Making Nonlinear Manifold Learning Models Interpretable: the Manifold Grand Tour

Paulo J.G. Lisboa[a], José D. Martín-Guerrero[b], Alfredo Vellido[c,d]

[a] *Liverpool John Moores University (LJMU), Liverpool, United Kingdom,*
E-Mail: p.j.lisboa@livjm.ac.uk
[b] *Universitat de València (UV), València, Spain,*
*E-Mail:*jose.d.martin@uv.es
[c] *Universitat Politècnica de Catalunya (UPC), Barcelona, Spain*
[d] *Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Cerdanyola del Vallès, Spain,*
E-Mail: avellido@lsi.upc.edu

Abstract

Dimensionality reduction is required to produce visualizations of high dimensional data. In this framework, one of the most straightforward approaches to visualising high dimensional data is based on reducing complexity and applying linear projections while tumbling the projection axes in a defined sequence which generates a Grand Tour of the data. We propose using smooth nonlinear topographic maps of the data distribution to guide the Grand Tour, increasing the effectiveness of this approach by prioritising the linear views of the data that are most consistent with global data structure in these maps. A further consequence of this approach is to enable direct visualisation of the topographic map onto projective spaces that discern structure in the data. The experimental results on standard databases reported in this paper, using Self-Organising Maps and Generative Topographic Mapping, illustrate the practical value of the proposed approach. It must be remarked the novelty of the proposed method that improves some of the aspects of previous approaces based on the Grand Tour.

Keywords: Manifold learning, Grand Tour, data visualisation, nonlinear dimensionality reduction, linear projections

1. Introduction

When exploring any environment, it would be unusual, or even counterintuitive, not to try to visualize it first. The same applies to the exploration of data (Vellido et al., 2011). Visualisation processes are straightforward when data sets comprise a handful of attributes, since data visualisation can be readily implemented, for instance with multiple scatter-plots.

However, high dimensional data require the application of more advanced methods. This may involve the application of projective or mapping algorithms and becomes an important, or even necessary, stage of data analysis. This is specially true when the interpretability of the results is a requirement of the analysis Vellido et al. (2012). Such techniques are data visualisation-oriented instances of the more general family of Dimensionality Reduction (DR) methods.

Some of the most frequently used DR methods involve only linear combinations of the covariates. These methods have the advantage over their non-linear counterparts that when a gap in the observed data is seen from a particular projection (revealing data grouping structure), then that gap is known to be present and cannot close when the dimensionality of the projection is increased. A popular such method is Principal Component Analysis (PCA), which is typically applied in practice using biplots Jolliffe (2002). This approach compensates for one of its limitations, in particular sensitivity to noise and the lack of a robust criterion for choosing the adequate number of PCs, by the straightforward interpretability of the resulting projections.

Alternatively, Non-linear Dimensionality Reduction (NLDR) (Lee and Verleysen, 2007) methods are potentially more powerful to model complex high-dimensional data. These methods are well-suited to map the topological structure of the data, especially when the regions of interest cannot be well-separated using linear discrimination functions, or, equivalently, whenever mean values are not representative of density functions due to deviations from normality.

Manifold learning methods are part of the NLDR family of techniques that attempt to represent multivariate data by assuming they can be closely approxi-

mated using low-dimensional manifolds, typically chosen to be 2-dimensional for visualisation purposes. However, these methods can generate complex surfaces with possible occurrence of folds, which are often the result of overfitting. Moreover, the projection of data onto the visualization maps is heavily conditioned by the assumed structure of the map and so does not necessarily provide a clear picture of the empirical data density. In order to increase the interpretability of manifold learning techniques, it is of interest to combine generative models and NLDR with linear projective methods.

An alternative approach is to produce different views of the data arising from a succession of linear projections. A framework to generate a comprehensive range of low-dimensional projections is the *Grand Tour* proposed by Diane Cook and colleagues (Buja et al., 2005; Cook and Swayne, 2007). In this approach, the data are effectively tumbled in a systematic way and viewed through the prism of low-dimensional linear projections, looking for indicators of structure, typically gaps between sub-population cohorts. Due to the usefulness of the visualisations obtained, and the need of an easy and straightforward way to obtain them, this method has been recently implemented in an *R* package (Wickham et al., 2011), with a user-friendly graphical user interface (GUI) (Huang et al., 2012). While this approach is powerful in principle, we reckon that the search procedure may be expedited by prioritising the most informative views of the data; this is the goal and main novelty of the proposed approach compared to (Buja et al., 2005; Cook and Swayne, 2007). In (Lecerf and Bouchard, 2012), a method based on selecting candidate projections from the space of all projections was proposed. Our proposed method also pursues that goal but from a different and more complete perspective, since it does not require any user-interaction and a two-dimensional track is employed to guide three-dimensional projections of the data without the need for space-filling patterns.

Our conjecture is hence that prioritising, we are effectively introducing an implicit *narrative* in the process of visual data analysis. This added contextual information becomes a way of storytelling that should potentially provide more actionable knowledge (Segel and Heer, 2010).

