Bayesian Gaussian Topographic Block Model

***Abstract*— Generative topographic mapping (GTM) is intensively studied for visualization of data tables with latent nonlinearities. This model is a probabilistic version of the artificial neural network called Kohonen's map or SOM. If high dimensional spaces can be reduced via pre-processing by an ad hoc reduction before training GTM or SOM, a clustering of the variables is a viable alternative for data tables with a block structure. A simultaneous clustering of the rows and the columns brings parsimony for the parameters in subspace modeling. Herein, the latent block mixture model is considered as a foundation for a new generative self-organizing map. We define the proposed new model called GBGTM for Gaussian data, its inference via the Block EM (BEM) algorithm. We present then its Bayesian variant with a hierarchical sparse prior and a variational EM algorithm for the inference. We illustrate empirically the behavior of the approach for the visualization two real biological datasets. We conclude with a summary of the contribution and perspectives.**

***Index Terms*— Latent block mixture model, Generative topographic mapping, Block expectation maximization, Variational expectation maximization, Hierarchical sparse prior.**

I. Introduction

In data analysis, the reduction of the dimensions of a numerical data table leads to synthetic and understandable representations in a low dimensional space. When the data table is large, a clustering might lead to a quicker and easier access to the data contents in comparison with a method which reduces the dimensionality by a projection. Combining clustering and reduction is often more informative. In visualization, it may be shown the data by points on a two dimensional map, and via different colors or symbols a clustering which is obtained from an additional method. Note that one can prefer for a large dataset to model the projection of row (resp. column) clusters rather than the (resp. column) row. But most of the time, mapping and clustering operate separately [3], and one has to decide which method must be devoted to each problem. On the contrary, it exists a family of methods which solves them simultaneously. In data visualization, the Kohonen's maps and more generally self-organizing maps (SOM) [1] are modified clustering methods. They are able to induce a strong relation of vicinity between the clusters and lead to a visualization from a latent nonlinear surface which is relevant for high dimensional spaces. They are generalizations to the principal component analysis (PCA) [2] and its linear planes. Some modified versions are able to provide more suitable results in particular situations like the analysis of discrete data. Moreover, a parametric model is very flexible and even scalable when it is defined and trained properly. Hence, a probabilistic model for SOM is interesting for these diverse reasons. The Generative Topographic Mapping (GTM) [4], [5] is a generative SOM with a more restricted set of parameter values than the Kohonen's map. It is intensively studied and improved ([6], [7], [8], [9]) recently for data analysis purposes. It formulates a self-organizing map by adding the constraints of vicinity between the clusters at the level of the expectations of a Gaussian mixture model (GMM) [10].

For a large amount of variables, co-clustering is a powerful approach in data analysis because it enables a reduction of the variables space simultaneously to the row clustering. This familly of methods is useful in many domains which need the analysis of a large amount of data such as often met today. For all these reasons, we propose an extension of GTM with the help of a co-clustering model, the latent block mixture model (LBM) [12] for data tables which have a block structure. Thus, our main contribution is to bring column clustering to a subspace method via probabilistic co-clustering: LBM is considered as a foundation for a new generative self-organizing map. The paper is organized as follows. In section 2, we review LBM with a Gaussian setting, and the related objective function to optimize for the parameters estimation. In sections $3-4$, we add the constraints in this model and we propose a learning algorithm. Two different approaches for the modeling of the parameters are considered: a first method named GBGTM is a Topographic LBM with a constant ℓ_2-norm penalization and a second one named GBGTM$_B$ is its Bayesian version with an adaptative ℓ_1-norm penalization for automatic sparsity constraints. In section 5 we present the numerical experiments for the comparison with GTM. Finally, in section 6 we summarize the contribution and perspectives.

II. Parameters estimation in LBM

Let's denote $\mathbf{x} = (x_{11}, x_{12}, \ldots, x_{ij}, \ldots, x_{nd})$, the data table in a latent block model. I is the set of the rows and J is the set of the columns. A possible assignment of I is modeled with the binary classification matrix $\mathbf{z} = (z_{ik})_{n\times g}$, it is such that $z_{ik} = 1$ indicates the component of the row i, and $\sum_{k=1}^{g} z_{ik} = 1$. Similarly for the assignments of J it is denoted the binary matrix $\mathbf{w} = (w_{j\ell})_{d\times m}$. The two sets of possible assignments $\mathbf{w}$ and $\mathbf{z}$ partition the cells of the table $\mathbf{x}$ into a number of contiguous, non-overlapping blocks. For a latent block model, the $n \times d$ random variables which generate the observed cells x_{ij} of the data table are assumed to be independent, once $\mathbf{z}$ and $\mathbf{w}$ are fixed, they permit 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.

