Mixture Models & EM

Lecture outline

- Brief review
  - what k-means CANNOT do
- How to model each cluster?
  - e.g., spherical Gaussians (brief review)
- Mixture Models
  - motivation, formulation
  - estimation, the EM algorithm
  - selecting the number of mixture components

Recall: K-means

- K-means clustering
  - initialize K different means
  - assign each data point to the closest cluster mean
  - re-estimate cluster means based on the points assigned to them
  - iterate until convergence

What K-means cannot do?

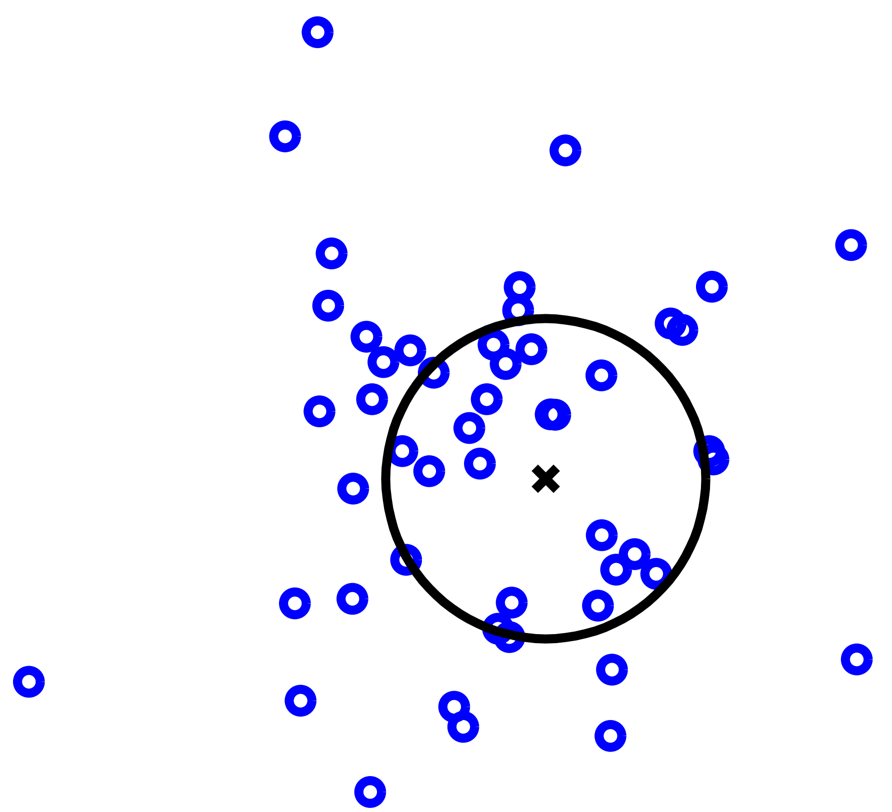
- It cannot handle overlapping clusters
- It cannot represent clusters with different “spreads”
- It cannot properly deal with clusters that have different numbers of points

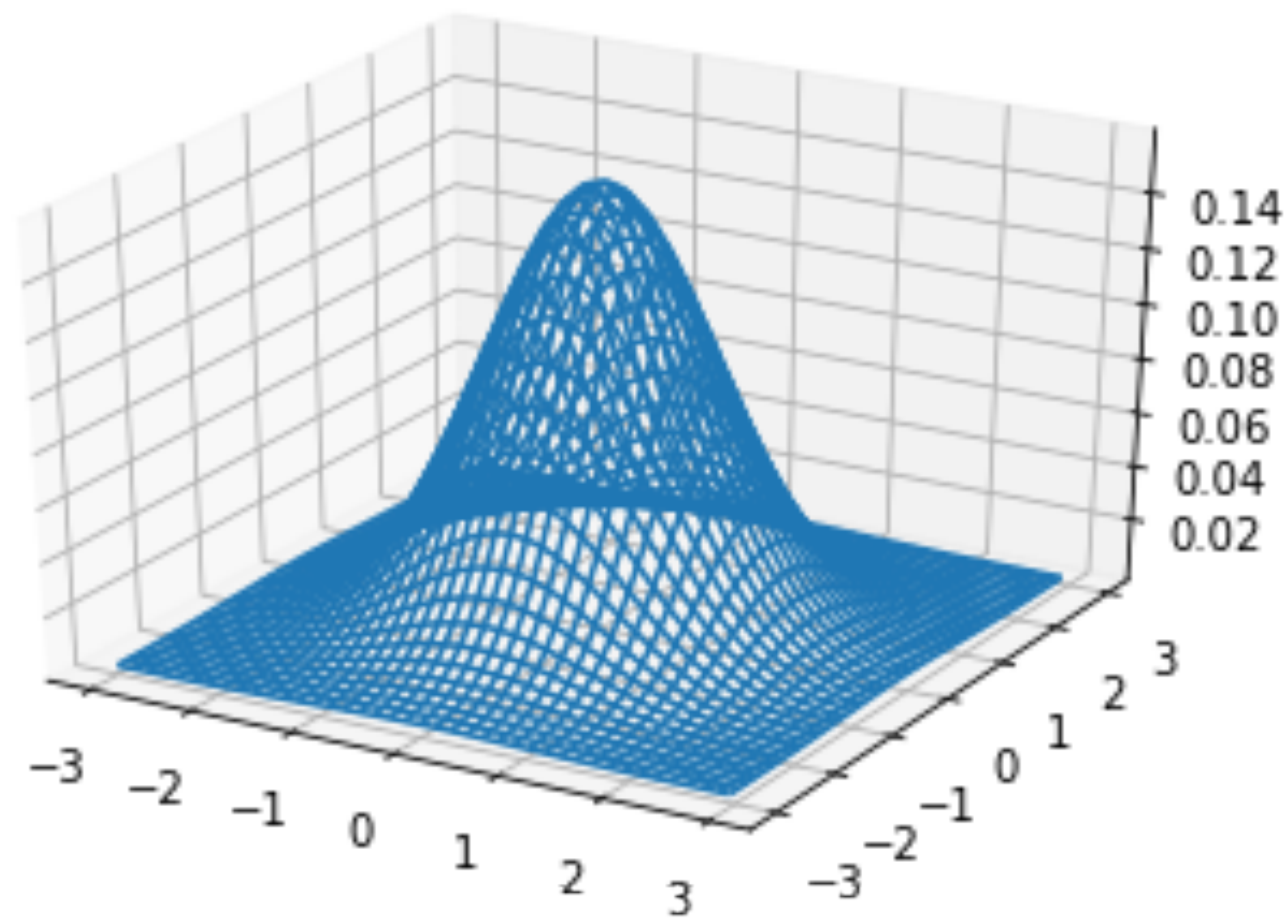
Recall: spherical Gaussian

- The pdf of a spherical Gaussian

$$N(x; \mu, \sigma^2 I) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \|x - \mu\|^2 \right)$$

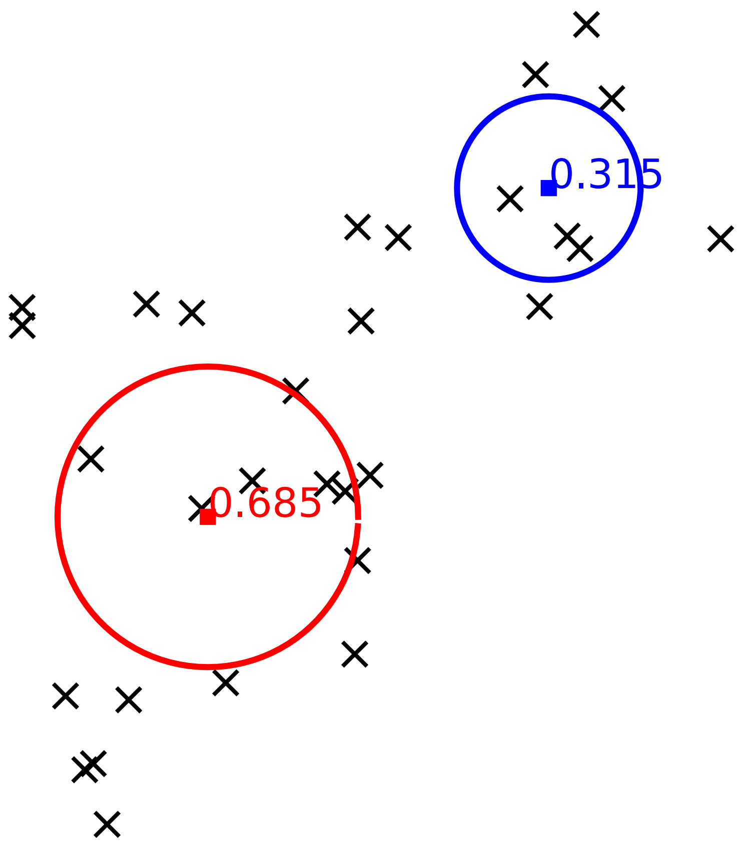
- Graphical representation (when dim d = 2) in terms of mean and stdv





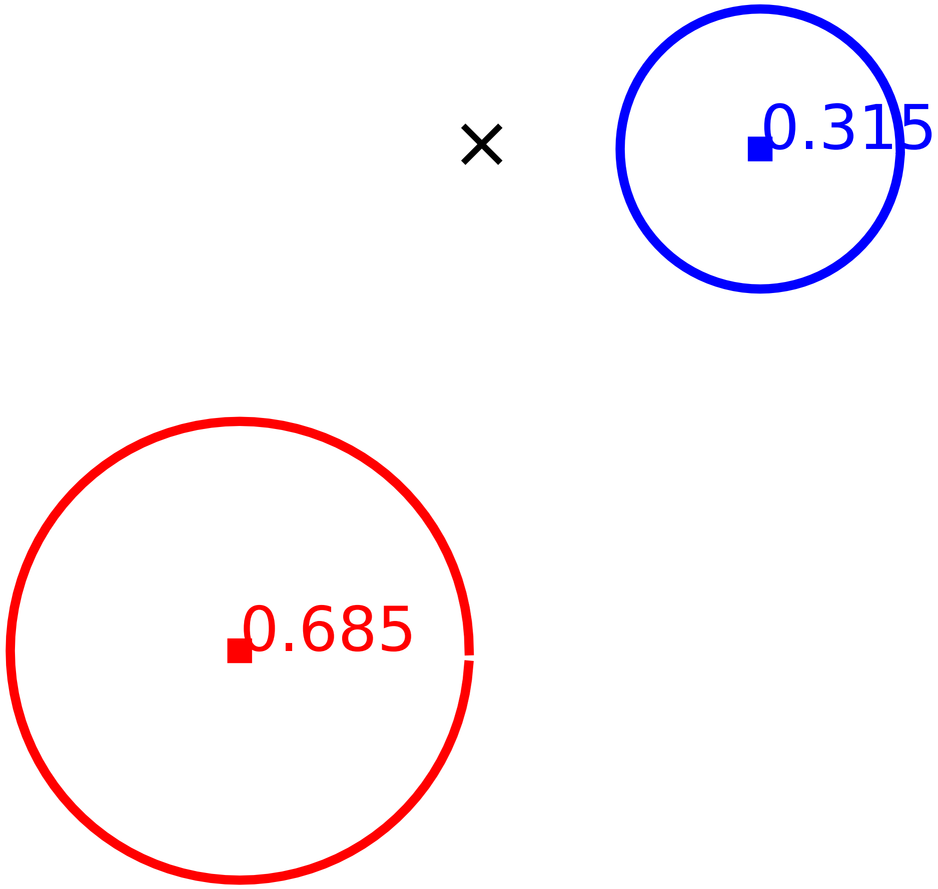
Mixture model: overview

- We use a spherical Gaussian to model each cluster
- These cluster models can have different means, variances (spreads), as well as "sizes"
- A mixture model combines these "components" into an overall probability model $P(x;\theta)$



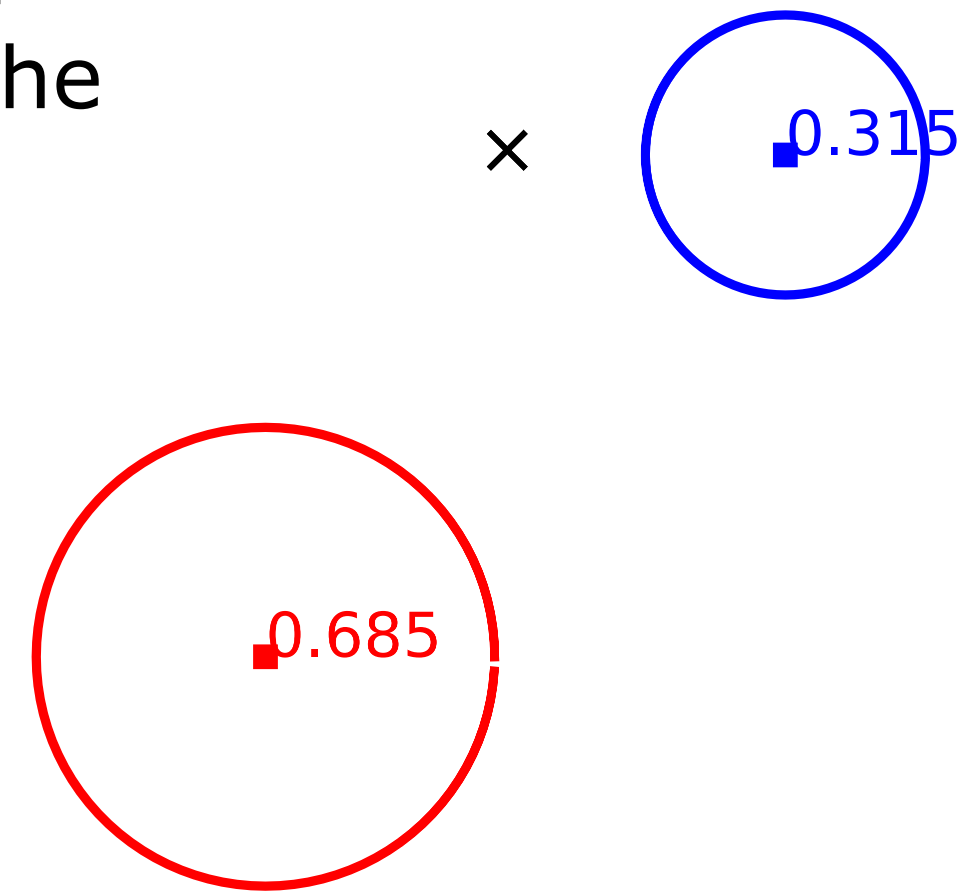
Mixture model: data generation

- We consider alternative ways that each data point x could have been created



Mixture model: data generation

- We consider alternative ways that each data point x could have been created:
  - select a cluster
  - select x from the cluster model



