Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis

Patricia Reynaud-Bouret, Vincent Rivoirard, Franck Grammont, Christine Tuleau-Malot

Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis

Patricia Reynaud-Bouret[*1] and Vincent Rivoirard[2] and Franck Grammont[1] and Christine Tuleau-Malot[1]

[1]Laboratoire J.-A. Dieudonné, UMR7351, UNS/CNRS, Université de Nice Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 2, France
[2]CEREMADE UMR CNRS 7534, Université Paris Dauphine, Place du Maréchal De Lattre De Tassigny, 75775 PARIS Cedex 16, France

Email: Patricia Reynaud-Bouret*- reynaudb@unice.fr; Vincent Rivoirard - rivoirard@ceremade.dauphine.fr; Franck Grammont - grammont@unice.fr; Christine Tuleau-Malot - malot@unice.fr;

*Corresponding author

Abstract

When dealing with classical spike train analysis, the practitioner often performs goodness-of-fit tests to test whether the observed process is a Poisson process, for instance, or if it obeys another type of probabilistic model [1–3]. In doing so, there is a fundamental plug-in step, where the parameters of the supposed underlying model are estimated. The aim of this article is to show that plug-in has sometimes very undesirable effects. We propose a new method based on subsampling to deal with those plug-in issues in the case of the Kolmogorov-Smirnov test of uniformity. The method relies on the plug-in of good estimates of the underlying model, that have to be consistent with a controlled rate of convergence. Some non parametric estimates satisfying those constraints in the Poisson or in the Hawkes framework are highlighted. Moreover they share adaptive properties that are useful from a practical point of view. We show the performance of those methods on simulated data. We also provide a complete analysis with these tools on single unit activity recorded on a monkey during a sensory-motor task.

1 Introduction

In neuroscience, the action potentials (spikes) are the main components for the real-time information processing in the brain. Moreover it is possible to record in vivo several neurons and to have access to simultaneous spike trains. The duration of each spike is very small, about one millisecond. Moreover the number and the position of each spike fluctuate from one trial to another trial. It is consequently quite natural to assimilate a spike to a random event. Therefore, in this article, we mathematically model spike trains as real-valued *point processes* that have been deeply described and studied for long in the literature (see [4] for a review) and often used in neuroscience (see for instance [2] and the references therein). However, except in very particular tests of independence (see for instance [5, 6]), it is most of the time necessary to describe spike trains as realizations of particular stochastic processes.

Most of the analyses start by answering a standard basic question. Is the process an homogeneous Poisson process or not? See for instance [7–9]. Indeed, for this simple model, extensively used in neuroscience, there is only one parameter to infer, namely the *firing rate*. The study of firing rates in neuroscience has lead to significative advances in the understanding of the coding of the direction of movements [10] for instance. But most of the time, spikes trains are more complex than homogeneous Poisson processes. Various studies have exhibited different kinds of correlations between some motor, sensory or cognitive events in a behaving animal and a variation of the firing rate of specific neurons, before, during or after this event [11, 12]. In particular, such data cannot be stationary. So, constraints on the previous model are relaxed and processes can be assumed to be *inhomogeneous Poisson processes*. In this setting, the firing rate is now time-dependent and is modeled by a function $\lambda(.)$ which is the intensity of the inhomogeneous Poisson process (see [8, 9]). Several studies have also established statistical evidence of dependence between the occurrences of the spikes of several neurons (see [5, 6, 13–15]) or even within a given neuron. In this case, standard homogeneous or inhomogeneous Poisson processes cannot be used and models based on *univariate or multivariate Hawkes processes* or variations of them are quite natural to capture dependence of spikes occurrences [16–21]. Hawkes processes, extensively described and discussed later on, generalize homogeneous Poisson processes by using functions quantifying interactions between spikes. These functions are called *interaction functions*. Such interaction functions are used in neuroscience to model excitation and inhibition phenomena [22].

Whatever the chosen model, this model has to be tested before any other inference based on this model. A plug-in step to infer unknown parameters is most of the time unavoidable to perform these tests. More precisely, for general models on point processes, the main ingredient consists in transforming the data so that the time changed process becomes a homogeneous Poisson process, fact which can be easily tested.

However the parameters of the transformation are usually unknown and are replaced by estimates. This plug-in trick has been widely popularized since [23]. It is widely used in neuroscience since [1] (see also the textbook of Tuckwell [24], [3] or [2]). The main goal of this article is to precisely show that the plug-in step may sometimes lead to undesirable effects and to propose the subsampling as a reasonable and quite universal solution. We focus here on the Kolmogorov-Smirnov (K.S.) test of uniformity. Indeed this K.S. test is usually considered as one of the three main tests on the first-order statistics that can be done to test the homogeneous Poisson hypothesis (see [1] and the references therein). More generally, the K.S. test (see [25] for its first use up to our knowledge) is one of the main omnibus tests [26], meaning that it is effective against a wide class of alternatives. However it is known that plug-in has to be taken with care for this test (see [27] for some brief discussion of this point). By using aggregated or cumulated tests, we propose 5 tests based on subsampling as goodness-of-fit tests, for which plug-in issues are solved. Note that, in neuroscience, plug-in problems have already been emphasized for another type of tests, namely the independence tests [22].

The second goal of this paper results from the first one: We have to develop statistical methods in the setting of point processes to estimate functions such as the intensity for the Poisson model or the interaction functions for the Hawkes model. Standard statistical procedures consist in assuming that these functions are parametrized by a few number of parameters, and in taking (for instance) the maximum likelihood estimator [28, 29]. This approach is called *parametric*. For instance assuming that a spike train is an homogeneous Poisson process, is equivalent to parametrizing the intensity by one parameter, namely the fixed constant firing rate. However in neuroscience, except in the particular case of homogeneous Poisson process, there is no a priori parametric shape for the functions to be estimated. These functions are most of the time unknown. Our second main contribution consists in proposing estimation procedures in a very flexible setting, once the probabilistic model is fixed. So we consider the setting of *nonparametric* statistics, which is designed to estimate functions when no parametric model can be assumed. In particular, this nonparametric setting allows us to weaken assumptions considerably. The estimates proposed in this paper are based on kernel rules, wavelets expansions or penalized criteria. Not only are they nonparametric but they also share the following features:

1. They are obtained by completely data-driven procedures that can be used even by neophytes in nonparametric statistics.

2. They achieve optimal convergence rates.

3. They do not assume light tails or any shape (exponential, unimodal,...) about the underlying function.

4. They adapt to the smoothness of the underlying function.

Furthermore, the developed strategies considerably extend the procedures proposed by [7,30]. In particular new data-driven kernel rules are introduced to estimate the intensity of inhomogeneous Poisson processes. We also derive a lasso-type estimate for recovering interaction functions of multivariate Hawkes processes when observing n trials. Some new interpretations of the estimate and connections with classical tools of the neuroscience literature such as joint peristimulus time histograms (JPSTH) and crosscorrelograms are also proposed. Theoretical results are established by using the *oracle* approach (see later).

The article is organized as follows. We first explain how subsampling can overcome the issues raised by plug-in for goodness-of-fit tests for the special case of the K.S. test. Then, we extensively discuss adaptive nonparametric estimation and its advantages with respect to parametric estimation. This is illustrated on Poisson or Hawkes processes and a wide range of nonparametric methods are proposed. Finally some simulations have been performed and real data sets coming from the recordings of a sensory-motor task (that can be found in [15], for instance) are analyzed thanks to these new methods. Most of the analysis has been performed with the software R. We refer to [7] for a complete list of its advantages.

