Lecture 7:

Data clustering (II)

Outline

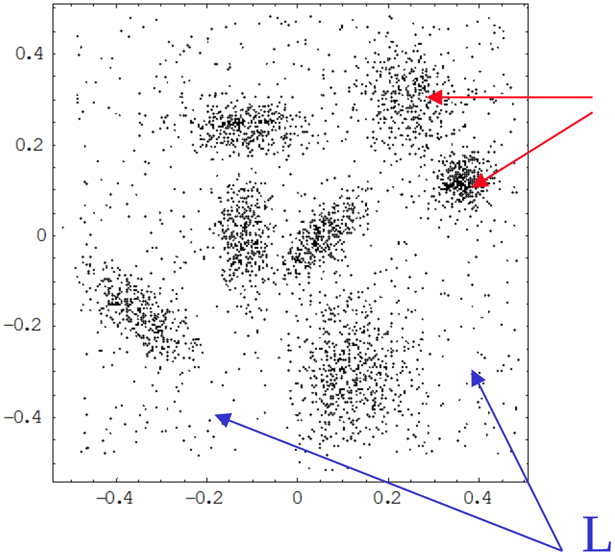
- Density based methods
  - DBSCAN
  - DENCLUE
- Probabilistic methods
  - Expectation Maximization

Density based clustering

- Clusters = dense groups of similar data separated by low density regions

- Basic idea: estimate the local density of data
  - either by determining the number of data in a given neighborhood of the analyzed point (DBSCAN)
  - or by using some influence functions (DENCLUE)

- Main aspects:
  - How is density estimated?
  - How is connectivity defined?
  - What data structures should be used?



High density

Low density

DBSCAN

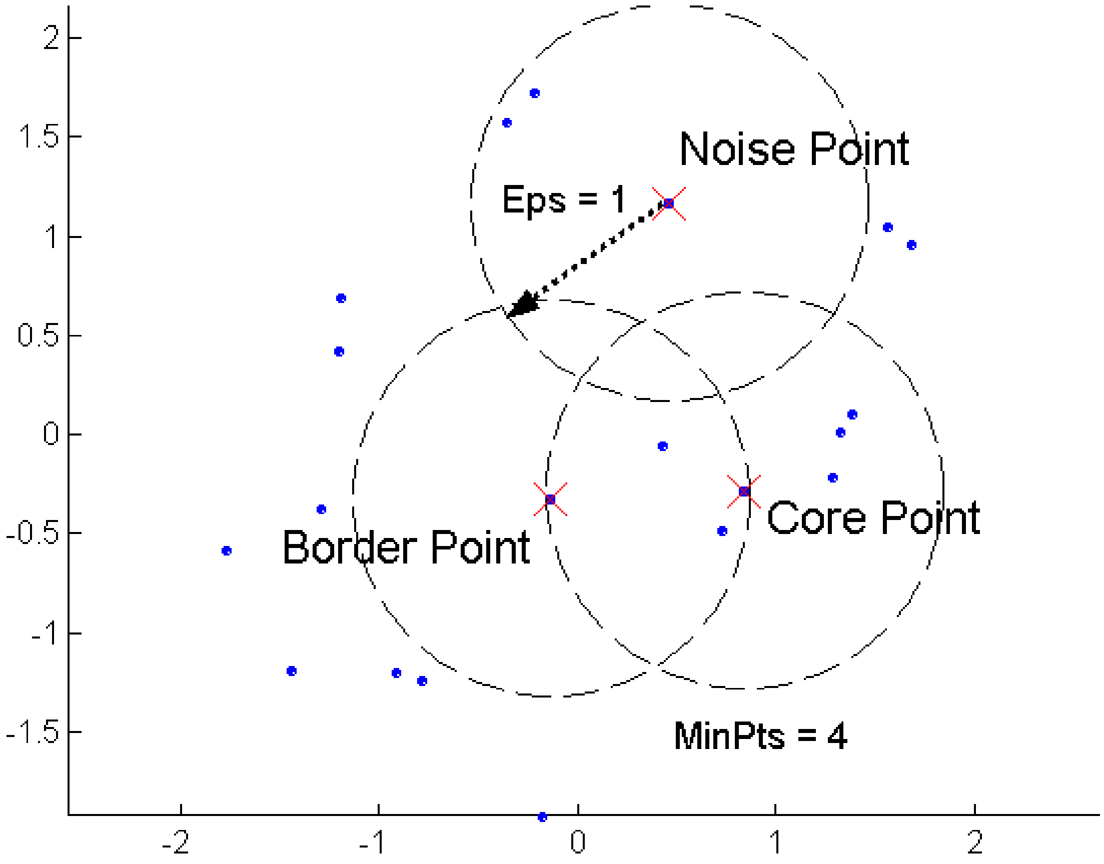
DBSCAN is a density-based algorithm

- Density measured at a point= number of points within a neighborhood of specified radius (Eps)

