A Scale Depedent Data Clustering Model by Direct Maximization of Homogeneity and Separation

Andy M. Yip

Department of Mathematics, UCLA

mhyip@math.ucla.edu

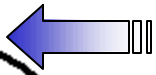
Joint work with Tony F. Chan (UCLA) and Tarek P. Mathew

Mathematical Challenges in Scientific Data Mining

IPAM, January 14-18, 2002

Supported by NSF and ONR

This work was initiated during the Functional Genomics program at IPAM in 2001



- Introduction
- General Framework of our model
- Application of our model to similarity data measured by Pearson correlation coefficients
- Conclusion

Clustering Problems

- To partition a data set according to some *a priori* knowledge or assumptions about the desired clustering
- *A priori* knowledge and assumptions are usually problem dependent

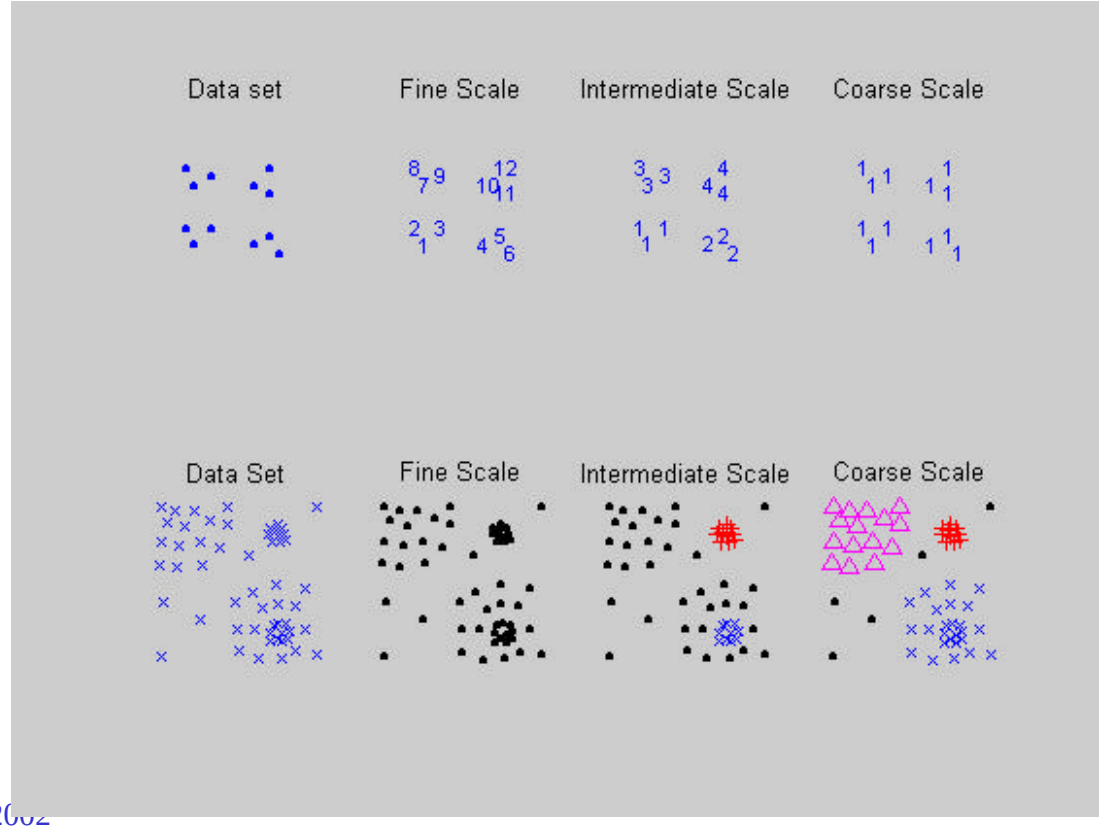
k-means Algorithm

1. Select *k* arbitrary points as initial centriods
2. Repeat until converge
   - 2.1 Assign each point to the closest cluster (closeness is measured by the distance between the point and the centriod)
   - 2.2 Update the centriods

- *A priori* knowledge of the number of clusters *k* is required

Our Objectives

- To identify clusterings at different scales without any knowledge of the number of clusters

