Visualizing clusters:

(if T>2) After clustering, 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.

(Relatedly) Dimension reduction and clustering:

dimension reduction techniques are NOT clustering tools.

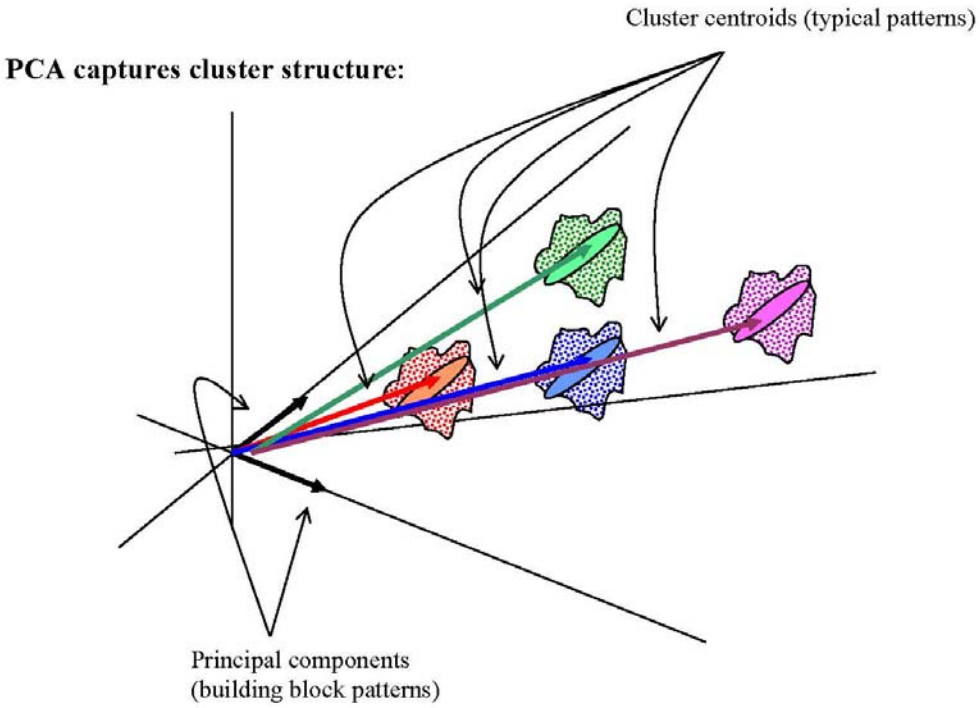
However, a dimension reduction may be performed prior to clustering (clustering occurs in terms of reduced representation; i.e. projection on a low-dim space), to:

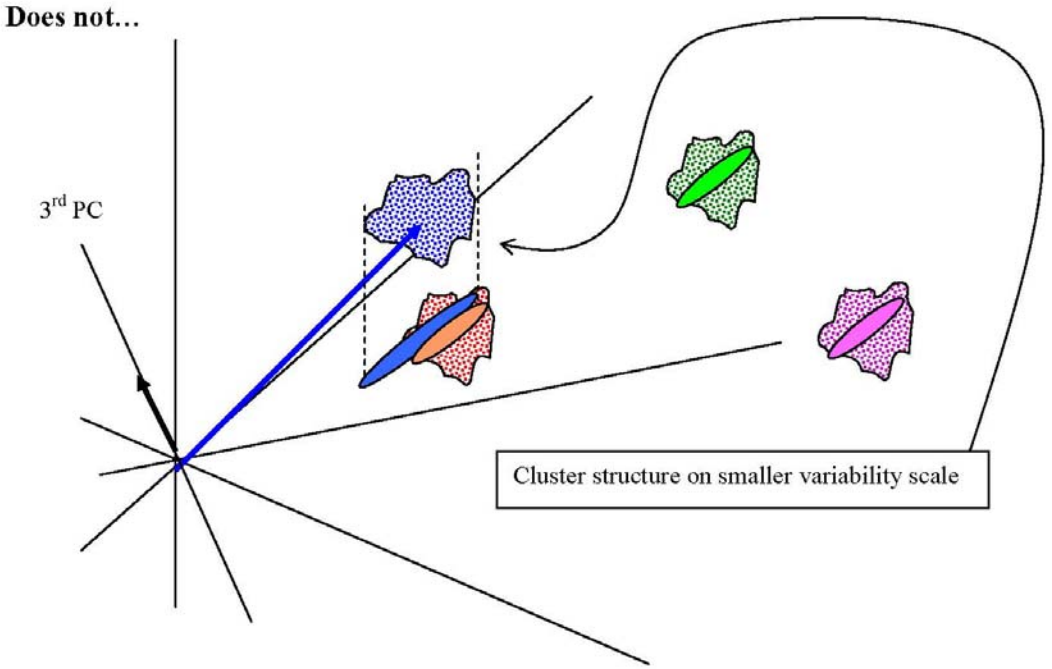
- Eliminate unwanted variation sources, artifacts, from the clustering exercise (care is needed on how much and what we are willing to "throw away").

