Data Mining Techniques

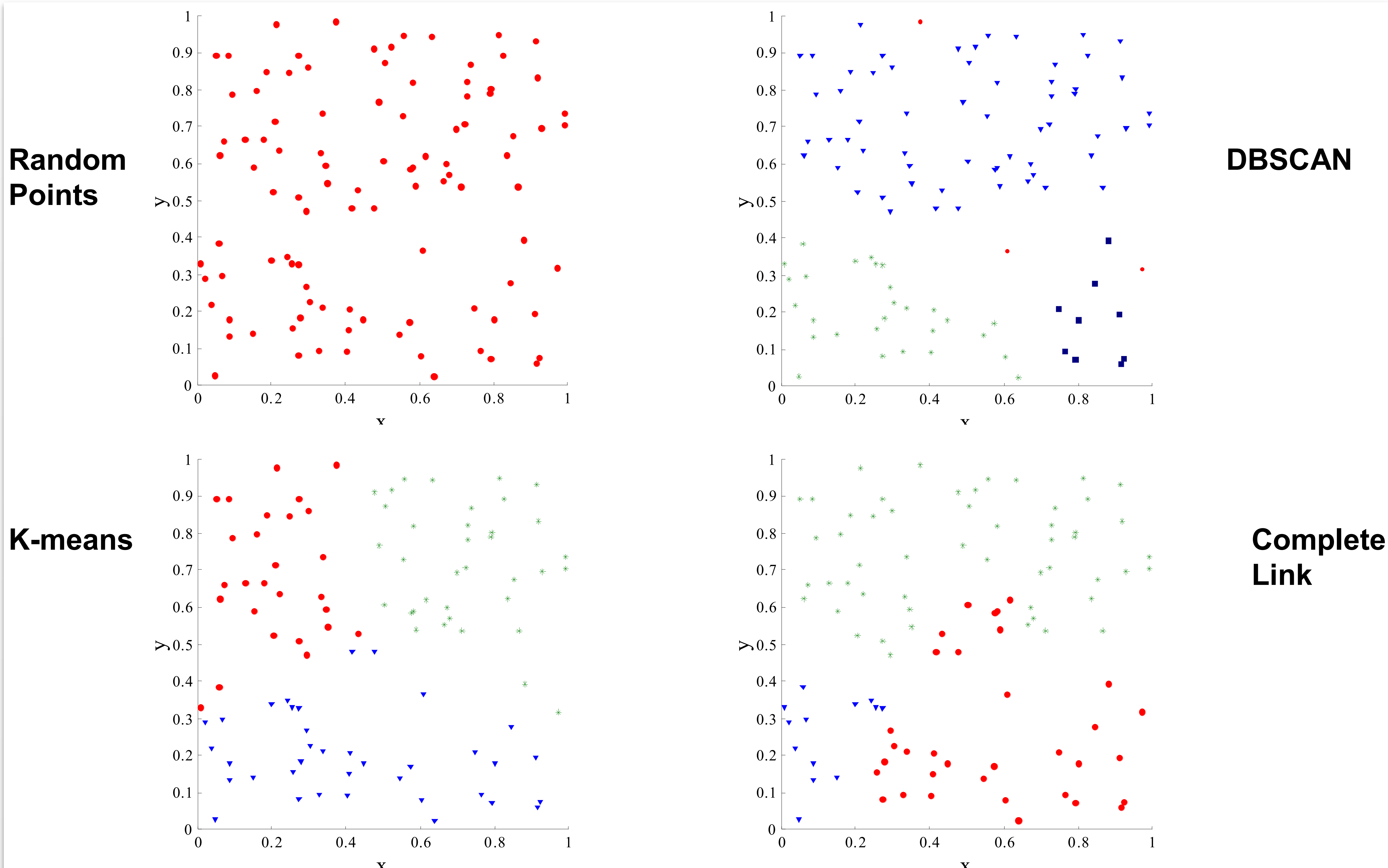
CS 6220 - Section 3 - Fall 2016

Lecture 10

Jan-Willem van de Meent
(*credit*: Yijun Zhao, Chris Bishop,
Andrew Moore, Hastie et al.)

Evaluation of Clustering

Clusters in Random Data



Clustering Criteria

- ***External Quality Criteria***
  - Precision-Recall Measure
  - Mutual Information

- ***Internal Quality Criteria***
  Measure compactness of clusters
  - Sum of Squared Error (SSE)
  - Scatter Criteria

Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Uncorrelated Variables

$$p(a,b) = p(a)p(b)$$

Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Uncorrelated Variables

$$p(a,b) = p(a)p(b)$$

$$I(A;B) = \sum_{a \in A, b \in B} p(a)p(b) \log \frac{p(a)p(b)}{p(a)p(b)} = 0$$

Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Perfectly Correlated Variables

$$p(A = a, B = b) = \delta(a,b) p(B = b)$$

Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Perfectly Correlated Variables

$$p(A=a, B=b) = \delta(a,b) p(B=b)$$

$$I(A;B) = \sum_{a \in A, B \in B} p(A=a, B=b) \log \frac{p(A=a, B=b)}{p(A=a)p(B=b)}$$

Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Perfectly Correlated Variables

$$p(A = a, B = b) = \delta(a,b) p(B = b)$$

$$I(A;B) = \sum_{b \in B} p(B = b) \log \frac{p(B = b)}{p(A = b) p(B = b)}$$

Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Perfectly Correlated Variables

$$P(A=k) = \sum_{l \in B} \delta(k,l) p(B=l) = p(B=k)$$

$$I(A;B) = \sum_{b \in B} p(B=b) \log \frac{p(B=b)}{p(A=b)p(B=b)}$$

Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Perfectly Correlated Variables

$$P(A=k) = \sum_{l \in B} \delta(k,l) p(B=l) = p(B=k)$$

$$I(A;B) = \sum_{b \in B} p(B=b) \log \frac{p(B=b)}{p(B=b)p(B=b)}$$

Mutual Information (External)

$$I(A;B) = \sum_{a\in A, b\in B} p(a,b)\log\frac{p(a,b)}{p(a)p(b)}$$

Perfectly Correlated Variables

$$P(A=k) = \sum_{l\in B}\delta(k,l)p(B=l) = p(B=k)$$

$$I(A;B) = -\sum_{b\in B} p(b)\log p(b) = H(B)$$

Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

y_n: True class label for example n

z_n: Clustering label for example n

Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

y_n: True class label for example n

z_n: Clustering label for example n

$$p(Y=k) = \frac{1}{N}\sum_n I(y_n = k) \qquad p(Z=l) = \frac{1}{N}\sum_n I(z_n = l)$$

$$p(Y=k, Z=l) = \frac{1}{N}\sum_n I(y_n = k \wedge z_n = l)$$

Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

$p(y,z)$	1	2	3	$p(y)$
cat	0.39	0.08	0.02	0.49
dog	0.06	0.31	0.01	0.38
parrot	0.01	0.01	0.11	0.13
$p(z)$	0.46	0.40	0.14	

Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

$p(y,z)$	1	2	3	$p(y)$
cat	0.39	0.08	0.02	0.49
dog	0.06	0.31	0.01	0.38
parrot	0.01	0.01	0.11	0.13
$p(z)$	0.46	0.40	0.14	

What happens to I(Y;Z) if we swap cluster labels?

Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

$p(y,z)$	1	2	3	$p(y)$
cat	0.08	0.39	0.02	0.49
dog	0.31	0.06	0.01	0.38
parrot	0.01	0.01	0.11	0.13
$p(z)$	0.40	0.46	0.14	

What happens to I(Y;Z) if we swap cluster labels?

Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

$p(y,z)$	1	2	3	$p(y)$
cat	0.08	0.39	0.02	0.49
dog	0.31	0.06	0.01	0.38
parrot	0.01	0.01	0.11	0.13
$p(z)$	0.40	0.46	0.14	

Mutual Information is ***invariant*** under label permutations

Scatter Criteria (Internal)

Let $\mathbf{x} = (x_1, \ldots, x_d)^T$

$C_1, \ldots, C_K$ be a clustering of $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Define

- Size of each cluster:

$$N_i = |C_i| \qquad i = 1, 2, \ldots, K$$

- Mean for each cluster:

$$\mu_i = \frac{1}{N_i} \sum_{x \in C_i} \mathbf{x} \qquad i = 1, 2, \ldots, K$$

- Total mean :

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \quad \text{OR} \quad \mu = \frac{1}{N} \sum_{i=1}^{K} N_i \mu_i$$

Scatter Criteria (Internal)

- Scatter matrix for the i^{th} cluster:

$$S_i = \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T \quad \text{(outer product)}$$

- Within cluster scatter matrix :

$$S_W = \sum_{i=1}^{K} S_i$$

- Between cluster scatter matrix :

$$S_B = \sum_{i=1}^{K} N_i(\mu_i - \mu)(\mu_i - \mu)^T \quad \text{(outer product)}$$

Scatter Criteria (Internal)

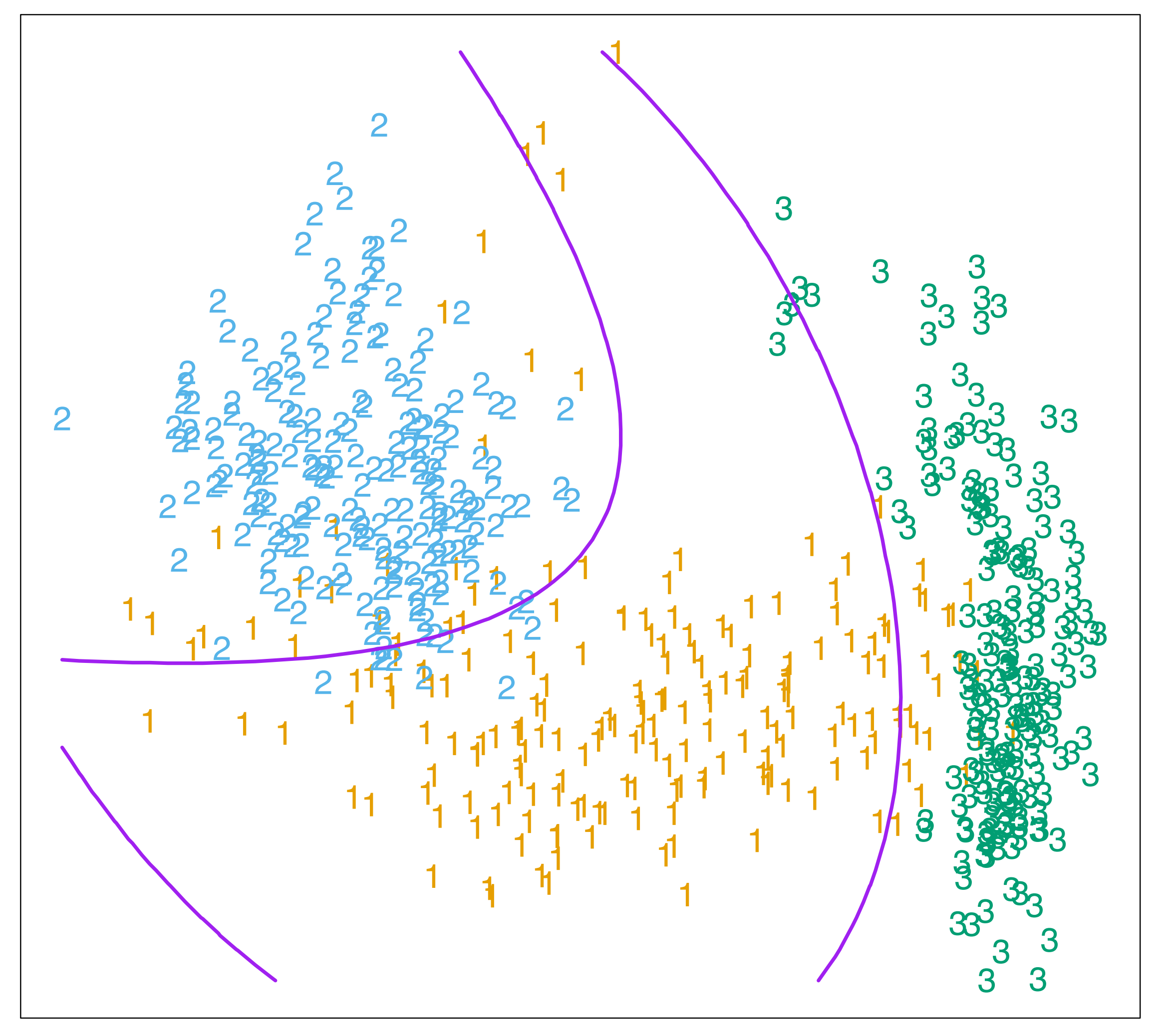
- The trace criteria: sum of the diagonal elements of a matrix
- A good partition of the data should have:
  - Low $tr(S_W)$: similar to minimizing SSE
  - High $tr(S_B)$
  - High $\frac{tr(S_B)}{tr(S_W)}$

Mixture Models

QDA: Gaussian Classification

Classify using posterior

$$y^* = \underset{k}{\operatorname{argmax}}\ p(y = k \mid \boldsymbol{x}, \boldsymbol{\theta})$$



Generative Model

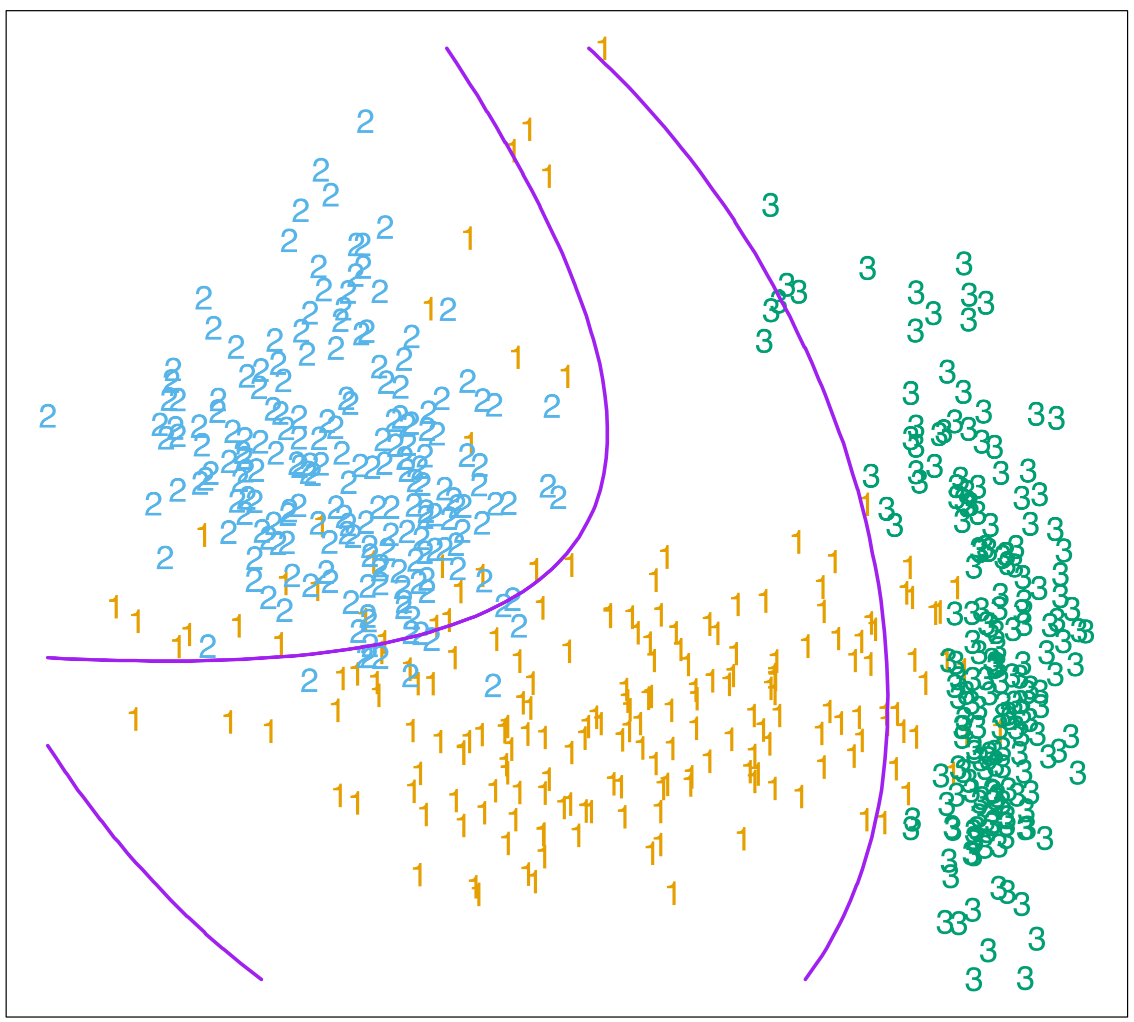
$$y_n \sim \text{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | y_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

(for n = 1, …, N)

QDA: Gaussian Classification

Classify using posterior

$$y^* = \underset{k}{\operatorname{argmax}}\; p(y = k \mid \boldsymbol{x}, \boldsymbol{\theta})$$



Generative Model

$$y_n \sim \text{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | y_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Joint Probability

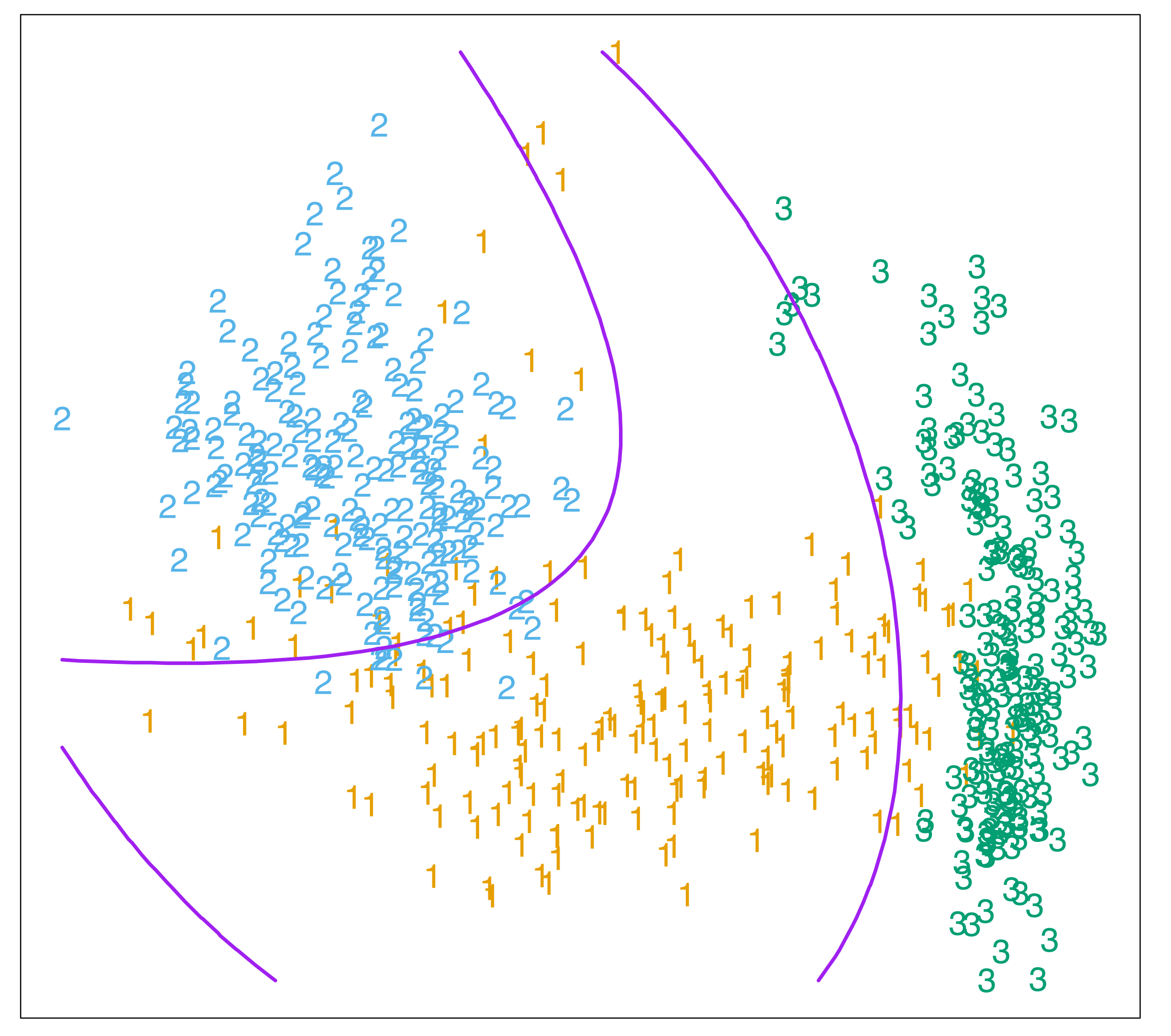
$$\boldsymbol{\theta} := \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$$
$$\boldsymbol{X} := (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top)$$
$$\boldsymbol{y} := (y_1, \ldots, y_N)$$

$$p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{X} \mid \boldsymbol{y}, \boldsymbol{\theta})\, p(\boldsymbol{y} \mid \boldsymbol{\theta})$$