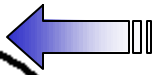


Clustering Algorithms

- Main Ingredients
  - Data Models/ Cluster Models (e.g. *k*-means assumes clusters should be as compact as possible)
  - Implementation of the Data Model (e.g. *k*-means minimizes the sum of all distances between each data point and the centroid of the cluster to which the data belongs)
  - Algorithm to form a clustering (e.g. *k*-means iteratively refines the centroids and the clusters)

Scale Dependent Model (SDM) for Data Clustering

- Main Ingredients
  - Data Models/ Cluster Models (At a fixed scale, we want homogeneous clusters which are well-separated from each other)
  - Implementation of the Data Model (Directly maximize a combination of homogeneity, separation and a scale parameter)
  - Algorithm to form a clustering (Greedy-based heuristic algorithms)

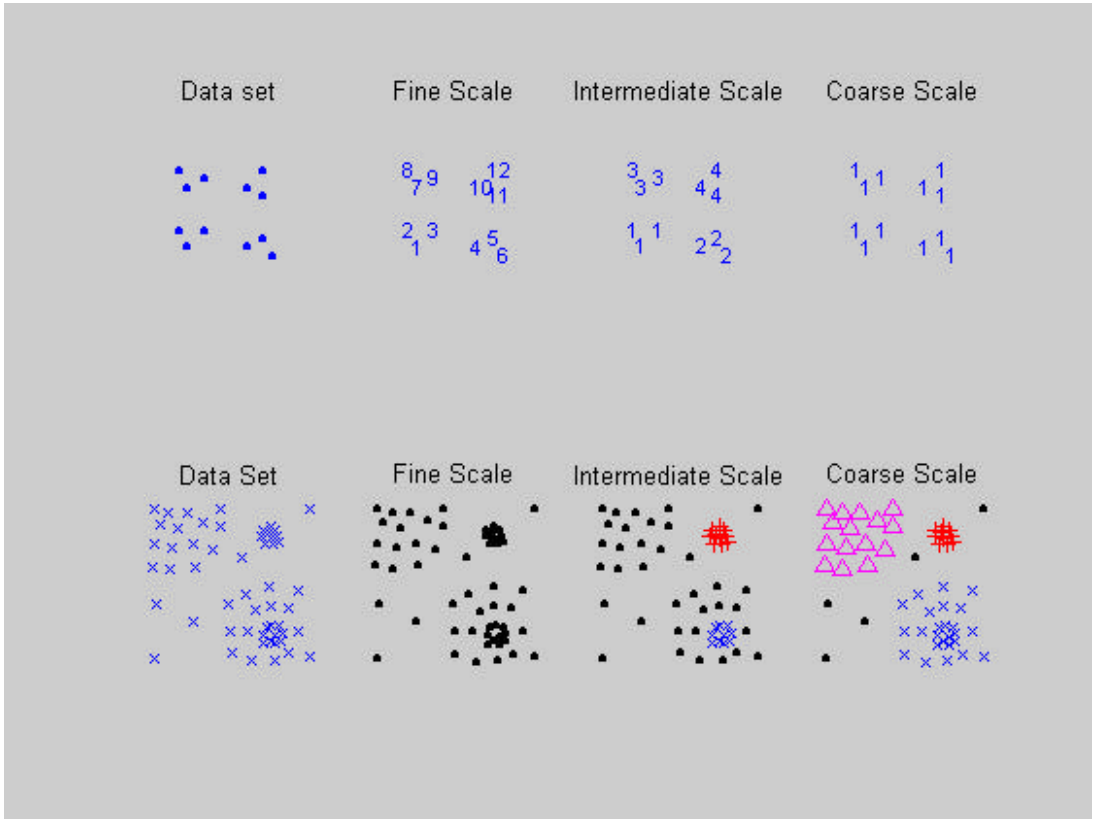


- Introduction
- General Framework of our model
- Application of our model to similarity data measured by Pearson correlation coefficients
- Conclusion

