Clustering Algorithms and their applications

S. Lakshmivarahan
School of Computer Science
University of Oklahoma
Email: varahan@ou.edu

Acknowledgement- Organizers

- Thanks to the workshop organizers especially to **Sue Ellen Haupt** and **Philippe Tissot** for this opportunity to participate in this workshop

- **Philippe Tissot** and his **staff at TAMUCC** deserves a special appreciation for all their hard work in putting together a wonderful program

Acknowledgements - collaborators

- Ahmad Alhamed (CS PhD) – Muti-model ensemble - SAMEX
- Mike Baldwin (Met PhD) – Rainfall classification
- Nusrat Yussouf (CS PhD) – New England Ensemble Project
- Prashant Kakani & Bharat Thanneru (CS MS) – Tornado classification using Linear Discriminant analysis
- Chris Calvert (CS MS) – Regime switching models for El Nino data
- Naren Papu (CS MS) – Time series analysis – volatility modeling
- Kyle Abott (CS MS) – Modeling electricity prices

- Mike Richman, Ted Trafalis, V. Lakshmanan – faculty colleagues at OU &NSSL

Note: **SAMEX** – Storm and Mesoscale Ensemble Experiment

References – Clustering based approach

- Ahmad Alhamed, S. Lakshmivarahan and David Stensrud (2002) "Cluster Analysis of Multimodel Ensemble Data from SAMEX", **Monthly Weather Review**, Vol 130, 226-256

- Nusrat Yussouf, David Stensrud,and S. Lakshmivarahan (2004) "Cluster Analysis of Multimodel Ensemble Data over New England", **Monthly Weather Review**, Vol 132, 2452 – 2462

- Mike Baldwin, S. Lakshmivarahan and John S. Kain (2005) " Development of an Automated Classification Procedure for Rainfall systems", **Monthly Weather Review**, Vol 133, 844-862

References – Linear Discrinimant Function based approach

- Prashant Kakani, S. Lakshmivarahan and M. B. Richman (2004) “A Tornado Detection Algorithm Using Empirical Orthogonal Functions**”, Proceedings of the ANNIE International Conference**, pp125-130

- Bharat Thanneru, S. Lakshmivarahan and M. B. Richman (2005) “ A Classification of Tornadoes using Linear Discriminant Functions”, **Proceedings of the ANNIE International Conference,** pp141-150

References – Time Series modeling in Finance

- S. Lakshmivarahan and Duane Stock (2006) “Explaining Municipal Bond Volatility: Implications for Practice”, **Municipal Finance Journal**, Vol 27, 1-33
- Naren Pappu, S. Lakshmivarahan and Duane Stock (2006) “Models of Conditional Variance for Bond Prices”, **Journal of Fixed Income**, Vol 16, 65-70

Data Sources

- **Science/Engineering**: Data/observations arise from measurements using instruments – Radar, Satellites, rain gauges, sensors
- **Medicine**: Blood tests, X-ray, MRI/Cat-scan
- **Business**: Sales records from credit card receipts
- **Psychology**: Opinion surveys
- **Consumer Protection**: Extensive testing of gadgets
- **Economics**: Unemployment data, Monthly Import/export
- **Finance**: Foreign exchange rate
- **Homeland Security**: Wire tapping, profiling

Organization of data – **time series**

Inherent Temporal variation

- **daily rain fall** in Corpus Christi
- **monthly unemployment** in the state of Texas
- **exchange rate** - US dollar/Japanese yen
- **hourly price** of a barrel of crude oil/natural gas/electricity – **spot/futures** prices

Note: Each item in the time series is a measurement of the same physical entity – all have same units

Organization of data - m*n **Data matrix**, X (m > n or m < n)

- **n Objects** (columns) – a severe weather event, a rain event, human with a specific ailment
- **m Attributes/Features** (rows)– spatial distribution of the rainfall, various features of a tornadic event extracted from the radar data, set of symptoms of a disease

Note: Selection of features is quite fundamental. The key lies in the answer to the question: **what to measure**?

An example

- Create an ensemble of forecasts using n=25 ensemble members (objects) created using 4 different models – CAPS, NCEP (Eta), NCEP (RSM), NSSL (MM5 modified) and different initial conditions
- Analyze how the output of one class of models compare with those of the others using clustering techniques to distinguish the behavior of the models

A set of field variables derived from the output of the dynamical models

Attributes are the spatial distribution of each of the field variable

Field (ten variables)	Abbreviation	Unit
Precipitation at SFC	PRCP	kg/m^2 *
Pressure at SFC	PRMSL	Pa
Height at 500 mb	HGT	gpm
Temperature at SFC	TMP	K
Wind at 250 mb	WIND	m/s
Absolute vorticity at 250 mb	ABS V	1/s
Absolute vorticity at 500 mb	ABS V	1/s
Absolute vorticity at 850 mb	ABS V	1/s
Pressure vertical velocity at 700 mb	V VEL	Pa/s
Convective available potential energy	CAPE	J/kg

Number of attributes/features

- Assume M*N*L computational grid for the model – which is a set of PDEs
- At a given levels there are m = M*N grid points at each of which we have an output – temperature, pressure, etc computed by the model
- In our example, M = 117, N= 81, L=20. In this case m = 9477
- The output of each of the models are available at multiples of 3 hours- 0, 3, 6,…24,…36.
- Thus, for each of the **ten field variable** we have 13 data matrices each of size 9477 * 25, a total of 130 data matrices on hand

Data matrix – n objects and m attributes

$$X = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ & & \ddots & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \end{array}$$

Cross sections of the data matrix

- Each object is denoted by the column. The Jth column x_{*J} denotes the Jth object
- Each row, X_{I*} denotes the distribution of a variable across the object
- An element x_{IJ} denotes the value of the Ith variable on the Jth object

