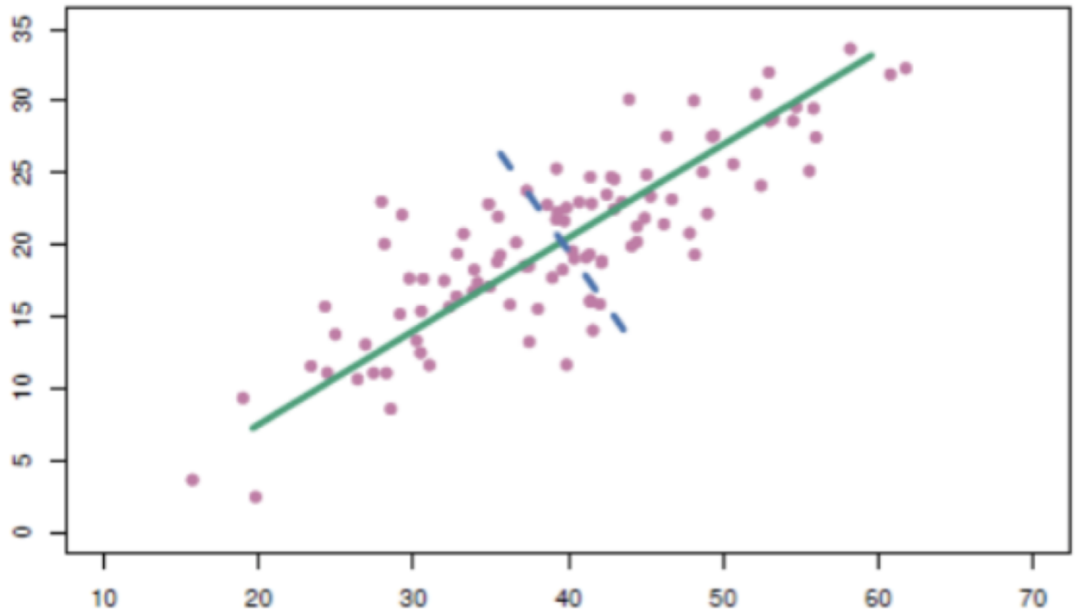
PCA and Clustering

MSBA7002 Business Statistics Tutorial 4

Table of Contents

Principal Component Analysis

- Unsupervised Learning
  - X supervises itself
- Geometry Intuition: ”rotation”



Principal Component Analysis

- ”Intrinsic” purpose: **Dimension Reduction**
  - Use reduced number of uncorrelated dimensions
  - Retain as much variance
- ”Extrinsic” purpose:
  i. Solve multicolinearity
  ii. Number of predictor is too large
- Assumption:
  - There are not any outliers
- Data-preprocessing procedure:
  – Normalization is recommended

$$\tilde{X} = \frac{X - \hat{\mu}_X}{\hat{\sigma}_X}$$

Principal Component Analysis

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}_{n \times p}$$

$$= \underbrace{\begin{pmatrix} s_{11} & s_{12} & \dots & s_{1k} \\ s_{21} & s_{22} & \dots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \dots & s_{nk} \end{pmatrix}_{n \times k}}_{\text{PC Score}} \underbrace{\begin{pmatrix} l_{11} & l_{12} & \dots & l_{1p} \\ l_{21} & l_{22} & \dots & l_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ l_{k1} & l_{k2} & \dots & l_{kp} \end{pmatrix}_{k \times p}}_{\text{PC loading}}$$

Principal Component Analysis

$$\overbrace{\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nk} \end{pmatrix}}^{\text{PC Score}}_{n\times k} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n\times p} \underbrace{\begin{pmatrix} l_{11} & l_{21} & \cdots & l_{k1} \\ l_{11} & l_{22} & \cdots & l_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ l_{1p} & l_{2p} & \cdots & l_{kp} \end{pmatrix}_{p\times k}}_{\text{PC loading}}$$

- Computation method
  - SVD (Singular Value Decomposition)