- A point is a core point if it has more than a specified number of points (MinPts) within Eps; these are points that are in the interior of a cluster

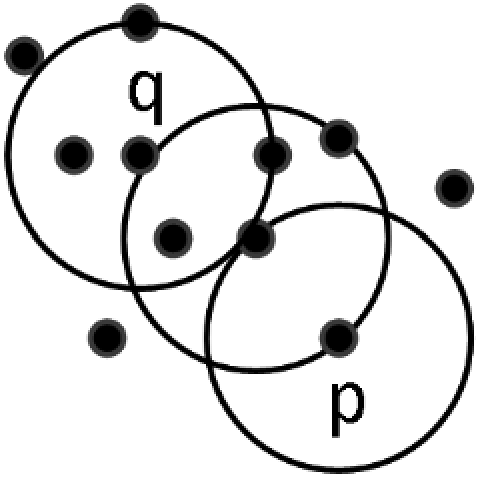
- A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point; Two points are connected if they are one in the neighborhood of the other
- A point q is directly density reachable from a core point p if it is in the neighborhood of p; density reachability is defined as transitive closure of direct density reachability (there is a chain of core points s.t. one point is directly reachable from the previous one)
- A noise point is any point that is not a core point or a border point.

DBSCAN

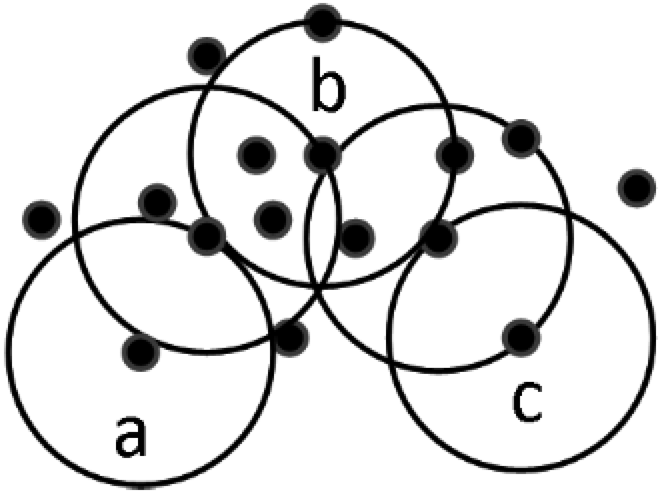


Noise Point
Eps = 1
Border Point
Core Point
MinPts = 4
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
0
1
2

