Visualisation of heterogeneous data with the Generalised Generative Topographic Mapping

Michel F. Randrianandrasana, Shahzad Mumtaz and Ian T. Nabney
Nonlinearity and Complexity Research Group, Aston University, Birmingham B4 7ET, UK
{*randrimf, mumtazs, i.t.nabney*}*@aston.ac.uk*

Keywords:
Data visualisation, GTM, LTM, heterogeneous and missing data

Abstract:
Heterogeneous and incomplete datasets are common in many real-world visualisation applications. The probabilistic nature of the Generative Topographic Mapping (GTM), which was originally developed for complete continuous data, can be extended to model heterogeneous (i.e. containing both continuous and discrete values) and missing data. This paper describes and assesses the resulting model on both synthetic and real-world heterogeneous data with missing values.

1 INTRODUCTION

Type-specific data analysis has been well studied in machine learning[1]. In the last couple of decades, the need to analyse mixed-type data has received some attention from the machine learning community because of the fact that real-world processes often generate data of mixed-type. An example of such mixed-type data could be a hospital's patient database where typical fields include age (continuous), gender (binary), test results (binary or continuous), height (continuous) etc. In practice a number of ad-hoc methods are used to analyse mixed-type data. For instance, if there is a mixture of continuous and discrete variables, then either all the discrete variables are converted to some numerical scoring equivalent or, on the other hand, all the continuous variables are discretised. Alternatively, both types of variables are analysed separately and then the results are combined using some criteria. According to (Krzanowski, 1983), "All these options involve some element of subjectivity, with possible loss of information, and do not appear very satisfactory in general". The ideal general solution for analysing such heterogeneous data is to specify a model that builds a joint distribution with an appropriate noise model for each type of feature (for example, a Bernoulli distribution for binary features, a multinomial distribution for multi-category features and a Gaussian distribution for continuous features) and then fit the model to data (de Leon and Chough, 2013).

A multivariate distribution that can model random variables of different types is not available. However, one possible way of jointly modelling discrete and continuous features is using a latent variable approach to model the correlation between features of different types. For example, a dataset consisting of continuous, binary and multi-category features can be modelled using a conditional distribution that is a product of Gaussian, Bernoulli and multinomial distributions. This approach has been previously discussed as a possible extension for GTM (Bishop and Svensen, 1998; Bishop et al., 1998) and PCA (Tipping, 1999) models. This idea was implemented in (Yu and Tresp, 2004) to visualise a mixture of continuous and binary data on a single continuous latent space by extending probabilistic principal component analysis (PPCA) and was called generalised PPCA (GPPCA). GPPCA is a linear probabilistic model and uses a variational Expectation-Maximisation (EM) algorithm for parameter estimation. There are other latent variable models for mixed-type datasets but to the best of our knowledge most of these are linear models (Moustaki, 1996; Sammel et al., 1997; Dunson, 2000; Teixeira-Pinto and Normand, 2009) and they either use numerical

[1]http://letdataspeak.blogspot.co.uk/2012/07/mixed-type-data-analysis-i-overview.html

integration or a sampling approach to handle the intractable integration for fitting a latent variable model of this type. It is important to mention that there is not much work reported in the literature for analysing mixed-type data using a latent variable formalism (de Leon and Chough, 2013). As a generalisation of GTM, a latent trait model (LTM) to handle discrete data was proposed in (Kabán and Girolami, 2001): the model used the exponential family of distributions. In this paper we describe and assess a probabilistic non-linear latent variable model to visualise a mixed-type dataset on a single continuous latent space. We shall refer to this model as a generalised GTM (GGTM).

The treatment of incomplete data for the standard GTM has been explored in (Sun et al., 2002) using an EM approach which estimates the parameters of the mixing components of the GTM and missing values at the same time. The same approach is used in this paper to visualise mixed-type data containing missing values with GGTM.

