Covariance asymptotics

I Markov Independent Model

Let M a Markov source and n an integer. The Markov Independent Model (MI) considers n infinite words generated by M. The choice of the starting symbol of each sequence is a parameter of the model. For example, all the sequences might start with the same letter of the alphabet $c \in \mathcal{A}$. Or the first symbol could be initialized using the stationary distribution of the Markov chain related to M.

This is an example for $n = 4$. These are the sequences :

$$\begin{aligned} \mathrm{X}(1) &= 00000\ldots \\ \mathrm{X}(2) &= 1010101\ldots \\ \mathrm{X}(3) &= 1001101\ldots \\ \mathrm{X}(4) &= 001100111\ldots \end{aligned}$$

These sequences are used to build a digital search tree by considering the shortest prefix of each sequence that has not appeared yet in the previously considered sequences.

On our example, it yields the parsed word :

$$()(0)(1)(10)(00)$$

which can be read as the DST :

INSERT TREE

II Tail symbols

1 Definition

Each of the n sequences possess a tail symbol. For each sequence, its tail symbol is the character that immediately follows the prefix phrase inserted into the DST. Therefore the tail symbol is a specific character of this sequence. If we only have the DST containing the prefix phrases, we cannot recover the tail symbols.

Visually, with the prefix phrases in red and the tail symbol in green :

$$\begin{aligned} \mathrm{X}(1) &= 0000000\ldots \\ \mathrm{X}(2) &= 1010101\ldots \\ \mathrm{X}(3) &= 1001101\ldots \\ \mathrm{X}(4) &= 001100111\ldots \end{aligned}$$

Let c be a character from our alphabet $\{a, b\}$. In the case when all the sequences start with a c, we define T_n^c the *number of times a is a tail symbol in the experiment.*

2 Recurrence relation

For $n \geqslant 0$, we have :

$$\boxed{\mathrm{T}_{n+1}^c = \delta_a + \widetilde{\mathrm{T}}_{\mathrm{N}_a}^a + \widetilde{\mathrm{T}}_{\mathrm{N}_b}^b}$$

where :

- $\delta_a = \begin{cases} 1 & \text{if } a \text{ is the tail symbol of the first sequence} \\ 0 & \text{else} \end{cases}$
- N_a is the random variable giving *the size of the left subtree which contains phrases whose second letter is a*
- $\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}$ is the number of times a is a tail symbol for the sequences that were used to build the subtree with phrases having a as second symbol.
- T^c_0 for all c by convention.

If we take $\{\mathrm{N}_a = k\}$, then $\{\mathrm{N}_b = n - k\}$, then the count of the tail symbols on the left tree is independent of the one on the right tree : *i.e.* these quantities are conditionnaly independent.

III Total path length

1 Definition

Defining L^c_n as the *total path length of the nodes of the DST that was built with MI model with n sequences starting with letter c.* It is the sum of the lengths of all the prefix phrases.

2 Recurrence relation

There is another recursive stochastic relation for this quantity, which is, for all $n \geqslant 0$:

$$\boxed{\mathrm{L}^c_{n+1} = n + \widetilde{\mathrm{L}}^a_{\mathrm{N}_a} + \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}}$$

with the convention that $\mathrm{L}^c_0 = 0$ for all c.
Same as for the number of tail symbols, this relation is found by considering the DST and its two main subtrees. Except that this time, we count the number of times an edge contributes to the path length. The root with its two nodes contributes as n. The two subtrees contribute respectively for $\widetilde{\mathrm{L}}^a_{\mathrm{N}_a}$ and $\widetilde{\mathrm{L}}^b_{\mathrm{N}_b}$.
It is convenient that these two quantities are conditionnaly independent in the same way as previously seen for the tail symbols.

IV Poisson tranform differential equation

1 Covariance recurrence relation

Using the previous recurrence relations, we have for all $n \geqslant 0$:

$$\mathrm{Cov}(\mathrm{T}^c_{n+1}, \mathrm{L}^c_{n+1}) = \mathrm{Cov}(\delta_a + \widetilde{\mathrm{T}}^a_{\mathrm{N}_a} + \widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, n + \widetilde{\mathrm{L}}^a_{\mathrm{N}_a} + \widetilde{\mathrm{L}}^b_{\mathrm{N}_b})$$

Since the covariance is a bilinear function which is equal to zero if its two terms are independent or if one is constant, we can ignore the term n and expand this quantity into six terms :

$$\begin{aligned} \mathrm{Cov}(\mathrm{T}^c_{n+1}, \mathrm{L}^c_{n+1}) &= \mathrm{Cov}(\delta_a, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) + \mathrm{Cov}(\delta_a, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) \\ &\quad + \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) \end{aligned}$$

Since δ_a is given by the tail symbol of the first sequence, which is independent from the rest of the process : $\mathrm{Cov}(\delta_a, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) = \mathrm{Cov}(\delta_a, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) = 0$

