MVA – Probabilistic Graphical Models
Homework 2

Wilson JALLET*

December 10, 2019

1 K-means and the EM algorithm

Question 1. We consider a mixture model of K components for a dataset (X_i) where $X_i \in \mathbb{R}^d$. We denote $Z_i \in \{1, \dots, K\}$ the latent hidden label. Each component occurs with probability $p_k = \mathbb{P}(Z_i = k)$ and is distributed as

$$X_i \sim \mathcal{N}(\mu_k, D_k)$$

i.e. $p(x|k) = \frac{1}{((2\pi)^d |D_k|)^{1/2}} \exp(-\frac{1}{2}(x-\mu_k)^T D_k^{-1}(x-\mu_k))$.

The data log-likelihood under parameters $\Theta = ((p_1, \mu_1, D_1), \dots, (p_K, \mu_K, D_K))$ is

$$\mathcal{L}(X_1, \dots, X_n; \Theta) = \sum_{i=1}^n \log\left(\sum_{k=1}^K p_k p(X_i|k; \mu_k, D_k)\right)$$

We seek to compute the MLE

$$\widehat{\Theta} \in \operatorname*{argmax}_{\Theta} \mathcal{L}(X_1, \dots, X_n; \Theta)$$

This optimization problem is intractable when using straightforward methods.

The EM algorithm goes as follows:

- *Expectation* Compute the posterior probability of the latent variables Z_i:

$$q_{k,i}^{(t)} = p(Z_i = k | X_i; \Theta^{(t)}) = \frac{p_k^{(t)} p(X_i | Z_i = k; \Theta^{(t)})}{\sum_{\ell=1}^K p_\ell^{(t)} p(X_i | Z_i = \ell; \Theta^{(t)})} \tag{1}$$

  and denote $w_k^{(t)} = \sum_{i=1}^n q_{k,i}^{(t)}$ – we then have $\sum_{k=1}^K w_k^{(t)} = n$.

- *Maximization* Update the parameters $\Theta^{(t)}$ by maximizing the lower bound objective:

$$\max_{\Theta} \mathcal{J}(q^{(t)}, \Theta) = \sum_{i=1}^n \left(\sum_{k=1}^K q_{k,i}^{(t)} \left(\log p_k - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|D_k| - \frac{1}{2}(X_i - \mu_k)^T D_k^{-1}(X_i - \mu_k)\right)\right)$$

*wilson.jallet@polytechnique.org

subject to $\sum_{k=1}^{K} p_k = 1$ (associated to a multiplier ν). The KKT conditions give the null gradient condition:

$$\frac{1}{p_k} w_k^{(t)} - \nu = 0 \tag{2a}$$

$$-\sum_{i=1}^{n} q_{k,i}^{(t)} D_k^{-1}(\mu_k - X_i) = 0 \tag{2b}$$

$$-\frac{1}{2}\sum_{i=1}^{n} q_{k,i}^{(t)} (D_k^{-1} - D_k^{-2} \operatorname{diag}((X_i - \mu_k))^2) = 0 \tag{2c}$$

Which leads to the updates:

