Chapter 4

Co-Clustering of Contingency Tables

The co-clustering methods have practical importance in a wide variety of applications such as document clustering where the data are often arranged as two-way contingency tables. In this situation (see the Introduction), the sets I and J are two categorical variables. Therefore the rows and columns correspond to different categories of the two variables and the data matrix, which displays the frequency distribution of the variables, is the contingency table. The sets I and J may also be any two sets with a data matrix defining a binary relation on $I \times J$. This chapter presents a coherent framework to understand some existing criteria and algorithms for analyzing contingency tables and to propose new tables. Two approaches of the problem of co-clustering are studied.

– The first approach is based on the minimization of an objective function based on the measures of association. Two algorithms, called CROKI2 and CROINFO, based on chi-squared statistic and mutual information are studied. Note that the proposed formalism can be extended to other measures of association. Furthermore, links between these

Co-Clustering: Models, Algorithms and Applications Gérard Govaert and Mohamed Nadif
© 2014 ISTE Ltd. Published 2014 by ISTE Ltd.

two algorithms can be established, and finally, we justify the use of correspondence analysis (CA) as a method of visualization.

– The second approach is based on a probabilistic approach. In this situation, two models are considered: the *block model* that gives a probabilistic interpretation of the criteria optimized by CROKI2 and CROINFO and the *latent block model* (LBM) that offers not only a probabilistic interpretation by using Poisson distributions, also but new criteria and algorithms in the family of the expectation-maximization algorithm [DEM 77], referred to hereafter as PLBVEM and PLBCEM. The latter can offer rich perspectives in the selection model context, a problem that we do not deal with here.

This chapter is organized as follows. In sections 4.1 and 4.2, we give the necessary background on measures of association used in this chapter. In section 4.3, we present the co-clustering approach based on these association measures. Section 4.4 deals with model-based co-clustering: a block model and an LBM recently developed in [GOV 10]. Fuzzy and hard co-clustering algorithms are described in detail and some connections between them are established. In section 4.5, we focus on the comparison and illustrations of our approach.

4.1. Measures of association

The contingency table characterizes the dependency links between two sets I and J, and measuring the strength of this association is a long tradition in statistics, going back at least to the work of Pearson [PEA 00]. In this chapter, two measures of association defined in the following sections will be used.

4.1.1. *Phi-squared coefficient*

Many measures of association arise from the standard chi-squared statistic upon which a test of independence is usually based

$$\chi^2(\mathbf{x}) = \sum_{i,j} \frac{(x_{ij} - x_{i.}x_{.j}/N)^2}{x_{i.}x_{.j}/N}$$

where $x_{i.}$, $x_{.j}$ and N are notations defined in Introduction.

The χ^2 measure usually provides statistical evidence of a significant association, or dependency, between rows and columns in the table. The phi-squared Pearson's coefficient of mean squared contingency is based on this χ^2 statistic as follows:

$$\begin{aligned}\Phi^2(P_{IJ}) &= \frac{\chi^2(\mathbf{x})}{N} = \sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} \\ &= \sum_{i,j} \frac{p_{ij}^2}{p_{i.}p_{.j}} - 1 = \sum_{i,j} p_{i.}p_{.j}(\frac{p_{ij}}{p_{i.}p_{.j}} - 1)^2.\end{aligned}$$

Considering $\mathcal{N}$ the cloud of $I \times J$, the set of $n \times d$ points belonging to $\mathbb{R}$ weighted by $p_{i.}p_{.j}$, the center of gravity of this cloud is therefore equal to 1. The last formulation shows that the discrepancy of the independence model is taken into account by considering $x_{ij} = \frac{p_{ij}}{p_{i.}p_{.j}} - 1$ as a general term of $\mathbf{x}$. More generally, the phi-squared coefficient can be seen as an estimation of the deviation between both the probabilities $\xi_{i.}\xi_{.j}$, which we would have if the two categorical random variables were independent, and the probabilities ξ_{ij}. The phi-squared coefficient corrects chi-squared sensitivity to sample size by dividing the chi-squared statistic by the number of cases. It can be shown that this coefficient ranges between 0, indicative of complete independence, and

$\min(n,d)-1$, indicative of perfect association. Since its range depends on the dimensions of the table and to overcome this difficulty, variations of this measure have been proposed as the Pearson's contingency coefficient $\sqrt{\frac{\Phi^2}{1+\Phi^2}}$, the Tschuprow's contingency coefficient $t = \frac{\Phi^2}{\sqrt{(n-1)(d-1)}}$ or the Cramer's contingency coefficient $v = \sqrt{\frac{\Phi^2}{min(n,d)-1}}$.

Note also that this chi-squared statistic measure of association is the "base" of the CA. The CA method is an exploratory multivariate technique that converts a contingency table into a particular type of graphical display in which the rows and the columns of the matrix are depicted as points [BEN 73b, GRE 07]. It aims to study the relationship between two categorical variables, precisely by describing the variability of the row and column profiles. It can be used on any two-way table, sparse or not, as the case may be. It projects the rows and columns of a data matrix onto points within a graph in a Euclidean space. The graphs or maps that are the factorial planes spawned by the different axes are then used to gain some understanding of the data and to extract information from them. Typically, this method of visualization proved its usefulness by giving a representation of the rows and columns. Hence, CA and co-clustering in this context can be more beneficial for contingency tables. For instance, consider that our data set in Table 4.1 is a contingency table resulting from the two sets of rows and columns $\{r1,\ldots,r6\}$ and $\{c1,\ldots,c5\}$. By applying CA onto this table, we obtain a good representation shown in Figure 4.1; the percentage of the total inertia explained by the factorial plane spawned by the first and second axes is equal to 98.87%. Note that we can observe some clusters such as $\{r_1,r_2\}$, $\{r3,r4\}$, $\{r5,r6\}$, $\{c1,c2,c3\}$ and $\{c4,c5\}$.

	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 4.1. *Example of contingency table and associated joint distribution*

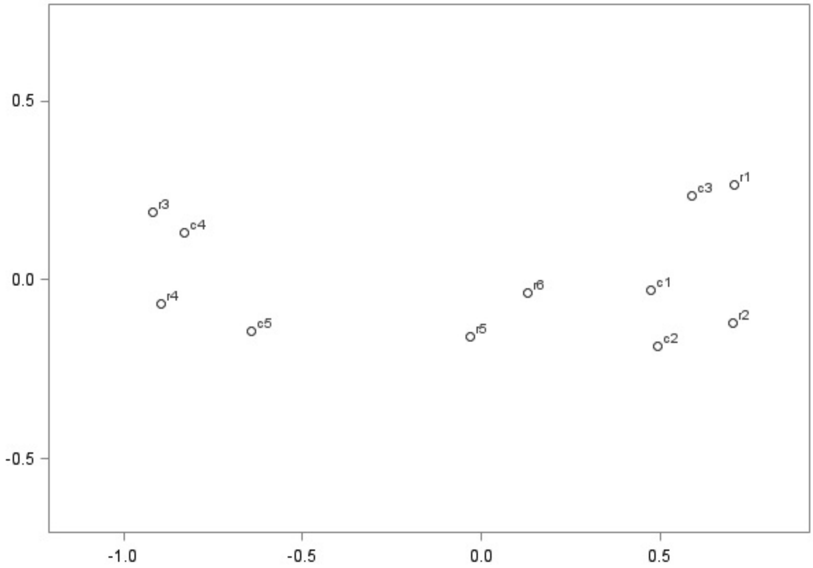


Figure 4.1. *CA: projection of the rows and columns into the factorial plane spawned by the first and second axes of CA*

4.1.2. *Mutual information*

To characterize the association between two categorical variables, the likelihood-ratio statistic, also called likelihood ratio chi-squared,

$$G^2(\mathbf{x}) = 2\sum_{i,j} x_{ij} \log \frac{x_{ij}}{x_{i.}x_{.j}/N}$$

is an alternative to the chi-squared statistic. It statistically compares the maximum likelihood of an unrestricted model with parameters ξ_{ij} with a restricted model with parameters $\xi_{i.}\xi_{.j}$. In this setting, the unrestricted model consists of the observed frequencies in the data and the restricted model consists of the expected frequencies under the null hypothesis of no association. The normalization of this measure yields the mutual information measure

$$\mathcal{I}(P_{IJ}) = \frac{G^2(\mathrm{x})}{2N} = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}},$$

an information-theoretic notion usually defined by

$$\mathcal{I}(P_{IJ}) = \mathrm{H}(P_I) + \mathrm{H}(P_J) - \mathrm{H}(P_{IJ})$$

where $\mathrm{H}(P_I) = -\sum_i p_{i.} \log p_{i.}$ and $\mathrm{H}(P_J) = -\sum_j p_{.j} \log p_{.j}$ are the marginal entropies and $\mathrm{H}(P_{IJ}) = -\sum_{i,j} p_{ij} \log p_{ij}$ is the joint entropy of I and J.

Note that there is a link between phi-squared coefficient and mutual information described in the following proposition:

PROPOSITION 4.1.– $\mathcal{I}(P_{IJ}) = \frac{1}{2}\Phi^2(P_{IJ}) + o(\sum_{ij}(p_{ij} - p_{i.}p_{.j})^2)$.

PROOF.– Denoting $e_{ij} = p_{ij} - p_{i.}p_{.j}$ for all i, j and using the following equation $\log(1+x) = x - x^2/2 + o(x^2)$, it can be written as

$$\log(1 + \frac{e_{ij}}{p_{i.}p_{.}}) = \frac{e_{ij}}{p_{i.}p_{.j}} - \frac{1}{2}\left(\frac{e_{ij}}{p_{i.}p_{.j}}\right)^2 + o(e_{ij}^2),$$

$$\begin{aligned}(p_{ij} + e_{ij}) \log(1 + \frac{e_{ij}}{p_{i.}p_{.}}) &= e_{ij} - \frac{1}{2}\frac{e_{ij}^2}{p_{i.}p_{.j}} + \frac{e_{ij}^2}{p_{i.}p_{.j}} - \frac{1}{2}\frac{e_{ij}^3}{p_{i.}^2 p_{.j}^2} + o(e_{ij}^2) \\ &= e_{ij} + \frac{1}{2}\frac{e_{ij}^2}{p_{i.}p_{.j}} + o(e_{ij}^2)\end{aligned}$$

and therefore

$$\sum_{i,j}(p_{ij}+e_{ij})\log(1+\frac{e_{ij}}{p_{i.}p_{.}}) = \underbrace{\sum_{i,j} e_{ij}}_{=0} + \frac{1}{2}\sum_{i,j}\frac{e_{ij}^2}{p_{i.}p_{.j}} + o(\sum_{i,j} e_{ij}^2),$$

which shows the proposition. ■

This link justifies the use of CA, which is based on Φ^2, as an appropriate visualization method for the rows and columns of contingency tables used in our examples.

