A Variational Formulation for GTM Through Time

Iván Olier and Alfredo Vellido

Abstract—**Generative Topographic Mapping (GTM) is a latent variable model that, in its original version, was conceived to provide clustering and visualization of multivariate, real-valued, i.i.d. data. It was also extended to deal with non-i.i.d. data such as multivariate time series in a variant called GTM Through Time (GTM-TT), defined as a constrained Hidden Markov Model (HMM). In this paper, we provide the theoretical foundations of the reformulation of GTM-TT within the Variational Bayesian framework and provide an illustrative example of its application. This approach handles the presence of noise in the time series, helping to avert the problem of data overfitting.**

I. 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. GTM [1], originally defined for the clustering and visualization of i.i.d. data, is one such model that can be ascribed to the field of Statistical Machine Learning. Its probabilistic setting eases the definition of principled extensions, such as GTM-TT [2] for the analysis of multivariate time series, a model that was assessed in detail in [3], [4].

One well-known potential drawback in the process of knowledge discovery from both static data and time series is that of the presence of uninformative noise and the associated problem of data overfitting. In its basic formulation, the GTM is trained within the Maximum Likelihood (ML) framework using the Expectation-Maximization (EM) algorithm, and overfitting may occur unless regularization methods are applied. In [5], [6], regularization of GTM was based on Bayesian evidence approaches, which require a number of modelling assumptions and approximations to be made.

An alternative for the formulation of GTM that confers the model with regularization capabilities, while avoiding such approximations, is that of using variational techniques [7]. A Variational GTM model based on the GTM with a Gaussian Process (GP) prior outlined in [5], with added Bayesian estimation of its parameters, was recently described in [8]. This Variational GTM was shown to limit the negative effect of data overfitting, improving on the performance of the standard GTM with GP prior, while retaining the data visualization capabilities of the model. In this paper we extend such Variational approach to the analysis of multivariate time series, defining the variational GTM-TT and illustrating its properties with a simple example.

Iván Olier and Alfredo Vellido are with the Department of Computing Languages and Systems, Technical University of Catalonia, C/. Jordi Girona 1-3, Edifici Omega, 08034 - Barcelona, Spain (email: {iaolier,avellido}@lsi.upc.edu).

The remaining of this paper is organized as follows: First, in section II, an introduction to the original GTM-TT [2] is provided. Section III provides a Bayesian framework for GTM-TT. This is followed, in section IV, by the description of the proposed Variational Bayesian inference method for GTM-TT in some detail. Its performance is illustrated by the experiments reported in section V.

II. The Standard Generative Topographic Mapping Through Time

GTM-TT can be seen as a GTM model in which the latent states are linked by transition probabilities in a similar fashion to HMMs. Therefore, GTM-TT can be understood as a topology-constrained HMM.

Assuming a sequence of N hidden states $\mathbf{Z} = \{z_1, z_2, \ldots, z_n, \ldots, z_N\}$ and the observed multivariate time series $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N\}$, the probability of the observations is given by:

$$p(\mathbf{X}) = \sum_{\text{all } \mathbf{Z}} p(\mathbf{Z}, \mathbf{X}) \tag{1}$$

where $p(\mathbf{Z}, \mathbf{X})$ defines the complete-data likelihood as in HMM models [9] and takes the following form:

$$p(\mathbf{Z}, \mathbf{X}) = p(z_1) \prod_{n=2}^{N} p(z_n|z_{n-1}) \prod_{n=1}^{N} p(\mathbf{x}_n|z_n) \tag{2}$$

The model parameters are $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta)$ where $\boldsymbol{\pi} = \{\pi_j\} : \pi_j = p(z_1 = j)$ are the initial state probabilities, $\mathbf{A} = \{a_{ij}\} : a_{ij} = p(z_n = j|z_{n-1} = i)$ are the transition state probabilities, and $\{\mathbf{Y}, \beta\} : p(\mathbf{x}_n|z_n = j) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2}\|\mathbf{x}_n - \mathbf{y}_j\|^2\right)$ are the emission probabilities, which are controlled by spherical Gaussian distributions with common inverse variance β and a matrix $\mathbf{Y}$ of K centroids $\mathbf{y}_j$, $1 \leq j \leq K$.

