An EM Algorithm for GTM-FS

Dharmesh M. Maniyar*
maniyard@aston.ac.uk

November 2, 2005

Abstract

We propose a generative topographic mapping (GTM) based data visualization with simultaneous feature selection (GTM-FS) approach which not only provides a better visualization by modeling irrelevant features ("noise") using a separate shared distribution but also gives a saliency value for each feature which helps the user to assess their significance. This technical report presents a varient of the Expectation-Maximization (EM) algorithm for GTM-FS.

1 GTM Architecture

In GTM-FS, the Gaussians in the constrained mixture of Gaussians have diagonal covariance. Roughly, GTM-FS Architecture can be displayed as below:

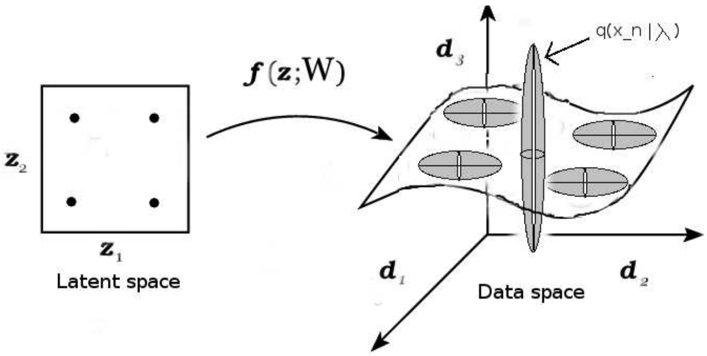

Figure 1: Schematic representation of the GTM model.

Following are the important dimension variables and indexes:

*Please note that this is an ad-hoc technical note. More structured report with clear notations will follow soon. Contact the author for a newer version.

N = Number of input data points. Index used : n.
M = Number of components (latent grid points). Index used : m.
D = Number of features (dimension of the data space). Index used : d.
K = Number of basis function for RBF mapping. Index used : k.

2 GTM with Feature Selection (GTM-FS)

GTM has a non-linear transformation from the latent space to the data space given by a linear combination of the basis functions. So that each point $\mathbf{z}_m$ in latent space is mapped to a corresponding point t_m in the D-dimensional data space (which acts as the centre of a Gaussian m) given by

$$\mathbf{T} = \mathbf{\Phi}(\mathbf{z})\mathbf{W}, \tag{1}$$

where $\mathbf{T}$ is an $M \times D$ matrix, $\mathbf{\Phi}$ is an $M \times K$ matrix, and $\mathbf{W}$ is a $K \times D$ matrix.

If we denote the node locations in latent space by $\mathbf{z}_m$, then eq. (1) defines a corresponding set of 'reference vectors' given by

$$t_{md} = \sum_{k=1}^{K} \phi_{mk}(\mathbf{z}_m) w_{kd}, \tag{2}$$

where t_{md} is a scalar and it represents estimated the dth feature of the mth component.

Each of the reference vectors then forms the centre of a Gaussian distribution in data space. For feature saliency purpose, we have one dimensional Gaussian for each feature,

$$p(x_{nd}|t_{md}, \sigma_{md}) = \frac{1}{\sqrt{2\pi\sigma_{md}^2}} \exp\left\{-\frac{(x_{nd} - t_{md})^2}{2\sigma_{md}^2}\right\}. \tag{3}$$

The probability density function for the GTM model is obtained by summing over all the Gaussian components, to give

$$p(\mathbf{x}|T, \Sigma^2) = \sum_{m=1}^{M} P(m) p(\mathbf{x}|\mathbf{t}_m, \sigma_m) \tag{4}$$

We assume that the features are conditionally independent given the (hidden) component label, so

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{m=1}^{M} \alpha_m \prod_{l=1}^{D} p(x_{nd}|\theta_{md}) \tag{5}$$

where $p(\cdot\,|\theta_{md}$ is the pdf of the dth feature for the mth component. $\theta_{md} = \{t_{md}, \sigma^2{}_d\}$ and α_m is $P(m)$ (prior).

