MONASH University

ETC3250

Business Analytics

Week 12
Clustering

21 October 2015

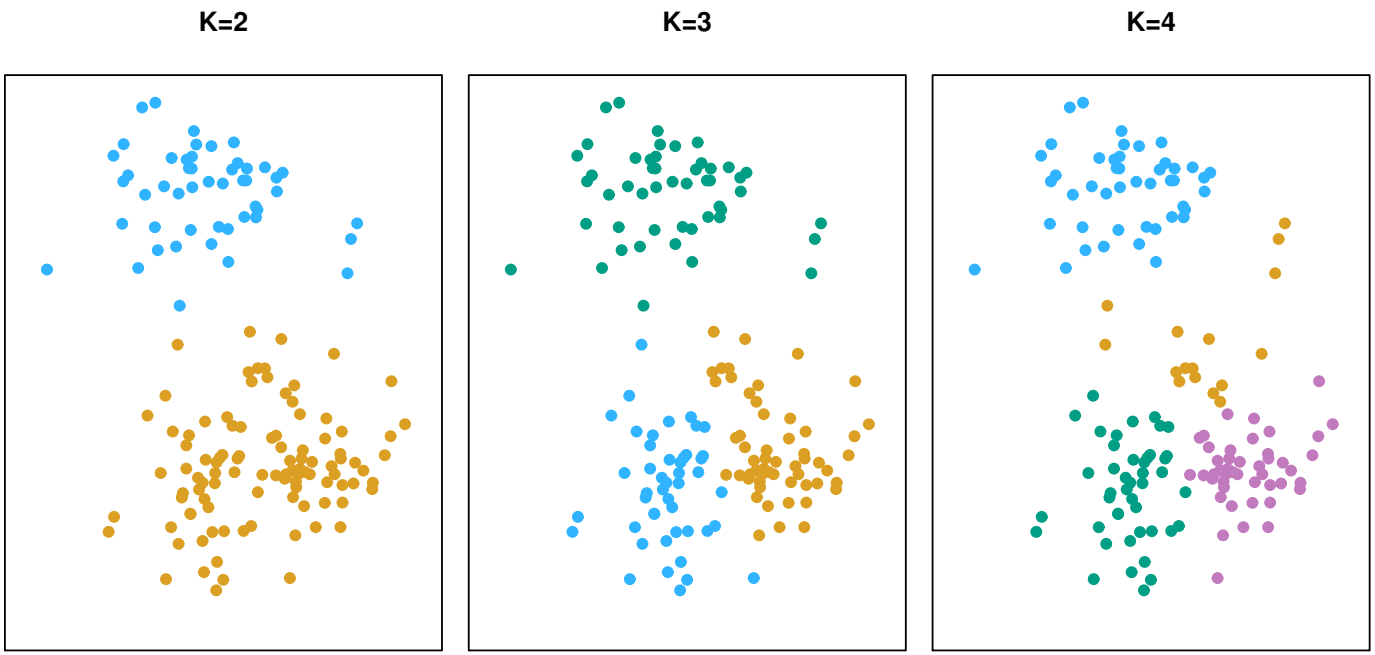
Outline

Week	Topic	Chapter	Lecturers
1	Introduction to business analytics & R	1	Rob,Souhaib
2	Statistical learning	2	Rob,Souhaib
3	Regression for prediction	3	Rob
4	Resampling	5	Rob, Souhaib
5	Dimension reduction	6,10	Rob, Souhaib
6	Visualization		Di
7	Visualization		Di
8	Classification	4	Souhaib, Di
9	Classification	4,9	Di, Souhaib
-	Semester Break		
10	Advanced classification	8	Di
11	Advanced regression	6	Di
12	Clustering	10	Di, Souhaib

K-means clustering

Find K clusters $C_1, \ldots, C_K$ where

- $C_1 \cup C_2 \cup C_K = \{1, \ldots, n\}$
- $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$



K-means clustering

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

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

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

- Which algorithm to solve this minimisation problem?
- There are almost K^n ways to partition n observations into K clusters

K-means clustering

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

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

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

- Which algorithm to solve this minimisation problem?
- There are almost K^n ways to partition n observations into K clusters

K-means clustering

This algorithm will converge to a local optimum:

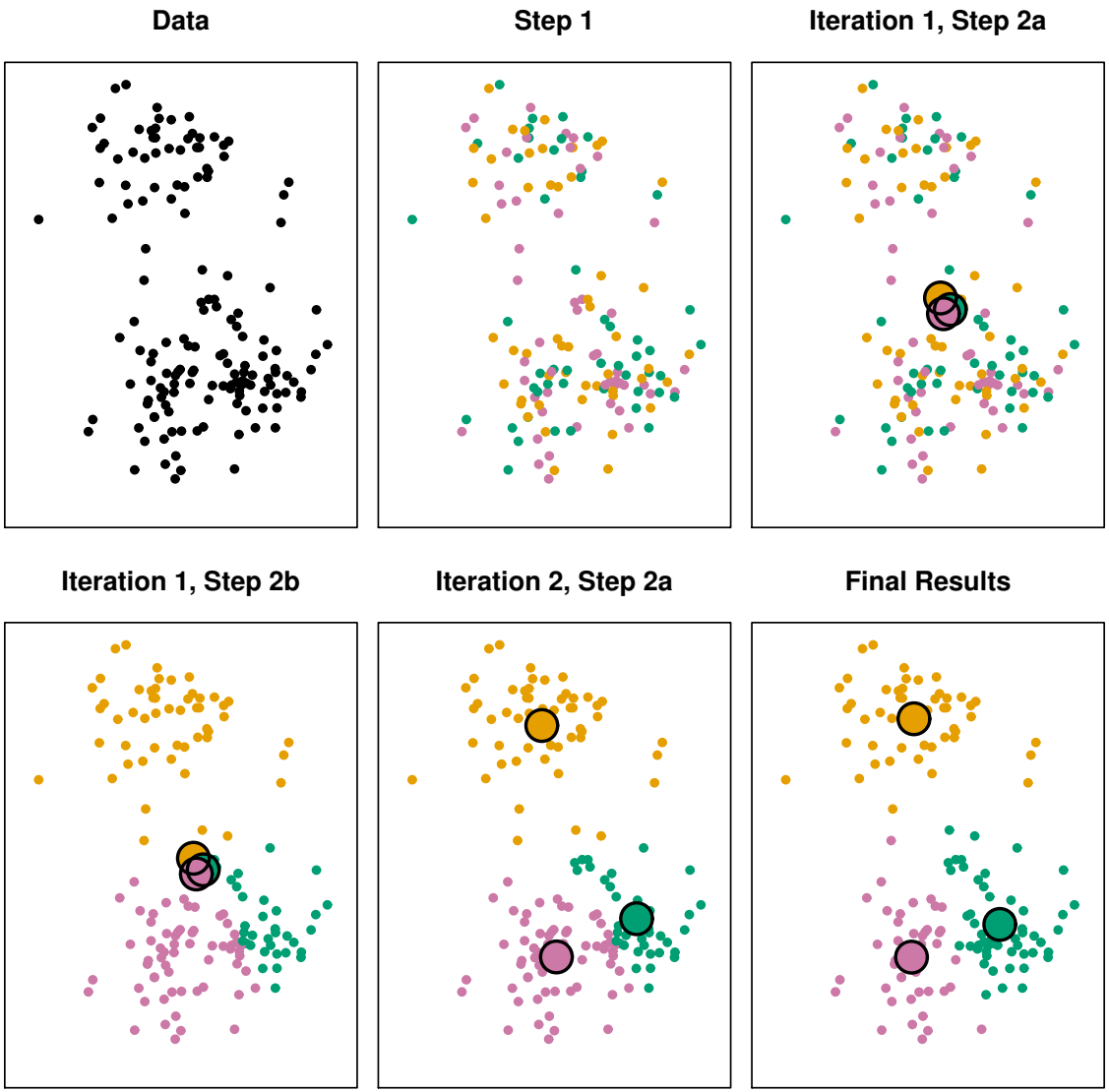
Algorithm 10.1 *K-Means Clustering*

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:
   (a) 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.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

