CS6220: Data mining techniques
Unsupervised analysis methods

Olga Vitek

November 4, 2015

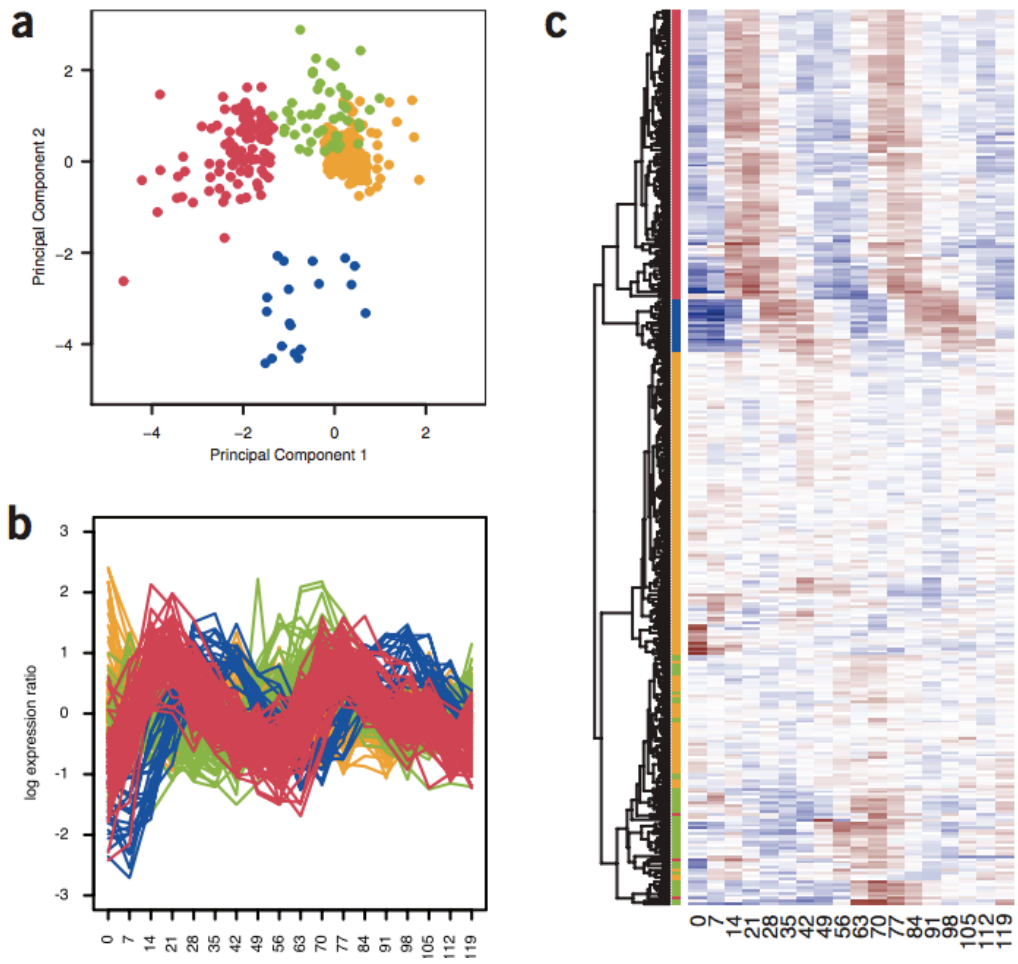
Outline

Principle component analysis

Heatmaps and hierarchical clustering

K-means

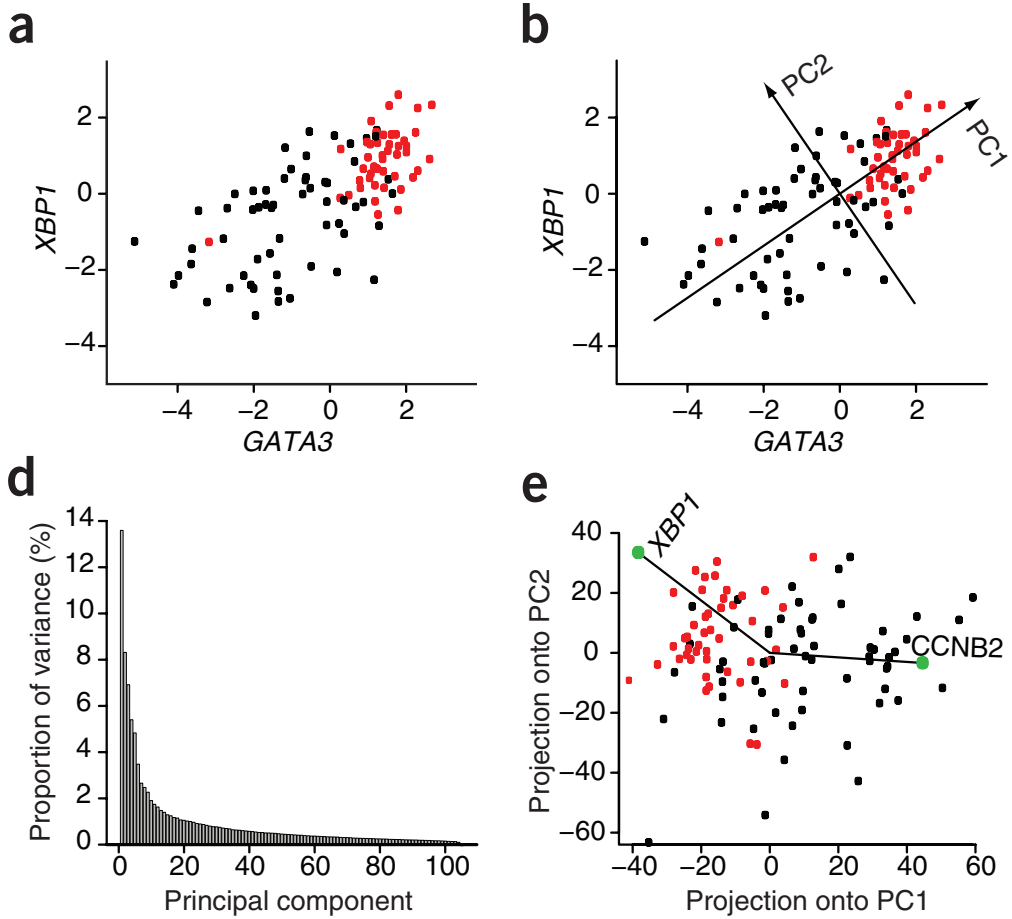
Can look for patterns in both samples and variables



Gehlenborg *et al*, Nature Methods, 2010

Principle component analysis

Overvew



Ringnér, Nature Biotechnology, 2008

Overview

- Each sample as an observation in a