$$p_k = \frac{1}{n} w_k^{(t)} \tag{3a}$$

$$\mu_k = \frac{1}{w_k^{(t)}} \sum_{i=1}^{n} q_{k,i}^{(t)} X_i \tag{3b}$$

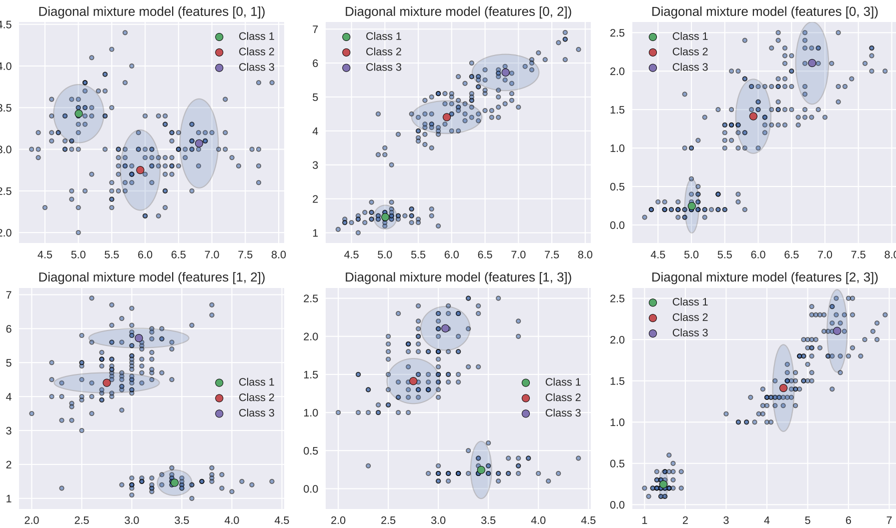
$$D_k = \frac{1}{w_k^{(t)}} \sum_{i=1}^{n} q_{k,i}^{(t)} \operatorname{diag}((X_i - \mu_k))^2 \tag{3c}$$

Question 2. The main advantage of this "reduced" covariance mixture model is that it is more sparse: it uses far fewer parameters ($K(2d+1)$) than the its full counterpart, which has $K(1 + d + d(d+1)/2)$ parameters. For datasets with relatively independent features (conditionally on the latent class), this can give performance very close to the full covariance while having a smaller, simpler model (meaning better AIC or BIC scores).

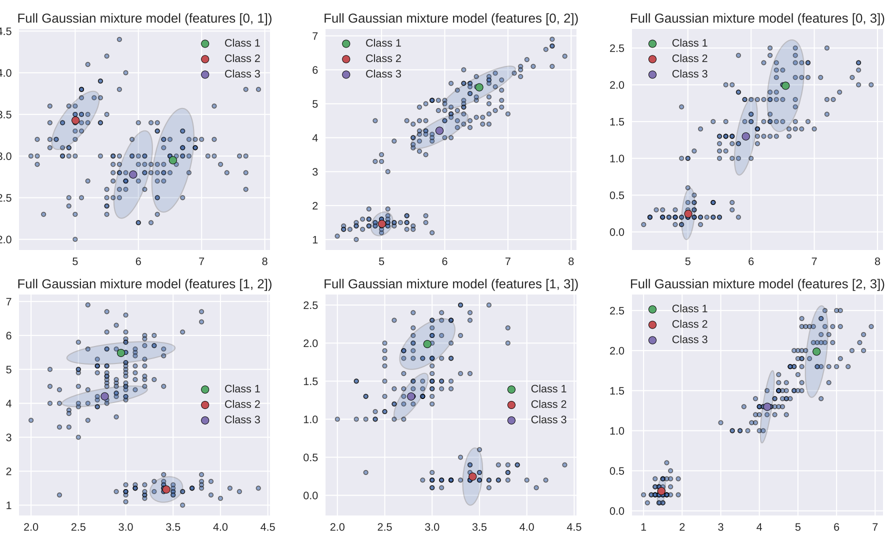
Question 3. Figure 1 compares the obtained latent class centroids and confidence ellipsoids (where applicable) for the diagonal and full covariance mixture models and K-means, on the Iris dataset, for a small number of classes $K = 3$ (the actual number of classes in the data). Figures 2 and 3 represent the same for $K = 2, 4$ classes.

Question 4.