Let us introduce succinctly the main notions. More mathematical insight on the subject can also be found in [31]. For more-to-the-point definitions in link with neuroscience, and heuristic interpretations, we refer the interested reader to the very limpid article of Brown et al. [2] on the time-rescaling theorem. In the sequel, a point process N is a random countable set of points. For all measurable subset A, $N(A)$ is the random variable giving the number of points of N in A. The associated point measure dN is defined as follows: for all measurable function f,

$$\int f(x)dN(x) = \sum_{T \in N} f(T).$$

To a finite point process N on the positive real line, one can associate the corresponding counting process $(N_t)_{t \geq 0} = (N([0,t]))_{t \geq 0}$ and its compensator $(\Lambda(t))_{t \geq 0}$ with respect to some given filtration (history). Most of the time, a conditional intensity $\lambda(.)$ depending on the past history exists and in this case

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

The function $\Lambda(.)$ is therefore continuous non decreasing. This is also the time-transformation on which the time-rescaling theorem is based [2]. In the sequel, $X_p \xrightarrow[p \to \infty]{\mathbb{P}} 0$ means that the sequence X_p converges in probability towards 0 when p tends to infinity; $X_p \xrightarrow[p \to \infty]{\mathcal{L}} X$ means that the distribution of X_p tends to the one of X when p tends to infinity.

4

2 Goodness-of-fit tests: the plug-in drawback and subsampling as a possible universal solution

Once spike trains have been obtained and sorted, neurophysiologists often perform a very basic data analysis, which consists in testing several features such as stationarity for instance among other statistical inferences [7]. Following Ventura et al. [8], the first step of a "good practice" is usually to test whether the observed spike train is homogeneous Poisson or not. But it is usually admitted that real spike trains cannot be that simple and this hypothesis is most of the time rejected. To explain the rejection, the next step, still following [8], is to impute it to a lack of stationarity or to something more complex. It means that we have to test whether the process is an inhomogeneous Poisson process or not. For this purpose, one uses the time-rescaling theorem (see [32] but also [4, 31]) under the hypothesis that the process is a Poisson process with deterministic intensity $\lambda(.)$. Its associated compensator $\Lambda(.)$ is in this case deterministic as well. The time-rescaling theorem, in its simplest version, states therefore that if N is a Poisson process with intensity $\lambda(.)$, observed on $[0, T_{max}]$, then $\mathcal{N} = \{X = \Lambda(T) : T \in N\}$ is an homogeneous Poisson process on $[0, \Lambda(T_{max})]$ with intensity 1, fact which can be tested by practitioners. However, there is a misspecification in the method since $\lambda(.)$ is unknown. The most popular and widely used method in neuroscience consists in plugging an estimate $\hat{\lambda}(.)$ in [8]. As explained in Introduction we first illustrate the drawbacks of non cautious plug-ins for goodness-of-fit tests on the K.S. test, which has already been observed by [27]. We then propose a remedy to overcome these drawbacks based on subsampling.

2.1 Elementary situation for illustration

Let us illustrate our purpose on a very basic situation. Assume that one observes $X_1, \ldots, X_n$ n independent and identically distributed (i.i.d.) real variables with cumulative distribution function (c.d.f.) $u \to F(u) = \mathbb{P}(X_1 \leq u)$. Given F_0 a c.d.f., we can test whether the hypothesis $H_0 :$ "$F = F_0$" is true or not. To do so, let us first define F_n the empirical distribution function associated with the X_i's by

$$u \to F_n(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq u\}}.$$

If n is large enough, $F_n(u)$ is close to $F(u)$ for any u. The K.S. test is therefore based on the statistic

$$KS_n = \sup_u |F_n(u) - F_0(u)|. \tag{1}$$

Under H_0, if F_0 is continuous, the distribution of KS_n is known and it does not depend on F_0, so it can be tabulated [27]. For any $\alpha \in (0, 1)$, let $k_{n,1-\alpha}$ be the $1 - \alpha$ quantile of this distribution. The classical

(without plug-in) K.S. test consists in rejecting H_0 whenever $KS_n > k_{n,1-\alpha}$ and this test is of exact level α. Note also that when n tends to ∞, the random variable $\sqrt{n}KS_n$ tends in distribution to a tabulated distribution $\mathcal{K}$ (see [33]). As a consequence, if $\tilde{k}_{1-\alpha}$ is the $1-\alpha$ quantile of $\mathcal{K}$, $\sqrt{n}k_{n,1-\alpha}$ tends to $\tilde{k}_{1-\alpha}$ and the approximation is valid as soon as $n > 45$ [34] (see also Durbin's modification in [27]).

Often, the c.d.f. F_0 is unknown since it depends on one or several unknown parameters and a natural idea consists in estimating it to use the previous procedure. This idea, extensively used in neuroscience, can lead to false results. For illustration, assume for example that we wish to test the hypothesis H_0 "The X_i's are exponential with unknown parameter λ". Note that this hypothesis is often tested on the interspike time intervals (ISI) [24] in order to test whether the observed spike process is an homogeneous Poisson process with unknown intensity λ. Following the scheme described previously, a natural procedure to test the exponentiality of the X_i's could be the following.

(i) Estimate λ by $\hat{\lambda} = 1/\bar{X}$, where $\bar{X}$ is the empirical mean of the X_i's: $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$.

(ii) Plug in the estimate $\hat{\lambda}$ and estimate F_0 by $u \to \hat{F}(u) = 1 - \exp(-\hat{\lambda}u)$.

(iii) Form the K.S. statistic (1) by replacing F_0 by $\hat{F}$. This leads to $KS^{(1)}$.

(iv) Reject H_0 whenever $KS^{(1)} > k_{n,1-\alpha}$.

The p-values of this test are represented in Figure 1. If the distribution of the test statistic was correctly predicted by the quantiles $k_{n,1-\alpha}$, then the repartition of the p-values should be close to the first diagonal of the graph (see [35]). Clearly the curve is above the diagonal and the test is too conservative, which means that the test will accept H_0 more than required. The previous procedure fails in obtaining good results since, roughly speaking, the same data are used to estimate λ and to compute the test statistic. For very specific c.d.f., this problem can be overcome by computing the distribution of $KS^{(1)}$ (see [27] for exponential and Gaussian cases). However, this is based on a trick that makes distributions, in those specific cases, independent of the unknown underlying parameter λ. Therefore this solution cannot be adapted to complex situations such as the inhomogeneous Poisson process framework described above in the neuroscience field [8].

To be more careful and to avoid dependencies between $\hat{\lambda}$ and F_n, we could use the following "split into two parts" procedure where n is assumed to be even.

(i) Estimate λ by $\tilde{\lambda} = 1/\bar{\bar{X}}$, where $\bar{\bar{X}}$ is the empirical mean of the first half of the X_i's: $\bar{\bar{X}} = 2/n \sum_{i=1}^{n/2} X_i$.

(ii) Plug in the estimate $\tilde{\lambda}$ and estimate F_0 by $u \to \tilde{F}(u) = 1 - \exp(-\tilde{\lambda}u)$.

(iii) Form the K.S. statistic (1) by replacing F_0 by $\tilde{F}$, but also by replacing F_n by the empirical cumulative distribution function only based on $X_{n/2+1}, \ldots, X_n$. This leads to $KS^{(2)}$.

(iv) Reject H_0 whenever $KS^{(2)} > k_{n/2, 1-\alpha}$.

The p-values of this test are represented on Figure 1. Surprisingly, the distribution of the p-values shows that the resulting test is not conservative enough. Indeed, the test will reject H_0 more than required and this procedure is even worse than the first strategy. Therefore, we turn towards a much more universal strategy, subsampling, thanks to the following result (see the Additional File 1 for the proof).

