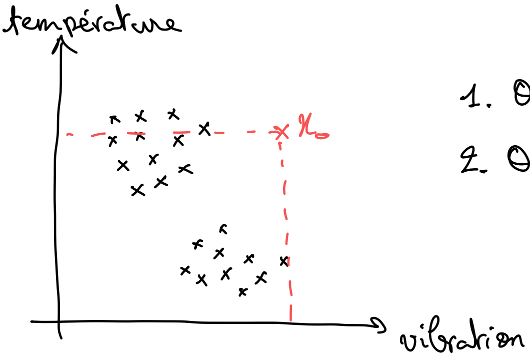
III. Algorithme EM appliqué aux mélanges gaussiens

Exemple d'application de l'estimation de modèle



1. On estime un modèle $p(x)$
2. On met en place une règle de décision

   si $p(x_0) < \varepsilon \to$ anomalie

Exemple introductif "simple" :

Estimation par MV d'une gaussienne simple $(K=1)$

$X = (x_1, \dots, x_N)^T$ iid $x_n \sim \mathcal{N}(\mu^*, \Sigma^*)$ $\forall n \in [\![1, N]\!]$

On cherche $\theta_{ML} = (\mu_{ML}, \Sigma_{ML})$ qui maximisent la log-vraisemblance $\ell(\theta) = \ln(p(X|\theta)) = \ln\left(\prod_{n=1}^{N} p(x_n|\theta)\right)$

$$= \sum_{n=1}^{N} \ln(p(x_n|\theta)).$$

En utilisant la pdf d'une gaussienne multivariée $(D>1)$:

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

On obtient :

$$\ell(\underbrace{\mu,\Sigma}_{=\theta}) = \frac{-ND}{2}\ln(2\pi) - \frac{N}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$$

Lemmes : $\det(A)$

(A) $$\frac{\partial \ln(|A|)}{\partial A_{ij}} = (A^{-1})_{ji}$$

B $\dfrac{\partial A^{-1}}{\partial A_{ij}} = -A^{-1}E_{ij}A^{-1} \quad (E_{ij} = \delta_{ij})$ $\begin{array}{cc} & \begin{matrix} & & & j \end{matrix} \\ i & \begin{bmatrix} (0) & \cdots & 1 & (0) \\ & (0) & \vdots & (0) \end{bmatrix} \end{array}$

C $\mathrm{Tr}(AE_{ij}) = A_{ji}$

D Si $f : x \mapsto \langle x | Ax \rangle$, $\dfrac{\partial f}{\partial x} = (A + A^T)x$.

Dérivation de $\ell(\mu, \Sigma)$:

1. Estimation de μ_{ML} :

$$\frac{\partial \ell}{\partial \mu} \overset{D}{=} -\frac{1}{2}\sum_{n=1}^{N}(-2)\Sigma^{-1}(x_n - \mu)$$

$$= \Sigma^{-1}\left(\sum_{n=1}^{N}(x_n - \mu)\right)$$

donc $\dfrac{\partial \ell}{\partial \mu} = 0 \iff \Sigma^{-1}\left(\sum_{n=1}^{N}(x_n - \mu)\right) = 0$

$$\overset{\Sigma^{-1}\text{ inv.}}{\iff} \left(\sum_{n=1}^{N} x_n\right) - N\mu = 0$$

$$\iff \mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Donc $\boxed{\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n}$

2. Estimation de Σ_{ML}

$$\frac{\partial \ell}{\partial \Sigma_{ij}} \overset{A,B}{=} -\frac{N}{2}(\Sigma^{-1})_{ji} + \frac{1}{2}\sum_{n=1}^{N}\underbrace{\overbrace{(x_n-\mu)^T\Sigma^{-1}}^{(1,D)}E_{ij}\overbrace{\Sigma^{-1}(x_n-\mu)}^{(D,1)}}_{:= S_n \quad (1,1)}$$

