Advanced machine learning

MTH793P 2024

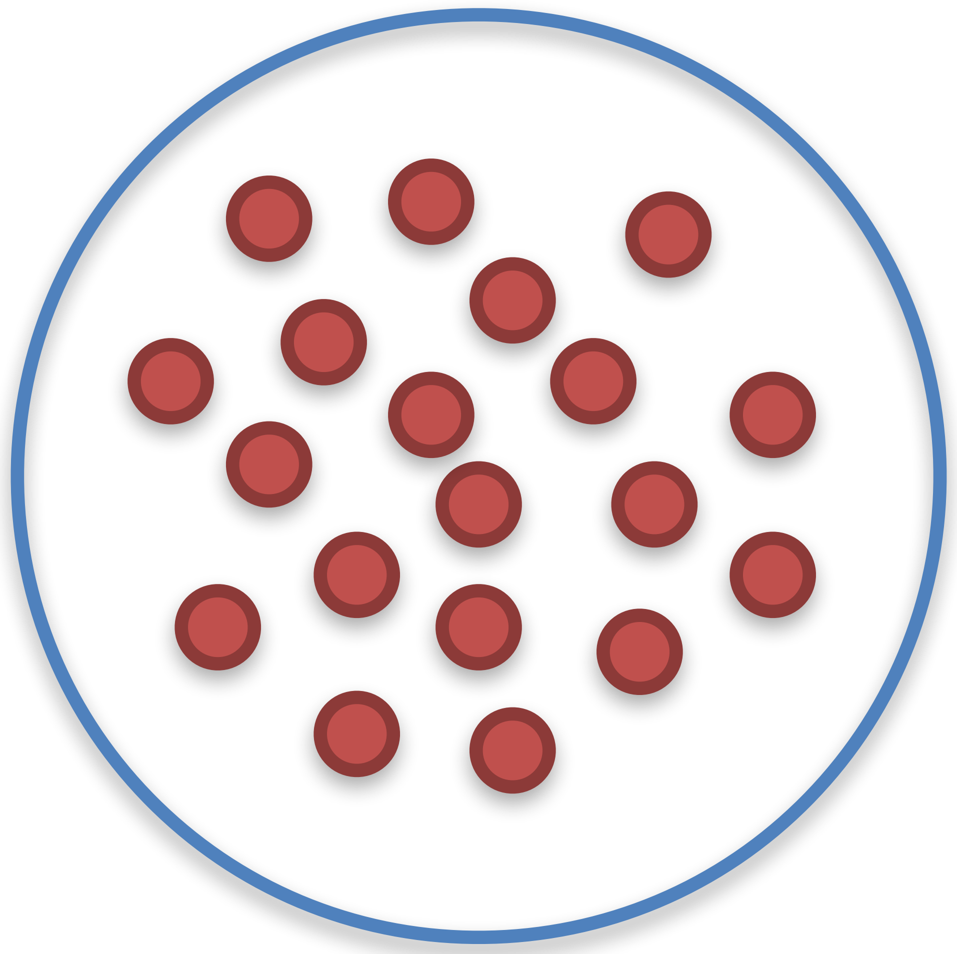
Omer Bobrowski, QMUL

GAUSSIAN MIXTURE MODELS

Gaussian mixture models

Motivation:

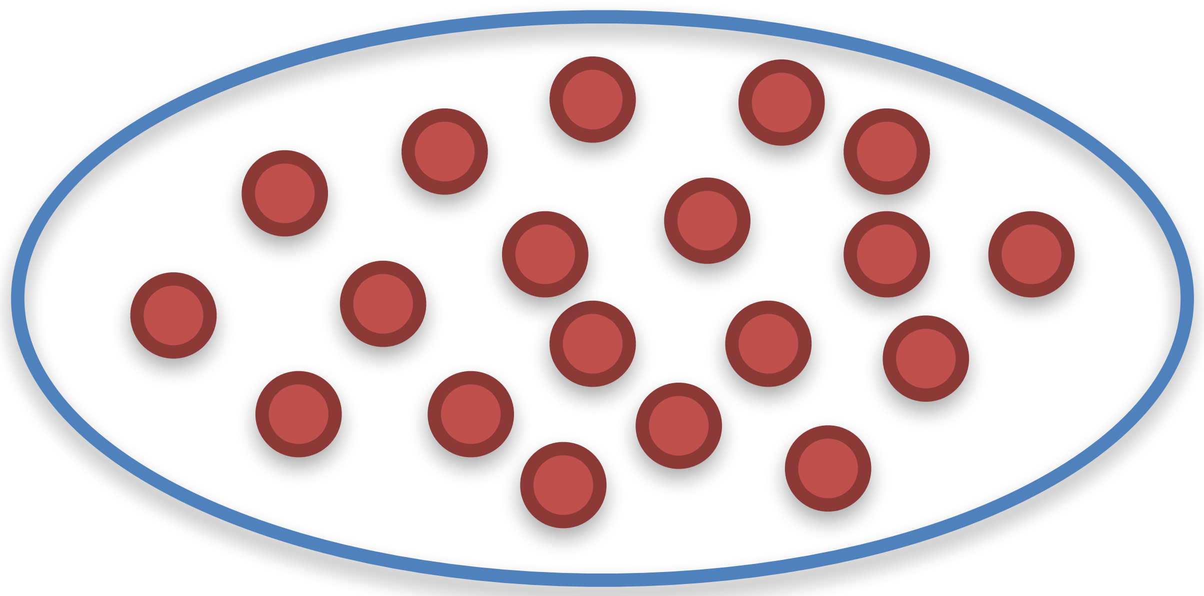
- K-means forces clusters to be "spherical"
- Sometimes it might be desirable to have elliptical clusters



