

Confidence Sets for Clusterings

Naomi S. Altman
Dept. of Statistics
Penn State University
March 19, 2004

The Clustering Problem

We want to partition a large collection of
nonhomogeneous objects into

Sets of homogeneous objects

Based on a measure of similarity

How should we do it?

How many sets do we need?

Clustering Noisy Data

When the data are noisy we also need to assess:

How stable are the sets?

What are alternative partitions that are compatible with the observed data?

Resampling for Clustering

Assessing clustering is often done as below:

“Resample” the data.

Recluster.

For each resample, compare the cluster structure with the actual data.

I will resample and recluster and then obtain a “confidence set of clusterings”.

The Data

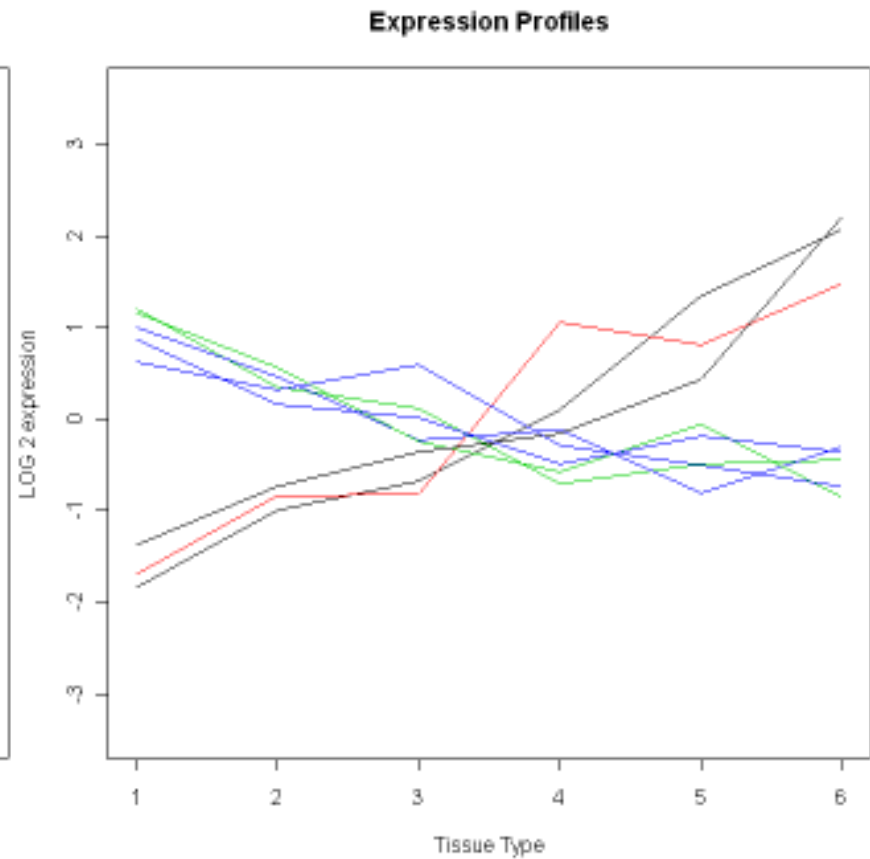
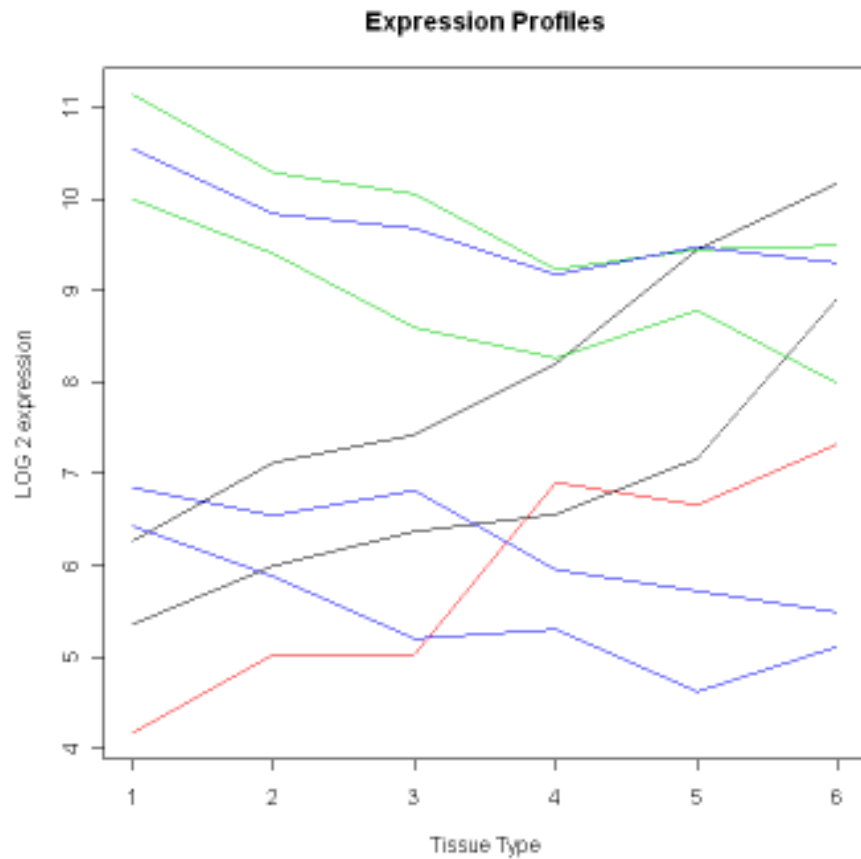
The method is applicable to any collection of objects with a (dis)similarity measure.