Some more details of the data matrix

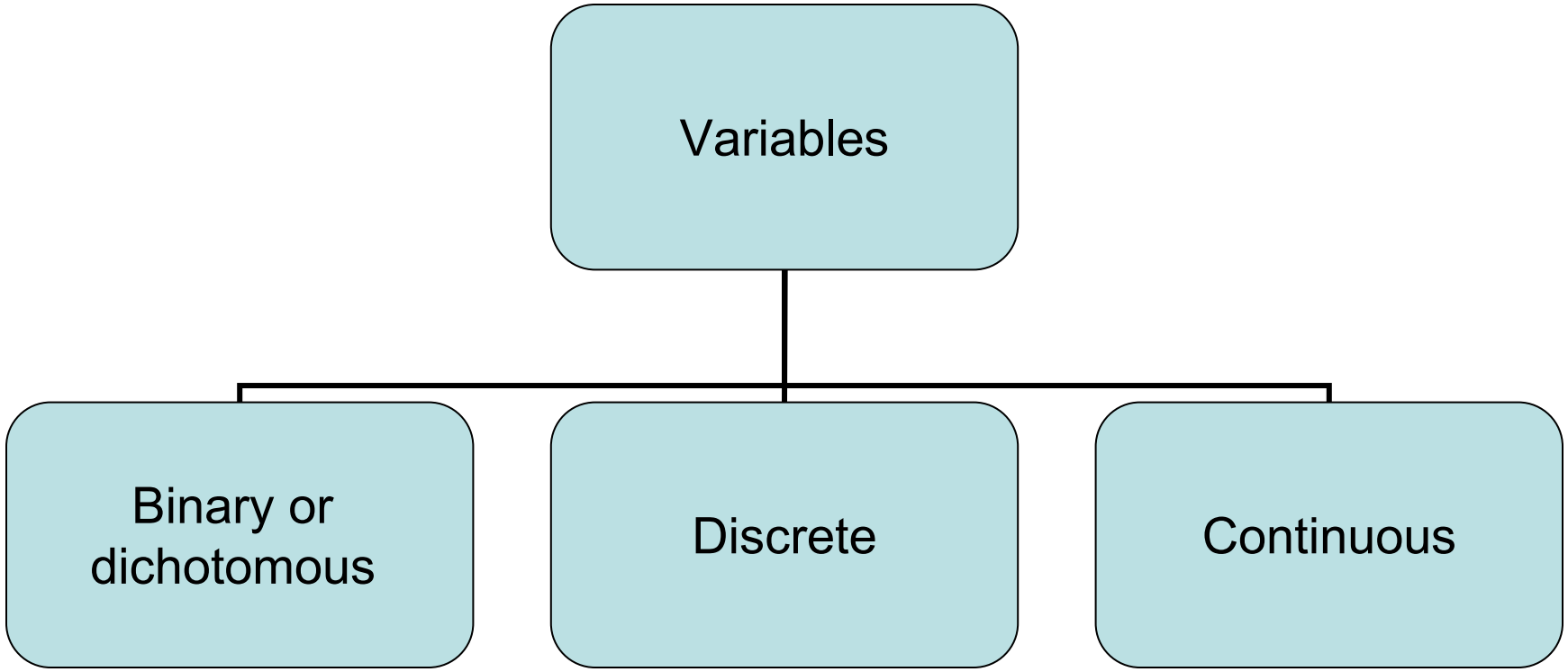
- The ensemble members (objects/columns) are numbered 1 through 25
- Objects 1-5 are from the CAPS model; 6,8,10,12 and14 are from NCEP (Eta) model; 7,9,11,13 and 15 are from NCEP (RSM) model and 16-25 are from NSSL MM5 (Version 2)
- Attributes/features, m = 9477
- We have one data matrix for each of the ten field variables listed above at every 3 hours from 0 to 36, a total of 13 matrices
- There are 10 field variables giving a total of 130 matrices

A geometric view of the data set

- You can visualize an **object as a point in** the m dimensional **Euclidean Space,** R^m **– multi-variate representation**
- We have a set of n points in an m – dimensional Euclidean space
- One of the questions is: how this set of n points align themselves in R^m ?
- We can tap into rich resources from the **geometry of the finite dimensional vector spaces** using matrix theory, multivariate calculus, etc to do the data analysis.

