

Advances in Data Visualisation

Ian T. Nabney

8 October 2003



<http://www.ncrg.aston.ac.uk>
Aston University, Birmingham, UK

Overview

- The role of visualisation in scientific data analysis.
- PCA and probabilistic PCA.
- Generative Topographic Mapping (GTM).
 - Magnitude and Curvature plots;
 - Hierarchical GTM;
- Neuroscale.
- Applications in chemometrics and bioinformatics.

Work with Peter Tiño, David d'Alimonte and Yi Sun. Funding from BBSRC and Pfizer Central Research.

Data Visualisation

Visualisation plays a key role in developing good models for data, especially when the quantity of data is large.

- It allows the user to **interact with** and **query** the data more effectively.
- It is an important aid in feature selection, gives information about local deviations in performance and provides a useful 'sanity check' for objective quantitative measures (such as generalisation performance).
- It plays an important role in the search for clusters of similar data points, which are most easily determined by eye.
- The quantity and complexity of many datasets means that simple visualisation methods, such as Principal Component Analysis, are not very effective.

PCA Primer

- A classical **linear** projection method that preserves as much data variance as possible. Fast and easy to compute.
- Suppose that we are trying to map a dataset of vectors \mathbf{x}^n for $n = 1, \dots, N$ in $V = \mathbb{R}^d$ to vectors \mathbf{z}^n in $U = \mathbb{R}^M$, a subspace of V .
- The quality of the approximation is measured by the **residual** sum-of-squares error

$$E = \frac{1}{2} \sum_{n=1}^N \|\mathbf{x}^n - \mathbf{z}^n\|^2 = \frac{1}{2} \sum_{i=M+1}^d \mathbf{u}_i^T \Sigma \mathbf{u}_i,$$

where Σ is the covariance matrix of the data.

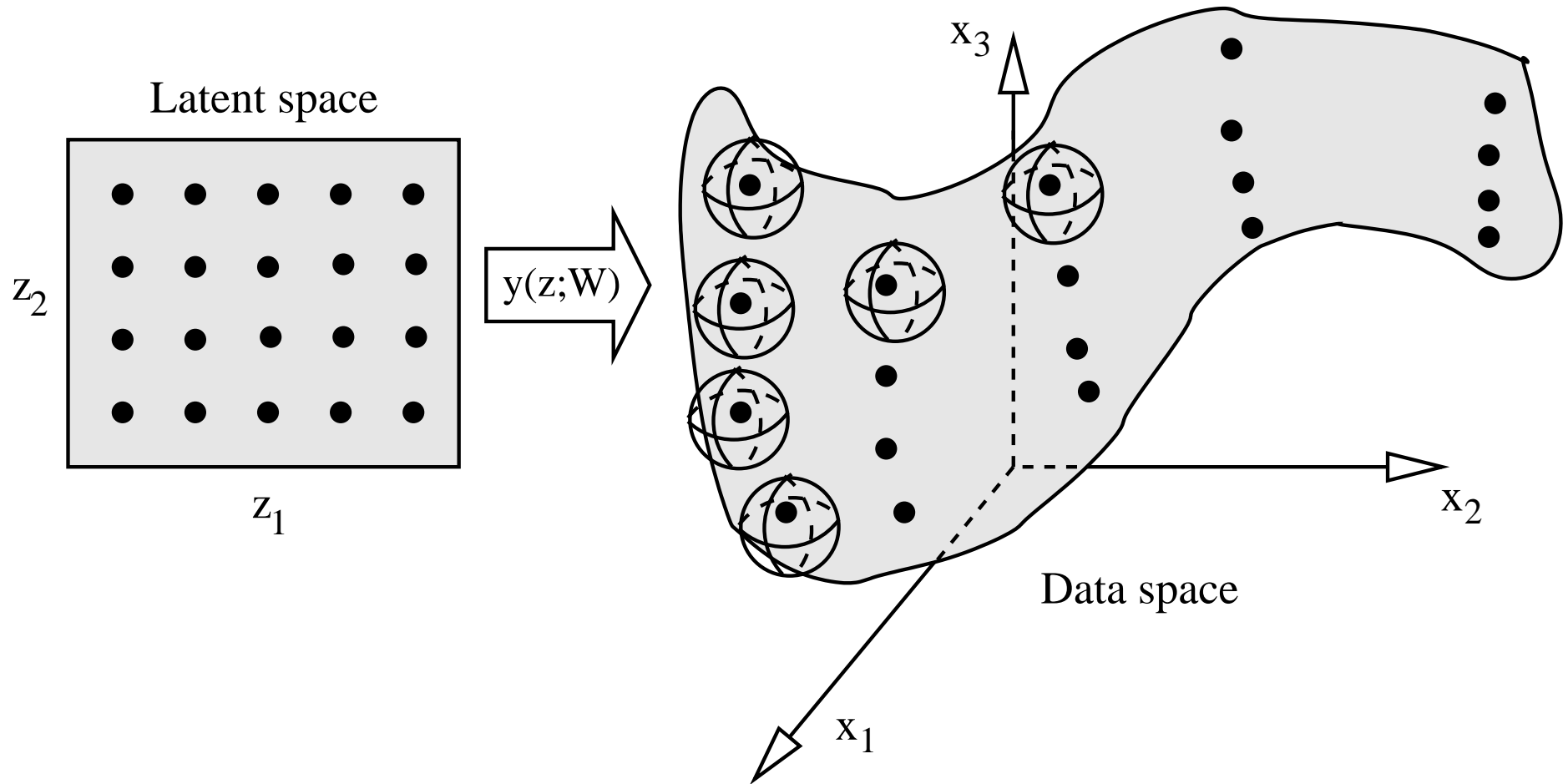
- The minimal error is achieved by projecting the data onto the space spanned by the eigenvectors corresponding to the largest M eigenvalues.
- It is possible to write down a probabilistic version of PCA based on **latent variables**.

Uncertainty

- Real data is noisy
- We are forced to deal with uncertainty, yet we need to be quantitative
- The optimal formalism for inference in the presence of uncertainty is **probability theory**
- We assume the presence of an underlying regularity to make predictions
- We usually use more **general** models than 'classical' statistics: e.g. non-Gaussian, non-linear.

