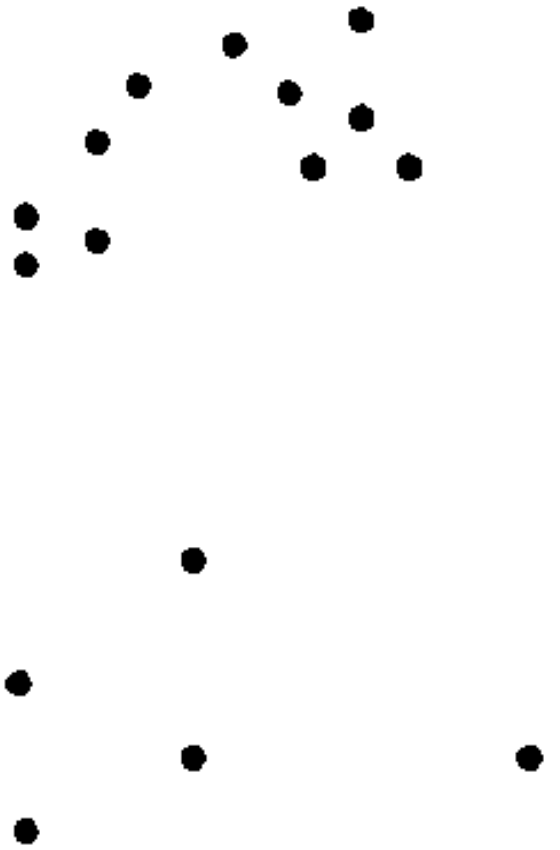
STATS 415: K-means Clustering

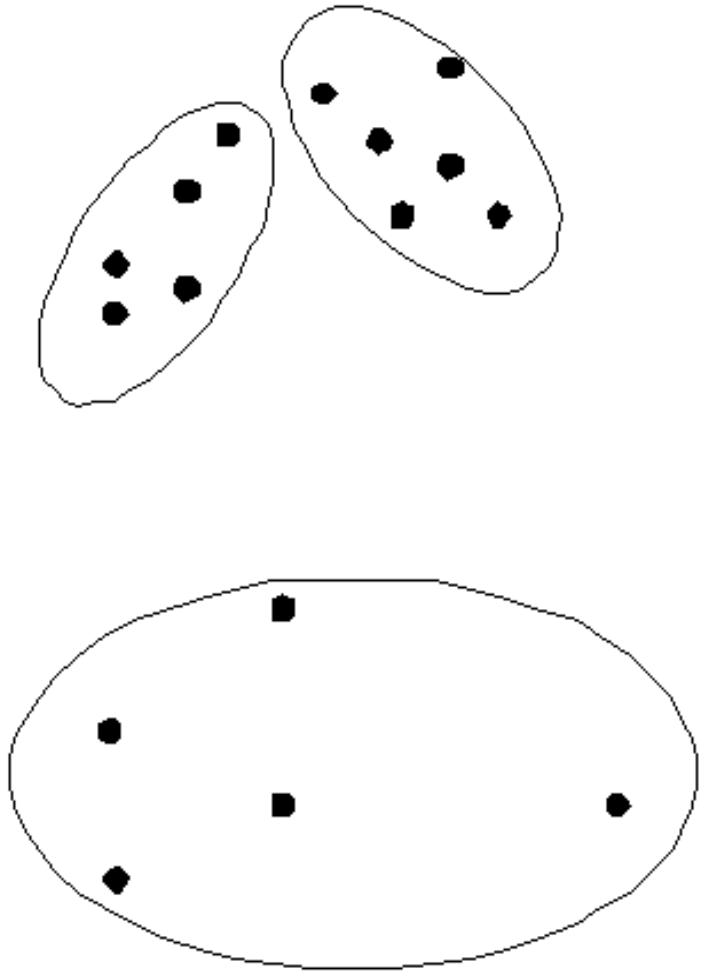
Prof. Liza Levina

Department of Statistics, University of Michigan

Partitional Clustering



Original Points



A Partitional Clustering

K-means Clustering

- Partitional clustering approach
- Number of clusters K must be specified in advance
- Each cluster is associated with a centroid (center point).
- Each point is assigned to the cluster with the closest centroid.

The K-means algorithm

- Initialize: Select K initial centroids
- Repeat:
  1. For each point, compute distance to all K centroids
  2. Assign each point to the cluster associated with the closest centroid
  3. Recompute the centroid of each cluster
- Until the centroids do not change

K-means: details

- Initial centroids are often chosen randomly, and then clusters will vary from one run to another.
- "Closeness" is most commonly measured by Euclidean distance, or another standard dissimilarity.
- The centroid is typically the mean of the points in the cluster, but there is a K-medians variant (PAM).
- K-means can be proven to converge for common similarity measures.
- Biggest improvements occur in the first few iterations

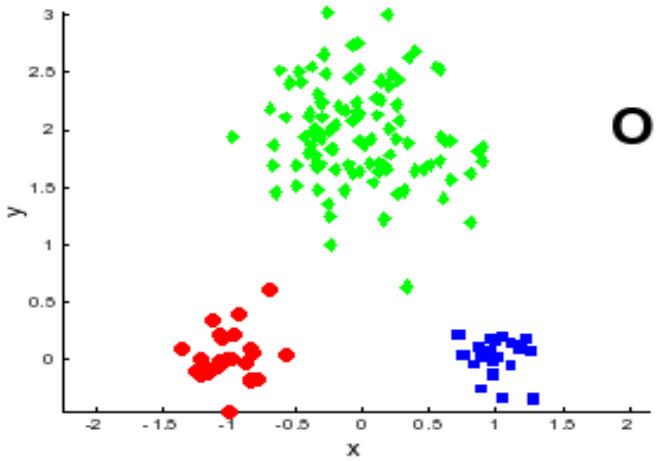
Evaluating K-means clusters

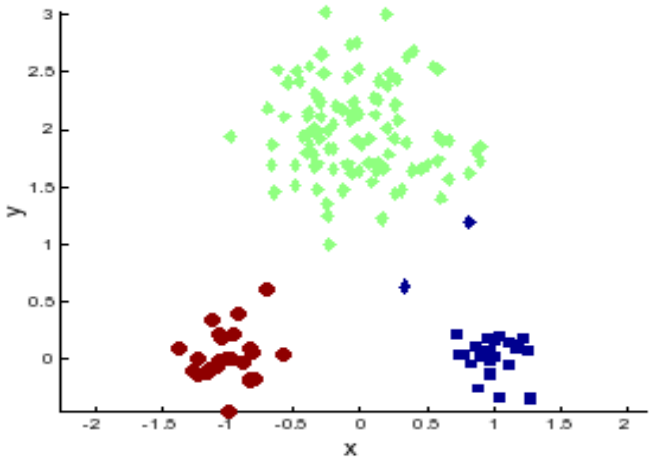
- Most common measure is the sum of squared errors (SSE).

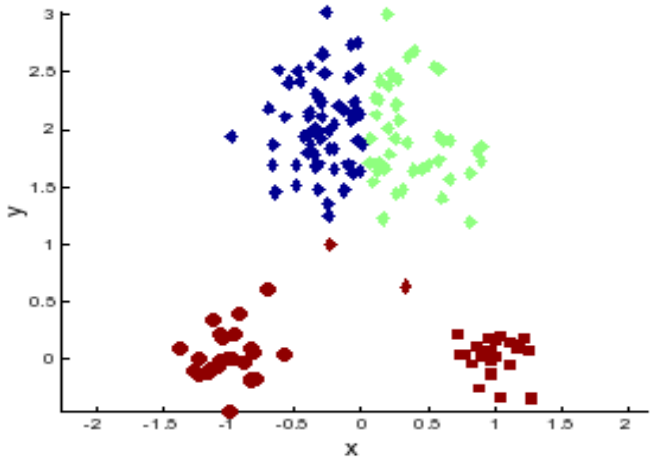
$$\text{SSE} = \sum_{k=1}^{K} \sum_{x \in C_k} \text{dist}^2(m_k, x)$$

  where x is a data point in cluster C_k and m_k is the centroid for cluster C_k.
- Comparing two partitions into clusters, we prefer the one with the smaller SSE.

