# Variational GTM

Iván Olier and Alfredo Vellido

Department of Computing Languages and Systems (LSI)
Technical University of Catalonia (UPC)
C/. Jordi Girona 1-3, Edifici Omega, Despatx S106
08034 - Barcelona, Spain
{iaolier,avellido}@lsi.upc.edu

**Abstract.** Generative Topographic Mapping (GTM) is a non-linear latent variable model that provides simultaneous visualization and clustering of high-dimensional data. It was originally formulated as a constrained mixture of distributions, for which the adaptive parameters were determined by Maximum Likelihood (ML), using the Expectation-Maximization (EM) algorithm. In this paper, we define an alternative variational formulation of GTM that provides a full Bayesian treatment to a Gaussian Process (GP)-based variation of GTM. The performance of the proposed Variational GTM is assessed in several experiments with artificial datasets. These experiments highlight the capability of Variational GTM to avoid data overfitting through active regularization.

## 1 Introduction

Manifold learning models attempt to describe multivariate data in terms of low dimensional representations, often with the goal of allowing the intuitive visualization of high-dimensional data. Generative Topographic Mapping (GTM) [1] is one such model, whose probabilistic setting and functional similarity make it a principled alternative to Self-Organizing Maps (SOM)[2]. In its basic formulation, the GTM is trained within the ML framework using EM, permitting the occurrence of data overfitting unless regularization is included, a major drawback when modelling noisy data. Its probabilistic definition, though, allows the formulation of principled extensions, such as those providing active model regularization to avoid overfitting [3,4].

The regularization methods in [3,4] were based on Bayesian evidence approaches. Alternatively, we could reformulate GTM within a fully Bayesian approach and endow the model with regularization capabilities based on variational techniques [5,6]. In this paper, we define a novel Variational GTM model based on the GTM with GP prior outlined in [3], to which a Bayesian estimation of its parameters is added.

Several preliminary experiments with noisy artificial data were designed to show how Variational GTM limits the negative effect of data overfitting, improving on the performance of the standard regularized GTM [3] and the standard GTM with GP prior, while retaining the data visualization capabilities of the model.

The remaining of the paper is organized as follows: First, in section 2, an introduction to the original GTM, the GTM with GP prior and a Bayesian approach for the GTM, are provided. This is followed, in section 3, by the description of the proposed Variational GTM. Several experiments for the assessment of the performance of the proposed model are described, and their results presented and discussed, in section 4. The paper wraps up with a brief conclusion section.

## 2 Generative Topographic Mapping

### 2.1 The Original GTM

The neural network-inspired GTM is a nonlinear latent variable model of the manifold learning family, with sound foundations in probability theory. It performs simultaneous clustering and visualization of the observed data through a nonlinear and topology-preserving mapping from a visualization latent space in $\Re^L$(with $L$ being usually 1 or 2 for visualization purposes) onto a manifold embedded in the $\Re^D$ space, where the observed data reside. The mapping that generates the manifold is carried out through a *regression function* given by:

$$\mathbf{y} = \mathbf{W}\mathbf{\Phi}\left(\mathbf{u}\right) \tag{1}$$

where $\mathbf{y} \in \Re^D$, $\mathbf{u} \in \Re^L$, $\mathbf{W}$ is the matrix that generates the mapping, and $\mathbf{\Phi}$ is a matrix with the images of $S$ basis functions $\phi_s$ (defined as radially symmetric Gaussians in the original formulation of the model). To achieve computational tractability, the prior distribution of $\mathbf{u}$ in latent space is constrained to form a uniform discrete grid of $K$ centres, analogous to the layout of the SOM units, in the form:

$$p\left(\mathbf{u}\right) = \frac{1}{K}\sum_{k=1}^{K}\delta\left(\mathbf{u} - \mathbf{u}_k\right) \tag{2}$$

