A Principled Approach to Interactive Hierarchical Non-Linear Visualization of High-Dimensional Data

Peter Tiňo, Ian Nabney, Yi Sun
Neural Computing Research Group
Aston University, Birmingham, B4 7ET, UK
`tinop,nabneyit,suny@aston.ac.uk`

Bruce S. Williams
Pfizer Global Research and Development
Ramsgate Road, Sandwich, CT13 9NT, UK

Abstract

Hierarchical visualization systems are desirable because a single two-dimensional visualization plot may not be sufficient to capture all of the interesting aspects of complex high-dimensional data sets. We extend an existing locally linear hierarchical visualization system PhiVis [1] in several directions: **(1)** we allow for *non-linear* projection manifolds (the basic building block is the Generative Topographic Mapping – GTM), **(2)** we introduce a general formulation of hierarchical probabilistic models consisting of local probabilistic models organized in a hierarchical tree, **(3)** we describe folding patterns of low-dimensional projection manifold in high-dimensional data space by computing and visualizing the manifold's local directional curvatures. Quantities such as magnification factors [3] and directional curvatures are helpful for understanding the layout of the nonlinear projection manifold in the data space and for further refinement of the hierarchical visualization plot. Like PhiVis, our system is statistically principled and is built interactively in a top-down fashion using the EM algorithm. We demonstrate the visualization system principle of the approach on a complex 12-dimensional data set and mention possible applications in the pharmaceutical industry.

1 Introduction

In general, a single two-dimensional projection of high-dimensional data, even if it is non-linear, may not be sufficient to capture all of the interesting aspects of the data. Indeed, when investigating a data set through low-dimensional projections, it is natural take a hierarchical approach: one first constructs a top-level plot and then concentrates on local regions of interest by recursively building the corresponding sub-projections. The sub-models are organized in a hierarchical tree.

The basic building block of our hierarchical visualization system is the *Generative Topographic Mapping* (GTM) introduced by Bishop, Svensén and Williams in [2]. Basically, GTM is a probabilistic re-formulation of the Kohonen self-organizing map (SOM), but unlike SOM, GTM defines an explicit probability density model of the data. This enables us to apply a consistent and *statistically principled* framework of [1] to formulate hierarchical *non-linear* visualization trees. Also, since GTM forms a *smooth* two-dimensional manifold in the data space, one can analytically compute quantities such as magnification factors and directional curvatures that are important for both understanding the visualization plots and constructing deeper levels of the visualization hierarchy. Magnification factors quantify the extent to which the areas are magnified on projection to the data space. Directional curvatures capture the local folding patterns of the projection manifold.

2 Generative Topographic Mapping

The Generative Topographic Mapping (GTM) [2] models a probability distribution in the (observable) high-dimensional data space $\mathcal{D} = \Re^D$ by means of low-dimensional latent, or hidden variables. The data is visualized in the latent space $\mathcal{H} \subset \Re^L$ (usually a two-dimensional interval $[-1, 1] \times [-1, 1]$).

We cover the latent space $\mathcal{H}$ with an array of K latent space centers $\mathbf{x}_i \in \mathcal{H}$, $i = 1, 2, ..., K$. In this study $K = 15 \times 15 = 225$. Non-linear GTM transformation $f : \mathcal{H} \to \mathcal{D}$ from the latent space to the data space is defined using a radial basis function network. To this end, we cover the latent space with a set of $M - 1$ fixed non-linear basis functions (here Gaussian functions of the same width σ) $\phi_j : \mathcal{H} \to \Re$, $j = 1, 2, ..., M - 1$, (here $\sigma = 1.0$, $M = 4 \times 4 + 1 = 17$, a single constant basis function accounts for the bias term) centered on a regular grid in the latent space. Given a point $\mathbf{x} \in \mathcal{H}$ in the latent space, its image under the map f is

$$f(\mathbf{x}) = \mathbf{W}\, \phi(\mathbf{x}), \tag{1}$$

where $\mathbf{W}$ is a $D \times M$ matrix of weight parameters and $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), ..., \phi_M(\mathbf{x}))^T$.

