Missing data imputation through Generative Topographic Mapping as a mixture of *t*-distributions: Theoretical developments

Alfredo Vellido

Department of Computing Languages and Systems (LSI). Polytechnic University of Catalonia (UPC). Barcelona, Spain.

Abstract

The Generative Topographic Mapping (GTM) was originally conceived as a probabilistic alternative to the well-known, neural network-inspired, Self-Organizing Map (SOM). The GTM can also be interpreted as a constrained mixture of distributions model. In recent years, much attention has been directed towards Student *t*-distributions as an alternative to Gaussians in mixture models due to their robustness towards outliers. In this report, the GTM is redefined as a constrained mixture of *t*-distributions: the *t*-GTM, and the Expectation-Maximization algorithm that is used to fit the model to the data is modified to provide missing data imputation.

Keywords: Missing data; Outliers; Generative topographic mapping; Student multivariate t-distributions; Robust imputation; Data visualization.

1. Introduction

Finite mixture models have settled in recent years as a standard for statistical modelling (McLachlan & Peel, 2000b). Their strength and flexibility has been attributed to the fact that they "offer natural models for unobserved population heterogeneity" (Böhning & Seidel, 2003). As such, they are being used in classical data analysis problems such as clustering, regression and probability distribution modelling. Gaussian mixture models have received especial attention for their computational convenience (McLachlan & Peel, 2000a) to deal with multivariate continuous data. The usefulness of these models is reinforced by the wide spectrum of their applications, from medicine (Yau, Lee, & Ng, 2003) to ecology (Ter Braak, Hoijtink, Akkermans, & Verdonschot, 2003) and marketing (Wedel & Kamakura, 2000) to name just a few. For more general reviews see, for instance, (Böhning, 1999; McLachlan & Peel, 2000b).

This report focuses on the Generative Topographic Mapping model (GTM: Bishop, Svensén, & Williams, 1998), conceived as a probabilistic alternative to the neural network-inspired Self-Organizing Map (SOM: Kohonen, 2000). The GTM can also be interpreted as a constrained mixture of distributions. This definition as a constrained model makes it less flexible than general mixtures, but this renounce to full flexibility is compensated by its multivariate data visualization capabilities. Being a non-linear latent variable model, it generates a description of the multivariate data in the form of a low-dimensional manifold embedded in data space, which allows for data visualizations comparable to those of the SOM, which have been widely illustrated (Vesanto, 1999). The GTM, unlike standard Gaussian Mixture Models, is computationally undemanding and its probabilistic setting enables the definition of principled model extensions for, amongst others, time series data (Bishop, Hinton, & Strachan, 1997), hierarchical structures (Tiňo & Nabney, 2002), incomplete data (Carreira-Perpiñan, 2000; Sun, Tiňo, & Nabney, 2001), regularized models (Bishop, Svensén, & Williams, 1998b; Vellido, El-Deredy, & Lisboa, 2003), and discrete data (Bishop, Svensén, & Williams, 1998b; Girolami, 2002).

The GTM was originally defined as a constrained mixture of Gaussian distributions. It is well known (Peel & McLachlan, 2000; Shoham, 2002) that Gaussian mixture models lack robustness in the presence of outlier observations in the data sample, which is a rather common feature on real-world applications (Last & Kandel, 2001) and one that has attracted considerable attention in recent literature (See, for instance, Bashir & Carter, 2005; Castejón Limas, Ordieres Meré, Martínez de Pisón Ascacibar, & Vergara González, 2004; Bullen, Cornford, & Nabney, 2003). Despite the fact that this limitation may also affect the GTM (Tiňo & Nabney, 2002), this model has been used, in its constrained mixture of Gaussians version, for outlier detection (Bullen, Cornford, & Nabney, 2003). An alternative strategy to deal with atypical data using the GTM was proposed by Tiňo & Nabney (2002), relying on the use of the model as the building block of an interactive hierarchical structure.

Starting from the seminal work by (McLachlan & Peel, 1998) and (Peel & McLachlan, 2000), several recent studies have suggested the use of multivariate Student *t*-distributions as a robust alternative to Gaussians for mixture models, as their longer tails prevent outliers from unduly affecting the estimation of the model parameters. Mixtures of *t*-distributions include models defined within a Bayesian approach (Archambeau, Vrins, & Verleysen, 2004; Bishop & Svensén, 2004), model extensions to deal explicitly with incomplete data (Wang, Zhang, Luo, & Wei, 2004), and variants of the Expectation-Maximization algorithm for robust data clustering (Shoham, 2002).