2 Visualisation of heterogeneous data with GGTM

The main goal of a latent variable model is to find a low-dimensional manifold, $\mathcal{H}$, with M dimensions (usually $M = 2$) for the distribution $p(\mathbf{x})$ of high-dimensional data space, $\mathcal{D}$, with D dimensions. Latent variable models have been developed to handle a dataset where all the features are of the same type.

Suppose that the D-dimensional data space is defined by $|\mathcal{R}|$ continuous, $|\mathcal{B}|$ binary and $|\mathcal{C}|$ multi-categorical features respectively. The link functions for continuous, binary and multi-category features are defined in equations (1), (2) and (3) respectively

$$\boldsymbol{\mu}^{\mathcal{R}} = \boldsymbol{\Phi}(\mathbf{z})\mathbf{W}^{\mathcal{R}}. \tag{1}$$

$$\begin{aligned}\boldsymbol{\mu}^{\mathcal{B}} &= g^{\mathcal{B}}(\boldsymbol{\Phi}(\mathbf{z})\mathbf{W}^{\mathcal{B}}) \\ &= \frac{\exp(\boldsymbol{\Phi}(\mathbf{z})\mathbf{W}^{\mathcal{B}})}{1+\exp(\boldsymbol{\Phi}(\mathbf{z})\mathbf{W}^{\mathcal{B}})}.\end{aligned} \tag{2}$$

$$\begin{aligned}\mu^{\mathcal{C}}_{s_d} &= g^{\mathcal{C}}(\boldsymbol{\Phi}(\mathbf{z})\mathbf{w}^{\mathcal{C}}_{s_d}) \\ &= \frac{\exp(\boldsymbol{\Phi}(\mathbf{z})\mathbf{w}^{\mathcal{C}}_{s_d})}{\sum_{s'_d=1}^{S_d}\exp(\boldsymbol{\Phi}(\mathbf{z})\mathbf{w}_{s'_d})}.\end{aligned} \tag{3}$$

We write each observation vector, $\mathbf{x}_n$ in terms of sub-vectors $\mathbf{x}^{\mathcal{R}}_n$, $\mathbf{x}^{\mathcal{B}}_n$ and $\mathbf{x}^{\mathcal{C}}_n$ for continuous, binary and multi-category features respectively. The likelihood of each type of feature is given by

$$\begin{aligned}p(\mathbf{x}^{\mathcal{R}}_{\mathbf{n}}|\mathbf{z},\mathbf{W}^{\mathcal{R}},\beta) &= p(\mathbf{x}^{\mathcal{R}}_{\mathbf{n}}|\boldsymbol{\mu}^{\mathcal{R}},\beta) \\ &= \left(\frac{\beta}{2\pi}\right)^{\frac{|\mathcal{R}|}{2}}\exp\left(-\frac{\beta}{2}||\boldsymbol{\mu}^{\boldsymbol{\mathcal{R}}}-\mathbf{x}^{\mathcal{R}}_n||^2\right).\end{aligned} \tag{4}$$

$$\begin{aligned}p(\mathbf{x}^{\mathcal{B}}_{\mathbf{n}}|\mathbf{z},\mathbf{W}^{\mathcal{B}}) &= p(\mathbf{x}^{\mathcal{B}}_{\mathbf{n}}|\boldsymbol{\mu}^{\mathcal{B}}) \\ &= \prod_{d=1}^{|\mathcal{B}|}\left(\mu^{\mathcal{B}}_d\right)^{x^{\mathcal{B}}_{nd}}\left(1-\mu^{\mathcal{B}}_d\right)^{(1-x^{\mathcal{B}}_{nd})}.\end{aligned} \tag{5}$$

$$\begin{aligned}p(\mathbf{x}^{\mathcal{C}}_{\mathbf{n}}|\mathbf{z},\mathbf{W}^{\mathcal{C}}) &= p(\mathbf{x}^{\mathcal{C}}_{\mathbf{n}}|\boldsymbol{\mu}^{\mathcal{C}}) \\ &= \prod_{d=1}^{|\mathcal{C}|}\prod_{s_d=1}^{S_d}\left(\mu^{\mathcal{C}}_{s_d}\right)^{x^{\mathcal{C}}_{ns_d}}.\end{aligned} \tag{6}$$