The Generative Topographic Mapping

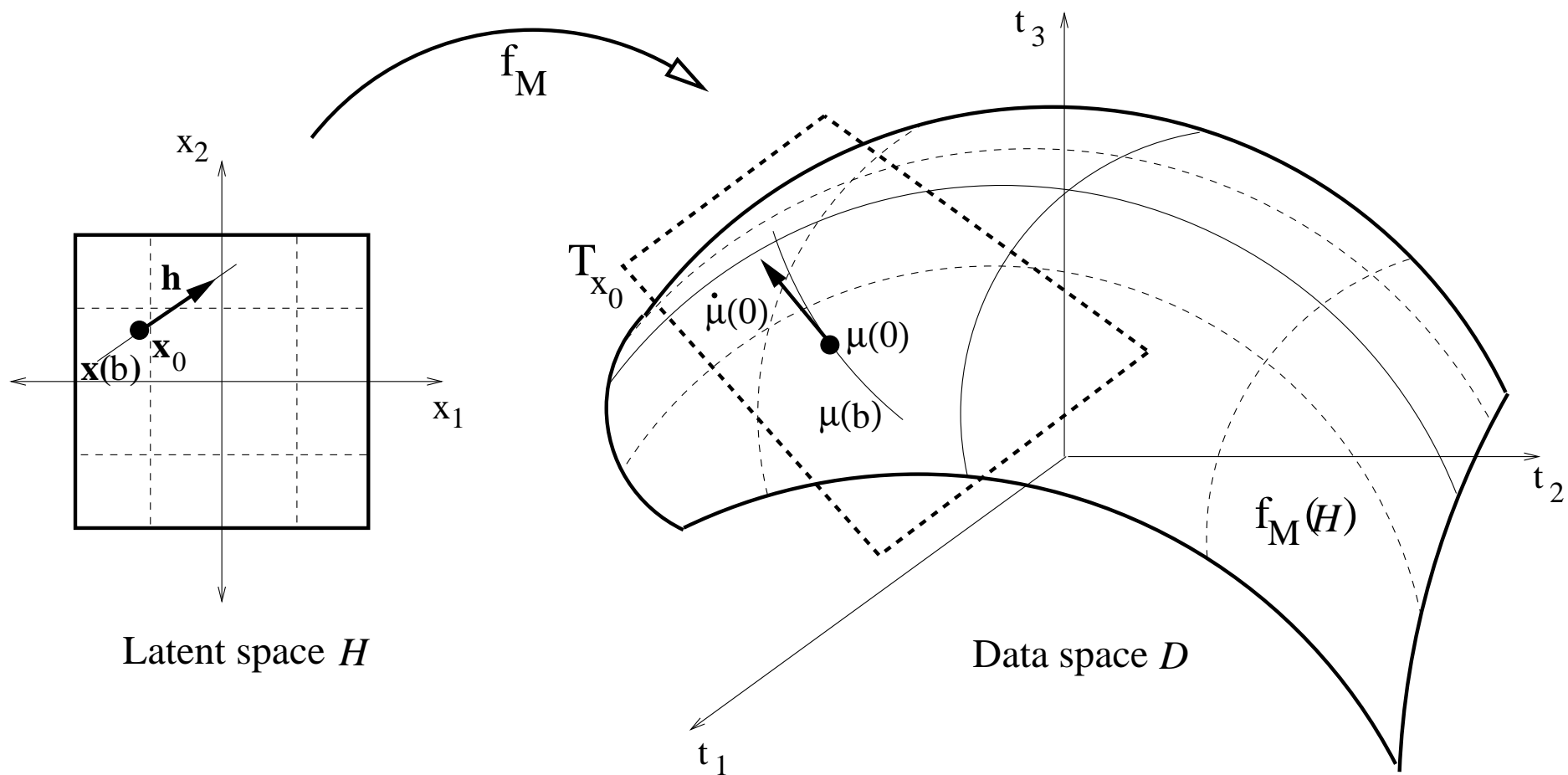
- GTM (Bishop, Svensén and Williams) is a **latent variable model** with a non-linear RBF $f_{\mathcal{M}}$ mapping a (usually two dimensional) latent space \mathcal{H} to the data space \mathcal{D} . This is a generative probabilistic model.
- For the purpose of data visualization, we use Bayes' theorem to invert the transformation $f_{\mathcal{M}}$.
- This model assumes that the data lies close to a two dimensional manifold; however, this is likely to be too simple a model for interesting data.
- We can measure the **stretch** in the sheet using **magnification factors**, and this can be used to detect the gaps between data clusters.



The data is modelled as a **constrained mixture of Gaussians**. GTM can be trained using an EM algorithm that is a generalisation of the standard EM for (unconstrained) Gaussian mixtures.

Stretch and Curvature

- Distribution of data in latent space should be close to uniform.
- Can plot magnification factors to show stretching of manifold
- If J is the RBF Jacobian, factors are $(\det JJ^T)^{1/2}$
- We can measure the **curvature** of the sheet, and this can be used to detect areas where the sheet fits the data poorly.
- This can be done locally, but is a directional measure.
- We can plot magnitude and direction of the largest curvatures to see where the manifold is most folded.

