CityU

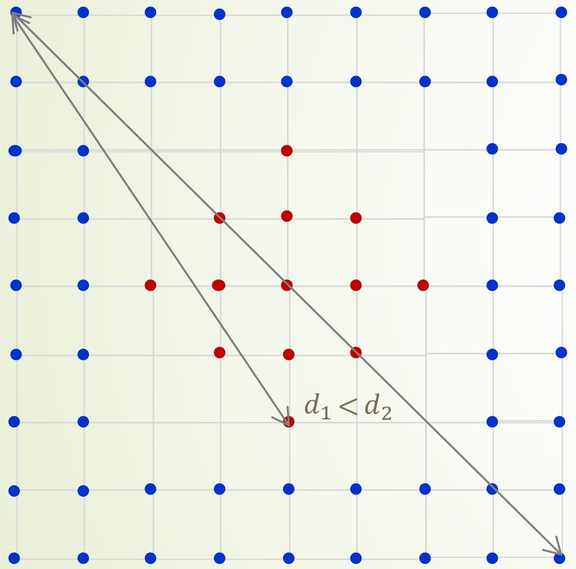
香港城市大學
City University of Hong Kong

專業 創新 胸懷全球
Professional·Creative
For The World

Cluster Analysis:
Density-Based Methods

CS5483 Data Warehousing and Data Mining

How to identify the following clusters?



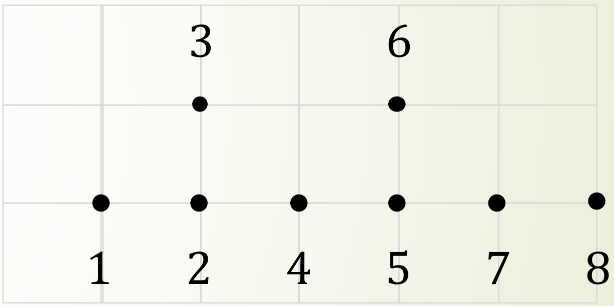
- Remedy?
- Why they should be clusters?
  They are **d**_________ regions.

- Why centroid-based method fails?
  - Both clusters have the same c___________.
  - Bias towards s___________ cluster.
- Why single-linkage fails?
  - The c______ distance of the two clusters is no larger than those between two points in the same cluster.
  - Chaining phenomenon
- Why complete-linkage method fails?
  - The f_______ distance of the two clusters is no larger than those between two points in one cluster.
  - Bias towards s__________ cluster.

Identify pillars of dense regions

- ε-neighbourhood of $\boldsymbol{p} \in D$ is the region:
  $$N_\varepsilon(\boldsymbol{p}) \coloneqq \{\boldsymbol{q} \in \mathbb{R}^d | \mathrm{dist}(\boldsymbol{p}, \boldsymbol{q}) \leq \varepsilon\}$$
  - Within r_______ $\varepsilon > 0$
  - from the c_______ $\boldsymbol{p}$.
- **C________ point**: $\boldsymbol{p} \in D$ such that
  $$|D \cap N_\varepsilon(\boldsymbol{p})| \geq \mathrm{MinPts}$$
- How to find clusters?

$\varepsilon = 1 \qquad \mathrm{MinPts} = 4$



Core points:___________

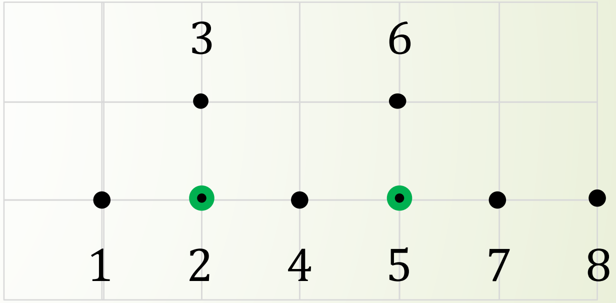
Density-reachability

- $\boldsymbol{q}$ is **directly density-reachable** from $\boldsymbol{p}$, denoted as $\boldsymbol{p} \rightarrow \boldsymbol{q}$, if
  1. $\boldsymbol{p}$ is a core point and
  2. $\boldsymbol{q} \in N_\varepsilon(\boldsymbol{p})$
- $\boldsymbol{q}$ is **density-reachable** from $\boldsymbol{p}$ if there is a **path**

$$\boldsymbol{p} \rightarrow \cdots \rightarrow \boldsymbol{q}.$$

  - All points in the path must be core points except q.
  - $\boldsymbol{q}$ is called a **border** point if it is not a core point.

