Learning to Cluster using Local Neighborhood Structure

University of Toronto

MIT

Rómer Rosales

Kannan Achan

Brendan Frey

Overview

- Introduction
  - A different clustering concept/properties
- Motivation
  - Using local neighborhood structure
  - Learning to cluster
- Clustering
  - Probability model
  - Learning to cluster
- Experiments-applications
  - Discovering noisy, sampled manifolds
  - Learning to find spatial patterns
  - Predicting gene function from gene expression
- Summary

Basic Clustering Problem

- Dataset
  - A finite set $\mathcal{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$
  - A measure or similarity between pairs of elements
- Class labels
  - A finite set of size M, *e.g.,* $\mathcal{C} = \{1, ..., M\}$
- Clustering/classification
  - Find labels $C = (c_1, ..., c_N) \in \mathcal{C}^N$ that optimize a certain function of the data points, measure, and labels.

Clustering Problem in This Work

- Dataset
  - A finite set $\mathcal{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$
  - No measure or similarity assumed beforehand
- Class labels
  - A finite set of size M, *e.g.*, $\mathcal{C} = \{1, ..., M\}$
- Clustering/classification
  - Find a posterior probability distribution over class labels given the dataset: $p(c_i | \mathcal{Z})$
- Learn to cluster from previously labeled data

  (labeled datasets are becoming increasingly popular)
- Neighborhood structure assumed relevant …

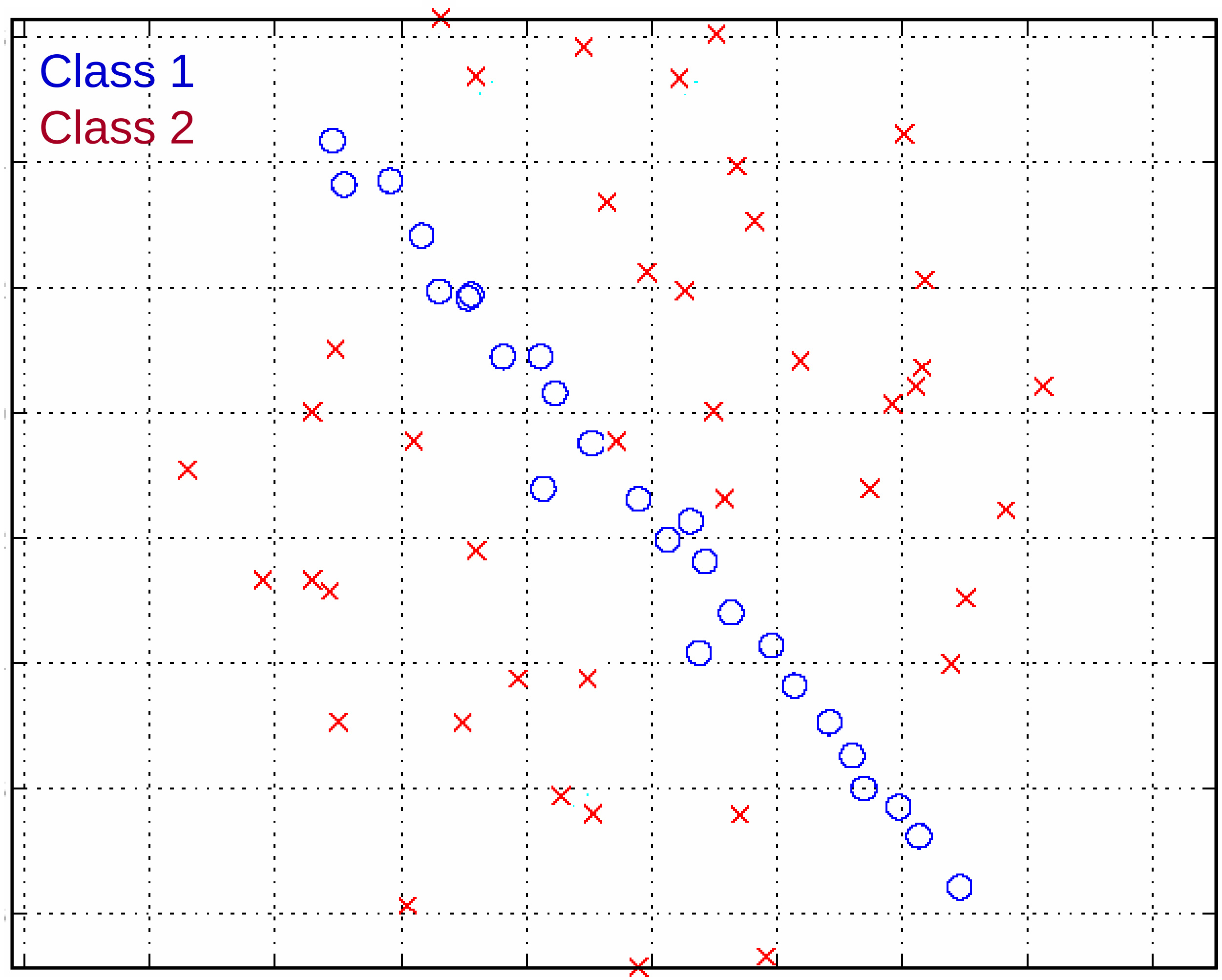
Main Conceptual Differences

- Classical clustering notions
  - Clusters should have high intra-cluster and low-inter cluster similarity
  - Clustering is defined based on pair-wise similarities between data points
  - Global measure
- Clustering using local structure
  - A cluster should be *structurally* similar *everywhere* (locally)
  - Clustering is defined based on the additional properties of the local structure of the data (in this work represented by the high-order neighborhood structure)
  - Class conditioned measure

Motivation I (Local Structure)

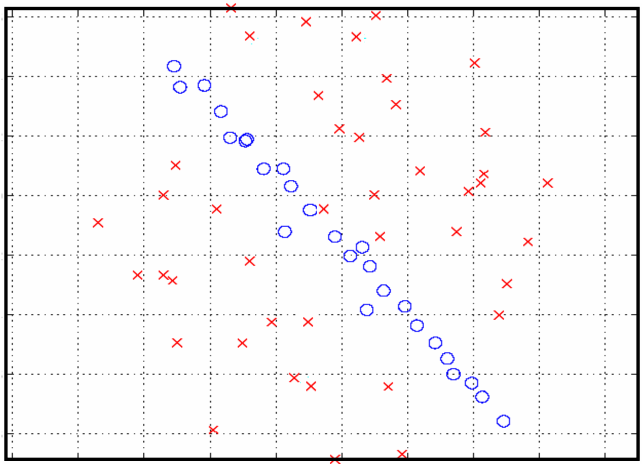
- Local structure
  - Commonly, affinities between pairs of data points are *enough* for classification
  - However, in some problems, the high-order local structure of the data is more relevant for classification

