Theorem: The Bell numbers satisfy the following recurrence relation:

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$$

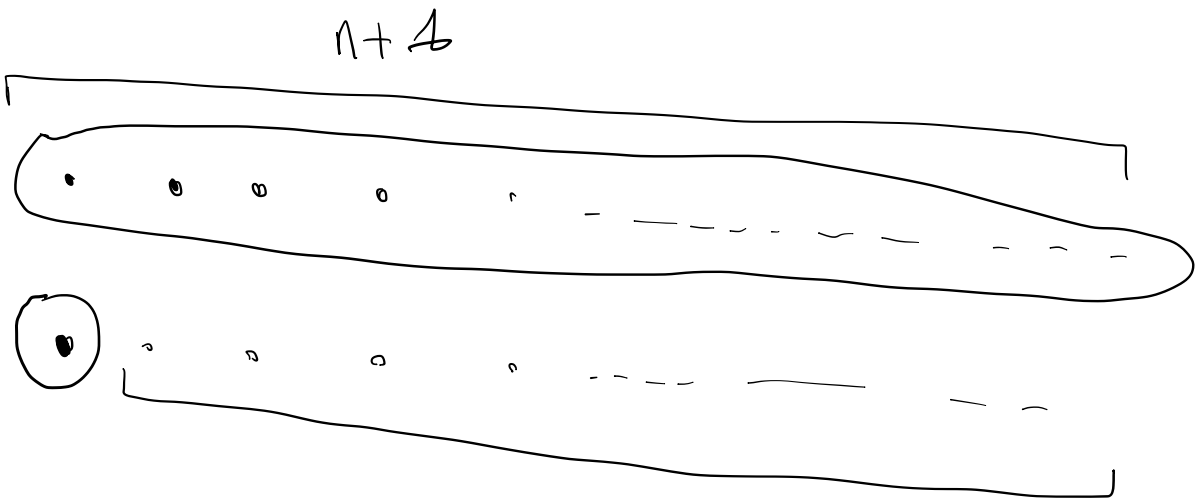
Proof: Proof of the recurrence relation:

First, consider a partition $\mathcal{P}$ of $(n+1)$ observation in $(C_k)_k$ clusters. The cluster containing the first observation is removed. The number of observations left varies from 0 (the cluster removed contained all the observations) to n (it contained only the first observation).

x_1 being always removed, the k observations left have to be chosen among n elements.

$\Rightarrow \binom{n}{k}$. Moreover, B_k is the number of ways to partition the k elements.

So: $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$



Remark: $S = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \hat{x}\|^2 = \sum_{j=1}^{p}\hat{\sigma}_j^2$ (with $x_i \in \mathbb{R}^p$ ∀i)

where $\hat{\sigma}_j^2$ is the empirical (biased) variance of variable j.

Proof:

$$S = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \hat{x}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \hat{x}_j)^2 = \sum_{j=1}^{p}\underbrace{\left\{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \hat{x}_j)^2\right\}}_{\hat{\sigma}_j^2}$$

$$\Rightarrow \boxed{S = \sum_{j=1}^{p}\hat{\sigma}_j^2}$$

Theorem (Huygens):

$$S = W + B$$

Proof:

$$\sum_{i=1}^{n} \|x_i - \hat{x}\|^2 = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|(x_i - \mu_k) + (\mu_k - \hat{x})\|^2$$

$$= \sum_{k=1}^{K} \sum_{x_i \in C_k} \left\{ \|x_i - \mu_k\|^2 + 2 (x_i - \mu_k)^T (\mu_k - \hat{x}) + \|\mu_k - \hat{x}\|^2 \right\}$$

$$= \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 + 2 \sum_{k=1}^{K} \underbrace{\left(\underbrace{\sum_{x_i \in C_k} x_i}_{n_k \mu_k} - \mu_k \underbrace{\sum_{x_i \in C_k} 1}_{n_k} \right)^T (\mu_k - \hat{x})}_{=0} + \sum_{k=1}^{K} n_k \|\mu_k - \hat{x}\|^2$$

$$= n W + n B$$

Property: Consider the random sample $(X_1, \dots, X_n)$ with $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$.
The MLE of μ and Σ are:

$$\rightarrow \hat{\mu} = \overline{x} \quad \text{(empirical mean)}$$

$$\rightarrow \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x})(x_i - \hat{x})^T$$

