NONLINEAR MAPPING BY CONSTRAINED CO-CLUSTERING

Rodolphe Priam[1], Mohamed Nadif[2] and Gérard Govaert[3]
[1]*S3RI, University of Southampton, University Road, SO17 1BJ, Southampton, U.K.*
[2]*LIPADE, Université Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France*
[3]*HEUDIASYC, Université de Technologie de Compiègne, 60205 Compiègne, France*

Keywords: Generative topographic mapping, Latent block model, Contingency table, Clustering.

Abstract: The latent block model is an efficient alternative to the mixture model for modelling a dataset when the number of rows or columns of the data matrix studied is large. For analyzing and reducing the spaces of a matrix, the methods proposed in the litterature are most of the time with their foundation in a non-parametric or a mixture model approach. We present an embedding of the projection of co-occurrence tables in the Poisson latent block mixture model. Our approach leads to an efficient way to cluster and reduce this kind of data matrices.

1 INTRODUCTION

Contingency tables, or co-occurrence matrices, are found in diverse domains. In these matrices each cell is a cross-product of two categorical variables I (n categories) and J (d categories.) The cells contain the number of occurrences for the corresponding cross-categories. Contingency tables appear in information retrieval (Deerwester et al., 1990) and document clustering (Hofmann, 1999), where I may correspond to a corpus of documents, J to a set of words, and so the frequency denotes the number of occurrences of a word in a document. Other examples from data mining, preference analysis, etc., show that analyzing contingency tables is in fact a very common and fundamental aspect of data analysis. Contingency tables are usually analyzed using one of the many categorical data analysis methods available in the literature.

When the data matrix is large, a clustering can give a quicker and easier access to the data content than a method for reducing the dimensionality of the features. Combining clustering and reduction for mapping clusters rather than rows or columns is therefore an interesting requirement for data analysis. One way to fulfill this purpose is by showing the clusters on a map after clustering the data by an ad'hoc algorithm and reducing the feature space, both separately. Alternatively, the Kohonen's self-organizing map (SOM) (Kohonen, 1997) is such that the clustering and the mapping of the clusters take place simultaneously while providing one final unique map. The SOM algorithm is not derived exactly through the optimization of an objective function, and several parameters have to to be set empirically.

A probabilistic model for SOM is appealing for several reasons, the principal one is that a parametric model is flexible and scalable when defined properly. We are interested in proposing an efficient parametric model for a bidimensional mapping of the clusters of I for a contingency table. Generative Topographic Mapping (GTM) (Bishop et al., 1998) is a parametric SOM with a number of advantages compared to the standard SOM. It re-formulates SOM by embedding the constraints of vicinity for the clusters in a Gaussian mixture model (GMM) (McLachlan and Peel, 2000). Classical mixture models, and in particular GMM, are generally not efficient for large datasets, and this is also true of GTM. One possible alternative to a clustering of rows or columns is a co-clustering approach that clusters the two dimensions of a matrix simultaneously and efficiently, with a competitive small number of parameters.

Here we turn to the latent block model (see (Govaert and Nadif, 2003)) with constraints in order to simultaneously cluster and visualize the clusters. In contrast to previous works like for instance (Kabán and Girolami, 2001), (Hofmann, 2000) or (Kában, 2005), the proposed method is parsimonious since the number of parameters remains constant when the size of the data matrix increases.

The paper is organized as follows. In Section 2 we review co-clustering for co-occurrence tables, and describe a Poisson latent block model (PLBM) (Govaert and Nadif, 2010). We add constraints in the model

Priam R., Nadif M. and Govaert G. (2012).
NONLINEAR MAPPING BY CONSTRAINED CO-CLUSTERING.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 63-68
DOI: 10.5220/0003764800630068

and propose an algorithm for the estimation of the parameters. In section 3 we present an evaluation of our new method named BlockGTM. Finally, the conclusion summarizes the advantages of our contribution.

