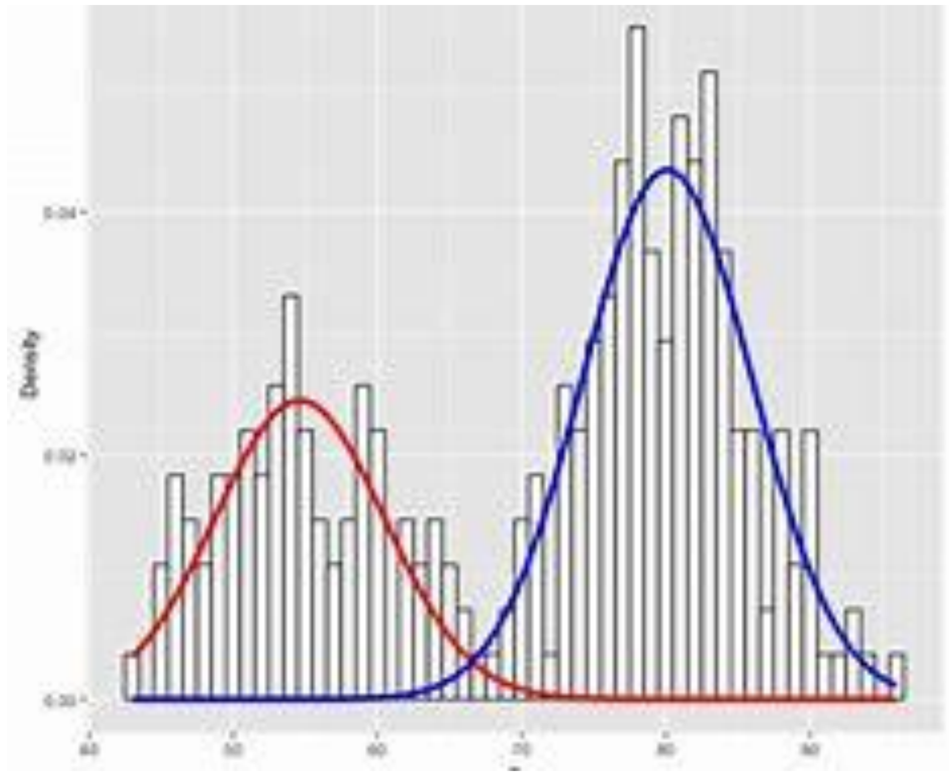
« BAYESIAN LEARNING »

2. MIXTURE MODEL

Stéphane Derrode, Dpt MI
Stephane.derrode@ec-lyon.fr

Contents for *2. Mixture model*

1. Definition / Sampling
2. Automatic parameter learning (EM algo)
3. 2D mixture models



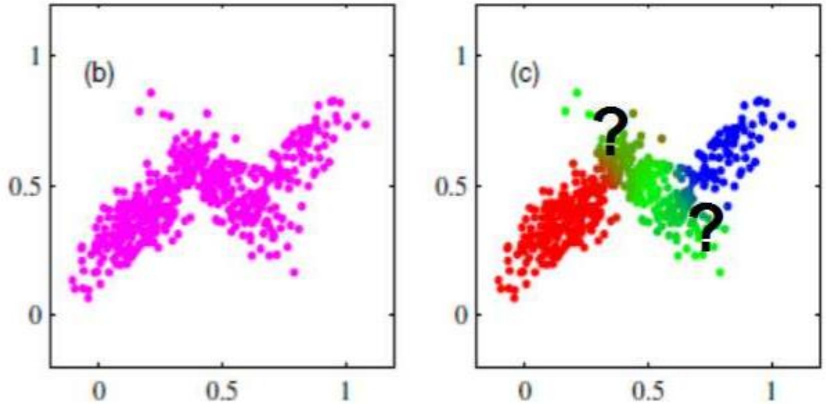
MIXTURE MODEL

Definition / sampling

Introduction

In statistics, a **mixture model** is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs.

Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-population identity information.



Definition

Suppose that we have a sample

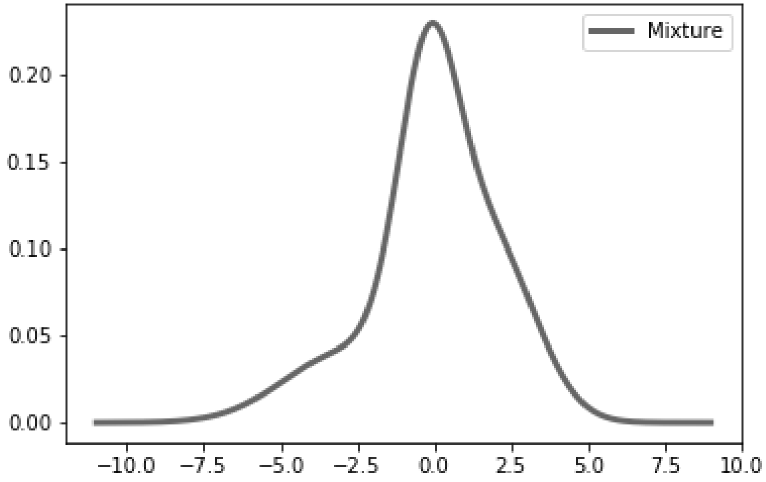
$$\boldsymbol{y} = \boldsymbol{y}_1^N = \{y_1, y_2, \ldots, y_n, \ldots, y_N\}$$

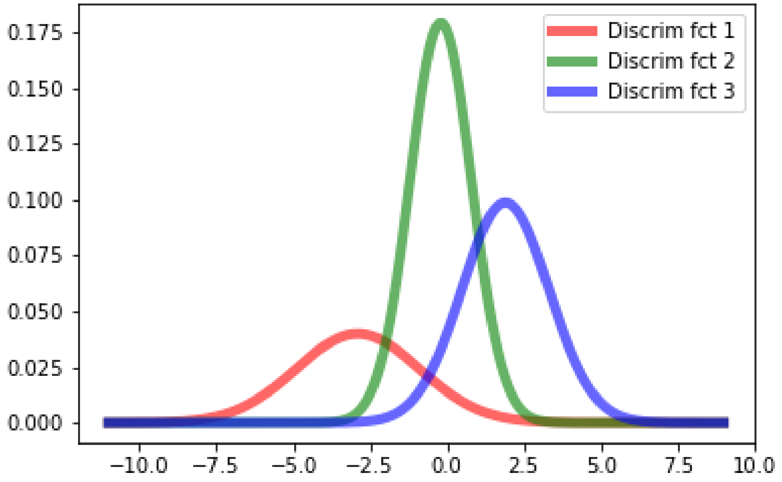
distributed according to a mixture of Gaussian distributions, so that all samples have the following with density:

$$P(Y_n = y_n) = f(y_n) = \sum_{k=1}^{K} \pi_k f_k(y_n)$$

A Gaussian mixture model is made with Gaussian f_k .

