Mean-field inference of Hawkes point processes

Emmanuel Bacry\(^1\), Stéphane Gaïffas\(^1\), Iacopo Mastromatteo\(^1,2\) and Jean-François Muzy\(^1,3\)

\(^1\)Centre de Mathématiques Appliquées, École Polytechnique and CNRS UMR 7641, F-91128 Palaiseau, France
\(^2\)Capital Fund Management, 21-23 Rue de l’Université, F-75007 Paris, France
\(^3\)Laboratoire Sciences Pour l’Environnement, CNRS, Université de Corse, UMR 6134, F-20250 Corté, France

E-mail: Iacopo.Mastromatteo@cfm.fr

Received 1 November 2015, revised 6 January 2016
Accepted for publication 11 February 2016
Published 22 March 2016

Abstract
We propose a fast and efficient estimation method that is able to accurately recover the parameters of a \(d\)-dimensional Hawkes point-process from a set of observations. We exploit a mean-field approximation that is valid when the fluctuations of the stochastic intensity are small. We show that this is notably the case in situations when interactions are sufficiently weak, when the dimension of the system is high or when the fluctuations are self-averaging due to the large number of past events they involve. In such a regime the estimation of a Hawkes process can be mapped on a least-squares problem for which we provide an analytic solution. Though this estimator is biased, we show that its precision can be comparable to the one of the maximum likelihood estimator while its computation speed is shown to be improved considerably. We give a theoretical control on the accuracy of our new approach and illustrate its efficiency using synthetic datasets, in order to assess the statistical estimation error of the parameters.

Keywords: Hawkes process, high-dimensional inference, mean-field theory

(Some figures may appear in colour only in the online journal)

1. Introduction
The use of point processes, in particular Hawkes processes \([1, 2]\) is ubiquitous in many fields of applications. Such applications include, among others, geophysics \([3]\), high frequency finance \([4–6]\), neuroscience \([7]\), predictive policing \([8]\) and social networks dynamics \([9–15]\).
A possible explanation for the success of this model is certainly its simplicity yet ability to account for several real-world phenomena, such as self-excitement, coexistence of exogenous and endogeneous factors or power-law distributions [16].

Various estimation methods of the Hawkes process parameters have been proposed in both parametric and non-parametric situations. The most commonly used is to maximize the likelihood function that can be written explicitly [17] while alternative linear methods such as the contrast function (CF) minimization [7], spectral methods [18] or estimation through the resolution of a Wiener–Hopf system [19, 20] have been proposed.

In many of the above mentioned applications, especially in the field of network activities, viral propagation or community detection one has to handle systems of very large dimensions for which these estimation methods can be heavy to implement. Motivated by the goal to devise a fast and simple estimation procedure, we introduce an alternative approach that is inspired by recent results [21] justifying a mean-field (MF) approximation of a Hawkes process that supposes small fluctuations of the stochastic intensities with respect to their mean values. This allows us to solve the estimation problem by replacing the original objective (log-likelihood) by a quadratic problem, which is much easier to optimize (see section 3 below). Note that this quadratic problem differs from the usual least-squares objective for counting processes [7] (see section 4 below for a precise comparison).

We show that in a wide number of situations this new framework allows a much faster estimation without inducing losses in precision. Indeed, we show that its bias can be negligible and its accuracy as good as the maximum likelihood provided the level of endogeneity of the process is sufficiently weak or the interactions are sufficiently ‘self-averaging’ meaning that they involve a large number of events over past times or over components of the system (i.e., in the large dimensional setting). Besides theoretical arguments, we give numerical illustrations of the fact that this approach leads to an improvement (that can be of several orders of magnitude) of the classical ones based on state-of-the-art convex solvers for the log-likelihood.

The organization of the paper is the following: in section 2 we formally introduce the Hawkes process and define the main notations of the paper. The MF inference method is defined in section 3. We show how this method is naturally obtained using a Baysian approach in a regime where the fluctuations of the stochastic intensity are very small. We conduct a theoretical analysis of the domain of validity of the method by comparing its accuracy to the maximum likelihood estimation (MLE). This notably allows us to provide a quantitative measure of what weakly endogenous or ‘self-averaging’ means for the interactions. In section 4 we describe the implementation of the algorithm and compare it with other state-of-the-art algorithms. Section 5 shows the numerical results that we obtain with this method on synthetic data. Section 6 is devoted to concluding remarks and to the discussion of the possible extensions of our method, particularly as far as penalization issues are concerned. The more technical parts of the discussion are relegated to the appendices.

2. The Hawkes process

2.1. Definition

Let us consider a network of \( d \) nodes, in which one observes a set of events encoded as a sequence \( \{(t_{mn}, u_m)\}_{m=1}^{\infty} \), where \( t_{mn} \) labels the time of the event number \( m \) and \( u_m \in \{1, \ldots, d\} \) denotes its corresponding node. Then we can define a set of counting functions \( \{N_t^d\}_{t \geq 0} \) as \( N_t^d = \sum_{m=1}^{\infty} \delta^{u_m} \mathbf{1}_{t_{mn} \leq t} \), where \( \delta^d \) indicates the Kronecker delta. These counting functions can be associated to a vector of stochastic intensities \( \lambda_t = [\lambda_t^1, \ldots, \lambda_t^d] \).
defined as
\[ \lambda_i^t = \lim_{dt \to 0} \frac{\mathbb{P}(N_{i+dt}^t - N_i^t = 1|\mathcal{F}_t)}{dt}, \]
where the filtration \( \mathcal{F}_t \) encodes the information available up to time \( t \). Then the process \( N_t \) is called a Hawkes process if the stochastic intensities can be written as
\[ \lambda_i^t = \mu^t + \sum_{j=1}^d \int_0^t dN_j^t \Phi^j(t - t'), \]
where \( \Phi(t - t') = [\Phi^j(t - t')_{1 \leq i, j \leq d}] \) is a component-wise positive, causal (i.e., whose support is in \( \mathbb{R}^+ \)), locally \( L^1 \)-integrable matrix kernel representing (after proper normalization) the probability for an event of type \( j \) occurring at time \( t' \) to trigger an event of type \( i \) at time \( t \), while \( \mu = [\mu^j]_{j=1} \) is a vector of positive exogenous intensities (see [22] for a more rigorous definition). It is well known that a sufficient condition for the intensity processes \( \lambda_i^t \) to be stationary (i.e., for the processes \( N_i^t \) to have stationary increments) is the so-called stability condition \( (SC) \) :
\[ ||\Phi|| < 1, \]
where \( || \cdot || \) stands for the spectral norm of the \( d \times d \) matrix \( \left[ \int_0^{+\infty} \Phi^i(t) dt \right]_{1 \leq i, j \leq d} \) made of the \( L^1 \) norms of each kernel \( \Phi^j(t) \).
In the following discussion, we will restrict our attention to stable Hawkes processes (in the sense of (SC)) for which the matrix \( \Phi(t) \) can be written as
\[ \Phi^j(t) = \sum_{q=1}^p \alpha_{ij}^q g_q(t) \]
where we have introduced a set of \( p > 0 \) known basis kernels \( g_q(t) \) satisfying
\[ \int_0^{+\infty} dt \ g_q(t) = 1. \]

2.2. Notations
For the sake of conciseness, it will be useful to preliminary introduce some notation.

- We will use the integer indexes \( a \in [1, ..., pd] \) in order to identify pairs \((i, a) \in \{1, ..., d\} \times \{1, ..., p\}\). Consequently, summing over \( a \) allows to run both over the \( d \) components and the \( p \) basis functions of the kernels. While the set of indices \( i, j, k, ... \) will be employed to label single nodes, the notation \( a, b, c, ... \) will be used to label a pair (node/kernel). In that respect we set
  \[ - \alpha_{ia} \equiv \alpha_{ab}, \]
  \[ - g^a(t - t') \equiv g_a(t - t') \] and
  \[ - dN_i^a \equiv dN_{ih}. \]
- For convenience we will also define the deterministic process \( N_i^0 = t \), which will be associated with the kernel \( g^0(t - t') = \delta(t - t' - dt) \) equal to a Dirac delta function shifted by an infinitesimal amount \( dt > 0 \).
- According to previous notations, equation (2) can be compactly rewritten as
  \[ \lambda_i^t = \sum_{a=0}^p \theta_{ia}^t \int_0^t dN_i^a g^a(t - t'), \]
where the parameters \( \theta_i \) refer to the Hawkes parameters, namely
\[
\theta^0 = \mu_i \quad \text{and} \quad \theta^a_i = \alpha^{ia}, \quad \forall \ a > 0.
\]

- We will adopt the notation \( x \) for scalars, \( y = [y^a]_{0 \leq a < dp} \) for vectors and \( Z = [Z_{ab}]_{0 \leq a, b < dp} \) for square matrices. Correspondingly, we define the usual matrix-vector composition by \( Z y = [\sum_{b=0}^{dp} Z_{ab} y^b]_{0 \leq a < dp} \) and the matrix-matrix product as \( W Z = [\sum_{b=0}^{dp} W_{ab} Z_{bc}]_{0 \leq a, b < dp} \).
- If \( y \) is a vector \( \|y\| \) will refer to the \( L^2 \) norm of \( y \) whereas if \( Z \) is a matrix, \( \|Z\| \) will refer to its spectral norm. Moreover if \( \Phi(t) = [\Phi^i(t)]_{0 \leq i,j < d} \) is a matrix of functions, \( \|\Phi\| \) will refer to the spectral norm of the matrix \( \left[ \int_0^{\infty} dr \|\Phi^i(t)\|_{i,j} \right] \).
- Hereafter, we will have to handle collections of \( d \) scalars, vectors or matrices indexed by \( i \in \{1,...,d\} \), that we will write as \( \mathbf{x}^i \) (for scalars), \( \mathbf{y}^i = [y^a]_{0 \leq a < dp} \) (for vectors) and \( \mathbf{Z}^i = [Z_{ab}]_{0 \leq a, b < dp} \) (for matrices).

According to these conventions, The goal of this paper is to present a MF framework for estimation of the \( d \) vectors of parameters \( \theta^i = [\theta^a_i]_{0 \leq a < dp} \) given a set of observations \( N_t \).

2.3. The likelihood function

The probability for a Hawkes process parametrized by \( \theta \) to generate a trajectory \( N_t \) in a time interval \( t \in [0, T] \) has been first computed in [17], and is given by
\[
P(\mathbf{N}|\theta) = \exp \left\{ -\sum_{i=1}^{d} \int_0^T \sum_{j} \lambda_i^j \lambda_{i}^{a} \right\}, \tag{6}
\]
which implicitly depends on \( \theta \) through the stochastic intensities \( \lambda_i^j \). We define the negative log-likelihood
\[
\mathcal{L}(\mathbf{N}, \theta) = -\log P(\mathbf{N}|\theta). \tag{7}
\]

Note that the maximum-likelihood estimator (that we will compare to our MF estimator to throughout this paper)
\[
\theta_{\text{MLE}} = \arg\min_{\theta} \mathcal{L}(\mathbf{N}, \theta), \tag{8}
\]
is the most commonly used estimator employed in the literature [17].