For mathematical convenience, it is useful defining a state in the vectorial form $\mathbf{z}_{j,n}$ such that it returns 1 if z_n is in state j, and zero otherwise. Using this notation, the initial state probabilities, the transition state probabilities and the emission probabilities are defined as:

$$p(z_1|\pi) = \prod_{j=1}^{K} \pi_j^{\mathbf{z}_{j,1}} \tag{3}$$

$$p(z_n|z_{n-1}, \mathbf{A}) = \prod_{i=1}^{K} \prod_{j=1}^{K} a_{ij}^{\mathbf{z}_{j,n}\mathbf{z}_{i,n-1}} \tag{4}$$

517

$$p(\mathbf{x}_n|z_n, \mathbf{Y}, \beta) =$$
$$\left(\frac{\beta}{2\pi}\right)^{D/2} \prod_{j=1}^{K} \left\{ \exp\left(-\frac{\beta}{2}\|\mathbf{x}_n - \mathbf{y}_j\|^2\right) \right\}^{\mathbf{z}_{j,n}} \qquad (5)$$

Eqs. 3 to 5 lead to the definition of the complete data log-likelihood as:

$$
\begin{aligned}
\ln p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\Theta}) &= \sum_{j=1}^{K} \mathbf{z}_{j,1} \ln \pi_j \\
&+ \sum_{n=2}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{z}_{i,n-1} \mathbf{z}_{j,n} \ln a_{ij} \\
&+ \frac{ND}{2} \ln\left(\frac{\beta}{2\pi}\right) \\
&- \frac{\beta}{2} \sum_{n=1}^{N} \sum_{j=1}^{K} \mathbf{z}_{j,n} \|\mathbf{x}_n - \mathbf{y}_j\|^2 \qquad (6)
\end{aligned}
$$

Parameter estimation can be accomplished in GTM-TT by maximum likelihood using the EM algorithm, in a similiar fashion to HMMs. Details can be found in [10].

III. BAYESIAN GTM THROUGH TIME

Although the ML framework is widely used for parameter optimization, it shows two important weaknesses: Its maximization process does not take into account the model complexity and it tends to overfit the model to the training data. The complexity in GTM-TT is related to the number of hidden states, their the degree of connectivity and the dimension of the hidden space. Usually, for visualization purposes, the dimension of the hidden space is limited to be less or equal to three. The number of hidden states and the maximum number of possible state transitions are strictly correlated by a squared power. In order to avoid overfitting, researchers have commonly limited the complexity of their models by restricting the number of possible state transitions [2] or by fixing the transition state probabilities a priori [11]. The alternative technique of cross-validation is computationally expensive and it could require large amounts of data to obtain low-variance estimates of the expected test errors.

A more elegant solution to control overfitting and complexity is providing a Bayesian formulation for the model [12], [13]. The Bayesian approach treats the parameters as unknown quantities and provides probability distributions for their priors. Bayes' theorem can then be used to infer the posterior distributions over the parameters. The model parameters can thus be considered as hidden variables and integrated out to describe the marginal likelihood as:

$$p(\mathbf{X}) = \int p(\boldsymbol{\Theta}) p(\mathbf{X}|\boldsymbol{\Theta}) d\boldsymbol{\Theta},$$
$$\text{where } \boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta) \qquad (7)$$

If an independent distribution is assumed for each parameter, then:

$$p(\boldsymbol{\Theta}) = p(\boldsymbol{\pi}) p(\mathbf{A}) p(\mathbf{Y}) p(\beta) \qquad (8)$$

Taking into account Eqs. 1, 7 and 8, the marginal likelihood in GTM-TT can be expressed, similarly to HMM [7], as:

$$
\begin{aligned}
p(\mathbf{X}) = &\int p(\boldsymbol{\pi}) \int p(\mathbf{A}) \int p(\mathbf{Y}) \int p(\beta) \\
&\sum_{\text{all } \mathbf{Z}} p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta) \, d\beta d\mathbf{Y} d\mathbf{A} d\boldsymbol{\pi} \qquad (9)
\end{aligned}
$$

