Unsupervised Learning

ddebarr@uw.edu

2017-03-09

Interviewer: “I heard you were extremely quick at math.”
Me: “Yes, as a matter of fact I am.”
Interviewer: “What’s 14 x 27?”
Me: “49”
Interviewer: “That’s not even close.”
Me: “Yeah, but it was fast.”
/u/RandomHuman1578

Course Outline

1. Introduction to Statistical Learning
2. Linear Regression
3. Classification
4. Resampling Methods
5. Linear Model Selection and Regularization
6. Moving Beyond Linearity
7. Tree-Based Methods
8. Support Vector Machines
9. Unsupervised Learning
10. Neural Networks and Genetic Algorithms

Agenda

In Practice …

We’re probably using these methods for visualization (e.g. exploratory analysis) or to support supervised learning

- For example, earlier we used principal component analysis for regression
- We can use also use cluster membership information as predictors for supervised learning

First Principal Component

- Loadings

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

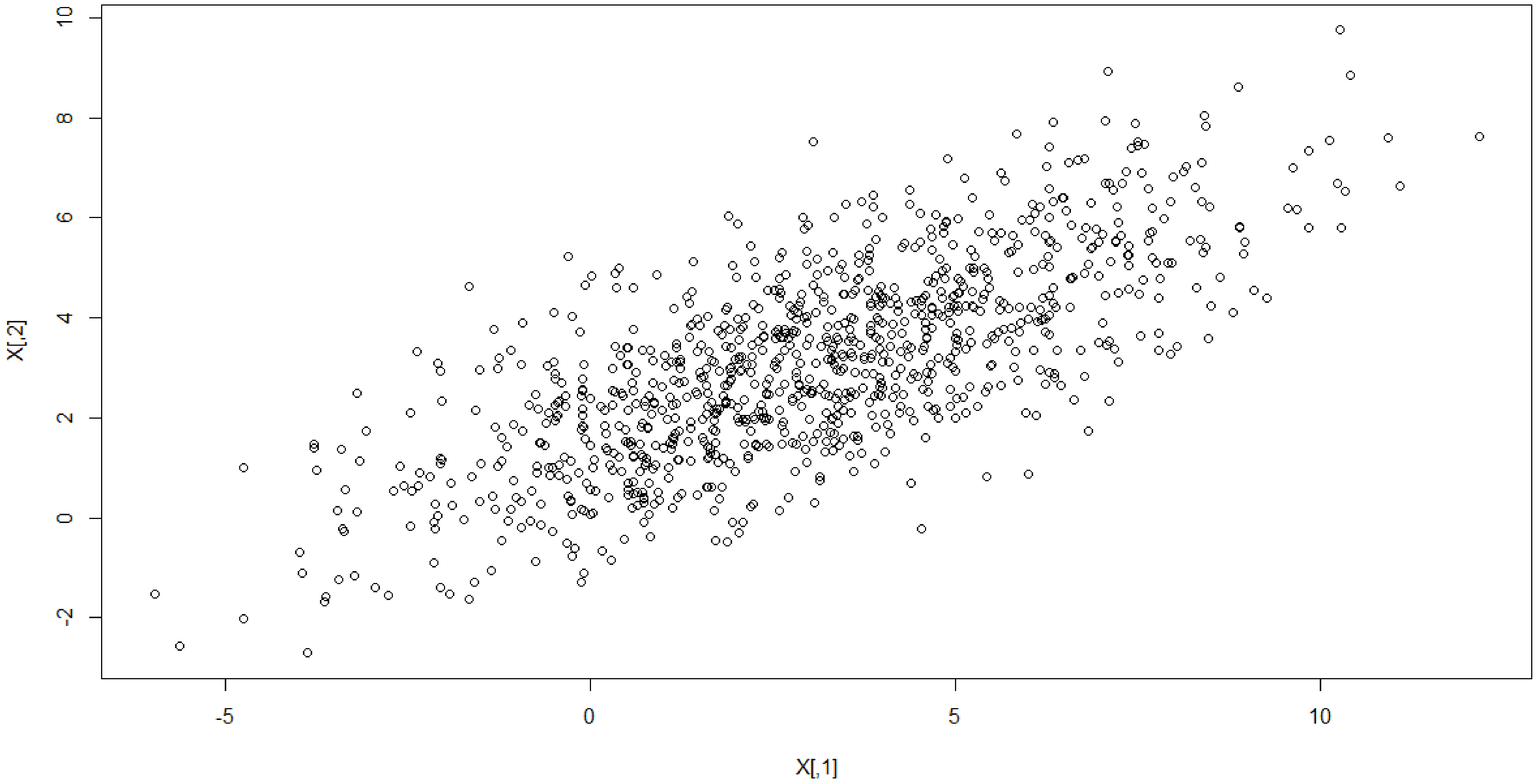
$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