- Facilitate cluster computation (some algorithms, e.g. mixture-based, depend strongly on dimension).

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.





Alternatives to PCA, in relation to clustering:

Multidimensional scaling:

Find directions, and thus low-dim projections, that preserve distances among data points. After clustering, 2D views obtained with multidimensional scaling may provide a better cluster visualization (in terms of displayed separation) than 2D views obtained from PCA. Before clustering, reducing the data with multidimensional scaling aims at preserving the “basis” for clustering (distances), and thus may be more effective than PCA.

Linear classifiers (Discriminant Analysis, Sliced Inverse Regression):

After clustering, treat cluster memberships as a (known) classification response. Find directions, and thus low-dim projections, that preserve separation (“distinguishability”) among the now given classes. These 2D views are “optimized” for cluster visualization.

Important for Yeung and Ruzzo (2001) and other papers:
Quantifying similarity between two partitions of a 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)

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

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

Other quantifications in Ben-Hur et al (2002).

Recall: 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 or furthest from the first direction, plane, 3Dspace, etc.

Example

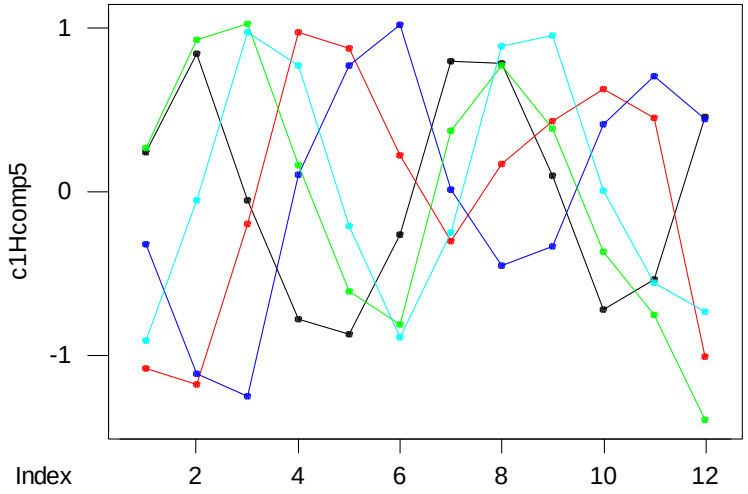
Yeast cell-cycle data

~679 genes (exhibiting periodic behavior)
first 12 time points from cdc-synchronized time course
normalized (to green; not synchronized) log-ratios, from spotted arrays
no missing entries
row-standardized columns (only first 12)

Spellman P.T., Sherlock G., Zhang M.Q., Vishwanath R.I., Anders K., Eisen M.B., Brown P.O., Botstein D. (1998), Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization, *Molecular Biology of the Cell* **9**:3273-3297.