General Framework

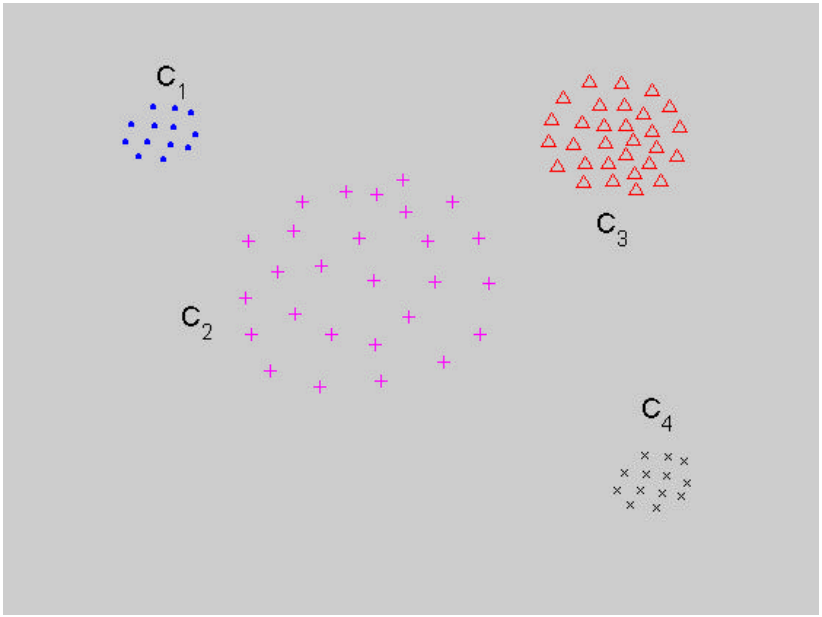
1. Scale
2. Homogeneity and Separation
3. Mathematical Formulation
4. Heuristic Algorithms

Clusterings at Various Scales (1)



Homogeneity and Separation (1)

- Intuition of *homogeneity* of a *single cluster* (*H*) and *separation* between a *pair of clusters* (*S*)



Example of *H* and *S*

H(C_i) = – variance of C_i

S(C_i,C_j) = distance between the means of C_i and C_j

$$H(C_1) \cong H(C_4) > H(C_3) > H(C_2) \qquad S(C_2, C_4) > S(C_1, C_2)$$

Homogeneity and Separation (2)

Let $C=\{C_1,C_2,\ldots,C_d\}$ be a clustering.

- We can define homogeneity for a *clustering*
- $H(C)=f(H(C_1), H(C_2), \ldots, H(C_d))$, i.e., a function of homogeneity of all clusters
- In particular, we may choose f to be a weighted average

$$H(C) = \frac{1}{|C|}\sum_{i=1}^{d}|C_i|\,H(C_1)$$

Homogeneity and Separation (3)

Let $C=\{C_1,C_2,\ldots,C_d\}$ be a clustering.

- We can also define separation for a *clustering*
- $S(C)=g(S(C_1, C_2), S(C_1, C_3), \ldots, S(C_{d-1}, C_d))$, i.e., a function of all pair-wise separation
- In particular, we may choose g to be a weighted average

$$S(C) = \frac{1}{\sum_{i<j} |C_i \| C_j|} \sum_{i<j} |C_i \| C_j| S(C_i, C_j)$$

Homogeneity and Separation (4)

Let $C=\{C_1,C_2,\ldots,C_d\}$ be a clustering.

