Expectation-Maximization derivation of CELLMA

Alexandre Coudray, Anya-Aurore Mauron, Justine Montavon

December 2021

From the problem described in CELLMA appendix [1], we would like to find the transition matrix $\Theta \in \mathbb{R}^{K\times K}$ that minimizes the following log-likelihood :

$$\min_{\Theta} -\sum_{r=1}^{N_k} \log \sum_{\mathbf{E}\in\mathcal{S}^{|\rho(i)|}} \exp\left(\sum_{i\in\mathcal{K}_r} \log(\mathbf{m}_i^T \Theta^{\mathcal{X}_i,\mu(i)} \epsilon_{\mu(i)}) + \sum_{u\in\mathcal{Q}_r} \log(\epsilon_u^T \Theta^{\mathcal{X}_u,\mu(u)} \epsilon_{\mu(u)})\right) \tag{1}$$

with the two following constraints : $\Theta^T \mathbb{1} = \mathbb{1}$ and $\Theta \geq 0$ and with $\mu(i)$ represent the mother of cell i.

We consider the above problem without the marginalisation over $\mathbf{E} \in \mathcal{S}^{|\rho(i)|}$. We want to resolve this problem using an Expectation-Maximization algorithm considering the state of non-observed cells $u \in \mathcal{Q}_r$ as latent variables. Latent variables are vector $\mathbf{z}_u$ of length K, equal to 1 if cell u is in state k for $k \in \{1, ..., K\}$, and equal to 0 otherwise. We consider the soft-assignment $\mathbf{s}_u$ of $\mathbf{z}_u$ for each state k:

$$s_{ku} = P(z_{ku} = 1 \mid \Theta, \mathbf{s}_{\mu(u)^{-1}}) \tag{2}$$

The equation above is telling us the probability of a mother cell u to be in state k, given the transition matrix Θ and the state probability distribution of its daughter $\mathbf{s}_{\mu(u)^{-1}}$. We therefore replaced ϵ variables in (1), representing every possible perfect realization of latent variables, by $\mathbf{s}_u$ variables in (15), which represent their soft assignment to each state k that we wish to infer by using the probability distribution of their daughter cells.

We can rewrite the log-likelihood as :

$$\min_{\Theta} -\sum_{r=1}^{N_k} \left(\sum_{i\in\mathcal{K}_r} \log(\mathbf{m}_i^T \Theta^{\mathcal{X}_i,j} \mathbf{s}_{\mu(i)}) + \sum_{u\in\mathcal{Q}_r} \log(\mathbf{s}_u^T \Theta^{\mathcal{X}_u,j} \mathbf{s}_{\mu(u)})\right) \tag{3}$$

From another perspective, we will consider each cell at each time step as a latent variable, instead of using only cells before division as in (1). Therefore we add latent variables to the problem (around 10 times more) but we make the resolution of Θ substantially easier as the power on Θ matrix disappears. Also, we introduce a sum over K different states (by definition of the matrix product) to be able to use the EM machinery :

$$-\max_{\Theta} \sum_{r=1}^{N_k} \left(\sum_{i\in\mathcal{K}_r} \log\left(\sum_{k=1}^{K} \left(\mathbf{m}_i^T\Theta\right)(k)\mathbf{s}_{\mu(i)}(k)\right) + \sum_{u\in\mathcal{Q}_r} \log\left(\sum_{k=1}^{K} \left(\mathbf{s}_u^T\Theta\right)(k)\mathbf{s}_{\mu(u)}(k)\right)\right) \tag{4}$$

where $\mathbf{v}(k)$ indicates the k^{th} component of the vector $\mathbf{v}$.
By using Jensen's inequality we know that $\log(\sum_i q_i g_i) \geq \sum_i q_i \log(g_i)$ whenever the condition $\sum_i q_i = 1$ is satisfied. Therefore we can get a lower bound function for (4) :

$$\begin{aligned}
&\sum_{r,i} \log\left(\sum_k (\mathbf{m}_i^T \Theta)(k)\mathbf{s}_{\mu(i)}(k)\right) + \sum_{r,u} \log\left(\sum_k (\mathbf{s}_u^T \Theta)(k)\mathbf{s}_{\mu(u)}(k)\right) \\
&\geq \sum_{r,i}\sum_k q_{ki} \log\left(\frac{(\mathbf{m}_i^T \Theta)(k)\mathbf{s}_{\mu(i)}(k)}{q_{ki}}\right) + \sum_{r,u}\sum_k q_{ku} \log\left(\frac{(\mathbf{s}_u^T \Theta)(k)\mathbf{s}_{\mu(u)}(k)}{q_{ku}}\right)
\end{aligned} \tag{5}$$