A. Latent block model

In LBM, the completed data are taken to be the vector $(\mathbf{x}, \mathbf{z}, \mathbf{w})$ where the latent vectors $\mathbf{z}$ and $\mathbf{w}$ are the random labels for the rows and the columns respectively. The classification log-likelihood can then be written:

$$L_C(\mathbf{z}, \mathbf{w}; \mathbf{x}, \boldsymbol{\theta}) = \log P(\mathbf{x}|\mathbf{z}, \mathbf{w}) + \log P(\mathbf{z}) + \log P(\mathbf{w})\,.$$

The different density and mass functions are:

$$\begin{array}{lcl} P(\mathbf{z}) & = & \prod_{i,k} (p_k)^{z_{ik}} \\ P(\mathbf{w}) & = & \prod_{j,\ell} (q_\ell)^{w_{j\ell}} \\ P(\mathbf{x}|\mathbf{z},\mathbf{w}) & = & \prod_{i,j,k,\ell} \left(\frac{1}{\sqrt{2\pi\sigma_{k\ell}^2}} e^{-\frac{(x_{ij}-\mu_{k\ell})^2}{2\sigma_{k\ell}^2}} \right)^{z_{ik}w_{j\ell}} \end{array} .$$

The set of parameters is denoted $\boldsymbol{\theta} = (\mathbf{p}, \mathbf{q}, \boldsymbol{\alpha})$. Here, the probabilities p_k (resp. q_ℓ) that a row (resp. a column) belongs to the k^{th} component (resp. ℓ^{th} component) are aggregated in $\mathbf{p} = (p_1, \ldots, p_g)$ (resp. $\mathbf{q} = (q_1, \ldots, q_m)$), while $\boldsymbol{\alpha}$ aggregates the parameters from all the p.d.f. of each block, say $\alpha_{k\ell}$. As in the Gaussian case of the non symmetric co-clustering model [13], the p.d.f. for the block $(k\ell)$ is such as $\alpha_{k\ell} = (\mu_{k\ell}, \sigma_{k\ell}^2)$ with the mean $\mu_{k\ell}$ and the variance $\sigma_{k\ell}^2$. The model is called GLBM with aggregating matrices $\boldsymbol{\Sigma} = (\sigma_{k\ell}^2)_{g\times m}$ and $\boldsymbol{\mu} = (\mu_{k\ell})_{g\times m}$.

The probability density function (p.d.f.) of a latent block model is defined as the following decomposition. It is obtained by independence of $\mathbf{z}$ and $\mathbf{w}$, by summing over all the assignments [12]:

$$f_{LBM}(\mathbf{x};\boldsymbol{\theta}) = \textstyle\sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} P(\mathbf{x}|\mathbf{z},\mathbf{w})P(\mathbf{z})P(\mathbf{w}) . \quad (1)$$

The set of all the possible assignments is denoted $\mathcal{Z}$ for I and $\mathcal{W}$ for J. The log-likelihood is as follows:

$$L(\mathbf{x};\boldsymbol{\theta}) = \log f_{LBM}(\mathbf{x};\boldsymbol{\theta}). \quad (2)$$

The block model is dramatically more parsimonious than the usual mixture model where each dimension of the data table is modeled separately. Next paragraph, we review the criterion and the algorithm for an estimation of the parameters.

B. Objective function, EM and BEM

For co-clustering, we aim to address the problem of parameters estimation by a maximum likelihood (ML) approach such that the log-likelihood L is maximized by:

$$\hat{\boldsymbol{\theta}} = argmax_{\boldsymbol{\theta}} L(\mathbf{x};\boldsymbol{\theta}) . \quad (3)$$

The log-likelihood of the completed data is considered for the inference by taking benefit of the introduced latent variables for the labelling of the cells. As EM [14] is intractable for the latent block model, the approach based on a generalized EM and a variational approximation has been proposed in [15] and named *Block EM* (BEM). This algorithm proceeds by maximizing a surrogate objective function which lower bounds L and converges towards a final solution which maximizes (locally) the log-likelihood of the latent block model.

In the following sections, we explain how to extend the latent block model and its algorithm in order to construct visual bidimensional maps from data samples within this probabilistic co-clustering framework.

III. Parsimonious topographic mapping

The parameters $\mu_{k\ell}$ in the co-clustering model are re-parameterized with two sets of vectors, one for the dimension k, and one for the dimension ℓ. By this way, we define a general model for a parsimonious parametric self-organizing map in order to map the rows of a numerical tables with a large number of columns into the plane. The method is named Gaussian Topographic latent block model or Block GTM (GBGTM).