- Scores

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

First Principal Component Optimization Problem

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


X[,2]
X[,1]
-2
0
2
4
6
8
10
-5
0
5
10

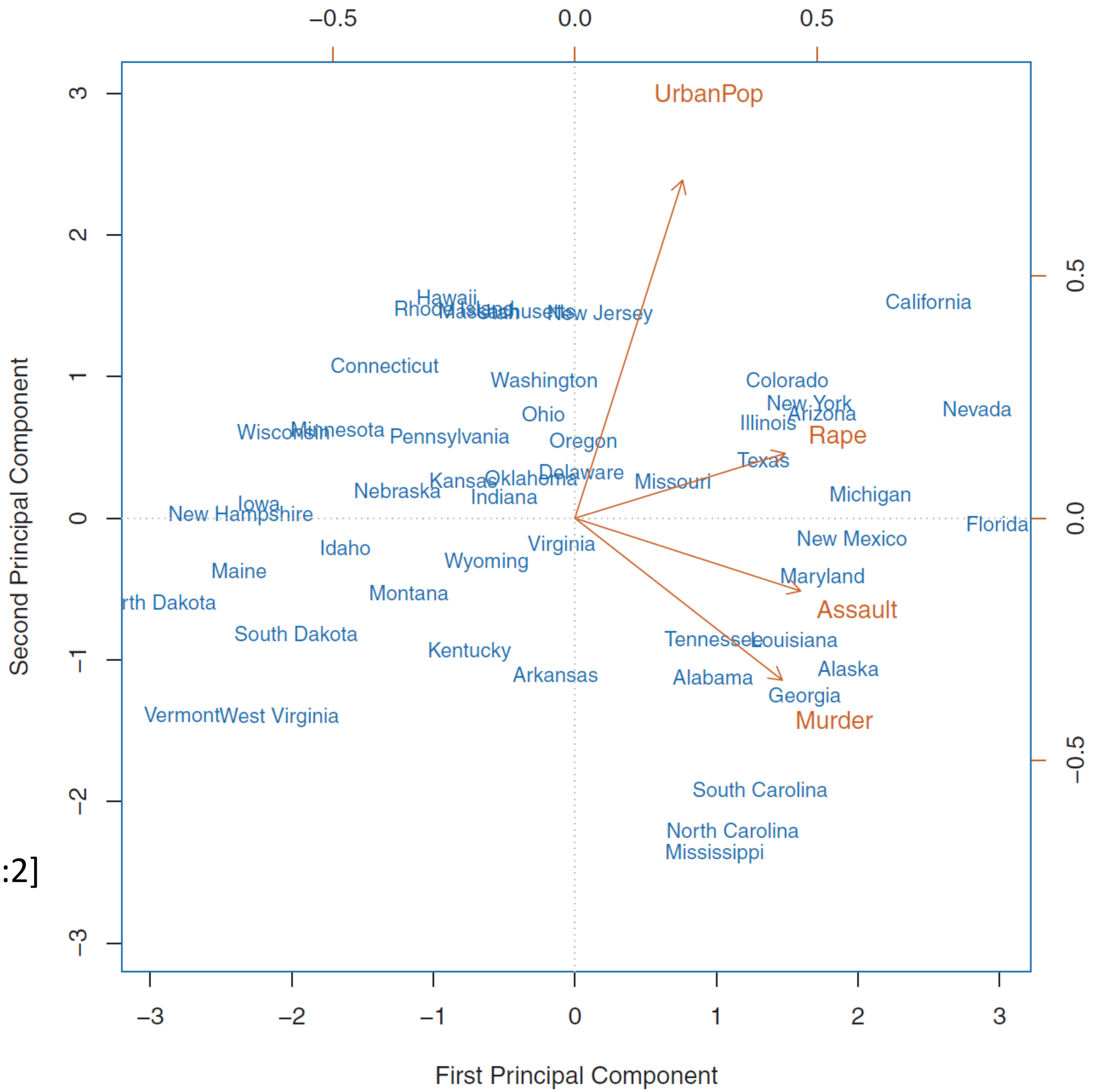
USArrests Data

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

USArrests Biplot

Q. What's better than one set of axes?

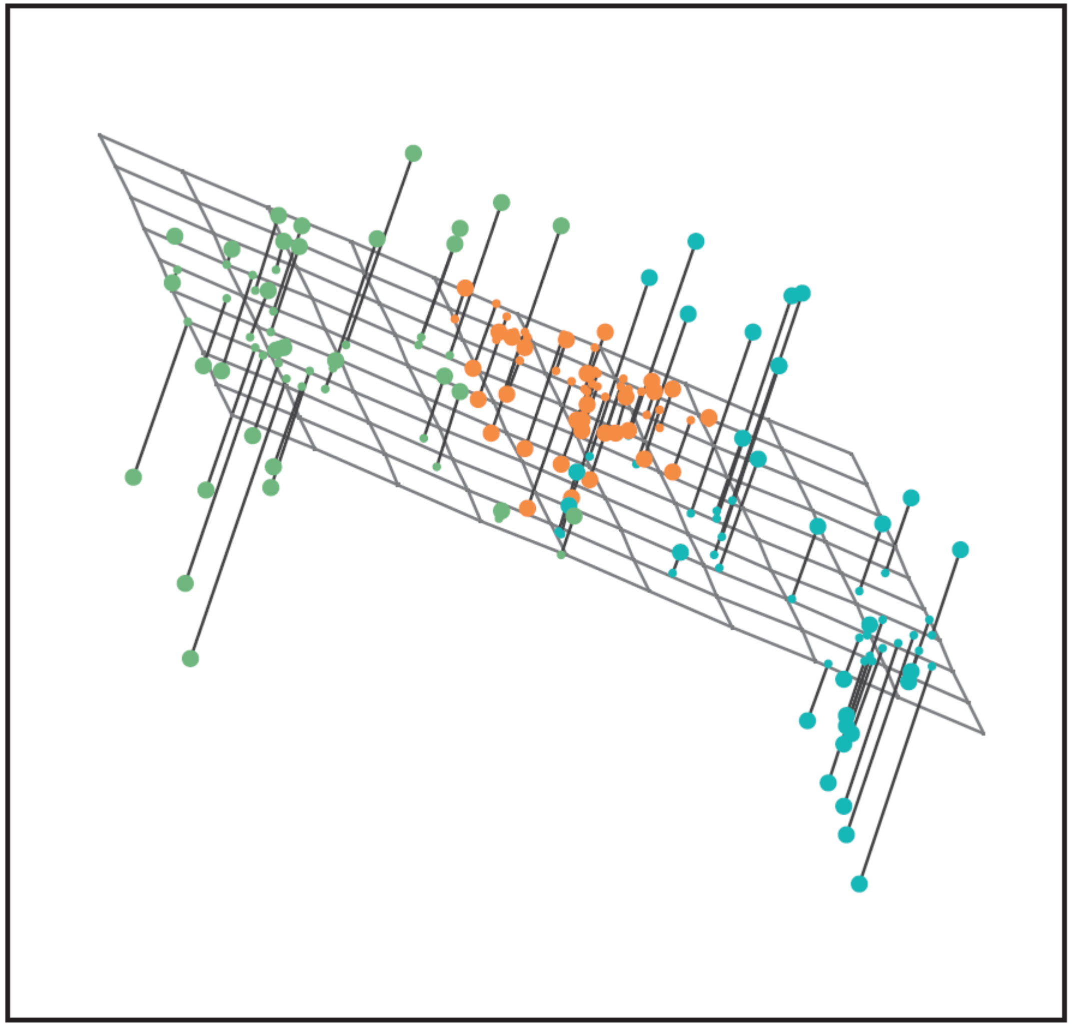
A. Two sets of axes ☺

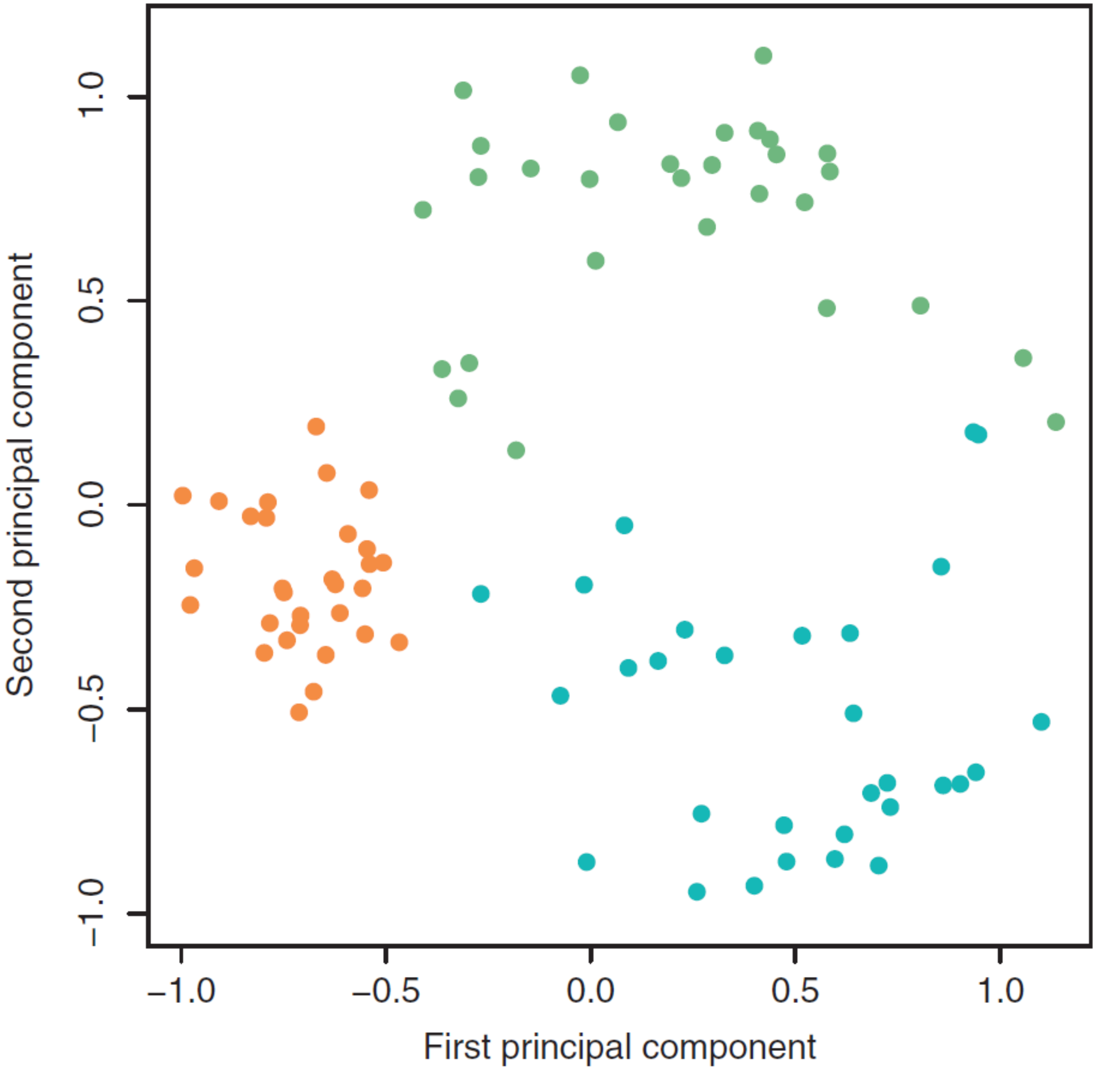


- prcomp(USArrests, scale = T)$rotation[,1:2]

- prcomp(USArrests, scale = T)$x[,1:2]

Minimizing the Sum of Squared Distances [Simulated Data]