We will consider microarray data.

Each microarray has 22746 genes. We have 6 tissues, 2 arrays per tissue.

466 genes have significant tissue effects
(1-way ANOVA)

Similarity of Expression Profiles



Outline of Talk

1. Forming a “confidence set” for clusterings.
2. Bootstrapping microarray data
3. The plant tissue example
4. Statistical issues raised by the example

Forming a Confidence Set: What is a confidence set?

Heuristically, a confidence set is a set of plausible values of a parameter.

Our parameter is the partition=clustering of the data.

The probability model is not clearly defined, so even if we know the true treatment means, there is some ambiguity about the “best” clustering.

We are looking for a set of partitions which are compatible with the observations.

Assessing Cluster Stability

Our objective is to understand how the clustering will change under small changes of the data.

Suppose we could redo the entire experiment and redo the clustering.

A cluster is stable if it appears every time we redo the experiment.

An item is stable if it appears in the “same” cluster every time we redo the experiment.

Perturbing to Assess Stability

Since we cannot redo the experiment, we perturb the observed data and redo the clustering.

Resampling arrays: Take a sample of the observed arrays for each treatment.

Parametric Bootstrap: Fit a model (e.g. Normal ANOVA model) to the experiment. Estimate the treatment means and data variance. Create new samples by adding Normal noise to the estimated treatment means.

Semiparametric Bootstrap: Fit a model (e.g. ANOVA model) to the experiment. Estimate the errors by the residuals. Create new samples by adding randomly selected residuals to the estimated treatment means. (Churchill and Kerr, 2003)

Using the Perturbed Data to Form Confidence Sets

Create and cluster B samples.

Using a metric on partitions, compute the distance between the clusterings.

For a confidence set, pick the clusterings that are “closest” to the clustering of the observed data.

For a “bagged” estimate, pick the clustering at the centroid of the samples.

