PSEUDO-LIKELIHOOD MAXIMIZATION WITH BETA-LEARNING

March 10, 2024

1 Pseudo-likelihood formalism

Here we want to devise a proper algorithm to perform an inverse Potts-model learning, in which the "temperature" belongs to the optimization parameters. To clarify what we mean let's consider the equilibrium ensemble distribution of a Potts model in which the temperature is explicitly present:

$$P(\sigma_1, ..., \sigma_N) = \frac{1}{Z} \exp\left\{ \beta \sum_{i=1}^{N} h_i(\sigma_i) + \beta \sum_{i<j=1}^{N} J_{ij}(\sigma_i, \sigma_j) \right\}.$$

Where the colors variables can assume q different values. Given this joint distribution the standard learning procedure via full likelihood maximization is not feasible, due to the necessity of computing the partition function. However we employ here the approximate method of pseudo-likelihood maximization, that is we deal with the single conditional distributions:

$$P(\sigma_r | \boldsymbol{\sigma}_{-r}) = \frac{\exp\left\{\beta \left[h_r(\sigma_r) + \sum_{i\neq r} J_{ri}(\sigma_r, \sigma_i)\right]\right\}}{\sum_{l=1}^{q} \exp\left\{\beta \left[h_r(l) + \sum_{i\neq r} J_{ri}(l, \sigma_i)\right]\right\}}.$$

Of particular importance is the function $g_r(\sigma_r)$, the minus single site log-pseudo-likelihood:

$$g_r(\sigma_r) = -\ln P(\sigma_r|\boldsymbol{\sigma}_{-r}) = -\beta \left[h_r(\sigma_r) + \sum_{i\neq r} J_{ri}(\sigma_r, \sigma_i) \right] + \ln \left[\sum_{l=1}^{q} \exp\left\{ \beta h_r(l) + \beta \sum_{i\neq r} J_{ri}(l, \sigma_i) \right\} \right].$$

Whose gradient with respect to β reads:

$$\partial_\beta g_r(\sigma_r) = -H_r(\sigma_r) + \langle H_r(\sigma_r) \rangle$$

Consequently, since the total pseudo-likelihood is given by the sum over all site we get:

$$g(\sigma) = -\sum_{r=1}^{N} \ln P\left(\sigma_r | \sigma_{-r}\right), \quad \partial_\beta g = \sum_{r=1}^{N} \left[\langle H_r(\sigma_r) \rangle - H_r(\sigma_r) \right]$$

It must be noticed that the sum over all site of the single site energies H_r is not the total energy of the system. Indeed, the single site pseudo-likelihood contains the sum over all the other sites different from $r$in the coupling contribution. This implies that the summation over all sites takes every coupling contribution twice. The same happens for the average site energy.

For the pseudo-likelihood method to be a good approximation of the actual likelihood it is necessary to have a certain ensemble of copies of the system. If these copies are statistically independent, then the joint probability is given by the product over replicas, and the likelihood is consequently given by a sum. By virtue of the law of large numbers the quantity:

$$g(\sigma) = -\frac{1}{M}\sum_{a=1}^{M}\sum_{r=1}^{N}\ln P\left(\sigma_r^{(a)}|\sigma_{-r}^{(a)}\right)$$

Converges for large sample size M to the actual total likelihood.

2 Continuous time and β-learning

In addition to the independent replicas of the system, which goes from $a = 1, ..., M$, we want to deal with a succession of temporal realization $t = 1, ..., T$. Specifically, we want to consider the case in which these times are not intrinsically discrete, but could be extracted from a continuous time process. Moreover, the precise time instant at which a specific sample is measured is often unknown. Thus, the necessity of treating time as one of the learning parameters emerges. Then, the temperature β is identified with time itself (or with a certain function of time). Unfortunately it is not possible to optimize together both the energy parameters $\{\mathbf{J}, \mathbf{h}\}$ and the temperature, since $\exp(-\beta H)$ is a non-convex function of both these variables. It is then necessary to implement a kind of alternate gradient descent algorithm, which leads to the minimum of the pseudo-likelihood. Before trying to devise such an algorithm, we want to discuss the issue of including "*round-0*" information into the optimization process.