Proposition 1. *Let $X_1, \ldots, X_p$ be p i.i.d. variables with c.d.f. F assumed to be continuous. Let F_p be the associated empirical distribution. Assume that $\hat{F}$ is a consistent estimate of F such that*

$$\sqrt{p} \sup_x |\hat{F}(x) - F(x)| \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{2}$$

Then

$$\sqrt{p} \sup_x |F_p(x) - \hat{F}(x)| \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{K}.$$

Therefore it remains to find $\hat{F}$ satisfying (2). In most of the parametric cases, and in particular in the exponential case, Assumption (2) does not hold if $\hat{F}$ is based on the same data as F_p. Assumption (2) may hold if p is much smaller than n, the whole sample size, as illustrated by the following strategy.

Test 1

1. Estimate λ by $\hat{\lambda} = 1/\bar{X}$, where $\bar{X}$ is the empirical mean of the X_i's on the whole sample size n.

2. Select randomly a subsample S of the trials with cardinality $p = p(n)$, such that $p(n)/n \longrightarrow_{n\to\infty} 0$ (for instance take $p(n) = \sqrt{n}$ or $p(n) = n^{2/3}$).

3. Compute on S the following empirical cumulative distribution function:

$$\forall x \geq 0, \quad F_S(x) = \frac{1}{p} \sum_{i \in S} \mathbf{1}_{\{X_i \leq x\}}$$

4. Take $\tilde{k}_{1-\alpha}$ the $1 - \alpha$ quantile of the asymptotic distribution $\mathcal{K}$.

5. Reject H_0: "The distribution of the X_i's is exponential" whenever

$$\sqrt{p} \sup_{x \in \mathbb{R}^+} |F_S(x) - \hat{F}(x))| > \tilde{k}_{1-\alpha},$$

where for any $x \geq 0$,
$$\hat{F}(x) = 1 - e^{-\hat{\lambda}x}.$$

Technical arguments of Additional File 1 prove that the previous test is of exact level α asymptotically. More importantly, in practice this conclusion remains true even for relatively small values of n as shown in Figure 1 illustrated with $n = 40$. Even if this test is not as powerful as the one described in [27], it has the main advantage to be almost universal. It can be adapted to most of parametric situations, since the use of subsampling makes the condition (2) quite easy to fulfill.

We want now to adapt this method to the more general scheme of goodness-of-fit tests for counting processes. From now on and whatever the situation, p will always correspond to the size of a subsample, i.e. a positive integer much smaller than n the total number of observations.

2.2   Aggregated test of H_0: "The observed processes are i.i.d. Poisson processes"

To fix notation, we consider in the sequel that we observe n i.i.d. trials. Consequently, we have access to $N_1, \ldots, N_n$, n i.i.d. point processes observed on $[0, T_{max}]$ representing the n i.i.d. spike trains of a fixed recorded neuron during T_{max} seconds.

It is not possible to assess on just one realization whether a point process is a (non necessarily homogeneous) Poisson process or not since the variations of the repartition of the points between different parts of one trial can either be due to non stationarity or to more complex dependency structures that cannot be studied on just one run.

The first simple way to use the repetition of the trials is to use aggregation, which can be viewed as an empirical sum on the point processes. More precisely, the aggregated process over the processes $N_1, \ldots, N_p$ is defined by

$$N^{a,p} = \cup_{i=1,\ldots,p} N_i \text{ or equivalently } dN^{a,p} = \sum_{i=1}^{p} dN_i.$$

By classical properties of Poisson processes [4], if the processes are i.i.d. Poisson processes with compensator $\Lambda(.)$, then $N^{a,p}$ is also a Poisson process with compensator $p\Lambda(.)$. This implies that conditionally to the event $\{N^{a,p}([0, T_{max}]) = n_{tot}\}$, the observed points of $N^{a,p}$ behave like an n_{tot} i.i.d. sample of c.d.f.

$$t \to F(t) = \frac{\Lambda(t)}{\Lambda(T_{max})}.$$

Since F is unknown in our present situation, one has to estimate it, which leads to exactly the same plug-in problem as before. Fortunately, we are able to prove the following result.

Proposition 2. *Let $N_1, \ldots, N_p$ be p i.i.d. Poisson processes with compensator $\Lambda(.)$, assumed to be contin-*

uous, on $[0, T_{max}]$. Let $F_{N^{a,p}([0,T_{max}])}$ be the associated empirical distribution, defined for any x by

$$F_{N^{a,p}([0,T_{max}])}(x) = \frac{1}{N^{a,p}([0,T_{max}])} \sum_{T \in N^{a,p}} \mathbf{1}_{\{T \leq x\}}, \tag{3}$$

where $N^{a,p}$ is the aggregated Poisson process. Assume that $\hat{F}(.)$ is a consistent estimate of $F(.) = \Lambda(.)/\Lambda(T_{max})$ such that

$$\sqrt{N^{a,p}([0,T_{max}])} \sup_{x \in [0,T_{max}]} |\hat{F}(x) - F(x)| \xrightarrow[p \to \infty]{\mathbb{P}} 0. \tag{4}$$

Then

$$\sqrt{N^{a,p}([0,T_{max}])} \sup_{x \in [0,T_{max}]} \left| F_{N^{a,p}([0,T_{max}])}(x) - \hat{F}(x) \right| \xrightarrow[p \to \infty]{\mathcal{L}} \mathcal{K}.$$

Once again, subsampling (i.e. choosing p much smaller than n) gives us estimates $\hat{F}$ satisfying (4). Two different approaches lead to two distinct tests. First let us use the empirical c.d.f. on the whole sample.

Test 2

1. Take $\hat{F}$ as $F_{N^{a,n}([0,T_{max}])}$, the empirical c.d.f. of the whole aggregated process $N^{a,n}$ over the n trials (see (3)).

2. Select randomly a subsample S of the trials with cardinality $p = p(n)$, such that $p(n)/n \longrightarrow_{n \to \infty} 0$ (for instance take $p(n) = \sqrt{n}$ or $p(n) = n^{2/3}$).

3. Aggregate the p processes in S to form $N^{a,p}$ and $F_{N^{a,p}([0,T_{max}])}$ as in Proposition 2.

4. Take $\tilde{k}_{1-\alpha}$ the $1 - \alpha$ quantile of the asymptotic distribution $\mathcal{K}$.

5. Reject H_0: "The n observed processes are i.i.d. (non necessarily homogeneous) Poisson processes" whenever

$$\sqrt{N^{a,p}([0,T_{max}])} \sup_{x \in [0,T_{max}]} \left| F_{N^{a,p}([0,T_{max}])}(x) - \hat{F}(x) \right| > \tilde{k}_{1-\alpha}.$$

In Additional File 1, we prove that this test is of exact asymptotic level α, as soon as the compensator $\Lambda(.)$ is continuous and this even if $\lambda(.)$ does not exist. However its practical performance are poor (see later). A slightly more useful test can be obtained by using smoother and more elaborate estimates $\hat{F}$ satisfying (4). We obtain the following testing procedure.