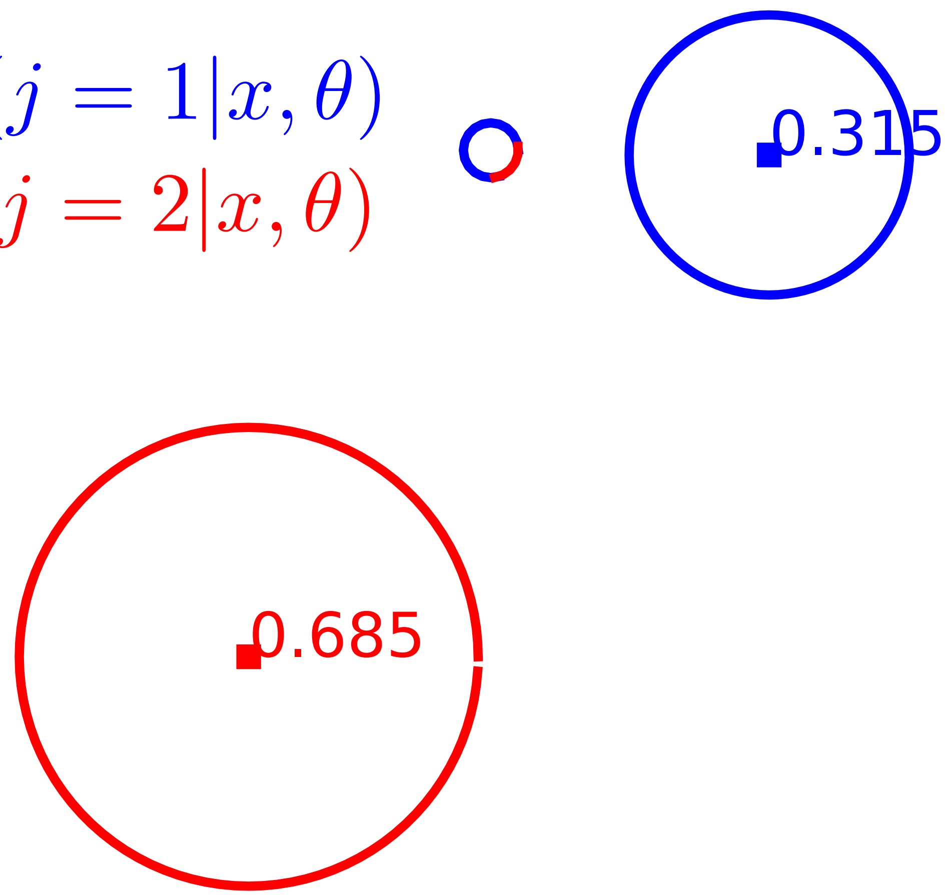
$$P(x;\theta) = p_1 N(x;\mu^{(1)},\sigma_1^2 I) + p_2 N(x;\mu^{(2)},\sigma_2^2 I)$$

Mixture model: posterior

- We can also infer (after the fact) which cluster likely generated each point by evaluating the posterior probability

$P(j = 1|x, \theta)$

$P(j = 2|x, \theta)$



$$P(j = 1|x, \theta) = \frac{p_1 N(x; \mu^{(1)}, \sigma_1^2 I)}{p_1 N(x; \mu^{(1)}, \sigma_1^2 I) + p_2 N(x; \mu^{(2)}, \sigma_2^2 I)}$$

Mixture models: estimation

- The goal is to find the parameters of the mixture model that maximize the log-likelihood that the data points came from the mixture distribution

$$l(D;\theta) = \sum_{i=1}^{n} \log P(x^{(i)};\theta)$$

$$= \sum_{i=1}^{n} \log \left[\sum_{j=1}^{K} p_j N(x^{(i)}; \mu^{(j)}, \sigma_j^2 I) \right]$$

- The difficulty lies in the fact that the Gaussian cluster models cannot be estimated independently (since we don't know which data points they should generate)

The EM algorithm: overview

- The EM algorithm solves the problem by iteratively re-assigning points to clusters (softly) and re-estimating the corresponding cluster models (cf. K-means)

- **Initialize** mixture

- **E-step** (complete the data)
  - evaluate the posterior probability that each data point came from a particular cluster

- **M-step** (maximize expected log-likelihood)
  - use the posterior probabilities (now fixed) as cluster specific weights on data points to separately re-estimate each cluster model

The EM algorithm (iterative)

- **Initialize:**
  - e.g. means as randomly selected points, all variances set to overall variance, uniform mixing proportions
- **E-step:** calculate posterior assignments

$$p(j|i) = \frac{p_j N(x^{(i)}; \mu^{(j)}, \sigma_j^2 I)}{P(x^{(i)}|\theta)}, \quad j = 1, \ldots, K, \quad i = 1, \ldots, n$$

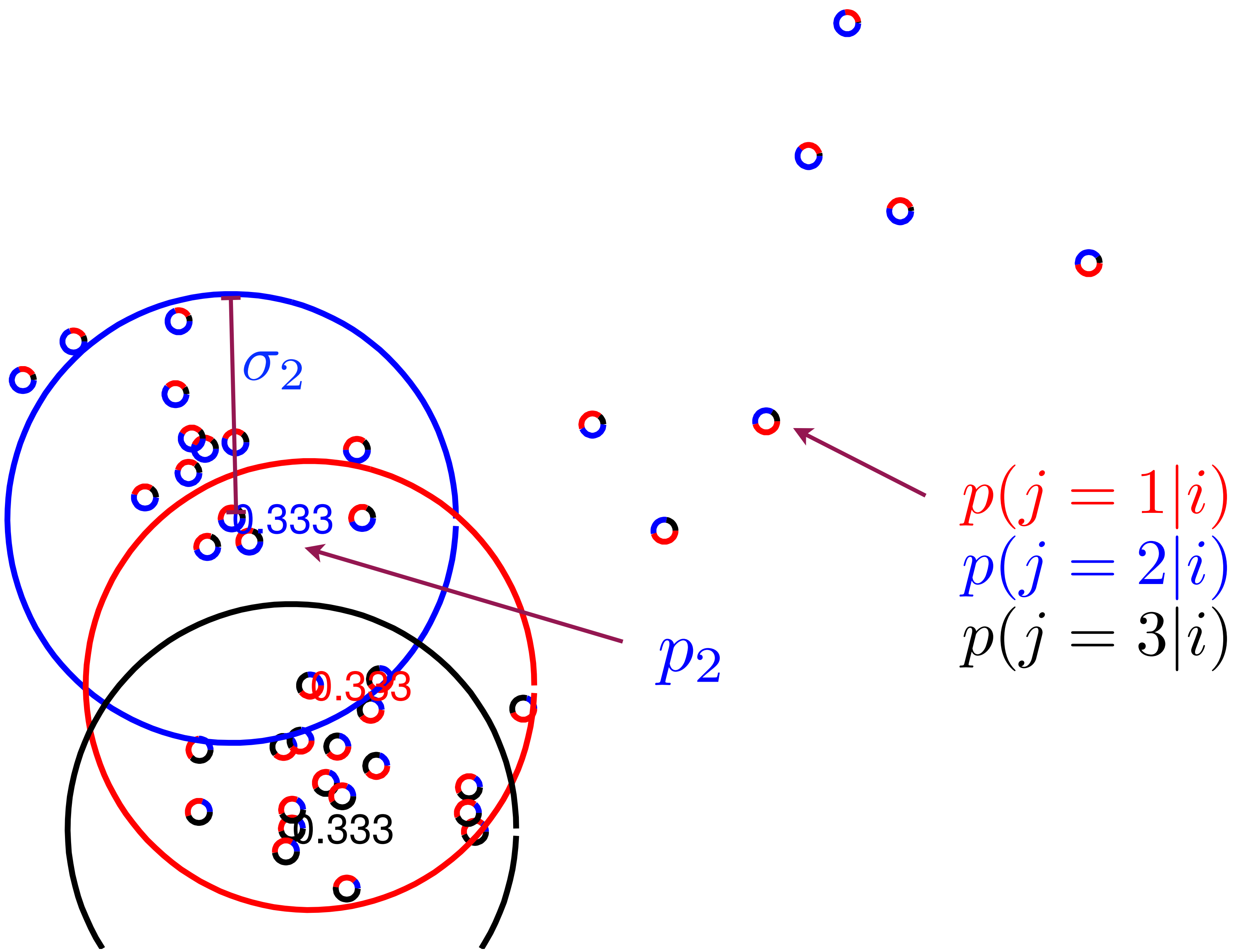
- **M-step:** maximize

$$\tilde{l}(D; \theta) = \sum_{j=1}^{K} \sum_{i=1}^{n} p(j|i) \log \left[\frac{p_j N(x^{(i)}; \mu^{(j)}, \sigma_j^2 I)}{p(j|i)} \right]$$

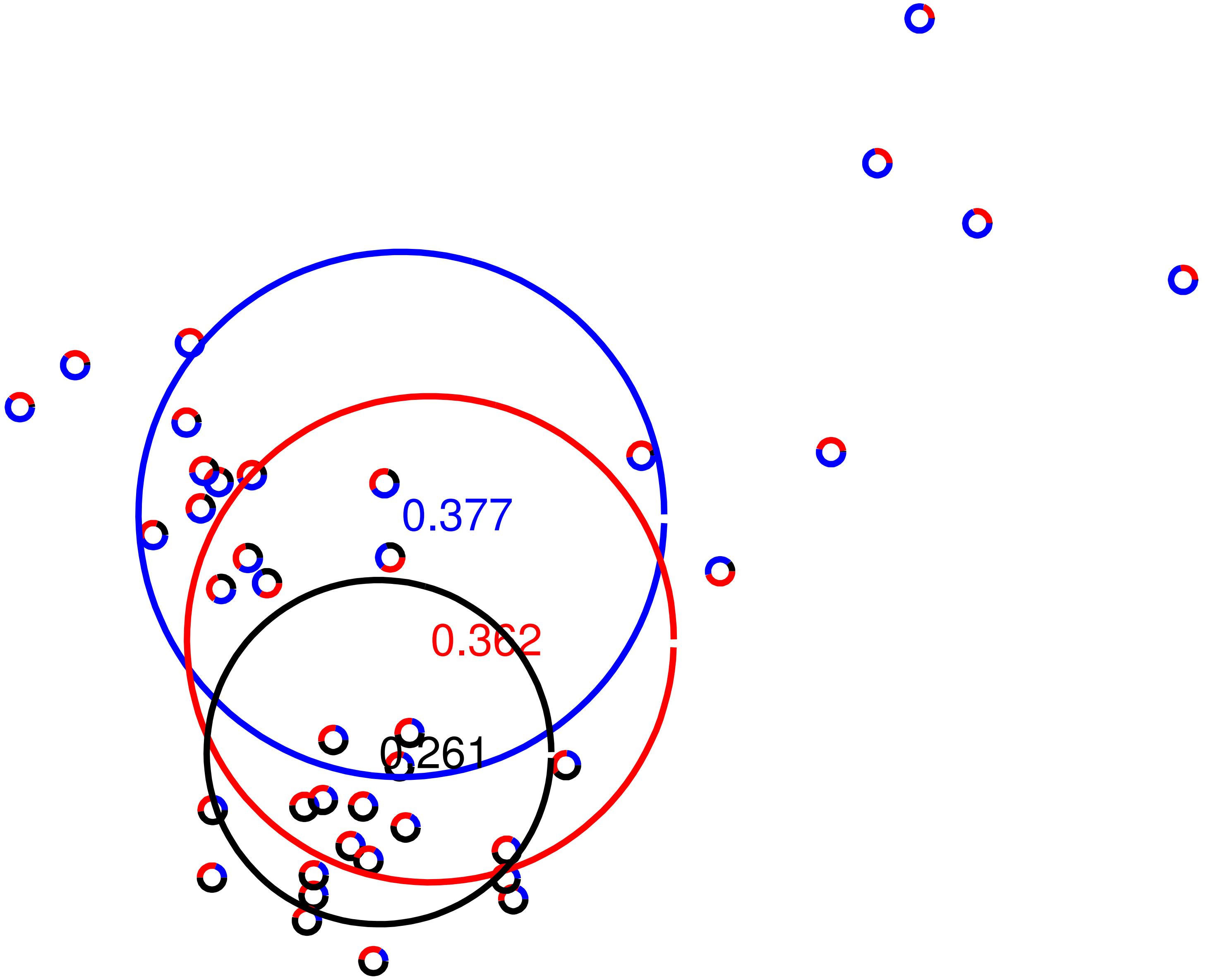
with respect to mixture parameters while keeping $p(j|i)$ fixed