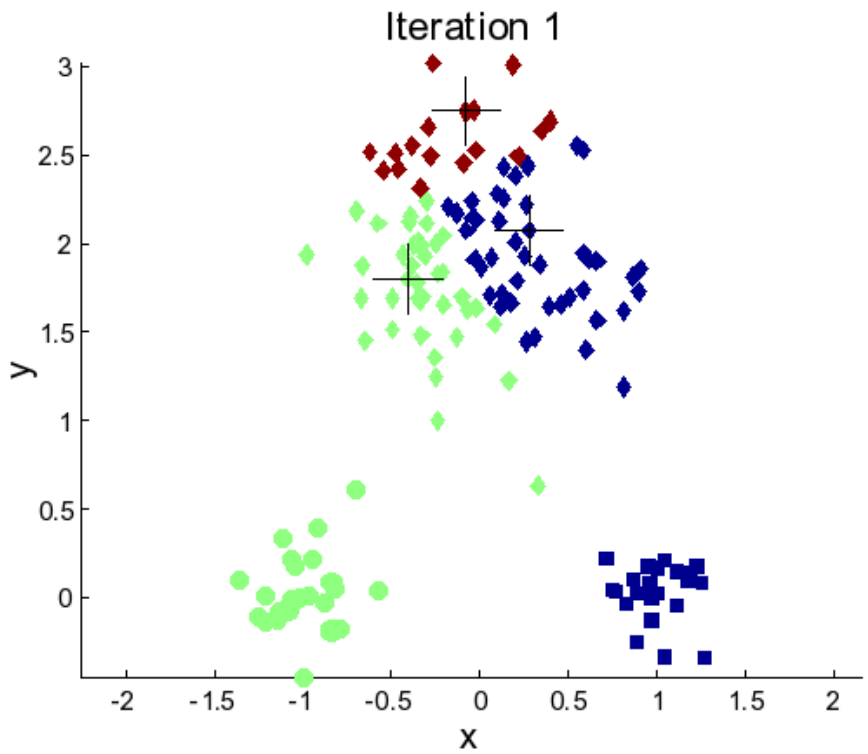
Two different K-means results on the same data