A. Parameter transformation

It is defined a set of constant bivariate vectors s_k, the coordinates of the nodes of an had hoc regular rectangular planar mesh. The coordinates of these nodes can be written for all k as follows:

$$\boldsymbol{s}_k = \begin{pmatrix} s_{k_1} \\ s_{k_2} \end{pmatrix} .$$

These nodes have a similar role than in the algorithm of Kohonen's maps. They lead to a discretization of the latent squared space where the data are projected. By attaching the parameters of one row cluster to the coordinates of one node, this may induce the wanted constraints for the self-organization of the row clusters. For nontrivial problems, the data cloud has a complex shape and higher dimensions than two are required for the modeling. Each coordinate $\boldsymbol{s}_k$ is projected into a space of h dimensions with the help of well defined basis functions ϕ^s. Thus, ϕ_k is written for all k as follows:

$$\left[1, s_k^T, e^{-\frac{||\boldsymbol{s}_k - \mu_{\phi^1}||^2}{2\times\left(\nu_{\phi^1}\right)^2}}, e^{-\frac{||\boldsymbol{s}_k - \mu_{\phi^2}||^2}{2\times\left(\nu_{\phi^2}\right)^2}}, \cdots, e^{-\frac{||\boldsymbol{s}_k - \mu_{\phi^{h-3}}||^2}{2\times\left(\nu_{\phi^{h-3}}\right)^2}} \right]^T .$$

Here, it is introduced an eventual intercept term, the coordinates s_k, and typical kernel Gaussian functions ϕ^s with mean centers $\mu_{\phi^s} \in \mathbb{R}^2$ and standard deviations $\nu_{\phi^s} \in \mathbb{R}_*$. Note that the first component induces an additive effect from the column clusters in each central block parameter and similary an effect from the row clusters might be added if required.

Finally, m latent vectors related to the clustering of the columns are denoted $w_\ell \in \mathbb{R}^h$ for all ℓ. They are aggregated in the matrix $\boldsymbol{\Omega} = [w_1|w_2|\cdots|w_m]$. Its estimation is required while the matrix for the nonlinearities is kept constant, $\boldsymbol{\Phi} = [\phi_1|\phi_2|\cdots|\phi_g]^T$. For modeling the dependence of each parameter $\mu_{k\ell}$ with ϕ_k and w_ℓ, it is considered their inner product such that:

$$\boldsymbol{\mu} = \boldsymbol{\Phi}\boldsymbol{\Omega} . \quad (4)$$

This re-parameterizes the latent block model for mapping the rows into a part of the plane. The reduced $g \times m$ parameter matrix $\boldsymbol{\mu}$ in the co-clustering model is replaced by the matrix $\boldsymbol{\Omega}$. The resulting model remains parsimonious because h is small, less than half of one hundred in general.

B. BEM and Gaussian Block GTM

For the Gaussian topographic block model, the set of parameters is $\boldsymbol{\theta} = (\boldsymbol{\Omega}, \boldsymbol{\Sigma}, \mathbf{p}, \mathbf{q})$. In this case, following the approach of BEM, the objective function can be written as follows:

$$\begin{array}{lcl} \tilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) & = & \sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \left[-\frac{(x_{ij}-\mu_{k\ell})^2}{2\sigma_{k\ell}^2} - \frac{\log(2\pi\sigma_{k\ell}^2)}{2} \right] \\ & - & \sum_{i,k} c_{ik}^{(t)} \log c_{ik}^{(t)} - \sum_{j,\ell} d_{j\ell}^{(t)} \log d_{j\ell}^{(t)} \\ & + & \sum_k c_k^{(t)} \log p_k + \sum_\ell d_\ell^{(t)} \log q_\ell . \end{array}$$

Hence when no regularization is in stake, it can be written:

$$\begin{aligned} w_\ell^{(t+1)} &= \tfrac{1}{d_\ell^{(t)}} \left(\mathbf{\Phi}^T \mathbf{\Upsilon}_\ell^{(t)} \boldsymbol{G}^{(t)} \mathbf{\Phi} \right)^{-1} \mathbf{\Phi}^T \mathbf{\Upsilon}_\ell^{(t)} \mathbf{y}_\ell^{(t)} ; \\ \mathbf{\Sigma}^{(t+1)} &= \left\{ \boldsymbol{C}^{(t)T} \mathbf{x} \odot \mathbf{x} \boldsymbol{D}^{(t)} - 2 \boldsymbol{C}^{(t)T} \mathbf{x} \boldsymbol{D}^{(t)} \odot \boldsymbol{\mu}^{(t)} \right. \\ &\quad \left. + \mathbf{c}^{(t)} \mathbf{d}^{(t)T} \odot \boldsymbol{\mu}^{(t)} \odot \boldsymbol{\mu}^{(t)} \right\} \oslash \mathbf{c}^{(t)} \mathbf{d}^{(t)T} . \end{aligned}$$