Mixture of Gaussians example

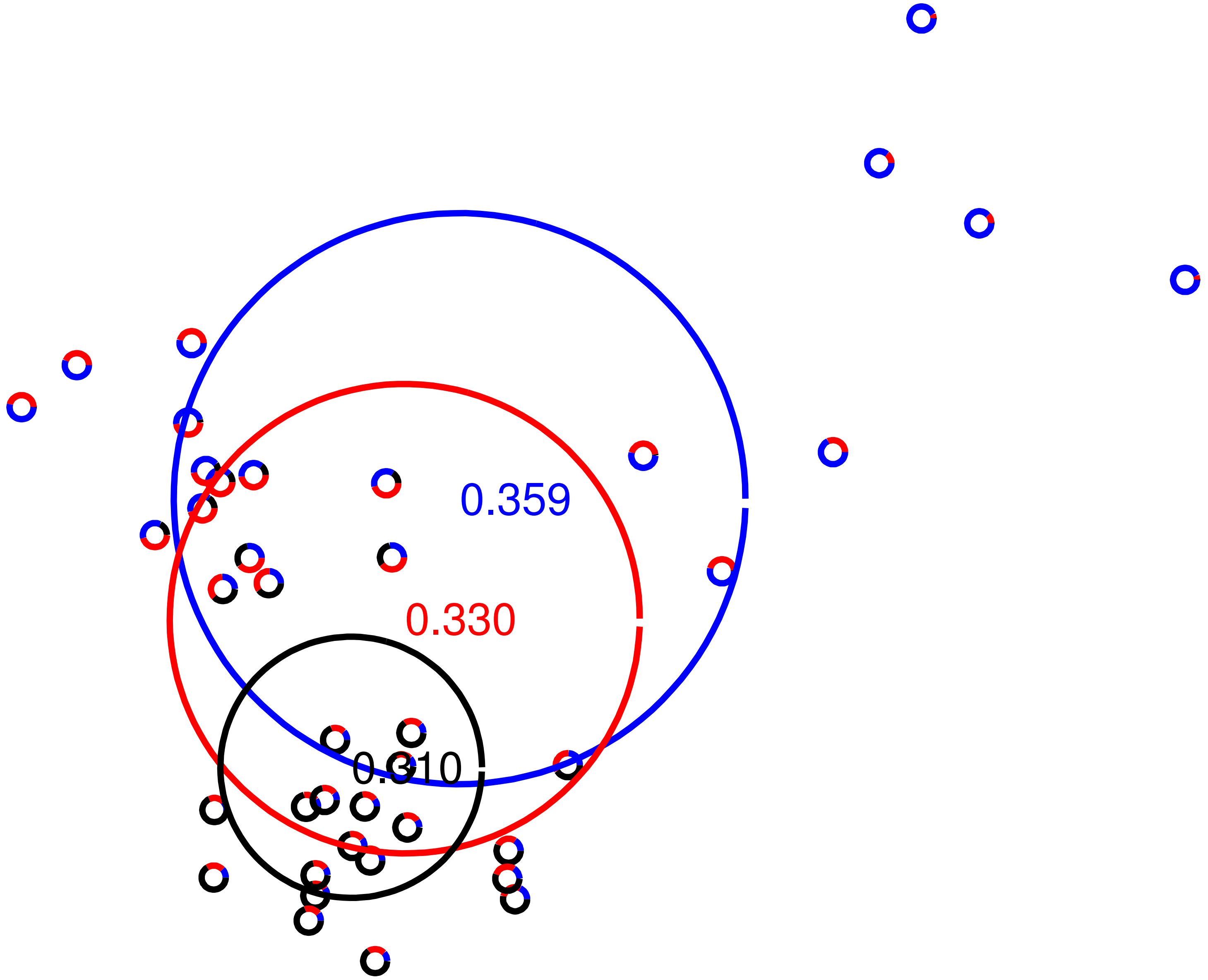
- initial 3-component mixture



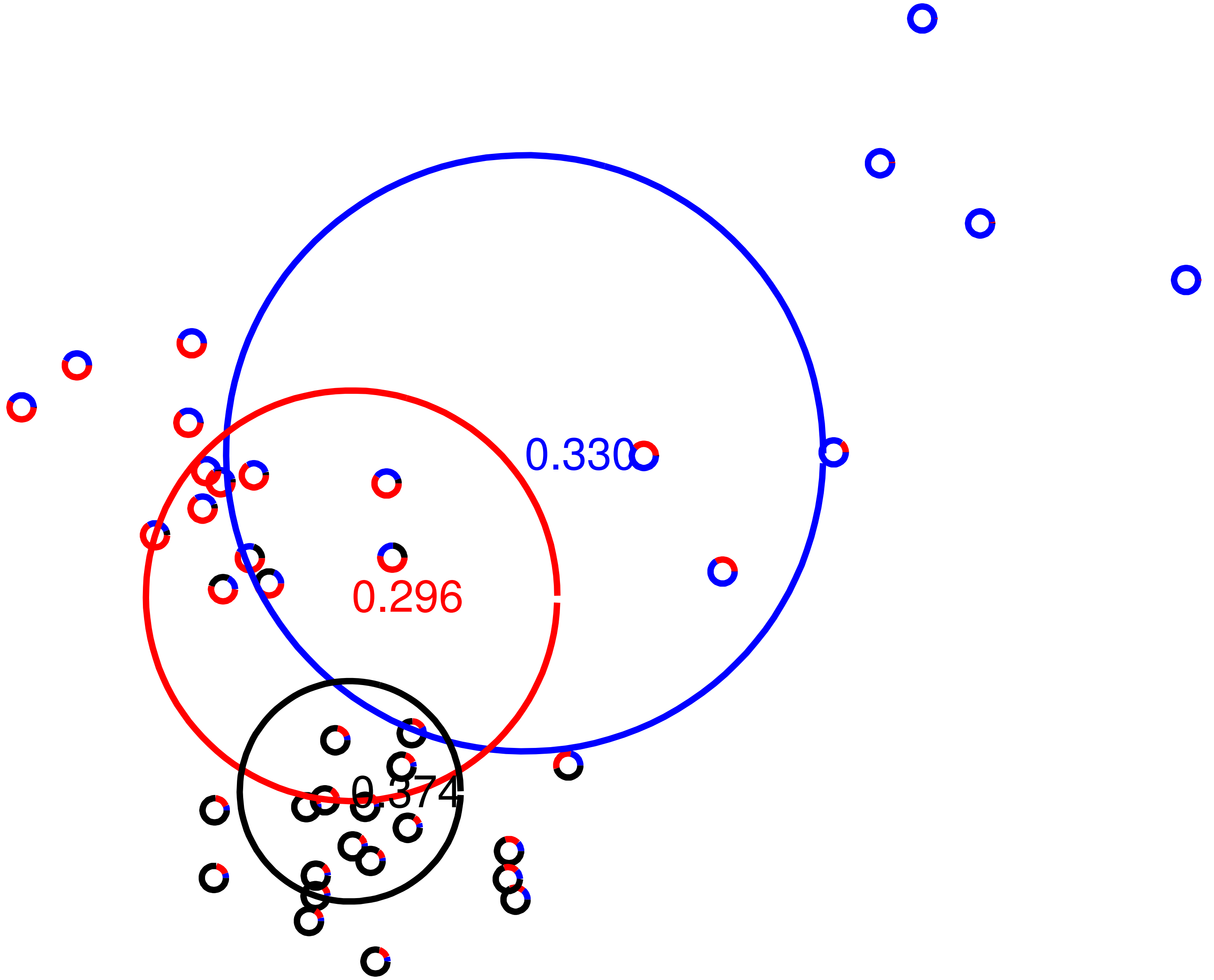
Mixture of Gaussians example



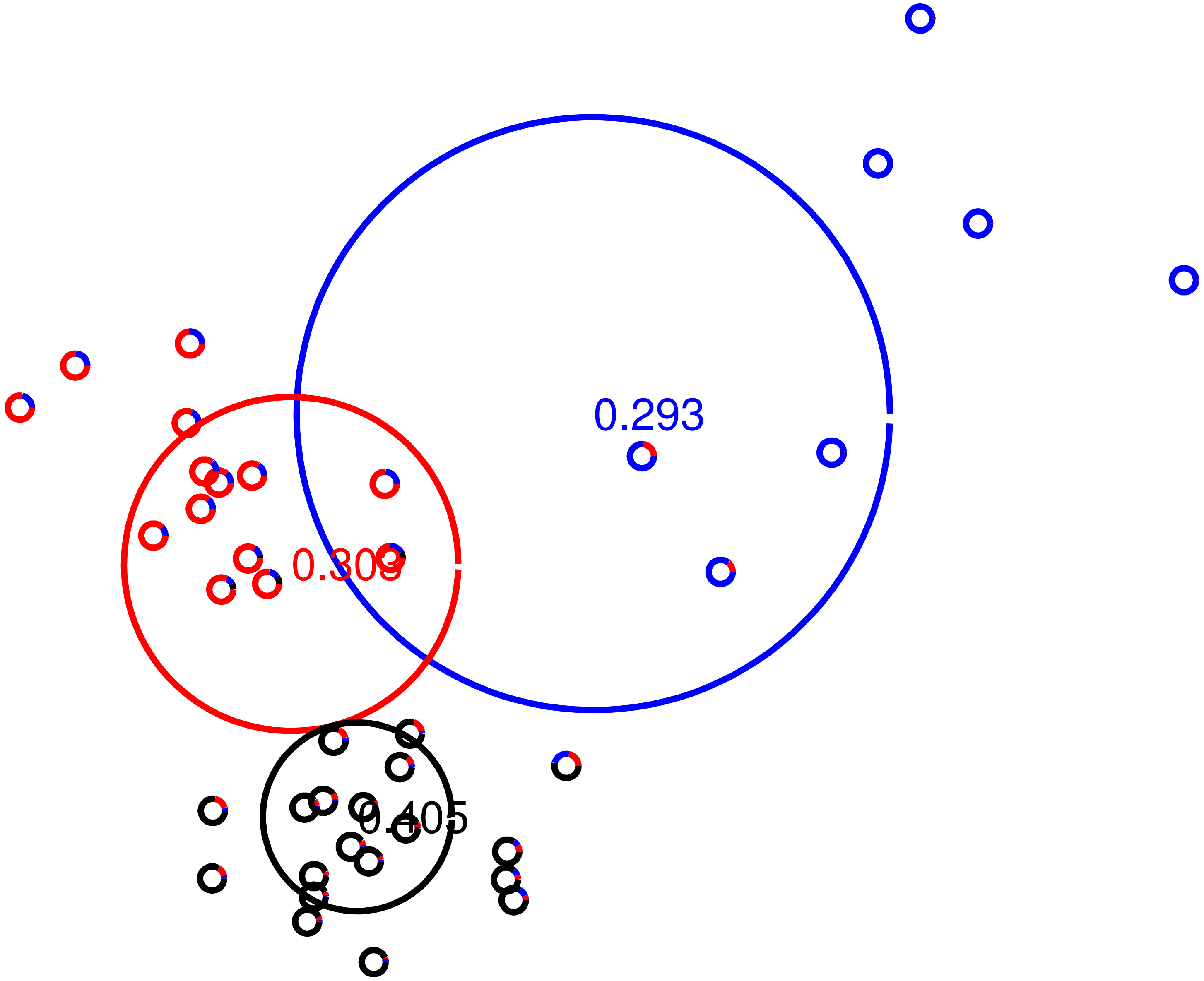
Mixture of Gaussians example



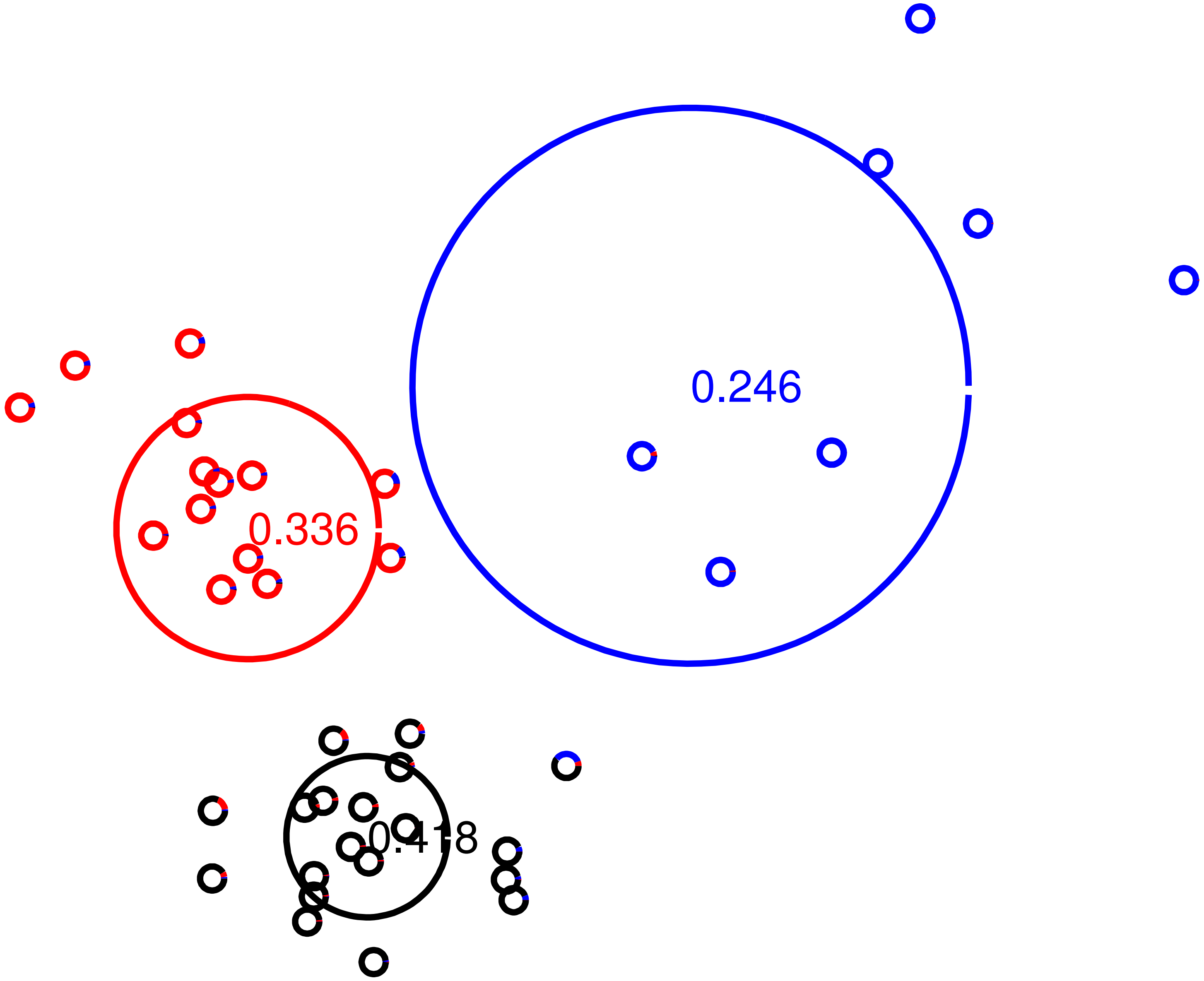
Mixture of Gaussians example



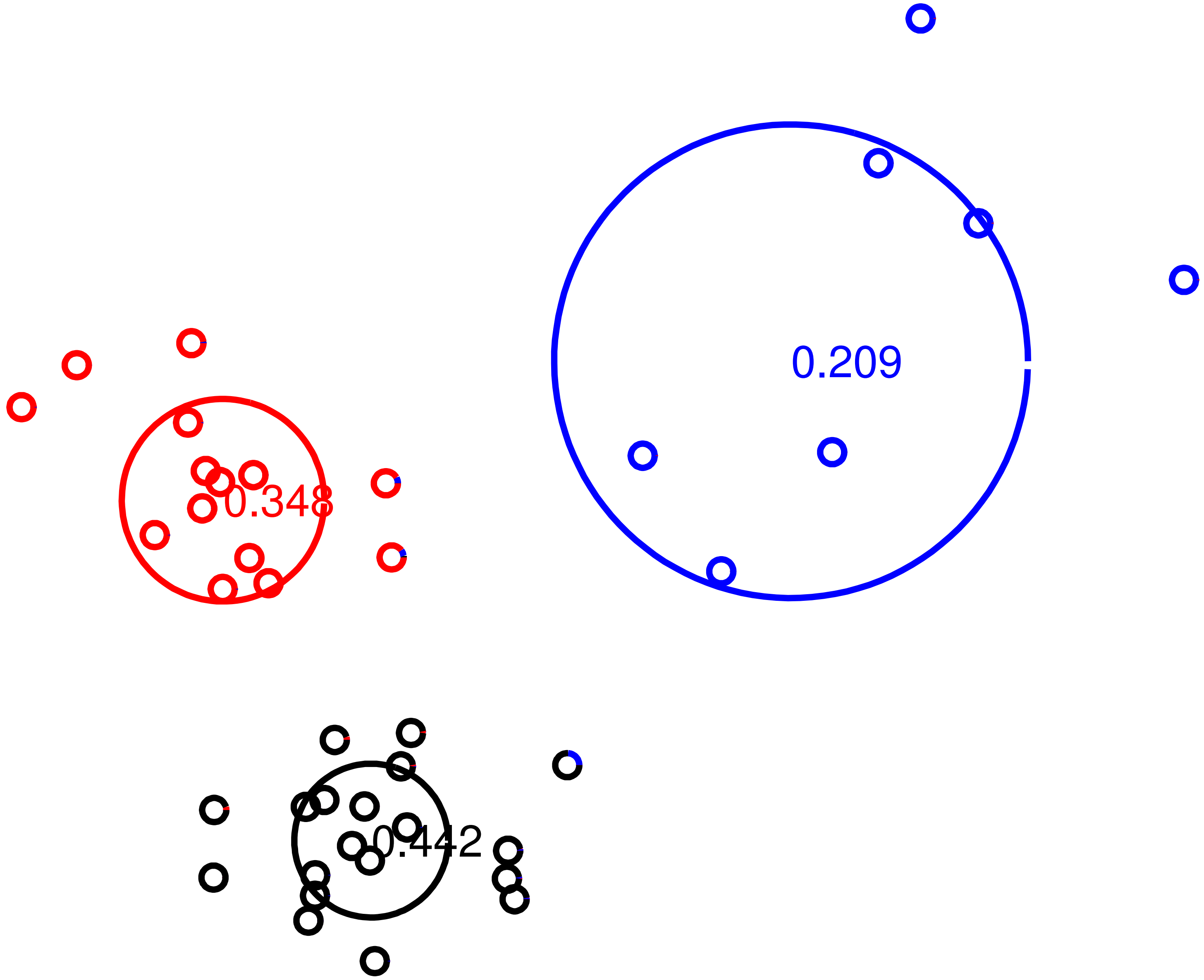