The occurrence of missing data is a pervasive problem in many application areas, and especially acute in domains such as surveys and census (Little & Rubin, 1987; Olynski, Chen, & Harlow, 2003) and, in general, in social and behavioural sciences and fields in which complex measurements are involved such as genetics and bioinformatics (Troyanskaya, Cantor, Sherlock, Brown, Hastie, Tibshirani, Botstein, & Altman, 2001), environmental sciences (Junninen, Niskaa, Tuppurainenc, Ruuskanena, & Kolehmainen, 2004; Vicente, Vellido, Martí, Comas, & Rodriguez-Roda, 2004), or signal processing (Cooke, Green, Josifovski, & Vizinho, 2001). Methods that impute the missing values are therefore of paramount importance for the

successful analysis of such data. Different methods are suitable for different types of data (continuous, discrete, categorical) and for different application fields, with no data imputation method being suitable and successful throughout the universe of data types and application areas. In this report, we provide details on how to integrate missing data imputation as part of the GTM model fitting to data, when GTM is defined as a constrained mixture of *t*-distributions. Data imputation arises naturally as part of the Maximum-Likelihood estimation of the GTM parameters via de Expectation-Maximization (E-M: Dempster, Laird, & Rubin, 1977) algorithm. The resulting GTM model plays a double role: it deals robustly with outliers while simultaneously imputes missing values, allowing the exploration of multivariate data through visualization at a reasonable computational cost.

The rest of the report is structured as follows. First, a brief introduction to the GTM as a constrained mixture of Gaussians is provided, together with details of the Maximum Likelihood estimation of its parameters within the E-M framework. This is followed by the re-definition of GTM as a constrained mixture of Student *t*-distributions (henceforth referred to as *t*-GTM). Finally, we describe the way missing data imputation can be naturally handled as part of the E-M algorithm used to determine the *t*-GTM adaptive parameters.

2. The standard Generative Topographic Mapping

The Generative Topographic Mapping (GTM: Bishop, Svensén, & Williams, 1998a), originally formulated as a statistically principled alternative to Self-Organizing Maps (SOM: Kohonen, 2000), is a non-linear latent variable model that defines a mapping from a low dimensional latent space onto the multidimensional space where the available data reside. The mapping is carried through by a set of basis functions generating a (mixture) density distribution. The functional form of this mapping is defined as a generalized linear regression model:

$$\mathbf{y} = \boldsymbol{\Phi}(\mathbf{u})\mathbf{W}\,, \tag{1}$$

where $\boldsymbol{\Phi}$ is a set of M basis functions $\boldsymbol{\Phi}(\mathbf{u}) = (\phi_1(\mathbf{u}),\ldots,\phi_M(\mathbf{u}))$ that can take diverse forms, depending on the data requirements (e.g., Gaussians for continuous data, Bernouilli distributions for binary data, or Multinomials for categorical data). These basis functions were originally defined (Svensén, 1998) as spherically symmetric Gaussians $\phi_m(\mathbf{u}) = exp\left\{-\frac{\|\mathbf{u}-\mu_m\|^2}{2\sigma^2}\right\}$ to deal with continuous data, with μ_m the centres of the basis functions and σ their common width; $\mathbf{W}$ is a matrix of adaptive weights w_{md} that defines the mapping, and $\mathbf{u}$ is a point in latent space. One of the main strengths of the model resides on its data exploration capabilities through visualization. In order to provide an alternative to the visualization space defined by the characteristic SOM lattice, and also to achieve computational tractability, the latent space of the GTM is discretized as a regular grid of K latent points $\mathbf{u}_k$ defined by the probability

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

where δ is the Kronecker's delta. The probability of a data point $\mathbf{x}$, given the latent space points $\mathbf{u}_k$ and the adaptive parameters of the model, which are the matrix $\mathbf{W}$ and the inverse variance of the Gaussians β, is:

$$P(\mathbf{x}|\mathbf{u},\mathbf{W},\beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} exp\left\{-\beta/2\,\|\mathbf{y}-\mathbf{x}\|^2\right\}. \tag{3}$$

Integrating the latent variables out, and using Eq. (2), we obtain

$$P(\mathbf{x}|\mathbf{W},\beta) = \int P(\mathbf{x}|\mathbf{u},\mathbf{W},\beta)P(\mathbf{u})d\mathbf{u} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\beta}{2\pi}\right)^{D/2} exp\left\{-\beta/2\,\|\mathbf{y}_k-\mathbf{x}\|^2\right\}. \tag{4}$$

According to this general description, the GTM is a constrained mixture of Gaussians in the sense that all the components of the mixture (where each latent point corresponds to a component) are equally weighted by the term $1/K$; all components share a common variance β^{-1} (therefore $\boldsymbol{\Sigma} = \beta^{-1}\mathbf{I}$); and the centres of the Gaussian components $\mathbf{y}_k = \boldsymbol{\Phi}(\mathbf{u}_k)\mathbf{W}$ do not

move independently from each other, as they are limited by the mapping definition to lie in a low dimensional manifold embedded in the D-dimensional space.

