HUDM5124 Session 9: Introduction to Clustering Methods

Overview:

- Types of cluster models
- Types of clustering algorithms
- Data for clustering
- Issues
- A. Partitioning Methods

A taxonomy of models for proximity data:

- Geometric models
  - Simple geometric (MDS) models
  - Weighted geometric (MDS) models
- Cluster/tree models
  - A. Partitions
  - B. Hierarchical/nested cluster models (= trees)
  - C. Overlapping clusters
- Graph/network models
  - Undirected graphs
  - Directed graphs (for asymmetric proximities)

Some general clustering references:

- Hartigan, John A. (1975). *Clustering Algorithms.* New York: Wiley.
- Jain, A.K., & Dubes, R.C. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.
- Aldenderfer, M.S., & Blashfield, R.K. (1984). *Cluster Analysis*. (Sage University Papers series: Quantitative Applications in the Social Sciences, series no. 07-44). Thousand Oaks CA: Sage.
- Sneath, P.H.A., & Sokal, R.R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W. H. Freeman.
- Arabie, P., & Hubert, L.J. (1996). An overview of combinatorial data analysis. In P. Arabie, L. Hubert, & G. De Soete (Eds.), *Clustering and Classification*. River Edge NJ: World Scientific.
- Gower, J.C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients*.* *Journal of Classification, 3*, 5-48.

The Clustering Problem

Given: multivariate data (either "multivariate" two-way two-mode data, or proximities) on N objects or entities,

Find: a set of clusters (where a cluster = a subset of objects) that tend to maximize within-cluster homogeneity, or between-category heterogeneity, or both (Cormack, 1971).

For proximity data, homogeneity or heterogeneity can be defined as high or low mean similarity.

For multivariate data, homogeneity or heterogeneity can be defined in terms of sums of squares: SSW and SSB, or by computed distances to the cluster "centroid".

Uses of cluster analysis:

- development of a taxonomy or classification scheme
- useful schemes for grouping entities
- theory generation (exploratory)
- theory confirmation (hypothesis testing)

Generally, cluster methods are exploratory methods for multivariate data

Clustering Algorithms:

(warning: this taxonomy mixes models and algorithms)

Aldenderfer & Blashfield (1984) taxonomy:

1. hierarchical agglomerative
2. hierarchical divisive
3. iterative partitioning
4. density search
5. factor analytic
6. clumping
7. graph theoretic

IT IS IMPORTANT TO DISTINGUISH MODELS FROM ALGORITHMS!

Three basic types of clustering models:

1. partitions
2. hierarchies (=trees)
3. overlapping clusters

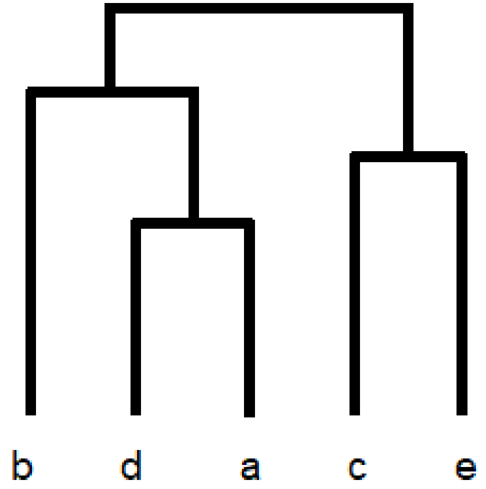
(these models are typically represented graphically in different ways; mathematically, they can be distinguished by the set relations among the clusters)

Three types of clustering model:

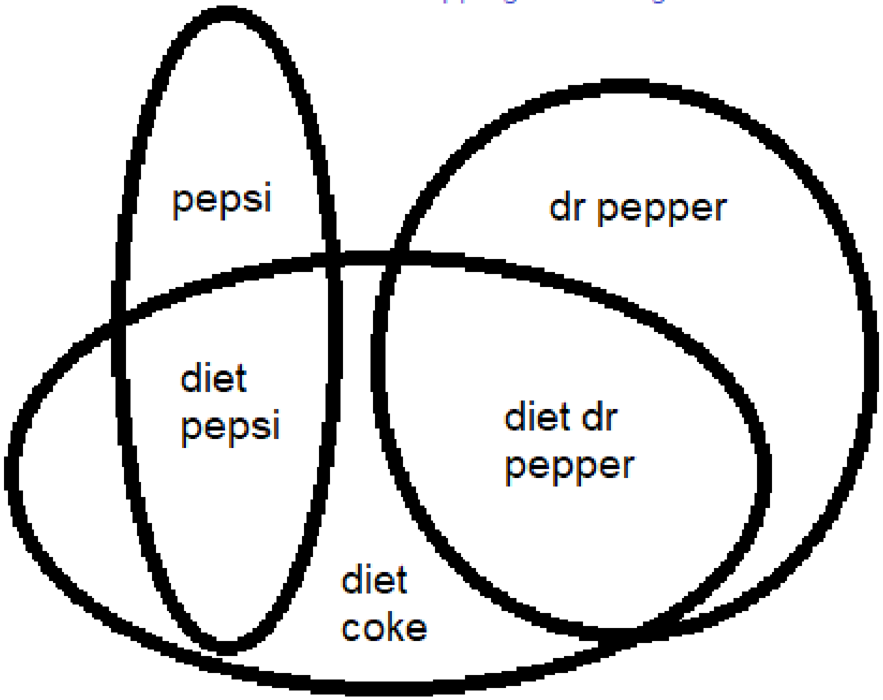
partitioning

P = [{dog,cat}{canary,parrot}{turtle}]

hierarchical clustering



overlapping clustering



Mathematical characterization of types of cluster models

Carroll & Corter (1995) pointed out that given a clustering C = {c_i}, the relationship between any two clusters c_k and c_l satisfies exactly one of the following relationships:

$$R1:\ c_k \cap c_l = \phi \quad \text{(disjointness)}$$

$$R2:\ c_k \subset c_l \vee c_k \supset c_l \quad \text{(inclusion)}$$

$$R3:\ c_k \cap c_l \neq \phi,\ c_k \not\subset c_l,\ c_k \not\subset c_l \quad \text{(overlap)}$$

Partitions

The set of clusters C = {ci} are mutually exclusive and exhaustive subsets of the objects to be clustered; thus, between any two distinct clusters *k* and *l*, R1 must hold.

Hierarchical clustering models (= trees)

For any two clusters, either R1 or R2 holds (Carroll & Corter, 1995)

Overlapping clustering

For any two clusters *k* and *l*, R1, R2, or R3 may hold.

Types of data for clustering:

Depends on particular algorithm used:

1. proximities
2. rectangular (multivariate) data

For the methods that begin with multivariate data, some of them compute proximities as an explicit step. Others use statistics other than proximity to decide which objects belong together.

For the algorithms that explicitly calculate proximities, the data analyst may need to choose the type of similarity coefficient to be used in computing the proximities. Sometimes the algorithm allows a lot of choices, sometimes none.

Types of explicit proximity measures

1. correlation measures (Pearson, rank-order, etc.)
2. distance measures (Euclidean, city-block; Mahalonobis distance)
3. association measures (phi, lambda, etc)
4. probabilistic similarity measures (e.g., based on information, uncertainty)