Then we compute the product of the likelihoods for the Gaussian (equation (4)), Bernoulli (equation (5)) and multinomial (equation (6)) distributions, and find the distribution of $\mathbf{x}$ by integrating over the latent variables, $\mathbf{z}$,

$$\begin{aligned}p(\mathbf{x}|\Omega) = \int & p(\mathbf{x}^{\mathcal{R}}_{\mathbf{n}}|\mathbf{z},\mathbf{W}^{\mathcal{R}},\beta) \\ & p(\mathbf{x}^{\mathcal{B}}_{\mathbf{n}}|\mathbf{z},\mathbf{W}^{\mathcal{B}})p(\mathbf{x}^{\mathcal{C}}_{\mathbf{n}}|\mathbf{z},\mathbf{W}^{\mathcal{C}})p(\mathbf{z})\,d\mathbf{z},\end{aligned} \tag{7}$$

where $\Omega = \{\mathbf{W}^{\mathcal{R}},\beta,\mathbf{W}^{\mathcal{B}},\mathbf{W}^{\mathcal{C}}\}$ contains all the model parameters. We use as prior distribution, $p(\mathbf{z})$, a sum of delta functions as for the standard GTM and LTM

$$p(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K}\delta(\mathbf{z}-\mathbf{z}_k). \tag{8}$$

The data distribution can now be derived from equations (7) and (8), where we use the same mixing co-efficient for all components (i.e. $\pi_k = \frac{1}{K}$),

$$p(\mathbf{x}|\Omega) = \sum_{k=1}^{K}\pi_k p(\mathbf{x}|\mathbf{z}_k,\Omega). \tag{9}$$

The log-likelihood of the complete data takes the form

$$\mathcal{L}(\Omega) = \sum_{n=1}^{N}\ln\sum_{k=1}^{K}\pi_k p(\mathbf{x}_n|\mathbf{z}_k,\Omega). \tag{10}$$

The choice of noise model is related to the corresponding type of data and also the link function mapping from latent to data space (Kabán and Girolami, 2001). The exponential family of distributions is used here to model mixed-type data under the latent variable framework. From here onward to simplify the notation, we use $\mathbf{x}^{\mathcal{M}}$, where $\mathcal{M}$ can represent either $\mathcal{R}$, $\mathcal{B}$ or $\mathcal{C}$, to indicate the type of feature for a data point $\mathbf{x}$.

2.1 An expectation maximization (EM) algorithm for GGTM

Our proposed model is based on a mixture of distributions where each component is a product of Gaussian, Bernoulli and/or multinomial distributions. The parameters of the mixture model can be determined using an EM algorithm: in the E-step, we use the current parameter set, Ω, to compute the posterior probabilities (responsibilities) using Bayes' theorem,

$$r_{kn} = p(\mathbf{z}_k|\mathbf{x}_n, \mathbf{W}) = \frac{\pi_k p(\mathbf{x}_n|\mathbf{z}_k, \mathbf{W})}{\sum_{k'=1}^{K} \pi_{k'} p(\mathbf{x}_n|\mathbf{z}_{k'}, \mathbf{W})}, \tag{11}$$

where

$$\begin{aligned} p(\mathbf{x}_n|\mathbf{z}_k, \mathbf{W}) =& p(\mathbf{x}_\mathbf{n}^\mathcal{R}|\mathbf{z}_k, \mathbf{W}^\mathcal{R}, \beta) \\ & p(\mathbf{x}_\mathbf{n}^\mathcal{B}|\mathbf{z}_k, \mathbf{W}^\mathcal{B}) p(\mathbf{x}_\mathbf{n}^\mathcal{C}|\mathbf{z}_k, \mathbf{W}^\mathcal{C}). \end{aligned} \tag{12}$$

