Cluster Analysis

Main sources:

Aldenderfer, M.S. and **Blashfield**, R.K. 1984. *Cluster Analysis*. Beverly Hills, CA: Sage Press

Everitt, B. S. 1980. *Cluster Analysis*. Second Edition, Heinemann Educational Books. London.

Gries, S.Th. 2007. *Cluster Analysis: A practical introduction with R*. [materials from workshop at the University of Sheffield, 21 May 2007]

Kaufmann, L. and **Rousseeuw**, P.J. 1990. *Finding Groups in Data*,

New York: John Wiley & Sons, Inc.

Summer 2008 - Daniel Wiechmann

What is cluster analysis (CA)?

- CA is a generic name for wide variety of procedures

- Def.: A clustering method is…
  - a multivariate statistical procedure
  - that starts with a data set containing information about a sample of entities and
  - attempts to reorganize these entities into relatively homogeneous groups

cluster analytical approaches

- **hierarchical approaches**
  - **agglomerative**
    - start off with any many clusters as there are objects in data set
    - merge succesively into larger clusters
  - divisive
    - start off with one cluster
    - split up successively
- partitioning approaches
  - k-means / k-medoids

Why clustering?

- Classification is a fundamental process in the practice of science
- Classification (categorization) is a basic human conceptual ability

How is CA used?

- **Development of a typology**
- Investigation of useful conceptual schemes for grouping entities
- Hypothesis generation through data exploration
- Hypothesis testing, or the attempt to determine if types defined through other procedures are in fact present in the data

Where has it been applied

- Information retrieval
  - Clustering documents so that users‘ queries can be matched against cluster centroids
- Text categorization and segmentation
  - Lexical macrostructure of …
    - Texts
    - Dialects
    - Genres
- Theoretical linguistics
  - Identifying semantically similar words on the based of syntactic and/or semantic distribution
  - Word sense disambiguation
  - Evaluating experimental results
- Linguistic typologies
  - Group languages into groups/families

What does it do exactly? Conceptual issues: similarity

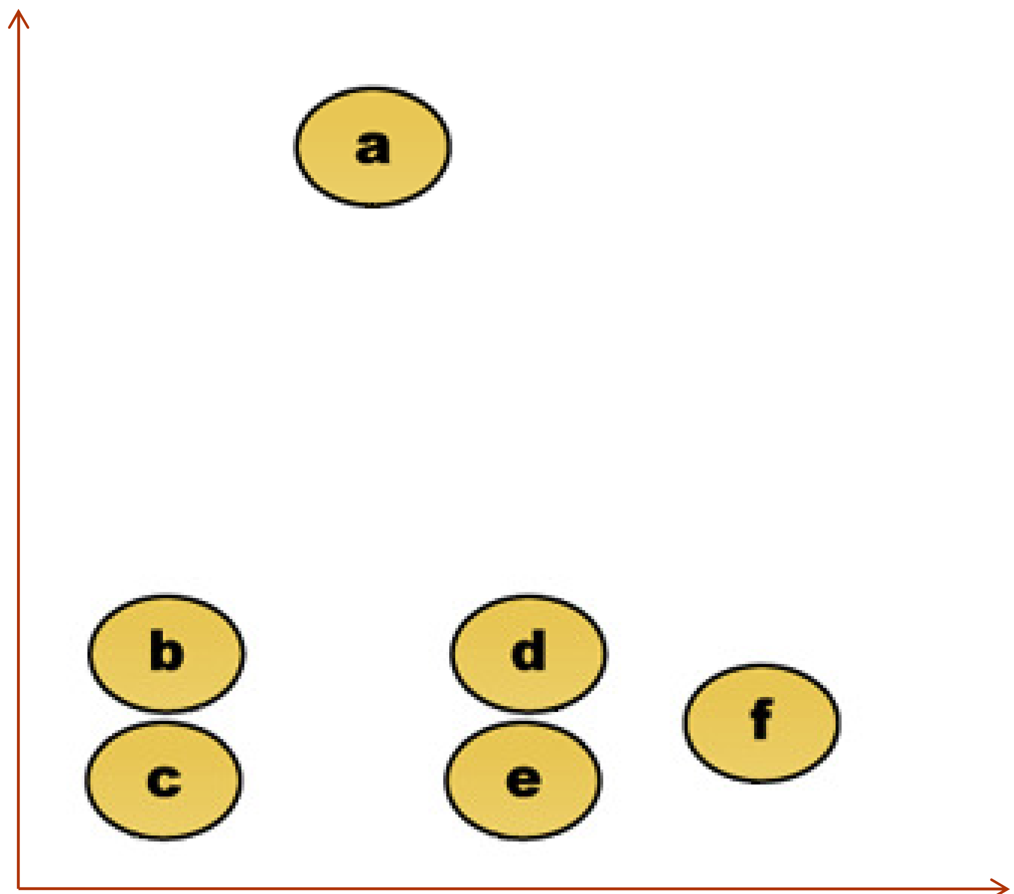
- Grouping object by (dis-)similarity
  - Maximize intra-cluster similarity
  - Maximize inter-cluster similarity
- But what exactly does it mean for two objects to be similar?

What does it do exactly? Conceptual issues: similarity

- Quantitative estimation dominated by concept of *metrics*
  - Cases are points in a coordinate space such that observerd **similarities** of the points correspond to **metric distances** between them
- Therefore, similarity is symmetric
  - $d(x,y) = d(y,x) \geq 0$
- *Philosophically* speaking, this is just one of many conceivable positions
- *Psychologically* speaking, this is controversial
  - Cf. Tversky 1977

Objects in metric space

a <- c(2,6)
b <- c(1,2)
c <- c(1,1)
d <- c(3,2)
e <- c(3,1)
f <- c(4,1.5)