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

$$\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i \in C_k} x_{ij}$$

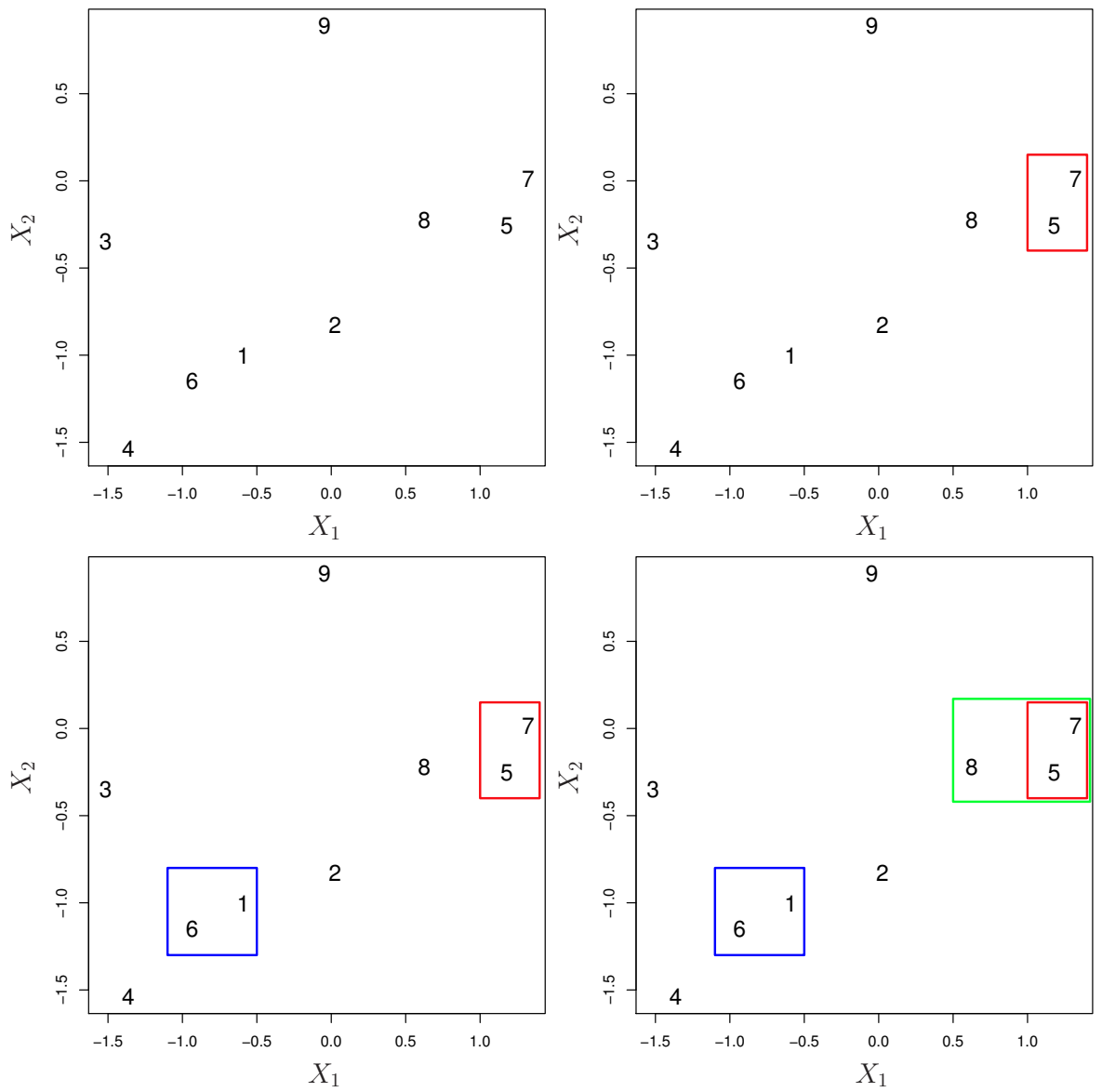
K-means clustering



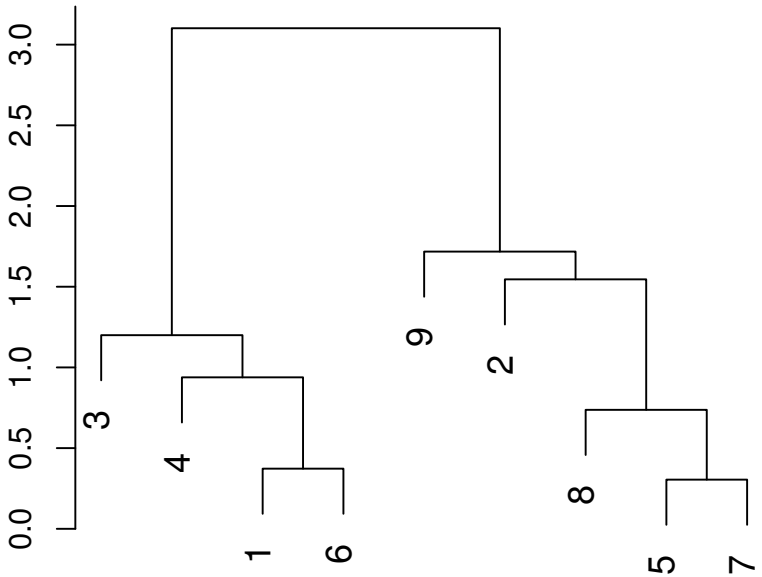
K-means clustering

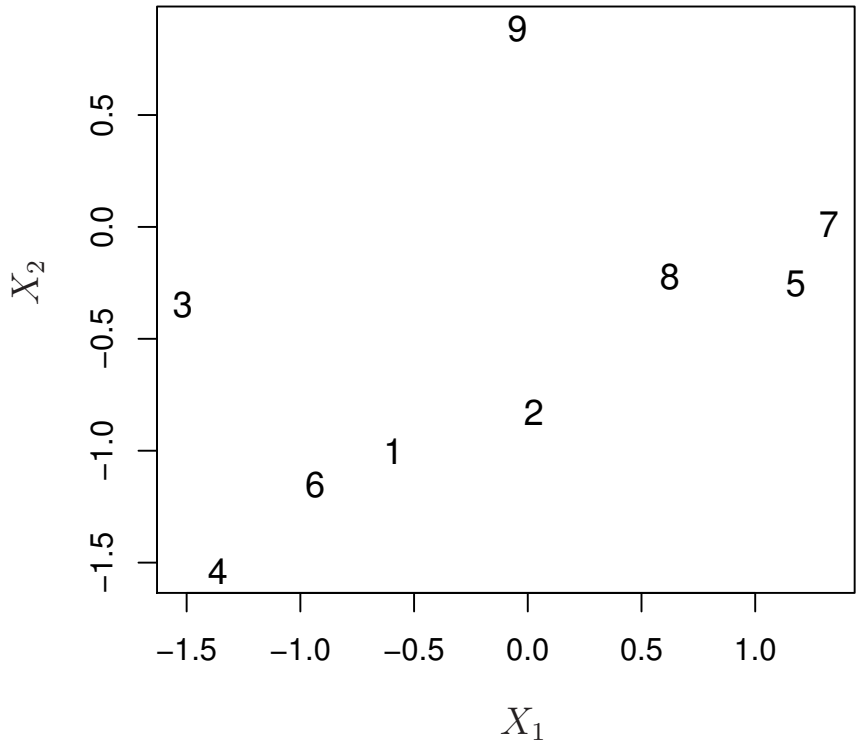
Different random initial configurations

Hierarchical clustering



