Supplementary Material: EM algorithm for estimating parameters of network evolution

Janusz Dutkowski, Jerzy Tiuryn

The $\mathcal{Q}$ function

We now derive the $\mathcal{Q}$ function for our model. For two parameter sets Θ and Θ^t we define

$$\mathcal{Q}(\Theta|\Theta^t) = \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}, \Theta^t) \log P(\mathbf{x}, \mathbf{y}|\Theta). \tag{1}$$

The probability of the data given the parameters Θ can be expressed as the product of the probabilities of the initial interactions and the transition probabilities:

$$P(\mathbf{x}, \mathbf{y}|\Theta) = \prod_{e \in E} \left(p_{e,\Theta}^{A_e(\mathbf{y})} \cdot (1 - p_{e,\Theta})^{A_{\neg e}(\mathbf{y})} \right).$$

Taking the logarithm we obtain:

$$\log P(\mathbf{x}, \mathbf{y}|\Theta) = \sum_{e \in E} (A_e(\mathbf{y}) \log p_{e,\Theta} + A_{\neg e}(\mathbf{y})(1 - \log p_{e,\Theta}))$$

which we can plug into (1):

$$\begin{aligned}
\mathcal{Q}(\Theta|\Theta^t) &= \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}, \Theta^t) \left[\sum_{e \in E} (A_e(\mathbf{y}) \log p_{e,\Theta} + A_{\neg e}(\mathbf{y})(1 - \log p_{e,\Theta})) \right] \\
&= \sum_{\mathbf{y}} \sum_{e \in E} P(\mathbf{y}|\mathbf{x}, \Theta^t) A_e(\mathbf{y}) \log p_{e,\Theta} + \sum_{\mathbf{y}} \sum_{e \in E} P(\mathbf{y}|\mathbf{x}, \Theta^t) A_{\neg e}(\mathbf{y}) \log(1 - p_{e,\Theta}) \\
&= \sum_{e \in E} \log p_{e,\Theta} \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}, \Theta^t) A_e(\mathbf{y}) + \sum_{e \in E} \log(1 - p_{e,\Theta}) \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}, \Theta^t) A_{\neg e}(\mathbf{y})
\end{aligned}$$

We thus obtain the following formula for the $\mathcal{Q}$

$$\mathcal{Q}(\Theta|\Theta^t) = \sum_{e} \left(\mathbb{E}(A_e) \log p_{e,\Theta} + \mathbb{E}(A_{\neg e}) \log(1 - p_{e,\Theta}) \right), \tag{2}$$

where the mean values are taken with respect to the probability distribution $P(-|\mathbf{x}, \Theta^t)$.

The E-step

In the E-step we compute the mean values $\mathbb{E}(A_e)$ and $\mathbb{E}(A_{\neg e})$, with e ranging over all types. The mean values are taken with respect to the probability distribution $P(-|\mathbf{x}, \Theta^t)$. They are computed using Pearl's message passing algorithm. Here we show only the slightly harder case when the event type e is associated either with duplications or speciation, i.e. when we have to consider both the parent and the child in the tree to determine the exact event.

Summing over all pairs of random variables (p, c) for which transition e can take place we have:

$$
\begin{aligned}
\mathbb{E}(A_e) &= \sum_{(p,c)} P(p, c|\mathbf{x}, \Theta^t) \\
&= \sum_{(p,c)} \frac{P(p, c, \mathbf{x}|\Theta^t)}{P(\mathbf{x}|\Theta^t)} \\
&= \sum_{(p,c)} \frac{P(c, \mathbf{x}|p, \Theta^t)P(p|\Theta^t)}{P(\mathbf{x}|\Theta^t)} \\
&= \sum_{(p,c)} \frac{P(c, \mathbf{x}|p, \Theta^t)P(p|\Theta^t)P(p, \mathbf{x}|\Theta^t)}{P(\mathbf{x}|\Theta^t)P(p, \mathbf{x}|\Theta^t)} \\
&= \sum_{(p,c)} \frac{P(c, \mathbf{x}|p, \Theta^t)P(p|\Theta^t)P(p|\mathbf{x}, \Theta^t)P(\mathbf{x}|\Theta^t)}{P(\mathbf{x}|\Theta^t)P(\mathbf{x}|p, \Theta^t)P(p|\Theta^t)} \\
&= \sum_{(p,c)} \frac{P(p|\mathbf{x}, \Theta^t)P(\mathbf{x}, c|p, \Theta^t)}{P(\mathbf{x}|p, \Theta^t)} = (*).
\end{aligned}
$$

