Robust Visual Mining of Data with Error Information

Jianyong Sun[1,2] and Ata Kabán[1] and Somak Raychaudhury[2]

[1] School of Computer Science, The University of Birmingham, Edgbaston, Birmingham, U.K. B15 2TT
[2] School of Physics and Astronomy, The University of Birmingham, Edgbaston, Birmingham, U.K. B15 2TT

Abstract. Recent results on robust density-based clustering have indicated that the uncertainty associated with the actual measurements can be exploited to locate objects that are atypical for a reason unrelated to measurement errors. In this paper, we develop a *constrained* robust mixture model, which, in addition, is able to nonlinearly map such data for visual exploration. Our robust visual mining approach aims to combine statistically sound density-based analysis with visual presentation of the density structure, and to provide visual support for the identification and exploration of 'genuine' peculiar objects of interest that are not due to the measurement errors. In this model, an exact inference is not possible despite the latent space being discretised, and we resort to employing a structured variational EM. We present results on synthetic data as well as a real application, for visualising peculiar quasars from an astrophysical survey, given photometric measurements with errors.

1 Introduction

Providing the users with a qualitative understanding of multivariate data, and automatically detecting atypical objects are among the most important tasks of data mining. Data visualisation techniques aim to capture the underlying structure of the dataset and preserve the structure in the low-dimensional space which can be readily visualised. However, care must be taken with appropriately handling outliers in the data set, to avoid biased parameter estimates [8] [1] and consequently to avoid misleading visual representations [11]. Moreover, outliers may occur for different reasons. Some of them are of interest in certain domain-specific applications, while others are not.

For example, one common reason for the occurrence of outliers or atypical data instances is due to measurement errors. These are inevitable in many areas and may arise from physical limitations of measuring devices and / or measurement conditions. In scientific areas, such as in Astrophysics [3], these errors are recorded, are available, and should be made use of. It is therefore essential to develop methods that are able to incorporate the existing knowledge of measurement errors in order to hunt for the potentially interesting peculiar objects. We have recently proposed a robust mixture modelling method that

takes the errors into account [9]. Here we constrain this model in the spirit of a generative topographic mapping [2], to enable a visual presentation of the data density structure and hence to provide visual support for the identification and exploration of 'genuine' peculiar objects that are not due to the measurement errors.

2 Robust GTM in the Presence of Measurement Errors

GTM The generative topographic mapping (GTM) [2] is a latent variable model for data visualisation. It expresses the probability distribution $p(\mathbf{t})$ of the data $\mathbf{t} \in \Re^D$ in terms of a small (typically two) number of continuous valued and uniformly distributed latent variables $\mathbf{x} \in \Re^d$ where $d \ll D$. To achieve this, GTM supposes that the data lies on a manifold, that is, a data point can be expressed as a mapping from the latent space to the data space as $\mathbf{t} = y(\mathbf{x}; \mathbf{W}) + \epsilon$; where $y(\mathbf{x}; \mathbf{W}) = \mathbf{W}\phi(x)$ is defined as a generalised linear regression model, and the elements $\phi(\mathbf{x})$ consist of M fixed basis functions $\phi_j(\mathbf{x})$, i.e. $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \cdots, \phi_M(\mathbf{x}))$ and $\mathbf{W}$ is the weight matrix. The basis functions may be chosen as RBF-s and ϵ is a Gaussian noise $\mathcal{N}(0, \sigma^{-1}\mathbf{I})$. The distribution $p(\mathbf{t})$ can then be obtained by integrating out the latent variable $\mathbf{x}$: $p(\mathbf{t}) = \int p(\mathbf{x})p(\mathbf{t}|\mathbf{x})d\mathbf{x}$. For tractability reasons, the continuous latent space is discretised in GTM, which results in a latent density $p(\mathbf{x})$ expressed as a finite sum of delta functions, each centered on the nodes of a regular grid in the latent space: $p(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K}\delta(\mathbf{x} - \mathbf{x}_i)$ and $p(\mathbf{t}|\mathbf{x}) = \mathcal{N}(\mathbf{t}|y(\mathbf{x}; \mathbf{W}), \sigma^{-1}\mathbf{I})$. This is convenient, since then the formalism of unconstrained mixture models is directly applicable, while a powerful nonlinear mapping is obtained. A robust extension of GTM, using Student t density components has been developed (t-GTM) in [11], for applications where outliers exist in the data set. However, the existing t-GTM cannot make use of the measurement error information, so it cannot differentiate between outliers that are due to a known cause as opposed to those that are of potential interest. Our purpose here is to address this issue.

