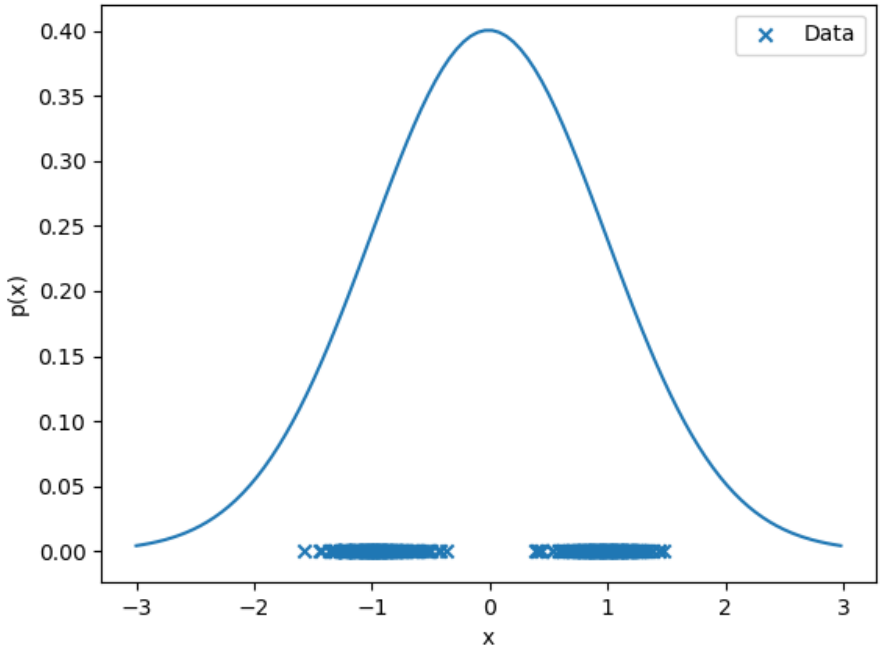
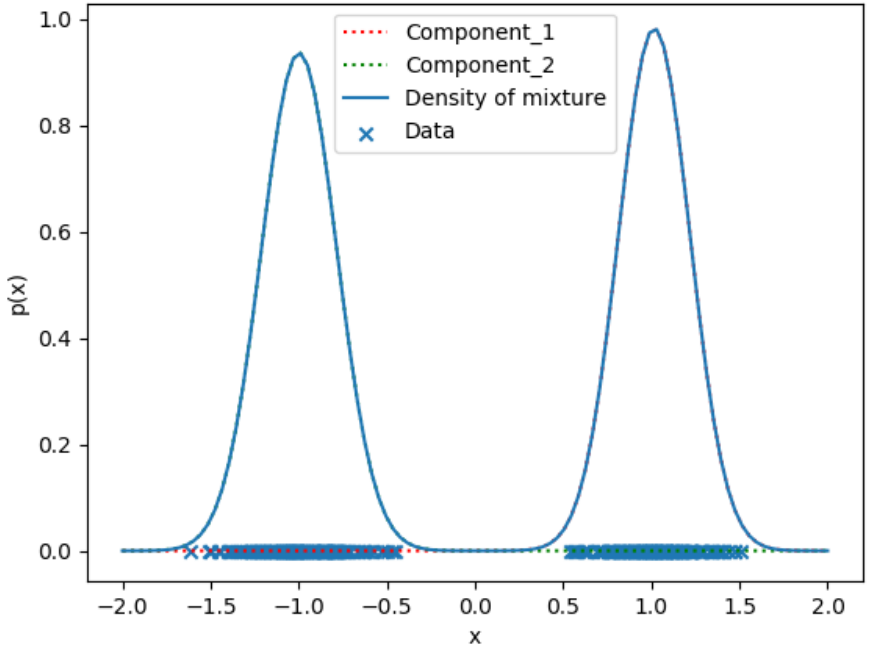
APPLIED MACHINE LEARNING