9

$\overline{}$ **Test 3**

1. Select randomly a subsample S of the trials with cardinality $p = p(n)$, such that $p(n)/n$ tends to 0.

2. Use all the n observed processes to obtain a $\hat{\lambda}(.)$ such that, if the processes are Poisson processes with intensity $\lambda(.)$, one can assume that

$$\sqrt{p(n)} \int_0^{T_{max}} |\hat{\lambda}(u) - \lambda(u)| \xrightarrow[p \to \infty]{\mathbb{P}} 0. \qquad (5)$$

3. Take

$$t \to \hat{F}(t) = \frac{\int_0^t \hat{\lambda}(u)du}{\int_0^{T_{max}} \hat{\lambda}(u)du}.$$

4. Aggregate the p processes in S to form $N^{a,p}$ and $F_{N^{a,p}([0,T_{max}])}$ as in Proposition 2.

5. Take $\tilde{k}_{1-\alpha}$ the $1 - \alpha$ quantile of the asymptotic distribution $\mathcal{K}$.

6. Reject H_0: "The n observed processes are i.i.d. (non necessarily homogeneous) Poisson processes" whenever

$$\sqrt{N^{a,p}([0,T_{max}])} \sup_{x \in [0,T_{max}]} \left| F_{N^{a,p}([0,T_{max}])}(x) - \hat{F}(x) \right| > \tilde{k}_{1-\alpha}.$$

In Additional File 1, we prove that the previous test is of asymptotic level α. Note that Condition (5) can be demanding and rejection can be due to nonfulfilment of this condition. For instance, estimates $\hat{\lambda}$ based on parametric estimates on a prescribed parametric model (such as maximum likelihood estimates for instance, see [8]) fulfil (5) if the prescribed model is true but cannot fulfill this condition if the prescribed parametric model is not true. Hence, using parametric estimates in this setting lead to test both H_0 and "the prescribed parametric model is correct", which is not satisfying. Therefore it is natural to make no parametric assumption on the underlying model for $\lambda(.)$ and to try to fulfill (5) by using nonparametric estimation.

Finally as already observed by [8], aggregation can dilute the dependencies between the points. Therefore Tests 2 and 3 cannot be really powerful as we will see later.

2.3 Cumulated goodness-of-fit tests

Another way to use the repetition of the trials, is to cumulate the p processes instead of aggregating them. This difference is made more explicit in Figure 2. When processes are aggregated, points of very different trials can be very close, which can dilute dependencies between occurrences. This cannot occur for cumulated processes. With this method, we can also test models that are more general than Poisson processes.

Those general models are usually described through their conditional intensity $\lambda(.)$, which represents the probability of occurrence of a point at time t given the past before t. So, defining a model through its conditional intensity is the easiest way to model the dependence between points. For instance, when we assume the conditional intensity λ to be a deterministic function f, we are assuming independence with respect to the past. This is equivalent to assuming that the process is an inhomogeneous Poisson process with intensity $\lambda = f$. Therefore testing H_0: " the observed processes are i.i.d. with conditional intensity $\lambda = f$ and unknown deterministic function f" is equivalent to testing H_0: "the observed processes are i.i.d. inhomogeneous Poisson processes".

More generally, we wish to test a nonparametric hypothesis on the conditional intensity. An example, more developed in the next section, is the multivariate Hawkes process which models the dependence between the spikes of different neurons via several interaction functions, for which we do not want to give a parametric form. Let us give just a simple expression of this process to illustrate our set-up, with only one process. The classical self-exciting Hawkes process has conditional intensity given by

$$\lambda(t) = \lambda_f(t) = \mu + \int_{-\infty}^{t-} h(t-u)dN(u), \tag{6}$$

where μ is a positive real parameter and h a non negative integrable function with support in $\mathbb{R}_+^*$ and with $f = (\mu, h)$. For instance if the function h is supported by the interval $(0, 2]$, then the probability of occurrence at time t randomly depends on the occurrences of the process on $[t-2, t)$. Testing whether the process is a classical self-exciting Hawkes or not can be rephrased as testing whether the process has conditional intensity given by the form λ_f defined in (6), with unknown f. Other famous examples in biomedical areas such as the multiplicative Aalen intensity or the Cox model can be found in [29].

As in the previous subsection, we use the time-rescaling theorem but in a deeper way. Remember that the general time-rescaling theorem [2] states that for any point process N on $[0, T_{max}]$ with compensator $\Lambda(.)$, the point process $\mathcal{N} = \{X = \Lambda(T) : T \in N\}$ is a Poisson process with intensity 1 on $[0, \Lambda(T_{max})]$. Therefore it is more interesting to cumulate the processes after time-rescaling than in the usual time space $[0, T_{max}]$. For general conditional intensity models, $\Lambda(.)$ is random. Therefore the state space $[0, \Lambda(T_{max})]$ is also random in general. Moreover when we are dealing with p i.i.d. processes $N_1, \ldots, N_p$, each N_i has a different compensator $\Lambda_i(.)$ which depends on the history of the ith trial. So except in the Poisson case where $\Lambda(.)$ is deterministic, we do not apply the same transformation to all the points. We finally have to deal with p processes $\mathcal{N}_i = \{X = \Lambda_i(T) : T \in N_i\}$ that are Poisson processes of intensity 1, and whose occurrences lie in $[0, \Lambda_i(T_{max})]$. Even if the $\Lambda_i(T_{max})$ are i.i.d., they are not equal in general.

11

This leads to two main remarks. First it is not possible to aggregate in general the time-transformed processes since we would aggregate processes with different lengths (see Figure 2). Therefore, Tests 2 and 3 cannot be transferred to the most general case straightforwardly. However one can cumulate those processes as done in Figure 2 and this even if the intervals have different lengths. The resulting process $\mathcal{N}^{c,p}$ is therefore a Poisson process with intensity 1 on the random interval $\mathcal{I} = [0, \sum_{i=1}^{p} \Lambda_i(T_{max})]$ (see also Additional File 1 for a more precise formula and a proof of this statement). The second remark consists in noting that $\sum_{i=1}^{p} \Lambda_i(T_{max})$ being a random quantity, it is not true in general that conditionally to the total number of points in $\mathcal{I}$, the points of $\mathcal{N}^{c,p}$ behave like an i.i.d. uniform sample and in the sequel, we shall need to restrict ourselves to an interval of the form $[0, p\theta]$ with a deterministic bound $p\theta$ which is, with high probability, smaller than $\sum_{i=1}^{p} \Lambda_i(T_{max})$.

Besides we have to deal with estimation of unknown transformations $\Lambda_i(.)$. For this purpose, we introduce estimates of the type $t \to \hat{\Lambda}_i(t) = \int_0^t \hat{\lambda}_i(u)du$, where $\hat{\lambda}_i(.)$ estimates $\lambda_i(.)$, the conditional intensity of the ith process N_i. We obtain a cumulate process $\hat{\mathcal{N}}^{c,p}$ built from the $\hat{\Lambda}_i(.)$'s. We have the following equivalent to Proposition 2.

Theorem 1. *Let $N_1, \ldots, N_p$ be p i.i.d. processes with respective conditional intensity $\lambda_i(.)$. Assume that there exist nonnegative estimates $\hat{\lambda}_i(.)$ of $\lambda_i(.)$ such that*

$$p^{-1/2} \left(\sum_{i=1}^{p} \int_0^{T_{max}} \left| \hat{\lambda}_i(u) - \lambda_i(u) \right| du \right) \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{7}$$

Then, for all $\theta > 0$ such that $\mathbb{E}(\Lambda_1(T_{max})) > \theta$,

$$\sqrt{\hat{\mathcal{N}}^{c,p}([0, p\theta])} \sup_{u\in[0,1]} \left| \frac{1}{\hat{\mathcal{N}}^{c,p}([0, p\theta])} \sum_{X\in\hat{\mathcal{N}}^{c,p}, X\leq p\theta} \mathbf{1}_{\{X/(p\theta)\leq u\}} - u \right| \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{K}.$$

It is now easy to turn this result into an operational test, using subsampling.

Test 4

1. Select randomly a subsample S of the trials with cardinality $p = p(n)$, such that $p(n)/n$ tends to 0.

2. Use all the n observed processes to obtain a $\hat{f}$ such that if $\lambda = \lambda_f$, one can assume that

$$p(n)^{-1/2} \left(\sum_{i=1}^{p(n)} \int_0^{T_{max}} \left| (\lambda_{\hat{f}})_i(u) - (\lambda_f)_i(u) \right| du \right) \xrightarrow[n \to \infty]{\mathbb{P}} 0. \qquad (8)$$

3. For all i in S, take

$$\hat{\Lambda}_i(t) = \int_0^t ((\lambda_{\hat{f}})_i(u))_+ du,$$

and change time of the ith process accordingly to obtain $\hat{\mathcal{N}}_i = \{X = \hat{\Lambda}_i(T) : T \in N_i\}$ on $[0, \hat{\Lambda}_i(T_{max})]$.