$\varepsilon = 1$ MinPts $= 4$

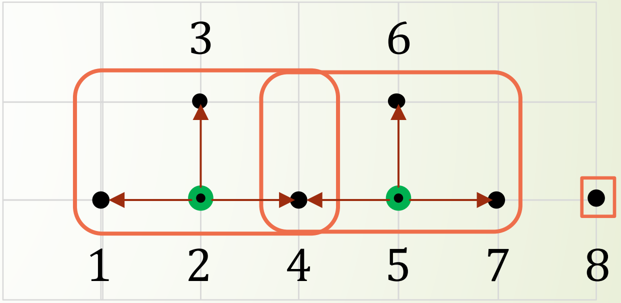


- $\boldsymbol{p}_1 \rightarrow \boldsymbol{p}_2$? ___________

Density-connectedness

- $\boldsymbol{q}$ is **density-connected** with $\boldsymbol{p}$, denoted as $\boldsymbol{p} \sim \boldsymbol{q}$, if $\boldsymbol{p}, \boldsymbol{q}$ are reachable from a c________ core point.
- Density-connected components as clusters? Y/N because ____

$\varepsilon = 1$ MinPts $= 4$



- $\boldsymbol{p}_1 \sim \boldsymbol{p}_2$? ___________

Border point: _____________

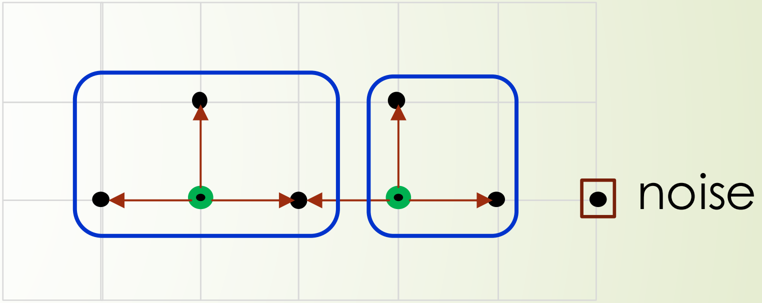
DBSCAN

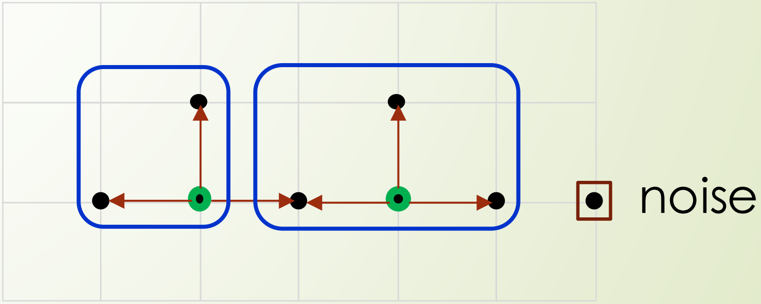
Density-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise

1. Identify core points and their density reachable points.
2. Return density-connected components of core points as clusters.
3. Assign border points to one of the cluster it is density-connected to.
4. Label the remaining points as noise.

- Clustering solution is not **u**______________.

$\varepsilon = 1 \qquad \mathrm{MinPts} = 4$

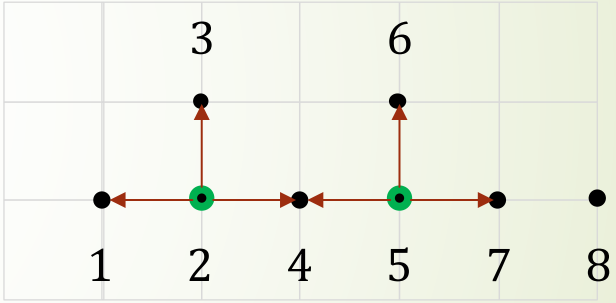




Uniqueness of clusters for core points

- DBSCAN give unique clusters for core points because density-connectedness is an equivalence relation on core points:
  - r_______________: $p \sim p$
  - s_______________: $p \sim q \Leftrightarrow q \sim p$
  - t_______________: $p \sim r, r \sim q \Rightarrow p \sim q$
- However, transitivity can fail on border points:
  - E.g., ______________________________
  - see the last slide for the mistake in [Han11].

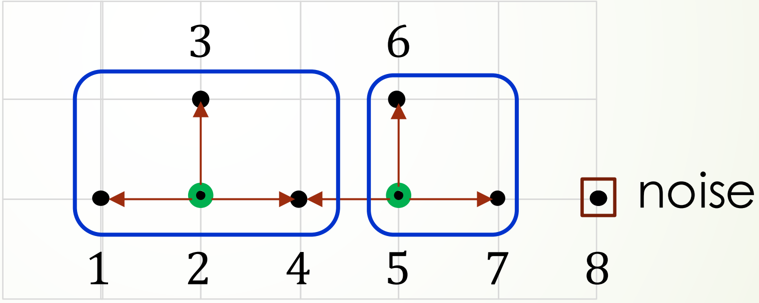
$\varepsilon = 1 \qquad \text{MinPts} = 4$



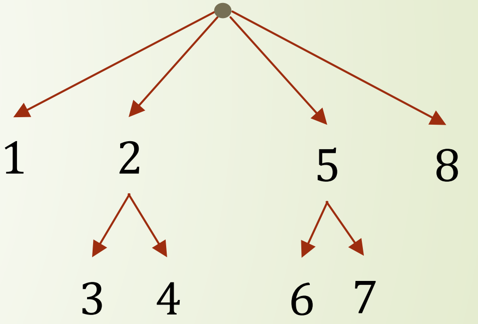
Implementation

- Repeatedly start a breadth first search (BFS) on a new point to add density-reachable points (not already assigned to another cluster) to a cluster.
- Label remaining points as noise.
- Complexity: ______________ (or $O(n \log n)$ with spatial indexing.)

