Cluster Analysis with SAS

(Some materials are based on SAS Documentation)

Chapter 1

Introduction to Clustering

Section 1.1

What is cluster analysis?

Definition

"Given a data set, find a partitioning of the data into groups (clusters), such that the patterns in a cluster are more similar to each other than patterns in different clusters."

--Jain AK & Dubes RC 1988

Clustering is an unsupervised learning

Learning without *a priori* knowledge about the classification of samples; learning without a teacher.

Kohonen (1995), "Self-Organizing Maps"

Characteristics of unsupervised learning:

- Only use independent variables
- No target variable is involved
- Does not predict a value
- Can be used to interpret

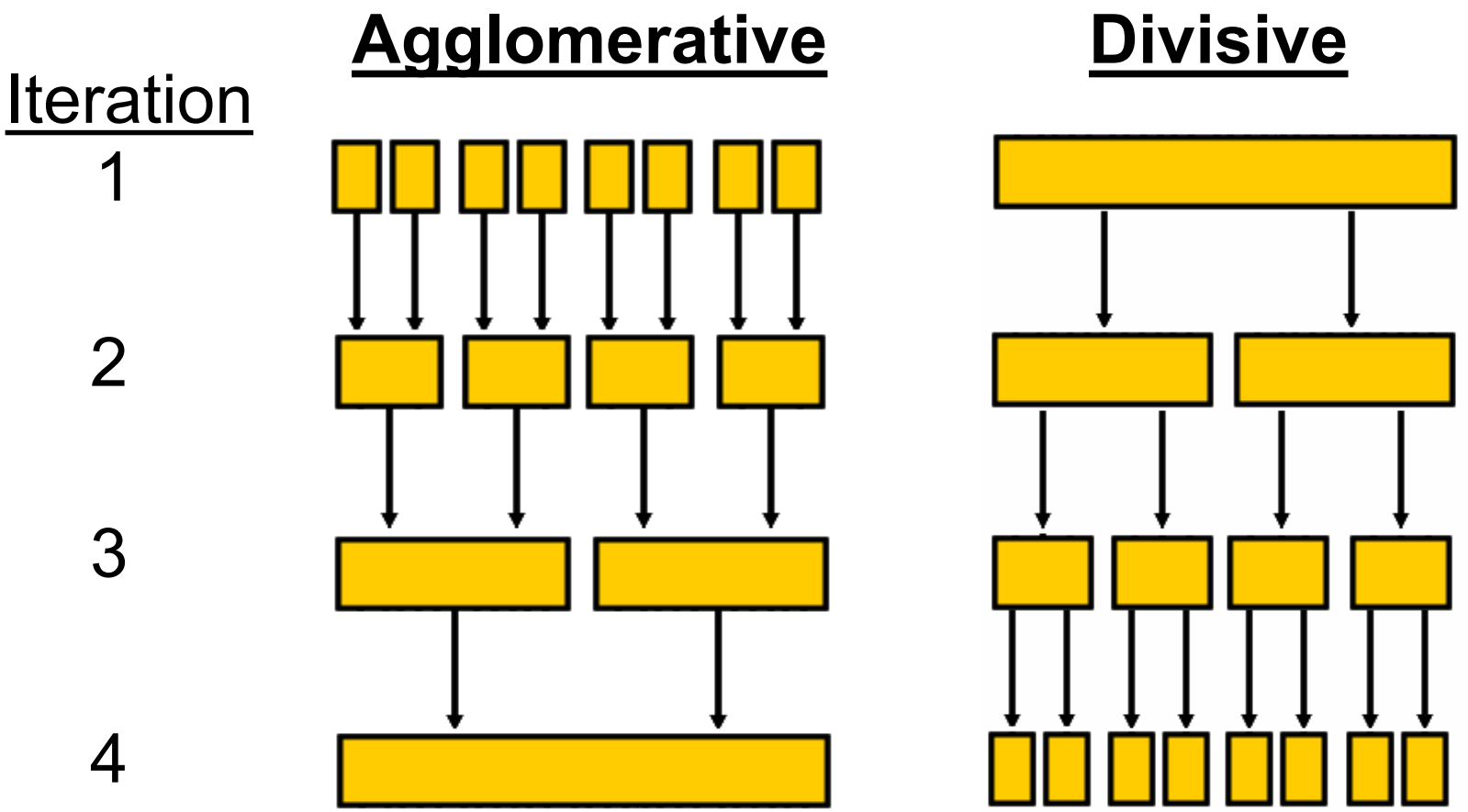
Section 1.2

Types of Clustering

Two Major Classes of Clustering Methods

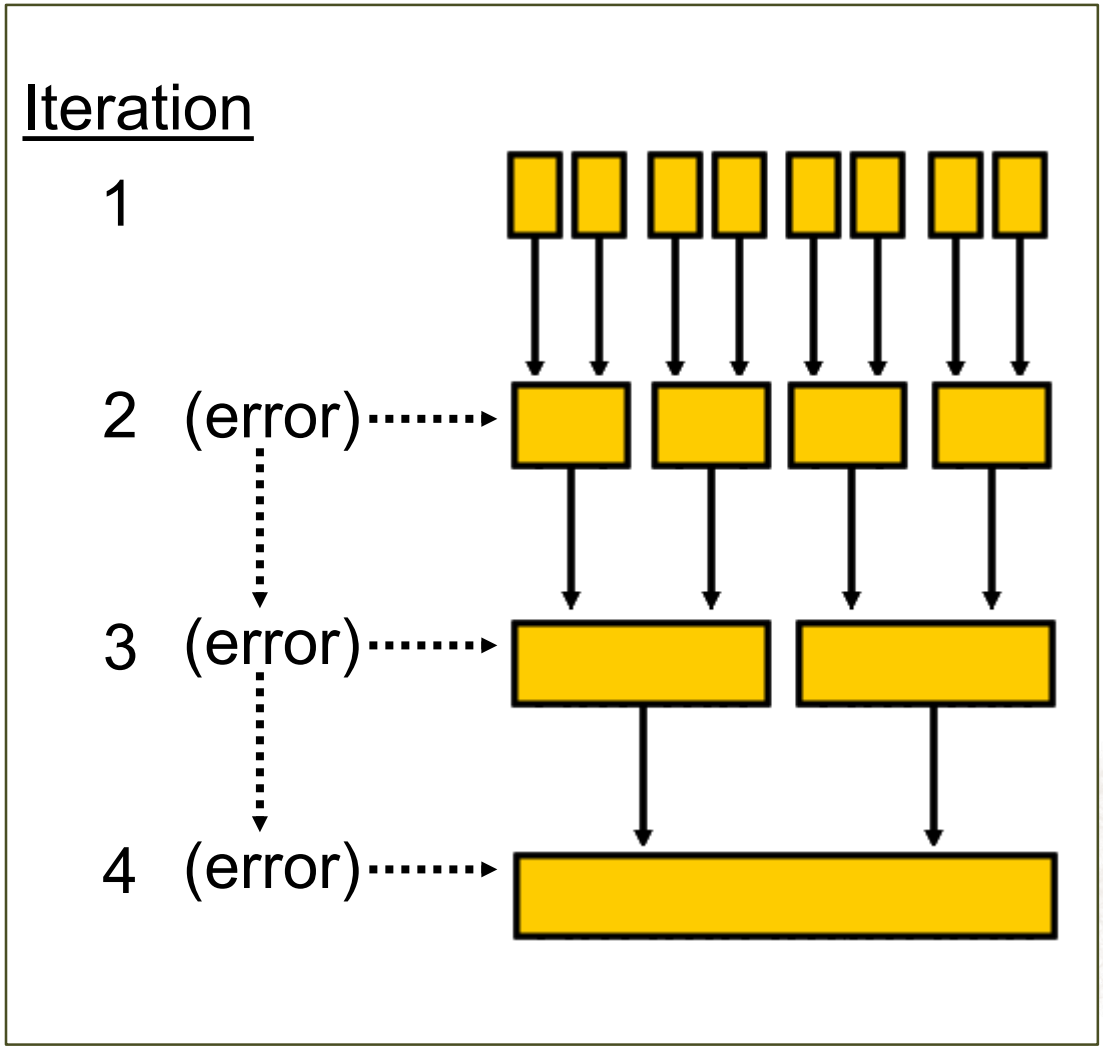
- Hierarchical clustering
- Optimization (partitive) clustering