Cluster centroids, Hcomp5

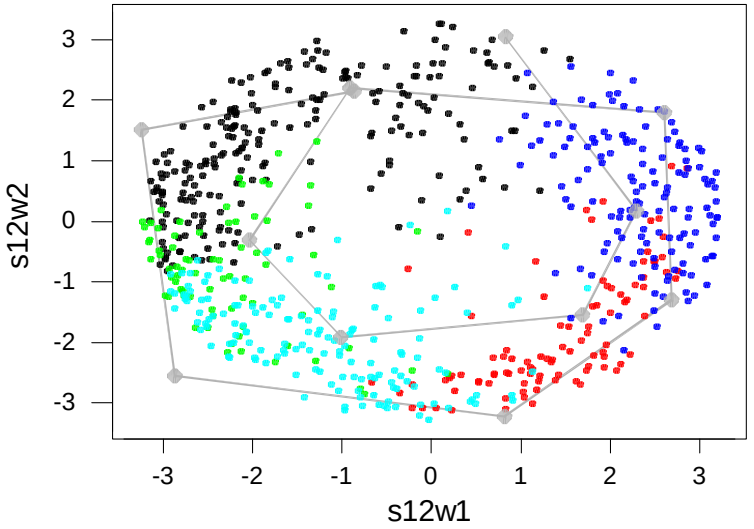
black=1 red=2 blue=3 green=4 cyan=5



(1)=218 (2)=96 (3)=139 (4)=75 (5)=150 genes

Clusters from Hcomp5 (1st PCA plane)

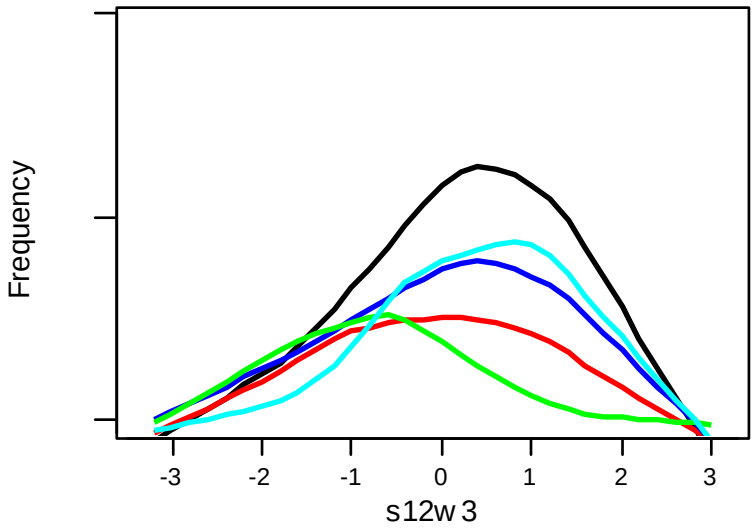
black=1 red=2 blue=3 green=4 cyan=5



(gray: original coord's projected on the plane, joined in time order and magnified)

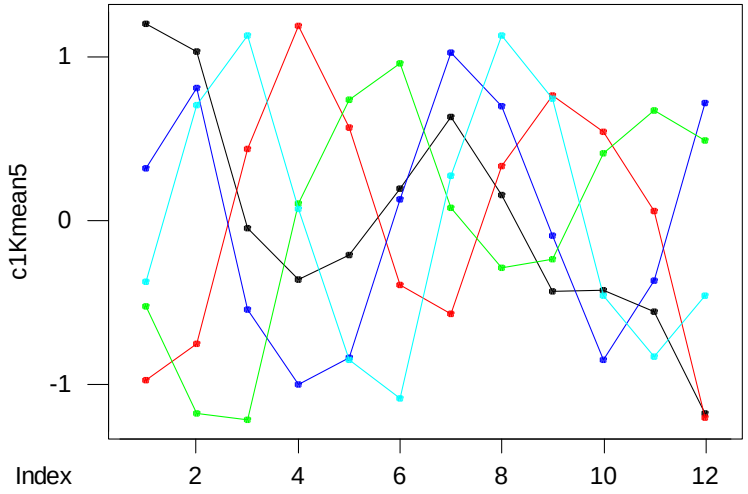
Smoothed freq. histograms for Hcomp5 clusters, s12w3

