

# Lasso and probabilistic inequalities for multivariate point processes

Niels Richard Hansen\*, Patricia Reynaud-Bouret† and Vincent Rivoirard‡

August 3, 2012

**Abstract:** Due to its low computational cost, Lasso is an attractive regularization method for high-dimensional statistical settings. In this paper, we consider multivariate counting processes depending on an unknown function to be estimated by linear combinations of a fixed dictionary. To select coefficients, we propose an adaptive  $\ell_1$ -penalization methodology, where data-driven weights of the penalty are derived from new Bernstein type inequalities for martingales. Oracle inequalities are established under assumptions on the Gram matrix of the dictionary. Non-asymptotic probabilistic results for multivariate Hawkes processes are proven, which allows us to check these assumptions by considering general dictionaries based on histograms, Fourier or wavelet bases. Motivated by problems of neuronal activities inference, we finally lead a simulation study for multivariate Hawkes processes and compare our methodology with the *adaptive Lasso procedure* proposed by Zou in [57]. We observe an excellent behavior of our procedure with respect to the problem of supports recovery. We rely on theoretical aspects for the essential question of tuning our methodology. Unlike adaptive Lasso of [57], our tuning procedure is proven to be robust with respect to all the parameters of the problem, revealing its potential for concrete purposes, in particular in neuroscience.

**Mathematics Subject Classification:** 62G05, 62M09, 60G55, 60E15.

**Keywords:** Multivariate counting process, Hawkes processes, adaptive estimation, Lasso procedure, Bernstein-type inequalities.

## 1 Introduction

The Lasso, proposed by [51], is a well established method that achieves sparsity of an estimated parameter vector via  $\ell_1$ -penalization. In this paper, we focus on using Lasso to select and estimate coefficients in the basis expansion of intensity processes for multivariate point processes.

Recent examples of applications of multivariate point processes include the modeling of multivariate neuron spike data, [41], [38], stochastic kinetic modeling, [6], and the modeling of the distribution of ChIP-seq data along the genome [19]. In the previous examples the intensity of a future occurrence of a point depends on the history of all or some of the coordinates of the point processes, and it is of particular interest to estimate this dependence. This can be achieved using a parametric family of models, as in several of the papers above. Our aim is to provide a non-parametric method based on the Lasso.

The statistical properties of Lasso are particularly well understood in the context of regression with i.i.d. errors or for density estimation where a range of *oracle inequalities* have been established. These inequalities,

---

\*Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark.  
*email:* Niels.R.Hansen@math.ku.dk

†CNRS, Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice cedex 02, France.  
*email:* Patricia.Reynaud-Bouret@unice.fr

‡CEREMADE, CNRS-UMR 7534, Université Paris Dauphine, Place Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France. INRIA Paris-Rocquencourt, projet Classic. *email:* Vincent.Rivoirard@dauphine.fr – Corresponding author

now widespread in the literature, provide theoretical error bounds that hold on events with a controllable (large) probability. See for instance [4, 5, 12, 13, 14, 15, 53]. We refer the reader to [11] for an excellent account on many state-of-the-art results. One main challenge in this context is to obtain as weak conditions as possible on the design – or Gram – matrix. The other important challenge is to be able to provide an  $\ell_1$ -penalization procedure that provides excellent performances from both theoretical and practical points of view. If standard Lasso proposed by [51] based on deterministic constant weights constitute a major contribution from the methodological point of view, underestimation due to its shrinkage nature may lead to poor practical performances in some contexts. Different two steps procedures have been suggested to overcome this drawback (see [37, 55, 57]). Zou in [57] also discusses problems for standard Lasso to cope with variable selection and consistency simultaneously. He overcomes these problems by introducing non-constant data-driven  $\ell_1$ -weights based on preliminary consistent estimates.

In this paper we consider an  $\ell_1$ -penalized least squares criterion for the estimation of coefficients in the expansion of a function parameter. As in [4, 31, 55, 57], we consider non-constant data-driven weights. However the setup is here that of multivariate point processes and the function parameter that lives in a Hilbert space determines the point process intensities. Even in this unusual context, the least squares criterion also involves a random Gram matrix as well, and in this respect, we present a fairly standard oracle inequality with a strong condition on this Gram matrix. Major contributions of this article is to provide probabilistic results that enable us to calibrate  $\ell_1$ -weights on the one hand and to deal with the assumption on the Gram matrix on the other hand.

## 1.1 Our probabilistic contribution

In an i.i.d. framework (see for instance [4]) classical concentration inequalities can be used to have access to the  $\ell_1$ -weights. In the counting processes framework, the data-driven calibrated form of these  $\ell_1$ -weights is naturally linked to sharp Bernstein type inequalities for martingales. In the literature, those kinds of inequalities generally provide an upper bound for the martingale that is deterministic and unobservable [50, 52]. More recently, there have been some attempts to use self-normalized processes in order to provide a more flexible and random upper bound [3, 22, 24, 23]. Nevertheless, those bounds are usually not (completely) observable when dealing with counting processes. We prove a result that goes further in this direction by providing a completely sharp random observable upper bound for the martingale in our counting process framework.

In another direction, we do not want to make assumptions that cannot be checked on the Gram matrix which is, in our case, generated by the process itself. When no i.i.d. structure underlies the process, this control may become very difficult to handle. We fully treat the multivariate Hawkes process as a main example of this case. Even if Hawkes processes have been largely studied in the literature (see [8, 21] for instance), very few is known about exponential inequalities and non asymptotic tail controls. In particular, up to our knowledge, no exponential inequality controlling the number of points per interval is known, except in [45] for the univariate case. We extend this type of results and other sharp controls of the convergence in the ergodic theorem to obtain a sharp control on the Gram matrix.

Before going further, let us specify our framework and detail some specific examples.

## 1.2 $M$ -dimensional counting process

We consider an  $M$ -dimensional counting process  $(N_t^{(m)})_{m=1,\dots,M}$ , which can also be seen as a random point measure on  $\mathbb{R}$  with marks in  $\{1, \dots, M\}$ , and a corresponding *predictable* intensity processes  $(\lambda_t^{(m)})_{m=1,\dots,M}$  under a probability measure  $\mathbb{P}$ . We will assume that each intensity  $\lambda_t^{(m)}$  can be written as a linear predictable transformation of a deterministic function parameter  $f^*$  in a Hilbert space  $\mathcal{H}$ . We denote this linear transformation by  $\psi(f) = (\psi^{(1)}(f), \dots, \psi^{(M)}(f))$ . Therefore, for any  $t$ ,

$$\lambda_t^{(m)} = \psi_t^{(m)}(f^*).$$

The goal is to estimate  $f^*$  based on observations of  $(N_t^{(m)})_{m=1,\dots,M}$  for  $t \in [0, T]$ . Given a dictionary of functions denoted  $\Phi$ , candidates for estimating  $f^*$  are linear combinations of functions of the dictionary:

$$f_a = \sum_{\varphi \in \Phi} a_\varphi \varphi,$$

where  $a = (a_\varphi)_{\varphi \in \Phi}$  belongs to  $\mathbb{R}^\Phi$ . Then, our Lasso procedure consists in selecting the vector  $\hat{a}$  by minimizing an  $\ell_1$ -penalized criterion (see (2.1)), where the penalty term takes the form  $\sum_{\varphi \in \Phi} d_\varphi |a_\varphi|$ . Using Bernstein type concentration inequalities for martingales, we propose an original methodology for deriving the data-driven weights  $d_\varphi$ .

We illustrate the general setup with three main examples. First, the case with i.i.d. observations of an inhomogeneous Poisson process on  $[0, 1]$  and unknown intensity. Second, the well known Aalen multiplicative intensity model and third, the central example of the multivariate Hawkes process.

### 1.2.1 The Poisson model

Let us start with a very simple example which will be somehow a toy problem here with respect to the other settings. In this example we take  $T = 1$  and assume that we observe  $M$  i.i.d. Poisson processes on  $[0, 1]$  with common intensity  $f^* : [0, 1] \mapsto \mathbb{R}_+$ . Asymptotic properties are obtained when  $M$  tends to infinity. In this case, for any  $m$ ,

$$\psi_t^{(m)}(f^*) = f^*(t),$$

and  $\mathcal{H} = \mathbb{L}_2([0, 1])$  is equipped with the classical norm defined by

$$\|f\| = \left( \int_0^1 f^2(t) dt \right)^{1/2}.$$

In this case, the support of  $f$ , namely  $[0, 1]$ , does not play a fundamental role. See [44] for adaptive wavelet estimation of non-compactly supported intensity functions.

### 1.2.2 The Aalen multiplicative intensity model

This is one of the most popular counting process because of its adaptivity to various situations (from Markov model to censored life times) and its various applications to biomedical data (see [2]). Given  $\mathcal{X}$  a Hilbert space, we consider  $f^* : [0, T] \times \mathcal{X} \mapsto \mathbb{R}_+$  and we set for any  $t \in \mathbb{R}$ ,

$$\psi_t^{(m)}(f^*) = f^*(t, X^{(m)}) Y_t^{(m)},$$

where  $Y^{(m)}$  is an observable predictable process and  $X^{(m)}$  is a covariate. In this case,  $\mathcal{H} = \mathbb{L}_2([0, T] \times \mathcal{X})$ . To fix ideas one can set  $T = 1$  and  $\mathcal{X} = [0, 1]$ . Hence  $\mathcal{H}$  can also be viewed as  $\mathbb{L}_2([0, 1]^2)$ . In right-censored data,  $f^*$  usually represents the hazard rate. The presence of covariates in this pure non parametric model is the classical generalization of the classical semi-parametric model of Cox (see [34] for instance).

The classical framework consists in assuming that the  $(X^{(m)}, Y^{(m)}, N^{(m)})_{m=1,\dots,M}$  are i.i.d. If there are no covariates, several adaptive approaches already exist (see [9, 10, 43]). In the presence of covariates, see [1, 2] for a parametric approach, see [20, 34] for a model selection approach and [26] for a Lasso approach.

### 1.2.3 The multivariate Hawkes process

Multivariate Hawkes processes are the point process equivalent to autoregressive models. They have extensively been used in sismology to model earthquakes and their aftershocks [56]. More recently they have been used to model favored or avoided distances between occurrences of motifs [28] or Transcription Regulatory Elements on the DNA [19]. Even more recently, they have emerged as a potential model for neuronal networks [18]. For this process, the intensity of a coordinate,  $N_t^{(m)}$ , depends on the history of this coordinate process as well as

the other coordinate processes through linear filters. In this example  $M$  is fixed and asymptotic properties are obtained when  $T \rightarrow \infty$ . With  $\nu^{(m)} \in \mathbb{R}$  and  $h_\ell^{(m)} : (0, \infty) \rightarrow \mathbb{R}$  for  $\ell, m = 1, \dots, M$  and with  $f^*$  the collection of  $\nu^{(m)}$ 's and  $h_\ell^{(m)}$ 's define

$$\psi_t^{(m)}(f^*) = \nu^{(m)} + \sum_{\ell=1}^M \int_{-\infty}^{t-} h_\ell^{(m)}(t-u) dN^{(\ell)}(u). \quad (1.1)$$

We will assume that the support of  $h_\ell^{(m)}$  is bounded. By rescaling we can then assume that the support is in  $(0, 1]$ , and we will do so throughout. Note that in this case we will need to observe the process on  $[-1, T]$  in order to compute  $\psi_t^{(m)}(f^*)$  for  $t \in [0, T]$ . The Hilbert space is

$$\mathcal{H} = (\mathbb{R} \times \mathbb{L}_2([0, 1])^M)^M = \left\{ f = \left( (\mu^{(m)}, (g_\ell^{(m)})_{\ell=1, \dots, M})_{m=1, \dots, M} \right) : g_\ell^{(m)} \text{ with support in } (0, 1] \right. \\ \left. \text{and } \|f\|^2 = \sum_m (\mu^{(m)})^2 + \sum_m \sum_\ell \int_0^1 g_\ell^{(m)}(t)^2 dt < \infty \right\}.$$

Taking the intensity to be  $\lambda_t^{(m)} = \psi_t^{(m)}(f^*)$  is only meaningful if the right hand side is non-negative, and this is the case if the  $\nu^{(m)}$ 's and  $h_\ell^{(m)}$ 's are non-negative. In this case the resulting process is known as the *linear* multivariate Hawkes process (see [30]). It is a well studied process from a probabilistic as well as a statistical point of view. For a parametric approach to the estimation of the interaction functions  $h_\ell^{(m)}$  see [39, 40]. For the use of an AIC criterion see [56]. A non-parametric model selection approach in the case  $M = 1$  is treated in [46] and for  $M = 2$  a combination of AIC and a spline basis expansion is considered in [28].

Note that in [28] and [46], the inhibition case where the functions  $h_\ell^{(m)}$  are negative has been partially studied and in this case  $\lambda_t^{(m)} = (\psi_t^{(m)}(f^*))_+$ . In [19], another parametric variant was studied where the process satisfies  $\lambda_t^{(m)} = \exp(\psi_t^{(m)}(f^*))$ .

### 1.3 Our statistical contribution

From the statistical point of view, our theoretical contribution consists in establishing oracle inequalities. Unlike many papers about theoretical performances of Lasso procedures, we do not wish to obtain assumptions on the dictionary that are as weak as possible but assumptions that can be checked. The first result we establish in Theorem 1 is a basic oracle inequality that clearly states assumptions we need on the Gram matrix  $G$  associated with the dictionary (see (2.2)) and on the weights of our methodology. From the first oracle inequality, we derive a more sophisticated one for general multivariate counting processes in Theorem 2 that gives the shape of data-driven weights by using the Bernstein type inequality of Theorem 3. Both oracle inequalities involve the tradeoff of two terms: an approximation term and a variation term measuring fluctuations of coefficient estimates. Of course, as usual, sparsity is a key point to realize the tradeoff. This general result is applied for the three previous examples of point processes where assumptions on the Gram matrix can be resumed to assumptions on the dictionary. So, unlike most of papers of the literature, these assumptions can be checked. Finally, we carry out a simulation study for the most intricate process, namely the multivariate Hawkes process. Using the framework of neuronal networks, we provide reconstructions of so called spontaneous rates and interactions functions. Data-driven weights for practical purposes are slight modifications of theoretical ones. These modifications essentially aim at reducing the number of tuning parameters to one. Table 1 in Section 6.3 shows that our methodology can easily and robustly be tuned by using limit values imposed by assumptions of Theorem 2. In particular, our tuning parameter is an absolute constant independent of  $T$ . The results for the problem of supports recovery, which is the main goal for high dimensional settings, are quite satisfying. However, due to non-negligible shrinkage that is unavoidable, in particular for large coefficients, we also propose a two steps procedure where estimation of coefficients is handled by using ordinary least squares estimation on the support preliminary determined by our Lasso methodology. We naturally compare our procedures with *adaptive Lasso* of [57] for which weights are proportional to the inverse of ordinary least squares estimates. The latter is very competitive for estimation aspects since shrinkage is all the more negligible as preliminary OLS

estimates are large. But adaptive Lasso has to cope with many difficulties for support recovery. Indeed, unlike our method, adaptive Lasso does not incorporate:

- the nature of the coefficients (our method handles differently the  $\nu^{(m)}$ 's and the coefficients of the interaction functions)

- random fluctuations of coefficient estimators.

In particular, tuning adaptive Lasso in the Hawkes setting is a difficult task, which cannot be tackled by using standard cross-validation. Our simulation study shows that performances of adaptive Lasso are very sensitive to the choice of the tuning parameter which highly depends on  $T$  in a complicated manner. Robustness with respect to tuning is another advantage of our method over adaptive Lasso.

## 1.4 Notation and overview of the paper

Some notation from the general theory of stochastic integration is useful to simplify the otherwise quite heavy notation. If  $H = (H^{(1)}, \dots, H^{(M)})$  is a multivariate process with locally bounded coordinates, say, and  $X = (X^{(1)}, \dots, X^{(M)})$  is a multivariate semi-martingale, we define the real valued process  $H \bullet X$  by

$$H \bullet X_t = \sum_{m=1}^M \int_0^t H_s^{(m)} dX_s^{(m)}.$$

Given  $\phi : \mathbb{R} \mapsto \mathbb{R}$  we use  $\phi(H)$  to denote the coordinatewise application of  $\phi$ , that is  $\phi(H) = (\phi(H^{(1)}), \dots, \phi(H^{(M)}))$ . In particular,

$$\phi(H) \bullet X_t = \sum_{m=1}^M \int_0^t \phi(H_s^{(m)}) dX_s^{(m)}.$$

With  $\psi_t^{(m)}$  as above we define the integrated process  $\Psi(f)$  by

$$\Psi_t^{(m)}(f) = \int_0^t \psi_s^{(m)}(f) ds.$$

With this notation

$$\langle f, g \rangle_T := \psi(f) \bullet \Psi(g)_T = \sum_{m=1}^M \int_0^T \psi_s^{(m)}(f) \psi_s^{(m)}(g) ds$$

is a bilinear form on  $\mathcal{H}$  where the associated quadratic form is denoted  $\|\cdot\|_T^2$ . The compensator  $\Lambda = (\Lambda^{(m)})_{m=1, \dots, M}$  of  $N = (N^{(m)})_{m=1, \dots, M}$  is defined for all  $t$  by

$$\Lambda_t^{(m)} = \int_0^t \lambda_s^{(m)} ds.$$

Section 2 gives our main oracle inequality and the choice of the  $\ell_1$ -weights in the general framework of counting processes. Section 3 provides the fundamental Bernstein-type inequality. Section 4 details the meaning of the oracle inequality in the Poisson and Aalen set-ups. The probabilistic results needed for the Hawkes processes as well as the interpretation of the oracle inequality in this framework is done in Section 5. Simulations on multivariate Hawkes processes are performed in Section 6. The last Section is dedicated to the proofs of our results.

## 2 Lasso estimate and oracle inequality

In the setting of Section 1.2, our goal is to estimate the parameter  $f^*$  non-parametrically. For this purpose we assume a dictionary of functions,  $\Phi$ , to be given, and we define  $f_a$  as a linear combination of the functions of  $\Phi$ , that is,

$$f_a := \sum_{\varphi \in \Phi} a_\varphi \varphi,$$

where  $a = (a_\varphi)_{\varphi \in \Phi}$  belongs to  $\mathbb{R}^\Phi$ . Then, since  $\psi$  is linear, we get

$$\psi(f_a) = \sum_{\varphi \in \Phi} a_\varphi \psi(\varphi).$$

To estimate  $a$  we introduce the quadratic contrast on  $\mathcal{H}$  by

$$\gamma(f) = -2 \psi(f) \bullet N_T + \|f\|_T. \quad (2.1)$$

Since  $\psi$  is linear we obtain

$$\gamma(f_a) = -2a'b + a'Ga$$

where  $a'$  denotes the transpose of the vector  $a$  and for  $\varphi_1, \varphi_2 \in \Phi$ ,

$$b_{\varphi_1} = \psi(\varphi_1) \bullet N_T, \quad G_{\varphi_1, \varphi_2} = \langle \varphi_1, \varphi_2 \rangle_T. \quad (2.2)$$

Note that the Gram matrix  $G$  of dimensions  $|\Phi| \times |\Phi|$  (where  $|\Phi|$  is the cardinality of  $\Phi$ ) may be random but nevertheless observable.

To estimate  $a$  we minimize the contrast,  $\gamma(f_a)$ , subject to an  $\ell_1$ -penalization on the  $a$ -vector. That is, we introduce the following  $\ell_1$ -penalized estimator

$$\hat{a} \in \operatorname{argmin}_{a \in \mathbb{R}^\Phi} \{-2a'b + a'Ga + 2d'|a|\} \quad (2.3)$$

where  $|a| = (|a_\varphi|)_{\varphi \in \Phi}$  and  $d \in \mathbb{R}_+^\Phi$ . With a good choice of  $d$  the solution of (2.3) will achieve both sparsity and have good statistical properties. Finally, we let  $\hat{f} = f_{\hat{a}}$  denote the Lasso estimate of the function  $f^*$  associated with  $\hat{a}$ .

Our first result establishes theoretical properties of  $\hat{f}$  by using the classical oracle approach. More precisely, we establish a bound on the risk of  $\hat{f}$  if some conditions are true. This is a non-probabilistic result that only relies on the definition of  $\hat{a}$  by (2.3). In the next section we will deal with this probabilistic aspect, which is to prove that the conditions are fulfilled with large probability.

**Theorem 1.** *Let  $c > 0$ . If*

$$G \succeq cI \quad (2.4)$$

and if for all  $\varphi \in \Phi$

$$|b_\varphi - \bar{b}_\varphi| \leq d_\varphi, \quad (2.5)$$

where

$$\bar{b}_\varphi = \psi(\varphi) \bullet \Lambda_T,$$

then there exists an absolute constant  $C$ , independent of  $c$ , such that

$$\|\hat{f} - f^*\|_T^2 \leq C \inf_{a \in \mathbb{R}^\Phi} \left\{ \|f^* - f_a\|_T^2 + c^{-1} \sum_{\varphi \in S(a)} d_\varphi^2 \right\}, \quad (2.6)$$

where  $S(a)$  is the support of  $a$ .

The proof of Theorem 1 is given in Section 7.1. Note that Assumption (2.4) ensures that  $G$  is invertible and then coordinates of  $\hat{a}$  are finite almost surely. Assumption (2.4) also ensures that  $\|f\|_T$  is a real norm on  $f$  at least when  $f$  is a linear combination of the functions of  $\Phi$ .

Two terms are involved on the right hand side of (2.6). The first one is an approximation term and the second one can be viewed as a variance term providing control of the random fluctuations of the  $b_\varphi$ 's around the  $\bar{b}_\varphi$ 's. Note that  $b_\varphi - \bar{b}_\varphi = \psi(\varphi) \bullet (N - \Lambda)_T$  is a martingale (see also the comments after Theorem 2 for more details). The approximation term can be small but the price to pay may be a large support of  $a$ , leading to large values for the second term. Conversely, a sparse  $a$  leads to a small second term. But in this case the

approximation term is potentially larger. Note that if the function  $f^*$  can be approximated by a sparse linear combination of the functions of  $\Phi$ , then we obtain a sharp control of  $\|\hat{f} - f^*\|_T^2$ . In particular, if  $f^*$  can be decomposed on the dictionary, so we can write  $f^* = f_{a^*}$  for some  $a^* \in \mathbb{R}^\Phi$ , then (2.6) gives

$$\|\hat{f} - f^*\|_T^2 \leq Cc^{-1} \sum_{\varphi \in S(a^*)} d_\varphi^2.$$

In this case, the right hand side can be viewed as the sum of the estimation errors made by estimating the components of  $a^*$ .