Mixture of Gaussians example



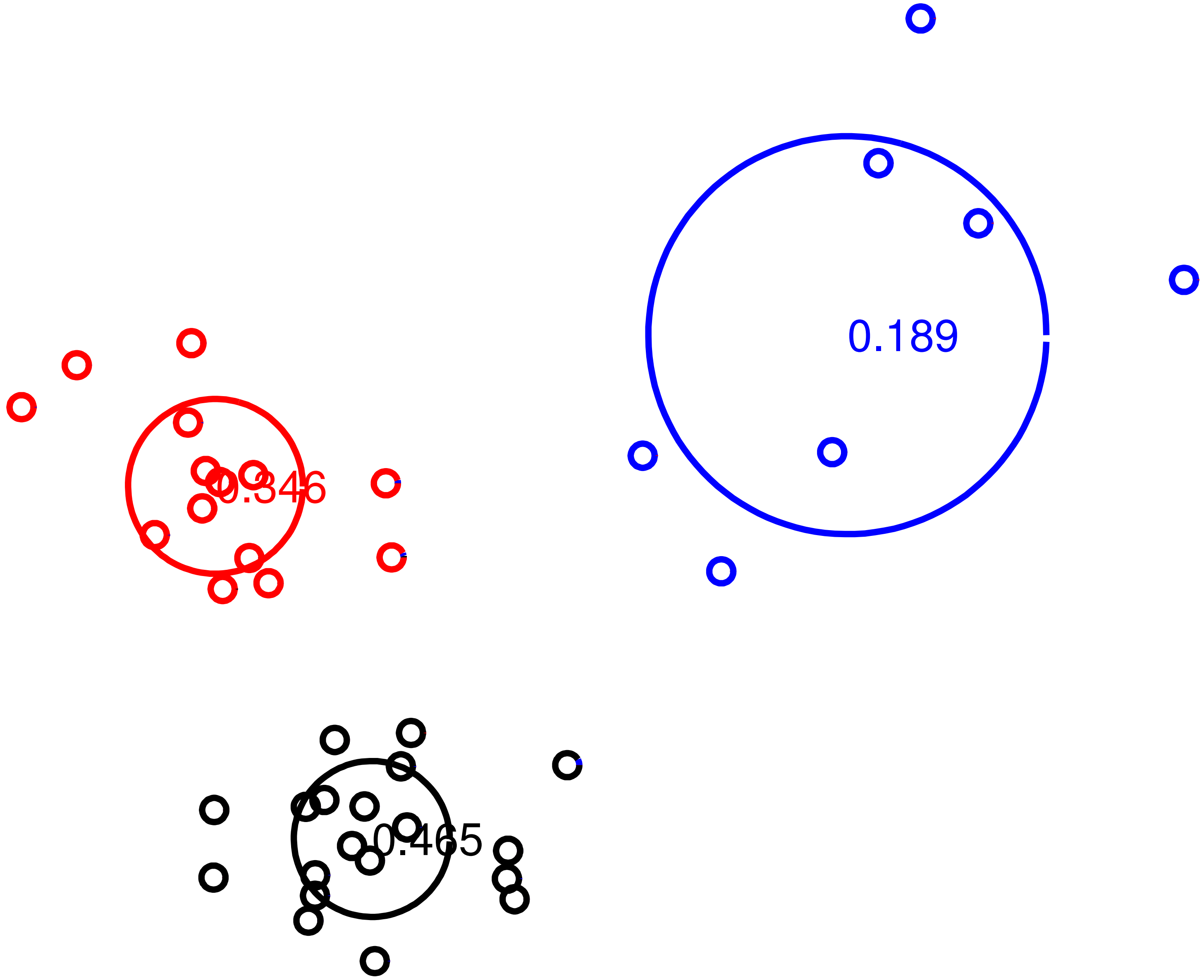
Mixture of Gaussians example



Mixture of Gaussians example



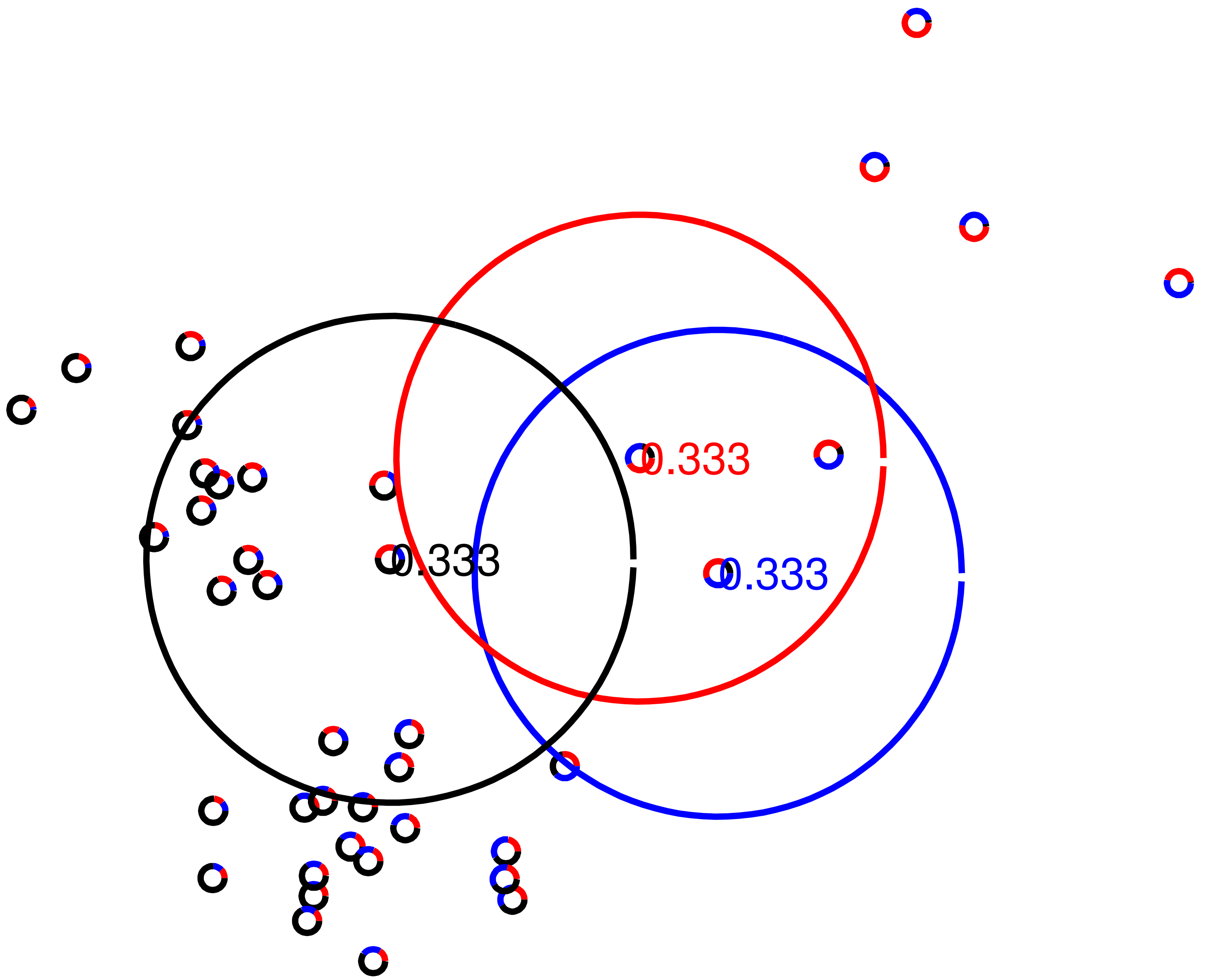
Mixture of Gaussians example



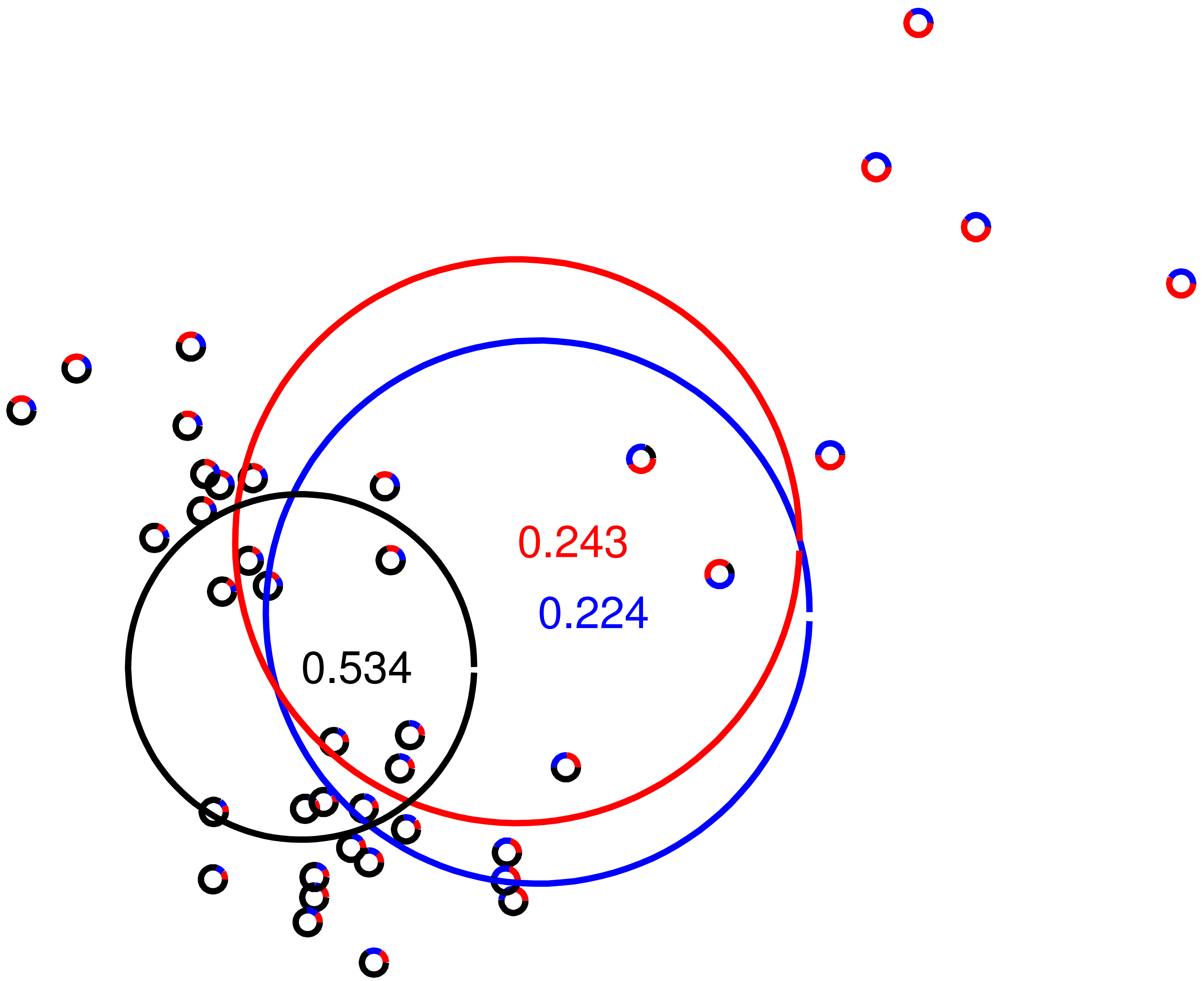
CSAIL

Mixture of Gaussians example

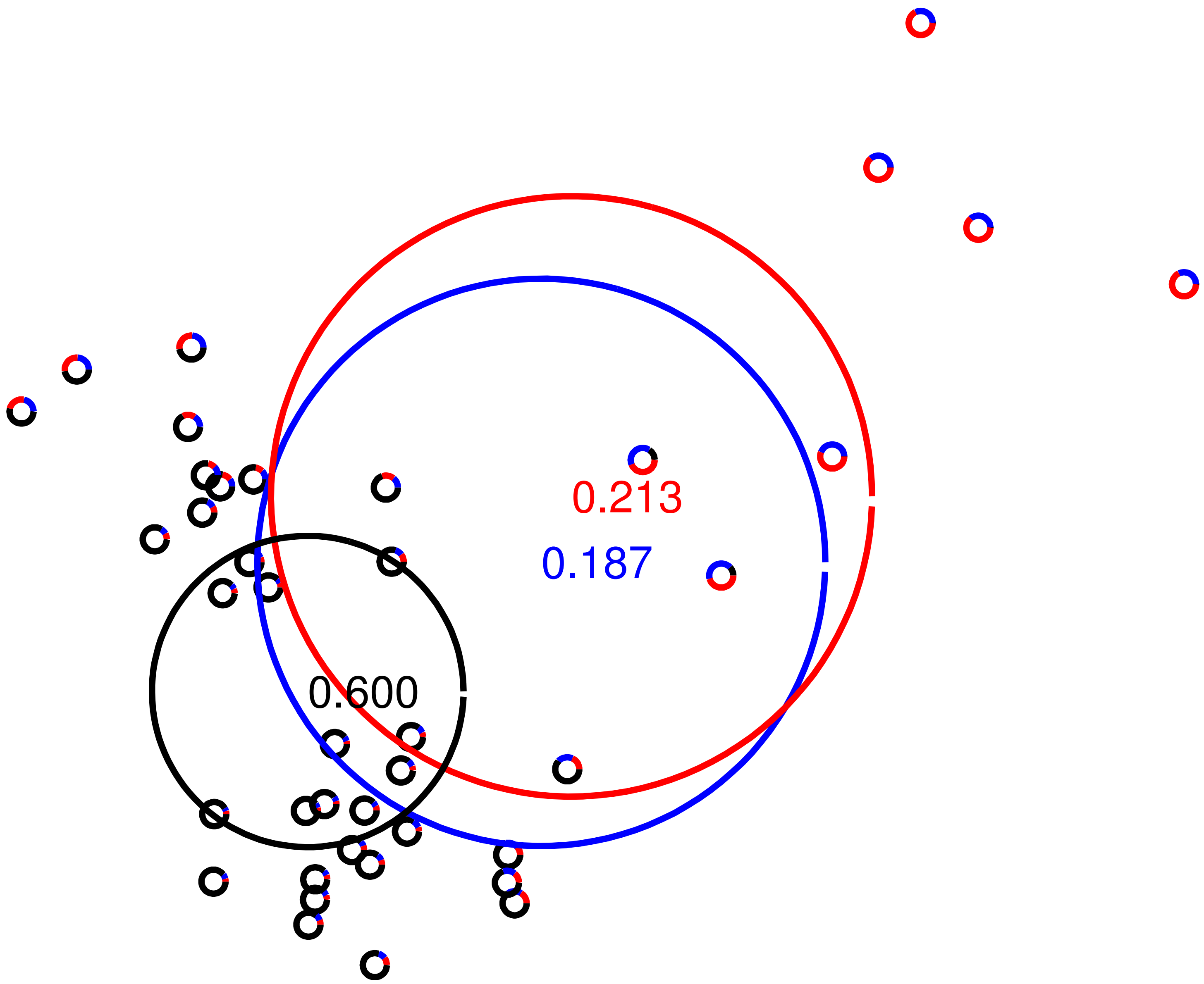
- initial 3-component mixture



