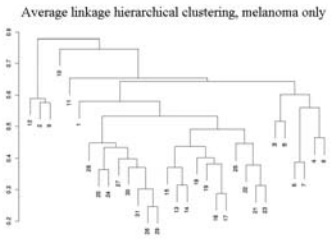
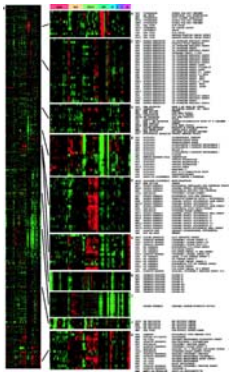
Multivariate Methods

Cluster Analysis





http://www.isrec.isb-sib.ch/~darlene/EMBnet/

Classification

- Historically, *objects* are classified into *groups*
  - periodic table of the elements (chemistry)
  - taxonomy (zoology, botany)
- Why classify?
  - organizational convenience, convenient summary
  - prediction
  - explanation
- *Note:* these aims do not necessarily lead to the same classification; e.g. *SIZE* of object in hardware store vs. *TYPE/USE* of object

Classification, cont

- Classification divides objects into groups based on a set of values
- Unlike a theory, a classification is neither true nor false, and should be judged largely on the usefulness of results (Everitt)
- However, a classification (clustering) may be useful for suggesting a theory, which could then be tested

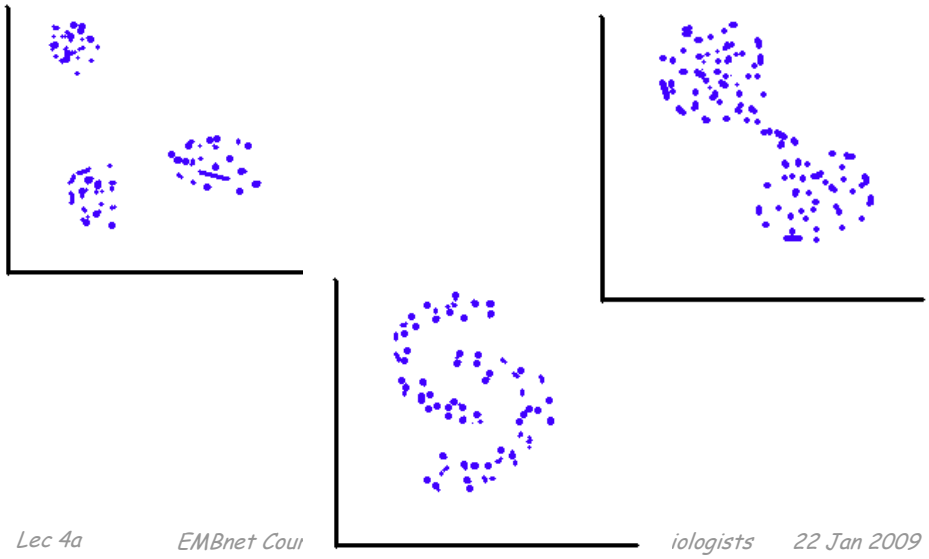
Classification

- *Task:* assign objects to classes (groups) on the basis of measurements made on the objects
- *Supervised:* classes are predefined, want to use a (training or learning) set of labeled objects to form a classifier for classification of future observations (discrimination analysis)
- *Unsupervised:* classes unknown, want to discover them from the data (cluster analysis)

Cluster analysis

- Addresses the problem: Given *n* objects, each described by *p* variables (or features), derive a useful division into a number of classes
- Often want a *partition* of objects
  - But also 'fuzzy clustering'
  - Could also take an exploratory perspective
- 'Unsupervised learning'
- Most clustering is not statistical

Difficulties in defining 'cluster'



Clustering Gene Expression Data

- Can cluster *genes* (rows), e.g. to (attempt to) identify groups of co-regulated genes
- Can cluster *samples* (columns), e.g. to identify tumors based on profiles
- Can cluster *both* rows and columns at the same time

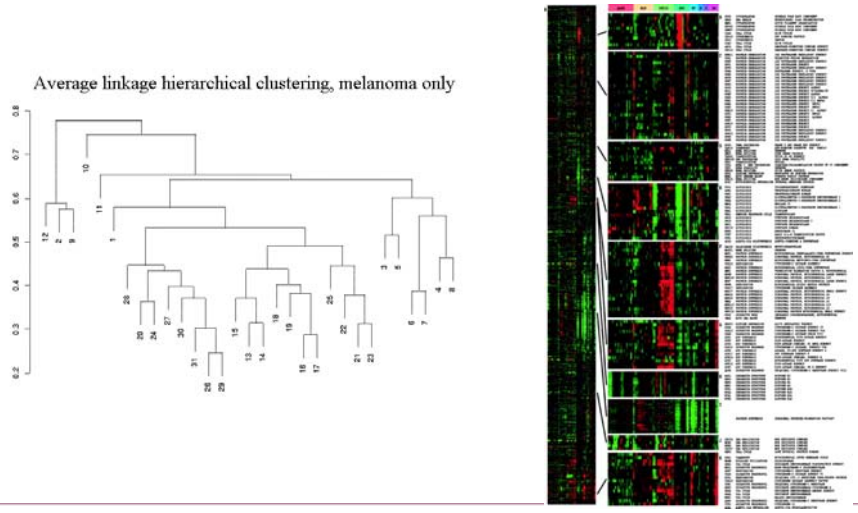
Clustering Gene Expression Data

- Leads to readily interpretable figures
- Can be helpful for identifying patterns in time or space
- Useful (essential?) when *seeking new subclasses* of samples
- Can be used for exploratory, quality assessment purposes