QDA: Gaussian Classification

Classify using posterior

$$y^* = \underset{k}{\operatorname{argmax}}\ p(y = k \mid \boldsymbol{x}, \boldsymbol{\theta})$$



Generative Model

$$y_n \sim \text{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | y_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Joint Probability

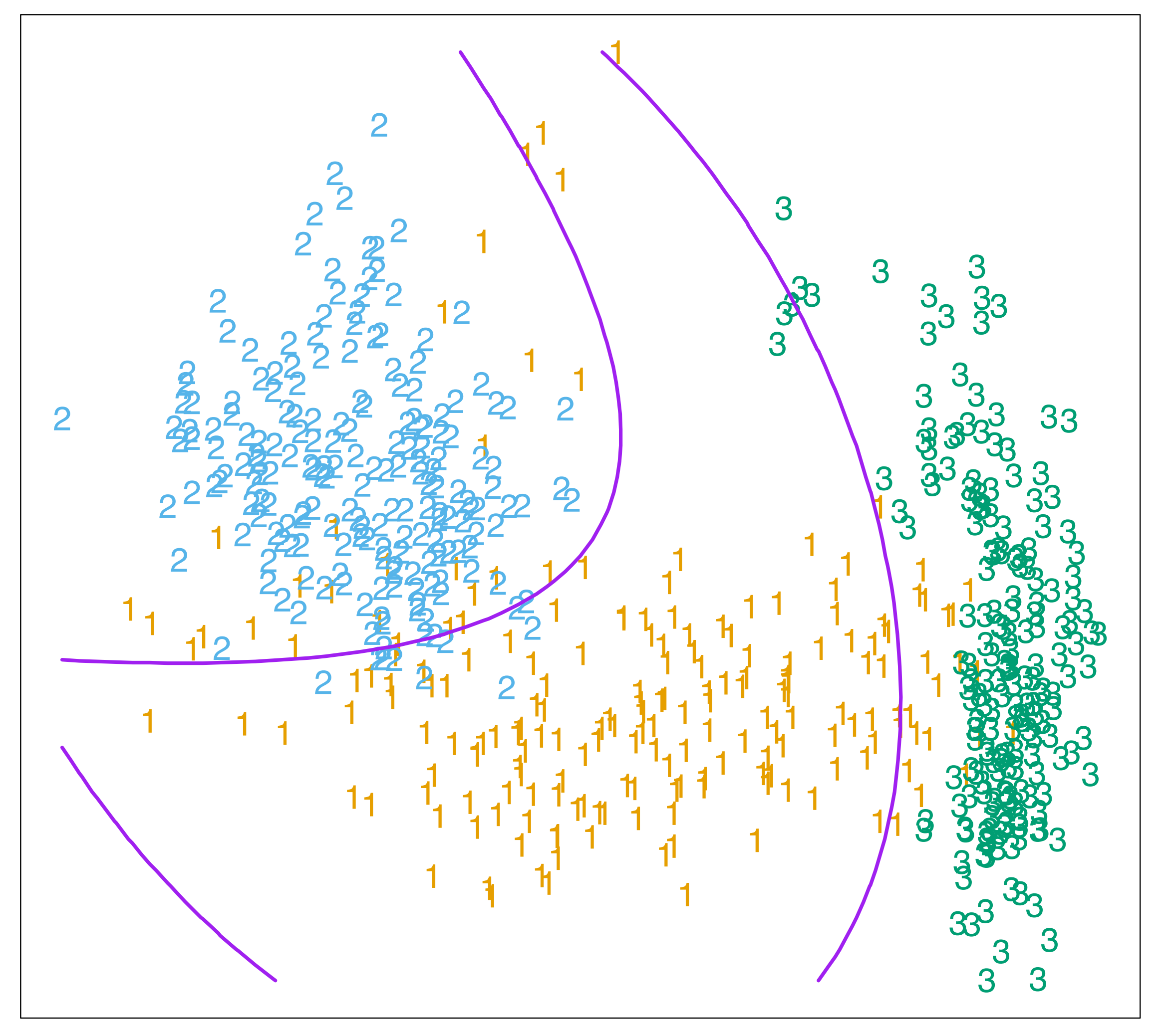
$$\boldsymbol{\theta} := \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$$
$$\boldsymbol{X} := (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top)$$
$$\boldsymbol{y} := (y_1, \ldots, y_N)$$

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta}) &= p(\boldsymbol{X} \mid \boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}) \\ &= p(\boldsymbol{X} \mid \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{y} \mid \boldsymbol{\pi}) \end{aligned}$$

QDA: Gaussian Classification

Classify using posterior / joint

$$y^* = \underset{k}{\operatorname{argmax}}\ p(y = k, \boldsymbol{x} \mid \boldsymbol{\theta})$$



Generative Model

$$y_n \sim \mathrm{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | y_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Joint Probability

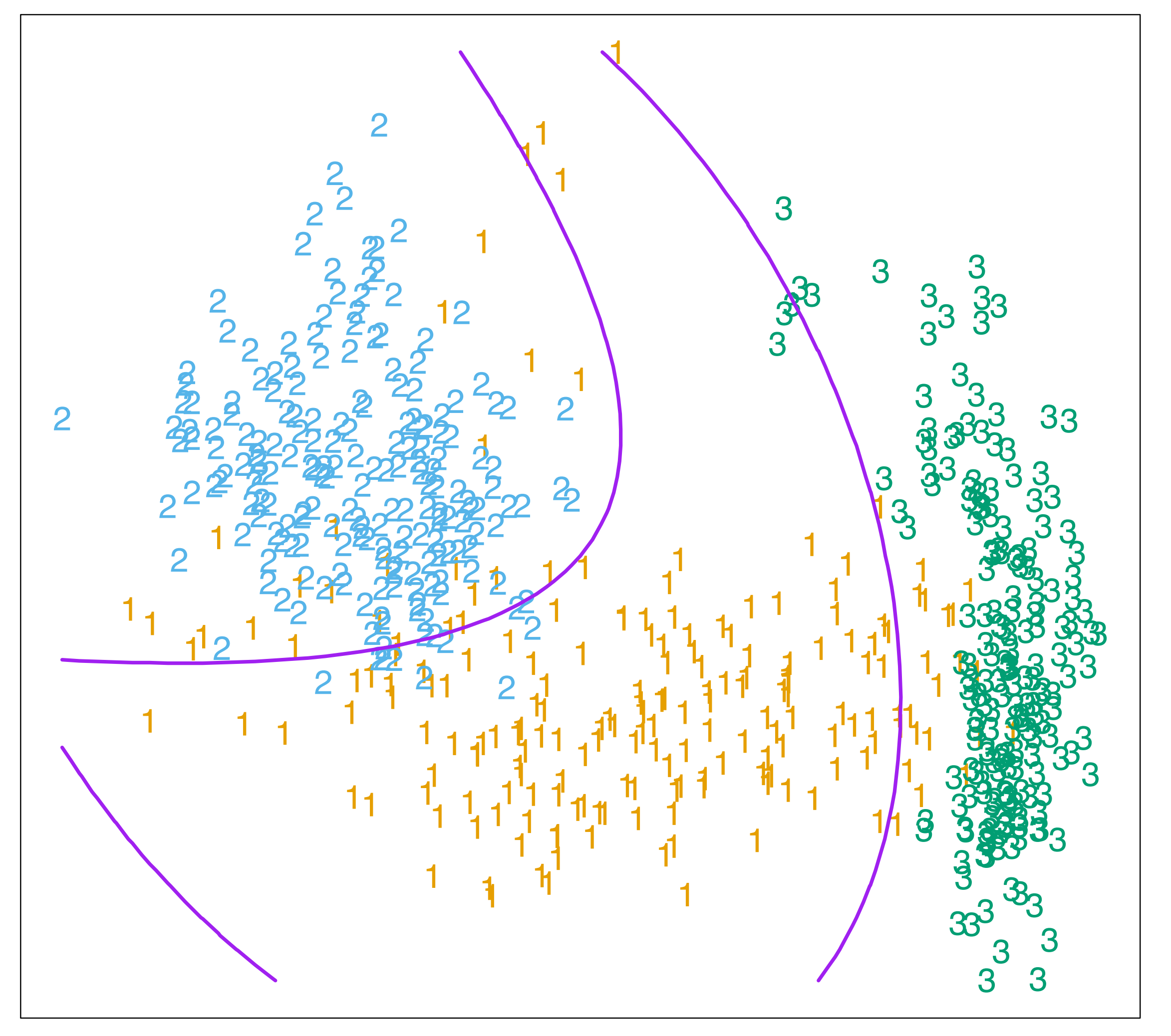
$$\boldsymbol{\theta} := \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$$
$$\boldsymbol{X} := (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top)$$
$$\boldsymbol{y} := (y_1, \ldots, y_N)$$

$$p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{X} \mid \boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta})$$
$$= p(\boldsymbol{X} \mid \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{y} \mid \boldsymbol{\pi})$$

QDA: Gaussian Classification

Classify using posterior / joint

$$y^* = \underset{k}{\operatorname{argmax}}\ p(y=k, \boldsymbol{x} \mid \boldsymbol{\theta}^*)$$



Generative Model

$$y_n \sim \text{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | y_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Use maximum likelihood params

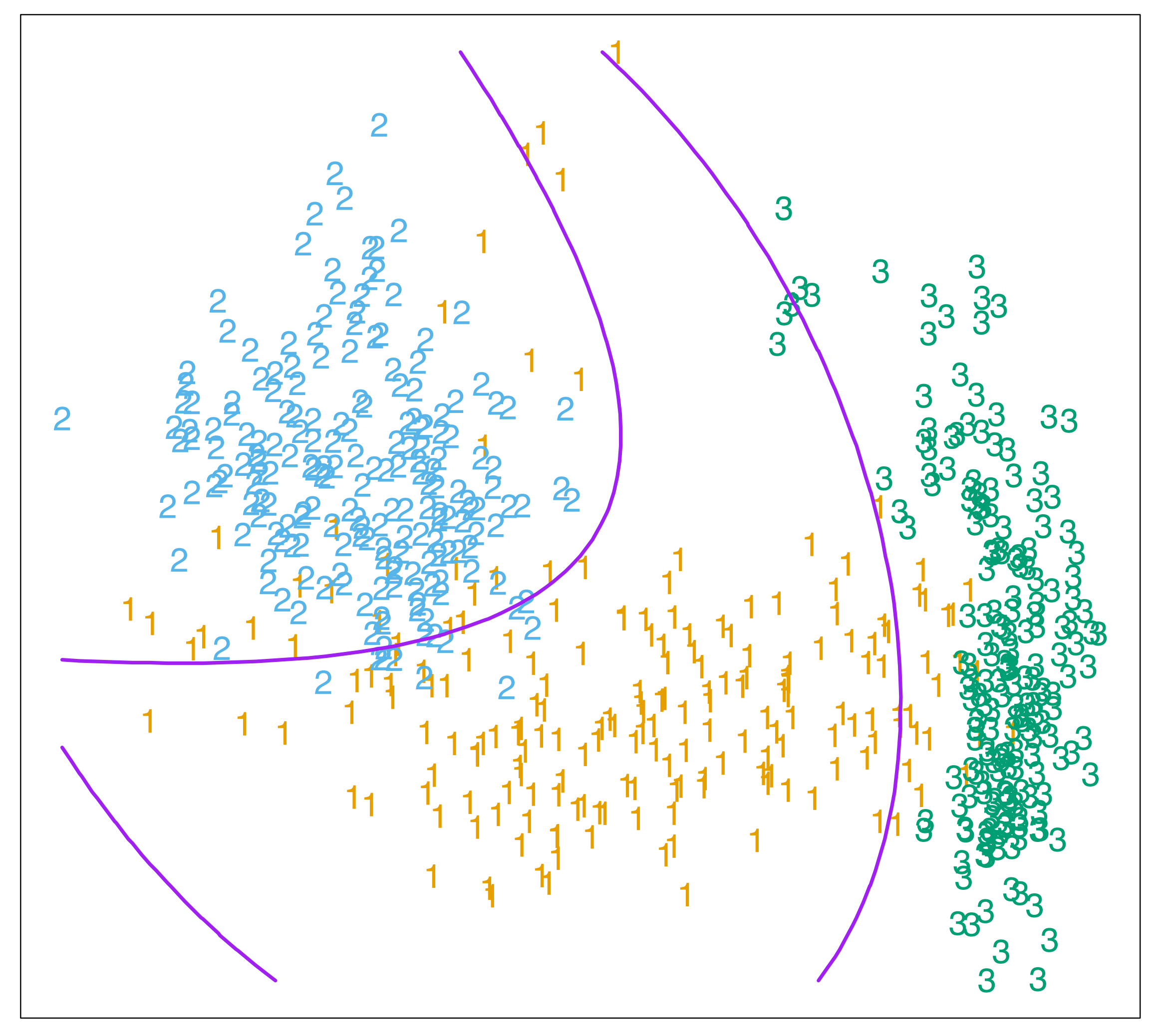
$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta})$$

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta}) &= p(\boldsymbol{X} \mid \boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}) \\ &= p(\boldsymbol{X} \mid \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{y} \mid \boldsymbol{\pi}) \end{aligned}$$

QDA: Gaussian Classification

Classify using posterior / joint

$$y^* = \underset{k}{\operatorname{argmax}}\ p(y{=}k, \boldsymbol{x} \mid \boldsymbol{\theta}^*)$$



Generative Model

$$y_n \sim \mathrm{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | y_n{=}k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Maximum Likelihood Parameters

$$\boldsymbol{\mu}_k^* = \frac{1}{N_k} \sum_{n=1}^{N} I[y_n = k]\, \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k^* = \frac{1}{N_k} \sum_{n=1}^{N} I[y_n = k]\, |\boldsymbol{x}_n - \boldsymbol{\mu}_k|^2$$

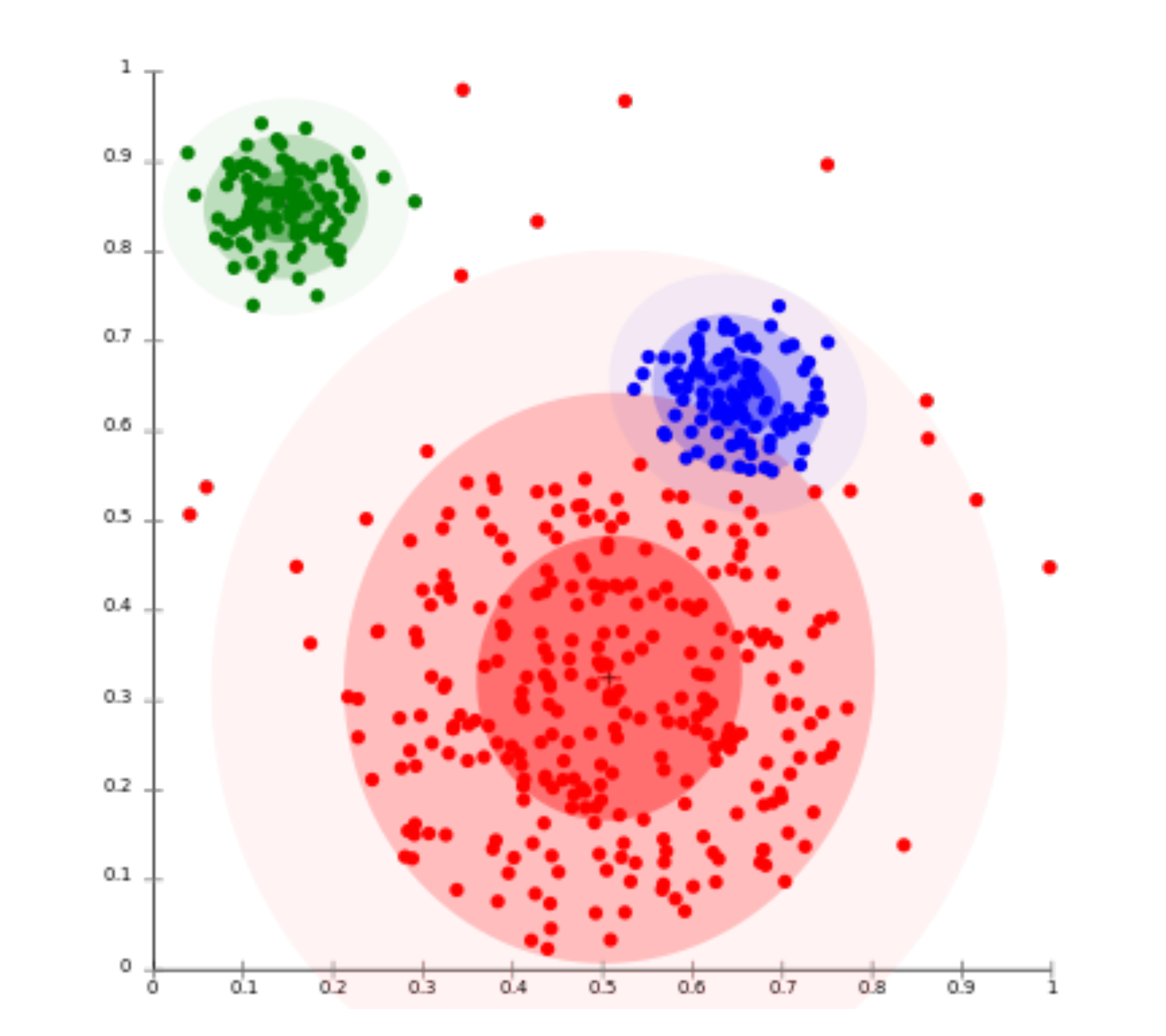
$$\boldsymbol{\pi}^* = (N_1/N, \ldots, N_K/N)$$

$$N_k = \sum_{n=1}^{N} I[y_n = k]$$

Gaussian Clustering

Maximum posterior clustering

$$z_n = \underset{k}{\operatorname{argmax}}\, p(z_n = k \mid \boldsymbol{x}_n, \boldsymbol{\theta})$$