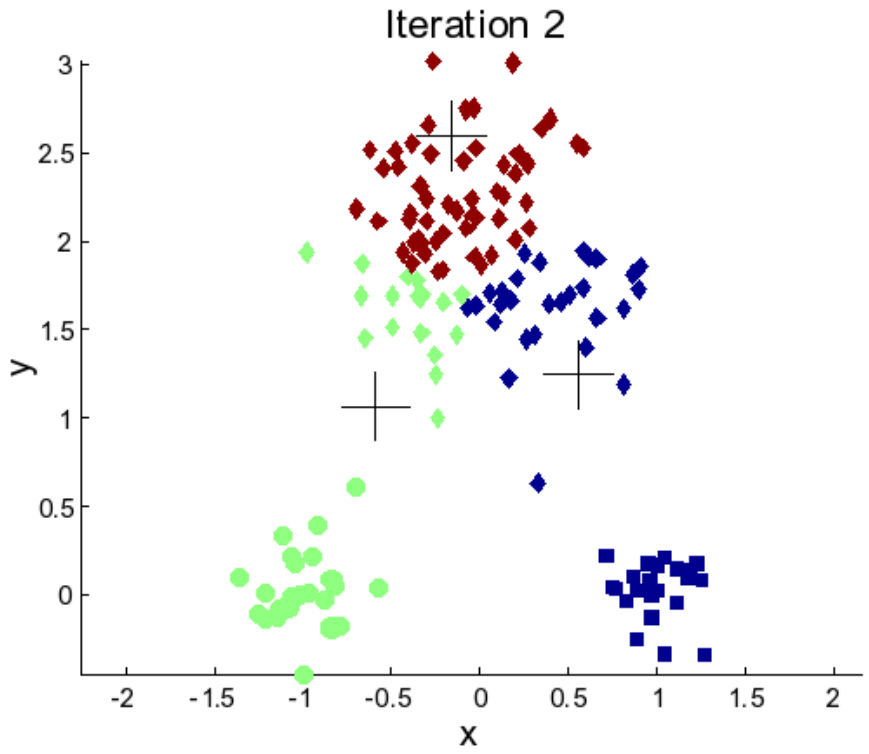


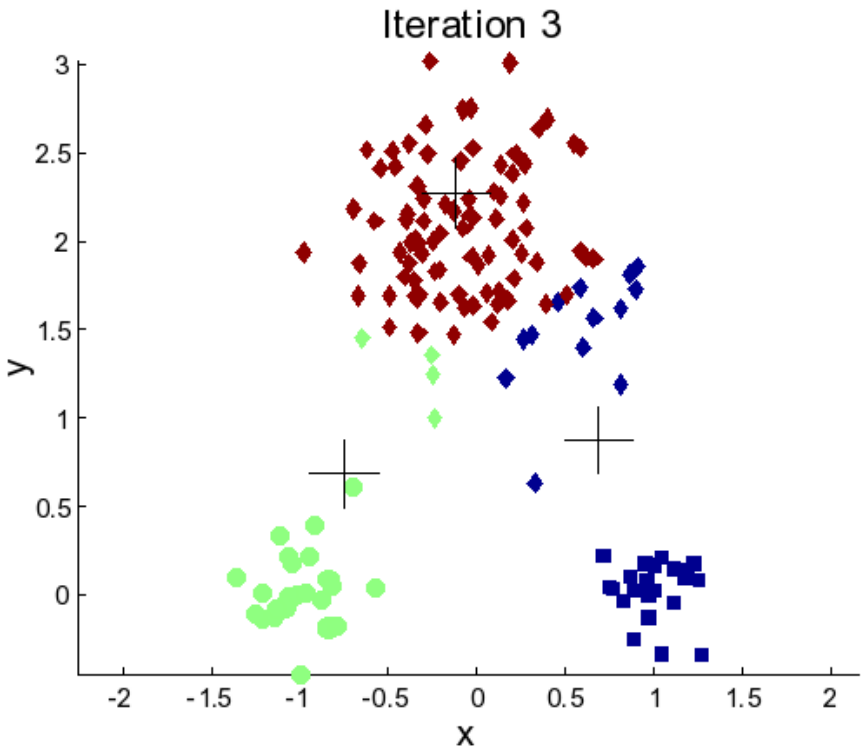


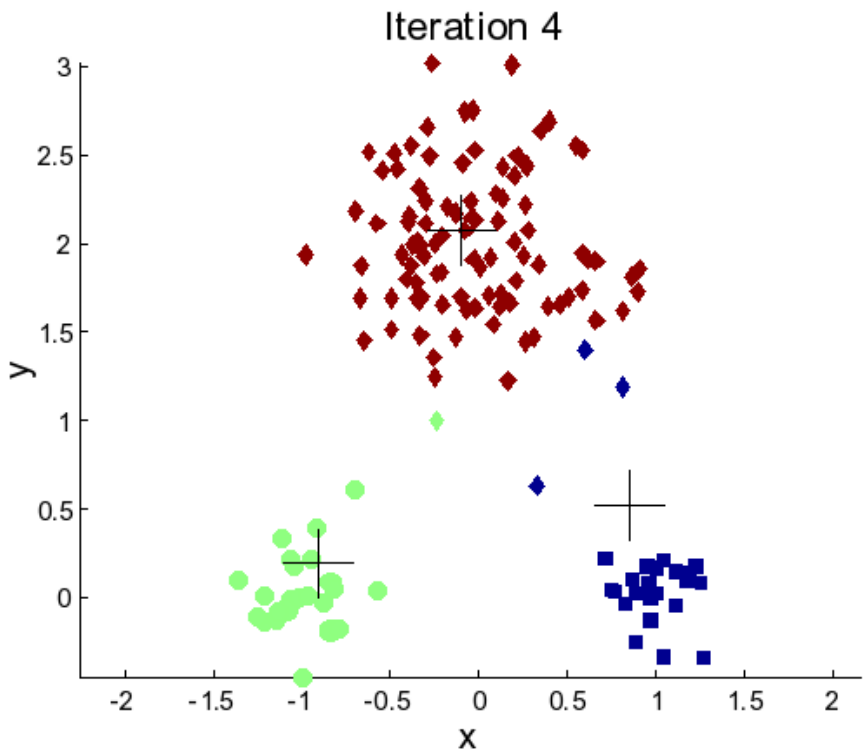


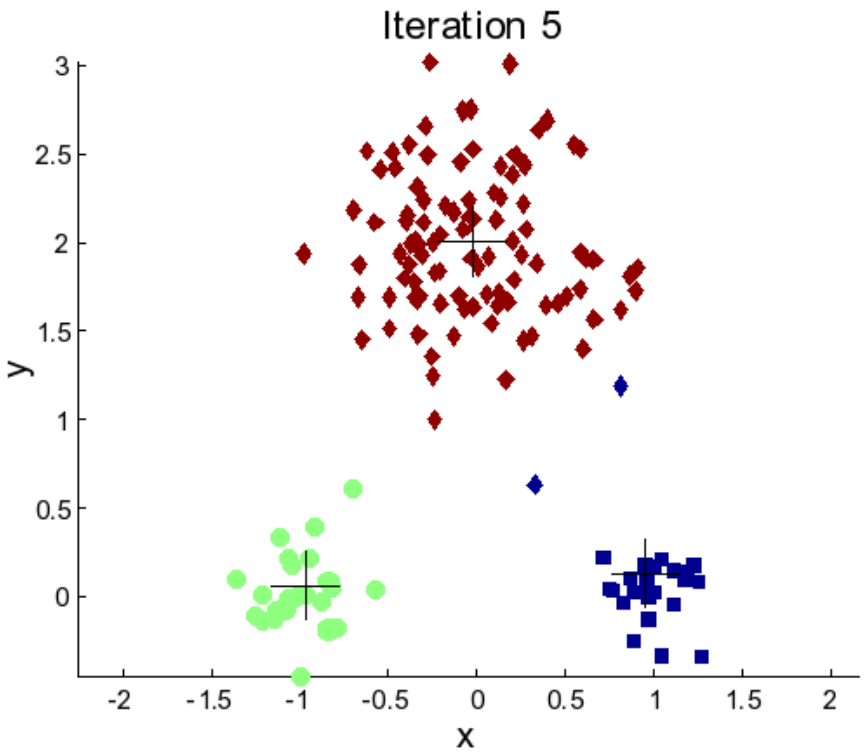
K-means in action

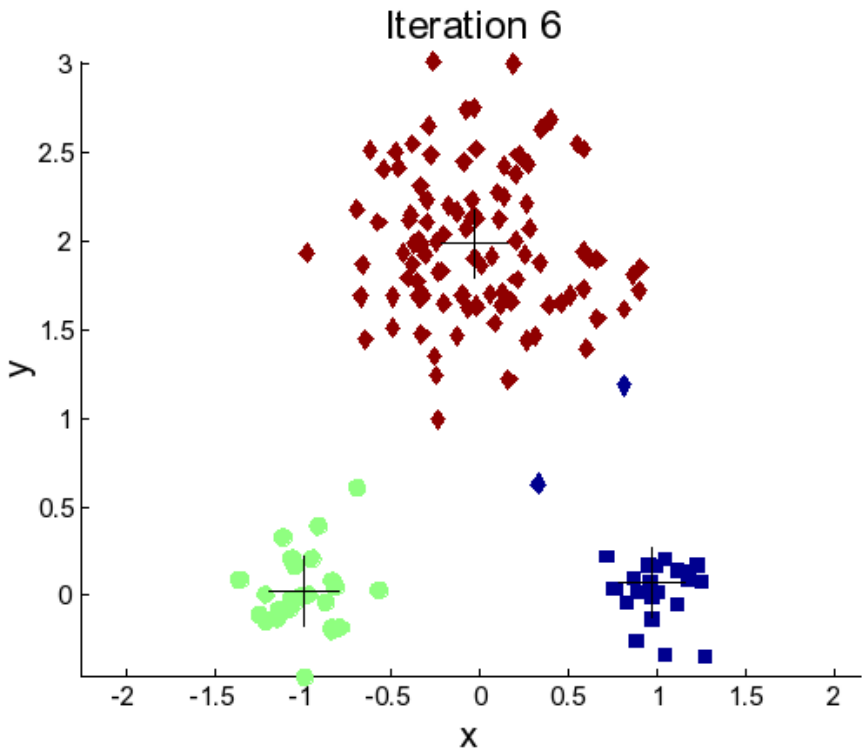


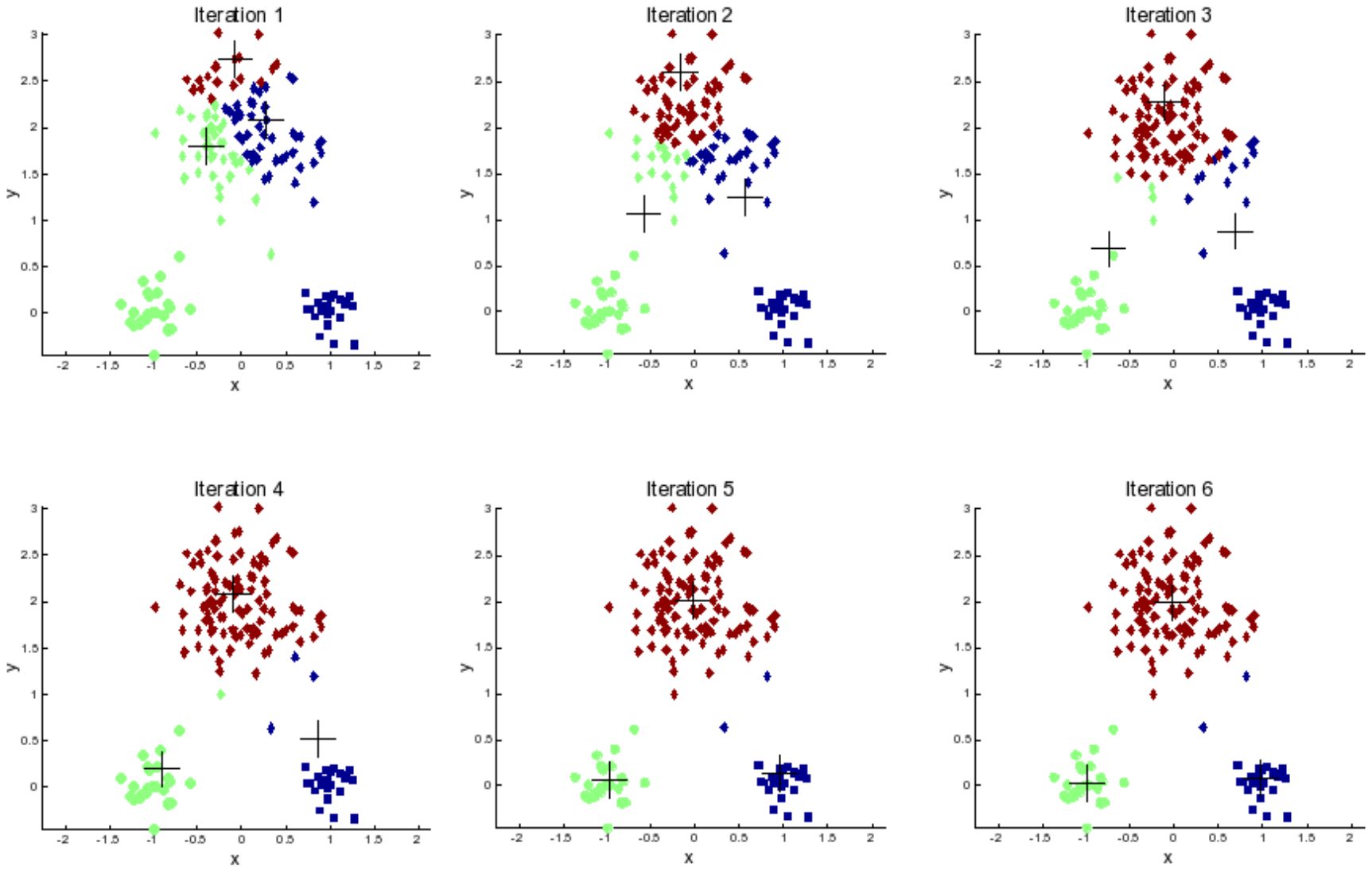



Iteration 3
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x

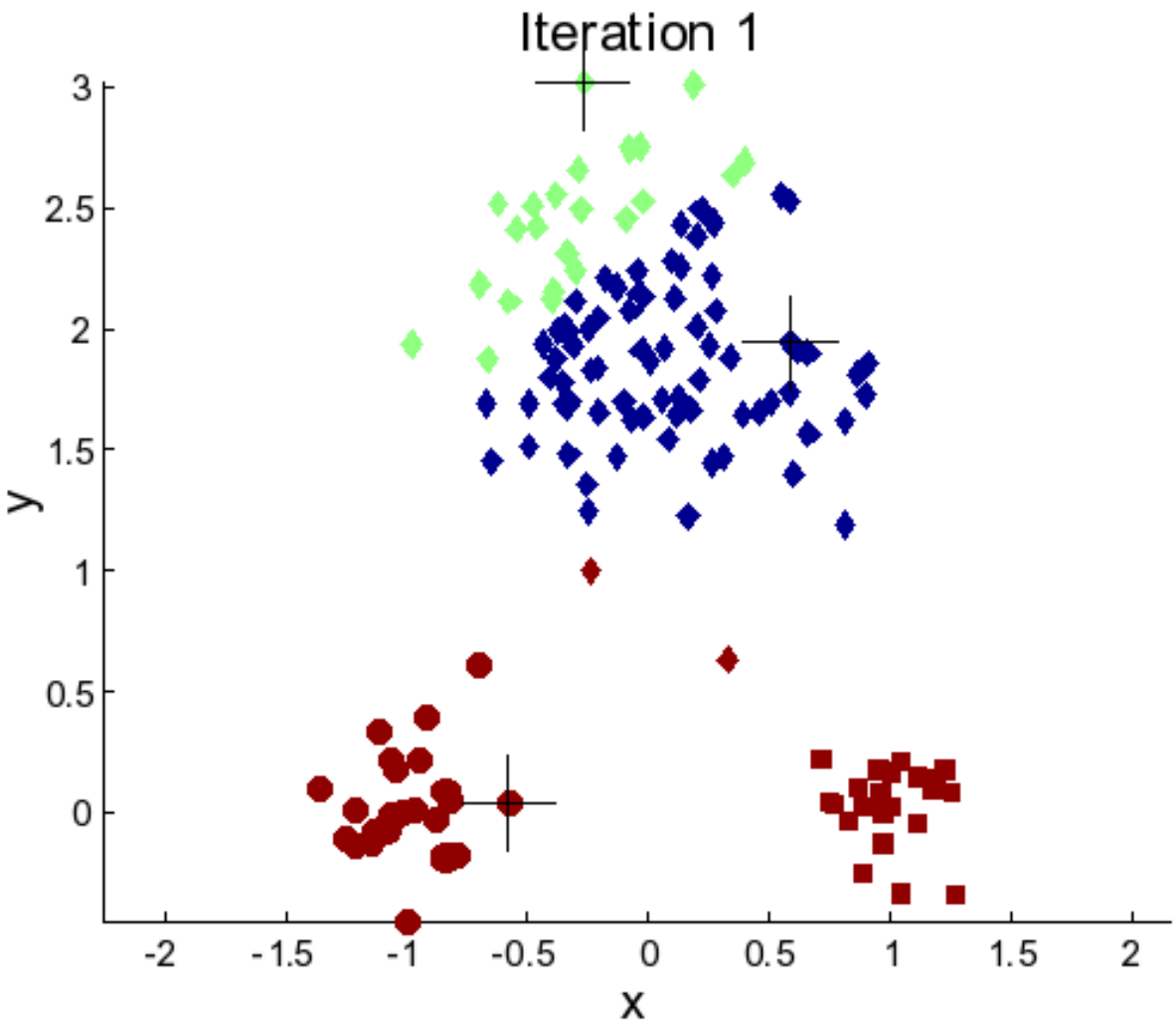

Iteration 4
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x

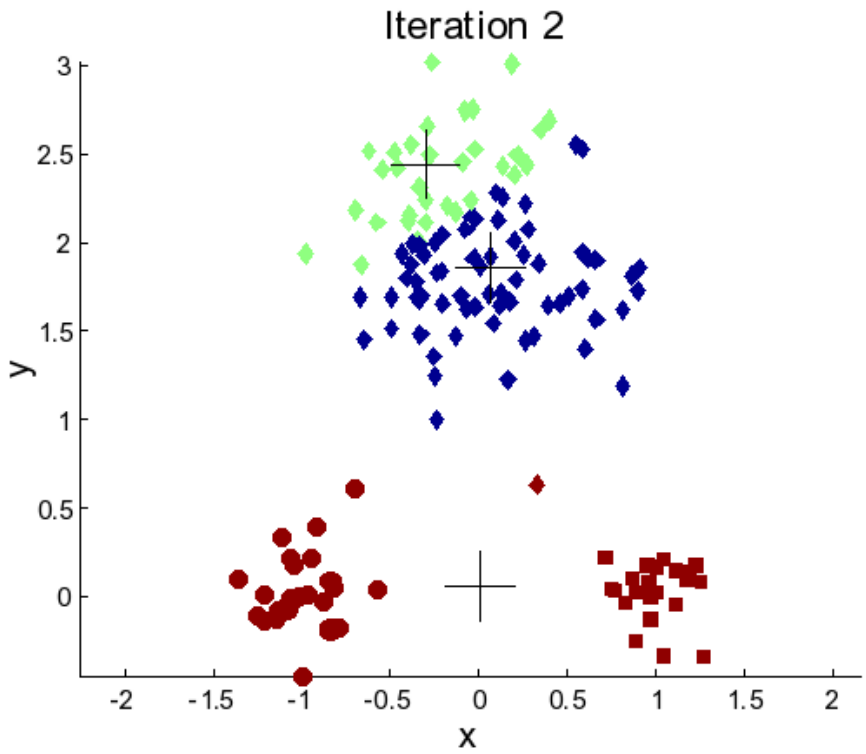

Iteration 5