Motivation I Example

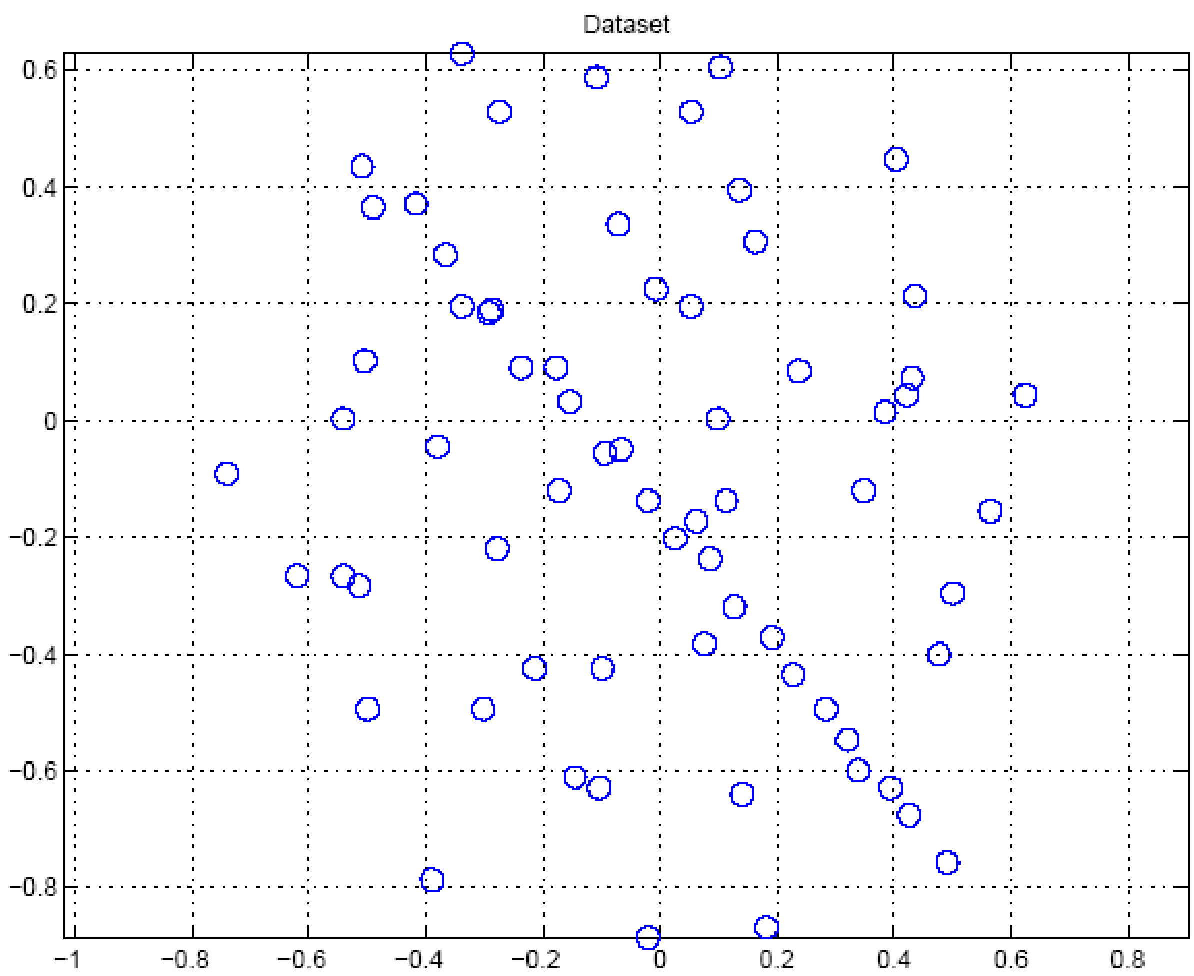


Motivation I

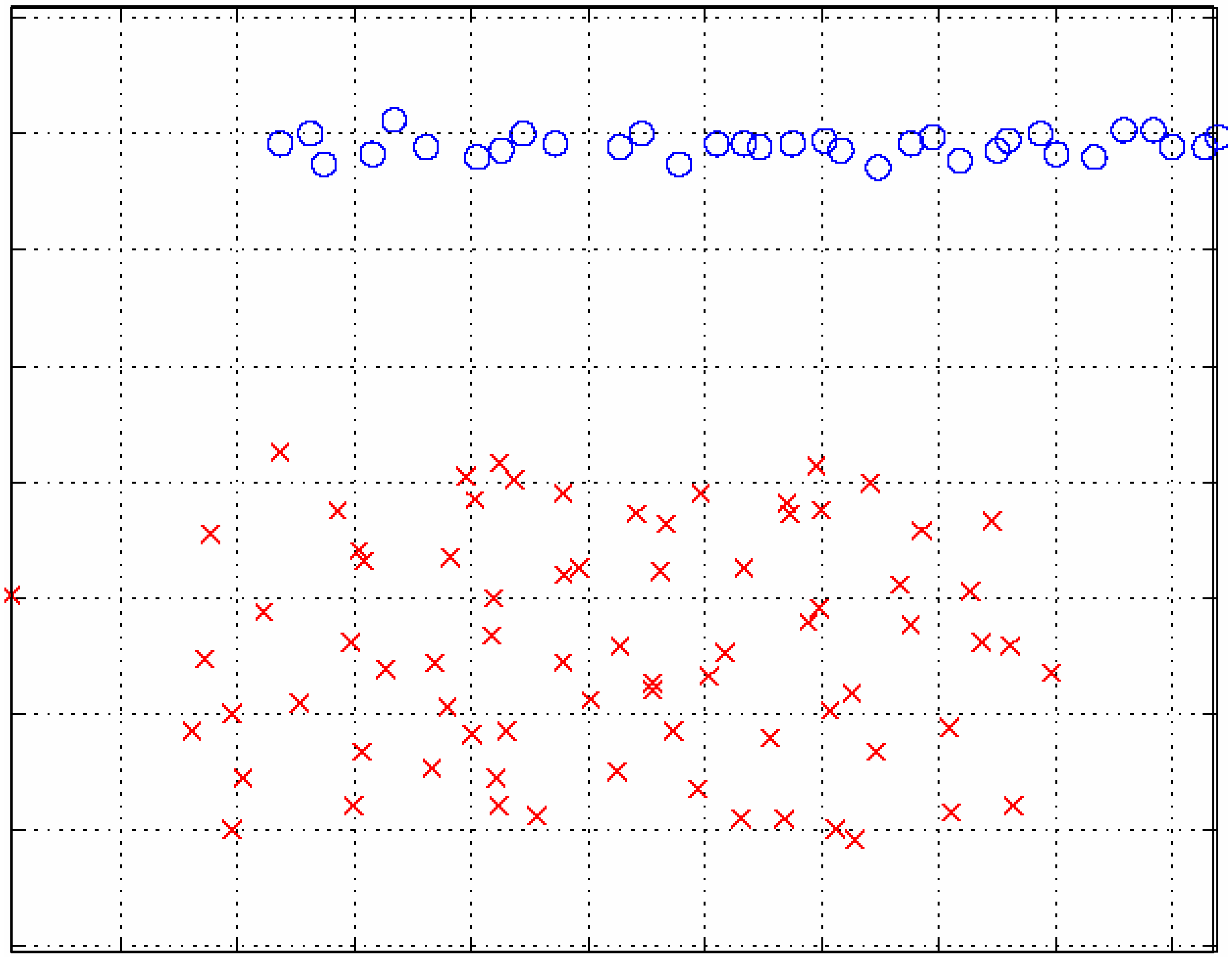
- Local structure
  - Commonly, affinities between pairs of data points are *enough* for classification
  - However, in some problems, the high-order local structure of the data is more relevant for classification
  - Concept allows to think of the notion of class-conditioned structure



Motivation II



Motivation II Example

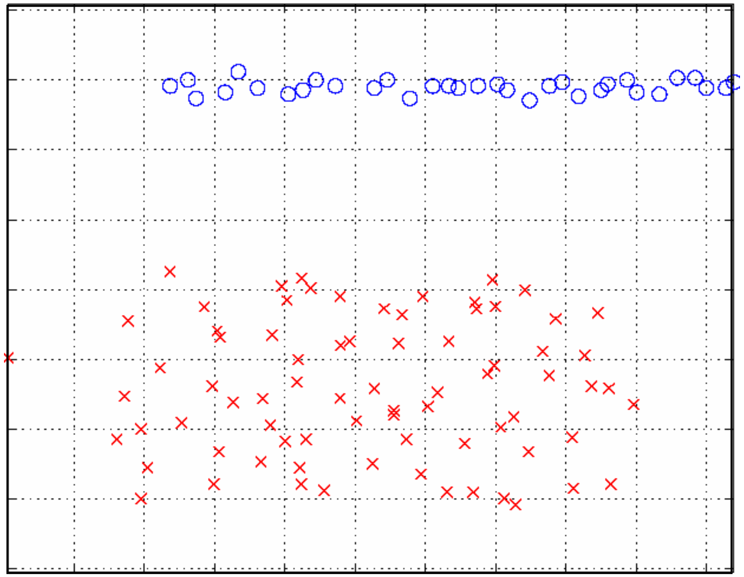


Motivation II (Learning to Cluster)

- Learning to cluster
  - A measure of similarity is rarely given
    - Hand-picked
    - Obtained after feature selection
  - Ideally, a way to *measure* likeness should be obtained directly from relevant labeled data

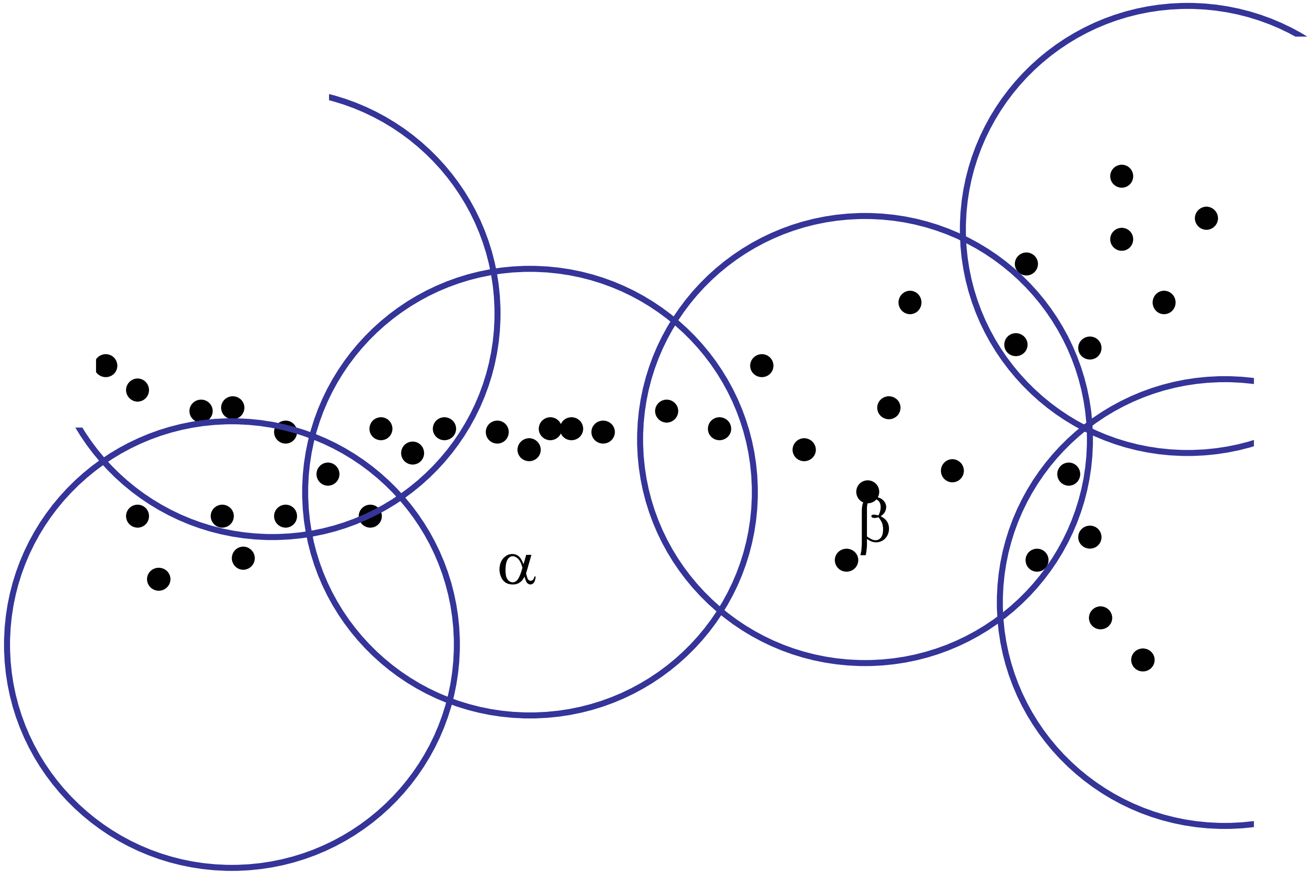
Motivation II

- Learning to cluster

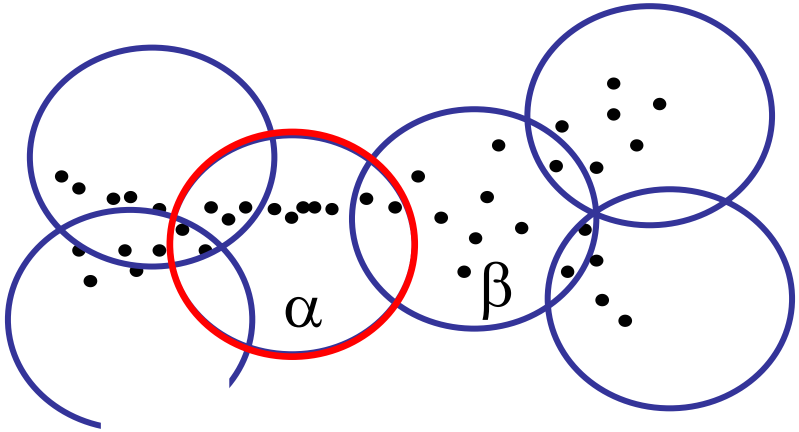


- Encoding prior knowledge
  - Use examples (*e.g.,* instead of analytical expression)
    - Example based clustering: simple/general
    - Labeled examples are becoming more readily available

