Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 19 (2006) 1624–1635
Neural Networks
www.elsevier.com/locate/neunet

Missing data imputation through GTM as a mixture of t-distributions

Alfredo Vellido*

Department of Computing Languages and Systems (LSI), Polytechnic University of Catalonia (UPC), C. Jordi Girona, 1-3. 08034, Barcelona, Spain

Received 12 November 2004; accepted 28 November 2005

Abstract

The Generative Topographic Mapping (GTM) was originally conceived as a probabilistic alternative to the well-known, neural network-inspired, Self-Organizing Maps. The GTM can also be interpreted as a constrained mixture of distribution models. 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 paper, 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 carry out missing data imputation. Several experiments show that the t-GTM successfully detects outliers, while minimizing their impact on the estimation of the model parameters. It is also shown that the t-GTM provides an overall more accurate imputation of missing values than the standard Gaussian GTM.

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

1. Introduction

Finite mixture models have become established in recent years as a standard for generic non-linear 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 special attention due to their computational convenience (McLachlan & Peel, 2000a) for dealing 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) and McLachlan and Peel (2000b).

Over recent decades, neural networks have steadily veered away from biologically plausible, mostly deterministic models, based on heuristic methods, towards stochastic models with solid grounds on probability theory (Bishop, 1995; MacKay, 1995). The model on which this paper focuses, Generative Topographic Mapping (GTM: Bishop, Svensén, & Williams, 1998a), is an example of this, as it was originally conceived as a probabilistic alternative to the originally bio-inspired Self-Organizing Maps (SOM: Kohonen, 2001). 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 compromise of 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 is less computationally demanding than standard Gaussian mixture models, 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 et al., 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

* Tel.: +34 93 4137796; fax: +34 93 4137833.
E-mail address: avellido@lsi.upc.edu.

0893-6080/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.11.003

robustness in the presence of outlier observations in the data sample, which is a rather common feature in real-world applications (Last & Kandel, 2001) and one that has attracted considerable attention in recent literature (see, for instance, Bashir and Carter (2005), Bullen, Cornford, and Nabney (2003) and Castejón Limas, Ordieres Meré, Martínez de Pisón Ascacibar, and Vergara González (2004)). Despite the fact that this limitation may also affect the GTM (Tiňo & Nabney, 2002), this model has been used—formulated as a constrained mixture of Gaussians—for outlier detection (Bullen et al., 2003). An alternative strategy for dealing with atypical data using the GTM was proposed by Tiňo and Nabney (2002), relying on the use of the model as the building block of an interactive hierarchical structure.

Starting with the seminal work by McLachlan and Peel (1998), 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. Among them are models defined within a Bayesian approach (Archambeau, Vrins, & Verleysen, 2004; Svensén & Bishop, 2005), extensions to subspace mixture models (de Ridder & Franc, 2003), and variants for dealing explicitly with incomplete data (Wang, Zhang, Luo, & Wei, 2004) and 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; Olinsky, 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 et al., 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 paper, 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 the Expectation–Maximization (EM: Dempster, Laird, & Rubin, 1977) algorithm. The resulting GTM model plays a double role: it deals robustly with outliers while it simultaneously imputes missing values, allowing the exploration of multivariate data through visualization at a reasonable computational cost.

This model is assessed in several experiments, aiming first to ascertain whether it can successfully detect outliers and whether it can minimize their impact on the estimation of the model parameters. Secondly, it aims to test the results yielded by the proposed missing data imputation procedure.

The rest of the paper 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 EM framework. This is followed by the re-definition of GTM as a constrained mixture of Student t-distributions (henceforth referred to as t-GTM). Next, we describe the way missing data imputation can be naturally handled as part of the EM algorithm used to determine the t-GTM adaptive parameters. Results of several experiments, designed to evaluate the robustness of the proposed model and the reliability of the missing data imputation, are then provided and discussed. The paper wraps up with some conclusions and directions for future research.

2. The standard generative topographic mapping

The Generative Topographic Mapping (GTM: Bishop et al., 1998a), originally formulated as a statistically principled alternative to Self-Organizing Maps (SOM: Kohonen, 2001), is a non-linear latent variable model that defines a mapping from a low-dimensional latent space onto the multi-dimensional 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 define 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 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:

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

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

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

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

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

and the EM 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

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

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

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

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),

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

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

$$\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{G}\boldsymbol{\Phi}\mathbf{W}^{\text{new}} - \boldsymbol{\Phi}^{\mathrm{T}}\hat{\mathbf{Z}}X = 0, \quad (9)$$

where $\boldsymbol{\Phi}$ 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 the responsibility matrix; and, finally, $\mathbf{G}$ is a diagonal square matrix with elements $g_{kk'} = \begin{cases}\sum_{n=1}^{N}\hat{z}_{kn}, & k=k'\\ 0 & k\neq k'\end{cases}$.

Maximizing L_c now with respect to β by setting the corresponding derivative to zero,

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

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

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

The GTM usually converges within a short number of iterations of the EM 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 for 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 $\boldsymbol{\Phi}$ are now Student t-distributions. Again assuming 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:

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

where $\Gamma(\cdot)$ is the gamma function and the parameter $\nu = (\nu_1,\ldots,\nu_K)^{\mathrm{T}}$ 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 distribution when $\nu\to\infty$.

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

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

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

Box I.

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

$$P(\mathbf{x}|\mathbf{W}, \beta, \nu) = \int P(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta, \nu)\, P(\mathbf{u})\, \mathrm{d}\mathbf{u}$$

$$= \frac{1}{K} \sum_{k=1}^{K} \frac{\Gamma\left(\frac{\nu_k}{2} + \frac{D}{2}\right) \beta^{D/2}}{\Gamma\left(\frac{\nu_k}{2}\right) (\nu_k \pi)^{D/2}} \times \left(1 + \frac{\beta}{\nu_k} \left\|\mathbf{y}_k - \mathbf{x}\right\|^2\right)^{-\frac{\nu_k+D}{2}}. \quad (13)$$