4.2. Contingency table associated with a couple of partitions

In this section, we study the effect of simultaneously aggregating the rows and the columns of a contingency table x according to a couple of partitions of the sets I and J. If z and w are partitions in g clusters and m clusters of the set I of the rows and the set J of columns of the contingency table x, a new two-way contingency table $\mathrm{x}^{\mathbf{zw}} = (x_{k\ell}^{\mathbf{zw}})$ can be associated with two categorical random variables that take values in sets $K = \{1,\ldots,g\}$ and $L = \{1,\ldots,m\}$ by merging the rows and columns according to the partitions z and w

$$x_{k\ell}^{\mathbf{zw}} = \sum_{i,j} z_{ik}w_{j\ell}x_{ij} \qquad \forall k \in K \quad \text{and} \quad \forall \ell \in L.$$

4.2.1. *Associated distributions*

The first distribution that can be associated with z and w is the distribution $P_{KL}^{\mathbf{zw}} = (p_{k\ell}^{\mathbf{zw}})$ defined on $K \times L$ by

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

The following equation

$$\sum_{k,\ell} p_{k\ell}^{\mathbf{zw}} = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} p_{ij} = \sum_{i,j} p_{ij} \underbrace{\sum_{k,\ell} z_{ik} w_{j\ell}}_{=1} = 1$$

shows that $P_{KL}^{\mathbf{zw}}$ is a distribution. Moreover, it can be noted that, as

$$\sum_{\ell} p_{k\ell}^{\mathbf{zw}} = \sum_{i,j,\ell} z_{ik} w_{j\ell}\, p_{ij} = \sum_{i} (z_{ik} \sum_{j} (p_{jj} \underbrace{\sum_{\ell} w_{j\ell}}_{=1})) = \sum_{i} z_{ik} p_{i.},$$

the row margins of this new distribution do not depend on the partition w and will be denoted as $p_{k.}^{\mathbf{z}}$. Similarly, the column margins $\sum_k p_{k\ell}^{\mathbf{zw}}$ are equal to $\sum_j w_{j\ell} p_{.j}$ and will be denoted as $p_{.\ell}^{\mathbf{w}}$.

The second distribution that can be associated with z and w is the distribution $Q_{IJ}^{\mathbf{zw}} = (q_{ij}^{\mathbf{zw}})$ defined on $I \times J$ by

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

The following equation

$$\sum_{i,j} q_{ij}^{\mathbf{zw}} = \sum_{i,j} p_{i.} p_{.j} \sum_{k,\ell} z_{ik} w_{j\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} = \sum_{k,\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} \underbrace{\sum_{i,j} p_{i.} p_{.j} z_{ik} w_{j\ell}}_{=p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}}$$

$$= \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}} = 1,$$

shows that $Q^{\mathbf{zw}}_{IJ}$ is a distribution. Moreover, as

$$q^{\mathbf{z}}_{i.} = \sum_j q^{\mathbf{zw}}_{ij} = \sum_j p_{i.}p_{.j} \sum_{k,\ell} z_{ik} w_{j\ell} \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}}$$

$$= p_{i.} \sum_k \frac{z_{ik}}{p^{\mathbf{z}}_{k.}} \underbrace{\sum_\ell \frac{\overbrace{\sum_j (w_{j\ell}p_{.j})}^{p^{\mathbf{w}}_{.\ell}} p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{w}}_{.\ell}}}_{=p^{\mathbf{z}}_{k.}} = p_{i.}$$

and, symmetrically, $q^{\mathbf{w}}_{.j} = p_{.j}$, this distribution has the same margins as the initial distribution P_{IJ}.

Distributions $Q^{\mathbf{zw}}_{IJ}$ and $P^{\mathbf{zw}}_{KL}$ can be illustrated using the example described in Table 4.1. For instance, the aggregation of the rows and columns of the data according to the partitions $\mathbf{z} = (1,1,2,2,3,3)$ and $\mathbf{w} = (1,1,1,2,2)$ leads to the contingency table $\mathbf{x}^{\mathbf{zw}}$ and the distribution $P^{\mathbf{zw}}_{KL}$ presented in Table 4.2.

	1	2	
1	30.0	2.00	32.0
2	4.00	23.0	27.0
3	25.0	16.0	41.0
	59.0	41.0	100

	1	2	
1	0.30	0.02	0.32
2	0.04	0.23	0.27
3	0.25	0.16	0.41
	0.59	0.41	1.00

Table 4.2. *Aggregated contingency table* $\mathbf{x}^{\mathbf{zw}}$*(left) and associated distribution* $P^{\mathbf{zw}}_{KL}$ *(right)*

Table 4.3 gives the original distribution P_{IJ} and the distribution $Q^{\mathbf{zw}}_{IJ}$ obtained after aggregating the rows and columns. It can be seen that, in this example, these two distributions are similar. Looking at the factorial plan in Figure 4.1, associations between row clusters and column clusters can be observed. We can then make sense of the row clusters and column clusters simultaneously: $\{r_1, r_2\}$

characterized by high values in $\{c1, c2, c3\}$ and small values in $\{c4, c5\}$, $\{r3, r4\}$ characterized by high values in $\{c4, c5\}$ and small values in $\{c1, c2, c3\}$ and $\{r5, r6\}$ characterized by high values in $\{c1, c2, c3\}$ and moderate values in $\{c4, c5\}$.

	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 4.3. *Distributions P_{IJ} (left) and $Q^{\mathbf{zw}}_{IJ}$ (right)*

4.2.2. *Associated measures of association*

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

$$\Phi^2(P^{\mathbf{zw}}_{KL}) = \sum_{k,\ell} \frac{(p^{\mathbf{zw}}_{k\ell} - p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell})^2}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}}, \quad \Phi^2(Q^{\mathbf{zw}}_{IJ}) = \sum_{i,j} \frac{(q^{\mathbf{zw}}_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

$$\mathcal{I}(P^{\mathbf{zw}}_{KL}) = \sum_{k,\ell} p^{\mathbf{zw}}_{k\ell} \log \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}} \text{ and } \quad \mathcal{I}(Q^{\mathbf{zw}}_{IJ}) = \sum_{i,j} q^{\mathbf{zw}}_{ij} \log \frac{q^{\mathbf{zw}}_{ij}}{p_{i.}p_{.j}}.$$

For the rest of this chapter, it is interesting to establish the relationship among the measures of association of the three distributions P_{IJ}, $P^{\mathbf{zw}}_{KL}$ and $Q^{\mathbf{zw}}_{IJ}$. These properties, which are similar for both types of measure of association, are grouped into the following two propositions.

PROPOSITION 4.2.–

$$\Phi^2(P^{\mathbf{zw}}_{KL}) = \Phi^2(Q^{\mathbf{zw}}_{IJ}),$$

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

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 distributions P_{IJ} and $Q_{IJ}^{\mathbf{zw}}$, $\Phi^2(Q_{IJ}^{\mathbf{zw}}) \leq \Phi^2(P_{IJ})$ or $\Phi^2(P_{KL}^{\mathbf{zw}}) \leq \Phi^2(P_{IJ})$.

LEMMA 4.1.–

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

PROOF.–

$$\sum_{i,j} \frac{p_{ij}q_{ij}^{\mathbf{zw}}}{p_{i.}p_{.j}} = \sum_{i,j} p_{ij} \frac{p_{i.}p_{.j}\sum_{k,\ell} z_{ik}w_{j\ell}\frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}}}{p_{i.}p_{.j}}$$

$$= \sum_{k,\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}} \sum_{i,j} z_{ik}w_{j\ell}p_{ij} = \sum_{k,\ell} \frac{(p_{k\ell}^{\mathbf{zw}})^2}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}}$$

$$\sum_{i,j} \frac{(q_{ij}^{\mathbf{zw}})^2}{p_{i.}p_{.j}} = \sum_{i,j} \frac{\left(p_{i.}p_{.j}\sum_{k,\ell} z_{ik}w_{j\ell}\frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}}\right)^2}{p_{i.}p_{.j}}$$

$$= \sum_{i,j} p_{i.}p_{.j} \sum_{k,\ell} z_{ik}w_{j\ell}\frac{(p_{k\ell}^{\mathbf{zw}})^2}{(p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}})^2} = \sum_{k,\ell} \frac{(p_{k\ell}^{\mathbf{zw}})^2}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}}.$$

■

PROOF.– The first equation can easily be deduced from lemma 4.1. We have

$$\Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{\mathbf{zw}}) = \left(\sum_{i,j} \frac{(p_{ij})^2}{p_{i.}p_{.j}} - 1\right) - \left(\sum_{i,j} \frac{(q_{ij}^{\mathbf{zw}})^2}{q_{i.}^{\mathbf{z}}q_{.j}^{\mathbf{w}}} - 1\right)$$

$$= \sum_{i,j} \frac{p_{ij}^2}{p_{i.}p_{.j}} - \sum_{i,j} \frac{(q_{ij}^{\mathbf{zw}})^2}{p_{i.}p_{.j}}$$

and using lemma [4.1], this expression can be written as

$$\sum_{i,j} \frac{p_{ij}^2}{p_{i.}p_{.j}} - \sum_{i,j} \frac{p_{ij}q_{ij}^{\mathbf{zw}}}{p_{i.}p_{.j}} = \sum_{i,j} p_{ij} \left(\frac{p_{ij}}{p_{i.}p_{.j}} - \frac{q_{ij}^{\mathbf{zw}}}{p_{i.}p_{.j}} \right)$$

and

$$\sum_{i,j} \frac{p_{ij}^2}{p_{i.}p_{.j}} - 2\sum_{i,j} \frac{p_{ij}q_{ij}^{\mathbf{zw}}}{p_{i.}p_{.j}} + \sum_{i,j} \frac{(q_{ij}^{\mathbf{zw}})^2}{p_{i.}p_{.j}} = \sum_{i,j} \frac{(p_{ij} - q_{ij}^{\mathbf{zw}})^2}{p_{i.}p_{.j}}$$

that shows the second equation. The third equation can easily be deduced from the previous equation. ■

PROPOSITION 4.3.–

$$\mathcal{I}(P_{KL}^{\mathbf{zw}}) = \mathcal{I}(Q_{IJ}^{\mathbf{zw}})$$

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

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 distance between the two distributions P_{IJ} and $Q_{IJ}^{\mathbf{zw}}$),

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

PROOF.–

$$\begin{aligned}
\mathcal{I}(Q_{IJ}^{\mathbf{zw}}) &= \sum_{i,j} q_{ij}^{\mathbf{zw}} \log \frac{q_{ij}^{\mathbf{zw}}}{p_{i.}p_{.j}} \\
&= \sum_{i,j} \left(p_{i.}p_{.j} \left(\sum_{k,\ell} z_{ik}w_{j\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}} \right) \log \left(\sum_{k,\ell} z_{ik}w_{j\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}} \right) \right) \\
&= \sum_{i,j} \left(p_{i.}p_{.j} \left(\sum_{k,\ell} z_{ik}w_{j\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}} \log \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}} \right) \right) \\
&= \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}} \log \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}} = \mathcal{I}(P_{KL}^{\mathbf{zw}})
\end{aligned}$$

$$\begin{aligned}
\mathcal{I}(P_{IJ}) - \mathcal{I}(Q^{\mathbf{zw}}_{IJ}) = \mathcal{I}(P_{IJ}) - \mathcal{I}(P^{\mathbf{zw}}_{KL}) &= \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{k,\ell} p^{\mathbf{zw}}_{k\ell} \log \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}} \\
&= \sum_{i,j,k,\ell} z_{ik}w_{j\ell}p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{i,j,k,\ell} (z_{ik}w_{j\ell}p_{ij}) \log \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}} \\
&= \sum_{i,j,k,\ell} z_{ik}w_{j\ell}p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}} \frac{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}}{p^{\mathbf{zw}}_{k\ell}} \\
&= \sum_{i,j,k,\ell} z_{ik}w_{j\ell}p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}\frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}}} \\
&= \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}\sum_{k,\ell} z_{ik}w_{j\ell}\frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}}} \\
&= \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}} = \mathrm{KL}(P_{IJ}||Q^{\mathbf{zw}}_{IJ}).
\end{aligned}$$