Example





$$\mathcal{N}\left(\mu_1 = -2.9, \sigma_1 = 2\right) \qquad \pi_1 = 0.10$$

$$\mathcal{N}\left(\mu_2 = -0.2, \sigma_2 = 1\right) \qquad \pi_2 = 0.55$$

$$\mathcal{N}\left(\mu_3 = 1.9, \sigma_3 = \sqrt{2}\right) \qquad \pi_3 = 0.35$$

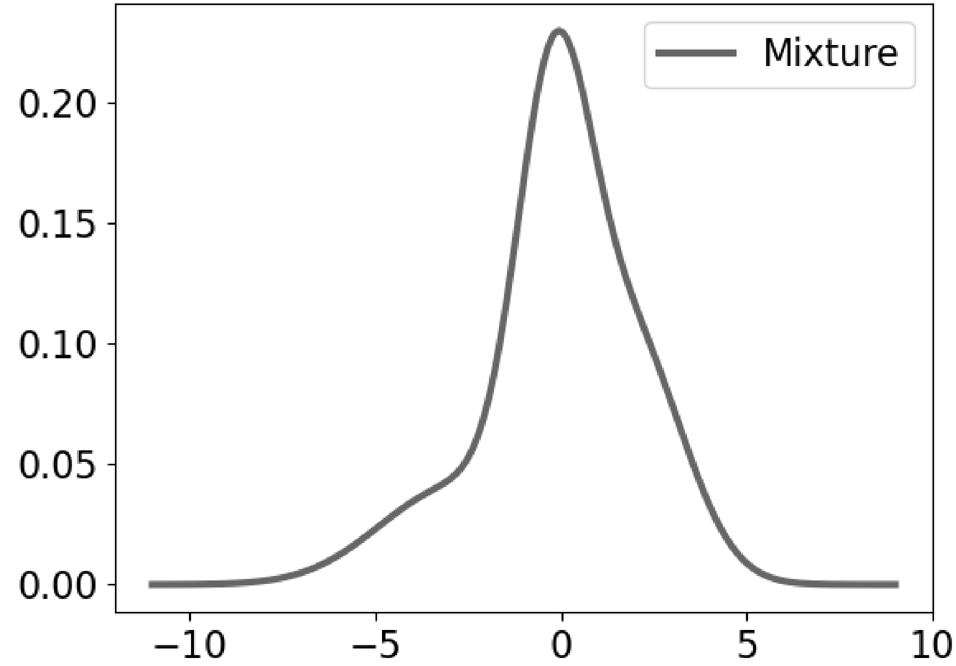
Question : How to draw a sample for a mixture ?

$$P(Y_n = y_n) = f(y_n) = \sum_{k=1}^{K} \pi_k f_k(y_n)$$

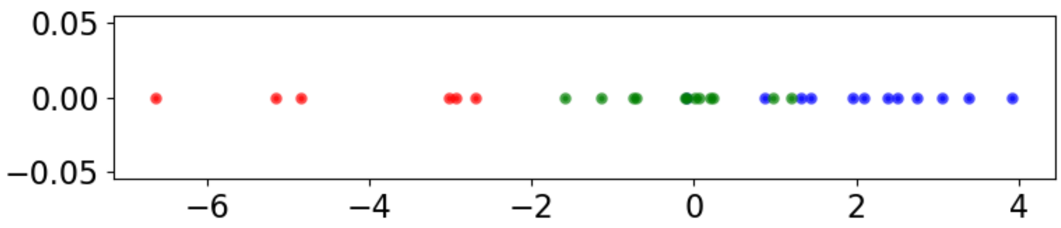
$$\mathcal{N}\left(\mu_1 = -2.9, \sigma_1 = 2\right) \quad \pi_1 = 0.10$$

$$\mathcal{N}\left(\mu_2 = -0.2, \sigma_2 = 1\right) \quad \pi_2 = 0.55$$

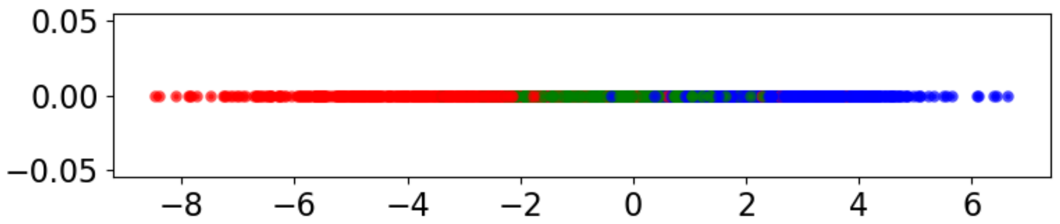
$$\mathcal{N}\left(\mu_3 = 1.9, \sigma_3 = \sqrt{2}\right) \quad \pi_3 = 0.35$$



30 échantillons



3000 échantillons



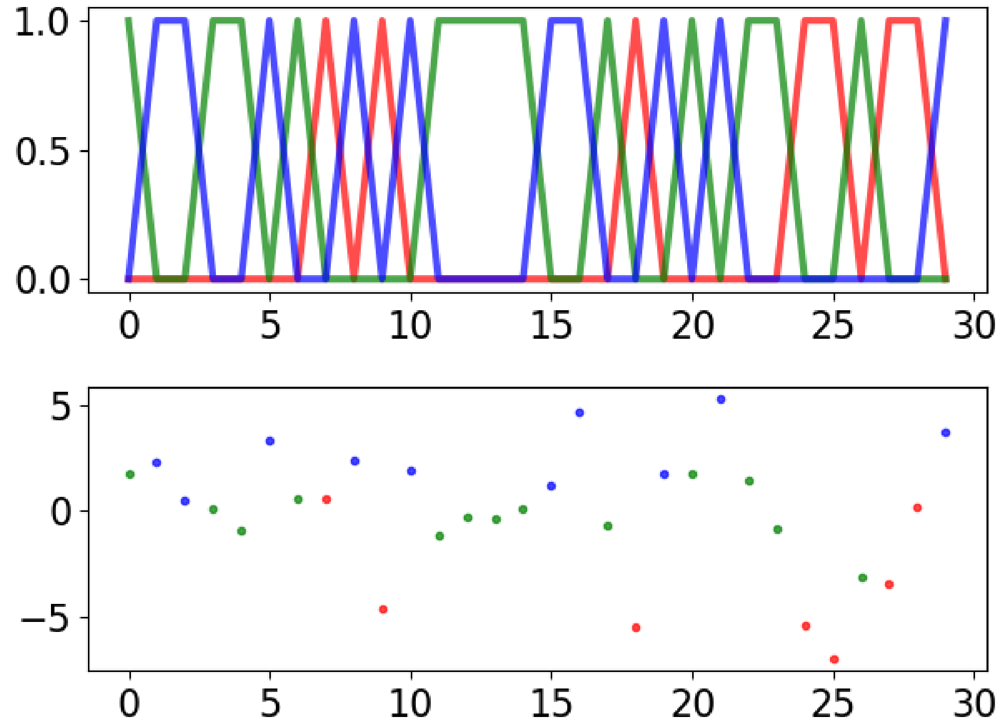
Question : How to draw a sample for a mixture ?

