# Multivariate Hawkes Processes for Large-scale Inference
# APPENDIX

Rémi Lemonnier[1,2] Kevin Scaman[1,3] Argyris Kalogeratos[1]
[1] CMLA – ENS Paris-Saclay, CNRS, Université Paris-Saclay, France
[2] Numberly, 1000Mercis group, Paris, France
[3] Microsoft Research – Inria Joint Center, Palaiseau, France
{lemonnier,scaman,kalogeratos}@cmla.ens-cachan.fr

**Abstract**

This is a document containing material that constitutes the Appendix for the paper entitled as shown above, which has been published in the *31st AAAI Conference on Artificial Intelligence* (2017). The included supplementary material consists of: i) an index with the basic notation used in the paper, ii) a detailed formula of the log-likelihood of our model, iii) algorithmic details regarding our inference procedure, iv) technical proofs.

# Appendix: supplementary material

## A. Index of notations

| Symbol | Description |
|---|---|
| $d$ | number of event types, i.e. dimensions of the multivariate Hawkes process |
| $r$ | rank of the low-dimensional approximation |
| $n$ | number of events of all realizations of the LRHP process |
| $K$ | number of triggering kernels |
| $G = \{\mathcal{V}, \mathcal{E}\}$ | a network of $d$ nodes, node set $\mathcal{V}$ and edge set $\mathcal{E}$ |
| $A$ | network's adjacency matrix |
| $\Delta$ | maximum node degree of $G$ |
| $u, v = 1, ..., d$ | indices on dimensions of the original space |
| $i, j = 1, ..., r$ | indices on dimensions of the low-dimensional embedding |
| $P$ | $d \times r$ event type-to-group projection matrix |
| $N(t) = [N_u(t)]_u$ | $d$-dimensional counting process ($t \geq 0, u = 1, ..., d$) |
| $\lambda_u(t)$ | non-negative occurrence rate for event type $u$ at time $t$ |
| $\mu_u(t)$ | natural occurrence rate for event type $u$ at time $t$ |
| $g_{vu}(\Delta t)$ | kernel function evaluating the affection of $\lambda_u$ due to events of type $v$ at time distance $\Delta t$ |
| $\alpha, \beta$ | parameters of the triggering kernels |
| $\gamma, \delta$ | hyperparameters of the triggering kernels |
| $h = 1, ..., H$ | realizations of the LRHP process ($d$-dimensional) |
| $m = 1, ..., n_h$ | events of the realization $h$, which may belong to any event type |
| $\mathcal{H}^h$ | history of $(t_m^h, u_m^h)_{m=1}^{n_h}$ events of the realization $h$, indicating (time of event, event type) |
| $\mathcal{H}$ | collection of the event histories of all $H$ realizations |
| $\sigma$ | maximum number of event types involved in a realization |
| $B, D$ | tensors with four and five dimensions, respectively, introduced to simplify our inference algorithm |

**Table 1.** Index of main notations.

## B. Formula for the log-likelihood

Following a simple calculation, Eq. 3 and Eq. 4 of the article lead to this analytic expression for the log-likelihood, which was left to the Appendix due to space constraints:

$$\mathcal{L}(P,\mathcal{H};\mu,g) = \sum_{h=1}^{H} \left[ \sum_{m=1}^{n_h} \ln \left( \sum_{i=1}^{r} P_{u_m^h i}\, \tilde{\mu}_i(t_m^h) + \sum_{i,j} \sum_{l:\, t_l^h < t_m^h} P_{u_m^h i}\, P_{u_l^h j}\, A_{u_l^h u_m^h}\, \tilde{g}_{ji}(t_m^h - t_l^h) \right) \right. \\ \left. - \sum_{u,i} P_{ui} \int_{T_-^h}^{T_+^h} \tilde{\mu}_i(s) ds - \sum_{u,v,i,j} P_{ui}\, P_{vj}\, A_{vu} \int_{T_-^h}^{T_+^h} \tilde{g}_{ji}(s - t_m^h) ds \right]. \tag{1}$$

## C. Details on the inference algorithm

***Computing $B$ and $D$ tensors.*** In order for the inference algorithm to be tractable, special attention has to be paid to the computation of $B$ and $D$ tensors. Alg. 2 describes the computation of the sparse tensors $B = (B_{h,u,v,k})$ and $D = (D_{h,m,u,v,k})$.