Here, the posterior probabilities c_{ik} and $d_{j\ell}$ are also written as in BEM (see [12], [15]) at time (t). The matrices $\boldsymbol{C}$, $\boldsymbol{D}$, and vectors $\mathbf{c}$, $\mathbf{d}$ stand respectively for $(c_{ik})_{n \times g}$, $(d_{j\ell})_{d \times m}$ $(c_k)_{g \times 1}$, $(d_\ell)_{m \times 1}$ where $c_k = \sum_i c_{ik}$ and $d_\ell = \sum_j d_{j\ell}$. Moreover $\boldsymbol{G}$ is a diagonal matrix with non null cells equal to the components of $\mathbf{c}$. For ℓ constant, $\mathbf{\Upsilon}_\ell^{(t)}$ is the diagonal matrix with non null cells equal to $\sigma_{k\ell}^{(t)-2}$, and $\mathbf{y}_\ell = (y_{k\ell})_{g \times 1}$ is a g-dimensional vector. Note also that a Gaussian Bayesian log-prior $-\frac{\lambda}{2}||w_\ell||^2$ with $\lambda \in \mathbb{R}^+$ can be added for the variables w_ℓ in order to regularize the parameters. A simple constant hyperparameter such as $\lambda = 0.001$ has been useful generally during our experiments with a classical Newton-Raphson procedure for increasing the likelihood but more advanced approaches might be considered (see [8]).

Next subsection, the model is extended with a Bayesian approach for a sparse regularization of the loading matrix.

IV. Variational learning

A Bayesian model is defined for extending BGTM to a more parsimonious setting and cancel out some component values of the vectors w_ℓ, it is called GBGTM$_{\mathbf{B}}$. Let's denote $\tilde{\boldsymbol{\theta}} = \{\mathbf{z}, \mathbf{w}, \mathbf{\Omega}, \boldsymbol{\beta}, \boldsymbol{\tau}\}$ for the random variables which replace the previous deterministic parameters, and also $P(\boldsymbol{x}; \tilde{\boldsymbol{\theta}})$ for the new conditional distribution involved in the sample modeling. By definition, we have:

$$P(\tilde{\boldsymbol{\theta}}) = P(\mathbf{\Omega}|\boldsymbol{\beta}) P(\boldsymbol{\beta}) P(\mathbf{z}) P(\mathbf{w}) P(\boldsymbol{\tau}) .$$

The matrix $\mathbf{\Omega}$ is now random as explained next subsections while $\boldsymbol{\beta}$ is a random parameter from a hierarchical prior for inducing the sparsity. The matrix $\boldsymbol{\tau}$ aggregates the inverses of the variance parameters $\sigma_{k\ell}^2$. Note that the distribution of the block model is recognized with a specific parameterization for handling the nonlinear mapping and its regularization. Hence, the new parameters can be found through the maximization of the marginal distribution of the observations:

$$L_{GBGTM_B}(\mathbf{x}) = \log \textstyle\int P(\boldsymbol{x}; \tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}} .$$

Following the extensive literature on Bayesian variational learning (see [16]), it can be written with this approach:

$$L_{GBGTM_B}(\mathbf{x}) \geq \mathbb{E}_{Q_{\boldsymbol{\xi}}(\tilde{\boldsymbol{\theta}})} \left[\log \tfrac{P(\boldsymbol{x}; \tilde{\boldsymbol{\theta}})}{Q_{\boldsymbol{\xi}}(\tilde{\boldsymbol{\theta}})} \right] . \quad (5)$$

The right member in (5) is the new function involved in the variational maximization, $\mathcal{F}(Q_{\boldsymbol{\xi}}; \tilde{\boldsymbol{\theta}})$ with an expectation w.r.t. the distribution $Q_{\boldsymbol{\xi}}$ with deterministic parameters $\boldsymbol{\xi}$. Next, we define precisely the new log-likelihood, its approximation $Q_{\boldsymbol{\xi}}$ with parameters $\boldsymbol{\xi}$, and a numerical procedure for the inference of the involved variational parameters.