black=1 red=2 blue=3 green=4 cyan=5



Cluster centroids, Kmeans5

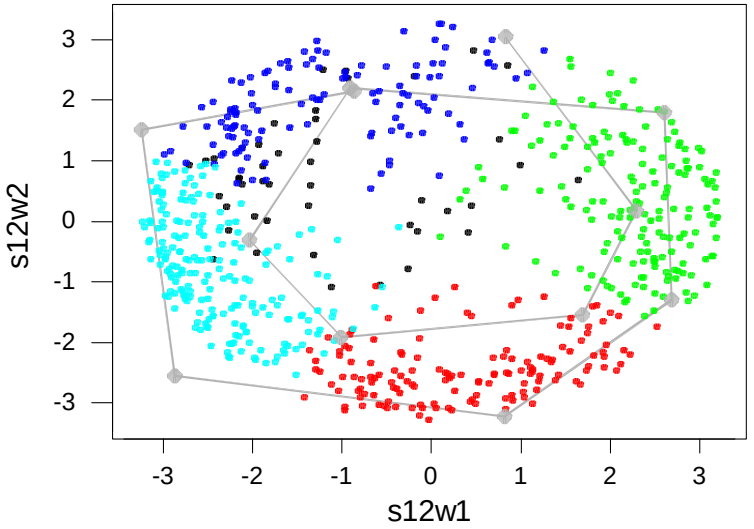
black=1 red=2 blue=3 green=4 cyan=5



(1)=45 (2)=143 (3)=119 (4)=169 (5)=202 genes

Clusters from Kmean5 (1st PCA plane)

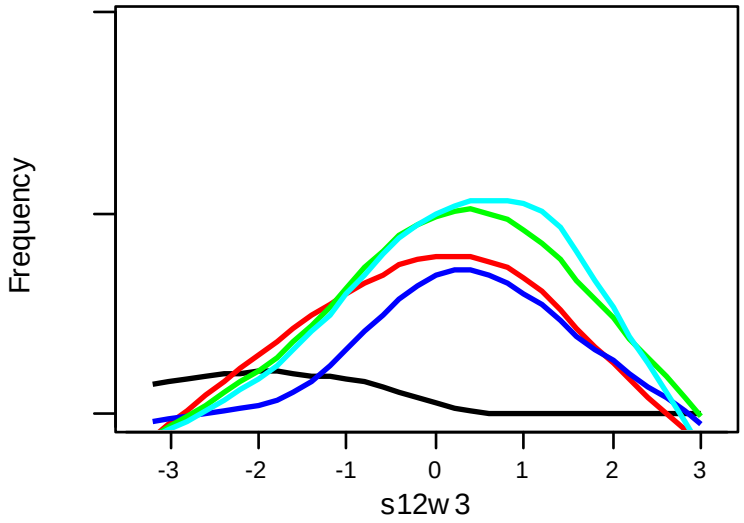
black=1 red=2 blue=3 green=4 cyan=5



(gray: original coord's projected on the plane, joined in time order and magnified)

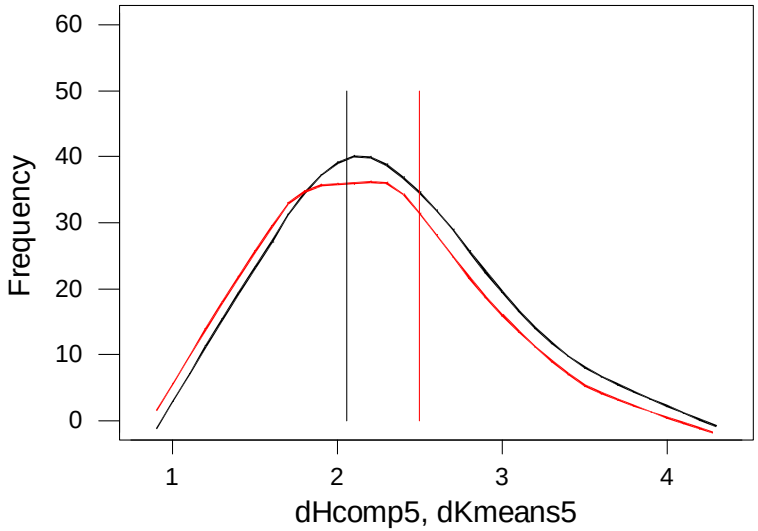
Smoothed freq. histograms for Kmean5 clusters, s12w3

