Unsupervised Learning

Unsupervised Learning

- Unsupervised vs Supervised Learning:
- Most of this course focuses on supervised learning methods such as regression and classification.
- In that setting we observe both a set of features $X_1, X_2, . . . , X_p$, for each object, as well as a response or outcome variable Y . The goal is then to predict Y using $X_1, X_2, . . . , X_p$.
- Here we instead focus on unsupervised learning, we where observe only the features $X_1, X_2, . . . , X_p$. We are not interested in prediction, because we do not have an associated response variable Y .

The Goals of Unsupervised Learning

- The goal is to discover interesting things about the measurements: is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?
- We discuss two methods:
  - principal components analysis, a tool used for data visualization or data pre-processing before supervised techniques are applied, and
  - clustering, a broad class of methods for discovering unknown subgroups in data.

The Challenge of Unsupervised Learning

- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.
- But techniques for unsupervised learning are of growing importance in a number of fields:
    - subgroups of breast cancer patients grouped by their gene expression measurements,
    - groups of shoppers characterized by their browsing and purchase histories,
    - movies grouped by the ratings assigned by movie viewers.

Another advantage

- It is often easier to obtain unlabeled data — from a lab instrument or a computer — than labeled data, which can require human intervention.
- For example it is difficult to automatically assess the overall sentiment of a movie review: is it favorable or not?

Principal Components Analysis

- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

Principal Components Analysis: details

- The first principal component of a set of features $X_1, X_2, . . . , X_p$is the normalized linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

- that has the largest variance. By P normalized, we mean that

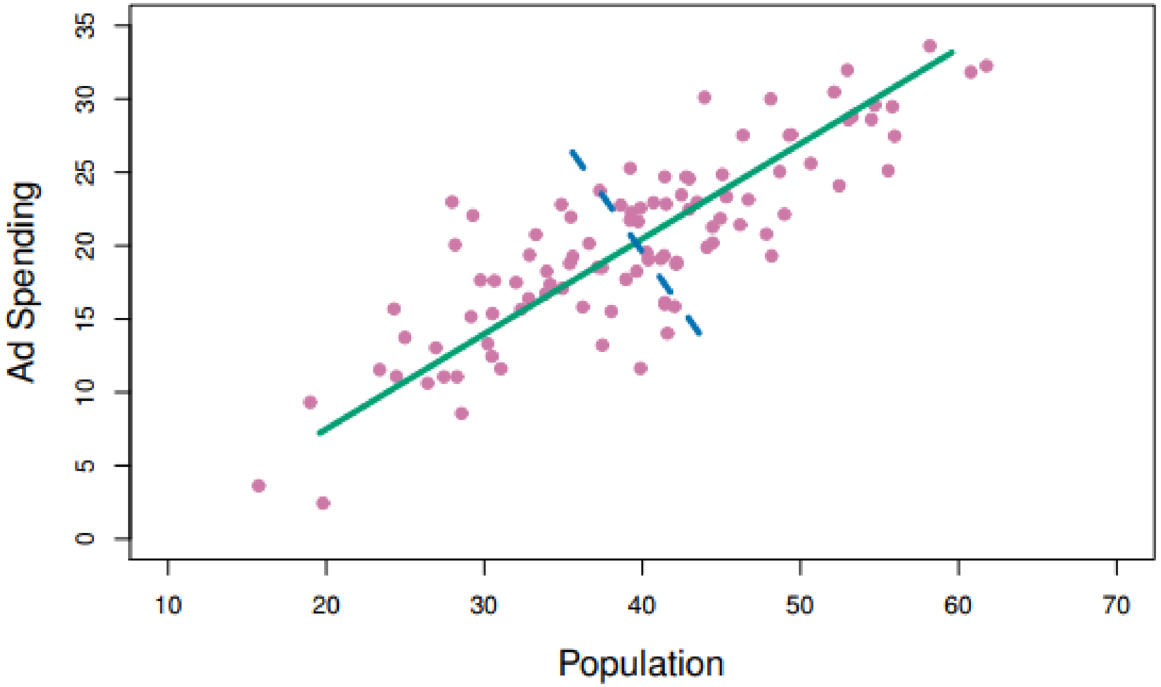
$\sum_{j=1}^{p} \phi_{j1}^2 = 1.$

- We refer to the elements $\phi_{11}, \ldots, \phi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector,

$\phi_1 = (\phi_{11}\ \phi_{21}\ \ldots\ \phi_{p1})^T.$

- We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

PCA: example



- The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.

Computation of Principal Components

Suppose we have a $n \times p$ data set X. Since we are only interested in variance, we assume that each of the variables in X has been centered to have mean zero (that is, the column means of X are zero).

We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip}$$

for i = 1, . . . , n that has largest sample variance, subject to the constraint that $\sum_{j=1}^{p} \phi_{j1}^2 = 1.$

Since each of the x_{ij} has mean zero, then so does z_{ij} (for any values of ϕ_{ij}). Hence the sample variance of the z_{i1} can be written as $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2.$

Computation: continued

Plugging in (1) the first principal component loading vector solves the optimization problem

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1.$$

This problem can be solved via a singular-value decomposition of the matrix $\boldsymbol{X}$, a standard technique in linear algebra.
We refer to Z_1as the first principal component, with realized values $Z_{11}, \ldots, Z_{n1}$

Geometry of PCA

- The loading vector ϕ_1 with elements $\phi_{11}, \phi_{21}, . . . , \phi_{p1}$ defines a direction in feature space along which the data vary the most.
- If we project the n data points $x_1, . . . , x_n$ onto this direction, the projected values are the principal component scores $z_{11}, . . . , \mathrm{z}_{n1}$ themselves.

Further principal components

- The second principal component is the linear combination of $X_1, X_2, . . . , X_p$ that has maximal variance among all linear combinations that are uncorrelated with Z_1.
- The second principal component scores $z_{12}, z_{22},. . . , z_{n2}$ take the form

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \ldots + \phi_{p2}x_{ip},$$

- where ϕ_2 is the second principal component loading vector, with elements $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$.

Further principal components: continued

- It turns out that constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction ϕ_2 to be orthogonal (perpendicular) to the direction 1. And so on.
- The principal component directions ϕ_1, ϕ_2, ϕ_3, . . . are the ordered sequence of right singular vectors of the matrix X, and the variances of the components are $\frac{1}{n}$ times the squares of the singular values. There are at most $\min(n-1, p)$ principal components.

Illustration

- USAarrests data: For each of the fifty states in the United States, the data set contains the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape. We also record UrbanPop (the percent of the population in each state living in urban areas).
- The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

USAarrests data: PCA plot

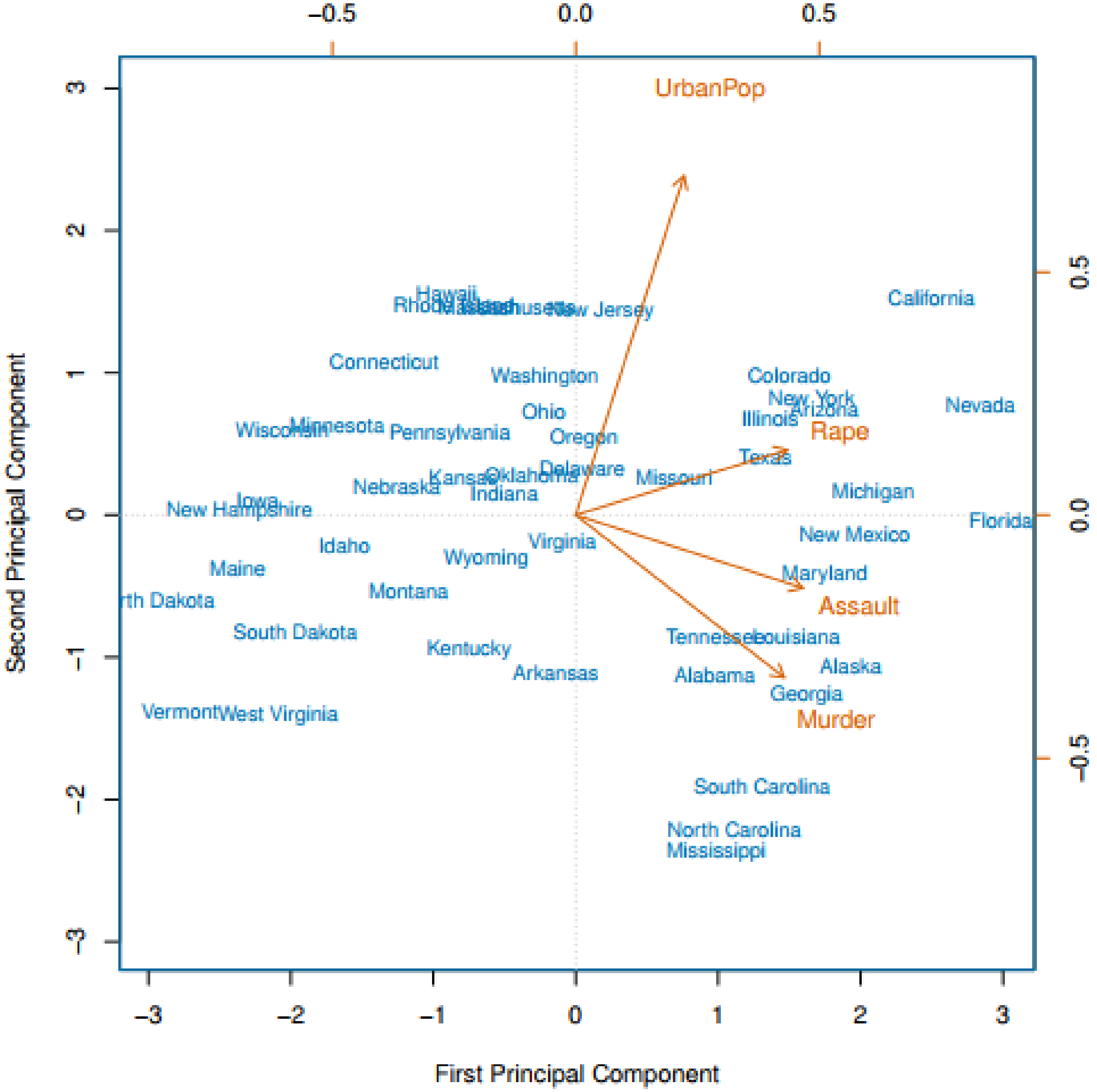
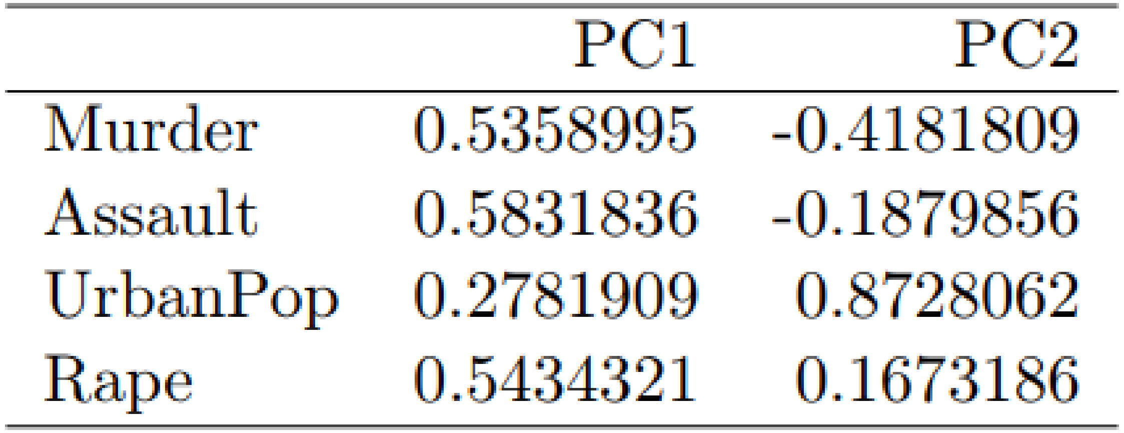


Figure details

The first two principal components for the USArrests data.

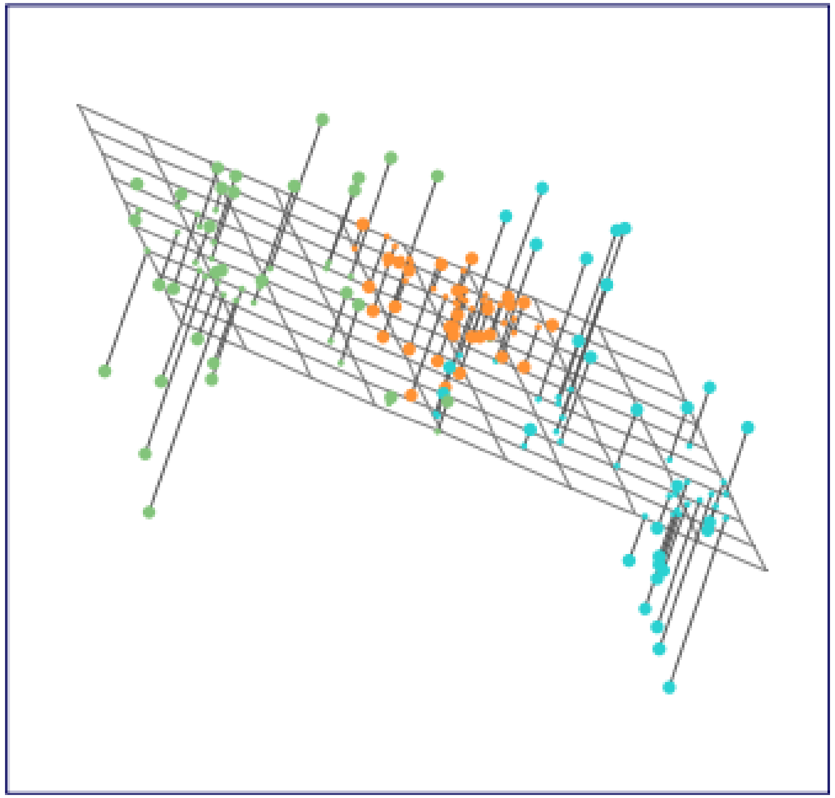
- The blue state names represent the scores for the first two principal components.
- The orange arrows indicate the first two principal component loading vectors (with axes on the top and right). For example, the loading for Rape on the first component is 0.54, and its loading on the second principal component 0.17 [the word Rape is centered at the point (0.54, 0.17)].
- This figure is known as a biplot, because it displays both the principal component scores and the principal component loadings.