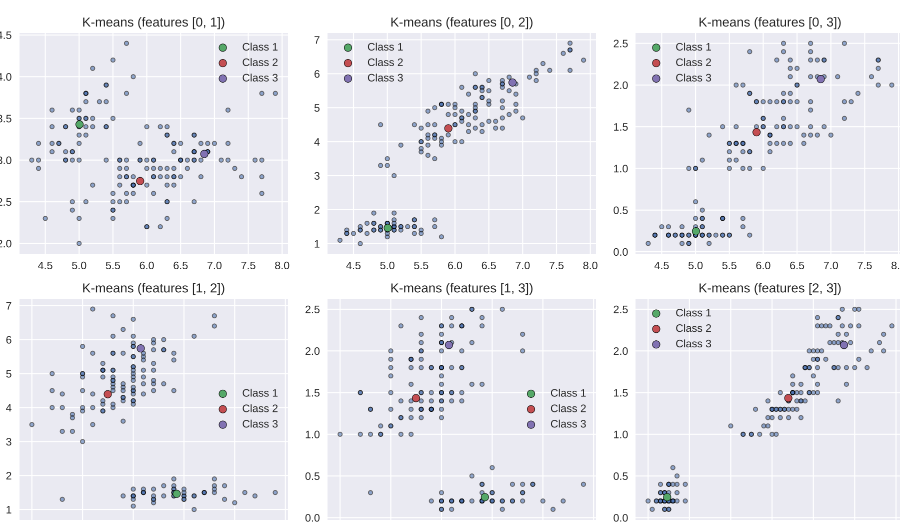
2 Graphs, algorithms and Ising



(a) Diagonal Gaussian mixture model on the Iris dataset (our implementation). $K = 3$ classes.

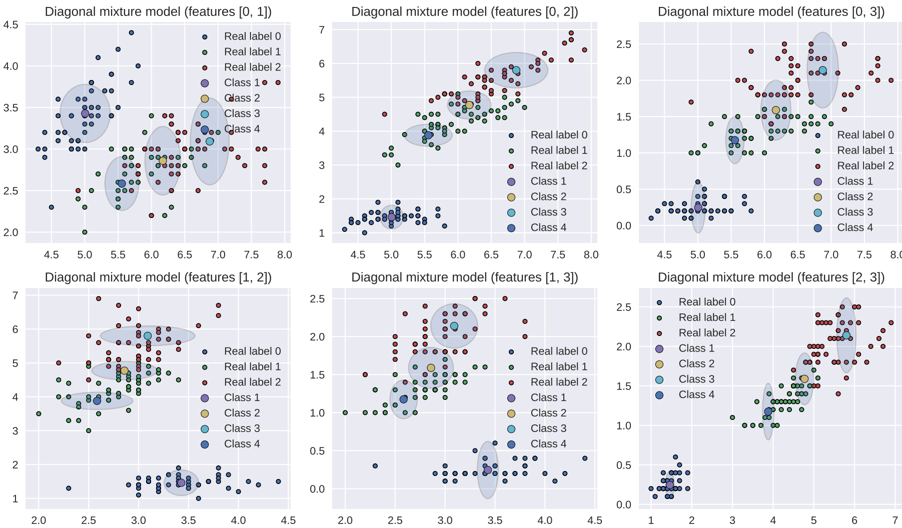


(b) Full Gaussian mixture on the Iris dataset using `scikit-learn`. $K = 3$ classes.

