2. **(K-means vs. EM/GMM, cazul uni-variat (π, μ, σ))**

Redăm mai jos pseudo-codul algoritmului EM — de fapt, doar partea sa iterativă, conţinând cei doi paşi, E şi M — pentru clusterizare prin modelare de mixturi de gaussiene (GMM), cazul uni-variat, cu toţi parametrii lăsaţi liberi. Îţi cerem să faci schimbările minimale(!) necesare pentru a-l transforma într-un pseudo-cod corespunzător buclei principale a algoritmului de clusterizare K-means. Rescrie paşii care trebuie modificaţi, folosind spaţiul lăsat disponibil sub fiecare pas. Dacă un pas nu necesită schimbări, scrie *„Nicio modificare“* în spaţiul disponibil sub pasul respectiv. Dacă un anumit pas nu este necesar, scrie *„Elimină acest pas“* în spaţiul disponibil sub pasul respectiv.

***Pasul E*:**
Calculează probabilitatea $p_{ij}^{(t)}$ de asignare a instanţei x_i la clusterul [corespunzător gaussienei] j, ţinând cont de valorile actuale ale parametrilor $\pi^{(t)}$, $\mu^{(t)}$, $\sigma^{(t)}$, unde t identifică iteraţia curentă a algoritmului EM.

$p_{ij}^{(t)} = P(z_i = j | x_i, \pi^{(t)}, \mu^{(t)}, \sigma^{(t)})$ **pentru** $i \in \{1, \dots, n\}$ **şi** $j \in \{1, \dots, K\}$

***Pasul M*:**
A. Re-calculează probabilităţile a priori pentru fiecare cluster / componentă a mixturii:

$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)}}{n}$ **pentru** $j \in \{1, \dots, K\}$

B. Re-calculează [centroizii clusterelor, reprezentaţi de] mediile distribuţiilor gaussiene care „modelează“ clusterele:

$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)} x_i}{\sum_{i=1}^n p_{ij}^{(t)}}$ **pentru** $j \in \{1, \dots, K\}$

C. Re-calculează varianţele distribuţiilor gaussiene care „modelează“ clusterele:

$(\sigma_j^2)^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^n p_{ij}^{(t)}}$ **pentru** $j \in \{1, \dots, K\}$

3. **(EM/GMM, cazul uni-variat: aplicarea manuală a unei iteraţii, pentru o mixtură de tipul $\mu_1 = \mu_2$ (liber), $\pi_1 = \pi_2, \sigma_1 = 1, \sigma_2 = 2$)**

Considerăm o mixtură de două gaussiene uni-variate, pentru care funcţia de densitate de probabilitate (p.d.f.) are pentru o „observaţie“ oarecare x expresia

$$\frac{1}{2}\mathcal{N}(x|\mu, 1) + \frac{1}{2}\mathcal{N}(x|\mu, 2^2).$$

În această formulă, $\mathcal{N}(x|\mu, \sigma^2)$ desemnează, pentru x, valoarea densităţii distribuţiei normale uni-variate de medie μ şi varianţă σ^2. Remarcaţi faptul că în acest model de mixtură probabilităţile de mixare / selecţie sunt egale, apoi că mediile celor două componente sunt egale şi, de asemenea, că deviaţiile standard ale celor două componente au valorile fixate 1 şi respectiv 2. Modelul acesta de mixtură are un singur parametru, μ.

Presupunem că dorim să estimăm valoarea parametrului μ prin metoda maximizării verosimilităţii, folosind algoritmul EM. Răspundeţi la următoarele întrebări privitoare la modul în care operează paşii E şi M ai acestui algoritm, atunci când considerăm cele trei instanţe / „observaţii“ de mai jos:

4.0, 4.6, 2.0.

Vă punem la dispoziţie un tabel cu anumite valori ale funcţiei de densitate pentru distribuţia normală standard, de care veţi avea probabil nevoie în rezolvarea acestui exerciţiu:

x	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\mathcal{N}(x\|0,1)$	.40	.40	.39	.38	.37	.35	.33	.31	.29	.27	.24
x		1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\mathcal{N}(x\|0,1)$		.22	.19	.17	.15	.13	.11	.09	.08	.07	.05

a. Găsiţi valoarea probabilităţilor condiţionale calculate la pasul E, presupunând că la execuţia precedentului pas M estimarea obţinută pentru parametrul modelului a fost $\mu = 4$ (iar $\sigma_1 = 1$ şi $\sigma_2 = 2$, totdeauna). Remarcaţi faptul că, întrucât probabilităţile condiţionale pentru cele două componente ale mixturii trebuie să se sumeze la valoarea 1, este suficient să se calculeze $p_{i1} \stackrel{not.}{=} P(componenta_1 | x_i)$ pentru $i = 1, 2, 3$.

b. Folosind probabilităţile pe care le-aţi calculat la punctul a, găsiţi estimarea pentru parametrul μ care va fi obţinută la următoarea execuţie a pasului M. Vă reamintim că pasul M maximizează media (engl., expected value) log-probabilităţii corelate pentru instanţele x_1, x_2 şi x_3 şi variabilele-indicator [pentru componentele mixturii], care sunt latente. Media aceasta se calculează în raport cu distribuţia de probabilitate condiţională a variabilelor-indicator [pentru componentele mixturii], care a fost determinată / calculată la precedentul pas E.

***Sugestie*: Nu este necesar ca în prealabil să faceţi elaborarea formulelor algoritmului EM pentru acest caz specific de mixtură de distribuţii gaussiene uni-variate. Este suficient să lucraţi cu funcţia „auxiliară“ Q corespunzătoare iteraţiei respective. Sau, mult mai simplu, puteţi folosi formula de actualizare a mediilor de la problema precedentă (fiindcă forma ei se menţine şi pentru acest caz).**