

## Lecture 15 - Imaging Data Sets

June 24, 2009

### Introduction

- ▶ We have produced signatures which give indications of geometric features, as a tool for “geometric pattern recognition”
- ▶ Humans are very good at pattern recognition when dimensions are  $\leq 3$
- ▶ Suggests we should develop methods for exploiting this capability
- ▶ This talk will present various existing methods for visualization and “conceptualization”
- ▶ Next talk will present a topological method, and discuss its features, along with sample applications

### Clustering

- ▶ The simplest way to visualize a space is by seeing its component structure
- ▶ Means that any clustering method can be regarded as a visualization method
- ▶ For “real” topology, there are two notions of components, connected and path-connected components. They usually agree for interesting spaces
- ▶ In the statistical version, there is a great multitude of methods

### Clustering - Single Linkage

- ▶ Recall that single linkage clustering for a finite metric space and scale parameter  $\epsilon$  was defined as  $\pi_0(VR(X, \epsilon))$ , where  $\pi_0(Z)$  is the set of path components of  $Z$
- ▶ The set of clusters are arranged as the nodes in a tree, or dendrogram
- ▶ Can be interpreted as the result of an iterative procedure

## Clustering - Single Linkage

- ▶ For any partition of a finite metric space, define the *linkage* of two blocks (i.e. equivalence classes)  $B_1$  and  $B_2$  to be
$$\lambda(B_1, B_2) = \min\{d(b, b') \mid b \in B_1 \text{ and } b' \in B_2\}$$
- ▶ Given any partition  $\Pi$  of  $X$ , let  $\mu(\Pi)$  be the minimum value of  $\lambda(B_1, B_2)$  which occurs for any pairs of blocks  $B_1$  and  $B_2$  of  $\Pi$
- ▶ For any partition  $\Pi$  of  $X$ , define a new partition  $\xi(\Pi)$  of  $X$  as the partition obtained from  $\Pi$  by merging any pair of blocks  $B_1$  and  $B_2$  for which  $\lambda(B_1, B_2) = \mu(\Pi)$
- ▶ Iterate, to obtain a sequence of partitions
- ▶ If we start with the discrete partition, it is clear that we produce the same nested sequence of partitions which occurs in single linkage hierarchical clustering, with height function given by the values  $\mu(\xi(\Pi))$ .

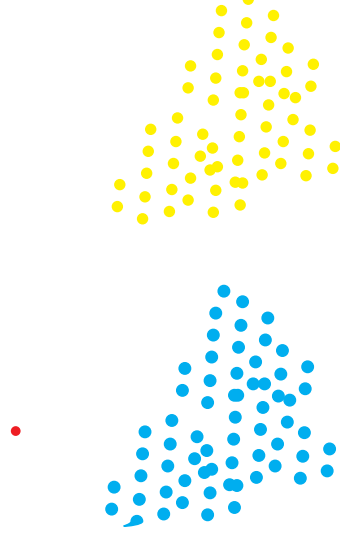
## Clustering - Average and Complete Linkage

- ▶ One could replace  $\lambda$  by another choice of linkage function, and obtain another sequence of partitions
- ▶ One choice is  $\lambda^{av}$ , defined by letting  $\lambda^{av}(B_1, B_2)$  be the average distance between any pairs of elements, one from  $B_1$  and the other from  $B_2$
- ▶ Another choice is  $\lambda^{comp}$ , defined by letting  $\lambda^{comp}(B_1, B_2)$  be the maximum distance between any pairs of elements from  $X$ , with the first in  $B_1$  and the other in  $B_2$
- ▶ Produces two new hierarchical clustering algorithms, called *average linkage clustering* and the other called *complete linkage clustering*

## Clustering - Average and Complete Linkage

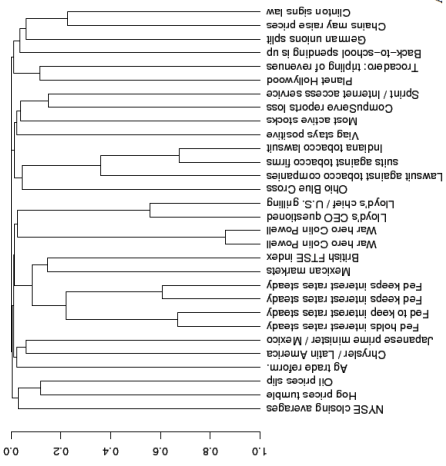
- ▶ Practitioners often prefer average and complete linkage to single linkage
- ▶ They feel that they obtain cleaner or more compact clusters this way
- ▶ Single linkage has “chaining” problems