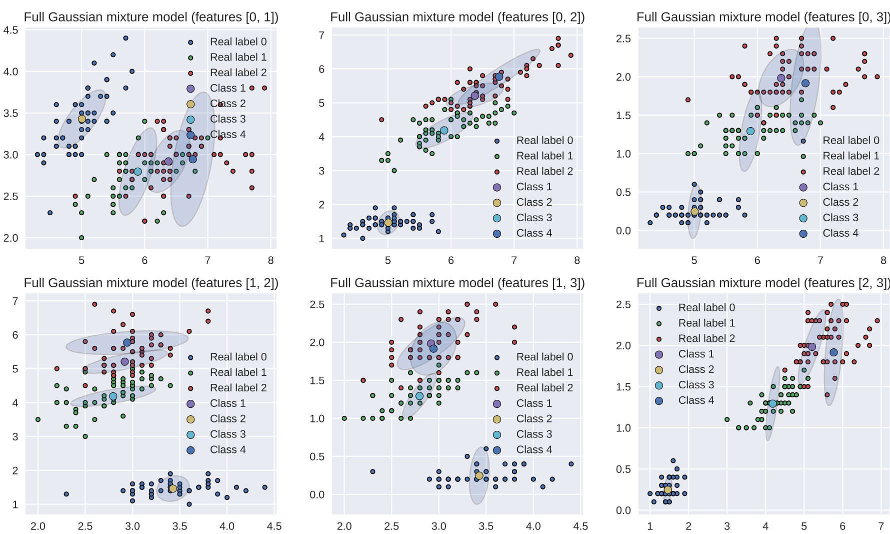


(c) Centroids of the K-means model.

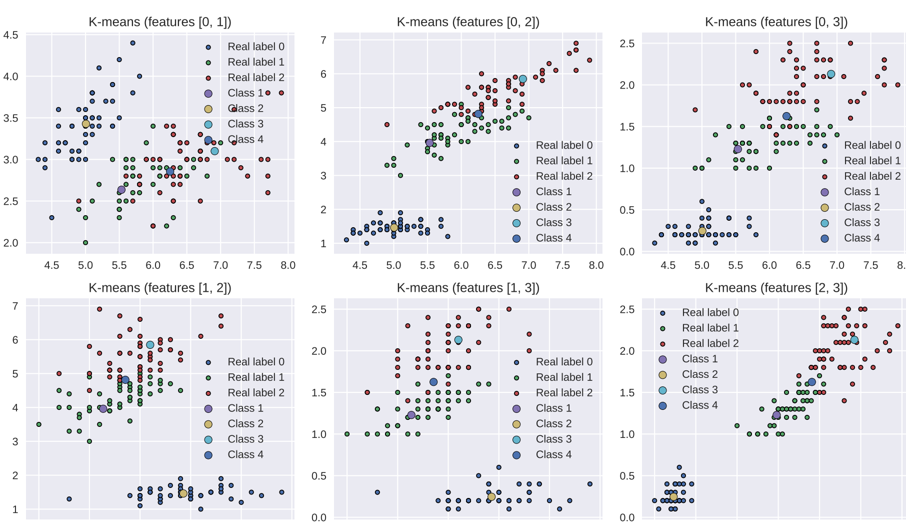
Figure 1: Comparison of the diagonal and full covariance mixture models and K-means for $K = 3$ classes.



(a) Diagonal Gaussian mixture model on the Iris dataset (our implementation).

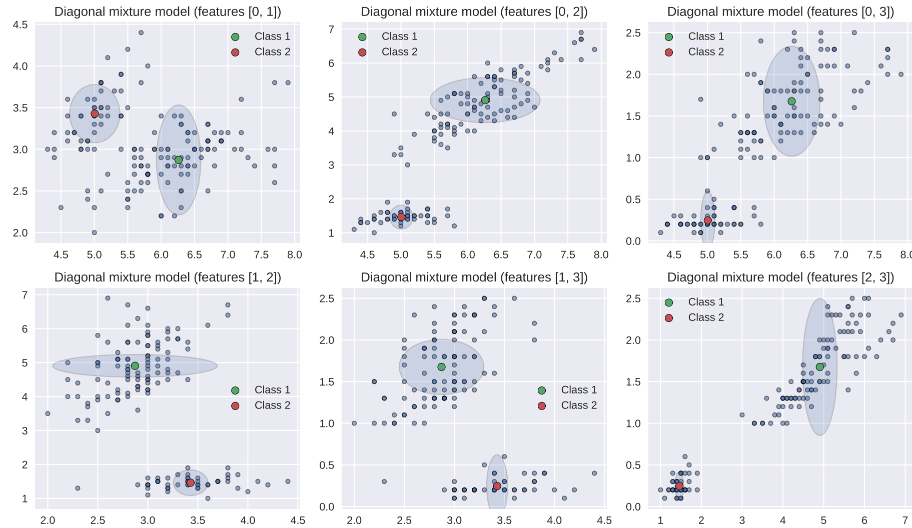


(b) Full Gaussian mixture on the Iris dataset using `scikit-learn`.