4. Cumulate the p processes $\hat{\mathcal{N}}_i$ for i in S to form $\hat{\mathcal{N}}^{c,p}$.

5. Take $\tilde{k}_{1-\alpha}$ the $1 - \alpha$ quantile of the asymptotic distribution $\mathcal{K}$.

6. Fix $\theta > 0$, strictly smaller than $p^{-1} \sum_{i \in S} \hat{\Lambda}_i(T_{max})$.

7. Reject H_0: "The n observed processes are i.i.d. processes with conditional intensity of the form λ_f and unknown f" whenever

$$\sqrt{\hat{\mathcal{N}}^{c,p}([0, p\theta])} \sup_{u \in [0,1]} \left| \frac{1}{\hat{\mathcal{N}}^{c,p}([0, p\theta])} \sum_{X \in \hat{\mathcal{N}}^c, X \leq p\theta} \mathbf{1}_{\{X/(p\theta) \leq u\}} - u \right| > \tilde{k}_{1-\alpha}.$$

In Additional File 1, we prove that the previous test is of exact asymptotic level α as soon as $\mathbb{E}(\Lambda_i(T_{max})) > \theta$. There exists a simpler form of this test when dealing with Poisson processes since in this case compensators do not depend on i.

13

Test 5

1. Select randomly a subsample S of the trials with cardinality $p = p(n)$, such that $p(n)/n$ tends to 0.

2. Use all the n observed processes to obtain a $\hat{\lambda}(.)$ such that, if the processes are Poisson processes with intensity $\lambda(.)$, one can assume that

$$\sqrt{p(n)} \left(\int_0^{T_{max}} \left| \hat{\lambda}(u) - \lambda(u) \right| du \right) \xrightarrow[n\to\infty]{\mathbb{P}} 0. \qquad (9)$$

3. Take
$$\hat{\Lambda}(t) = \int_0^t (\hat{\lambda}(u))_+ du,$$

and change time of the ith process accordingly to obtain $\hat{\mathcal{N}}_i = \{X = \hat{\Lambda}(T) : T \in N_i\}$ on $[0, \hat{\Lambda}(T_{max})]$.

4. Cumulate the p processes in S to form $\mathcal{N}^{c,p}$.

5. Take $\tilde{k}_{1-\alpha}$ the $1 - \alpha$ quantile of the asymptotic distribution $\mathcal{K}$.

6. Fix $\theta > 0$, strictly smaller than $\hat{\Lambda}(T_{max})$.

7. Reject H_0: "The n observed processes are i.i.d. non necessarily stationary Poisson processes" whenever

$$\sqrt{\hat{\mathcal{N}}^{c,p}([0, p\theta])} \sup_{u \in [0,1]} \left| \frac{1}{\hat{\mathcal{N}}^{c,p}([0, p\theta])} \sum_{X \in \hat{\mathcal{N}}^{c,p}, X \leq p\theta} \mathbf{1}_{\{X/(p\theta) \leq u\}} - u \right| > \tilde{k}_{1-\alpha}.$$

This test, as a special case of Test 4, is also of exact asymptotic level α as soon as $\Lambda(T_{max}) > \theta$. Tests 4 and 5 are more powerful to detect dependencies or to reject the Poisson assumption than Tests 2 or 3, as we will see later.

As for Test 3, and for exactly the same reasons, we want to find non parametric estimates satisfying (8) or (9). We provide in the next section powerful tools to deal with this problem and theoretical guarantees of performance of these estimates.

3 Non parametric and adaptive estimation
3.1 Why is adaptive estimation useful ?

Nonparametric estimation, and in particular nonparametric estimation of Poisson process intensity, is at the root of most of the data analyses performed on spike trains. Indeed, peristimulus time histograms (PSTH) [36] are usually the first graphical representations of an experiment. Those histograms have usually a fixed

length for each interval (typically 10 ms) and are quite noisy from a statistical point of view (see for instance the representations of [8]). Therefore there have been several attempts to provide smoother estimates, either by kernel estimates (see for instance [30]) or by projection on an orthonormal basis (see for instance [7] for the use of splines). These methods provide a first illustration of the data with as few assumptions as possible on the underlying "true" firing rate. They are originally not linked at all to any statistical or probabilistic models and constitute descriptive statistics. In particular no parametric assumption on the underlying intensity is made at this step, the parametric model and its associated maximum likelihood estimator (MLE) being given in a second time once the shape of the curve is guessed [8]. Because of this lack of parametric assumption, those estimates seem to be the best candidates at first glance for the estimate $\hat{\lambda}$ that needs to be plugged in Test 3 or 5.

However, the problem of the convergence rate remains. In all these methods, there is a tuning parameter that needs to be chosen: it is the length of the interval for histograms, the bandwidth in kernel rules or the number of coefficients in the orthonormal expansion. The problem of the choice of this parameter has first been tackled very roughly in the neuroscience literature by choosing a fixed value. On the real data presented here or on the ones in [8], it was usually considered that a bandwidth of 50 or 100 ms was a good choice. However such a very rough choice cannot guarantee a convergence rate when n goes to ∞. Indeed let us look more closely at the kernel estimate.

For the i.i.d. observed point processes $N_1, \ldots, N_n$, the kernel estimate with kernel K and bandwidth h is given by

$$\hat{\lambda}_n^{K_h}(x) = \frac{1}{n} \sum_{i=1}^{n} \int K_h(x-u) dN_i(u) = \frac{1}{n} \int K_h(x-u) dN^{a,n}(u), \tag{10}$$

where $N^{a,n}$ is the corresponding aggregated process and where $K_h(u) = (1/h)K(u/h)$. If we assume that the observed processes are inhomogeneous Poisson processes with intensity λ, $N^{a,n}$ is also an inhomogeneous Poisson process with intensity $n\lambda$ and therefore

$$\mathbb{E}[\hat{\lambda}_n^{K_h}(x)] = (K_h \star \lambda)(x), \quad \forall x \in \mathbb{R},$$

where $\star$ denotes the convolution product. So, the expectation of $\hat{\lambda}_n^{K_h}$ constitutes a regularized approximation of $\lambda(.)$. To measure the performance of $\hat{\lambda}_n^{K_h}$, we compute its $\mathbb{L}_2$-risk (see further details in Additional File 2):

$$\mathbb{E} \|\hat{\lambda}_n^{K_h} - \lambda\|_2^2 = \|K_h \star \lambda - \lambda\|_2^2 + \frac{\|\lambda\|_1}{nh} \|K\|_2^2, \tag{11}$$

which is classically interpreted as a bias-variance decomposition. Therefore if h is fixed, the variance term goes to 0 whereas the bias remains fixed so that the $\mathbb{L}_2$-risk of the estimate does not go to 0. Consequently

a fixed choice for the bandwidth is not convenient and it is essential to choose $h = h(n)$ tending to 0 with n. The dependence of h with respect to n is a problem that has been extensively studied in the density framework, a setting close to the present one since conditionally to the total number of points, the observed points of $N^{a,n}$ behave like an i.i.d. sample of density $\lambda(.)/\Lambda(T_{max})$. We refer the reader to [37] for a review. The main conclusion of such a study is that if λ is regular and if the regularity is known then we are able to choose $h(n)$ such that the $\mathbb{L}_2$-risk tends to 0 at a known rate of convergence depending on the regularity. Furthermore, the larger the regularity, the faster the rate. Typically, if the rth derivative of λ is bounded in the $\mathbb{L}_2$ sense, then it is possible to choose K and[1] $h(n) \asymp n^{-\frac{1}{2r+1}}$ such that the $\mathbb{L}_2$-risk behaves as

$$\mathbb{E} \left\| \hat{\lambda}_n^{K_h} - \lambda \right\|_2^2 \asymp n^{-\frac{2r}{2r+1}}.$$

In this setting, this choice can be applied to Tests 3 and 5, since the Markov inequality implies that (5) or (9) are satisfied with $p(n) = n^\delta$ and $\delta < (2r)/(2r+1)$. The choice $r = 1$ gives $\delta < 2/3$ and $r = 2$ gives $\delta < 4/5$.

