GENERATIVE TOPOGRAPHIC MAPPING AND FACTOR ANALYZERS

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

Keywords: Generative topographic mapping, Random factors, Expectation-maximization.

Abstract: By embedding random factors in the Gaussian mixture model (GMM), we propose a new model called faGTM. Our approach is based on a flexible hierarchical prior for a generalization of the generative topographic mapping (GTM) and the mixture of principal components analyzers (MPPCA). The parameters are estimated with expectation-maximization and maximum a posteriori. Empirical experiments show the interest of our proposal.

1 INTRODUCTION

In data analysis (Bishop, 1995), partioning the space of 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 literature, the Kohonen's map (Kohonen, 1997) or more generally the family of the self-organizing maps (SOM) yield informative results. Indeed, they make possible to synthesize efficiently the distribution of a set of high dimensional vectors with an unique two dimensional map. These methods construct a discretized surface by constraining the clusters which are laid over the mapping 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 et al., 1998).

As usually, it is considered the sample of n continuous i.i.d vectors $\mathcal{D} = \{x_1, x_2, \cdots, x_n\}$. Each x_i is a d-dimensional random vector $[x_{i1}, x_{i2}, \cdots, x_{id}]^T$ with a probability density function (pdf) of parameter θ. In the following, the random variables are not be in bold font and are named as their observed values for lighter notation. In GTM, the densities of the components of a Gaussian Mixture model (GMM) (McLachlan and Peel, 2000) have same spherical covariance matrices $\Sigma_k = \sigma^2 \mathbb{I}_d$ with $\mathbb{I}_d$ the d-dimensional identity matrix. They have same prior probabilities $\pi_k = g^{-1}$ and are denoted $f(x_i|k;\theta) \sim \mathcal{N}(\mu_k, \Sigma_k)$ where θ is the vector or set of parameters and μ_k is the mean center. The means are constrained by considering a grid discretizing $[-1;1] \times [-1;1]$. The bidimensional coordinates of this mesh are kept constant and denoted:

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

The mean centers are parameterized by $\mu_k = W\xi_k$ where W is a matrix for a linear projection while ξ_k comes from a nonlinear transformation of the s_k by h kernel functions $\phi_\ell(s_k)$ such as:

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

Like the Mixture of Factor Analyzers (MFA) (Ghahramani and Hinton, 1996) and the Mixture of PPCA (MPPA) (Tipping and Bishop, 1999a), GTM is a particular model of Linear Latent Gaussian Model. The GTM model is often presented as a crude Monte-Carlo of the probabilistic PCA (PPCA) (Tipping and Bishop, 1999b), by writing the model with a marginalization over a discrete random variable equally distributed for the g values s_k. The constraints on its centers derive from an underlying regular mesh. Its factors ξ_k are shared in the clusters as MFA with common loading matrix (Baek et al., 2009) but they are constant. In the following, we introduce a random noise over the ξ_k by a hierarchical prior for modelling random factors and the resulting coordinates s_k are no longer fixed. Without loss of generality, the data are supposed centered hereafter.

The paper is organized as follows. In section 2, we present the proposed prior and the new method named *faGTM*. In section 3, we propose an estimation of the

Priam R. and Nadif M. (2012).
GENERATIVE TOPOGRAPHIC MAPPING AND FACTOR ANALYZERS.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 284-287
DOI: 10.5220/0003765202840287

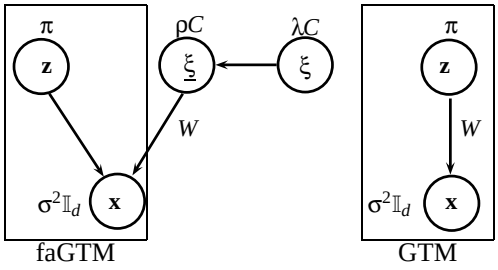

Figure 1: Representation by the plate notation for GTM and faGTM with corresponding variables. In faGTM, a factor is modeled by the random variable denoted $\underline{\xi}$ while the variable ξ becomes its random expectation.

parameters of the model. Then we present a way to perform the mapping in section 4 and the results of our experiments in section 5. Finally, we conclude with perspectives.

2 GTM AND HIERARCHICAL FACTOR PRIOR

In the following, the fixed coordinates s_k of GTM are denoted $s_k^{(0)}$ while the vectors $\xi_k^{(0)}$ are the constant initial basis of GTM with corresponding matrix $\Psi^{(0)}$.