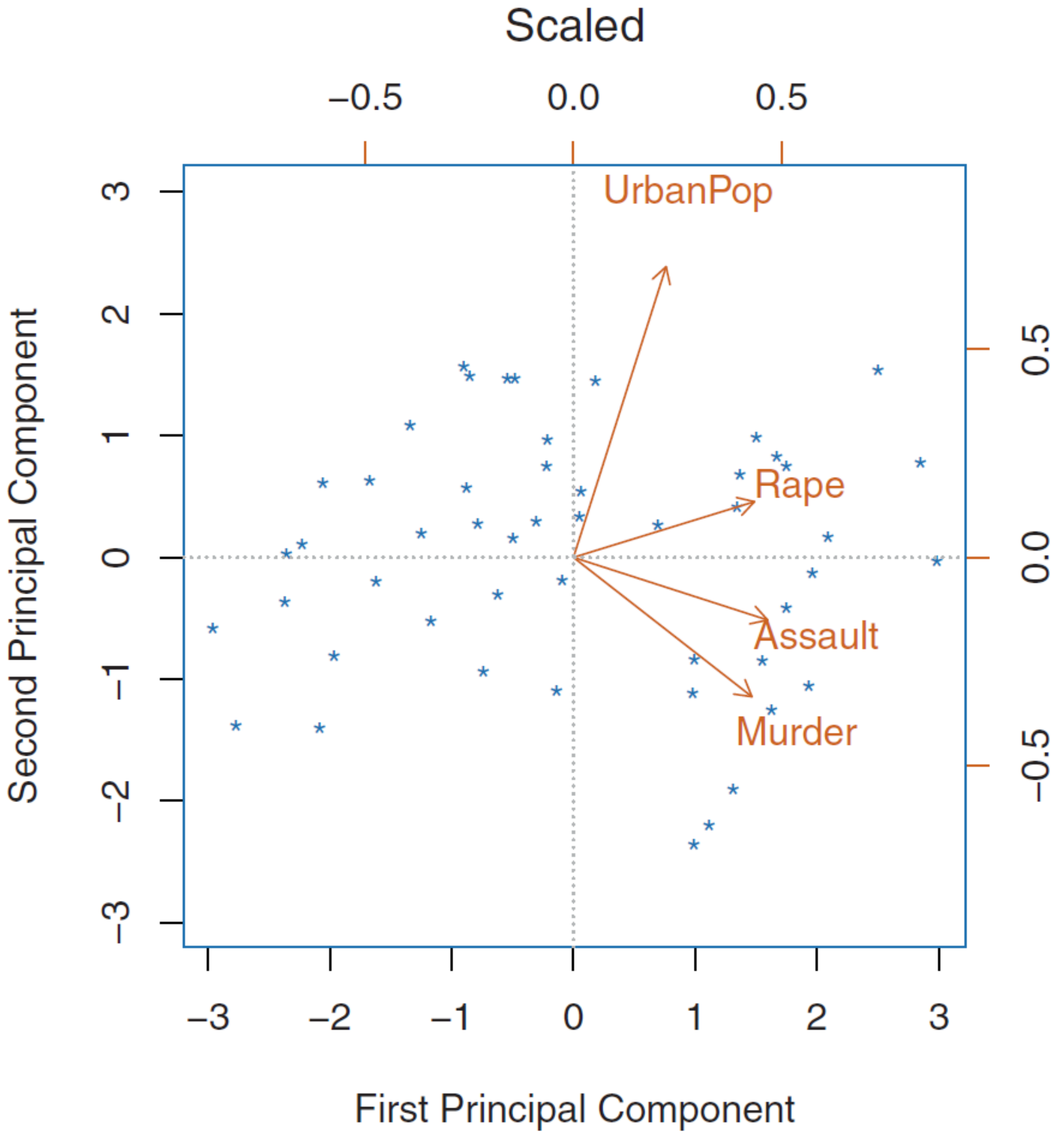


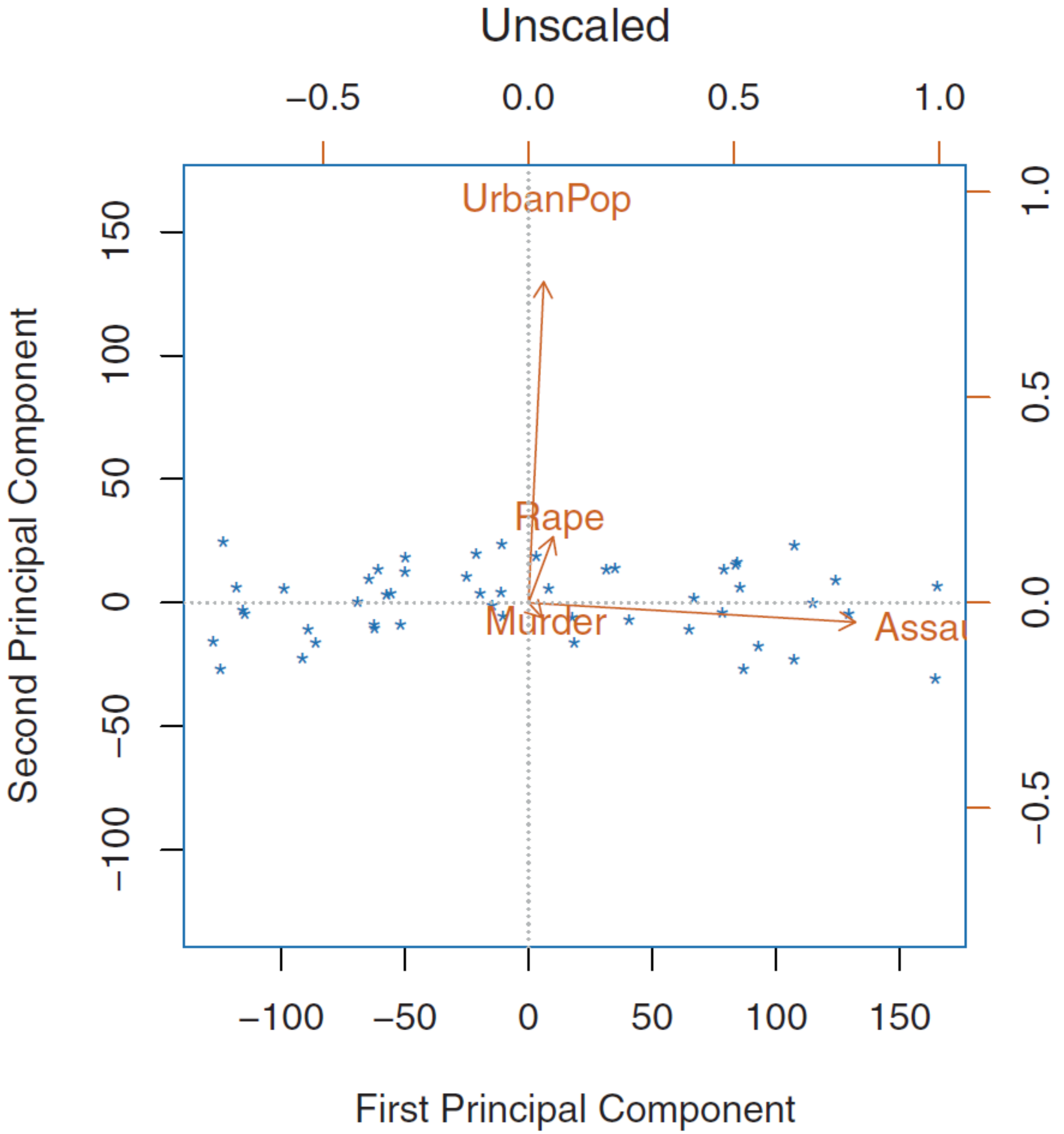
Reconstruction

Multiplying the scores by the loadings to approximate the original data

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}$$

USArrests: Scaled Versus Unscaled Solutions





Scaling Variables

Variables with larger variance can drive the output

```
> apply(USArrests, 2, mean)
  Murder  Assault UrbanPop     Rape 
   7.788  170.760   65.540   21.232 
> apply(USArrests, 2, var)
    Murder    Assault   UrbanPop       Rape 
  18.97047 6945.16571  209.51878   87.72916 
```

Proportion of Variance Explained (PVE)

