Visualisation of heterogeneous data with simultaneous feature saliency using Generalised Generative Topographic Mapping

Shahzad Mumtaz, Michel F. Randrianandrasana, Gurjinder Bassi and Ian T. Nabney

System Analytics Research Institute (SARI),
Aston University, Birmingham, B4 7ET, United Kingdom.
Email: mumtazs;randrimf;t-bassig;i.t.nabney@aston.ac.uk

Abstract. Most machine-learning algorithms are designed for datasets with features of a single type whereas very little attention has been given to datasets with mixed-type features. We recently proposed a model to handle mixed types with a probabilistic latent variable formalism. This proposed model describes the data by type-specific distributions that are conditionally independent given the latent space and is called generalised generative topographic mapping (GGTM). It has often been observed that visualisations of high-dimensional datasets can be poor in the presence of noisy features. In this paper we therefore propose to extend the GGTM to estimate feature saliency values (GGTMFS) as an integrated part of the parameter learning process with an *expectation-maximisation* (EM) algorithm. The efficacy of the proposed GGTMFS model is demonstrated both for synthetic and real datasets.

1 Introduction

Type-specific data analysis has been well studied in the machine learning community [6]. In the recent couple of decades, the need to analyse mixed-type data is gaining a lot of attention from machine learning experts because of the fact that real world processes often generate a data of mixed-type. An example of such a mixed-type data could be a hospital's patients' dataset where typical fields include age (discrete or continuous), gender (binary), test results (binary or continuous), height (continuous) etc. In practice a number of ad-hoc solutions are used to handle mixed-type data [6]. However, the ideal general solution for analysing such heterogeneous data is to specify a model that builds a joint distribution with the assumption of an appropriate noise distribution for each type of feature set (for example a Bernoulli for modelling binary, a multinomial for modelling multi-category features and a Gaussian for modelling continuous features) and then fitting the model to data where the parameter estimates are used to draw inferences [6].

In the literature there is no multivariate distribution that can model random variables of different types. However, one possible way of jointly modelling discrete and continuous features is using a latent variable approach to understand the correlation between features of different types in combination. Type-specific latent variable models have already been proposed such as a generative topographic mapping (GTM) appropriate for continuous features and a latent trait model (LTM) appropriate for discrete

type features [7] (an LTM was proposed as a generalisation of GTM model). This has encouraged us to recently propose to combine GTM and LTM [13] in a probabilistic non-linear latent variable model in a principled way to visualise mixed-type data on a single continuous latent space under a unified proposed framework of conditional independence criteria: we called this model a generalised-GTM (GGTM).

In principle, the machine learning algorithms assume to perform well in cases where we have more information about data instances. This suggests that the use of more features is important for the learning algorithms. However, in practice it is observed that not all the features are important. It is therefore useful to select a subset of features which are relevant thereby ignoring the irrelevant (noisy) features which compromise performance of the learning algorithm [8,9]. An understanding of which features are relevant is valuable in its own right. In the exploratory phases of analysis (which is when data visualisation is most used) it is usual to measure as many variables as is feasible, since it is not known which features are relevant to the task. Feature selection then plays an important role in simplifying the task and making data collection cheaper and faster.

Feature selection (FS) has been widely used in supervised learning problems where the search is guided by the known target values. FS methods can be categorized into four classes [1,14]: filters, wrappers, hybrid and embedded. Details of each of them are given in [10,14]. FS for unsupervised learning algorithms is a challenging task as there are no target values to guide the search. Very few attempts have been made to estimate the importance of features in the unsupervised learning algorithms. A brief review of feature selection in a clustering perspective is given in [1,4,8,14,18] and details of some previous attempts in the latent variable formalism are given in [5,9,11,15,16,17]. To the best of our knowledge, there is no similar approach in the literature for estimating feature saliency when modelling mixed-type data, though [4] did discuss this as a possible extension in the clustering perspective.

Our focus in this paper is to demonstrate an extension of the GGTM model to estimate feature saliency values not only for discrete type features but also for mixed-type features in a dataset, as an integrated part of the parameter learning process, under the latent variable formalism. The structure of the remainder of the paper is as follows. In Section 2, we explain our proposed GGTMFS model and derive the EM parameter learning process. Section 3 describes our experiments to demonstrate the effectiveness of the proposed approach. We conclude the paper in Section 4.

