Co-clustering for large datasets

Mohamed Nadif

LIPADE, Université Paris Descartes, France

Travaux menés avec G. Govaert et L. Lazhar

Outline

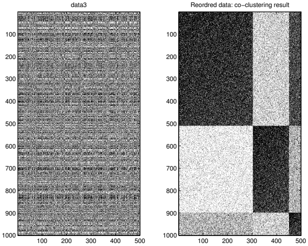
Simultaneous clustering on both dimensions

- The co-clustering methods have attracted much attention in recent years
- The block clustering had an influence in applied mathematics from the sixties (Jennings, 1968)
- First works in J.A. Hartigan, Direct Clustering of a Data Matrix (1972)
- Works of Govaert (1983)
- Referred in the literature as bi-clustering, co-clustering, double clustering, direct clustering, coupled clustering
- Different approaches (see for instance chapter 1, Govaert and Nadif 2013),
- These approaches can differ in the pattern they seek and the types of data they apply to
- Organization of the data matrix into homogeneous blocks or extraction of co-clusters
  - no-overlapping co-clustering
  - overlapping co-clustering

Aim

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

Example of co-clustering



Why co-clustering ?

- (1) : Utilizing duality of clustering
- (2) : Reducing running time
- (3) : Discovering hidden latent patterns and generating compact representation
- (4) : Reducing dimensionality implicitly
- (5) : High dimension

Applications and approaches

Fields

- Text mining: clustering of documents and words simultaneously
- Bioinformatics: clustering of genes and tissus simultaneously
- Collaborative Filtering
- Social Network Analysis

Approaches

- Spectral
- Factorization
- Latent block models
- etc.

Softwares

- Package {biclust} in **R**, Bicat, etc.
- R {blockcluster}

Notations

- Let be $\mathbf{x} = (x_{ij})$ of size $n \times d$, $i \in I$ set of n rows, $j \in J$ set of d columns

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

- $\mathbf{z} = (z_1, \ldots, z_n) \longrightarrow (z_{ik})$
- z_i cluster indicator of $i \Longrightarrow z_{ik} = 1$ if $i \in k^{th}$ cluster and $z_{ik} = 0$ otherwise
- $z_{.k}$ cardinality of k^{th} cluster, $k \in \{1, \ldots, g\}$

z_i	z_{i1}	z_{i2}	z_{i3}
3	0	0	1
2	0	1	0
3	0	0	1
2	0	1	0
1	1	0	0

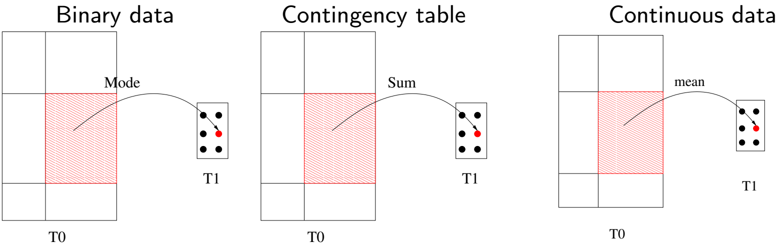
Partition $\mathbf{w}$ of J in m clusters

- $\mathbf{w} = (w_1, \ldots, w_d) \longrightarrow (w_{j\ell})$
- w_j cluster indicator of $j \Longrightarrow w_{j\ell} = 1$ if $j \in \ell^{th}$ cluster and $w_{j\ell} = 0$ otherwise
- $w_{.\ell}$ cardinality of ℓ^{th} cluster, $\ell \in \{1, \ldots, m\}$

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

- Block formed by the couple k^{th} and ℓ^{th} clusters is defined by the x_{ij}'s with $z_{ik} w_{j\ell} = 1$

General principle



Criteria

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

Notations and example

	1	2	3	4	5	6	7	8	9	10
a	1	0	1	0	1	0	0	1	0	1
b	0	1	0	1	0	1	1	0	1	0
c	1	0	0	0	0	0	0	1	1	0
d	1	0	1	0	0	0	0	1	0	0
e	0	1	0	1	0	1	1	0	1	0
f	0	1	0	0	0	1	1	0	1	0
g	0	1	0	0	0	0	0	1	0	1
h	1	0	1	0	1	1	0	1	1	1
i	1	0	0	1	0	0	0	0	0	1
j	0	1	0	1	0	0	1	0	0	0

Binary data x

				1					2	
		1	3	5	8	10	2	4	6	7
A	a	1	1	1	1	1	0	0	0	0
	d	1	1	0	1	0	0	0	0	0
	h	1	1	1	1	1	0	0	1	0
B	b	0	0	0	0	0	1	1	1	1
	e	0	0	0	0	0	1	1	1	1
	f	0	0	0	0	0	1	0	1	1
	j	0	0	0	0	0	1	1	0	1
C	c	1	0	0	1	0	0	0	0	0
	g	0	0	0	1	1	1	0	0	0
	i	1	0	0	0	1	0	1	0	0