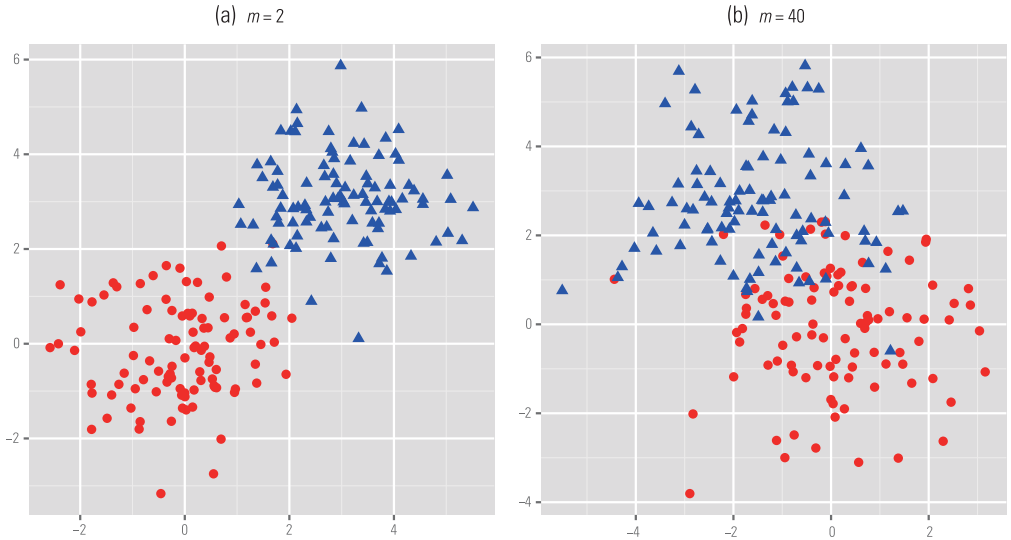
  G-dimensional space

  - Use the 'traditional' representation of the data (rows=observations; columns=variables)
  - X is the $I \times G$ matrix of centered variable expressions

	g_1	$\ldots$	g_G
S_1	$x_{11} - \bar{x}_{\cdot 1}$	$\ldots$	$x_{1G} - \bar{x}_{\cdot G}$
$\ldots$		$\ldots$	
S_I	$x_{I1} - \bar{x}_{\cdot 1}$	$\ldots$	$x_{IG} - \bar{x}_{\cdot G}$
$S_\cdot$	$\bar{x}_{\cdot 1}$	$\ldots$	$\bar{x}_{\cdot G}$

