GENERATIVE TOPOGRAPHIC MAPPING AND FACTOR ANALYZERS

Rodolphe Priam¹ and Mohamed Nadif²

1S3RI, University of Southampton, University Road, SO17 1BJ, Southampton, U.K.
2LIPADE, Université Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France

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), partitioning 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 $\Omega = \{x_1,x_2,\ldots,x_n\}$. Each $x_i$ is a d-dimensional random vector $[x_{i1},x_{i2},\ldots,x_{id}]^T$ with a probability density function (pdf) of parameter $\theta$. 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 $\Sigma_k = \sigma^2 I_d$ with $I_d$ the d-dimensional identity matrix. They have same prior probabilities $g_k = g^{-1}$ and are denoted $f(x_i|\theta, \Sigma_k)$ where $\theta$ is the vector or set of parameters and $\mu_k$ 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: $s = [s(k,1),s(k,2)]^T, 1 \leq k \leq g$.

The mean centers are parameterized by $\mu_k = W\xi_k$, where $W$ is a matrix for a linear projection while $\xi_k$ comes from a nonlinear transformation of the $s_k$ by $h$ kernel functions $\phi_h(s_k)$ such as: $\xi_k = [\phi_h(s(k,1)),\phi_h(s(k,2)),\ldots,\phi_h(s(k,d))]^T$.

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 $\xi_k$ 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 $\xi_k$ 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

Keywords: Generative topographic mapping, Random factors, Expectation-maximization.
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 $\rho$ be a positive value for parameterization of the prior pdf and the symmetric matrix $C$ chosen such as:

$$C = \left[ \exp \left( -\frac{1}{2\nu_c} ||\xi_0^{(0)} - \xi_j^{(0)}||^2 \right) \right]_{j \in J},$$

with $\nu_c$ a positive real scalar and $\xi_j^{(0)}$ the $j$-th row of $\Psi^{(0)}$. The quantity $\nu_c$ 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 $\xi_k$ is then defined conditionally to the values of $\xi_k$ as:

$$f(\xi_k|\xi_0; \theta) \sim \mathcal{N}(\xi_k, \rho C).$$

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

$$f(x_i|\xi_k; \theta) = \mathcal{N}(W\xi_k, \sigma^2 I_d).$$

If no constraint is further added, then the model reduces to a MPPCA with its factor having their components non independent. The parameter $p$ helps to keep a slow convergence for $\xi_k$ 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 $\xi_k$ are bound. In order to constrain the $\xi_k$ 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 $\xi_k$ parameterized with a positive constant $\lambda$, and:

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

Such a hierarchical prior with a chain of three variables $(\xi_k, \xi_k, \xi_k^{(0)})$ has never been proposed for generative self-organizing maps. The $g \times h$ dimensional matrix of basis functions is unknown and denoted $\Psi = [\xi_1, \xi_2, \ldots, \xi_g]$. In Figure 1, the proposed model called $faGTM$ and GTM are graphically pictured with a plate notation. In the proposed model, $\rho, C, \lambda$, and $(\pi_1, \ldots, \pi_g)$, 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:

$$f(\theta;\Psi, W, \Psi^{(0)}) = \prod_i \sum_j \pi_j f(x_i; \xi_j; \sigma, W) \times \prod_k f(\xi_k; \Psi^{(0)}).$$

In order to estimate the unknown parameters $\theta$, 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 $\hat{\theta} = \arg \max_{\theta} \log f(\theta, \Psi, W, \Psi^{(0)})$. 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 $t(i)*_{k|z_i}$ the posterior probability that the $i$-th datum is generated by the $z_i$-th component having:

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

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

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

Then, for the $i$-th datum the projection $s_i^{(0)} \gamma_{aGT M}$ is written with the projected expectation:

$$\tilde{s}_i = \mathcal{P}_{2d} \left\{ \sum_{k=1}^{g} \mathbb{E}_{\theta_k} \left[ \xi_k \right] \right\} = \sum_{k=1}^{g} \mathbb{E}_{\xi_k} \left[ \left( \tilde{s}_{k,1}, \tilde{s}_{k,2} \right) \right] + \rho \mathcal{P}_{2d} \left\{ \Gamma^T (x_i - \bar{W} \xi_k) \right\}$$

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 the faGTM the evolution with the time step $t$ of the positions of the nodes $s_i^{(t)}$ 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:

$$\xi_k = \left\{ s_k^{(0)}, s_k^{(1)}, s_k^{(2)}, \ldots, s_k \right\}$$

As $\rho$ 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, $[0,0,3.5], [-3.5,0,0], [3.5,0,0], [0,0,-3.5], [0,0,0]$ and diagonal covariance matrices $\Sigma_k$ 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 $B_1^T = [-5,-9,3,6,2,-7,0,0,0,0]$, $B_2^T = [-5,-9,3,6,2,-7,0,0,0,0]$ and $B_2^T = [-5,-9,3,6,2,-7,0,0,0,0]$. An uniform noise supported on the interval $[0,1]$ is also added. Finally each resulting data vector is completed with $5$ variables which are i.i.d. from an
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 $C, \rho$ and $\lambda$ 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; Tiño and Gianniotis, 2007) can be studied further.

REFERENCES