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

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

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

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

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

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

where

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

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

$$\boldsymbol{\Phi}^{\mathrm{T}} \mathbf{G}^* \boldsymbol{\Phi} \mathbf{W}^{\mathrm{new}} - \boldsymbol{\Phi}^{\mathrm{T}} \hat{\mathbf{Z}}^* \mathbf{X} = 0, \quad (18)$$

where

$$\hat{z}^*_{kn} = \frac{\nu_k + D}{\nu_k + \beta \left\|\mathbf{y}_k^{\mathrm{old}} - \mathbf{x}_n\right\|^2} \hat{z}_{kn} \quad (19)$$

and $\hat{z}_{kn}$ is defined by Eq. (16). Matrix $\mathbf{G}^*$ has values $g^*_{kk'} = \begin{cases} \sum_{n=1}^{N} \hat{z}^*_{kn}, & k = k' \\ 0 & k \neq k' \end{cases}$. The new terms in Box I 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 updated formula for general mixtures of t-distributions:

$$\left(\beta^{\mathrm{new}}\right)^{-1} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{z}_{kn} (\nu_k + D) \times \left(\nu_k + \beta^{\mathrm{old}} \left\|\mathbf{y}_k^{\mathrm{new}} - \mathbf{x}_n\right\|^2\right)^{-1} \left\|\mathbf{y}_k^{\mathrm{new}} - \mathbf{x}_n\right\|^2, \quad (20)$$

where $\mathbf{y}_k^{\mathrm{new}} = \boldsymbol{\Phi}(\mathbf{u}_k)\mathbf{W}^{\mathrm{new}}$. 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 mixture component centres, where this distance is weighted by the posterior probabilities $\hat{z}_{kn}$. Notice that Eq. (20) implies the existence of a further weighting term $(\nu_k + D)(\nu_k + \beta^{\mathrm{old}} \|\mathbf{y}_k^{\mathrm{new}} - \mathbf{x}_n\|^2)^{-1}$ for the t-GTM, which, according to Peel and 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 is due to the overall longer distances between outlier observations and the centres of the mixture components $\mathbf{y}_k^{\text{new}}$, which, in turn, are the result of the same weighting term operating in Eq. (19) and affecting the re-estimation of matrix $\mathbf{W}$. 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 and selecting that which maximized the complete log-likelihood.

3.1. Related work

The use of multivariate Student t-distributions as a robust alternative to Gaussians for finite mixture models was first proposed by McLachlan and Peel (1998). In this seminal work, details of the Maximum Likelihood estimation of the mixture using the EM algorithm (and constrained versions, such as ECM) were provided.

Since then, several recent studies have suggested the use of multivariate Student t-distributions as a robust alternative to Gaussians for mixture models, following different approaches and with diverse goals: The work of Shoham (2002) focused on improved versions of the classic EM algorithm based on deterministic annealing (DAEM: Ueda & Nakano, 1998) to improve algorithm convergence. In Archambeau et al. (2004), the regularized Mahalanobis distance was proposed for finite mixtures of t-distributions to avoid common numerical difficulties encountered with standard EM. A variant of this model that goes beyond EM and is defined within a Bayesian approach was presented by Svensén and Bishop (2005); here, a tractable variational inference algorithm for the model is derived, which also allows for the automatic determination of the appropriate number of mixture components; the model was tested only on univariate data sets. Model extensions to deal explicitly with incomplete data have also been developed (Wang et al., 2004), taking advantage of the inherent capability of the EM algorithm to deal with missing data. This is the same approach as followed in this paper, with the main difference that t-GTM allows for the visualization of the regenerated full set of multivariate data. The work of de Ridder and Franc (2003) is conceptually closer to t-GTM, as they combine the ideas of robust modelling by t-distributions and probabilistic subspace mixture models (of which GTM is a nonlinear example), with a focus on mixtures of Probabilistic Principal Component Analyzers (PPCA: Tipping & Bishop, 1999).

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 EM algorithm. As stated in Ghahramani and 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 EM procedure for the t-GTM.

For that, we follow Sun et al. (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 constraint has been imposed on the pattern followed by the missing values. The expectation step of the EM 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:

$$L_c\left(\mathbf{W}, \beta, \nu | \mathbf{X}^o, \mathbf{X}^m, \mathbf{Z}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn} \times \log\left\{ C_k \left[1 + \frac{\beta}{\nu_k}\left(\left\| \mathbf{y}_k^o - \mathbf{x}_n^o \right\|^2 + \left\| \mathbf{y}_k^m - \mathbf{x}_n^m \right\|^2 \right) \right]^{-\frac{\nu_k + D}{2}} \right\}, \tag{21}$$

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 $E[z_{kn}|\mathbf{x}_n^o, \mathbf{W}, \beta, \nu_k]$, which are precisely the posterior probabilities in Eq. (16), calculated using only the observed data:

$$\hat{z}_{kn} = P\left(k|\mathbf{x}_n, \mathbf{W}, \beta, \nu_k\right) = \frac{C_k \left(1 + \frac{\beta}{\nu_k}\left\| \mathbf{y}_k^o - \mathbf{x}_n^o \right\|^2\right)^{-\frac{\nu_k + D}{2}}}{\sum_{k'=1}^{K} C_{k'} \left(1 + \frac{\beta}{\nu_{k'}}\left\| \mathbf{y}_{k'}^o - \mathbf{x}_n^o \right\|^2\right)^{-\frac{\nu_{k'} + D}{2}}}, \tag{22}$$

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

$$E\left[\mathbf{x}_n^m | z_{kn} = 1, \mathbf{x}_n^o, \mathbf{W}, \beta, \nu_k\right] = \hat{\mathbf{x}}_{kn}^m = \left(\mathbf{y}_k^m\right)^{\text{old}}, \tag{23}$$

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

$$E\left[z_{kn}\mathbf{x}_n^m | \mathbf{x}_n^o, \mathbf{W}, \beta, \nu_k\right] = \hat{z}_{kn}\hat{\mathbf{x}}_{kn}^m \tag{24}$$

and

$$E\left[z_{kn}\mathbf{x}_n^m \mathbf{x}_n^{m\mathrm{T}} | \mathbf{x}_n^o, \mathbf{W}, \beta, \nu_k\right] = \hat{z}_{kn}\left(\left(\beta^{-1}\right)^{\text{old}} + \hat{\mathbf{x}}_{kn}^m \hat{\mathbf{x}}_{kn}^{m\mathrm{T}}\right), \tag{25}$$

where, for both Eqs. (24) and (25), $\hat{z}_{kn}$ is given by Eq. (22). The missing data imputation is now straightforward—it is

performed according to:

$$E\left[\mathbf{x}_n^m|\mathbf{x}_n^o, \mathbf{W}, \beta, \nu_k\right] = \sum_{k=1}^{K} \hat{z}_{kn} E\left[\mathbf{x}_n^m|z_{kn}=1, \mathbf{x}_n^o, \mathbf{W}, \beta, \nu_k\right] = \sum_{k=1}^{K} \hat{z}_{kn} \left(\mathbf{y}_k^m\right)^{\text{old}}. \quad (26)$$