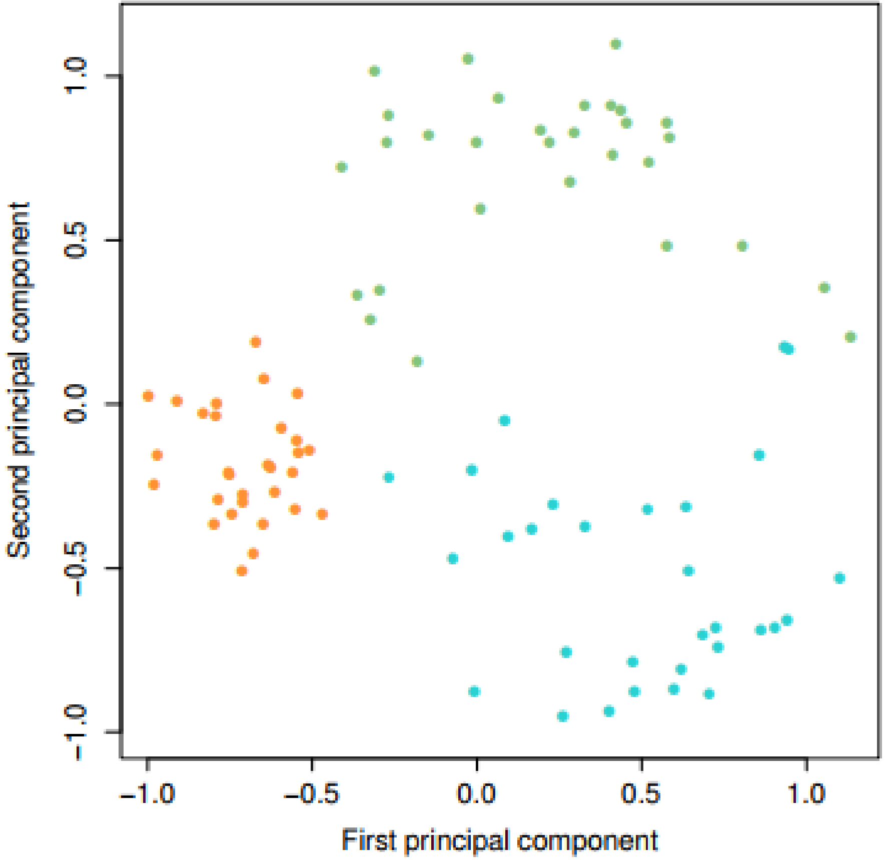
PCA loadings



	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Another Interpretation of Principal Components



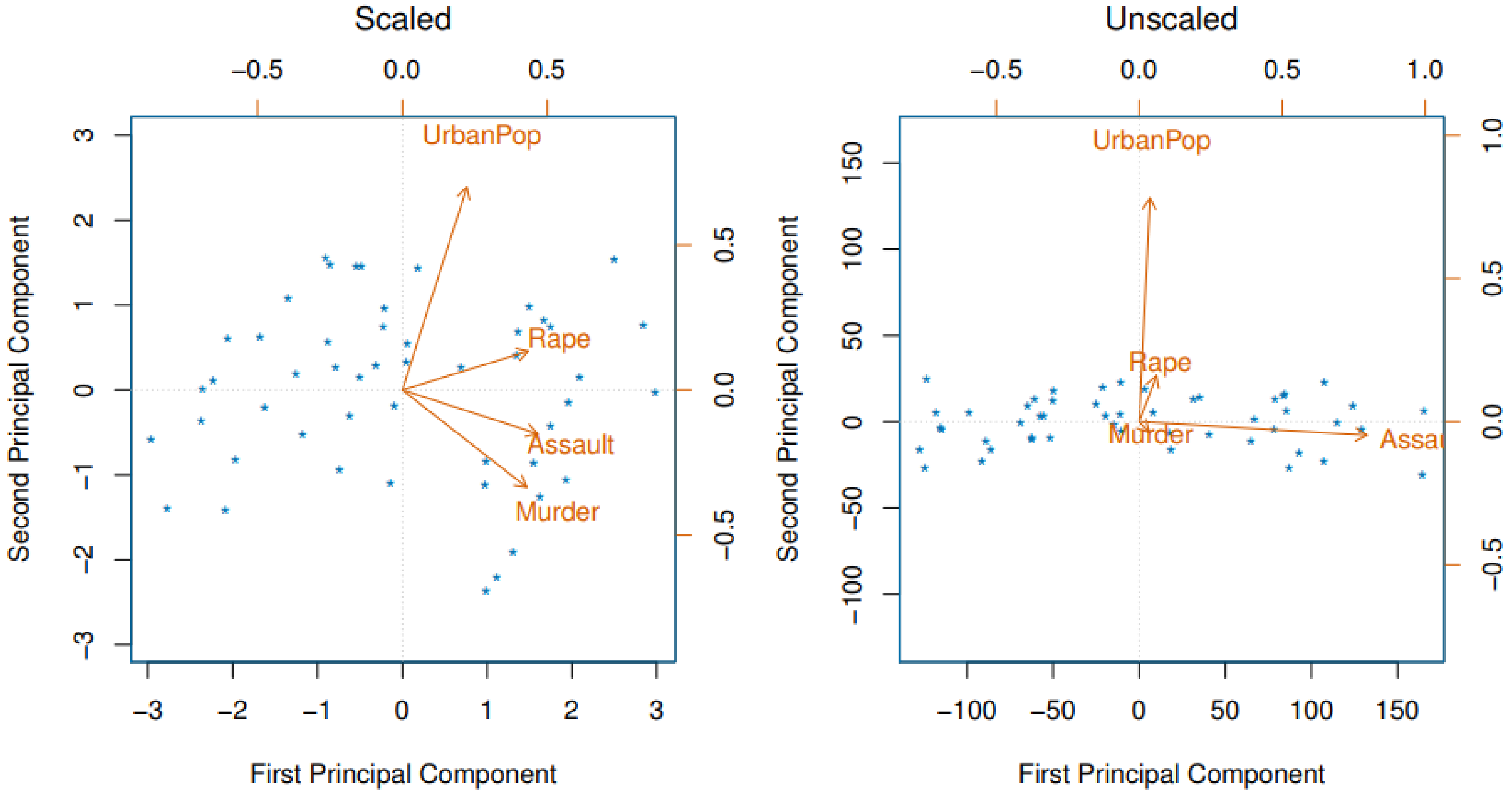


PCA find the hyperplane closest to the observations

- The first principal component loading vector has a very special property: it defines the line in p-dimensional space that is closest to the n observations (using average squared Euclidean distance as a measure of closeness)
- The notion of principal components as the dimensions that are closest to the n observations extends beyond just the first principal component.
- For instance, the first two principal components of a data set span the plane that is closest to the n observations, in terms of average squared Euclidean distance.

Scaling of the variables matters

- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.
- If they are in the same units, you might or might not scale the variables.



Kernel Trick

- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.
- The total variance present in a data set (assuming that the variables have been centered to have mean zero) is defined as

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2,$$

and the variance explained by the mth principal component is

$$\mathrm{Var}(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2.$$

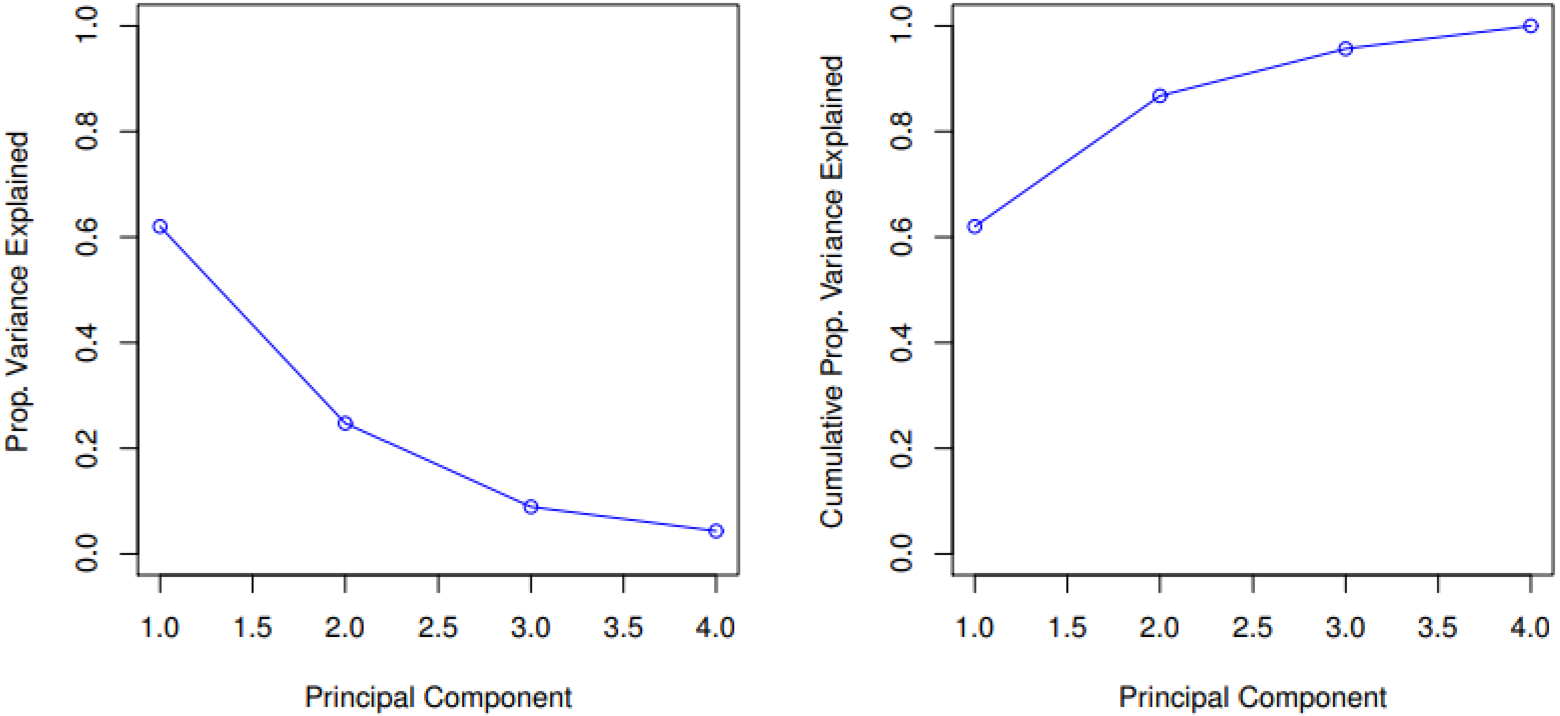
It can be shown that $\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{m=1}^{M} \mathrm{Var}(Z_m)$,
with $M = \min(n-1, p)$.

Proportion Variance Explained: continued

- Therefore, the PVE of the mth principal component is given by the positive quantity between 0 and 1

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}.$$

- The PVEs sum to one. We sometimes display the cumulative PVEs.

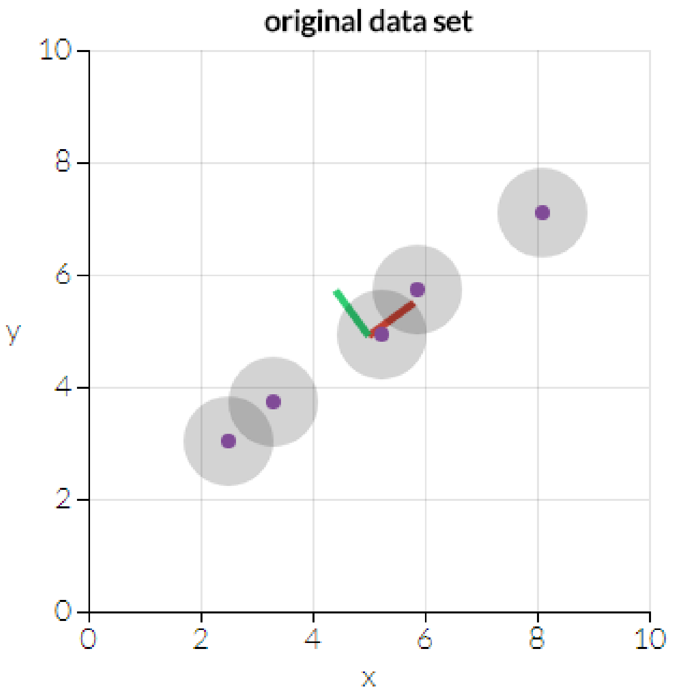


How many principal components should we use?

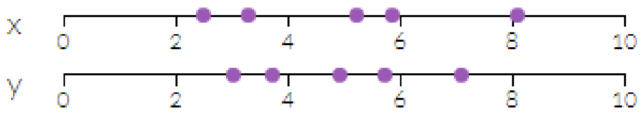
If we use principal components as a summary of our data, how many components are sufficient?

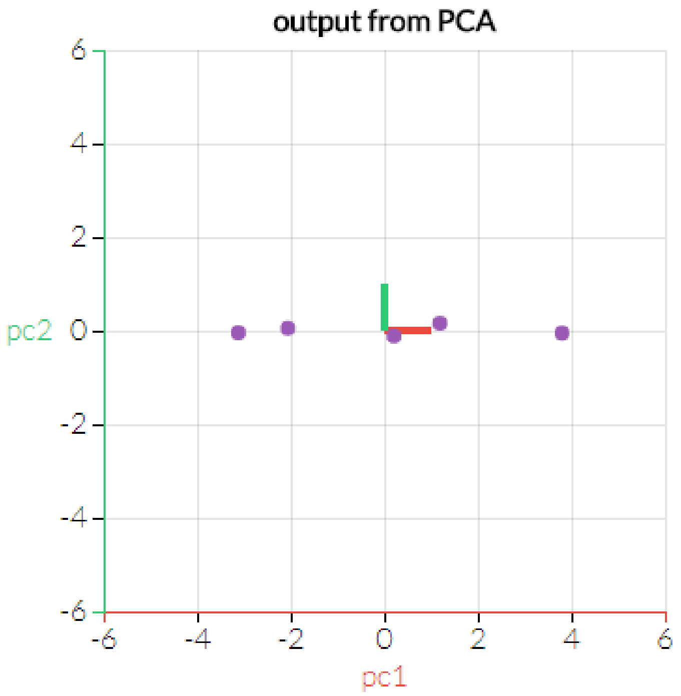
- No simple answer to this question, as cross-validation is not available for this purpose.
  - Why not?
  - When could we use cross-validation to select the number of components?
- the “scree plot” on the previous slide can be used as a guide: we look for an “elbow”.

Example

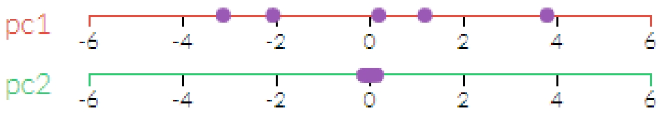

original data set

PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.

