbf{w}}_{.\ell})}$;

until the change in objective function value $W_{\Phi^2}(\mathbf{z},\mathbf{w},\delta)$ is "small" (say 10^{-6})
return $\mathbf{z}$ and $\mathbf{w}$

Mutual information criterion and Algorithm

$$\widetilde{W}_{\mathrm{I}}(\mathbf{z},\mathbf{w},\delta)=\mathrm{I}(P_{IJ})-\mathrm{I}(R_{IJ}^{\mathbf{zw}\delta})=\sum_{i,j}p_{ij}\log\frac{p_{ij}}{p_{i.}p_{.j}}-\sum_{k,\ell}p_{k\ell}^{\mathbf{zw}}\log\delta_{k\ell}.$$

Algorithm 2 Croinfo

input: contingency table $\mathbf{X}$, g and s the desired numbers of row column clusters;
output: partitions $\mathbf{z}$ and $\mathbf{w}$;
initialization:
. start with some initial partitions $\mathbf{z}$, $\mathbf{w}$; $\delta_{k\ell} \leftarrow \frac{p_{k\ell}^{\mathbf{zw}}}{(p_{k.}^{\mathbf{z}})(p_{.\ell}^{\mathbf{w}})}$;
repeat
step 1. $\mathbf{z}$: each row i is assigned to the cluster k minimizing $\sum_\ell(\sum_j w_{j\ell}p_{ij})\log\delta_{k\ell}$;
step 2. $\delta_{k\ell} \leftarrow \frac{p_{k\ell}^{\mathbf{zw}}}{(p_{k.}^{\mathbf{z}})(p_{.\ell}^{\mathbf{w}})}$;
step 3. $\mathbf{w}$: each column j is assigned to the cluster ℓ minimizing $\sum_k(\sum_i z_{ik}p_{ij})\log\delta'_{k\ell}$;
step 4. $\delta_{k\ell} \leftarrow \frac{p_{k\ell}^{\mathbf{zw}}}{(p_{k.}^{\mathbf{z}})(p_{.\ell}^{\mathbf{w}})}$;
until the change in objective function value $W_{\mathrm{I}}(\mathbf{z},\mathbf{w},\delta)$ is "small" (say 10^{-6})
return $\mathbf{z}$ and $\mathbf{w}$

This algorithm monotonically decreases the criterion: $W_{\mathrm{I}}(\mathbf{z}^{(t)},\mathbf{w}^{(t)})\geq W_{\mathrm{I}}(\mathbf{z}^{(t+1)},\mathbf{w}^{(t+1)})$ and that its convergence properties are the same that the properties of Croki2.

Example

	1	2	3	4	5	
1	5	4	6	1	0	16
2	6	5	4	0	1	16
3	1	0	1	7	5	14
4	1	1	0	6	5	13
5	4	5	3	4	5	21
6	5	4	4	3	4	20
	22	19	18	21	20	100

	1	2	3	4	5	
1	0.05	0.04	0.06	0.01	0.00	0.16
2	0.06	0.05	0.04	0.00	0.01	0.16
3	0.01	0.00	0.01	0.07	0.05	0.14
4	0.01	0.01	0.00	0.06	0.05	0.13
5	0.04	0.05	0.03	0.04	0.05	0.21
6	0.05	0.04	0.04	0.03	0.04	0.20
	0.22	0.19	0.18	0.21	0.20	1.00

Table: Example of contingency table and associated joint distribution

Approximation of P_{IJ} by $R_{IJ}^{\mathbf{zw}\delta}$

	1	2	3	4	5	
1	0.050	0.040	0.060	0.010	0.000	0.160
2	0.060	0.050	0.040	0.000	0.010	0.160
3	0.010	0.000	0.010	0.070	0.050	0.140
4	0.010	0.010	0.000	0.060	0.050	0.130
5	0.040	0.050	0.030	0.040	0.050	0.210
6	0.050	0.040	0.040	0.030	0.040	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

