EuroVis 2017
Machine Learning Methods in Visualization for Big Data 2017
12 June 2017, Barcelona, Spain

Machine Learning Methods in Visualisation for Big Data 2017

Daniel Archambault[1] Ian Nabney[2]
Jaakko Peltonen[3]

1 Swansea University

2 Aston University

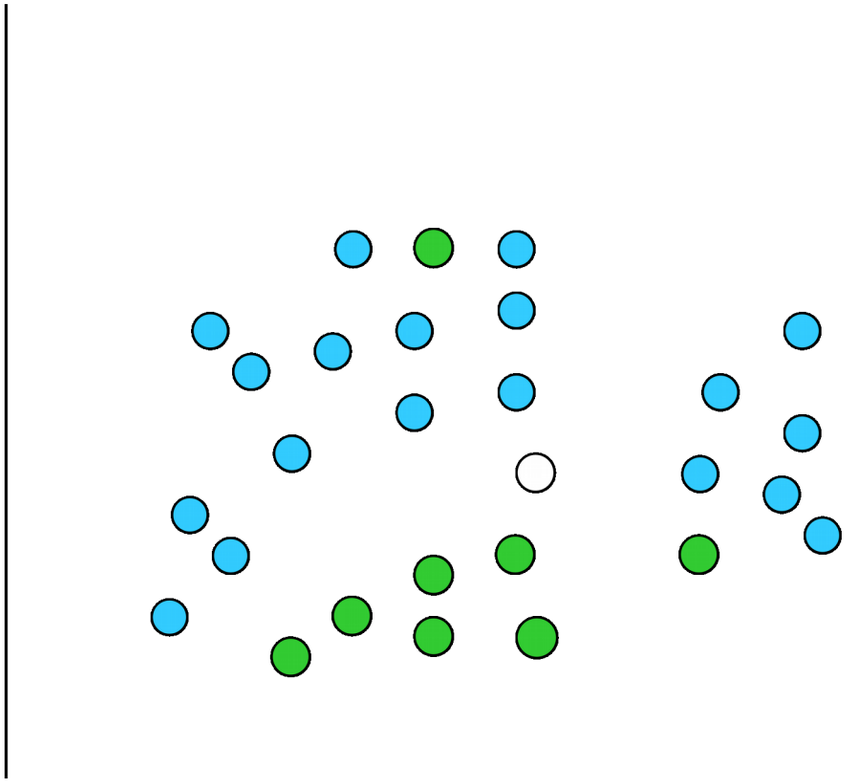
3 University of Tampere, Aalto University

Semi-Supervised Models

- In a *supervised* task we know the outcome for each example (e.g. a class or continuous value) and we try to develop a model that can predict that outcome. Classification or regression
- In an *unsupervised* task we have data, but no variable represents a single outcome for each example and we try to develop a model that looks for groups in the data. Clustering or visualisation
- In some unsupervised tasks we want a target variable to influence the output: *semi-supervised* or *relative supervision*.

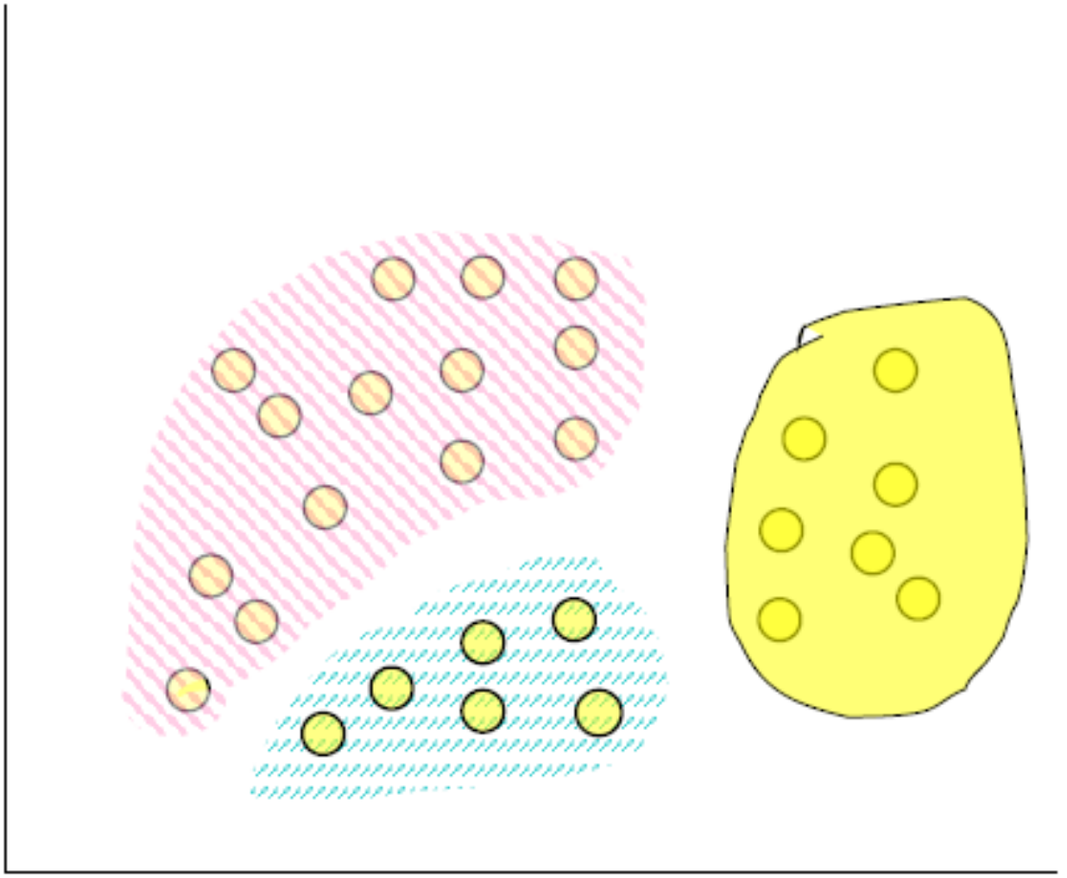
Supervised Task: Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances



Unsupervised Models

- Find natural grouping of instances given unlabelled data



Metric learning

- Many statistical methods rely on distances as much or more than they do on feature values:
  - nearest neighbor regression/classification uses distances to find the nearest neighbors
  - many clustering approaches such as k-means use distances as part of the algorithm to optimize the clustering
  - in information retrieval, “best” results are often the ones most similar to the query according to some distance
- Dimensionality reduction methods such as multidimensional scaling, Sammon mapping, Self-organizing maps, Stochastic Neigbor Embedding, Neighbor Retrieval Visualizer, and others are distance-based
- In many cases distances from a new distance function can be just plugged in to dimensionality reduction methods. (In some cases more is needed.)