**Algorithm 2** Construction of $D$ and $B$ tensors

Initialize $j = 0$
**for all** $h$ **do**
  Initialize $(C_v^k = \mathbb{1}_{\{v=d+1\}})_{v\geq 0, k\geq 0}$ ; $t_0^h = T_-^h$ ; $(B'_{h,u,k} = 0)_{u\geq 0, k\geq 0}$
  $B'_{h,d+1,k} \leftarrow \frac{1-\exp(-k\gamma(T_+^h - T_-^h))}{k\gamma}$
  **for all** $m \in [1...n_h]$ **do**
    $dt \leftarrow t_m^h - t_{m-1}^h$
    **for all** $k$,$v$ s.t $C_v^k > 0$ **do**
      $C_v^k \leftarrow C_v^k \exp(-\mathbb{1}_{\{v>0\}}(k+1)\delta dt - \mathbb{1}_{\{v=0\}}\gamma dt)$
    **end for**
    **for all** $k$ **do**
      $D_{h,m,u,v,k} \leftarrow \mathbb{1}_{\{u=u_m\}} \sum_{v\geq 0} A_{u_m v} C_v^k$
      $B'_{h,u_m,k} \leftarrow B'_{h,u_m,k} + \frac{1-\exp(-k\delta(T_+^h - t_m^h))}{k\delta}$
      $C_{u_m}^k \leftarrow C_{u_m}^k + 1$
    **end for**
    $j \leftarrow j + 1$
  **end for**
  $B_{h,u,v,k} \leftarrow A_{uv} B'_{h,v,k}$
**end for**
**return** $B$, $D$

The most expensive operation in this algorithm is the multiplicative update of all $C_v^k$ with the exponential decay $\exp(-(k+\mathbb{1}_{\{v>0\}})\gamma dt)$. Fortunately, this update only has to be performed for every node $v$ that already appeared in the cascade, which are at most $\sigma \leq d$ (by definition). The complexity of this operation is therefore $O(nK\sigma)$. The number of non-zero elements of $D$ and $B$ is $O(nK\min(\Delta, \sigma))$, where $\Delta$ is the maximum number of neighbors of a node in the underlying network $\mathcal{G}$. If $\mathcal{G}$ is sparse, which is usually the case for social networks for instance, then $\Delta \ll d$ and therefore $O(nK\Delta) \ll O(nKd)$.

Thus, storing and computing $B$ and $D$ is tractable for large dense graphs and for particularly large sparse graphs. Note that, since computing the log-likelihood requires the computation of occurrence rates at each event time, which depends on the occurrences of all preceding events, the linear complexity in the number of events is only possible because of the memoryless property of the decomposition over a basis of exponential functions. Otherwise, the respective complexity would have been at least $\Theta(\sum_{h=1}^{H} n_h^2 K\sigma)$, with $\sum_{h=1}^{H} n_h^2 \gg n$.

## D. Proof of Proposition 1

For this proof we will make use of the concept of *auxiliary functions*.

**Definition 1.** *Let $g\colon \mathcal{X}^2 \to \mathcal{R}$ is an auxiliary function for $f\colon \mathcal{X} \to \mathcal{R}$ iff $\forall (x,y) \in \mathcal{X}^2$, $g(x,y) \geq f(x)$ and $\forall x \in \mathcal{X}, g(x,x) = f(x)$.*

The reason why these functions are an important tool for deriving iterative optimization algorithms is given by the following lemma.

**Lemma 1.** *If $g$ is an auxiliary function for $f$, then*

$$f\left(\operatorname*{argmin}_x g(x,y)\right) \leq f(y). \tag{2}$$

*Proof.* Let $z = \operatorname{argmin}_x g(x,y)$. Then

$$f(z) = g(z,z) \leq g(z,y) \leq g(y,y) = f(y).$$

where the first inequality comes from the definition of $g$ and the second from the definition of $z$. □

Therefore, if an auxiliary function $g$ is available, constructing the sequence $y_{t+1} = \operatorname{argmin}_x g(x, y_t)$ that verifies $f(y_{t+1}) \leq f(y_t)$ for all $t$ constitutes a candidate method for finding the minimum of $f$. In our case, we are able to make use of the following result.