GTM creates a generative probabilistic model in the data space by placing a radially-symmetric Gaussian with zero mean and inverse variance β around images, under f, of the latent space centers $\mathbf{x}_i \in \mathcal{H}$, $i = 1, 2, ..., K$. We refer to the Gaussian density associated with the center $\mathbf{x}_i$ by $P(\mathbf{t}|\, \mathbf{x}_i, \mathbf{W}, \beta)$.

Defining a uniform prior over $\mathbf{x}_i$, the density model in the data space provided by the GTM is $P(\mathbf{t}|\, \mathbf{W}, \beta) = 1/K \sum_{i=1}^{K} P(\mathbf{t}|\, \mathbf{x}_i, \mathbf{W}, \beta)$.

For the purpose of data visualization, we use Bayes' theorem to invert the transformation f from the latent space $\mathcal{H}$ to the data space $\mathcal{D}$. The posterior distribution on $\mathcal{H}$, given a data point $\mathbf{t}_n \in \mathcal{D}$, is a sum of delta functions centered at centers $\mathbf{x}_i$, with coefficients equal to the posterior probability R_{in} that the i-th Gaussian (corresponding to the latent space center $\mathbf{x}_i$) generated $\mathbf{t}_n$ [2],

$$R_{i,n} = \frac{P(\mathbf{t}_n|\, \mathbf{x}_i, \mathbf{W}, \beta)}{\sum_{j=1}^{K} P(\mathbf{t}_n|\, \mathbf{x}_j, \mathbf{W}, \beta)}. \tag{2}$$

The latent space representation of the point $\mathbf{t}_n$, i.e. the *projection of* $\mathbf{t}_n$, is taken to be the mean $\sum_{i-1}^{K} R_{in}\, \mathbf{x}_i$ of the posterior distribution on $\mathcal{H}$.

The f–image of the latent space $\mathcal{H}$, $\Omega = f(\mathcal{H}) = \{f(\mathbf{x}) \in \Re^D |\, \mathbf{x} \in \mathcal{H}\}$, forms a *smooth* L-dimensional manifold in the data space. We refer to the manifold Ω as the *projection manifold* of the GTM.

Magnification factors of Ω were calculated as the Jacobian of the GTM map f in [3]. In [4] we derived a closed-form formula for directional curvature of the projection manifold Ω for a latent space point $\mathbf{x} \in \mathcal{H}$ and a directional vector $\mathbf{h} \in \mathcal{H}$.

3 Hierarchical GTM

The hierarchical GTM arranges a set of GTMs and their corresponding plots in a tree structure $\mathcal{T}$. In this section we give a general formulation of hierarchical mixture models: more detail can be found in [5].

The *Root* is at level 1, i.e. $Level(Root) = 1$. Children of a model $\mathcal{N}$ with $Level(\mathcal{N}) = \ell$ are at level $\ell + 1$, i.e. $Level(\mathcal{M}) = \ell + 1$, for all $\mathcal{M} \in Children(\mathcal{N})$.

Each model $\mathcal{M}$ in the hierarchy, except for *Root*, has an associated parent-conditional mixture coefficient, or prior $\pi(\mathcal{M}|Parent(\mathcal{M}))$. The priors are non-negative and satisfy the consistency condition: $\sum_{\mathcal{M} \in Children(\mathcal{N})} \pi(\mathcal{M}|\,\mathcal{N}) = 1$. Unconditional priors for the models are recursively calculated as follows: $\pi(Root) = 1$, and for all other models

$$\pi(\mathcal{M}) = \prod_{i=2}^{Level(\mathcal{M})} \pi(Path(\mathcal{M})_i|\, Path(\mathcal{M})_{i-1}), \tag{3}$$

where $Path(\mathcal{M}) = (Root, ..., \mathcal{M})$ is the N-tuple $(N = Level(\mathcal{M}))$ of nodes defining the path in $\mathcal{T}$ from *Root* to $\mathcal{M}$.

Distribution given by the hierarchical model is a mixture of leaf models of $\mathcal{T}$,

$$P(\mathbf{t}|\,\mathcal{T}) = \sum_{\mathcal{M} \in Leaves(\mathcal{T})} \pi(\mathcal{M})\, P(\mathbf{t}|\,\mathcal{M}). \tag{4}$$

Non-leaf models not only play their role in the process of creating the hierarchical model, but in the context of data visualization can be useful for determining the relationship between related subplots in the hierarchy.