2 BLOCK EM MAPPING

In latent block model (LBM) the $n \times d$ random variables generating the observed x_{ij} cells of the data matrix are assumed to be independent, once $\mathbf{z}$ and $\mathbf{w}$ are fixed where the set of all possible assignments $\mathbf{w}$ of J (resp. $\mathbf{z}$ of I) is denoted $\mathcal{W}$ (resp. $\mathcal{Z}$). The data matrix $\mathbf{x}$ is therefore a set of cells:

$$(x_{11}, x_{12}, \ldots, x_{ij}, \ldots, x_{nd}),$$

rather than the sample of d-dimensional vectors in the more classical mixture setting. The two sets of possible assignments $\mathbf{w}$ and $\mathbf{z}$ cluster the cells of the matrix $\mathbf{x}$ into a number of contiguous, non-overlapping blocks. A block $k\ell$ is defined as the set of cells $\{x_{ij}; z_i = k, w_j = \ell\}$. The binary classification matrix $\mathbf{z} = (z_{ik})_{n\times g}$ is such that $\sum_{k=1}^{g} z_{ik} = 1$ and $z_{ik} = 1$ indicates the component of the row i, and similarly for the columns with $\mathbf{w} = (w_{j\ell})_{d\times m}$.

The following decomposition is obtained (Govaert and Nadif, 2003) by independence of $\mathbf{z}$ and $\mathbf{w}$, by summing over all the assignments:

$$\begin{aligned} f_{LBM}(\mathbf{x};\theta) &= \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} p_k^{z_{ik}} \prod_{j,\ell} q_\ell^{w_{j\ell}} \\ &\times \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}, \end{aligned}$$

where $\varphi(.;\alpha_{k\ell})$ is a density function defined on the set of reals $\mathbb{R}$ and $\{\alpha_{k\ell}\}$ are unknown parameters. The vectors of the probabilities p_k and q_ℓ that a row and a column belong to the k-th component and to the ℓ-th component are respectively denoted $\mathbf{p} = (p_1, \ldots, p_g)$ and $\mathbf{q} = (q_1, \ldots, q_m)$. The set of parameters is denoted θ and is compound of $\mathbf{p}$ and $\mathbf{q}$ plus α which aggregates all the $\alpha_{k\ell}$. Hereafter, to simplify the notation, the sums and the products relating to rows, columns or clusters will be subscripted respectively by the letters i, j or k, ℓ without indicating the limits of variation, which are implicit. Next, PLBM is described for contincency tables and the constraints are added.

2.1 Poisson Latent Block Model

For co-occurrence tables, PLBM assumes that the observed values x_{ij} in a block $k\ell$ are drawn from a Poisson distribution $P(\lambda_{k\ell}^{ij})$ with parameter $\lambda_{k\ell}^{ij} = \mu_i \nu_j \alpha_{k\ell}$ where the effects $\mu = (\mu_1, \ldots, \mu_n)$ and $\nu = (\nu_1, \ldots, \nu_d)$ are assumed equal to the margin totals $\{\mu_i = \sum_j x_{ij}; 1 \leq i \leq n\}$ and $\{\nu_j = \sum_i x_{ij}; 1 \leq j \leq d\}$. Then φ for the block $k\ell$ is defined as follows:

$$\varphi(x_{ij};\mu_i,\nu_j,\alpha_{k\ell}) = \frac{exp(-\mu_i\nu_j\alpha_{k\ell})(\mu_i\nu_j\alpha_{k\ell})^{x_{ij}}}{x_{ij}!}.$$

Given that $x_{ij} \in \mathbb{N}_+$, the unknown parameter $\alpha_{k\ell}$ of the block $k\ell$ is in $[0;1]$, since $x_{ij} < \mu_i\nu_j$. The set of parameters θ of the model can be estimated by maximizing the log-likelihood:

$$L(\mathbf{x};\theta) = \log f_{LBM}(\mathbf{x};\theta).$$

2.2 Constrained Parameters

To induce a quantization with a large number of clusters, the probabilities p_k and q_ℓ are fixed and equiproportional such that $\{p_k = 1/g; 1 \leq k \leq g\}$ and $\{q_\ell = 1/m; 1 \leq \ell \leq m\}$. The parameters of the Poisson LBM are parameterized with the fixed vectors $\{\xi_k\}$ defined hereafter for the mapping of I, and the unknown vectors $\{w_\ell \in \mathbb{R}^h, 1 \leq \ell \leq m, h \in \mathbb{N}_+^*\}$ because $\alpha_{k\ell}$ is dependent on the index k and ℓ. The parameters $\{w_\ell\}$ are estimated by maximum likelihood. For defining the vectors $\{\xi_k\}$, it is considered the bidimensional coordinates:

$$\mathcal{S} = \{s_k = (s_{k1}, s_{k2}); k = 1, ..., g\},$$

from the nodes of a regular mesh discretizing the square where the data are projected $[-1;1] \times [-1;1]$. $\mathcal{S}$ is similar to the set of nodes of SOM. As in GTM, each coordinate s_k is nonlinearly transformed by h basis functions ϕ such as:

$$\xi_k = \Phi(s_k) = (\phi_1(s_k), \phi_2(s_k), \ldots, \phi_h(s_k))^T,$$

where each basis function ϕ is a kernel-like function:

$$\phi(s_k) \propto exp[-||s_k - \mu_\phi||^2 / 2\nu_\phi^2],$$

with a mean center $\mu_\phi \in \mathbb{R}^2$ and a standard deviation ν_ϕ. It is then considered the inner products:

$$\{w_\ell^T \xi_k; 1 \leq k \leq g, 1 \leq \ell \leq m\}.$$

To map the inner product ($w_\ell^T \xi_k \in \mathbb{R}$) onto its corresponding parameter ($\alpha_{k\ell} \in [0;1]$) it is used a sigmoidal function $\sigma(.)$ as in (Girolami, 2001) such that for $1 \leq k \leq g, 1 \leq \ell \leq m$, we have:

$$\alpha_{k\ell} = \sigma(w_\ell^T \xi_k) = \frac{exp(w_\ell^T \xi_k)}{1 + exp(w_\ell^T \xi_k)}.$$

The relative ordering of the coordinates $\{s_k = (s_{k1}, s_{k2}); k = 1, ..., g\}$ remains, at least locally, after the transformation. The reduced $g \times m$ matrix α in PLBM is replaced by an $h \times m$ matrix:

$$\Omega = [w_1 | w_2 | \cdots | w_m].$$

The model remains parsimonious because h is small, less than half of one hundred in practice.

Below we present an algorithm for the estimation of the parameters $\theta = \Omega$, the matrix for the constraints. The optimization problem is slightly different from the unconstrained case, as we shall explain in the next section.

2.3 Parameter Estimates

For the proposed model with the introduced constraints, we aim to address the problem of parameters estimation by a maximum likelihood (ML) approach such that:

$$\hat{\theta} = argmax_{\theta} L(\mathbf{x};\theta).$$

For finding a suitable value of θ for the constrained PLBM, the Block EM (BEM) (Govaert and Nadif, 2005) results in the following criterion (denoted $\tilde{Q}$ for short) which is maximized iteratively:

$$\begin{aligned} &\tilde{Q}_{BlockGTM}(\theta,\theta^{(t)}) \\ &= \sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \log\varphi(x_{ij};\alpha_{k\ell}) \\ &= \sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \left\{x_{ij}\log\alpha_{k\ell} - \mu_i\nu_j\alpha_{k\ell}\right\} + cte \\ &= \sum_{k,\ell} y_{k\ell}^{(t)} \log\alpha_{k\ell} - \mu_k^{(t)}\nu_\ell^{(t)}\alpha_{k\ell}) + cte. \qquad (1) \end{aligned}$$

Here cte is a constant independent of the parameters, the index (t) permits to denote a current estimation of a parameter or a function of the parameters. It is also denoted $y_{k\ell}^{(t)} = \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} x_{ij}$, $\mu_k^{(t)} = \sum_i c_{ik}^{(t)}\mu_i$, and $\nu_\ell^{(t)} = \sum_j d_{j\ell}^{(t)}\nu_j$, while given $\theta^{(t)}$, the quantities $c_{ik}^{(t)}$ (resp. $d_{j\ell}^{(t)}$) are the posterior probabilities that a row (resp. a column) belongs to the block $k\ell$. Here, the posterior probabilities are estimated by using the dependent equations:

$$c_{ik}^{(t)} \propto \exp\left(\sum_{j\ell} d_{j\ell}^{(t)} \log\varphi(x_{ij};\alpha_{k\ell}^{(t)})\right), \qquad (2)$$

$$d_{j\ell}^{(t)} \propto \exp\left(\sum_{ik} c_{ik}^{(t)} \log\varphi(x_{ij};\alpha_{k\ell}^{(t)})\right). \qquad (3)$$

At the ML, they are denoted $\{\hat{c}_{ik}\}$ and $\{\hat{d}_{j\ell}\}$. The parameters are estimated in an iterative way. The BEM algorithm proceeds by an alternated maximization of $\tilde{Q}$. At each iteration the posterior probabilities $\{c_{ik}\}$ and $\{d_{j\ell}\}$ are evaluated for all rows and all columns, and just after the maximization of the function $\tilde{Q}$ is obtained with respect to the parameters. As a remark, this induces a maximization with a variational approximation at each iteration. Another approximation of the resulting criterion $\tilde{Q}$ is also required at the maximization step as explained in the following paragraphs.