On a $S_n \in \mathbb{R}$ donc $S_n = \mathrm{Tr}(S_n)$,

ie $(x_n-\mu)^T\Sigma^{-1}E_{ij}\Sigma^{-1}(x_n-\mu) = \mathrm{Tr}\left((x_n-\mu)^T\Sigma^{-1}E_{ij}\Sigma^{-1}(x_n-\mu)\right)$

$= \mathrm{Tr}\left(\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}E_{ij}\right)$

$$\overset{\text{ⓒ}}{=} \left(\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}\right)_{ji}$$

Donc $$\frac{\partial \ell}{\partial \Sigma_{ij}} = \left[-\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\sum_{n=1}^{N}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}\right]_{ji}$$

Enfin, $$\left[\frac{\partial \ell}{\partial \Sigma_{ij}} = 0 \quad \forall i,j\right] \iff -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\sum_{n=1}^{N}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1} = 0$$

multiplier à g. et à d. par Σ.

$$\iff -\frac{N}{2}\Sigma + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)(x_n-\mu)^T = 0$$

$$\iff \Sigma = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)(x_n-\mu)^T$$

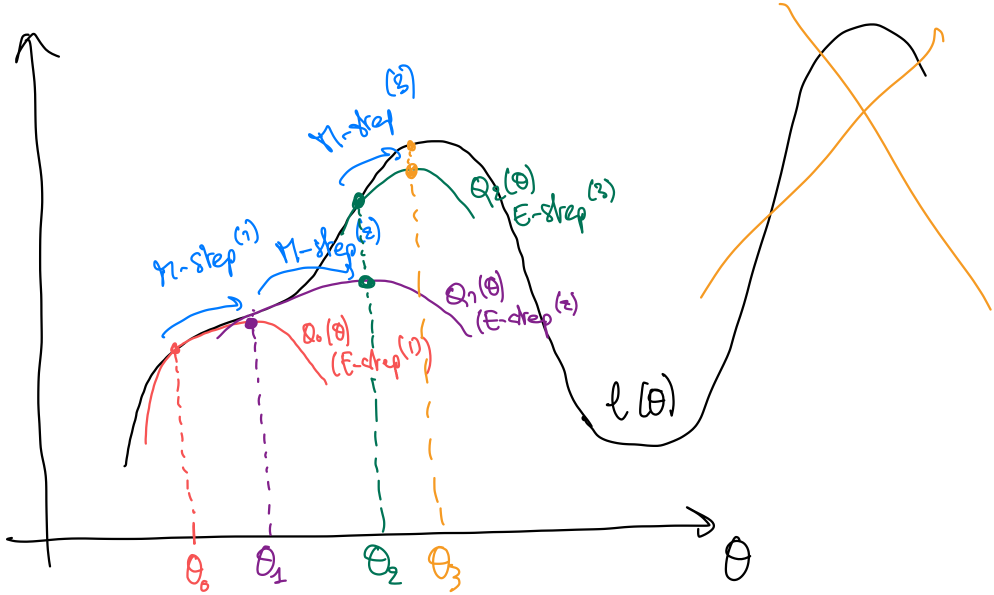
Donc $$\boxed{\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})(x_n-\mu_{ML})^T}$$

Estimation pour un mélange de K gaussiennes

<u>Contexte</u>: On suppose que des observations $x_1, \dots, x_N \in \mathbb{R}^D$ sont issues d'un mélange gaussien de paramètres $\theta^* = (\pi^*, \mu^*, \Sigma^*)$.

On cherche θ qui maximise la log-vraisemblance

$$\ell(\theta) = \ln\left(\prod_{n=1}^{N} p(x_n|\theta)\right)$$



Pseudo-code :

Initialisation : initialiser $\Theta = (\pi, \mu, \Sigma)$

Répéter :

- **E-step** : calculer les postérieurs $\gamma_{nk} = p(z_n = k \mid x = x_n)$ avec les valeurs actuelles de Θ pour en déduire la fonction $Q(\Theta)$.
- **M-step** : mise à jour des paramètres $\Theta = (\pi, \mu, \Sigma)$ pour maximiser $Q(\Theta)$.