Reorganized data matrix x

	1	2
A	1	0
B	0	1
C	0	0

Summary matrix a

Matrix	Size	Definition
$\mathbf{x}^{\mathbf{z}} = (x^{\mathbf{z}}_{kj})$	$(g \times d)$	$x^{\mathbf{z}}_{kj} = \sum_i z_{ik} x_{ij}$
$\mathbf{x}^{\mathbf{w}} = (x^{\mathbf{w}}_{i\ell})$	$(n \times m)$	$x^{\mathbf{w}}_{i\ell} = \sum_j w_{j\ell} x_{ij}$
$\mathbf{x}^{\mathbf{zw}} = (x^{\mathbf{zw}}_{k\ell})$	$(g \times m)$	$x^{\mathbf{zw}}_{k\ell} = \sum_{i,j} z_{ik} w_{j\ell} x_{ij}$

Reduced matrices, sizes and definitions of $\mathbf{x}^{\mathbf{z}}$, $\mathbf{x}^{\mathbf{w}}$ and $\mathbf{x}^{\mathbf{zw}}$

Intermediate data matrices x^z, x^w and x^{zw}

		1	3	1 5	8	10	2	4	2 6	7	9
A	a	1	1	1	1	1	0	0	0	0	0
	d	1	1	0	1	0	0	0	0	0	0
	h	1	1	1	1	1	0	0	1	0	1
B	b	0	0	0	0	0	1	1	1	1	1
	e	0	0	0	0	0	1	1	1	1	1
	f	0	0	0	0	0	1	0	1	1	1
	j	0	0	0	0	0	1	1	0	1	0
C	c	1	0	0	1	0	0	0	0	0	1
	g	0	0	0	1	1	1	0	0	0	0
	i	1	0	0	0	1	0	1	0	0	0

$$x^w = \begin{pmatrix} 5 & 0 \\ 3 & 0 \\ 5 & 2 \\ 0 & 5 \\ 0 & 5 \\ 0 & 4 \\ 0 & 3 \\ 2 & 1 \\ 2 & 1 \\ 2 & 1 \end{pmatrix}$$

$$\mathbf{x}^z = \begin{pmatrix} 3 & 3 & 2 & 3 & 2 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 4 & 3 & 3 & 4 & 3 \\ 2 & 0 & 0 & 2 & 2 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{x}^{zw} = \begin{pmatrix} 13 & 2 \\ 0 & 17 \\ 6 & 3 \end{pmatrix}$$

Minimization of the following criterion

$$\mathcal{C}(\mathbf{z}, \mathbf{w}, \mathbf{a}) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|,$$

where $a_{k\ell} \in \{0, 1\}$

Algorithm

Minimization of $\mathcal{C}(\mathbf{z}, \mathbf{w}, \mathbf{a})$ by alternated minimization of $\mathcal{C}(\mathbf{z}, \mathbf{a}|\mathbf{w})$ and $\mathcal{C}(\mathbf{w}, \mathbf{a}|\mathbf{z})$

Crobin (here $\lfloor x \rceil$ is the nearest integer function)

input: $\mathbf{x}$, g, m
initialization: $\mathbf{z}$, $\mathbf{w}$, $a_{k\ell} = \lfloor \frac{x^{\mathbf{zw}}_{k\ell}}{z_{.k} w_{.\ell}} \rceil$
repeat
  $x^{\mathbf{w}}_{i\ell} = \sum_j w_{j\ell} x_{ij}$
  repeat
    step 1. $z_i = \text{argmin}_k \sum_\ell w_{j\ell} |x^{\mathbf{w}}_{i\ell} - w_{.\ell} a_{k\ell}|$
    step 2. $a_{k\ell} = \lfloor \frac{\sum_k z_{ik} x^{\mathbf{w}}_{i\ell}}{z_{.k} w_{.\ell}} \rceil$
  until convergence
  $x^{\mathbf{z}}_{kj} = \sum_i z_{ik} x_{ij}$
  repeat
    step 3. $w_j = \text{argmin}_\ell \sum_k z_{ik} |x^{\mathbf{z}}_{kj} - z_{.k} a_{k\ell}|$
    step 4. $a_{k\ell} = \lfloor \frac{\sum_j w_{j\ell} x^{\mathbf{z}}_{kj}}{z_{.k} w_{.\ell}} \rceil$
  until convergence
until convergence
return $\mathbf{z}$, $\mathbf{w}$, $\mathbf{a}$

Two geometrical representations

- Each individual i is weighted by p_i and each column j is weighted by q_j

$$\mathbf{d}^2(i,i') = \sum_{j=1}^{d} q_j (x_{ij} - x_{i'j})^2 \text{ and } \mathbf{d}^2(j,j') = \sum_{i=1}^{n} p_i (x_{ij} - x_{ij'})^2$$

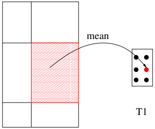
In the sequel, and only to simplify the notation, we assume that $p_i = \frac{1}{n}$ for all i and $q_j = 1$ for all j.