2.4 Algorithm

The algorithm for maximizing $\tilde{Q}$ proceeds iteratively by increasing an approximation of the log-likelihood at each step. At the Maximization step we estimate the next current value for $\theta^{(t+1)}$ by:

$$\theta^{(t+1)} = argmax_{\theta}\ \tilde{Q}(\theta|\theta^{(t)}).$$

Let us have ε a small positive real. The algorithm for finding the maximum likelihood solution is given in Figure 1 (see Appendix for $\tilde{Q}_\ell^{(t)}$ and $\tilde{H}_\ell^{(t)}$).

```
Learning algorithm for BlockGTM:

 - Initialization:
   Initialize $\{c_{ik}^{(0)}\}$, $\{d_{j\ell}^{(0)}\}$ and $\Omega^{(0)}$.
 - E-Step:
   Compute $\{c_{ik}^{(t)}\}$ by (2), and $\{d_{j\ell}^{(t)}\}$ by (3) in a loop.
 - M-Step:
   Compute the new parameters for $\ell = 1,\ldots,m$
       $w_\ell^{(t+1)} = w_\ell^{(t)} + \left[\tilde{H}_\ell^{(t)}\right]^{-1} \nabla\tilde{Q}_\ell^{(t)},$     (4)
   with $\nabla\tilde{Q}_\ell^{(t)}$ by (5) and $\tilde{H}_\ell^{(t)}$ by (6).
 - End:
   If $|\Omega^{(t+1)} - \Omega^{(t)}| < \varepsilon$ then stop else return E-Step.
```

Figure 1: Iterations for *BlockGTM*.

Next, we evaluate the performance of BlockGTM for several real datasets.

3 NUMERICAL EXPERIMENTS

In order to test the proposed method, we construct the bidimensional projections of the obtained clusters by the proposed method for four textual datasets.

3.1 Bi-dimensional Mapping

The set of bidimensional coordinates $\mathcal{S}$ for the g clusters are used to find the final projection $\hat{s}_i$ of each category i in the latent space. When a row i has a higher posterior probability $\hat{c}_{ik}$ for a cluster k then it belongs to this cluster and the label for the i-th row is estimated by $\hat{z}_i = k$. This same row can then be represented at the bidimensional coordinates $\hat{s}_i^{MAP} = s_{\hat{z}_i} = (s_{\hat{z}_i 1}, s_{\hat{z}_i 2})^T$. By performing this procedure for each row i, the model builds a reduced view of the n categories of I. Moreover, when two nodes have their coordinates s_k and $s_{k'}$ near in the latent space, their corresponding clusters should have similar parameters $\alpha_{k\ell}$ and $\alpha_{k'\ell}$, so their corresponding contents should be also similar. A fuzzy projection can be obtained by computing an average position of each row i from its posterior probabilities $\hat{c}_{ik}$. This is written $\hat{s}_i = \sum_{k=1}^{g} \hat{c}_{ik} (s_{k1}, s_{k2})^T$. If the vector of probabilities $(\hat{c}_{i1}, \hat{c}_{i2}, \cdots, \hat{c}_{ig})$ is binary, then the row i is in the cluster $\hat{z}_i$, and $\hat{s}_i = \hat{s}_i^{MAP}$. This is generally the case for GTM for a large part of the dataset. In the experimental part, it is constructed only a tabular view after binarizing these vectors of probabilities, except a small illustrative example.

3.2 Datasets

The characteristics of the four real datasets are described below.