$\varepsilon = 1 \qquad \text{MinPts} = 4$

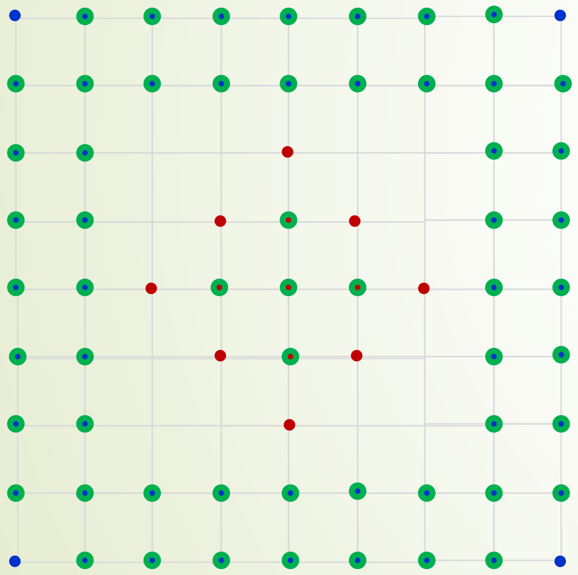


Nodes visited by BFS



Clusters can be non-convex

$\varepsilon = 1 \qquad \text{MinPts} = 4$

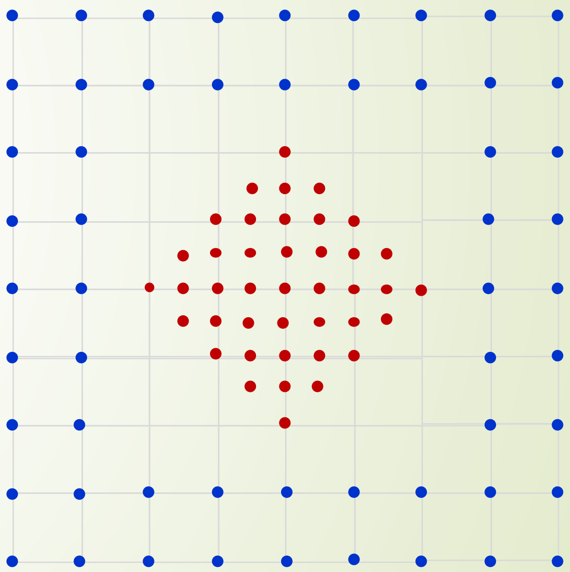


- Identify the core points, clusters, ambiguous border points and the noise.

Limitations

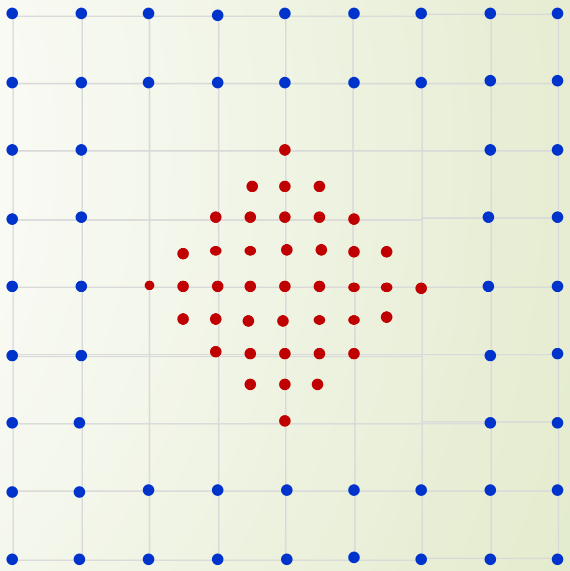
- How to choose MinPts and ε?
  - For the outer ring, want $\varepsilon \geq$___.
  - But the corner point of the inner diamond will be ____.
- What about clusters with different densities?

ε? MinPts?



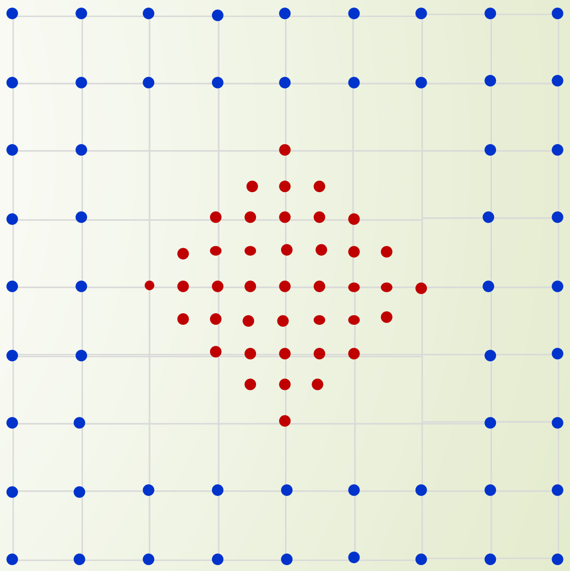
Density-based method

- Can DBSCAN recover the two clusters?
  - How to choose the parameters ε and MinPts?
- How to handle clusters with different densities?
  - Varying ε to obtain a h__________ of clusters.
  - How to do this efficiently?



OPTICS

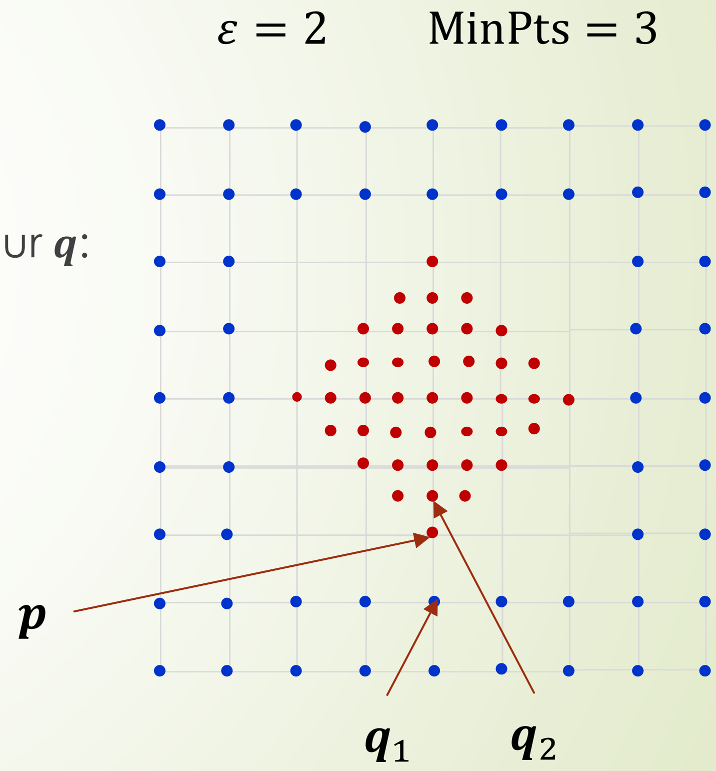
- Ordering Points to Identify the Clustering Structure.
  - Idea: Apply DBSCAN but visit c__________ points first.
- Choice of parameters:
  - MinPts normally chosen as dimension + 1. Why?
    [Optional] See Carathéodory's theorem.
  - Choose the worst-case (largest) radius ε for the neighborhood.
- For the example, we can choose
  - $\mathrm{MinPts} =$______________
  - $\varepsilon \geq$________, where all the points are density-connected.



Core and reachability distances

- When a core point $\boldsymbol{p}$ is reached,
  - Calculate the true density:
    $\text{core−distance}(\boldsymbol{p}) \coloneqq \min\{0 \leq \varepsilon' \leq \varepsilon \,|\, |D \cap N_{\varepsilon'}(\boldsymbol{p})| \geq \text{MinPts}\}$
  - Calculate the distance to its density-reachable neighbour $\boldsymbol{q}$:
    $\text{reachability−distance}(\boldsymbol{q}) \coloneqq \max\{\text{dist}(\boldsymbol{p}, \boldsymbol{q}), \text{core−distance}(\boldsymbol{p})\}$
- Visit neighbors with smaller reachability distance first.
- Why use reachability distance instead of distance?