- Goal: find at most I linear combinations of variables that best characterize the total between-sample variation

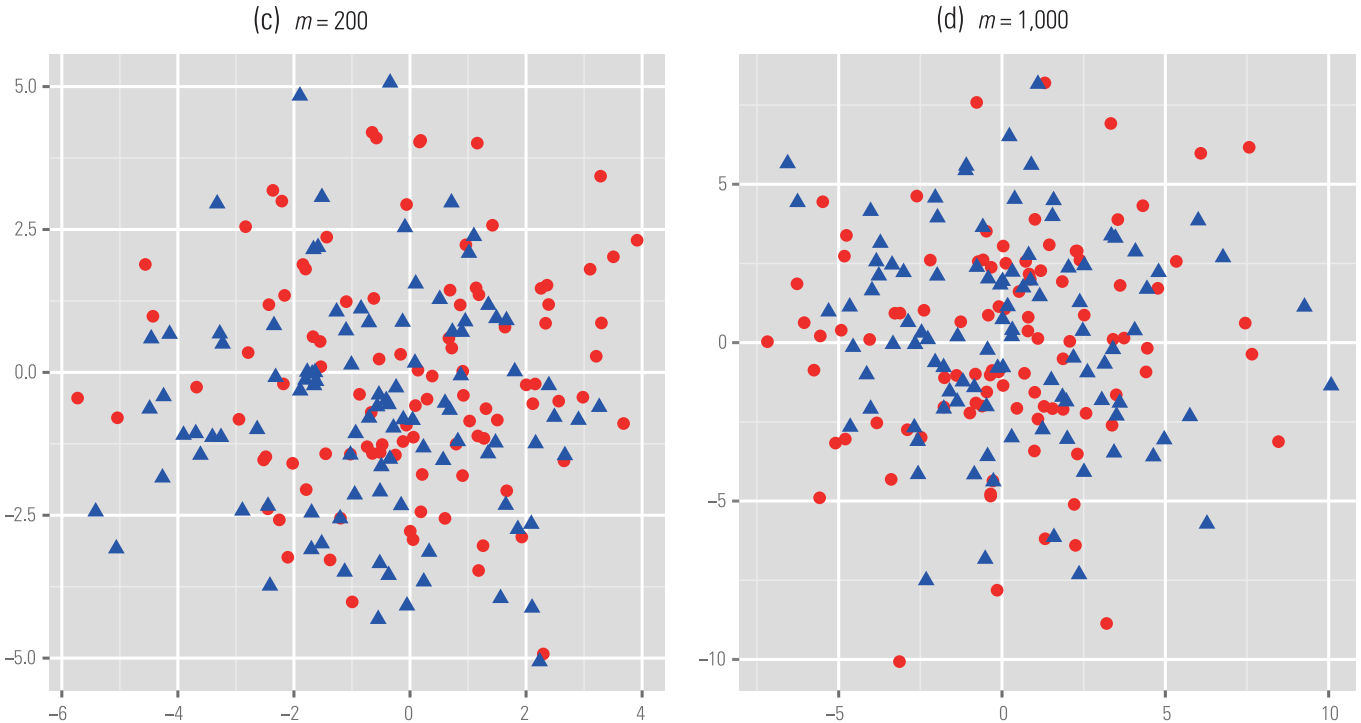
With PCA, signal is lost when data has many unrelated dimensions



A simulation study

- Simulate $n = 100$ observations from 2 classes
- Each observation is a point in $m = 2, 40, 200, 1000$ dimensions
- Only first 10 dimensions are informative
- Plot first 2 principle components (i.e., eigenvectors)
- Informative data should show a good separation between the two classes