3. Inference and MF approximation

3.1. A Bayesian approach

From a Bayesian standpoint, the probability for an observed sample \( N_t \) to be generated by a Hawkes process parametrized by \( \theta \) is given by the posterior distribution
\[
P(\theta|\mathbf{N}) = \frac{P(\mathbf{N}|\theta)P(\theta)}{P(\mathbf{N})}, \tag{9}
\]

\footnote{Note that, in order for the parameters \( \theta^a_i \) to be dimensionally homogeneous, we need to assume a dimensionless \( \mu \), while \( g(t-t') \) is taken to be of dimension \( \text{[t]}^{-1} \). Moreover, we implicitly assume \( N^i_t = \omega t = t \) to be dimensionless through a suitable choice of a unit (i.e., we take \( \omega = 1 \)).}
where $P_{0}(\theta)$ is a prior which we assume to be flat (see section 6 below for other choices). Hence, we will be interested in averaging the inferred couplings over the posterior, so to compute their averages and their covariances

$$E\theta^{ia} = \int d\theta P(\theta|N_{i}) \theta^{ia},$$  \hspace{1cm} \text{(10)}$$

$$C(\theta^{ia}, \theta^{jb}) = \int d\theta P(\theta|N_{i}) \theta^{ia}\theta^{jb} - E\theta^{ia}E\theta^{jb},$$  \hspace{1cm} \text{(11)}$$

where we are writing $\int d\theta = \int_{\mathbb{R}^{(d-1)p+1}} \prod_{1 \leq i \leq d, 0 \leq a \leq dp} d\theta^{ia}$. This can be technically done by introducing the partition function $Z(s)$ defined as the Laplace transform of the posterior

$$Z(s) = \int d\theta P(\theta|N_{i}) e^{-T \sum_{i=1}^{d} s^{i} \theta^{i}},$$  \hspace{1cm} \text{(12)}$$

where $s = \{s^{i}\}_{1 \leq i \leq d}$ is a collection of $d$ Laplace-parameter vectors $s^{i} = [s^{ia}]_{0 \leq a \leq dp}$. Its computation allows one to obtain equations (10) and (11) by differentiating the free-energy density $f(s) = -\log Z(s)/T$, so that

$$E\theta^{ia} = \partial_{s^{ia}} f(s)|_{s=0},$$  \hspace{1cm} \text{(13)}$$

$$C(\theta^{ia}, \theta^{jb}) = -\frac{1}{T} \partial_{s^{ia}} \partial_{s^{jb}} f(s)|_{s=0}.\hspace{1cm} \text{(14)}$$

We will be interested in finding an approximated expression for the free-energy density $f(s)$ allowing to compute efficiently equations (10) and (11).

**Remark.** Let us point out that, since we are interested in derivatives of the free energy, we can drop additive $s$-independent terms in its computation. This will be done implicitly all along the computation.

Let us first use equation (6) so to write explicitly the partition function as

$$Z(s) = \int d\theta e^{-T \sum_{i=1}^{d} (s^{i} + h^{i} \theta^{i}) + \sum_{m=1}^{n} \log \lambda_{m}^{i}},$$  \hspace{1cm} \text{(15)}$$

where we have dropped multiplicative $s$-independent terms, and introduced auxiliary vector $h = [h^{i}]_{0 \leq a \leq dp}$ equal to

$$h^{a} = \frac{1}{T} \int_{0}^{T} df \int_{0}^{T} dN_{t}^{a} S^{a}(t - t').$$  \hspace{1cm} \text{(16)}$$

Note that in the limit of large sample size $T \to \infty$ and in the vicinity of $s = 0$, the integral is dominated by the maximum of the exponential corresponding to the maximum likelihood estimator $\theta_{MLE}$ (8) which is commonly employed in the literature relating to parametric inference of Hawkes processes [17].

Let us also point out that $Z(s)$ can be rewritten as:

$$Z(s) = \int d\theta e^{-T (s^{i} + h^{i} \theta^{i}) + \int_{0}^{T} dN_{t}^{i} \log \lambda_{t}^{i}} = \prod_{i=1}^{d} \int d\theta^{i} e^{-T (s^{i} + h^{i} \theta^{i}) + \int_{0}^{T} dN_{t}^{i} \log \lambda_{t}^{i}}$$  \hspace{1cm} \text{(17)}$$

and therefore the partition function is a product of independent partition functions associated with each node $i \in \{1, \ldots, d\}$. This means that the inference problem factorizes in single node problems associated with each vector $\theta^{i}$. 

5
3.2. MF approximation

Our strategy in order to approximate equation (15) is to suppose the fluctuations of the \( \lambda_i^t \) to be small with respect to their empirical averages

\[
\Lambda = \{ \bar{\lambda}_i \}_{1 \leq i \leq d} = \left\{ \frac{N_i^t}{T} \right\}_{1 \leq i \leq d}.
\]

(18)

This allows us to approximate the partition function appearing in equation (15) by a quadratic expansion of the logarithmic functions around the empirical averages \( \bar{\lambda} \). Let us first suppose that the Hawkes process is stable in the sense of (SC) (see (3)). We have seen that it means that the processes \( \{ \lambda_i^t \}_{1 \leq i \leq d} \) are stationary. The MF expansion basically assumes that each process \( \lambda_i^t \) has small fluctuations around its empirical averaged value \( \bar{\lambda}_i \). More precisely, let us suppose that the following mean-field hypothesis holds

\[
(\text{MFH}) : \quad r_i = \left| \frac{\lambda_i^t - \bar{\lambda}_i}{\bar{\lambda}_i} \right| \ll 1,
\]

(19)

then

\[
\log \lambda_i^t \approx \log \bar{\lambda}_i + \frac{\lambda_i^t - \bar{\lambda}_i}{\bar{\lambda}_i} - \frac{(\lambda_i^t - \bar{\lambda}_i)^2}{2(\bar{\lambda}_i)^2}.
\]

(20)

In order to get \( Z(s) \) (see (17)), we need to compute \( \int_0^T dN_i^t \log \lambda_i^t \) for all \( i = 1,\ldots,d \). Substituting (20) in the expression of (17), we are left with three terms corresponding to the constant, linear and quadratic contribution in \( \lambda_i^t \). More precisely, we have:

- The constant term

\[
\int_0^T dN_i^t \log \bar{\lambda}_i = T \sum_{a=0}^{d} \bar{g}_{aa} \bar{k}_{aa} = 0,
\]

is independent of both \( \theta \) and \( s \), thus it can be dropped for the rest of this section.

- The linear term, after dropping \( \theta \) and \( s \)-independent terms, leads to

\[
\int_0^T dN_i^t \left\{ \frac{\lambda_i^t - \bar{\lambda}_i}{\bar{\lambda}_i} \right\} = T \sum_{a=0}^{d} \bar{g}_{aa} \bar{k}_{aa},
\]

where we have introduced a collection of \( d \) vectors \( k^i = [k_{aa}^i]_{0 \leq a \leq dp} \) defined as:

\[
k_{aa}^i = \frac{1}{N_i^t} \int_{t'=0}^t dt' \int_{0 \leq t'=0}^{d} \bar{g}_{aa} \bar{k}_{aa} \, (t - t')
\]

(21)

- The quadratic term (after straightforward computations, dropping again \( \theta \) and \( s \)-independent terms) leads to

\[
-\frac{1}{2} \int_0^T dN_i^t \left( \frac{\lambda_i^t - \bar{\lambda}_i}{\bar{\lambda}_i} \right)^2 = -\frac{T}{2} \sum_{a,b=0}^{dp} \bar{g}_{aa} \bar{g}_{bb} J_{ab} \quad \text{and} \quad \frac{T}{2} \sum_{a=0}^{d} \bar{g}_{aa} \bar{k}_{aa},
\]

where we have introduced a collection of \( d \) matrices \( J^i = [j_{ab}^i]_{0 \leq a,b \leq dp} \) defined as:

\[
j_{aa}^i = \frac{1}{N_i^t} \int_{t'=0}^t dt' \int_{0 \leq t'=0}^{d} \bar{g}_{aa} \bar{k}_{aa} \, (t - t')
\]

(21)

5 Had we introduced \( \bar{\lambda} \) as a variational parameter to estimate self-consistently, this term would have contributed to fixing it to its optimal value. This is non-necessary in this case, since we are implicitly forcing the empirical value of \( \bar{\lambda} \) to be the MF parameter.
It then results from (17) that, up to a constant factor

$$Z(s) \simeq \int d\theta \ e^{-\sum_{i=1}^{d} (s^i + h^i \theta^i + \frac{1}{2} \theta^i \theta^i)}.$$  

(23)

The integral appearing in equation (23) can be evaluated analytically as a simple Gaussian integral\(^6\), so that one can write the free-energy density as

$$f(s) = -\frac{1}{2} \sum_{i=1}^{d} [(2k^i - h^i - s^i)\mathbf{C}^{-1}(2k^i - h^i - s^i)],$$

(24)

where \(\mathbf{C}^{-1}\) denotes the inverse of matrix \(\mathbf{J}\). Under this approximation, one obtains

$$\mathbb{E}[\theta^{ia}] \simeq \sum_{b=0}^{d} C_{iab} (2k^b - h^b) \equiv \theta_{\text{MF}}^{ia},$$

(25)

$$\mathbb{C}(\theta^{ia}, \theta^{ib}) \simeq \frac{\delta^{ij} C_{iab}}{T}.$$  

(26)

These are the central equations of our paper: equation (25) expresses the definition of our MF estimator while equation (26) expresses the convergence rate to the expected value of the estimator, so that \(\theta - \mathbb{E}[\theta] \sim T^{-1/2}\).

3.3. Validity conditions for the MF approximation

As discussed previously, the MF approximation is likely to be pertinent in the regime described by (MFH) (19), i.e., when the \(\lambda_i\) calculated under the inferred parameters have small fluctuations with respect to the empirical averages \(\bar{\lambda}\). However the computation we performed in the previous section (relying on a second order truncation of \(\log \lambda_i\)) can be hardly used to obtain a control on the accuracy of this approximation or to define its domain of validity. These problems are addressed in appendix D where we show that the error of the MF estimates \(\theta_{\text{MF}}\) with respect to the maximum-likelihood estimates \(\theta_{\text{MLE}}\) (8)

$$\delta\theta = \theta_{\text{MF}} - \theta_{\text{MLE}}$$

(27)

is directly bounded by the ratio of the empirical variance of \(\lambda_i\) to its empirical mean. In particular, at leading order in the fluctuations and in the limit of large \(T\), one has

$$||\delta\theta|| \leq \frac{\bar{\lambda}(\lambda_0^2)}{(\lambda_{\text{MLE}})^2} ||\mathbf{C}|| ||\bar{\lambda}||,$$

(28)

where |||| refers to the \(L^2\) norm in case of vectors and spectral norm in case of matrices, \(\bar{\lambda} = [\bar{\lambda}_{a}^{b}]_{a=0}^{d},\) and where the variance corresponds to the \textit{empirical} one computed under the saddle-point parameters \(\theta_{\text{MLE}}\). We see that when the fluctuation ratio \(\bar{\lambda}(\lambda_0^2)/(\lambda_{\text{MLE}})^2\) is very small, the results of the MF estimate can be very close to the maximum-likelihood estimate.

