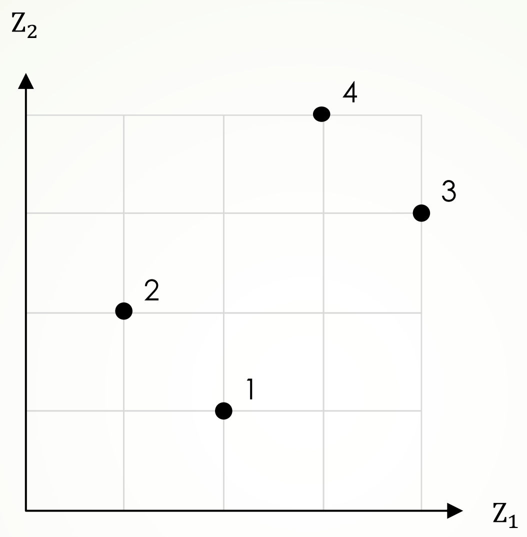
香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional · Creative
For The World

Cluster Analysis:
Partitioning Methods

CS5483 Data Warehousing and Data Mining

Group similar tuples together

	Z_1	Z_2
1.	2	1
2.	1	2
3.	3	4
4.	4	3

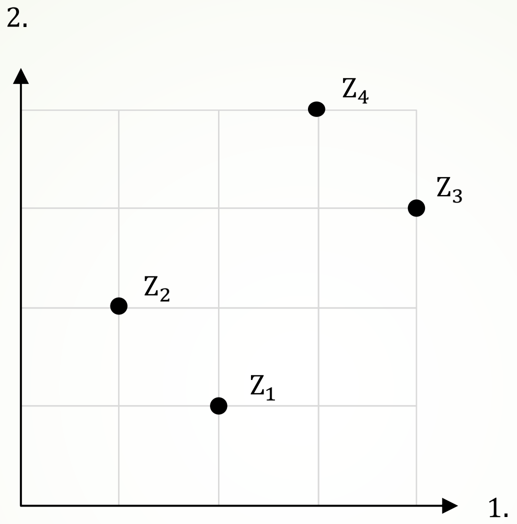


Example:

- Z_i: intensity of i-th pixel in a picture
- Clustering the tuples identifies images of same/similar objects.

Group similar features together

	Z_1	Z_2	Z_3	Z_4
1.	2	1	3	4
2.	1	2	4	3



Example:

- Z_i: expression level of gene i
- Clustering the features identifies co-expressed genes.

Partitioning method

- Input: A set $D \coloneqq \{\boldsymbol{p}_i\}_{i=1}^{n}$ of data points
- Output: A set $\{C_j\}_{j=1}^{k}$ of non-empty disjoint clusters that partition D.
- Challenges:
  - there are often **too many data points**, and
  - the **d**__________ can be **too high** to visualize.
- Need a mathematical criteria to automate clustering.

Centroid-based method

Model assumption

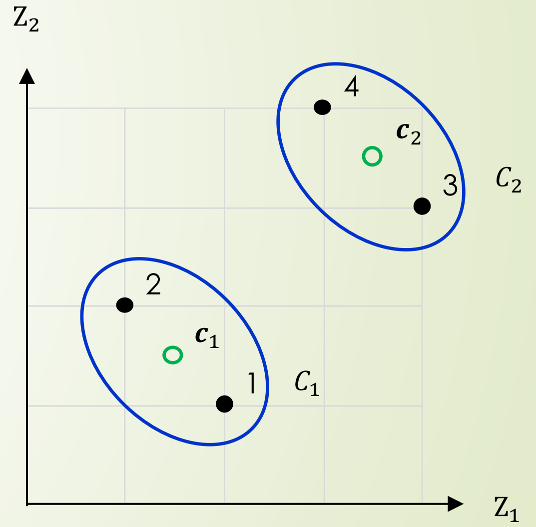
- Suppose:
  1. There is a typical point (**c**__________ **c**________) in each cluster.
  2. The **variations** of the points in the same cluster are **due to noise**.
- How to recover $\boldsymbol{c}_j$ given C_j?

$$\min_{\mathbf{c}_j} \sum_{\boldsymbol{p} \in C_j} \operatorname{dist}(\boldsymbol{p}, \boldsymbol{c}_j)^2$$

- For Euclidean distance, the solution is the **c**________________

$$\boldsymbol{c}_j = \frac{1}{|C_j|} \sum_{\boldsymbol{p} \in C_j} \boldsymbol{p}$$

- How to find C_j's?