2 A GGTM with Simultaneous Feature Saliency (GGTMFS)

The main goal of a latent variable model is to find an M-dimensional manifold, $\mathcal{H}$, (usually $M = 2$) for the distribution $p(\mathbf{x})$ in a high-dimensional data space, $\mathcal{D}$, with D-dimensions. We write each observation vector, $\mathbf{x}_n$ in terms of sub-vectors $\mathbf{x}_n^{\mathcal{R}}$, $\mathbf{x}_n^{\mathcal{B}}$ and $\mathbf{x}_n^{\mathcal{C}}$ for continuous, binary and multi-category features respectively. In the rest of this paper we use superscript $\mathcal{R}$ for continuous features, superscript $\mathcal{B}$ for binary features and superscript $\mathcal{C}$ for categorical features. The symbol $|.|$ is used to indicate the number of features in each type of data space. We also use $\mathcal{M}$ to indicate either $\mathcal{R}$ or $\mathcal{B}$ or $\mathcal{C}$.

In this paper, we now propose an extension of the GGTM visualisation model described in [10] to simultaneously estimate feature saliencies (we call this extension GGTMFS) and learn the model parameters. To estimate feature saliency values, we assume that each feature is independent of the component label under the appropriate noise model distribution. As a special case for the Gaussian noise model, the feature independence assumption is modelled by adopting diagonal covariance matrices (as used in [8,9]) instead of spherical covariance (as used in [3] and GGTM). Now the probability density function of the GGTMFS model takes the form

$$p(\mathbf{x}_n|\pi,\Theta) = \sum_{k=1}^{K} \pi_k \left[\prod_{\mathcal{M}} \left[\prod_{d=1}^{|\mathcal{M}|} p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}}) \right] \right], \tag{1}$$

where $p(.|\Theta_{kd}^{\mathcal{M}})$ is the probability density functions of the dth feature for the kth component and π_k is the mixing coefficient of the kth component and is taken to be fixed to $\frac{1}{K}$ for all the components in the mixture model and $\mathcal{M} \in \{\mathcal{R}, \mathcal{B}, \mathcal{C}\}$) indicates type of data space (and the corresponding distributional assumption).

We make the definition that $\Psi^{\mathcal{M}} = (\psi_1^{\mathcal{M}}, \cdots, \psi_{|\mathcal{M}|}^{\mathcal{M}})$ (where $\mathcal{M} \in \{\mathcal{R}, \mathcal{B}, \mathcal{C}\}$), is the set of binary indicators $\psi_d^{\mathcal{M}} = 1$ for a relevant feature and $\psi_d^{\mathcal{M}} = 0$ otherwise. Combining $\psi_d^{\mathcal{M}}$ for each type of variable, we obtain $\Psi = \{\Psi^{\mathcal{R}}, \Psi^{\mathcal{B}}, \Psi^{\mathcal{C}}\}$. Now the probability density of our mixture model takes the form

$$p(\mathbf{x}_n|\pi,\Theta,\lambda,\Psi) = \sum_{k=1}^{K} \pi_k \left[\prod_{\mathcal{M}} \left[\prod_{d=1}^{|\mathcal{M}|} [p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}})]^{\psi_d^{\mathcal{M}}} [q(x_{nd}^{\mathcal{M}}|\lambda_d^{\mathcal{M}})]^{(1-\psi_d^{\mathcal{M}})} \right] \right]. \tag{2}$$

The common distribution $q(x_{nd}^{\mathcal{M}}|\lambda_d^{\mathcal{M}})$ is designed to explain all the data that is poorly explained by the GGTM model. The notion of feature saliency is modelled as follows: we first treat $\psi_d^{\mathcal{M}}$ as a missing variable in the EM algorithm and as a second step we estimate the feature saliency, $\rho_d^{\mathcal{M}} = p(\psi_d^{\mathcal{M}} = 1)$, which is the probability that the dth feature is relevant. The resulting model now takes the form,

$$p(\mathbf{x}_n|\Omega) = \sum_{k=1}^{K} \pi_k \left[\prod_{\mathcal{M}} \left[\prod_{d=1}^{|\mathcal{M}|} [\rho_d^{\mathcal{M}} p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}})] + [(1-\rho_d^{\mathcal{M}}) q(x_{nd}^{\mathcal{M}}|\lambda_d^{\mathcal{M}})] \right] \right], \tag{3}$$