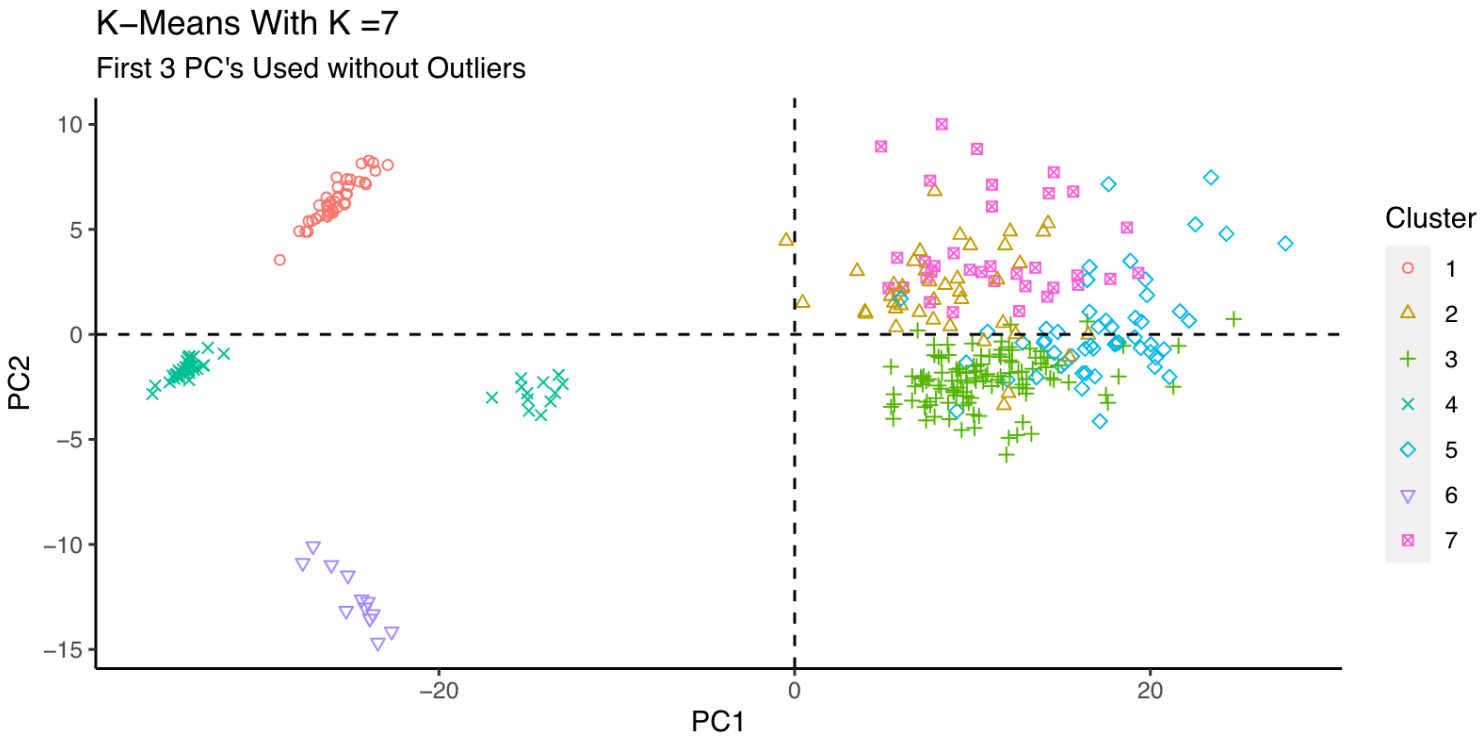
Principal Component Analysis

- Popular outputs used
  i. PC score
  Be used do further research
  → Data Visualization
  → Linear Discriminant Analysis
  → Clustering
  → Principal Component Regression

Principal Component Analysis

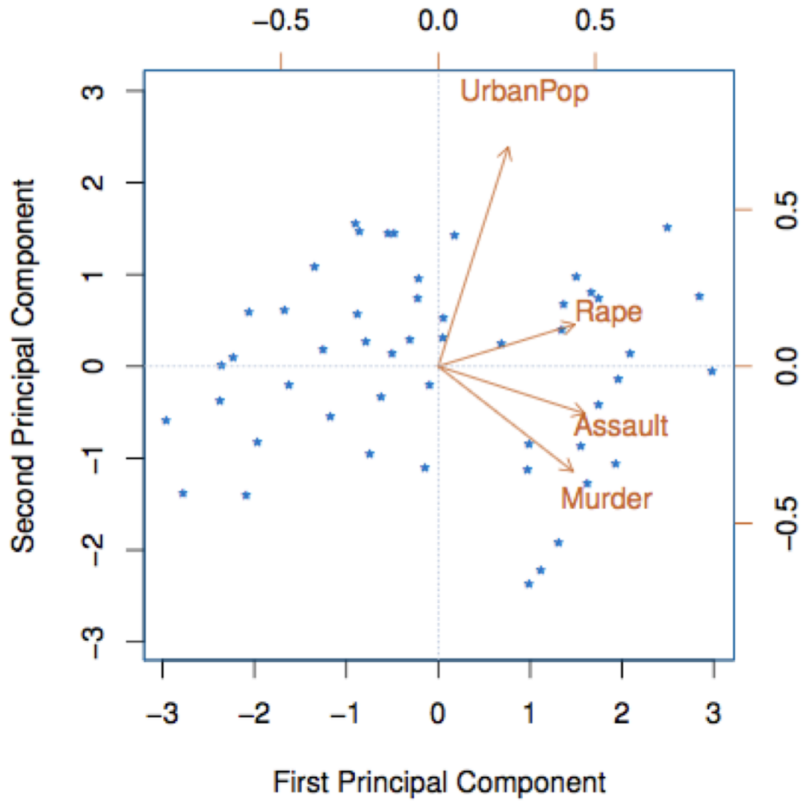
- Popular outputs used
  - i. PC score

    Be used do further research
    - → Data Visualization ✓
    - → Linear Discriminant Analysis
    - → Clustering ✓
    - → Principal Component Regression ✓