Neighborhoods



Neighborhoods



- Element set η_{α} composed of K elements
  - E.g., randomly pick reference points and find its K-NN
- Structure representation

$$\mathbf{y}_{\alpha} = f(\{\mathbf{z}_i\}_{i \in \eta_{\alpha}})$$

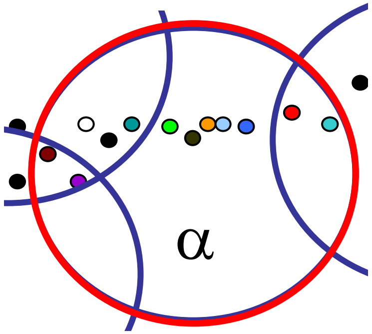
We will look $\mathbf{y}_{\alpha}$ (structure) as a random variable

Probability Models of Local Structure

- Main idea: conditioning structure on class label

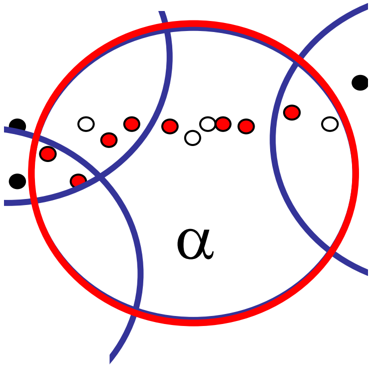
$$p(\mathbf{y}_\alpha | \mathbf{x}_\alpha)$$

- Domain $\mathcal{S}$ of $\mathbf{x}$:
  - Worst case



$$|\mathcal{S}| = |\mathcal{C}|^K$$

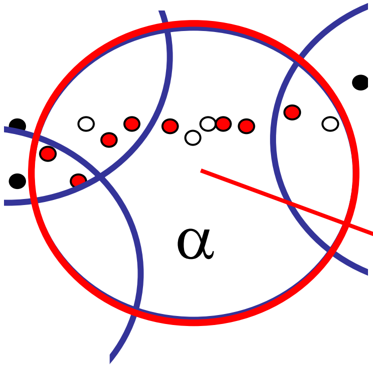
  - A more structured representation



$$|\mathcal{S}| = |\mathcal{C}|\binom{K}{K_{out}}$$

Efficient Representation of Class Labels

- A more economical representation



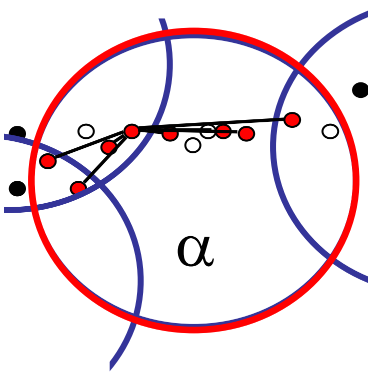
$$\mathbf{x}_\alpha = (\ell_\alpha, s_\alpha)$$

Class label Binary indicator

$$\mathbf{x}_\alpha = (c;1,1,0,1,1,1,0,0,1,1,1,0)$$

$$|\mathcal{S}| : \mathcal{O}(MK^{\min(K_{in},K_{out})})$$

- Conditional probability distribution



$$p(\mathbf{y}_\alpha|\mathbf{x}_\alpha) = p(\mathbf{d}|\ell_\alpha)$$

$$\mathbf{d} = \{d_{ij}|s_i = 1, s_j = 1\}$$

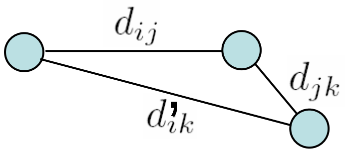
(In-class points only)

Representing Local Structure

■ High order relationships

- Collection of pair-wise relationships is appropriate to describe local structure

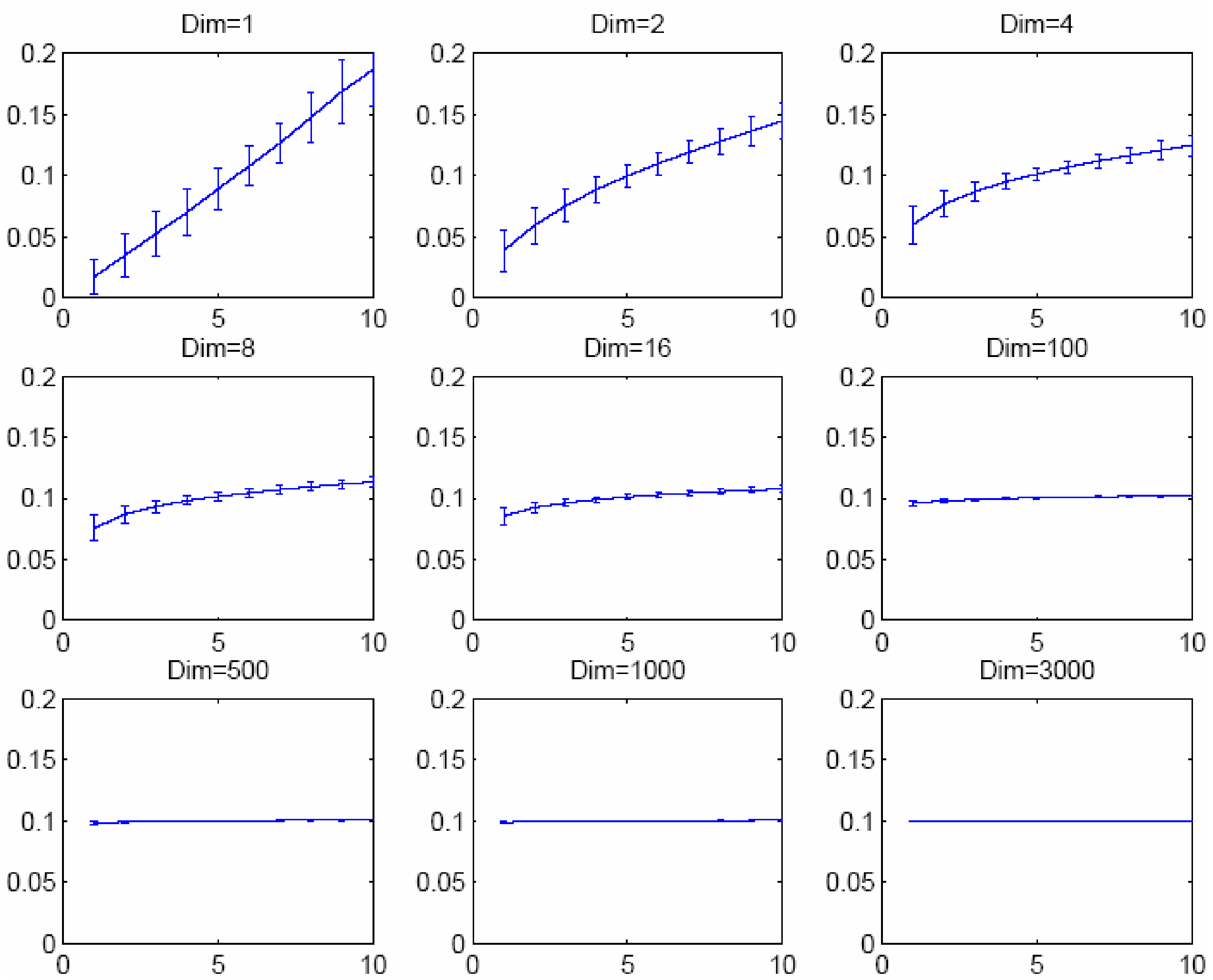
$$f(\{\mathbf{z}_k\})_k = \{f(\mathbf{z}_i, \mathbf{z}_j)\}_{(i,j)} \quad \sim$$