where $\Omega = \left\{ \pi_k, \left\{\Theta_{kd}^{\mathcal{M}}\right\}, \left\{\lambda_d^{\mathcal{M}}\right\}, \left\{\rho_d^{\mathcal{M}}\right\} \right\}$ is the set of all the parameters of the model. A simple way to understand how Equation (3) is obtained is to observe that $[p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}})]^{\psi_d^{\mathcal{M}}}$ $[q(x_{nd}^{\mathcal{M}}|\lambda_d^{\mathcal{M}})]^{1-\psi_d^{\mathcal{M}}}$ can be re-written as $\psi_d^{\mathcal{M}}[p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}})] + (1-\psi_d^{\mathcal{M}})[q(x_{nd}^{\mathcal{M}}|\lambda_d^{\mathcal{M}})]$ given that $\psi_d^{\mathcal{M}}$ is a binary indicator variable (see the proof in [10,12]). The log-likelihood now takes the form

$$\mathcal{L}(\Omega) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\Omega). \tag{4}$$

2.1 An EM algorithm for GGTMFS

The latent structure of the GGTM model can be exploited to estimate feature saliencies, in a similar way as previously exploited for the standard GTM [9]. For this purpose, we consider flipping of a biased coin with probability $\rho_d^{\mathcal{M}}$; if the coin is a head then the feature is generated from the mixture component, $p(.|\Theta_{kd}^{\mathcal{M}})$, otherwise the 'background component' , $q(.|\lambda_d^{\mathcal{M}})$, is responsible.

We treat $\mathbf{Y}$ (i.e. component labels) and Ψ as missing variables and we can derive an EM algorithm for estimating model parameters (see details in [10,12]). In the **E-step**, we use the current set of parameters, Ω, to compute the posterior probability (i.e. responsibility) $r_{nk} = p(y_n = k|\mathbf{x}_n)$ using Bayes' theorem,

$$\frac{\pi_k \left[\prod_{\mathcal{M}} \left[\prod_{d=1}^{|\mathcal{M}|}[\rho_d^{\mathcal{M}} p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}})] + [(1-\rho_d^{\mathcal{M}})q(x_{nd}^{\mathcal{M}}|\lambda_d^{\mathcal{M}})]\right]\right]}{\sum_{k=1}^{K} \pi_k \left[\prod_{\mathcal{M}} \left[\prod_{d=1}^{|\mathcal{M}|}[\rho_d^{\mathcal{M}} p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}})] + [(1-\rho_d^{\mathcal{M}})q(x_{nd}^{\mathcal{M}}|\lambda_d^{\mathcal{M}})]\right]\right]}. \quad (5)$$

The responsibility matrix, $\mathbf{R}$, is used to compute $u_{nkd}^{\mathcal{M}} = p(\psi_d^{\mathcal{M}} = 1, y_n = k|\mathbf{x}_n^{\mathcal{M}})$, which is a measure of the importance of the nth data point relating to the kth component using the dth feature of the $\mathcal{M}$ type observation space and $v_{nkd}^{\mathcal{M}} = p(\psi_d^{\mathcal{M}} = 0, y_n = k|\mathbf{x}_n^{\mathcal{M}})$.

$$u_{nkd}^{\mathcal{M}} = \frac{\rho_d^{\mathcal{M}} p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}})}{\rho_d^{\mathcal{M}} p(x_{nd}^{\mathcal{M}}|\Theta_{kd}^{\mathcal{M}})] + [(1-\rho_d^{\mathcal{M}})q(x_{nd}|\lambda_d^{\mathcal{M}})} r_{nk}, \quad (6)$$

$$v_{nkd}^{\mathcal{M}} = r_{nk} - u_{nkd}^{\mathcal{M}}. \quad (7)$$

M-step: We can use $\mathbf{U}^{\mathcal{M}}$ to re-estimate the weight matrix $\mathbf{W}^{\mathcal{M}}$ (i.e. $\mathcal{M}$ indicate type of observation space) using a set of linear equations. Both for binary and multinomial cases we use gradient-based approach as used in [7]. The weight vector $\mathbf{w}_d^{\mathcal{M}}$ of each dth feature can be updated using