Some association coefficients for binary (categorical) variables

(attr. k)		Case *j*		
		1	0	
Case *i*	1	a	b	(a+b)
	0	c	d	(c+d)
		(a+c)	(b+d)	*K*

- phi coefficient

  $\Phi = (ad-bc)/(\surd(a+b)(c+d)(a+c)(b+d))$
- matching coefficient

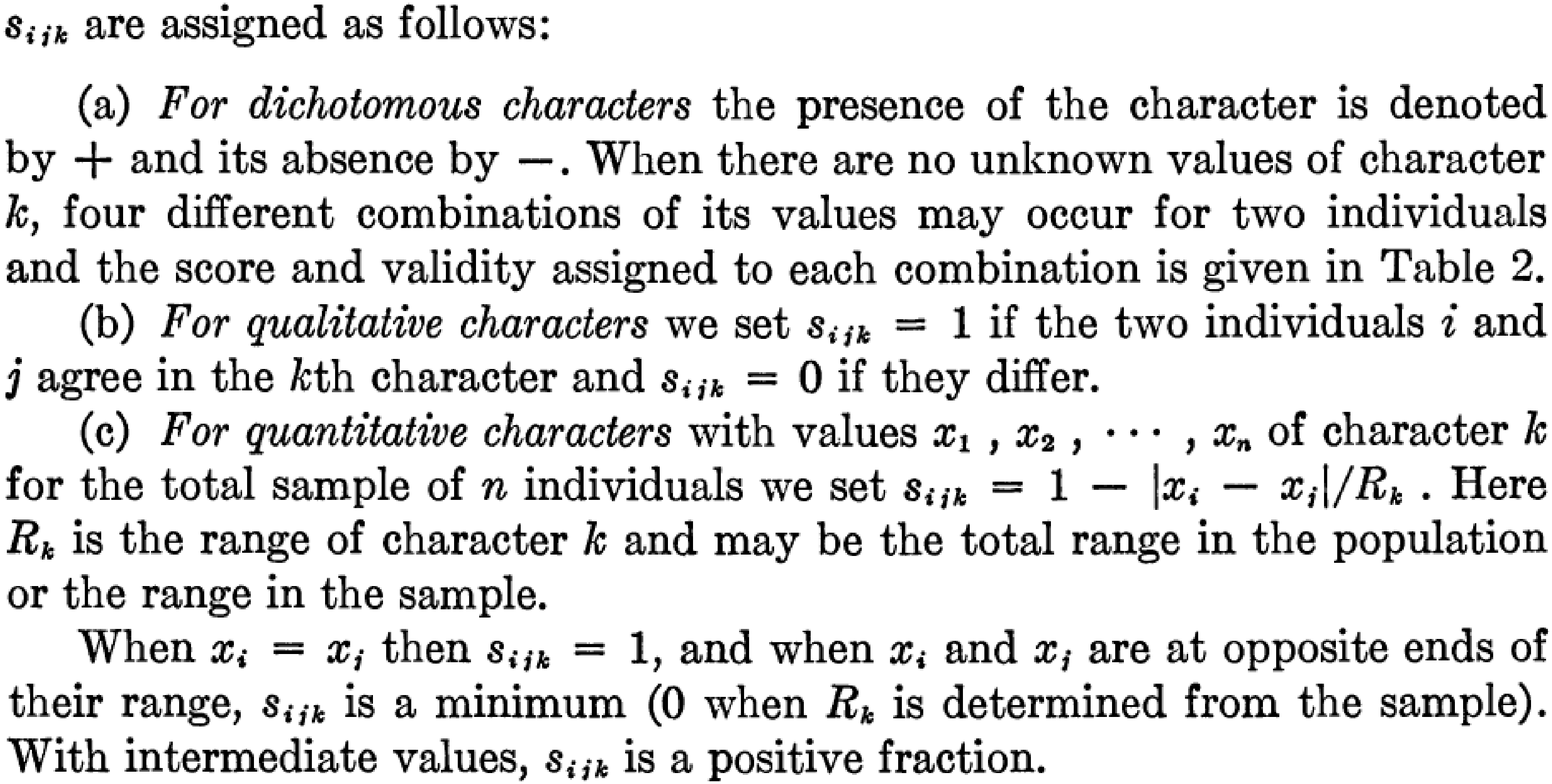
  $S = (a+d)/(a+b+c+d)$
- Jaccard's coefficient

  $S = a/(a+b+c)$
- Gower's coefficient (Gower, 1971) (for mixed data type)

$$S_{ij} = \frac{\sum_k^K w_{ijk}\, s_{ijk}}{\sum_k^K w_{ijk}}$$

(w_{ijk} =1 if the comparison is “valid”, 0 otherwise)

Note: How s_{ijk} is computed in Gower's coefficient (Gower, 1971)



s_{ijk} are assigned as follows:

(a) *For dichotomous characters* the presence of the character is denoted by + and its absence by −. When there are no unknown values of character k, four different combinations of its values may occur for two individuals and the score and validity assigned to each combination is given in Table 2.

(b) *For qualitative characters* we set $s_{ijk} = 1$ if the two individuals i and j agree in the kth character and $s_{ijk} = 0$ if they differ.

(c) *For quantitative characters* with values $x_1, x_2, \cdots, x_n$ of character k for the total sample of n individuals we set $s_{ijk} = 1 - |x_i - x_j|/R_k$. Here R_k is the range of character k and may be the total range in the population or the range in the sample.

When $x_i = x_j$ then $s_{ijk} = 1$, and when x_i and x_j are at opposite ends of their range, s_{ijk} is a minimum (0 when R_k is determined from the sample). With intermediate values, s_{ijk} is a positive fraction.

Discrete-Feature Models of Similarity

- Contrast model of similarity (Tversky, 1977)

  $s(x,y) = \theta\, f(A \cap B) - \alpha\, g(A-B) - \beta\, g(B-A)$

  Special cases:

  $s(x,y) = \theta\, f(A \cap B)$ common-feature model

  $s(x,y) = -\alpha\, g(A-B) - \beta\, g(B-A)$ distinctive-features model

SOME ISSUES IN SELECTING A PROXIMITY MEASURE:

- what do you want to measure? (e.g. distance vs. profile level vs. profile similarity vs. ??)
- level of measurement of data (may be mixed)
- missing data
- meaningfulness
- structured objects

Also:

- selection of variables
- standardization of variables

Variable selection

Problem: Some variables in a multivariate data set may be relevant, others irrelevant. In big data / data mining applications, the vast majority of possible variables may be irrelevant, or it is *unknown* if they are useful.

SO, perhaps we want to let the data itself decide what variables are relevant (analogous to exploratory stepwise regression, or "structure learning" in Bayesian Network or SEM models). For example, what if some variables are basically "flat" (i.e., constant), or contain simply noise, or are completely redundant with other variable already in the set?