We can rewrite our lower bound function by transforming the log of a division into a subtraction of log, therefore we get :

$$\sum_{r,i}\sum_k q_{ki} \log((\mathbf{m}_i^T \Theta)(k)\mathbf{s}_{\mu(i)}(k)) - q_{ki}\log q_{ki} + \sum_{r,u}\sum_k q_{ku} \log\left(\left(\mathbf{s}_u^T \Theta\right)(k)\mathbf{s}_{\mu(u)}(k)\right) - q_{ku}\log q_{ku} \tag{6}$$

While iterating over the Expectation-Maximization algorithm, we will alternate between the E-step where we recalculate our lower bound by defining q_{ki} and q_{ku} at step (t) by using Θ and $\mathbf{s}$ defined at step $(t-1)$. Therefore we can maximize (6) during M-step with respect to Θ, and redefine $\mathbf{s}_u$ with respect to $\mathbf{q}_u$.

Initialization

First our parameter Θ needs to be initialized to some Θ_0 values to be able to launch the first E-step. We chose to use a transition matrix $\Theta_0 = \mathbf{I}$ plus some noise (where $\mathbf{I}$ is the identity matrix). For each non-observed cells, we initialize its soft-assigment of latent variable $\mathbf{s}_u$ to the average of all its observed daughter cells at the leaf of the tree.

E-step

During the E-step, we will generate a new estimate of q_{ki} and q_{ku} at step (t+1) based on previous calculation of Θ and $\mathbf{s}$ at step (t). We therefore define a lower bound at step (t+1) by defining q_{ki} and q_{ku}, which corresponds to the posterior distribution of the latent variables. It turns out that q_{ki} satisfy :

$$q_{ki} = P(z_{ki} = 1 \mid \Theta, \mathbf{m}_{\mu(i)^{-1}}) \tag{7}$$

Which also corresponds to the definition of the soft assignment s_{ki}. Using Bayes rule, we can rewrite it as :

$$P(z_{ki} = 1 \mid \mathbf{m}_{\mu(i)^{-1}}, \Theta) = \frac{P(\mathbf{m}_{\mu(i)^{-1}} \mid z_{ki} = 1, \Theta)P(z_{ki} = 1|\Theta)}{\sum_{k=1}^{K} P(\mathbf{m}_{\mu(i)^{-1}} \mid z_{ki} = 1, \Theta)P(z_{ki} = 1|\Theta)} \tag{8}$$

Where $P(z_{ki} = 1)$ corresponds to the prior distribution π_k, and where the probability of observing the daughter cell knowing the mother state is :

$$P(\mathbf{m}_{\mu(i)^{-1}} \mid z_{ki}=1,\Theta) = (\mathbf{m}_{\mu(i)^{-1}}^T\Theta)(k)\mathbf{s}_i(k) \tag{9}$$

Therefore we know all the terms in q_{ki} which can be rewritten as:

$$q_{ki} = \frac{\pi_k\ (\mathbf{m}_i^T\Theta)(k)\mathbf{s}_{\mu(i)}(k)}{\sum_k \pi_k\ (\mathbf{m}_i^T\Theta)(k)\mathbf{s}_{\mu(i)}(k)} \tag{10}$$

In the case of a division, the conditional probability depends on two cells, and should therefore be written as the probability of the mother to be in state k, knowing the observed state of its two daughter cells $\mathbf{m}_{\mu_1(i)^{-1}}$ and $\mathbf{m}_{\mu_2(i)^{-1}}$:

$$P(z_{ki}=1 \mid \Theta, \mathbf{m}_{\mu_1(i)^{-1}}, \mathbf{m}_{\mu_2(i)^{-1}}) = \frac{P(\mathbf{m}_{\mu_1(i)^{-1}}, \mathbf{m}_{\mu_2(i)^{-1}} \mid z_{ki}=1,\Theta)P(z_{ki}=1)}{\sum_{k=1}^K P(\mathbf{m}_{\mu_1(i)^{-1}}, \mathbf{m}_{\mu_2(i)^{-1}} \mid z_{ki}=1,\Theta)P(z_{ki}=1)} \tag{11}$$