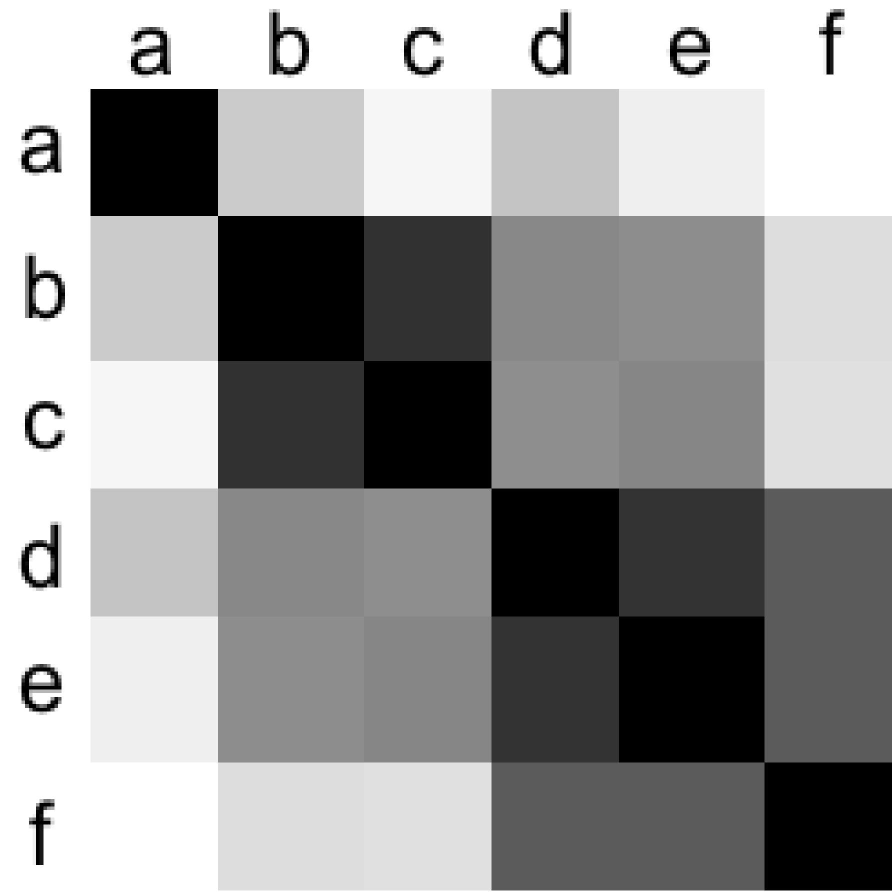


Distance matrix

	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203

Distance -> Euclidean distance
= square root of the sum of squared distances of a pair of objects
In R: sqrt(sum((a-b)^2))

Distance matrix (heat map)



…let‘s pause a second

A few precautionary generalizations…

A few precautionary generalizations…

- Most CA methods are relatively *simple procedures* that in most cases, are *not supported by an extensive body of statistical reasoning*
  - Cf. Aldenderfer & Bleshfield 1984, Jardon and Sibson 1971

- *Different methods* can and do generate *different solutions*

- Strategy of cluster analysis is *structure-seeking* although its operation is *structure-imposing*

…and a moral

1. Do not fall in love with a given clustering solution
2. Do not (blindly) trust studies that use some clustering method, but don‘t tell you why exactly that one was chosen
   1. if they do not spend much time on the justification of their choices of algorithms, chances are they are fishing in the dark
3. Do not commit the **buzzword fallacy**:
   - **data-driven**, **buttom up** methods do not necessarily constitute good science
   - …in fact the can be rather ~~stupid~~ unwise (**naive empiricism**)

Issues in clustering

Problem 1:
Choice of variables

- Most critical step in research process...
- Theory guides choice of variables (theory driven)

Problem 2:
Variable controversies

- Weighting
  - Motivated, i.e. informed by theory
    - Often missing in data-driven multivariate approaches to word meaning (semantic profiling; cf. Gries, Divjak, …)
  - Danger: Unmotivated due to correlated descriptors
    - -> Implicit weighting
    - Possible solution: Factor analysis or principle component analysis

Problem 3:
Variable controversies

- Standardization
  - yes or no? Well, it depends…
    - Standardization prevents undesired implicit weighting
    - …but maybe we do not always want to counter such effect…

Procedure

Four steps in cluster analysis

Procedure: four steps

- STEP 1: Choose **measure of (dis)similarity** to generate a (dis)similarity matrix
  - Depends on information value & nature of the variables describing the objects to be clustered
- STEP 2: Choose **amalgamation rule** to determine how elements are merged
  - Depends on the structure one suspects the objects to exhibit
  - Characteristics of almagamation rules
- STEP 3: Interpreting the results
- STEP 4: validating the results

When should I use what similarity measure? (STEP 1)

- IF object are dissimilar when the exhibit widely different values
- THEN use distance measure
  - Euclidean distance
  - Manhattan distance
  - Maximum distance
- IF objects are dissimilar when the exhibit different slopes
- THEN use correlational measures
  - Pearson
  - Spearman
  - Cosine (of angle between vectors)

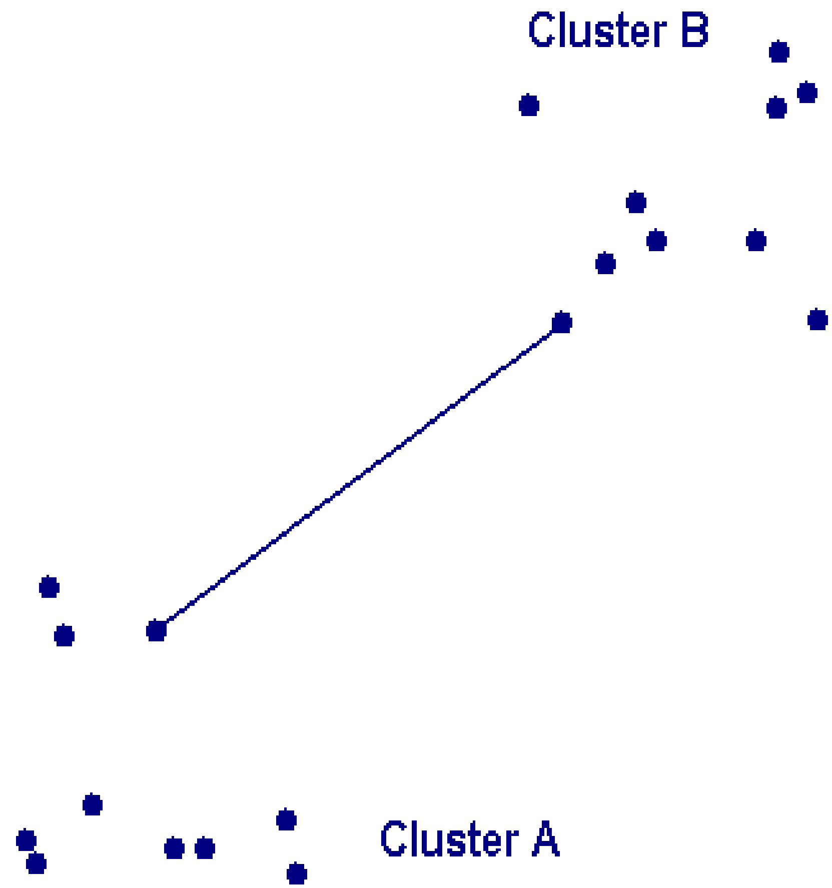
How to generate cluster structures (STEP 2)

- Single linkage
- Complete linkage
- Average linkage
- Ward‘s method

Single linkage (nearest neighbor)

Distance of two groups x,y is defined as *minimal* distance between any one element of x and any one element of y

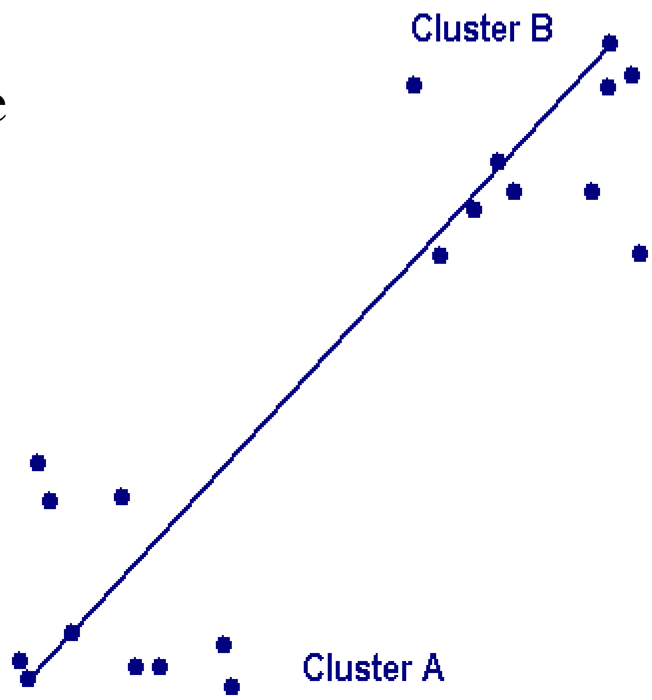
Tends to generate elongated cluster chains, can identify outliers



