A parameterization via random factor for generative topographic mapping

Rodolphe Priam, M. Nadif

A parameterization via random factor for generative topographic mapping

R. Priam, M. Nadif

Abstract— **The model proposed is based on a flexible hierarchical prior for a generalization of the generative topographic mapping (GTM) and the mixture of principal components analyzers (MPCA). By introducing random factors with correlated features in the Gaussian mixture model (GMM) this makes more flexible the center means such that the resulting projection is written with a new term coming from the prior. The parameters are estimated with a generalized expectation-maximization and maximum a posteriori. Empirical experiments illustrate positively the interest of our proposal.**

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

I. Introduction

In data analysis [1], partionning the space the rows or columns of a numerical data matrix and reducing its dimension lead to synthetic and understandable representations. Among the existing methods in the litterature, the Kohonen's map [2] or more generally the family of Self-Organizing maps (SOM) yield informative results; indeeed, they make possible to synthesize efficiently the whole distribution of a set of high dimensional vectors with a unique two dimensional map. These methods construct a discretized surface by using constrained clusters which are laid over the plane.

The family of the SOM methods includes several parametric alternative models with particular constraints over their parameters. Different methods have been developed in the literature. One of the most efficient is the Generative Topographic Mapping (GTM) model of Bishop and al. [3]. In this paper, it is proposed a more flexible model by a hierarchical prior for modelling randon factors in the Gaussian Mixture model (GMM). The obtained method called faGTM is unifying GTM and the mixture principal components analyzers (MPCA) by one more general model.

This paper is organized as follows. In section 2, we review GTM. In section 3, we introduce the method *faGTM* (see [4]). Section 4 presents the estimation of the parameters of the model. In Section 5, it is described the links between our model with the state of art and an extension with a Markov field is presented. Then we illustrate the non linear mapping with *faGTM* in Section 6. Finally, we summarize the results and conclude with perspectives.

II. The GTM model

Let us denote a set of n continuous vectors $\mathcal{D} = \{x_1, x_2, \cdots, x_n\}$ where x_i is a random vector $[x_{i1}, x_{i2}, \cdots, x_{id}]^T$ with a probability density function (pdf) of parameter $\boldsymbol{\theta}$. In the following, the random variables will not be in bold font and will be named as their observed values for lighter notation. Assuming $\mathcal{D}$ a i.i.d sample, the corresponding loglikelihood $L(\boldsymbol{\theta})$ is equal to:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

It is assumed that the data are generated by a GMM where each component k of the mixture represents a cluster. The i-st datum belongs to one class among the g classes according to a hidden random variable z_i in $\{1, \cdots g\}$ with a priori probability $\pi_k = P(z_i = k)$ such as $\pi_k \geq 0$ and $\sum_{k=1}^{g} \pi_k = 1$. The density of x_i is then defined in the corresponding classical way:

$$f(x_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f(x_i; \alpha_k).$$

where $f(x_i; \alpha_k)$ is the density of an observation x_i generated from the k-st *component*, a Gaussian density $\mathcal{N}(\mu_k, \Sigma_k)$ with the parameter $\alpha_k = (\mu_k, \Sigma_k)$ compound of the mean μ_k and the covariance matrix Σ_k. It is assumed that the data are generated by a Gaussian mixture model where each component k of the mixture represents a cluster. The density of x_i is then defined in the corresponding classical way as the weigthed sum of the Gaussian densities

$$f(x_i; \alpha_k) = \mathcal{N}(\mu_k, \Sigma_k)$$

for the k-st *component* with weight π_k such as $\pi_k > 0$ and $\sum_k \pi_k = 1$, chosen here equiprobable; the parameters $\alpha_k = (\mu_k, \Sigma_k)$ are the mean μ_k and the covariance matrix Σ_k. A regular mesh is defined by a set of nodes with bidimensional coordinates denoted:

$$\mathcal{S} = \{s_k = [s_{(k,1)}, s_{(k,2)}]^T, 1 \leq k \leq g\}.$$

These coordinates discretize the latent space which is the part of the plane is where the data sample is projected with the model:

$$\mathbb{S} = [-1; 1] \times [-1; 1].$$

Indeed, the components of the mean centers of the Gaussian mixture are typically modelled such as their relative multidimensional positions μ_k in the data space reflect as much as possible the relative bidimensional positions of the nodes s_k. Let us have h kernel functions $\{\phi_\ell, 1 \leq \ell \leq h\}$, like h bivariate Gaussian distributions for instance. Let us define g vectors ξ_k as the combination Φ of the h smooth nonlinear basis functions $\{\phi_\ell\}$ evaluated on the g constant coordinates of the grid $\mathcal{S}$:

$$\xi_k = [\phi_1(s_k), \phi_2(s_k), \cdots, \phi_h(s_k)]^T.$$

R. Priam, rpriam@gmail.com.

A linear transformation by a product with an unknown $d \times h$ matrix W provides mean centers in the data space. For a lighter notation, the global mean of the data cloud, denoted μ, can be introduced as a column of the matrix W when a bias term equal to 1 is among the components of ξ_k. Denoting $\mathbb{I}_d$ the identity matrix, the variance matrix chosen isotropic with parameter σ and the mean centers are written:

$$\mu_k = W\xi_k \text{ and } \Sigma_k = \sigma^2 \mathbb{I}_d.$$

The estimation of the unknown vector of parameter $\boldsymbol{\theta} = (W, \sigma)$ is performed by maximizing the loglikelihood $L(\boldsymbol{\theta})$ of the data $\mathcal{D}$ by the Expectation-Maximization or EM algorithm [5] with solution $\hat{\boldsymbol{\theta}}$.

The GTM model is often presented as a crude Monte-Carlo of probabilistic PCA (PPCA) [6], by writing the model with a marginalization over a discrete random variable equally distributed for the g values s_k. Like the Mixture of Factor Analyzers (MFA)[7] and the mixture of PPCA (MPPCA) [6], GTM is a particular model of Linear Latent Gaussian Model. The constraints on its centers derive from an underlying regular mesh; its factors ξ_k are shared in the clusters as the MFA with common loading matrix [8] but they are constant. On Figure 1 a graphical representation of the mapping by GTM is given.

In the next subsection, we introduce a random noise over the ξ_k. Without loss of generality, the data are supposed centered hereafter and the previously defined constant coordinates for the nodes of the mesh are denoted by:

$$s_k^{(0)} = \left(s_{(k,1)}^{(0)}, s_{(k,2)}^{(0)}\right)^T,$$

while the vectors $\xi_k^{(0)}$ are the constant initial basis of GTM with corresponding matrix $\Psi^{(0)}$. A more general $g \times h$ matrix of basis functions is denoted and will be random in the following of this paper:

$$\Psi = [\xi_1 | \xi_2 | \cdots | \xi_g].$$

The constraints over the underlying regular mesh of the Generative Topographic Mapping are freed by embedding random factors in the model as explained in the next section.

III. GTM WITH HIERARCHICAL FACTOR PRIOR

By introducing the appropriate prior over the variables ξ_k, the flexibility of the resulting mesh is increased and therefore leads to a novel projection map. So in the following, the basis vectors are supposed distributed according to independent Gaussian random variables. Their variances are chosen small in order to induce slow updates of the mean parameters during learning, and the covariances are not null between components. Let ρ be a positive value for parameterization of the prior pdf and the symmetric matrix C chosen such as:

$$C = \left[\exp\left(-\frac{1}{2\nu_C}||\xi_{(j)}^{(0)} - \xi_{(j')}^{(0)}||^2\right)\right]_{j,j'},$$

with ν_C a positive real well chosen and $\xi_{(j)}$ as the j-st row of Ψ. The quantity ν_C is automatically chosen by maximizing the entropy of the vector of probability defined by the normalized cell values of the matrix C, except its diagonal. An alternative for C is the sample correlation matrix, for instance. A random variable $\underline{\xi_k}$ is then defined conditionally to the values of ξ_k as:

$$f(\underline{\xi_k}|\xi_k; \boldsymbol{\theta}) \sim \mathcal{N}(\xi_k, \rho C).$$

The variables $\underline{\xi_k}$ are so random version of the fixed basis vector $\xi_k^{(0)}$ in the previous section. According to these hypotheses, for $x_i \in \mathcal{D}$, the proposed model is written using the variables $\underline{\xi_{z_i}}$ such as:

$$f(x_i|\underline{\xi_{z_i}}; \boldsymbol{\theta}) = \mathcal{N}(W\underline{\xi_{z_i}}, \sigma^2 \mathbb{I}_d).$$

If no constraint is further added, then the model reduces to a MPPCA with its factor having their components non independent. The parameter ρ helps to keep a slow convergence for ξ_k during the learning phase when it is chosen small enough. Then the induced self-organization of the center means behaves like in GTM if the updates of the mean vectors ξ_k are bound. In order to constrain the ξ_k basis vectors, we suppose these variables random and distributed as a Gaussian pdf with an expectation equal to the initial $\xi_k^{(0)}$. The variance of the noise is modeled with the same correlation matrix C as for $\underline{\xi_k}$ parameterized with a positive constant λ:

$$f(\xi_k; \Psi^{(0)}) = \mathcal{N}(\xi_k^{(0)}, \lambda C).$$

Such a hierarchical prior with a chain of three variables $(\underline{\xi_k}, \xi_k, {\xi_k}^{(0)})$ was never proposed for generative self-organizing maps. The introduced small variance parameter ρ and its consequence is also shown useful.