- *N4*. This dataset is composed of 400 documents selected from a textual corpus of 20000 usenet posts from 20 original newsgroups. From each group among the 4 retained, 100 posts are selected and 100 terms are filtered by mutual information (Kabán and Girolami, 2001).
- *Binary*$_1$. This dataset in (Slonim et al., 2000) consists of 500 posts separated into two clusters for the newsgroups *talk.politics.mideast* and *talk.politics.misc*. A preprocessing was carried out by the authors to reduce the number of words by ignoring all file headers, stop words and numeric characters. Moreover, using the mutual information, the top 2000 words were selected.
- $Multi5_1$. This dataset in (Slonim et al., 2000), consists of 500 posts separated into five clusters *comp.graphics*, *sci.space*, *rec.motorcycles*, *res.sports.baseball* and *talk.politics.mideast*. The same pre-processing than for *Binary*$_1$ was performed.
- *C3*. This dataset in (Dhillon et al., 2003), also known as *Classic3*, is often used as a benchmark for co-clustering. This dataset is a contingency table of size 3891 x 4303 and it is compound of three classes denoted *Medline*, *Cisi* and *Cranfield* as in the larger complete data sample not considered here.

These four datasets studied in our experiments are of increasing size.

3.3 Results

Table 1 summarizes the characteristics of the datasets and the parameters for BlockGTM. The four constructed maps are squares of size $g = 9 \times 9$ for the clustering in rows, while the number of clusters m in columns and the dimension h were chosen after a few tries. Each map is represented as following. For each k-th cluster, a barplot corresponding to the true labels of the data in the cluster is constructed at position s_k after fitting the model. The results are shown in Figure 2 for $Multi5_1$ and *C3*. So, for a given dataset the map shows a matrix of 9×9 barplots such as if two nodes are close on the latent space they should have similar barplots. This is a tabular view of the data (categories I) which confirms also that the nearest clusters have their texts with similar topics as expected.

Table 1: Summary where $n \times d$ is the size of the contingency table, m is the number of clusters in columns, h is the number of basis functions, and E_{r1} (resp. E_{r2}) is the accuracy in percent from BlockGTM (resp. PLBM).

Data	n	d	m	h	E_{r1}	E_{r2}
N4	400	100	10	12	96.5	93.4
Binary$_1$	400	100	10	19	91.2	92.4
$Multi5_1$	500	2000	20	19	90.6	89.0
C3	3891	4303	20	28	99.1	99.3

In this section we are interested on measuring how well the co-clustering can reveal the inherent structure of a given textual dataset. We consider the accuracy which is usually derived from the confusion matrix or the cluster purity. Specifically, we measure the quality of the clustering for the obtained clusters comparatively to the real categories of the documents. The columns E_{r1} and E_{r2} of Table 1 give, in percent, the accuracy obtained respectively for BlockGTM and PLBM initialized with the final parameters of BlockGTM.

- For *N4*, the categories of I are projected by the Correspondence Analysis (CA) method (Benzecri, 1980). The coordinates from CA are used to compute the positions of the mean centers in a 3-dimensional space thanks to the quantities $\hat{c}_{ik}$. Figure 3 shows the result. It is interesting to note that the original mesh compound of the nodes $\mathcal{S}$ in the latent space is easily recognized in this 3-dimensional space. Here each class is quantized by a subset of clusters from the map, and the subset usually includes only data with their corresponding projections close in the space of projection as expected.
- For *C3*, the proposed method extracts the origi-

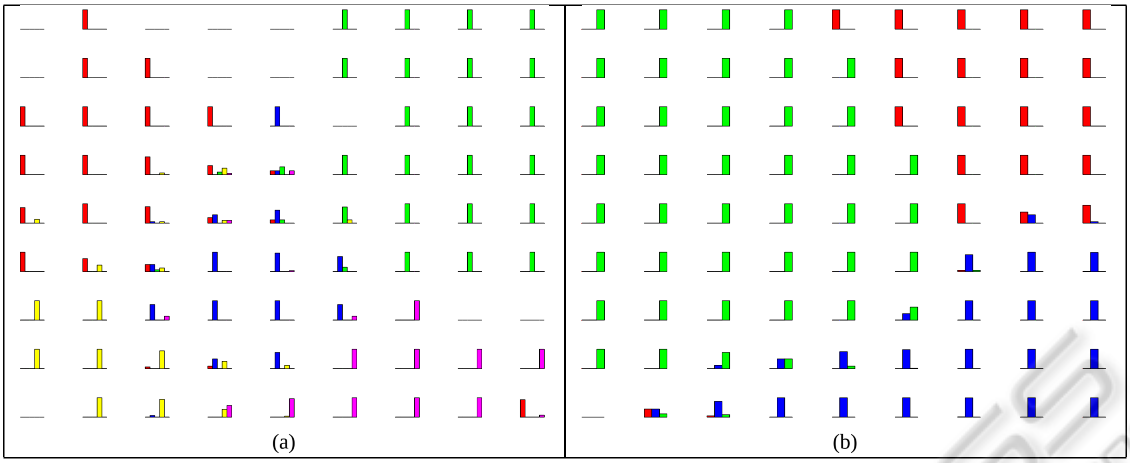

Figure 2: The result from the proposed method for the datasets (a) $Multi5_1$, and (b) *C3*.