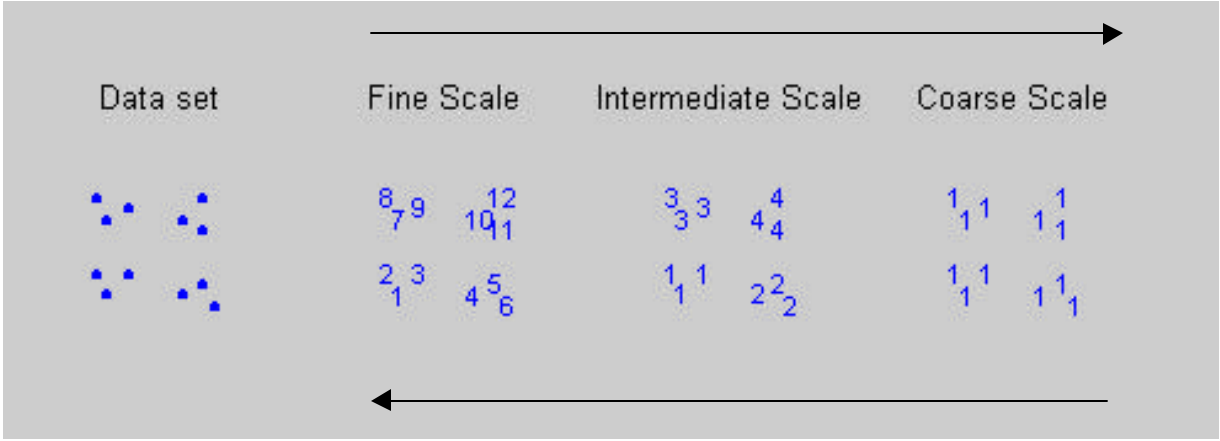
Properties of *H(C)* and *S(C)*

- *H*(*C*) is maximized when each cluster consists of a singleton only
- *S*(*C*) is maximized when every point collapses to a single cluster
- Usually, there is a trade-off between maximizing *H*(*C*) and maximizing *S*(*C*)

Homogeneity and Separation (5)

- Trade-off between maximizing homogeneity and maximizing separation

Increasing Separation

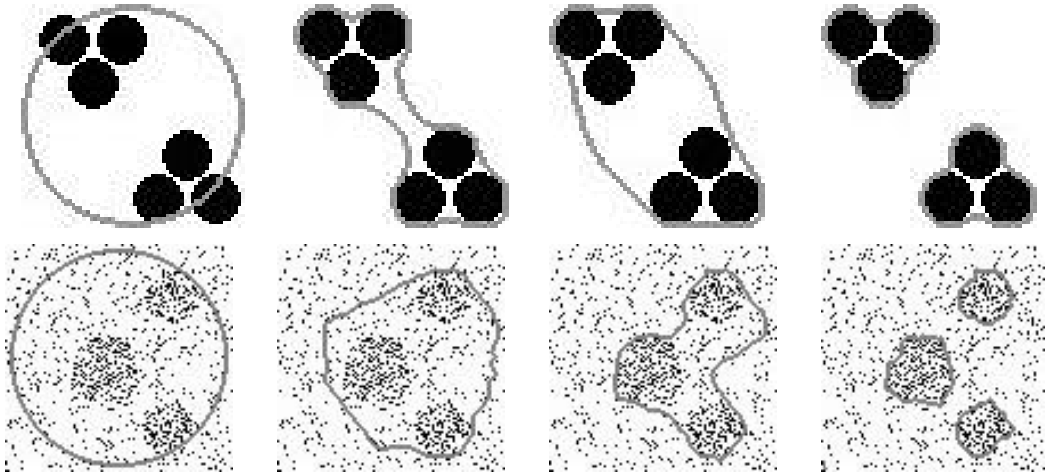


Increasing Homogeneity

Motivation of our Mathematical Formulation

- A closely related problem: Image Segmentation

Active Contour without Edges (Chan and Vese, 2001)



Model: $\min_C \|u_0 - u(C)\|_2 + m \text{ Length}(C)$

where μ is a scale parameter

Mathematical Formulation (1)

Homogeneous and Separated Clusters

+

Scale

=

Maximize $F_m(C) = H(C) + mS(C)$

Mathematical Formulation (2)