This way defined, the GTM can also be understood as a constrained mixture of Gaussians. A density model in data space is therefore generated for each component $k$ of the mixture, which, assuming that the observed data set $\mathbf{X}$ is constituted by $\mathbf{N}$ independent, identically distributed (i.i.d.) data points $\mathbf{x}_n$, leads to the definition of a complete likelihood in the form:

$$P\left(\mathbf{X}|\mathbf{W},\beta\right) = \left(\frac{\beta}{2\pi}\right)^{ND/2}\prod_{n=1}^{N}\left\{\frac{1}{K}\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\left\|\mathbf{x}_n - \mathbf{y}_k\right\|^2\right)\right\} \tag{3}$$

where $\mathbf{y}_k = \mathbf{W}\mathbf{\Phi}\left(\mathbf{u}_k\right)$. From Eq. 3, the adaptive parameters of the model, which are $\mathbf{W}$ and the common inverse variance of the Gaussian components, $\beta$, can be optimized by ML using the EM algorithm. Details can be found in [1].

### 2.2 Gaussian Process Formulation of GTM

The original formulation of GTM described in the previous section has a hard constraint imposed on the mapping from the latent space to the data space due to the finite number of basis functions used. An alternative approach is introduced in [3], where the regression function using basis functions is replaced by a smooth mapping carried out by a GP prior. This way, the likelihood takes the form:

$$P\left(\mathbf{X}|\mathbf{Z},\mathbf{Y},\beta\right)=\left(\frac{\beta}{2\pi}\right)^{ND/2}\prod_{n=1}^{N}\prod_{k=1}^{K}\left\{\exp\left(-\frac{\beta}{2}\left\|\mathbf{x}_n-\mathbf{y}_k\right\|^2\right)\right\}^{z_{kn}} \tag{4}$$

where: $\mathbf{Z}=\{z_{kn}\}$ are binary membership variables complying with the restriction $\sum_{k=1}^{K} z_{kn}=1$ and $\mathbf{y}_k=\left(y_{k1},\ldots,y_{kD}\right)^T$ are the column vectors of a matrix $\mathbf{Y}$ and the centroids of spherical Gaussian generators. Note that the spirit of $\mathbf{y}_k$ in this approach is similar to the regression version of GTM (Eq. 1) but with a different formulation: A GP formulation is assumed introducing a prior multivariate Gaussian distribution over $\mathbf{Y}$ defined as:

$$P\left(\mathbf{Y}\right)=\left(2\pi\right)^{-KD/2}\left|\mathbf{C}\right|^{-D/2}\prod_{d=1}^{D}\exp\left(-\frac{1}{2}\mathbf{y}_{(d)}^T\mathbf{C}^{-1}\mathbf{y}_{(d)}\right) \tag{5}$$

where $\mathbf{y}_{(d)}$ is each one of the row vectors of the matrix $\mathbf{Y}$ and $\mathbf{C}$ is a matrix where each of its elements is a covariance function that can be defined as

$$\mathbf{C}\left(i,j\right)=\mathbf{C}\left(\mathbf{u}_i,\mathbf{u}_j\right)=\nu\exp\left(-\frac{\left\|\mathbf{u}_i-\mathbf{u}_j\right\|^2}{2\alpha^2}\right),\quad i,j=1\ldots K \tag{6}$$

and where parameter $\nu$ is usually set to 1. The $\alpha$ parameter controls the flexibility of the mapping from the latent space to the data space. An extended review of covariance functions can be found in [7]. An alternative GP formulation was introduced in [8], but this approach had the disadvantage of not preserving the topographic ordering in latent space, being therefore inappropiate for data visualization purposes.

Note that Eqs. 3 and 4 are equivalent if a prior multinomial distribution over $\mathbf{Z}$ in the form $P\left(\mathbf{Z}\right)=\prod_{n=1}^{N}\prod_{k=1}^{K}\left(\frac{1}{K}\right)^{z_{kn}}=\frac{1}{K^N}$ is assumed.