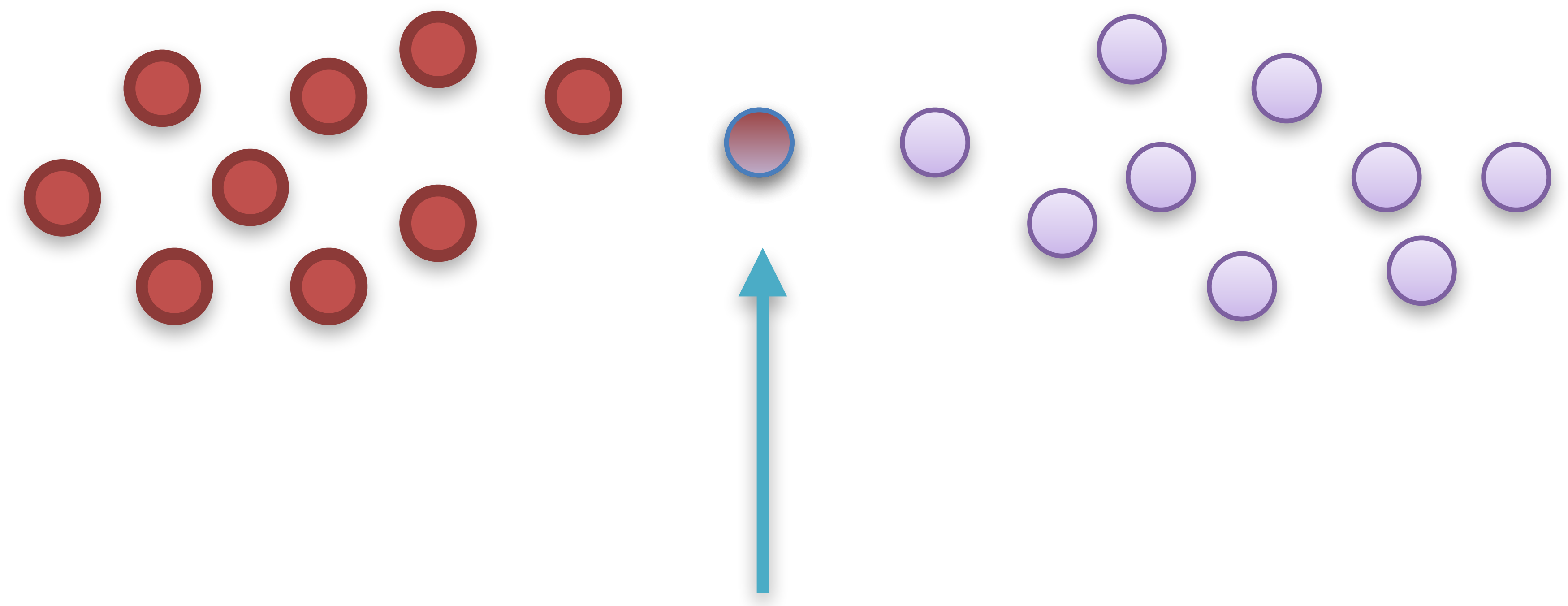
Gaussian mixture models

Motivation:

- K-means forces clusters to be "spherical"
- Sometimes it might be desirable to have elliptical clusters



Gaussian mixture models



In class 1 or in class 2?

Hard partition: every point is associated with a single cluster
(K-means: $z_{ik} \in \{0,1\}$)

Soft partition: for every point and every cluster we have a "score" in [0,1]
(GMM, next)

Gaussian mixture models

n-dimensional Gaussian density:

mean vector

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \qquad x,\mu \in \mathbb{R}^d \quad \Sigma \in \mathbb{R}^{d\times d}$$

covariance matrix

Gaussian mixture:

$$p(x) = \sum_{k=1}^{K} \rho_k \mathcal{N}(x;\mu_k,\Sigma_k), \qquad \rho_k \geq 0, \text{ and } \sum_{k=1}^{K} \rho_k = 1$$

Gaussian mixture models

$$p(x) = \sum_{k=1}^{K} \rho_k \mathcal{N}(x; \mu_k, \Sigma_k), \qquad \rho_k \geq 0, \text{ and } \sum_{k=1}^{K} \rho_k = 1$$

Interpretation

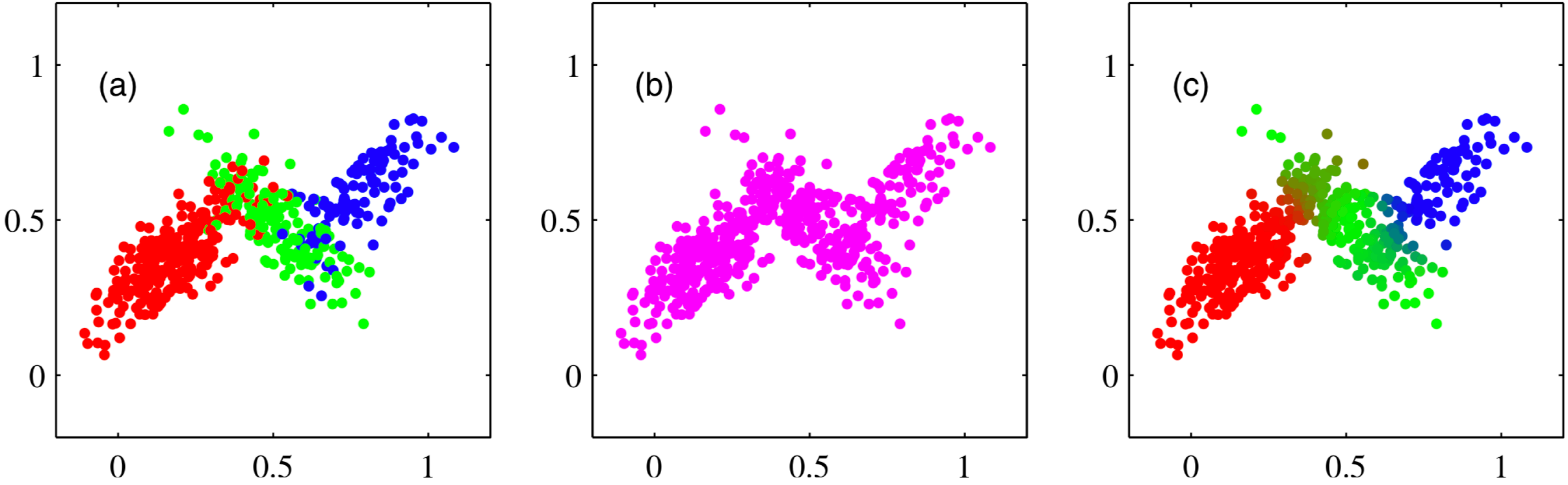
Every $X \sim p$ can be generated as follows:

1. Generate a random value $Z \in \{1,2,\ldots,K\}$, using $P(Z = k) = \rho_k$
2. If $Z = k$: sample X from $\mathcal{N}(\mu_k, \Sigma_k)$

Gaussian mixture models

From Bishop. Pattern Recognition & Machine Learning

Gaussian mixture models



From Bishop. Pattern Recognition & Machine Learning

Gaussian mixture models

$$p(x) = \sum_{k=1}^{K} \rho_k \mathcal{N}(x; \mu_k, \Sigma_k), \qquad \rho_k \geq 0, \text{ and } \sum_{k=1}^{K} \rho_k = 1$$

Likelihood

- Parameters: $\theta = (\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \rho_1, \ldots, \rho_K)$
- Likelihood: $p(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \rho_k \mathcal{N}(x_i; \mu_k, \Sigma_k)$
- Log-likelihood: $\log p(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \rho_k \mathcal{N}(x_i; \mu_k, \Sigma_k) \right)$

Gaussian mixture models

$$\log p(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \rho_k \mathcal{N}(x_i; \mu_k, \Sigma_k) \right)$$

can we maximize?

Solution (take derivatives, compare to zero):

Define:

$$\gamma_{i,k} = \frac{\rho_k \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \rho_j \mathcal{N}(x_i; \mu_j, \Sigma_j)}$$

probability that i^{th} data-point is in k^{th} cluster

$$n_k = \sum_{i=1}^{n} \gamma_{i,k}$$

total "mass" of k^{th} cluster

Maximum:

$$\rho_k = \frac{n_k}{n} \qquad \mu_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} x_i \qquad \Sigma_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} (x_i - \mu_k)(x_i - \mu_k)^T$$

no explicit solution!

Gaussian mixture models

Iterative solution (similar to K-means):

Step 1 - associate clusters (soft):

$$\gamma_{i,k} = \frac{\rho_k \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \rho_j \mathcal{N}(x_i; \mu_j, \Sigma_j)}, \quad n_k = \sum_{i=1}^{n} \gamma_{i,k}$$

(ρ_k, μ_k, Σ_k are fixed)

Step 2 - update GMM parameters:

$$\rho_k = \frac{n_k}{n} \quad \mu_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} x_i \quad \Sigma_k = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} (x_i - \mu_k)(x_i - \mu_k)^T$$

($\gamma_{i,k}, n_k$ are fixed)

Algorithm 1 GMM clustering

Input: $x_1, \dots, x_n \in \mathbb{R}^d$

Initialise: $\theta^0 = (\mu_1^0, \dots, \mu_K^0, \Sigma_1^0, \dots, \Sigma_K^0, \rho_1^0, \dots, \rho_K^0)$

while stopping condition not met **do**

step 1

for $i = 1, \dots, n$ **do**

  for $k = 1, \dots, K$ **do**

    $\gamma_{i,k}^{t+1} \leftarrow \frac{\rho_k^t \mathcal{N}(x_i; \mu_k^t, \Sigma_k^t)}{\sum_{j=1}^K \rho_j^t \mathcal{N}(x_i; \mu_j^t, \Sigma_j^t)}$ $\qquad n_k^{t+1} \leftarrow \sum_{i=1}^n \gamma_{i,k}^{t+1}$

  end for

end for

associate clusters (soft)

step 2

for $k = 1, \dots, K$ **do**

  $\mu_k^{t+1} \leftarrow \frac{1}{n_k^{t+1}} \sum_{i=1}^n \gamma_{i,k}^{t+1} x_i \,,$