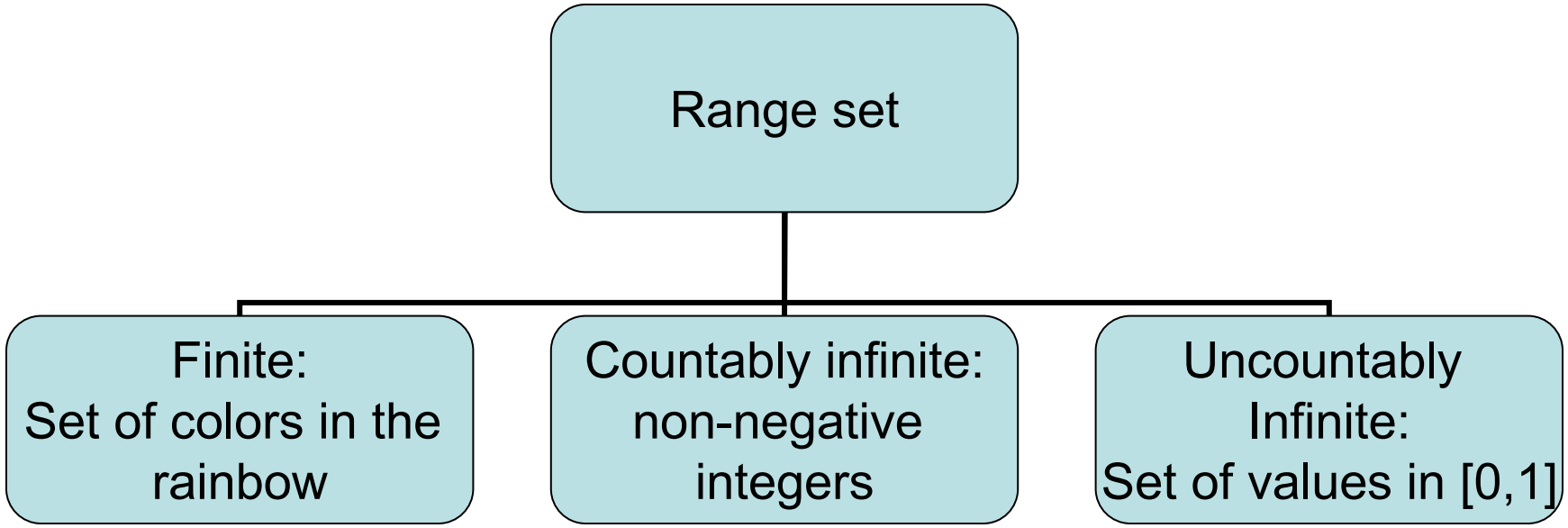
Variables and scales

- A classification of variables



Variables and Scales

- A classification of variables based on the size of the range set



A hierarchy of scales

- **Nominal scale**: We simply distinguish the classes. Variables taking binary values, colors of a rainbow. We can only test for equality – if $x_1 = x_2$ or not
- **Ordinal scale**: In addition to distinguishing the objects, it allows for an ordering of objects. $x = y$ or not and $x <$ or $> y$. In an opinion survey, the rating is on a scale of 1 to 10, Grades in a class A, B, C, D, F

A hierarchy of scales

- **Interval scale**: Assigns a meaningful measure of the difference. In addition to saying x > y, we can also perceive the difference (x – y). Temperature x = 65 in Corpus Christi is 30 degrees warmer than the temperature y = 35 in Oklahoma City
- **Ratio scale**: An interval scale with a fixed zero point or origin. In addition to (x - y), we can also say x is (x / y) time superior. The salary of a CEO is 200 times that of a janitor or temperature in absolute scale

A classification of variables

- Variables with nominal / ordinal scales are called **categorical / qualitative** variables
- Variables with interval / ratio scales are known as **quantitative** variables
- Conversion of scales may be necessary. Convert numerical grade from continuous (ratio scale) to discrete (ordinal scale) letter grade
- Temperature measured in **Centigrade** scale is of the interval type but when measured in **Absolute scale** becomes ratio type

Summary: know thy data

- It is clear from this discussion that the type of operations – test for **equality, compare/order**, compute the **difference** or the **ratio** critically depends on the scales
- Consequently, algorithms designed for one type of data many not be useful for a different type

Data Transformations

- In the m*n data matrix X, each row represents the values of a given variable on different objects. Hence, n elements of a given row are of the same units but elements of different rows may be of different types and units
- If all the elements of the matrices are of similar type and order of magnitude, then no transformation is needed.
- Otherwise, we need to transform the data type and range so that they have a comparable range
- For example, if one measurements gives values in **single digits** but another in **thousands**, this would call for data transformation

A typical data transformation

- **Normalize each column so that it has mean zero and unit variance.**
- This is done by subtracting from the elements of each row the mean of that row and then dividing the difference by the standard deviation of that row
- With this transformation, each row of data will share the common property of having **mean zero and with unit** variance
- There are other transformations – log, square-root, etc

Data reduction- Principal component Analysis (PCA)

- If the given set of attributes exhibit **correlation**, then there is scope for reducing the number of attributes to a smaller uncorrelated set.
- This is done by (linearly) transforming the given data set to obtain a new set of attributes that are **orthogonal** (uncorrelated)
- The transformation consists in changing the **standard basis** to a **new basis** formed by the (mutually orthogonal) **eigen vectors of the covariance matrix C** = $(1/n)(xx^T)$ where it is assumed that X is the normalized data matrix
- Express $X = FA^T$ where the columns of the matrix A are the scaled eigen vectors of C and F denotes the (orthogonal) uncorrelated representation
- The elements of A are called the **loading** and those of F are called **scores**

Data Reduction

- The eigen values of C denote the variance and for a given data matrix X, the total variance (sum of the eigen values) is fixed
- We can order the eign values from the largest to the smallest
- By considering only the first r (<m) largest eigen values we can reduce the number of attributes from m to r
- In this case we obtain a new reduced representation as given below using only the first r eigen values

$$X = \overleftarrow{F}\overleftarrow{A}^T$$

Cluster Analysis (CA)- References

- M. R. Anderberg (1973) **Cluster Analysis for Applications**, Academic Press, NY
- J. A. Hartigan (1975) **Clustering Algorithms**, John Wiley & Sons, NY
- H. C. Romesberg (1984) **Cluster Analysis for Researchers,** Life Time Learning Publications, Belmont, CA
- A.K.Jain and R. C. Dubes (1988) **Algorithms for Clustering Data**, Prentice Hall, Englewood Cliff, NJ
- R. O. Duda, P. E. Hart and D. G. Stork (2001) **Pattern Classification**, Wiley, NY (Second Edition)
- D. S. Wilks (2006) **Statistical Methods in Atmospheric Sciences**, Academic Press, NY

Software for CA

- **MATLAB** as part of the Statistical Tool box has several clustering algorithms
- **IMSL** has a very nice package for clustering
- We have developed a comprehensive set of clustering package written in **FORTRAN 95**

Applications of clustering

- Psychology
- Anthropology
- Biology, Medicine
- Pattern recognition
- Machine Intelligence
- Geosciences

Cluster Analysis (CA)

- CA is the **simplest** of the multivariate methods
- CA is a **descriptive** method for gauging similarities
- CA **does not test** but it helps to **generate hypotheses**
- CA methods are based on **intelligent heuristics**
- Since the samples may not be drawn at random, **extension** of the conclusions must be done with great care **using analogy**
- There are **no formal methods for evaluating the risk** in extrapolation – **false alarm vs. miss**

A classification of CA

- **Agglomerative** Algorithm
- To start with **each object is in a cluster**
- Progressively **combine** these clusters **until a single cluster** is obtained
- Leads to **hierarchical** Methods

- **Divisive** algorithm
- **Opposite** of agglomerative algorithm
- Starts with a **single cluster** and **progressively divide** it into smaller clusters