Visualizing Gene Expression Data

- Dendrogram (tree diagram)
- Heat Diagram
  - available as R function `heatmap()`
  - `http://rana.lbl.gov/EisenSoftware.htm`
- Need to *reduce number of genes* first for figures to be legible/interpretable (at most a few hundred genes, not a whole array)
- A visual representation for a given clustering (e.g. dendrogram) is *not unique*
- Beware the influence of representation on apparent structure (e.g. color scheme)

Cluster visualization

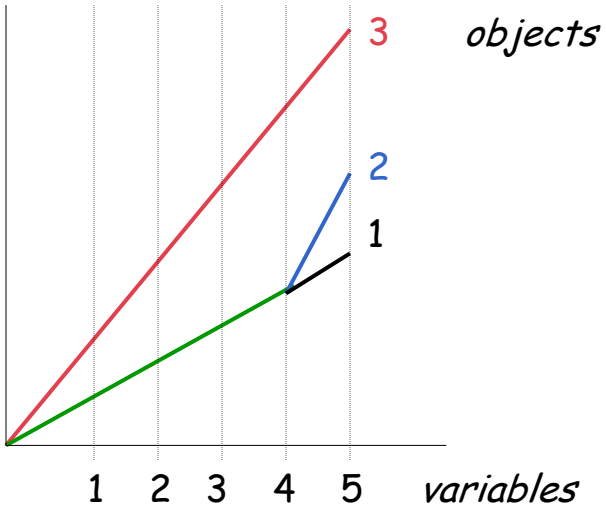


Eisen, Michael B. et al. (1998)
Proc. Natl. Acad. Sci. USA 95, 14863-14868

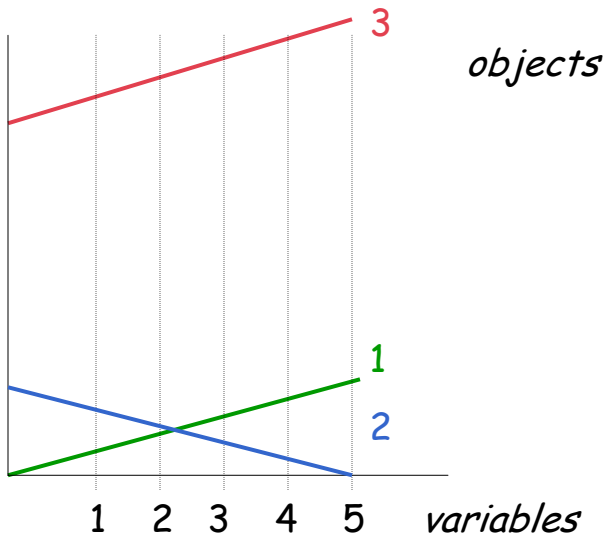
Similarity

- *Similarity* s_{ij} indicates the strength of relationship between two objects i and j
- Usually $0 \leq s_{ij} \leq 1$
- Correlation-based similarity ranges from –1 to 1
- Use of correlation-based similarity is quite common in gene expression studies but is in general contentious...

Problems using correlation



A more extreme example



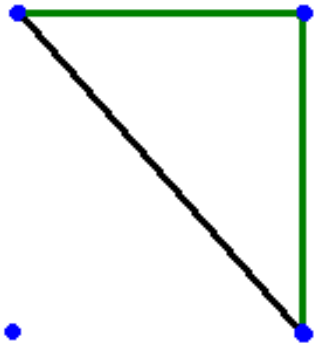
Dissimilarity and Distance

- Associated with similarity measures s_{ij} bounded by 0 and 1 is a *dissimilarity* $d_{ij} = 1 - s_{ij}$
- *Distance* measures have the metric property ($d_{ij} + d_{ik} \geq d_{jk}$)
- Many examples: Euclidean ('as the crow flies'), Manhattan ('city block'), *etc.*
- Distance measure has a large effect on performance
- Behavior of distance measure related to *scale* of measurement

Distance example

Euclidean

Manhattan



What distance should I use?

- This is like asking: *What tool should I buy?*
- It depends on what similarities you are interested in finding
- With Euclidean distance, larger values will tend to dominate; not useful if large value is simply a result of using smaller units (*e.g.*, grams vs Kilos)
- Can get around this (if desired) by scaling or standardizing variables

Partitioning Methods

- Partition the objects into a *prespecified* number of groups K
- Iteratively reallocate objects to clusters until some criterion is met (e.g. minimize within cluster sums of squares)
- Examples: k-means, self-organizing maps (SOM), partitioning around medoids (PAM; more robust and computationally efficient than k-means), model-based clustering