**Lemma 2.** *Let $f(p) = -\sum_{k=1}^{K} \ln\left(p^\top \Xi^k p\right) + p^\top \Psi p$ where $p \in \mathbb{R}_+^K$, $\Xi^1$, ..., $\Xi^K$ are positive symmetric matrices and $\Psi$ is a symmetric matrix, then an auxiliary function for $f$ is the following:*

$$g(p,q) = -\sum_{k=1}^{K} \left( \frac{2q^\top \Xi^k [q \ln(p/q)]}{q^\top \Xi^k q} + \ln\left(q^\top \Xi^k q\right) \right) + q^\top \Psi [p^2/q] \tag{3}$$

In the lemma above, the vectors $[q\ln(p/q)]$ and $[p^2/q]$ are to be understood as coordinate-wise operations, i.e. $(q_i \ln(p_i/q_i))_i$ and $(p_i^2/q_i)_i$.

*Proof.* It is clear that $g(p,p) = f(p)$ so the proof reduces to showing that $g(p,q) \geq f(p)$. Let $k \leq K$. By concavity of the logarithm function, we have for every weight matrix $(\alpha_{ij})_{ij}$ such that $\sum_{i,j} \alpha_{ij} = 1$,

$$\ln\left(p^\top \Xi^k p\right) \geq \sum_{i,j} \alpha_{ij} \ln\left(\frac{p_i \Xi_{ij}^k p_j}{\alpha_{ij}}\right).$$

Note that the right-hand side term of the equation is well-defined because of the positivity constraint imposed on each $\Xi_{ij}^k$. By choosing $\alpha_{ij} = q_i \Xi_{ij}^k q_j / q^\top \Xi^k q$, and using the symmetry of $\Xi^k$, we get:

$$\ln\left(p^\top \Xi^k p\right) \geq \frac{2q^\top \Xi^k [q \ln(p/q)]}{q^\top \Xi^k q} + \ln\left(q^\top \Xi^k q\right).$$

For the right-hand side of the above equation, we use the fact that for every $i, j$ it holds:

$$p_i p_j \leq \frac{p_i^2 q_j}{2q_i} + \frac{p_j^2 q_i}{2q_j},$$

and the symmetry of $\Psi$, in order to conclude that $p^\top \Psi p \leq q^\top \Psi [p^2/q]$. □

Using Lemma 2, we are now in position to prove Proposition 1 by showing that the proposed update $p^{t+1}$ is indeed the global minimum of $g(p, p^t)$. $g$ being the sum of univariate convex functions of the $p_i$, it is sufficient to show that for every $i$, the partial derivative of $g(p, p^t)$ with respect to $p_i$ vanishes in $p_i^{t+1}$. We therefore need:

$$-\sum_k \frac{p_i^t (\Xi^k p^t)_i}{p_i^{t+1} {p^t}^\top \Xi^k p^t} + \frac{p_i^{t+1} (\Psi p^t)_i}{p_i^t} = 0,$$

which only positive solution is given by:

$$p_i^{t+1} = p_i^t \left( \sum_k \frac{(\Xi^k p^t)_i}{{p^t}^\top \Xi^k p^t (\Psi p^t)_i} \right)^{1/2}. \tag{4}$$

Finally, if $p$ is a stable fixed point of Eq. 4, then, by definition, there exists $\epsilon > 0$ such that, $\forall p'$ s.t. $||p - p'||_2 \leq \epsilon$, the iterative algorithm starting at $p^0 = p'$ converges to $p$. However, since $f$ is continuous, a simple iteration of the inequality of Lemma 1 implies that $f(p') \geq f(p^1) \geq ... \geq \lim_{t \to +\infty} f(p^t) = f(p)$, and $p$ is a local minimum of $f$.