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

In the maximization step of the EM algorithm, we use those now reconstructed data consisting of the combination of the observed and imputed subsets, which we call $\mathbf{X}^{\text{rec}}$ (where rec stands for reconstructed), to obtain $\mathbf{W}^{\text{new}}$ as the solution of a modified version of Eq. (18):

$$\boldsymbol{\Phi}^{\text{T}}\mathbf{G}^*\boldsymbol{\Phi}\mathbf{W}^{\text{new}} - \boldsymbol{\Phi}^{\text{T}}\hat{\mathbf{Z}}^*\mathbf{X}^{\text{rec}} = 0. \quad (27)$$

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

$$\hat{z}_{kn}^* = \frac{\nu_k + D}{\nu_k + \beta\left(\left\|\mathbf{y}_k^{o,\text{old}} - \mathbf{x}_n^o\right\|^2 + \left\|\mathbf{y}_k^{m,\text{old}} - \mathbf{x}_n^m\right\|^2\right)}\hat{z}_{kn}. \quad (28)$$

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

$$\begin{aligned}\left(\beta^{\text{new}}\right)^{-1} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{k=1}^{K}\hat{z}_{kn}(\nu_k + D) &\times \left\{\nu_k + \beta^{\text{old}}\left(\left\|\left(\mathbf{y}_k^o\right)^{\text{new}} - \mathbf{x}_n^o\right\|^2 + E\left[\left\|\left(\mathbf{y}_k^m\right)^{\text{new}} - \mathbf{x}_n^m\right\|^2|z_{kn}=1\right]\right)\right\}^{-1} \\ &\times \left\{\left\|\left(\mathbf{y}_k^o\right)^{\text{new}} - \mathbf{x}_n^o\right\|^2 + E\left[\left\|\left(\mathbf{y}_k^m\right)^{\text{new}} - \mathbf{x}_n^m\right\|^2|z_{kn}=1\right]\right\}, \end{aligned} \quad (29)$$

where

$$E\left[\left\|\left(\mathbf{y}_k^m\right)^{\text{new}} - \mathbf{x}_n^m\right\|^2|z_{kn}=1\right] = \left(\beta^{-1}\right)^{\text{old}} + \hat{\mathbf{x}}_{kn}^{m^{\text{T}}}\hat{\mathbf{x}}_{kn}^m + \left(\mathbf{y}_k^m\right)^{\text{new}^{\text{T}}}\left(\mathbf{y}_k^m\right)^{\text{new}} - 2\hat{\mathbf{x}}_{kn}^{m^{\text{T}}}\left(\mathbf{y}_k^m\right)^{\text{new}}. \quad (30)$$

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

5. Experiments

The experiments, performed on several data sets, were divided according to two main goals. The first was to illustrate the effect of using t-distributions instead of Gaussians in the definition of the GTM model. Therefore, it focused on the model's robustness towards outliers, ignoring, at this stage, the possible incompleteness of the data. In accordance with Peel and McLachlan (2000), we expected to see less extreme estimates of the posterior probability of the mixture components (latent space points of the t-GTM) given the data, expressed by Eq. (16), than those obtained for the GTM described as a constrained mixture of Gaussians, expressed by Eq. (7). For this experiment, and following Wang et al. (2004), we used an augmented version of Fisher's Iris data set, available from the U.C.I. Machine Learning Repository.[1] We also used a data set of single-voxel Magnetic Resonance Spectra (MRS) measured in vivo from several human brain tumours and cystic growths.

The second goal concerned two aspects. It first aimed to assess whether the missing data imputation procedure was robust enough to avoid the adverse effect of data incompleteness on the identification of outliers. At this stage, we used the augmented version of the Iris data set of the first experiment, as well as a second, bigger, augmented version. Moreover, the brain tumour set, the blue crab data set of Campbell and Mahon (1974), and the multi-phase flow pipeline data set,[2] were also used. Secondly, it aimed to investigate how the t-GTM imputed the missing values and whether there was any difference between the way it did it for values corresponding to outliers and the way it did it for values corresponding to data dense regions. We also benchmarked the t-GTM with the standard Gaussian GTM.

All the aforementioned data sets are described in summary next.

5.1. Data sets

The Iris data set is considered first. It consists of 150 observations and 4 variables (sepal and petal lengths and widths) describing three iris flower species (*Setosa, Versicolor* and *Virginica*). For the experiments corresponding to the first goal described in the previous paragraphs, it was augmented by five artificially added outliers, prior to data normalization, sampled from a uniform distribution and with at least the value for one of the variables falling neatly outside the range spanned by the original variable. For the experiments corresponding to the second goal, a second set based on the Iris data, augmented this time by 20% of outlier instances, sampled from an artificially generated uniform distribution, was created.

For both types of experiments, a data set of single-voxel Magnetic Resonance Spectra from several human brain tumours and cystic growths was also used. These data consist of 98 spectra acquired in vivo for five tumour types: Astrocytes, Glioblastomas, Metastases, Meningiomas, and Oligodendrogliomas, and for cystic growths. The latter have a distinctive metabolic profile, quite different from the tumours themselves, characterized by high levels of lactate. Given their different but inhomogeneous composition, most of these cystic regions are likely to be outliers with respect to the tumours. The spectra were originally digitised, sampling the region known

[1] www.ics.uci.edu/~mlearn/MLRepository.html.

[2] Available from the GTM homepage at Aston University, UK: www.ncrg.aston.ac.uk/GTM/.

to contain clinically relevant metabolic information, into 194 frequency intensity values. In Huang, Lisboa, and El-Deredy (2003), a method based on Multivariate Bayesian Variable Selection provided a parsimonious and predictive description of the data in the form of six frequency intensities, assigned to Fatty Acids, Lactate, a compound-unassigned peak, Glutamine, Choline, and Taurine-Inositol. The data used in this study consist of these six variables.