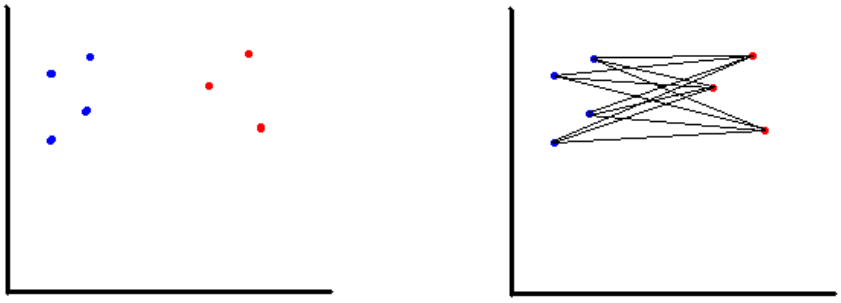
Hierarchical Clustering

- Produce a *dendrogram* (tree diagram)
- Avoid prespecification of the number of clusters K
- The tree can be built in two distinct ways:
  - Bottom-up: *agglomerative* clustering
  - Top-down: *divisive* clustering

Agglomerative Methods

- Start with *n* mRNA sample (or G gene) clusters
- At each step, *merge* the two closest clusters using a measure of between-cluster dissimilarity
- Examples of *between-cluster* dissimilarities:
  - *Average linkage (Unweighted Pair Group Method with Arithmetic Mean (UPGMA)):* average of pairwise dissimilarities
  - *Single-link (NN):* min of pairwise dissimilarities
  - *Complete-link (FN):* max of pairwise dissimilarities

Between cluster distances



Divisive Methods

- Start with only *one* cluster
- At each step, *split* clusters into two parts
- Advantage: Obtain the main structure of the data (*i.e.* focus on upper levels of dendrogram)
- Disadvantage: Computational difficulties when considering all possible divisions into two groups

Partitioning vs. Hierarchical

- *Partitioning*
  - Advantage: Provides clusters that satisfy some optimality criterion (approximately)
  - Disadvantages: Need initial K, long computation time
- *Hierarchical*
  - Advantage: Fast computation (agglomerative)
  - Disadvantages: Rigid, cannot correct later for erroneous decisions made earlier

R: clustering

- A number of R packages contain functions to carry out clustering, including:
  - **stats: hclust**
  - **cluster (Kaufman and Rousseeuw)**
  - **cclust**
  - **mclust**
  - **e1071**

Generic Clustering Tasks

- Estimating number of clusters
- Assigning each object to a cluster
- Assessing strength/confidence of cluster assignments for individual objects
- Assessing cluster homogeneity
- *(Interpretation of the resulting clusters)*

Estimating how many clusters

- Many suggestions for how to decide this!
- Indices based on homogeneity and/or separation (within and between cluster sums of squares)
- Milligan and Cooper (Psychometrika 50:159-179, 1985) studied performance of 30 such methods in a large simulation
- R package `fpc` (Christian Hennig) has function `cluster.stats` which computes many of these

Additional methods

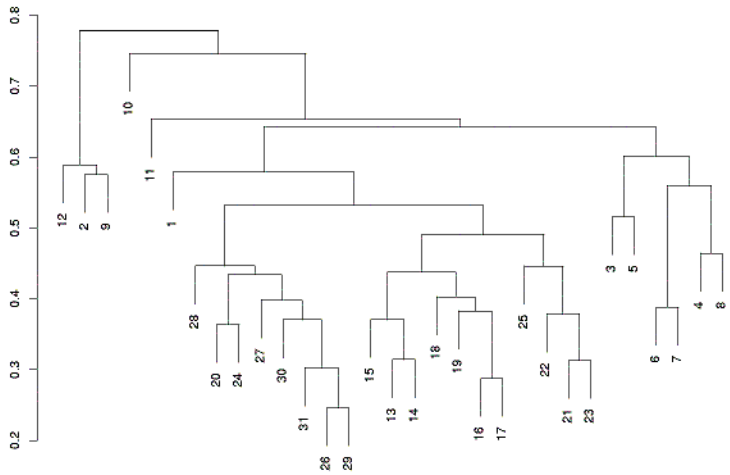
- Model-based criteria (AIC, BIC, MDL) when using model-based clustering
- GAP, GAP-PC (Tibshirani et al.)
- Average silhouette width (Kaufman and Rousseuw)
- mean silhouette split (Pollard and van der Laan)
- clest (Dudoit and Fridlyand)

Example: Bittner et al.

It has been proposed (by many) that a *cancer taxonomy* can be identified from *gene expression experiments*.

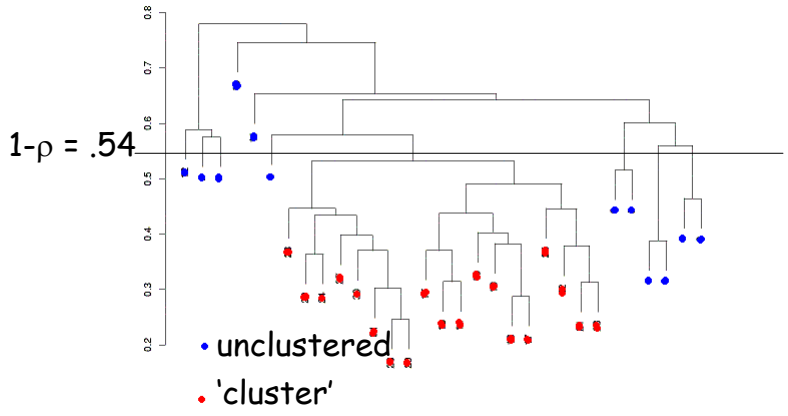
- 31 melanomas (from a variety of tissues/cell lines)
- 7 controls
- 8150 cDNAs
- 6971 unique genes
- 3613 genes 'strongly detected'

Average linkage hierarchical clustering, melanoma only



How many clusters are present?

