Advanced Data Analytics
Week 3: GTM

Ian T. Nabney

University of BRISTOL

Overview

- Understand a probabilistic version of PCA and continuous latent variable models
- Able to use mixture of probabilistic PCA
- Understand how the GTM works in theory and practice
- Aware of some extensions of GTM

Based on Section 12.2 of Bishop and chapter 7 of the Netlab book. A reminder that the PDF can be downloaded from `https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/`

University of BRISTOL

Probabilistic PCA

- The examples of latent variable models we have considered so far used discrete latent variables. In this lecture we will look at continuous latent variables.
- We start with probabilistic PCA.
- Probabilistic PCA represents a constrained Gaussian distribution in which the number of free parameters can be restricted while still allowing the model to capture the dominant correlations in a data set.
- We will then look at the Generative Topographic Mapping (GTM) which incorporates a non-linear mapping into a latent variable model.

Benefits of probabilistic PCA

- We can derive an EM algorithm for PCA that is computationally efficient in situations where only a few leading eigenvectors are required and that avoids having to evaluate the data covariance matrix as an intermediate step.
- The combination of a probabilistic model and EM allows us to deal with missing values in the data set.
- Mixtures of probabilistic PCA models can be formulated in a principled way and trained using the EM algorithm.
- Probabilistic PCA forms the basis for a Bayesian treatment of PCA in which the dimensionality of the principal subspace can be found automatically from the data.
- The existence of a likelihood function allows direct comparison with other probabilistic density models. By contrast, conventional PCA will assign a low reconstruction cost to data points that are close to the principal subspace even if they lie arbitrarily far from the training data.
- Probabilistic PCA can be used to model class-conditional densities and hence be applied to classification problems.
- The probabilistic PCA model can be run generatively to provide samples from the distribution.

Structure of probabilistic PCA

- We introduce an explicit latent variable $\mathbf{z}$ corresponding to the principal-component subspace with the prior distribution over $\mathbf{z}$ given by a zero-mean unit-covariance Gaussian

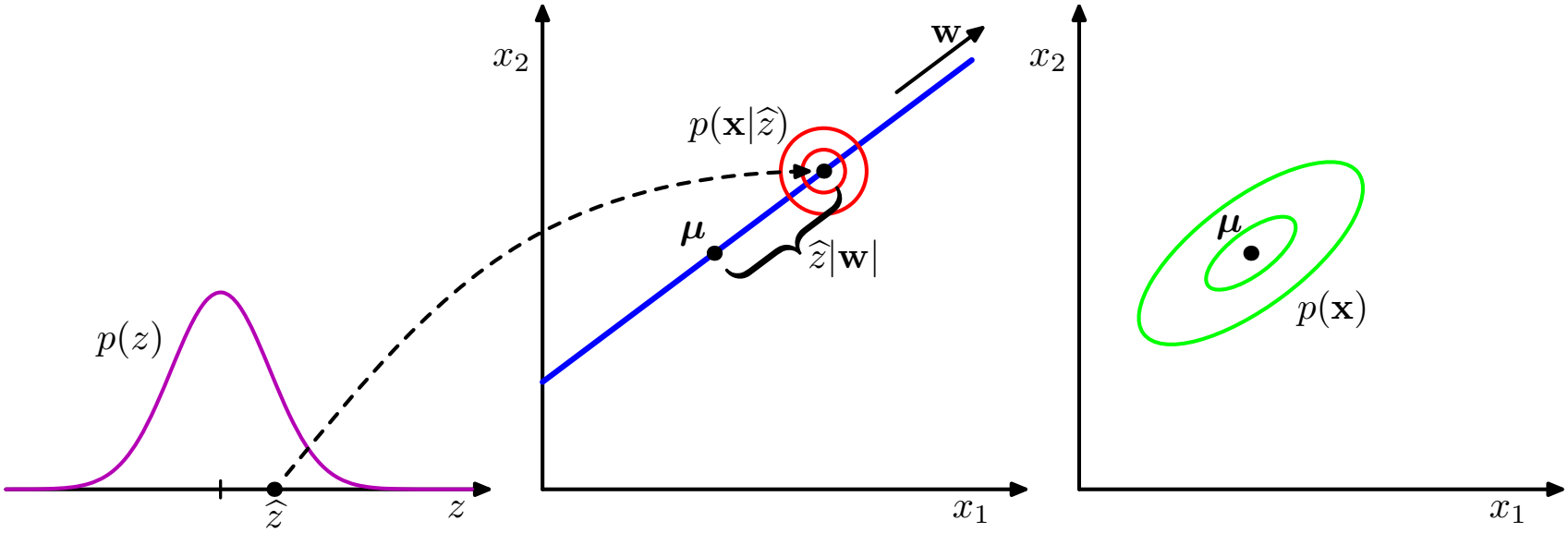
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}). \tag{1}$$

- Next we define a a Gaussian conditional distribution $p(\mathbf{x}|\mathbf{z})$ for the observed variable $\mathbf{x}$ conditioned on the value of the latent variable

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{2}$$

  in which the mean of $\mathbf{x}$ is a general linear function of $\mathbf{z}$ governed by the $D \times M$ matrix $\mathbf{W}$ and the D-dimensional vector $\boldsymbol{\mu}$.

PPCA as a generative model



The green ellipses show the density contours for the marginal distribution $p(\mathbf{x})$.

