Cluster Analysis – Beyond Basics

Peerapon S.

Machine Learning (1/67)

Topics

Gaussian mixture model

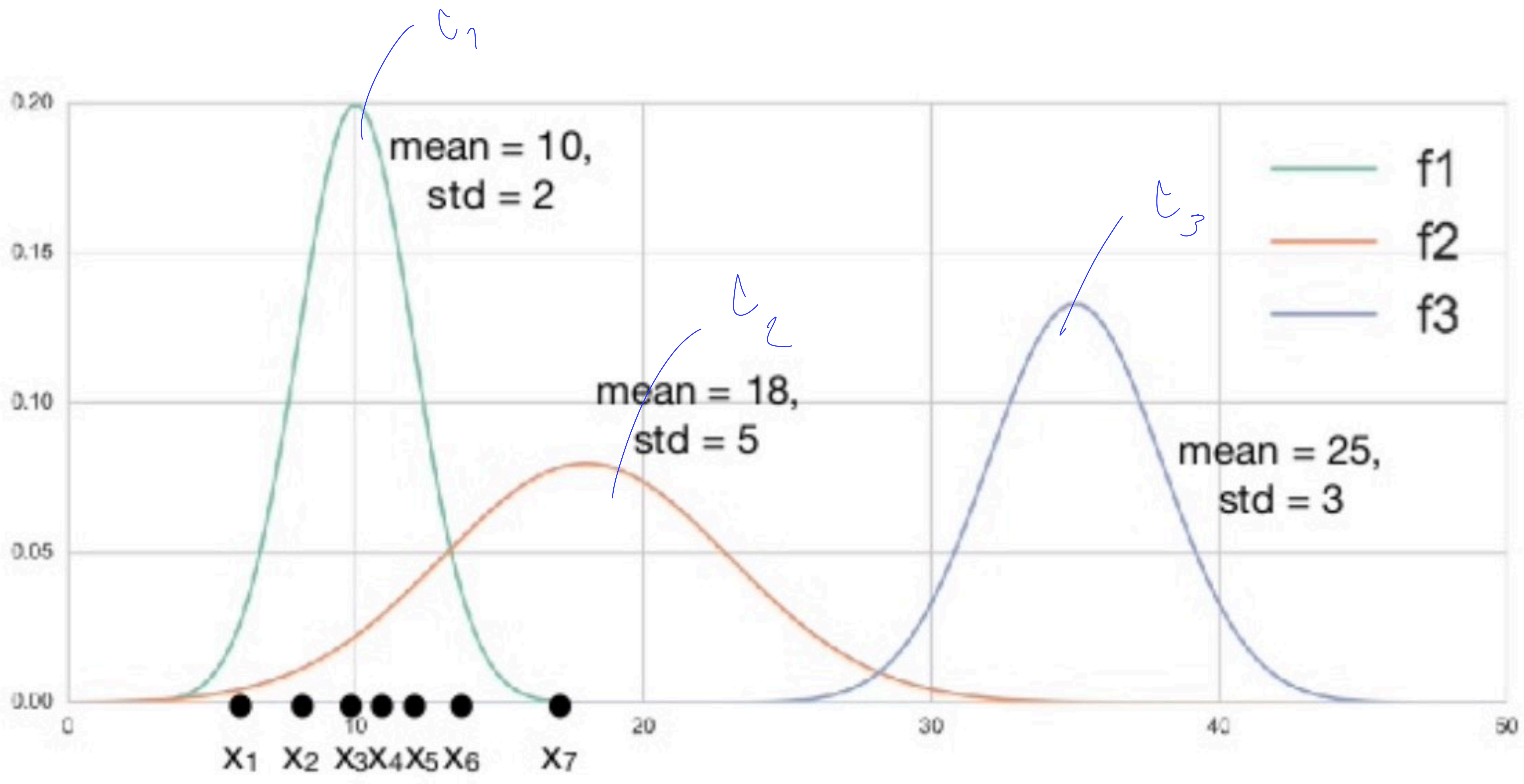
Hierarchical clustering (Dendrogram)

Density-based clustering

Gaussian Distribution

Consider a univariate Gaussian distribution with parameters $\boldsymbol{\theta} = (\mu, \sigma)$

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Maximum Likelihood Estimation (MLE)

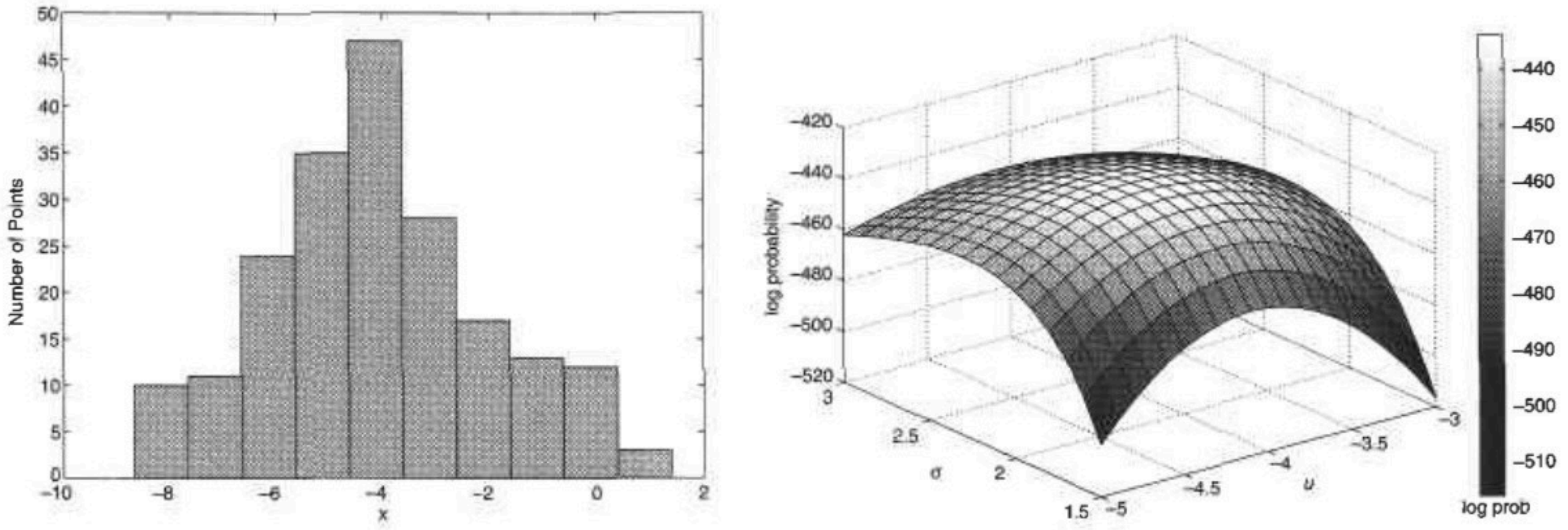
Given a set of 1D data $\mathbb{X} = \{x_1, x_2, \ldots, x_n\}$, define the ***likelihood function*** and ***log-likelihood function*** as

$$L(\theta|\mathbb{X}) = \prod_{i=1}^{n} f(x_i, \theta), \quad \theta = (\mu, \sigma)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\log L(\theta|\mathbb{X}) = -\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} - 0.5n\log(2\pi) - n\log(\sigma)$$

Maximizing the log-likelihood function gives the parameters (μ, σ) best matched the data.



(a) Histogram of 200 points from a Gaussian distribution.

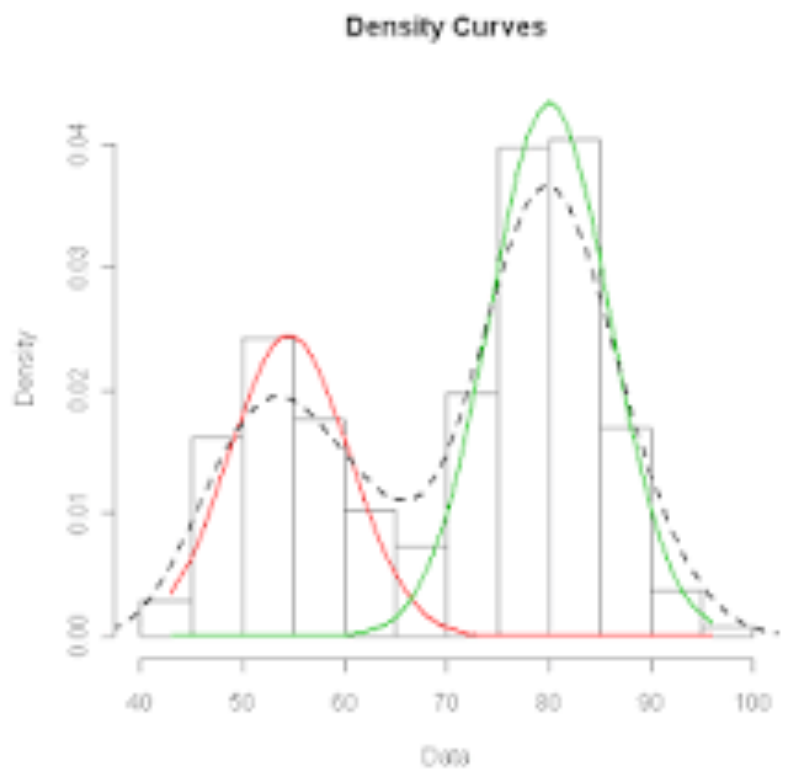
(b) Log likelihood plot of the 200 points for different values of the mean and standard deviation.

Gaussian Mixture Model (GMM) in 1D

Data as a set of observations from a weighted sum of K gaussian distributions $f_k(x;\theta)$.

$$f(x) = w_1 f_1(x) + w_2 f_2(x) + \cdots + w_K f_K(x), \quad f_i(x) \sim \text{Normal}(\mu_i, \sigma_i)$$

$$f_i(x|\boldsymbol{\theta}_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}, \quad \boldsymbol{\theta}_i = (\mu_i, \sigma_i)$$



Each distribution describes a different group/cluster.

Each data point is probabilistically associated with all clusters in the clustering.

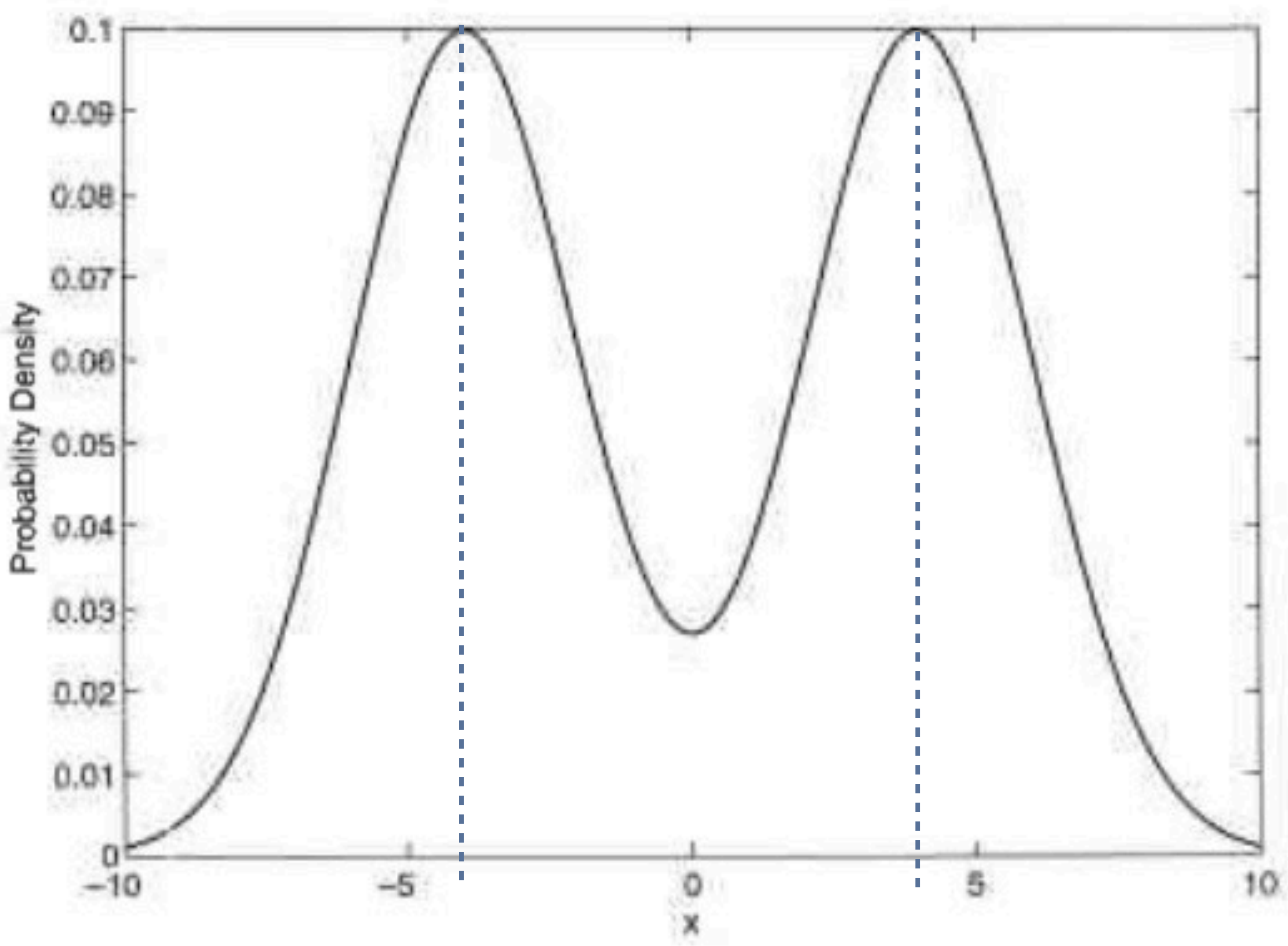
Consider a set of n observations $\mathbb{X} = \{x_1, x_2, \ldots, x_n\}$ and two gaussian distributions with parameter set $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$.

Then, for $w_1 + w_2 = 1$, and $0 < w_k < 1$, the probability distribution of x from the mixture model is

$$f(x|\Theta) = w_1 f(x|\theta_1) + w_2 f(x|\theta_2)$$

Ex: $w_1 = w_2 = 0.5$, $\boldsymbol{\theta}_1 = (\mu_1, \sigma_1) = (-4{,}2)$, $\boldsymbol{\theta}_2 = (\mu_2, \sigma_2) = (4{,}2)$,

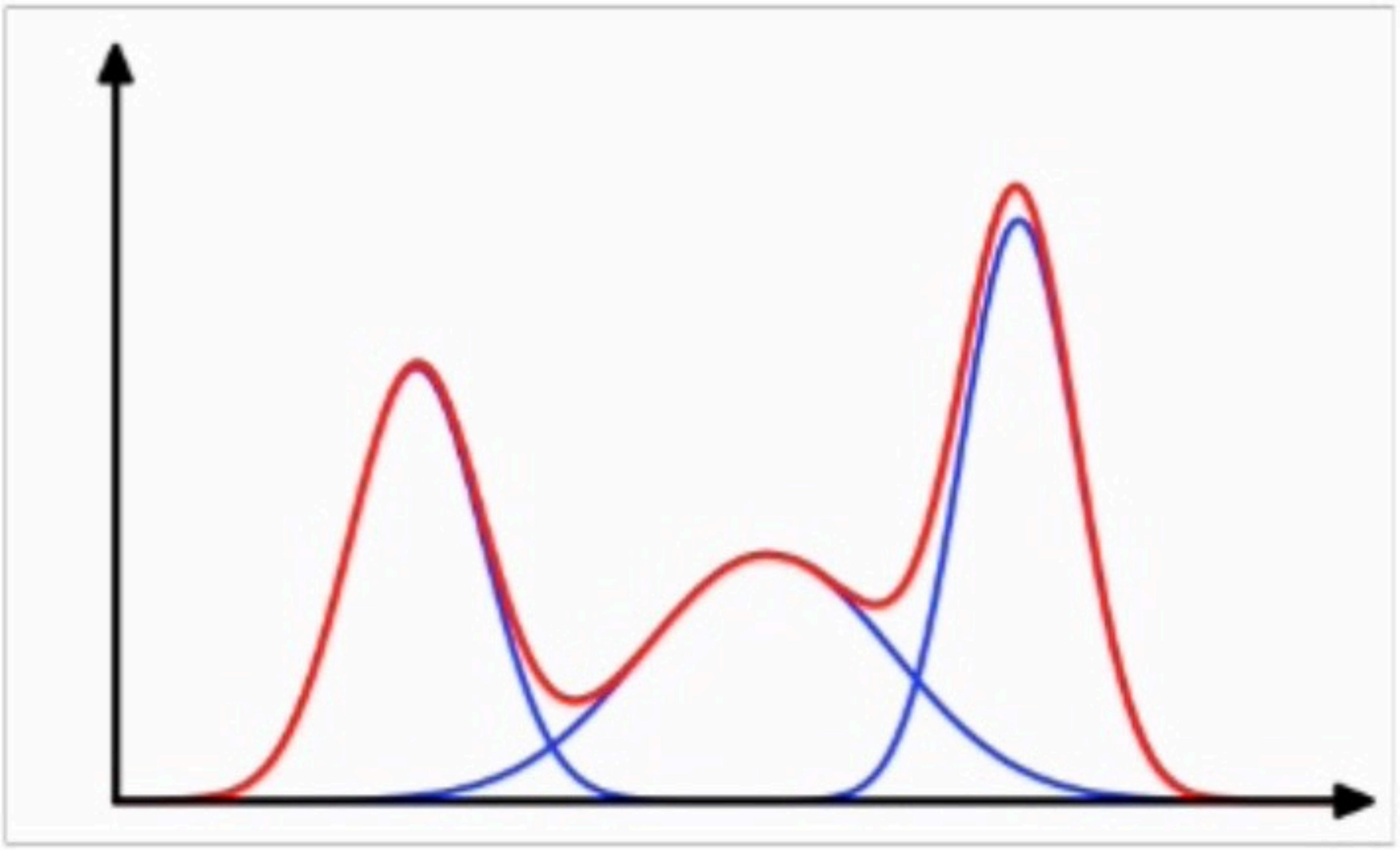
$$f(x|\Theta) = 0.5\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x+4)^2}{8}} + 0.5\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-4)^2}{8}}$$



For $K = 3$ clusters,

$$f(x|\Theta) = w_1 f_1(x|\boldsymbol{\theta_1}) + w_2 f_2(x|\boldsymbol{\theta_2}) + w_3 f_3(x|\boldsymbol{\theta_3})$$

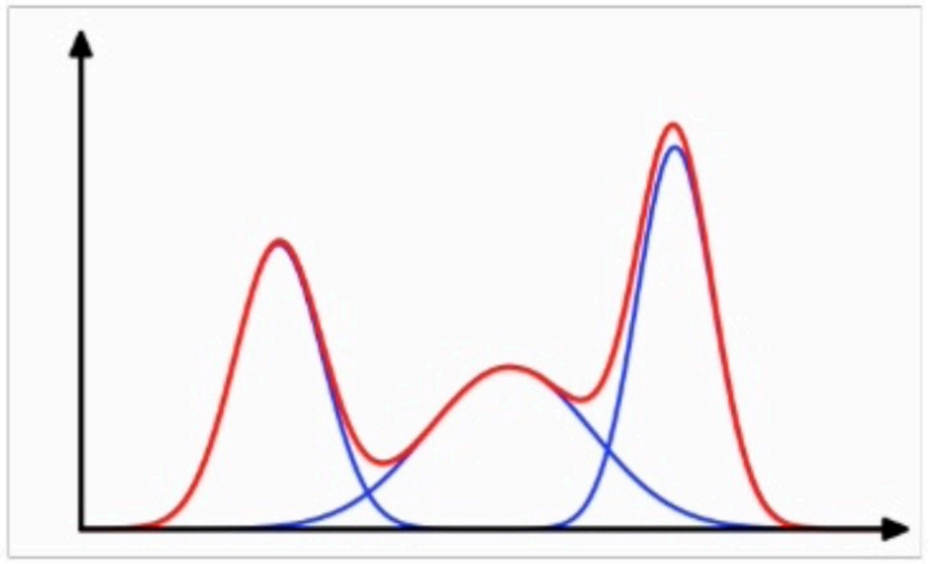
$$f_i(x|\boldsymbol{\theta_i}) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}$$



Estimating Model Parameters in 1D Mixture Model

Given a data set $\mathbb{X} = \{x_1, x_2, \ldots, x_n\}$, the likelihood function from a mixture of $K = 2$ gaussian distributions is

$$
\begin{aligned}
L(\boldsymbol{\Theta}, \boldsymbol{w} | \mathbb{X}) &= \prod_{i=1}^{n} f(x_i | \boldsymbol{\Theta}) \\
&= \prod_{i=1}^{n} \sum_{k=1}^{K} \overset{\text{gaussian dist.}}{\underline{w_k f(x_i; \theta_k)}} \\
&= \prod_{i=1}^{n} \left[w_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + w_2 \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}} \right]
\end{aligned}
$$



x_i

No closed-form solution for MLE unless w_i's are known.