For the experiments corresponding to the second goal described in the previous paragraphs, two extra sets were used, which had also been considered in similar studies: the blue crab data set, used in Peel and McLachlan (2000), and the multi-phase flow pipeline data set, used, for instance, in Tiňo and Nabney (2002). The crab data[3] consist of five morphological measurements on 200 specimen observations of crabs of genus *Leptograpsus*. The pipeline data set, generated synthetically, simulates the flow in an oil pipeline, which takes one out of three possible configurations: horizontally stratified, nested annular, or homogeneous mixture flow. It consists of 1000 observations and 12 variables. Following a similar procedure to the one sketched in Svensén and Bishop (2005), we added 20% of outliers to both data sets (after normalization), drawn from a uniform distribution on $[-10, 10]$ along each dimension.

5.2. Experimental settings

For both the GTM as a constrained mixture of Gaussians and the t-GTM, the adaptive parameters $\mathbf{W}$ and β were initialized, following a standard procedure (Bishop et al., 1998a). Matrix $\mathbf{W}$ was initialized to minimize the difference between the centres of the Gaussian distributions in data space $\mathbf{y}_i$ and the projections into data space that would be generated by a partial PCA, $\mathbf{y}'_i = \mathbf{V}_2\mathbf{u}_i$, where the columns of matrix $\mathbf{V}_2$ are the two principal eigenvectors (given that the latent space considered here was two-dimensional). This way, the replicability of the results was ensured. The variance β^{-1} was initialized as the larger of either the third principal component obtained in the PCA procedure or half the average minimum distance between latent points.

For all experiments outlined in the previous section, the grid of GTM latent centres was fixed to a square layout of 5×5 nodes (i.e., 25 constrained mixture components). The corresponding grid of basis functions ϕ_m was fixed to a 3×3 layout. Alternative layouts, in terms of shapes and sizes, are indeed possible and several were tested without significant differences (concerning the goals of the current analyses) being observed.

In the second set of experiments, several levels of data incompleteness were considered (5%, 10%, 15% and 20%) in order to test the limits of the robustness of the missing data imputation methods. Given that all the data sets selected for the second experiment were originally complete, the selection of those values that were artificially made missing for each of them was randomly varied 30 times, in order not to bias the results. Therefore, for each level of incompleteness, each GTM model was fitted to the data 30 times. This procedure was skipped over for the brain tumour data set for illustrative purposes. In all experiments, a fixed value of ν, common to all GTM mixture components, was used; preliminary tests were run for a range of its possible values in order to select that which maximized the complete log-likelihood in Eq. (21).

5.3. Results and discussion

Fig. 1 confirms the expectations regarding the first experiment stated in the previous section. In accordance with Peel and McLachlan (2000), we find that the posterior probabilities of the mixture components (latent space points of the t-GTM) given the data, expressed by Eq. (16), for the outliers artificially added to the Iris data set are much smaller than those obtained for the GTM described as a constrained mixture of Gaussians, expressed by Eq. (7). The same behaviour can also be observed, in Fig. 2, for an outlier cystic growth region of the brain tumour data set. In comparison, both variants of the GTM generate narrowly peaked posterior probabilities for instances belonging to the three classes of the original Iris data set. It can also be seen that these probabilities are narrower for the GTM defined as a constrained mixture of Gaussians. In fact, it is common, for this model, to find that the posterior probability in Eq. (7) is concentrated in a single component (Bishop et al., 1998b). This is not uncommon in the case of the t-GTM, although it is not unusual either to find posterior probabilities that are slightly spread across two or three components in neighbouring positions of the latent space.

As mentioned previously, a multivariate t-distribution converges to a multivariate normal distribution in the limit $\nu \to \infty$. Let us first calculate the maximum of the posterior probability in Eq. (16) across all t-GTM mixture components for each of the five outlier data instances added to the Iris data set, and then calculate their mean over these instances. This is illustrated in Fig. 3, where such a mean is displayed against increasing values of parameter ν. As expected, these maximum posterior probabilities increase monotonically as we get closer to a multivariate normal distribution. Therefore, as ν increases, the t-GTM loses its capability to minimize the influence of the outliers on the overall parameter estimation process.

The focus is now shifted to the second goal of the experiments, which first aims to answer the following question: Is the missing data imputation process robust enough to avoid the adverse effect of data incompleteness on the identification of outliers? As mentioned in the previous section, according to Peel and McLachlan (2000) a given data instance could be considered as an outlier if the value of

$$O_n = \sum_k \hat{z}_{kn} \frac{\nu + D}{\nu + \beta \left\| \mathbf{y}_k - \mathbf{x}_n \right\|^2} \tag{31}$$

was sufficiently small or, equivalently, the value of

$$O_n^* = \sum_k \hat{z}_{kn} \beta \left\| \mathbf{y}_k - \mathbf{x}_n \right\|^2 \tag{32}$$

[3] Available from Professor B. Ripley's web site at www.stats.ox.ac.uk/pub/PRNN/.