The proposed model Each individual measurement $\mathbf{t}_n$ is given with an associated error. It is conceptually justified to assume a Gaussian error model (e.g.[10]), where the square of the (known) errors are arranged on the diagonals of variance matrices denoted by $\mathbf{S}_n$, so we have:

$$p(\mathbf{t}_n|\mathbf{w}_n) = \mathcal{N}(\mathbf{t}_n|\mathbf{w}_n, \mathbf{S}_n) \tag{1}$$

The unknown mean values $\mathbf{w}_n$ represent the clean, error-free version of the data. The genuine outliers, which we are interested in, must be those of the density of $\mathbf{w}_n$ rather than those of the density of $\mathbf{t}_n$. We also assume that it is the clean data rather than the contaminated data, who lies on a manifold. We will therefore model the hidden clean density as a GTM with Student t components $p(\mathbf{w}|k) = S_t(\mathbf{w}; \mathbf{W}, \sigma, \nu_k)$ which can be re-written as a convolution of a Gaussian with a Gamma ($\mathcal{G}(u|a, b) = b^a u^{a-1} e^{-bu}/\mathbf{\Gamma}(a)$) placed on the Gaussian precisions [7]. So in our model, the distribution of $\mathbf{t}$ can be obtained by integration over $\mathbf{w}$:

$$p(\mathbf{t}; \mathbf{W}, \sigma, \nu) = \frac{1}{K}\sum_k \iint \mathcal{N}(\mathbf{t}|\mathbf{w}, \mathbf{S})\mathcal{N}(\mathbf{w}|\mathbf{W}\phi(\mathbf{x}_k), \frac{1}{u\sigma})\mathcal{G}(u|\frac{\nu_k}{2}, \frac{\nu_k}{2})dud\mathbf{w} \tag{2}$$

3 Structured variational EM solution

Since the integration in Eq. (2) is not tractable, we develop a generalised EM (GEM) algorithm (see e.g. [6]), with approximate E-step. In general terms, for each data point $\mathbf{t}_n$, its log-likelihood can be bounded as follows:

$$\log p(\mathbf{t}_n|\theta) \geq \int q(h_n) \log \frac{p(h_n, \mathbf{t}_n|\theta)}{q(h_n)} dh \equiv \mathcal{F}(\mathbf{t}_n|q, \theta) \tag{3}$$

where q is the free-form variational posterior, $\mathcal{F}$ is called variational free energy function, h_n is the set of latent variables associated with $\mathbf{t}_n$, and θ is the set of parameters of the model. In our case, $h_n = (z_n, \mathbf{w}_n, u_n)$ and $\theta = (\mathbf{W}, \sigma, \{\nu_k\})$. Some tractable form needs to be chosen for q, e.g. a fully factorial form [6] is the most common choice. However, under our model definitions, it is feasible to keep some of the posterior dependencies by choosing the following tree-structured variational distribution: $q(\mathbf{w}, u, z = k) = q(z = k)q(\mathbf{w}|z = k)q(u|z = k) \equiv q(k)q(\mathbf{w}|k)q(u|k)$. The free energy function $\mathcal{F}(\mathbf{t}|q, \theta)$ can be evaluated as:

$$\mathcal{F}(\mathbf{t}|q, \theta) = \sum_k q(k) \left[\langle \log p(\mathbf{t}, \mathbf{w}, u, k) - \log \left(q(u|k)q(\mathbf{w}|k)q(k) \right) \rangle_{\mathbf{w},u|k} \right] \equiv \sum_k q(k) A_{\mathbf{t},k}.$$