The model fitting maximizes the log likelihood function (LLF) by iteratively updating the model parameters.

Expectation Maximization (EM) Algorithm in 1D

Define a discrete variable $Z \in \{1,2,\ldots,K\}$ where $w_k \triangleq P\{Z=k\}$ or $P\{Z_k\}$ is the probability that a randomly selected data point comes from distribution/cluster k.

Given $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\} = \{(\mu_1,\sigma_1),(\mu_2,\sigma_2)\}$ of two distributions in the mixture with weights w_1, w_2, each data point x_i belongs to cluster k with probability

$$
\begin{aligned}
w_{ki} = \mathbb{P}\{Z=k|x_i\} &= \frac{\mathbb{P}\{Z_k, x_i\}}{\mathbb{P}\{x_i\}}, \quad k=1,2 \\
&= \frac{\mathbb{P}\{x_i|Z_k\}\mathbb{P}\{Z_k\}}{\sum_j \mathbb{P}\{x_i|Z_j\}\mathbb{P}\{Z_j\}}, \quad \text{(Bayes' theorem)} \\
&\approx \frac{w_k f(x_i|\theta_k)}{\sum_j w_j f(x_i|\theta_j)}
\end{aligned}
$$

Ex: With initial values of $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ and $w_1 = w_2 = 0.5$, we have

$$
w_{ki} \approx \frac{0.5 f(x_i|\theta_k)}{0.5 f(x_i|\theta_1) + 0.5 f(x_i|\theta_2)}
$$

$$
\hat{w}_k \leftarrow (1/n)\sum_i w_{ki} \qquad \text{(Expectation Step)}
$$

1 cluster มี 6 parameters

$\boxed{\mu_{1i}, \mu_{2i}, \sigma_1 \sigma_2, \sigma_{1i}, \sigma_{2i}, \hat{w}_i}$

case: data มี 2มิติ

Plugging the estimated values of w_1, w_2 into the likelihood function:

unknown

constant

constant

$$L(\Theta \mid \hat{\boldsymbol{w}}, \mathbb{X}) = \prod_{i=1}^{n} \left[\hat{w}_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}} + \hat{w}_2 \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_i - \mu_2)^2}{2\sigma_2^2}} \right]$$

It turns out that $L(\Theta \mid \hat{\boldsymbol{w}}, \mathbb{X})$ is maximized at

$$\left.\begin{aligned} \hat{\mu}_k &= \frac{1}{n} \sum_i \frac{w_{ki}}{\hat{w}_k} x_i \\ \hat{\sigma}_k^2 &= \frac{1}{n} \sum_i \frac{w_{ki}}{\hat{w}_k} (x_i - \hat{\mu}_k)^2 \end{aligned}\right\} \quad \text{(Maximization Step)}$$

Substituting the updated parameters $\hat{\mu}_k$, $\hat{\sigma}_k^2$ to $f(x \mid \boldsymbol{\theta}_i)$, recompute $\hat{w}_{ki}$, and repeat until $\hat{\mu}_k$ converges.

How to determine the cluster membership for a given data point ?

ใช้ threshold

Algorithm (Univariate) Gaussian Mixture EM Algorithm

Input: n data points $\{x_1, x_2, \dots, x_n\}$ and the number of clusters K

Output: K model parameters $\{(\mu_k, \sigma_k)\}$, $1 \leq k \leq K$

Initialization

Initialize $\hat{\mu}_k$

$\hat{\sigma}_k^2 \leftarrow \frac{1}{n} \sum_i (x_i - \bar{x})^2$

$w_k \leftarrow 1/K$

repeat

    Expectation Step

    Compute for all data points x_i, $w_{ki} \leftarrow \frac{w_k f(\boldsymbol{x}_i | \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} w_j f(\boldsymbol{x}_i | \boldsymbol{\theta}_j)}$, $\forall k$

    $w_k \leftarrow (1/n) \sum_i w_{ki}$, $\forall k$

    → w ของทุก data points

    Maximization Step

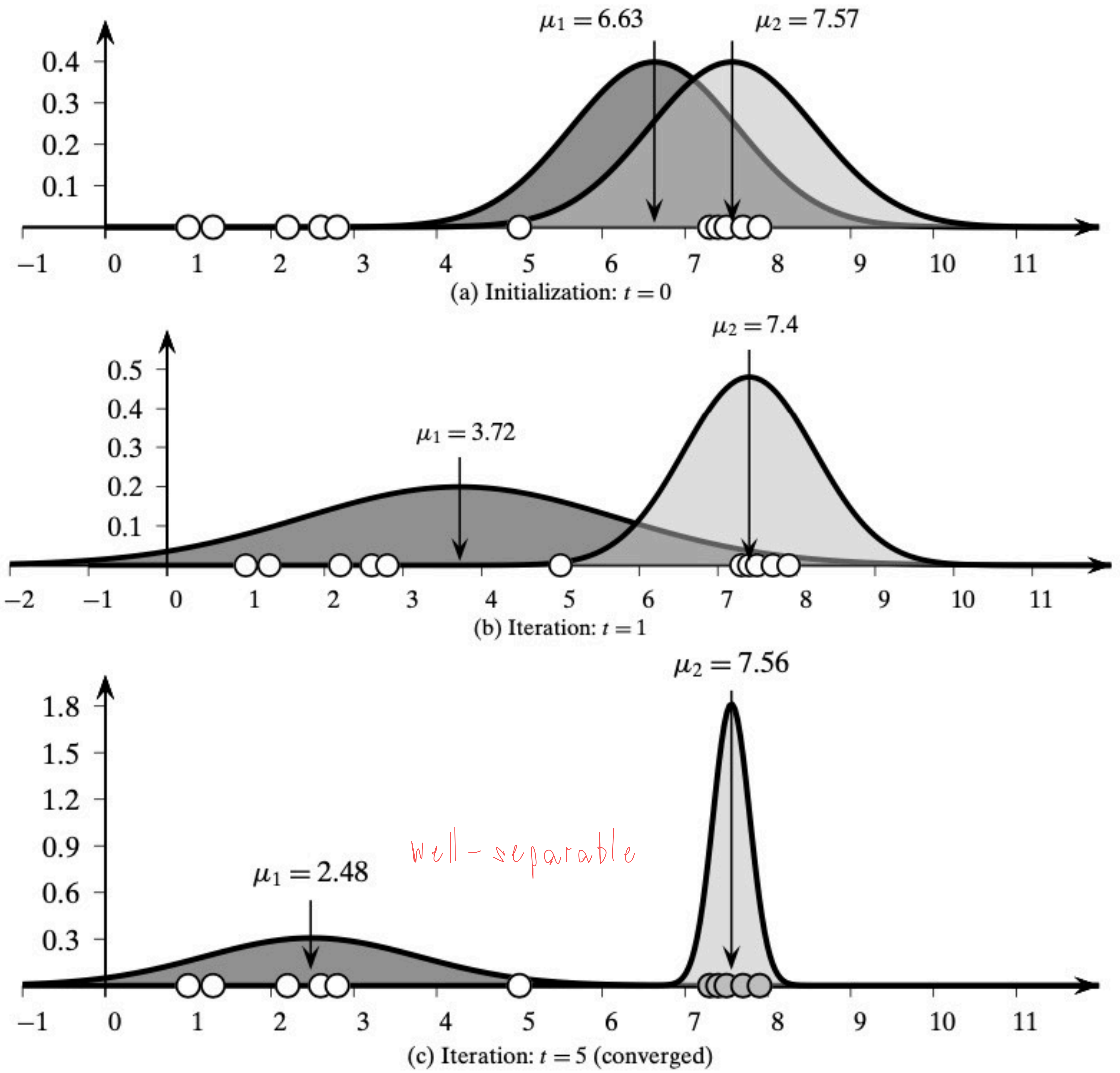
        $\hat{\mu}_k \leftarrow \frac{1}{n} \sum_i \frac{w_{ki}}{w_k} x_i$

        $\hat{\sigma}_k^2 \leftarrow \frac{1}{n} \sum_i \frac{w_{ki}}{w_k} (x_i - \hat{\mu}_k)^2$

    maximize likelihood

until $\hat{\mu}_k$ converges

return



$\mu_1 = 6.63$
$\mu_2 = 7.57$
(a) Initialization: $t = 0$
$\mu_2 = 7.4$
$\mu_1 = 3.72$
(b) Iteration: $t = 1$
$\mu_2 = 7.56$
well-separable
$\mu_1 = 2.48$
(c) Iteration: $t = 5$ (converged)

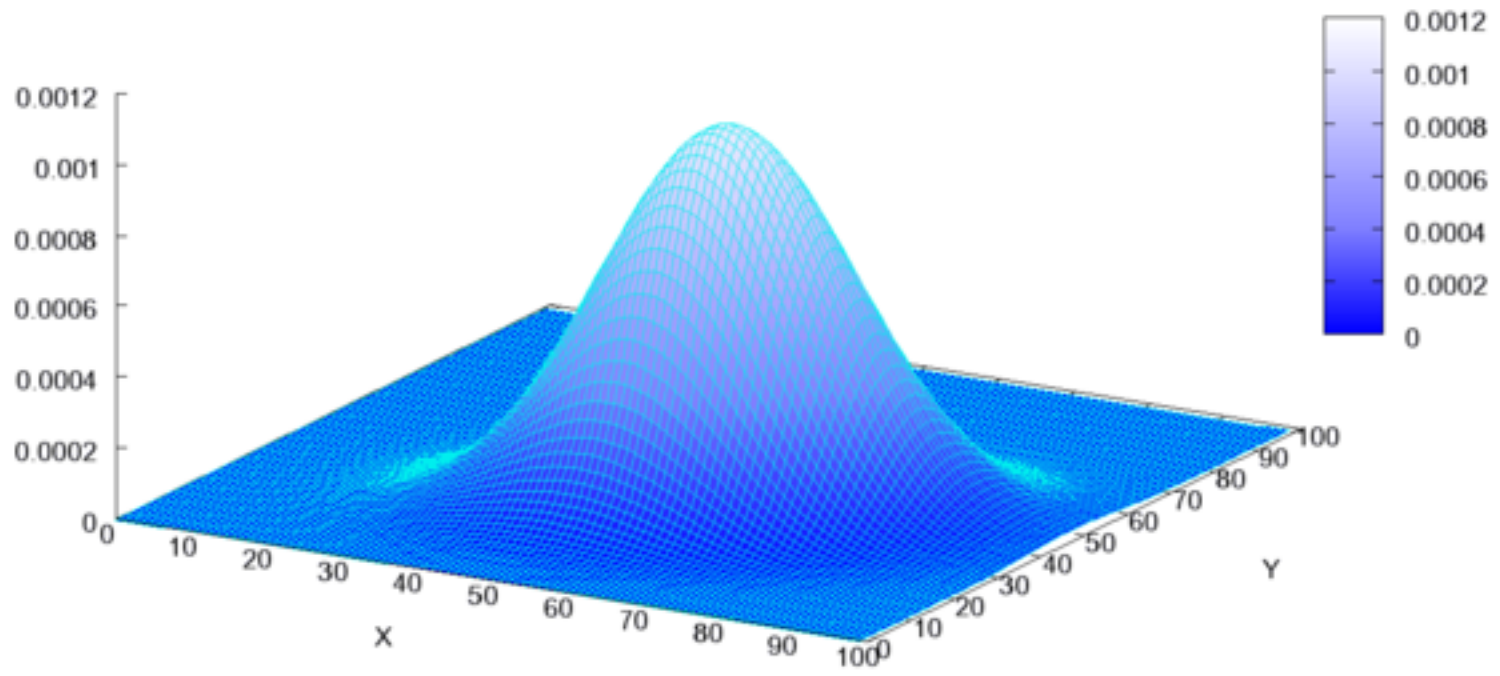
$\dim = 2;\ f(x_1, x_2 \mid \theta) = \dfrac{\exp\left(-\frac{1}{2}(x_1 - \mu_1)^T \cdot \Sigma^{-1}(x_1 - \mu_1) - \frac{1}{2}(x_2 - \mu_2)^T \cdot \Sigma^{-1}(x_2 - \mu_2)\right)}{\sqrt{(2\pi)^2 |\Sigma|}}$

data

หา det

$\Sigma = \text{cov\_mat} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$

Multivariate Normal Distribution



columns/variables

$$X = (X_1, X_2, \ldots, X_d)$$

covariance

$$\boxed{f(\boldsymbol{x}|\theta) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d |\Sigma|}}}, \quad \boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \cdots & \sigma_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \sigma_{d3} & \sigma_d^2 \end{bmatrix}$$

Parameters : $\theta = (\boldsymbol{\mu}, \Sigma)$ ของแต่ละ distribution ใน mixture

correlation

$$\text{For } d = 2: \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho_{12}^2)$$

$$f(x_1, x_2|\theta) = \frac{\exp\left\{-\frac{1}{2(1-\rho)^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

22:21

forms.office.com

* Required

1

Consider the likelihood function of a mixture of three bivariate Gaussian distributions below. For a dataset with n rows, how many model parameters to estimate in this mixture model ? * (1 Point)

cluster

$$L(\Theta, \boldsymbol{w}|\mathbb{X}) = \prod_{i=1}^{n} \boxed{\sum_{k=1}^{3}} w_k f(\boldsymbol{x}_i; \theta_k)$$

$$f(\boldsymbol{x}; \theta_k) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^2 |\Sigma|}}, \quad \theta_k = (\mu_k, \Sigma_k)$$

- ( ) 3
- ( ) 6
- ( ) 9
- ( ) 12
- (•) 18
- ( ) 6n
- ( ) 9n
- ( ) No idea

assume 2D-data

$$|\Sigma| = \sigma_1^2 \sigma_2^2 (1 - \rho_{12}^2)$$

$$f(x_1, x_2 | \theta) = \frac{\exp\left\{-\frac{1}{2(1-\rho)^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

2 มิติ

ใน 1 cluster เกิด unknown param. 6 ตัว

∴ 3 cluster = 3×6 = 18 ตัว

Assume a set of n observations $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^d$ and K multivariate normal distributions with parameter set $\Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K)$,

Then, for $\sum_i w_k = 1$, and $0 < w_k < 1$, the probability distribution of $\mathbf{x}$ given the mixture model is

$$f(\boldsymbol{x}|\Theta) = \sum_{k=1}^{K} w_k f(\boldsymbol{x}|\boldsymbol{\theta}_k)$$

Given a data set $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the likelihood function is

$$L(\Theta, \boldsymbol{w}|\mathbb{X}) = \prod_{i=1}^{n} \underbrace{f(\boldsymbol{x}_i|\Theta)}^{\text{likelihood}}$$

$$= \prod_{i=1}^{n} \sum_{k=1}^{K} w_k f(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$$

self-test: if k=3 model will have 18 parameters

each cluster has $\mu_1, \mu_2, \sigma_1, \sigma_2, \sigma_1\sigma_2, w_1$

Algorithm (Multivariate) Gaussian Mixture EM Algorithm

Input: n data points $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{X} \in \mathbb{R}^{d \times n}$, and the number of clusters K

Output: K model parameters $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$, $1 \leq k \leq K$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$

Initialization

Initialize $\hat{\boldsymbol{\mu}}_k$ (e.g., random, K-means)

$\hat{\boldsymbol{\Sigma}}_k \leftarrow \frac{1}{n}(\boldsymbol{X} - \bar{\boldsymbol{x}})(\boldsymbol{X} - \bar{\boldsymbol{x}})^T$

$w_k \leftarrow 1/K$

$\boldsymbol{\theta}_k \leftarrow (\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$

repeat

    Expectation Step

    Compute for all data points x_i, $w_{ki} \leftarrow \frac{w_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{\sum_{j=1}^{K} w_j f(\boldsymbol{x}_i|\boldsymbol{\theta}_j)}$

