A Variational Formulation for GTM Through Time: Theoretical Foundations

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 latent variable model that, in its standard 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 technical report, we provide the theoretical foundations of the reformulation of GTM-TT within the Variational Bayesian framework. This approach, in its application, should naturally handle the presence of noise in the time series, helping to avert the problem of data overfitting.

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. 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, 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.

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, 8]. 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 [9]. 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 technical report we extend such Variational approach to the analysis of multivariate time series, defining the theoretical foundations of a model known as Variational GTM-TT.

The remaining of this report is organized as follows: First, in section 2, an introduction to the original GTM-TT [2] is provided. Section 3 provides a Bayesian framework for GTM-TT. This is followed, in section 4, by the description of the proposed Variational Bayesian inference method for GTM-TT in some detail.

2 The Standard Generative Topographic Mapping Through Time

2.1 The GTM-TT Model

The GTM-TT was introduced in [2] as a way to extend the standard GTM [1] model for the analysis of context-dependent data sets such as multivariate time series. 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:

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

where $p\left(\mathbf{Z}, \mathbf{X}\right)$ defines the complete-data likelihood as in HMM models [10] and takes the following form:

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

The model parameters are $\mathbf{\Theta} = \left(\boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta\right)$ where $\boldsymbol{\pi} = \{\pi_j\} : \pi_j = p\left(z_1 = j\right)$ are the initial state probabilities, $\mathbf{A} = \{a_{ij}\} : a_{ij} = p\left(z_n = j | z_{n-1} = i\right)$ are the transition state probabilities, and $\{\mathbf{Y}, \beta\} : p\left(\mathbf{x}_n | z_n = j\right) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2} \left\|\mathbf{x}_n - \mathbf{y}_j\right\|^2\right)$ 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:

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

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

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

Note that Eqs. 3 and 4 are multinomial distributions, while Eq. 5 defines a mixture of Gaussian distributions in which the parameter $\mathbf{z}_{j,n}$ is a *selector* of these Gaussians. Thus, the log-complete-data likelihood is defined using the above formulation as follows:

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

2.2 Learning through Maximum Likelihood

Parameter estimation in GTM-TT can be accomplished by ML using the EM algorithm in a similar fashion to HMMs [10, 11]. The following expectations $\langle\cdot\rangle$ are estimated in the E-step:

$$\gamma_{j,n} = \langle \mathbf{z}_{j,n} \rangle = \frac{\alpha\left(j,n\right)\beta\left(j,n\right)}{\sum_{j'=1}^{K}\alpha\left(j',n\right)\beta\left(j',n\right)} \tag{7}$$

$$\xi_{i,j,n} = \langle \mathbf{z}_{i,n-1}\mathbf{z}_{j,n} \rangle = \frac{\alpha\left(i,n-1\right)a_{ij}p\left(\mathbf{x}_n|\mathbf{z}_{j,n}\right)\beta\left(j,n\right)}{\sum_{i'=1}^{K}\sum_{j'=1}^{K}\alpha\left(i',n-1\right)a_{i'j'}p\left(\mathbf{x}_n|\mathbf{z}_{j',n}\right)\beta\left(j',n\right)} \tag{8}$$

where $\alpha\left(j,n\right) = p\left(\mathbf{z}_n|\mathbf{X}_1^n\right)$ and $\beta\left(j,n\right) = p\left(\mathbf{X}_{n+1}^N|\mathbf{z}_n\right)$ are obtained using the forward-backward recursion algorithm. The notations $\mathbf{X}_1^n$ and $\mathbf{X}_{n+1}^N$ represent, in turn, the subsequences from 1 to n and from $n+1$ to N.

