A General Framework for a Principled Hierarchical Visualization of Multivariate Data

Ata Kabán[1], Peter Tiňo[2] and Mark Girolami[3]

[1] Department of Information Systems, Eötvös Loránd University, Budapest, Hungary
[2] Neural Computing Research Group, Aston University, Birmingham B4 7ET, UK
[3] School of Information & Communication Technology, University of Paisley
ata@ullman.inf.elte.hu, tinop@aston.ac.uk, mark.girolmi@paisley.ac.uk

Abstract. We present a general framework for interactive visualization and analysis of multi-dimensional data points. The proposed model is a hierarchical extension of the latent trait family of models developed in [4] as a generalization of GTM to noise models from the exponential family of distributions. As some members of the exponential family of distributions are suitable for modeling discrete observations, we give a brief example of using our methodology in interactive visualization and semantic discovery in a corpus of text-based documents. We also derive formulas for computing local magnification factors of latent trait projection manifolds.

1 Introduction

The topographic visualization of multi-dimensional data has been an important method of data analysis and data mining [1, 4, 5]. Hierarchical extensions of topographic visualization methods [2, 7, 9] have been developed for capturing informative aspects of the data in a more flexible manner: successive 2-D plots are interactively constructed to account for the level of detail required by the user. In particular, hierarchical GTM [9] has been proposed as a statistically principled construction of local non-linear projection manifolds. However, by imposing a Gaussian noise model in the original GTM, the hierarchical GTM is restricted to continuous data types. Similar tools are needed for other data types, such as discrete data. Application examples may include organization of digital libraries encoded in binary and/or count-based arrays.

Building upon the latent trait models [4] — a generalization of GTM to noise models from the exponential family of distributions - we develop a general probabilistic framework for interactive hierarchical visualization of multi-dimensional data sets in a top-down fashion. Some members of the exponential family of distributions are suitable for modeling discrete observations and we briefly elaborate upon the possibility of using our methodology in interactive visualization and semantic discovery in collections of text-based documents.

We also give formulas for computing local geometric properties, namely magnification factors, of latent trait projection manifolds, which, due to the smooth character of the latent trait mapping, can be computed using the tools of differential geometry. Magnification factors provide potentially valuable additional visual information regarding the structure of the data and may aid the user to better understand the visualization plot and provide useful hints in the process of interactive construction of deeper levels of the hierarchy (see [1, 9]).

2 Probabilistic Model of Hierarchic Visualization

2.1 Latent Trait Models

The latent trait model [4] is a generalized version of GTM capable of modeling outcomes from the exponential family of distributions.

Consider an L-dimensional latent space, which, for visualization purposes is typically a bounded 2-D Euclidean domain, e.g. $[-1,1] \times [-1,1]$. The aim is to represent, using the latent space, multi-dimensional data points $\{\boldsymbol{d}_n\}_{n=1..N}$, such that the "important" structural characteristics are revealed. For tractability reasons, the latent space is discretized into a regular grid of K latent points $\boldsymbol{x}_1, ...\boldsymbol{x}_K$ (which are analogous to the nodes of the SOM [5]). The uniform latent space prior distribution is a mixture of delta functions and so $p(\boldsymbol{d})$ can be evaluated as $p(\boldsymbol{d}) = \sum_{k=1}^{K} p(\boldsymbol{d}|\boldsymbol{x}_k)p(\boldsymbol{x}_k)$.

The conditional distribution over the data is modeled as a member of the exponential family in a parameterized functional form

$$p_G(\boldsymbol{d}_n|\boldsymbol{x}_k, \boldsymbol{A}) = \exp\left\{\boldsymbol{f}_{\boldsymbol{A}}(\boldsymbol{x}_k)\boldsymbol{d}_n - G(\boldsymbol{f}_{\boldsymbol{A}}(\boldsymbol{x}_k))\right\} p_0(\boldsymbol{d}_n), \tag{1}$$