- A sampled value of the observed variable is obtained by first choosing a value for the latent variable and then sampling the observed variable conditioned on this latent value.
- The D-dimensional observed variable $\mathbf{x}$ is defined by a linear transformation of the M-dimensional latent variable $\mathbf{z}$ plus additive Gaussian 'noise', so that

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{z}$ is an M-dimensional Gaussian latent variable, and $\boldsymbol{\epsilon}$ is a D-dimensional zero-mean Gaussian-distributed noise variable with covariance $\sigma^2\mathbf{I}$.

Related distributions

- The marginal distribution $p(\mathbf{x})$ of the observed variable

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}. \tag{4}$$

Because this corresponds to a linear-Gaussian model, this marginal distribution is again Gaussian

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C}) \tag{5}$$

where the $D \times D$ covariance matrix $\mathbf{C}$ is defined by

$$\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}. \tag{6}$$

- We also require the posterior distribution $p(\mathbf{z}|\mathbf{x})$, which is

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2}\mathbf{M}\right). \tag{7}$$

where the $M \times M$ matrix $\mathbf{M}$ is defined by

$$\mathbf{M} = \mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}. \tag{8}$$

University of BRISTOL

Data projection

- Probabilistic PCA is most naturally expressed as a mapping from the latent space into the data space. For applications such as visualization and dimensionality reduction, we can reverse this mapping using Bayes' theorem.
- Any point $\mathbf{x}$ in data space can then be summarized by its posterior mean and covariance in latent space. The mean is given by

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}}) \tag{9}$$

  This projects to a point in data space given by

$$\mathbf{W}\mathbb{E}[\mathbf{z}|\mathbf{x}] + \boldsymbol{\mu}. \tag{10}$$

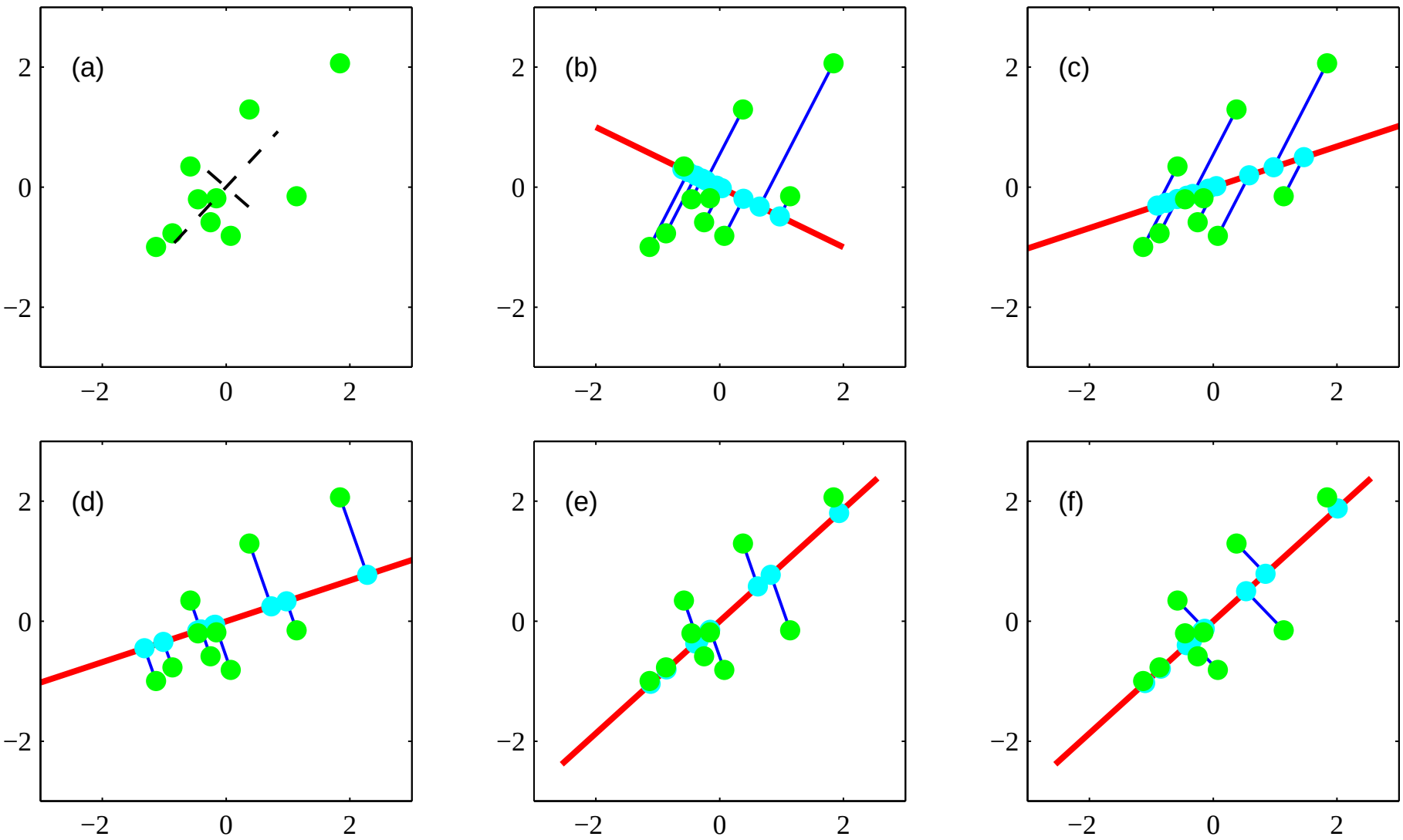
- Similarly, the posterior covariance is given from (7) by $\sigma^2\mathbf{M}^{-1}$ and is independent of $\mathbf{x}$. If we take the limit $\sigma^2 \to 0$, then the posterior mean reduces to

$$(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}})^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}}) \tag{11}$$

  which is an orthogonal projection of the data point onto the latent space, and so we recover the standard PCA model.
- For $\sigma^2 > 0$, the latent projection is shifted towards the origin, relative to the orthogonal projection.