- Total Variance

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

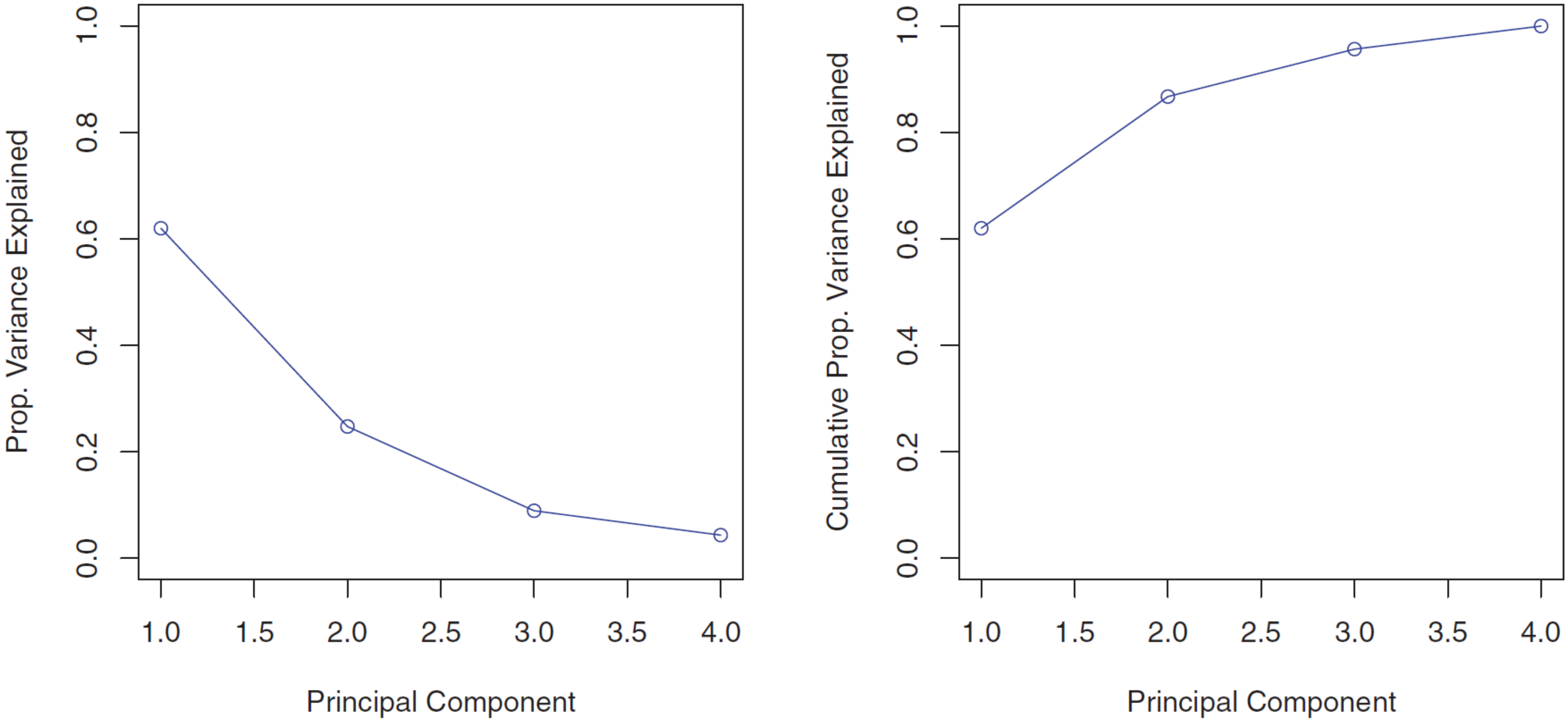
- Variance Explained by the mth Principal Component

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

- Proportion of Variance Explained

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

Scree and Cumulative PVE Plots



Crisp Clustering [Disjoint Clusters]

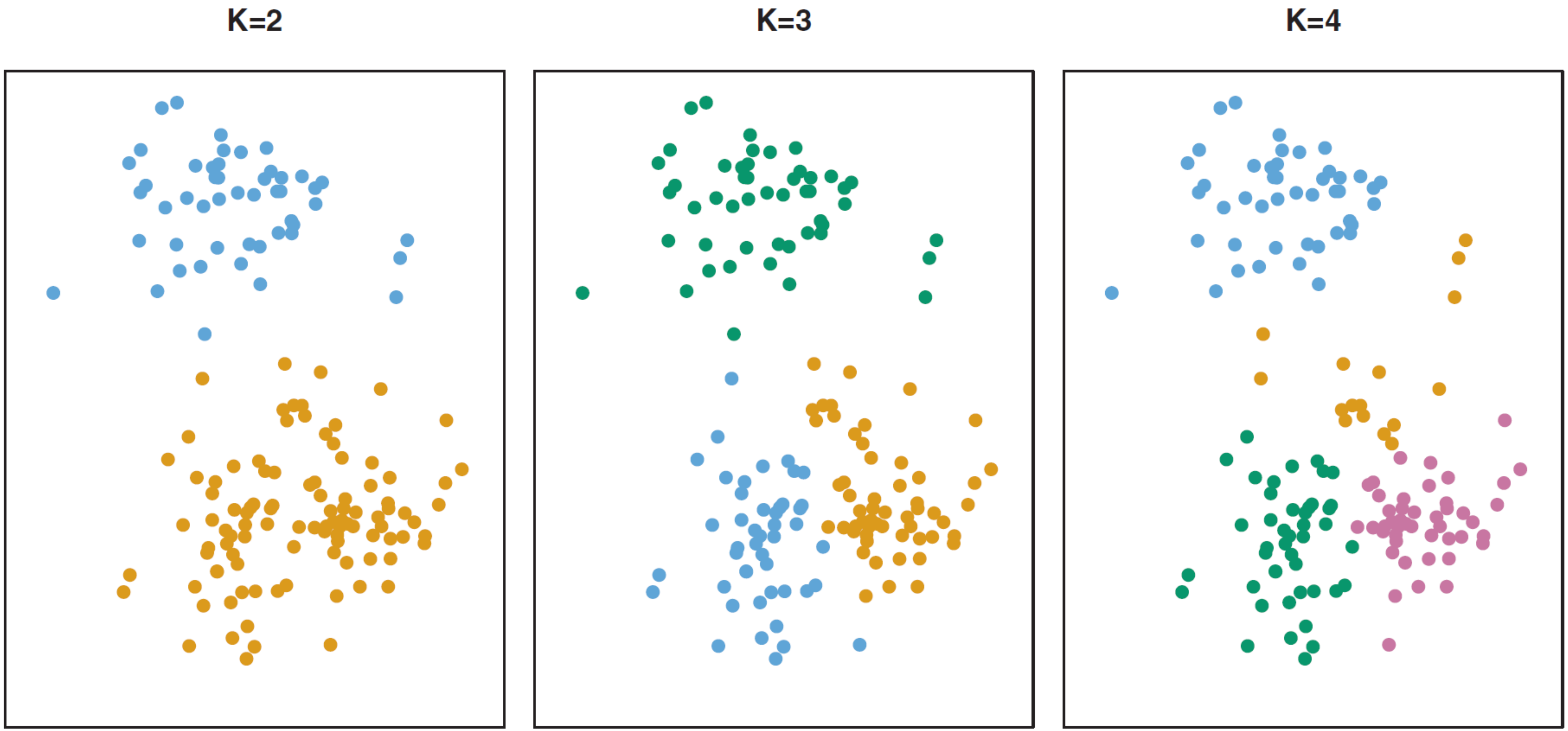
Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster.

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 $i \in C_k$.

Example Clusterings for Simulated Data



Objective Function

Minimize within class variation …

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

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

Alternative Within Cluster Variation Expression [first expression equals last expression]

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

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

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

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

$$=2\sum_{j=1}^{p}\sum_{i\in C_k}\left(x_{ij}^2-\bar{x}_{kj}x_{ij}\right)=2\sum_{j=1}^{p}\sum_{i\in C_k}\left[x_{ij}\left(x_{ij}-\bar{x}_{kj}\right)\right]=2\sum_{j=1}^{p}\sum_{i\in C_k}\left[\left(x_{ij}-\bar{x}_{kj}+\bar{x}_{kj}\right)\left(x_{ij}-\bar{x}_{kj}\right)\right]$$

$$=2\sum_{j=1}^{p}\sum_{i\in C_k}\left(x_{ij}-\bar{x}_{kj}\right)^2+2\sum_{j=1}^{p}\sum_{i\in C_k}\left[\bar{x}_{kj}\left(x_{ij}-\bar{x}_{kj}\right)\right]=2\sum_{j=1}^{p}\sum_{i\in C_k}\left(x_{ij}-\bar{x}_{kj}\right)^2+2\sum_{j=1}^{p}\left[\bar{x}_{kj}\sum_{i\in C_k}\left(x_{ij}-\bar{x}_{kj}\right)\right]$$

$$=2\sum_{j=1}^{p}\sum_{i\in C_k}\left(x_{ij}-\bar{x}_{kj}\right)^2+2\sum_{j=1}^{p}\left[\bar{x}_{kj}\left(\sum_{i\in C_k}x_{ij}-|C_k|\,\bar{x}_{kj}\right)\right]=2\sum_{j=1}^{p}\sum_{i\in C_k}\left(x_{ij}-\bar{x}_{kj}\right)^2+2\sum_{j=1}^{p}\left[\bar{x}_{kj}\left(\sum_{i\in C_k}x_{ij}-\sum_{i\in C_k}x_{ij}\right)\right]$$

$$=2\sum_{j=1}^{p}\sum_{i\in C_k}\left(x_{ij}-\bar{x}_{kj}\right)^2+0=2\sum_{j=1}^{p}\sum_{i\in C_k}\left(x_{ij}-\bar{x}_{kj}\right)^2=2\sum_{i\in C_k}\sum_{j=1}^{p}\left(x_{ij}-\bar{x}_{kj}\right)^2$$

Minimize squared distance to the mean

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:

   (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).