where $G(\cdot)$ denotes the cumulant function, $\boldsymbol{d}_n \in \mathcal{R}^T$ denotes the n-th observed data point, $n = 1 : N$, $\boldsymbol{x}_k \in \mathcal{R}^L$ is the k-th sample from a uniformly distributed latent variable over the continuous visualization space $[-1,1]^2$, $k = 1 : K$, and the nonlinearity $\boldsymbol{f}(.)$ is (conveniently) chosen to be of the form $\boldsymbol{f}_{\boldsymbol{A}}(\boldsymbol{x}_k) = \boldsymbol{A}\boldsymbol{\phi}(\boldsymbol{x}_k)$, where $\boldsymbol{A} \in \mathcal{R}^{T\times M}$ is a parameter matrix and $\boldsymbol{\phi}(.) = (\phi_1(.), ..., \phi_M(.))^T, \phi_m(.) : \mathcal{R}^L \to \mathcal{R}$ is a set of M non-parametric nonlinear smooth basis functions (typically RBFs). This formulation allows for a convenient formalism which includes a principled treatment of discrete observations. Projection of a data point $\boldsymbol{d}$ in the latent space is obtained as the mean of the posterior distribution over the latent space, given the observation $\boldsymbol{d}$. Details regarding the derivation of an EM-based parameter estimation algorithm and application examples for some of the discrete members of this model family can be found in [4].

2.2 Hierarchical Latent Trait Models

When dealing with large and complex data sets, a single global visualization plot is often not sufficient. Therefore we introduce a hierarchical mixture extension of the latent trait model family. A set of latent trait projections is arranged in a tree structure in a recursive top-down fashion [2, 3, 9]: provided that one has already trained models up to level $\ell - 1$ (by convention, the *root*'s level is 1), training of the ℓ-th level models proceeds on the data indicated by the user as "regions of interest" on the previous level parent plot (see [2, 9]).

Denote by c_r the indicator variables for tree nodes at level r. In particular, $c_1 = root$. $c_{1:r} = (c_1, c_2, ..., c_r)$ denotes a path from the *root* to a level-r node identified by c_r. Given a particular node $\mathcal{N}$ at level r, i.e. $c_r = \mathcal{N}$, there is only one path $c_{1:r}$ in the tree from the *root* to $\mathcal{N}$, such that $c_r = \mathcal{N}$.

The expectation of the *complete* data likelihood of an ℓ-level hierarchical model, under the *joint* posterior density of latent variables $c_{1:\ell}$ and $\boldsymbol{x}_k$ is

$$Q^{(\ell)} = \sum_{n=1}^{N} \sum_{c_{1:\ell}} \sum_{k} p(c_{1:\ell}|\boldsymbol{d}_n) p(\boldsymbol{x}_k|c_\ell, \boldsymbol{d}_n) \log \left\{ p(c_{1:\ell}) p(\boldsymbol{x}_k|c_\ell) p(\boldsymbol{d}_n|\boldsymbol{x}_k, \boldsymbol{A}, c_\ell) \right\},$$

where the tree-structure leads to the following recursive decompositions:

$$p(c_{1:\ell}|\boldsymbol{d}_n) = p(c_{1:\ell-1}|\boldsymbol{d}_n) p(c_\ell|c_{\ell-1}, \boldsymbol{d}_n) \quad \text{and} \quad p(c_{1:\ell}) = p(c_{1:\ell-1}) p(c_\ell|c_{\ell-1}).$$

Maximizing $Q^{(\ell)}$ under the constraints $\sum_k p(\boldsymbol{x}_k|c_\ell) = 1$ and $\sum_{c_l} p(c_\ell|c_{\ell-1}) = 1$, we obtain the following updates for the parameters (**M-step**)

$$p(c_\ell|c_{\ell-1}) = \frac{\sum_{n=1}^{N} p(c_{1:\ell-1}|\boldsymbol{d_n}) p(c_\ell|c_{\ell-1}, \boldsymbol{d}_n)}{\sum_{n=1}^{N} p(c_{1:\ell-1}|\boldsymbol{d}_n)} \tag{2}$$

$$\boldsymbol{g}(\boldsymbol{A\Phi}) \boldsymbol{G}^{(\ell)} \boldsymbol{\Phi}^T = \boldsymbol{D} \boldsymbol{R}^{(\ell)T} \boldsymbol{\Phi}^T, \qquad p(\boldsymbol{x}_k|c_\ell) = \frac{1}{N} \sum_{n=1}^{N} r_{kn}^{(\ell)} \tag{3}$$