Complete linkage (farthest neighbor)

Distance of two groups x,y is defined as the *maximal* distance between any one element of x and any one element of y

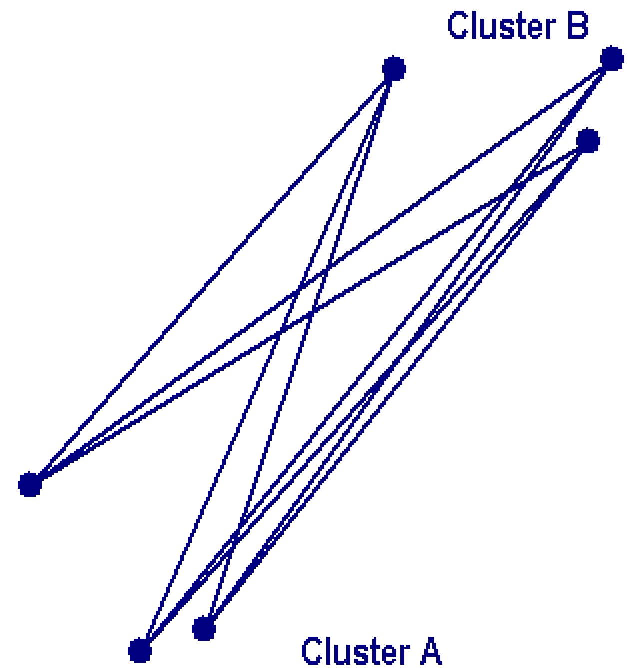
Good if data do consist of distinct clusters, produces compact clusters, problems with outliers



Average linkage

Distance of two groups x,y is defined as the *average* distance between any one element of x and any one element of y

Creates ball-shaped clusters with similar variances



Ward's method

- *Minimize information loss* associated with each grouping

- Information loss is defined in terms of error sum of squares crierion (ESS)
  - At eachstep, union of every possilble cluster pair is considered
  - merge those two elements, whose merging least increases their sums of squared difference from the mean

- Creates small and even sized clusters
- Computationally intensive

Ward's method example

- 10 objects have scores (2, 6, 5, 6, 2, 2, 2, 2, 0, 0, 0) on some particular variable.
- The loss of information that would result from treating the ten scores as one group with a mean of 2.5 is represented by ESS given by,
- ESS One group = (2 -2.5)2 + (6 -2.5)2 + ……. + (0 -2.5)2 = 50.5
- On the other hand, if the 10 objects are classified according to their scores into four sets,
- {0,0,0}, {2,2,2,2}, {5}, {6,6}
- The ESS can be evaluated as the sum of squares of four separate error sums of squares
  - ESS group1 + ESSgroup2 + ESSgroup3 + ESSgroup4 = 0.0
- **Thus, clustering the 10 scores into 4 clusters results in no loss of information.**

Application (examples)

Applications: Typology

- Altmann (1971) calculates difference for every pair of languages (using Euclidean distance)

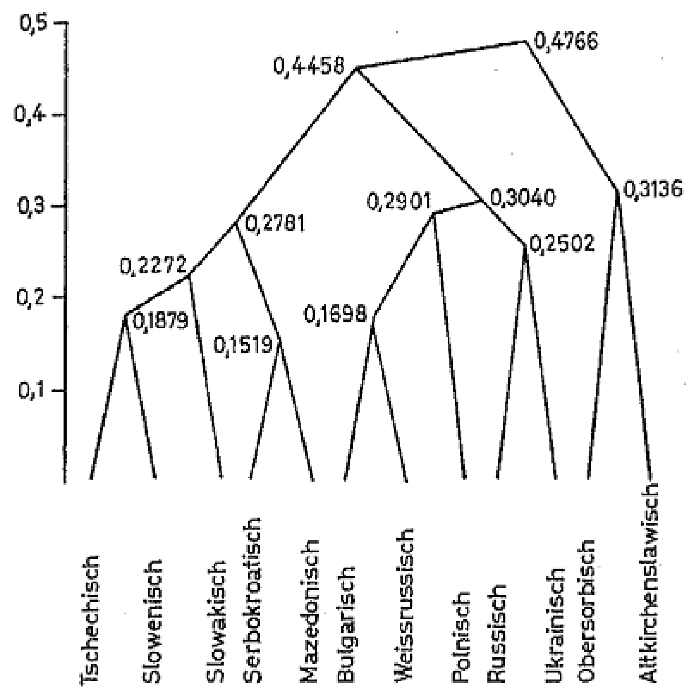


Figure 1: Hierarchical classification of Slavic phonological profiles (Altmann 1971: 19)

Applications

- Cysouw (2006) questions the adequacy of rooted trees for typological classification
- Proposes unrooted phylogenetic trees (neighbor joining algorithm instead of Euclidean distance)

Unrooted phylogenetic trees (Cysouw 2006)