Proof:

$$\underset{(x_1,\dots,x_n)}{L}(\mu, \Sigma) = \log p(x_1, \dots, x_n \mid \mu, \Sigma)$$

$$= \sum_{i=1}^{n} \log p(x_i \mid \mu, \Sigma)$$

$$= \frac{-nd}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

for μ:

$$\underset{(x_1,\dots,x_n)}{L}(\mu, \Sigma) = -\frac{1}{2} \sum_{i=1}^{n} \left\{ x_i^T \Sigma^{-1} x_i - 2 x_i^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu \right\} + \text{cte}$$

(cte does not depend on μ)

$$= -\frac{1}{2} \sum_{i=1}^{n} \left\{ -2 x_i^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu \right\} + \text{cte}$$

Recall: consider $f : \mathbb{R}^d \to \mathbb{R}$, $f(x) = y^T x = \sum_{j=1}^{d} y_j x_j \Rightarrow \boxed{\nabla_x f(x) = y}$

Moreover: consider $g : \mathbb{R}^d \to \mathbb{R}$, $g(x) = x^T A x = \sum_{i=1}^{d} \sum_{j=1}^{d} x_i A_{ij} x_j = \sum_{i=1}^{d} x_i^2 A_{ii} + \sum_{i \neq j} x_i A_{ij} x_j$

$$\Rightarrow \frac{\partial g(x)}{\partial x_k} = 2 x_k A_{kk} + \underbrace{\sum_{j \neq a} A_{aj} x_j + \sum_{i \neq a} x_i A_{ia}}$$

$$\sum_{j \neq a} (A_{aj} + A_{ja}) x_j$$

$$= 2 \sum_{j \neq a} A_{aj} x_j \quad \text{if } A \text{ is symmetric}$$

$$\Rightarrow \frac{\partial g(x)}{\partial x_a} = 2 x_a A_{aa} + 2 \sum_{j \neq a} x_j A_{aj} = 2 \sum_{j=1}^{d} x_j A_{aj} = 2 A_{a\cdot} x$$

$$\Rightarrow \boxed{\nabla_x g(x) = 2Ax}$$

Going back to our problem, we have:

$$\nabla_\mu \underset{(x_1,\dots,x_n)}{L}(\mu, \Sigma) = \sum_{i=1}^{n} \Sigma^{-1} x_i - n \Sigma^{-1} \mu$$

We solve: $\nabla_\mu \underset{(x_1,\dots,x_n)}{L}(\mu, \Sigma) = 0 \iff \Sigma^{-1} \sum_{i=1}^{n} x_i = n \Sigma^{-1} \mu$

$$\Rightarrow \boxed{\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}} \quad \text{(only if } \Sigma \text{ can be inverted)}$$

For Σ: $\underset{(x_1,\dots,x_n)}{L}(\mu,\Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) + \text{cste}$

Recall: $x^TAx = \sum_{q,l} x_q A_{ql} x_l = \sum_{q,l} A_{ql} B_{lq}$ where $B_{lq} = x_l x_q \Leftrightarrow B = x x^T$

Moreover: $(AB)_{ij} = \sum_k A_{ik} B_{kj} \Rightarrow \mathrm{TR}(AB) = \sum_i (AB)_{ii} = \sum_i \sum_k A_{ik} A_{ki}$

$\Rightarrow$ so: $\boxed{x^TAx = \mathrm{TR}(Axx^T)}$

Therefore: $\underset{(x_1,\dots,x_n)}{L}(\mu,\Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{TR}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\right) + \text{cste}$

