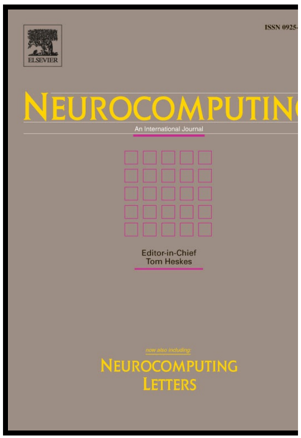


www.elsevier.com

Generalized topographic block model

Rodolphe Priam, Nadif Mohamed, Gérard Govaert

PII: S0925-2312(15)01279-5
DOI: http://dx.doi.org/10.1016/j.neucom.2015.04.115
Reference: NEUCOM16039

To appear in: *Neurocomputing*

Received date: 6 July 2014
Revised date: 24 February 2015
Accepted date: 8 April 2015

Cite this article as: Rodolphe Priam, Nadif Mohamed and Gérard Govaert Generalized topographic block model, *Neurocomputing* http://dx.doi.org/10.1016/j.neucom.2015.04.115

Elsevier Editorial System(tm) for Neurocomputing
Manuscript Draft

Manuscript Number: NEUCOM-D-14-01310R3

Title: Generalized topographic block model

Article Type: SI:Visual Big Data

Keywords: Latent Block Mixture Model, Exponential Family, Generative Topographic Mapping, Block expectation-maximization, Visualisation.

Corresponding Author: Dr. Rodolphe Priam,

Corresponding Author's Institution: University of Southampton

First Author: Rodolphe Priam

Order of Authors: Rodolphe Priam; Mohamed Nadif; Gerard Govaert

Abstract: Co-clustering leads to parsimony in data visualisation with a number of parameters dramatically reduced in comparison to the dimensions of the data sample. Herein, we propose a new generalized approach for nonlinear mapping by a re-parameterization of the latent block mixture model. The densities modeling the blocks are in an exponential family such that the Gaussian, Bernoulli and Poisson laws are particular cases. The inference of the parameters is derived from the block expectation-maximization algorithm with a Newton-Raphson procedure at the maximization step. Empirical experiments with textual data validate the interest of our generalized model.

The authors have very carefully revised the document by following every comments from the editors and the two anonymous reviewers. It has been tried every possible effort to solve each remark that has been addressed. Below a detailed summary of the updates is provided. We would like to thank the editors and the reviewers for their constructive comments on this manuscript and positive support.

To Editors:

Once more, we would very much like to invite you to revise your paper, seriously taking into account the comments of the reviewers, and to resubmit your revised version by 02/25/2015 (mm/dd/yy). Any revision received after that may be treated as a new submission.

Authors' response:
The paper has been revised according to the comments and suggestions of reviewer 2.

2) To Reviewer #1:

The revised manuscript is sufficient to Neurocomputing publication standards, and I suggest accepting this manuscript.

Authors' response:
Thanks for the positive comments and the opportunity of publishing the document in Neurocomputing.

3) To Reviewer #2:

Q1. I thank the authors for the revised version of their manuscript.

Authors' response:
Thanks for the positive comments.

Q2. They open the abstract with the statement: "Parametric methods for data visualisation are most of the time founded on an usual mixture model framework."
Even a light-hearted revision of existing parametric methods for multivariate data visualization (See, for instance, Lee & Verleysen, 2007) would reveal that this is
not the case. Therefore, I think this statement should be either removed or revised.

Authors' response:
Thanks for this suggestion. Indeed, the term « parametric methods » was meaning « probabilistic methods » or « parametric model » in a statistical framework and could have been read as any methods with parameters on the contrary to svd for instance. This sentence has been removed, and the summary updated for complying with other comments in the review.

Q3. Co(Bi)-clustering in general and co(bi)-clustering with visualization-oriented self-organizing models are more adequately introduced in the new version.

Authors' response:
Thanks for this remark.

Q4. I am a bit puzzled by the new introduction "storyline", though. It roughly goes like this:

a) - Co-clustering was first proposed in the seventies and some more works [6-11], reviewed in [12-13], have been produced. All this references are prior to 2004.
b) - Only [11] (2004) deals with SOM for visualization (and one of the strong points of the reviewed manuscript is precisely visualization).
c) - GTM is a good alternative to SOM (correct)
d) - co-clustering with SOM has been used, mostly in bio- areas [23-28]
e) - authors propose a more principled co-clustering model based on GTM
f) If this is the storyline, the second point and the last two are at odds with each other, because some of the works in [23-28] (all of them post-2004) do indeed deal with data visualization. This running theme (co-clustering/SOM>M/Visualization) should be more consistently told.

Authors' response:
Thanks for this concern. Indeed, after a reconsideration of the references, it is really better to introduce co-clustering at first without visualization (see a) and then introduce its combination with visualisation (see b).
It is now clear in the text that many methods exist with an hybridization from a co-clustering framework (see a) and a SOM, not only in [11] (see b). These combinations do not exist only in bioinformatics (see d).
Note that BGTM is not exactly a new principle co-clustering model based on GTM (see e) but a variant of GTM with a clustering of the columns.
Hence, the story line has been revised for more clarity as follows in the introduction:
- First paragraph, the former co-clustering methods are briefly listed and justified for a reduction of the dimension for large number of columns.
- Second paragraph, the combination of a co-clustering with a SOM is presented.
- Third paragraph, SOM is presented with its probabilistic variant, GTM.
- Fourth paragraph, the proposal by a probabilistic method by extending LBM to data visualisation. This can be seen as a principle method indeed because the model takes benefice of the properties of the statistical tools.
- Fifth paragraph, the plan of the paper.

Q5. Also in the introduction, you state that "A parametric model is flexible and scalable when it is defined properly." Well, the concept of "properly defined" is quite
vague, to say the least. I reckon that Kohonen and colleagues would not be too happy with this comment. I would feel far more confortable with something along the
lines of "A parametric model is more flexible and reliable when it is defined according to sound statistical principles".

Authors' response:
Thanks for this suggestion. The idea was only to comment on probabilistic models and also to justify co-clustering model in comparison to clustering model because in certain cases indeed scalability is obtained jointly with parsimony. But the sentence has been removed and only the term parsimony remains in the text.

Q6. I am afraid that it is also unclear what does scalability have to do this. In what sense a model is more scalable when it is "properly defined"?

Authors' response:
Thanks for this concern. The idea was once again to recall the possible scalability of co-clustering in comparison with row clustering. The sentence have been removed.

Q7. Authors go on to say that "the Generative Topographic Mapping (GTM) [18] is a parametric SOM with a set of possible values for its parameters which are more restricted than Kohonen's maps." I truly ignore what do the authors mean by "a set of possible values for its parameters which are more restricted than Kohonen's maps" Why are they more restricted? In what sense? Do they mean that the values of some of the GTM parameters can be adaptively estimated and, therefore, there is no need of trial-and-error choices for the values of at least these parameters?

Authors' response:

Thanks for this remark. Indeed, it seems less needed of performing trial-and-error choices in GTM because the width of the neirboorhood function is not explicit, while in SOM, this is a function which needs to be chosen. The sentence has been removed and the new one is as follows :

« In GTM the auto-organization of the clusters is directly induced by the parameterization. The algorithm of Kohonen's map is re-formulated by embedding the auto-organization process at the level of the means of a Gaussian mixture model. »

Hence, in this new version, it is not longer possible to see any blow against the Kohonen's map, but just a short recall about the differences of the modeling. Note that trial-and-error choices is not a problem per se.

Q8. Another unfair blow against SOM is contained in the following comment: "Self-organizing maps and co-clustering have been combined and illustrated with bioinformatics or biological data in several methods [23, 24, 25, 26, 27], but they are non generative and use an Euclidian distance which may not be relevant in certain cases."
Well, being generative certainly endows GTM with a number of advantages over SOM, but both SOM and GTM have been re-defined in existing literature to make use of non-Euclidean metrics, therefore, I do not see how this argument can be used as a disadvantage of SOM for co-clustering.

Authors' response:

Thanks for this suggestion. This was not a justification of the paper, the paper deals with LBM and how to do visualisation with LBM via SOM. There was no any blow, just a short remark but the sentence has been removed.

Q9. A more serious concern is that, even if the overall structure and style of the text has admittedly improved, it is still very difficult to separate in this paper what is novel contribution and what is existing work.

Authors' response:

Thanks for this concern. A new sentence has been inserted in the introduction part :

« If the previous models of block generative topographic mapping have been proposed for only one particular distribution for the blocks, the new generalization is able to provide a unified framework for the visualisation of data block matrices and can help for implementing and comparing alternative distributions in future. »

And the summary is more explicit :

« A co-clustering leads to parsimony in data visualisation with a number of parameters dramatically reduced in comparison with the dimensions of the data sample. Herein, we propose a new generalized approach for nonlinear mapping by a re-parameterization of the latent block mixture model. The approach is related to probabilistic factorisation where the two dimensions of the matrix are clustered but one of the two sets of latent vectors is fixed. The densities modeling the blocks are in an exponential family such that the Gaussian, Bernoulli and Poisson laws are particular cases. The inference of the parameters is derived from the block expectation-maximization algorithm with a Newton-Raphson procedure at the maximization step. Empirical experiments with textual data validate the interest of our generalized model. »

Q10. This worries me in particular because, despite onerous self-citation, authors fail to cite some of their own recent work such as: Pattern Anal Applic (2014) 17:839-847, doi:10.1007/s10044-014-0368-8 which has an obvious relation with the current manuscript.

Authors' response:
Thanks for this suggestion. The reference has been added.

Q11. It has also come to my attention recent work by Sarazin, Lebbah, Azzag and Chaibi:
Sarazin, T., Lebbah, M., Azzag, H., & Chaibi, A. (2014, January). Feature Group Weighting and Topological Biclustering. In Neural Information Processing, pp.369-376, Springer
Chaibi, A., Lebbah, M. and Azzag, H. A new bi-clustering approach using topological maps. The 2013 International Joint Conference on Neural Networks (IJCNN),IEEE,2013.
.. that the authors must now account for to put their own work in perspective.