With PCA, signal is lost when data has many unrelated dimensions



Conclusion: As we add new unrelated variables, we lose information

Heatmaps and hierarchical clustering

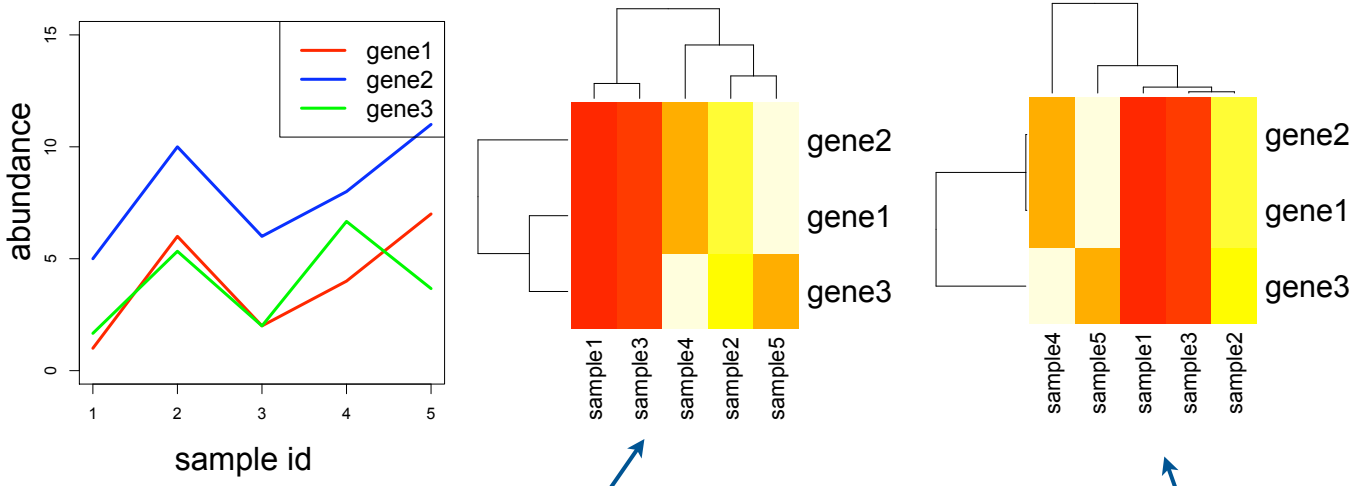
Define dissimilarity between multivariate data points

- $\mathbf{x} = (x_1, \ldots, x_P),\ \mathbf{y} = (y_1, \ldots, y_P)$
- Eucledian distance
  $d_{euc}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{P}(x_i - y_i)^2}$
- Pearson sample correlation distance
  $d_{cor}(\mathbf{x}, \mathbf{y}) = 1 - r(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{P}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{P}(y_i - \bar{y})^2}}$
- Spearman sample correlation distance
  $d_{spear}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{P}(x_i' - \bar{x}')(y_i' - \bar{y}')}{\sqrt{\sum_{i=1}^{P}(x_i' - \bar{x}')^2 \sum_{i=1}^{P}(y_i' - \bar{y}')^2}}$
  where $x_i' = rank(x_i)$ and $y_i' = rank(y_i)$

Important points of the algorithm

- Agglomerative vs divisive
  - Agglomerative: group smaller clusters
  - Divisive: split clusters (more comp. intensive)
- Linkage
  - Single (min. pairwise distance)
  - Complete (max pairwise distance)
  - Average (average of pairwise distances)
- Standardizations
  - Standardize variables: values in each variable on a same scale. **Affects dendrograms (i.e. order of rows and columns).**
  - Standardize colors: colors in each variable on a same scale. **Affects the color distribution of the matrix.**
  - In many implementations. the option 'scale' affects colors but not expression values

Effect of choice of distance



$$Euclidian(\vec{e}_i, \vec{e}_j) = \left\| \vec{e}_i - \vec{e}_j \right\|_2 = \sqrt{\sum_{k=1}^{n} (e_{ik} - e_{jk})^2}$$

$$Pearson(\vec{e}_i, \vec{e}_j) = \frac{Cov[\vec{e}_i, \vec{e}_j]}{\sqrt{Var[\vec{e}_i] Var[\vec{e}_j]}} = \frac{\sum_{k=1}^{n} (e_{ik} - \mu_i)(e_{jk} - \mu_j)}{\sigma_i \sigma_j}$$

Conclusion:

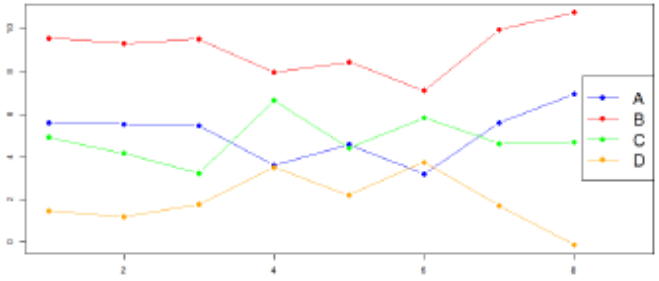
Eucledian: cluster variables with similar values

Pearson: cluster variables with similar profile patterns

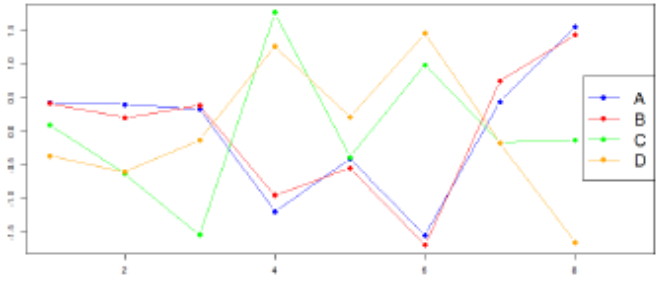
Caution:

The dendrograms are not unique

Effect of standardization



(a) original scale



(b) standardized scale

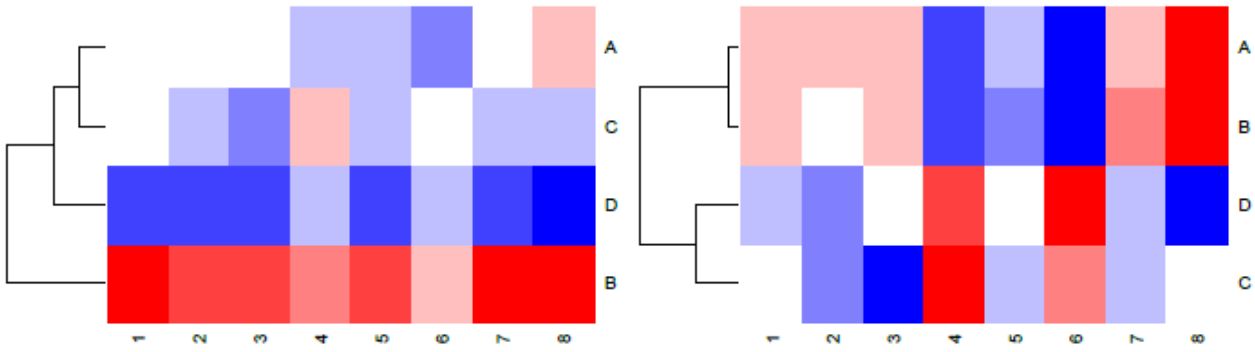
Measurement 1	Measurement 2	Euclidean distance		Correlation distance	
		original	standard	original	standard
A	B	11.4058	0.4980	0.0177	0.0177
A	C	5.4447	4.7830	0.3659	0.3659
A	D	10.9919	5.2709	0.0155	0.0155
B	C	13.1881	4.7092	0.4159	0.4159
B	D	21.2118	5.2096	0.0615	0.0615
C	D	8.5689	2.3094	0.3809	0.3809

Conclusions:

- **Original scale:** A and C are closest in location, while A and B are most correlated.
- **Standardized scale:** correlation distance does not change. A and B have similar standardized values.
- D has a large negative correlation with the A and B, so its correlation distance is low