Using a partition $\mathbf{z}$ of I and a partition $\mathbf{w}$ of J, the initial data is summarized by two sets of weights $\mathbf{p}^{\mathbf{z}} = (p_1^{\mathbf{z}}, \ldots, p_g^{\mathbf{z}})$ and $\mathbf{q}^{\mathbf{w}} = (q_1^{\mathbf{w}}, \ldots, q_m^{\mathbf{w}})$ and a $g \times m$ matrix $\mathbf{x}^{\mathbf{zw}} = (x_{k\ell}^{\mathbf{zw}})$ defined by

$$p_k^{\mathbf{z}} = \frac{\sum_i z_{ik}}{n} = \frac{z_{.k}}{n}, \qquad q_\ell^{\mathbf{w}} = \sum_j w_{j\ell} = w_{.\ell}$$

and

$$x_{k\ell}^{\mathbf{zw}} = \frac{\sum_{i,j} z_{ik} w_{j\ell} p_i q_j x_{ij}}{\sum_{i,j} z_{ik} w_{j\ell} p_i q_j} = \frac{\sum_{i,j} z_{ik} w_{j\ell} x_{ij}}{z_{.k} w_{.\ell}}.$$



Example

$$\mathbf{x} = \begin{pmatrix} 1 & 2 & 8 \\ 2 & 1 & 7 \\ 2 & 4 & 7 \\ 4 & 4 & 6 \end{pmatrix}$$

$$\mathbf{p} = (1/4, 1/4, 1/4, 1/4) \text{ and } \mathbf{q} = (1, 1, 1)$$

Let be $\mathbf{z} = (1, 1, 2, 2)$ and $\mathbf{w} = (1, 1, 2)$, we obtain the summary $\mathbf{x}^{\mathbf{zw}}$ with weights

$$\mathbf{p}^{\mathbf{z}} = (1/2, 1/2) \text{ and } \mathbf{q}^{\mathbf{w}} = (2, 1)$$

$\mathbf{x}^{\mathbf{w}} = (x^{\mathbf{w}}_{i\ell})$ of size (4×2) and $\mathbf{x}^{\mathbf{z}} = (x^{\mathbf{z}}_{kj})$ of size (2×3) can be defined

$$x^{\mathbf{w}}_{i\ell} = \frac{\sum_{j,\ell} w_{j\ell} q_j x_{ij}}{\sum_{j,\ell} w_{j\ell} q_j} = \frac{\sum_{j,\ell} w_{j\ell} x_{ij}}{w_{.\ell}} \quad \text{and} \quad x^{\mathbf{z}}_{kj} = \frac{\sum_{i,k} z_{ik} p_i x_{ij}}{\sum_{i,k} p_i z_{ik}} = \frac{\sum_{i,k} z_{ik} x_{ij}}{z_{.k}}$$

$$\mathbf{x}^{\mathbf{z}} = \begin{pmatrix} 1.5 & 1.5 & 7.5 \\ 3 & 4 & 6.5 \end{pmatrix}, \quad \mathbf{x}^{\mathbf{w}} = \begin{pmatrix} 1.5 & 8 \\ 1.5 & 7 \\ 3 & 7 \\ 4 & 6 \end{pmatrix} \quad \text{and} \quad \mathbf{x}^{\mathbf{zw}} = \begin{pmatrix} 1.5 & 7.5 \\ 3.5 & 6.5 \end{pmatrix}$$

Information measures

Let be $(\mathbf{x}^{\mathbf{zw}}, \mathbf{p}^{\mathbf{z}}, \mathbf{q}^{\mathbf{w}})$ associated to $(\mathbf{z}, \mathbf{w})$ and having the same structure that the initial data $(\mathbf{x}, \mathbf{p}, \mathbf{q})$. We can define the information measure

$$\mathcal{I}(\mathbf{x}^{\mathbf{zw}}, \mathbf{p}^{\mathbf{z}}, \mathbf{q}^{\mathbf{w}}) = \sum_{k,\ell} p_k^{\mathbf{z}} q_\ell^{\mathbf{w}} (x_{k\ell}^{\mathbf{zw}})^2 = \frac{1}{n} \sum_{k,\ell} z_{.k} w_{.\ell} (x_{k\ell}^{\mathbf{zw}})^2$$

and the chosen information to approximate

$$\mathcal{I}(\mathbf{x}, \mathbf{p}, \mathbf{q}) = \sum_{i,j} p_i q_j x_{ij}^2 = \frac{1}{n} \sum_{i,j} x_{ij}^2$$

