Preliminary theoretical results on a feature relevance determination method for Generative Topographic Mapping

Alfredo Vellido

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

C. Jordi Girona, 1-3. 08034, Barcelona, Spain.

Abstract

Feature selection (FS) has long been studied in classification and regression problems, following diverse approaches and resulting on a wide variety of methods, usually grouped as either *filters* or *wrappers*. In comparison, FS for unsupervised learning has received far less attention. For many real problems concerning unsupervised multivariate data clustering, FS becomes an issue of paramount importance as results have to meet interpretability and actionability requirements. A FS method for Gaussian mixture models was recently defined in Law et al. (2004). Mixture models are well established as clustering methods, but their multivariate data visualization capabilities are limited. The Generative Topographic Mapping (Bishop et al. 1998a), a constrained mixture of distributions, was originally defined to overcome such limitation. In this brief report we provide the theoretical development of a feature relevance determination method for Generative Topographic Mapping, based on that defined in Law et al. (2004); with this method, the clustering results can be visualized on a low dimensional latent space and interpreted in terms of a reduced subset of selected relevant features.

Keywords: Feature Selection; Feature Relevance Determination; Generative Topographic Mapping; Finite Mixture Models; Clustering; Expectation-Maximization

1. Introduction

Finite mixture models have settled in recent years as a standard for statistical modelling (McLachlan and Peel, 2000a). They can be used in classical data analysis problems such as clustering, regression and probability distribution modelling. This report focuses on their clustering capabilities. Gaussian mixture models (GMM), in particular, have received especial attention for their computational convenience (McLachlan and Peel, 2000b) to deal with multivariate continuous data. The usefulness of these models is reinforced by the wide spectrum of their applications (see, for instance, Wedel and Kamakura, 2000; McLachlan et al., 2004).

In practice, general finite mixture models suffer from several shortcomings that may limit their applicability; one of them is their lack of multivariate data visualization capabilities. Data visualization can be especially important in the exploratory stages of an analytical data mining process (Wong, 1999). The Generative Topographic Mapping (GTM) was originally defined by Bishop et al (1998a) as a constrained GMM allowing for multivariate data visualization on a low dimensional space. The model is constrained in that mixture components are equally weighted, share a common variance and their centres do not move independently from each other. This last feature also makes GTM an alternative, founded on probability theory, to the widely used (Kaski et al, 1998; Oja et al, 2002) Self-Organizing Maps (SOM: Kohonen, 2000). What makes the GTM especially useful is its combination of a readily interpretable clustering model with strong visualization capabilities (an extension of those of the SOM) and computational tractability. Its probabilistic setting ensures the existence of a proper error function and the convergence of its parameter optimization procedure, as well as enables the definition of principled extensions (Bishop et al, 1998b).

The interpretability of the clustering results provided by the GTM, even in terms of exploratory visualization, can be hampered when the data sets under analysis consist of a large number of features. This situation is not uncommon in real problems concerning clustering in areas such as, for instance, bioinformatics, chemometrics, or web mining. The data analyst would benefit from any method that allowed ranking the features according to their relative relevance and, ultimately, from a feature selection method. Feature selection (FS) has for long been the preserve of supervised methods for classification and regression problems. Diverse approaches have been followed, resulting on a wide variety of methods usually grouped as either *filters* or *wrappers*. Reviews of such methods can be found, for instance, in George (2000) and Kudo and Sklansky (2000). In comparison, FS for unsupervised learning has received far less attention, and initial strands of research have only started to shed light on this matter. For many real problems concerning unsupervised multivariate data clustering, FS becomes an issue of paramount importance as results have to meet interpretability and actionability requirements. Interpretability of clusters would be improved by their description in terms of a reduced subset of relevant variables, while clustering actionability (understood as the capability to act upon the clustering results), most important in managerial decision making problems such as market segmentation (Wedel and Kamakura, 2000), would be improved by enabling actions based only on a parsimonious subset of relevant features.

A recent main advance on feature selection in unsupervised model-based clustering has been presented in Law et al. (2004) for GMM. It provides a definition of unsupervised feature saliency and a method for its estimation as part of the Expectation-Maximization (EM: Dempster et al., 1977) algorithm. In this report we follow this approach to provide the theoretical development of a feature relevance determination (FRD) method for GTM; with this method, the clustering results can be analysed on a low dimensional visualization space and interpreted only in terms of a parsimonious subset of selected relevant features.