  $\Sigma_k^{t+1} \leftarrow \frac{1}{n_k^{t+1}} \sum_{i=1}^n \gamma_{i,k}^{t+1} \left(x_i - \mu_k^{t+1}\right) \left(x_i - \mu_k^{t+1}\right)^\top ,$

  $\rho_k^{t+1} = \frac{n_k^{t+1}}{n}$

end for

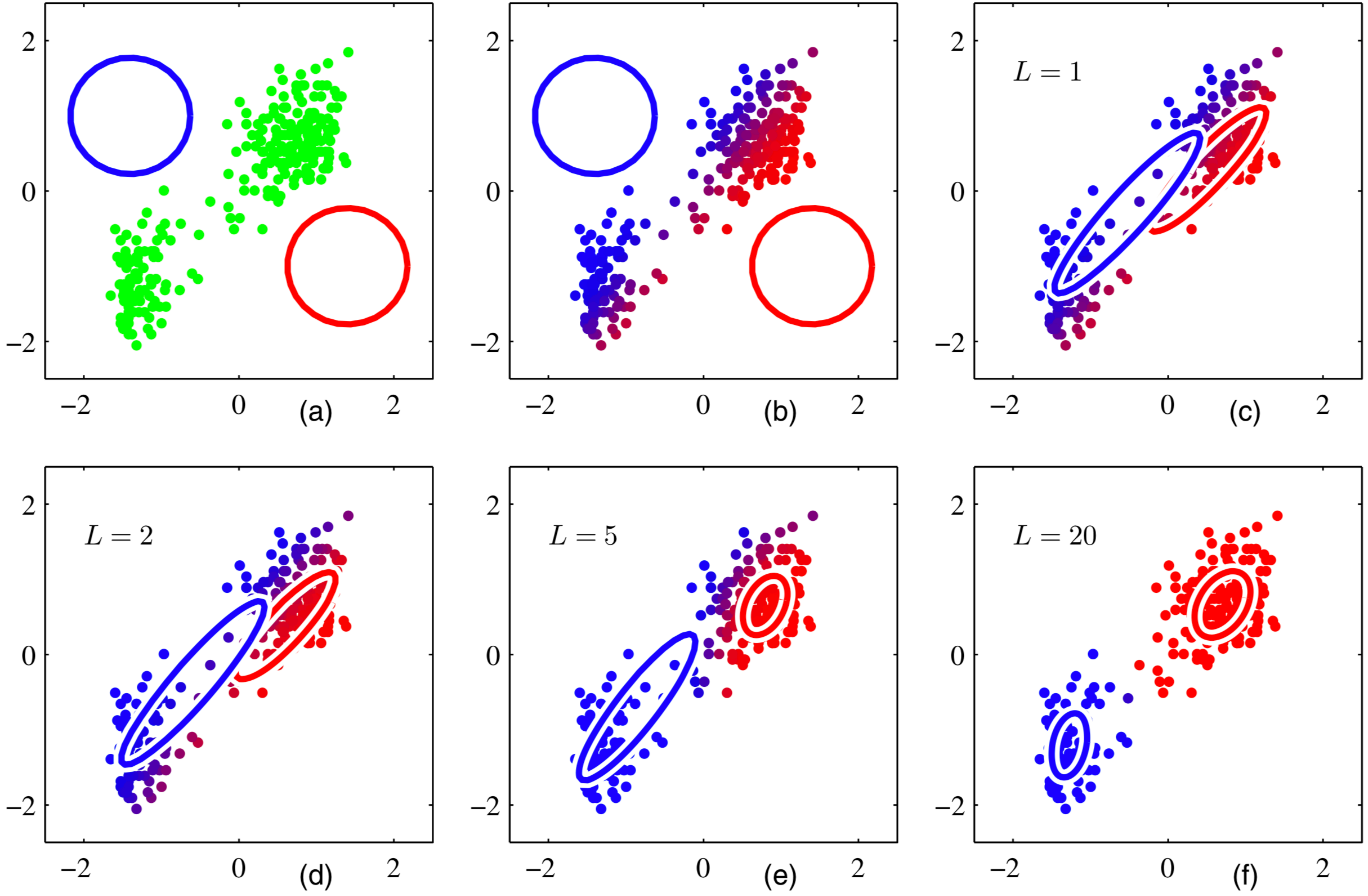
update GMM parameters

  $t \leftarrow t + 1$

end while