Imagine we have a certain sequence $s = \{\sigma\}$, whose associated probability density is labeled as $p_s(t)$, where time flows continuously. The stochastic nature of this evolution lies in the fact the sequences are randomly muted during evolution. Moreover, we have in mind for instance the case in which a certain population of bacteria may grow or die according to a rate Γ. This evolution is by itself intrinsically random. We can consequently suppose a time evolution for the probability distribution of the sequence s:

$$\dot{p}^{(t)}(s) = -E_s p^{(t)}(s), \;\; p^{(t)}(s) \propto p^{(0)}(s)\exp(-E_s t)$$

Where the initial distribution is given by $p_s(0) = 1/Z_0 \exp(-G_s)$, with G_s being the Potts-model like energy of the considered sequence. An initial learning is performed over the parameters defining this energy. Then, when dealing with the subsequent time evolution, the parameters to be learnt are time $t \to \beta$, and those defining rate Γ_s which we consider to possess a Potts-like structure itself. Finally probability density would read:

$$p^{(\beta)}(s) = \frac{1}{Z} \exp\left[-\beta E_s - G_s\right] \tag{1}$$

In other words this equation expresses the conditional probability of going from a certain initial distribution defined by the set of parameters $\{G\}$ to a certain probability distribution at "*time*" β, through a selection process which is conveyed by the set of parameters $\{E\}$.

Now in general we do not consider a single time, but rather and ensemble of different instants. Each one would have associated a certain conditional probability of going from an initial distribution to another one up to time β. The idea is then the following. In addition to the average over sequences $s = 1, ..., M$, single conditional probability referring to different times will be multiplied together so to yield a joint probability of observing a certain sequences of time samples. The set of parameters $\{E\}$ and $\{G\}$ are supposed to be the same for every time, and are consequently learnt altogether.

We want now to extend the formalism of the first section to the case in which *round-0* information is taken into account. To do so, let's rewrite the expression of the single site conditioned probability:

$$P^{(\beta)}(\sigma_r = \sigma_r^{(\beta)}(s) | \boldsymbol{\sigma}_{-r} = \boldsymbol{\sigma}_{-r}^{(\beta)}(s)) = \frac{\exp\left\{\beta\left[h_r^{(E)}(\sigma_r^{(\beta)}(s)) + \sum_{i\neq r} J_{ri}^{(E)}(\sigma_r^{(\beta)}(s), \sigma_i^{(\beta)}(s))\right] + h_r^{(G)}(\sigma_r^{(\beta)}(s)) +\right.}{\sum_{l=1}^{q} \exp\left\{\beta\left[h_r^{(E)}(l) + \sum_{i\neq r} J_{ri}^{(E)}(l, \sigma_i^{(\beta)}(s))\right] + h_r^{(G)}(l) + \sum\right.}$$

Which directly comes from 1. From the previous expression it is straightforward to derive the single site pseudo-likelihood at time β related to sequence s:

$$\begin{aligned} g_r^{(\beta)}(\sigma_r^{(\beta)}(s)) = -\beta\left[h_r^{(E)}(\sigma_r^{(\beta)}(s)) + \sum_{i\neq r} J_{ri}^{(E)}(\sigma_r^{(\beta)}(s), \sigma_i^{(\beta)}(s))\right] - h_r^{(G)}(\sigma_r^{(\beta)}(s)) - \sum_{i\neq r} J_{ri}^{(G)}(\sigma_r^{(\beta)}(s), \sigma_i^{(\beta)}(s)) \\ + \ln\left[\sum_{l=1}^{q} \exp\left\{\beta\left[h_r^{(E)}(l) + \sum_{i\neq r} J_{ri}^{(E)}(l, \sigma_i^{(\beta)}(s))\right] + h_r^{(G)}(l) + \sum_{i\neq r} J_{ri}^{(G)}(l, \sigma_i^{(\beta)}(s))\right\}\right], \end{aligned} \tag{2}$$

and thus its derivative with respect to β:

$$\begin{aligned} \partial_\beta g_r^{(\beta)}(\sigma_r^{(\beta)}(s)) = -\left[h_r^{(E)}(\sigma_r^{(\beta)}(s)) + \sum_{i\neq r} J_{ri}^{(E)}(\sigma_r^{(\beta)}(s), \sigma_i^{(\beta)}(s))\right] + \\ + \frac{1}{Z_r^{(\beta)}(s)} \sum_{l=1}^{q}\left[h_r^{(E)}(\sigma_r^{(\beta)}(s)) + \sum_{i\neq r} J_{ri}^{(E)}(\sigma_r^{(\beta)}(s), \sigma_i^{(\beta)}(s))\right] \exp\left\{\beta\left[h_r^{(E)}(l) + \sum_{i\neq r} J_{ri}^{(E)}(l, \sigma_i^{(\beta)}(s\right.\right. \end{aligned}$$