Average linkage, melanoma only



Issues in Clustering

- Pre-processing (Image analysis and Normalization)
- Which *variables* are used
- Which *samples* are used
- Which *distance measure* is used
- Which *algorithm* is applied
- How to decide the *number of clusters K*

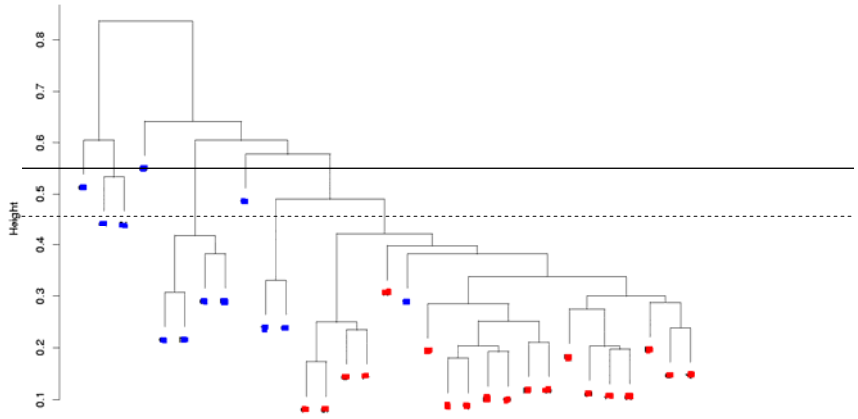
Issues in Clustering

- Pre-processing (Image analysis and Normalization)
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

Filtering Genes

- All genes (i.e. don't filter any)
- At least k (or a proportion p) of the samples must have expression values larger than some specified amount, *A*
- Genes showing 'sufficient' variation
  - a gap of size *A* in the central portion of the data
  - a interquartile range of at least *B*
  - 'large' SD, CV, ...

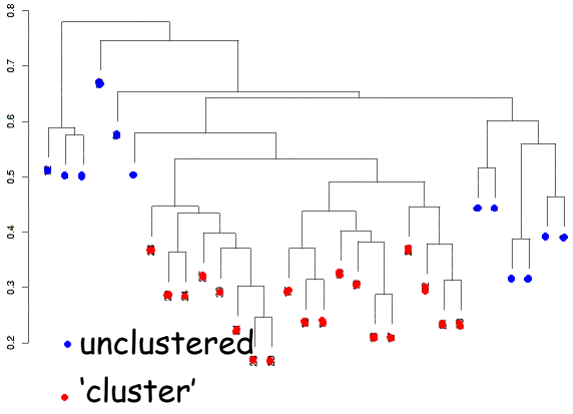
Average linkage, top 300 genes in SD



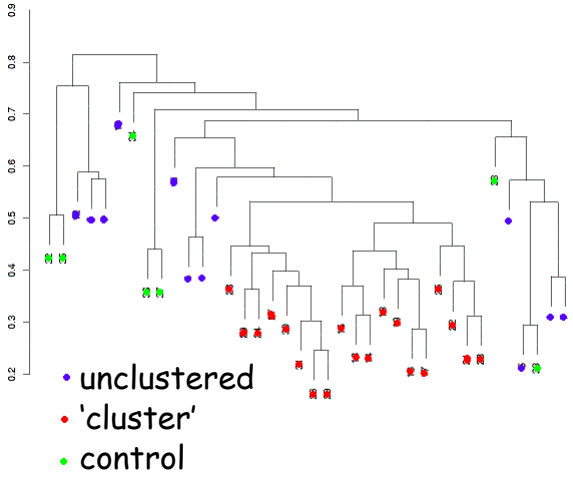
Issues in Clustering

- Pre-processing (Image analysis and Normalization)
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