We split the evidence $\mathbf{x}$ into the evidence below p ($\mathbf{d}_p$) and the evidence above p ($\mathbf{n}_p$) and take advantage of the implied conditional independencies:

$$
\begin{aligned}
(*) &= \sum_{(p,c)} \frac{P(p|\mathbf{x}, \Theta^t)P(\mathbf{d}_p, \mathbf{n}_p, c|p, \Theta^t)}{P(\mathbf{d}_p, \mathbf{n}_p|p, \Theta^t)} \\
&= \sum_{(p,c)} \frac{P(p|\mathbf{x}, \Theta^t)P(\mathbf{d}_p, c|p, \Theta^t)P(\mathbf{n}_p|p, \Theta^t)}{P(\mathbf{d}_p|p, \Theta^t)P(\mathbf{n}_p|p, \Theta^t)} \\
&= \sum_{(p,c)} \frac{P(p|\mathbf{x}, \Theta^t)P(\mathbf{d}_p, c|p, \Theta^t)}{P(\mathbf{d}_p|p, \Theta^t)}, \qquad (3)
\end{aligned}
$$

where $P(\mathbf{d}_p, c|p, \Theta^t)$ can be written as:

$$\begin{aligned} P(\mathbf{d}_p, c|p, \Theta^t) =& P(\mathbf{d}_c, \mathbf{d}_{p-c}, c|p, \Theta^t) = P(\mathbf{d}_{p-c}|p, \Theta^t) P(\mathbf{d}_c, c|p, \Theta^t) \\ =& P(\mathbf{d}_{p-c}|p, \Theta^t) P(\mathbf{d}_c|c, \Theta^t) P(c|p, \Theta^t), \end{aligned}$$

where $\mathbf{d}_c$ is the observed evidence below c and $\mathbf{d}_{p-c}$ is the evidence below p which is not below c. We observe that each of the probabilities in (3) can be easily computed as part of the Pearls message passing algorithm [Pearl, 1988], either as the posterior probabilities or from the appropriate λ messages and λ values (for details see [Pearl, 1988, Neapolitan, 2003]).

The M-step

The M-step of the algorithm determines new parameter values that maximize the $\mathcal{Q}$ function. The terms of the expression (2) can be maximized separately with respect to one of the model parameters:

$$\mathbb{E}(A_e(\mathbf{y})) \log(p_e) + \mathbb{E}(A_{\neg e}(\mathbf{y})) \log(1 - p_e), \text{ for each type } e.$$

For a given e, we find p_e^* for which the derivative is equal to 0:

$$\begin{aligned} \frac{\partial \mathcal{Q}}{\partial p_e} &= \frac{\mathbb{E}(A_e)}{p_e^*} - \frac{\mathbb{E}(A_{\neg e})}{1 - p_e^*} = 0 \\ p_e^* &= \frac{\mathbb{E}(A_e)}{\mathbb{E}(A_e) + \mathbb{E}(A_{\neg e})}. \end{aligned}$$

Notice that $\frac{\partial \mathcal{Q}}{\partial p_e} > 0$ for $p_e < p_e^*$ and $\frac{\partial \mathcal{Q}}{\partial p_e} < 0$ for $p_e > p_e^*$. Thus p_e^* is the optimal value and is selected as p_e in the new set of parameters Θ^{t+1}.

References

[Neapolitan, 2003] Neapolitan, R. E. (2003). *Learning Bayesian Networks.* Prentice Hall.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann.