Visualisation of heterogeneous data with the Generalised Generative Topographic Mapping

First Author Name\textsuperscript{1}, Second Author Name\textsuperscript{1} and Third Author Name\textsuperscript{2}

\textsuperscript{1}Institute of Problem Solving, XYZ University, My Street, MyTown, MyCountry
\textsuperscript{2}Department of Computing, Main University, MySecondTown, MyCountry

\{f\_author, s\_author\}@ips.xyz.edu, t\_author@dc.mu.edu

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

Abstract:
Heterogeneous and incomplete datasets are common in many real-world applications. The probabilistic nature of the Generative Topographic Mapping (GTM), which only handles complete continuous data originally, offers the ability to extend it to also visualise mixed-type and missing data as suggested in (Bishop et al., 1998a). This paper describes this generalisation of GTM 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\textsuperscript{1}. In the last couple of decades, the need to analyse mixed-type data gained a lot of attention from machine learning experts 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 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 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 the assumption that an appropriate noise distribution exists for each type of feature (for example a Bernoulli distribution for modelling binary features, a multinomial distribution for modelling multi-category features and a Gaussian distribution for modelling continuous features) and then fit the model to data where the parameter estimates are used to draw inferences (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 understand the correlation between features of different type. For example, a dataset consisting of continuous, binary and multi-category features can be modelled in a latent variable model using the conditional distribution as a product of Gaussian, Bernoulli and multinomial distributions in the data space. Such a latent variable model follows a conditional independence criterion for each type of the observed variable to find a correlation between observed features and unobserved latent features in a unified framework. This idea of handling a mixed-type dataset under the latent

\textsuperscript{1}http://letdataspeak.blogspot.co.uk/2012/07/mixed-type-data-analysis-i-overview.html
variable formalism has been previously discussed as a possible extension for GTM-like (Bishop and Svensen, 1998; Bishop et al., 1998b) and PCA-like (Tipping, 1999) models. About a decade ago, 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 a generalised PPCA (GPPCA). It was influenced by the generalisation of PPCA to binary data (Tipping, 1999). 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) and they either use numerical integration or a sampling approach to handle the intractable integration for fitting a latent variable model of this type. So far a GTM-like non-linear model has not been implemented to visualising mixed-type data.

A standard GTM is a probabilistic reformulation of the self-organising map (SOM) (Bishop et al., 1997). As a generalisation of GTM, a latent variable model was proposed in (Kabán and Girolami, 2001) with a goal of handling different types of data by considering the exponential family of distributions. However the main focus was to visualise discrete datasets it is called a latent trait model (LTM) and all these variants were only able to model one type of features at a time in the latent variable framework. 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 by combining a GTM a LTM in a principled way under a unified framework of conditional independence criterion. We shall refer to this model as a generalised GTM (GGTM) for which we consider different noise distributions depending on the type of features.

The treatment of incomplete data for the standard GTM has been explored in (Sun et al., 2002). This is done through an EM approach which estimates the parameters of the mixing components of the GTM and missing values at the same time. The idea is based on a similar work to learn from incomplete data for clustering and classification purposes using mixture models (Ghahramani and Jordan, 1994). The same approach will be used in this paper in order to visualise mixed-type data having 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 $p(x)$ of high-dimensional data space, $\mathcal{D}$, with $D$-dimensions. Latent variable models like GTM or LTM have been developed to handle a dataset where all the features are of the same type (either continuous, binary or multi-categorical).

We consider the task of modelling a $D$-dimensional data space defined by $|\mathcal{R}|$ continuous, $|\mathcal{B}|$ binary and/or $|\mathcal{C}|$ multi-categorical number of features respectively. The link functions for mapping between the latent space and the data space for continuous, binary and multi-category features are defined in equations (1), (2) and (3) respectively

$$m^R = \Phi(z)W^R. \quad (1)$$

$$m^B = g^B(\Phi(z)W^B) = \frac{\exp(\Phi(z)W^B)}{1 + \exp(\Phi(z)W^B)}. \quad (2)$$

$$m^C_{s,d} = g^C(\Phi(z)w^C_{s,d}) = \frac{\exp(\Phi(z)w^C_{s,d})}{\sum_{s,d'=1}^{S_d} \exp(\Phi(z)w^C_{s,d'})}. \quad (3)$$