\(^6\) As the region of integration in \(\theta\) is the positive orthant \(\mathbb{R}^{d(d+1)}_{+}\), equation (24) should in principle include extra terms preventing the emergence of negative couplings. We prefer to extend the region of integration to the whole space \(\mathbb{R}^{d(d+1)}\), so to include potentially negative couplings. A saddle-point expansion of equation (23) reveals indeed that these differences are subleading in \(T\).
3.3.1. Analysis of the fluctuation ratio. The natural question that follows these considerations concerns the characterization of the situations when this fluctuation ratio is very small. In appendix A, we establish these conditions in the particular case of a perfectly homogeneous Hawkes process. We show that in the limit of small $||\Phi||_i$ (weak endogeneity case (i) below), or in the limit of large dimension $d$ or slow interactions (self-averaging interactions case (ii) below) the fluctuation ratio is small. More precisely (see equation (67)), we find that it is directly controlled by the spectral norm $||\Phi||_i$, the system dimension $d$ and the interaction characteristic time-scale $\tau_g$ (defined in appendix A):

$$\frac{\nabla(\lambda_i')}{(N)^2} \sim \frac{1}{N\tau_g} \left( \frac{||\Phi||_i^2}{d(1 - ||\Phi||_i)^{2 - 1/\eta}} \right).$$

Let us recall that in the exponential kernel case, one has $\eta = 1$ (see appendix A). Intuitively, this characterization indicates that the MF approximation is valid whenever one of these conditions is met:

(i) **Weak endogeneity** ($||\Phi||_i \ll 1$): If the spectral norm $||\Phi||_i$ (that controls the level of endogeneity of the Hawkes process) is much smaller than one, then the intensity vector $\lambda_i$ is dominated by the exogenous component, and one can expect a very small fluctuation ratio.

(ii) **Self-averaging interactions** ($d\bar{\lambda}\tau_g \gg 1$): Even if the system is not weakly endogenous (i.e. $||\Phi||_i$ is not necessarily small), if the component-wise intensities are determined by a very large number of events with comparable contributions, a law of large numbers leads to small fluctuation ratios. This can notably occur in two different situations. First in large dimensional ‘quasi-homogenous’ systems where a large number of events associated with each of the components contribute to the intensity. This regime corresponds to large $d$ in equation (29) and has been studied notably in [21]. The second situation is the case of ‘slow interactions’ when a large number of past events equally impact the intensity. If the characteristic time scale of the kernel $g(t)$ is large with respect to the typical inter-event distance, then one expects the past contribution to the instantaneous intensity $\lambda_i'$ to have small fluctuations. This regime corresponds to a large $\bar{\lambda}\tau_g$ in equation (29).

The perfect homogeneous case is a very good toy model, it has the main ingredients that make the MFH valid. The result we so-obtained in the appendix A reproduce much more complex situations. Indeed, in the following we considered a Hawkes process with a two-block interaction matrix and exponential kernels fully specified in section 5. Figures 1 and 2 show that the formula for the fluctuation ratio $r_i = \sqrt{\frac{\nabla(\lambda_i')}{(N)^2}}$ (see equation (29) or equivalently (67) for general kernel case and equation (72) for exponential kernels) lead to a perfect prediction of the simulated curve.

The right plot of figure 1 shows that the fluctuation ratio $r_i$ of the first block ($i = 1$) behaves in very good agreement with equation (72). It scales like $d^{-1/2}$ when the dimension varies. Figure 2 shows the same ratio as a function of both the spectral norm $||\Phi||_i$ and the dimension $d$. Let us point out that, when varying the norm $||\Phi||_i$ and/or the dimension $d$, the average expectation $\Lambda$ is held fixed. One can notably see that, both (i) for a fixed value of the dimension, there exists a sufficiently small interaction $||\Phi||_i$ such that the fluctuation ratio can be arbitrarily small and (ii) for a fixed value of the norm (even large), there exists a sufficiently large system dimension $d$ such that the fluctuation ratio can be arbitrarily small. The displayed contour plots are the ones predicted by our analytical considerations (72) proved in appendix A (see equations (67) and (72)). The theoretical prediction appears to be in good agreement with simulated data.
Figure 1. Fluctuation ratio \( r = \sqrt{\frac{V}{\Lambda}} \) of the intensity of a simulated Hawkes process as a function of the number of components \( d \) for a fixed value of \( \Lambda \) and the interaction parameter \( ||\Phi||\) = 0.5 against the theoretical prediction of equation (65). By increasing the size of the system with a fixed value of the intensity, the fluctuations are found to decrease as \( d^{-1/2} \). Indeed, the theoretical predictions (dashed blue line) perfectly match the results of simulation (dots). The parameters of the simulation are defined as in section 5, and are equal to \( \beta = \mu = 1 \). The interaction matrix has the two-block structure as illustrated in section 5.

Figure 2. Fluctuation ratio \( r = \sqrt{\frac{V}{\Lambda}} \) for a system with variable number of components \( d = \{1, \ldots, 128\} \) and interaction strengths \( ||\Phi|| \in [0, 0.8] \), given a fixed value of the intensities \( \Lambda \). We have chosen \( p = 1 \) and \( \Phi(t) = 1, \alpha^2 e^{-\beta t} \) (with \( \beta = 1 \)) and chosen a two-block structure for the matrix \( \alpha_t \) with entries equal to either \( \alpha^2 = 0 \) or \( \alpha^2 = ||\Phi||/c \), with \( c \in \{[d/2], [d/2]\} \), as described in section 5. The other simulation parameters are set to \( \mu = 1 - ||\Phi||, \beta_t = 1 \). The figure shows that larger values of \( ||\Phi|| \) increase the size of the fluctuations. Conversely, the fluctuations can be reduced by an arbitrary amount by distributing the interaction among a large number of components \( d \). The theoretical predictions of equation (65) correspond to the contour lines, which accurately reproduce the results of simulation.
Finally, let us remark that while above considerations are useful for assessing the validity of the MF approximation under some specific assumption for the model, it is possible to bound a priori the approximation error $\delta \theta$ by exploiting the results of appendix E, that allow one to estimate the MF error by supplying the empirical cumulants of the data, without requiring any assumption about the underlying model.

3.3.2. Estimation error and validity condition. Since it basically amounts in solving a linear system, it is clear that the MF estimator will perform much faster than any MLE based estimator (see next section for detailed analysis on the complexity). The idea is to get, while being much faster, an estimation as precise as the MLE. This is true provided $||\Delta \theta||$ is smaller than the statistical error $||\delta \theta||$ associated with the maximum likelihood estimator. One can obtain a self-consistent estimation of the regime in which the error is dominated by its statistical component by using equation (26):

$$||\Delta \theta|| \approx \left( \frac{\text{tr}(C)}{T} \right)^{1/2}.$$  \hspace{1cm} (30)

that in principle is valid under the MF approximation. Figure 5 below shows that this assumption is correct under a large choice of parameter values. Then a simple condition for the MF method to perform as well as the MLE is

$$||\delta \theta|| < \left( \frac{\text{tr}(C)}{T} \right)^{1/2}.$$  \hspace{1cm} (31)

Equation (28) provides a sufficient condition for this to hold (after identifying empirical averages and statistical averages), namely

$$\frac{\text{tr}(C)}{(N^2)} ||C|| ||\Lambda|| \lesssim \left( \frac{\text{tr}(C)}{T} \right)^{1/2}.$$  \hspace{1cm} (32)

Equation (29) has been established in the perfectly homogeneous case for which $||\Lambda|| = \sqrt{d} N$. Moreover, we use very conservatively the inequality $||C|| \lesssim \text{tr}(C)$, which is likely to be largely underestimating the statistical error, so that in practice the performance

![Figure 3. Illustrative plot of the results of the reconstruction of a Hawkes model with a three-block structure for the matrix $\alpha$. We used the parameters $||\Phi|| = 0.5$, $\mu = 1$ and $\beta = 1$, and ran the simulation for $T = 10^4$. The true underlying matrix (left) is reconstructed both with the MF algorithm (center) and with a maximum likelihood estimation (right), yielding very similar results.](image-url)
of the MF algorithm is much better than what appears from equation (34) below (see figure 4). Indeed, plugging (29) and assuming the scaling $||C|| \sim (N)^{-1}$ in the last equation leads to the sufficient condition

$$\frac{1}{\sqrt{N} r_g} \left( \frac{||\Phi||^2}{d^{1/2}(1 - ||\Phi||^2)^{1/2}} \right) \lesssim \left( \frac{1}{T} \right)^{1/2}. \tag{33}$$

Therefore

$$T \lesssim T^* = \frac{N r_g}{d^{1/2}(1 - ||\Phi||^2)^{1/2}} \left( \frac{||\Phi||^2}{||\Phi||^2} \right). \tag{34}$$

Although far from being optimal, this rough bound provides a qualitative idea of how the different values of $d$, $||\Phi||$ and $r_g$ can contribute in determining the performance of the MF approximation. We will see in section 5, that in practice $T$ can be very large for moderate system dimension and values of the spectral norm. In other words, the situations where the performance of the MF method is comparable to those of the maximum likelihood are very easy to meet, in particular for large system sizes. In section 5 we will illustrate this point by showing that in a broad range of regimes of $(||\Phi||, d, T)$ we will be able to exploit the condition $T < T^*$, thus outperforming the maximum-likelihood algorithm through the superior numerical performance of the MF approximation.
4. Implementation and complexity

In this section we want to illustrate the numerical implementation of the MF algorithm, by comparing its complexity with the one of other state-of-the art algorithm. We will express all the complexities in terms of

- $d$ is the dimension of the Hawkes process.
- $p$ is the number of basis kernels.
- $L$ is the maximum average intensity over all the components ($\Lambda = \text{sup}_t \lambda_t$).
- $T$ is the sample size (expressed in physical time).
- $t$ is a characteristic time above which the kernels can be considered null.
- In the case of an iterative algorithm $R$ will denote the number of iterations needed.

Before focusing on the complexity of the MF method, let us briefly review the main alternative methods and discuss their corresponding complexities.