DBSCAN



p is density reachable from q

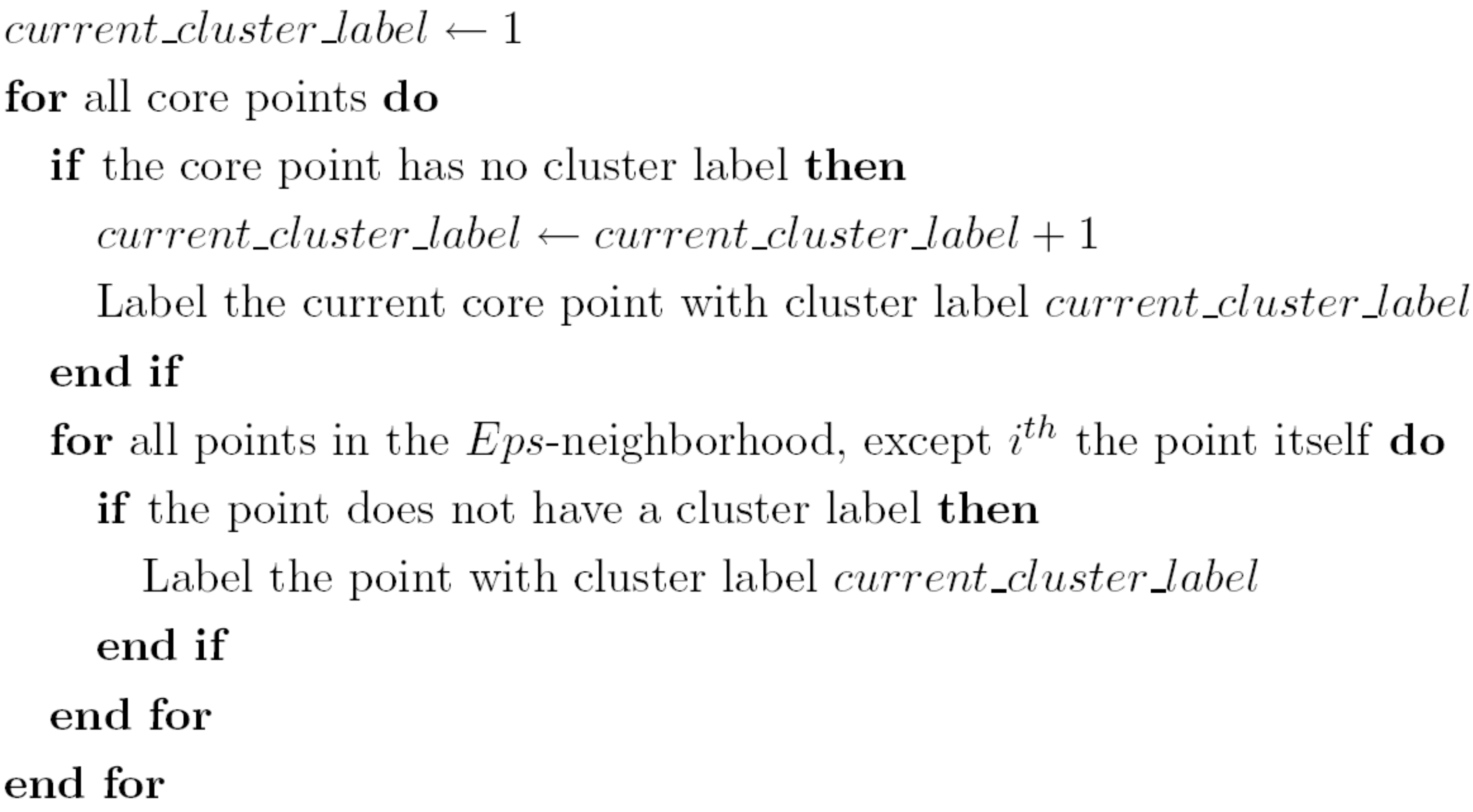


a is density reachable from b
c is density reachable from b
=> a and c are density-connected

Remark:

- Two points, a and b, are density-connected if there exist a third point, c, such that c is reachable both from a and from b
- Two density-connected points belong to the same cluster =>
  a density based cluster is a maximal set of density-connected data