$$\widehat{\mathbf{w}_d^{\mathcal{R}}} = (\boldsymbol{\Phi}^T \mathbf{E}_d^{\mathcal{R}} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{U}_d^{\mathcal{R}} \mathbf{x}_d^{\mathcal{R}}, \quad (8) \quad \Delta \mathbf{w}_d^{\mathcal{B}} \propto \boldsymbol{\Phi}^T \left[\mathbf{U}_d^{\mathcal{B}} \mathbf{x}_d^{\mathcal{B}} - \mathbf{E}_d^{\mathcal{B}} g^{\mathcal{B}}(\boldsymbol{\Phi} \mathbf{w}_d^{\mathcal{B}})\right], \quad (9)$$

$$\Delta \mathbf{W}_{\mathbf{S}_d}^{\mathcal{C}} \propto \boldsymbol{\Phi}^T \left[\mathbf{U}_d^{\mathcal{C}} \mathbf{X}_{\mathbf{S}_d}^{\mathcal{C}} - \mathbf{E}_d^{\mathcal{C}} g^{\mathcal{C}}(\boldsymbol{\Phi} \mathbf{W}_{\mathbf{S}_d}^{\mathcal{C}})\right], \quad (10)$$

where $\boldsymbol{\Phi}$ is a $K \times L$ matrix, $\mathbf{U}_d^{\mathcal{M}}$ is a $K \times N$ matrix calculated using Equation (6), $\mathbf{x}_d^{\mathcal{M}}$ is an $N \times 1$ data vector of real/binary values (the $\mathbf{X}_{\mathbf{S}_d}^{\mathcal{C}}$ is binary encoded matrix of dth multi-category feature) and the diagonal matrix $\mathbf{E}_d^{\mathcal{M}}$ has values $e_{kkd}^{\mathcal{M}} = \sum_{n=1}^{N} u_{nkd}^{\mathcal{M}}$.

Now we can straightforwardly re-estimate parameters of the mixture model using the re-estimated weight matrix of each type, $\widehat{\mathbf{W}}^{\mathcal{M}}$: first we re-estimate the centres (for each type features) of the mixture model in the data space (see Equations (11), (12) and (13))

$$\widehat{Mean\Theta}_k^{\mathcal{R}} = \widehat{\mathbf{m}_k^{\mathcal{R}}} = \boldsymbol{\Phi}(\mathbf{z}_k)\widehat{W^{\mathcal{R}}}, \quad (11)$$

$$\widehat{\mathbf{m}_k^{\mathcal{B}}} = g^{\mathcal{B}}(\boldsymbol{\Phi}(\mathbf{z}_k)\widehat{\mathbf{W}^{\mathcal{B}}}), \qquad (12) \qquad \widehat{\mathbf{m}_{kS_d}^{\mathcal{C}}} = g^{\mathcal{C}}(\boldsymbol{\Phi}(\mathbf{z}_k)\mathbf{w}_{S_d}^{\mathcal{C}}), \qquad (13)$$

where $\widehat{\mathbf{m}_k^{\mathcal{M}}}$ is a $1 \times |\mathcal{M}|$ vector, $g^{\mathcal{B}}(.)$ is a logistic sigmoid and $g^{\mathcal{C}}(.)$ is a *softmax* function. We use re-estimated centre to update the diagonal Gaussian width in each direction (for each continuous feature): see Equation (14) (similar to standard GTMFS [9])

$$\frac{1}{\widehat{\beta_d^{\mathcal{R}}}} = \frac{\sum_k \sum_n u_{nkd}^{\mathcal{R}} (x_{nd}^{\mathcal{R}} - \widehat{m_{kd}^{\mathcal{R}}})^2}{\sum_k \sum_n u_{nkd}^{\mathcal{R}}}. \qquad (14)$$