	1	2	3	4	5	
1	0.056	0.048	0.046	0.005	0.005	0.160
2	0.056	0.048	0.046	0.005	0.005	0.160
3	0.008	0.007	0.006	0.061	0.058	0.140
4	0.007	0.006	0.006	0.057	0.054	0.130
5	0.048	0.041	0.039	0.042	0.040	0.210
6	0.045	0.039	0.037	0.040	0.038	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

Table: Distributions P_{IJ} (left) and $Q_{IJ}^{\mathbf{zw}}$ (right)

Outline

1. Co-clustering
2. Latent block model
   - Bernoulli LBM
   - Gaussian LBM
3. CML and ML approaches
   - CML approach
   - Links between BCEM and classical co-clustering algorithms
   - Gaussian latent block models
4. Contingency Table
   - Criteria and Algorithms
5. **Latent block model**
   - Poisson Latent Block Model
   - Algorithms
6. Conclusion

Latent block model

Definition

The pdf of **X**:

$$f(\mathbf{X};\Theta) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_i \pi_{z_i} \prod_j \rho_{w_j} \prod_{i,j} \varphi(X_{ij};\boldsymbol{\alpha}_{z_i w_j})$$

where $\Theta = (\pi_1,\ldots,\pi_g;\rho_1,\ldots,\rho_s;\boldsymbol{\alpha}_{11},\ldots,\boldsymbol{\alpha}_{gs})$

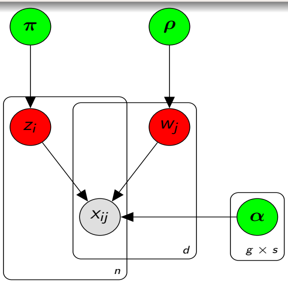


Figure: LBM as a graphical model

Advantages in this case; Govaert and Nadif, 2016

- Parsimonious models giving probabilistic interpretations of classical criteria

Definition

In the Poisson latent block mixture model (PLBM), **z** and **w** are considered as latent random variables and it is assumed that for each block k, ℓ the values x_{ij} are distributed according to the Poisson distribution $\mathcal{P}(\lambda_{ij})$ with

$$\lambda_{ij} = \mu_i \nu_j \sum_{k,\ell} z_{ik} w_{j\ell} \gamma_{k\ell}$$

for a row effect μ_i, a column effect ν_j and a block effect $\gamma_{k\ell}$.

Approaches and criteria

1. Maximization of the complete data log-likelihood can be written, up to a constant, as $(z_{.k} = \sum_i z_{ik}$ and $w_{.\ell} = \sum_j w_{j\ell})$

$$\mathrm{L_C}(\Theta, \mathbf{Z}, \mathbf{W}) = \sum_k z_{.k} \log \pi_k + \sum_\ell w_{.\ell} \log \rho_\ell + \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} \log \gamma_{k\ell} - x_{i.} x_{.j} \gamma_{k\ell})$$

Maximization of $\mathrm{L_C}(\Theta, \mathbf{Z}, \mathbf{W})$

Algorithm 3 Poisson LBCEM

input: contingency table $\mathbf{X}$, and g, s the desired numbers of row and column clusters.
output: parameters $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and γ
initialization: start with some initial partitions $\mathbf{z}$, $\mathbf{w}$;

$$\pi_k \leftarrow \frac{z_{.k}}{n},\ \rho_\ell \leftarrow \frac{w_{.\ell}}{d} \text{ and } \gamma_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{k\ell} x_{ij}}{\sum_i z_{ik} x_{i.} \sum_j w_{j\ell} x_{.j}};$$