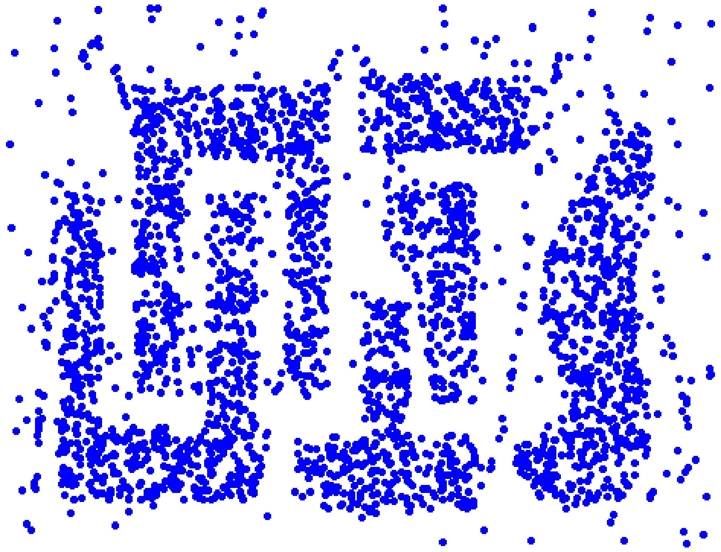
DBSCAN

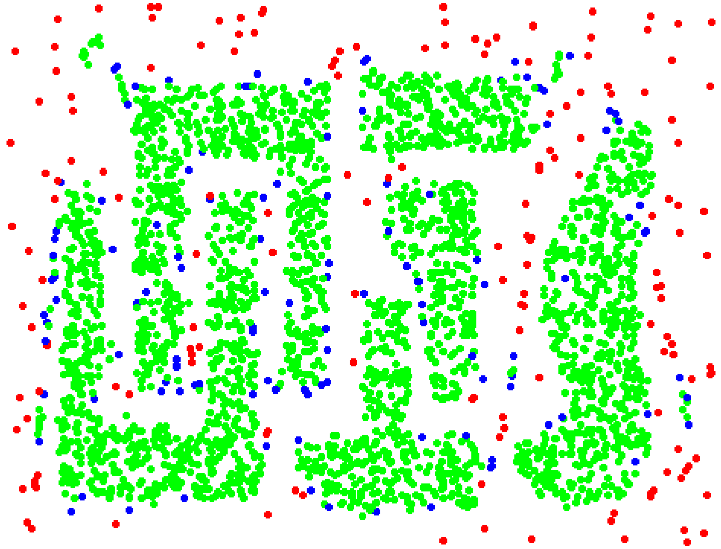


```
$current\_cluster\_label \leftarrow 1$
for all core points do
    if the core point has no cluster label then
        $current\_cluster\_label \leftarrow current\_cluster\_label + 1$
        Label the current core point with cluster label $current\_cluster\_label$
    end if
    for all points in the $Eps$-neighborhood, except $i^{th}$ the point itself do
        if the point does not have a cluster label then
            Label the point with cluster label $current\_cluster\_label$
        end if
    end for
end for
```

[images from slides by Kumar, 2004]