A. Completed log-likelihood

Let's have $\mathcal{N}(y; \mu, \Sigma)$ for y distributed as a Gaussian p.d.f. of mean μ and variance Σ, while $\mathcal{G}(y; d_0, c_0)$ is for y distributed as a Gamma distribution with two real parameters d_0 and c_0. Then, a hierarchical prior is defined from Gaussian and Gamma distributions for a penalization in the new completed log-likelihood which can be written:

$$\begin{aligned} \tilde{L}_C(\tilde{\boldsymbol{\theta}}; \mathbf{x}) &= L_C(\mathbf{z}, \mathbf{w}; \mathbf{x}, \boldsymbol{\theta}) + \log P(\boldsymbol{\tau}) \\ &+ \log P(\mathbf{\Omega}|\boldsymbol{\beta}) + \log P(\boldsymbol{\beta}) , \end{aligned}$$

where,

$$\begin{aligned} P(\mathbf{\Omega}|\boldsymbol{\beta}) &= \textstyle\prod_{s\ell} \mathcal{N}(w_{s\ell}; 0, \beta_{s\ell}^{-1}) \\ P(\boldsymbol{\beta}) &= \textstyle\prod_{s\ell} \mathcal{G}(\beta_{s\ell}; d_{\beta_s^0}, c_{\beta_s^0}) \\ P(\boldsymbol{\tau}) &= \textstyle\prod_{s\ell} \mathcal{G}(\tau_{s\ell}; d_{\tau_s^0}, c_{\tau_s^0}) . \end{aligned}$$

Here, $\boldsymbol{\beta} = (\beta_{s\ell})_{h \times m}$, $w_{s\ell}$ is a cell of $\mathbf{\Omega}$, while $\beta_{s\ell}^{-1}$ is a variance parameter. For large values of $\beta_{s\ell}$, the corresponding cells in $\mathbf{\Omega}$ might cancel out, leading to a parsimonious matrix as expected. The parameter $\boldsymbol{\theta}$ stands for the set of parameters of GBGTM but $\mathbf{\Omega}$ and $\mathbf{\Sigma}$ are now random variables. The completed log-likelihood for the proposed Bayesian model is the written:

$$\begin{aligned} & \tilde{L}_C(\tilde{\boldsymbol{\theta}}; \mathbf{x}) \\ = & \textstyle\sum_{ijk\ell} z_{ik} w_{j\ell} \left\{ -\frac{\tau_{k\ell}}{2} (x_{ij} - w_\ell^T \phi_k)^2 + \frac{1}{2} \ln(\frac{\tau_{k\ell}}{2\pi}) \right\} \\ + & \textstyle\sum_{s,\ell}^{h,m} \left\{ -\frac{1}{2} \beta_{s\ell} w_{s\ell}^2 - \frac{1}{2} \ln(2\pi) + \frac{1}{2} \ln(\beta_{s\ell}) \right\} \\ + & \textstyle\sum_{s,\ell}^{h,m} \left\{ d_{\beta_0} \ln s_{\beta_0} + (d_{\beta_0} - 1) \ln \beta_{s\ell} - s_{\beta_0} \beta_{s\ell} \right\} \\ + & \textstyle\sum_{k,\ell}^{g,m} \left\{ d_{\tau_0} \ln s_{\tau_0} + (d_{\tau_0} - 1) \ln \tau_{k\ell} - s_{\tau_0} \tau_{k\ell} \right\} \\ + & \textstyle\sum_{i,k}^{n,g} z_{ik} \ln(p_k) + \sum_{j,\ell}^{d,m} w_{j\ell} \ln(q_\ell) \\ - & h\, m\, \ln \Gamma(d_{\beta_0}) - g\, m\, \ln \Gamma(d_{\tau_0}) . \end{aligned}$$

The components of $\tilde{\boldsymbol{\theta}}$ defined for the ℓ_1-norm regularization $-\lambda |w_\ell|$ with $\lambda \in \mathbb{R}^+$ are independent and the hierarchical setting can induce parsimony [17], while the hyperparameters $(d_{\beta_s^0})$ and $(c_{\beta_s^0})$ remain constant (for instance 10^{-3}). Note that a related hierarchical prior has been proposed recently (see [18] and [9] for instance) for GTM.

B. Variational distributions

For the distributions defined in the variational bound, the law of $(\mathbf{z}, \mathbf{w})$ remains identical to BEM. The different variables in $\tilde{\boldsymbol{\theta}}$ are supposed independent such that the related density $Q_{\boldsymbol{\xi}}$ is a product. The random variables in $\tilde{\boldsymbol{\theta}}$ have related distributions as defined just as follows,

$$\begin{aligned} Q(\mathbf{z}) &= \textstyle\prod_{i,k} (c_{ik})^{z_{ik}} \\ Q(\mathbf{w}) &= \textstyle\prod_{j,\ell} (d_{j\ell})^{w_{j\ell}} \\ Q(\mathbf{\Omega}) &= \textstyle\prod_\ell \mathcal{N}(w_\ell; \mu_\ell, S_\ell) \\ Q(\boldsymbol{\beta}) &= \textstyle\prod_{\ell,s} \mathcal{G}(\beta_{s\ell}; d_{\beta_{s\ell}}, c_{\beta_{s\ell}}) \\ Q(\boldsymbol{\tau}) &= \textstyle\prod_{\ell,s} \mathcal{G}(\tau_{s\ell}; d_{\tau_{s\ell}}, c_{\tau_{s\ell}}) \end{aligned}$$