The total pseudolikelihood (that a pseudolikelihood is not) would be given by the sum over sequences and different times of 2, weighted by the normalized number of reads of each sequence at that round $w^{(\beta)}(s)$. Finally, total pseudolikelihood would read

$$g(\{\sigma\}) = -\sum_{\beta=1}^{T}\sum_{s=1}^{M}\sum_{r=1}^{N} w^{(\beta)}(s)\left\{\beta\left[h_r^{(E)}(\sigma_r^{(\beta)}(s)) + \sum_{i\neq r} J_{ri}^{(E)}(\sigma_r^{(\beta)}(s),\sigma_i^{(\beta)}(s))\right] + h_r^{(G)}(\sigma_r^{(\beta)}(s)) - \sum_{i\neq r} J_{ri}^{(G}\right.$$
$$\left. + \ln\left[\sum_{l=1}^{q}\exp\left\{\beta\left[h_r^{(E)}(l) + \sum_{i\neq r} J_{ri}^{(E)}(l,\sigma_i^{(\beta)}(s))\right] + h_r^{(G)}(l) + \sum_{i\neq r} J_{ri}^{(G)}(l,\sigma_i^{(\beta)}(s))\right\}\right]\right]\right\} \tag{3}$$

3 Gauge

If the statistics of data sample is not complete, a regularization term has to be added to 3. We choose a l_2 regularization, that is:

$$R\left(\left\{\mathbf{J}^{(E)},\mathbf{h}^{(E)},\mathbf{J}^{(G)},\mathbf{h}^{(G)}\right\}\right) = \lambda_J^{(E)}\sum_{1\leq i<j\leq N}||J_{ij}^{(E)}||^2 + \lambda_h^{(E)}\sum_{i=1}^{N}||h_i^{(E)}||^2 + \lambda_J^{(G)}\sum_{1\leq i<j\leq N}||J_{ij}^{(G)}||^2 + \lambda_h^{(G)}\sum_{i=1}^{N}||$$

This should impose a specific gauge to couplings and fields, depending if the optimization process is either symmetric or asymmetric. In the first case the J's are meant to be symmetric, that is, function 3 is minimized altogether with the condition $J_{ij}(\sigma_i,\sigma_j) = J_{ji}(\sigma_j,\sigma_i)$ for both E and G part. On the other hand, in the asymmetric case each g_r is minimized separately, thus yielding a different estimate for J_{ij} and J_{ji}, implying the necessity to combine them together so to obtain a single estimate of the parameter.

Let's which are the imposed gauge in both cases, starting from the asymmetric. Let's compute the derivative of g_r with respect to couplings and fields of the selection part:

$$\frac{\partial g_r}{\partial h_r^{(E)}(m)} = -\sum_{\beta=1}^{T}\sum_{s=1}^{M} w^{(\beta)}(s)\left[\beta\mathbb{I}\left[\sigma_r^{(\beta)}(s) = m\right] - \beta P\left(\sigma_r = m|\boldsymbol{\sigma}_{-r} = \boldsymbol{\sigma}_{-r}^{(\beta)}(s)\right)\right] + 2\lambda_h^{(E)}h_r^{(E)}(m) = 0,$$

$$\frac{\partial g_r}{\partial J_{ri}^{(E)}(m,n)} = -\sum_{\beta=1}^{T}\sum_{s=1}^{M} w^{(\beta)}(s)\left\{\beta\mathbb{I}\left[\sigma_i^{(\beta)}(s) = n\right]\left[\mathbb{I}\left[\sigma_r^{(\beta)}(s) = m\right] - P\left(\sigma_r = m|\boldsymbol{\sigma}_{-r} = \boldsymbol{\sigma}_{-r}^{(\beta)}(s)\right)\right]\right\} + 2\lambda_J^{(E}$$