Authors' response:
Thanks for this suggestion. The two references have been added. Done for perspective with further comparisons in future.

Q12. Concerning the experiments, I would like to see an explicit expression of what the authors call "the usual classification error rate, denoted error-rate, [...] obtained from the estimated labels ..." It is very unclear from the authors' description.

Authors' response:
Thanks for this concern. Done.

Q13. Some of the settings are also unclear: in the last paragraph of p.14, what is m=20? Also, you say that the number of bfs. is h=28. These bfs. are usually set as a square grid, so, where does the value of 28 come from? According to what rationale were the values of g, m and h chosen? Were they compared with others and no significant changes were found? (a similar thing can be said about the experiment reported in section 5.3)

Authors' response:
Thanks for this suggestion. Three new sentences have been added.

« A visual inspection of the final map and the values of the indicators lead to an empirical choice of the parameters g, m, and h in BGTM for the results presented hereafter with three datasets. »

« The frequencies are meaningful in clustering textual data. If this is clearly true for *N4*, in the case of *C3-s* outliers seem to perturbate the empirical result obtained from only one sample and a more robust model might be preferred. »

And a sentence at the sub-section 4.1 for describing further the vectors of bfs. These three terms have been introduced in other references in the litterature like .

Q14. I am not convinced with the comparison experiments: you just compare PBGTM with BBGTM and GBGTM. Is that fair? wouldn't each of those models be more suited to certain types of data that to others (thus explaining the advantages of the Poisson model)? Why not compare with a SOM-based co-clustering technique?

Authors' response:
Thanks for this concern. This was mostly motivated with the presented large dataset which is difficult to be handled by usual approaches because of the large number of variables and also the non existence of available implementations for the existing models in the litterature.
Anyway, a new method called BCASOM is proposed by restricting the parameters of CASOM, a multinomial SOM. The column clustering is fixed and pre-computed with a Poisson LBM, and only the row clustering is required in the training procedure. Please, see page 9 and 10 plus the footnote number 4.

Q15. In the experiment reported in section 5.3 you say that "the initialization was performed with a first mapping of only 2500 documents followed by a nearest neighbour rule." Why did you choose this type of initialization? How did you sample those 2,500 cases? Why 2,500?

Authors' response:
Thanks for this suggestion. As this is possible to deal with more than 10000 rows for handling the SVD actually, even with a regular function in R langage now, the sentence has been changed into :

« The initialization is performed with the help of the first principal plane of CA. »

Q16. The experiment shows that a sensible data visualization can be generated ... so what?

Authors' response:
Thanks for this remark. This is an illustration. If the small (toy) datasets are 'easy' to visualize, this is clearly not true for this larger one.

Q17. In my previous review, I said that results seemed to be evaluated only in terms of error measures, but no qualitative comment was made on the (potential) meaning
of the blocks of data variables obtained (in the same way we could comment on the data prototypes of the blocks of data points)
I still cannot find any comment on that (although I agree that the parsimony of the resulting model makes, by itself, a point).

Authors' response:
Thanks for this suggestion. Some elements in that direction have been added at the end of the subsection 4.3 but no empirical results have been provided in this document which deals mainly with the generalization and not a double representation. Note that this could be possible to provide usual results such that a table of the more frequent or meaningful words in each clusters of the map for instance.

Q18. Minor things:
p.2 "Euclidian distance": Euclidean
p.4 "modeled separatly":separately
p.11 "Connexion to GTM" Connection
p.11 "in close form": closed
p.12 "The pdf of the latent block model becomes as:"
p.12 "The normal equations equating [...]"

p.12 "4.3. visualisation"; Visualization (the full text should be revised to make consistent use of US vs. British English)
p.13 "for the of contingency tables" ... ???
p.13 "... They are brievly described ...": briefly
p.14 "This indicator decreases with more compact and more separated clusters such that it is preferred minimal.": awkward sentence (The description of the S-index would
also require some re-phrasing.
p.14 "... and it is compound of 3 classes"

Authors' response:
Thanks for this concern. Done.

Q19. The full name of the model "generative topographic mapping" is used throughout the text despite the fact that the acronym has been defined almost at the onset.

Authors' response:
Thanks for this suggestion. Some acronyms have been added but note that it is not readable to have to much acronyms (in uppercase) too.

Q20. The format of the references requires some patient attention.

Authors' response:
Thanks for this remark. If the paper is accepted, we will be happy to correct the references under the checking of the edition process.

We tried our best to address all comments and concerns raised by the reviewers and believe that the paper is improved considerably. In the meantime, if there are still concerns or additional comments related to our answers and clarifications, we will be happy to address them.

Generalized topographic block model

Abstract

Co-clustering leads to parsimony in data visualisation with a number of parameters dramatically reduced in comparison to the dimensions of the data sample. Herein, we propose a new generalized approach for nonlinear mapping by a re-parameterization of the latent block mixture model. The densities modeling the blocks are in an exponential family such that the Gaussian, Bernoulli and Poisson laws are particular cases. The inference of the parameters is derived from the block expectation-maximization algorithm with a Newton-Raphson procedure at the maximization step. Empirical experiments with textual data validate the interest of our generalized model.

Keywords: Latent block mixture model, Exponential family, Generative topographic mapping, Block expectation-maximization, Visualisation.

1. Introduction

For the visualisation [1, 2] of a data matrix, the main proximities or the higher correlations are summarized by a comprehensible and low dimensional graphical view. When the number of variables is large, the visualisation may combine a preprocessing stage by selection or linear transformation [3, 4, 5]. In a co-clustering method, both sides of the matrix are partionned [6], hence the reduction of the variables space and the row clustering occur simultaneously. A earliest co-clustering formulation called direct clustering was introduced by Hartigan [7] who proposed a greedy algorithm for hierarchical co-clustering. We can also mention the following works [8, 9, 10, 11, 12] and the reviews in [13, 14, 15]. These methods are dedicated to a simultaneous clustering but not to visualisation.

Co-clustering is combined to self-organizing maps (SOM) for visualisation or clustering purposes in many ways in the litterature [16, 17, 18, 19, 20, 21, 22], with illustrations to biological or textual data. Such combination can improve the quality of the clustering [23, 24], with two contributing modeling factors or four sub-ones. Roughly speaking, the co-clustering leads to (a) the parsimony of the parameters and (b) the groups of variables. And, the auto-organization leads to (c) the partition of each class into several clusters and (d) the connections between neighboor clusters. Note that the subfactors (c) can enhance the classification [25], while (a) and (b) the regression [26].

The family of methods SOM counts the variants and the extensions of the Kohonen's map [27] which is a sequential clustering algorithm with decreasing connections of vicinity between the clusters for mapping continuous data. Modified versions are adapted to the analysis of discrete, sequential or block matrices for instance. Moreover, generative models for self-organizing maps has been justified [28, 29, 30]. The Generative Topographic Mapping (GTM) [31] is a probabilistic model of SOM for data visualisation [32, 33, 34, 35]. In GTM the auto-organization of the clusters is directly induced by the parameterization. The algorithm of Kohonen's map is re-formulated by embedding the auto-organization process at the level of the means of a Gaussian mixture model (GMM) [36].

Herein we are interested on a probabilistic co-clustering model, the latent block mixture model (LBM) [12, 37, 38], in order to visualize the natural classes in a block matrix with a parameterization similar to GTM. First, we define a general model of LBM with the help of an univariate exponential familly [39] which is well suited for most kinds of numerical variables. Then we introduce a parameterization of the central parameters in order to simultaneously perform the clustering and the reduction of the obtained clusters in a low dimensional space. The model is general enough to be related not only to self-organizing maps but also to recent approaches in factorization [40, 41, 42]. This offers a broad perspective for data analysis as illustrated through a generalized method for block generative topographic mapping or block GTM (BGTM) [43, 44]. If the previous models of block generative topographic mapping have been proposed for only one particular distribution for the blocks, the new generalization is able to provide a unified framework for the visualisation of data block matrices and can help for implementing and comparing alternative distributions in future.

The paper is organized as follows. In section 2, we introduce the latent block model for an exponential family and add the constraints. In section 3, we present the related objective function to optimize for the estimation of parameters. We deduce the learning algorithm in a general setting, from the block expectation-maximization (BEM) [45]. In section 4 we present the connection of our approach with GTM and discuss the resulting nonlinear visualisation. In section 5 we present the numerical experiments for testing the proposed approach. Finally, in section 6 we summarize our contribution.

2. Generalized LBM

Let us have $\mathbf{x} := \{x_{ij}; i = 1, \ldots, n; j = 1, \ldots, d\}$ stand for a data matrix of size $n \times d$. When $\mathbf{x}$ is a two-way contingency table it is associated to two categorical variables that take values in sets $I = \{1, \ldots, n\}$ and $J = \{1, \ldots, d\}$. In this case, the entries x_{ij} are co-occurrences of row and column categories, each of them counts the number of entities that fall simultaneously in the corresponding row and column categories. Let $\mathbf{z}$ and $\mathbf{w}$ be partitions in g row clusters and m column clusters of I and J of $\mathbf{x}$. The partition $\mathbf{z}$ will be represented by the vector of labels $(z_1, \ldots, , z_n)$ where $z_i \in \{1, \ldots, g\}$ or, by the classification matrix $\{z_{ik}; i = 1, \ldots, n; k = 1, \ldots, g\}$ where $z_{ik} = 1$ if i belongs to the k^{th} cluster and 0 otherwise. A similar notation will be used for the partition $\mathbf{w}$ which will be represented by the vector $(w_1, \ldots, w_j, \ldots, w_d)$ where $w_j \in \{1, \ldots, m\}$ or the classification matrix $\{w_{j\ell}; j = 1, \ldots, d; \ell = 1, \ldots, m\}$. Note that $z_{ik}w_{j\ell} = 1$ if x_{ij} belongs to the $(k\ell)^{th}$ block and 0 otherwise. For a latent block model, the $n \times d$ random variables that generate the observed cells x_{ij} are assumed to be independent, once $\mathbf{z}$ and $\mathbf{w}$ are fixed, they make it possible to define a co-clustering model. 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, k, or ℓ without indicating the limits of variation, which are implicit.