Gaussian Mixture Models

Expectation-Maximization

Mixture of Gauss Functions





Combination of K Gauss functions

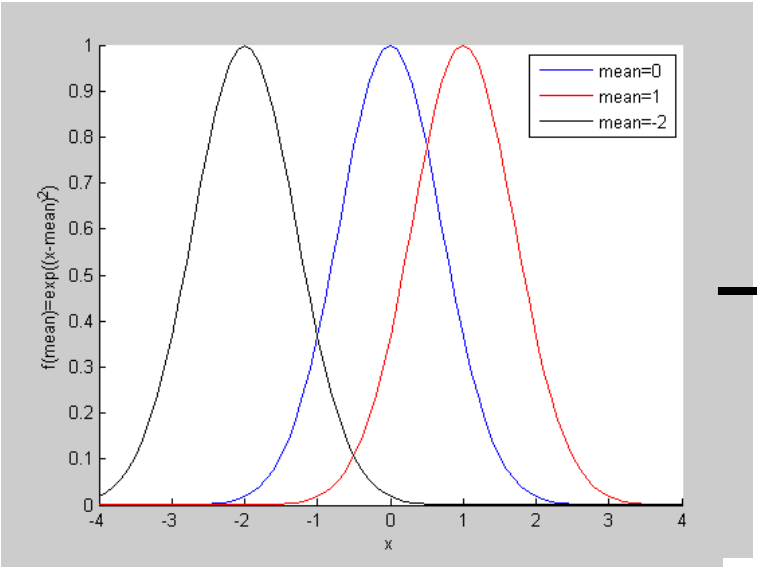
$$p(x;\Theta) = \sum_{k=1}^{K} \alpha_k p_k\left(x;\mu^k,\Sigma^k\right) \text{ with } \Theta = \{\mu^1,\Sigma^1,\ldots\mu^K,\Sigma^K\},\ \alpha_k \in [0,1].$$

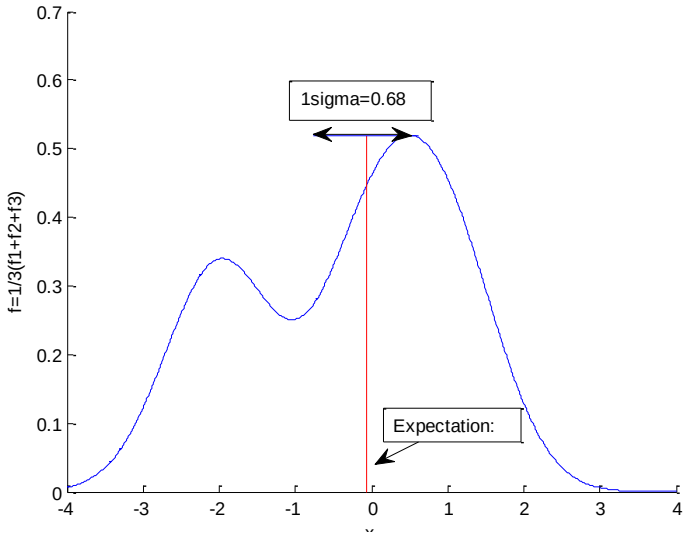
Linear weighted combination

K Gaussian Functions

Mixture of Gauss Functions

Here we use $K = 3,\ \alpha_1 = \alpha_2 = \alpha_3$





Superposition of the 3 Gauss functions with equal weight.

To find the optimal parameters:

$$\max_{\Theta} L(\Theta \mid X) = \max_{\Theta} p(X \mid \Theta)$$

Combination of K Gauss functions

$$p(x;\Theta) = \sum_{k=1}^{K} \alpha_k p_k\left(x;\mu^k,\Sigma^k\right) \text{ with } \Theta = \{\mu^1,\Sigma^1,\ldots\mu^K,\Sigma^K\},\ \alpha_k \in [0,1].$$

Linear weighted combination

K Gaussian Functions

Gaussian Mixture Modeling with E-M

E-M searches for optimum of the likelihood of the model given the data, i.e.:

$$\max_{\Theta} \; L(\Theta \mid X) = \max_{\Theta} \; p(X \mid \Theta)$$

The parameters of a GMM are the means, covariance matrices and priors:

$$\Theta = \left\{ \mu^1, ... \mu^K, \sum^1, ... \sum^K, \alpha_1, ... \alpha_K \right\}$$