4.1. Maximum-likelihood estimation (MLE)

This class of methods, following [17], builds upon the direct maximization of the likelihood function $\log L(\theta|N)$. This problem is concave, so efficient solvers like BFGS are able to quickly find a solution. Indeed, the main drawback for this approach is the computational cost of each of the iterations of the algorithm, as the complexity of each evaluation of the gradient function scales with the number of points in the sample $n$. More precisely, this strategy requires to minimize the function

$$-\log L(\theta|N) = \gamma(\theta) = \sum_{i=1}^{d} h^T \theta - \sum_{m=1}^{n} \log \sum_{a=0}^{dp} \theta_{ma} G_{ma}. \quad (35)$$

Where

$$G_{ma} = \int_{0}^{T} dt \; g^{a}(t_m - t). \quad (36)$$

One can reduce the cost of each iteration by precomputing the coefficients $G$ on the right hand side of equation (35).

The corresponding complexity is $O(d^2p^2L^2T)$. Once done, the complexity for computing the log likelihood (or its gradient) is dominated by the rightest term in (35) and is of the order of $O(d^2p^2L^2T)$. If the overall gradient algorithm involves $R$ iterations, the overall complexity writes

$$\text{MLE-complexity} \sim d^2pL^2T + Rd^2pL. \quad (37)$$

Let us point out that if $R$ is large, its corresponding term dominates. This is typically the case, as seen in figure 6.

4.2. Expectation-maximization (EM)

In the case of the model we focus on, EM is actually as complex as MLE, as it requires calculating a gradient of the likelihood function at each iteration [23]. According to the type of MLE algorithm adopted, the relative speed of EM with respect to MLE can change. Taking BFGS as an example, the EM algorithm suffers actually from a slower convergence, as it is unable to exploit second-order information, as done instead by Newton and quasi-Newton methods.
4.3. Contrast function (CF)

The closest method to the one that we have presented is a generalization of the contrast function approach proposed in [7]. In this framework, the $\theta$ should minimize a loss function defined as

$$C(\theta) = \frac{1}{2} \sum_i \int_0^T dt \left( \frac{dN_i^t}{dt} - \lambda_i^t \right)^2,$$

resulting in

$$C(\theta) = T \left( \sum_{i=1}^d \left( -\tilde{\lambda}^t k_i^T \theta^i + \frac{1}{2} \theta^i \mathbf{J}^i \theta^i \right) \right),$$

where the collection $k_i^t$ is defined as in equation (21), while the matrix $\mathbf{J}^i$ is given by

$$J^i_{ab} = \frac{1}{T} \int_0^T dt dN_i^a dN_i^b g^a(t - t') g^b(t - t').$$

This approach also maps the inference problem on a linear system, whose solution is given by

$$\theta^i = \tilde{\lambda}^i \mathbf{C}' k_i^t,$$

where $\mathbf{C}' = (\mathbf{J}')^{-1}$. This solution is numerically close to equation (25). The computation of the $\mathbf{J}'$ has a complexity of $O(d^2 p^2 N^2 T)$ (we cannot use the trick illustrated in appendix B for the computation of $\mathbf{J}$). When using an optimized linear algebra package, the time of computation is typically dominated by the preprocessing phase (we can neglect the matrix inversions). Thus, the overall complexity writes

$$\text{CF-complexity} \sim d^2 p^2 N^2 T.$$

4.4. MF algorithm

The simplest procedure to use in order to obtain the estimator equation (25) and its associated covariance matrix is summarized in algorithm 1.

**Algorithm 1.** Calculation of the mean-field estimator $\mathbb{E}\theta^i$ and its covariance $\mathbb{C}(\theta^i, \theta^j)$.

| Input: $N_i$, $g(t)$ |
|-----------------------|
| Output: $\mathbb{E}\theta^i$, $\mathbb{C}(\theta^i, \theta^j)$ |
| 1: Compute the auxiliary functions $h$, $k^i$ and $\mathbf{J}'$ as shown in appendix B |
| 2: Invert the $d$ matrices $\mathbf{J}'$ |
| 3: Evaluate equations (25) and (26) |
| 4: return $\mathbb{E}\theta^i$, $\mathbb{C}(\theta^i, \theta^j)$ |

Note that the complexity of the preprocessing phase (computing the auxiliary functions $h$, $k^i$ and $\mathbf{J}'$) is dominated by the computation of the $\mathbf{J}'$. As proved at the end of appendix B the corresponding complexity is

$$\text{MF-complexity} \sim d^2 p \Lambda T \times \max(\Lambda t, dp).$$

The inversion of the matrices can be performed with a complexity bounded by $O(d^2 p^3)$. In order to speed up the overall computation, one can use some tricks such as the one presented in appendix B. Other prescriptions can be adopted to speed up the computation:
• If one is not interested in the covariances \( \mathbb{C}(\theta^a, \theta^b) \), the inversion of the matrix can be avoided, by replacing it with the solution of the \( d \) linear systems \( \mathbf{J}^t \theta_{\mathbf{MF}} = 2k^t - \mathbf{h} \). This substantially reduces the time required to find \( \theta_{\mathbf{MF}} \) in step 2 and 3 of algorithm 1, as the matrix inversion can be traded with a faster linear solver (e.g., Cholesky decomposition, or iterative algorithms such as BFGS [24]).

• The case of weak endogeneity described above allows one to perform an approximate inversion of the matrices \( \mathbf{J}^t \), as shown in appendix C, reducing the time required for the solution of the \( d \) linear systems from \( O(d^3p^3) \) to \( O(d^3p^2) \) (corresponding to \( d \) matrix-vector compositions). Note indeed that this is expected to yield inaccurate results when the fluctuations of the \( N_t \) are of a larger order than \( d^{-1} \), even though the elements of the \( \alpha \) matrix are individually small (see again appendix C).

The time of computation is typically dominated by the preprocessing phase up to very large sizes. As an illustrative example, a parallel implementation in python + c on a four-core machine with a 3.40 Hz CPU the calculation of the auxiliary functions takes \( \sim 6 \times 10^2 \) s for \( d = 128, p = 1 \) and \( n = 2 \times 10^3 \) events, while the matrix inversions can be performed in a time \( \sim 10^{-1} \) s. Finally, we remark that the algorithm can be straightforwardly parallelized across components, meaning that a factor \( d \) can be gained in both the preprocessing and the matrix inversion phase.

Comparing the various complexities (37), (42) and (43), one can see that, as long as the dimension is not too large and the kernel support is not too small, the MF method performs better than both CF and MLE (see figure 6). Our numerical simulations show that \( R \) depends on all the other factors (notably \( \Lambda \) and \( t \)). The precise relationship is hard to obtain and calls for heavy numerical experimentations that will be the topic of a forthcoming publication. However, it is clear that, as for the number of iterations of the EM algorithm (see [20]), \( R \) should increase very strongly with \( \Lambda t \) and lead consequently to a very slow algorithm.

More systematically, we find that for dimensions up to \( d = 100 \), when considering \( p = 1, 2 \) exponential basis kernel functions and \( \alpha, \beta \) of order unity, the MF method is still the fastest one in comparison with all the others.

5. Numerical examples

We report in this section the results obtained by calibrating a Hawkes process from synthetic data generated by a known model, and by comparing the performance of our algorithm with the one of known methods.

5.1. Single exponential basis kernel

First, we have considered as a benchmark the case in which \( p = 1 \), so that \( g^a(t) = g(t) \equiv g(t) \). For \( j > 0 \), we have assumed the basis kernel to have the exponential structure

\[
g(t) = \beta e^{-\beta t}1_{t > 0}.
\]

(44)

The topology of the matrix \( \alpha \) has been chosen from the following ensemble:

• A block structure in which each nodes belong to one out of several clusters. In particular we studied the case in which \( \alpha^i = 0 \) (if \( i \) and \( j \) do not belong to the same cluster) or \( \alpha^i = \alpha/c \) (if they do belong to the same cluster, with size given by \( c \));
Regardless of the structure, in both cases one has $\|\Phi\|_1 = \|\alpha\|_1 = \alpha$, so that we used the parameter $\alpha \in \mathbb{R}_+$ in order to interpolate between the non-interacting case $\alpha = 0$ and the critical one $\alpha = 1$. We have chosen for the purpose of these numerical experiments the values $\mu = [1, \ldots, 1]$.

As an illustrative example, we first plot in figure 3 the results of our algorithm in the case of a three block structure for the $\alpha$ matrix. The results are for this choice of value essentially the same as the ones obtained by MLE, with a considerable reduction in the computational time.

In order to systematically assess the performance of the algorithm under various circumstances, we have varied $\alpha$ uniformly in the interval $[0, 1)$, and addressed the problem of studying the scaling of the results in $d$ (we used $d = 2, 4, 8, \ldots, 128$) and in $T$ (we considered $T \in [10^3, \ldots, 10^5]$). Unless specified otherwise throughout the text, we have assumed block structures with two clusters, so that $c \in \{[d/2], [d/2]\}$. We measured the quality of the reconstruction by using

- The negative log-likelihood of the sample $\mathcal{L}(N, \theta) = -\log \mathbb{P}(N|\theta_{\text{inf}})$ under the inferred couplings
- The relative error on the non-zero couplings

\[
\delta \alpha_{\text{rel}}^2 = \sum_{i,j|\alpha_{\text{true}}^{ij} \neq 0} \left( \frac{\alpha_{\text{inf}}^{ij}}{\alpha_{\text{true}}^{ij}} - 1 \right)^2. \tag{45}
\]

- The absolute error on all the couplings

\[
\delta \alpha_{\text{abs}}^2 = \sum_{i,j} (\alpha_{\text{inf}}^{ij} - \alpha_{\text{true}}^{ij})^2. \tag{46}
\]

The results that we have found for the relative error $\delta \alpha_{\text{rel}}$ in the two-block case are summarized in figure 4, where we have shown them for $\alpha = 0.3$ and $\alpha = 0.7$. What one finds is that, as expected, the error of the maximum likelihood estimator decreases as $T^{-1/2}$ by increasing the sample size $T$. Indeed, the same plot also shows that the MF estimator is able to match the results of the maximum likelihood estimator up to a maximum value $T_*$ beyond which the statistical error no longer dominates $\delta \alpha_{\text{rel}}$ for the MF estimator, and the performance of MF deteriorates. At that point, the approximation error becomes comparable with the statistical one, inducing a plateau in the curve of the error as a function of $T$. What is interesting is that—even for moderate values of $d$—the value of $T_*$ is quite large, indicating that the MF approximation is extremely effective in a broad range of regimes. By increasing the value of $\|\Phi\|_1 = \alpha$, even though the relative error decreases due to the increase of the signal $\alpha$, the quality of the MF approximation decreases, and the value of $T_*$ becomes smaller. Indeed, for a fixed value of $T$ and $\|\Phi\|_1$, it is always possible to find a sufficiently large $d$ so that $T \ll T_*$, and the MF approximation is effective. This is confirmed in figure 5, where we have represented the absolute error $\delta \alpha_{\text{abs}}$ obtained by MLE against the one estimated by using equation (26).