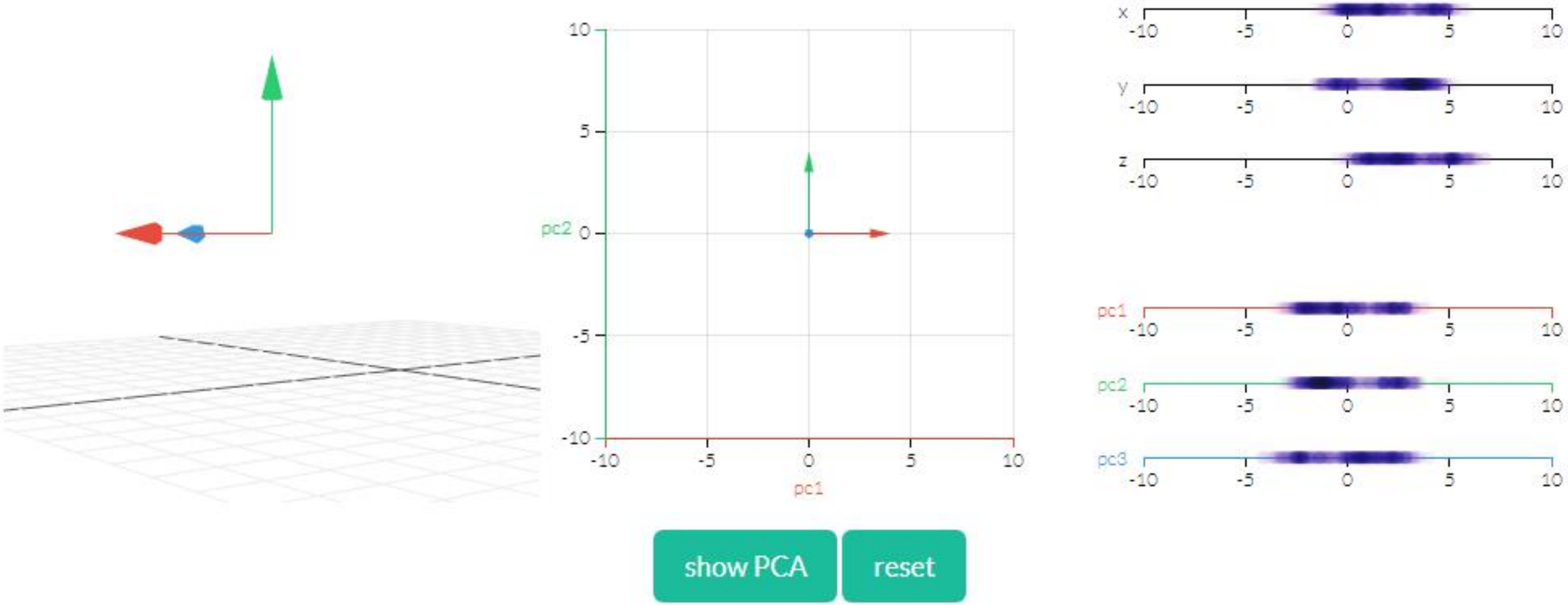



output from PCA

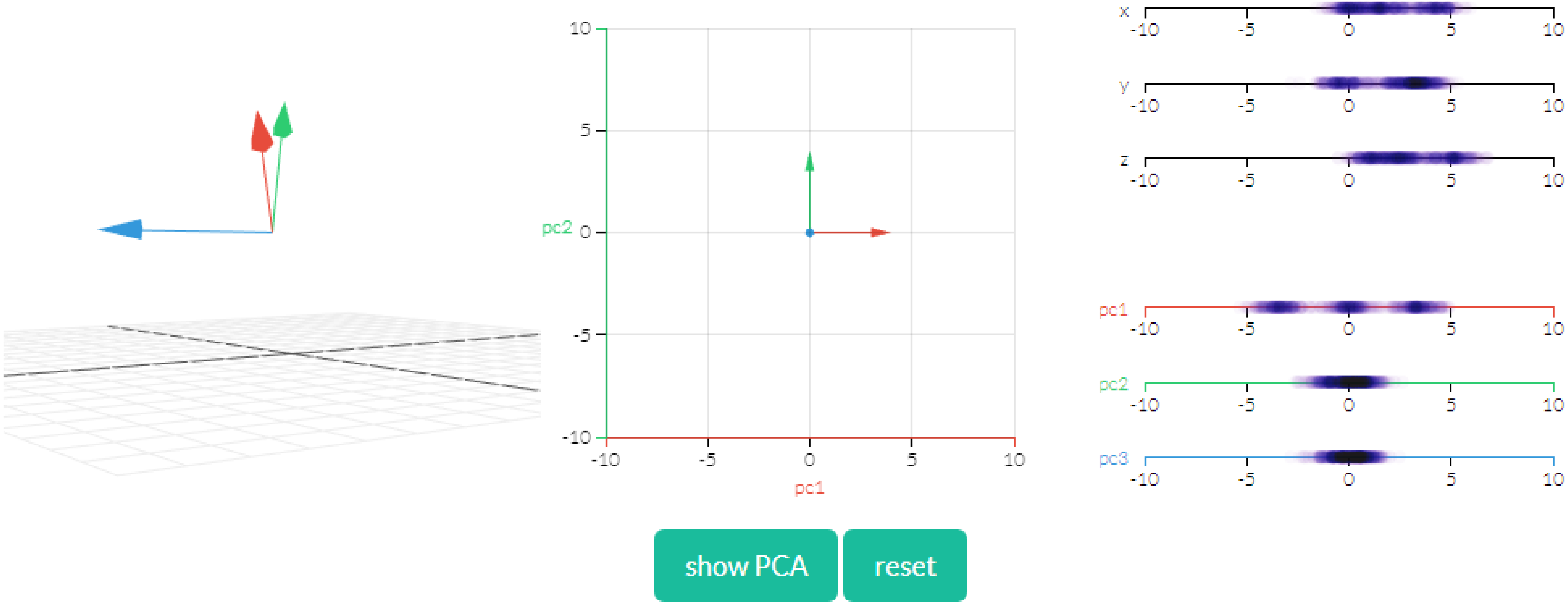
If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.



3D example


10
5
pc2 0
-5
-10
-10
-5
0
5
10
pc1
x
y
z
pc1
pc2
pc3
-10
-5
0
5
10
show PCA
reset

3 D example


pc2
10
5
0
-5
-10
-10
-5
0
5
10
pc1
x
y
z
pc1
pc2
pc3
-10
-5
0
5
10
show PCA
reset

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other,
- It make this concrete, we must define what it means for two or more observations to be similar or different.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

PCA vs Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.

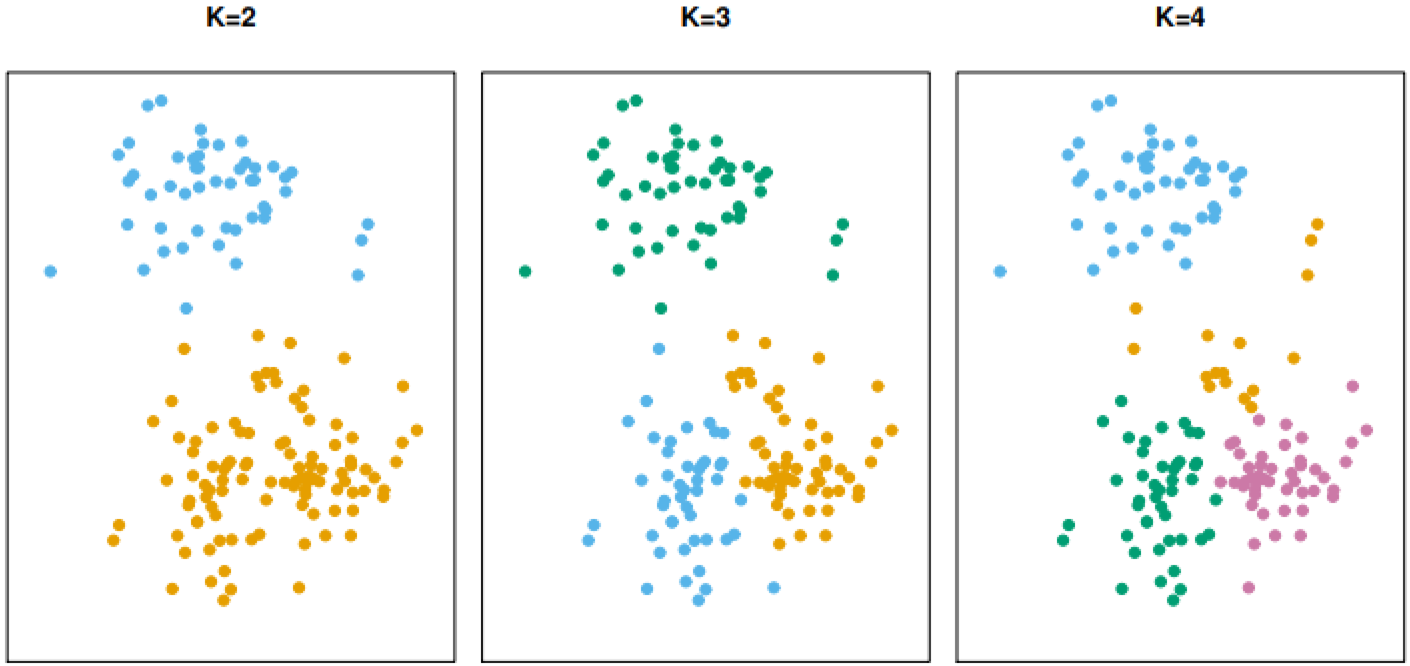
Clustering for Market Segmentation

- Suppose we have access to a large number of measurements (e.g. median household income, occupation, distance from nearest urban area, and so forth) for a large number of people.
- Our goal is to perform market segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product.
- The task of performing market segmentation amounts to clustering the people in the data set.

Two clustering methods

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- In hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.

K-means clustering



- A simulated data set with 150 observations in 2-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.

Details of K-means clustering

- Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:
  1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.
  2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.
- For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

Details of K-means clustering: continued

- The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.
- The within-cluster variation for cluster $\mathrm{C_k}$ is a measure WCV($\mathrm{C_k}$) of the amount by which the observations within a cluster differ from each other.
- Hence we want to solve the problem

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \mathrm{WCV}(C_k) \right\}.$$

- In words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

How to define within-cluster variation?

- Typically we use Euclidean distance

$$\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

- where $|\mathrm{C_k}|$ denotes the number of observations in the kth cluster.
- Combining (2) and (3) gives the optimization problem that defines K-means clustering,

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

K-Means Clustering Algorithm

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
    1. For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
    2. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Properties of the Algorithm

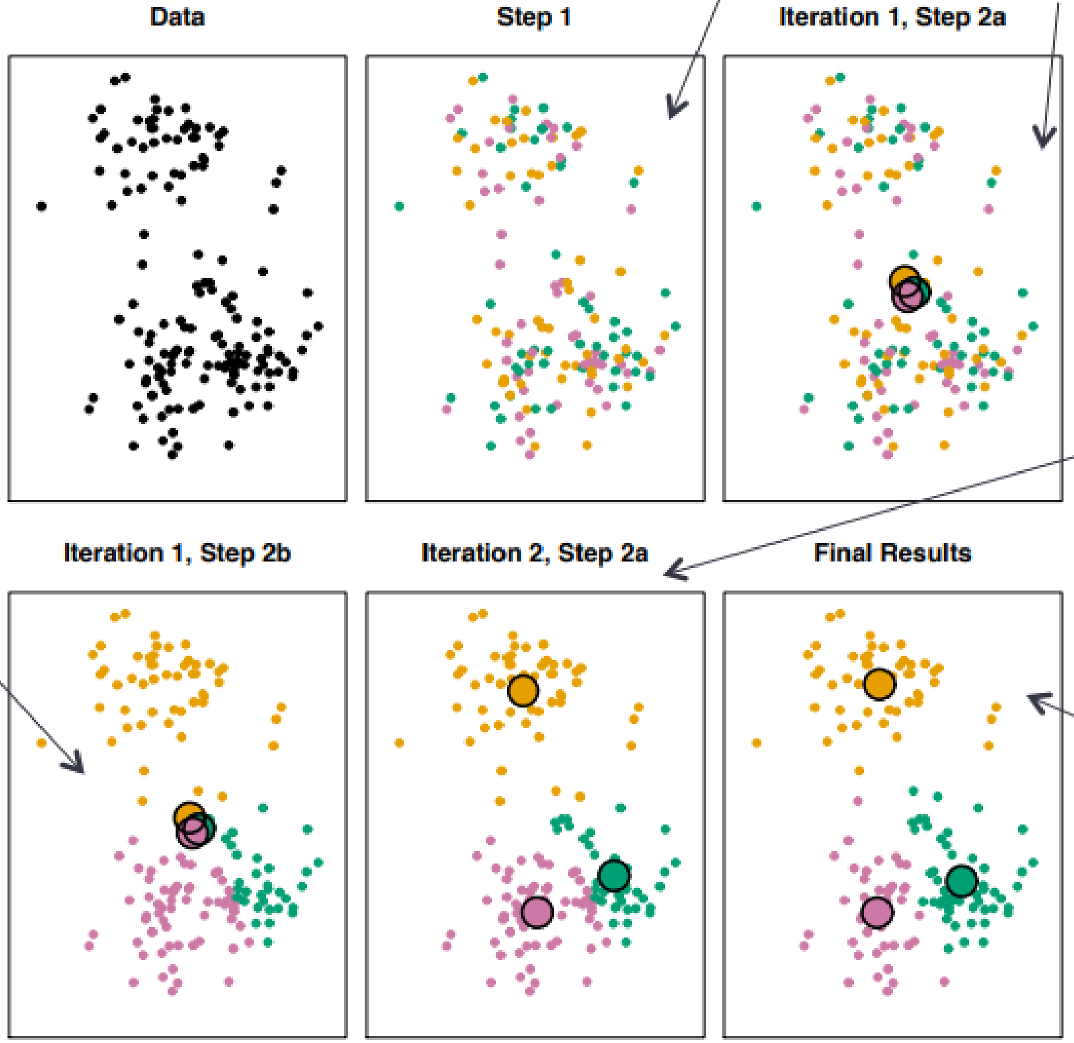
- This algorithm is guaranteed to decrease the value of the objective (4) at each step. Why? Note that

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2,$$

where $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ is the mean for feature *j* in cluster C_k

- however it is not guaranteed to give the global minimum. Why not?