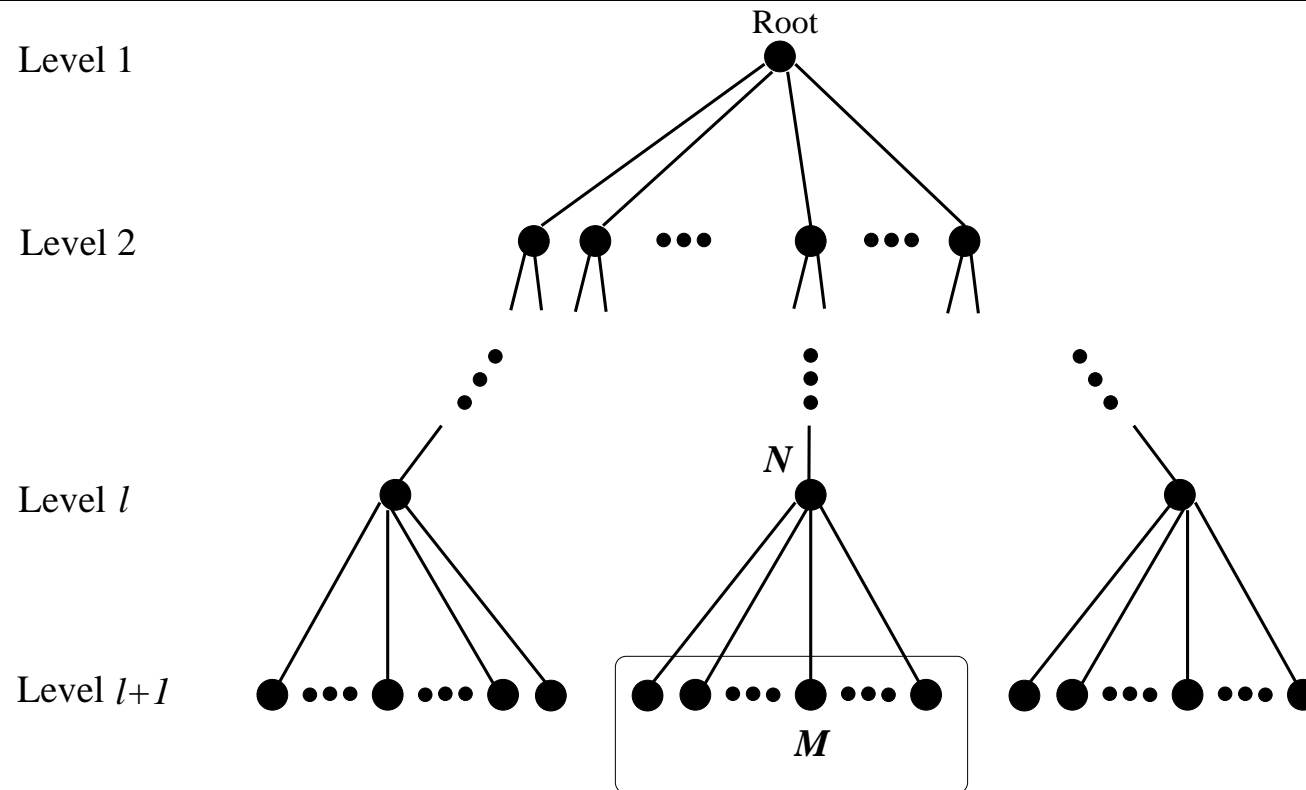


The tangent vector $\dot{\mu}(0)$ to μ at $\mu(0)$ lies in $T_{\mathbf{x}_0}$ (dashed rectangle), the tangent plane of the manifold $f_M(\mathcal{H})$ at $\mu(0)$.

Hierarchical GTM: Drilling Down

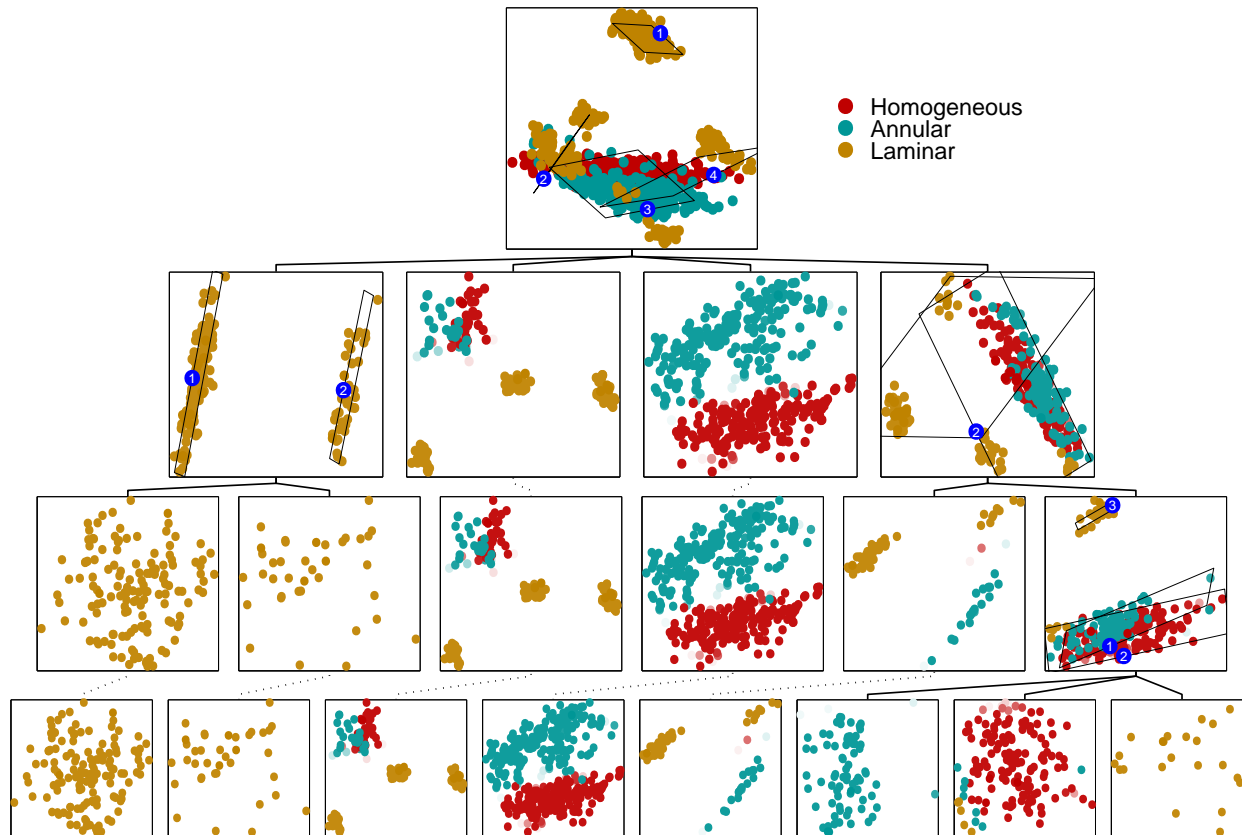
- Bishop and Tipping introduced the idea of hierarchical visualisation for probabilistic PCA. We have developed a general framework for arbitrary latent variable models.
- Because GTM is a generative latent variable model, it is 'straightforward' to train hierarchical mixtures of GTMs.
- We model the whole data set with a GTM at the top level, which is broken down into clusters at deeper levels of the hierarchy.
- Because the data can be visualised at each level of the hierarchy, the selection of clusters, which are used to train GTMs at the next level down, can be carried out interactively by the user.

Tree Structure

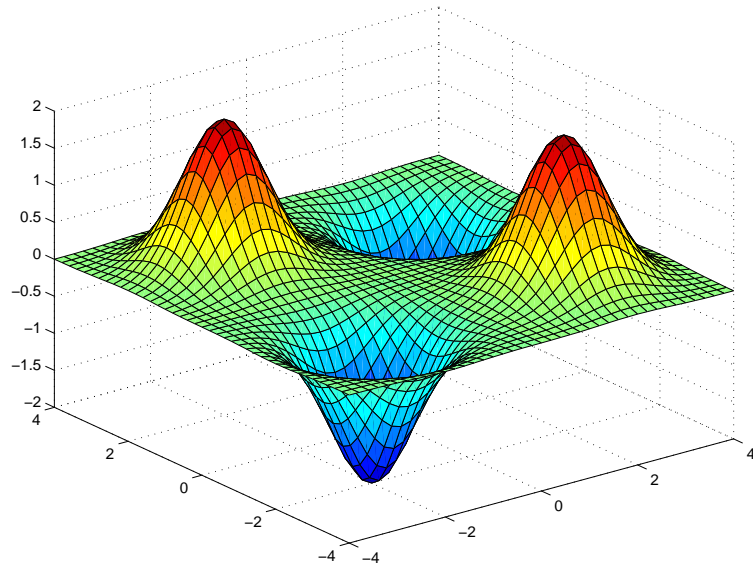


- Models corresponding to internal (i.e. non-leaf) nodes of \mathcal{T} play a role only in the creation of the hierarchical model.
- The tree segments the data: this can be used to develop local models.