On the other hand, the computational time required in order to run the MF algorithm is considerably smaller, as summarized in figure 6. In such figure we have represented the computational time required in order to run the different algorithm up to a target negative log-likelihood $\mathcal{L}(N, \theta)$, on a machine with the same specifications as above, and shown that by taking into account both the pre-processing phase and the solution of the linear system via
matrix inversion, the MF algorithm achieves within a single iteration a value of the target function which is very close to the asymptotic one got from a BFGS minimization.

Figure 5. Absolute error couplings $\delta_{\text{abs}}$ as a function of the sample size $T$ for the maximum likelihood estimator (dashed lines) and the estimation of the error provided by equation (26) (solid lines), for a fixed value of the empirical intensity $\bar{L}$ and the same choice of parameters as in the previous figure. These figures show that the mean-field estimation of the absolute error provides, up to a factor independent of $T$ and $d$, a good estimation of the true MLE absolute error. Notice that both errors are very close for small values of $||\Phi||_h$ (left figure) while the MLE error is underestimated by the MF error when approaching the critical point $||\Phi||_h = 1$ (right figure).

Figure 6. Time required by the learning algorithms described in section 4 in order to achieve fixed values of accuracy, as measured by the negative log-likelihood (left panel) and the relative error (right panel). We have considered the two block case described in section 5, and adopted the same parameters as in figure 4, taking in particular a dimension equal to $d = 25$. 
5.2. Multiple exponential basis kernels

We have also considered the more general case in which the $F(t)$ is given by a sum of kernels of different nature. Even though the implementation of an algorithm with multiple basis kernels is technically more challenging, there are no qualitative differences with respect to what we have stated above. That is why we have chosen to present an illustrative example showing how the MF algorithm performs in the multiple basis case. In particular, we have considered an example in which $p = 2$, in which the matrix $\alpha_1$ has a uniform two-block structure as in the previous case, while for the second matrix $\alpha_2$ only the complementary blocks are uniformly filled (see figure 7). For both $\alpha_1$ and $\alpha_2$, the non-zero entries have value $\alpha/c$, being $c$ the size of the block. Finally, we chose $\beta_1 = 1$ and $\beta_2 > 1$, so that for $\beta_1 = \beta_2$ one recovers a completely homogeneous interaction structure. Figure 7 shows an illustrative plot of the results of the reconstruction in the multiple basis kernel scenario.

5.3. The role of $\beta$

We have also studied the effect of the misspecification of $\beta$ by studying the effect of an input kernel $g(t)$ different from the $g_{\text{true}}(t)$ used to generate the data. We did it in the case of $p = 1$, with a block $\alpha$ matrix as described in the paragraph above relating to the single
exponential basis kernel scenario. Also in this case we have chosen $\mu = 1 = \beta_{\text{true}} = 1$, with $\beta_{\text{true}}$ denoting the true decay constant for the exponential kernels, whereas $\beta_{\text{in}}$ will denote the decay constant chosen to perform the inference. Our results are summarized in figure 8, where we illustrate them in a specific case. We find that the likelihood of the model under the inferred couplings is not strongly affected by the misspecification of $\beta$, as the inferred intensity turns out to be very weakly dependent on $\beta_{\text{in}}$. What actually happens is that by changing $\beta_{\text{in}}$ it is possible to tune a bias changing $\mu_{\text{in}}$ in favor of $\alpha$. This behavior can be understood in the framework of our MF approximation:

- On one hand, appendix C shows that the value of the inferred couplings $\alpha$ is proportional to the fluctuations of the auxiliary function $\delta k$. According to equation (113), these fluctuations are obtained by integrating the empirical correlation function against the kernel $g_{in}(t)$ in order to obtain $\delta k$. Hence, when $\beta_{\text{in}}$ is large $g_{in}(t)$ is peaked close to the origin, where the correlations are larger, resulting in a greater value for $\delta k$. Conversely, for large $\beta_{\text{in}}$ such integral assumes a small value, because the correlations are small at the timescales in which $g_{in}(t)$ has large mass. In any case, the norm of $\delta k$ is bounded from above, and is maximum for $\beta_{\text{in}}$ large.
- On the other hand, the entries of $C$ decrease with $\beta_{\text{in}}$. For example, equation (86) shows that all the entries but the component $a = b = 0$ of the first order contribution $C_{00}^{ab}$ to $C^{\alpha \beta}$ have a $\beta_{\text{in}}^{-1}$ dependence. Hence, in this case they decrease unboundedly with $\beta_{\text{in}}$.

The latter effect is dominant in our numerical examples, so that a small value of $\beta_{\text{in}}$ leads to a bias in favor of $\alpha$, while a large value favors $\mu$. Independently of $\beta_{\text{in}}$, the combination of $\mu$ and $\alpha$ leading to the inferred value of $\Lambda$ is almost unchanged (i.e., the inferred event rate is always correct).
Even though these conclusions depend on the choice of an exponential kernel for $g_{ik}$, we believe the two effects identified above to play an analogous role even in more general scenarios.

6. Conclusion and prospects

We have presented an approximated method for the calibration of a Hawkes process from empirical data. This method allows one to obtain in close form the inferred parameters for such a process and its implementation is considerably faster than the algorithms customarily employed in this field (MLE+BFGS, EM, CF). In particular, we map the problem of the estimation of a Hawkes process on a least-square regression, for which extremely efficient techniques are available [24, 25]. Although this method is asymptotically biased, we have shown numerically and analytically that the bias is negligible in situations when the interactions of the system are sufficiently weak or self-averaging, which is thought to be the case in many applications to high-dimensional inference.

We believe that our method is not limited to the framework that we have presented, but can be extended to more general settings like marked or nonlinear Hawkes processes. Another prospect of particular interest is the case in which a non trivial prior is added to the Bayesian framework equation (9), which allows one to embed regularization in our framework. Popular priors that are relevant in the field of optimization include for example the Gaussian prior

$$P^L_0(\theta) \propto \exp\left( \sum_{ia} \kappa_{ia}(\theta_{ia})^2 \right)$$

and the Laplace prior

$$P^L_0(\theta) \propto \exp\left( \sum_{ia} \kappa_{ia}|\theta_{ia}| \right).$$

The inference problem analogous to the evaluation of the partition function (12) associated with these priors is obviously more complex than the one with a flat prior that we have presented. Nevertheless, our MF approach allows even in those cases a mapping on simpler problems, for which efficient solution methods exists [24]. In particular,

- The case of a Gaussian prior can be mapped to the one of the minimization of the quadratic form

$$\mathcal{H}^L(\theta) = \sum_{i=1}^{d} \frac{1}{2} \theta_i^{T} J \theta_i - (2 k_i - h_i) \theta_i + \sum_{ia} \kappa_{ia}(\theta_{ia})^2.$$  

- The case of a Laplace prior can be mapped to the LASSO problem [26]

$$\mathcal{H}^L(\theta) = \sum_{i=1}^{d} \frac{1}{2} \theta_i^{T} J \theta_i - (2 k_i - h_i) \theta_i + \sum_{ia} \kappa_{ia}|\theta_{ia}|.$$  

Other types of regularizers might be considered by applying this same scheme.

We believe the approach we propose in this paper to be of particular interest for the problem of high-dimensional inference as, by construction, it is particularly effective in the regime of large $d$ (especially relevant for big-data analysis) where traditional numerical methods can be computationally expensive. Last but not least, the MF method can provide an
easy and efficient way to set a good starting point for standard iterative algorithms that maximize the likelihood objective.

Acknowledgments

This research benefited from the support of the ‘Chair Markets in Transition’, under the aegis of ‘Louis Bachelier Finance and Sustainable Growth’ laboratory, a joint initiative of École Polytechnique, Université d’Évry Val d’Essonne and Fédération Bancaire Française.

Appendix A. Fluctuation ratio of the intensity on the homogeneous case

In order to get an intuition about the behavior of a Hawkes process in a very stylized setting, we present an estimation of the square root of the ratio of the variance of \( \lambda_t \) over its squared mean. Thus, we will localize in the phase space of the model the points in which the regime of large dimension \( d \) (quasi-homogeneous) and/or weak interactions (see (i) and (ii) in section 3.3) can be detected. In order to do this, we study the quantity

\[ r_t = \sqrt{\frac{\text{Var}(\lambda_t)}{\mathbb{E}(\lambda_t^2)}}, \tag{51} \]

in the simplest possible framework, that is, a perfectly homogeneous Hawkes process of the form

\[ \mu = \mu \begin{bmatrix} 1, & \ldots, & 1 \end{bmatrix}, \tag{52} \]

\[ \Phi(t) = \Phi(t) \begin{bmatrix} 1 \end{bmatrix}_{1 \leq i \leq d}. \tag{53} \]

Let us point out that in this appendix we calculate the observables of a Hawkes process by averaging over its stationary measure, instead of using the empirical averages. Hence, when using these result in section 3.3, one should keep in mind that we are replacing all the empirical averages by expectations.

1) Mean of \( \lambda_t \). The mean value is easy to compute once one particularizes the results of [1, 2] in the homogeneous case, obtaining

\[ \Lambda = \mathbb{E}(\lambda_t) = \left( I - \left[ \int \Phi(t) dt \right]_{1 \leq i, j \leq d} \right)^{-1} \mu, \tag{54} \]

\[ = \left( I - ||\Phi||_1 [1]_{1 \leq i, j \leq d} \right)^{-1} \mu, \tag{55} \]

\[ = \frac{\mu}{I - d ||\Phi||_1 [1]_{1 \leq i \leq d}}. \tag{56} \]

Let us point out that we want to consider the focus on a regime in which the SC hold, so that we hold the averaged intensity \( \Lambda \) constant when varying the dimension \( d \). In order to do so, we choose to keep the exogeneous intensity \( \mu \) constant and let \( \Phi(t) \) vary so that \( d ||\Phi||_1 \) is constant. Thus we set

\[ \Phi(t) = g(t) \frac{||\Phi||_1}{d}, \quad \text{with} \quad ||g||_1 = 1. \]
In that case, one gets
\[ \Lambda = \left[ \frac{\mu}{1 - \|\Phi\|_1} \right]_{1 \leq i, j \leq d} = [\Lambda]_{1 \leq i, j \leq d}. \]  
(57)