Example



Random Assignment of points

Compute cluster centers from initial assignments

Computer new cluster centers

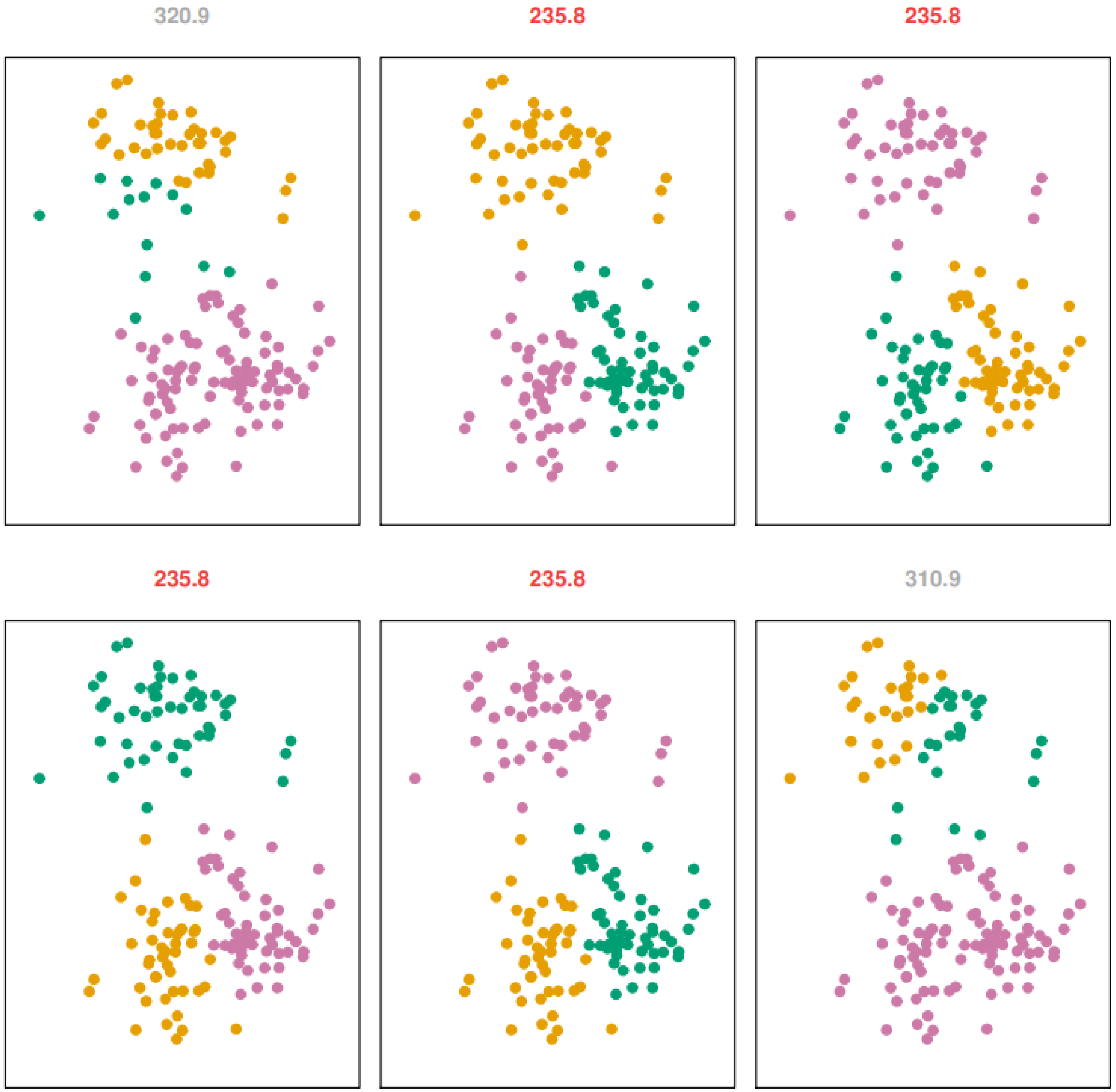
Assign points to closest cluster center

Now there is no further change so stop

Details of Previous Figure

- The progress of the K-means algorithm with K=3.
- Top left: The observations are shown.
- Top center: In Step 1 of the algorithm, each observation is randomly assigned to a cluster.
- Top right: In Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.
- Bottom left: In Step 2(b), each observation is assigned to the nearest centroid.
- Bottom center: Step 2(a) is once again performed, leading to new cluster centroids.
- Bottom right: The results obtained after 10 iterations.

Example: different starting values



Details of Previous Figure

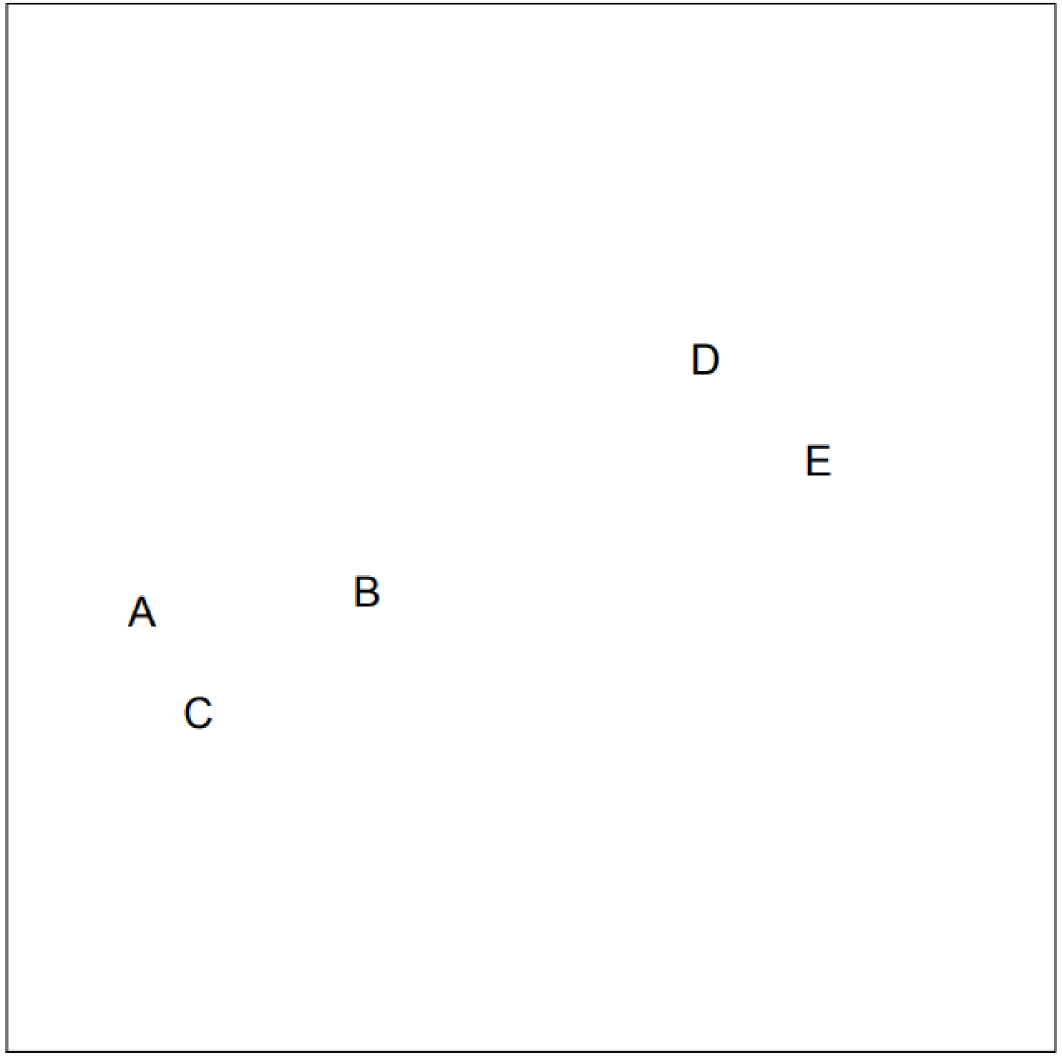
- K-means clustering performed six times on the data from previous figure with K = 3, each time with a different random assignment of the observations in Step 1 of the K-means algorithm.
- Above each plot is the value of the objective (4).
- Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides better separation between the clusters.
- Those labeled in red all achieved the same best solution, with an objective value of 235.8

Hierarchical Clustering

- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage (later we discuss strategies for choosing K)
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.
- In this section, we describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

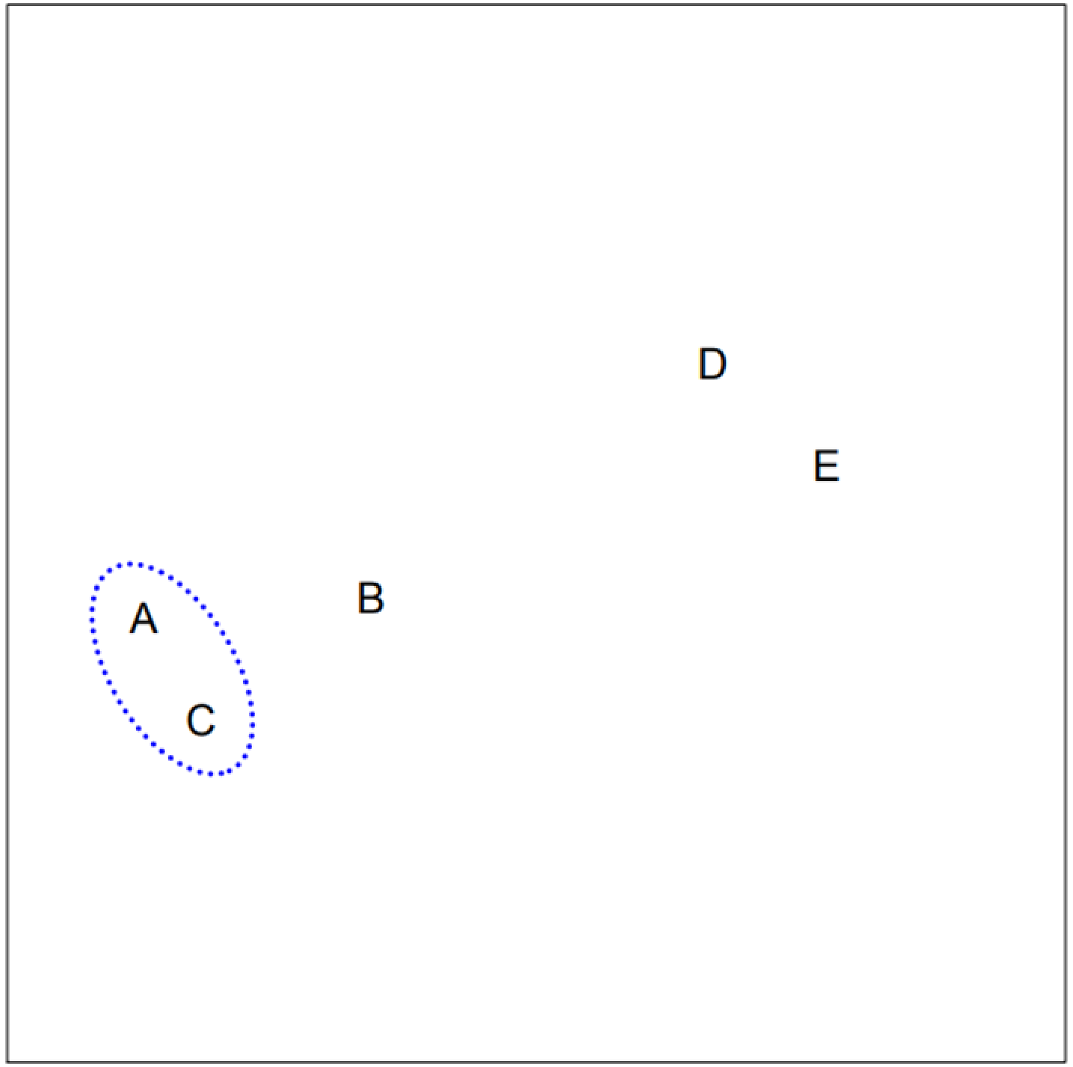
Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...



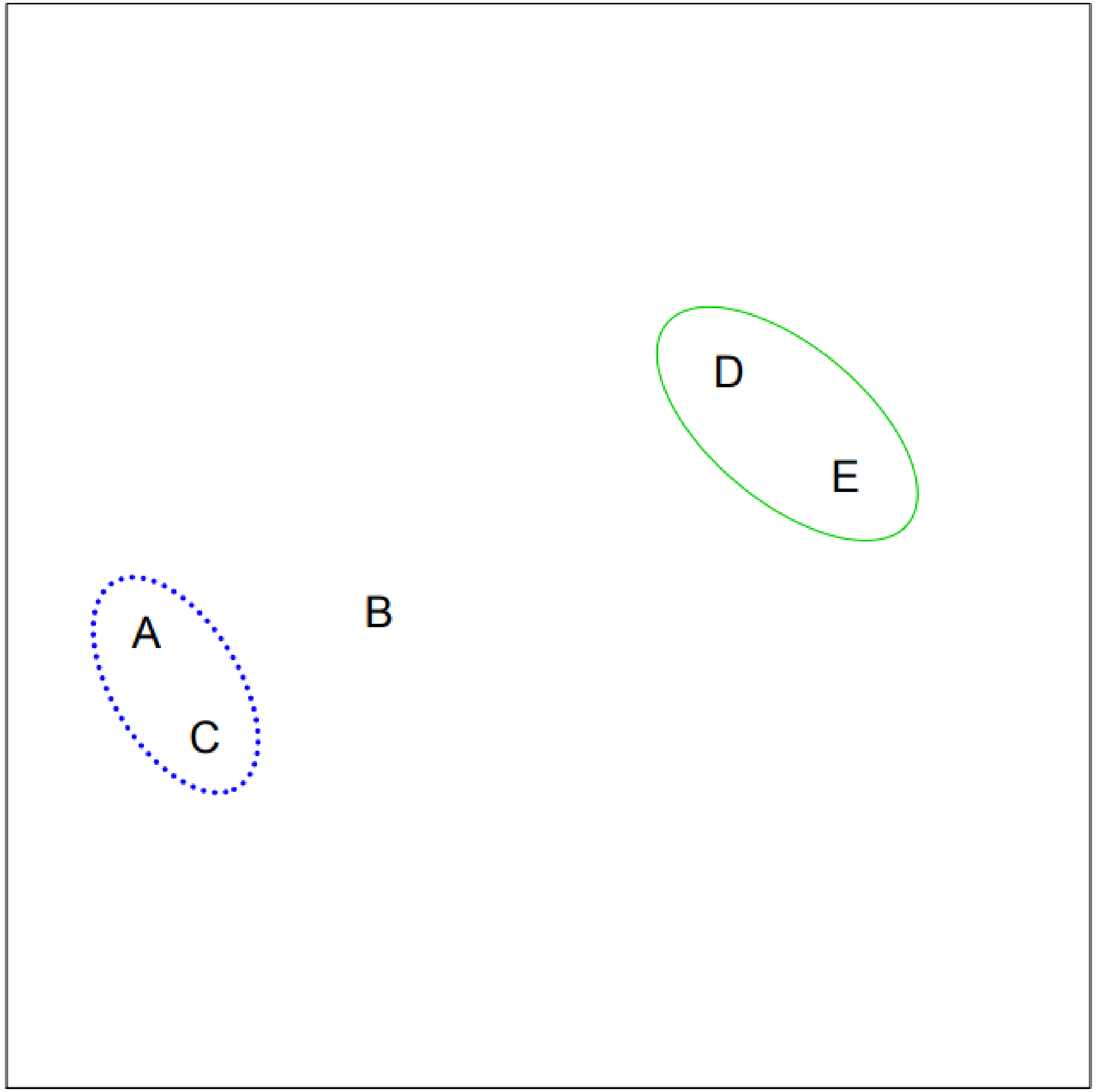
Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...