Such oracle inequalities are now classical in the huge literature of Lasso procedures. See for instance [4, 5, 12, 13, 14, 15, 33, 53], who established oracle inequalities in the same spirit as in Theorem 1. We bring out the paper [17], which gives technical and heuristic arguments for justifying optimality of such oracle inequalities (see Section 1.3 of [17]). Most of these papers, that deal with independent data, aim at establishing oracle inequalities under assumptions as weak as possible on the design matrix. We refer the reader to [54] or [11] for a good review and a hierarchy of these assumptions. Assumption (2.4), that can also be found in [15], is not the weakest one since it involves *simultaneously* all columns of  $G$  unlike assumptions based on restricted isometry constants. Remember that for any positive integer  $S$ , the  $S$ -restricted isometry constant associated with a matrix  $G$  is the smallest quantity  $\delta_S$  satisfying

$$(1 - \delta_S)\|x\|_{\ell_2} \leq \|Gx\|_{\ell_2} \leq (1 + \delta_S)\|x\|_{\ell_2},$$

for any  $S$ -sparse vector  $x$  (see the seminal paper [16]). As mentioned, the main contributions of this paper is not to obtain assumptions as weak as possible on the matrix  $G$ , but rather to prove that Assumption (2.4) is satisfied with large probability. We adapt the same approach as [48, 49] and to a lesser extent as Section 2.1 of [17] or [47]. Section 5 is in particular mainly devoted to show that (2.4) holds with large probability for the multivariate Hawkes processes.

For Theorem 1 to be of interest, the condition on the martingale, condition (2.5), needs to hold with large probability as well. Therefore, one of the main contribution of the paper is to provide new sharp concentration inequalities that are satisfied by multivariate point processes. This is the main goal of Theorem 3 in Section 3 where we establish Bernstein type inequalities for martingales. We apply it to the control of (2.5). This allows us to derive the following result, which specifies the choice of the  $d_\varphi$ 's needed to obtain the oracle inequality with large probability.

**Theorem 2.** *Let  $N = (N^{(m)})_{m=1,\dots,M}$  be a multivariate counting process with predictable intensities  $\lambda_t^{(m)}$  and almost surely finite corresponding compensator  $\Lambda_t^{(m)}$ . Define*

$$\Omega_{V,B} = \left\{ \text{for any } \varphi \in \Phi, \sup_{t \in [0,T], m} |\psi_t^{(m)}(\varphi)| \leq B_\varphi \text{ and } (\psi(\varphi))^2 \bullet N_T \leq V_\varphi \right\},$$

for positive deterministic constants  $B_\varphi$  and  $V_\varphi$  and

$$\Omega_c = \{G \succeq cI\}.$$

Let  $x$  and  $\varepsilon$  be strictly positive constants and define for all  $\varphi \in \Phi$ ,

$$d_\varphi = \sqrt{2(1 + \varepsilon)\hat{V}_\varphi^\mu x + \frac{B_\varphi x}{3}}, \quad (2.7)$$

with

$$\hat{V}_\varphi^\mu = \frac{\mu}{\mu - \phi(\mu)} (\psi(\varphi))^2 \bullet N_T + \frac{B_\varphi^2 x}{\mu - \phi(\mu)}$$

for a real number  $\mu$  such that  $\mu > \phi(\mu)$ , where  $\phi(u) = \exp(u) - u - 1$ . Let us consider the Lasso estimator  $\hat{f}$  of  $f^*$  defined in Section 2. Then, with probability larger than

$$1 - 4 \sum_{\varphi \in \Phi} \left( \frac{\log \left( 1 + \frac{\mu V_\varphi}{B_\varphi^2 x} \right)}{\log(1 + \varepsilon)} + 1 \right) e^{-x} - \mathbb{P}(\Omega_{V,B}^c) - \mathbb{P}(\Omega_c^c),$$

inequality (2.6) is satisfied, i.e.

$$\|\hat{f} - f^*\|_T^2 \leq C \inf_{a \in \mathbb{R}^{\Phi}} \left\{ \|f^* - f_a\|_T^2 + c^{-1} \sum_{\varphi \in S(a)} d_{\varphi}^2 \right\},$$

where  $C$  is a constant independent of  $c$ ,  $\Phi$ ,  $T$  and  $M$ .

Of course, the smaller the  $d_{\varphi}$ 's the better the oracle inequality. So for the choice of  $x$ , we have to realize a compromise to obtain a meaningful oracle inequality on an event with large probability. Let us discuss more deeply the definition of  $d_{\varphi}$  (derived from subsequent Theorem 3) which seems intricate. Up to a constant depending on the choice of  $\mu$  and  $\varepsilon$ ,  $d_{\varphi}$  is of same order as  $\max(\sqrt{x(\psi(\varphi))^2 \bullet N_T}, B_{\varphi}x)$ . To give more insight on the values of  $d_{\varphi}$ , let us consider the very special case where for any  $m \in \{1, \dots, M\}$  for any  $s$ ,  $\psi_s^{(m)}(\varphi) = c_m 1_{\{s \in A_m\}}$ , where  $c_m$  is a positive constant and  $A_m$  a compact set included into  $[0, T]$ . In this case, by naturally choosing  $B_{\varphi} = \max_{1 \leq m \leq M} c_m$ , we have:

$$\sqrt{x(\psi(\varphi))^2 \bullet N_T} \geq B_{\varphi}x \iff \sum_{m=1}^M c_m^2 N_{A_m}^{(m)} \geq x \max_{1 \leq m \leq M} c_m^2,$$

where  $N_{A_m}^{(m)}$  represents the number of points of  $N^{(m)}$  falling in  $A_m$ . For more general vector functions  $\psi(\varphi)$ , the term  $\sqrt{x(\psi(\varphi))^2 \bullet N_T}$  will dominate  $B_{\varphi}x$  if the number of points of the process lying where  $\psi(\varphi)$  is large, is significant. In this case, the leading term in  $d_{\varphi}$  is expected to be the quadratic term  $\sqrt{2(1+\varepsilon)\frac{\mu}{\mu-\phi(\mu)}x(\psi(\varphi))^2 \bullet N_T}$  and the linear terms in  $x$  can be viewed as residual terms. Furthermore, note that when  $\mu$  tends to 0,

$$\frac{\mu}{\mu - \phi(\mu)} = 1 + \frac{\mu}{2} + o(\mu), \quad \frac{x}{\mu - \phi(\mu)} \sim \frac{x}{\mu} \rightarrow +\infty$$

since  $x > 0$ . So, if  $\mu$  and  $\varepsilon$  tend to 0, the quadratic term tends to  $\sqrt{2x(\psi(\varphi))^2 \bullet N_T}$  but the price to pay is the explosion of the linear term in  $x$ . In any case, it is possible to make the quadratic term as close to  $\sqrt{2x(\psi(\varphi))^2 \bullet N_T}$  as desired.

Let us emphasize the importance of this last quadratic term. Since this corresponds to the rate given by the central limit theorem, this means that we have some chance to have sharp values for the components of  $d_{\varphi}$ . Remember that the smaller the  $d_{\varphi}$ 's, the better the oracle inequality. Furthermore, in more classical contexts such as density estimation (see [4]), it is shown that if the components of  $d_{\varphi}$  are chosen smaller than the analog of  $\sqrt{2x(\psi(\varphi))^2 \bullet N_T}$  then the resulting estimator is definitely a bad one, but simulations show that, to some extent, if the components of  $d$  are larger than the analog of  $\sqrt{2x(\psi(\varphi))^2 \bullet N_T}$ , then the estimator deteriorates too. A similar result is out of reach in our setting, but similar conclusions may remain valid here since density estimation often provides some clues about what happens for more intricate heteroscedastic models. See also the simulation study in Section 6.

Finally, it remains to control  $\mathbb{P}(\Omega_{V,B})$  and  $\mathbb{P}(\Omega_c)$ . This is the goals of Section 4 for Poisson and Aalen models and Section 5 for multivariate Hawkes processes.

### 3 Bernstein type inequalities for multivariate point processes

We establish a Bernstein type concentration inequality based on boundedness assumptions. This result, which has an interest per se from the probabilistic point of view, was the key result to derive the convenient values for the vector  $d$  in Theorem 2 and so is capital from the statistical perspective.

**Theorem 3.** *Let  $N = (N^{(m)})_{m=1, \dots, M}$  be a multivariate counting process with predictable intensities  $\lambda_t^{(m)}$  and corresponding compensator  $\Lambda_t^{(m)}$  with respect to some given filtration. Let  $B > 0$ . Let  $H = (H^{(m)})_{m=1, \dots, M}$  be a multivariate predictable process such that for all  $\xi \in (0, 3)$ , for all  $t$ ,*

$$\exp(\xi H/B) \bullet \Lambda_t < \infty \text{ a.s. and } \exp(\xi H^2/B^2) \bullet \Lambda_t < \infty \text{ a.s.} \quad (3.1)$$



Let us consider the martingale defined for all  $t \geq 0$  by

$$M_t = H \bullet (N - \Lambda)_t.$$

Let  $v > w$  and  $x$  be positive constants and let  $\tau$  be a bounded stopping time. Let us define

$$\hat{V}^\mu = \frac{\mu}{\mu - \phi(\mu)} H^2 \bullet N_\tau + \frac{B^2 x}{\mu - \phi(\mu)}$$

for a real number  $\mu \in (0, 3)$  such that  $\mu > \phi(\mu)$ , where  $\phi(u) = \exp(u) - u - 1$ . Then, for any  $\varepsilon > 0$ ,

$$\mathbb{P} \left( M_\tau \geq \sqrt{2(1 + \varepsilon)\hat{V}^\mu x} + \frac{Bx}{3} \text{ and } w \leq \hat{V}^\mu \leq v \text{ and } \sup_{m, t \leq \tau} |H_t^{(m)}| \leq B \right) \leq 2 \left( \frac{\log(v/w)}{\log(1 + \varepsilon)} + 1 \right) e^{-x}. \quad (3.2)$$

This result is based on the exponential martingale for counting processes, which has been used for a while in the context of the counting process theory. See for instance [7], [50] or [52]. This basically gives a concentration inequality taking the following form (see (7.7)) (the result is stated here in its univariate form for comparison purposes): for any  $x > 0$ ,

$$\mathbb{P} \left( M_\tau \geq \sqrt{2\rho x} + \frac{Bx}{3} \text{ and } \int_0^\tau H_s^2 d\Lambda_s \leq \rho \right) \leq e^{-x}. \quad (3.3)$$

Typically, in (3.3),  $\rho$  is not random and  $B$  is a deterministic upper bound of  $\sup_{s \in [0, \tau]} |H_s|$ . The leading term for moderate values of  $x$  and  $\tau$  large enough is  $\sqrt{2\rho x}$  where the constant  $\sqrt{2}$  is not improvable since this coincides with the rate of the central limit theorem for martingales. Theorem 3 consists in plugging the estimate  $\hat{v} = H^2 \bullet N_\tau$  instead of a non sharp deterministic upper bound of  $v = H^2 \bullet \Lambda_\tau$ . The proof is based on a peeling argument that was first introduced in [35] for Gaussian processes.

Note that there exist also inequalities that seem nicer than (7.7) which constitutes the basic brick for our purpose. For instance, [24] establish that for any deterministic positive real number  $\theta$ , for any  $x > 0$ ,

$$\mathbb{P} \left( M_\tau \geq \sqrt{2\theta x} \text{ and } \int_0^\tau H_s^2 d\Lambda_s + \int_0^\tau H_s^2 dN_s \leq \theta \right) \leq e^{-x}. \quad (3.4)$$

At first sight, this seems better than Theorem 3 because no linear term depending on  $B$  appears, but if we want to use the estimate  $2 \int_0^\tau H_s^2 dN_s$  instead of  $\theta$  in the inequality, we will have to bound  $|H_s|$  by some  $B$  in any case. Moreover, by doing so, the quadratic term will be of order  $\sqrt{4\hat{v}x}$  which is worse than the term  $\sqrt{2\hat{v}x}$  derived in Theorem 3, even if this constant  $\sqrt{2}$  can only be reached asymptotically in our case.

There exists a better result if the martingale  $M_t$  is conditionally symmetric (see [24] but also [22] and [3] for the discrete time case): for any  $x > 0$ ,

$$\mathbb{P} \left( M_\tau \geq \sqrt{2\kappa x} \text{ and } \int_0^\tau H_s^2 dN_s \leq \kappa \right) \leq e^{-x}, \quad (3.5)$$

which almost seems to be the ideal one. But there are actually two major flaws in this inequality. First, one would need to assume that the martingale is conditionally symmetric, which cannot be the case in our situation for general counting processes and general dictionaries. Secondly, we have the deterministic upper bound  $\kappa$  instead of  $\hat{v}$ . To replace it by  $\hat{v}$  and apply peeling arguments as in the proof of Theorem 3, we need to assume the existence of a positive constant  $w$  such that  $\hat{v} \geq w$ . But if the process happens to be empty, then  $\hat{v} = 0$ , so we cannot generally find such a lower bound, whereas in our theorem, we can always take  $w = \frac{B^2 x}{\mu - \phi(\mu)}$  as a lower bound for  $\hat{V}^\mu$ .

Finally, note that in Proposition 5 (see Section 7.3), we also derived a similar bound where  $\hat{V}^\mu$  is replaced by  $\int_0^\tau H_s^2 d\Lambda_s$ . Basically, it means that the same type of results hold for quadratic characteristic instead of quadratic variation. If this quadratic characteristic result is of little use here since the quadratic characteristic is not observable, we think that it may be of interest for readers looking for self-normalized results as in [23].

## 4 Applications to the Poisson and Aalen models

We apply Theorem 2 to the Poisson and Aalen models. The case of the multivariate Hawkes process, which is much more intricate, will be the subject of the next section.

### 4.1 The Poisson model

Let us recall that in this case, we observe  $M$  i.i.d. Poisson processes with intensity  $f^*$  supported by  $[0, 1]$  and that the meaningful norm is given by  $\|f\|^2 = \int_0^1 f^2(x)dx$ . We assume that  $\Phi$  is an orthonormal system for  $\|\cdot\|$ . In this case,

$$\|\cdot\|_T^2 = M\|\cdot\|^2 \quad \text{and} \quad G = MI,$$

where  $I$  is the identity matrix. One applies Theorem 2 with  $c = M$  (so  $\mathbb{P}(\Omega_c^c) = 0$ ) and

$$B_\varphi = \|\varphi\|_\infty, \quad V_\varphi = \|\varphi\|_\infty^2(1 + \delta)Mm_1,$$

for  $\delta > 0$  and  $m_1 = \int_0^1 f^*(t)dt$ . Note that here  $T = 1$  and therefore  $N_T^{(m)} = N_1^{(m)}$  is the total number of observed points for the  $m$ th process. Using

$$\psi(\varphi)^2 \bullet N_T \leq \|\varphi\|_\infty^2 \sum_{m=1}^M N_1^{(m)}$$

and since the distribution of  $\sum_{m=1}^M N_1^{(m)}$  is the Poisson distribution with parameter  $Mm_1$ , Cramer-Chernov arguments give:

$$\mathbb{P}(\Omega_{V,B}^c) \leq \mathbb{P}\left(\sum_{m=1}^M N_1^{(m)} > (1 + \delta)Mm_1\right) \leq \exp(-\{(1 + \delta)\ln(1 + \delta) - \delta\}Mm_1).$$

For  $\alpha > 0$ , by choosing  $x = \alpha \log(M)$ , we finally obtain the following corollary derived from Theorem 2.

**Corollary 1.** *With probability larger than  $1 - C_1 \frac{|\Phi| \log(M)}{M^\alpha} - e^{-C_2 M}$ , where  $C_1$  is a constant depending on  $\mu$ ,  $\varepsilon$ ,  $\alpha$ ,  $\delta$  and  $m_1$  and  $C_2$  is a constant depending on  $\delta$  and  $m_1$ , we have:*

$$\|\hat{f} - f^*\|^2 \leq C \inf_{a \in \mathbb{R}^\Phi} \left\{ \|f^* - f_a\|^2 + \frac{1}{M^2} \sum_{\varphi \in S(a)} \left( \log(M) \sum_{m=1}^M \int_0^1 \varphi^2(x) dN_x^{(m)} + \log^2(M) \|\varphi\|_\infty^2 \right) \right\},$$

where  $C$  is a constant depending on  $\mu$ ,  $\varepsilon$ ,  $\alpha$ ,  $\delta$  and  $m_1$ .

To shed some lights on this result, consider an asymptotic perspective by assuming that  $M$  is large. Assume also, for sake of simplicity, that  $f^*$  is bounded below from 0 on  $[0, 1]$ . If the dictionary  $\Phi$  (whose size may depend on  $M$ ) satisfies

$$\max_{\varphi \in \Phi} \|\varphi\|_\infty = o\left(\sqrt{\frac{M}{\log M}}\right),$$

then, since, almost surely,

$$\frac{1}{M} \sum_{m=1}^M \int_0^1 \varphi^2(x) dN_x^{(m)} \xrightarrow{M \rightarrow \infty} \int_0^1 \varphi^2(x) f^*(x) dx,$$

almost surely,

$$\frac{1}{M^2} \sum_{\varphi \in S(a)} \left( \log(M) \sum_{m=1}^M \int_0^1 \varphi^2(x) dN_x^{(m)} + \log^2(M) \|\varphi\|_\infty^2 \right) = \log M \sum_{\varphi \in S(a)} \frac{1}{M} \int_0^1 \varphi^2(x) f^*(x) dx \times (1 + o(1)).$$

The right hand term corresponds, up to the logarithmic term, to the sum of variance terms when estimating  $\int_0^1 \varphi(x) f^*(x) dx$  with  $\frac{1}{M} \sum_{m=1}^M \int_0^1 \varphi(x) dN_x^{(m)}$  for  $\varphi \in S(a)$ . This means that the estimator adaptively achieves the best trade-off between a bias term and a variance term. The logarithmic term is the price to pay for adaptation. We refer the reader to [44] for a deep discussion on optimality of such results.

## 4.2 The Aalen model

Similar results presented in this paragraph can be found in [26] under alternative assumptions on the dictionary. Unlike the previous model, Assumption (2.4) is hard to check here since the intensity depends on covariates and variables  $Y^{(m)}$ 's. [26] use restricted eigenvalues conditions instead of (2.4) but this similarly expresses some orthogonality properties of columns of  $G$ , that are non-mild conditions as well.

Recall that we observe an  $M$ -sample  $(X^{(m)}, Y^{(m)}, N^{(m)})_{m=1, \dots, M}$ , with  $Y^{(m)} = (Y_t^{(m)})_{t \in [0;1]}$  and  $N^{(m)} = (N_t^{(m)})_{t \in [0;1]}$ . We assume that  $X^{(m)} \in [0, 1]$  and that the intensity of  $N_t^{(m)}$  is  $f^*(t, X^{(m)})Y_t^{(m)}$  and we set

$$\|f\|^2 := \mathbb{E} \left( \int_{[0,1]^2} f^2(t, X^{(1)})(Y_t^{(1)})^2 dt \right).$$

We assume that  $\|\cdot\|$  is a true norm. For instance if there are no covariates, it is equivalent to assuming that  $\mathbb{E}((Y_t^{(1)})^2) \neq 0$  on  $[0, 1]$  i.e.  $Y_t^{(1)}$  cannot be zero almost surely and this for all  $t$  in  $[0, 1]$ . This is natural since of course one cannot estimate  $f^*(t)$  if  $Y_t^{(1)} = 0$  almost surely. If  $Y_t^{(1)}$  is deterministic and non zero on  $[0, 1]$  then we are in the case of a Cox process ( $N^{(m)}$  is a Poisson process given the covariates  $X^{(m)}$ ), and it is natural to say that we will be able to measure  $f^*$  only on the support of the variables  $X^{(m)}$ . Note that  $\|f\|_{emp}$  defined by

$$\|f\|_{emp}^2 := \frac{1}{M} \|f\|_T^2 = \frac{1}{M} \sum_{m=1}^M \int_0^1 f^2(t, X^{(m)})(Y_t^{(m)})^2 dt$$

corresponds to the empirical version of  $\|f\|$ . We assume that  $\Phi$  is an orthonormal system for  $\|\cdot\|_2$  (the classical norm on  $\mathbb{L}_2([0, 1]^2)$ ) and we assume that there exists a positive constant  $r$  such that for all  $f \in \mathbb{L}_2([0, 1]^2)$ ,  $\|f\| \geq r\|f\|_2$ .

The control of  $\Omega_c$  is much more cumbersome for the Aalen case, even if it is less intricate than the control for Hawkes processes (see Section 5). To avoid another set of tedious computations, we just give here a brief sketch of what one could do. To control  $\Omega_c$ , we only need to concentrate the elements of  $G$  around their mean since they are sum of i.i.d. variables and use the fact that  $\mathbb{E}(G) \succeq Mr^2I$ . Then the probability of  $\Omega_c^c$  can be proved to be smaller than  $\frac{|\Phi|^2}{M^\alpha}$  up to a constant if one chooses  $c = Mr^2(1 - \delta)$  and if one assumes that  $|\Phi| = o(\sqrt{T} \log(T)^{-\beta})$  (where of course  $\alpha, \beta$  and  $\delta$  are convenient positive constants).

For the sequel, we use two classical assumptions (see [43] for instance):

- $\sup_{t \in [0;1]} \max_{m \in \{1, \dots, M\}} Y_t^{(m)} \leq 1$  almost surely.
- For some positive constant  $R$ ,  $\max_{m \in \{1, \dots, M\}} N_1^{(m)} \leq R$  almost surely.

Therefore, almost surely,

$$\psi(\varphi)^2 \bullet N_T = \sum_{m=1}^M \int_0^1 [Y_t^{(m)}]^2 \varphi^2(t, X^{(m)}) dN_t^{(m)} \leq \sum_{m=1}^M \int_0^1 \varphi^2(t, X^{(m)}) dN_t^{(m)} \leq MR \|\varphi\|_\infty^2.$$

So, we apply Theorem 2 with  $B_\varphi = \|\varphi\|_\infty$ ,  $V_\varphi = MR \|\varphi\|_\infty^2$  (so  $\mathbb{P}(\Omega_{V,B}) = 1$ ) and  $x = \alpha \log(M)$  for  $\alpha > 0$ . We finally obtain the following corollary.