____



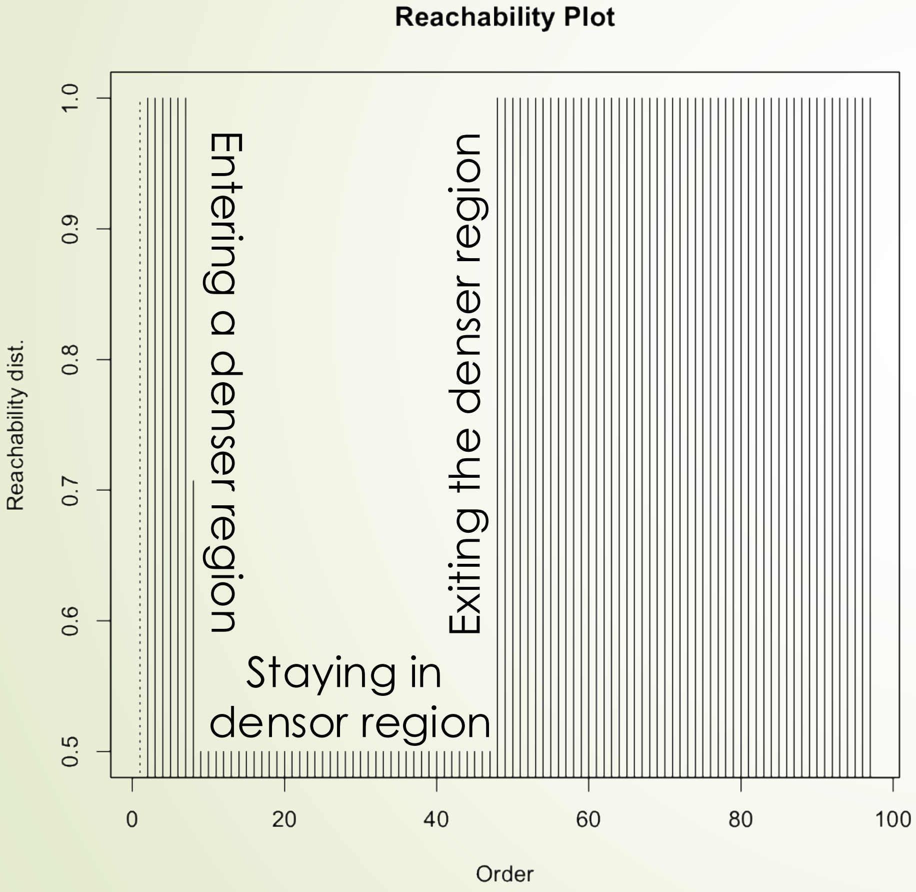
$\text{core−distance}(\boldsymbol{p}) =$ ___________

$\text{reachability−distance}(\mathbf{p}, \boldsymbol{q}_1) =$ _________

$\text{reachability−distance}(\mathbf{p}, \boldsymbol{q}_2) =$ _________

Reachability plot

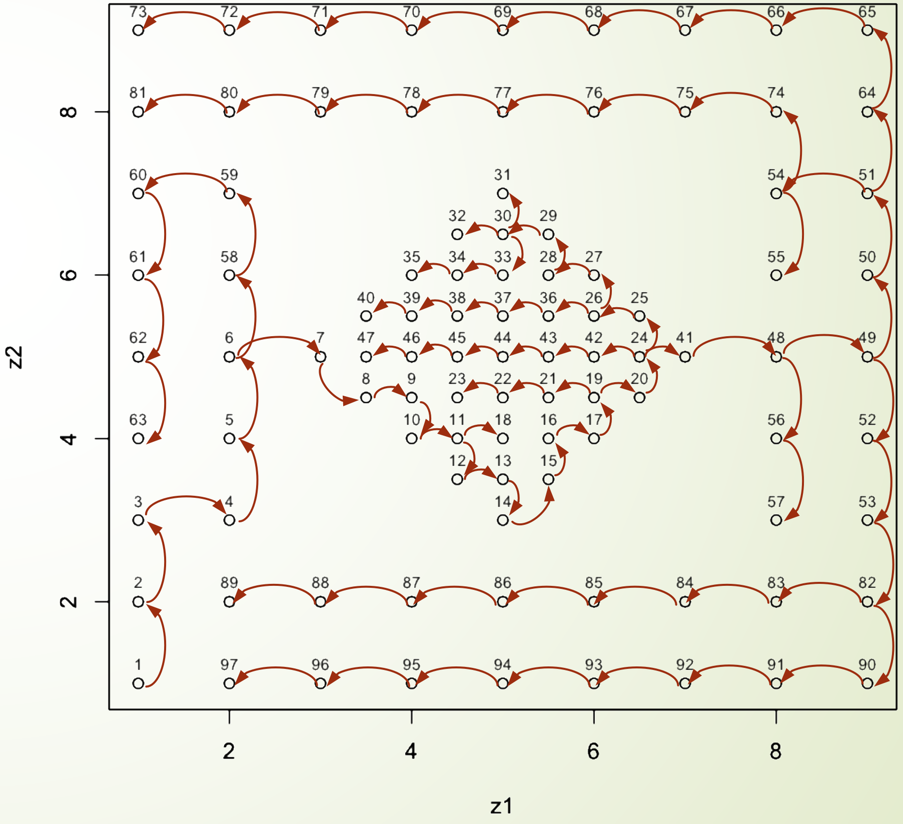
- Plot the reachability distances of the sequences of visited points.