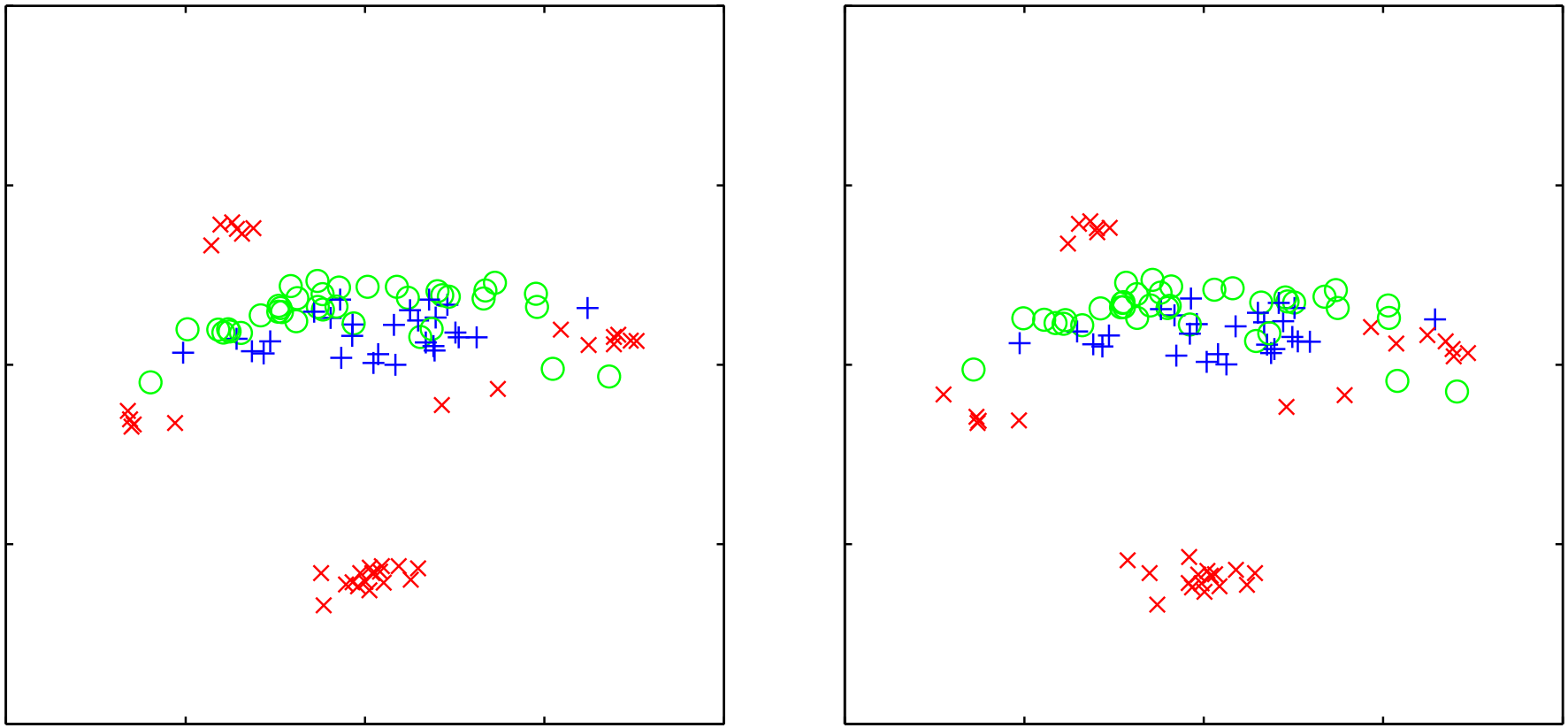
University of BRISTOL

EM algorithm for PCA



University of BRISTOL

PPCA with missing data



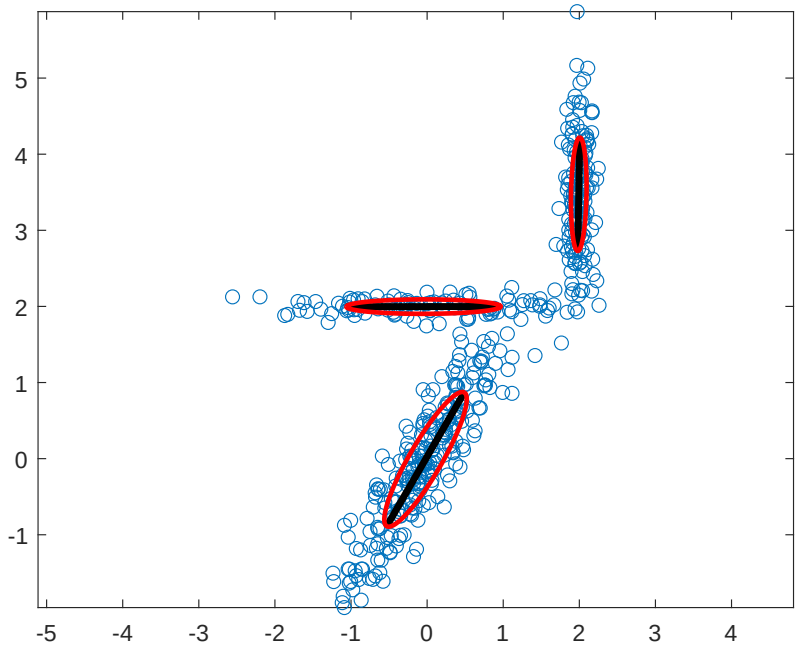
Probabilistic PCA visualization of a portion of the oil flow data set for the first 100 data points. The left-hand plot shows the posterior mean projections of the data points on the principal subspace. The right-hand plot is obtained by first randomly omitting 30% of the variable values and then using EM to handle the missing values. Note that each data point then has at least one missing measurement but that the plot is very similar to the one obtained without missing values.