DBSCAN



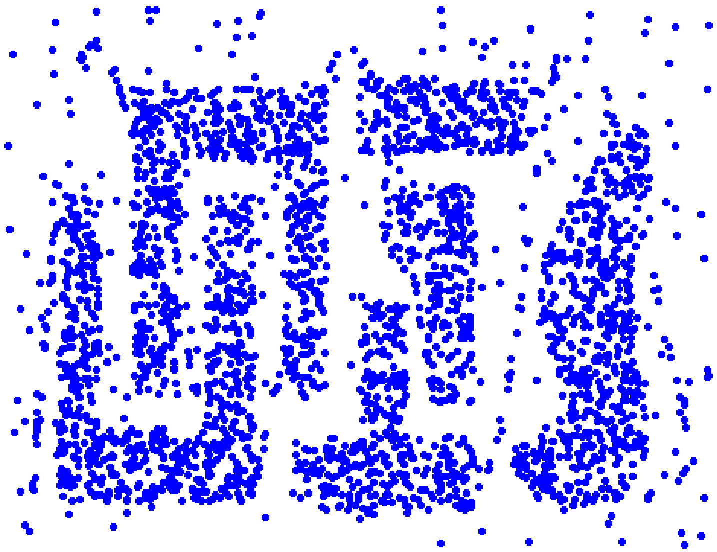


Original Points

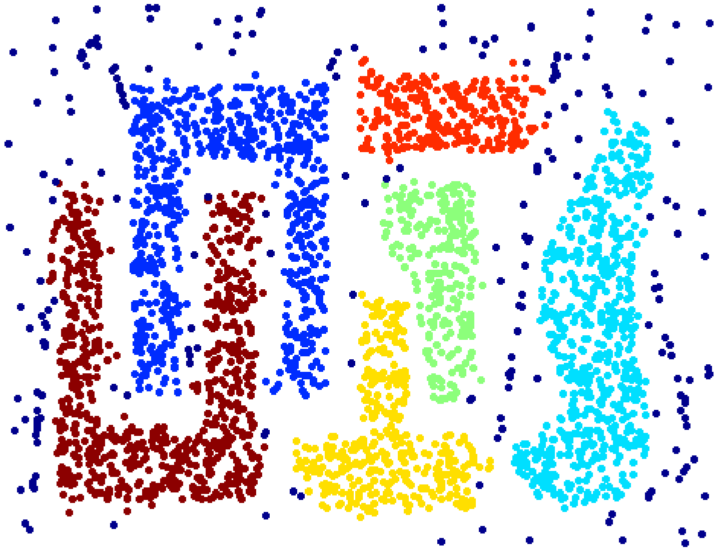
Point types: core, border and noise

Eps = 10, MinPts = 4

DBSCAN



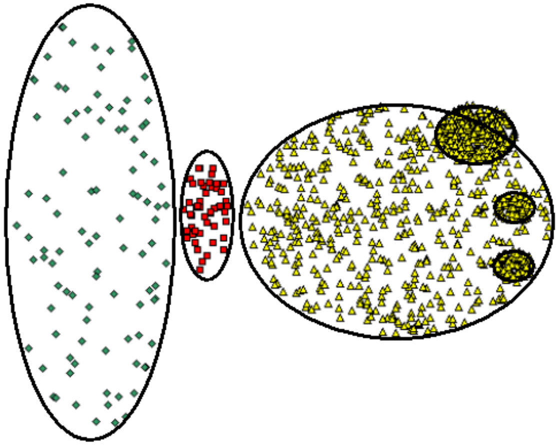
Original Points



Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

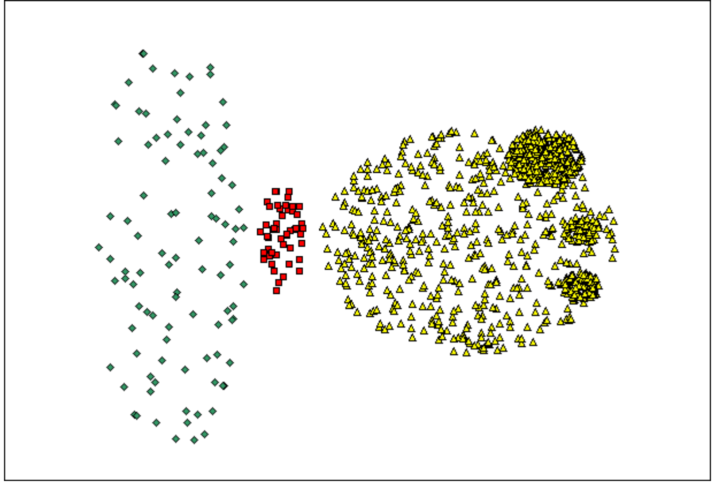
DBSCAN



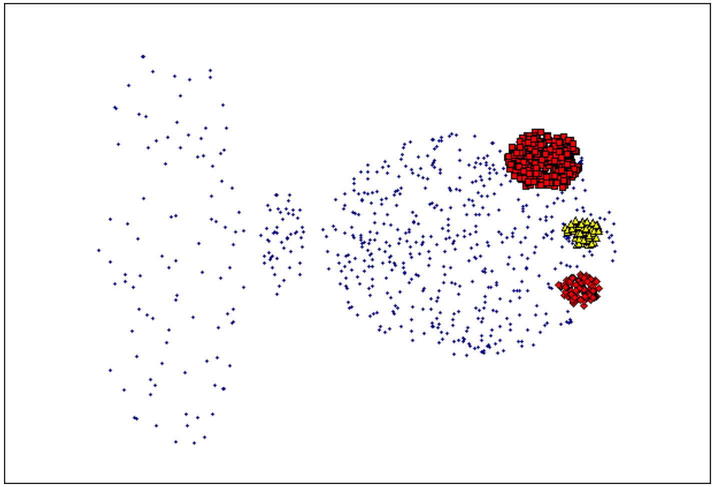
Original Points

It does not work well:

- Varying densities
- High-dimensional data

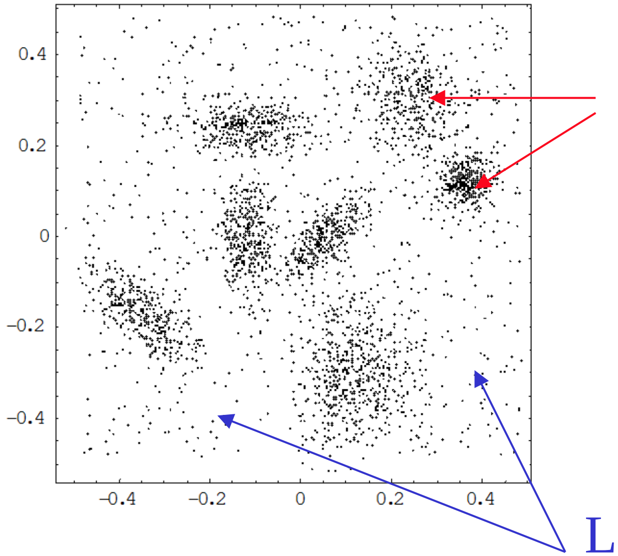


(MinPts=4, Eps=9.75).



