Chapter 3

Co-Clustering of Binary and Categorical Data

To justify and detail the search for simultaneous partitions of binary data, let us cite some ideas proposed by Lerman about the notion of polythetic cluster [LER 81]. He first recalled the notion of polythetic cluster: "A polythetic cluster G of a natural clustering refers to a subset B of attributes in such a way that: each element of the cluster has an important proportion of attributes from B; each attribute of B is present in an important proportion; an attribute is not necessarily shared by all elements of G". Lerman generalized this notion: "In the more general situation of a good clustering on E with a good clustering on A, each cluster E_k of the partition $(E_1, ..., E_g)$ corresponds to the union B of clusters of the partition $(A_1, ..., A_m)$. Conversely, each cluster A_ℓ corresponds to the union G of clusters of the partition $(E_1, ..., E_g)$". This situation may be represented by a binary table reorganized according to row and column partitions (Figure 3.1) where the shaded blocks correspond to regions with high density of ones and the unshaded blocks correspond to regions with high density of zeros.

Co-Clustering: Models, Algorithms and Applications Gérard Govaert and Mohamed Nadif

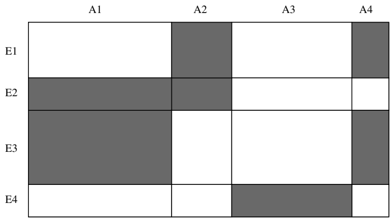

Figure 3.1. *Binary table reorganized according to row and column partitions*

3.1. Example and notation

This purpose can be shown precisely with a simple example. Let us consider data matrix x in Figure 3.2(a). The rows correspond to a set of 10 micro computers and the columns to 10 properties that these computers may (value 1) or may not have (value 0). With this example, if the partitions z and w are, respectively, $\{\{a, d, h\}, \{b, e, f, j\}, \{c, g, i\}\}$ and $\{\{1, 3, 5, 8, 10\}, \{2, 4, 6, 7, 9\}\}$, we obtain reorganized data matrix x (see Figure 3.2(b)) by reorganizing rows and columns according to these two partitions and the initial binary matrix can be summarized by the binary matrix a (see Figure 3.2(c)) when crossing the two partitions. In doing so, the 100 binary values of the initial data matrix x were summarized by the six binary values of the matrix a.

	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

(a) Binary data x

		1					2				
		1	3	5	8	10	2	4	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

(b) Reorganized data matrix x

	1	2
A	1	0
B	0	1
C	0	0

(c) Summary matrix a

Figure 3.2. *Microcomputer data*

To obtain a summary of initial data, we will see that the algorithms proposed are based, at each step, on intermediate matrices $\mathbf{x}^{\mathbf{z}}$, $\mathbf{x}^{\mathbf{w}}$ and $\mathbf{x}^{\mathbf{zw}}$ of reduced size defined in Table 3.1 whose values obtained with the previous example are given in Table 3.2.

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

Table 3.1. *Reduced matrices, sizes and definitions of* $\mathbf{x}^{\mathbf{z}}$, $\mathbf{x}^{\mathbf{w}}$ *and* $\mathbf{x}^{\mathbf{zw}}$

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

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

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

Table 3.2. *Intermediate data matrices* $\mathbf{x}^{\mathbf{z}}$, $\mathbf{x}^{\mathbf{w}}$ *and* $\mathbf{x}^{\mathbf{zw}}$ *obtained with the microcomputer data*

Moreover, we will also use the notation $x^{\mathbf{z}}_{k.} = \sum_i z_{ik} x_{i.}$ and $x^{\mathbf{w}}_{.\ell} = \sum_j w_{j\ell} x_{.j}$. In a similar way, we can define, with the fuzzy partitions $\widetilde{z}$ and $\widetilde{\mathrm{w}}$, the matrices $\mathrm{x}^{\widetilde{\mathbf{z}}}$, $\mathrm{x}^{\widetilde{\mathbf{w}}}$ and $\mathrm{x}^{\widetilde{\mathbf{z}}\widetilde{\mathbf{w}}}$.

To obtain these homogeneous blocks (or biclusters), most methods used in this context can be grouped into two families: metric methods and model-based clustering. Sections 3.2 and 3.3 will be, respectively, devoted to these two kinds of approaches.

3.2. Metric approach

To obtain such homogeneous blocks, we must minimize the number of times where the value x_{ij} is different from the value $a_{k\ell}$ associated with the clusters pair (k, ℓ) to which (i, j) belongs. This quantity represents the difference between the initial data and the summary data. Then, the problem consists of optimizing the following objective function

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

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

The minimization of this criterion leads to obtaining homogeneous blocks of 0 or 1 by reorganizing rows and columns according to the partitions $\mathbf{z}$ and $\mathbf{w}$. Hence, each block $k\ell$, defined by the elements x_{ij} for i belonging to the kth cluster and j to the ℓth cluster, is characterized by $a_{k\ell}$, which is the highest frequency value. To interpret the results, some statistics can be computed to evaluate the quality of the partition into blocks. For instance, we can define the proportion of block $k\ell$ values equal to $a_{k\ell}$ and therefore measure the degree of homogeneity of this block.

When we only deal with the research of one partition, we obtain the well-known maximal predictive classification criterion proposed by Gower [GOW 74]. Let us mention that we also find this criterion when we use a classification maximum likelihood criterion [CEL 92] or an entropy criterion [GYL 94] on the Bernoulli mixture model.

Besides, if we constrain the summary to have a diagonal structure (same number of clusters for the two partitions, 1 on the diagonal and 0 everywhere else), we find the criterion proposed by Garcia and Proth [GAR 86]. Let us note that introducing such a constraint in the following algorithm is very easy.

In algorithmic terms, to minimize this objective function and using the strategy described in section 2.1 of Chapter 2, Govaert [GOV 83, GOV 95] proposed the following CROBIN algorithm 3.1.

Algorithm 3.1 CROBIN (here $\lfloor x \rceil$ is the nearest integer function)

```
input: x, $g$, $m$
initialization: z, 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. $z_i = \arg\min_k \sum_\ell w_{j\ell}|x^{\mathbf{w}}_{i\ell} - w_{.\ell}a_{k\ell}|$
    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. $w_j = \arg\min_\ell \sum_k z_{ik}|x^{\mathbf{z}}_{kj} - z_{.k}a_{k\ell}|$
    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 z, w, a
```

REMARK 3.1.– Steps 1 and 2 can be seen as a dynamic cluster algorithm [DID 79] applied to the $n \times m$ matrix x^{w} with the L_1 distance and centroids

$$(w_{.1}a_{k1}, \ldots, w_{.m}a_{km})^t,$$

which minimizes the criterion $\mathrm{W}(\mathbf{z}, \mathbf{a}|\mathbf{w}) = \sum_{i,k} z_{ik} \sum_{\ell} |x^{\mathbf{w}}_{i\ell} - w_{.\ell}a_{k\ell}|$. Similarly, steps 3 and 4 can be seen as a dynamic cluster algorithm applied to the $d \times n$ matrix $\mathbf{x}^{\mathbf{z}}$ with the L_1 distance and centroids

$$(z_{.1}a_{1\ell}, \ldots, z_{.g}a_{g\ell})^t,$$

which minimizes the criterion $\mathrm{W}(\mathbf{w}, \mathbf{a}|\mathbf{z}) = \sum_{j,\ell} w_{j\ell} \sum_{k} |x^{\mathbf{z}}_{kj} - z_{.k}a_{k\ell}|$.