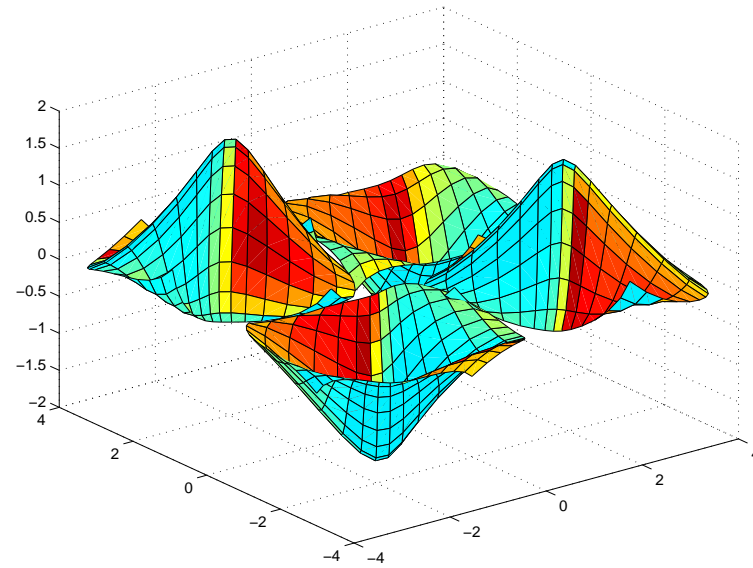
Hierarchical PPCA



Interpretation of Hierarchy



Original Data Generator



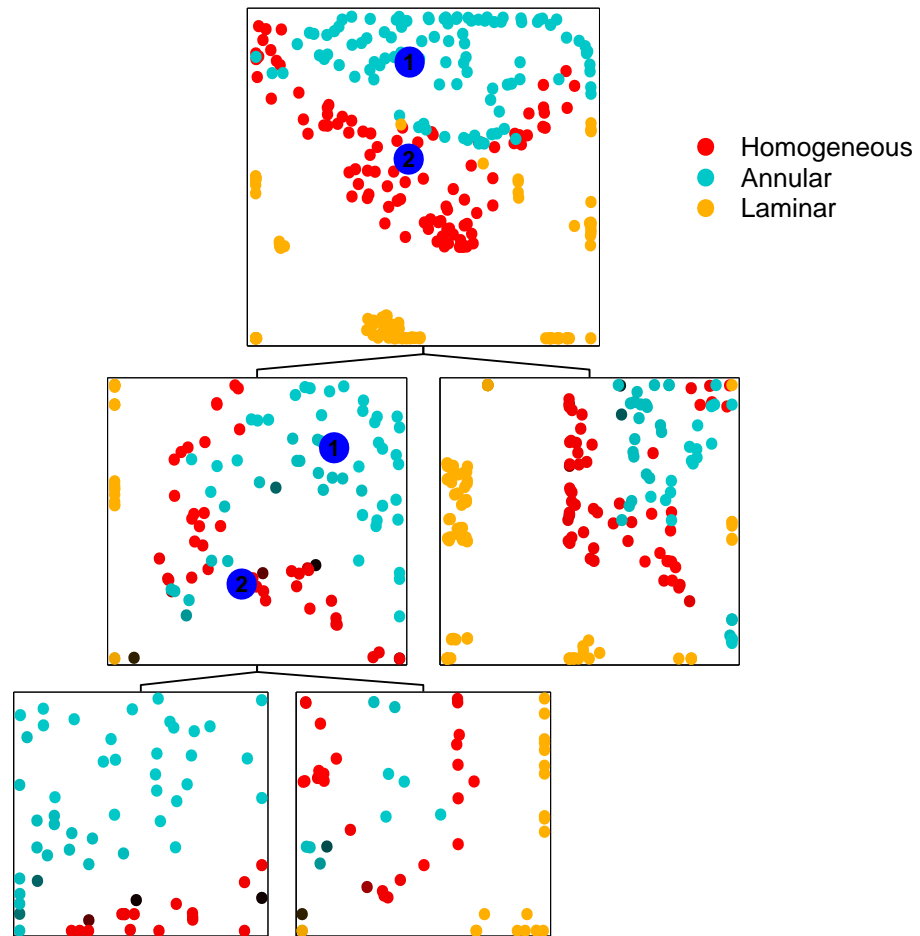
Manifolds of GTM Hierarchy

Training a Hierarchy of GTMs

Training of a hierarchy of GTMs proceeds in a recursive fashion.

1. A root GTM is trained and used to visualize the data.
2. The user identifies interesting regions on the visualization plot.
3. These “regions of interest” are transformed into the data space and form the basis for building a collection of new, child GTMs.
4. The EM algorithm works as before, with responsibilities moderated by the parent-conditional prior.
5. After 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.