3.1 Training

The hierarchical GTM is trained using EM to maximize its likelihood with respect to the data sample $\zeta = \{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N\}$. Training of a hierarchy of GTMs proceeds in a recursive fashion. First, a base (*Root*) GTM is trained and used to visualize the data. Then the user identifies interesting regions on the visualization plot that they would like to model in a greater detail. In particular, the user chooses a collection of points $\mathbf{c}_i \in \mathcal{H}$ by clicking on the plot. These "regions of interest" are then transformed into the data space as Voronoi compartments defined by the mapped points $f_{Root}(\mathbf{c}_i) \in \mathcal{D}$, where f_{Root} is the map (1) of the *Root* GTM. The child GTMs are initiated by local PCA in the corresponding Voronoi compartments. After training the child GTMs and seeing the lower level visualization plots, the user may decide to proceed further and model in a greater detail some portions of the lower level plots, etc. At each stage of hierarchical GTM construction, the EM algorithm alternates between the E- and M-steps until convergence is satisfactory (typically after 10–20 iterations). A detailed derivation of the training equations can be found in [5].

3.1.1 E-step

In the E-step, we estimate the posterior over all hidden variables, using the "old" values of GTM parameters. Given a data point $\mathbf{t}_n \in \mathcal{D}$, compute the model responsibilities corresponding to the competition among models belonging to the same parent,

$$P(\mathcal{M}|\ Parent(\mathcal{M}), \mathbf{t}_n) = \frac{\pi(\mathcal{M}|\ Parent(\mathcal{M}))\ P(\mathbf{t}_n|\ \mathcal{M})}{\sum_{\mathcal{N} \in [\mathcal{M}]} \pi(\mathcal{N}|\ Parent(\mathcal{M}))\ P(\mathbf{t}_n|\ \mathcal{N})}, \tag{5}$$

where $[\mathcal{M}] = Children(Parent(\mathcal{M}))$.

Imposing $P(Root|\ \mathbf{t}_n) = 1$, the unconditional (on parent) model responsibilities are recursively determined by

$$P(\mathcal{M}|\ \mathbf{t}_n) = P(\mathcal{M}|\ Parent(\mathcal{M}), \mathbf{t}_n)\ P(Parent(\mathcal{M})|\ \mathbf{t}_n). \tag{6}$$

Responsibilities of the latent space centers $\mathbf{x}_i$, $i = 1, 2, ..., K$, corresponding to the competition among the latent space centers within each model $\mathcal{M}$, are calculated using (2).

3.1.2 M-step

In the M-step, we estimate the free parameters using the posterior over hidden variables computed in the E-step.

Parent-conditional mixture coefficients are determined by

$$\pi(\mathcal{M}|Parent(\mathcal{M})) = \frac{\sum_{n=1}^{N} P(\mathcal{M}|\ \mathbf{t}_n)}{\sum_{n=1}^{N} P(Parent(\mathcal{M})|\ \mathbf{t}_n)}. \tag{7}$$

Weight matrices **W** are calculated by solving (using the pseudoinverse to allow for possible ill-conditioning)

$$(\mathbf{\Phi}^T\ \mathbf{B}\ \mathbf{\Phi})\ \mathbf{W}^T = \mathbf{\Phi}^T\ \mathbf{R}\ \mathbf{T}, \tag{8}$$

where $\mathbf{\Phi}$ is a $K \times M$ matrix with elements $(\mathbf{\Phi})_{ij} = \phi_j(\mathbf{x}_i)$, $\mathbf{T}$ is a $N \times D$ matrix storing the data points $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_N$ as rows, $\mathbf{R}$ is a $K \times N$ matrix containing, for each latent space center $\mathbf{x}_i$, and each data point $\mathbf{t}_n$, *scaled responsibilities* $(\mathbf{R})_{in} = P(\mathcal{M}|\ \mathbf{t}_n)\ R_{i,n}$, and $\mathbf{B}$ is a $K \times K$ diagonal matrix with diagonal elements corresponding to responsibilities of latent space centers for the whole data sample, $(\mathbf{B})_{ii} = \sum_{n=1}^{N} P(\mathcal{M}|\ \mathbf{t}_n)\ R_{i,n}$.