**Corollary 2.** *With probability larger than  $1 - C_1 \frac{|\Phi| \log(M)}{M^\alpha} - \mathbb{P}(\Omega_c^c)$ , where  $C_1$  is a constant depending on  $\mu, \varepsilon, \alpha$  and  $R$ , we have:*

$$\|\hat{f} - f^*\|_{emp}^2 \leq C \inf_{a \in \mathbb{R}^\Phi} \left\{ \|f^* - f_a\|_{emp}^2 + \frac{1}{M^2} \sum_{\varphi \in S(a)} \left( \log(M) \sum_{m=1}^M \int_0^1 \varphi^2(t, X^{(m)}) dN_t^{(m)} + \log^2(M) \|\varphi\|_\infty^2 \right) \right\}$$

where  $C$  is a constant depending on  $\mu, \varepsilon, \alpha$  and  $R$ .

To shed lights on this result, assume that the density of the  $X^{(m)}$ 's is upper bounded by a constant  $\tilde{R}$ . In an asymptotic perspective with  $M \rightarrow \infty$ , we have almost surely,

$$\frac{1}{M} \sum_{m=1}^M \int_0^1 \varphi^2(t, X^{(m)}) dN_t^{(m)} \rightarrow \mathbb{E} \left( \int_0^1 \varphi^2(t, X^{(1)}) f^*(t, X^{(1)}) Y^{(1)} dt \right).$$

But

$$\mathbb{E} \left( \int_0^1 \varphi^2(t, X^{(1)}) f^*(t, X^{(1)}) Y^{(1)} dt \right) \leq \|f^*\|_\infty \mathbb{E} \left( \int_0^1 \varphi^2(t, X^{(1)}) dt \right) \leq \tilde{R} \|f^*\|_\infty.$$

So, if the dictionary  $\Phi$  (whose size may depend on  $M$ ) satisfies

$$\max_{\varphi \in \Phi} \|\varphi\|_\infty = O \left( \sqrt{\frac{M}{\log M}} \right),$$

then, almost surely, the variance term is asymptotically smaller than  $\log(M) \frac{|S(a)| \|f^*\|_\infty}{M}$  up to constants. So, we can draw the same conclusions as for the Poisson model.

## 5 Applications to the case of multivariate Hawkes process

### 5.1 Identification of the parameters

For a multivariate Hawkes model, the parameter  $f^*$  belongs to

$$\mathcal{H} = \mathbb{H}^M = \left\{ f = (\mathbf{f}^{(m)})_{m=1, \dots, M} \mid \mathbf{f}^{(m)} \in \mathbb{H} \text{ and } \|f\|^2 = \sum_{m=1}^M \|\mathbf{f}^{(m)}\|^2 \right\}$$

where

$$\mathbb{H} = \left\{ \mathbf{f} = (\mu, (g_\ell)_{\ell=1, \dots, M}) \mid \mu \in \mathbb{R}, g_\ell \text{ with support in } (0, 1] \text{ and } \|\mathbf{f}\|^2 = \mu^2 + \sum_{\ell=1}^M \int_0^1 g_\ell^2(x) dx < \infty \right\}.$$

If one defines  $\kappa$  the linear predictable transformation of  $\mathbb{H}$  defined by

$$\kappa_t(\mathbf{f}) = \mu + \sum_{\ell=1}^M \int_{t-1}^{t-} g_\ell(t-u) dN_u^{(\ell)}, \quad (5.1)$$

then the transformation  $\psi$  on  $\mathcal{H}$  is just defined by

$$\psi_t^{(m)}(f) = \kappa_t(\mathbf{f}^{(m)}).$$

Before stating oracle inequalities for Lasso estimates, we need to prove some probabilistic results. They will be useful to deal with  $\mathbb{P}(\Omega_{V,B})$  and  $\mathbb{P}(\Omega_c)$ .

### 5.2 Some useful probabilistic results for multivariate Hawkes processes

In this paragraph, we present some particular exponential results and tail controls for Hawkes processes. Up to our knowledge, these results are new: They constitute the generalization of [45] to the multivariate case. In this paper, they are used to control  $\mathbb{P}(\Omega_c^c)$  and  $\mathbb{P}(\Omega_{V,B}^c)$  but they may be of self-interest.

Since the functions  $h_\ell^{(m)}$ 's are nonnegative, a cluster representation exists. We can indeed construct the Hawkes process by the Poisson cluster representation (see [21]) as follows:

- Distribute *ancestral points* with marks  $\ell = 1, \dots, M$  according to homogeneous Poisson processes with intensities  $\nu^{(\ell)}$  on  $\mathbb{R}$ .

- For each ancestral point, form a cluster of descendant points. More precisely, starting with an ancestral point at time 0 of a certain type, we successively build new generations as Poisson processes with intensity  $h_\ell^{(m)}(\cdot - T)$ , where  $T$  is the parent of type  $\ell$  (the corresponding children being of type  $m$ ). We are in the situation where this process extinguishes and we denote by  $H$  the last children of all generations, which also represents the length of the cluster. Note that the number of descendants is a multitype branching process (and there exists a branching cluster representation (see [8, 21, 30]) with offspring distributions being Poisson variables with means

$$\gamma_{\ell,m} = \int_0^1 h_\ell^{(m)}(t) dt.$$

The essential part we need is that the expected number of offsprings of type  $m$  from a point of type  $\ell$  is  $\gamma_{\ell,m}$ . With  $\Gamma = (\gamma_{\ell,m})_{\ell,m=1,\dots,M}$  the matrix of expectations the theory of multitype branching processes gives that the clusters are finite almost surely if and only if the spectral radius of  $\Gamma$  is smaller than or equal to 1. In this case, there is a stationary version of the Hawkes process by the Poisson cluster representation.

Below we will need the stronger requirement that  $\Gamma$  has spectral radius strictly smaller than 1 to ensure a bound on the number of points in a cluster. We denote by  $\mathbb{P}_\ell$  the law of the cluster whose ancestral point is of type  $\ell$ ,  $\mathbb{E}_\ell$  is the corresponding expectation.

The following lemma is very general and holds even if the function  $g_\ell^{(m)}$  have infinite support as long as the spectral radius  $\Gamma$  is strictly less than 1.

**Lemma 1.** *If  $W$  denotes the total number of points of any type in the cluster whose ancestral point is of type  $\ell$  then if the spectral radius of  $\Gamma$  is strictly smaller than 1 there exists  $\vartheta_\ell > 0$ , only depending on  $\ell$  and on  $\Gamma$ , such that*

$$\mathbb{E}_\ell(e^{\vartheta_\ell W}) < \infty.$$

This easily leads to the following result, which provides the existence of the Laplace transform of the total number of points in an arbitrary bounded interval, when the function  $g_\ell^{(m)}$  have bounded support.

**Proposition 1.** *Let  $N$  be a stationary multivariate Hawkes process, with bounded support interactions functions and such that the spectral radius of  $\Gamma$  is strictly smaller than 1. For any  $A > 0$ , let us define  $N_{[-A,0]}$  the total number of points of  $N$  in  $[-A, 0)$ , all marks included. Then there exists a constant  $\theta > 0$ , depending on the distribution of the process and on  $A$  such that*

$$\mathcal{E} := \mathbb{E}(e^{\theta N_{[-A,0]}}) < \infty,$$

which implies that for all positive  $u$

$$\mathbb{P}(N_{[-A,0]} \geq u) \leq \mathcal{E} e^{-\theta u}.$$

Moreover one can precise the ergodic theorem in a non-asymptotic way.

**Proposition 2.** *Let  $A > 0$  and let  $Z(N)$  be a function depending on the points lying in  $[-A, 0)$  of a stationary multivariate Hawkes process,  $N$ , with parameter  $f^* \in \mathcal{H}$ . Assume that there exist  $b$  and  $\eta$  non-negative constants such that*

$$|Z(N)| \leq b(1 + N_{[-A,0]}^\eta),$$

where  $N_{[-A,0]}$  represents the total number of points of  $N$  in  $[-A, 0)$ , all marks included. We denote  $\theta$  the shift operator, meaning that  $Z \circ \theta_t(N)$  depends now in the same way as  $Z(N)$  on some points that are now the points of  $N$  lying in  $[t - A, t)$ .

We assume  $\mathbb{E}[|Z(N)|] < \infty$  and for short, we denote  $\mathbb{E}(Z) = \mathbb{E}[Z(N)]$ . Then, for any  $\alpha > 0$ , there exists a constant  $\mathcal{T}(\alpha, \eta, f^*, A) > 1$  such that for  $T > \mathcal{T}(\alpha, \eta, f^*, A)$ , there exist  $C_1, C_2, C_3$  and  $C_4$  positive constants depending on  $\alpha, \eta, A$  and  $f^*$  such that

$$\mathbb{P}\left(\int_0^T [Z \circ \theta_t(N) - \mathbb{E}(Z)] dt \geq C_1 \sigma \sqrt{T \log^3(T)} + C_2 b (\log(T))^{2+\eta}\right) \leq \frac{C_4}{T^\alpha},$$

with  $\tilde{\mathcal{N}} = C_3 \log(T)$  and  $\sigma^2 = \mathbb{E}([Z(N) - \mathbb{E}(Z)]^2 \mathbb{1}_{N_{[-A,0]} \leq \tilde{\mathcal{N}}})$ .

Finally, to deal with the control of  $\mathbb{P}(\Omega_c)$ , we shall need the next result. First, we define a quadratic form  $Q$  on  $\mathbb{H}$  by

$$Q(\mathbf{f}, \mathbf{g}) = \mathbb{E}_{\mathbb{P}} (\kappa_1(\mathbf{f})\kappa_1(\mathbf{g})) = \mathbb{E}_{\mathbb{P}} \left( \frac{1}{T} \int_0^T \kappa_t(\mathbf{f})\kappa_t(\mathbf{g})dt \right), \quad \mathbf{f}, \mathbf{g} \in \mathbb{H}. \quad (5.2)$$

We have:

**Proposition 3.** *For a stationary Hawkes process with intensities given by (1.1), which fulfill*

$$\min_{m \in \{1, \dots, M\}} \nu^{(m)} > 0 \quad \text{and} \quad \max_{l, m \in \{1, \dots, M\}} \sup_{t \in [0, 1]} h_\ell^{(m)}(t) < \infty \quad (5.3)$$

and where the spectral radius of  $\Gamma$  is strictly smaller than 1, there is a constant  $\zeta > 0$  such that for any  $\mathbf{f} \in \mathbb{H}$ ,

$$Q(\mathbf{f}, \mathbf{f}) \geq \zeta \|\mathbf{f}\|^2.$$

We are now ready to establish oracle inequalities for multivariate Hawkes processes.

### 5.3 Lasso for Hawkes processes

In the sequel, we still consider the main assumptions of the previous subsection: stationarity, (5.3) and the fact that the spectral radius of  $\Gamma$  is strictly smaller than 1. We recall that the components of  $\Gamma$  are the  $\gamma_{\ell, m}$ 's with

$$\gamma_{\ell, m} = \int_0^1 h_\ell^{(m)}(t)dt.$$

One of the main result of this section is to link properties of the dictionary (mainly orthonormality but also more involved assumptions) to properties of  $G$  (the control of  $\Omega_c$ ). To do so let us define for all  $f \in \mathcal{H}$ ,

$$\|f\|_\infty = \max \left\{ \max_{m=1, \dots, M} |\mu^{(m)}|, \max_{m, \ell=1, \dots, M} \|g_\ell^{(m)}\|_\infty \right\}.$$

Then, let us define by  $\|\Phi\|_\infty := \max\{\|\varphi\|_\infty, \varphi \in \Phi\}$ , and recall that  $|\Phi|$  is the cardinality of  $\Phi$ .

The next result is based on the probabilistic results of Section 5.2.

**Proposition 4.** *Assume that the Hawkes process is stationary, that (5.3) is satisfied and that the spectral radius of  $\Gamma$  is strictly smaller than 1. Let  $r_\Phi$  be the spectral radius of the matrix  $\Lambda$  defined by*

$$\Lambda = \left( \sum_m \left[ |\mu_\varphi^{(m)}| |\mu_\rho^{(m)}| + \sum_{\ell=1}^M \int_0^1 |(g_\varphi)_\ell^{(m)}| |(g_\rho)_\ell^{(m)}|(u)du \right] \right)_{\varphi, \rho \in \Phi}.$$

Assume that  $\Phi$  is orthonormal and that

$$A_\Phi(T) := r_\Phi \|\Phi\|_\infty^2 |\Phi| [\log(\|\Phi\|_\infty) + \log(|\Phi|)] \frac{\log^5(T)}{T} \rightarrow 0 \quad (5.4)$$

when  $T \rightarrow \infty$ . Then, for any  $\alpha > 0$ , there exists  $C_1 > 0$  depending on  $\alpha$  and  $f^*$  such that with  $c = C_1 T$ , we have

$$\mathbb{P}(\Omega_c^c) = O(T^{-\alpha}).$$

Now, let us deal with the choice of the dictionary  $\Phi$ . The easiest case, and the only one we will consider here for sake of simplicity, is built via a dictionary  $(\Upsilon_k)_{k=1, \dots, K}$  of functions of  $\mathbb{L}_2((0, 1])$  (that may depend on  $T$ ) in the following way. A function  $\varphi = (\mu_\varphi^{(m)}, ((g_\varphi)_\ell^{(m)})_\ell)_m$  belongs to  $\Phi$  if and only if only one of its  $M + M^2$  components is non zero and in this case,

- if  $\mu_\varphi^{(m)} \neq 0$ , then  $\mu_\varphi^{(m)} = 1$ ,

- if  $(g_\varphi)_\ell^{(m)} \neq 0$ , then there exists  $k \in \{1, \dots, K\}$  such that  $(g_\varphi)_\ell^{(m)} = \Upsilon_k$ .

Note that  $|\Phi| = M + KM^2$ . Furthermore, assume from now on that  $(\Upsilon_k)_{k=1, \dots, K}$  is orthonormal in  $\mathbb{L}_2([0, 1])$ . Then  $\Phi$  is also orthonormal in  $\mathcal{H}$  endowed with  $\|\cdot\|$ .

Before going further, let us discuss Assumption (5.4). First note that the matrix  $\Lambda$  is block diagonal. The first block is the identity matrix of size  $M$ . The other  $M^2$  blocks are identical to the matrix:

$$\tilde{\Lambda}_K = \left( \int |\Upsilon_{k_1}(u)| |\Upsilon_{k_2}(u)| du \right)_{1 \leq k_1, k_2 \leq K}.$$

So, if we denote  $\tilde{r}_K$  the spectral radius of  $\tilde{\Lambda}_K$ , we have:

$$r_\Phi = \max(1, \tilde{r}_K).$$

We analyze the behavior of  $\tilde{r}_K$  with respect to  $K$ . Note that for any  $k_1$  and any  $k_2$ ,

$$(\tilde{\Lambda}_K)_{k_1, k_2} \geq 0.$$

Therefore,

$$\tilde{r}_K \leq \sup_{\|x\|_{\ell_1}=1} \|\tilde{\Lambda}_K x\|_{\ell_1} \leq \max_{k_1} \sum_{k_2} (\tilde{\Lambda}_K)_{k_1, k_2}.$$

We now distinguish three types of orthonormal dictionaries (remember that  $M$  is viewed as a constant):

- Let us consider regular histograms. The basis is composed of the functions  $\Upsilon_k = \delta^{-1/2} \mathbb{1}_{((k-1)\delta, k\delta]}$  with  $K\delta = 1$ . Therefore  $\|\Phi\|_\infty = \delta^{-1/2} = \sqrt{K}$ . But  $\tilde{\Lambda}_K$  is the identity matrix and  $\tilde{r}_K = 1$ . Hence (5.4) is satisfied as soon as

$$\frac{K^2 \log(K) \log^5(T)}{T} \rightarrow 0$$

when  $T \rightarrow \infty$ , which is satisfied if  $K = o\left(\frac{\sqrt{T}}{\log^3(T)}\right)$ .

- Assume that  $\|\Phi\|_\infty$  is bounded by an absolute constant (Fourier dictionaries satisfy this assumption). Since  $\tilde{r}_K \leq K$ , (5.4) is satisfied as soon as

$$\frac{K^2 \log(K) \log^5(T)}{T} \rightarrow 0$$

when  $T \rightarrow \infty$ , which is satisfied if  $K = o\left(\frac{\sqrt{T}}{\log^3(T)}\right)$ .

- Assume that  $(\Upsilon_k)_{k=1, \dots, K}$  is a compactly supported wavelet dictionary where resolution levels belong to the set  $\{0, 1, \dots, J\}$ . In this case,  $K$  is of the same order as  $2^J$ ,  $\|\Phi\|_\infty$  is of the same order as  $2^{J/2}$  and it can be seen that  $\tilde{r}_K \leq C2^{J/2}$  where  $C$  is a constant only depending on the choice of the wavelet system (see [29] for further details). Then, (5.4) is satisfied as soon as

$$\frac{K^{5/2} \log(K) \log^5(T)}{T} \rightarrow 0$$

when  $T \rightarrow \infty$ , which is satisfied if  $K = o\left(\frac{T^{2/5}}{\log^{12/5}(T)}\right)$ .

To apply Theorem 2, it remains to control  $\Omega_{V,B}$ . Note that

$$\psi_t^{(m)}(\varphi) = \begin{cases} 1 & \text{if } \mu_\varphi^{(m)} = 1 \\ \int_{t-1}^{t-} \Upsilon_k(t-u) dN_u^{(\ell)} & \text{if } (g_\varphi)_\ell^{(m)} = \Upsilon_k. \end{cases}$$

Let us define

$$\Omega_{\mathcal{N}} = \left\{ \text{for all } t \in [0, T], \text{ for all } m \in \{1, \dots, M\} \text{ we have } N_{[t-1, t]}^{(m)} \leq \mathcal{N} \right\}.$$

We therefore set

$$B_\varphi = 1 \text{ if } \mu_\varphi^{(m)} = 1 \text{ and } B_\varphi = \|\Upsilon_k\|_\infty \mathcal{N} \text{ if } (g_\varphi)_\ell^{(m)} = \Upsilon_k. \quad (5.5)$$

Note that on  $\Omega_{\mathcal{N}}$ , for any  $\varphi \in \Phi$ ,

$$\sup_{t \in [0, T], m} |\psi_t^{(m)}(\varphi)| \leq B_\varphi.$$

Now, for each  $\varphi \in \Phi$ , let us determine  $V_\varphi$  that constitutes an upper bound of

$$M_\varphi = \sum_{m=1}^M \int_0^T [\psi_t^{(m)}(\varphi)]^2 dN_t^{(m)}.$$

Note that only one term in this sum is non-zero.

$$V_\varphi = \lceil T \rceil \mathcal{N} \text{ if } \mu_\varphi^{(m)} = 1 \text{ and } V_\varphi = \|\Upsilon_k\|_\infty^2 \lceil T \rceil \mathcal{N}^3 \text{ if } (g_\varphi)_\ell^{(m)} = \Upsilon_k. \quad (5.6)$$

With this choice of  $B_\varphi$  and  $V_\varphi$ , one has that  $\Omega_{\mathcal{N}} \subset \Omega_{V, B}$ , which leads to the following result.

**Corollary 3.** *Assume that the Hawkes process is stationary, that (5.3) is satisfied and that the spectral radius of  $\Gamma$  is strictly smaller than 1. With the choices (5.5) and (5.6) of the  $B_\varphi$ 's and of the  $V_\varphi$ 's,*

$$\mathbb{P}(\Omega_{V, B}) \geq \mathbb{P}(\Omega_{\mathcal{N}}) \geq 1 - C_1 T \exp(-C_2 \mathcal{N}),$$

where  $C_1$  and  $C_2$  are positive constants depending on  $f^*$ .

If  $\mathcal{N} \gg \log(T)$ , then for all  $\beta > 0$ ,

$$\mathbb{P}(\Omega_{V, B}^c) \leq \mathbb{P}(\Omega_{\mathcal{N}}^c) = o(T^{-\beta}).$$

We are now in position to apply Theorem 2.

**Corollary 4.** *Assume that the Hawkes process is stationary, that (5.3) is satisfied and that the spectral radius of  $\Gamma$  is strictly smaller than 1. Assume that the dictionary  $\Phi$  is built as previously from an orthonormal family  $(\Upsilon_k)_{k=1, \dots, K}$ . With the notations of Theorem 2, let  $B_\varphi$  be defined by (5.5) and  $d_\varphi$  be defined accordingly with  $x = \alpha \log(T)$ . Then, with probability larger than*

$$1 - 4(M + M^2 K) \left( \frac{\log \left( 1 + \frac{\mu \lceil T \rceil \mathcal{N}}{\alpha \log(T)} \right)}{\log(1 + \varepsilon)} + 1 \right) T^{-\alpha} - \mathbb{P}(\Omega_{\mathcal{N}}^c) - \mathbb{P}(\Omega_c^c),$$

$$\frac{1}{T} \|\hat{f} - f^*\|_T^2 \leq C \inf_{a \in \mathbb{R}^\Phi} \left\{ \frac{1}{T} \|f^* - f_a\|_T^2 + \sum_{\varphi \in \mathcal{S}(a)} \left( \frac{\log(T) (\psi(\varphi))^2 \bullet N_T}{T^2} + \frac{B_\varphi^2 \log^2(T)}{T^2} \right) \right\},$$

where  $C$  is a constant depending on  $f^*$ ,  $\mu$ ,  $\varepsilon$ , and  $\alpha$ .

From an asymptotic point of view, if the dictionary also satisfies (5.4), and if  $\mathcal{N} = \log^2(T)$  in (5.5), then for  $T$  large enough with probability larger than  $1 - C_1 K \log(T) T^{-\alpha}$

$$\frac{1}{T} \|\hat{f} - f^*\|_T^2 \leq C_2 \inf_{a \in \mathbb{R}^\Phi} \left\{ \frac{1}{T} \|f^* - f_a\|_T^2 + \frac{\log^3(T)}{T} \sum_{\varphi \in \mathcal{S}(a)} \left[ \frac{1}{T} \|\varphi\|_T^2 + \frac{\log^{7/2}(T)}{\sqrt{T}} \|\Phi\|_\infty^2 \right] \right\},$$

where  $C_1$  and  $C_2$  are constants depending on  $M$ ,  $f^*$ ,  $\mu$ ,  $\varepsilon$ , and  $\alpha$ .