We use the maximization of the relative likelihood (Bishop, 1995), which does not require the computation of the log of a sum. The relative likelihood between the old and new set of parameters can be calculated as

$$\begin{aligned} Q &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{kn} \log\left\{p(\mathbf{x}_n|\mathbf{z}_k, \mathbf{W}) p(\mathbf{z}_k)\right\} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{kn} \left\{ \begin{array}{l} \left\{\mathbf{x}_n^\mathcal{R}\theta_k^\mathcal{R} - \mathcal{G}\left(\theta_k^\mathcal{R}\right) + \log(p_0(\mathbf{x}_n^\mathcal{R}))\right\} \\ + \left\{\mathbf{x}_n^\mathcal{B}\theta_k^\mathcal{B} - \mathcal{G}\left(\theta_k^\mathcal{B}\right) + \log(p_0(\mathbf{x}_n^\mathcal{B}))\right\} \\ + \left\{\mathbf{x}_n^\mathcal{C}\theta_k^\mathcal{C} - \mathcal{G}\left(\theta_k^\mathcal{C}\right) + \log(p_0(\mathbf{x}_n^\mathcal{C}))\right\} \\ + \left\{\log(p(\mathbf{z}_k))\right\} \end{array} \right\} \end{aligned} \tag{13}$$

where $\theta_k^\mathcal{M} = \mathbf{\Phi}(\mathbf{z}_k)\mathbf{W}^\mathcal{M}$. In the M-step we maximize the function Q with respect to each type of weight sub-matrix $\mathbf{W}^\mathcal{M}$ as

$$\frac{\partial Q}{\partial \mathbf{W}^\mathcal{M}} = \mathbf{\Phi}^T\left[\mathbf{R}\mathbf{X}^\mathcal{M} - \mathbf{E}g(\mathbf{\Phi}\mathbf{W}^\mathcal{M})\right], \tag{14}$$

where $\mathbf{\Phi}$ is a $K \times L$ matrix, $\mathbf{R}$ is a $K \times N$ matrix calculated using equation (11), $\mathbf{X}^\mathcal{M}$ is an $N \times |\mathcal{M}|$ data sub-matrix and the diagonal matrix $\mathbf{E}$ contains the values

$$e_{kk} = \sum_{n=1}^{N} r_{kn}. \tag{15}$$

In the case of an isotropic Gaussian with unit variance, the link function $g(.)$ is the identity and by setting the derivative to zero we obtain, as in the standard GTM (Bishop and Svensen, 1998),

$$\widehat{W^\mathcal{R}} = (\mathbf{\Phi}^T E \mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{R}\mathbf{X}^\mathcal{R}. \tag{16}$$

For other link functions, a Generalised EM (GEM) (McLachlan and Krishnan, 1997) algorithm is used because convergence to the local maximum is guaranteed without maximizing the relative likelihood (Kabán and Girolami, 2001). A simple gradient-based update can be obtained for $\mathbf{W}^\mathcal{M}$ from Equation (14)

$$\Delta\mathbf{W}^\mathcal{M} \propto \mathbf{\Phi}^T\left[\mathbf{R}\mathbf{X}^\mathcal{M} - \mathbf{E}g(\mathbf{\Phi}\mathbf{W}^\mathcal{M})\right], \tag{17}$$

where this can be used as an inner loop in the M-step. The correlations between the dimensions of $\boldsymbol{\phi}_l$ responsible for preserving the neighbourhood are required for a topographic organisation given that the natural parameter $\theta^\mathcal{M}$ is being updated under the gradient update of the weight matrix $\mathbf{W}^\mathcal{M}$ (Kabán and Girolami, 2001):

$$\widehat{\theta_k^\mathcal{M}} = \boldsymbol{\phi}_k\mathbf{W}^\mathcal{M} + \eta\sum_{n=1}^{N}\sum_{k'=1}^{K} r_{k'n}\boldsymbol{\phi}_k\boldsymbol{\phi}_{k'}^T(\mathbf{x}^\mathcal{M} - \boldsymbol{\mu}_{k'}^\mathcal{M}).$$

3 Visualisation of missing data with GGTM

The EM framework supports the treatment of missing values in the GGTM model.