The vectors of basis functions 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 scalar and $\xi_{(j)}^{(0)}$ the j-th row of $\Psi^{(0)}$. 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:

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

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

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

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 when it is chosen small enough. Then the induced self-organization of the mean centers 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, and:

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

Such a hierarchical prior with a chain of three variables $(\underline{\xi_k}, \xi_k, \xi_k{}^{(0)})$ has never been proposed for generative self-organizing maps. The $g \times h$ dimensional matrix of basis functions is unknown and denoted $\Psi = [\xi_1|\xi_2|\cdots|\xi_g]$. In Figure 1, the proposed model called *faGTM* and GTM are graphically pictured with a plate notation. In the proposed model, ρ, C, λ, and $(\pi_1,...,\pi_g)$, are constant, while $\theta = (\sigma, W, \Psi)$ needs to be estimated. Finally, the whole parametric pdf of our proposed flexible model *faGTM* is written in summary:

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

In order to estimate the unknown parameters θ, it is proposed an a posteriori maximization, by processing the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) over the pdf of the model for solving $\hat{\theta} = \text{argmax}_\theta \log f(\mathcal{D}, \Psi; \sigma, W, \Psi^{(0)})$. 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.

3 ESTIMATION BY EM

In this section we denote $t_{z_i|x_i}^{(t)}$ the posterior probability that the i-th datum is generated by the z_i-th component having:

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

Then it can be written for the posterior joint distribution of the component and the vector of basis functions $t_{k,\underline{\xi_k}|x_i}^{(t)} = f(\underline{\xi_k}|x_i,\theta^{(t)}) f(z_i = k|x_i;\theta^{(t)})$. The function that we maximize, up to an additive constant, can be written:

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

where:

$$\begin{aligned} q^{ik}_{W|\theta^{(t)}} &= \text{trace}(W^T W u^{(t)}_{ik}) + x_i^T x_i - 2x_i^T W e^{(t)}_{ik}, \\ q^{ik}_{\Psi|\theta^{(t)}} &= (\rho+\lambda)\xi_k{}^T C^{-1}\xi_k \\ &- 2\xi_k{}^T C^{-1}\left(\lambda e^{(t)}_{ik} + \rho\xi_k{}^{(0)}\right). \end{aligned}$$

Here $e_{ik} = \xi_k + \rho\Gamma^T x_{ik}$, $u_{ik} = \rho(I - \rho\Gamma^T W)C + e_{ik}e_{ik}{}^T$, $\Gamma = (\sigma^2\mathbb{I}_d + \rho WCW^T)^{-1}WC$, and $x_{ik} = x_i - W\xi_k$ at the t-th step of EM.

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

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

With $\beta = \rho/\lambda$, derivation of the criterion and resolving $\frac{\partial Q}{\partial \xi_k} = 0$ provides the updates for the vectors of basis functions such as:

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

Evaluating the $t_{k|x_i}, e_{ik}$, u_{ik} and Γ from $\theta^{(t)}$ is the t-th 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 $\hat{\theta}$ of θ, while $\hat{t}_{k|x_i}$ are the final posterior probabilities $t^{(t)}_{k|x_i}$ at the end of the learning.

In the next sections, we construct several nonlinear maps with the faGTM method for three datasets, after introducing a way to perform the projection of a dataset with the method.

4 MAPPING WITH THE MODEL

During the learning, the vectors $\xi_k = \Phi(s_k)$ are updated and the positions s_k are also indirectly updated. It is proposed an approach to retrieve the not constant positions s_k of the clusters by using $s_k^{(0)}$ as first components of $\xi_k^{(0)}$. Let $\mathcal{P}_{2d}(u)$ be the projection of the vector u to its two first components. The 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-th datum the projection $\tilde{s}_i^{faGTM}$ is written with the projected expectation:

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

In comparison with the GTM, the coordinates discretizing the projection space are flexible and an additive smoothing term appears in the mapping.

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 $\mathcal{P}_{2d}$ of $\xi_k^{(t)}$ after each EM iteration. It is then drawn the g curves passing through the g sets of points:

$$\mathcal{T}_k = \left\{s_k^{(0)}, s_k^{(1)}, s_k^{(2)}, \cdots, \hat{s}_k\right\}$$

As ρ is chosen small, the difference between two consecutive positions $s_k^{(t-1)}$ and $s_k^{(t)}$ should be small too, and these g curves should be smooth as observed next section.