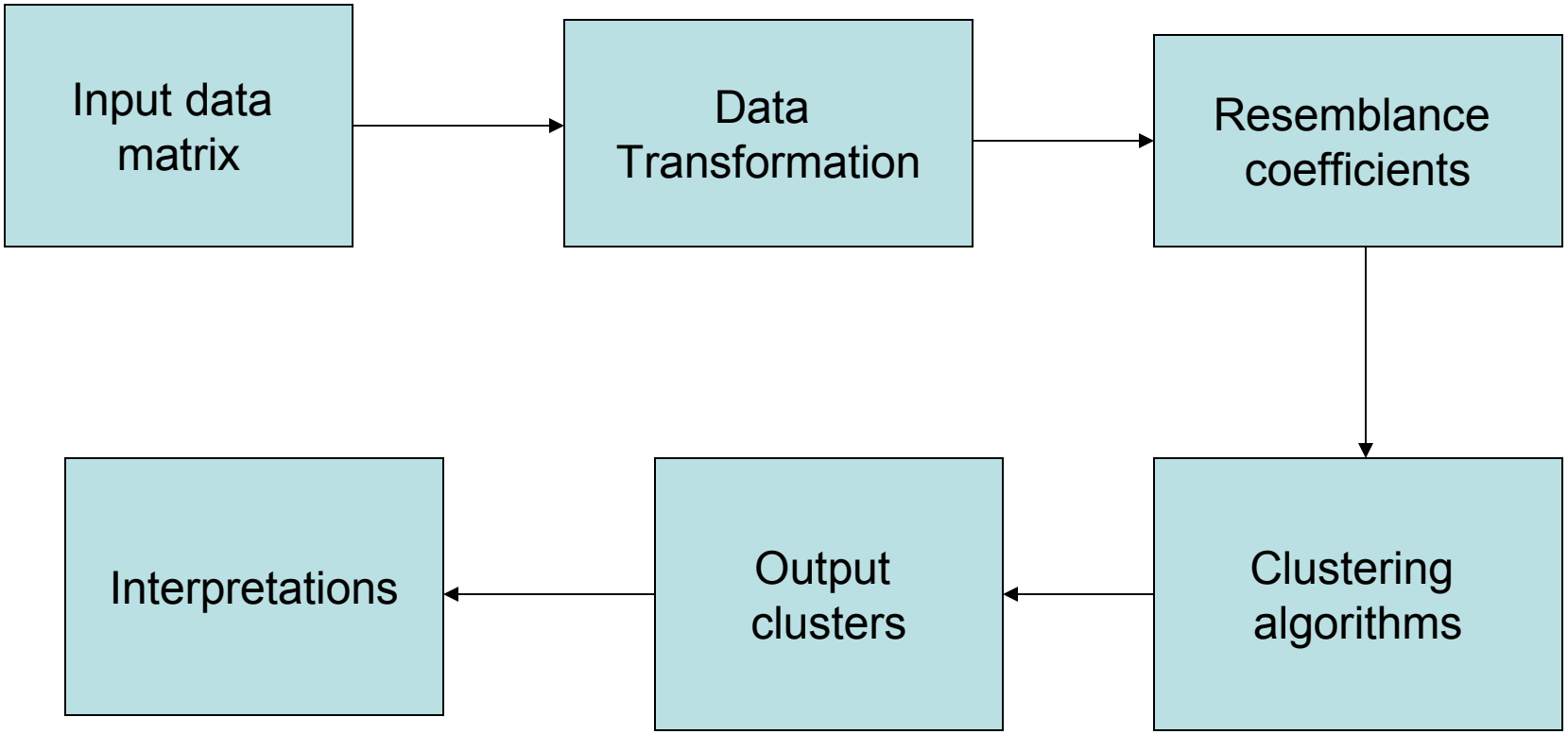
A classification of CA

- **Serial algorithm** that handles one object at a time
- **Monothetic** algorithm that uses one attribute at a time
- **Parallel algorithm** handling multiple objects at a time
- **Polythetic** algorithms capable of handling several attributes at a time

Clustering methodology

- **Preprocessing** phase
- Extract the **resemblance coefficients** using intelligent heuristics
- **Classification** phase
- **Interpretation/validation**

A view of clustering methodology



Preprocessing phase

- Collect data – arrange in the **matrix form**
- Know the **scales** of various measurements
- Decide if data **transformation** is needed
- Decide if data **reduction** is needed
- Call the resulting processed data matrix, X
- In the following we will be working with such a **processed data set**

Resemblance Coefficients: Similarity measure

- **Correlation** – larger the correlation between objects larger is the similarity
- **Cosine of the angle between the vectors**: Recall that objects are points in the feature space R^m. Each object represents a vector from the origin to the point corresponding to it. Smaller the angle between vectors, larger is the similarity. This is given the **inner product** of the **unit vectors** in the direction representing the objects

Resemblance coefficients - Dissimilarity measures

- **Euclidean distance** between two objects: $d(i, j) = (X_{*I} – X_{*j})^T (X_{*i}- X_{*j})$ – square of the distance between the ith and jth objects. Larger the distance, smaller is the similarity between the objects
- **Mahalanabis distance**:

  $MD(i,j) = (X_{*I} – X_{*j})^T \; S^{-1} (X_{*I} – X_{*j})$ where S is called the **pooled sample covariance** matrix

Clustering based on Scatter analysis

- Consider a set of n objects divided into two clusters consisting of n_1 and $n_2 = n - n_1$ objects

$$X = \begin{pmatrix} X_{11} & \cdots & X_{1n_1} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn_1} \end{pmatrix} \qquad Y = \begin{pmatrix} Y_{11} & \cdots & Y_{1n_2} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \cdots & Y_{mn_2} \end{pmatrix}$$

$$\bar{X}_{i*} = \frac{1}{n_1} \sum_{j=1}^{n_1} X_{ij} \qquad \bar{Y}_{i*} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{ij}$$

Mean of each cluster

- Mean of each cluster

$$M_1 = \begin{bmatrix} \bar{X}_{1*} \\ \bar{X}_{2*} \\ . \\ . \\ . \\ \bar{X}_{m*} \end{bmatrix} = \frac{1}{n_1} \sum_{j=1}^{n_1} X_{*j} \qquad M_2 = \begin{bmatrix} \bar{Y}_{1*} \\ \bar{Y}_{2*} \\ . \\ . \\ . \\ \bar{Y}_{m*} \end{bmatrix} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{*j}$$