Of course, in practice the choice of the bandwidth is capital. Since the smoothness of λ is unknown, the practitioner cannot use the previous choice. Furthermore, guessing the order of magnitude of $h(n)$ is not enough to achieve good performance since the leading constant plays an essential role. Hence the theoretical considerations developed before do not solve the practical problem. Several directions have been proposed to overcome this problem. One of the most famous ones consists in using leave-one-out or other cross-validation methods [30, 38]: among a finite family of fixed bandwidths, such methods choose the best one in an asymptotic setting. However, up to our knowledge, nothing can be said when the family of bandwidths is not fixed and some bandwidths tend to 0 with n. It is not clear at all that the resulting estimate achieves a prescribed rate and therefore it cannot be used for the proposed tests in particular. Other methods based on the rule of the thumb (and variations of it) have been proposed in the density or the Poisson setting [8, 39], and in this case the resulting bandwidth is of the form $h(n) = Cn^{-1/5}$ for various possible choices of the constant C. Generally those choices lead to poor results as noted by [8] (see also our the simulation study).

Adaptive estimation [37] aims at tuning in a data-driven way the unknown parameters of those methods (kernels, histograms,...) such that the resulting estimate has good practical performance and a guaranteed convergence rate. The adaptive estimates are usually mathematically proved to achieve the best possible rate of convergence and this even if the regularity is unknown. Moreover, they do not depend on any restrictive assumption such as, for instance, some parametric assumption. The only assumption lies in the

---

[1] If $(a_n)_{n \geq 0}$ and $(b_n)_{n \geq 0}$ are two sequences, $a(n) \asymp b(n)$ means that there exists two positive constants c_1, c_2 such that for all $n \geq 0$, $c_1 a_n \leq b_n \leq c_2 a_n$.

underlying probabilistic model (for instance, one assumes that the processes are inhomogeneous Poisson processes). Their reconstructions are therefore much more trustworthy than other methods for which those extra assumptions may not be fulfilled. As a conclusion, adaptive estimates constitute ideal candidates to be plugged in Tests 3, 4 or 5.

The main aim of next subsections is therefore to present adaptive estimates in the Poisson or in the Hawkes model that will have these good properties.

3.2 Adaptive estimation for Poisson processes
3.2.1 Kernel estimates

As mentioned previously, the Poisson setting is very close to the density setting. In the density setting, the main adaptive method for choosing a bandwidth is the Lepski method, which has been recently updated to the multidimensional framework and to deal with the problem of choosing the leading constant of the bandwidth. Due to Goldenshluger and Lepski [40] it is referred in the sequel as the GL method. We propose here to adapt this method to the Poisson setting in the following way and to prove its adaptive properties.

We consider a set of bandwidths $\mathcal{H}$ and their corresponding kernel estimates $\hat{\lambda}_n^{K_h}$. The bias-variance decomposition shows that the parameter $\bar{h}$ which minimizes the right hand side of (11) with respect to $h \in \mathcal{H}$ is the best possible choice. It is called the oracle bandwidth: since it depends on $\lambda(.)$, it cannot be used in practice. To propose a data-driven choice of the bandwidth by a GL method, we define for any $h, h' \in \mathcal{H}$:

$$\hat{\lambda}_n^{h,h'}(x) := \frac{1}{n} \sum_{T \in N^{a,n}} (K_h \star K_{h'})(x - T) = (K_h \star \hat{\lambda}_n^{K_{h'}})(x), \quad \forall x \in \mathbb{R},$$

then for $\eta > 0$, we set:

$$A(h) := \sup_{h' \in \mathcal{H}} \left\{ \|\hat{\lambda}_n^{h,h'} - \hat{\lambda}_n^{K_{h'}}\|_2 - \frac{(1+\eta)(1+\|K\|_1)\|K\|_2\sqrt{N^{a,n}([0,T_{max}])}}{n\sqrt{h'}} \right\}_+ .$$

The Additional File 2 shows that $A(h)$ constitutes a good estimate of the bias term (see (11)). Finally, we select the data-driven bandwidth as follows:

$$\hat{h} := \arg\min_{h \in \mathcal{H}} \left\{ A(h) + \frac{(1+\eta)(1+\|K\|_1)\|K\|_2\sqrt{N^{a,n}([0,T_{max}])}}{n\sqrt{h}} \right\}, \tag{12}$$

which allows us to estimate $\lambda(.)$ by using

$$\hat{\lambda}_n^{GL} := \hat{\lambda}_n^{K_{\hat{h}}}. \tag{13}$$

Note that in (12), $\|K\|_2^2 N^{a,n}([0,T_{max}])/(n^2 h)$ is an unbiased estimate of the variance term in (11) and therefore the previous criterion mimics the bias-variance decomposition of the risk of $\hat{\lambda}_n^{K_h}$ up to some

multiplicative constant. Once K, $\mathcal{H}$ and η are chosen, we obtain a turnkey procedure. The following theoretical result justifies our procedure.

Theorem 2. *If $\mathcal{H} \subset \{D^{-1} : D = 1, ..., D_{\max}\}$ with $D_{\max} = \delta n$ for some $\delta > 0$, and if $\|\lambda\|_\infty < \infty$, then,*

$$\mathbb{E} \|\hat{\lambda}_n^{GL} - \lambda\|_2^2 \leq C_1 \inf_{h \in \mathcal{H}} \left\{ \|K_h \star \lambda - \lambda\|_2^2 + \frac{\|\lambda\|_1}{nh} \|K\|_2^2 \right\} + C_2 n^{-1},$$

where C_1 is a constant depending on $\|K\|_1$ and η and C_2 is a constant depending on δ, η, $\|K\|_2$, $\|K\|_1$, $\|\lambda\|_1$ and $\|\lambda\|_\infty$.

Theorem 2 combined with (11) shows that our procedure mimics the performance of the oracle up to the constant C_1 and up to the term $C_2 n^{-1}$ which is negligible when n goes to $+\infty$. It is classically called an oracle inequality, which is the main property of adaptive estimates. In particular, one can take the family $\mathcal{H} = \{1, ..., \lfloor \delta n \rfloor^{-1}\}$ which grows with n and it is possible to select a bandwidth tending to 0 with n. If the rth derivarive of $\lambda(.)$ is bounded in $\mathbb{L}_2$, then the choice $h(n) \asymp n^{-\frac{1}{2r+1}}$ is in the family $\mathcal{H}$ and the oracle inequality gives straightforwardly that

$$\mathbb{E} \|\hat{\lambda}_n^{GL} - \lambda\|_2^2 \asymp n^{-\frac{2r}{2r+1}},$$

which is the optimal rate of convergence over such spaces. This rate is achieved, even if we do not know in advance the regularity r of λ, which is from a theoretical point of view the main improvement with respect to the theory described in the previous subsection.