Generative Model

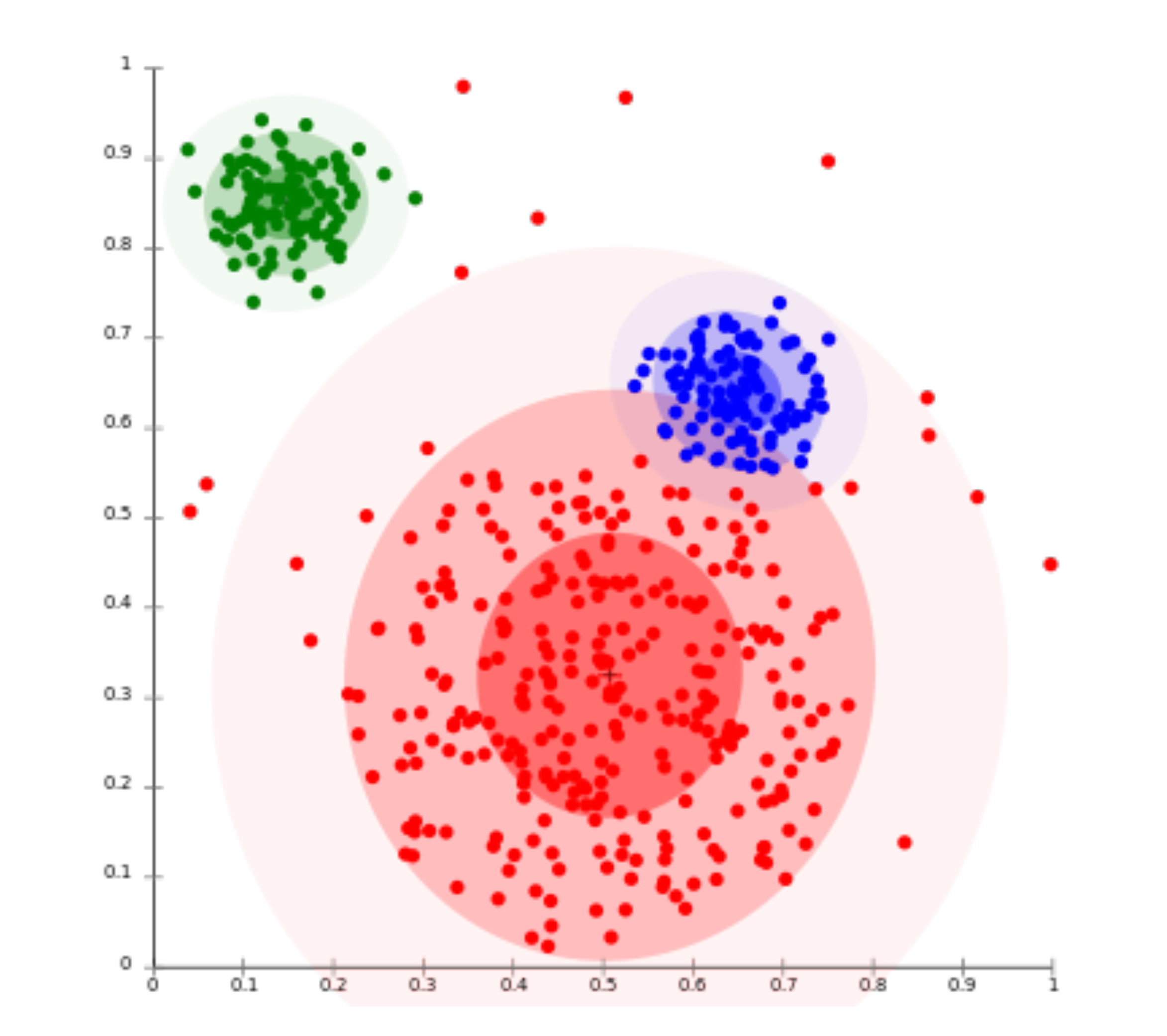
$$z_n \sim \mathrm{Discrete}(\boldsymbol{\pi})$$

$$\boldsymbol{x}_n | z_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Gaussian Clustering

Maximum posterior clustering

$$z_n = \underset{k}{\operatorname{argmax}}\, p(z_n = k \,|\, \boldsymbol{x}_n, \boldsymbol{\theta})$$



Generative Model

$$z_n \sim \text{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | z_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Maximum Likelihood Parameters

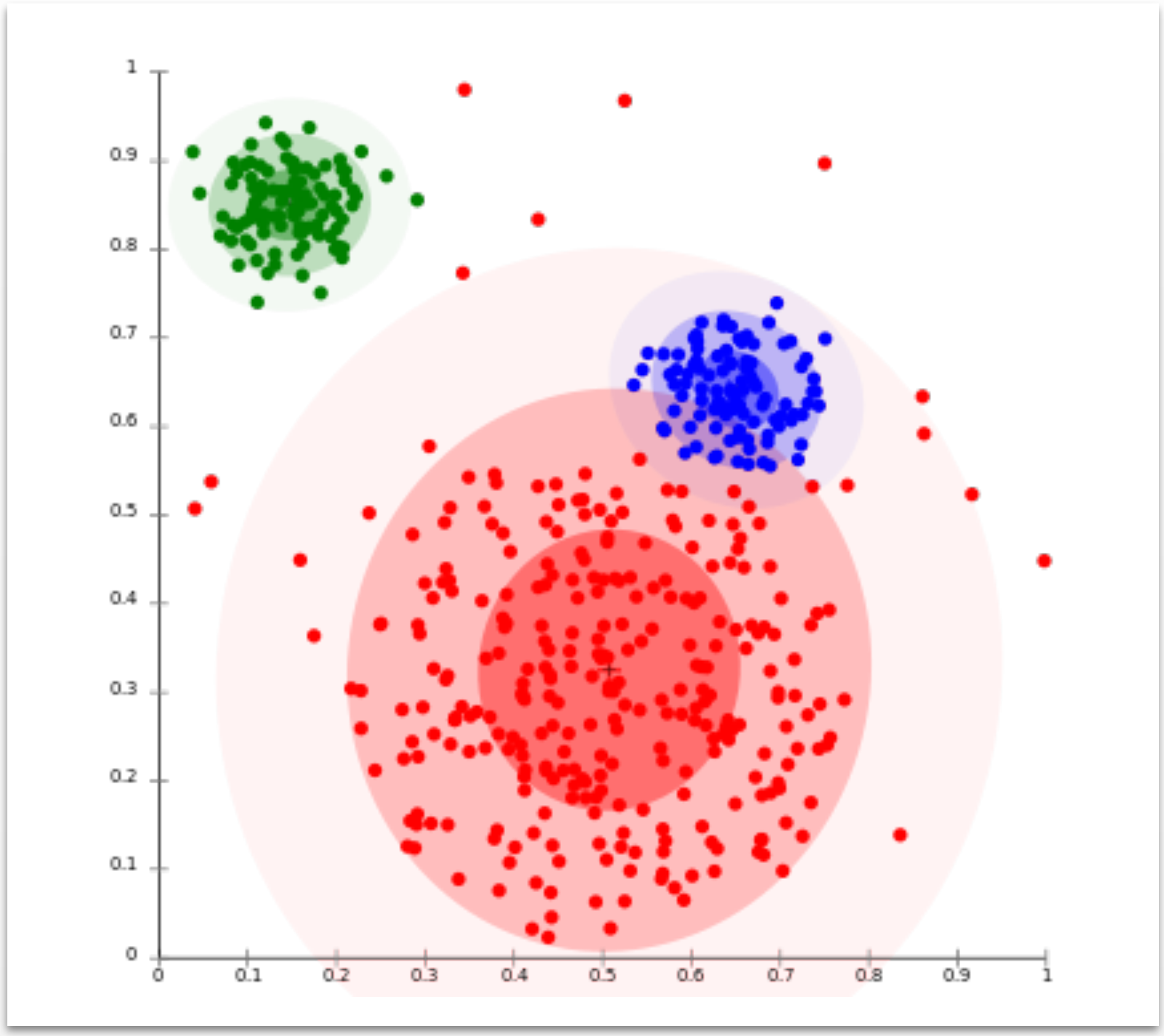
$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} I[z_n = k]\, \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} I[z_n = k]\, |\boldsymbol{x}_n - \boldsymbol{\mu}_k|^2$$

$$\boldsymbol{\pi} = (N_1/N, \ldots, N_K/N)$$

$$N_k = \sum_{n=1}^{N} I[z_n = k]$$

Gaussian Clustering



Algorithm

Initialize parameters to $\boldsymbol{\theta}^0$

Repeat until convergence

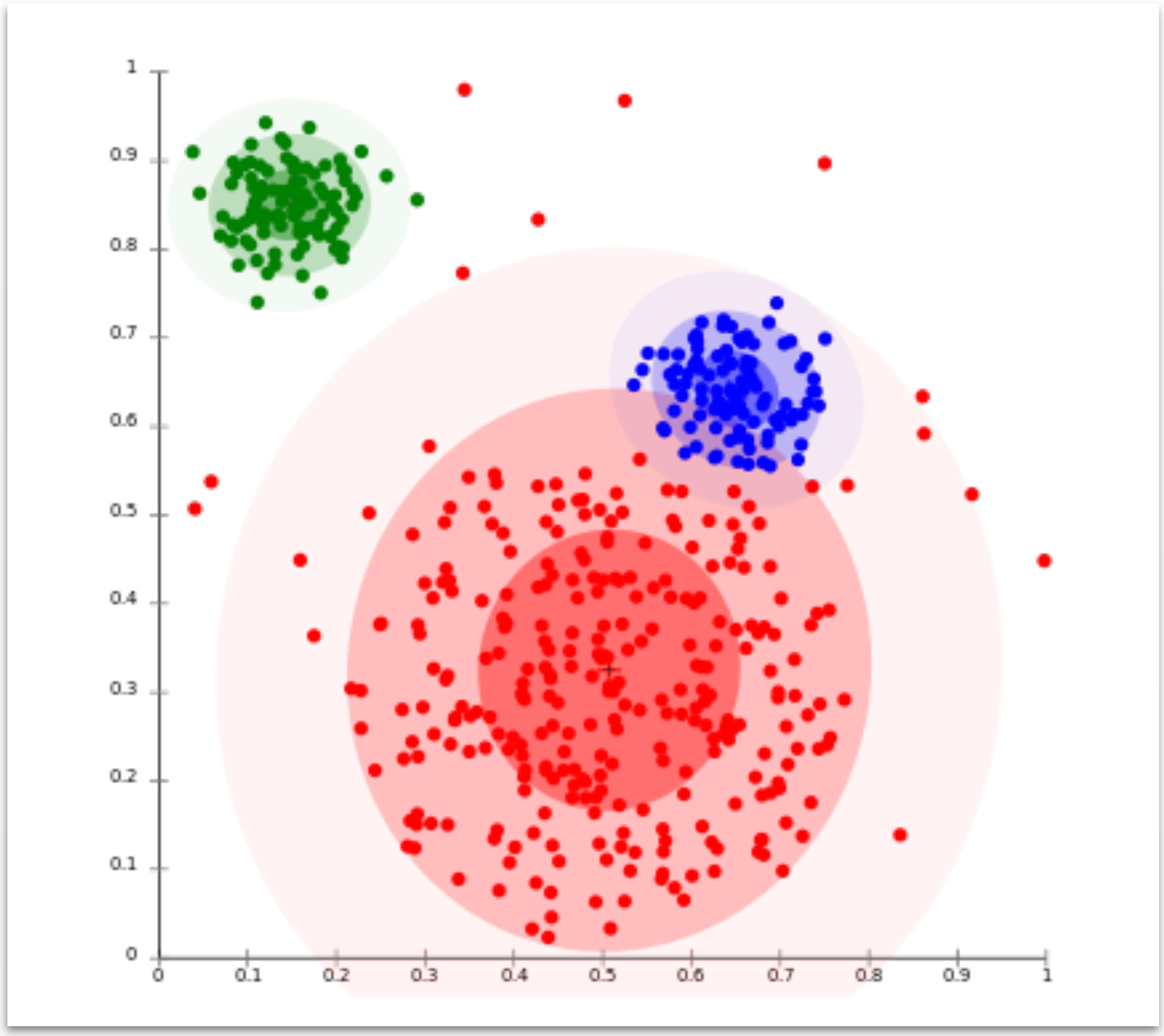
1. Update cluster assignments

$$\boldsymbol{z}^i = \underset{\boldsymbol{z}}{\operatorname{argmax}}\ p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}^{i-1})$$

2. Update parameters

$$\boldsymbol{\theta}^i = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{X}, \boldsymbol{z}^i \mid \boldsymbol{\theta})$$

Gaussian Clustering



Algorithm

Initialize parameters to $\boldsymbol{\theta}^0$

Repeat until convergence

1. Update cluster assignments

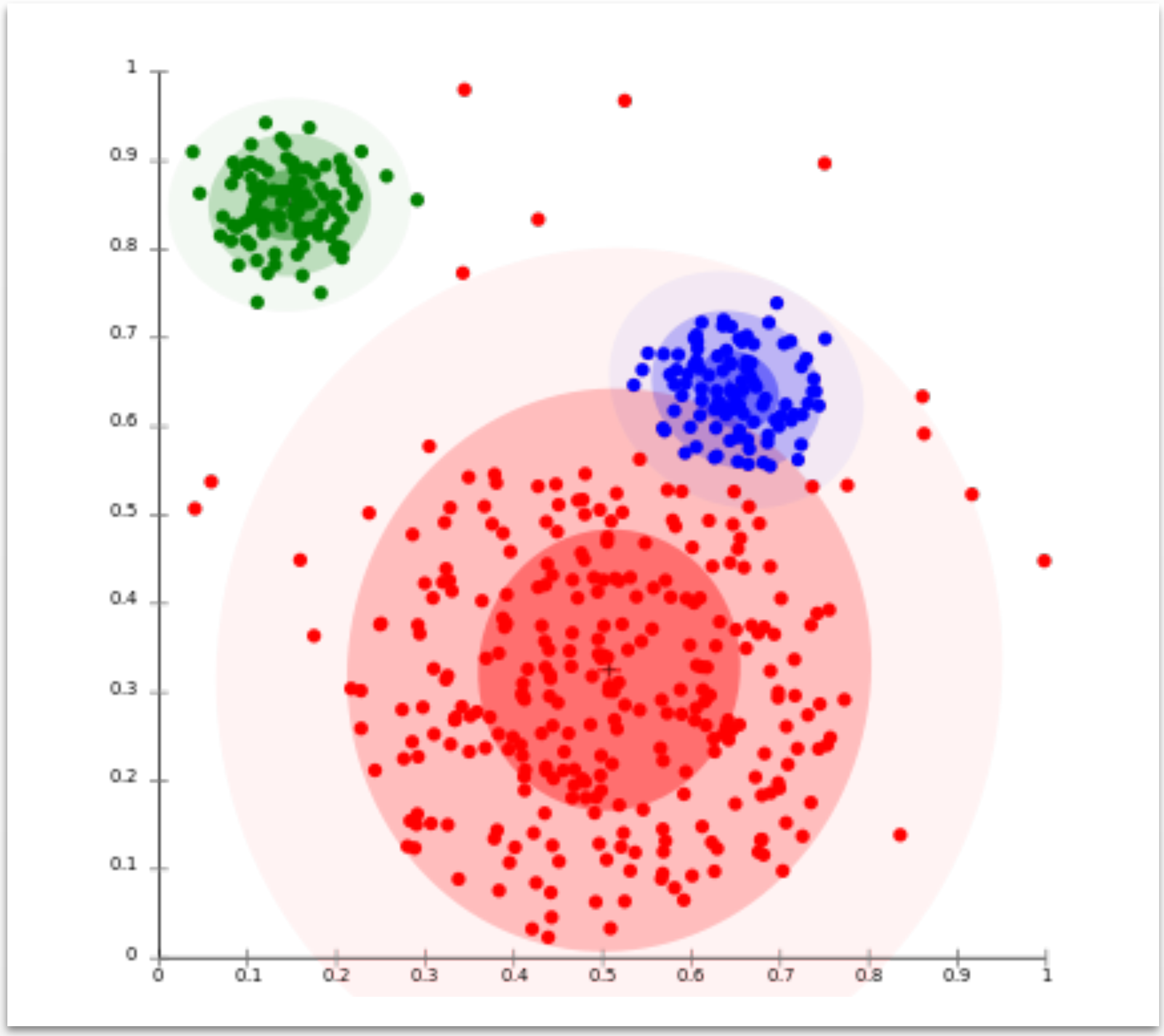
$$\boldsymbol{z}^i = \underset{\boldsymbol{z}}{\operatorname{argmax}}\; p(\boldsymbol{X}, \boldsymbol{z} \,|\, \boldsymbol{\theta}^{i-1})$$

2. Update parameters

$$\boldsymbol{\theta}^i = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\; p(\boldsymbol{X}, \boldsymbol{z}^i \,|\, \boldsymbol{\theta})$$

How does this algorithm relate to K-means?

Gaussian Clustering



Algorithm

Initialize parameters to $\boldsymbol{\theta}^0$

Repeat until convergence

1. Update cluster assignments

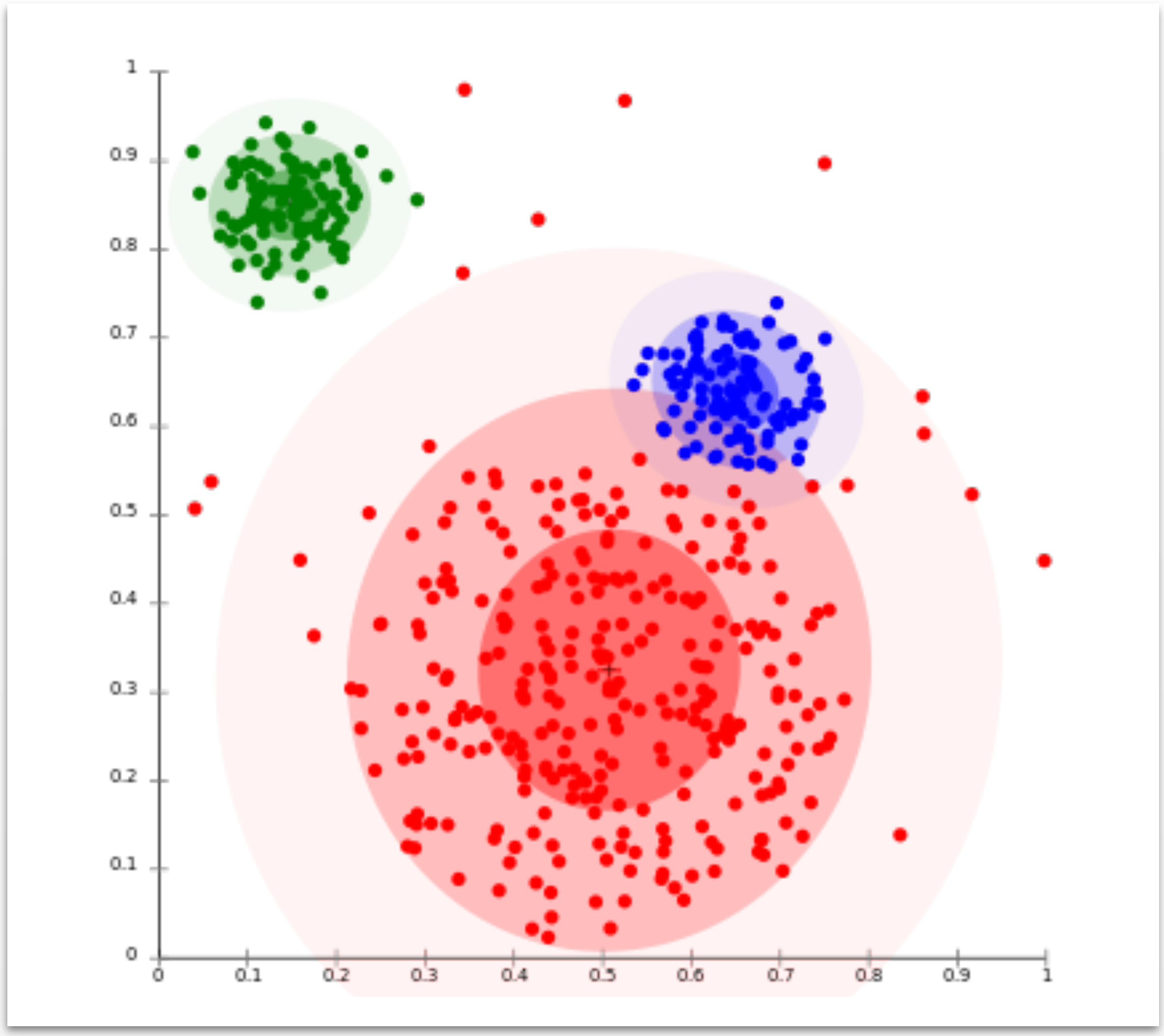
$$\boldsymbol{z}^i = \underset{\boldsymbol{z}}{\operatorname{argmax}}\, p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}^{i-1})$$

2. Update parameters

$$\boldsymbol{\theta}^i = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{X}, \boldsymbol{z}^i \mid \boldsymbol{\theta})$$

How can we deal with overlapping clusters in a better way?