where $r_{kn}^{(\ell)} = p(c_{1:\ell-1}|\boldsymbol{d}_n) p(c_\ell|c_{\ell-1}, \boldsymbol{d}_n) p(\boldsymbol{x}_k|c_\ell, \boldsymbol{d}_n)$, $\boldsymbol{R}^{(\ell)} = (r_{kn}^{(\ell)})_{k=1:K, n=1:N}$, $\boldsymbol{G}^{(\ell)}$ is a diagonal matrix with elements $g_{kk}^{(\ell)} = \sum_{n=1}^{N} r_{kn}^{(\ell)}$, $\boldsymbol{\Phi}$ is an $M \times K$ matrix the k-th column of which is $\boldsymbol{\phi(x_k)}$ and $\boldsymbol{D}$ is the data matrix. The factor $p(c_{1:\ell-1}|\boldsymbol{d}_n)$ is the posterior "inherited" from the previous level. For consistency of the training equations, we adopt the convention that $p(c_1|\boldsymbol{d}_n) \equiv 1, p(c_1|c_0) \equiv 1$ and $p(c_0|\boldsymbol{d}_n) \equiv 1$. The (non-linear) inverse link function $\boldsymbol{g}(.)$ is the derivative of the cumulant function $G(.)$ [6]. In general, non-linear optimization techniques are needed to solve for $\boldsymbol{A}$ in (3). Posterior estimates over the hidden variables are computed using the "old" values of the parameters (**E-step**)

$$p(c_\ell|c_{\ell-1}, \boldsymbol{d}_n) = \frac{p(c_\ell|c_{\ell-1}) p(\boldsymbol{d}_n|c_\ell)}{\sum_{c'_\ell} p(c'_\ell|c_{\ell-1}) p(\boldsymbol{d}_n|c'_\ell)} \tag{4}$$

$$p(\boldsymbol{x}_k|c_\ell, \boldsymbol{d}_n) = \frac{p(\boldsymbol{x}_k|c_\ell) p(\boldsymbol{d}_n|\boldsymbol{x}_k, c_\ell, \boldsymbol{A})}{\sum_{k'} p(\boldsymbol{x}_{k'}|c_\ell) p(\boldsymbol{d}_n|\boldsymbol{x}_{k'}, c_\ell, \boldsymbol{A})} \tag{5}$$

At each level, the E and M steps are iterated until convergence to a local optimum.

3 Local Geometric Analysis of the Latent Trait Manifolds

The probabilistic latent trait model defines a density model (1), using a smooth mapping from the latent space into the 'data space' (projection space) $\boldsymbol{z} : \mathcal{R}^L \rightarrow \mathcal{R}^T, \boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{A\Phi}(\boldsymbol{x}))$, where $\boldsymbol{g}()$ is the the inverse link function. We refer to the image of the latent space under $\boldsymbol{z}$ as the *projection manifold* [9]. For visualization purposes, $L = 2$, and so the projection manifold is a 2-D surface embedded in the data space $\mathcal{R}^T$.

The term *magnification factor* [1] refers to the degree of stretching/compression of the latent space when embedded into the data space. Consider a Cartesian coordinate system defined in the latent space and the corresponding curvilinear coordinate system on the projection manifold. The local magnification factor corresponding to a latent space point $\boldsymbol{x}_0$ is the ratio between the areas of an infinitesimal rectangle in the latent Cartesian space and its image (under $\boldsymbol{z}$) in the projection manifold. It is computed as ([1]) $\sqrt{|\boldsymbol{T}(\boldsymbol{x}_0)|}$, where $|\boldsymbol{T}(\boldsymbol{x}_0)|$ is the determinant of the metric tensor $\boldsymbol{T}(\boldsymbol{x}_0) = \boldsymbol{\Gamma}^T \boldsymbol{\Gamma}$ [9], with $\boldsymbol{\Gamma}$ denoting (the transpose of) the Jacobian of the mapping $\boldsymbol{z}$. In our case

$$\boldsymbol{\Gamma} = \frac{\partial \boldsymbol{z}(\boldsymbol{x}_0)}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{g}(\boldsymbol{A}\boldsymbol{\phi}(\boldsymbol{x}_0))}{\partial \boldsymbol{x}} = \boldsymbol{Q}\boldsymbol{A}\boldsymbol{V}, \tag{6}$$