The structure of this program allows us to control the size of the matrices used in each step making it possible to use this program for large, and also to take into account sparse, data. This particular structure is an important feature in all of the programs described in this book.

3.3. Bernoulli latent block model and algorithms

3.3.1. *The model*

Using the latent block model (LBM) described in section 2.3 of Chapter 2, the binary data situation leads us to assume that for each block $k\ell$, the values x_{ij} are distributed according to the Bernoulli distribution $\mathcal{B}(\alpha_{k\ell})$ ($\alpha_{k\ell} \in \mathbb{R}$ and $0 < \alpha_{k\ell} < 1$) for which the pdf is

$$f(x_{ij}; \alpha_{k\ell}) = (\alpha_{k\ell})^{x_{ij}}(1 - \alpha_{k\ell})^{(1-x_{ij})}.$$

Denoting $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$ and $\boldsymbol{\alpha} = (\alpha_{11}, \ldots, \alpha_{gm})$, we obtain the following pdf according to Govaert and Nadif [GOV 03]

$$f(\mathrm{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \times \prod_{i,j,k,\ell} \Big((\alpha_{k\ell})^{x_{ij}} (1-\alpha_{k\ell})^{(1-x_{ij})} \Big)^{z_{ik}w_{j\ell}},$$

the complete-data log-likelihood becomes

$$\mathrm{L_C}(\mathbf{z},\mathbf{w},\boldsymbol{\theta}) = \sum_{i,k} z_{ik}\log\pi_k + \sum_{j,\ell} w_{j\ell}\log\rho_\ell + \sum_{i,k,j,\ell} z_{ik}w_{j\ell}x_{ij}\log\frac{\alpha_{k\ell}}{1-\alpha_{k\ell}} + \sum_{k,\ell} z_{k.}w_{.\ell}\log(1-\alpha_{k\ell}),$$

and then, the following fuzzy clustering criterion can be deduced

$$\mathrm{F_C}(\widetilde{\mathbf{z}},\widetilde{\mathbf{w}};\boldsymbol{\theta}) = \mathrm{L_C}(\widetilde{\mathbf{z}},\widetilde{\mathbf{w}},\boldsymbol{\theta}) + \mathrm{H}(\widetilde{\mathbf{z}}) + \mathrm{H}(\widetilde{\mathbf{w}}). \qquad [3.1]$$

3.3.2. *Model identifiability*

It is well-known that simple multivariate Bernoulli mixtures are not identifiable [GYL 94], regardless of the invariance to relabeling. Allman *et al.* [ALL 09] set a sufficient condition to their identifiability that cannot be applied to LBM. The following theorem on sufficient conditions ensuring the identifiability of the Bernoulli LBM can be established [KER 13].

THEOREM 3.1.– With $\boldsymbol{\alpha}$, the matrix of the Bernoulli coefficients, $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$, the row and column mixing proportions of the mixture, assume that $n \geq 2m-1$ and $d \geq 2g-1$ and

– (H_1) for all $1 \leq k \leq g$, $\pi_k > 0$ and the coordinates of vector $\boldsymbol{\tau} = \boldsymbol{\alpha}\boldsymbol{\rho}$ are distinct;

– (H_2) for all $1 \leq \ell \leq m$, $\rho_\ell > 0$ and the coordinates of vector $\boldsymbol{\sigma} = \boldsymbol{\pi}^t\boldsymbol{\alpha}$ are distinct (where $\boldsymbol{\pi}^t$ is the transpose of $\boldsymbol{\pi}$);

then, the binary LBM is identifiable.

H_1 and H_2 are not strongly restrictive since the set of vectors $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}$ violating them is of Lebesgue measure 0. Therefore, theorem 3.1 asserts the generic identifiability of the Bernoulli LBM, which is a "practical" identifiability, explaining why it works in the applications [CAR 00]. These assumptions could appear to be somewhat unnatural; however, (1) it is not surprising that the number of row labels g (respectively, column labels m) is constrained by the number of columns d (respectively, rows n). In case of a simple finite mixture of g different Bernoulli products with d components, the more clusters you define in the mixture, the more components for the multivariate Bernoulli you need in order to ensure the identifiability: see also the following condition $d > 2\lceil \log_2 g \rceil + 1$ of Allman *et al.* [ALL 09] for simple Bernoulli mixtures, where $\lceil . \rceil$ is the ceil function. (2) Assumptions H_1 and H_2 are the extensions of an assumption made to ensure the identifiability of the stochastic block model [CEL 12], where $n = d$ and $\mathbf{z} = \mathbf{w}$.

3.3.3. *Binary* LBVEM *and* LBCEM *algorithms*

Taking into account the definition of the pdf $f(x_{ij}; \alpha_{k\ell})$, the variational approximation of the expectation–maximization (EM) algorithm described in section 2.4.1 of Chapter 2 and maximizing $\mathrm{F_C}(\widetilde{\mathrm{z}}, \widetilde{\mathrm{w}}; \boldsymbol{\theta})$ can be completed by the computation of the block parameters $\alpha_{k\ell}$s. We obtain for all k, ℓ

$$\begin{aligned}\alpha_{k\ell} &= \arg\max_{\alpha_{k\ell}} \sum_{i,j} \widetilde{z}_{ik}\widetilde{\mathrm{w}}_{j\ell} \log f(x_{ij}, \boldsymbol{\alpha}_{k\ell}) \\ &= \arg\max_{\alpha_{k\ell}} \sum_{i,j} \widetilde{z}_{ik}\widetilde{\mathrm{w}}_{j\ell} \left(x_{ij} \log \alpha_{k\ell} + (1 - x_{ij}) \log(1 - \alpha_{k\ell})\right) \\ &= \arg\max_{\alpha_{k\ell}} \left(x_{k\ell}^{\widetilde{\mathbf{z}}\widetilde{\mathbf{w}}} \log \alpha_{k\ell} + (\widetilde{z}_{.k}\widetilde{\mathrm{w}}_{.\ell} - x_{k\ell}^{\widetilde{\mathbf{z}}\widetilde{\mathbf{w}}}) \log(1 - \alpha_{k\ell})\right),\end{aligned}$$

which yields $\alpha_{k\ell} = \frac{x_{k\ell}^{\widetilde{\mathbf{z}}\widetilde{\mathbf{w}}}}{\widetilde{z}_{.k}\widetilde{\mathrm{w}}_{.\ell}}$ and finally algorithm 3.2.

