The Annals of Statistics
2020, Vol. 48, No. 5, 2698–2727
https://doi.org/10.1214/19-AOS1903
© Institute of Mathematical Statistics, 2020

NONPARAMETRIC BAYESIAN ESTIMATION FOR MULTIVARIATE HAWKES PROCESSES

BY SOPHIE DONNET[1], VINCENT RIVOIRARD[2,*] AND JUDITH ROUSSEAU[2,**]

[1] *INRA, Université Paris Saclay, sophie.donnet@inra.fr*
[2] *CEREMADE, CNRS, UMR 7534, Université Paris–Dauphine, * Vincent.Rivoirard@dauphine.fr;*
*** Judith.Rousseau@stats.ox.ac.uk*

This paper studies nonparametric estimation of parameters of multivariate Hawkes processes. We consider the Bayesian setting and derive posterior concentration rates. First, rates are derived for $\mathbb{L}_1$-metrics for stochastic intensities of the Hawkes process. We then deduce rates for the $\mathbb{L}_1$-norm of interactions functions of the process. Our results are exemplified by using priors based on piecewise constant functions, with regular or random partitions and priors based on mixtures of Betas distributions. We also present a simulation study to illustrate our results and to study empirically the inference on functional connectivity graphs of neurons

1. Introduction.

In this paper we study the properties of Bayesian nonparametric procedures in the context of multivariate Hawkes processes. The aim of this paper is to give some general results on posterior concentration rates for such models and to study some families of nonparametric priors.

1.1. *Hawkes processes.*

Hawkes processes, introduced by Hawkes [27] and Hawkes and Oakes [28], are specific point processes which are extensively used to model data whose occurrences depend on previous occurrences of the same process. First introduced in the univariate setting, they can be easily extended to model marked or multivariate point processes [18].

To describe multivariate Hawkes processes, we consider a multivariate point process $(N_t)_t := (N_t^1, \dots, N_t^K)_t$, each component N_t^k recording the time of occurrences of events of the kth component of a system or equivalently the number of events of the kth component until time t. Under mild general assumptions, a multivariate counting process is characterized by its intensity process $(\lambda_t^1, \dots, \lambda_t^K)$, informally given by

$$\lambda_t^k \, dt = \mathbb{P}\big(N_k^t \text{ has a jump in } [t, t+dt] | \mathcal{G}_{t^-}\big),$$

where $\mathcal{G}_{t^-}$ denotes the sigma-field generated by $N = (N^1, \dots, N^K)$ up to time t (excluding t). In this paper we concentrate on linear Hawkes processes. The intensities associated with the class of linear mutually exciting Hawkes processes are of the form

$$\lambda_t^k = \nu_k + \sum_{\ell=1}^K \int_{-\infty}^{t^-} h_{\ell,k}(t-u) \, dN_u^\ell,$$

where $h_{\ell,k}$, is nonnegative, supported by $\mathbb{R}_+$ and is called the *interaction function* of N^ℓ on N^k and $\nu_k > 0$ is the *spontaneous rate* associated with the process N^k. We recall that the

Received March 2018; revised March 2019.
MSC2020 subject classifications. Primary 62G20, 60G55; secondary 62G05.
Key words and phrases. Multivariate counting process, Hawkes processes, nonparametric Bayesian estimation, posterior concentration rates.

previous integral means

$$\int_{-\infty}^{t^-} h_{\ell,k}(t-u)\,dN_u^\ell = \sum_{T_i^\ell \in N^\ell : T_i^\ell < t} h_{\ell,k}(t - T_i^\ell),$$

where the T_i^ℓ's are the random points of N^ℓ (see [2]).

Hawkes processes have been extensively used in a wide range of applications. They are used to model earthquakes [33, 45, 48], interactions in social networks [5, 17, 30, 31, 44, 46, 47], financial data [1, 4, 6, 7, 20], violence rates [32, 37], genomes [12, 25, 42] or neuronal activities [11, 16, 26, 34–36, 39, 40], to name but a few.

Parametric inference for Hawkes models based on the likelihood is the most common in the literature, and we refer the reader to [12, 33] for instance. Nonparametric estimation has first been considered by Reynaud-Bouret and Schbath [42] who proposed a procedure based on minimization of an ℓ_2-criterion penalized by an ℓ_0-penalty for univariate Hawkes processes. Their results have been extended to the multivariate setting by Hansen, Reynaud-Bouret and Rivoirard [26] where the ℓ_0-penalty is replaced with an ℓ_1-penalty. The resulting Lasso-type estimate leads to an easily implementable procedure providing sparse estimation of the structure of the underlying connectivity graph. To generalize this procedure to the high-dimensional setting, Chen, Witten and Shojaie [15] proposed a simple and computationally inexpensive edge screening approach, whereas Bacry, Gaïffas and Muzy [5] combine ℓ_1 and trace norm penalizations to take into account the low rank property of their self-excitement matrix. Very recently, to deal with nonpositive interaction functions, Chen, Shojaie, Shea-Brown and Witten [14] combine the thinning process representation and a coupling construction to bound the dependence coefficient of the Hawkes process. Other alternatives based on spectral methods [3] or estimation through the resolution of a Wiener–Hopf system [8] can also been found in the literature. These are all frequentist methods; Bayesian approaches for Hawkes models have received much less attention. To the best of our knowledge, the only contributions for the Bayesian inference are due to Rasmussen [38] and Blundell, Beck and Heller [9] who explored parametric approaches and used MCMC to approximate the posterior distribution of the parameters.

1.2. *Our contribution.* In this paper, we study nonparametric posterior concentration rates, when $T \to +\infty$, for estimating the parameter $f = ((\nu_k)_{k=1,\dots,K}, (h_{\ell,k})_{k,\ell=1,\dots,K})$ by using realizations of the multivariate process $(N_t^k)_{k=1,\dots,K}$ for $t \in [0, T]$. Analyzing asymptotic properties in the setting where $T \to +\infty$ means that the observation time becomes very large, hence providing a large number of observations. Note that along the paper, K, the number of observed processes, is assumed to be fixed and can be viewed as a constant. Considering $K \to +\infty$ is a very challenging problem beyond the scope of this paper. Using the general theory of Ghosal and van der Vaart [22], we express the posterior concentration rates in terms of simple and usual quantities associated to the prior on f and under mild conditions on the true parameter. Two types of posterior concentration rates are provided: the first one is in terms of the $\mathbb{L}_1$-distance on the stochastic intensity functions $(\lambda^k)_{k=1,\dots,K}$, and the second one is in terms of the $\mathbb{L}_1$-distance on the parameter f (see precise notation below). To the best of our knowledge, these are the first theoretical results on Bayesian nonparametric inference in Hawkes models. Moreover, these are the first results on $\mathbb{L}_1$-convergence rates for the interaction functions $h_{\ell,k}$. In the frequentist literature, theoretical results are given in terms of either the $\mathbb{L}_2$-error of the stochastic intensity, as in [5] and [8], or in terms of the $\mathbb{L}_2$-error on the interaction functions themselves, the latter being much more involved, as in [42] and [26]. In [42], the estimator is constructed using a frequentist model selection procedure with a specific family of models based on piecewise constant functions. In the multivariate setting of [26], more generic families of approximation models are considered (wavelets of Fourier dictionaries) and then combined with a Lasso procedure but under a somewhat restrictive assumption on the type and size of the models that can be used to construct their estimators (see Section 5.2 of [26]). Our general results do not involve such strong conditions and, therefore, allow us to work with approximating families of models that are quite general. Our conditions are very similar to the conditions proposed in the context of density estimation in [21] so that most of the priors which have been studied in the context of density estimation can now be easily adapted to the context of the interaction functions of multivariate Hawkes processes. In particular, we have applied these conditions to two families of prior models on the interaction functions $h_{\ell,k}$: priors based on piecewise constant functions, with regular or random partitions, and priors based on mixtures of Betas distributions. From the posterior concentration rates we also deduce a frequentist convergence rate for the posterior mean, seen as a point estimator. We finally propose an MCMC algorithm to simulate from the posterior distribution for the priors constructed from piecewise constant functions, and a simulation study is conducted to illustrate our results.

1.3. *Formal definitions*, *notation and assumptions*. We first recall the formal definition of multivariate Hawkes processes, and we define our setup. In the sequel, for any $\mathbb{R}$-valued function h, we denote by $\|h\|_p$ its $\mathbb{L}_p$-norm. Consider the probability space $(\mathcal{X}, \mathcal{G}, \mathbb{P})$. For any k and any set A, we denote by $N^k(A)$ the number of occurrences of N^k in A. We can define linear multivariate Hawkes processes as follows.

DEFINITION 1. Let $T > 0$. We consider $f = ((\nu_k)_{k=1,\dots,K}, (h_{\ell,k})_{k,\ell=1,\dots,K})$ such that for all k, ℓ, $\nu_k > 0$ and $h_{\ell,k}$ is nonnegative and integrable. Let $(N_t)_t = (N_t^1, \dots, N_t^K)_t$, and assume that $\mathcal{G}_T \subset \mathcal{G}$ with $\mathcal{G}_t = \mathcal{G}_0 \vee \sigma(N_s, s \leq t)$, for some $\mathcal{G}_0 \subset \mathcal{G}$. Then, the process $(N_t)_t$ adapted to $(\mathcal{G}_t)_t$ is a linear multivariate Hawkes process with parameter f if:

- almost surely, for all $k \neq \ell$, $(N_t^k)_t$ and $(N_t^\ell)_t$ never jump simultaneously;
- for all k, the intensity process $(\lambda_t^k(f))_t$ of $(N_t^k)_t$ is given by

$$\lambda_t^k(f) = \nu_k + \sum_{\ell=1}^K \int_{-\infty}^{t^-} h_{\ell,k}(t-u)\,dN_u^\ell. \tag{1.1}$$

Conditions of Definition 1 on f ensure existence and uniqueness of a pathwise Hawkes process $(N_t)_t = (N_t^1, \dots, N_t^K)_t$ such that $N_t^k < \infty$ almost surely for any k and any t. Furthermore, Theorem 7 of [10] shows that if the $K \times K$ matrix ρ, with

$$\rho_{\ell,k} = \int_0^{+\infty} h_{\ell,k}(t)\,dt, \quad \ell, k = 1, \dots, K, \tag{1.2}$$

has a spectral radius strictly smaller than 1, then there exists a unique stationary distribution for the multivariate process $N = (N^k)_{k=1,\dots,K}$ with intensities given by (1.1) and finite average intensity.

Given a parameter $f = ((\nu_k)_{k=1,\dots,K}, (h_{\ell,k})_{k,\ell=1,\dots,K})$, we denote by $\|\rho\|$ the spectral norm of the matrix ρ associated with f and defined in (1.2). We recall that $\|\rho\|$ provides an upper bound of the spectral radius of ρ.

Let $A > 0$ be a given known constant, set

$$\mathcal{H} = \{(h_{\ell,k})_{k,\ell=1,\dots,K};\, h_{\ell,k} \text{ is integrable}, h_{\ell,k} \geq 0, \|h_{\ell,k}\|_\infty < \infty,$$
$$\text{support}(h_{\ell,k}) \subset [0, A], \forall k, \ell \leq K\}$$

and

$$\mathcal{F} = \{f = ((\nu_k)_{k=1,\dots,K}, (h_{\ell,k})_{k,\ell=1,\dots,K}); 0 < \nu_k < \infty, \forall k \leq K,$$
$$(h_{\ell,k})_{k,\ell=1,\dots,K} \in \mathcal{H}\}.$$

In the sequel, for $T > 0$, we assume that we observe N, a linear Hawkes process with true parameter $f_0 = ((\nu_k^0)_{k=1,\dots,K}, (h_{\ell,k}^0)_{k,\ell=1,\dots,K}) \in \mathcal{F}$, until time T. Denote by ρ^0 the matrix such that $\rho_{\ell,k}^0 = \int_0^A h_{\ell,k}^0(t)\,dt$, and assume that $\|\rho^0\| < 1$. For the sake of simplicity, we assume $\sigma(N_s, s < 0) \subset \mathcal{G}_0$ so $\mathcal{G}_0 = \mathcal{G}_0 \vee \sigma(N_s, s < 0)$, and we denote by $\mathbb{P}_0$ the stationary distribution of N (associated to f_0) and by $\mathbb{P}_0(\cdot|\mathcal{G}_0)$ the conditional distribution of N given $\mathcal{G}_0$. Finally, $\mathbb{E}_0$ is the expectation associated with $\mathbb{P}_0$.

Now, let $f = ((\nu_k)_{k=1,\dots,K}, (h_{\ell,k})_{k,\ell=1,\dots,K}) \in \mathcal{F}$, and we define $\lambda_t(f) = (\lambda_t^k(f))_{k=1,\dots,K}$ for all $t \geq 0$ where

$$\lambda_t^k(f) = \nu_k + \sum_{\ell=1}^{K} \int_{t-A}^{t^-} h_{\ell,k}(t-u)\,dN_u^\ell.$$

From Chapter II of [2], if

$$L_T(f) := \sum_{k=1}^{K} \left[\int_0^T \log(\lambda_t^k(f))\,dN_t^k - \int_0^T \lambda_t^k(f)\,dt\right], \tag{1.3}$$

and

$$d\mathbb{P}_f(\cdot|\mathcal{G}_0) = e^{L_T(f) - L_T(f_0)}\,d\mathbb{P}_0(\cdot|\mathcal{G}_0),$$

then $\mathbb{P}_f(\cdot|\mathcal{G}_0)$ is a conditional probability distribution on $(\mathcal{X}, \mathcal{G})$ and, under $\mathbb{P}_f$, N is a multivariate Hawkes process with intensity process $(\lambda_t(f))_{0\leq t\leq T}$. Note that if the spectral radius of ρ is less than 1, then, under $\mathbb{P}_f$, N is a stationary multivariate Hawkes process. With a slight abuse of notation, we also denote, at times, $L_T(\lambda)$ in place of $L_T(f)$. In the sequel, $\mathbb{E}_f$ is the expectation with respect to $\mathbb{P}_f$.

Now, given two parameters $f = ((\nu_k)_{k=1,\dots,K}, (h_{\ell,k})_{k,\ell=1,\dots,K})$ and $f' = ((\nu_k')_{k=1,\dots,K}, (h_{\ell,k}')_{k,\ell=1,\dots,K})$ belonging to $\mathcal{F}$, we set

$$\|f - f'\|_1 = \sum_{k=1}^{K} |\nu_k - \nu_k'| + \sum_{k=1}^{K}\sum_{\ell=1}^{K} \|h_{\ell,k} - h_{\ell,k}'\|_1 \tag{1.4}$$

and

$$d_{1,T}(f, f') = \frac{1}{T} \sum_{k=1}^{K} \int_0^T |\lambda_t^k(f) - \lambda_t^k(f')|\,dt.$$

Note that $d_{1,T}$ is a data-dependent pseudodistance on $\mathcal{F}$. We denote by $\mathcal{N}(u, \mathcal{H}_0, d)$ the covering number of a set $\mathcal{H}_0$ by balls with respect to a metric d with radius u. We set for any ℓ, μ_ℓ^0 the mean of $\lambda_t^\ell(f_0)$ under $\mathbb{P}_0$:

$$\mu_\ell^0 = \mathbb{E}_0[\lambda_t^\ell(f_0)].$$

We also write $u_T \lesssim v_T$ if $|u_T/v_T|$ is bounded when $T \to +\infty$ and, similarly, $u_T \gtrsim v_T$ if $|v_T/u_T|$ is bounded. Finally, if Ω is a set Ω^c denotes its complement.