Oil Data Revisited

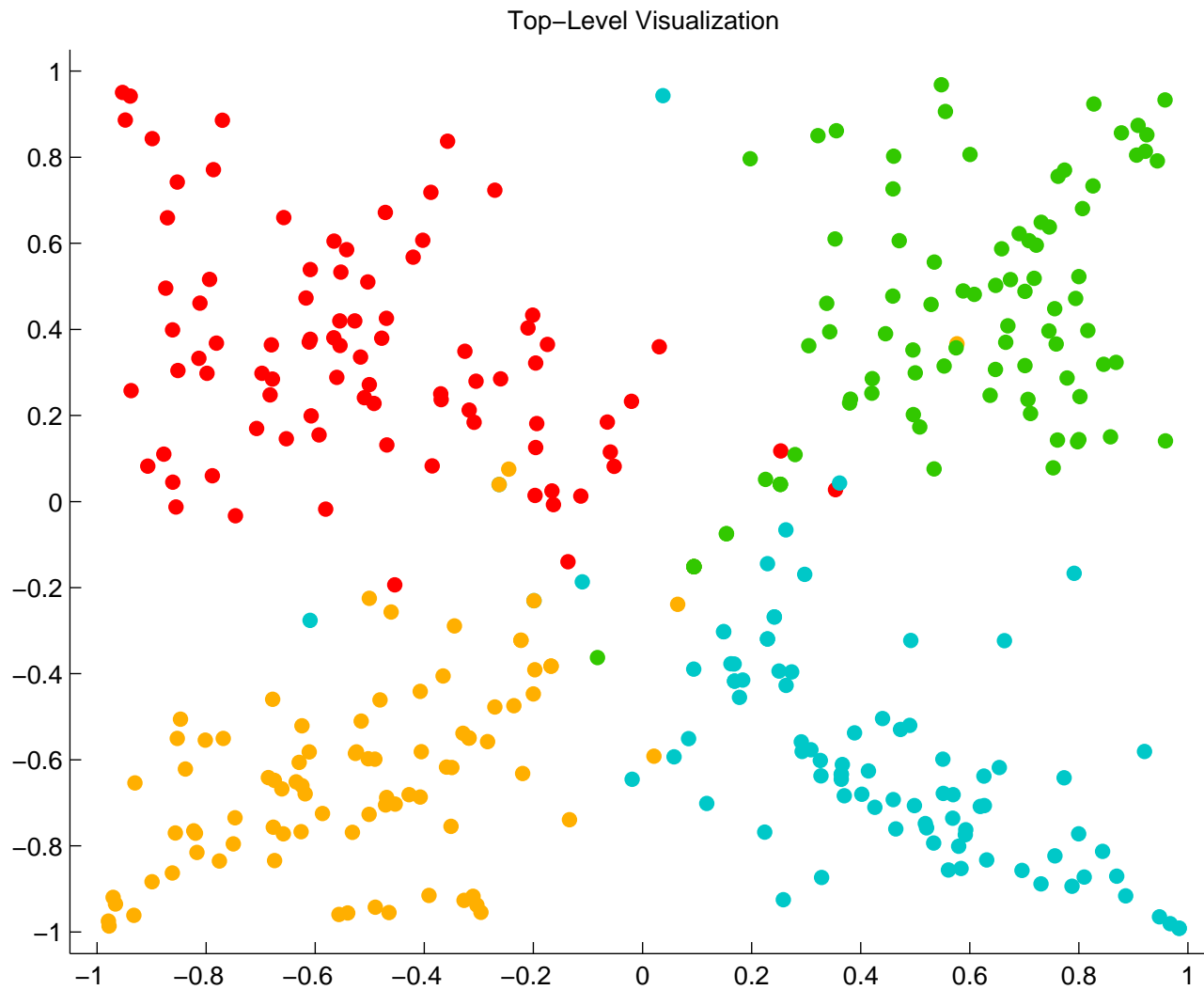


But My Data is Discrete . . .

That's OK; we can do all this for Bernoulli and multinomial noise models (any exponential family distribution).

- Latent Trait Model (LTM) in place of GTM (Kaban and Girolami). EM algorithm still applies.
- Magnification and curvature can still be computed.
- A hierarchy can also be constructed.
- Automated initialisation can still be performed.

Document Dataset



Neuroscale

- Given a dissimilarity matrix d_{ij} , we want to map points y_i to a different space such that their dissimilarities in that space, \tilde{d}_{ij} , are as close as possible to the d_{ij} .
- The **stress** measure is used as objective function

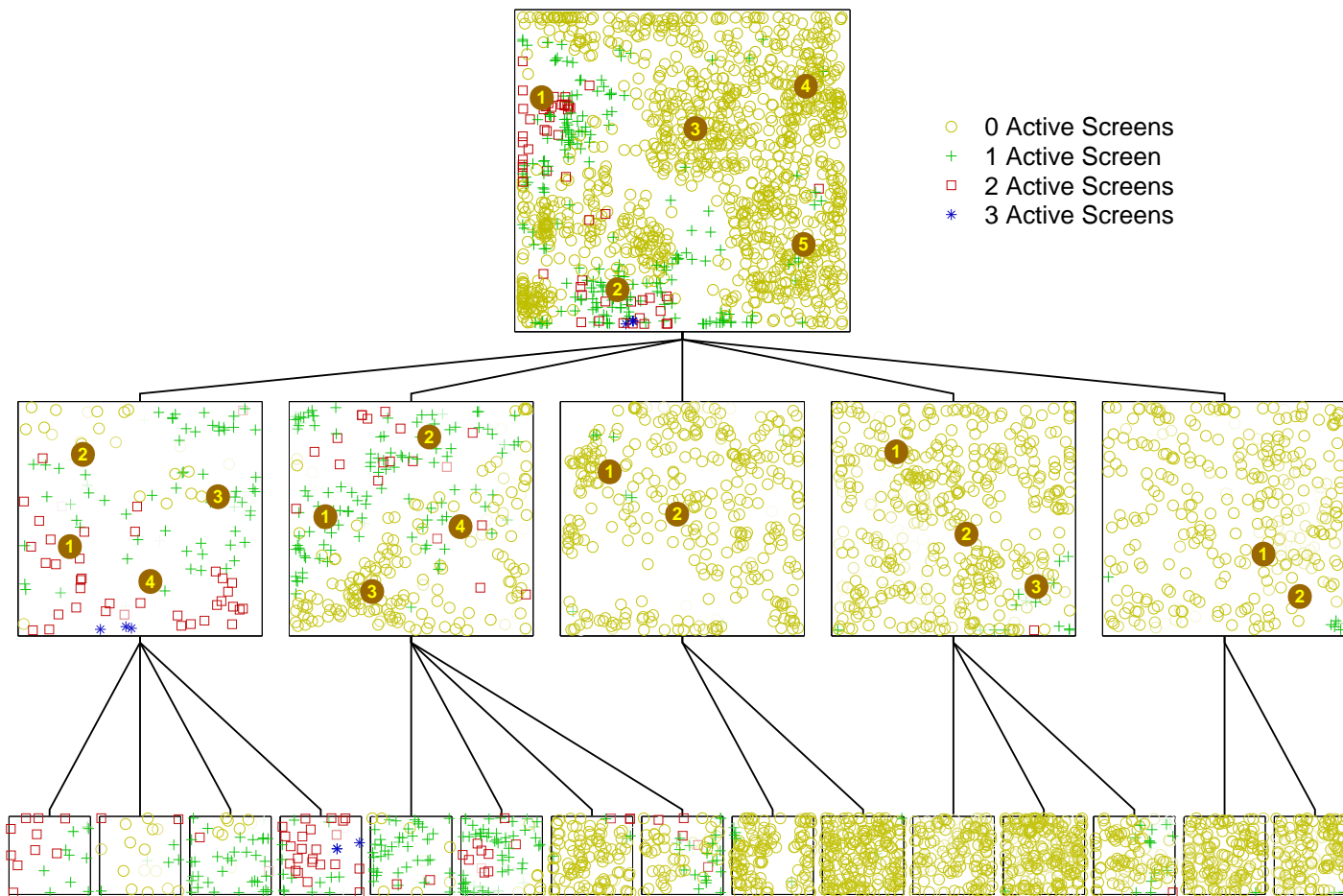
$$E = \frac{1}{\sum_{ij} d_{ij}} \sum_{i < j} \frac{(d_{ij} - \tilde{d}_{ij})^2}{d_{ij}} \quad (1)$$