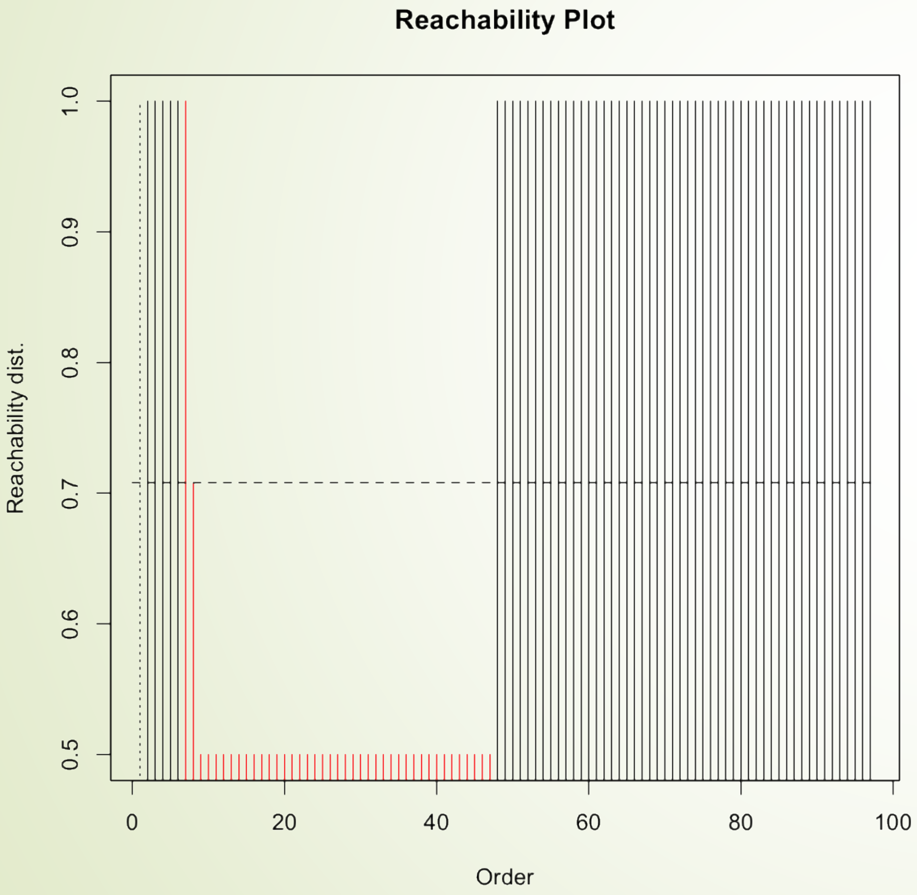
Entering a denser region

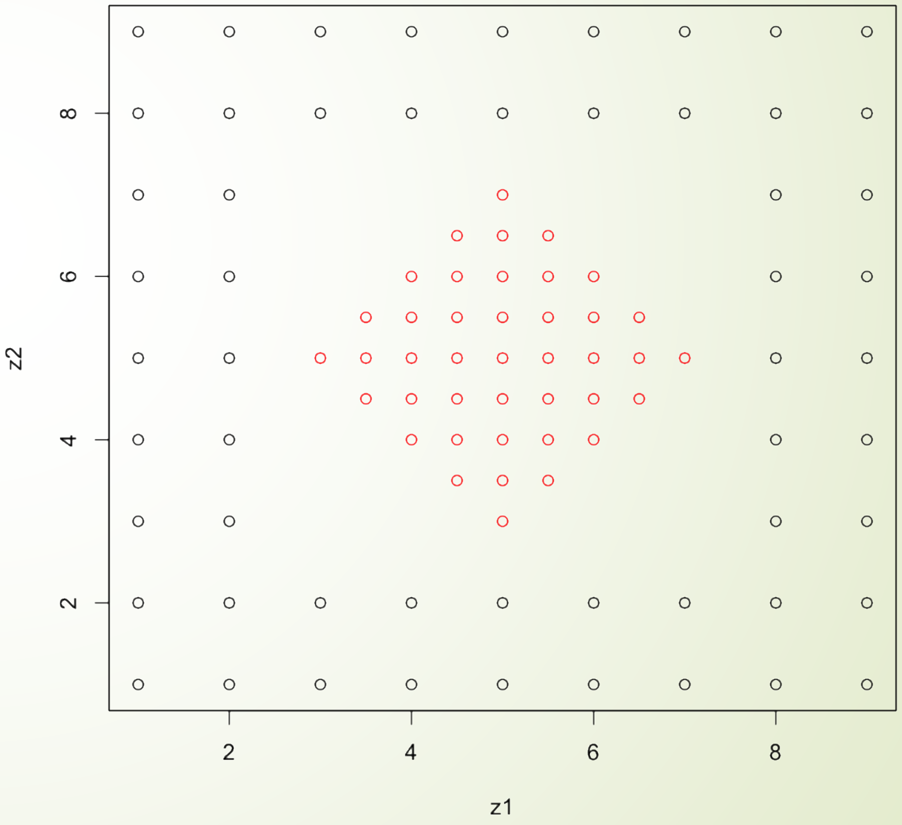
Exiting the denser region

Staying in densor region

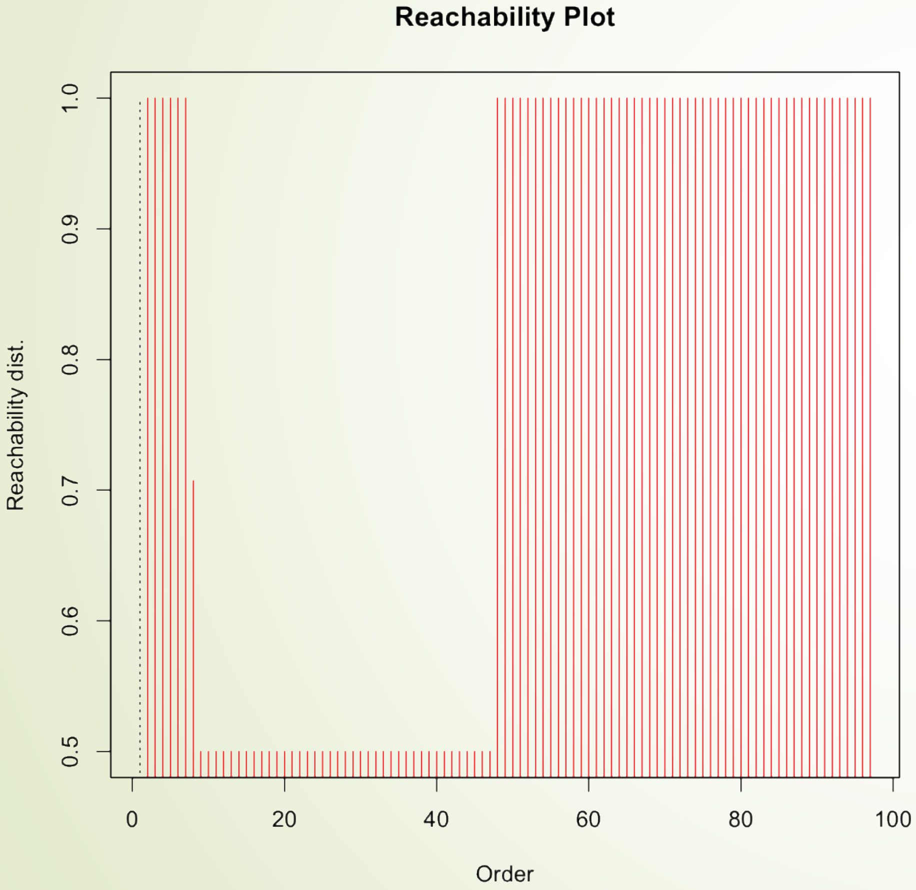


