STAT 802 Applied Multivariate Analysis Oct. 31, 2005

Today: Remarks on Factor Analysis Assignment
Overview of Ch 12: Clustering, Distance Methods, and Ordination

Remarks on Assignment:

It was a badly designed assignment! However, many students could have done better on it, if they had remembered that the reason we used simulated data was that we could judge results against the known correct model. Many students used the comparison criteria appropriate to assessment of the techniques based on real data, but this was not appropriate in this case. For example, having large communalities is not good if they are generated by fictitious factors! Small specific variances are not good if they are generated by error variables. Reproducing the correlation matrix is not the same as determining the model correctly.

The modal mark was 7/10. That is because there were a couple of students who did the correct comparison, and they needed to be rewarded, However, I do appreciate that everyone spent considerable time on the exercise and hopefully learned something from it.

Overview of Ch 12: Clustering, Distance Methods, and Ordination

12.1 Card Example is just to make the point that the identification of clusters will depend on how similarity of items is defined.

One other point is that to try all possible arrangements into k subsets while computing some measure of goodness-of-clustering for each arrangement, is not feasible even for modern computers. See the footnote on p 669 for details.

12.2 Similarity can be defined in terms of a distance Similarity=(1/(1+d)) for example.

Euclidean distance usually used for objects, correlation describes similarity of variables (would you use r or |r|?

Other metrics are possible (pp 670-671)

Binary variables can be handled (0-1 variables) by counting matches for example. See other posibilities Table 12.2 p 674.

Note connection of binary table p 677 and chi sq intedependence statistic

Ad hoc methods as Example 12.3 re languages.

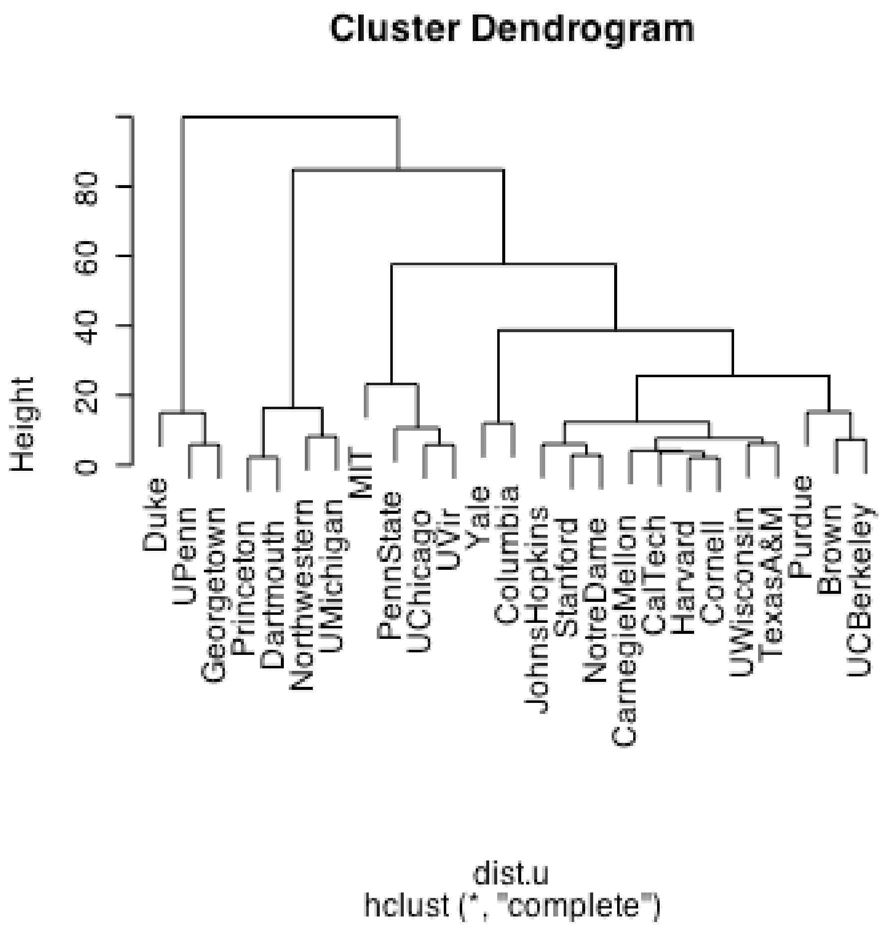
12.3 Hierarchical Clustering (hclust command in R)