5 EXPERIMENTS

In this section we test the method with several datasets, two simulated ones and a real one:

- *Art1.* This dataset is a sample drawn from five Gaussian pdf in a high dimensional space. The data are generated from a mixture with the prior probabilities $(0.15, 0.2, 0.15, 0.2, 0.3)$, transposed center means, $[0.0, 3.5]$, $[-3.5, 0.0]$, $[3.5, 0.0]$, $[0.0, -3.5]$, $[0.0, 0.0]$ and diagonal covariance matrices Σ_k with diagonal $[0.10, 0.45]$ $[0.45, 0.10]$, $[0.45, 0.10]$, $[0.10, 0.45]$ for $k = 1, \ldots, 4$ and a fifth matrix equal to a correlation matrix with non-diagonal components equal to 0.90. A sample of 1000 data from the mixture is projected in a space of dimension 10 by the matrix $B = [B_1|B_2]^T$, and $B_1^T = [.5, -.9, .3, .6, .2, -.7, .0, .0, .0, .0]$ and $B_2^T = [.0, .0, .0, .8, -.7, .5, .6, -.4, .3, -.5]$. An uniform noise supported on the interval $[0; 0.1]$ is also added. Finally each resulting data vector is completed with 5 variables which are i.i.d. from an

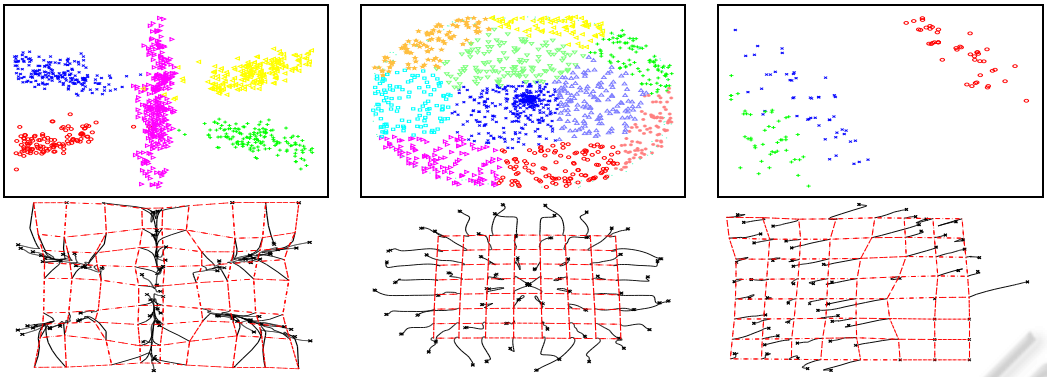

Figure 2: The results for the three datasets are given in column 1 for *Art1*, column 2 for *Art2* and column 3 for *Iris*. The 1-th row is for the map from the faGTM model. The 2-nd row is for the graphs of the g sets of curves τ_k. The mesh resulting of the first EM step with coordinates $s_k^{(1)}$ is in red dot line.

uniform distribution on $[0;0.15]$. This resulting dataset counts $n = 1000$ vectors with $d = 15$ features.

- *Art2*. This dataset is a random sample from one half of a sphere centered at origin in $\mathbb{R}^3$ with radius 1, plus a circular band surrounding the 2-th hemisphere near the great circle. This dataset counts $n = 1479$ vectors of $d = 3$ features. The sample from the hemisphere is clustered artificially into 10 non-overlapping classes.
- *Iris*. The dataset of the Iris is compound of 150 vectors in a 4-dimensional space and 3 classes. The trajectory plot is less relevant in this situation to reveal the 3 clusters which are less separated.

The projections for the three datasets are shown in Figure 2. The points for the different classes have different colors on the graphics. The results are very encouraging, the method adds flexibility to the vectors of basis function, and leads to a novel graphical representation for the GTM.

6 CONCLUSIONS AND PERSPECTIVE

We have proposed a hierarchical factor prior with parameters C, ρ and λ for generalizing MPPCA and GTM. The faGTM and its prior offer several perspectives. For instance, the trajectory map as a complement to the magnification factors (Bishop et al., 1997; Maniyar and Nabney, 2006; Tiňo and Gianniotis, 2007) can be studied further.

REFERENCES

Baek, J., McLachlan, G., and Flack, L. (2009). 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*.

Bishop, C., Svensen, M., and Williams, C. (1997). Magnification factors for the gtm algorithm. In *Fifth International Conference on Artificial Neural Networks*, pages 64 –69.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press.

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

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

Ghahramani, Z. and Hinton, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1.

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

Maniyar, D. M. and Nabney, I. T. (2006). 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*, KDD '06, pages 643–648. ACM.

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

Tipping, M. E. and Bishop, C. M. (1999a). Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482.

Tipping, M. E. and Bishop, C. M. (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):pp. 611–622.

Tiňo, P. and Gianniotis, N. (2007). Metric properties of structured data visualizations through generative probabilistic modeling. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1083–1088.