By using the fact that $P(A\cap B|C) = P(A|C)P(B|C)$ and therefore that daughter cells are conditionally independent given their mother state, we can rewrite the above relation as :

$$P(z_{ki}=1 \mid \Theta, \mathbf{m}_{\mu_1(i)^{-1}}, \mathbf{m}_{\mu_2(i)^{-1}}) = \frac{P(\mathbf{m}_{\mu_1(i)^{-1}} \mid z_{ki}=1,\Theta)P(\mathbf{m}_{\mu_2(i)^{-1}} \mid z_{ki}=1,\Theta)P(z_{ki}=1)}{\sum_{k=1}^K P(\mathbf{m}_{\mu_1(i)^{-1}} \mid z_{ki}=1,\Theta)P(\mathbf{m}_{\mu_2(i)^{-1}} \mid z_{ki}=1,\Theta)P(z_{ki}=1)} \tag{12}$$

As we know all the terms, we can define q_{ki} as :

$$q_{ki} = \frac{\pi_k\ (\mathbf{m}_1^T\Theta)(k)\mathbf{s}_{\mu(1)}(k)(\mathbf{m}_2^T\Theta)\mathbf{s}_{\mu(2)}(k)}{\sum_k \pi_k(\mathbf{m}_1^T\Theta)(k)\mathbf{s}_{\mu(1)}(k)(\mathbf{m}_2^T\Theta)\mathbf{s}_{\mu(2)}(k)} \tag{13}$$

where $\mu(1)$ and $\mu(2)$ are the two daughters of cell i.

As we are dealing with a tree-like structure, state probabilities of mother cells are conditionally dependent of their lineage. Therefore when one wish to compute the probability of a mother cell with latent state k, we need to take all daugthers and grand-daughters into account. This can be done by chaining probabilities are therefore the probability q_{ku} depends on $q_{k\mu^{-1}(u)}$, namely the probability distribution of its daugther $\mu^{-1}(u)$. We can therefore define q_{ku} when there is no division as being :

$$q_{ku} = \frac{\pi_k\ (\mathbf{s}_{\mu^{-1}(u)}^T\Theta)(k)\mathbf{s}_u(k)}{\sum_k \pi_k\ (\mathbf{s}_{\mu^{-1}(u)}^T\Theta)(k)\mathbf{s}_u(k)} \tag{14}$$

where s_{ku} is the soft assignment, in other words the probability of cell u to be in state k :

$$s_{ku} = P(z_{ku}=1 \mid \Theta, \mathbf{s}_{\mu(u)^{-1}}) \tag{15}$$

Using conditional probability for two daughter cells, similarly than in (13), we can derived q_{ku} for the case of a mother with two daugthers :

$$q_{ku} = \frac{\pi_k\ (\mathbf{s}_{\mu^{-1}(1)}^T\Theta)(k)\mathbf{s}_1(k)(\mathbf{s}_{\mu^{-1}(2)}^T\Theta)\mathbf{s}_2(k)}{\sum_k \pi_k(\mathbf{s}_{\mu^{-1}(1)}^T\Theta)(k)\mathbf{s}_1(k)(\mathbf{s}_{\mu^{-1}(2)}^T\Theta)\mathbf{s}_2(k)} \tag{16}$$

M-step

As we have estimated q_{ki} and q_{ku} during E-step, we can estimate Θ using $\mathbf{s}$ from step $(t-1)$. Therefore we wish to maximize (6) with respect to Θ. Note that the terms $q_{ki}\log q_{ki}$ and $q_{ku}\log q_{ku}$ do not depend on Θ and can therefore be ignored. We obtain the objective function $f(\Theta)$ to optimize:

$$\begin{aligned} -\max_{\Theta\in\mathbb{R}^{K\times K}} f(\Theta) = -\max_{\Theta\in\mathbb{R}^{K\times K}} &\sum_{r,i}\sum_{k} q_{ki}\log\left((\mathbf{m}_i^T\Theta)(k)\mathbf{s}_{\mu(i)}(k)\right) \\ &+\sum_{r,u}\sum_{k} q_{ku}\log\left(\left(\mathbf{s}_u^T\Theta\right)(k)\mathbf{s}_{\mu(u)}(k)\right) \end{aligned} \tag{17}$$

with the two following constraints : $\Theta^T\mathbb{1}=\mathbb{1}$ and $\Theta\geq 0$.