In the proposed model, ρ, C, λ, and π, are constant, while $\boldsymbol{\theta} = (\sigma, W, \Psi)$ needs to be estimated. Thus each factor ξ_k is shared by the observations belonging to the same cluster as in MFA with a common loading matrix [8]. In Factor Analysis and PPCA, each factor has its features which are indepent in order to reduce the dimensionality of the data to a smaller space with uncorrelated directions. In the Figure 1, the proposed model called *faGTM* and the original GTM are graphically pictured with a plate notation.

Finally, the whole parametric pdf of the flexible model *faGTM* is written in summary:

$$\begin{aligned} & f(\mathcal{D}, \Psi; \sigma, W, \Psi^{(0)}) \\ = \;& \textstyle\prod_i \sum_k \pi_k f(x_i|\xi_k; \sigma, W) \times \prod_k f(\xi_k; \Psi^{(0)}). \end{aligned}$$

In order to estimate the unknown parameters $\boldsymbol{\theta}$, we propose an a posteriori maximization, by processing the EM algorithm over the a posteriori law $\Psi|\mathcal{D}$ which leads to:

$$\hat{\boldsymbol{\theta}} = argmax_{\boldsymbol{\theta}} \; \log f(\Psi|\mathcal{D}; \sigma, W, \Psi^{(0)}).$$

The corresponding numerical problem is how to find a (local) maximum a posteriori to the proposed parametric distribution. In the next section, the expressions for the iterative updates of the parameter values are presented in closed-form.

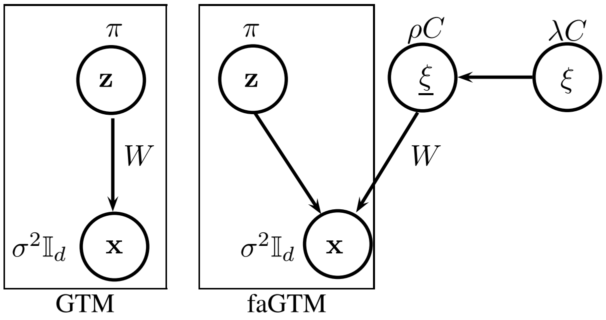

Fig. 1: Representation by the plate notation for GTM and faGTM with corresponding variables. In faGTM, a factor is modeled by the random variable denoted $\underline{\xi}$ while the variable ξ becomes its random expectation.
figure

IV. Estimation by EM

It is denoted $t^{(t)}_{z_i|x_i}$ the a posteriori probability that the i-st datum is generated by the z_i-st component having:

$$f(x_i|z_i = k; \boldsymbol{\theta}) = \mathcal{N}(W\xi_k, \sigma^2\mathbb{I}_d + \rho W C W^T).$$

Then inside a cluster it can be written for the a posteriori joint distribution for the cluster and basis functions:

$$t^{(t)}_{k,\underline{\xi_k}|x_i} = f(\underline{\xi_k}|x_i, \boldsymbol{\theta}^{(t)}) f(z_i = k|x_i; \boldsymbol{\theta}^{(t)}).$$

The function that we maximize, up to an additive constant, takes this form:

$$Q_{\sigma,W,\Psi|\boldsymbol{\theta}^{(t)}} = \textstyle\sum_{i,k} t^{(t)}_{k|x_i} \left[-d\log\sigma - \frac{q^{ik}_{W|\boldsymbol{\theta}^{(t)}}}{2\sigma^2} - \frac{q^{ik}_{\Psi|\boldsymbol{\theta}^{(t)}}}{2\rho}\right],$$

where,

$$\begin{aligned} q^{ik}_{W|\boldsymbol{\theta}^{(t)}} &= x_i^T x_i - 2x_i^T W e^{(t)}_{ik} + \mathrm{trace}(W^T W u^{(t)}_{ik}), \\ q^{ik}_{\Psi|\boldsymbol{\theta}^{(t)}} &= {\xi_k}^T C^{-1} \xi_k - 2{\xi_k}^T C^{-1} e^{(t)}_{ik}, \end{aligned}$$

and,

$$\begin{aligned} e^{(t)}_{ik} &= \xi^{(t)}_k + \rho {\Gamma^{(t)}}^T x^{(t)}_{ik} = \mathbb{E}_{\underline{\xi_k}|x_i;\hat{\boldsymbol{\theta}}}\left[\, \underline{\xi_k}\, \right], \\ u^{(t)}_{ik} &= \rho(I - \rho {\Gamma^{(t)}}^T W^{(t)})C + e^{(t)}_{ik} {e^{(t)}_{ik}}^T, \\ x_{ik} &= x_i - W\xi_k, \\ \Gamma &= (\sigma^2\mathbb{I}_d + \rho W C W^T)^{-1} W C. \end{aligned}$$