The inverse variances are re-estimated using

$$\frac{1}{\beta} = \frac{\sum_{n=1}^{N} P(\mathcal{M}|\ \mathbf{t}_n)\ \sum_{i=1}^{K}\ R_{i,n}\ \ \|\mathbf{W}^{new}\ \phi(\mathbf{x}_i) - \mathbf{t}_n\|^2}{D\ \sum_{n=1}^{N} P(\mathcal{M}|\ \mathbf{t}_n)}, \tag{9}$$

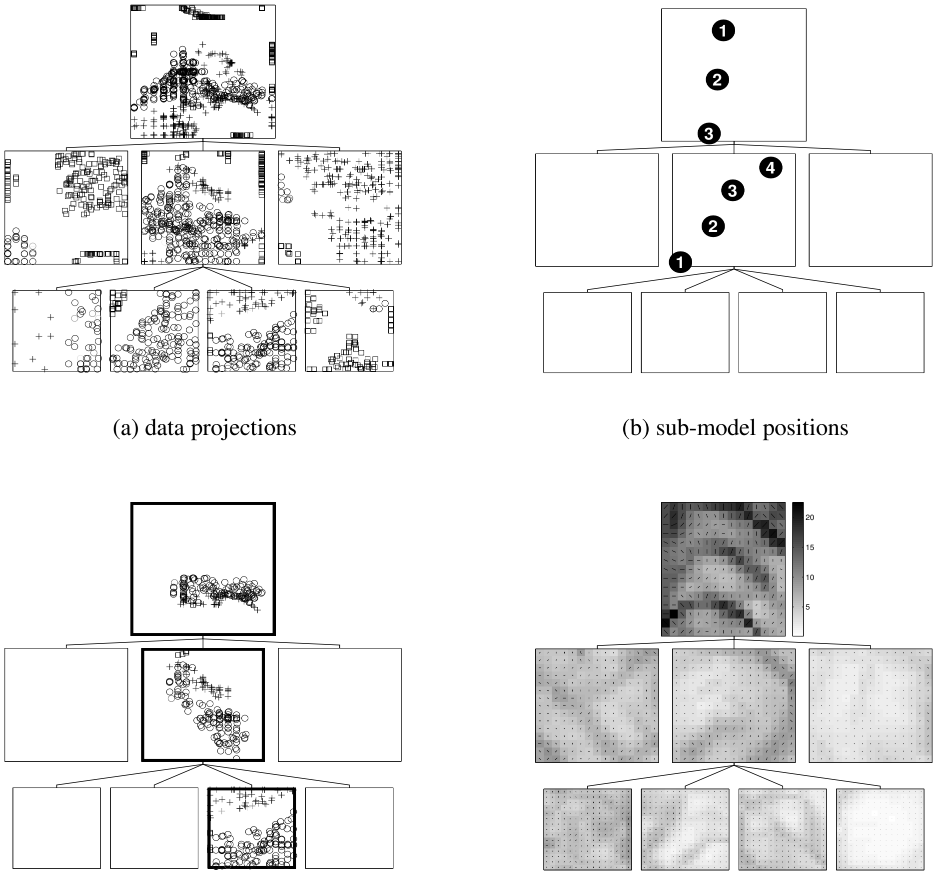
where $\mathbf{W}^{new}$ is the "new" weight matrix computed from (8).

To make expressions for training individual models $\mathcal{M}$ consistent throughout the hierarchy, we introduce a virtual model $Parent(Root)$ by postulating $\pi(Root|Parent(Root)) = 1$, $Children(Parent(Root)) = \{Root\}$, and $P(Parent(Root)|\ \mathbf{t}_n) = 1$. Details concerning parameter initialization and regularization can be found in [5].