(2) Variance of \( \lambda_t \). The variance can be obtained by using the stationary version of the martingale representation for the Hawkes process (see [18])
\[ \lambda_t = \Lambda + \int_{-\infty}^0 \Psi(t - t') \, dM_t, \]  
(58)
where \( M_t \) is the martingale \( M_t = N_t - \int_0^t \lambda_s \, ds \) (one has \( E[dM_t | F_s] = 0 \)), and \( \Psi(t) \) can be obtained from the Fourier transform of \( \Phi(t) \), that we denote as \( \hat{\Phi}(\omega) = \int \Phi(t) e^{-i\omega t} \):
\[ \hat{\Psi}(\omega) = (1 - \hat{\Phi}(\omega))^{-1} - 1, \]  
(59)
as shown in [1, 2] and generalized in [18]. Then one can write
\[ \mathbb{V}(\lambda_t^2) = \sum_{1 \leq i, j \leq d} E \left[ \int_{-\infty}^{t'} dM^j_{t'} \int_{-\infty}^{t'} dM^i_{t'} \Psi(t - s) \, \Psi(t - s') \right], \]  
(60)
\[ = \sum_{j=1}^d \int_0^\infty dt \left( \Psi(t)^2 \right) (t) = d\Lambda \, \|\Psi\|^2 = \frac{d\Lambda}{2\pi} \|\hat{\Psi}\|^2, \]  
(61)
where \( \| \cdot \| \) stands for the \( L^2 \) norm. However, since
\[ \hat{\Psi}(\omega) = \left( 1 - \frac{\hat{g}(\omega) \|\Phi\|_1}{d} \right)^{-1} - 1 = \frac{1}{d} \left( \frac{\hat{g}(\omega) \|\Phi\|_1}{1 - \hat{g}(\omega) \|\Phi\|_1} \right) [1]_{1 \leq i, j \leq d}, \]  
(62)
one has
\[ \hat{\Psi}(\omega) = \frac{1}{d} \left( \frac{\hat{g}(\omega) \|\Phi\|_1}{1 - \hat{g}(\omega) \|\Phi\|_1} \right). \]  
(63)
Thus
\[ \mathbb{V}(\lambda_t^2) = \frac{\Lambda \|\Phi\|_1^2}{d} \int \frac{d\omega}{2\pi} \left( 1 - \frac{\|\hat{g}(\omega)\|^2}{\|\|\Phi\|_1 \hat{g}(\omega)\|^2} \right). \]  
(64)
So the scaling in either limit \( \|\Phi\|_1 \to 0 \) (case (i) in section 3.3) or \( d \to +\infty \) (case (ii) in section 3.3) is thus of the form \( \mathbb{V}(\lambda_t^2) \sim \Lambda \|\Phi\|_1^2 / d \). Let us try to be more accurate and characterize the scaling when approaching the critical case \( \|\Phi\|_1 \to 1 \). It will naturally depend on the behavior of \( \hat{g}(\omega) \) around 0. We set
\[ \hat{g}(\omega) = 1 + K\omega^{\eta} + o(\omega^{\eta}). \]  
Let us point out that in the case \( g \) is exponential, one has \( \eta = 1 \) and in the case \( g(x) \sim x^{-\gamma} \) (when \( x \to +\infty \)) then, using a Tauberian theorem (see theorems 3 and 4 in section XIII.5 of [27]), one can prove that \( \eta = \gamma + 1 \). Let us set \( \epsilon = 1 - \|\Phi\|_1 \). Then, focusing on \( |\omega| < a \) (\( a \) fixed), the integral in (64) becomes
This last result along with (64) gives that

$$\mathbb{V}(\lambda'_i) = \frac{\Lambda}{\tau_g(\|\Phi\|_i)} \left( \frac{\|\Phi\|_i^2}{d (1 - \|\Phi\|_i)^{2-1/\eta}} \right),$$

(65)

where

$$\tau_g(x) = (1 - x)^{-2+1/\eta} \left( \int \frac{d\omega}{2\pi} \frac{|\hat{g}(\omega)|^2}{|1 - x\hat{g}(\omega)|^2} \right)^{-1}$$

(66)

is a function on $[0, 1]$ which is both bounded and bounded away from zero. As it is illustrated below in the exponential case, the constant $\tau_g(\|\Phi\|_i)$ can be interpreted as a characteristic time scale of the kernel.

(3) Fluctuations of $\lambda_i$ and (MFH). Thus, the fluctuations of $\lambda_i$ behave, in either limit (i) weak interactions ($\|\Phi\|_i \ll 1$, case (i) in section 3.3) or (ii) large dimension $(d \gg 1$ case (ii) in section 3.3), like

$$\frac{\mathbb{V}(\lambda'_i)}{\mathbb{E}^2(\lambda'_i)} \sim \frac{1}{\Lambda \tau_g} \left( \frac{\|\Phi\|_i^2}{d (1 - \|\Phi\|_i)^{2-1/\eta}} \right).$$

(67)

and the MFH simply writes

$$(\text{MFH}) : \quad \eta' = \sqrt{\frac{\mathbb{V}(\lambda'_i)}{\mathbb{E}^2(\lambda'_i)}} \sim \frac{\|\Phi\|_i}{\sqrt{d \Lambda \tau_g (1 - \|\Phi\|_i)^{1-1/2\eta}}} \ll 1.$$ 

(68)

(4) Exponential case. In the simpler case in which

$$g(t) = \beta e^{-\beta t} t_{\geq 0}$$

(69)

we get

$$\hat{g}(\omega) = \frac{1}{1 + i\omega/\beta} \quad \text{and} \quad \eta' = 1,$$

(70)

which implies

$$\tau_g(\|\Phi\|_i) = \frac{2}{\beta}.$$ 

(71)

We see here that $\tau_g$ does correspond to the characteristic time scale of the kernel. The MFH simply writes

$$(\text{MFH}) : \quad \eta' = \sqrt{\frac{\mathbb{V}(\lambda'_i)}{\mathbb{E}^2(\lambda'_i)}} \sim \frac{\|\Phi\|_i \sqrt{\beta}}{\sqrt{2d \Lambda (1 - \|\Phi\|_i)}} \ll 1.$$ 

(72)

This result is illustrated in figure 1.
Appendix B. Auxiliary functions for MF-algorithm

This appendix is devoted to a more detailed analysis of the auxiliary functions, so to provide a recipe for an efficient numerical implementation. We start by reminding that

\[ h^a = \frac{1}{T} \int_0^T dt \int_0^t dN_t^a g^a(t - t'), \quad (73) \]

\[ k^{ia} = \frac{1}{N_f^i} \int_{t'=t}^t dN_t^i dN_t^j g^a(t - t'), \quad (74) \]

\[ J_{iab} = \frac{T}{(N_f^i)^2} \int_{t'=t}^t dN_t^i dN_t^j dN_t^b \times g^a(t - t') g^b(t - t'') \quad (75) \]

required in order to compute the quantities equations (25) and (26) defined in the main text. First, note that while in order to compute equation (74) one needs to discard the point \( t = t' \), in the case of equation (75) it is necessary to consider the points \( t = t' = t'' \), so that it reads more explicitly

\[ J_{iab} = \frac{T}{(N_f^i)^2} \int_{t'=t}^t dN_t^i dN_t^j dN_t^b \times g^a(t - t') g^b(t - t'') + \frac{T}{(N_f^i)^2} \delta^{i\beta} \delta^{ab} \int_{t'=t}^t dN_t^i dN_t^j g^a(t - t') g^b(t - t'). \quad (76) \]

B.1. Implementation and complexity of MF-algorithm

The non-zero components of the auxiliary functions can be computed efficiently by exploiting the relations

\[ h^a = \frac{1}{T} \sum_{m=1}^n w_{ta}^a, \quad (77) \]

\[ k^{ia} = \frac{1}{N_f^i} \sum_{m=1}^n y_{ia}^m, \quad (78) \]

\[ J_{iab} = \frac{T}{(N_f^i)^2} \sum_{m=1}^n y_{ia}^m y_{ib}^m, \quad (79) \]

where one has defined the quantities \( w_{ta}^a \) and \( y_{ia}^m \) as\(^7\)

\[ w_{ta}^a = \int_{t_a}^T dt' g^a(t - t_m), \quad (80) \]

\[ y_{ia}^m = \sum_{m'=1}^{m-1} g^a(t_m - t_{m'}) \delta_{i\alpha} \delta_{a\alpha'}, \quad (81) \]

while for \( a = 0 \) or \( b = 0 \), one can use the relations \( J^{i00} = J^{00} = k^{ia} \) and \( k^{00} = N_f^i \).

\(^7\) Note that The values of \( w_{ta}^a \) and \( y_{ia}^m \) do not have to be stored in memory and can rather be computed on-line during the evaluation of the auxiliary functions.
The basis kernels $g(t_m - t_n)$ can be cut-off when they fall below a target value of $\epsilon$ in order to gain computational speed. If we call $t$ the characteristic scale above which the basis kernels are null, then the complexity for computing the $\{y_{n_{m,i,u}}\}$ is $d^3p^2\Lambda T$. The sum corresponding to (79) then needs $d^3p^2\Lambda T$ operations. Thus the total complexity is

$$\text{MF-complexity} \sim d^2p^2\Lambda T + d^3p^2\Lambda T \sim d^2p^2\Lambda T \times \max(\Lambda t, dp).$$

(82)

Appendix C. Small fluctuations

We want to calculate here the value of the inferred couplings under the MF approximation in the case in which the fluctuations of the process $N_t$ are small. Accordingly, we define the quantities

$$\delta k^{ia} = k^{ia} - \bar{\Lambda}^a,$$  
$$\delta J^{iab} = J^{iab} - \delta^{iab} (1 - \delta^{iab}) \frac{\bar{\Lambda}^a \bar{\Lambda}^b}{\bar{\Lambda}^i},$$

$$= J^{iab} - J_0^{iab},$$

which are expected to be small if the condition (MFH) holds (explicit bounds on the rest terms are available in appendix E). In above equation we have also chosen the notation $\nu^{ab} = \int_0^\infty dt \ g^a(t) g^b(t)$, that we will employ extensively in the following. Then it is possible to perturbatively invert the matrix $J$ by using the formula

$$C^i = C_0 \sum_{n=0}^{\infty} (-\delta J C_0)^n,$$

(85)

where we have defined the collection $\{C_n\}$ with elements

$$C_0 = \bar{\Lambda} \begin{bmatrix} 1 + \sum_{j=1}^d (\bar{\Lambda}^j)^2 \sum_{q,q'=0}^p Z_{qq'} \{y^j\}_{1 \leq j \leq d} \{y^j\}_{1 \leq j \leq d} \{Z^b Z\}_{1 \leq j,k \leq d} \end{bmatrix},$$

(86)

and introduced a collection of $d$ vectors $\{y^i\}_{1 \leq i \leq d}$ of dimension $p$ together with a collection of $d$ matrices of size $p \times p$ denoted as $\{Z^i\}_{1 \leq i \leq d}$

$$y^i = [-\bar{\Lambda} \sum_{q=0}^p Z_{qq}]_{1 \leq q \leq p},$$

$$Z^i = [\bar{\Lambda} \nu^{ii}]_{1 \leq q,q' \leq p}^{-1},$$

(87)

(88)

In above equation we have recovered the original notation of section 2 which identifies kernel basis functions with subscripts, while the symbol $^{-1}$ above indicates the matrix inversion with respect to the $p \times p$ block. One can easily verify by using above equation and equation (84) defining $J_0$ and $\delta J$ that

$$C_0 J_0 = 1,$$

$$C_0 A = \Sigma^i,$$

(89)

(90)
With \( \Sigma' = [N^{\delta i ds}]_{i \in a, d \in dp} \). Thanks to these last two equations one can prove that
\[
\theta_{\text{MF}}^i = C'(2k^i - h),
\]  
(92) \[
\Sigma' + C'((\delta k^i - \delta h))
\]  
(93) at any order in \( \delta k \)! Together with the explicit expression for \( C_0^i \), this is the main result of this appendix, that allows us to conclude that:

1. The values of the interaction parameters \( \alpha \) are proportional to the fluctuations of \( \delta k^i \) and \( \delta h \) around their average values.