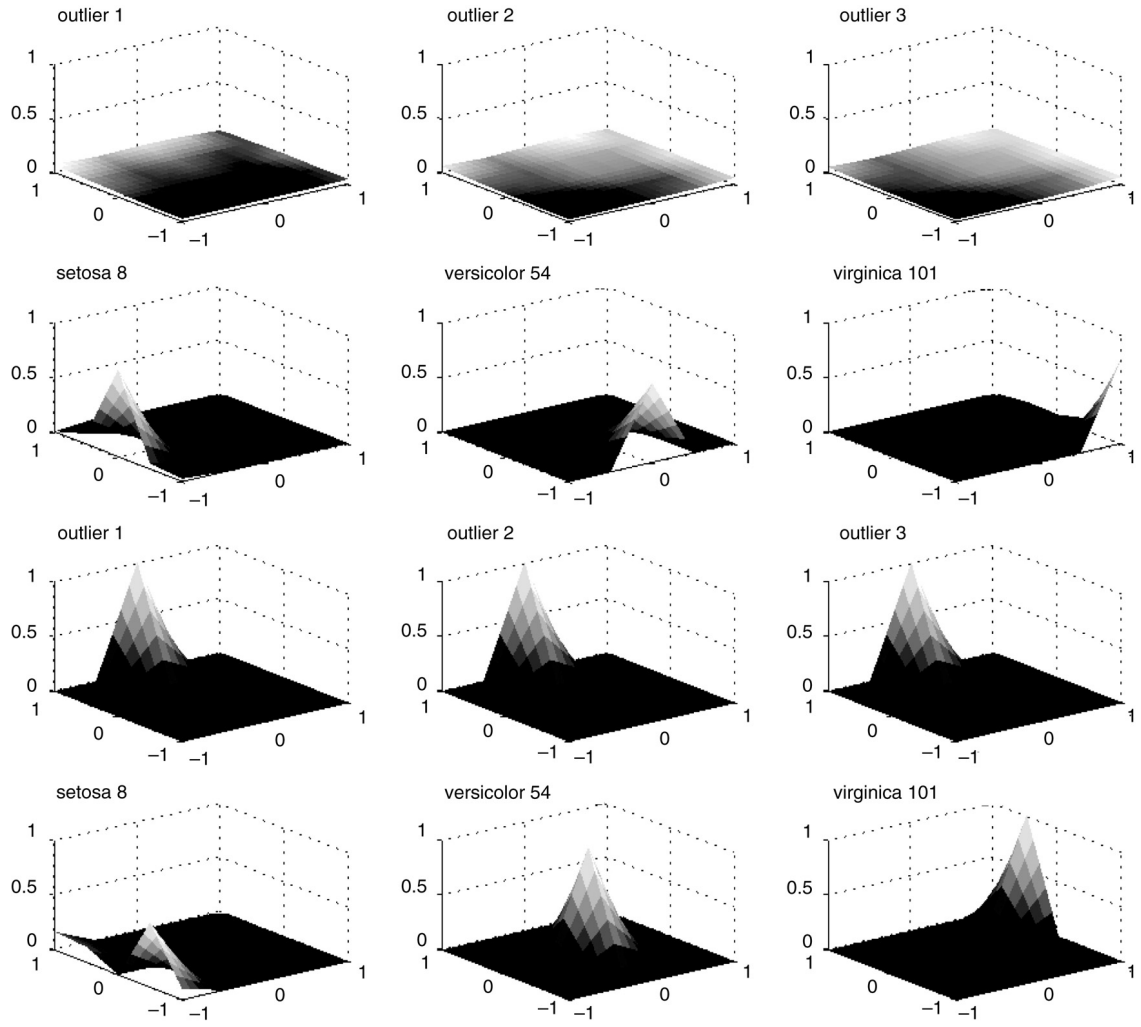

Fig. 1. Posterior probabilities of all 25 GTM components (latent nodes forming a 5×5 square grid), given several data instances. The top row corresponds to the t-GTM. It displays, on the vertical axis, the posterior probabilities from Eq. (16) for three out the five outliers added to the original Iris data. The third row displays the posterior probabilities for the same three outliers, as estimated from Eq. (7) by the Gaussian GTM. Whereas no mixture component takes the main responsibility for the outliers in the first case, the responsibility for the outliers is highly peaked around a single component in the case of the Gaussian GTM. The second and fourth rows, in turn for the t-GTM and the Gaussian GTM, correspond to the posterior probabilities for all components, given three data instances (one from each of the Iris data set classes: *Setosa, Versicolor* and *Virginica*) of the original data set.

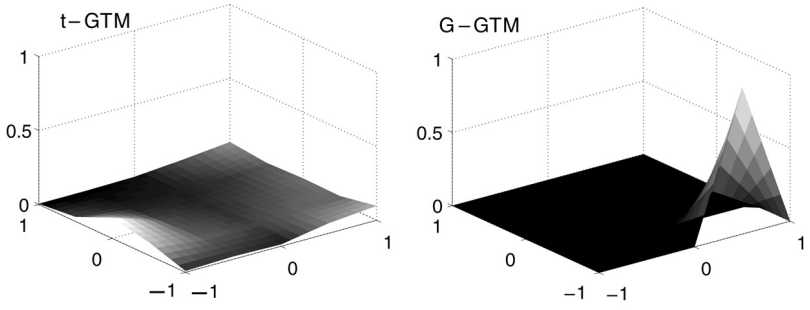

Fig. 2. Posterior probabilities of all 25 GTM components, given an extreme outlier of the brain tumour data set: *cystic growth 5*: left, t-GTM; right, Gaussian GTM (G-GTM). Once again, the responsibility taken by all components of t-GTM is almost flat compared to the narrowly peaked one corresponding to the Gaussian GTM.

was sufficiently large. Fig. 4 displays the histograms for the statistic in Eq. (32) at different levels of data incompleteness for the first augmented Iris data set. Bearing in mind that a histogram is just a simplification for illustrative purposes, it can be said that all five outliers are neatly isolated for levels of data incompleteness of 5%, 10%, and 15%. At an incompleteness level of 20%, one of the outliers is clearly not recognized as such, whereas one of the instances belonging to the Virginica

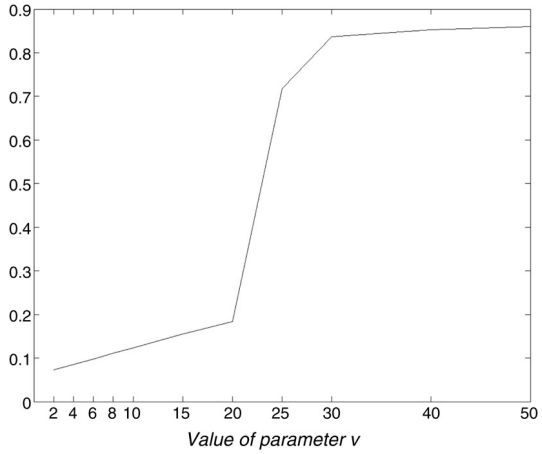

Fig. 3. Mean, over the five outliers in the first augmented Iris data set, of the máximum of the posterior probability across all GTM mixture components (i.e. $\text{mean}_{n=151,\ldots,155}\{\max_{k=1,\ldots,K}\{\hat{z}_{kn}\}\}$) as a function of parameter ν. It is calculated for a limited range of ν values (from 2 to 50). Note the almost sigmoidal shape of the function.