Critère d'arrêt : par exemple quand les paramètres ont "convergé".

On tente un maximum vraisemblance pour un mélange de K gaussiennes

$X = (x_1, \dots, x_N)^T$: observations ($x_n \in \mathbb{R}^D \ \forall n \in [\![1, N]\!]$)

$Z = (z_1, \dots, z_N)$: variables ($z_n \in [\![1, K]\!] \ \forall n \in [\![1, N]\!]$)

$$\theta = \begin{cases} \pi = (\pi_1, \dots, \pi_K) \text{ coefficients de mixage } \pi_k = p(z=k) \in]0,1[\text{ et } \sum_{k=1}^{K} \pi_k = 1 \\ \mu = (\mu_1, \dots, \mu_K) \text{ moyennes des K gaussiennes } (\mu_k \in \mathbb{R}^D) \\ \Sigma = (\Sigma_1, \dots, \Sigma_K) \text{ matrices de covariance des K gaussiennes } (\Sigma_k \in \mathbb{R}^{D \times D}) \end{cases}$$

On cherche les valeurs de π, μ, Σ qui maximisent :

$$\ell(\pi, \mu, \Sigma) = \ln\left(\prod_{n=1}^{N} p(x_n \mid \pi, \mu, \Sigma)\right)$$

$$= \sum_{n=1}^{N} \ln\left(p(x_n \mid \pi, \mu, \Sigma)\right)$$

formule des proba totales

$$= \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} p(x = x_n, z_n = k \mid \pi, \mu, \Sigma)\right)$$

$$= \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} p(z_n = k)\, p(x = x_n \mid z_n = k, \pi, \mu, \Sigma)\right)$$

$$= \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \underbrace{\mathcal{N}(x_n \mid \mu_k, \Sigma_k)}_{= \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right)}\right)$$

<u>Problème</u> : On ne peut pas différencier par rapport aux μ_k et Σ_k (log de somme d'exponentielles).

<u>Idée</u> : Rendre la log-vraisemblance séparable.

<u>Rappel</u> : Si $f : (x,y) \mapsto g(x) + h(y)$

alors $\underset{(x,y)}{\operatorname{argmax}} f = \left(\underset{x}{\operatorname{argmax}}\, g \,,\, \underset{y}{\operatorname{argmax}}\, h \right)$

$$\ell(\theta) = \ln\left(p(X|\theta) \right)$$

On introduit la log-vraisemblance "complétée" :

$$\ell'(\theta) = \ln\left(p(X,Z|\theta) \right) = \sum_{n=1}^{N} \ln\left(p(x_n, z_n | \theta) \right)$$

$$= \sum_{n=1}^{N} \ln\left(p(z=z_n)\, p(x=x_n | z=z_n, \theta) \right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}_{z_n=k} \ln\Big(\underbrace{p(z=k)}_{=\pi_k}\, \underbrace{p(x=x_n | z=k, \theta)}_{=\mathcal{N}(x_n | \mu_k, \Sigma_k)} \Big)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}_{z_n=k} \ln\Big(\pi_k \underbrace{\mathcal{N}(x_n | \mu_k, \Sigma_k)}_{\frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp(\dots)} \Big)$$

Pour maximiser cette quantité $\ell'(\theta)$, on va maximiser son espérance (heuristique ¿) :

$$\mathbb{E}\left(\ell'(\theta) \right) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}\left[\mathbb{1}_{z_n=k} \ln\left(\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right) \right]$$

<u>Lemme</u> : Si X, Y 2 v.a.r tq (X,Y) admet une espérance sur $\mathbb{R}^2$, alors $\mathbb{E}_X(X) = \mathbb{E}_Y\left(\mathbb{E}_{X|Y}(X|Y) \right)$.