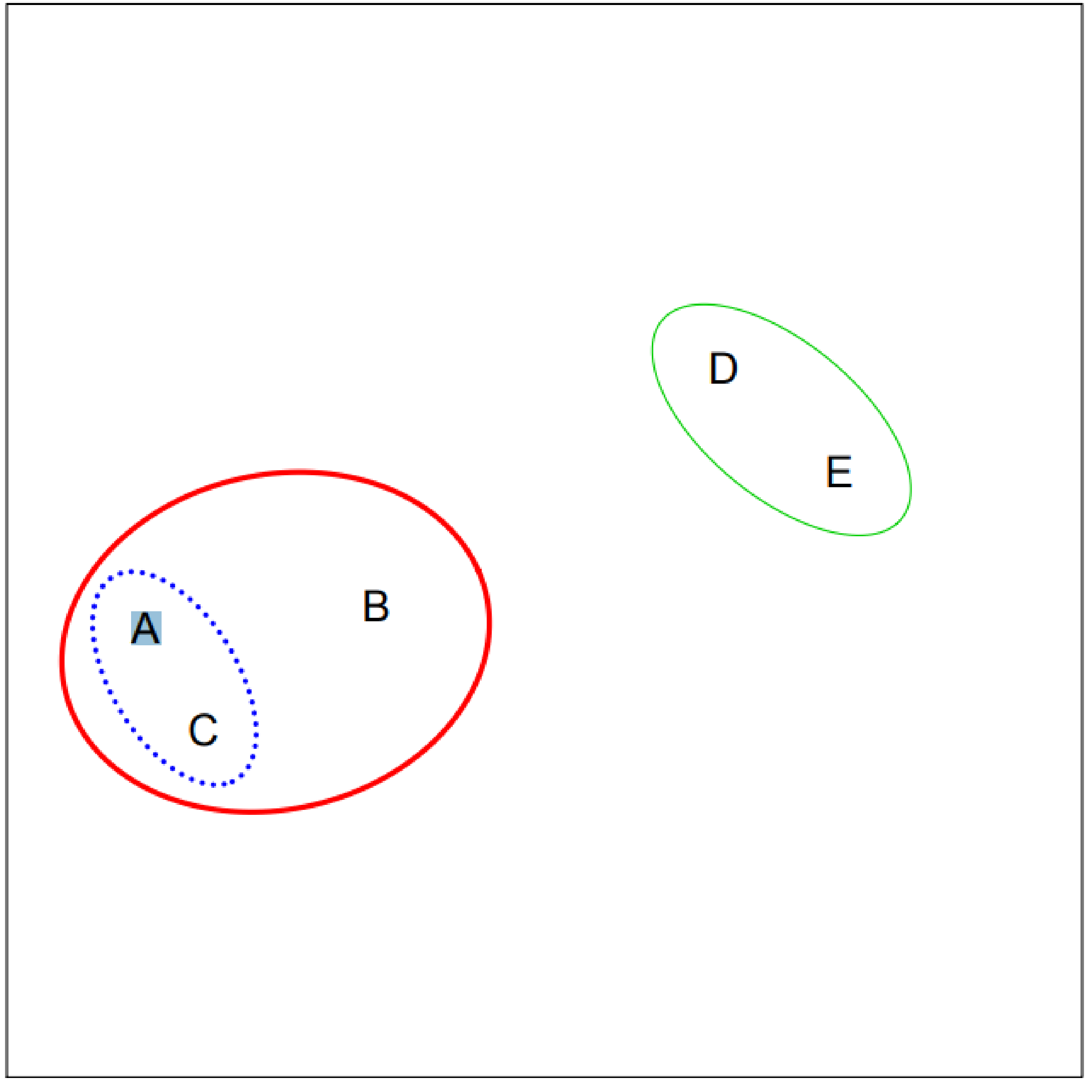
Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...



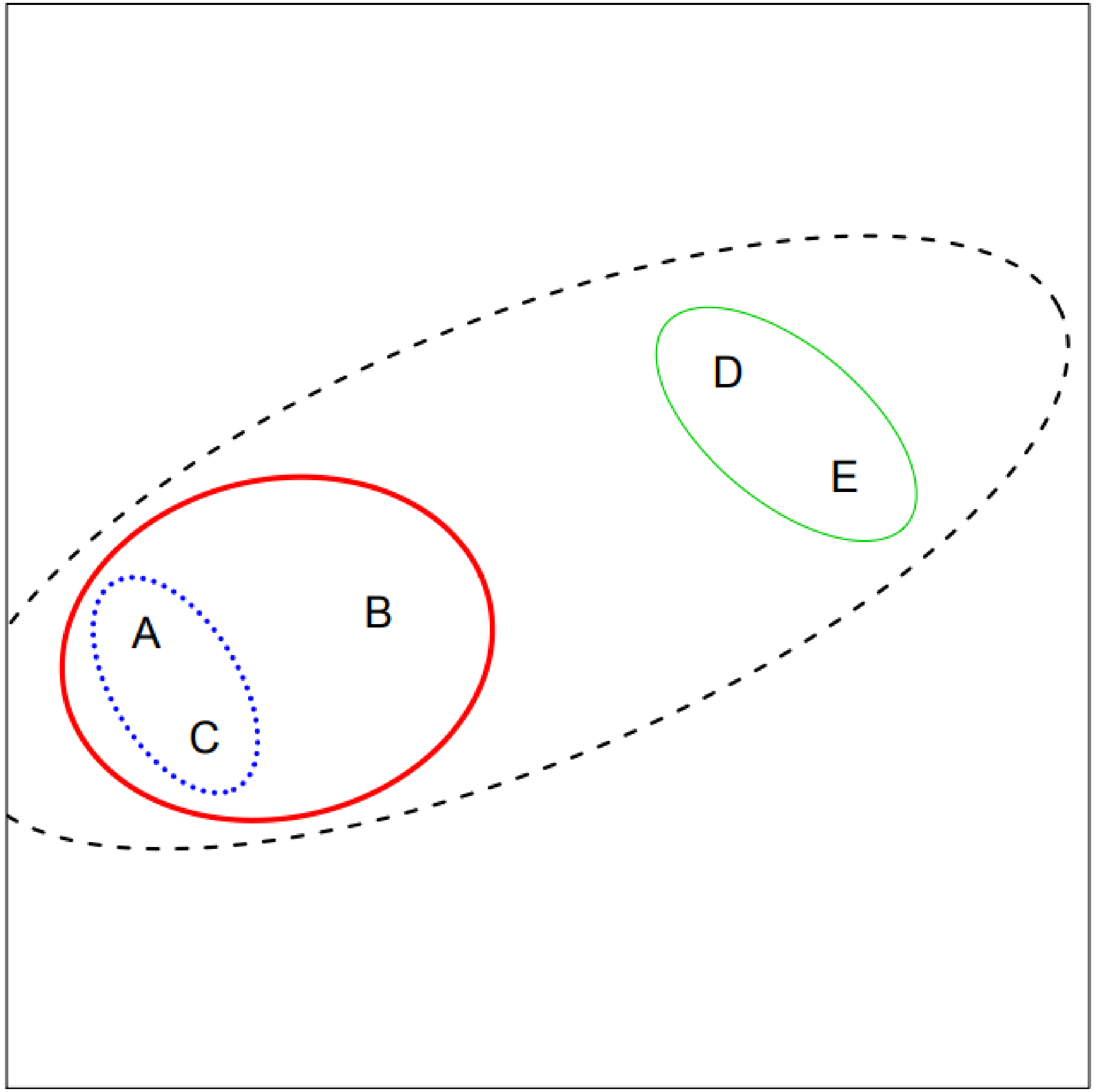
Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...



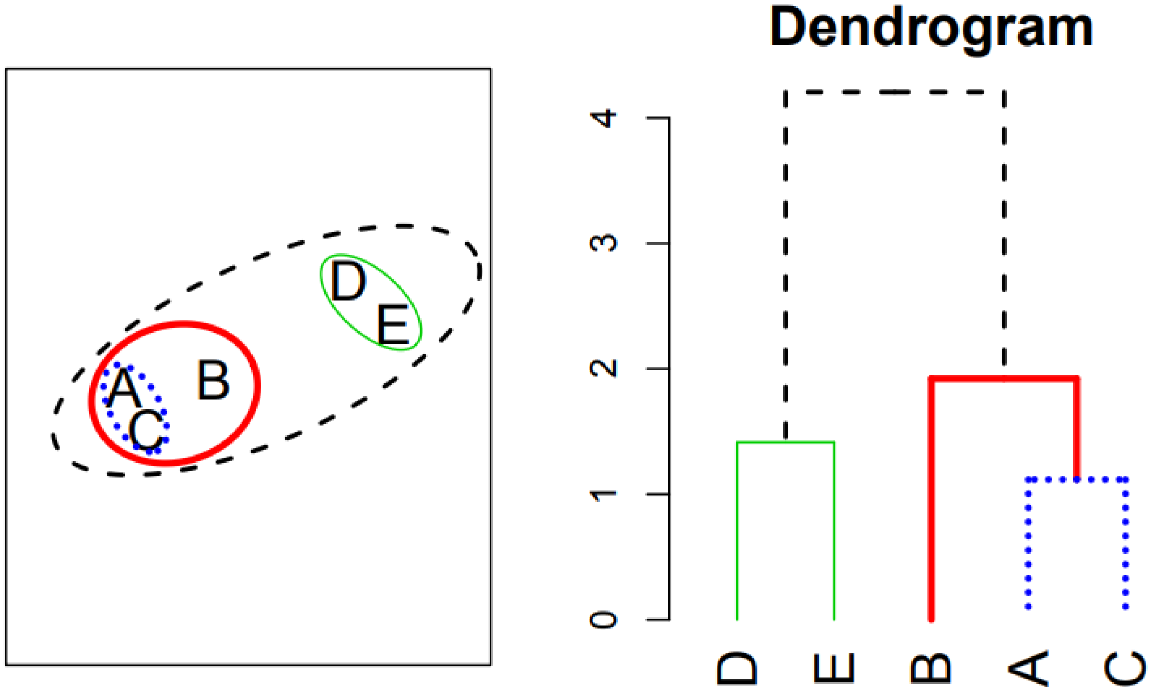
Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...