Algorithm 3.2 Bernoulli LBVEM

```
input: x, $g$, $m$
initialization: $\widetilde{\mathbf{z}}$, $\widetilde{\mathbf{w}}$, $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\rho_\ell = \frac{\widetilde{\mathrm{w}}_{.\ell}}{d}$, $\alpha_{k\ell} = \frac{x_{k\ell}^{\widetilde{\mathbf{z}}\widetilde{\mathbf{w}}}}{\widetilde{z}_{.k}\widetilde{\mathrm{w}}_{.\ell}}$
repeat
  $x_{i\ell}^{\widetilde{\mathbf{w}}} = \sum_j \widetilde{\mathrm{w}}_{j\ell} x_{ij}$
  repeat
    step 1. $\widetilde{z}_{ik} \propto \pi_k \exp \sum_\ell \left(x_{i\ell}^{\widetilde{\mathbf{w}}} \ln \frac{\alpha_{k\ell}}{1-\alpha_{k\ell}} + \widetilde{\mathrm{w}}_{.\ell} \ln(1 - \alpha_{k\ell})\right)$
    step 2. $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\alpha_{k\ell} = \frac{\sum_i \widetilde{z}_{ik} x_{i\ell}^{\widetilde{\mathbf{w}}}}{\widetilde{z}_{.k}\widetilde{\mathrm{w}}_{.\ell}}$
  until convergence
  $x_{kj}^{\widetilde{\mathbf{z}}} = \sum_i \widetilde{z}_{ik} x_{ij}$
  repeat
    step 3. $\widetilde{\mathrm{w}}_{j\ell} \propto \rho_\ell \exp \sum_k \left(x_{jk}^{\widetilde{\mathbf{z}}} \ln \frac{\alpha_{k\ell}}{1-\alpha_{k\ell}} + \widetilde{z}_{.k} \ln(1 - \alpha_{k\ell})\right)$
    step 4. $\rho_\ell = \frac{\widetilde{\mathrm{w}}_{.\ell}}{d}$, $\alpha_{k\ell} = \frac{\sum_j \widetilde{\mathrm{w}}_{j\ell} x_{kj}^{\widetilde{\mathbf{z}}}}{\widetilde{z}_{.k}\widetilde{\mathrm{w}}_{.\ell}}$
  until convergence
until convergence
return $\boldsymbol{\pi}$, $\boldsymbol{\rho}$, $\boldsymbol{\alpha}$
```

Knowing that we have the following two relations shown in [GOV 08]

$$\mathrm{F_C}(\widetilde{\mathbf{z}}, \widetilde{\mathbf{w}}, \boldsymbol{\theta}) = \mathrm{F_C}(\widetilde{\mathbf{z}}, \boldsymbol{\alpha}, \boldsymbol{\pi} | \widetilde{\mathbf{w}}, \boldsymbol{\rho}) + \sum_{j,\ell} \widetilde{\mathrm{w}}_{j\ell} \log \rho_\ell + \mathrm{H}(\widetilde{\mathbf{w}})$$

and

$$\mathrm{F_C}(\widetilde{\mathbf{z}}, \widetilde{\mathrm{w}}, \boldsymbol{\theta}) = \mathrm{F_C}(\widetilde{\mathrm{w}}, \boldsymbol{\alpha}, \boldsymbol{\rho}|\widetilde{\mathbf{z}}, \boldsymbol{\pi}) + \sum_{i,k} \widetilde{z}_{j\ell} \log \pi_k + \mathrm{H}(\widetilde{\mathbf{z}}),$$

where $\mathrm{F_C}(\widetilde{\mathbf{z}}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\widetilde{\mathrm{w}}, \boldsymbol{\rho})$ and $\mathrm{F_C}(\widetilde{\mathrm{w}}, \boldsymbol{\alpha}, \boldsymbol{\rho}|\widetilde{\mathbf{z}}, \boldsymbol{\pi})$ correspond, respectively, to Hathaway's criteria defined in section 1.4.5 of Chapter 1 for the row clustering of the data matrix x^{w} and for the column clustering of the data matrix x^{z}, steps 1 and 2 can be seen as steps of EM applied on the $n \times m$ matrix $\mathrm{x}^{\widetilde{\mathrm{w}}}$ and maximizing $\mathrm{F_C}(\widetilde{\mathbf{z}}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\widetilde{\mathrm{w}}, \boldsymbol{\rho})$. Similarly, steps 3 and 4 can be seen as steps of EM applied on the $g \times d$ matrix $\mathrm{x}^{\widetilde{\mathrm{z}}}$ and maximizing $\mathrm{F_C}(\widetilde{\mathrm{w}}, \boldsymbol{\alpha}, \boldsymbol{\rho}|\widetilde{\mathbf{z}}, \boldsymbol{\pi})$.

In a similar way, the maximization of the complete-data log-likelihood $\mathrm{L_C}$ can be performed by the alternated maximization of conditional complete-data log-likelihoods $\mathrm{L_C}(\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\pi}|\mathrm{w}, \boldsymbol{\rho})$ and $\mathrm{L_C}(\mathrm{w}, \boldsymbol{\alpha}, \boldsymbol{\rho}|\mathrm{z}, \boldsymbol{\pi})$ which correspond, respectively, to the complete-data log-likelihoods for the row clustering of the data matrix x^{w} and for the column clustering of the data matrix x^{z}. Then, the principal steps of the algorithm used for this task are reported in algorithm 3.3.

Different comparisons were performed by Govaert and Nadif [GOV 05, GOV 08] confirming that LBVEM outperforms LBCEM in terms of estimation. In Table 3.3, we illustrate this observation by reporting a comparison between the two algorithms LBCEM and LBVEM on 200 × 120 data matrix. The quality of estimate measured by the deviance between the estimated parameter $\boldsymbol{\theta}$ and the true parameter $\boldsymbol{\theta}^0$, whose direct calculation is not possible, is approximated by $D^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)$ (D^2 denotes the Euclidean distance) expressed by the following formula

$$D^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0) = \sum_{k,\ell} (\alpha_{k\ell} - \alpha^0_{k\ell})^2 + \sum_k (\pi_k - \pi^0_k)^2 + \sum_\ell (\rho_\ell - \rho^0_\ell)^2.$$

We observe that the estimated parameters are closer to the true values with LBVEM than with LBCEM. The value of $D^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)$ is equal to 0.0252 for LBVEM and is equal to 0.0824 for LBCEM. In terms of co-clustering, in [GOV 05, GOV 08], the authors showed that LBVEM outperforms LBCEM even when the clusters are well separated. However, the latter is more scalable and it converges within 30 iterations.