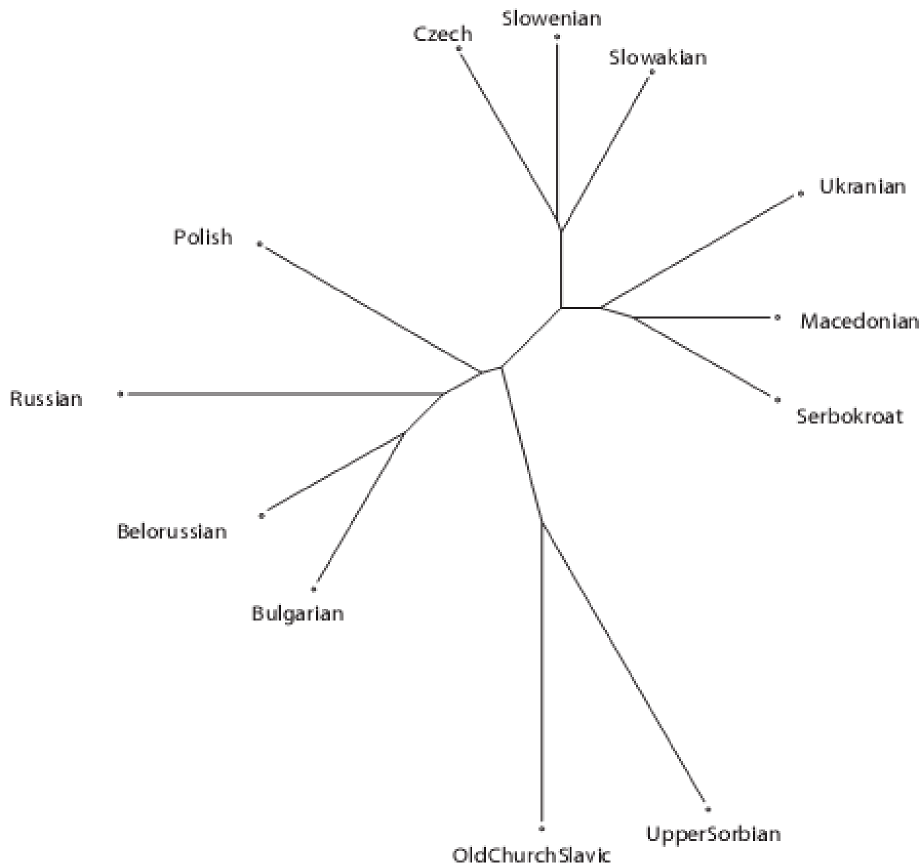
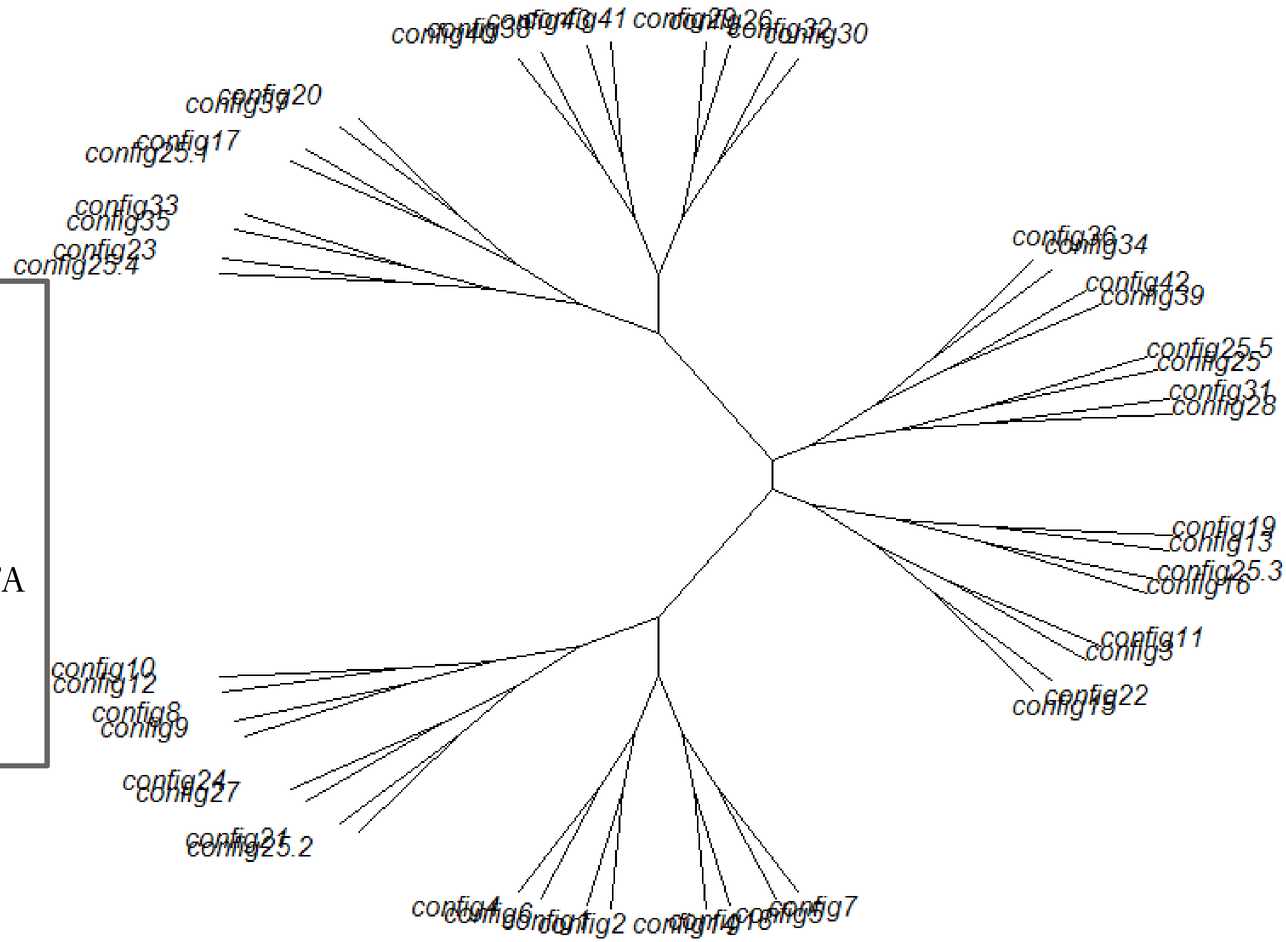


Figure 2: Unrooted tree of Slavic similarities, using the neighbour joining algorithm

Unrooted phylogenetic trees Wiechmann (in progress)

Similarity of constructions

Objects are TYPES$_{CFA}$ of relative clause construction



Do different parameters really make that much of a difference?

Wiechmann (to appear)

SCENARIO:

- We are interested in association strength (collostruction strength)
- this quantity is important for theory development
- lots of measures of that quantity have been suggested in the computational and corpus linguistic literature

QUESTION:

- How do the measures' outputs relate to each other?

TASK:

- Assess degrees of similarity the output of 47 measures of association

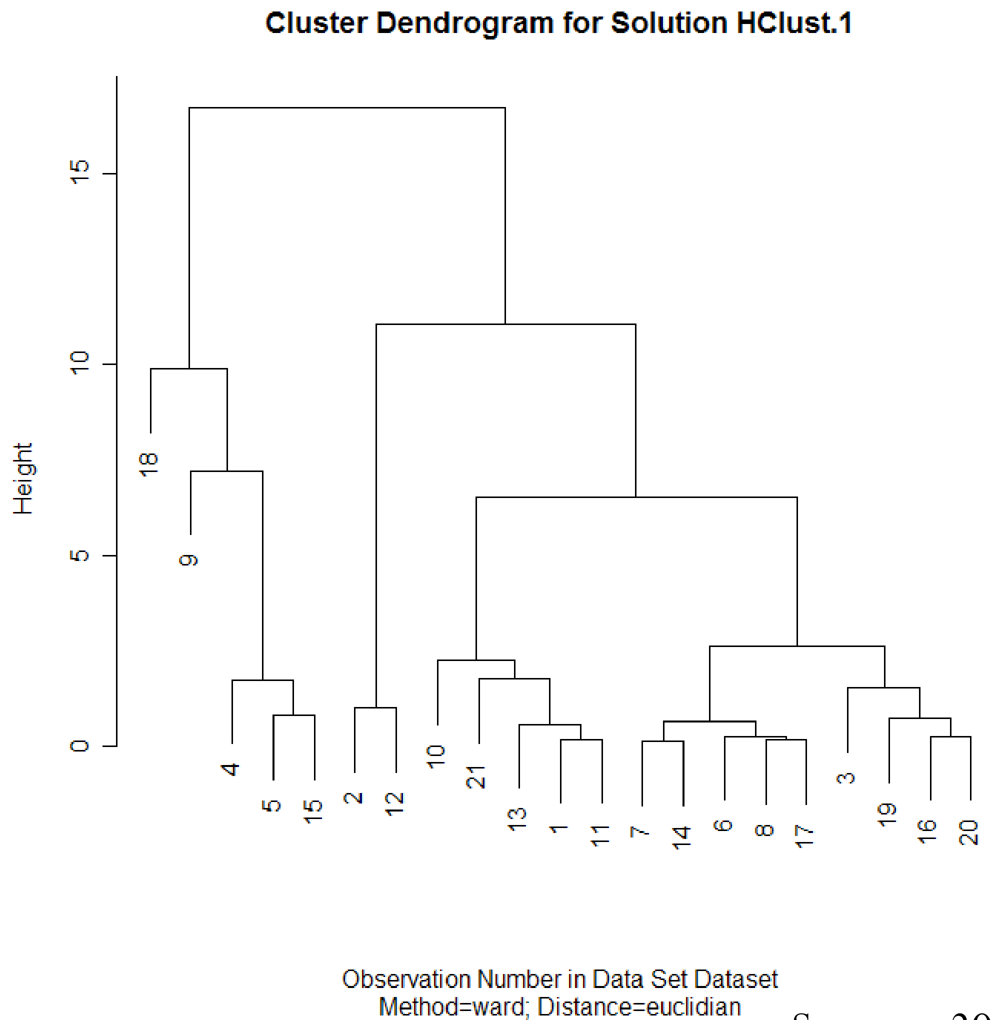
An example: comparing 47 column vectors

am.MI3	am.MS	am.Poisson.Stirling
0.749415659	0.410358961	-0.020680797
1.493657045	2.130066207	-2.147689106
-0.762153491	-0.729778419	0.493547173
-2.106709137	-1.109824235	0.778951194
-1.500297066	-1.014812772	0.83622818
0.240854228	-0.207215443	0.524033039
0.062586293	-0.349732637	0.591497289
0.35455516	-0.140707426	0.599819276
-0.544487573	-0.77728415	1.011055346
0.147044034	0.115823482	-1.048740493
0.817198929	0.467365852	0.050545439
1.823982324	2.842652109	-1.972060588
0.587632034	0.296345247	-0.22580183
0.127573567	-0.302226906	0.575871329
-1.391740027	-0.872295613	0.270828972
-0.214864909	-0.577760099	0.833772541

…and so on

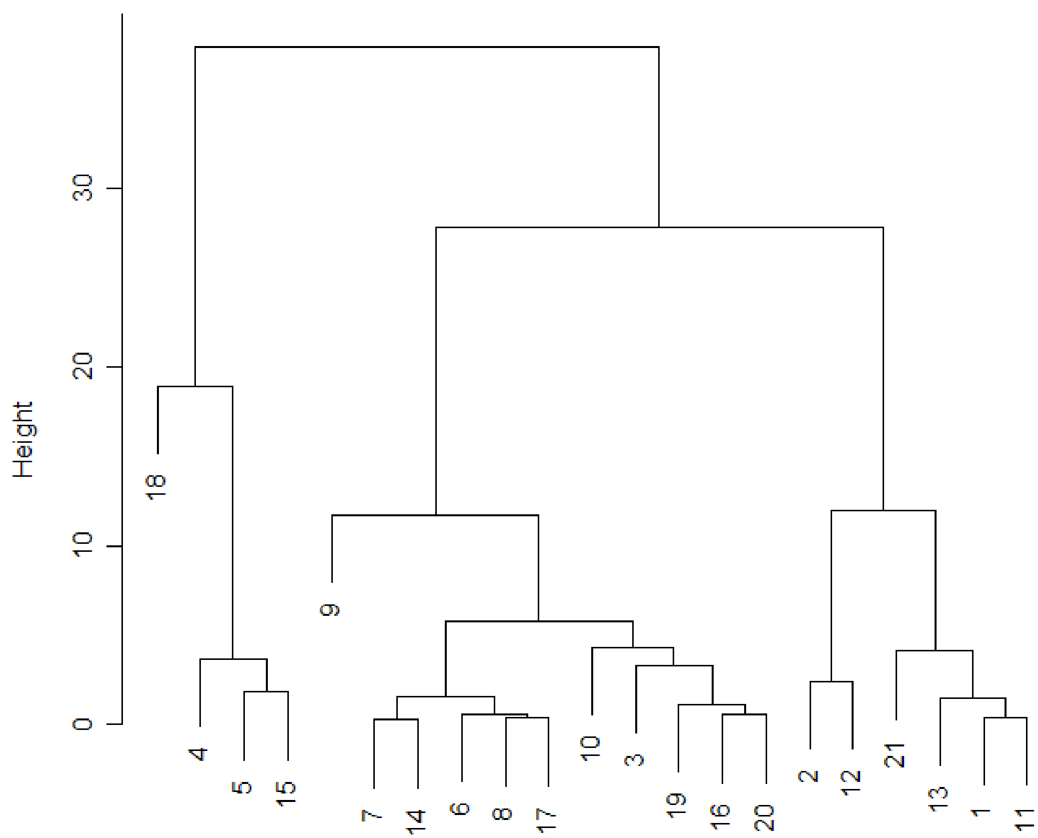
z-standardized* association scores for 21 verbs towards nominal complementation pattern
(*better always scale to avoid that VAR with greatest range dominates results)

Parameter settings and cluster solutions



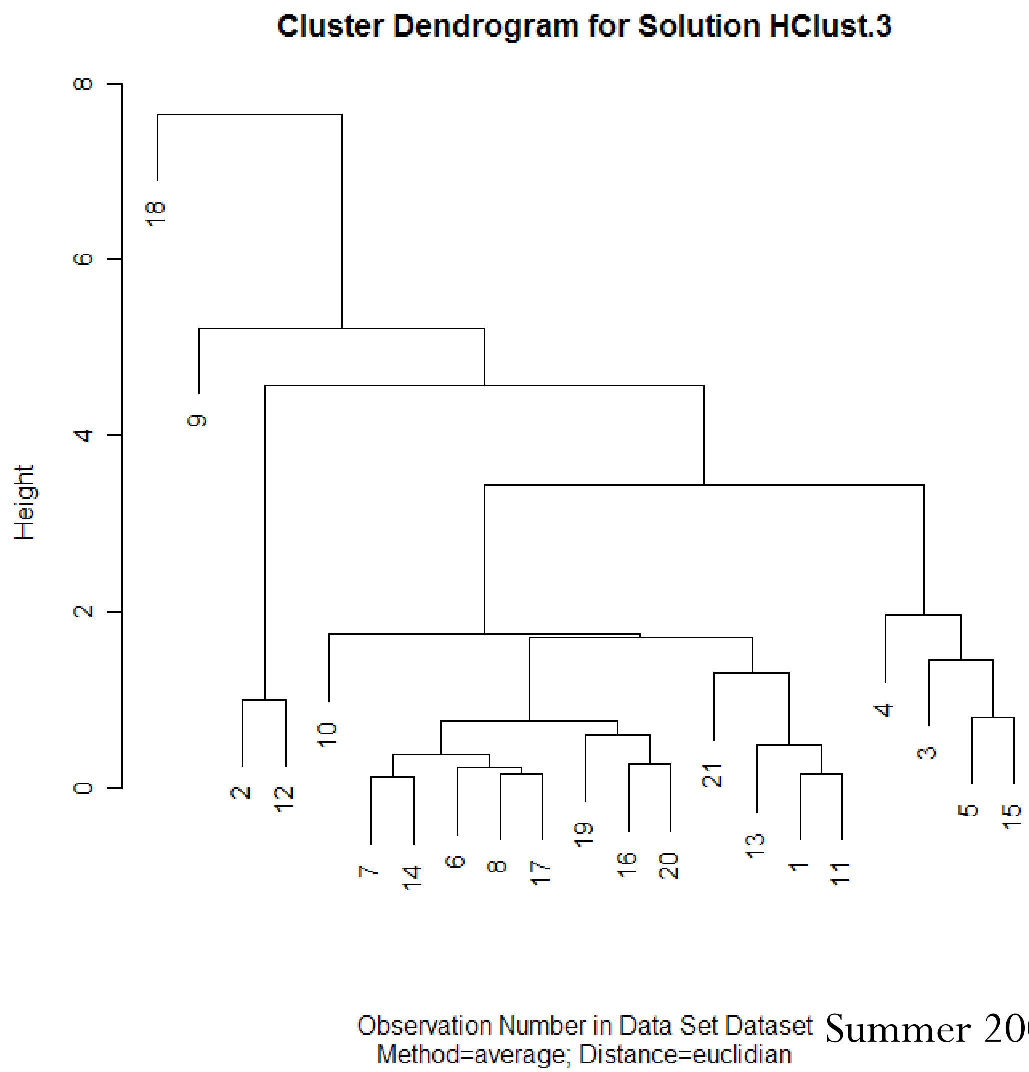
Parameter settings and cluster solutions

