Clustering (B)

Visualizing clusters if T>2:

Plot the points, color coded according to cluster membership, on the 1st principal components plane. This 2D view is "most representative" of the data, in the sense that it maximizes the share of captured overall variation, but is not necessarily the best to separate clusters.

Alternatives:

Plot the points, color coded according to cluster membership, on the 1st plane from multidimensional scaling. This 2D view is the one that best preserves distances among data points, and it may be better to separate clusters.

Treat cluster memberships as a classification response, and find the 1st discriminant (or SIR) plane relative to it. This 2D view is the one that maximizes cluster separation.

(Relatedly) Dimension reduction and clustering:

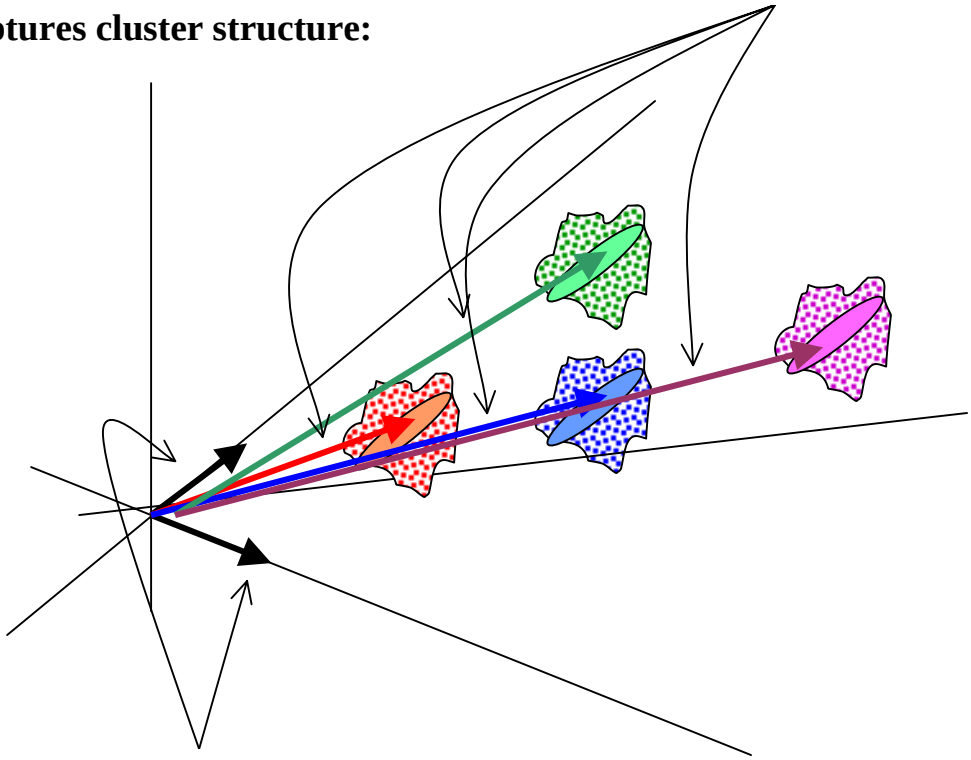
dimension reduction techniques are NOT clustering tools.

However, a dimension reduction may be performed prior to clustering (clustering occurs in terms of the reduced representation; e.g. projecion on a low-dimensional space)

- To eliminate unwanted variation sources, artifacts, from the clustering exercise (then PCA may be a good idea, but care is needed on how much and what we are willing to "throw away")
- To facilitate cluster computation (then MDS seems definitely a smarter option).