When $\mathbf{x}$ is "column-centered" this information represents in $\mathbb{R}^d$ the inertia of the set I relative to the center of gravity and in $\mathbb{R}^n$ the inertia of the set J relative to the origin. This information measure is the measure used by PCA

Objective function

$$\mathcal{I}(\mathbf{x}, \mathbf{p}, \mathbf{q}) - \mathcal{I}(\mathbf{x}^{\mathbf{zw}}, \mathbf{p}^{\mathbf{z}}, \mathbf{q}^{\mathbf{w}}) = \frac{1}{n} \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - x_{k\ell}^{\mathbf{zw}})^2$$

Let be $(\mathbf{x}^{\mathbf{w}}, \mathbf{p}, \mathbf{q}^{\mathbf{w}})$ obtained when $\mathbf{z}$ is the singleton partition and $(\mathbf{x}^{\mathbf{z}}, \mathbf{p}^{\mathbf{z}}, \mathbf{q})$ obtained when $\mathbf{w}$ is the singleton partition. Hence, we obtain the associated measures of association

$$\mathcal{I}(\mathbf{x}^{\mathbf{z}}, \mathbf{p}^{\mathbf{z}}, \mathbf{q}) = \frac{1}{n} \sum_{k,j} z_{.k} (x^{\mathbf{z}}_{kj})^2 \quad \text{and} \quad \mathcal{I}(\mathbf{x}^{\mathbf{w}}, \mathbf{p}, \mathbf{q}^{\mathbf{w}}) = \frac{1}{n} \sum_{i,\ell} w_{.\ell} (x^{\mathbf{w}}_{i\ell})^2$$

When $\mathbf{w}$ is the partition of singletons, this criterion can be expressed as the loss of information due to $\mathbf{z}$ and, by using the Huygens theorem, it can be shown that

$$\mathcal{I}(\mathbf{x}, \mathbf{p}, \mathbf{q}) - \mathcal{I}(\mathbf{x}^{\mathbf{z}}, \mathbf{p}^{\mathbf{z}}, \mathbf{q}) = \frac{1}{n} \widetilde{\mathrm{W}}(\mathbf{z}|J)$$

where $\widetilde{\mathrm{W}}(\mathbf{z}|J) = \sum_{i,k} z_{ik} \sum_j (x_{ij} - x^{\mathbf{z}}_{kj})^2$ is the intra-class inertia, or within-group sum of squares, minimized by the classical k-means algorithm. Similarly, when $\mathbf{z}$ is the partition of singletons, we have

$$\mathcal{I}(\mathbf{x}, \mathbf{p}, \mathbf{q}) - \mathcal{I}(\mathbf{x}^{\mathbf{w}}, \mathbf{p}, \mathbf{q}^{\mathbf{w}}) = \frac{1}{n} \widetilde{\mathrm{W}}(\mathbf{w}|I)$$

where $\widetilde{\mathrm{W}}(\mathbf{w}|I) = \sum_{j,\ell} w_{j\ell} \sum_i (x_{ij} - x^{\mathbf{w}}_{i\ell})^2$

The minimization of the objective function can be solved by an iterative alternating least-squares optimization procedure. Several equivalent variants of double k-means

Double k-means

Input: $\mathbf{x}$, g, m
Initialization: $\mathbf{z}$, $\mathbf{w}$, $x_{k\ell}^{\mathbf{zw}} = \sum_{i,j} \frac{z_{ik} w_{j\ell} x_{ij}}{z_{.k} w_{.\ell}}$
repeat
  step 1. $z_i = \text{argmin}_k \sum_{j,\ell} w_{j\ell}(x_{ij} - x_{k\ell}^{\mathbf{zw}})^2$
  step 2. $w_j = \text{argmin}_\ell \sum_{i,k} z_{ik}(x_{ij} - x_{k\ell}^{\mathbf{zw}})^2$
  step 3. $x_{k\ell}^{\mathbf{zw}} = \sum_{i,j} \frac{z_{ik} w_{j\ell} x_{ij}}{z_{.k} w_{.\ell}}$
until convergence
return $\mathbf{z}$, $\mathbf{w}$

- Croeuc algorithm (Govaert, 1983)
- As for Crobin, Croeuc is based on reduced intermediate matrices

$$\mathbf{x}^{\mathbf{w}} = (x_{i\ell}^{\mathbf{w}}) \text{ and } \mathbf{x}^{\mathbf{z}} = (x_{kj}^{\mathbf{z}})$$