Mixtures of probabilistic PCAs

- The covariance matrix is the sum of two terms: one is diagonal in a q-dimensional subspace spanned by the first q principal components and the other is spherical.
- M step

$$\mathbf{S}_j = \frac{1}{\pi_j^{(m+1)} N} \sum_{n=1}^{N} P^{(m)}(j|\mathbf{x}_n) (\mathbf{x}_n - \boldsymbol{\mu}_j^{(m+1)})(\mathbf{x}_n - \boldsymbol{\mu}_j^{(m+1)})^T.$$

where $\mathbf{S}_j$ is the covariance matrix computed for data weighted by the responsibility of the jth component.



The first cluster has axis aligned variance and centre $(0, 2)$. The variance parallel to the x-axis is significantly greater than that parallel to the y-axis. The second cluster has variance axes rotated by 30 degrees and centre $(0, 0)$. The third cluster has significant variance parallel to the y-axis and centre $(2, 3.5)$.

GTM: the Generative Topographic Mapping

- A single low-dimensional linear space is not a realistic model for many real datasets. The aim of the Generative Topographic Mapping (GTM) is to allow a non-linear transformation from latent space to data space but still make the model computationally tractable.
- The data is modelled by a mixture of Gaussians, in which the centres of the Gaussians are constrained to lie on a lower dimensional manifold.
- The topographic nature of the mapping comes about because the kernel centres in the data space preserve the structure of the latent space.
- By careful selection of the form of the non-linear mapping (using an RBF network), it is possible to train the model using a generalisation of the EM algorithm.

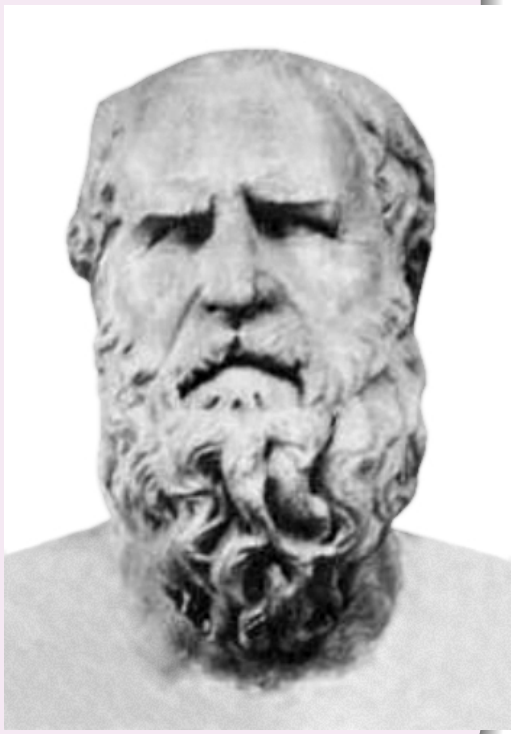
University of BRISTOL

Latent variables

Hidden Connections

A hidden connection is stronger than an obvious one.