Maximize $F_m(C) = H(C) + mS(C)$

When $m \to 0$, $H(C)$ is essentially maximized

$\Rightarrow$ Each cluster contains a singleton only

When $m \to \infty$, $S(C)$ is essentially maximized

$\Rightarrow$ Every points collapses to a single cluster

When m is something between

$\Rightarrow$ We get a clustering at an intermediate scale

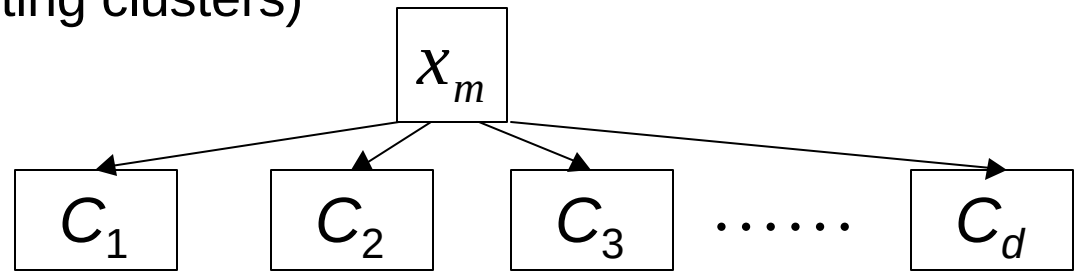
Heuristic Algorithms (1)

A two-stage greedy based algorithm:

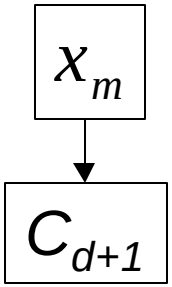
Stage 1: Assign the points sequentially to an existing cluster or a new cluster

- Suppose $x_1,\ldots,x_m$ are assigned to some clusters $C_1,\ldots,C_d$

Case 1: (Existing clusters)

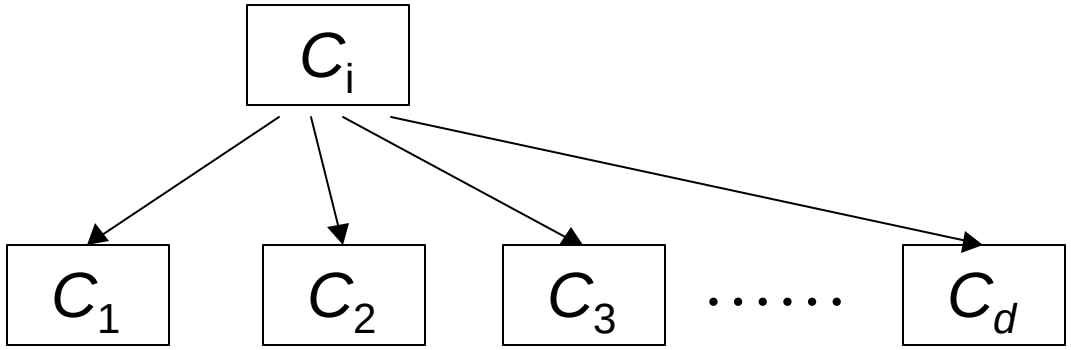


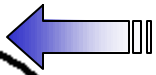
Case 2: (A new cluster)



Heuristic Algorithms (2)

Stage 2: Merge existing clusters





- Introduction
- General Framework of our model
- Application of our model to similarity data measured by Pearson correlation coefficients
- Conclusion

Measuring Pair-wise Similarity by the Pearson Correlation Coefficients (1)

- Pearson Correlation Coefficient:

$$r_P(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\| \|x_j\|}$$

- Applications in clustering gene expression data and text documents

Measuring Pair-wise Similarity by the Pearson Correlation Coefficients (2)

- Normalize to $\|x\| = 1$ during preprocessing steps

- Pearson Correlation Coefficient

$$r_P(x_i, x_j) = x_i^T x_j$$

we have

$$-1 \le r_P(x_i, x_j) \le 1$$

Average Homogeneity and Average Separation (1)

Denote the usual inner product in R^n by

$$r(x_i, x_j) = x_i^T x_j$$

Mean:

$$F(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Average Homogeneity and Average Separation (2)

Homogeneity of a single cluster:

$$H(C_i) = \frac{1}{|C_i|}\sum_{x \in C_i} r\left(x, \frac{F(C_i)}{\|F(C_i)\|}\right) = \|F(C_i)\|$$

Average Homogeneity:

$$H_{AVE}(C) = \frac{1}{\sum_{i=1}^{d}|C_i|}\sum_{i=1}^{d}|C_i|H(C_i) = \frac{1}{|C|}\sum_{i=1}^{d}|C_i|\|F(C_i)\|$$