M. Key, BMC Bioinformatics, 2012

Role of color mapping



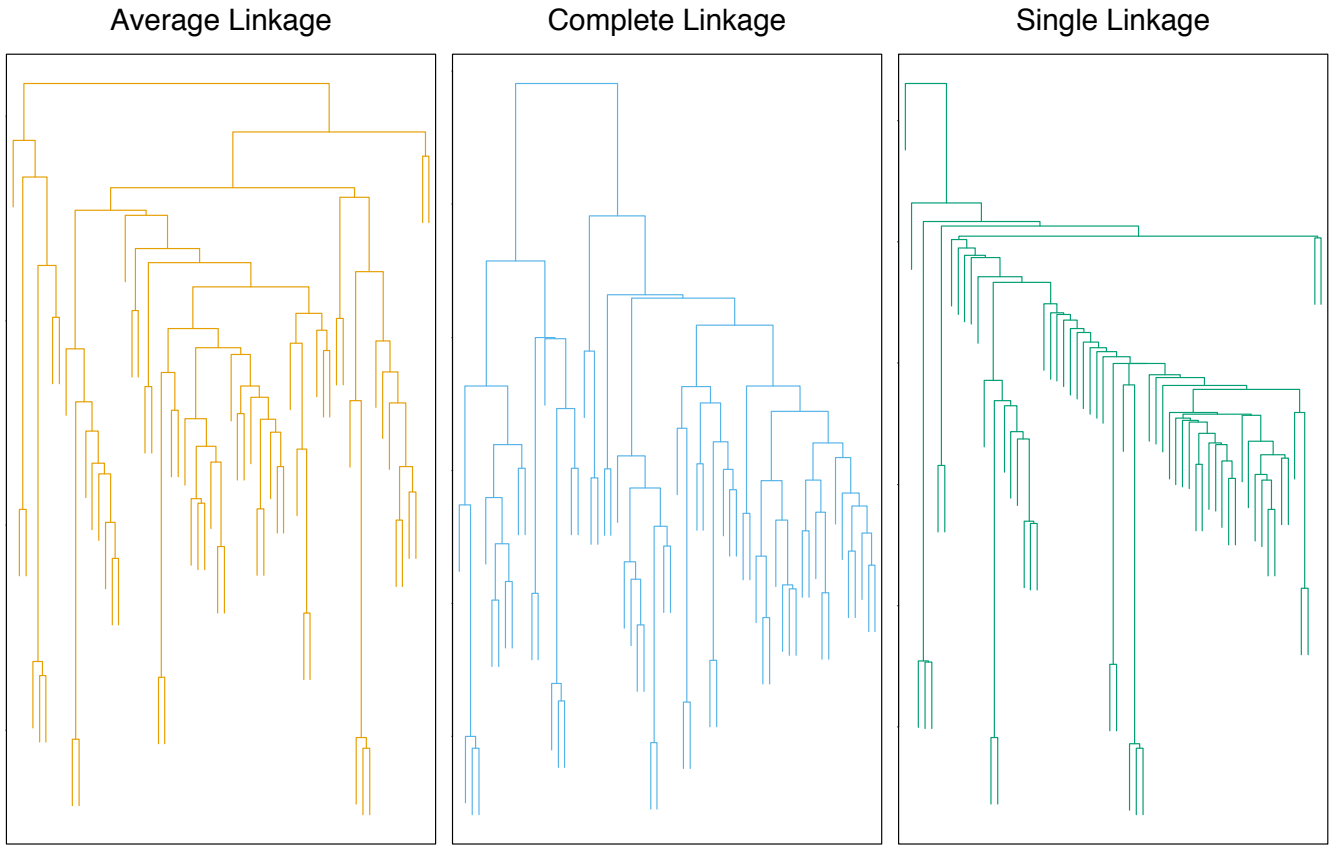
(a) original scale (b) standardized scale

- (a): clusters & colors are applied to original data
- (b): clusters & colors are applied to row-scaled data

Conclusion: With row-scaled data, it is easier to see that the patterns in A and B are the same.
Conclusion: Need to make standardization decisions for both values and colors