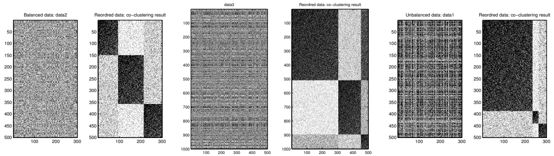
Croeuc

input: $\mathbf{x}$, g, m
initialization: $\mathbf{z}$, $\mathbf{w}$
repeat
$x^{\mathbf{w}}_{i\ell} = \frac{1}{w_{.\ell}} \sum_j w_{j\ell} x_{ij}$, $x^{\mathbf{zw}}_{k\ell} = \frac{1}{z_{.k}} \sum_i z_{ik} x^{\mathbf{w}}_{i\ell}$
repeat
step 1. $z_i = \text{argmin}_k \sum_\ell w_{.\ell}(x^{\mathbf{w}}_{i\ell} - x^{\mathbf{zw}}_{k\ell})^2$
step 2. $x^{\mathbf{zw}}_{k\ell} = \frac{\sum_i z_{ik} x^{\mathbf{w}}_{i\ell}}{z_{.k}}$
until convergence
$x^{\mathbf{z}}_{kj} = \frac{1}{z_{.k}} \sum_i z_{ik} x_{ij}$, $x^{\mathbf{zw}}_{k\ell} = \frac{1}{w_{.\ell}} \sum_j z_{j\ell} x^{\mathbf{z}}_{kj}$
repeat
step 3. $w_j = \text{argmin}_\ell \sum_k z_{.k}(x^{\mathbf{z}}_{kj} - x^{\mathbf{zw}}_{k\ell})^2$
step 4. $x^{\mathbf{zw}}_{k\ell} = \frac{\sum_j w_{j\ell} x^{\mathbf{z}}_{kj}}{w_{.\ell}}$
until convergence
until convergence
return $\mathbf{z}$, $\mathbf{w}$

Weaknesses

Limits of classical co-clustering methods

- $\sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|$, $\sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - a_{k\ell})^2$, $\mathcal{I}(\mathbf{z}, \mathbf{w}) = \sum_{k,\ell} p_{k\ell} \log \frac{p_{k\ell}}{p_{k.} p_{.\ell}}$
- 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

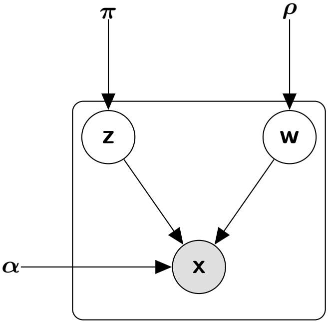
Outline

Definition

The pdf of $\mathbf{x}$:

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

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



Advantages

- Parsimonious models
- Gives probabilistic interpretations of classical criteria via Classification ML approach
- Allows a rigorous simulation (degree of mixtures, proportions)

Binary data: Classical Bernoulli Mixture model

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

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

$\alpha_{k\ell} = (a_{k\ell}, \varepsilon_{k\ell})$ where $a_{k\ell} \in \{0,1\}$ and $\varepsilon_{k\ell} \in]0, 1/2[$

More parsimonious than classical mixture models on I and J

- $n = 10000,\ d = 5000,\ g = 4,\ m = 3$
- Bernoulli latent block model : $4 \times 3 + 3 + 2 = 17$ parameters, Two mixture models : $(4 \times 5000 + 3) + (3 \times 10000 + 2)$ parameters

Classification likelihood

The criterion

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

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

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

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

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

Link between LBCEM and Crobin

Parsimonious models

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

- Fixed proportions: $\pi_1 = \ldots = \pi_g$ and $\rho_1 = \ldots = \rho_m$
- Bernoulli latent model : $\alpha_{k\ell} \to (a_{k\ell}, \varepsilon_{k\ell})$ where $a_{k\ell} \in \{0,1\}$ and $\varepsilon \in]0,1/2[$
- Different models with ε, ε_k, ε_ℓ, $\varepsilon_{k\ell}$

Aim

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

$$L_C(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) = \log(\frac{\varepsilon}{1-\varepsilon}) \sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}| + cst$$

Summary

- Maximization of L_C equivalent to minimization of $\sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|$
- The optimization of $\mathcal{C}$ by *Crobin* assumes strong constraints on the heterogenity of blocks and their proportions
- BCEM=Crobin

Continuous data

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

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

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

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

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

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

Gaussian LBCEM

input: $\mathbf{x}$, g, m

initialization: $\mathbf{z}$, $\mathbf{w}$, $\pi_k = \frac{z_{.k}}{n}$ $\rho_\ell = \frac{w_{.\ell}}{d}$, $\mu_{k\ell} = \frac{x_{k\ell}^{\mathbf{zw}}}{z_{.k}w_{.\ell}}$. $\sigma_{k\ell}^2 = \frac{\sum_{ij} z_{ik}w_{j\ell}x_{ij}^2}{z_{.k}w_{.\ell}} - \mu_{k\ell}^2$