2. When the fluctuations are very small, the MF approximation explains all the events by forcing the exogenous intensities \( \mu \) to be equal to the empirical ones \( \Lambda \).

3. One could be tempted to expand equation (93) as
\[
\theta_{\text{MF}}^i = \Sigma' + C_0^i(\delta k^i - \delta h) - C_0^i\delta J^iC_0^i(\delta k^i - \delta h) + \cdots.
\]  
(94) This is not always appropriate, as the truncation can lead to large errors even when the term \( \delta J^i \) is extremely small.

In order to understand this last point, let us note that if one could take the limit of small \( \delta J^i \) independently of the number of summands \( dp \) in the tensor compositions, the approximation would be correct (this is appropriate in the regime of weak interactions described in section 5). On the other hand if the fluctuations are suppressed by the system size \( d \), it is important to control the scaling in \( d \) of \( \delta J^i \), as one typically expects \( \delta J^i \sim O(d^{-1}) \), so that \( \delta J^i C_0^i \) can be estimated to be of order 1 (this should be the case of well-mixed interactions mentioned in section 5) and the truncation leads to large errors.

Appendix D. MLE versus MF solution

In this appendix we want to construct a bound on the error induced by the MF approximation, so to be able to compare it with the typical value of the couplings and the statistical error due to the finite value of \( T \). In particular, we are interested in controlling the distance between the exact saddle-point solution \( \theta_{\text{MLE}}^i \) and the MF one \( \theta_{\text{MF}}^i \) given by equation (25), that are given by

\[
\theta_{\text{MLE}}^i = \arg\min_{\theta} \mathcal{L}(N, \theta),
\]  
(95) \[
\theta_{\text{MF}}^i = C'(2k^i - h).
\]  
(96) The idea is to bound their difference
\[
\delta \theta = \theta_{\text{MLE}} - \theta_{\text{MF}}.
\]  
(97) where \( \mathcal{L}(N, \theta) = -\log p(N|\theta) \) is the negative log-likelihood function equal to
\[
\mathcal{L}(N, \theta) = h^T \theta - \sum_{m=1}^{n} \log \lambda^m_{nm}.
\]  
(98)
Minimizing it with respect to $\theta^{ia}$ produces the relation:

$$0 = h^{ia} - \frac{1}{T} \sum_{m=1}^{n} \frac{\partial \lambda^{ua}}{\partial \theta^{ia}} - \frac{1}{T} \sum_{m=1}^{n} \frac{\partial \lambda^{ua}}{\partial \theta^{ia}} \left( \frac{1}{\Lambda^{ua}} - \frac{\lambda^{ua} - \bar{\Lambda}^{ua}}{(\bar{\Lambda}^{ua})^2} + \frac{(\xi^{ua} - \bar{\Lambda}^{ua})^2}{(\bar{\Lambda}^{ua})^3} \right)$$

$$= h^{ia} - 2 \lambda^{ia} + \sum_{k \in \mathcal{A}} f^{ab} \theta^{kb}_{\text{MLE}} - \frac{1}{T} \sum_{m=1}^{n} \frac{\partial \lambda^{ua}}{\partial \theta^{ia}} \left( \frac{\xi^{ua} - \bar{\Lambda}^{ua}}{(\bar{\Lambda}^{ua})^3} \right),$$

(99)

where in the second line we have expanded $\lambda^{ua}$ around $\bar{\Lambda}^{ua}$, and we defined the rest of the Taylor expansion $\xi^{ia}$, satisfying $|\xi^{ia}/\bar{\Lambda}^{ia} - 1| \leq |\lambda^{ia}/\bar{\Lambda}^{ia} - 1|$.

One can make some progress by first substituting the definition of $\delta \theta^{i}$ into above equation, and then using the definition of $\theta^{ia}_{\text{MLE}}$ in order to get:

$$\delta \theta^{i} = C^{i} v^{i},$$

(100)

where

$$v^{ia} = \frac{1}{T} \sum_{m=1}^{n} \frac{\partial \lambda^{ua}}{\partial \theta^{ia}} \left( \frac{\xi^{ua} - \bar{\Lambda}^{ua}}{(\bar{\Lambda}^{ua})^3} \right)$$

$$= \frac{1}{N_{\mathcal{J}}^{\theta}} \int dN^{a}_{\mathcal{J}} dN^{a}_{\mathcal{J}} g^{a}(t - t') (\xi^{ia}/\bar{\Lambda}^{ia} - 1)^2.$$

(101)

This allows us to construct an estimate for $v$ in order to obtain a bound on $\delta \theta$.

D.1. Expansion of $\delta \theta$

One can first use a bound on the spectral norm of the $C^{i}$ matrices in order to constrain $\delta \theta$. More precisely

$$||\delta \theta^{i}|| \leq ||C^{i}|| ||v^{i}||$$

(102)

with

$$|v^{ia}| = \frac{1}{N_{\mathcal{J}}^{\theta}} \int_{t, t'} dN^{a}_{\mathcal{J}} dN^{a}_{\mathcal{J}} g^{a}(t - t') (\xi^{ia}/\bar{\Lambda}^{ia} - 1)^2$$

$$\leq \frac{1}{N_{\mathcal{J}}^{\theta}} \int_{t, t'} dN^{a}_{\mathcal{J}} dN^{a}_{\mathcal{J}} g^{a}(t - t') (\lambda^{ia}/\bar{\Lambda}^{ia} - 1)^2$$

$$= k^{ia} - 2 \sum_{b=0}^{d} f^{ab} \theta^{ib}_{\text{MLE}} + \sum_{b, c=0}^{d} \theta^{ib}_{\text{MLE}} \theta^{ic}_{\text{MLE}} Q^{abc},$$

(103)

where we have used the relation

$$\lambda^{ia} = \sum_{a=0}^{d} \theta^{ia}_{\text{MLE}} \int dN^{a}_{\mathcal{J}} g^{a}(t - t').$$

(104)

Calculated at the saddle-point, and introduced a further auxiliary function

$$Q^{abc} = \frac{2}{(N_{\mathcal{J}}^{\theta})^3} \int_{t, t', t''} dN^{a}_{\mathcal{J}} dN^{b}_{\mathcal{J}} dN^{c}_{\mathcal{J}} g^{a}(t - t') g^{b}(t - t'') g^{c}(t - t'').$$

(105)
This allows to rewrite more compactly

\[ |v^{ij}| \leq \sum_{0 \leq b, c \leq \text{dp}} (\theta_{\text{MLE}}^{ab} - \Sigma^b) Q^{abc} (\theta_{\text{MLE}}^{bc} - \Sigma^c), \tag{106} \]

where \( \Sigma^i = [\tilde{\Lambda}^i \delta^{ai} \delta_{ai}]_{0 \leq a \leq \text{dp}} \) has been introduced in appendix C. The next step will be to perform a cumulant expansion of \( Q^{abc} \) so to take into account the different contributions to equation (106).

Before starting the inspection of the various cumulants of \( Q^{abc} \) one-by-one, let us expand first the higher order differentials

\[ \int_{t,t'} dN_t^a dN_t^b dN_t^c g^a(t - t') g^b(t - t'') g^c(t - t''), \]

by writing

\[ Q^{abc} = \frac{T^2}{(N_t^a)^3} \int_{t,t',t''=t'} dN_t^a dN_t^b dN_t^c g^a(t - t') g^b(t - t'') g^c(t - t''), \tag{107} \]

\[ + \delta^{abc}(1 - \delta^{ab}) \frac{T^2}{(N_t^a)^3} \int_{t,t',t''=t'} dN_t^a dN_t^b dN_t^c g^a(t - t') g^b(t - t'') g^c(t - t''), \]

\[ + \delta^{abc}(1 - \delta^{ab}) \frac{T^2}{(N_t^a)^3} \int_{t,t',t''=t'} dN_t^a dN_t^b dN_t^c g^a(t - t') g^b(t - t'') g^c(t - t''), \]

\[ + \delta^{abc} \delta^{abc}(1 - \delta^{ab}) \frac{T^2}{(N_t^a)^3} \int_{t,t',t''=t'} dN_t^a dN_t^b dN_t^c g^a(t - t') g^b(t - t'') g^c(t - t'). \]

\[ \]

\[ \]

\[ \]

\[ \]

\[ \]

\[ \]

\[ \]

\[ \]

\[ \]
which is expected to go to zero with $T$ independently of the MF approximation. The leading order term is the one associated with the second line of equation (110), as

- The first and the third line are associated with terms that go to zero with $T$ due to the presence of $\delta \tilde{\Lambda}^i$.
- The fourth line term is bound by a quantity the order of $p^2 [\theta_{\text{MLE}}^2]$, and hence is expected to be of the order of $p^2 ||c_{\text{max}}||^2$—as one can show by using the bounds proven in appendix E—where the functions $c_i(t)$ are the empirical connected cross-correlation functions, which are expected to be of order $d^{-1}$.
- The second line term is instead of the order of $d^{-1}$. Hence, if $||c_{\text{max}}|| \sim d^{-1}$, as one expects in the large dimensional setting (see e.g. Appendix A which proves this scaling in the homogeneous case), that is of the order of $d^{-1}$.

D.3. Higher order contributions

At the second order in the cumulants (i.e., terms containing the first power of the cross-correlation function $c(t)$), all terms but one contain contributions smaller than $\sim d^{-1}$. In fact, all terms but one do either

- contain one extra power of $||c_{\text{max}}||$,
- or they are proportional to $\delta \tilde{\Lambda}^i$.