## Clustering - Average and Complete Linkage



In this case, average and complete linkage will merge the red point into the blue cluster rather than merging the blue and yellow

## Clustering - Average and Complete Linkage



Complete Linkage

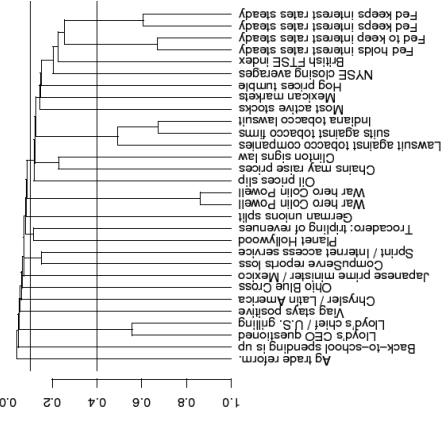
## Clustering - k-Means

- ▶ Method for clustering Euclidean data
- ▶ Requires a choice of  $k$ , the number of clusters to be constructed
- ▶ Operates by optimizing and objective function. Given a set  $S \subseteq \mathbb{R}^n$ , and a partition  $S = S_1 \cup \dots \cup S_k$ , the objective function is given by

$$\arg \min_{\Pi} \sum_{i=1}^k \sum_{x_j \in S_j} \|x_j - \mu_j\|$$

where  $\mu_j$  is the mean of the cluster  $S_j$ , and  $\Pi$  denotes the set of all partitions of  $S$  into  $k$  blocks.

## Clustering - Average and Complete Linkage



Single Linkage

## Clustering - Clique Method

- ▶ Method for graph clustering
- ▶ Two points  $p$  and  $q$  connected if there is a sequence of  $n$ -cliques (complete graphs), first of which contains  $p$  and the second of which contains  $q$
- ▶ Can be adapted to metric space context by replacing cliques by a metric space on  $n$  points in which all pairwise distances are a fixed  $\epsilon$
- ▶ Used in social network analysis

## Clustering - Density Cluster Trees (W. Stuetzle)

- ▶ Fix a threshold  $\epsilon$
- ▶ Form clustering by considering only the points whose density is greater than a threshold  $\delta$ , where density is measured by some estimator
- ▶ Set of points increases as  $\delta$  decreases, obtain a tree
- ▶ Produces dendrograms just as in the hierarchical clustering case
- ▶ Reflects the density as well as the geometry

## Clustering - Theory

- ▶ Many different methods lead to lack of clarity in the theory of clustering
- ▶ One should try to develop a theory which provides more understanding of different methods
- ▶ Initially, J. Kleinberg proved a negative result

## Clustering - Kleinberg's Theorem

- ▶ Proves non-existence of clustering scheme satisfying certain natural hypotheses
- ▶ Clustering scheme: method for assigning a partition to any finite metric space
- ▶ We say the scheme is *scale invariant* if multiplying the metric by a fixed factor doesn't alter the partition
- ▶ It is *rich* if any partition of a set  $X$  can be obtained as the result of the scheme applied to some metric on  $X$
- ▶ Given a metric space  $(X, d)$  and a partition  $\Xi$ , a *K-modification* of the metric on  $X$  is a new metric  $d'$  on  $X$  which has the property that distances within the clusters of  $\Xi$  are  $\leq$  the original distances and intra-cluster distances computed by  $d'$  are  $\geq$  the distances computed in  $d$

## Clustering - Kleinberg's Theorem

**Theorem (Kleinberg):** There is no clustering scheme satisfying scale invariance and richness, together with the consistency hypothesis that the clustering associated to any  $K$ -modification of a metric  $d$  on  $X$  yields the same partition on  $X$  as does  $d$ .

Reminiscent of Arrow's impossibility theorem.