repeat

$x_{i\ell}^{\mathbf{w}} = \frac{1}{w_{.\ell}} \sum_j w_{j\ell}x_{ij}$, $u_{i\ell}^{\mathbf{w}} = \frac{1}{w_{.\ell}} \sum_j w_{j\ell}x_{ij}^2$

repeat

step 1. $z_i = \text{argmax}_k \log \pi_k - \frac{1}{2}\sum_\ell w_{.\ell}\left(\log \sigma_{k\ell}^2 + \frac{u_{i\ell}^{\mathbf{w}} - 2\mu_{k\ell}x_{i\ell}^{\mathbf{w}} + \mu_{k\ell}^2}{\sigma_{k\ell}^2}\right)$

step 2. $\pi_k = \frac{z_{.k}}{n}$, $\mu_{k\ell} = \frac{\sum_i z_{ik}x_{i\ell}^{\mathbf{w}}}{z_{.k}}$, $\sigma_{k\ell}^2 = \frac{\sum_i z_{ik}u_{i\ell}^{\mathbf{w}}}{z_{.k}} - \mu_{k\ell}^2$

until convergence

$x_{kj}^{\mathbf{z}} = \frac{1}{z_{.k}} \sum_i z_{ik}x_{ij}$, $v_{kj}^{\mathbf{z}} = \frac{1}{z_{.k}} \sum_i z_{ik}x_{ij}^2$

repeat

step 3. $w_j = \text{argmax}_\ell \log \rho_\ell - \frac{1}{2}\sum_k z_{.k}\left(\log \sigma_{k\ell}^2 + \frac{v_{kj}^{\mathbf{z}} - 2\mu_{k\ell}x_{kj}^{\mathbf{z}} + \mu_{k\ell}^2}{\sigma_{k\ell}^2}\right)$

step 4. $\rho_\ell = \frac{w_{.\ell}}{d}$, $\mu_{k\ell} = \frac{\sum_j w_{j\ell}x_{kj}^{\mathbf{z}}}{w_{.\ell}}$, $\sigma_{k\ell}^2 = \frac{\sum_j w_{j\ell}v_{kj}^{\mathbf{z}}}{w_{.\ell}} - \mu_{k\ell}^2$

until convergence

until convergence

return $\mathbf{z}$, $\mathbf{w}$, $\boldsymbol{\pi}$, $\boldsymbol{\rho}$,

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 $(\pi_k = 1/g, \rho_\ell = 1/m)$

$$L_C(\mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}) = -\frac{nd}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i,j,k,\ell} z_{ik}w_{j\ell}(x_{ij}-\mu_{k\ell})^2 - n\log g - d\log m$$

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

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

Assignation steps

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

$$z_i = \operatorname*{argmax}_k \log\pi_k - \frac{1}{2}\sum_{\ell} w_{.\ell}\left(\log\sigma_{k\ell}^2 + \frac{u_{i\ell}^{\mathbf{w}} - 2\mu_{k\ell}x_{i\ell}^{\mathbf{w}} + \mu_{k\ell}^2}{\sigma_{k\ell}^2}\right).$$

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

Model

Hereafter, we use a classical mixture model in which the partition $\mathbf{w}$ of the variables is considered as a parameter of the model. The pdf is therefore

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_k \pi_k f(\mathbf{x}_i;\mathbf{w},\boldsymbol{\alpha})$$

with $f(\mathbf{x}_i;\mathbf{w},\boldsymbol{\alpha}) = \prod_{j,\ell}\left(\frac{1}{\sqrt{2\pi\sigma^2_{k\ell}}}e^{-\frac{1}{2\sigma^2_{k\ell}}(x_{ij}-a_{k\ell})^2}\right)^{w_{j\ell}}$. The unknown parameter $\boldsymbol{\theta}$ is formed now by $\boldsymbol{\pi}$, $\mathbf{w}$ and $\boldsymbol{\alpha}$ where $= (\mathbf{a},\Sigma)$ with $\mathbf{a}$ and Σ being $g \times m$ matrices representing the means and the variances of blocks

$$\mathbf{a} = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{g1} & \dots & a_{gm} \end{pmatrix},\ \Sigma = \begin{pmatrix} \sigma^2_{11} & \dots & \sigma^2_{1m} \\ \vdots & \ddots & \vdots \\ \sigma^2_{g1} & \dots & \sigma^2_{gm} \end{pmatrix},$$

or

$$= \begin{pmatrix} (a_{11},\sigma^2_{11}) & \dots & (a_{1m},\sigma^2_{1m}) \\ \vdots & \ddots & \vdots \\ (a_{g1},\sigma^2_{g1}) & \dots & (a_{gm},\sigma^2_{gm}) \end{pmatrix}.$$