Variational E-step Due to discretisation of the latent space, the E-step expressions follow analogously to the unconstrained case [9], and we obtain the following variational posteriors:

$$q(\mathbf{w}|k) = \mathcal{N}(\mathbf{w}|\langle \mathbf{w} \rangle_k, \mathbf{\Sigma}_{\mathbf{w}|k}); \quad q(u|k) = \mathcal{G}(u|a_k, b_k); \quad q(k) = \frac{\exp(A_{\mathbf{t},k})}{\sum_{k'} \exp(A_{\mathbf{t},k'})} \tag{4}$$

$$\mathbf{\Sigma}_{\mathbf{w}|k} = \left[\sigma \langle u \rangle_{u|k} + \mathbf{S}^{-1} \right]^{-1}; \qquad \langle \mathbf{w} \rangle_k = \mathbf{\Sigma}_{\mathbf{w}|k} \left[\langle u \rangle_{u|k} \sigma \mathbf{W} \phi(\mathbf{x}_k) + \mathbf{S}^{-1} \mathbf{t} \right] \tag{5}$$

$$a_k = \frac{\nu_k + D}{2}; b_k = \frac{\nu_k + C_k}{2}; C_k = \sigma \left(\| \langle \mathbf{w} \rangle_k - \mathbf{W} \phi(\mathbf{x}_k) \|^2 + \mathbf{Tr} \left(\mathbf{\Sigma}_{\mathbf{w}|k} \right) \right)$$

Using these, the likelihood bound is straightforward to evaluate (omitted for space constraints).

M-step Taking derivatives of $\mathcal{F}$ w.r.t $\mathbf{W}$, and solving the stationary equations, we obtain the following equation in matrix notation:

$$\mathbf{\Phi}^T \mathbf{G} \mathbf{\Phi} \mathbf{W}^T = \mathbf{\Phi}^T \mathbf{A} \tag{6}$$

where $\mathbf{\Phi}$ is a $K \times L$ matrix with element $\mathbf{\Phi}_{ij} = \phi_j(x_i)$, $\mathbf{A}$ is a $K \times d$ matrix, its (k, i) element is $\sum_n q(z_n = k) \langle u_n \rangle_k \langle \mathbf{w}_n \rangle_{ki}$, $\mathbf{G}$ is a diagonal $K \times K$ matrix with elements $\mathbf{G}_{kk} = \sum_n q(z_n = k) \langle u_n \rangle_k$. Eq. (6) can be solved by using standard matrix inversion techniques. Similarly, maximising the likelihood bound w.r.t σ, we can re-estimate the inverse variance σ as:

$$\frac{1}{\sigma} = \frac{1}{ND} \sum_n \sum_k q(z_n = k) \langle u_n \rangle_k \left\langle \| \mathbf{w}_n - \mathbf{W} \phi(x_k) \|^2 \right\rangle_{\mathbf{w}_n|k}.$$

Time Complexity The E-step scales as $\mathcal{O}(\max\{D, M\}KN)$, the M-step is $\mathcal{O}(\max\{D, M\}NK + MKd + M^3)$. In conclusion, the theoretical complexity per iteration is the same as that of the t-GTM.

4 Deriving interpretations from the model

Outlier Detection Criteria. In addition to mapping the density of non-outliers to 2D, our main goal includes the ability to visualise the genuine outliers. Similarly to the case of unconstrained density modelling [9], the posterior expectation of u is of interest:

$$e \equiv \sum_k q(k) \frac{\nu_k + D}{\nu_k + \sigma \left(\| \langle \mathbf{w} \rangle_k - \mathbf{W}\phi(\mathbf{x}_k) \|^2 + \mathbf{Tr}\left(\mathbf{\Sigma}_{\mathbf{w}|k} \right) \right)} \tag{7}$$

where a_k and b_k are given in Eq. (6). A data point is considered to be an outlier not due to errors, if its e value is sufficiently small, or equivalently, the value

$$v \equiv \sum_k q(k) \sigma \left(\| \langle \mathbf{w} \rangle_k - \mathbf{W}\phi(\mathbf{x}_k) \|^2 + \mathbf{Tr}\left(\mathbf{\Sigma}_{\mathbf{w}|k} \right) \right) \tag{8}$$

is sufficiently large. By contrast, in the t-GTM [11], the outlier criterion is

$$m \equiv \sum_k p(k|\mathbf{t}) \sigma \| \mathbf{t} - \mathbf{W}\phi(\mathbf{x}_k) \|^2 . \tag{9}$$

Comparing Eq. (9) with Eq. (8), we can see that the degree of outlierness will differ when measurement errors are taken into consideration.