If K is the Gaussian kernel, then $\|K\|_1 = 1$ and $\|K\|_2 = 2^{-1/2}\pi^{-1/4}$. Moreover $K_h \star K_{h'} = K_{\sqrt{h^2 + h'^2}}$ and a straightforward computation shows that explicit formula for $\|\hat{\lambda}_n^{h,h'} - \hat{\lambda}_n^{K_{h'}}\|_2$ are also available. It is consequently very easy to implement the method, the computational cost being almost of the same order as cross-validation. We will see in the simulation study that this practical choice is also quite robust.

3.2.2 Histograms

In the Poisson set-up, there are several ways to select data-driven partitions that lead to adaptive histogram estimates. For instance, one can use model selection as in [41]. Model selection can either select a regular partition or an irregular partition on a grid. When regular partitions are considered, the resulting estimator satisfies an oracle inequality similar to the oracle inequality established in Theorem 2 for the kernel rule combined with the GL method. Indeed the bin for the histograms plays exactly the same role as the kernel bandwidth. Therefore it leads to similar theoretical performance, except that the histograms cannot become smooth enough to guarantee an optimal convergence rate for regular intensities (namely $r > 1$). Therefore,

the choice of regular partitions is probably not the best one and one may prefer the GL method. The data-driven choice of the partition becomes much more interesting when the partition is not forced to be regular. Indeed irregular partitions can capture a fast increase of the firing rate followed very quickly by a fast decrease at some particular moment of the experiment, without leading to too noisy estimates as the classical PSTH, since over smoother periods, the length of the interval can be much larger. However, the method of [41] is too time consuming to be really considered in practice. Another possible direction is the context of Markov modulated Poisson processes [42], where the algorithms are also quite time consuming without ensuring any adaptive property in terms of convergence rate (despite some possible interpretation with respect to hidden Markov processes).

However, and as already noticed in [41], it is possible in certain cases to interpret a model selection estimate as a thresholding rule. We hereafter illustrate in a simpler case, the method developed in [43]: If $\lambda(.) \in \mathbb{L}_2$, we can decompose it on the Haar basis:

$$\lambda = \sum_{j=-1}^{+\infty} \sum_{k \in \mathbb{Z}} \beta_{j,k} \psi_{j,k},$$

where $\psi_{-1,k}(.) = \phi(. - k)$ with $\phi = \mathbf{1}_{[0,1)}$ the Haar father wavelet and where $\psi_{j,k}(.) = 2^{j/2} \psi \left(2^j(. - k) \right)$ for $j \geq 0$ with $\psi = \mathbf{1}_{[0,1/2)} - \mathbf{1}_{[1/2,1)}$ the Haar mother wavelet. The $\beta_{j,k}$'s are the unknown coefficients of $\lambda(.)$ and are given by

$$\forall j \geq -1, k \in \mathbb{Z}, \quad \beta_{j,k} = \int \psi_{j,k}(x)\lambda(x)dx.$$

These coefficients can therefore be unbiasedly and consistently estimated by

$$\forall j \geq -1, k \in \mathbb{Z}, \quad \hat{\beta}_{j,k} = \frac{1}{n} \int \psi_{j,k}(x)dN^{a,n}(x).$$

Given a fixed finite subset of indices m, we obtain an easily computable estimate of $\lambda(.)$:

$$\hat{\lambda}_n^m = \sum_{(j,k) \in m} \hat{\beta}_{j,k} \psi_{j,k}.$$

Since the Haar basis is piecewise constant, the previous estimate is also piecewise constant on a certain partition $\mathcal{P}$ depending on m. A data-driven choice of m therefore leads to a data-driven choice of the partition that can be irregular. Let us fix an arbitrary highest level of resolution j_0 such that $2^{j_0} \leq n < 2^{j_0+1}$ and let us consider the $\mathbb{L}_2$-risk of $\hat{\lambda}_n^m$ such that if $(j, k) \in m$ then $j \leq j_0$. The bias-variance decomposition

19

of $\hat{\lambda}_n^m$ can be written as follows:

$$
\begin{aligned}
\mathbb{E}[\|\hat{\lambda}_n^m - \lambda\|^2] &= \sum_{(j,k)\notin m} \beta_{j,k}^2 + \sum_{(j,k)\in m} \mathrm{Var}(\hat{\beta}_{j,k}) \\
&= \sum_{j>j_0} \sum_k \beta_{j,k}^2 + \sum_{j\leq j_0} \sum_k [\beta_{j,k}^2 \mathbf{1}_{(j,k)\notin m} + v_{j,k}\mathbf{1}_{(j,k)\in m}],
\end{aligned}
\tag{14}
$$

where

$$
v_{j,k} := \mathrm{Var}(\hat{\beta}_{j,k}) = \frac{1}{n}\int \psi_{j,k}^2(x)\lambda(x)dx.
$$

Hence the best subset m is the set of indices (j,k) such that $\beta_{j,k} > \sqrt{v_{j,k}}$. This is the oracle choice. A possible data-driven way to choose the indices (j,k) is to choose the indices such that $\hat{\beta}_{j,k}$ are larger than a certain threshold η_{jk} depending on an estimate of the variance $v_{j,k}$. The choice advertised in practice in [43] is:

$$
\eta_{j,k} = \sqrt{2\gamma \ln(n)\hat{v}_{j,k}} + \frac{\gamma \ln(n)2^{j/2}}{3n} \quad \text{where} \quad \hat{v}_{j,k} = \frac{1}{n^2}\int \psi_{j,k}^2(x)dN^{a,n}(x).
\tag{15}
$$

Then, we obtain the following thresholding estimator:

$$
\hat{\lambda}_n^{Th} = \sum_{j=-1}^{j_0} \sum_k \hat{\beta}_{j,k}\mathbf{1}_{\{|\hat{\beta}_{j,k}|>\eta_{j,k}\}}\psi_{j,k}.
\tag{16}
$$

In [43], it has been proved that a slight modification of this estimate satisfies an oracle inequality in the same spirit as Theorem 2. It also generalizes this estimate by considering general biorthogonal bases instead of the Haar basis, leading to smooth estimates (see [43, 44]). In this case, for a slight modification of the threshold, the resulting estimate has the same convergence rates as the kernel estimate combined with the GL method, up to some logarithmic term, as soon $\gamma > 1$. The choice $\gamma < 1$ has been shown to lead to bad convergence rates and the choice $\gamma = 1$ has been shown to work well on extensive simulations in both [43, 44]. This method is easily implementable leading to very fast algorithms that are in particular faster than algorithms based on the GL method.

3.2.3   More sophisticated procedures

Thresholding rules and irregular partitions overcome a drawback of kernel estimates that suffer from a lack of spatial adaptivity on the time axis. Several attempts have been proposed to build more local choices of the bandwidth (see [30] for instance), but up to our knowledge no mathematical proof of this spatial adaptation has been established, whereas histograms and in particular the previous Haar thresholding estimator can adapt the length of the bin to the heterogeneity of the data. But the resulting estimator is not smooth at

all. As explained, we can consider a smoother wavelet basis, but this extension does not completely address the issue.

The best alternative, up to our knowledge, when the support of $\lambda(.)$ is known and bounded (here $[0, T_{max}]$) and when $\lambda(.)$ does not vanish for a significant period of time, is due to Willett and Nowak [45]. Their method is quite intricate to describe. Informally, a penalized log-likelihood criterion is used to select a piecewise polynomial. Both the partition and the degree of each polynomial on each interval of the partition are free (on a very refined grid of resolution). Willett and Nowak have proved that such an estimator achieves optimal rates of convergence for various classes of regularity and in an adaptive way. From a practical point of view, a dyadic tree algorithm is used. Its complexity is much smaller than a full model selection method on the same piecewise polynomial family of models. It is a bit more complex than a thresholding algorithm, but there exist a program (`FreeDegree`) in `Matlab` interfaced with `C` which makes its use in practice quite easy. For a more complete description of the method, we refer to [45]. Note that in practice because of its adaptive properties, this estimator is able to be piecewise constant when the true intensity is piecewise constant but also very smooth (with high degree for the polynomials) when the underlying intensity is smooth and when the number of points is sufficient. It is also able to be spatially adaptive, the underlying data-driven partition being irregular. In the sequel, we denote this method $\hat{\lambda}_n^{WN}$.