Hierarchical clustering

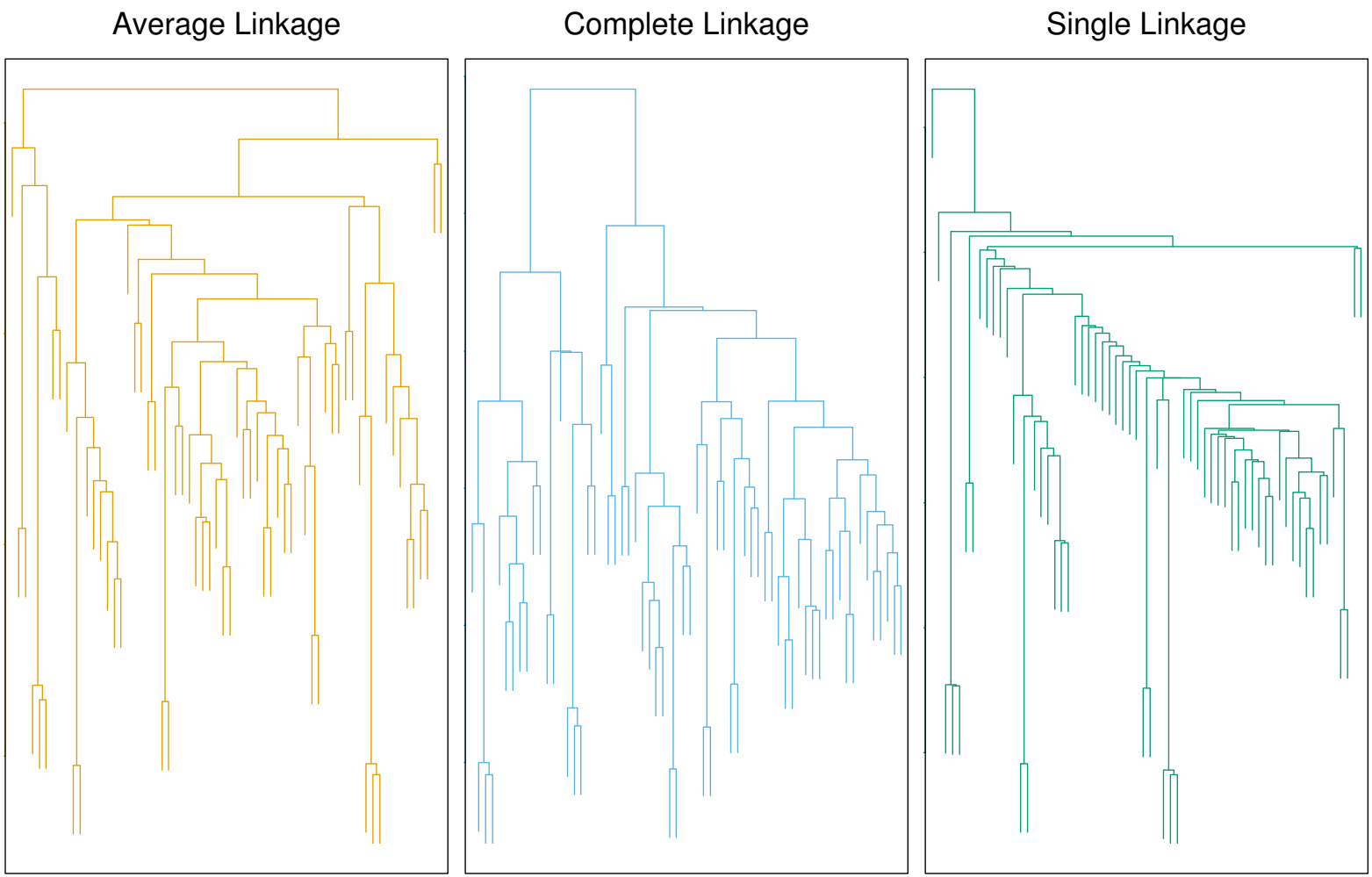




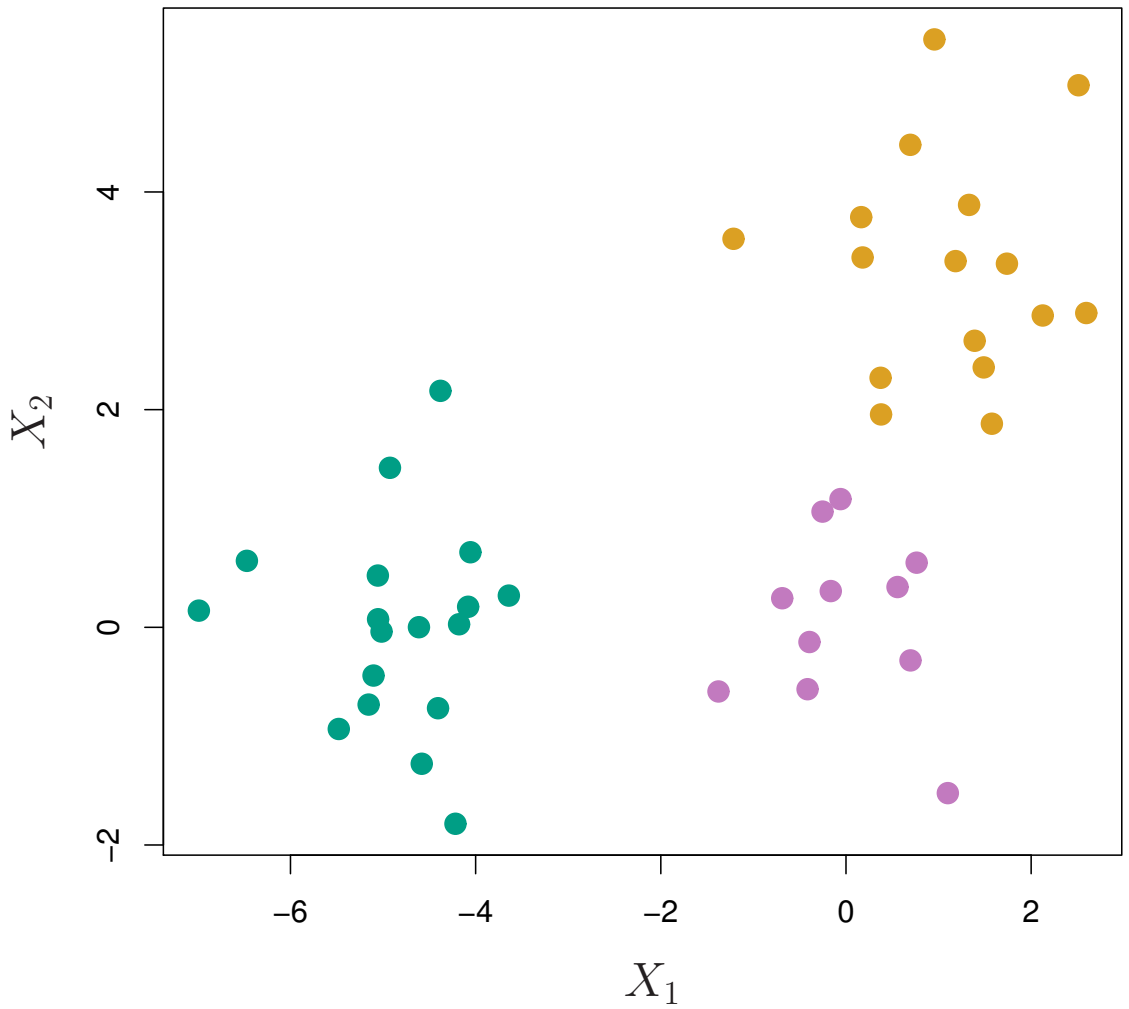
Hierarchical clustering

Linkage	*Description*
Complete	Maximal intercluster 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 intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster 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*.