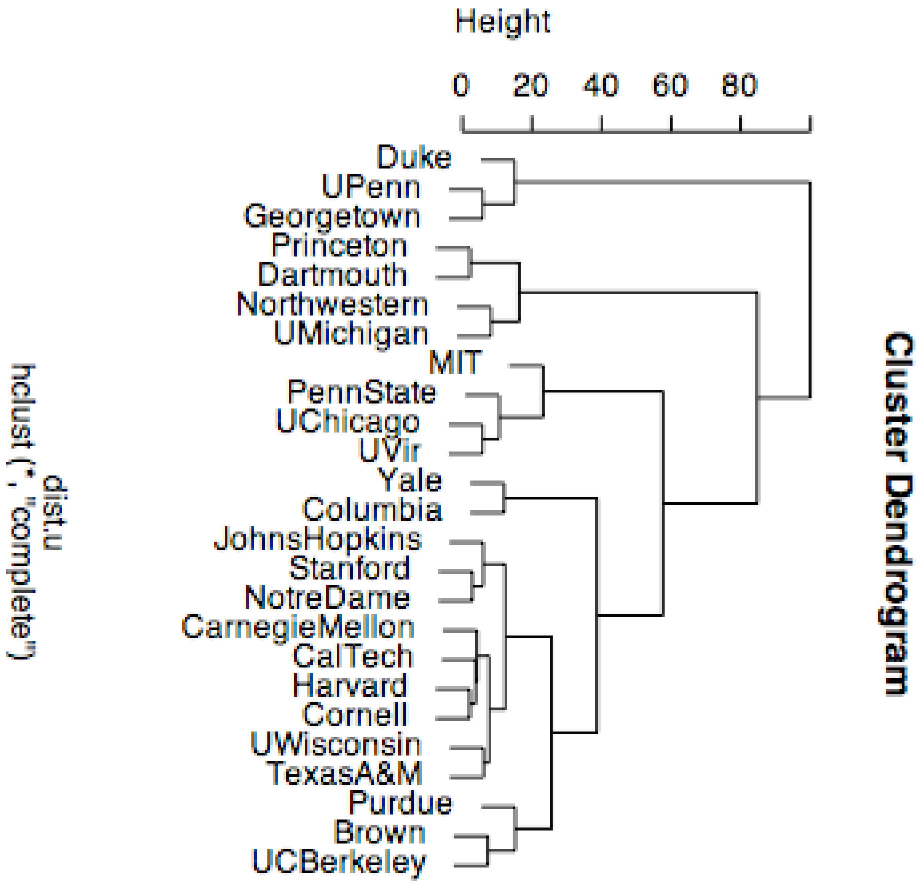
start every pt a cluster, merge 2 closest, then recompute cluster to cluster distances, and repeat. eventually get all points in one cluster. Need information about how "closest" distance jumps at merge.

More than one way to use pt-wise paired distances into a way to measure cluster-to-cluster distances. See Fig 12.3 p 680. Implications p 684. Language example changes slightly with different definition.

Can test stability of cluster result by perturbations, or by bootstrap.

```
>dist.u=dist(T12.9.df)       # creates Euclidean distances from data matrix
>clust.out=hclust(dist.u)   # does the cluster analysis with default method "complete"
> plot(clust.out)              # makes output tree.
```





12.4 Non hierarchical Clustering (kmeans command in R)

Choose typical cases and gather others around. K-means method – assign pts to nearest centroid, then re-compute centroid.