We express the oracle inequality by using  $\frac{1}{T} \|\cdot\|_T$  simply because, when  $T$  goes to  $+\infty$ , by ergodicity of the process (see for instance [21], and Proposition 2 for a non asymptotic statement)

$$\frac{1}{T} \|f\|_T^2 = \sum_{m=1}^M \frac{1}{T} \int_0^T (\kappa_t(\mathbf{f}^{(m)}))^2 dt \longrightarrow \sum_{m=1}^M Q(\mathbf{f}^{(m)}, \mathbf{f}^{(m)})$$



under assumptions of Proposition 4. Note that the right hand side is a true norm on  $\mathcal{H}$  by Proposition 3. Note also that

$$\frac{\log^{7/2}(T)}{\sqrt{T}} \|\Phi\|_\infty^2 \xrightarrow{T \rightarrow \infty} 0,$$

as soon as (5.4) is satisfied for Fourier and compactly supported wavelets. It is also the case for histograms as soon as  $K = o\left(\frac{\sqrt{T}}{\log^{7/2}(T)}\right)$ . Therefore, with respect to the previous remark, this term should be considered as a residual one. In those cases, the last inequality can be rewritten as

$$\frac{1}{T} \|\hat{f} - f^*\|_T^2 \leq C \inf_{a \in \mathbb{R}^\Phi} \left\{ \frac{1}{T} \|f^* - f_a\|_T^2 + \frac{\log^3(T)}{T} \sum_{\varphi \in \mathcal{S}(a)} \frac{1}{T} \|\varphi\|_T^2 \right\},$$

for a different constant  $C$ , the probability of this event tending to 1 as soon as  $\alpha \geq 1/2$  in the Fourier and histogram cases and  $\alpha \geq 2/5$  in the compactly supported wavelet basis. Once again, as mentioned for the Poisson or Aalen models, the right hand side corresponds to a classical "bias-variance" trade off and a classical shape of oracle inequality up to the logarithmic terms. Note that this time, the asymptotic is done in  $T$  and not in  $M$ , as for Poisson or Aalen models but the same result, namely Theorem 2, is capable, depending on the framework, to lead to both potential asymptotics.

## 6 Simulations for the multivariate Hawkes process

This section is devoted to illustrations of our procedure on simulated data of multivariate Hawkes processes and comparisons with the well-known adaptive Lasso procedure proposed by [57].

### 6.1 Description of the Data

As mentioned in the introduction, Hawkes processes can be used in Neuroscience to model the Unitary Event Activity of individual neurons (see [27]). So, we perform simulations whose parameters are close, to some extent, to real neuronal data. For a given neuron  $m \in \{1, \dots, M\}$ , its activity is modeled by a point process  $N^{(m)}$  whose intensity is

$$\lambda_t^{(m)} = \nu^{(m)} + \sum_{\ell=1}^M \int_{-\infty}^{t-} h_\ell^{(m)}(t-u) dN^{(\ell)}(u).$$

The *interaction function*  $h_\ell^{(m)}$  represents the influence of the past activity of the neuron  $\ell$  on the neuron  $m$ . The *spontaneous rate*  $\nu^{(m)}$  may somehow represent the external excitation linked to all the other neurons that are not recorded. It is consequently of crucial importance not only to correctly infer the interaction functions, but also to reconstruct the spontaneous rates accurately. Usually, activity up to 10 neurons can be recorded in a "stationary" phase during a few seconds (sometimes up to one minute). Typically, the points frequency is of the order of 10-80 Hz and the interaction range between points is of the order of a few milliseconds (up to 20 or 40 ms). We lead three experiments by simulating multivariate Hawkes processes (two with  $M = 2$ , one with  $M = 8$ ) based on these typical values. More precisely, for all experiments, we take for any  $m \in \{1, \dots, M\}$ ,  $\nu^{(m)} = 20$  and the interaction functions  $h_\ell^{(m)}$  are defined as follows (supports of all the functions are assumed to lie in the interval  $[0, 0.04]$ ):

- **Experiment 1:  $M = 2$  and piecewise constant functions.**

$$h_1^{(1)} = 30 \times \mathbb{1}_{(0,0.02]}, \quad h_2^{(1)} = 30 \times \mathbb{1}_{(0,0.01]}, \quad h_1^{(2)} = 30 \times \mathbb{1}_{(0.01,0.02]}, \quad h_2^{(2)} = 0.$$

In this case, each neuron depends on the other one. The spectral radius of the matrix  $\Gamma$  is 0.725.

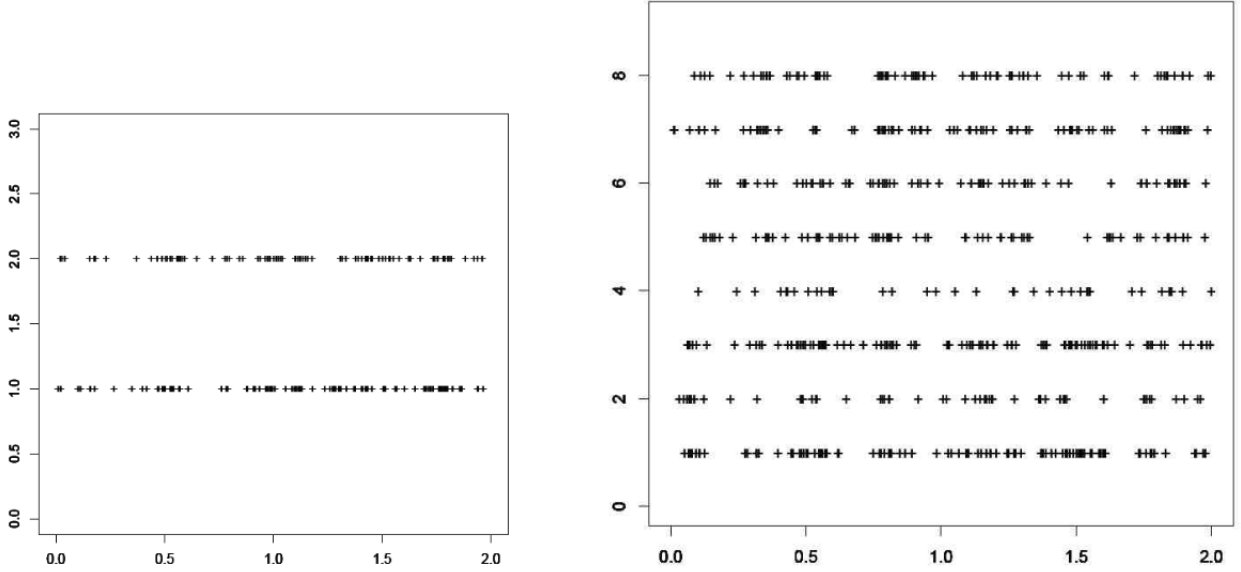


Figure 1: Raster plots of two data sets with  $T = 2$  corresponding to Experiment 2 on the left and Experiment 3 on the right. The  $x$ -axis correspond to the time of the experiment. Each line with ordinate  $m$  corresponds to the points of the process  $N^{(m)}$ . From bottom to top, we observe 124 and 103 points for Experiment 2 and 101, 60, 117, 38, 73, 75, 86 and 86 points for Experiment 3.

- **Experiment 2:  $M = 2$  and "smooth" functions.** In this experiment,  $h_1^{(1)}$  and  $h_1^{(2)}$  are not piecewise constant.

$$h_1^{(1)}(x) = 100 e^{-200x} \times \mathbb{1}_{(0,0.04]}(x), \quad h_2^{(1)}(x) = 30 \times \mathbb{1}_{(0,0.02]}(x),$$

$$h_1^{(2)}(x) = \frac{1}{0.008\sqrt{2\pi}} e^{-\frac{(x-0.02)^2}{2 \cdot 0.004^2}} \times \mathbb{1}_{(0,0.04]}(x), \quad h_2^{(2)}(x) = 0.$$

In this case, each neuron depends on the other one as well. The spectral radius of the matrix  $\Gamma$  is 0.711.

- **Experiment 3:  $M = 8$  and piecewise constant functions.**

$$h_2^{(1)} = h_3^{(1)} = h_2^{(2)} = h_1^{(3)} = h_2^{(3)} = h_8^{(5)} = h_5^{(6)} = h_6^{(7)} = h_7^{(8)} = 25 \times \mathbb{1}_{(0,0.02]}$$

and all the other 55 interaction functions are equal to 0. Note in particular that this leads to 3 independent groups of dependent neurons  $\{1, 2, 3\}$ ,  $\{4\}$  and  $\{5, 6, 7, 8\}$ . The spectral radius of the matrix  $\Gamma$  is 0.5.

In all the simulations, we let the process "warm up" during 1 second to reach the stationary state<sup>1</sup>. Then the data are collected by taking records during the next  $T$  seconds. For instance, this leads to roughly about 100 points per neuron when  $T = 2$  and 1000 points when  $T = 20$ . Figure 1 shows two instances of data sets with  $T = 2$ .

## 6.2 Description of the methods

To avoid approximation errors by computing the matrix  $G$ , we focus on a dictionary  $(\Upsilon_k)_{k=1,\dots,K}$  whose functions are piecewise constant. More precisely, we take  $\Upsilon_k = \delta^{-1/2} \mathbb{1}_{((k-1)\delta, k\delta]}$  with  $\delta = 0.04/K$  and  $K$ , the size of the

<sup>1</sup>Note that since the size of the support of the interaction functions is less or equal to 0.04, the "warm up" period is 25 times the interaction range.

dictionary, is chosen later.

Our practical procedure strongly relies on the theoretical one with the natural choice  $x$  in (2.7) of the form  $x = \alpha \log(T)$ . Three hyperparameters (namely  $\alpha$ ,  $\mu$  and  $\varepsilon$ ) would need to be tuned if we directly used the proposed Lasso parameters of Theorem 2 (see also Corollary 4). So, for simplifications, we implement our procedure by replacing the Lasso parameters  $d_\varphi$  given in (2.7) with

$$\tilde{d}_\varphi(\gamma) = \sqrt{2\gamma \log(T)(\psi(\varphi))^2 \bullet N_T} + \frac{\gamma \log(T)}{3} \sup_{t \in [0, T], m} |\psi_t^{(m)}(\varphi)|,$$

where  $\gamma$  is a constant to be tuned. Besides taking  $x = \alpha \log(T)$ , our modification consists in neglecting the linear part  $\frac{B_\varphi^2 x}{\mu - \phi(\mu)}$  in  $\hat{V}^\mu$  and replacing  $B_\varphi$  with  $\sup_{t \in [0, T], m} |\psi_t^{(m)}(\varphi)|$ . Then, note that, up to these modifications, the choice  $\gamma = 1$  corresponds to the limit case where  $\alpha \rightarrow 1$ ,  $\varepsilon \rightarrow 0$  and  $\mu \rightarrow 0$  in the definition of the  $d_\varphi$ 's (see the comments after Theorem 2). Note also that, under the slight abuse consisting in identifying  $B_\varphi$  with  $\sup_{t \in [0, T], m} |\psi_t^{(m)}(\varphi)|$ , for every parameter  $\mu$ ,  $\varepsilon$  and  $\alpha$  of Theorem 2 with  $x = \alpha \ln(T)$ , one can find two parameters  $\gamma$  and  $\gamma'$  such that

$$\tilde{d}_\varphi(\gamma) \leq d_\varphi \leq \tilde{d}_\varphi(\gamma').$$

Therefore, this practical choice is consistent with the theory and tuning hyperparameters reduces to only tuning  $\gamma$ .

We compute the Lasso estimate by using the shooting method of [25] and the R-package `Lassoshooting`. Note in particular that to do so, we need to invert the matrix  $G$ . In all simulations, this matrix has always been invertible, which is consistent with the fact that  $\Omega_c$  happens with large probability. Note also that the value of  $c$ , namely the smallest eigenvalue of  $G$ , can be very small (about  $10^{-4}$ ) whereas the largest eigenvalue is potentially as large as  $10^5$ , both values highly depending on the simulation and on  $T$ . Fortunately, those values are not needed to compute our Lasso estimate. Since it is based on *Bernstein type inequalities*, our Lasso method is denoted **B** in the sequel.

Due to their soft thresholding nature, Lasso methods are known to underestimate the coefficients [37, 57]. To overcome biases in estimation due to shrinkage, we propose a two steps procedure, as usually suggested in the literature: Once the support of the vector has been estimated by **B**, we compute the ordinary least-square estimator among the vectors  $a$  having the same support, which provides the final estimate. This method is denoted **BO** in the sequel.

Another popular method is *adaptive Lasso* proposed by Zou [57]. This method overcomes the flaws of standard Lasso by taking  $\ell_1$ -weights of the form

$$d_\varphi^a(\gamma) = \frac{\gamma}{2|\hat{a}_\varphi^o|^p},$$

where  $p > 0$ ,  $\gamma > 0$  and  $\hat{a}_\varphi^o$  is a preliminary consistent estimate of the true coefficient. Even if the shape of the weights are different, the latter are data-driven and this method constitutes a natural competitive method with ours. The most usual choice, which is adopted in the sequel, consists in taking  $p = 1$  and the ordinary least squares estimate for the preliminary estimate (see [31, 55, 57]). Then, penalization is stronger for coefficients that are preliminary estimated by small values of the ordinary least square estimate. In the literature, the parameter  $\gamma$  of adaptive Lasso is usually tuned by cross-validation, but this does not make sense for Hawkes data that are fully dependent. Therefore, a preliminary study has been performed to provide meaningful values for  $\gamma$ . Results are given in the next section. This adaptive Lasso method is denoted **A** in the sequel and **AO** when combined with ordinary least squares in the same way as for **BO**.

Simulations are performed in R. The computational time is weak (merely few seconds for one estimate even when  $M = 8$ ,  $T = 20$  and  $K = 8$  on a classical laptop computer), which constitutes a clear improvement with respect to existing adaptive methods for Hawkes processes. For instance, the "Islands" method<sup>2</sup> of [46] was limited due to extreme computational time for estimating one or two dozens of coefficients at the most whereas here when  $M = 8$  and  $K = 8$ , we have to deal with  $M + KM^2 = 520$  coefficients.

<sup>2</sup>This method developed for  $M = 1$  could easily be theoretically adapted for larger values of  $M$ , but its extreme computational cost prevents us from using it in practice.

M=2, T=2	Our Lasso Method			Adaptive Lasso			M=2, T=20	Our Lasso Method			Adaptive Lasso		
$\gamma$	0.5	1	2	2	200	1000	$\gamma$	0.5	1	2	2	200	1000
DG	<b>100</b>	<b>100</b>	98	<b>100</b>	<b>100</b>	98	DG	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
S	*	*	*	2	2	1	S	*	*	*	*	*	*
F+	<b>0</b>	<b>0</b>	<b>0</b>	1	<b>0</b>	<b>0</b>	F+	<b>0</b>	<b>0</b>	<b>0</b>	1	<b>0</b>	<b>0</b>
F-	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	F-	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
Coeff+	2	1	<b>0</b>	11	2	<b>0</b>	Coeff+	1	<b>0</b>	<b>0</b>	11	2	<b>0</b>
Coeff-	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	Coeff-	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
SpontMSE	108	140	214	150	193	564	SpontMSE	22	37	69	14	12	27
	+ols	104	96	95	151	154		516	+ols	11	10	9	14
InterMSE	7	9	15	13	8	11	InterMSE	2	3	6	1.4	0.6	0.5
	+ols	7	7	7	14	10		10	+ols	0.6	0.5	0.4	1.4

M=8, T=2	Our Lasso Method			Adaptive Lasso			M=8, T=20	Our Lasso Method			Adaptive Lasso		
$\gamma$	0.5	1	2	2	200	1000	$\gamma$	0.5	1	2	2	200	1000
DG	0	<b>32</b>	24	0	0	<b>32</b>	DG	63	99	<b>100</b>	0	0	90
S	*	*	*	8	7	5	S	*	*	*	*	*	*
F+	17	6	1	55	13	<b>0.5</b>	F+	3	1	<b>0</b>	55	10	<b>0</b>
F-	<b>0</b>	<b>0</b>	2	<b>0</b>	<b>0</b>	2	F-	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
Coeff+	22	7	<b>1</b>	199.5	17	<b>1</b>	Coeff+	4	1	<b>0</b>	197	13	<b>0</b>
Coeff-	0.5	2	7	<b>0</b>	2	7	Coeff-	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
SpontMSE	295	428	768	1445	1026	1835	SpontMSE	82	166	355	104	43	64
	+ols	1327	587	859	1512	1058		1935	+ols	41	26	24	107
InterMSE	38	51	79	214	49	65	InterMSE	10	19	39	16	2.9	3.17
	+ols	63	45	61	228	84		70	+ols	3	2.1	1.9	17

Table 1: Numerical results of both procedures over 100 runs with  $K = 4$ . Results for Experiment 1 (top) and Experiment 3 (bottom) are given for  $T = 2$  (left) and  $T = 20$  (right). "DG" gives the number of correct identifications of dependency groups over 100 runs. "S" gives the median number of non-zero spontaneous rate estimates, "\*" means that all the spontaneous rate estimates are non-zero over all the simulations. "F+" gives the median number of additional non-zero interaction functions w.r.t. the truth. "F-" gives the median number of missing non-zero interaction functions w.r.t. the truth. "Coeff+" and "Coeff-" are defined in the same way for the coefficients. "SpontMSE" is the Mean Square Error for the spontaneous rates with or without the additional "ordinary least squares step". "InterMSE" is the analog for the interaction functions. In red, we give the optimal values for the qualitative criteria.

### 6.3 Results

A study over 100 simulations has been carried out corresponding to Experiments 1 and 3 for which we can precisely check if the support of the vector  $\hat{a}$  is the correct one. Results for our method and for adaptive Lasso can be found in Table 1. For each method, we have selected 3 values for the hyperparameter  $\gamma$  based on results of preliminary simulations. Two types of criterion are discussed: qualitative ones based on supports recovery and quantitative ones based on Mean Square Errors.

Let us first review the qualitative ones. The first main purpose of the method is to correctly guess the dependency groups, which is essential from the neurobiological point of view since knowing interactions between two neurons is of capital importance. So the line "DG", which gives the number of correct identifications of dependency groups, is very relevant. For instance, for  $M = 8$ , "DG" gives the number of simulations for which the 3 dependency groups  $\{1, 2, 3\}$ ,  $\{4\}$  and  $\{5, 6, 7, 8\}$  are recovered by the methods. When  $M = 2$ , both methods correctly find that neurons 1 and 2 are dependent, even if  $T = 2$ . When 8 neurons are considered, the estimates should find 3 dependency groups. We see that even with  $T = 2$ , our method with  $\gamma = 1$  correctly guesses the dependency groups for 32% of the simulations. It's close or equal to 100% when  $T = 20$  with  $\gamma = 1$  or  $\gamma = 2$ . The adaptive Lasso has to take  $\gamma = 1000$  for  $T = 2$  and  $T = 20$  to obtain as convincing results.

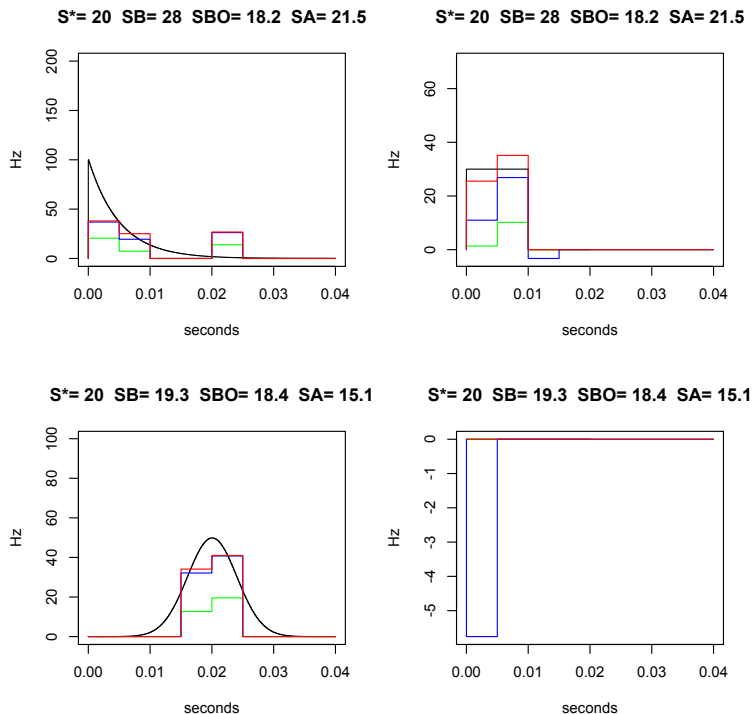


Figure 2: Reconstructions corresponding to Experiment 2 with  $T = 2$  and  $K = 8$ . Each line  $m$  represents the function  $h_\ell^{(m)}$ , for  $\ell = 1, 2$ . The spontaneous rates estimation associated with each line  $m$  is given above the graphs:  $S^*$  denotes the true spontaneous rate and its estimators computed by using **B**, **BO** and **A** respectively are denoted by SB, SBO and SA. The true interactions functions (in black) are reconstructed by using **B**, **BO** and **A** providing reconstructions in green, red and blue respectively. We use  $\gamma = 1$  for **B** and **BO** and  $\gamma = 200$  for **A**.

Clearly, smaller choices of  $\gamma$  for adaptive Lasso leads to bad estimations of the dependency groups. Next, the main point is to see whether the methods are able to guess the correct number of non-zero spontaneous rates. Whatever the experiment and the parameter  $\gamma$ , our method is optimal whereas adaptive Lasso misses some non-zero spontaneous rates when  $T = 2$ . Under this criterion, for adaptive Lasso, the choice  $\gamma = 1000$  is clearly bad when  $T = 2$  (the optimal value of  $S$  is  $S = 2$  when  $M = 2$  and  $S = 8$  when  $M = 8$ ) on both experiments, whereas  $\gamma = 2$  or  $\gamma = 200$  is better. Not surprisingly, the number of additional non-zero functions and additional non-zero coefficients decreases when  $T$  grows and when  $\gamma$  grows, whatever the method whereas the number of missing functions or coefficients increases. We can conclude from these facts and from further analysis of Table 1 that the choice  $\gamma = 0.5$  for our method and the choice  $\gamma = 2$  for the adaptive Lasso are wrong choices of the tuning parameters. In conclusion of the qualitative aspects, our method with  $\gamma = 1$  or  $\gamma = 2$  seems a good choice and is robust with respect to  $T$ . When  $T = 20$ , the optimal choice for adaptive Lasso is  $\gamma = 1000$ . When  $T = 2$ , the choice is not so clear and depends on the qualitative criterion we wish to favor.