1. Sampling according to the a priori proba to get the class number.
2. Sampling according to the selected Gaussian.

$$\mathcal{N}\left(\mu_1 = -2.9, \sigma_1 = 2\right) \qquad \pi_1 = 0.10$$

$$\mathcal{N}\left(\mu_2 = -0.2, \sigma_2 = 1\right) \qquad \pi_2 = 0.55$$

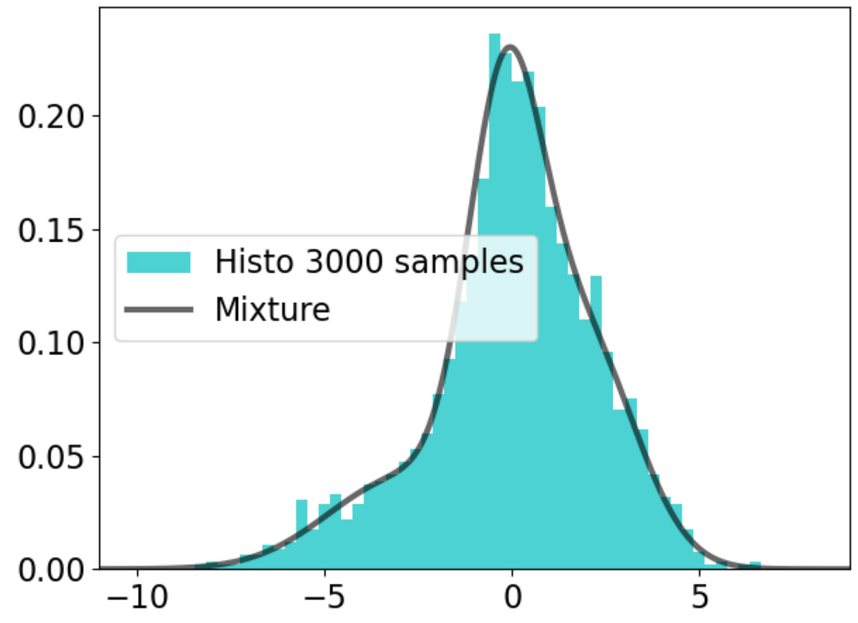
$$\mathcal{N}\left(\mu_3 = 1.9, \sigma_3 = \sqrt{2}\right) \qquad \pi_3 = 0.35$$



Question : How to draw a sample for a mixture ?

1. Sampling according to the a priori proba to get the class number.
2. Sampling according to the selected Gaussian.

$$\mathcal{N}\left(\mu_1 = -2.9, \sigma_1 = 2\right) \qquad \pi_1 = 0.10$$
$$\mathcal{N}\left(\mu_2 = -0.2, \sigma_2 = 1\right) \qquad \pi_2 = 0.55$$
$$\mathcal{N}\left(\mu_3 = 1.9, \sigma_3 = \sqrt{2}\right) \qquad \pi_3 = 0.35$$



MIXTURE MODEL

Automatic parameter learning

Likelihood

In statistics, the **likelihood** expresses how probable a given set of observations is for different values of statistical parameters. It is equal to the joint probability distribution of the random sample evaluated at the given observations, and it is, thus, solely a function of parameters that index the family of those probability distributions.

For MM $$\mathcal{L}_{\Theta}(\boldsymbol{y}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \, f_k(y_n)$$

considered as a function of Θ.

Sir Ronald Fisher

Maximum likelihood estimator

Mapping from the parameter space to the real line, the likelihood function presents a peak, if it exists, which represents the combination of model parameter values that maximize the probability of drawing the sample actually obtained.

The procedure for obtaining these arguments of the maximum of the likelihood function is known as maximum likelihood estimation (MLE), which for computational convenience is usually done using the natural logarithm of the likelihood, known as the **log-likelihood function**.

Question: How to maximise $\mathcal{L}_{\Theta}(\boldsymbol{y})$?

Question: How to maximise $\mathcal{L}_\Theta(\boldsymbol{y})$?

- Direct maximization is not possible.
- Solution : Expectation-Maximization (EM) algorithm

We define the "joint likelihood"

$$\mathcal{H}_\Theta(\boldsymbol{y}, \boldsymbol{X}) = \prod_{n=1}^{N} \pi_{X_n}\, f_{X_n}(y_n) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k\, f_k(y_n)\, \mathbb{I}_{(X_n=k)}$$

This a random function of

$$\boldsymbol{X} = \boldsymbol{X}_1^N = \{X_1, X_2, \ldots, X_n, \ldots, X_N\}$$

EM algorithm

The EM algorithm: this is an iterative algorithm to estimate the maximum of the likelihood function, by computing iteratively two steps:

$$\mathcal{Q}\left(\Theta;\Theta^{(\ell)}\right) = E\left[\ln \mathcal{H}_{\Theta}(\boldsymbol{y},\boldsymbol{X})|\boldsymbol{y},\Theta^{(\ell)}\right]$$

1. <u>Expectation of the auxiliary function</u>

where

- Θ is the set of true parameters (we are looking for).
- $\Theta^{(\ell)}$ is the estimated parameters set at iteration ℓ .

2. <u>Maximization of the auxiliary function</u>

$$\Theta^{(\ell+1)} = \arg\max_{\Theta} \mathcal{Q}\left(\Theta;\Theta^{(\ell)}\right)$$

Properties of the EM algorithm (not proven)

1. Construction of a series of estimators for which the likelihood is increasing.

$$\mathcal{L}_{\boldsymbol{\Theta}^{(\ell+1)}}(\boldsymbol{y}) \geq \mathcal{L}_{\boldsymbol{\Theta}^{(\ell)}}(\boldsymbol{y})$$

The likelihood is always increasing (this is a sufficient condition to ensure the convergence of the EM algorithm).