Average linkage, *melanoma only*



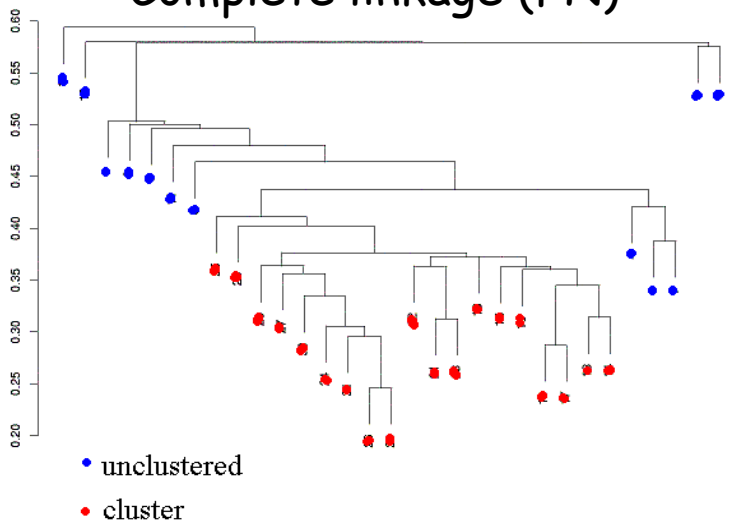
Average linkage, *melanoma & controls*



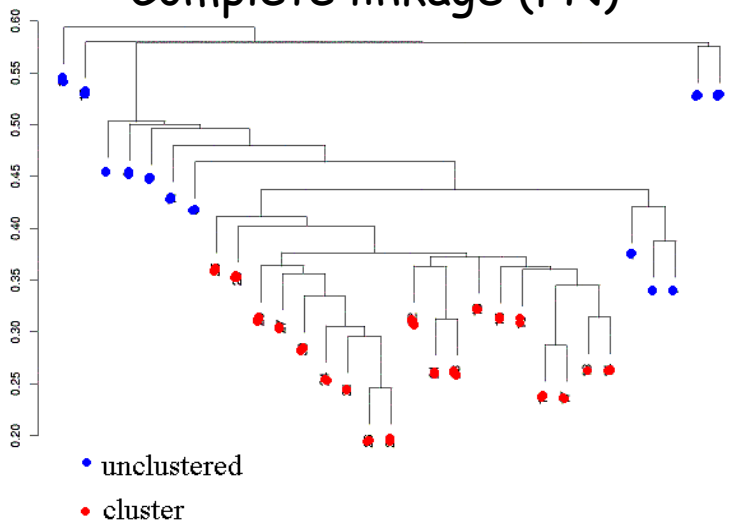
Issues in clustering

- Pre-processing
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

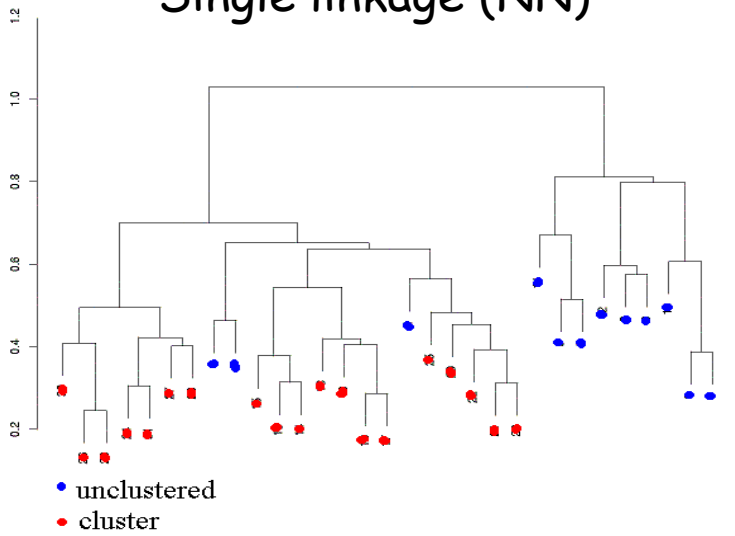
Complete linkage (FN)



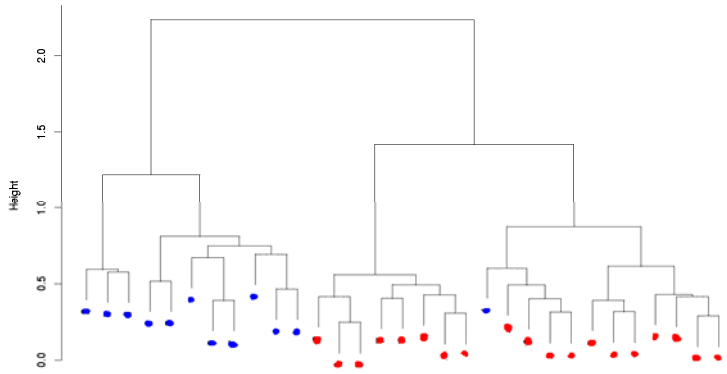
Complete linkage (FN)



Single linkage (NN)



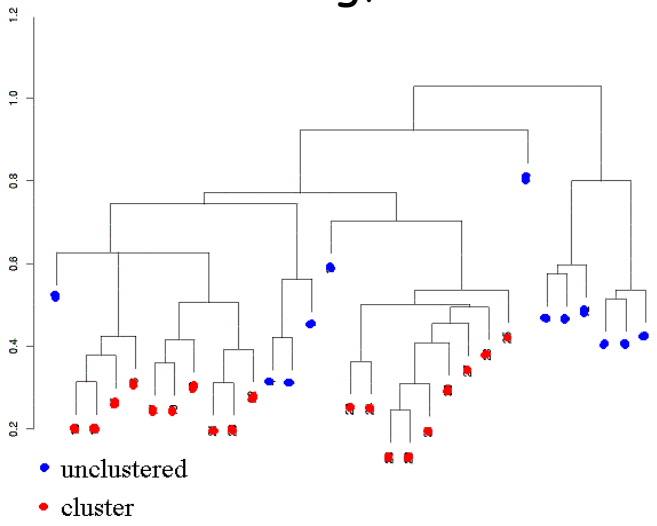
Ward's method (information loss)