4 Experiments

4.1 Hierarchical visualization of oil flow data

We illustrate our system on a 12-dimensional oil flow data set generated from a physics-based simulation of a non-invasive monitoring system, used to determine the quantity of oil in a multi-phase pipeline containing a mixture of oil, water and gas. The data set consists



(a) data projections

(b) sub-model positions

(c) child-modulated ancestor plots

(d) directional curvatures

Figure 1: Visualization of oil flow data. Points shown as circles, pluses and squares correspond to flow configurations *homogeneous*, *annular* and *laminar*, respectively.

of 1000 points obtained synthetically by simulating the physical process in the pipe. Points in the data set are classified into three different multi-phase flow configurations, namely *homogeneous*, *annular* and *laminar*.

The projection plot is organized in a hierarchical tree $\mathcal{T}$ shown in figure 1(a). If a node (subplot) in figure 1(a) is not a leaf, we show in the corresponding node in figure 1(b) the latent space points $\mathbf{c}_i$ that were chosen to be the "centers" of the regions of interest for the child GTMs (see section 3.1). The centers are shown as circles labeled by numbers. The numbers determine the order of the corresponding child GTM subplots (left-to-right).

We make use of the probabilistic character of GTM and plot all the data points on every plot, but modify the intensity in proportion to the responsibility which each plot (submodel $\mathcal{M}$) has for the data point $\mathbf{t}_n$. Points that are not well captured by $\mathcal{M}$ will appear with low intensity. The user can visualize the regions captured by a particular child GTM $\mathcal{M}$, by modifying the plot of its parent (or, alternatively, the plots of all its ancestors), so that instead of the parent responsibilities, the responsibilities of the model $\mathcal{M}$ are used. This is done by simply clicking with a mouse on a chosen child GTM plot. As an example, we show in figure 1(c) points captured by the third level-three GTM $\mathcal{M}$ in the hierarchy $\mathcal{T}$. In the *Root* plot, points visualized by $\mathcal{M}$ form a dense cluster, with overlapping *homogeneous* and *annular* classes. The situation improves a bit at level two, but only at level three are the two classes nicely separated.

The hierarchical structure of plots used for plotting the GTMs' projections is also used to show the local directional curvatures (figure 1(d)). For every GTM $\mathcal{M}$ and each latent space center $\mathbf{x}_i$ of $\mathcal{M}$, we evaluate the local directional curvatures of the projection manifold Ω (see section 2) at the mapped centers $f(\mathbf{x}_i)$ with respect to latent space directions $\mathbf{h}_j$, $j = 1, 2, ..., N_h$ (see [2]). The directions $\mathbf{h}_j$, $j = 1, 2, ..., N_h$, correspond to the N_h equidistant points on the unit circle, subject to the constraint that the first direction is $(1, 0)$. We set $N_h = 16$. In the final plot we show, for each latent space center $\mathbf{x}_i$, the maximal norm of the curvature across the different "probing" directions $\mathbf{h}_j$. The direction of the maximal curvature is shown as a black line of length proportional to the curvature's value. The intensity of curvatures in the hierarchy of GTMs is scaled by the minimal and maximal curvatures found in the whole hierarchy $\mathcal{T}$. A locally scaled plot of curvatures can be obtained by clicking on a chosen plot corresponding to a local GTM. Although not shown here, we visualize in a similar manner local magnification factors of GTMs in the hierarchy $\mathcal{T}$.

The curvature plot of the $Root$ GTM in figure 1(d) reveals that the two-dimensional projection manifold folded three times in order to "capture" the distribution of points in the 12-dimensional space. Interestingly, the three flow configurations seem to be roughly separated by the folds (compare the top level visualization plot in figure 1(a) with the corresponding curvature plot). We confirmed this hypothesis by constructing three local second-level visualization plots initialized in the regions between the folds. The second plot of the second-level projections was further detailed in four third-level plots. Curvature (and magnification factor – not shown here) plots of the lower level GTMs reveal that, compared with the $Root$ GTM, the lower level projection manifolds do not significantly stretch/contract and are almost flat.

4.2 Investigation of pharmaceutical data through visualization

In co-operation with Pfizer Global Research and Development we performed a series of experiments with a data set of molecular compounds (represented as 16 dimensional feature vectors) labelled by their biological activity against a set of targets. The data set contained all 16 dimensional feature vectors for a total of 49500 chemical entities. It did not contain any descriptors of the chemical structure, but contained features describing biological activity in 4 tests in addition to a number of whole molecule properties for the chemical entities (compounds) e.g. Molecular Weight, cLogP, number of atoms.