→ This is the "variable selection problem"

Some classic references in the statistics / clustering literature:

- Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification, 5(2),*205-228.
- Steinley, D., & Brusco, M. J. (2008). A new variable weighting and selection procedure for k-means cluster analysis. *Multivariate Behavioral Research, 43(1)*, 77-108.
- Friedman, J.H., & Meulman, J.J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 66(4)*, 815-849.

And lots of recent work in data mining / machine learning / pattern recognition:

E.g., Guyon & Elisseeff (2003); Raferty & Dean (2006); Witten & Tibshirani (2013)

Variable standardization

Problem: different variables in a multivariate data set may be measured on different scales – then the variables with large values (e.g. income in $) will dominate those with small values (e.g., number of siblings) in computing proximity.

Solution(?): do PCA on the variables, use component scores as input to clustering algorithm. This also 'adjusts" for correlation of variables (but this may obscure cluster structure, because based on spatial / continuous model)

Solution(?): standardize all variables to z-scores before computing proximities - but this may give too much weight to trivial "flat" variables.

Solution(?): standardize by range (e.g., Milligan; Steinley & Brusco)

Some references:

- Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 5,* 181-204.
- Steinley, D., & Brusco, M. J. (2008). A new variable weighting and selection procedure for k-means cluster analysis. *Multivariate Behavioral Research, 43(1)*, 77-108.

Partitioning

A partition of a set of objects X is a set of mutually exclusive and exhaustive subsets of X.

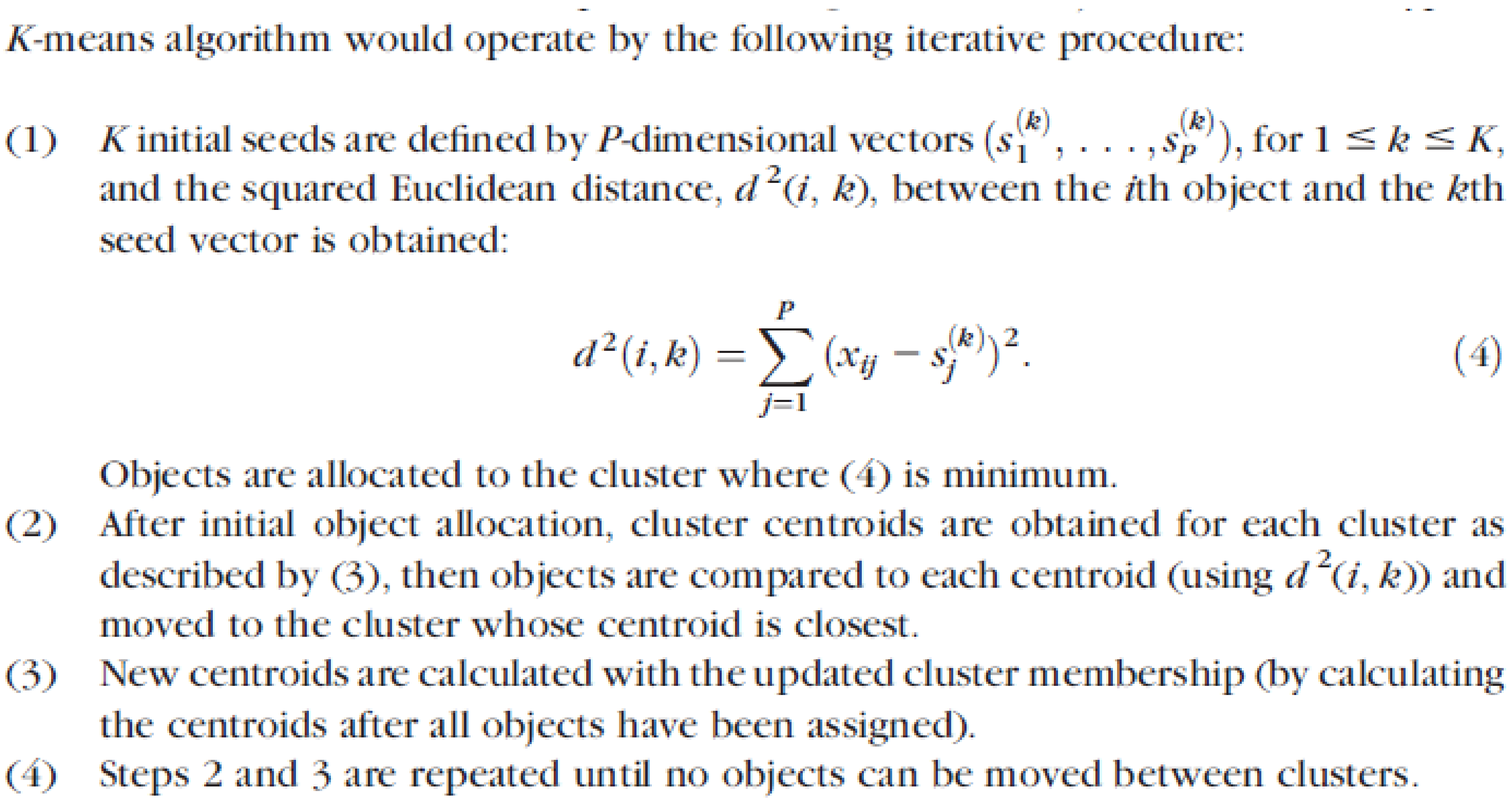
The partitioning (algorithm) problem has many variants, but the general idea is to find the optimal partition into k classes, that maximizes mean intra-class similarity (and inter-class distinctiveness). This method begins with “multivariate” two-way two-mode data, not proximities. See Steinley (2006) for a review.