The previous Q function computed with previous parameters at step t is maximized in order to get the new current estimate $\boldsymbol{\theta}^{(t+1)}$. By resolving $\frac{\partial Q}{\partial W} = 0$ and $\frac{\partial Q}{\partial \sigma} = 0$, updates are derived for W and σ. Then, it is obtained:

$$\begin{aligned} W^{(t+1)} &= \left(\sum_{i,k} t^{(t)}_{k|x_i} x^{(t)}_i {e^{(t)}_{ik}}^T\right) \left(\sum_{i,k} t^{(t)}_{k|x_i} u^{(t)}_{ik}\right)^{-1} \\ \sigma^{(t+1)} &= \sqrt{\sum_{i,k} \frac{t^{(t)}_{k|x_i}}{nd} q^{ik}_{W^{(t)}|\boldsymbol{\theta}^{(t)}}}. \end{aligned}$$

Derivation of the criterion and resolving $\frac{\partial Q}{\partial \xi_k} = 0$ provides the following updates that are used for estimating the basis vectors updates, with:

$$\beta = \rho/\lambda.$$

This must be seen as a maximum a posteriori estimation of the mean parameter of the random variable $\underline{\xi_k}$ such that for all k:

$$\xi^{(t+1)}_k = \frac{1}{\sum_i t^{(t)}_{k|x_i} + \beta} \left(\sum_i t^{(t)}_{k|x_i} e^{(t)}_{ik} + \beta\, \xi^{(0)}_k\right).$$

Evaluating the $t_{k|x_i}, e_{ik},\ u_{ik}$ and Γ from $\boldsymbol{\theta}^{(t)}$ is the t-st E-step of EM which provides the Q function to be maximized. Solutions of the resulting null equations give new values for W and ξ_k for the M-step which completes an EM step at time $t+1$. Iterating this process converges to a stable solution for the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. The resulting updates for the g basis vectors illustrate the common observation in Bayesian statistics that the EM solution without any prior is averaged with an a priori value $\xi^{(0)}_k$. In faGTM, the hierarchical prior leads to the quotient of two real values ρ and λ for weigthing the additional value. At the t-st E-step of EM, it is evaluated $t^{(t)}_{k|x_i}$, $e^{(t)}_{ik}$, $u^{(t)}_{ik}$ and $\Gamma^{(t)}$ from current estimate of the parameters $\boldsymbol{\theta}^{(t)}$. At the $(t+1)$-st M-step the maximization provides the new estimates $W^{(t+1)}$, $\sigma^{(t+1)}$, and $\Psi^{(t+1)}$. Iterating the two steps, the EM algorithm converges to a stable solution for the maximum likelihood estimation of $\boldsymbol{\theta}$ denoted $\hat{\boldsymbol{\theta}}$.

The next section presents a discussion about the final estimates of the parameters by underlying the link with GTM and by extending the prior pdf with a Markov random field (MRF).

V. Discussion on the model and extension

The parameters ρ and λ (or β) are of great interest in regulating the evolution of the coordinated of the basis vectors during the learning because they permit to choose the behaviour of the constraints in faGTM. Stronger constraints in the model can be wanted, such situation is handled with a MRF model.

A. A generalizing model

The particular case where $\rho >> 0$ and $\lambda > 0$, are without any practical interest, while three other cases with particular values of these parameters have a noticable behavior:

i) For $\beta = 0$ with $\rho = 0$ and $\lambda > 0$, similar updates to GTM are obtained for W and σ.

ii) For $\beta = 0$ with $\rho > 0$ and $\rho << \lambda$, updates similar to MFA are obtained for the Ψ matrix except that factors are shared by data belonging to a same cluster, with a common loading matrix [8] and isotropic noises in the clusters, so like a mixture of PPCA.

iii) For $\beta > 0$ with $\rho > 0$ and $\lambda > 0$, this case is the more useful for data projection.

In summary, the model is generalizing the GTM and MPPCA. The approach leads to several other extensions of GTM as explained in this section.