The complete log-likelihood can now be defined as

$$L_c(\mathbf{W},\beta|\mathbf{X}) = \sum_{n=1}^{N} log\left\{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\beta}{2\pi}\right)^{D/2} exp\left\{-\beta/2\,\|\mathbf{y}_k - \mathbf{x}_n\|^2\right\}\right\} \tag{5}$$

and the E-M algorithm can be used to obtain the Maximum Likelihood estimates of the adaptive parameters $\mathbf{W}$ and β. Let us first define, in the usual way, the matrix $\mathbf{Z}$, whose indicators z_{kn} describe our lack of knowledge of which latent point $\mathbf{u}_k$ is responsible for the generation of data point $\mathbf{x}_n$. With this, the complete log-likelihood in Eq. (5) can be re-defined as

$$L_c(\mathbf{W},\beta|\mathbf{X},\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\, log\left[\left(\frac{\beta}{2\pi}\right)^{D/2} exp\left\{-\beta/2\,\|\mathbf{y}_k - \mathbf{x}_n\|^2\right\}\right]. \tag{6}$$

The expected value of z_{kn} can be obtained in the E-step of the algorithm using Bayes' formula and Eq. (4):

$$\hat{z}_{kn} = P(k|\mathbf{x}_n,\mathbf{W},\beta) = \frac{exp\left\{-\frac{\beta}{2}\|\mathbf{y}_k - \mathbf{x}_n\|^2\right\}}{\sum_{k'=1}^{K} exp\left\{-\frac{\beta}{2}\|\mathbf{y}_k - \mathbf{x}_n\|^2\right\}}\,. \tag{7}$$

Let us now rewrite Eq. (1) for each data dimension d as $\mathbf{y}_d = \sum_{m=1}^{M}\phi_m(\mathbf{u})w_{md}$. In the M-step, by setting the derivative of L_c from Eq. (6) with respect to w_{md} to zero, and using Eq. (7),

$$\frac{\partial L_c}{\partial w_{md}} = \sum_{n=1}^{N}\sum_{k=1}^{K}\hat{z}_{kn}\left(\sum_{m'=1}^{M}\phi_{m'}(\mathbf{u}_k)w_{m'd} - x_{nd}\right)\phi_m(\mathbf{u}_k) = 0\,, \tag{8}$$

we obtain $\mathbf{W}^{new}$ as the solution of the following system of equations in matricial form

$$\mathbf{\Phi}^T\mathbf{G}\mathbf{\Phi}\mathbf{W}^{new} - \mathbf{\Phi}^T\hat{\mathbf{Z}}\mathbf{X} = 0, \tag{9}$$

where $\mathbf{\Phi}$ is a $K \times M$ matrix with elements $\phi_{km} = \phi_m(\mathbf{u}_k)$; $\hat{\mathbf{Z}}$ is a matrix with elements $\hat{z}_{kn}$ that in the GTM literature is know as responsibility matrix; and, finally, $\mathbf{G}$ is a square matrix with elements $g_{kk'} = \begin{cases} \sum_{n=1}^{N} \hat{z}_{kn}, k = k' \\ 0 \quad k \neq k' \end{cases}$.

Maximazing L_c now with respect to β by setting the corresponding derivative to zero:

$$\frac{\partial L_c}{\partial \beta} = \frac{\partial \left[\sum_{n=1}^{N} \sum_{k=1}^{K} \hat{z}_{kn} \left(D/2 \, log(\beta/2) - \beta/2 \left\| \mathbf{y}_k - \mathbf{x}_n \right\|^2 \right) \right]}{\partial \beta} =$$

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \hat{z}_{kn} \left(\frac{D}{\beta} - \left\| \mathbf{y}_k - \mathbf{x}_n \right\|^2 \right) = 0 \ , \tag{10}$$

we obtain the update expression for the remaining adaptive parameter, the inverse variance β:

$$\left(\beta^{new} \right)^{-1} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{z}_{kn} \left\| \mathbf{y}_k - \mathbf{x}_n \right\|^2 \tag{11}$$

The GTM usually converges within a short number of iterations of the E-M algorithm.