3.1 Continuous data

The data points $\mathbf{x}_n$ are written as $(\mathbf{x}_n^o, \mathbf{x}_n^m)$, where m and o represent subvectors and submatrices of the parameters matching the missing and observed components of the data (Ghahramani and Jordan, 1994). Binary indicator variables ζ_{nk} are introduced to specify which component of the mixture model generated the data point. Both the indicator variables ζ_{nk} and the missing inputs $\mathbf{x}_n^m$ are treated as hidden variables in the EM algorithm. The changes made to the EM algorithm for GTM are detailed in (Sun et al., 2002).

3.2 Discrete data

The missing values are inferred in the E-step using the usual posterior means with responsibility r_{kn} computed on the observed data,

$$E[\mathbf{x}_n^m|\mathbf{x}_n^o, \boldsymbol{\mu}^\mathcal{D}] = \sum_{k=1}^{K} r_{kn}\mu_k^\mathcal{D}, \tag{18}$$

where $\mathcal{D} = \{\mathcal{B}$ or $\mathcal{C}\}$. In the M-step, the weight matrix $\widehat{\mathbf{W}^\mathcal{D}}$ is updated first using the complete training data and we then update $\widehat{\boldsymbol{\mu}_k^\mathcal{D}}$ with

$$\widehat{\boldsymbol{\mu}_k^\mathcal{D}} = g^\mathcal{D}(\mathbf{\Phi}(\mathbf{z}_k)\widehat{\mathbf{W}^\mathcal{D}}). \tag{19}$$

4 Visualisation quality evaluation measures

Algorithms based on GTM are examples of unsupervised learning which always give a result when applied to a particular dataset. Thus we cannot tell *a priori* what is the expected or desired outcome. This makes it difficult to judge which method is the best (i.e. tells us the most about a certain dataset). Here we use metrics that measure the degree of local neighbourhood similarity between data space and latent space which can be calculated even if 'ground truth' is not known.

4.1 Trustworthiness, continuity and mean relative rank errors (MRREs)

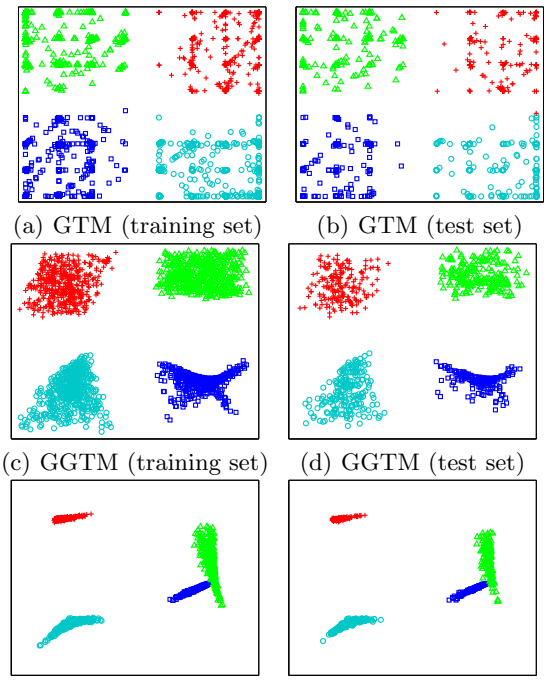
Two well-known visualisation quality measures based on comparing neighbourhoods in the data space $\mathbf{x}$ and projection space $\mathbf{z}$ are *trustworthiness* and *continuity* (Venna and Kaski, 2001). A mapping is said to be *trustworthy* if k-neighbourhood in the visualised space matches that in the data space but if the k-neighbourhood in the data space matches that in the visualised space it maintains *continuity*. The higher the measure the better the visualisation, as this implies that local neighbourhoods are better preserved by the projection. We also use mean relative rank errors with respect to data and latent spaces ($\text{MRRE}^{\mathbf{x}}$ and $\text{MRRE}^{\mathbf{z}}$), which measure the preservation of the rank of the k-nearest neighbours contrary to the trustworthiness and continuity which only consider matches in the k-neighbourhood (Lee and Verleysen, 2008). Note that the lower the MRRE the better the projection quality.