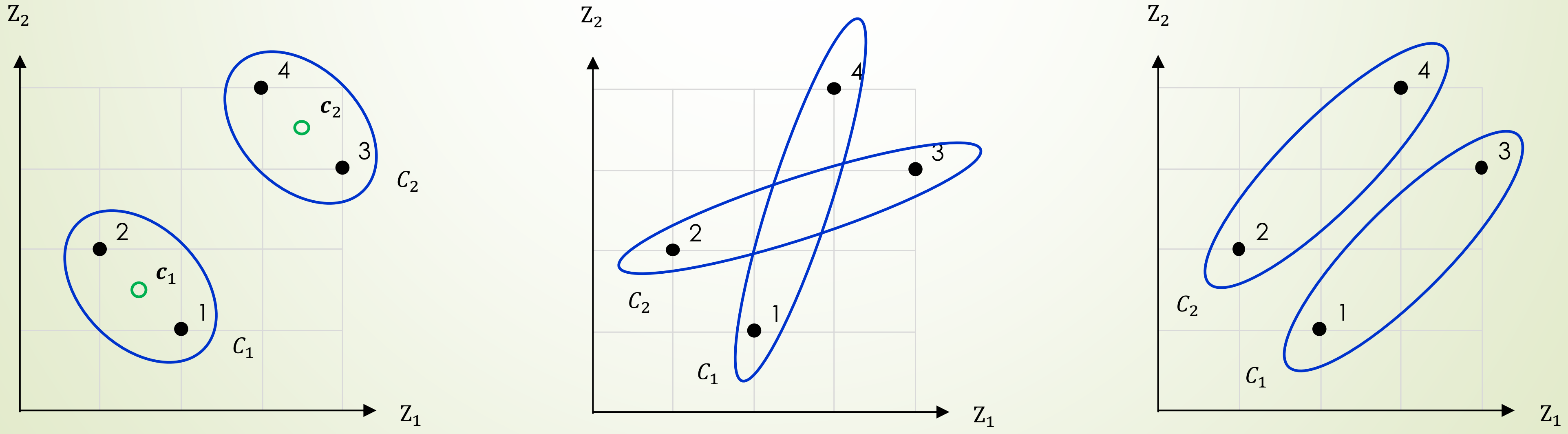


Centroid-based method
Mathematical criteria

- Given the number k of clusters, solve

$$\min_{\{C_j\}_{j=1}^{k}} \sum_{j=1}^{k} \min_{\boldsymbol{c}_j} \sum_{\boldsymbol{p}\in C_j} \operatorname{dist}(\boldsymbol{p}, \boldsymbol{c}_j)^2$$

- Example: Left/Middle/Right is the optimal clustering solution.

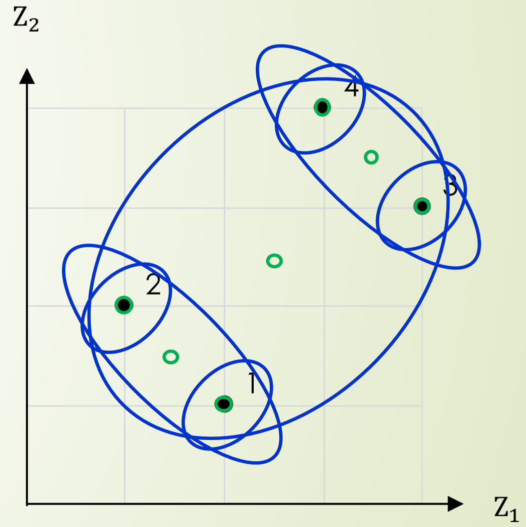


Centroid-based method
Challenges

- Given the number k of clusters, solve

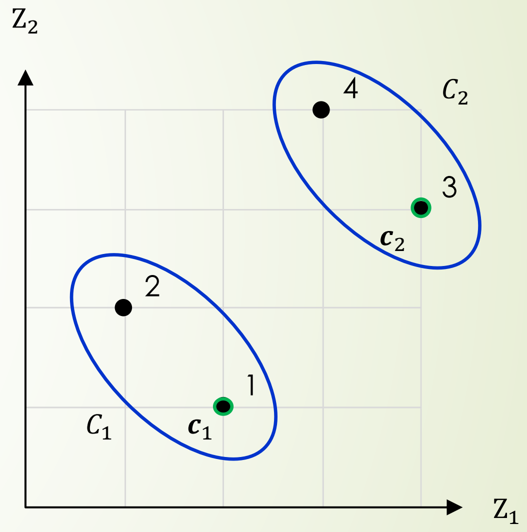
$$\min_{\{C_j\}_{j=1}^{k}} \sum_{j=1}^{k} \min_{\boldsymbol{c}_j} \sum_{\boldsymbol{p}\in C_j} \mathrm{dist}(\boldsymbol{p},\boldsymbol{c}_j)^2$$