2.1. Latent block model (LBM)

The probability density function (pdf) of a latent block model is denoted $f_{LBM}(\mathbf{x}; \boldsymbol{\theta})$ and defined as the following decomposition. It is obtained by independence of $\mathbf{z}$ and $\mathbf{w}$, by summing over all the assignments [12] and takes the following form:

$$\sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} \prod_i p_{z_i} \prod_j q_{w_j} \prod_{i,j} \varphi(x_{ij}; \alpha^{ij}_{z_i w_j}),$$

where the set of all the possible assignments is denoted $\mathcal{Z}$ for I and $\mathcal{W}$ for J, while $\varphi(.; \alpha^{ij}_{k\ell})$ is a probability density function defined for cell (ij) on the set of reals $\mathbb{R}$ while $\alpha^{ij}_{k\ell}$ depends on the parameter $\alpha_{k\ell}$ as given in (1). The vectors of the probabilities p_k and q_ℓ that a row (resp. a column) belongs to the k^{th} component (resp. ℓ^{th} component) are denoted $\mathbf{p} = (p_1, \ldots, p_g)$ (resp.

Model	pdf of cell $(k\ell)$	φ	$A(\alpha_{k\ell}^{ij})$	$B(\alpha_{k\ell}^{ij})$	β_{ij}	$\alpha_{k\ell}$	x_{ij}
GLBM	$\mathcal{N}(\alpha_{k\ell};\sigma_{k\ell 0})$	$\frac{exp(-\lvert x_{ij}-\alpha_{k\ell}\rvert^2/2\sigma_{k\ell 0}^2)}{\sqrt{2\pi}\sigma_{k\ell 0}}$	$\alpha_{k\ell}/\sigma_{k\ell 0}^2$	$\alpha_{k\ell}^2/2\sigma_{k\ell 0}^2$	1	$\mathbb{R}$	$\mathbb{R}$
BLBM	$\mathcal{B}(\alpha_{k\ell})$	$(\alpha_{k\ell})^{x_{ij}}(1-\alpha_{k\ell})^{1-x_{ij}}$	$\log\frac{\alpha_{k\ell}}{1-\alpha_{k\ell}}$	$-\log(1-\alpha_{k\ell})$	1	$[0;1]$	$\{0,1\}$
PLBM	$\mathcal{P}(\beta_{ij}\alpha_{k\ell})$	$\frac{exp(-\beta_{ij}\alpha_{k\ell})(\beta_{ij}\alpha_{k\ell})^{x_{ij}}}{x_{ij}!}$	$\log\alpha_{k\ell}$	$\beta_{ij}\alpha_{k\ell}$	$\mu_i\nu_j$	$[0;1]$	$\mathbb{N}_+$

Table 1: Table with the three cases of distribution for the matricial cells modeled with ELBM.

$\mathbf{q}=(q_1,\ldots,q_m)$). The set of parameters is denoted $\boldsymbol{\theta}$ and is a compound of $\mathbf{p}$ and $\mathbf{q}$ plus $\boldsymbol{\alpha}$ which aggregates all the parameters from the pdf of the cells, $\boldsymbol{\theta}=\{\mathbf{p},\mathbf{q},\boldsymbol{\alpha}\}$. The set of parameters $\boldsymbol{\theta}$ of the model can be estimated by maximizing the log-likelihood:

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

The block model is dramatically more parsimonious than the usual mixture model where each dimension of the data table is modeled separately. Next, we describe the latent block model where φ is in an exponential family.

2.2. Univariate exponential family of distributions

When the cells are generated with an exponential family, the latent block model is denoted in the following ELBM and the density function for the $(k\ell)^{th}$ block is written:

$$\varphi(x_{ij};\alpha_{k\ell}^{ij})=exp\left(x_{ij}A(\alpha_{k\ell}^{ij})-B(\alpha_{k\ell}^{ij})+C(x_{ij})\right)\ ,$$

where $A(\alpha_{k\ell}^{ij})$ is the natural parameter, while $B(\alpha_{k\ell}^{ij})$ and $C(x_{ij})$ ensure that φ is a probability density function. The considered form of distributions is defined without nuisance parameter and without loss of generality. Note that a more general expression is possible for modeling more particular distributions. For instance a function of x_{ij} could be used instead of the identity one or A could be chosen multivariate. It is also supposed that the quantities $\alpha_{k\ell}^{ij}$ are written as a function of a fixed parameter depending on the data β_{ij} and an unknown parameter named $\alpha_{k\ell}$, such that:

$$\alpha_{k\ell}^{ij}=\beta_{ij}\alpha_{k\ell}\,. \tag{1}$$

The model can be represented by a graphical model depicted in Figure 1. Here two aggregating matrices are involved, $\boldsymbol{\alpha} = (\alpha_{k\ell})_{g\times m}$ for the parameters and $\boldsymbol{\beta} = (\beta_{ij})_{n\times d}$ for the multiplicative effects. Three cases of distributions which belong to this family for discrete and continuous matrices are considered. The different distributions are listed on Table 1, where the cells are drawn from one particular distribution. For a Bernoulli law with the parameters $\alpha_{k\ell}$, the model is denoted BLBM [12]. For a Poisson law with the parameters $\alpha_{k\ell}$, it is denoted PLBM [46], with $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^T$ where $\mu_i = \sum_j x_{ij}$ and $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_d)^T$ where $\nu_j = \sum_i x_{ij}$. For a normal law, it is denoted GLBM [47] with the means $\alpha_{k\ell}$ and the variances $\sigma_{k\ell 0}$ assumed constant here. For the three cases, the support for the variables generating the observation x_{ij} and the parameter range of $\alpha_{k\ell}$ are defined in Table 1. Note that in the case of a Poisson law, the parameters $\alpha_{k\ell}$ can be chosen unconstrained as introduced in [46], and the quantities β_{ij} can be optimized too in certain cases. Next, each parameter $\alpha_{k\ell}$ is written with a link function as explained.

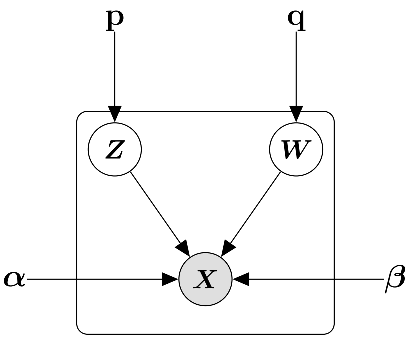


Figure 1: Graphical notation for the Latent block model with random variables $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ generating the observations and latent labels.

2.3. Re-parameterization of the model

The parameters of the exponential latent block model are parameterized with two sets of (unknown) vectors,

$$\begin{array}{l}\left\{\xi_k \in \mathbb{R}^h, 1 \leq k \leq g\right\}, \\ \left\{w_\ell \in \mathbb{R}^h, 1 \leq \ell \leq m\right\}.\end{array} \tag{2}$$

where $h \in \mathbb{N}_+^*$ is the dimension of the latent space. These two sets of vectors are used for modeling the blocks $(k\ell)$ because each $\alpha_{k\ell}$ is dependent on two indices, k and ℓ. As an effect from the k^{th} and ℓ^{th} clusters, the inner products are then considered as

$$\{w_\ell^T \xi_k; 1 \leq k \leq g, 1 \leq \ell \leq m\}. \tag{3}$$

To map the inner product ($w_\ell^T \xi_k \in \mathbb{R}$) onto its corresponding parameter ($\alpha_{k\ell} \in [0;1]$) a link function $\varrho(.)$ is used. For instance, for the Bernoulli law the sigmoid function can be chosen. For all k, and ℓ, we have:

$$\alpha_{k\ell} = \varrho(w_\ell^T \xi_k). \tag{4}$$

For the Poisson law, the sigmoid or the exponential function may be selected. For the Gaussian law, the canonical identity function can be used for ϱ. The reduced $g \times m$ matrix $\boldsymbol{\alpha}$ with cells defined by the parameters $\alpha_{k\ell}$ in the previous co-clustering model is re-parameterized with the two matrices:

$$\begin{array}{rcl} \boldsymbol{\Phi} & = & [\xi_1|\xi_2|\cdots|\xi_g]^T, \\ \boldsymbol{\Omega} & = & [w_1|w_2|\cdots|w_m]. \end{array} \tag{5}$$

The resulting model is more parsimonious than a GTM with an usual mixture model because the loading matrix $\boldsymbol{\Omega}$ counts only m columns instead of d ones and $m \ll d$. Next, we propose a generalized criterion and an algorithm for the inference of $\boldsymbol{\theta}$ in the exponential latent block model.

$\beta_{ij} = 1$	$\nabla\tilde{Q}_\ell$	$=$	$\boldsymbol{\Phi}^T \mathbf{G}'_\ell \{\mathbf{M}_{A'} Y_\ell - d_\ell \mathbf{M}_{B'} \mathbf{G}_c\} \mathbf{1}_g$
	$\tilde{H}_\ell$	$=$	$\boldsymbol{\Phi}^T \mathbf{G}'^{\times 2}_\ell \{\mathbf{M}_{A''} Y_\ell - d_\ell \mathbf{M}_{B''} \mathbf{G}_c\} \boldsymbol{\Phi} + \boldsymbol{\Phi}^T \mathbf{G}''_\ell \{\mathbf{M}_{A'} Y_\ell - d_\ell \mathbf{M}_{B'} \mathbf{G}_c\} \boldsymbol{\Phi}$
$\beta_{ij} \neq 1$	$\nabla\tilde{Q}_\ell$	$=$	$\boldsymbol{\Phi}^T \mathbf{G}'_\ell \mathbf{C}^T \{\mathbf{T}_{A'} \times_2 (\mathbf{x} \odot \boldsymbol{\beta} \times \mathbf{D}_\ell) - \mathbf{T}_{B'} \times_2 (\boldsymbol{\beta} \times \mathbf{D}_\ell)\} \mathbf{1_g}$
	$\tilde{H}_\ell$	$=$	$\boldsymbol{\Phi}^T \mathbf{G}'^{\times 2}_\ell \mathbf{C}^T \left\{\mathbf{T}_{A''} \times_2 (\mathbf{x} \odot \boldsymbol{\beta}^{\times 2} \times \mathbf{D}_\ell) - \mathbf{T}_{B''} \times_2 \boldsymbol{\beta}^{\times 2} \times \mathbf{D}_\ell\right\} \boldsymbol{\Phi}$
		$+$	$\boldsymbol{\Phi}^T \mathbf{G}''_\ell \mathbf{C}^T \{\mathbf{T}_{A'} \times_2 (\mathbf{x} \odot \boldsymbol{\beta} \times \mathbf{D}_\ell) - \mathbf{T}_{B'} \times_2 (\boldsymbol{\beta} \times \mathbf{D}_\ell)\} \boldsymbol{\Phi}$

Table 2: Derivatives of the constrained criterion at t-th step of EM for the exponential family.