It can be shown that the optimization problem (17) is easier to solve than the very first optimization (1). The objective function $f(\Theta)$ is a sum of concave functions with positive coefficients, therefore a concave function. We have that $S=\{\Theta\in\mathbb{R}^{K\times K} \quad \text{such that} \quad \Theta^T\mathbb{1}=\mathbb{1},\Theta\geq 0\}$ is convex. Hence, the optimization problem (17) admits a global maximum.
Let's show that $f(\Theta)$ is indeed concave. The terms

$$\log((\mathbf{m}_i^T\Theta)(k)\mathbf{s}_{\mu(i)}(k)) \quad \text{and} \quad \log\left(\left(\mathbf{s}_u^T\Theta\right)(k)\mathbf{s}_{\mu(u)}(k)\right) \tag{18}$$

are concave with respect to Θ. Let's show it for $\log((\mathbf{m}_i^T\Theta)(k)\mathbf{s}_{\mu(i)}(k))$ (it is the same principle for the second term $\log\left(\left(\mathbf{s}_u^T\Theta\right)(k)\mathbf{s}_{\mu(u)}(k)\right)$). Let $\Theta_1,\,\Theta_2\in\mathbb{R}^{K\times K}$ and $t\in[0,1]$. We have:

$$\begin{aligned} &\log((\mathbf{m}_i^T(t\Theta_1+(1-t)\Theta_2))(k)\mathbf{s}_{\mu(i)}(k)) \\ &=\log((\mathbf{m}_i^T(t\Theta_1+(1-t)\Theta_2)(k))+\log(\mathbf{s}_{\mu(i)}(k)) \\ &=\log(t\mathbf{m}_i^T\Theta_1(k)+(1-t)\mathbf{m}_i^T\Theta_2(k))+\log(\mathbf{s}_{\mu(i)}(k)) \\ &\geq t\log(\mathbf{m}_i^T\Theta_1(k))+(1-t)\log((\mathbf{m}_i^T\Theta_2)(k))+\log(\mathbf{s}_{\mu(i)}(k)) \end{aligned} \tag{19}$$

The last inequality is given by the concavity of the logarithm. Now, by writing

$$\log(\mathbf{s}_{\mu(i)}(k))=t\log(\mathbf{s}_{\mu(i)}(k))+(1-t)\log(\mathbf{s}_{\mu(i)}(k)) \tag{20}$$

we have:

$$\begin{aligned} &\log((\mathbf{m}_i^T(t\Theta_1+(1-t)\Theta_2))(k)\mathbf{s}_{\mu(i)}(k)) \\ &\geq t\log((\mathbf{m}_i^T\Theta_1)(k))+t\log(\mathbf{s}_{\mu(i)}(k))+(1-t)\log((\mathbf{m}_i^T\Theta_2)(k))+(1-t)\log(\mathbf{s}_{\mu(i)}(k)) \\ &\geq t\log((\mathbf{m}_i^T\Theta_1)(k)\mathbf{s}_{\mu(i)}(k))+(1-t)\log((\mathbf{m}_i^T\Theta_2)(k)\mathbf{s}_{\mu(i)}(k)) \end{aligned} \tag{21}$$

which proves the concavity of $\log((\mathbf{m}_i^T\Theta)(k)\mathbf{s}_{\mu(i)}(k))$. Since $f(\Theta)$ is a sum of concave functions with positive coefficients, it is itself concave.

Now, together with the constraints $\Theta^T\mathbb{1}=\mathbb{1}$ and $\Theta\geq 0$, the optimization problem (17) can be easily resolved with an interior point method. The jacobian with respect to Θ is computable, which results to a lot less computational time (with different trials, we managed to reduce the computation time to a few seconds instead to a few minutes, to give a broad idea). Let's define the jacobian $\nabla_\theta f(\Theta)\in\mathbb{R}^{K\times K}$ with $a,b\in\{1,...,K\}$ as

$$\left(\nabla_\theta f(\Theta)\right)_{ab}=\partial\Theta_{ab}f(\Theta) \tag{22}$$