The third equation can easily be deduced from the previous equation. ■

4.3. Co-clustering of contingency table

4.3.1. *Two equivalent approaches*

Propositions 4.2 and 4.3 show that the properties of the co-clustering defined by the couple of partitions z and w can equally be expressed in terms of the distribution $P^{\mathbf{zw}}_{KL}$ or distribution $Q^{\mathbf{zw}}_{IJ}$. Two approaches of co-clustering of contingency table can be then considered:

– 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. This aim can be reached by collapsing the rows and columns according to row and column partitions to obtain a reduced contingency $\mathbf{x}^{\mathbf{zw}}$ table that preserves a relationship of interest. Using the two measures of association previously defined, the objective of co-clustering can be based on maximizing

$$\Phi^2(P_{KL}^{\mathbf{zw}}) \quad \text{or} \quad \mathcal{I}(P_{KL}^{\mathbf{zw}}),$$

or, equally, on minimizing

$$\Phi^2(P_{IJ}) - \Phi^2(P_{KL}^{\mathbf{zw}}) \quad \text{or} \quad \mathcal{I}(P_{IJ}) - \mathcal{I}(P_{KL}^{\mathbf{zw}}). \qquad [4.1]$$

– Co-clustering obtained by approximating the initial distribution: the co-clustering problem can be viewed as an approximation of the distribution P_{IJ} by a distribution according to co-clustering termed $Q_{IJ}^{\mathbf{zw}}$ by minimizing the difference between the measures of information of the initial distribution and of the new distribution

$$\Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{\mathbf{zw}}) \quad \text{or} \quad \mathcal{I}(P_{IJ}) - \mathcal{I}(Q_{IJ}^{\mathbf{zw}}). \qquad [4.2]$$

Finally, in both situations, the problem is finding the partitions z and w minimizing the criterion for the Φ^2 measure of association

$$\begin{aligned} \mathrm{W}_{\Phi^2}(\mathbf{z},\mathbf{w}) &= D_{\Phi^2}(P_{IJ}||P_{KL}^{\mathbf{zw}}) = \Phi^2(P_{IJ}) - \Phi^2(P_{KL}^{\mathbf{zw}}) \\ &= \Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{\mathbf{zw}}) \end{aligned}$$

or minimizing the criterion for the $\mathcal{I}$ measure of association

$$\begin{aligned} \mathrm{W}_{\mathcal{I}}(\mathbf{z},\mathbf{w}) &= \mathrm{KL}(P_{IJ}||P_{KL}^{\mathbf{zw}}) = \mathcal{I}(P_{IJ}) - \mathcal{I}(P_{KL}^{\mathbf{zw}}) \\ &= \mathcal{I}(P_{IJ}) - \mathcal{I}(Q_{IJ}^{\mathbf{zw}}). \end{aligned}$$

It can be shown that there is no loss of information ($\mathrm{W}_{\Phi^2}(\mathbf{z},\mathbf{w}) = \mathrm{W}_{\mathcal{I}}(\mathbf{z},\mathbf{w}) = 0$) if the merged rows and columns have the same profile, that is if $(\frac{p_{i1}}{p_{i.}},\ldots,\frac{p_{id}}{p_{i.}})$ is constant for all i belonging to a same row cluster and if $(\frac{p_{1j}}{p_{.j}},\ldots,\frac{p_{nj}}{p_{.j}})$ is constant for all j belonging to a same column cluster. This property of contingency tables, which is a fundamental property of CA [BEN 73b, GRE 88a] is the so-called principle of *distributional equivalence*. In this situation, it can be noted that the initial distribution P_{IJ} and the final distribution $Q_{IJ}^{\mathbf{zw}}$ are the same.

Next, we focus on two algorithms of co-clustering based on both equivalent approaches. The first algorithm consists of maximizing the expression [4.2] with phi-squared coefficient and the second algorithm consists of minimizing the expression [4.2] with mutual information. But before developing these algorithms, we introduce a third distribution to facilitate the optimization of the criteria.

4.3.2. *Parameter modification of criteria*

Parameter modification [WIN 87, BRY 88] consists of increasing the number of parameters in a criterion so that the optimal value of the original parameters is not changed but is easier to optimize. The k-means algorithm is the most standard example using this principle: replacing the trace criterion $g(\mathbf{z}) = \sum_{i,k} d^2(\boldsymbol{x}_i, \mathrm{g}_k)$ with the criterion

$$\widetilde{g}(\mathbf{z}, \mathbf{a}) = \sum_{i,k} d^2(\boldsymbol{x}_i, \mathbf{a}_k)$$

with the new parameter $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_g)$, the alternating optimization of this new criterion leads to the k-means algorithm.

Here, we introduce a new parameter $\boldsymbol{\delta} = (\delta_{k\ell})$, a matrix of size (g, m) where each $\delta_{k\ell}$ plays the role of the centroid of the block $k\ell$ and such that

$$\delta_{k\ell} > 0 \quad \forall k, \ell \quad \text{and} \quad \sum_{k,\ell} p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}} \delta_{k\ell} = 1. \qquad [4.3]$$

We will note $\Delta(\mathbf{z},\mathbf{w})$ the set of matrices $\boldsymbol{\delta}$ which verify these constraints. Using this parameter, a new distribution $R_{IJ}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}} = (r_{ij}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}})$ depending on partitions $\mathbf{z}$ and $\mathbf{w}$ and parameter $\boldsymbol{\delta}$ can be defined by

$$r_{ij}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}} = p_{i.}p_{.j} \sum_{k,\ell} z_{ik} w_{j\ell} \delta_{k\ell}.$$

The constraints [4.3] ensure that $R_{IJ}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}}$ is a distribution

$$\sum_{i,j} r_{ij}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}} = \sum_{i,j} p_{i.}p_{.j} \sum_{k,\ell} z_{ik} w_{j\ell} \delta_{k\ell} = \sum_{k,\ell} \delta_{k\ell} \sum_{i,j} p_{i.}p_{.j} z_{ik} w_{j\ell}$$
$$= \sum_{k,\ell} p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}} \delta_{k\ell} = 1.$$

This distribution $R_{IJ}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}}$ has also the same column and row margins that the distributions P_{IJ} and $Q_{IJ}^{\mathbf{z}\mathbf{w}}$

$$r_{i.}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}} = \sum_{j} r_{ij}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}} = \sum_{j} p_{i.}p_{.j} \sum_{k,\ell} z_{ik} w_{j\ell} \delta_{k\ell} = p_{i.} \sum_{j} p_{.j} = p_{i.}$$

and, symmetrically, $r_{.j}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}} = p_{.j}$.

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

$$\widetilde{W}_{\Phi^2}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \Phi^2(P_{IJ}) - \Phi^2(R_{IJ}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}}) \qquad [4.4]$$

or the new criterion for the $\mathcal{I}$ measure of association

$$\widetilde{W}_{\mathcal{I}}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \mathcal{I}(P_{IJ}) - \mathcal{I}(R_{IJ}^{\mathbf{z}\mathbf{w}\boldsymbol{\delta}}). \qquad [4.5]$$

Finally, two equations which will be used in the next section can be established

$$
\begin{aligned}
\widetilde{W}_{\Phi^2}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = D_{\Phi^2}(P_{IJ}||R_{KL}^{\mathbf{zw}\boldsymbol{\delta}}) &= \sum_{i,j} \frac{(p_{ij}-r_{ij})^2}{p_{i.}p_{.j}} \\
&= \sum_{i,j} \frac{(p_{ij}-p_{i.}p_{.j}\sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell})^2}{p_{i.}p_{.j}} \\
&= \sum_{i,j} p_{i.}p_{.j}\left(\frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell}\right)^2 \\
&= \sum_{i,j} p_{i.}p_{.j}\left(\frac{(\sum_{k,\ell} z_{ik}w_{j\ell})p_{ij}}{p_{i.}p_{.j}} - \sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell}\right)^2 \\
&= \sum_{i,j,k,\ell} z_{ik}w_{j\ell}p_{i.}p_{.j}\left(\frac{p_{ij}}{p_{i.}p_{.j}} - \delta_{k\ell}\right)^2,
\end{aligned}
\qquad [4.6]
$$

and

$$
\begin{aligned}
\widetilde{W}_{\mathcal{I}}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = KL(P_{IJ}||R_{IJ}^{\mathbf{zw}\boldsymbol{\delta}}) &= \sum_{i,j} p_{ij}\log\frac{p_{ij}}{r_{ij}} \\
&= \sum_{i,j} p_{ij}\log\frac{p_{ij}}{p_{i.}p_{.j}\sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell}} \\
&= \sum_{i,j} p_{ij}\log\frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{i,j} p_{ij}\log\sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell} \qquad [4.7] \\
&= \sum_{i,j} p_{ij}\log\frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{k\ell}\log\delta_{k\ell}\sum_{i,j} z_{ik}w_{ij}p_{ij} \\
&= \sum_{i,j} p_{ij}\log\frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}}\log\delta_{k\ell}.
\end{aligned}
$$

4.3.3. *Co-clustering with the phi-squared coefficient*

The maximization of the criterion $\widetilde{\mathrm{W}}_{\Phi^2}(\mathbf{z}, \mathrm{w}, \boldsymbol{\delta})$ can be obtained by alternating the three computations: $\mathbf{z} = \arg\min_{\mathbf{z}} \widetilde{\mathrm{W}}_{\Phi^2}(\mathbf{z}, \mathbf{w}, \boldsymbol{\delta})$, $\mathbf{w} = \arg\min_{\mathbf{w}} \widetilde{\mathrm{W}}_{\Phi^2}(\mathbf{z}, \mathbf{w}, \boldsymbol{\delta})$ and $\boldsymbol{\delta} = \arg\min_{\boldsymbol{\delta}} \widetilde{\mathrm{W}}_{\Phi^2}(\mathbf{z}, \mathrm{w}, \boldsymbol{\delta})$. Using equation [4.6], the criterion can be written

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

and therefore, in the computation of z, the minimization of $\widetilde{\mathrm{W}}_{\Phi^2}(\mathbf{z}, \mathrm{w}, \boldsymbol{\delta})$ for fixed w and $\boldsymbol{\delta}$ is obtained by assigning each row i to the cluster k maximizing $\sum_{j,\ell} w_{j\ell} p_{.j} \left(\frac{p_{ij}}{p_{i.} p_{.j}} - \delta_{k\ell} \right)^2$. Similarly, in the computation of w, the minimization of $\widetilde{\mathrm{W}}_{\Phi^2}(\mathbf{z}, \mathrm{w}, \boldsymbol{\delta})$ for fixed z and $\boldsymbol{\delta}$ is obtained by assigning each column j to the cluster ℓ maximizing $\sum_{i,k} z_{ik} p_{i.} \left(\frac{p_{ij}}{p_{i.} p_{.j}} - \delta_{k\ell} \right)^2$. Finally, for the computations of $\boldsymbol{\delta}$, using equation [4.6], the problem can be formulated as $\arg\min_{\delta_{k\ell}} F(\delta_{k\ell})$ for all k, ℓ, where

$$\begin{aligned} F(\delta_{k\ell}) &= \sum_{i,j} z_{ik} w_{j\ell} p_{i.} p_{.j} \left(\frac{p_{ij}}{p_{i.} p_{.j}} - \delta_{k\ell} \right)^2 \\ &= p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}} \delta_{k\ell}^2 - 2 p_{k\ell}^{\mathbf{zw}} \delta_{k\ell} + \sum_{i,j} z_{ik} w_{j\ell} \frac{p_{ij}^2}{p_{i.} p_{.j}} \end{aligned}$$

which yields to

$$\delta_{k\ell} = -\frac{-2 p_{k\ell}^{\mathbf{zw}}}{2(p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}})} = \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} \qquad \forall k, \ell.$$

Using the strategy described in section 2.4.1 of Chapter 2, we obtain the CROKI2 algorithm described in algorithm 4.1.

Algorithm 4.1 CROKI2

```
input: x, $g$, $m$
initialization: z, w, $\delta_{k\ell} = \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.} p^{\mathbf{w}}_{.\ell}}$
repeat
  repeat
    step 1. $z_i = \arg\min_k \sum_{j,\ell} w_{j\ell} p_{.j} (\frac{p_{ij}}{p_{i.} p_{.j}} - \delta_{k\ell})^2$
    step 2. $\delta_{k\ell} = \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.} p^{\mathbf{w}}_{.\ell}}$
  until convergence
  repeat
    step 3. $w_j = \arg\min_\ell \sum_{i,k} z_{ik} p_{i.} (\frac{p_{ij}}{p_{i.} p_{.j}} - \delta_{k\ell})^2$
    step 4. $\delta_{k\ell} = \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.} p^{\mathbf{w}}_{.\ell}}$
  until convergence
until convergence
return z, w, $\boldsymbol{\delta}$
```

This algorithm defines a sequence $(\mathbf{z}^{(t)}, \mathbf{w}^{(t)}, \boldsymbol{\delta}^{(t)})$ which monotonically decreases the criterion $\widetilde{\mathrm{W}}_{\Phi^2}(\mathbf{z}^{(t)}, \mathbf{w}^{(t)}, \boldsymbol{\delta}^{(t)})$ and, as the parameter is equal to $\delta^{(t)}_{k\ell} = \frac{(p^{\mathbf{zw}}_{k\ell})^{(t)}}{(p^{\mathbf{z}}_{k.})^{(t)} (p^{\mathbf{w}}_{.\ell})^{(t)}}$, we obtain

$$\begin{aligned}
\widetilde{\mathrm{W}}_{\Phi^2}(\mathbf{z}^{(t)}, \mathbf{w}^{(t)}, \boldsymbol{\delta}^{(t)}) &= \sum_{i,j,k,\ell} z^{(t)}_{ik} w^{(t)}_{j\ell}\, p_{i.} p_{.j} \left(\frac{p_{ij}}{p_{i.} p_{.j}} - \delta^{(t)}_{k\ell} \right)^2 \\
&= \sum_{i,j,k,\ell} z^{(t)}_{ik} w^{(t)}_{j\ell}\, p_{i.} p_{.j} \left(\frac{p_{ij}}{p_{i.} p_{.j}} - \frac{(p^{\mathbf{zw}}_{k\ell})^{(t)}}{(p^{\mathbf{z}}_{k.})^{(t)} (p^{\mathbf{w}}_{.\ell})^{(t)}} \right)^2 \\
&= \sum_{i,j} \frac{(p_{ij} - (q^{\mathbf{zw}}_{ij})^{(t)})^2}{p_{i.} p_{.j}} = W_{\Phi^2}(\mathbf{z}^{(t)}, \mathbf{w}^{(t)})
\end{aligned}$$

which proves that this algorithm also monotonically decreases the initial criterion $W_{\Phi^2}(\mathbf{z}^{(t)}, \mathbf{w}^{(t)})$. It can be shown that it converges to a local optimum of the criterion. CROKI2 is computationally efficient even for sparse data and its complexity can be shown to be $O(nz * it * (g + m))$ where nz is the number of non-zero values in the input contingency table and it the number of iterations. Empirically, 20 iterations are seen to suffice.

Hereafter, we propose to illustrate CROKI2 by applying it to time-budget data sets used in Chapter 1. Let us recall that we have a data matrix of size 28×10 where x_{ij} represents the amount of time spent on a variety of activities i by a population j during a given time period. The initial $\Phi^2(I, J)$ value is 0.14392 and the resulting $\Phi^2(\mathbf{z}, \mathrm{w})$ value is 0.11993. The percentage of Φ^2 accounted for by the co-clustering is very good in this example: more than 83% of Φ^2 is preserved. The clusters are the following:

– Row partition:

- waus wcus wawe wcwe wcyo wces;

- wayo waes wmes;

- wmus wmwe wmyo;

- wnau wnaw wnay wnae;

- maus mmus mcus mawe mmwe mcwe may0 mmyo mcyo maes mmes mces.

– Column partition:

- home child;

- prof tran;

- shop wash meal sleap tv leis.

and the reorganized data matrix is presented in Table 4.4. In Figure 4.5, we present data matrix $\mathrm{x}^{\mathbf{zw}}$ and its associated data matrix $\mathbf{P}^{\mathbf{zw}} = (\frac{p_{k\ell}}{p_{k.}p_{.\ell}}) \times 1,000$. With a multiple coefficient $p_{k.}p_{.\ell}$, each cell $\mathbf{P}^{\mathbf{zw}}$ measures the deviance between the centroid of a block and the centroid of the initial data that is equal to 1. The most interesting values are therefore those which are far from the mean 1 (that is 1000 in our case).

These values characterize the 15 blocks. For instance, column clusters 1 and 2 are features of row clusters 1, 4 and 5. Row cluster 1 has a high value for column cluster 1 and a small value for column cluster 2, while row clusters 4 and 5 have high values for column cluster 2 and small values for cluster 1. Note that reciprocally these row clusters are characteristic of these column clusters. At the same time, we note that row cluster 3 and column cluster 3 do not have any characteristics; the values are close to 1,000. This observation is confirmed by looking at the planes obtained by CA in Figures 4.2 and 4.3. The clusters are located near the center of gravity. Furthermore, the simultaneous visualization of planes confirms the opposition of row cluster 1 and row clusters 4 and 5 according to column clusters 1 and 2.

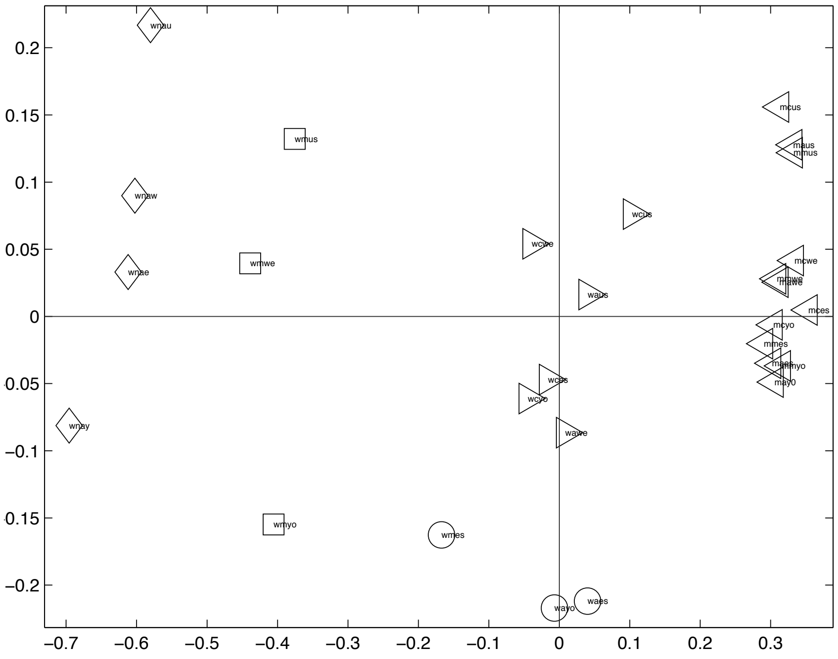


Figure 4.2. *Projection of the columns into the factorial plane spawned by the first and second axes that account for 84% of ϕ^2*

	home	child	prof	tran	shop	wash	meal	sleap	tv	leis
waus	250	30	475	90	140	120	100	775	115	305
wcus	196	18	482	94	141	130	96	775	132	336
wawe	307	30	510	70	80	95	142	815	87	262
wcwe	262	14	389	34	92	97	147	848	84	392
wcyo	318	23	413	89	112	96	102	774	45	409
wces	296	21	433	86	128	102	94	758	58	379
wayo	375	45	560	105	90	90	95	745	60	235
waes	338	42	578	106	106	94	52	752	64	228
wmes	431	60	434	77	117	88	105	770	73	229
wmus	421	87	179	29	161	112	119	776	143	373
wmwe	529	69	168	22	102	83	174	825	119	392
wmyo	576	59	260	52	116	85	117	775	65	295
wnau	495	110	10	0	170	110	130	785	160	430
wnaw	567	87	20	7	112	90	180	842	125	367
wnay	710	55	10	10	145	85	130	815	60	380
wnae	594	72	24	8	158	92	128	840	86	398
maus	60	10	610	140	120	95	115	760	175	315
mmus	65	10	615	141	115	90	115	765	180	305
mcus	50	0	585	115	150	105	100	760	150	385
mawe	95	7	652	100	57	85	150	807	115	330
mmwe	97	10	655	97	52	85	152	807	122	320
mcwe	72	0	642	105	62	77	140	812	100	387
may0	120	15	650	140	85	90	105	760	70	365
mmyo	112	15	650	145	85	90	105	760	80	357
mcyo	95	0	615	125	115	90	85	760	40	475
maes	122	22	650	142	76	94	100	764	96	334
mmes	134	22	652	133	68	94	102	762	122	310
mces	68	0	627	148	88	92	86	770	58	463

Table 4.4. *Co-clustering obtained with* CROKI2 *on a time-budget data set*

Finally, in Figure 4.4, a synthetic representation of initial and reorganized data, and the reorganized averaged data are shown.