Gaussian Mixture Modeling with E-M

One usually can safely assume that the datapoints are i.i.d. (identically and independently distributed).

$$\max_{\Theta} \; p(X \mid \Theta) = \max_{\Theta} \; \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k \cdot p\left(x^i; \mu^k, \Sigma^k\right)$$

Computing the log of the likelihood yields the same optimum:

$$\max_{\Theta} \; p(X \mid \Theta) = \max_{\Theta} \; \log p(X \mid \Theta)$$

$$\max_{\Theta} \; \log \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k \cdot p\left(x^i; \mu^k, \Sigma^k\right) = \max_{\Theta} \; \sum_{i=1}^{M} \log \left(\sum_{k=1}^{K} \alpha_k \cdot p\left(x^i; \mu^k, \Sigma^k\right) \right)$$

No closed-form solution → Solve through Expectation-Maximization (E-M)
E-M is an *iterative* procedure to estimate the best set of parameters
It converges to **a local optimum** → Sensitive to initialization!
See derivation of E-M for GMM in the annexes posted on the website

Expectation-Maximization (E-M)

EM is an iterative procedure:

0) Make a guess, pick a set of $\hat{\Theta}$ (initialization)

1) Compute likelihood $L\left(\hat{\Theta} \mid X, Z\right)$ (E-Step)

2) Update Θ by gradient ascent on $L\left(\Theta \mid X, Z\right)$

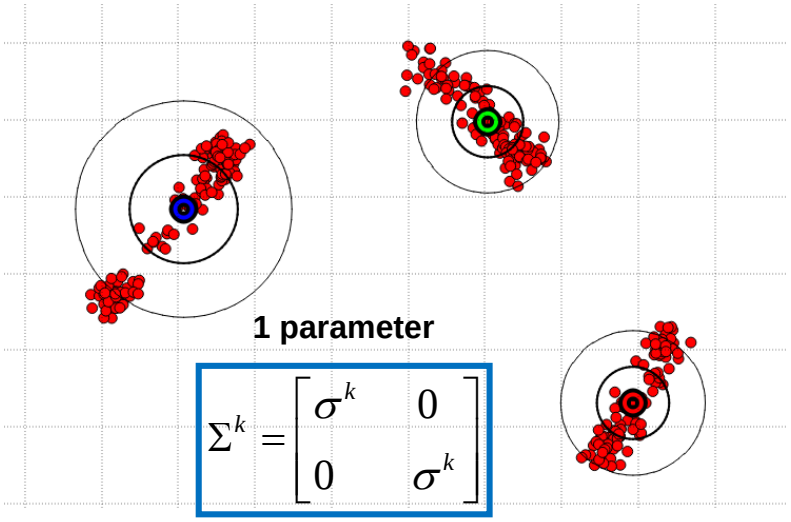
3) Iterate between steps 1 and 2 until reach plateau (no improvement on likelihood)

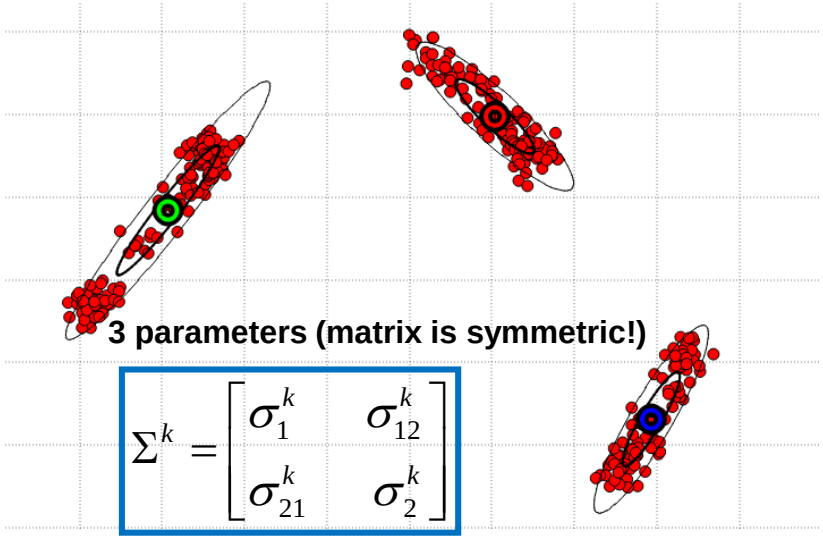
Ensured to converge to a local optimum only!