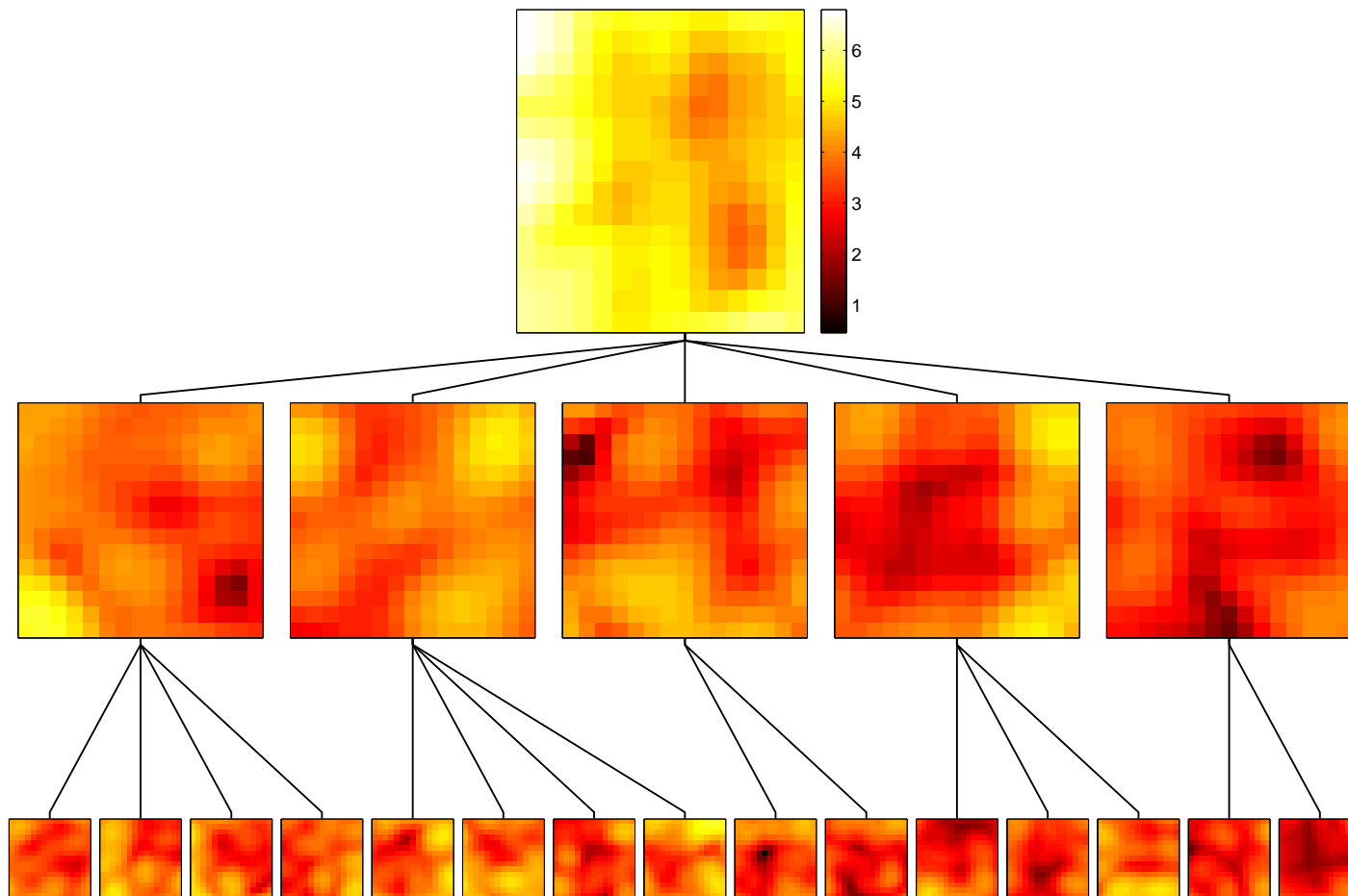
Neuroscale is an RBF network that is trained with the stress error measure.

- The stress can be differentiated and so non-linear optimisation techniques can be used to find y_i .
- A generalisation of the **Sammon mapping** and **multi-dimensional scaling**.

Chemometric Application I: HTS Data Exploration

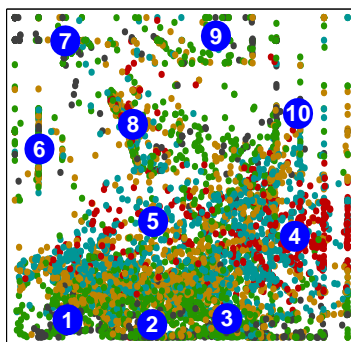
- Chemists at Pfizer searching for active compounds can now screen hundreds of thousands of compounds in a fortnight.
- Gain a better understanding of the results of multiple screens through the use of novel data visualisation and modelling techniques.
- Find **clusters** of similar compounds (measured in terms of biological activity) and using a representative subset to reduce the number of compounds in a screen.