Hence, $\boldsymbol{\xi}$ is compound of all the parameters of these distributions. Note also that $\sigma_{S,s\ell}$ stands for the s^{th} diagonal component of S_ℓ while $\mu_{s\ell}$ is the s^{th} component of μ_ℓ. The distribution of the variational parameters in $Q_{\boldsymbol{\xi}}$ are found by maximizing the free energy $\mathcal{F}$ after marginalizing out other corresponding random variables in $\tilde{\boldsymbol{\theta}}$. An algorithm for the inference of the parameters is presented next paragraphs.

$$
\begin{aligned}
c_{ik} &\propto p_k\, e^{\left\{\frac{-1}{2}\sum_{j\ell} d_{j\ell}\tilde{\tau}_{k\ell}(x_{ij}-\mu_\ell^T\phi_k)^2-\frac{1}{2}\sum_\ell d_\ell\tilde{\tau}_{k\ell}\phi_k^T S_\ell\phi_k\right\}} \\
&\quad \times e^{\left\{\frac{1}{2}\sum_\ell d_\ell[\psi(d_{\tau_{k\ell}})-\ln(s_{\tau_{k\ell}})]\right\}}\ ; \\
d_{j\ell} &\propto q_\ell\, e^{\left\{\frac{-1}{2}\sum_{ik} c_{ik}\tilde{\tau}_{k\ell}(x_{ij}-\mu_\ell^T\phi_k)^2-\frac{1}{2}\sum_k c_k\tilde{\tau}_{k\ell}\phi_k^T S_\ell\phi_k\right\}} \\
&\quad \times e^{\left\{\frac{1}{2}\sum_k c_k[\psi(d_{\tau_{k\ell}})-\ln(s_{\tau_{k\ell}})]\right\}}\ ; \\
\tilde{\tau}_{k\ell} &= \frac{d_{\tau_0}+\frac{c_k d_\ell}{2}}{s_{\tau_0}+\frac{1}{2}\sum_{ij} c_{ik}d_{j\ell}[(x_{ij}-\mu_{s\ell}^T\phi_k)^2+\phi_k^T S_\ell\phi_k]}\ ; \\
\tilde{\beta}_{s\ell} &= \frac{d_{\beta_s^0}+\frac{1}{2}}{c_{\beta_s^0}+\frac{1}{2}\left(\mu_{s\ell}^2+\sigma_{S,s\ell}^2\right)}\ ; \\
S_\ell &= \left\{d_\ell\Phi^T\Upsilon_\ell G\Phi+\Lambda_{\beta,\ell}\right\}^{-1}\ ; \\
\mu_\ell &= S_\ell\Phi^T\Upsilon_\ell y_\ell\,.
\end{aligned}
$$

Fig. 1. Updates in the variational EM Algorithm for GBGTM$_B$ at time (t).

C. Detail of the algorithm for GBGTM$_B$

Let's have Υ_ℓ a diagonal matrix with non null elements equal the expectations of $\tau_{k\ell}$ where ℓ is constant. Let's $\Lambda_{\boldsymbol{\beta},\ell}$ stands for a regularizing diagonal matrix with its non elements equal to the expectation of $\beta_{k\ell}$. Moreover, $\tilde{\tau}_{k\ell} = d_{\tau_{k\ell}}/s_{\tau_{k\ell}}$ is the expectation of $\tau_{k\ell}$ while $\tilde{\beta}_{s\ell} = d_{\beta_{s\ell}}/s_{\beta_{s\ell}}$ is the expectation for $\beta_{s\ell}$. The parameters and the required intermediate quantities in the approximating distribution $Q_{\boldsymbol{\xi}}$ from the variational bound are updated as given in Figure 1. The mixing probabilities may be also updated, as in BEM as no prior were introduced for them:

$$p_k = \frac{1}{n}\sum_k c_{ik} \text{ and } q_\ell = \frac{1}{d}\sum_\ell d_{j\ell}\,.$$

A completly full Bayesian setting would model also these probabilities with typically a Dirichlet distribution which can be considered as an important perspective and might be useful for model selection which is out of the scope of the paper.