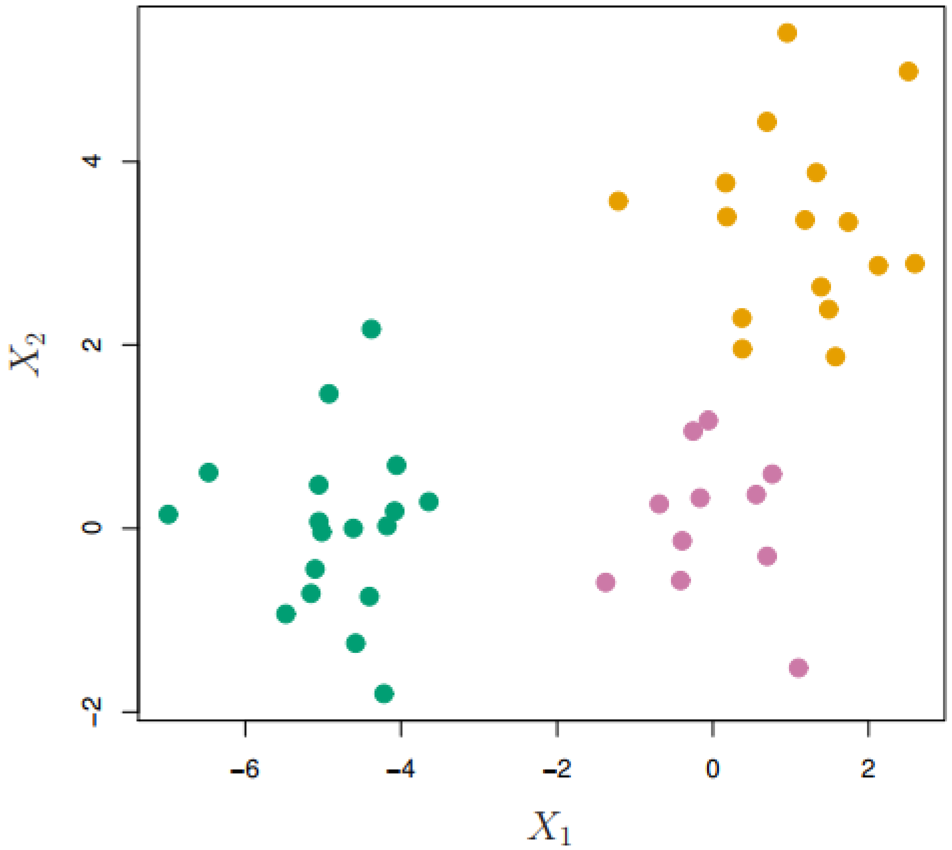


Hierarchical Clustering Algorithm

- The approach in words:
- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

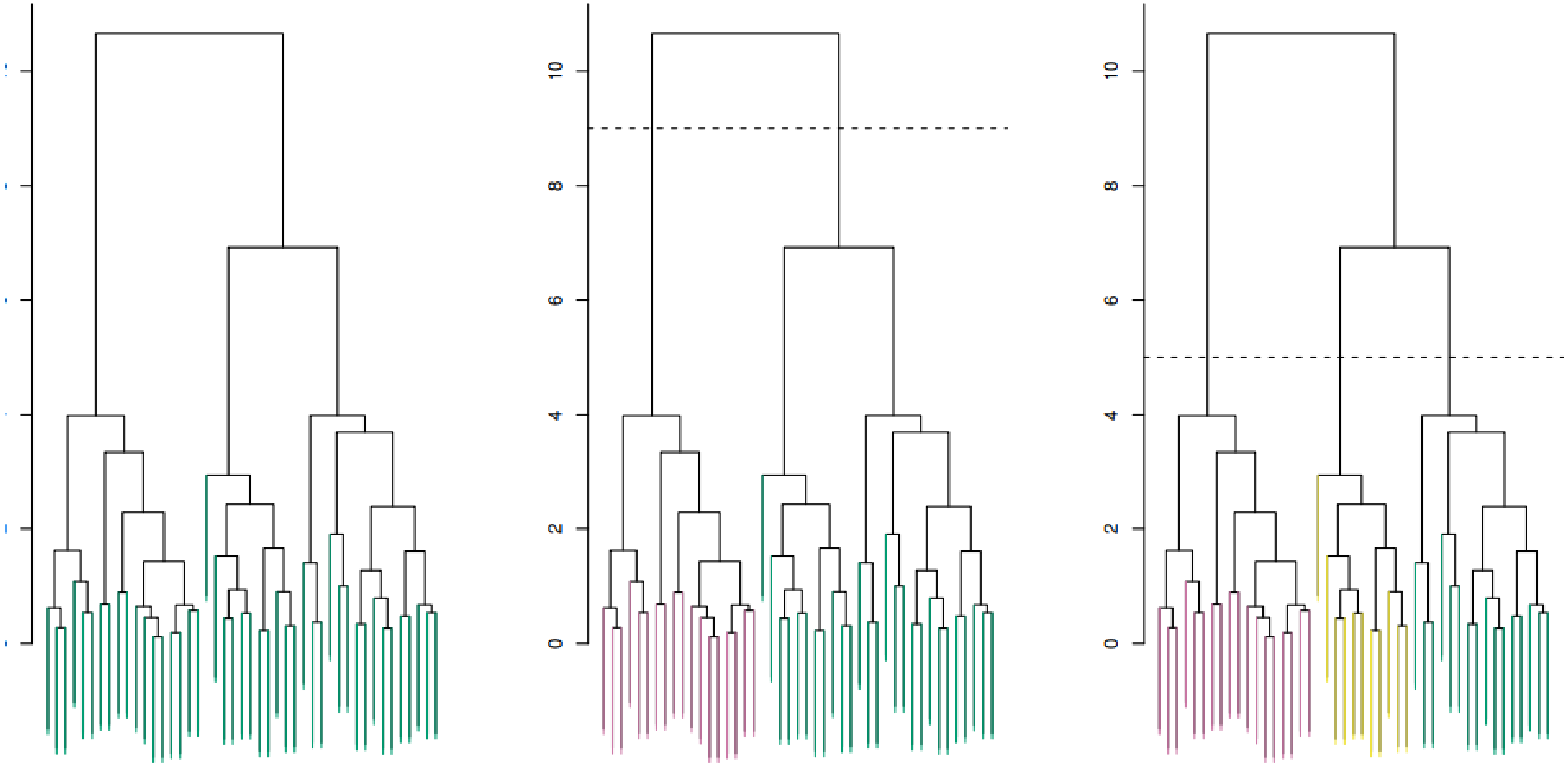


An Example



45 observations generated in 2-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.

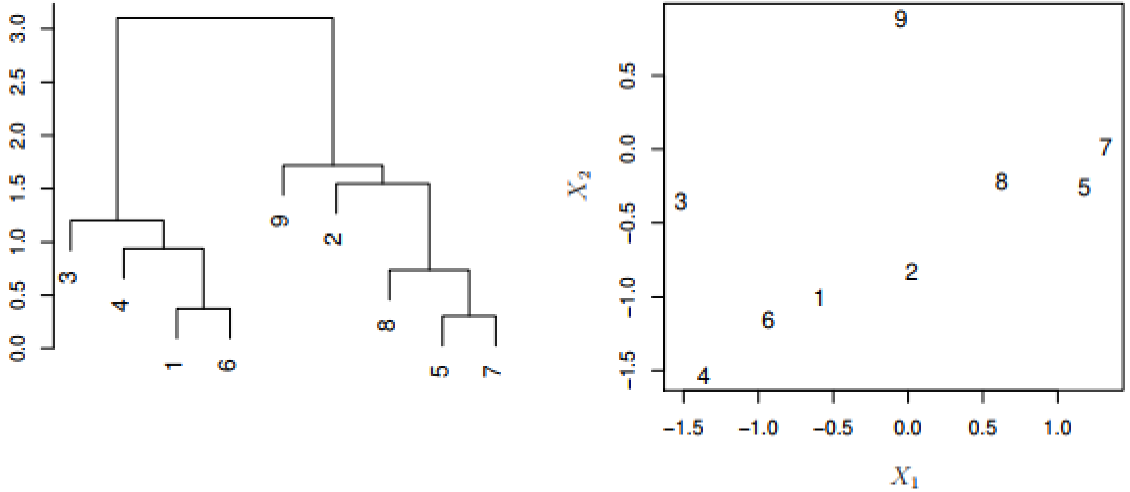
Application of hierarchical clustering



Details of previous figure

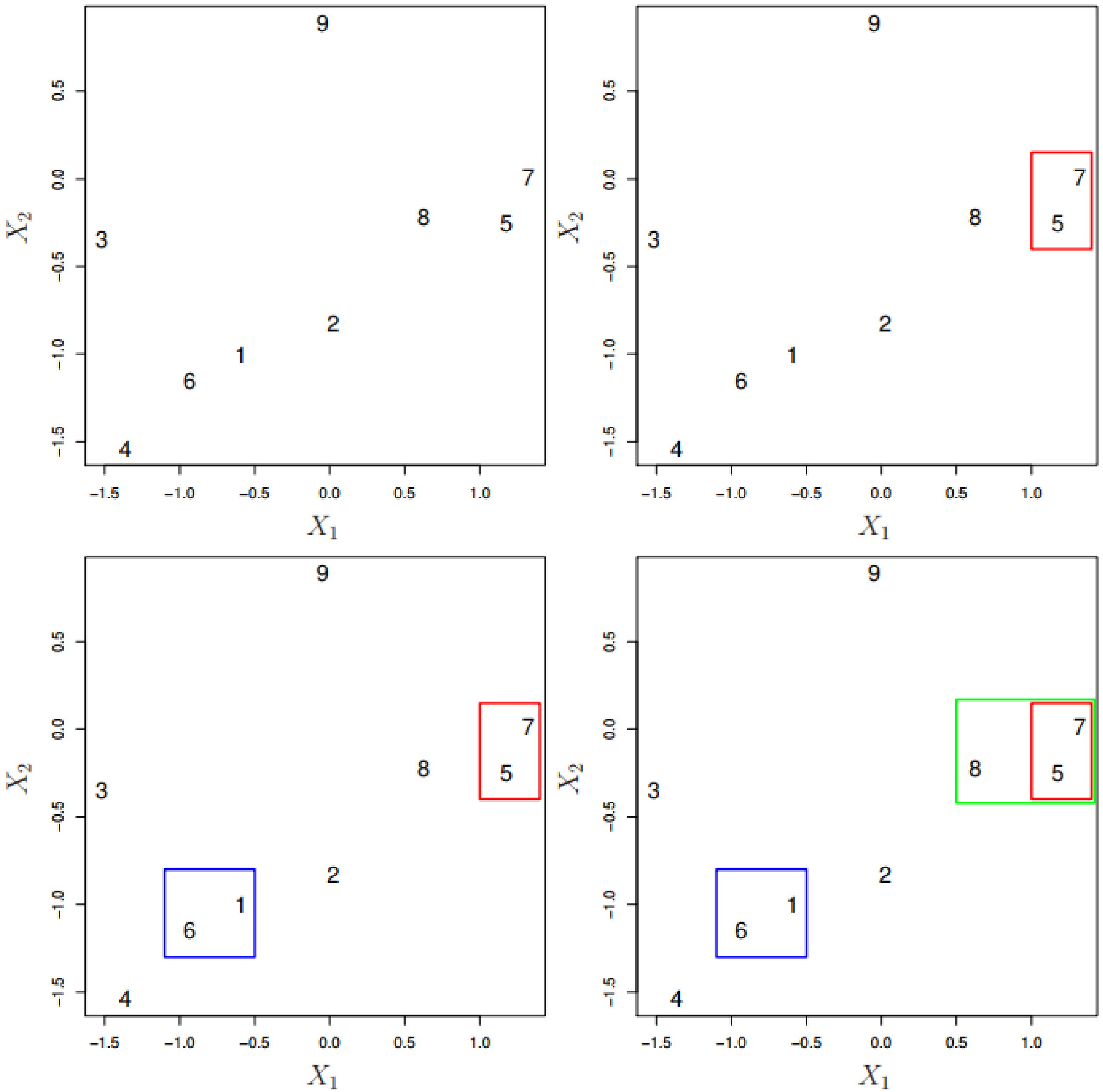
- Left: Dendrogram obtained from hierarchically clustering the data from previous slide, with complete linkage and Euclidean distance.
- Center: The dendrogram from the left-hand panel, cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.
- Right: The dendrogram from the left-hand panel, now cut at a height of 5. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure

Another Example

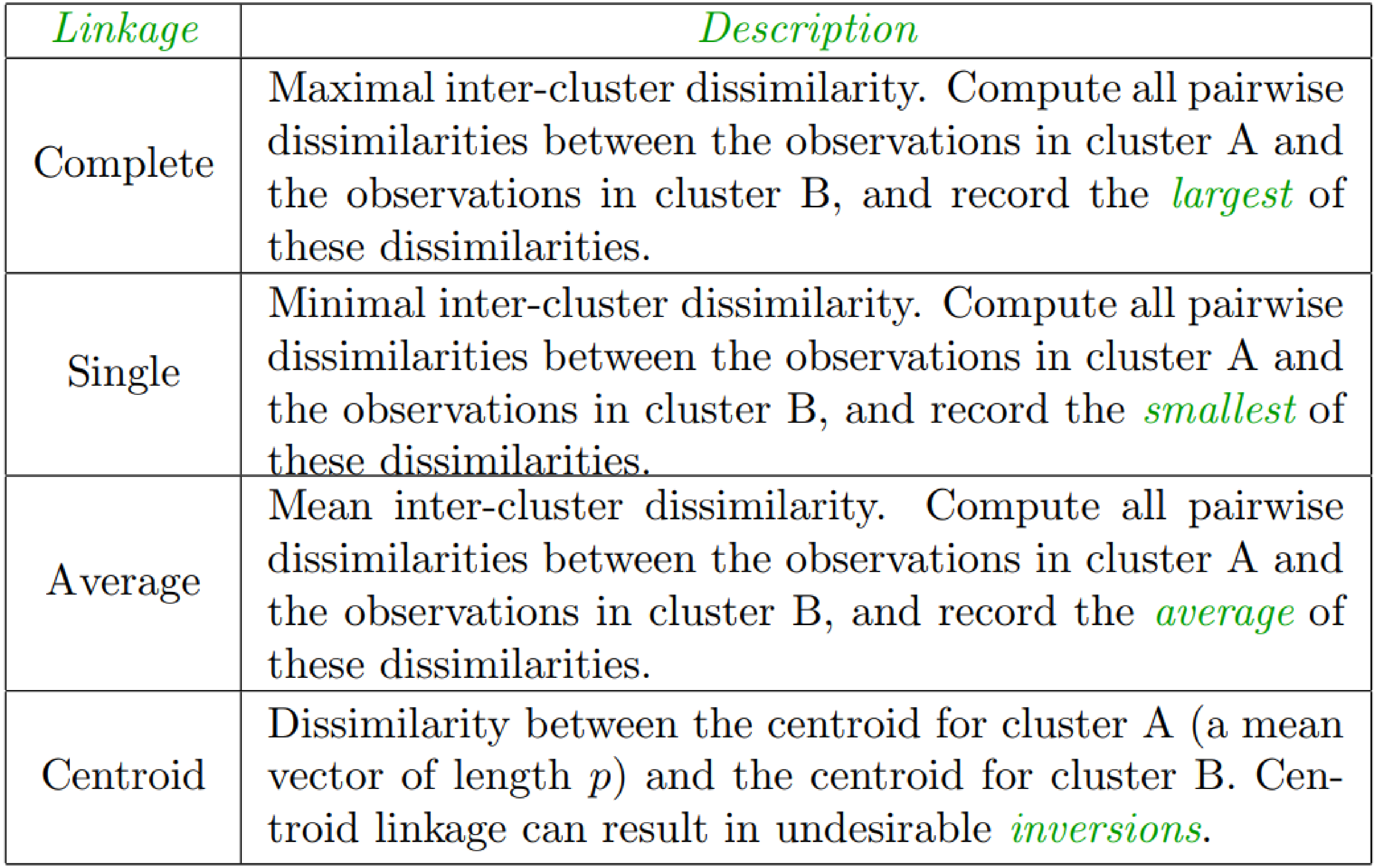


- An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space. The raw data on the right was used to generate the dendrogram on the left.
- Observations 5 and 7 are quite similar to each other, as are observations 1 and 6.
- However, observation 9 is no more similar to observation 2 than it is to observations 8, 5, and 7, even though observations 9 and 2 are close together in terms of horizontal distance.
- This is because observations 2, 8, 5, and 7 all fuse with observation 9 at the same height, approximately 1.8.