Heraclitus



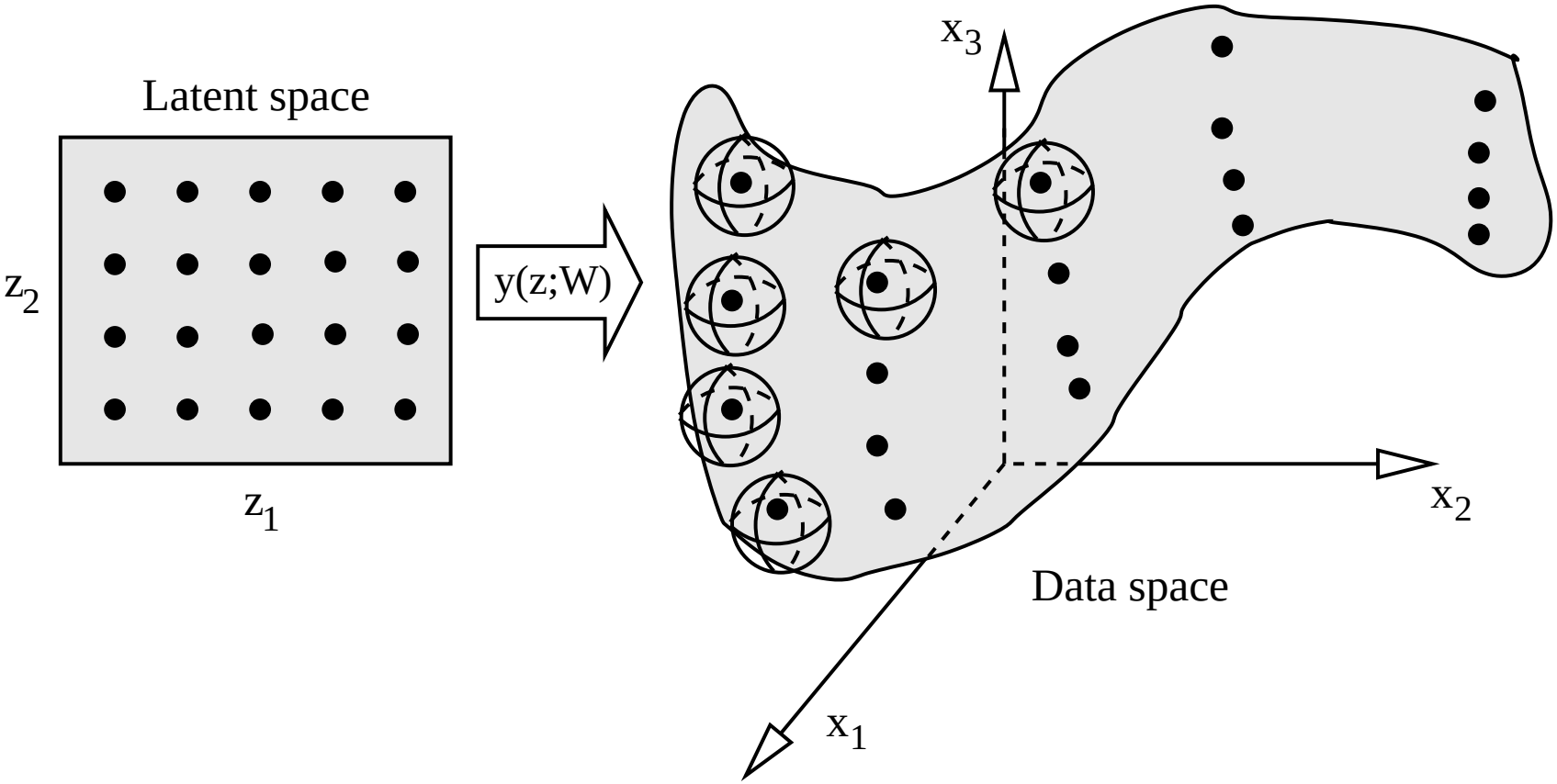
University of BRISTOL

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 $\mathcal{D}$.
- Data doesn't live exactly on manifold, so smear it with Gaussian noise. Introduce latent space density $p(\mathbf{x})$: approximate by a data sample. This is a generative probabilistic model.
- 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 non-linearity of the sheet and use this to understand the visualisation plot.
- Train the model in maximum likelihood framework using an iterative algorithm (EM).

University of BRISTOL

GTM schematic



Model for GTM

- We shall represent the data $\mathbf{x} = (x_1, \ldots, x_d)$ in a d-dimensional space using a q-dimensional latent variable space $\mathbf{z} = (z_1, \ldots, z_q)$.
- The two spaces are linked by a (non-linear) function $\mathbf{y}(\mathbf{z}; \mathbf{W})$ which maps $\mathbf{z}$ to $\mathbf{y}(\mathbf{z}; \mathbf{W})$ and is parameterised with the matrix $\mathbf{W}$.
- We add a noise model for $\mathbf{x}$: a spherical Gaussian with variance σ^2 so that the data density conditional on the latent variables is given by

$$p(\mathbf{x}|\mathbf{z}, \mathbf{W}, \sigma) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{\|y(\mathbf{z}; \mathbf{W}) - \mathbf{x}\|^2}{2\sigma^2}\right\}. \tag{12}$$

- The density in data space is then obtained by integrating out the latent variables:

$$p(\mathbf{x}|\mathbf{W}, \sigma) = \int p(\mathbf{x}|\mathbf{z}, \mathbf{W}, \sigma) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}. \tag{13}$$

  However, for a general model $\mathbf{y}(\mathbf{z}; \mathbf{W})$, this integral is analytically intractable.

University of BRISTOL

Latent space for GTM

- Let the density $p(\mathbf{z})$ be given by a sum of delta functions centred on nodes $\mathbf{z}_1, \ldots, \mathbf{z}_M$ in latent space:

$$p(\mathbf{z}) = \frac{1}{M} \sum_{j=1}^{M} \delta(\mathbf{z} - \mathbf{z}_j). \tag{14}$$

  If the nodes are uniformly spread in latent space, this is an approximation to a uniform distribution.

- Equation (13) is now tractable, and becomes a simple sum of M Gaussians:

$$p(\mathbf{x}|\mathbf{W}, \sigma) = \frac{1}{M} \sum_{j=1}^{M} p(\mathbf{x}|\mathbf{z}_j, \mathbf{W}, \sigma). \tag{15}$$

  This is a mixture model where all the kernels have the same mixing coefficient $1/M$ and variance σ^2, and the jth centre is given by $\mathbf{y}(\mathbf{z}_j; \mathbf{W})$.

- It is a constrained mixture model because the centres are not independent but are related by the mapping $\mathbf{y}$. If this mapping is smooth, then the centres will necessarily be topographically related in the sense that two points $\mathbf{z}_a$ and $\mathbf{z}_b$ which are close in latent space will be mapped to points $\mathbf{y}(\mathbf{z}_a; \mathbf{W})$ and $\mathbf{y}(\mathbf{z}_b; \mathbf{W})$ which are close in data space.