- Clustering is a function of structure relationships between neighborhoods

Other representations possible

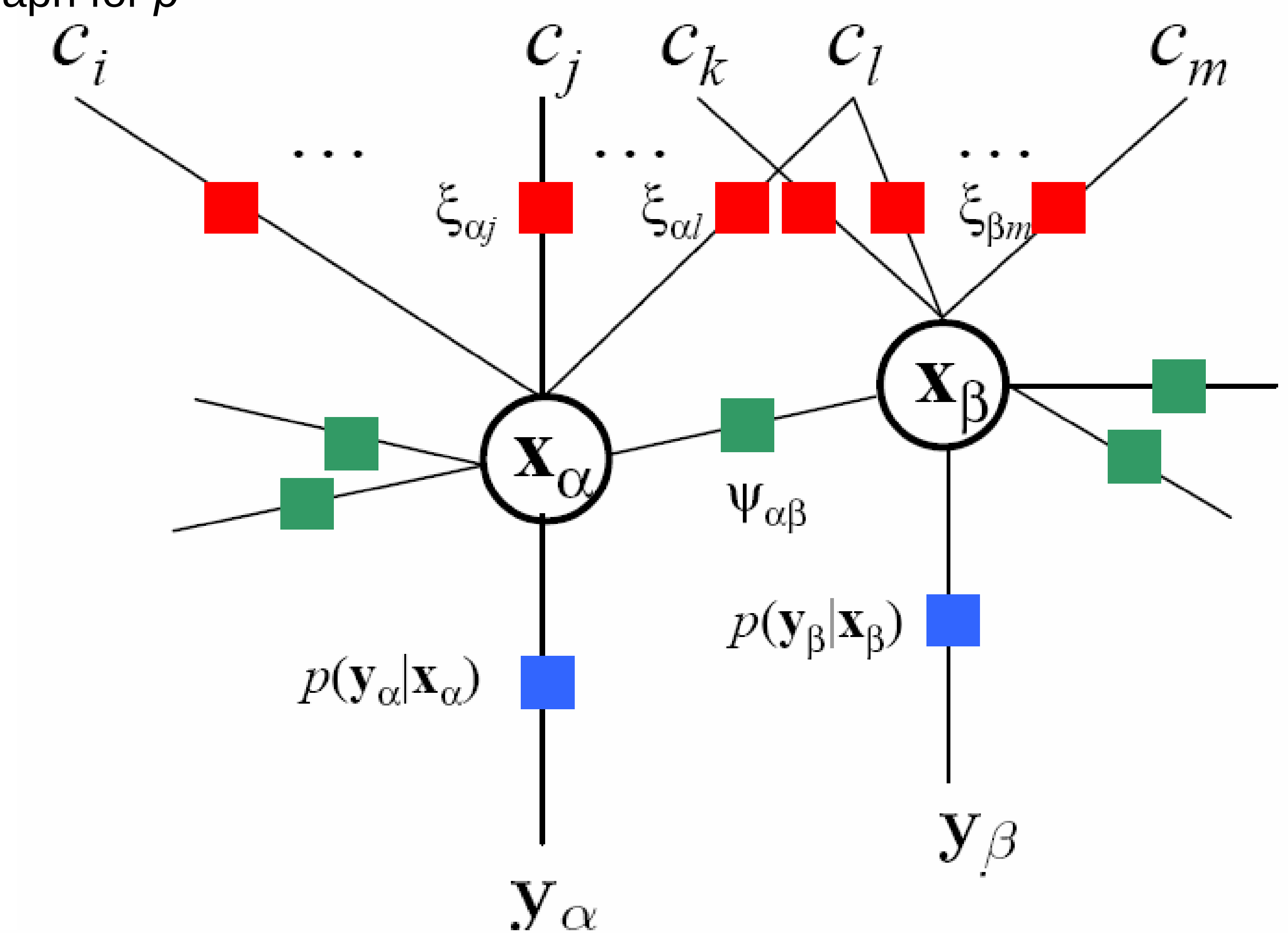
Local Structure Example



Mean distance to K=[1…10] nearest neighbors (normalized) for planar surfaces of various dimensions

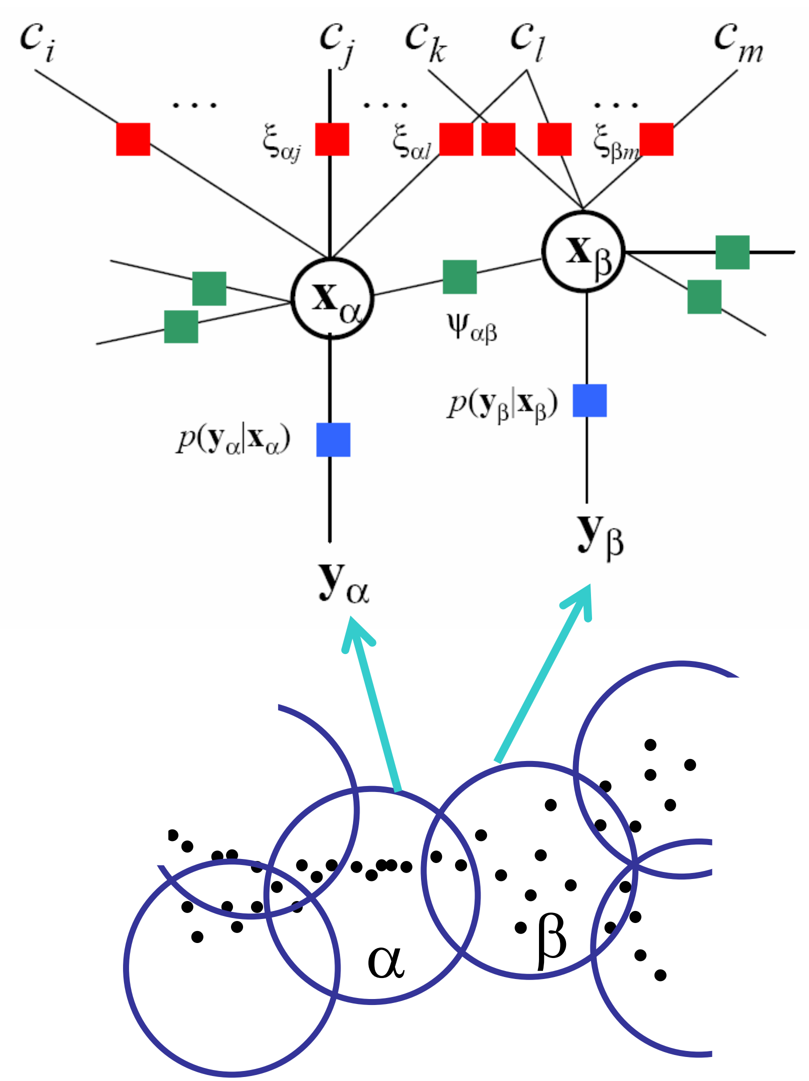
From Neighborhoods to Labels

Factor graph for *p*



$$P(\mathbf{y}, \mathbf{x}, C) \;=\; \frac{1}{Z} \prod_{\alpha} p(\mathbf{y}_\alpha|\mathbf{x}_\alpha) \prod_{(\alpha,\beta)\in \text{proxim.}} \psi(\mathbf{x}_\alpha, \mathbf{x}_\beta) \prod_{(\alpha,i)} \xi(c_i, \mathbf{x}_\alpha)$$

From Neighborhoods to Labels



Individual Labels

Neighborhood assignments and labels

$$P(\mathbf{y},\mathbf{x},C) = \frac{1}{Z}\underbrace{\prod_{\alpha} p(\mathbf{y}_\alpha|\mathbf{x}_\alpha)}_{\text{Structure conditioned on Class+ N. assignment}} \underbrace{\prod_{(\alpha,\beta)\in\text{proxim.}} \psi(\mathbf{x}_\alpha,\mathbf{x}_\beta)}_{\text{Neighborhood compatibility}}$$

Structure conditioned on Class+ N. assignment

Neighborhood compatibility

$$\underbrace{\prod_{(\alpha,i)} \xi(c_i,\mathbf{x}_\alpha)}_{\text{Individual point class constraint}}$$

Individual point class constraint

Additional Model Description

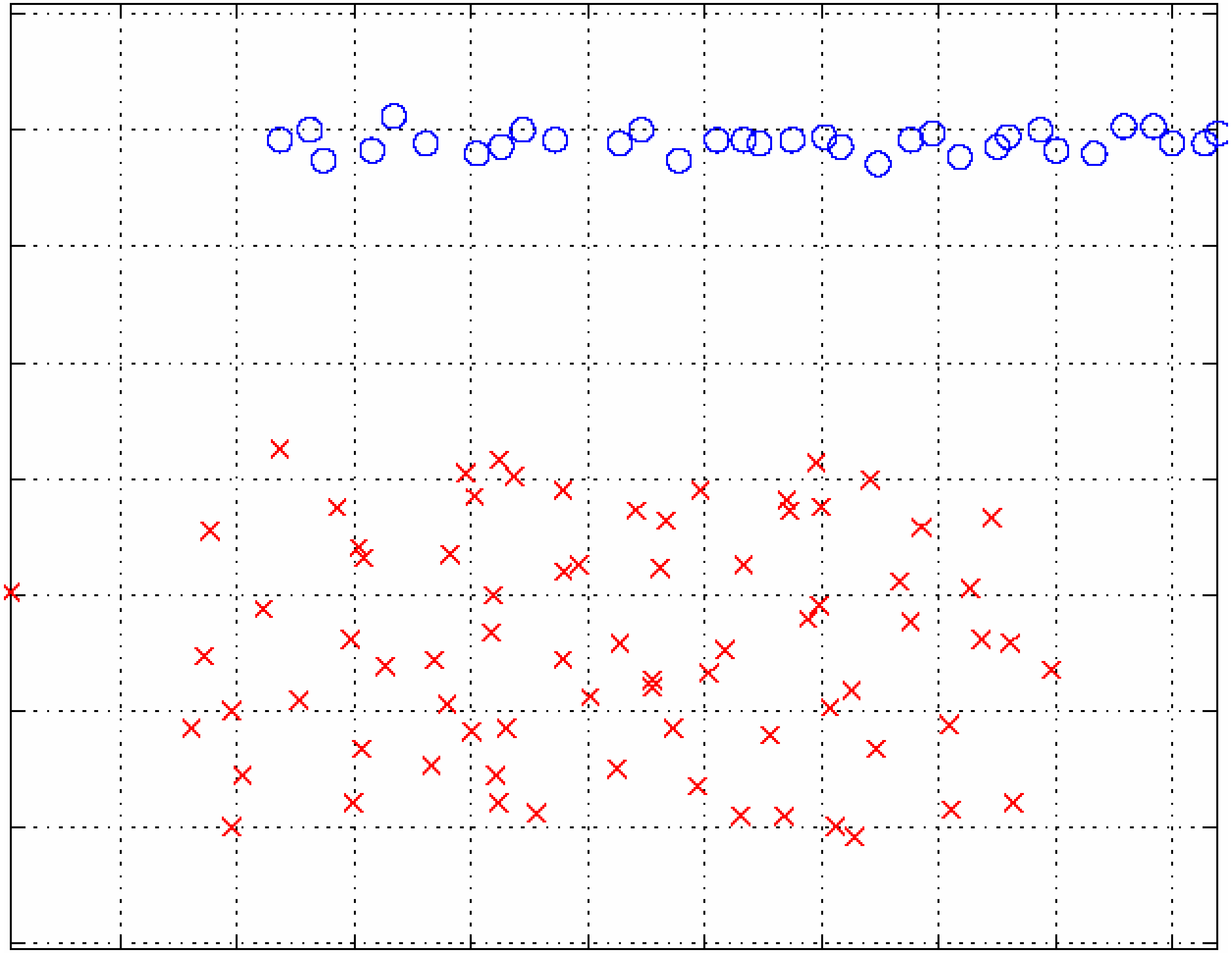
- The compatibility of two neighborhoods is inversely proportional to the number of common elements that *disagree*

$$\psi(\mathbf{x}_\alpha, \mathbf{x}_\beta) \propto \exp\{-\sum_{(i,j)\in\mathcal{P}_{\alpha\beta}} \phi(s_{\alpha i}, s_{\beta j})\}^{\delta(\ell_\alpha \neq \ell_\beta)}$$