## Clustering - Theory

- ▶ Joint work with F. Memoli
- ▶ Can one modify the hypotheses so as to obtain an existence and uniqueness result instead?
- ▶ We will adopt the notion that functoriality is desirable, and should be part of the hypotheses
- ▶ Will require introduction of persistence

## Clustering - Theory

- ▶ Make the category of finite metric spaces into a category  $\underline{\mathcal{FM}}$
- ▶ The morphisms are *distance non-decreasing maps*, i.e.  $f : X \rightarrow Y$  so that  $d_Y(f(x), f(y)) \leq d_X(x, y)$  for all  $x, y \in X$
- ▶ Other choices for the morphisms are possible, for example maps which are injections on points

## Clustering - Theory

- ▶ Persistence sets form a category  $\underline{\mathcal{P}} - \text{sets}$ , just as persistence vector spaces do
- ▶ A *functorial clustering scheme* is a functor

$$\xi : \underline{\mathcal{FM}} \rightarrow \underline{\mathcal{P}} - \text{sets}$$

together with a surjective *natural transformation*  
 $X \rightarrow \xi(X)$

## Clustering - Theory

- ▶  $X$  can be regarded as a constant persistence set
- ▶ Naturality means the diagram

$$\begin{array}{ccc} X & \longrightarrow & \xi(X) \\ \downarrow & & \downarrow \\ Y & \longrightarrow & \xi(Y) \end{array}$$

commutes

## Clustering - Theory

**Theorem (Memoli-C.):** There is a unique functorial clustering scheme  $\xi$  for which the following conditions hold. It is standard hierarchical clustering.

1. For any finite metric space  $X$ , we let  $\text{sep}(X)$  denote the minimum non-zero distance in  $X$ . We assume that for any metric space with  $\text{sep}(X) > \delta$ , we have that  $\xi(X)_t$  gives the discrete clustering on  $X$  for all  $t < \delta$
2.  $\xi$  has the scale invariance property that the result of multiplying the metric on  $X$  by a fixed constant  $c$  results in the rescaled persistence set  $\xi(X)_{ct}$ .
3. If  $E$  denotes the two point metric space  $\{p, q\}$ , with  $d(p, q) = 1$ , then  $\xi(E)$  is the persistent set with  $\xi(E)_t = E$  for  $t < 1$ , and  $\xi(E)$  is a one point set for  $t \geq 1$ .

## Clustering - Theory

- ▶ We say that a clustering scheme is *excisive* if for every finite metric space  $X$ , each of the clusters in  $\xi(X)$  is a  $\xi$ -connected metric space
- ▶ One can prove that every excisive scheme which is functorial for the injective maps of metric spaces can be written as  $\xi_\Omega$  for some set

## Clustering - Theory

- ▶ Restricting the functoriality to only injective morphisms of metric spaces produces many more interesting possibilities
- ▶ For any clustering scheme  $\xi$  with injective functoriality, we say a finite metric space  $X$  is  $\xi$ -connected if  $\xi(X)$  is the one point set
- ▶ Any family  $\Omega$  of finite metric spaces produces a clustering scheme, by declaring two points  $p, q$  in a finite metric space  $X$  to be in the same component if there are elements  $\{\omega_i\}_{i=0}^N \subseteq \Omega$  and distance non-increasing inclusions  $\omega_i \xrightarrow{\rho_i} X$ , so that  $p$  and  $q$  are in the images of  $\rho_0$  and  $\rho_N$  respectively, and so that the images of the  $\rho_i$  and  $\rho_{i+1}$  overlap. Denote this scheme by  $\xi_\Omega$ .

## Clustering - Discussion

- ▶ What is the reason for the great variety of clustering schemes?
- ▶ Perhaps everyone is looking for a way to incorporate density into the decisions about how to cluster
- ▶ Practitioners want “compact” clusters
- ▶ This is stated directly for density cluster trees and clique clustering
- ▶ Why not use two dimensional persistent clustering, using scale parameter and density?

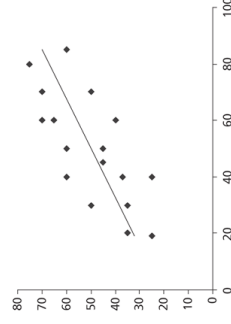
## Clustering - Density Estimation

- ▶ Wide variety of methods
- ▶ For linear or Euclidean data, histograms or higher dimensional variants a good option
- ▶ Naive method simply counts number of data points within a ball of fixed radius about a given point, and then normalizes. Applicable to general metric spaces
- ▶ Kernel estimator also popular - depends on choice of a kernel real valued function of a real variable
- ▶  $K(r)$  should decrease with  $r$