Topographic Mappings

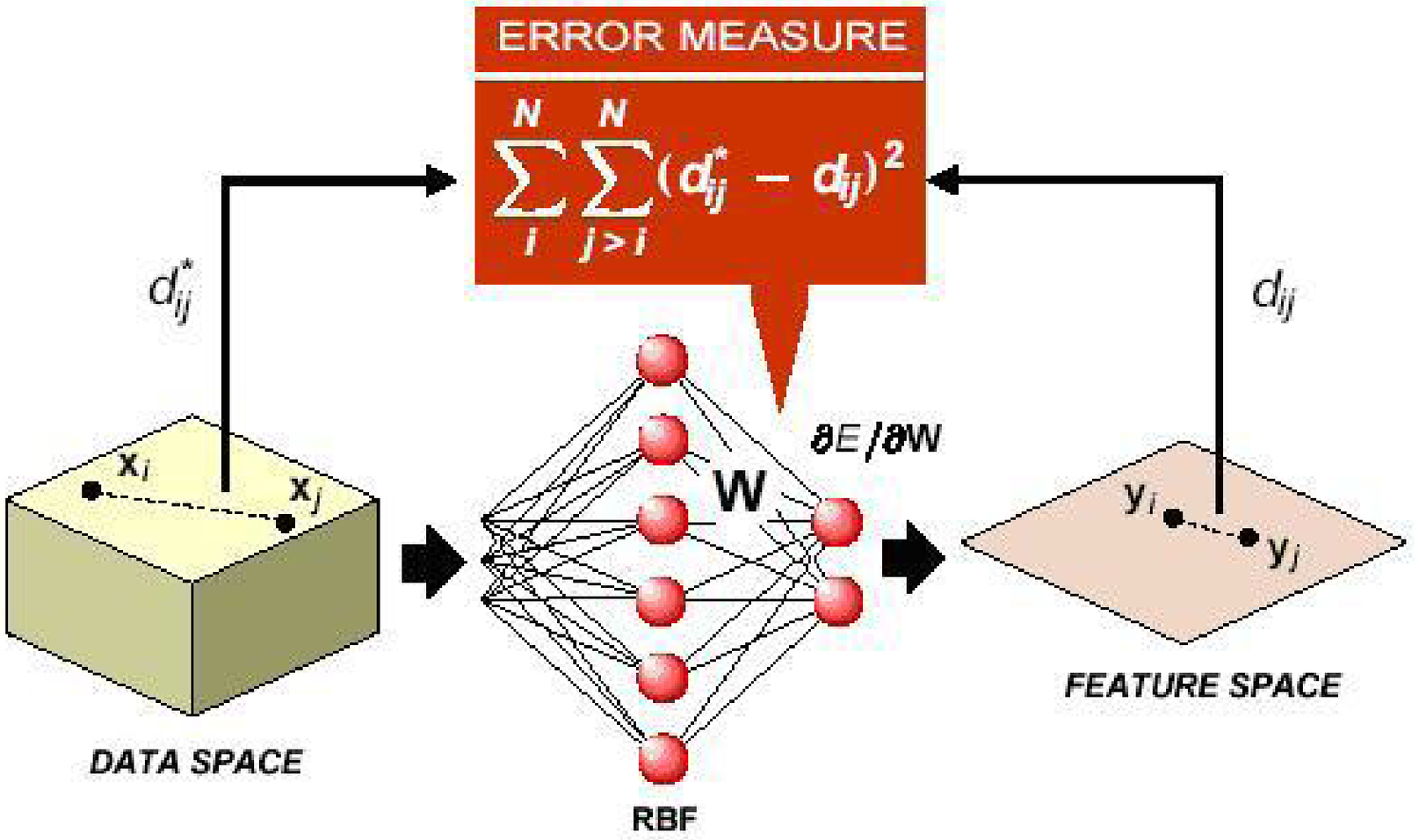
- Basic aim is that distances in the visualisation space are as close a possible to those in original data space.
- Given a dissimilarity matrix $d_{ij}|$ we want to map data points x_i to points y _i in a feature space such that their dissimilarities $\tilde{d}_{ij}$ are as close as possible to $d_{ij}|$
- The map is said to preserve similarities. The stress measure is used as objective function.

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

Multi-dimensional Scaling

- Given distances or dissimilarities between every pair of observations try to preserve these as far as possible in lower-dimensional space.
- In classical scaling, the distance between the objects is assumed to be Euclidean. A linear projection then corresponds to PCA.
- The Sammon mapping is a non-linear multidimensional scaling technique more general (and more widely used) than classical scaling.
- Neuroscale is a neural network based scaling technique that has the advantage of actually giving a map that generalises!

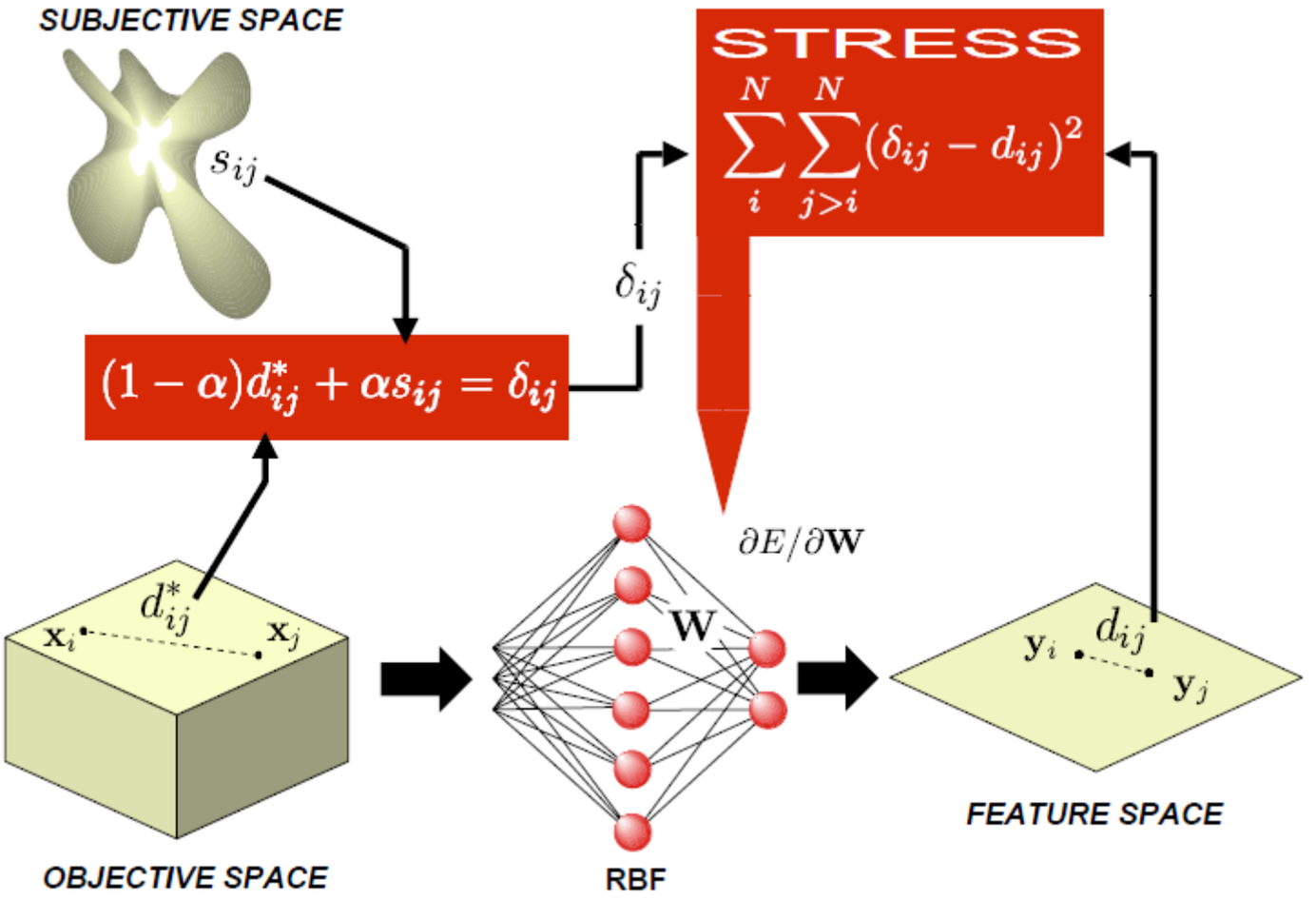
Neuroscale


ERROR MEASURE
$\sum_{i}^{N}\sum_{j>i}^{N}(d_{ij}^{*} - d_{ij})^2$
d_{ij}^{*}
d_{ij}
x_i
x_j
$\partial E / \partial W$
W
y_i
y_j
DATA SPACE
RBF
FEATURE SPACE