Morever, when the parameter ρ is small enough, an approximation holds for the parameters. Expansions of the estimates lead to similar expressions to GTM, for their principle term, except the basis vectors which are non constant. For instance, at the first order it is given an analytical expression for such approximation of the update equation of the W matrix:

$$W^{(t+1)} \quad \approx \quad A_{(t)}\, V_{(t)}^{-1} + \rho\Delta_W^{(t+1)},$$

where, $\Delta_W^{(t+1)}$ is a matrix coming from the prior. It is recognized the update equations of GTM for the linear transformation, plus the term coming from the prior introduced in the model for the basis functions. Similarly, with $\delta^{(t+1)}$ obtained by appropriate Taylor expansion for ρ small, it is written:

$$\sigma^{(t+1)} \quad \approx \quad \sqrt{\sum_{i,k} \frac{t_{k|x_i}^{(t)}}{nd} \left|\left| x_i - W^{(t)}\xi_k^{(t)} \right|\right|^2} + \frac{1}{2}\rho\delta^{(t+1)}.$$

This is the update equation of GTM for the standard-deviation, plus the term coming from the prior introduced in the model for the basis functions. For a small ρ introduced by the faGTM model, the sign of $\hat{\delta}$, computed at the end of the maximum likelihood of GTM is informative about the variation of the corresponding intra-variance.

B. Constraints with a Markov random field

Eventually, for greater control of the flexibility, an alternative pdf of the proposed hyper prior is a MRF by constraining [3], [9] the mean centers $\{\mu_k = W\xi_k, 1 \leq k \leq g\}$. In the faGTM, constraining the set of basis vectors $\{\xi_k, 1 \leq k \leq g\}$ is performed by a suitable prior, say a Gaussian MRF with a covariance matrix C well chosen in order to keep the required correlations between neighboors features. With this regularization, the improvement of the clustering remains while a more controlled flexibility at the level of the model's grid is induced.

Indeed, the Q function at the E-step is written with another term for penalizing the likelihood of the mixture of Gaussian distributions:

$$q_{\Psi|\boldsymbol{\theta}^{(t)}}^{ik} = (\xi_{(j)} - \xi_{(j)}^{(0)})^T C^{-1} (\xi_{(j)} - \xi_{(j)}^{(0)}),$$

such as finally the update equations for Ψ are also different. Derivation of the new criterion and resolving $\frac{\partial Q}{\partial \xi_k} = 0$ leads to the g null equations, for $k = 1, \dots, g$:

$$C^{-1}\xi_k + d_{kk}^{-}\,\xi_k + \beta \sum_{l \neq k} d_{kl}^{-}\,\xi_l \quad = \quad C^{-1} \sum_i t_{k|x_i}^{(t)} e_{ik}^{(t)} + \beta \sum_k d_{kl}^{-} \xi_l^{(0)}.$$

It is denoted $vect(\Psi) = [\xi_1^T, \xi_2^T, ..., \xi_g^T]^T$, while the vector $e_{\beta,\Psi^{(0)}}^{(t)}$ is compound of the g right hand side terms from the previous g normal equations. The $gh \times gh$ matrix $M_{(t)}$ is compound of $g \times g$ cells of dimension $h \times h$. For the k-st row o cells, the contents of the (k,ℓ)-st column-cell is equal to $\left[d_{k\ell}^{-}\mathbb{I}_h\right]$ except the k-st cell with $[C^{-1}]$ added:

$$\begin{aligned} M^{(t)} \quad &= \quad \begin{pmatrix} d_{11}^{-}\mathbb{I}_h & d_{12}^{-}\mathbb{I}_h & \cdots & d_{1g}^{-}\mathbb{I}_h \\ d_{21}^{-}\mathbb{I}_h & d_{22}^{-}\mathbb{I}_h & \cdots & d_{2g}^{-}\mathbb{I}_h \\ \vdots & \vdots & \vdots & \vdots \\ d_{g1}^{-}\mathbb{I}_h & d_{g2}^{-}\mathbb{I}_h & \cdots & d_{gg}^{-}\mathbb{I}_h \end{pmatrix} \\ &+ \quad \begin{pmatrix} C^{-1} & 0_h & 0_h & 0_h \\ 0_h & C^{-1} & 0_h & 0_h \\ \vdots & \vdots & \vdots & \vdots \\ 0_h & 0_h & \cdots & C^{-1} \end{pmatrix}. \end{aligned}$$

The new parameters $\Psi^{(t+1)}$ at step t are then updated by:

$$vect\left(\Psi^{(t+1)}\right) = \left(M^{(t)}\right)^{-1} e_{\beta,\Psi^{(0)}}^{(t)}$$

This induces the inversion of a matrix of size $gh \times gh$ where gh is about several hundred in practice. This is numerically more demanding when comparing with the first version of the model in the previous section. In the next section, several non linear maps are constructed with the faGTM method for three datasets.

VI. Illustration

The questions that naturally arise are how are constructed the non linear maps by faGTM and how the approach is different to the original GTM. This section first introduces the projection with the proposed method and compares the resulting mapping with GTM. The following experiments illustrate the interest of our method.

A. Mapping with the model

Contrary to the GTM, the position of the clusters s_k are not constant for faGTM. Indeed, during the EM algorithm, the vectors $\xi_k = \Phi(s_k)$ are updated and the positions s_k are also indirectly updated. To find their final position after estimation, it is presented an appropriate parameterization by adding their original bidimensional coordinates $s_k^{(0)}$ among the features of $\xi_k^{(0)}$.