    $w_k \leftarrow (1/n) \sum_i w_{ki}$

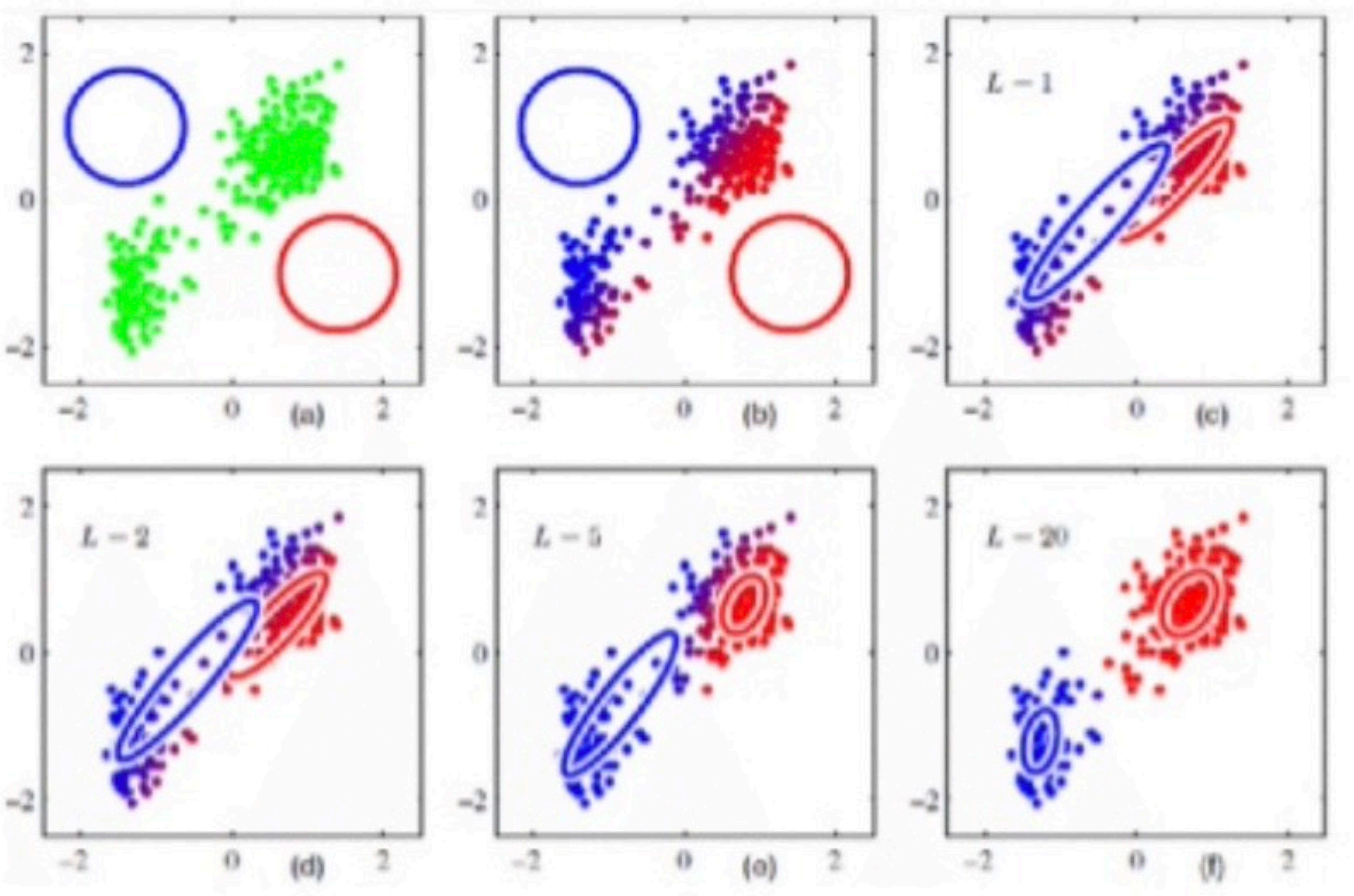
    Maximization Step

        $\hat{\boldsymbol{\mu}}_k \leftarrow \frac{1}{n} \sum_i \frac{w_{ki}}{w_k} \boldsymbol{x}_i$

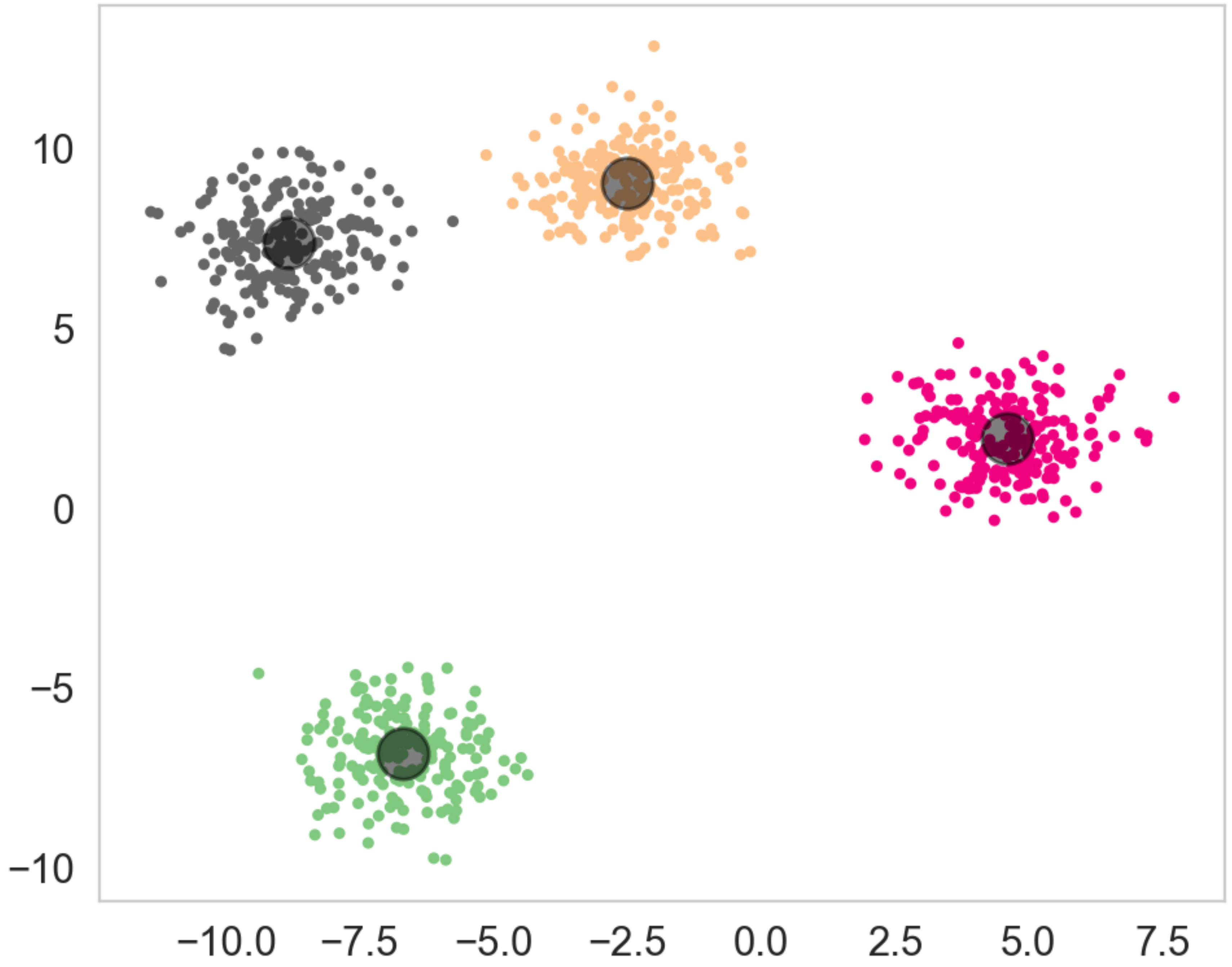
        $\hat{\boldsymbol{\Sigma}}_k \leftarrow \frac{1}{n} \frac{\sum_{i=1}^{n} w_{ki} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T}{w_k}$

until $\hat{\boldsymbol{\mu}}_k$ converges

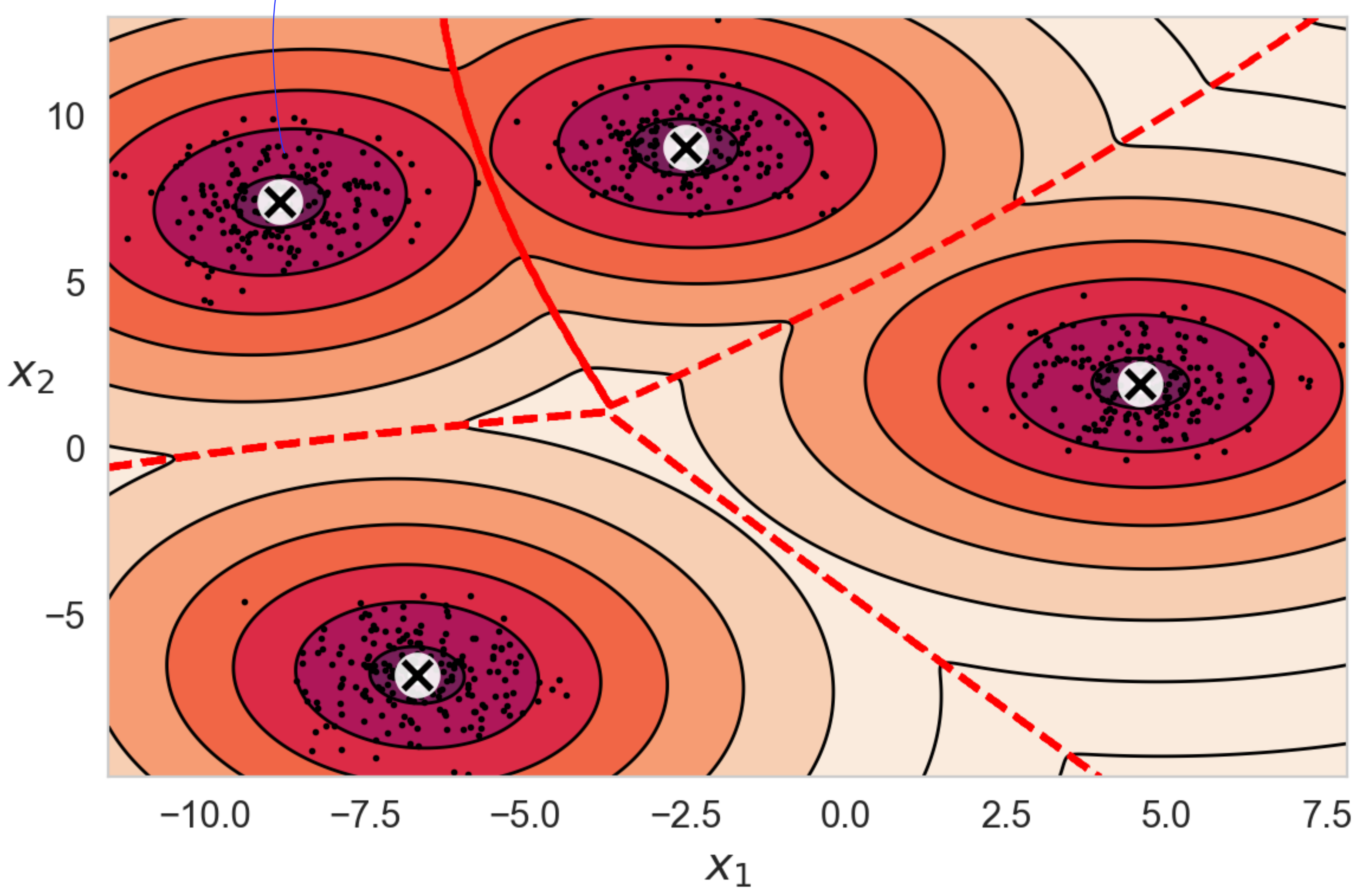
return



```python
from sklearn.mixture import GaussianMixture
gm = GaussianMixture(n_components=4, n_init=10, random_state=42)
gm.fit(X)
```



contour: shows density



Determining the number of clusters in GMM

Common criteria to assess how well the model fits the data are AIC and BIC:

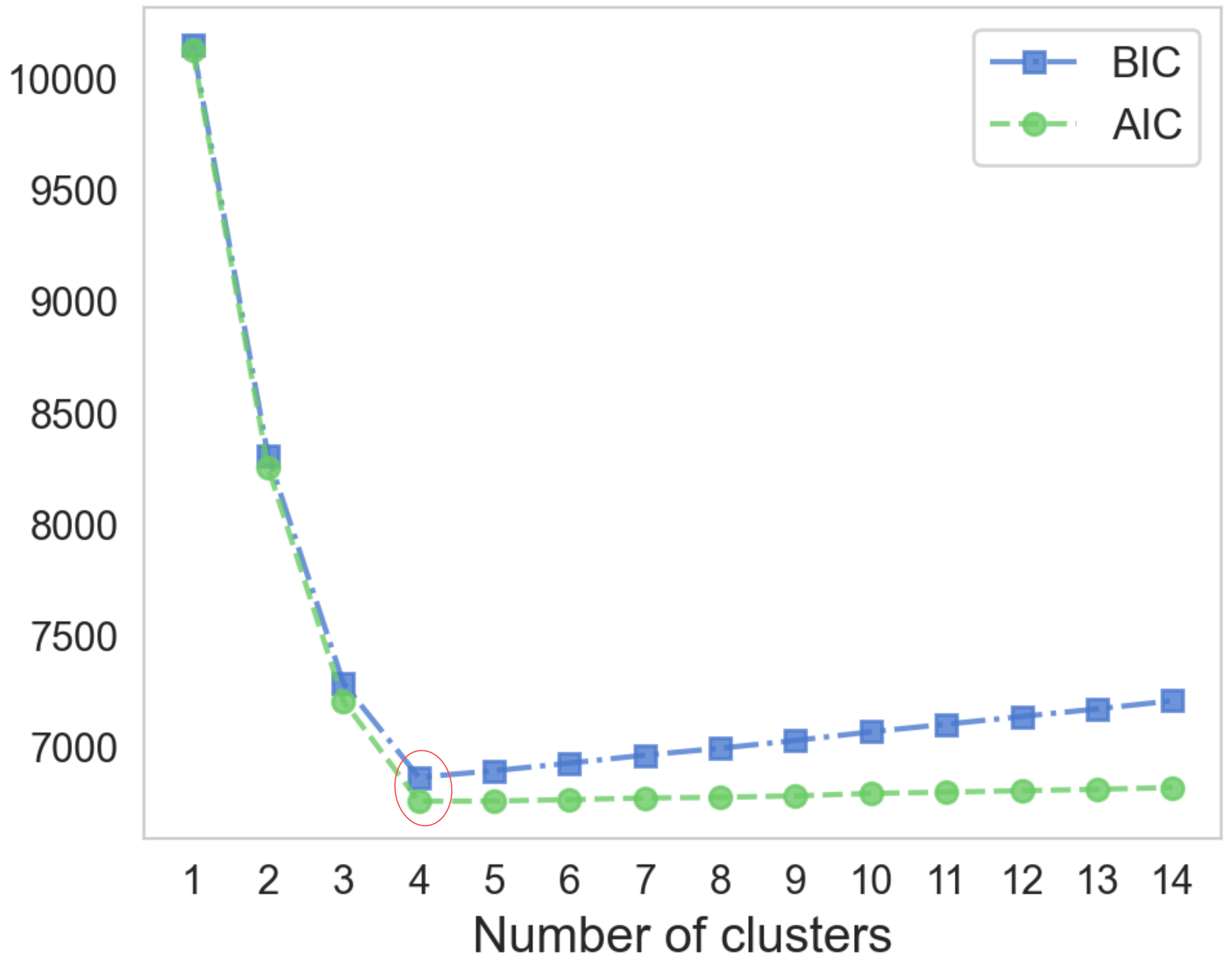
model parameters   log likelihood

$$\mathrm{AIC} = 2m - 2\log(L)$$

$$\mathrm{BIC} = m\log(N) - 2\log(L)$$

data points

- L: Maximum likelihood value
- m: Number of model parameters
- N: Number of data points