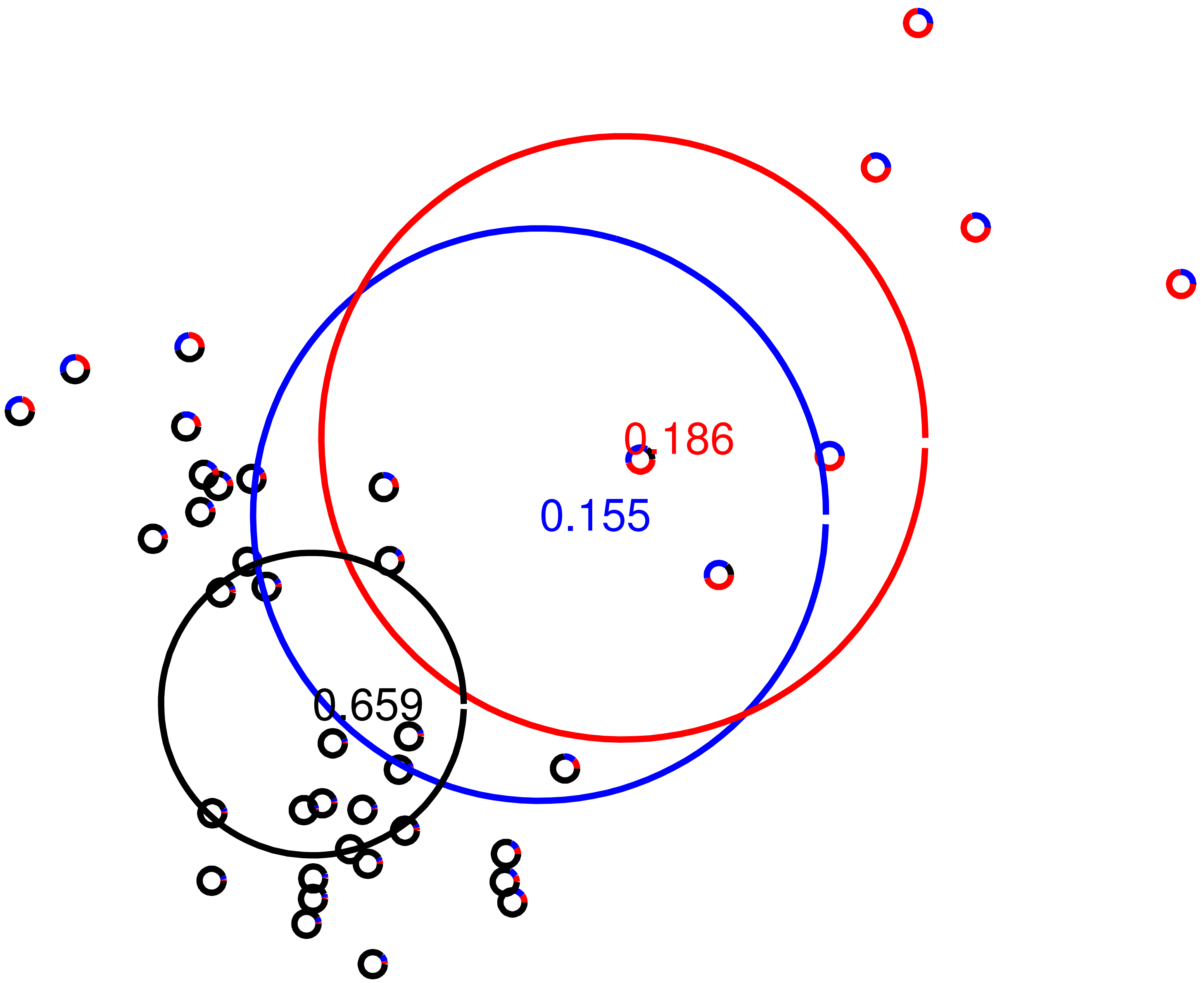
Mixture of Gaussians example



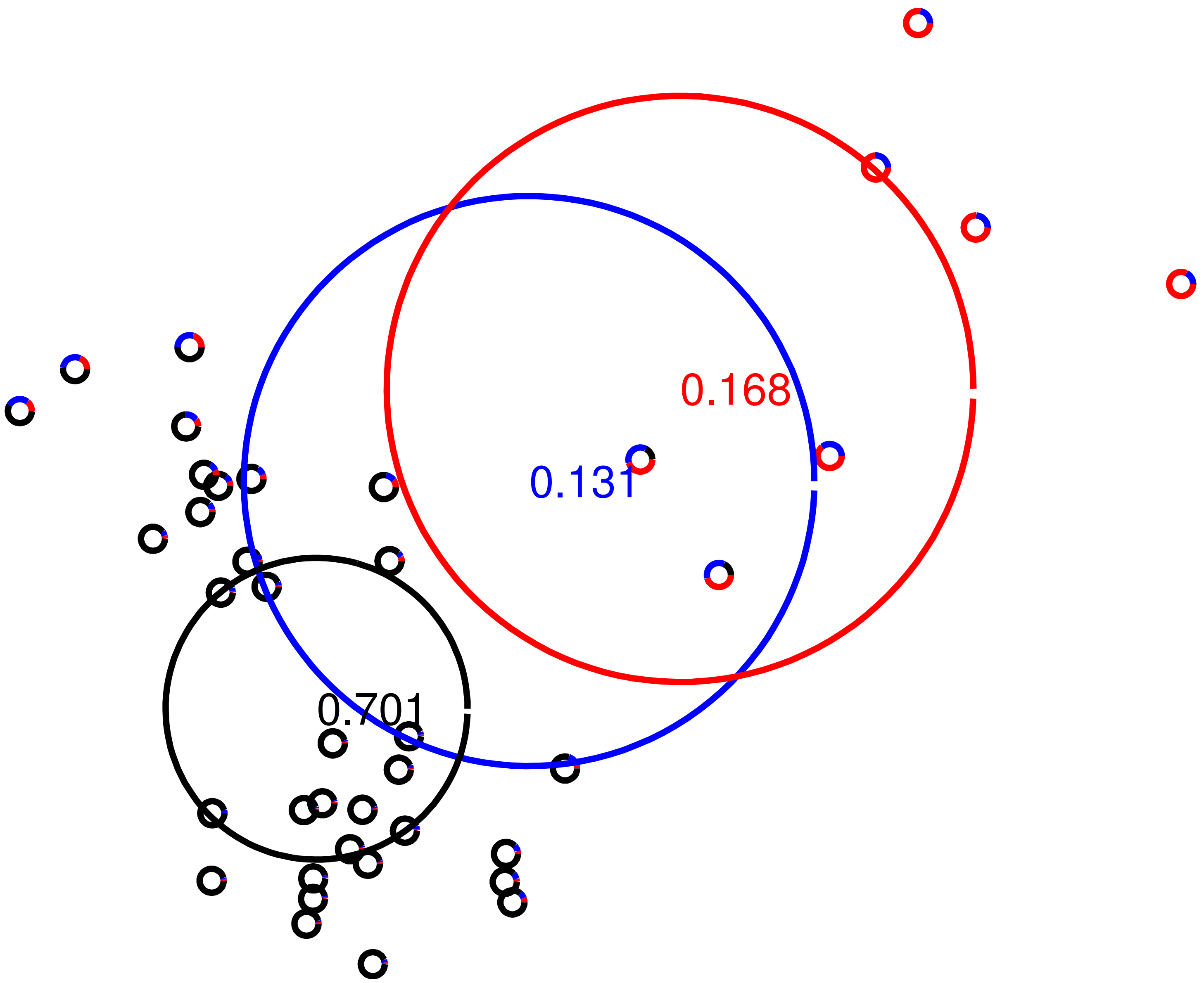
Mixture of Gaussians example



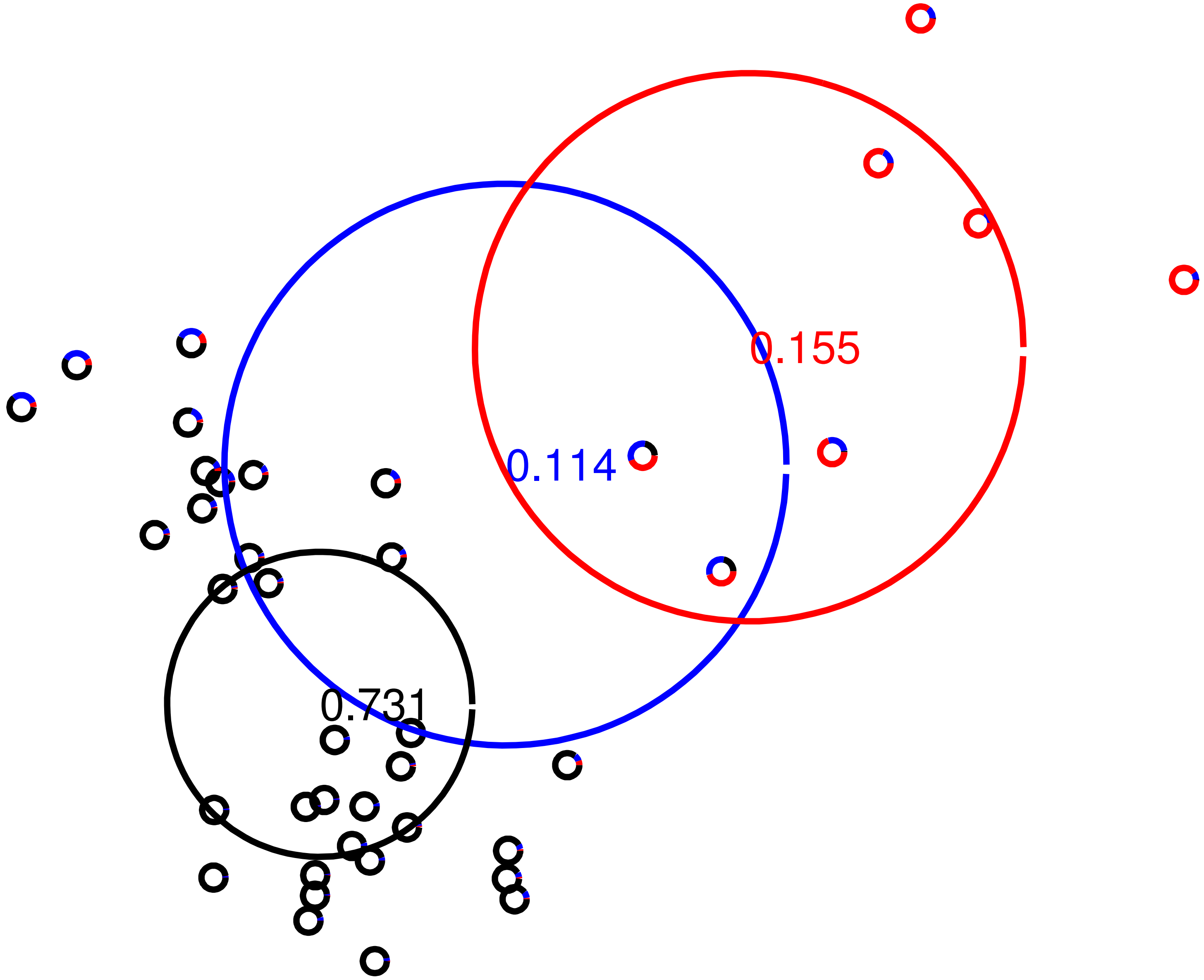
Mixture of Gaussians example



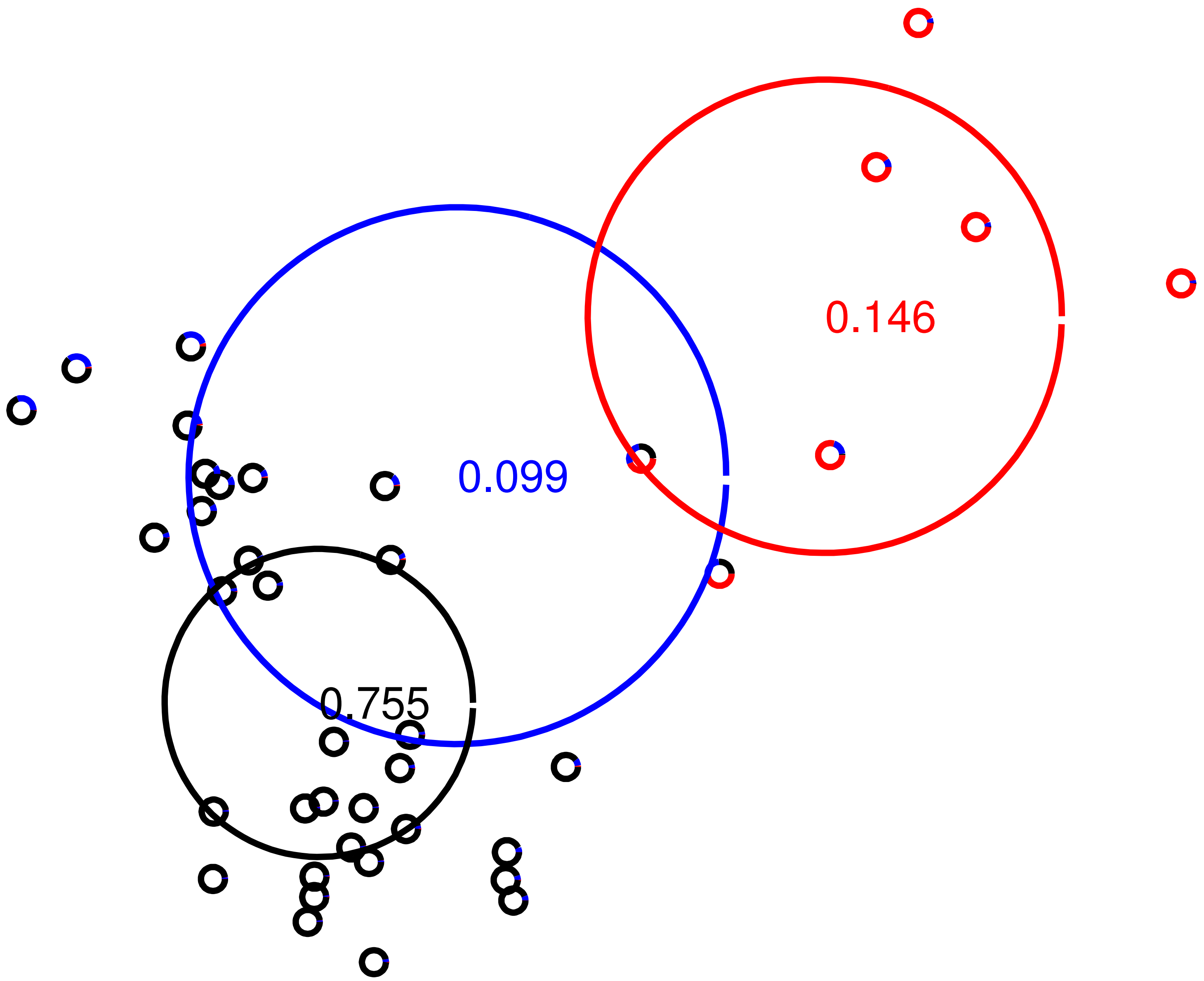
Mixture of Gaussians example



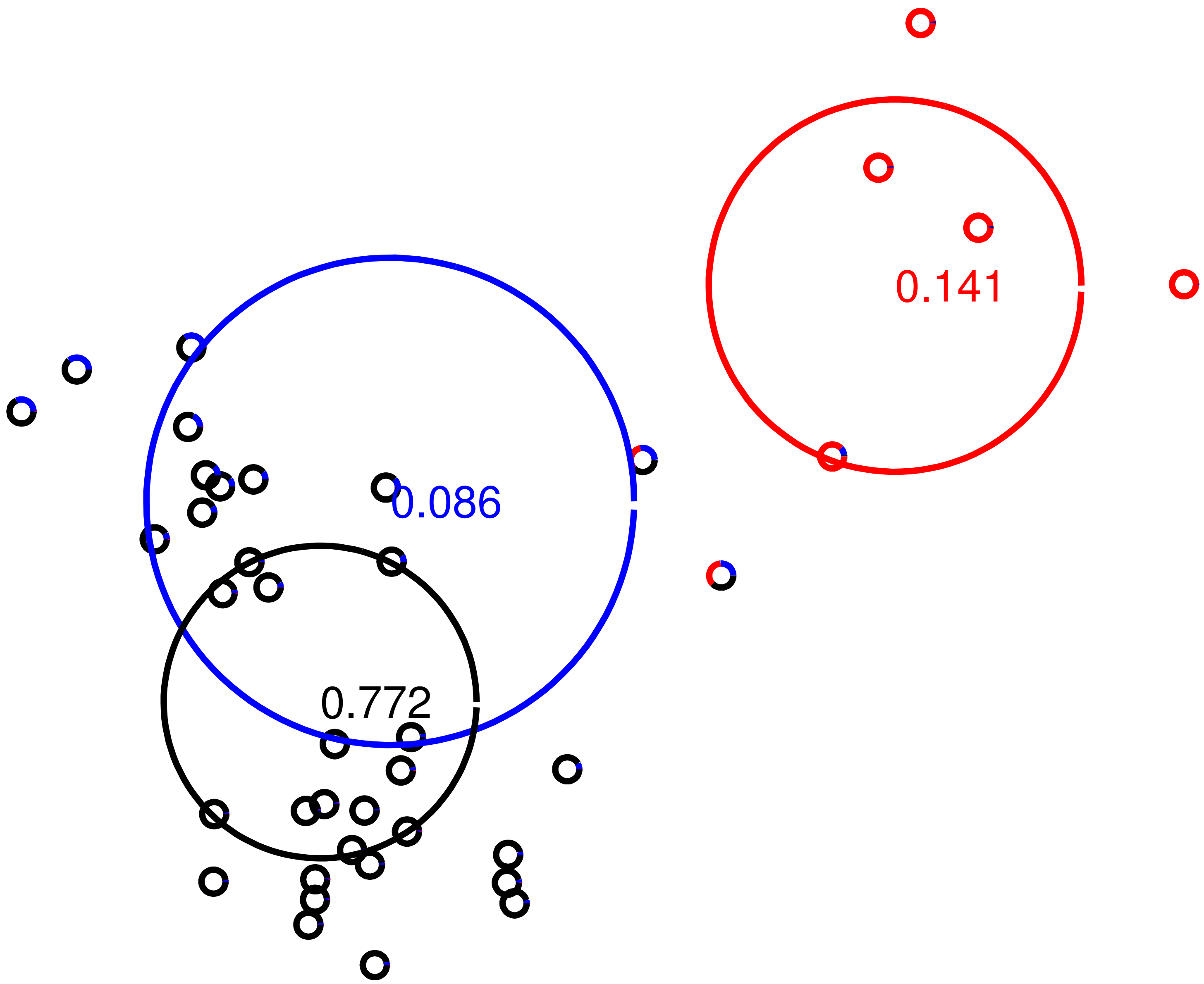