The only term which can contribute at order $d^{-1}$ is the one associated with the first line of equation (110). Such term induces a contribution

$$
|v^{\text{st}}(\bar{c})| = \frac{\tilde{\Lambda}^a}{(\bar{\Lambda}^i)^2} \sum_{b,c=0}^d \theta_{\text{MLE}}^b \theta_{\text{MLE}}^c \int \text{d}t \text{d}t' c^{bc}(t - t') g^b(t) g^c(t') + \text{(higher order)}.
$$

Terms associated with cumulants of order larger than the second do not contribute at leading order. Hence, the dominant contribution is of the order of $d^{-1}$. Putting this information together, we obtain

$$
|v^{\text{st}}| \leq \frac{\tilde{\Lambda}^a}{(\bar{\Lambda}^i)^2} \left[ \sum_{b,c=0}^d \theta_{\text{MLE}}^b \theta_{\text{MLE}}^c \int \text{d}t \text{d}t' (c^{bc}(t - t') + \tilde{\Lambda}^b \delta^{bc} \delta(t - t')) g^b(t) g^c(t') \right] + \text{(higher order)}.
$$

That can be more compactly written as

$$
|v^{\text{st}}| \leq \frac{\tilde{\Lambda}^a}{(\bar{\Lambda}^i)^2} \mathcal{V}(\bar{\Lambda}_i^2) + \text{(higher order)}, \quad (111)
$$

where the variance corresponds to the empirical one computed under the saddle-point parameters $\theta_{\text{MLE}}$. The term in parenthesis corresponds exactly to the fluctuations of $N_t$, hence the error induced by the MF approximation is bound by the variance of the intensity $\Lambda$, as anticipated in the main text. We finally get:

$$
||\delta \theta|| \leq ||C|| ||\bar{\Lambda}|| \frac{\mathcal{V}(\bar{\Lambda}_i^2)}{(\bar{\Lambda}^i)^2} + \text{(higher order)} \quad (112)
$$

that is the formula used in section 3.3.
Appendix E. Cumulant expansion

In this appendix we will be interesting in providing \textit{a priori} bounds on the value of the inferred parameters \( \theta_{MF} \) and on their associated errors \( \delta \theta' \). We will be able to relate with quantities with the empirical cumulants of the process \( N_t \), thus allowing to obtain bounds on the error that do not require assumptions about the underlying model.

We start by recalling that after neglecting border effects (i.e., in the limit of large \( T \)), one has:

\[
h^{ia} = \tilde{\Lambda}^a, 
\]
\[
k^{ia} = \tilde{\Lambda}^a + \frac{1}{\Lambda^i} \int_0^\infty dt \, c^{ia}(t)g^a(t) ,
\]

\[
J_{iab} = \delta^{iab}(1 - \delta^{a0}) \left( \frac{\tilde{\Lambda}^a}{\Lambda} \int_0^\infty dt \, g^a(t)g^b(t) + \frac{1}{(\Lambda^i)^2} \int_0^\infty dt \, c^{ia}(t)g^a(t)g^b(t) \right) 
\]
\[
+ \frac{\tilde{\Lambda}^a}{\Lambda^i} \int_0^\infty dt \, c^{ib}(t)g^b(t) + \frac{\tilde{\Lambda}^b}{(\Lambda^i)^2} \int_0^\infty dt \, c^{ia}(t)g^a(t) 
\]
\[
+ \frac{1}{(\Lambda^i)^2} \int_0^\infty dt dt' \, K_{iab}(t, t')g^a(t)g^b(t'),
\]

where the vectors \( c^i(t) \) and the matrices \( K(t, t') \) correspond respectively to the second and the third empirical cumulant of the process \( N_t \), which for \( t, t' > 0 \) and \( a, b > 0 \) are defined as

\[
c^i(t) = \frac{1}{T} \sum_{m' < m} \delta^{i m'} \delta^{j m'} \delta(t - t_m + t_{m'}) - \tilde{\Lambda}^i \tilde{\Lambda}^j ,
\]

\[
K^{ij}(t, t') = \frac{1}{T} \sum_{m' = m < m''} \delta^{i m'} \delta^{j m''} \delta(t - t_m + t_{m'}) \delta(t' - t_m + t_{m''}) 
\]
\[
- \tilde{\Lambda}^i \tilde{\Lambda}^j - \tilde{\Lambda}^i c^{jk}(t - t') - \tilde{\Lambda}^j c^{ik}(t') - \tilde{\Lambda}^i \tilde{\Lambda}^j c^{ik}(t)
\]

while one obviously has \( c^{0i}(t) = K^{i00}(t, t') = K^{i00}(t, t') = 0 \). As in the last appendix, we write \( \nu^{ab} = \int_0^\infty dt \, g^a(t)g^b(t) \), and we will also employ the quantities

\[
\delta k^{ia} = k^{ia} - \tilde{\Lambda}^a ,
\]

\[
\delta J_{iab} = J_{iab} - \delta^{iab}(1 - \delta^{a0}) \frac{\tilde{\Lambda}^a}{\Lambda^i} \nu^{ab} - \frac{\tilde{\Lambda}^a \tilde{\Lambda}^b}{\Lambda^i} ,
\]

which are proportional to the empirical cumulants. The quantities \( \delta k^i \) and \( \delta J^i \) are in fact bound by

\[
|\delta k^{ia}| \leq \frac{g_{\text{max}} ||\tilde{\epsilon}_{\text{max}}||}{\lambda_{\text{min}}} ,
\]
\[ |\delta J^{ab}| \leq \delta^{ab}(1 - \delta^{ab}) \left( \frac{g_{\text{max}}^2 ||K_{\text{max}}||_1}{\hat{\lambda}_{\text{min}}^2} \right), \]  
(122)

\[ + 2 \left( \frac{g_{\text{max}} ||e_{\text{max}}||_1}{\hat{\lambda}_{\text{min}}^2} \right) + g_{\text{max}}^2 \left( \frac{||K_{\text{max}}||_1}{\hat{\lambda}_{\text{min}}^2} \right), \]

where \( \hat{\lambda}_{\text{min}} = \min_{1 \leq i, j \leq d} \hat{\lambda}_i \), \( g_{\text{max}} = \max_{a,d} g^a(t) \) and

\[ ||e_{\text{max}}||_1 = \max \int_0^\infty dt \, |e^i(t)| \]  
(123)

\[ ||K_{\text{max}}||_1 = \max \int_0^\infty dr dr' |K^{ijk}(t, r')|. \]  
(124)

By particularizing these inequalities to equation (111) one can obtain a bound on the MF error that is of the form

\[ |\nu^{ij}| \leq \frac{\hat{\lambda}_a}{(\hat{\lambda}_j)^2} \left( \frac{\nu_{\text{max}} \hat{\lambda}_{\text{max}} \max \sum_{b=0}^{dp} \theta^{ij}_{b,\text{MLE}} \sum_{b=0}^{dp} \theta^{ik}_{b,\text{MLE}} + \nu_{\text{max}} \hat{\lambda}_{\text{max}} \max \sum_{b=0}^{dp} \theta^{ij}_{b,\text{MLE}} \sum_{b=0}^{dp} \theta^{ik}_{b,\text{MLE}} \delta^{ij}. \right) \]  
(125)

By constructing the norm of the error, and comparing them with the bound on the inferred couplings equation (93), yields the expressions

\[ ||C^{ij}|| \leq \left( \frac{g_{\text{max}} ||e_{\text{max}}||_1}{\hat{\lambda}_{\text{min}}^2} \right), \]  
(126)

References

[1] Hawkes A G 1971 Point spectra of some mutually exciting point processes J. R. Stat. Soc. B 33 438–43
[2] Hawkes A G 1971 Spectra of some self-exciting and mutually exciting point processes Biometrika 58 83–90
[3] Ogata Y 1998 Space–time point-process models for earthquake occurrences Ann. Inst. Stat. Math. 50 379–402
[4] Bauwens L and Hautsch N 2009 Stochastic conditional intensities processes J. Financ. Economometrics 4 450–93
[5] Bowsher C G 2007 Modelling security market events in continuous time: intensity based, multivariate point process models J. Econometrics 141 876–912
[6] Bacry E, Mastromatteo I and Muzy J-F 2015 Hawkes processes in finance Market Microstructure Liquidity 01 1550005
[7] Reynaud-Bouret P, Rivoirard V and Tuleau-Malot C 2013 Inference of functional connectivity in neurosciences via Hawkes processes 1st IEEE Global Conf. on Signal and Information Proc. pp 317–20
[8] Mohler G O, Short M B, Brantingham P J, Schoenberg F P and Tita G E 2011 Self-exciting point process modeling of crime J. Am. Stat. Assoc. (doi:10.1198/jasa.2011.ap09546)
Blundell C, Beck J and Heller K A 2012 Modelling reciprocating relationships with Hawkes processes Adv. Neural Inf. Process. Syst. 2600–8

Farajtabar M, Du N, Gomez-Rodriguez M, Valera I, Zha H and Song L 2014 Shaping social activity by incentivizing users Adv. Neural Inf. Process. Syst. 2474–82

Linderman S W and Adams R P 2014 Discovering latent network structure in point process data arXiv:1402.0914

Iwata T, Shah A and Ghahramani Z 2013 Discovering latent influence in online social activities via shared cascade poisson processes Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (New York: ACM) pp 266–74

Zhou K, Zha H and Song L 2013 Learning triggering kernels for multi-dimensional Hawkes processes Proc. 30th Int. Conf. on Machine Learning (ICML-13) pp 1301–8

Tran L, Farajtabar M, Song L and Zha H 2015 Netcodec: community detection from individual activities SIAM Int. Conf. on Data Mining (2015) pp 91–99

Farajtabar M, Wang Y, Gomez-Rodriguez M, Li S, Zha H and Song L 2015 Coevolve: a joint point process model for information diffusion and network co-evolution 29th Ann. Conf. on Neural Information Processing Systems (NIPS) 2015 (arXiv:1507.02293) at press

Crane R and Sornette D 2008 Robust dynamic classes revealed by measuring the response function of a social system Proc. Natl. Acad. Sci. 105 15649–53

Ogata Y and Akaike H 1982 On linear intensity models for mixed doubly stochastic poisson and self-exciting point processes J. R. Stat. Soc. B 44 102–7

Emmanuel B, Khalil D and Jean-François M 2012 Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data Eur. Phys. J. B 85 187–199

Emmanuel B and Jean-François M 2014 Hawkes model for price and trades high-frequency dynamics Quant. Finance 14 1147–66

Emmanuel B and Jean-François M 2014 Second order statistics characterization of Hawkes processes and non-parametric estimation Trans. in Inf. Theory (arXiv:1401.0903) at press

Sylvain D, Nicolas F and Marc H 2014 High dimensional Hawkes processes arXiv:1403.5764

Daley D J and Vere-Jones D 1988 An Introduction to the Theory of Point Processes vol 2 (Berlin: Springer)

Lewis E and Mohler G 2011 A nonparametric em algorithm for multiscale Hawkes processes unpublished

Dong C L and Jorge N 1989 On the limited memory bfgs method for large scale optimization Math. Program. 45 503–28

John E D Jr and Jorge J M 1977 Quasi-Newton methods, motivation and theory SIAM Rev. 19 46–89

Robert T 1996 Regression shrinkage and selection via the lasso J. R. Stat. Soc. B 267–88

Feller W 1971 An Introduction to Probability Theory and its Applications vol II, 2nd edn (New York: Wiley)