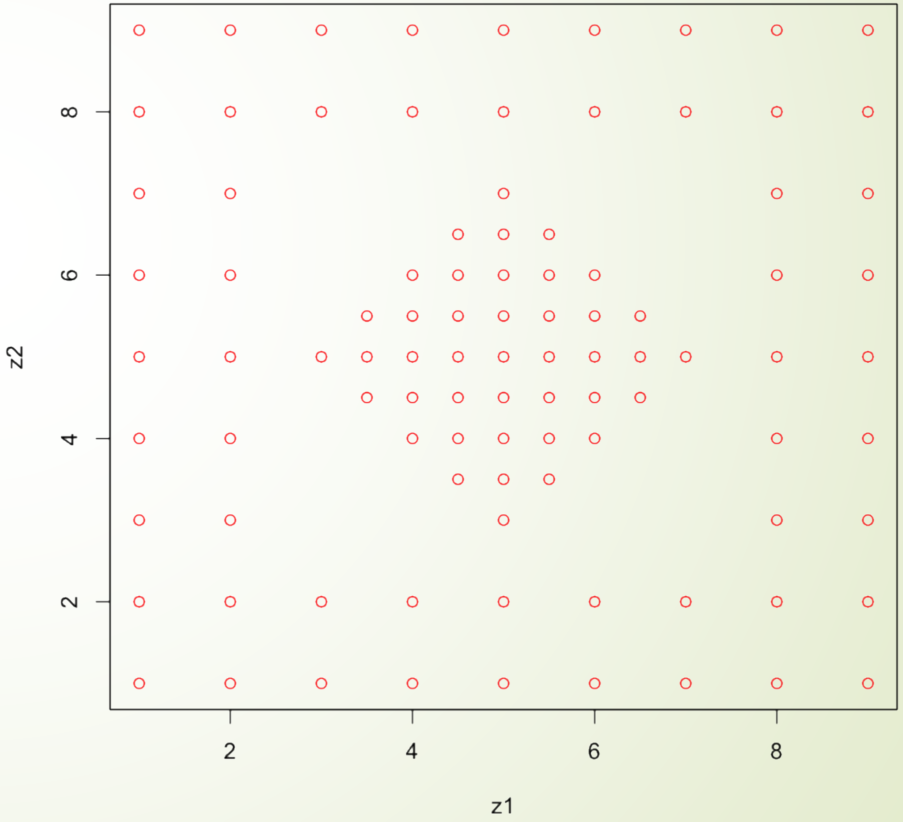
- A cluster at threshold ε is discovered as the valley of points with reachability distance below ε except for the first point.