Within cluster scatter

- Compute the scatter within each clusters S_1 and S_2 and the total with in cluster scatter S_w

Scatter Matrix for Group 1:

$$S_1 = \sum_{j=1}^{n_1} (X_{*j} - M_1)\,(X_{*j} - M_1)^T$$

Scatter Matrix for Group 2:

$$S_2 = \sum_{j=1}^{n_2} (Y_{*j} - M_2)\,(Y_{*j} - M_2)^T$$

$$S_w = S_1 + S_2$$

Pooled mean vector

- Pooled mean vector M

$$M = \frac{1}{n_1 + n_2}[n_1 M_1 + n_2 M_2]$$

Total scatter of all samples

- Center each data set with respect to its row mean and the overall scatter matrix S are given by

$$\hat{X}_{*j} = X_{*j} - M \qquad \hat{Y}_{*j} = Y_{*j} - M$$

$$S = \sum_{j=1}^{n_1} \hat{X}_{*j} \hat{X}_{*j}^T + \sum_{j=1}^{n_2} \hat{Y}_{*j} \hat{Y}_{*j}^T$$

Scatter between the clusters

- Between cluster scatter S_B is given by

$$S_B = \sum_{j=1}^{n_1} (M_1 - M)(M_1 - M)^T + \sum_{j=1}^{n_2} (M_2 - M)(M_2 - M)^T$$

A decomposition

- The total scatter of all the samples S is the sum of the within cluster scatter S_W and between cluster scatter S_B. That is,
- $$S = S_W + S_B$$
- $$Tr(S) = Tr(S_W) + Tr(S_B)$$
- Trace of S denotes the sum of the diagonal elements which denotes the sum of the variance
- Thus, by reducing the $Tr(S_w)$, we can increase $Tr(S_B)$, a measure of the separation
- The well known **Ward's method** exploits this idea

Hierarchical Method: Algorithmic framework

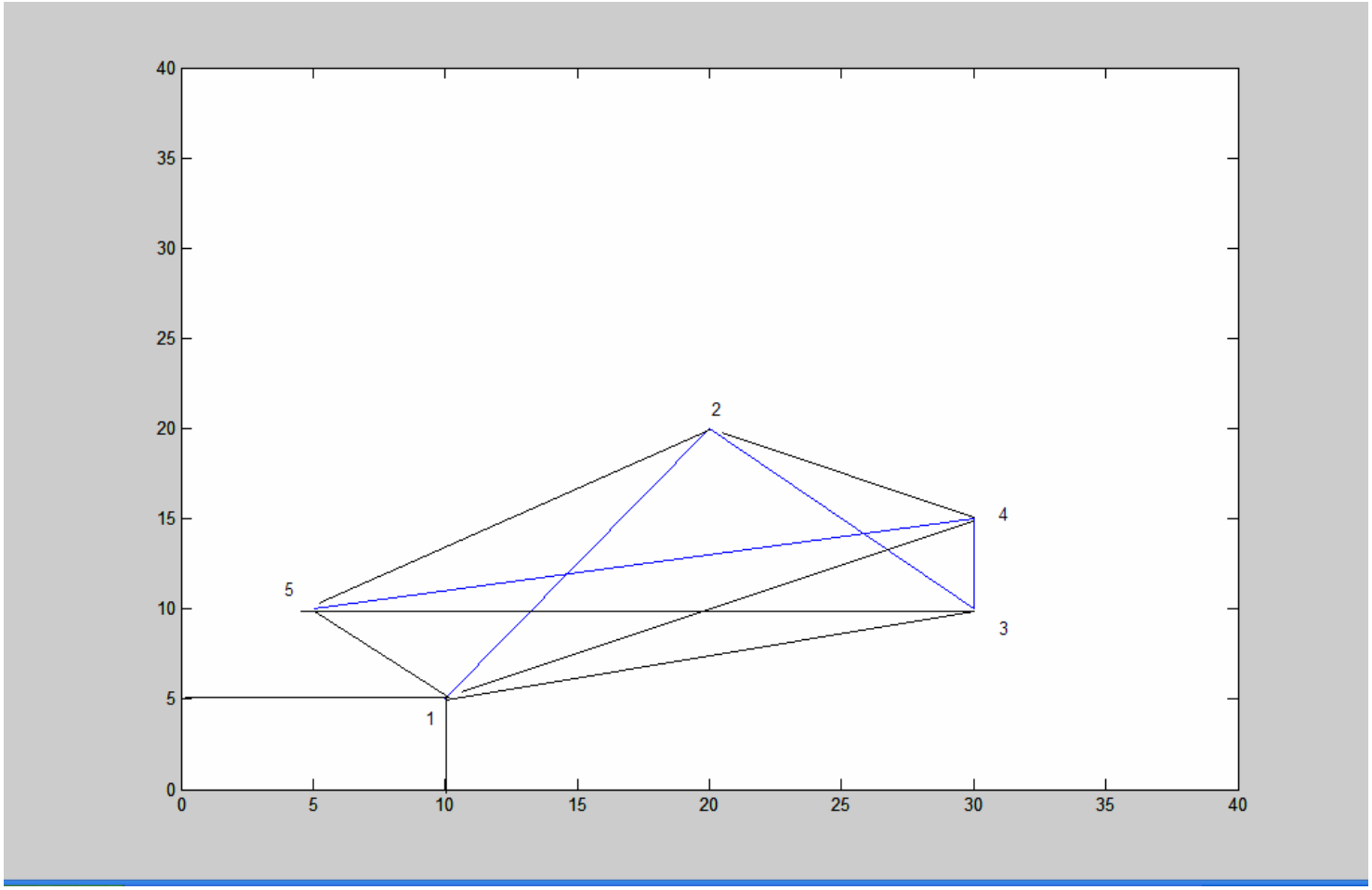
1) Begin with n clusters each containing <u>one object</u>. Let the clusters be named 1 through n.
2) Search the similarity matrix S for the most similar pair of clusters. Let p and q be the two similar clusters, with S_{pq} as their similarity measure p > q.
3) Reduce the number of clusters by 1 by merging the two clusters p and q. Label the new cluster q. Update the entries of S to reflect the revised similarity between the new cluster q and other clusters other than p. Delete the row and column of S that corresponds to the cluster p.
4) Perform steps 2 and 3 a total of (n-1) times. At each stage record the element of each cluster and keep track of all similarity measures at each stage to have a complete record