These link functions map a latent point $z \in \mathcal{H}$ to a point in the data space. We write each observation vector, $x_n$ in terms of sub-vectors $x^R_n$, $x^B_n$ and $x^C_n$ for continuous, binary and multi-category features respectively. In the rest of the paper we use $\mathcal{R}$, $\mathcal{B}$ and $\mathcal{C}$ to denote continuous, binary and categorical features respectively. The likelihood
of each type of features is given by
\[ p(x^R_n | z, W^R, \beta) = p(x^R_n | m^R, \beta) \]
\[ = \left( \frac{\beta}{2\pi} \right)^{|R|/2} \exp \left( -\frac{\beta}{2} |m^R - x^R_n|^2 \right) \]  
\[ p(x^B_n | z, W^B) = p(x^B_n | m^B) \]
\[ = \prod_d \left( m_d^B \right)^{x_n^B} \left( 1 - m_d^B \right)^{1-x_n^B} \]  
\[ p(x^C_n | z, W^C) = p(x^C_n | m^C) \]
\[ = \prod_d \prod_{s_d=1}^{s_d} \left( m_{d,s_d}^C \right)^{z_{n,d}^C} \]  

In order to model mixed-type data under the latent variable formalism, we compute the product of the likelihoods for the Gaussian (equation (4)), Bernoulli (equation (5)) and multinomial (equation (6)) distributions. The distribution of \( x \) is obtained by integrating over the latent variables, \( z \), as
\[ p(x | \Omega) = \int p(x^R_n | z, W^R, \beta) p(x^B_n | z, W^B) p(x^C_n | z, W^C) p(z) \, dz. \]  
(7)

where \( \Omega = (W^R, \beta, W^B, W^C) \) is a set containing all the parameters of the model. We consider the prior distribution, \( p(z) \), as a sum of delta functions similarly to the standard GTM (Bishop and Svensen, 1998) and LTM (Kabán and Girolami, 2001) as
\[ p(z) = \frac{1}{K} \sum_{k=1}^{K} \delta(z - z_k). \]  
(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. \( \pi_k = \frac{1}{K} \)), as
\[ p(x | \Omega) = \sum_{k=1}^{K} \pi_k p(x | z_k, \Omega). \]  
(9)

The log-likelihood of the complete data takes the form
\[ L(\Omega) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k p(x_n | z_k, \Omega). \]  
(10)

2.1 Distributional Assumptions over the Noise Model

The distributional assumption over the noise model is important for the type of data the model can handle and also for the link function required for modelling different types of data (Kabán and Girolami, 2001). It has been argued in (Kabán and Girolami, 2001; Bishop et al., 1998b) that modelling continuous data can be achieved with the assumption that the noise is independent and identically distributed (i.i.d.) Gaussian. An exponential family of distributions is assumed to handle different type of features in the derivation of the LTM algorithm. The same idea is adopted here to model mixed-type data under the latent variable framework. To simplify the notation, we use \( x^M \), where \( M \) can represent either \( R \), \( B \) or \( C \), to indicate the type of feature for a data point \( x \). The functional form of the exponential family of distributions can be defined by
\[ p_M(x^M | \theta^M) = \exp \left\{ x^M \theta^M - \mathcal{G} \left( \theta^M \right) \right\} p_0(x^M). \]  
(11)

In our case, the conditional probability distribution of a data point \( x^M_n \) given the latent point \( z_k \) and the weight matrix \( W^M \) can be defined as,
\[ p_M(x^M_n | z_k, W^M) = \exp \left\{ x^M_n \Phi(z_k) W^M - \mathcal{G} \left( \Phi(z_k) W^M \right) \right\} p_0(x^M_n). \]  
(12)

where \( \mathcal{G}(.) \) is the cumulant function and is defined as
\[ \mathcal{G}(\Phi(z_k) W^M) = \ln \left( \int \exp \left\{ x^M \Phi(z_k) W^M \right\} p_0(x^M_n) \, dx^M_n \right). \]  
(13)