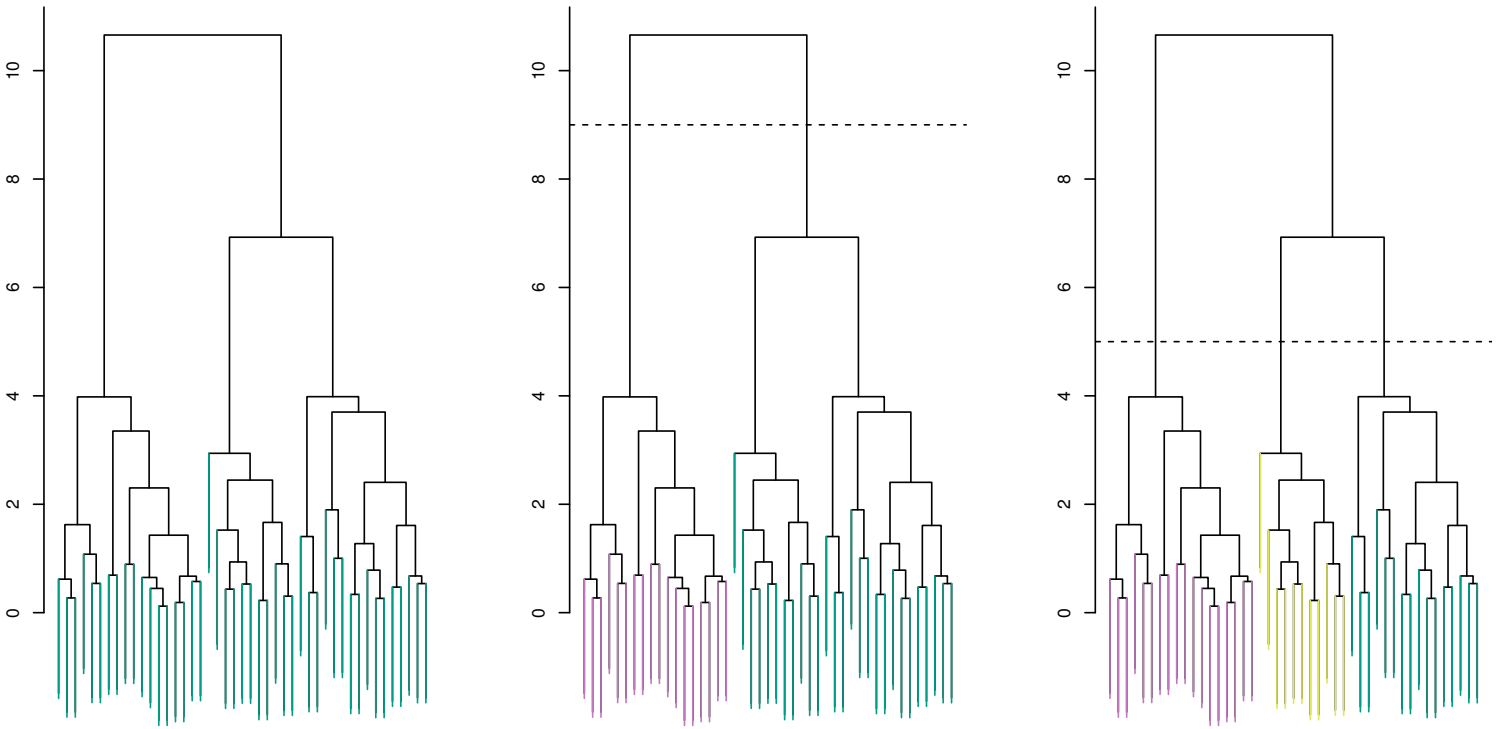
Hierarchical clustering



Hierarchical clustering



Hierarchical clustering



Hierarchical clustering

Algorithm 10.2 *Hierarchical Clustering*

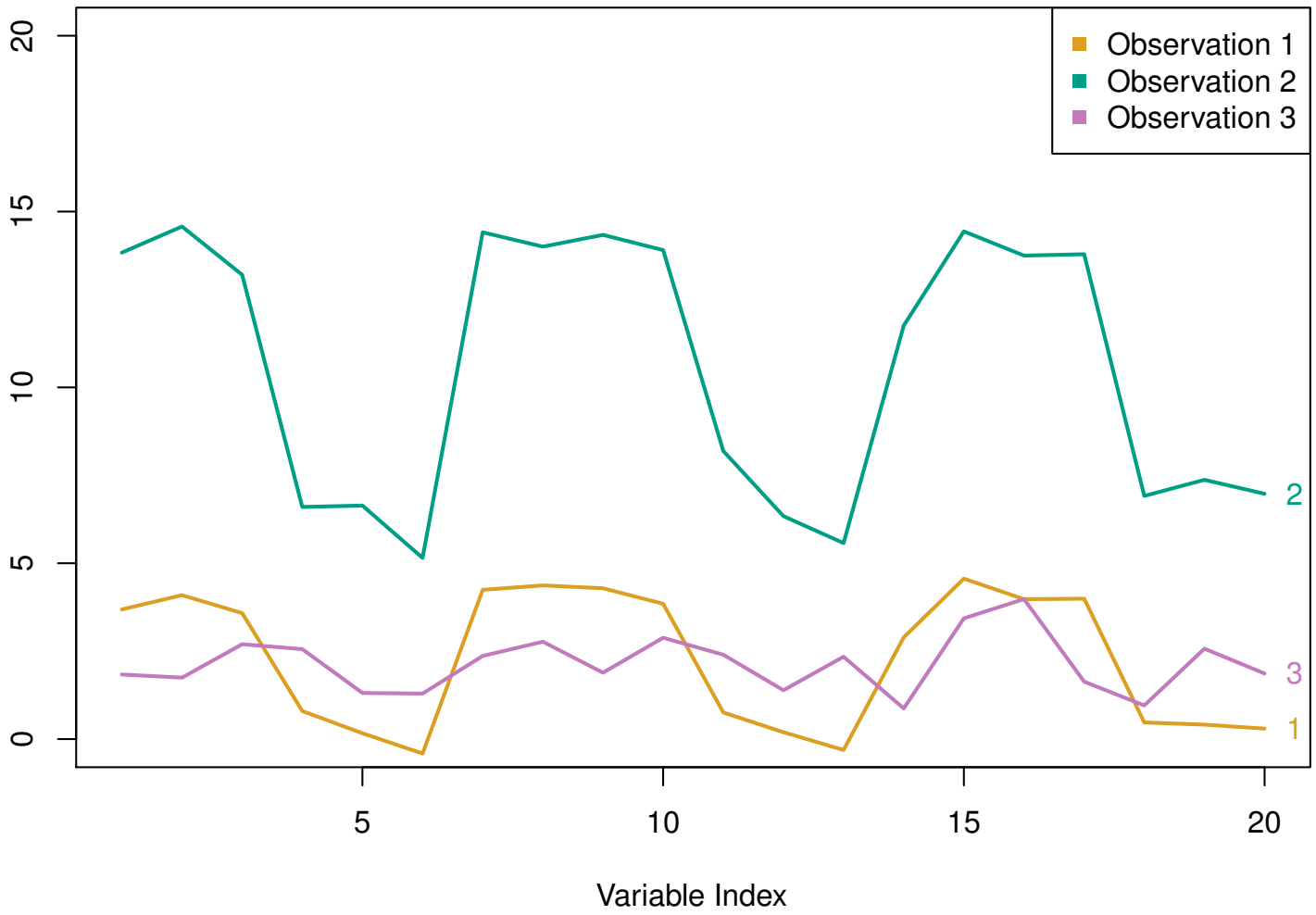
1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