Eq. 4 leads to the definition of a log-likelihood and parameters $\mathbf{Y}$ and $\beta$ of this model can be optimized using the EM algorithm, in a similar way to the parameters $\mathbf{W}$ and $\beta$ in the regression formulation. Some basic details are provided in [3].

### 2.3 Bayesian GTM

The specification of a full Bayesian model of GTM can be completed by defining priors over the parameters $\mathbf{Z}$ and $\beta$. Since $z_{kn}$ are defined as binary values, a multinomial distribution can be chosen for $\mathbf{Z}$:

$$P\left(\mathbf{Z}\right) = \prod_{n=1}^{N}\prod_{k=1}^{K} p_{kn}^{z_{kn}} \tag{7}$$

where $p_{kn}$ is the parameter of the distribution.

As in [9], a Gamma distribution[1] is chosen to be the prior over $\beta$:

$$P\left(\beta\right) = \Gamma\left(\beta|d_\beta, s_\beta\right) \tag{8}$$

where $d_\beta$ and $s_\beta$ are the parameters of the distribution. Therefore, the joint probability $P\left(\mathbf{X},\mathbf{Z},\mathbf{Y},\beta\right)$ is given by:

$$P\left(\mathbf{X},\mathbf{Z},\mathbf{Y},\beta\right) = P\left(\mathbf{X}|\mathbf{Z},\mathbf{Y},\beta\right)P\left(\mathbf{Z}\right)P\left(\mathbf{Y}\right)P\left(\beta\right) \tag{9}$$

This expression can be maximized through evidence methods using the Laplace approximation [10] or, alternatively, using Markov Chain Monte Carlo [11] or variational [5,6] methods.

## 3 Variational GTM

### 3.1 Motivation of the Use of Variational Inference

A basic problem in statistical machine learning is the computation of the marginal likelihood $P\left(\mathbf{X}\right) = \int P\left(\mathbf{X},\Theta\right)d\Theta$, where $\Theta = \{\theta_i\}$ is the set of parameters defining the model. Depending of the complexity of the model, the analytical computation of this integral could be intractable. Variational inference allows approximating the marginal likelihood through Jensen's inequality as follows:

$$\begin{aligned}\ln P\left(\mathbf{X}\right) &= \ln\int P\left(\mathbf{X},\Theta\right)d\Theta = \ln\int Q\left(\Theta\right)\frac{P\left(\mathbf{X},\Theta\right)}{Q\left(\Theta\right)}d\Theta \\ &\geq \int Q\left(\Theta\right)\ln\frac{P\left(\mathbf{X},\Theta\right)}{Q\left(\Theta\right)}d\Theta = F\left(Q\right)\end{aligned} \tag{10}$$

The function $F\left(Q\right)$ is a lower bound function such that its convergence guarantees the convergence of the marginal likelihood. The goal in variational methods is choosing a suitable form for the density $Q\left(\Theta\right)$ in such a way that $F\left(Q\right)$ can be readily evaluated and yet which is sufficiently flexible that the bound is reasonably tight. A reasonable approximation for $Q\left(\Theta\right)$ is based on the assumption that it factorizes over each one of the parameters as $Q\left(\Theta\right) = \prod_i Q_i\left(\theta_i\right)$. That assumed, $F\left(Q\right)$ can be maximized leading the optimal distributions:

$$Q_i\left(\theta_i\right) = \frac{\exp\left\langle\ln P\left(\mathbf{X},\Theta\right)\right\rangle_{k\neq i}}{\int\exp\left\langle\ln P\left(\mathbf{X},\Theta\right)\right\rangle_{k\neq i}d\theta_i} \tag{11}$$

where $\langle\ .\ \rangle_{k\neq i}$ denotes an expectation with respect to the distributions $Q_k\left(\theta_k\right)$ for all $k \neq i$.