The dth feature is irrelevant if its distribution is independent of the class labels, i.e., if it follows a common density, denoted by $q(x_{nd}|\lambda_d)$. Let $\Psi =$

$(\psi_1, ..., \psi_D)$ be an ordered set of binary parameters, such that $\psi_d = 1$ if feature d is relevant and $\psi_d = 0$, otherwise. The mixture density in eq. (5) is now:

$$p(\mathbf{x}_n|\Psi, \alpha_m, \theta_{md}, \lambda_d) = \sum_{m=1}^{M} \alpha_m \prod_{l=1}^{D} [p(x_{nd}|\theta_{md})]^{\psi_d} [q(x_{nd}|\lambda_d)]^{(1-\psi_d)} \tag{6}$$

Our notion of feature saliency is summarised in the following steps:

1. We treat the ψ_ds as missing variables
2. We define the feature saliency as $\rho_d = P(\psi_d = 1)$, the probability that the dth feature is relevant.

So the resulting model can be written as

$$p(\mathbf{x}_n|\mathbf{\Theta}) = \sum_{m=1}^{M} \alpha_m \prod_{l=1}^{D} (\rho_d p(x_{nd}|\theta_{md}) + (1-\rho_d) q(x_{nd}|\lambda_d)) \tag{7}$$

where $\Theta = \alpha_m, \theta_{md}, \lambda_d, \rho_d$ is the set of all the parameters of the model.
The complete-data log-likelihood for the model in eq. (7) is

$$P(\mathbf{x}_n, y_n = m, \mathbf{\Theta}) = \alpha_m \prod_{l=1}^{D} (\rho_d p(x_{nd}|\theta_{md}))^{\psi_d} ((1-\rho_d) q(x_{nd}|\lambda_d))^{(1-\psi_d)} \tag{8}$$

We can define the following quantities

$$s_{nm} = P(y_n = m|\mathbf{x}_n), \tag{9}$$
$$u_{nmd} = P(y_n = m, \psi_d = 1|\mathbf{x}_n), \tag{10}$$
$$v_{nmd} = P(y_n = m, \psi_d = 0|\mathbf{x}_n) \tag{11}$$

They are calculated using the current parameter estimate Θ^{new}. Now that $u_{nmd} + v_{nmd} = s_{nm}$ and $\sum_{n=1}^{N} \sum_{m=1}^{M} w_{nm} = N$. The expected complete data log-likelihood based on Θ^{old} we get

$$\begin{aligned} E_{\theta^{new}}[\ln P(X, \mathbf{z}, \Theta)] = & \sum_m (\sum_n s_{nm}) \ln \alpha_m + \\ & \sum_{md} \sum_n u_{nmd} \ln p(x_{nd}|\theta_{md}) + \\ & \sum_d \sum_{nm} v_{nmd} \ln q(x_{nd}|\lambda_d) + \\ & \sum_d \left(\ln \rho_d \sum_{nm} u_{nmd} + \ln(1-\rho_d) \sum_{nm} v_{nmd} \right) \end{aligned} \tag{12}$$

The four parts in the equation above can be maximised separately.

3 EM Algorithm

E-Steps: Compute the following quantities:

$$a_{nmd} = P(\psi_d = 1, x_{nd}|z_n = m) = \rho_d p(x_{nd}|\theta_{md}), \tag{13}$$
$$b_{nmd} = P(\psi_d = 0, x_{nd}|z_n = m) = (1 - \rho_d) q(x_{nd}|\lambda_d), \tag{14}$$
$$c_{nmd} = P(x_{nm}|z_n = m) = a_{nmd} + b_{nmd}, \tag{15}$$
$$s_{nm} = P(z_n = m|\mathbf{x}_n) = \frac{\alpha_m \prod_d c_{nmd}}{\sum_m \alpha_m \prod_d c_{nmd}}, \tag{16}$$
$$u_{nmd} = P(\psi_d = 1, z_n = m|\mathbf{x}_n) = \frac{a_{nmd}}{c_{nmd}} s_{nm}, \tag{17}$$
$$v_{nmd} = P(\psi_d = 0, z_n = m|\mathbf{x}_n) = s_{nm} - u_{nmd}. \tag{18}$$

To obtain re-estimation of the parameters, we consider complete log likelihood (eq. (12)) and using eq. (3) and eq. (2), we get following for the second term in eq. (12):

$$\mathcal{L}_{2ndpart} = \sum_{md} \sum_{n} u_{nmd} \ln p(x_{nd}|\theta_{md}), \tag{19}$$
$$\mathcal{L}_{2ndpart} = \sum_{md} \sum_{n} u_{nmd} \ln \left\{ \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp \left\{ -\frac{(x_{nd} - t_{md})^2}{2\sigma_d^2} \right\} \right\}, \tag{20}$$
$$\mathcal{L}_{2ndpart} = \sum_{md} \sum_{n} u_{nmd} \left[(-\frac{1}{2} \ln(\sigma_d^2)) - \frac{(x_{nd} - \Phi_m \mathbf{w}_d)^2}{2\sigma_d^2} \right]. \tag{21}$$