1.4. *Overview of the paper.* In Section 2, Theorem 1 first states the posterior convergence rates obtained for stochastic intensities. Theorem 2 constitutes a variation of this first result. From these results we derive $\mathbb{L}_1$-rates for the parameter f (see Theorem 3) and for the posterior mean (see Corollary 1). Examples of prior models satisfying conditions of these theorems are given in Section 2.3. In Section 3 numerical results are provided. Finally, Section 4 provides the proof of Theorem 3 (Section 4.3). Before that, to deal with the posterior distributions, we construct specific tests (Lemma 1 in Section 4.1) and provide a general control of the Kullback–Leibler divergence between two given functions (Section 4.2). Proofs of other results are given in the Supplementary Material [19].

2. Main results.

This section contains the main results of the paper. We first provide an expression for the posterior distribution.

2.1. *Posterior distribution.*

Recall that we restrict ourselves to the setup where for all ℓ, k, $h_{\ell,k}$ has support included in $[0, A]$ for some fixed known $A > 0$. This hypothesis is very common in the context of Hawkes processes; see [26].

Hence, in the following we assume that we observe the process $(N^k)_{k=1,\dots,K}$ on $[-A, T]$, but we base our inference on the log-likelihood (1.3) which is associated to the observation of $(N^k)_{k=1,\dots,K}$ on $[0, T]$. We consider a Bayesian nonparametric approach and denote by Π the prior distribution on the parameter $f = ((\nu_k)_{k=1,\dots,K}, (h_{\ell,k})_{k,\ell=1,\dots,K})$. The posterior distribution is then formally equal to

$$\Pi(B|N, \mathcal{G}_0) = \frac{\int_B \exp(L_T(f))\, d\Pi(f|\mathcal{G}_0)}{\int_{\mathcal{F}} \exp(L_T(f))\, d\Pi(f|\mathcal{G}_0)}.$$

We approximate it by the following pseudo-posterior distribution, which we write $\Pi(\cdot|N)$

$$\Pi(B|N) = \frac{\int_B \exp(L_T(f))\, d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f))\, d\Pi(f)}, \tag{2.1}$$

which thus corresponds to choosing $d\Pi(f) = d\Pi(f|\mathcal{G}_0)$.

2.2. *Posterior convergence rates for $d_{1,T}$ and $\mathbb{L}_1$-metrics.*

In this section we give two results of posterior concentration rates, one in terms of the stochastic distance $d_{1,T}$ and another one in terms of the $\mathbb{L}_1$-distance, which constitutes the main result of this paper. We define

$$\Omega_T = \Big\{ \max_{\ell\in\{1,\dots,K\}} \sup_{t\in[0,T]} N^\ell[t-A, t) \le C_\alpha \log T \Big\} \cap \Bigg\{ \sum_{\ell=1}^K \bigg| \frac{N^\ell[-A, T]}{T} - \mu_\ell^0 \bigg| \le \delta_T \Bigg\}$$

with $\delta_T = \delta_0 (\log T)^{3/2}/\sqrt{T}$ and $\delta_0 > 0$ and C_α two positive constants not depending on T. From Lemmas 2 and 3 in Section 2.5 in the Supplementary Material [19], we have that for all $\alpha > 0$ there exist $C_\alpha > 0$ and $\delta_0 > 0$ only depending on α and f_0 such that

$$\mathbb{P}_0(\Omega_T^c) \le T^{-\alpha}, \tag{2.2}$$

when T is large enough. In the sequel, we take $\alpha > 1$ and C_α accordingly. Note, in particular, that on Ω_T, $\sum_{\ell=1}^K N^\ell[-A, T] \le N_0 T$, with $N_0 = 1 + \sum_{\ell=1}^K \mu_\ell^0$, when T is large enough. We then have the following theorem.

THEOREM 1. *Consider the multivariate Hawkes process $(N^k)_{k=1,\dots,K}$ observed on $[-A, T]$ with likelihood given by (1.3). Let Π be a prior distribution on $\mathcal{F}$. Let ϵ_T be a* *positive sequence such that* $\epsilon_T = o(1)$ *and* $\log\log(T)\log^3(T)/(T\epsilon_T^2) = o(1)$. *For* $B > 0$, *we consider*

$$B(\epsilon_T, B) := \Big\{ (\nu_k, (h_{\ell,k})_\ell)_k; \max_k |\nu_k - \nu_k^0| \le \epsilon_T, \\ \max_{\ell,k} \|h_{\ell,k} - h_{\ell,k}^0\|_2 \le \epsilon_T, \max_{\ell,k} \|h_{\ell,k}\|_\infty \le B \Big\}$$

and assume the following conditions are satisfied for T *large enough*:

(i) *There exists* $c_1 > 0$ *and* $B > 0$ *such that*

$$\Pi\big(B(\epsilon_T, B)\big) \ge e^{-c_1 T \epsilon_T^2}.$$

(ii) *There exists a subset* $\mathcal{H}_T \subset \mathcal{H}$, *such that*

$$\frac{\Pi(\mathcal{H}_T^c)}{\Pi(B(\epsilon_T, B))} \le e^{-(2\kappa_T + 3)T\epsilon_T^2},$$

where $\kappa_T := \kappa \log(r_T^{-1}) \asymp \log\log T$, *with* r_T *defined in* (4.4) *and* κ *defined in* (4.2).

(iii) *There exist* $\zeta_0 > 0$ *and* $x_0 > 0$ *such that*

$$\log \mathcal{N}(\zeta_0 \epsilon_T, \mathcal{H}_T, \|\cdot\|_1) \le x_0 T \epsilon_T^2.$$

Then, there exist $M > 0$ *and* $C > 0$ *such that*

$$\mathbb{E}_0\big[\Pi\big(d_{1,T}(f_0, f) > M\sqrt{\log\log T}\,\epsilon_T | N\big)\big] \\ \le \frac{C \log\log(T)\log^3(T)}{T\epsilon_T^2} + \mathbb{P}_0(\Omega_T^c) + o(1) = o(1).$$

Assumptions (i), (ii) and (iii) are very common in the literature about posterior convergence rates. As expressed by Assumption (ii), some conditions are required on the prior on $\mathcal{H}_T$ but not on the parameters ν_k. Except the usual concentration property of ν around ν^0 expressed in the definition of $B(\epsilon_T, B)$, which is in particular satisfied if ν has a positive continuous density with respect to Lebesgue measure, we have no further condition on the tails of the distribution of ν.

REMARK 1. As appears in the proof of Theorem 1, the term $\sqrt{\log\log T}$ appearing in the posterior concentration rate can be dropped if $B(\epsilon_T, B)$ is replaced by

$$B_\infty(\epsilon_T, B) = \Big\{ (\nu_k, (h_{\ell,k})_\ell)_k; \max_k |\nu_k - \nu_k^0| \le \epsilon_T, \max_{\ell,k} \|h_{\ell,k} - h_{\ell,k}^0\|_\infty \le \epsilon_T \Big\},$$

in Assumption (i). In this case, $r_T = 1/2$ in Assumption (ii) and κ_T does not depend on T. This is used for instance in Section 2.3.1 to study random histograms priors whereas mixtures of Beta priors are controlled using the $\mathbb{L}_2$-norm.

Similar to other general theorems on posterior concentration rates, we can consider some variants. Since the metric $d_{1,T}$ is stochastic, we cannot use slices in the form $d_{1,T}(f_0, f) \in (j\epsilon_T, (j+1)\epsilon_T)$ as in Theorem 1 of Ghosal and van der Vaart [22], however, we can consider other forms of slices, using a similar idea as in Theorem 5 of Ghosal and van der Vaart [23]. This is presented in the following theorem:

THEOREM 2. *Consider the setting and assumptions of Theorem 1 except that assumption (iii) is replaced by the following one. There exists a sequence of sets $(\mathcal{H}_{T,i})_{i\geq 1} \subset \mathcal{H}$ with $\bigcup_i \mathcal{H}_{T,i} = \mathcal{H}_T$ and $\zeta_0 > 0$ such that*

$$\sum_{i=1}^{\infty} \mathcal{N}(\zeta_0 \epsilon_T, \mathcal{H}_{T,i}, \|\cdot\|_1)\sqrt{\Pi(\mathcal{H}_{T,i})} e^{-x_0 T \epsilon_T^2} = o(1), \tag{2.3}$$

for some positive constant $x_0 > 0$. Then, there exists $M > 0$ such that

$$\mathbb{E}_0\big[\Pi\big(d_{1,T}(f_0, f) > M\sqrt{\log\log T}\epsilon_T | N\big)\big] = o(1).$$

The posterior concentration rates of Theorems 1 and 2 are in terms of the metric $d_{1,T}$ on the intensity functions which are data dependent and therefore not completely satisfying to understand concentration around the objects of interest namely f_0. We now use Theorem 1 to provide a general result to derive a posterior concentration rate in terms of the $\mathbb{L}_1$-norm.

THEOREM 3. *Assume that the prior Π satisfies the following assumptions*:

(i) *There exists $\varepsilon_T = o(1)$ such that $\varepsilon_T \geq \delta_T$ (see the definition of Ω_T) and $c_1 > 0$ such that*

$$\mathbb{E}_0\big[\Pi\big(A_{\varepsilon_T}^c | N\big)\big] = o(1) \quad \text{and} \quad \mathbb{P}_0\big(D_T < e^{-c_1 T \varepsilon_T^2}\big) = o(1), \tag{2.4}$$

where $D_T = \int_{\mathcal{F}} e^{L_T(f) - L_T(f_0)} \, d\Pi(f)$ and $A_{\varepsilon_T} = \{f; d_{1,T}(f_0, f) \leq \varepsilon_T\}$.

(ii) *The prior on ρ satisfies the following property*: *for all $u_0 > 0$, when T is large enough,*

$$\Pi\big(\|\rho\| > 1 - u_0 (\log T)^{1/6} \varepsilon_T^{1/3}\big) \leq e^{-2c_1 T \varepsilon_T^2}. \tag{2.5}$$

Then, for any $w_T \to +\infty$,

$$\mathbb{E}_0\big[\Pi\big(\|f - f_0\|_1 > w_T \varepsilon_T | N\big)\big] = o(1). \tag{2.6}$$

REMARK 2. Condition (i) of Theorem 3 is, in particular, verified under the assumptions of Theorem 1, with $\varepsilon_T = M\epsilon_T\sqrt{\log\log T}$ for M a constant.

REMARK 3. Compared to Theorem 1, we also assume (ii), that is, that the prior distribution puts very little mass near the boundary of space $\{f; \|\rho\| < 1\}$. In particular, if under Π, $\|\rho\|$ has its support included in $[0, 1 - \epsilon]$ for a fixed small $\epsilon > 0$ then (2.5) is verified.

REMARK 4. A close inspection of the proofs shows that all convergence results of Theorems 1, 2 and 3 are uniform over the class of parameters satisfying the following condition: there exist $c_0 > 0$, $C_0 > 0$ and $e_0 \in (0, 1)$ such that

$$\begin{aligned} &c_0 \leq \min_k \nu_k^0 \leq \max_k \nu_k^0 \leq C_0, \qquad \max_{k,\ell} \|h_{\ell,k}^0\|_\infty \leq C_0, \quad \text{and} \\ &\|\rho^0\| \leq 1 - e_0. \end{aligned} \tag{2.7}$$

A consequence of previous theorems is that the posterior mean $\hat{f} = \mathbb{E}^\pi[f|N]$ is converging to f_0 at the rate ε_T which is described by the following corollary:

COROLLARY 1. *Under the assumptions of Theorem 1 or Theorem 2, together with (2.5) with $\varepsilon_T = \sqrt{\log\log T}\epsilon_T$ and if $\int_{\mathcal{F}} \|f\|_1 \, d\Pi(f) < +\infty$, then for any $w_T \to +\infty$*

$$\mathbb{P}_0\big(\|\hat{f} - f_0\|_1 > w_T \varepsilon_T\big) = o(1).$$

The proof of Corollary 1 is given in Section 2.3 in the Supplementary Material [19].

The results of Theorem 3 and Corollary 1 lead to $\mathbb{L}_1$ convergence results which are weaker than the $\mathbb{L}_2$ convergence results of [26]. But our results allow for a much wider range of possible dictionaries (prior models in the Bayesian formulation) since, contrariwise to [26], we do not require the stringent (lower bound) condition on the Gram matrix G made of the scalar products between $\lambda_T(\varphi_j)$ and $\lambda_T(\varphi_{j'})$ with $(\varphi_j)_{j\le J}$ denoting the dictionary used to construct the estimator (see Inequality (2.4) of Theorem 1 of [26]). It is assumed, in particular, in [26] (see Proposition 5) that this dictionary has to be an orthonormal basis and some stringent conditions on J are considered. We see in the following section that no such a condition is required to apply Theorems 1 to 3, and priors based on overcomplete continuous dictionaries are easily allowed. Indeed, our assumptions resemble the type of assumptions considered for density estimation for i.i.d. models, for which a large literature already exists.

2.3. *Examples of prior models.* The advantage of Theorems 1 and 3 is that the conditions required on the priors on the functions $h_{k,\ell}$ are quite standard, in particular, if the functions $h_{k,\ell}$ are parameterized in the following way:

$$h_{k,\ell} = \rho_{k,\ell}\bar{h}_{k,\ell}, \qquad \int_0^A \bar{h}_{k,\ell}(u)\,du = 1.$$

We thus consider priors on $\theta = (\nu_\ell, \rho_{k,\ell}, \bar{h}_{k,\ell}, k, \ell \le K)$ following the scheme

$$\nu_\ell \overset{\text{i.i.d.}}{\sim} \Pi_\nu, \qquad \rho = (\rho_{k,\ell})_{k,\ell\le K} \sim \Pi_\rho, \qquad \bar{h}_{k,\ell} \overset{\text{i.i.d.}}{\sim} \Pi_{\bar{h}}, \tag{2.8}$$

with Π_ν, Π_ρ and $\Pi_{\bar{h}}$ independent. We consider Π_ν absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_+$ with positive and continuous density π_ν, Π_ρ a probability distribution on the set of matrices with positive entries and spectral norm $\|\rho\| < 1$, with positive density with respect to Lebesgue measures and satisfying (2.5). We now concentrate on the nonparametric part, namely, the prior distribution $\Pi_{\bar{h}}$. Then, from Theorems 1 and 3 it is enough that $\Pi_{\bar{h}}$ satisfies for all $1 \le k, \ell \le K$,

$$\Pi_{\bar{h}}(\|\bar{h} - \bar{h}^0_{k,\ell}\|_2 \le \epsilon_T, \|\bar{h}\|_\infty \le B) \ge e^{-cT\epsilon_T^2},$$

for some $B > 0$ and $c > 0$ such that there exists $\overline{\mathcal{H}_T}$ with

$$\overline{\mathcal{H}_T} \subset \left\{h : [0, A] \to \mathbb{R}^+, \int_0^A h(x)\,dx = 1\right\}$$

satisfying

$$\Pi_{\bar{h}}(\overline{\mathcal{H}_T}^c) \le e^{-CT\epsilon_T^2 \log\log T}, \qquad \log \mathcal{N}(\zeta\epsilon_T; \overline{\mathcal{H}_T}; \|\cdot\|_1) \le x_0 T\epsilon_T^2, \tag{2.9}$$

for $\zeta > 0$, $x_0 > 0$ and $C > 0$ large enough. Note that from Remark 1, if we have that for all ℓ, k

$$\Pi_{\bar{h}}(\|\bar{h} - \bar{h}^0_{k,\ell}\|_\infty \le \epsilon_T, \|\bar{h}\|_\infty \le B) \ge e^{-cT\epsilon_T^2},$$

then condition (2.9) can be replaced by

$$\Pi_{\bar{h}}(\overline{\mathcal{H}_T}^c) \le e^{-CT\epsilon_T^2}, \qquad \log \mathcal{N}(\zeta\epsilon_T; \overline{\mathcal{H}_T}; \|\cdot\|_1) \le x_0 T\epsilon_T^2. \tag{2.10}$$

These conditions have been checked for a large selection of types of priors on the set of densities. We discuss here two cases: one based on random histograms, these priors make sense in particular in the context of modeling neuronal interactions, and the second based on mixtures of Betas, because it leads to adaptive posterior concentration rates over a large collection of functional classes. To simplify the presentation we assume that $A = 1$ but generalization to any $A > 0$ is straightforward. Following Remark 4, we also assume in the sequel that there exist $c_0 > 0$, $C_0 > 0$ and $e_0 \in (0, 1)$ such that Condition (2.7) is satisfied.