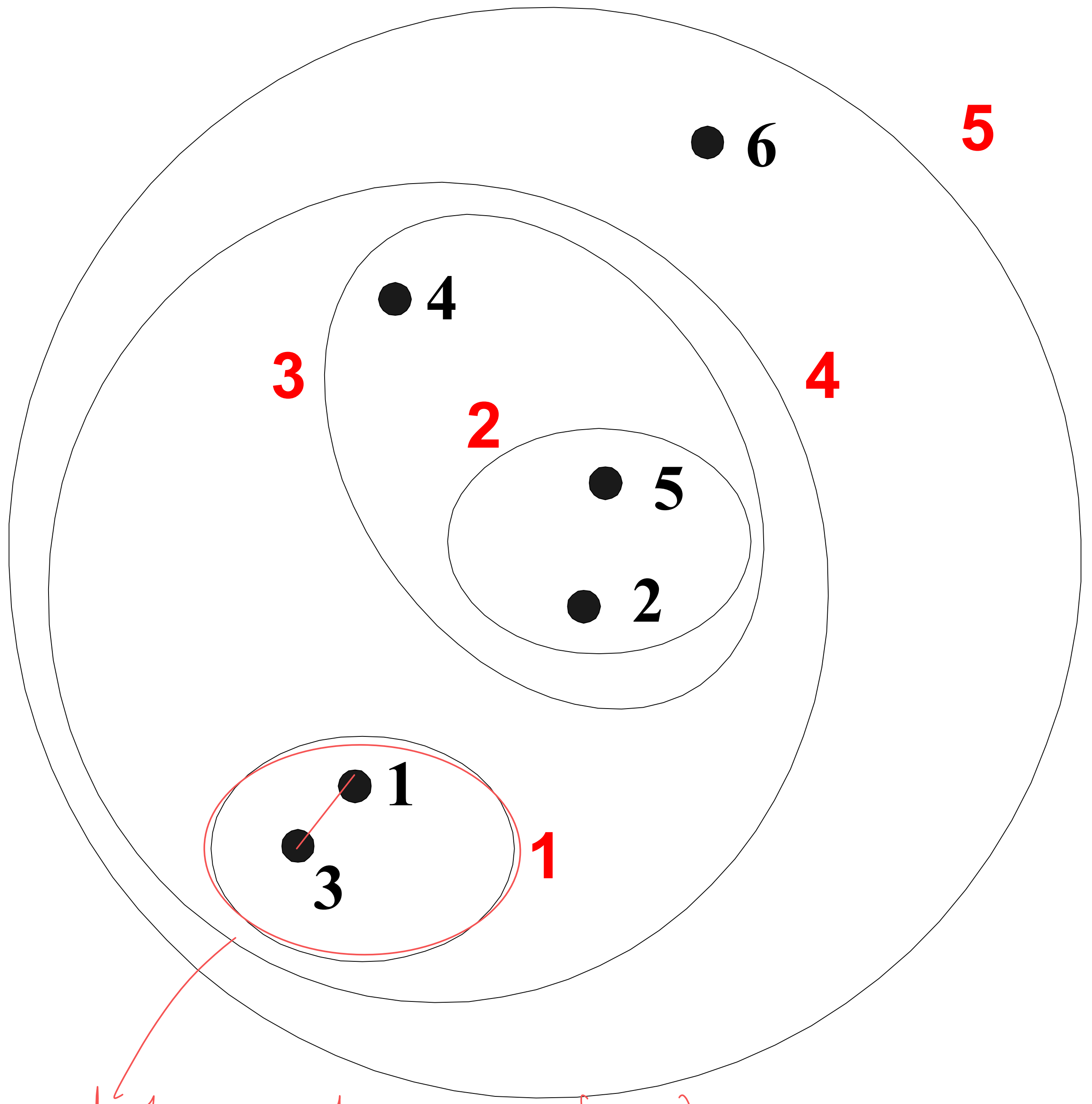
(Agglomerative) Hierarchical Clustering

Data point = Singleton cluster

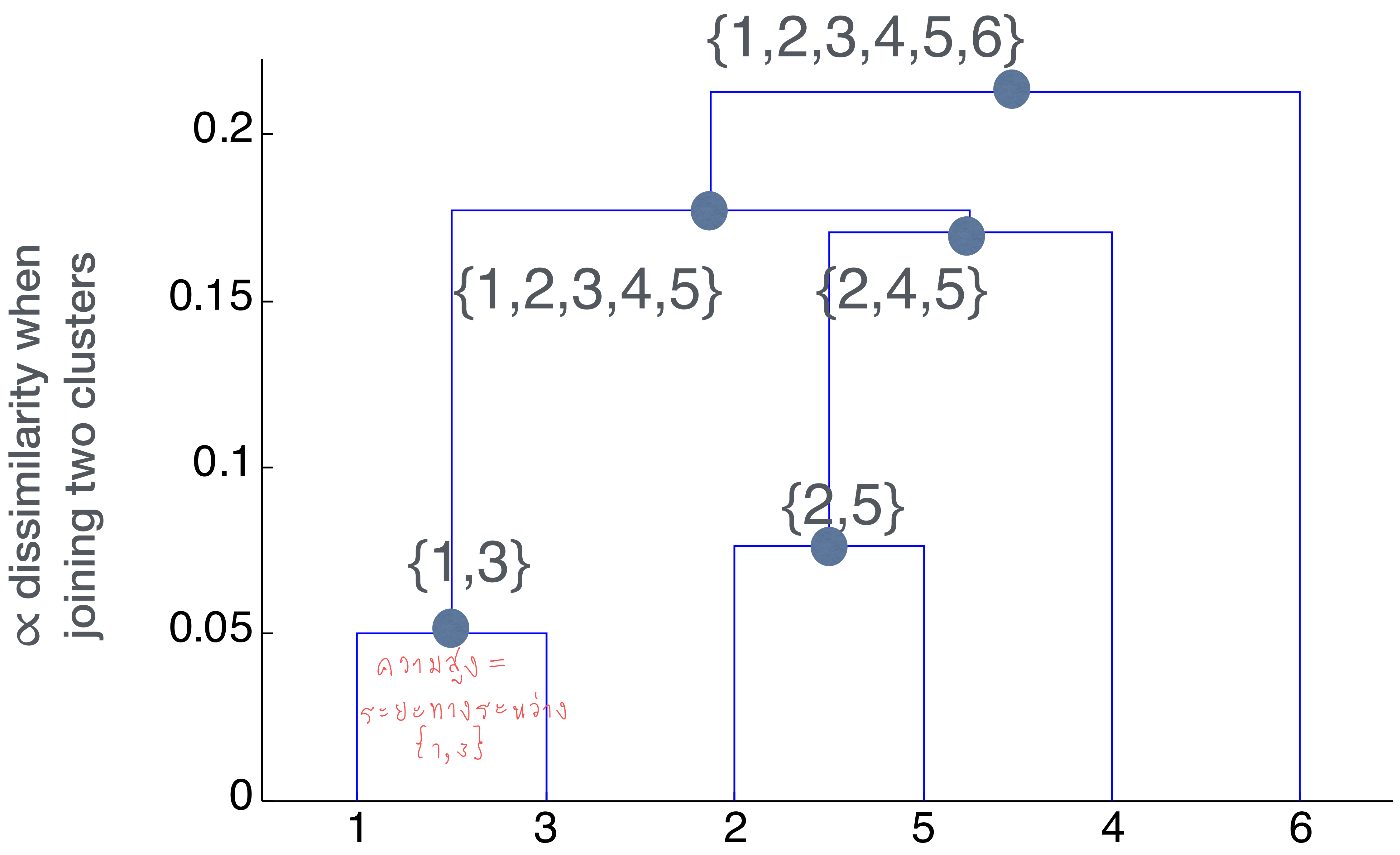
Iteratively merge two ***least dissimilar clusters*** until getting one cluster containing all points.

Dissimilarities of merged clusters monotonically increases with the level of the merger.

merge เสร็จ มองเป็น 1 data point แทน

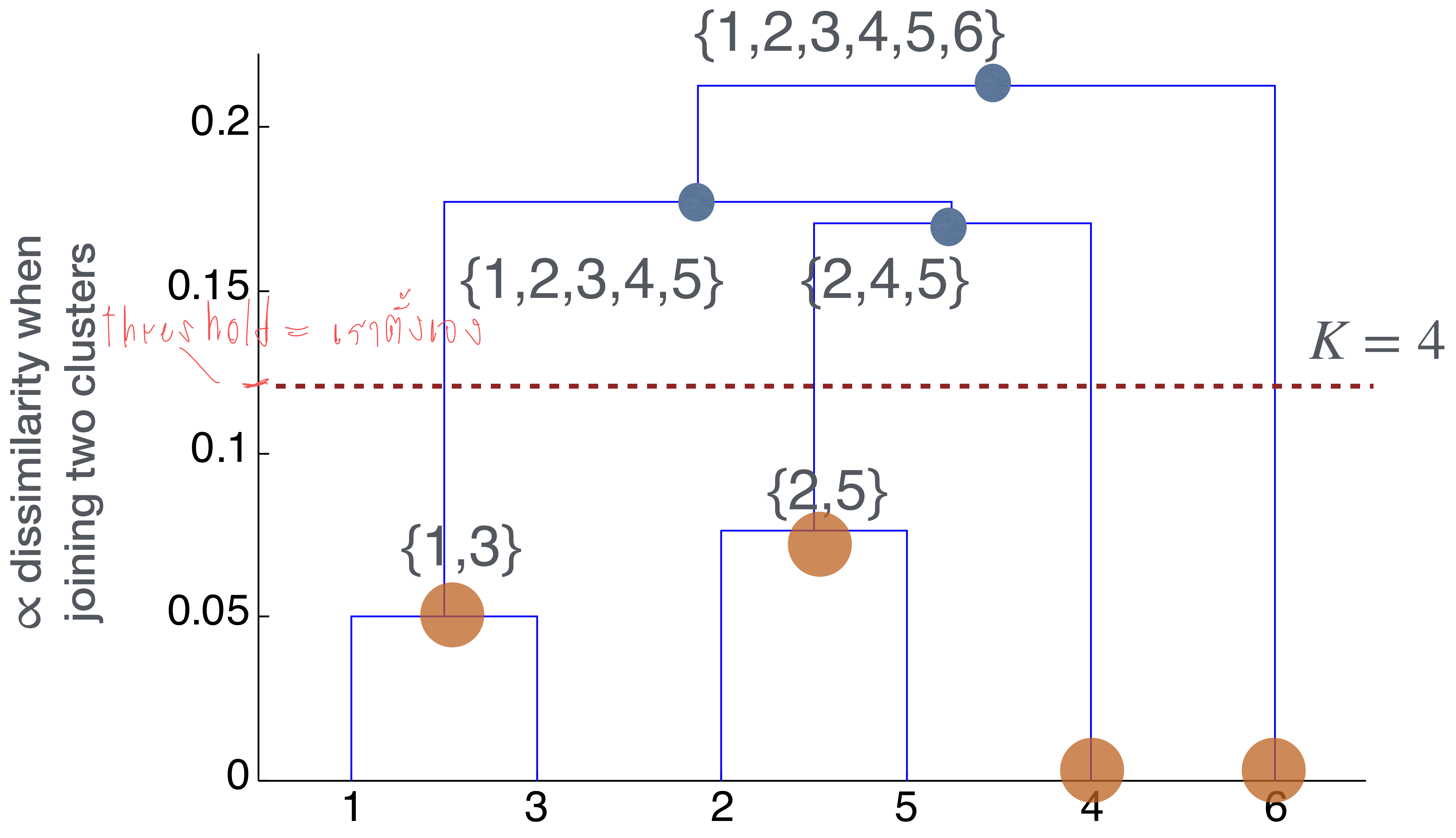


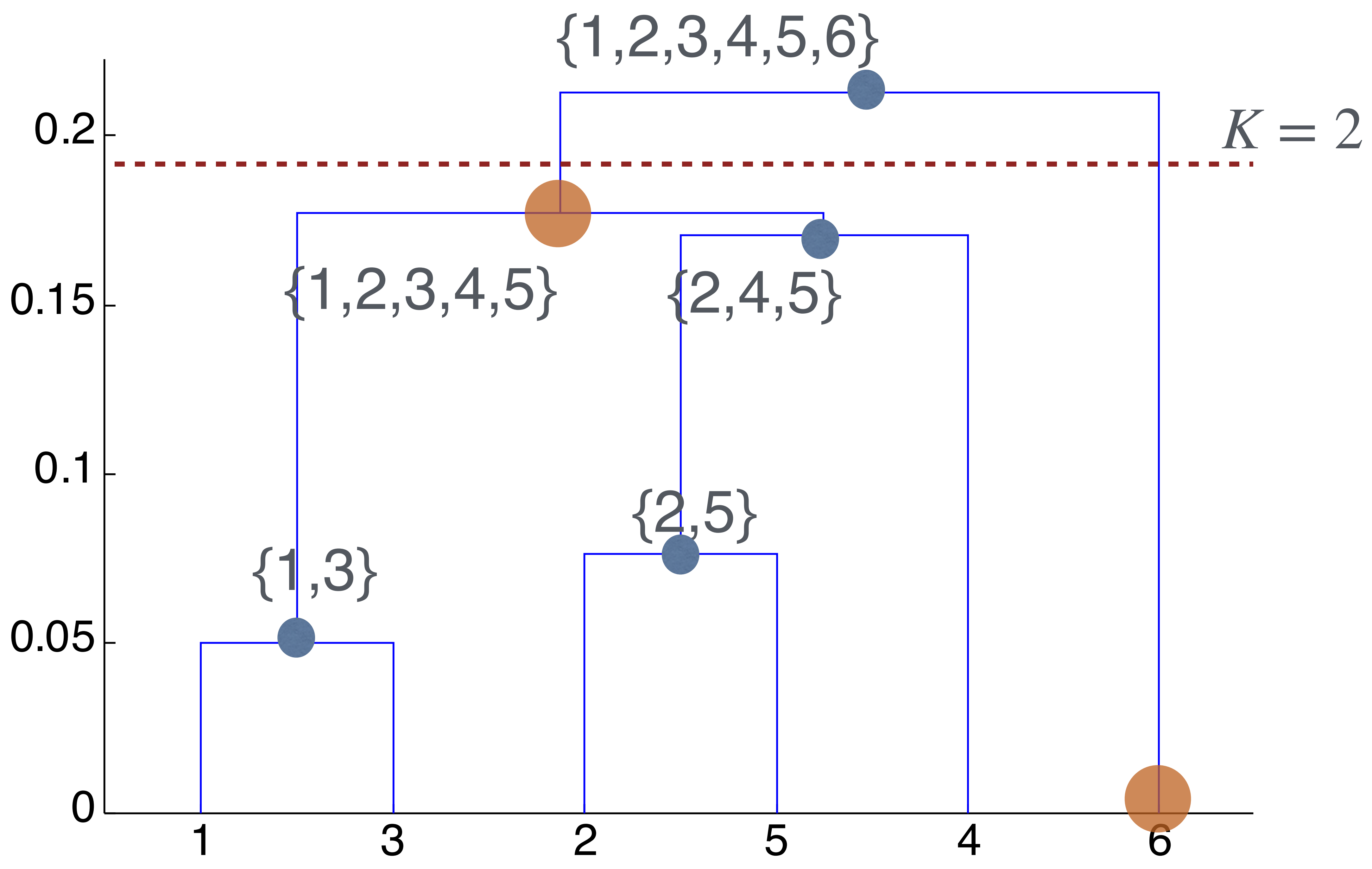
new 1 data point = mean({1,3})