The parameters a_{ij} and π_j are estimated in the M-step as:

$$\pi_j = \gamma_{j,1} \tag{9}$$

$$a_{ij} = \frac{\sum_{n=2}^{N}\xi_{i,j,n}}{\sum_{n=2}^{N}\gamma_{i,n}} \tag{10}$$

Finally, parameters $\mathbf{Y}$ and β are estimated as in the standard GTM using a GP for the mapping from the hidden states to the data space setting a prior distribution over $\mathbf{Y}$ defined by:

$$P\left(\mathbf{Y}\right) = \left(2\pi\right)^{-KD/2}\left|\mathbf{C}\right|^{-D/2}\prod_{d=1}^{D}\exp\left(-\frac{1}{2}\mathbf{y}_{(d)}^T\mathbf{C}^{-1}\mathbf{y}_{(d)}\right) \tag{11}$$

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

$$C_{ij} = C\left(\mathbf{u}_i,\mathbf{u}_j\right) = \nu\exp\left(-\frac{\left\|\mathbf{u}_i-\mathbf{u}_j\right\|^2}{2\alpha^2}\right), \quad i,j = 1\ldots K \tag{12}$$

and where parameter ν is usually fixed a priori. The α 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 [12]. 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. Consequently, the updating expressions for $\mathbf{Y}$ and β are as for the standard GTM.

The parameter $\mathbf{Y}$ is estimated from:

$$\left(\mathbf{G}+\beta^{-1}\mathbf{C}^{-1}\right)\mathbf{Y} = \mathbf{\Gamma}\mathbf{X} \tag{13}$$

where $\mathbf{\Gamma}$ is the matrix of state expectations with elements $\langle\mathbf{z}_{j,n}\rangle$ that were previously defined by Eq. 7, $\mathbf{G}$ is a diagonal matrix formed by the elements $g_{jj} = \sum_n^N \gamma_{j,n}$, and $\mathbf{C}$ is the matrix of covariance functions.

The parameter β is estimated as:

$$\beta^{-1} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{j=1}^{K}\gamma_{j,n}\|\mathbf{x}_n-\mathbf{y}_j\|^2 \tag{14}$$

3 Bayesian GTM Through Time

Although the ML framework is widely used for parameter optimization, it shows two significant 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 [13]. 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 [14, 15]. 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:

$$p(\mathbf{X}) = \int p(\boldsymbol{\Theta})\, p(\mathbf{X}|\boldsymbol{\Theta})\, d\boldsymbol{\Theta}, \quad \text{where } \boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta) \tag{15}$$

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

$$p(\boldsymbol{\Theta}) = p(\boldsymbol{\pi})\, p(\mathbf{A})\, p(\mathbf{Y})\, p(\beta) \tag{16}$$

Taking into account Eqs. 1, 15 and 16, the marginal likelihood in GTM-TT can be expressed, similarly to HMM [7], as:

$$p(\mathbf{X}) = \int p(\boldsymbol{\pi}) \int p(\mathbf{A}) \int p(\mathbf{Y}) \int p(\beta) \sum_{\text{all } \mathbf{Z}} p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta)\, d\beta d\mathbf{Y} d\mathbf{A} d\boldsymbol{\pi} \tag{17}$$

Although there are many possible prior distributions to choose from, the conjugates of the distributions defined in Eqs. 3 and 4 and the GP defined in Eq. 11 are a good choice. In this way, a set of prior distributions is defined as follows:

$$p(\boldsymbol{\pi}) = \mathtt{Dir}\left(\{\pi_1, \ldots, \pi_K\}\,|\boldsymbol{\nu}\right) \tag{18}$$

$$p(\mathbf{A}) = \prod_{j=1}^{K} \mathtt{Dir}\left(\{a_{j1}, \ldots, a_{jK}\}\,|\boldsymbol{\lambda}\right) \tag{19}$$

$$p(\beta) = \Gamma\left(\beta|d_\beta, s_\beta\right) \tag{20}$$

where $\mathtt{Dir}(\cdot)$ represents the Dirichlet distribution; and $\Gamma(\cdot)$ is the Gamma distribution. The prior over the parameter $\mathbf{Y}$ was previously defined by Eq. 11.

Unfortunately, Eq. 17 is analytically intractable. In the following section of the report, we provide the details of its approximation using Variational inference techniques.