2.3.1. *Random histogram prior.* These priors are motivated by the neuronal application, where one is interested in characterizing time zones when neurons are or are not interacting (see Section 3). Random histograms have been studied quite a lot recently for density estimation, both in semi and nonparametric problems. We consider two types of random histograms: regular partitions and random partitions histograms. Random histogram priors are defined as follows. For $J \geq 1$,

$$\begin{aligned} &\bar{h}_{w,t,J} = \delta \sum_{j=1}^{J} \frac{w_j}{t_j - t_{j-1}} \mathbb{1}_{I_j}, \qquad I_j = (t_{j-1}, t_j), \\ &\sum_{j=1}^{J} w_j = 1, \qquad \delta \sim \mathcal{B}ern(p) \end{aligned} \tag{2.11}$$

and

$$t_0 = 0 < t_1 < \cdots < t_J = 1.$$

In both cases, the prior is constructed in the following hierarchical manner:

$$\begin{aligned} &J \sim \Pi_J, \qquad e^{-c_1 x L_1(x)} \lesssim \Pi_J(J = x), \\ &\Pi_J(J > x) \lesssim e^{-c_2 x L_1(x)}, \\ &L_1(x) = 1 \quad \text{or} \quad L_1(x) = \log x, \\ &(w_1, \dots, w_J) | J \sim \Pi_w, \end{aligned} \tag{2.12}$$

where c_1 and c_2 are two positive constants. Denoting $\mathcal{S}_J$ the J-dimensional simplex, we assume that the prior on $(w_1, \dots, w_J)$ satisfies the following property. For all $M > 0$, for all $w_0 \in \mathcal{S}_J$ with for any j, $w_{0j} \leq M/J$ and all $u > 0$ small enough, there exists $c > 0$ such that

$$\begin{aligned} &\Pi_w\big((w_{01} - u/J^2, w_{01} + u/J^2) \\ &\quad \times \cdots \times (w_{0J} - u/J^2, w_{0J} + u/J^2)\big) > e^{-cJ \log J}. \end{aligned} \tag{2.13}$$

Many probability distributions on $\mathcal{S}_J$ satisfy (2.13). For instance, if Π_w is the Dirichlet distribution $\mathcal{D}(\alpha_{1,J}, \dots, \alpha_{J,J})$ with $c_3 J^{-a} \leq \alpha_{i,J} \leq c_4$, for a, c_3 and c_4 three positive constants, then (2.13) holds; see, for instance, Castillo and Rousseau [13]. Also, consider the following hierarchical prior allowing some of the w_j's to be equal to 0. Set $Z_j \overset{\text{i.i.d.}}{\sim} \mathcal{B}ern(p)$, $j \leq J$, $s_z = \sum_{j=1}^{J} Z_j$ and $(j_1, \dots, j_{s_z})$ the indices corresponding to $Z_j = 1$. Then,

$$\begin{aligned} &(w_{j_1}, \dots, w_{j_{s_z}}) \sim \mathcal{D}(\alpha_{1,J}, \dots, \alpha_{s_z,J}), \qquad c_3 J^{-a} \leq \alpha_{i,J} \leq c_4, \\ &w_j = 0 \quad \text{if } Z_j = 0. \end{aligned}$$

Regular partition histograms correspond to $t_j = j/J$ for $j \leq J$, in which case we write $\bar{h}_{w,J}$ instead of $\bar{h}_{w,t,J}$; while in random partition histograms, we put a prior on $(t_1, \dots, t_J)$. We now consider Hölder balls of smoothness $\beta \in (0, 1]$ and radius $L_0 > 0$, denoted $\mathcal{H}(\beta, L_0) := \{g; |g(x) - g(y)| \leq L_0 |x - y|^\beta\}$ and prove that the posterior concentration rate associated with both types of histogram priors is bounded by $\epsilon_T = \epsilon_0 (\log T/T)^{\beta/(2\beta+1)}$ for $0 < \beta \leq 1$, where ϵ_0 is a constant large enough. From Remark 1 we use the version of assumption (i) based on

$$B_\infty(\epsilon_T, B) = \Big\{ \big(\nu_k, (h_{\ell,k})_\ell\big)_k; \max_k |\nu_k - \nu_k^0| \leq \epsilon_T, \max_{\ell,k} \|h_{\ell,k} - h_{\ell,k}^0\|_\infty \leq \epsilon_T \Big\}$$

and need to verify (2.10). Then, applying Lemma 4 of the Supplementary Material of Castillo and Rousseau [13], we obtain that for each $1 \leq k, \ell \leq K$ such that $\bar{h}^0_{k,\ell} \neq 0$ and $\bar{h}^0_{k,\ell} \in \mathcal{H}(\beta, L_0)$,

$$\Pi(\|\bar{h}_{w,J} - \bar{h}^0_{\ell,k}\|_\infty \leq 2L_0 J^{-\beta} | J) \gtrsim p e^{-cJ \log T}$$

for some $c > 0$ and $\Pi_J(J = J_0 \lfloor (T/\log T) \rfloor^{1/(2\beta+1)}) \gtrsim e^{-c_1 J_0 (T/\log T)^{1/(2\beta+1)} L_1(T)}$ if J_0 is a constant. If $\bar{h}^0_{\ell,k} = 0$, then $\Pi(\|\bar{h}_{w,J} - \bar{h}^0_{\ell,k}\|_\infty = 0) = 1 - p$, so that

$$\Pi\big(B_\infty(\epsilon_T, B)\big) \gtrsim \epsilon_T^K \times \big[(1-p)p\big]^{K^2} \times e^{-K^2 c_1 J_0 (T/\log T)^{1/(2\beta+1)} L_1(T)} \gtrsim e^{-c' T \epsilon_T^2},$$

for some $c' > 0$. This result holds both for the regular grid and random grid histograms with a prior on the grid points $(t_1, \dots, t_J)$ given by $(u_1, \dots, u_J) \sim \mathcal{D}(\alpha, \dots, \alpha)$ with $u_j = t_j - t_{j-1}$. Then, condition (2.5) is verified if $\Pi(\|\rho\| > 1 - u) \lesssim e^{-a' u^{-a}}$ with $a > 3/\beta$ and $a' > 0$, for u small enough. This condition holds for any $\beta \in (0, 1]$ if there exist $a', \tau > 0$ such that when u is small enough

$$\Pi\big(\|\rho\| > 1 - u\big) \lesssim e^{-a' e^{-1/u^\tau}}. \tag{2.14}$$

Moreover, set $\overline{\mathcal{H}_T} = \{\bar{h}_{w,J}, J \leq J_1 (T/\log T)^{1/(2\beta+1)}\}$ for J_1 a constant, then for all $\zeta > 0$, $\log \mathcal{N}(\zeta \epsilon_T, \overline{\mathcal{H}_T}, \|\cdot\|_1) \lesssim J_1 (T/\log T)^{1/(2\beta+1)} \log T$. Therefore, (2.10) is checked. We finally obtain the following corollary:

COROLLARY 2 (Regular partition). *Under the random histogram prior* (2.11) *based on a regular partition and verifying* (2.12) *and* (2.13) *and if* (2.14) *is satisfied, then if for any* $k, \ell = 1, \dots, K$, $h^0_{k,\ell}$ *belongs to* $\mathcal{H}(\beta, L)$ *for* $0 < \beta \leq 1$, *for any* $w_T \to +\infty$,

$$\mathbb{E}_0\big[\Pi\big(\|f - f_0\|_1 > w_T (T/\log T)^{-\beta/(2\beta+1)} | N\big)\big] = o(1).$$

To extend this result to the case of random partition histogram priors, we consider the same prior on $(J, w_1, \dots, w_J)$ as in (2.12) and the following condition on the prior on $\underline{t} = (t_1, \dots, t_K)$. Writing $u_1 = t_1$, $u_j = t_j - t_{j-1}$, we have that $\underline{u} = (u_1, \dots, u_J)$ belongs to the J-dimensional simplex $\mathcal{S}_J$, and we consider a Dirichlet distribution on $(u_1, \dots, u_J)$, $\mathcal{D}(\alpha, \dots, \alpha)$ with $\alpha \geq 6$. The arguments used to the regular partition apply also to the case of the random partition apart from the computation of the entropy, which is more involved here.

COROLLARY 3. *Consider the random histogram prior* (2.11) *based on random partition with a prior on* $\underline{w} = (w_1, \dots, w_J)$ *satisfying* (2.12) *and* (2.13) *and with a Dirichlet prior on* $\underline{u} = (t_j - t_{j-1}, j \leq J)$ *with parameter* $\alpha \geq 6$. *If* (2.14) *is satisfied, then if for any* $k, \ell = 1, \dots, K$, $h^0_{k,\ell}$ *belongs to* $\mathcal{H}(\beta, L)$ *for* $0 < \beta \leq 1$, *for any* $w_T \to +\infty$,

$$\mathbb{E}_0\big[\Pi\big(\|f - f_0\|_1 > w_T (T/\log T)^{-\beta/(2\beta+1)} | N\big)\big] = o(1).$$

The proof of this corollary is given in Section 2.6 in the Supplementary Material [19]. In the following section we consider another family of priors suited for smooth functions $h_{k,\ell}$ and based on mixtures of Beta distributions.

2.3.2. *Mixtures of betas.* The following family of prior distributions is inspired by Rousseau [43]. Consider functions

$$h_{k,\ell} = \rho_{k,\ell} \left(\int_0^1 g_{\alpha_{k,\ell},\epsilon} \, dM_{k,\ell}(\epsilon) \right)_+,$$

$$g_{\alpha,\epsilon}(x) = \frac{\Gamma(\alpha/(\epsilon(1-\epsilon)))}{\Gamma(\alpha/\epsilon)\Gamma(\alpha/(1-\epsilon))} x^{\frac{\alpha}{1-\epsilon}-1} (1-x)^{\frac{\alpha}{\epsilon}-1},$$

where $M_{k,\ell}$ are bounded signed measures on $[0, 1]$ such that $|M_{k,\ell}| = 1$. In other words, the above functions are the positive parts of mixtures of Betas' distributions with parameterization $(\alpha/\epsilon, \alpha/(1-\epsilon))$ so that ϵ is the mean parameter. The mixing random measures $M_{k,\ell}$ are allowed to be negative. The reason for allowing $M_{k,\ell}$ to be negative is that $h_{k,\ell}$ is then allowed to be null on sets with positive Lebesgue measure. The prior is then constructed in the following way. Writing $h_{k,\ell} = \rho_{k,\ell}\bar{h}_{k,\ell}$, we define a prior on $\bar{h}_{k,\ell}$ via a prior on $M_{k,\ell}$ and on $\alpha_{k,\ell}$. In particular, we assume that $M_{k,\ell} \overset{\text{i.i.d.}}{\sim} \Pi_M$ and $\alpha_{k,\ell} \overset{\text{i.i.d.}}{\sim} \pi_\alpha$. As in Rousseau [43], we consider a prior on α absolutely continuous with respect to Lebesgue measure and with density such that there exists $b_1, c_1, c_2, c_3, A, C > 0$ such that for all u large enough,

$$\begin{aligned} \pi_\alpha(c_1 u < \alpha < c_2 u) &\geq Ce^{-b_1 u^{1/2}}, \\ \pi_\alpha(\alpha < e^{-Au}) + \pi_\alpha(\alpha > c_3 u) &\leq Ce^{-b_1 u^{1/2}}. \end{aligned} \tag{2.15}$$

For instance, if $\sqrt{\alpha}$ follows a Gamma distribution, then (2.15) is verified. There are many ways to construct discrete signed measures on $[0, 1]$, for instance, writing

$$M = \sum_{j=1}^{J} r_j p_j \delta_{\epsilon_j}, \tag{2.16}$$

the prior on M is then defined by $J \sim \Pi_J$ and conditionally on J,

$$r_j \overset{\text{i.i.d.}}{\sim} \text{Ra}(1/2), \qquad \epsilon_j \overset{\text{i.i.d.}}{\sim} G_\epsilon, \qquad (p_1, \dots, p_J) \sim \mathcal{D}(a_1, \dots, a_J),$$

where Ra denotes the Rademacher distribution taking values $\{-1, 1\}$ each with probability $1/2$. Assume that G_ϵ has positive continuous density on $[0, 1]$ and that there exists $A_0 > 0$ such that $\sum_{j=1}^{J} a_j \leq A_0$. Recall that when $\beta > 1$, the Hölder ball $\mathcal{H}(\beta, L_0)$ is defined as the set of functions g such that

$$\|g\|_\infty + \sum_{\ell=1}^{\lfloor\beta\rfloor} \|g^{(\ell)}\|_\infty + \sup_{x \neq y} \frac{|g^{(\lfloor\beta\rfloor)}(x) - h^{(\lfloor\beta\rfloor)}(y)|}{|x-y|^{\beta-\lfloor\beta\rfloor}} \leq L_0,$$

where the last term disappears if β is an integer. We have the following corollary:

COROLLARY 4. *Consider a prior as described above. Assume that for all $k, \ell \leq K$ $h_{k,\ell}^0 = (g_{k,\ell}^0)_+$ for some functions $g_{k,\ell}^0 \in \mathcal{H}(\beta, L_0)$ with $\beta > 0$. If condition* (2.14) *holds and if G_ϵ has density with respect to Lebesgue measure verifying*

$$x^{A_1}(1-x)^{A_1} \lesssim g_\epsilon(x) \lesssim x^3(1-x)^3 \quad \textit{for some } A_1 \geq 3,$$

then, for any $w_T \to +\infty$,

$$\mathbb{E}_0\big[\Pi\big(\|f - f_0\|_1 > w_T T^{-\beta/(2\beta+1)}(\log T)^{5\beta/(4\beta+2)}\sqrt{\log\log T} | N\big)\big] = o(1).$$

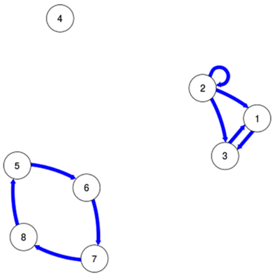
Note that in the context of density estimation, $T^{-\beta/(2\beta+1)}$ is the minimax rate, and we expect that it is the same for Hawkes processes. Indeed, since $\mathbb{P}_0(\Omega_T^c)$ goes to 0, the number of observations is of order T.

3. Numerical illustration in the neuroscience context. It is now well-known that neurons receive and transmit signals as electrical impulses called *action potentials*. Although action potentials can vary somewhat in duration, amplitude and shape, they are typically treated as identical stereotyped events in neural coding studies. Therefore, an action potentials sequence, or *spike train*, can be characterized simply by a series of all-or-none point events in time. Multivariate Hawkes processes have been used in neuroscience to model spike trains of several neurons and in particular to model functional connectivity between them through mutual excitation or inhibition [29]. In this section, we conduct a simulation study mimicking the neural context, through appropriate choices of parameters. The protocol is similar to the setting proposed in Section 6 of [26].

3.1. *Simulation scenarios.* We consider three simulation scenarios involving respectively $K = 2$ and $K = 8$ neurons. The scenarios are roughly similar to the one tested in [26]. Following the notation introduced in the previous sections, for any $(k, \ell) \in \{1, \dots, K\}^2$, $h_{k,\ell}$ denotes the interaction function of neuron k over neuron ℓ. We now describe the three scenarios. The upper bound of each $h_{k,\ell}$'s support, denoted $[0, A]$, is set equal to $A = 0.04$ seconds.