Algorithm 3.3 Bernoulli LBCEM

```
input: x, $g$, $m$
initialization: z, w, $\pi_k = \frac{z_{.k}}{n}$, $\rho_\ell = \frac{w_{.\ell}}{d}$, $\alpha_{k\ell} = \frac{x^{\tilde{\mathbf{z}}\tilde{\mathbf{w}}}_{k\ell}}{\tilde{z}_{.k}\tilde{\mathrm{w}}_{.\ell}}$
repeat
    $x^{\mathbf{w}}_{i\ell} = \sum_j w_{j\ell} x_{ij}$
    repeat
        step 1. $z_i = \arg\max_k \{\sum_\ell \{x^{\mathbf{w}}_{i\ell} \ln \frac{\alpha_{k\ell}}{1-\alpha_{k\ell}} + w_{.\ell} \ln(1-\alpha_{k\ell})\} + \ln \pi_k\}$
        step 2. $\pi_k = \frac{z_{.k}}{n}$, $\alpha_{k\ell} = \frac{\sum_i z_{ik} x^{\mathbf{w}}_{i\ell}}{z_{.k} w_{.\ell}}$
    until convergence
    $x^{\mathbf{z}}_{kj} = \sum_i z_{ik} x_{ij}$
    repeat
        step 3. $w_j = \arg\max_\ell \{\sum_k \{x^{\mathbf{z}}_{kj} \ln \frac{\alpha_{k\ell}}{1-\alpha_{k\ell}} + z_{.k} \ln(1-\alpha_{k\ell})\} + \ln \rho_\ell\}$
        step 4. $\rho_\ell = \frac{w_{.\ell}}{d}$, $\alpha_{k\ell} = \frac{\sum_j w_{j\ell} x^{\mathbf{z}}_{kj}}{z_{.k} w_{.\ell}}$
    until convergence
until convergence
return z, w, $\boldsymbol{\pi}$, $\boldsymbol{\rho}$, $\boldsymbol{\alpha}$
```

Furthermore, the complexity of these algorithms depends on the average time of execution of an iteration. The number of iterations, depends on the degree of mixture, the size of the data and the complexity of the model. For the average time of an iteration, we checked on many examples that this one was independent of the degree of the mixture and that it depends, in a roughly linear way, on the sizes n and d of the data and the number g and m of clusters.

$\boldsymbol{\theta}$	True values ($\boldsymbol{\theta}^0$)	Parameter estimations values by LBVEM	Parameter estimations by LBCEM
π_1	0.2	0.1979	0.1900
π_2	0.3	0.3140	0.3400
π_3	0.5	0.4881	0.4700
ρ_1	0.3	0.2929	0.2583
ρ_2	0.7	0.7071	0.7417
$\boldsymbol{\alpha}$	$\begin{pmatrix} 0.60 & 0.40 \\ 0.40 & 0.60 \\ 0.60 & 0.65 \end{pmatrix}$	$\begin{pmatrix} 0.6067 & 0.4026 \\ 0.4089 & 0.6041 \\ 0.5989 & 0.6565 \end{pmatrix}$	$\begin{pmatrix} 0.6188 & 0.4063 \\ 0.3861 & 0.6000 \\ 0.6095 & 0.6559 \end{pmatrix}$
$D^2(\boldsymbol{\theta}, \boldsymbol{\theta}^0)$	0	**0.0252**	0.0824

Table 3.3. *True values and their estimates by* LBVEM *and* LBCEM *for* $n \times d = 200 \times 120$ *data matrix*

3.4. Parsimonious Bernoulli LBMs

As in section 1.7.2 of Chapter 1, a parsimonious model can be defined by breaking down the block parameter into a "center" parameter and a "dispersion" parameter and by imposing constraints on the dispersion parameter. More precisely, each parameter $\alpha_{k\ell}$ is replaced by $a_{k\ell} \in \{0, 1\}$ and $\varepsilon_{k\ell} \in [0, 1/2]$ with

$$\begin{cases} a_{k\ell} = 1 \text{ and } \varepsilon_{k\ell} = 1 - \alpha_{k\ell} \text{ if } \alpha_{k\ell} \in [1/2, 1] \\ a_{k\ell} = 0 \text{ and } \varepsilon_{k\ell} = \alpha_{k\ell} \text{ if } \alpha_{k\ell} \in [0, 1/2[. \end{cases}$$

Thus, the Bernoulli pdf can be written as

$$f(x_{ij}; (a_{k\ell}, \varepsilon_{k\ell})) = (\varepsilon_{k\ell})^{|x_{ij} - a_{k\ell}|}(1 - \varepsilon_{k\ell})^{1 - |x_{ij} - a_{k\ell}|},$$

where

– $a_{k\ell} \in \{0, 1\}$, which is the most frequent binary value of the block, represents the center of the block, and;

– $\varepsilon_{k\ell} \in]0, 1/2[^d$, which is the probability of any particular variable having a value different from that of the center of the block, represents the dispersion of the component.

Starting from this formulation, we arrive at parsimonious situations by stipulating certain constraints: so, the $[\varepsilon]$ model is defined by stipulating that the dispersion should not depend either on the row cluster or on the column cluster, the $[\varepsilon_k]$ model by stipulating that it should depend only on the row cluster and the $[\varepsilon^j]$ model by stipulating that it should depend only on the column cluster.

This parsimonious parameterization allows us to study the relationship between the CROBIN algorithm described in section 3.2 and the LBM. Indeed, in the simplest case, in the $[\varepsilon]$ model, given identical proportions ($\pi_k = 1/g$ and $\rho_\ell = 1/m$), the complete-data log-likelihood is equal to

$$\log \frac{\varepsilon}{1-\varepsilon} \mathrm{W}(\mathbf{z}, \mathbf{w}, \mathbf{a}) + nd \log(1-\varepsilon) - n \log g - d \log m,$$

where $\mathrm{W}(\mathbf{z}, \mathbf{w}, \mathbf{a}) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|$ is the criterion optimized by the algorithm CROBIN described in section 3.2 so that maximizing $\mathrm{L_C}(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta})$ is equivalent to minimizing $\mathrm{W}(\mathbf{z}, \mathbf{w}, \mathrm{a})$. In addition, it clearly appears that the steps maximizing alternatively $\mathrm{L_C}(\mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\pi} | \mathrm{w}, \boldsymbol{\rho})$ and $\mathrm{L_C}(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\rho} | \mathbf{z}, \boldsymbol{\pi})$ correspond to steps of CROBIN minimizing $\mathrm{W}(\mathbf{z}, \mathrm{a} | \mathrm{w})$ and $\mathrm{W}(\mathbf{w}, \mathbf{a} | \mathbf{z})$.