Let $\mathcal{P}_{2d}(u)$ be the projection of the vector u to its two first components. The initial positions $s_k^{(0)}$ are supposed to be inserted in the 2 first components of $\xi_k^{(0)}$. Their final positions at the maximum likelihood are:

$$\hat{s}_k = [\hat{s}_{(k,1)}, \hat{s}_{(k,2)}]^T = \mathcal{P}_{2d}(\hat{\xi_k}).$$

Then, for the i-st datum the projection $\tilde{s}_i^{faGTM}$ is written by using the appropriate projected expectation:

$$\begin{aligned} \tilde{s}_i \quad &= \quad \mathcal{P}_{2d} \left\{ \sum_{k=1}^{g} \mathbb{E}_{\underline{\xi_k}|x_i;\hat{\boldsymbol{\theta}}} \left[\, \underline{\xi_k} \,\right] \right\} \\ &= \quad \sum_{k=1}^{g} \hat{t}_{k|x_i} \begin{pmatrix} \hat{s}_{(k,1)} \\ \hat{s}_{(k,2)} \end{pmatrix} + \rho\,\hat{\Delta}_i \end{aligned}$$

where,

$$\hat{\Delta}_i = \sum_{k=1}^{g} \hat{t}_{k|x_i} \mathcal{P}_{2d} \left\{ \hat{\Gamma}^T (x_i - \hat{W}\hat{\xi_k}) \right\}$$

A similar weighted expectation is used in [8] but the corresponding model has its properties very different from GTM and faGTM. The obtained expression induces that faGTM adds a continuous term to the original projection. Mapping by GTM is reduced to only the first term in the right member with $\hat{s}_k = s_k^{(0)}$ kept constant.

The $\tilde{s}_i$ breaks into two terms contributing to a possible enhancement of the projection. The first one is an averaged discrete quantity resulting from the new positions of the cluster $\tilde{s}_{z_i}$. The second one is an averaged linear and continuous quantity which induces variability of the projection around a cluster position. This quantity denoted as $\hat{\Delta}_i$ recalls a linear Principal Component Analysis [10] and should be small.

In the case of faGTM the evolution with the time step t of the positions of the nodes $s_k^{(t)}$ during EM is also informative. As the proposed algorithm is able to move the positions s_k during the learning process, the trajectory of these quantities can be observed by using the projection $\mathcal{P}_{2d}$ after each EM iteration. It is then drawn the g curves passing through the g sets of plots:

$$\mathcal{T}_k = \left\{ s_k^{(0)}, s_k^{(1)}, s_k^{(2)}, \cdots, \hat{s}_k \right\} \text{ where } s_k^{(t)} = \mathcal{P}_{2d} \left\{ \xi_k^{(t)} \right\} .$$

As ρ is chosen small, the difference between two consecutive s_k positions should be small too, and these g curves should be smooth and informative about the learning process.

B. Empirical results

In this subsection, the method is illustrated with several datasets. Two datasets are simulated for two situations, a sample from clusters and a sample from a surface. The third dataset is a sample of real data, the so-called Iris data with three classes.

a) Artificial classes dataset:: To illustrate the method, a sample is drawn from five Gaussian classes in a high dimensional space; this dataset is generated from a mixture of mixing a priori probabilities, $(0.15, 0.2, 0.15, 0.2, 0.3)$, center means, $\mu_1 = [0.0, 3.5]$, $\mu_2 = [-3.5, 0.0]$, $\mu_3 = [3.5, 0.0]$, $\mu_4 = [0.0, -3.5]$, $\mu_5 = [0.0, 0.0]$ and covariance matrices Σ_k diagonal for $k = 1, \ldots, 4$ with non null elements $[0.10, 0.45]$ $[0.45, 0.10]$, $[0.45, 0.10]$, $[0.10, 0.45]$ and a fifth matrix equal to a correlation matrix with a covariance value equal to 0.90. A sample of 1000 data from the mixture is projected in a space of dimension 10 by the matrix $B = [B_1^T B_2^T]$, and

$$\begin{aligned} B_1^T &= [.5\,, -.9\,, .3\,, .6\,, .2\,, -.7\,, .0\,, .0\,, .0\,, .0] \\ B_2^T &= [.0\,, .0\,, .0\,, .8\,, -.7\,, .5\,, .6\,, -.4\,, .3\,, -.5] \end{aligned}$$