4 Variational Bayesian Inference for GTM-TT

4.1 The Variational Bayesian EM Algorithm

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

$$
\begin{aligned}
\ln p(\mathbf{X}) &= \ln \int \sum_{\text{all } \mathbf{Z}} p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\Theta})\, p(\boldsymbol{\Theta})\, d\boldsymbol{\Theta} \\
&= \ln \int \sum_{\text{all } \mathbf{Z}} q(\boldsymbol{\Theta}, \mathbf{Z}) \frac{p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{q(\boldsymbol{\Theta}, \mathbf{Z})} d\boldsymbol{\Theta} \\
&\geq \int \sum_{\text{all } \mathbf{Z}} q(\boldsymbol{\Theta}, \mathbf{Z}) \ln \frac{p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{q(\boldsymbol{\Theta}, \mathbf{Z})} d\boldsymbol{\Theta} \\
&= F(q(\boldsymbol{\Theta}, \mathbf{Z}))
\end{aligned}
\tag{21}
$$

The function $F(q(\boldsymbol{\Theta}, \mathbf{Z}))$ 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 density $q(\boldsymbol{\Theta}, \mathbf{Z})$ in such a way that $F(q)$ can be readily evaluated and yet which is sufficiently flexible that the bound is reasonably tight. A reasonable approximation for $q(\boldsymbol{\Theta}, \mathbf{Z})$ is based on the assumption that the hidden states $\mathbf{Z}$ and the parameters $\boldsymbol{\Theta}$ are independently distributed, i.e. $q(\boldsymbol{\Theta}, \mathbf{Z}) = q(\boldsymbol{\Theta})\, q(\mathbf{Z})$. Thereby, a Variational EM algorithm can be derived [7]:

VBE-Step:

$$
q(\mathbf{Z})^{(\text{new})} \leftarrow \underset{q(\mathbf{Z})}{\text{argmax}}\, F\left(q(\mathbf{Z})^{(\text{old})}, q(\boldsymbol{\Theta})\right) \tag{22}
$$

VBM-Step:

$$
q(\boldsymbol{\Theta})^{(\text{new})} \leftarrow \underset{q(\boldsymbol{\Theta})}{\text{argmax}}\, F\left(q(\mathbf{Z})^{(\text{new})}, q(\boldsymbol{\Theta})\right) \tag{23}
$$

4.2 Variational Bayesian EM for GTM-TT

4.2.1 The VBE Step

The expression $q(\mathbf{Z})$ is estimated using Eq. 6 in Eq. 22, so that:

$$
\begin{aligned}
\ln q(\mathbf{Z}) = {} & \left\langle \sum_{j=1}^{K} \mathbf{z}_{j,1} \ln \pi_j \right\rangle_{q(\boldsymbol{\pi})} + \left\langle \sum_{n=2}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{z}_{i,n-1} \mathbf{z}_{j,n} \ln a_{ij} \right\rangle_{q(\mathbf{A})} \\
& + \left\langle \frac{ND}{2} \ln\left(\frac{\beta}{2\pi}\right) \right\rangle_{q(\beta)} - \left\langle \frac{\beta}{2} \sum_{n=1}^{N} \sum_{j=1}^{K} \mathbf{z}_{j,n} \|\mathbf{x}_n - \mathbf{y}_j\|^2 \right\rangle_{q(\mathbf{Y},\beta)} \\
& - \ln \tilde{\mathcal{Z}}(\mathbf{X})
\end{aligned}
\tag{24}
$$

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

$$
\alpha(j, n) = \frac{1}{\tilde{\zeta}(\mathbf{x}_n)} \left[\sum_{i=1}^{K} \alpha(i, n-1)\, \tilde{a}_{ij} \right] \tilde{p}(\mathbf{x}_n | z_n = j) \quad \text{with } \alpha(j, 1) = \tilde{\pi}_j \tag{25}
$$

$$
\beta(j, n) = \sum_{i=1}^{K} \beta(i, n+1)\, \tilde{a}_{ij} \tilde{p}(\mathbf{x}_{n+1} | z_{n+1} = i) \quad \text{with } \beta(j, N) = 1 \tag{26}
$$

where $\tilde{\pi}_j$ and $\tilde{a}_{ij}$ are the estimated parameters; $\tilde{p}(\mathbf{x}_n | z_n = j)$ and $\tilde{p}(\mathbf{x}_{n+1} | z_{n+1} = i)$ are the emission probabilities calculated using the estimated parameters $\mathbf{Y}$ and β; and $\tilde{\zeta}(\mathbf{x}_n)$ is the normalization constant, which is related to the normalization constant of Eq. 24 by the expression:

$$\prod_{n=1}^{N} \tilde{\zeta}\left(\mathbf{x}_n\right) = \tilde{\mathcal{Z}}\left(\mathbf{X}\right) \tag{27}$$