repeat

- $z_{ik} \leftarrow \operatorname{argmax}_k \pi_k \exp(\sum_\ell (\sum_j w_{j\ell} x_{.j}) \log \gamma_{k\ell});$
- $\pi_k \leftarrow \frac{z_{.k}}{n},\ \gamma_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{k\ell} x_{ij}}{\sum_i z_{ik} x_{i.} \sum_j w_{j\ell} x_{.j}};$
- $w_{j\ell} \leftarrow \operatorname{argmax}_\ell \rho_\ell \exp(\sum_k (\sum_i z_{ik} x_{i.}) \log \gamma_{k\ell});$
- $\rho_\ell \leftarrow \frac{w_{.\ell}}{d}$ and $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{j\ell} x_{ij}}{\sum_i z_{ik} x_{i.} \sum_j w_{j\ell} x_{.j}};$

until the change in objective function value $\mathrm{L_C}(\Theta, \mathbf{Z}, \mathbf{W})$ is "small" (say 10^{-6}).
return $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and γ

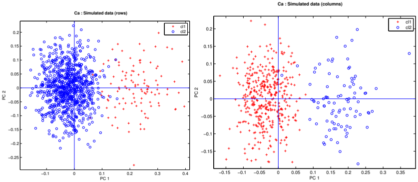
Poisson LBM

- Constraints: $\pi_1 = \ldots = \pi_g,\ \rho_1 = \ldots = \rho_s$

$$\underset{\mathbf{Z},\Theta,\mathbf{W}}{\operatorname{argmax}} L_{RC}(\Theta, \mathbf{Z}, \mathbf{W}) \equiv \underset{\mathbf{Z},\mathbf{W}}{\operatorname{argmax}} \mathrm{I}(P_{IJ}^{\mathbf{zw}}) \approx \underset{\mathbf{Z},\mathbf{W}}{\operatorname{argmax}} \boldsymbol{\Phi}^2(P_{IJ}^{\mathbf{zw}})$$

- Poisson LBCEM=Croinfo $\approx$ Croki2

Croinfo vs Poisson LBCEM We simulate a dataset with proportions dramatically different ($\pi_1 = \frac{110}{1000}, \pi_2 = \frac{990}{1000}$) and ($\rho_1 = \frac{89}{500}, \rho_2 = \frac{411}{500}$) we observe the projection of the sets of rows and columuns by CA.



From confusion matrices obtained by the application of Croinfo, Poisson LBCEM, we note the good performance of Poisson BCEM due to the role of proportions in this situation

	$\mathbf{z}_1^T$	$\mathbf{z}_2^T$
$\hat{\mathbf{z}}_1$	**109**	**1**
$\hat{\mathbf{z}}_2$	**213**	**677**

	$\mathbf{w}_1^T$	$\mathbf{w}_2^T$
$\hat{\mathbf{w}}_1$	**336**	**75**
$\hat{\mathbf{w}}_2$	**1**	**88**

Table: Croinfo: $(\mathbf{z}, \mathbf{w})$ vs $(\mathbf{z}^T, \mathbf{w}^T)$

	$\mathbf{z}_1^T$	$\mathbf{z}_2^T$
$\hat{\mathbf{z}}_1$	**103**	**7**
$\hat{\mathbf{z}}_2$	**1**	**889**

	$\mathbf{w}_1^T$	$\mathbf{w}_2^T$
$\hat{\mathbf{w}}_1$	**411**	**0**
$\hat{\mathbf{w}}_2$	**2**	**87**

Table: Poisson LBCEM: $(\mathbf{z}, \mathbf{w})$ vs $(\mathbf{z}^T, \mathbf{w}^T)$

ML approach

Variationnal EM

Variational approximation by imposing a constraint on the joint distribution of the labels (Govaert and Nadif, 2005, 2008, 2010, 2016)

Algorithm 4 Poisson VEM

input: contingency table $\mathbf{X}$, and g, s the desired numbers of row and column clusters.
output: parameters $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and γ
initialization: start with some initial partitions $\widetilde{\mathbf{z}}$, $\widetilde{\mathbf{w}}$;

$$\pi_k \leftarrow \frac{\widetilde{z}_{.k}}{n},\ \rho_\ell \leftarrow \frac{\widetilde{w}_{.\ell}}{d} \text{ and } \gamma_{k\ell} \leftarrow \frac{\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{k\ell}x_{ij}}{\sum_i \widetilde{z}_{ik}x_{i.}\sum_j \widetilde{w}_{j\ell}x_{.j}};$$