[1] The Gamma distribution is defined as follows: $\Gamma\left(\nu|d_\nu,s_\nu\right) = \frac{s_\nu^{d_\nu}\nu^{d_\nu-1}\exp^{-s_\nu\nu}}{\Gamma(d_\nu)}$

### 3.2 A Bayesian Approach of GTM Based on Variational Inference

In order to apply the variational principles to the Bayesian GTM within the framework described in the previous section, a $Q$ distribution of the form:

$$Q(\mathbf{Z}, \mathbf{Y}, \beta) = Q(\mathbf{Z})\, Q(\mathbf{Y})\, Q(\beta) \tag{12}$$

is assumed, where natural choices of $Q(\mathbf{Z})$, $Q(\mathbf{Y})$ and $Q(\beta)$ are similar distributions to the priors $P(\mathbf{Z})$, $P(\mathbf{Y})$ and $P(\beta)$, respectively. Thus, $Q(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \tilde{p}_{kn}^{z_{kn}}$, $Q(\mathbf{Y}) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_{(d)} | \tilde{\mathbf{m}}^{(d)}, \tilde{\mathbf{\Sigma}}\right)$, and $Q(\beta) = \mathit{\Gamma}\left(\beta | \tilde{d}_\beta, \tilde{s}_\beta\right)$. Using these expressions in Eq. 11, the following formulation for the variational parameters $\tilde{\mathbf{\Sigma}}, \tilde{\mathbf{m}}^{(d)}, \tilde{p}_{kn}, \tilde{d}_\beta$ and $\tilde{s}_\beta$ can be obtained:

$$\tilde{\mathbf{\Sigma}} = \left( \langle \beta \rangle \sum_{n=1}^{N} \mathbf{G}_n + \mathbf{C}^{-1} \right)^{-1} \tag{13}$$

$$\tilde{\mathbf{m}}^{(d)} = \langle \beta \rangle\, \tilde{\mathbf{\Sigma}} \sum_{n=1}^{N} x_{nd} \langle \mathbf{z}_n \rangle \tag{14}$$

$$\tilde{p}_{kn} = \frac{\exp\left\{ -\frac{\langle \beta \rangle}{2} \left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle \right\}}{\sum_{k'=1}^{K} \exp\left\{ -\frac{\langle \beta \rangle}{2} \left\langle \|\mathbf{x}_n - \mathbf{y}_{k'}\|^2 \right\rangle \right\}} \tag{15}$$

$$\tilde{d}_\beta = d_\beta + \frac{ND}{2} \tag{16}$$

$$\tilde{s}_\beta = s_\beta + \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{kn} \rangle \left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle \tag{17}$$

where $\mathbf{z}_n$ corresponds to each row vector of $\mathbf{Z}$ and $\mathbf{G}_n$ is a diagonal matrix of size $K \times K$ with elements $\langle \mathbf{z}_n \rangle$. The moments in the previous equations are defined as: $\langle z_{kn} \rangle = \tilde{p}_{kn}$, $\langle \beta \rangle = \frac{\tilde{d}_\beta}{\tilde{s}_\beta}$, and $\left\langle \|\mathbf{x}_n - \mathbf{y}_k\|^2 \right\rangle = D\tilde{\mathbf{\Sigma}}_{kk} + \sum_{d=1}^{D} \left( x_{nd} - \tilde{m}^{(kd)} \right)^2$.

Finally, and according to Eq. 10, the lower bound function $F(Q)$ is derived from:

$$F(Q) = \int Q(\mathbf{Z})\, Q(\mathbf{Y})\, Q(\beta) \ln \frac{P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta)\, P(\mathbf{Z})\, P(\mathbf{Y})\, P(\beta)}{Q(\mathbf{Z})\, Q(\mathbf{Y})\, Q(\beta)} d\mathbf{Z} d\mathbf{Y} d\beta \tag{18}$$