Mixture of Gaussians example



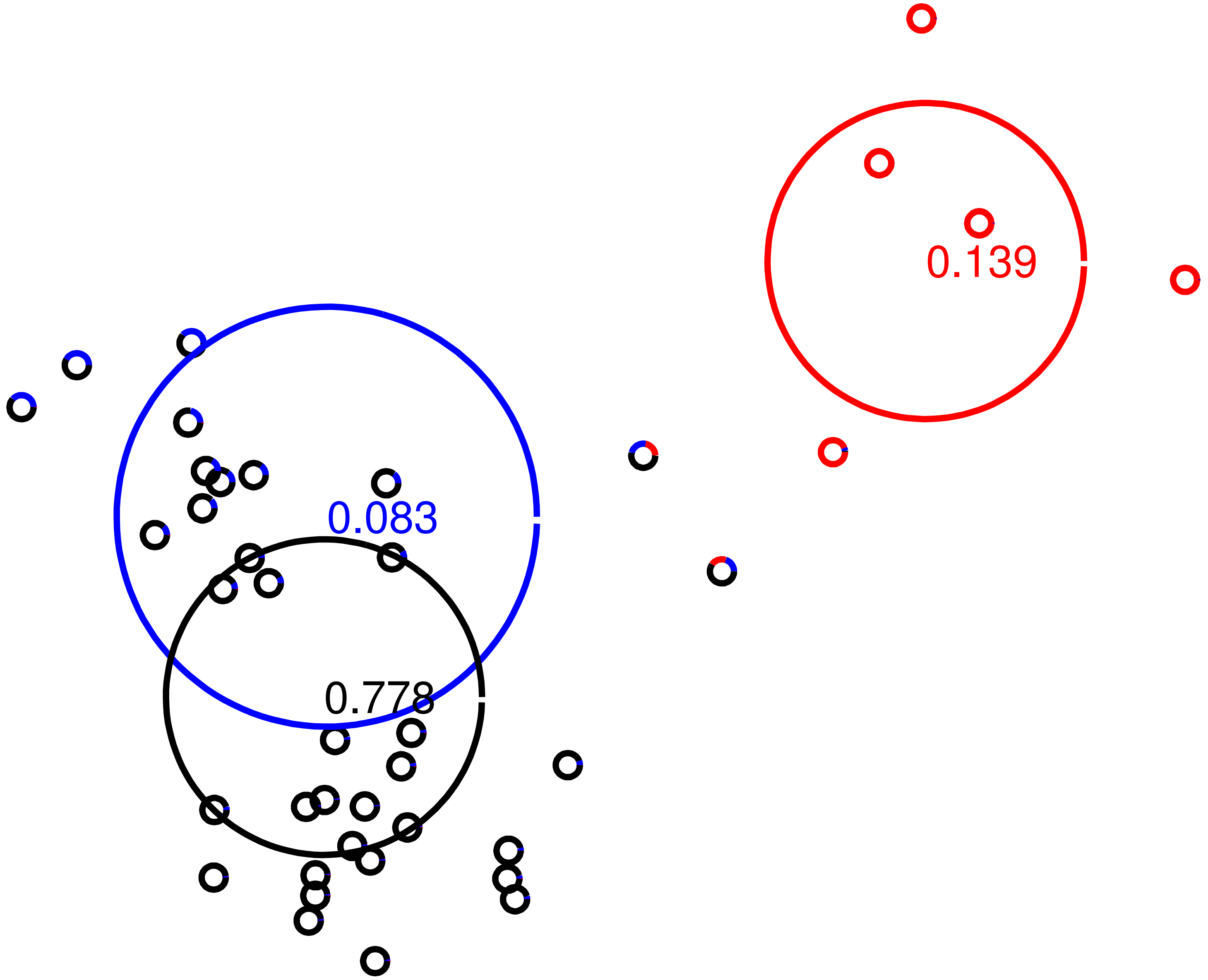
Mixture of Gaussians example



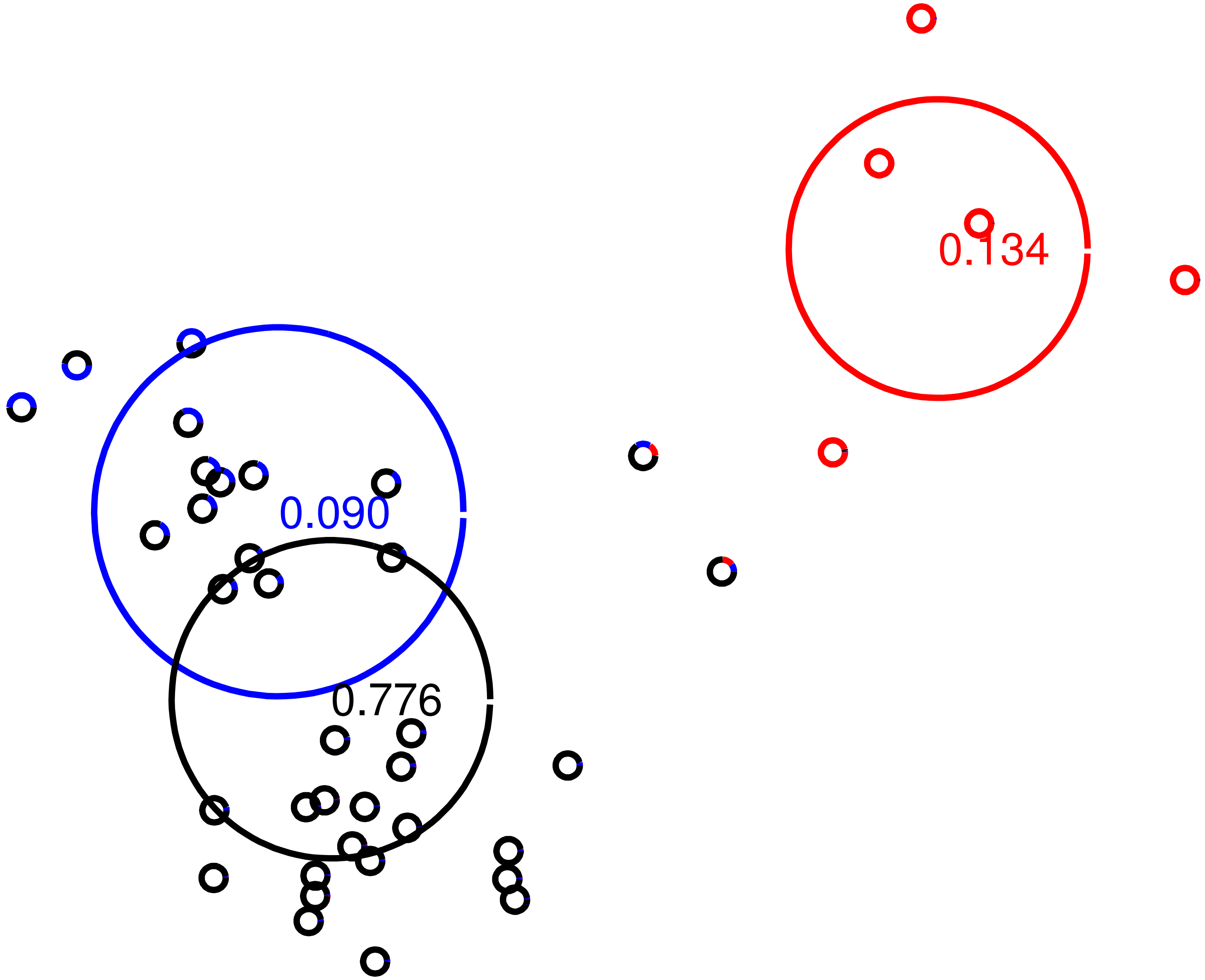
Mixture of Gaussians example



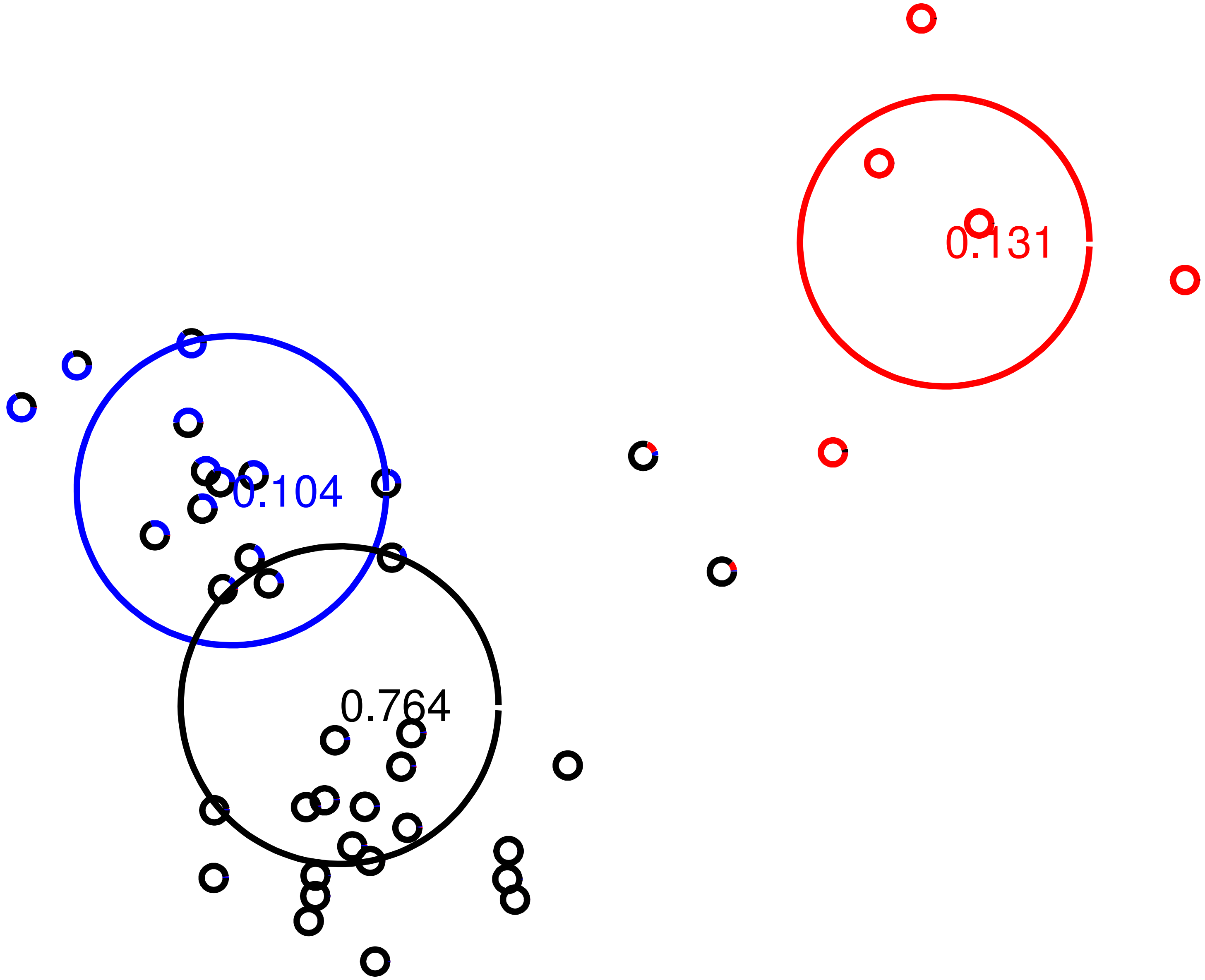
Mixture of Gaussians example



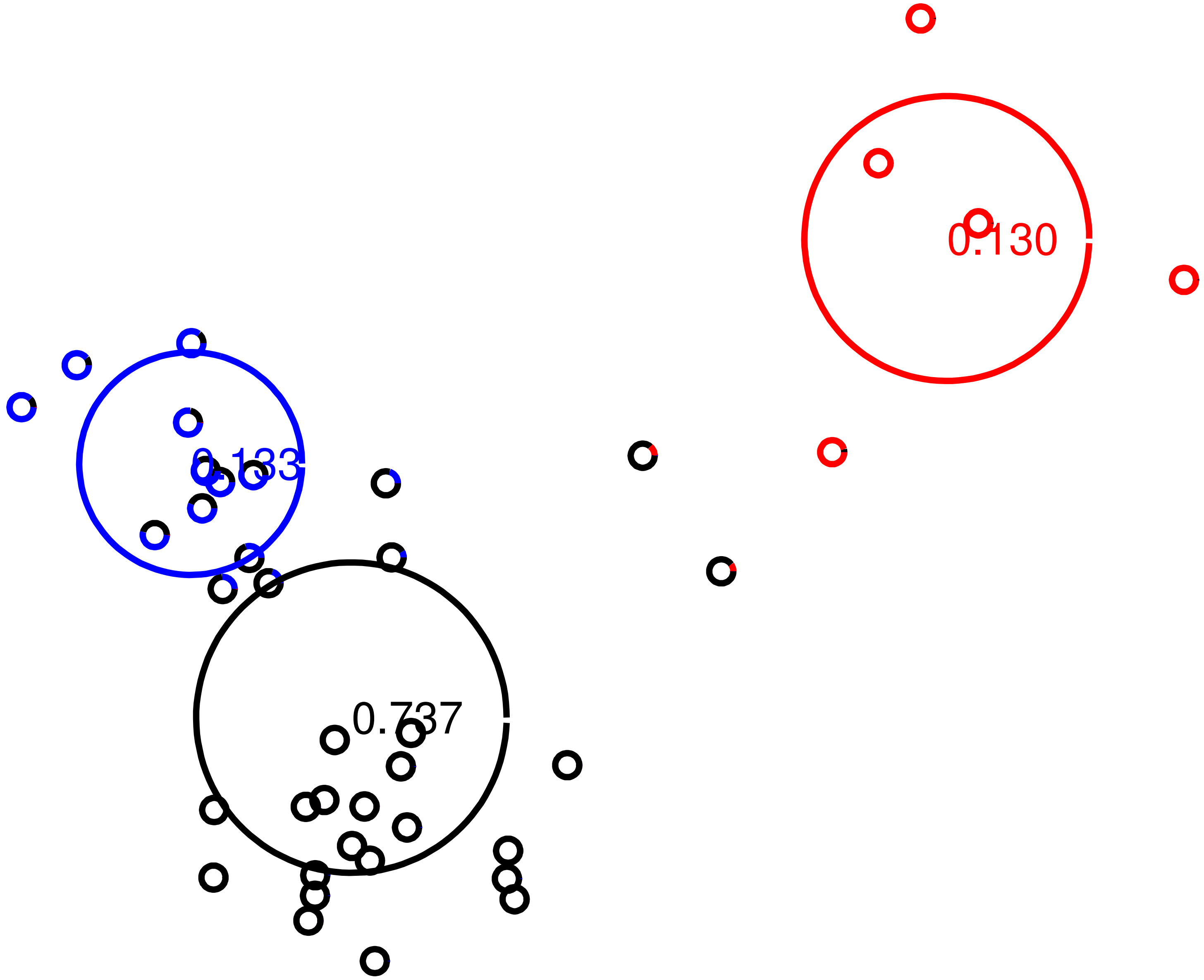
Mixture of Gaussians example



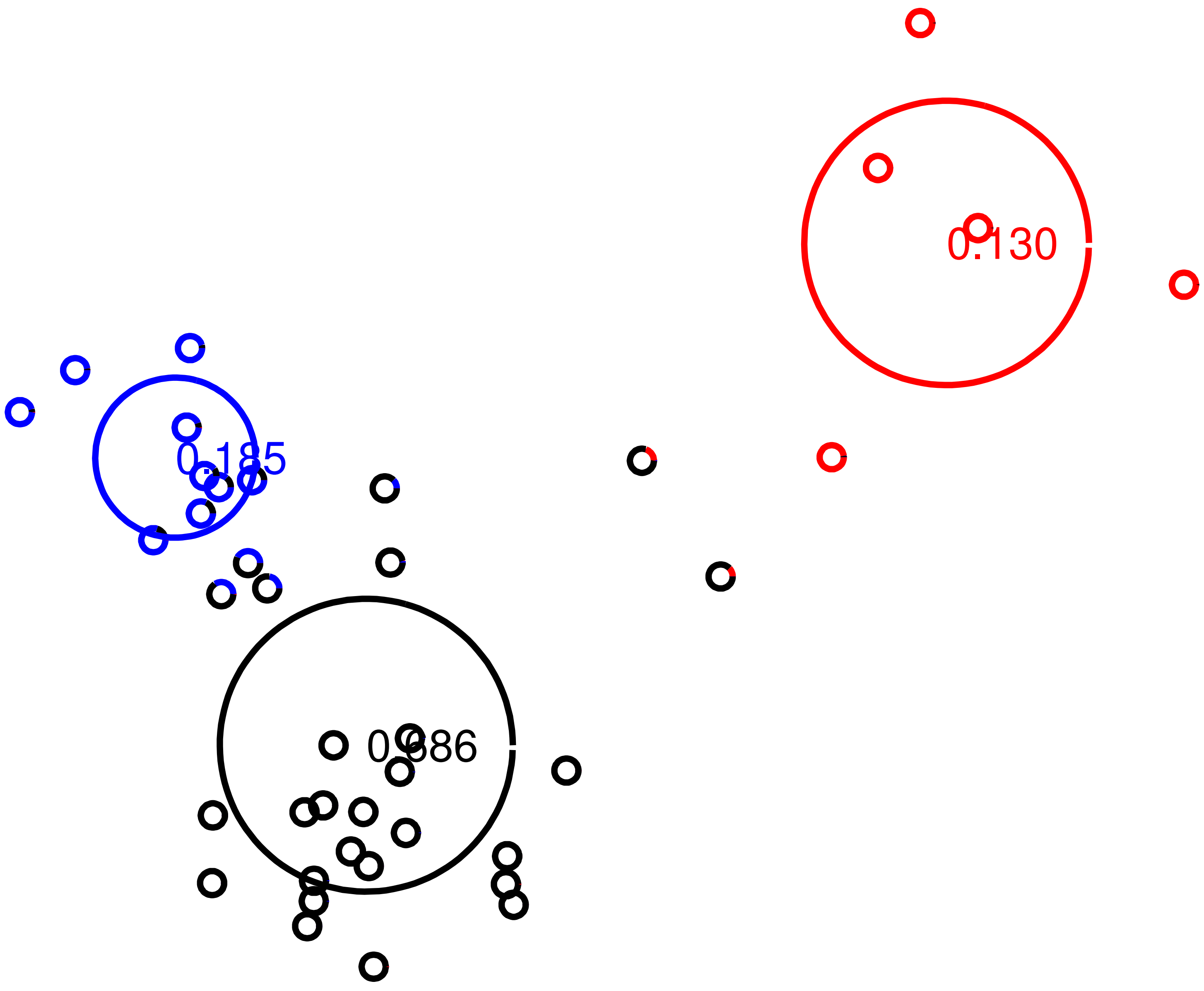