Cluster Dendrogram for Solution HClust.2



Observation Number in Data Set Dataset
Method=ward; Distance=city-block

Parameter settings and cluster solutions



MORAL

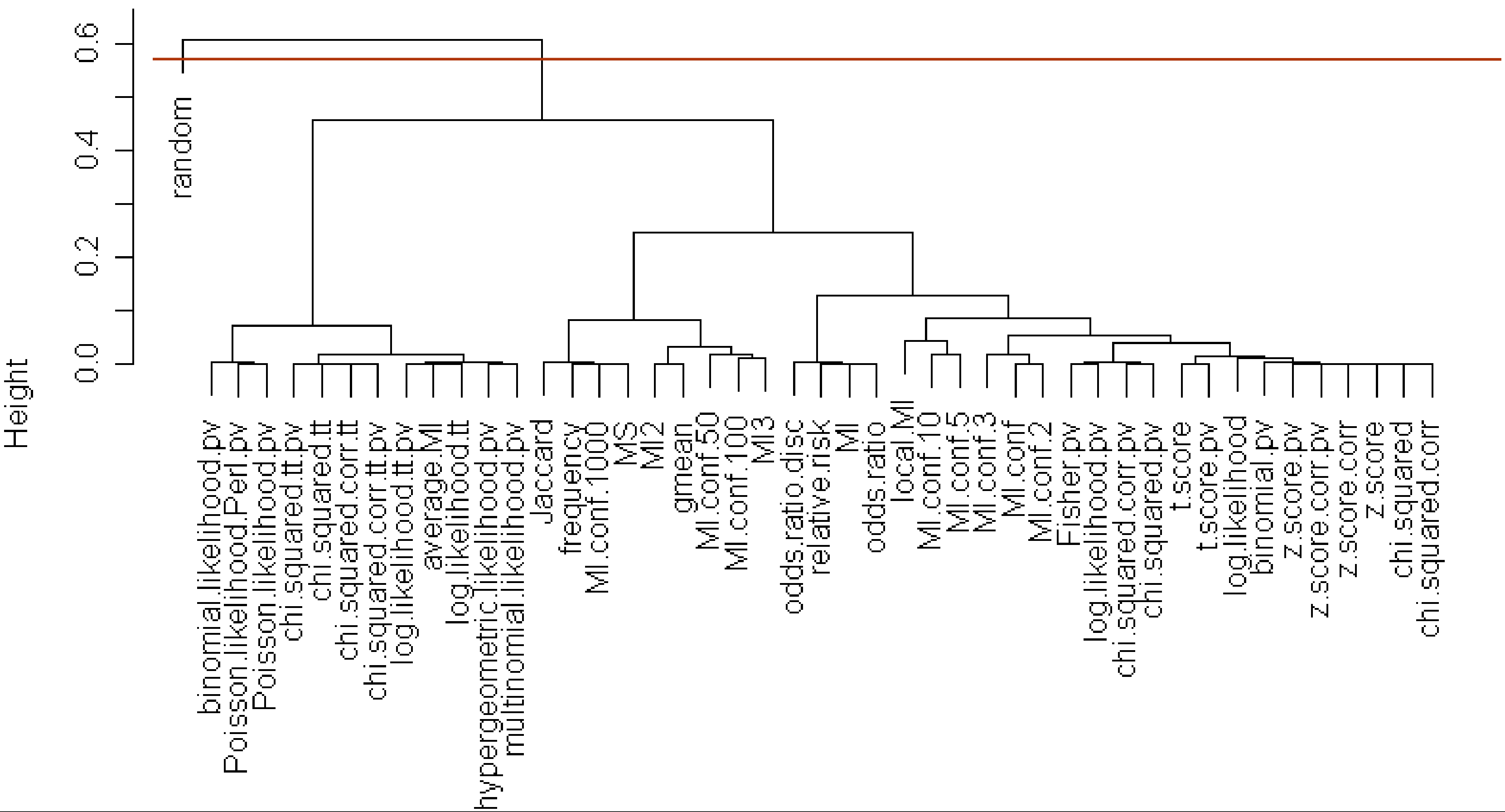
- With each setting of a parameter, we **influence** the form of the cluster solution
- We effectively determine what structure we **impose** on the data
- This is why we need to think about these things before we calculate the solution and let our **theories** guide our choices

PART II:

Interpretation and validation

Interpreting the solution

Wiechmann (to appear)
task: classify AM output



Where to cut the tree, so that the optimal number of groups is found?

Split evaluation

- Graph number of clusters implied by a tree against almagamation coefficient (e.g. Ward) & and look for flattening of curve
  - (cf. scree test for factor analysis)
- '**average silhouette width**' (cf. Roossseuw 1987; Kaufman & Roosseeuw 1990: Chapter 2)

'average silhouette width'

- ASW coefficient assesses the *optimal ratio* of the intra-cluster dissimilarity of the objects within their clusters and the dissimilarity between elements of objects between clusters

Inter-clusters distance ⇒ maximized
Intra-clusters distance ⇒ minimized

Silhouette width (SW)

- SW is a way to assess strength of clusters
  - SW of a point measures how well the individual was clustered
- $SW_i = (b_i - a_i) / \max(a_i, b_i)$
  - Where a_i is the **average disstance** from **point i** to **all other points in i's cluster**, and $\mathbf{b_i}$ is is the **minimum average distance** from **point i** to **all points in another cluster**
  - $-1 < SW_i < 1$

Average Silhouette Width (ASW)

- ASW measures the global goodness of clustering
  - $ASW = (\sum_i SW_i) / n$
  - $0 < ASW < 1$
  - The larger ASW the better the split