- Each point class label must agree with its neighborhood(s) label(s)
  - Care about in-class points
  - Do not care about out-of-class points (wildcards)

$$\xi(c_i, \mathbf{x}_\alpha) = \delta(c_i - \ell_\alpha)[1 - \delta(s_i)]$$

Learning to Cluster



Learning to Cluster

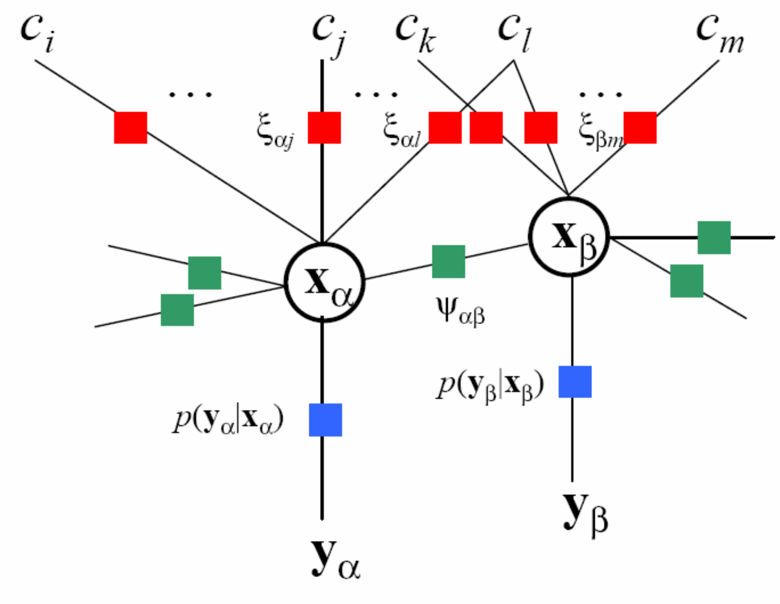
- Conceptual differences
  - Familiar clustering concepts
    - Learn a similarity measure between pairs of points (e.g., affinity matrix)
  - Clustering using local structure
    - Learn the local structure of clusters
- Learning local structure
  - Learning local structure from labeled (or partially labeled) datasets
  - Learning is equivalent to estimating $p(\mathbf{y}_\alpha | \mathbf{x}_\alpha)$!
    - Well defined task
    - Because labels are given, this can be done easily for a number of distributions (in contrast to other popular clustering models)

Extension to Unsupervised Clustering

- Familiar clustering methods
  - Changes in class label should occur in areas of low data density
- Clustering using local structure
  - Changes in class label should occur in areas where there is a change in local structure of the data (*e.g.*, where the observed structure has low probability)

Inference Problem

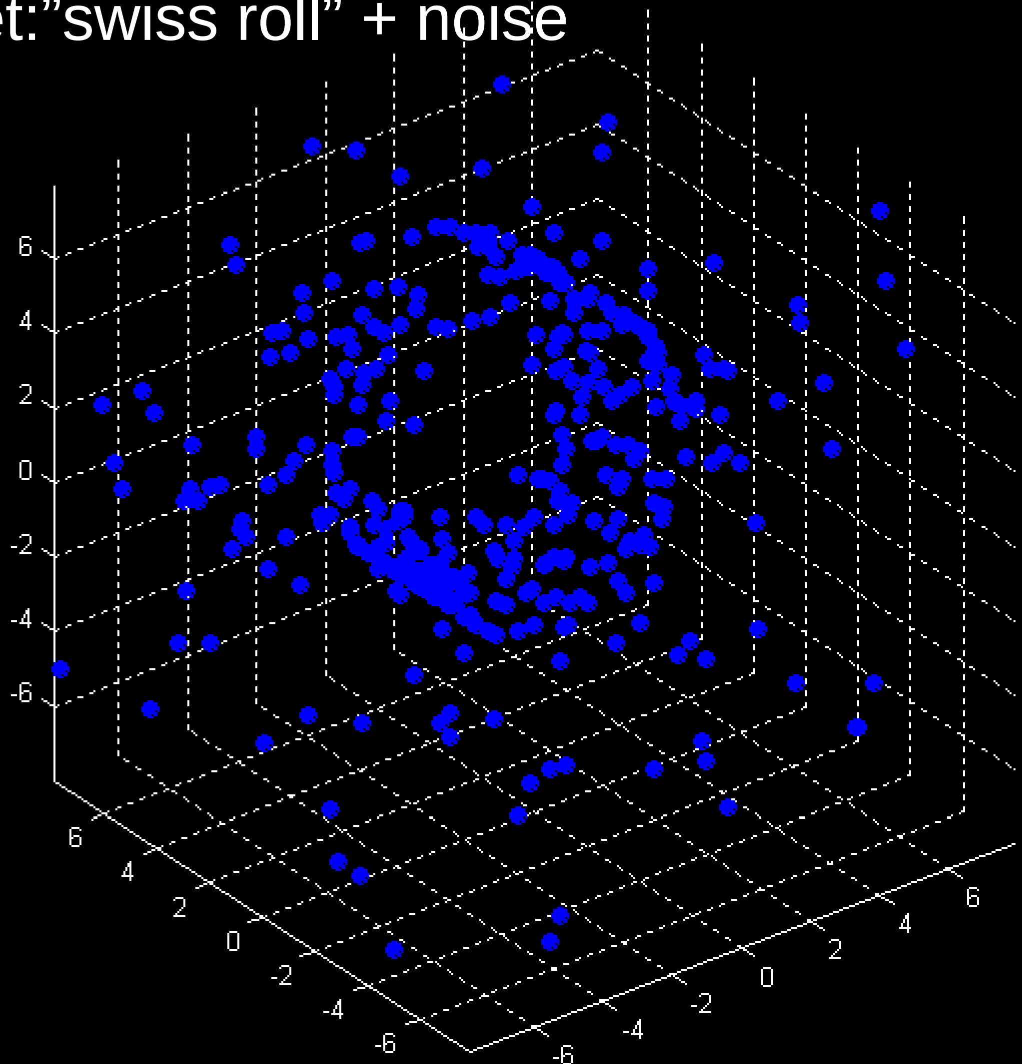
- Given the neighborhoods:
  - Infer class labels c_i
  - Infer neighborhood labels and point ownership $\mathbf{x}_\alpha = (\ell_\alpha, s_\alpha)$



- In our experiments:
  - Approximate solution by using the sum-product algorithm

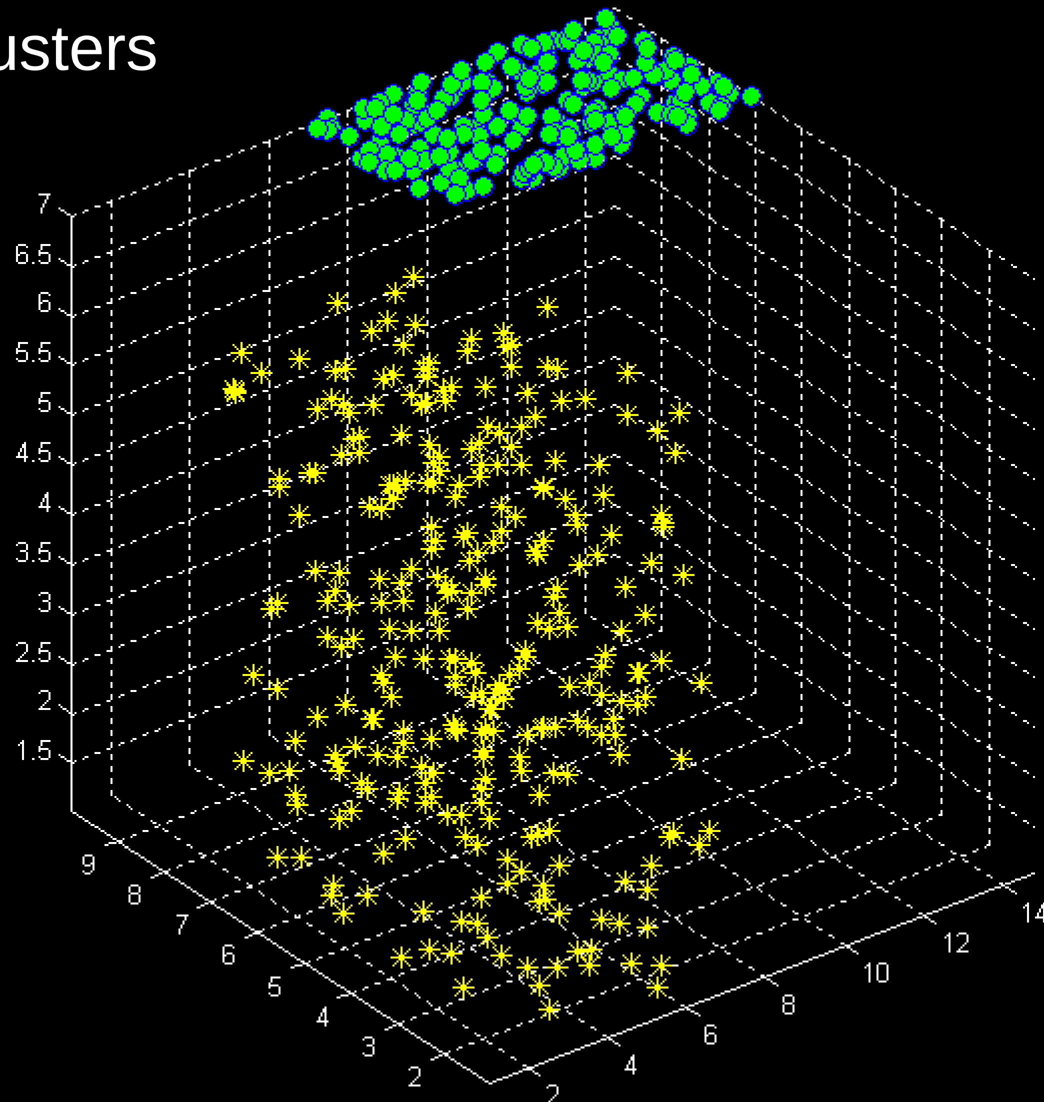
Experiments (Manifold Discovery)