DENCLUE

- Clusters = dense groups of similar data separated by low density regions
- Basic idea: estimate the local density of data
  - either by determining the number of data in a given neighborhood of the analyzed point (DBSCAN)
  - or by using some influence functions (DENCLUE)



High density

Low density

Influence function

$$I_y(x) = \frac{1}{\sigma^{n/2}} \exp\left(- \frac{\sum_{j=1}^{n} (x_j - y_j)^2}{2\sigma^2} \right)$$

Density function

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} I_{x_i}(x)$$

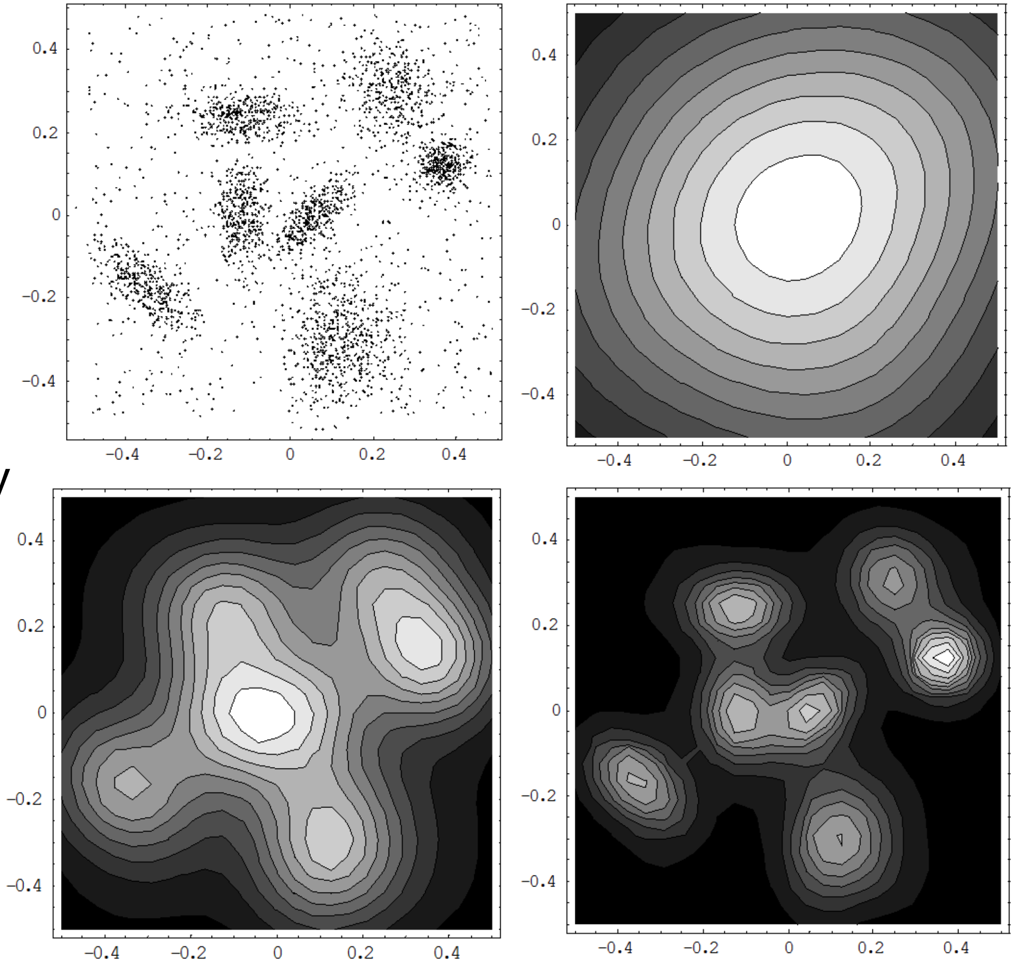
DENCLUE

- The landscape of the density function is highly dependent on the value of parameter σ

- For appropriate values of σ, the local maxima of the density function correspond to clusters