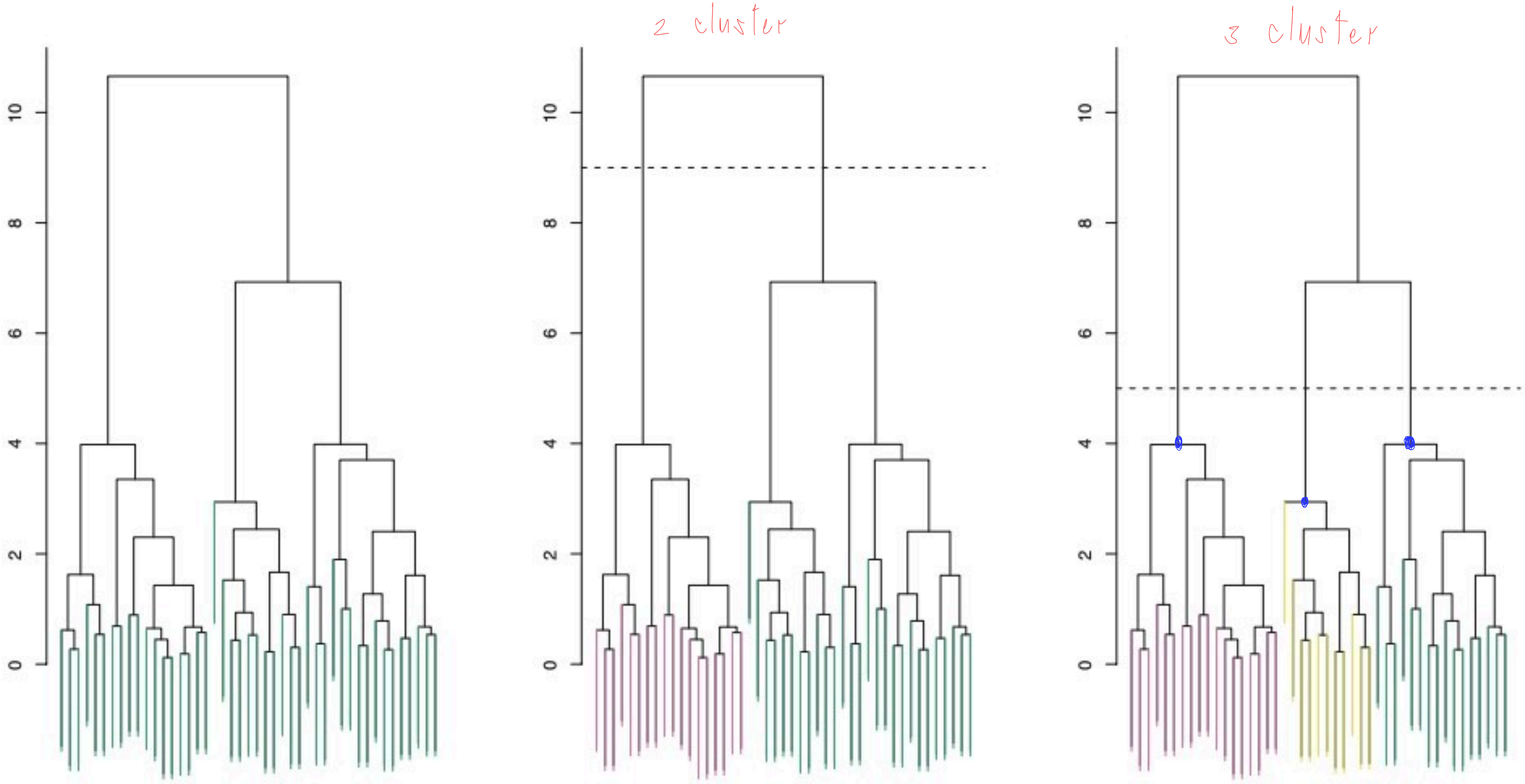
Dendrogram

Nodes directly under cut line = n cluster



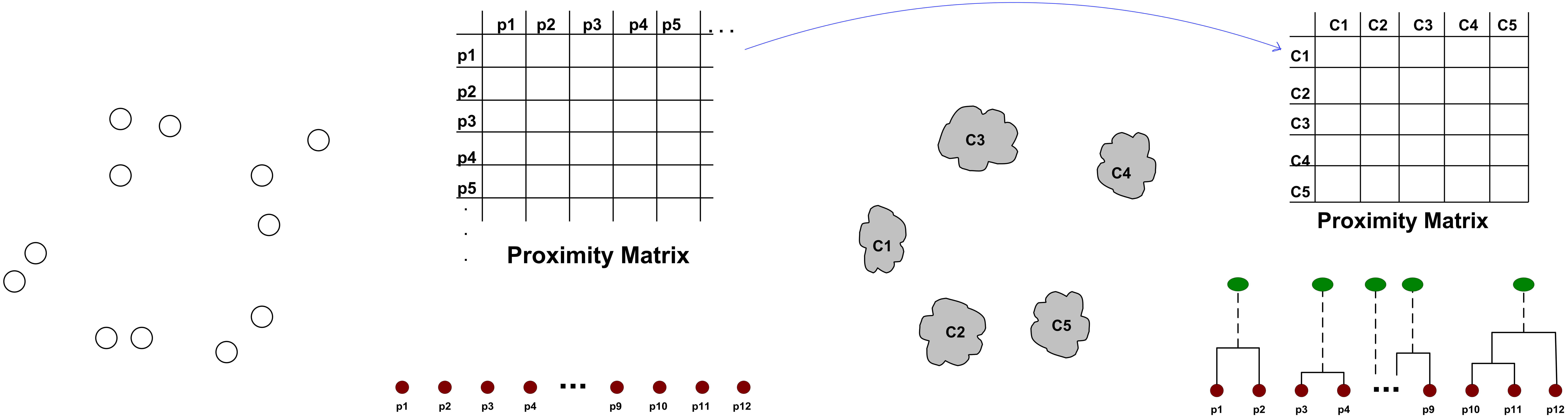


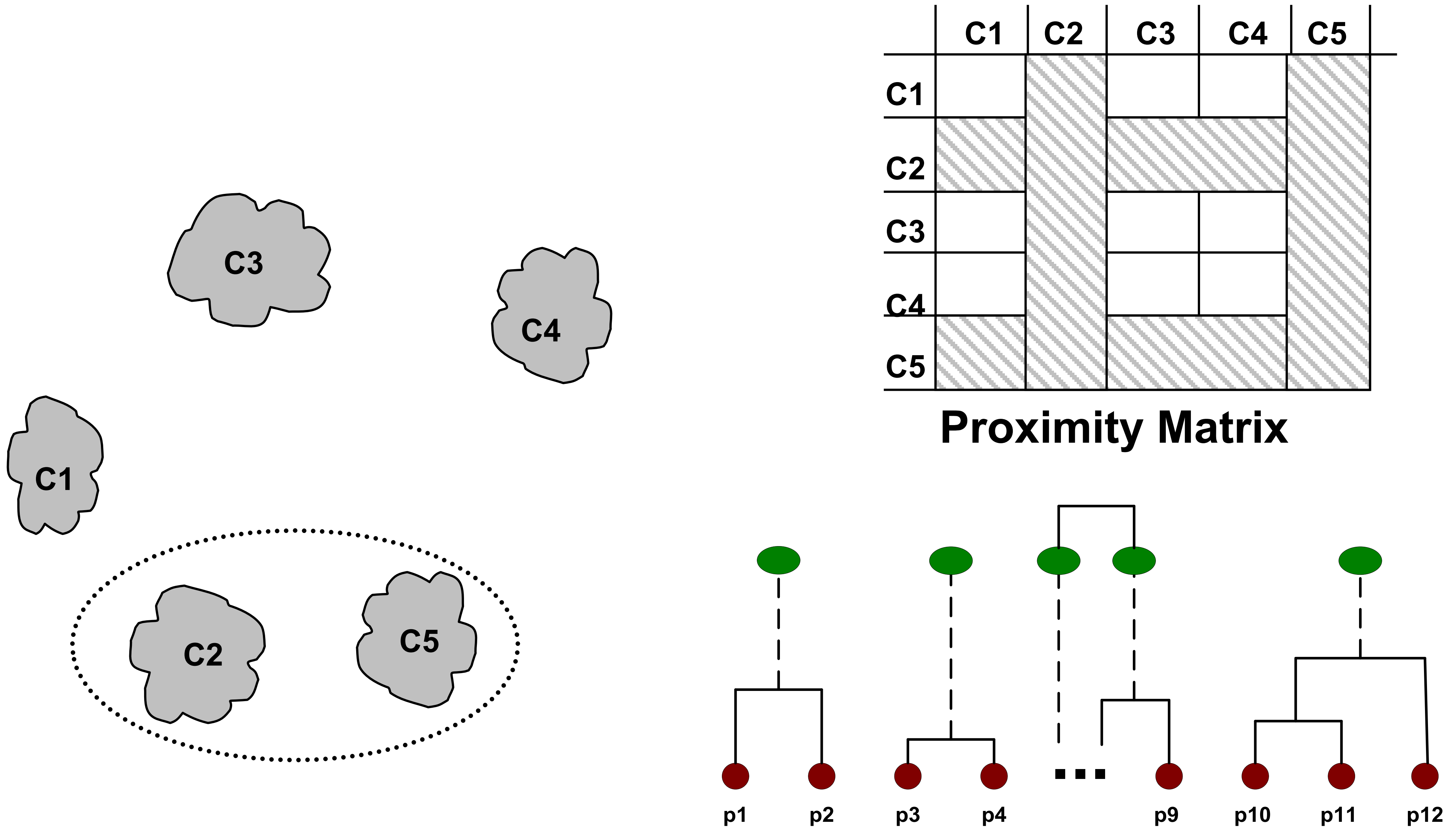
Node height = Dissimilarity of two merged clusters

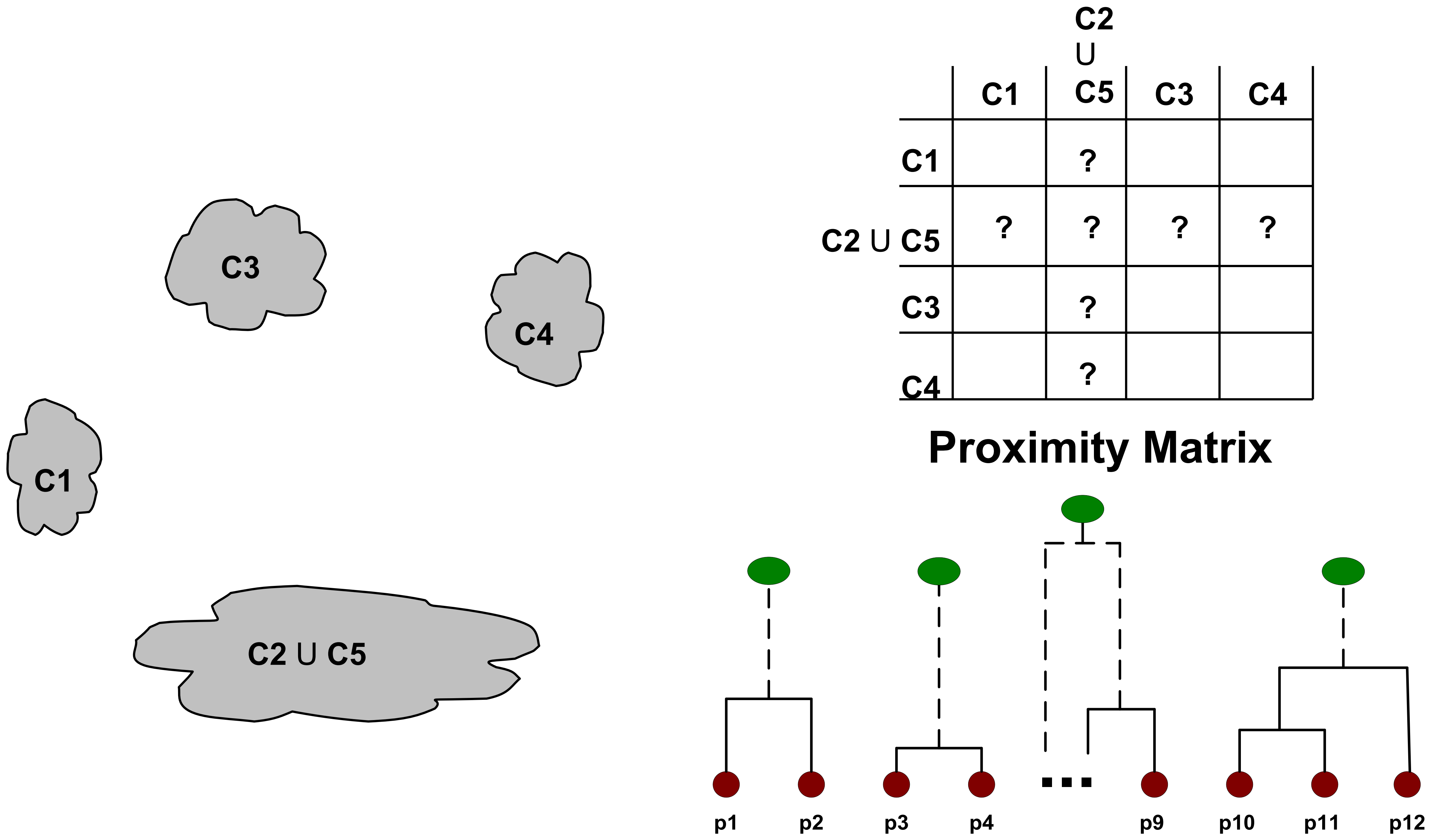


2 cluster
3 cluster
0
2
4
6
8
10

Merging Clusters




C1
C2
C3
C4
C5
Proximity Matrix
p1
p2
p3
p4
p9
p10
p11
p12


C1
C3
C4
C2 U C5
C2 U C5
?
Proximity Matrix
p1
p2
p3
p4
p9
p10
p11
p12

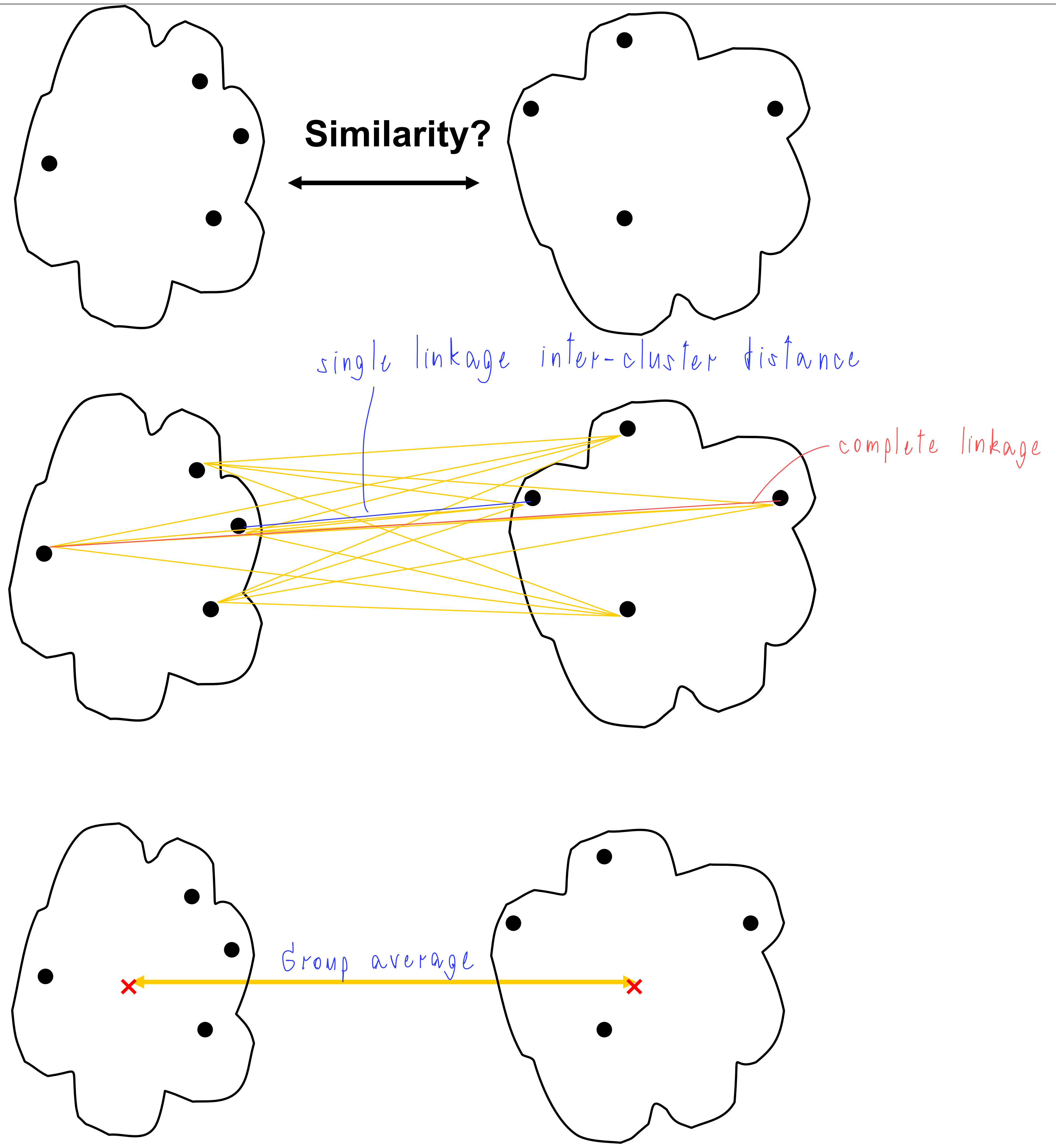
Measures of Inter-Cluster Distance

MIN (Single linkage)

MAX (Complete linkage)

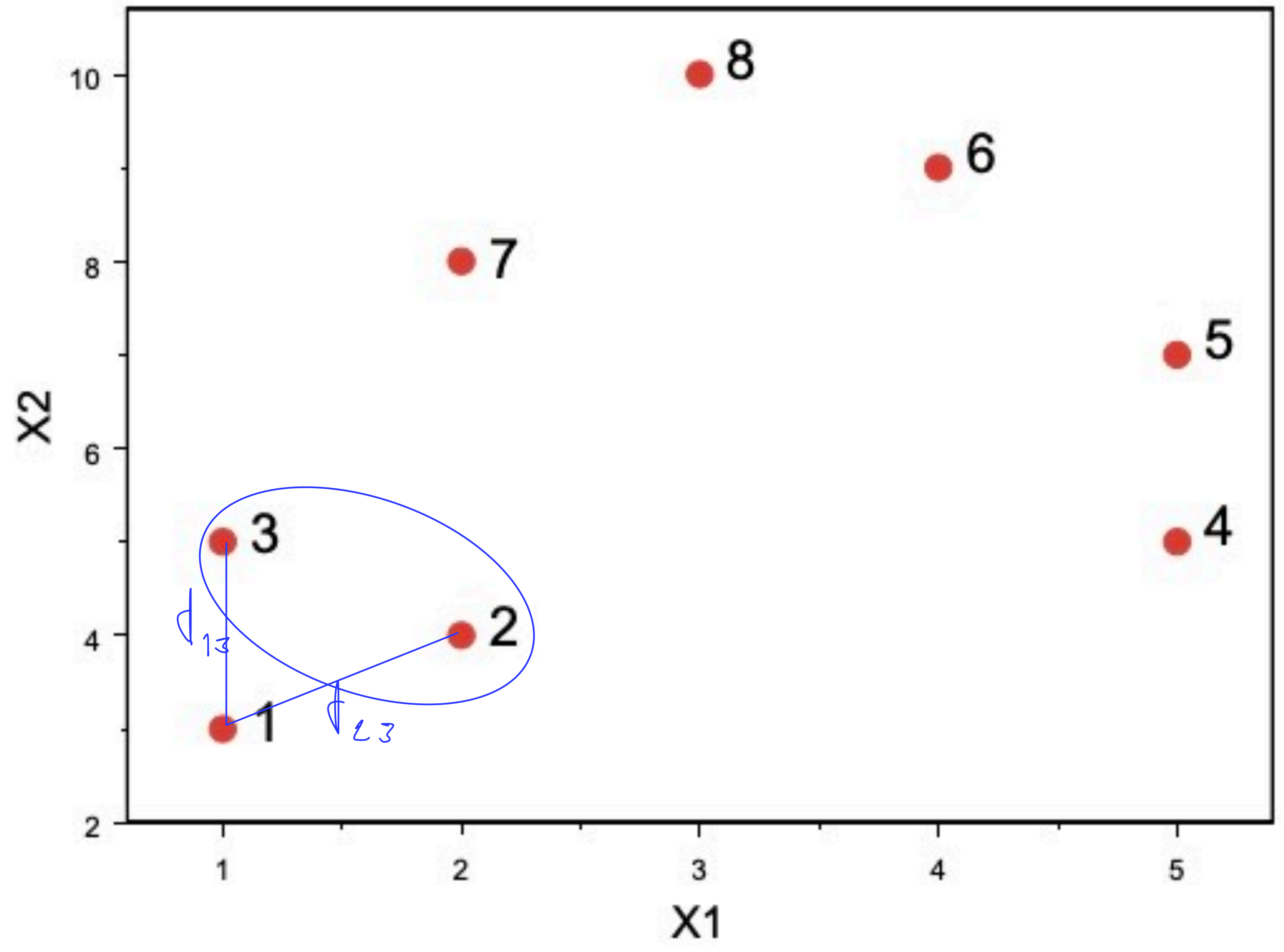
Group Average (Average linkage)

Ward's Method



$$\mathbf{x}_1 = (1,3)^\tau, \mathbf{x}_2 = (2,4)^\tau, \mathbf{x}_3 = (1,5)^\tau, \mathbf{x}_4 = (5,5)^\tau,$$
$$\mathbf{x}_5 = (5,7)^\tau, \mathbf{x}_6 = (4,9)^\tau, \mathbf{x}_7 = (2,8)^\tau, \mathbf{x}_8 = (3,10)^\tau$$

Ex: Single Linkage



Single linkage: $(p_2, p_{23}) = \min(d_{13}, d_{23})$

Pair-wise distance matrix – static ตลอด

	1	2	3	4	5	6	7	8
1	0	1.414	2.000	4.472	5.657	6.708	5.099	7.280
2		0	1.414	3.162	4.243	5.385	4.000	6.083
3			0	4.000	4.472	5.000	3.162	5.385
4				0	2.000	4.123	4.243	5.385
5					0	2.236	3.162	3.606
6						0	2.236	1.414
7							0	2.236
8								0

merge {2,3}

Find pairs with smallest dissimilarities

$$d_{12} = d_{23} = d_{68} = 1.414$$

Merge $\mathbf{x}_2$ and $\mathbf{x}_3$ to {23} cluster and update the proximity matrix

Updated proximity matrix

	1	23	4	5	6	7	8
1	0	1.414	4.472	5.657	6.708	5.099	7.280
23		0	3.162	4.243	5.000	3.162	5.385
4			0	2.000	4.123	4.243	5.385
5				0	2.236	3.162	3.606
6					0	2.236	1.414
7						0	2.236
8							0