The natural parameter \( \theta^M \) of the exponential family of the distribution is taken to be a linear mixing of the latent vectors with respect to the weight matrix \( W^M \),
\[ \theta^M_k = \Phi(z_k) W^M, \]  
(14)

where \( W^M \) is the weight matrix of size \( L \times |M| \). The gradient of the cumulant function with respect to the natural parameter (i.e. \( \Phi(z_k) W^M \)) is
\[ m^M_k = g^M(\Phi(z_k) W^M) = \nabla_{\theta^M} \mathcal{G}(\Phi(z_k) W^M), \]  
(15)

where \( \nabla \) represents the gradient operation and the function \( g(.) \) is the link function (Kabán and
\( \textbf{m}_k^M \) represents the mean of the \( k \)th class under the distribution \( p(x^M|z_k) \) (Barndorff-Nielsen, 1978)

\[
\textbf{m}_k^M = E\{x^M|z_k\},
\]

where \( x^M \) represents the observation vectors and \( E\{.\} \) represent the expectation operator. The expected value of the Hessian of the cumulant function with respect to the natural parameters \( \theta_k^M = \Phi(z_k)W^M \) represents the covariance matrix of the \( k \)th class for the distribution \( p(x^M|z_k) \) (Barndorff-Nielsen, 1978). Such a matrix is represented as

\[
G_k^M = \nabla_{\theta_k^M} \textbf{m}_k^M = Var\{x^M|z_k\},
\]

2.2 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 which can be formulated as follows: in the E-step, we use the current parameter set, \( \Omega \), to compute the posterior probabilities (responsibilities) using Bayes’ theorem as,

\[
r_{kn} = p(z_k|x_n, W) = \frac{\pi_k p(x_n|z_k, W)}{\sum_{k'} \pi_{k'} p(x_n|z_{k'}, W)},
\]

where

\[
p(x_n|z_k, W) = p(x_n^R|z_k, W^R, \beta)p(x_n^C|z_k, W^C).
\]

In the context of EM algorithm methodology, we can use the maximization of the relative likelihood (McLachlan and Krishnan, 1997; Kabán and Girolami, 2001; Bishop, 1995; Dempster et al., 1977), instead of maximizing the log-likelihood, 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

\[
Q = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{kn} \log \{ p(x_n|z_k, W)p(z_k) \}
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{kn} \left\{ \begin{array}{l}
\{ x_n^R \theta_k^R - G(\theta_k^R) + \log(p_0(x_n^R)) \} \\
+ \{ x_n^C \theta_k^C - G(\theta_k^C) + \log(p_0(x_n^C)) \} \\
+ \{ \log(p(z_k)) \}
\end{array} \right\}
\]

where \( \theta_k^M = \Phi(z_k)W^M \). In the M-step we can maximize the function \( Q \) with respect to each type of weight sub-matrix \( W^M \) as

\[
\frac{\partial Q}{\partial W^M} = \Phi^T [RX^M - E_g(\Phi W^M)],
\]

where \( \Phi \) is a \( K \times L \) matrix, \( R \) is a \( K \times N \) matrix calculated using equation (18), \( X^M \) is an \( N \times |M| \) data sub-matrix and the diagonal matrix \( E \) contains the values

\[
e_{kk} = \sum_{n=1}^{N} r_{kn}.
\]

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

\[
\hat{W}^R = (\Phi^T E \Phi)^{-1} \Phi^T RX^R,
\]

In general a non-linear optimisation approach, such as iterative least-square, may be required. However, a Generalised EM (GEM) (McLachlan and Krishnan, 1997) algorithm is a more appropriate choice because of the fact that the 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 \( W^M \) from Equation (21) as

\[
\Delta W^M \propto \Phi^T \left[ RX^M - E_g(\Phi W^M) \right],
\]

where this can be used as an inner loop in the \( M \)-step. We in this paper employed a gradient inner loop \( M \)-step and the correlations between dimensions of \( \phi_i \) responsible for preserving the neighbourhood are required for a topographic organization. We explain here how the natural parameter \( \theta^M \) modifies under the gradient update.
of the weight matrix $W^M$ as,
\[
\theta_k^M = \phi_k W^M = \phi_k W^M + n \sum_{n=1}^N \sum_{k'=1}^K r_{kn} \phi_k \phi^T_{k'},
\]
\[
(x^M - m_k^M).
\]
This is analogous to the Self Organising Map (SOM) (Kohonen, 1995) update.

3 Visualisation of missing data with GGTM

The EM algorithm framework allows the treatment of missing values in the GGTM model. This will be described next for continuous and discrete data.