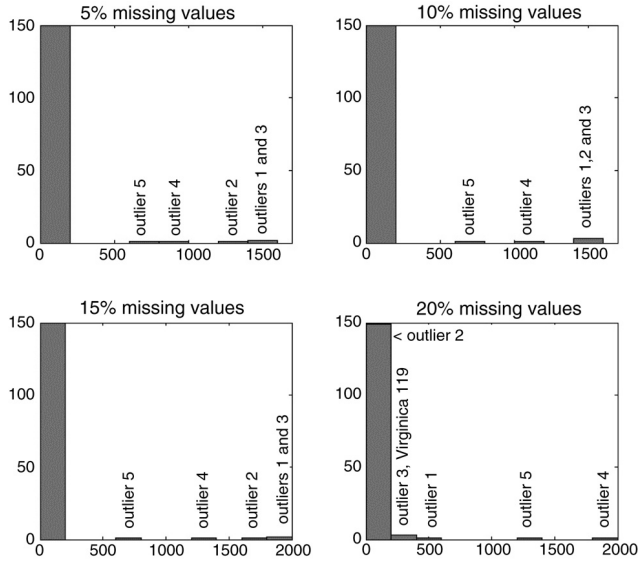

Fig. 4. Histograms of the statistic in Eq. (32), at different levels of data incompleteness, for the first augmented Iris data set. They illustrate the robustness of the model for outlier detection when data are incomplete. Results are commented on in the main text. In a practical implementation, a threshold value of this statistic might be set up to differentiate potential outliers.

class might well be considered as an outlier. It must be pointed out that this specific data instance (119 in the original Iris data set) has the highest observed value for variable *Petal Length*, which is, most probably, the reason for the misclassification. This is an indication of the fair robustness of the proposed model for simultaneous missing data imputation and outlier detection. To qualify this statement further, we now resort to several other data sets. Paired histograms for the statistic in Eq. (32), at different levels of data incompleteness from 5% to 20%, for the second augmented version of the Iris data set, which includes 20% of outliers, are displayed on the first two rows of Fig. 5. The histograms depicting outliers and non-outliers (original data) are neatly separated, with no overlapping, at all the evaluated levels of missing values present, confirming the robustness of the proposed model regarding outlier identification, even at rather high levels of both data incompleteness and outlier presence. The last two rows show similar and consistent results, in a more summarized form, for the crabs and pipeline data sets.

The same experiment was carried out for the brain tumour data set and for several levels of incompleteness. The corresponding histograms of Eq. (32) for some of these levels show in Fig. 6 that, even at rather high levels of incompleteness, the imputation behaves in a reasonably robust manner, as similar cystic growth regions are singled out as the most extreme outliers, which is discussed in more detail elsewhere (Vellido, Lisboa, & Vicente, 2006). The existing literature regards Lactate as a generically good discriminator of tumour types and grades (Huang et al., 2003; Preul et al., 1996), but it is known (Howe & Opstad, 2003) to discriminate especially well between tumours and cystic growth regions. The bottom plot in Fig. 6 partially illustrates this through a representation of Lactate versus Glutamine, where the most extreme cystic outliers singled out in the previous histograms have been labelled.

We might expect that the missing data imputation procedure associated with the t-GTM yielded different results depending on whether the missing values corresponded to outliers or non-outliers. This was indeed the case. The first two rows of Fig. 7 illustrate this by displaying, for different levels of data incompleteness from 5% to 20% of the second augmented Iris data set, the paired histograms of normalized errors $\left\|\mathbf{x}_n - \mathbf{x}_n^{\text{rec}}\right\|$ for both outliers and non-outliers, where $\mathbf{x}_n^{\text{rec}}$ is data instance n as reconstructed by the model and $\mathbf{x}_n$ is the observed complete data instance n. The majority of outliers show large errors associated with the imputation of their missing values. The level of data incompleteness, up to 20%, does not seem to have a significant effect on the distribution of these errors. Similar and consistent results, displayed in a more summarized manner in the last two rows of Fig. 7, were found for the crabs and the pipeline data sets. The interpretation of these results can be eased by using an overall measure, such as the normalized root mean squared error $\text{RMSE} = \left(\frac{1}{N}\sum_n^N \left\|\mathbf{x}_n - \mathbf{x}_n^{\text{rec}}\right\|^2\right)^{1/2}$, provided in two different circumstances: for all data, including outlier instances, and for the added outliers only. The results shown in Table 1 for the second augmented Iris set, the crabs set, and the pipeline set, consistently summarize those displayed in Fig. 7. From these results, it becomes clear that, beyond identifying the outliers in a data set, the t-GTM is capable of inhibiting their effect on the overall data model even when data are incomplete. To qualify this statement further, we compared the performance of t-GTM and the standard Gaussian GTM. These results are also added to Table 1. Overall, the t-GTM is shown to produce a more accurate missing data reconstruction but, equally, if we pay attention only to the reconstruction of the outliers (in brackets in Table 1), the t-GTM also imputes the missing values better than the Gaussian GTM. In spite of this, the percentage of the overall error corresponding to the reconstruction of the outliers is consistently smaller for the Gaussian model. In conclusion, the

Table 1
Normalized root mean square error of the overall missing data reconstruction, and the outliers missing data reconstruction (in brackets, together with the percentage of the former represented by the latter) for three data sets (Iris, crabs, and pipeline), two levels of missing values (10% and 20%), and two model variants (t-GTM and Gaussian GTM)

	10% missing values		20% missing values	
	t-GTM	Gaussian GTM	t-GTM	Gaussian GTM
Iris data	6,401 (5,947; 92.91%)	7,944 (7,025; 87.88%)	6,944 (6,347; 91.40%)	8,257 (7,163; 86.75%)
Crabs data	6,588 (6,302; 95.66%)	8,795 (7,881; 89.61%)	7,495 (7,159; 95.52%)	9,619 (8,492; 88.28%)
Pipeline data	8,247 (7,493; 90.86%)	9,175 (7,980; 86.98%)	10,288 (9,327; 90.66%)	11,272 (9,768; 86.66%)

Mean results over 30 different random selections of the missing values.