Gaussian Clustering



Algorithm

Initialize parameters to $\boldsymbol{\theta}^0$

Repeat until convergence

1. Update cluster assignments

$$\boldsymbol{z}^i = \underset{\boldsymbol{z}}{\operatorname{argmax}}\ p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}^{i-1})$$

2. Update parameters

$$\boldsymbol{\theta}^i = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{X}, \boldsymbol{z}^i \mid \boldsymbol{\theta})$$

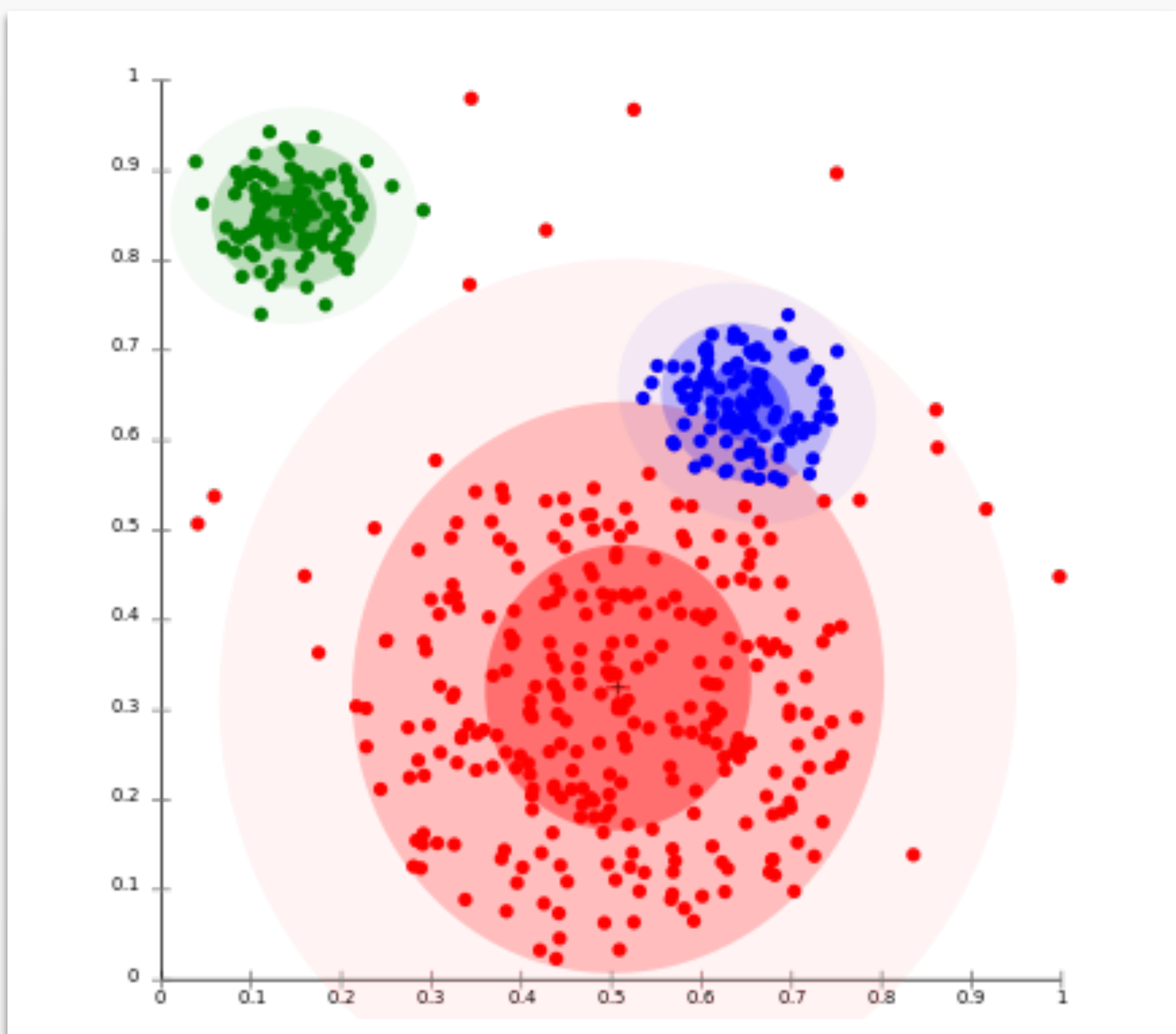
How can we deal with overlapping clusters in a better way?

Idea: Perform *soft* clustering using weighted assignments

Gaussian Clustering

Maximum posterior clustering

$$\boxed{z_n = \operatorname*{argmax}_k p(z_n = k \,|\, \boldsymbol{x}_n, \boldsymbol{\theta})}$$



Generative Model

$$z_n \sim \text{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | z_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Maximum Likelihood Parameters

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \boxed{I[z_n = k]} \boldsymbol{y}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \boxed{I[z_n = k]} |\boldsymbol{y}_n - \boldsymbol{\mu}_k|^2$$

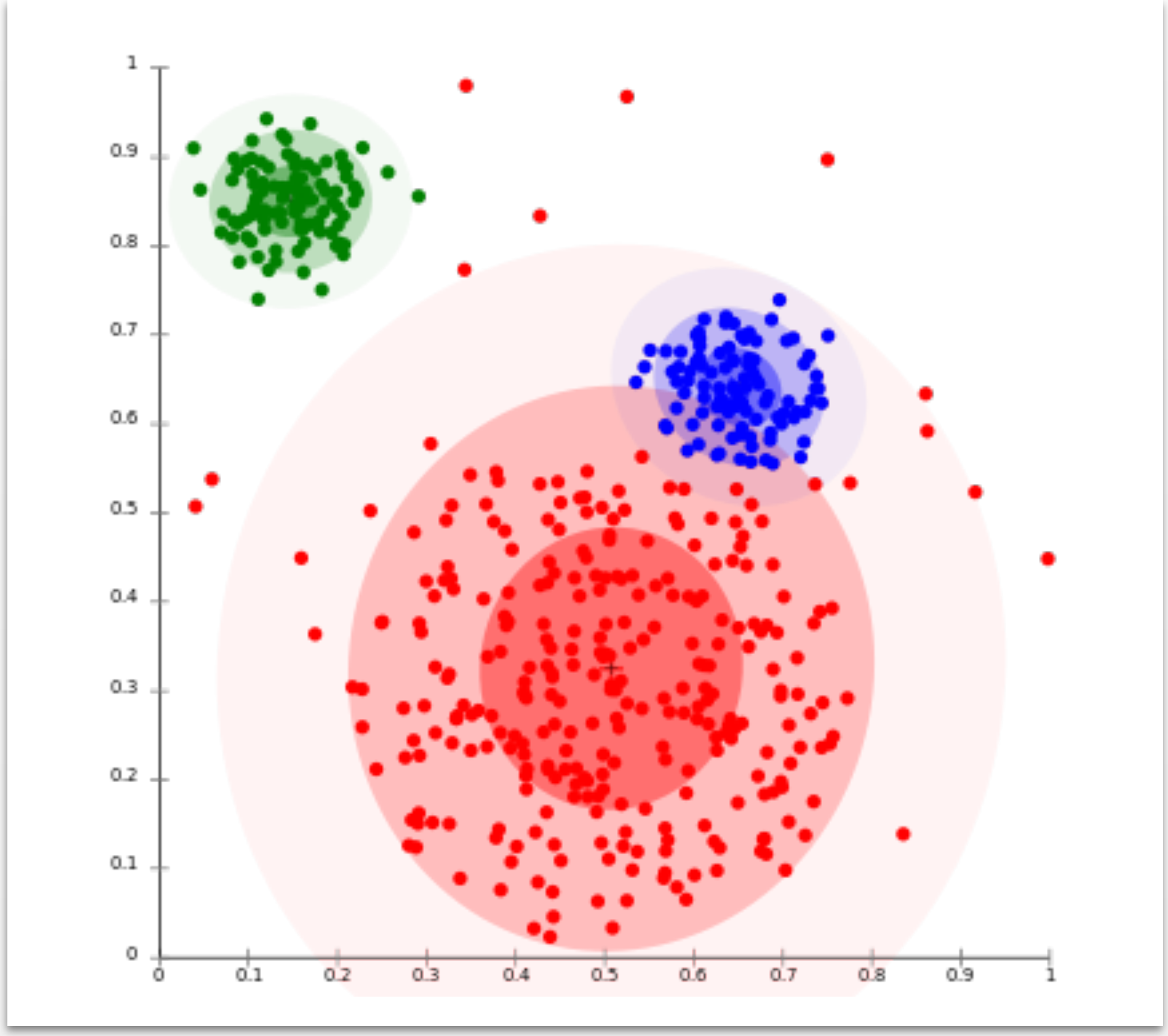
$$\boldsymbol{\pi} = (N_1/N, \ldots, N_K/N)$$

$$N_k = \sum_{n=1}^{N} \boxed{I[z_n = k]}$$

Gaussian *Soft* Clustering

Posterior weights

$$\gamma_{nk} := p(z_n = k \,|\, \boldsymbol{x}_n, \boldsymbol{\theta})$$



Generative Model

$$z_n \sim \text{Discrete}(\boldsymbol{\pi})$$

$$\boldsymbol{x}_n | z_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Parameter Estimates

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \boxed{\gamma_{nk}} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \boxed{\gamma_{nk}} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$$

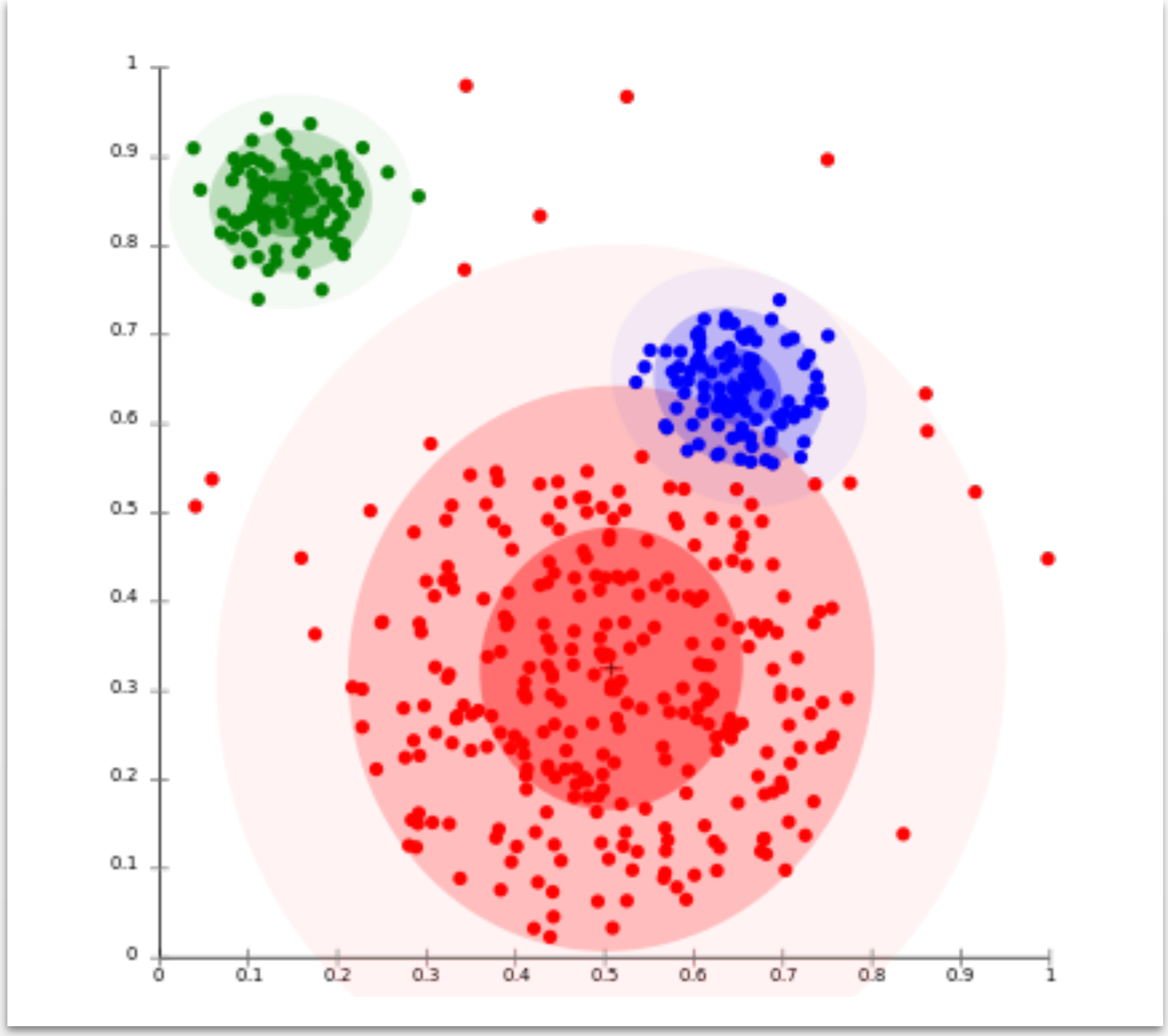
$$\boldsymbol{\pi} = (N_1/N, \ldots, N_K/N)$$

$$N_k = \sum_{n=1}^{N} \boxed{\gamma_{nk}}$$

Gaussian *Soft* Clustering

Posterior weights

$$\gamma_{nk} := p(z_n = k \mid \boldsymbol{x}_n, \boldsymbol{\theta})$$



Generative Model

$$z_n \sim \mathrm{Discrete}(\boldsymbol{\pi})$$

$$\boldsymbol{x}_n | z_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Parameter Estimates

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk}\, (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$$

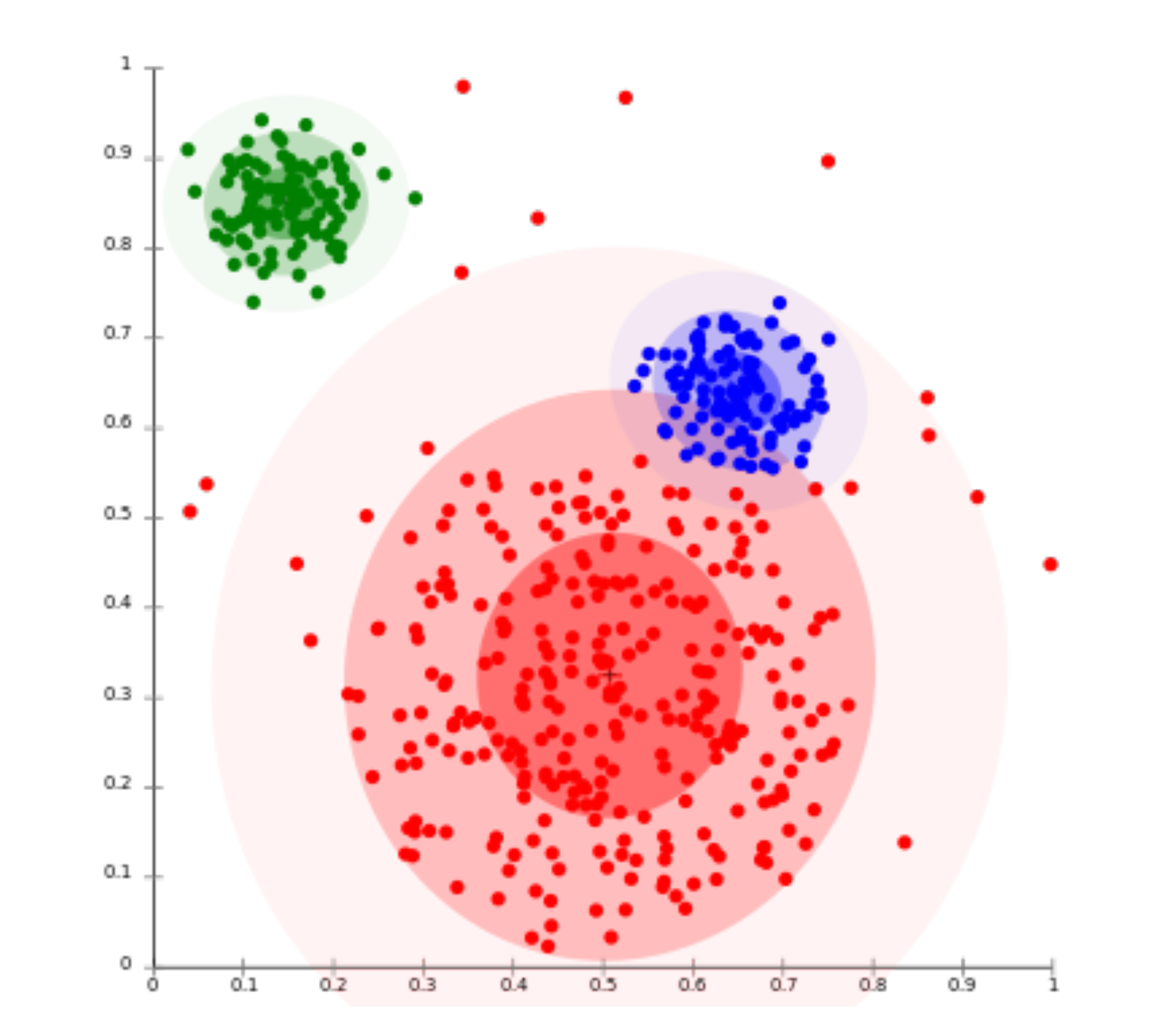
$$\boldsymbol{\pi} = (N_1/N, \ldots, N_K/N)$$

$$N_k = \textstyle\sum_{n=1}^{N} \gamma_{nk}$$

Gaussian Mixture Model

Generative Model

$$z_n \sim \text{Discrete}(\pi)$$
$$\boldsymbol{x}_n | z_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



Expectation Maximization (*sketch*)