3.1 Continuous data

Missing values can be incorporated into the EM algorithm for the GTM model. The data points $x_n$ are written as $(x_n^o, x_n^m)$, where $m$ and $o$ represent subvectors and submatrices of the parameters matching the missing and observed components of the data (Sun et al., 2002; Ghahramani and Jordan, 1994). Binary indicator variables $z_{nj}$ are introduced to specify which component of the mixture model generated the data point. $z_{nj} = 1$ if and only if $x_n$ is generated by component $j$, otherwise $z_{nj} = 0$. Both the indicator variables $z_{nj}$ and the missing inputs $x_n^m$ are treated as hidden variables in the EM algorithm. The following changes are applied to the EM algorithm for GTM:

\[
\hat{x}_{nj}^m = E(x_n^m | x_n^o = 1, \theta_j) = \beta^{-1} y_j^m + (\Sigma_j^{-1} - \beta^{-1} I)^{-1} \sum_j \beta^{-1} y_j^m - 2 \beta^{-1} y_j^m, \]
\[
\hat{x}_{nj}^o = E(x_n^o | x_n^m = 0, \theta_j) = y_j^o, \]
\[
\hat{x}_{nj}^m = \hat{x}_{nj}^o = (W_{old} \Phi(z_j))^m \text{ and } (y_j^o)^{old} = (W_{old} \Phi(z_j))^o.
\]

Since the covariance matrix is isotropic, $\Sigma_j = \beta^{-1} I$, and the covariance of missing and observed values $\Sigma_j^{mo} = 0$. Consequently,
\[
\hat{x}_{nj}^m = (y_j^m)^{old}, \tag{27}
\]
The missing values are filled in using the posterior means,
\[
E[x_n^m | x_n^o, \theta_j] = \sum_{j=1}^K r_{nj} \hat{x}_{nj}^m, \tag{28}
\]
in the M-step, the new weights are updated to $W_{new}$ as in the case for complete training data. The variance is updated as follows:
\[
\beta^{-1} = \frac{1}{ND} \sum_{n=1}^N \sum_{j=1}^K r_{nj} (||x_n^o - y_j^o||^2 + E[z_{nj} ||x_n^m - y_j^m||^2]), \tag{29}
\]
where
\[
E[z_{nj} ||x_n^m - y_j^m||^2] = M(\beta^{-1})^{old} + (\hat{x}_{nj}^m)^T (\hat{x}_{nj}^m - 2(y_j^m)^{old} + (y_j^m)^{old}). \tag{30}
\]

3.2 Discrete data

Similarly, in the E-step, the responsibilities $r_{nj}$ are computed using the observed data in the discrete case. The missing values are filled in using the usual posterior means,
\[
E[x_n^m | x_n^o, \theta_j] = \sum_{j=1}^K r_{nj} \hat{x}_{nj}^m. \tag{31}
\]

In the M-step, we have
\[
\hat{x}_{nj}^m = \frac{\sum_{n=1}^N r_{nj} x_n}{\sum_{n=1}^N r_{nj}}. \tag{32}
\]

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). However in the simple
case of artificial data one can use prior knowledge about the structure of the data in the original space to quantify the error on the projection. This is exploited in some of the following measures. Other metrics measure the degree of local neighbourhood similarity between data space and latent space.

### 4.1 Trustworthiness and Continuity

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.

To measure the trustworthiness, we consider the rank \( r_{i,j}^x \) of the \( j \)th data point from the corresponding \( i \)th data point with respect to the distance measure in the high-dimensional data space \( \mathbf{t} \), and \( P_k(i) \) represents the data points in the \( k \)-nearest neighbourhood of the \( i \)th data point in the latent space \( \mathbf{z} \) but not in the \( k \)-nearest neighbourhood of the data space \( \mathbf{t} \). The trustworthiness with \( k \) neighbours can be calculated as

\[
1 - \frac{2}{\gamma_k} \sum_{i=1}^{N} \sum_{j \in P_k(i)} (r_{i,j}^x - k).
\]

(33)

To measure the continuity, we consider \( r_{i,j}^z \) as the rank of the \( j \)th data point from the \( i \)th data point with respect to the distance measure in the visualisation space \( \mathbf{z} \) and \( Q_k(i) \) to be the set of data points in the \( k \)-nearest neighbourhood of the \( i \)th data point in the data space \( \mathbf{x} \) but not in the visualisation space \( \mathbf{z} \). The continuity with \( k \) neighbours can be calculated as

\[
1 - \frac{2}{\gamma_k} \sum_{i=1}^{N} \sum_{j \in Q_k(i)} (r_{i,j}^z - k).
\]

(34)

For both trustworthiness and continuity we take the normalising factor \( \gamma_k \) as

\[
\gamma_k = \begin{cases} 
Nk(2N - 3k - 1) & \text{if } k < N/2, \\
N(N - k)(N - k - 1) & \text{if } k \geq N/2,
\end{cases}
\]

(35)

where the \( \gamma_k \) ensures that the trustworthiness and continuity values lie between 0 and 1. The higher the measure the better the visualisation, as this implies that local neighbourhoods are better preserved by the projection.