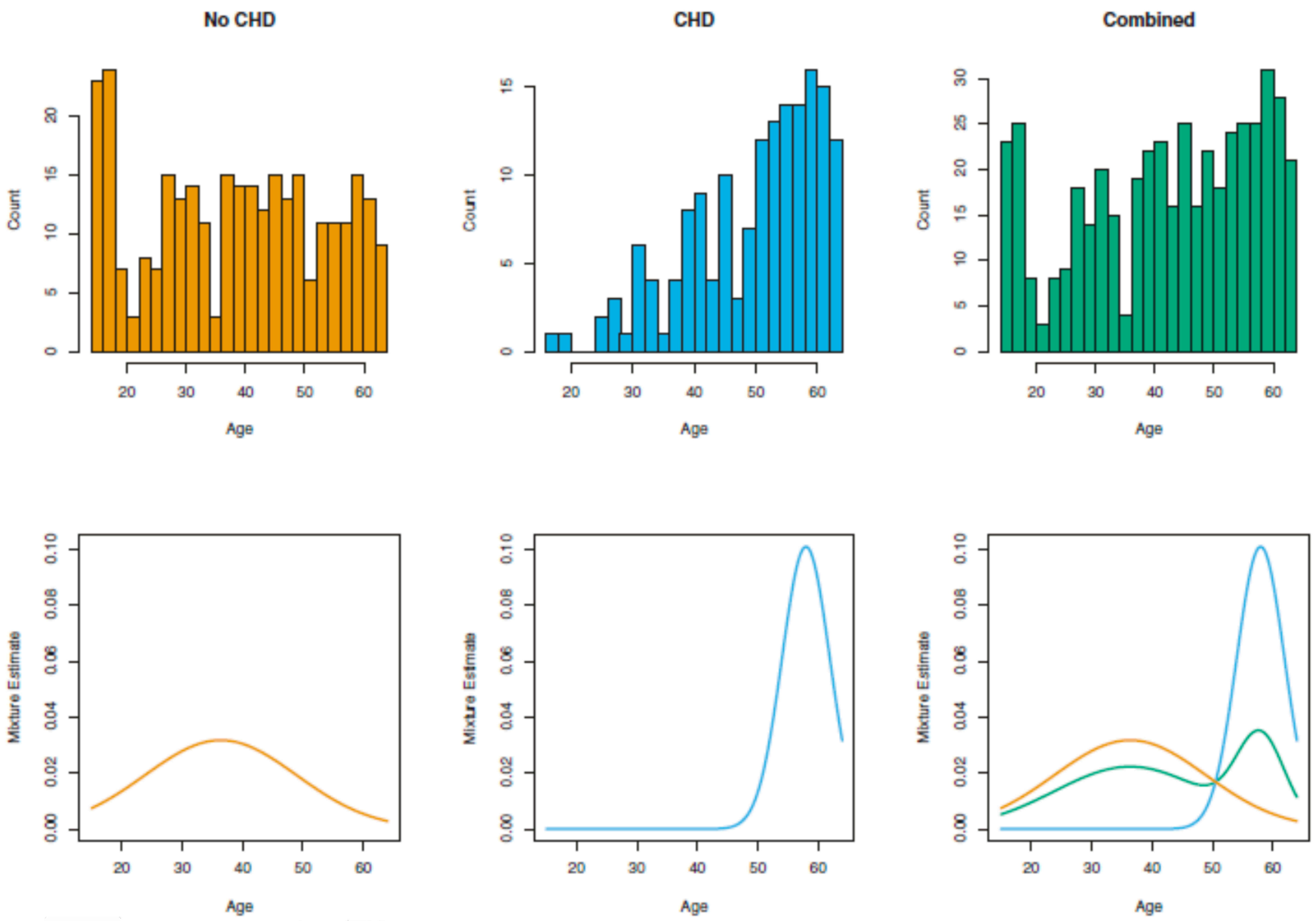
return θ^M

Gaussian mixture models



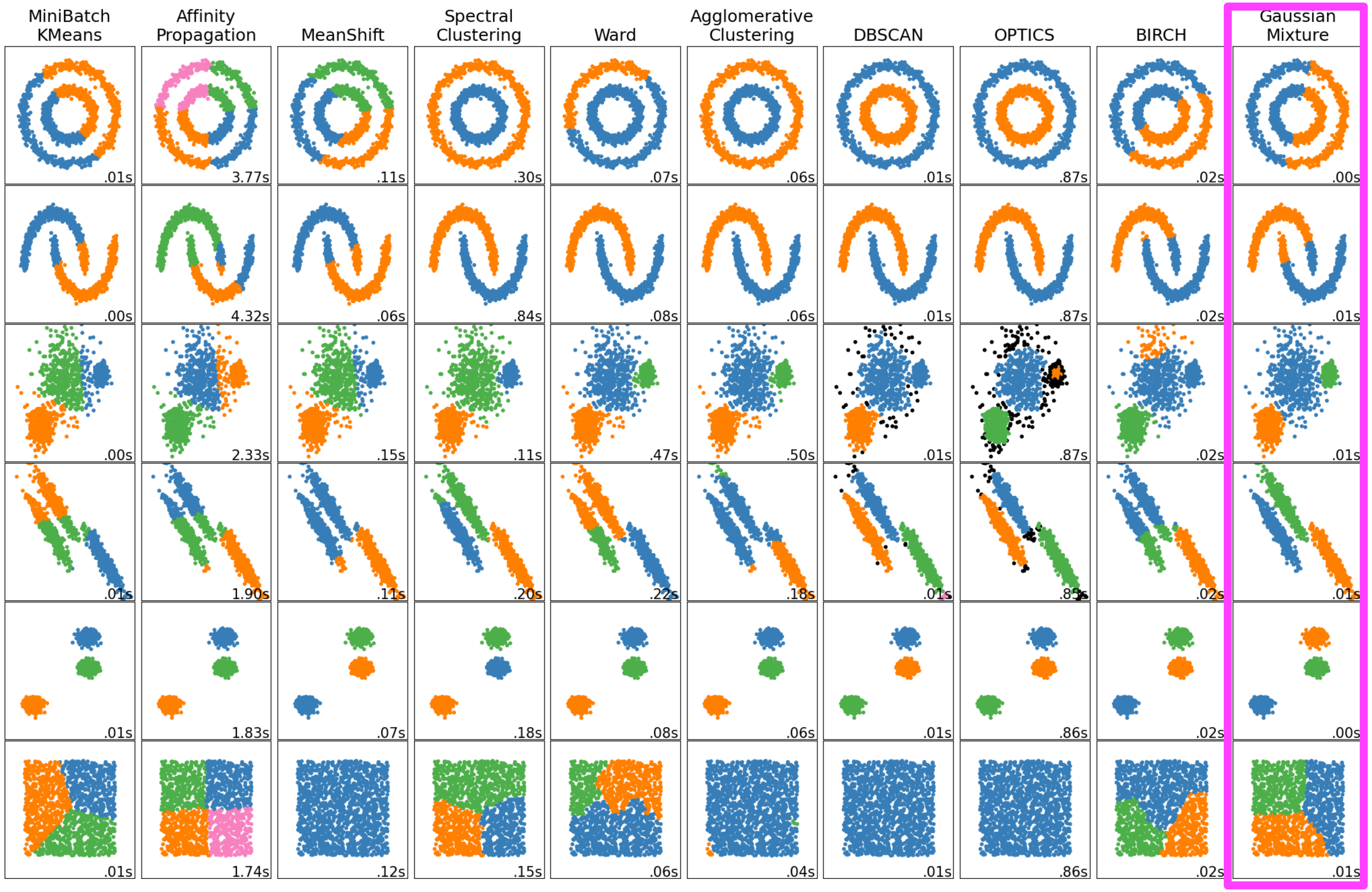
From Bishop. Pattern Recognition & Machine Learning