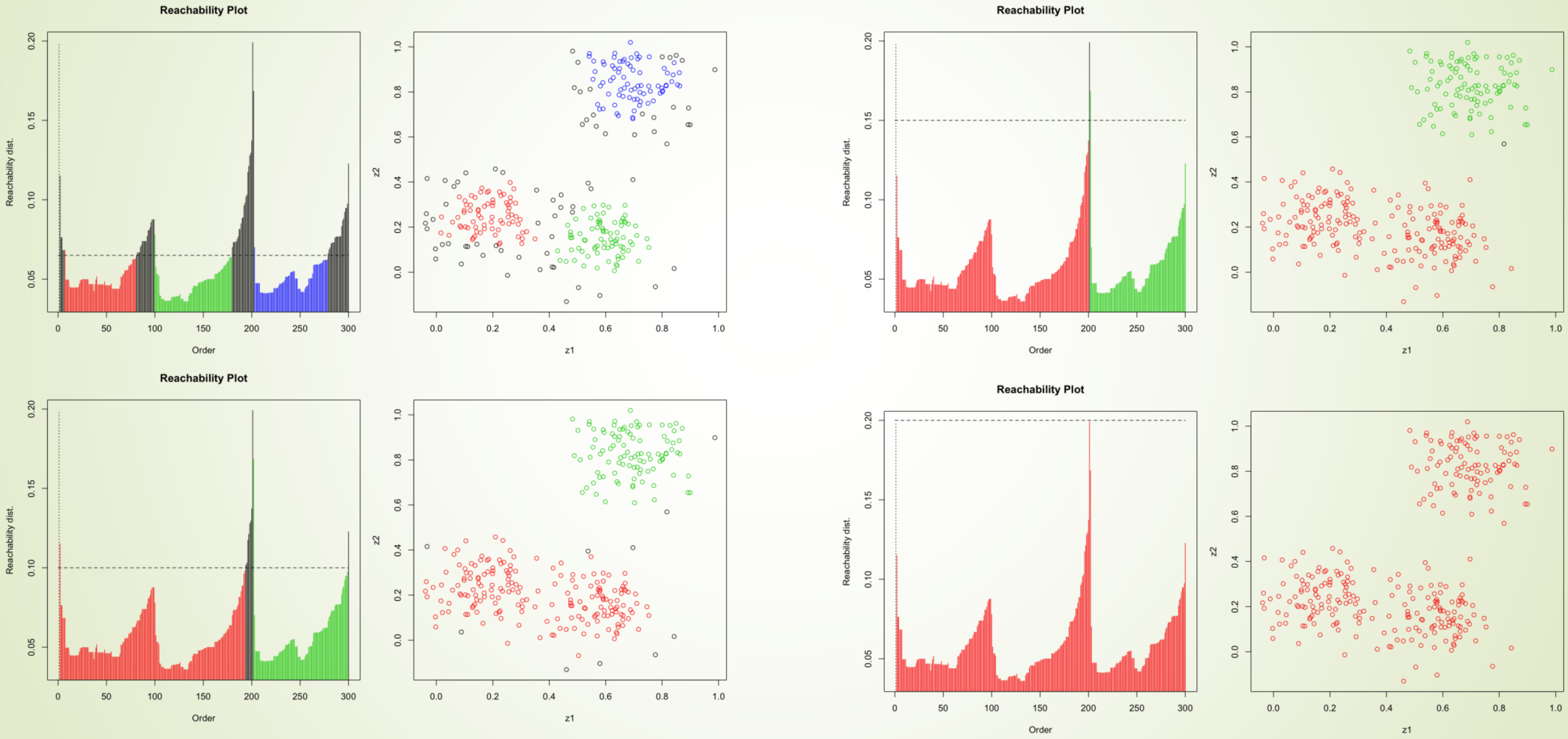


- With the threshold of reachability raised to __ or above, the valley covers all the points.
- Note that the outer ring is not identified as a separate cluster from the inner diamond.





An example with multiple clusters



References

- 10.4.1 DBSCAN: Density-Based Clustering Based on Connected Regions with High Density
    - Errata on p.472: Density-connectedness is NOT an equivalence relation unless restricted to only core points.
- McInnes, Leland, and John Healy. "Accelerated Hierarchical Density Based Clustering." *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE, 2017.