- *Scenario 1.* We first consider $K = 2$ neurons and piecewise constant interactions: $h_{1,1} = 30 \cdot \mathbb{1}_{(0,0.02]}$, $h_{2,1} = 30 \cdot \mathbb{1}_{(0,0.01]}$, $h_{1,2} = 30 \cdot \mathbb{1}_{(0.01,0.02]}$, $h_{2,2} = 0$.
- *Scenario 2.* In this scenario, we mimic $K = 8$ neurons belonging to three independent groups. The non-null interactions are the piecewise constant functions defined as: $h_{2,1} = h_{3,1} = h_{2,2} = h_{1,3} = h_{2,3} = h_{8,5} = h_{5,6} = h_{6,7} = h_{7,8} = 30 \cdot \mathbb{1}_{(0,0.02]}$.

  We plot the subsequent interactions directed graph between the 8 neurons: the vertices represent the K neurons and an oriented edge is plotted from vertex k to vertex ℓ if the interaction function $h_{k,\ell}$ is non-null.



- *Scenario 3.* Setting $K = 2$, we consider non-piecewise constant interactions functions defined as

$$h_{1,1}(t) = 100 \cdot e^{-100t} \mathbb{1}_{(0,0.04]}(t), \qquad h_{2,1}(t) = 30 \cdot \mathbb{1}_{(0,0.02]}(t),$$

$$h_{1,2}(t) = \frac{1}{2 \times 0.004\sqrt{2\pi}} e^{-\frac{(t-0.02)^2}{2\times 0.004^2}} \cdot \mathbb{1}_{(0,0.04]}(t), \qquad h_{2,2}(t) = 0.$$

For any scenario, we consider $\nu_\ell = 20$, $\ell = 1, \dots, K$. For each scenario, we simulate 25 datasets on the time interval $[0, 22]$ seconds. The Bayesian inference is performed considering recordings on three possible periods of length $T = 5$ seconds, $T = 10$ seconds and $T = 20$ seconds. For any dataset, we remove the initial period of 2 seconds corresponding to 50 times the length of the support of the $h_{k,\ell}$-functions, assuming that, after this period, the Hawkes processes have reached their stationary distribution. Note that the chosen parameters induce that the mean number of events per neuron and per period of five seconds is approximatively 321 for *Scenario 1*, 472 for *Scenario 2* and 317 for *Scenario 3*. More details on the simulated dataset are supplied in the supplementary material [19].

3.2. *Prior distribution on $f = (\nu_\ell, h_{k,\ell})_{k,\ell\in\{1,\dots,K\}}$.* We use the prior distribution described in Section 2.3 setting a log-normal prior distribution on the ν_ℓ's of parameter μ_ν, s_ν^2. About the interaction functions $(h_{k,\ell})_{k,\ell\in\{1,\dots,K\}}$, the prior distribution is defined on the set of piecewise constant functions, $h_{k,\ell}$ being written as follows:

$$h_{k,\ell}(t) = \delta^{(k,\ell)} \sum_{j=1}^{J^{(k,\ell)}} \beta_j^{(k,\ell)} \mathbb{1}_{[t_{j-1}^{(k,\ell)}, t_j^{(k,\ell)}]}(t), \qquad \rho_{k,\ell} = \int_0^\infty h_{k,\ell}(t)\, dt \tag{3.1}$$

with $t_0^{(k,\ell)} = 0$ and $t_{J^{(k,\ell)}}^{(k,\ell)} = A$. Using the notation in Section 2.3, we have if $\delta^{(k,\ell)} \neq 0$, $\beta_j^{(k,\ell)} = \rho_{k,\ell} \frac{\omega_j^{(k,\ell)}}{(t_j^{(k,\ell)} - t_{j-1}^{(k,\ell)})}$. Here, $\delta^{(k,\ell)}$ is a global parameter of nullity for $h_{k,\ell}$ and model the graph of interactions: for all $(k, \ell) \in \{1, \dots, K\}^2$,

$$\delta^{(k,\ell)} \sim_{\text{i.i.d.}} \mathcal{B}ern(p). \tag{3.2}$$

For all $(k, \ell) \in \{1, \dots, K\}^2$, the number of steps $(J^{(k,\ell)})$ follows a translated Poisson prior distribution:

$$J^{(k,\ell)} | \{\delta^{(k,\ell)} = 1\} \sim_{\text{i.i.d.}} 1 + \mathcal{P}(\eta). \tag{3.3}$$

To minimize the influence of η on the posterior distribution, we consider a hyperprior distribution on the hyperparameter η:

$$\eta \sim \Gamma(a_\eta, b_\eta). \tag{3.4}$$

Given $J^{(k,\ell)}$, we consider a spike and slab prior distribution on $(\beta_j^{(k,\ell)})_{j=1,\dots,J^{(k,\ell)}}$. Let us introduce $Z_j^{(k,\ell)} \in \{0, 1\}$ such that $\forall j \in \{1, \dots, J^{(k,\ell)}\}$,

$$\begin{aligned} &\mathbb{P}(Z_j^{(k,\ell)} = z | \delta^{(k,\ell)} = 1) = \pi_z \quad \forall z \in \{0, 1\}, \\ &\beta_j^{(k,\ell)} | \{\delta^{(k,\ell)} = 1\} \sim Z_j^{(k,\ell)} \times \log\mathcal{N}(\mu_\beta, s_\beta^2). \end{aligned} \tag{3.5}$$

We consider two prior distributions on $(t_j^{(k,\ell)})_{j=1,\dots,J^{(k,\ell)}}$. The first one (referred as the regular histogram prior) is a regular partition of $[0, A]$:

$$t_j^{(k,\ell)} = \frac{j}{J^{(k,\ell)}} A \quad \forall j = 0, \dots, J^{(k,\ell)}. \tag{3.6}$$

The second prior distribution is referred as the random histogram prior and

$$\begin{aligned} &(u_1, \dots, u_{J^{(k,\ell)}}) \sim \mathcal{D}(\alpha'_1, \dots, \alpha'_{J^{(k,\ell)}}), \\ &t_j^{(k,\ell)} = A \sum_{r=1}^{j} u_r \quad \forall j = 1, \dots, J^{(k,\ell)}; \qquad t_0^{(k,\ell)} = 0 \end{aligned} \tag{3.7}$$

Equations (3.2)–(3.6) (or (3.7)) define a prior distribution $\mathbb{P}$ on $(h_{k,\ell})_{k,\ell}$, without any constraint on $\|\rho\|$. The prior is defined by truncating this distribution to the set $\{\|\rho\| \leq 1 - \epsilon\}$ for an arbitrarily small ϵ. In practice we have chosen $\epsilon = 10^{-16}$, which is the precision of the machine. In the simulation studies, we set the following hyperparameters:

$$\mu_\beta = 3.5, \qquad s_\beta = 1, \qquad \mu_\nu = 3.5, \qquad s_\nu = 1,$$
$$\mathbb{P}(Z_j^{(k,\ell)} = 1) = 1/2, \qquad \mathbb{P}(\delta^{(k,\ell)} = 1) = p = 1/2; \qquad \alpha'_j = 2 \quad \forall j.$$

3.3. *Posterior sampling.* The posterior distribution is sampled using a standard reversible-jump Markov chain Monte Carlo (RJ-MCMC). Considering the current parameter $(\boldsymbol{\nu}, \boldsymbol{h})$, $\boldsymbol{\nu}^{(c)}$ is proposed using a Metropolis-adjusted Langevin proposal. For a fixed $J^{(k,\ell)}$, the heights $\beta_j^{(k,\ell)}$ are proposed using a random walk proposing null or non-null candidates. Changes in the number of steps $J^{(k,\ell)}$ are generated by standard birth and death moves [24]. In this simulation study, we generate chains of length 30,000 removing the first 10,000 burn-in iterations. The algorithm is implemented in R on an Intel(R) Xeon(R) CPU E5-1650 v3 @ 3.50 GHz.

TABLE 1
Mean computation time (in seconds) for the reversible-jump MCMC algorithms as a function of the scenario, the observation time interval and the prior distribution (random or regular histogram). The mean is computed over the 25 simulated datasets

	$K=2$		$K=8$	$K=2$ with smooth $h_{k,\ell}$
Prior on t:	Regular	Random	Regular	Random
$T=5$	1508.44	1002.45		823.84
$T=10$	1383.72	1459.55	37,225.19	1284.93
$T=20$	2529.19	2602.48	49,580.18	1897.17

REMARK 5. Note that, in order to get a better mixing Markov Chain, we first sample the posterior distribution of f on the unconstraint parameter set, i.e. not taking into account the constraint $\|\rho\| \leq 1-\epsilon$, and we discard all iterations where $\|\rho\| > 1-\epsilon$.

The computation times (mean over the 25 datasets) are given in Table 1. First note that the computation time increases roughly as a linear function of T. This is due to the fact that the heavier task in the algorithm is the integration of the conditional likelihood and the computation time of this operation is roughly a linear function of the length of the integration (observation) time interval. Besides, because we implemented a reversible-jump MCMC algorithm, the computation time is a stochastic quantity: the algorithm can explore parts of the domain where the number of bins $J^{(k,\ell)}$ is large, thus increasing the computation time. Moreover, we remark that the computation time explodes as K increases (due to the fact that K^2 interaction functions have to be estimated), reaching computation times greater than a day.

3.4. *Results.* We describe here the results for each scenario. We consider the previous scenarios, three observation durations T and two prior distributions. In Table 2, we supply the estimated $\mathbb{L}_1$-distances on the λ^k's and the $h_{k,\ell}$'s. More precisely, we evaluate the estimated average posterior expectation of the $\mathbb{L}_1$-distances on the $h_{k,\ell}$'s:

$$D_h = \frac{1}{25}\sum_{sim=1}^{25} \widehat{\mathbb{E}}\Big[\frac{1}{K^2}\sum_{k,\ell=1}^{K} \left\|h_{k,\ell} - h^0_{k,\ell}\right\|_1 |(N_t^{sim})_{t\in[0,T]}\Big], \tag{3.8}$$

and the estimated average posterior expectation of the renormalized pseudo-distance $d_{1,T}$ on the parameters:

$$D_\lambda = \frac{1}{25}\sum_{sim=1}^{25} \widehat{\mathbb{E}}\left[\frac{1}{K} d_{1,T}(f, f_0) \middle| (N_t^{sim})_{t\in[0,T]}\right], \tag{3.9}$$

TABLE 2
Posterior expectations of the distances D_λ and D_h

		$K=2$		$K=8$	$K=2$ with smooth $h_{k,\ell}$
	Prior	Regular	Random	Regular	Random
D_λ	$T=5$	5.79	4.79		5.87
	$T=10$	3.74	3.16	0.70	4.74
	$T=20$	2.70	2.05	0.39	3.95
D_h	$T=5$	0.1423	0.0996		0.1431
	$T=10$	0.0844	0.0578	0.1199	0.1131
	$T=20$	0.0564	0.0336	0.0616	0.0945

where $\widehat{\mathbb{E}}$ refers to the Monte Carlo estimator obtained as a by product of the RJ-MCMC algorithm. To compute the Monte Carlo posterior expectations given in Equations (3.8) and (3.9), we consider the outputs of the reversible jumps MCMC algorithm, then evaluate the functions $h_{k,\ell}$ and λ^k on a fine grid and finally compute the mean. Observe that the distances have been normalized by the number of estimated functions (K^2 for the $h_{k,\ell}$'s and K for the λ^k's). As a consequence, we can compare results obtained in the three scenarios and reported in Table 2.

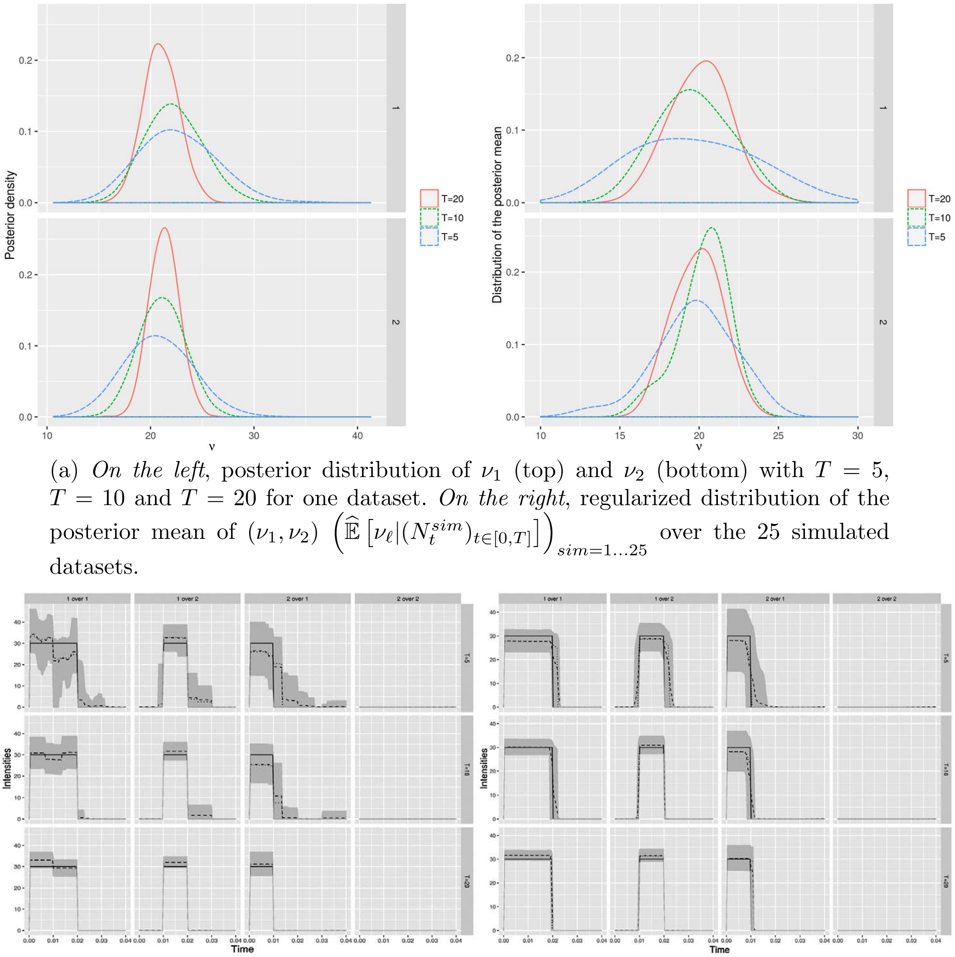
As expected, Table 2 illustrates the fact that the error decreases as T increases. As we will detail later, the random histogram prior gives better results than the regular prior. Finally, performances are better when the true interaction function $h_{k,\ell}$ are step functions (due to the form of the prior distribution).

3.4.1. *Results for scenario* 1: $K = 2$ *with step functions.* When $K = 2$, we estimate the parameters using both *regular* and *random* prior distributions on $(t_j^{(k,\ell)})$ (equations (3.6) and (3.7)). One typical posterior distribution of ν_ℓ is given in Figure 1a (left), for a randomly chosen dataset and the regular histogram, clearly showing a smaller variance when the length of the observation interval increases. We also present the global estimation results, over the 25 simulated datasets for the regular prior. The regularized distribution of the posterior mean estimators for (ν_1, ν_2) computed for the 25 simulated datasets $(\widehat{\mathbb{E}}[\nu_\ell|(N_t^{sim})_{t\in[0,T]}])_{sim=1\dots25}$ is given in Figure 1a on the right panel, showing an expected decreasing variance for the estimator as T increases. We only supply the plots for the regular histogram prior. The plots corresponding to the random histogram prior are supplied in the supplementary material [19] and are similar to the one presented her.