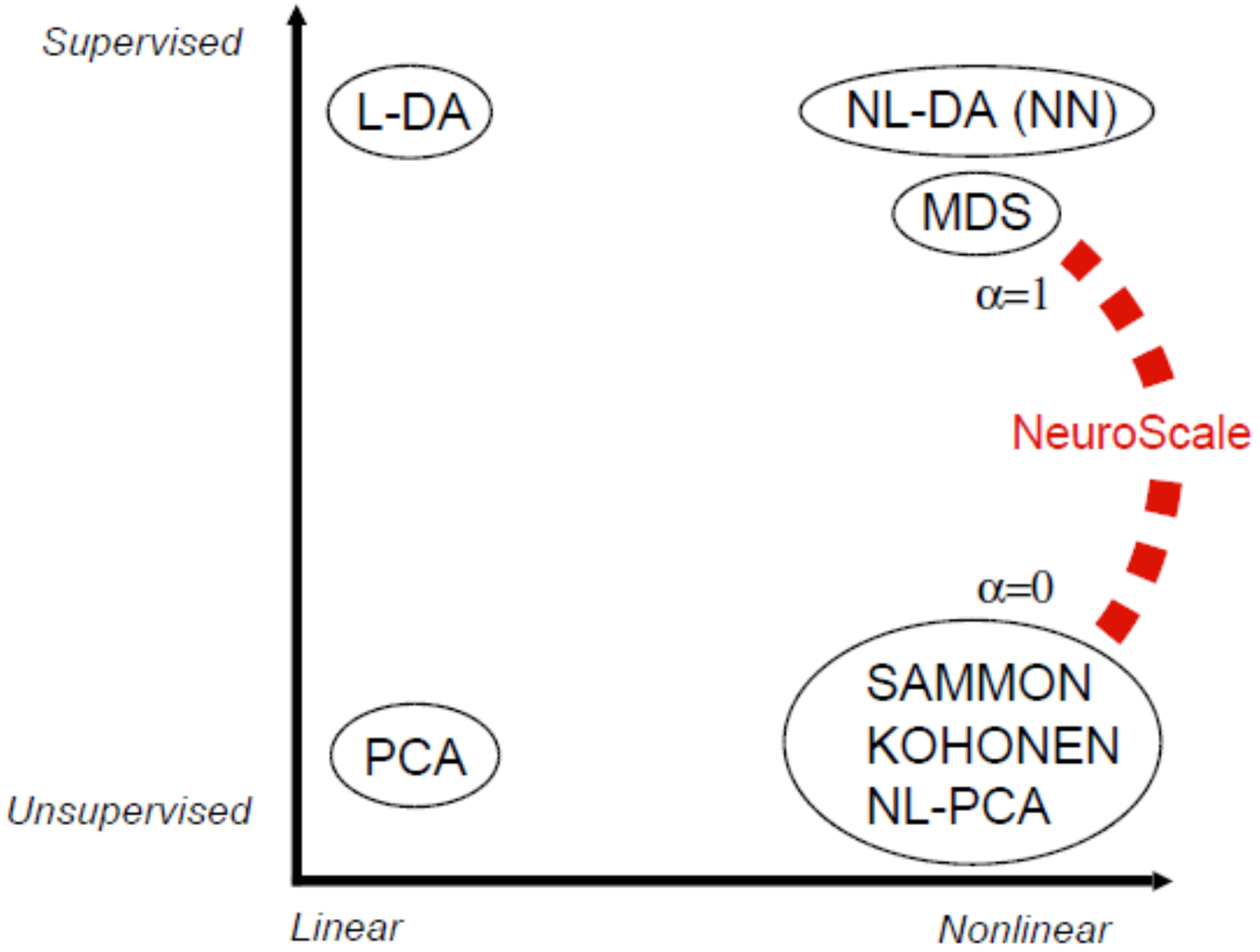
Subjective metrics

- Modify the stress measure:
- .
- $$E' = \sum_{i}^{N} \sum_{j}^{N} (\delta_{ij} - \| \mathbf{y}_i - \mathbf{y}_j \|)^2$$
-
- Inter-point distances for pairs of points in different classes are modified by the addition of some constant term k, such that their separation should be exaggerated in the resultant map.
-
- $$\delta_{ij} = \begin{cases} d_{ij}^* & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are in the same class,} \\ d_{ij}^* + k & \text{otherwise.} \end{cases}$$
-
- Other formulations are possible – can use a dissimilarity matrix for classes or distance for an auxiliary continuous variable. The relative weight of objective and subjective elements can be controlled by a parameter.

Neuroscale Operation

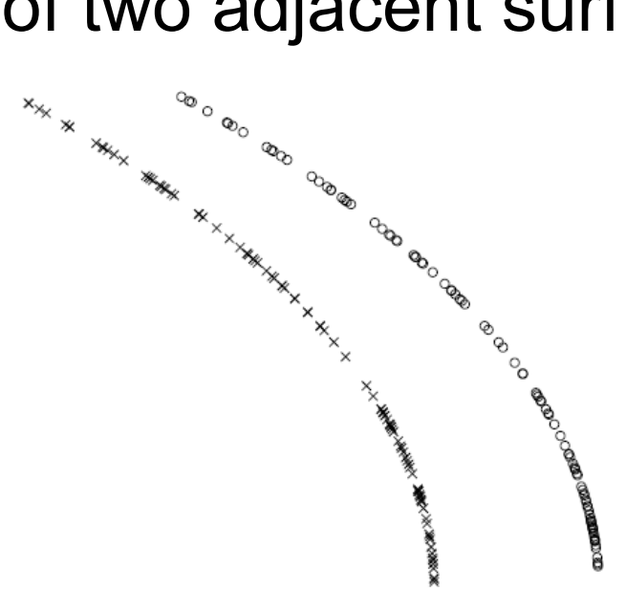


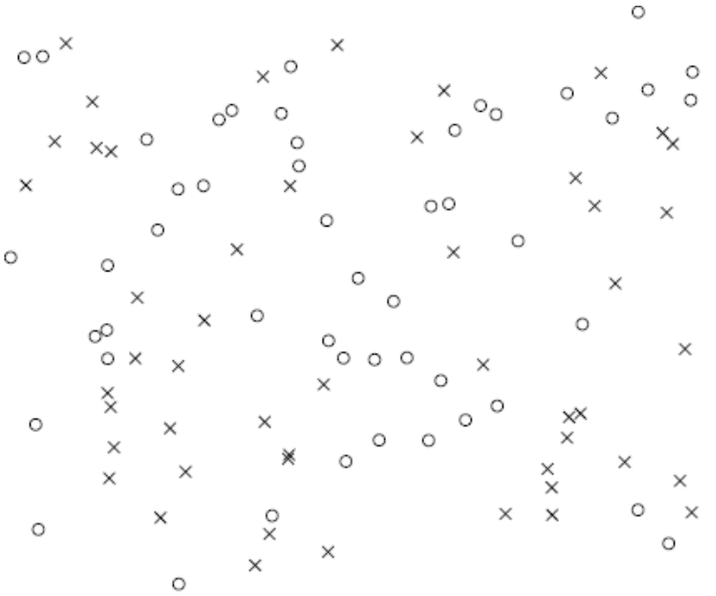
Model Comparison



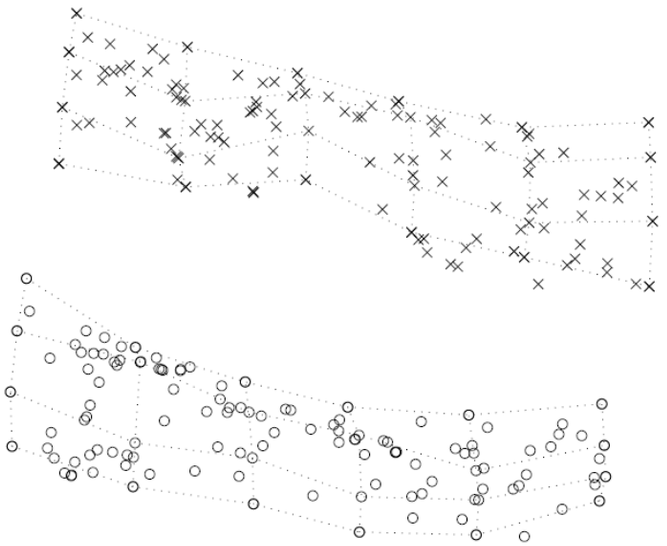
Synthetic Example

- 50 data points were distributed uniformly at random over each of two adjacent surfaces.