Dissimilarity measure

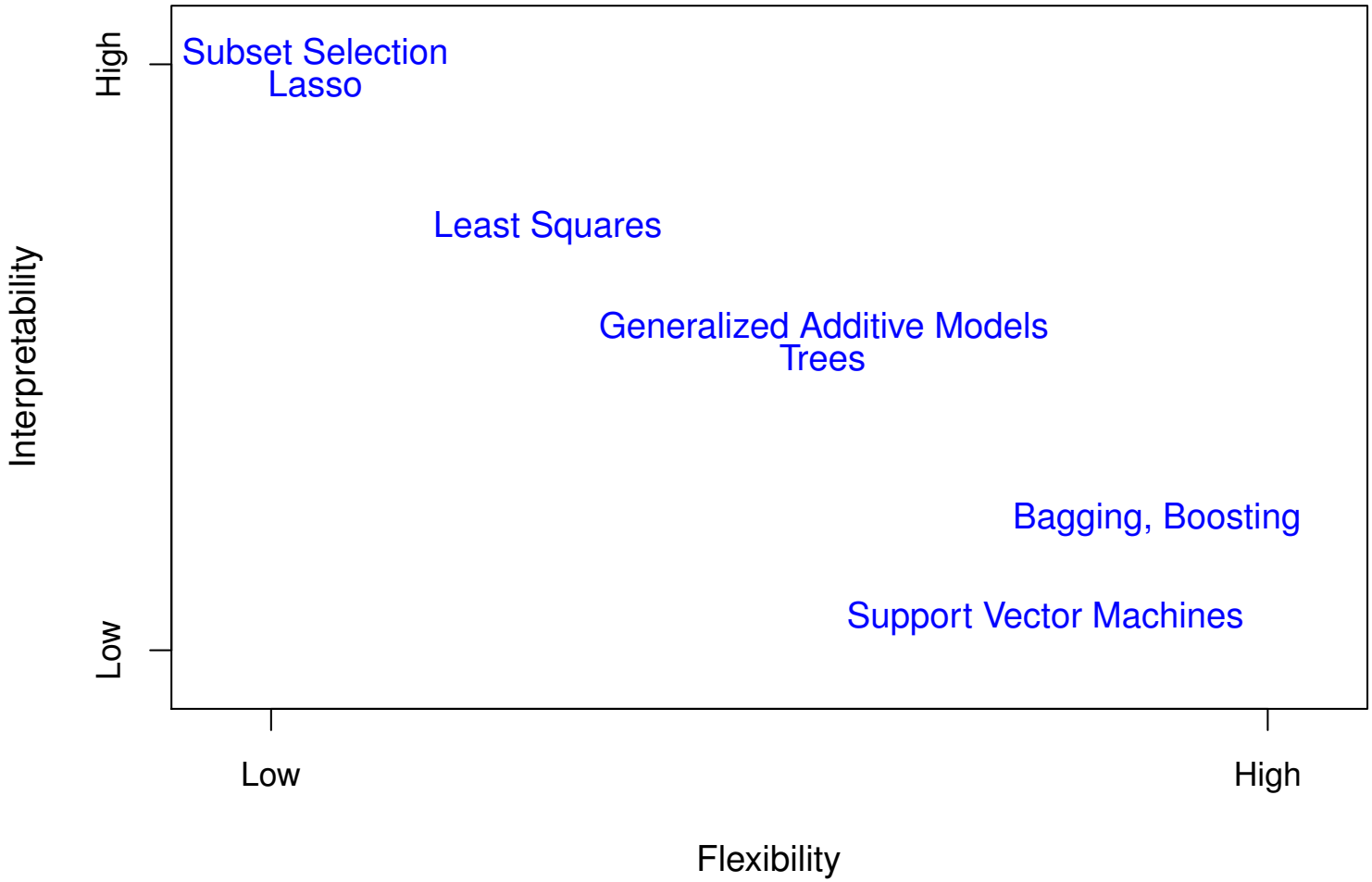


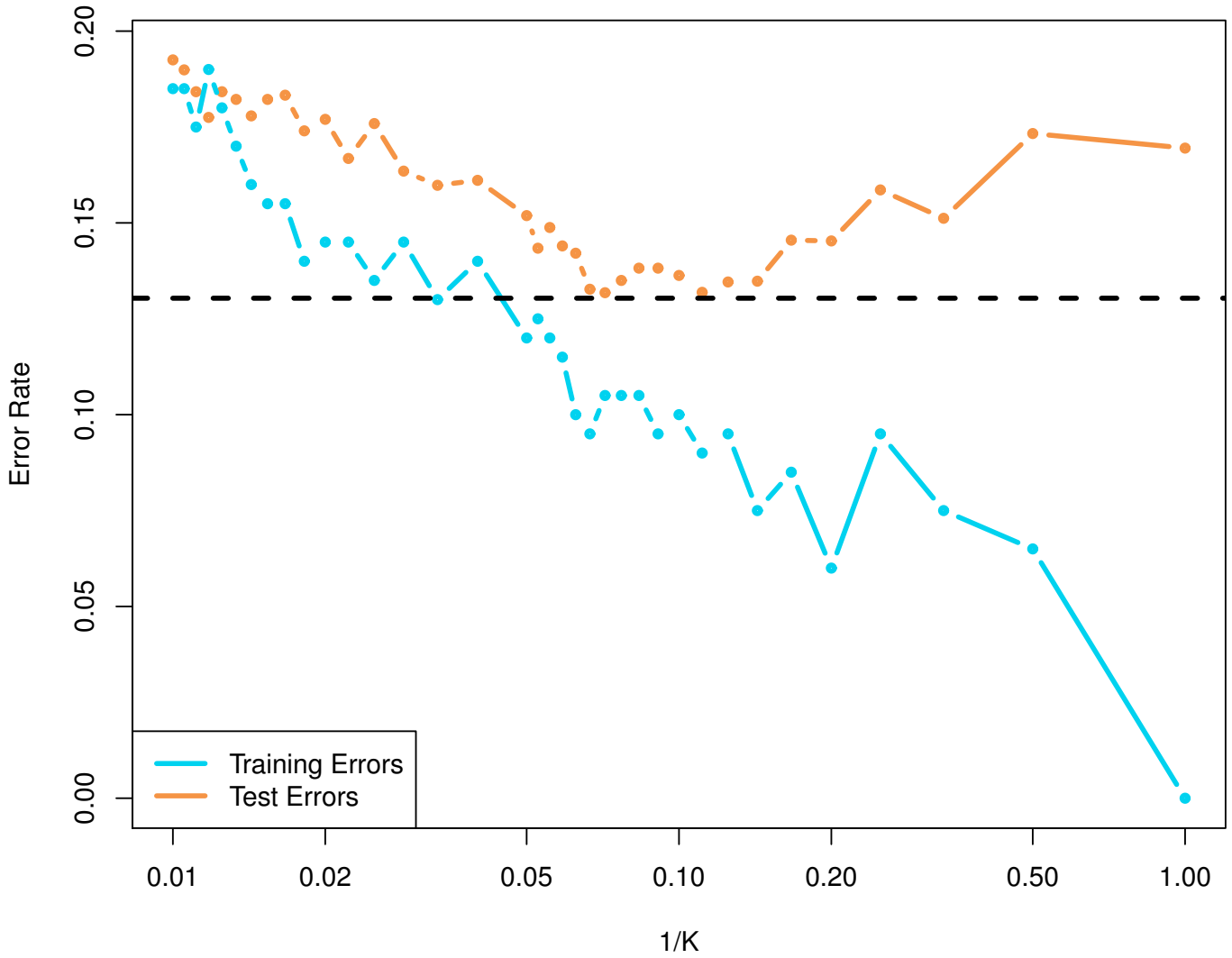
Practical issues in clustering

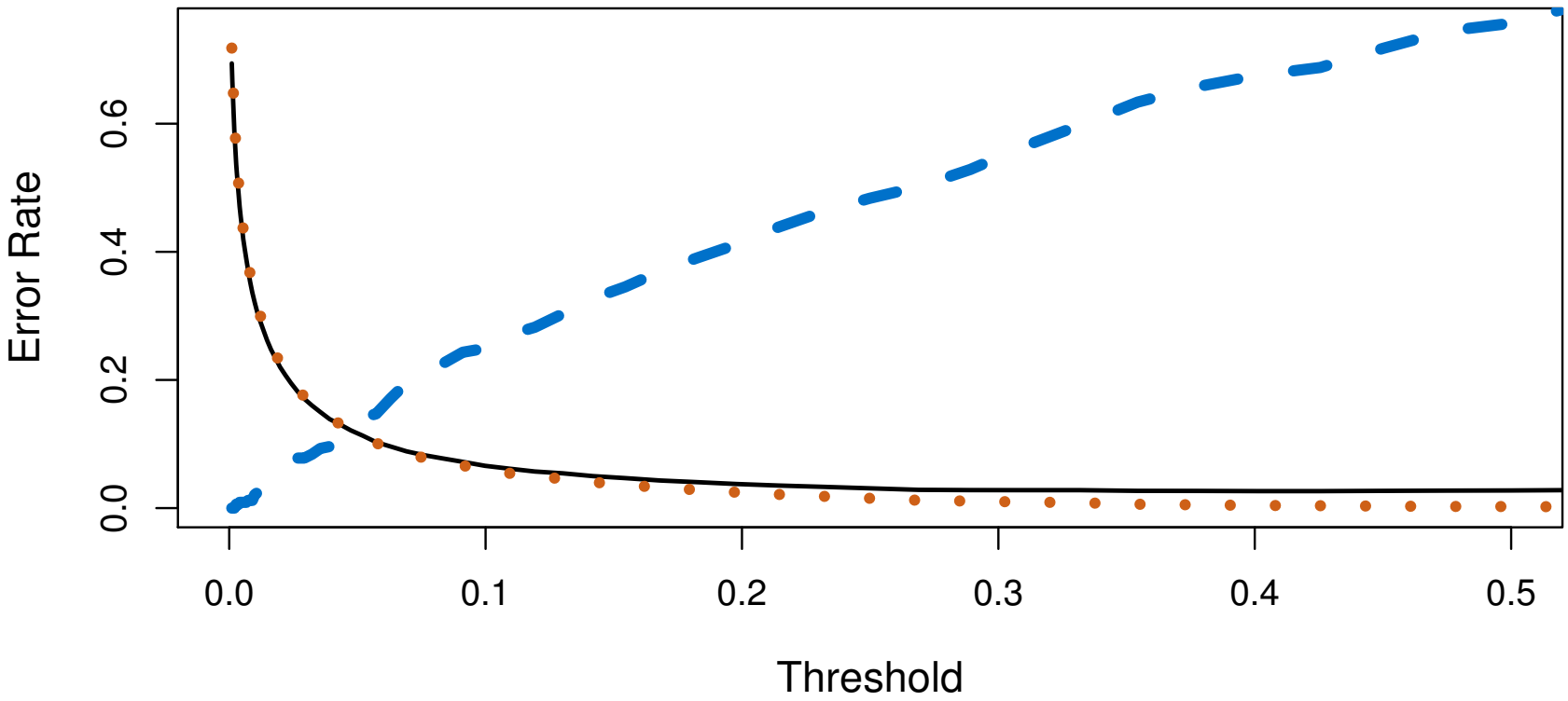
- Should we standardise the data?
- For K-means clustering:
  - How many clusters should we use?
- For hierarchical clustering:
  - Which dissimilarity measure?
  - What type of linkage?
  - Where should we cut the dendrogram?
- Other considerations
  - Soft clustering (e.g. using mixture models)
  - Clustering high-dimensional data