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x

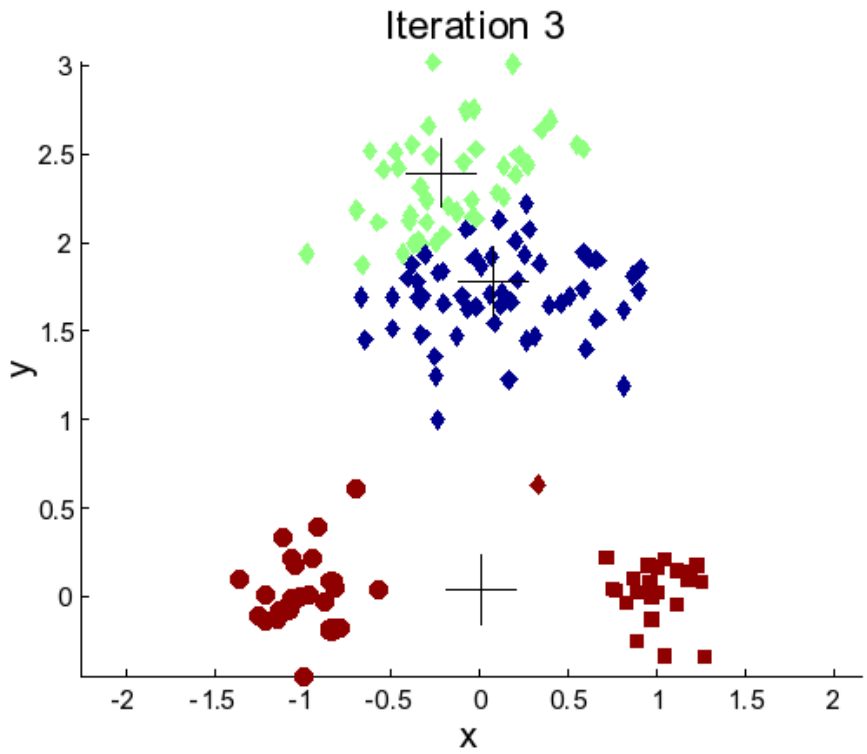

Iteration 6
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x

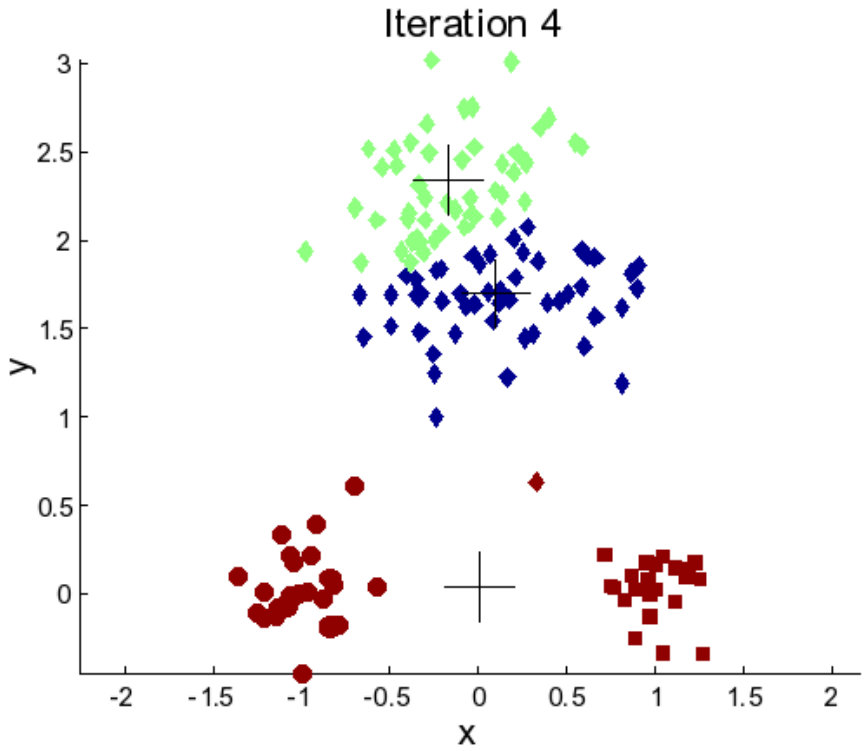

Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
x
y

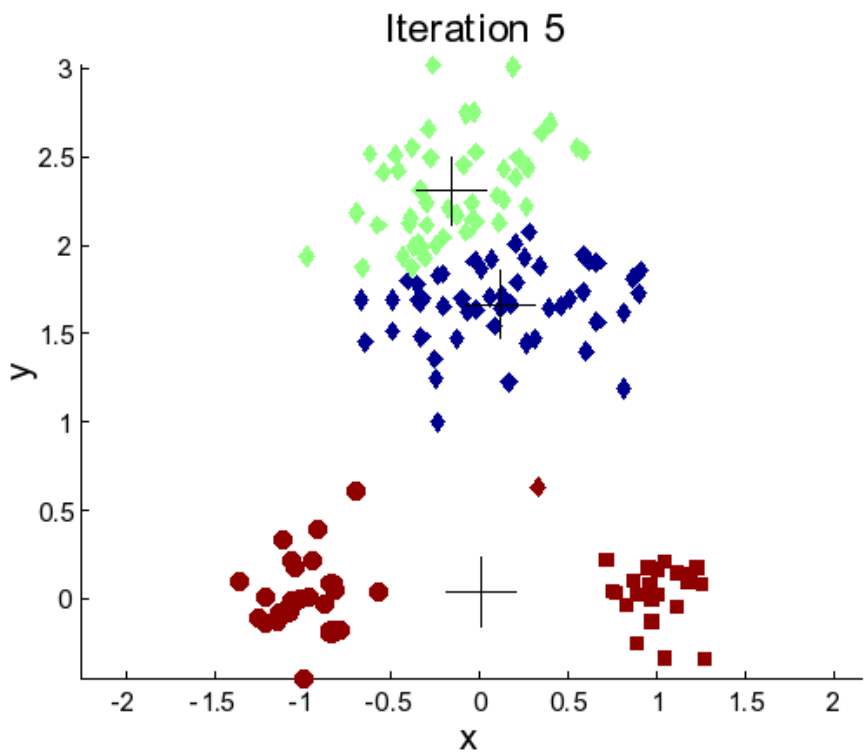
Importance of initial values

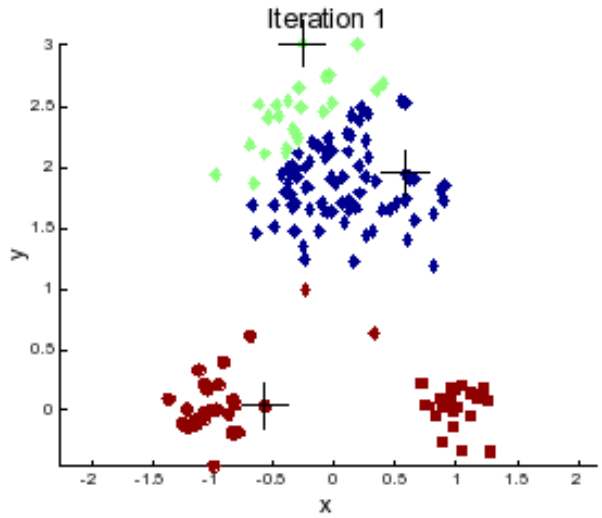


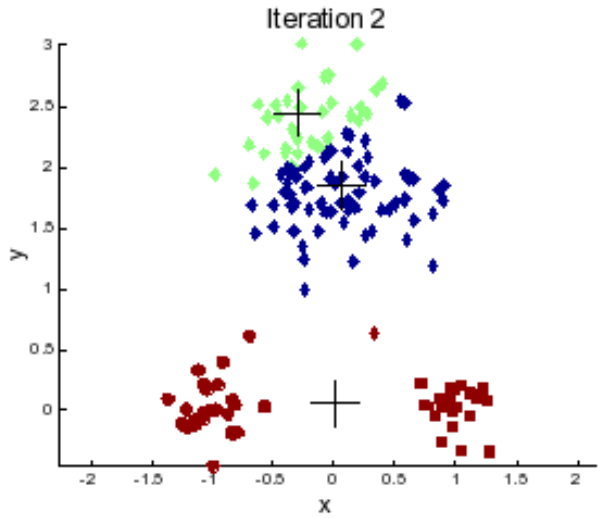

Iteration 2
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x