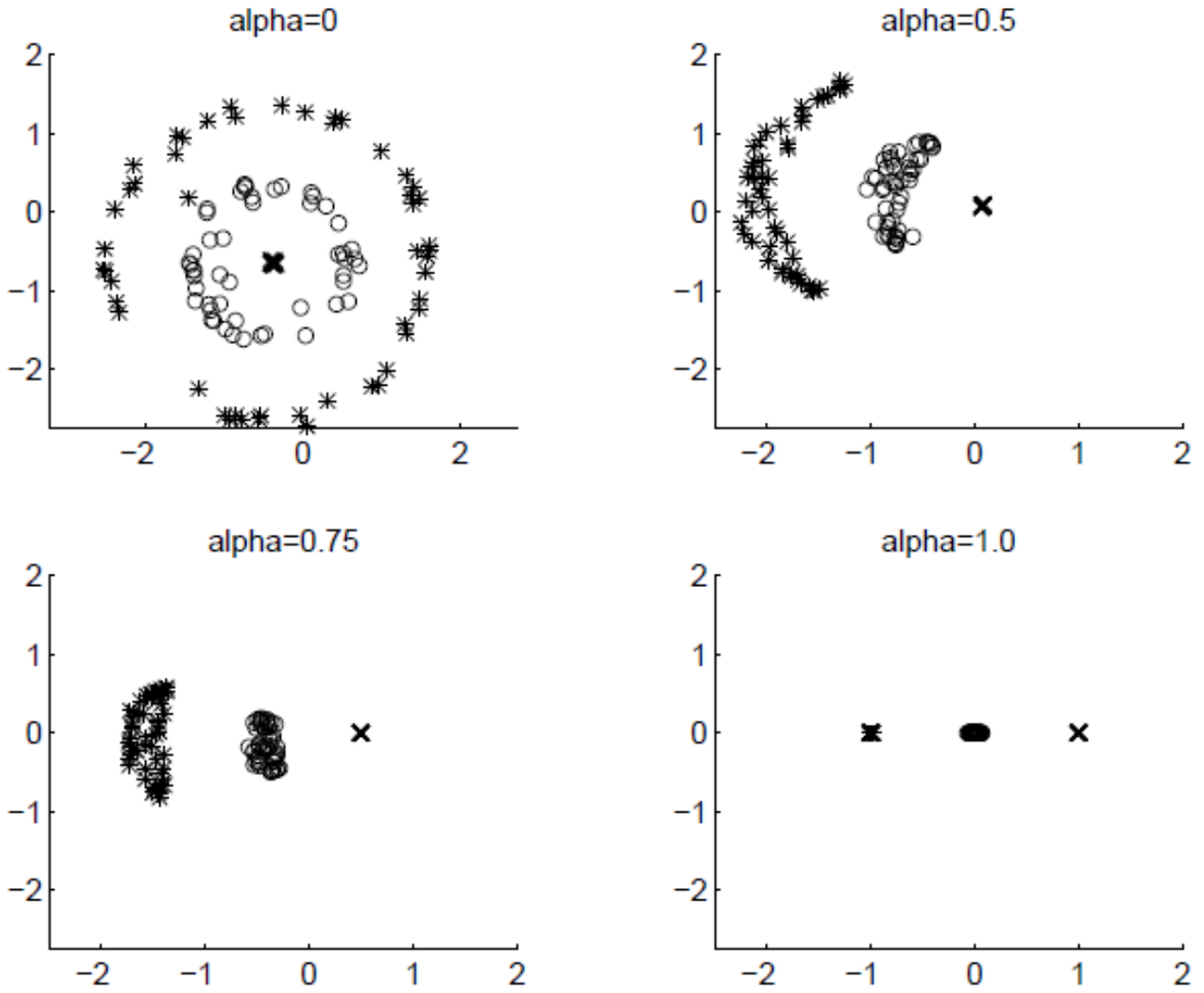
Neuroscale



Neuroscale:
alpha = 0.5

Synthetic Example 2

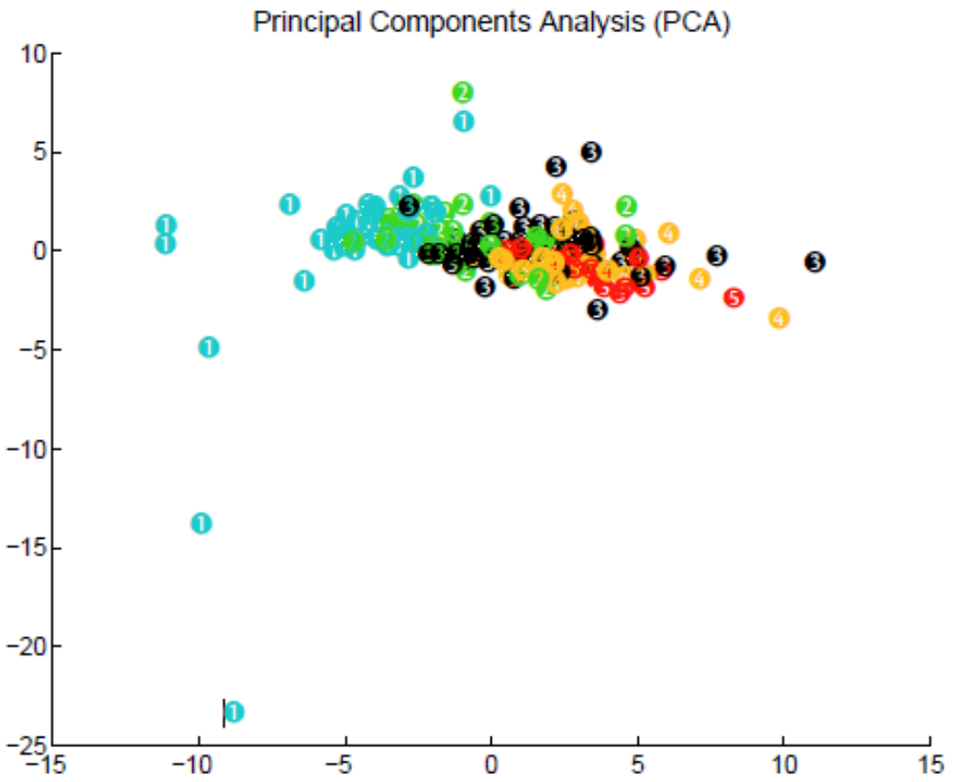
- Three concentric spheres.
-

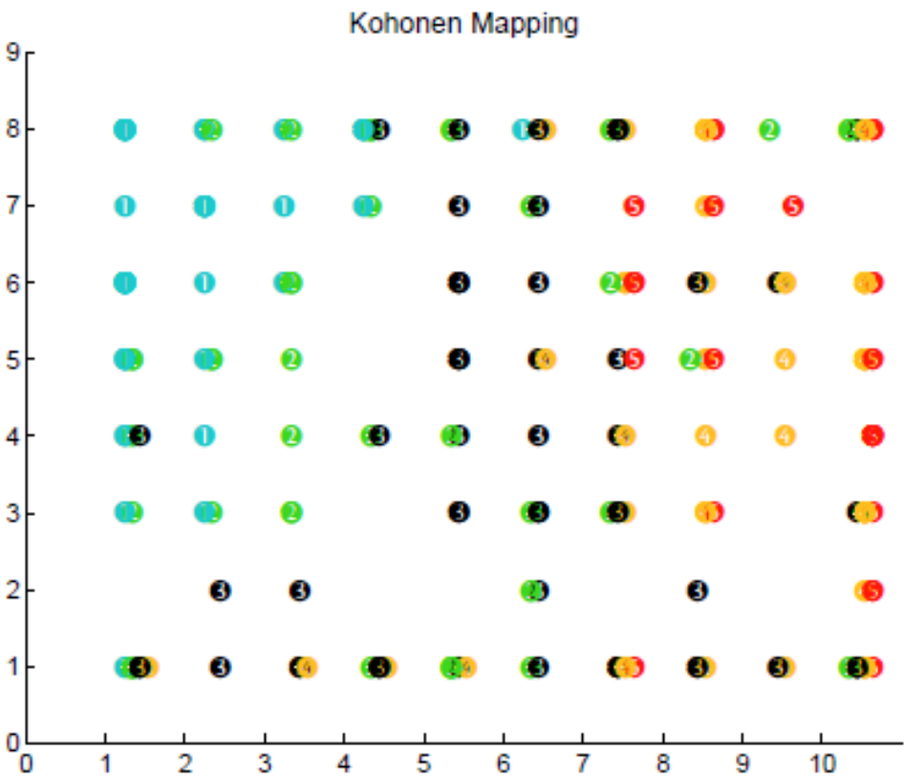


RAE 1992 Dataset

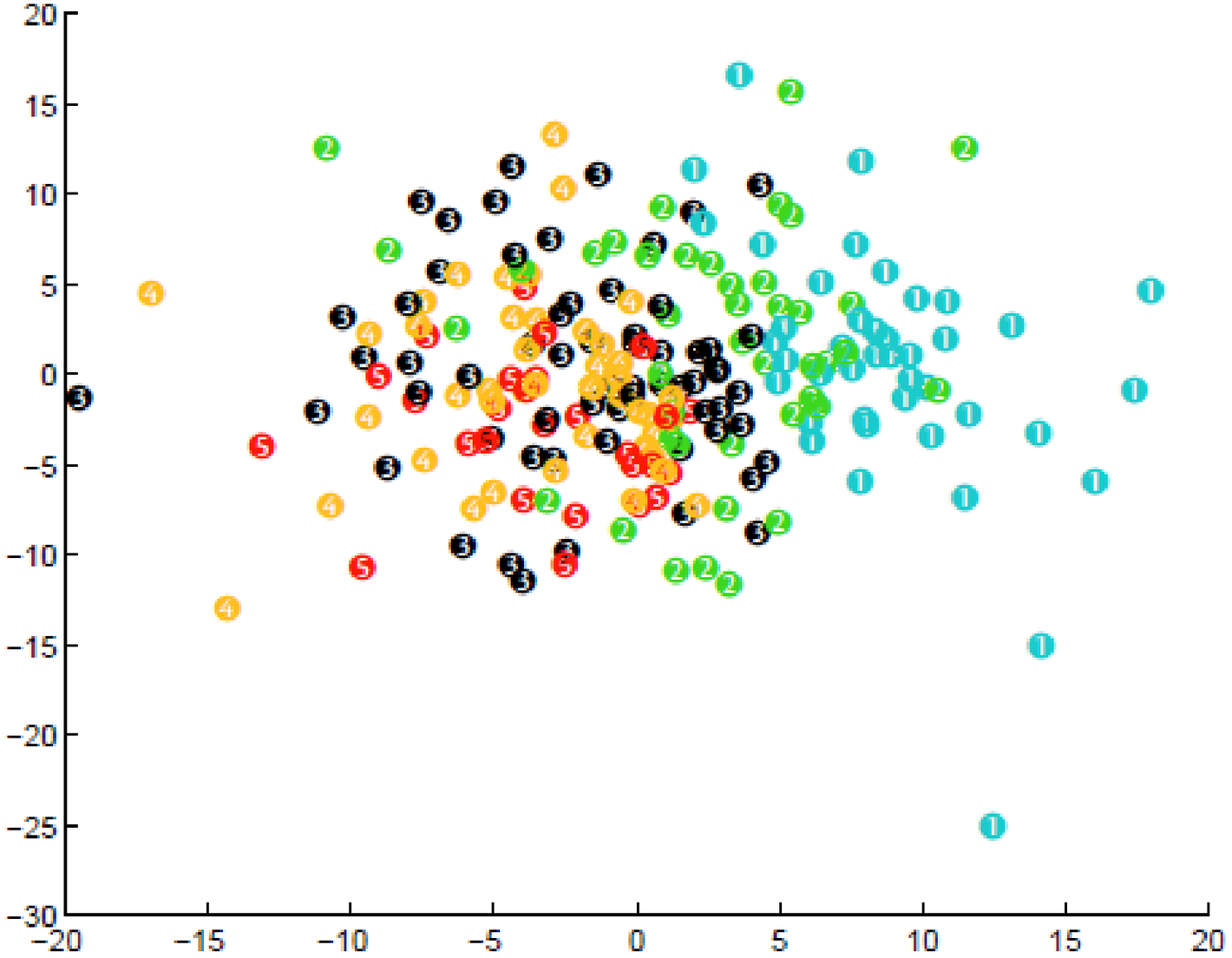
- Departments in UK universities rated by an assessment panel on research every 6 years. 150 variables and a ranking (1-5): number and value of grants, number of publications, …
- Accumulated variables measured over multiple years.
- Interest in predicting outcome from numeric variables only.
- Combined data from chemistry, physics, biology panels (217 instances). Applied Maths used as an independent test set (67 examples).
- Aim is to extract features to improve accuracy of ranking prediction. Used neural network classification models.