3. GTM as a constrained mixture of Student *t*-distributions: The *t*-GTM

The definition of the GTM as a constrained mixture of Gaussians limits its capability of handling outliers in a data sample consisting of continuous, real-valued variables: The presence of outliers is likely to negatively bias the estimation of parameters $\mathbf{W}$ and β, and it is also likely to result in extreme estimates of the posterior probabilities of component membership (Peel & McLachlan, 2000). Here, the GTM is redefined as a constrained mixture of Student *t*-distributions, the *t*-GTM, aiming to increase the robustness of the model towards outliers. The *t*-GTM is a constrained mixture for the same reasons described in the previous section.

The mapping described by the generalized linear regression model in Eq. (1) remains, and the basis functions $\mathbf{\Phi}$ are now Student *t*-distributions. Assuming again a single common inverse

variance β ($\boldsymbol{\Sigma} = \beta^{-1}\mathbf{I}$) and equal weightings $1/K$ for all components, the data distribution is defined as:

$$P(\mathbf{x}|\mathbf{u},\mathbf{W},\beta,\nu) = \frac{\Gamma\left(\nu/2 + D/2\right)\beta^{D/2}}{\Gamma\left(\nu/2\right)(\nu\pi)^{D/2}}\left(1 + \beta/\nu \left\|\mathbf{y} - \mathbf{x}\right\|^2\right)^{-\frac{\nu+D}{2}}, \tag{12}$$

where $\boldsymbol{\Gamma}(\cdot)$ is the gamma function and the parameter $\nu = (\nu_1,\ldots,\nu_K)^T$ represents the degrees of freedom for each component k of the mixture, so that it can be viewed as a tuner that adapts the level of robustness (divergence from normality) for each component. A multivariate t-distribution converges to a multivariate normal one when $\nu \to \infty$.

Integrating the latent variables out, and using the discretized latent space prior described by Eq. (2):

$$P(\mathbf{x}|\mathbf{W},\beta,\nu) = \int P(\mathbf{x}|\mathbf{u},\mathbf{W},\beta,\nu)P(\mathbf{u})d\mathbf{u} = \frac{1}{K}\sum_{k=1}^{K}\frac{\Gamma\left(\nu_k/2 + D/2\right)\beta^{D/2}}{\Gamma\left(\nu_k/2\right)(\nu_k\pi)^{D/2}}\left(1 + \beta/\nu_k \left\|\mathbf{y}_k - \mathbf{x}\right\|^2\right)^{-\frac{\nu_k+D}{2}} \tag{13}$$

With this, the complete log-likelihood is expressed as:

$$L_c(\mathbf{W},\beta,\nu|\mathbf{X}) = \sum_{n=1}^{N} log\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{\Gamma\left(\nu_k/2 + D/2\right)\beta^{D/2}}{\Gamma\left(\nu_k/2\right)(\nu_k\pi)^{D/2}}\left(1 + \beta/\nu_k \left\|\mathbf{y}_k - \mathbf{x}_n\right\|^2\right)^{-\frac{\nu_k+D}{2}}\right\}. \tag{14}$$

Again, the use of the E-M algorithm for the estimation of parameters $\mathbf{W}$, β and possibly ν, requires re-writing the complete log-likelihood as

$$L_c(\mathbf{W},\beta,\nu|\mathbf{X},\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\, log\left\{\frac{\Gamma\left(\nu_k/2 + D/2\right)\beta^{D/2}}{\Gamma\left(\nu_k/2\right)(\nu_k\pi)^{D/2}}\left(1 + \beta/\nu_k \left\|\mathbf{y}_k - \mathbf{x}_n\right\|^2\right)^{-\frac{\nu_k+D}{2}}\right\}, \tag{15}$$

where indicator variables $\mathbf{Z}$ have once more been introduced. In the E-step, the *responsibilities* $\hat{z}_{kn}$ now follow the expression:

$$\hat{z}_{kn} = P(k|\mathbf{x}_n, \mathbf{W}, \beta, \nu_k) = \frac{C_k\left(1 + \beta/\nu_k \left\|\mathbf{y}_k - \mathbf{x}_n\right\|^2\right)^{-\frac{\nu_k + D}{2}}}{\sum_{k'=1}^{K} C_{k'}\left(1 + \beta/\nu_{k'} \left\|\mathbf{y}_{k'} - \mathbf{x}_n\right\|^2\right)^{-\frac{\nu_{k'} + D}{2}}} \ , \tag{16}$$

where

$$C_k = \Gamma\left(\nu_k/2 + D/2\right)\beta^{D/2}\left[\Gamma\left(\nu_k/2\right)(\nu_k\pi)^{D/2}\right]^{-1} . \tag{17}$$