This problem involves combinatorial optimization, and there are approximately $k^N/k!$ possible solutions. So exhaustive search is not practical. Instead, heuristic approaches have been proposed. The k-means algorithm (Hartigan & Wong, 1979) is perhaps the best-known partitioning algorithm.

Software: QUICK CLUSTER in SPSS; PROC FASTCLUS in SAS; kmeans in R

The generic k-means algorithm

(Steinley, 2006)



K-means algorithm would operate by the following iterative procedure:

(1) *K* initial seeds are defined by *P*-dimensional vectors $(s_1^{(k)}, \ldots, s_P^{(k)})$, for $1 \leq k \leq K$, and the squared Euclidean distance, $d^2(i, k)$, between the *i*th object and the *k*th seed vector is obtained:

$$d^2(i,k) = \sum_{j=1}^{P} (x_{ij} - s_j^{(k)})^2. \qquad (4)$$

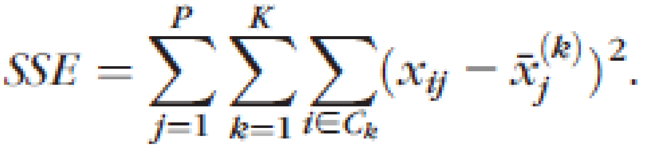
Objects are allocated to the cluster where (4) is minimum.

(2) After initial object allocation, cluster centroids are obtained for each cluster as described by (3), then objects are compared to each centroid (using $d^2(i, k)$) and moved to the cluster whose centroid is closest.

(3) New centroids are calculated with the updated cluster membership (by calculating the centroids after all objects have been assigned).

(4) Steps 2 and 3 are repeated until no objects can be moved between clusters.

The object is to minimize the SSW or SSE:



$$SSE = \sum_{j=1}^{P} \sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_j^{(k)})^2.$$

This is equivalent to minimizing tr(**W**)

Variants and options for k-means

Initial cluster seeds:

a. randomly selected cases, or random points in p-space (use only with a large number of restarts)

b. “well-separated” cases (e. g., Ball & Hall, 1967)

c. user-specified seeds

d. seeds suggested by a prior hierarchical clustering (e.g. Milligan, 1980)

Updating of cluster centroids:

-do this after each case is reassigned, OR only at the end of each “pass” through the entire set of cases

Variants and options for k-means

Some software implementations of iterative partitioning algorithms offer a choice of distance metrics (used to assign a case to “nearest” cluster). However, use of anything other than Euclidean or squared Euclidean distance makes it less clear what the algorithm is attempting to optimize, thus is non-standard.

Methods to decide on the number of clusters, k:

See Steinley (2006) for a brief review relevant to k-means.

(A simulation study of different methods: Milligan & Cooper, 1985)

NOTE: Such general clustering issues as variable selection, standardization, weighting, etc. arise in k-means.

REFERENCE: Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*, 1–34.

Validating Cluster Solutions

RELIABILITY:

-do different samples / different methods / different “runs” yield the same cluster solution?

VALIDITY:

-correspondence to an external criterion clustering solution

-face validity: do the clusters “make sense”?

-predictive power of the clusters (re external “DVs”)

COMPARING PARTITIONS

(to assess reliability, validity, replicability)