Summing over n in the second equation one gets $\sum_{n=1}^{q} J_{ri}^{(E)}(m,n) = \frac{\lambda_h^{(E)}}{\lambda_J^{(E)}}h_r^{(E)}(m)$, while summing over m yields $\sum_{m=1}^{q} J_{ri}^{(E)}(m,n) = 0$. Analogously summing over m in the first equation yields $\sum_{m=1}^{q} h_r^{(E)}(m) = 0$. The same procedure can be applied to the parameters related to the $\{G\}$ part, obtaining:

$$\frac{\partial g_r}{\partial h_r^{(G)}(m)} = -\sum_{\beta=1}^{T}\sum_{s=1}^{M} w^{(\beta)}(s)\left[\mathbb{I}\left[\sigma_r^{(\beta)}(s)=m\right] - P\left(\sigma_r = m|\boldsymbol{\sigma}_{-r} = \boldsymbol{\sigma}_{-r}^{(\beta)}(s)\right)\right] + 2\lambda_h^{(G)} h_r^{(G)}(m) = 0,$$

$$\frac{\partial g_r}{\partial J_{ri}^{(G)}(m,n)} = -\sum_{\beta=1}^{T}\sum_{s=1}^{M} w^{(\beta)}(s)\left\{\mathbb{I}\left[\sigma_i^{(\beta)}(s)=n\right]\left[\mathbb{I}\left[\sigma_r^{(\beta)}(s)=m\right] - P\left(\sigma_r = m|\boldsymbol{\sigma}_{-r} = \boldsymbol{\sigma}_{-r}^{(\beta)}(s)\right)\right]\right\} + 2\lambda_J^{(G)}$$

Which yield exactly the same conditions as above. We can finally conclude that in the asymmetric case parameters fullfil the following gauge conditions:

$$\begin{cases} \sum_{n=1}^{q} J_{ri}^{(E)}(m,n) = \frac{\lambda_h^{(E)}}{\lambda_J^{(E)}} h_r^{(E)}(m) & , \\ \sum_{m=1}^{q} J_{ri}^{(E)}(m,n) = 0 & , \\ \sum_{m=1}^{q} h_r^{(E)}(m) = 0 & , \\ \sum_{n=1}^{q} J_{ri}^{(G)}(m,n) = \frac{\lambda_h^{(G)}}{\lambda_J^{(G)}} h_r^{(G)}(m) & , \\ \sum_{m=1}^{q} J_{ri}^{(G)}(m,n) = 0 & , \\ \sum_{m=1}^{q} h_r^{(G)}(m) = 0 & . \end{cases}$$

In the symmetric case the function to be derived is the total pseudolikelihood 3 and moreover it holds $J_{ij}(\sigma_i,\sigma_j) = J_{ji}(\sigma_j,\sigma_i)$. The only difference with the previous case lies in the coupling derivative:

$$\frac{\partial g}{\partial J_{kj}^{(E)}(m,n)} = -\sum_{\beta=1}^{T}\sum_{s=1}^{M} w^{(\beta)}(s)\left\{\beta\left[\mathbb{I}\left[\sigma_k^{(\beta)}(s)=m\right]\mathbb{I}\left[\sigma_j^{(\beta)}(s)=n\right] + \mathbb{I}\left[\sigma_j^{(\beta)}(s)=n\right]\mathbb{I}\left[\sigma_k^{(\beta)}(s)=m\right] - \right.\right.$$
$$\left.\left. -\mathbb{I}\left[\sigma_k^{(\beta)}(s)=m\right] P\left(\sigma_j = n|\boldsymbol{\sigma}_{-j} = \boldsymbol{\sigma}_{-j}^{(\beta)}(s)\right)\right]\right\} + 2\lambda_J^{(E)} J_{ri}^{(E)}(m,n) = 0,$$

which summing either over m or n yields the gauge conditions for the symmetric case:

$$\begin{cases} \sum_{n=1}^{q} J_{ri}^{(E)}(m,n) = \frac{\lambda_h^{(E)}}{\lambda_J^{(E)}} h_r^{(E)}(m) & , \\ \sum_{m=1}^{q} J_{ri}^{(E)}(m,n) = \frac{\lambda_h^{(E)}}{\lambda_J^{(E)}} h_r^{(E)}(n) & , \\ \sum_{m=1}^{q} h_r^{(E)}(m) = 0 & , \\ \sum_{n=1}^{q} J_{ri}^{(G)}(m,n) = \frac{\lambda_h^{(G)}}{\lambda_J^{(G)}} h_r^{(G)}(m) & , \\ \sum_{m=1}^{q} J_{ri}^{(G)}(m,n) = \frac{\lambda_h^{(G)}}{\lambda_J^{(G)}} h_r^{(G)}(n) & , \\ \sum_{m=1}^{q} h_r^{(G)}(m) = 0 & . \end{cases}$$