## Clustering - Density Estimation

$$\rho^{ker}(x) = \sum_{i=1}^M K(d(x, x_i))$$

## Linear Regression



- ▶ Attempt to fit data to a real line or hyperplane, expressing a given coordinate in terms of the others using linear equations
- ▶ Least squares difference between fitted points and actual points used

## Projection Pursuit

- ▶ Starting point - high dimensional Euclidean data
- ▶ Look for low dimensional linear projections which are “informative”
- ▶ Random projections in a single direction are typically near Gaussian
- ▶ Strategy: search for directions in which the projected data is far from Gaussian

## Projection Pursuit

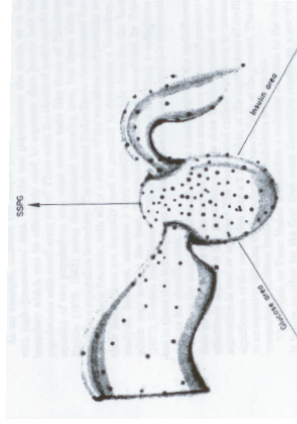
### How do we quantify how far from Gaussian a distribution on the real line is?

- ▶ Associated to any distribution on the real line are moments
- ▶ The mean  $\mu$  is  $E[X]$ , where  $X$  is the standard variable on the real line
- ▶ The variance is given as  $E[(X - \mu)^2]$ .
- ▶ More generally, the  $k$ -th central moment is  $E[(X - \mu)^k]$

## Projection Pursuit

- ▶ The  $k$ -th central moment of a normal distribution with parameters  $\mu$  and  $\sigma$  is  $= 0$  for  $k$  odd, and equals  $\frac{(2s)!}{2^s s!} \sigma^k$  for  $k = 2s$ .
- ▶ One uses the third and fourth central moments as measures of difference between a given distribution and the normal distribution
- ▶ The third central moment measures *skewness*
- ▶ *Kurtosis* is defined to be the difference  $\frac{\mu_4}{\sigma^4} - 3$ , where  $\mu_4$  is the 4th central moment. This measure vanishes on any normal distribution. It measures peakedness of the distribution
- ▶ One chooses coordinates which maximizes either of these in absolute value, or some combination of them, in order to obtain coordinates containing maximal amount of “information”
- ▶ Other objectives, such as measures of multimodality, are also used

## Projection Pursuit



Miller-Reaven Diabetes Study

## Principal Component Analysis

- ▶ Method of “dimensionality reduction” for Euclidean data
- ▶ Key ingredient: singular value decomposition
- ▶ Every  $m \times n$  matrix  $A$  can be written as

$$A = UDV$$

where  $U$  and  $V$  are unitary, and  $D$  is generalized diagonal, i.e. of the form

$$\begin{bmatrix} d_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

with  $d_1 > d_2 > d_3$

## Principal Component Analysis

- ▶ The vector corresponding to  $d_1$  is regarded as the most significant coordinate,  $d_2$  the next most significant, etc.
- ▶ Euclidean data can be arranged in a matrix, and one can apply PCA to such a data matrix
- ▶ Projecting to the  $k$  most significant eigenvalues provides a map from the original data to a lower dimensional data set which in favorable situations does not distort distances too much

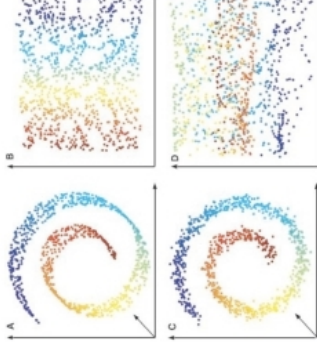
## Multidimensional Scaling

- ▶ Applies to any distance matrix - data need not be Euclidean
- ▶ When data is Euclidean, the method is essentially PCA
- ▶ In other situations, produces a map to Euclidean space, with an ordered set of coordinates
- ▶ First two or three coordinates frequently produce useful low dimensional embeddings of data set, with small distortion
- ▶ Works best when one is dealing with an intrinsically flat Riemannian manifold

## Multidimensional Scaling - ISOMAP

- ▶ Particular version using graph length distances instead of Euclidean distances is called ISOMAP
- ▶ Some very nice examples from data in the form of images

## Multidimensional Scaling - ISOMAP



# Multidimensional Scaling - ISOMAP