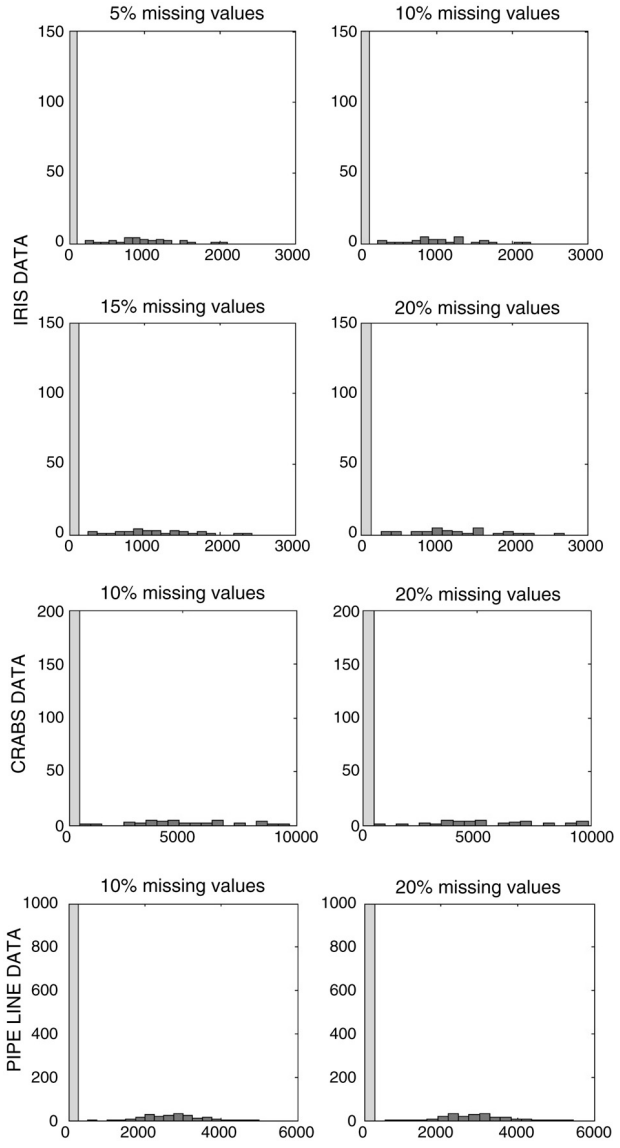

Fig. 5. Paired histograms of the statistic in Eq. (32), at different levels of data incompleteness, for the second augmented Iris set (first two rows) and for the crabs and pipeline data sets (last two rows). Light grey identifies the original data; dark grey identifies the added outliers. For all data incompleteness levels, these are mean results over 30 different random selections of the missing values. These histograms provide further evidence of the robustness of the model for outlier detection when data are incomplete.

t-GTM would be the model choice if we aimed to minimize the negative impact of outliers, or even to provide an overall better missing data reconstruction. Nevertheless, it should be born in mind that the t-GTM is likely to reconstruct outliers and non-outliers in a more different way than its Gaussian counterpart.

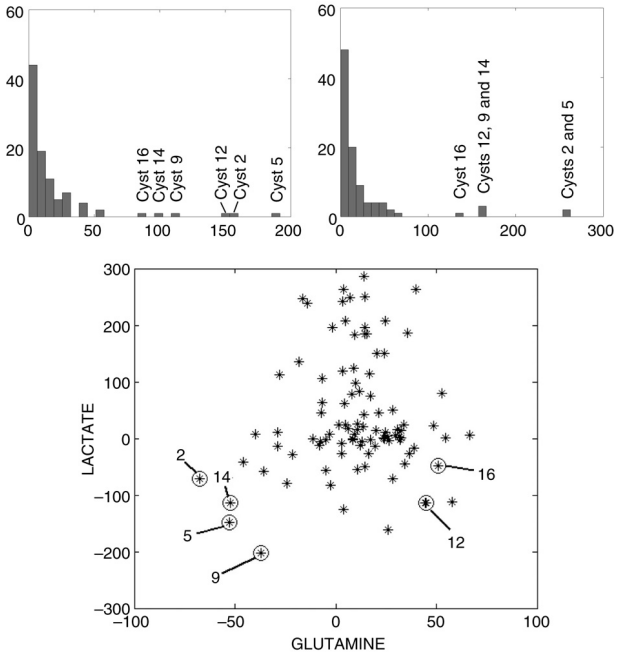

Fig. 6. Top row: histograms of the statistic in Eq. (32) at two levels of data incompleteness (10%: top left; 20%: top right) for the brain tumour data set. The most extreme cystic growth region outliers have been labelled. Bottom row: this is accompanied by a bi-plot, using the original data, of Glutamine (horizontal axis) versus Lactate (vertical axis) where the same outliers have been located.

6. Conclusions

Probabilistic models offer a consistent framework for dealing with problems that entail uncertainty. When probability theory lies at the foundation of a learning algorithm, the risk that the reasoning performed in it is inconsistent in some cases is lessened (Jaynes, 2003; Cerquides, 2004). For SOM, that lack of a probabilistic framework is a limitation. The GTM was defined as a probabilistic alternative to SOM precisely to overcome such a limitation, while preserving its multivariate data visualization and clustering advantages. Nevertheless, the GTM, defined as a constrained mixture of Gaussians, shares with its unconstrained counterparts (Gaussian mixture models) its lack of robustness towards outliers.

In this paper, the GTM has been redefined as a mixture of t-distributions, which are known to provide such robustness towards outlier data. This redefinition simultaneously provides