The prioritisation of the most informative views of the data can effectively be provided by the NLDR methods. Therefore, in this study, we propose using NLDR manifolds as the base surface over which we roll-out sequences of linear projections, knowing that they will cross regions of high data density. This approach is intended to use the power of linear projections and leverage it on the data coverage generated by NLDR methods, in particular Self-Organising Map (SOM) networks (Kohonen, 2000) and Generative Topographic Mapping (GTM) (Bishop et al., 1998a).

In particular, this method aims to quickly discover gaps in the data distribution, which may be consistent with a hierarchical structure that may not be explicitly available even with prior clustering. The proposed methodology is the Manifold Grand Tour.

The remaining of the paper starts with a summary description of manifold learning models such as SOM and GTM, together with an overview of cohort-based linear visualisation, which improves on PCA by using data labels from cluster or class membership, whenever this information is available. This is followed by a detailed description of the Manifold Grand Tour procedure. Empirical results for two public domain data sets illustrate the application of the method.

2. Methods

2.1. Overview of Topographic Maps

The last decade has witnessed a quick development of nonlinear manifold learning methods for the analysis of multivariate data. Some examples include Locally Linear Embedding (Roweis and Saul, 2000) and Laplacian Eigenmaps (Belkin and Niyogi, 2003). Surveying such methods is beyond the focus of this paper. We instead focus on two consolidated techniques with similar goals but very different formalisation, namely SOM (Kohonen, 2000) and GTM (Bishop et al., 1998a).

2.1.1. Self-Organising Maps

Probably the best-known and widely used NLDR method for data visualisation is Kohonen's SOM, in its many variants. Although not strictly a manifold learning model, this method attempts to model data through a discrete version of a low-dimensional manifold consisting of a topologically ordered grid of prototypes.

SOM is an algorithmic procedure that simultaneously performs a combination of vector quantisation and topographic representation. Its nonlinearity has not prevented SOM from becoming mainstream in many application fields.

A SOM consists of a discrete layer (map) of units or neurons arranged in a low dimensional regular grid (often 2D, for visualisation). Each of these neurons k ($k = 1, \ldots, K$) is related, through an embedding function, with a $d-$dimensional vector $\mathbf{y}$, usually called prototype or weight vector.

Let $X = \{\mathbf{x}_n\}_{n=1}^{N}$ be a data set with vectors $\mathbf{x}$ of dimension d. After initialising the weight vectors $\mathbf{y}_k$, the algorithm finds the closest prototype to each data vector $\mathbf{x}_j$ ($j = 1, \ldots, N$), which is also known as best matching unit (BMU) $\mathbf{y}_{k_j}$ of index k_j, computed as $k_j = argmin_k \{d(\mathbf{x}_j, \mathbf{y}_k)\}$, where $d(\cdot, \cdot)$ is commonly defined as the Euclidean distance $L_2(\mathbf{x}_j, \mathbf{m}_k) = \|\mathbf{x}_j - \mathbf{m}_k\|$, although alternatives such as L_1 or L_∞, for instance, can also be considered.

Each BMU relates to its closest neighbours through a neighbourhood function $h(\cdot, \cdot)$. Different functions can be considered, being the Gaussian the most common choice. The prototype $\mathbf{y}_i$ is updated according to $\mathbf{y}_i^{(t+1)} = \mathbf{y}_i^{(t)} + \alpha^{(t)} h^{(t)}(\mathbf{x}_i, \mathbf{y}_c) \left(\mathbf{x}^{(t)} - \mathbf{y}_i^{(t)}\right)$, where t is time, $\mathbf{x}^{(t)} \in X$ is randomly selected at time t, and $0 \leq \alpha^{(t)} \leq 1$ denotes the learning rate.

The original version of SOM makes a separate update of the model parameters for each data point, taken one at a time, whereas its batch version makes the update on the basis of all data points. In this latter variant of the algorithm, the update equation can be rewritten in a kernel regression form (Mulier and

Cherkassky, 1995), for a given iteration, as:

$$\mathbf{y}_k = \sum_{k'} (F(\mathbf{u}_k, \mathbf{u}_{k'})\bar{\mathbf{x}}_{k'}) \tag{1}$$

where $\bar{\mathbf{x}}_{k'} = \frac{1}{n_{V'_k}} \sum_{j \in G'_k} \mathbf{x}_j$ is the mean of the group $G_{k'}$ of $n_{V'_k}$ data points assigned to a given node k', and $F(\mathbf{u}, \mathbf{u}_k) = N_k h(\mathbf{u}, \mathbf{u}_k) / \sum_{k'} N_{k'} h(\mathbf{u}, \mathbf{u}_{k'})$

2.1.2. Generative Topographic Mapping

The mostly heuristic definition of SOM inspired the development of a method that, while retaining its many functional advantages, was set within a principled probability theory framework. The resulting GTM (Bishop et al., 1998a) is a manifold learning model that, as SOM, has its main appeal in the simultaneous provision of multivariate data clustering and exploratory data visualisation. Its basic formulation has been extended to target goals as diverse as time series modelling (Olier and Vellido, 2008) , outlier detection (Vellido et al., 2009), unsupervised feature selection (Etchells et al., 2006), or semi-supervised learning (Cruz and Vellido, 2011), amongst others.