An example: 5 objects each with 2 attributes

n = 5 and m = 2

	1	2	3	4	5
X1	10	20	30	30	5
X2	5	20	10	15	10

An example: A geometric view of objects in R^2



Euclidean distance matrix of order 5 * 5 (dissimilarity measure) # of clusters 5. Objects 3 & 4 are most similar

	1	2	3	4	5
1	x	x	x	x	x
2	18	x	x	x	x
3	20.6	14.1	x	x	x
4	22.4	11.2	5	x	x
5	7.07	18.0	25.0	25.5	x

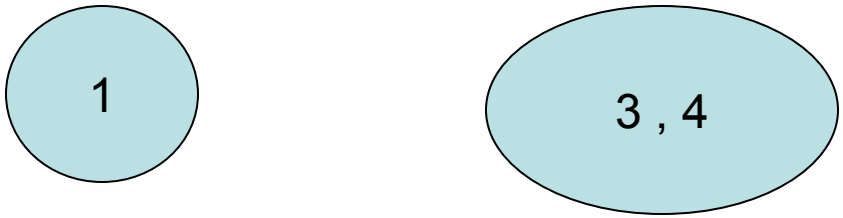
Merging of clusters based on similarity: # of clusters 4

$$e_{(34)1} = \frac{1}{2}[e_{31} + e_{41}] = \frac{1}{2}[20.6 + 22.4] = 21.5$$

	1	2	5	(34)
1	x	x	x	x
2	18	x	x	x
5	7.07	18	x	x
(34)	21.5	12.7	25.3	x

Objects 1 and 5 are most similar

Variations on the theme



- There are at least three ways to compute the distance between the clusters (1) and (3,4)
- Maximum of the distance from 1 to 3 and 4 – C-link
- Minimum of the distance from 1 to 3 and 4- s-link
- Average of the distance 1 to 3 and 4 - UPGMA
- Accordingly we get different algorithms

Merging of clusters: # of clusters 3

	2	(34)	(15)
2	x	x	x
(34)	12.7	x	x
(15)	18.0	23.4	x

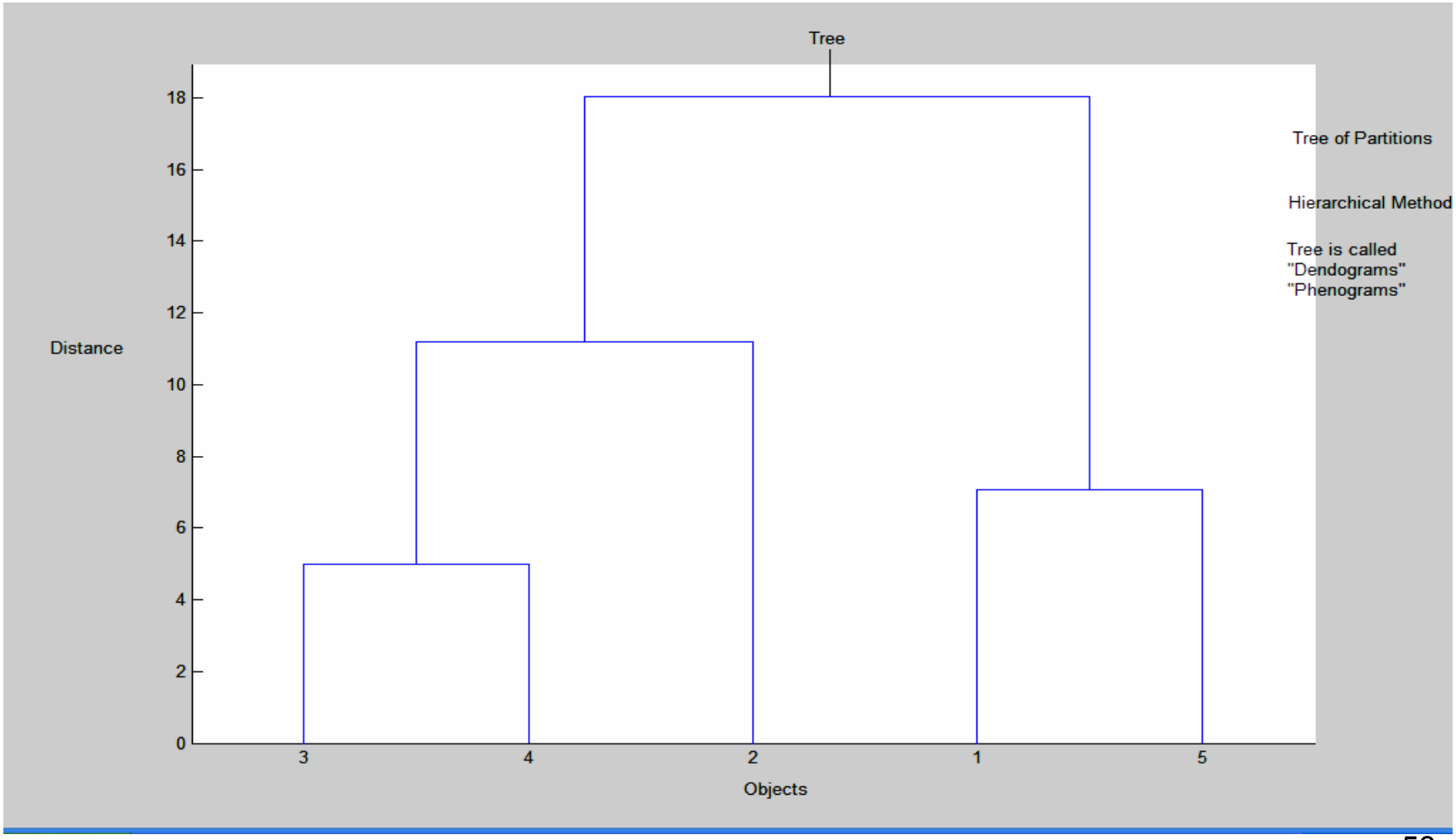
- Now 2 and (34) are very similar

Merging of clusters: # of clusters 2

	(15)	(234)
(15)	x	x
(234)	21.6	x

- Clusters (15) and (234) are similar

Merging of clusters: # of cluster 1
Dendrogram, a pictorial view of the nesting of clusters