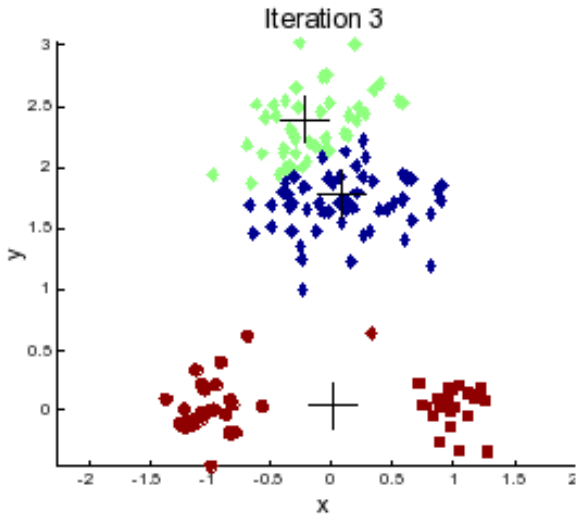

Iteration 3
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x

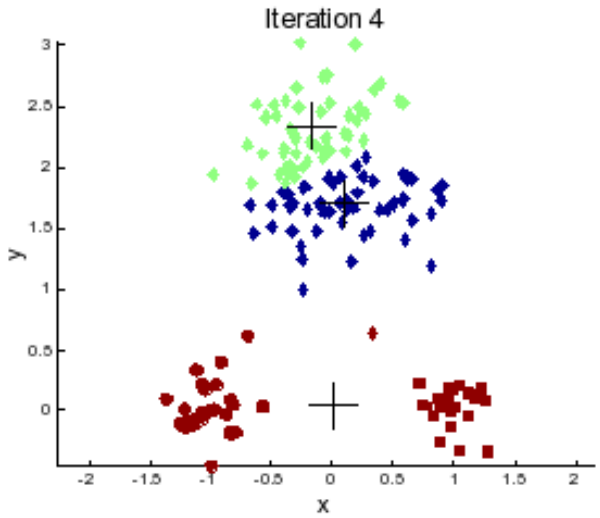

Iteration 4
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x

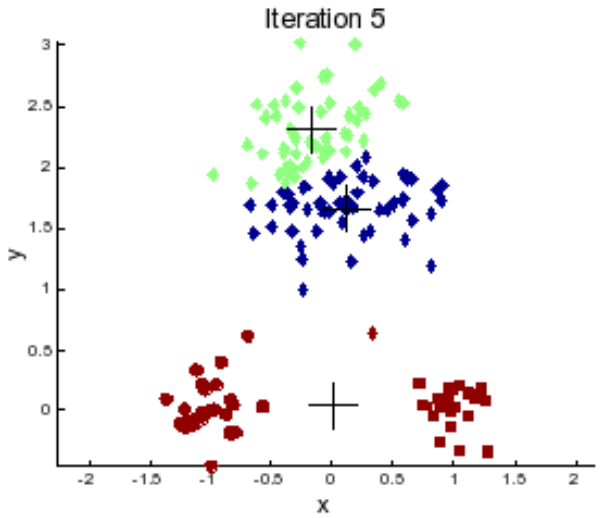

Iteration 5
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x


Iteration 1
x
y


Iteration 2
x
y


Iteration 3
x
y


Iteration 4
x
y


Iteration 5
x
y

Overcoming the initial centroids problem

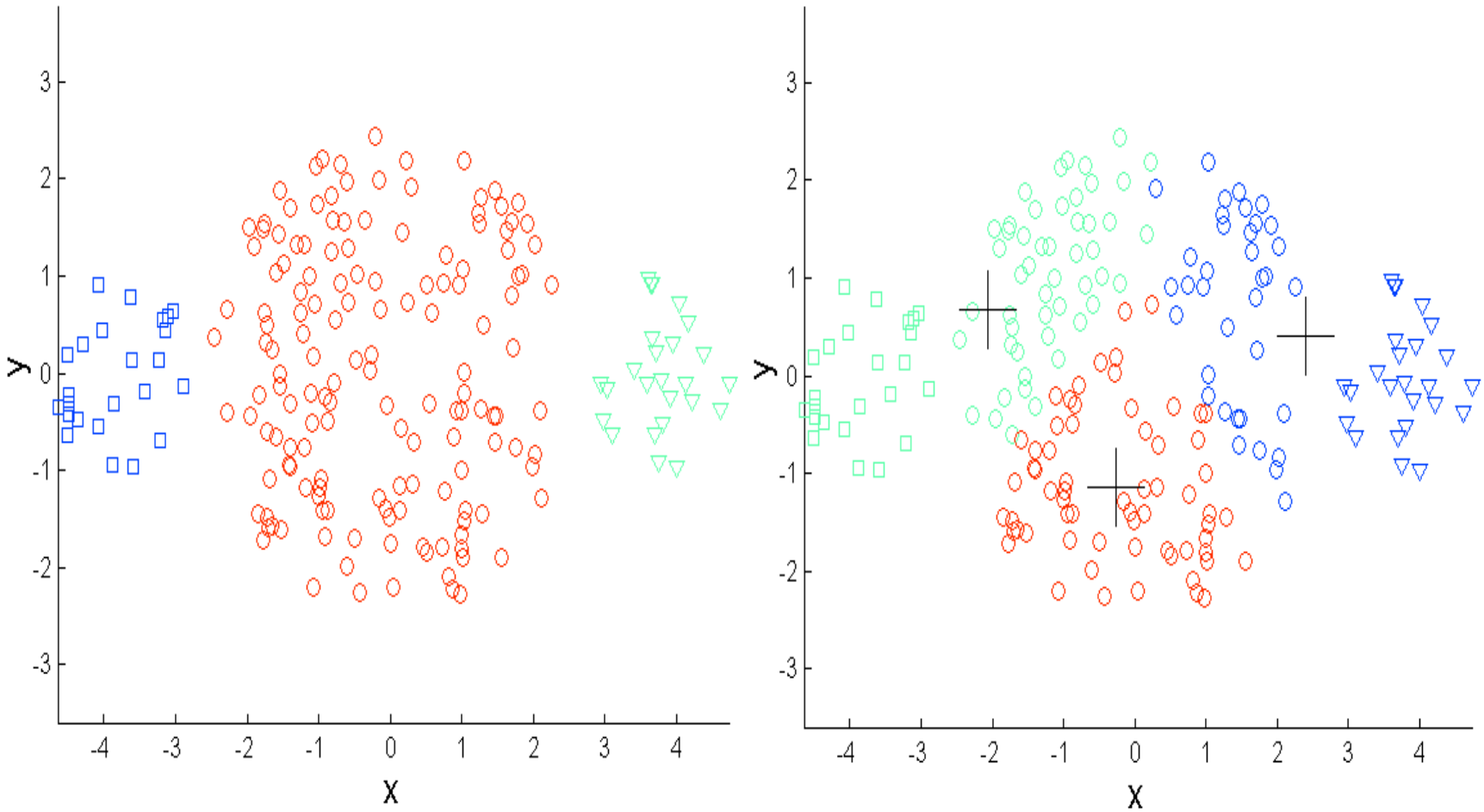