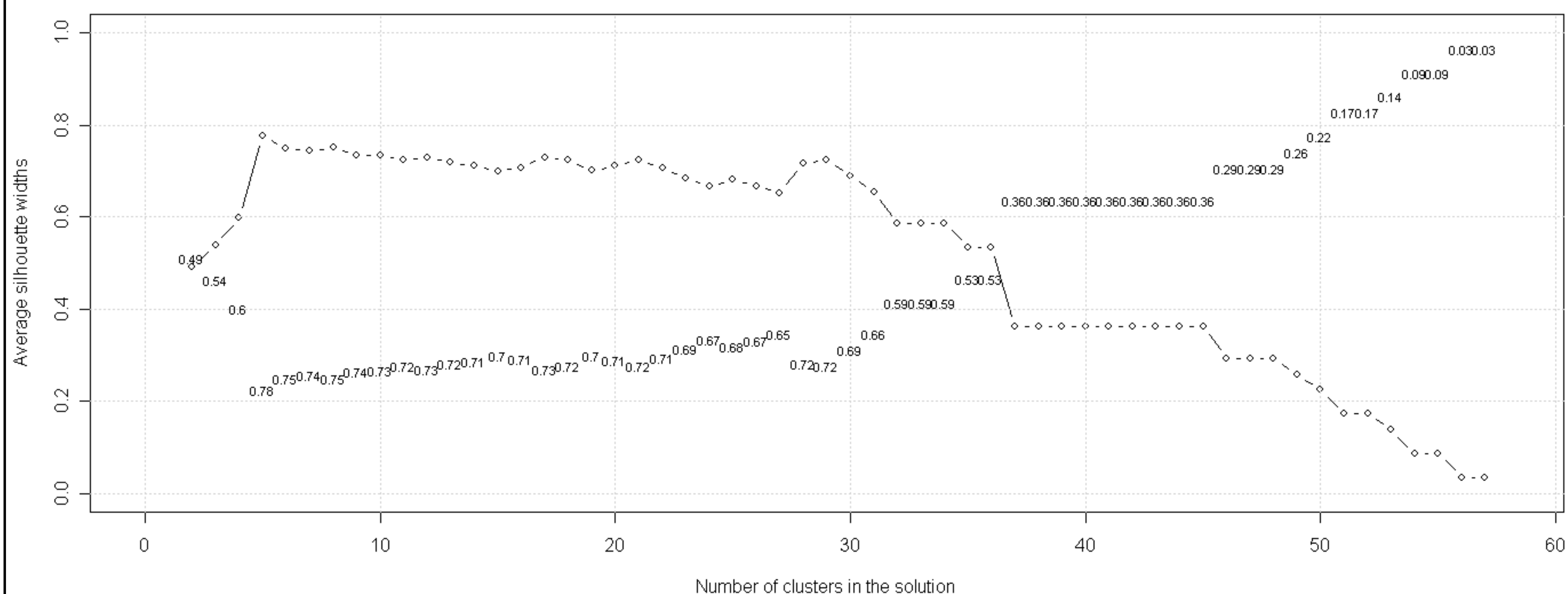
Average Silhouette Width (ASW) Interpretation

I	0.71 – 1.00	A strong structure has been found (excellent split)
II	0.51 - 0.70	A reasonable structure has been found
III	0.26 - 0.50	The structure is weak and could be artificial
IV	$\leq$ 0.25	No substantial structure has been found (horrible split)

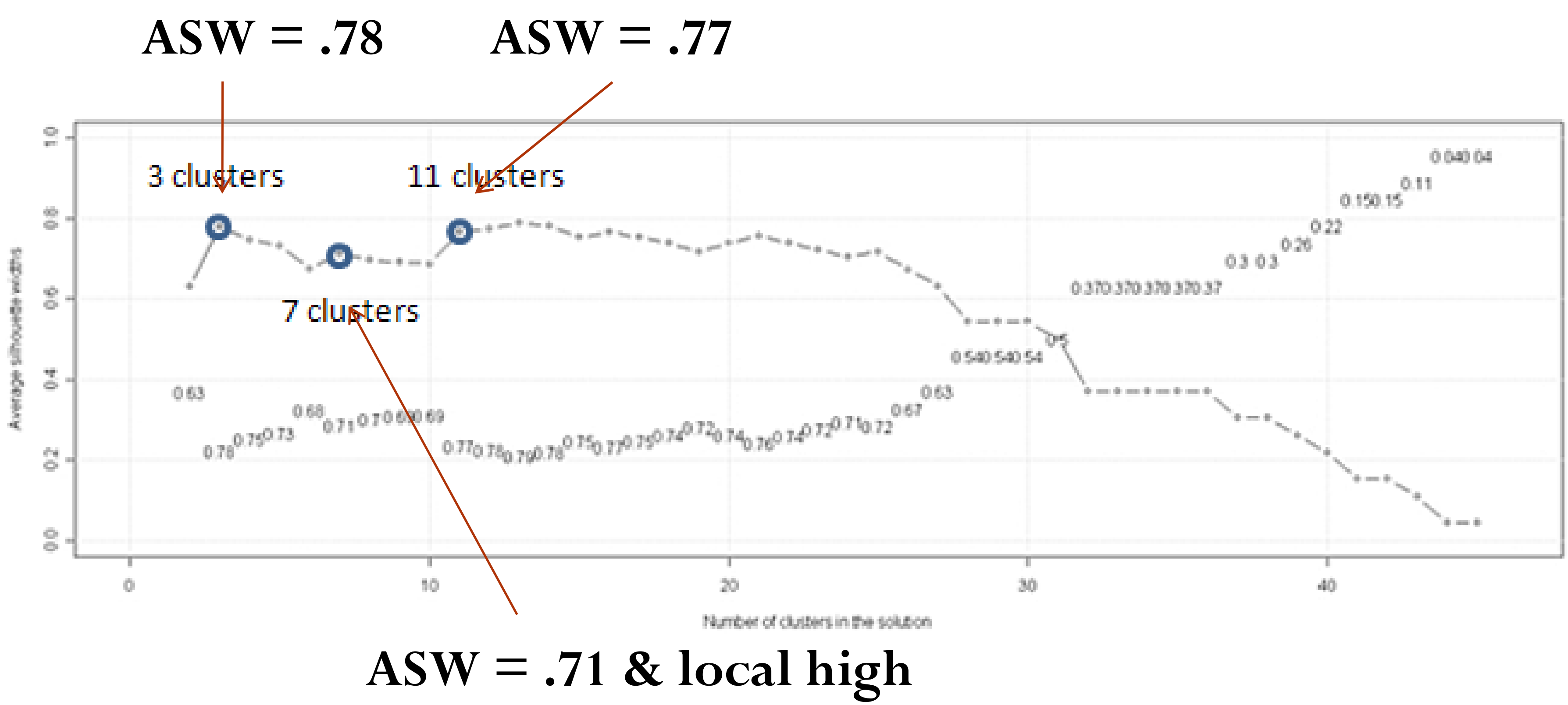
Computing ASW

- for all partitioning solutions
  - beginning with the minimal one that consists of just two groups
  - to the most detailed one, which consists of $N_{objects} - 1$
    - here 48 – 1 = 47
- Compare ASW
  - Look for **highest values**
  - Look for **local highs**