The GTM is also a Latent Variable Model (LVM). An LVM attempts to model observed data through the definition of a parsimonious set of non-observable, or latent variables (Bishop, 1998). Specifically, an LVM expresses the distribution $p(\mathbf{x})$ of the variables $x^1, \ldots, x^D$ of the observed data X in terms of a smaller number of latent variables $u^1, \ldots, u^L$, where $L < D$ and, if used for visualisation, $L \leq 3$. For that, the joint distribution $p(\mathbf{x}, \mathbf{u})$ is decomposed into the product of the marginal distribution $p(\mathbf{u})$ of the latent variables and the conditional $p(\mathbf{x}|\mathbf{u})$ of the observed data given the latent variables. The conditional distribution $p(\mathbf{x}|\mathbf{u})$ can be expressed in terms of a mapping from the latent space to the data space that involves a noise process. The definition of an LVM involves describing this conditional distribution as well as the mapping function itself and the marginal distribution $p(\mathbf{u})$. From these, the distribution $p(\mathbf{x})$ of

the data can be obtained by marginalising over the latent variables:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \tag{2}$$

In the GTM, a finite number of latent points $k = 1, \ldots, K$, usually spaced in a regular lattice, are mapped into the observed data space, each of them defining a prototype point. This prototype is the image of the former according to a mapping function in the form of a generalized regression model, so that each of the D-dimensional prototypes, $\mathbf{y}_k$, is defined as:

$$\mathbf{y}_k = \mathbf{W}\Phi(\mathbf{u}_k), \tag{3}$$

where Φ is a set of M basis functions ϕ_m (Gaussians in the standard model) that introduce the nonlinearity in the model, and $\mathbf{W}$ is a $D \times M$ matrix of adaptive weight parameters w_{dm}, each associated to a basis function m and to an observed data dimension d.

The prototype vector $\mathbf{y}_k$ can be considered as a representative of those data points $\mathbf{x}_n$ which are closer to it than to any other prototype. In that sense, this model clusters the data set as the result of a vector quantisation process. The set of prototypes resides in a smooth manifold (where such smoothness is conferred by the mapping function itself) that wraps around the observed data $X = \{\mathbf{x}_n\}_{n=1}^N$. The conditional distribution of the observed data variables, given the latent variables, $p(\mathbf{x}|\mathbf{u})$, involves a noise model with variance β^{-1}, defined as:

$$p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta) = (\frac{\beta}{2\pi})^{D/2} \exp\{-\frac{\beta}{2}\sum_{d=1}^{D}(x^d - y^d(\mathbf{u}))^2\}, \tag{4}$$

In order to integrate the latent variables out, we first need to define the marginal distribution $p(\mathbf{u})$. A regular square lattice of K latent points will be distributed according to $p(\mathbf{u}) = \sum_{k=1}^{K} \delta(\mathbf{u} - \mathbf{u}_k)$. This definition makes the integration in

Eq.(2) analytically tractable. The data distribution thus becomes:

$$p(\mathbf{x}|\mathbf{W},\beta) = \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{x}|\mathbf{u}_k,\mathbf{W},\beta)\} \tag{5}$$

From this expression, a complete model likelihood can be defined, and a maximum likelihood approach can be used for the estimation of the adaptive parameters of the model, usually through expectation-maximisation (EM) (Dempster et al., 1977). Details of the complete procedure can be found in (Bishop et al., 1998a,b).

For data visualisation, one of the results obtained in the maximisation step of the EM algorithm can be used through a direct application of Bayes' theorem that inverts the mapping from latent space to observed data space, producing the conditional probability of each latent point given each observed data point:

$$p(\mathbf{u}_k|\mathbf{x}_n) = \frac{p(\mathbf{x}_n|\mathbf{u}_k,\mathbf{W},\beta)}{\sum_{k'=1}^{K} p(\mathbf{x}_n|\mathbf{u}_{k'},\mathbf{W},\beta)}, \tag{6}$$

which is often referred to as the *responsibility* of each latent point for the generation of each observed data point, $r_{kn} \equiv p(\mathbf{u}_k|\mathbf{x}_n)$. This responsibility can be used to obtain data visualisation in the form of a posterior mode projection of $\mathbf{x}_n$: $k_n^{mode} = \arg\max_{\{k_n\}} r_{kn}$ (which implies assigning each observed data point to that latent point with the highest responsibility for its generation), or a posterior mean projection $\mathbf{u}_n^{mean} = \sum_{k=1}^{K} r_{kn}\mathbf{u}_k$ (placing the observed data point at a location in latent space that results from a responsibility-weighted combination of all latent point locations).