	1	2	3
1	1765	3165	9363
2	1291	1860	3993
3	1741	710	4832
4	2690	89	6818
5	1201	9134	18456

	1	2	3
1	954	993	1011
2	1396	1168	863
3	1846	437	1024
4	2165	42	1097
5	322	1423	990

Table 4.5. $\mathrm{x}^{\mathbf{zw}}$*: Summary of initial data matrix and* $\mathbf{P}^{\mathbf{zw}}$*: deviance between each centroid and 1*

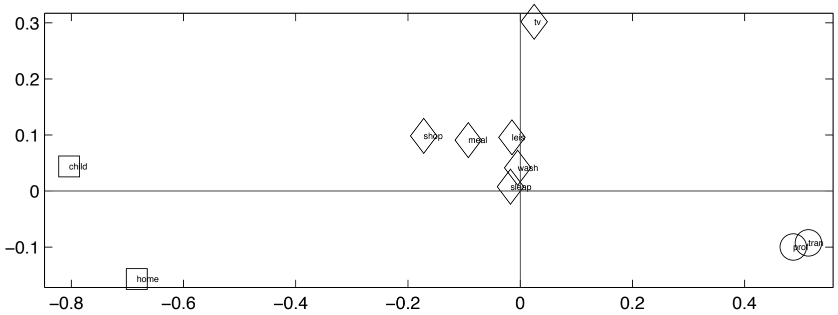


Figure 4.3. *Projection of the columns into the factorial plane spawned by the first and second axes that account for 84% of* ϕ^2

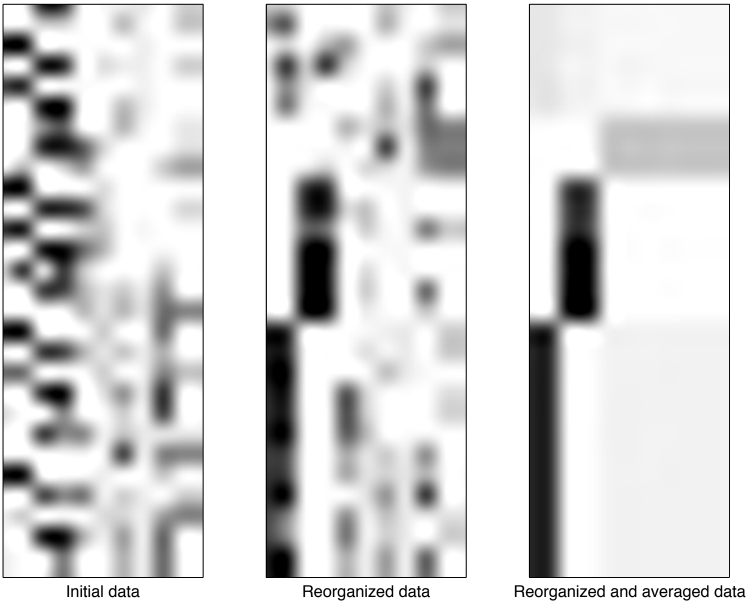
4.3.4. *Co-clustering with the mutual information*

The maximization of the criterion $\widetilde{\mathrm{W}}_{\mathcal{I}}(\mathbf{z},\mathrm{w},\boldsymbol{\delta})$ can be obtained by alternating the three computations: $\mathbf{z} = \arg\min_{\mathbf{z}} \widetilde{\mathrm{W}}_{\mathcal{I}}(\mathbf{z},\mathrm{w},\boldsymbol{\delta})$, $\mathbf{w} = \arg\min_{\mathbf{w}} \widetilde{\mathrm{W}}_{\mathcal{I}}(\mathbf{z},\mathrm{w},\boldsymbol{\delta})$ and $\boldsymbol{\delta} = \arg\min_{\boldsymbol{\delta}} \widetilde{\mathrm{W}}_{\mathcal{I}}(\mathbf{z},\mathrm{w},\boldsymbol{\delta})$. Using equation [4.7], the criterion can be written as

$$\widetilde{\mathrm{W}}_{\mathcal{I}}(\mathbf{z},\mathrm{w},\boldsymbol{\delta}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{i,k} z_{ik} \sum_{\ell} p_{i\ell}^{\mathbf{w}} \log \delta_{k\ell}$$

and therefore, in the computation of z, the minimization of $\widetilde{\mathrm{W}}_{\mathcal{I}}(\mathbf{z},\mathrm{w},\boldsymbol{\delta})$ for fixed w and $\boldsymbol{\delta}$ is obtained by assigning each row i to the cluster k maximizing

$$\sum_{\ell} (\sum_{j} w_{j\ell} p_{ij}) \log \delta_{k\ell}.$$



Initial data

Reorganized data

Reorganized and averaged data

Figure 4.4. *Visualization of initial data and results after co-clustering*

Similarly, in the computation of w, the minimization of $\widetilde{\mathrm{W}}_{\mathcal{I}}(\mathbf{z}, \mathrm{w}, \boldsymbol{\delta})$ for fixed z and $\boldsymbol{\delta}$ is obtained by assigning each column j to the cluster ℓ maximizing

$$\sum_k (\sum_i z_{ik} p_{ij}) \log \delta_{k\ell}.$$

Finally, for the computations of $\boldsymbol{\delta}$, the problem can be formulated as

$$\arg\max_{\boldsymbol{\delta}} \sum_{k,\ell} p^{\mathbf{zw}}_{k\ell} \log \delta_{k\ell}$$

with $\sum_{k,\ell} p^{\mathbf{z}}_{k.} p^{\mathbf{w}}_{.\ell} \delta_{k\ell} = 1$ which yields to $\delta_{k\ell} = \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.} p^{\mathbf{w}}_{.\ell}}$ for all k, ℓ. A strategy of alternated minimization similar to the CROKI2 algorithm leads to the CROINFO algorithm (algorithm 4.2) .

Algorithm 4.2 CROINFO

```
input: x, $g$, $m$
initialization: z, w, $\delta_{k\ell} = \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}}$
repeat
  repeat
    step 1. $z_i = \arg\min_k \sum_\ell p^{\mathrm{w}}_{i\ell} \log \delta_{k\ell}$
    step 2. $\delta_{k\ell} = \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}}$
  until convergence
  repeat
    step 3. $w_j = \arg\min_\ell \sum_k p^{\mathbf{z}}_{kj} \log \delta_{k\ell}$
    step 4. $\delta_{k\ell} = \frac{p^{\mathbf{zw}}_{k\ell}}{p^{\mathbf{z}}_{k.}p^{\mathbf{w}}_{.\ell}}$
  until convergence
until convergence
return z and w
```

It can also be shown that this algorithm monotonically decreases the criterion $\mathrm{W}_{\mathcal{I}}(\mathbf{z}^{(t)}, \mathrm{w}^{(t)})$ and its convergence properties are the same as the properties of the CROKI2 algorithm.

To illustrate the CROINFO algorithm, we present in Table 4.6 the distribution $R^{\mathbf{zw}\delta}_{IJ}$ obtained after the different steps of this algorithm are applied to the contingency table of Table 4.1. In Table 4.7, the values of phi-squared coefficient and mutual information of the initial distribution P_{IJ} and final distribution $Q^{\mathbf{zw}}_{IJ}$ are presented, and the corresponding criterion W expressing a loss information. It can be observed that this algorithm gives progressively better clustering and approximations to reach a loss information equal to 0.04.

4.4. Model-based co-clustering

Model-based clustering assumes that the data were generated by a model and tries to recover the original model from the data. For the co-clustering problem, these models (see Chapter 2) assumed that there exists a partition z into g

clusters on I and a partition w into m clusters on J, such that the random variables x_{ij} are conditionally independent knowing z and w. Two approaches can be considered:

After step **z** (iteration 1)

$R_{IJ}^{\mathbf{zw}\delta}$	1	2	3	4	5	**z**
1	0.035	0.030	0.029	0.034	0.032	2
2	0.037	0.032	0.027	0.035	0.030	1
3	0.029	0.025	0.027	0.028	0.030	3
4	0.029	0.025	0.023	0.027	0.026	2
5	0.046	0.040	0.038	0.044	0.042	2
6	0.042	0.036	0.039	0.040	0.043	3
w	2	2	1	2	1	1

After step δ (iteration 1)

$R_{IJ}^{\mathbf{zw}\delta}$	1	2	3	4	5	**z**
1	0.035	0.030	0.029	0.034	0.032	2
2	0.039	0.034	0.024	0.037	0.026	1
3	0.029	0.025	0.027	0.028	0.030	3
4	0.029	0.025	0.023	0.027	0.026	2
5	0.046	0.040	0.038	0.044	0.042	2
6	0.042	0.036	0.039	0.040	0.043	3
w	2	2	1	2	1	1

After step **w** (iteration 1)

$R_{IJ}^{\mathbf{zw}\delta}$	1	2	3	4	5	**z**
1	0.035	0.030	0.029	0.034	0.032	2
2	0.039	0.034	0.032	0.028	0.026	1
3	0.029	0.025	0.024	0.032	0.030	3
4	0.029	0.025	0.023	0.027	0.026	2
5	0.046	0.040	0.038	0.044	0.042	2
6	0.042	0.036	0.034	0.046	0.043	3
w	2	2	2	1	1	1

After step δ (iteration 1)

$R_{IJ}^{\mathbf{zw}\delta}$	1	2	3	4	5	**z**
1	0.035	0.030	0.028	0.034	0.033	2
2	0.056	0.048	0.046	0.005	0.005	1
3	0.023	0.020	0.019	0.040	0.038	3
4	0.028	0.024	0.023	0.028	0.027	2
5	0.045	0.039	0.037	0.045	0.043	2
6	0.033	0.028	0.027	0.057	0.055	3
w	2	2	2	1	1	1

After step **z** (iteration 2)

$R_{IJ}^{\mathbf{zw}\delta}$	1	2	3	4	5	**z**
1	0.056	0.048	0.046	0.005	0.005	1
2	0.056	0.048	0.046	0.005	0.005	1
3	0.023	0.020	0.019	0.040	0.038	3
4	0.021	0.018	0.017	0.037	0.035	3
5	0.045	0.039	0.037	0.045	0.043	2
6	0.043	0.037	0.035	0.043	0.041	2
w	2	2	2	1	1	1

After step δ (iteration 2)

$R_{IJ}^{\mathbf{zw}\delta}$	1	2	3	4	5	**z**
1	0.056	0.048	0.046	0.005	0.005	1
2	0.056	0.048	0.046	0.005	0.005	1
3	0.008	0.007	0.006	0.061	0.058	3
4	0.007	0.006	0.006	0.057	0.054	3
5	0.048	0.041	0.039	0.042	0.040	2
6	0.045	0.039	0.037	0.040	0.038	2
w	2	2	2	1	1	1

After step **w** (iteration 2)

$R_{IJ}^{\mathbf{zw}\delta}$	1	2	3	4	5	**z**
1	0.056	0.048	0.046	0.005	0.005	1
2	0.056	0.048	0.046	0.005	0.005	1
3	0.008	0.007	0.006	0.061	0.058	3
4	0.007	0.006	0.006	0.057	0.054	3
5	0.048	0.041	0.039	0.042	0.040	2
6	0.045	0.039	0.037	0.040	0.038	2
w	2	2	2	1	1	1

After step δ (iteration 2)

$R_{IJ}^{\mathbf{zw}\delta}$	1	2	3	4	5	**z**
1	0.056	0.048	0.046	0.005	0.005	1
2	0.056	0.048	0.046	0.005	0.005	1
3	0.008	0.007	0.006	0.061	0.058	3
4	0.007	0.006	0.006	0.057	0.054	3
5	0.048	0.041	0.039	0.042	0.040	2
6	0.045	0.039	0.037	0.040	0.038	2
w	2	2	2	1	1	1

Table 4.6. *First two iterations of* CROINFO

	Φ^2	$\mathcal{I}$
P_{IJ}	0.415	0.254
$P_{KL}^{\mathbf{zw}}$ or $Q_{IJ}^{\mathbf{zw}}$	0.378	0.214
W	0.037	0.040

Table 4.7. *Phi-squared coefficient and mutual information of initial distribution P_{IJ} and final distribution $Q_{IJ}^{\mathbf{zw}}$, and corresponding loss criterion W*

– In the first approach, the most common, the data are considered as a sample of size N of a bivariate categorical random variable with n and d levels and the contingency table x is an $n \times d$ matrix in which the cells contain frequency counts of the $n \times d$ possible outcomes. Therefore, the contingency table has the multinomial distribution $\mathcal{M}(N, p_{11}, \ldots, p_{nd})$ characterized by the probability p_{ij} of each cell. In this situation, all probabilistic models consist of giving the form of the probabilities p_{ij}, and the best-known ones for contingency tables are the log-linear models. The model proposed here, the *block model*, used the partitions of the levels of the two categorical variables as parameters of the model.