Hierarchical Clustering

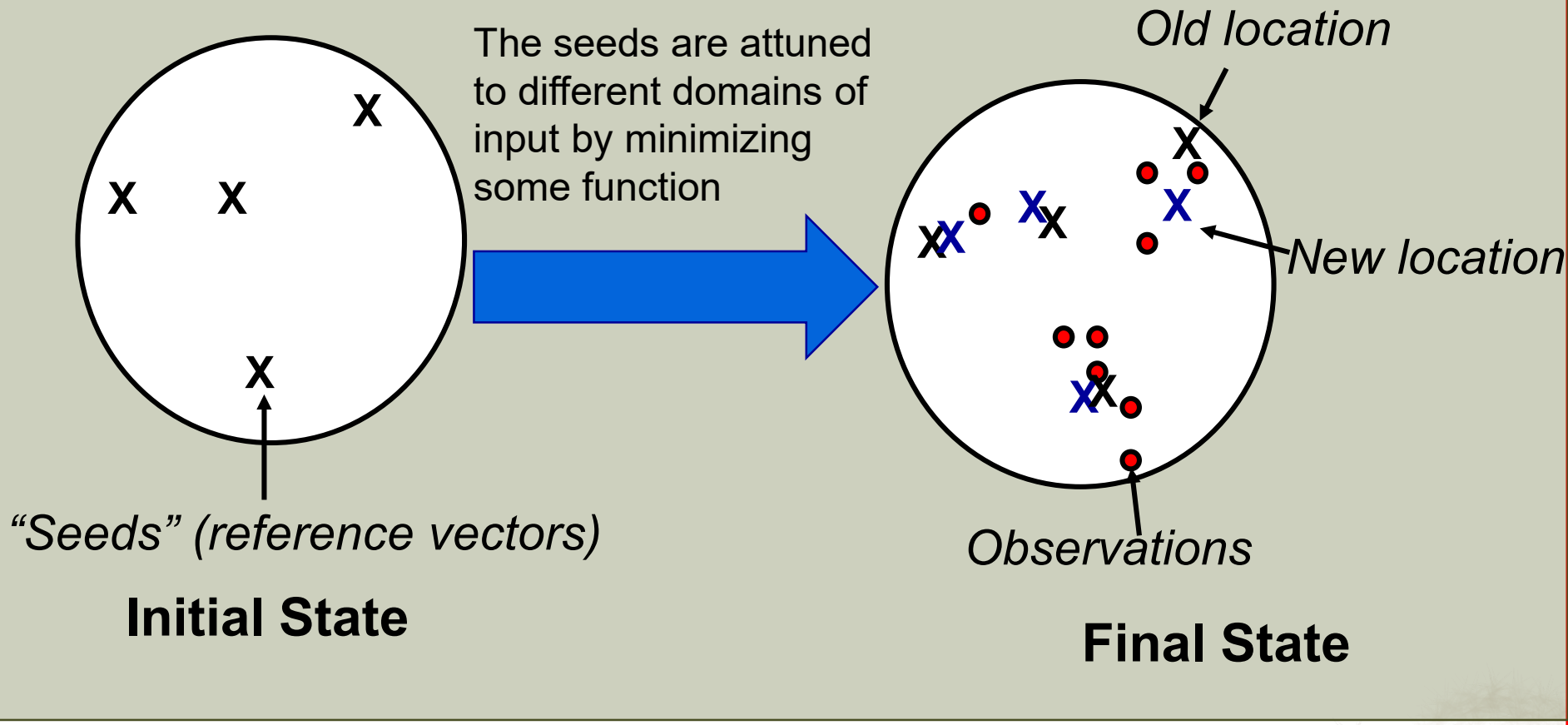


Hierarchical clustering is not perfect. The main drawback is propagation of errors

- Divisions are irrevocable: it can never repair the errors that were done in previous steps
- It does not scale up well with the number of observations



Optimization (Partitive) Clustering



Advantages:

- Does not depend on previously found clusters
- Scales up linearly with the number of observations

Shortcomings of Optimization Clustering

- Requires initial guess at the number of clusters.
- Is influenced by
  - the choice of initial seeds,
  - the presence of outliers and
  - the order in which the seeds are read.
- Makes explicit assumptions about the shape of clusters.

The number of possible partitions of dividing n observations into g clusters

$$N(n,g) = \frac{1}{g!}\sum_{m=1}^{g}(-1)^{g-m}\binom{g}{m}m^{n}$$

Examples:

- $N(50,4) = 5.3*10^{28}$
- $N(100,4) = 6.7*10^{58}$

So complete enumeration is (currently) impossible.

Heuristic Search

1. Find an initial partition of the *n* observations into *g* groups.
2. Calculate the change in some *function* by making some changes, e.g., moving each observation from its own cluster to another group.
3. Make the change resulting in the greatest improvement (or minimum loss) in the above function.
4. Repeat steps 2 and 3 until no move results in improvement.

Section 1.3

Similarity Metrics

How similar are they?

Which is more similar to a duck: a crow or a penguin?

How about these two?

What is similarity?

- Although the concept of similarity is fundamental to our thinking, it is also often difficult to precisely quantify.

- The metric that you choose to operationalize similarity often impacts the clusters you recover (for example, Euclidean distance or Pearson's correlation coefficient) .

What makes a good similarity metric?