3.1 Direct observation

ACHTUNG: while parameters inferred from the asymmetric code seems to fullfil precisely the gauge relation obtained previously, the same cannot be said for the symmetric code. Indeed, the actual gauge observed is rather:

$$\begin{cases} \sum_{n=1}^{q} J_{ri}^{(E)}(m,n) = \frac{\lambda_h^{(E)}}{2\lambda_J^{(E)}} h_r^{(E)}(m) & , \\ \sum_{m=1}^{q} J_{ri}^{(E)}(m,n) = \frac{\lambda_h^{(E)}}{2\lambda_J^{(E)}} h_r^{(E)}(n) & , \\ \sum_{m=1}^{q} h_r^{(E)}(m) = 0 & , \\ \sum_{n=1}^{q} J_{ri}^{(G)}(m,n) = \frac{\lambda_h^{(G)}}{2\lambda_J^{(G)}} h_r^{(G)}(m) & , \\ \sum_{m=1}^{q} J_{ri}^{(G)}(m,n) = \frac{\lambda_h^{(G)}}{2\lambda_J^{(G)}} h_r^{(G)}(n) & , \\ \sum_{m=1}^{q} h_r^{(G)}(m) = 0 & . \end{cases}$$

The reason of this modification is unknown...

4 Another approach to PLM

In the previous section there is an uncertainty upon how the probability of observing a certain sequence as a function of time is defined. Indeed, whereas it looks reasonable when a single time instant t is taken into account, the situation becomes quite more fuzzy when a time series is considered. Right now we have to make a decision. Either we take subsequent time as being statistically independent, or we decide to deal with a Markov process, in which the probability of any time instant depends solely on the previous one. Let's begin with the first hypothesis.

4.1 Independent times

In this framework subsequent sampling times are dealt as being statistically independent. This implies that a certain time series $\mathcal{T} = \{t_0, t_1, ..., t_n\}$, with $t_n > t_{n-1} > ... > t_1 > t_0$ has a realization probability which factorizes (e.g. the probability of a certain sequence to appear in subsequent rounds):

$$P_s(t_0, t_1, ..., t_n) = \prod_{k=0}^{n} P_s(t_k)$$

How can we model the probabilty that a certain sequence shows up at time t? We can think it having form:

$$P_s(t) = \frac{1}{Z} \mathrm{e}^{-t\Gamma_s - E_s}$$

Which depends upon two quantities, a certain *selection rate* associated to the specific sequence s, which is gradually more important as the time goes by, and a contribution E_s that does not depend on time and is necessary to heal spurious effects due for instance to an uneven population bias. To learn parameters defining this quantity all rounds are used! On the other hand, in order to compute Γ_s parameters all rounds but zero must considered. This becomes clear if we set $t_0 = 0$, consequently making time zero selection contribution disappear.

A possibility for the choice of the functional form of Γ and E is the following. Both are meant to be Potts-like energy, but whether E includes single site contribution, Γ posseses epistatic ones as well, thus yielding:

$$\Gamma = -\sum_{i<j,i=1}^{N} J_{ij}(\sigma_i, \sigma_j) - \sum_{i=1}^{N} h_i(\sigma_i), \quad E = -\sum_{i=1}^{N} \tilde{h}_i(\sigma_i)$$

In this framework the partition functions associated to each single time instant are $Z(t) = \sum_{\{s\}} \mathrm{e}^{-t\Gamma_s - E_s}$. This approach has the great advantage that single time contributions to pseudo-likelihood are factored out, and thus single sites parameter estimates can be computed summing over both times and sequences.

4.2 Markov-like process

Another possible way is to considered the whole time series as being statistically identified with a Markov process. In this case the joint probability can be written as:

$$P_s(t_0, t_1, ..., t_n) = P_s(t_n|t_{n-1})...P_s(t_1|t_0)P_s(t_0)$$

We might consider a sligthly different quantity, that is the probability conditioned to the initial configuration $P_s(t_1, ..., t_n|t_0) = P_s(t_n|t_{n-1})...P_s(t_2|t_1)P_s(t_1|t_0)$. In a manner similar to what we outlined before, we suppose the single propagator having form:

$$P_s(t_n|t_{n-1}) = \frac{1}{Z}\mathrm{e}^{-(t_n - t_{n-1})\Gamma_s - E_s^{(t_{n-1})}}$$

The whole time sequence can consequently be rewritten as:

$$P_s(t_1, ..., t_n|t_0) = \frac{1}{Z}\mathrm{e}^{-t_n\Gamma_s}\mathrm{e}^{-\sum_{k=0}^{n-1} E_s^{(t_k)}}$$

Where once again we consider that Γ posses both epistatic and single site contributions, whereas $E^{(t)}$ has just the fields one (or does it?).