4.2.2 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:

$$q\left(\boldsymbol{\Theta}\right) = q\left(\boldsymbol{\pi}\right) q\left(\mathbf{A}\right) q\left(\mathbf{Y}\right) q\left(\beta\right) \tag{28}$$

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,

$$q\left(\boldsymbol{\pi}\right) = \mathtt{Dir}\left(\left\{\pi_1, \ldots, \pi_K\right\} | \tilde{\boldsymbol{\nu}}\right) \tag{29}$$

$$q\left(\mathbf{A}\right) = \prod_{j=1}^{K} \mathtt{Dir}\left(\left\{a_{j1}, \ldots, a_{jK}\right\} | \tilde{\boldsymbol{\lambda}}\right) \tag{30}$$

$$q\left(\mathbf{Y}\right) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_{(d)} | \tilde{\mathbf{m}}^{(d)}, \tilde{\boldsymbol{\Sigma}}\right) \tag{31}$$

$$q\left(\beta\right) = \Gamma\left(\beta | \tilde{d}_\beta, \tilde{s}_\beta\right) \tag{32}$$

Now, using Eqs. 29 to 32 in Eq. 23, 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:

$$\begin{aligned}
\tilde{\nu}_j &= \nu_j + \left\langle \mathbf{z}_{j,1} \right\rangle & (33) \\
\tilde{\lambda}_{i,j} &= \lambda_{i,j} + \sum_{n=2}^{N} \left\langle \mathbf{z}_{i,n-1} \mathbf{z}_{j,n} \right\rangle & (34) \\
\tilde{\boldsymbol{\Sigma}} &= \left(\left\langle \beta \right\rangle \sum_{n=1}^{N} \mathbf{G}_n + \mathbf{C}^{-1} \right)^{-1} & (35) \\
\tilde{\mathbf{m}}_{(d)} &= \left\langle \beta \right\rangle \tilde{\boldsymbol{\Sigma}} \sum_{n=1}^{N} x_{nd} \left\langle \mathbf{z}_n \right\rangle & (36) \\
\tilde{d}_\beta &= d_\beta + \frac{ND}{2} & (37) \\
\tilde{s}_\beta &= s_\beta + \frac{1}{2} \sum_{n=1}^{N} \sum_{j=1}^{K} \left\langle \mathbf{z}_{j,n} \right\rangle \left\langle \left\| \mathbf{x}_n - \mathbf{y}_j \right\|^2 \right\rangle & (38)
\end{aligned}$$

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 $\left\langle \mathbf{z}_n \right\rangle$. The moments in the previous equations are defined as:

$$\begin{aligned}
\langle\beta\rangle &= \frac{\tilde{d}_\beta}{\tilde{s}_\beta} && (39)\\
\langle\|\mathbf{x}_n-\mathbf{y}_j\|\rangle &= \left\langle\left(\mathbf{x}_n-\mathbf{y}_j\right)^T\left(\mathbf{x}_n-\mathbf{y}_j\right)\right\rangle\\
&= \sum_{d=1}^{D} x_{nd}^2 - 2x_{nd}\left\langle y_{jd}\right\rangle + \left\langle y_{jd}^2\right\rangle\\
&= \sum_{d=1}^{D} x_{nd}^2 - 2x_{nd}\tilde{m}_j^{(d)} + \tilde{\Sigma}_{jj} + \left(\tilde{m}_j^{(d)}\right)^2\\
&= D\tilde{\tilde{\Sigma}}_{jj} + \|\mathbf{x}_n-\tilde{\mathbf{m}}_j\|^2 && (40)
\end{aligned}$$

Details on these calculations are provided in Appendix A.