Test dataset:”swiss roll” + noise



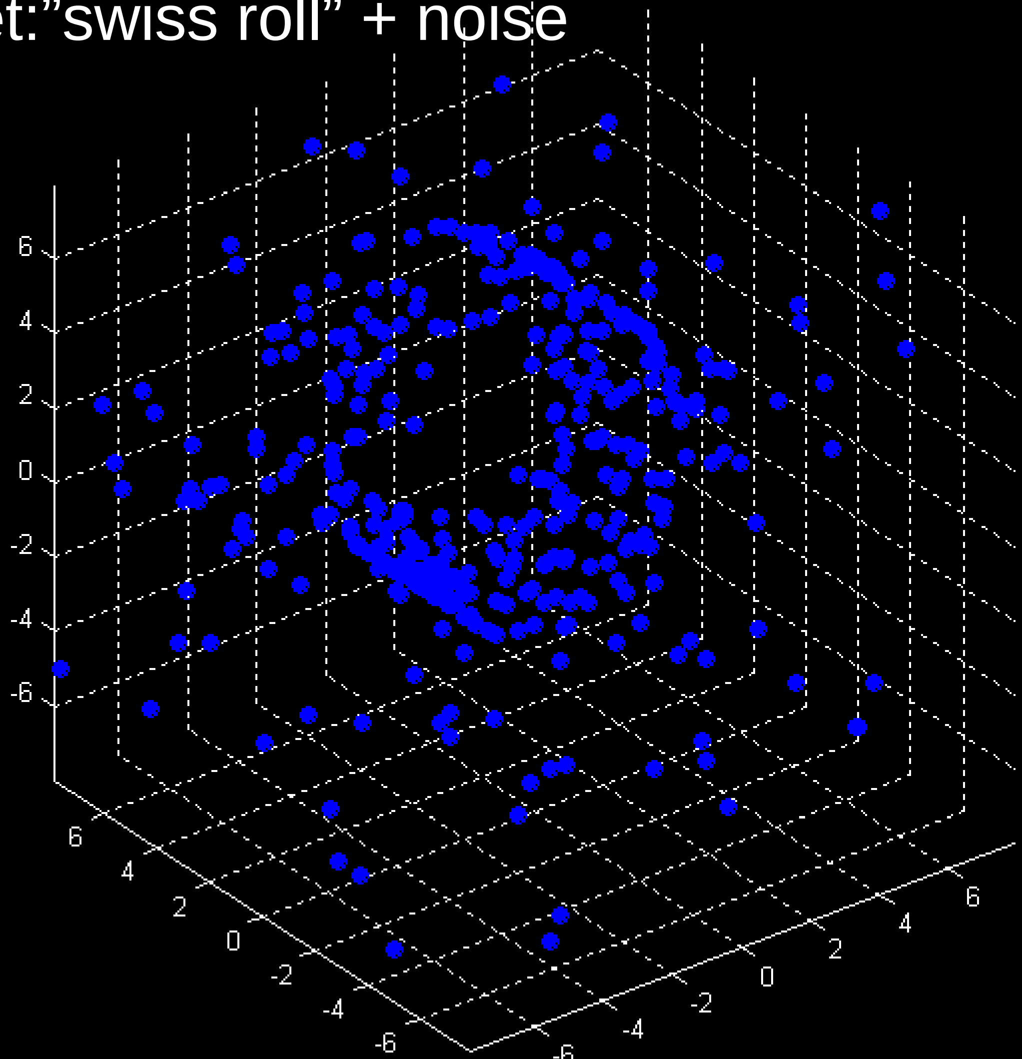
Experiments (Manifold Discovery)

Training clusters



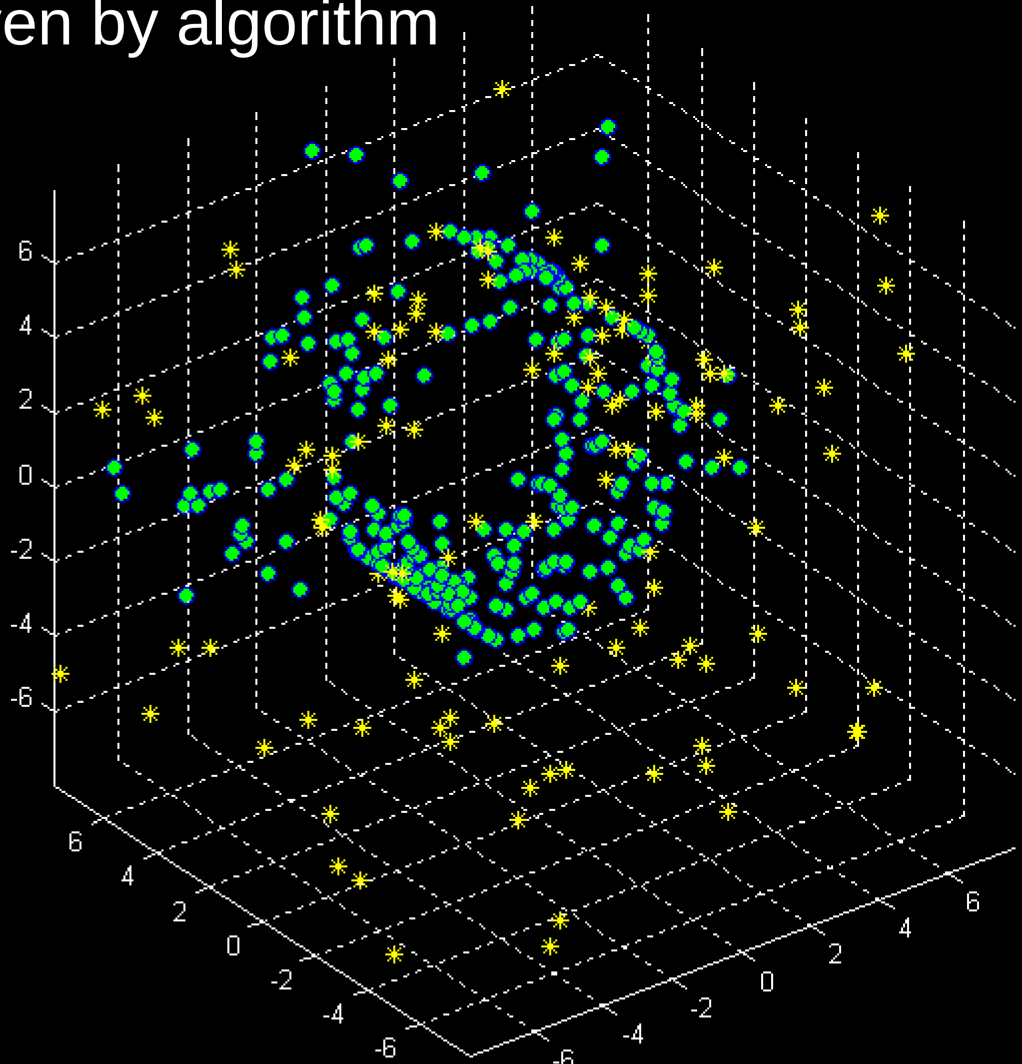
Experiments (Manifold Discovery)

Test dataset:”swiss roll” + noise



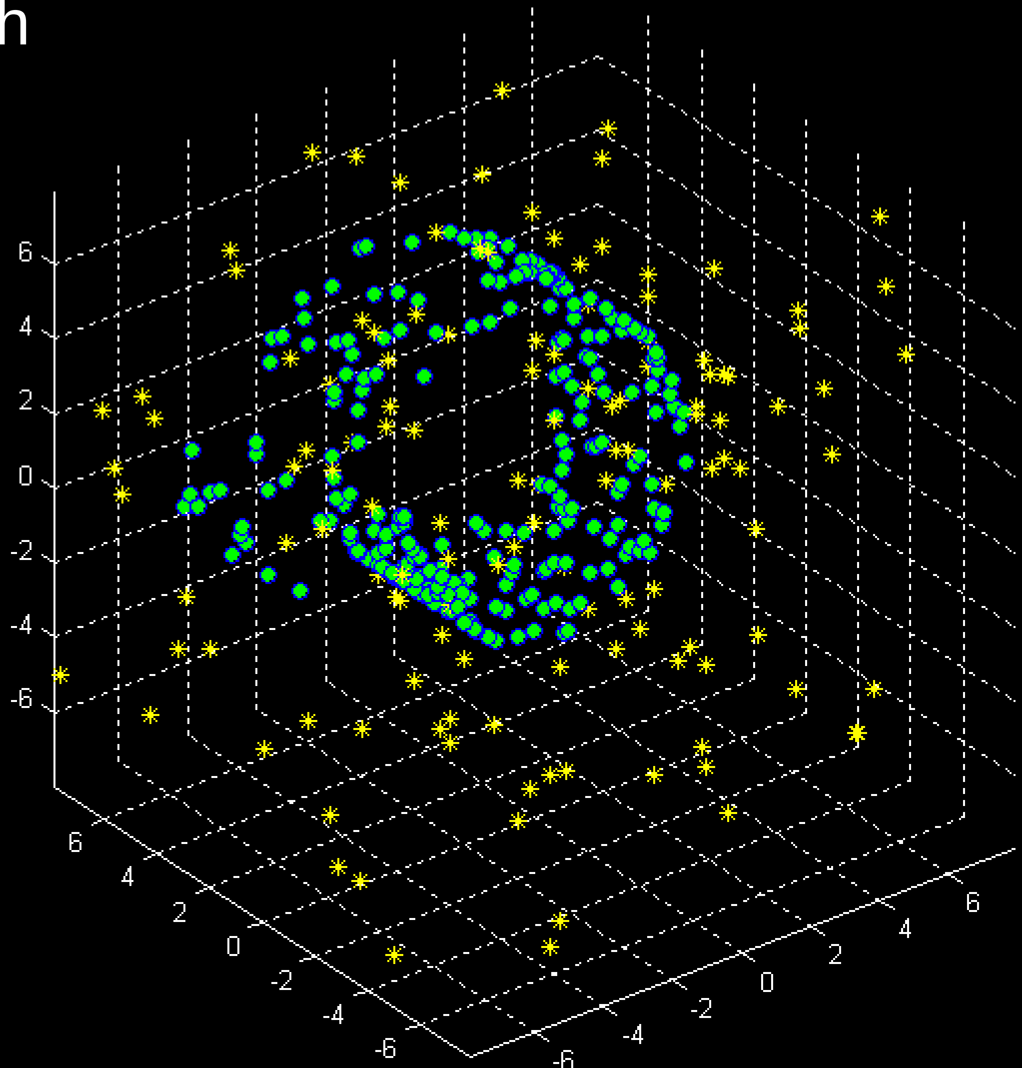
Experiments (Manifold Discovery)