The following three principles have been identified as a foundation of any good similarity metric:

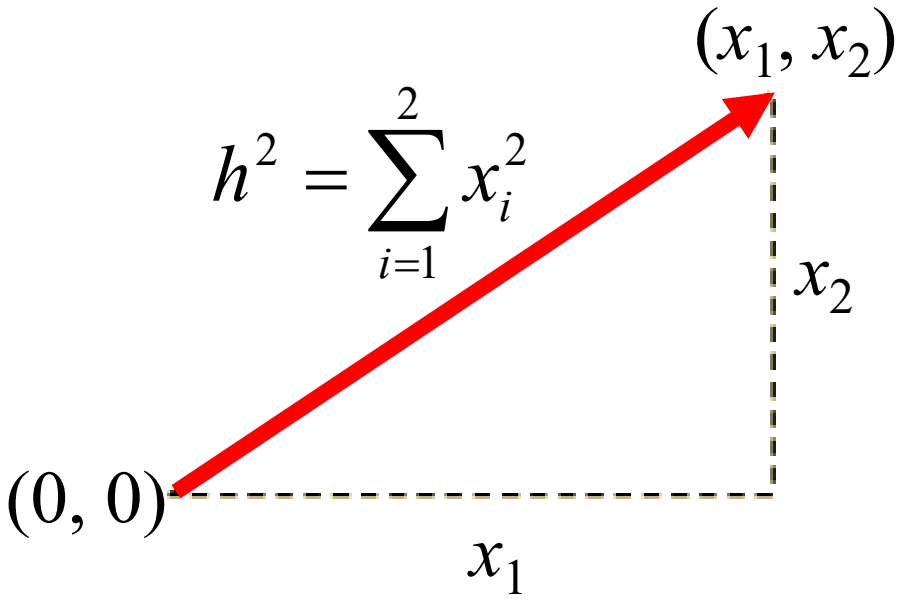
1. symmetry: $d(x,y) = d(y,x)$
2. non-identical distinguishability: if $d(x,y) \neq 0$ then $x \neq y$
3. identical non-distinguishability: if $d(x,y) = 0$ then $x = y$

Some Common Similarity Metrics

Euclidean Distance Similarity Metric (EUCLID)

$$D_E = \sqrt{\sum_{i=1}^{d} (x_i - w_i)^2}$$

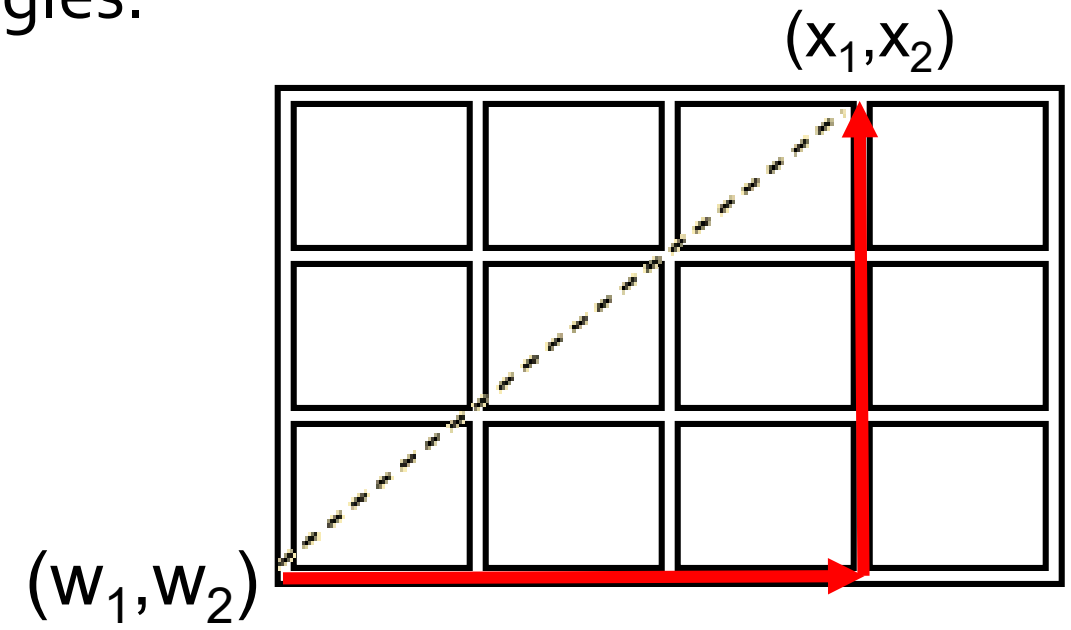
Pythagorean Theorem: The square of the hypotenuse is equal to the sum of the squares of the other two sides.



City Block Distance Similarity Metric (CITYBLOCK)

$$D_{M_1} = \sum_{i=1}^{d} |x_i - w_i|$$

City block (Manhattan) distance is the distance between two points measured along axes at right angles.



Minkowski Distance of Order λ

$$D_{M\lambda} = \left(\sum_{i=1}^{d} |w_i - x_i|^{\lambda} \right)^{1/\lambda}$$

When λ = 2, D_{M2} is the Euclidean distance

$$D_{M2} = \left(\sum_{i=1}^{d} |w_i - x_i|^{2} \right)^{1/2} \equiv \sqrt{\sum_{i=1}^{d} (w_i - x_i)^2}$$

When λ = 1, D_{M1} is the city block distance

$$D_{M1} = \sum_{i=1}^{d} |w_i - x_i|$$

Pearson's correlation coefficient (CORR)

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

i^{th} observation in cluster x

x cluster centroid

i^{th} observation in cluster y

y cluster centroid

$$r_{xy} \in [-1,1]$$

r_{xy} = -1, dissimilar
r_{xy} = 1, similar
r_{xy} = 0, no similarity.