It is given by:

$$\nabla_\theta f(\Theta) = \sum_{r,i}\sum_k \frac{q_{ki}}{(\mathbf{m}_i^T\Theta)(k)}\mathcal{M}_i^k + \sum_{r,u}\sum_k \frac{q_{ku}}{(\mathbf{s}_u^T\Theta)(k)}\mathcal{S}_u^k \tag{23}$$

where $\mathcal{M}_i^k$ and $\mathcal{S}_u^k$ are matrices in $\mathbb{R}^{K\times K}$ with null elements, except in the k^{th} column where $\mathbf{m}_i$, respectively $\mathbf{s}_u$, lie. Indeed, by rearranging the terms of $f(\Theta)$:

$$\begin{aligned} f(\Theta) &= \sum_{r,i}\sum_k q_{ki}\log\{(\mathbf{m}_i^T\Theta)(k)\mathbf{s}_{\mu(i)}(k)\} + \sum_{r,u}\sum_k q_{ku}\log\{(\mathbf{s}_u^T\Theta)(k)\mathbf{s}_{\mu(u)}(k)\} \\ &= \sum_{r,i}\sum_k q_{ki}\log\{(\mathbf{m}_i^T\Theta)(k)\} + \sum_{r,u}\sum_k q_{ku}\log\{(\mathbf{s}_u^T\Theta)(k)\} \\ &+ \sum_{r,i}\sum_k q_{ki}\log\{\mathbf{s}_{\mu(i)}(k)\} + \sum_{r,u}\sum_k q_{ku}\log\{\mathbf{s}_{\mu(u)}(k)\} \end{aligned} \tag{24}$$

We do not consider the two last double sums computing the gradient, since they don't depend on Θ. Denoting $\partial\Theta_{ab}$ the derivative with respect to Θ_{ab}, for $a, b \in \{1, ..., K\}$, we have:
for $b = k$

$$\begin{aligned} \partial\Theta_{ab}f(\Theta) &= \sum_{r,i}\sum_k q_{ki}\partial\Theta_{ab}\log((\mathbf{m}_i^T\Theta)(k)) + \sum_{r,u}\sum_k q_{ku}\partial\Theta_{ab}\log((\mathbf{s}_u^T\Theta)(k)) \\ &= \sum_{r,i}\sum_k q_{ki}\partial\Theta_{ab}\log(\sum_{l=1}^K \mathbf{m}_i(l)\Theta_{lk}) + \sum_{r,u}\sum_k q_{ku}\partial\Theta_{ab}\log(\sum_{l=1}^K \mathbf{s}_u(l)\Theta_l k) \\ &= \sum_{r,i}\sum_k q_{ki}\frac{\mathbf{m}_i(a)}{(\mathbf{m}_i^T\Theta)(k)} + \sum_{r,u}\sum_k q_{ku}\frac{\mathbf{s}_u(a)}{(\mathbf{s}_u^T\Theta)(k)} \end{aligned} \tag{25}$$

The second equality is given by the definition of matrix product. The third equality is obtained by the derivative chain rule.
For $b \neq k$

$$\partial\Theta_{ab}f(\Theta) = 0 \tag{26}$$

since all deratives of log are null in this case; there is no component Θ_{ab}. This finally yields the result at equation (23).

During the M-step, we are also interested into getting an estimate of the soft-assignment $\mathbf{s}$ of our latent variables. Interestingly, it can be shown that $q_{ku}^{(t-1)}$ and $\mathbf{s}_{ku}^{(t)}$ both correspond to the posterior probability of our latent variables at step (t) :

$$s_{ku}^{(t)} = P(z_{ku}^{(t)} = 1 \mid \Theta^{(t)}, \mathbf{s}_{k\mu^{-1}(u)}^{(t)}, \mathbf{m}) = q_{ku}^{(t-1)} \tag{27}$$

The above relation holds for a single cell connected between two EM steps.

Convergence

We stop the iterations over the Expectation-Maximization algorithm whenever the maximum number of iterations or the following convergence criterion are reached :

$$\left\| \Theta^{(t+1)} - \Theta^{(t)} \right\| \leq \text{tol} \tag{28}$$

References

[1] Gioele La Manno and Alex Lederer. "CELLMA: Cell-state transition Estimation by Lineage Leaf-state Markov Analysis". In: (2021).