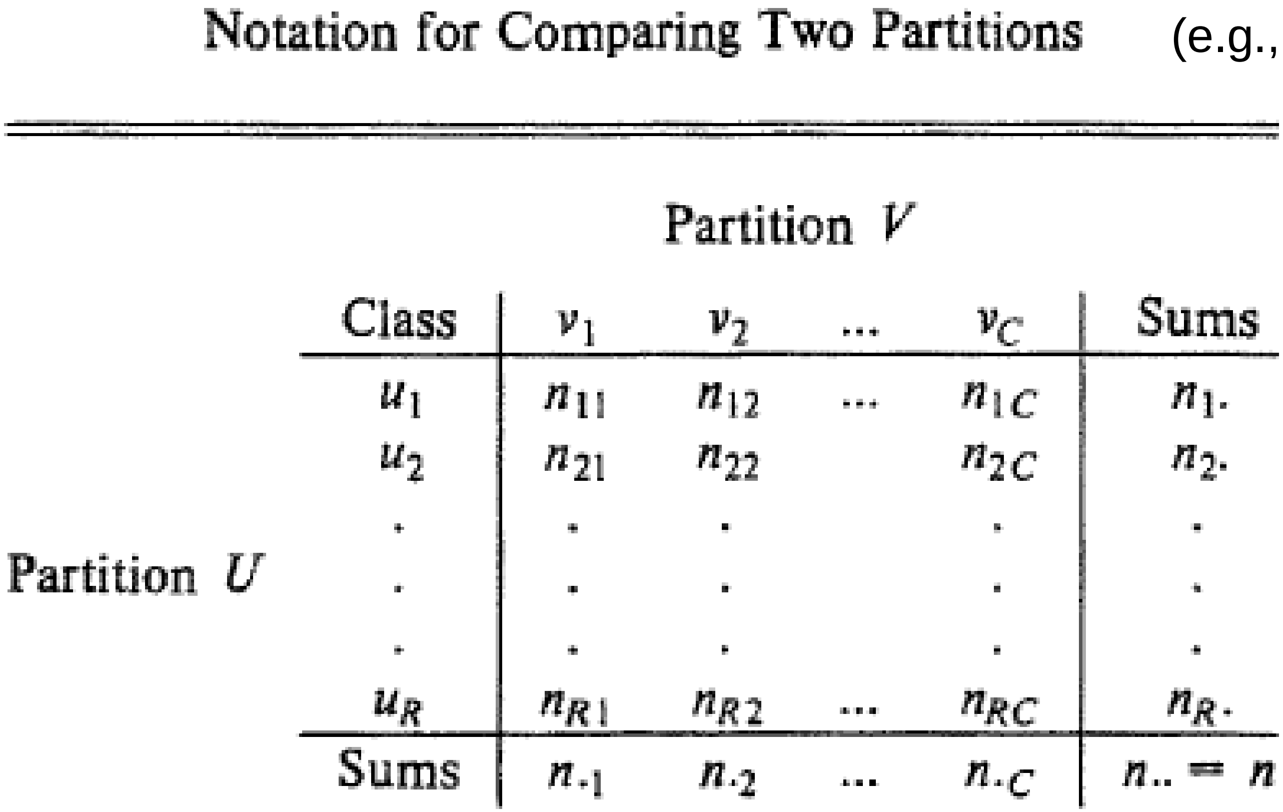
REFERENCES:

Morey, L., & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement, 44*, 33-37.

Hubert, L.J., and Arabie, P. (1985). Comparing partitions*.* *Journal of Classification, 2*, 193-218.

Warrens, M. J. (2008). On the equivalence of Cohen’s Kappa and the Hubert-Arabie Adjusted Rand Index. *Journal of Classification, 25*, 177-183.