The remaining of the report is structured as follows. First, a brief definition of the standard GMM is provided, accompanied by the description of the estimation of its parameters within the EM framework. This is followed by a description of the FS method for GMM developed by Law et al. (2004). A self-contained introduction to the standard Gaussian GTM is then provided, followed by the presentation of the main contribution of this report: a FRD method for GTM, accompanied by a summary of the Maximum Likelihood estimation of its parameters within the EM framework; the corresponding details are presented in an appendix. The report wraps up with some brief conclusions and directions for future research.

2. Gaussian Mixture Models and the EM estimation of their parameters

In mixture models, the observed data are assumed to be samples of a combination or finite mixture of k=1,…,K components or underlying distributions, weighted by unknown priors $P(k)$. Given a D-dimensional dataset $\mathbf{X}=\{\mathbf{x}_n\}_{n=1}^{N}$, consisting of N random observations, the corresponding mixture density is defined as:

$$p(\mathbf{x})=\sum_{k=1}^{K} p(\mathbf{x}|k;\theta_k)P(k), \tag{1}$$

where each mixture component k is parameterized by θ_k. For continuous data, the choice of Gaussian distributions is a rather straightforward option due to their computational convenience (McLachlan and Peel, 2000), in which case

$$p(\mathbf{x}|k;\mu_k,\boldsymbol{\Sigma}_k)=(2\pi)^{-D/2}|\boldsymbol{\Sigma}_k|^{-1/2}\, exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)\right\}, \tag{2}$$

where the adaptive parameters θ_k are the mean vector and the covariance matrix of the D-variate distribution for each mixture component, namely μ_k and $\boldsymbol{\Sigma}_k$. Their Maximum Likelihood estimates can be obtained using the EM algorithm and, for that, first we define the complete log-likelihood as

$$L_c(\mu,\boldsymbol{\Sigma}|\mathbf{X})=log\prod_{n=1}^{N}p(\mathbf{x}_n)=\sum_{n=1}^{N}log\sum_{k=1}^{K}p(\mathbf{x}_n|k;\mu_k,\boldsymbol{\Sigma}_k)P(k). \tag{3}$$

In the context of the EM algorithm, we can introduce the binary indicator variables $\mathbf{Z}=\{\mathbf{z}_k\}_{k=1}^{K}$, with $\mathbf{z}_k=(z_{k1},\dots,z_{kN})$, which reflect our ignorance of which mixture component k is responsible for the generation of data observation *n*. The complete expected log-likelihood can now be expressed as

$$L_c(\mu,\boldsymbol{\Sigma}|\mathbf{X},\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\, log\left[p(\mathbf{x}_n|k;\mu_k,\boldsymbol{\Sigma}_k)P(k)\right]. \tag{4}$$

The indicators **Z** are effectively treated as missing data and, following the iterative EM procedure, the re-estimation of the adaptive parameters $\mu_k,\boldsymbol{\Sigma}_k$ requires the maximization of the expected log-likelihood $E\left[L_c(\mu,\boldsymbol{\Sigma}|\mathbf{X},\mathbf{Z})|\mathbf{X},\mu_k,\boldsymbol{\Sigma}_k\right]$.

The expectation of each of the indicators in **Z**, which is the probability of a mixture component *k* being responsible for data observation *n* (also known as responsibility r_{kn}) can be written as:

$$r_{kn} = p(k|\mathbf{x}_n,\mu_k,\boldsymbol{\Sigma}_k) = \frac{|\boldsymbol{\Sigma}_k|^{-1/2}\, exp\left\{-\frac{1}{2}(\mathbf{x}_n-\mu_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n-\mu_k)\right\}P(k)}{\sum_{k'=1}^{K}|\boldsymbol{\Sigma}_{k'}|^{-1/2}\, exp\left\{-\frac{1}{2}(\mathbf{x}_n-\mu_{k'})^T\boldsymbol{\Sigma}_{k'}^{-1}(\mathbf{x}_n-\mu_{k'})\right\}P(k')} \tag{5}$$

With this, in the maximization step, the update formulae for $\mu_k,\boldsymbol{\Sigma}_k$ are obtained as:

$$\hat{\mu}_k = \frac{\sum_{n=1}^{N} r_{kn}\mathbf{x}_n}{\sum_{n=1}^{N} r_{kn}} \tag{6}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{n=1}^{N} r_{kn}(\mathbf{x}_n-\hat{\mu}_k)(\mathbf{x}_n-\hat{\mu}_k)^T}{\sum_{n=1}^{N} r_{kn}} \tag{7}$$