Regarding the estimate of the interaction functions, for the same given dataset, the estimation of the $h_{k,\ell}$'s is plotted in Figure 1b (left panel) for the regular prior, with its credible interval. Its corresponding estimation with the random prior is given in Figure 1b (right panel). For both prior distributions, the functions are globally well estimated, showing a clear concentration when T increases. The regions where the interaction functions are null are also well identified. The estimation associated with the random histogram prior is in general better than the one supplied by the regular prior. This may be due to several factors. First, the random histogram prior leads to a sparser estimation than the regular one. Secondly, it is easier to design a proposal move in the reversible-jump MCMC algorithm in the former case than in the latter context.

Moreover, the interaction graph is perfectly inferred since the posterior probability for $\delta^{(2,2)}$ to be 0 is almost 1. For the 25 datasets, we estimate the posterior probabilities $\widehat{\mathbb{P}}(\delta^{(k,\ell)} = 1|(N_t^{sim})_{t\in[0,T]})$ for $k, \ell = 1, 2$ and $sim = 1, \dots, 25$. In Table 3, we display the mean of these posterior quantities. Even for the shorter observation time interval ($T = 5$) these quantities—defining completely the connexion graph—are well recovered. These results are improved when T increases. Once again, the random histogram prior (3.7) gives slightly better results.

Finally, we also have a look at the conditional intensities λ_t^k. On Figure 2, we plot 50 realizations of the conditional intensity from the posterior distributions. More precisely, for one given dataset, for 50 parameters $\theta^{(i)} = ((h_{k,\ell}^{(i)})_{k,\ell}, (\nu_k^{(i)})_{k=1\dots K})$ sampled from the posterior distribution (issued from the RJ-MCMC chain), we compute the corresponding $(\lambda_t^{k(i)})$ and plot them. For the sake of clarity, only the conditional intensity of the first process ($k = 1$) is plotted and we restrict the graph to a short time interval $[3.2, 3.6]$. As noticed before, the conditional intensity is well reconstructed, with a clear improvement of the precision as the length of the observation time T increases.



(a) *On the left*, posterior distribution of ν_1 (top) and ν_2 (bottom) with $T = 5$, $T = 10$ and $T = 20$ for one dataset. *On the right*, regularized distribution of the posterior mean of (ν_1, ν_2) $\left(\widehat{\mathbb{E}}\left[\nu_\ell|(N_t^{sim})_{t\in[0,T]}\right]\right)_{sim=1\ldots25}$ over the 25 simulated datasets.

(b) Estimation of the $(h_{k,\ell})_{k,\ell=1,2}$ using the regular prior (left panel) and the random histogram prior (right panel). The gray region indicates the credible region for $h_{k,\ell}(t)$ (delimited by the 5% and 95% percentiles of the posterior distribution). The true $h_{k,\ell}$ is in plain line, the posterior expectation and posterior median for $h_{k,\ell}(t)$ are in dotted and dashed lines respectively.

FIG. 1. *Results of* scenario 1: *estimation of* $(h_{k,\ell})_{k,\ell=1,2}$ *and* $(\nu_k)_{k=1,2}$.

TABLE 3

Results of Scenario 1. *Mean of the posterior estimations*: $\frac{1}{25}\sum_{sim=1}^{25}\widehat{\mathbb{P}}(\delta^{(k,\ell)} = 1|(N_t^{sim})_{t\in[0,T]})$, *for the three observation time intervals and the two prior distributions*

ℓ over k:		1 over 1	1 over 2	2 over 1	2 over 2
True value of $\delta^{(k,\ell)}$:		1	1	1	0
	Prior				
$T = 5$	Regular	1.0000	0.8970	1.0000	0.0071
	Random	1.0000	0.9812	1.0000	0.0196
$T = 10$	Regular	1.0000	0.9954	1.0000	0.0047
	Random	1.0000	1.0000	1.0000	0.0102
$T = 20$	Regular	1.0000	1.0000	1.0000	0.0099
	Random	1.0000	1.0000	1.0000	0.0102

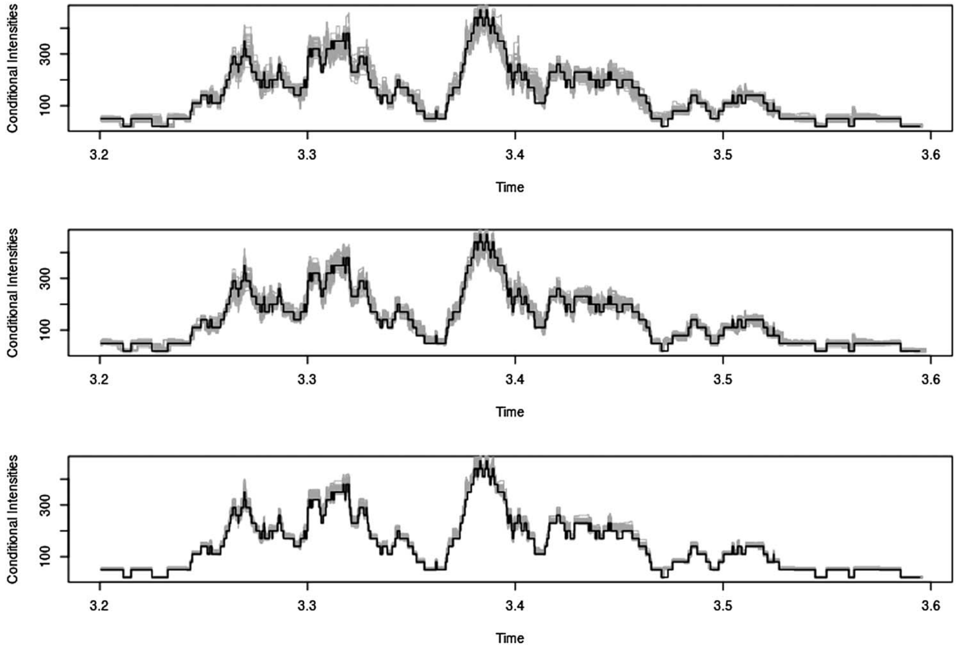


FIG. 2. *Results for* scenario 1. *Conditional intensity* λ_t^1: 50 *realizations of* λ_t^1 *from the posterior distribution for one particular dataset and* 3 *lengths of observation interval* ($T = 5$ *on the first row*, $T = 10$ *on the second row, and* $T = 20$ *on the third row*). *True conditional intensity in black plain line.*

3.4.2. *Results for scenario* 2: $K = 8$. In this scenario, we perform the Bayesian inference using only the regular prior distribution on $(\mathbf{t}^{(k,\ell)})_{(k,\ell)\in\{1,\dots,K\}^2}$ and two lengths of observation interval ($T = 10$ and $T = 20$). Here we set $a_\eta = 3$ and $b_\eta = 1$.

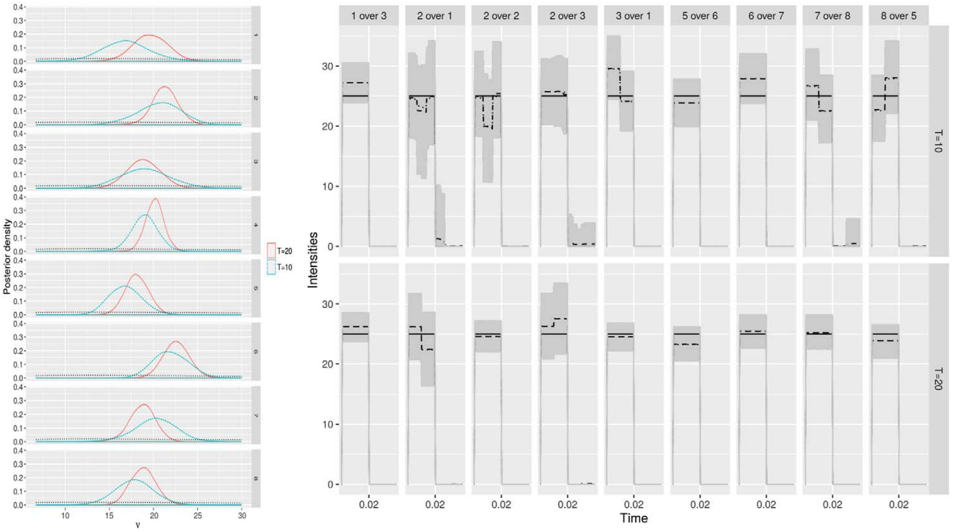
The posterior distribution of the ν_k's for a randomly chosen dataset is plotted in Figure 3a. The prior distribution is in dotted line and is flat. The posterior distribution concentrates around the true value (here 20) with a smaller variance when T increases.

In Figure 3b, we plot the posterior means (with credible regions) of the non-null interaction functions for the same randomly chosen dataset. The time intervals where the interaction functions are null are again perfectly recovered. The posterior uncertainty around the non-null functions $h_{k,\ell}$ decreases when T increases.

In the context of neurosciences, we are especially interested in recovering the interaction graph of the $K = 8$ neurons. In Figure 4a, we consider the same dataset as the one used in Figures 3a and 3b and plot the posterior estimation of the interaction graph, for respectively $T = 10$ on the left and $T = 20$ on the right. The width and the gray level of the edges are proportional to the estimated posterior probability $\widehat{\mathbb{P}}(\delta^{(k,\ell)} = 1|(N_t)_{t\in[0,T]})$. The global structure of the graph is recovered (to be compared to the true graph plotted before). We observe that the false positive edges appearing when $T = 10$ disappear when $T = 20$. In Figure 4b, we consider the mean of the estimates of the graph over the 25 datasets. The resulting graph for $T = 10$ is on the left and for $T = 20$ on the right.

Note that, in this example, for any (k,ℓ) such that the true $\delta^{(k,\ell)} = 1$, the estimated posterior probability $\widehat{\mathbb{P}}(\delta^{(k,\ell)} = 1|(N_t^{sim})_{t\in[0,T]})$ is equal to 1, for any dataset and any length of observation interval. In other words, the non-null interactions are perfectly recovered. In a simulation scenario with other interaction functions, the results could have been different.

3.4.3. *Results for scenario* 3: $K = 2$ *with smooth functions*. In this context, we perform the inference using the random histogram prior distribution (3.7). In this case, we set $a_\eta = 10$

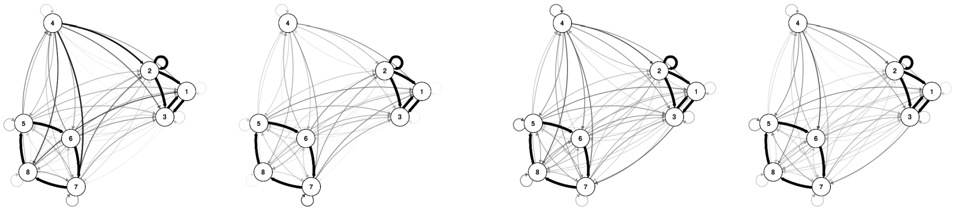


(a) Results on $(\nu_\ell)_{k=1\dots K}$ for a particular dataset. Prior distribution (dotted line), Posterior distributions for $T = 10$ (dashed line) and $T = 20$ (plain line).

(b) Estimation of the non null interaction functions $(h_{k,\ell})_{k,\ell=1,\dots,8}$ using the regular prior for $T = 10$ (upper panel) and $T = 20$ (bottom). The gray region indicates the credible region for $h_{k,\ell}(t)$ (delimited by the 5% and 95% percentiles of the posterior distribution). The true $h_{k,\ell}$ is in plain line, the posterior expectation and posterior median for $h_{k,\ell}(t)$ are in dotted and dashed lines respectively (often indistinguishable).

FIG. 3. *Results of* Scenario 2 *for one given dataset.*

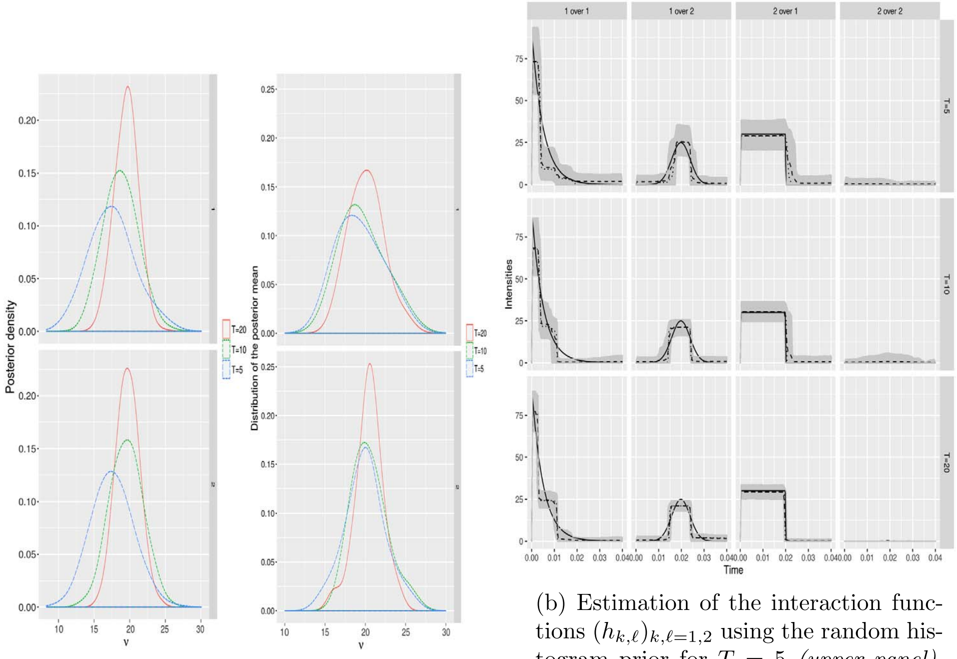
and $b_\eta = 1$, thus encouraging a larger number of steps in the interactions functions. The behavior of the posterior distribution of ν_k is the same as in the other examples. In Figure 5a, we plot the distribution of $(\mathbb{E}[\nu_k|(N_t^{sim})_{t\in[0,T]}])_{sim=1\dots25}$ for $T = 5, 10, 20$ seconds and clearly observe a decrease of the bias and the variance as the length of the observation period increases. Some estimation of the interaction functions is given in Figure 5b. Due to the choice of the prior distribution of these quantities, we get a sparse posterior inference. Note that, like in the other scenarios, the null interaction is clearly identified, making possible to recover to true posterior graph of interactions.



(a) For one given dataset. Posterior estimation of the interaction graph for $T = 10$ *on the left* and $T = 20$ *on the right*, for one randomly chosen dataset. Level of grey and width of the edges proportional to $\widehat{\mathbb{P}}(\delta^{(k,\ell)} = 1|(N_t^{sim})_{t\in[0,T]})$, $sim = 1$.

(b) For the 25 simulated datasets. Posterior estimation of the interaction graph for $T = 10$ *on the left* and $T = 20$ *on the right*, averaged over all the datasets. Level of grey and width of the edges proportional to $\frac{1}{25}\sum_{sim=1}^{25}\widehat{\mathbb{P}}(\delta^{(k,\ell)} = 1|(N_t^{sim})_{t\in[0,T]})$.

FIG. 4. *Results of* Scenario 2: *interaction graphs.*



(a) *On the left*, posterior distribution of ν_1 (top) and ν_2 (bottom) with $T = 5$, $T = 10$ and $T = 20$ for one dataset. *On the right*, regularized distribution of the posterior mean of (ν_1, ν_2) $\left(\widehat{\mathbb{E}}\left[\nu_\ell | (N_t^{sim})_{t\in[0,T]}\right]\right)_{sim=1\dots 25}$ over the 25 simulated datasets.