Data Visualisation. Analogously to the original GTM [2], we can derive 2D coordinates of each multivariate data point $\mathbf{t}_n$, by computing its posterior expectation in the latent space. This is calculated as follows:

$$\langle \mathbf{x}|\mathbf{t}_n, \mathbf{W}^*, \sigma^* \rangle = \int p(\mathbf{x}|\mathbf{t}_n; \mathbf{W}^*, \sigma^*) \mathbf{x} d\mathbf{x} = \int \frac{p(\mathbf{t}_n|\mathbf{x}; \mathbf{W}^*, \sigma^*) p(\mathbf{x})}{p(\mathbf{t}_n)} \mathbf{x} d\mathbf{x} = \sum_k q(k) \mathbf{x}_k$$

where $\mathbf{W}^*$ and σ^* are the parameters after training. Each multivariate data point $\mathbf{t}_n$ will have its image in the latent space (in case $d = 2$), as $\langle \mathbf{x}|\mathbf{t}_n, \mathbf{W}^*, \sigma^* \rangle$.

It should be noted that in the case of an atypical object, outside the manifold defined by the bulk of data density, the posterior responsibilities, i.e. the $q(z_n = k)$, will tend to be approximately equal for all $k, 1 \leq k \leq K$. Hence these outliers will tend to lie in the center of the latent space. It is therefore not particularly convenient to visualise these outliers in the 2D space together with the non-outliers. To deal with this this problem, while also trating outlierness as a continuous notion, in this paper we propose to use a third dimension for representing outlierness on the visual image. In addition, we can use markers of different sizes to indicate the extent of measurement error for each object. These pieces of information together have the potential to provide a both comprehensive and intuitive visual representation of complex realistic data sets, which combines theoretical soundness and rigour with intuitive expressiveness.

Visualising a Held-out Set. Since the model is fully generative, it can also be applied to new, previously unseen data from the same source. For a given held-out data set, we need to calculate the posterior distributions of $\mathbf{w}_n$ and u_n associated with each test point $\mathbf{t}_n$. To this end, we fix the parameters $\mathbf{W}$ and σ, obtained from the training set and perform the E-step iterations until convergence. This typically converges at least an order of magnitude faster than the full training procedure.

5 Experiments and Results

Illustrative experiments on synthetic data First, we create a synthetic clean data set, sampled from a mixture of three well separated Gaussians in $\Re^{10}$ and a uniform distribution simulates the presence of genuine outliers. Then we add Gaussian noise to all sample points, in order to simulate measurement errors. The resulting data set is the input of our algorithm. The aim is to recover and display the genuine outliers, along with the density of non-outliers, despite the added errors. Moreover, we demonstrate the advantage of our 3D display. The leftmost plot of Fig. 1 shows a conventional 2D visualisation of the input data based on posterior means $\langle \mathbf{x} \rangle$. The rightmost plot of Fig.1 provides a 3D image with the outlierness (cf. eq. (7)) being displayed on a third, vertical dimension. On both plots, different markers are used for points in different Gaussian density-components, the outliers are highlighted by star markers and the marker sizes are proportional to the size of errors. We see, since outlierness is a continuous quantity, it is hard to distinguish the outliers based on 2D posterior means only. In the 3D display in turn they are nicely revealed.

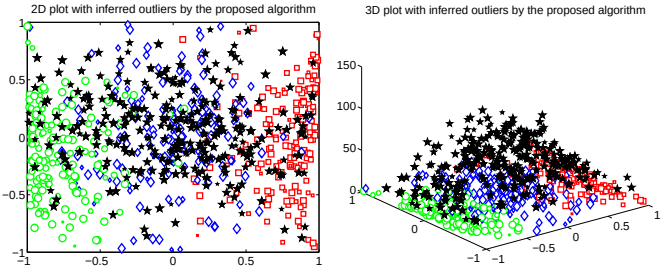

Fig. 1. Synthetic data sets with cluster structure and outliers.