### 4.2 Mean relative rank errors (MRREs)

We also use other quantitative quality measures such as mean relative rank errors with respect to data and latent spaces (MRRE\(^x \) and MRRE\(^z \)), which preserve 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). The MRREs are defined with respect to the data space,

\[
\text{MRRE}^x(k) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in N^x_k(i)} \frac{|r_{i,j}^x - \tau_{i,j}^x|}{\tau_{i,j}^x},
\]

and with respect to latent space as

\[
\text{MRRE}^z(k) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in N^z_k(i)} \frac{|r_{i,j}^z - \tau_{i,j}^z|}{\tau_{i,j}^z},
\]

where the normalisation factor for both types of MRREs is \( \tau_k = N \sum_{k' = 1}^{k} \frac{[N - 2k' + 1]}{k'} \) and \( N^x_k(i) \) and \( N^z_k(i) \) represent the set of \( k \)-nearest neighbours of data point \( i \) in the data space and latent space respectively. However, in this case, the lower the MRRE the better the projection quality.

### 5 Experimental results

The GGTM is evaluated on both complete and missing synthetic and real-world datasets. The weight matrix \( \mathbf{W} \) is initialised using principal component analysis (PCA). We use the quality measures described in Section 4 in order to evaluate the quality of the visualisation results. The data space is assumed to be of mixed-type, therefore we compute pair-wise distances using hamming distances for the binary features and Euclidean distances for the continuous features. For each distance matrix, we divide each column by its standard deviation in order to make both distance matrices on an equivalent scale before adding them up. A 10-fold cross-validation is applied on the different data. The visualisation quality measures are computed with different number of neighbours (5, 10, 15, 20) and the mean of these measures over the different neighbours and cross-validation runs is computed.
Table 1: GGTM visualisation quality metrics of the 12-dimensional (3 continuous and 9 binary features) mixed-type complete and missing synthetic datasets. Each figure represents the average over a 10-fold cross-validation with one standard deviation on the test sets.

<table>
<thead>
<tr>
<th></th>
<th>GGTM complete</th>
<th>GGTM missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trustworthiness</td>
<td>0.921 ± 0.033</td>
<td>0.946 ± 0.005</td>
</tr>
<tr>
<td>Continuity</td>
<td>0.947 ± 0.021</td>
<td>0.967 ± 0.002</td>
</tr>
<tr>
<td>MRRE$^x$</td>
<td>0.045 ± 0.002</td>
<td>0.047 ± 0.001</td>
</tr>
<tr>
<td>MRRE$^z$</td>
<td>0.040 ± 0.001</td>
<td>0.041 ± 0.001</td>
</tr>
</tbody>
</table>

Table 2: GGTM visualisation quality metrics of the 14-dimensional (3 continuous, 9 binary and 2 multi-category features) mixed-type complete and missing synthetic datasets. Each figure represents the average over a 10-fold cross-validation with one standard deviation on the test sets.

<table>
<thead>
<tr>
<th></th>
<th>GGTM complete</th>
<th>GGTM missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trustworthiness</td>
<td>0.948 ± 0.023</td>
<td>0.974 ± 0.008</td>
</tr>
<tr>
<td>Continuity</td>
<td>0.953 ± 0.017</td>
<td>0.981 ± 0.005</td>
</tr>
<tr>
<td>MRRE$^x$</td>
<td>0.043 ± 0.001</td>
<td>0.116 ± 0.004</td>
</tr>
<tr>
<td>MRRE$^z$</td>
<td>0.042 ± 0.000</td>
<td>0.127 ± 0.004</td>
</tr>
</tbody>
</table>

5.1 Synthetic dataset

The synthetic dataset is generated from an equiprobable mixture of two Gaussians, $N(m_k, I)$ (with $k = 1, 2$) with means $m_1 = \left(\frac{2}{3}, \frac{3}{2}\right)$, and $m_2 = \left(\frac{3}{2}, \frac{1}{3}\right)$. A dataset with 9-dimensional binary features having four classes is also generated. The labels we use indicate the four binary classes. Both continuous and binary data are combined to make a dataset of 12 features with 2,800 data points.