Principal Component Analysis

- Popular outputs used
  - ii. PC loading
    Reveals the relationship
    between PCs and original variables
    → Biplot

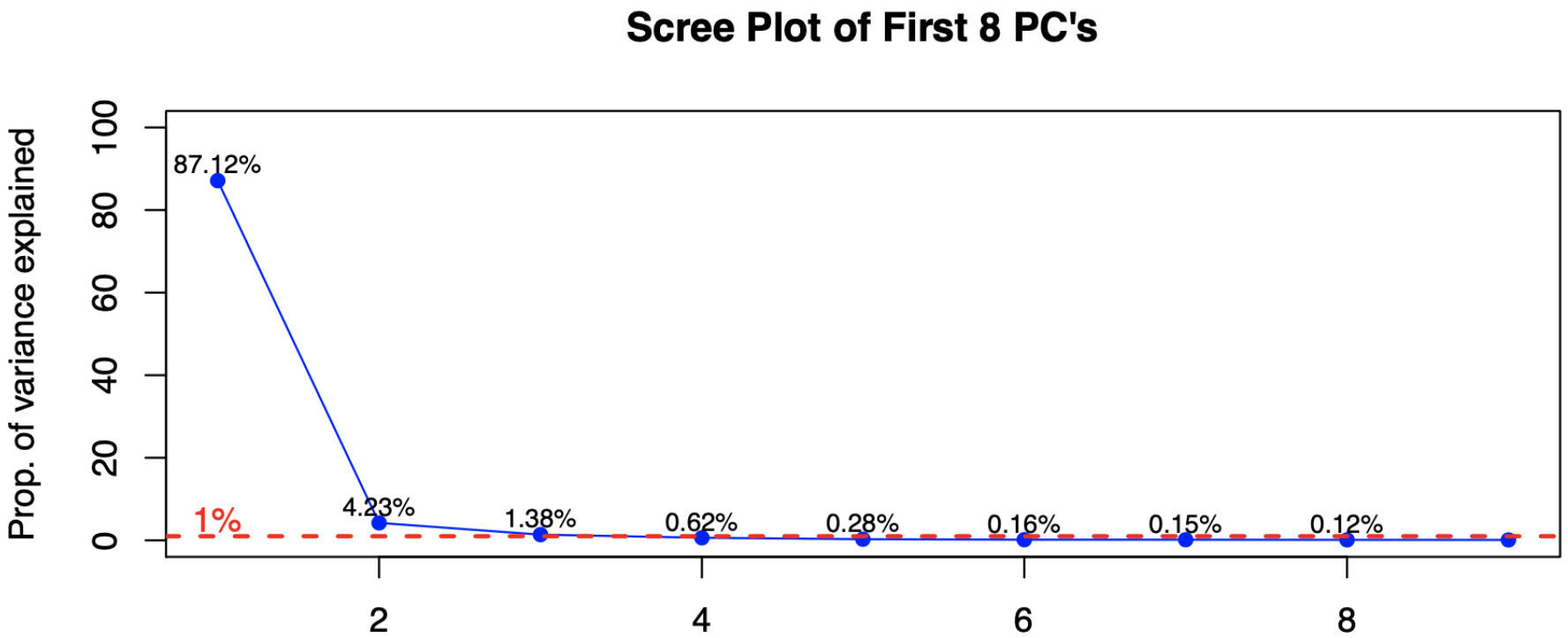


Princial Component Analysis

- Number of PC
  - Scree plot
  - Percentage of Variance Explained
  - Prior information
  - Other models (e.g. Error of PC regression)

Princial Component Analysis

- Number of PC
  - **Scree plot**
  - Percentage of Variance Explained
  - Prior information
  - Other models (e.g. Error of PC regression)



Principal Component Analysis

Comparison with LDA

	LDA	PCA
	Supervised Learning	Unsupervised Learning
y required?	yes	no
Normalization?	not necessary	strongly recommended
Rule for rotation	seperate the classes	capture variance