Although there are many possible prior distributions to choose from, the conjugates of the distributions defined in Eqs. 3 to 5 are a reasonable choice. In this way, a set of prior distributions is defined as follows:

$$
\begin{aligned}
p(\boldsymbol{\pi}) &= \texttt{Dir}\left(\{\pi_1, \ldots, \pi_K\} | \boldsymbol{\nu}\right) \\
p(\mathbf{A}) &= \prod_{j=1}^{K} \texttt{Dir}\left(\{a_{j1}, \ldots, a_{jK}\} | \boldsymbol{\lambda}\right) \\
p(\mathbf{Y}) &= \left[(2\pi)^K |\mathbf{C}|\right]^{-D/2} \prod_{d=1}^{D} \exp\left(-\frac{1}{2}\mathbf{y}_{(d)}^T \mathbf{C}^{-1} \mathbf{y}_{(d)}\right) \\
p(\beta) &= \Gamma\left(\beta | d_\beta, s_\beta\right)
\end{aligned}
$$

where $\texttt{Dir}(\cdot)$ represents the Dirichlet distribution; and $\Gamma(\cdot)$ is the Gamma distribution. The vector $\boldsymbol{\nu}$, the matrix $\boldsymbol{\lambda}$ and the scalars d_β and s_β correspond to the hyperparameters of the model which are fixed a priori. The prior over the parameter $\mathbf{Y}$ defines the mapping from the hidden states to the data space as a GP, where $\mathbf{y}_{(d)}$ is each of the row vectors (centroids) of the matrix $\mathbf{Y}$ and $\mathbf{C}$ is a matrix where each element is a covariance function that can be defined as

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

The α parameter controls the flexibility of the mapping from the latent space to the data space. The vector $\mathbf{u}_j$, $j = 1 \ldots K$ corresponds to the state j in a latent space of usually lower dimension than that of the data space. Thus, a topography over the states is defined by the GP as in the standard GTM.

Unfortunately, Eq. 9 is analytically intractable. In the following section, we provide the details of its approximation using Variational inference techniques.

IV. VARIATIONAL BAYESIAN INFERENCE FOR GTM-TT

A. The Variational Bayesian EM Algorithm

Variational inference allows approximating the marginal log-likelihood through Jensen's inequality as follows:

$$
\begin{aligned}
\ln p(\mathbf{X}) &= \ln \int \sum_{\text{all } \mathbf{Z}} p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}) d\boldsymbol{\Theta} \\
&\geq \int \sum_{\text{all } \mathbf{Z}} q(\boldsymbol{\Theta}, \mathbf{Z}) \ln \frac{p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta})}{q(\boldsymbol{\Theta}, \mathbf{Z})} d\boldsymbol{\Theta} \\
&= F(q(\boldsymbol{\Theta}, \mathbf{Z}))
\end{aligned}
$$

The function $F\left(q\left(\boldsymbol{\Theta}, \mathbf{Z}\right)\right)$ is a lower bound such that its convergence guarantees the convergence of the marginal likelihood. The goal in variational inference is choosing a suitable form for the approximate density $q\left(\boldsymbol{\Theta}, \mathbf{Z}\right)$ 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 $q\left(\boldsymbol{\Theta}, \mathbf{Z}\right)$ is based on the assumption that the hidden states $\mathbf{Z}$ and the parameters $\boldsymbol{\Theta}$ are independently distributed, i.e. $q\left(\boldsymbol{\Theta}, \mathbf{Z}\right) = q\left(\boldsymbol{\Theta}\right) q\left(\mathbf{Z}\right)$. Thereby, a Variational EM algorithm can be derived [7]:

VBE-Step:

$$q\left(\mathbf{Z}\right)^{(\mathrm{new})} \leftarrow \underset{q(\mathbf{Z})}{\operatorname{argmax}} F\left(q\left(\mathbf{Z}\right)^{(\mathrm{old})}, q\left(\boldsymbol{\Theta}\right)\right) \qquad (10)$$