Update expressions for the adaptive parameters are calculated in the M-step of the algorithm. Maximizing with respect to w_{md}, by setting the derivatives of Eq. (14) with respect to w_{md} to zero, we obtain:

$$\frac{\partial L_c}{\partial w_{md}} = \sum_{n=1}^{N} \frac{\partial \, log\left\{1/K \sum_{k=1}^{K} C_k\left(1 + \beta/\nu_k \left\|\mathbf{y}_k - \mathbf{x}_n\right\|^2\right)^{-\frac{\nu_k + D}{2}}\right\}}{\partial w_{md}} =$$

$$\sum_{n=1}^{N} \frac{1/K \sum_{k=1}^{K} C_k\left(-\frac{\nu_k + D}{2}\right)\left(1 + \beta/\nu_k \left\|\mathbf{y}_k - \mathbf{x}_n\right\|^2\right)^{-\frac{\nu_k + D + 2}{2}}\left(2\beta/\nu_k\right)(\phi(\mathbf{u}_k)\mathbf{w}_d - x_{nd})(-\phi_m(\mathbf{u}_k))}{1/K \sum_{k'}^{K} C_{k'}\left(1 + \beta/\nu_{k'} \left\|\mathbf{y}_{k'} - \mathbf{x}_n\right\|^2\right)^{-\frac{\nu_{k'} + D}{2}}} =$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{(\nu_k + D)\beta}{\nu_k} \hat{z}_{kn} \frac{\left(\sum_{m'=1}^{M} \phi_{m'}(\mathbf{u}_k) w_{m'd} - x_{nd}\right)\phi_m(\mathbf{u}_k)}{1 + \beta/\nu_k \left\|\mathbf{x}_n - \mathbf{y}_k\right\|^2} = 0 \ . \tag{18}$$

This leads to an equation, in matrix form, for the update of $\mathbf{W}$ that is similar to Eq. (9):

$$\mathbf{\Phi}^T \mathbf{G}^* \mathbf{\Phi} \mathbf{W}^{new} - \mathbf{\Phi}^T \hat{\mathbf{Z}}^* \mathbf{X} = 0 \ , \tag{19}$$

where

$$\hat{z}^*_{kn} = \frac{\nu_k + D}{\nu_k + \beta\left\|\mathbf{x}_n - \mathbf{y}_k^{old}\right\|^2} \hat{z}_{kn} \tag{20}$$

and $\hat{z}_{kn}$ is defined by Eq. (16). Matrix $\mathbf{G}^*$ has values $g^*_{kk'} = \begin{cases} \sum_{n=1}^{N} \hat{z}^*_{kn}, k = k' \\ 0 \quad k \neq k' \end{cases}$. The new terms in Eq. (19) do not add any extra computational burden with respect to Eq. (16), as they have already been calculated in previous steps of the algorithm.

The maximization with respect to parameter β leads to a special case of the update formula for general mixtures of *t*-distributions:

$$\left(\beta^{new}\right)^{-1} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{z}_{kn} (\nu_k + D) \left(\nu_k + \beta^{old} \left\| \mathbf{y}_k^{new} - \mathbf{x}_n \right\|^2 \right)^{-1} \left\| \mathbf{y}_k^{new} - \mathbf{x}_n \right\|^2, \tag{21}$$

where $\mathbf{y}_k^{new} = \boldsymbol{\Phi}(\mathbf{u}_k)\mathbf{W}^{new}$. For the standard Gaussian GTM (Svensén, 1998), Eq. (11) can be interpreted as the off-manifold variance of the model being updated to the averaged distance between data points and latent points (or mixture components), where this distance is weighted by the posterior probabilities $\hat{z}_{kn}$. Notice that Eq. (21) implies the existence of a further weighting term for the *t*-GTM, which, according to (Peel & McLachlan, 2000), will be small for data outliers. As a result, the impact of outliers on the estimation of the variance parameter will be effectively minimized. This leaves us with parameter ν, for which optimization is less straightforward. Different approaches might be considered: an approximation for general mixture models was proposed by (Shoham, 2002) for a common ν for all mixture components (i.e. $\forall k, \nu_k = \nu$). Alternatively, ν might be kept fixed, running experiments for a range of its possible values.

4. Missing data imputation through *t*-GTM

It has been shown how the GTM model, defined as a constrained mixture of either Gaussian or Student *t*-distributions, can be fitted to the data using the E-M algorithm. As stated in (Ghahramani & Jordan, 1994), "the problem of estimating mixture densities can itself be viewed as a missing data problem". In the previous sections, the matrix $\mathbf{Z}$ of indicators -describing our lack of knowledge of which latent point $\mathbf{u}_k$ is responsible for the generation of data point $\mathbf{x}_n$-

was treated as missing data. In this section, we see how the missing data themselves can be explicitly dealt with and imputed as part of the own E-M procedure for the *t*-GTM.