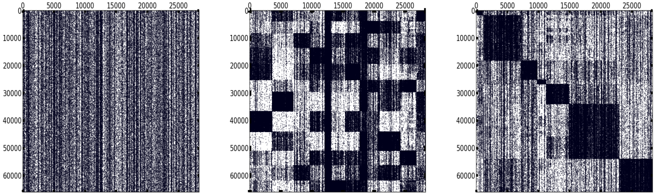
repeat

- $\tilde{z}_{ik} \leftarrow \frac{\pi_k \exp(\sum_\ell(\sum_j \widetilde{w}_{j\ell}X_{.j}) \log \gamma_{k\ell})}{\sum_{k'} \pi_{k'} \exp(\sum_\ell(\sum_j \widetilde{w}_{j\ell}x_{.j}) \log \gamma_{k'\ell})}$;
- $\pi_k \leftarrow \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{k\ell}x_{ij}}{\sum_i \widetilde{z}_{ik}x_{i.}\sum_j \widetilde{w}_{j\ell}x_{.j}}$;
- $\tilde{w}_{j\ell} \leftarrow \frac{\rho_\ell \exp(\sum_k(\sum_i \widetilde{z}_{ik}x_{i.}) \log \gamma_{k\ell})}{\sum_{\ell'} \rho_{\ell'} \exp(\sum_k(\sum_i \widetilde{z}_{ik}x_{i.}) \log \gamma_{k\ell})}$;
- $\rho_\ell \leftarrow \frac{\widetilde{w}_{.\ell}}{d}$ and $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{j\ell}x_{ij}}{\sum_i \widetilde{z}_{ik}x_{i.}\sum_j \widetilde{w}_{j\ell}x_{.j}}$;

until the change in objective function value $F_C(\tilde{\mathbf{Z}}, \tilde{\mathbf{W}}, \Theta)$ is "small" (say 10^{-6}).
return $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and γ

Handling large datasets

65991 $\times$ 28327, sparsity=99.75%, balance=0.006



LBM

- Poisson LBM via Poisson LBCEM/VEM detects the second structure
- However, LBM can be adapted to detect the third structure (Ailem et al. 2016)
    - document clusters
    - term clusters

Outline

1. Co-clustering
2. Latent block model
   - Bernoulli LBM
   - Gaussian LBM
3. CML and ML approaches
   - CML approach
   - Links between BCEM and classical co-clustering algorithms
   - Gaussian latent block models
4. Contingency Table
   - Criteria and Algorithms
5. Latent block model
   - Poisson Latent Block Model
   - Algorithms
6. **Conclusion**

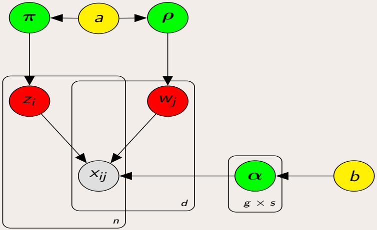
Approaches

- Factorization, spectral, graph, fuzzy, PLSA, LDA and LBM
- Poisson LBM is flexible and parsimonious
- Unlike factorization approaches, the Poisson LBM does not require any transformation of data. They give coherent document and term clusters
- `https://pypi.python.org/pypi/coclust` (Role et al. 2016)

Limits of LBM and derived algorithms

- Symmetric model
- LBCEM, VEM: sensitive to strating values and tend to provide empty clusters

Graphical model (Keribin et al. 15)



- V-Bayes algorithm
- $\mathcal{D}(a, \ldots, a)$, $\mathcal{D}(b, \ldots, b)$
- a=b=1 involves no regularisation (VEM)