VBM-Step:

$$q\left(\boldsymbol{\Theta}\right)^{(\mathrm{new})} \leftarrow \underset{q(\boldsymbol{\Theta})}{\operatorname{argmax}} F\left(q\left(\mathbf{Z}\right)^{(\mathrm{new})}, q\left(\boldsymbol{\Theta}\right)\right) \qquad (11)$$

B. Variational Bayesian EM for GTM-TT

1) The VBE Step: The expression $q\left(\mathbf{Z}\right)$ is estimated using Eq. 6 in Eq. 10, so that:

$$
\begin{aligned}
\ln q\left(\mathbf{Z}\right) = & \left\langle \sum_{j=1}^{K} \mathbf{z}_{j,1} \ln \pi_j \right\rangle_{q(\boldsymbol{\pi})} \\
& + \left\langle \sum_{n=2}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{z}_{i,n-1} \mathbf{z}_{j,n} \ln a_{ij} \right\rangle_{q(\mathbf{A})} \\
& + \left\langle \frac{ND}{2} \ln \left(\frac{\beta}{2\pi} \right) \right\rangle_{q(\beta)} \\
& - \left\langle \frac{\beta}{2} \sum_{n=1}^{N} \sum_{j=1}^{K} \mathbf{z}_{j,n} \|\mathbf{x}_n - \mathbf{y}_j\|^2 \right\rangle_{q(\mathbf{Y}, \beta)} \\
& - \ln \tilde{\mathcal{Z}}\left(\mathbf{X}\right)
\end{aligned} \qquad (12)
$$

where $\ln \tilde{\mathcal{Z}}\left(\mathbf{X}\right)$ is a normalization constant that depends on $\mathbf{X}$. Although it is expressed here in terms of the mean of the parameters of the model, this equation has a similar form to Eq. 6. Furthermore, a modified forward-backward procedure [7] can be used to solve it as follows:

$$\alpha\left(j, n\right) = \frac{1}{\tilde{\zeta}\left(\mathbf{x}_n\right)} \left[\sum_{i=1}^{K} \alpha\left(i, n-1\right) \tilde{a}_{ij} \right] \tilde{p}\left(\mathbf{x}_n | z_n = j\right)$$
$$\text{with } \alpha\left(j, 1\right) = \tilde{\pi}_j$$

$$\beta\left(j, n\right) = \sum_{i=1}^{K} \beta\left(i, n+1\right) \tilde{a}_{ij} \tilde{p}\left(\mathbf{x}_{n+1} | z_{n+1} = i\right)$$
$$\text{with } \beta\left(j, N\right) = 1$$

where $\tilde{\pi}_j$ and $\tilde{a}_{ij}$ are the estimated parameters; $\tilde{p}\left(\mathbf{x}_n | z_n = j\right)$ and $\tilde{p}\left(\mathbf{x}_{n+1} | z_{n+1} = i\right)$ are the emission probabilities calculated using the estimated parameters $\mathbf{Y}$ and β; and $\tilde{\zeta}\left(\mathbf{x}_n\right)$ is the normalization constant, related to the normalization constant of Eq. 12 by the expression: $\prod_{n=1}^{N} \tilde{\zeta}\left(\mathbf{x}_n\right) = \tilde{\mathcal{Z}}\left(\mathbf{X}\right)$

2) The VBM Step: The variational distribution $q\left(\boldsymbol{\Theta}\right)$ can be approximated to the product of the variational distribution of each one of the parameters if they are assumed to be independent and identically distributed. If so, $q\left(\boldsymbol{\Theta}\right)$ is expressed as:

$$q\left(\boldsymbol{\Theta}\right) = q\left(\boldsymbol{\pi}\right) q\left(\mathbf{A}\right) q\left(\mathbf{Y}\right) q\left(\beta\right)$$

where natural choices of $q\left(\boldsymbol{\pi}\right)$, $q\left(\mathbf{A}\right)$, $q\left(\mathbf{Y}\right)$ and $q\left(\beta\right)$ are similar distributions to the priors $p\left(\boldsymbol{\pi}\right)$, $p\left(\mathbf{A}\right)$, $p\left(\mathbf{Y}\right)$ and $p\left(\beta\right)$, respectively. Thus,