12.5 Multidimensional Scaling (cmdscale in R)

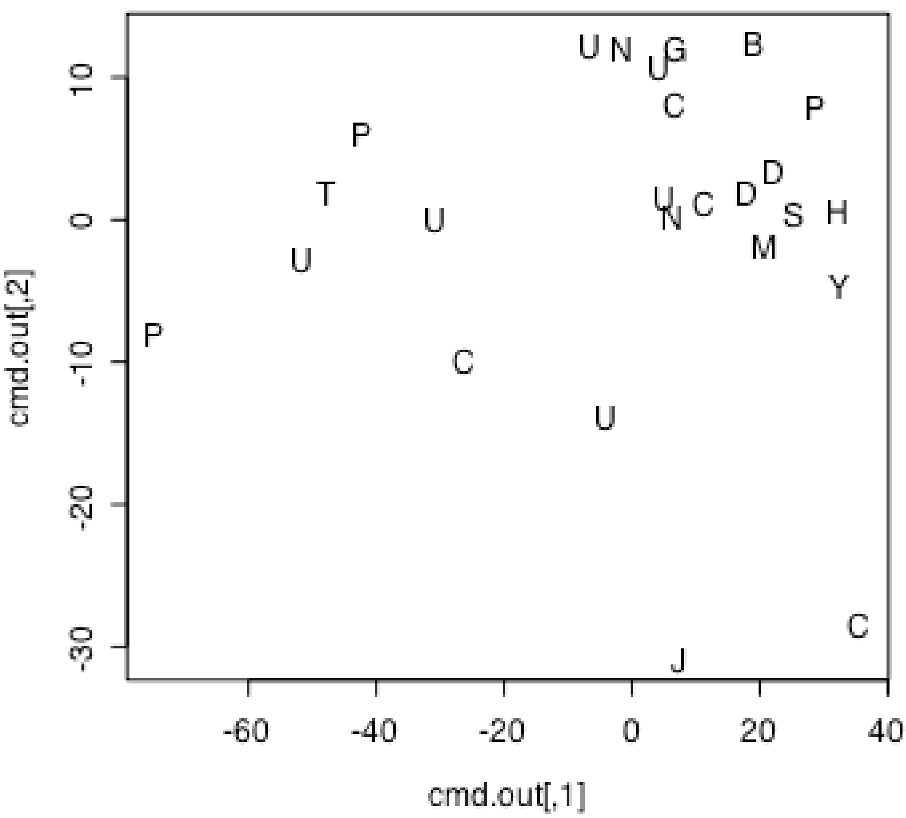
pairwise-distances -> usually 2 dimensional representation
Sometimes only have ranks of distances (non-metric MDS) (isoMDS in R)

want resulting distances as near as possible to original distances. stress reduction. p 701 bottom. Based on ordered similarities of pairs.

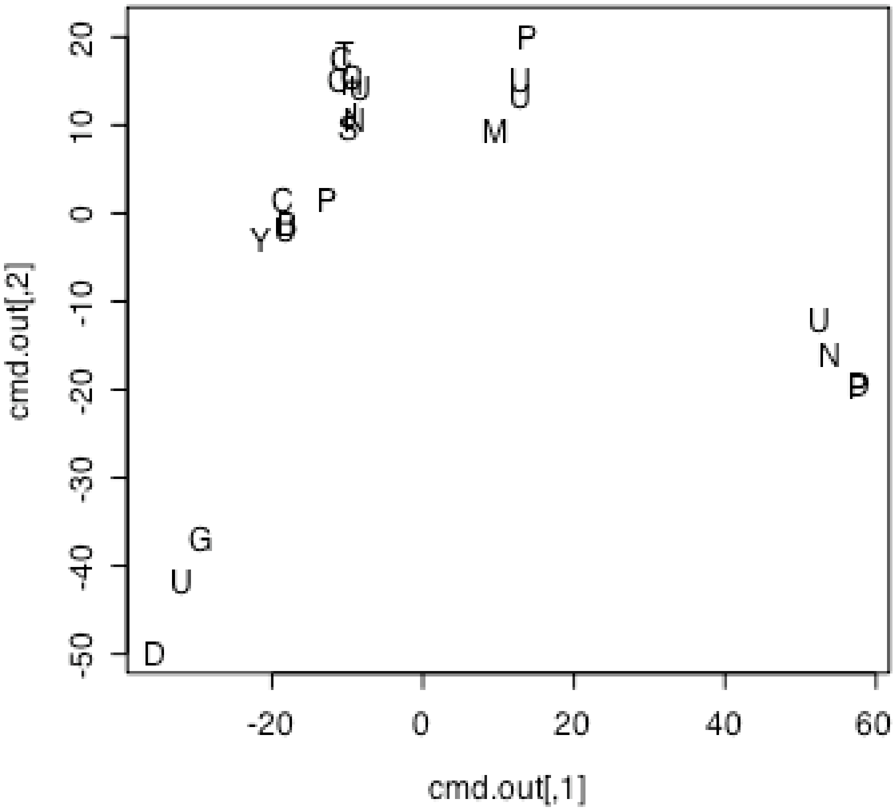
US cities example p 703-705

US universiyies example p 706-708 metric and non-metric options. (Data are tabulated on p 722)

```
dist.u=dist(T12.9.df)       # creates Euclidean distances from data matrix
cmd.out=cmdscale(dist.u,2)
cmd.out[,1]=-cmd.out[,1]
plot(cmd.out,pch=c(u.labels[[1]]))
```



Note same thing based on standardized data:



12.6 Correspondence Analysis – a bit like MDS, PC combined - more later
A way of plotting aa contingency table in 2-D

12.7 Biplot Plot objects in two dim space showing both cases and variables!
Universities example p 723.

Data Mining? Later.