Some Metrics on Partitions

Meila (2002) proposes 2 measures of distance among clusterings

These distance measures have some good properties:

e.g. splitting or merging clusters produce "close" clusterings, but scrambling produces "distant" clusterings

e.g. clusterings that differ because a few elements moved between clusters are closer than clusterings that differ because lots of elements moved

A Metric for Clusterings

C - a set in the cluster

\mathcal{C} - a clustering

n - the number of objects

$$P(C) = \#C/n \quad h(\mathcal{C}) = \sum P(C) \log(P(C))$$

$$P(C_i, C_j) = \#(C_i \cap C_j)/n$$

$$IC(\mathcal{C}_1, \mathcal{C}_2) = \sum \sum P(C_{1i}, C_{2j}) \log(P(C_{1i}, C_{2j}) / P(C_{1i})P(C_{2j}))$$

$$d(\mathcal{C}_1, \mathcal{C}_2) = h(\mathcal{C}_1) + h(\mathcal{C}_2) - 2 IC(\mathcal{C}_1, \mathcal{C}_2)$$

Algorithm

1. Do the 1-way ANOVA on each gene to generate fitted values and residuals= 2-way ANOVA with gene by tissue interaction.
2. Create B semi-parametric bootstrap samples, leading to B sets of fitted values.
3. Create $B+1$ clusterings of the genes.
4. Consider the subset of clusterings closest to the clustering from the actual fitted values.

Some Notes on the Algorithm

For clustering, I used “Partitioning on Mediods” a method that requires prespecifying K , the number of clusters.

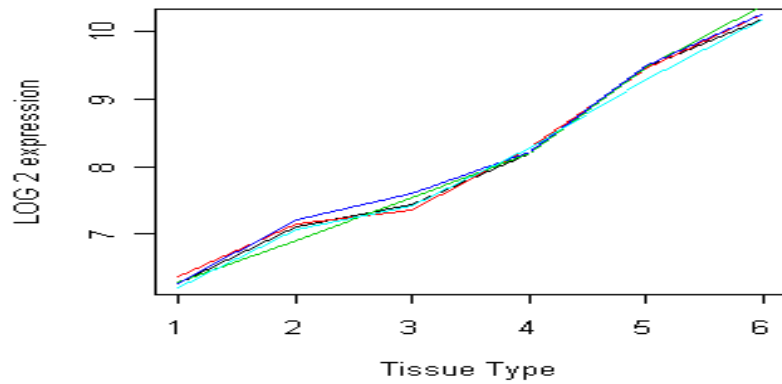
I tried $K=10, 15, 20, 25, 40$.

I used 50 bootstrap samples. For each sample, I clustered at each value of K .

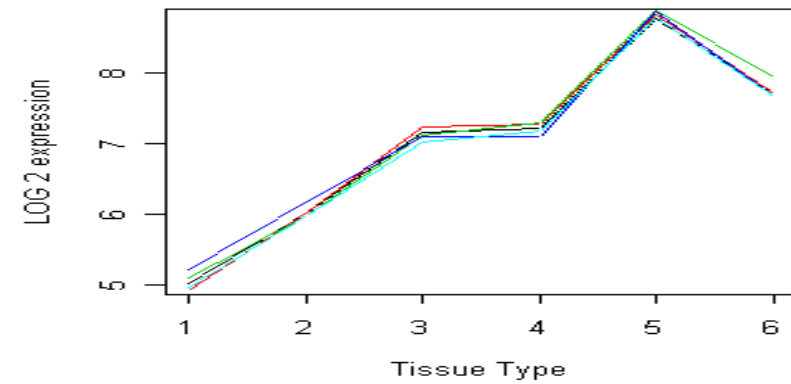
To visualize the distribution of the clusterings, I used multidimensional scaling in 2 dimensions – a method of taking the matrix of all pairwise distances and finding the 2-D locations of the objects that best reproduces the distance.