Properties of the EM algorithm (not proven)

2. Convergence towards one of the (local) maxima of likelihood since we have

$$\left.\frac{\partial \mathcal{Q}\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(\ell)}\right)}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^{(\ell)}} = \left.\frac{\partial \mathcal{L}_{\boldsymbol{\Theta}}(\boldsymbol{y})}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^{(\ell)}}$$

Initialization: Biernacki, C., Celeux, G. and Govaert, G. (2003). *Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models*. Computational Statistics and Data Analysis 41, 561–575.

The joint log-likelihood is written

$$\ln \mathcal{H}_\Theta(\boldsymbol{y}, \boldsymbol{X}) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \ f_k(y_n) \ \mathbb{I}_{(X_n = k)} \right)$$

So

$$\begin{aligned} \mathcal{Q}\left(\Theta; \Theta^{(\ell)}\right) &= E\left[\ln \mathcal{H}_\Theta(\boldsymbol{y}, \boldsymbol{X}) | \boldsymbol{y}, \Theta^{(\ell)}\right] \\ &= \sum_{n=1}^{N} E\left[\ln \left(\sum_{k=1}^{K} \pi_k \ f_k(y_n) \ \mathbb{I}_{(X_n = k)} \right) \middle| \boldsymbol{y}, \Theta^{(\ell)}\right] \end{aligned}$$

Given

$$E\left[f(X_n) \middle| \boldsymbol{y}, \Theta^{(\ell)}\right] = \sum_{i=1}^{K} f(i) \ p\left(X_n = i \middle| \boldsymbol{y}, \Theta^{(\ell)}\right)$$

We get

$$\mathcal{Q}\left(\Theta; \Theta^{(\ell)}\right) = \sum_{n=1}^{N} \underbrace{\sum_{i=1}^{K} \left[\ln \left(\sum_{k=1}^{K} \pi_k^{(\ell)} \ f_k^{(\ell)}(y_n) \ \mathbb{I}_{(i = k)} \right) \ p\left(X_n = i \middle| \boldsymbol{y}, \Theta^{(\ell)}\right)\right]}_{E\left[f(X_n) \middle| \boldsymbol{y}, \Theta^{(\ell)}\right]}$$

So

$$\mathcal{Q}\left(\Theta;\Theta^{(\ell)}\right) = \sum_{n=1}^{N} \underbrace{\sum_{i=1}^{K} \left[\ln\left(\sum_{k=1}^{K} \pi_k^{(\ell)}\ f_k^{(\ell)}(y_n)\ \mathbb{I}_{(i=k)}\right) p\left(X_n = i \,\middle|\, \boldsymbol{y}, \Theta^{(\ell)}\right)\right]}_{E\left[f(X_n)\middle|\boldsymbol{y},\Theta^{(\ell)}\right]}$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{K} \ln\left(\pi_i^{(\ell)}\ f_i^{(\ell)}(y_n)\right)\ p\left(X_n = i \,\middle|\, \boldsymbol{y}, \Theta^{(\ell)}\right)$$

As $p\left(X_n = i \,\middle|\, \boldsymbol{y}, \Theta^{(\ell)}\right) = p\left(X_n = i \,\middle|\, y_n, \Theta^{(\ell)}\right)$

We get

$$p\left(X_n = i \,\middle|\, y_n, \Theta^{(\ell)}\right) = \frac{\pi_i^{(\ell)}\ f_i^{(\ell)}(y_n)}{\sum_{j=1}^{K} \pi_j^{(\ell)}\ f_j^{(\ell)}(y_n)} = \gamma_n^{(\ell)}(i)$$

<u>Gaussian mixture:</u> as an exercise, proof that EM-based re-estimation formulas for parameters of a MM can be written:

$$\pi_k^{(\ell+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_n^{(\ell)}(k)$$

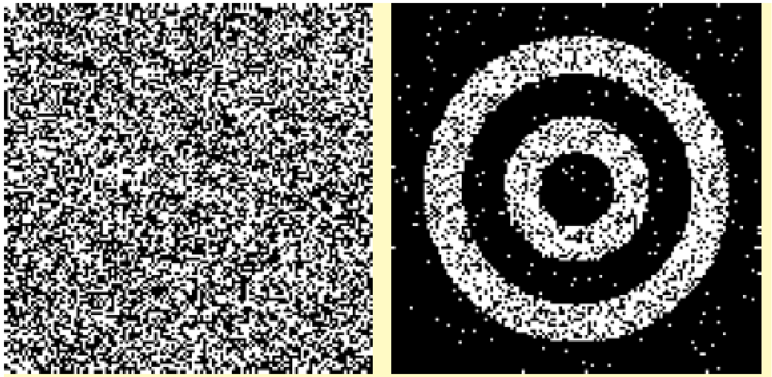
$$\sigma_k^{2,(\ell+1)} = \frac{\sum_{n=1}^{N} \gamma_n^{(\ell)}(k) \, \left(y_n - \mu_k^{(\ell+1)}\right)^2}{\sum_{n=1}^{N} \gamma_n^{(\ell)}(k)}$$

$$\mu_k^{(\ell+1)} = \frac{\sum_{n=1}^{N} \gamma_n^{(\ell)}(k) \; y_n}{\sum_{n=1}^{N} \gamma_n^{(\ell)}(k)}$$