2.2. The Grand Tour

The concept of the Grand Tour was published in its original form by Asimov and colleagues (Asimov, 1985; Buja and Asimov, 1985) to explore high-dimensional data by travelling along a series of 2D-planes in which the data would be projected. In this way, the Grand Tour introduces a sequential element to the exploration process with the aim of obtaining new insights from the 2D visualisation of the data that might remain occult otherwise. As pointed

out in (Wegman, 2003), the grand tour is, in effect, an animation of the data. This animation requires some illusion of continuity that can only be achieved by ensuring that the navigation changes smoothly over time.

This is achieved in (Wegman, 2003) with a continuous geometric transformation of the coordinate system through all possible orientations of the coordinate axes.

The Grand Tour defines a trajectory through the Grassmannian manifold $G(2, D)$, which is the space of all 2D planes through the origin. This is an efficient calculation of a space-filling curve in the manifold of low-dimensional projections of high-dimensional data spaces. A number of different algorithms to implement the Grand Tour have been developed over time and their description is beyond the scope of this paper. The reader is referred to (Wegman et al., 2002), where a discussion of several approaches can be found, and to a more recent publication (Buja et al., 2005) on Grand Tours and related data visualisation methods.

In this paper, we alternatively propose restricting the sliding of the viewfinder of the Grand Tour to a trajectory through the previously obtained smooth manifold model of the data distribution, obtained with a nonlinear topographic model. This will generate a limited but faster visualisation process that aims to visit the most informative perspectives by simply complementing the NLDR manifold methods through the addition of linear projections of the data, where the interpretation of structural features such as gaps between clusters or marginal shape profiles, is more straightforward.

A further improvement will arise if the linear projective axes are chosen with as much knowledge about the data as possible, for instance cohort labels arising from clustering or class tags. This is the subject of the next section.

2.3. Cohort-Based Visualisation with Scatter Matrices

Purely linear DR methods for visualisation frequently utilize singular values spanning the largest variance in the data, as in PCA biplots (Jolliffe, 2002). While this approach is useful to visually verify known correlations between at-

tributes, the first two or three PCs that could be used for visualisation may well explain only a relatively small proportion of the data variance in the data. As already mentioned in the introduction, there is no guarantee that those PCs will provide a faithful enough representation of the data. As a result, true compact groups of data are severely mixed in the representation space due to the loss of information incurred by the projection.

When population cohorts are labelled, it is straightforward to decompose the data covariance matrix using the cohort means and the variance of each cohort with respect to the corresponding mean point. This is justified for linear modelling of discriminant features to separate the cohorts, on the basis that second order statistics are sufficient for the parameterisation of multivariate normal distributions which, in turn, are consistent with the assumption of linear separating surfaces.

The cohort-based visualisation with scatter matrices method described in (Lisboa et al., 2008) starts with the following identity showing that the total variance matrix, S_T, can be expressed as the sum of within- and between-group scatter matrices defined around the cohort means m_i:

$$S_T = S_W + S_B, \tag{7}$$

where

$$S_T = \sum_{i=1}^{N}((X_i - m)^T(X_i - m)), \tag{8}$$

$$S_W = \sum_{j=1}^{N_c}\sum_{i=1}^{N_j}((X_i - m_j)^T(X_i - m_j)), \tag{9}$$

$$S_B = \sum_{j=1}^{N_c}(N_j(m_j - m)^T(m_j - m)) \tag{10}$$

and m is the overall data mean; N_c is the number of labelled cohorts and N_j is the number of items in cohort j.

The so-called separating matrix is defined by extending the intuitive concept of the ratio of the variance of the means over the within-covariance matrix, as follows

$$M_T = S_W^{-1} S_B. \tag{11}$$

This matrix replaces the data covariance matrix in the calculation of the eigenvectors with the largest eigenvalues, which form the projection directions, now informed by the cohort labels. An extension of this method to the case where the covariance matrix of the data is singular can be found in (Lisboa et al., 2008).

2.4. The Manifold Grand Tour (MGT)

The proposed visualisation of the data is now straightforward. Given a topographic map of the data, which passes through the peaks in the data density distribution, and assuming a 2D structure to the map, the MGT procedure can be described as follows:

- Fit a topographic map to the data (a GTM in the experiments reported in this paper, although variants of SOM or alternative methods could be used).
- Start at an arbitrary node, e.g. one of the corners of the map, and the direction along the edges of the node defined by that node and its nearest neighbours. The two-dimensional structure of the first square cell defines a plane, for which orthonormal spanning coordinate axes can be obtained using Gram-Schmidt orthogonalisation.
- With the cohort-based visualisation method described in Section 2.3 (or, for instance, with PCA), find the direction of maximum spread of the data and with Gram-Schmidt and define a third projective axis.
- The complete data can now be displayed, along with a projection of the manifold and coordinate axes, if required, by projecting onto the linear

3D space spanned by the axes defined above.

- Move onto the next node and repeat the previous steps.