3. Parameters inference and algorithm

We aim to address the problem of parameters estimation by a maximum likelihood (ML) approach, in Expectation and Maximization steps, such that:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\mathbf{x}; \boldsymbol{\theta}) \,.$$

To induce a quantization, the mixing probabilities can be chosen constant and equidistributed. For visualisation like in the generative topographic mapping, $\boldsymbol{\Phi}$ may be kept constant and not optimized. Hence, the set of parameters is reduced to $\boldsymbol{\theta} = \boldsymbol{\Omega}$ in the following.

3.1. Expectation step

For the estimation of a suitable value of $\boldsymbol{\theta}$ by the maximum likelihood, the block EM algorithm or BEM (see [45]) leads to an objective function denoted $\tilde{Q}$ for short. It is maximized instead of the original log-likelihood function and written:

$$\begin{aligned} & \tilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\ = & \sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \log \varphi(x_{ij}; \alpha_{k\ell}^{ij}) \\ = & \sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \left\{ x_{ij} A\left(\alpha_{k\ell}^{ij}\right) - B\left(\alpha_{k\ell}^{ij}\right\}\right) + cte \,. \end{aligned}$$

Here $\boldsymbol{\theta}^{(t)}$ is a current value of the parameters at the t-th step, cte is a constant independent of the parameters, the superscript (t) permits to denote a current estimation of the parameters or a function of them but is removed in the following when the notation is made lighter. The quantities c_{ik} (resp. $d_{j\ell}$) are the posterior probabilities that a row (resp. a column) belongs to the $(k\ell)^{th}$ block and at the current time (t). It is also denoted $y_{k\ell}^{(t)} = \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} x_{ij}$, $c_k^{(t)} = \sum_i c_{ik}^{(t)}$, and $d_\ell^{(t)} = \sum_j d_{j\ell}^{(t)}$. The function $\tilde{Q}$ comes from an approach type EM [48] which often makes possible to obtain values of the parameters in a closed form as opposed to a direct optimization with function L. The algorithm proceeds by alternating two steps at each iteration $(t+1)$. At the first step called Expectation or E-step, the posterior probabilities are computed knowing the data and the current value of the parameter $\boldsymbol{\theta}^{(t)}$. These probabilities are estimated by maximizing the same objective function

$\tilde{Q}$ where their entropies are added. They are solution of the dependent equations:

$$\begin{array}{lcl} c_{ik} & \propto & \exp \sum_{j\ell} d_{j\ell} \log \varphi(x_{ij}; \alpha_{k\ell}) \,, \\ d_{j\ell} & \propto & \exp \sum_{ik} c_{ik} \log \varphi(x_{ij}; \alpha_{k\ell}) \,. \end{array} \tag{6}$$

A new current value of the parameters $\boldsymbol{\theta}^{(t+1)}$ is obtained at a second step called Maximization or M-step where the new parameters are written as weighted means of the sufficient statistics and the weights are the posterior probabilities. At each iteration $(t+1)$, the algorithm increases the function $\tilde{Q}$ as explained in [12, 46, 47].

The two next sections present the algorithm for the estimation of the parameters by optimizing the objective function parameterized with the two matrices.

3.2. Derivatives of the objective function

The expression for the gradient vector $\tilde{Q}_\ell^{(t)}$ and the Hessian matrix $\tilde{H}_\ell^{(t)}$ for the function $\tilde{Q}$ w.r. to the parameter w_ℓ, may be directly written[1] according to the value of β_{ij} in a matricial format as given in Table 2. Here it is also denoted $\mathbf{1}_a = (1)_{a\times 1}$ with an integer a, $\mathbf{C} = (c_{ik})_{n\times g}$ and $\mathbf{D} = (d_{j\ell})_{d\times m}$ while other matrices or tensors are given in Table 3. The operators are as follows: $M^{\times 2} = M \times M$, $M \odot N$ is the Hadamard product of M and N, while $T_M \times_2 N = (\sum_j M_{ijk} N_{ij})_{a\times c}$ where $T_M = (M_{ijk})_{a\times b\times c}$ is a tensor with three modes and $N = (N_{ij})_{a\times b}$ is a matrix with a rows and b columns. $diag_k(v_k)$ is the diagonal matrix with non nul elements v_k. These general expressions could permit to differentiate between the cell distributions for continuous or binary variables ($\beta_{ij} = 1$) and categorical or counting variables ($\beta_{ij} \neq 1$).

From the Table 2 or by a direct derivation from each particular function $\tilde{Q}$, the formula for updating the matrix $\boldsymbol{\Omega}$ for the three univariate distributions given in Table 1 can be deduced. As the provided expression is general, other distributions or link functions are also possible. Moreover, the Hessian matrix from the model can be not definite negative in certain cases like for the Poisson law and an approximation can be derived as in [44]. It is supposed in the following that the Hessian matrix has been altered into a well defined approximation if required. An increase of the diagonal from the matrix by

[1] We denote: $\frac{\partial \alpha_{k\ell}}{\partial w_\ell} = \alpha'_{k\ell}\xi_k$ and $\frac{\partial^2 \alpha_{k\ell}}{\partial w_\ell \partial w_\ell^T} = \alpha''_{k\ell}\xi_k\xi_k^T$, while A' and A'' (resp. B' and B'') are the first and second order one-dimensional derivative of A (resp. B).

$$
\begin{array}{lcl}
\mathbf{M}_{A'} & = & (A'(\alpha_{k\ell}))_{g\times m} \\
\mathbf{M}_{A''} & = & (A''(\alpha_{k\ell}))_{g\times m} \\
\mathbf{M}_{B'} & = & (B'(\alpha_{k\ell}))_{g\times m} \\
\mathbf{M}_{B''} & = & (B''(\alpha_{k\ell}))_{g\times m} \\
\mathbf{T}_{A'} & = & (A'(\beta_{ij}\alpha_{k\ell}))_{n\times d\times g} \\
\mathbf{T}_{A''} & = & (A''(\beta_{ij}\alpha_{k\ell}))_{n\times d\times g} \\
\mathbf{T}_{B'} & = & (B'(\beta_{ij}\alpha_{k\ell}))_{n\times d\times g} \\
\mathbf{T}_{B''} & = & (B''(\beta_{ij}\alpha_{k\ell}))_{n\times d\times g} \\
\mathbf{G}'_{\ell} & = & diag_k\,(\alpha'_{k\ell}) \\
\mathbf{G}''_{\ell} & = & diag_k\,(\alpha''_{k\ell}) \\
\mathbf{Y}_{\ell} & = & diag_k\,(y_{k\ell}) \\
\mathbf{D}_{\ell} & = & diag_j(d_{j\ell}) \\
\mathbf{G}_{\mathbf{c}} & = & diag_k\,(c_k)
\end{array}
$$

Table 3: Matrices and tensors for the derivatives.

an additive or multiplicative way may be useful, at least for exact Hessian matrices and their regularization. Note also that numerical approximations of the Hessian matrix exist in the literature, and from [49] a justification of the approach can be derived.

3.3. Maximization step and learning algorithm

The algorithm for maximizing $\tilde{Q}$ depending on the new set of parameters $\boldsymbol{\theta}$ proceeds iteratively as previously explained but at the M-step the next current value of the parameters is estimated by:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \tilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\,. \tag{7}$$

Here, a Newton-Raphson procedure is considered for this maximization step because of the non linearities. In [32], a model which is a particular case of our proposal was introduced. Indeed, when cancelling the column clustering, the objective functions are the same. Our constrained model is more general than this former generalizing approach. In particular, the Bernoulli case is identical when $m = d$. It must be noticed that in the original paper [32], the Poisson case is slightly different because instead of the sigmoidal function, an exponential transformation was chosen. This alternative unbounded link function has been tested for block GTM in pratice, but was less stable in its current implementation.

In order to find a local solution to (7) in the general case, the proposed algorithm is BEM with a new M-step. In order to decide when the algorithm should be stopped, a small positive constant ϵ_{BEM} is taken to indicate the $\tilde{Q}$ value change is small or there is no change. After initializing the parameters, the algorithm is an incremental procedure repeating two steps:

- E-step where the posterior probabilities $\{c_{ik}\}$ or $\{d_{j\ell}\}$ are updated (see update formula (6)).

- M-step where the parameters are updated for all columns of $\mathbf{\Omega}$, such that $\tilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is increased with respect to w_ℓ as follows:

$$w_\ell^{(t+1)} \quad = \quad w_\ell^{(t)} - \left[\tilde{H}_\ell^{(t)}\right]^{-1} \nabla\tilde{Q}_\ell^{(t)} .$$

If any, nuisance parameters need to be updated here, such as in BEM for the unconstrained LBM. Note also that for factorization, $\mathbf{\Phi}$ would be also updated similarly than $\mathbf{\Omega}$ by symmetry of the minimized criterion. By repeating the computational iterations, the agorithm converges towards a solution where the parameters reach a stable value. They are denoted with a hat while the final matrices of posterior probabilites are respectively $\hat{\mathbf{C}} = (\hat{c}_{ik})$ and $\hat{\mathbf{D}} = (\hat{d}_{j\ell})$.