Interpretation of the GTM visualizations showed areas of particular interest where compounds, which demonstrated an effect against one or more biological targets, were spatially clustered with compounds that showed little or no effect. The clusters of compounds could be described by 3 GTM sub-models and totalled 34 specific compounds. The models constructed by the GTM technique did not use any structural descriptors. These groups of compounds were investigated further. The full chemical descriptors for the structure were obtained and using a structure based clustering technique the 34 compounds were analyzed using a proprietary clustering tool which determines homology of compounds based on the most prominent ring system. This tool partitioned the 34 compounds into 3 groups that had significant overlap with the clustering observed in the GTM visualizations. The key distinction between the two techniques is that the GTM utilizes mainly biological data and no structural information and the proprietary clustering tool uses only structural descriptors and no biological information.

Preliminary conclusions suggests that the hierarchical visualization using GTM provide meaningful clustering of compound related data based on biological information and a limited number of physiochemical parameters. But more importantly the analysis suggests that there is potential for clustering compounds based on biological data which is meaningful in the light of grouping by purely structural parameters. This implies some similarity between how a Medicinal Chemist would view the compounds and the GTM visualizations.

Further work will be performed to determine the application of the GTM visualizations based on biological data for identification of compounds that are falsely positive in a biological test, or are incorrectly classified by structural clustering algorithms.

5 Discussion

This study was motivated by the fact that most visualization algorithms projecting points from a high-dimensional data space onto a two-dimensional projection space suffer from two major drawbacks: **(1)** A single visualization plot is often not sufficient. Encompassing all the data points in a single low-dimensional projection can make the plot confusing and difficult to read. **(2)** The low dimensional non-linear projection manifold onto which the data points are projected can form complicated folds and/or significant contractions/stretchings in the high-dimensional data space. When presented with the projection plot alone, it is difficult for the user to judge the actual "layout" of the points in the data space.

Instead of a single plot, we construct a whole *hierarchy* of localized *non-linear* visualization plots by extending the locally linear hierarchical visualization system PhiVis [1] to allow for non-linear projection manifolds. Like PhiVis, our system is *statistically principled* and is built *interactively* in a top-down fashion using the EM algorithm. Compared to linear projection methods such as PCA, non-linear visualization techniques are more capable of revealing the nature of data distribution in a high-dimensional space.

But, as mentioned above (point **(2)**), there is a price to pay for going from linear to non-linear visualization methods. We equip the user with first- and second-order geometric information in the form of local magnification factors and directional curvatures of the projection manifold. Local curvature and magnification factor patterns provide information about dominant folds and contraction/expansion regions of the manifold. This information is helpful for **(1)** determining the geometry of the projection manifold in the high-dimensional data space, **(2)** changing the amount of regularization in non-linear projection manifolds, and for **(3)** choosing regions of interest when constructing detailed lower-level visualization plots.

Taking advantage of the probabilistic character of local GTMs we can, for each lower-level plot $\mathcal{M}$, modulate its ancestor plots by shading the projections of data points $\mathbf{t}$ according to the responsibilities $P(\mathcal{M}|\mathbf{t})$. Understanding where the points in a lower-level plot come from in the higher-level plots helps the user to relate plots at lower levels with their ancestors in the visualization hierarchy.

References

[1] C.M. Bishop and M.E. Tipping. A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), pp. 281–293, 1998.

[2] C.M. Bishop, M. Svensén and C.K.I. Williams. GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1), pp. 215–235, 1998.

[3] C.M. Bishop, M. Svensén and C.K.I. Williams. Magnification Factors for the GTM Algorithm. *Proceedings IEE Fifth International Conference on Artificial Neural Networks*, IEE, London, pp. 64–69, 1997.

[4] P. Tiňo, I. Nabney and Yi Sun. Using Directional Curvatures to Visualize Folding Patterns of the GTM Projection Manifolds. In *Artificial Neural Networks - ICANN 20001*, (eds) G. Dorffner, H. Bischof and K. Hornik. Springer-Verlag, pp. 421-428, 2001.

[5] P. Tiňo and I. Nabney. Hierarchical GTM: constructing localized non-linear projection manifolds in a principled way. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in print, 2001.