3.5. Categorical data

Developments presented in the first part of this chapter can be extended to a population of n observations described with d categorical variables of the same nature with r levels being available. Saying that the categorical variables are of the same nature means that it is possible to code them in the same (and natural) way. This assumption is needed to ensure that decomposing the data set in a block structure makes sense. Let $\mathrm{x} = (x_{ij}, i = 1, \ldots, n; j = 1, \ldots, d)$ be the data matrix defined on $I \times J$ where $x_{ij} = h$, $1 \leq h \leq r$, r is the number of levels of the J variables. In the following, an alternative representation of the data set with binary indicator vectors will often be used for mathematical convenience: $\mathrm{x}_{ij} = (x_{ij}^h, h = 1, \ldots, r)$ with $x_{ij}^h = 1$ when $x_{ij} = h$ and $x_{ij}^h = 0$.

To extend results on binary data to categorical data, a categorical LBM is considered: the conditional distribution of the outcome x_{ij} knowing the labels z_{ik} and $w_{j\ell}$ is a categorical distribution $\mathcal{C}(\alpha_{k\ell})$, of parameter $\alpha_{k\ell} = (a_{k\ell}^h)_{h=1,\ldots,r}$, where $a_{k\ell}^h \in (0;1)$ and $\sum_h a_{k\ell}^h = 1$. Using the binary indicator vector x_{ij}, the pdf per block is

$$f(\mathrm{x}_{ij}; \alpha_{k\ell}) = \prod_h (a_{k\ell}^h)^{x_{ij}^h}$$

and the mixture pdf is

$$f(\mathrm{x}; \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}) = \sum_{\mathbf{z},\mathbf{w}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell,h} (a_{k\ell}^h)^{z_{ik} w_{j\ell} x_{ij}^h}.$$

The $g + m + (r-1)gm - 2$ parameters to be estimated are $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$.

For the identifiability, theorem 3.1 is easily extended to the categorical case, where the assumptions are defined on vectors $\boldsymbol{\tau}^h = \boldsymbol{\alpha}^h \boldsymbol{\rho}$ and $\boldsymbol{\sigma}^h = \boldsymbol{\pi}^t \boldsymbol{\alpha}^h$, with $h = 1, \ldots, r$ and $\boldsymbol{\alpha}^h = (\alpha_{k\ell}^h)_{k=1,\ldots,g;\ell=1,\ldots,m}$.

The algorithms 3.2 and 3.3 can be easily extended by replacing the computation of $\alpha_{k\ell}$ by the following computations of $a_{k\ell}^h$

$$a_{k\ell}^h = \frac{\sum_{i,j} \widetilde{z}_{ik} \widetilde{\mathrm{w}}_{j\ell} x_{ij}^h}{\widetilde{z}_{.k} \widetilde{\mathrm{w}}_{.\ell}}$$

in the variational EM (VEM) situation and

$$a_{k\ell}^h = \frac{\sum_{i,j} z_{ik} w_{j\ell} x_{ij}^h}{z_{.k} w_{.\ell}}$$

in the classification EM (CEM) situation.

3.6. Bayesian inference

Bayesian inference in statistics can be regarded as a well-grounded tool for regularizing maximum likelihood estimates in a poorly posed setting. In the LBM setting, Bayesian inference could be thought of as useful for avoiding spurious solutions and thus attenuating the "empty cluster" problem. In particular, for the categorical LBM, it is possible to consider non-informative prior distribution for the model parameters (see Figure 3.3)

$$\boldsymbol{\pi} \sim \mathcal{D}(a, \dots, a), \; \boldsymbol{\rho} \sim \mathcal{D}(a, \dots, a), \; \boldsymbol{\alpha} \sim \prod_{k,\ell} \mathcal{D}(b, \dots, b),$$

$\mathcal{D}(v, \dots, v)$ denoting a Dirichlet distribution with parameter v. Obviously, since Dirichlet prior distributions are conjugate priors for the multinomial distribution, full conditional posterior distributions of the LBM parameters are closed form and Gibbs sampling is easy to implement.

Using Bayesian inference from a regularization perspective, the model parameter may be estimated by maximizing the posterior probability $p(\boldsymbol{\theta}|\mathbf{x})$ that leads to the so-called maximum *a posteriori* (MAP) estimate

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}).$$

The Bayes formula

$$\log p(\boldsymbol{\theta}|\mathbf{x}) = \log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{x})$$

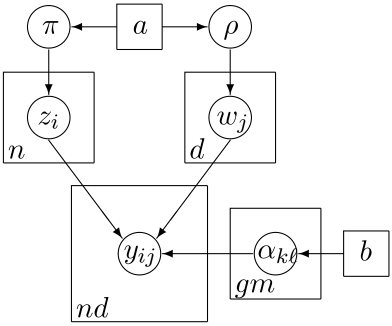

Figure 3.3. *Bayesian latent block model*

allows us to straightforwardly define an EM algorithm for the computation of the MAP estimate:

– The E step relies on the computation of the conditional expectation of the complete log-likelihood $Q(\boldsymbol{\theta},\boldsymbol{\theta}')$ as for the maximum likelihood estimator

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}') = \mathbb{E}\big(\log p(\mathrm{x},\mathbf{z},\mathrm{w}|\boldsymbol{\theta}')|\mathrm{x},\boldsymbol{\theta}'\big).$$

– The M step differs in that the objective function for the maximization process is equal to the $Q(\boldsymbol{\theta},\boldsymbol{\theta}')$ function augmented by the logarithm of the prior probability

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \big(Q(\boldsymbol{\theta},\boldsymbol{\theta}') + \log p(\boldsymbol{\theta})\big).$$

This M step forces an increase in the log posterior function $p(\boldsymbol{\theta}|\mathrm{x})$ [MCL 07, Chapter 5, p. 231]. For LBM, the Bayesian approach combined with the variational approximation leads to the V-Bayes algorithm, with the following M step:

M step (V-Bayes algorithm: estimation of the posterior mode)

$$\pi_k = \frac{a-1+\widetilde{z}_{.k}}{n+g(a-1)}, \quad \rho_\ell = \frac{a-1+\widetilde{\mathrm{w}}_{.\ell}}{d+m(a-1)},$$

$$a^h_{k\ell} = \frac{b-1+\sum_{i,j}\widetilde{z}_{ik}\widetilde{\mathrm{w}}_{j\ell}\, x^h_{ij}}{r(b-1)+\widetilde{z}_{.k}\widetilde{\mathrm{w}}_{.\ell}}.$$

The hyperparameters a and b are acting as regularization parameters. From this perspective, the choices of a and b are important. It clearly appears from the updating equations of this M step that V-Bayes is useless for uniform prior ($a = b = 1$) and worse than useless for Jeffreys prior ($a = b = 1/2$). Frühwirth-Schnatter [FRÜ 11] showed the great influence of a for Bayesian inference in the Gaussian mixture context. Based on an asymptotic analysis by Rousseau and Mergensen [ROU 11] and on a thorough finite sample analysis, they advocated taking $a = 4$ for moderate dimensions ($g < 8$) and $a = 16$ for larger dimensions to avoid empty clusters.

