

**STAT 597D, SPRING 2001
FINAL PROJECT
DUE WEEK OF APRIL 23**

The final project will consist of a group presentation that your group will give to the class on either Tuesday, April 24 or Thursday, April 26. Plan on approximately 20 minutes for the presentation. The presentation should coherently explain what you did, how you did it, and what you found in a way that your peers can understand. The grade, of course, will depend on how well you meet that objective and on the quality of your work; however, we will probably not grade especially harshly. If you need transparencies or any other supplies, let us know. Obviously, you'll have access to the computer in the lectern as well if you wish.

For a topic, please select one of A through F. If your group would like to pursue a topic not on this list, that should be okay but please check it out with us first. We don't mind if two groups decide on the same topic, since different presentations will undoubtedly cover different points anyway. However, if you could email Dave as soon as your group settles on a topic, we will keep a list on the web just in case the other groups want to avoid doubling up on one topic.

Project A:

Analyze the pedigree data from Litt et al (Am J of Hum Gen 1994; 55:702-709) on episodic ataxia and graph a location score curve. This may be based on an exact computation (very time-consuming computationally) or an MCMC scheme (more difficult to set up but quicker computationally). You can do this analysis on a different pedigree if you want to, as long as it is a sensible use of lod scores. The episodic ataxia data set may be found at www.stat.psu.edu/~dhunter/genetics/ataxia.ped.

Project B:

Read the paper on TDT for multi-allele marker loci by Sham and Curtis (Annals of Hum Gen 1995; 59:323-336) and show how the model of Bradley and Terry (1952) is used in this context. Find a relevant dataset and fit the model; draw appropriate conclusions.

Project C:

Using Lange (1997), chapter 12 as a reference, reanalyze the drosophila dataset of Assignment 3 (also Table 12.1, p. 222) using one of the other three methods summarized in Table 12.2 (count-location, chi-square, or mixture). Keep in mind that the values in Table 12.2 may not be correct, since the Haldane values are known to be wrong. Explain the model, do the computations, and compare with the Haldane model.

Projects D-F (Microarray data analysis):

Consider again the data on cdc yeast cell-cycle data from Stanford (files available online; see Assignment #2). This time, the objective is to try to partition the N=800 genes on the basis of expression similarity over the available time course. This can be done using a clustering algorithm.

Again you will have to decide

1. Whether or not to consider all time points (take $T=15$, or less)
2. How to deal with missing entries (possibly deleting genes, and thus taking an N smaller than 800)
3. Whether to center and standardize the data by row and/or column.

Apply a hierarchical or a K-means algorithm, and select a K by “eyeballing” the “fit” statistics (if you are going to work on Project F below with K-means, stay on “the large side” – select a larger K rather than a smaller one).

Visualize the clusters in the principal components plane, and/or in the the discriminant plane (the categorical response to build the latter is provided by the cluster memberships is output of the algorithm).

If you are using hierarchical clustering, visualizing the dendrogram for hundreds of genes will be problematic. If you are using K-means, do visualize the centroids as (time) profiles superimposed on the same plot.

Create any other output summaries that you find interesting (e.g. cluster frequencies, within-cluster square distances, etc.) and comment. You may also use both hierarchical and K-means algorithms to obtain partitions in the same number of clusters, and compare them by (i) cross-tabulating the cluster memberships, (ii) verifying whether known genes that are expected to have similar expression profiles are indeed clustered together by one or both of the algorithms, etc.

Now pick one of the following three Projects:

Project D:

Some authors (see for example Holter et al., 2000) claim that this data set does not present an obvious “lumpiness”, which might signify continuity in expression. In order to verify this, we could study how strong is the “association” of genes to the clusters, overall. Here is one proposal: for each gene, compute the ratio between (i) the distance of the gene profile from the cluster it is attributed to, and (ii) the average distance of the gene profile from the clusters

$$q(x) = \frac{D(x, C_{m(x)})}{\frac{1}{K} \sum_{k=1}^K D(x, C_k)}$$

$$q(x) = \frac{d(x, \bar{x}_{m(x)})}{\frac{1}{K} \sum_{k=1}^K d(x, \bar{x}_k)}$$

(the two formulae are relative to, respectively, hierarchical and K-means clustering – see class notes). If the association between genes and clusters is strong, the q(x)’s will tend to be small. Produce a histogram and summary statistics on these quantities, and comment. (If you want, you can perform a similar analysis using a quantity other than the ratio we proposed above).

Project E:

It is well known that traditional clustering algorithms are very sensitive to anomalous observations. Concentrate on K-means with your choice of K. What we will attempt here is an analysis of the stability of the K cluster centroids when we perturb the data with deletions. In T-dimensions, compute the traditional “leverages” of each point/gene. These are distances of each point from the overall mean (a T-vector), according to the metric provided by the inverse of the

overall T by T covariance matrix. As an alternative, compute simple distances of the genes from the overall mean (it is not clear that the overall covariance matrix provides a good metric to order genes in terms of likely influence on the clustering). Rank the genes according to leverage (or simple distance from the overall mean). Delete 1%, 2%, ... L% (you pick L) of the genes along the ranking, and repeat the cluster analysis. This will produce L sets of K centroids (each a T-vector)

$$\bar{x}_1^{(1)} \dots \bar{x}_K^{(1)}, \bar{x}_1^{(2)} \dots \bar{x}_K^{(2)}, \dots \bar{x}_1^{(L)}, \dots \bar{x}_K^{(L)}$$

If the cluster centroids are stable under deletion of potentially influential points, this set of LxK vectors ought to cluster into K very tight clusters with approximately L observations each. Run a K-means with the same K on them, compute within-cluster square distances and cluster frequencies, and comment.

Project F:

Concentrate again on K-means with your choice of K. What we will attempt here is an analysis of the stability of the K cluster centroids, when we resample at random from our data points. If the choice of K is reasonable, randomly selected subsets of the data ought to produce fairly similar centroid structures. On the other hand, if K is “too large”, a subset of its centroid structure may remain fairly stable under random resampling, but one or more centroids will “jump all over the place”. Produce L random samples of size M<N from the data (you pick M and L), and repeat the cluster analysis on each. This will produce L sets of K centroids (each a T-vector)

$$\bar{x}_1^{(1)} \dots \bar{x}_K^{(1)}, \bar{x}_1^{(2)} \dots \bar{x}_K^{(2)}, \dots \bar{x}_1^{(L)}, \dots \bar{x}_K^{(L)}$$

If K is reasonable, this set of LxK vectors ought to cluster into K very tight clusters with approximately L observations each. Run a K-means with the same K on them, compute within-cluster square distances and cluster frequencies, and comment.