Recall that: $\boxed{\nabla_X \log|X| = (X^{-1})^T = (X^T)^{-1}}$

Moreover: $\boxed{\nabla_X a^T X b = ab^T}$ and $\boxed{\nabla_X \mathrm{TR}(AX^{-1}B) = -X^{-T}A^TB^TX^{-T}}$

So, we find: $\nabla_\Sigma \underset{(x_1,\dots,x_n)}{L}(\mu,\Sigma) = -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}$

We solve: $\nabla_\Sigma \underset{(x_1,\dots,x_n)}{L}(\mu,\Sigma) = 0_{d\times d} \Leftrightarrow -\frac{n}{2}\Sigma + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T = 0_{d\times d}$

$$\Rightarrow \boxed{\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T}$$

Property: Given the observations (and the parameters), all the Z_i are independent.

Proof:

$$p((z_i)_i \mid (x_i)_i, \pi, \theta) = \frac{p((x_i, z_i)_i \mid \pi, \theta)}{p((x_i)_i \mid \pi, \theta)} \quad \text{(Bayes rule)}$$

$$= \frac{\prod_{i=1}^{n} p(x_i, z_i \mid \pi, \theta)}{\prod_{i=1}^{n} p(x_i \mid \pi, \theta)} = \prod_{i=1}^{n} \frac{p(x_i, z_i \mid \pi, \theta)}{p(x_i \mid \pi, \theta)}$$

$$= \prod_{i=1}^{n} p(z_i \mid x_i, \pi, \theta)$$

Property: $p(z_i \mid x_i, \pi, \theta) = \mathcal{M}(z_i; 1, \tau_i)$ where $\tau_i^T = (\tau_{i1}, \dots, \tau_{ik})$ and $\tau_{ik} = \frac{\pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i; \mu_l, \Sigma_l)}$

Proof:

$$p(z_i \mid x_i, \pi, \theta) = \frac{p(x_i \mid z_i, \theta)\, p(z_i \mid \pi)}{p(x_i \mid \pi, \theta)}$$

$$= \frac{\prod_{k=1}^{k} \left(\pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k)\right)^{z_{ik}}}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i; \mu_l, \Sigma_l)}$$

$$= \prod_{k=1}^{k} \left(\frac{\pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i; \mu_l, \Sigma_l)} \right)^{z_{ik}}$$

Property: if $Y \sim \mathcal{M}(1, \mu)$ where $\mu^T = (\mu_1 \dots \mu_k)$

then: $E[Y_k] = \mu_k$

Proof:

$$E[Y] = \sum_y y\, P(Y = y) = \sum_y y \prod_{k=1}^{k} \mu_k^{y_k} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}$$

$\Rightarrow$ in particular: $E[Y_k] = \mu_k$.

Property: The EM algorithm does increase the log-likelihood $L_{(x_i)_i}(\pi, \theta)$ at each iteration.

Proof: (general): for any law $R((z_i)_i)$, the log-likelihood can always be decomposed in two terms:

(1) $$L_{(x_i)_i}(\mu, \theta) = \log p((x_i)_i | \mu, \theta)$$
$$= \mathcal{L}(R(z_i)_i; \mu, \theta) + KL(R((z_i)_i) \| p((z_i)_i | (x_i)_i; \pi, \theta))$$

where: $$\mathcal{L}(R(z_i)_i, \mu, \theta) = \sum_{(z_i)_i} R((z_i)_i) \log \frac{p((x_i, z_i)_i | \pi, \theta)}{R((z_i)_i)}$$

and $$KL(R((z_i)_i) \| p((z_i)_i | (x_i)_i, \pi, \theta) = -\sum_{(z_i)_i} R((z_i)_i) \log \frac{p((z_i)_i | (x_i)_i, \pi, \theta)}{R((z_i)_i)}$$