- For large values of σ the density function landscape has a single maximum

- For small values of σ the local maxima are steep, thus difficult to find



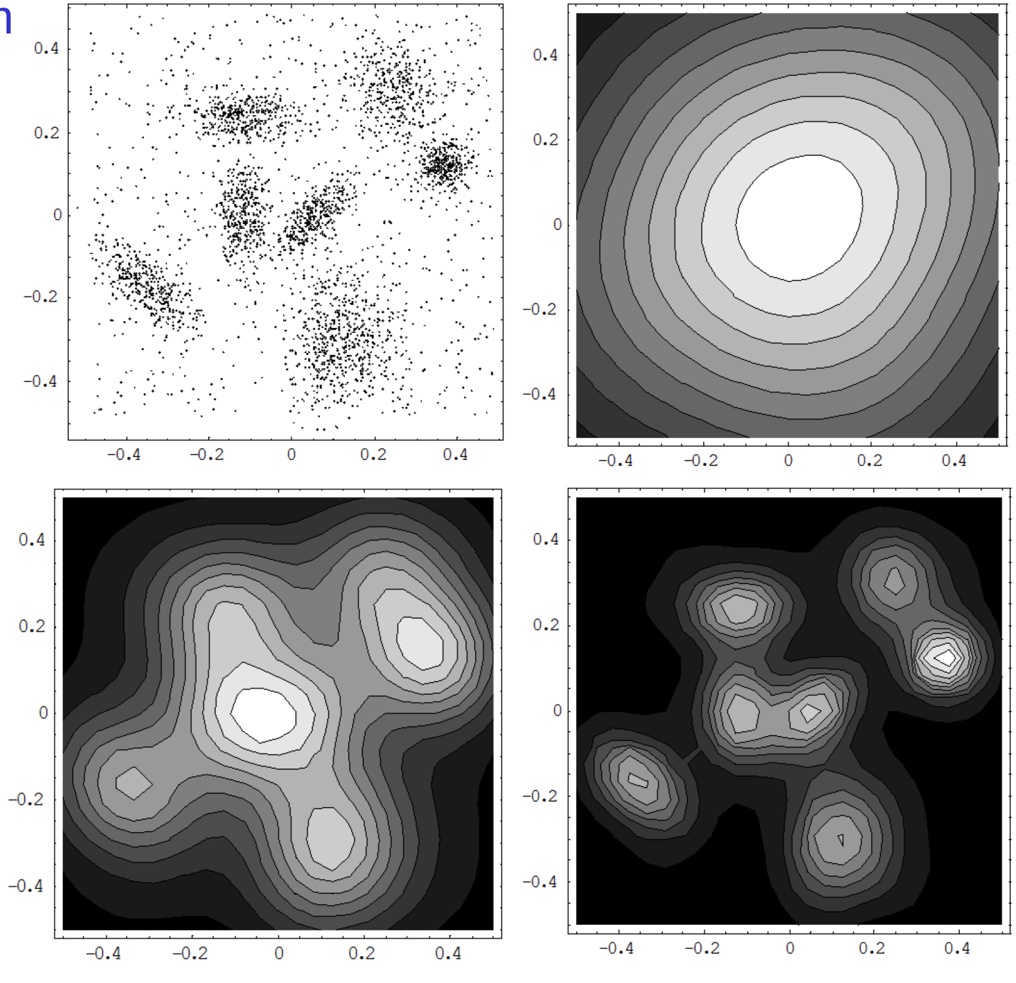
σ=0.1

σ=0.05

DENCLUE

Idea of DENCLUE [Hinneburg, Keim – 1998]: apply a gradient search to find local maxima starting from data points

- identify center-defined clusters (one cluster correspond to one local maxima)
- identify arbitrary-shaped clusters (one cluster corresponds to a set of “connected” local maxima)



σ=0.1

σ=0.05

Probabilistic methods

Main idea:

- The data are generated by a stochastic process (a mixture of probability distributions, each one being in correspondence with a cluster)
- The aim of the clustering algorithm is to discover the probabilistic model, i.e. identify the probability distributions

Example:

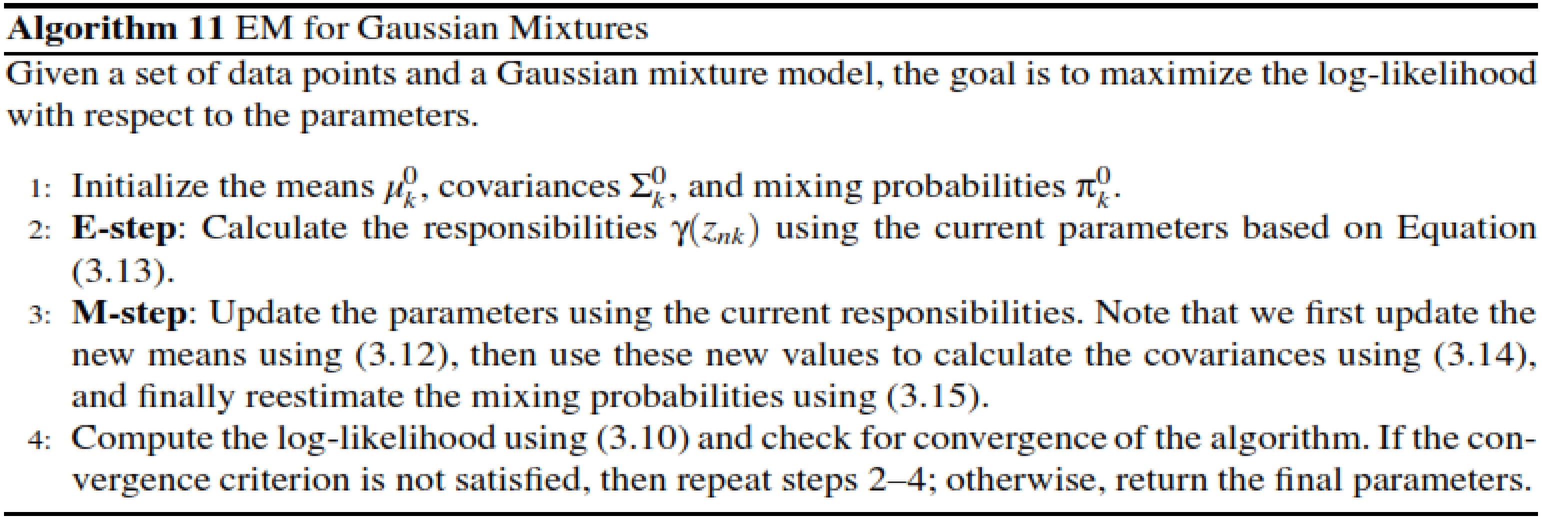
- Expectation–Maximization (EM) algorithm; it is based on the following assumptions:
  - each data has been generated by a probability distribution
  - in the generative process, the probability distribution corresponding to each data is selected according to a selection probability