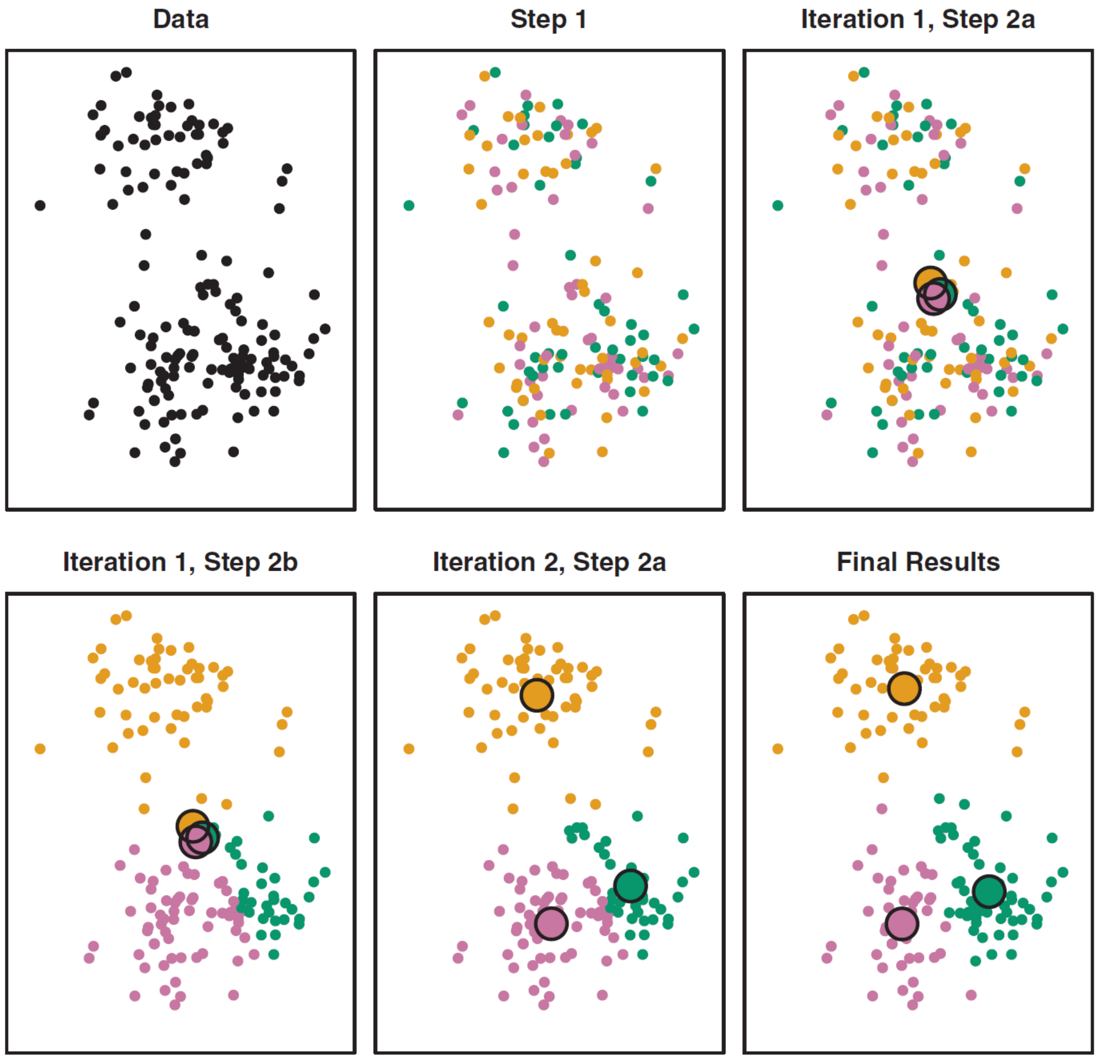
Implementation Note

Clusters are initialized to randomly selected observations

```
> stats::kmeans
centers <- x[sample.int(n, k), ]
```

Iterative Expectation Maximization (EM)

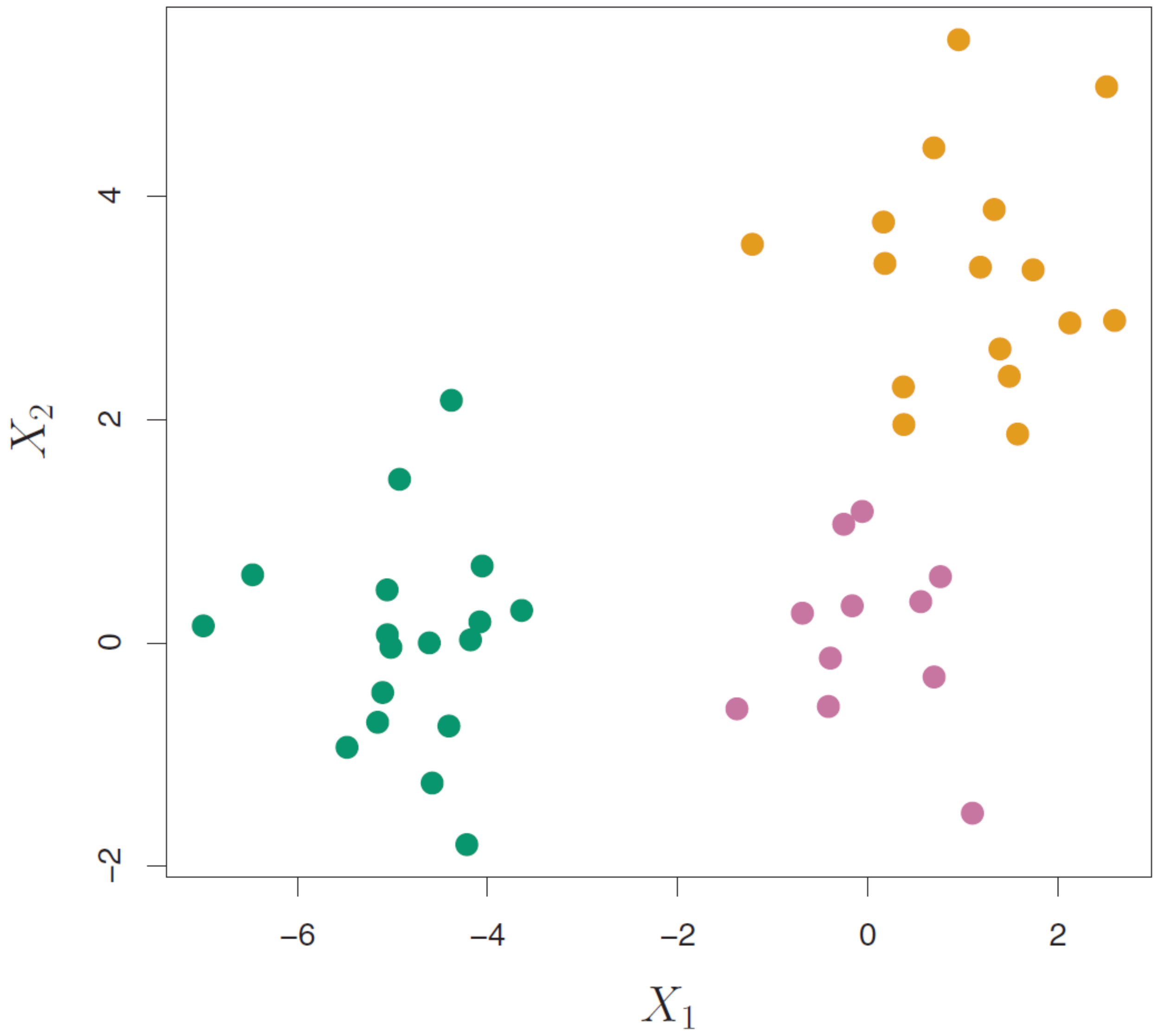
- Step 2b: assign observations to clusters [expectation]
- Step 2a: update the cluster centroids [maximization]
- Neither step will increase the value of the objective function [they're designed to reduce it]