4. From GTM to block GTM

A semi-flexible model is obtained for data visualisation with our generalized setting and a matrix $\mathbf{\Phi}$ constant as defined in [31]. We present the former model of generative topographic mapping next, after a brief description of its particular matrix for inducing an auto-organization of the central parameters during the inference. We also discuss the method for the construction of a nonlinear projection after the parameters estimation.

4.1. Set of basis functions of GTM

The constant matrix $\mathbf{\Phi}$ in the generative topographic mapping is defined as follows. Let us consider a set of g two-dimensional vectors of coordinates,

$$\mathcal{S} = \left\{ \boldsymbol{s}_k = \begin{pmatrix} s_{k1} \\ s_{k2} \end{pmatrix} ; k = 1, ..., g \right\} .$$

They come from the nodes of a regular mesh which discretizes the latent space for the projection, a regular square on the plane. $\mathcal{S}$ is directly related to the set of nodes in the Kohonen's maps. Each coordinate $\boldsymbol{s}_k$ is nonlinearly transformed as follows. This is a vector of h kernel Gaussian functions, $\psi_o(\boldsymbol{s}_k) = e^{-||\boldsymbol{s}_k - \mu_{\psi^o}||^2/2\nu_{\psi^o}}$ with mean centers $\mu_{\phi^o} \in \mathbb{R}^2$, variances $\nu_{\phi^o} \in \mathbb{R}^*_+$ and $1 \leq o \leq h$, completed with an intercept equal to one and the two coordinates of s_k, such that for all k:

$$\xi_k = \left[1, s_k^T, \psi_1(\boldsymbol{s}_k), \psi_2(\boldsymbol{s}_k), \cdots, \psi_h(\boldsymbol{s}_k)\right]^T .$$

The three first components are added here for the empirical results for BGTM otherwise they may be omitted for the original GTM. The relative positions between the bidimensional coordinates in $\mathcal{S}$ are kept for the transformed vectors at least locally.

Introducing the vectors ξ_k at the level of the central parameters of the generative model with the inner products $\xi_k^T w_\ell$ induces an organization of the cluster centers as a discretized surface. It is able to summarize the data cloud and its unfolding leads to the projection on the latent space discretized by $\mathcal{S}$. In the re-parameterized ELBM, the obtained model is called the block generative topographic mapping (block GTM). In [43, 44], the discrete cases for binary and counting features have been presented and are called respectively Poisson block generative topographic mapping (PBGTM) and Bernoulli block generative topographic mapping (BBGTM). In the Gaussian case, the obtained model is called Gaussian block generative topographic mapping (GBGTM). This model has an inference in closed form[2] for linearity reasons and derived from the proposed algorithm in the previous section.

4.2. Connection to GTM

In the next paragraphs, we present the generative topographic mapping with its parametric model, its training algorithm and its link with our general framework. Our proposed model generalizes the generative topographic

[2] This is: $w_\ell \leftarrow \frac{1}{d_\ell}\left(\boldsymbol{\Phi}^T \Upsilon_\ell G_c \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \Upsilon_\ell \mathbf{Y}_\ell \mathbf{1}_g$. The nuisance parameters $\sigma^2_{k\ell 0}$ are estimated as $\sigma^2_{k\ell} \leftarrow \sum c_{ik} d_{j\ell}(x_{ij} - \alpha_{k\ell})^2 / c_k d_\ell$ while $\Upsilon_\ell = diag_k(\sigma^{-2}_{k\ell})$ and $\alpha_{k\ell} = w_\ell^T \xi_k$. Note that a non fuzzy column clustering could be preferred for GBGTM as the non constrained model [47] and the batch procedure for a Kohonen's map with a block setting in [23].

mapping. Let us have $m = d$, the i^{th} row of $\mathbf{x}$ denoted $\mathbf{x}_i$ in a column-vector format and all $\sigma_{k\ell 0}$ equal to σ. In this case, when $p_k = \frac{1}{g}$ and the terms coming from the mixing probabilities q_ℓ are removed, the pdf of the latent block model can be written as follows:

$$\prod_i \sum_k p_k \frac{\exp(-||\mathbf{x}_i - \mathbf{\Omega}^T \xi_k||^2/2\sigma^2)}{(2\pi)^{d/2}\sigma^{2d}} \,,$$

which is the likelihood of the generative topographic mapping. With the gradient equal to zero, this leads to the update equation of the generative topographic mapping which is written as follows in matricial form:

$$\mathbf{\Phi}^T G_c \mathbf{\Phi}\, \mathbf{\Omega} = \mathbf{\Phi}^T C^T \mathbf{x} \,,$$

which is exactly[3] the solution in [31] and can be directly found by expectation-maximization. In generative topographic mapping (resp. block generative topographic mapping), the estimation of the variance term is identical to its corresponding unconstrained model with the following update:

$$\sigma^2 = \frac{1}{n \times d} \sum_{i,k} ||\mathbf{x}_i - \mathbf{\Omega}^T \xi_k||^2 \,.$$

4.3. Visualisation

A nonlinear projection of a given data sample is obtained after the parameters inference. This is the set of the two-dimensional coordinates or projection on the plane for all the row data. For the k^{th} cluster it corresponds a two-dimensional position s_k on the plane, such that an averaged position is written:

$$\hat{\boldsymbol{s}}_i = \sum_k \hat{c}_{ik} s_k \,.$$

From an auto-organization point of view of the probabilities $\alpha_{k\ell}$, only a projection at the maximum a posteriori position can also be preferred. Indeed, when a row i has a higher probability in a given cluster then it belongs to

[3] Constraints corresponding to a non fuzzy clustering of the rows can be defined by a linear matrix $\mathbf{R}$ such that when $m < d$, it can be written $\mathbf{R\Omega} = \mathbf{0}_{h \times m}$ where the right member is a null matrix. In this case a solution for the update with the constraints may be obtained as in restricted least squares.

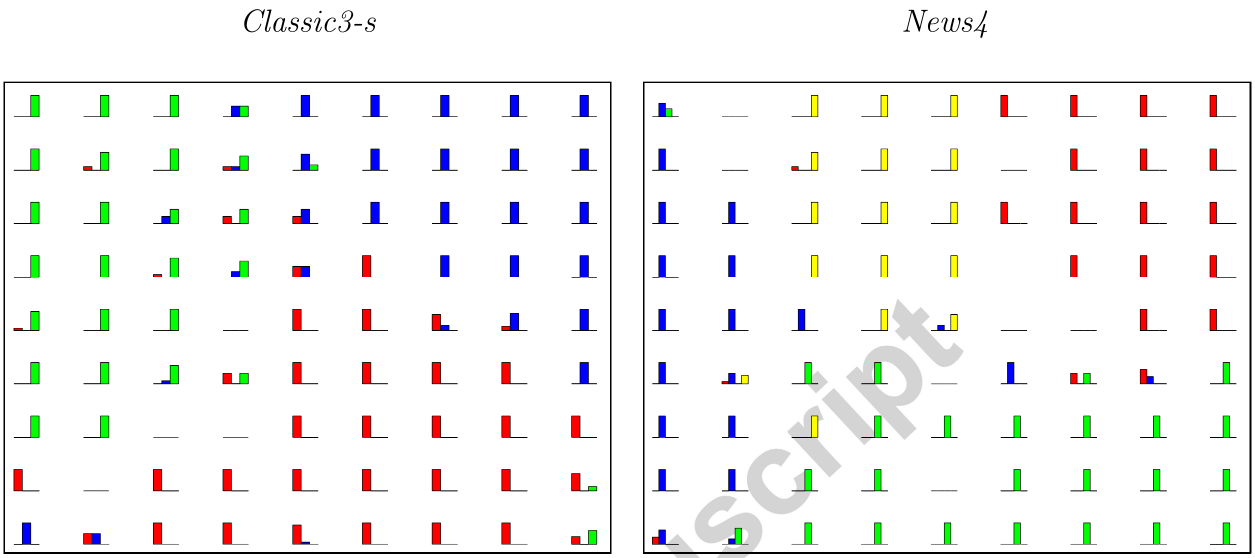


Figure 2: Maps with the method PBGTM.

this cluster. An estimated label for the i^{th} row maximizes the posterior probabilities and is denoted $\hat{z}_i$. The corresponding datum can be represented at a $\hat{z}_i^{\text{th}}$ node with coordinates $\hat{\boldsymbol{s}}_i^{MAP} = \boldsymbol{s}_{\hat{z}_i}$. By performing this procedure for each datum, the model builds a reduced view of the dataset in tabular form. As our model is defined for the projection of the rows, a direct double representation with both rows and variables is not available, and a methodology is required for adding a projection of the variables. For instance for a discrete law, the more meaningful columns for each cluster k^{th} can be added at the coordinates of the corresponding node s_k or a continuous projection can be deduced from the parameters (see [43]).

Next, experiments with the visualisation of real data illustrate the generalized model in order to validate the proposal.

5. Experiments

In this section, we are interested in illustrating our model for textual data with diverse functions φ for a nonlinear mapping of contingency tables in order to observe how the constrained latent block model behaves. We present our experiments with two small datasets and one large dataset.

5.1. Experimental settings

The results of the methods are compared with three indicators. For an evaluation of the quality of the clustering, an error rate is obtained from the estimated labels $\{\hat{z}_i\}$ and denoted error-rate. For an evaluation of the quality of the nonlinear mapping, two indicators are obtained by a measure of the separation of the original classes from the true labels and the projection points $\{\hat{\boldsymbol{s}}_i\}$ in the latent space. They are briefly described below.