Common density parameters, $\lambda_d^{\mathcal{R}}$, can be updated using

$$\widehat{Mean\lambda_d^{\mathcal{R}}} = \frac{\sum_n (\sum_k v_{nkd}^{\mathcal{R}}) x_{nd}^{\mathcal{R}}}{\sum_{nk} v_{nkd}^{\mathcal{R}}}. \qquad (15)$$

$$\widehat{Mean\lambda_d^{\mathcal{B}}} = \frac{\sum_n (\sum_k v_{nkd}^{\mathcal{B}}) x_{nd}^{\mathcal{B}}}{\sum_{nk} v_{nkd}^{\mathcal{B}}}. \quad (16) \quad \widehat{Mean\lambda_{S_d}^{\mathcal{C}}} = \frac{\sum_n (\sum_k v_{nkd}^{\mathcal{C}}) x_{nS_d}^{\mathcal{C}}}{\sum_{nk} v_{nkd}^{\mathcal{C}}}. \quad (17)$$

$$\widehat{Var\lambda_d^{\mathcal{R}}} = \frac{\sum_n (\sum_k v_{nkd}^{\mathcal{R}} (x_{nd}^{\mathcal{R}} - \widehat{Mean\lambda_d^{\mathcal{R}}})^2}{\sum_{nk} v_{nkd}^{\mathcal{R}}}. \qquad (18)$$

For the feature saliency parameter update, we use prior distributions for each type of variable separately as explained in [12]. The resultant feature saliency updates are

$$\widehat{\rho_d^{\mathcal{R}}} = \frac{\max(\sum_{nk} u_{nkd}^{\mathcal{R}} - \frac{KP}{2}, 0)}{\max(\sum_{nk} u_{nkd}^{\mathcal{R}} - \frac{KP}{2}, 0) + \max(\sum_{nk} v_{nkd}^{\mathcal{R}} - \frac{T}{2}, 0)}, \qquad (19)$$

where P and T are the number of parameters in $\Theta_{kd}^{\mathcal{R}}$ and $\lambda_d^{\mathcal{R}}$ respectively.

$$\widehat{\rho_d^{\mathcal{B}}} = \frac{\max(\sum_{nk} u_{nkd}^{\mathcal{B}} + \alpha_d - 1, 0)}{\max(\sum_{nk} u_{nkd}^{\mathcal{B}} + \alpha_d - 1, 0) + \max(\sum_{nk} v_{nkd}^{\mathcal{B}} + \beta_d - 1, 0)}. \qquad (20)$$

$$\widehat{\rho_d^{\mathcal{C}}} = \frac{\max\left(\sum_{nk} u_{nkd}^{\mathcal{C}} - \frac{K(c_d-1)}{2}, 0\right)}{\max\left(\sum_{nk}, u_{nkd}^{\mathcal{C}} - \frac{K(c_d-1)}{2}, 0\right) + \max\left(\sum_{nk} v_{nkd}^{\mathcal{C}} - \frac{(c_d-1)}{2}, 0\right)}, \qquad (21)$$

where c_d represents number of categories for the dth feature. We also extend GGTMFS by deriving an *expectation-maximisation* (EM) variant to incorporate missing values (for details see the technical report [12]).

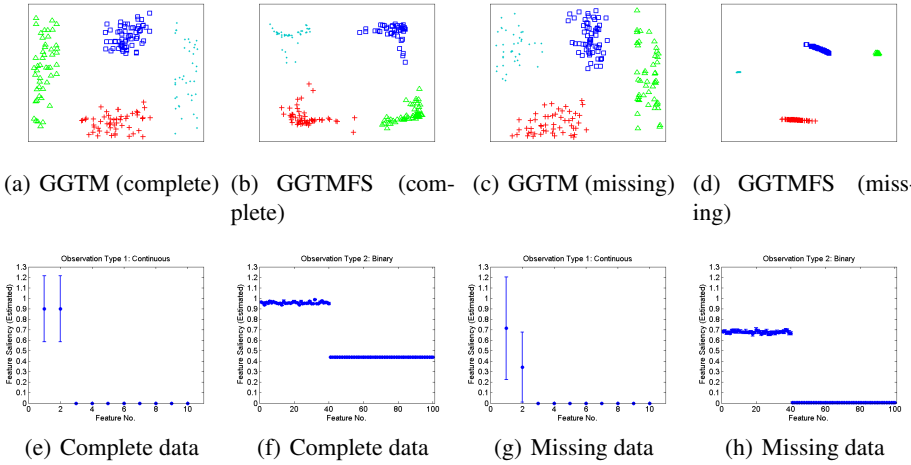
3 Experiments

A series of experiments was performed to demonstrate the effectiveness of the proposed GGTMFS model for both synthetic and real datasets. Each weight sub-matrix (i.e. $\mathbf{W}^{\mathcal{R}}$, $\mathbf{W}^{\mathcal{B}}$ and $\mathbf{W}^{\mathcal{C}}$) was initialised using principal component analysis (PCA). On average, 500 iterations of EM were sufficient for convergence. We used a latent grid of size 8×8 and an RBF grid of size 4×4.