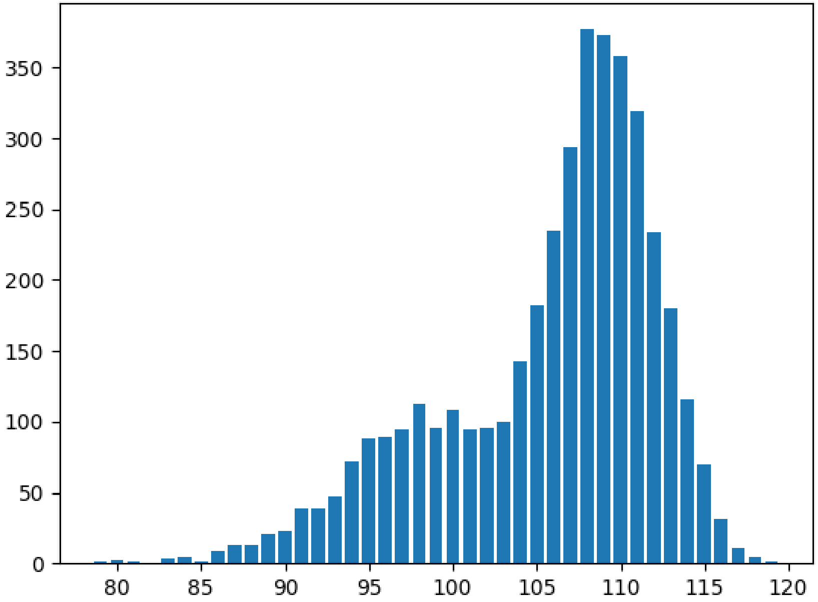
Image processing

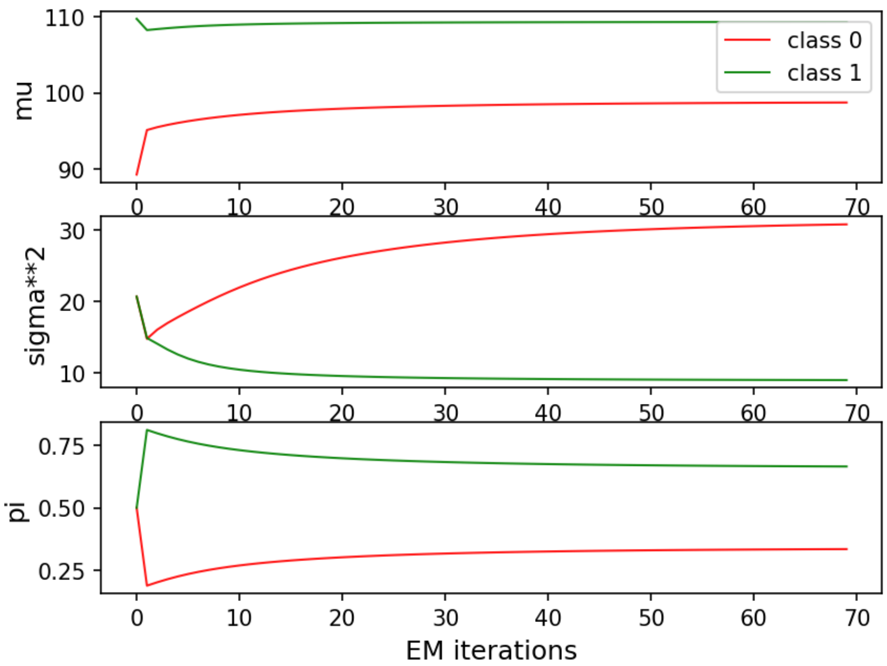
The EM algorithm looks for a local maxima of the likelihood: it requires the parameters to be initialized “not so far” from the true values.

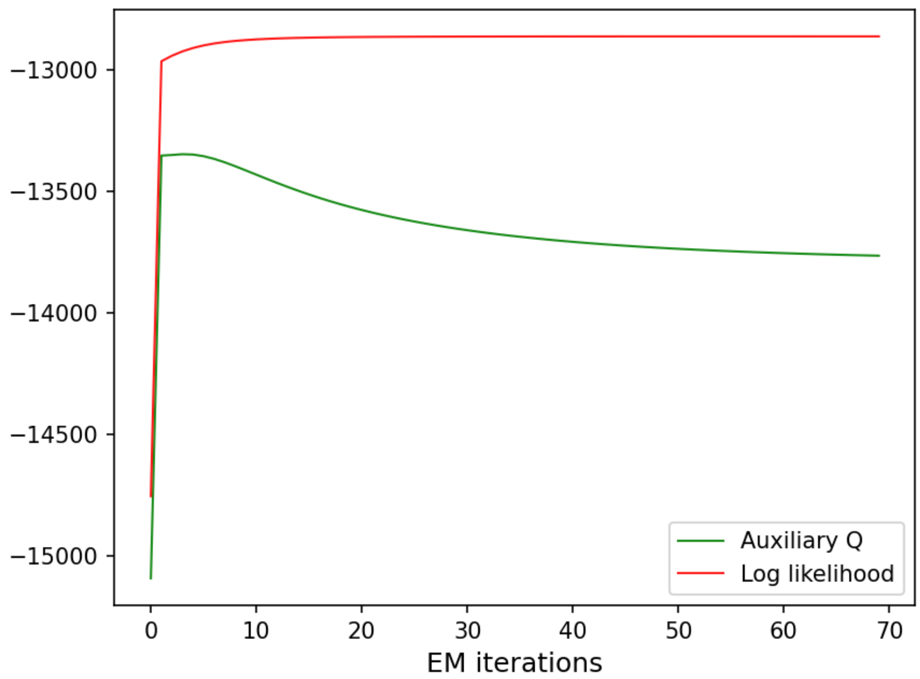
Any idea to initialize parameters ?

