Patent mining

Antonin Bergeaud, Yoann Potiron and Juste Raimbault

April 19, 2016

We define N the number of patents in the database. We let $t_1, \cdots, t_N$ their corresponding submission date by increasing-time order. For each patent $i = 1, \cdots, N$, we consider C_i the number of citations made by the ith patent as well as $C_i^{(tec)}$ the number of citations made to patents of its own technological class and $C_i^{(sem)}$ the number of citations made to patents of its own semantic class. Finally, we define respectively $N_i^{(tec)}$ and $N_i^{(sem)}$ the number of patents which belonged to the technological (semantic) class of the ith patent when it was submitted.

Let $z \in \{tec, sem\}$. If we assume that the type of citations made $C_i^{(z)}$ are independent of the technological or semantic classes, we can expect $C_i^{(z)}$ to follow a $Bin(C_i, \frac{N_i^{(z)}}{i-1})$ conditioned on C_i and $N_i^{(z)}$. This assumption is not satisfied in practice. Consequently, we define the parameter $0 < \theta^{(z)} < 1$ which indicates the propensity for any patent to cite patents of its own technological or semantic class. We assume that $C_i^{(z)}$ follows a $Bin\big(C_i, \min(1, \frac{N_i^{(z)}}{i-1} + \theta)\big)$.

We assume that the number of citations C_i are a sequence of IID random variables following the discrete distribution C, and also independent of any other quantity. This will facilitate the likelihood analysis in the following. We have

$$\log \mathcal{L}\big(C_N, C_N^{(z)}, \cdots, C_1, C_1^{(z)}\big) = \sum_{i=2}^{N} \log \mathcal{L}\big(C_i, C_i^{(z)} | C_{i-1}, C_{i-1}^{(z)} \cdots, C_1, C_1^{(z)}\big) + \log \mathcal{L}\big(C_1, C_1^{(z)}\big).$$

Note that in view of our assumptions we have

$$\begin{aligned}
\log \mathcal{L}\big(C_i, C_i^{(z)} | C_{i-1}, C_{i-1}^{(z)} \cdots, C_1, C_1^{(z)}\big) &= \log \mathcal{L}\big(C_i, C_i^{(z)} | N_i^{(z)}\big) \\
&= \log \mathcal{L}\big(C_i^{(z)} | C_i, N_i^{(z)}\big) + \log \mathcal{L}\big(C_i | N_i^{(z)}\big) \\
&= \log \mathcal{L}\big(C_i^{(z)} | C_i, N_i^{(z)}\big) + \log \mathcal{L}\big(C_i\big)
\end{aligned}$$

Thus, the log likelihood is equal to

$$\log \mathcal{L}\big(C_N, C_N^{(z)}, \cdots, C_1, C_1^{(z)}\big) = \sum_{i=1}^{N} \log \mathcal{L}\big(C_i^{(z)} | C_i, N_i^{(z)}\big) + \log \mathcal{L}\big(C_i\big). \tag{1}$$

Note that on the right-hand side in (1), the left term depends on the parameter $\theta^{(z)}$ whereas the right term doesn't depend on $\theta^{(z)}$. Thus we can remove it for the maximization procedure.