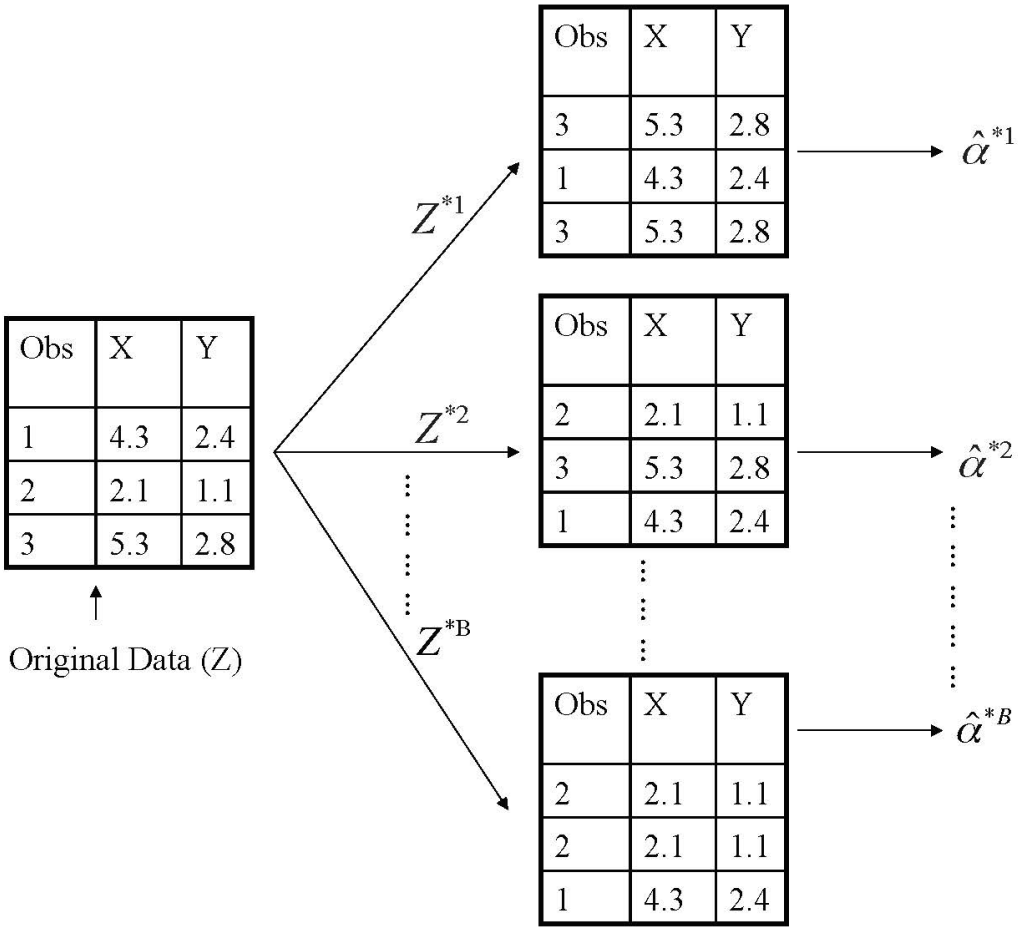
Summary

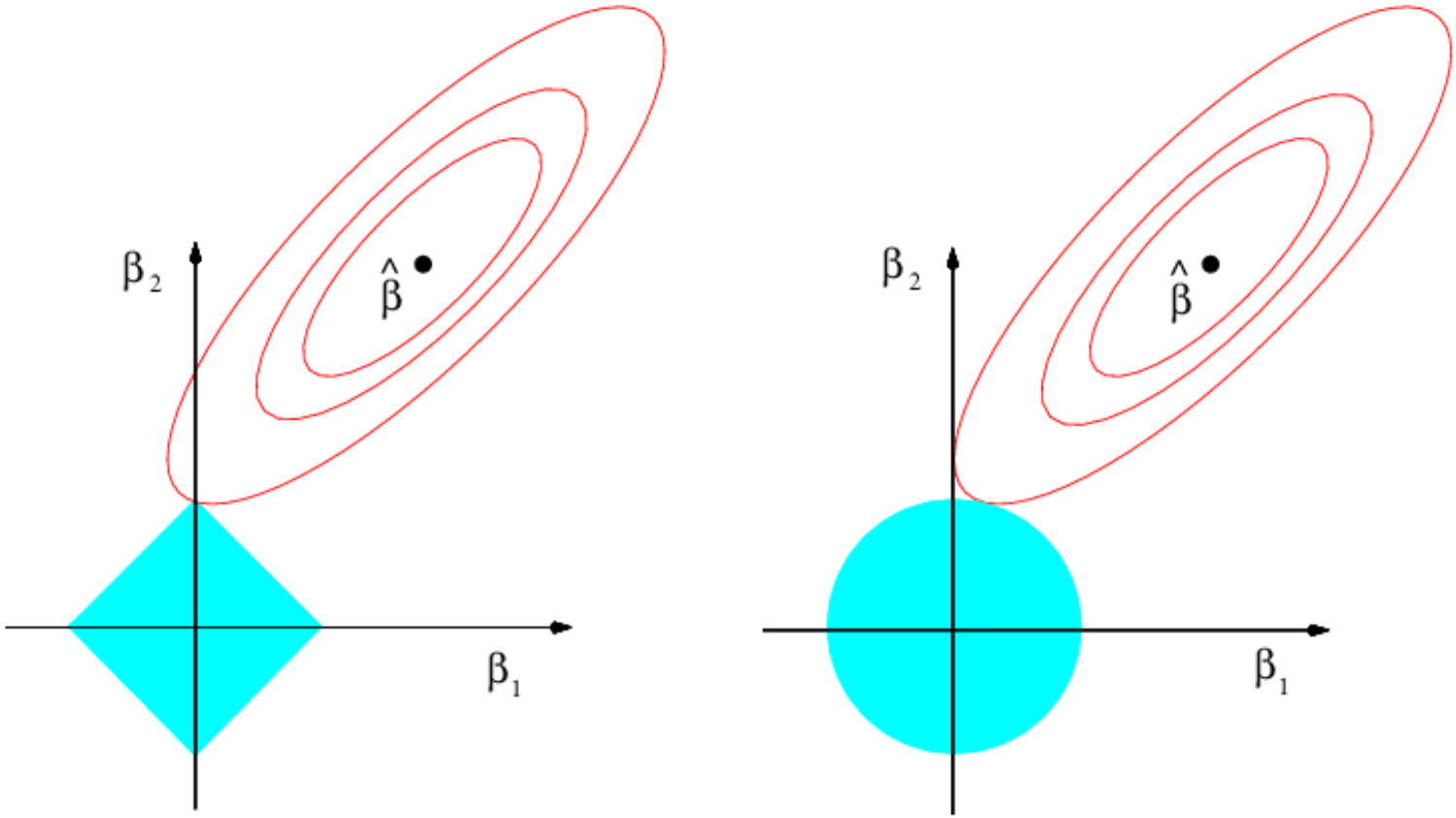
Week	Topic	Chapter	Lecturers
1	Introduction to business analytics & R	1	Rob,Souhaib
2	Statistical learning	2	Rob,Souhaib
3	Regression for prediction	3	Rob
4	Resampling	5	Rob, Souhaib
5	Dimension reduction	6,10	Rob, Souhaib
6	Visualization		Di
7	Visualization		Di
8	Classification	4	Souhaib, Di
9	Classification	4,9	Di, Souhaib
-	Semester Break		
10	Advanced classification	8	Di
11	Advanced regression	6	Di
12	Clustering	10	Di, Souhaib

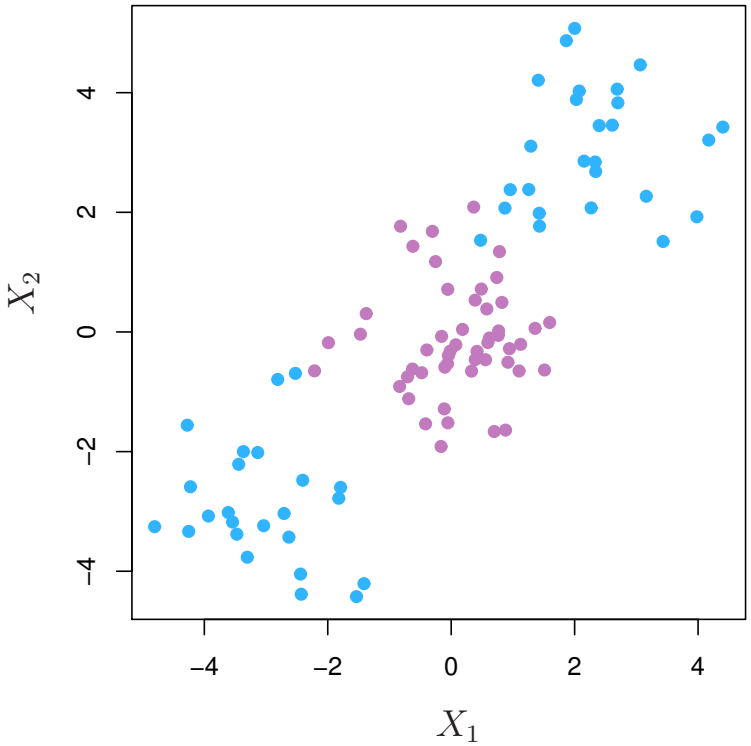

Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
High
Low
Interpretability
Low
High
Flexibility