$$
\begin{aligned}
q\left(\boldsymbol{\pi}\right) &= \texttt{Dir}\left(\{\pi_1, \ldots, \pi_K\} | \tilde{\boldsymbol{\nu}}\right) & (13) \\
q\left(\mathbf{A}\right) &= \prod_{j=1}^{K} \texttt{Dir}\left(\{a_{j1}, \ldots, a_{jK}\} | \tilde{\boldsymbol{\lambda}}\right) & (14) \\
q\left(\mathbf{Y}\right) &= \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_{(d)} | \tilde{\mathbf{m}}^{(d)}, \tilde{\boldsymbol{\Sigma}}\right) & (15) \\
q\left(\beta\right) &= \Gamma\left(\beta | \tilde{d}_\beta, \tilde{s}_\beta\right) & (16)
\end{aligned}
$$

Now, using Eqs. 13 to 16 in Eq. 11, the following expressions for the variational parameters $\tilde{\boldsymbol{\nu}}$, $\tilde{\boldsymbol{\lambda}}$, $\tilde{\boldsymbol{\Sigma}}$, $\tilde{\mathbf{m}}$, $\tilde{d}_\beta$ and $\tilde{s}_\beta$ can be obtained:

$$
\begin{aligned}
\tilde{\nu}_j &= \nu_j + \langle \mathbf{z}_{j,1} \rangle \\
\tilde{\lambda}_{i,j} &= \lambda_{i,j} + \sum_{n=2}^{N} \langle \mathbf{z}_{i,n-1} \mathbf{z}_{j,n} \rangle \\
\tilde{\boldsymbol{\Sigma}} &= \left(\langle \beta \rangle \sum_{n=1}^{N} \mathbf{G}_n + \mathbf{C}^{-1} \right)^{-1} \\
\tilde{\mathbf{m}}_{(d)} &= \langle \beta \rangle \tilde{\boldsymbol{\Sigma}} \sum_{n=1}^{N} x_{nd} \langle \mathbf{z}_n \rangle \\
\tilde{d}_\beta &= d_\beta + \frac{ND}{2} \\
\tilde{s}_\beta &= s_\beta + \frac{1}{2} \sum_{n=1}^{N} \sum_{j=1}^{K} \langle \mathbf{z}_{j,n} \rangle \left\langle \|\mathbf{x}_n - \mathbf{y}_j\|^2 \right\rangle
\end{aligned}
$$

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:

$$
\begin{aligned}
\langle \beta \rangle &= \frac{\tilde{d}_\beta}{\tilde{s}_\beta} \\
\langle \|\mathbf{x}_n - \mathbf{y}_j\| \rangle &= D \tilde{\Sigma}_{jj} + \|\mathbf{x}_n - \tilde{\mathbf{m}}_j\|^2
\end{aligned}
$$

Details on these calculations are provided in [10].

C. Lower Bound Function

The lower bound function for GTM-TT is obtained through a similar procedure to the one described in [7], although, here, we must take into account the variational distributions of the parameters $\mathbf{Y}$ and β. The solution for the lower bound is:

$$F\left(q\left(\mathbf{\Theta}\right),q\left(\mathbf{Z}\right)\right) = \int q\left(\boldsymbol{\pi}\right)\ln\frac{p\left(\boldsymbol{\pi}\right)}{q\left(\boldsymbol{\pi}\right)}d\boldsymbol{\pi}$$
$$+ \int q\left(\mathbf{A}\right)\ln\frac{p\left(\mathbf{A}\right)}{q\left(\mathbf{A}\right)}d\mathbf{A}$$
$$+ \int q\left(\mathbf{Y}\right)\ln\frac{p\left(\mathbf{Y}\right)}{q\left(\mathbf{Y}\right)}d\mathbf{Y}$$
$$+ \int q\left(\beta\right)\ln\frac{p\left(\beta\right)}{q\left(\beta\right)}d\beta + \ln\tilde{\mathcal{Z}}\left(\mathbf{X}\right)$$

This equation implies that only the computation of the KL-divergence between the variational and the prior distribution for each parameter and the normalization constant is necessary to evaluate the lower bound function. Furthermore, the computation of the KL-divergence is straightforward because the distributions are known.