An additive uniform noise in the range $[0; 0.1]$ is also added. Finally each resulting data vector is completed with 5 variables which are noises from an uniform distribution in the range $[0; 0.15]$. This resulting dataset counts $n = 1000$ vectors with $d = 15$ features. The projection for this simulated sample is shown in Figure 2. In this simulation, faGTM leads to a result very similar to GTM as expected. The trajectory map of the $s_k^{(t)}$, which plots the trajectory of the positions during learning of the ξ_k, is also able to help the understanding of the final map thanks to a concentration of the mean centers inside each natural class. Such a property is related to an enhanced clustering.

b) Artificial sphere dataset:: A second illustration in Figure 3 is the mapping of a random noisy sample from half of a sphere surface in $\mathbb{R}^3$ centered in the origin, with rayon 1, plus a small cut of the second half-sphere; this dataset counts $n = 1479$ vectors of $d = 3$ features. This sample was partitioned artificially into 10 non-overlapping classes which are shown in the graphics with different colors. The result is then a nice view of a spherical shape with ten separated clusters of points plus a narrow band around the disq. It is noticed that the projection of the border of the second half-sphere has small overlapping with the first sphere due to boundary effects due to binary posterior probabilities.

c) Iris dataset:: A third illustration in Figure 4 is the projection of the dataset of the Iris with 150 vectors in a 4-dimensional space and 3 classes. The result is also very encouraging as the projection shows by a nice display the 3 classes almost all separated. Two classes are very near in the data space and by consequence are also their relative projections on the map with moreover no overlapping induced. The trajectory plot less relevant in this situation to reveal the 3 clusters is able to show how the s_k have moved during the learning.

Further experiments with the MRF are not reported in the document, and have shown to keep more faithfully the underlying mesh with its nodes compound of the positions s_k. These results are very encouraging, the method adds flexibility to the vectors of basis function, and leads to a novel graphical representation for the Generative Topographic Mapping.

VII. Conclusion and perspective

The paper presents a hierarchical prior pdf for unifying the MPPCA and GTM methods. A variant with a Gaussian Markov random field is introduced with an update equation more computationally intensive for updating the set of basis functions. The resulting hidden mesh is also more regular in this case than for the product of Gaussian distributions. The method offers several perspectives. For instance, the trajectory map can be associated to the previous mapping methods [11], [12], [13] for revealing the hidden structure of the data cloud.

References

[1] C. M. Bishop, *Neural Networks for Pattern Recognition.* Clarendon Press, 1995.

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

[3] C. M. Bishop, M. Svensén, and C. K. I. Williams, “Developpements of generative topographic mapping,” *Neurocomputing*, vol. 21, pp. 203–224, 1998.

[4] R. Priam and M. Nadif, “Generative topographic mapping and factor analyzers,” in *ICPRAM (1)*, 2012, pp. 284–287.

[5] A. Dempster, N. Laird, and D. Rubin, “Maximum-likelihood from incomplete data via the EM algorithm,” *J. Royal Statist. Soc. Ser. B., 39*, pp. 1–38, 1977.

[6] M. E. Tipping and C. M. Bishop, “Mixtures of probabilistic principal component analyzers,” *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

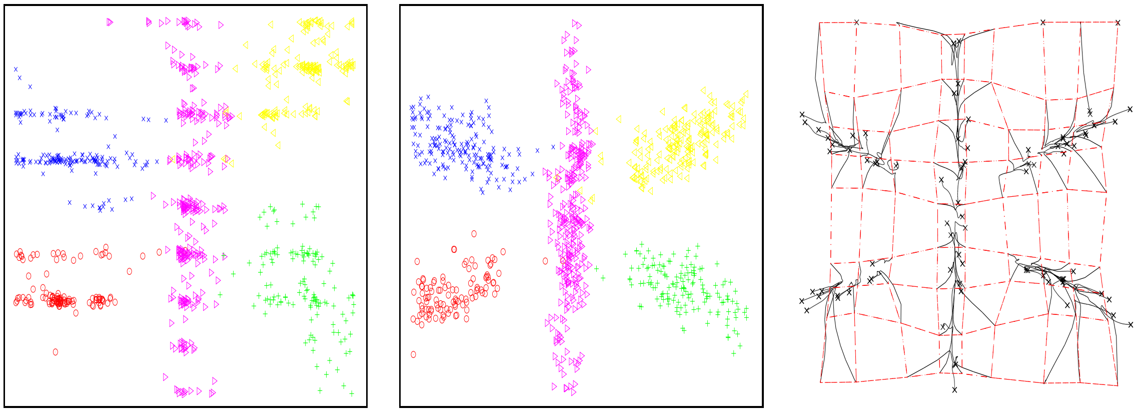

Fig. 2: The results for the simulated dataset from five classes, a) the GTM mapping, b)the faGTM mapping, c) the graphs of the g sets $\{s_k^{(0)}, s_k^{(1)}, \cdots, \hat{s}_k\}$ with in red dot line the mesh resulting of the first EM step with coordinates $s_k^{(1)}$.
figure