- The chance of randomly selecting one centroid from each cluster is small
- Multiple runs of the algorithm: helps, but probability is not on your side.
- Use the solution from some hierarchical algorithm as initial value
- Select more than K initial centroids and then select K best separated.

Limitations of K-means

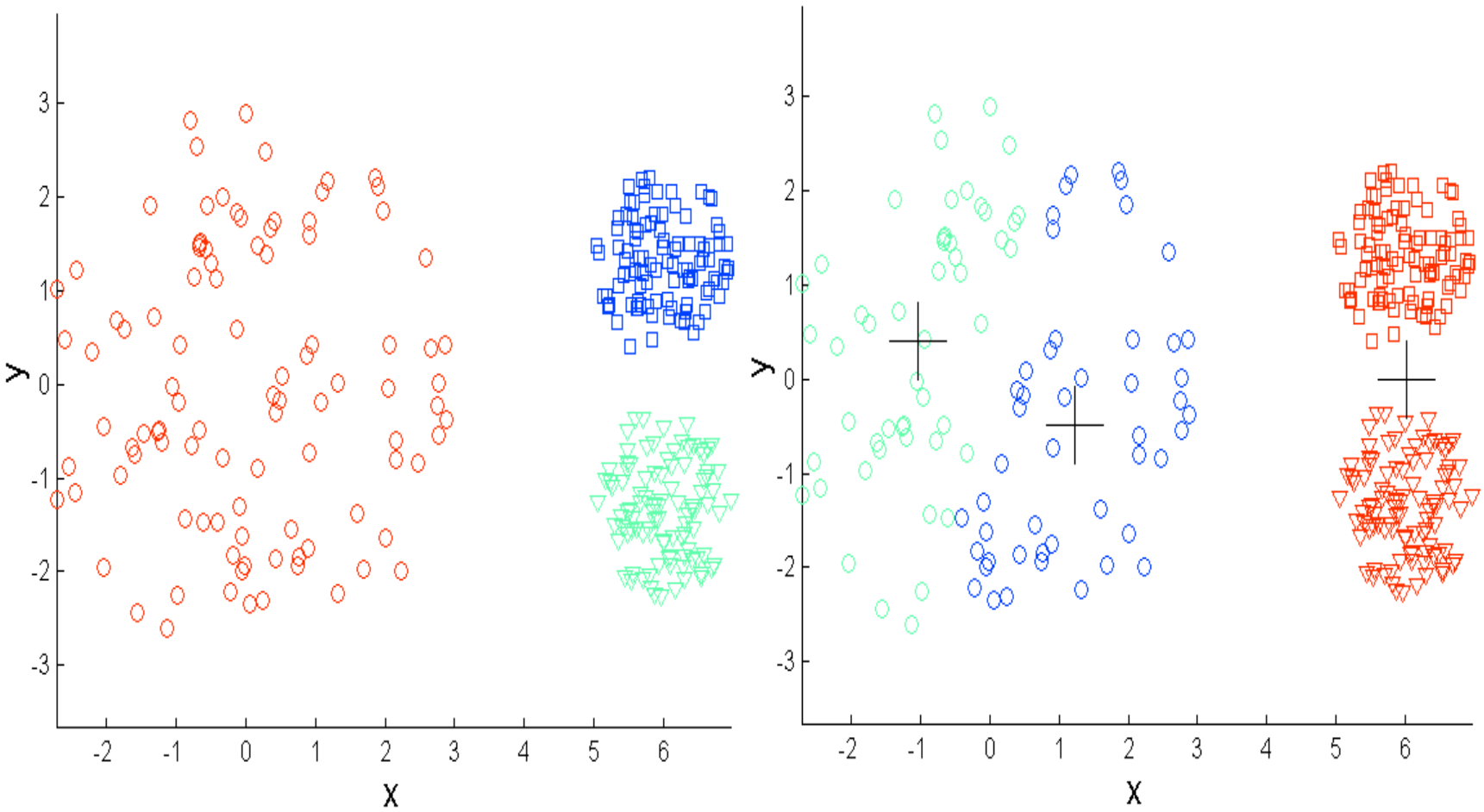
K-means has problems when

- Clusters are of different sizes
- Clusters have different densities
- Clusters have non-spherical shapes
- Data contain outliers