Integrating out, we obtain:

$$\begin{aligned} F(Q) = {} & \langle \ln P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta) \rangle + \langle \ln P(\mathbf{Z}) \rangle + \langle \ln P(\mathbf{Y}) \rangle + \langle \ln P(\beta) \rangle \\ & - \langle \ln Q(\mathbf{Z}) \rangle - \langle \ln Q(\mathbf{Y}) \rangle - \langle \ln Q(\beta) \rangle \end{aligned} \tag{19}$$

where the moments are expressed as:

$$\begin{aligned}\langle \ln P(\mathbf{X}|\mathbf{Z},\mathbf{Y},\beta)\rangle &= \frac{ND}{2}\langle \ln\beta\rangle - \frac{ND}{2}\ln 2\pi \\ &\quad -\frac{\langle\beta\rangle}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{kn}\rangle\left\langle \|\mathbf{x}_n-\mathbf{y}_k\|^2\right\rangle \qquad (20)\end{aligned}$$

$$\langle \ln P(\mathbf{Z})\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{kn}\rangle \ln p_{kn} \qquad (21)$$

$$\langle \ln P(\mathbf{Y})\rangle = -\frac{KD}{2}\ln 2\pi - \frac{D}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{d=1}^{D}\left\langle \mathbf{y}_{(d)}^{T}\mathbf{C}^{-1}\mathbf{y}_{(d)}\right\rangle \qquad (22)$$

$$\langle \ln P(\beta)\rangle = d_\beta \ln s_\beta - \ln \mathit{\Gamma}(d_\beta) + (d_\beta - 1)\langle\ln\beta\rangle - s_\beta\langle\beta\rangle \qquad (23)$$

$$\langle \ln Q(\mathbf{Z})\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{kn}\rangle \ln \tilde{p}_{kn} \qquad (24)$$

$$\langle \ln Q(\mathbf{Y})\rangle = -\frac{KD}{2}\ln 2\pi - \frac{D}{2}\ln\left|\tilde{\mathbf{\Sigma}}\right| - \frac{KD}{2} \qquad (25)$$

$$\langle \ln Q(\beta)\rangle = \tilde{d}_\beta \ln \tilde{s}_\beta - \ln \mathit{\Gamma}\left(\tilde{d}_\beta\right) + \left(\tilde{d}_\beta - 1\right)\langle\ln\beta\rangle - \tilde{s}_\beta\langle\beta\rangle \qquad (26)$$

and

$$\langle\ln\beta\rangle = \psi\left(\tilde{d}_\beta\right) - \ln\tilde{s}_\beta \qquad (27)$$

$$\left\langle \mathbf{y}_{(d)}^{T}\mathbf{C}^{-1}\mathbf{y}_{(d)}\right\rangle = \mathrm{tr}\left[\mathbf{C}^{-1}\left(\tilde{\mathbf{\Sigma}} + \tilde{\mathbf{m}}^{(d)}\left(\tilde{\mathbf{m}}^{(d)}\right)^{T}\right)\right] \qquad (28)$$

In the previous expressions, $\mathit{\Gamma}(\cdot)$ are Gamma functions, and $\psi(\cdot)$ is the Digamma function. Details of these calculations can be found in [12].

## 4 Experiments

### 4.1 Experimental Design

The main aim of the set of experiments presented and discussed in this section is the preliminary assessment of the robustness of the proposed model in the presence of noise. Moreover, the performance of Variational GTM is compared with that of the standard GTM (with a GP formulation).

The models used in all the experiments were initialized in the same way to allow straightforward comparison. The matrix centroids of the Gaussian generators $\mathbf{Y}$ and the inverse of the variance $\beta$ were set through PCA-based initialization [1] and the parameters $\{p_{kn}\}$ are fixed and were initialized using the posterior selection probability of the latent node $k$ given data point $\mathbf{x}_n$, defined using Bayes' theorem as:

$$p_{kn} = \frac{\exp\left(-\frac{\beta}{2}\left\|\mathbf{x}_n - \mathbf{y}_k^*\right\|^2\right)}{\sum_{k=1}^{K}\exp\left(-\frac{\beta}{2}\left\|\mathbf{x}_n - \mathbf{y}^*{}_k\right\|^2\right)} \tag{29}$$

where $\mathbf{y}_k^*$ is the initial value obtained previously for each centroid $k$. The parameter $s_\beta$ was set to $d_\beta/\beta$ and $d_\beta$ was initialized to a small value close to 0. For each set of experiments, several values of $K$ and $\alpha$ were used.