Visualisations

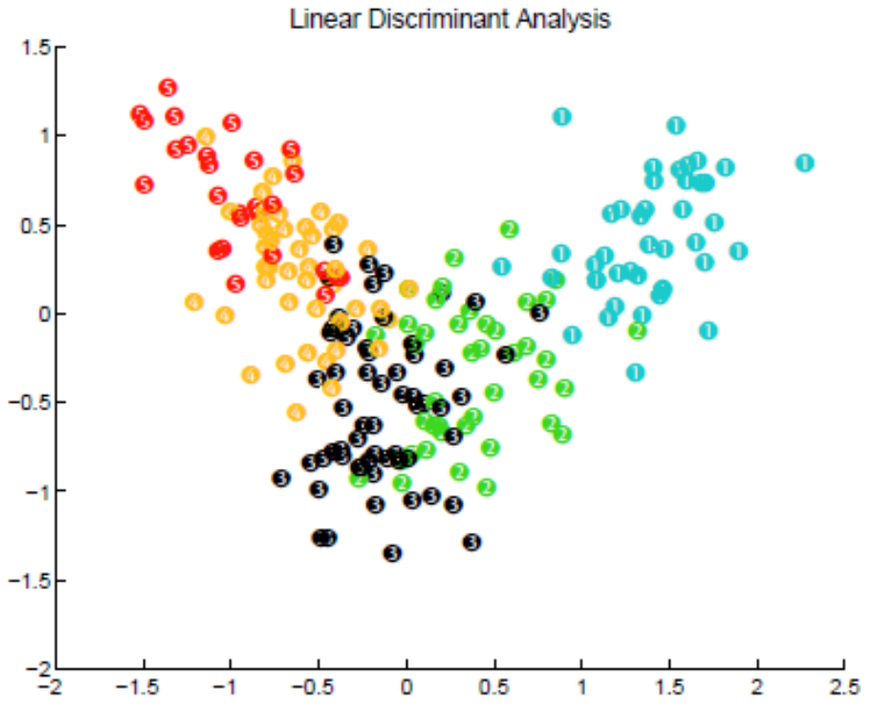


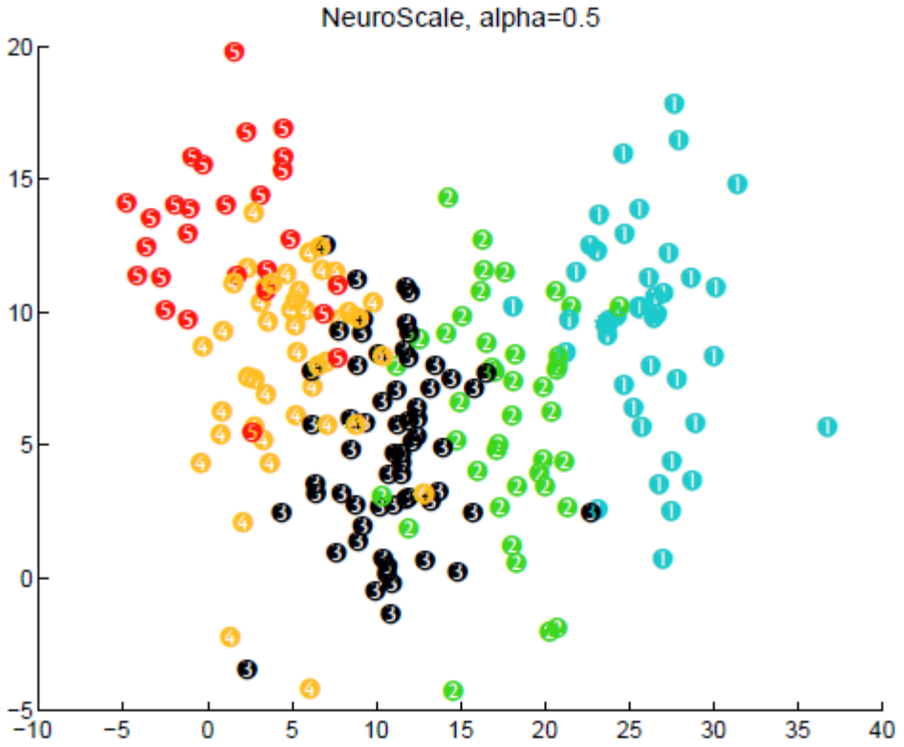


Visualisations: Neuroscale



Subjective Visualisations

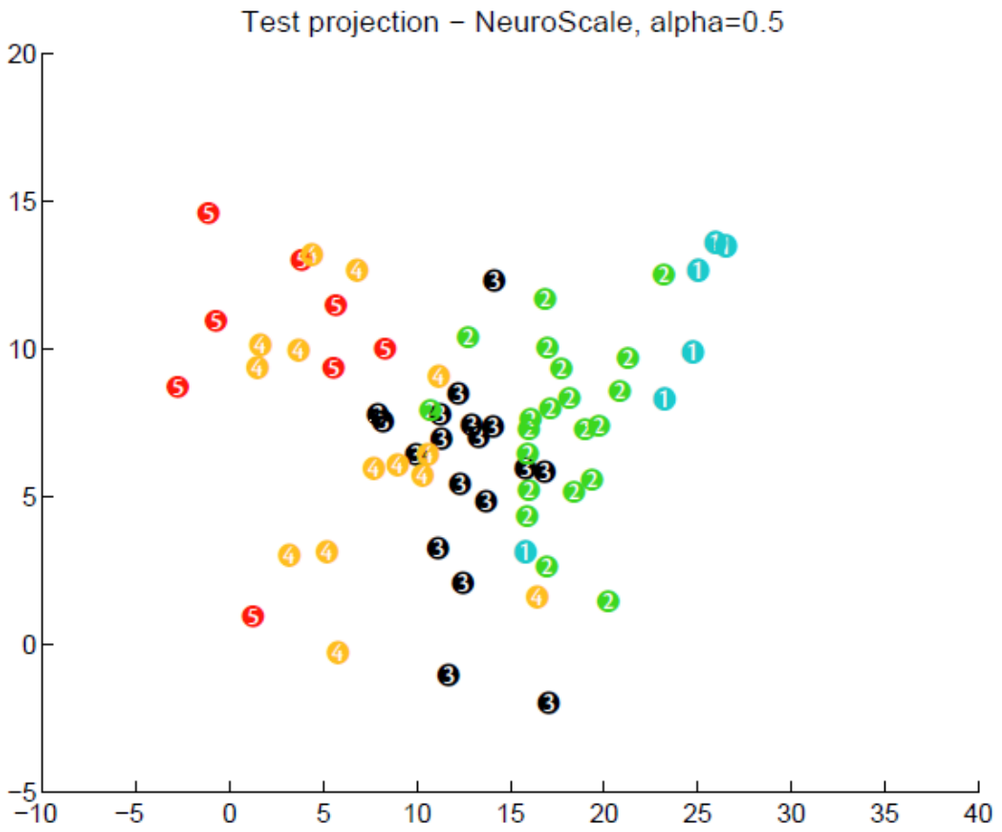


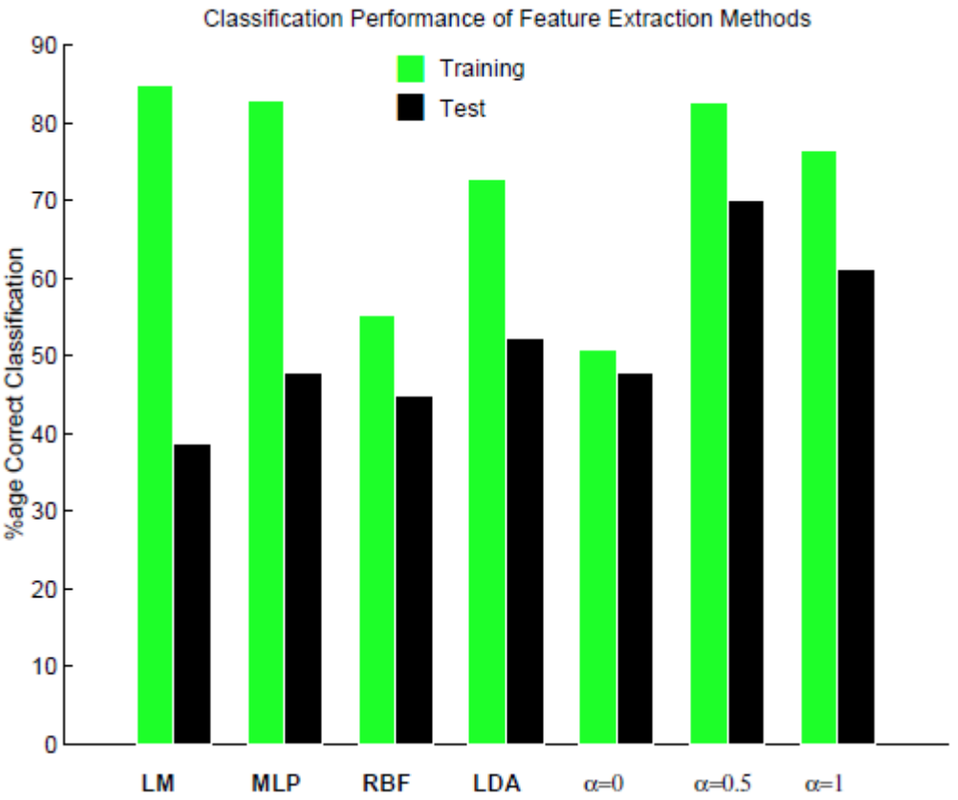


	Actual Rating	Linear	MLP	RBF
Physics, Heriot-Watt	*5*	5	4	3
Physics, Queen's	*5*	4	4	3
Physics, Stirling	*4*	4	4	3
Physics, Westminster	*3*	1	2	1