The Bootstrap Data

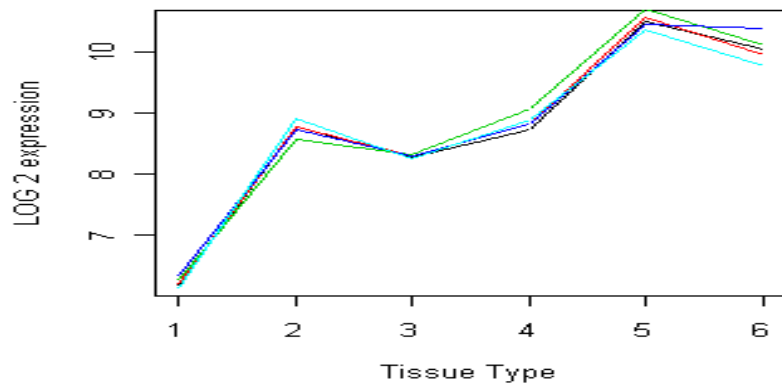
Gene 1 Profiles



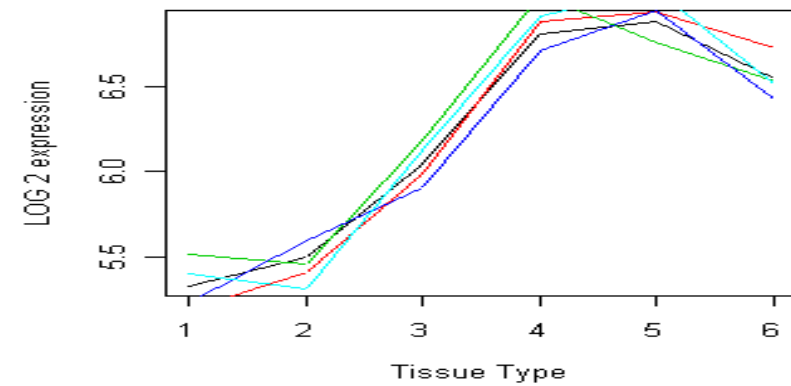
Gene 47 Profiles



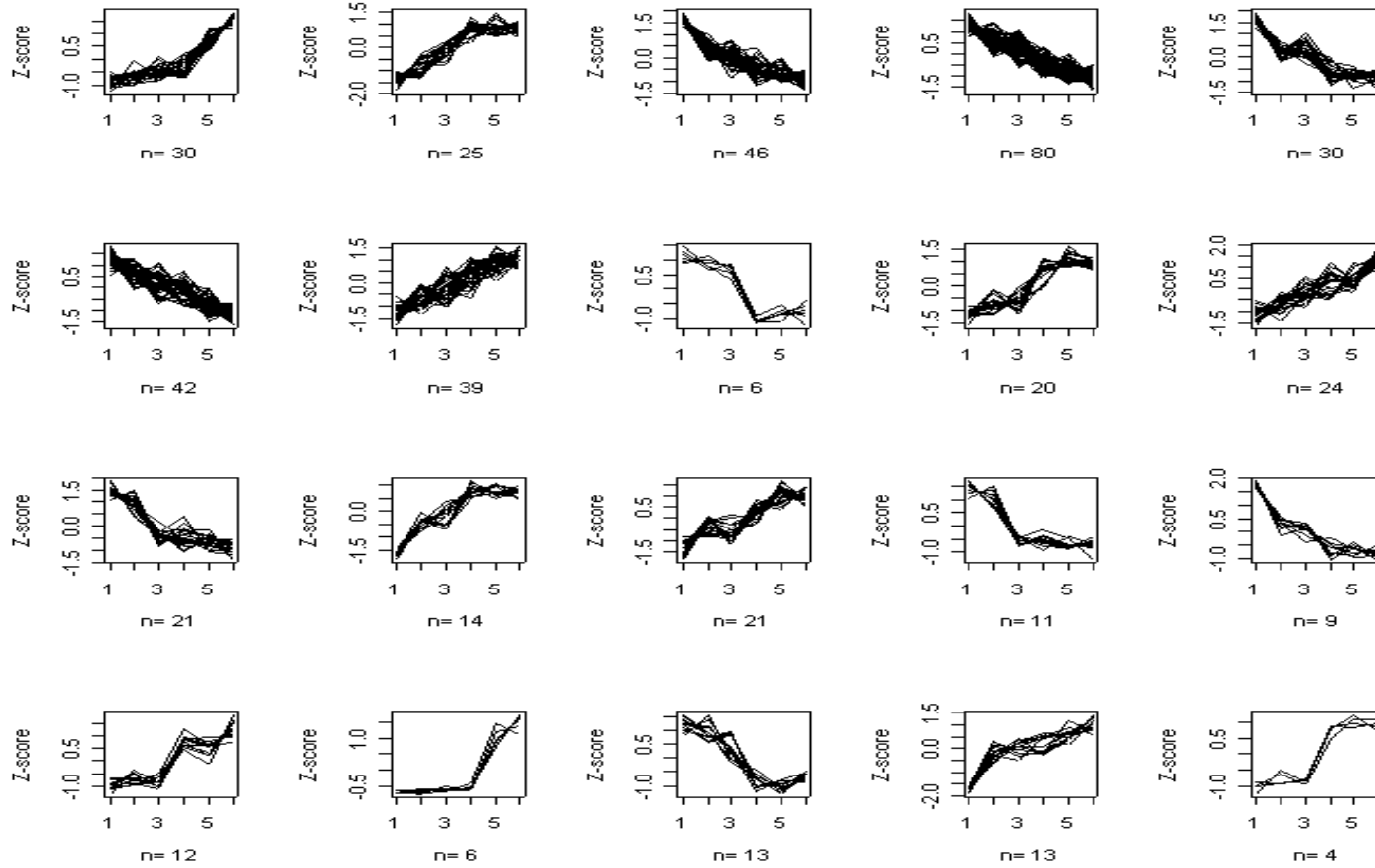