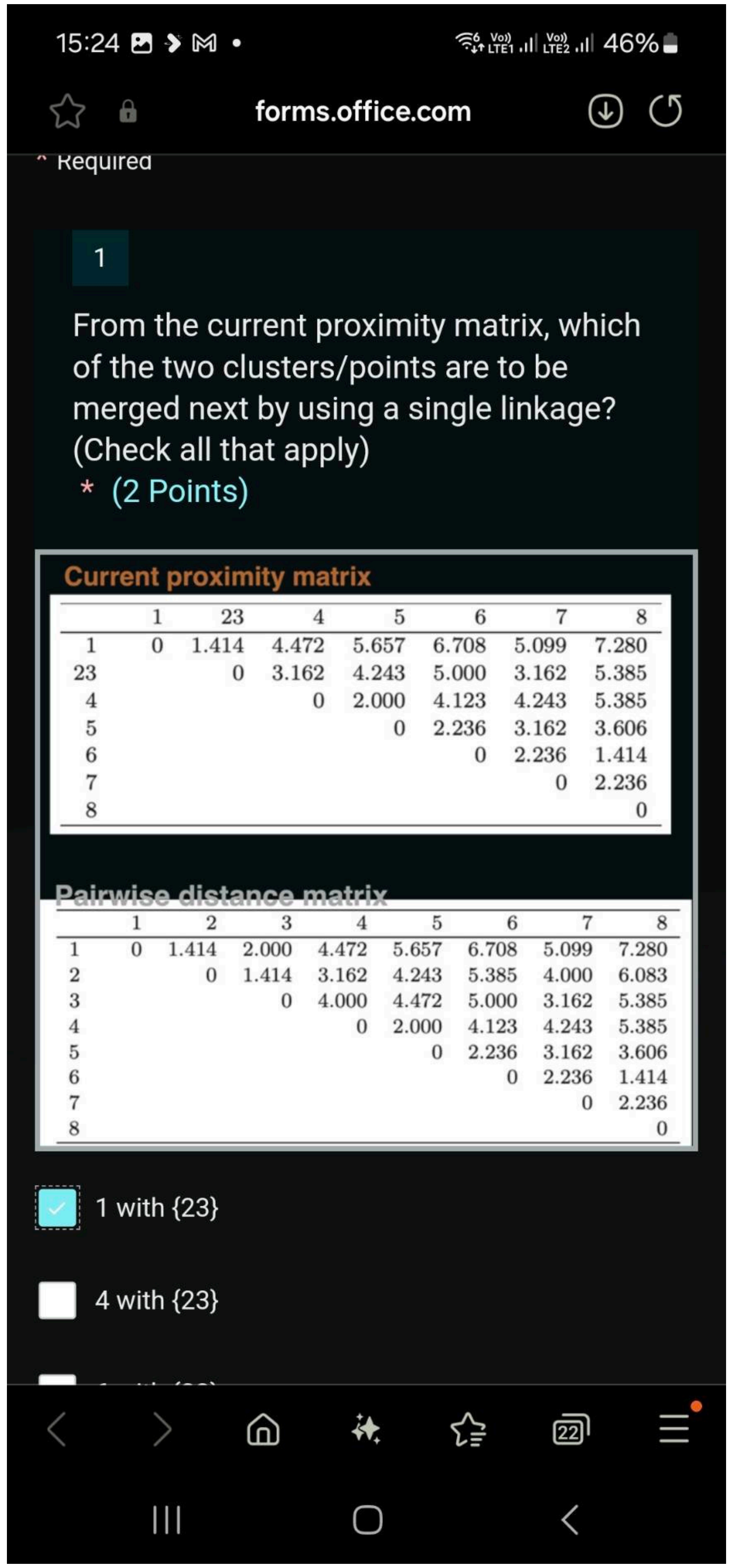
Current proximity matrix

	1	23	4	5	6	7	8
1	0	1.414	4.472	5.657	6.708	5.099	7.280
23		0	3.162	4.243	5.000	3.162	5.385
4			0	2.000	4.123	4.243	5.385
5				0	2.236	3.162	3.606
6					0	2.236	1.414
7						0	2.236
8							0

Ans. Group {1, 23} หรือ {6, 8} ก็ได้

Pairwise distance matrix

	1	2	3	4	5	6	7	8
1	0	1.414	2.000	4.472	5.657	6.708	5.099	7.280
2		0	1.414	3.162	4.243	5.385	4.000	6.083
3			0	4.000	4.472	5.000	3.162	5.385
4				0	2.000	4.123	4.243	5.385
5					0	2.236	3.162	3.606
6						0	2.236	1.414
7							0	2.236
8								0



Ward's Method

Merge two clusters that results in a smallest increase in the total sum of squared errors.

Let ESS_i be the SS error of every points from the centroid of cluster i.

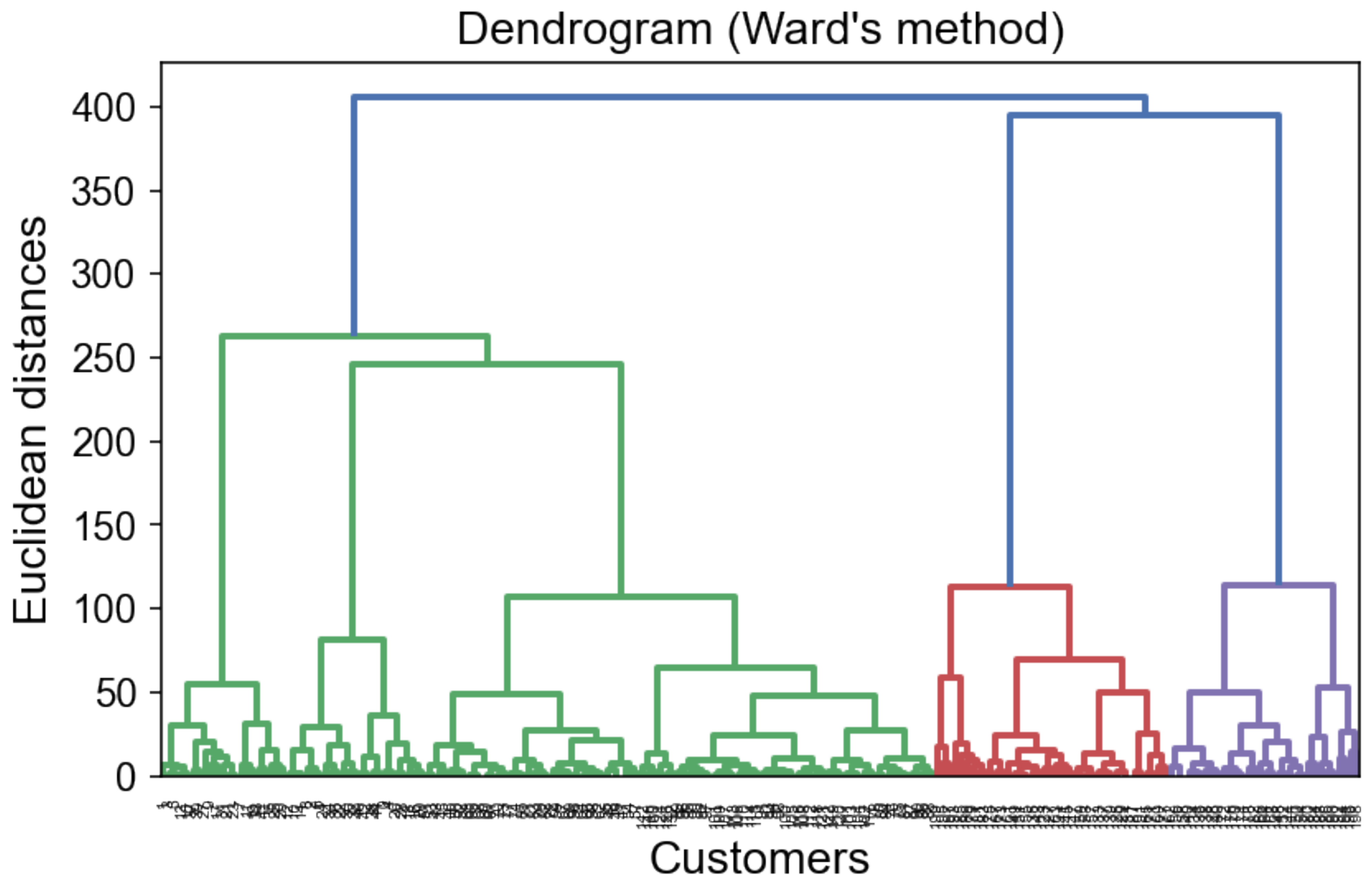
For K clusters, the total SS is $\mathrm{ESS}_K = \sum_{i=1}^{K} \mathrm{ESS}_i$

- Determine all $\binom{K}{2}$ combinations of clusters to join such that $\underbrace{\mathrm{ESS}_{K-1}}_{\text{after merging}} - \overbrace{\mathrm{ESS}_K}^{\text{before merging}}$ is smallest.
- Repeat until only one cluster is left.

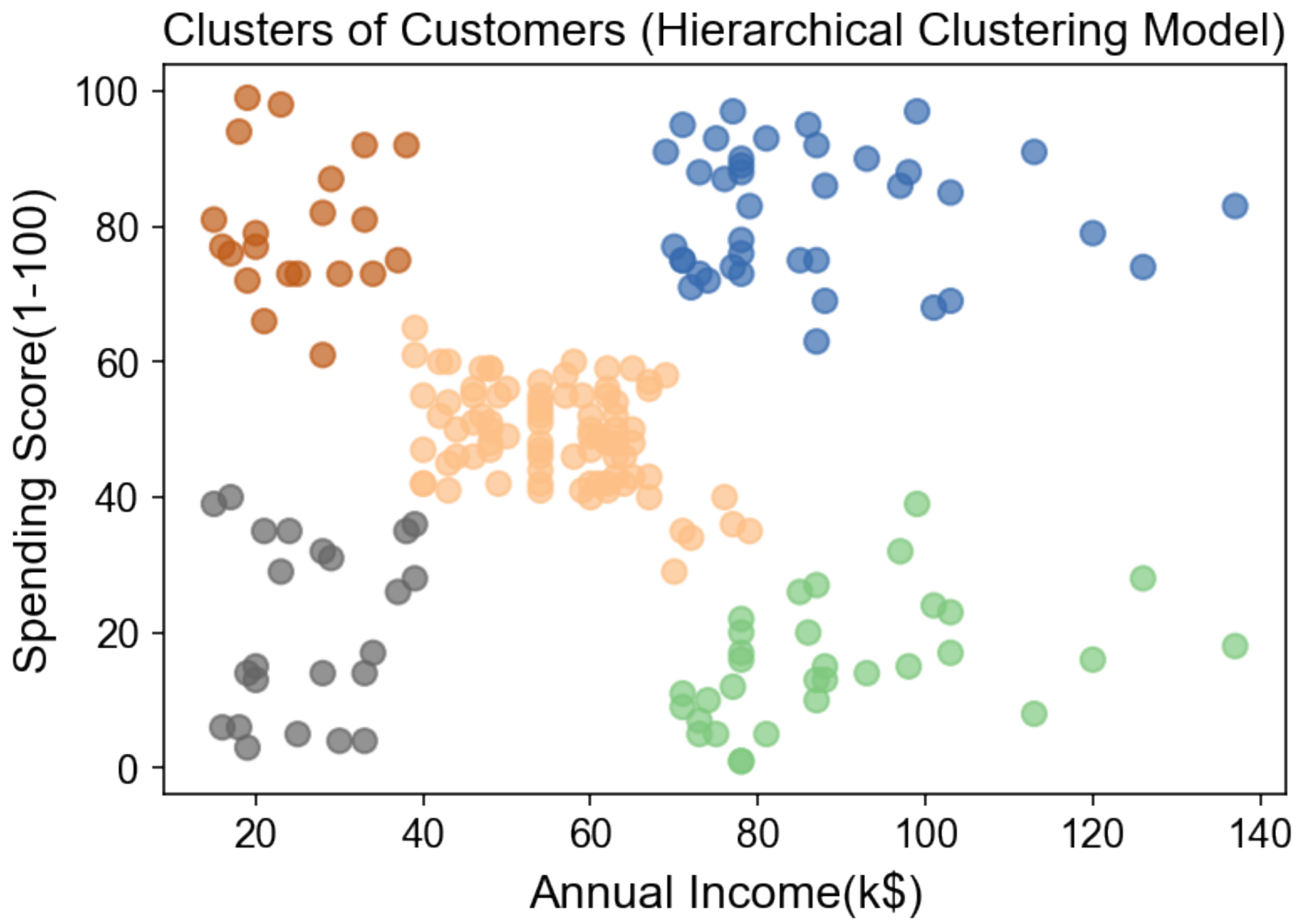
Data from sheet MALLCUST in clustering-adv.xlsx

	Gender	Age	Annual Income	SpendingScore
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

Only three features used: Age, Annual income, Spending Score

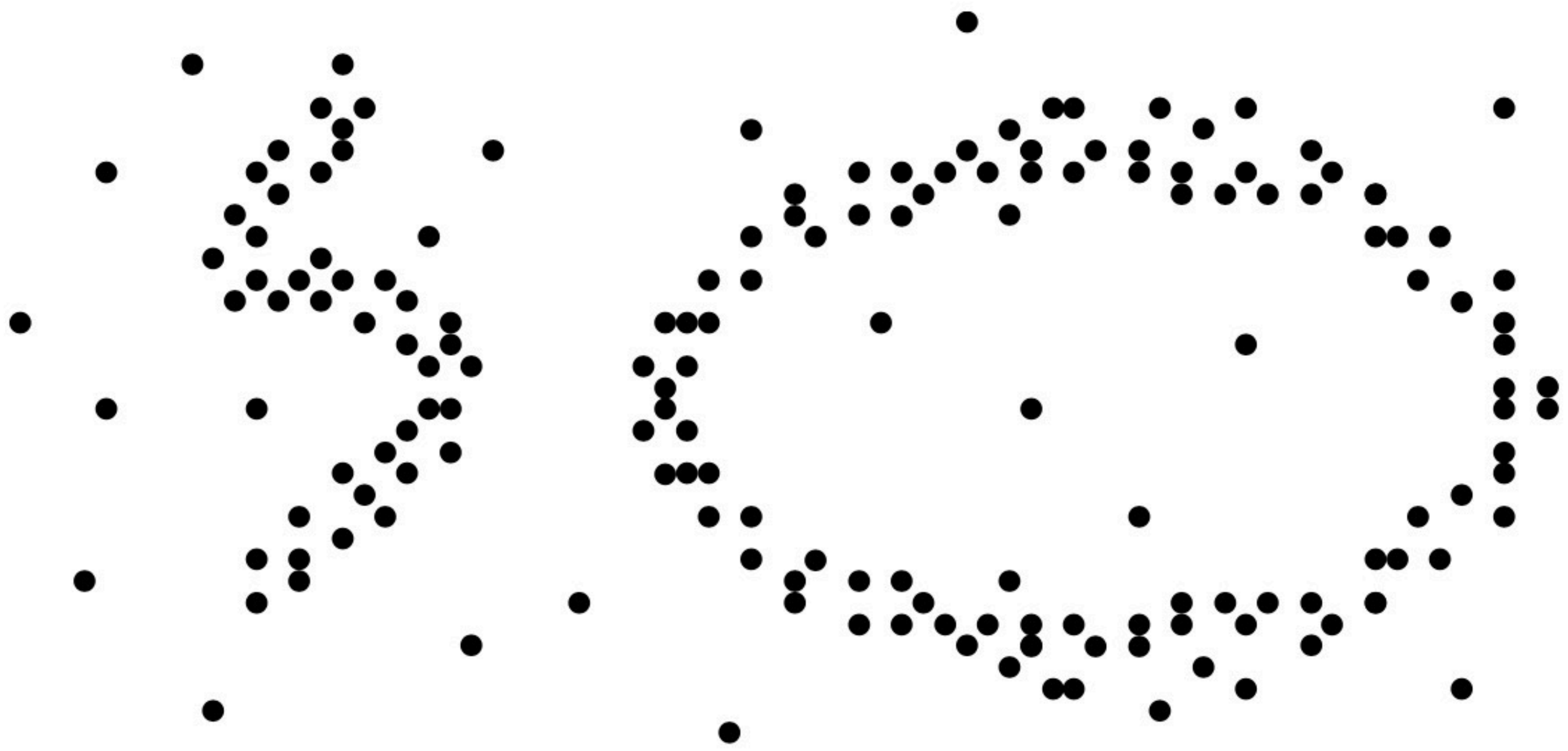


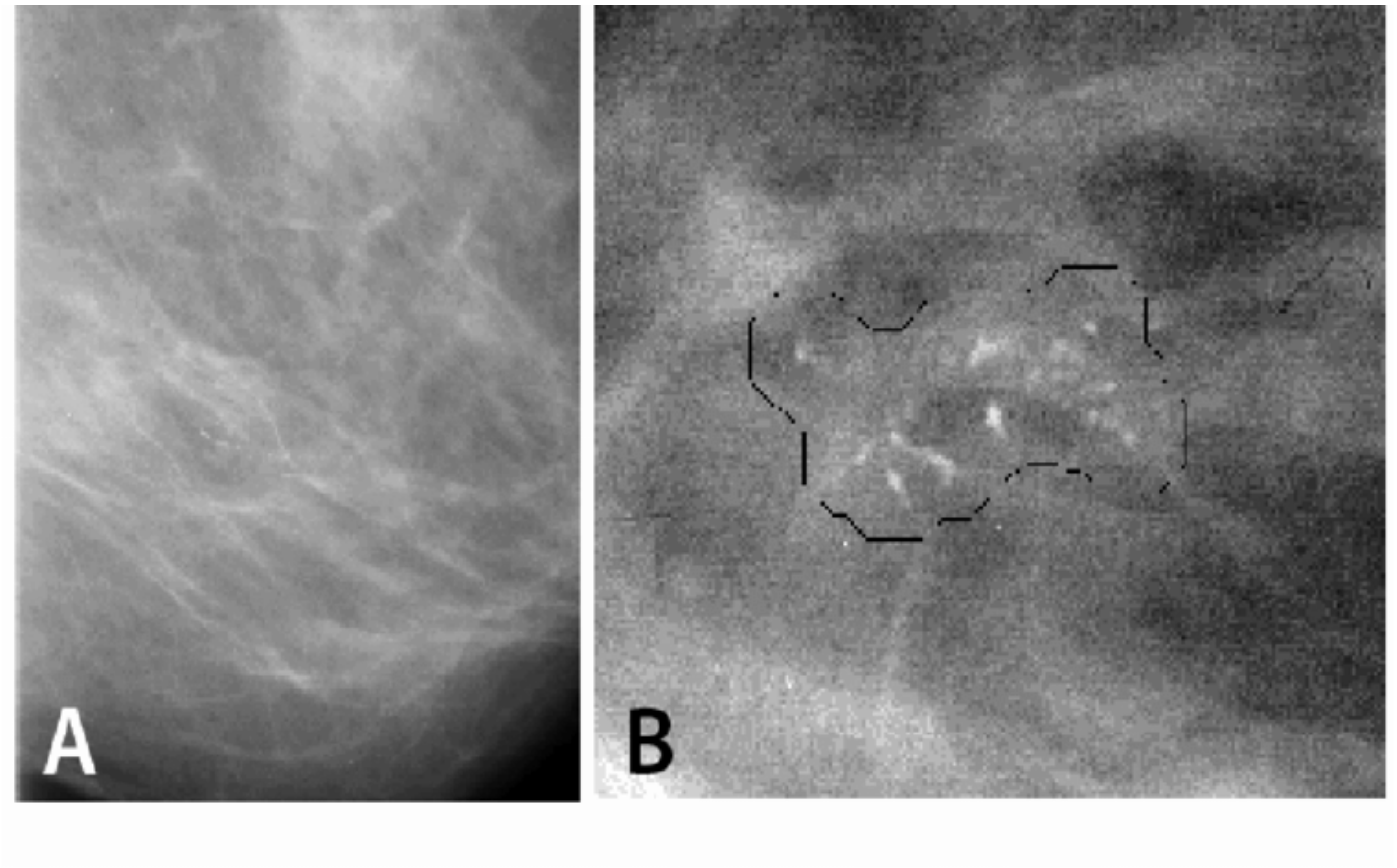

```python
import scipy.cluster.hierarchy as sch
dendrogram = sch.dendrogram(sch.linkage(X, method  = "ward"))
plt.title('Dendrogram (Ward\'s method)')
plt.xlabel('Customers')
plt.ylabel('Euclidean distances')
plt.tight_layout();
```




```python
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = 5, affinity = 'euclidean', linkage ='ward')
y_hc=hc.fit_predict(X)
```