University of BRISTOL

EM algorithm for GTM

- We write

$$\mathbf{y}(\mathbf{z};\mathbf{W}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{z}), \tag{16}$$

  where $\boldsymbol{\phi}(\mathbf{z})$ are K fixed basis functions $\phi_i(\mathbf{z})$, and $\mathbf{W}$ is a $d \times K$ matrix.
- The E-step of the algorithm is the same as that for the Gaussian mixture model; recall that GTM is a constrained mixture of Gaussians.
- The M-step consists of solving this equation

$$\boldsymbol{\Phi}^T\mathbf{G}^{(m)}\boldsymbol{\Phi}(\mathbf{W}^{(m+1)})^T = \boldsymbol{\Phi}^T\mathbf{R}^{(m)}\mathbf{X}, \tag{17}$$

  where $\boldsymbol{\Phi}$ is the $M \times K$ RBF design matrix with elements $\Phi_{ji} = \phi_i(\mathbf{z}_j)$, $\mathbf{X}$ is the $N \times d$ data matrix, $\mathbf{R}$ is an $M \times N$ responsibility matrix with elements R_{jn}, and $\mathbf{G}$ is an $M \times M$ diagonal matrix with elements

$$G_{jj} = \sum_{n=1}^{N} R_{jn}(\mathbf{W}, \sigma). \tag{18}$$

University of BRISTOL

EM algorithm implementation

- It is straightforward to solve (17) using standard techniques from linear algebra, providing some care is taken over ill-conditioning (using the pseudo-inverse).
- It can be shown that the data likelihood increases at each step of the algorithm until a local maximum is reached.
- The RBF basis function parameters control the complexity, or smoothness, of the map from latent space to data space.
- Increasing the number of latent space sample points $\mathbf{z}_i$ can only improve the data model. If there are too few sample points compared to the number of basis functions, then the Gaussian components in data space become relatively independent and there is effectively no manifold.
- For initialisation, we use a q-dimensional linear subspace as the initial manifold, and this can be found easily using PCA. Then we determine $\mathbf{W}$ by minimising the error

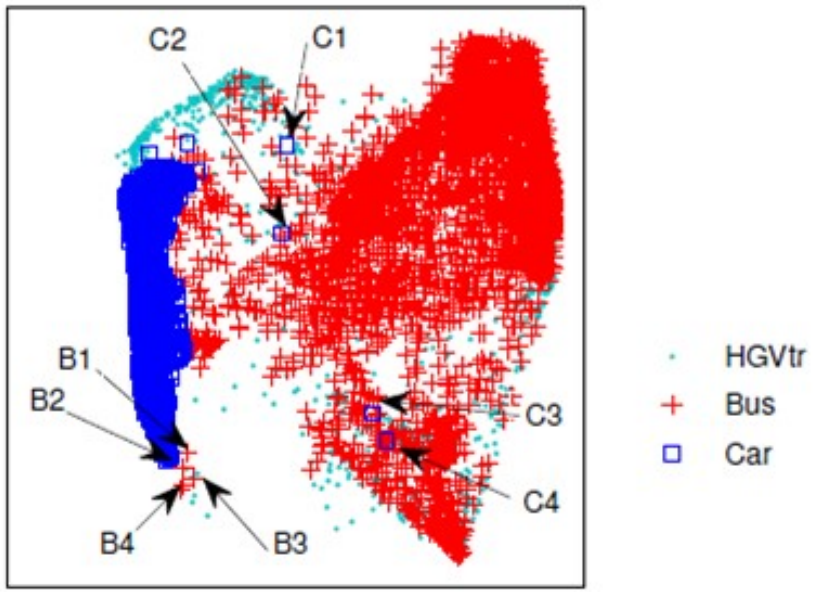
$$E = \frac{1}{2} \sum_{j=1}^{M} \|\mathbf{W}\phi(\mathbf{z}_j) - \mathbf{U}\mathbf{z}_j\|^2 \tag{19}$$

where the columns of $\mathbf{U}$ are the relevant eigenvectors of the data covariance matrix.

University of BRISTOL

GTM example

- Typically vehicle tyres are 10% underinflated.
- This wears the tyres 8% faster leading to £500M per year in additional fuel costs.
- Road based pressure sensors provide many measurements as the tyre rolls over them.



Magnification factors

- GTM is a powerful visualisation tool but there are some aspects of data structure that it does not show clearly in its standard form. In particular, even if the data consists of well-separated clusters of points, the latent space representation will be much closer to a uniform distribution.
- The EM algorithm will attempt to place the mixture components in regions of high data density and will move the components away from regions of low data density. It can do this because the non-linear map from latent space to data space enables the manifold to stretch across regions of low data density.
- This stretching (or magnification) can be measured using techniques of differential geometry, and plotting the magnification factors in latent space allows the user to see separation between clusters.
- We need to work out the change in a small volume $\mathrm{d}V$ in latent space mapped to a small volume $\mathrm{d}V'$ on $\mathcal{M}$.