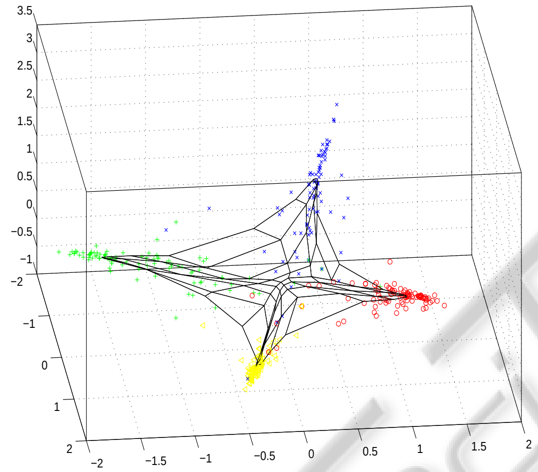

Figure 3: A result from BlockGTM for *N4* in the 3-dimensional space of the projection with the 3 first principle factorial directions of CA.

nal clusters almost correctly. The accuracy of the method is $1-33/3891 = 99.15\%$, while the co-clustering based on (Dhillon et al., 2003) has an accuracy of $97.74 = 1-88/3891$ so the obtained error is smaller. The macro-clustering comes after a finer clustering and is better able to separate the different classes.

- For $Binary_1$ and $Multi5_1$, Table 2 reports the resulting accuracies for BlockGTM, PLBM, IB_{double} (Slonim et al., 2000), and IDC-15 (El-yaniv and Souroujon, 2001). This helps for the comparison between the different results. Despite a slightly different error rate, the proposed method is able to map the whole datasets and separate the natural classes. The main difference with the alternative approaches is not only the efficient co-clustering, but also the capacity to provide a quick visual overview of the proximities between the clusters.

Table 2: Accuracy for BlockGTM, PLBM, IB_{double}, IDC.

	BlockGTM	PLBM	IB_{double}	IDC-15
$Binary_1$	91.2	92.4	70	85
$Multi5_1$	90.6	89.0	59	86

4 CONCLUSIONS

We have proposed an embedding of the projection of co-occurrence tables in the Poisson latent block mixture model. The presented model is parsimonious when compared to the existing alternatives in the domain. The empirical results obtained show that BlockGTM is able to present a quick summary of the dataset contents. So the approach is interesting for data analysis of large contingency tables.

ACKNOWLEDGEMENTS

This research was supported by the CLasSel ANR project ANR-08-EMER-002.

REFERENCES

Benzecri, J. P. (1980). *L'analyse des données tome 1 et 2 : l'analyse des correspondances*. Dunod.

Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). Developpements of generative topographic mapping. *Neurocomputing*, 21:203–224.

Böhning, D. and Lindsay, B. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41(6):391-407*.

Dhillon, I. S., Mallela, S., and Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pages 89–98.

El-yaniv, R. and Souroujon, O. (2001). Iterative double clustering for unsupervised and semi-supervised learning. In *In Advances in Neural Information Processing Systems (NIPS*, pages 121–132.

Girolami, M. (2001). The topographic organization and visualization of binary data using multivariate-bernoulli latent variable models. *IEEE Transactions on Neural Networks*, 20(6):1367–1374.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473.

Govaert, G. and Nadif, M. (2005). An EM algorithm for the block mixture model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(4):643–647.

Govaert, G. and Nadif, M. (2010). Latent block model for contingency table. *Communications in Statistics-theory and Methods*, 39:416–425.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *SIGIR'99*, pages 50–57.

Hofmann, T. (2000). Probmap - a probabilistic approach for mapping large document collections. *Intell. Data Anal.*, 4(2):149–164.