5 Experimental results

The GGTM was evaluated on both complete and missing synthetic and real-world datasets and compared with standard GTM for complete data. The weight matrix $\mathbf{W}$ was initialised using principal component analysis (PCA). For the metrics in Section 4, we computed pair-wise distances using Hamming distances for the binary features and Euclidean distances for the continuous features. For each distance matrix, we divided each column by its standard deviation. All experiments used 10-fold cross-validation. The visualisation quality measures were computed with a range of neighbourhood sizes $(5, 10, 15, 20)$ and the mean of these measures over the different sizes and cross-validation runs was computed.

5.1 Synthetic dataset

The synthetic dataset was generated from an equiprobable mixture of two Gaussians, $\mathcal{N}(\mathbf{m}_k, I)$ (with $k = 1, 2$) with means $\mathbf{m}_1 = \begin{pmatrix} 2.0 \\ 3.5 \\ 3.5 \end{pmatrix}$, and $\mathbf{m}_2 = \begin{pmatrix} 3.5 \\ 4.5 \\ 4.5 \end{pmatrix}$. A dataset with 9-dimensional binary features from four classes was also generated (these classes were not used as inputs to the visualisation). Both continuous and binary data were combined to make a dataset of 12 features with $2,800$ data points. The visualisation results of the complete and missing datasets (10% randomly removed) are shown in Figure 1 and the quality metrics are given in Table 1. We also generated a dataset with two multi-category features with 8 and 16 categories in the first and second features respectively. We appended the multi-category features to the previous 12-dimensional dataset and used a 1-of-S encoding scheme for the



(a) GTM (training set) (b) GTM (test set)

(c) GGTM (training set) (d) GGTM (test set)

(e) GGTM missing (training set) (f) GGTM missing (test set)

Figure 1: GTM and GGTM visualisations of the synthetic 12-dimensional datasets with 3 continuous and 9 binary features.

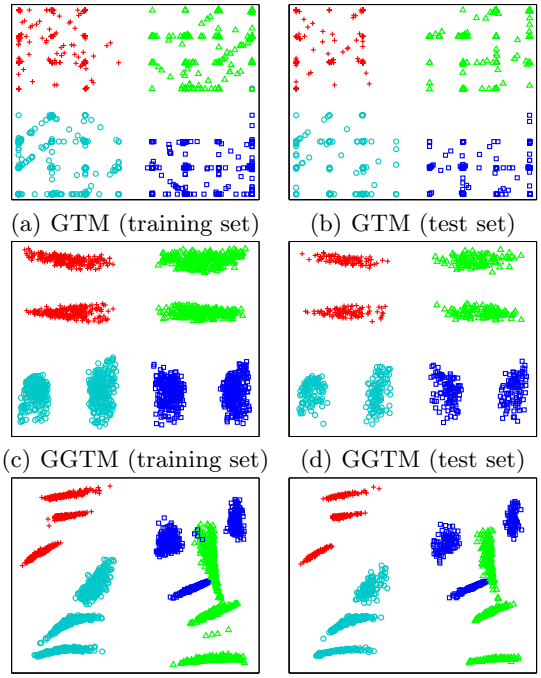
	GTM complete	GGTM complete	GGTM missing
Trustworthiness	0.969 ± 0.003	0.949 ± 0.024	0.947 ± 0.027
Continuity	0.964 ± 0.003	0.970 ± 0.013	0.969 ± 0.014
$\text{MRRE}^{\mathbf{x}}$	0.040 ± 0.000	0.043 ± 0.003	0.042 ± 0.003
$\text{MRRE}^{\mathbf{z}}$	0.004 ± 0.000	0.038 ± 0.002	0.037 ± 0.002

Table 1: GTM and GGTM visualisation quality metrics of the 12-dimensional synthetic datasets. Each figure represents the average over a 10-fold cross-validation with one standard deviation on the test sets.