For that, we follow (Sun, Tiňo, & Nabney, 2001) and consider two separate submatrices: $\mathbf{X}^o$, consisting of the observed data represented by superscript *o*, and $\mathbf{X}^m$, consisting of the missing data represented by superscript *m*. No constrain has been imposed on the pattern followed by the missing values, although either a Missing Completely At Random (MCAR) or a Missing At Random (MAR) situation is assumed.

The Expectation step of the E-M algorithm includes the calculation of the expected complete log-likelihood. The definition of submatrices $\mathbf{X}^o$ and $\mathbf{X}^m$ entails a modification of Eq. (15), which now becomes:

$$L_c\left(\mathbf{W},\beta,\nu\middle|\mathbf{X}^o,\mathbf{X}^m,\mathbf{Z}\right)=\sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\, log\left\{C_k\left[1+\beta/\nu_k\left(\left\|\mathbf{y}_k^o-\mathbf{x}_n^o\right\|^2+\left\|\mathbf{y}_k^m-\mathbf{x}_n^m\right\|^2\right)\right]^{-\frac{\nu_k+D}{2}}\right\}, \quad (22)$$

given that we are defining a common variance for all mixture components and, therefore, using an isotropic covariance matrix $\boldsymbol{\Sigma}=\beta^{-1}\mathbf{I}$ that excludes values involving both observed and missing data. The sufficient statistics that must be calculated prior to the M-step are: the expected values of the unknown indicator variables $E\left[z_{kn}\middle|\mathbf{x}_n^o,\mathbf{W},\beta,\nu_k\right]$, which are precisely the posterior probabilities in Eq. (16), calculated using only the observed data:

$$\hat{z}_{kn}=P\left(k\middle|\mathbf{x}_n,\mathbf{W},\beta,\nu_k\right)=\frac{C_k\left(1+\beta/\nu_k\left\|\mathbf{y}_k^o-\mathbf{x}_n^o\right\|^2\right)^{-\frac{\nu_k+D}{2}}}{\sum_{k'=1}^{K}C_{k'}\left(1+\beta/\nu_{k'}\left\|\mathbf{y}_{k'}^o-\mathbf{x}_n^o\right\|^2\right)^{-\frac{\nu_{k'}+D}{2}}}, \quad (23)$$

and the interactions between the indicator variables and the first and second moments of $\mathbf{x}_n^m$: $E\left[z_{kn}\mathbf{x}_n^m\middle|\mathbf{x}_n^o,\mathbf{W},\beta,\nu_k\right]$ and $E\left[z_{kn}\mathbf{x}_n^m\mathbf{x}_n^{mT}\middle|\mathbf{x}_n^o,\mathbf{W},\beta,\nu_k\right]$. We first define (Ghahramani & Jordan, 1994; Sun, Tiňo, & Nabney, 2001) the expectation

$$E\left[\mathbf{x}_n^m \middle| z_{kn}=1, \mathbf{x}_n^o, \mathbf{W}, \boldsymbol{\beta}, \nu_k\right] = \hat{\mathbf{x}}_{kn}^m = \left(\mathbf{y}_k^m\right)^{old}, \tag{24}$$

where *old* stands for calculations obtained in the previous algorithm iteration. This way, we obtain

$$E\left[z_{kn}\mathbf{x}_n^m \middle| \mathbf{x}_n^o, \mathbf{W}, \boldsymbol{\beta}, \nu_k\right] = \hat{z}_{kn}\hat{\mathbf{x}}_{kn}^m \tag{25}$$

and

$$E\left[z_{kn}\mathbf{x}_n^m \mathbf{x}_n^{mT} \middle| \mathbf{x}_n^o, \mathbf{W}, \boldsymbol{\beta}, \nu_k\right] = \hat{z}_{kn}\left(\left(\boldsymbol{\beta}^{-1}\right)^{old} + \hat{\mathbf{x}}_{kn}^{mT}\hat{\mathbf{x}}_{kn}^m\right), \tag{26}$$

where, for both Eq. (25) and Eq. (26), $\hat{z}_{kn}$ is given by Eq. (23). The missing data imputation is now straightforward: it is performed according to:

$$E\left[\mathbf{x}_n^m \middle| \mathbf{x}_n^o, \mathbf{W}, \boldsymbol{\beta}, \nu_k\right] = \sum_{k=1}^{K}\hat{z}_{kn}E\left[\mathbf{x}_n^m \middle| z_{kn}=1, \mathbf{x}_n^o, \mathbf{W}, \boldsymbol{\beta}, \nu_k\right] = \sum_{k=1}^{K}\hat{z}_{kn}\left(\mathbf{y}_k^m\right)^{old} \tag{27}$$

This imputation procedure completes the data and allows their full visualization on the low-dimensional latent space.