Output clusters

- By **cutting** the tree at a suitable similarity level we can the generate a set of clusters
- In the above example, **a cut at 14** will produce three clusters: {3,4,2} and {1,5}
- If we **cut it at 10** instead, we will get three clusters {3,4}, {2} and {1,5}

Application I
Multimodal ensemble- SAMEX data
(Storm and Mesoscale Ensemble Experiment)

- Consider the field variable say, sea level pressure for all the 25 cases (using 4 models)
- At time t=0, the data matrix X is of the order 9477 *25
- Each column describes the spatial variation of the field variable at time t=0.
- Thus the entire matrix describes the collective spatial variation of the field variable due to all the models and for all the chosen initial conditions.
- Apply a clustering algorithm to this data matrix (**Alhamed et. al., 2002)**

Analysis of SAMEX data

- Redo the clustering experiment for the same field variable for t= 3,6, 9, …36
- Our goal is to quantify the temporal evolution of these clusters analyze
- You may think of this evolution as tubes containing different clusters
- It turns out outputs from the same models cluster together and out put of different models occupy different regions

Analysis of SAMEX data

- We repeated this for several field variables – 3hr accumulated rain, convective available potential energy, 500h-Pa geopotential height, 250-hPa wind speed
- The forecasts cluster by the model and this clustering occurs very early in the game
- There are very few intermodel clusters
- This result high lights the importance of the model physics in determining the resulting forecasts
- It also brings out the importance of having ensembles with model physics diversity and initial condition diversity to capture a wide range of possibilities in near term forecasts.
- For more results refer to Nusrat Yussouf et. al., (2004)

Application II
Rainfall classification system: data set

- One hour accumulated rain fall data on a spatial grid of size 128 * 128 covering roughly 500km by 500km – (Baldwin et. al., 2005)
- 48 cases for **training** and 100 cases for **testing**
- The late summer-fall 2000 season was chosen
- Events contained **two** classes– **convective and non-convective cases**
- **Convective case** with two sub classes **cellular**(19) and **linear forms**(18)
- **Non-convective case** with one subclass called the stratiform (11)

Rainfall system: selection of attributes

- **Intensity based attributes**:
- Histograms of typical rain fall systems suggest that the distribution of the rain fall behave more like Gamma variate

Gamma PDF depends on 2 parameters: a,b

f(x;a,b) = (x/b)a-1 [exp(-x/b)] [bGamma(a)]-1

$x \geq 0$, a,b>0

a is the **shape** parameter, b is the **scale** parameter

- Our goal is to first estimate these two parameters a and b using the generalized method of moments (GMM)

Spatial attributes

- **Spatial continuity related attributes**:
- Plot the spatial correlogram of the rain data: variogram, covariance and correlation respectively

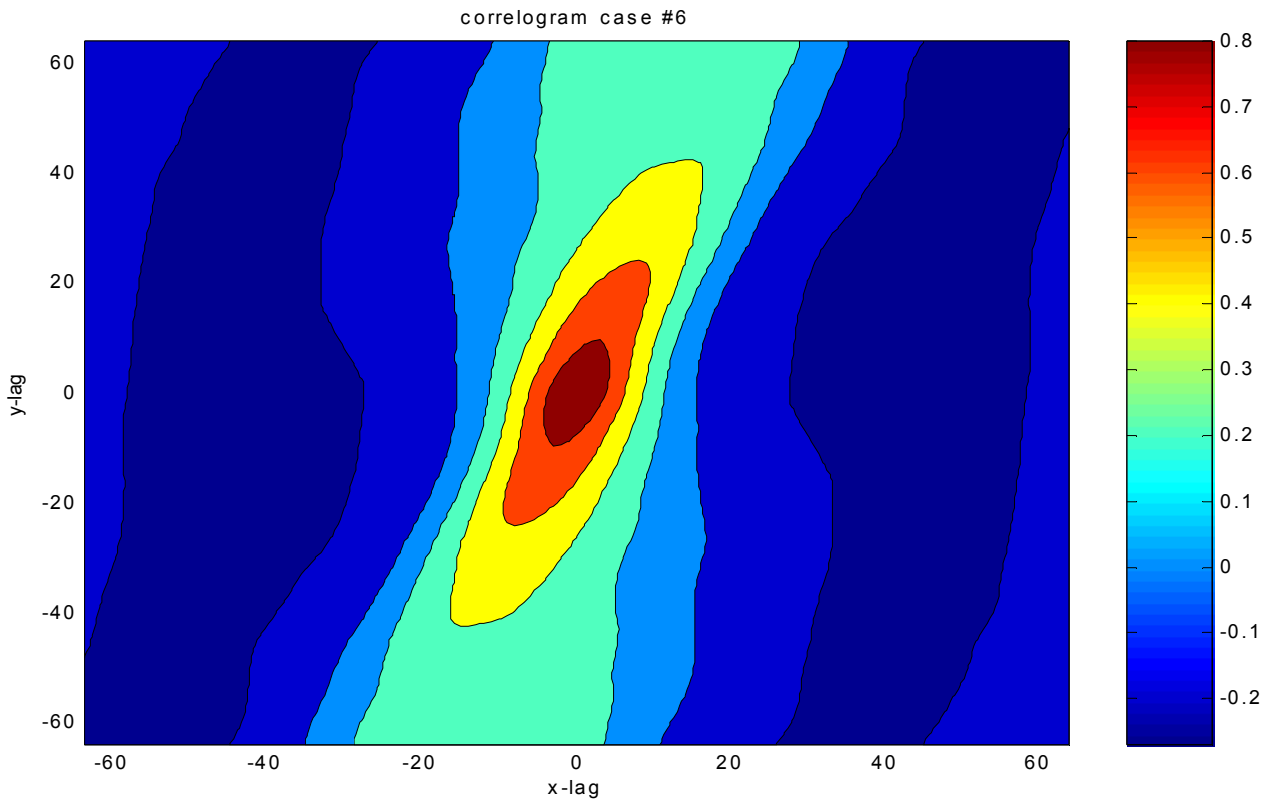
$$\gamma(h) = \left(\frac{1}{2N(h)}\right)\sum_{N(h)}\left(x_t - x_h\right)^2$$