Merges in previous example



Types of Linkage



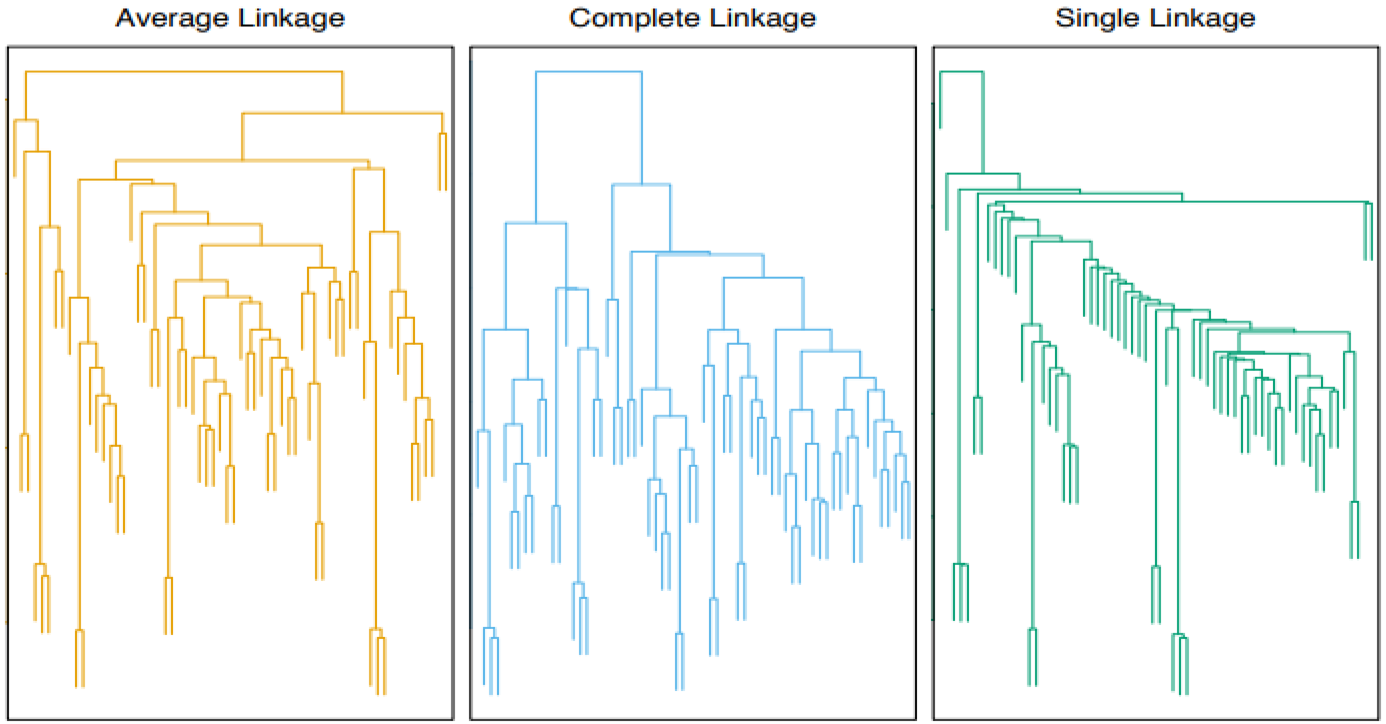
Linkage	*Description*
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities.
Average	Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Linkage Methods: Distance Between Clusters

- Complete Linkage: Largest distance between observations
- Single Linkage: Smallest distance between observations
- Average Linkage: Average distance between observations
- Centroid: distance between centroids of the observations

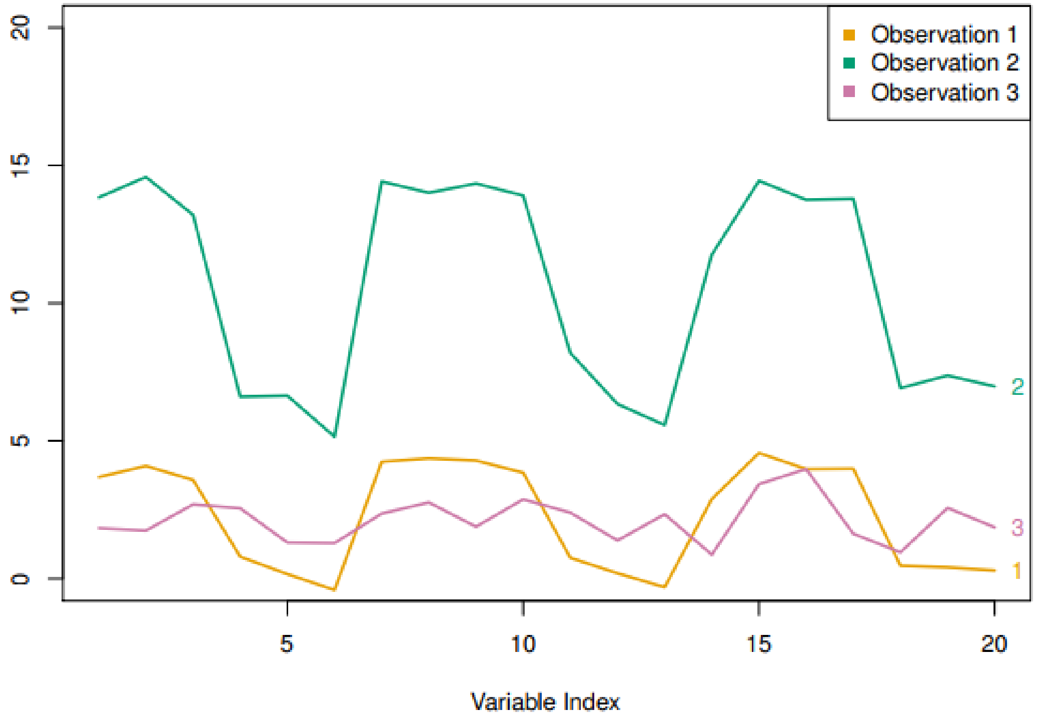
Linkage Can be Important

- Here we have three clustering results for the same data
- The only difference is the linkage method but the results are very different
- Complete and average linkage tend to yield evenly sized clusters whereas single linkage tends to yield extended clusters to which single leaves are fused one by one.



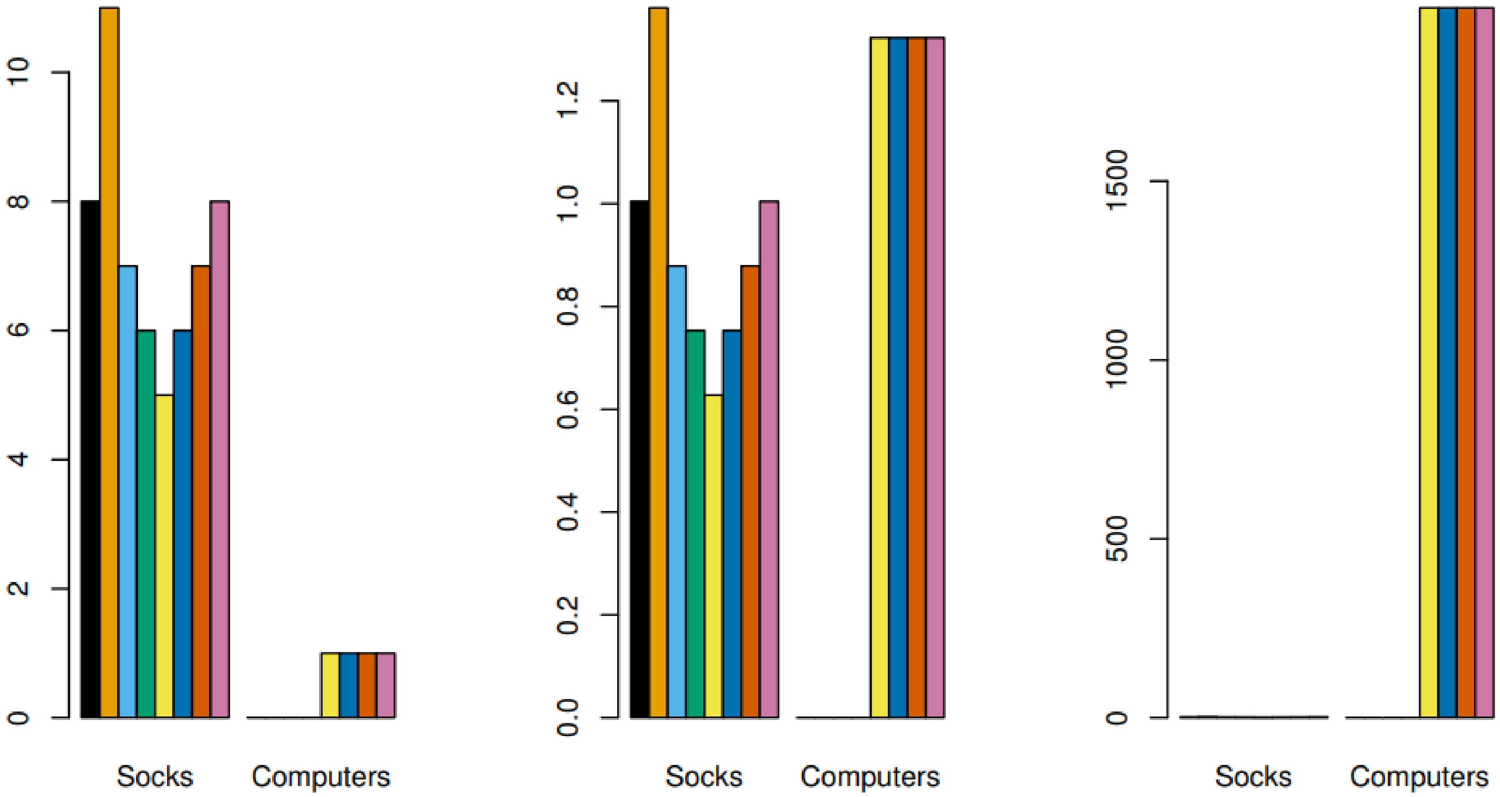
Choice of Dissimilarity Measure

- So far have used Euclidean distance.
- An alternative is correlation-based distance which considers two observations to be similar if their features are highly correlated.
- This is an unusual use of correlation, which is normally computed between variables; here it is computed between the observation profiles for each pair of observations.



Scaling of the variables matters

- Consider an online shop that sells two items: socks and computers
  - Left: In terms of quantity, socks have higher weight
  - Center: After standardizing, socks and computers have equal weight
  - Right: In terms of dollar sales, computers have higher weight



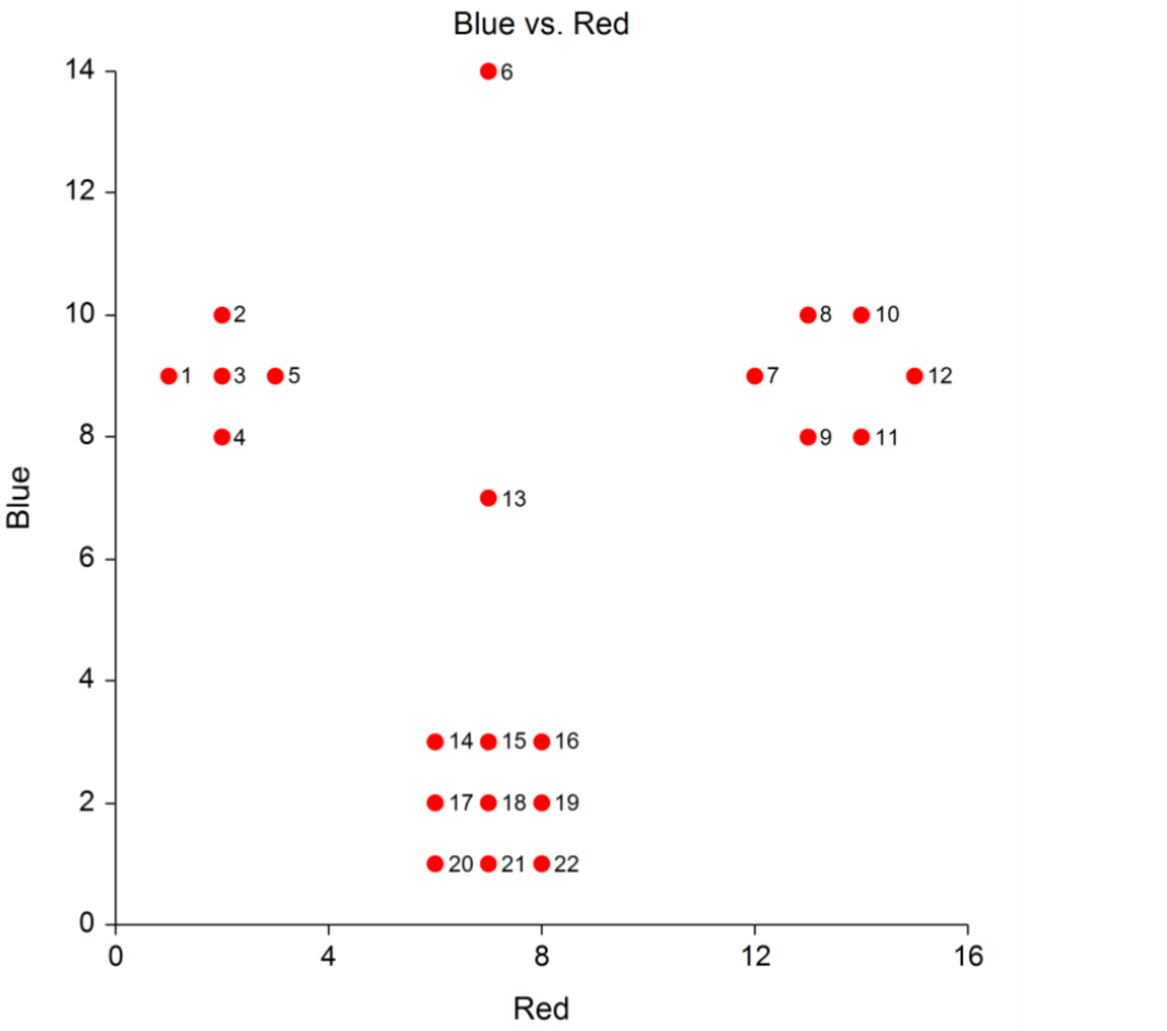
Practical issues

- Should the observations or features first be standardized in some way? For instance, maybe the variables should be centered to have mean zero and scaled to have standard deviation one.
- In the case of hierarchical clustering,
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendogram in order to obtain clusters?
- In case of K-means clustering:
  - How many clusters should we look for the data?
- In practice, we try several different choices, and look for the one with the most useful or interpretable solution. There is no single right answer!