Magnification factors implementation

- We denote by $\mathbf{J}$ the $q \times d$ Jacobian matrix of the map $\mathbf{y}(\mathbf{z}; \mathbf{W})$:

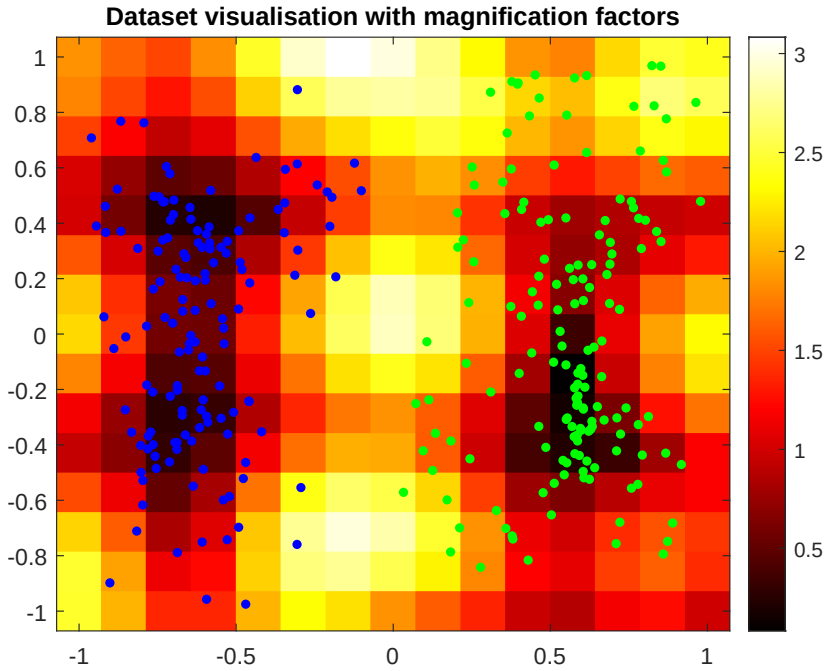
$$\mathbf{J} = (J_{kl}) = \frac{\partial y_k}{\partial z_l}. \tag{20}$$

- $\mathrm{d}V'$ is equal to $\sqrt{\det(\mathbf{J}\mathbf{J}^T)}$.
- Because the centres of the basis functions are fixed, $\mathbf{J} = \boldsymbol{\psi}\mathbf{W}$ where $\boldsymbol{\psi}$ has elements $\psi_{ji} = \partial\phi_j/\partial z^i$. Hence the magnification factors are given by

$$\frac{\mathrm{d}V'}{\mathrm{d}V} = \det{}^{1/2}\left(\boldsymbol{\psi}\mathbf{W}^T\mathbf{W}\boldsymbol{\psi}^T\right), \tag{21}$$

Magnification factors example

- We superimpose the magnification factors on the latent space visualisation.
- There is a vertical area with large magnification that clearly divides the two classes of data and makes the clustering apparent.



Enhancements to GTM

Currently a very active area of research:

- Curvatures 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.

University of BRISTOL

Enhancements to GTM

Currently a very active area of research:

- Curvatures 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.

University of BRISTOL

Enhancements to GTM

Currently a very active area of research:

- Curvatures 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.

University of BRISTOL

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.

