STAT 802 Multivariate Analysis November 13, 2003

Cluster Analysis

Intro:

An exploratory technique:

Note: Univariate cluster analysis has no "standard" methods – there is some theory for mixtures of distributions of like shape, but even that is incomplete. (e.g. Usually require specification of number of clusters sought).

Lots of room for creativity and making use of data context to improve methods.

Some notes from Ch 12:

p. 668 – no a priori groups – aim is to find homogeneous subsets – evidence of separate subpopulations? Note usual overlap of subpopulations and the indirect way we have to detect these. Problem even for unimodal data.

Can cluster variables as well as iems

Similarity between items or between variables – like inverse distance. But not all similarities can be converted into a "proper" distance. (p 37)

p 669 - Evaluating all possible subgroupings possible only in theory, not in practice. See calculation of number of subgroups bottom p 669.

p 670 Various point-to-point metrics affect how clusters are detected.

p 671 Categoirize pairwise distances into a few ranges – sort to give pic like p 673

p 673-679 Dealing with character (or binary categorical) data. Measures of distance vary. When clustering variables, connection with correlation and chi-square p 677

Hierarchical:

p 680 Agglomerative vs divisive: check that nearest neighbour agglomerative technique cannot be reversed to give the same tree. Consider how to divide????

cluster-to-cluster distances – several strategies – single linkage, complete linkage, avg linkage – lead to different shaped clusters – 2-D paradigm useful for understanding. Join distance = inverse of "strength of clustering" - trees p 690

p 693 - Stability – bootstrap? – or jitter data.
(Note: Take a look at the online bootstrap notes for Nov 7 and 10 in STAT 400).

Ward'd Method; Note that sum of ESS is a measure of lack of information – the method tries to add points such that the ESS increases as little as possible.

Non-hierarchical:

p 694 ff k – means

start with k means, and assign points nearest. Then recomputed means, and repeat.

Note dependence on starting means.

Note also that if every re-assignment led to a recalculation of the centroids (not the usual way – usually complete one pass before recalculation) then order of consideration could be an additional source of variation of outcome.

Multidimensional scaling:

An "ordination" technique:

Try to reproduce in a few (usually 2) dimensions the distances in p dimensions. "Stress" is a measure of how much info is lost in this. Useful example for understanding – geographic pairwise distances. p 704.

Metric and non-metric. Metric same as pple components.

Correspondence Analysis:

Plot rows and columns on same plot. Shows how subgroups of individuals can be defined by certain variables, and how subgroups of variables can be defined by certain individuals. (see para 2 p 709). Seems most useful for contingency tables. See also para in middle of p 712 for a definition of the correspondence matrix P.

Depends on singular value decomposition of P-rc', where r and c are row and column total vectors. P 714. See also direct approach to biplot from SVD on p 721. This shows how the biplot is related to the eigenanalysis of S.