If further views of the data are required over each node in the topographic map, then the third axis can rotate from each eigenvector of the covariance (or separating) matrix to the next. This can be done either in order, or reducing the size of the corresponding eigenvalue, returning from the last to the first eigenvalue before proceeding to the next node, for which the first two dimensions change slowly, due to the smoothness of the topographic map. Each successive iteration will be less informative since the separation between data cohorts will gradually reduce.

A limitation of the method is that the views of the data are bound to lie in the space spanned by the edges linking successive nodes in the topographic map and the span of the matrix used to define the third axis for each visualisation perspective. If this matrix is the separating matrix, then the dimensionality of this space is limited by the rank of that matrix, which is the number of distinct cohorts minus one. However, if the variance matrix is used, then this is clearly of full rank.

In both cases, the quality of the visualisation depends on how well the topographic maps cover the data. In each case, the eigenvector structure of the matrices derived from the second-order statistics take over from the Grassmannian manifolds as the "tour guides".

In the following experiments, it was sufficient to show the first iteration where the eigenvector with the largest eigenvalue of the separating matrix was used to define the orthogonal direction in each cell of the topographic surface covering the data.

3. Experiments

3.1. Materials

The proposed methodology for multivariate data visualisation was tested in two different real data sets: *Italian olive oil* (Cook and Swayne, 2007; Forina

et al., 1983) and *music* (Cook and Swayne, 2007).

The *Italian olive oil* data set consists of 572 samples and 10 variables. Eight variables describe the percentage composition of fatty acids found in the lipid fraction of these oils, which is used to determine their authenticity. The remaining two variables contain information about the classes, which are of two kinds: three "super-classes" at country level: North, South, and the island of Sardinia; *and* nine collection area classes: three from the Northern region (Umbria, East and West Liguria), four from the South (North and South Apulia, Calabria, and Sicily), and two from the island of Sardinia (inland and coastal Sardinia).

The goal is to distinguish the oils from different regions and areas in Italy based on their combinations of the fatty acids. The clusters corresponding to classes all have different shapes in the eight-dimensional data space defined by the concentration of fatty acids.

The *music* data set consists of 62 samples and seven variables. Data were produced by reading different songs using the music editing software Amadeus II®, and then snipping and saving the first 40-second clip of each as a WAV file. Audio was converted into numeric data using the R programming language. The meaning of the variables is the following:

- Artist: Abba, Beatles, Eels, Vivaldi, Mozart, Beethoven, Enya.
- Type: rock, classical, or new wave.
- Average, variance and maximum of the frequencies of the left channel (three variables).
- Amplitude of the loudness of the sound.
- Median of the location of the 15 highest peaks in the periodogram.

The analysis goal for this data set is to group the tracks into a small number of clusters according to their similarity in terms of audio characteristics, thus enlightening whether, for instance, rock and classical tracks are distinguishable. This knowledge can be applied, for instance, to arrange tracks on a digital music player, or to make recommendations based on track similarity.

3.2. Experimental Settings

The adaptive parameters of the GTM models used to analyze the data described in the previous section were initialized following a standard procedure described in (Bishop et al., 1998a): The weight matrix $\mathbf{W}$, which embodies the mapping from the latent to the observed data space, was defined so as to minimize the difference between the prototype vectors $\mathbf{y}_k$ defined in Eq.(3) and the vectors that would be generated in the observed space by a partial PCA process. The inverse noise model variance parameter β is initialized as the inverse of the 3^{rd} PCA eigenvalue. This initialisation procedure has been shown to be reliable while ensuring the replicability of the results that could not be guaranteed by a random initialisation of parameters.

Different GTM square lattice sizes were explored but, in the end, it is convenient to achieve a trade-off between detail (which would be proportional to the size of the lattice) and practical visual interpretability. For the analysed data, a suitable layout for the GTM lattice was found to be a 15×15 grid, which was thus chosen for all the reported experiments.

In order to avoid data overfitting, a regularized version of GTM was used. Regularisation encourages smoother manifolds in what, in fact, becomes a complexity control process that is achieved with the addition of a regularisation term to the log-likelihood of the model, which becomes:

$$L_{reg} = \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{W}, \beta) - \frac{1}{2}\alpha\|\mathbf{w}\|^2. \tag{12}$$

Here, α is a regularisation coefficient and $\mathbf{w}$ is the vector resulting from the concatenation of the different column vectors of the weight matrix $\mathbf{W}$. The optimisation of the parameters can be accomplished using the Bayesian formalism and, more in particular, the evidence approximation (Mackay, 1991; Vellido et al., 2003).

Using the manifolds yielded by GTM, visualisations were produced according to the procedure described in section 2.4.

3.3. Results and Discussion

3.3.1. Italian Olive Oil data set

We expected the GTM-based MGT to produce useful visualisations when projecting the *data* into the axes defined in some of the GTM nodes using the procedure described in Section 2.4. Figure 1 shows a selection of three representative *projections* for illustration: the top plot represents a projection in which the three clusters are mixed up and extensively overlap. It is difficult to separate the three regions from mere visual inspection. The bottom plot, instead, shows a projection in which the three clusters are neatly separated, while an intermediate case (neither so well-separated as the bottom plot, nor mixed-up as the top one) is shown in the middle plot.