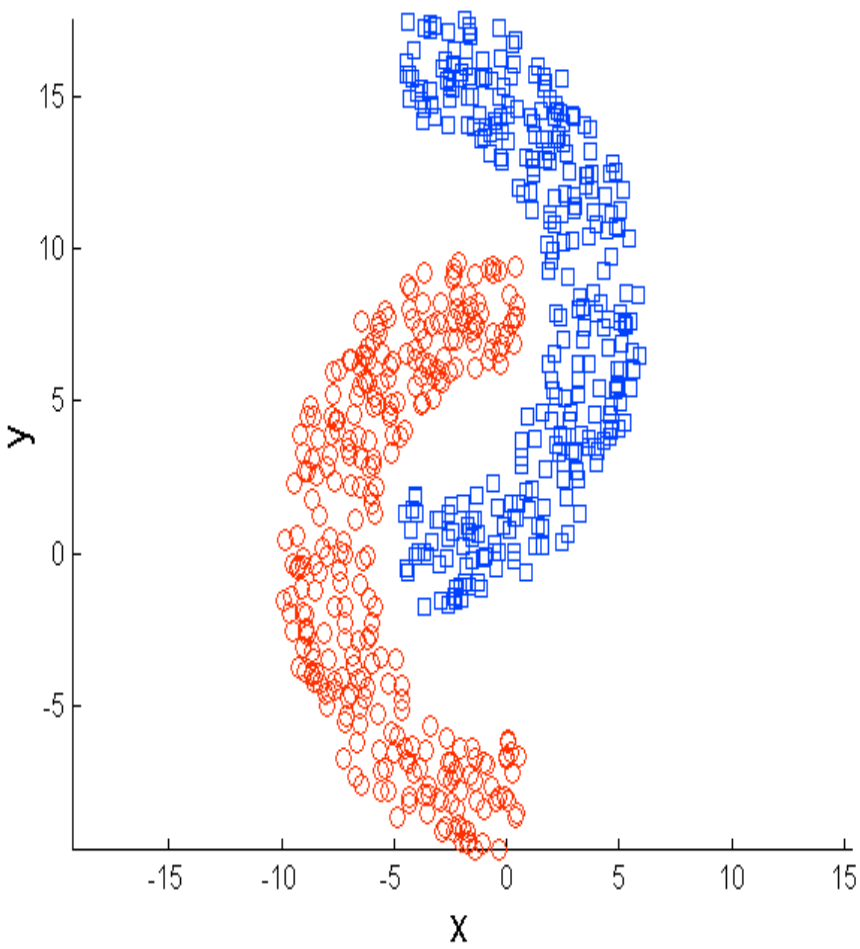
Limitations of K-means: Different sizes

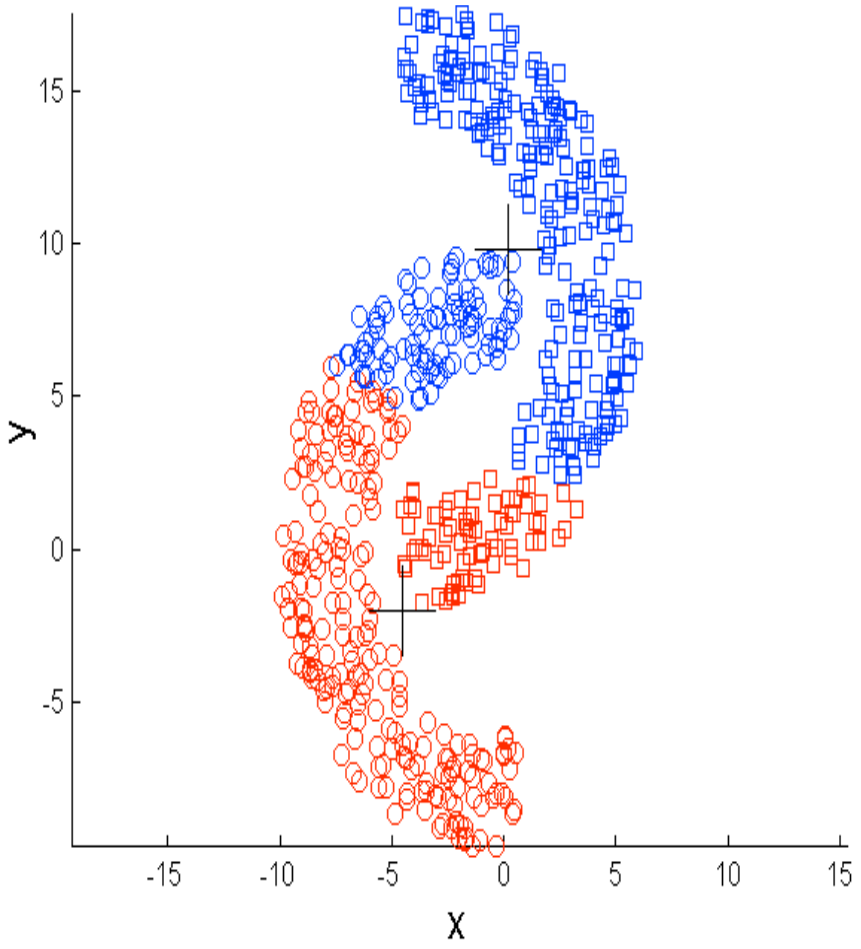


Limitations of K-means: Different densities



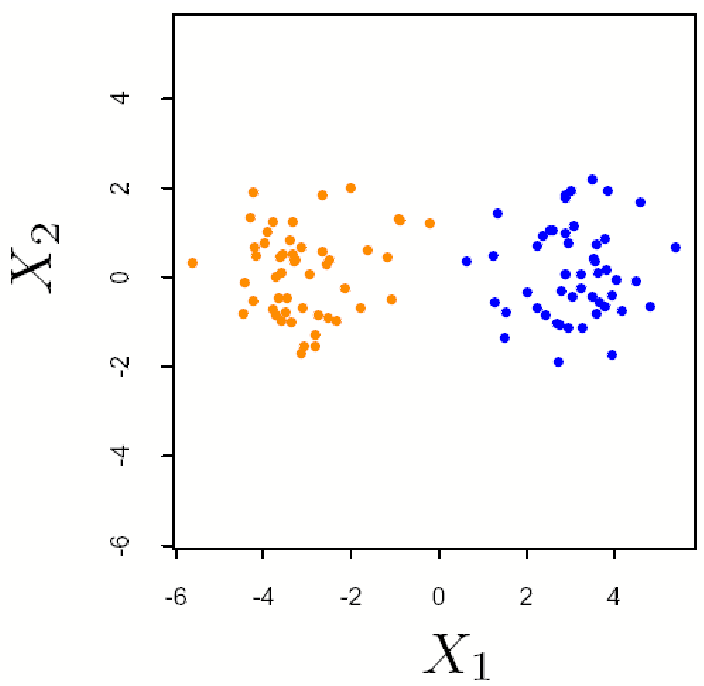
Limitations of K-means: Non-spherical shapes



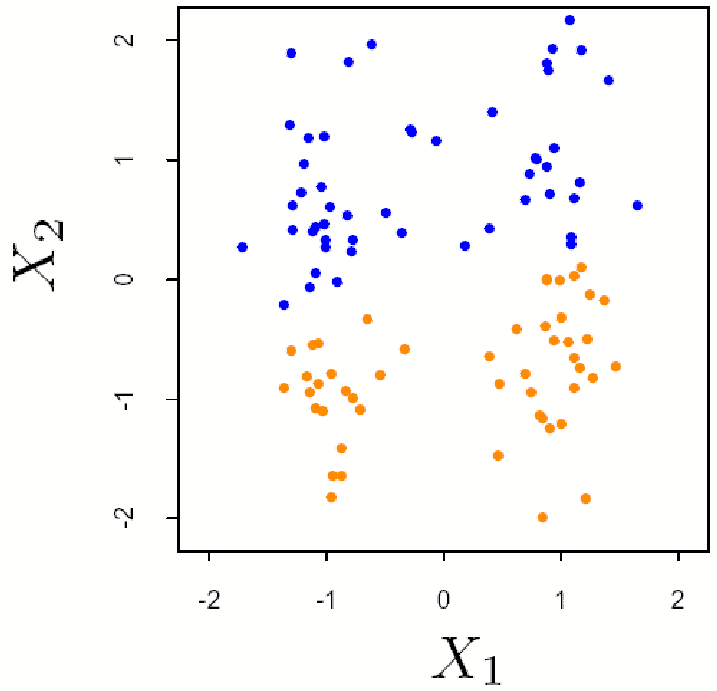


To standardize or not standardize?