V. Experiments

A. Experimental Design

The performance of the proposed Variational GTM-TT model will be illustrated using synthetic and real datasets: The synthetic one is a simple 3-variate time series, each one consisting of 500 points sampled from a different linear combination of sinusoidal functions. The real one is the *Shuttle* dataset (available from www.cs.ucr.edu/~eamonn). These 6-variate time series consist of 1000 data points obtained from various inertial sensors from Space Shuttle mission STS-57. They contain subsequences of little variability followed by sudden transition periods. Both datasets were contaminated by Gaussian noise of increasing standard deviation, from 0.01 to 0.20, for model training. The uncontaminated datasets were used for test. The resulting datasets were clustered and visualized ten times using both the standard GTM-TT with GP prior and the proposed Variational GTM-TT, with random initialization.

B. Robustness of the Variational GTM Through Time in the Presence of Noise

The goal of this experiment is assessing and comparing the robustness of the standard GTM-TT with GP prior and the Variational GTM-TT in the presence of different levels of noise. Such assessment and comparison is made in terms of a test log-likelihood, defined as $\ln p\left(\mathbf{X}'|\mathbf{X}\right)$, where $\mathbf{X}'$ is the test dataset. Different numbers of states were used, from 9 to 144 for the *Shuttle* dataset and from 4 to 36 for the synthetic dataset (squares of natural numbers, corresponding to square visualization grids), in order to assess the effect of this parameter in the performance of the models. Figs. 1 and 2 display the test log-likelihood results for both models, for different numbers of states and for data contaminated by two levels of noise (many other intermediate levels were explored, conforming to the pattern described next). The results are consistent and their interpretation is straightforward: The variational formulation consistently outperforms the standard GTM-TT with GP prior, for all levels of added noise and across the whole range of number of states, for both the

synthetic and real datasets. This indicates that the Variational GTM-TT has dealt better with the problem of overfitting, being able to generalize better as a result.
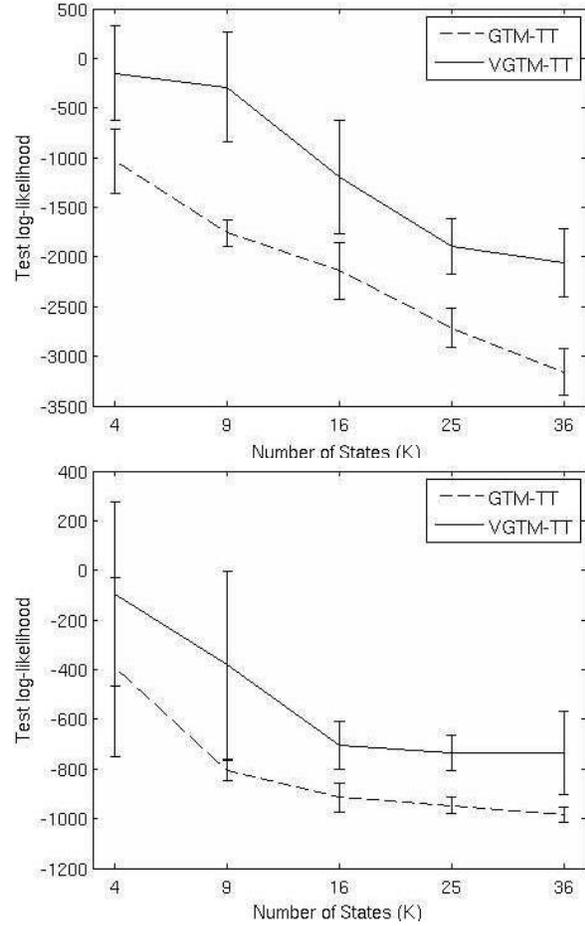

Fig. 1. Average test log-likelihood for the proposed variational GTM-TT (solid line) and standard GTM-TT with GP prior (dashed line) over 10 runs of the algorithms, for different numbers of states and different levels of added Gaussian noise using the synthetic dataset. Top: 0.05 standard deviation, bottom: 0.20. The vertical bars span from the average minus one standard deviation of the 10 runs, to the average plus one standard deviation.