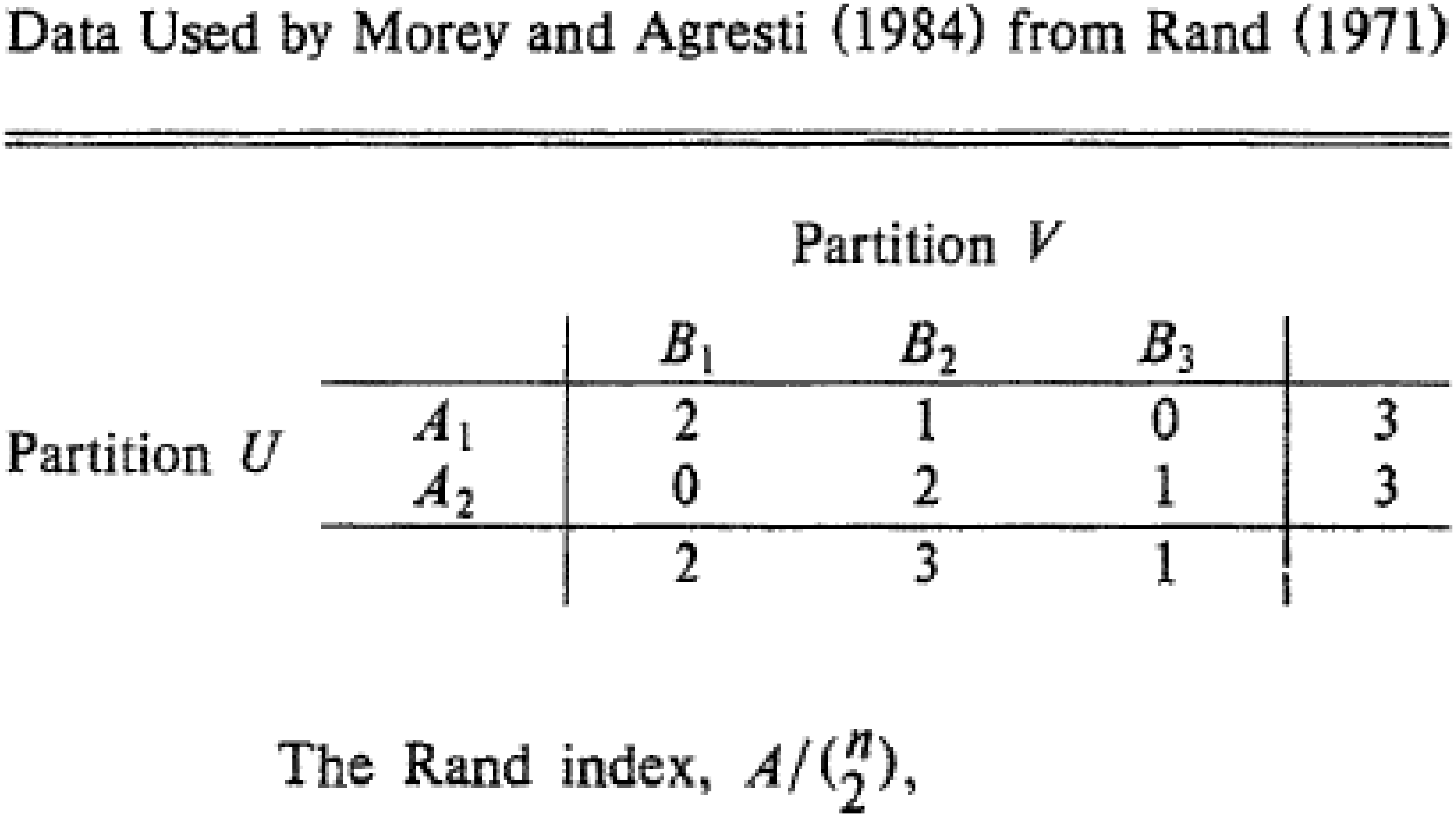
Comparing Partitions:
Hubert & Arabie's (1985) adjusted Rand index



Notation for Comparing Two Partitions

(e.g., grade changes)

		Partition V				
	Class	v_1	v_2	...	v_C	Sums
	u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1\cdot}$
	u_2	n_{21}	n_{22}		n_{2C}	$n_{2\cdot}$
	.	.	.		.	.
Partition U	.	.	.		.	.
	.	.	.		.	.
	u_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R\cdot}$
	Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

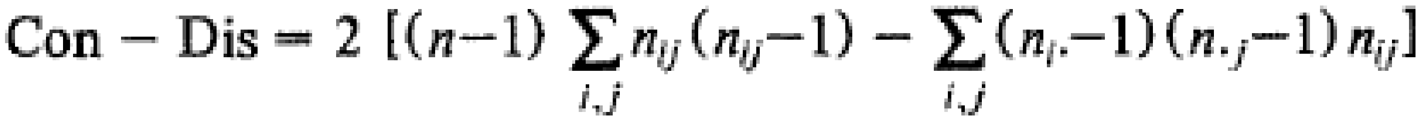


Data Used by Morey and Agresti (1984) from Rand (1971)

		Partition V			
		B_1	B_2	B_3	
Partition U	A_1	2	1	0	3
	A_2	0	2	1	3
		2	3	1	

The Rand index, $A/\binom{n}{2}$,

Hubert & Arabie (1985) measure adjusts for chance agreement:



$$\text{Con} - \text{Dis} = 2\left[(n-1)\sum_{i,j} n_{ij}(n_{ij}-1) - \sum_{i,j}(n_{i.}-1)(n_{.j}-1)\,n_{ij}\right]$$

Warrens (2008) showed that H&A’s adjusted Rand index is related to Cohen’s Kappa:

If we represent in a 2x2 table the number of agreements and disagreements between two partitions (i.e. the number of object pairs that are grouped similarly or differently in the two partitions), then the HA adjusted Rand index can be seen to be equivalent to Cohen’s *K*:

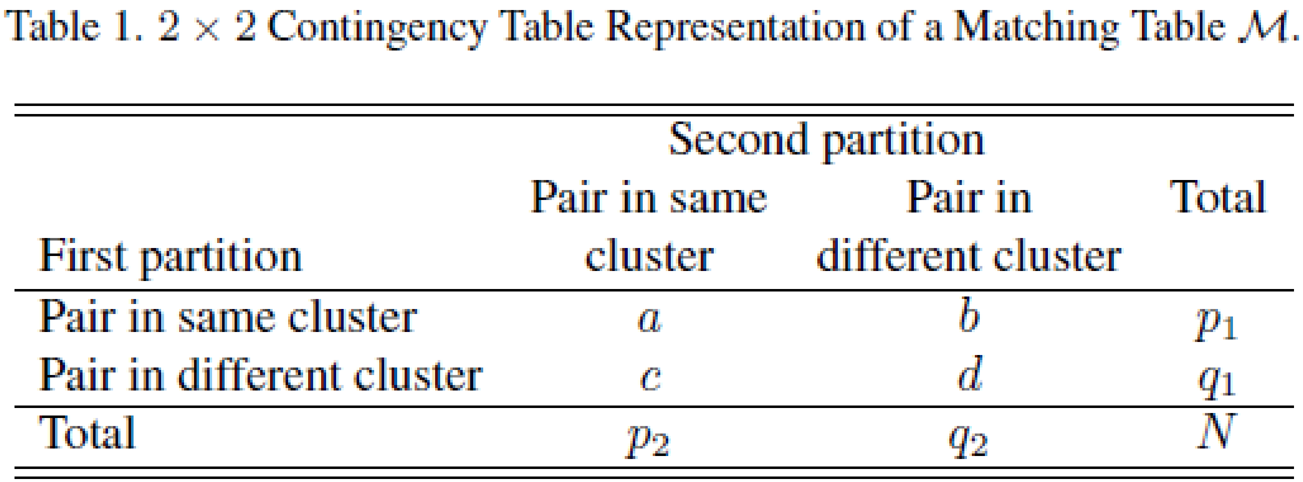


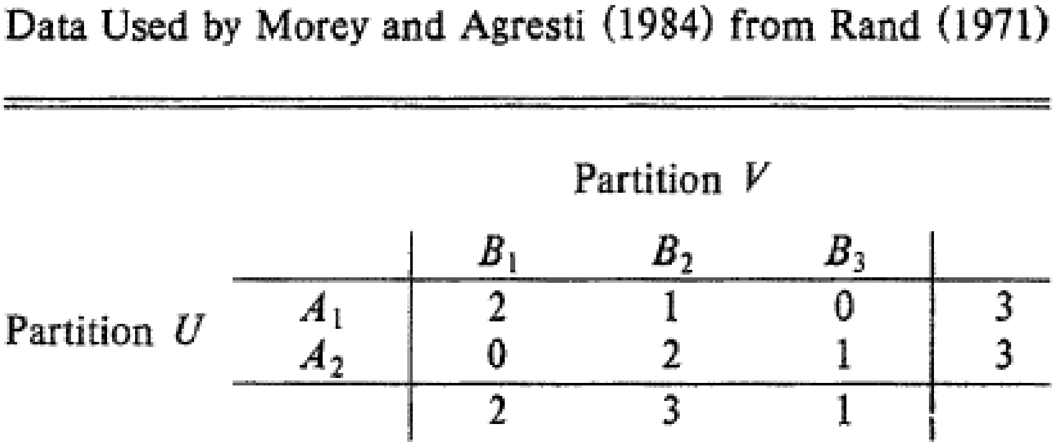
Table 1. 2 × 2 Contingency Table Representation of a Matching Table $\mathcal{M}$.

	Second partition		
First partition	Pair in same cluster	Pair in different cluster	Total
Pair in same cluster	a	b	p_1
Pair in different cluster	c	d	q_1
Total	p_2	q_2	N

HA adjusted Rand index = *K* $= \dfrac{2(ad - bc)}{p_1 q_2 + p_2 q_1}.$

where p_1=(a+b), q_1=(c+d) etc.

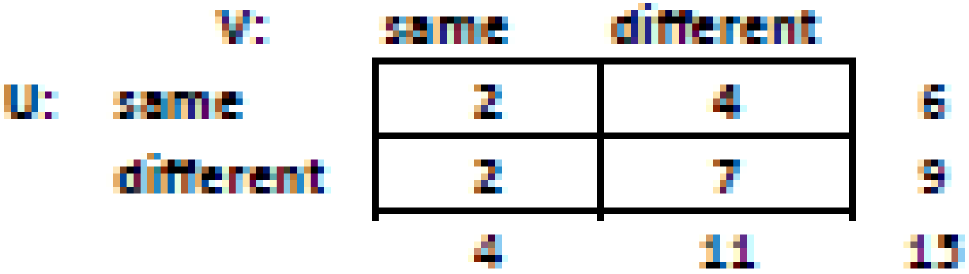
Example 1:



Data Used by Morey and Agresti (1984) from Rand (1971)

		Partition V			
		B_1	B_2	B_3	
Partition U	A_1	2	1	0	3
	A_2	0	2	1	3
		2	3	1	

Example 2: (from Warrens (2008)



	V:	same	different	
U:	same	2	4	6
	different	2	7	9
		4	11	15

HA adj. Rand = K $= \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}.$ = $\frac{2[(2)(7)-(4)(2)]}{[(6)(11)+(4)(9)]}$ = 12/ 98 = .122