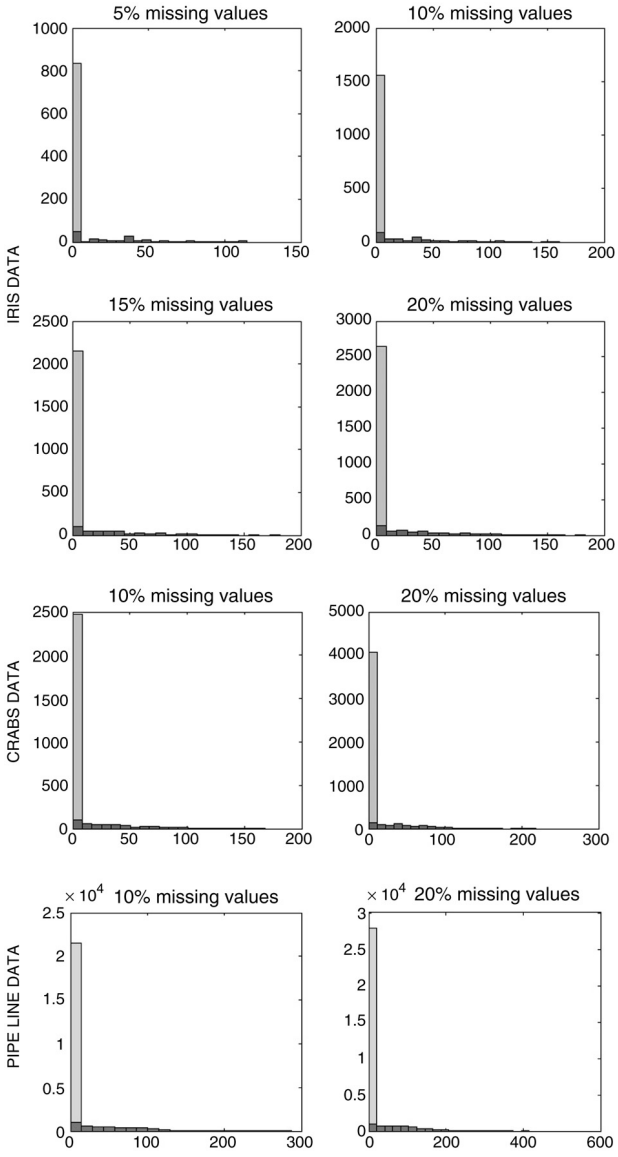

Fig. 7. Paired histograms of missing values estimation error, for the second augmented Iris set, at different levels of data incompleteness (first two rows) and for the crabs and pipeline data sets (last two rows). Light grey colour identifies the original data; dark grey identifies the added outliers. For all data incompleteness levels, these are cumulative results over 30 different random selections of the missing values (notice that the different vertical scales of the histograms, increasing with the percentage of missing values, are due to the fact that observations with no missing values are not included in the error calculation).

a procedure to impute missing information in case of data incompleteness, based on a modification of the EM algorithm for maximum likelihood estimation of the t-GTM parameters.

The experiments have highlighted the capability of the t-GTM not only to identify outliers, but also to effectively inhibit their possibly negative influence on the fitting of the model to the data. It has also been shown that the missing data imputation procedure embedded in t-GTM provides reasonably accurate estimates. This is an interesting result from the point of view of possible practical applications of the model in which the available data were incomplete: for some applications, the existence of outliers might be an undesirable feature, and a model that identified and inhibited their effect would be advantageous. Other applications, though, might still consider the information provided by outliers to be of qualitative interest; in such cases, the t-GTM would still provide a reasonably accurate imputation of their missing values.

An avenue for future research is the definition of criteria to estimate automatically the adequate number of mixture components for the t-GTM. This model can be used for simultaneous multivariate data visualization and clustering, but the number of latent points (or constrained mixture components, with their associated data clusters) is largely unconstrained, precisely for data visualization purposes. The automatic estimation of the number of components, in this case, could be a basis for an agglomerative procedure for merging individual components into bigger clusters. These bigger clusters are likely to be a more stable outcome for practical data clustering applications.

Also, research should be devoted to overcome the problem of convergence towards local maxima associated with the classical EM algorithm, which affects both the standard GTM and the t-GTM. In this direction, techniques based on deterministic annealing EM (DAEM: Ueda & Nakano, 1998), split-and-merge EM (SMEM: Ueda, Nakano, Ghahramani, & Hinton, 2000; Ueda & Ghahramani, 2002), or competitive EM (CEM: Zhang, Zhang, & Yi, 2004) might be explored.

Acknowledgements

Alfredo Vellido is a research fellow within the Ramón y Cajal program of the Spanish Ministry of Science and Education. The author would like to thank Professor Paulo J.G. Lisboa at Liverpool John Moores University, UK, Dr. Wael El-Deredy at The University of Manchester, UK, Mr. Iván Olier at Polytechnic University of Catalonia, Spain, as well as the anonymous referees, for useful discussions, comments and suggestions on earlier versions of the manuscript. The author also gratefully acknowledges Professor C. Arús at Universitat Autónoma de Barcelona for making available the brain tumour data set.

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. (1995). *Neural networks for pattern recognition*. Oxford University Press.

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* (pp. 111–116). London: IEE.

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: Chapman and 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.

Campbell, N. A., & Mahon, R. J. (1974). A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. *Australian Journal of Zoology*, *22*, 417–425.

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.

Cerquides, J. (2004). Improving Bayesian network classifiers. Ph.D. thesis. Barcelona, Spain: Polytechnic University of Catalonia (U.P.C.).

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.

de Ridder, D., & Franc, V. (2003). Robust subspace mixture models using t-distributions. In R. Harvey, & A. Bangham (Eds.), *Proceedings of the British machine vision conference 2003*. UK: Norwich.

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.

Howe, F. A., & Opstad, K. S. (2003). 1H MR spectroscopy of brain tumours and masses. *NMR in Biomedicine*, *16*, 123–131.

Huang, Y., Lisboa, P. J. G., & El-Deredy, W. (2003). Tumour grading from magnetic resonance spectroscopy: a comparison of feature extraction with variable selection. *Statistics in Medicine*, *22*, 147–164.

Jaynes, E. (2003). *Probability theory: the logic of science*. Cambridge: Cambridge University Press.

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. (2001). *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).

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

MacKay, D. J. C. (1995). Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, *6*, 469–505.

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, VA: 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.

Preul, M. C., Caramanos, Z., Collins, D. L., Villemure, J. -G., Leblanc, R., Olivier, A., et al. (1996). Accurate, non-invasive diagnosis of human brain tumours by using Proton Magnetic Resonance Spectroscopy. *Nature Medicine*, *2*, 323–325.

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, UK: NCRG, Aston University.

Svensén, M. (1998). GTM: The generative topographic mapping. Ph.D. thesis. Birmingham, UK: Aston University.

Svensén, M., & Bishop, C. M. (2005). Robust Bayesian mixture modelling. *Neurocomputing*, *64*, 235–252.

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.

Tipping, M. E., & Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural Computation*, *11*, 443–482.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*, 520–525.

Ueda, N., & Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, *15*, 1223–1241.

Ueda, N., & Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks*, *11*, 271–282.

Ueda, N., Nakano, R., Ghahramani, Z., & Hinton, G. E. (2000). SMEM algorithm for mixture models. *Neural Computation*, *12*, 2109–2128.

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.

Vellido, A., Lisboa, P. J. G., & Vicente, D. (2006). Robust analysis of MRS brain tumour data using t-GTM. *Neurocomputing*, *69*(7–9), 754–768.

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.

Zhang, B., Zhang, C., & Yi, X. (2004). Competitive EM algorithm for finite mixture models. *Pattern Recognition*, *37*, 131–144.