Initialize $\boldsymbol{\theta}$

Repeat until convergence

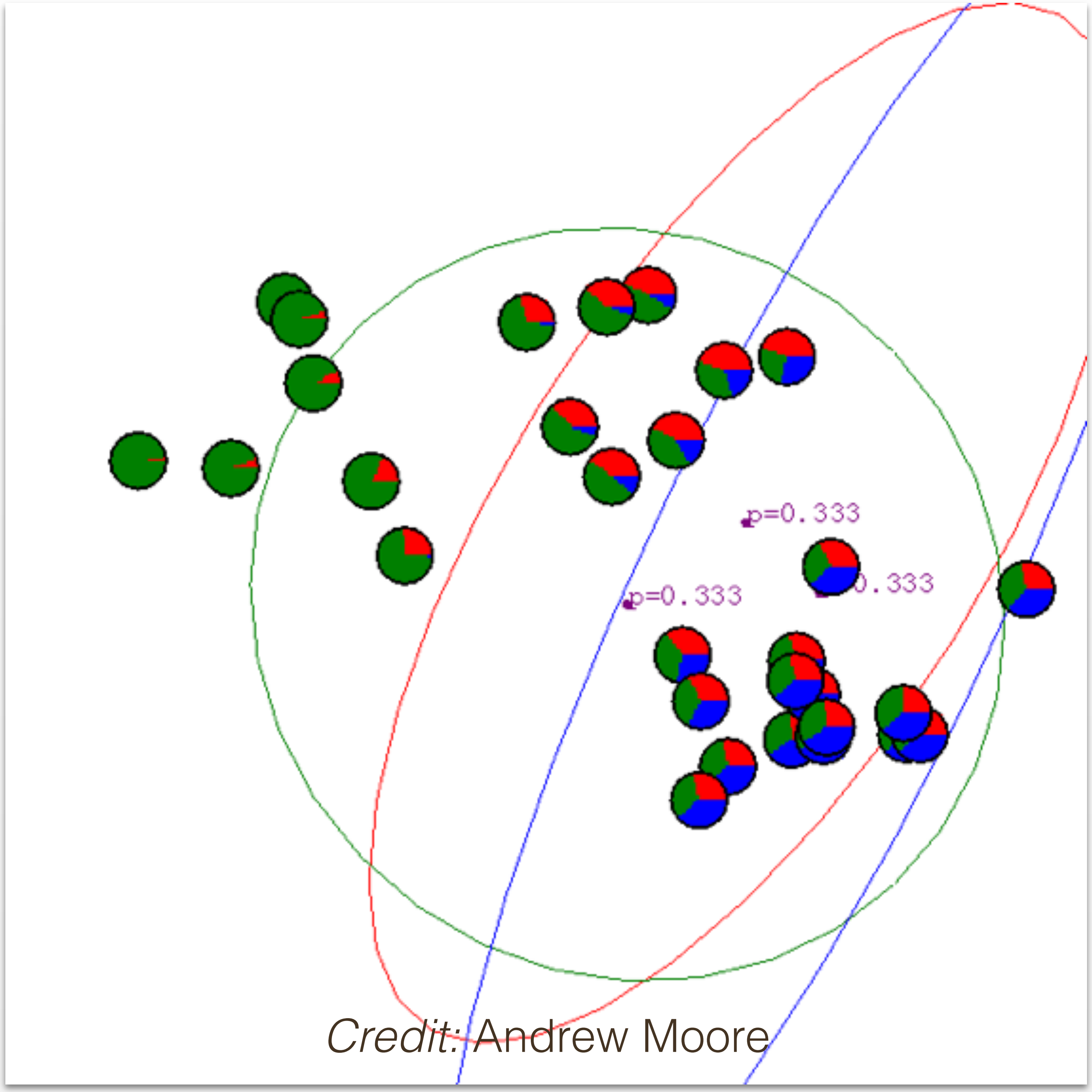
1. Expectation Step

   "calculate $\boldsymbol{\gamma}$ from $\boldsymbol{\theta}$"

2. Maximization Step

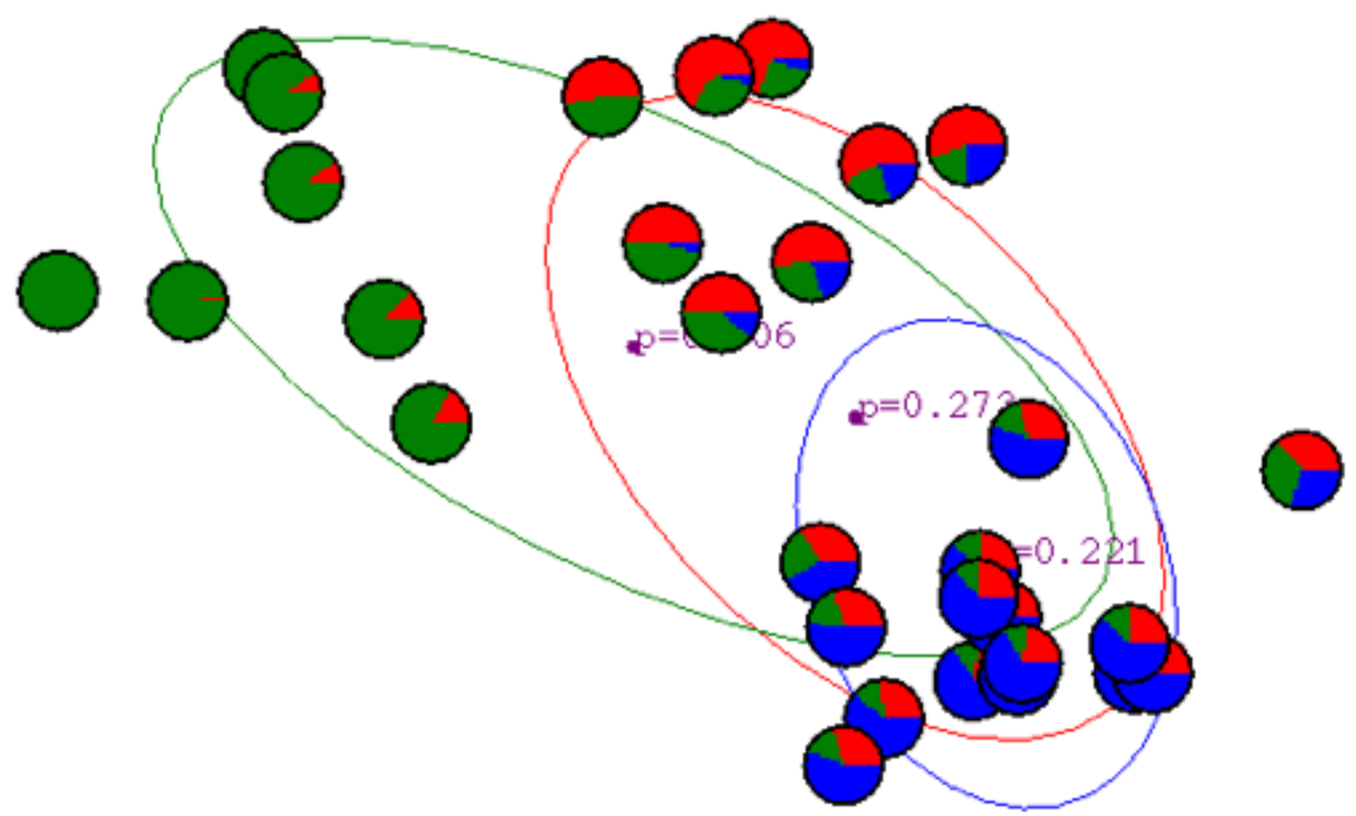
   "calculate $\boldsymbol{\theta}$ from $\boldsymbol{\gamma}$"