Proof of (1): $$\mathcal{L}(R((z_i)_i); \mu, \theta) + KL(R((z_i)_i) \| p((z_i)_i | (x_i)_i, \pi, \theta)$$
$$= \sum_{(z_i)_i} R((z_i)_i) \log \frac{p((x_i, z_i)_i | \pi, \theta)}{p((z_i)_i | (x_i)_i, \pi, \theta)}$$
$$= \sum_{(z_i)_i} R((z_i)_i) \log p((x_i)_i | \pi, \theta)$$
$$= \log p((x_i)_i | \pi, \theta)$$
$$= L_{(x_i)_i}(\pi, \theta)$$

2) Therefore: since $\underset{(x_i)_i}{L}(\pi,\theta)$ does not depend on $R((z_i)_i)$, when $\mathcal{L}\uparrow$ then $KL\downarrow$

In fact: $\mathcal{L}$ acts as a lower bound of the log-likelihood.
$\Rightarrow$ we focus the optimisation on $\mathcal{L}$ with respect to $R((z_i)_i)$, π,θ.

EM: E-step: π,θ fixed. $\mathcal{L}$ maximized with respect to $R((z_i)_i)$. In the case of Gaussian mixture models, $p((z_i)_i|(x_i)_i,\pi,\theta)$ has an analytical form, so KL can reach 0 by setting $R((z_i)_i) = p((z_i)_i|(x_i)_i,\pi,\theta)$ [Optimal]

$$= \prod_{i=1}^{n} \mathcal{M}(z_i;1,\tau_i)$$

$\Leftrightarrow$ characterized by the τ_i.

M-step: τ_i fixed, $\mathcal{L}$ maximized with respect to π and θ:

$$(\hat{\pi},\hat{\theta}) = \underset{\pi,\theta}{\operatorname{argmax}}\, \mathcal{L}(R((z_i)_i),\pi,\theta) = \underset{\pi,\theta}{\operatorname{argmax}} \sum_{(z_i)_i} R((z_i)_i)\log p((x_i,z_i)_i|\pi,\theta)$$

$$= \underset{\pi,\theta}{\operatorname{argmax}}\, \underset{(z_i)_i}{E}\left[\underset{(x_i,z_i)_i}{L}(\pi,\theta)\right]$$

(3) Recall that after the E-step, $KL=0$ so $\underset{(x_i)_i}{L}(\pi,\theta) = \mathcal{L}(R((z_i)_i),\pi,\theta)$

Thus: Init: τ_i^0

after M-step: $\pi^0,\theta^0 \to \mathcal{L}((\tau_i^0)_i,\pi^0,\theta^0)$

E-step: $\tau_i^1 \to \mathcal{L}((\tau_i^1)_i,\pi^0,\theta^0) = \underset{(x_i)_i}{L}(\pi^0,\theta^0)$

M-step: $\pi^1,\theta^1 \to \mathcal{L}((\tau_i^1)_i,\pi^1,\theta^1) \geq \mathcal{L}((\tau_i^1),\pi^0,\theta^0)$

E-step: $\tau_i^2 \to \mathcal{L}((\tau_i^2)_i,\pi^1,\theta^1) = \underset{(x_i)_i}{L}(\pi^1,\theta^1) \geq \mathcal{L}((\tau_i^1)_i,\pi^1,\theta^1)$

So, in summary: $\underset{(x_i)_i}{L}(\pi^1,\theta^1) = \mathcal{L}((\tau_i^2)_i,\pi^1,\theta^1) \geq \mathcal{L}((\tau_i^1)_i,\pi^1,\theta^1) \geq \mathcal{L}((\tau_i^1)_i,\pi^0,\theta^0) = \underset{(x_i)_i}{L}(\pi^0,\theta^0)$

(E) (M)

that is $\boxed{\underset{(x_i)_i}{L}(\pi^1,\theta^1) \geq \underset{(x_i)_i}{L}(\pi^0,\theta^0)}$