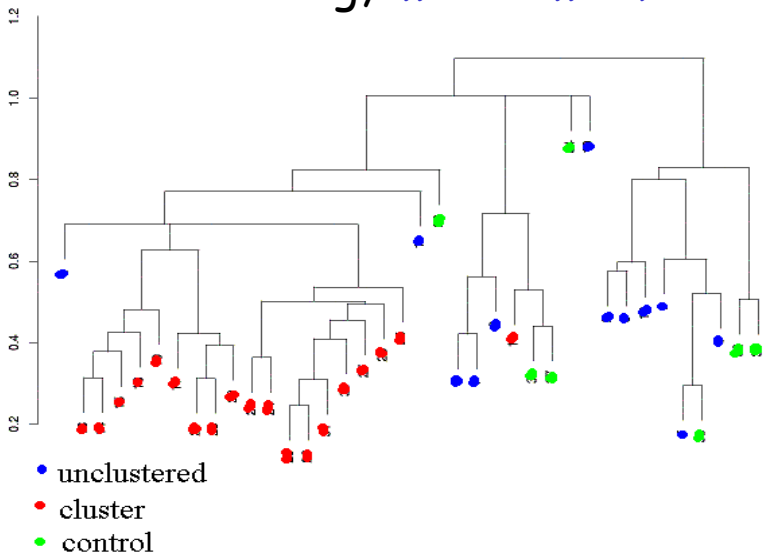
Issues in clustering

- Pre-processing
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

Divisive clustering, *melanoma only*



Divisive clustering, *melanoma & controls*



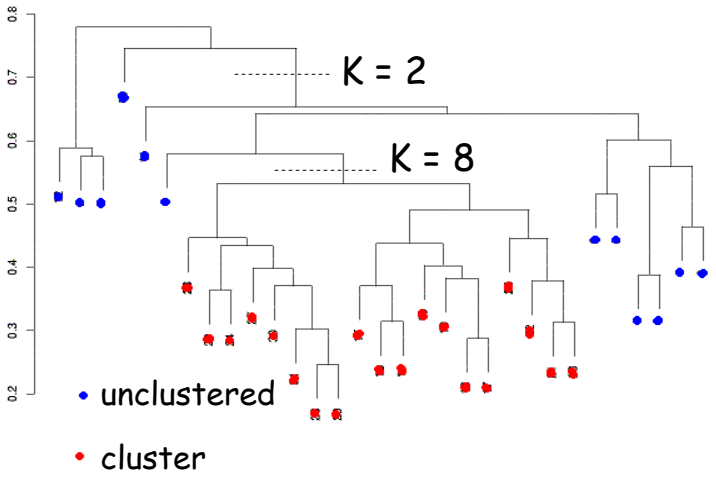
Issues in clustering

- Pre-processing
- Which genes (variables) are used
- Which samples are used
- Which distance measure is used
- Which algorithm is applied
- How to decide the number of clusters *K*

How many clusters *K*?

- Applying several methods yielded estimates of *K* = 2 (largest cluster has 27 members) to *K* = 8 (largest cluster has 19 members)

Average linkage, melanoma only

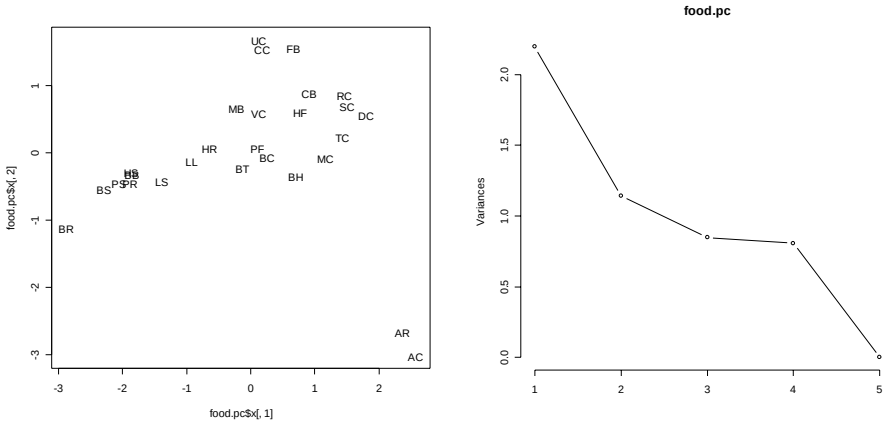


Summary

- Buyer beware – results of cluster analysis should be treated with GREAT CAUTION and ATTENTION TO SPECIFICS, because…
- Many things can vary in a cluster analysis
- If covariates/group labels are known, then clustering is usually inefficient

Multivariate Methods

Principal Components Analysis

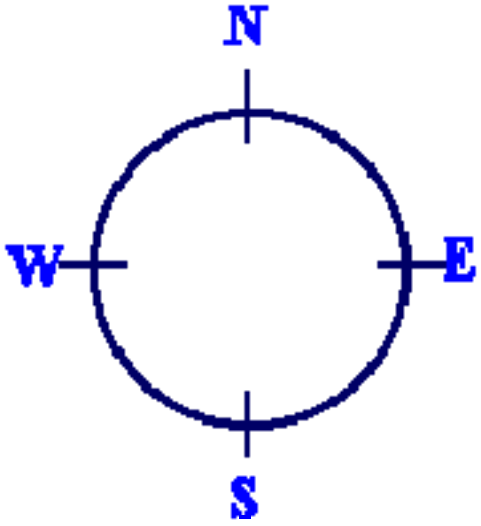


http://www.isrec.isb-sib.ch/~darlene/EMBnet/

Locating a point in the plane

- We can describe the *location of a point in the plane* by saying how much we move in the horizontal (X) direction, then how much we move in the vertical (Y) direction
- As an example, think of describing how to get to some particular place from where you are (for example, how to get from the train station to the BioZentrum)
- One way to do this is to say how far you go NORTH, then how far you go EAST