EM for Gaussian Mixtures



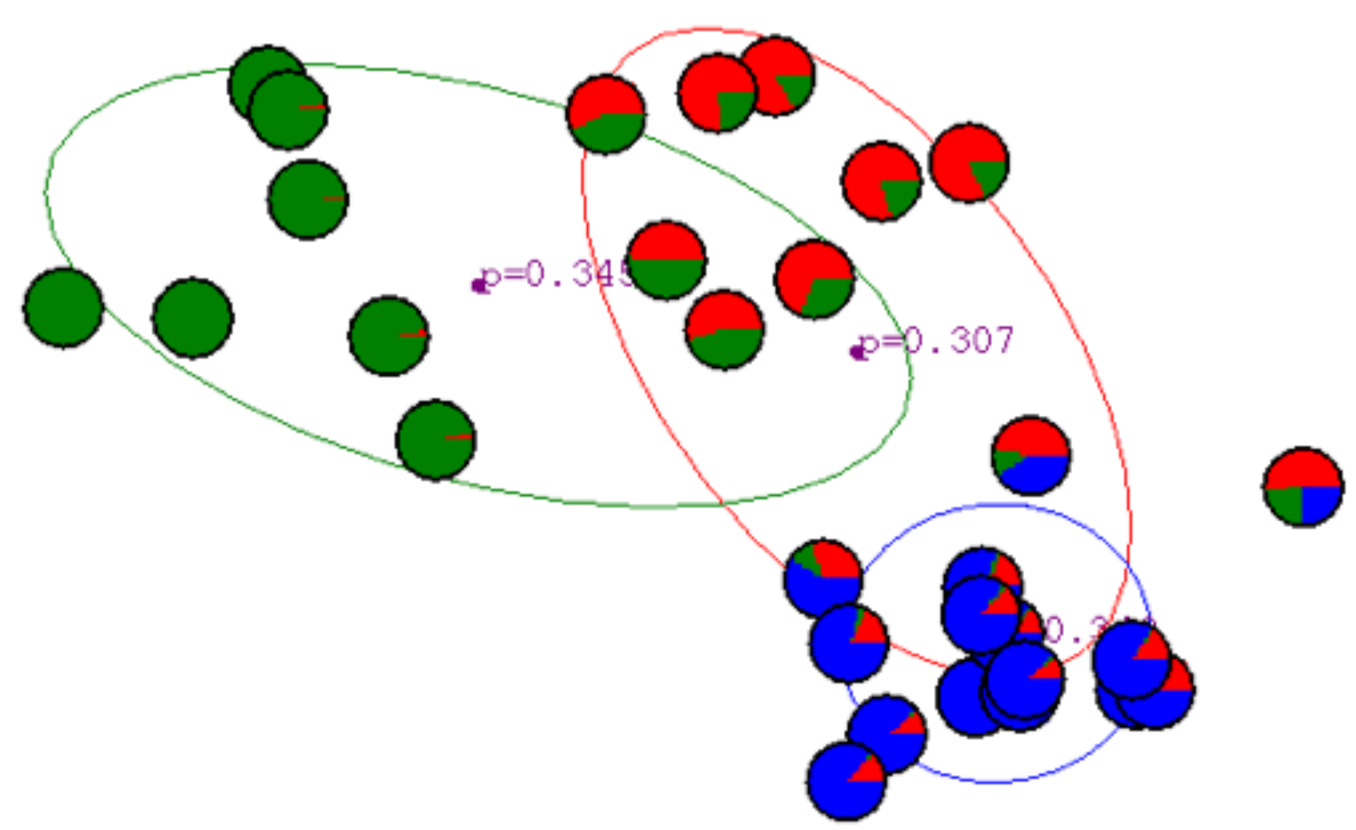
Credit: Andrew Moore

EM for Gaussian Mixtures



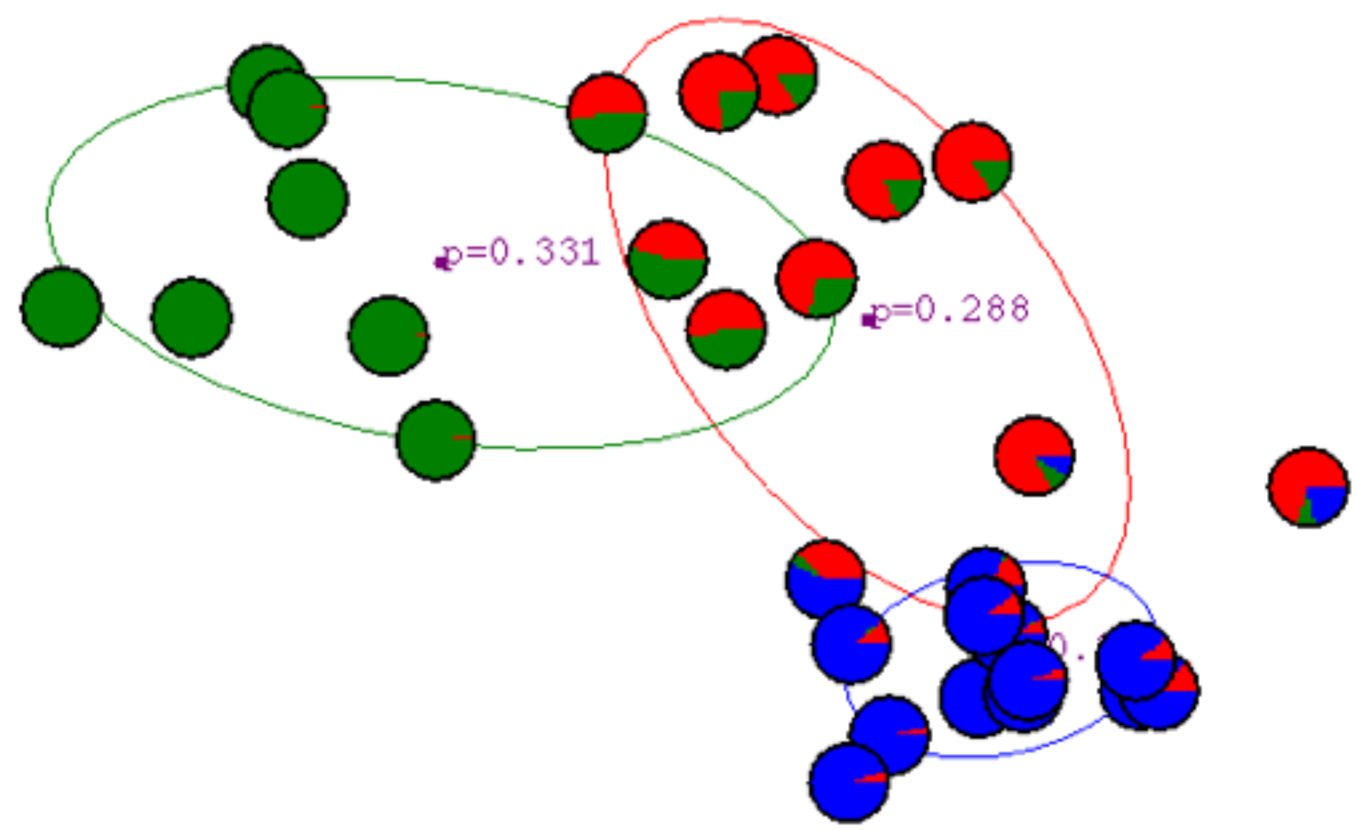
Credit: Andrew Moore

EM for Gaussian Mixtures



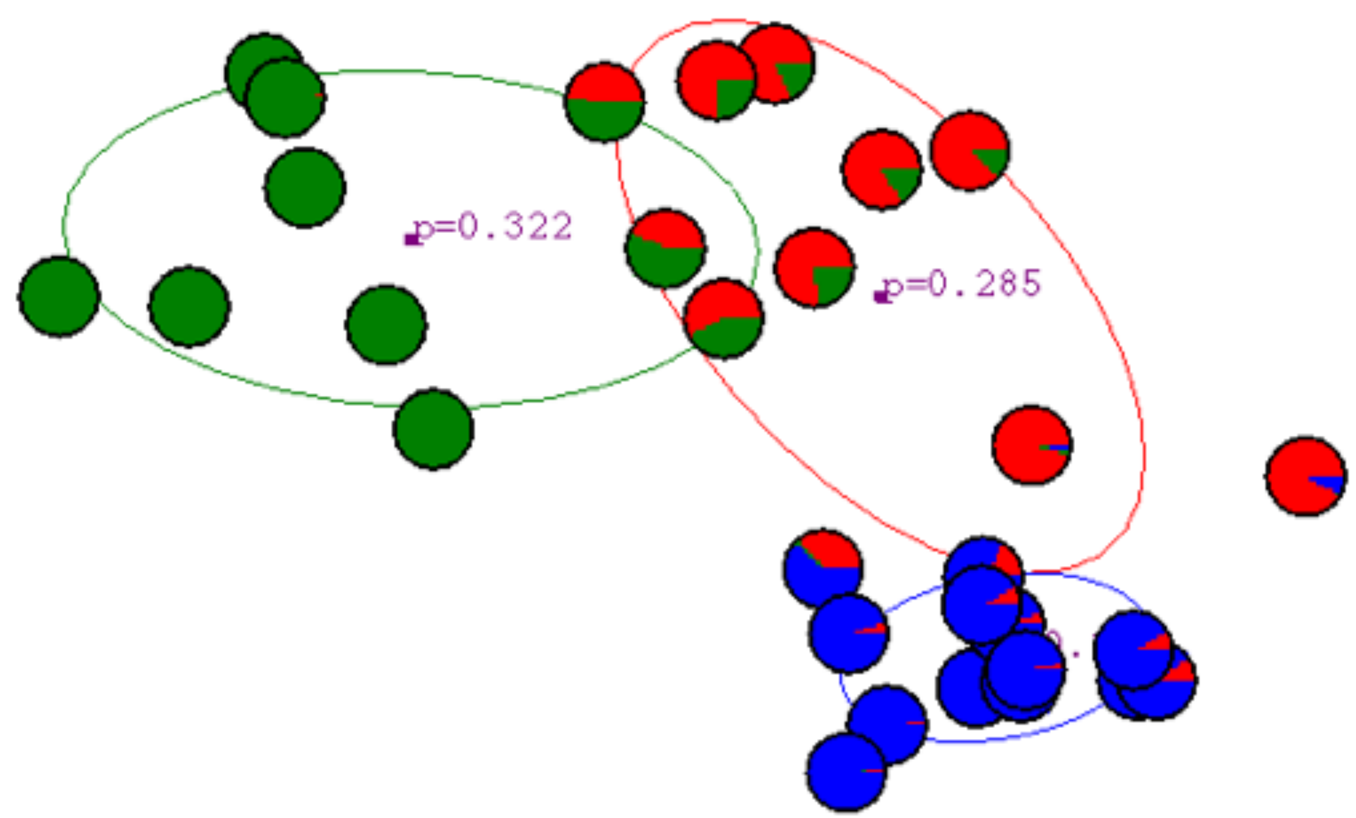
Credit: Andrew Moore

EM for Gaussian Mixtures



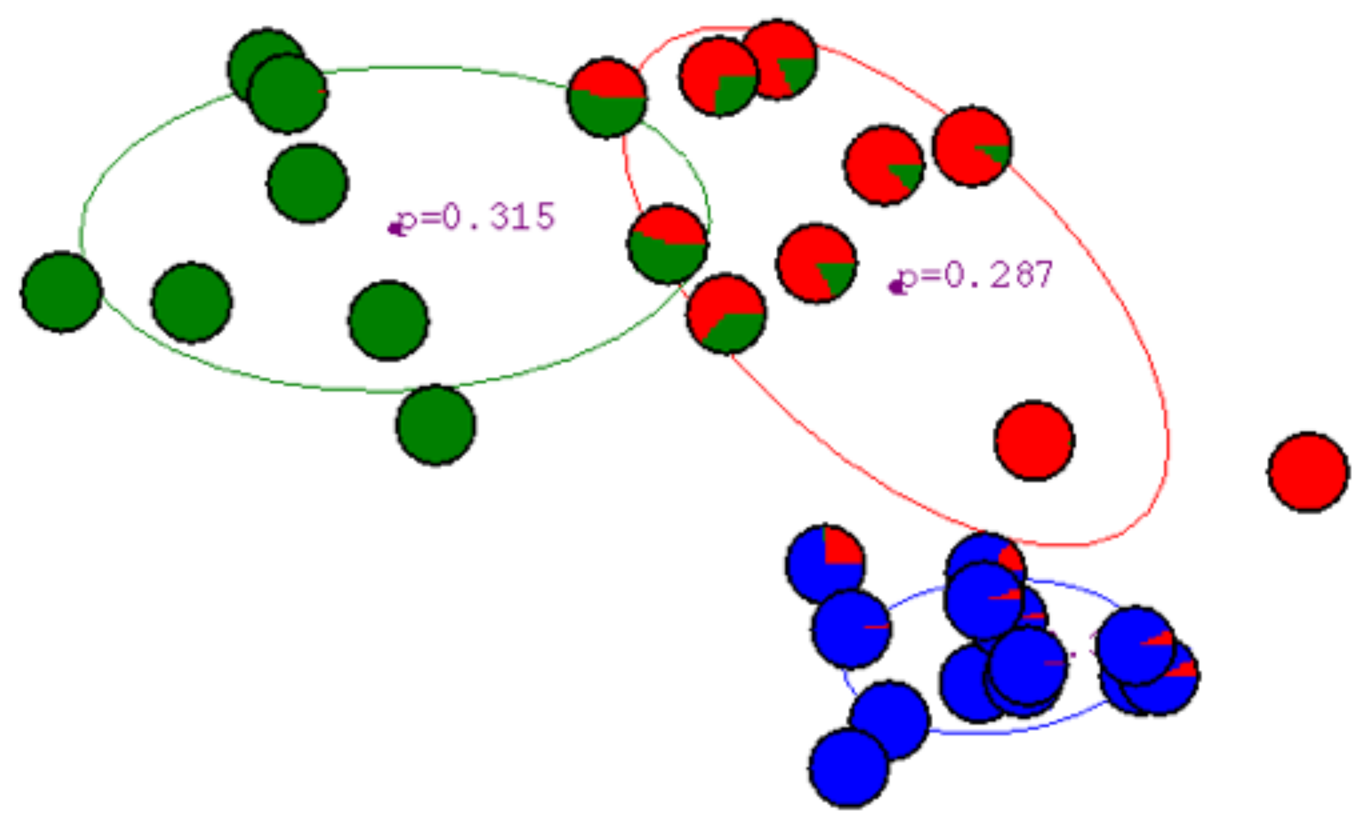
Credit: Andrew Moore

EM for Gaussian Mixtures



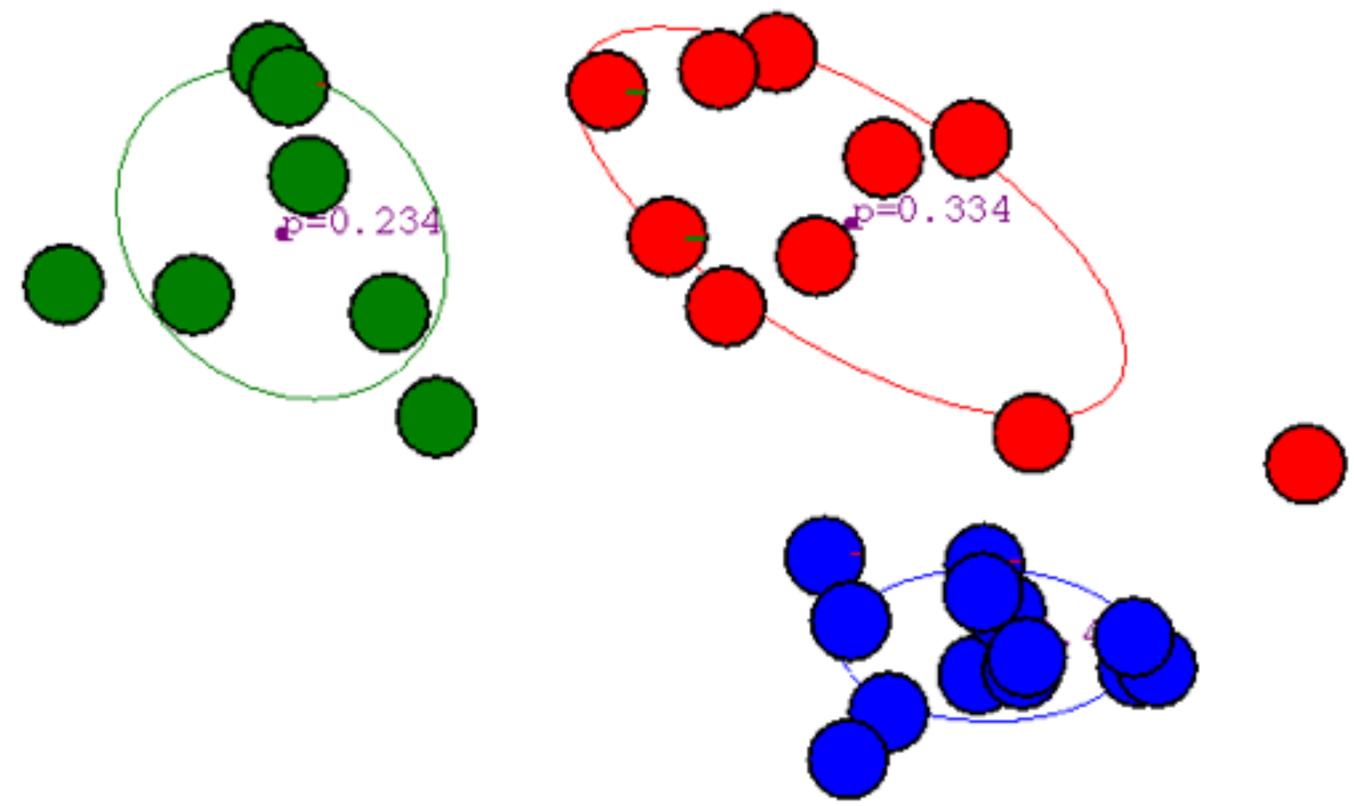
Credit: Andrew Moore

EM for Gaussian Mixtures



Credit: Andrew Moore

EM for Gaussian Mixtures



Credit: Andrew Moore

Expectation Maximization

Maximum Likelihood Estimation

Supervised (e.g. QDA)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta})$$

Unsupervised (e.g. GMM)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \textstyle\sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})$$

Maximum Likelihood Estimation

Supervised (e.g. QDA)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta})$$

Unsupervised (e.g. GMM)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \textstyle\sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})$$

Maximum Likelihood Estimation

Supervised (e.g. QDA)

$$\begin{aligned}\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n, y_n \mid \boldsymbol{\theta})\end{aligned}$$

Unsupervised (e.g. GMM)

$$\begin{aligned}\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \log \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})\end{aligned}$$

Maximum Likelihood Estimation

Supervised (e.g. QDA)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta})$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \boxed{\textstyle\sum_{n=1}^{N} \log p(\boldsymbol{x}_n, y_n \mid \boldsymbol{\theta})}$$

Solve for zero gradient to find maximum

Unsupervised (e.g. GMM)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \textstyle\sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \textstyle\sum_{n=1}^{N} \log \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})$$

Maximum Likelihood Estimation

Supervised (e.g. QDA)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta})$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \boxed{\textstyle\sum_{n=1}^{N} \log p(\boldsymbol{x}_n, y_n \mid \boldsymbol{\theta})}$$

Solve for zero gradient to find maximum

Unsupervised (e.g. GMM)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \textstyle\sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \boxed{\textstyle\sum_{n=1}^{N} \log \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})}$$

Not so easy here, because of sum inside logarithm

Lower Bound on Log Likelihood

$$\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{X} \mid \boldsymbol{\theta})$$

(multiplication by 1)

Lower Bound on Log Likelihood

$$\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{X} \mid \boldsymbol{\theta})$$ (multiplication by 1)

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[p(\boldsymbol{X} \mid \boldsymbol{\theta}) \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right]$$ (multiplication by 1)

Lower Bound on Log Likelihood

$$\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{X} \mid \boldsymbol{\theta}) \qquad \text{(multiplication by 1)}$$

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[p(\boldsymbol{X} \mid \boldsymbol{\theta}) \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] \qquad \text{(multiplication by 1)}$$

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] \qquad \text{(Bayes rule)}$$

Lower Bound on Log Likelihood

$$\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{X} \mid \boldsymbol{\theta})$$ (multiplication by 1)

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[p(\boldsymbol{X} \mid \boldsymbol{\theta}) \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right]$$ (multiplication by 1)

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right]$$ (Bayes rule)

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{q(\boldsymbol{z})} \right] + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{q(\boldsymbol{z})}{p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})} \right]$$

Lower Bound on Log Likelihood

$$\begin{aligned}
\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{X} \mid \boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[p(\boldsymbol{X} \mid \boldsymbol{\theta}) \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] \\
&= \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{q(\boldsymbol{z})} \right]}_{\text{Lower bound: } \mathscr{L}(q, \boldsymbol{\theta})} + \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{q(\boldsymbol{z})}{p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})} \right]}_{\text{KL divergence: } KL(q \,||\, p)}
\end{aligned}$$

Lower Bound on Log Likelihood

$$
\begin{aligned}
\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{X} \mid \boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[p(\boldsymbol{X} \mid \boldsymbol{\theta}) \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})} \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \right] \\
&= \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{q(\boldsymbol{z})} \right]}_{\text{Lower bound: } \mathscr{L}(q, \boldsymbol{\theta})} + \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left[\frac{q(\boldsymbol{z})}{p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta})} \right]}_{\text{KL divergence: } KL(q \,||\, p)}
\end{aligned}
$$

Claim: $\mathscr{L}(q, \boldsymbol{\theta}) \leq \log p(\boldsymbol{X} \mid \boldsymbol{\theta})$

Intermezzo: KL Divergence

KL Divergence

$$KL(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

Properties

- KL(q || p) ≥ 0
- If KL(q || p) = 0, then q = p
- KL(q || p) ≠ KL(p || q)

Intermezzo: Information Theory

KL Divergence

$$KL(q\,||\,p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

Properties

- $\boxed{KL(q\,||\,p) \geq 0}$
- If $KL(q\,||\,p) = 0$, then $q = p$
- $KL(q\,||\,p) \neq KL(p\,||\,q)$

Proof

$$\begin{aligned}
-D(p||q) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \\
&= \log \sum_{x \in A} q(x) \\
&\leq \log \sum_{x \in \mathcal{X}} q(x) \\
&= \log 1
\end{aligned}$$

Intermezzo: Information Theory

KL Divergence

$$KL(q\,||\,p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

Entropy

$$H(X) = -\sum_x p(x) \log p(x)$$

Mutual Information

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$I(X;Y) = H(Y) - H(Y|X)$$

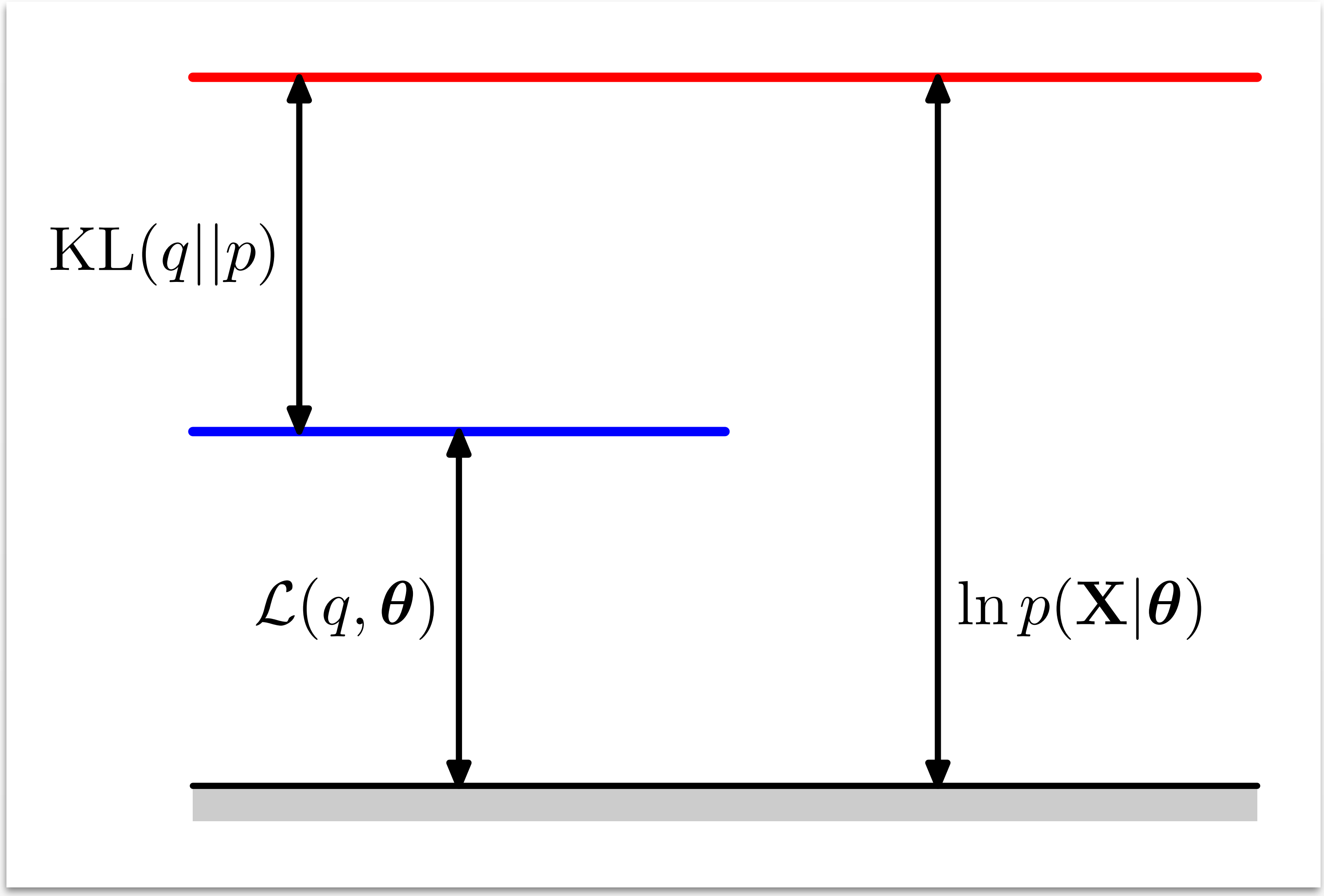
$$I(X;Y) = KL(p(x,y)\,||\,p(x)p(y))$$

Lower Bound on Log Likelihood

$$
\begin{aligned}
\log p(\mathbf{X} \mid \boldsymbol{\theta}) &= \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{X} \mid \boldsymbol{\theta}) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[p(\mathbf{X} \mid \boldsymbol{\theta}) \frac{q(\mathbf{z})}{q(\mathbf{z})} \right] \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{p(\mathbf{X}, \mathbf{z} \mid \boldsymbol{\theta})}{p(\mathbf{z} \mid \mathbf{X}, \boldsymbol{\theta})} \frac{q(\mathbf{z})}{q(\mathbf{z})} \right] \\
&= \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{p(\mathbf{X}, \mathbf{z} \mid \boldsymbol{\theta})}{q(\mathbf{z})} \right]}_{\text{Lower bound: } \mathcal{L}(q, \boldsymbol{\theta})} + \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{X}, \boldsymbol{\theta})} \right]}_{\text{KL divergence: } KL(q \,||\, p)}
\end{aligned}
$$

Claim: $\mathcal{L}(q, \boldsymbol{\theta}) \leq \log p(\mathbf{X} \mid \boldsymbol{\theta})$

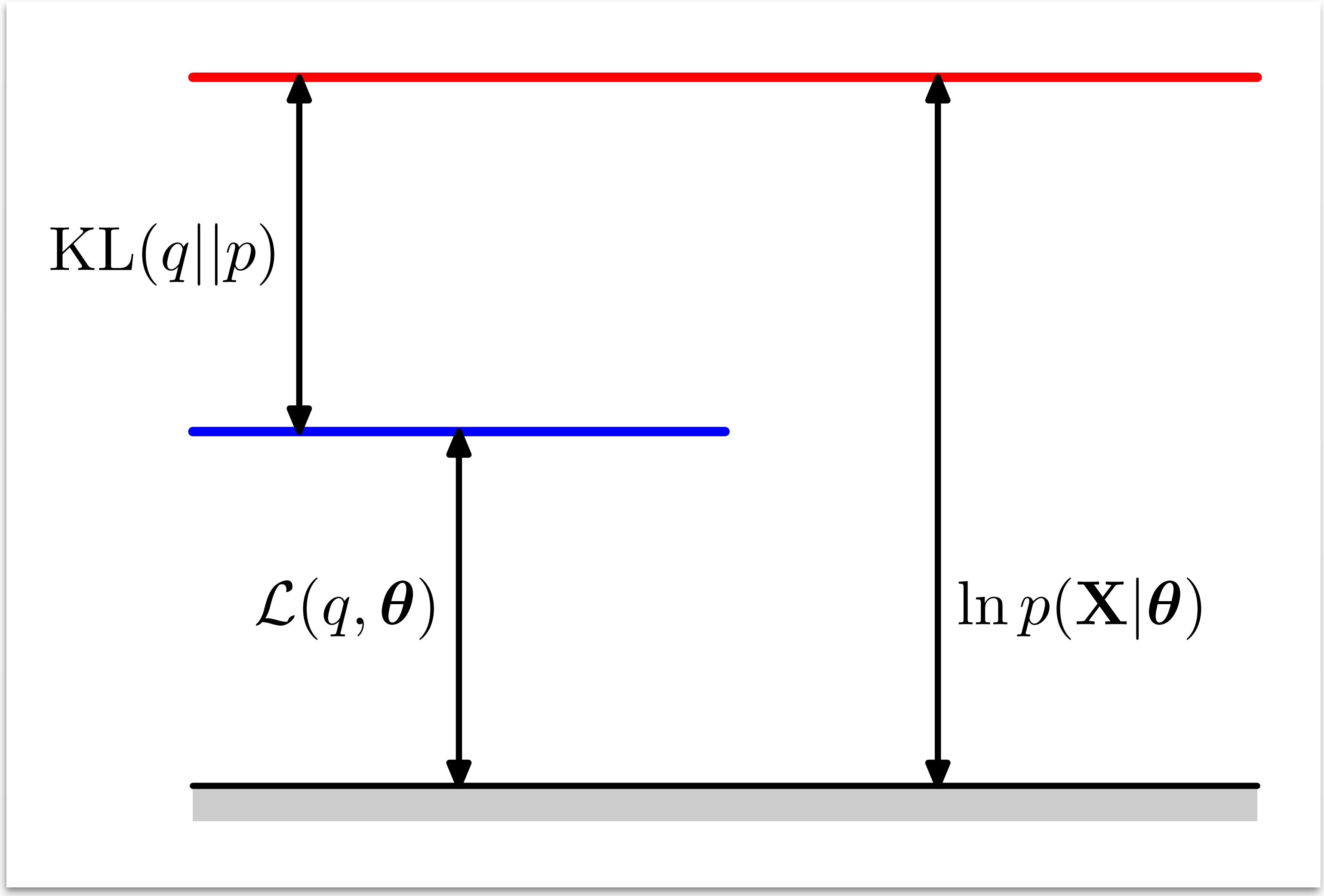
Generalized EM



1. Lower bound is sum over log, not log of sum

$$\mathscr{L}(q(\boldsymbol{z}), \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})}{q(\boldsymbol{z})} \leq \log \sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta})$$