Although, for the sake of brevity, many results are omitted, it should be emphasized that the projections into many of the nodes produced quite a few very meaningful plots that showed the difference between the three main classes (*South*, *Sardinia* and *North*) clearly. Moreover, results matched those achieved in (Cook and Swayne, 2007), revealing the presence of internal structure specially in Cluster 2 (Sardinian origin), which is shown to be formed by two subclusters (*Inland* or *Coastal Sardinia*). This is clearly revealed by the detailed nine sub-classes (Umbria, East and West Liguria, North and South Apulia, Calabria, Sicily, and inland and coastal Sardinia) representation in Figure 2, where *Inland* is represented by blue triangles and *Coast* by black pentagrams.

3.3.2. Music data set

As for the *Italian Olive Oil* data set, the regularized GTM corresponding to the *music* data set also generated a manifold with some degree of folding, hence relevant visualisations were produced when projecting the data into the axes defined by the GTM nodes according to the proposed MGT procedure. Again summarily, Figure 3 shows three projections of the data into different nodes of the GTM. While the top and middle plots represent visualisations that are not especially helpful, since the three different clusters (rock, classical, new wave) do not appear clearly separated, the situation is reversed in the bottom plot,

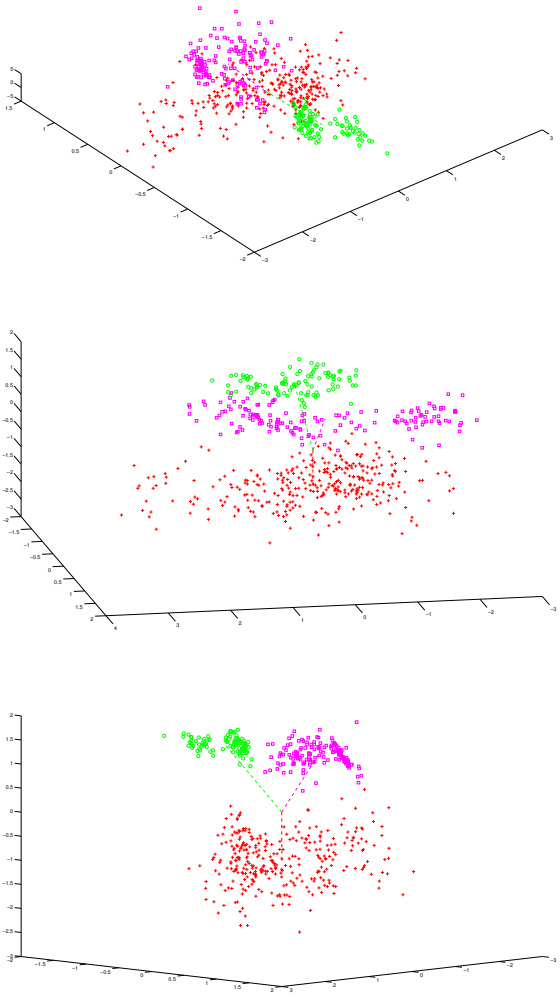

Figure 1: Three chosen illustrative MGT projections of the *Italian Olive Oil* data set, colour-labelled to show the main three classes (*South*: red crosses, *Sardinia*: green circles and *North*: magenta squares), defined in three different GTM nodes. The bottom plot exemplifies a projection in which the three clusters are clearly separated, while the top and middle plots correspond to projections in which it is more difficult to visually disentangle the cluster structure.

which represents one of the projections in which the three clusters can be easily differentiated.