Density-based Clustering



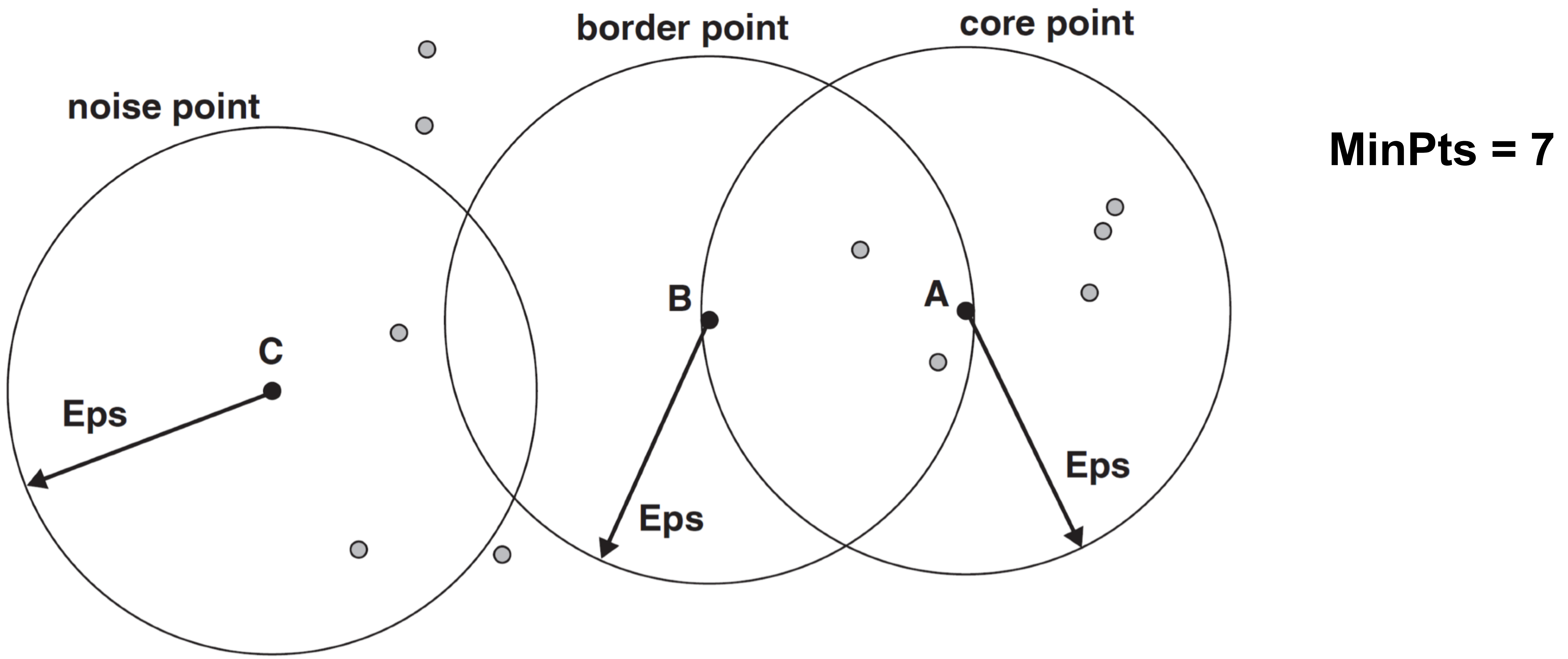


DBSCAN : Density-Based Spatial Clustering of Applications with Noise

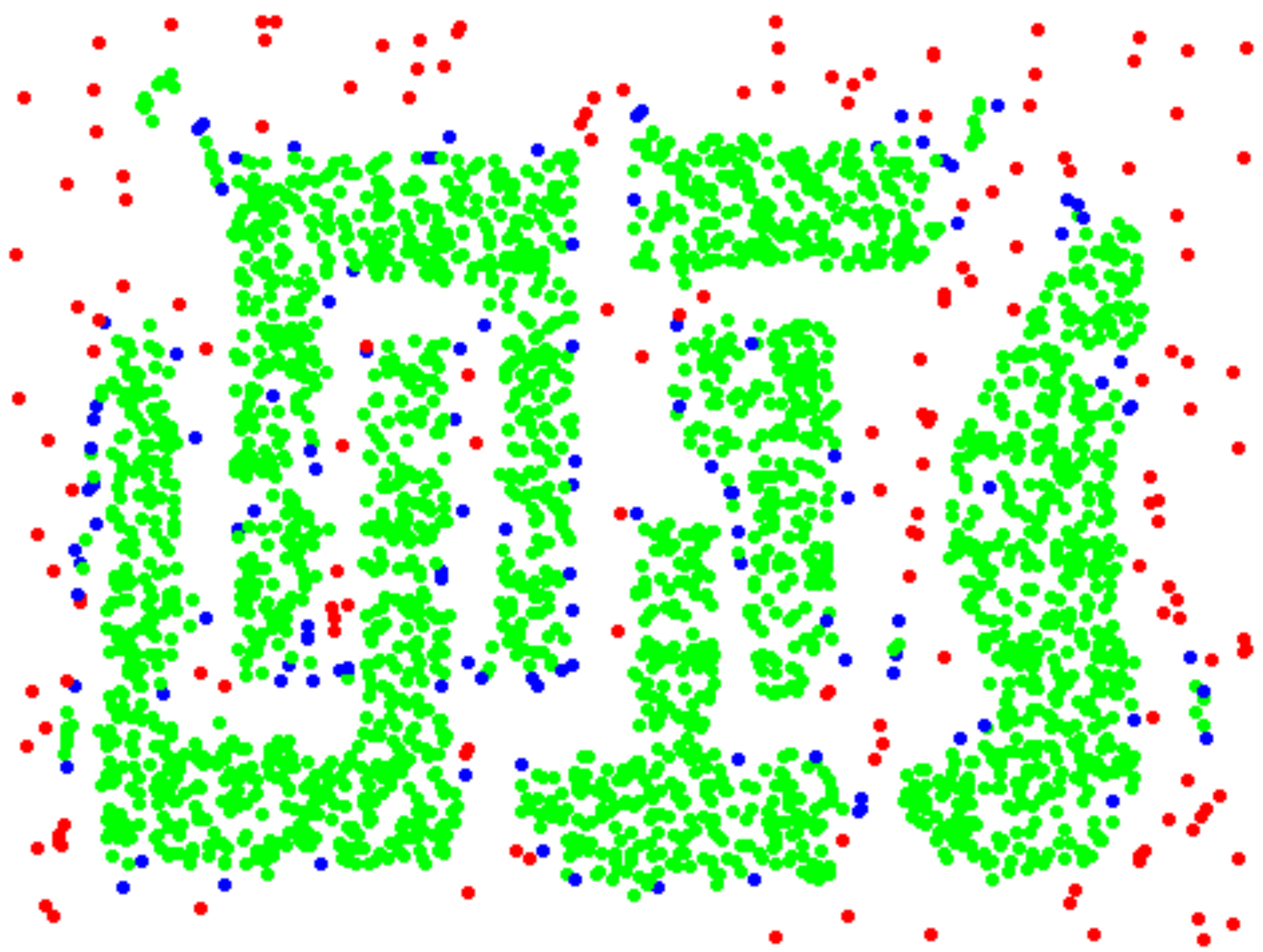
Find objects with dense neighborhood and connect them to form dense regions as clusters.

Two hyper-parameters

- Density threshold of dense region (MinPts) core border noise
- Radius of neighborhood (ε or Eps)

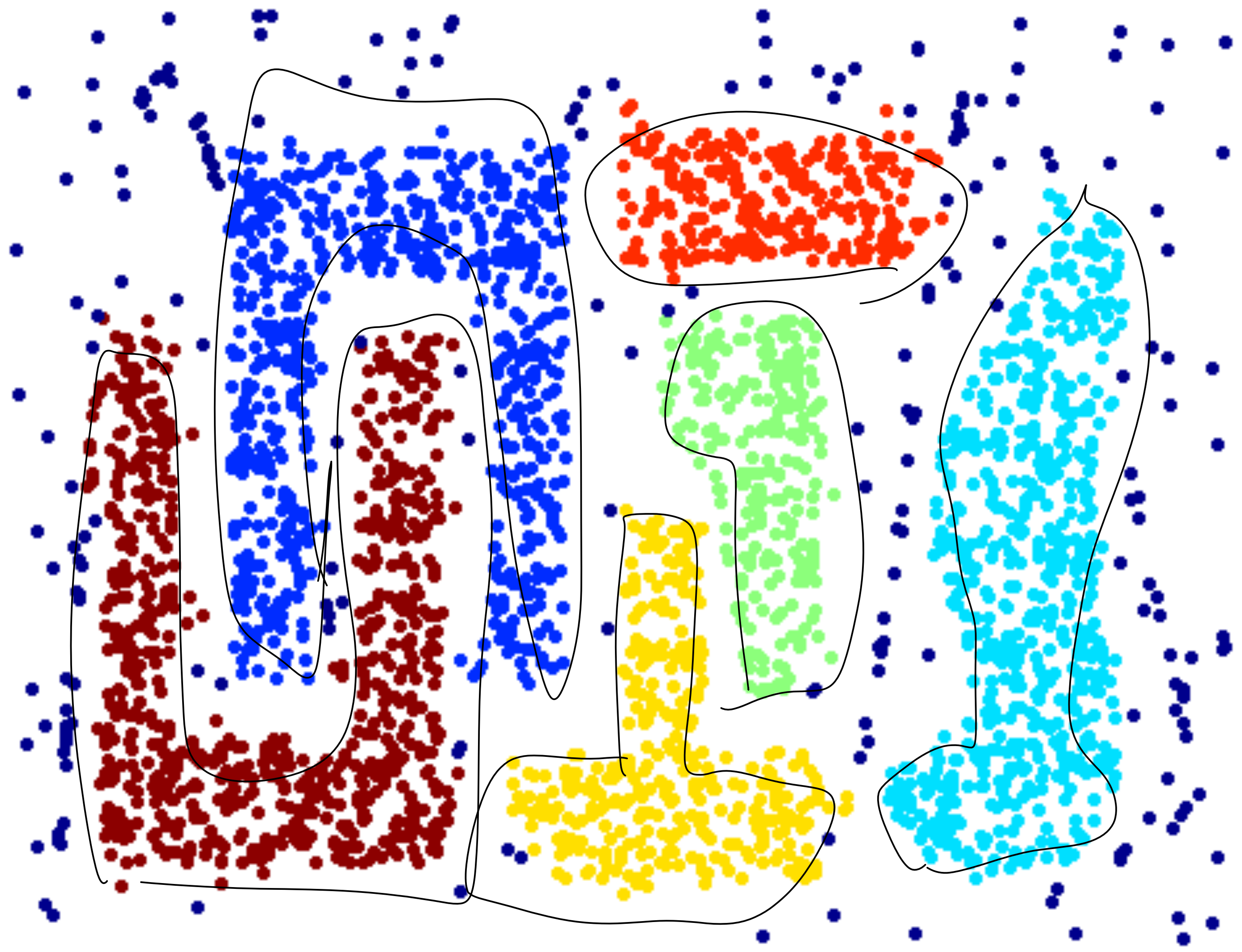


Original Points



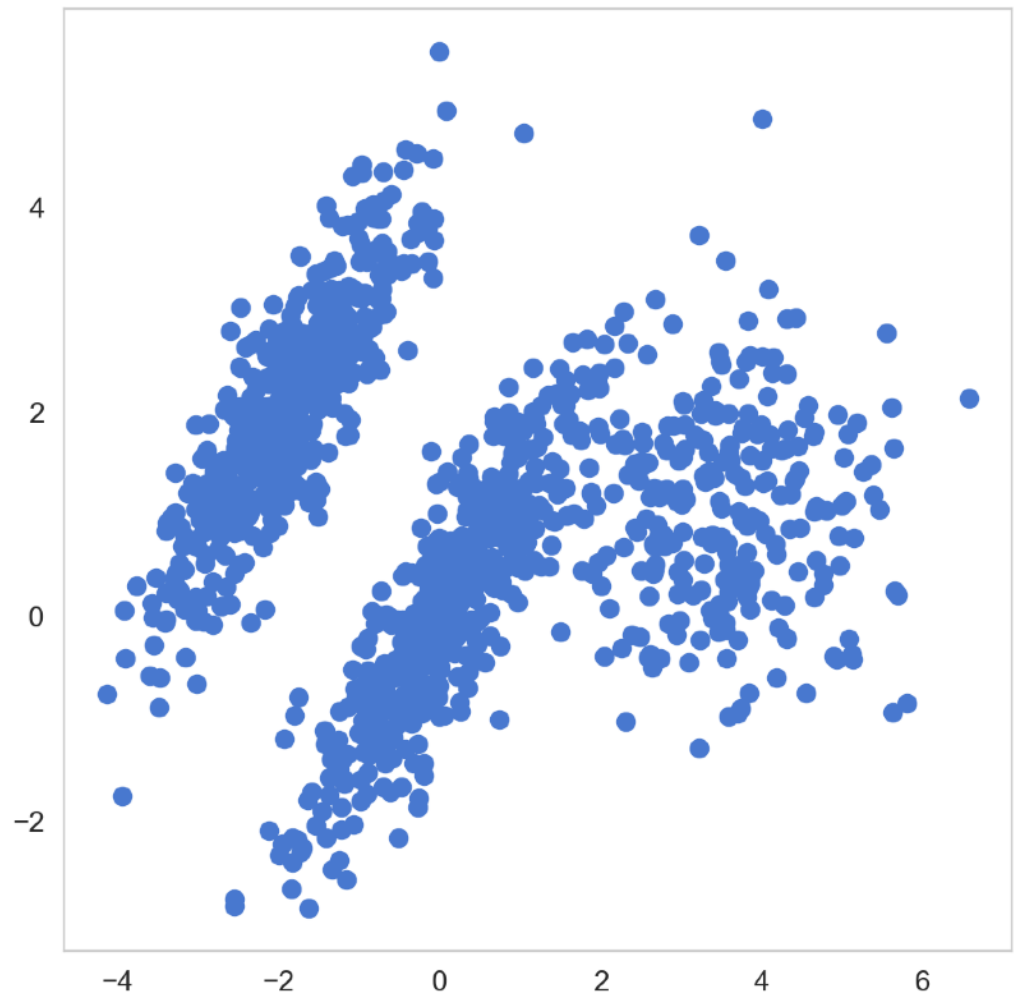
Point types: core, border and noise

Eps = 10, MinPts = 4



Cluster = Core points + Neighbors (Core/ Border points)

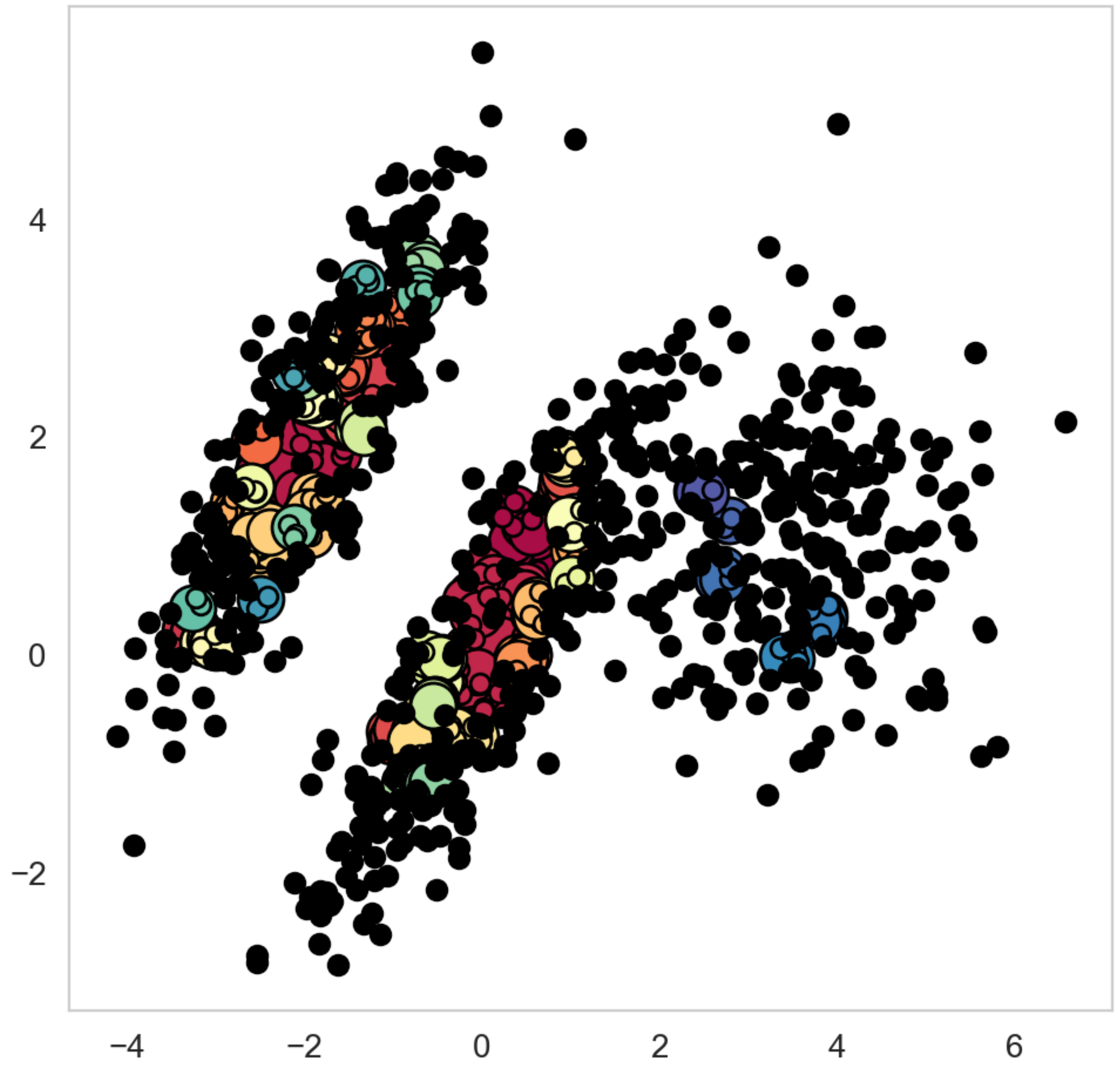
For a given MinPts, what happen if Eps is too small or too large ?