In the Maximization step of the E-M algorithm, we use those now reconstructed data consisting on the combination of the observed and imputed subsets, which we call $\mathbf{X}^{rec}$ (where *rec* stands for reconstructed), to obtain $\mathbf{W}^{new}$ as the solution of a modified version of Eq. (19):

$$\boldsymbol{\Phi}^T\mathbf{G}^*\boldsymbol{\Phi}\mathbf{W}^{new} - \boldsymbol{\Phi}^T\hat{\mathbf{Z}}^*\mathbf{X}^{rec} = 0\ . \tag{28}$$

Note that the elements $\hat{z}_{kn}^*$ of $\mathbf{Z}^*$, also basis of the calculation of the elements of $\mathbf{G}^*$, are now calculated as

$$\hat{z}_{kn}^* = \frac{\nu_k + D}{\nu_k + \beta\left(\left\|\mathbf{x}_n^o - \mathbf{y}_k^{o,old}\right\|^2 + \left\|\mathbf{x}_n^m - \mathbf{y}_k^{m,old}\right\|^2\right)}\hat{z}_{kn} \tag{29}$$

This matrix of weights $\mathbf{W}^{new}$ can be used to update the generated mixture component centres as $\left(\mathbf{y}_k^m\right)^{new} = \left(\mathbf{W}^{new}\boldsymbol{\Phi}(\mathbf{u}_k)\right)^m$ and $\left(\mathbf{y}_k^o\right)^{new} = \left(\mathbf{W}^{new}\boldsymbol{\Phi}(\mathbf{u}_k)\right)^o$, which, in turn, are used to update the mixture component-common inverse variance:

$$\left(\beta^{new}\right)^{-1} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{k=1}^{K}\hat{z}_{kn}\left(\nu_k + D\right)\left\{\nu_k + \beta^{old}\left(\left\|\left(\mathbf{y}_k^o\right)^{new} - \mathbf{x}_n^o\right\|^2 + E\left[z_{kn}\left\|\left(\mathbf{y}_k^m\right)^{new} - \mathbf{x}_n^m\right\|^2\right]\right)\right\}^{-1}$$

$$\left\{\left\|\left(\mathbf{y}_k^o\right)^{new} - \mathbf{x}_n^o\right\|^2 + E\left[z_{kn}\left\|\left(\mathbf{y}_k^m\right)^{new} - \mathbf{x}_n^m\right\|^2\right]\right\}, \tag{30}$$

where

$$E\left[z_{kn}\left\|\left(\mathbf{y}_k^m\right)^{new} - \mathbf{x}_n^m\right\|^2\right] = \left(\beta^{-1}\right)^{old} + \hat{\mathbf{x}}_{kn}^{m\,T}\hat{\mathbf{x}}_{kn}^m + \left(\mathbf{y}_k^m\right)^{new^T}\left(\mathbf{y}_k^m\right)^{new} - 2\hat{\mathbf{x}}_{kn}^{m\,T}\left(\mathbf{y}_k^m\right)^{new}. \tag{31}$$

This completes the account of modifications of the E-M procedure described in the previous section that are necessary to implement missing data imputation as an integral part of it.

References

Archambeau, C., Vrins, F., & Verleysen, M. (2004). Flexible and robust Bayesian classification by finite mixture models. In M. Verleysen (Ed.), *Proceedings Twelfth European Symposium on Artificial Neural Networks*, (pp. 75-80). Evere, Belgium: D-Side Publications.

Bashir, S., & Carter, E.M. (2005). High breakdown mixture discriminant analysis. *Journal of Multivariate Analysis*, **93**, 102-111.

Bishop, C.M., Hinton G.E., & Strachan, I.G.D. (1997). GTM through time. In *Proceedings of IEE Fifth International Conference on Artificial Neural Networks, Cambridge, U.K.* (pp. 111-116). London: IEE.

Bishop, C.M., & Svensén, M. (2004). Robust Bayesian mixture modelling. In M. Verleysen (Ed.), *Proceedings Twelfth European Symposium on Artificial Neural Networks*, (pp. 69-74). Evere, Belgium: D-Side Publications.

Bishop, C.M., Svensén, M., & Williams, C.K.I. (1998a). GTM: The Generative Topographic Mapping. *Neural Computation*, **10**, 215-234.

Bishop, C.M., Svensén, M., & Williams, C.K.I. (1998b). Developments of the Generative Topographic Mapping. *Neurocomputing*, **21**, 203-224.

Böhning, D. (1999). Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others. London, UK: Chapman & Hall /CRC.