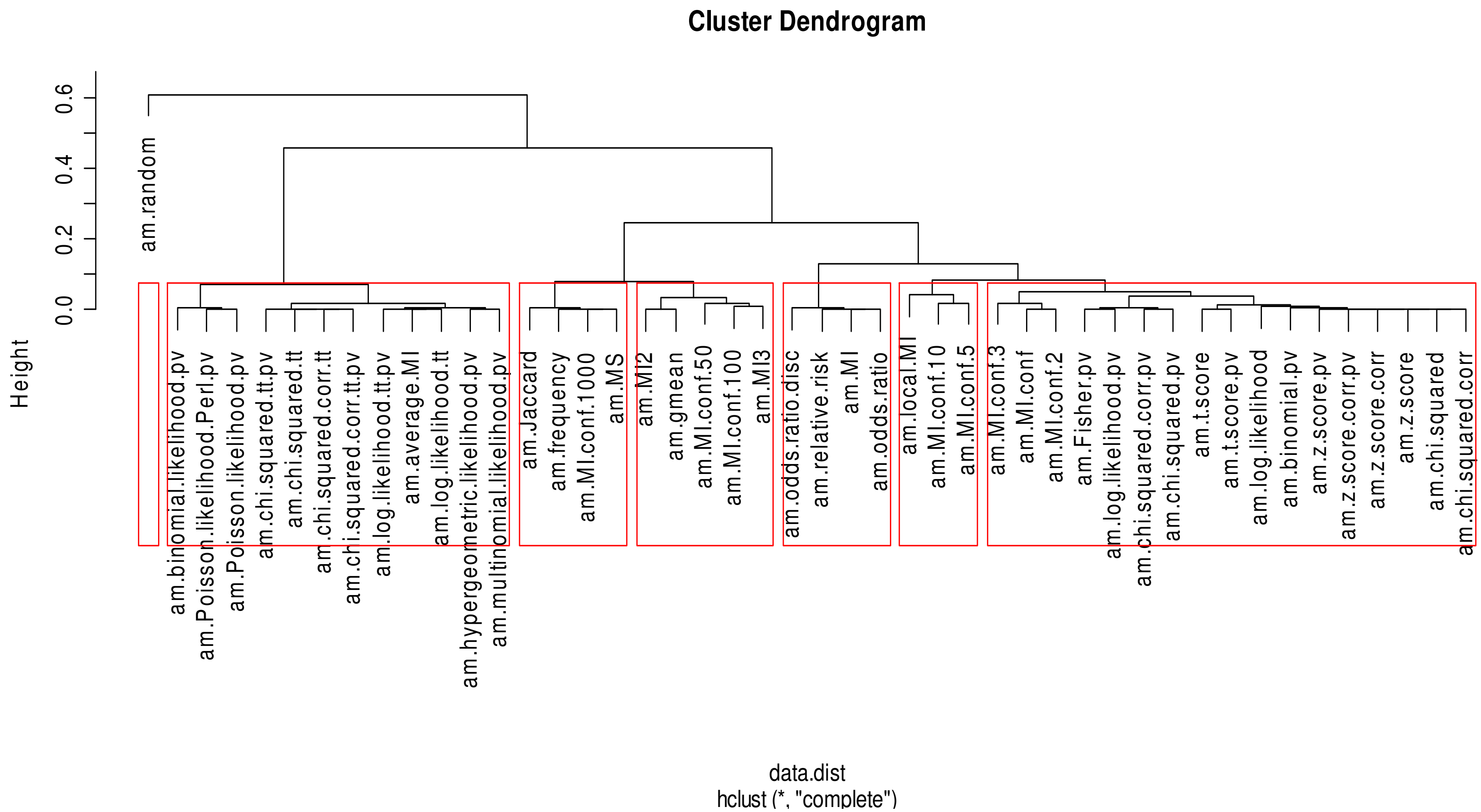
Cluster solutions by average silhouette width



Cluster validation graph



7 cluster solution



Validation techniques

Validation techniques

1. Cophenetic correlation
2. Significance tests on vaiables used to create clusters
3. Significance test on independent variables
4. **Monte Carlo**
5. **Replication**

cf,. Aldenderfer & Blashfield 1984 for a discussion of these techniques

Monte carlo procedures

- Uses random number generators to generate data sets with general characteristics matching the overall characteristics of original data
- Same clustering methods are applied
- Results are compared

Replication

- Split up your data set into random subsamples and apply the same methodologies
- Checks internal consistency of a solution
  - If a cluster solution is repeatedly discovered across different sample from the same population, then it is plausible to conclude that this solution has some generality
- Replicability is necessary but not sufficient
  - Failure of replication -> bad solution
  - Successful replication -> chances are it is a good solution

Practical issues in clustering

Cluster analysis and scales of measurement

Dissimilarity and scales of measurement

- Interval (we have talked about this case already)
- **Binary**
- Nominal
- Ordinal
- Ratio -> **counts**

- Mix types

Interval-valued variables

- <u>similarity</u> is expressed as distance between objects
- *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance
- If q = 2, d is Euclidean distance

Practical issues:
(Dis)similarity measures and scales of measurement

- **Binary data**
  - object_1 = c(1,1,1,0,0,1,0,1,1,0)
  - object_2= c(0,1,0,0,0,1,0,1,1,1)
  - object_3 = c(1,0,1,0,0,1,0,1,1,0)
  - …

Practical issues:

(Dis)similarity measures and scales of measurement

Binary variable

	Object_2: F is present	Object_2: F is absent	
Object_1: F is present	a	b	a+b
Object_1: F is absent	c	d	c+d
	a+c	b+d	m

Practical issues:

(Dis)similarity measures and scales of measurement

- **Similarity of two objects:**

  (parameters for w_1 and w_2 dependent on sim_coef choice)

  $$\mathbf{a + (w_1 * d) / (a + (w_1 * d)) + (w_2\ (b+c))}$$

- IF presence or absence of variable level have same information value (= symmetric, i.e. $d(i,j) = \frac{b+c}{a+b+c+d}$, e.g. animacy),

  THEN use ***simple matching***
  ($w_1 = 1; w_2 = 1$)

- Otherwise, (asymmetric, $d(i,j) = \frac{b+c}{a+b+c}$, use either *Jaccard* or *Dice*

Practical issues:
(Dis)similarity measures and scales of measurement

- **Nominal variables**
  - Well, they can be handled by generalizing over what we just said about binary variables
  - Recode VAR as dummies and proceed as just described

Practical issues:

(Dis)similarity measures and scales of measurement

- **Ordinal variables**
  - can be treated like interval-scaled variables
  - Replace x by their rank
  - Recode VAR as dummies and proceed as just described

Practical issues:
(Dis)similarity measures and scales of measurement

- **Ratio-scaled**
  - averages
  - lengths
  - **counts**
    - object_1 = c(10,12,123,60,70,11,50,31,11,10)
    - object_2 = c(1,15,130,62,75,21,40,24,11,18)
    - …

Practical issues:
(Dis)similarity measures and scales of measurement

- For mixed variables…
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio
- …we may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - f is **binary** or **nominal**:
    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$
  - f is **interval-based**: use the normalized distance
  - f is **ordinal** or **ratio-scaled**
    - compute ranks r_{if} and
    - and treat z_{if} as interval-scaled $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

How to do all this...with SPSS

- Try this:
  - create a (fictive) data set in Excel
    - =RANDBETWEEN(1,100) # random number between 1 and 100
  - import this set to your favorite stat soft
  - In SPSS: Classify -> Hierarchical Cluster… ->
    - Choose variables
      - Tick:
        - o Cluster: cases
        - o Display: statistics & plots
        - o Statistics -> (Agglomeration schedule) & proximity matrix
        - o Plots -> Dendrogram
        - o Method -> some cluster method & counts -> Chi squared
      - You should get something like this SPSS_demo_out

How to do all this...with R

- R is - of course - way more powerful
  - more algorithms
  - new techniques get implemented as they are developed
  - R graphics are much more versatile and look way cooler ;)
- this is what you get if you search for >>cluster<<

Fuzzy search