The visualisation results of the 12-dimensional mixed-type complete and missing datasets using GGTM are shown in Figure 1 and the visualisation quality metrics are given in Table 1.

Table 2: GGTM visualisation quality metrics of the 14-dimensional (3 continuous, 9 binary and 2 multi-category features) mixed-type complete and missing synthetic datasets. Each figure represents the average over a 10-fold cross-validation with one standard deviation on the test sets.

We also generated a dataset with two multi-category features in which there are 8 and 16 categories in the first and second features respectively. We appended the multi-category features to the previous 12-dimensional dataset yielding a data with 14 features. It is important to note that we use a 1-of-D encoding scheme on the multi-category features ($D$ is the number of categories).

The visualisation results of the 14-dimensional complete and missing datasets using GGTM are shown in Figure 2 and the corresponding visualisation quality metrics are given in Table 2.

Figure 1: GGTM visualisations of the synthetic mixed-type complete and missing datasets with 3 continuous and 9 binary features. 10% of the continuous and binary complete data have been randomly removed to produce the missing data.

(c) GGTM missing (training set)  (d) GGTM missing (testing set)

Figure 2: GGTM visualisations of the synthetic mixed-type complete and missing datasets with 3 continuous, 9 binary and 2 multi-category features. 10% of the continuous, binary and categorical complete data have been randomly removed to produce the missing data. The subfigures are assigned colours from 4 classes defined in the binary features.

The 10% missing values tend to create more compact clusters on both training and test sets of the synthetic data. However, this does not occur.
when only 1% of the data is missing. The proportion of missing values has also been increased to 30%, 50%, 70% and 90% without substantially degrading the visualisation quality measures.

5.2 Hypothyroid dataset

This real-world dataset is publicly available from the UCI data repository (Bache and Lichman, 2013). The dataset consists of mixed types: 15 binary and 6 continuous features. It contains three classes: primary thyroid, compensated thyroid and normal. The dataset is 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 using GGTM are shown in Figure 3 and the visualisation quality metrics are given in Table 3.

![GGTM (training set)](image1)

![GGTM (test set)](image2)

![GGTM missing (training set)](image3)

![GGTM missing (test set)](image4)

Figure 3: GGTM visualisations of the thyroid disease mixed-type complete and missing datasets. 10% of the continuous and binary complete data have been randomly removed to produce the missing data. The cyan circles, red plus sign and blue squares represent primary hypothyroid, compensated hypothyroid and normal respectively.

Unlike the synthetic data, the 10% missing values do not create tight clusters on the training and test sets of the real-world hypothyroid data.

<table>
<thead>
<tr>
<th></th>
<th>GGTM complete</th>
<th>GGTM missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trustworthiness</td>
<td>0.695±0.014</td>
<td>0.687±0.013</td>
</tr>
<tr>
<td>Continuity</td>
<td>0.803±0.008</td>
<td>0.838±0.007</td>
</tr>
<tr>
<td>MRRE_x</td>
<td>0.018±0.000</td>
<td>0.018±0.000</td>
</tr>
<tr>
<td>MRRE_z</td>
<td>0.016±0.000</td>
<td>0.016±0.000</td>
</tr>
</tbody>
</table>

Table 3: GGTM visualisation quality metrics of the hypothyroid disease mixed-type complete and missing datasets. Each figure represents the average over a 10-fold cross-validation with one standard deviation on the test sets.

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 multinomial distributions respectively. The conditional distribution of the combined model is obtained by multiplying each previous distribution. These various GTM extensions have been suggested in (Bishop et al., 1998a).

It has been observed that randomly removing 10% of the synthetic data tend to create more compact clusters on both training and test sets. However this does not occur when only 1% of the data is missing. This phenomenon will be the object of further investigations. The proportion of missing values does not substantially affect the visualisation quality measures. Unlike the synthetic data, the 10% missing values do not create tight clusters on the training and test sets of the real-world hypothyroid data.
REFERENCES