Böhning, D. & Seidel. W. (2003). Recent developments in mixture models. *Computational Statistics and Data Analysis*, **41**, 349-357.

Bullen, R.J., Cornford, D., & Nabney, I.T. (2003). Outlier detection in scatterometer data: neural network approaches. *Neural Networks,* **16**, 419-426.

Carreira-Perpiñan, M.A. (2000). Reconstruction of sequential data with probabilistic models and continuity constraints. In: S.A. Solla, T.K. Leen, & K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems Vol. 12* (pp. 414-420). Cambridge, MA: MIT Press.

Castejón Limas, M., Ordieres Meré, J.B., Martínez de Pisón Ascacibar, F.J., & Vergara González, E.P. (2004). Outlier detection and data cleaning in multivariate non-normal samples: the *PAELLA* algorithm. *Data Mining and Knowledge Discovery*, **9**, 171-187.

Cooke, M.P., Green, P.D. Josifovski, L. & Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, **34**, 267–285.

Dempster, A.P., Laird, M.N., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Ghahramani, Z., & Jordan, M.I. (1994). Learning from incomplete data. Technical Report, AI Laboratory, MIT.

Girolami, M. (2002). Latent variable models for the topographic organisation of discrete and strictly positive data. *Neurocomputing*, **48**, 185-198.

Junninen, H., Niskaa, H., Tuppurainenc, K., Ruuskanena, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, **38**, 2895-2907.

Kohonen, T. (2000). Self-organizing Maps (3rd ed.). Berlin: Springer-Verlag.

Last, M., & Kandel, A. (2001). Automated detection of outliers in real-world data. In *Proceedings of the Second International Conference on Intelligent Technologies* (pp. 292-301). Bangkok, Thailand.

Little, R.J.A., & Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

McLachlan, G.J., & Peel, D. (1998). Robust cluster analysis via mixtures of multivariate *t*-distributions. In A. Amin, D. Dori, P. Pudil, & H. Freeman (Eds.), *Lecture Notes in Computer Science Vol. 1451* (pp. 658–666). Berlin: Springer-Verlag.

McLachlan, G.J., & Peel, D. (2000a). On computational aspects of clustering via mixtures of normal and t-components. In *Proceedings of the American Statistical Association (Bayesian Statistical Science Section)*. Alexandria, Virginia: American Statistical Association.

McLachlan, G.J., & Peel, D. (2000b). Finite Mixture Models. New York: John Wiley & Sons.

Olinsky, A., Chen, S. & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modelling. *European Journal of Operational Research*, **151**, 53–79.

Peel, D., & McLachlan, G.J. (2000). Robust mixture modelling using the t distribution. Statistics and Computing, **10**, 339–348.

Shoham, S. (2002). Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. *Pattern Recognition*, **35**, 1127–1142.

Sun, Y., Tiňo, P., & Nabney, I. (2001). GTM-based data visualization with incomplete data. Technical Report, NCRG, Aston University, UK.

Svensén, M. (1998). GTM: The Generative Topographic Mapping. PhD Thesis, Aston University, Birmingham, U.K.

Ter Braak, C.J.F., Hoijtink, H., Akkermans, W. & Verdonschot, P.F.M. (2003). Bayesian model-based cluster analysis for predicting macrofaunal communities. *Ecological Modelling*, **160**, 235-248.

Tiňo, P., & Nabney, I. (2002). Hierarchical GTM: constructing localized non-linear projection manifolds in a principled way. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 639-656.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520-525.

Vellido, A., El-Deredy, W., & Lisboa, P.J.G. (2003). Selective smoothing of the Generative Topographic Mapping. *IEEE Transactions on Neural Networks*, **14**, 847-852.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, **3**, 111-126.

Vicente, D., Vellido, A., Martí, E., Comas, J., & Rodriguez-Roda, I. (2004). Exploration of the ecological status of mediterranean rivers: Clustering, visualizing and reconstructing streams data using Generative Topographic Mapping. In A. Zanasi, N.F.F. Ebecken, & C.A. Brebbia (Eds.) *WIT Transactions on Information and Communication Technologies, Vol.33* (pp. 121-130). Southampton: WIT Press.

Wang, H.X., Zhang, Q.B., Luo, B., & Wei, S. (2004). Robust mixture modelling using multivariate t-distribution with missing information. *Pattern Recognition Letters*, **25**, 701–710.

Wedel, M., & Kamakura, W.A. (2000). Market Segmentation: Conceptual and Methodological Foundations (2nd ed.). Boston: Kluwer Academic Publishers.

Yau, K.K.W., Lee, A.H., & Ng, A.S.K. (2003). Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics and Data Analysis*, **41**, 359-366.