### 4.2 Robustness of the Variational GTM in the Presence of Noise

The goal of this first set of experiments was assessing and comparing the robustness of both the standard GTM using GP and the proposed Variational GTM models in the presence of increasing levels of noise, as well as comparing it to the robustness of the standard regularized GTM with single regularization term [3] trained by EM (GTM-SRT). The artificial data sets used to this end consisted of 700 points sampled from a circumference to which different levels of random Gaussian noise were added (standard deviations of $\{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35\}$). For each noise level, 10 data sets were randomly generated and used to train every model. All training runs used the following settings: $K = 36$ for all models, $\alpha = 0.1$ for the GTM-GP and the Variational GTM and $d_\beta = 0.01$ for the Variational GTM. Furthermore, the number of basis functions for GTM-SRT was set to 25. Different values of $K$ and $\alpha$ were considered with similar results.

Two measures were employed to gauge the regularization capabilities of the models: The mean square error between the centroids $\{\mathbf{y}_k\}$ and the underlying circumference without noise, and the standard deviation of the square error. The results for these measures, displayed in Fig. 1, indicate that, as the levels of noise increase, the mean and standard deviation square errors grow to be much higher for the standard GTM using GP than for the proposed Variational GTM, although in the case of the mean error this difference cannot be clearly appreciated for very low levels of noise. Furthermore, Variational GTM is shown to outperform GTM-SRT at all noise levels, while being far less sensitive to the increase of such levels.

These results are a preliminary but clear indication that the proposed Variational GTM provides better regularization performance than both the standard GTM using GP and GTM-SRT. This is neatly illustrated in Fig. 2, for the first two models, where two samples of the artificial data sets used in this experiments and their corresponding results (represented by the connected centroids) are displayed. Although at low noise levels, both models perform similarly, at higher levels the standard GTM using GP fits the noise to a great extent, whereas Variational GTM is much less affected by it and is capable of reproducing the underlying data generator far more faithfully. This should lead to a model with better generalization capabilities.

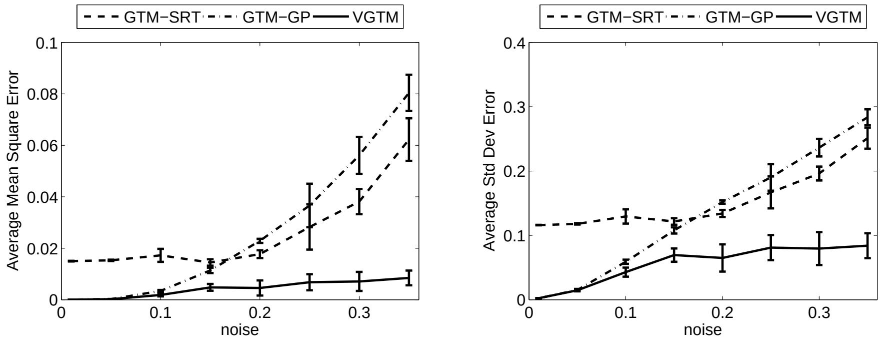

**Fig. 1.** Plots of the average mean square error between the centroids $\{\mathbf{y}_k\}$ and the theorical circumference whithout noise (left plot) and the average standard deviation of the square error (right plot) for GTM-SRT (dashed line), for the standard GTM using GP (dashed-dotted) and the proposed Variational GTM (solid). The vertical bars indicate the standard deviation of these averages.