Table of Contents

Clustering

Clustering

- Core idea
  - The observations in the same **cluster** are more simlar
- Models
  - Connectivity-based model
    - **Hierarchy Clustering**
  - Centroid model
    - **K-means Clustering**
  - Distribution model
  - Graph-based model
  - Others
- Data-preprocessing procedure:
  - Normalization is recommended

$$\tilde{X} = \frac{X - \hat{\mu}_X}{\hat{\sigma}_X}$$

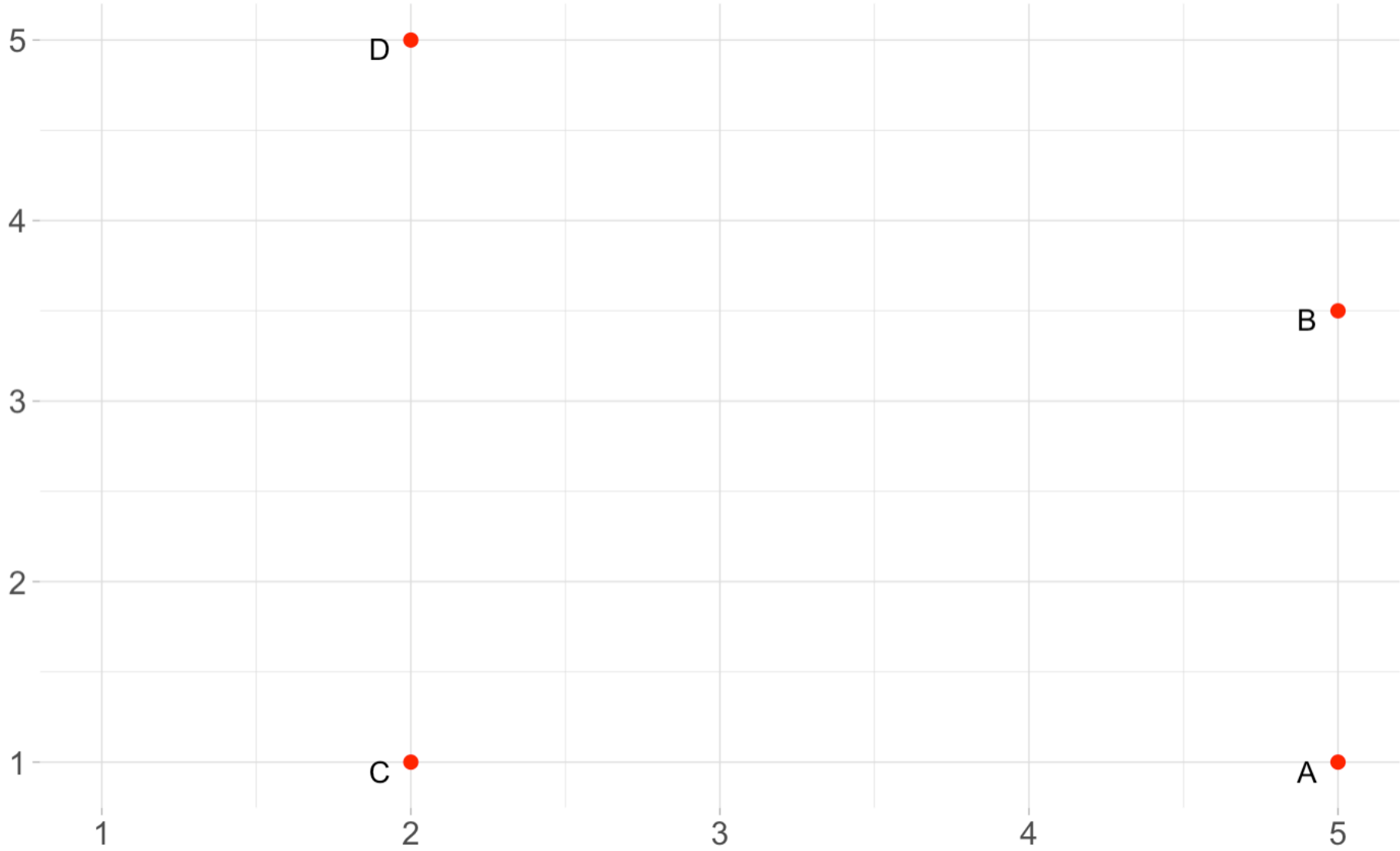
Clustering

Hierarchy Clustering

- Core idea
  - Observations being more related to nearby observations than to observations farther away
- Procedures
  - Measure the distance among observations
  - Build cluster dendrogram
  - Determine number of cluster and get prediction