4.3 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:

$$\begin{aligned}
F\left(q\left(\boldsymbol{\Theta}\right),q\left(\mathbf{Z}\right)\right) &= \int q\left(\boldsymbol{\pi}\right)\ln\frac{p\left(\boldsymbol{\pi}\right)}{q\left(\boldsymbol{\pi}\right)}d\boldsymbol{\pi} + \int q\left(\mathbf{A}\right)\ln\frac{p\left(\mathbf{A}\right)}{q\left(\mathbf{A}\right)}d\mathbf{A}\\
&\quad + \int q\left(\mathbf{Y}\right)\ln\frac{p\left(\mathbf{Y}\right)}{q\left(\mathbf{Y}\right)}d\mathbf{Y} + \int q\left(\beta\right)\ln\frac{p\left(\beta\right)}{q\left(\beta\right)}d\beta\\
&\quad + \ln\tilde{\mathscr{Z}}\left(\mathbf{X}\right) && (41)
\end{aligned}$$

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.

5 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 report, 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.

Future research will be devoted to test the model in detail, using both artificial and real datasets of various characteristics.

References

[1] Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The Generative Topographic Mapping. Neural Comput. **10**(1) (1998) 215–234

[2] Bishop, C.M., Hinton, G., Strachan, I.: GTM Through Time. In: IEE Fifth Int. Conf. on Artif. Neural Net., Cambridge, U.K. (1997) 111–116

[3] Olier, I., Vellido, A.: Capturing the dynamics of multivariate time series through visualization using Generative Topographic Mapping Through Time. In: Proceedings of IEEE ICEIS 2006, Islamabad, Pakistan. (2006) 492–497

[4] Olier, I., Vellido, A.: 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. Volume 4224. (2006) 40–47

[5] Bishop, C.M., Svensén, M., Williams, C.K.I.: Developments of the Generative Topographic Mapping. Neurocomputing **21**(1–3) (1998) 203–224

[6] 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

[7] Beal, M.: Variational algorithms for approximate Bayesian inference. PhD thesis, The Gatsby Computational Neuroscience Unit, Univ. College London (2003)

[8] Jakkola, T., Jordan, M.I.: Bayesian parameter estimation via variational methods. Stat. Comput. **10** (2000) 25–33

[9] Olier, I., Vellido, A.: Variational GTM. In: The 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07). Lect. Notes Comput. Sc. (2007)

[10] Rabiner, L.: A tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proc. of the IEEE. Volume 77. (1989) 257–285

[11] Baum, L., Egon, J.: An inequality with applications to statistical estimation for probabilistic functions for a Markov process and to a model for ecology. B. Am. Meteorol. Soc. **73** (1967) 360–363

[12] Abrahamsen, P.: A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center, Oslo, Norway (1997)

[13] Kabán, A., Girolami, M.: A dynamic probabilistic model to visualise topic evolution in text streams. J. Intell. Inf. Syst. **18**(2–3) (2002) 107–125

[14] Stolcke, A., Omohundro, S.: Hidden Markov model induction by Bayesian model merging. In Hanson, S.J., Cowan, J.D., Giles, C.L., eds.: Advances in Neural Information Processing Systems. Volume 5., San Francisco, CA. Morgan Kauffmann (1993) 11–18

[15] MacKay, D.J.C.: Ensemble learning for Hidden Markov Models. Technical report, Cavendish Laboratory, University of Cambridge (1997) Unpublished Manuscript.

A Computation of the Variational Parameters

Taking into account the approximation $q\left(\boldsymbol{\Theta},\mathbf{Z}\right)=q\left(\boldsymbol{\Theta}\right)q\left(\mathbf{Z}\right)$ and the approximation of the variational distributions over the parameters in Eq. 28, the lower bound can be expressed as:

$$\begin{aligned} F\left(q\left(\boldsymbol{\pi}\right),q\left(\mathbf{A}\right),q\left(\mathbf{Y}\right),q\left(\beta\right),q\left(\mathbf{Z}\right)\right) &= \int q\left(\boldsymbol{\pi}\right)\int q\left(\mathbf{A}\right)\int q\left(\mathbf{Y}\right)\int q\left(\beta\right)\left[\ln\frac{p\left(\boldsymbol{\pi}\right)p\left(\mathbf{A}\right)p\left(\mathbf{Y}\right)p\left(\beta\right)}{q\left(\boldsymbol{\pi}\right)q\left(\mathbf{A}\right)q\left(\mathbf{Y}\right)q\left(\beta\right)}\right.\\ &\left.+\sum_{\text{all path}}q\left(\mathbf{Z}\right)\ln\frac{p\left(\mathbf{X},\mathbf{Z}|\boldsymbol{\pi},\mathbf{A},\mathbf{Y},\beta\right)}{q\left(\mathbf{Z}\right)}\right]d\beta d\mathbf{Y}d\mathbf{A}d\boldsymbol{\pi} \qquad (42)\end{aligned}$$

In the following subsections we will proceed to obtain the expressions of the variational hyperparameters by taking functional derivates over F with respect to $q\left(\boldsymbol{\pi}\right)$, $q\left(\mathbf{A}\right)$, $q\left(\mathbf{Y}\right)$, and $q\left(\beta\right)$ (Eq. 23).

A.1 Derivation of $\tilde{\boldsymbol{\nu}}$

The functional derivative of F with respect to $q\left(\boldsymbol{\pi}\right)$ is given by:

$$\frac{\partial F}{\partial q\left(\boldsymbol{\pi}\right)}=\ln p\left(\boldsymbol{\pi}\right)+\sum_{\text{all path}}q\left(\mathbf{Z}\right)\ln p\left(\mathbf{X},\mathbf{Z}|\boldsymbol{\pi},\mathbf{A},\mathbf{Y},\beta\right)-\ln q\left(\boldsymbol{\pi}\right)+c \qquad (43)$$

where c groups all constant expressions with respect to $q\left(\boldsymbol{\pi}\right)$. If this expression is equated to zero, it yields:

$$\ln q\left(\boldsymbol{\pi}\right) = \ln p\left(\boldsymbol{\pi}\right) + \left\langle \ln p\left(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta\right)\right\rangle_{q(\mathbf{Z})} + c \tag{44}$$

Now, taking into account Eqs. 2 and 3, the expression $\left\langle \ln p\left(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta\right)\right\rangle_{q(\mathbf{Z})}$ can be simplified to $\left\langle \ln p\left(z_1 | \boldsymbol{\pi}\right)\right\rangle_{q(z_1)}$. Given that $p\left(\boldsymbol{\pi}\right)$ and $q\left(\boldsymbol{\pi}\right)$ are Dirichlet distributions, then Eq. 44 is solved as follows:

$$\sum_{j=1}^{K} \left(\tilde{\nu_j} - 1\right) \ln \pi_j = \sum_{j=1}^{K} \left(\nu_j - 1\right) \ln \pi_j + \left\langle \ln p\left(z_1 | \boldsymbol{\pi}\right)\right\rangle_{q(z_1)} + c \tag{45}$$

Each hyperparameter $\tilde{\nu_j}$ is obtained by matching the terms dependent on $\ln \pi_j$ as follows:

$$\tilde{\nu_j} = \nu_j + \left\langle \mathbf{z}_{j,1} \right\rangle \tag{46}$$

A.2 Derivation of $\tilde{\boldsymbol{\lambda}}$

In this case, the functional derivative of F is taken with respect to $q\left(\mathbf{A}\right)$, with the aim to obtain the variational hyperparameter $\tilde{\boldsymbol{\lambda}}$ as follows:

$$\frac{\partial F}{\partial q\left(\mathbf{A}\right)} = \ln p\left(\mathbf{A}\right) + \sum_{\text{all path}} q\left(\mathbf{Z}\right) \ln p\left(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta\right) - \ln q\left(\mathbf{A}\right) + c \tag{47}$$

Following a similar procedure to the one presented in the previous subsection, we obtain:

$$\sum_{i=1}^{K} \sum_{j=1}^{K} \left(\tilde{\lambda_{i,j}} - 1\right) \ln a_{ij} = \sum_{i=1}^{K} \sum_{j=1}^{K} \left(\lambda_{i,j} - 1\right) \ln a_{ij} + \left\langle \ln p\left(\mathbf{Z}_2^N | z_1, \mathbf{A}\right)\right\rangle_{q(\mathbf{Z})} + c \tag{48}$$