black=1 red=2 blue=3 green=4 cyan=5



Smoothed freq. histograms, distance from centroid

black: Hcomp5, red: Kmeans5



min dist between centroids: Hcomp5=2.0563, Kmeans5=2.4974

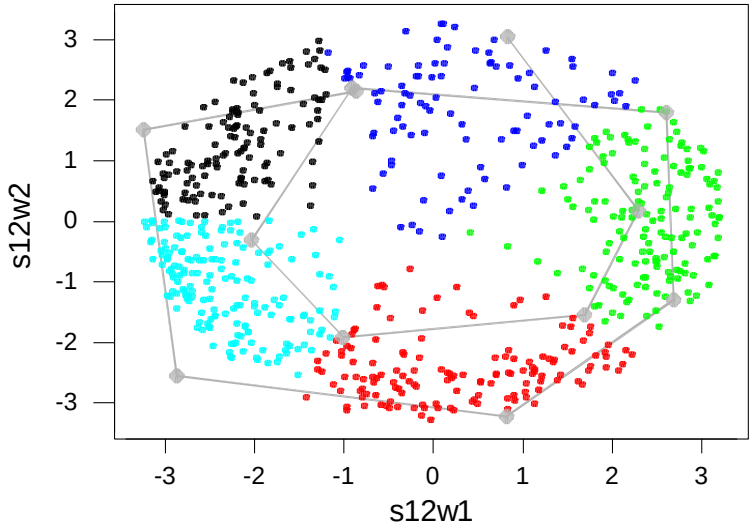
One possible visual comparison of two partitions:

Histogram of the distances between each point and the centroid of the cluster it belongs to.

What share is on the right of the minimum distance between centroids?

Clusters from Kmean5pc12 (1st PCA plane)

black=1 red=2 blue=3 green=4 cyan=5

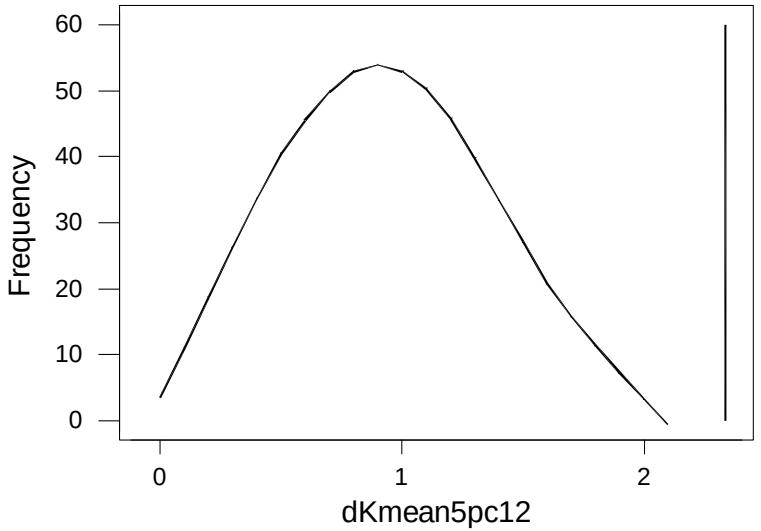


(gray: original coord's projected on the plane, joined in time order and magnified)

<u>Clustering after emiliorating artifacts</u> (amplitude dampening, trend), i.e. using the PC(1,2) projection.

But care needed: Yeung & Ruzzo (2001)

Smoothed freq. histogram, distance from centroids



min distance between centroids 2.3351