Practical issues

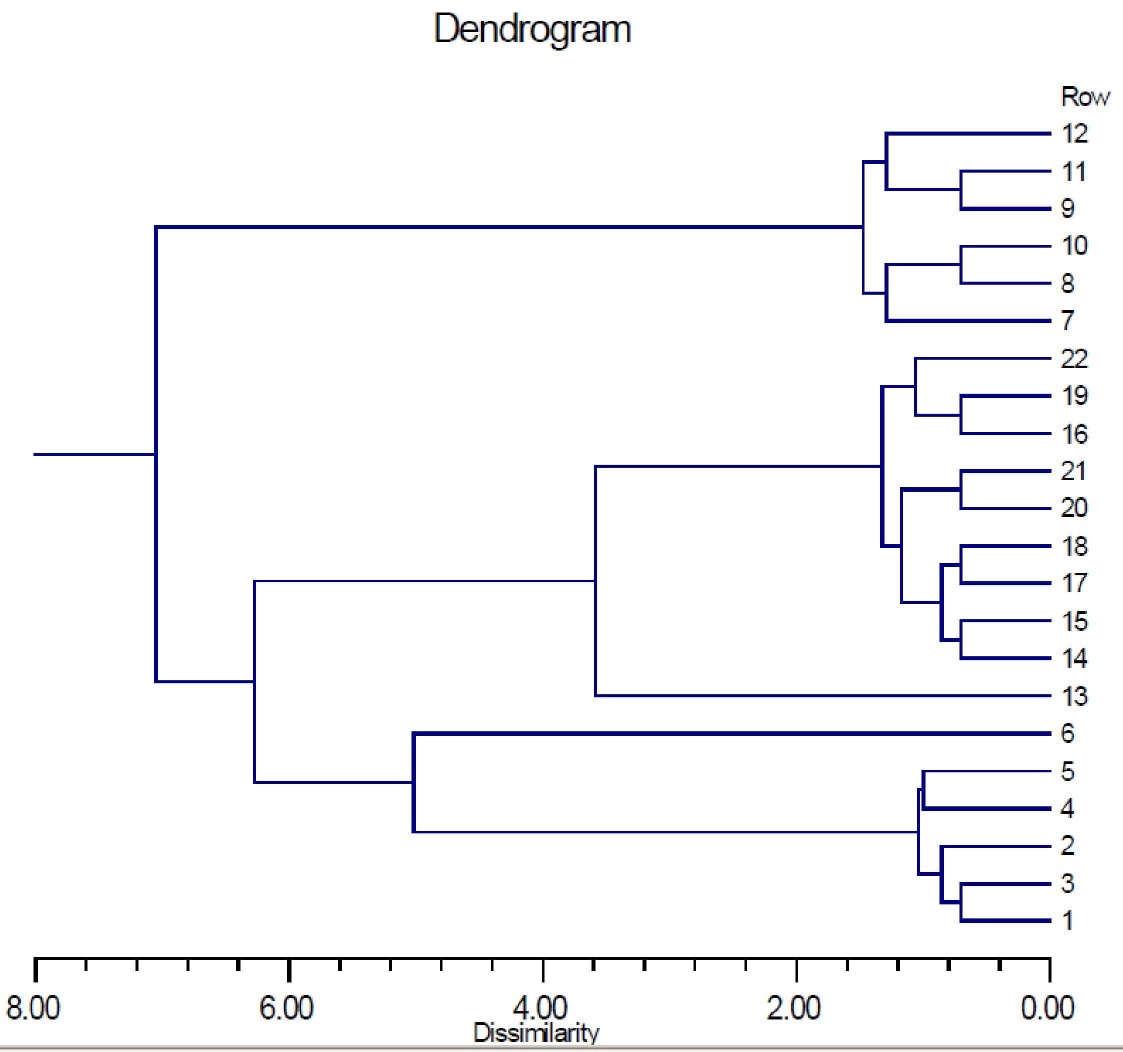
- Most importantly, one must be careful about how the results of a clustering analysis are reported
- These results should not be taken as the absolute truth about a data set
- Rather, they should constitute a starting point for the developments of a scientific hypothesis and further study, preferably on independent data

Examples



Suppose we wish to cluster the bivariate data shown in the following scatter plot. In this case, the clustering may be done visually. The data have three clusters and two singletons, 6 and 13.

Following is a dendrogram of the results of running these data through the Group Average clustering algorithm.



The algorithm used by all eight of the clustering methods is outlined as follows. Let the distance between clusters *i* and *j* be represented as d_{ij} and let cluster *i* contain n_i objects. Let ***D*** represent the set of all remaining d_{ij}. Suppose there are *N* objects to cluster.

1. Find the smallest element d_{ij} remaining in ***D***.
2. Merge clusters *i* and *j* into a single new cluster, *k*.
3. Calculate a new set of distances d_{km} using the following distance formula.

   $$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma \left| d_{im} - d_{jm} \right|$$

   Here *m* represents any cluster other than *k*. These new distances replace d_{im} and d_{jm} in ***D***. Also let $n_k = n_i + n_j$.

   Note that the eight algorithms available represent eight choices for α_i, α_j, β, and γ.
4. Repeat steps 1 - 3 until ***D*** contains a single group made up off all objects. This will require *N-1* iterations.

We will now give brief comments about each of the eight techniques.

Single Linkage

Also known as *nearest neighbor* clustering, this is one of the oldest and most famous of the hierarchical techniques. The distance between two groups is defined as the distance between their two closest members. It often yields clusters in which individuals are added sequentially to a single group.

The coefficients of the distance equation are $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = -0.5.$

Complete Linkage

Also known as furthest neighbor or maximum method, this method defines the distance between two groups as the distance between their two farthest-apart members. This method usually yields clusters that are well separated and compact.

The coefficients of the distance equation are $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = 0.5.$

Simple Average

Also called the weighted pair-group method, this algorithm defines the distance between groups as the average distance between each of the members, weighted so that the two groups have an equal influence on the final result.

The coefficients of the distance equation are $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = 0.$

Centroid

Also referred to as the unweighted pair-group centroid method, this method defines the distance between two groups as the distance between their centroids (center of gravity or vector average). The method should only be used with Euclidean distances.

The coefficients of the distance equation are $\alpha_i = \frac{n_i}{n_k}, \alpha_j = \frac{n_j}{n_k}, \beta = -\alpha_i\alpha_j, \gamma = 0.$

Backward links may occur with this method. These are recognizable when the dendrogram no longer exhibits its simple tree-like structure in which each fusion results in a new cluster that is at a higher distance level (moves from right to left). With backward links, fusions can take place that result in clusters at a lower distance level (move from left to right). The dendrogram is difficult to interpret in this case.

Median

Also called the weighted pair-group centroid method, this defines the distance between two groups as the weighted distance between their centroids, the weight being proportional to the number of individuals in each group. Backward links (see discussion under Centroid) may occur with this method. The method should only be used with Euclidean distances.

The coefficients of the distance equation are $\alpha_i = \alpha_j = 0.5, \beta = -0.25, \gamma = 0.$

Group Average

Also called the unweighted pair-group method, this is perhaps the most widely used of all the hierarchical cluster techniques. The distance between two groups is defined as the average distance between each of their members.

The coefficients of the distance equation are $\alpha_i = \dfrac{n_i}{n_k}, \alpha_j = \dfrac{n_j}{n_k}, \beta = 0, \gamma = 0.$

Ward's Minimum Variance

With this method, groups are formed so that the pooled within-group sum of squares is minimized. That is, at each step, the two clusters are fused which result in the least increase in the pooled within-group sum of squares.

The coefficients of the distance equation are $\alpha_i = \dfrac{n_i + n_m}{n_k + n_m}, \alpha_j = \dfrac{n_j + n_m}{n_k + n_m}, \beta = \dfrac{-n_m}{n_k + n_m}, \gamma = 0.$

Flexible Strategy

Lance and Williams (1967) suggested that a continuum could be made between single and complete linkage. The program lets you try various settings of these parameters which do not conform to the constraints suggested by Lance and Williams.

The coefficients of the distance equation should conform to the following constraints
$\alpha_i = 1 - \beta - \alpha_j, \alpha_j = 1 - \beta - \alpha_i, -1 \leq \beta \leq 1, \gamma = 0.$

One interesting exercise is to vary these values, trying to find the set that maximizes the cophenetic correlation coefficient.

Measures

- Euclidean distance

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

- Squared Euclidean distance

$$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$$

- Manhattan distance

$$\|a-b\|_1 = \sum_i |a_i - b_i|$$

- Maximum distance

$$\|a-b\|_\infty = \max_i |a_i - b_i|$$

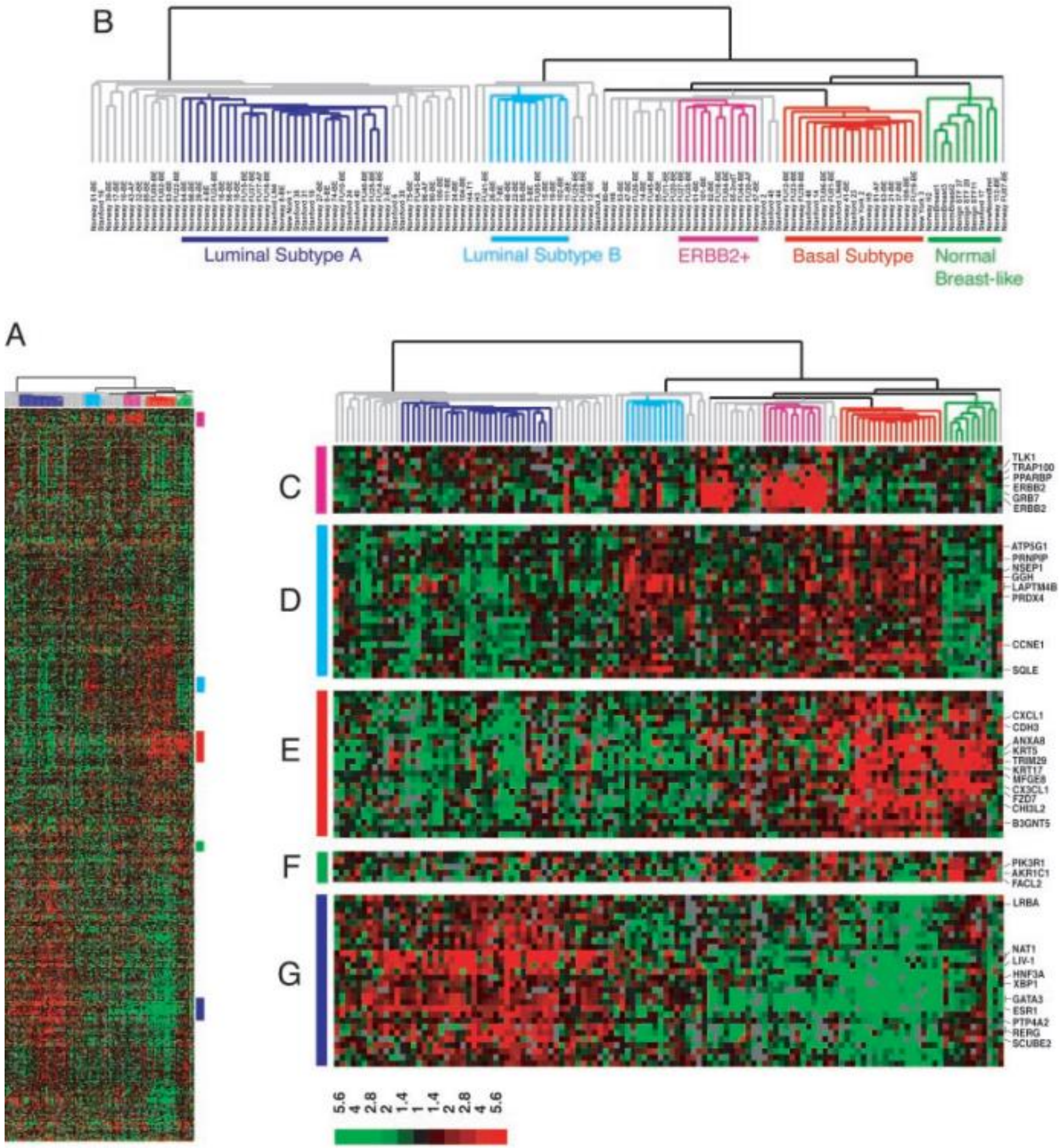
- Mahalanobis distance

$$\sqrt{(a-b)^\top S^{-1}(a-b)}$$

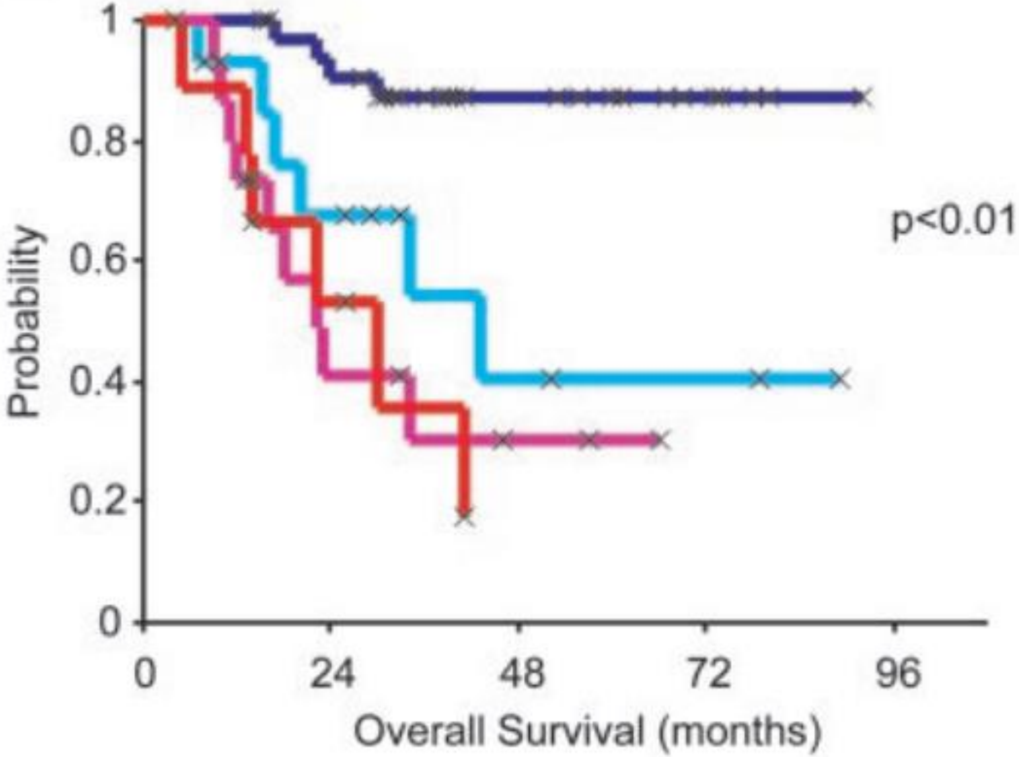
where S is the covariance matrix

Example: breast cancer microarray study

- “Repeated observation of breast tumor subtypes in independent gene expression data sets;” Sorlie at el, PNAS 2003
- Average linkage, correlation metric
- Clustered samples using 500 intrinsic genes: each woman was measured before and after chemotherapy. Intrinsic genes have smallest within/between variation.


B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal Breast-like
A
C
D
E
F
G
TLK1
TRAP100
PPARBP
ERBB2
GRB7
ERBB2
ATP5G1
PRNPIP
NSEP1
GGH
LAPTM4B
PRDX4
CCNE1
SQLE
CXCL1
CDH3
ANXA8
KRT5
TRIM29
KRT17
MFGE8
CX3CL1
FZD7
CHI3L2
B3GNT5
PIK3R1
AKR1C1
FACL2
LRBA
NAT1
LIV-1
HNF3A
XBP1
GATA3
ESR1
PTP4A2
RERG
SCUBE2
5.6
4
2.8
2
1.4
1
1.4
2
2.8
4
5.6

B Norway/Stanford data set



Conclusions

- Unsupervised learning is important for understanding the variation and grouping structure of a set of unlabeled data, and can be a useful pre-processor for supervised learning
- It is intrinsically more difficult than supervised learning because there is no gold standard (like an outcome variable) and no single objective (like test set accuracy)
- It is an active field of research, with many recently developed tools such as self-organizing maps, independent components analysis and spectral clustering. See The Elements of Statistical Learning, chapter 14.