2.1 Feature Relevance Determination in Gaussian Mixture Models

The problem of feature relative relevance determination for GMM was recently addressed by Law et al. (2004). Feature relevance in this unsupervised setting is understood as the likelihood of a feature being useful to define the data clustering structure. In that sense, it becomes a soft version of a FS method: no feature is actually meant to be discarded because none is likely to be either completely useful or useless. However, the resulting relevance ranking can be the basis of an *a posteriori* selection. A similar counterpart procedure for supervised models is Automatic Relevance Determination (ARD: MacKay, 1994; Qi et al., 2004)

Formally, the saliency of feature *d* is defined as $\rho_d = P(\eta_d = 1)$, where $\boldsymbol{\eta} = (\eta_1,...,\eta_D)$ is a further set of binary indicators that, like **Z**, can be integrated in the EM algorithm as missing variables. A value of $\eta_d = 1$ indicates the full relevance of feature *d*. According to this definition, the mixture density in Eq.1 can be rewritten as:

$$p(\mathbf{x})=\sum_{k=1}^{K}P(k)\prod_{d=1}^{D}\{\rho_d p(x_d|k;\theta_k)+(1-\rho_d)q(x_d|\lambda_d)\} \tag{8}$$

Notice that this entails the assumption that features are conditionally independent given a mixture component, which is equivalent to the assumption of a diagonal covariance matrix. The distribution p would be a univariate version of Eq.2, and the relevance of feature d would be given by ρ_d; consequently, a feature d would be considered as irrelevant, with *irrelevance* $(1-\rho_d)$, if, for all mixture components, $p(x_d|k;\theta_{kd})=q(x_d|\lambda_d)$, where $q(x_d|\lambda_d)$ is a common density followed by feature d, or common mixture component. Notice that this is tantamount to say that the distribution for feature d does not follow the cluster structure defined by the GMM. This common component should reflect any prior knowledge we might have regarding irrelevant features, or otherwise take the form of a general, uninformative distribution.

The maximum likelihood criterion can now be made explicit as the estimation of those model parameters that maximize the complete log-likelihood

$$L_c=\sum_{n=1}^{N}log\sum_{k=1}^{K}P(k)\prod_{d=1}^{D}(\rho_d p(x_d|k;\theta_k)+(1-\rho_d)q(x_d|\lambda_d)), \tag{9}$$

which can be accomplished using the EM algorithm (For details, see Law et al., 2004). The probability of a component k being the generator of observation n: r_{kn}, is computed in the expectation step of the algorithm as:

$$r_{kn}=\frac{P(k)\prod_d(\rho_d p(x_{nd}|k;\theta_{kd})+(1-\rho_d)q(x_{nd}|\lambda_d))}{\sum_{k'}P(k')\prod_d(\rho_d p(x_{nd}|k';\theta_{k'd})+(1-\rho_d)q(x_{nd}|\lambda_d))}. \tag{10}$$

Then, the maximization step provides update expressions for the components' priors $P(k)\equiv\alpha_k$, for the means and variances associated to each feature d in $p(\cdot|\cdot)$ and $q(\cdot|\cdot)$, as well as for the relevance parameter ρ_d:

$$\hat{\alpha}_k={\sum_n r_{kn}}\big/{N} \tag{11}$$

$$\hat{\mu}_{\theta_{kd}}=\frac{\sum_n\frac{\rho_d p(x_{nd}|k;\theta_{kd})}{\rho_d p(x_{nd}|k;\theta_{kd})+(1-\rho_d)q(x_{nd}|\lambda_d)}r_{kn}x_{nd}}{\sum_n\frac{\rho_d p(x_{nd}|k;\theta_{kd})}{\rho_d p(x_{nd}|k;\theta_{kd})+(1-\rho_d)q(x_{nd}|\lambda_d)}r_{kn}} \tag{12}$$

$$\hat{\boldsymbol{\Sigma}}_{\theta_{kd}}=\frac{\sum_n\frac{\rho_d p(x_{nd}|k;\theta_{kd})}{\rho_d p(x_{nd}|k;\theta_{kd})+(1-\rho_d)q(x_{nd}|\lambda_d)}r_{kn}(x_{nd}-\hat{\mu}_{\theta_{kd}})^2}{\sum_n\frac{\rho_d p(x_{nd}|k;\theta_{kd})}{\rho_d p(x_{nd}|k;\theta_{kd})+(1-\rho_d)q(x_{nd}|\lambda_d)}r_{kn}} \tag{13}$$