<u>Preuve</u> : $\mathbb{E}_Y\left(\mathbb{E}_X (X|Y) \right) = \int_y \mathbb{E}_X(X | Y=y)\, f_Y(y)\, dy$

$$= \int_y \left[\int_x x \, f_{X|Y}(x) \, dx \right] f_Y(y) \, dy$$

Fubini (cf. ci-dessous)

$$= \int_x x \left[\int_y \underbrace{f_{X|Y}(x) f_Y(y)}_{= f_{(X,Y)}(x,y)} \, dy \right] dx$$

$$\left[\text{“} \mathbb{P}(X|Y) \mathbb{P}(Y) = \mathbb{P}(X,Y) \text{”} \right]$$

$$= \int_x x \, f_X(x) \, dx \qquad \mathbb{E}_{X,Y}(X,Y) = \iint_{(x,y)} f_{X,Y}(x,y) \binom{x}{y} dx\,dy$$

$$= \mathbb{E}_X(X)$$

<u>Fubini</u> : $\left| x \underbrace{f_{X|Y}(x) f_Y(y)}_{= f_{(X,Y)}(x,y)} \right| \leq \underbrace{\left\| \binom{x}{y} \right\| \times \left| f_{(X,Y)}(x,y) \right|}_{\text{intégrable car } (X,Y) \text{ admet une espérance}}$

$\boxed{\text{CQFD}}$

On a donc

$$\mathbb{E}_{(X,Z)}(\ell'(\theta)) = \mathbb{E}_X\left(\mathbb{E}_{(X,Z)}(\ell'(\theta) \mid X) \right)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{x_n}\left[\mathbb{E}\left[\mathbb{1}_{z_n=k} \left(\ln(\pi_k) + \ln(\mathcal{N}(x_n \mid \mu_k, \Sigma_k)) \right) \right] \mid x = x_n \right]$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{x_n}\left[\underbrace{\mathbb{E}\left(\mathbb{1}_{z_n=k} \mid x = x_n \right)}_{= p(z_n = k \mid x = x_n) \; = \gamma_{nk}} \left[\ln(\pi_k) + \ln(\mathcal{N}(x_n \mid \mu_k, \Sigma_k)) \right] \right]$$

$$= \mathbb{E}_X \Big[\underbrace{\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left[\ln(\pi_k) + \ln(\mathcal{N}(x_n | \mu_k, \Sigma_k)) \right] \Big]}_{:= Q(\theta)}$$

Q est bien séparable à γ_{nk} fixés : en fait on va considérer que les postérieurs γ_{nk} (qui ont été calculés avec les valeurs actuelles des paramètres) sont fixés.

Retour sur les idées à l'origine de l'expression de Q :

Quand on a une seule gaussienne, on maximise $\ell(\theta) = \ln(p(X|\theta))$. Pour un mélange à K gaussiennes,

$$\ell(\theta) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right)$$

↳ pas tractable

Idée : utiliser la log-vraisemblance "complétée" :

$$p(X, Z | \theta) = p(X|\theta)\, p(Z | X, \theta)$$

ie
$$\underbrace{\ln(p(X, Z | \theta))}_{\approx Q(\theta) \text{ moralement}} = \underbrace{\ln(p(X|\theta))}_{\ell(\theta)} + \underbrace{\ln(p(Z | X, \theta))}_{\leq 0}$$

Pour formaliser, on introduit la distribution $q(Z)$:

$$\ell(\theta) = \ln(p(X|\theta)) = \ln(p(X, Z|\theta)) - \ln(p(Z | X, \theta))$$

$$= \ln\left(\frac{p(X,z|\theta)}{q(z)}\right) - \ln\left(\frac{p(z|X,\theta)}{q(z)}\right)$$