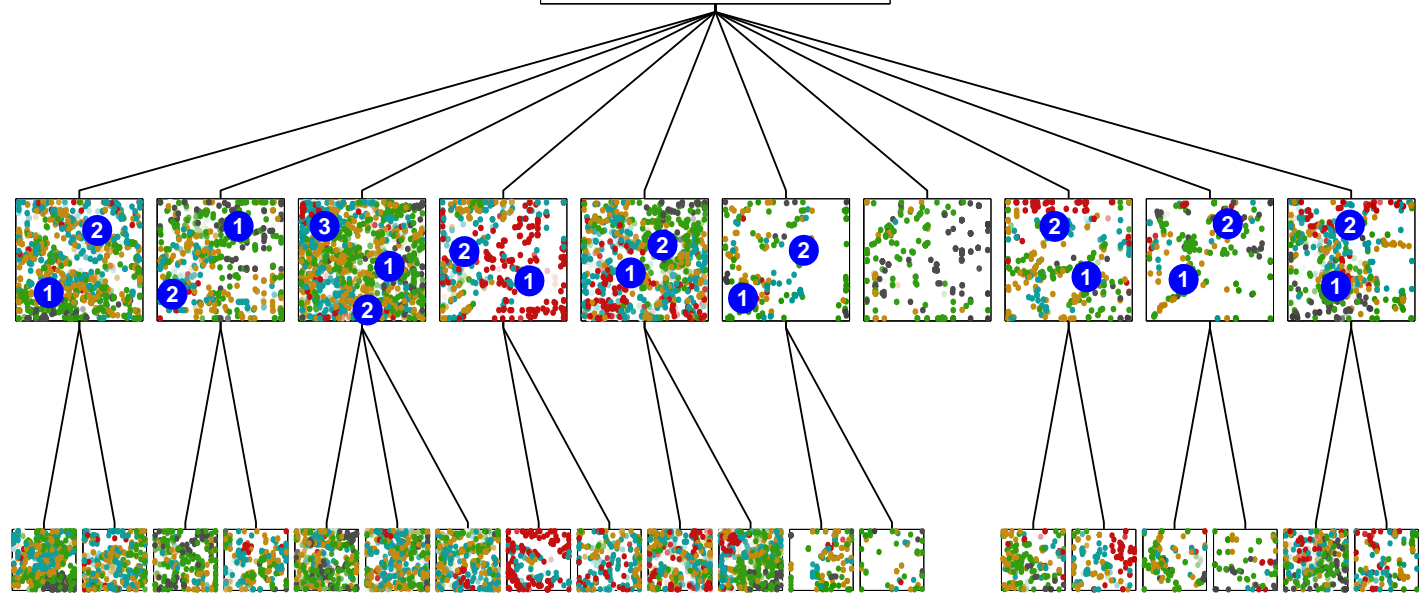


Chemometric Application II: Predicting Chemical Properties from Topological Features

- We have taken data from Jens Lösel (Pfizer) which consists of 6912 14-dimensional vectors representing chemical compounds using topological indexes developed at Pfizer.
- The task is to predict LogP.
- Plots segment the data (by responsibility) which can be used to build **local** predictive models which are often more accurate than **global** models.
- Only 14 inputs, compared with c. 1000 for other methods of predicting logP.
- Results better than other algorithms.

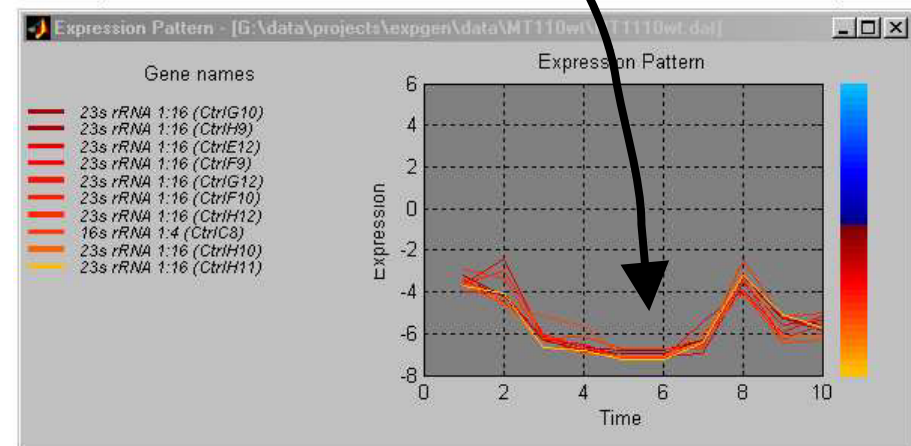
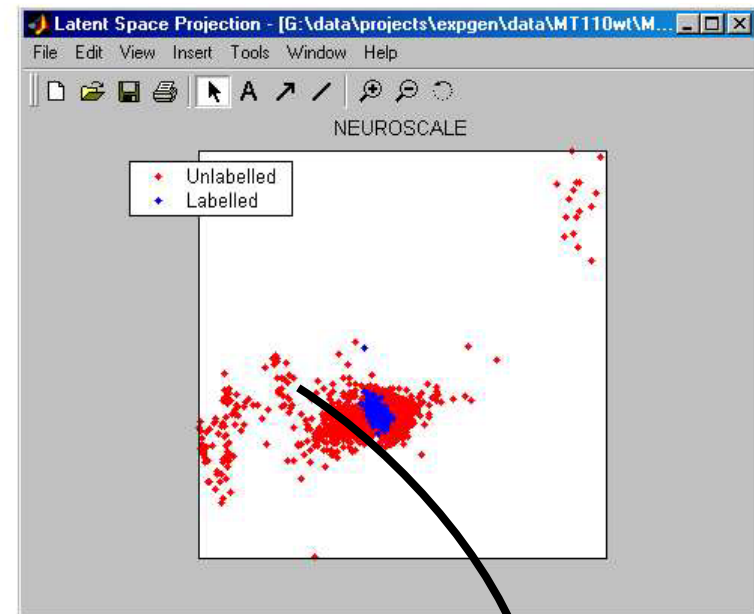


- [-6,0)
- [0,1.5)
- [1.5,2.5)
- [2.5,4.2)
- [4.2,12)