	GTM complete	GGTM complete	GGTM missing
Trustworthiness	0.962 ± 0.004	0.977 ± 0.009	0.973 ± 0.014
Continuity	0.946 ± 0.008	0.980 ± 0.007	0.976 ± 0.013
$\text{MRRE}^{\mathbf{x}}$	0.045 ± 0.001	0.044 ± 0.001	0.116 ± 0.005
$\text{MRRE}^{\mathbf{z}}$	0.045 ± 0.001	0.041 ± 0.002	0.132 ± 0.005

Table 2: GTM and GGTM visualisation quality metrics of the 14-dimensional synthetic datasets.

multi-category features. Labels were based on the four classes in the binary data.The visualisation results of the 14-dimensional complete and missing datasets are shown in Figure 2 and the corresponding quality metrics are given in Table 2. The proportion of missing values has also been increased to 30%, 50%, 70% and 90% without substantially degrading the visualisation quality measures.



(a) GTM (training set) (b) GTM (test set)

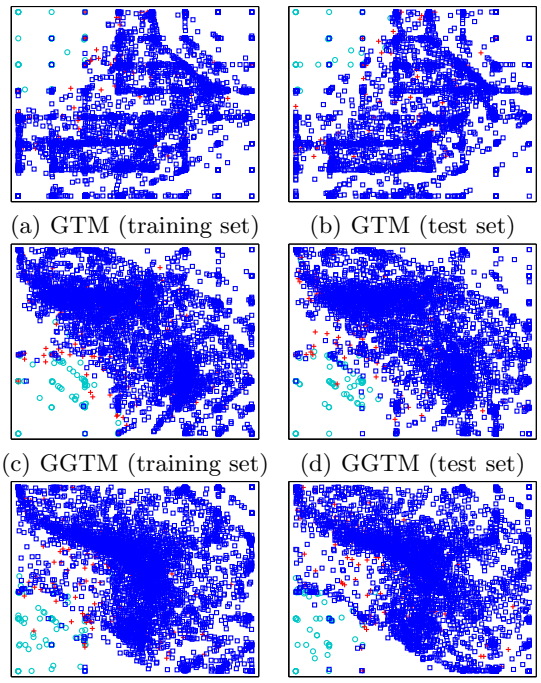
(c) GGTM (training set) (d) GGTM (test set)

(e) GGTM missing (training set) (f) GGTM missing (test set)

Figure 2: GTM and GGTM visualisations of the synthetic 14-dimensional datasets with 3 continuous, 9 binary and 2 multi-category features.

5.2 Hypothyroid dataset

This real-world dataset is publicly available from the UCI data repository (Bache and Lichman, 2013). The dataset consists of two variable types: 15 binary and 6 continuous features. It contains three classes: primary thyroid, compensated thyroid and normal. The dataset was originally divided into a training set of $3,772$ data points (93 with primary hypothyroid, 191 with compensated hypothyroid and 3488 normal) and a test set of $3,428$ data points (73 with primary hypothyroid, 177 with compensated hypothyroid and 3178 normal). These training and test sets have been merged prior to running a 10-fold cross-validation. The visualisation results of the complete and missing datasets are shown in Figure 3 and the quality metrics are given in Table 3.



(a) GTM (training set) (b) GTM (test set)

(c) GGTM (training set) (d) GGTM (test set)

(e) GGTM missing (training set) (f) GGTM missing (test set)

Figure 3: GTM and GGTM visualisations of the thyroid disease datasets. The cyan circles, red plus sign and blue squares represent primary hypothyroid, compensated hypothyroid and normal respectively.