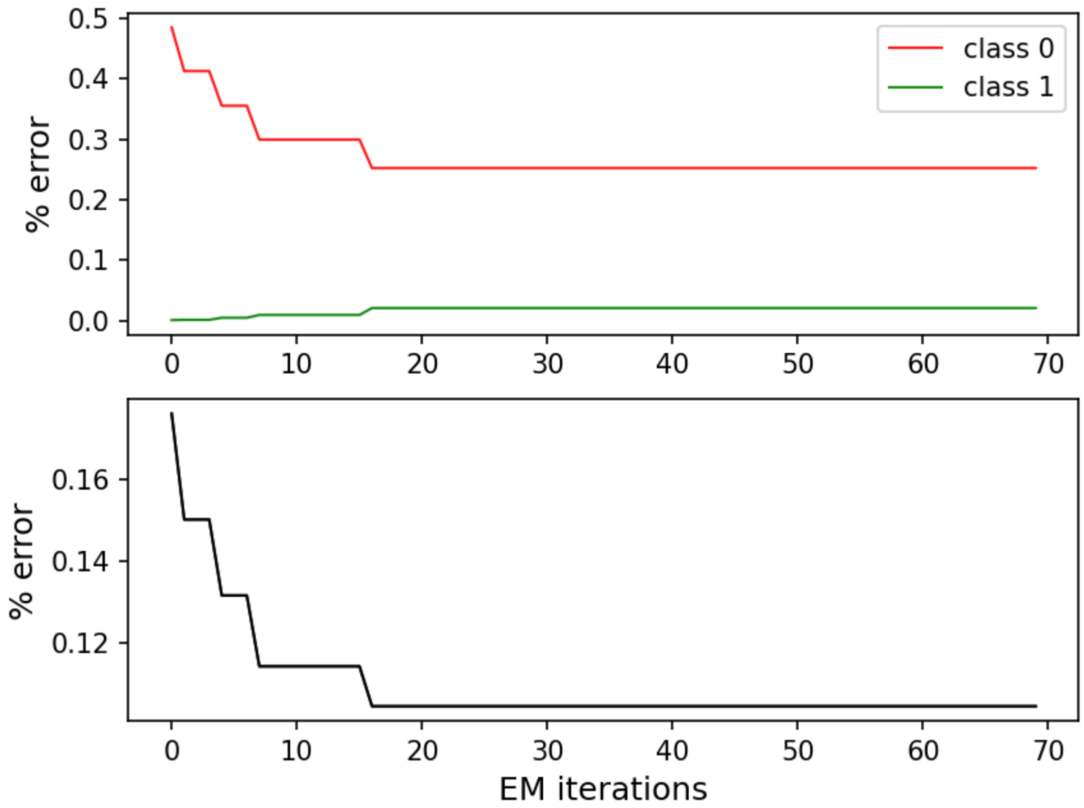


(a) Hazard (1) (b) K-means (2)

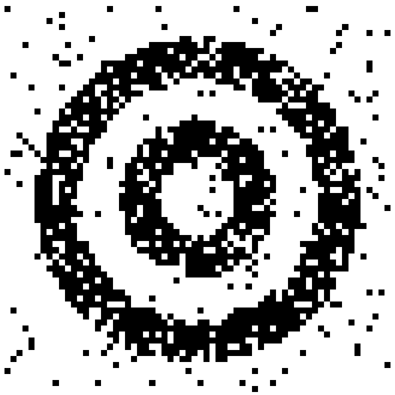



mu
110
100
90
class 0
class 1
sigma**2
30
20
10
pi
0.75
0.50
0.25
0
10
20
30
40
50
60
70
EM iterations

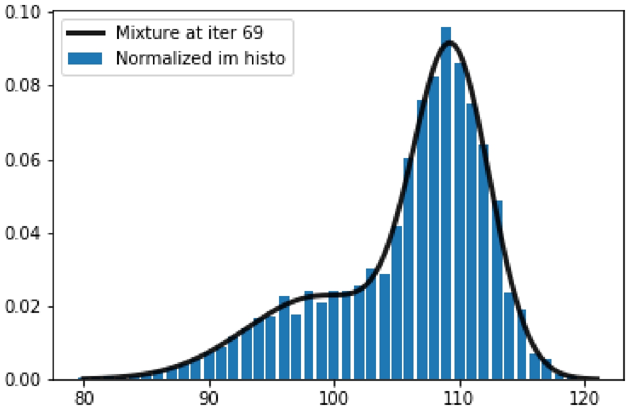

−13000
−13500
−14000
−14500
−15000
Auxiliary Q
Log likelihood
0
10
20
30
40
50
60
70
EM iterations



0.5
0.4
0.3
0.2
0.1
0.0
% error
class 0
class 1
0
10
20
30
40
50
60
70
0.16
0.14
0.12
% error
EM iterations



$$\xi_1 = 0.261$$
$$\xi_2 = 0.025$$
$$\xi = 0.11$$



$$\pi_1 = 0.334 \quad \mathcal{N}\left(\mu_1 = 98.85, \sigma_1 = 5.91\right)$$
$$\pi_2 = 0.666 \quad \mathcal{N}\left(\mu_2 = 109.39, \sigma_2 = 3.06\right)$$

MIXTURE MODEL

2D mixture models

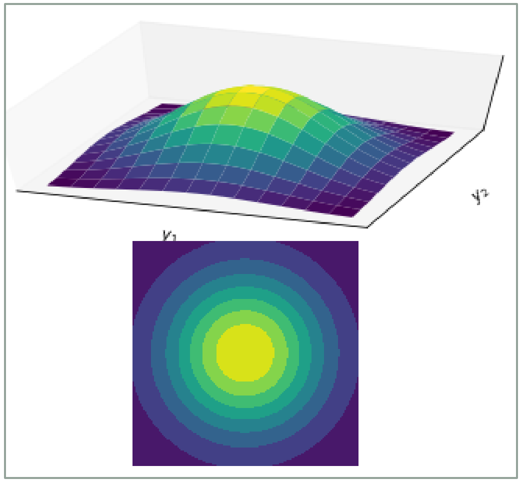
2D mixture model

$$f(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \, f_k(\boldsymbol{y}) \longrightarrow \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

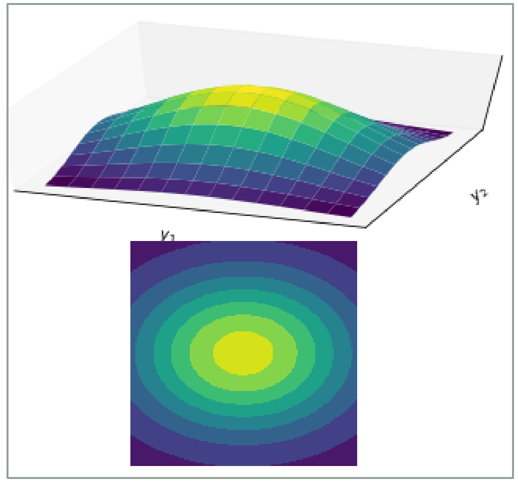
2D Gaussian

$$f(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \, e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}$$

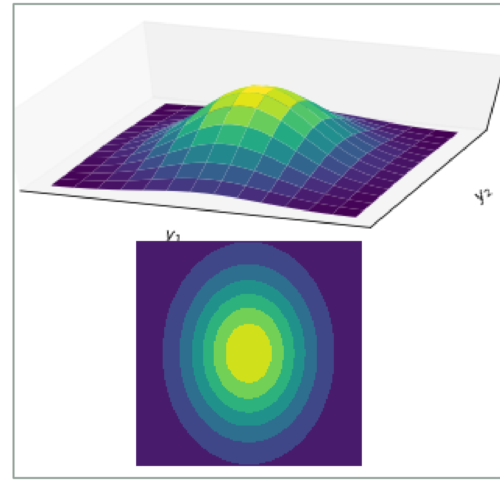
$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$



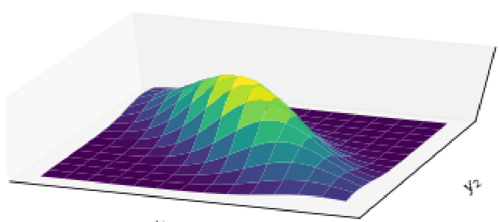
$$\Sigma = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$