3.1 Synthetic data

A synthetic data was used to assess the GGTMFS model: a combination of continuous and binary features. A comparison of the resulting projections with those given by the GGTM model is also shown on both complete and incomplete data where 10% of the data was removed at random for each observation space. We first generated 2 feature dataset with 2000 data points from an equiprobable mixture of four Gaussians (for details see technical report [12]) and then generated 8 noisy features (where each feature was sampled independently from $\mathcal{N}(0, I)$ distribution) and combined them yielding a 10-feature dataset. We then generated a binary dataset of 100 features where the first 40 features were drawn from four equiprobable clusters and the remaining 60 features are noisy (with random distribution of 1s). A small amount of noise (5%) was added by inserting random 0s in the informative features. For the uninformative features, we added a random distribution of 1s with different percentages by $20\%, 40\%, 60\%, 80\%$ and also with no or all 1s in the uninformative features (and we report here results of binary uninformative features with no $1s$). We then combined both continuous and binary features yielding a dataset with 110 features.

Visualisation results for GGTM and GGTMFS and saliency values estimated from the GGTMFS are presented in Figure 1 for both complete and incomplete data.



(a) GGTM (complete) (b) GGTMFS (complete) (c) GGTM (missing) (d) GGTMFS (missing)

(e) Complete data (f) Complete data (g) Missing data (h) Missing data

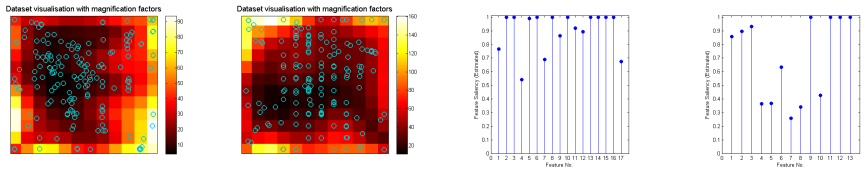
Fig. 1. GGTM and GGTMFS visualisations and saliencies of the synthetic mixed-type complete and missing datasets with 10 continuous and 100 binary features. 10% of the continuous and binary complete data have been randomly removed to produce the missing data. GGTMFS visualisations quite often show more compact clusters compared to GGTM visualisations. Saliencies plots show results as error bars from our cross-validation results. (e) and (g) show FS values of continuous features whereas (f) and (h) show FS values of binary features.

3.2 Real-world data: oil exploration

We applied the GGTMFS to two sets of oil exploration data from the Barents Sea: oil maturity and environmental parameters. The maturity data consists of 17 continu-

ous features with a large fraction (34%) of missing values and the environment data contains 13 continuous features with 16% of missing values. Both datasets consist of 168 samples. All the variables in the maturity dataset are important except feature 7, which has an environment influence that might make it behave differently, and feature 17, which is an environmental parameter. The features in the environment data are a combination of geochemical properties with variable importance. Features 6, 7, 11 and 12 are more influenced by maturity than environment and hence their saliency values should be low.

The resulting GGTMFS visualisation and saliency values plots are shown in Figure 2. The GGTMFS plots are superimposed with magnification factor plots which enable the user to observe the amount of stretching of the data-space manifold at different parts of the latent space [2]. This is useful to understand how the data is embedded in the data space, detect outliers and separate clusters. The magnification factors are represented by colour shading in the projection manifold: the lighter the colour, the more stretch in the projection manifold. The GGTMFS visualisations on the oil data show



(a) Projection (maturity) (b) Projection (environment) (c) Saliencies (maturity) (d) Saliencies (environment)

Fig. 2. GGTMFS plot of the maturity (see (a)) and environment data (see (b)) and estimated feature saliency values of continuous features (see (c) and (d)).

that there is relatively little discrete structure (no clear clusters) in the data. The model was able to give a sensible saliency value (0.675) to the feature 17 in the maturity data as this variable is actually an environmental parameter. However, feature 1 should have a high saliency value according to the domain experts. In the environment data, the low saliency values of the features 6 and 7 make sense given that these features have a strong maturity influence. However, features 11 and 12 should also have low saliecny values, and feature 10 should have a high saliency value.