- What if we further minimize over k?
  - $k = \underline{\quad}$, $\boldsymbol{c}_j = \underline{\quad}$, $C_j = \underline{\quad}$ (good? Why or why not?)
  - Not the right objective to find k.
  - Remedy? Assume k is given for now.
- Another Issue: Minimization is ______________________________.
  - Best bound: $O(n^{d(k+1)} \log n)$, exponential in k and the dimension d.



k-means clustering
Greedy algorithm

1. Select k tuples randomly as cluster centers initially
2. Calculate the clusters given the cluster centers
   for each $\boldsymbol{p} \in D$
   assign $\boldsymbol{p}$ to C_j where j minimizes $\mathrm{dist}(\boldsymbol{p}, \boldsymbol{c}_j)$
3. Calculate the cluster centers given the clusters
   for each j from 1 to k
   $\boldsymbol{c}_j \leftarrow \frac{1}{|C_j|} \sum_{\boldsymbol{p} \in C_j} \boldsymbol{p}$
4. Repeat 2 to 3 until no change in clusters.

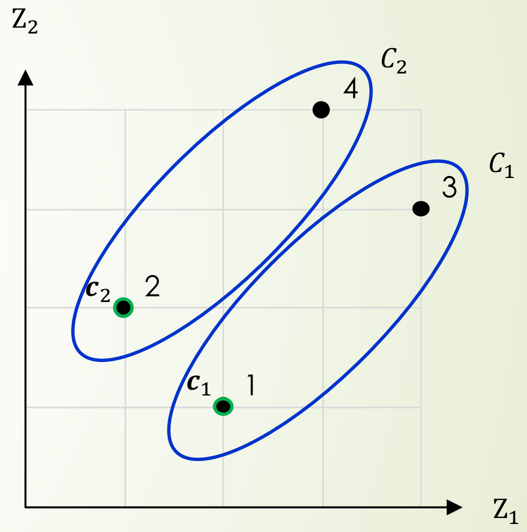


Complexity

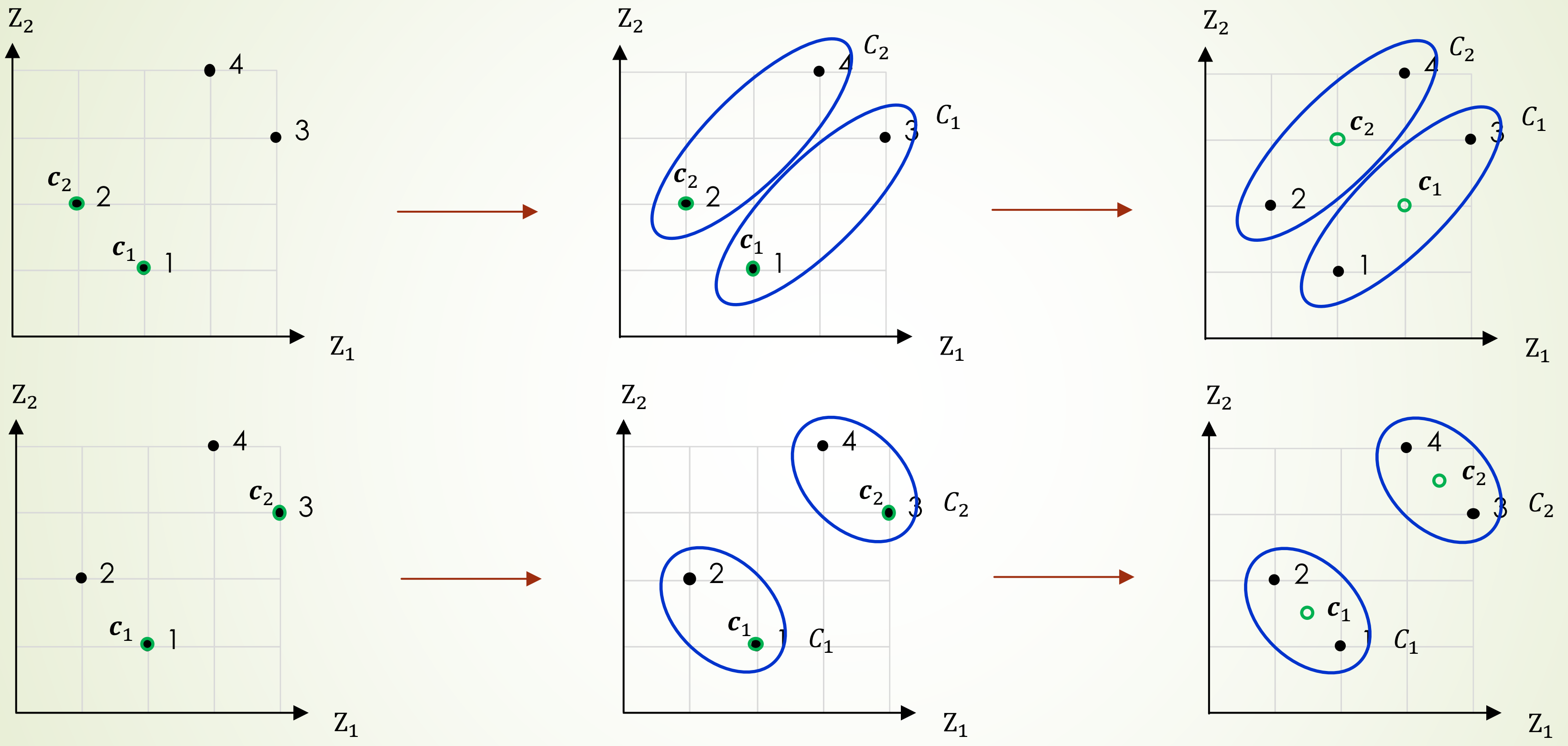
- $O(\_\_\_)$ where t is the number of iterations.
- Efficient when $k, t \ll n$.
- Does the algorithm always converge to an optimal solution?