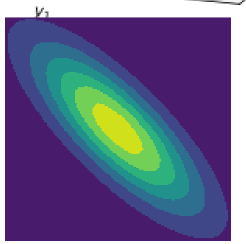
$$\Sigma = \begin{bmatrix} 0.6 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$

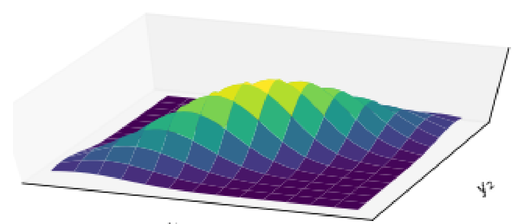


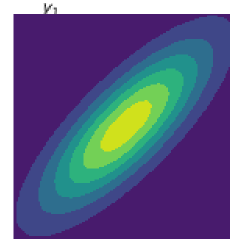
$$\Sigma = \begin{bmatrix} 2.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1.0 & -0.8 \\ -0.8 & 1.0 \end{bmatrix}$$



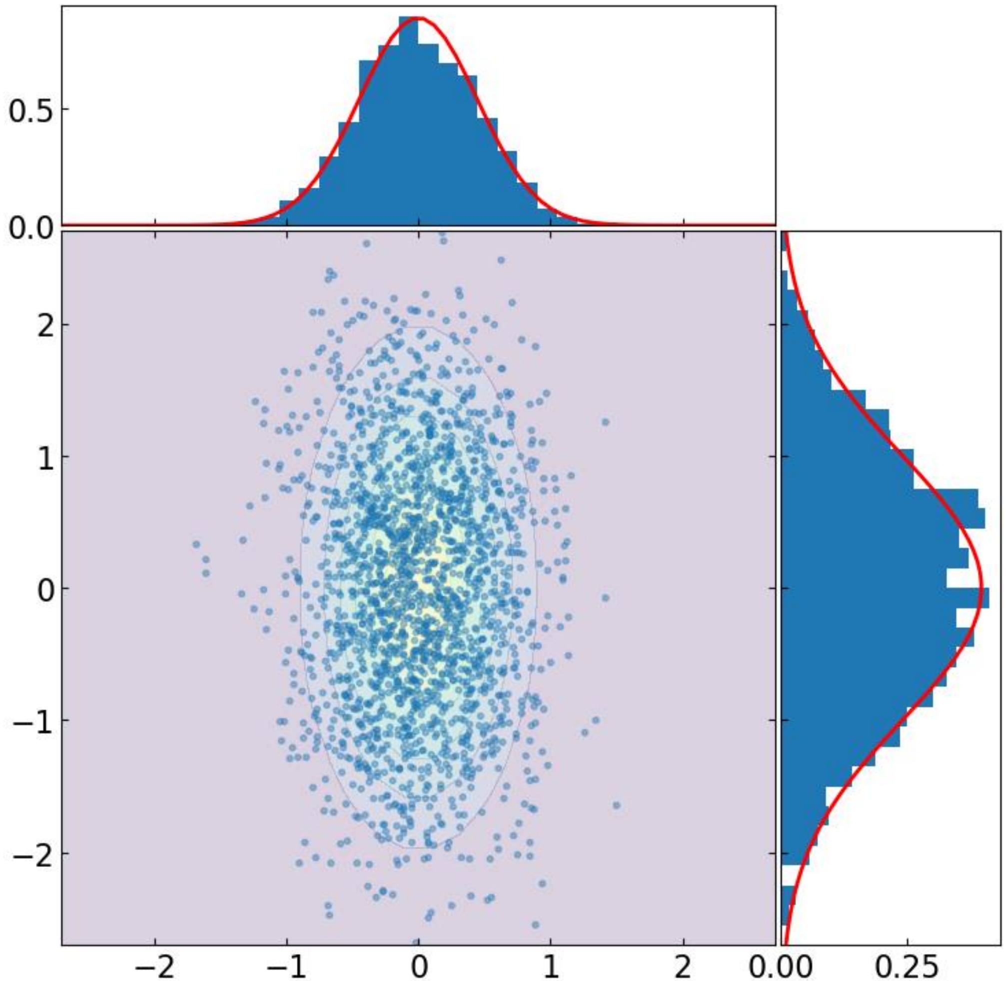




$$\Sigma = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix}$$

4. 2D Mixture Model

Margins, conditional laws, empirical estimation of parameters



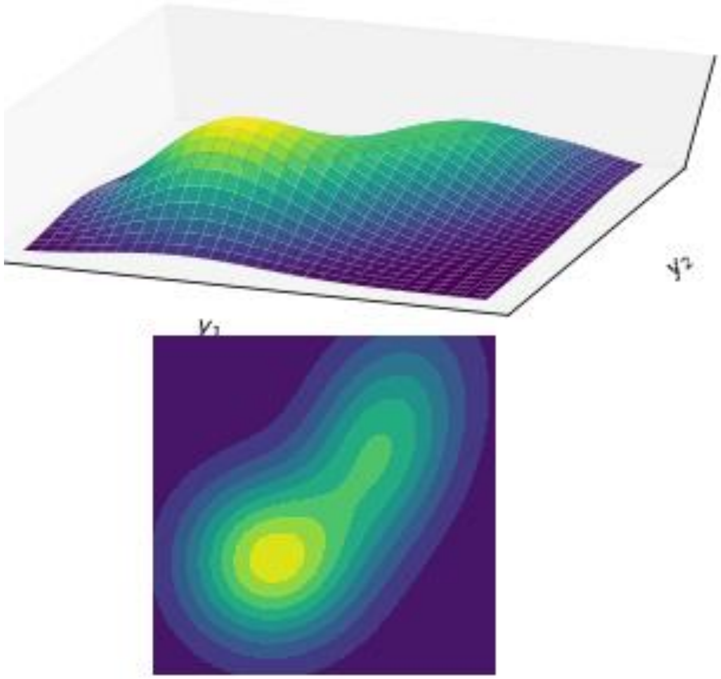
4. 2D Mixture Model

$$f(\boldsymbol{y}) = \sum_{k=1}^{2} \pi_k \, f_k(\boldsymbol{y})$$

$$\pi_1 = \pi_2 = \frac{1}{2}$$

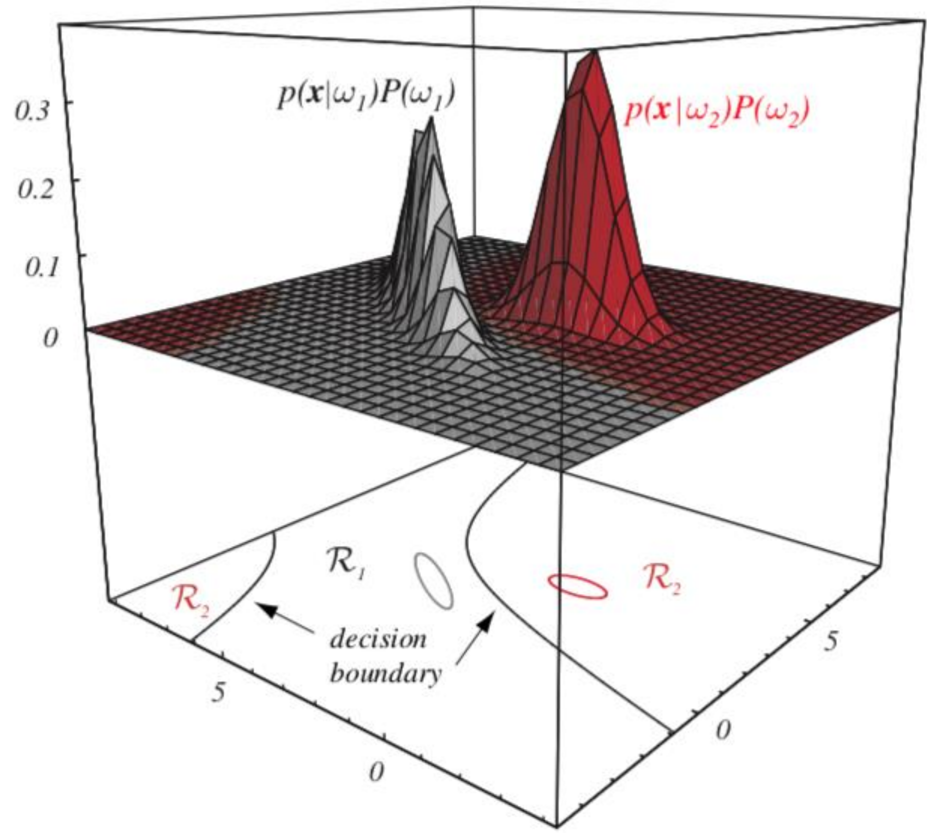
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$



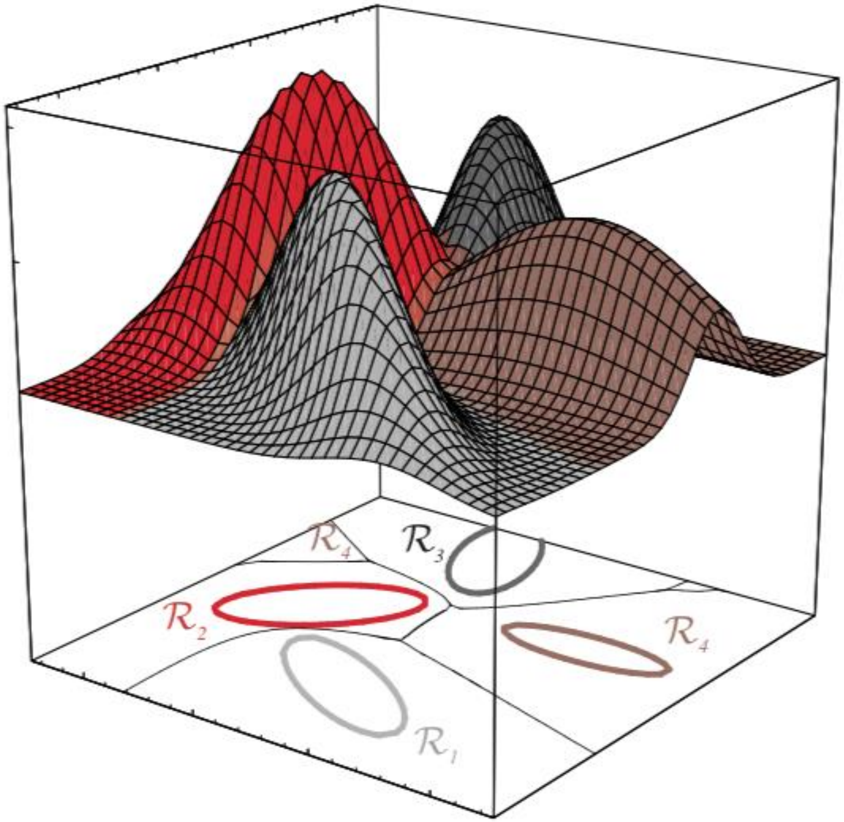
4. 2D Mixture Model

Decision boundary can be complex!!! (i.e. not always linear)



4. 2D Mixture Model

Multi-class decision boudaries



4. 2D Mixture Model

EM-based re-estimation formulas for parameters of a 2D MM are essentially the same:

$$\pi_k^{(\ell+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_n^{(\ell)}(k)$$