Note that in Figure 1 the two first quantities are updated jointly in a E-step while the remaining ones are updated with $\mathbf{p}$ and $\mathbf{q}$ just after in a M-step. For the variational parameters related to $\boldsymbol{\Omega}$, the presented updates might have been obtained by a Laplace approximation around the solution from the previous step if the model was not linear and Gaussian. The stopping rule may be the relative difference between two consecutive values of the objective fonction. Finally, the procedure converges towards a solution for the variational distributions with:

$$\hat{\boldsymbol{\xi}} = (\{\hat{\mu}_\ell\}\{\hat{S}_\ell\}, \{\hat{d}_{\beta_{s\ell}}\}, \{\hat{c}_{\beta_{s\ell}}\}, \{\hat{d}_{\tau_{k\ell}}\}, \{\hat{c}_{\tau_{k\ell}})\}, \{\hat{c}_{ik}\}, \{\hat{d}_{j\ell}\})\,,$$

the estimated values of the variational parameters. The final mixing probabilities are also obtained as $\{\hat{p}_k\}$ and $\{\hat{q}_\ell\}$.

V. Experiments

In this section, we are interested on the comparison of GTM and GBGTM. Experiments with real datasets illustrate the interest of our approach for a parsimonious and sparse GTM.

A. Post-treatment and experimental settings

Let's define $\hat{z}_i = argmax_k \hat{c}_{ik}$ such that a row $i \in I$ might have a point representative with coordinates $\hat{\boldsymbol{s}}_i^{MAP} = \boldsymbol{s}_{\hat{z}_i}$. Alternatively, a continuous projection my be obtained as usually in GTM by the average positions:

$$\hat{\boldsymbol{s}}_i = \sum_{k=1}^{g} \hat{c}_{ik}\boldsymbol{s}_k\,.$$

From the estimated labels $\hat{z}_i$, the error rate is the percentage of missclassified data when each node is labelled by majority vote, denoted ER and the extended error rate to nearest neighboor nodes, denoted ED [1]. From the estimated coordinates $\{\hat{s}_i\}$, it can be computed other indicators in order to reveal the quality of the projection onto the plane, the average of the Silhouettes [20] denoted SN and the average of the percentage of nearest neighboors from the same class and computed according to the graph in the original space with $K = 7$ neighboors [21], denoted CN.

B. Output for a biological data table ($Data1$)

In bioinformatics, the Kohonen's map is useful [11] for the analysis of microarrays and their clustering. As an illustration, we consider a microarray data [22] which is generally used for numerical experiment. This is of type Affymetrix with here two classes of tissues for 104 subjects and 182 remaining probes which were selected among the 22283 original ones. The obtained results illustrate the interest of our approach because it is empirically observed that the column clustering is able to preserve well the stucture of the natural classes.

For this dataset, the Table I presents the average values of the indicators for 10 random sampling with replacement of the original data table. The parameters are $h = 19$, while the number of clusters were arbitrary chosen among $g \in \{9, 16, 25, 36\}$ and $m \in \{3, 4, 5\}$. Note that a few outliers exist

	GTM	GBGTM	GBGTM$_B$
ER	0.05	0.02	0.02
ED	0.16	0.19	0.15
SN	0.49	0.43	0.47
CN	0.90	0.95	0.93

TABLE I: Indicators for the dataset $Data1$.

apparently in the dataset according to the first principal plane of PCA. For this data, the topographic block model is able to slightly outperform the usual GTM with a small number of column clusters and thus very few parameters. Moreover, the two algorithms (Bayesian and not) lead to very similar results for the quality of the mapping. The reduction factor for the number of parameters is about $1 - m/d = 98\%$.

C. Output for a textual data table ($Data2$)

In textmining, the dataset CSTR [23] is widely used for applications. It is compound of 475 documents which are coded with 1000 terms. Four classes have a non balanced

distribution 101, 71, 178, 125. The transformation tfidf[1] (term frequency inverse document frequency) is applied to the co-occurence matrice.

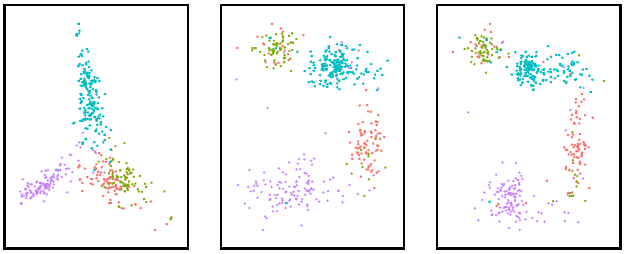

Fig. 2. Maps with the coordinates $\{\hat{s}_i\}$ from PCA, GTM, and GBGTM$_B$ from left to right.