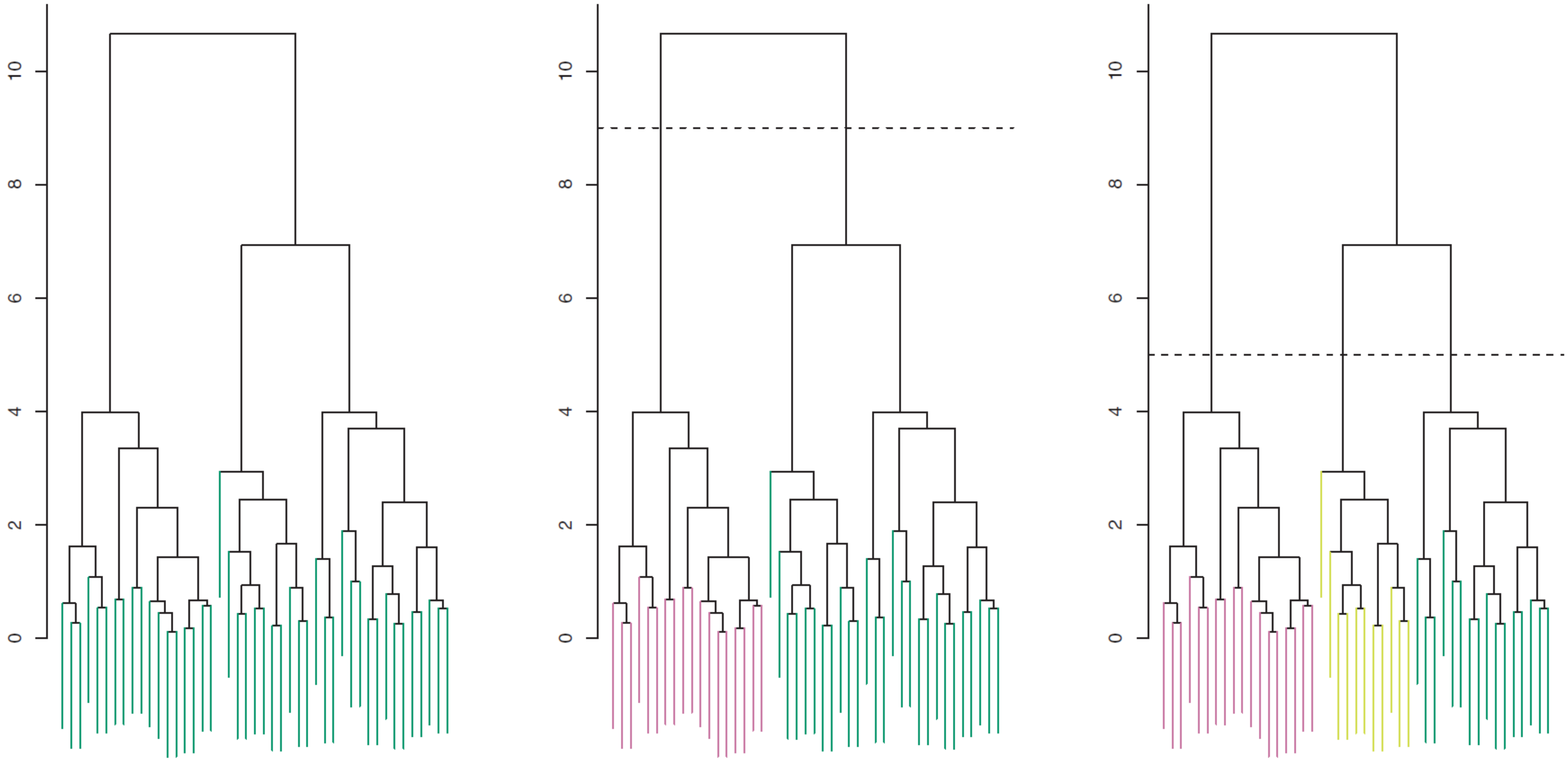
Multiple Starts [random initializations]

Perform the k-means clustering procedure multiple times and select the model that produces the lowest value for the objective function

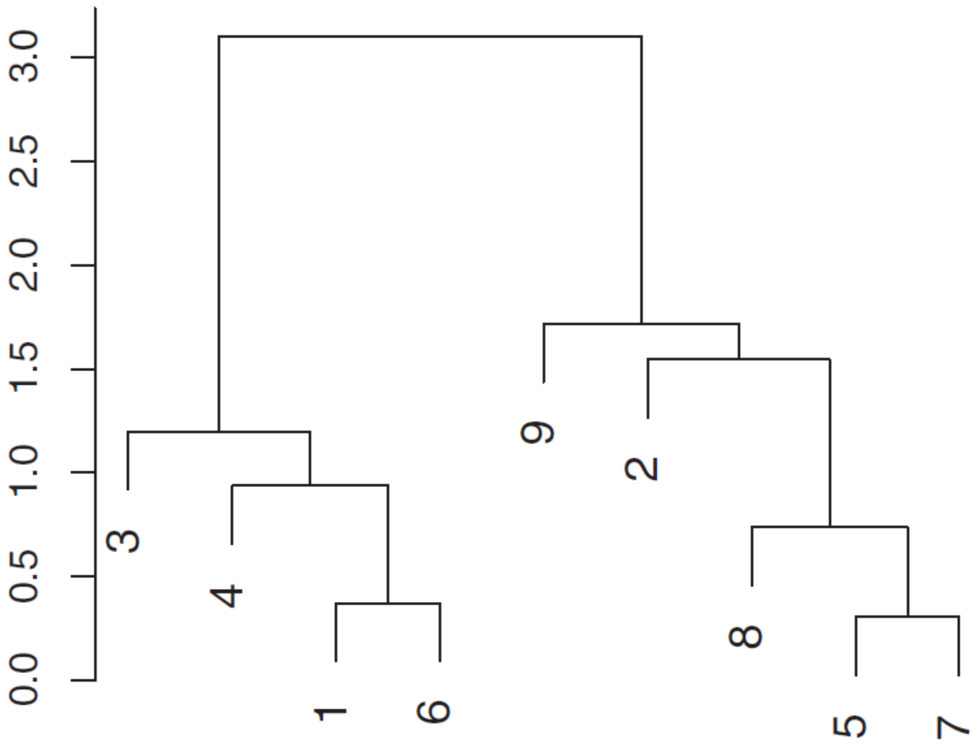
Simulated Data for Hierarchical Clustering