Now let us look at some more quantitative criteria. Since the spontaneous rates do not behave like the other coefficients, we split the Mean Square Error in two parts: one for the spontaneous rates:

$$\text{SpontMSE} = \sum_{m=1}^M (\hat{\nu}^{(m)} - \nu^{(m)})^2,$$

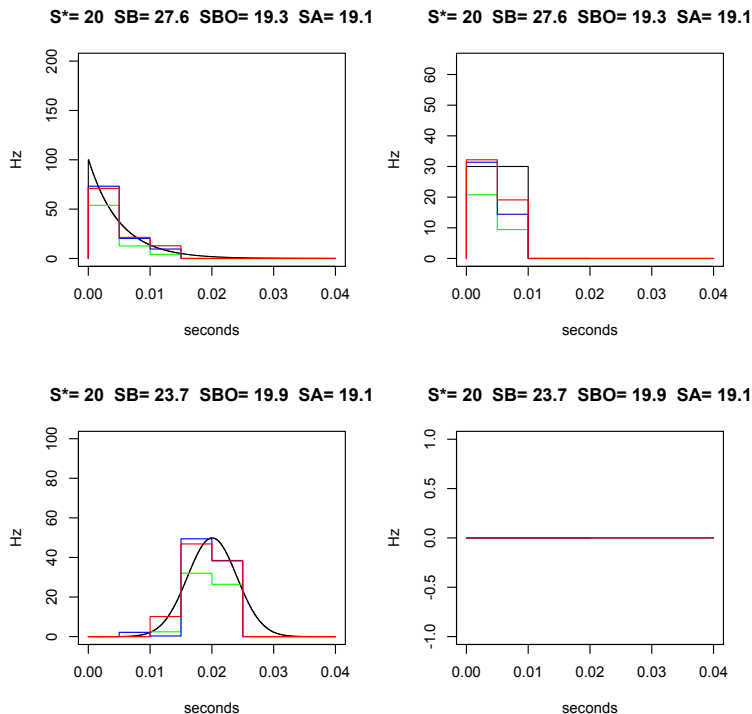


Figure 3: Reconstructions corresponding to Experiment 2 with  $T = 20$  and  $K = 8$ . Each line  $m$  represents the function  $h_\ell^{(m)}$ , for  $\ell = 1, 2$ . The spontaneous rates estimation associated with each line  $m$  is given above the graphs:  $S^*$  denotes the true spontaneous rate and its estimators computed by using **B**, **BO** and **A** respectively are denoted by SB, SBO and SA. The true interactions functions (in black) are reconstructed by using **B**, **BO** and **A** providing reconstructions in green, red and blue respectively. We use  $\gamma = 1$  for **B** and **BO** and  $\gamma = 1000$  for **A**.

and one for interactions:

$$\text{InterMSE} = \sum_{m=1}^M \sum_{\ell=1}^M \int (\hat{h}_\ell^{(m)}(t) - h_\ell^{(m)}(t))^2 dt.$$

We report the results for **B**, **BO**, **A** and **AO**. We mostly focus on cases where supports are correctly estimated. In this case, results are better by using the second step. MSE are increasing with  $\gamma$  for **B** and **A**, since underestimation is stronger when  $\gamma$  increases. This phenomenon does not appear for two steps procedures, which leads to more stable MSE when the support is correct. One of the main differences between both methods can be seen by analyzing SpontMSE. Since adaptive Lasso does not detect all non-zero spontaneous rates, the corresponding MSE cannot be good and this cannot be improved via the OLS transformation. This comforts us in the fact that the choice  $\gamma = 1000$  is a wrong choice for  $T = 2$  and adaptive Lasso. The choice  $\gamma = 200$  leads to good MSE, but the MSE are smaller for **BO** with  $\gamma = 1$ . When  $T = 20$ , the choice  $\gamma = 1000$  for **AO** leads to results that are of the same magnitude as the ones obtained by **BO** with  $\gamma = 1$  or 2. Still for  $T = 20$ , results for the estimate **B** are worse than results for **A**. It is due to the fact that the shrinkage is larger in our method for the coefficients we want to keep than shrinkage of adaptive Lasso that becomes negligible as soon as the true coefficients are large enough. However the second step overcomes this problem.

Note also that a more thorough study of the tuning parameter  $\gamma$  has been performed by [4] who mathematically prove that the choice  $\gamma < 1$  leads to very degenerate estimates in the density setting. Their method for choosing Lasso parameters being analogous to ours, it seems coherent to obtain worse MSE for  $\gamma = 0.5$  than

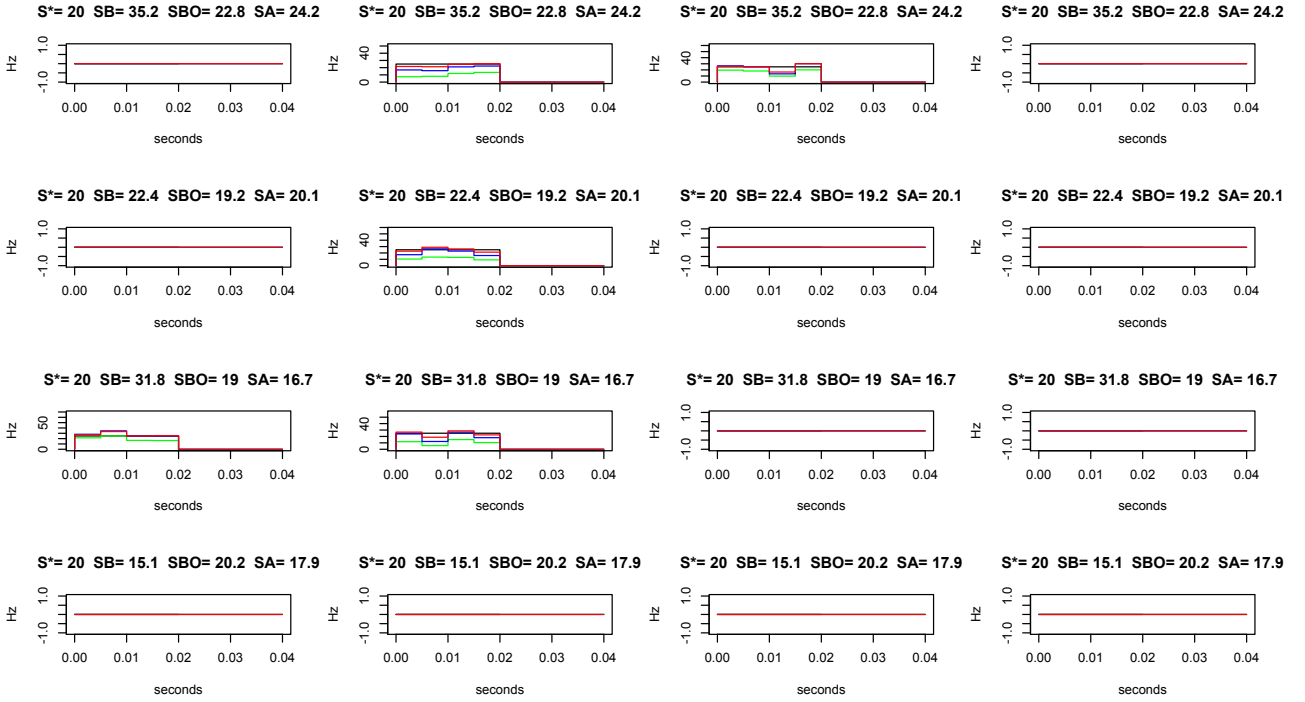


Figure 4: Reconstructions corresponding to Experiment 3 with  $T = 20$  and  $K = 8$  and for the first 4 neurons. Each line  $m$  represents the function  $h_\ell^{(m)}$ , for  $\ell = 1, 2$ . The spontaneous rates associated with each line  $m$  are given above the graphs where  $S^*$  denotes the true spontaneous rate and its estimators computed by using **B**, **BO** and **A** respectively and denoted by SB, SBO and SA. The true interactions functions (in black) are reconstructed by using **B**, **BO** and **A** providing reconstructions in green, red and blue respectively. We use  $\gamma = 1$  for **B** and **BO** and  $\gamma = 1000$  for **A**.

for  $\gamma = 1$  or  $\gamma = 2$ , at least for **BO**. The boundary  $\gamma = 1$  in their simulation study seems to be a robust choice, and it seems to be the case here too.

We now provide some reconstructions. Figures 2 and 3 give the reconstructions corresponding to Experiment 2 ( $M = 2$ ) with  $K = 8$  for  $T = 2$  and  $T = 20$  respectively. The reconstructions are quite satisfying. Of course, the quality improves when  $T$  grows. We also note improvements by using **BO** instead of **B**. For adaptive Lasso, improvements by using the second step are not significant and this is the reason why we do not represent reconstructions with **AO**. Graphs of the right hand side of Figure 2 illustrate the difficulties of adaptive Lasso to recover the exact support of interactions functions, namely  $h_2^{(1)}$  and  $h_2^{(2)}$  for  $T = 2$ . Figure 4 provides another illustration in the case of Experiment 3 ( $M = 8$ ) with  $K = 8$  for  $T = 20$ . For the sake of clarity, we only represent reconstructions for the first 4 neurons. Supports of coefficients are well recovered by all the methods. From the estimation point of view, this illustration provides a clear hierarchy between the methods: **BO** seems to achieve the best results and **B** the worst.

## 6.4 Conclusions

With respect to the problem of tuning our methodology based on Bernstein type inequalities, our simulation study is coherent with theoretical aspects since we achieve our best results by taking  $\gamma = 1$ , which constitutes the limit case of assumptions of Theorem 2. For practical aspects, we recommend the choice  $\gamma = 1$  even if  $\gamma = 2$  is acceptable. More important, this choice is robust with respect to the duration of records, which is not the

case for adaptive Lasso. Implemented with  $\gamma = 1$ , our method outperforms adaptive Lasso for supports recovery since it is able to recover the dependency groups, the non-zero spontaneous rates, the non-zero functions and even the non-zero coefficients as soon as  $T$  is large enough. Most of the time, the two step procedure **BO** seems to achieves the best results for parameter estimations.

It is important to note that the question of tuning adaptive Lasso remains open. Some values of  $\gamma$  allow us to obtain very good results but they are not robust with respect to  $T$ , which may constitute a serious problem for practitioners. In the standard regression setting, this problem may be overcome by using cross-validation on independent data, which somehow estimates random fluctuations. But in this multivariate Hawkes set-up, independence assumptions on data cannot be made and this explains the problems for tuning adaptive Lasso. Our method based on Bernstein type concentration inequalities take into account those fluctuations. It also takes into account the nature of the coefficients and the variability of their estimates which differ for spontaneous rates on the one hand and coefficients of interaction functions on the other hand. The shape of weights of adaptive Lasso does not incorporate this difference, which explains the contradictions for tuning the method when  $T = 2$ . For instance, in some cases, adaptive Lasso tends to estimate some spontaneous rate by zero in order to achieve better performances on the interaction functions.

## 7 Proofs

This section is devoted to the proofs of the results of the paper. Throughout,  $C$  is a constant whose value may change from line to line.

### 7.1 Proof of Theorem 1

We use  $\|\cdot\|_{\ell^2}$  for the Euclidian norm of  $\mathbb{R}^\Phi$ . Given  $a$  recall that

$$f_a = \sum_{\varphi \in \Phi} a_\varphi \varphi.$$

Then, we have  $\hat{f} = f_{\hat{a}}$ ,

$$a'b = \psi(f_a) \bullet N_T$$

and

$$a'Ga = \|f_a\|_T^2.$$

Then,

$$-2\psi(f_{\hat{a}}) \bullet N_T + \|f_{\hat{a}}\|_T^2 + 2d'|\hat{a}| \leq -2\psi(f_a) \bullet N_T + \|f_a\|_T^2 + 2d'|a|.$$

So,

$$\begin{aligned} \|f_{\hat{a}} - f^*\|_T^2 &= \|f_{\hat{a}}\|_T^2 + \|f^*\|_T^2 - 2 \langle f_{\hat{a}}, f^* \rangle_T \\ &\leq \|f_a\|_T^2 + \|f^*\|_T^2 + 2\psi(f_{\hat{a}} - f_a) \bullet N_T + 2d'(|a| - |\hat{a}|) - 2 \langle f_{\hat{a}}, f^* \rangle_T \\ &= \|f_a - f^*\|_T^2 + 2 \langle f_a - f_{\hat{a}}, f^* \rangle_T + 2\psi(f_{\hat{a}} - f_a) \bullet N_T + 2d'(|a| - |\hat{a}|) \\ &= \|f_a - f^*\|_T^2 + 2\psi(f_a - f_{\hat{a}}) \bullet (\Psi(f^*) - N)_T + 2d'(|a| - |\hat{a}|) \\ &= \|f_a - f^*\|_T^2 + 2 \sum_{\varphi \in \Phi} (a_\varphi - \hat{a}_\varphi) \psi(\varphi) \bullet (\Psi(f^*) - N)_T + 2d'(|a| - |\hat{a}|) \\ &\leq \|f_a - f^*\|_T^2 + 2 \sum_{\varphi \in \Phi} |a_\varphi - \hat{a}_\varphi| \times |\bar{b}_\varphi - b_\varphi| + 2d'(|a| - |\hat{a}|). \end{aligned}$$

Using (2.5), we obtain:

$$\begin{aligned} \|f_{\hat{a}} - f^*\|_T^2 &\leq \|f_a - f^*\|_T^2 + 2 \sum_{\varphi \in \Phi} d_\varphi |a_\varphi - \hat{a}_\varphi| + 2 \sum_{\varphi \in \Phi} d_\varphi (|a_\varphi| - |\hat{a}_\varphi|) \\ &\leq \|f_a - f^*\|_T^2 + 2 \sum_{\varphi \in \Phi} d_\varphi (|a_\varphi - \hat{a}_\varphi| + |a_\varphi| - |\hat{a}_\varphi|). \end{aligned}$$



Now, if  $\varphi \notin S(a)$ ,  $|a_\varphi - \hat{a}_\varphi| + |a_\varphi| - |\hat{a}_\varphi| = 0$ , and

$$\begin{aligned} \|f_{\hat{a}} - f^*\|_T^2 &\leq \|f_a - f^*\|_T^2 + 2 \sum_{\varphi \in S(a)} d_\varphi (|a_\varphi - \hat{a}_\varphi| + |a_\varphi| - |\hat{a}_\varphi|) \\ &\leq \|f_a - f^*\|_T^2 + 4 \sum_{\varphi \in S(a)} d_\varphi (|a_\varphi - \hat{a}_\varphi|) \\ &\leq \|f_a - f^*\|_T^2 + 4\|\hat{a} - a\|_{\ell_2} \left( \sum_{\varphi \in S(a)} d_\varphi^2 \right)^{1/2}. \end{aligned}$$

We now use the assumption on the Gram matrix given by (2.4) and the triangular inequality for  $\|\cdot\|_T$ , which yields

$$\begin{aligned} \|\hat{a} - a\|_{\ell_2}^2 &\leq c^{-1} (\hat{a} - a)' G (\hat{a} - a) \\ &= c^{-1} \|f_{\hat{a}} - f_a\|_T^2 \\ &\leq 2c^{-1} (\|f_{\hat{a}} - f^*\|_T^2 + \|f_a - f^*\|_T^2). \end{aligned}$$

Let us take  $\alpha \in (0; 1)$ . Since for any  $x \in \mathbb{R}$  and any  $y \in \mathbb{R}$ ,  $2xy \leq \alpha x^2 + \alpha^{-1}y^2$ , we obtain:

$$\begin{aligned} \|f_{\hat{a}} - f^*\|_T^2 &\leq \|f_a - f^*\|_T^2 + 4\sqrt{2}c^{-1/2} \sqrt{\|f_{\hat{a}} - f^*\|_T^2 + \|f_a - f^*\|_T^2} \left( \sum_{\varphi \in S(a)} d_\varphi^2 \right)^{1/2} \\ &\leq \|f_a - f^*\|_T^2 + \alpha (\|f_{\hat{a}} - f^*\|_T^2 + \|f_a - f^*\|_T^2) + 8\alpha^{-1}c^{-1} \sum_{\varphi \in S(a)} d_\varphi^2 \\ &\leq (1 - \alpha)^{-1} \left( (1 + \alpha)\|f_a - f^*\|_T^2 + 8\alpha^{-1}c^{-1} \sum_{\varphi \in S(a)} d_\varphi^2 \right). \end{aligned}$$

The theorem is proved just by taking an arbitrary absolute value for  $\alpha \in (0; 1)$ .

## 7.2 Proof of Theorem 2

Let us first define

$$\mathcal{T} = \{t \geq 0 \mid \sup_m |\psi_t^{(m)}(\varphi)| > B_\varphi\}. \quad (7.1)$$

Let us define the stopping time  $\tau' = \inf \mathcal{T}$  and the predictable process  $H$  by

$$H_t^{(m)} = \psi_t^{(m)}(\varphi) \mathbb{1}_{t \leq \tau'}.$$

Let us apply Theorem 3 to this choice of  $H$  with  $\tau = T$  and  $B = B_\varphi$ . The choice of  $v$  and  $w$  will be given later on. To apply this result, we need to check that for all  $t$  and all  $\xi \in (0, 3)$ ,  $\sum_m \int_0^t e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} ds$  is a.s. finite. But if  $t > \tau'$ , then

$$\int_0^t e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} ds = \int_0^{\tau'} e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} ds + \int_{\tau'}^t \lambda_s^{(m)} ds,$$

where the second part is obviously finite (it is just  $\Lambda_t^{(m)} - \Lambda_{\tau'}^{(m)}$ ). Hence it remains to prove that for all  $t \leq \tau'$ ,

$$\int_0^t e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} ds$$

is finite. But for all  $s < t$ ,  $s < \tau'$  and consequently  $s \notin \mathcal{T}$ . Therefore  $|H_s^{(m)}| \leq B_\varphi$ . Since we are integrating with respect to the Lebesgue measure, the fact that it eventually does not hold in  $t$  is not a problem and

$$\int_0^t e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} ds \leq e^{\xi} \Lambda_t^{(m)},$$

which is obviously finite a.s. The same reasoning can be applied to show that a.s.  $\exp(\xi H^2/B^2) \bullet \Lambda_t < \infty$ . We can also apply Theorem 3 to  $-H$  in the same way. We obtain at the end that for all  $\varepsilon > 0$

$$\mathbb{P} \left( |H \bullet (N - \Lambda)_T| \geq \sqrt{2(1 + \varepsilon) \hat{V}^\mu x} + \frac{B_\varphi x}{3} \text{ and } w \leq \hat{V}^\mu \leq v \text{ and } \sup_{m, t \leq T} |H_t^{(m)}| \leq B_\varphi \right) \leq 4 \left( \frac{\log(v/w)}{\log(1 + \varepsilon)} + 1 \right) e^{-x}. \quad (7.2)$$

But on  $\Omega_{V,B}$  it is clear that  $\forall t \in [0, T], t \notin \mathcal{T}$ . Therefore  $\tau' \geq T$ . Therefore for all  $t \leq T$ , one also has  $t \leq \tau'$  and  $H_t^{(m)} = \psi_t^{(m)}(\varphi)$ . Consequently, on  $\Omega_{V,B}$ ,

$$H \bullet (N - \Lambda)_T = b_\varphi - \bar{b}_\varphi \text{ and } \hat{V}^\mu = \hat{V}_\varphi^\mu.$$

Moreover, on  $\Omega_{V,B}$ , one has that

$$\frac{B_\varphi^2 x}{\mu - \phi(\mu)} \leq \hat{V}_\varphi^\mu \leq \frac{\mu}{\mu - \phi(\mu)} V_\varphi + \frac{B_\varphi^2 x}{\mu - \phi(\mu)}.$$

So, we take  $w$  and  $v$  as respectively the left and right hand side of the previous inequality. Finally note that on  $\Omega_{V,B}$ ,

$$\sup_{m, t \leq T} |H_t^{(m)}| = \sup_{m, t \leq T} |\psi_t^{(m)}(\varphi)| \leq B_\varphi.$$

Hence, we can rewrite (7.2) as follows

$$\mathbb{P} \left( |b_\varphi - \bar{b}_\varphi| \geq \sqrt{2(1 + \varepsilon) \hat{V}_\varphi^\mu x} + \frac{B_\varphi x}{3} \text{ and } \Omega_{V,B} \right) \leq 4 \left( \frac{\log \left( 1 + \frac{\mu V_\varphi}{B_\varphi^2 x} \right)}{\log(1 + \varepsilon)} + 1 \right) e^{-x}. \quad (7.3)$$

Apply this to all  $\varphi \in \Phi$ , we obtain that

$$\mathbb{P} (\exists \varphi \in \Phi \text{ s.t. } |b_\varphi - \bar{b}_\varphi| \geq d_\varphi \text{ and } \Omega_{V,B}) \leq 4 \sum_{\varphi \in \Phi} \left( \frac{\log \left( 1 + \frac{\mu V_\varphi}{B_\varphi^2 x} \right)}{\log(1 + \varepsilon)} + 1 \right) e^{-x}.$$

Now on the event  $\Omega_c \cap \Omega_{V,B} \cap \{\forall \varphi \in \Phi, |b_\varphi - \bar{b}_\varphi| \leq d_\varphi\}$ , one can apply Theorem 1. To obtain Theorem 2, it remains to bound the probability of the complementary event by

$$\mathbb{P}(\Omega_c^c) + \mathbb{P}(\Omega_{V,B}^c) + \mathbb{P}(\exists \varphi \in \Phi \text{ s.t. } |b_\varphi - \bar{b}_\varphi| \geq d_\varphi \text{ and } \Omega_{V,B}).$$

### 7.3 Proof of Theorem 3

First, replacing  $H$  with  $H/B$ , we can always assume that  $B = 1$ .