In order to evaluate the accuracy of the data representation created in some objective manner, as well as to assess the benefits of including measurement error information, we evaluate k-nearest neighbour (KNN) classifiers on the obtained visual representation of a held-out set, sampled in the same way as the training set. The leftmost plot of Fig.2 shows the visual image of the held-out set, as obtained with our approach. The rightmost plot presents the correct KNN classification rates in comparison with those achieved with t-GTM (i.e. without knowledge of the measurement errors), varying the neighbourhood size as shown on the horisontal axis. At all neighbourhood sizes tested, the proposed algorithm achieves better performance compared with t-GTM. This demonstrates that inclusion of measurement error information is beneficial for achieving a more accurate representation of the data, and hence the visual exploration of this representation is more meaningful. We notice that KNNs with small neighbourhoods perform better in both cases, probably because of the high degree

of compression from the original data space and since the smoothness of the generative mapping preserves the *local* topology.

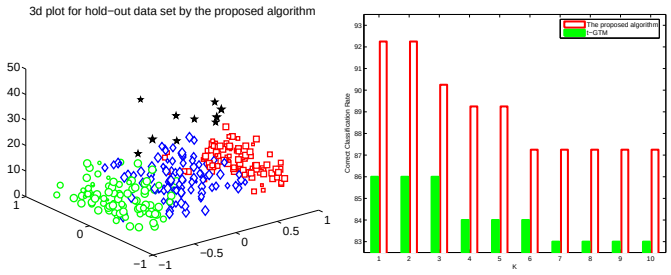

Fig. 2. Left: Visualisation of held-out data. Right: Comparison of KNN classification accuracies in the latent space, computed on held-out data.

Visualising high-redshift quasars from the SDSS quasar In this section, we apply the proposed algorithm on a well-studied data set in astrophysics – the SDSS quasar photometric catalogue [12]. The initial features are five observed magnitudes (measured brightness with a particular filter, in logarithmic units), u_{mag}, g_{mag}, i_{mag}, r_{mag} and z_{mag}, from which we created four features relating to colour, by subtracting r_{mag} from each. The reason for this is to avoid discriminating on the basis of brightness and the choice of r_{mag} as a reference magnitude is because that is most reliably measured. Further, the measurement errors are known for each feature and each object. In addition, spectroscopic redshift estimates are available. These are not used within the algorithm, but are useful for relating our results to known physical quantities, as a way of validating the interpretability of our results. The redshift relates to the distance of the object from us. As there are fewer distant quasars in the catalogue than closer ones, and given that with higher redshift the entire spectral pattern is shifted more towards lower frequencies, this provides a physical reason for high redshift quasars to be perceived as outliers in the overall density of quasars in the colour space. This observation was exploited in a number of previous studies by astronomers for finding high redshift quasars in various 2D projections of the data [4]. However through our developments, in addition to having a model based approach which takes the multivariate feature space and also takes principled account of the measurement errors [9], we are now also able to visualise the structure of the data so that domain experts may explore and understand the automated analysis such as the detection of atypical objects. In this visualisation, an optional step is to determine cluster assignments using the robust mixture modelling method for data with errors [9], in order to display the objects in the various clusters found in different colours or with different markers.

Here we apply our method to a sample of 2,000 quasars. The 3D visualisation produced by the proposed algorithm on the SDSS quasars subset is displayed in

Fig. 3. In this plot, we detail two selected outliers and two prototypical objects. The error bars represent their measurement errors, and the numbers shown on the plot are the redshift estimates for the selected outliers. As mentioned, we would expect the genuine outliers to be related to an increased redshift, in average, so we compute the area under the ROC curve (AUC) [5] of our outlierness estimates against a varying redshift threshold. The resulting relation between these two quantities is shown on Fig. 4. The y-coordinate of each point indicates the probability of detecting quasars of redshift greater than its x-coordinate. As expected, in a similar manner to the unconstrained analogue of the model presented [9], our principled method in four-colour space, using errors, can identify as outliers an overwhelming fraction of quasars already at a redshift of 2.5 (or higher). The main advantage of the constrained model over the unconstrained one is allowing us to visually understand these detections in the context of the structure of the entire data set. The rightmost plot on Fig. 4 shows the distribution of outlierness estimates with and without taking measurement errors into account. Upon zooming the denser region we notice the relative ranking produced by the two approaches is indeed quite different.