We have

$$0 \le H_{AVE}(C) \le 1$$

Average Homogeneity and Average Separation (3)

Separation of a pair of clusters:

$$S(C_i, C_j) = \frac{1}{2}\left[1 - r\left(\frac{F(C_i)}{\|F(C_i)\|}, \frac{F(C_j)}{\|F(C_j)\|}\right)\right]$$

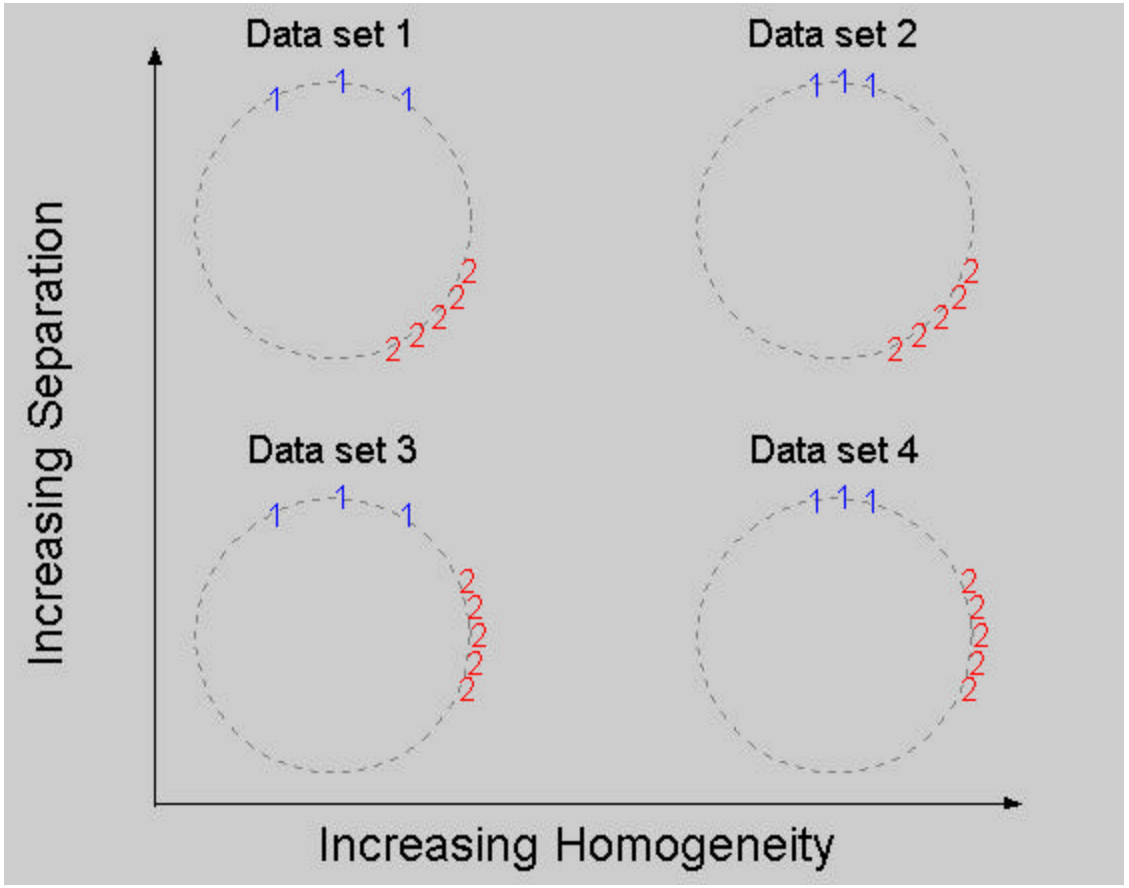
Average Separation

$$S_{AVE}(C) = \frac{1}{\sum_{i \neq j} |C_i \| C_j|} \sum_{i=1}^{d} \sum_{i \neq j} |C_i \| C_j| S(C_i, C_j)$$

We have

$$0 \leq S_{AVE}(C) \leq 1$$

Average Homogeneity and Average Separation (4)



Fast Evaluation and Updating of the Objective Function (1)

- In case we use $H_{AVE}(C)$ and $S_{AVE}(C)$ to measure the quality of a clustering, we derived fast method to evaluate and update the objective function.
- In the first stage of our heuristic algorithms, we need to compute the objective function value of merging a point to an existing cluster.