$$C = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{bmatrix}$$

Feature Extraction





Visualisation and Uncertainty

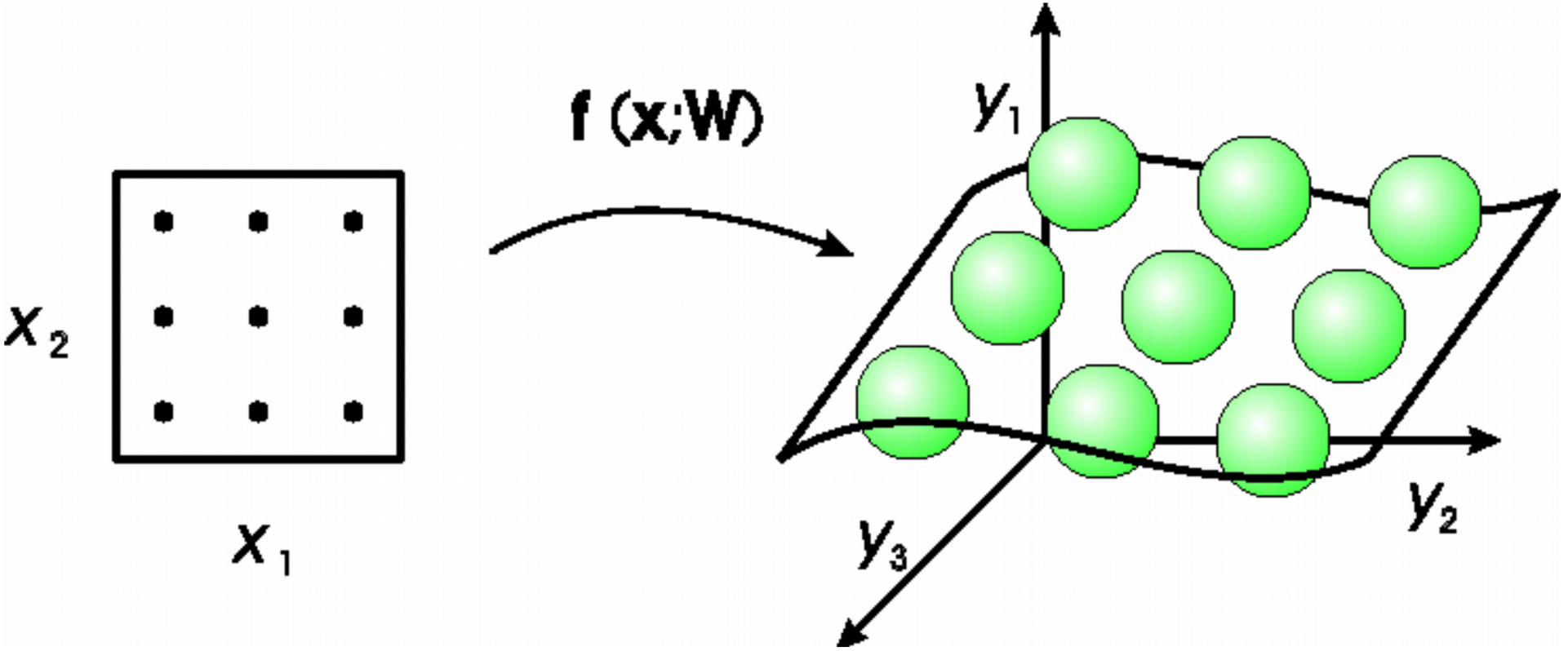
- Real data is noisy.
- We are forced to deal with uncertainty, yet we need to be quantitative.
- The optimal formalism for inference in the presence of uncertainty is probability theory.
- We assume the presence of an underlying regularity to make predictions.
- Bayesian inference allows us to reason probabilistically about the model as well as the data.

Doubt is not a pleasant condition, but certainty is absurd.

Voltaire

Generative Topographic Mapping

- Mapping from latent space to data space
- A thick rubber sheet studded with tennis balls. GTM defines p(y|x;W); use Bayes' theorem to compute p(x|y*;W) for a given point y* in data space.



Enhancements to GTM

- Curvatures and magnification factors give more information about shape of manifold.
- Hierarchy allows the user to drill down into data; either user-defined or automated (MML) selection of sub-model positions.
- Temporal dependencies in data handled by GTM through Time.
- Discrete data handled by Latent Trait Model (LTM): all the other goodies work for it as well.
- Can cope with missing data in training and visualisation.
- MML methods for feature selection.
- Structured covariance.
- Uncertainty measures.

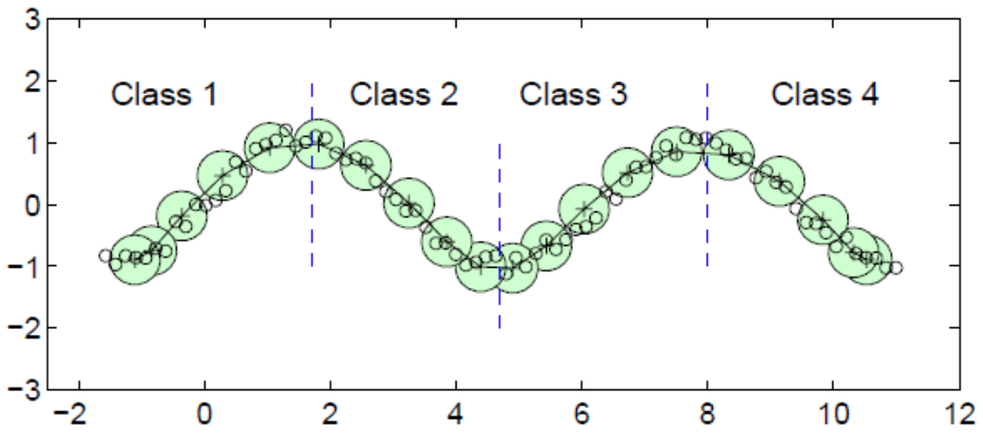
Incomplete Data

- The training algorithm for the Generative Topographic Mapping (GTM) can be modied to use class information to improve results on incomplete data.
- The approach is based on an EM method which estimates the parameters of the mixture components and missing values at the same time.
- Furthermore, if we know the class membership of each pattern, we can improve the generic algorithm by eliminating multi-modalities in the posterior distribution over the latent space centres.
- The algorithm can help to construct informative visualisation plots, even when many of the training points are incomplete.

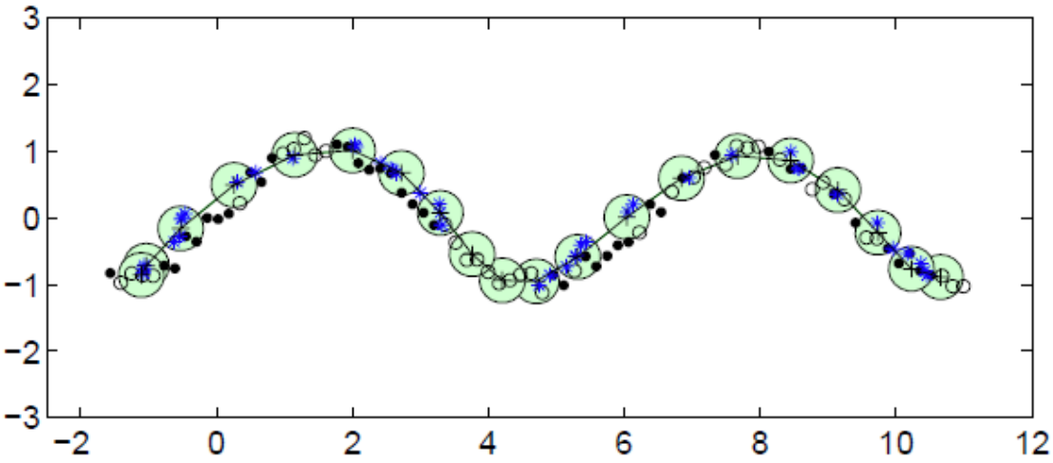
Algorithm

- EM algorithm involves expectation (E-step): extend this to missing values as well as missing kernel.
- $$\hat{\mathbf{t}}_{nj}^{m} = \mathrm{E}(\mathbf{t}_n^m | z_{nj} = 1, \mathbf{t}_n^o, \theta_j) = (\mathbf{y}_j^m)^{old} + \Sigma_j^{mo} \Sigma_j^{oo^{-1}} (\mathbf{t}_n^o - (\mathbf{y}_j^o)^{old})$$
- GTM model uses spherical covariance, hence this inference is quite uninformative
- $$\hat{\mathbf{t}}_{nj}^{m} = (\mathbf{y}_j^m)^{old}$$
- Class membership (if available) can provide more accurate inference.
- $$r_{njc} = P(\mathbf{x}_j | \mathbf{t}_n^o, c_n) = \frac{P(\mathbf{x}_j, c_n | \mathbf{t}_n^o)}{\sum_{k=1}^{K} P(\mathbf{x}_k, c_n | \mathbf{t}_n^o)}$$