(b) Estimation of the interaction functions $(h_{k,\ell})_{k,\ell=1,2}$ using the random histogram prior for $T = 5$ *(upper panel)*, $T = 10$ *(middle) and* $T = 20$ *(bottom)*. The gray region indicates the credible region for $h_{k,\ell}(t)$ (delimited by the 5% and 95% percentiles of the posterior distribution). The true $h_{k,\ell}$ is in plain line, the posterior expectation and posterior median for $h_{k,\ell}(t)$ are in dotted and dashed lines respectively (often undistinguishable).

FIG. 5. *Results of* Scenario 3.

4. Proofs of theorems. In the sequel, specific tests to deal with the numerator of posterior distributions are first built in Section 4.1. The denominator is controlled by using upper bounds of Section 4.2. We finally provide the proof of Theorem 3 in Section 4.3. Other technical proofs are provided in the Supplementary Material [19] which contains, in particular, the proofs of Theorems 1 and 2, the proof of Corollary 1 and the proofs of results of Section 2.3.

4.1. *Construction of tests.* As usual, the control of the posterior distributions is based on specific tests. We build them in the following lemma whose proof is given in the Supplementary Material [19]. Our tests are based on ideas similar to $\mathbb{L}_1$-tests for density estimation but adapted to the more complex framework of Hawkes processes. To build them, we use specific Bernstein-type concentration inequalities for martingales established in [26], which leads to the natural use of the $\mathbb{L}_1$-loss. Moreover, the subsequent tests also take into account the fact that the metric is random, which makes their construction slightly more involved.

LEMMA 1. *Let $j \geq 1$, $f_1 \in \mathcal{F}_j$ and define the test*

$$\phi_{f_1,j} = \max_{\ell=1,\dots,K} \left(\mathbb{1}_{\{N^\ell(A_{1,\ell}) - \Lambda^\ell(A_{1,\ell}; f_0) \geq jT\epsilon_T/8\}} \vee \mathbb{1}_{\{N^\ell(A^c_{1,\ell}) - \Lambda^\ell(A^c_{1,\ell}; f_0) \geq jT\epsilon_T/8\}}\right),$$

with for all $\ell \le K$, $A_{1,\ell} = \{t \in [0,T]; \lambda_t^\ell(f_1) \ge \lambda_t^\ell(f_0)\}$, $\Lambda^\ell(A_{1,\ell}; f_0) = \int_0^T \mathbb{1}_{A_{1,\ell}}(t)\lambda_t^\ell(f_0)\,dt$ *and* $\Lambda^\ell(A_{1,\ell}^c; f_0) = \int_0^T \mathbb{1}_{A_{1,\ell}^c}(t)\lambda_t^\ell(f_0)\,dt$. *Then*

$$\begin{aligned}\mathbb{E}_0[\mathbb{1}_{\Omega_T}\phi_{f_1,j}] &+ \sup_{\|f-f_1\|_1 \le j\epsilon_T/(6N_0)} \mathbb{E}_0[\mathbb{E}_f[\mathbb{1}_{\Omega_T}\mathbb{1}_{f\in S_j}(1-\phi_{f_1,j})|\mathcal{G}_0]]\\ &\le (2K+1)\max_\ell e^{-x_{1,\ell}Tj\epsilon_T(\sqrt{\mu_\ell^0}\wedge j\epsilon_T)},\end{aligned}$$

with N_0 *is defined in Section* 2 *and* $x_{1,\ell} = \min(36, 1/(4096\mu_\ell^0), 1/(1024K\sqrt{\mu_\ell^0}))$.

4.2. *Control of the denominator.*

LEMMA 2. *Let*

$$\mathrm{KL}(f_0, f) = \mathbb{E}_0[L_T(f_0) - L_T(f)].$$

On $B(\epsilon_T, B)$,

$$0 \le \mathrm{KL}(f_0, f) \le \kappa \log(r_T^{-1})T\epsilon_T^2, \tag{4.1}$$

for T *larger than* T_0, *with* T_0 *some constant depending on* f_0, *with*

$$\kappa = 4\sum_{k=1}^K (\nu_k^0)^{-1}\left(3 + 4K\sum_{\ell=1}^K (A\mathbb{E}_0[(\lambda_0^\ell(f_0))^2] + \mathbb{E}_0[\lambda_0^\ell(f_0)])\right) \tag{4.2}$$

and r_T *is defined in* (4.4). *Then*,

$$\mathbb{P}_0(L_T(f_0) - L_T(f) \ge (\kappa\log(r_T^{-1}) + 1)T\epsilon_T^2) \le \frac{C\log\log(T)\log^3(T)}{T\epsilon_T^2}, \tag{4.3}$$

for C *a constant only depending on* f_0 *and* B.

PROOF. We consider the set $\tilde{\Omega}_T$ defined in Lemma 2 in the Supplementary Material [19], and we set $\mathcal{N}_T = C_\alpha \log T$. We have

$$\begin{aligned}\mathrm{KL}(f^0, f) &= \sum_{k=1}^K \mathbb{E}_0\left[\int_0^T \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)dN_t^k - \int_0^T (\lambda_t^k(f_0) - \lambda_t^k(f))\,dt\right]\\ &= \sum_{k=1}^K \mathbb{E}_0\left[\int_0^T \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)\,dt - \int_0^T (\lambda_t^k(f_0) - \lambda_t^k(f))\,dt\right]\\ &= \sum_{k=1}^K \mathbb{E}_0\left[\int_0^T \Psi\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)\lambda_t^k(f_0)\,dt\right],\end{aligned}$$

where for $u > 0$, $\Psi(u) := -\log(u) - 1 + u \ge 0$. First, observe that on $\tilde{\Omega}_T \cap B(\epsilon_T, B)$,

$$\begin{aligned}\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)} &\ge \frac{\nu_k}{\nu_k^0 + \sum_{\ell=1}^K \int_{t-A}^{t-} h_{\ell,k}^0(t-u)\,dN^\ell(u)}\\ &\ge \frac{\min_k \nu_k^0 - \epsilon_T}{\max_k \nu_k^0 + \max_{\ell,k}\|h_{\ell,k}^0\|_\infty K\mathcal{N}_T} =: r_T.\end{aligned} \tag{4.4}$$

Furthermore, observe that for $u \in [r_T, 1/2)$, $\Psi(u) \le \log(r_T^{-1})$, since $r_T = o(1)$. And for all $u \ge 1/2$, $\Psi(u) \le (u-1)^2$. Finally, for any $u \ge r_T$,

$$\Psi(u) \le 4\log(r_T^{-1})(u-1)^2.$$

Therefore, on $B(\epsilon_T, B)$, we have

$$\begin{aligned}0 \le \mathrm{KL}(f^0, f) &\le 4\log(r_T^{-1}) \sum_{k=1}^K \mathbb{E}_0\left[\int_0^T \frac{(\lambda_t^k(f_0) - \lambda_t^k(f))^2}{\lambda_t^k(f_0)} \mathbb{1}_{\tilde{\Omega}_T}\, dt\right] + R_T \\ &\le 4\log(r_T^{-1}) \sum_{k=1}^K (\nu_k^0)^{-1} \mathbb{E}_0\left[\int_0^T (\lambda_t^k(f_0) - \lambda_t^k(f))^2\, dt\right] + R_T,\end{aligned}$$

where

$$R_T = \sum_{k=1}^K \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_0^T \left(-\log\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right) - 1 + \frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)\lambda_t^k(f_0)\, dt\right].$$

We first deal with the first term. Using stationarity of the process and Proposition 2 of [26],

$$\begin{aligned}&\mathbb{E}_0\left[\int_0^T (\lambda_t^k(f_0) - \lambda_t^k(f))^2\, dt\right] \\ &\quad\le 2T(\nu_k^0 - \nu_k)^2 + 2\int_0^T \mathbb{E}_0\left[\left(\sum_{\ell=1}^K \int_{t-A}^{t^-} (h_{\ell,k} - h_{\ell,k}^0)(t-u)\, dN^\ell(u)\right)^2\right] dt \\ &\quad\le 2T\epsilon_T^2 + 4K\int_0^T \mathbb{E}_0\left[\sum_{\ell=1}^K \left(\int_{t-A}^{t^-} (h_{\ell,k} - h_{\ell,k}^0)(t-u)\lambda_u^\ell(f_0)\, du\right)^2\right] dt \\ &\qquad + 4K\int_0^T \mathbb{E}_0\left[\sum_{\ell=1}^K \left(\int_{t-A}^{t^-} (h_{\ell,k} - h_{\ell,k}^0)(t-u)(dN_u^\ell - \lambda_u^\ell(f_0)\, du)\right)^2\right] dt \\ &\quad\le 2T\epsilon_T^2 + 4K\sum_{\ell=1}^K \|h_{\ell,k} - h_{\ell,k}^0\|_2^2 \int_0^T \int_{t-A}^{t^-} \mathbb{E}_0[(\lambda_u^\ell(f_0))^2]\, du\, dt \\ &\qquad + 4K\int_0^T \sum_{\ell=1}^K \int_{t-A}^{t^-} (h_{\ell,k} - h_{\ell,k}^0)^2(t-u)\mathbb{E}_0[\lambda_u^\ell(f_0)]\, du\, dt \\ &\quad\le 2T\epsilon_T^2 + 4KT\sum_{\ell=1}^K \|h_{\ell,k} - h_{\ell,k}^0\|_2^2 (A\mathbb{E}_0[(\lambda_0^\ell(f_0))^2] + \mathbb{E}_0[\lambda_0^\ell(f_0)]) \\ &\quad\le T\epsilon_T^2\left(2 + 4K\sum_{\ell=1}^K (A\mathbb{E}_0[(\lambda_0^\ell(f_0))^2] + \mathbb{E}_0[\lambda_0^\ell(f_0)])\right).\end{aligned}$$

We now deal with R_T. We have, on $B(\epsilon_T, B)$,

$$\tag{4.5}\begin{aligned}\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)} &\le (\nu_k^0)^{-1}\left(\nu_k + \sum_{\ell=1}^K \|h_{\ell,k}\|_\infty \sup_{t\in[0,T]} N^\ell([t-A, t))\right) \\ &\le (\nu_k^0)^{-1}\left(\nu_k^0 + \epsilon_T + B\sum_{\ell=1}^K \sup_{t\in[0,T]} N^\ell([t-A, t))\right).\end{aligned}$$

Conversely,

$$\tag{4.6}\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)} \ge (\nu_k^0 - \epsilon_T)\left(\nu_k^0 + \sum_{\ell=1}^K \|h_{\ell,k}^0\|_\infty \sup_{t\in[0,T]} N^\ell([t-A, t))\right)^{-1}.$$

So, using Lemma 2 in the Supplementary Material [19], if α is an absolute constant large enough, $R_T = o(1)$ and $R_T = o(T\epsilon_T^2)$. Choosing $\kappa = 4\sum_{k=1}^K (\nu_k^0)^{-1}(3 + 4K\sum_{\ell=1}^K (A \times \mathbb{E}_0[(\lambda_0^\ell(f_0))^2] + \mathbb{E}_0[\lambda_0^\ell(f_0)]))$ terminates the proof of (4.1). Note that if $B(\epsilon_T, B)$ is replaced with $B_\infty(\epsilon_T, B)$ (see Remark 1), then

$$\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)} \leq 1 + \frac{|\nu_k - \nu_k^0| + \sum_\ell \|h_{\ell,k} - h_{\ell,k}\|_\infty \mathcal{N}_T}{\nu_k^0}$$

and

$$\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)} \geq 1 - \frac{|\nu_k - \nu_k^0| + \sum_\ell \|h_{\ell,k} - h_{\ell,k}\|_\infty \mathcal{N}_T}{\nu_k^0}$$

so that we can take $r_T = 1/2$ and $R_T = o(T\epsilon_T^2)$.

We now study

$$\mathcal{L}_T := L_T(f_0) - L_T(f) - \mathbb{E}_0\big[L_T(f_0) - L_T(f)\big].$$

We have for any integer Q_T such that $x := T/(2Q_T) > A$,

$$\begin{aligned}
&L_T(f_0) - L_T(f)\\
&\quad= \sum_{k=1}^K \biggl(\int_0^T \log\biggl(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\biggr)\,dN_t^k - \int_0^T \bigl(\lambda_t^k(f_0) - \lambda_t^k(f)\bigr)\,dt\biggr)\\
&\quad= \sum_{q=0}^{Q_T-1} \int_{2qx}^{2qx+x} \sum_{k=1}^K \biggl(\log\biggl(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\biggr)\,dN_t^k - \bigl(\lambda_t^k(f_0) - \lambda_t^k(f)\bigr)\,dt\biggr)\\
&\qquad+ \sum_{q=0}^{Q_T-1} \int_{2qx+x}^{2qx+2x} \sum_{k=1}^K \biggl(\log\biggl(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\biggr)\,dN_t^k - \bigl(\lambda_t^k(f_0) - \lambda_t^k(f)\bigr)\,dt\biggr)\\
&\quad=: \sum_{q=0}^{Q_T-1} F_q + \sum_{q=0}^{Q_T-1} \tilde{F}_q.
\end{aligned}$$

Note that F_q is a measurable function of the points of N appearing in $[2qx - A; 2qx + x)$ denoted by $\mathcal{F}(N_{|[2qx-A;2qx+x)})$. Using Proposition 3.1 of [41], we consider an i.i.d. sequence $(M_q^x)_{q=0,\dots,Q_T-1}$ of Hawkes processes with the same distribution as N but restricted to $[2qx - A; 2qx + x)$ and such that for all q, the variation distance between M_q^x and $N_{|[2qx-A;2qx+x)}$ is less than $2\mathbb{P}_0(T_e > x - A)$, where T_e is the extinction time of the process. We then set for any q, $G_q = \mathcal{F}(M_q^x)$. We have built an i.i.d. sequence $(G_q)_{q=0,\dots,Q_T-1}$ with the same distributions as the F_q's. Furthermore, for any q,

$$\mathbb{P}_0(F_q \neq G_q) \leq 2\mathbb{P}_0(T_e > x - A).$$

We now have, by stationarity,

$$\begin{aligned}
\mathbb{P}_0(\mathcal{L}_T \geq T\epsilon_T^2) &= \mathbb{P}_0\bigl(L_T(f_0) - L_T(f) - \mathbb{E}_0\big[L_T(f_0) - L_T(f)\big] \geq T\epsilon_T^2\bigr)\\
&= \mathbb{P}_0\Biggl(\sum_{q=0}^{Q_T-1} (F_q - \mathbb{E}_0[F_q]) + \sum_{q=0}^{Q_T-1} (\tilde{F}_q - \mathbb{E}_0[\tilde{F}_q]) \geq T\epsilon_T^2\Biggr)\\
&\leq 2\mathbb{P}_0\Biggl(\sum_{q=0}^{Q_T-1} (F_q - \mathbb{E}_0[F_q]) \geq T\epsilon_T^2/2\Biggr)
\end{aligned}$$

$$\leq 2\mathbb{P}_0\left(\sum_{q=0}^{Q_T-1}(G_q - \mathbb{E}_0[G_q]) \geq T\epsilon_T^2/2\right) + 2\mathbb{P}_0(\exists q;\, F_q \neq G_q)$$

$$\leq 2\mathbb{P}_0\left(\sum_{q=0}^{Q_T-1}(G_q - \mathbb{E}_0[G_q]) \geq T\epsilon_T^2/2\right) + 4Q_T\mathbb{P}_0(T_e > x - A).$$

We first deal with the first term of the previous expression:

$$\begin{aligned}\mathbb{P}_0\left(\sum_{q=0}^{Q_T-1}(G_q - \mathbb{E}_0[G_q]) \geq T\epsilon_T^2/2\right) &\leq \frac{4}{T^2\epsilon_T^4}\operatorname{Var}_0\left(\sum_{q=0}^{Q_T-1} G_q\right)\\ &\leq \frac{4}{T^2\epsilon_T^4}\sum_{q=0}^{Q_T-1}\operatorname{Var}_0(G_q) \leq \frac{4Q_T}{T^2\epsilon_T^4}\operatorname{Var}_0(G_0)\\ &= \frac{4Q_T}{T^2\epsilon_T^4}\operatorname{Var}_0(F_0).\end{aligned}$$

Now, by setting $d\mathcal{M}_t^{(k)} = dN_t^k - \lambda_t^k(f_0)\,dt$,

$$\begin{aligned}\operatorname{Var}_0(F_0) &\leq \mathbb{E}_0[F_0^2]\\ &\leq \mathbb{E}_0\Bigg[\Bigg(\sum_{k=1}^K\int_0^{\frac{T}{2Q_T}}\log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)dN_t^k\\ &\qquad - \sum_{k=1}^K\int_0^{\frac{T}{2Q_T}}(\lambda_t^k(f_0) - \lambda_t^k(f))\,dt\Bigg)^2\Bigg]\\ &\lesssim \sum_{k=1}^K\mathbb{E}_0\Bigg[\Bigg(\int_0^{\frac{T}{2Q_T}}\Psi\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)\lambda_t^k(f_0)\,dt\\ &\qquad + \int_0^{\frac{T}{2Q_T}}\log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)d\mathcal{M}_t^{(k)}\Bigg)^2\Bigg]\\ &\lesssim \sum_{k=1}^K\mathbb{E}_0\left[\left(\int_0^{\frac{T}{2Q_T}}\Psi\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)\lambda_t^k(f_0)\,dt\right)^2\right]\\ &\qquad + \mathbb{E}_0\left[\left(\int_0^{\frac{T}{2Q_T}}\log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)d\mathcal{M}_t^{(k)}\right)^2\right]\\ &\lesssim \sum_{k=1}^K\frac{T}{Q_T}\mathbb{E}_0\left[\int_0^{\frac{T}{2Q_T}}\Psi^2\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)(\lambda_t^k(f_0))^2\,dt\right]\\ &\qquad + \mathbb{E}_0\left[\int_0^{\frac{T}{2Q_T}}\log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)\,dt\right].\end{aligned}$$

Note that on $\tilde{\Omega}_T$, for any $t \in [0; T/(2Q_T)]$,

$$0 \leq \Psi\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)\lambda_t^k(f_0) \leq C_1(B, f_0)\mathcal{N}_T^2,$$

where $C_1(B, f_0)$ only depends on B and f_0. Then,

$$\begin{aligned}\mathbb{E}_0\biggl[\mathbb{1}_{\tilde{\Omega}_T}\int_0^{\frac{T}{2Q_T}}\Psi^2\biggl(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\biggr)\bigl(\lambda_t^k(f_0)\bigr)^2\,dt\biggr]&\\ \le C_1(B, f_0)\mathcal{N}_T^2\times\mathbb{E}_0\biggl[\mathbb{1}_{\tilde{\Omega}_T}\int_0^{\frac{T}{2Q_T}}\Psi\biggl(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\biggr)\lambda_t^k(f_0)\,dt\biggr]&\end{aligned}$$

and using the same arguments as for the bound of $\mathrm{KL}(f^0, f)$, the previous term is bounded by $\log(r_T^{-1})\mathcal{N}_T^2 \times (T/Q_T)\epsilon_T^2$ up to a constant. Since for any $u \ge 1/2$, we have $|\log(u)| \le 2|u-1|$; we have for any $u \ge r_T$,

$$\bigl|\log(u)\bigr| \le 2\log\bigl(r_T^{-1}\bigr)|u-1|$$

and

$$\begin{aligned}\mathbb{E}_0\biggl[\mathbb{1}_{\tilde{\Omega}_T}\int_0^{\frac{T}{2Q_T}}\log^2\biggl(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\biggr)\lambda_t^k(f_0)\,dt\biggr]&\\ \le 4\log^2\bigl(r_T^{-1}\bigr)\bigl(\nu_k^0\bigr)^{-1}\mathbb{E}_0\biggl[\mathbb{1}_{\tilde{\Omega}_T}\int_0^{\frac{T}{2Q_T}}\bigl(\lambda_t^k(f_0)-\lambda_t^k(f)\bigr)^2\,dt\biggr]&\\ \lesssim \log^2\bigl(r_T^{-1}\bigr)(T/Q_T)\epsilon_T^2.&\end{aligned}$$

By taking $\alpha \ge 2$ and using Lemma 2 in the Supplementary Material [19], we obtain

$$\begin{aligned}\mathbb{E}_0\biggl[\mathbb{1}_{\tilde{\Omega}_T^c}\int_0^{\frac{T}{2Q_T}}\Psi^2\biggl(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\biggr)\bigl(\lambda_t^k(f_0)\bigr)^2\,dt\biggr]&\\ +\mathbb{E}_0\biggl[\mathbb{1}_{\tilde{\Omega}_T^c}\int_0^{\frac{T}{2Q_T}}\log^2\biggl(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\biggr)\lambda_t^k(f_0)\,dt\biggr]&=o\bigl(TQ_T^{-1}\epsilon_T^2\bigr).\end{aligned}$$

Finally,

$$\mathrm{Var}_0(F_0) \le C_2(B, f_0)\log\bigl(r_T^{-1}\bigr)\mathcal{N}_T^2 \times (T/Q_T)^2\epsilon_T^2,$$

for $C_2(B, f_0)$ a constant only depending on B and f_0, and

$$\begin{aligned}\mathbb{P}_0\bigl(\mathcal{L}_T \ge T\epsilon_T^2\bigr) \le 8C_2(B, f_0)\log\bigl(r_T^{-1}\bigr)\mathcal{N}_T^2 \times (T/Q_T) \times \bigl(1/\bigl(T\epsilon_T^2\bigr)&\\ + 4Q_T\mathbb{P}_0(T_e > x - A).&\end{aligned}$$

It remains to deal with the last term of the previous expression. The proof of Proposition 3 of [26] shows that there exists a constant D only depending on f_0 such that if we take $x = D\log T$, which is larger than A for T large enough, then $4Q_T\mathbb{P}_0(T_e > x - A) = o(T^{-1})$. We now have $\log(r_T^{-1})\mathcal{N}_T^2 \times (T/Q_T) = O(\log\log(T)\log^3(T))$, which ends the proof of the lemma. □

4.3. *Proof of Theorem* 3. Define

$$A_{L_1}(w_T\varepsilon_T) = \bigl\{f \in \mathcal{F};\ \|f - f_0\|_1 \le w_T\varepsilon_T\bigr\},$$

then $\Pi(A_{L_1}(w_T\varepsilon_T)^c|N) \le \Pi(A_{\varepsilon_T}^c|N) + \Pi(A_{L_1}(w_T\varepsilon_T)^c \cap A_{\varepsilon_T}|N)$. Using Assumption (i), we just need to prove that

$$\mathbb{E}_0\bigl[\mathbb{1}_{\Omega_{1,T}}\Pi\bigl(A_{L_1}(w_T\varepsilon_T)^c \cap A_{\varepsilon_T}|N\bigr)\bigr] = o(1) \tag{4.7}$$

for some well chosen set $\Omega_{1,T} \subset \Omega_T$ such that

$$\mathbb{P}_0\bigl(\Omega_{1,T}^c \cap \Omega_T\bigr) = o(1). \tag{4.8}$$

Using (2.2) in the Supplementary Material [19], there exists C_0 such that for all $f \in A_{\varepsilon_T}$, on Ω_T, $\sum_\ell \nu_\ell + \sum_{\ell,k} \rho_{\ell,k} \le C_0$. Therefore, on Ω_T,

$$A_{L_1}(w_T\varepsilon_T)^c \cap A_{\varepsilon_T} \subset \left\{ f \in \mathcal{F}; \|f - f_0\|_1 > w_T\varepsilon_T, \sum_\ell \left(\nu_\ell + \sum_k \rho_{\ell,k} \right) \le C_0 \right\}.$$

We set $u_T := u_0(\log T)^{1/6}\varepsilon_T^{1/3}$ with u_0 a large constant to be chosen later. Let $\mathcal{F}_T = \{f \in \mathcal{F}; \|\rho\| \le 1 - u_T\}$. From Assumption (ii),

$$\Pi(\mathcal{F}_T^c) \le e^{-2c_1 T\varepsilon_T^2}$$

for T large enough. Following the same lines as in the proof of Theorem 1, we then have

$$\begin{aligned} &\mathbb{E}_0[\mathbb{1}_{\Omega_{1,T}} \Pi(A_{L_1}(w_T\varepsilon_T)^c \cap A_{\varepsilon_T}|N)] \\ &\quad \le \mathbb{P}_0(D_T < e^{-c_1 T\varepsilon_T^2}) \\ &\qquad + e^{c_1 T\varepsilon_T^2} \int_{A_{L_1}(w_T\varepsilon_T)^c \cap \mathcal{F}_T} \mathbb{E}_0[\mathbb{P}_f(\Omega_{1,T} \cap \{d_{1,T}(f, f_0) \le \varepsilon_T\}|\mathcal{G}_0)]\, d\Pi(f) \\ &\qquad + e^{-c_1 T\varepsilon_T^2}, \end{aligned} \tag{4.9}$$

where $\mathbb{P}_f$ denotes the stationary distribution when the true parameter is f. We will now prove that

$$\sup_{f \in A_{L_1}(w_T\varepsilon_T)^c \cap \mathcal{F}_T} \mathbb{P}_f(\Omega_{1,T} \cap \{d_{1,T}(f, f_0) \le \varepsilon_T\}|\mathcal{G}_0) = o(e^{-c_1 T\varepsilon_T^2}).$$

Let $Z_{m,\ell}$ be defined by

$$Z_{m,\ell} = \int_{2mT/(2J_T)}^{(2m+1)T/(2J_T)} \left| \nu_\ell - \nu_\ell^0 + \sum_{k=1}^K \int_{t-A}^{t^-} (h_{k,\ell} - h_{k,\ell}^0)(t - s)\, dN_s^k \right| dt$$

with J_T such that $J_T = \lfloor \kappa_0 (\log T)^{-1} T u_T^2 \rfloor$ and κ_0 a constant chosen later. Note that $J_T \to +\infty$ and $T/J_T \to +\infty$ when $T \to +\infty$. Since $T d_{1,T}(f, f_0) \ge \max_{1\le\ell\le K} \sum_{m=1}^{J_T-1} Z_{m,\ell}$, we have that

$$\begin{aligned} &\mathbb{P}_f(\Omega_{1,T} \cap \{d_{1,T}(f, f_0) \le \varepsilon_T\}|\mathcal{G}_0) \\ &\quad \le \min_{1\le\ell\le K} \mathbb{P}_f\left(\Omega_{1,T} \cap \left\{ \sum_{m=1}^{J_T-1} Z_{m,\ell} \le \varepsilon_T T \right\} \Big| \mathcal{G}_0 \right) \\ &\quad \le \min_{1\le\ell\le K} \mathbb{P}_f\Bigg(\Omega_{1,T} \cap \Bigg\{ \sum_{m=1}^{J_T-1} (Z_{m,\ell} - \mathbb{E}_f[Z_{m,\ell}]) \\ &\qquad \le \varepsilon_T T - (J_T - 1)\mathbb{E}_f[Z_{1,\ell}] \Bigg\} \Big| \mathcal{G}_0 \Bigg). \end{aligned}$$

From Lemma 4 in the Supplementary Material [19], we have that there exists ℓ (depending on f and f^0) such that $\mathbb{E}_f[Z_{1,\ell}] \ge CT\|f - f_0\|_1/J_T$ for some $C > 0$ so that if $f \in A_{L_1}(w_T\varepsilon_T)^c$, then, since $w_T \to +\infty$,

$$\begin{aligned} &\mathbb{P}_f(\Omega_{1,T} \cap \{d_{1,T}(f, f_0) \le \varepsilon_T\}|\mathcal{G}_0) \\ &\quad \le \max_\ell \mathbb{P}_f\left(\Omega_{1,T} \cap \left\{ \sum_{m=1}^{J_T-1} [Z_{m,\ell} - \mathbb{E}_f[Z_{m,\ell}]] \le -\frac{CT\|f - f_0\|_1}{2} \right\} \Big| \mathcal{G}_0 \right). \end{aligned}$$

The problem in dealing with the right-hand side of the above inequality is that the $Z_{m,\ell}$'s are not independent. We therefore show that we can construct independent random variables $\tilde{Z}_{m,\ell}$ such that, conditionally on $\mathcal{G}_0$, $\sum_{m=1}^{J_T-1}(Z_{m,\ell} - \mathbb{E}_f[Z_{m,\ell}])$ is close to $\sum_{m=1}^{J_T-1}(\tilde{Z}_{m,\ell} - \mathbb{E}_f[\tilde{Z}_{m,\ell}])$ on $\Omega_{1,T}$. For all $1 \le m \le J_T - 1$, define $N^{0,m}$ the subcounting measure of N generated from the ancestors of any type born on $[(2m-1)T/(2J_T), (2m+1)T/(2J_T)]$ and the K-multivariate point process $\bar{N}^m$ defined by

$$\bar{N}^m = N - N^{0,m}.$$

Denote

$$\tilde{Z}_{m,\ell} = \int_{2mT/(2J_T)}^{(2m+1)T/(2J_T)} \left| \nu_\ell - \nu_\ell^0 + \sum_{k=1}^K \int_{t-A}^{t^-} (h_{k,\ell} - h_{k,\ell}^0)(t-s)\, dN_s^{0,m,k} \right| dt,$$

where $N^{0,m,k}$ if the kth coordinate of $N^{0,m}$. Observe that if $I_m = [2mT/(2J_T) - A, (2m+1)T/(2J_T)]$, then $\bar{N}^m(I_m)$ is the number of points of $\bar{N}^m$ lying in I_m. We have

$$\begin{aligned}
&|Z_{m,\ell} - \tilde{Z}_{m,\ell}| \\
&\quad = \left| \int_{2mT/(2J_T)}^{(2m+1)T/(2J_T)} \left(\left| \nu_\ell - \nu_\ell^0 + \sum_{k=1}^K \int_{t-A}^{t^-} (h_{k,\ell} - h_{k,\ell}^0)(t-s)\, dN_s^k \right| \right.\right. \\
&\qquad \left.\left. - \left| \nu_\ell - \nu_\ell^0 + \sum_{k=1}^K \int_{t-A}^{t^-} (h_{k,\ell} - h_{k,\ell}^0)(t-s)\, dN_s^{0,m,k} \right| \right) dt \right| \qquad (4.10) \\
&\quad \le \mathbb{1}_{\bar{N}^m(I_m) \ne 0} \sum_{k=1}^K \int_{2mT/(2J_T)}^{(2m+1)T/(2J_T)} \int_{t-A}^{t^-} |(h_{k,\ell} - h_{k,\ell}^0)(t-s)|\, d\bar{N}_s^{m,k}\, dt \\
&\quad \le \mathbb{1}_{\bar{N}^m(I_m) \ne 0} \sum_{k=1}^K \|h_{k,\ell} - h_{k,\ell}^0\|_1 \bar{N}^{m,k}(I_m) \le \|f - f_0\|_1 \bar{N}^m(I_m).
\end{aligned}$$

Let $\Omega_{1,T} = \Omega_T \cap \{\sum_{m=1}^{J_T-1} \bar{N}^m(I_m) \le CT/8\}$. In Lemma 6 in the Supplementary Material [19], we prove that there exists $\tilde{c}_0$ such that $\mathbb{P}_0(\Omega_{1,T}^c \cap \Omega_T) \le e^{-C\tilde{c}_0 T}$, and (4.8) is satisfied. Using (4.10), we have on $\Omega_{1,T}$

$$|Z_{m,\ell} - \tilde{Z}_{m,\ell}| \le \|f - f_0\|_1 CT/8. \qquad (4.11)$$

Lemma 6 in the Supplementary Material [19] proves that there exists a constant $\kappa_0 > 0$ (see the definition of J_T) such that $\sum_{m=1}^{J_T-1} \mathbb{E}_f[\bar{N}^m(I_m)] \le CT/8$, so that

$$\begin{aligned}
\sum_{m=1}^{J_T-1} \left| \mathbb{E}_f[Z_{m,\ell}] - \mathbb{E}_f[\tilde{Z}_{m,\ell}] \right| &\le \sum_{m=1}^{J_T-1} \mathbb{E}_f |Z_{m,\ell} - \tilde{Z}_{m,\ell}| \\
&\le \|f - f_0\|_1 \sum_{m=1}^{J_T-1} \mathbb{E}_f[\bar{N}^m(I_m)] \le C\|f - f_0\|_1 T/8
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_f\big(\Omega_{1,T} \cap \{d_{1,T}(f, f_0) \le \varepsilon_T\} | \mathcal{G}_0\big) \\
&\quad \le \max_\ell \mathbb{P}_f\left(\Omega_{1,T} \cap \left\{ \sum_{m=1}^{J_T-1} [Z_{m,\ell} - \mathbb{E}_f[Z_{m,\ell}]] \le -\frac{CT\|f - f_0\|_1}{2} \right\} \middle| \mathcal{G}_0 \right) \\
&\quad \le \mathbb{P}_f\left(\sum_{m=1}^{J_T-1} (-\tilde{Z}_{m,\ell} + \mathbb{E}_f(\tilde{Z}_{m,\ell})) \ge CT\|f - f_0\|_1/4 \middle| \mathcal{G}_0 \right).
\end{aligned}$$

Since by construction the $\tilde{Z}_{m,\ell}$ are positive, independent, identically distributed and independent of $\mathcal{G}_0$, the Bernstein inequality gives

$$\mathbb{P}_f\left(\sum_{m=1}^{J_T-1}(-\tilde{Z}_{m,\ell}+\mathbb{E}_f(\tilde{Z}_{m,\ell}))\geq CT\|f-f_0\|_1/4\Big|\mathcal{G}_0\right)\leq e^{-\frac{C^2T^2\|f-f_0\|_1^2}{32(J_T-1)\mathbb{E}_f(\tilde{Z}_{1,\ell}^2)}}.$$

We have to bound $\mathbb{E}_f(\tilde{Z}_{1,\ell}^2)$. Observe that

$$\begin{aligned}\tilde{Z}_{m,\ell}&\leq\int_{2mT/(2J_T)}^{(2m+1)T/(2J_T)}|\nu_\ell-\nu_\ell^0|\,dt\\&\quad+\int_{2mT/(2J_T)}^{(2m+1)T/(2J_T)}\sum_{k=1}^K\int_{t-A}^{t^-}|(h_{k,\ell}-h_{k,\ell}^0)(t-s)|\,dN_s^{0,m,k}\,dt\\&\leq\frac{T}{2J_T}|\nu_\ell-\nu_\ell^0|+\sum_{k=1}^K\|h_{k,\ell}-h_{k,\ell}^0\|_1N^{0,m,k}(I_m)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_f[\tilde{Z}_{1,\ell}^2]&\leq\frac{T^2}{2J_T^2}|\nu_\ell-\nu_\ell^0|^2+2K\sum_{k=1}^K\|h_{k,\ell}-h_{k,\ell}^0\|_1^2\mathbb{E}_f[N^{0,1,k}(I_1)^2]\\&\leq\frac{T^2}{J_T^2}\|f-f_0\|_1^2\left(\frac{1}{2}+\frac{2K\max_k\mathbb{E}_f[N^{0,1,k}(I_1)^2]J_T^2}{T^2}\right).\end{aligned}$$

We then have to bound $T^{-2}J_T^2\max_k\mathbb{E}_f[N^{0,1,k}(I_1)^2]$. Using notation of Lemma 6 in the Supplementary Material [19], we have

$$\begin{aligned}\mathbb{E}_f[N^{0,1,k}(I_1)^2]&\leq\mathbb{E}_f\left[\left(\sum_{\ell=1}^K\sum_{T/(2J_T)\leq p\leq 3T/(2J_T)}\sum_{k=1}^{B_{p,\ell}}W_{k,p}^\ell\right)^2\right]\\&\leq\frac{KT}{J_T}\sum_{\ell=1}^K\sum_{T/(2J_T)\leq p\leq 3T/(2J_T)}\mathbb{E}_f\left[\left(\sum_{k=1}^{B_{p,\ell}}W_{k,p}^\ell\right)^2\right]\\&\leq\frac{KT}{J_T}\sum_{\ell=1}^K\sum_{T/(2J_T)\leq p\leq 3T/(2J_T)}\mathbb{E}_f\left[\mathbb{E}_f\left[\left(\sum_{k=1}^{B_{p,\ell}}W_{k,p}^\ell\right)^2\Big|B_{p,\ell}\right]\right]\\&\leq\frac{KT^2}{J_T^2}\sum_{\ell=1}^K(\nu_\ell^2+\nu_\ell)\mathbb{E}_f[(W^\ell)^2].\end{aligned}$$

We now bound $\mathbb{E}_f[(W^\ell)^2]$ by using Lemma 5 in the Supplementary Material [19]. Without loss of generality, we can assume that $\|\rho\|>1/2$. We take $t=\frac{1-\|\rho\|}{2\sqrt{K}}\log(\frac{1+\|\rho\|}{2\|\rho\|})$ and

$$\mathbb{E}_f[(W^\ell)^2]\leq 2t^{-2}\mathbb{E}_f[\exp(tW^\ell)]\lesssim t^{-2}\lesssim(1-\|\rho\|)^{-4}$$

and

$$T^{-2}J_T^2\max_k\mathbb{E}_f[N^{0,1,k}(I_1)^2]\lesssim(1-\|\rho\|)^{-4}.$$

Therefore, since $f \in \mathcal{F}_T$, there exists a constant C'_K only depending on K such that

$$\begin{aligned}
&\mathbb{P}_f\left(\sum_{m=1}^{J_T-1}(-\tilde{Z}_{m,\ell}+\mathbb{E}_f(\tilde{Z}_{m,\ell})) \geq CT\|f-f_0\|_1/4 \middle| \mathcal{G}_0\right) \\
&\quad \leq e^{-C'_K J_T(1-\|\rho\|)^4} \\
&\quad \leq e^{-C'_K J_T u_T^4} \lesssim e^{-C'_K \kappa_0 (\log T)^{-1} T u_T^6} \lesssim e^{-C'_K \kappa_0 u_0^6 T \epsilon_T^2},
\end{aligned}$$

where the last inequality follows from the definition of u_T and J_T. We obtain the desired bound as soon as u_0 is large enough, namely,

$$\sup_{f \in A_{L_1}(w_T \varepsilon_T)^c \cap \mathcal{F}_T} \mathbb{P}_f\big(\Omega_{1,T} \cap \{d_{1,T}(f, f_0) \leq \varepsilon_T\} | \mathcal{G}_0\big) = o\big(e^{-c_1 T \varepsilon_T^2}\big).$$

Using (4.9) and Assumption (i), we then have that (4.7) is true which proves the theorem.

SUPPLEMENTARY MATERIAL

Supplement to "Nonparametric Bayesian estimation for multivariate Hawkes processes" (DOI: 10.1214/19-AOS1903SUPP; .pdf). The supplement material contains additional numerical results and the proofs of all theoretical results needed previously.

REFERENCES

[1] AÏT-SAHALIA, Y., CACHO-DIAZ, J. and LAEVEN, R. J. (2015). Modeling financial contagion using mutually exciting jump processes. *J. Financ. Econ.* **117** 585–606.

[2] ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes. Springer Series in Statistics.* Springer, New York. MR1198884 https://doi.org/10.1007/978-1-4612-4348-9

[3] BACRY, E., DAYRI, K. and MUZY, J. F. (2012). Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *Eur. Phys. J. B* **85** 157.

[4] BACRY, E., DELATTRE, S., HOFFMANN, M. and MUZY, J. F. (2013). Modelling microstructure noise with mutually exciting point processes. *Quant. Finance* **13** 65–77. MR3005350 https://doi.org/10.1080/14697688.2011.647054

[5] BACRY, E., GAÏFFAS, S. and MUZY, J.-F. (2015). A generalization error bound for sparse and low-rank multivariate Hawkes processes. ArXiv e-prints.

[6] BACRY, E., JAISSON, T. and MUZY, J. (2016). Estimation of slowly decreasing Hawkes kernels: Application to high-frequency order book dynamics. *Quant. Finance* **16** 1179–1201. MR3520321 https://doi.org/10.1080/14697688.2015.1123287

[7] BACRY, E., MASTROMATTEO, I. and MUZY, J.-F. (2015). Hawkes processes in finance. *Mark. Microstruct. Liq.* **1** 1550005.

[8] BACRY, E. and MUZY, J.-F. (2016). First- and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Trans. Inform. Theory* **62** 2184–2202. MR3480107 https://doi.org/10.1109/TIT.2016.2533397

[9] BLUNDELL, C., BECK, J. and HELLER, K. A. (2012). Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) **25** 2600–2608. Curran Associates, Red Hook.

[10] BRÉMAUD, P. and MASSOULIÉ, L. (1996). Stability of nonlinear Hawkes processes. *Ann. Probab.* **24** 1563–1588. MR1411506 https://doi.org/10.1214/aop/1065725193

[11] BRILLINGER, D. R. (1988). Maximum likelihood analysis of spike trains of interacting nerve cells. *Biol. Cybernet.* **59** 189–200. https://doi.org/10.1007/bf00318010

[12] CARSTENSEN, L., SANDELIN, A., WINTHER, O. and HANSEN, N. R. (2010). Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC Bioinform.* **11** 456. https://doi.org/10.1186/1471-2105-11-456

[13] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

[14] CHEN, S., SHOJAIE, A., SHEA-BROWN, E. and WITTEN, D. (2017). The multivariate Hawkes process in high dimensions: Beyond mutual excitation. ArXiv e-prints.

[15] CHEN, S., WITTEN, D. and SHOJAIE, A. (2017). Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. *Electron. J. Stat.* **11** 1207–1234. MR3634334 https://doi.org/10.1214/17-EJS1251

[16] CHORNOBOY, E. S., SCHRAMM, L. P. and KARR, A. F. (1988). Maximum likelihood identification of neural point process systems. *Biol. Cybernet.* **59** 265–275. MR0961117 https://doi.org/10.1007/BF00332915

[17] CRANE, R. and SORNETTE, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. USA* **105** 15649–15653.

[18] DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431

[19] DONNET, S., RIVOIRARD, V. and ROUSSEAU, J. (2020). Supplement to "Nonparametric Bayesian estimation for multivariate Hawkes processes." https://doi.org/10.1214/19-AOS1903SUPP.

[20] EMBRECHTS, P., LINIGER, T. and LIN, L. (2011). Multivariate Hawkes processes: An application to financial data. *J. Appl. Probab.* **48A** 367–378. MR2865638 https://doi.org/10.1239/jap/1318940477

[21] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[22] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274 https://doi.org/10.1214/009053606000001172

[23] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864 https://doi.org/10.1214/009053606000001271

[24] GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810 https://doi.org/10.1093/biomet/82.4.711

[25] GUSTO, G. and SCHBATH, S. (2005). FADO: A statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 24, 28. MR2170440 https://doi.org/10.2202/1544-6115.1119

[26] HANSEN, N. R., REYNAUD-BOURET, P. and RIVOIRARD, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* **21** 83–143. MR3322314 https://doi.org/10.3150/13-BEJ562

[27] HAWKES, A. G. (1971). Point spectra of some mutually exciting point processes. *J. Roy. Statist. Soc. Ser. B* **33** 438–443. MR0358976

[28] HAWKES, A. G. and OAKES, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Probab.* **11** 493–503. MR0378093 https://doi.org/10.2307/3212693

[29] LAMBERT, R. C., TULEAU-MALOT, C., BESSAIH, T., RIVOIRARD, V., BOURET, Y., LERESCHE, N. and REYNAUD-BOURET, P. (2018). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *J. Neurosci. Methods* **297** 9–21. https://doi.org/10.1016/j.jneumeth.2017.12.026

[30] LI, L. and ZHA, H. (2014). Learning parametric models for social infectivity in multi-dimensional Hawkes processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, *AAAI'*14 101–107. AAAI Press, Menlo Park.

[31] MITCHELL, L. and CATES, M. E. (2010). Hawkes process as a model of social interactions: A view on video dynamics. *J. Phys. A* **43** 045101, 11. MR2578723 https://doi.org/10.1088/1751-8113/43/4/045101

[32] 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. Amer. Statist. Assoc.* **106** 100–108. MR2816705 https://doi.org/10.1198/jasa.2011.ap09546

[33] OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

[34] OKATAN, M., WILSON, M. A. and BROWN, E. N. (2005). Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Comput.* **17** 1927–1961.

[35] PANINSKI, L., PILLOW, J. and LEWI, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. *Prog. Brain Res.* **165** 493–507. https://doi.org/10.1016/S0079-6123(06)65031-0

[36] PILLOW, J. W., SHLENS, J., PANINSKI, L., SHER, A., LITKE, A. M., CHICHILNISKY, E. and SIMONCELLI, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454** 995–999.

[37] PORTER, M. D. and WHITE, G. (2012). Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* **6** 106–124. MR2951531 https://doi.org/10.1214/11-AOAS513

[38] RASMUSSEN, J. G. (2013). Bayesian inference for Hawkes processes. *Methodol. Comput. Appl. Probab.* **15** 623–642. MR3085883 https://doi.org/10.1007/s11009-011-9272-5

[39] REYNAUD-BOURET, P., RIVOIRARD, V., GRAMMONT, F. and TULEAU-MALOT, C. (2014). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *J. Math. Neurosci.* **4** Art. 3, 41. MR3197017 https://doi.org/10.1186/2190-8567-4-3

[40] REYNAUD-BOURET, P., RIVOIRARD, V. and TULEAU-MALOT, C. (2013). Inference of functional connectivity in neurosciences via Hawkes processes. In *Global Conference on Signal and Information Processing* (*GlobalSIP*), 2013 *IEEE* 317–320. IEEE.

[41] REYNAUD-BOURET, P. and ROY, E. (2006). Some non asymptotic tail estimates for Hawkes processes. *Bull. Belg. Math. Soc. Simon Stevin* **13** 883–896. MR2293215

[42] REYNAUD-BOURET, P. and SCHBATH, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *Ann. Statist.* **38** 2781–2822. MR2722456 https://doi.org/10.1214/10-AOS806

[43] ROUSSEAU, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.* **38** 146–180. MR2589319 https://doi.org/10.1214/09-AOS703

[44] SIMMA, A. and JORDAN, M. I. (2012). Modeling events with cascades of Poisson processes. ArXiv e-prints.

[45] VERE-JONES, D. and OZAKI, T. (1982). Some examples of statistical estimation applied to earthquake data I: Cyclic Poisson and self-exciting models. *Ann. Inst. Statist. Math.* **34** 189–207.

[46] YANG, S.-H. and ZHA, H. (2013). Mixture of mutually exciting processes for viral diffusion. *ICML* (2) **28** 1–9.

[47] ZHOU, K., ZHA, H. and SONG, L. (2013). Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the* 30*th International Conference on Machine Learning* (*ICML*-13) (S. Dasgupta and D. Mcallester, eds.). *JMLR Workshop and Conference Proceedings* **28** 1301–1309.

[48] ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2002). Stochastic declustering of space–time earthquake occurrences. *J. Amer. Statist. Assoc.* **97** 369–380. MR1941459 https://doi.org/10.1198/016214502760046925