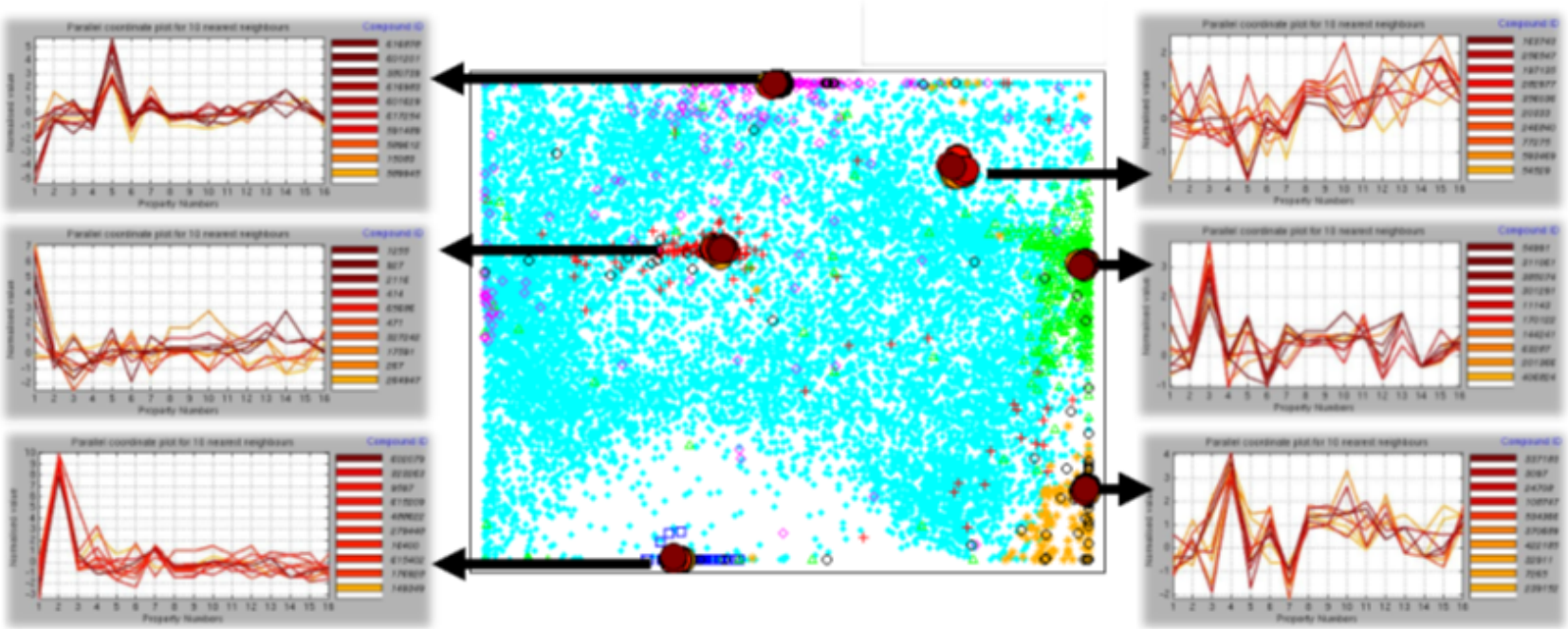
University of BRISTOL

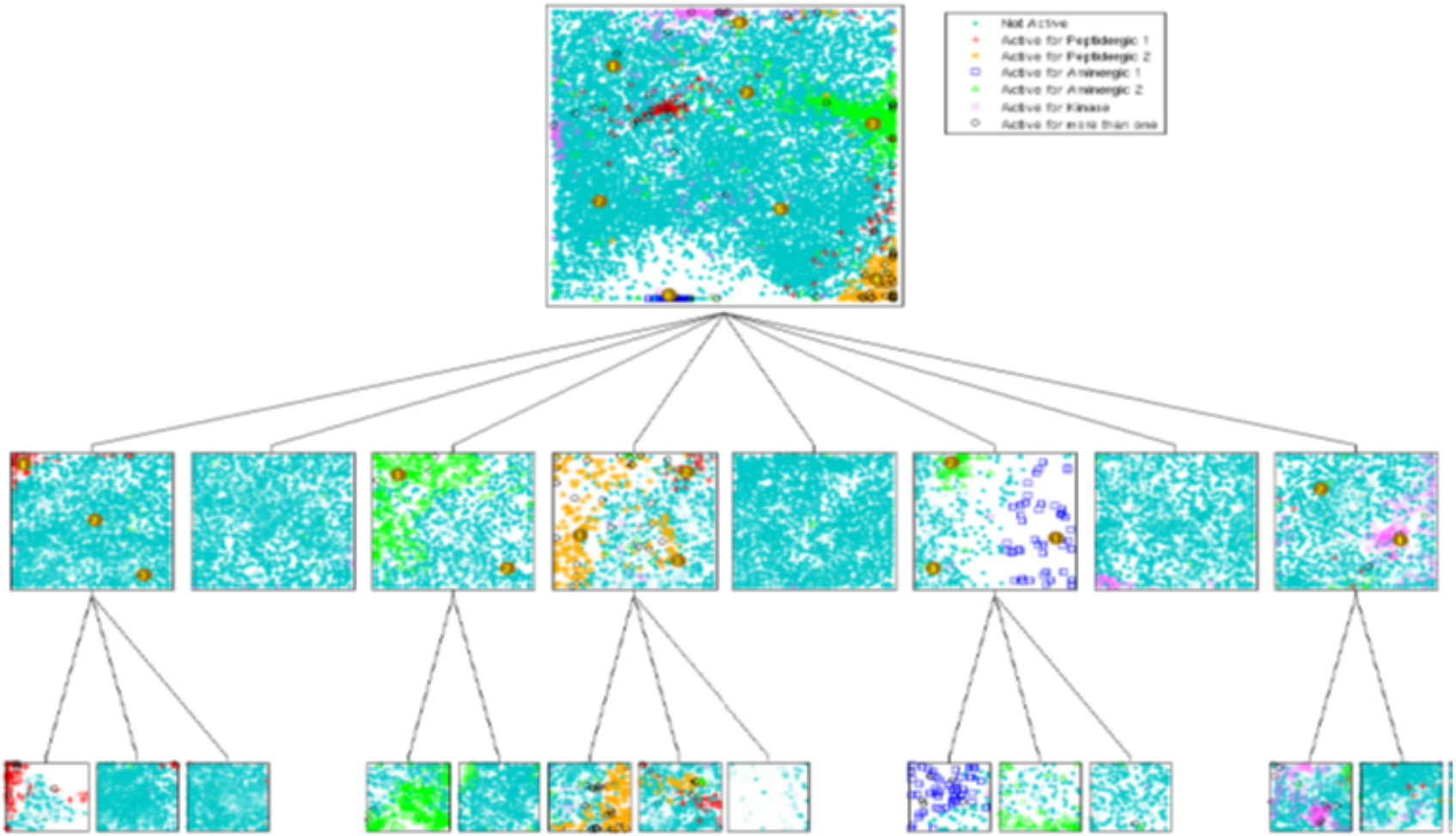
Chemometric Application: HTS Data Exploration

- Scientists at Pfizer searching for active compounds can now screen millions 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.
- Build local prediction models.

University of BRISTOL

- We have taken data from Jens Lösel (Pfizer) which consists of 6912 14-dimensional vectors representing chemical compounds using topological indices 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 comparable with other algorithms for logP.




Not Active
Active for Peptidergic 1
Active for Peptidergic 2
Active for Aminergic 1
Active for Aminergic 2
Active for Kinase
Active for more than one

Overview

- Understand a probabilistic version of PCA and continuous latent variable models
- Able to use mixture of probabilistic PCA
- Understand how the GTM works in theory and practice
- Aware of some extensions of GTM

University of BRISTOL