Clustering

Hierarchy Clustering



Step 1

Measure the distance among observations

$$d(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

	A	B	C	D
A		2.5	3	5
B	2.5		3.9	3.4
C	3	3.9		4
D	4	3.4	4	

Step 2

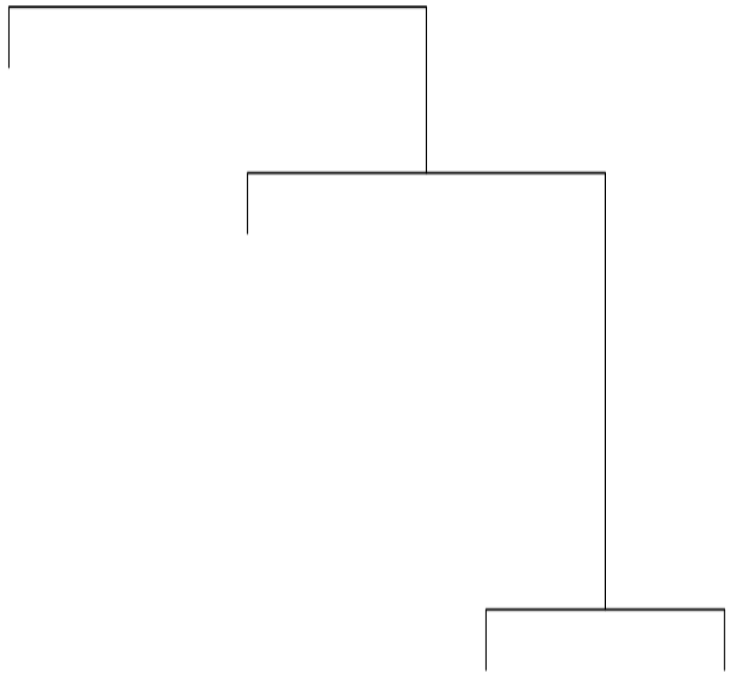
Build cluster dendrogram

	A	B	C	D
$\Rightarrow$ A		2.5	3	5
$\Rightarrow$ B	2.5		3.9	3.4
$\Rightarrow$ C	3	3.9		4
$\Rightarrow$ D	5	3.4	4	

Step 2

Build cluster dendrogram

	A	B	C	D
$\Rightarrow$ A		2.5	3	5
$\Rightarrow$ B	2.5		3.9	3.4
$\Rightarrow$ C	3	3.9		4
$\Rightarrow$ D	5	3.4	4	