$\to \times q(z)$

$\to \sum_z$

$$= \underbrace{\sum_z q(z) \ln\left(\frac{p(X,z|\theta)}{q(z)}\right)}_{\mathcal{L}(q,\theta)} + \underbrace{\sum_z -q(z) \ln\left(\frac{p(z|X,\theta)}{q(z)}\right)}_{= KL(q \| p(z|X,\theta))}$$

$\sum_z q(z) \times cst$
$= cst \underbrace{\sum_z q(z)}_{=1}$

Rappel sur la Kullback-Leibler

Si P et Q sont 2 var de densités p et q,

alors $KL(P \| Q) = \int_{-\infty}^{+\infty} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx = \mathbb{E}_P\left(\ln\left(\frac{P}{Q}\right)\right)$

"KL de P par % à Q"

Rmq :
- mesure de l'information qu'on perdrait si on utilisait Q pour approximer P
- KL n'est pas symétrique
- $KL(P \| Q) \geq 0$ avec $=$ ssi $P = Q$

Donc on réécrit ce qu'on avait :

$$\ell(\theta) = \mathcal{L}(q,\theta) + KL(q \| p(z|X,\theta))$$

E-step : On cherche à maximiser $\mathcal{L}(q, \theta^{old})$ en fct° de q, or $\ell(\theta)$ ne dépend pas de q, donc cela revient à minimiser $KL(q \| p(z|X,\theta))$, ie de forcer

$$q = p(z|X,\theta^{old}).$$

M-step : $\ell(\theta) = \mathcal{L}(p(z|X,\theta^{old}), \theta)$

$$= \sum_{z} p(z|x,\theta^{old}) \ln\left(\frac{p(x,z|\theta)}{p(z|x,\theta^{old})}\right)$$

$$= \underbrace{\sum_{z} p(z|x,\theta^{old}) \ln(p(x,z|\theta))}_{Q(\theta)} - \underbrace{\sum_{z} p(z|x,\theta^{old}) \ln(p(z|x,\theta^{old}))}_{\text{ne dépend pas de } \theta}$$

On cherche à maximiser :

$$Q(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left[\ln(\pi_k) + \ln(\mathcal{N}(x_n | \mu_k, \Sigma_k))\right]$$

M-step en calculs :

① Maximisation de Q en fct° de π

On doit respecter $\sum_{k=1}^{K} \pi_k = 1$ donc il s'agit d'une optimisation sous contrainte, donc on introduit le lagrangien

$$\mathcal{L}(\pi, \lambda) = Q(\pi, \mu, \Sigma) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\gamma_{nk}}{\pi_k} + \lambda = \frac{1}{\pi_k} \sum_{n=1}^{N} \gamma_{nk} + \lambda$$

$$\left[\forall k \in [\![1, K]\!],\ \frac{\partial \mathcal{L}}{\partial \pi_k} = 0\right] \Rightarrow \sum_{k=1}^{K} \pi_k \frac{\partial \mathcal{L}}{\partial \pi_k} = 0$$

$$\Rightarrow \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{nk} + \underbrace{\sum_{k=1}^{K} \pi_k}_{=1} \lambda = 0$$

$$\Rightarrow \sum_{n=1}^{N} \underbrace{\sum_{k=1}^{K} \underbrace{\gamma_{nk}}_{p(z_n = k \mid x_n)}}_{=1} + \lambda = 0$$

$$\Rightarrow \lambda = -N$$

En réinjectant la valeur de λ dans $\frac{\partial \mathcal{L}}{\partial \pi_k}$, on obtient $\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{1}{\pi_k} \sum_{n=1}^{N} \gamma_{nk} - N$.

Donc $\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \iff \pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}$

$$\boxed{\pi_k^{M\text{-step}} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}}$$

Rmq: C'est la moyenne sur toutes les observations de leur probabilité (postérieure) d'appartenir à la $k^{\text{ème}}$ gaussienne.