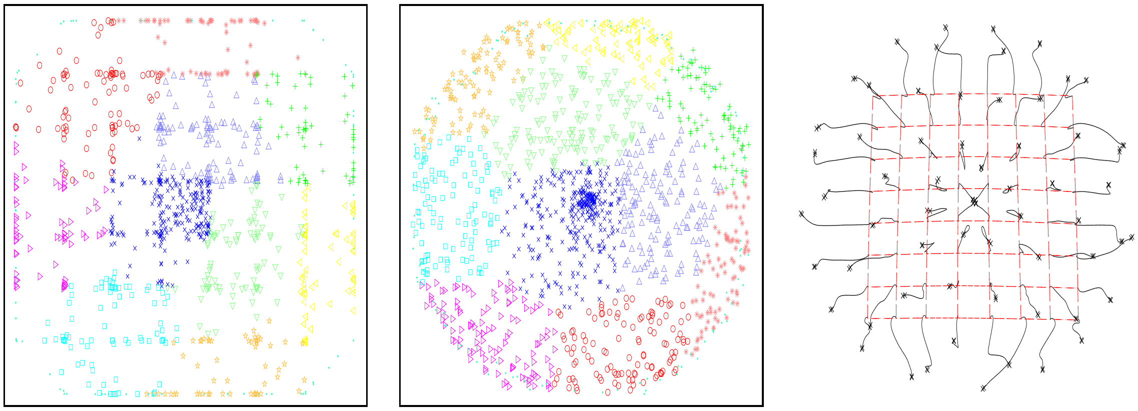

Fig. 3: The results for the simulated dataset from a surface, a) the GTM mapping, b)the faGTM mapping, c) the graphs of the g sets $\{s_k^{(0)}, s_k^{(1)}, \cdots, \hat{s}_k\}$ with in red dot line the mesh resulting of the first EM step with coordinates $s_k^{(1)}$.
figure

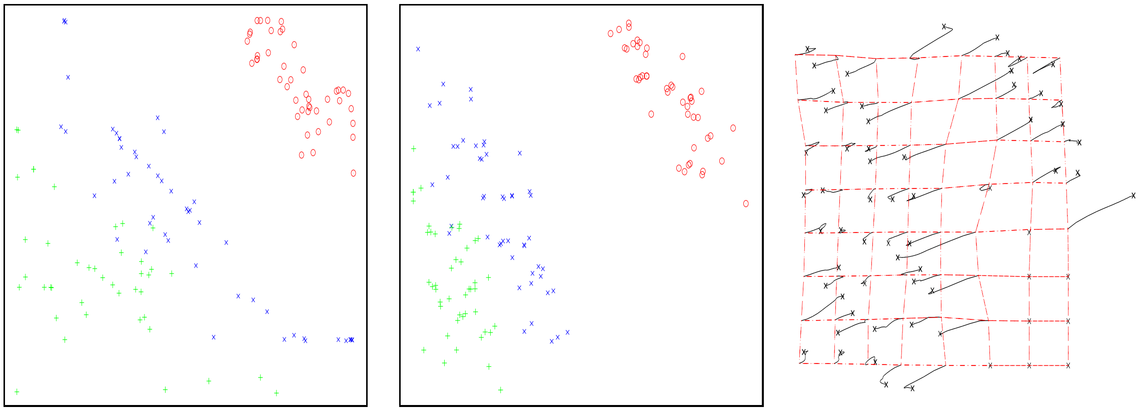

Fig. 4: The results for the Iris dataset, a) the GTM mapping, b)the faGTM mapping, c) the graphs of the g sets $\{s_k^{(0)}, s_k^{(1)}, \cdots, \hat{s}_k\}$ with in red dot line the mesh resulting of the first EM step with coordinates $s_k^{(1)}$.
figure

[7] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Tech. Rep. CRG-TR-96-1, 1996.

[8] J. Baek, G. McLachlan, and L. Flack, "Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[9] A. Utsugi, "Hyperparameter selection for self-organizing maps," *Neural Comp*, vol. 9, pp. 623–635, 1997.

[10] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 2002.

[11] C. Bishop, M. Svensen, and C. Williams, "Magnification factors for the gtm algorithm," in *Fifth International Conference on Artificial Neural Networks*, 1997, pp. 64 –69.

[12] D. M. Maniyar and I. T. Nabney, "Visual data mining using principled projection algorithms and information visualization techniques,"

in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '06. ACM, 2006, pp. 643–648.

[13] P. Tiňo and N. Gianniotis, "Metric properties of structured data visualizations through generative probabilistic modeling," in *Proceedings of the 20th international joint conference on Artifical intelligence*, 2007, pp. 1083–1088.