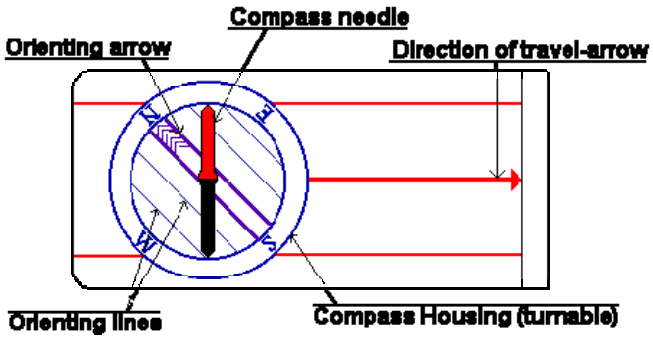
Directions



North = 1st?

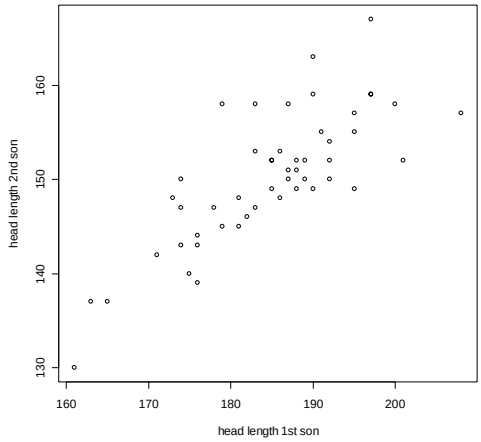
- There is no rule that says we must first say how far to go NORTH – for example, we could instead say first how far to go SOUTH (can think of as 'negative NORTH')
- We could even say first how far to go NORTH-EAST, then how far to go ...

Alternate Directions



A small data set

- Head length (in mm) for each of the first two adult sons in 50 families



Variance-Covariance matrix

- Consider a data set consisting of *p* variables measured on *n* cases
- How the variables change together is summarized by the variance-covariance matrix (or by the correlation matrix)
- For our simple example:

```
> cov(head)
         [,1]     [,2]
[1,] 96.95061 54.48939
[2,] 54.48939 49.57918
```

```
> cor(head)
        [,1]   [,2]
[1,] 1.0000  .7859
[2,] 0.7859  1.0000
```


Principal Component Analysis (PCA)

- One aim of principal component analysis (PCA) is to *reduce the dimensionality* from *p* variables
- This has the effect of *simplifying a dataset*, by reducing multidimensional data to a *lower dimension* (i.e. have smaller number of variables)
- Try to explain the variance-covariance structure through *linear combinations (principal components)* of the (original) variables

Principal Component Analysis (PCA)

- Another aim is to interpret the first few principal components in terms of the original variables to give greater insight into the data structure

More on PCA

- Another aim is to interpret the first few components in terms of the original variables => *greater insight into data structure*
- Each PC *accounts for* a certain amount of the *variation in the data*
- The 1st PC is the *linear combination* that accounts for ('explains') the *most variation*
- Subsequent PCs account for as much as possible of the remaining variation, while being *uncorrelated* with earlier PCs
- *Aubergine ...*

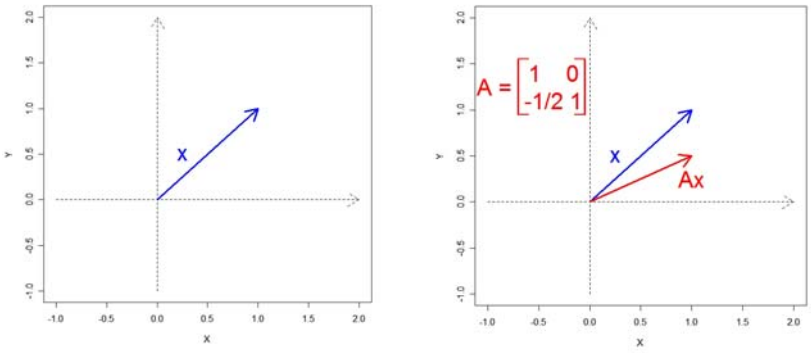
What is a linear combination?

- Say we have 2 variables, height and length
- Create *new variables* from these by summing multiples of the original values
- *Examples:*
  - V = 12*height + 4*length
  - W = π*height – 4.2*length
  - X = -√3*height +0.75*length

What is a linear transformation?

- The main example of a linear transformation is given by *matrix multiplication*
- Say we have a matrix A, of dimension p x p
- We can multiply a vector x by A to form a new vector Ax
- For most vectors x, applying A to x changes both the *length* and the *direction* of the original vector

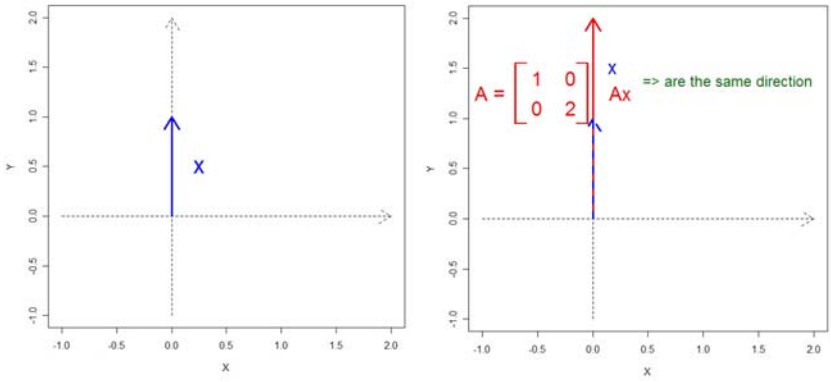
Example linear transformation



Not all vectors are the same

- Usually, applying A to x changes both the length and the direction of the vector
- *But for some special vectors x, the result is just an expansion or contraction with no change of direction (except possibly flipped)*