### 4.3 Data Visualization Using Variational GTM

A second set of experiments was carried out with the aim of verifying the topographic preservation capabilities of the proposed Variational GTM and consequently, its data visualization capabilities on a low-dimensional discrete latent space. For that, an artificial data set consisting of 12 hetereogenously separated clusters was generated by means of an equivalent number of radial Gaussian distributions. The following settings were used to train the model: $K = 64$, $\alpha = 0.1$ and $d_\beta = 0.01$. The resulting data visualization is accomplished through the membership map generated by means of the mode projection [1] of the data into the latent space, given by $\mathbf{u}_n^{\texttt{mode}} = \underset{k}{\texttt{argmax}}\,(\tilde{p}_{kn})$, where the variational parameter $\tilde{p}_{kn}$ was used.

The data set and its corresponding membership map are displayed in Fig. 3, where several interesting data points, some of these placed well within the clusters and others in the edge between two clusters, are singled out for illustration. It is clear that their representation in latent space faithfully preserves the existing topographic ordering and neighbouring relations in data space.

## 5 Conclusions

Details of a variational formulation of GTM have been provided in this paper. Through several experiments, Variational GTM has been shown to endow the model with effective regularization properties, enabling it to avoid, at least partially, fitting the noise and, therefore, enhancing its generalization capabilities. This regularization has been shown to be more effective than that provided by the standard GTM with GP formulation and the standard regularized GTM.

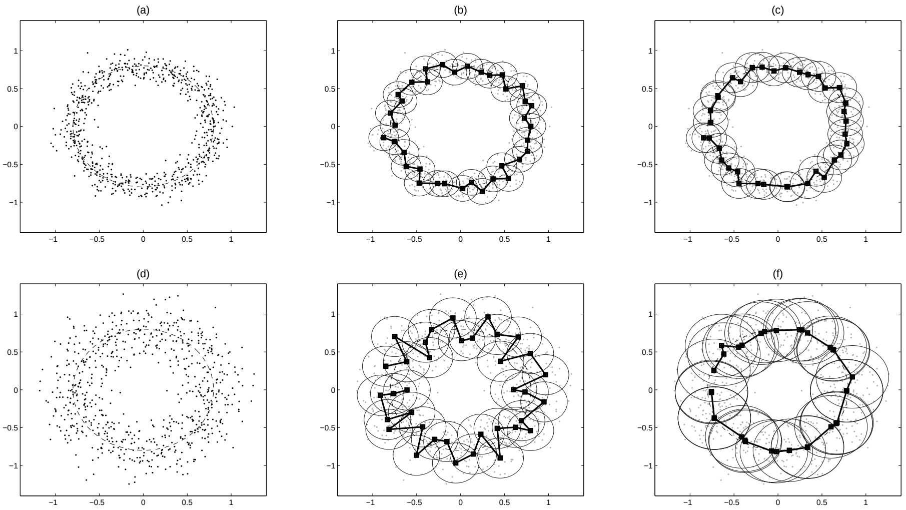

**Fig. 2.** Left column: two of the artificial data sets generated from a circumference (dashed line) to which noise of levels 0.1 (top row) and 0.25 (bottom row) was added. Middle column: including results after training using standard GTM with GP prior. Right column: including results after training using the proposed Variational GTM. The resulting manifold embedded in the data space is represented by the connected centroids $\{\mathbf{y}_k\}$ (filled squares) in the centres of circles of radius $2\sqrt{\beta^{-1}}$) (common standard deviation).

The experiments reported in this brief paper are necessarily limited by space availability and therefore preliminary. A much more detailed experimental design, including more datasets spanning a wider range of characteristics, as well an explicit testing of its generalization capabilities, would be required to complete the assessment of the model. The current study should be understood as a first step towards that end.

A variational treatment of parameter $\alpha$ is difficult and, therefore, it was fixed a priori in the reported experiments. However, an interesting approach to its calculation in the context of variational GP classifiers, using lower and upper bound funtions, was presented in [13] and could be considered in future work with the proposed Variational GTM.