Asymmetric Gaussian LBCEM

input: $\mathbf{x}$, g, m

initialization: $\mathbf{z}$, $\mathbf{w}$, $\pi_k = \frac{z_{.k}}{n}$ $\rho_\ell = \frac{w_{.\ell}}{d}$, $\mu_{k\ell} = \frac{x^{\mathbf{zw}}_{k\ell}}{z_{.k}w_{.\ell}}$. $\sigma^2_{k\ell} = \frac{\sum_{ij} z_{ik}w_{j\ell}x^2_{ij}}{z_{.k}w_{.\ell}} - \mu^2_{k\ell}$

repeat

$x^{\mathbf{w}}_{i\ell} = \frac{1}{w_{.\ell}}\sum_j w_{j\ell}x_{ij}$, $u^{\mathbf{w}}_{i\ell} = \frac{1}{w_{.\ell}}\sum_j w_{j\ell}x^2_{ij}$

repeat

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

step 2. $\pi_k = \frac{z_{.k}}{n}$, $\mu_{k\ell} = \frac{\sum_i z_{ik}x^{\mathbf{w}}_{i\ell}}{z_{.k}}$, $\sigma^2_{k\ell} = \frac{\sum_i z_{ik}u^{\mathbf{w}}_{i\ell}}{z_{.k}} - \mu^2_{k\ell}$

until convergence

$x^{\mathbf{z}}_{kj} = \frac{1}{z_{.k}}\sum_i z_{ik}x_{ij}$, $v^{\mathbf{z}}_{kj} = \frac{1}{z_{.k}}\sum_i z_{ik}x^2_{ij}$

repeat

step 3. $w_j = \text{argmax}_\ell \log \rho_\ell - \frac{1}{2}\sum_k z_{.k}\left(\log \sigma^2_{k\ell} + \frac{v^{\mathbf{z}}_{kj} - 2\mu_{k\ell}x^{\mathbf{z}}_{kj} + \mu^2_{k\ell}}{\sigma^2_{k\ell}}\right)$

step 4. $\rho_\ell = \frac{w_{.\ell}}{d}$, $\mu_{k\ell} = \frac{\sum_j w_{j\ell}x^{\mathbf{z}}_{kj}}{w_{.\ell}}$, $\sigma^2_{k\ell} = \frac{\sum_j w_{j\ell}v^{\mathbf{z}}_{kj}}{w_{.\ell}} - \mu^2_{k\ell}$

until convergence

until convergence

return $\mathbf{z}$, $\mathbf{w}$, $\boldsymbol{\pi}$, $\boldsymbol{\rho}$,

Comparisons

- LBVEM: Variational EM
- LBCEM: Classification version of LBVEM.
- EM: EM applied only on the rows.
- CEM: Classification version of EM applied on the rows and columns separately.
- EM-w: Classical EM applied with optimal partition **w** obtained by CEM.
- CEM-w: Classification version of EM-w.

Comparison on 5000×2000 with different degrees of mixtures

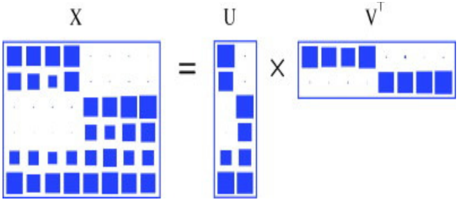
error	Models	LBVEM	LBCEM	CEM	EM	EM-w	CEM-w
$\delta(\mathbf{z}, \mathbf{z}')$	M1	1	1	0	0	1	1
	M2	11	12	21	19	15	15
	M3	29	41	41	39	44	42
$\delta(\mathbf{w}, \mathbf{w}')$	M1	0	0	0	–	0	0
	M2	5	5	30	–	30	30
	M3	20	35	48	–	47	48

- LBCEM $>$ CEM, CEM-w
- LBVEM $>$ EM, EM-w
- LBVEM outperforms all the other variants

Outline

NMF: Nonnegative Matrix Factorization (Lee and Seung, 1999, 2001)

- Problem : $\text{argmin}_{\mathbf{U},\mathbf{V}\geq 0} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||^2$ where factor matrices, $\mathbf{U} \in \mathbb{R}_+^{n\times g}$ and $\mathbf{V} \in \mathbb{R}_+^{d\times m}$
- Other measures can be used as an error measures (for instance, KL divergence)
- The clustering problem is not the main objective of NMF



NMF: Nonnegative Matrix Factorization

- Each column of $\mathbf{X}$ is treated as a data point in n-dimensional space
- Each u_{ik} of $\mathbf{U}$ corresponds to the degree to which row i belongs to k*th* cluster
- Each column of U is associated with a prototype vector for the k*th* cluster
- Problems: Uniqueness, initialization

Expressions of U and V

A typical constrained optimization problem, and can be solved using the Lagrange multiplier method: $u_{ik} \leftarrow u_{ik} \frac{(\mathbf{XV})_{ik}}{(\mathbf{UV}^T\mathbf{V})_{ik}}$ and $v_{kj} \leftarrow v_{kj} \frac{(\mathbf{X}^T\mathbf{U})_{kj}}{(\mathbf{VU}^T\mathbf{U})_{kj}}$