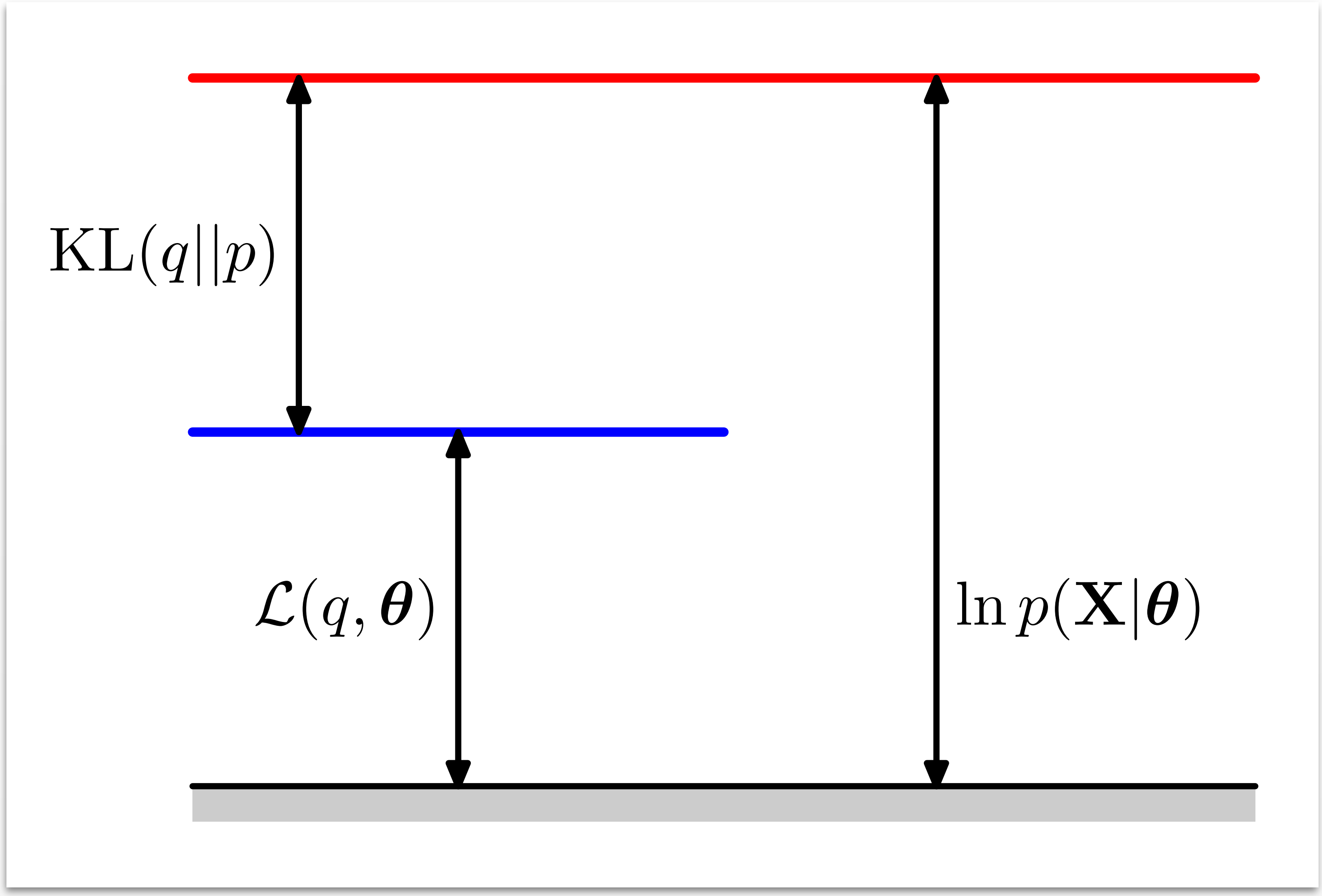
Generalized EM



1. Lower bound is sum over log, not log of sum

$$\mathscr{L}(q(\boldsymbol{z}), \boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_n = k) \log \frac{p(\boldsymbol{x}_n, z_n = k \,|\, \boldsymbol{\theta})}{q(z_n = k)}$$

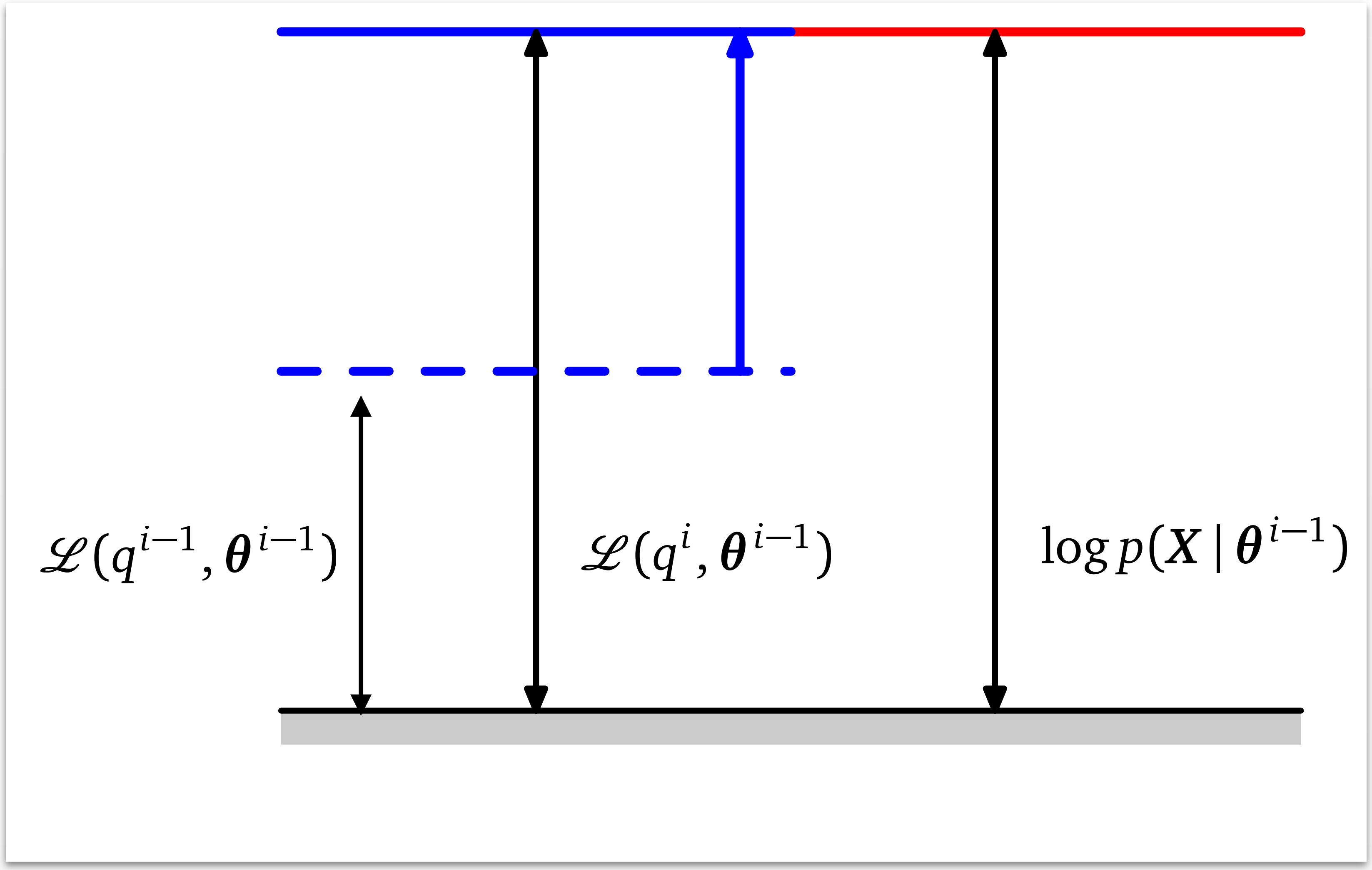
Generalized EM



2. Bound is tight when q(z) = p(z | X, θ)

$$\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \mathscr{L}(q(\boldsymbol{z}), \boldsymbol{\theta}) + KL(q(\boldsymbol{z}) \,||\, p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}))$$

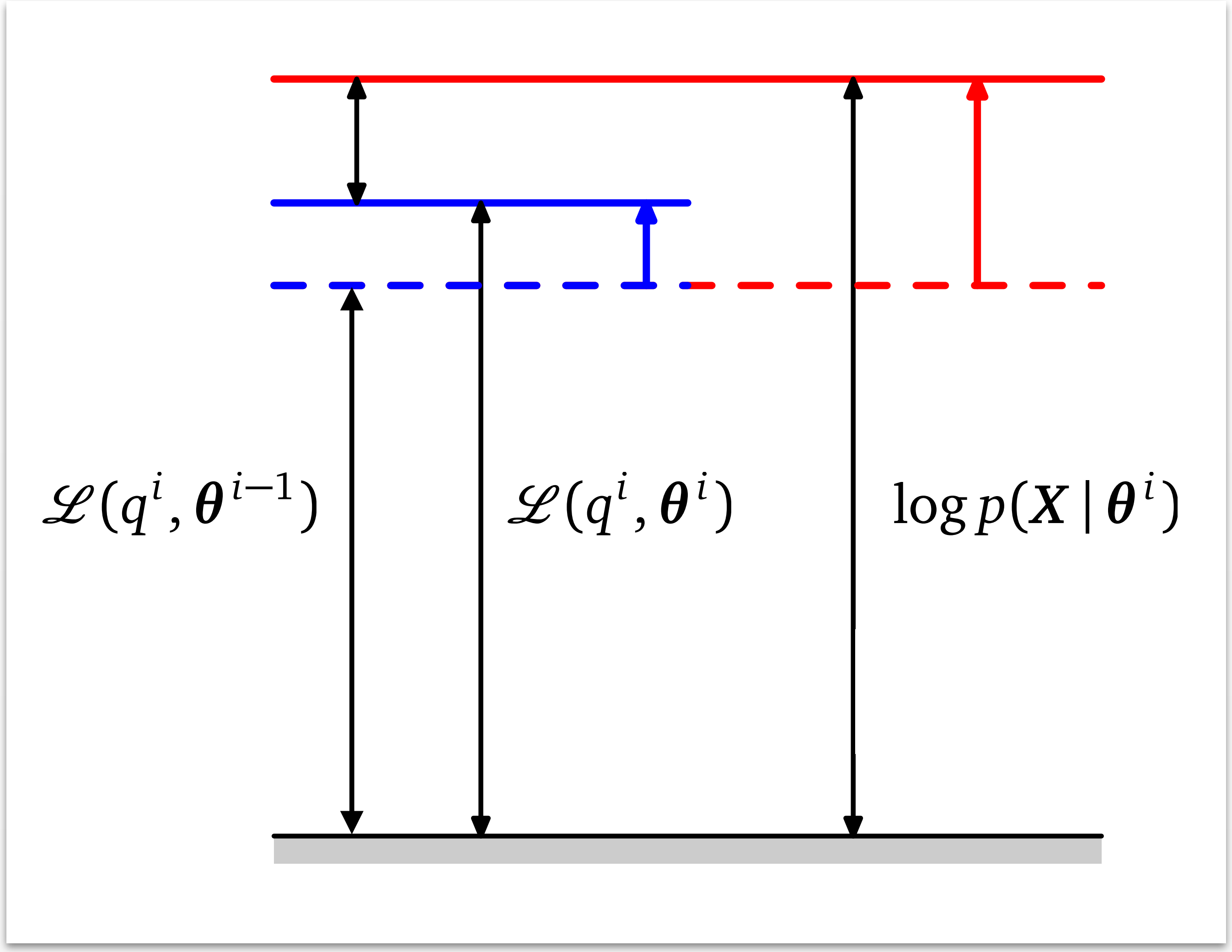
Generalized EM



E-step: maximize with respect to q(*z*)

$$q^i(\boldsymbol{z}) = \underset{q(\boldsymbol{z})}{\operatorname{argmax}} \mathscr{L}(q(\boldsymbol{z}), \boldsymbol{\theta}^{i-1}) = p(\boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{i-1})$$

Generalized EM



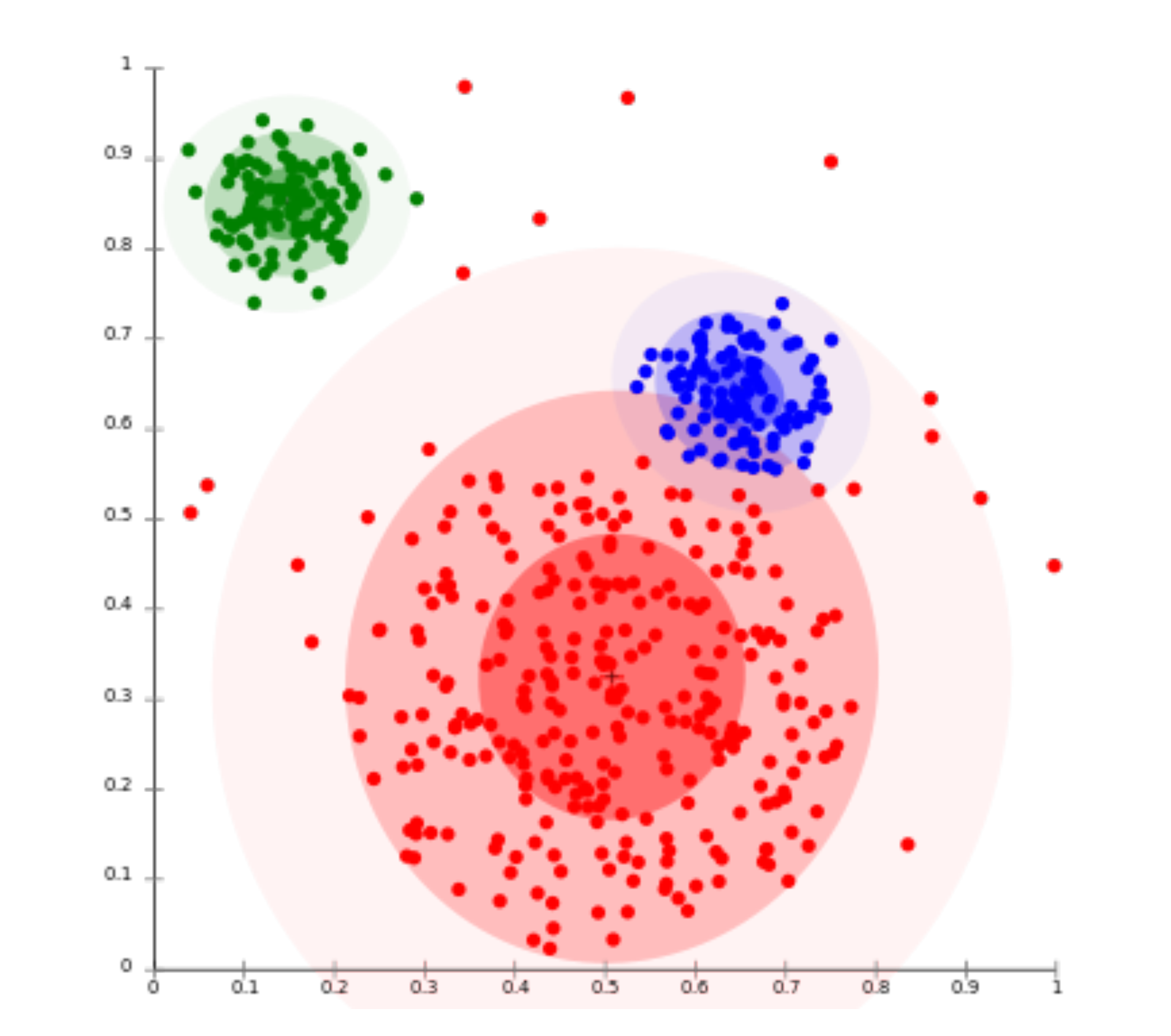
M-*step*: maximize with respect to $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^i = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathscr{L}(q^i(\boldsymbol{z}), \boldsymbol{\theta})$$

Gaussian Mixture Model

Generative Model

$$z_n \sim \mathrm{Discrete}(\boldsymbol{\pi})$$
$$\boldsymbol{x}_n | z_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



Expectation Maximization

Initialize $\boldsymbol{\theta}$

Repeat until convergence

1. Expectation Step

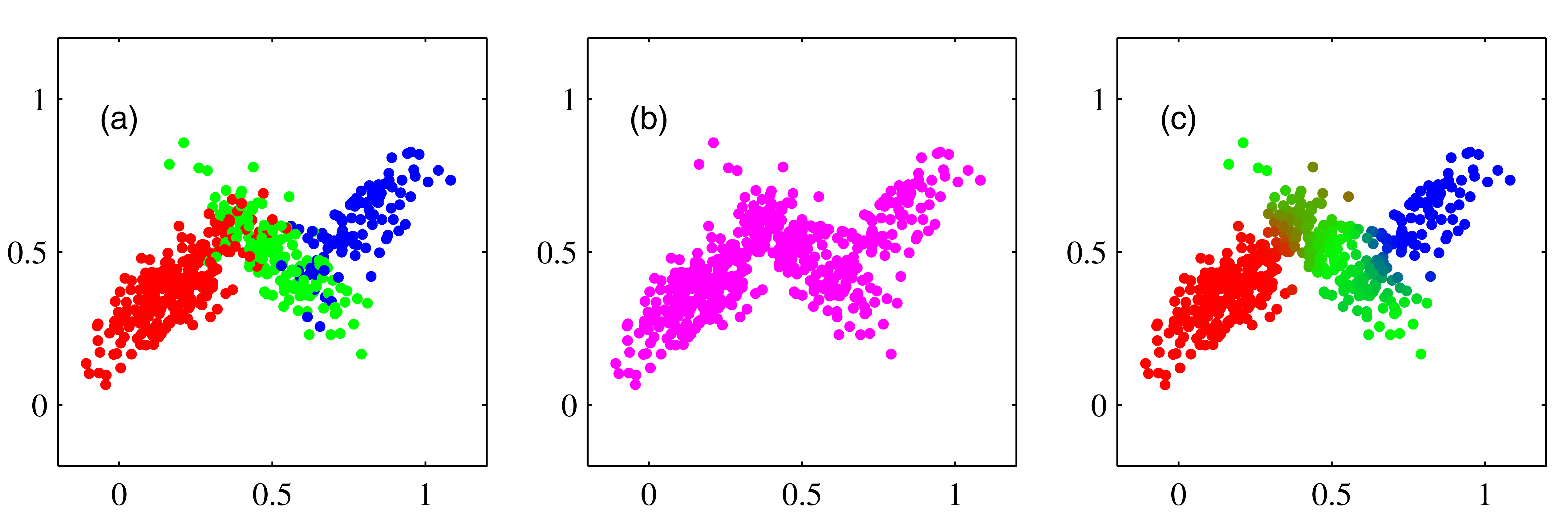
$$q^i(\boldsymbol{z}) = \underset{q(\boldsymbol{z})}{\operatorname{argmax}} \mathcal{L}(q(\boldsymbol{z}), \boldsymbol{\theta}^{i-1})$$

2. Maximization Step

$$\boldsymbol{\theta}^i = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathcal{L}(q^i(\boldsymbol{z}), \boldsymbol{\theta})$$

$$\mathcal{L}(q(\boldsymbol{z}), \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{z} \,|\, \boldsymbol{\theta})}{q(\boldsymbol{z})}$$