C. Time Series Visualization Using Variational GTM Through Time

As already mentioned, one of the most interesting capabilities of Variational GTM-TT is that of providing a visualization of the multivariate time series and their evolution over the model states. According to the results provided in section V-B, we would expect the state representation resulting from the training of the standard GTM-TT with GP prior to be more affected by the presence of noise than that of its variational counterpart. This would entail a profusion of small states reflecting the noise in the former, paired with a jittery trajectory through them, while a more parsimonious
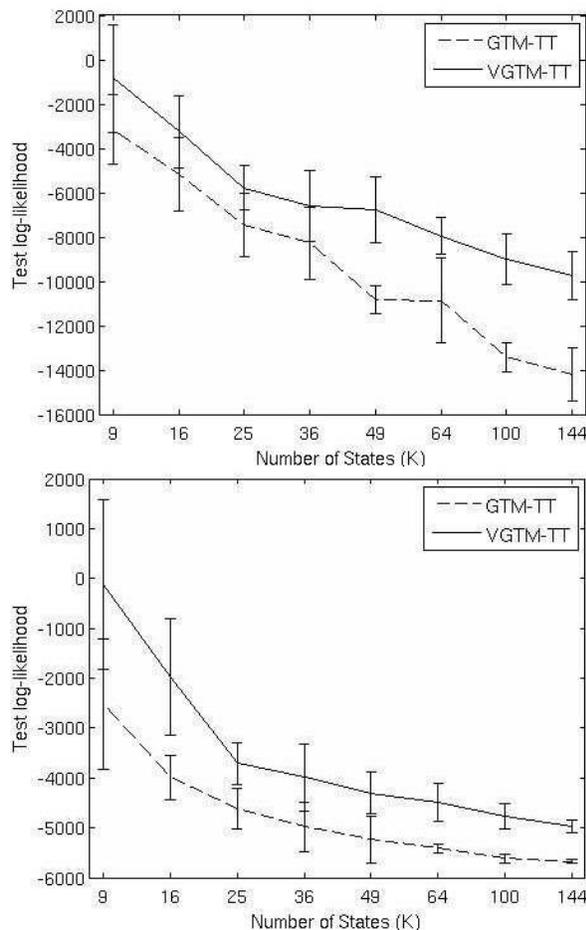
2008 International Joint Conference on Neural Networks (IJCNN 2008)

Fig. 2. Average test log-likelihood for *Shuttle* dataset, using the same settings of Fig. 1
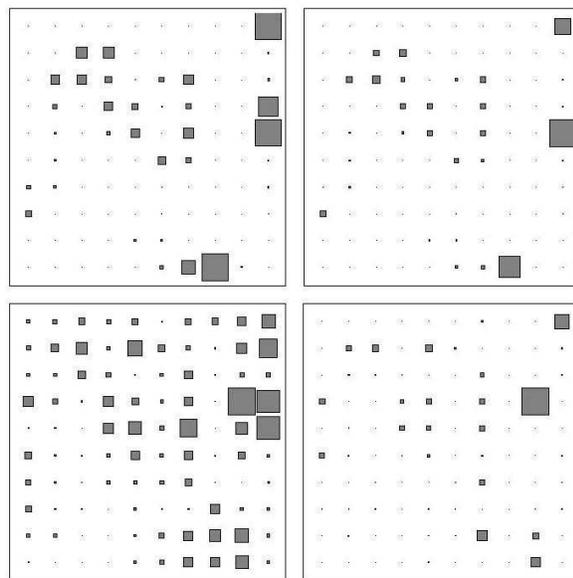

Fig. 3. State membership maps generated by means of the mode projection of the *Shuttle* time series into the latent space of states, as described in the main text. Top row: Variational GTM-TT results for training (left) and test (right). Bottom row: GTM-TT with GP prior results for training (left) and test (right). In all representations, the relative size of each square (state) is proportional to the number of data points assigned to it.

state distribution and a more stable trajectory through the state space would be expected from the training results of the latter.