$$\boldsymbol{\Sigma}_k^{(\ell+1)} = \frac{\sum_{n=1}^{N} \gamma_n^{(\ell)}(k) \left(\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(\ell+1)}\right)^T \left(\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(\ell+1)}\right)}{\sum_{n=1}^{N} \gamma_n^{(\ell)}(k)}$$

$$\boldsymbol{\mu}_k^{(\ell+1)} = \frac{\sum_{n=1}^{N} \gamma_n^{(\ell)}(k)\ \boldsymbol{y}_n}{\sum_{n=1}^{N} \gamma_n^{(\ell)}(k)}$$

References

- *Theory and Use of the EM Algorithm* By Maya R. Gupta and Yihua Chen, Book pdf.
- *The EM algorithm and related statistical models* By Michiko Watanabe and Kazunori Yamaguchi, Book pdf.
- *Pattern classification* by Richard 0. Duda, Peter E. Hart and David G. Stork, 2015, Book pdf, Slides of the book.
- Finite Mixture model By Geoffrey McLachlan and D. Peel, Book pdf.
- Python library for MM : Pymix, sklearn.mixture.

Not seen!

- Variations about EM

  GEM, CEM, SEM -- On-line EM , by O. Cappé

- Mixture of non-gaussian type:

  M. of generalized hyperbolic distribution

  M. of skew-normal distribution, M. of t-distribution

- Choosing the number of clusters via model selection criteria

  BIC: Bayesian Information Criterion,

  AIK: Akaike Information Criterion,

  ICL: Integrated completed likelihood criterion

C. Biernacki, G. Celeux, and G. Govaert (2000). "*Assessing a mixture model for clustering with the integrated completed likelihood*". IEEE Trans. On PAMI, Vol 22(7), pp. 719–725.