$$C(h) = \left(\frac{1}{N(h)}\right)\sum_{N(h)}\left(x_t x_h - m_t m_h\right)$$

$$\rho(h) = \frac{C(h)}{\sigma_t \sigma_h}$$

An example correlogram

- An example of the correlogram showing the contours at various levels



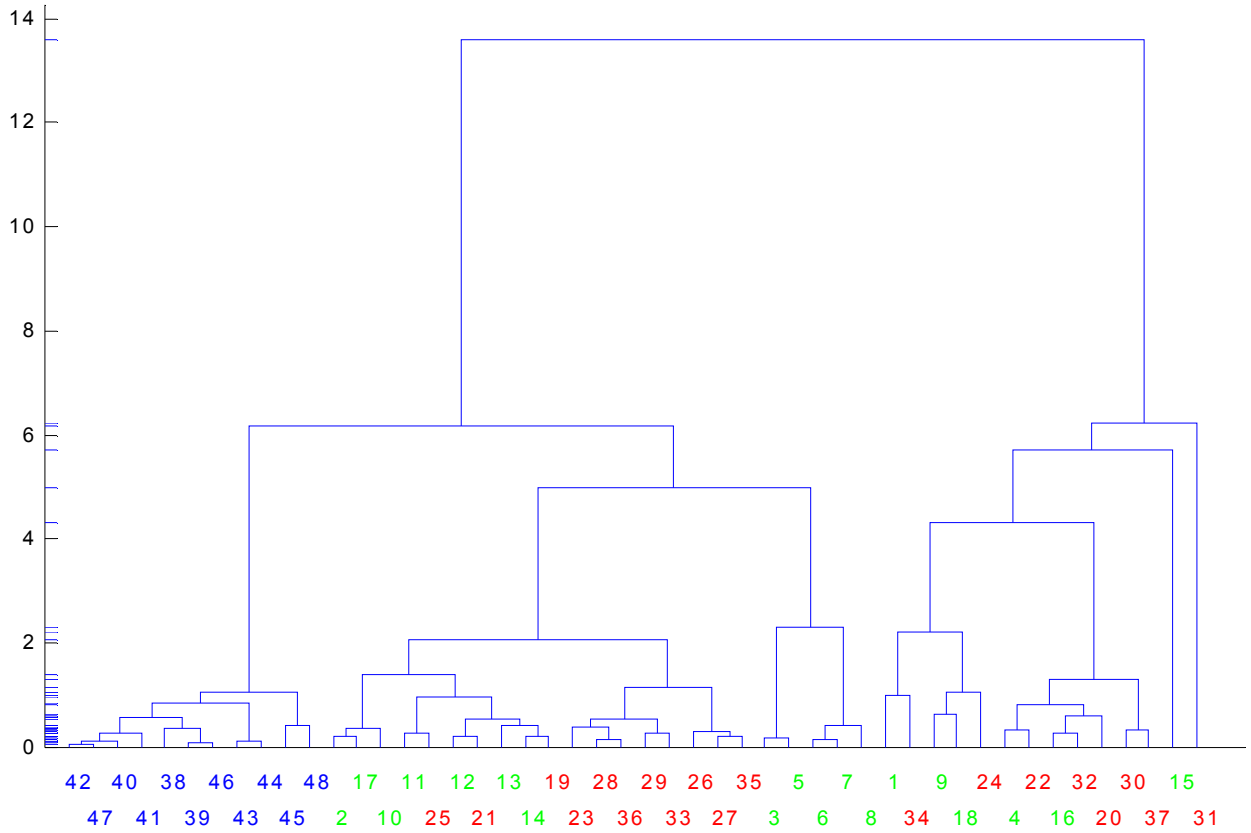
Extract two spatial attributes

- Identify the ellipse corresponding to the 60% level correlation
- Measure the semi-major axis A and the semi-minor axis B of this ellipse.
- Then, the product AB is a measure of the area of this ellipse and the ratio A/B is a measure of its eccentricity

A set of four attributes

- The intensity parameters a and b
- Spatial parameters AB and A/B
- Create an m * n data matrix X with m =4 and n=48
- Perform clustering experiments on this data matrix to obtain the classification

Dendrogram obtained by the Wards method



Automated system for rainfall classification system

- Based on the success of this experiment we have used these attributes to classify the 100 test cases from the year 2002
- This system can classify convective and non-convective with 89% accuracy (compared with those of human experts)
- This system can identify cellular, linear and stratiform with 85% accuracy

Summary

- We have reviewed the basic principles of the **clustering methodology**
- Provided applications of this methodology to two cases – analysis of **ensemble forecast** and **rain fall classification** system
- Demonstrated the development of an **automated classification** of rainfall system
- This approach is readily applicable to the analysis of a whole host of problems

Optimal partition is infeasible

- Given n objects, how many ways in which we can partition them into k non-empty subsets.
- Example: n=4, k=2
- {1,2,3},{4}; {1,2,4},{3}; {1,3,4},{2}; {2,3,4},{1}; {1,2},{3,4}; {1,3},{2,4}; {1,4},{2,3}
- There are seven ways

Number of distinct partitions

- The number of partitions obtained in this fashion is called the Stirling number of the second kind and is denoted by $T(n, k)$
- $T(n, 1) = 1$, $T(0, 0)$
- $T(0, 1) = 0$ $T(n, 0)$
- $T(n, k) = k\,T(n-1, k) + T(n-1, k-1)$
- $= 0$ for all $n > 0$

Infeasibility

- It can be shown that the value of
- $T(19, 3) = 6.9 * 10^{11}$
- Thus, the computational problem of obtaining the optimal clustering is infeasible