EM algorithm

- Input: data set $D=\{x_1,x_2,\ldots,x_N\}$, K = number of clusters
- Output: a partition $P=\{C_1,C_2,\ldots,C_K\}$ of D

- (E-Step) Determine the expected probability of assignment of data points to clusters with the use of current model parameters.
- (M-Step) Determine the optimum model parameters of each mixture by using the assignment probabilities as weights.

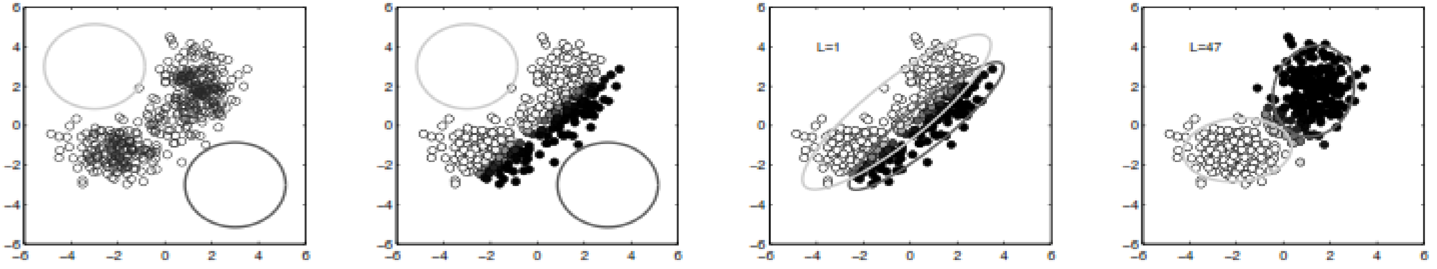
EM algorithm



Algorithm 11 EM for Gaussian Mixtures

Given a set of data points and a Gaussian mixture model, the goal is to maximize the log-likelihood with respect to the parameters.

```
1: Initialize the means $\mu_k^0$, covariances $\Sigma_k^0$, and mixing probabilities $\pi_k^0$.
2: E-step: Calculate the responsibilities $\gamma(z_{nk})$ using the current parameters based on Equation (3.13).
3: M-step: Update the parameters using the current responsibilities. Note that we first update the new means using (3.12), then use these new values to calculate the covariances using (3.14), and finally reestimate the mixing probabilities using (3.15).
4: Compute the log-likelihood using (3.10) and check for convergence of the algorithm. If the convergence criterion is not satisfied, then repeat steps 2–4; otherwise, return the final parameters.
```



(a) (b) (c) (d)

Data Clustering Algorithms and applications (ed. CC Aggarwal, CK Reddy, 2014)

EM algorithm

Equations involved in the EM algorithm

$$p(\mathbf{x}_n|\Theta) = p(\mathbf{x}_n|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k). \qquad (3.10)$$

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\}$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \qquad (3.12)$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)}. \qquad (3.13)$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}. \qquad (3.14)$$

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} \qquad (3.15)$$

Data Clustering Algorithms and applications (ed. CC Aggarwal, CK Reddy, 2014)