– In the second approach, co-occurrence data are defined in two samples (authors and publications, words and documents, and so on) of size n and d which, unlike the first approach are not fixed. Therefore, it is difficult to consider that the partitions, the dimensions of which will vary with the size of the data, are model parameters. The model proposed here, LBM, solves this problem by considering that the partitions are latent variables.

In this section, we examine these two approaches and show the links with the algorithms described in the previous section.

4.4.1. *Block model for contingency tables*

4.4.1.1. *Definition of the model*

This model considers that the contingency table x is distributed according to a multinomial distribution with parameters $(N, \xi_{11}, \ldots, \xi_{nd})$ where cell probabilities ξ_{ij} of the $I \times J$ contingency table take the following form

$$\xi_{ij} = \alpha_i \beta_j \sum_{k,\ell} z_{ik} w_{j\ell} \delta_{k\ell}$$

for a row effect α_i, a column effect β_j and a block effect $\delta_{k\ell}$. Its pdf takes the following form

$$\begin{aligned} f(\mathbf{x};\boldsymbol{\theta}) &= \frac{N!}{\prod_{i,j} x_{ij}!}\prod_{i,j}\xi_{ij}^{x_{ij}} \\ &= \frac{N!}{\prod_{i,j} x_{ij}!}\prod_{i,j}\left(\alpha_i\beta_j\sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell}\right)^{x_{ij}}, \end{aligned} \qquad [4.8]$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\mathbf{z},\mathbf{w})$ with $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)$, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)$, $\boldsymbol{\delta} = (\delta_{11},\ldots,\delta_{gm})$ and $\mathbf{z}$ and $\mathbf{w}$ are partitions of the rows I and columns J. As for the log-linear models for contingency tables, identifiability requires constraints. Denoting $\alpha_k = \sum_i z_{ik}\alpha_i$ and $\beta_\ell = \sum_j w_{j\ell}\beta_j$, the constraints

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \alpha_k\delta_{k\ell} = \sum_\ell \beta_\ell\delta_{k\ell} = 1 \quad \forall k,\ell,$$

which ensure that $\sum_{i,j}\xi_{ij} = 1$, will be chosen.

4.4.1.2. *Maximum likelihood estimation*

The maximum likelihood estimates are parameter values that maximize the likelihood [4.8] or the log-likelihood which can be written as

$$\begin{aligned} L(\mathbf{z},\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta}) &= \sum_{i,j} x_{ij}\log\left(\alpha_i\beta_j\sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell}\right) + C \\ &= \sum_i x_{i.}\log\alpha_i + \sum_j x_{.j}\log\beta_j \\ &\quad + \sum_{k,\ell} x_{k\ell}^{\mathbf{zw}}\log\delta_{k\ell} + C \end{aligned}$$

where C does not depend on z, w, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. It can be easily shown that $\alpha_i = p_{i.}$, $\beta_j = p_{.j}$ and the estimation of the parameters z, w and $\boldsymbol{\delta}$ are obtained by maximization of

$$\sum_{k,\ell} x_{k\ell}^{\mathbf{zw}} \log \delta_{k\ell} = N \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}} \log \delta_{k\ell}$$

which is exactly the problem of co-clustering with the mutual information discussed in section 4.3.4, formulated as

$$\arg\min_{\mathbf{z},\mathbf{w},\boldsymbol{\delta}} \widetilde{W_{\mathcal{I}}}(\mathbf{z}, \mathbf{w}, \boldsymbol{\delta}) = \arg\max_{\mathbf{z},\mathbf{w},\boldsymbol{\delta}} \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}} \log \delta_{k\ell}$$

and solved by the CROINFO algorithm. Thus, we give a probabilistic interpretation of the minimization problem of mutual information loss function by this algorithm. Note that we have used multinomial distribution and assumed N to be known. In certain situations, N can be unknown and in that case, we can use a Poisson distribution for counts of events that occur randomly. With this distribution, the same calculus can easily be performed and leads to the same conclusion.

4.4.1.3. *Minimum chi-squared estimation*

In the minimum chi-squared estimation [HAR 83], which is an alternative to maximum likelihood estimation, the parameter estimations are obtained by minimizing

$$A_1 = N \sum_{i,j} \frac{(p_{ij} - \xi_{ij})^2}{\xi_{ij}}.$$

The expected frequency $N\xi_{ij}$ in the denominator causes certain difficulties and a modification has been suggested which makes the computation easier. For instance, the modified chi-squared statistic proposed by Neyman [NEY 49]

$$A_2 = N \sum_{i,j} \frac{(p_{ij} - \xi_{ij})^2}{p_{ij}}$$

leads to estimators that have the same properties as those provided by the minimum chi-squared. Another solution is to use the modified chi-squared statistic

$$A_3 = N \sum_{i,j} \frac{(p_{ij} - \xi_{ij})^2}{p_{i.}p_{.j}}.$$

No theoretical property seems to have been proved but a very good approximation of the expression A_1 obtained by A_3 (much better than that obtained by A_2) on many examples led us to use this expression to establish the minimum chi-squared estimation.

Unfortunately, the estimation of row and column effects is difficult. One solution is to estimate these effects by their sufficient statistics. Under the multinomial model, the sufficient statistic is expressible as a linear combination of the corresponding cell proportions, and therefore a sufficient statistic of $\xi_{i.}$ is $p_{i.} = \frac{x_{i.}}{N}$ and, as we have

$$\xi_{i.} = \sum_j \alpha_i \beta_j \sum_{k,\ell} z_{ik} w_{j\ell} \delta_{k\ell} = \alpha_i \sum_k z_{ik} \underbrace{\sum_\ell \beta_\ell \delta_{k\ell}}_{1} = \alpha_i,$$

the $p_{i.}$s are sufficient statistics of row effects α_i. In a similar way, we obtain that the $p_{.j}$s are sufficient statistics of row effects β_j. Therefore, the estimations of parameters z, w and $\boldsymbol{\delta}$ are obtained by minimizing

$$\sum_{i,j} \frac{(x_{ij} - Np_{i.}p_{.j} \sum_{k\ell} z_{ik} w_{j\ell} \delta_{k\ell})^2}{Np_{i.}p_{.j}} = N \sum_{i,j} \frac{(p_{ij} - r_{ij}^{\mathbf{zw}\boldsymbol{\delta}})^2}{p_{i.}p_{.j}}$$

which is equal, up to the multiplicative constant N, to the criterion $\widetilde{W}_{\Phi^2}(\mathbf{z}, \mathrm{w}, \boldsymbol{\delta})$ and therefore the estimation can be solved by the CROKI2 algorithm.

Finally, the two approaches of co-clustering of a contingency table defined in the previous section clustering can be viewed as estimating the parameters of a probabilistic block model by two different estimation methods.

4.4.2. *Poisson latent block model*

4.4.2.1. *The model*

Using the LBM described in section 2.3 of Chapter 2 in the contingency table situation and assuming that for each block $k\ell$ the values x_{ij} are distributed according to the Poisson distribution $\mathcal{P}(\lambda_{ij})$ where the parameters λ_{ij} take the following form

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

for a row effect μ_i, a column effect ν_j and a block effect $\gamma_{k\ell}$, we obtain the Poisson LBM [GOV 10] with the following pdf

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \times \prod_{i,j} \frac{e^{-\mu_i\nu_j \sum_{k,\ell} z_{ik} w_{j\ell}\gamma_{k\ell}} \left(\mu_i\nu_j \sum_{k,\ell} z_{ik} w_{j\ell}\gamma_{k\ell}\right)^{x_{ij}}}{x_{ij}!} \qquad [4.9]$$

parameterized by $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma})$ with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)$ and $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{gm})$.

As for the previous block models, identifiability requires constraints. Denoting $\mu_k = \sum_i z_{ik}\mu_i$ and $\nu_\ell = \sum_j w_{j\ell}\nu_j$, the constraints

$$\sum_i \mu_i = \sum_j \nu_j = 1/\sum_k \pi_k \gamma_{k\ell} = 1/\sum_\ell \rho_\ell \gamma_{k\ell} = M \quad \forall k, \ell$$

are chosen. Under these constraints, it can be shown that $\mathbb{E}(x_{i.}) = \mu_i$ and $\mathbb{E}(x_{.j}) = \nu_j$ and μ_i and ν_j can be replaced by the margins $x_{i.}$ and $x_{.j}$; the parameters to be estimated are therefore reduced to $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and $\boldsymbol{\gamma}$. With this model, the complete-data log-likelihood is given, up to a constant, by

$$\begin{aligned}\mathrm{L_C}(\mathbf{z},\mathbf{w},\boldsymbol{\theta}) &= \sum_k z_{.k}\log\pi_k + \sum_\ell w_{.\ell}\log\rho_\ell \\ &\quad + \sum_{i,j,k,\ell} z_{ik}w_{j\ell}(x_{ij}\log\gamma_{k\ell} - x_{i.}x_{.j}\gamma_{k\ell}).\end{aligned}$$

4.4.3. *Poisson* Lbvem *and* Lbcem *algorithms*

Taking into account the definition of the pdf $f(x_{ij};\alpha_{k\ell})$, the variational approximation of the EM algorithm described in section 2.4.1 of Chapter 2 is obtained by alternating the three computations $\arg\max_{\widetilde{\mathbf{z}}} F_C(\widetilde{\mathbf{z}},\widetilde{\mathbf{w}},\boldsymbol{\theta})$, $\arg\max_{\widetilde{\mathbf{w}}} F_C(\widetilde{\mathbf{z}},\widetilde{\mathbf{w}},\boldsymbol{\theta})$ and $\arg\max_{\boldsymbol{\theta}} F_C(\widetilde{\mathbf{z}},\widetilde{\mathbf{w}},\boldsymbol{\theta})$.