Run again

1. Select k tuples randomly as cluster centers initially
2. Calculate the clusters given the cluster centers
   for each $\boldsymbol{p} \in D$
   assign $\boldsymbol{p}$ to C_j where j minimizes $\mathrm{dist}(\boldsymbol{p}, \boldsymbol{c}_j)$
3. Calculate the cluster centers given the clusters
   for each j from 1 to k
   $\boldsymbol{c}_j \leftarrow \frac{1}{|C_j|} \sum_{\boldsymbol{p} \in C_j} \boldsymbol{p}$
4. Repeat 2 to 3 until no change in clusters.



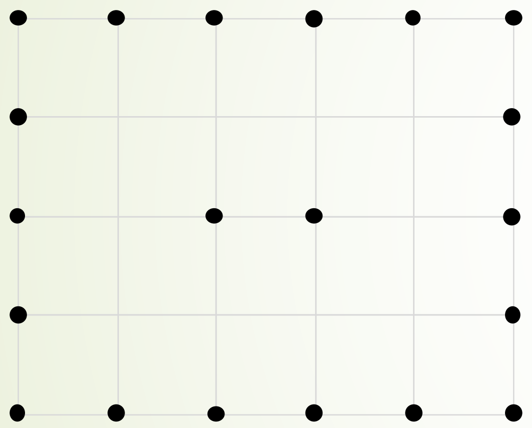
Can fail to converge to the optimum



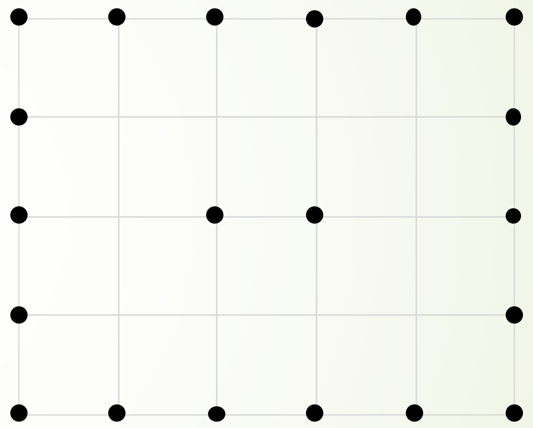
- What went wrong? ____________________
- What is the chance of failure? ____________________

Limitation of centroid-based methods

Desired



Centroid-based



- Fails because the two clusters have the same **c**_____________.
- Fail more generally when the cluster shape is **non-c**____________/**non-s**____________.
- How to resolve?

References

- 10.1 Cluster Analysis
- 10.2 Partitioning Methods