Next, let us fix for the moment  $\xi \in (0, 3)$ . If one assumes that almost surely for all  $t > 0$ ,  $\sum_{m=1}^M \int_0^t e^{\xi H_s^{(m)}} \lambda_s^{(m)} ds < \infty$  (ie that the process  $e^{\xi H} \bullet \Lambda$  is well defined) then one can apply Theorem 2 of [7, p165], stating that the process  $(E_t)_{t \geq 0}$  defined for all  $t$  by

$$E_t = \exp(\xi H \bullet (N - \Lambda)_t - \phi(\xi H) \bullet \Lambda_t)$$

is a supermartingale. It is also the case for  $E_{t \wedge \tau}$  if  $\tau$  is a bounded stopping time. Hence for any  $\xi \in (0, 3)$  and for any  $x > 0$ , one has that

$$\mathbb{P}(E_{t \wedge \tau} > e^x) \leq e^{-x} \mathbb{E}(E_{t \wedge \tau}) \leq e^{-x},$$

which means that

$$\mathbb{P}(\xi H \bullet (N - \Lambda)_{t \wedge \tau} - \phi(\xi H) \bullet \Lambda_{t \wedge \tau} > x) \leq e^{-x}.$$

Therefore

$$\mathbb{P}(\xi H \bullet (N - \Lambda)_{t \wedge \tau} - \phi(\xi H) \bullet \Lambda_{t \wedge \tau} > x \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1) \leq e^{-x}.$$

But if  $\sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1$ , then for any  $\xi > 0$  and any  $s$ ,

$$\phi(\xi H_s^{(m)}) \leq (H_s^{(m)})^2 \phi(\xi).$$

So, for every  $\xi \in (0, 3)$ , we obtain:

$$\mathbb{P} \left( M_\tau \geq \xi^{-1} \phi(\xi) H^2 \bullet \Lambda_\tau + \xi^{-1} x \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1 \right) \leq e^{-x}. \quad (7.4)$$

Now let us focus on the event  $H^2 \bullet \Lambda_\tau \leq v$  where  $v$  is a deterministic quantity. We have that consequently

$$\mathbb{P} \left( M_\tau \geq \xi^{-1} \phi(\xi) v + \xi^{-1} x \text{ and } H^2 \bullet \Lambda_\tau \leq v \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1 \right) \leq e^{-x}.$$

It remains to choose  $\xi$  such that  $\xi^{-1} \phi(\xi) v + \xi^{-1} x$  is minimal. But this expression has no simple form. However, since  $0 < \xi < 3$ , one can bound  $\phi(\xi)$  by  $\xi^2(1 - \xi/3)^{-1}/2$ . Hence we can start with

$$\mathbb{P} \left( M_\tau \geq \frac{\xi}{2(1 - \xi/3)} H^2 \bullet \Lambda_\tau + \xi^{-1} x \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1 \right) \leq e^{-x} \quad (7.5)$$

and also

$$\mathbb{P} \left( M_\tau \geq \frac{\xi}{2(1 - \xi/3)} v + \xi^{-1} x \text{ and } H^2 \bullet \Lambda_\tau \leq v \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1 \right) \leq e^{-x}. \quad (7.6)$$

It remains now to minimize  $\xi \mapsto \frac{\xi}{2(1 - \xi/3)} v + \xi^{-1} x$ .

**Lemma 2.** *Let  $a, b$  and  $x$  be positive constants and let us consider on  $(0, 1/b)$ ,*

$$g(\xi) = \frac{a\xi}{(1 - b\xi)} + \frac{x}{\xi}.$$

*Then  $\min_{\xi \in (0, 1/b)} g(\xi) = 2\sqrt{ax} + bx$  and the minimum is achieved in  $\xi(a, b, x) = \frac{xb - \sqrt{ax}}{xb^2 - a}$ .*

**Proof.** The limits of  $g$  in  $0^+$  and  $(1/b)^-$  are  $+\infty$ . The derivative is given by

$$g'(\xi) = \frac{a}{(1 - b\xi)^2} - \frac{x}{\xi^2}$$

which is null in  $\xi(a, b, x)$  (remark that the other solution of the polynomial does not lie in  $(0, 1/b)$ ). Finally it remains to evaluate the quantity in  $\xi(a, b, x)$  to obtain the result.  $\square$

Now, we apply (7.6) with  $\xi(v/2, 1/3, x)$  and we obtain this well known formula which can be found in [50] for instance:

$$\mathbb{P} \left( M_\tau \geq \sqrt{2vx} + x/3 \text{ and } H^2 \bullet \Lambda_\tau \leq v \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1 \right) \leq e^{-x}. \quad (7.7)$$

Now we would like first to replace  $v$  by its random version  $H^2 \bullet \Lambda_\tau$ . Let  $w, v$  be some positive constants and let us concentrate on the event

$$w \leq H^2 \bullet \Lambda_\tau \leq v. \quad (7.8)$$

For all  $\varepsilon > 0$  we introduce  $K$  a positive integer depending on  $\varepsilon, v$  and  $w$  such that  $(1 + \varepsilon)^K w \geq v$ . Note that  $K = \lceil \log(v/w) / \log(1 + \varepsilon) \rceil$  is a possible choice. Let us denote  $v_0 = w, v_1 = (1 + \varepsilon)w, \dots, v_K = (1 + \varepsilon)^K w$ . For any  $0 < \xi < 3$  and any  $k$  in  $\{0, \dots, K - 1\}$ , one has, by applying (7.5),

$$\mathbb{P} \left( M_\tau \geq \frac{\xi}{2(1 - \xi/3)} H^2 \bullet \Lambda_\tau + \xi^{-1} x \text{ and } v_k \leq H^2 \bullet \Lambda_\tau \leq v_{k+1} \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1 \right) \leq e^{-x}.$$

This implies that

$$\mathbb{P} \left( M_\tau \geq \frac{\xi}{2(1 - \xi/3)} v_{k+1} + \xi^{-1} x \text{ and } v_k \leq H^2 \bullet \Lambda_\tau \leq v_{k+1} \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1 \right) \leq e^{-x}.$$

Using the previous lemma, with  $\xi = \xi(v_{k+1}/2, 1/3, x)$ , this gives

$$\mathbb{P}\left(M_\tau \geq \sqrt{2v_{k+1}x} + x/3 \text{ and } v_k \leq H^2 \bullet \Lambda_\tau \leq v_{k+1} \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}.$$

But if  $v_k \leq H^2 \bullet \Lambda_\tau$ ,  $v_{k+1} \leq (1 + \varepsilon)v_k \leq (1 + \varepsilon)H^2 \bullet \Lambda_\tau$ , so

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1 + \varepsilon)(H^2 \bullet \Lambda_\tau)x} + x/3 \text{ and } v_k \leq H^2 \bullet \Lambda_\tau \leq v_{k+1} \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}.$$

Finally summing on  $k$ , this gives

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1 + \varepsilon)(H^2 \bullet \Lambda_\tau)x} + x/3 \text{ and } w \leq H^2 \bullet \Lambda_\tau \leq v \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq Ke^{-x}. \quad (7.9)$$

This leads to the following result that has interest per se.

**Proposition 5.** *Let  $N = (N^{(m)})_{m=1, \dots, M}$  be a multivariate counting process with predictable intensities  $\lambda_t^{(m)}$  and corresponding compensator  $\Lambda_t^{(m)}$  with respect to some given filtration. Let  $B > 0$ . Let  $H = (H^{(m)})_{m=1, \dots, M}$  be a multivariate predictable process such that for all  $\xi \in (0, 3)$ ,  $e^{\xi H/B} \bullet \Lambda_t < \infty$  a.s. for all  $t$ . Let us consider the martingale defined for all  $t$  by*

$$M_t = H \bullet (N - \Lambda)_t.$$

Let  $v > w$  be positive constants and let  $\tau$  be a bounded stopping time. Then for any  $\varepsilon, x > 0$

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1 + \varepsilon)(H^2 \bullet \Lambda_\tau)x} + \frac{Bx}{3} \text{ and } w \leq H^2 \bullet \Lambda_\tau \leq v \text{ and } \sup_{m, t \leq \tau} |H_t^{(m)}| \leq B\right) \leq \left(\frac{\log(v/w)}{\log(1 + \varepsilon)} + 1\right) e^{-x}. \quad (7.10)$$

Next, we would like to replace  $H^2 \bullet \Lambda_\tau$ , the quadratic characteristic of  $M$ , with its estimator  $H^2 \bullet N_\tau$ , i.e. the quadratic variation of  $M$ . For this purpose, let us consider  $W_t = -H^2 \bullet (N - \Lambda)_t$  which is still a martingale since the  $-(H_s^{(m)})^2$ 's are still predictable processes. We apply (7.4) with  $\mu$  instead of  $\xi$ , noticing that on the event  $\{\sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\}$ , one has that  $H^4 \bullet \Lambda_\tau \leq H^2 \bullet \Lambda_\tau$ . This gives that

$$\mathbb{P}\left(H^2 \bullet \Lambda_\tau \geq H^2 \bullet N_\tau + \{\phi(\mu)/\mu\}H^2 \bullet \Lambda_\tau + x/\mu \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x},$$

which means that

$$\mathbb{P}\left(H^2 \bullet \Lambda_\tau \geq \hat{V}^\mu \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}. \quad (7.11)$$

So we use again (7.5) combined with (7.11) to obtain that for all  $\xi \in (0, 3)$

$$\begin{aligned} \mathbb{P}\left(M_\tau \geq \frac{\xi}{2(1 - \xi/3)} \hat{V}^\mu + \xi^{-1}x \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) &\leq \\ \mathbb{P}\left(M_\tau \geq \frac{\xi}{2(1 - \xi/3)} \hat{V}^\mu + \xi^{-1}x \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1 \text{ and } H^2 \bullet \Lambda_\tau \leq \hat{V}^\mu\right) &+ \\ + \mathbb{P}\left(H^2 \bullet \Lambda_\tau \geq \hat{V}^\mu \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) &\leq 2e^{-x}. \end{aligned}$$

This new inequality replaces (7.5) and it remains to replace  $H^2 \bullet \Lambda_\tau$  by  $\hat{V}^\mu$  in the peeling arguments to obtain as before that

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1 + \varepsilon)\hat{V}^\mu x} + x/3 \text{ and } w \leq \hat{V}^\mu \leq v \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq 2Ke^{-x}. \quad (7.12)$$

## 7.4 Proofs of the probabilist results for Hawkes processes

### 7.4.1 Proof of Lemma 1

Let  $K(n)$  denote the vector of the number of descendants in the  $n$ 'th generation from a single ancestral point of type  $\ell$ , define  $K(0) = \mathbf{e}_\ell$  and let  $W(n) = \sum_{k=0}^n K(k)$  denote the total number of points in the first  $n$  generations. Define for  $\theta \in \mathbb{R}^M$

$$\phi_\ell(\theta) = \log \mathbb{E}_\ell e^{\theta^T K(1)}.$$

Thus,  $\phi_\ell(\theta)$  is the log-Laplace transform of the distribution of  $K(1)$  given that there is a single initial ancestral point of type  $\ell$ . We define the vector  $\phi(\theta)$  by  $\phi(\theta)' = (\phi_1(\theta), \dots, \phi_M(\theta))$ . Note that  $\phi$  only depends on the law of the number of children per parent, ie it only depends on  $\Gamma$ . Then

$$\begin{aligned} \mathbb{E}_\ell e^{\theta^T W(n)} &= \mathbb{E}_\ell \left( e^{\theta^T W(n-1)} \mathbb{E} \left( e^{\theta^T K(n)} \mid K(n-1), \dots, K(1) \right) \right) \\ &= \mathbb{E}_\ell \left( e^{\theta^T W(n-1)} e^{\phi(\theta)^T K(n-1)} \right) \\ &= \mathbb{E}_\ell e^{(\theta + \phi(\theta))^T K(n-1) + \theta^T W(n-2)} \end{aligned}$$

Defining  $g(\theta) = \theta + \phi(\theta)$  we arrive by recursion at the formula

$$\begin{aligned} \mathbb{E}_\ell e^{\theta^T W(n)} &= \mathbb{E}_\ell e^{g^{\circ(n-1)}(\theta)^T K(1) + \theta^T W(0)} \\ &= e^{\phi(g^{\circ(n-1)}(\theta))_\ell + \theta_\ell} \\ &= e^{g^{\circ n}(\theta)_\ell}. \end{aligned}$$

Or, in other words, we have the following representation

$$\log \mathbb{E}_\ell e^{\theta^T W(n)} = g^{\circ n}(\theta)_\ell$$

of the log-Laplace transform of  $W(n)$ .

Below we show that  $\phi$  is a contraction in a neighborhood containing 0, that is, for some  $r > 0$  and a constant  $C < 1$  (and a suitable norm),  $\|\phi(s)\| \leq C\|s\|$  for  $\|s\| \leq r$ . If  $\theta$  is chosen such that

$$\frac{\|\theta\|}{1-C} \leq r$$

we have  $\|\theta\| \leq r$ , and if we assume that  $g^{\circ k}(\theta) \in B(0, r)$  for  $k = 1, \dots, n-1$  then

$$\begin{aligned} \|g^{\circ n}(\theta)\| &\leq \|\theta\| + \|\phi(g^{\circ(n-1)}(\theta))\| \\ &\leq \|\theta\| + C\|g^{\circ(n-1)}(\theta)\| \\ &\leq \|\theta\| (1 + C + C^2 + \dots + C^{n-1}) \\ &\leq r \end{aligned}$$

Thus, by induction,  $g^{\circ n}(\theta) \in B(0, r)$  for all  $n \geq 1$ . Since  $W_m(n) \nearrow W_m(\infty)$  monotonely for  $n \rightarrow \infty$ , with  $W_m(\infty)$  the total number of points in a cluster of type  $m$ , and since  $W = \sum_m W_m(\infty) = \mathbf{1}^T W(\infty)$ , we have by monotone convergence that for  $\vartheta \in \mathbb{R}$

$$\log \mathbb{E}_\ell e^{\vartheta W} = \lim_{n \rightarrow \infty} g^{\circ n}(\vartheta \mathbf{1})_\ell.$$

By the previous result, the right hand side is bounded if  $|\vartheta|$  is sufficiently small. This completes the proof up to proving that  $\phi$  is a contraction.

To this end we note that  $\phi$  is continuously differentiable (on  $\mathbb{R}^M$  in fact, but a neighborhood around 0 suffice) with derivative  $D\phi(0) = \Gamma$  at 0. Since the spectral radius of  $\Gamma$  is strictly less than 1 there is a  $C < 1$  and, by the Householder theorem, a norm  $\|\cdot\|$  on  $\mathbb{R}^M$  such that for the induced operator norm of  $\Gamma$  we have

$$\|\Gamma\| = \max_{x: \|x\| \leq 1} \|\Gamma x\| < C$$

Since the norm is continuous and  $D\phi(s)$  is likewise there is an  $r > 0$  such that

$$\|D\phi(s)\| \leq C < 1$$

for  $\|s\| \leq r$ . This, in turn, implies that  $\phi$  is Lipschitz continuous in the ball  $B(0, r)$  with Lipschitz constant  $C$ , and since  $\phi(0) = 0$  we get

$$\|\phi(s)\| \leq C\|s\|$$

for  $\|s\| \leq r$ . This ends the proof of the lemma.

Note that we have not at all used the explicit formula for  $\phi$  above, which is obtainable and simple since the offspring distributions are Poisson. The only thing we needed was the fact that  $\phi$  is defined in a neighborhood around 0, thus that the offspring distributions are sufficiently light-tailed.

### 7.4.2 Proof of Proposition 1

We use the cluster representation, and we note that any cluster with ancestral point in  $[-n-1, -n]$  must have at least  $n+1 - \lceil A \rceil$  points in the cluster if any of the points are to fall in  $[-A, 0)$ . This follows from the assumption that all the  $h_\ell^{(m)}$ -functions have support in  $[0, 1]$ . With  $\tilde{N}_{A,\ell}$  the number of points in  $[-A, 0)$  from a cluster with ancestral points of type  $\ell$  we thus have the bound

$$\tilde{N}_{A,\ell} \leq \sum_n \sum_{k=1}^{A_n} \max\{W_{n,k} - n + \lceil A \rceil, 0\}$$

where  $A_n$  is the number of ancestral points in  $[-n-1, -n]$  of type  $\ell$  and  $W_{n,k}$  is the number of points in the respective clusters. Here the  $A_n$ 's and the  $W_{n,k}$ 's are all independent, the  $A_n$ 's are Poisson distributed with mean  $\nu_\ell$  and the  $W_{n,k}$ 's are iid with the same distribution as  $W$  in Lemma 1. Moreover,

$$H_n(\vartheta_\ell) := \mathbb{E}_\ell e^{\vartheta_\ell \max\{W - n + \lceil A \rceil, 0\}} \leq \mathbb{P}_\ell(W \leq n - \lceil A \rceil) + e^{-\vartheta_\ell(n - \lceil A \rceil)} \mathbb{E}_\ell e^{\vartheta_\ell W},$$

which is finite for  $|\vartheta_\ell|$  sufficiently small according to Lemma 1. Then we can compute an upper bound on the Laplace transform of  $\tilde{N}_{A,\ell}$ :

$$\begin{aligned} \mathbb{E} e^{\vartheta_\ell \tilde{N}_{A,\ell}} &\leq \prod_n \mathbb{E} \prod_{k=1}^{A_n} \mathbb{E} \left( e^{\vartheta_\ell \max\{W_{n,k} - n + \lceil A \rceil, 0\}} \mid A_n \right) \\ &\leq \prod_n \mathbb{E} H_n(\vartheta_\ell)^{A_n} \\ &= \prod_n e^{\nu_\ell (H_n(\vartheta_\ell) - 1)} \\ &= e^{\nu_\ell \sum_n (H_n(\vartheta_\ell) - 1)} \end{aligned}$$

Since  $H_n(\vartheta_\ell) - 1 \leq e^{-\vartheta_\ell(n - \lceil A \rceil)} \mathbb{E}_\ell e^{\vartheta_\ell W}$  we have  $\sum_n (H_n(\vartheta_\ell) - 1) < \infty$ , which shows that the upper bound is finite. To complete the proof, observe that  $N_{[-A, 0)} = \sum_\ell \tilde{N}_{A,\ell}$  where  $\tilde{N}_{A,\ell}$  for  $\ell = 1, \dots, M$  are independent. Since all variables are positive, it is sufficient to take  $\theta = \min_\ell \vartheta_\ell$ .

### 7.4.3 Proof of Proposition 2

In this paragraph, the notation  $\square$  simply denotes a generic positive absolute constant that may change from line to line. The notation  $\square_{\theta_1, \theta_2, \dots}$  denotes a positive constant depending on  $\theta_1, \theta_2, \dots$  that may change from line to line.

Let

$$u = C_1 \sigma \log^{3/2}(T) \sqrt{T} + C_2 b(\log(T))^{2+\eta}, \quad (7.13)$$

where the choices of  $C_1$  and  $C_2$  will be given later. For any positive integer  $k$  such that  $x := T/(2k) > A$ , we have by stationarity:

$$\begin{aligned} \mathbb{P}\left(\int_0^T [Z \circ \theta_t(N) - \mathbb{E}(Z)]dt \geq u\right) &= \mathbb{P}\left(\sum_{q=0}^{k-1} \int_{2qx}^{2qx+x} [Z \circ \theta_t(N) - \mathbb{E}(Z)]dt + \int_{2qx+x}^{2qx+2x} [Z \circ \theta_t(N) - \mathbb{E}(Z)]dt \geq u\right) \\ &\leq 2\mathbb{P}\left(\sum_{q=0}^{k-1} \int_{2qx}^{2qx+x} [Z \circ \theta_t(N) - \mathbb{E}(Z)]dt \geq \frac{u}{2}\right). \end{aligned}$$

Similarly to [45], we introduce  $(\tilde{M}_q^x)_q$  a sequence of independent Hawkes processes, each being stationary with intensities per mark given by  $\psi_t^{(m)}$ . For each  $q$ , we then introduce  $M_q^x$  the truncated process associated with  $\tilde{M}_q^x$ , where truncation means that we only consider the points lying in  $[2qx - A, 2qx + x]$ . So, if we set

$$\begin{aligned} F_q &= \int_{2qx}^{2qx+x} [Z \circ \theta_t(M_q^x) - \mathbb{E}(Z)]dt, \\ \mathbb{P}\left(\int_0^T [Z \circ \theta_t(N) - \mathbb{E}(Z)]dt \geq u\right) &\leq 2\mathbb{P}\left(\sum_{q=0}^{k-1} F_q \geq \frac{u}{2}\right) + 2k\mathbb{P}\left(T_e > \frac{T}{2k} - A\right), \end{aligned} \quad (7.14)$$

where  $T_e$  represents the time to extinction of the process. More precisely  $T_e$  is the last point of the process if in the cluster representation only ancestral points before 0 are appearing. For more details, see section 3 of [45]. So, denoting  $a_l$  the ancestral points with marks  $l$  and  $H_{a_l}^l$  the length of the corresponding cluster whose origin is  $a_l$ , we have:

$$T_e = \max_{l \in \{1, \dots, M\}} \max_{a_l} \{a_l + H_{a_l}^l\}.$$

But, for any  $a > 0$ ,

$$\begin{aligned} \mathbb{P}(T_e \leq a) &= \mathbb{E}\left[\prod_{l=1}^M \prod_{a_l} \mathbb{E}\left[1_{\{a_l + H_{a_l}^l \leq a\}} | a_l\right]\right] \\ &= \mathbb{E}\left[\prod_{l=1}^M \prod_{a_l} \exp(\log(\mathbb{P}(H_0^l \leq a - a_l)))\right] \\ &= \mathbb{E}\left[\prod_{l=1}^M \exp\left(\int_{-\infty}^0 \log(\mathbb{P}(H_0^l \leq a - x)) d\tilde{N}_x^{(l)}\right)\right], \end{aligned}$$

where  $\tilde{N}^{(l)}$  denotes the process associated with the ancestral points with marks  $l$ . So,

$$\begin{aligned} \mathbb{P}(T_e \leq a) &= \exp\left(\sum_{l=1}^M \int_{-\infty}^0 (\exp(\log(\mathbb{P}(H_0^l \leq a - x))) - 1) \nu^{(l)} dx\right) \\ &= \exp\left(-\sum_{l=1}^M \nu^{(l)} \int_a^{+\infty} \mathbb{P}(H_0^l > u) du\right). \end{aligned}$$

Now, by Lemma 1, there exists some  $\vartheta_l > 0$ , such that  $c_l = \mathbb{E}_\ell(e^{\vartheta_l W}) < +\infty$ , where  $W$  is the number of points in the cluster. But if all the interaction functions have support in  $[0, 1]$ , one always have that  $H_0^l < W$ . Hence

$$\begin{aligned} \mathbb{P}(H_0^l > u) &\leq \mathbb{E}[\exp(\vartheta_l H_0^l)] \exp(-\vartheta_l u) \\ &\leq c_l \exp(-\vartheta_l u). \end{aligned}$$

So,

$$\begin{aligned}
\mathbb{P}(T_e \leq a) &\geq \exp\left(-\sum_{l=1}^M \nu^{(l)} \int_a^{+\infty} c_l \exp(-\vartheta_l u) du\right) \\
&= \exp\left(-\sum_{l=1}^M \nu^{(l)} c_l / \vartheta_l \exp(-\vartheta_l a)\right) \\
&\geq 1 - \sum_{l=1}^M \nu^{(l)} c_l / \vartheta_l \exp(-\vartheta_l a).
\end{aligned}$$

So, there exists a constant  $C_{\alpha, f^*, A}$  depending on  $\alpha, A$ , and  $f^*$  such that if we take  $k = \lfloor C_{\alpha, A, f^*} T / \log(T) \rfloor$ , then

$$k\mathbb{P}\left(T_e > \frac{T}{2k} - A\right) \leq T^{-\alpha}.$$

In this case  $x = \frac{T}{2k} \approx \log(T)$  is larger than  $A$  for  $T$  large enough (depending on  $A, \alpha, f^*$ ).