Now differentiating above equation w.r.t w_{id} (where $i \in 1, ..., K$, we get

$$\frac{\partial \mathcal{L}_{2ndpart}}{\partial w_{id}} = \sum_{m} \sum_{n} u_{nmd} \left[\frac{(x_{nd} - \Phi_m \mathbf{w}_d)}{\sigma_d^2} \phi_{mi} \right],$$

setting above equation to 0 and solving it we get

$$\sum_{m} \sum_{n} u_{nmd} [(x_{nd} - \Phi_m \mathbf{w}_d) \phi_{mi}] = 0. \tag{22}$$

This can be written in matrix notation in the form

$$\Phi_i^T \mathbf{U}_d \mathbf{x}_d = \Phi_i^T \mathbf{G}_d \Phi_m \mathbf{w}_d, \tag{23}$$

where Φ_m is a $1 \times K$ vector, $\mathbf{w}_d$ is a $K \times 1$ weight vector for the feature d, $\mathbf{R}_d$ is a $M \times N$ responsibility matrix for the feature d, $\mathbf{x}_d$ is a $N \times 1$ data vector for the feature d, and $\mathbf{G}_d$ is a $M \times M$ diagonal matrix with elements

$$g_{mmd} = \sum_{n}^{N} u_{nmd}. \tag{24}$$

So for all $i \in \{1, 2, ..., K\}$, we have,

$$\mathbf{\Phi}^T \mathbf{U}_d \mathbf{x}_d = \mathbf{\Phi}^T \mathbf{G}_d \mathbf{\Phi} \mathbf{w}_d, \tag{25}$$

Similarly, differentiating eq. (21) w.r.t σ_d, we get

$$\frac{\partial \mathcal{L}_{2ndpart}}{\partial \sigma_d} = \sum_m \sum_n u_{nmd} \left[-\frac{1}{2\hat{\sigma}_d^2} + \frac{(x_{nd} - \Phi_m \hat{\mathbf{w}})^2}{2(\hat{\sigma}_d^2)^2} \right] \tag{26}$$

setting above equation to 0 and solving it, we get

$$\hat{\sigma}_d = \frac{\sum_m \sum_n u_{nmd}(x_{nd} - \Phi_m \hat{\mathbf{w}}_d)^2}{\sum_m \sum_n u_{nmd}} \tag{27}$$

M-Steps: Reestimate the parameters according to following expressions:

$$\hat{\alpha_m} = \frac{\sum_n s_{nm}}{\sum_{nm} s_{nm}} = \frac{\sum_n s_{nm}}{N}, \tag{28}$$

$$\mathbf{\Phi}^T \mathbf{U}_d \mathbf{x}_d = \mathbf{\Phi}^T \mathbf{G}_d \mathbf{\Phi} \mathbf{w}_d, \text{Solve this to find the updated } \mathbf{w}_d \tag{29}$$

$$\widehat{\text{Mean in}} \theta_{md} = \Phi_m \hat{\mathbf{w}}_d, \tag{30}$$

$$\widehat{\text{Var in}} \theta_{md} = \frac{\sum_m \sum_n u_{nmd}(x_{nd} - \Phi_m \hat{\mathbf{w}}_d)^2}{\sum_m \sum_n u_{nmd}}, \tag{31}$$

$$\widehat{\text{Mean in}} \lambda_d = \frac{\sum_n (\sum_m v_{nmd}) x_{nd}}{\sum_{nm} v_{nmd}}, \tag{32}$$

$$\widehat{\text{Var in}} \lambda_d = \frac{\sum_n (\sum_m v_{nmd}) x_{nd}}{\sum_{nm} v_{nmd}}, \tag{33}$$

$$\hat{\rho}_d = \frac{\sum_n u_{nmd}}{\sum_{nm} u_{nmd} + \sum_{nm} v_{nmd}} = \frac{\sum_n u_{nmd}}{N} \tag{34}$$

More later ...