However, as VEM, a V-Bayes algorithm could be expected to be highly dependent on its initial values. Thus, it could be of significance to initiate V-Bayes with the solution derived from the following Gibbs sampler:

1) Simulation of z according to $p(\mathrm{z}|\mathrm{x},\mathrm{w};\boldsymbol{\theta})$ as in SEM-Gibbs.

2) Simulation of w according to $p(\mathrm{w}|\mathrm{x},\mathrm{z};\boldsymbol{\theta})$ as in SEM-Gibbs.

3) Simulation of $\boldsymbol{\pi}$ according to $\mathcal{D}(a+z_{.1},\ldots,a+z_{.g})$.

4) Simulation of $\boldsymbol{\rho}$ according to $\mathcal{D}(a+w_{.1},\ldots,a+w_{.m})$.

5) Simulation of $\boldsymbol{\alpha}_{k\ell}$ according to a $\mathcal{D}(b+x^1_{k\ell},\ldots,b+x^r_{k\ell})$ for $k=1,\ldots,g;\ell=1,\ldots,m$ and with $x^h_{k\ell}=\sum_{i,j} z_{ik}w_{j\ell}x^h_{ij}$.

As Gibbs sampling explores the whole distribution, it is subject to label switching. This problem can be sensibly solved for the categorical LBM by using the identifiability conditions of theorem 3.1: these conditions define a natural order of row and column labels except on a set of parameters of measure zero. Hence, column (respectively, row) labels are reordered according to the ascending values of $\boldsymbol{\tau}^h = \boldsymbol{\alpha}^h\boldsymbol{\rho}$

(respectively, $\boldsymbol{\sigma}^h = \boldsymbol{\pi}^t\boldsymbol{\alpha}^h$) coordinates for a given h, after each steps 3 and 5 (respectively, 4 and 5).

3.7. Model selection

Choosing relevant numbers of clusters in an LBM is obviously of crucial importance. In the previous chapter, we mentioned the reasons that pose difficulties to dealing with the selection problem. However, it is possible to compute the exact integrated completed log-likelihood (ICL) of the categorical LBM and an asymptotic approximation of the integrated likelihood could be derived from this ICL criterion.

3.7.1. *The integrated completed log-likelihood (ICL)*

The ICL [BIE 00] is the logarithm of the integrated completed likelihood, which takes the following form

$$p(\mathbf{x}, \mathbf{z}, \mathbf{w} \mid g, m) = \int_{\Theta_{g,m}} p(\mathbf{x}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\theta}_{g,m})p(\boldsymbol{\theta}_{g,m})d\boldsymbol{\theta}_{g,m}$$

with

$$p(\mathbf{x}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\theta}_{g,m}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} p(\mathbf{z})p(\mathbf{w})f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}_{g,m}),$$

where $f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}_{g,m}) = \prod_{i,j,k,\ell} f(x_{ij}; \alpha_{k\ell})^{z_{ik}w_{j\ell}}$, $\Theta_{g,m}$ being the parameter space of the model with $g \times m$ blocks and $p(\boldsymbol{\theta}_{g,m})$ a non-informative or a weakly informative prior distribution on $\boldsymbol{\theta}$. By taking into account the missing data, ICL is focusing on the clustering view of the model. For this very reason, ICL could be expected to select a stable model allowing us to partition the data with the greatest evidence [BIE 10].

For the categorical LBM, proper non-informative priors are available and ICL can be computed without requiring asymptotic approximations. Using the conjugacy properties of the prior Dirichlet distributions and the conditional independence of the x_{ij} knowing the latent vectors z and w and the LBM parameters, it can be shown (see [KER 13]) that

$$\begin{aligned}\mathrm{ICL}(g,m) = {} & \log\Gamma(ga) + \log\Gamma(ma) - (m+g)\log\Gamma(a)\\ & + mg(\log\Gamma(rb) - r\log\Gamma(b))\\ & - \log\Gamma(n+ga) - \log\Gamma(d+ma)\\ & + \sum_k \log\Gamma(z_{.k}+a) + \sum_\ell \log\Gamma(w_{.\ell}+a)\\ & + \sum_{k,\ell}\Big[\big(\sum_h \log\Gamma(N^h_{k\ell}+b)\big) - \log\Gamma(z_{.k}w_{.\ell}+rb)\Big].\end{aligned}$$

In practice, the missing labels $\mathbf{z}, \mathbf{w}$ are to be chosen. Following [BIE 00], they are replaced by

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

$\hat{\boldsymbol{\theta}}$ being the maximum likelihood estimate of the LBM parameter derived from the SEM-Gibbs algorithm followed by the VEM algorithm as described in section 3.3.3.

3.7.2. *Penalized information criteria*

It could be of significance to analyze the behavior of standard information criteria as Bayesian information criterion (BIC) and $\mathrm{ICL_{BIC}}$ derived from asymptotic approximations. Using the Stirling formula

$$\Gamma(z) \underset{+\infty}{\sim} z^{z-1/2}e^{-z}\sqrt{2\pi},$$

the asymptotic approximation of ICL as n and d tend to infinity is

$$\mathrm{ICL}_{\mathrm{BIC}}(g,m) = \max_{\boldsymbol{\theta}} \log p(\mathbf{x},\widehat{\mathbf{z}},\widehat{\mathbf{w}};\boldsymbol{\theta}) - \frac{g-1}{2}\log n - \frac{m-1}{2}\log d - \frac{gm(r-1)}{2}\log(nd).$$

This result is a generalization of the categorical case of the result proved in [KER 13] for the binary LBM. Moreover, BIC can be conjectured to have the following expression

$$\mathrm{BIC}(g,m) = \max_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) - \frac{g-1}{2}\log n - \frac{m-1}{2}\log d - \frac{gm(r-1)}{2}\log(nd).$$

As the maximized likelihood is unavailable for the LBM, it could be approximated by the maximized fuzzy function $\mathrm{F_C}$. But there is no reason to replace any criterion when available with its asymptotic approximation. Thus, ICL can be preferred to $\mathrm{ICL_{BIC}}$ or BIC.