As in the previous data set, the presence of an internal structure, which

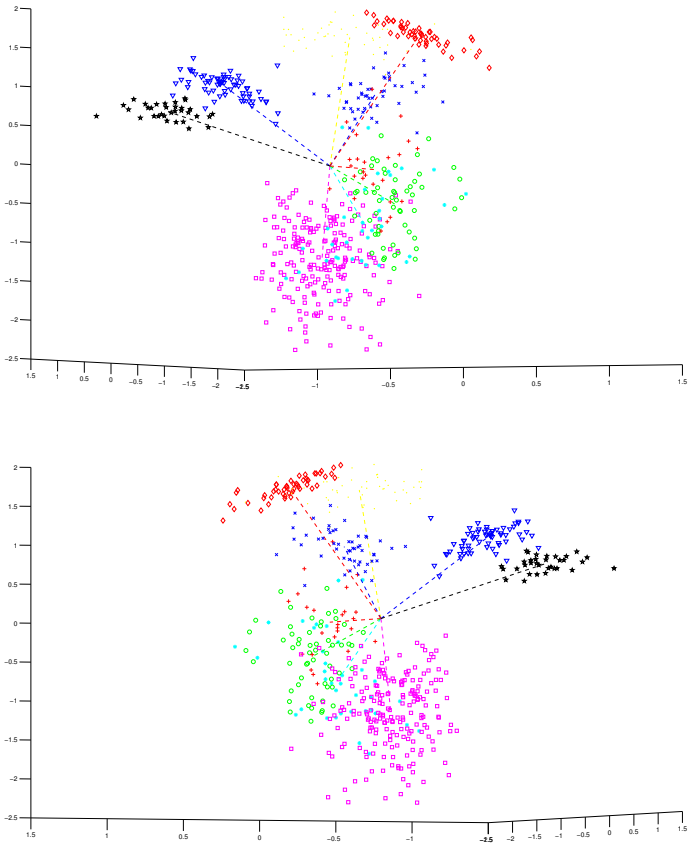

Figure 2: Two chosen illustrative MGT projections of the *Italian Olive Oil* data set, this time colour-labelled to show the detailed nine sub-class structure, defined in two different GTM nodes. The internal structure of Sardinian cluster is shown in two sub-clusters (black pentagrams and blue triangles)

would not be obvious from the single flat visualisation of the data provided by the GTM, is remarkable. This is particularly true for *rock*, but a certain internal structure within *classical* can also be visually discerned. This might be explained by the fact that classical music is a more normative style than rock and, as a result, the influence of the artist is not as relevant in the former as it is in the latter for the definition of the internal structure. Such an effect is likely to be more predominant in small data sets such as the one analyzed in these experiments.

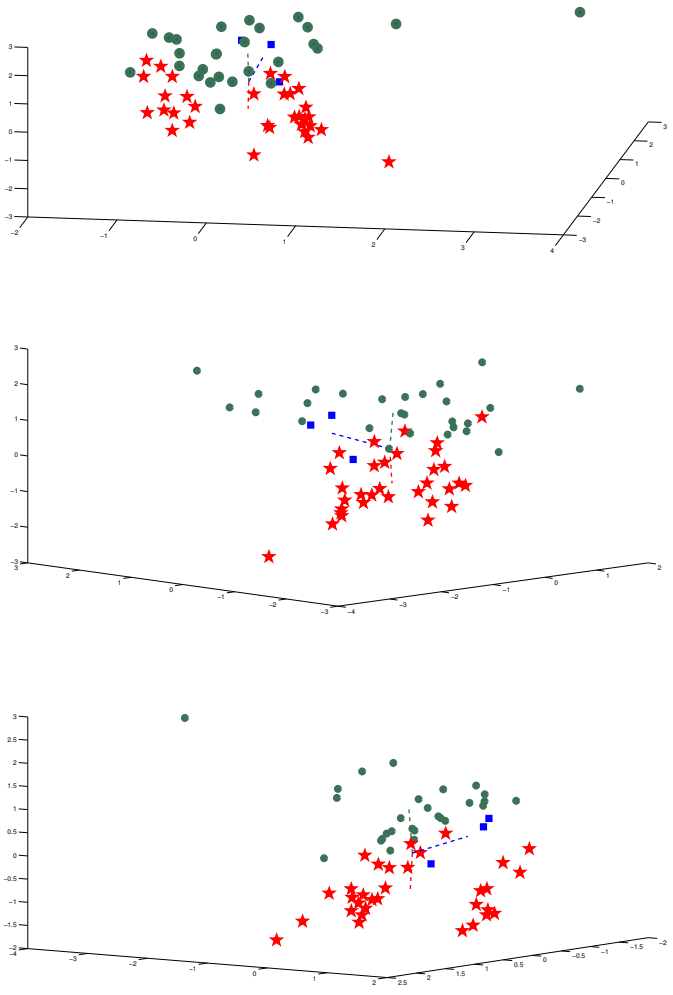

Figure 3: Three chosen illustrative MGT projections of the *Music* data set, colour-labelled to show the main three classes (*Rock* is represented by red stars, *classical* by green circles and *new wave* by blue squares), defined in three different GTM nodes. The bottom plot exemplifies a projection in which the three clusters are clearly separated, while the top and middle plots once again correspond to projections with different degree of visual cluster overlapping.

4. Conclusions

This paper has presented a new approach for NLDR methods oriented to multivariate exploratory data visualisation that combines the modelling flex-

ibility of one such method, namely GTM, with the interpretability of linear models for data visualisation. In the reported experiments, GTM has been used to guide a *Grand Tour* of some real data sets that uses a recently proposed linear DR method for data visualisation, which is based on a clustering approach. The achieved results illustrate the suitability of the proposed method to produce useful representations that intuitively reveal the internal data structure.

This paper involves a relevant theoretical advance with respect to the standard Grand Tour since views are not random but selected according a smart guide, such us GTM. This work also improves and completes the approach presented in (Lecerf and Bouchard, 2012) since neither user interaction nor a two-dimensional track is required to guide three dimensional projections. The main limitation of the study is related to the number of class structures since a high number of classes might difficult the visualization; this is however a common problem in this kind of visualizations. It is finally remarkable that an appealing advantage of the proposed method is that it could straightforwardly be extended to alternative manifold learning algorithms.