Mixture of Gaussians example



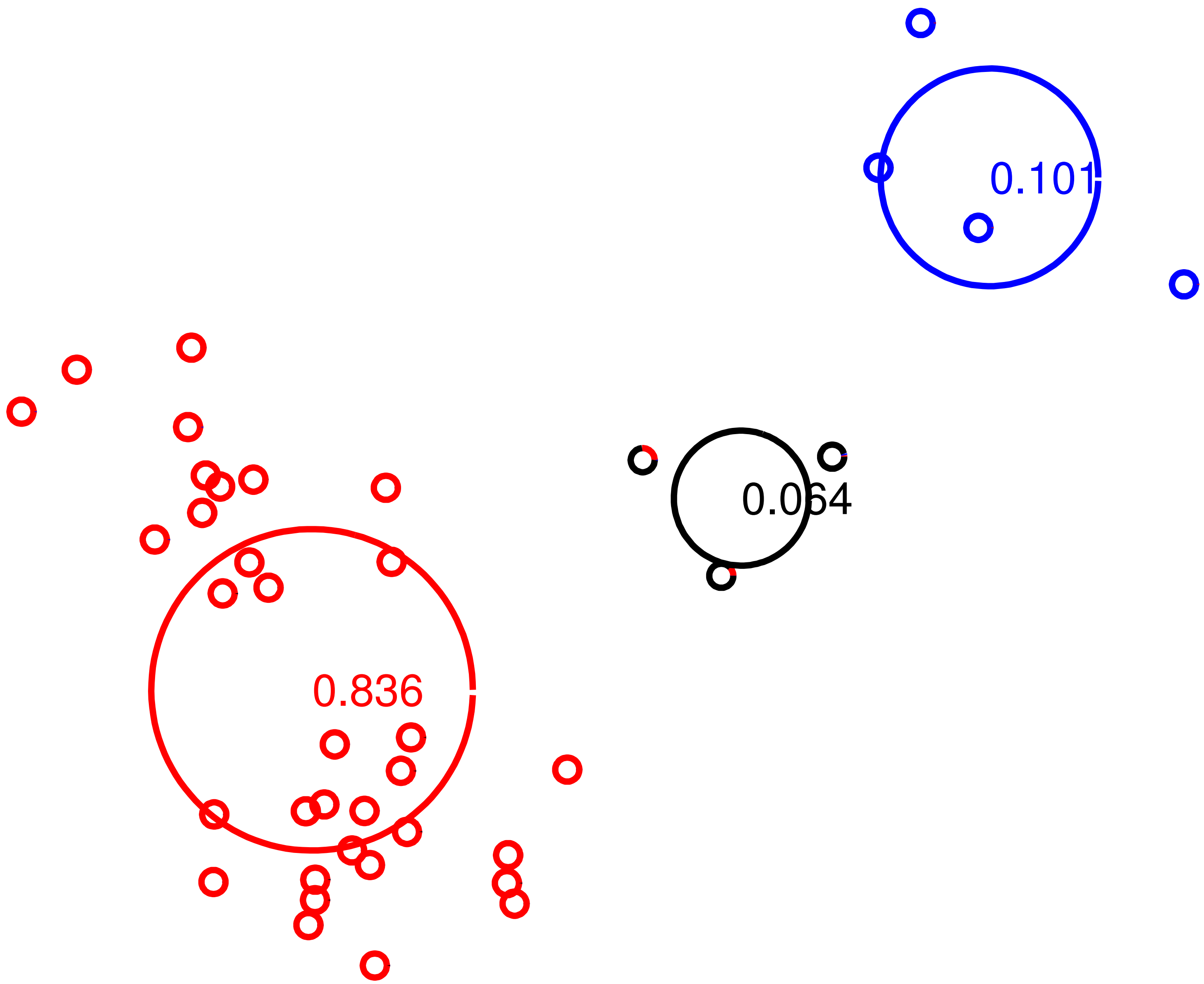
Mixture of Gaussians example



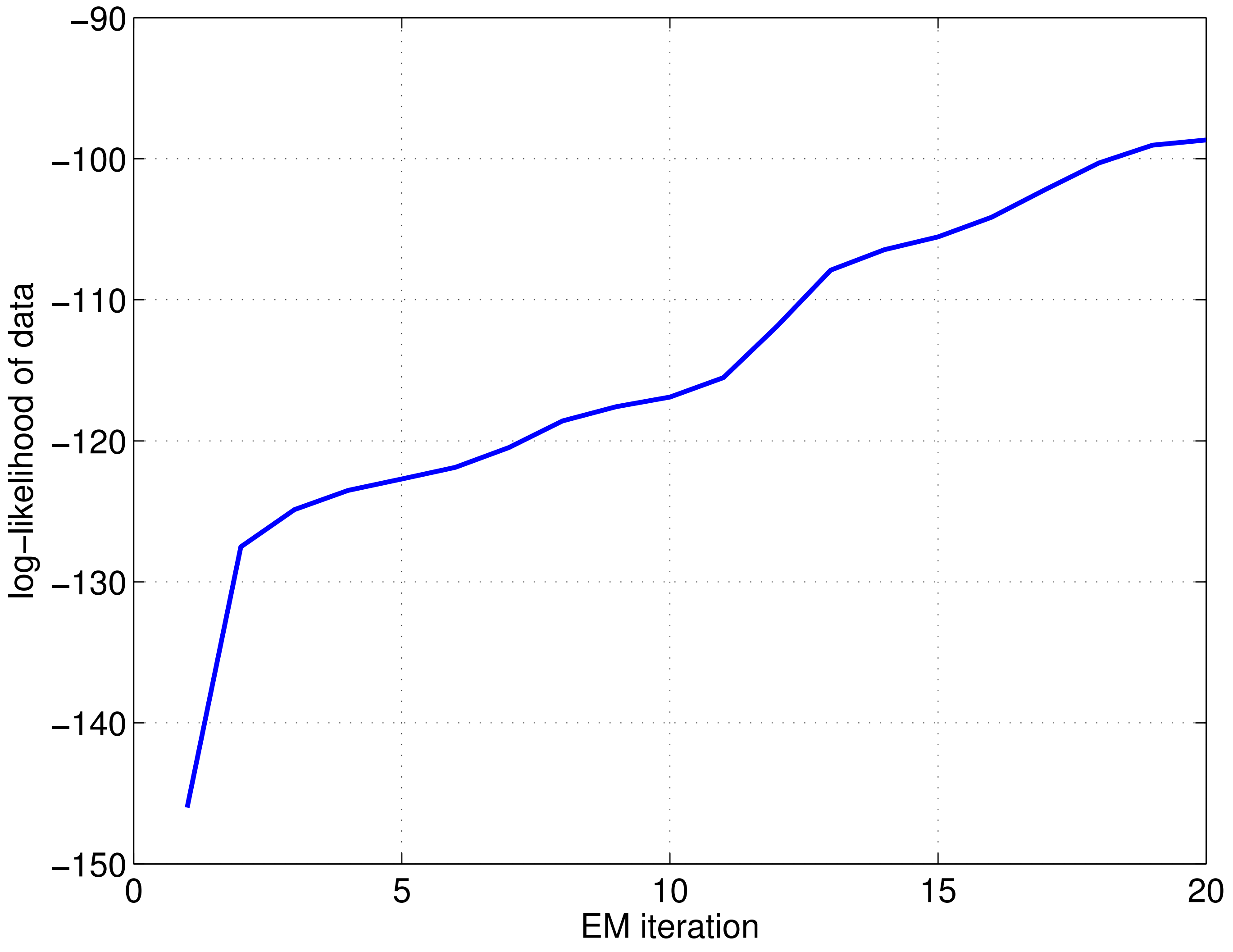
Mixture of Gaussians example



Doesn’t always work… well



The EM algorithm

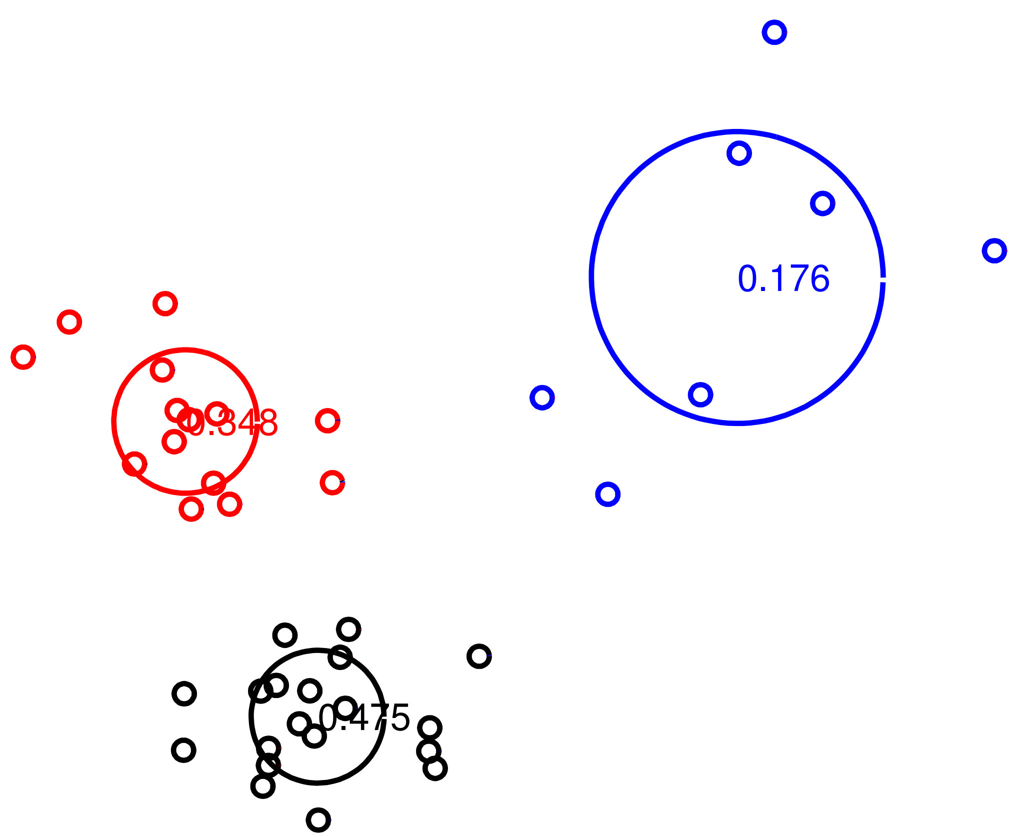


- The EM-algorithm monotonically increases the log-likelihood of the training data (cf. K-means)