For this dataset, the Table II presents the values of the indicators without sampling, $h = 19$, $g = 81$ and $m = 50$. Hence, GTM and GBGTM$_B$ have nearly identical quantitative indicators even if there is small loss with a block wise hypothesis while GBGTM has not performed well. Visually in Figure 2, the new method seems to show the classes more compact on the map, hence the loss might come from the missclassified data even for the visual indicators SN and CN. Note that a jittering with variance 0.01 has been added to the plot for better lisibility. The reduction factor for the number of parameters is about $1 - m/d = 95\%$. Note that the available implementation was too slow to handle any sampling.

	GTM	GBGTM$_B$
ER	0.09	0.10
ED	0.11	0.15
SN	0.47	0.45
CN	0.84	0.80

TABLE II: Indicators for the dataset $Data2$.

VI. Conclusion and discussion

Herein, we propose a new GTM for the projection of continuous data tables with a block structure. The approach is parsimonious and flexible. We also present brievly the links with GTM, PCA and least squares for contextual reasons. We observe in practice with two datasets of real data that GBGTM is able to outperform GTM or eventually lead to nearly similar results. A main property of our approach is to bring parsimony for the parameters in subspace modeling which makes it promising to manage with the current increasing size of the available datasets.

Several perspectives are possible for future developments of our approach. The column clusters offer additional information on the relations between the variables and can be discussed further. Some issues need to be solved. In particular convergence towards local minima must be reduced for GTM and for our proposed variant too. Other ways to improve the results might be alternatives to the training procedure with an eventual removal of the Bayesian treatment. Robustness might also dramatically improve even more the results and could be included in our model (see [24]).

[1]If f_{ij} is the original count, d_j is the number of documents containing the j^{th} term, f_i is the document total, then, $x_{ij} = \frac{f_{ij}}{f_i} \log \frac{n}{d_j+1}$.

References

[1] Kohonen, T.: Self-organizing maps. Springer (1997)

[2] Lebart, L., Morineau, A., Warwick, K. M.: Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices Wiley (1984)

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

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

[5] Bishop, C., Williams, C. K. I.: Developments of the Generative Topographic Mapping. Neurocomputing 21 203–224

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

[7] Tino, P., Nabney, I.: Hierarchical gtm: Constructing localized nonlinear projection manifolds in a principled way. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5) (2002) 639–656

[8] Vellido, A.: Selective smoothing of the generative topographic mapping. IEEE Transactions on Neural Networks **14**(3) (july 2003) 847–852

[9] Olier, I., Vellido, A.: Variational Bayesian Generative Topographic Mapping. Journal of Mathematical Modelling and Algorithms **7**(4) (December 2008) 371–387

[10] McLachlan, G.J., Peel, D.: Finite Mixture Models. John Wiley and Sons, New York (2000)

[11] Hautaniemi, S., Yli-Harja, O., Astola, J., Kauraniemi, P., Kallioniemi, A, Wolf, M., Ruiz, J., Mousses, S., Kallioniemi, O-P.: Analysis and Visualization of Gene Expression Microarray Data in Human Cancer Using Self-Organizing Maps. Mach. Learn. **52**(1-2) (July-August 2003) 45–66

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

[13] Nadif, M., Govaert, G.: Model-Based Co-clustering for Continuous Data. In: ICMLA, IEEE Computer Society. (2010) 175–180

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

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

[16] Jaakkola, T. S., Jordan, M. I : Bayesian parameter estimation via variational methods. Statistics and Computing **10**(1) (January 2000) 25–37

[17] Neal, R.M.: Bayesian Learning for Neural Networks, Vol. 118 of Lecture Notes in Statistics. Springer-Verlag (1996)

[18] Yamaguchi, N.: Variational bayesian inference with automatic relevance determination for generative topographic mapping. In: SCIS-ISIS'12. (2012) 2124–2129

[19] Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2) (April 1979) 224–227

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

[21] Owen, J. R., Nabney, I. T., Medina-Franco, J. L., Lopez-Vallejo, F.: Visualization of molecular fingerprints. Journal of chemical information and modeling **51**(7) (June 2011) 1552–1563

[22] Chowdary D., Lathrop J., Skelton J., Curtin K., Briggs T., Zhang Y., Yu J. Wang Y., Mazumder A.: Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. J Mol Diagn **8**(1) (Feb 2006) 31–39

[23] Tao, L., Ding, C.: The Relationships Among Various Nonnegative Matrix Factorization Methods for Clustering. ICDM '06. Sixth International Conference on Data Mining (Dec. 2006) 362–371 ,

[24] Vellido, A.: Missing data imputation through GTM as a mixture of t-distributions. Neural Networks **19** (November 2005) 1624–1635

[25] Zou, H., Hastie, T.: Regularization and variable selection via the Elastic Net. J. Royal Statist. Soc. Ser. B., 67 (2005) 301–320