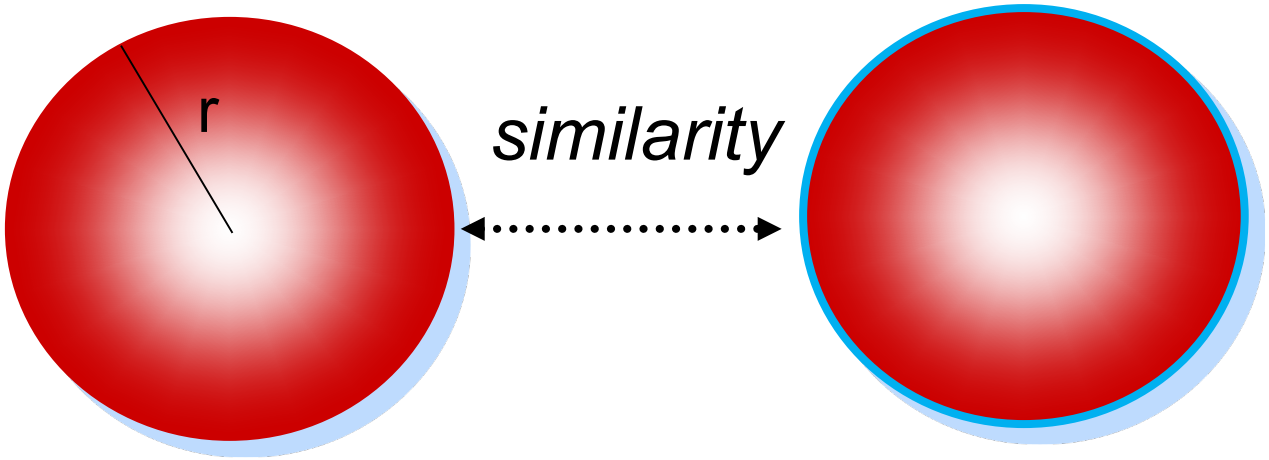
The Problem with Correlation

Variable	Observation 1	Observation 2
x_1	5	51
x_2	4	42
x_3	3	33
x_4	2	24
x_5	1	15
Mean	3	33
Std. Dev.	1.5811	14.2302

The correlation between observations 1 and 2 is a perfect 1.0, but are the observations **really** similar?

Density Estimate Based Similarity Metrics

Clusters can be seen as areas of increased observation density. Similarity is a function of the distance between the identified density bubbles (hyper-spheres).



Density at Point (i)

$$\hat{f}_i = \frac{n_i}{nv_i}$$

n_i, # of observations within the radius (r) of the hyper-sphere centered at i

n, total # of observations

v_i, volume of the i^{th} hyper-sphere

$$v(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} r^d$$

d, dimensionality

Γ, the gamma function

Similarity Metrics of Discrete Variables

- Hamming distance (HAMMING)
- Simple matching coefficient (MATCH)
- Nonmissing mismatches (X)
- Nonmissing matches (M)
- Jaccard similarity coefficient (JACCARD)
- Russell and Rao coefficient (RR)

Hamming Distance Similarity Metric

$$D_H = \sum_{i=1}^{d} |x_i - w_i|$$

	1 2 3 4 5 … 17
Gene A	0 1 1 0 0 1 0 0 1 0 0 1 1 1 0 0 1
Gene B	0 1 1 1 0 0 0 0 1 1 1 1 1 1 0 1 1

(low=0, high=1)

D_H = 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0 1 0 = 5

Gene expression levels under 17 conditions

The DISTANCE Procedure

General form of the DISTANCE procedure:

```
PROC DISTANCE  DATA=dataset METHOD=method <options> ;
        VAR level (variables < / option-list >) ;
RUN;
```

Both the PROC DISTANCE statement and the VAR statement are required.

An Example

Dividend yields of 15 US utility stocks in the period of 1986-1990

```
%let inputs=div_1986 div_1987 div_1988 div_1989
    div_1990;
title 'Stock Dividends';
title2 'The STOCK Data Set';
proc print data=work.stock;
   var company &inputs;
run;
```

Stock Dividends 17:52 Sunday, November 16, 2008 1

The STOCK Data Set

Obs	company	div_1986	div_1987	div_1988	div_1989	div_1990
1	Cincinnati G&E	8.4	8.2	8.4	8.1	8.0
2	Texas Utilities	7.9	8.9	10.4	8.9	8.3
3	Detroit Edison	9.7	10.7	11.4	7.8	6.5
4	Orange & Rockland Utilitie	6.5	7.2	7.3	7.7	7.9
5	Kentucky Utilities	6.5	6.9	7.0	7.2	7.5
6	Kansas Power & Light	5.9	6.4	6.9	7.4	8.0
7	Union Electric	7.1	7.5	8.4	7.8	7.7
8	Dominion Resources	6.7	6.9	7.0	7.0	7.4
9	Allegheny Power	6.7	7.3	7.8	7.9	8.3
10	Minnesota Power & Light	5.6	6.1	7.2	7.0	7.5
11	Iowa-Ill Gas & Electric	7.1	7.5	8.5	7.8	8.0
12	Pennsylvania Power & Light	7.2	7.6	7.7	7.4	7.1
13	Oklahoma Gas & Electric	6.1	6.7	7.4	6.7	6.8
14	Wisconsin Energy	5.1	5.7	6.0	5.7	5.9
15	Green Mountain Power	7.1	7.4	7.8	7.8	8.3

Generating Distances (Euclidean)

```
proc distance data=work.stock method=Euclid
        out=distances;
  id company;
  var interval(&inputs/std=range);
run;
```

id statements specify a single variable to be copied to the OUT= dataset and used to generate names for the distance variables

```
title2 'Euclidean Distance Matrix';
proc print data=distances;
  id company;
run;
```

Euclidean Distance Matrix

Euclidean Distance Matrix

company	Cincinnati_ G_E	Texas_ Utilities	Detroit_ Edison	Orange___ Rockland_ Utilitie	Kentucky_ Utilities	Kansas_ Power___ Light
Cincinnati G&E	0.00000	.	.	.	.	.
Texas Utilities	0.49670	0.00000	.	.	.	.
Detroit Edison	1.01879	0.99884	0.00000	.	.	.
Orange & Rockland Utilitie	0.51910	0.84035	1.37538	0.00000	.	.
Kentucky Utilities	0.65416	1.02098	1.39076	0.24265	0.00000	.
Kansas Power & Light	0.74161	1.04618	1.58831	0.24213	0.27325	0.00000
Union Electric	0.35197	0.65306	1.13477	0.26463	0.37514	0.47420
Dominion Resources	0.67235	1.05719	1.36747	0.31596	0.08679	0.34455
Allegheny Power	0.44817	0.70706	1.37683	0.20627	0.43497	0.36120
Minnesota Power & Light	0.87055	1.17406	1.57540	0.40330	0.26298	0.26451
Iowa-Ill Gas & Electric	0.32954	0.60517	1.18664	0.26965	0.43259	0.46890
Pennsylvania Power & Light	0.53641	0.90027	1.10787	0.39363	0.30207	0.54776

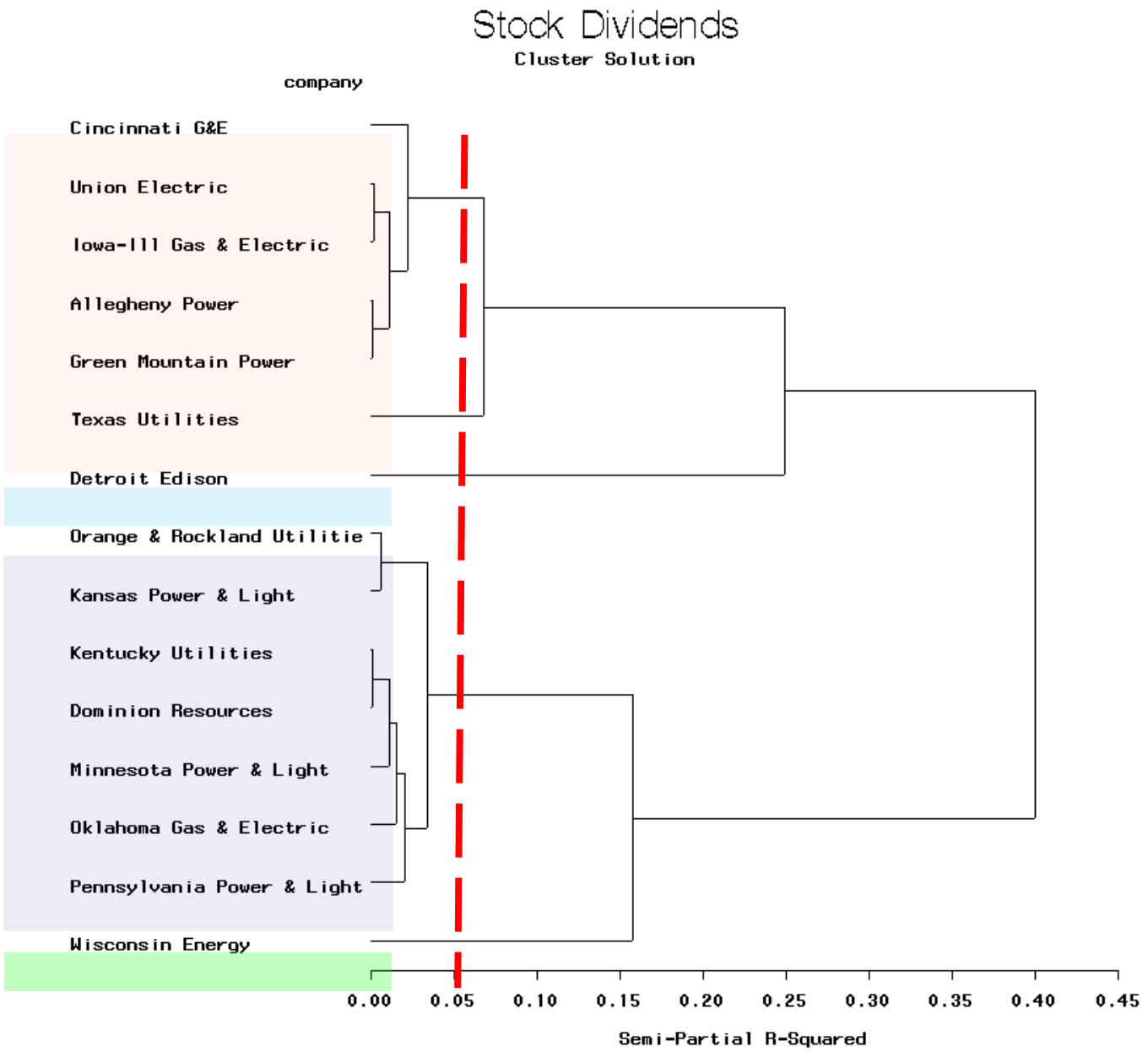
company	Union_ Electric	Dominion_ Resources	Allegheny_ Power	Minnesota_ Power___ Light	Iowa_Ill_ Gas___ Electric	Pennsylvania_ Power___Light
Cincinnati G&E	.	.	.	.	.	.
Texas Utilities	.	.	.	.	.	.
Detroit Edison	.	.	.	.	.	.
Orange & Rockland Utilitie	.	.	.	.	.	.
Kentucky Utilities	.	.	.	.	.	.
Kansas Power & Light	.	.	.	.	.	.
Union Electric	0.00000	.	.	.	.	.
Dominion Resources	0.40903	0.00000	.	.	.	.
Allegheny Power	0.29152	0.49807	0.00000	.	.	.
Minnesota Power & Light	0.55096	0.29307	0.56333	0.00000	.	.
Iowa-Ill Gas & Electric	0.12636	0.47342	0.20632	0.59042	0.00000	.
Pennsylvania Power & Light	0.30952	0.28190	0.53868	0.51280	0.42317	0.00000

company	Oklahoma_ Gas___ Electric	Wisconsin_ Energy	Green_ Mountain_ Power
Cincinnati G&E	.	.	.
Texas Utilities	.	.	.
Detroit Edison	.	.	.
Orange & Rockland Utilitie	.	.	.
Kentucky Utilities	.	.	.
Kansas Power & Light	.	.	.
Union Electric	.	.	.
Dominion Resources	.	.	.
Allegheny Power	.	.	.
Minnesota Power & Light	.	.	.
Iowa-Ill Gas & Electric	.	.	.
Pennsylvania Power & Light	.	.	.

From Distance Matrix to Tree

```
proc cluster data=distances method=ward
    outtree=tree

        noprint;

  id company;

run;


title2 "Cluster Solution";

proc tree data=tree horizontal;

  id company;

run;
```

The Dendrogram (Tree)



Section 1.4

Clustering Performance

Quality of the Cluster Solution as Indicated by a Confusion Matrix

Class \ Cluster	1	2	3	Total
A	50	0	0	50
B	50	0	0	50
C	50	0	0	50
Total	150	0	0	150

No Solution

Class \ Cluster	1	2	3	Total
A	50	0	0	50
B	0	40	10	50
C	10	5	35	50
Total	60	45	45	150

Typical Solution

Class \ Cluster	1	2	3	Total
A	50	0	0	50
B	0	50	0	50
C	0	0	50	50
Total	50	50	50	150

Perfect Solution

Probability of Cluster Assignment

The probability that a cluster number represents a given class is given by the cluster's proportion of the row total.