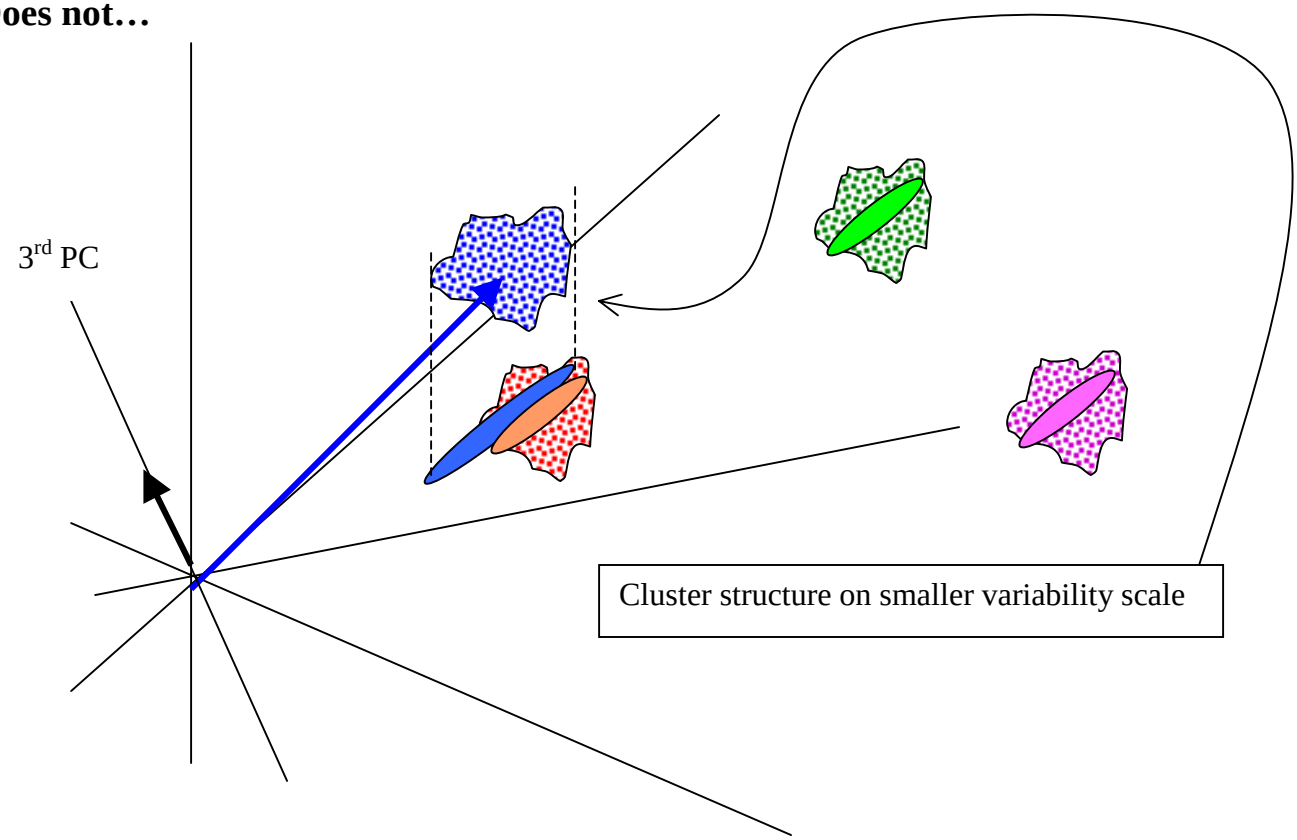
Cluster centroids (typical patterns)

PCA captures cluster structure:



Principal components
(building block patterns)

Does not…



Yeung K.Y., Ruzzo W.L. (2001): Principal component analysis for clustering gene expression data. *Bioinformatics* 17 (9) 762-744.

Using more than one clustering method, more than one underlying metrics choice, and both actual and simulated data, they show how clustering based on the first few principal components may significantly degrade the clustering results.

Important for this paper and others to come:

Quantifying the similarity between two partitions of the same set of N "objects" (e.g. genes)

$\binom{N}{2}$ pairs of objects

$$Rand = \frac{\#\text{ pairs together in both partitions } + \#\text{ pairs not together in both partitions}}{\binom{N}{2}} \in [0,1]$$

(expected value in the case of corresponding random partitions is not 0)

$$CorrRand = \frac{Rand - Rand(\text{two corresponding random partitions})}{Max\_Rand - Rand(\text{two corresponding random partitions})} \in [0,1]$$

(expected value in the case of corresponding random partitions is 0)

other quantifications are possible (we will encounter another later)

A side note: although dimension reduction techniques do not produce clusters, they can be used to form groups of genes as for instance

- the closest to the first, second, third etc. direction;
- the closest and furthest from the first direction, plane, 3Dspace, etc.