3.8. Illustrative experiments

3.8.1. *Townships*

Hereafter, we illustrate the co-clustering aim by an example which consists of the characteristics (rows) of 16 townships {A,...,P} (columns), each cell indicating the presence 1 or absence 0 of a characteristic of a township (Table 3.4). The set of rows consists of the following characteristics: High School (hsco), Agricult Coop (agri), Railway Station (rail), One Room School (osco), Veterinary (vete), No Doctor (nodo), No Water Supply (nwat), Police Station (poli) and Land Reallocation (land). This example has been used by Niermann [NIE 05] for data ordering tasks, and the author aimed to reveal a diagonal of homogeneous blocks by using a genetic algorithm.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	$x_{i.}$
hsco	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	2
agri	0	1	1	0	0	0	1	0	0	0	0	1	0	0	1	0	5
rail	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	2
osco	1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	1	8
vete	0	1	1	1	0	0	1	0	0	0	0	1	0	0	1	0	6
nodo	1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	1	8
nwat	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	4
poli	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	2
land	0	1	1	1	0	0	1	0	0	0	0	1	0	0	1	0	6
$x_{.j}$	2	3	3	2	2	2	3	3	3	3	3	3	3	3	3	2	43

Table 3.4. *Townships data matrix*

After having applied LBVEM and LBCEM with different initializations, we retain the best result corresponding to the higher log-likelihood for LBVEM or complete-data log-likelihood for LBCEM. Both algorithms give the same results in terms of co-clustering. The parameters, the partitions, the summary matrix and the reorganized data matrix obtained with LBCEM and the model $[\pi, \rho, \varepsilon_{k\ell}]$ are reported, respectively, in Tables 3.5–3.8.

$$a_{k\ell} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \varepsilon_{k\ell} = \begin{pmatrix} 0.00 & 0.06 & 0.00 \\ 0.00 & 0.00 & 0.00 \\ 0.17 & 0.00 & 0.00 \end{pmatrix} \quad \alpha_{k\ell} = \begin{pmatrix} 0.00 & 0.94 & 0.00 \\ 0.00 & 0.00 & 1.00 \\ 0.83 & 0.00 & 0.00 \end{pmatrix}$$

Table 3.5. *Parameters obtained by* LBVEM

1: agri,vete,land	1: A, E, F, I, J, M, N, P
2: hsco, rail, poli	2: B, C, D, G, L, O
3: osco, nodo, nwat	3: H, K

Table 3.6. *Row (left) and column (right) partitions obtained by* LBVEM

$$\mathbf{x}^{\mathbf{zw}} = \begin{pmatrix} 0 & 17 & 0 \\ 0 & 0 & 6 \\ 20 & 0 & 0 \end{pmatrix}$$

Table 3.7. *Summary matrix obtained by* LBVEM

	AEF IJMNP	BCDGLO	HK
agri		1 1 1 1 1	
vete		1 1 1 1 1 1	
land		1 1 1 1 1 1	
hsco			1 1
rail			1 1
poli			1 1
osco	1 1111 1 1 1		
nodo	1 1111 1 1 1		
nwat	11 1 1		

Table 3.8. *Reorganized data matrix obtained by* LBVEM

It clearly appears that we can characterize each cluster of townships by a cluster of characteristics: {H, K} by {High School, Railway Station, Police Station}, {B, C, D, G, L, O} by {Agricult Coop, Veterinary, Land Reallocation} and {M, N, J, I, A, P, F, E} by {One Room School, No Doctor, No Water Supply}.

To quickly view the results, a graphical representation of the results can be used (see Figure 3.4).

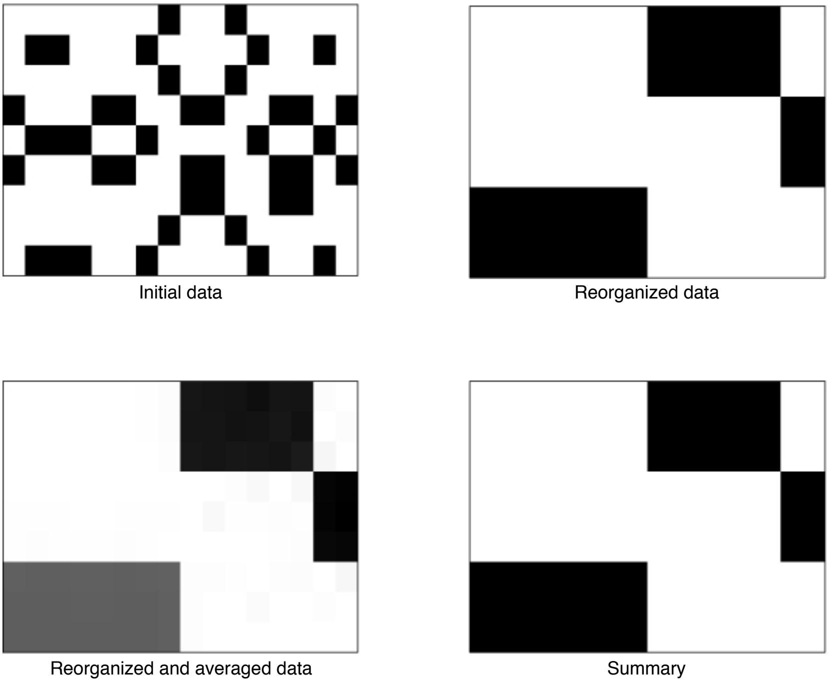

Figure 3.4. *Initial and reorganized data according to row and column partitions (top), reorganized averaged data expressing the degree of homogeneity and summary (bottom)*

3.8.2. *Mero*

The Mero data set consists of 59 Merovingian belt buckles {bu1,...,bu59} from north-eastern France, ranging from the late sixth Century and the early eighth Century on which the presence or absence of 26 technical criteria of manufacturing, shape and decor {ch1,...,ch26} was observed [LER 80]. The parameters, the row partition, the column partition and the summary matrix obtained with the algorithm LBCEM and the model $[\pi, \rho, \varepsilon_{k\ell}]$ are, respectively, reported in Tables 3.9–3.12 and in Figure 3.5.

$$\alpha_{k\ell} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \quad \varepsilon_{k\ell} = \begin{pmatrix} 0.03 & 0.09 & 0.00 & 0.00 & 0.02 \\ 0.00 & 0.22 & 0.16 & 0.18 & 0.03 \\ 0.00 & 0.35 & 0.06 & 0.03 & 0.31 \\ 0.15 & 0.02 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.07 & 0.07 & 0.00 & 0.29 \end{pmatrix}$$

$$\alpha_{k\ell} = \begin{pmatrix} 0.03 & 0.91 & 0.00 & 1.00 & 0.02 \\ 0.00 & 0.78 & 0.16 & 0.82 & 0.03 \\ 0.00 & 0.35 & 0.06 & 0.97 & 0.31 \\ 0.85 & 0.02 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.07 & 0.93 & 0.00 & 0.29 \end{pmatrix}$$