Kabán, A. (2005). A scalable generative topographic mapping for sparse data sequences. In *ITCC '05: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I*, pages 51–56, Washington, DC, USA. IEEE Computer Society.

Kabán, A. and Girolami, M. (2001). A combined latent class and trait model for analysis and visualisation of discrete data. *IEEE Trans. Pattern Anal. and Mach. Intell.*, pages 859–872.

Kohonen, T. (1997). *Self-organizing maps*. Springer.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York.

Slonim, N., Tishby, N., and Y, Y. I. (2000). Document clustering using word clusters via the information bottleneck method. In *In ACM SIGIR 2000*, pages 208–215. ACM press.

APPENDIX

We have to find a zero for the $\tilde{Q}$ function such that for all ℓ, we have $\frac{\partial\tilde{Q}(\theta|\theta^{(t)})}{\partial w_\ell}\Big|_{w_\ell^{(t+1)}} = 0$. Considering the usual Newton-Raphson algorithm for the proposed model, the maximizing step w.r. w_ℓ is written as in Formula (4).

Then, keeping only the sum on k for ℓ constant, the score for the criterion maximized at the t-th iteration with respect to w_ℓ can be written:

$$\begin{aligned}
&\nabla\tilde{Q}_\ell^{(t)}\\
&=\frac{\partial\tilde{Q}_{BlockGTM}(\theta,\theta^{(t)})}{\partial w_\ell}\\
&=\sum_k\left\{y_{k\ell}^{(t)}\frac{\partial\log\alpha_{k\ell}}{\partial w_\ell}-\mu_k^{(t)}\nu_\ell^{(t)}\frac{\partial\alpha_{k\ell}}{\partial w_\ell}\right\}\\
&=\sum_k\left\{y_{k\ell}^{(t)}(1-\alpha_{k\ell})\xi_k-\mu_k^{(t)}\nu_\ell^{(t)}\alpha_{k\ell}(1-\alpha_{k\ell})\xi_k\right\}\\
&=\sum_k(1-\alpha_{k\ell})\left\{y_{k\ell}^{(t)}-\mu_k^{(t)}\nu_\ell^{(t)}\alpha_{k\ell}\right\}\xi_k\\
&=\Phi^T(I_g-A_\ell)\left[y_\ell-\nu_\ell^{(t)}A_\ell\mu\right] \qquad (5)
\end{aligned}$$

where we denote $\mu=(\mu_1,\mu_2,\cdots,\mu_g)^T$ and the diagonal matrix $A_\ell=diag_{1\leq k\leq g}(\alpha_{k\ell})$ at step (t).

Similarly, the second-order derivative of the criterion gives the Hessian matrix which is written:

$$\begin{aligned}
&\nabla^2\tilde{Q}_\ell\\
&=\frac{\partial\tilde{Q}_{BlockGTM}(\theta,\theta^{(t)})}{\partial w_\ell^T\partial w_\ell}\\
&=-\sum_k(1-\alpha_{k\ell})\alpha_{k\ell}\left\{y_{k\ell}+(1-2\alpha_{k\ell})\mu_k^{(t)}\nu_\ell^{(t)}\right\}\xi_k^T\xi_k\\
&=-\Phi^TA_\ell(I_g-A_\ell)\left\{Y_\ell+\nu_\ell(I_g-2A_\ell)M_\mu\right\}\Phi,
\end{aligned}$$

where we have at step (t), $Y_\ell=diag_{1\leq k\leq g}(y_{k\ell})$ and $M_\mu=diag_{1\leq k\leq g}(\mu_k)$.

A lower bound of this matrix is proposed in order to improve the maximization step. This symmetric matrix $\tilde{H}_\ell$ is strictly negative-definite for all parameters $\{\alpha_{k\ell},1\leq k\leq g\}$ remaining in $]0;1[$, and satisfies the inequality:

$$\begin{aligned}
\tilde{H}_\ell &= -\Phi^TA_\ell(I_g-A_\ell)\left\{Y_\ell+\nu_\ell M_\mu\right\}\Phi \qquad (6)\\
&\leq \nabla^2\tilde{Q}_\ell.
\end{aligned}$$

This new matrix is able to increase the criterion to be maximized (see (Böhning and Lindsay, 1988)) in the Newton-Raphson algorithm, while providing a more stable learning behavior than the original Hessian matrix, so this solution has been preferred in the experiments.