Uniqueness

If **U** and **V** are solutions, then, **UD**, $\mathbf{VD}^{-1}$ will also form a solution for any positive diagonal matrix **D**. Generally to eliminate this uncertainty, in practice one will further require that the Euclidean length of each column vector in **U** or **V** is 1.
$u_{ik} \leftarrow \frac{u_{ik}}{\sqrt{\sum_i u_{ik}^2}}$ and $v_{kj} \leftarrow v_{kj}\sqrt{\sum_i u_{ik}^2}$

NMF towards clustering

1. Perform the NMF on **X** to obtain **U** and **V**
2. Normalize **U** and **V**
3. Use matrix **V** to determine the cluster label of each column. More precisely, examine each row of matrix **V**. Assign a column j to cluster k^* if $k^* = \arg\max_k v_{kj}$

Orthogonal NMF

$\text{argmin}_{U,V \geq 0} ||\mathbf{X} - \mathbf{UV}^T||^2$ where factor matrices, $U \in \mathbb{R}_+^{n\times g}$, $V \in \mathbb{R}_+^{d\times m}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$

NBVD: Nonegative Block Value Decomposition (Long et al. 2005)

- For co-clustering, it consists in seeking a 3-factor decomposition:

$$\underset{R,A,C\geq 0}{\text{argmin}} ||\mathbf{X} - \mathbf{RAC}^T||^2 \text{ where } R \in \mathbb{R}_+^{n\times g}, A \in \mathbb{R}_+^{g\times m}, C \in \mathbb{R}_+^{d\times m}$$

- R and C play the roles of row and column memberships
- A makes it possible to absorb the scales of R, C and $\mathbf{X}$

NMTF: Nonnegative Matrix Tri-Factorization (Ding et al., 2006), (Wang et al. 2011)

$$\underset{R,A,C\geq 0, R^TR=I_g, C^TC=I_m}{\text{argmin}} ||\mathbf{X} - RAC^T||^2$$

Double *k*means towards NMTF (Lazhar and Nadif, 2011)

- Convert the double kmeans criterion to an optimization problem under NMF
- R and C are cluster indicators

$$\underset{\mathbf{R},\geq 0,\mathbf{R}^T\mathbf{R}=I_g,\mathbf{C}^T\mathbf{C}=I_m}{\text{argmin}} ||\mathbf{X} - \mathbf{RR}^T\mathbf{XCC}^T||^2 \text{ with } \mathbf{R} = RD_r^{-0.5} \text{ and } \mathbf{C} = CD_c^{-0.5}$$

where $D_r^{-0.5} = Diag(\frac{1}{\sqrt{r_1}}, \ldots, \frac{1}{\sqrt{r_g}})$ and $D_c^{-0.5} = Diag(\frac{1}{\sqrt{c_1}}, \ldots, \frac{1}{\sqrt{c_m}})$

Dyadic Analysis

- Document clustering, term-document co-clustering
- Even if the objective is the clustering of documents, the co-clustering is beneficial
- TF-IDF $x_{ij} \leftarrow x_{ij} \log \frac{n}{n^j}$ where $n^j = \sum_{i|x_{ij} \neq 0}$

Datasets

- Classic30 is an extract of Classic3 which counts three classes denoted Medline, Cisi, Cranfield as their original database source. It consists of 30 random documents described by 1000 words
- Classic150 consists of 150 random documents described by 3652 words
- NG2 is a subset of 20-Newsgroup data NG20, it is composed by 500 documents concerning talk.politics.mideast and talk.politics.misc described by 2000 words

Results

dataset	performance measure	DNMF	ODNMF	ONM3F	ONMTF	NBVD
Classic30	Acc	96.67	100	100	100	96.67
	NMI	89.97	100	100	100	89.97
Classic150	Acc	98.66	98.66	99.33	98.66	98.66
	NMI	94.04	94.04	97.02	94.04	94.04
NG2	Acc	77.6	86.2	74.6	74.2	77.4
	NMI	19.03	43.47	18.27	16.03	23.31

Outline

Conclusion

Principal points

- Different approches exist
- Latent Block Models offer different co-clustering algorithms: LBCEM, LBVEM
- LBVBEM is more efficient in terms of clustering and estimation
- Document clustering: LBVEM, LBCEM on document-term matrix without any **normalization**
- Case of continuous data: Connections between LBCEM and NMTF

Works related to co-clustering

- KL divergence as an error measure: Connections between NMF and PLSA (Gaussier and Goutte, 2005), NMTF and *Aspect model* (Yoo and Choi, 2012).
- Visualization by GTM using LBM (Priam et al., 2013, 2014)
- Constraint co-clustering in Bioinformatics and document clustering