– Computation of $\boldsymbol{\theta}$: we obtain $\pi_k = \frac{\widetilde{z}_{.k}}{n}$, $\rho_\ell = \frac{\widetilde{\mathrm{w}}_{.\ell}}{n}$ and $\gamma_{k\ell}$ is defined as follows

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

and therefore $\gamma_{k\ell} = \frac{x_{k\ell}^{\widetilde{\mathbf{z}}\widetilde{\mathbf{w}}}}{x_{k.}^{\widetilde{\mathbf{z}}}x_{.\ell}^{\widetilde{\mathbf{w}}}}$.

– Computation of $\widetilde{\mathbf{z}}$: we obtain

$$\begin{aligned}\widetilde{z}_{ik} = \arg\max_{\widetilde{z}_{ik}} \sum_k \Bigg(&\widetilde{z}_{ik}\log\pi_k \\ &+ \widetilde{z}_{ik}\sum_\ell (x_{i\ell}^{\widetilde{\mathbf{w}}}\log\gamma_{k\ell} - x_{i.}x_{.\ell}^{\widetilde{\mathbf{w}}}\gamma_{k\ell}) - \widetilde{z}_{ik}\log\widetilde{z}_{ik}\Bigg)\end{aligned}$$

and, as $\gamma_{k\ell} = \frac{x^{\widetilde{\mathbf{z}}'\widetilde{\mathbf{w}}}_{k\ell}}{x^{\widetilde{\mathbf{z}}'}_{k.}x^{\widetilde{\mathbf{w}}}_{.\ell}}$ where $\widetilde{\mathbf{z}}'$ is the previous probabilities, we have $\sum_\ell x_{i.}x^{\widetilde{\mathbf{w}}}_{.\ell}\gamma_{k\ell} = \sum_\ell x_{i.}x^{\widetilde{\mathbf{w}}}_{.\ell}\frac{x^{\widetilde{\mathbf{z}}'\widetilde{\mathbf{w}}}_{k\ell}}{x^{\widetilde{\mathbf{z}}'}_{k.}x^{\widetilde{\mathbf{w}}}_{.\ell}} = x_{i.}\frac{\sum_\ell x^{\widetilde{\mathbf{z}}'\widetilde{\mathbf{w}}}_{k\ell}}{x^{\widetilde{\mathbf{z}}}_{k.}} = x_{i.}$ and therefore $\widetilde{z}_{ik} = \arg\max_{\widetilde{z}_{ik}} \sum_k \left(\widetilde{z}_{ik}\log\pi_k + \widetilde{z}_{ik}\sum_\ell x^{\widetilde{\mathbf{w}}}_{i\ell}\log\gamma_{k\ell}\right.$ $\left.-\widetilde{z}_{ik}\log\widetilde{z}_{ik}\right)$. The constraint $\sum_k \widetilde{z}_{ik} = 1$ yields to $\widetilde{z}_{ik} \propto \pi_k \exp(\sum_\ell x^{\widetilde{\mathbf{w}}}_{i\ell}\log\gamma_{k\ell})$.

– Computation of $\widetilde{\mathbf{w}}$: similarly, we have $\widetilde{\mathrm{w}}_{j\ell} \propto \rho_\ell \exp(\sum_k x^{\widetilde{\mathbf{z}}}_{kj}\log\gamma_{k\ell})$.

Finally, we obtain algorithm 4.3. In a similar way, the maximization of the $\mathrm{L_C}$ criterion can be performed by algorithm 4.4.

Algorithm 4.3 Poisson LBVEM

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

REMARK 4.1.– When the proportions are assumed to be equal ($\pi_1 = \ldots = \pi_g$ and $\rho_1 = \ldots = \rho_m$), it can be observed that the Poisson LBCEM algorithm is equivalent to the CROINFO algorithm. Moreover, after the step of computation of γ, the complete data log-likelihood $\mathrm{L_C}(\mathbf{z},\mathbf{w},\boldsymbol{\gamma}(\mathbf{z},\mathbf{w}))$ becomes

$$\begin{aligned}\mathrm{L_C}(\mathbf{z},\mathbf{w},\boldsymbol{\gamma}(\mathbf{z},\mathbf{w})) &= \sum_{i,j,k,\ell} z_{ik}w_{j\ell}(x_{ij}\log\gamma_{k\ell} - x_{i.}x_{.j}\gamma_{k\ell})\\ &= \sum_{k,\ell} x_{k\ell}^{\mathbf{zw}}\log\frac{x_{k\ell}^{\mathbf{zw}}}{x_{k.}^{\mathbf{z}}x_{.\ell}^{\mathbf{w}}} - N\\ &= N\sum_{k,\ell} p_{k\ell}^{\mathbf{zw}}\log\frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}p_{.\ell}^{\mathbf{w}}} + N(\log N - 1)\\ &= N\mathcal{I}(P_{KL}^{\mathbf{zw}}) + C\end{aligned}$$

where C is a constant and therefore the maximization of the complete data log-likelihood of the Poisson LBM with equal proportions is equivalent to the maximization of the information criterion $\mathcal{I}(P_{KL}^{\mathbf{zw}})$ which shows that the maximization of $\mathcal{I}$ assumes implicitly that the proportions of rows and rows are equal. Hence, considering the unknown proportions, Poisson LBCEM clearly appears as an extension of CROINFO. The assignation and maximization are equivalent. This assumption (equality of proportions) usually not true in practice can lead to bad results.

4.5. Comparison of all algorithms

In this chapter, two unified frameworks have been presented: in the first framework, based on measures of association, the phi-squared coefficient and mutual information have been retained. In the second framework, using the model-based co-clustering approach, the block model and Poisson LBM have been considered. Both approaches yielded four algorithms, CROKI2, CROINFO, Poisson LBVEM and LBCEM summarized in Table 4.8.

Algorithm 4.4 Poisson LBCEM

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

Algorithm	Criterion
CROKI2	$\widetilde{W}_{\Phi^2}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \sum_{i,j,k,\ell} z_{ik}w_{j\ell}\, p_{i.}p_{.j}\left(\frac{p_{ij}}{p_{i.}p_{.j}} - \delta_{k\ell}\right)^2$
CROINFO	$\widetilde{W}_{\mathcal{I}}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \sum_{i,j} p_{ij}\log\frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{k,\ell} p^{\mathbf{zw}}_{k\ell}\log\delta_{k\ell}$
PLBCEM	$L_C(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = \sum_{i,j,k,\ell} z_{ik}w_{j\ell}(x_{ij}\log\gamma_{k\ell} - x_{i.}x_{.j}\gamma_{k\ell})$ $+ \sum_{i,k} z_{ik}\log\pi_k + \sum_{j,\ell} w_{j\ell}\log\rho_\ell$
PLBVEM	$F_C(\tilde{\mathbf{z}},\tilde{\mathbf{w}};\boldsymbol{\theta}) = \sum_{i,j,k,\ell} \tilde{z}_{ik}\tilde{w}_{j\ell}(x_{ij}\log\gamma_{k\ell} - x_{i.}x_{.j}\gamma_{k\ell})$ $+\sum_{i,k}\tilde{z}_{ik}\log\pi_k + \sum_{j,\ell}\tilde{w}_{j\ell}\log\rho_\ell$ $-\sum_{i,k}\tilde{z}_{ik}\log\tilde{z}_{ik} - \sum_{j,\ell}\tilde{w}_{j\ell}\log\tilde{w}_{j\ell}$

Table 4.8. *Algorithms and criteria*

In this section, we present case studies to demonstrate the difference among these algorithms. First, we compare CROKI2 and CROINFO in terms of convergence. Second, we show the impact of the proportions on the co-clustering performance of CROINFO. Third, we compare Poisson LBCEM and LBVEM in terms of co-clustering and estimation. Finally, we compare all the algorithms discussed in this chapter.

4.5.1. CROKI2 *versus* CROINFO

First, we study the behavior of the different criteria Φ^2 and $\mathcal{I}$ when applying CROKI2 and CROINFO. As co-clustering has been a topic of much interest in document clustering, we propose to use the Classic3 data set, a document collection from the SMART project at Cornell University. This data set consists of 3,891 documents and 4,303 words after removing stop words. It is classified into three classes denoted as *Medline* (1,033 documents), *Cisi* (1,460 documents) and *Cranfield* (1,398 documents). Note that the proportions of clusters are not dramatically different. To visualize this data set, we use the CA and we present, in Figure 4.5, the projection of rows and columns into the factorial plane spawned by the first and second axes.

We applied the two algorithms with $g = 3$ and $m = 3$. The CROKI2 algorithm naturally increases the objective function Φ^2 and the CROINFO algorithm increases $\mathcal{I}$. However, in our experiments, we have observed that each algorithm monotonically increases the two objective functions. We illustrate this behavior in Figure 4.6.

4.5.2. CROINFO *versus Poisson* LBCEM

As we discussed in section 4.4.3, the CROINFO and PLBCEM algorithms are the same when the proportions are assumed to be equal. We simulate a data set with dramatically different proportions ($\pi_1 = 0.1$, $\pi_2 = 0.9$, $\rho_1 = 0.178$ and $\rho_2 = 0.822$). We observe the projection of the sets of rows and columns by CA in Figure 4.7. In Tables 4.9 and 4.10, we present the confusion matrices between estimated $(\hat{\mathbf{z}}, \hat{\mathbf{w}})$, obtained by the application of both algorithms with $g = 2$ and $m = 2$, and the true partitions $(\mathbf{z}^T, \mathbf{w}^T)$. We note the good performance of Poisson LBCEM and therefore the impact of absence of proportions in the CROINFO criteria.

