Université Paris Saclay	**MSc AI & MSc Data Science.**
T3A : Machine Learning Algorithms	2024/2025

Gaussian Mixture Model (GMM) and Expectation-Maximization (EM) Algorithm

1 Generating Data from a Gaussian Mixture Model

- Generate a two-dimensional dataset from a K-component Gaussian mixture density with different means and covariance matrices :

  Data generating process : for each dataum :

  1. Generate the class label z_i according to a multinomial distribution with mixing proportions $(\pi_1, \ldots, \pi_K)$
  2. Given the class label z_i, generate a data vector x_i from the corresponding Gaussian component $\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$.
- Store the class labels for each generated data point for later comparison.

2 Expectation-Maximization (EM) Algorithm for GMM

- Implement the EM algorithm to estimate a K-component Gaussian mixture density.
- Initialize :
  - Mixing proportions equally.
  - Covariance matrices as identity matrices.
  - Means randomly (or using K-means clustering).
- In the EM training loop :
  - Store the value of the observed-data log-likelihood at each iteration.
  - At convergence, plot the log-likelihood curve and estimated density.
  - Plot the corresponding MAP partition (using scatter plots, gscatter, and density ellipses).

3 Model Selection

- Select the number of mixture components by computing a model selection criterion (BIC, AIC, AIC3, ICL, etc.).
- Vary K from 1 to 10, and compute the criterion for each EM run.
- Compare results with ground truth :
  - Number of mixture components.
  - Classification error rate for $K = 3$.

4 Application on other data

- take another data set used last week for K-meas : https://chamroukhi.com/data/K-means/X.txt
- Perform the same steps as above (GMM estimation, EM algorithm, and model selection).

5 Optimize your code by vectorizing as much as possible.