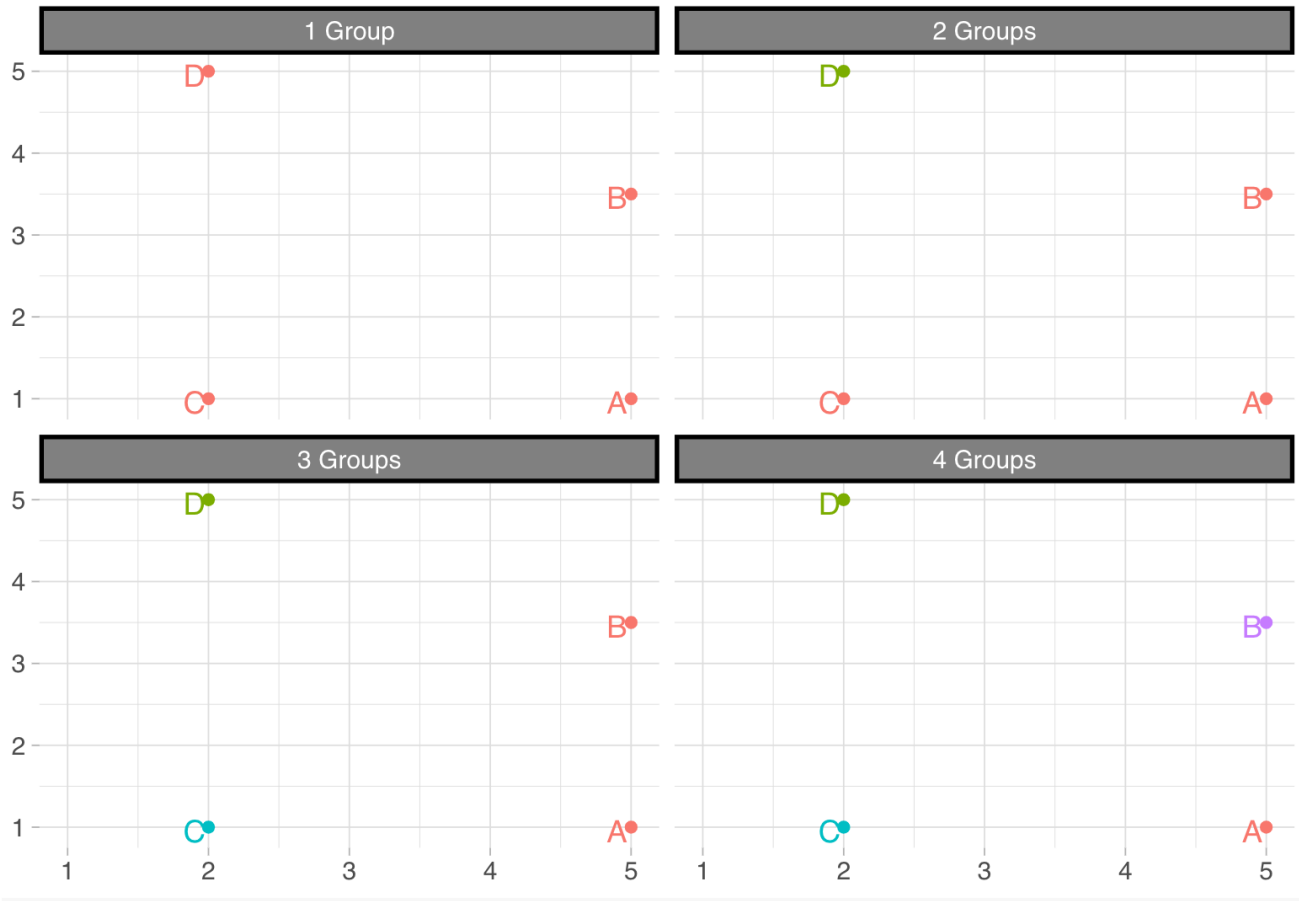


Clustering

Hierarchy Clustering

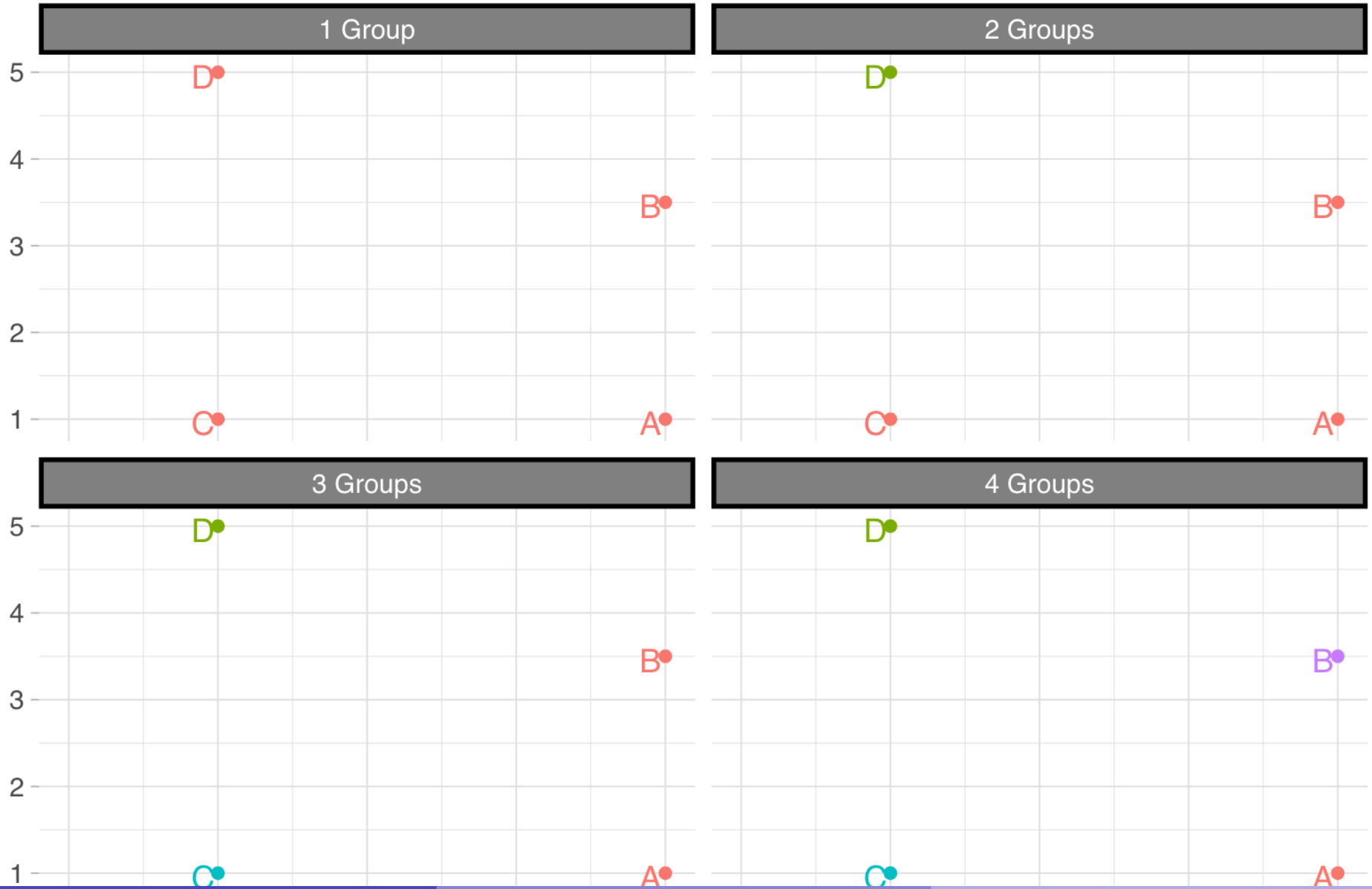
Step 3

Determine number of cluster and get prediction



Clustering

Hierarchy Clustering



Clustering

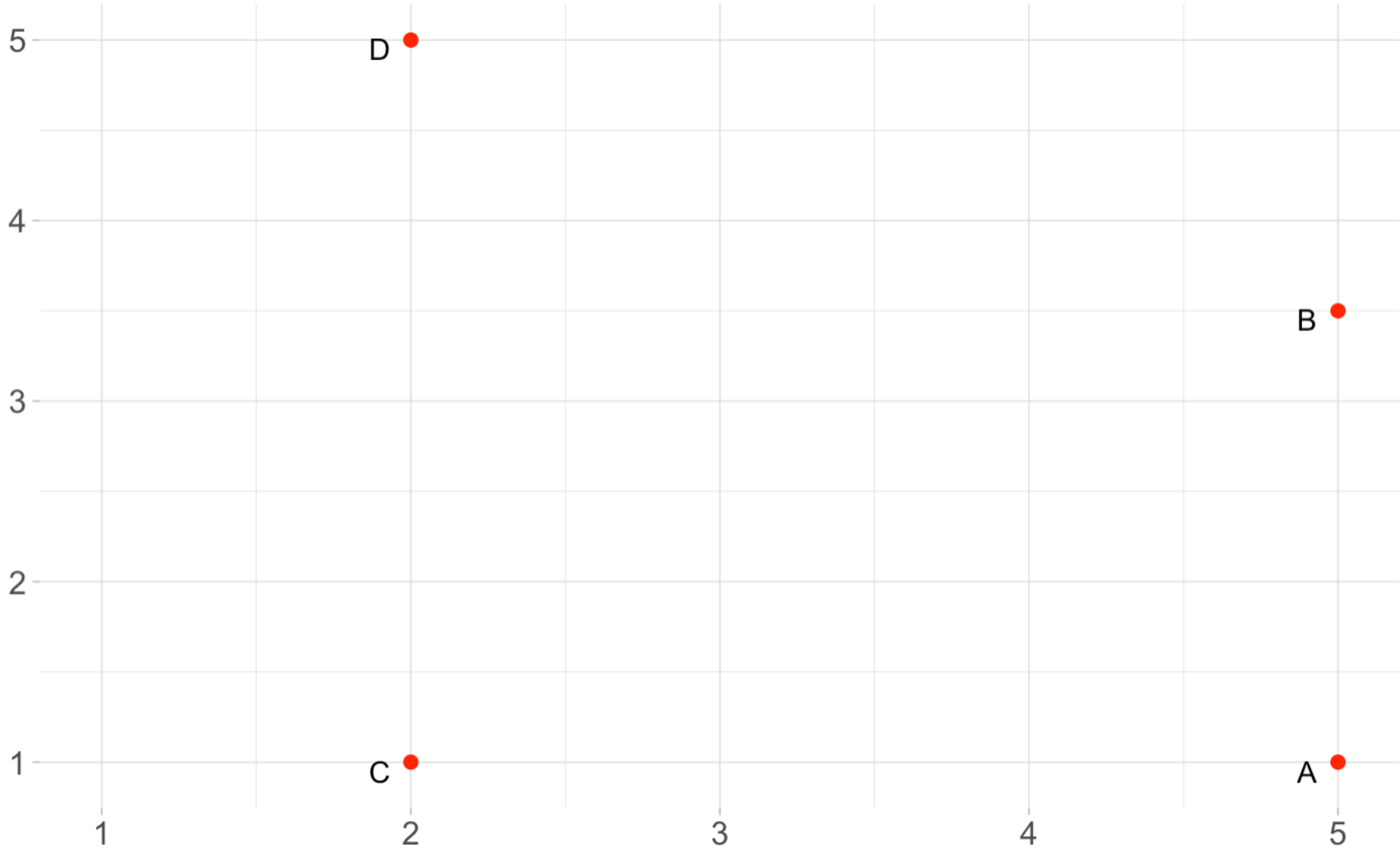
K-means Clustering

- Core idea
  - Each observation belongs to the cluster with the nearest cluster centroid
- Procedures

Step 1 Determine number of clusters K
Step 2 Select cluster centroids
Step 3 Cluster observations
Step 4 Use mean as centroid for each cluster
Step 5 If centroids are almost the same, stop.
If not, go back to step 2.