Gene 175 Profiles



Gene 413 Profiles

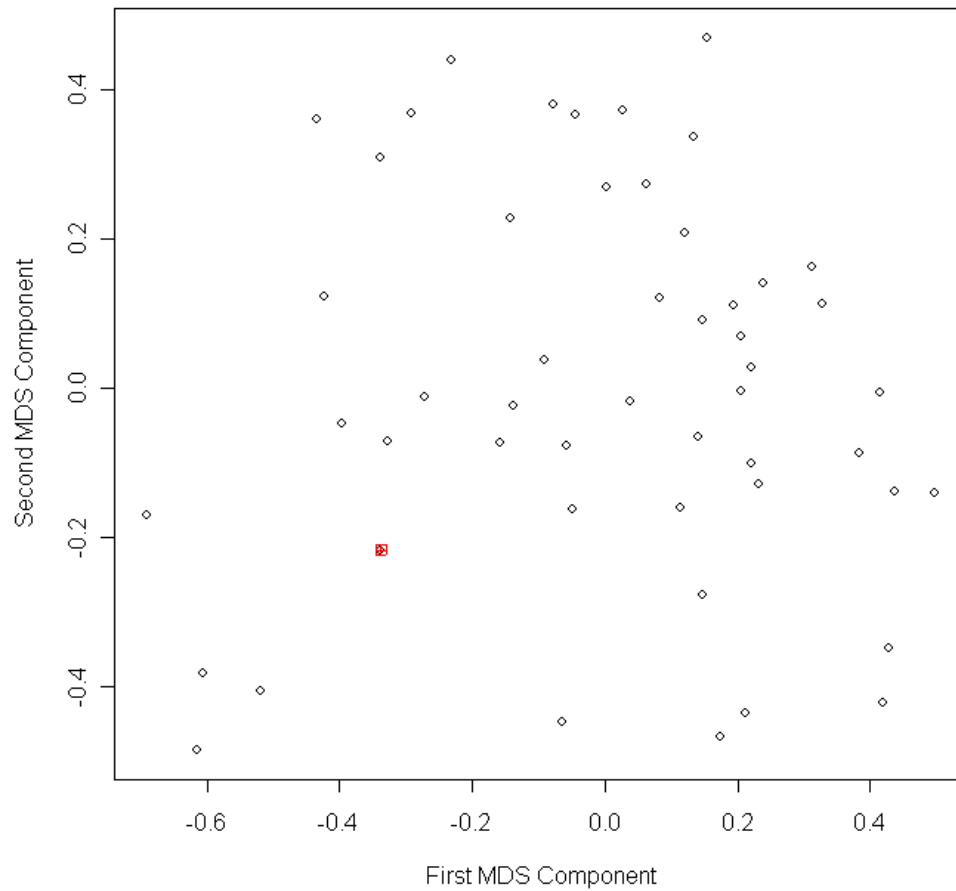


The 20 cluster profiles



K=20 Bootstrap Clusterings

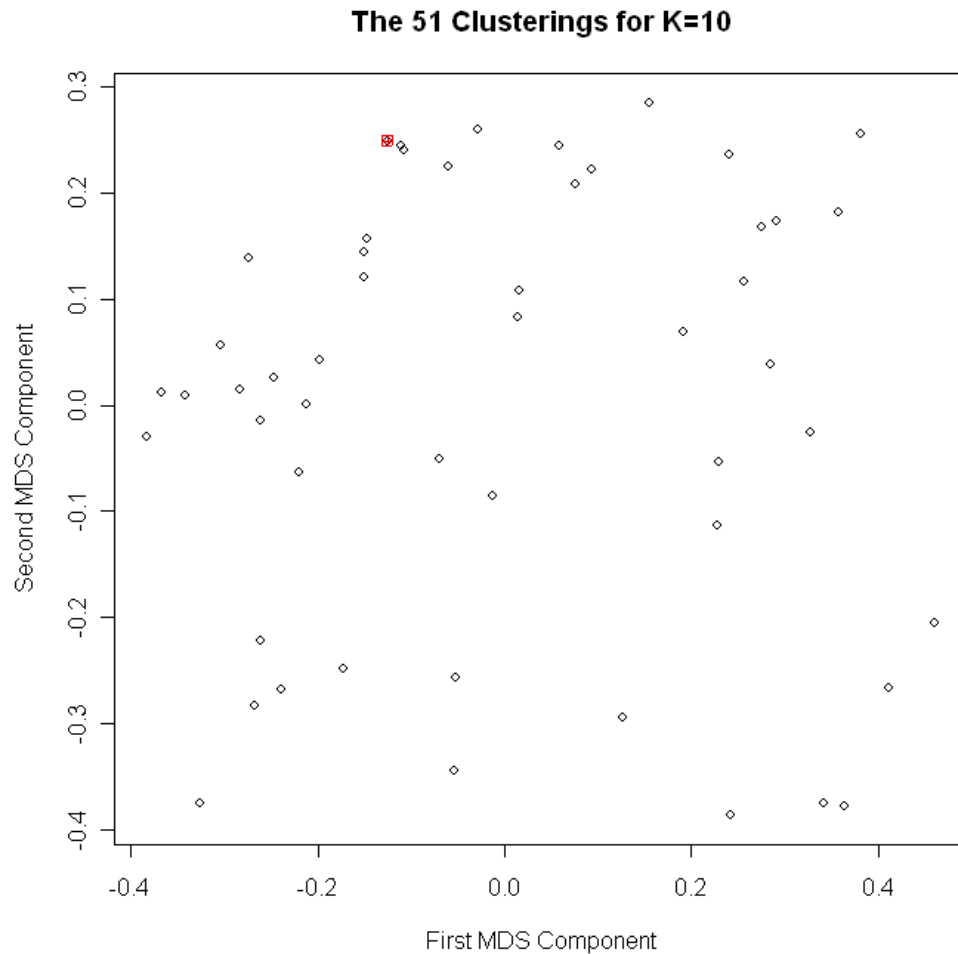
The 51 Clusterings for K=20



The clustering of the actual data is not centrally located.

The clustering of the actual data is isolated

K=10



The clustering of the actual data is even less centrally located.