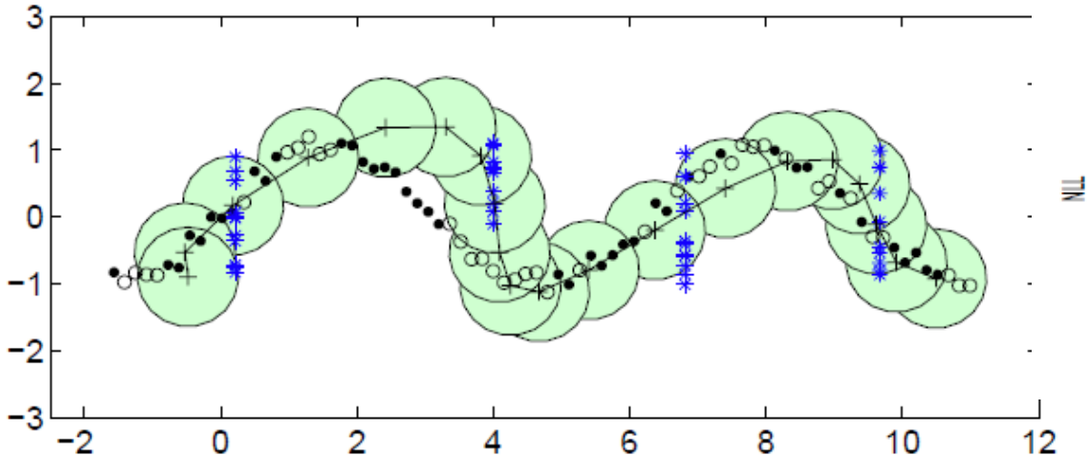
Synthetic Data Results



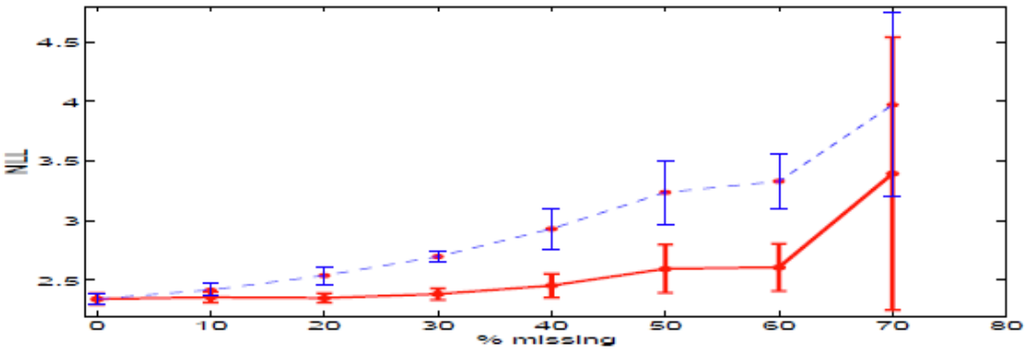
GTM with complete data



GTM-EM with missing data



GTM with missing data and MI

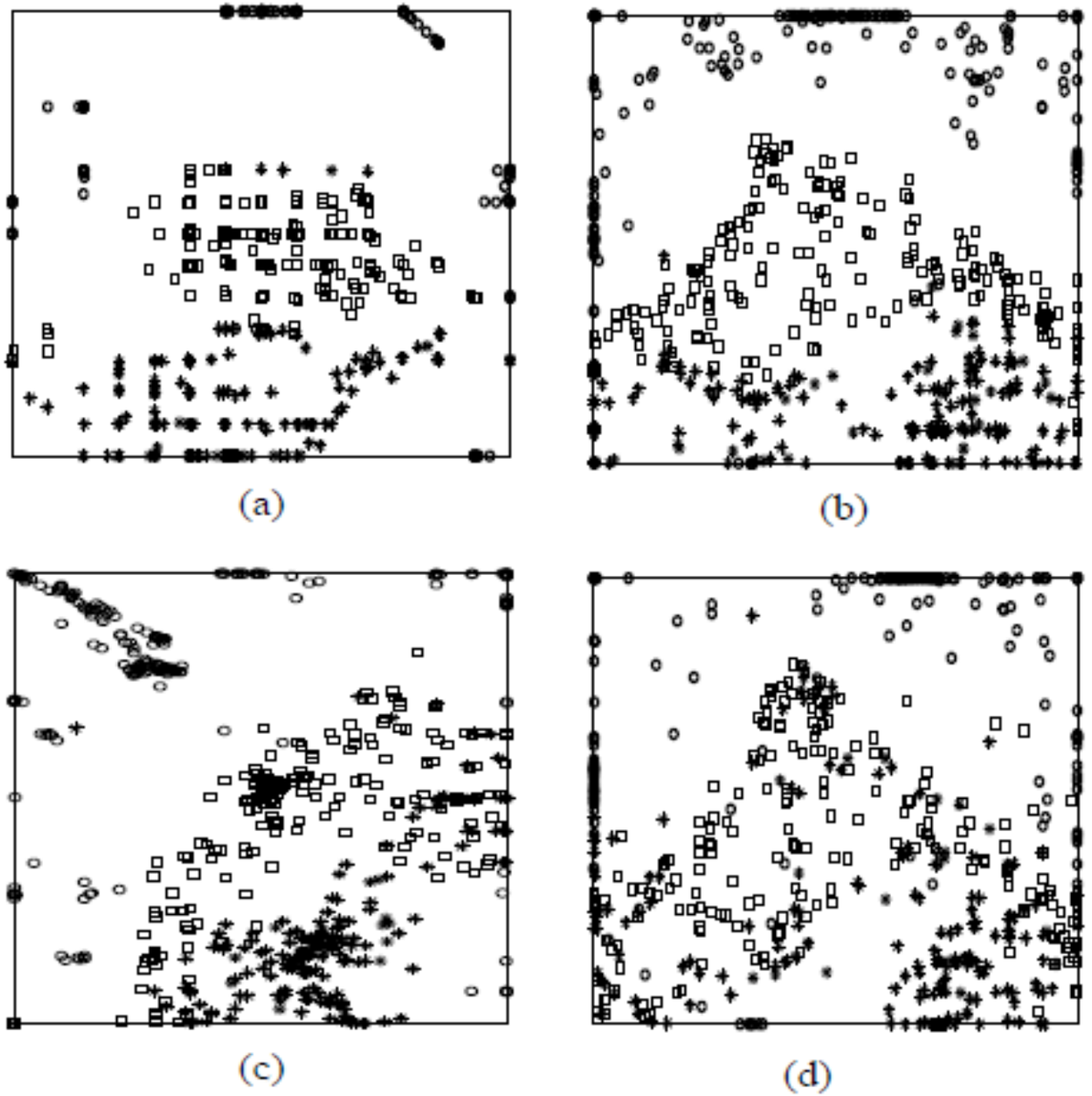


Accuracy as function of missing percentage

Oilflow Dataset

- 12 measured variables and 3 states: homogeneous, annular and laminar.
- In the training set, 50% of the data points in each class are incomplete, with between 6 and 9 values removed.
- Plot (b) shows better separation of classes and matches better to the result obtained from the complete data set (plot (a)).
- After using class-conditional MI, some strongly overlapped clusters appear in plot (c) since the same means are substituted for missing values of the same class.
- As for plot (d), which was obtained just by the generic algorithm, the homogeneous and annular classes are not separated well as we did not use the class-conditional prior knowledge in the training process.

Visualisations: GTM



(a) Complete data
(b) Missing data: class-conditional EM
(c) Missing data: EM conditional MI
(d) Missing data: generic EM

Homogeneous, annular and laminar are represented by square, star and circle signs