The results displayed in Fig. 3 confirm these expectations. The *Shuttle* time series visualization is accomplished through state membership maps generated by means of the mode projection [1] of the data into the latent space of states, given by $\mathbf{u}_n^{mode} = \mathrm{argmax}_j \; z_{jn}$, where the variational parameter $\langle z_{jn} \rangle$ was used. The state membership map resulting from the training of the noisy data with Variational GTM-TT reproduces, quite faithfully, the corresponding map of the noise-free test data. On the contrary, the map resulting from the training of the noisy data with the standard GTM-TT with GP prior suffers of an undue proliferation of states that reflect the occurrence of overfitting.

VI. Conclusions

The presence of noise is commonplace in multivariate time series. In many real applications, it may shadow the informative patterns that might be present in the signal, making the process of knowledge extraction difficult. This could entail poorer predictions over time, or more ambiguous signal source separation and identification. For these reasons, time series analysis should benefit from the definition of models that behave robustly in the presence of noise, preventing data overfitting. In this brief paper, we have laid the theoretical foundations of Variational GTM-TT, an unsupervised model with those characteristics, capable of clustering and visualizing the underlying structure of multivariate time series in the presence of noise. Its performance has been illustrated with an experimental example.

Future research will be devoted to test the model in detail, using both artificial and real datasets of various characteristics, as well as to compare it with alternative techniques.

Finally, we remark that the computational complexity of Variational GTM-TT does not increase with respect to that of the standard GTM-TT with GP prior. On the other hand, the formulation of Variational GTM-TT introduces a heavier computational load as compared to the standard GTM-TT, as usual in most formulations involving Bayesian inference. However, there was no significant increase in the running times for the experiments reported in this paper. A more thorough study of the computational efficiency of the method will also be the matter of future research.

References

[1] C. M. Bishop, M. Svensén, and C. K. I. Williams, "GTM: The Generative Topographic Mapping," *Neural Comput.*, vol. 10, no. 1,

pp. 215–234, 1998.

[2] C. M. Bishop, G. Hinton, and I. Strachan, "GTM Through Time," in *IEE Fifth Int. Conf. on Artif. Neural Net., Cambridge, U.K.*, 1997, pp. 111–116.

[3] I. Olier and A. Vellido, "Capturing the dynamics of multivariate time series through visualization using Generative Topographic Mapping Through Time," in *Proceedings of IEEE ICEIS 2006, Islamabad, Pakistan*, 2006, pp. 492–497.

[4] ——, "Time Series Relevance Determination through a topology-constrained Hidden Markov model," in *The 7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'06). Lect. Notes Comput. Sc.*, vol. 4224, 2006, pp. 40–47.

[5] C. M. Bishop, M. Svensén, and C. K. I. Williams, "Developments of the Generative Topographic Mapping," *Neurocomputing*, vol. 21, no. 1–3, pp. 203–224, 1998.

[6] A. Vellido, W. El-Deredy, and P. J. G. Lisboa, "Selective smoothing of the Generative Topographic Mapping," *IEEE T. Neural Networ.*, vol. 14, no. 4, pp. 847–852, 2003.

[7] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, The Gatsby Computational Neuroscience Unit, Univ. College London, 2003.

[8] I. Olier and A. Vellido, "Variational GTM," in *The 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07). Lect. Notes Comput. Sc.*, vol. 4881, 2007, pp. 77–86.

[9] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," in *Proc. of the IEEE*, vol. 77, no. 2, 1989, pp. 257–285.

[10] I. Olier and A. Vellido, "A variational formulation for GTM Through Time: Theoretical foundations," Technical University of Catalonia (UPC), Tech. Rep. LSI-07-38-R, 2007.

[11] A. Kabán and M. Girolami, "A dynamic probabilistic model to visualise topic evolution in text streams," *J. Intell. Inf. Syst.*, vol. 18, no. 2–3, pp. 107–125, 2002.

[12] A. Stolcke and S. Omohundro, "Hidden Markov model induction by Bayesian model merging," in *Advances in Neural Information Processing Systems*, S. J. Hanson, J. D. Cowan, and C. L. Giles, Eds., vol. 5. San Francisco, CA. Morgan Kauffmann, 1993, pp. 11–18.

[13] D. J. C. MacKay, "Ensemble learning for Hidden Markov Models," Cavendish Laboratory, University of Cambridge, Tech. Rep., 1997, unpublished Manuscript.