- The error rate is the percentage of missclassified samples, say $\frac{\#\left\{z_i \neq \hat{z}^{\mathcal{S}}_{\hat{z}_i}\right\}}{n}$. Here, $\hat{z}^{\mathcal{S}}_k$ denotes the estimated class label by majority vote of the k^{th} cluster from the map while z_i is the true class label and $\#\{.\}$ is the cardinality of a set. This indicator may decrease with the size of the map g until a limit if the classes are not perfectly separated.

- The Davies-Bouldin index [50], denoted DB-index, is the average similarity between each class and its most similar one. The similarity between the k_1^{th} class and the k_2^{th} one is measured by the quotient $(v_{k_1} + v_{k_2})/d_{k_1k_2}$ where v_{k_1} and v_{k_2} are the intra-variance in each class while $d_{k_1k_2}$ is the Euclidean distance between the class centers. This indicator is preferred minimal as it decreases if the classes are more separated.

- The average of the Silhouettes [51], denoted S-index, is the mean value of $(b_i - a_i)/\max(a_i, b_i)$ where a_i and b_i are as follows. The first quantity is the average dissimilarity between the i^{th} datum and the other ones in the $\hat{z}_i^{\text{th}}$ class. The second one is the minimal average dissimilarity from the other classes. This indicator is confined in the interval $[-1; 1]$ and is preferred maximal for more compact classes.

A visual inspection of the final map and the values of the indicators lead to an empirical choice of the parameters g, m, and h in BGTM for the results presented hereafter with three datasets.

5.2. Output for two small datasets

When $\mathbf{x}$ is a contingency table, I corresponds to a corpus of documents, and J to the vocabulary, so the frequency x_{ij} denotes the number of occurrences of a word in a document. The two datasets considered for comparing the methods are:

- The dataset *News4* or *N4* which consists 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 [32].

- The dataset *Classic3-s* or *C3-s* is a sample from *Classic3* resulting to a contingency table of size 450 × 171. The texts are distributed among 3 topics named respectively *Medline*, *Cisi* and *Cranfield*. The dataset *Classic3* or *C3* is a commonly used dataset [11] for experiments in co-clustering.

In this experiment, for the three versions of the block generative topographic mapping, the map is a square with a size equal to $g = 9^2$, while the number of column clusters is $m = 20$. The number of basis functions for the nonlinear mapping is $h = 28$, with 5^2 nonlinear basis functions plus the three linear components.

The empirical visualisations are shown in Figure 2 for the case of a Poisson law. Each map is represented as follows. For each cluster, a barplot corresponding to the true labels of the data is constructed after fitting the model. Hence, for a given dataset the map shows a table of 9 × 9 barplots such that if two clusters are near in the data space and on the map they should have similar barplots. This is a tabular view of the dataset which also confirms that the nearest clusters have texts with similar topics as expected for a self-organizing map. Moreover, most of the clusters with misclassification are at the frontier between two classes which confirms that the classes are well detected by the method.

The statistics for a comparison of the different methods are summarized in Table 4 with the obtained results per method and per dataset. The last two indicators are informative of how compact are the projections of the classes, but we are mostly interested in an auto-organization of the clusters. Globally, the Poisson case performs better as expected because the binary version loses the information of the true number of occurrences. Moreover, the Gaussian case behaves slightly like the Bernoulli one because the clusters do not overlap much. To sum up, the Poisson law with parametric constraints is clearly more valuable than the other laws considered in our experiments. The frequencies are meaningful in clustering textual data. If this is clearly true for *N4*, in the case of *C3-s* outliers seem to perturbate the empirical result obtained from only one sample and a more robust model might be preferred.

Algorithm	error-rate (%)		S-index		DB-index	
	N4	*C3-s*	*N4*	*C3-s*	*N4*	*C3-s*
PBGTM	3.8	5.3	0.44	0.36	0.82	0.94
BBGTM	5.6	5.1	0.27	0.36	1.68	0.89
GBGTM	8.3	5.1	0.27	0.28	1.14	1.55
BCASOM	6.5	3.6	0.51	0.36	0.76	1.38

Table 4: Error rate in percent, S-index and DB-index for the two datasets considered, *N4* and *C3-s*.

For comparing our approach with an alternative combination of co-clustering with SOM, we have also computed the indicators with a variant[4] of CASOM [52]. The parameters in each multinomial distribution are reduced to only m values where a fixed pre-clustering was obtained from the Poisson latent block model. The results from this alternative[5] method, denoted BCASOM, are presented in Table 4. Hence, BGTM leads to less compact projections of the classes with the current implementation for these small datasets. Note that a regularisation of the matrix $\mathbf{\Omega}$ as in [53] for our generalized model, or an early stopping and another setting of the neighborhood function for BCASOM may improve the empirical results.

5.3. *Output for a table of size* 12648 × 6034

The dataset *PubMed*5 comes from the collection 10Pubmed [54], with approximately 15500 medical abstracts from the database *Medline* published between the years 2000 and 2008 for 10 classes. While the size of the original data table is 15565 × 22437, only the five largest classes are included here for this experiment with small overlapping. We end with 12648 documents and only 6034 terms after a selection of the vocabulary: when selecting the

[4] Let's have $\mathbf{P} = (P_{k\ell})_{m\times g}$ the matrix of component parameters in the $(k\ell)^{\text{th}}$ blocks, and $\mathbf{H} = (h_{kk'})_{g\times g}$ the matrix with the values of the neighborhood function where the width is decreasing with the iteration (t). $\mathbf{C}$ and $\mathbf{D}$ (here binary) are the matrices of the posterior probabilities. $\mathbf{N}$ is the diagonal matrix with non null elements $\mathbf{1}_d^T\mathbf{D}$. The update equations are as follows. a) $\mathbf{C} \propto \exp\{\mathbf{xD}\log\{\mathbf{P}\}\mathbf{H}\}$ with an usual normalisation from the diagonal elements equal to the inverse of the components of $\mathbf{C1}_g$, while $\exp\{.\}$ and $\log\{.\}$ are respectively the logarithmic and exponential transformation of each cell of the matrix. b) $\mathbf{P}^T \propto \mathbf{HC}^T\mathbf{xD}$ with a normalisation from the diagonal elements equal to the inverse of the components of $\mathbf{1}_m^T(\mathbf{NP})$.

[5] Another possible model is the multinomial GTM [32] with restricted parameters.

five larger classes, the terms which occur in less than 6 different texts are removed. The initialization is performed with the help of the first principal plane of CA.

The parameters of the model PBGTM are chosen equal to $m = 30$, $g = 9^2$, and $h = 19$. The value of the S-index is 0.42 for this dataset. The confusion matrix for the proposed model is presented in Table 5, while the overall accuracy is equal to 97.1%. In the Figure 3, the five classes are almost perfectly separated on the map.

		PBGTM		
1485	2	2	20	8
1	1499	15	27	7
7	8	3220	41	7
25	18	40	3609	11
59	17	7	46	2467

Table 5: Confusion matrix for $PubMed5$.

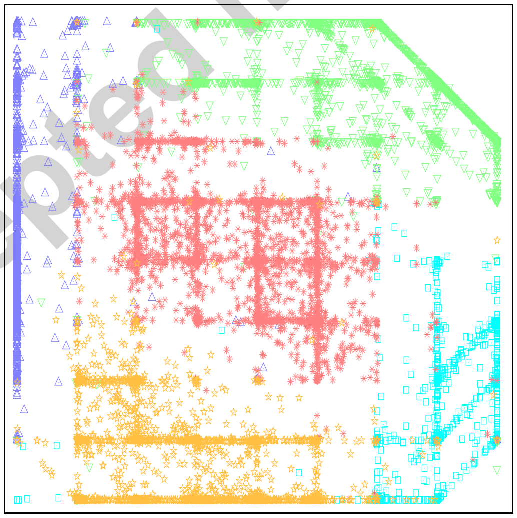


Figure 3: Nonlinear map with the method PBGTM, with the coordinates $(\hat{s}_i)$ for the real dataset $PubMed5$.

In conclusion, we observe that the method is able to deal with a large table, with a large number of columns. For a comparison with an usual mixture model where the columns are not clustered, the number of parameters would be $d \times g = 488754$ instead of $m \times h = 570$. An alternative model such as in [32] would need $d \times h = 114646$ parameters. Hence our proposal is dramatically more parsimonious with a dramatic reduction factor of respectively 0.001 and 0.005.

6. Conclusion

Herein, we have proposed a family of models for the reduction and projection of numerical tables with a block structure. In the experiments for discrete data we observe that block GTM is able to present a quick summary for three datasets. Our proposal brings parsimony, flexibility and more generality than the existing models.

As a perspective, the parameter selection or the clustering of the nodes of the map [55, 56] could be addressed. Local maxima in the training process is a serious concern in GTM, this may be worse for BGTM hence this issue needs to be reduced in future for making the model available for further extensive experiments, by adding constraints [57, 58] for instance. Other distributions and link functions for the cells are also interesting to explore in order to help improving the robustness and the fitting.

Acknowledgment

The authors would like to thanks the editor and the reviewers for contributing to the improvement of the earlier versions of the paper. This research was partially supported by the CLasSel ANR project ANR-08-EMER-002 and the BLUE-ETS project G.A. n.244767.

References

[1] L. Lebart, A. Morineau, K. Warwick, Multivariate Descriptive Statistical Analysis, J. Wiley, 1984.

[2] J. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer, 2007.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (1990) 391–407.

[4] E. Bingham, H. Mannila, Random projection in dimensionality reduction: Applications to image and text data, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '01, ACM, New York, NY, USA, 2001, pp. 245–250.

[5] T. Liu, S. Liu, Z. Chen, W.-Y. Ma, An evaluation on feature selection for text clustering., in: T. Fawcett, N. Mishra (Eds.), ICML, AAAI Press, 2003, pp. 488–495.

[6] A. K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognition Letters 31 (8) (2010) 651–666.

[7] J. A. Hartigan, Direct Clustering of a Data Matrix, Journal of the American Statistical Association 67 (337) (1972) 123–129.

[8] H. Bock, Simultaneous clustering of objects and variables, in: E. Diday (Ed.), Analyse des Données et Informatique, INRIA, 1979, pp. 187–203.