Fast Evaluation and Updating of the Objective Function (2)

- In the second stage, we need to compute the objective function value of merging clusters.

- Total complexity of our heuristic algorithm is

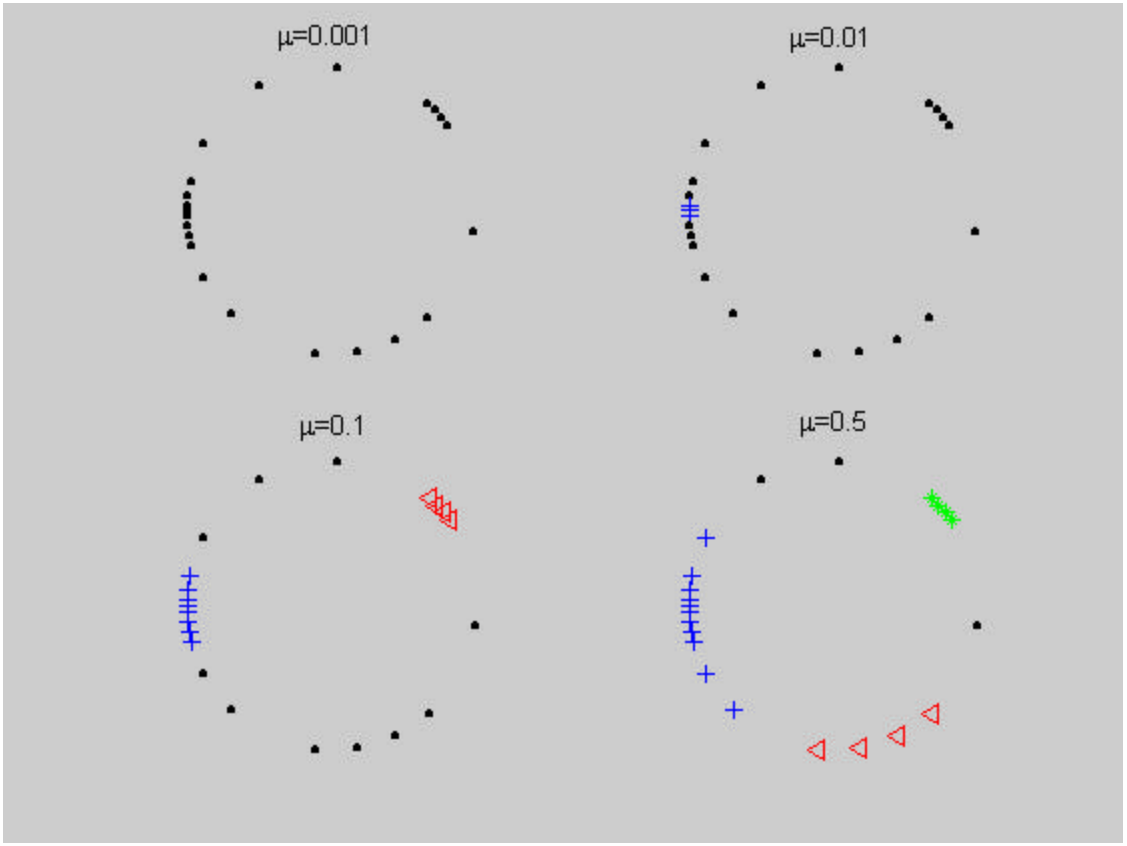
$$O(npk^2)$$

where k=number of clusters detected

n=number of data points

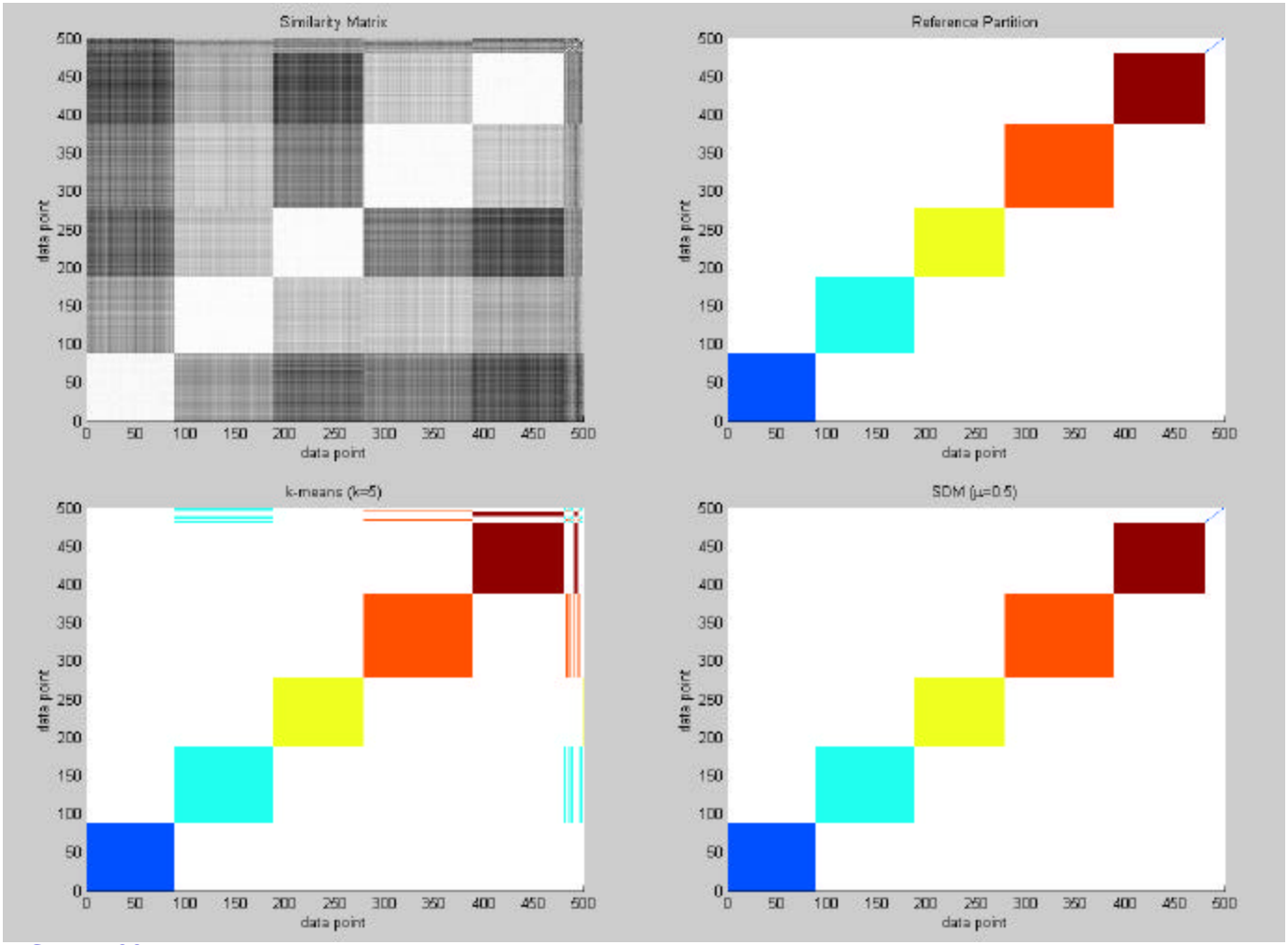
p=dimension of the data

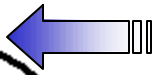
Example 1: Clustering at Various Scales



- For a fixed μ, clusters of the same scale are identified

Example 2: Comparison with *k*-means





- Introduction
- General Framework of our model
- Application of our model to similarity data measured by Pearson correlation coefficients
- Conclusion

Conclusions

- A Scale Dependent Model (SDM) is developed which can capture clusterings at various scales through a choice of the scale parameter without any *a priori* knowledge about the number of clusters.
- A combination of Homogeneity and Separation are directly maximized.
- Fast Evaluation and Updating methods are derived for the case of using average homogeneity and average separation.