M. Key, BMC Bioinformatics, 2012

Role of linkage



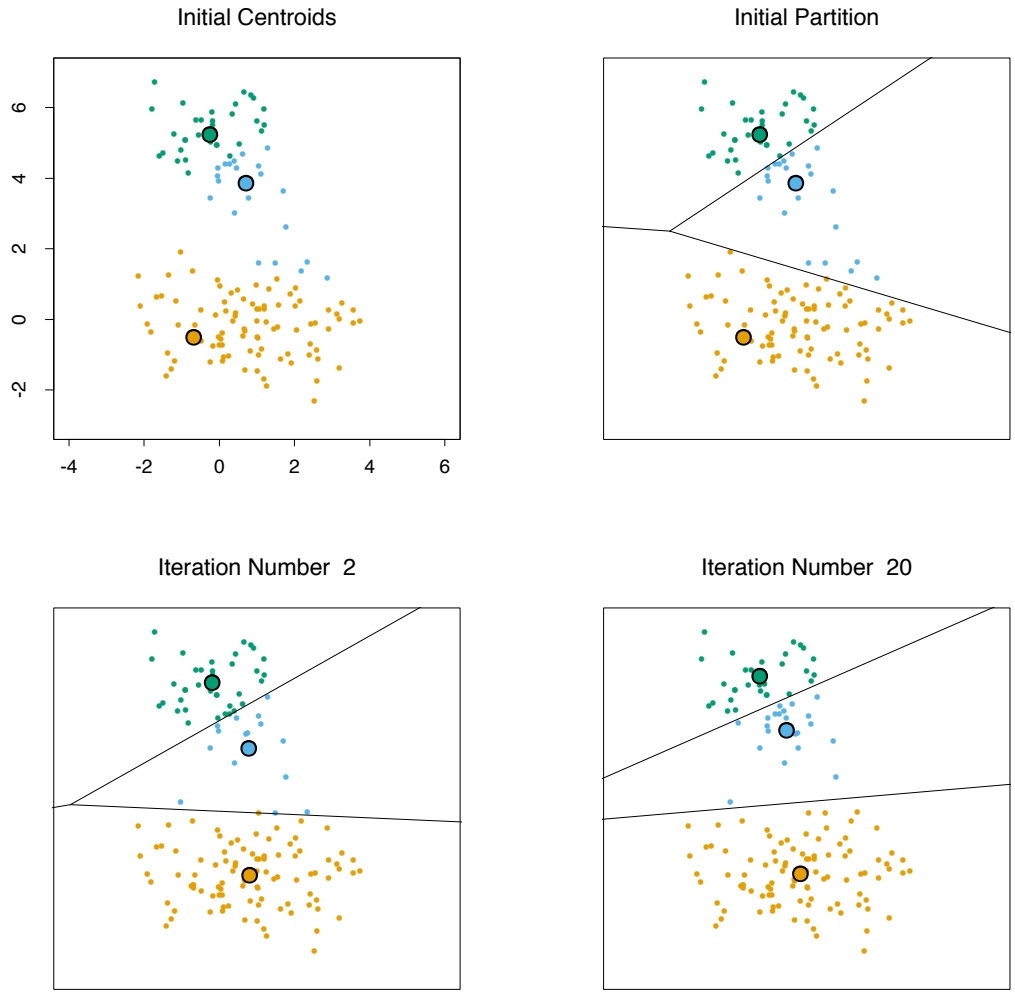
Conclusion: Average linkage leads to most 'balanced' dendrograms
Hastie, Tibshirani, Friedman, *The elements of Statistical Learning* 2008

K-means

Algorithm 'pseudocode'

- Input:
  - K (the number of clusters)
  - I observations in P quantitative dimensions
  - I.e. subjects in the space of variables, or variables in the space of subjects
- Randomly assign a number from 1 to K to the observations. These are initial clusters.
- Iterate until no more changes in clusters
  - For each of the K clusters, compute its *centroid*, i.e. the mean vector of all the observations in the cluster
  - Assign each observation to the cluster whose centroid is the closest
    - 'Closest' is defined with respect to a metric, typically Eucledian distance
- Output:
  - Allocation of each multivariate observation to a cluster

Example



Hastie, Tibshirani, Friedman, *The elements of Statistical Learning* 2008

More formal details

- Dissimilarity $d(\mathbf{x}_i, \mathbf{x}_{i'})$ between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$
  - Assume d additive in features: $d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{P} d_{i,i',j}$
  - E.g. Euclidean distance: $d_{i,i',j} = (x_{ij} - x_{i'j})^2$
- K-means partitions observations into K sets
  - minimize the sum of average within-cluster dissimilarities:

$$W_K = \sum_{k=1}^{K} \frac{1}{2\, n_k} \sum_{i,i' \in C_k} d(\mathbf{x}_i, \mathbf{x}_{i'}), \text{ where } i, i' \text{ are in cluster } k$$

and n_k is the # of observations in cluster k

  - Equivalently, minimizes pooled within-cluster sum of squares:

$$WCSS_K = \sum_{k=1}^{K} \sum_{i:i \in C_k} \sum_{p=1}^{P} (x_{ip} - \bar{x}^k_{\cdot p})^2$$

  - If observations are multivariate Normal, then $W_k = -\text{logLikelihood}$