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, 91128 Palaiseau, France
[2]Capital Fund Management, 21-23 Rue de l'Université, 75007 Paris, France
[3]Laboratoire Sciences Pour l'Environnement, CNRS, Université de Corse,
UMR 6134, 20250 Corté, France

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.

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, 5, 6], neuroscience [7], predictive policing [8] and social networks dynamics [9, 10, 11, 12, 13, 14, 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 explicitely [17] while alternative linear methods such as the contrast function 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* 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 Sec. 3 below). Note that this quadratic problem differs from the

usual least-squares objective for counting processes [7] (see Sec. 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 suffciently "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 our 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 Sec. 2 we formally introduce the Hawkes process and define the main notations of the paper. The mean-field inference method is defined In Sec. 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. This notably allows us to provide a quantitative measure of what weakly endogenous or "self-averaging" means for the interactions. In Sec. 4 we describe the implementation of the algorithm and compare it with other state-of-the-art algorithms. Sec. 5 shows the numerical results that we obtain with this method on synthetic data. Sec. 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 $\{(t_m, u_m)\}_{m=1}^n$, where t_m 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 $\boldsymbol{N}_t = [N_t^1, \ldots, N_t^d]^\top$ as $N_t^i = \sum_{m=1}^n \delta^{iu_m} \mathbb{1}_{t_m \leq t}$, where δ^{ij} indicates the Kronecker delta. These counting functions can be associated to a vector of stochastic intensities $\boldsymbol{\lambda}_t = [\lambda_t^1, \ldots, \lambda_t^d]^\top$ defined as

$$\lambda_t^i = \lim_{dt \to 0} \frac{\mathbb{P}\left(N_{t+dt}^i - N_t^i = 1 \middle| \mathcal{F}_t\right)}{dt} \tag{1}$$

where the filtration $\mathcal{F}_t$ encodes the information available up to time t. Then the process $\boldsymbol{N}_t$ is called a Hawkes process if the stochastic intensities can be written as

$$\lambda_t^i = \mu^i + \sum_{j=1}^d \int_0^t dN_{t'}^j \Phi^{ij}(t - t'), \tag{2}$$

where $\boldsymbol{\Phi}(t - t') = [\Phi^{ij}(t - t')]_{1 \leq i,j \leq d}$ 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 $\boldsymbol{\mu} = [\mu^i]_{i=1}^d$ is a vector of *positive* exogenous intensities (see Ref. [22] for a more rigorous definition). It is well known that a sufficient condition for the intensity processes λ_t^i to be stationnary (i.e., for the processes N_t^i to have stationnary increments) is the so-called **Stability Condition**

$$(\mathbf{SC}): \quad ||\boldsymbol{\Phi}||_1 < 1 \tag{3}$$

where $||.||_1$ stands for the spectral norm of the $d \times d$ matrix $\left[\int_0^{+\infty} \Phi^{ij}(t)\mathrm{d}t\right]_{1 \leq i,j \leq d}$ made of the L^1 norms of each kernel $\Phi^{ij}(t)$.

In the following discussion, we will restrict our attention to stable Hawkes processes (in the sense of (SC)) for which the matrix $\boldsymbol{\Phi}(t)$ can be written as

$$\Phi^{ij}(t) = \sum_{q=1}^{p} \alpha_q^{ij} g_q(t) \tag{4}$$

where we have introduced a set of $p > 0$ known *basis kernels* $g_q(t)$ satisfying

$$\int_0^\infty \mathrm{d}t \, g_q(t) = 1$$

.

2.2 Notations

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

- We will use the integer indexes $a \in \{1, \dots, pd\}$ in order to identify pairs $(j_a, q_a) \in \{1, \dots, d\} \times \{1, \dots p\}$. 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 $i, j, k \dots$ will be employed to label single nodes, the notation $a, b, c, \dots$ will be used to label a pair (node/basis kernel). In that respect we set

  - $\alpha^{ia} \equiv \alpha_{q_a}^{ij_a}$,
  - $g^a(t - t') \equiv g_{q_a}(t - t')$ and
  - $\mathrm{d}N_t^a \equiv \mathrm{d}N_t^{j_a}$.

- For convenience we will also define the deterministic process $N_t^0 = t$, which will be associated with the kernel $g^0(t - t') = \delta(t - t' - \mathrm{d}t)$ equal to a Dirac delta function shifted by an infinitesimal amount $\mathrm{d}t > 0$.

- According to previous notations, Eq. (2) can be compactly rewritten as

$$\lambda_t^i = \sum_{a=0}^{dp} \theta^{ia} \int_0^t \mathrm{d}N_{t'}^a g^a(t - t') \tag{5}$$

  where the parameters θ^{ia} refer to the Hawkes parameters, namely[1]

$$\theta^{i0} = \mu^i \quad \text{and} \quad \theta^{ia} = \alpha^{ia}, \quad \forall a > 0.$$

- We will adopt the notation x for scalars , $\boldsymbol{y} = [y^a]_{0 \leq a \leq dp}$ for vectors and $\mathbf{Z} = [Z^{ab}]_{0 \leq a,b \leq dp}$ for square matrices. Correspondingly, we define the usual matrix-vector composition by $\mathbf{Z}\boldsymbol{y} = [\sum_{b=0}^{dp} Z^{ab} y^b]_{0 \leq a \leq dp}$ and the matrix-matrix product as $\mathbf{W}\mathbf{Z} = [\sum_{b=0}^{dp} W^{ab} Z^{bc}]_{0 \leq a,c \leq dp}$.

- If $\boldsymbol{y}$ is a vector $||\boldsymbol{y}||$ will refer to the L^2 norm of $\boldsymbol{y}$ whereas if $\mathbf{Z}$ is a matrix, $||\mathbf{Z}||$ will refer to its spectral norm. Moreover if $\boldsymbol{\Phi}(t) = [\Phi^{ij}(t)]_{1 \leq i,j \leq d}$ is a matrix of functions, $||\boldsymbol{\Phi}||_1$ will refer to the spectral norm of the matrix $[\int_0^{+\infty} \mathrm{d}t \, |\Phi^{ij}(t)|]_{1 \leq i,j \leq d}$.

---

[1]Note that, in order for the parameters θ^{ia} to be dimensionally homogeneous, we need to assume a dimensionless μ, while $g^0(t - t')$ is taken to be of dimension $[t]^{-1}$. Moreover, we implicitly assume $N_t^0 = \omega t = t$ to be dimensionless through a suitable choice of a unit (i.e, we take $\omega = 1$).

- Hereafter, we will have to handle collections of d scalars, vectors or matrices indexed by $i \in \{1, \ldots, d\}$, that we will write as x^i (for scalars), $\boldsymbol{y}^i = [y^{ia}]_{0 \leq a \leq dp}$ (for vectors) and $\mathbf{Z}^i = [Z^{iab}]_{0 \leq a,b \leq dp}$ (for matrices).

According to these conventions, The goal of this paper is to present a mean-field framework for estimation of the d vectors of parameters $\boldsymbol{\theta}^i = [\theta^{ia}]_{0 \leq a \leq dp}$ given a set of observations N_t. In the following the set of all the $\boldsymbol{\theta}^i$ parameters will be often referred to as $\boldsymbol{\theta} = \{\boldsymbol{\theta}^i\}_{1 \leq i \leq d}$.

2.3 The Likelihood function

The probability for a Hawkes process parametrized by $\boldsymbol{\theta}$ to generate a trajectory N_t in a time interval $t \in [0, T]$ has been first computed in Ref. [17], and is given by

$$\mathbb{P}(\boldsymbol{N}_t | \boldsymbol{\theta}) = e^{-\sum_{i=1}^d \int_0^T \mathrm{d}t\, \lambda_t^i} \prod_{m=1}^n \lambda_{t_m}^{u_m} \tag{6}$$

which implicitly depends on $\boldsymbol{\theta}$ through the stochastic intensities λ_t^i. We define the negative log-likelihood

$$\mathcal{L}(\boldsymbol{N}_t, \boldsymbol{\theta}) = -\log \mathbb{P}(\boldsymbol{N}_t | \boldsymbol{\theta}). \tag{7}$$

Note that the maximum-likelihood estimator (that we will compare to our mean-field estimator to throughout this paper)

$$\boldsymbol{\theta}_{MLE} = \mathrm{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{N}_t, \boldsymbol{\theta}), \tag{8}$$

is the most commonly used estimator employed in the literature [17].

3 Inference and mean-field approximation

3.1 A Bayesian approach

From a Bayesian standpoint, the probability for an observed sample $\boldsymbol{N}_t$ to be generated by a Hawkes process parametrized by $\boldsymbol{\theta}$ is given by the *posterior distribution*

$$\mathbb{P}(\boldsymbol{\theta} | \boldsymbol{N}_t) = \frac{\mathbb{P}(\boldsymbol{N}_t | \boldsymbol{\theta}) \mathbb{P}_0(\boldsymbol{\theta})}{\mathbb{P}(N_t)} \tag{9}$$

where $\mathbb{P}_0(\boldsymbol{\theta})$ is a *prior* which we assume to be flat (see Sec. 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

$$\mathbb{E}\theta^{ia} = \int \mathrm{d}\boldsymbol{\theta}\, \mathbb{P}(\boldsymbol{\theta} | \boldsymbol{N}_t)\, \theta^{ia} \tag{10}$$

$$\mathbb{C}(\theta^{ia}, \theta^{jb}) = \int \mathrm{d}\boldsymbol{\theta}\, \mathbb{P}(\boldsymbol{\theta} | \boldsymbol{N}_t) \theta^{ia} \theta^{ib} - \mathbb{E}\theta^{ia} \mathbb{E}\theta^{jb}, \tag{11}$$

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

$$Z(\boldsymbol{s}) = \int \mathrm{d}\boldsymbol{\theta}\, \mathbb{P}(\boldsymbol{\theta} | \boldsymbol{N}_t) e^{-T \sum_{i=1}^d \boldsymbol{s}^{i\top} \boldsymbol{\theta}^i} \tag{12}$$

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

$$\mathbb{E}\theta^{ia} = \partial_{s^{ia}} f(\boldsymbol{s})|_{\boldsymbol{s}=0} \tag{13}$$

$$\mathbb{C}(\theta^{ia}, \theta^{jb}) = -\frac{1}{T}\partial_{s^{ia}}\partial_{s^{jb}} f(\boldsymbol{s})|_{\boldsymbol{s}=0} \tag{14}$$

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

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

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

$$Z(\boldsymbol{s}) = \int \mathrm{d}\boldsymbol{\theta} \, e^{-T \sum_{i=1}^{d} (\boldsymbol{s}^i + \boldsymbol{h})^\top \boldsymbol{\theta}^i + \sum_{m=1}^{n} \log \lambda_{t_m}^{u_m}} \tag{15}$$

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

$$h^a = \frac{1}{T}\int_0^T \mathrm{d}t \int_0^t \mathrm{d}N_{t'}^a g^a(t - t'). \tag{16}$$

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

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

$$Z(\boldsymbol{s}) = \int \mathrm{d}\boldsymbol{\theta} \, e^{\sum_{i=1}^{d} \left(-T(\boldsymbol{s}^i + \boldsymbol{h})^\top \boldsymbol{\theta}^i + \int_0^T dN_t^i \log \lambda_t^i\right)} = \prod_{i=1}^{d} \int \mathrm{d}\boldsymbol{\theta}^i e^{-T(\boldsymbol{s}^i + \boldsymbol{h})^\top \boldsymbol{\theta}^i + \int_0^T dN_t^i \log \lambda_t^i} \tag{17}$$

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 $\boldsymbol{\theta}^i$.

3.2 Mean-field approximation

Our strategy in order to approximate Eq. (15) is to suppose the fluctuations of the λ_t^i to be small with respect to their empirical averages

$$\bar{\boldsymbol{\Lambda}} = \{\bar{\Lambda}^i\}_{1 \le i \le d} = \left\{ \frac{N_T^i}{T} \right\}_{1 \le i \le d}. \tag{18}$$

This allows us to approximate the partition function appearing in Eq. (15) by a quadratic expansion of the logarithmic functions around the empirical averages $\bar{\boldsymbol{\Lambda}}$. 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 $[\lambda_t^i]_{1 \le i \le d}$ are stationary. The mean-field expansion basically assumes that each process λ_t^i has small fluctuations around its empirical averaged value $\bar{\Lambda}^i$. More precisely, let us suppose that the following **Mean-Field Hypothesis** holds

$$\textbf{(MFH) :} \qquad r_i = \frac{|\lambda_t^i - \bar{\Lambda}^i|}{\bar{\Lambda}^i} \ll 1, \tag{19}$$

then

$$\log \lambda_t^i \simeq \log \bar{\Lambda}^i + \frac{\lambda_t^i - \bar{\Lambda}^i}{\bar{\Lambda}^i} - \frac{(\lambda_t^i - \bar{\Lambda}^i)^2}{2(\bar{\Lambda}^i)^2} \ . \tag{20}$$

In order to get $Z(\boldsymbol{s})$ (see (17)), we need to compute $\int_0^T dN_t^i \log \lambda_t^i$ 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 λ_t^i. More precisley, we have:

- The constant term

$$\int_0^T \mathrm{d}N_t^i \log \bar{\Lambda}^i = T \bar{\Lambda}^i \log \bar{\Lambda}^i$$

  is independent of both $\boldsymbol{\theta}$ and $\boldsymbol{s}$, thus it can be dropped.[2]

- The linear term, after dropping $\boldsymbol{\theta}$ and $\boldsymbol{s}$-independent terms, leads to

$$\int_0^T \mathrm{d}N_t^i \frac{\lambda_t^i - \bar{\Lambda}^i}{\bar{\Lambda}^i} = T \sum_{a=0}^{dp} \theta^{ia} k^{ia}$$

  where we have introduced a collection of d vectors $\boldsymbol{k}^i = [k^{ia}]_{0 \leq a \leq dp}$ defined as:

$$k^{ia} = \frac{1}{N_T^i} \int_{t \neq t'} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a g^a(t - t') \tag{21}$$

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

$$-\frac{1}{2} \int_0^T \mathrm{d}N_t^i \left(\frac{\lambda_t^i - \bar{\Lambda}^i}{\bar{\Lambda}^i} \right)^2 = -\frac{T}{2} \sum_{a,b=0}^{dp} \theta^{ia} \theta^{ib} J^{iab} + T \sum_{a=0}^{dp} \theta^{ia} k^{ia},$$

  where we have introduced a collection of d matrices $\mathbf{J}^i = [J^{iab}]_{0 \leq a,b \leq dp}$ defined as:

$$J^{iab} = \frac{T}{(N_T^i)^2} \int_t \int_{t' \neq t} \int_{t'' \neq t} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \mathrm{d}N_{t''}^b g^a(t - t') g^b(t - t'') \ . \tag{22}$$

It then results from (17) that, up to a constant factor,

$$Z(\boldsymbol{s}) \simeq \int \mathrm{d}\boldsymbol{\theta} \, e^{-T \sum_{i=1}^d \left((\boldsymbol{s}^i + \boldsymbol{h} - 2\boldsymbol{k}^i)^\top \boldsymbol{\theta}^i + \frac{1}{2} \boldsymbol{\theta}^{i\top} \mathbf{J}^i \boldsymbol{\theta}^i \right)}. \tag{23}$$

The integral appearing in Eq. (23) can be evaluated analytically as a simple Gaussian integral[3], so that one can write the free-energy density as

$$f(\boldsymbol{s}) = -\frac{1}{2} \sum_{i=1}^d \left[(2\boldsymbol{k}^i - \boldsymbol{h}^i - \boldsymbol{s}^i)^\top \mathbf{C}^i (2\boldsymbol{k}^i - \boldsymbol{h}^i - \boldsymbol{s}^i) \right] , \tag{24}$$

where $\mathbf{C}^i$ denotes the inverse of matrix $\mathbf{J}^i$. Under this approximation, one obtains

$$\mathbb{E}\theta^{ia} \simeq \sum_{b=0}^{dp} C^{iab}(2k^{ib} - h^b) \equiv \theta_{MF}^{ia} \tag{25}$$

$$\mathbb{C}(\theta^{ia}, \theta^{jb}) \simeq \frac{\delta^{ij} C^{iab}}{T} \ . \tag{26}$$

---

[2]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 mean-field parameter.

[3]As the region of integration in $\mathrm{d}\boldsymbol{\theta}$ is the positive orthant $\mathbb{R}_+^{d(dp+1)}$, Eq. (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 $\mathbb{R}^{d(dp+1)}$, so to include potentially negative couplings. A saddle-point expansion of Eq. (23) reveals indeed that these differences are subleading in T.

These are the central equations of our paper: **Eq. (25) expresses the definition of our mean-field estimator** while Eq. (26) expresses the convergence rate to the expected value of the estimator, so that $\boldsymbol{\theta} - \mathbb{E}\boldsymbol{\theta} \sim T^{-1/2}$.

3.3 Validity conditions for the mean-field approximation

As discussed previously, the mean-field approximation is likely to be pertinent in the regime described by **(MFH)** (19), i.e., when the $\boldsymbol{\lambda}_t$ calculated under the inferred parameters have small fluctuations with respect to the empirical averages $\bar{\boldsymbol{\Lambda}}$. However the computation we performed in the previous section (relying on a second order truncation of $\log \lambda_t^i$) 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 mean-field estimates $\boldsymbol{\theta}_{MF}^i$ with respect to the maximum-likelihood estimates $\boldsymbol{\theta}_{MLE}^i$ (8)

$$\boldsymbol{\delta\theta}^i = \boldsymbol{\theta}_{MF}^i - \boldsymbol{\theta}_{MLE}^i \tag{27}$$

is directly bounded by the ratio of the empirical variance of λ_t^i to its empirical mean. In particular, at leading order in the fluctuations and in the limit of large T, one has

$$||\boldsymbol{\delta\theta}^i|| \leq \frac{\bar{\mathbb{V}}(\lambda_t^i)}{(\bar{\Lambda}^i)^2}||\mathbf{C}^i||||\bar{\boldsymbol{\Lambda}}|| \tag{28}$$

where $||.||$ refers to the L^2 norm in case of vectors and spectral norm in case of matrices, $\bar{\boldsymbol{\Lambda}} = [\bar{\Lambda}^a]_{a=0}^{dp}$ and where the variance corresponds to the *empirical* one computed under the saddle-point parameters $\boldsymbol{\theta}_{MLE}$. We see that when the *fluctuation ratio* $\bar{\mathbb{V}}(\lambda_t^i)/(\bar{\Lambda}^i)^2$ is very small, the results of the mean-field estimate can be very close to the maximum-likelihood estimate.

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 $||\boldsymbol{\Phi}||_1$ (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 Eq. (61)), we find that it is directly controlled by the spectral norm $||\boldsymbol{\Phi}||_1$, the system dimension d and the interaction characteristic time-scale τ_g (defined in Appendix A):

$$\frac{\mathbb{V}(\lambda_t^i)}{(\Lambda^i)^2} \sim \frac{1}{\Lambda^i \tau_g}\left(\frac{||\boldsymbol{\Phi}||_1^2}{d(1 - ||\boldsymbol{\Phi}||_1)^{2-1/\eta}}\right). \tag{29}$$

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

(i) *Weak endogeneity* ($||\boldsymbol{\Phi}||_1 \ll 1$): If the spectral norm $||\boldsymbol{\Phi}||_1$ (that controls the level of endogeneity of the Hawkes process) is much smaller than one, then the intensity vector λ_t 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. $||\boldsymbol{\Phi}||_1$ 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 Eq. (29) and has been studied notably in Ref. [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 $\boldsymbol{g}(t)$ is large with respect to the typical inter-event distance, then one expects the past contribution to the instantaneous intensity λ_t^i to have small fluctuations. This regime corresponds to a large $\bar{\Lambda}\tau_g$ in Eq. (29).
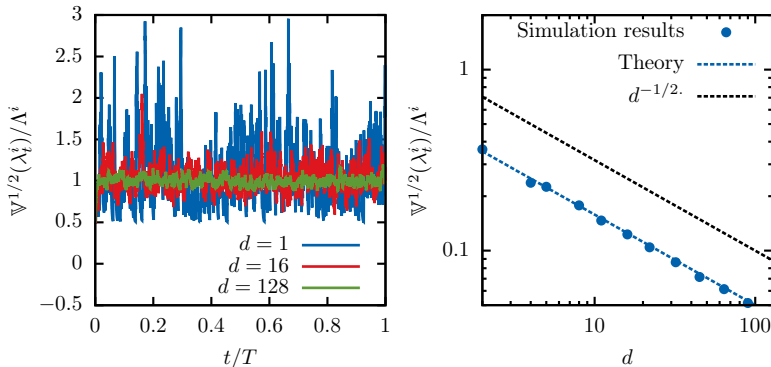

Figure 1: Fluctuation ratio $r_i = \bar{\mathbb{V}}^{1/2}(\lambda_t^i)/\bar{\Lambda}^i$ of the intensity of a simulated Hawkes process as a function of the number of components d for a fixed value of $\boldsymbol{\Lambda}$ and the interaction parameter $||\boldsymbol{\Phi}||_1 = 0.5$ against the theoretical prediction of Eq. (66). 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 Sec. 5, and are equal to $\beta = \mu = 1$. The interaction matrix has the two-block structure as illustrated in Sec. 5.

The perfect homogeneous case is a very good toy model, it has the main ingredients that make the mean-field hypothesis 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 2-block interaction matrix and exponential kernels fully specified in Sec. 5. Figs 1 and 2 show that the formula for the fluctuation ratio $r_i = \sqrt{\frac{\mathbb{V}(\lambda_t^i)}{(\Lambda^i)^2}}$ (see Eq. (29) or equivalently (61) for general kernel case and Eq. (66) for exponential kernels) lead to a perfect prediction of the simulated curve.

The right plot of Fig. 1 shows that the fluctuation ratio r_1 of the first block ($i = 1$) behaves in very good agreement with Eq. (66). It scales like $d^{-1/2}$ when the dimension varies. Fig. 2 shows the same ratio as a function of both the spectral norm $||\boldsymbol{\Phi}||_1$ and the dimension d. Let us point out that, when varying the norm $||\boldsymbol{\Phi}||_1$ and/or the dimension d, the average expectation $\boldsymbol{\Lambda}$ is held fixed. One can notably see that, both (i) for a fixed value of the dimension, there exists a sufficiently small interaction $||\boldsymbol{\Phi}||_1$ 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 (66) proved in Appendix A (see Eqs (61) and (66)). The theoretical prediction appears to be in good agreement with simulated data.

Finally, let us remark that while above considerations are useful for assessing the validity of the mean-field approximation under some specific assumption for the model,

it is possible to bound *a priori* the approximation error $\boldsymbol{\delta\theta}$ by exploiting the results of App. E, that allow one to estimate the mean-field error by supplying the empirical cumulants of the data, without requiring any assumption about the underlying model.
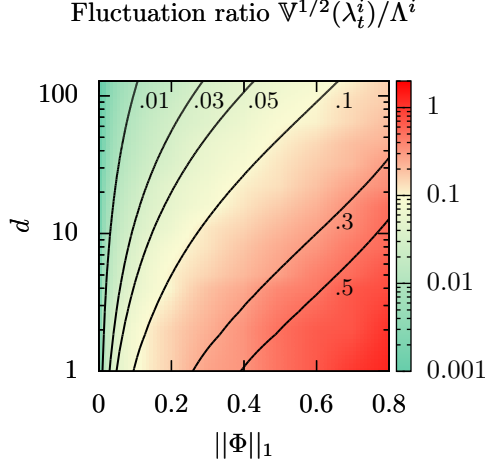

Fluctuation ratio $\mathbb{V}^{1/2}(\lambda_t^i)/\Lambda^i$

Figure 2: Fluctuation ratio $r_i = \bar{\mathbb{V}}^{1/2}(\lambda_t^i)/\bar{\Lambda}^i$ for a system with variable number of components $d = \{1, \ldots, 128\}$ and interaction strengths $||\boldsymbol{\Phi}||_1 \in [0, 0.8)$, given a fixed value of the intensities $\bar{\boldsymbol{\Lambda}}$. We have chosen $p = 1$ and $g_1(t) = \mathbb{1}_{t>0}\beta e^{-\beta t}$ (with $\beta = 1$) and chosen a two-block structure for the matrix $\boldsymbol{\alpha}_1$ with entries equal to either $\alpha_1^{ij} = 0$ or $\alpha_1^{ij} = ||\boldsymbol{\Phi}||_1/c$, with $c \in \{\lceil d/2 \rceil, \lfloor d/2 \rfloor\}$, as described in Sec. 5. The other simulation parameters are set to $\mu = 1 - ||\boldsymbol{\Phi}||_1$, $\beta_1 = 1$. The figure shows that larger values of $||\boldsymbol{\Phi}||_1$ 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 Eq. (66) correspond to the contour lines, which accurately reproduce the results of simulation.

Estimation error and validity condition Since it basically amounts in solving a linear system, it is clear that the mean-field estimator will perform much faster than any MLE based estimator (see next section for deatiled analysis on the complexity). The idea is to get, while being much faster, an estimation as precise as the MLE. This is true provided $||\boldsymbol{\delta\theta}^i||$ is smaller than the statistical error $||\boldsymbol{\Delta\theta}^i|| = ||\boldsymbol{\theta}_{MLE}^i - \boldsymbol{\theta}_{true}^i||$ 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 Eq. (26)):

$$||\boldsymbol{\Delta\theta}^i|| \simeq \left(\frac{\text{tr}(\mathbf{C}^i)}{T}\right)^{1/2},$$

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

$$||\boldsymbol{\delta\theta}^i|| < \left(\frac{\text{tr}(\mathbf{C}^i)}{T}\right)^{1/2}.$$

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

$$\frac{\mathbb{V}(\lambda_t^i)}{(\Lambda^i)^2}||\mathbf{C}^i||||\mathbf{\Lambda}|| \lesssim \left(\frac{\text{tr}(\mathbf{C}^i)}{T}\right)^{1/2}.$$

Eq. (29) has been established in the perfectly homogeneous case for which $||\mathbf{\Lambda}|| = \sqrt{d}\Lambda^i$. Moreover, we use very conservatively the inequality $||\mathbf{C}^i|| \leq \text{tr}(\mathbf{C}^i)$, which is likely to be largely underestimating the statistical error, so that in practice the performance of the MF algorithm is much better than what appears from Eq. (30) below (see Fig. 4). Indeed, plugging (29) and assuming the scaling $||\mathbf{C}^i|| \sim (\Lambda^i)^{-1}$ in the last equation leads to the sufficient condition

$$\frac{1}{\sqrt{\Lambda^i}\tau_g}\left(\frac{||\mathbf{\Phi}||_1^2}{d^{1/2}(1-||\mathbf{\Phi}||_1)^{2-1/\eta}}\right) \lesssim \left(\frac{1}{T}\right)^{1/2}.$$

Therefore

$$T \lesssim T^\star = \Lambda^i\tau_g^2\left(\frac{d(1-||\mathbf{\Phi}||_1)^{4-2/\eta}}{||\mathbf{\Phi}||_1^4}\right). \tag{30}$$

Although far from being optimal, this rough bound provides a qualitative idea of how the different values of d, $||\mathbf{\Phi}||_1$ and τ_g can contribute in determining the performance of the mean-field approximation. We will see in Sec. 5, that in practive 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 mean-field method is comparable to those of the maximum likelihood are very easy to meet, in particular for large system sizes. In Sec. 5 we will illustrate this point by showing that in a broad range of regimes of $(||\mathbf{\Phi}||_1, d, T)$ we will be able to exploit the condition $T < T^\star$, thus outperforming the maximum-likelihood algorithm through the superior numerical performance of the mean-field approximation.

4 Implementation and complexity

In this section we want to illustrate the numerical implementation of the mean-field 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 : the dimension of the Hawkes process

- p : the number of basis kernels

- Λ : the maximum average intensity over all the components ($\Lambda = \sup_i \Lambda^i$).

- T : the sample size (expressed in physical time)

- t : 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 Ref. [17], builds upon the direct maximization of the likelihood function $\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{N}_t)$. 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

$$-\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{N}_t) = \mathcal{H}(\boldsymbol{\theta}) = \sum_{i=1}^{d} \boldsymbol{h}^\top \boldsymbol{\theta} - \sum_{m=1}^{n} \log \sum_{a=0}^{dp} \theta^{u_m a} G^{ma} \, . \tag{31}$$

where

$$G^{ma} = \int_0^T \mathrm{d}t \, g^a(t_m - t) \tag{32}$$

One can reduce the cost of each interation by precomputing the coefficients **G** on the right-hand-side of Eq. (31).

The corresponding complexity is $O(d^2 p \Lambda^2 T t)$. Once done, the complexity for computing the log likelihood (or its gradient) is dominated by the rightest term in (31) and is of the order of $O(d^2 p \Lambda T)$. If the overall gradient algorithm involves R iterations, the overall complexity writes

$$\text{MLE-Complexity} \sim d^2 p \Lambda^2 T t + R d^2 p \Lambda T. \tag{33}$$

Let us point out that if R is large, its corresponding term dominates. This is typically the case, as seen in Fig. 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 constrast function approach proposed in [7]. In this framework, the $\boldsymbol{\theta}$ should minimize a loss function defined as

$$\mathcal{C}(\boldsymbol{\theta}) = \frac{1}{2} \sum_i \int_0^T \mathrm{d}t \left(\frac{\mathrm{d}N_t^i}{\mathrm{d}t} - \lambda_t^i \right)^2 , \tag{34}$$

resulting in

$$\mathcal{C}(\boldsymbol{\theta}) = T \left(\sum_{i=1}^{d} (-\bar{\Lambda}^i \boldsymbol{k}^{i\top} \boldsymbol{\theta}^i + \frac{1}{2} \boldsymbol{\theta}^{i\top} \mathbf{J}' \boldsymbol{\theta}^i) \right) , \tag{35}$$

where the collection $\boldsymbol{k}^i$ is defined as in Eq. (21), while the matrix $\mathbf{J}'$ is given by

$$J'^{ab} = \frac{1}{T} \int_0^T \mathrm{d}t \, \mathrm{d}N_{t'}^a \, \mathrm{d}N_{t''}^b \, g^a(t - t') g^b(t - t'') \, . \tag{36}$$

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

$$\boldsymbol{\theta}^i = \bar{\Lambda}^i \mathbf{C}' \boldsymbol{k}^i \tag{37}$$

where $\mathbf{C'} = (\mathbf{J'})^{-1}$. This solution is numerically close to Eq. (25). The computation of the $\mathbf{J'}$ has a complexity of $O(d^2 p^2 \Lambda^2 T t)$ (we cannot use the trick illustrated in App. B for the computation of $\mathbf{J}^i$). 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

$$\text{CF-Complexity} \sim d^2 p^2 \Lambda^2 T t. \tag{38}$$

4.4   Mean-field (MF) algorithm

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

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

Input: $\boldsymbol{N}_t, \boldsymbol{g}(t)$
Output: $\mathbb{E}\theta^{ia}, \mathbb{C}(\theta^{ia}, \theta^{jb})$
  1: Compute the auxiliary functions $\boldsymbol{h}, \boldsymbol{k}^i$ and $\mathbf{J}^i$ as shown in Appendix B
  2: Invert the d matrices $\mathbf{J}^i$
  3: Evaluate Eqs. (25) and (26)
  4: **return** $\mathbb{E}\theta^{ia}$, $\mathbb{C}(\theta^{ia}, \theta^{jb})$

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

$$\text{MF-Complexity} \sim d^2 p \Lambda T \times \max(\Lambda t, dp). \tag{39}$$

The inversion of the matrices can be performed with a complexity bounded by $O(d^4 p^3)$. 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 $\mathbb{C}(\theta^{ia}, \theta^{ib})$, the inversion of the matrix can be avoided, by replacing it with the solution of the d linear systems $\mathbf{J}^i \boldsymbol{\theta}^i_{MF} = 2\boldsymbol{k}^i - \boldsymbol{h}$. This substantially reduces the time required to find $\boldsymbol{\theta}_{MF}$ 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 $\mathbf{J}^i$, as shown in Appendix C, reducing the time required for the solution of the d linear systems from $O(d^4 p^3)$ to $O(d^3 p^2)$ (corresponding to d matrix-vector compositions). Note indeed that this is expected to yield inaccurate results when the fluctuations of the $\boldsymbol{N}_t$ are of a larger order than d^{-1}, even though the elements of the $\boldsymbol{\alpha}$ 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 (33), (38) and (39), 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 Fig. 6). Our numerical simulations show that R depends on all the other factors (notably Λ 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 Λt 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 α, β 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.

Single exponential basis kernel First, we have considered as a benchmark the case in which $p = 1$, so that $g^a(t) = g_1(t) \equiv g(t)$. For $j > 0$, we have assumed the basis kernel to have the exponential structure

$$g(t) = \beta e^{-\beta t} \mathbb{1}_{t>0} \,. \tag{40}$$

The topology of the matrix $\boldsymbol{\alpha}$ 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 α^{ij} is zero (if i and j do not belong to the same cluster) or $\alpha^{ij} = \alpha/c$ (if they do belong to the same cluster, with size given by c);

Regardless of the structure, in both cases one has $||\boldsymbol{\Phi}||_1 = ||\boldsymbol{\alpha}||_1 = \alpha$, 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 Fig. 3 the results of our algorithm in the case of a three block structure for the $\boldsymbol{\alpha}$ matrix. The results are for this choice of value essentially the same as the ones obtained by maximum likelihood estimation, with a considerable reduction in the computational time.

In order to systematically assess the performance of the algorithm under various circumstances, we have varied α 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 specifed otherwise throughout the text, we have assumed block structures with two clusters, so that $c \in \{\lceil d/2 \rceil, \lfloor d/2 \rfloor\}$. We measured the quality of the reconstruction by using

- The negative log-likelihood of the sample $\mathcal{L}(\boldsymbol{N}_t, \boldsymbol{\theta}) = -\log \mathbb{P}(\boldsymbol{N}_t|\boldsymbol{\theta}_{inf})$ under the inferred couplings

- The relative error on the non-zero couplings

$$\delta \alpha_{rel}^2 = \sum_{i,j|\alpha_{true}^{ij} \neq 0} \left(\frac{\alpha_{inf}^{ij}}{\alpha_{true}^{ij}} - 1 \right)^2 \,. \tag{41}$$
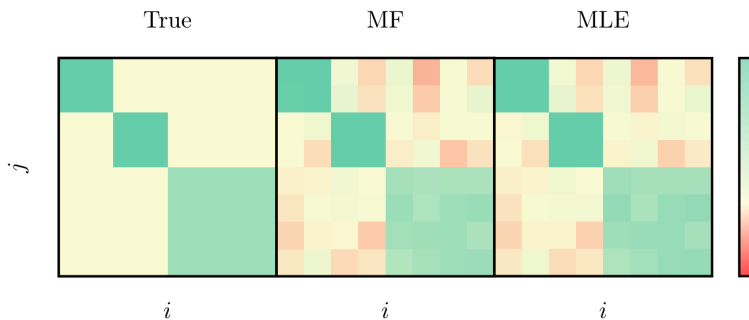
Figure 3: Illustrative plot of the results of the reconstruction of a Hawkes model with a three-block structure for the matrix $\boldsymbol{\alpha}$. We used the parameters $||\boldsymbol{\Phi}||_1 = 0.5$, $\mu = 1$ and $\beta = 1$, and ran the simulation for $T = 10^4$. The true underlying matrix (left) is reconstructed both with the MF algorithm (center) and with a maximum likelihood estimation (right), yielding very similar results.

- The absolute error on all the couplings

$$\delta\alpha_{abs}^2 = \sum_{i,j} \left(\alpha_{inf}^{ij} - \alpha_{true}^{ij} \right)^2 . \tag{42}$$

The results that we have found for the relative error $\delta\alpha_{rel}$ in the two-block case are summarized in Fig. 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 $T_\star$, beyond which the statistical error no longer dominates $\delta\alpha_{rel}$ for the mean-field 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 $T_\star$ is quite large, indicating that the MF approximation is extremely effective in a broad range of regimes. By increasing the value of $||\boldsymbol{\Phi}||_1 = \alpha$, even though the relative error decreases due to the increase of the signal α, the quality of the MF approximation decreases, and the value of $T_\star$ becomes smaller. Indeed, for a fixed value of T and $||\boldsymbol{\Phi}||_1$, it is always possible to find a sufficiently large d so that $T \ll T_\star$, and the MF approximation is effective. This is confirmed in Fig. 5, where we have represented the absolute error $\delta\alpha_{abs}$ obtained by MLE against the one estimated by using Eq. 26.

On the other hand, the computational time required in order to run the MF algorithm is considerably smaller, as summarized in Fig. 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 $\mathcal{L}(\boldsymbol{N_t}, \boldsymbol{\theta})$, 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 the asymptotic one got from a BFGS minimization.
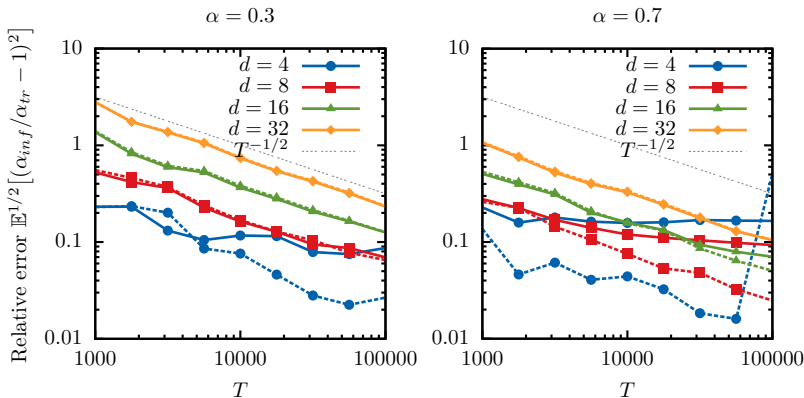
Figure 4: Relative error on the non-zero couplings $\delta\alpha_{rel}$ as a function of the sample size T for the maximum likelihood estimator (dashed lines) and the MF approximation (solid lines) for a fixed value of the empirical intensity $\bar{\Lambda}$ in the two-block case as described in Sec. 5 for various values of α and d. The other paramters are set to $\mu = \beta = 1$. The figure shows that even for moderate values of d the error related to the MF approximation is much smaller than the statistical error in the estimation of the couplings induced by the finite value of T affecting the maximum likelihood estimator. By comparing the left and the right panel, one sees that large values of $||\boldsymbol{\Phi}||_1 = \alpha$ worsen the quality of the MF algorithm.

Multiple exponential basis kernels We have also considered the more general case in which the $\boldsymbol{\Phi}(t)$ 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 $\boldsymbol{\alpha}_1$ has a uniform two-block structure as in the previous case, while for the second matrix $\boldsymbol{\alpha}_2$ only the complementary blocks are uniformly filled (see Fig. 7). For both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, the non-zero entries have value α/c, being c the size of the block. Finally, we chose $\beta_1^{ij} = 1$ and $\beta_2^{ij} > 1$, so that for $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ one recovers a completely homogeneous interaction structure. Fig. 7 shows an illustrative plot of the results of the reconstruction in the multiple basis kernel scenario.

The role of β We have also studied the effect of the misspecification of β by studying the effect of an input kernel $\boldsymbol{g}_{in}(t)$ different from the $\boldsymbol{g}_{true}(t)$ used to generate the data. We did it in the case of $p = 1$, with a block $\boldsymbol{\alpha}$ matrix as described in the paragraph above relating to the single exponential basis kernel scenario. Also in this case we have chosen $\mu = 1 = \beta_{true} = 1$, with β_{true} denoting the true and unique decay constant for the exponential kernels, whereas β_{in} will denote the decay constant chosen to perform the inference. Our results are summarized in Fig. 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 β, as the inferred intensity turns out to be very weakly dependent on β_{in}. What actually happens is that by changing β_{in} is possible to tune a bias changing $\boldsymbol{\mu}$ in favor of $\boldsymbol{\alpha}$. This behavior can be understood in the framework of our MF approximation:

- One one hand, App. C shows that the value of the inferred couplings $\boldsymbol{\alpha}$ is pro-
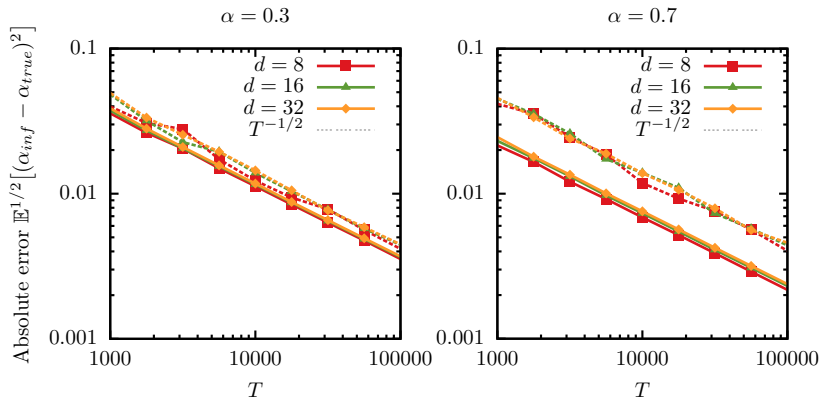
15

Figure 5: Absolute error couplings $\delta\alpha_{abs}$ as a function of the sample size T for the maximum likelihood estimator (dashed lines) and the estimation of the error provided by Eq. (26) (solid lines), for a fixed value of the empirical intensity $\bar{\Lambda}$ 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 $||\mathbf{\Phi}||_1$ (left figure) while the MLE error is underestimated by the MF error when approaching the critical point $||\mathbf{\Phi}||_1 = 1$ (right figure).

portional to the fluctuations of the auxiliary function $\boldsymbol{\delta k}$. According to Eq. (102), these fluctuations are obtained by integrating the empirical correlation function against the kernel $\boldsymbol{g}_{in}(t)$ in order to obtain $\boldsymbol{\delta k}$. Hence, when β_{in} is large $\boldsymbol{g}_{in}(t)$ is peaked close to the origin, where the correlations are larger, resulting in a greater value for $\boldsymbol{\delta k}$. Conversely, for large β_{in} such integral assumes a small value, because the correlations are small at the timescales in which $\boldsymbol{g}_{in}(t)$ has large mass. In any case, the norm of $\boldsymbol{\delta k}$ is **bounded from above**, and is maximum for β_{in} large.

- On the other hand, the entries of C decrease with β_{in}. For example, Eq. (82) shows that all the entries but the component $a = b = 0$ of the first order contribution C_0^{iab} to C^{iab} have a β_{in}^{-1} dependence. Hence, in this case they **decrease unboundedly** with β_{in}.

The latter effect is dominant in our numerical examples, so that a small value of β_{in} leads to a bias in favor of $\boldsymbol{\alpha}$, while a large value favors $\boldsymbol{\mu}$. Independently of β_{in}, the combination of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ leading to the inferred value of $\boldsymbol{\Lambda}$ 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 $\boldsymbol{g}_{in}$, 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
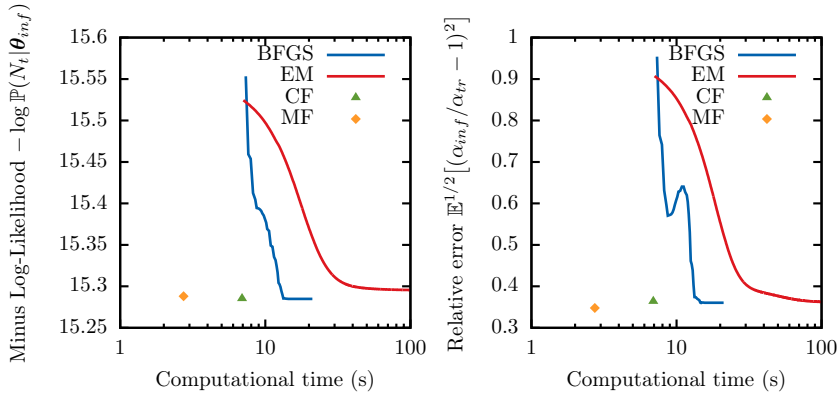
Figure 6: Time required by the learning algorithms described in Sec. 4 in order to achieve fixed values of accuracy, as measured by the negative log-likelihood (left panel) and the relative error (right panel). We have considered the two block case described in Sec. 5, and adopted the same parameters as in Fig. 4, taking in particular a dimension equal to $d = 25$.

extremely efficients techniques are available [25, 24]. 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 non-linear Hawkes processes. Another prospect of particular interest is the case in which a non trivial prior is added to the Bayesian formula Eq. (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

$$\mathbb{P}_0^{L^2}(\boldsymbol{\theta}) \propto \exp\left(\sum_{ia} \kappa^{ia}(\theta^{ia})^2\right) \tag{43}$$

and the Laplace prior

$$\mathbb{P}_0^{L^1}(\boldsymbol{\theta}) \propto \exp\left(\sum_{ia} \kappa^{ia}|\theta^{ia}|\right) . \tag{44}$$

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 mean-field 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

$$\mathcal{H}^{L^2}(\boldsymbol{\theta}) = \sum_{i=1}^{d} \left(\frac{1}{2}\boldsymbol{\theta}^{i\top}\mathbf{J}^i\boldsymbol{\theta}^i - (2\boldsymbol{k}^i - \boldsymbol{h})^\top\boldsymbol{\theta}^i\right) + \sum_{ia} \kappa^{ia}(\theta^{ia})^2 \tag{45}$$
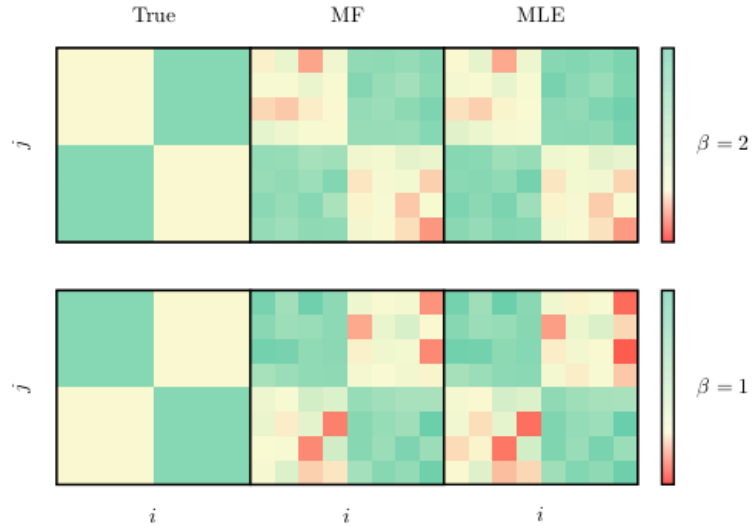
Figure 7: Reconstruction of a Hawkes process in which two different exponential kernels are mixed together. In this example we have chosen $d = 8, p = 2$, and we have chosen the matrices $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ as in the left panels of the figure. They both have a block structure with uniform entries equal to either zero or $\alpha = 0.5/c$, where $c = 4$ is the connectivity of the block. The other parameters are set to $\mu = 1$, $\beta_1 = 1$, $\beta_2 = 2$ and $T = 10^5$. Also in this case our MF algorithm (central panels) is able to reconstruct the interaction network approximately as well as the maximum likelihood estimators (right panels).

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

$$\mathcal{H}^{L^1}(\boldsymbol{\theta}) = \sum_{i=1}^{d} \left(\frac{1}{2} \boldsymbol{\theta}^{i\top} \mathbf{J}^i \boldsymbol{\theta}^i - (2\boldsymbol{k}^i - \boldsymbol{h})^{\top} \boldsymbol{\theta}^i \right) + \sum_{ia} \kappa^{ia} |\theta^{ia}| \tag{46}$$

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 mean-field method can provide an easy and efficient way to set a good starting point for standard iterative algorithms that maximize the likelihood objective.
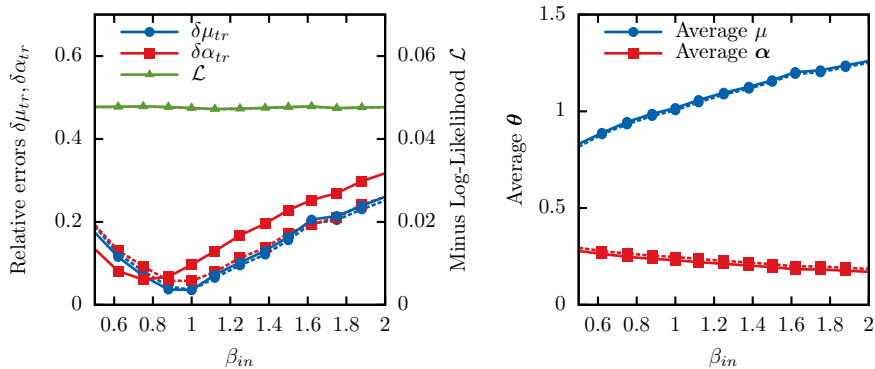
Acknowledgements

Figure 8: Effect of the mispecification of β on the quality of inference. We simulated a two-block structure for the matrix $\boldsymbol{\alpha}$ as described in Sec. 5, with $d = 8$, $\mu = \beta_{true} = 1$ and $T = 10^5$, where β_{true} denotes the true value of the decay constant β. We subsequently inferred the parameters μ and $\boldsymbol{\alpha}$ under the maximum likelihood estimator (dashed line) and the MF one (solid line) under various values for the decay constant β_{in} of the exponential kernel $\boldsymbol{g}_{in}(t)$. We find that, even though the likelihood is barely changed by the shift in β_{in} (left panel), its change reflects in a bias in the inferred couplings (right panel).

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 $\boldsymbol{\lambda}_t$ 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 Sec. 3.3) can be detected. In order to do this, we study the quantity

$$r_i = \sqrt{\frac{\mathbb{V}(\lambda_t^i)}{\mathbb{E}^2(\lambda_t^i)}}\,, \tag{47}$$

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

$$\boldsymbol{\mu} \;=\; \mu\,[1, \ldots, 1] \tag{48}$$
$$\boldsymbol{\Phi}(t) \;=\; \Phi(t)\,[1]_{1 \le i,j \le d} \tag{49}$$

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 Sec. 3.3, one should keep in mind that we are replacing all the empirical averages by expectations.

1. Mean of $\boldsymbol{\lambda}_t$ The mean value is easy to compute once one particularizes the results of [1, 2] in the homogeneous case, obtaining

$$\boldsymbol{\Lambda} = \mathbb{E}\boldsymbol{\lambda}_t = \left(\mathbb{I} - \left[\int \Phi^{ij}(t)\mathrm{d}t \right]_{1 \leq i,j \leq d} \right)^{-1} \boldsymbol{\mu} \tag{50}$$

$$= \left(\mathbb{I} - ||\Phi||_1 [1]_{1 \leq i,j \leq d} \right)^{-1} \boldsymbol{\mu} \tag{51}$$

$$= \frac{\mu}{1 - d||\Phi||_1}[1]_{1 \leq i \leq d}, \tag{52}$$

$$\tag{53}$$

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

$$\Phi(t) = g(t)||\boldsymbol{\Phi}||_1/d, \quad \text{with} \ \ ||g||_1 = 1.$$

In that case, one gets

$$\boldsymbol{\Lambda} = \left[\frac{\mu}{1 - ||\boldsymbol{\Phi}||_1} \right]_{1 \leq i \leq d} = [\Lambda]_{1 \leq i \leq d} \tag{54}$$

2. Variance of $\boldsymbol{\lambda}_t$ The variance can be obtained by using the stationary version of the martingale representation for the Hawkes process (see [18])

$$\boldsymbol{\lambda}_t = \boldsymbol{\Lambda} + \int_{-\infty}^t \boldsymbol{\Psi}(t - t') \, \mathrm{d}\boldsymbol{M}_{t'} \,, \tag{55}$$

where $\boldsymbol{M}_t$ is the martingale $\boldsymbol{M}_t = \boldsymbol{N}_t - \int_0^t \boldsymbol{\lambda}_t \mathrm{d}t$, (one has $\mathbb{E}[\mathrm{d}\boldsymbol{M}_t|\mathcal{F}_t] = 0$), and $\boldsymbol{\Psi}(t)$ can be obtained from the Fourier transform of $\boldsymbol{\Phi}(t)$, that we denote as $\hat{\boldsymbol{\Phi}}(\omega) = \int \mathrm{d}t\boldsymbol{\Phi}(t)e^{-i\omega t}$:

$$\hat{\boldsymbol{\Psi}}(\omega) = (\mathbb{I} - \hat{\boldsymbol{\Phi}}(\omega))^{-1} - \mathbb{I}, \tag{56}$$

as shown in [1, 2] and generalized in [18]. Then one can write

$$\mathbb{V}(\lambda_t^i) = \sum_{1 \leq j,k \leq d} \mathbb{E}\left[\int_{-\infty}^t \mathrm{d}M_s^j \int_{-\infty}^t \mathrm{d}M_{s'}^k \Psi^{ij}(t-s)\Psi^{ik}(t-s') \right] \tag{57}$$

$$= \sum_{j=1}^d \Lambda^j \int_0^\infty \mathrm{d}t \, (\Psi^{ij})^2(t) = d\Lambda ||\Psi||^2 = \frac{d\Lambda}{2\pi}||\hat{\Psi}||^2, \tag{58}$$

where $||.||$ stands for the L^2 norm. However, since

$$\hat{\boldsymbol{\Psi}}(\omega) = \left(\mathbb{I} - \frac{\hat{g}(\omega)||\boldsymbol{\Phi}||_1}{d}[1]_{1 \leq i,j \leq d} \right)^{-1} - \mathbb{I} = \frac{1}{d}\left(\frac{\hat{g}(\omega)||\boldsymbol{\Phi}||_1}{1 - \hat{g}(\omega)||\boldsymbol{\Phi}||_1} \right) [1]_{1 \leq i,j \leq d},$$

one has

$$\hat{\Psi}(\omega) = \frac{1}{d}\left(\frac{\hat{g}(\omega)||\boldsymbol{\Phi}||_1}{1 - \hat{g}(\omega)||\boldsymbol{\Phi}||_1} \right).$$

Thus

$$\mathbb{V}(\lambda_t^i) = \frac{\Lambda ||\boldsymbol{\Phi}||_1^2}{d} \int \frac{\mathrm{d}\omega}{2\pi}\left(\frac{|\hat{g}(\omega)|^2}{|1 - ||\boldsymbol{\Phi}||_1 \hat{g}(\omega)|^2} \right). \tag{59}$$

So the scaling in either limit $||\boldsymbol{\Phi}||_1 \to 0$ (case (i) in Sec. 3.3) or $d \to +\infty$ (case (ii) in Sec. 3.3) is thus of the form $\mathbb{V}(\lambda_t^i) \sim \Lambda ||\boldsymbol{\Phi}||_1^2/d$. Let's try to be more accurate and

20

characterize the scaling when approching the critical case $||\mathbf{\Phi}||_1 \to 1$. 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 Theorem 3 and Theorem 4 in Sec. XIII.5 of [27]), one can prove that $\eta = \gamma + 1$. Let's set $\epsilon = 1 - ||\mathbf{\Phi}||_1$. Then, focusing on $|\omega| < a$ (a fixed), the integral in (59) becomes

$$
\begin{aligned}
\int_{|\omega|<a} \frac{d\omega}{2\pi} \frac{|\hat{g}(\omega)|^2}{|1 - (1-\epsilon)\hat{g}(\omega)|^2} &\simeq \int_{|\omega|<a} \frac{d\omega}{2\pi} \left(\frac{1}{|1 - (1-\epsilon)(1+K\omega^\eta)|^2} \right) \\
&\simeq \int_{|\omega|<a} \frac{d\omega}{2\pi} \left(\frac{1}{\epsilon^2 + \Re(K)^2\omega^{2\eta} - 2\Re(K)\epsilon\omega^\eta + \Im(K)^2\omega^{2\eta}} \right) \\
&\simeq \epsilon^{1/\eta-2} \int_{|\omega|<a\epsilon^{-1/\eta}} \frac{d\omega}{2\pi} \left(\frac{1}{1 + \Re(K)^2\omega^{2\eta} - 2\Re(K)\omega^\eta + \Im(K)^2\omega^{2\eta}} \right) \\
&\sim \epsilon^{1/\eta-2}
\end{aligned}
$$

This last result along with (59) gives that

$$\mathbb{V}(\lambda_t^i) = \frac{\Lambda}{\tau_g(||\mathbf{\Phi}||_1)} \left(\frac{||\mathbf{\Phi}||_1^2}{d(1 - ||\mathbf{\Phi}||_1)^{2-1/\eta}} \right), \tag{60}$$

where

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

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 $\tau_g(||\mathbf{\Phi}||_1)$ can be interpreted as a characteristic time scale of the kernel.

3. Fluctuations of $\boldsymbol{\lambda}_t$ and (MFH) Thus, the fluctuations of $\boldsymbol{\lambda}_t$ behave, in either limit (i) weak interactions ($||\mathbf{\Phi}||_1 \ll 1$, case (i) in Sec. 3.3) or (ii) large dimension ($d \gg 1$ case (ii) in Sec. 3.3), like

$$\frac{\mathbb{V}(\lambda_t^i)}{\mathbb{E}^2(\lambda_t^i)} \sim \frac{1}{\Lambda\tau_g} \left(\frac{||\mathbf{\Phi}||_1^2}{d(1 - ||\mathbf{\Phi}||_1)^{2-1/\eta}} \right), \tag{61}$$

and the mean-field hypothesis simply writes

$$(\textbf{MFH}) \ : \quad r_i = \sqrt{\frac{\mathbb{V}(\lambda_t^i)}{\mathbb{E}^2(\lambda_t^i)}} \sim \frac{||\mathbf{\Phi}||_1}{\sqrt{d\Lambda\tau_g}(1 - ||\mathbf{\Phi}||_1)^{1-1/2\eta}} \ll 1, \tag{62}$$

4. Exponential case In the simpler case in which

$$g(t) = \beta e^{-\beta t} \mathbb{1}_{t>0} \tag{63}$$

we get

$$\hat{g}(\omega) = \frac{1}{1 + i\omega/\beta} \quad \text{and} \quad \eta = 1, \tag{64}$$

which implies

$$\tau_g(||\mathbf{\Phi}||_1) = \frac{2}{\beta}. \tag{65}$$

We see here that τ_g does correspond to the characteristic time scale of the kernel. The mean-field hypothesis simply writes

$$(\textbf{MFH}) \;:\;\;\; r_i = \sqrt{\frac{\mathbb{V}(\lambda_t^i)}{\mathbb{E}^2(\lambda_t^i)}} \sim \frac{||\boldsymbol{\Phi}||_1 \sqrt{\beta}}{\sqrt{2d\Lambda(1 - ||\boldsymbol{\Phi}||_1)}} \ll 1 \,. \tag{66}$$

This result is illustrated in Fig. 1.

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

$$h^a \;=\; \frac{1}{T} \int_0^T \mathrm{d}t \int_0^t \mathrm{d}N_{t'}^a g^a(t - t') \tag{67}$$

$$k^{ia} \;=\; \frac{1}{N_T^i} \int_{t \neq t'} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a g^a(t - t') \tag{68}$$

$$J^{iab} \;=\; \frac{T}{(N_T^i)^2} \int_{t \neq t', t''} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \mathrm{d}N_{t''}^b \tag{69}$$

$$\times \;\; g^a(t - t') g^b(t - t'') \tag{70}$$

required in order to compute the quantities Eq. (25) and (26) defined in the main text. First, note that while in order to compute Eq. (68) one needs to discard the point $t = t'$, in the case of Eq. (69) it is necessary to consider the points $t \neq t' = t''$, so that it reads more explicitly

$$\begin{aligned} J^{iab} \;=\;& \frac{T}{(N_T^i)^2} \int_{t \neq t' \neq t''} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \mathrm{d}N_{t''}^b \\ \times\;& g^a(t - t') g^b(t - t'') \\ +\;& \frac{T}{(N_T^i)^2} \delta^{i_a i_b} \int_{t \neq t'} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a g^a(t - t') g^b(t - t') \,. \end{aligned} \tag{71}$$

Implementation and complexity of MF-algorithm   The non-zero components of the auxiliary functions can be computed efficiently by exploiting the relations

$$h^a \;=\; \frac{1}{T} \sum_{m=1}^n w_{t_m}^a \tag{72}$$

$$k^{ia} \;=\; \frac{1}{N_T^i} \sum_{m=1}^n y_{t_m}^{ia} \tag{73}$$

$$J^{iab} \;=\; \frac{T}{(N_T^i)^2} \sum_{m=1}^n y_{t_m}^{ia} y_{t_m}^{ib} \tag{74}$$

where one has defined the quantities $w_{t_m}^{ia}$ and $y_{t_m}^{ia}$ as[4]

$$w_{t_m}^a \;=\; \int_{t_m}^T \mathrm{d}t \, g^a(t - t_m) \tag{75}$$

$$y_{t_m}^{ia} \;=\; \sum_{m'=1}^{m-1} g^a(t_m - t_{m'}) \, \delta^{i u_m} \delta^{a u_{m'}} \,, \tag{76}$$

---

[4]Note that The values of $w_{t_m}^a$ and $y_{t_m}^{ia}$ don't have to be stored in memory and can rather be computed on-line during the evaluation of the auxiliary functions

while for $a = 0$ or $b = 0$, one can use the relations $J^{i0a} = J^{ia0} = k^{ia}$ and $k^{i0} = \bar{\Lambda}^i$.

The basis kernels $\boldsymbol{g}(t_m - t_n)$ can be cut-off when they fall below a target value of ϵ 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 $\{y_{t_m}^{ia}\}_{m,i,a}$ is $d^2 p \Lambda^2 T t$. The sum corresponding to (74) then needs $d^3 p^2 \Lambda T$ operations. Thus the total complexity is

$$\text{MF-Complexity} \sim d^2 p \Lambda^2 T t + d^3 p^2 \Lambda T \sim d^2 p \Lambda T \times \max(\Lambda t, dp). \tag{77}$$

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

$$\delta k^{ia} = k^{ia} - \bar{\Lambda}^a \tag{78}$$

$$\delta J^{iab} = J^{iab} - \delta^{i_a i_b}(1 - \delta^{a0})\frac{\bar{\Lambda}^a}{\bar{\Lambda}^i}\nu^{ab} - \frac{\bar{\Lambda}^a \bar{\Lambda}^b}{\bar{\Lambda}^i} \tag{79}$$

$$= J^{iab} - J_0^{iab}, \tag{80}$$

which are expected to be small if the condition **(MFH)** holds (explicit bounds on the rest terms are available in App. 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 $\mathbf{J}$ by using the formula

$$\mathbf{C}^i = \mathbf{C}_0^i \sum_{n=0}^{\infty} (-\boldsymbol{\delta J}^i \mathbf{C}_0^i)^n \tag{81}$$

where we have defined the collection $\{\mathbf{C}_0^i\}_{i=1}^d$ with elements

$$\mathbf{C}_0^i = \bar{\Lambda}^i \begin{bmatrix} 1 + \sum_{j=1}^d (\bar{\Lambda}^j)^2 \sum_{q,q'=0}^p Z_{qq'}^j & \{\boldsymbol{y}^j\}_{1 \le j \le d}^\top \\ \{\boldsymbol{y}^j\}_{1 \le j \le d} & \{\delta^{jk}\mathbf{Z}^j\}_{1 \le j,k \le d} \end{bmatrix}, \tag{82}$$

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

$$\boldsymbol{y}^i = [-\bar{\Lambda}^i \sum_{q'=0}^p Z_{qq'}^i]_{1 \le q \le p} \tag{83}$$

$$\mathbf{Z}^i = [\bar{\Lambda}^i \nu_{qq'}^{ii}]_{1 \le q,q' \le p}^{-1}, \tag{84}$$

In above equation we have recovered the original notation of Sec. 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 Eq. (80) defining $\mathbf{J}_0$ and $\boldsymbol{\delta J}$ that

$$\mathbf{C}_0^i \mathbf{J}_0^i = \mathbb{I} \tag{85}$$

$$\mathbf{C}_0^i \bar{\boldsymbol{\Lambda}} = \boldsymbol{\Sigma}^i \tag{86}$$

$$\boldsymbol{\delta J}^i \mathbf{C}_0^i \bar{\boldsymbol{\Lambda}} = \bar{\Lambda}^i \boldsymbol{\delta k}^i. \tag{87}$$

with $\boldsymbol{\Sigma}^i = [\Lambda^i \delta^{a0}]_{0 \le a \le dp}$. Thanks to these last two equations one can prove that

$$\boldsymbol{\theta}_{MF}^i = \mathbf{C}^i(2\boldsymbol{k}^i - \boldsymbol{h}) \tag{88}$$

$$= \boldsymbol{\Sigma}^i + \mathbf{C}^i(\boldsymbol{\delta k}^i - \boldsymbol{\delta h}) \tag{89}$$

at any order in $\boldsymbol{\delta k}$! Together with the explicit expression for $\mathbf{C}_0^i$, this is the main result of this appendix, that allows us to conclude that:

1. The values of the interaction parameters $\boldsymbol{\alpha}$ are proportional to the fluctuations of $\boldsymbol{\delta k}^i$ and $\boldsymbol{\delta h}$ around their average values.

2. When the fluctuations are very small, the MF approximation explains all the events by forcing the exogenous intensities $\boldsymbol{\mu}$ to be equal to the empirical ones $\bar{\boldsymbol{\Lambda}}$.

3. One could be tempted to expand Eq. (89) as

$$\boldsymbol{\theta}^i_{MF} = \boldsymbol{\Sigma}^i + \mathbf{C}^i_0(\boldsymbol{\delta k}^i - \boldsymbol{\delta h}) - \mathbf{C}^i_0 \boldsymbol{\delta J}^i \mathbf{C}^i_0(\boldsymbol{\delta k}^i - \boldsymbol{\delta h}) + \dots \tag{90}$$

This is not always appropriate, as the truncation can lead to large errors even when the term $\boldsymbol{\delta J}^i$ is extremely small.

In order to understand this last point, let us note that if one could take the limit of small $\boldsymbol{\delta J}^i$ 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 Sec. 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 $\boldsymbol{\delta J}^i$, as one typically expects $\boldsymbol{\delta J}^i \sim O(d^{-1})$, so that $\boldsymbol{\delta J}^i \mathbf{C}^i_0$ can be estimated to be of order 1 (this should be the case of *well-mixed interactions* mentioned in Sec. 5) and the truncation leads to large errors.

D   MLE versus MF solution

In this appendix we want to construct a bound on the error induced by the mean-field 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 $\boldsymbol{\theta}^i_{MLE}$ and the mean-field one $\boldsymbol{\theta}^i_{MF}$ given by Eq. (25), that are given by

$$\begin{aligned}
\boldsymbol{\theta}^i_{MLE} &= \text{argmin}_{\boldsymbol{\theta}^i} \mathcal{L}(\boldsymbol{N}_t, \boldsymbol{\theta}) \tag{91} \\
\boldsymbol{\theta}^i_{MF} &= \mathbf{C}^i(2\boldsymbol{k}^i - \boldsymbol{h}) \tag{92}
\end{aligned}$$

The idea is to bound their difference

$$\boldsymbol{\delta\theta} = \boldsymbol{\theta}_{MLE} - \boldsymbol{\theta}_{MF} \,. \tag{93}$$

where $\mathcal{L}(\boldsymbol{N}_t, \boldsymbol{\theta}) = -\log \mathbb{P}(\boldsymbol{N}_t|\boldsymbol{\theta})$ is the negative log-likelihood function equal to

$$\mathcal{L}(\boldsymbol{N}_t, \boldsymbol{\theta}) = \boldsymbol{h}^\top \boldsymbol{\theta} - \sum_{m=1}^{n} \log \lambda^{u_m}_{t_m} \,. \tag{94}$$

Minimizing it with respect to θ^{ia} produces the relation:

$$\begin{aligned}
0 &= h^{ia} - \frac{1}{T} \sum_{m=1}^{n} \frac{\partial \lambda^{u_m}_{t_m}}{\partial \theta^{ia}} \frac{1}{\lambda^{u_m}_{t_m}} \\
&= h^{ia} - \frac{1}{T} \sum_{m=1}^{n} \frac{\partial \lambda^{u_m}_{t_m}}{\partial \theta^{ia}} \left(\frac{1}{\bar{\Lambda}^{u_m}} - \frac{\lambda^{u_m}_{t_m} - \bar{\Lambda}^{u_m}}{(\bar{\Lambda}^{u_m})^2} + \frac{(\xi^{u_m}_{t_m} - \bar{\Lambda}^{u_m})^2}{(\bar{\Lambda}^{u_m})^3} \right) \\
&= h^{ia} - 2k^{ia} + \sum_{b \in \mathcal{A}^i} J^{iab} \theta^{ib}_{MLE} - \frac{1}{T} \sum_{m=1}^{n} \frac{\partial \lambda^{u_m}_{t_m}}{\partial \theta^{ia}} \frac{(\xi^{u_m}_{t_m} - \bar{\Lambda}^{u_m})^2}{(\bar{\Lambda}^{u_m})^3}
\end{aligned}$$

where in the secion line we have expanded $\lambda^{u_m}_{t_m}$ around $\bar{\Lambda}^{u_m}$, and we defined the rest of the Taylor expansion ξ^i_t, satisfying $|\xi^i_t/\bar{\Lambda}^i - 1| \leq |\lambda^i_t/\bar{\Lambda}^i - 1|$.

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

$$\boldsymbol{\delta\theta}^i = \mathbf{C}^i \boldsymbol{v}^i \,, \tag{95}$$

where

$$
\begin{aligned}
v^{ia} &= \frac{1}{T} \sum_{m=1}^{n} \frac{\partial \lambda_{t_m}^{u_m}}{\partial \theta^{ia}} \frac{(\xi_{t_m}^{u_m} - \bar{\Lambda}^{u_m})^2}{(\bar{\Lambda}^{u_m})^3} \\
&= \frac{1}{N_T^i} \int \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \, g^a(t-t')(\xi_t^i/\bar{\Lambda}^i - 1)^2
\end{aligned}
$$

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

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

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

with

$$
\begin{aligned}
|v^{ia}| = v^{ia} &= \frac{1}{N_T^i} \int_{t \neq t'} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \, g^a(t-t')(\xi_t^i/\bar{\Lambda}^i - 1)^2 \\
&\leq \frac{1}{N_T^i} \int_{t \neq t'} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \, g^a(t-t')(\lambda_t^i/\bar{\Lambda}^i - 1)^2 \\
&= k^{ia} - 2 \sum_{b=0}^{dp} J^{iab} \theta_{MLE}^{ib} + \sum_{b,c=0}^{dp} \theta_{MLE}^{ib} \theta_{MLE}^{ic} Q^{iabc} \,,
\end{aligned}
$$

where we have used the relation

$$\lambda_t^i = \sum_{a=0}^{dp} \theta_{MLE}^{ia} \int \mathrm{d}N_{t'}^a \, g^a(t-t') \,. \tag{96}$$

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

$$Q^{iabc} = \frac{T^2}{(N_T^i)^3} \int_{t \neq t',t''.t'''} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \mathrm{d}N_{t''}^b \mathrm{d}N_{t'''}^c g^a(t-t')g^b(t-t'')g^c(t-t''') \,. \tag{97}$$

This allows to rewrite more compactly

$$|v^{ia}| \leq \sum_{0 \leq b,c \leq dp} (\theta_{MLE}^{ib} - \Sigma^{ib}) Q^{iabc} (\theta_{MLE}^{ic} - \Sigma^{ic}) \,, \tag{98}$$

where $\boldsymbol{\Sigma}^i = [\bar{\Lambda}^i \delta^{a0}]_{0 \leq a \leq dp}$ has been introduced in App. C. The next step will be to perform a cumulant expansion of Q^{iabc} so to take into account the different contributions to Eq. (98).

Before starting the inspection of the various cumulants of Q^{iabc} one-by-one, let us expand first the higher order differentials $\int_{t,t'} \mathrm{d}N_t^a \mathrm{d}N_{t'}^b = \int_{t \neq t'} \mathrm{d}N_t^a \mathrm{d}N_{t'}^b + \int_t \mathrm{d}N_t^a \delta^{i_a i_b}(1 - \delta^{a0})$

by writing

$$
\begin{aligned}
Q^{iabc} &= \frac{T^2}{(N_T^i)^3} \int_{t\neq t'\neq t''\neq t'''} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \mathrm{d}N_{t''}^b \mathrm{d}N_{t'''}^c g^a(t-t')g^b(t-t'')g^c(t-t''') \\
&+ \delta^{i_b i_c}(1-\delta^{b0}) \frac{T^2}{(N_T^i)^3} \int_{t\neq t'\neq t''} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \mathrm{d}N_{t''}^b g^a(t-t')g^b(t-t'')g^c(t-t'') \\
&+ \delta^{i_a i_b}(1-\delta^{a0}) \frac{T^2}{(N_T^i)^3} \int_{t\neq t'\neq t'''} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \mathrm{d}N_{t'''}^c g^a(t-t')g^b(t-t')g^c(t-t''') \\
&+ \delta^{i_a i_c}(1-\delta^{a0}) \frac{T^2}{(N_T^i)^3} \int_{t\neq t'\neq t''} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a \mathrm{d}N_{t''}^b g^a(t-t')g^b(t-t'')g^c(t-t') \\
&+ \delta^{i_a i_b}\delta^{i_a i_c}(1-\delta^{a0}) \frac{T^2}{(N_T^i)^3} \int_{t\neq t'} \mathrm{d}N_t^i \mathrm{d}N_{t'}^a g^a(t-t')g^b(t-t')g^c(t-t')
\end{aligned}
$$

First order contribution At first order in the cumulants of N_t, obtained by substituting $\mathrm{d}N_t = \bar{\Lambda}\mathrm{d}t$, we have

$$
\begin{aligned}
Q_{(1)}^{iabc} &= \frac{\bar{\Lambda}^a}{(\bar{\Lambda}^i)^2} \Big(\bar{\Lambda}^b \bar{\Lambda}^c + \delta^{i_b i_c} \bar{\Lambda}^b (1-\delta^{b0}) \nu^{bc} \\
&+ \delta^{i_a i_b}(1-\delta^{a0}) \bar{\Lambda}^c \nu^{ab} + \delta^{i_a i_c}(1-\delta^{a0}) \bar{\Lambda}^b \nu^{ac} + \delta^{i_a i_b}\delta^{i_a i_c}(1-\delta^{a0}) \rho^{abc} \Big) ,
\end{aligned}
$$

where $\rho^{abc} = \int_0^\infty \mathrm{d}t\, g^a(t)g^b(t)g^c(t)$. By using the properties of $\boldsymbol{\Sigma}$ proved in App. C, we can check that

$$
\begin{aligned}
|v_{(1)}^{ia}| &= \frac{\bar{\Lambda}^a}{(\bar{\Lambda}^i)^2} \bigg((\delta\bar{\Lambda}^i)^2 \\
&+ \sum_{b,c=0}^{dp} \theta_{MLE}^{ib} \theta_{MLE}^{ic} \bar{\Lambda}^b \nu^{bc} \delta^{i_b i_c} \\
&+ 2(1-\delta^{a0})\delta\bar{\Lambda}^i \sum_{b=0}^{dp} \nu^{ab} \delta^{i_a i_b} \theta_{MLE}^{ib} \\
&+ \sum_{b,c=0}^{dp} (1-\delta^{a0}) \rho^{abc} \theta_{MLE}^{ib} \theta_{MLE}^{ic} \delta^{i_a i_b} \delta^{i_a i_c} \bigg)
\end{aligned}
$$

where we have defined

$$
\delta\bar{\Lambda}^i = \boldsymbol{\theta}_{MLE}^{i\top}\bar{\boldsymbol{\Lambda}} - \bar{\Lambda}^i \tag{99}
$$

which is expected to go to zero with T independently of the mean-field approximation. The leading order term is the one associated with the second line of Eq. (99), as

- The first and the third line are associated with terms that go to zero with T due to the presence of $\delta\bar{\Lambda}^i$.

- The fourth line term is bound by a quantity the order of $p^2 |\theta_{MLE}^{\max}|^2$, and hence is expected to be of the order of $p^2\|c_{\max}\|_1^2$ – as one can show by using the bounds proven in App. E – where the functions $c^{ij}(t)$ 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 $dp^2\|c_{\max}\|_1^2$

Hence, if $\|c_{\max}\|_1 \sim d^{-1}$, as one expects in the large dimensional setting (see e.g. App. A which proves this scaling in the homogeneous case), that is of the order of d^{-1}.

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 $||c_{\max}||_1$,

- or they are proportional to $\delta\bar{\Lambda}^i$.

The only term which can contribute at order d^{-1} is the one associated with the first line of Eq. (99). Such term induces a contribution

$$
|v^{ia}_{(2)}| = \frac{\bar{\Lambda}^a}{(\bar{\Lambda}^i)^2} \sum_{b,c=0}^{dp} \theta^{ib}_{MLE}\theta^{ic}_{MLE} \int dtdt'\, c^{bc}(t-t')g^b(t)g^c(t')
$$
$$
+ \quad \text{(higher order)}
$$

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

$$
|v^{ia}| \leq \left| \frac{\bar{\Lambda}^a}{(\bar{\Lambda}^i)^2} \left[\sum_{b,c=0}^{dp} \theta^{ib}_{MLE}\theta^{ic}_{MLE} \int dtdt' \left(c^{bc}(t-t') + \bar{\Lambda}^b \delta^{i_b i_c}\delta(t-t') \right) g^b(t)g^c(t') \right] \right|
$$
$$
+ \quad \text{(higher order)}.
$$

that can be more compactly written as

$$
|v^{ia}| \leq \left| \frac{\bar{\Lambda}^a}{(\bar{\Lambda}^i)^2} \bar{\mathbb{V}}(\lambda^i_t) \right| + \text{(higher order)}
$$

where the variance corresponds to the *empirical* one computed under the saddle-point parameters $\boldsymbol{\theta}_{MLE}$. The term in parenthesis corresponds exactly to the fluctuations of $\boldsymbol{N}_t$, hence **the error induced by the mean-field approximation is bound by the variance of the intensity $\boldsymbol{\lambda}_t$** as anticipated in the main text. We finally get:

$$
||\boldsymbol{\delta\theta^i}|| \leq ||\mathbf{C}^i|| ||\bar{\mathbf{\Lambda}}|| \frac{\bar{\mathbb{V}}(\lambda^i_t)}{(\bar{\Lambda}^i)^2} + \text{(higher order)} \tag{100}
$$

that is the formula used in Sec. 3.3.

E   Cumulant expansion

In this appendix we will be interesting in providing *a priori* bounds on the value of the inferred parameters $\boldsymbol{\theta}^i_{MF}$ and on their associated errors $\boldsymbol{\delta\theta^i}$. We will be able to relate with quantities with the empirical cumulants of the process $\boldsymbol{N}_t$, 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:

$$h^{ia} = \bar{\Lambda}^a \tag{101}$$

$$k^{ia} = \bar{\Lambda}^a + \frac{1}{\bar{\Lambda}^i} \int_0^\infty \mathrm{d}t\, c^{ia}(t) g^a(t) \tag{102}$$

$$
\begin{aligned}
J^{iab} &= \delta^{i_a i_b}(1 - \delta^{a0}) \left(\frac{\bar{\Lambda}^a}{\bar{\Lambda}^i} \int_0^\infty \mathrm{d}t\, g^a(t) g^b(t) + \frac{1}{(\bar{\Lambda}^i)^2} \int_0^\infty \mathrm{d}t\, c^{ia}(t) g^a(t) g^b(t) \right) \\
&+ \frac{\bar{\Lambda}^a \bar{\Lambda}^b}{\bar{\Lambda}^i} + \frac{1}{\bar{\Lambda}^i} \int_0^\infty \mathrm{d}t\mathrm{d}t'\, c^{ab}(t - t') g^a(t) g^b(t') \\
&+ \frac{\bar{\Lambda}^a}{(\bar{\Lambda}^i)^2} \int_0^\infty \mathrm{d}t\, c^{ib}(t) g^b(t) + \frac{\bar{\Lambda}^b}{(\bar{\Lambda}^i)^2} \int_0^\infty \mathrm{d}t\, c^{ia}(t) g^a(t) \\
&+ \frac{1}{(\bar{\Lambda}^i)^2} \int_0^\infty \mathrm{d}t\mathrm{d}t'\, K^{iab}(t, t') g^a(t) g^b(t')
\end{aligned}
\tag{103}
$$

where the vectors $\boldsymbol{c}^i(t)$ and the matrices $\mathbf{K}^i(t, t')$ 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

$$c^{ij}(t) = \frac{1}{T} \sum_{m' < m} \delta^{i u_m} \delta^{j u_{m'}} \delta(t - t_m + t_{m'}) - \bar{\Lambda}^i \bar{\Lambda}^j \tag{104}$$

$$
\begin{aligned}
K^{ijk}(t, t') &= \frac{1}{T} \sum_{m' \neq m'' < m} \delta^{i u_m} \delta^{j u_{m'}} \delta^{k u_{m''}} \delta(t - t_m + t_{m'}) \delta(t' - t_m + t_{m''}) \\
&- \bar{\Lambda}^i \bar{\Lambda}^j \bar{\Lambda}^k - \bar{\Lambda}^i c^{jk}(t - t') - \bar{\Lambda}^j c^{ik}(t') - \bar{\Lambda}^k c^{ij}(t)
\end{aligned}
\tag{105}
$$

while one obviously has $c^{i0}(t) = K^{i0a}(t, t') = K^{ia0}(t, t') = 0$. As in the last appendix, we write $\nu^{ab} = \int_0^\infty \mathrm{d}t\, g^a(t) g^b(t)$, and we will also employ the quantities

$$\delta k^{ia} = k^{ia} - \bar{\Lambda}^a \tag{106}$$

$$\delta J^{iab} = J^{iab} - \delta^{i_a i_b}(1 - \delta^{a0}) \frac{\bar{\Lambda}^a}{\bar{\Lambda}^i} \nu^{ab} - \frac{\bar{\Lambda}^a \bar{\Lambda}^b}{\bar{\Lambda}^i} \tag{107}$$

$$= J^{iab} - J_0^{iab}, \tag{108}$$

which are proportional to the empirical cumulants. The quantities $\boldsymbol{\delta k}^i$ and $\boldsymbol{\delta J}^i$ are in fact bound by

$$|\delta k^{ia}| \leq \frac{g_{\max} ||c_{\max}||_1}{\bar{\Lambda}_{\min}} \tag{109}$$

$$|\delta J^{iab}| \leq \delta^{i_a i_b}(1 - \delta^{a0}) \left(\frac{g_{\max}^2 ||c_{\max}||_1}{\bar{\Lambda}_{\min}^2} \right) \tag{110}$$

$$+ 2 \left(\frac{g_{\max} ||c_{\max}||_1}{\bar{\Lambda}_{\min}} \right) + \frac{g_{\max}^2 ||K_{\max}||_1}{\bar{\Lambda}_{\min}^2} \tag{111}$$

where $\bar{\Lambda}_{\min} = \min_{1 \leq i \leq d} \bar{\Lambda}^i$, $g_{\max} = \max_{a,t} g^a(t)$ and

$$||c_{\max}||_1 = \max_{ij} \int_0^\infty \mathrm{d}t\, |c^{ij}(t)| \tag{112}$$

$$||K_{\max}||_1 = \max_{ijk} \int_0^\infty \mathrm{d}t\mathrm{d}t'\, |K^{ijk}(t, t')|. \tag{113}$$

By particularizing these inequalities to Eq. (D) one can obtain a bound on the mean-field

error that is of the form

$$
\begin{aligned}
|v^{ia}| &\leq \frac{\bar{\Lambda}^a}{(\bar{\Lambda}^i)^2} \left(||c_{\max}||_1 g_{\max} \left(\sum_{b=0}^{dp} \theta_{MLE}^{ib} \right)^2 + \nu_{\max} \bar{\Lambda}_{\max} \sum_{b,c=0}^{dp} \theta_{MLE}^{ib} \theta_{MLE}^{ic} \delta^{i_b i_c} \right) \\
&\leq \frac{||c_{\max}||_1 g_{\max} \bar{\Lambda}_{\max}}{\bar{\Lambda}_{\min}^2} \left(\sum_{b=0}^{dp} \theta_{MLE}^{ib} \right)^2 + \frac{\nu_{\max} \bar{\Lambda}_{\max}^2}{\bar{\Lambda}_{\min}^2} \sum_{b,c=0}^{dp} \theta_{MLE}^{ib} \theta_{MLE}^{ic} \delta^{i_b i_c}
\end{aligned}
$$

By constructing the norm of the error, and comparing them with the bound on the inferred couplings Eq. (89), yields the expressions

$$
\begin{aligned}
||\boldsymbol{\theta}_{MF}^i|| &\leq \sqrt{dp} ||\mathbf{C}^i|| \left(\frac{g_{\max} ||c_{\max}||_1}{\bar{\Lambda}_{\min}} \right) \\
||\boldsymbol{\delta\theta}^i|| &\leq \sqrt{dp} ||\mathbf{C}^i|| \left(\frac{||c_{\max}||_1 g_{\max} \bar{\Lambda}_{\max}}{\bar{\Lambda}_{\min}^2} \left(\sum_{b=0}^{dp} \theta_{MLE}^{ib} \right)^2 + \frac{\nu_{\max} \bar{\Lambda}_{\max}^2}{\bar{\Lambda}_{\min}^2} \sum_{b,c=0}^{dp} \theta_{MLE}^{ib} \theta_{MLE}^{ic} \delta^{i_b i_c} \right) .
\end{aligned}
$$

References

[1] A.G. Hawkes. Point spectra of some mutually exciting point processes. *J. R. Statist. Soc. B*, 33:438–443, 1971.

[2] A.G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[3] Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

[4] L. Bauwens and N. Hautsch. Stochastic conditional intensities processes. *Journal of Financial Econometrics*, 4:450–493, 2009.

[5] C.G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

[6] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 01(01):1550005, 2015.

[7] P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Inference of functional connectivity in neurosciences via hawkes processes. In *1st IEEE Global Conference on Signal and Information Processing*, 2013.

[8] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 2011.

[9] C. Blundell, J. Beck, and K. A. Heller. Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

[10] M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. In *Advances in neural information processing systems*, pages 2474–2482, 2014.

[11] S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*, 2014.

[12] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–274. ACM, 2013.

[13] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1301–1309, 2013.

[14] L. Tran, M. Farajtabar, L. Song, and H. Zha. Netcodec: Community detection from individual activities. In *In SDM*, 2015.

[15] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song. Co-evolve: A joint point process model for information diffusion and network co-evolution. *arXiv preprint arXiv:1507.02293*, 2015.

[16] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41), 2008.

[17] Y. Ogata and H. Akaike. On linear intensity models for mixed doubly stochastic poisson and self- exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):102–107, January 1982. ArticleType: research-article / Full publication date: 1982 / Copyright Â© 1982 Royal Statistical Society.

[18] Emmanuel Bacry, Khalil Dayri, and Jean-François Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5):1–12, 2012.

[19] Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14:1147–1166, 2014.

[20] Emmanuel Bacry and Jean-Francois Muzy. Second order statistics characterization of Hawkes processes and non-parametric estimation. *arXiv preprint arXiv:1401.0903*, 2014.

[21] Sylvain Delattre, Nicolas Fournier, and Marc Hoffmann. High dimensional hawkes processes. *arXiv preprint arXiv:1403.5764*, 2014.

[22] D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes*, volume 2. Springer, 1988.

[23] E. Lewis and G. Mohler. A nonparametric em algorithm for multiscale hawkes processes. *preprint*, 2011.

[24] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[25] John E Dennis, Jr and Jorge J Moré. Quasi-newton methods, motivation and theory. *SIAM review*, 19(1):46–89, 1977.

[26] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[27] W. Feller. *An introduction to probability theory and its applications*, volume II. John Wiley & Sons Inc., 2 edition, 1971.