Now, let us focus on the first term  $B$  of (7.14), where

$$B = \mathbb{P}\left(\sum_{q=0}^{k-1} F_q \geq \frac{u}{2}\right).$$

Let us consider some  $\tilde{\mathcal{N}}$  where  $\tilde{\mathcal{N}}$  will be fixed later and let us define the measurable events

$$\Omega_q = \left\{ \sup_t \{M_q^x|_{[t-A, t]}\} \leq \tilde{\mathcal{N}} \right\},$$

where  $M_q^x|_{[t-A, t]}$  represents the set of points of  $M_q^x$  lying in  $[t-A, t)$ . Let us also consider  $\Omega = \cap_{1 \leq q \leq k} \Omega_q$ . Then

$$B \leq \mathbb{P}\left(\sum_q F_q \geq u/2 \text{ and } \Omega\right) + \mathbb{P}(\Omega^c).$$

We have  $\mathbb{P}(\Omega^c) \leq \sum_q \mathbb{P}(\Omega_q^c)$ . Each  $\Omega_q$  can also be easily controlled. Indeed it is sufficient to split  $[2qx-A, 2qx+x]$  in intervals of size  $A$  (there are about  $\square_{\alpha, A, f^*} \log(T)$  of those) and require that the number of points in each subinterval is smaller than  $\tilde{\mathcal{N}}/2$ . By stationarity, we obtain that

$$\mathbb{P}(\Omega_q^c) \leq \square_{\alpha, A, f^*} \log(T) \mathbb{P}(N_{[-A, 0]} > \tilde{\mathcal{N}}/2).$$

Using Proposition 1 with  $u = \lceil \tilde{\mathcal{N}}/2 \rceil + 1/2$ , we obtain:

$$\mathbb{P}(\Omega_q^c) \leq \square_{\alpha, A, f^*} \log(T) \exp(-\square_{\alpha, A, f^*} \tilde{\mathcal{N}}) \text{ and } \mathbb{P}(\Omega^c) \leq \square_{\alpha, A, f^*} T \exp(-\square_{\alpha, A, f^*} \tilde{\mathcal{N}}). \quad (7.15)$$

Note that this control holds for any positive choice of  $\tilde{\mathcal{N}}$ . Hence this gives also the following Lemma that will be used later.

**Lemma 3.** *For any  $\mathcal{R} > 0$ ,*

$$\mathbb{P}\left(\text{there exists } t \in [0, T] \mid M_q^x|_{[t-A, t]} > \mathcal{R}\right) \leq \square_{\alpha, A, f^*} T \exp(-\square_{\alpha, A, f^*} \mathcal{R}).$$

Hence by taking  $\tilde{\mathcal{N}} = C_3 \log(T)$  for  $C_3$  large enough this is smaller than  $\square_{\alpha, A, f^*} T^{-\alpha'}$ , where  $\alpha' = \max(\alpha, 2)$ .

It remains to obtain the rate of  $D := \mathbb{P}(\sum_q F_q \geq u/2 \text{ and } \Omega)$ . For any positive constant  $\theta$  that will be chosen later, we have:

$$\begin{aligned}
D &\leq e^{-\frac{\theta u}{2}} \mathbb{E}\left(e^{\theta \sum_q F_q} \prod_q \mathbb{1}_{\Omega_q}\right) \\
&\leq e^{-\frac{\theta u}{2}} \prod_q \mathbb{E}(e^{\theta F_q} \mathbb{1}_{\Omega_q})
\end{aligned} \quad (7.16)$$



since the variables  $(M_q^x)_q$  are independent. But

$$\mathbb{E}(e^{\theta F_q} \mathbb{1}_{\Omega_q}) = 1 + \theta \mathbb{E}(F_q \mathbb{1}_{\Omega_q}) + \sum_{j \geq 2} \frac{\theta^j}{j!} \mathbb{E}(F_q^j \mathbb{1}_{\Omega_q})$$

and  $\mathbb{E}(F_q \mathbb{1}_{\Omega_q}) = \mathbb{E}(F_q) - \mathbb{E}(F_q \mathbb{1}_{\Omega_q^c}) = -\mathbb{E}(F_q \mathbb{1}_{\Omega_q^c})$ .

Next note that if for any integer  $l$ ,

$$l\tilde{\mathcal{N}} < \sup_t M_q^x|_{[t-A, t]} \leq (l+1)\tilde{\mathcal{N}}$$

then

$$|F_q| \leq xb[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1] + x\mathbb{E}(f).$$

Hence, cutting  $\Omega_q^c$  in slices of the type  $\{l\tilde{\mathcal{N}} < \sup_t M_q^x|_{[t-A, t]} \leq (l+1)\tilde{\mathcal{N}}\}$  and using Lemma 3, we obtain by taking  $C_3$  large enough,

$$\begin{aligned} |\mathbb{E}(F_q \mathbb{1}_{\Omega_q})| = |\mathbb{E}(F_q \mathbb{1}_{\Omega_q^c})| &\leq \sum_{l=1}^{+\infty} x(b[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1] + |\mathbb{E}(Z)|) \mathbb{P}(\text{there exists } t \in [0, T] \mid \{M_q^x|_{[t-A, t]}\} > l\tilde{\mathcal{N}}) \\ &\leq \square_{\alpha, A, f^*} \sum_{l=1}^{+\infty} x(b[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1] + |\mathbb{E}(Z)|) \log(T) e^{-\square_{\alpha, A, f^*} l\tilde{\mathcal{N}}} \\ &\leq \square_{\alpha, A, f^*} \sum_{l=1}^{+\infty} x(b\tilde{\mathcal{N}}^\eta + |\mathbb{E}(Z)|) \log(T) 2^{l\eta} e^{-\square_{\alpha, A, f^*} l\tilde{\mathcal{N}}} \\ &\leq \square_{\alpha, \eta, A, f^*} \log^2(T) b\tilde{\mathcal{N}}^\eta \frac{e^{-\square_{\alpha, A, f^*} \tilde{\mathcal{N}}}}{1 - 2^\eta e^{-\square_{\alpha, A, f^*} \tilde{\mathcal{N}}}} \\ &\leq z_1 := \square_{\alpha, \eta, A, f^*} bT^{-\alpha'}. \end{aligned}$$

Note that in the previous inequalities, we have bounded  $|\mathbb{E}(Z)|$  by  $b\mathbb{E}[N_{[-A, 0]}^\eta]$ . In the same way, one can bound

$$\mathbb{E}(F_q^j \mathbb{1}_{\Omega_q}) \leq \mathbb{E}(F_q^2 \mathbb{1}_{\Omega_q}) z_b^{j-2},$$

with  $z_b := xb[\tilde{\mathcal{N}}^\eta + 1] + x\mathbb{E}(Z) = \square_{\alpha, \eta, A, f^*} b \log(T)^{1+\eta}$ . One can also note that by stationarity,

$$\begin{aligned} \mathbb{E}(F_q^2 \mathbb{1}_{\Omega_q}) &\leq x\mathbb{E} \left[ \int_{2qx}^{2qx+x} [Z \circ \theta_s(M_q^x) - \mathbb{E}(Z)]^2 \mathbb{1}_{\{\text{for all } t, M_q^x|_{[t-A, t]} \leq \tilde{\mathcal{N}}\}} ds \right] \\ &\leq x\mathbb{E} \left[ \int_{2qx}^{2qx+x} [Z \circ \theta_s(M_q^x) - \mathbb{E}(Z)]^2 \mathbb{1}_{\{M_q^x|_{[s-A, s]} \leq \tilde{\mathcal{N}}\}} ds \right] \\ &\leq x^2 \mathbb{E}([Z(N) - \mathbb{E}(Z)]^2 \mathbb{1}_{N_{[-A, 0]} \leq \tilde{\mathcal{N}}}) \\ &\leq z_v := \square_{\alpha, \eta, A, f^*} (\log(T))^2 \sigma^2. \end{aligned}$$

Now let us go back to (7.16). We have that

$$\begin{aligned} D &\leq \exp \left[ -\frac{\theta u}{2} + k \ln \left( 1 + \theta z_1 + \sum_{j \geq 2} z_v z_b^{j-2} \frac{\theta^j}{j!} \right) \right] \\ &\leq \exp \left[ -\theta \left( \frac{u}{2} - k z_1 \right) + k \sum_{j \geq 2} z_v z_b^{j-2} \frac{\theta^j}{j!} \right], \end{aligned}$$

using that  $\ln(1+u) \leq u$ . It is sufficient now to recognize a step of the proof of the Bernstein inequality (weak version see [36, p25]). Since  $kz_1 = \square_{\alpha, \eta, s} bT^{1-\alpha'} / (\log(T))$ , one can choose  $\alpha' > 1, C_1$  and  $C_2$  in the definition

(7.13) of  $u$  (not depending on  $b$ ) such that  $u/2 - kz_1 \geq \sqrt{2kz_v z} + \frac{1}{3}z_b z$  for some  $z = C_4 \log(T)$ , where  $C_4$  is a constant. Hence

$$D \leq \exp \left[ -\theta \left( \sqrt{2kz_v z} + \frac{1}{3}z_b z \right) + k \sum_{j \geq 2} z_v z_b^{j-2} \frac{\theta^j}{j!} \right].$$

One can choose accordingly  $\theta$  (as for the proof of the Bernstein inequality) to obtain a bound in  $e^{-z}$ . It remains to choose  $C_4$  large enough and only depending on  $\alpha, \eta, A$  and  $f^*$  to guarantee that  $D \leq e^{-z} \leq \square_{\alpha, \eta, A, f^*} T^{-\alpha}$ . This concludes the proof of the proposition.

#### 7.4.4 Proof of Proposition 3

Let  $\mathbb{Q}$  denote a measure such that under  $\mathbb{Q}$  the distribution of the full point process restricted to  $(-\infty, 0]$  is identical to the distribution under  $\mathbb{P}$  and such that on  $(0, \infty)$  the process consists of independent components each being a homogeneous Poisson process with rate 1. Furthermore, the Poisson processes should be independent of the process on  $(-\infty, 0]$ . From Corollary 5.1.2 in [32] the likelihood process is given by

$$\mathcal{L}_t = \exp \left( Mt - \sum_m \int_0^t \lambda_u^{(m)} du + \sum_m \int_0^t \log \lambda_u^{(m)} dN_u^{(m)} \right)$$

and we have for  $t \geq 0$  the relation

$$\mathbb{E}_{\mathbb{P}} \kappa_t(\mathbf{f})^2 = \mathbb{E}_{\mathbb{Q}} \kappa_t(\mathbf{f})^2 \mathcal{L}_t, \quad (7.17)$$

where  $\mathbb{E}_{\mathbb{P}}$  and  $\mathbb{E}_{\mathbb{Q}}$  denote the expectation with respect to  $\mathbb{P}$  and  $\mathbb{Q}$  respectively. Let, furthermore,  $\tilde{N}_1 = N_{[-1, 0]}$  denote the total number of points on  $[-1, 0)$ . Proposition 3 will be an easy consequence of the following lemma.

**Lemma 4.** *If the point process is stationary under  $\mathbb{P}$ , if*

$$e^d \leq \lambda_t^{(m)} \leq a(N_1 + \tilde{N}_1) + b$$

for  $t \in [0, 1]$  and for constants  $d \in \mathbb{R}$  and  $a, b > 0$ , and if  $\mathbb{E}_{\mathbb{P}}(1 + \varepsilon)^{\tilde{N}_1} < \infty$  for some  $\varepsilon > 0$  then for any  $\mathbf{f}$ ,

$$Q(\mathbf{f}, \mathbf{f}) \geq \zeta \|\mathbf{f}\|^2 \quad (7.18)$$

for some constant  $\zeta > 0$ .

**Proof.** We use Hölders inequality on  $\kappa_1(\mathbf{f})^{\frac{2}{p}} \mathcal{L}_1^{\frac{1}{p}}$  and  $\kappa_1(\mathbf{f})^{\frac{2}{q}} \mathcal{L}_1^{-\frac{1}{p}}$  to get

$$\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \leq (\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \mathcal{L}_1)^{\frac{1}{p}} \left( \mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \mathcal{L}_1^{-\frac{q}{p}} \right)^{\frac{1}{q}} = Q(\mathbf{f}, \mathbf{f})^{\frac{1}{p}} \left( \mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \mathcal{L}_1^{1-q} \right)^{\frac{1}{q}} \quad (7.19)$$

where  $\frac{1}{p} + \frac{1}{q} = 1$ . We choose  $q \geq 1$  (and thus  $p$ ) below to make  $q - 1$  sufficiently small. For the left hand side we have by independence of the homogeneous Poisson processes that if  $\mathbf{f} = (\mu, (g_\ell)_{\ell=1, \dots, M})$ ,

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 &= (\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f}))^2 + \mathbb{V}_{\mathbb{Q}} \kappa_1(\mathbf{f}) \\ &= \left( \mu + \sum_{\ell} \int_0^1 g_\ell(u) du \right)^2 + \sum_{\ell} \int_0^1 g_\ell(u)^2 du. \end{aligned}$$

Exactly as on page 32 in [46] there exists  $c' > 0$  such that

$$\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \geq c' \left( \mu^2 + \sum_{\ell} \int_0^1 g_\ell^2(u) du \right) = c' \|\mathbf{f}\|^2. \quad (7.20)$$

To bound the second factor on the right hand side in (7.19) we observe, by assumption, that we have the lower bound

$$\mathcal{L}_1 \geq e^{M(1-b)} e^{(d-aM)N_1} e^{-aM\tilde{N}_1}$$

on the likelihood process. Under  $\mathbb{Q}$  we have that  $(\kappa_1(\mathbf{f}), N_1)$  and  $\tilde{N}_1$  are independent, and with  $\rho = e^{(q-1)(aM-d)}$  and  $\tilde{\rho} = e^{(q-1)aM}$  we get that

$$\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \mathcal{L}_1^{1-q} \leq e^{(q-1)M(b-1)} \mathbb{E}_{\mathbb{Q}} \tilde{\rho}^{\tilde{N}_1} \mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \rho^{N_1}.$$

Here we choose  $q$  such that  $\tilde{\rho}$  is sufficiently close to 1 to make sure that  $\mathbb{E}_{\mathbb{Q}} \tilde{\rho}^{\tilde{N}_1} = \mathbb{E}_{\mathbb{P}} \tilde{\rho}^{\tilde{N}_1} < \infty$ . Moreover, by Cauchy-Schwarz' inequality

$$\kappa_1^2(\mathbf{f}) \leq \left( \mu^2 + \sum_{\ell} \int_0^{1-} g_{\ell}^2(1-u) dN_u^{(\ell)} \right) (1 + N_1). \quad (7.21)$$

Under  $\mathbb{Q}$  the point processes on  $(0, \infty)$  are homogeneous Poisson processes with rate 1 and  $N_1$ , the total number of points, is Poisson. This implies that conditionally on  $(N_1^{(1)}, \dots, N_1^{(M)}) = (n^{(1)}, \dots, n^{(M)})$  the  $n^{(m)}$ -points for the  $m$ 'th process are uniformly distributed on  $[0, 1]$ , hence

$$\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \mathcal{L}_1^{1-q} \leq \left( \mu^2 + \sum_{\ell} \int_0^1 g_{\ell}^2(u) du \right) \underbrace{e^{(q-1)M(b-1)} \mathbb{E}_{\mathbb{Q}} \tilde{\rho}^{\tilde{N}_1} \mathbb{E}_{\mathbb{Q}} (1 + N_1)^2 \rho^{N_1}}_{c''} = c'' \|\mathbf{f}\|^2. \quad (7.22)$$

Combining (7.20) and (7.22) with (7.19) we get that

$$c' \|\mathbf{f}\|^2 \leq (c'')^{\frac{1}{q}} \|\mathbf{f}\|^{\frac{2}{q}} Q(\mathbf{f}, \mathbf{f})^{\frac{1}{p}}$$

or by rearranging that

$$Q(\mathbf{f}, \mathbf{f}) \geq \zeta \|\mathbf{f}\|^2$$

with  $\zeta = (c')^p / (c'')^{p-1}$ .  $\square$

For the Hawkes process it follows that if  $\nu^{(m)} > 0$  and if

$$\sup_{t \in [0,1]} h_{\ell}^{(m)}(t) < \infty$$

for  $l, m = 1, \dots, M$  then for  $t \in [0, 1]$  we have  $e^d \leq \lambda_t^{(m)} \leq a(N_1 + \tilde{N}_1) + b$  with

$$d = \log \nu^{(m)}, \quad a = \max_l \sup_{t \in [0,1]} h_{\ell}^{(m)}(t), \quad b = \nu^{(m)}.$$

Proposition 1 proves that there exists  $\varepsilon > 0$  such that  $\mathbb{E}_{\mathbb{P}}(1 + \varepsilon)^{\tilde{N}_1} < \infty$ . This ends the proof of Proposition 3.

## 7.5 Proofs of the results of Section 5.3

### 7.5.1 Proof of Proposition 4

As in the proof of Proposition 2, we use the notation  $\square$ . Note that for any  $\varphi_1$  and any  $\varphi_2$  belonging to  $\Phi$ ,

$$G_{\varphi_1, \varphi_2} = \psi(\varphi_1) \bullet \Psi(\varphi_2)_T = \sum_{m=1}^M \int_0^T \kappa_t(\varphi_1^{(m)}) \kappa_t(\varphi_2^{(m)}) dt$$

and  $\mathbb{E}(G_{\varphi_1, \varphi_2}) = T \sum_{m=1}^M Q(\varphi_1^{(m)}, \varphi_2^{(m)})$  by using (5.2). This implies that

$$\mathbb{E}(a'Ga) = a' \mathbb{E}(G)a = T \sum_m Q(\mathbf{f}_a^{(m)}, \mathbf{f}_a^{(m)}).$$

Hence by Proposition 3,  $\mathbb{E}(a'Ga) \geq T\zeta \sum_m \|\mathbf{f}_a^{(m)}\|^2 = T\zeta \|f_a\|^2$  by definition of the norm on  $\mathcal{H}$ . Since  $\Phi$  is an orthonormal system, this implies that  $\mathbb{E}(a'Ga) \geq T\zeta \|a\|_{\ell^2}$ . Hence, to show that  $\Omega_c$  is a large event for some  $c > 0$ , it is sufficient to show that for some  $0 < \epsilon < \zeta$ , with high probability, for any  $a \in \mathbb{R}^{\Phi}$ ,

$$|a'Ga - a' \mathbb{E}(G)a| \leq T\epsilon \|a\|_{\ell^2}. \quad (7.23)$$

Indeed, (7.23) implies that, with high probability, for any  $a \in \mathbb{R}^\Phi$ ,

$$a'Ga \geq a'\mathbb{E}(G)a - T\epsilon\|a\|_{\ell^2} \geq T(\zeta - \epsilon)\|a\|_{\ell^2},$$

and the choice  $c = T(\zeta - \epsilon)$  is convenient. So, first one has to control all the coefficients of  $G - \mathbb{E}(G)$ . For all  $\varphi, \rho \in \Phi$ , we apply Proposition 2 to

$$Z(N) = \sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho).$$

Note that  $Z$  only depends on points lying in  $[-1, 0)$ . Therefore,  $|Z(N)| \leq 2M\|\varphi\|_\infty\|\rho\|_\infty(1 + N_{[-1,0)}^2)$ . This leads to

$$\mathbb{P}\left(\frac{1}{T}\left|G_{\varphi,\rho} - \mathbb{E}(G_{\varphi,\rho})\right| \geq x_{\varphi,\rho}\right) \leq \square_{\alpha,f^*}T^{-\alpha}$$

with

$$x_{\varphi,\rho} = \square_{\alpha,f^*,M}[\sigma_{\varphi,\rho} \log^{3/2}(T)T^{-1/2} + \|\varphi\|_\infty\|\rho\|_\infty \log^4(T)T^{-1}]$$

and

$$\sigma_{\varphi,\rho}^2 = \mathbb{E}\left[\left[\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho) - \mathbb{E}\left(\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)\right)\right]^2 \mathbb{1}_{N_{[-1,0)} \leq \tilde{\mathcal{N}}}\right].$$

Hence, with probability larger than  $1 - \square_{\alpha,f^*}|\Phi|^2T^{-\alpha}$  one has that

$$|a'Ga - a'\mathbb{E}(G)a| \leq \square_{\alpha,f^*} \left( \sum_{\varphi,\rho \in \Phi} |a_\varphi||a_\rho| [\sigma_{\varphi,\rho} \log^{3/2}(T)T^{1/2} + \|\varphi\|_\infty\|\rho\|_\infty \log^4(T)] \right).$$

Hence, for any positive constant  $\delta$  chosen later,

$$|a'Ga - a'\mathbb{E}(G)a| \leq \square_{\alpha,f^*} \left[ T \sum_{\varphi,\rho \in \Phi} |a_\varphi||a_\rho| \left[ \delta \frac{\sigma_{\varphi,\rho}^2}{\|\varphi\|_\infty\|\rho\|_\infty} + \left[ \frac{1}{\delta \log(T)} + 1 \right] \|\varphi\|_\infty\|\rho\|_\infty \frac{\log^4(T)}{T} \right] \right]. \quad (7.24)$$

Now let us focus on  $E := \sum_{\varphi,\rho \in \Phi} |a_\varphi||a_\rho| \frac{\sigma_{\varphi,\rho}^2}{\|\varphi\|_\infty\|\rho\|_\infty}$ . First, we have:

$$E \leq 2 \sum_{\varphi,\rho \in \Phi} |a_\varphi||a_\rho| \frac{\mathbb{E}([\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)]^2 \mathbb{1}_{N_{[-1,0)} \leq \tilde{\mathcal{N}}}) + (\mathbb{E}[\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)])^2}{\|\varphi\|_\infty\|\rho\|_\infty}$$

with  $\tilde{\mathcal{N}} := \square_{\alpha,f^*} \log(T)$ . Next,

$$\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho) \leq 2M\|\varphi\|_\infty\|\rho\|_\infty(1 + N_{[-1,0)}^2).$$

Hence, if  $N_{[-1,0)} \leq \tilde{\mathcal{N}} = \square_{\alpha,f^*} \log(T)$ , for  $T$  large enough,

$$\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho) \leq \square_{\alpha,M,f^*} \|\varphi\|_\infty\|\rho\|_\infty \log^2(T)$$

and

$$\mathbb{E}\left(\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)\right) \leq \square_{\alpha,M,f^*} \|\varphi\|_\infty\|\rho\|_\infty \log^2(T).$$