Clustering

K-means Clustering



Clustering

K-means Clustering

Step 1

Determine number of clusters $K = 2$

Step 2

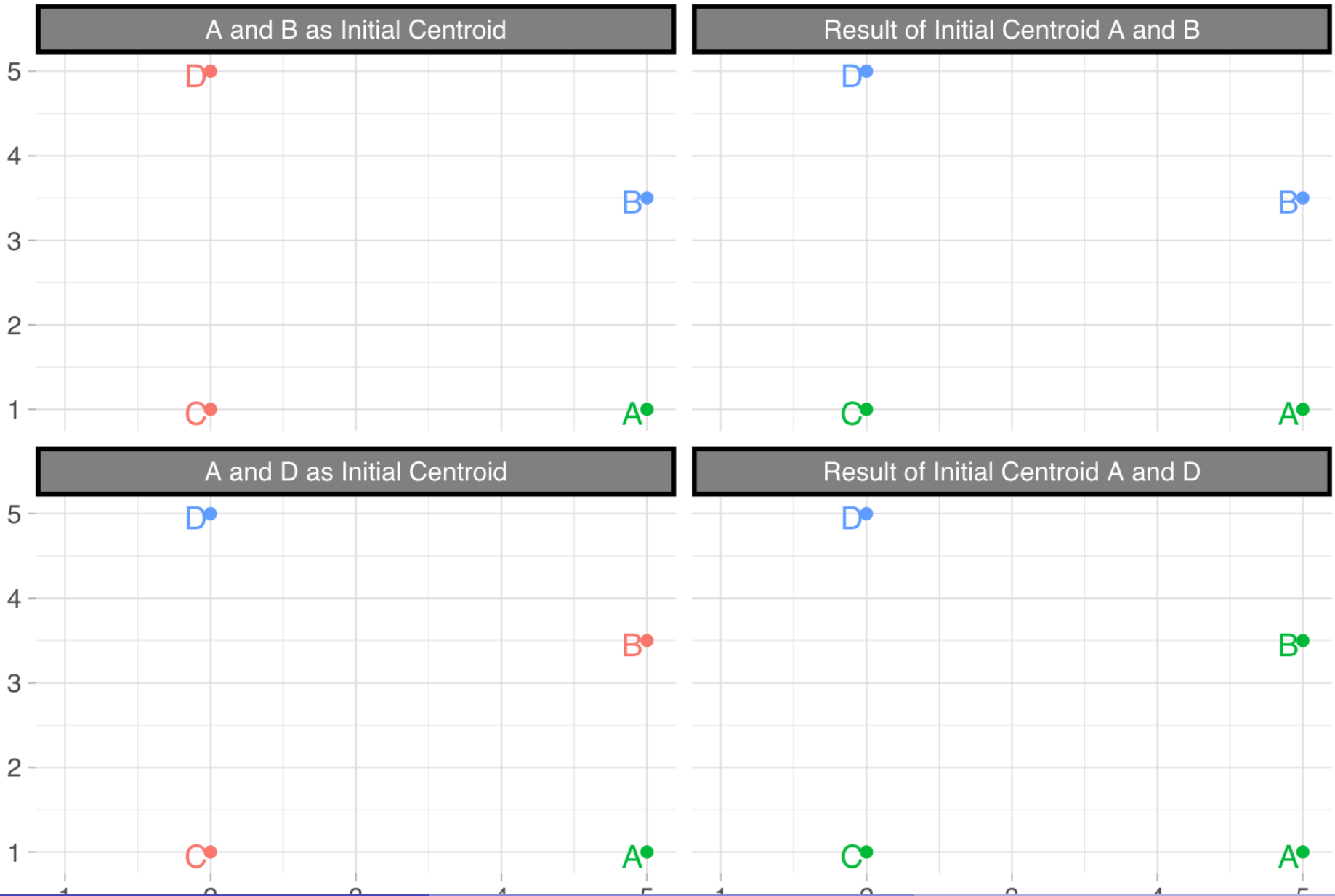
Select cluster centroids

i. A and B

ii. A and D

Clustering

K-means Clustering



Summary

	Class	Model
Supervised Learning	Regression	Linear Regression Lasso Regression Ridge Regression Possion Regression
	Classification	Logistic Regression Linear Discriminant Analysis Quadratic Discriminant Analysis Support Vector Machines
Unsupervised Learning		Principal Component Analysis K-means Clustering Hierarchy Clustering