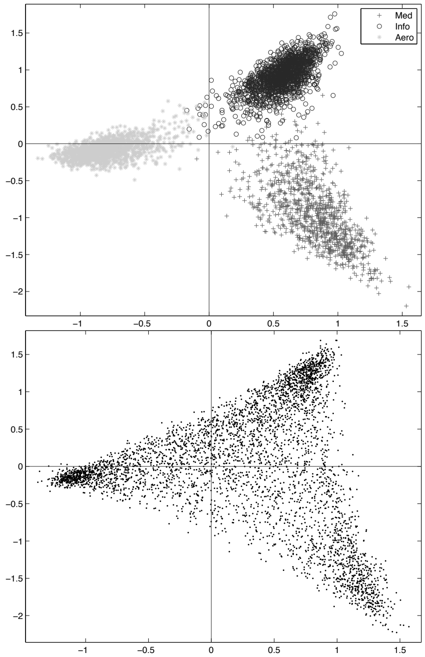


Figure 4.5. *a) Projection of the rows and b) the columns into the factorial plane spawned by the first and second axes*

	$\mathbf{z}_1^T$	$\mathbf{z}_2^T$
$\hat{\mathbf{z}}_1$	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 4.9. *Confusion matrices between estimated partitions* $(\hat{\mathbf{z}}, \hat{\mathbf{w}})$ *obtained by* CROINFO *and true partitions*$(\mathbf{z}^T, \mathbf{w}^T)$

	$\mathbf{z}_1^T$	$\mathbf{z}_2^T$
$\hat{\mathbf{z}}_1$	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 4.10. *Confusion matrices between estimated* $(\hat{\mathbf{z}}, \hat{\mathbf{w}})$ *obtained by Poisson* LBCEM *and true* $(\mathbf{z}^T, \mathbf{w}^T)$

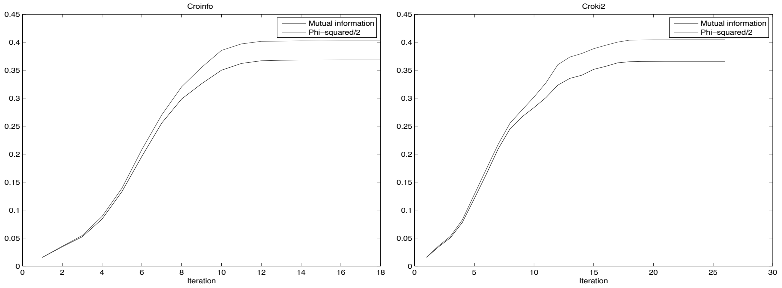


Figure 4.6. *Behavior of mutual information and phi-squared coefficient as a function of iterations by* CROINFO *and* CROKI2 *on Classic3 data*

4.5.3. *Poisson* LBVEM *versus Poisson* LBCEM

In this section, to illustrate the difference between the LBVEM and LBCEM algorithms in terms of behavior, we studied their performances on simulated data. In our experiments, we selected 12 types of data arising from 3×2 component mixture model corresponding to three degrees of cluster overlap (well separated, moderately separated and ill-separated), and four data dimensions ($n \times d = 50 \times 50$, $n \times d = 100 \times 60, n \times d = 200 \times 120$ and $n \times d = 300 \times 180$).

The concept of cluster separation is difficult to visualize for Poisson LBMs, but the degree of overlap can be measured by the true error rate, which is defined as the average misclassification probability $\mathbf{E}(\Gamma((\mathbf{z}, \mathbf{w}), r_B(\mathbf{x})))$ where r_B is the optimal Bayes' rule $r_B(\mathbf{x}) = \arg\max_{\mathbf{z},\mathbf{w}} P(\mathbf{z}, \mathbf{w}|\mathbf{x})$ associated with the LBM and $\Gamma((\mathbf{z}, \mathbf{w}), (\mathbf{z}^T, \mathbf{w}^T))$ is the proportion of misclassified items. Its computation being theoretically difficult, we used Monte Carlo simulations and approximated this error rate by comparing the partitions simulated with those we obtained by applying a classification step. Parameters were selected so as to obtain error rates, respectively, in [0.04, 0.06], for the well-separated (+), in [0.14, 0.16], for the moderately (++), and in [0.23, 0.26], for the ill-separated situations (+++).

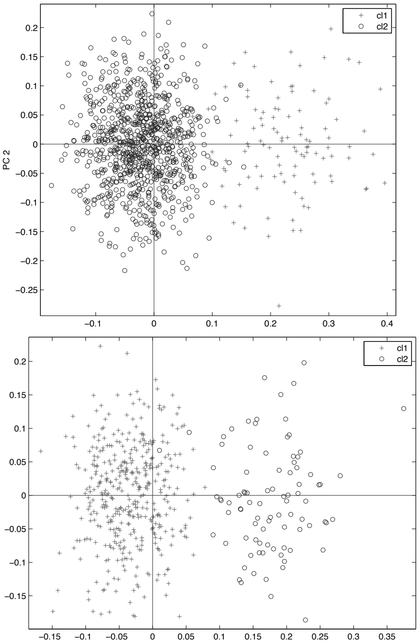


Figure 4.7. *a) Row representation onto plane 1,2 and b) column representation onto plane 1,2*

For each of these 12 data structures, 30 samples were generated and, for each sample, Poisson LBVEM and Poisson LBCEM algorithms were run 20 times starting from the same random situations. Then, the best solution for each method was selected. The simulation results are summarized in Tables 4.11 and 4.12. The first one displays the mean error rates and the second displays the mean Euclidean distance between true parameters and estimated parameters for each situation.

Degree of overlap	Size	PLBCEM	PLBVEM
+	50 × 50	0.0690	0.0709
	100 × 60	0.0647	0.0594
	200 × 120	0.0554	0.0539
	300 × 180	0.0576	0.0561
++	50 × 50	0.2740	0.2900
	100 × 60	0.1870	0.1768
	200 × 120	0.1617	0.1533
	300 × 180	0.1459	0.1372
+++	50 × 50	0.3418	0.3467
	100 × 60	0.2895	0.2937
	200 × 120	0.2488	0.2262
	300 × 180	0.2635	0.2341

Table 4.11. *Comparison of Poisson* LBVEM *and Poisson* LBCEM *in terms of mean error rates between estimated* $(\mathbf{z}, \mathbf{w})$ *and true partitions* $(\mathbf{z}^T, \mathbf{w}^T)$

Degree of overlap	Size	PLBCEM	PLBVEM
+	50 × 50	0.0061	0.0060
	100 × 60	0.0032	0.0030
	200 × 120	0.0015	0.0013
	300 × 180	0.0012	0.0011
++	50 × 50	0.0195	0.0178
	100 × 60	0.0048	0.0039
	200 × 120	0.0024	0.0018
	300 × 180	0.0013	0.0011
+++	50 × 50	0.0131	0.0109
	100 × 60	0.0068	0.0062
	200 × 120	0.0027	0.0019
	300 × 180	0.0024	0.0015

Table 4.12. *Comparison of Poisson* LBVEM *and Poisson* LBCEM *in terms of mean Euclidean distance between true parameters and estimated parameters*

The main findings arising from these experiments are the following. In terms of co-clustering, the Poisson LBVEM algorithm, even though it is slower than LBCEM, generally gives the best results, especially when the clusters are not well separated; not surprisingly, its performance increases with the size of the data. In terms of quality of estimation, LBVEM outperforms LBCEM (Table 4.12). This quality increases with the size of the data. This point provides an additional argument for its use in the selection model context.

4.5.4. *Behavior of* CROKI2, CROINFO, LBCEM *and* LBVEM

The four algorithms LBCEM, LBVEM, CROINFO and CROKI2 were evaluated on Classic3 data set in varying the number of column clusters m. The values of Φ^2, $\mathcal{I}$ and the number of documents misclassified denoted as e are presented in Table 4.13.

– We observe that the approximation of $\mathcal{I}$ by $\frac{1}{2}\Phi^2$ is very good when m is high enough, which is consistent with proposition 4.1.

– Unsurprisingly, on the one hand, CROKI2 maximizes Φ^2 and, on the other hand, CROINFO maximizes $\mathcal{I}$. We note that in all situations, CROINFO outperforms CROKI2 in terms of clustering. The mutual information appears more adapted to co-clustering.

– We observe that implicit dimensionality reduction by co-clustering actually gives better document clusters, in the sense that the cluster labels agree more with the true class labels with fewer word clusters. However, we note that the values of Φ^2 and $\mathcal{I}$ increase with m even though this growth becomes smaller from $m = 30$. The quality of clustering by CROKI2 and CROINFO starts to decrease from $m = 100$.

m	Algorithm	Φ^2	$\mathcal{I}$	e
3	CROINFO	0.8047575	0.3682842	52
	CROKI2	0.8094602	0.3645942	64
	PLBCEM	0.8044226	0.3682617	52
	PLBVEM	0.8036757	0.3678993	52
5	CROINFO	0.9248306	0.4522520	31
	CROKI2	0.9393426	0.4452273	51
	PLBCEM	0.9240040	0.4522505	32
	PLBVEM	0.9225218	0.4513789	28
10	CROINFO	0.9862210	0.4990168	29
	CROKI2	0.9973430	0.4924405	40
	PLBCEM	0.9844507	0.4988628	28
	PLBVEM	0.9838697	0.4965514	29
30	CROINFO	1.0193666	0.5270878	28
	CROKI2	1.0233885	0.5204656	47
	PLBCEM	1.0095892	0.5195290	26
	PLBVEM	1.0095892	0.5195290	26
40	CROINFO	1.0226414	0.5302415	32
	CROKI2	1.0266164	0.5257123	36
	PLBCEM	1.0124342	0.5220182	25
	PLBVEM	1.0124342	0.5220182	25
50	CROINFO	1.0240256	0.5320448	28
	CROKI2	1.0282436	0.5282652	37
	PLBCEM	1.0142280	0.5232399	28
	PLBVEM	1.0142280	0.5232399	28
100	CROINFO	1.0289786	0.5350597	30
	CROKI2	1.0310344	0.5319551	48
	PLBCEM	1.0159995	0.5255252	26
	PLBVEM	1.0159995	0.5255252	26

Table 4.13. *Performances of the four algorithms on Classic3 data set; e denotes the number of documents misclassified*

– Poisson LBCEM which is an extension of CROINFO appears more efficient, although the proportions are not dramatically different. We have discussed the impact of these parameters in section 4.5.2.

– As should be expected, the values of $\mathcal{I}$ obtained by LBCEM are smaller than those obtained by CROINFO but in all cases, the values obtained by these algorithms are very close. This is due to the relation between the criteria optimized by Poisson LBCEM and CROINFO.

– In all situations, LBVEM outperforms all the other algorithms. This performance was observed for other data sets and already confirmed by intensive numerical experiments on binary data (see, for instance, [GOV 05, GOV 08]).

4.6. Conclusion

In this chapter, two unified frameworks have been studied. In the first framework, based on measures of association, we have retained phi-squared coefficient and mutual information criteria. In the second framework, based on model-based co-clustering, we have considered the block model and Poisson LBM. Both approaches yielded four algorithms: CROKI2 CROINFO and Poisson LBVEM and Poisson LBCEM . We have established important connections between these algorithms. Note that our two approaches could be extended to other measures of association and other LBMs. In terms of performance, we have noted that Poisson LBVEM outperforms all the other algorithms. We have also selected some situations showing the importance of proportions in co-clustering and therefore the weakness of algorithms omitting this situation such as CROKI2 and CROINFO .

In an objective visualization, combining clustering and reduction for mapping clusters rather than individuals is therefore an appealing requirement for data analysis. One way to perform this task is by showing the clusters on a map after performing an *ad hoc* algorithm for partitioning the data set and reducing the data space, both separately as we have performed by using co-clustering and CA. Alternatively, Kohonen's self-organizing map (SOM) [KOH 82] enables clustering and mapping of clusters while providing one final single map. The SOM algorithm is not derived exactly through the optimization of an objective function, and some crucial parameters need to be chosen empirically. A probabilistic model for SOM is appealing for several reasons, the principal reason being that a parametric model is flexible and scalable when defined appropriately. Generative topographic mapping (GTM) [BIS 98] is a parametric SOM that offers a number of advantages compared to the standard SOM. Recently, in [PRI 12], the authors presented a GTM based on the Poisson latent block mixture model. The empirical results obtained showed that the method proposed is able to present a quick summary of the data set contents. In addition, the proposed model is parsimonious when compared to the existing alternative in the domain.