Another linear transformation



A little (more) linear algebra

- In these cases, the new vector is just a (scalar) multiple λ of the original vector
- The values λ satisfying $Ax = \lambda x$ are called *eigenvalues* (also characteristic values or latent roots) of the matrix A
- The corresponding (nonzero) vector x is called an *eigenvector* (or characteristic vector, latent vector)
- Usually convenient to scale eigenvectors to have length 1 ('unit norm')

What does this have to do with PCA?

- Consider the variance-covariance matrix A of your dataset
- The eigenvectors of A provide sets of coefficients defining *p* linear functions of the original variables
- *These functions are the PCs*
- If A has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ then the PCs have variances $\lambda_1, \lambda_2, \ldots, \lambda_p$ and *zero covariances* (i.e. they are uncorrelated, and are hence giving independent information)

Cautions

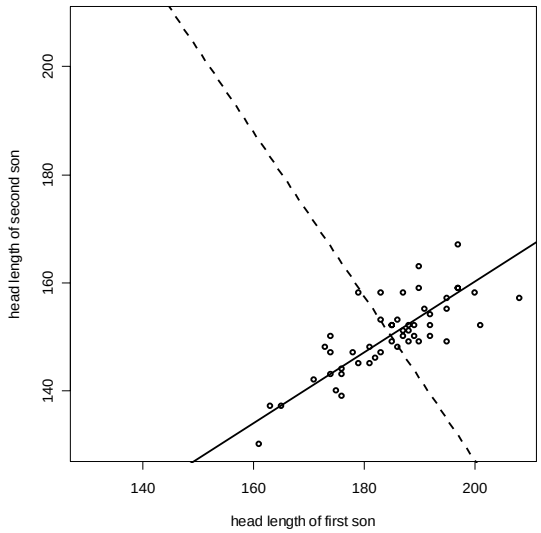
- Sometimes used as a method for *simplifying data* because PCs associated with smaller eigenvalues have smaller variances and might therefore be 'ignored'
- *This assumption requires caution*
- When variables are on *different scales*, it is customary to use the *correlation matrix* (rather than the covariance matrix)
- *These two formulations give different results* : the eigenvalues for the two matrices are not related in a simple way
- Theory not simple for correlation-based PCA

R: PCA (I)

```
> head.pc <- prcomp(head)
> head.pc
Standard deviations:
[1] 11.518663  3.721586

Rotation:
             PC1        PC2
[1,] -0.8362568 -0.5483381
[2,] -0.5483381  0.8362568
```


Principal axes



How many PCs?

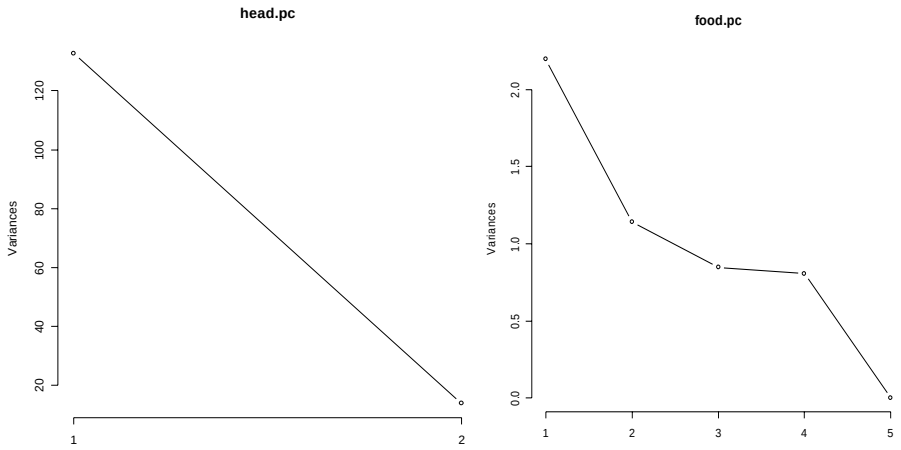
- Retain the number required to *explain some percentage of the total variation* (e.g. 90%)
- *Number of eigenvalues > average* (1 if correlation matrix is used)
- Look for 'elbow' in *scree plot*
  - scree plot shows proportion of variance (or just variance) explained by each component
- Compromise between these

R: PCA (II)

```
> summary(head.pc)
Importance of components:
                            PC1    PC2
Standard deviation     11.519 3.7216
Proportion of Variance  0.905 0.0945
Cumulative Proportion   0.905 1.0000

> screeplot(head.pc,type="lines")
> head.pc$sdev^2
[1] 132.6796  13.8502
> head.pc$sdev^2/sum(head.pc$sdev^2)
[1] 0.90547862 0.09452138
```

R: scree plots



(BREAK)