Table 3.9. *Parameters obtained by* LBVEM *on the Mero data*

1: bu3,bu9,bu14,bu24,bu25,bu26,bu30,bu32,bu33,bu36,bu46,bu49,bu53

2: bu4,bu20,bu21,bu22,bu31

3: bu7,bu11,bu17,bu27,bu28,bu34,bu35,bu48,bu50,bu51,bu52, bu59

4: bu5,bu15,bu23,bu47,bu54,bu55,bu56,bu57,bu58

5: bu1,bu2,bu6,bu8,bu10,bu12,bu13,bu16,bu18,bu19,bu29,bu37,bu38, bu39,bu40,bu41,bu42,bu43,bu44,bu45

Table 3.10. *Row partition obtained by* LBVEM *on the Mero data*

1: ch1,ch6,ch12

2: ch5,ch13,ch17,ch22,ch23

3: ch4,ch8,ch10,ch20,ch21,ch26

4: ch3,ch7,ch9

5: ch2,ch11,ch14,ch15,ch16,ch18,ch19,ch24,ch25

Table 3.11. *Column partition obtained by* LBVEM *on the Mero data*

$$\mathbf{x}^{\mathbf{zw}} = \begin{pmatrix} 1 & 58 & 0 & 39 & 2 \\ 0 & 20 & 7 & 11 & 1 \\ 0 & 22 & 4 & 35 & 33 \\ 23 & 1 & 0 & 27 & 0 \\ 0 & 7 & 112 & 0 & 52 \end{pmatrix}$$

Table 3.12. *Summary matrix obtained by* LBVEM *on the Mero data*

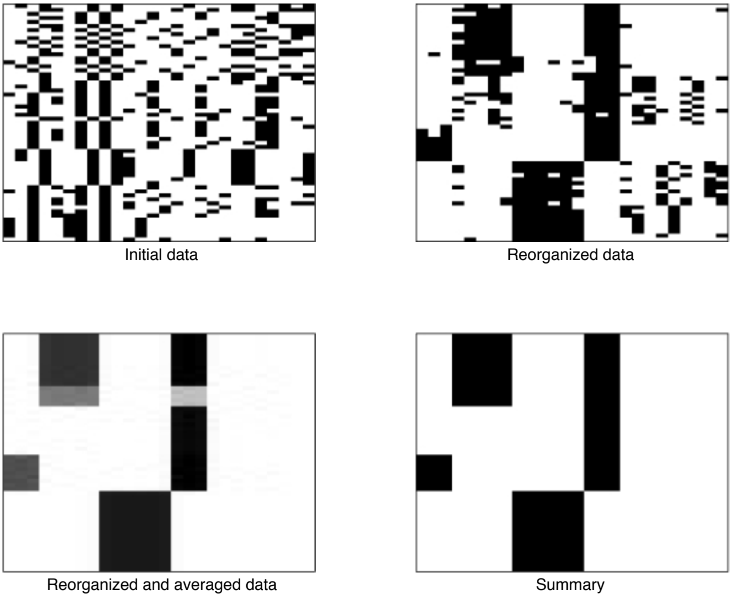

Figure 3.5. *Initial and reorganized data according to row and column partitions (top), reorganized averaged data expressing the degree of homogeneity and summary (bottom)*

First, we see the initial data and their reorganization according to the row and column clusters highlighting a structure into homogeneous blocks. Second, we note that this homogeneity can be evaluated with the reorganized averaged data and their summary. Furthermore, we complete the co-clustering results by the projections of row and column clusters obtained with the correspondence analysis (CA) [BEN 73b] in Figures 3.6 and 3.7. Observing the row and column clusters in both planes, we can interpret the associations between the clusters; however, we only have part of the information. Note that the co-clustering directly offers this interpretation.

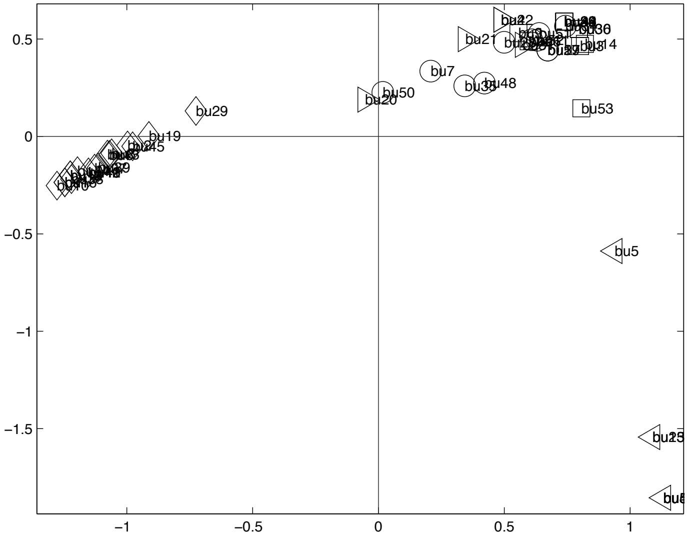

Figure 3.6. *Projection of the row clusters into the factorial plane spawned by the first and second axes*

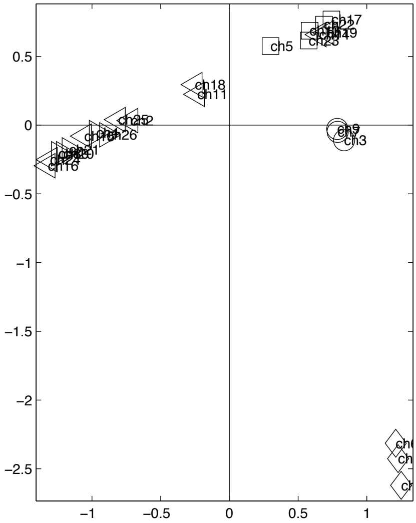

Figure 3.7. *Projection of the column clusters into the factorial plane spawned by the first and second axes*

3.9. Conclusion

Placing the problem of co-clustering within the maximum likelihood and classification maximum likelihood approaches, we have described two algorithms LBVEM and LBCEM. In terms of co-clustering and estimation, LBVEM appears more efficient than LBCEM. However, as we have emphasized, LBCEM has the advantage of the speed of convergence. Besides, note that although LBVEM does not maximize the likelihood, as in the classical mixture model situation, but only maximizes a fuzzy criterion, the variational approximation seems good. However, and despite the quality of the results obtained by the approximation proposed and commonly used in physics as a means to avoid the heavy computation involved in the E step of EM (see, for instance, [HOF 99a]), the link between the fuzzy criterion and the true log-likelihood, and not its approximation, deserves future investigation.