Solution given by algorithm

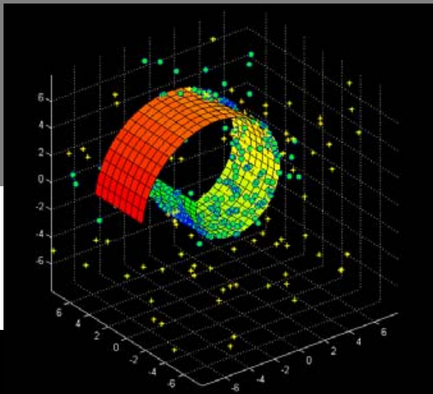


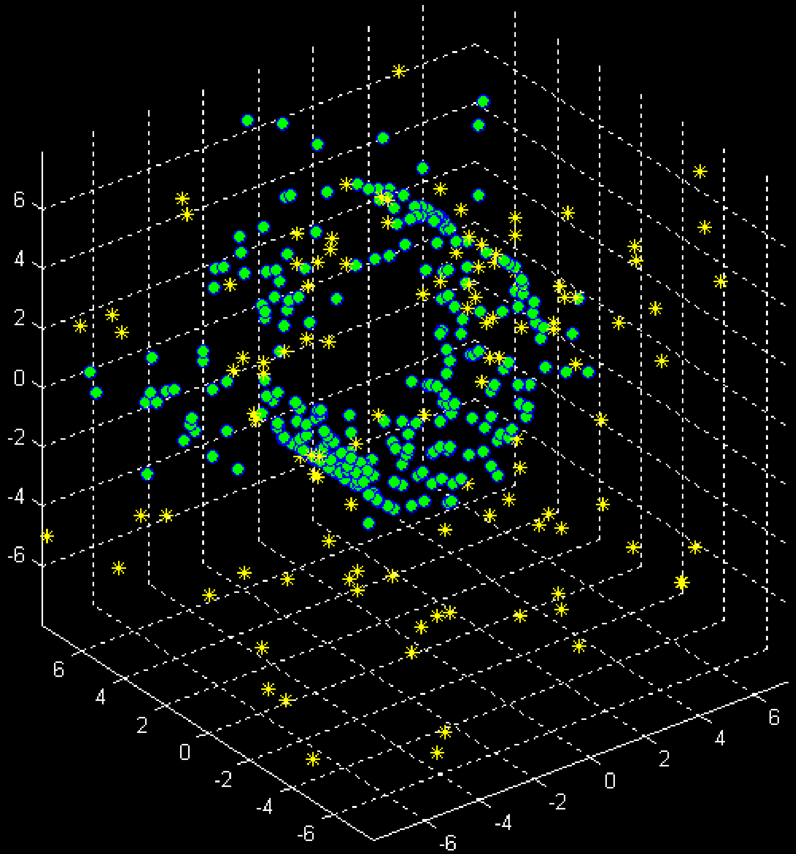
Experiments (Manifold Discovery)

Ground truth

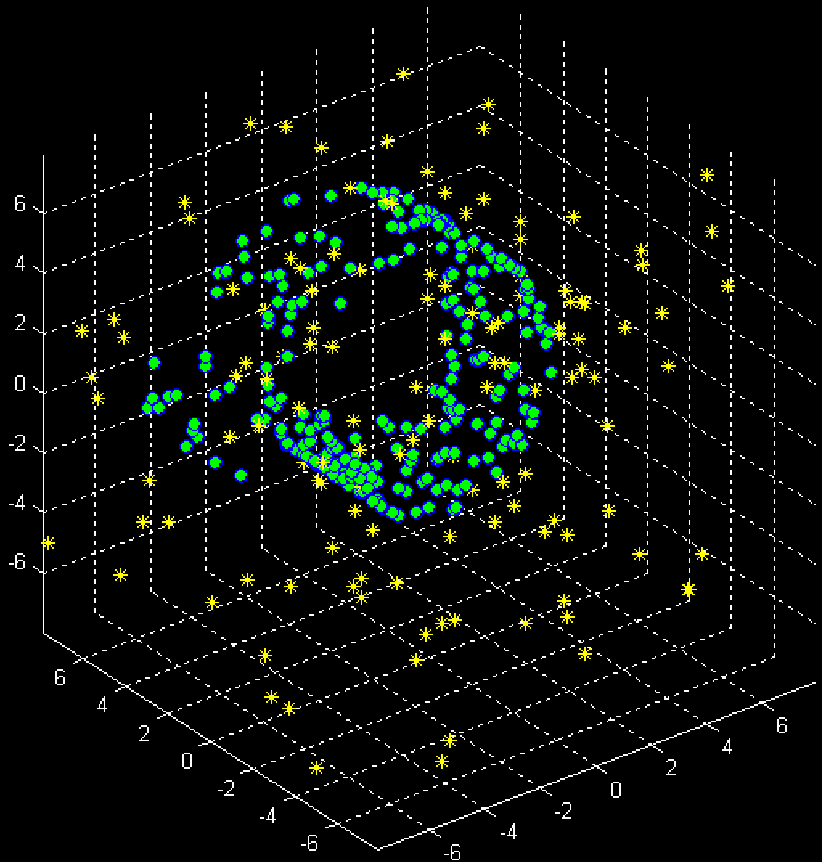


Experiments (M. Discovery)



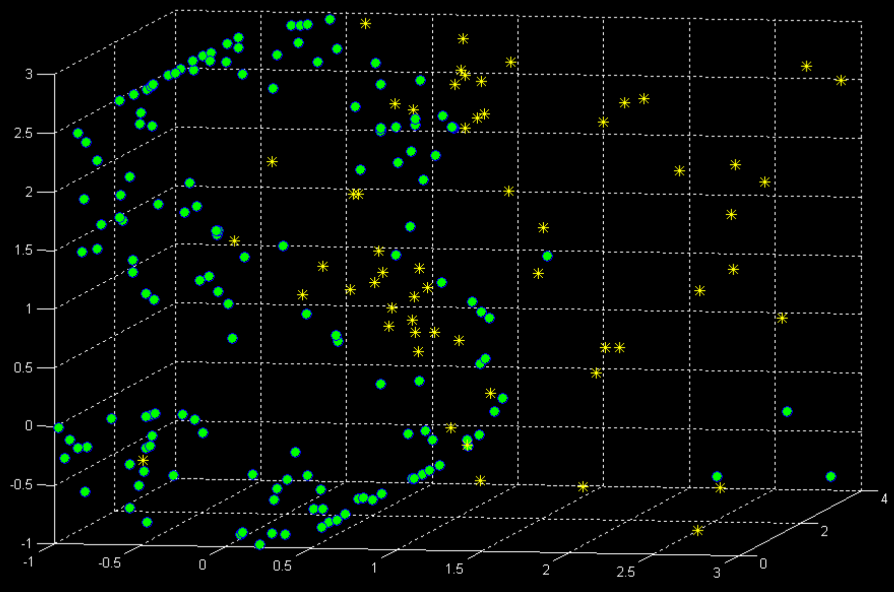


Solution given by algorithm

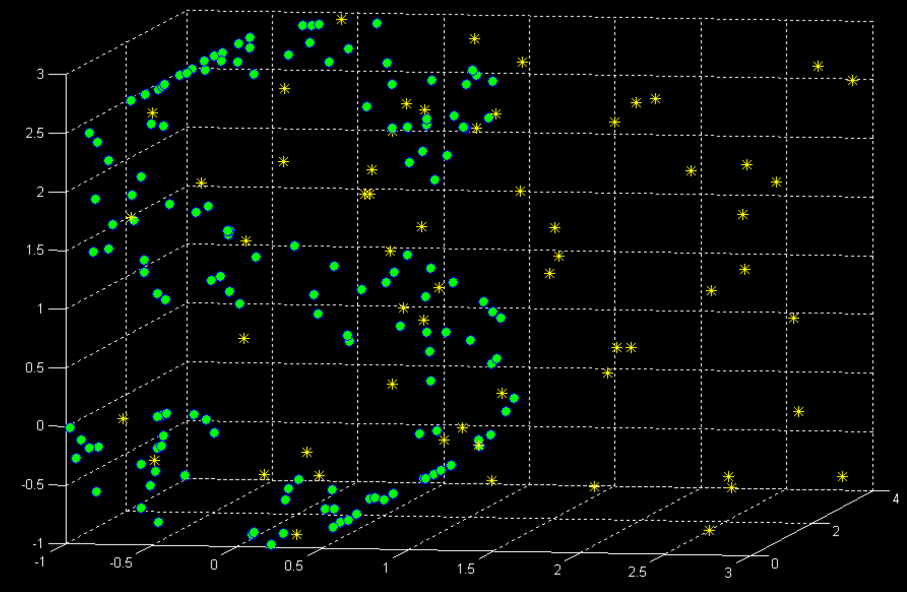


Ground truth

Experiments (Manifold Discovery)