Class	Cluster 1	2	3	Total
A	50	0	0	50
B	0	40	10	50
C	10	5	35	50
Total	60	45	45	150

Frequency

P = cell freq/row total

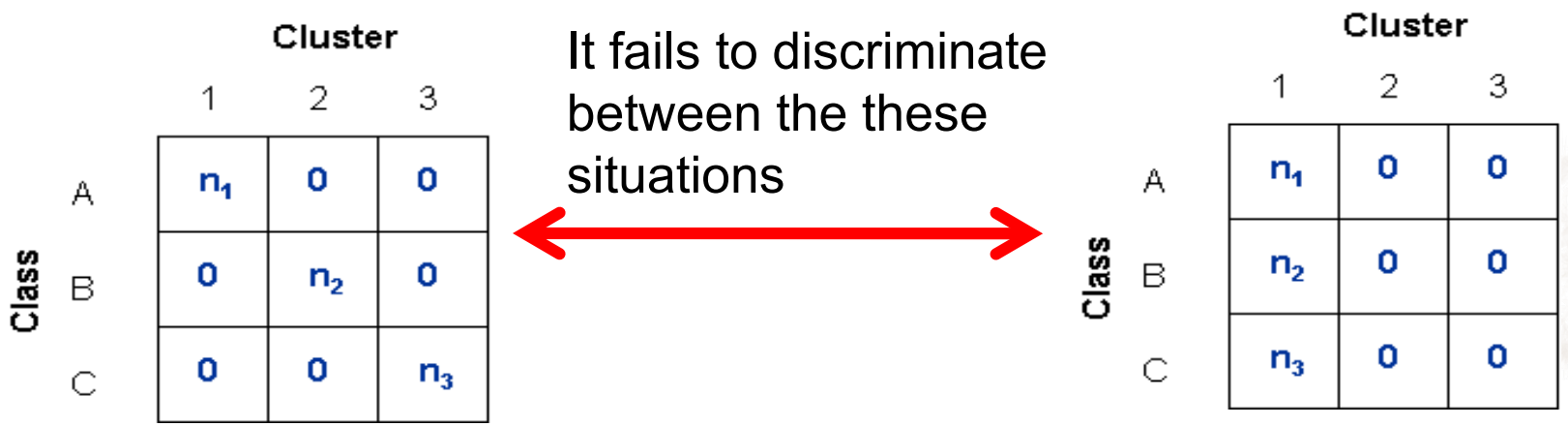
Class	Cluster 1	2	3	Total
A	1	0	0	1
B	0	0.8	0.2	1
C	0.2	0.1	0.7	1

Probability

The mode of a given class is its most likely cluster assignment.

Misclassification Metric

$$\sum_{i=1}^{\text{classes}} (\text{row total}_i - \text{row mode}_i)$$



The Chi-square Statistic

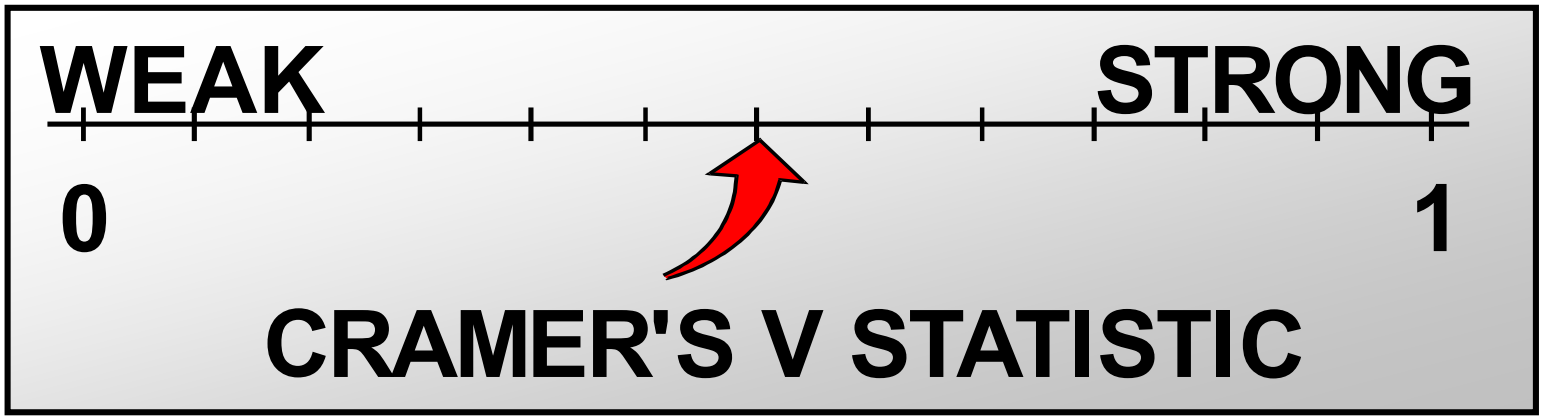
$$\chi^2 = \sum_i \sum_j \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}}$$

The chi-square statistic (and associated probability)

- determines whether an association exists
- depends on sample size
- but does **not** measure the strength of the association.

Measuring Strength of an Association

$$Cramer\ 's\ V = \sqrt{\frac{\chi^2 / n}{\min(r-1, c-1)}}$$



r , number of rows

c, number of columns

n, overall total of the cell frequencies, $n = \sum_i \sum_j n_{ij}$