Gaussian mixture models

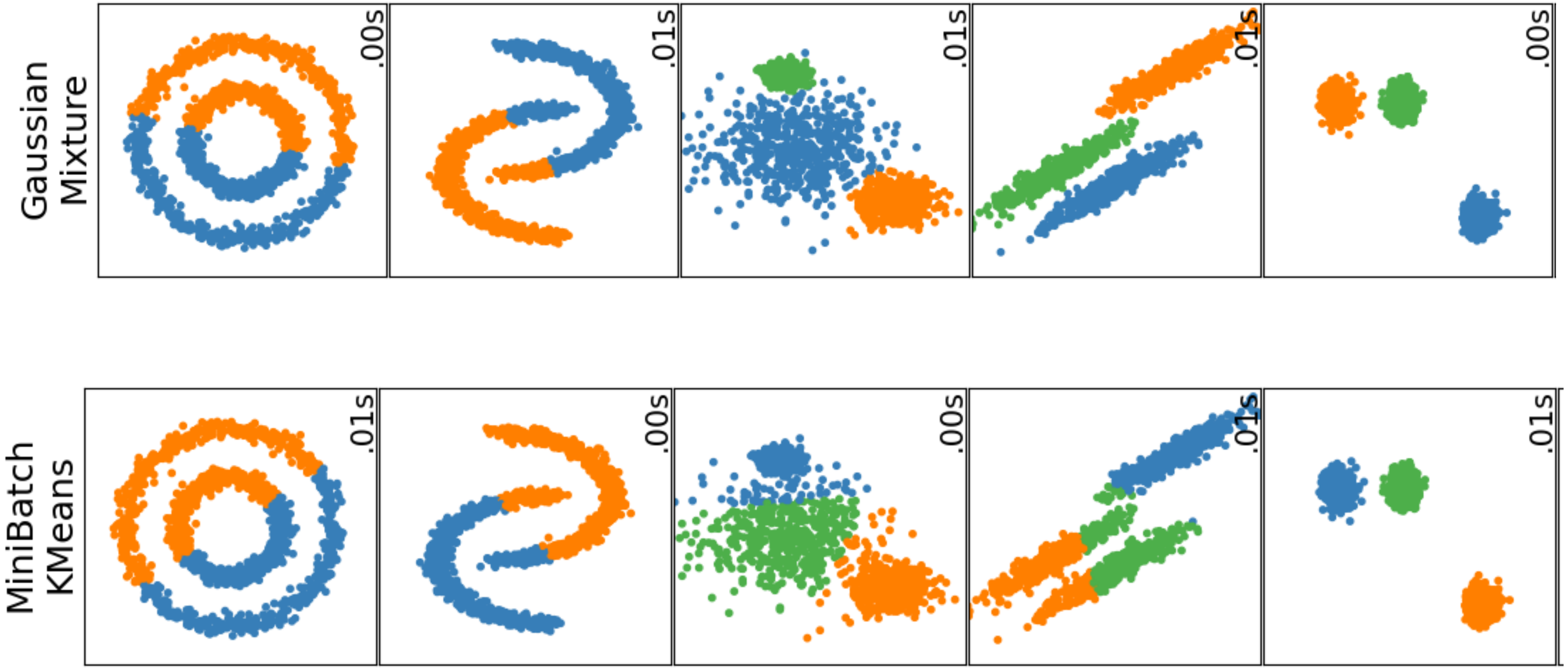


From Hastie, Tibshirani and Friedman, The Elements of Statistical Learning, 2nd edition, 2009

Various Clustering Algorithms



Taken from scikit-learn python package documentation


Gaussian Mixture
.00s
.01s
.01s
.01s
.00s
MiniBatch KMeans
.01s
.00s
.00s
.01s
.01s

Comments

- Allows for more precise clustering
- Heavier computation than K-means
- Takes longer to converge than K-means
- Likelihood increases in every iteration, but may converge to local maximum
- Common practice:
  - Run K-means
  - Use results to initialise the parameter for the GMM