- May help but often does not
- Unless there is a special reason, don't



Original data (2-means)

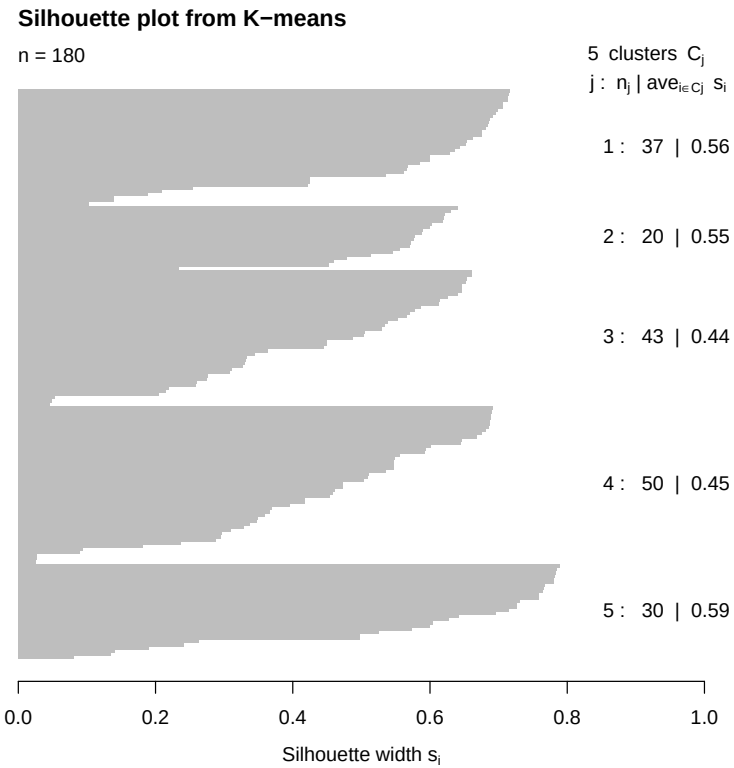


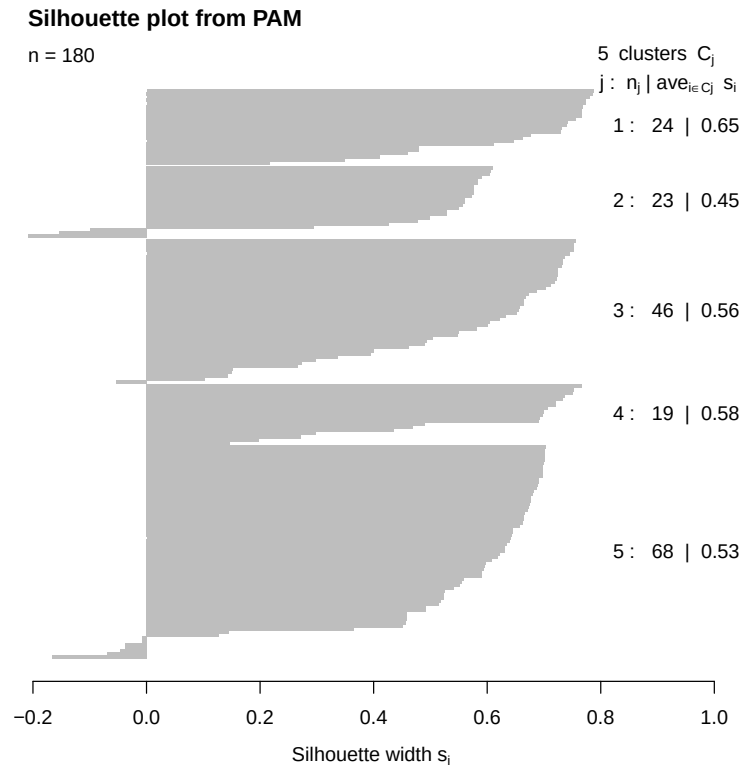
Standardized data (2-means)

PAM: robust version of K-means

- PAM = partitioning around medoids (multivariate median)
- Can take dissimilarity matrix as input (unlike regular K-means)

K-means and PAM on the wine data





Spectral clustering: an application of K-means

1. perform PCA on the data X and replace the $n \times p$ data matrix by the scores on the first K components ($n \times K$)
2. apply regular K-means clustering to the rows of the new $n \times K$ matrix

- A very popular general method
- Allows to reduce dimension before clustering
- Inherits most of advantages and disadvantages of K-means itself (e.g. sensitive to unbalanced clusters)
- The eigenvectors onto which the data points are projected can be computed from general similarity matrices (PCA is equivalent to computing them from the correlation matrix)

1.PCA X_n+p 2.cov: XX^T

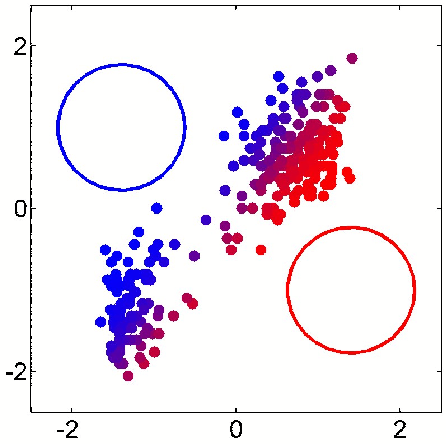
Model-based clustering

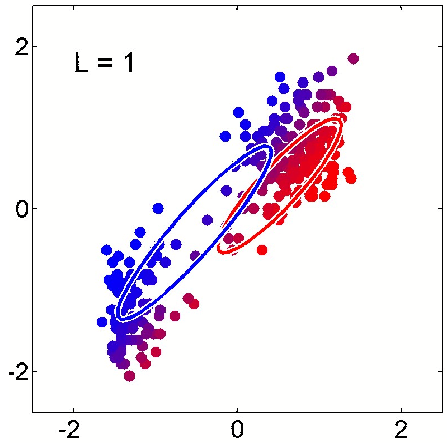
- A principled modeling approach to clustering
- Assume each point has a cluster assignment, but it is latent (unobserved)
- Assume a particular distribution for the labels, and for cluster densities conditional on the labels (with unknown parameters)
- Estimate the parameters and "fill-in" the missing labels

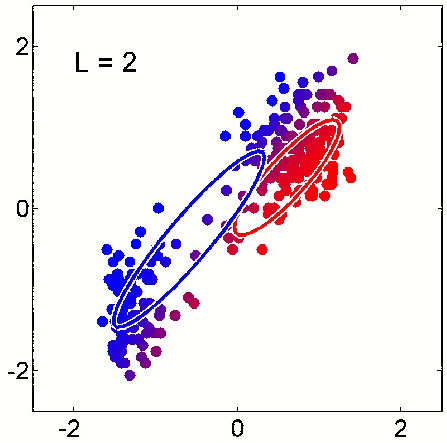
Gaussian mixture models (GMM)

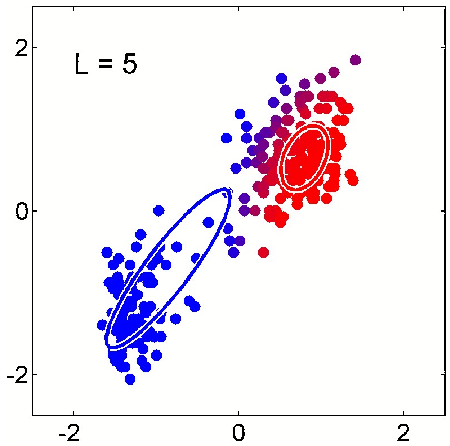
- Assume there are K possible labels, and each point i gets its label Z_i assigned independently of others, with probability π_k
- Conditional on $Z_i = k$ (which we don't observe), the points in each cluster come from the normal distribution, with mean μ_k and covariance matrix Σ_k
- Parameters (π_k, μ_k, Σ_k), $k = 1, \ldots, K$ and labels Z_i, $i = 1, \ldots, n$ can be estimated by the Expectation-Maximization (EM) algorithm (details omitted)
- K-means is a special case: amounts to assuming the same diagonal Σ_k for all clusters
- The general GMM allow for clusters of different shapes, orientations, and sizes: more details in R package `mclust`

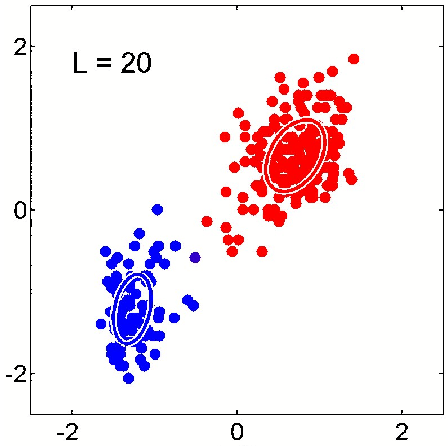
EM in action











Summary

- A variety of clustering methods are available
- Almost every clustering method works well in some situations and fails in others
- Determining the “best” clustering is hard (no objective measure of success)
- Determining the “best” number of clusters is equally hard
- A clustering that remains stable for multiple choices of similarity measures, methods, and small random perturbations of the data is more convincing
- Most importantly, clustering results should not be taken as the absolute truth about a data set, but rather as a starting point for developing scientific hypotheses and further study, preferably on independent data.