[9] G. Govaert, Simultaneous clustering of rows and columns, Control and Cybernetics 24 (4) (1995) 437–458.

[10] I. S. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '01, ACM, New York, NY, USA, 2001, pp. 269–274.

[11] I. S. Dhillon, S. Mallela, D. S. Modha, Information-theoretic co-clustering, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, ACM, New York, NY, USA, 2003, pp. 89–98.

[12] G. Govaert, M. Nadif, Clustering with block mixture models, Pattern Recognition 36 (2) (2003) 463–473.

[13] I. V. Mechelen, H. H. Bock, P. D. Boeck, Two-mode clustering methods: a structured overview, Statistical methods in medical research 13 (5) (2004) 363–394.

[14] S. C. Madeira, A. L. Oliveira, Biclustering algorithms for biological data analysis: A survey, IEEE/ACM Trans. Comput. Biol. Bioinformatics 1 (1) (2004) 24–45.

[15] M. Charrad, M. Ben Ahmed, Simultaneous clustering: A survey, in: PReMI 2011, LNCS 6744, 2011, pp. 370–375.

[16] Y. Kluger, R. Basri, J. T. Chang, M. Gerstein, Spectral biclustering of microarray cancer data: Co-clustering genes and conditions, Genome Research 13 (2003) 703–716.

[17] M. Cottrell, S. Ibbou, P. Letremy, Som-based algorithms for qualitative variables, Neural Networks 17 (8-9) (2004) 1149–1167.

[18] T. Hoang, M. Olteanu, Som biclustering - coupled self-organizing maps for the biclustering of microarray data, in: IDAMAP 03, Workshop notes, 2003, pp. 40–46.

[19] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler, A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics 22 (9) (2006) 1122–1129.

[20] M. Brameier, C. Wiuf, Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae using self-organizing maps, Journal of Biomedical Informatics 40 (2) (2007) 160 – 173.

[21] G. Cabanes, Y. Bennani, D. Fresneau, 2012 special issue: Enriched topological learning for cluster detection and visualization, Neural Netw. 32 (2012) 186–195.

[22] K. Benabdeslem, K. Allab, Bi-clustering continuous data with self-organizing map., Neural Computing and Applications 22 (7-8) (2013) 1551–1562.

[23] A. Chaibi, M. Lebbah, H. Azzag, A new bi-clustering approach using topological maps, in: Neural Networks (IJCNN), The 2013 International Joint Conference on, 2013, pp. 1–7.

[24] T. Sarazin, M. Lebbah, H. Azzag, A. Chaibi, Feature group weighting and topological biclustering, in: 21st International Conference, ICONIP, 2014, pp. 369–376.

[25] Y. Pang, S. Wang, Y. Yuan, Learning regularized lda by clustering, Neural Networks and Learning Systems, IEEE Transactions on 25 (12) (2014) 2191–2201.

[26] L. Yengo, J. Jacques, C. Biernacki, Variable clustering in high dimensional linear regression models, Journal de la Société Française de Statistique 155 (2) (2014) 38–56.

[27] T. Kohonen, Self-organizing maps, Springer, 1997.

[28] T. Heskes, Self-organizing maps, vector quantization, and mixture modeling, Neural Networks, IEEE Transactions on 12 (6) (2001) 1299–1305.

[29] C. Ambroise, G. Govaert, Constrained clustering and kohonen self-organizing maps, Journal of Classification 13 (2) (1996) 299–313.

[30] M. M. Van Hulle, Kernel-Based Topographic Maps: Theory and Applications, John Wiley & Sons, Inc., 2007.

[31] C. M. Bishop, M. Svensén, C. K. I. Williams, GTM: A principled alternative to the self-organizing map, in: M. C. Mozer, M. I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems 9, The MIT Press, Cambridge, MA, 1997, pp. 354–360.

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

[33] P. Tino, I. Nabney, Hierarchical gtm: constructing localized non-linear projection manifolds in a principled way, IEEE transactions on pattern analysis and machine intelligence 24 (5) (2002) 639–656.

[34] A. Vellido, Assessment of an unsupervised feature selection method for generative topographic mapping, in: Artificial Neural Networks ICANN 2006, Vol. 4132 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2006, pp. 361–370.

[35] D. Maniyar, I. Nabney, Data visualization with simultaneous feature selection, in: Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB '06. 2006 IEEE Symposium on, 2006, pp. 1–8.

[36] G. J. McLachlan, D. Peel, Finite Mixture Models, John Wiley and Sons, New York, 2000.

[37] G. Govaert, M. Nadif, Block clustering with Bernoulli mixture models: Comparison of different approaches, Computational Statistics and Data Analysis 52 (6) (2008) 3233–3245.

[38] M. Govaert, M. Nadif, Co-clustering: models, algorithms and applications, Wiley ISTE, 2013.

[39] P. McCullagh, J. Nelder, Generalized linear models, London: Chapman and Hall, 1983.

[40] R. Salakhutdinov, A. Mnih, Probabilistic matrix factorization, in: Advances in Neural Information Processing Systems, Vol. 20, 2008.

[41] H. Ma, H. Yang, M. R. Lyu, I. King, Sorec: Social recommendation using probabilistic matrix factorization, in: CIKM'08, Napa Valley, California, USA, 2008.

[42] M. Collins, S. Dasgupta, R. E. Schapire, A Generalization of Principal Components Analysis to the Exponential Family, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, MIT Press, 2002.

[43] R. Priam, M. Nadif, G. Govaert, The block generative topographic mapping, in: ANNPR, Vol. 5064 of Lecture Notes in Computer Science, Springer, 2008, pp. 13–23.

[44] R. Priam, M. Nadif, G. Govaert, Nonlinear mapping by constrained co-clustering, in: ICPRAM'2012, 2012, pp. 63–68.

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

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

[47] M. Nadif, G. Govaert, Model-based co-clustering for continuous data, in: ICMLA, IEEE Computer Society, 2010, pp. 175–180.

[48] A. Dempster, N. Laird, D. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, J. Royal Statist. Soc. Ser. B., 39 (1977) 1–38.

[49] D. Bohning, B. Lindsay, Monotonicity of quadratic-approximation algorithms, Annals of the Institute of Statistical Mathematics 40 (4) (1988) 641–663.

[50] D. L. Davies, D. W. Bouldin, A cluster separation measure, Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1 (2) (1979) 224 –227.

[51] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[52] R. Priam, CASOM: Som for contingency tables and biplot, in: 5th Workshop on Self-Organizing Maps (WSOM'05), 2005, pp. 379–385.

[53] R. Priam, M. Nadif, G. Govaert, Topographic bernoulli block mixture mapping for binary tables, Pattern Analysis and Applications 17 (4) (2014) 839–847.

[54] Y. Chen, L. Wang, M. Dong, J. Hua, Exemplar-based visualization of large document corpus (infovis2009-1115), IEEE Transactions on Visualization and Computer Graphics 15 (2009) 1161–1168.

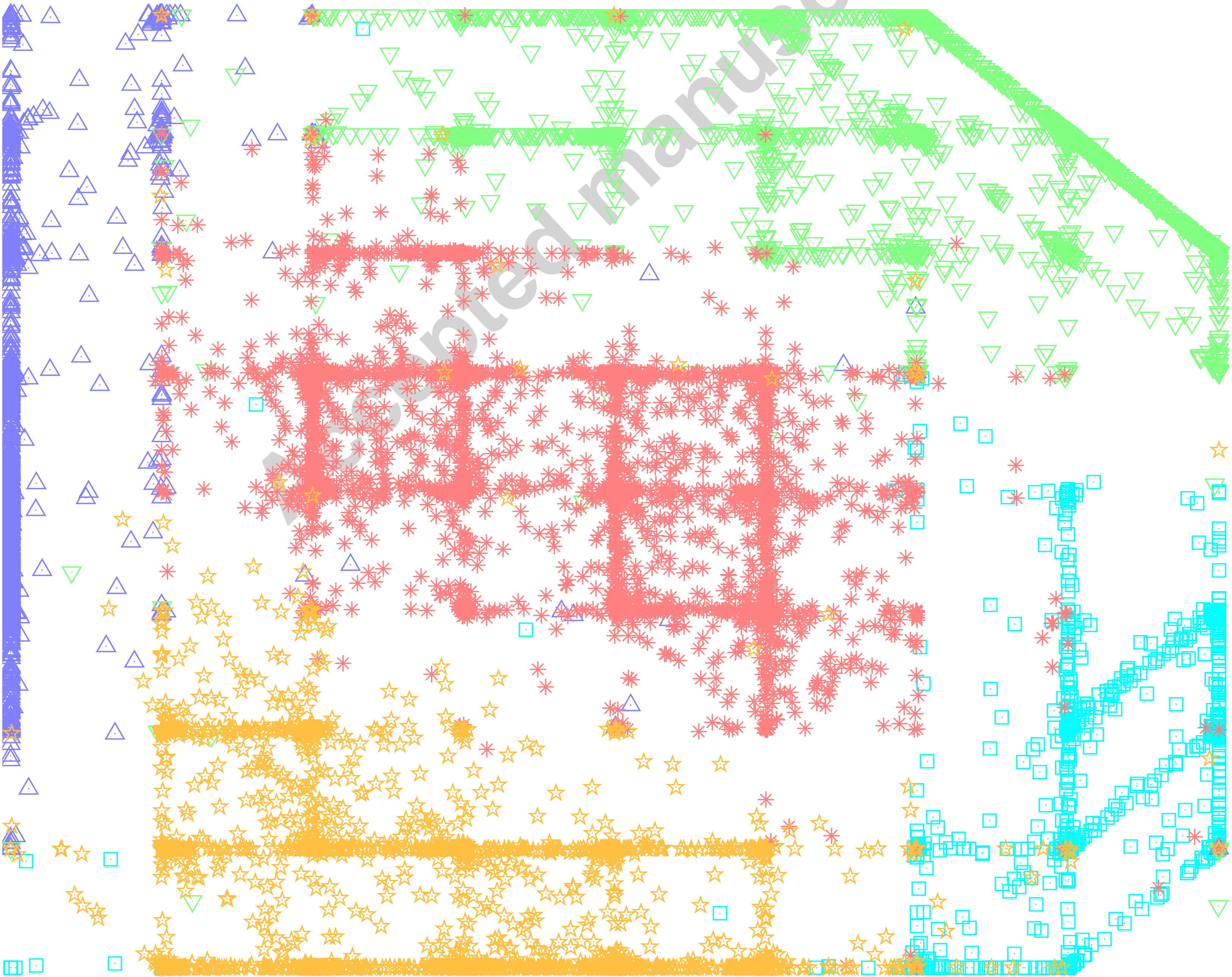
[55] A. M. Newman, J. B. Cooper, Autosome: a clustering method for identifying gene expression modules without prior knowledge of cluster number, BMC Bioinformatics 11 (2010) 117.

[56] D. Brugger, M. Bogdan, W. Rosenstiel, Automatic cluster detection in kohonen's som, Neural Networks, IEEE Transactions on 19 (3) (2008) 442–459.

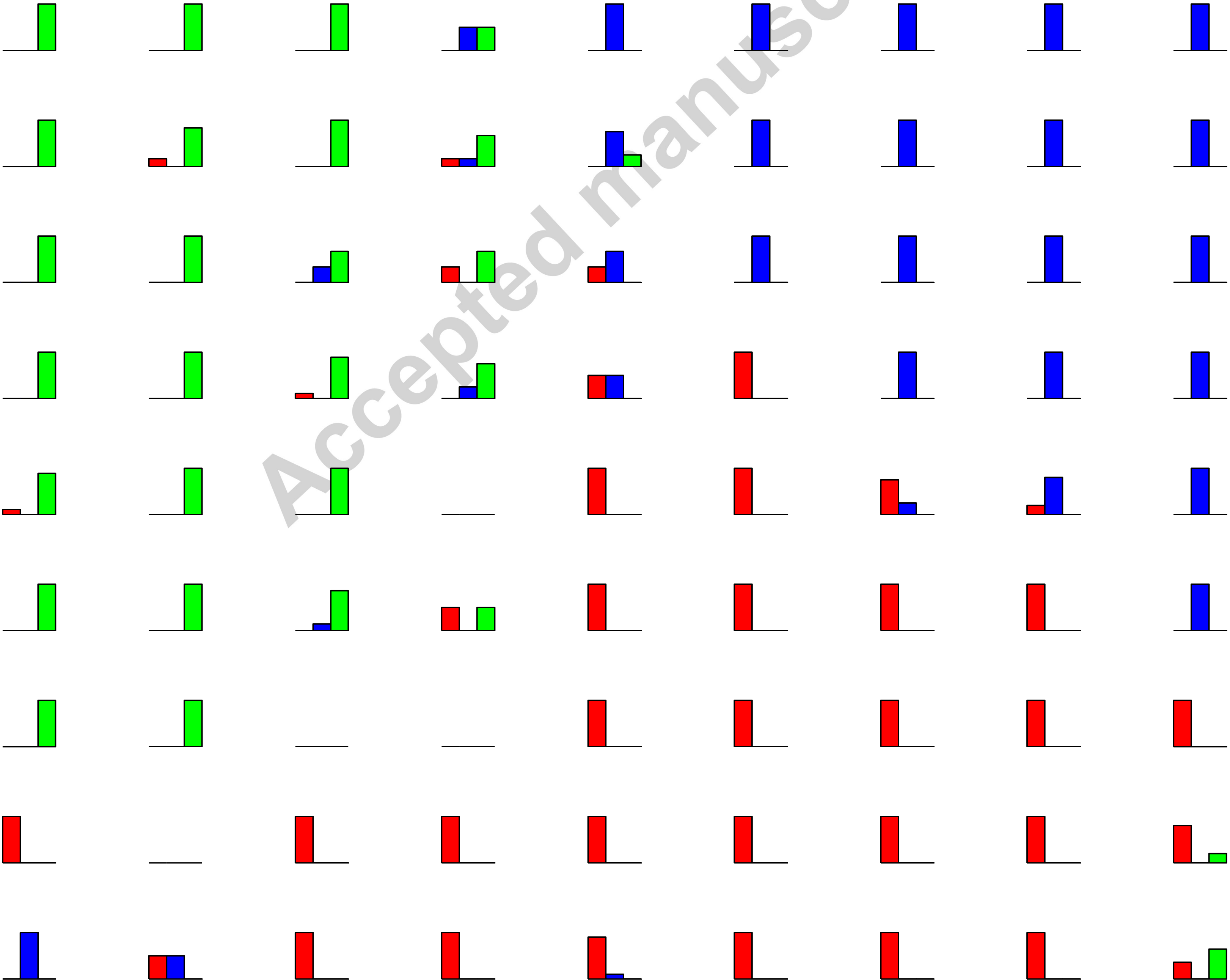
[57] Y. Pang, Z. Ji, P. Jing, X. Li, Ranking graph embedding for learning to rerank, Neural Networks and Learning Systems, IEEE Transactions on 24 (8) (2013) 1292–1303.

[58] J. Shen, J. Bu, B. Ju, T. Jiang, H. Wu, L. Li, Refining gaussian mixture model based on enhanced manifold learning, Neurocomputing 87 (2012) 19–25.

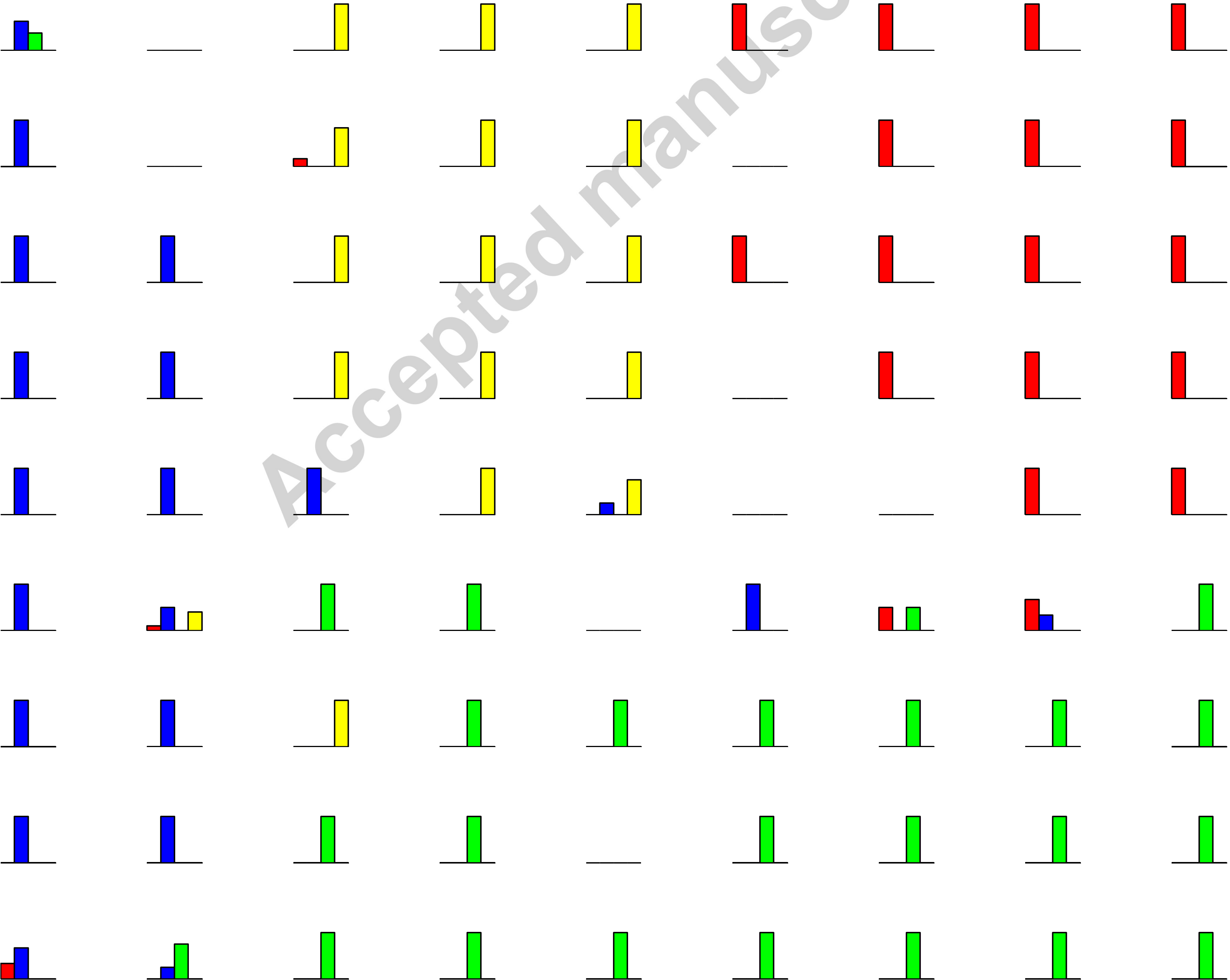
Figure



Figure



Figure



Rodolphe Priam has been assistant professor in Nantes and associate professor in Poitiers from 2004 to 2008, and research assistant at Southampton university from 2009. His current research interests are surveys and data mining.
Mohamed Nadif is full Professor of the University of Paris Descartes. He received his HDR in computer science in 2004 from the University of Metz. His current research research interests are in data mining, mixture models, cluster analysis, missing data and visualization.
Gérard Govaert is Professor of the University of Technology of Compiègne and researcher at the CNRS Laboratory Heudiasyc (Heuristic and diagnostic of complex systems). He received his "thèse d'Etat" in computer science in 1983 from the University Pierre et Marie Curie, Paris. His current research interests include cluster analysis, statistical pattern recognition and spatial data analysis.

*Photo of the author(s)

Click here to download high resolution image

*Photo of the author(s)