② Maximisation de Q en fct° de μ

$$\ln(\mathcal{N}(x_n, \mu_k, \Sigma_k)) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_k|) - \frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)$$

$$\frac{\partial \ln(\mathcal{N}(x_n|\mu_k, \Sigma_k))}{\partial \mu_k} = \Sigma_k^{-1}(x_n-\mu_k)$$

$f : x \mapsto \langle x | Ax \rangle$

$\Rightarrow \nabla f_x = (A + A^T)x$

et ici Σ_k^{-1} symétrique

$$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \gamma_{nk}\Sigma^{-1}(x_n-\mu_k)$$

$$= \Sigma^{-1}\left(\sum_{n=1}^{N}\gamma_{nk}x_n - \mu_k\sum_{n=1}^{N}\gamma_{nk}\right)$$

donc $\frac{\partial Q}{\partial \mu_k} = 0 \Leftrightarrow \boxed{\mu_k = \frac{\sum_{n=1}^{N}\gamma_{nk}x_n}{\sum_{n=1}^{N}\gamma_{nk}}}$

Rmq : C'est la moyenne de toutes les observations pondérées par leur probabilité (postérieure) d'appartenir à la $k^{\text{ème}}$ gaussienne.

③ Maximisation en fct° de Σ

Lemmes :

ⓐ $$\frac{\partial \ln(|A|)}{\partial A_{ij}} = (A^{-1})_{ji}$$

ⓑ $$\frac{\partial A^{-1}}{\partial A_{ij}} = -A^{-1} E_{ij} A^{-1} \quad (E_{ij} = \delta_{ij})$$

ⓒ $$\mathrm{Tr}(A E_{ij}) = A_{ji}$$

$$\frac{\partial \ln(\mathcal{N}(x_n \mid \mu_k, \Sigma_k))}{\partial \Sigma_{k,ij}} =$$

$$\frac{1}{2}(\Sigma_k^{-1})_{ji} + \frac{1}{2}\underbrace{\overbrace{(x_n-\mu_k)^T}^{1\times D}\,\overbrace{\Sigma_k^{-1}}^{D\times D}\,\overbrace{E_{ij}}^{D\times D}\,\overbrace{\Sigma_k^{-1}}^{D\times D}\,\overbrace{(x_n-\mu_k)}^{D\times 1}}_{1\times 1}$$

$$= \frac{1}{2}(\Sigma_k^{-1})_{ji} + \frac{1}{2}\mathrm{Tr}\Big(\underbrace{(x_n-\mu_k)^T \Sigma_k^{-1} E_{ij}}_{1\times D}\,\underbrace{\Sigma_k^{-1}(x_n-\mu_k)}_{D\times 1}\Big)$$

$$= \frac{1}{2}(\Sigma_k^{-1})_{ji} + \frac{1}{2}\mathrm{Tr}\left(\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T \Sigma_k^{-1} E_{ij}\right)$$

$$= \frac{1}{2}(\Sigma_k^{-1})_{ji} + \frac{1}{2}\left[\Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T \Sigma_k^{-1}\right]_{ji}$$

$$= \frac{1}{2}\left[\Sigma_k^{-1} + \Sigma_k^{-1}(x_n-\mu_k)(x_n-\mu_k)^T \Sigma_k^{-1}\right]_{ji}$$

$$\Rightarrow \frac{\partial \ln(\mathcal{N}(x_n \mid \mu_k, \Sigma_k))}{\partial \Sigma_k} = \frac{1}{2}\left[\Sigma_k^{-1} + \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}\right]$$

$$\frac{\partial Q}{\partial \Sigma_k} = 0 \Leftrightarrow \sum_{n=1}^{N} \gamma_{nk}\left[\Sigma_k^{-1} + \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}\right] = 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \gamma_{nk} \Sigma_k + \sum_{n=1}^{N} \gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T$$

$$\Leftrightarrow \boxed{\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma_{nk}}}$$