Notes and Comments on HICLUS

APMC 9 May 2001

HICLUS (Hierarchical Clustering Schemes)

provides analysis of:

- **two-way, one-mode (dis) similarity *data***
- **by means of an ultrametric distance *mode*l (hierarchical clustering) scheme**
- **using a monotonic *transformation* of the data.**

---

1. HICLUS is a Bell Laboratories Program implementing Johnson's (1967) Hierarchical Clustering Schemes (HCS).

Strictly, HICLUS is not a dimensional program, but an agglomerative clustering model, giving rise to an set of nested (hierarchical) clusterings, from the "splitter" clustering where each object is a cluster to the "lumper" cluster, where all objects are in the one cluster. Each level is a finer clustering than the level above it. Only if the data perfectly obey the "ultrametric inequality" is there one single solution. In the usual non-perfect case, ambiguities arise in defining "the" distance between a point and an existing cluster, and Johnson's HICLUS provides the two extreme possibilities: taking the the maximum (aka diameter, complete link) and minimum (aka connectedness, single-link) options. (An "average" solution can be obtained by using a centroId or median-link method (see refs)

Hierarchical Clustering Schemes (HCS) are almost always used in parallel with MINISSA, and the HCS level clusterings are embedded in the 2-dimensional space as iso-contours (see 4.3.3.1 in the documentation)

2. **MDSX DOCUMENTATION:**
   MDS(X) Users Manual, Edinburgh 1981, ch. 3 *(HICLUS_TUM.pdf)*
   The User' Guide to MDS, Heinemann 1982 , 4.3.3.1 and 6.1.6
   (*HCS4331_TUM.pdf; HCS616_TUM.pdf*)
3. MDSX DATA:
   TEST INPUT: (*TESTHCS_INPUT.txt*)
   Co-occurrence data derived from 68 individuals' free-sorting of 28 drugs
   TESTOUTPUT: (*TESTHCS_OUT.txt)*
4. COMMENTS:

Hierarchical Clustering Schemes (HCS) are almost always used in parallel with MINISSA, and

the HCS level clusterings are embedded (by hand) in the 2-dimensional space as contour levels (see 4.3.3.1 in the documentation)

5. HINTS:

Very extensively used program. The connectedness method is known to tend to produce the chaining effect (successive addition of single objects to a group) and users therefore tend to prefer the diameter method. The "graphical" output – designed for the line-printer – can easily be inverted ("lumper" on the top of the diagram) or re-drawn as a tree. Note that there is no measure of fit. One suggestion for a goodness-of-fit measure is to create a matrix of lowest upper bounds (i.e. the level at which two points are joined in a cluster), and rank-correlate those matrix entries with the order of the original data.

6. REFERENCES

BASIC REFERENCE:

Johnson,S.C. 1967, Hierarchical Clustering Schemes, Psychometrika, 2, 241-254.

Other:

Borgatti, S.P. (1994) How to ExplainHierarchical Clustering, Connections 17(2):78-80

Murtagh,F (1983) . A survey of recent advances in hierarchical clustering algorithms. The Computer Journal, 26:354--359, 1983.

7. STATUS

The algorithm is stable , reliable and fast. The SHADE option depends on local implementation and should not be relied upon. Usage: High