Tradeoff between computation costs and better fit

- A GMM can fit very closely the local distribution of datapoints.
- But this comes at the cost of an increase in the number of parameters

Full covariance matrices require N(N+1)/2 parameters against N for diagonal matrices and 1 for spherical matrices.*





How to determine the best mixtures of Gaussians?

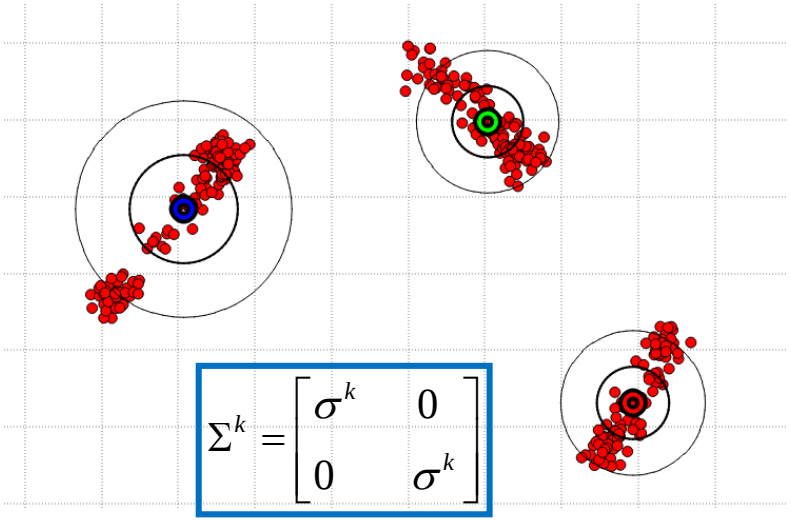
Hyper-parameter optimization in GMMs

The selection is performed using the AIC and BIC criteria.

$$AIC = -2\ln(L) + 2B$$

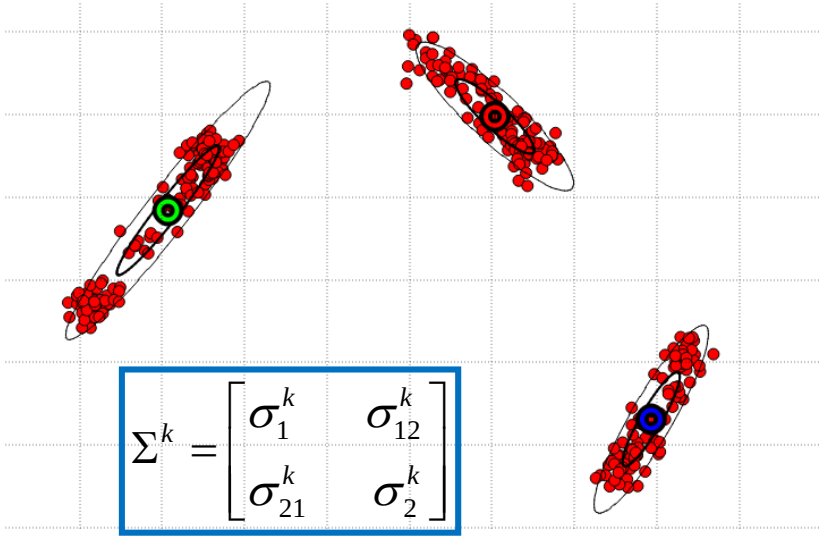
$$BIC = -2\ln(L) + \ln(M)B$$

B: number of parameters of the mixture.



$$\Sigma^k = \begin{bmatrix} \sigma^k & 0 \\ 0 & \sigma^k \end{bmatrix}$$

$B = 3*(2+1) = 9$



$$\Sigma^k = \begin{bmatrix} \sigma_1^k & \sigma_{12}^k \\ \sigma_{21}^k & \sigma_2^k \end{bmatrix}$$

$B = 3*(2+3) = 15$

B depends on the choice of covariance matrix

Hyper-parameter optimization in GMMs