## References

1. Bishop, C.M., Svensen, M., Williams, C.R.I.: GTM: The Generative Topographic Mapping. Neural Comput. **10**(1) (1998) 215–234
2. Kohonen, T.: Self-Organizing Maps (3rd ed). Springer-Verlag, Berlin (2001)
3. Bishop, C.M., Svensen, M., Williams, C.R.I.: Developments of the Generative Topographic Mapping. Neurocomputing **21**(1–3) (1998) 203–224

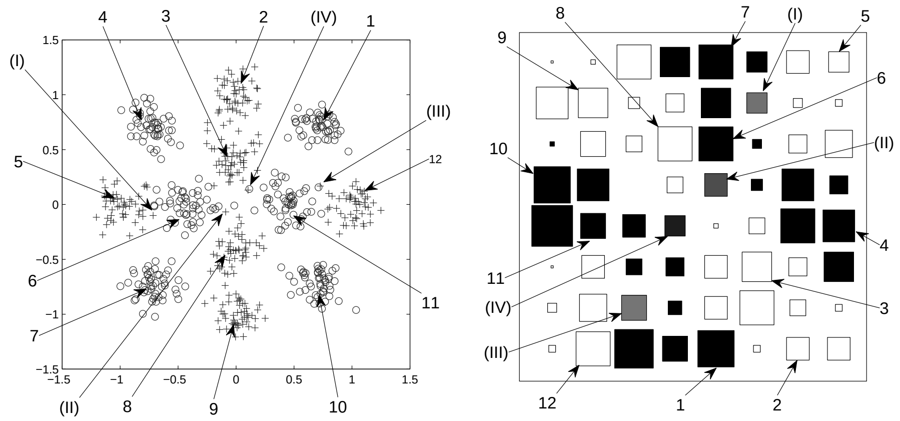

**Fig. 3.** (Left) 600 points randomly sampled for 12 Gaussians, forming clusters artificially labeled as 'o' and '+'. These labels are only used for visualization purposes and were not included in the training. (Right) The resulting membership map corresponding to the mode projection into a latent space of $K = 64$ latent points (represented as squares). The relative size of each square is proportional to the number of data assigned to its respective latent point and the hue of gray indicates the ratio of the data belonging to each cluster label: white for '+' and black for 'o'. The arabic numbers (1 to 12), signaling specific data points, show how their neighbouring relations are preserved in the latent space. The romanic numbers (I to IV) signal with ambiguous cluster allocation. They are all mapped into *grey points* in latent space.

4. Vellido, A., El-Deredy, W., Lisboa, P.J.G.: Selective smoothing of the Generative Topographic Mapping. IEEE T. Neural Networ. **14**(4) (2003) 847–852
5. Beal, M.: Variational algorithms for approximate Bayesian inference. PhD thesis, The Gatsby Computational Neuroscience Unit, Univ. College London (2003)
6. Jakkola, T., Jordan, M.I.: Bayesian parameter estimation via variational methods. Stat. Comput. **10** (2000) 25–33
7. Abrahamsen, P.: A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center, Oslo, Norway (1997)
8. Utsugi, A.: Bayesian sampling and ensemble learning in Generative Topographic Mapping. Neural Process. Lett. **12** (2000) 277–290
9. Bishop, C.M.: Variational principal components. In: Proceedings Ninth Intern. Conf. on Artificial Neural Networks. Volume 1. (1999) 509–514
10. MacKay, D.J.C.: A practical Bayesian framework for back-propagation networks. Neural Comput. **4**(3) (1992) 448–472
11. Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.I.: An introduction to MCMC for machine learning. Mach. Learn. **50** (2003) 5–43
12. Olier, I., Vellido, A.: A variational Bayesian formulation for GTM: Theoretical foundations. Technical report, Technical University of Catalonia (UPC) (2007)
13. Gibbs, M., MacKay, D.J.C.: Variational Gaussian process classifiers. IEEE T. Neural Networ. **11**(6) (2000) 1458–1464