	GTM complete	GGTM complete	GGTM missing
Trustworthiness	0.718 ± 0.022	0.718 ± 0.015	0.716 ± 0.014
Continuity	0.804 ± 0.017	0.843 ± 0.014	0.835 ± 0.007
$\text{MRRE}^{\mathbf{x}}$	0.018 ± 0.000	0.019 ± 0.000	0.019 ± 0.000
$\text{MRRE}^{\mathbf{z}}$	0.016 ± 0.000	0.016 ± 0.000	0.016 ± 0.000

Table 3: GTM and GGTM visualisation quality metrics of the hypothyroid disease datasets.

6 CONCLUSIONS

A generalisation of the GTM to heterogeneous and missing data has been described and assessed in this paper. This involves modelling the continuous and discrete data with Gaussian and Bernoulli/multinomial distributions respectively. These extensions have been suggested in (Bishop et al., 1998) but this is the first time the mathematical details have been worked out and an implementation written and evaluated.

Visualisation results for synthetic data using the GGTM have shown more compact clusters for each class compared to the standard GTM whereas for the real dataset no significant difference was observed. For synthetic datasets with missing values, GGTM visualisations have greater compactness for each class. In terms of visualisation quality evaluation metrics, we observed that for a mix of continuous and binary data, the trustworthiness and $\text{MRRE}^{\mathbf{x}}$ are slightly better for standard GTM compared to GGTM whereas the continuity and $\text{MRRE}^{\mathbf{z}}$ were better for GGTM compared to standard GTM. However, for a mix of continuous, binary and multi-category features, all the quality evaluation measures were better for GGTM compared to the standard GTM. Missing values have caused limited deterioration in results compared to the complete data case.

REFERENCES

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Bishop, C. M. and Svensen, M. (1998). GTM: The generative topographic mapping. *Neural Compuatation*, 10(1):215–234.

Bishop, C. M., Svensen, M., and Williams, C. K. I. (1998). Developments of the generative topographic mapping. *Neurocomputing*, 21(1):203–224.

de Leon, A. R. and Chough, K. C. (2013). *Analysis of Mixed Data: Methods & Applications*. Taylor & Fracis Group. Chapman and Hall/CRC.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(2):355–366.

Ghahramani, Z. and Jordan, M. I. (1994). Learning from incomplete data. Technical Report AIM-1509.

Kabán, A. and Girolami, M. (2001). A combined latent class and trait model for the analysis and visualization of discrete data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(8):859–872.

Krzanowski, W. J. (1983). Distance between populations using mixed continuous and categorical variables. *Biometrika*, 70(1):235–243.

Lee, J. A. and Verleysen, M. (2008). Rank-based quality assessment of nonlinear dimensionality reduction. In *ESANN*, pages 49–54.

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley, New York.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, 49(2):313–334.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):667–678.

Sun, Y., Tino, P., and Nabney, I. (2002). Visualisation of incomplete data using class information constraints. In Winkler, J. and Niranjan, M., editors, *Uncertainty in Geometric Computations*, volume 704 of *The Springer International Series in Engineering and Computer Science*, pages 165–173. Springer US.

Teixeira-Pinto, A. and Normand, S. T. (2009). Correlated bivariate continuous and binary outcomes: issues and applications. *Statistics in Medicine*, 28(13):1753–1773.

Tipping, M. E. (1999). Probabilistic visualisation of high-dimensional binary data. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pages 592–598, Cambridge, MA, USA. MIT Press.

Venna, J. and Kaski, S. (2001). Neighborhood preservation in nonlinear projection methods: an experimental study. In *Proceedings of the International Conference on Artificial Neural Networks*, ICANN '01, pages 485–491, London, UK. Springer-Verlag.

Yu, K. and Tresp, V. (2004). Heterogenous data fusion via a probabilistic latent-variable model. In Müller-Schloer, C., Ungerer, T., and Bauer, B., editors, *ARCS*, volume 2981 of *Lecture Notes in Computer Science*, pages 20–30. Springer.