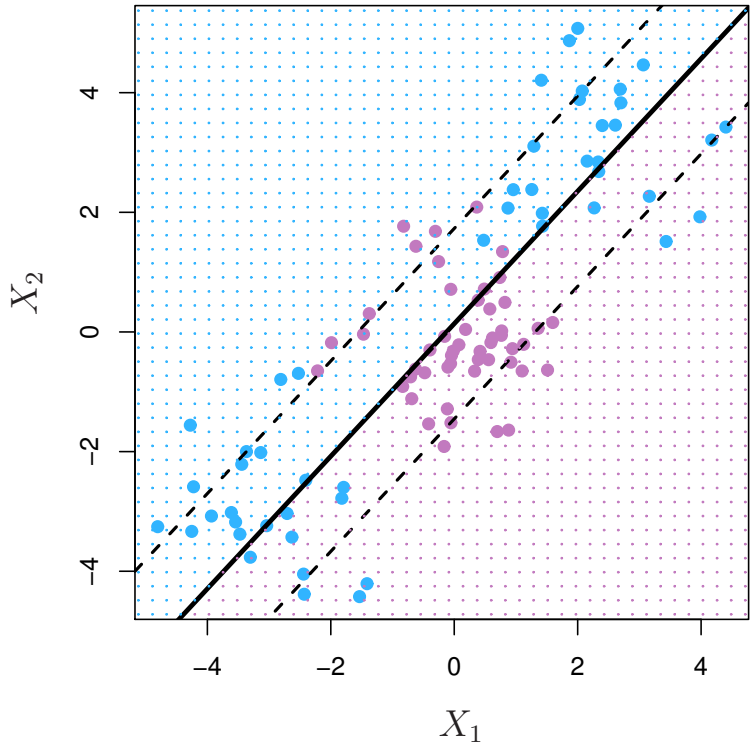

Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors

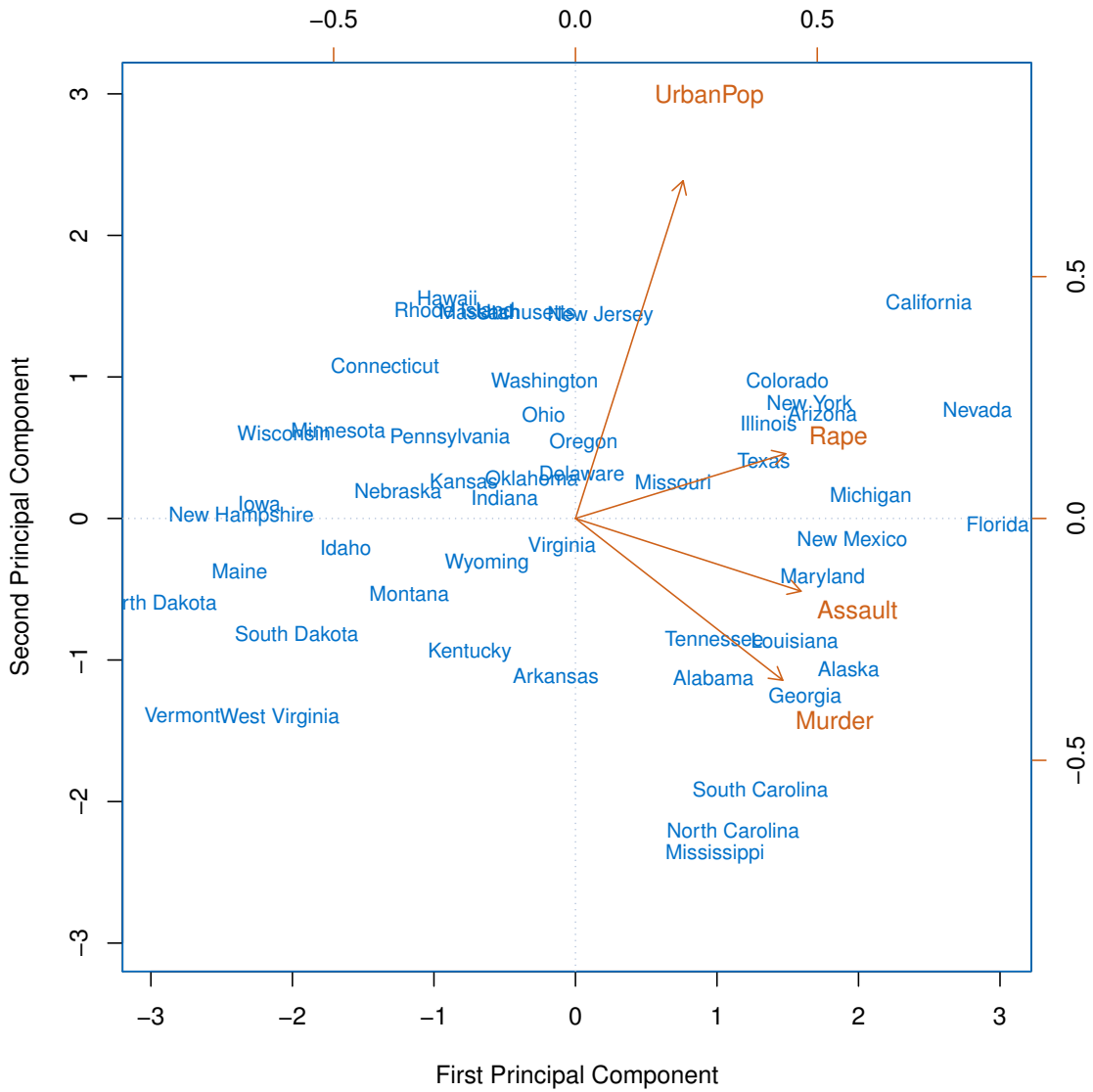

Error Rate
0.0
0.2
0.4
0.6
0.0
0.1
0.2
0.3
0.4
0.5
Threshold


Obs X Y
1 4.3 2.4
2 2.1 1.1
3 5.3 2.8
Original Data (Z)
Z^{*1}
Z^{*2}
Z^{*B}
Obs X Y
3 5.3 2.8
1 4.3 2.4
3 5.3 2.8
Obs X Y
2 2.1 1.1
3 5.3 2.8
1 4.3 2.4
Obs X Y
2 2.1 1.1
2 2.1 1.1
1 4.3 2.4
$\hat{\alpha}^{*1}$
$\hat{\alpha}^{*2}$
$\hat{\alpha}^{*B}$


β_2
$\hat{\beta}$
β_1
β_2
$\hat{\beta}$
β_1


X_2
4
2
0
−2
−4
−4
−2
0
2
4
X_1


X_2
4
2
0
−2
−4
−4
−2
0
2
4
X_1


−0.5
0.0
0.5
UrbanPop
Hawaii
New Jersey
California
Connecticut
Washington
Colorado
New York
Arizona
Nevada
Ohio
Illinois
Wisconsin
Minnesota
Pennsylvania
Oregon
Rape
Texas
Delaware
Kansas
Oklahoma
Missouri
Nebraska
Indiana
Michigan
Iowa
New Hampshire
Florida
Idaho
Virginia
New Mexico
Wyoming
Maine
Maryland
Montana
Assault
South Dakota
Tennessee
Louisiana
Kentucky
Alaska
Arkansas
Alabama
Georgia
Vermont
West Virginia
Murder
South Carolina
North Carolina
Mississippi
Second Principal Component
First Principal Component
3
2
1
0
−1
−2
−3
0.5
0.0
−0.5