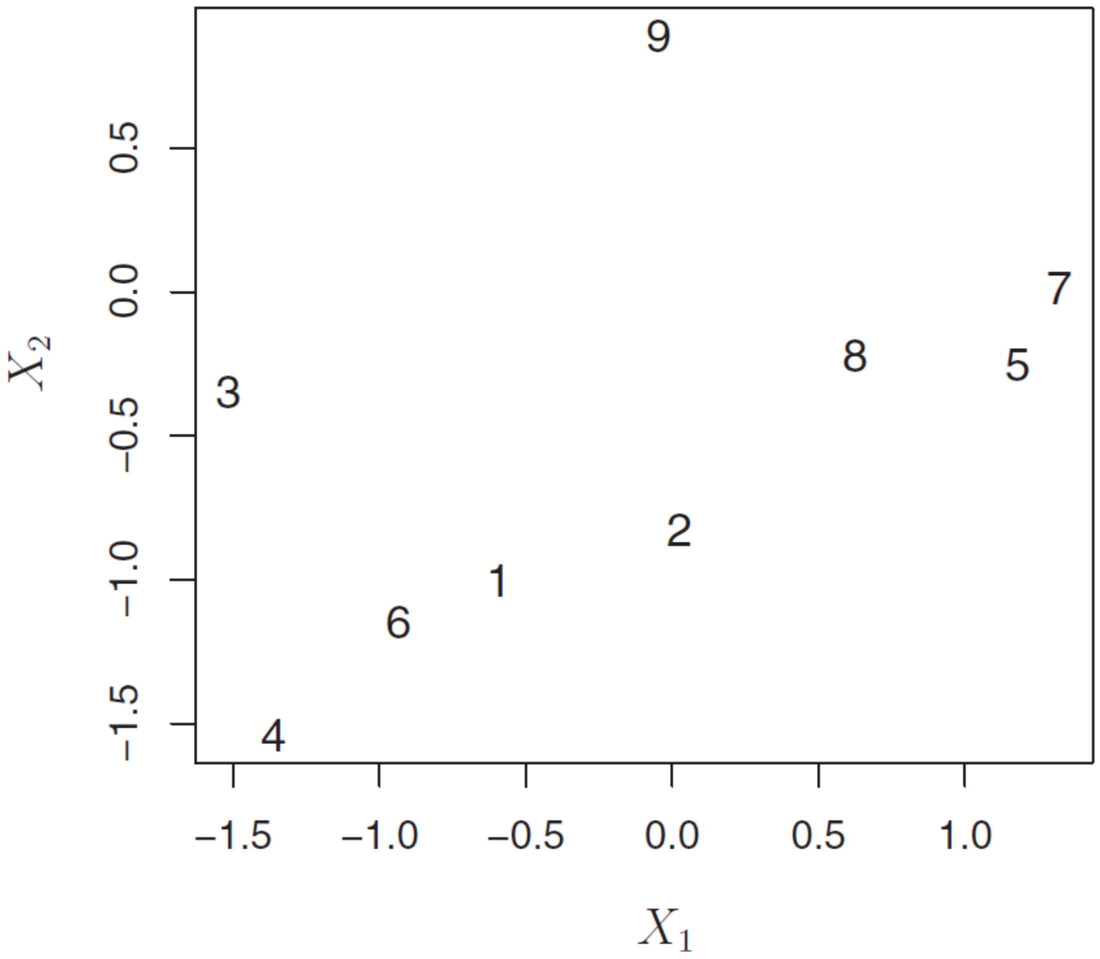


Hierarchical Cluster Analysis Dendrogram



Dendrogram Example





Hierarchical Clustering Algorithm

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

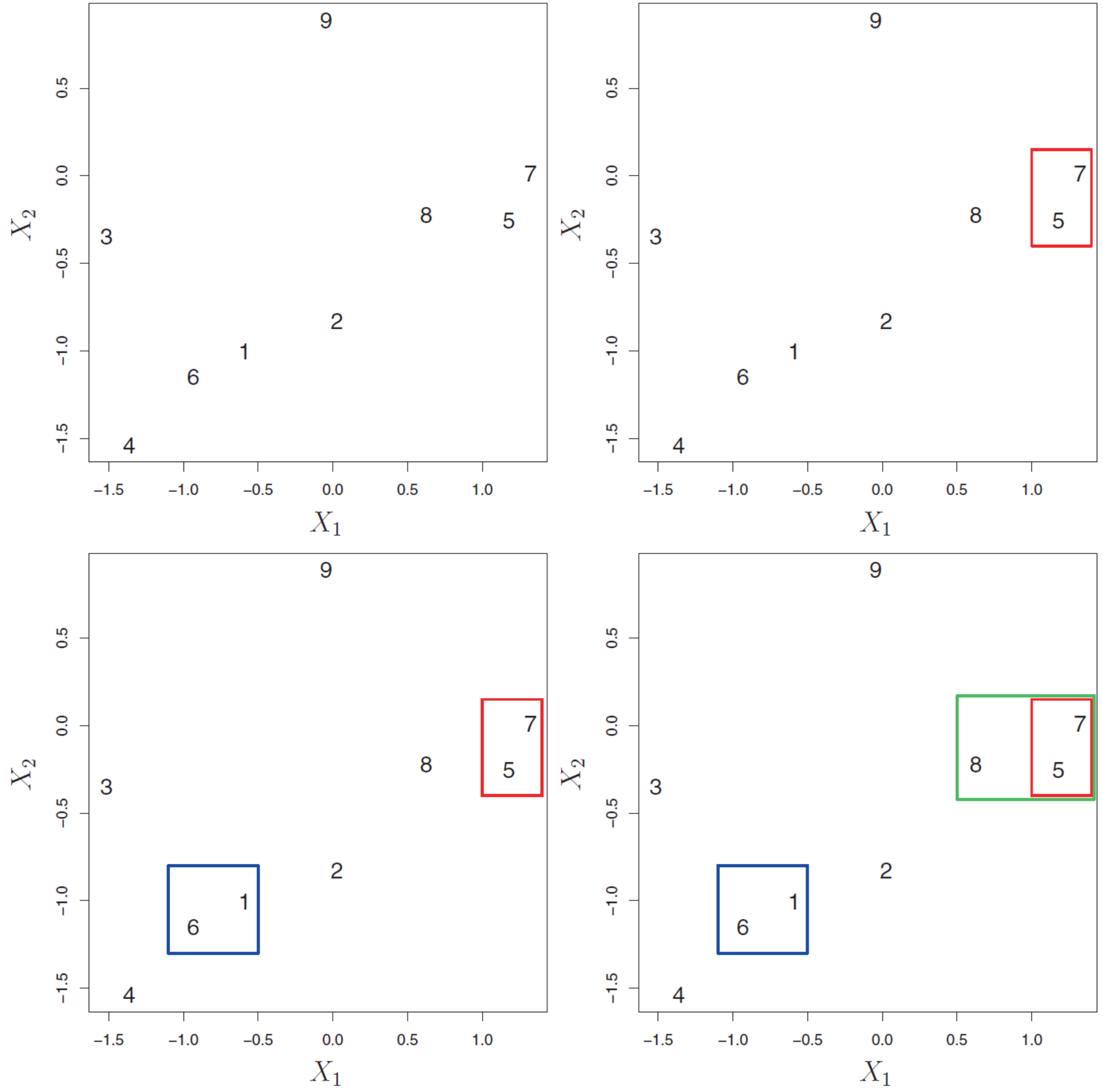
   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

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

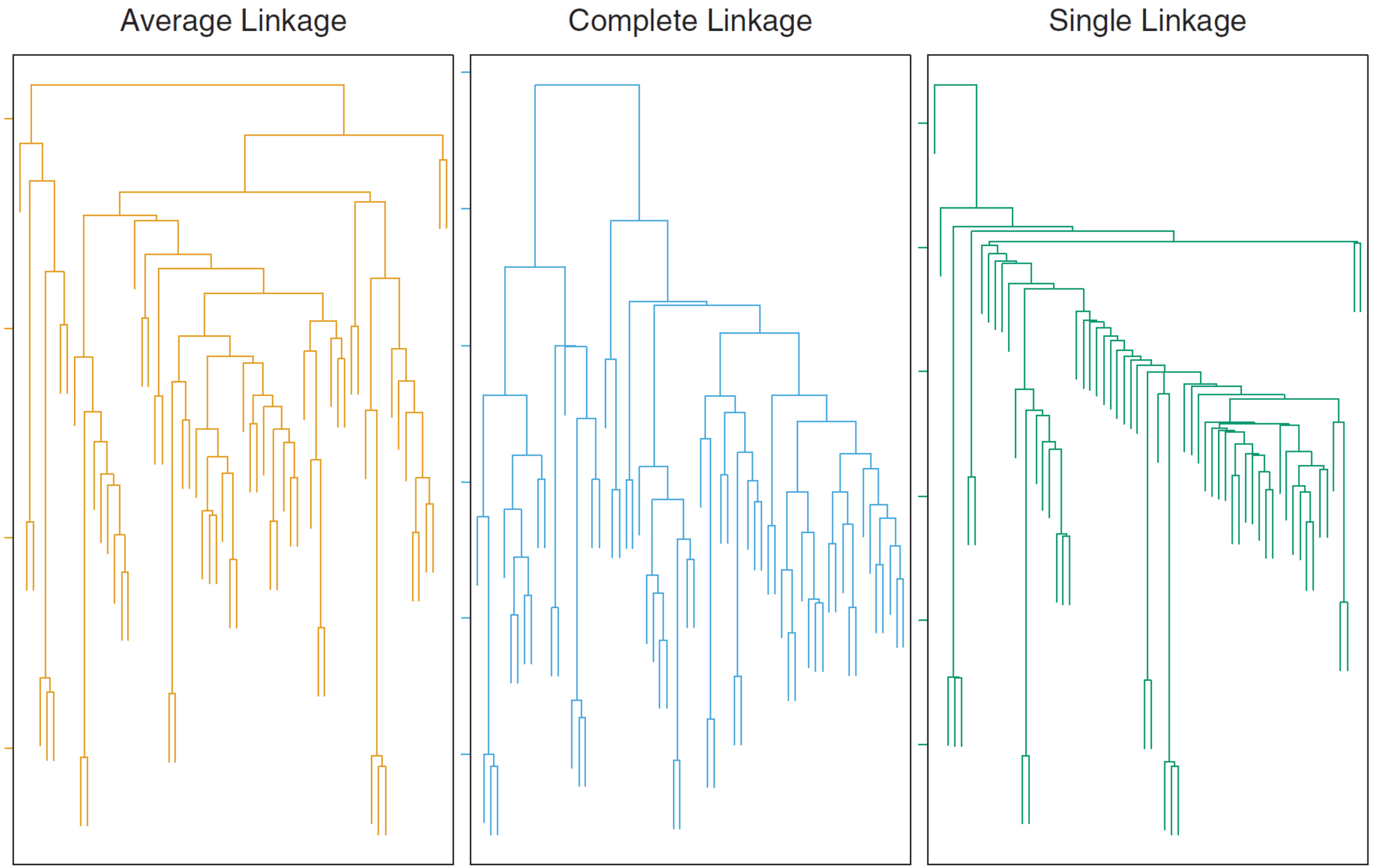
Hierarchical Clustering Linkage [distance between groups]

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*.