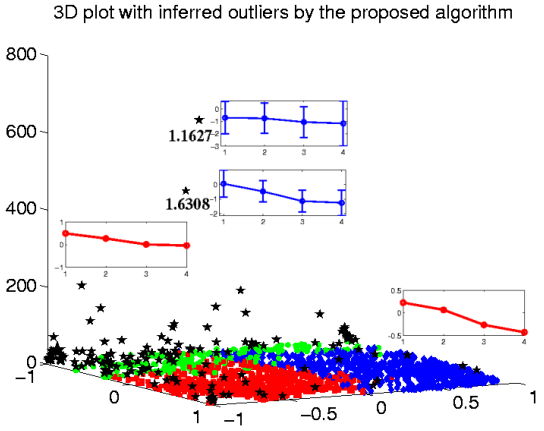

Fig. 3. The 3D plot of the proposed algorithm on the subset of SDSS quasar catalogue.

To relate our developments to existing methods in this application domain, it should be noted the prevalence of 2D projection methods plotting two features at a time against each other, e.g. [4]. Such a method can only manage to detect objects whose redshift is $z > 3.5$, which are extremely rare, and obvious from naive projections. In turn, by being able to map the multivariate data density to a visualisation space in a principled and robust manner, and taking into account measurement error information, we provide domain experts with a framework and methodology for further exploring realistic data sets in a more comprehensive and goal-directed manner.

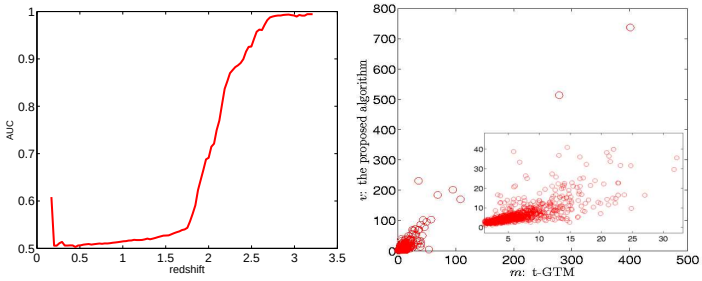

Fig. 4. Left: AUC vs. possible redshift thresholds. Right: The comparison of the distribution of outlierness values with and without taking into account measurement errors.

6 Conclusions

We presented a robust visual mining method for detecting and visualising interesting outliers in the presence of known measurement errors. A generalised EM algorithm was derived for inference and parameter estimation. The resulting algorithm was then demonstrated on both synthetic data and a real application.

References

1. C. Archambeau, N. Delannay, and M. Verleysen. Robust probabilistic projections. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.
2. C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–235, 1998.
3. S. Djorgovski, A. Mahabal, R. Brunner, R. Gal, and S. Castro. Searches for rare and new types of objects. *Virtual Observatories of the Future, ASP Conference Series*, 225, 2001.
4. X. Fan. A survey of $z > 5.7$ quasars in the Sloan Digital Sky Survey. IV. discovery of seven additional quasars. *The Astronomical Journal*, 131:1203–1209, 2006.
5. T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 2004.
6. M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.
7. C. Liu and D. Rubin. ML estimation of the t distribution using EM and its extensions: ECM and ECME. *Statistica Sinica*, 5:19–39, 1995.
8. D. Peel and G. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.
9. J. Sun, A. Kabán, and S. Raychaudhury. Robust mixtures in the presence of measurement errors. In *Proceedings of The 24th International Conference on Machine Learning*, 2007.
10. J. Taylor. *An introduction to error analysis*. University Science Books, 1996.
11. A. Vellido, P. Lisboa, and D. Vicente. Handling outliers and missing data in brain tumor clinical assessment using t-GTM. *Computers in Biology and Medicine*, 2006.
12. D. York. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120:1579–1587, 2000.