Hence,

$$E \leq \square_{\alpha,M,f^*} \log^2(T) \sum_{\varphi,\rho \in \Phi} |a_\varphi||a_\rho| \mathbb{E}\left(\left|\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)\right|\right).$$

But note that for any  $f$ ,  $|\psi_0^{(m)}(f)| \leq \psi_0^{(m)}(|f|)$  where  $|f| = ((|\mu^{(m)}|, (|g_\ell^{(m)}|)_{\ell=1, \dots, M})_{m=1, \dots, M})$ . Therefore,

$$\begin{aligned} E &\leq \square_{\alpha, M, f^*} \log^2(T) \sum_{\varphi, \rho \in \Phi} |a_\varphi| |a_\rho| \mathbb{E} \left( \sum_m \psi_0^{(m)}(|\varphi|) \psi_0^{(m)}(|\rho|) \right) \\ &\leq \square_{\alpha, M, f^*} \log^2(T) \sum_m \mathbb{E} \left( \left[ \sum_{\varphi \in \Phi} |a_\varphi| \psi_0^{(m)}(|\varphi|) \right]^2 \right) \\ &\leq \square_{\alpha, M, f^*} \log^2(T) \sum_m \mathbb{E} \left( \left[ \psi_0^{(m)} \left( \sum_{\varphi \in \Phi} |a_\varphi| |\varphi| \right) \right]^2 \right). \end{aligned}$$

But if  $\varphi = (\mu_\varphi^{(m)}, (g_\varphi)_\ell^{(m)})_\ell$ , then

$$\left[ \psi_0^{(m)} \left( \sum_{\varphi \in \Phi} |a_\varphi| |\varphi| \right) \right]^2 = \left[ \sum_{\varphi} |a_\varphi| \mu_\varphi^{(m)} + \sum_{\ell=1}^M \int_{-1}^{0-} \sum_{\varphi} |a_\varphi| |(g_\varphi)_\ell^{(m)}| (-u) dN_u^{(\ell)} \right]^2.$$

If one creates artificially a process  $N^{(0)}$  with only one point and if we decide that  $(g_\varphi)_0^{(m)}$  is the constant function equal to  $\mu_\varphi^{(m)}$ , this can also be rewritten as

$$\left[ \psi_0^{(m)} \left( \sum_{\varphi \in \Phi} |a_\varphi| |\varphi| \right) \right]^2 = \left[ \sum_{\ell=0}^M \int_{-1}^{0-} \sum_{\varphi} |a_\varphi| |(g_\varphi)_\ell^{(m)}| (-u) dN_u^{(\ell)} \right]^2.$$

Now we apply the Cauchy-Schwarz inequality for the measure  $\sum_\ell dN^{(\ell)}$ , which gives

$$\left[ \psi_0^{(m)} \left( \sum_{\varphi \in \Phi} |a_\varphi| |\varphi| \right) \right]^2 \leq (N_{[-1, 0)} + 1) \sum_{\ell=0}^M \int_{-1}^{0-} \left[ \sum_{\varphi} |a_\varphi| |(g_\varphi)_\ell^{(m)}| (-u) \right]^2 dN_u^{(\ell)}.$$

Consequently,

$$\begin{aligned} E &\leq \square_{\alpha, M, f^*} \log^2(T) \sum_{m=1}^M \sum_{\ell=0}^M \mathbb{E} \left( (N_{[-1, 0)} + 1) \int_{-1}^{0-} \left[ \sum_{\varphi} |a_\varphi| |(g_\varphi)_\ell^{(m)}| (-u) \right]^2 dN_u^{(\ell)} \right) \\ &\leq \square_{\alpha, M, f^*} \log^2(T) \sum_{m=1}^M \sum_{\ell=0}^M \sum_{\varphi, \rho \in \Phi} |a_\varphi| |a_\rho| \mathbb{E} \left( \int_{-1}^{0-} (N_{[-1, 0)} + 1) |(g_\varphi)_\ell^{(m)}| (-u) |(g_\rho)_\ell^{(m)}| (-u) dN_u^{(\ell)} \right). \end{aligned}$$

Now let us use the fact that for every  $x, y \geq 0$ ,  $\eta, \theta > 0$  that will be chosen later,

$$xy - \eta e^{\theta x} \leq \frac{y}{\theta} [\log(y) - \log(\eta\theta) - 1],$$

with the convention that  $y \log(y) = 0$  if  $y = 0$ . Let us apply this to  $x = N_{[-1, 0)} + 1$  and  $y = |(g_\varphi)_\ell^{(m)}| (-u) |(g_\rho)_\ell^{(m)}| (-u)$ . We obtain that

$$\begin{aligned} E &\leq \square_{\alpha, M, f^*} \eta \log^2(T) \sum_{m=1}^M \sum_{\varphi, \rho \in \Phi} |a_\varphi| |a_\rho| \mathbb{E} \left( (N_{[-1, 0)} + 1) e^{\theta(N_{[-1, 0)} + 1)} + \right. \\ &\quad \left. \square_{\alpha, M, f^*} \theta^{-1} \log^2(T) \sum_{m=1}^M \sum_{\ell=0}^M \sum_{\varphi, \rho \in \Phi} |a_\varphi| |a_\rho| \mathbb{E} \left( \int_{-1}^{0-} |(g_\varphi)_\ell^{(m)}| |(g_\rho)_\ell^{(m)}| (-u) [\log(|(g_\varphi)_\ell^{(m)}| |(g_\rho)_\ell^{(m)}| (-u)) - \log(\eta\theta) - 1] dN_u^{(\ell)} \right) \right). \end{aligned}$$

Since for  $\ell > 0$ ,  $dN_u^{(\ell)}$  is stationary, one can replace  $\mathbb{E}(dN_u^{(\ell)})$  by  $\square_{f^*} du$ . Moreover since by Proposition 1,  $N_{[-1,0]}$  has some exponential moments there exists  $\theta = \square_{f^*}$  such that  $\mathbb{E}((N_{[-1,0]} + 1)e^{\theta(N_{[-1,0]} + 1)}) = \square_{f^*}$ . With  $|\Phi|$  the size of the dictionary, this leads to

$$E \leq \square_{\alpha, M, f^*} \eta |\Phi| \log^2(T) \|a\|_{\ell_2}^2 + \square_{\alpha, M, f^*} \log^2(T) \sum_{m=1}^M \left[ \sum_{\varphi, \rho \in \Phi} |a_\varphi| |a_\rho| |\mu_\varphi^{(m)}| |\mu_\rho^{(m)}| \left[ \log(|\mu_\varphi^{(m)}| |\mu_\rho^{(m)}|) - \log(\eta\theta) - 1 \right] + \sum_{\ell=1}^M \sum_{\varphi, \rho \in \Phi} |a_\varphi| |a_\rho| \int_0^1 |(g_\varphi)_\ell^{(m)}| |(g_\rho)_\ell^{(m)}|(u) \left[ \log(|(g_\varphi)_\ell^{(m)}| |(g_\rho)_\ell^{(m)}|(u)) - \log(\eta\theta) - 1 \right] du \right].$$

Consequently, using  $\|\Phi\|_\infty$  and  $r_\Phi$ ,

$$E \leq \square_{\alpha, M, f^*} \eta |\Phi| \log^2(T) \|a\|_{\ell_2}^2 + \square_{\alpha, M, f^*} \log^2(T) r_\Phi [2 \log(\|\Phi\|_\infty) - \log(\eta\theta) - 1] \|a\|_{\ell_2}^2.$$

We choose  $\eta = |\Phi|^{-1}$  and obtain that

$$E \leq \square_{\alpha, M, f^*} \log^2(T) r_\Phi [\log(\|\Phi\|_\infty) + \log(|\Phi|)] \|a\|_{\ell_2}^2.$$

Now, let us choose  $\delta = \omega / (\log^2(T) r_\Phi [\log(\|\Phi\|_\infty) + \log(|\Phi|)])$  where  $\omega$  depends only on  $\alpha, M$  and  $f^*$  and will be chosen later and let us go back to (7.24):

$$\begin{aligned} \frac{1}{T} |a'Ga - a'\mathbb{E}(G)a| &\leq \square_{\alpha, M, f^*} \omega \|a\|_{\ell_2}^2 + \square_{\alpha, f^*, \omega} r_\Phi [\log(\|\Phi\|_\infty) + \log(|\Phi|)] \sum_{\varphi, \rho \in \Phi} |a_\varphi| |a_\rho| \|\varphi\|_\infty \|\rho\|_\infty \frac{\log^5(T)}{T} \\ &\leq \square_{\alpha, M, f^*} \omega \|a\|_{\ell_2}^2 + \square_{\alpha, f^*, \omega} \|a\|_{\ell_2}^2 A_\Phi(T). \end{aligned}$$

Under assumptions of Proposition 4, for  $T_0$  large enough and  $T \geq T_0$ ,

$$\frac{1}{T} |a'Ga - a'\mathbb{E}(G)a| \leq \square_{\alpha, M, f^*} \omega \|a\|_{\ell_2}^2.$$

It is now sufficient to take  $\omega$  small enough and then  $T_0$  large enough to obtain (7.23) with  $\epsilon < \zeta$ .

### 7.5.2 Proof of Corollary 3

First let us cut  $[-1, T]$  in  $\lfloor T \rfloor + 2$  intervals  $I$ 's of the type  $[a, b)$  such that the first  $\lfloor T \rfloor + 1$  intervals are of length 1 and the last one is of length strictly smaller than 1 (eventually it is just a singleton). Then, any interval of the type  $[t - 1, t]$  for  $t$  in  $[0, T]$  is included into the union of two such intervals. Therefore the event where all the  $N_I$ 's are smaller than  $u = \mathcal{N}/2$  is included into  $\Omega_{\mathcal{N}}$ . It remains to control the probability of the complementary of this event. By stationarity, all the first  $N_I$ 's have the same distribution and satisfy Proposition 1. The last one can also be viewed as the truncation of a stationary point process to an interval of length smaller than 1. Therefore the exponential inequality of Proposition 1 also applies to the last interval. It remains to apply  $\lfloor T \rfloor + 2$  times this exponential inequality and to use a union bound.

### 7.5.3 Proof of Corollary 4

As in the proof of Proposition 2, we use the notation  $\square$ . The non-asymptotic part of the result is just a pure application of Theorem 2, with the choices of  $B_\varphi$  and  $V_\varphi$  given by (5.5) and (5.6). The next step consists in controlling the martingale  $\psi(\varphi)^2 \bullet (N - \Lambda)_T$  on  $\Omega_{V, B}$ . To do so, let us apply (7.7) to  $H$  such that for any  $m$ ,

$$H_t^{(m)} = \psi_t^{(m)}(\varphi)^2 \mathbb{1}_{t \leq \tau'},$$

with  $B = B_\varphi^2$  and  $\tau = T$  and where  $\tau'$  is defined in (7.1) (see the proof of Theorem 2). The assumption to be fulfilled is checked as in the proof of Theorem 2. But as previously, on  $\Omega_{V,B}$ ,  $H \bullet (N - \Lambda)_T = \psi(\varphi)^2 \bullet (N - \Lambda)_T$  and also  $H^2 \bullet \Lambda_T = \psi(\varphi)^4 \bullet \Lambda_T$ . Moreover on  $\Omega_{\mathcal{N}} \subset \Omega_{V,B}$

$$H^2 \bullet \Lambda_T = \psi(\varphi)^4 \bullet \Lambda_T \leq v := TM(\max_m \nu^{(m)} + \mathcal{N} \max_{m,\ell} h_\ell^{(m)})B_\varphi^4.$$

Recall that  $x = \alpha \log(T)$ . So on  $\Omega_{V,B}$ , with probability larger than  $1 - (M + KM^2)e^{-x} = 1 - (M + KM^2)T^{-\alpha}$ , one has that for all  $\varphi \in \Phi$ ,

$$\psi(\varphi)^2 \bullet N_T \leq \psi(\varphi)^2 \bullet \Lambda_T + \sqrt{2vx} + \frac{B_\varphi^2 x}{3}.$$

So that for all  $\varphi \in \Phi$ ,

$$\psi(\varphi)^2 \bullet N_T \leq \square_{M,f^*} \left[ \mathcal{N} \|\varphi\|_T^2 + \|\Phi\|_\infty^2 \mathcal{N}^2 \sqrt{T \mathcal{N} \log(T)} \right].$$

Also, since  $\mathcal{N} = \log^2(T)$ , one can apply Corollary 3, with  $\beta = \alpha$ . We finally choose  $c$  as in Proposition 4. This leads to the result.

**Acknowledgements:** We are very grateful to Christine Tuleau-Malot who allowed us to use her R programs simulating Hawkes processes. The research of Patricia Reynaud-Bouret and Vincent Rivoirard is partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration).

## References

- [1] Aalen, O. (1980) *A model for nonparametric regression analysis of counting processes*. Mathematical statistics and probability theory (Proc. Sixth Internat. Conf., Wisla, 1978). Lecture Notes in Statist, **2**, 1–25.
- [2] Andersen, Per Kragh, Borgan, Ørnulf, Gill, Richard D. and Keiding, Niels (1993) *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York.
- [3] Bercu, Bernard and Touati, Abderrahmen (2008) *Exponential inequalities for self-normalized martingales with applications*. The Annals of Applied Probability, **18**(5), 1848–1869.
- [4] Bertin, Karine, Le Pennec, Erwan and Rivoirard, Vincent (2011) *Adaptive Dantzig density estimation*, Annales de l’Institut Henri Poincaré Probabilités et Statistiques, **47**(1), 43–74.
- [5] Bickel, Peter J., Ritov, Ya’acov and Tsybakov, Alexandre B. (2009) *Simultaneous analysis of Lasso and Dantzig selector*. The Annals of Statistics, **37**(4), 1705–1732.
- [6] C. G. Bowers (2010). Stochastic kinetic models: Dynamic independence, modularity and graphs. *Annals of Statistics*, **38**(4):2242–2281.
- [7] Brémaud, Pierre (1981) *Point processes and queues*. Springer-Verlag, New York.
- [8] Brémaud, Pierre and Massoulié, Laurent (1996) *Stability of nonlinear Hawkes processes*. The Annals of Probability, **24**(3), 1563–1588.
- [9] Brunel, Elodie and Comte, Fabienne (2005) *Penalized contrast estimation of density and hazard rate with censored data*. Sankhyā, **67**(3), 441–475.
- [10] Brunel, Elodie and Comte, Fabienne (2008) *Adaptive estimation of hazard rate with censored data* Communications in Statistics. Theory and Methods, **37**(8-10), 1284–1305.
- [11] Bühlmann, Peter and van de Geer, Sara (2011) *Statistics for high-dimensional data*. Springer, Heidelberg.

- [12] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2006) *Aggregation and sparsity via  $\ell_1$  penalized least squares*, Proceedings of 19th Annual Conference on Learning Theory (COLT 2006), Lecture Notes in Artificial Intelligence v.4005 (Lugosi, G. and Simon, H.U., eds.), Springer-Verlag, Berlin-Heidelberg.
- [13] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007) *Sparse density estimation with  $l_1$  penalties*, Lecture Notes in Artificial Intelligence, **4539**, 530–543.
- [14] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007) *Aggregation for Gaussian regression*. The Annals of Statistics, **35**(4), 1674–1697.
- [15] Bunea F., Tsybakov A.B. and Wegkamp M.H. (2007) *Sparsity Oracle Inequalities for the Lasso*. Electronic Journal of Statistics, **1**, 169–194.
- [16] Candes, Emmanuel and Tao, Terence (2005) *Decoding by linear programming*, IEEE Trans. Inform. Theory **51**(12), 4203–4215.
- [17] Candes, Emmanuel and Tao, Terence (2007) *The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$* , Ann. Statist. **35**(6), 2313–2351.
- [18] Cardanobile, S., Rotter, S. *Multiplicatively interacting point processes and applications to neural modeling*, arXiv (2010).
- [19] Carstensen, Lisbeth, Sandelin, Albin, Winther, Ole and Hansen, Niels Richard (2010) *Multivariate Hawkes process models of the occurrence of regulatory elements and an analysis of the pilot ENCODE regions*. BMC Bioinformatics, **11**(456).
- [20] Comte, Fabienne, Gaïffas, Stéphane and Guilloux Agathe (2011) *Adaptive estimation of the conditional intensity of marker-dependent counting processes*. Preprint. <http://lanl.arxiv.org/abs/0810.4263>.
- [21] Daley, D. J. and Vere-Jones, D. (2003) *An introduction to the theory of point processes. Vol. I. Probability and its Applications*. Second edition. Springer-Verlag, New York.
- [22] de la Peña, Victor H. (1999) *A general class of exponential inequalities for martingales and ratios*. The Annals of Probability, **27**(1), 537–564.
- [23] de la Peña, V. H., Lai, T. L., Shao, Qi-Man *Self-Normalized Processes* Springer (2009).
- [24] Dzhaparidze, K. and van Zanten, J. H. (2001) *On Bernstein-type inequalities for martingales*. Stochastic Processes and their Applications, **93**(1), 109–117.
- [25] Fu, Wenjiang J. (1998) *Penalized regressions: the bridge versus the Lasso*. J. Comput. Graph. Statist. **7**(3) 397–416.
- [26] Gaïffas, Stéphane and Guilloux Agathe (2011) *High-dimensional additive hazard models and the Lasso*. Preprint. <http://lanl.arxiv.org/abs/1106.4662>
- [27] Grün, S., Diesmann, M., Grammont, F., Riehle, A., Aertsen, A. *Detecting unitary events without discretization in time*. Journal of Neuroscience methods **94** (1999).
- [28] Gusto, Gaelle and Schbath, Sophie (2005). *FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes’ model*. Statistical Applications in Genetics and Molecular Biology, **4**.
- [29] Härdle, W., Kerkycharian, G., Picard, D. and Tsybakov A. (1998) *Wavelets, Approximation and Statistical Applications*. Springer-Verlag, Berlin.
- [30] Hawkes, A.G. (1971) *Point spectra of some mutually exciting point processes*. Journal of the Royal Statistical Society. Series B. Methodological, **33**, 438–443.



- [31] J. Huang, S. Ma, and C.-H. Zhang. (2008) *Adaptive Lasso for sparse high-dimensional regression models*. *Statistica Sinica*, **18** 160–1618.
- [32] Jacobsen, Martin (2006) *Point process theory and applications. Marked point and piecewise deterministic processes*. Probability and its Applications. Birkhäuser Boston, Inc., Boston, MA
- [33] Vladimir Koltchinskii, Karim Lounici and Alexandre B. Tsybakov (2011). *Nuclear norm penalization and optimal rates for noisy low rank matrix completion*. Preprint. <http://lanl.arxiv.org/abs/1011.6256>
- [34] Letue, Frédérique (2000). *Modèle de Cox : Estimation par sélection de modèle et modèle de chocs bivarié*. PhD thesis.
- [35] R. Liptser, V. Spokoiny, Deviation probability bound for martingales with applications to statistical estimation, *Statist. Probab. Lett.*, 46 (2000) pp. 347–357.
- [36] Massart, P. (2007) *Concentration inequalities and model selection*. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Springer, Berlin
- [37] Meinshausen, Nicolai (2007) *Relaxed Lasso*. *Comput. Statist. Data Anal.* **52**(1) 374-393
- [38] Mohammad Shahed Masud and Roman Borisyuk (2011). Statistical technique for analysing functional connectivity of multiple spike trains. *Journal of Neuroscience Methods*, **196**(1):201 – 219.
- [39] Ogata, Yoshihiko and Akaike, Hirotugu (1982). On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes. *Journal of the Royal Statistical Society. Series B. Methodological*, **44**(1), 102–107.
- [40] Ozaki, T. (1979) *Maximum likelihood estimation of Hawkes' self-exciting point processes*. *Annals of the Institute of Statistical Mathematics*, **31**(1), 145–155.
- [41] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli 2008. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, **454**, 995–999.
- [42] Reynaud-Bouret, Patricia (2003) *Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities*, *Probability Theory and Related Fields*, **126**(1), 103–153.
- [43] Reynaud-Bouret, Patricia (2006) *Penalized projection estimators of the Aalen multiplicative intensity*, *Bernoulli*, **12**(4), 633–661.
- [44] Reynaud-Bouret, Patricia and Rivoirard, Vincent (2010) *Near optimal thresholding estimation of a Poisson intensity on the real line*. *Electronic Journal of Statistics*, **4**, 172–238.
- [45] Reynaud-Bouret, Patricia and Roy Emmanuel (2007) *Some non asymptotic tail estimates for Hawkes processes*. *Bull. Belg. Math. Soc. Simon Stevin*, **13**(5), 883–896.
- [46] Reynaud-Bouret, Patricia and Schbath, Sophie (2010) *Adaptive estimation for Hawkes processes; application to genome analysis*. *The Annals of Statistics*, **38**(5), 2781–2822.
- [47] Rudelson, Mark and Vershynin, Roman (2008) *On sparse reconstruction from Fourier and Gaussian measurements*. *Comm. Pure Appl. Math.* **61**(8), 1025–1045.
- [48] Rudelson, Mark and Vershynin, Roman (2009) *Smallest singular value of a random rectangular matrix*. *Comm. Pure Appl. Math.* **62**(12), 1707–1739.
- [49] Rudelson, Mark and Vershynin, Roman(2010) *Non-asymptotic theory of random matrices: extreme singular values*. *Proceedings of the International Congress of Mathematicians. Volume III*, Hindustan Book Agency, New Delhi, 1576–1602.

- [50] Shorack, Galen R. and Wellner, Jon A. (1986) *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.
- [51] Tibshirani, Robert. (1996) *Regression shrinkage and selection via the Lasso*. J. Roy. Statist. Soc. Ser. B **58**(1) 267–288.
- [52] van de Geer, S. (1995) *Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes*. The Annals of Statistics, **23**(5), 1779–1801.
- [53] van de Geer, S. (2008) *High dimensional generalized linear models and the Lasso*. The Annals of Statistics, **36**(2), 614–645.
- [54] van de Geer, S. and Bühlmann, P. (2009) *On the conditions used to prove oracle results for the Lasso*. Electronic Journal of Statistics, **3**, 1360–1392.
- [55] Sara van de Geer, Peter Bühlmann, Shuheng Zhou (2010) *The adaptive and the thresholded Lasso for potentially misspecified models*. Manuscript
- [56] Vere-Jones, D. and Ozaki, T. (1982) *Some examples of statistical estimation applied to earthquake data I: cyclic Poisson and self-exciting models*. Annals of the Institute of Statistical Mathematics, **34**(1), 189–207.
- [57] Zou, Hui (2006) *The adaptive Lasso and its oracle properties*. J. Amer. Statist. Assoc. **101**(476), 1418–1429.