$$\hat{\mu}_{\lambda_d} = \frac{\sum_n \sum_k \left(\frac{(1-\rho_d) q(x_{nd}|\lambda_d)}{\rho_d p(x_{nd}|k;\theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn} \right) x_{nd}}{\sum_n \sum_k \frac{(1-\rho_d) q(x_{nd}|\lambda_d)}{\rho_d p(x_{nd}|k;\theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn}} \tag{14}$$

$$\hat{\boldsymbol{\Sigma}}_{\lambda_d} = \frac{\sum_n \sum_k \left(\frac{(1-\rho_d) q(x_{nd}|\lambda_d)}{\rho_d p(x_{nd}|k;\theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn} \right) \left(x_{nd} - \hat{\mu}_{\lambda_d} \right)^2}{\sum_n \sum_k \frac{(1-\rho_d) q(x_{nd}|\lambda_d)}{\rho_d p(x_{nd}|k;\theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn}} \tag{15}$$

$$\hat{\rho}_d = \frac{1}{N} \sum_{n,k} \frac{\rho_d p(x_{nd}|k;\theta_{kd})}{\rho_d p(x_{nd}|k;\theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn} \tag{16}$$

3. GTM as a constrained GMM

The GTM (Bishop et al., 1998a) was originally formulated both as a probabilistic alternative to SOM (Kohonen, 1995) and as a constrained mixture of distributions. It is precisely its constrained definition that allows overcoming the data and cluster visualization limitations of general finite mixture models. The GTM is a non-linear latent variable model that defines a mapping from a low dimensional latent space onto the multivariate data space. 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:

$$y_d(\mathbf{u}, \mathbf{W}) = \sum_{m}^{M} \phi_m(\mathbf{u}) w_{md} \, , \tag{17}$$

where $\boldsymbol{\Phi}$ is a set of M basis functions $\boldsymbol{\Phi}(\mathbf{u}) = (\phi_1(\mathbf{u}),...,\phi_M(\mathbf{u}))$ that were originally defined as spherically symmetric Gaussians $\phi_m(\mathbf{u}) = exp\left\{ -\frac{\|\mathbf{u} - \mu_m\|^2}{2\sigma^2} \right\}$, with μ_m the centres of the Gaussians and σ their common width; **W** is the matrix of adaptive weights w_{md} that defines the mapping; and **u** is a point in latent space. In order to achieve computational tractability and to provide an alternative to the clustering and visualization space defined by the characteristic SOM lattice, the latent space of the GTM is discretized as a regular grid of K latent points $\mathbf{u}_k$ defined by the probability

$$P(\mathbf{u}) = 1/K \sum_{k=1}^{K} \delta(\mathbf{u} - \mathbf{u}_k), \tag{18}$$

The probability distribution for a data point $\mathbf{x}$ takes the form of isotropic Gaussian noise and, given the adaptive parameters of the model, which are the matrix $\mathbf{W}$ and the inverse variance of the Gaussians β, it can be written as:

$$p(\mathbf{x}|\mathbf{u},\mathbf{W},\beta)=\left(\frac{\beta}{2\pi}\right)^{D/2} exp\left\{-\beta/_2 \|\mathbf{x}-\mathbf{y}\|^2\right\} \tag{19}$$

Marginalizing over the latent points and using Eq.18, we obtain

$$p(\mathbf{x}|\mathbf{W},\beta)=\int p(\mathbf{x}|\mathbf{u},\mathbf{W},\beta)P(\mathbf{u})d\mathbf{u}=\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\beta}{2\pi}\right)^{D/2} exp\left\{-\beta/_2 \|\mathbf{x}-\mathbf{y}_k\|^2\right\} \tag{20}$$

According to this general description, the GTM is a constrained mixture of Gaussians in the sense that all the components of the mixture 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. Notice that, given the common variance constrain, the GTM complies by definition with the assumption that features are conditionally independent given a mixture component, expressed in section 2.1.