Now, it is not obvious if the pairs $(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b})$ and $(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a})$ are independent or uncorrelated, because the random variable N_a is not fixed. However they are conditionnaly independent, therefore :

$$\begin{aligned}
\mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) &= \sum_{k=0}^{n} \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b} \,|\, \mathrm{N}_a = k)\mathrm{P}(\mathrm{N}_a = k) \\
&= \sum_{k=0}^{n} \mathrm{Cov}(\widetilde{\mathrm{T}}^a_k, \widetilde{\mathrm{L}}^b_{n-k})\mathrm{P}(\mathrm{N}_a = k) \\
&= \sum_{k=0}^{n} 0 \cdot \mathrm{P}(\mathrm{N}_a = k) \\
&= 0
\end{aligned}$$

Samely, $\mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) = 0$. Yielding :

$$\boxed{\mathrm{Cov}(\mathrm{T}^c_{n+1}, \mathrm{L}^c_{n+1}) = \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b})}$$

2 Poisson transform

Defining

$$\boxed{\mathrm{C}_c(z) = \sum_{n\geqslant 0} \mathrm{Cov}(\mathrm{T}^c_n, \mathrm{L}^c_n)\frac{z^n}{n!}\mathrm{e}^{-z}}$$

Computing, with $p = \mathrm{P}(a|c)$ and $q = 1 - p$:

$$\begin{aligned}
\sum_{n\geqslant 0} \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a})\frac{z^n}{n!}\mathrm{e}^{-z} &= \sum_{n\geqslant 0}\sum_{k=0}^{n} \mathrm{P}(\mathrm{N}_a = k)\,\mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a} \,|\, \mathrm{N}_a = k)\frac{z^n}{n!}\mathrm{e}^{-z} \\
&= \sum_{n\geqslant 0}\sum_{k=0}^{n} \tbinom{n}{k} p^k q^{n-k}\,\mathrm{Cov}(\widetilde{\mathrm{T}}^a_k, \widetilde{\mathrm{L}}^a_k)\frac{z^n}{n!}\mathrm{e}^{-z}
\end{aligned}$$

In this case, $\widetilde{\mathrm{T}}^a_k$ and T^a_k as well as $\widetilde{\mathrm{L}}^a_k$ and L^a_k have the same distribution, hence :

$$\begin{aligned}
\sum_{n\geqslant 0} \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a})\frac{z^n}{n!}\mathrm{e}^{-z} &= \sum_{n\geqslant 0}\sum_{k=0}^{n} \tbinom{n}{k} p^k q^{n-k}\,\mathrm{Cov}(\mathrm{T}^a_k, \mathrm{L}^a_k)\frac{z^n}{n!}\mathrm{e}^{-z} \\
&= \sum_{n\geqslant 0}\sum_{k=0}^{n} \left(\frac{(zp)^k}{k!}\,\mathrm{Cov}(\mathrm{T}^a_k, \mathrm{L}^a_k)\mathrm{e}^{-zp}\right)\left(\frac{(zq)^{n-k}}{(n-k)!}\mathrm{e}^{-zq}\right) \\
&= \underbrace{\left(\sum_{n\geqslant 0} \frac{(zp)^n}{n!}\,\mathrm{Cov}(\mathrm{T}^a_n, \mathrm{L}^a_n)\mathrm{e}^{-zp}\right)}_{=\mathrm{C}_a(zp)}\underbrace{\left(\sum_{n\geqslant 0} \frac{(zq)^n}{n!}\mathrm{e}^{-zq}\right)}_{=1} \\
&= \mathrm{C}_a(zp)
\end{aligned}$$

A similar computation gives $\sum_{n\geqslant 0} \mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b})\frac{z^n}{n!}\mathrm{e}^{-z} = \mathrm{C}_b(zq)$, this time conditionning on $\mathrm{P}(\mathrm{N}_b = k) = \binom{n}{k} q^k p^{n-k}$.

From what we've seen, when derivating $\mathrm{C}_c(z)$ we get :

$$\begin{aligned}
\partial_z \mathrm{C}_c(z) &= \sum_{n\geqslant 0} \mathrm{Cov}(\mathrm{T}^c_n, \mathrm{L}^c_n) n \frac{z^{n-1}}{n!} \mathrm{e}^{-z} - \mathrm{C}_c(z) \\
&= \sum_{n\geqslant 0} \mathrm{Cov}(\mathrm{T}^c_{n+1}, \mathrm{L}^c_{n+1}) \frac{z^n}{n!} \mathrm{e}^{-z} - \mathrm{C}_c(z) \\
&= \sum_{n\geqslant 0} \left[\mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) \right] \frac{z^n}{n!} \mathrm{e}^{-z} - \mathrm{C}_c(z) \\
&= \mathrm{C}_a(zp) + \mathrm{C}_b(zq) - \mathrm{C}_c(z)
\end{aligned}$$

Finally the equation for $\mathrm{C}_c(z)$ is :

$$\boxed{\partial_z \mathrm{C}_c(z) + \mathrm{C}_c(z) = \mathrm{C}_a(zp) + \mathrm{C}_b(zq)}$$