Mixture of balls with different volumes

The project will be emailed (to `christophe.ambroise@genopole.cnrs.fr`) as a `Rmd` file with possibility to build a pdf file. 'You will detail the calculations.

Context

Let us consider a vector of p random variables $x^1, \dots, x^p$ independent, normal, all with mean 0 and variance σ^2. The the random vector $\boldsymbol{x} = (x^1, \dots, x^p)'$ is normal with mean vector $(0, \cdots, 0)^t$, and covariance matrix $\sigma^2 I_p$. This distribution defines a gaussian ball with mean vector $(0, \cdots, 0)^t$, and covariance matrix $\sigma^2 I_p$.
Let us consider a mixture of K gaussian balls

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\mu_1}, \cdots, \boldsymbol{\mu_K}, \sigma_1, \cdots, \sigma_K) = \sum_{k=1}^{K} \pi_k \mathcal{N}_p(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k} = \sigma_k^2 I_p),$$

where $\boldsymbol{\pi} = \{\pi_k\}$ are the proportions of the mixture.
In the following, we will consider

- a sample $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ from the above ddefined mixture,
- latent variables $Z = \{z_1, \cdots, z_n\}$ indicating from which component of the mixture each $\boldsymbol{x}_i$ originates.
- the vector of parameters is denoted

$$\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu_1}, \cdots, \boldsymbol{\mu_K}, \sigma_1, \cdots, \sigma_K\}$$

Problem

Exercise 1 Simulation

1. Simulate sample of 1000 vectors from a 2 dimensional mixture with 2 components
   - mixed in proportion $\pi_1 = \pi_2 = \frac{1}{2}$.
   - with mean vectors $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (1, 2)^t$.
   - with covariance matrices $\boldsymbol{\Sigma}_1 = I$ and $\boldsymbol{\Sigma}_2 = 4I$.
2. Display the sample.
3. Display the coutour plot of the two dimensional density.

Exercise 2 Mclust versus kmeans

1. Run `Mclust` on the simulated data from the first exercise and comment the result.
2. Estimate the parameters of the simulated data from the first exercise using `mclust`.
3. Find a partition of the simulated data into two classes using `mclust`.
4. Find a partition of the simulated data into two classes using `kmeans`.
5. Compare the two partitions (from `kmeans` and `mclust`). Comment your result.

Exercise 3 EM algorithm for a Mixture of balls

1. Detail the computation of the $t_{ik}^q = \mathbb{E}[Z_{ik}]$ with respect to $p_{\theta^q}(Z|X)$ where $Z_{ik} = \mathbb{I}_{(Z_i=k)}$.
2. Express $Q(\boldsymbol{\theta}^q|\boldsymbol{\theta})$ the expectation of the complete log-likelihood with respect to $p_{\theta^q}(Z|X)$.
3. Detail the computation of $\boldsymbol{\theta}^{q+1} = argmax_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}^q|\boldsymbol{\theta})$.
4. Write the pseudo-code of an EM algorithm for estimating $\boldsymbol{\theta}$.
5. Write a `E-step` function that produces the t_{ik} from $\boldsymbol{\theta}$. Check the results by injecting the real parameters of your simulation and comparing the t_{ik} estimated against the latent variables Z in your simulation.
6. Write a `M-step` function that produces $\boldsymbol{\theta}^{q+1}$ from the sample and the t_{ik}^qs.
7. Program the EM algorithm (you could check that each step increases the log-likelihood.)

Exercise 4 Data `iris`

1. Run your algorithm with the `iris` dataset and compare the results to the one obtained using the `kmeans` algorithm.
2. Comment.