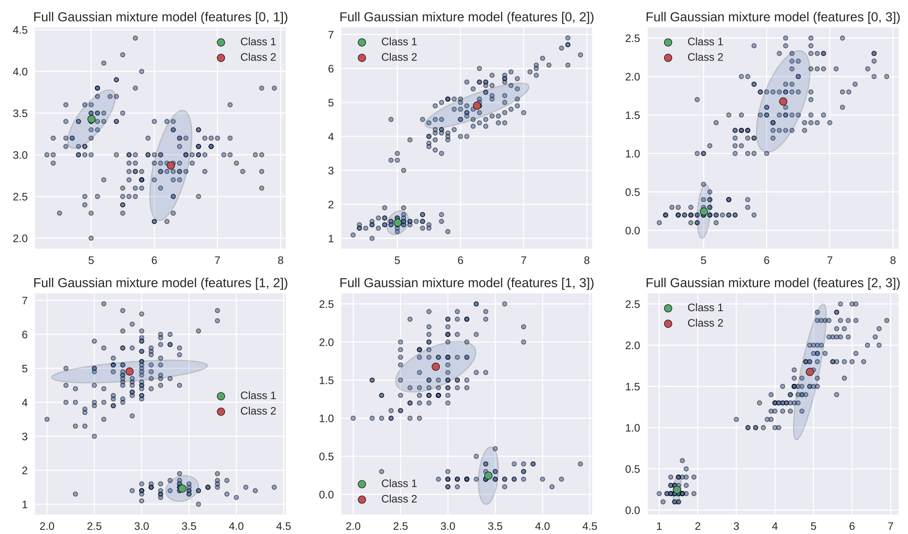


(c) K-means.

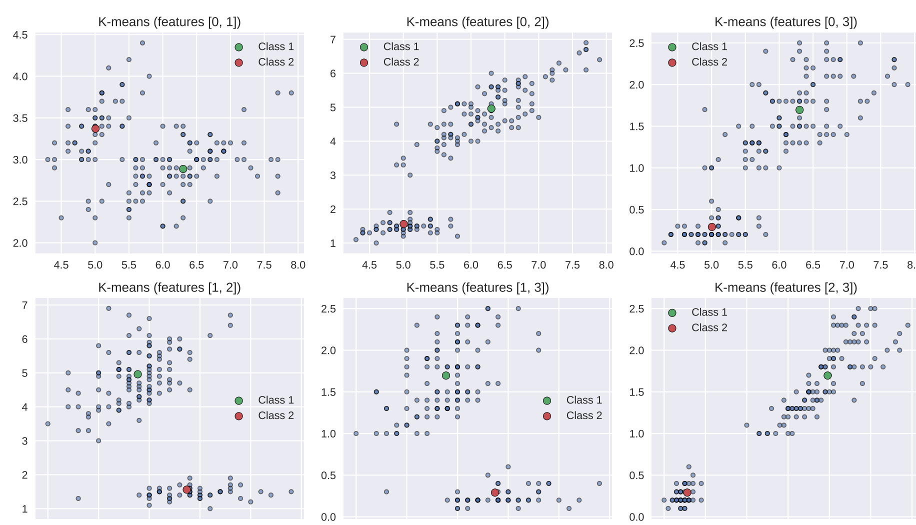
Figure 2: Comparison of the models for $K = 4$ classes.



(a) Diagonal Gaussian mixture model on the Iris dataset (our implementation).



(b) Full Gaussian mixture on the Iris dataset using `scikit-learn`.



(c) Full Gaussian mixture on the Iris dataset using `scikit-learn`.

Figure 3: Comparison of the models for $K = 2$ classes.