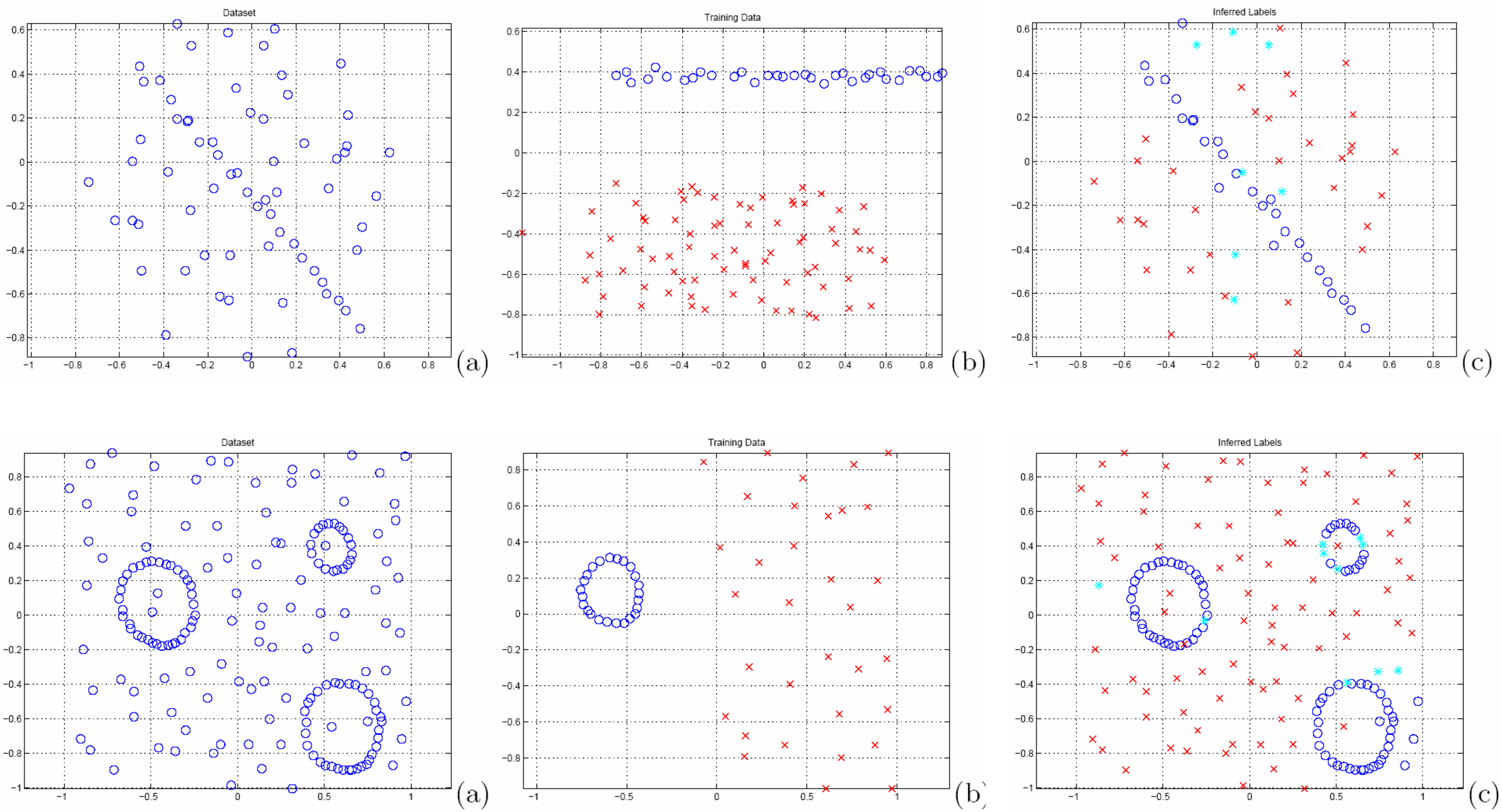
Solution given by algorithm



Ground truth

Experiments (Learning Spatial Patterns)

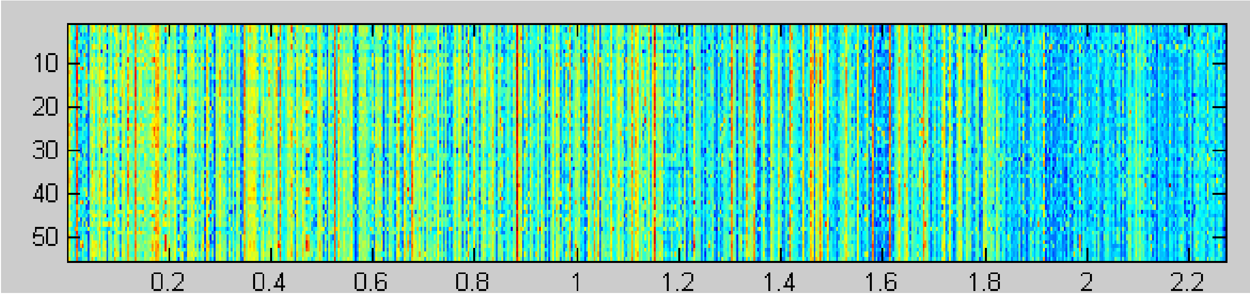
Input | Training Set | Result



Experiments (Functional Gene Classification)

- Functional categories (GO-BP) [Ashburner et. al. 2000]
  - E.g.:
    - cell homeostasis [GO:0019725] Total genes:111
    - anti-apoptosis [GO:0006916] Total genes:112
    - secretory pathway [GO:0045045] Total genes:112
    - hemopoiesis [GO:0030097] Total genes:113
    - humoral defense mechanism (sensu Vertebrata) [GO:0016064] Total genes:114
    - translational initiation [GO:0006413] Total genes:119
    - amino acid biosynthesis [GO:0008652] Total genes:124
    - muscle development [GO:0007517] Total genes:126
- Mouse gene expression data*

Experiments



Genes

*[Hughes Lab,
Banting and Best Institute
University of Toronto]

Experiments (Functional Gene Classification)

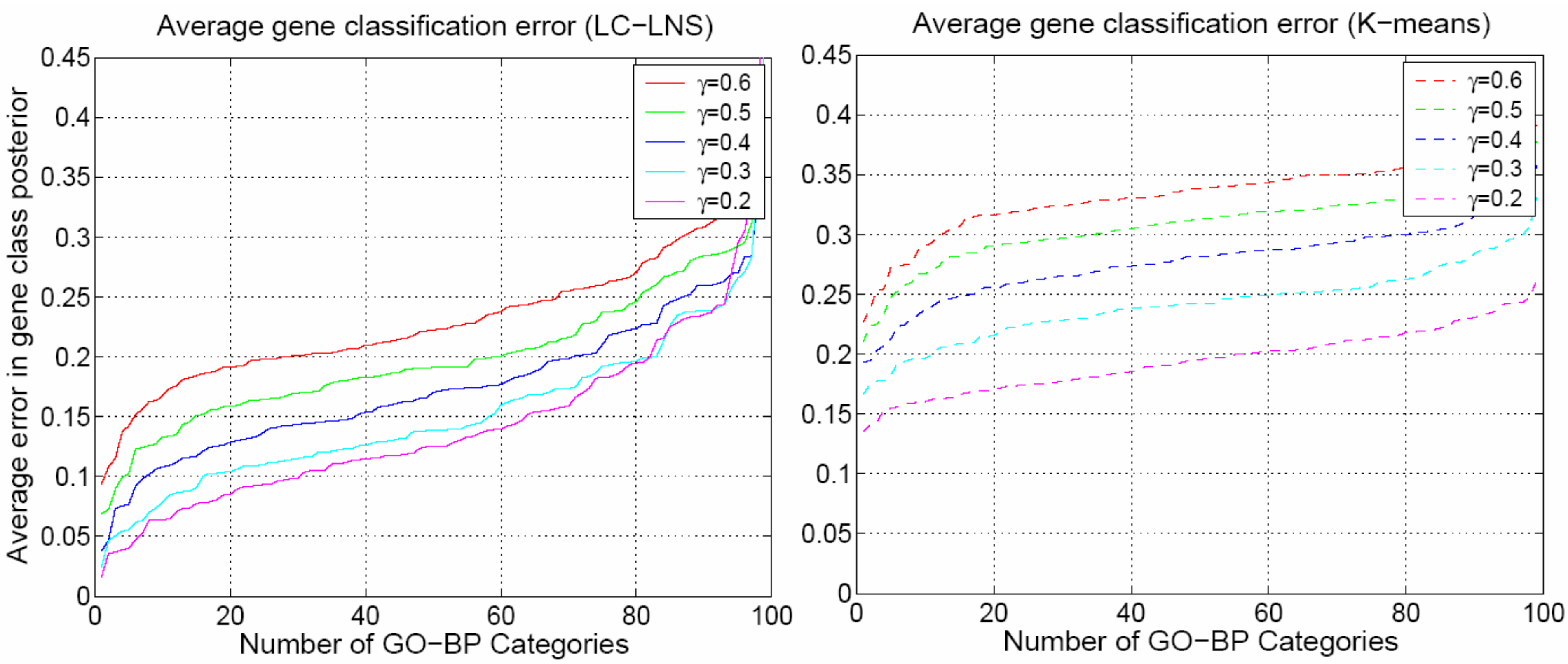
Underlying assumptions

- It might be possible to predict gene function based on the pattern of gene expression in which they are involved
- This pattern might be shared by same function genes
- Thus, different classes could be distinguished by their collective pattern of gene expression

Experiments (Functional Gene Classification)

- Experimental set-up
  - Considered the 99 GO-BP categories with over 80 labeled genes
  - Partition data: train 80% - test 20%
  - Absolute error curves based on γ = proportion of genes that **should** be classified

Experiments (Functional Gene Classification)



Summary

- Clustering/classification based on alternative concept
  - Higher order properties of local structure of the data are more relevant for certain tasks
  - Class dependent cluster structure
- Probabilistic formulation yielded well defined concepts regarding
  - Learning to cluster
  - Inferring clusters
  - Extension to unsupervised clustering
- Concept can be related to more standard clustering ideas
- Negative aspect: Inference algorithm does not in general converge to *good* solutions (the correct posteriors)
- Demonstrated on several applications
  - Learning and finding coherent spatial patterns
  - Separating low dimensional (sampled) manifolds from higher dimensional noise
  - Predicting gene function via collective pattern of expression