The clustering of the actual data is not isolated but ... the nearest clusterings are still not very close

What about other values of K

None of $K=10, 15, 20, 25,$ or 40 have very nice “confidence sets”

Note that we use the same bootstrap samples at every value of K . So the nearest neighbors of the data should be similar.

10 nearest neighbors:

K=10 44 48 5 39 2 41 8 50 3 20

K=15 1 44 10 26 23 3 39 6 21 14

K=20 27 31 3 38 20 22 48 45 29 44

K=25 8 2 42 27 39 14 34 49 29 44

K=40 15 23 8 20 26 39 13 11 43 9

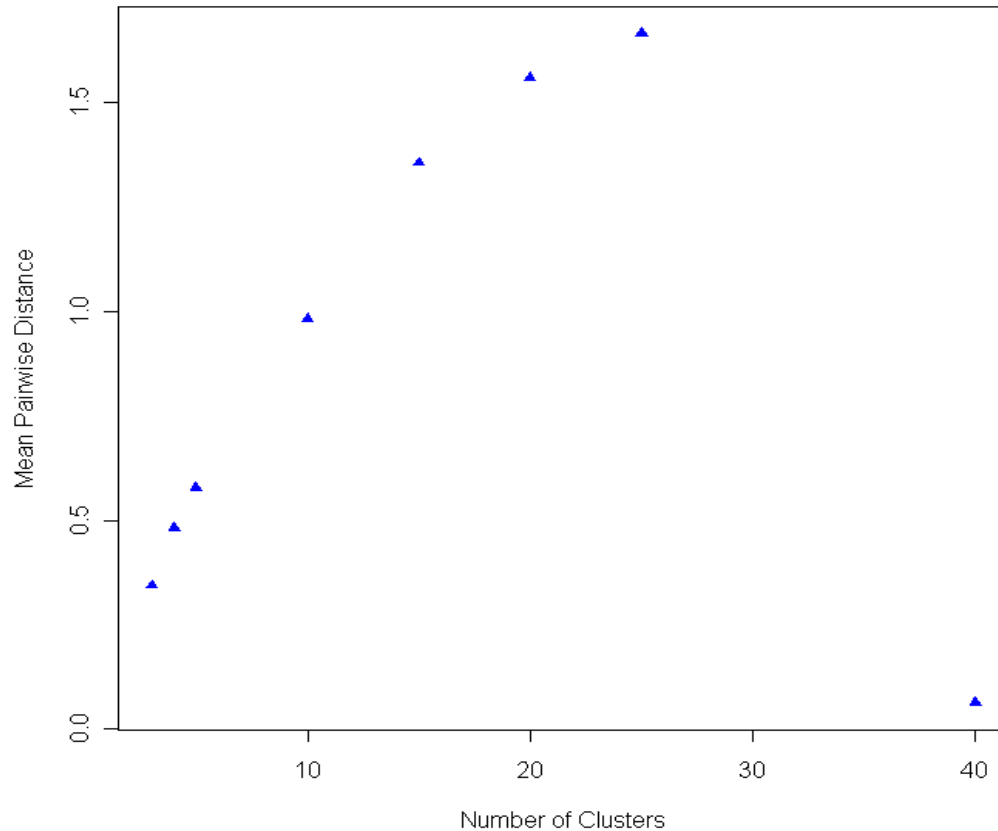
What went wrong?

Possibilities:

- Data Quality – only 2 reps (but the bootstraps look good).
- Choice of K (see next slide)
- The distance measure is poor.
- The idea just doesn't work.