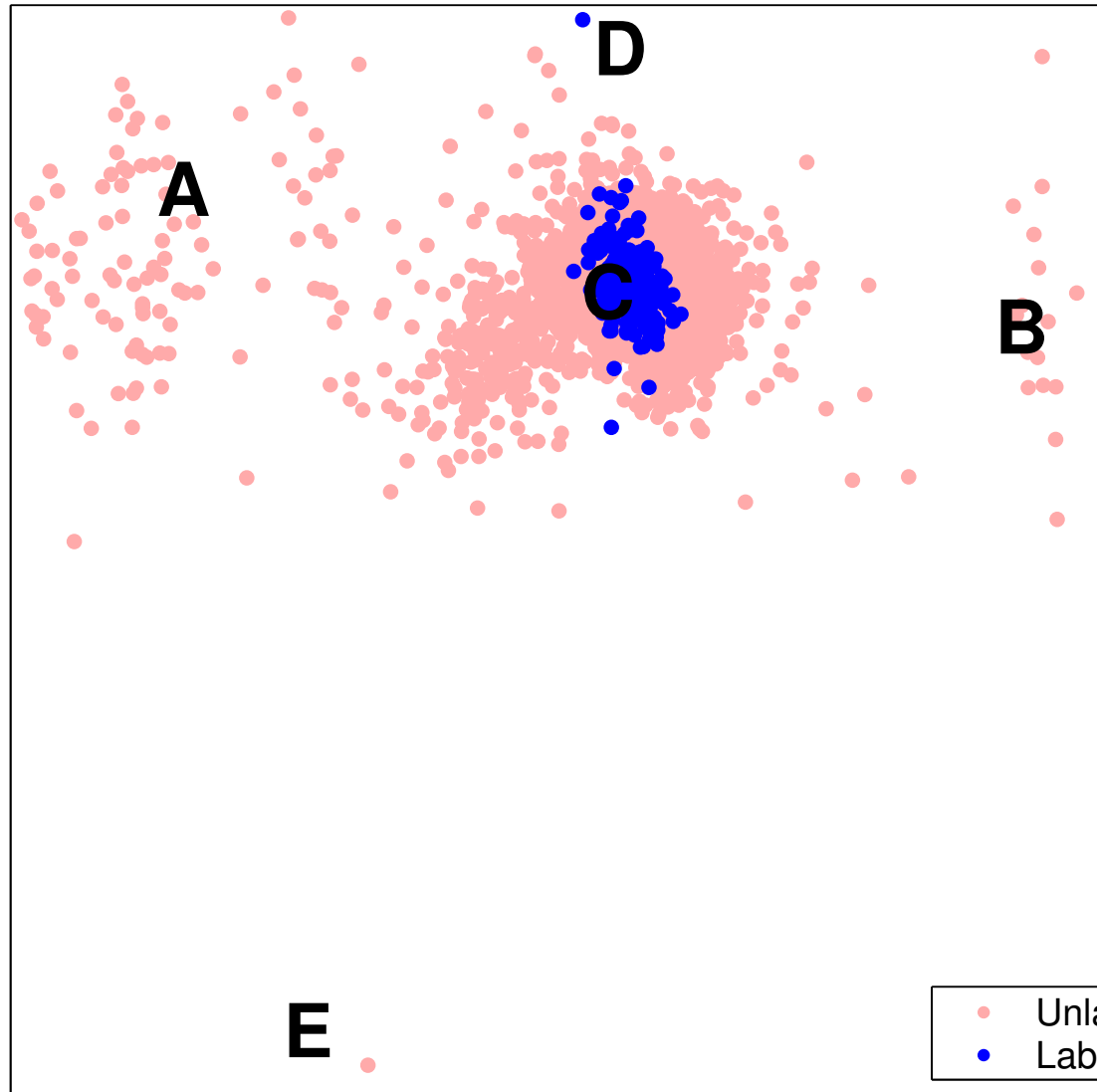


Biological Application: Gene Expression Data

- Exploring cell dynamics through time-course gene expression experiments.
- Use visualisation in a user-friendly tool to explore the data:
 - Find significant clusters.
 - Detect anomalies.
 - Separate significantly expressed genes.



NEUROSCALE



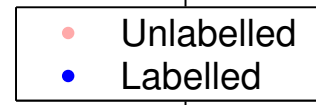
A

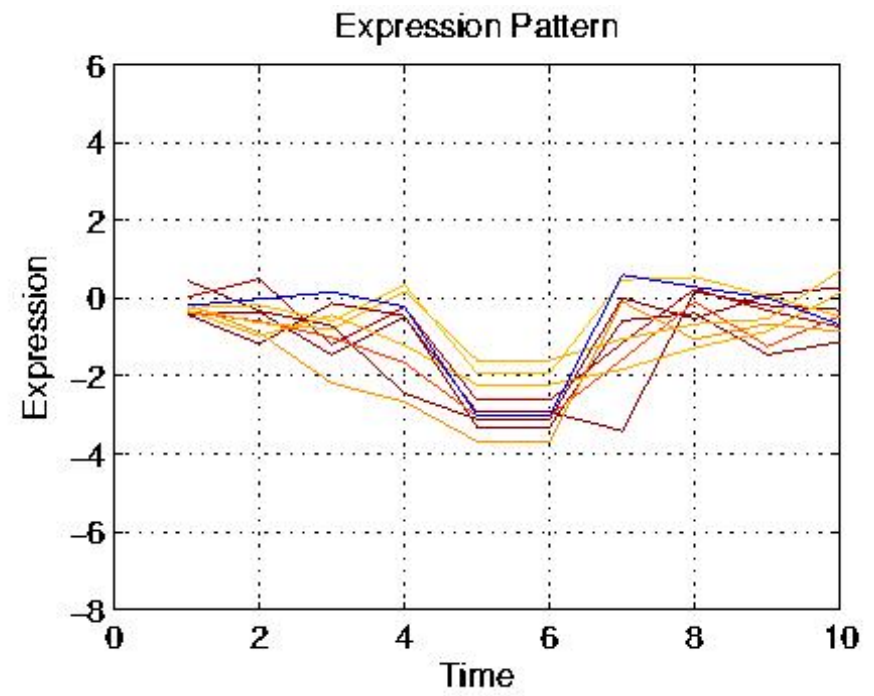
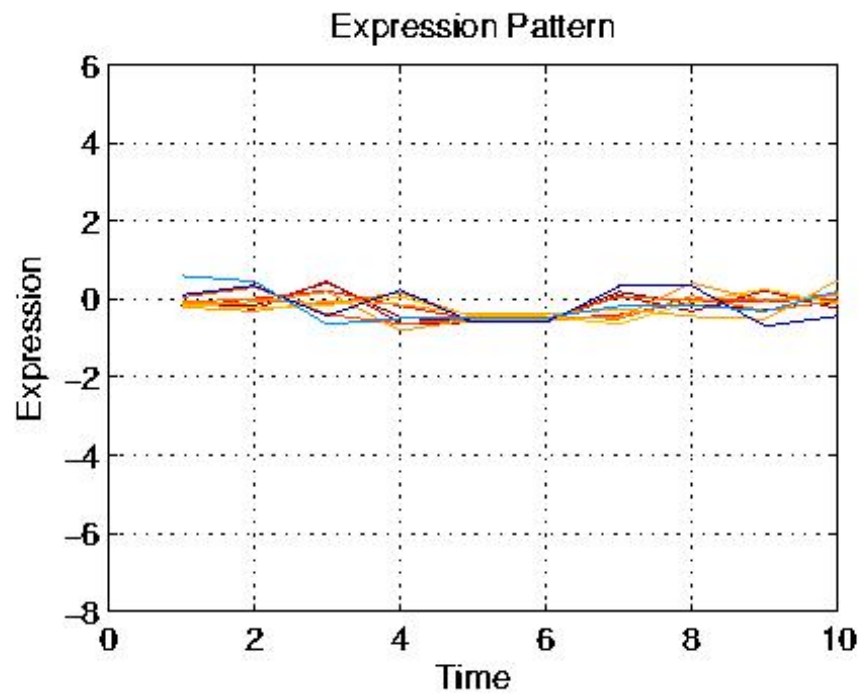
D

C

B

E





Further Applications and Developments

- Can combine data from multiple sources (cf. HTS data): other assays, information from the literature.
- Training and inference algorithms can cope with missing data.
- Temporal dependencies can be incorporated (GTM through Time).

Conclusions

- Visualisation is an important tool for all types of user; the domain expert **must** be involved in the process.
- A single plot is not enough for large, complicated datasets.
- Hierarchies allow the user to drill down into data using either supervised or unsupervised placement of sub-models.
- **Interaction** with the plots allows the user to query the data more effectively.
- All the software is available in Netlab toolbox and its extensions. This contains other goodies, such as neural networks, Gaussian mixture models, and the Neuroscale topographic visualisation technique.

<http://www.ncrg.aston.ac.uk/netlab>