where $\boldsymbol{V} = \left(\frac{\partial \phi_m(\boldsymbol{x})}{\partial x_l}{}_{|\boldsymbol{x}=\boldsymbol{x}_0}\right)_{m=1:M,l=1:L}$. The Fisher information matrix $\boldsymbol{Q} = \left(\frac{\partial g_{t'}(\boldsymbol{y})}{\partial y_t}{}_{|\boldsymbol{y}=\boldsymbol{A}\boldsymbol{\phi}(\boldsymbol{x}_0)}\right)_{t'=1:T,t=1:T}$ happens to be the identity matrix only in the special case of the original GTM with spherical Gaussian noise model. In this case the magnification factors derived here coincide with those previously obtained for the original GTM model [1].

4 Hierarchical Visualization of Text Documents

We demonstrate our system on an interactive visualization of a text-collection formed by 10 topic classes from the newsgroups[4] text corpus. 800 instances were taken from each topic. The instances were binary encoded over a dictionary of $T = 300$ words [8]. The initial preprocessing, word-stemming and removal of 'stop-words' was done using the Bow toolkit[5]. To account for binary encoding, the Bernoulli noise model was employed at the expert nodes of our model, i.e. the cumulant function had the form $G(\boldsymbol{A\Phi}(\boldsymbol{x})) = \log\{1 + exp(\boldsymbol{A\Phi}(\boldsymbol{x}))\}$. The latent (topics') space has been discretized into a $K = 15 \times 15$ grid, and $M = 16$ spherical RBF basis functions of unit width were positioned on a regular 4×4 grid. An example of hierarchically organized projection plot of the posterior means is shown in Figure 1. The *root* plot contains a dense cluster with overlapping projection classes and is recursively detailed at deeper levels of the hierarchy. The level of detail in the leaf plots is solely dictated by the user and his idea of "interesting" areas in the data space.

In the case of the Bernoulli model, the proximity of local regions of the low-dimensional projection manifold to vertices of the T-dimensional hypercube $\{0, 1\}^T$ determines what areas of the latent space will be devoted to which keyword occurrences. Hence, when two documents are mapped to neighboring positions in the latent plane, but the area between them is characterized by high

[4] http://www.cs.cmu.edu/˜textlearning
[5] http://www-2.cs.cmu.edu/˜mccalum/bow/

magnification factors, the user can infer that the two documents correspond to different clusters. This leads recursively to ("semantically-driven") selection of regions of interest for more and more detailed plots on lower levels of the visualization hierarchy[6]. Other tools aimed at improving our understanding of the plots, like listing the most probable dictionary words for each latent point $\mathbf{x_k}$ [4], are also incorporated in the system.

5 Conclusions

We have developed a statistically principled general framework for interactive visualization and analysis of multi-dimensional data in a hierarchical manner and generalized the local geometric analysis of projection manifolds (magnification factors) to accommodate for non-Gaussian members of the exponential noise model family. We demonstrated a topic-driven interactive visualization of a real-world text corpus.

Acknowledgements

P. Tiňo is supported by the BBSRC grant BIO/12093 and Pfizer Research. A. Kabán and M. Girolami has been funded by *re*:source, The Council for Museums Archives and Libraries, Grant Number RE/092.

References

1. C. Bishop, M. Svensén, and C. Williams, "Magnification factors for the SOM and GTM algorithms," in *Proceedings 1997 Workshop on Self-Organizing Maps, Helsinki, Finland*, 1997.
2. C. Bishop and M. E. Tipping, "A Hierarchic Latent Variable Model for data Visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), pp. 281–293, 1998.
3. M. I. Jordan, "Hierarchical Mixture of Experts and the EM Algorithm". *Neural Computation*, (6), pp.181–214, 1994.
4. A. Kabán and M. Girolami, "A combined latent class and trait model for the analysis and visualization of discrete data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8), pp. 859 -872, 2001.
5. T. Kohonen, *Self-organized formation of topographically correct feature maps.* Biological Cybernetics, vol.43, pp. 59–69, 1982.
6. P. McCullagh and L. Nelder, *Generalized Linear Models.* Chapman and Hall, 1985.
7. R. Miikkulainen, *Script recognition with hierarchical feature maps* Connection Science, vol.2, pp. 83–101,1990.
8. Sahami, *Using Machine Learning to Improve Information Access*, PhD Thesis, Stanford University, 1998.
9. P. Tiňo and I. Nabney, "Constructing localized non-linear projection manifolds in a principled way: hierarchical generative topographic mapping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in print.