Distance Among Bootstrap Samples

Mean pairwise distance as a function of K



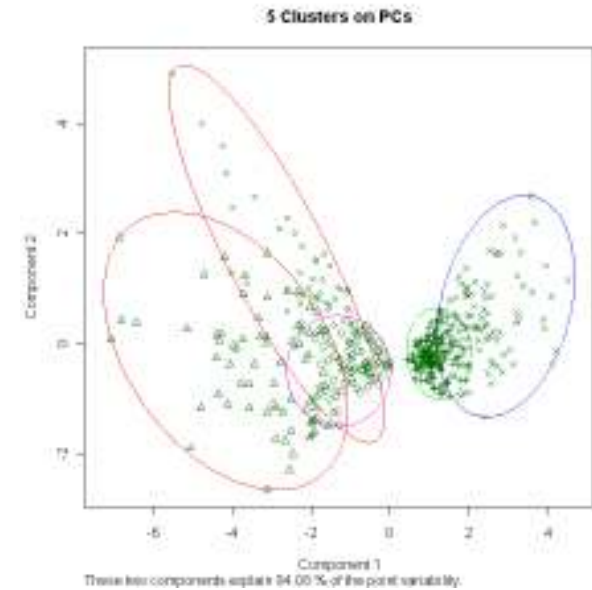
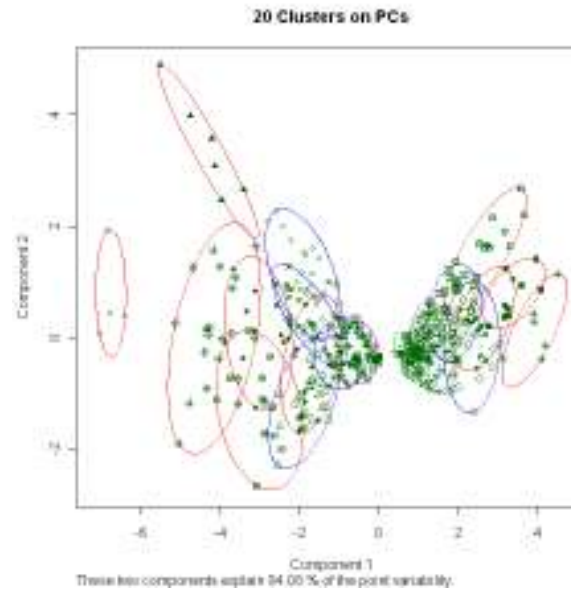
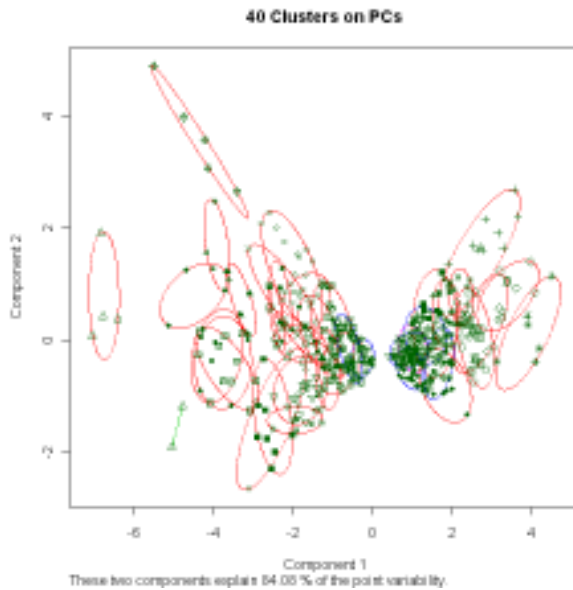
The number of genes that changed cluster

K=3 m=7

K=4 m=9

K=5 m=18

How Many Sets in the Clustering?



What went wrong?

Current hypothesis:

Poor choice of K.

For the rest of the story ...

For the biologists: Biological meaning is more important than statistical stability

For the statistician: A firm statistical basis for choice K and consensus clustering or confidence set of partitions is needed

In practice, we expect that biologically meaningful clusters will be supported by statistical evidence.

Many thanks to:

Xiaohong Zhang and Hong Ma who collected that data and spent many hours discussing the meaning of “gene profile similarity”

Claude de Pamphilis, Jim Leebens-Mack and Kerr Wall of the Floral Genome Project who spent many hours discussing the meaning of “statistical support for a cluster” based on sequence analysis

The Dept. of Statistics at PSU for providing a supportive environment for collaborative research