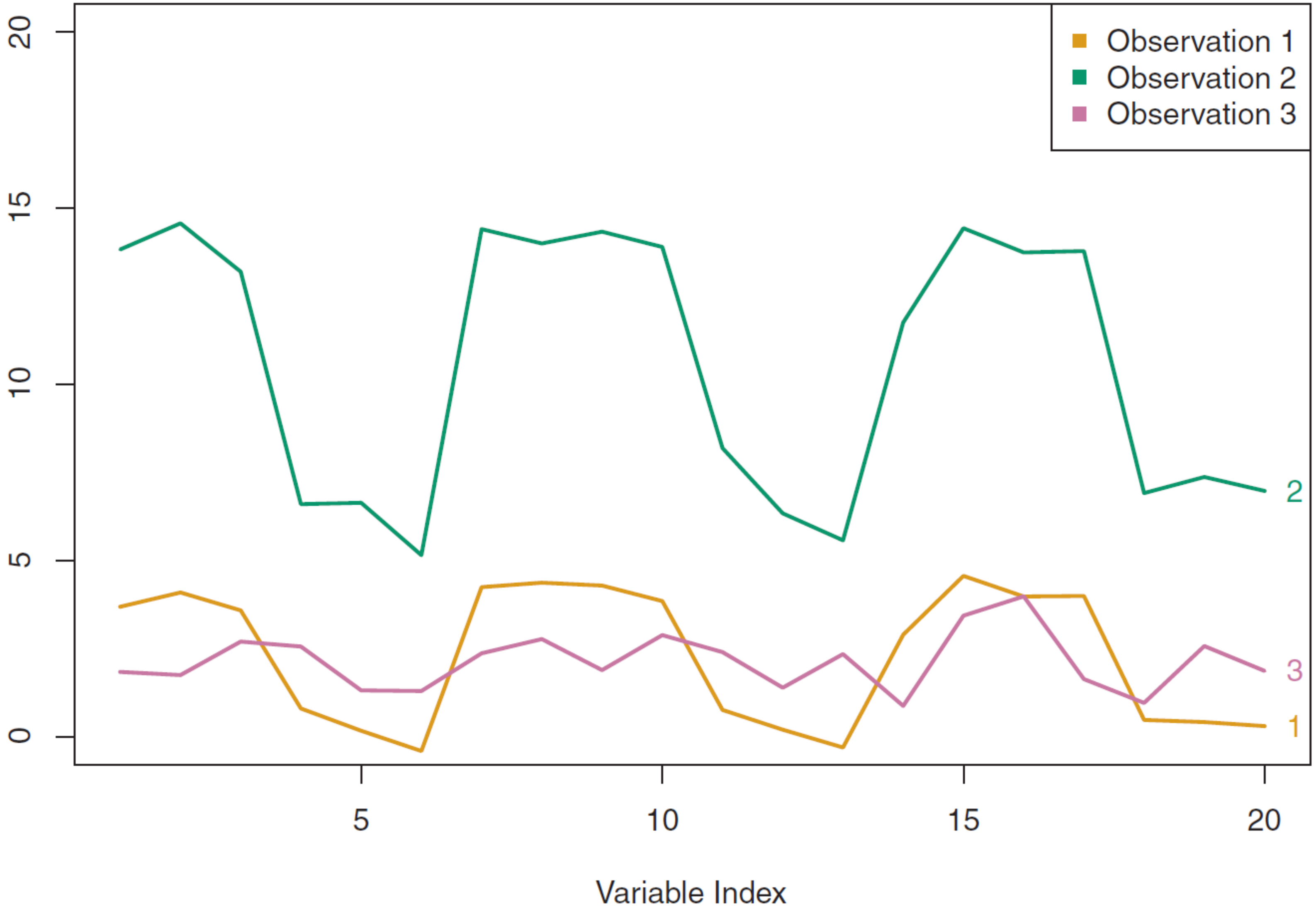
First Three Steps for Hierarchical Clustering



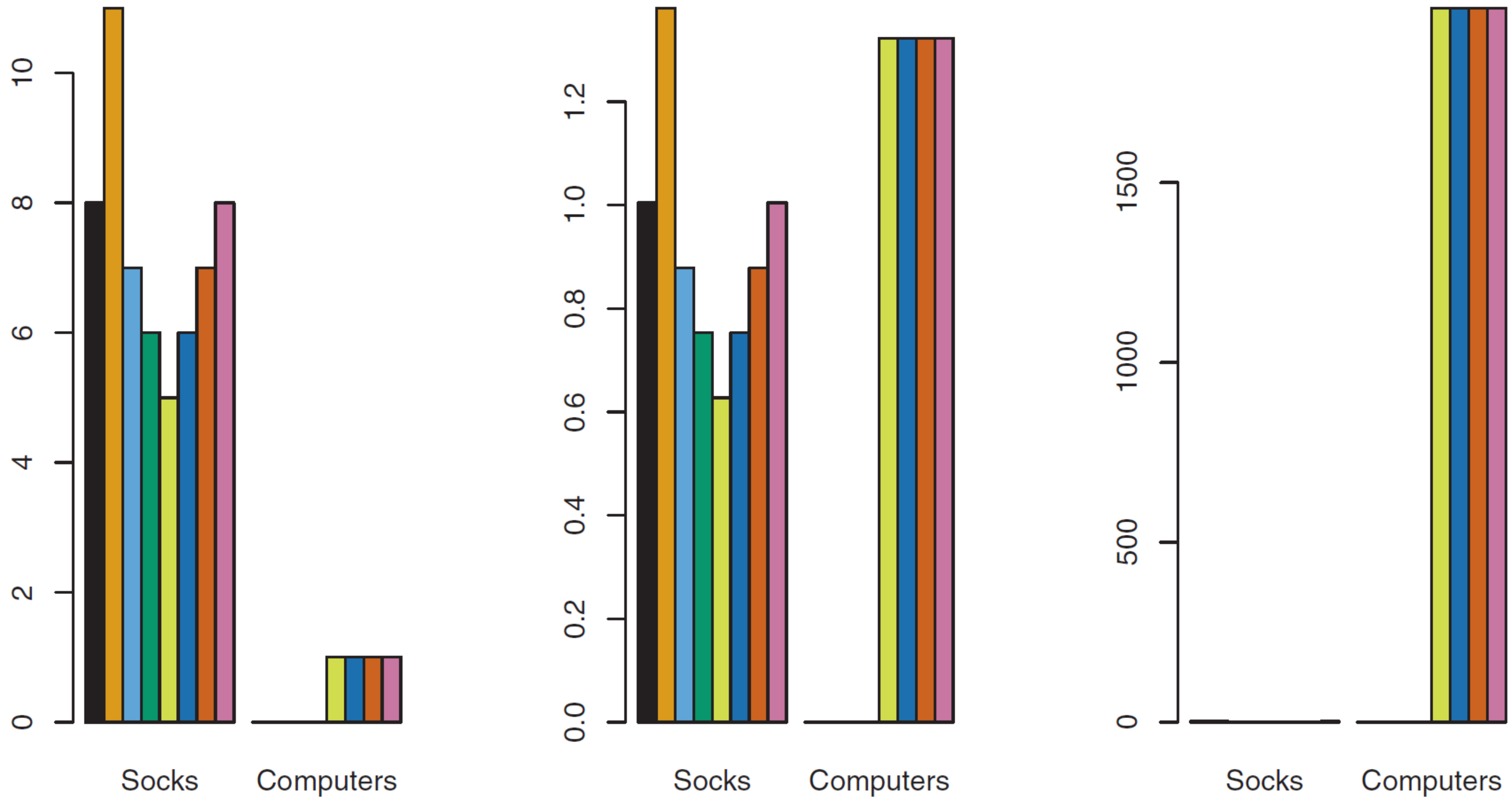
Linkage:
Average versus Complete versus Single



Euclidean versus Correlation Based Distance



To Scale or Not To Scale?



Decisions for Clustering

- Should the observations be standardized? (e.g. centered, scaled)
- What dissimilarity measure should be used?
- For hierarchical clustering, what type of linkage should be used?
- How many clusters?

Agenda