References

References

Asimov, D., 1985. The grand tour: A tool for viewing multidimensional data. SIAM Journal of Science and Statistical Computing 6, 128–143.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15 (6), 1373–1396.

Bishop, C. M., 1998. Learning in Graphical Models. The M.I.T. Press, Ch. Latent variable models, pp. 371–404.

Bishop, C. M., Svensén, M., Williams, C., 1998a. GTM: The generative topographic mapping. Neural Computation 10 (1), 215–234.

Bishop, C. M., Svensén, M., Williams, C. K. I., 1998b. Advances of the generative topographic mapping. Neurocomputing 21 (1–3), 203–224.

Buja, A., Asimov, D., 1985. Grand tour methods: an outline. In: Computing Science and Statistics: Proceedings of the Seventeenth Symposium on the Interface. pp. 63–67.

Buja, A., Cook, D., Asimov, D., Hurley, C., 2005. Handbook of Statistics Volume 24: Data Mining and Data Visualization. Elsevier, Ch. Computational Methods for High-Dimensional Rotations in Data Visualization, pp. 391–414.

Cook, D., Swayne, D. F., 2007. Interactive and Dynamic Graphics for Data Analysis. Springer Verlag, Berlin, Germany.

Cruz, R., Vellido, A., 2011. Semi-supervised analysis of human brain tumours from partially labeled MRS information, using manifold learning models. International Journal of Neural Systems 21 (1), 17–29.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of The Royal Statistical Society Series B-statistical Methodology 39 (1), 1–38.

Etchells, T. A., Nebot, A., Vellido, A., Lisboa, P. J. G., Mugica, F., 2006. Learning what is important: feature selection and rule extraction in a virtual course. In: Proceedings of the ESANN 2006. pp. 401–406.

Forina, M., Armanino, C., Lanteri, S., Tiscornia, E., 1983. Food Research and Data Analysis. Applied Science Publishers, Ch. Classification of Olive Oils from their Fatty Acid Composition, pp. 189–214.

Huang, B., Cook, D., Wickham, H., 2012. tourrGui: A gWidgets GUI for the tour to explore high-dimensional data using low-dimensional projections. Journal of Statistical Software 49 (6), 1–12.

Jolliffe, I. T., 2002. Principal Component Analysis, 2nd Edition. Springer Series in Statistics, Springer Verlag, Berlin, Germany.

Kohonen, T., 2000. Self-Organizing Maps. Information Science Series, Springer, Berlin, Germany.

Lecerf, L. M., Bouchard, G. M., 06 2012. Adaptive Grand Tour. URL https://www.lens.org/lens/patent/US_8194077_B2

Lee, J. A., Verleysen, M., 2007. Nonlinear Dimensionality Reduction, Information Science and Statistics. Springer, Berlin, Germany.

Lisboa, P. J. G., Ellis, I. O., Green, A. R., Ambrogi, F., Dias, M. B., 2008. Cluster-based visualisation with scatter matrices. Pattern Recognition Letters 29 (13), 1814–1823.

Mackay, D., 1991. Bayesian Methods for Adaptive Models. Ph.D. thesis, California Institute of Technology, U.S.A.

Mulier, F., Cherkassky, V., 1995. Self-organization as an iterative kernel smoothing process. Neural Computation 7 (6), 1165–1177.

Olier, I., Vellido, A., 2008. Advances in clustering and visualization of time series using gtm through time. Neural Networks 21 (7), 904–913.

Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Segel, E., Heer, J., 2010. Narrative visualization: Telling stories with data. IEEE Transactions on Visualization and Computer Graphics 16 (6), 1139–1148.

Vellido, A., El-Deredy, W., Lisboa, P., 2003. Selective smoothing of the generative topographic mapping. IEEE Transactions on Neural Networks 14 (4), 847–852.

Vellido, A., Martín, J. D., Lisboa, P. J. G., 2012. Making machine learning models interpretable. In: Proceedings of the ESANN 2012. pp. 163–172.

Vellido, A., Martín, J. D., Rossi, F., Lisboa, P. J. G., 2011. Seeing is believing: The importance of visualization in real-world machine learning applications. In: Proceedings of the ESANN 2011. pp. 219–226.

Vellido, A., Romero, E., González-Navarro, F. F., Belanche-Muñoz, L., Julià-Sapé, M., Arús, C., 2009. Outlier exploration and diagnostic classification of a multi-centre ^{1}H-MRS brain tumour database. Neurocomputing 72 (13–15), 3085–3097.

Wegman, E. J., , Solka, J. L., 2002. On some mathematics for visualizing high dimensional data. The Indian Journal of Statistics, Series A 64 (2), 429–452.

Wegman, E. J., 2003. Visual data mining. Statistics in Medicine 22 (9), 1383–1397.

Wickham, H., Cook, D., Hofmann, H., Buja, A., 2011. tourr: An R package for exploring multivariate data with projections. Journal of Statistical Software 40 (2), 1–18.