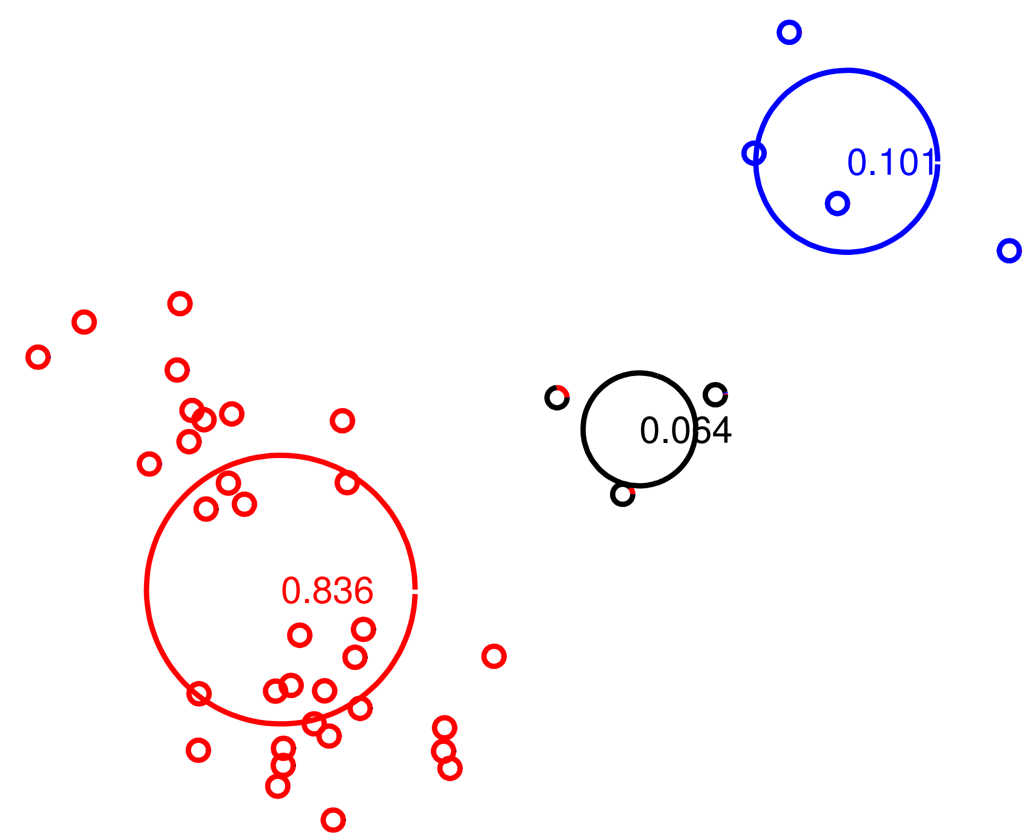
$$l(D;\theta) < l(D;\theta') < l(D;\theta'') < \dots$$

Locally optimal solutions

- The EM-algorithm is guaranteed to find a locally optimal solution by monotonically increasing the log-likelihood (the estimation problem with respect to θ is typically not convex)
- Whether the algorithm converges to the globally optimal solution depends on the initialization



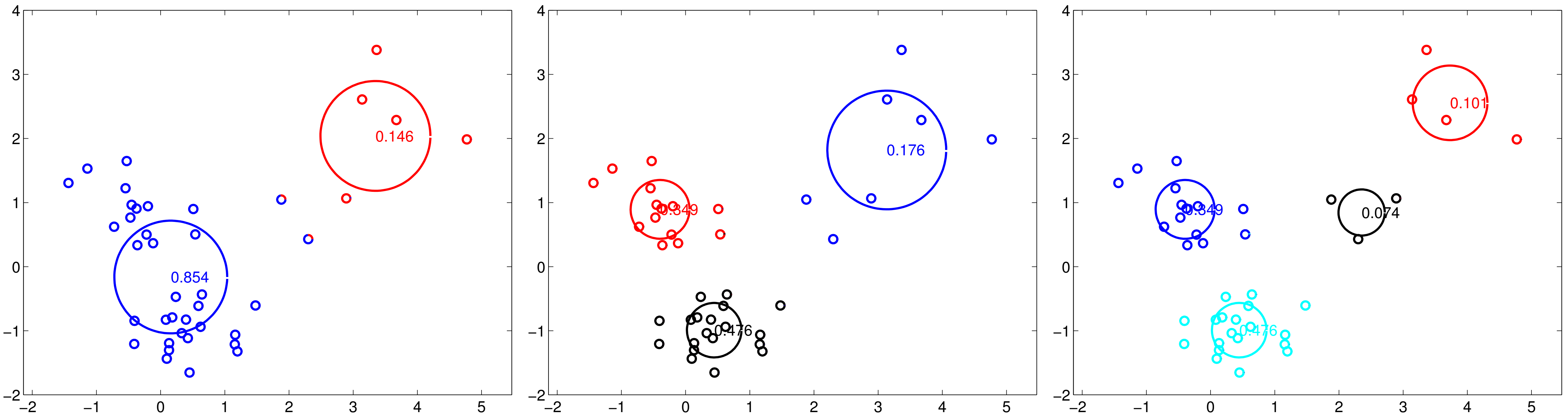
$$l(D;\hat{\theta}) = -98.64$$



$$l(D;\hat{\theta}) = -114.82$$

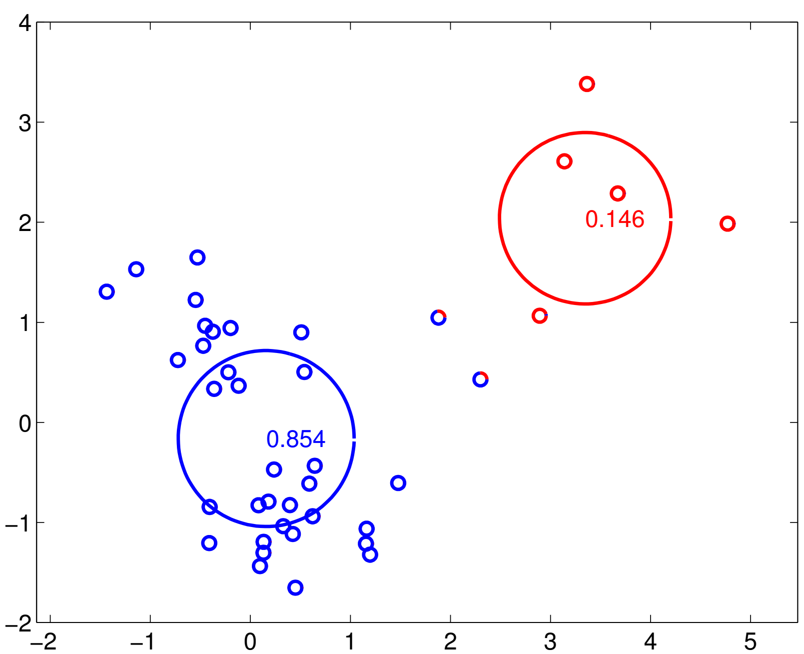
Model selection

- We can run the EM-algorithm with different numbers of components. Need to specify a criterion for selecting among the different models.

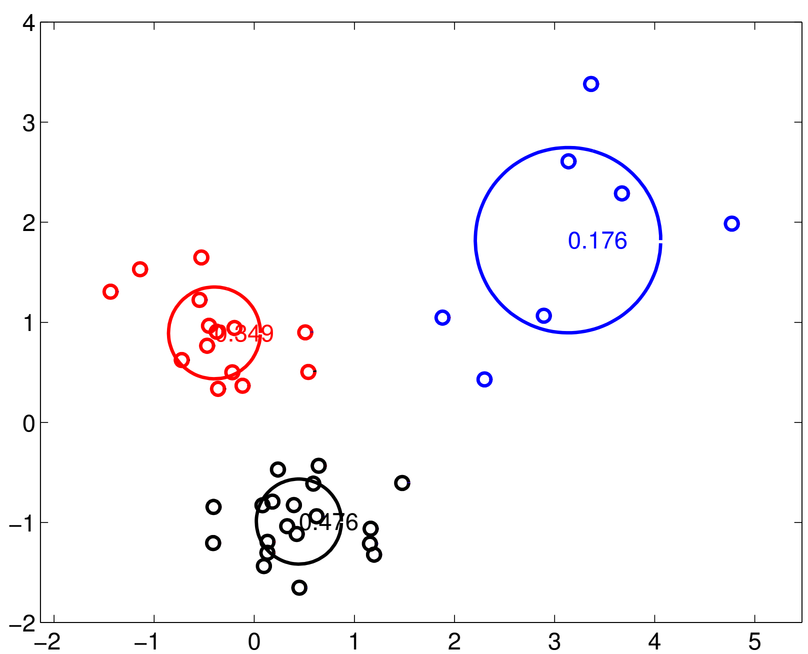


Model selection

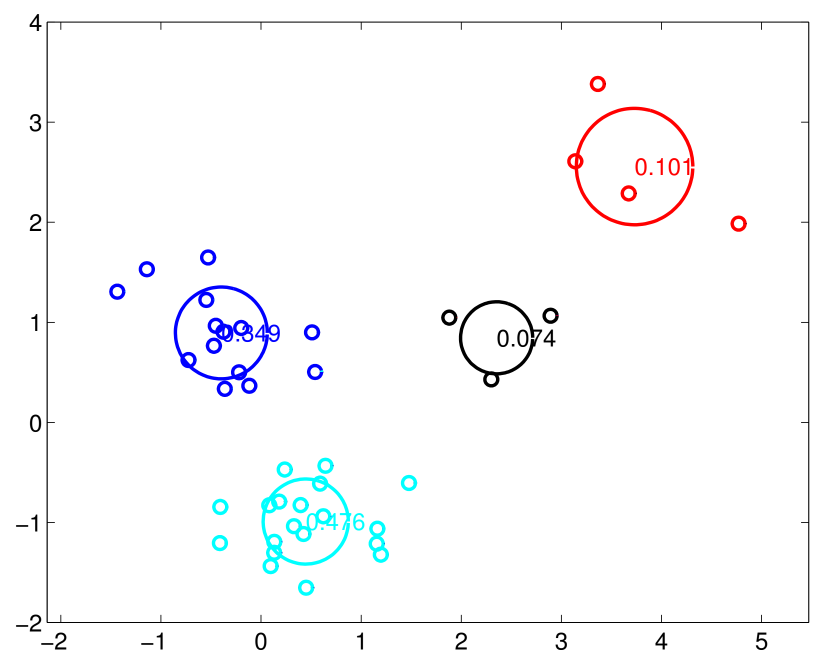
- We can run the EM-algorithm with different numbers of components. Need to specify a criterion for selecting among the different models.



$$l(D; \hat{\theta}) = -118.25$$



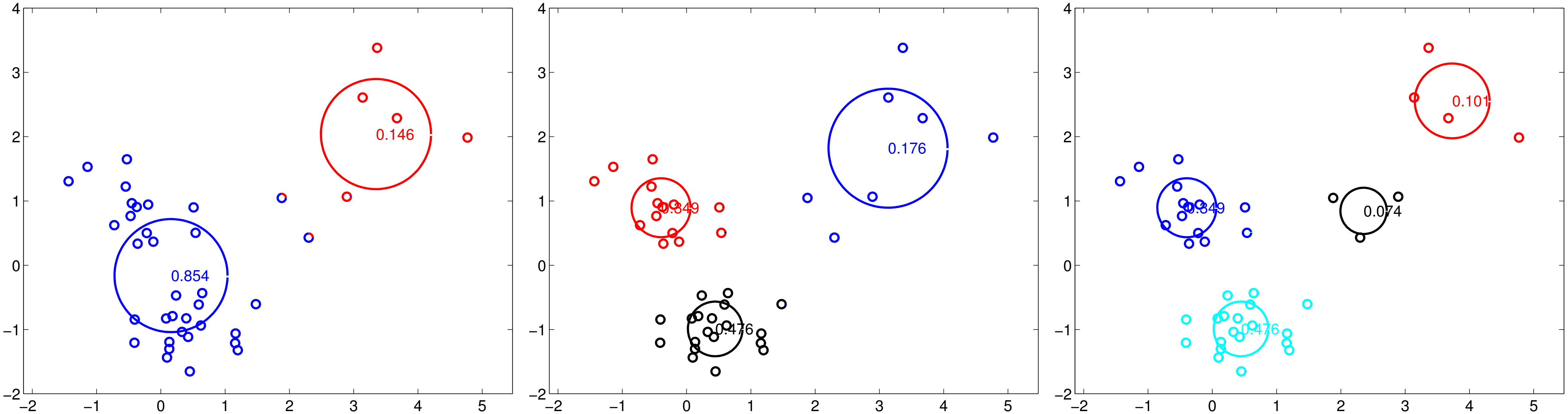
$$l(D; \hat{\theta}) = -98.64$$



$$l(D; \hat{\theta}) = -94.11$$

Model selection

- We can run the EM-algorithm with different numbers of components. Need to specify a criterion for selecting among the different models.



$l(D;\hat{\theta}) = -118.25$ $\quad l(D;\hat{\theta}) = -98.64$ $\quad l(D;\hat{\theta}) = -94.11$

- Basing the selecting on the value of log-likelihood would invariably lead to the largest number of components

Key things to know

- K-means failures
- Mixture model as a latent variable generative model
- Evaluating posterior probabilities
- Mixture estimation
  - ML criterion
  - complete data case
  - EM algorithm
- Why ML cannot be used to select the number of mixture components