The complete log-likelihood can now be defined as:

$$L_c(\mathbf{W},\beta|\mathbf{X})=\sum_{n=1}^{N} log\left\{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\beta}{2\pi}\right)^{D/2} exp\left\{-\beta/_2 \|\mathbf{x}_n-\mathbf{y}_k\|^2\right\}\right\} \tag{21}$$

As for GMM, we can resort to the EM algorithm to obtain the Maximum Likelihood estimates of the adaptive parameters $\mathbf{W}$ and β. Defining once again as $\mathbf{Z}$ the indicators describing our lack of knowledge of which latent point $\mathbf{u}_k$ is responsible for the generation of data point $\mathbf{x}_n$, the complete expected log-likelihood is defined as

$$L_c(\mathbf{W},\beta|\mathbf{X},\mathbf{Z})=\sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\, log\left[\left(\frac{\beta}{2\pi}\right)^{D/2} exp\left\{-\beta/_2 \|\mathbf{x}_n-\mathbf{y}_k\|^2\right\}\right] \tag{22}$$

The expected value of z_{kn} is now an special case of Eq.5

$$r_{kn}=P(k|\mathbf{x}_n,\mathbf{W},\beta)=\frac{exp\left\{-\frac{\beta}{2}\|\mathbf{x}_n-\mathbf{y}_k\|^2\right\}}{\sum_{k'=1}^{K} exp\left\{-\frac{\beta}{2}\|\mathbf{x}_n-\mathbf{y}_k\|^2\right\}} \tag{23}$$

The update expressions for $\mathbf{W}$ and β are computed in the maximization step. We obtain $\mathbf{W}^{new}$ as the solution of the following system of equations in matricial form:

$$\mathbf{\Phi}^T \mathbf{G\Phi W}^{new} - \mathbf{\Phi}^T \mathbf{RX} = 0\,, \tag{24}$$

where $\mathbf{\Phi}$ is a $K \times M$ matrix with elements $\phi_{km} = \phi_m(\mathbf{u}_k)$; $\mathbf{R}$ is the responsibility matrix, with elements r_{kn}; and $\mathbf{G}$ is a matrix with values $g_{kk'} = \begin{cases} \sum_{n=1}^{N} r_{kn}\,, k = k' \\ 0 \quad k \neq k' \end{cases}$.

Notice that Eq.24 is equivalent to Eq.6, given that the component centres for the GTM are described by $\mathbf{Y} = \mathbf{\Phi W}$.

The update expression for β is:

$$\left(\beta^{new}\right)^{-1} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{kn} \left\| \mathbf{x}_n - \mathbf{y}_k \right\|^2 \tag{25}$$

See Bishop et al. (1998a) for further details on these calculations.

3.1 Feature Relevance Determination in Generative Topographic Mapping: the FRD-GTM

The approach to feature relevance determination (FRD) described in section 2.1 can be transferred to the standard Gaussian GTM. It has to be born in mind, though, that, to some extent, the relevance of a feature depends on the number of clusters defined by a given solution. Considering the GTM strictly from its definition as a constrained mixture model, each of the points of the latent space sampling defined by Eq.18 can be thought as the generator of a single data cluster. For data visualization purposes, the number of latent points is left rather unconstrained in the usual GTM definition. Therefore, the FRD method applied to GTM should be understood as a constrained one in as far as it is meant to reach a compromise between its own ability as detector of feature relevance in clustering structure, and the data visualization capabilities of the GTM. In other words, for FRD-GTM, individual features are relevant in the sense that they explain the specific clustering structure provided by GTM, and not necessarily the unconstrained clustering structure.

For FRD-GTM, the complete log-likelihood in Eq.21 becomes:

$$L_c\left(\mathbf{W}, \mathbf{w}_o, \beta, \beta_o \middle| \mathbf{X}\right) = \sum_{n=1}^{N} \log \left\{ \frac{1}{K} \sum_{k=1}^{K} \prod_{d=1}^{D} \left(a_{knd} + b_{knd}\right) \right\}, \tag{26}$$

where

$$a_{knd} = \rho_d (\beta/2\pi)^{1/2} exp\left(-\beta/2\left(x_{nd} - \sum_m \phi_m(\mathbf{u}_k) w_{md}\right)^2\right) \tag{27}$$

and

$$b_{knd} = (1-\rho_d)(\beta_{o,d}/2\pi)^{1/2} exp\left(-\beta_{o,d}/2\left(x_{nd} - \phi_o(\mathbf{u}_o)\mathbf{w}_o\right)^2\right). \tag{28}$$

The common component requires the definition of two extra adaptive parameters $\mathbf{w}_o$ and $\boldsymbol{\beta}_o$, so that $\mathbf{y}_o = \phi_o(\mathbf{u}_o)\mathbf{w}_o$.