This article was processed using the LaTeX macro package with LLNCS style

[6] due to the lack of space not demonstrated here

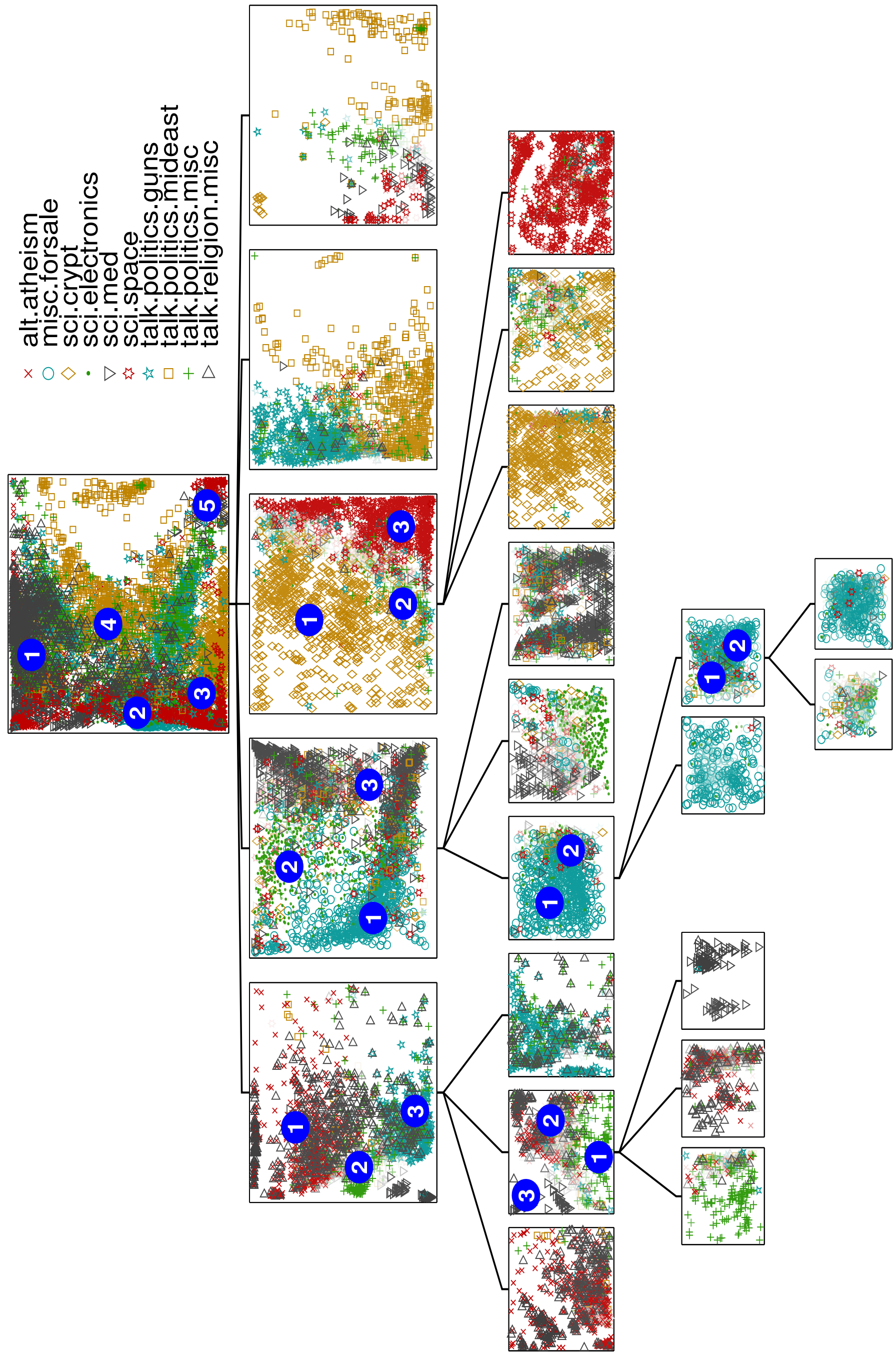

Fig. 1. Hierarchical visualization of text documents marked by 10 discussion topics.