4 Conclusion

We derived a non-linear model for visualising a mixed-type dataset to simultaneously estimate saliency values both for complete and incomplete datasets. We called this model a generalised GTM with simultaneous feature saliency estimation (GGTMFS). Experimental visualisation results for both synthetic and real mixed-type datasets have shown that this model, unlike GGTM, provided more compact clusters especially in the presence of missing values and irrelevant features. More detailed results with other datasets are available in a technical report [12].

References

1. S. Alelyani, J. Tang, and H. Liu. Feature selection for clustering: A review. In *Data Clustering: Algorithms and Applications*, pages 29–60. Chapman and Hall/CRC, 2013.
2. C. Bishop, M. Svensén, and C. K. I. Williams. Magnification factors for the GTM algorithm. In *In Proceedings IEE Fifth International Conference on Artificial Neural Networks*, pages 64–69, 1997.
3. C. M. Bishop and M. Svensen. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.
4. N. Bouguila. On multivariate binary data clustering and feature weighting. *Comput. Stat. Data Anal.*, 54(1):120–134, 2010.
5. I. O. Caparroso. *Variational Bayesian algorithms for generative topographic mapping and its extensions*. PhD thesis, Universitat Politècnica de Catalunya, 2008.
6. A. R. de Leon and K. C. Chough. *Analysis of Mixed Data: Methods & Applications*. Taylor & Francis Group. Chapman and Hall/CRC, 2013.
7. A. Kabán and M. Girolami. A combined latent class and trait model for the analysis and visualization of discrete data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):859–872, 2001.
8. M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004.
9. D. M. Maniyar and I. T. Nabney. Data visualization with simultaneous feature selection. In *Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB '06. 2006 IEEE Symposium on*, pages 1–8, 2006.
10. S. Mumtaz. *Visualisation of bioinformatics datasets*. PhD thesis, Aston University, 2015.
11. S. Mumtaz, I. T. Nabney, and D. R. Flower. Novel visualization methods for protein data. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012*, pages 198 –205, May 2012.
12. S. Mumtaz, M. F. Randrianandrasana, and I. T. Nabney. Mixed-type data visualisation and simultaneous feature saliency estimation using generalised generative topographic mapping. Technical report, Systems Analytics and Research Institute (SARI), Aston University, Birmingham, United Kingdom, 2015.
13. M. F. Randrianandrasana, S. Mumtaz, and I. T. Nabney. Visualisation of heterogeneous data with the generalised generative topographic mapping. In *Proceedings of the Tenth International Conference on Information Visualization Theory and Application*, pages 233–238, 2015.
14. C. M. V. Silvestre, M. M. G. Cardoso, and M. A. T. Figueiredo. Clustering and selecting categorical features. In *EPIA*, volume 8154 of *Lecture Notes in Computer Science*, pages 331–342. Springer, 2013.
15. A. Vellido. Preliminary theoretical results on a feature relevance determination method for generative topographic mapping. Technical report, Universitat Politecnica de Catalunya (UPC)LSI-05-13-R, Barcelona, Spain, 2005.
16. A. Vellido. Assessment of an unsupervised feature selection method for generative topographic mapping. In *Proceedings of the 16th International Conference on Artificial Neural Networks - Volume Part II*, ICANN'06, pages 361–370, Berlin, Heidelberg, 2006. Springer-Verlag.
17. A. Vellido, P. J. G. Lisboa, and D. Vicente. Robust analysis of MRS brain tumour data using *t*-GTM. *Neurocomputing*, 69(7-9):754–768, 2006.
18. X. Wang and A. Kabán. Finding uninformative features in binary data. In *Intelligent Data Engineering and Automated Learning - IDEAL 2005*, volume 3578 of *Lecture Notes in Computer Science*, pages 40–47. Springer Berlin Heidelberg, 2005.