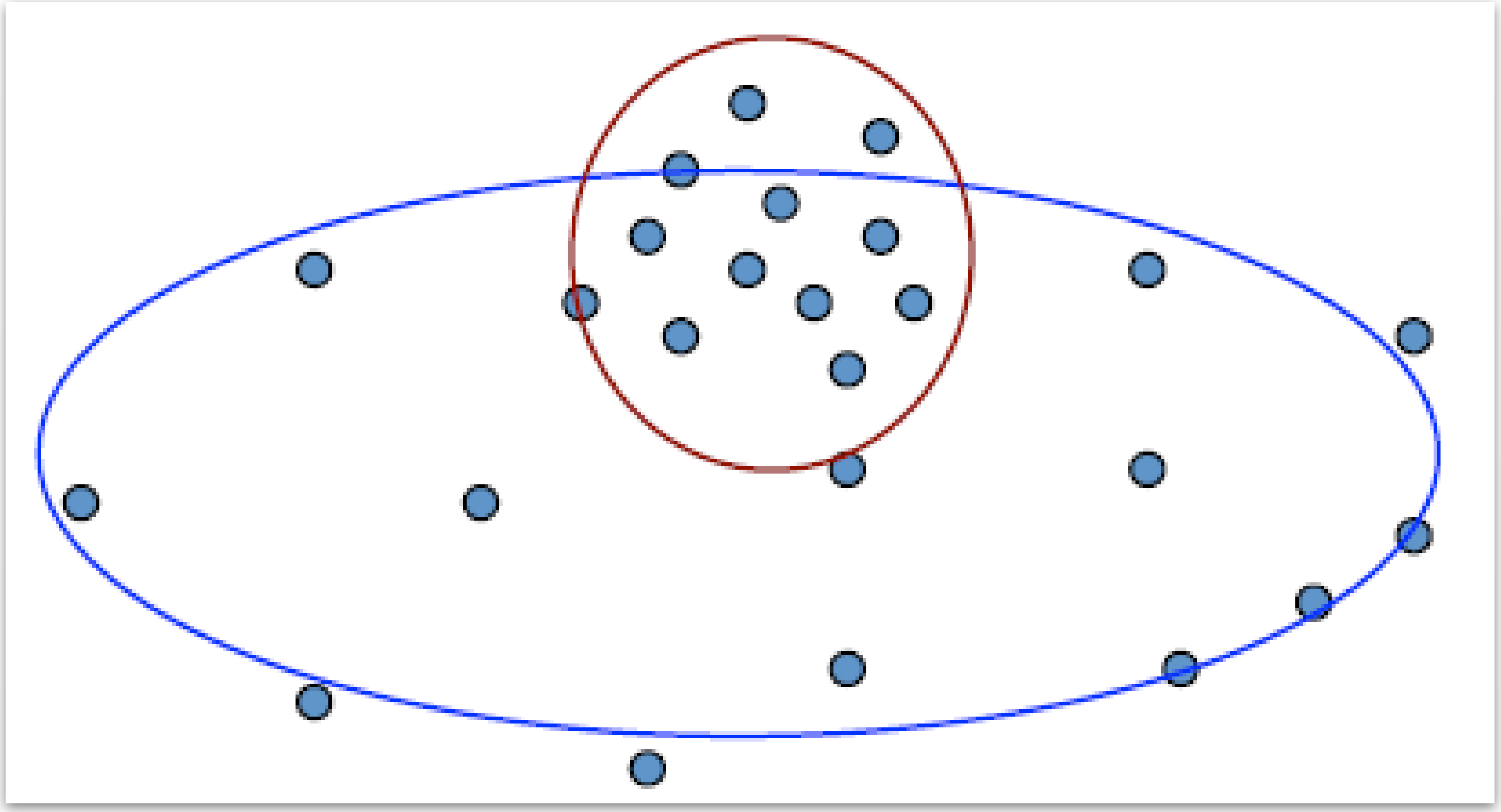
GMM Advantages / Disadvantages



+ Works with overlapping clusters
+ Works with clusters of different densities
+ Same complexity as K-means

- Can get stuck in local maximum
- Need to set number of components

GMM Advantages / Disadvantages



+ Works with overlapping clusters
+ Works with clusters of different densities
+ Same complexity as K-means

- Can get stuck in local maximum
- Need to set number of components

Model Selection

$$p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n | z_n, \boldsymbol{\theta}) p(z_n \mid \boldsymbol{\theta})$$

Need to specify two components

1. Likelihood

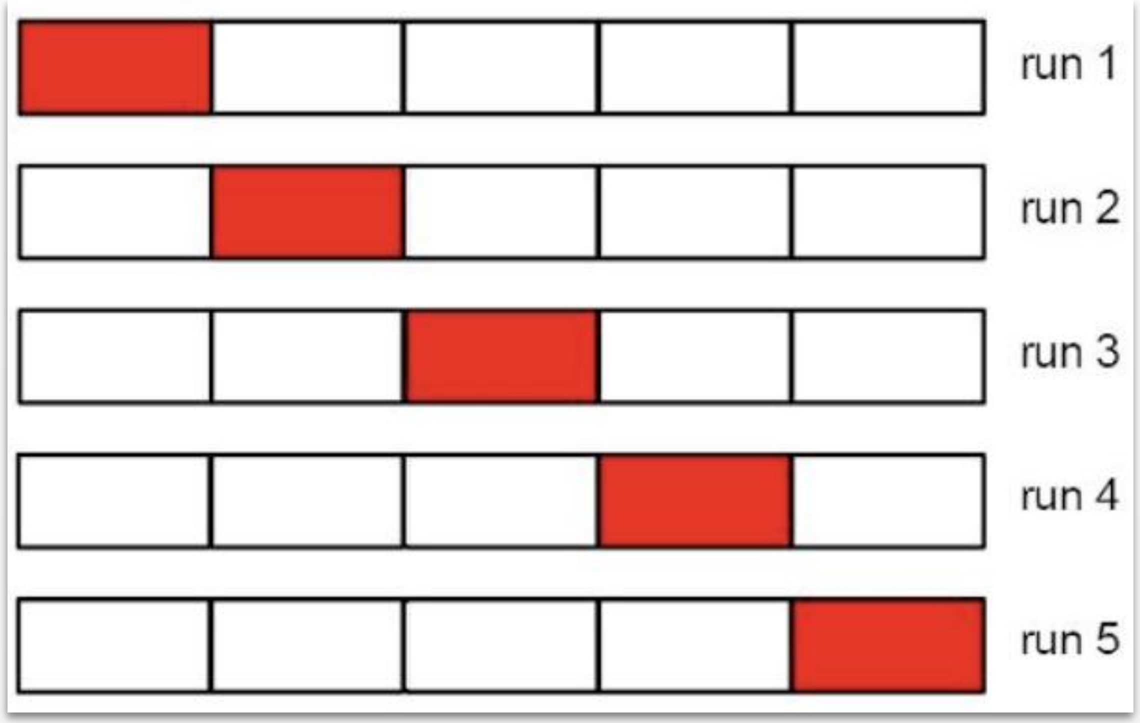
$$p(\boldsymbol{x}_n | z_n, \boldsymbol{\theta})$$

2. Mixture distribution

$$p(z_n \mid \boldsymbol{\theta})$$

How do we know that we have made “good” choices?

Model Selection



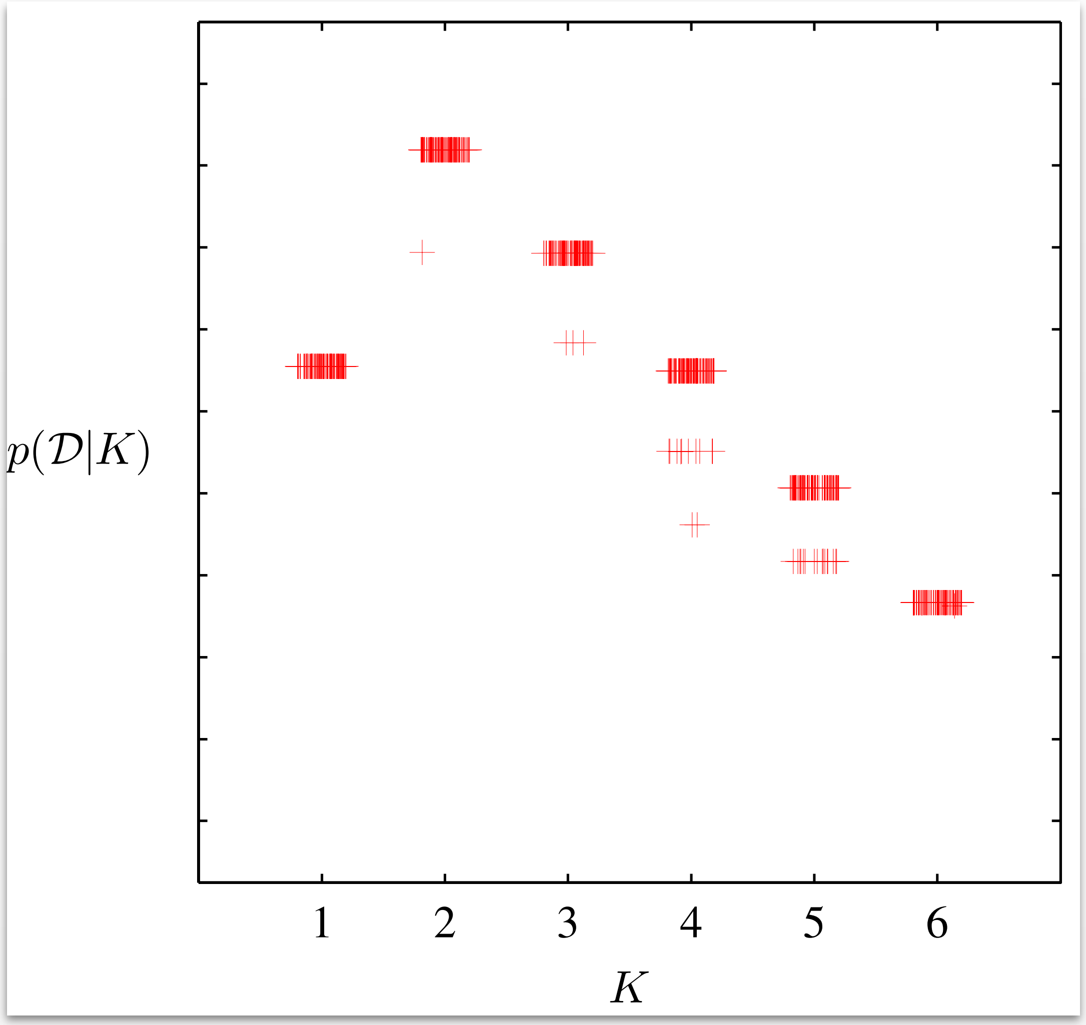
Strategy 1: Cross-validation

Split data in to K folds.

For each fold k

- Perform EM to learn θ from training set X^{train}
- Calculate test set likelihood $p(X^{test} \mid \theta)$

Model Selection



Strategy 2: Model Evidence

Define a prior p($\boldsymbol{\theta}$) and evaluate the marginal likelihood

$$p(\boldsymbol{X}) = \int d\boldsymbol{\theta}\, p(\boldsymbol{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$$

Two families of methods

- Variational Inference
- Importance Sampling

Variational Inference (Sketch)

Lower bound on Log Evidence

$$
\begin{aligned}
\mathcal{L}(q(\boldsymbol{z}), q(\boldsymbol{\theta})) &= \int d\boldsymbol{\theta} \sum_{\boldsymbol{z}} q(\boldsymbol{\theta}) q(\boldsymbol{z}) \log \frac{p(\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}) q(\boldsymbol{z})} \\
&= \log p(\boldsymbol{X}) - KL(q(\boldsymbol{\theta}) q(\boldsymbol{z}) \,||\, p(\boldsymbol{z}, \boldsymbol{\theta} \,|\, \boldsymbol{X}))
\end{aligned}
$$

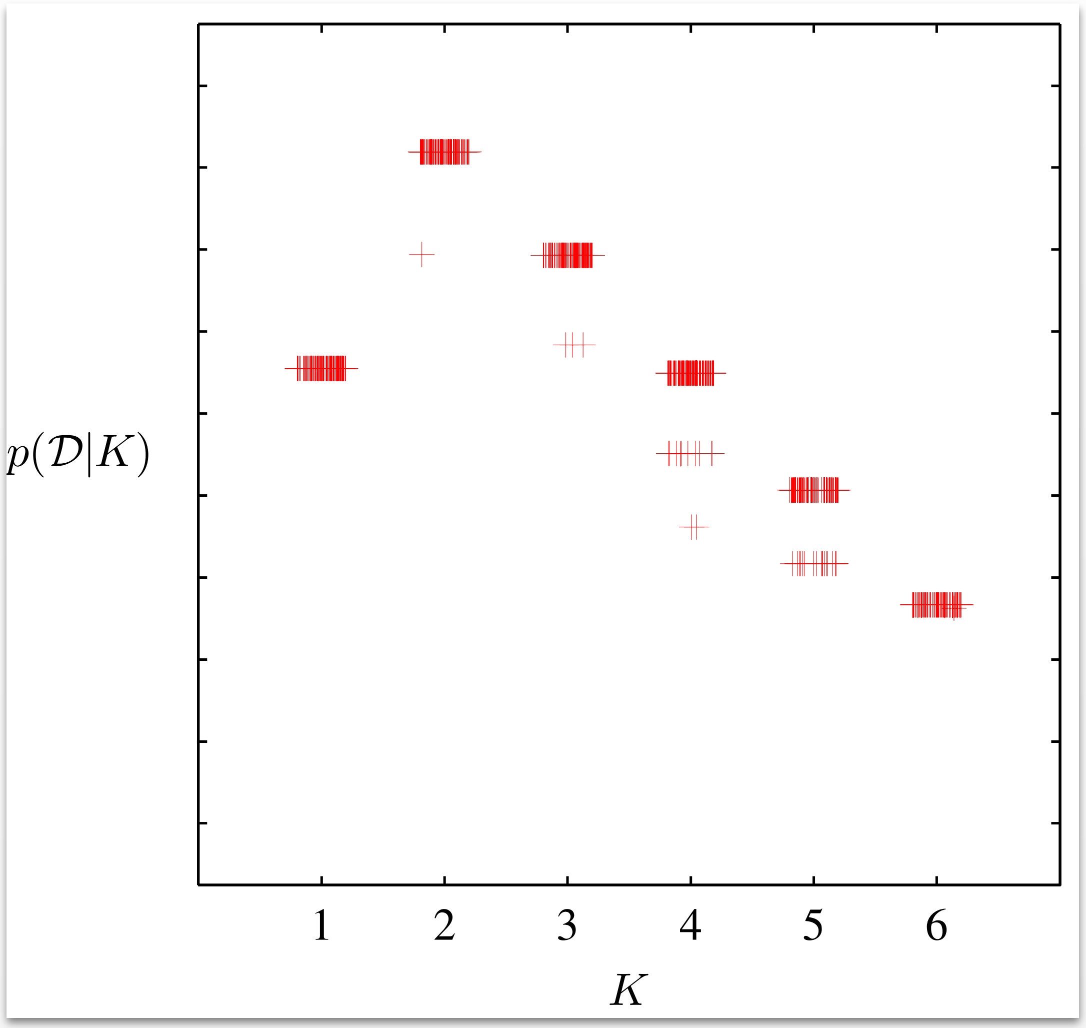
Variational E-step

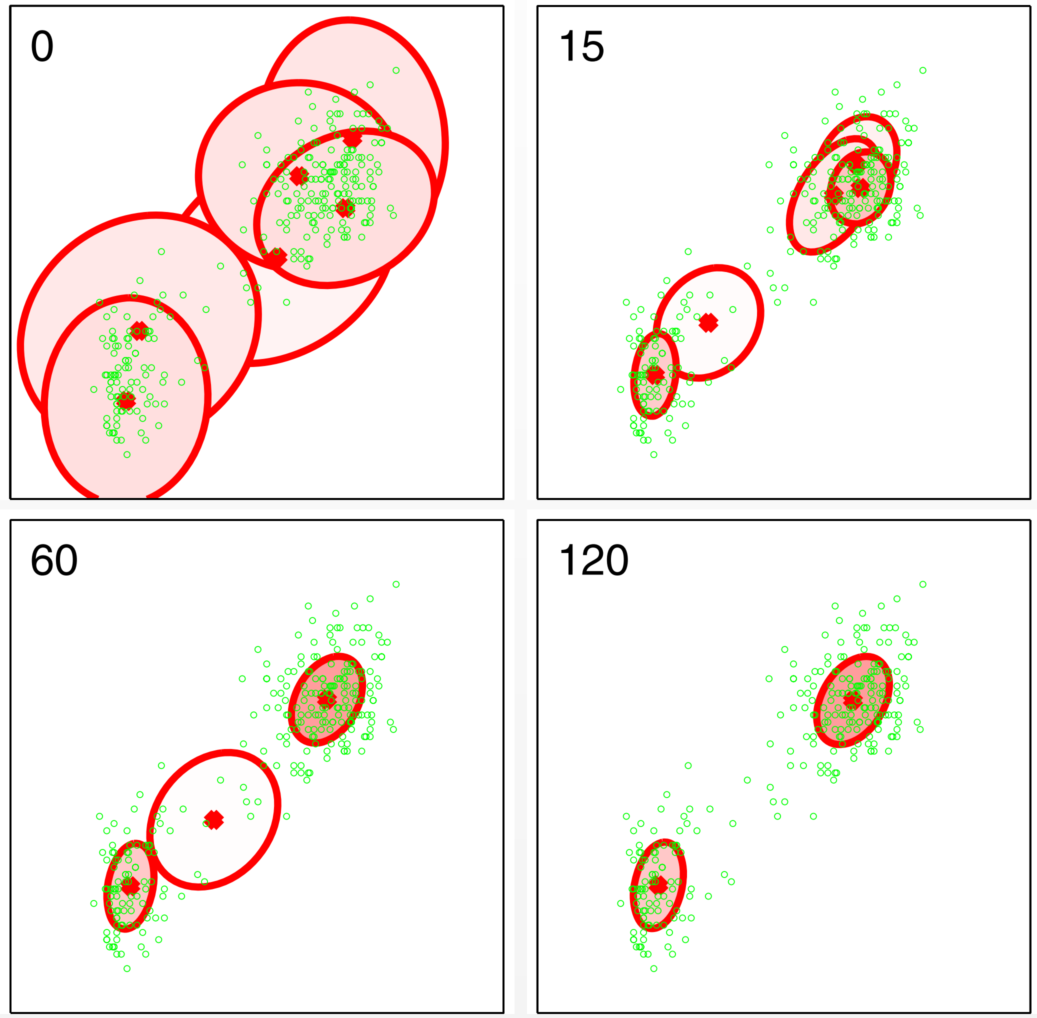
$$
q^i(\boldsymbol{z}) = \underset{q(\boldsymbol{z})}{\operatorname{argmax}}\ \mathcal{L}(q(\boldsymbol{z}), q^{i-1}(\boldsymbol{\theta}))
$$

Variational M-step

$$
q^i(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})}{\operatorname{argmax}}\ \mathcal{L}(q^i(\boldsymbol{z}), q(\boldsymbol{\theta}))
$$

Variational Inference (Sketch)





Can use lower bound on evidence to select best model

Variational inference for often assigns zero weight to superfluous components