Then, for each of the variational hyperparameters $\tilde{\lambda}_{i,j}$:

$$\tilde{\lambda}_{i,j} = \lambda_{i,j} + \sum_{n=2}^{N} \left\langle \mathbf{z}_{i,n-1} \mathbf{z}_{j,n} \right\rangle \tag{49}$$

A.3 Derivation of $\tilde{\Sigma}$ and $\tilde{\mathrm{m}}$

The functional derivate of F with respect to $q\left(\mathbf{Y}\right)$ is given by:

$$\frac{\partial F}{\partial q\left(\mathbf{Y}\right)} = \ln p\left(\mathbf{Y}\right) + \int q\left(\beta\right) \sum_{\text{all path}} q\left(\mathbf{Z}\right) \ln p\left(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta\right) d\beta - \ln q\left(\mathbf{Y}\right) + c \tag{50}$$

Now, using Eqs. 6, 11 and 31, Eq. 50 is solved first for $\tilde{\Sigma}$ and then for $\tilde{\mathbf{m}}_{(d)}$ as follows:

$$\begin{aligned} \sum_{d=1}^{D} \left(\mathbf{y}_{(d)} - \tilde{\mathbf{m}}_{(d)}\right)^T \tilde{\boldsymbol{\Sigma}}^{-1} \left(\mathbf{y}_{(d)} - \tilde{\mathbf{m}}_{(d)}\right) \quad = \quad & \left\langle \beta \right\rangle \sum_{n=1}^{N} \sum_{j=1}^{K} \left\langle \mathbf{z}_{j,n} \right\rangle \left\| \mathbf{x}_n - \mathbf{y}_j \right\|^2 \\ & + \sum_{d=1}^{D} \mathbf{y}_{(d)}^T \mathbf{C}^{-1} \mathbf{y}_{(d)} + c \end{aligned} \tag{51}$$

This leads to:

$$\tilde{\mathbf{\Sigma}} = \left(\langle \beta \rangle \sum_{n=1}^{N} \mathbf{G}_n + \mathbf{C}^{-1} \right)^{-1} \tag{52}$$

and

$$\tilde{\mathbf{m}}_{(d)} = \langle \beta \rangle \tilde{\mathbf{\Sigma}} \sum_{n=1}^{N} x_{nd} \langle \mathbf{z}_n \rangle \tag{53}$$

A.4 Derivation of $\tilde{d}_\beta$ and $\tilde{s}_\beta$

Once again, Eq. 42 is solved as in previous subsections, but now with respect to $q(\beta)$ as follows:

$$\frac{\partial F}{\partial q(\beta)} = \ln p(\beta) + \int q(\mathbf{Y}) \sum_{\text{all path}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\pi}, \mathbf{A}, \mathbf{Y}, \beta)\, d\mathbf{Y} - \ln q(\beta) + c \tag{54}$$

$$\begin{aligned} \tilde{d}_\beta \ln \tilde{s}_\beta + \left(\tilde{d}_\beta - 1\right) \ln \beta - \tilde{s}_\beta \beta \quad &= \quad \frac{ND}{2} \ln \beta - \frac{\beta}{2} \sum_{n=1}^{N} \sum_{j=1}^{K} \langle \mathbf{z}_{j,n} \rangle \left\langle \|\mathbf{x}_n - \mathbf{y}_j\|^2 \right\rangle \\ &\quad + d_\beta \ln s_\beta + (d_\beta - 1) \ln \beta - s_\beta \beta + c \end{aligned} \tag{55}$$

Equating the factors of $\ln \beta$, the variational parameter $\tilde{d}_\beta$ is obtained as:

$$\tilde{d}_\beta = \frac{ND}{2} + d_\beta \tag{56}$$

Similarly, equating with respect to β the variational parameter $\tilde{s}_\beta$, is obtained:

$$\tilde{s}_\beta = \frac{1}{2} \sum_{n=1}^{N} \sum_{j=1}^{K} \langle \mathbf{z}_{j,n} \rangle \left\langle \|\mathbf{x}_n - \mathbf{y}_j\|^2 \right\rangle \tag{57}$$