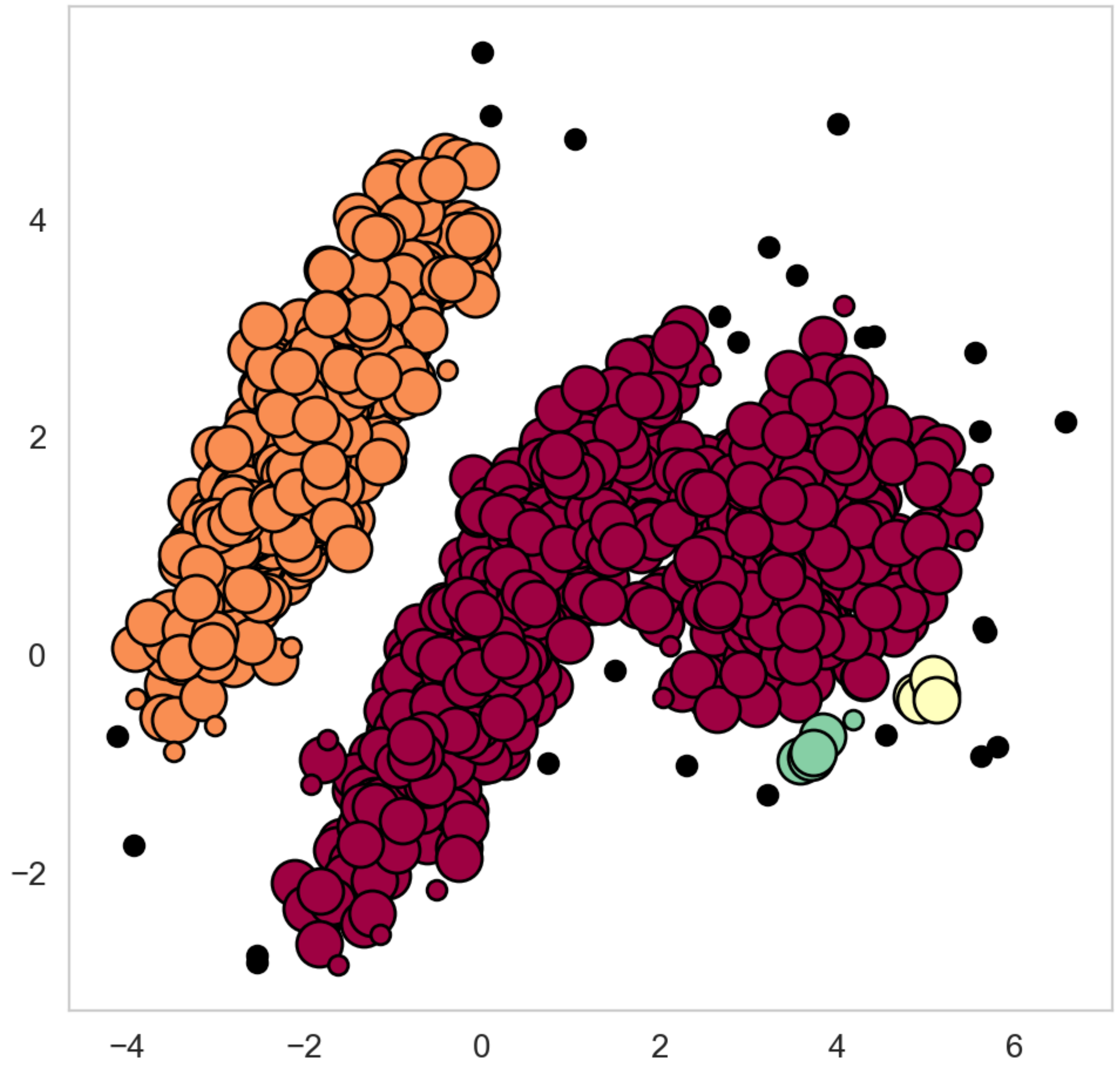

```python
from sklearn.cluster import DBSCAN

MinPts,Eps=4,1.2
dbc = DBSCAN(eps=Eps, min_samples=MinPts).fit(X)
labels = dbc.labels_
```

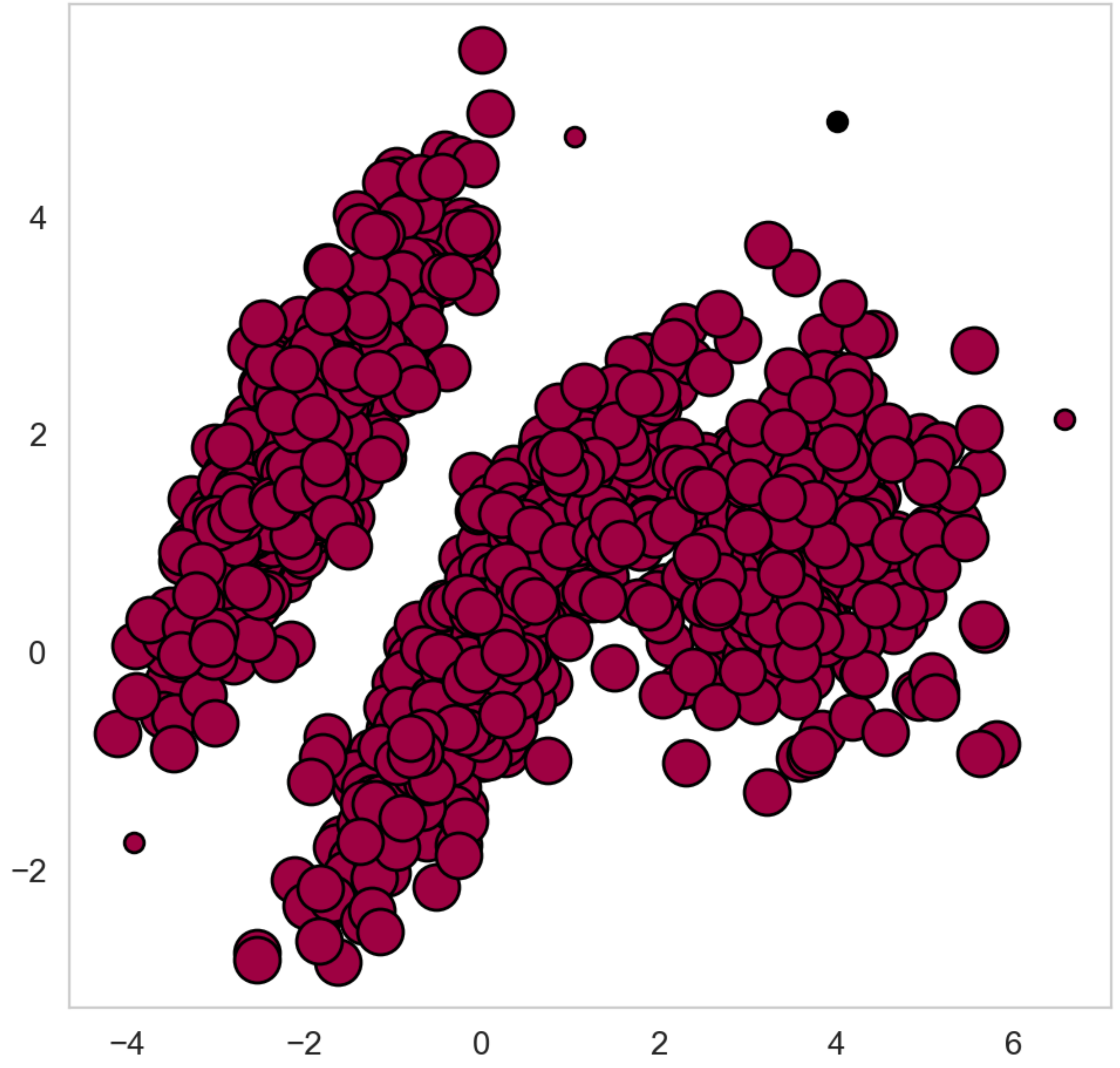
MinPts = 4, Eps=0.1, # Clusters = 46, # Noise points = 608



MinPts = 4, Eps=0.4, # Clusters = 4, # Noise points = 26



MinPts = 4, Eps=1.2, # Clusters = 1, # Noise points = 1

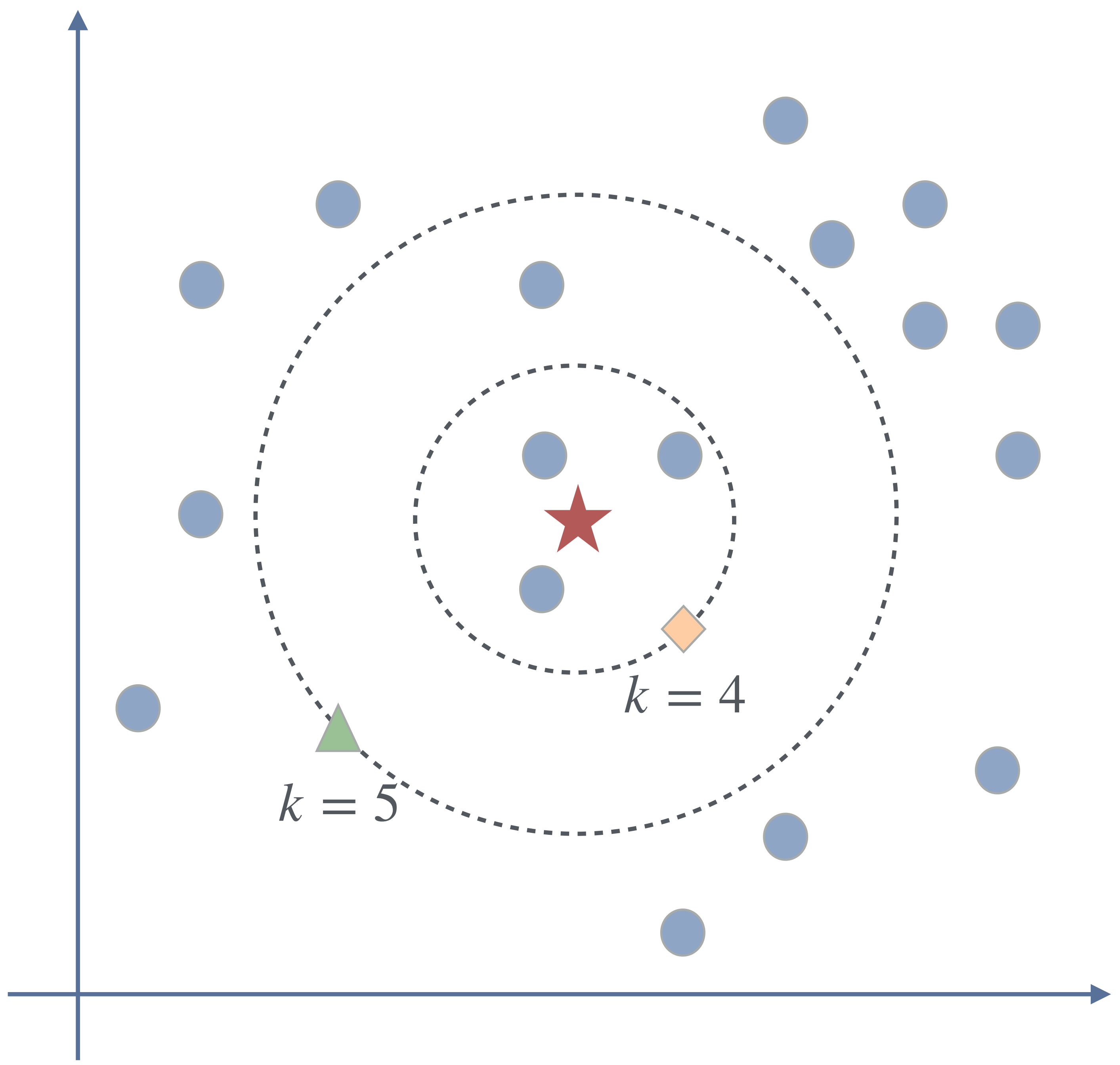


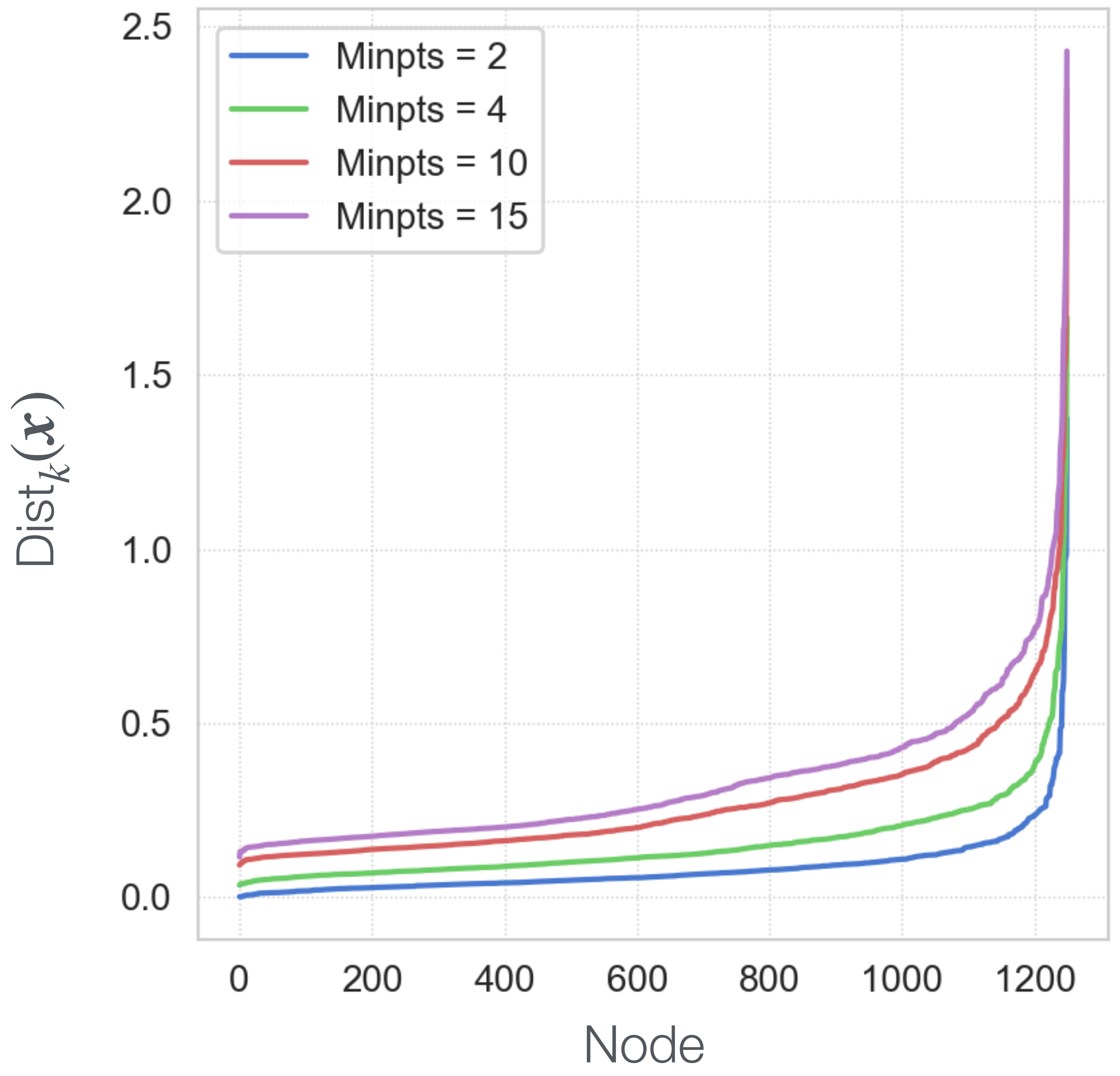
Choice of MinPts and Eps

Pick a value of MinPts and set k = MinPts.

For every point $\boldsymbol{x}$, look at $\text{dist}_k(\mathbf{x})$, distance to the k^{th} nearest neighbor.

- Small values of $\text{dist}_k(\mathbf{x})$ for points in cluster and relatively large for noise points.
- Knee point of $\text{dist}_k(\mathbf{x})$ indicates suitable value of Eps for a given MinPts $= k$.





Summary

GMM clustering

- Soft assignment based on the mixture model of Gaussian distributions
- EM algorithm to find best distribution parameters fitted to the data.
- Find elliptical-shape clusters with different sizes.

Hierarchical clustering

- Iteratively merge "closest" clusters to produces a set of nested clusters organized as a hierarchical tree.
- Clusters obtained from cutting at dendrogram at the proper level.

Density-based clustering

- Group points based on neighborhood density.
- Handle non-convex clusters and noisy data.
- Good at identifying outliers in dataset.