The core computational problem in the use of point processes for statistical modeling is the optimization of the minus-log-likelihood function, which is given as

$$ l(\theta) = \sum_{j=1}^{n} \log \lambda_\theta(t_j) - \int_{0}^{T} \lambda_\theta(s)ds $$

where $0 < t_1 < t_2 < \ldots < t_n < T$ are observations and $\lambda_\theta$ is a parameterized family of intensities. Typically $\theta \in \Theta \subseteq \mathbb{R}^p$. For the Hawkes family of generalized linear point process models in ppstat we consider situations where

$$ \lambda_\theta(t) = \varphi \left( \alpha^T X(t) + \sum_{m=1}^{K} \sum_{i=1}^{n(m)} h_{\beta m}^m(t - s_{i}^m) \right) $$

where $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is a given function,

$$ \theta = \begin{pmatrix} \alpha \\ \beta^1 \\ \vdots \\ \beta^K \end{pmatrix} $$

and $s_{i1}^m < \ldots < s_{in(m)}^m$ for $m = 1, \ldots, K$ are observations of point processes (one of these sets of points could be the $t_i$ observations above). The process $X(t)$ is an auxiliary, $d(0)$-dimensional observed processes – observed at least discretely. The processes

$$ \sum_{i=1}^{n(m)} h_{\beta m}^m(t - s_{i}^m) $$

are linear filters using the (parameterized) filter function $h_{\beta m}^m$, which are given via a basis expansion

$$ h_{\beta m}^m(t) = (\beta^m)^T B(t) = \sum_{l=1}^{d(m)} \beta_{l}^m B_l(t), $$

and $\beta^m \in \mathbb{R}^{d(m)}$. Collecting these ingredients – and interchanging two sums – the intensity function can be written as

$$ \lambda_\theta(t) = \varphi \left( \alpha^T X(t) + \sum_{m=1}^{K} \sum_{i=1}^{d(m)} \beta_{l}^m \sum_{l=1}^{n(m)} B_l(t - s_{i}^m) \right) = \varphi \left( \theta^T Z(t) \right) $$
with $Z(t)$ a process of dimension $p = d(0) + d(1) + \ldots + d(K)$. Each of the linear filter components

$$\sum_{i=1}^{n(m)} B_i(t - s_i^{(m)})$$

are computable from the observations and the fixed choice of basis. The minus-log-likelihood function that we want to minimize reads

$$l(\theta) = \int_0^T \varphi(\theta^T Z(s)) \, ds - \sum_{j=1}^n \log \varphi(\theta^T Z(t_j)).$$

The integral is not in general analytically computable. We discretize time to have a total of $N$ time points and let $Z$ denote the $N \times p$ matrix of the $Z(t)$-process values at the discretization points. With $\Delta$ the $N$-dimensional vector of interdistances from the discretization we arrive at the approximation of the minus-log-likelihood function that we seek to minimize:

$$l(\theta) \simeq \Delta^T \varphi(\theta^T Z) - \sum_{j=1}^n \log \varphi(\theta^T Z(t_j)).$$

We have used the convention that $\varphi$ applied to a vector means coordinate-wise applications of $\varphi$. Using this expression a precomputation of the $Z$ matrix will allow for a rapid computation of (the approximation to) $l$. The derivatives are likewise approximated as

$$Dl(\theta) \simeq [\Delta \circ \varphi'(\theta^T Z)]^T Z - \sum_{j=1}^n \frac{\varphi'(\theta^T Z(t_j))}{\varphi(\theta^T Z(t_j))} Z(t_j)^T,$$

with $\circ$ the Hadamard (or coordinate-wise) matrix product and

$$D^2l(\theta) \simeq Z^T [\Delta \circ \varphi''(\theta^T Z) \circ Z] - \sum_{j=1}^n \frac{\varphi''(\theta^T Z(t_j)) \varphi(\theta^T Z(t_j)) - \varphi'(\theta^T Z(t_j))^2}{\varphi(\theta^T Z(t_j))^2} Z(t_j) Z(t_j)^T.$$