This common component accounts for data observations that the mixture components cannot explain well; in other words, data observations that do not fit with the cluster structure described by these components. This approach is not unlike the one commonly used to deal with the presence of atypical data observations, or outliers, when fitting Gaussian mixtures, which entails the inclusion of an additional component with a uniform distribution. This can be circumvented by the fitting of Student *t*-distributions mixtures (Peel and McLachlan, 2000), which has also been done for GTM (Vellido et al., 2005). The FRD method presented in this report, though, differs from the former on its featurewise approach.

Resorting again to the EM algorithm, we rewrite the complete log-likelihood of the model as:

$$L_c\left(\mathbf{W}, \mathbf{w}_o, \beta, \boldsymbol{\beta}_o \middle| \mathbf{X}, \mathbf{Z}\right) = \sum_{n,k} r_{kn} \sum_{d=1}^{D} \log(a_{knd} + b_{knd}) \tag{29}$$

where the expected *responsibiblity* in Eq.23 becomes:

$$r_{kn} = p\left(k \middle| \mathbf{x}_n, \mathbf{W}, \mathbf{w}_o, \beta, \boldsymbol{\beta}_o\right) = \frac{\prod_{d=1}^{D}(a_{knd} + b_{knd})}{\sum_{k'=1}^{K} \prod_{d=1}^{D}(a_{k'nd} + b_{k'nd})}. \tag{30}$$

The maximization of the expected log-likelihood for GTM yields the following update formulae for parameters ρ_d, $\mathbf{W}$, β, $\mathbf{w}_o$ and $\boldsymbol{\beta}_o$:

$$\rho_d^{new} = \frac{1}{N} \sum_{n,k} r_{kn} u_{knd}, \tag{31}$$

where

$$u_{knd} = \frac{a_{knd}}{a_{knd} + b_{knd}}. \tag{32}$$

$$\beta^{new} = \frac{\sum_{n.k} r_{kn} \sum_d u_{knd}}{\sum_{n.k} r_{kn} \sum_d u_{knd} \left(x_{nd} - \sum_m \phi_m(\mathbf{u}_k) w_{md}\right)^2} \tag{33}$$

$$\beta_{o,d}^{new} = \frac{\sum_{n.k} r_{kn} v_{knd}}{\sum_{n.k} r_{kn} v_{knd} \left(x_{nd} - \phi_o(\mathbf{u}_o) w_{o,d}\right)^2}, \tag{34}$$

where

$$v_{knd} = \frac{b_{knd}}{a_{knd} + b_{knd}}. \tag{35}$$

For fully relevant ($\rho_d \to 1$) features, the common component variance $\left(\beta_{o,d}\right)^{-1} \to 0$. We now obtain, for each feature *d*, the elements of matrix $\mathbf{W}^{new}$ as the solution of the following system of equations in matricial form:

$$\mathbf{\Phi}^T \mathbf{G}^* \mathbf{\Phi} \mathbf{W}_d{}^{new} - \mathbf{\Phi}^T \mathbf{R}^* \mathbf{X}_d = 0, \tag{36}$$

where $\mathbf{R}^*$ has elements $r_{kn}^* = u_{knd^*} r_{kn}$ for a given feature d^* with r_{kn} given by Eq.30, and $\mathbf{G}^*$ has elements $g_{kk'}^* = \begin{cases} \sum_{n=1}^{N} r_{kn}^*, k = k' \\ 0 \quad k \neq k' \end{cases}$. Notice the similarity of Eq.36 and Eq.12. Similarly, we obtain $\mathbf{w}_o^{new}$, featurewise, as the solution of:

$$\phi_o{}^T g^* \phi_o \mathbf{w}_{o,d}^{new} - \phi_o{}^T \mathbf{r}^* \mathbf{X}_d = 0 \tag{37}$$

where $\mathbf{r}^*$ has elements $r_n^* = \sum_k r_{kn}^* = \sum_k v_{knd^*} r_{kn}$ for a given feature d^*, and $g^* = \sum_{n,k} \overset{*}{r_{kn}}$. Details of all these calculations can be found in the appendix.

Note that the expression $u_{knd} r_{kn}$ could be considered as the *responsibility* of mixture component *k* for generating feature *d* of a data observation *n*. Correspondingly, expression $v_{knd} r_{kn}$ could actually be considered as the *irresponsibility* of mixture component *k* for generating feature *d* of a data observation *n*.

4. Conclusion

A definition of feature saliency for unsupervised clustering with GMM was recently provided by Law et al. (2004). In this report, we have detailed some preliminary theoretical developments concerning the extension of this method to the constrained mixture GTM model. The result is the definition of a feature relevance determination method for unsupervised clustering with GTM. The FRD-GTM model is capable of simultaneous multivariate data clustering and data visualization based upon relevant features. Future work will include the model implementation and its test using synthetic and real data. Further developments of FRD-GTM might include its extension to *t*-GTM: a constrained mixture of *t*-distributions (Vellido et al, 2005).

Acknowledgements

A. Vellido is a research fellow within the Ramón y Cajal program of the Spanish Ministry of Education and Science.

References

C.M. Bishop, M. Svensén, C.K.I. Williams, GTM: The Generative Topographic Mapping, Neural Computation 10(1) (1998a) 215-234.

C.M. Bishop, M. Svensén, C.K.I. Williams, Developments of the Generative Topographic Mapping, Neurocomputing 21(1-3) (1998b) 203-224.

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

E.I. George, The variable selection problem. Journal of American Statistical Association 95 (2000) 1304-1308.

S. Kaski, J. Kangas, T. Kohonen, Bibliography of self-organizing map (SOM) papers: 1981–1997, Neural Computing Surveys 1(3-4) (1998) 1–176.

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

M. Kudo, J. Sklansky, Comparison of algorithms that select features for pattern classifiers, Pattern Recognition 33 (2000) 25-41.

D.J.C MacKay, Bayesian methods for back-propagation networks, in E. Domany, J.L. van Hemmen, K. Schulten (eds.) *Models of Neural Networks III,* ch.6, Springer, New York (1994).

G.J. McLachlan, D. Peel, Finite Mixture Models, John Wiley & Sons, New York, 2000a.

G.J. McLachlan, D. Peel, On computational aspects of clustering via mixtures of normal and t-components, in: Proc. American Statistical Association (Bayesian Statistical Science Section) Alexandria, Virginia: American Statistical Association, 2000b.

G.J. McLachlan, K.-A. Do, C. Ambroise, Analyzing Microarray Gene Expression Data, John Wiley & Sons, New York, 2004.

M. Oja, S. Kaski, T. Kohonen, T., Bibliography of Self-Organizing Map (SOM) papers: 1998-2001 addendum, Neural Computing Surveys 3 (2002) 1-156.

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

Y. Qi, T.P. Minka, R.W. Picard, Z. Ghahramani, Predictive automatic relevance determination by expectation propagation, in: Proc. 21st International Conference on Machine Learning ICML, Banff, Canada, 2004.

A. Vellido, P.J.G. Lisboa, D. Vicente, Handling outliers and missing data in brain tumor clinical assessment using t-GTM, in: Proc. ESANN 2005, Bruges, Belgium, 2005.

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

P.C. Wong, Visual data mining, IEEE Computer Graphics and Applications 19(5) (1999) 20-21.

Appendix

In this appendix we provide a more detailed account of the calculations to obtain the update Eqs. 31, 33, 34, 36 and 37 in section 3.1 for the FRD-GTM, within the EM framework. Starting from the expression in Eq.29 for the complete log-likelihood, here in extended form

$$L_c = \sum\nolimits_{n,k} r_{kn} \sum_{d=1}^{D} log\left\{ \begin{array}{l} \rho_d \left(\frac{\beta}{2\pi}\right)^{1/2} exp\left(-\frac{\beta}{2}\left(x_{nd} - \sum_{m}^{M} \phi_m(\mathbf{u}_k) w_{md}\right)^2\right) + \\ (1-\rho_d)\left(\frac{\beta_{o,d}}{2\pi}\right)^{1/2} exp\left(-\frac{\beta_{o,d}}{2}\left(x_{nd}-\phi_o(\mathbf{u}_o) w_{o,d}\right)^2\right) \end{array}\right\}, \qquad \text{(A.1)}$$

update equations are obtained through maximization with respect to the various parameters. Maximization with respect to the elements of **W**, using differentiation rules and Eqs.27 and 28, implies:

$$\frac{\partial L_c}{\partial w_{ij}} = 0 = \sum\nolimits_{n,k} r_{kn} \frac{a_{knj} \cdot 2 \cdot \left(-\frac{\beta}{2}\right)\left(x_{nj} - \sum\nolimits_m \phi_m(\mathbf{u}_k) w_{mj}\right)\phi_i(\mathbf{u}_k)}{a_{knj} + b_{knj}}, \qquad \text{(A.2)}$$

and then

$$\sum\nolimits_{n,k} r_{kn} \frac{a_{knj}}{a_{knj} + b_{knj}} \left(x_{nj} - \sum\nolimits_m \phi_m(\mathbf{u}_k) w_{mj}\right)\phi_i(\mathbf{u}_k) = 0 . \qquad \text{(A.3)}$$

This leads, for each feature *d,* to Eq.36 in matricial form. Now, for the common component, the maximization with respect to the elements of $\mathbf{w}_o$:

$$\frac{\partial L_c}{\partial w_i} = 0 = \sum\nolimits_{n,k} r_{kn} \frac{b_{kni} \cdot 2 \cdot \left(-\frac{\beta_{o,i}}{2}\right)\left(x_{ni} - \phi_o(\mathbf{u}_o) w_{o,i}\right)\phi_o(\mathbf{u}_o)}{a_{kni} + b_{kni}} \qquad \text{(A.4)}$$

which implies

$$\sum\nolimits_{n,k} r_{kn} \frac{b_{kni}}{a_{kni} + b_{kni}} \left(x_{ni} - \phi_o(\mathbf{u}_o) w_i\right)\phi_o(\mathbf{u}_o) = 0 , \qquad \text{(A.5)}$$

leading, for each feature *d*, to Eq.37.

Expressions for the two inverse variance parameters: β and $\boldsymbol{\beta}_o$, are obtained as follows:

$$\frac{\partial L_c}{\partial \beta} = 0 = \sum\nolimits_{n,k} r_{kn} \sum_{d}^{D} \frac{1}{2} \frac{{}^{a_{knd}}\!\big/\!_{\beta} - a_{knd}\left(x_{nd} - \sum\nolimits_m \phi_m(\mathbf{u}_k) w_{md}\right)^2}{a_{knd} + b_{knd}} . \qquad \text{(A.6)}$$

Then,

$$\frac{1}{\beta}\sum_{n,k} r_{kn} \sum_{d} \frac{a_{knd}}{a_{knd}+b_{knd}} = \sum_{n,k} r_{kn} \sum_{d} \frac{a_{knd}\left(x_{nd}-\sum_{m}\phi_{m}(\mathbf{u}_{k})w_{md}\right)^{2}}{a_{knd}+b_{knd}}, \tag{A.7}$$

which leads to Eq.33. Similarly, for the common component:

$$\frac{\partial L_{c}}{\partial \beta_{o,i}} = 0 = \sum_{n,k} r_{kn} \frac{1}{2} \frac{b_{kni}/\beta_{o.i} - b_{kni}\left(x_{ni}-\phi_{o}(\mathbf{u}_{o})w_{o,i}\right)^{2}}{a_{kni}+b_{kni}}. \tag{A.8}$$

Then,

$$\frac{1}{\beta_{o,i}}\sum_{n,k} r_{kn} \frac{b_{kni}}{a_{kni}+b_{kni}} = \sum_{n,k} r_{kn} \frac{b_{kni}\left(x_{ni}-\phi_{o}(\mathbf{u}_{o})w_{o,i}\right)^{2}}{a_{kni}+b_{kni}}, \tag{A.9}$$

which leads to Eq.34.

Finally, maximizing with respect to ρ_{d}, and using Eqs. 32 and 35,

$$\frac{\partial L_{c}}{\partial \rho_{i}} = 0 = \sum_{n,k} r_{kn} \frac{a_{kni}/\rho_{i} - b_{kni}/(1-\rho_{i})}{a_{kni}+b_{kni}} = \sum_{n,k} r_{kn}\left(\frac{u_{kni}}{\rho_{i}} - \frac{v_{kni}}{(1-\rho_{i})}\right), \tag{A.10}$$

we obtain (recall that $\sum_{n,k} r_{kn} = N$)

$$\left(1-\rho_{i}^{new}\right) = \frac{1}{N}\sum_{n,k} r_{kn} v_{kni}, \tag{A.11}$$

which, given that $u_{kni} + v_{kni} = 1$, leads to Eq.31, as:

$$\rho_{i}^{new} = \frac{1}{N}\left(N - \sum_{n,k} r_{kn} v_{kni}\right) = \frac{1}{N}\sum_{n,k} r_{kn} u_{kni}. \tag{A.12}$$