3.3 Adaptive estimation for Hawkes processes

If inhomogeneous Poisson processes can model non-stationary data, they are not appropriate to model dependencies between points. However, several studies have established potential dependence of spike occurrences for different neurons. This has been detected via descriptive statistics, via independence tests for a given fixed model or via model-free independence tests based on permutations (also called trials-shuffling) [5, 6, 13, 15, 22, 46].

One simple model of dependency is the multivariate Hawkes process, which is the point process equivalent to the auto-regressive model. It has first been introduced by Hawkes [47], as a self-exciting point process, that is useful in particular in sismology (see for instance [23]). It has also been used to model positions of motifs along the DNA molecule [48, 49]. In neuroscience, it explicitly appears in the 80's with [19] and is close in spirit to [50, 51], with the additional advantage of modelling potential feed-back between the neurons.

The multivariate Hawkes process (see for instance [52] or [4]) models the instantaneous firing rates of M different neurons, with spike trains $N^{(1)}, \ldots, N^{(M)}$, where the conditional intensity of the mth point process

is defined for any $t \geq 0$ by

$$\lambda^{(m)}(t) = \left(\nu^{(m)} + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_{\ell}^{(m)}(t-u)dN^{(\ell)}(u) \right)_{+} = \left(\nu^{(m)} + \sum_{\ell=1}^{M} \sum_{T_{\ell} \in N^{(\ell)}, \; T_{\ell}<t} h_{\ell}^{(m)}(t-T_{\ell}) \right)_{+}. \quad (17)$$

In (17), the $\nu^{(m)}$'s are positive parameters representing the spontaneous firing rates and the $h_{\ell}^{(m)}$'s are the interaction functions and have support included into $\mathbb{R}_{+}^{*}$. More precisely, before the first occurrence of the multivariate process, the $N^{(m)}$'s behave like homogeneous Poisson processes with constant intensities $\nu^{(m)}$. The first occurrence (and the next ones) affects all the processes by increasing or decreasing the conditional intensity via the interaction functions $h_{\ell}^{(m)}$'s. For instance, if $h_{\ell}^{(m)}$ takes large positive values in the neighborhood of the delay d and is null elsewhere, then after the delay d of one occurrence of $N^{(\ell)}$, the probability to have a new occurrence of $N^{(m)}$ will significantly increase: the process $N^{(\ell)}$ excites the process $N^{(m)}$. On the contrary, if $h_{\ell}^{(m)}$ is negative around d, then after the delay d of one occurrence of $N^{(\ell)}$, the probability to have a new occurrence of $N^{(m)}$ will significantly decrease: the process $N^{(\ell)}$ inhibits the process $N^{(m)}$. Note in particular that the functions $h_{m}^{(m)}$'s model self-interactions.

The Hawkes process as described above cannot really model non-stationary data. Indeed, when t grows (and under conditions on the interaction functions), the process converges quite quickly towards an equilibrium, which is stationary (see for instance [52], [53] and the references therein). If these conditions are not satisfied, the number of points in the process grows too fast to be a realistic model for spike trains anyway. Hence Hawkes processes as defined in (17) cannot model non-stationary data but can model dependent data.

Therefore we fix an interval $[T_1, T_2] \subset [0, T_{max}]$, typically an interval where all the estimated mean firing rates seem constant. The aim is to estimate on this interval

$$f^* = \left((\nu^{(m)})_{m=1,\ldots,M}, (h_{\ell}^{(m)})_{\ell,m=1,\ldots,M} \right),$$

where it is assumed that the interaction functions are bounded with support in $[0, A]$ with $T_1 > A$.

Inference for Hawkes models based on the likelihood can be found in the literature, in particular for parametric models [23,49]. However in neuroscience, for flexibility, the used parametric models are based on a large number of parameters. Therefore they require several thousand spikes per neuron to be observed in a stationary way to achieve good estimation [19]. Classical model selection based on AIC and BIC criteria have also been used to select the number of knots for the spline estimate [21,48,54]. However these criteria do not adapt well to irregular functions. This is the reason why alternative nonparametric adaptive inference has recently been developed in such models. The univariate case ($M = 1$) has been studied in [55], where rates of convergence depending on the underlying regularity of the self-interaction function have been derived.

We can also mention the alternatives proposed in [20, 56] but no theoretical validation is provided in those works.

A multivariate approach, valid for very general counting processes including Hawkes processes and based on ℓ_1 penalties, has been recently developed in [57]. Based on minimization of convex criteria, its computational cost is more reasonable than procedures proposed in [55] and it is also proved to satisfy oracle inequalities. We shall detail this method in the case of Hawkes processes and with piecewise constant estimates of the underlying interaction functions.

In the next section that can be skipped at first reading, we describe the method in a technical way. Then we give heuristic arguments to understand more deeply the presented method (see also [58] for a quicker view on this estimate). In particular the method does not rely on the likelihood but on a least-square contrast, which can be reinterpreted in terms of JPSTH [59].

3.3.1 *Intensity candidates and least-square contrast on one trial*

We first propose a conditional intensity candidate. So for any $f \in \mathcal{H}$ with

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

we consider the predictable transformation $\psi(f) = (\psi^{(1)}(f), \ldots, \psi^{(M)}(f))$ such that

$$\forall t > 0, \quad \psi_t^{(m)}(f) = \mu^{(m)} + \sum_{\ell=1}^M \int_{-\infty}^{t-} g_\ell^{(m)}(t-u) dN_\ell(u). \tag{18}$$

Note that $\lambda^{(m)} = [\psi^{(m)}(f^*)]_+$. Therefore, for each m, $\psi^{(m)}(f)$ can be considered as a good intensity candidate as long as it is close enough to the conditional intensity $\lambda^{(m)}$ (even if $\psi^{(m)}(f)$ takes negative values). We measure the distance between $\psi(f)$ and λ by using the classical $\mathbb{L}_2$-norm $\|\cdot\|$:

$$\|\psi(f) - \lambda\|^2 = \sum_{m=1}^M \int_{T_1}^{T_2} [\psi_t^{(m)}(f) - \lambda^{(m)}(t)]^2 dt = \sum_{m=1}^M \int_{T_1}^{T_2} [\psi_t^{(m)}(f) - [\psi_t^{(m)}(f^*)]_+]^2 dt. \tag{19}$$

Depending on f^*, the right hand side is not observable. But minimizing the last expression with respect to f is equivalent to minimizing $f \longmapsto \tilde{\gamma}(f)$ with

$$\tilde{\gamma}(f) = -2 \sum_{m=1}^M \int_{T_1}^{T_2} \psi_t^{(m)}(f) \lambda^{(m)}(t) dt + \sum_{m=1}^M \int_{T_1}^{T_2} [\psi_t^{(m)}(f)]^2 dt.$$

But by definition of the conditional intensity, $\tilde{\gamma}(f)$ is close to $\gamma(f)$ with

$$\gamma(f) = -2 \sum_{m=1}^M \int_{T_1}^{T_2} \psi_t^{(m)}(f) dN^{(m)}(t) + \sum_{m=1}^M \int_{T_1}^{T_2} [\psi_t^{(m)}(f)]^2 dt, \tag{20}$$

23

which is called the least-square contrast. This expression is observable and can be minimized if f is parametrized by a fixed number of parameters.

One particular parametrization, that is used in practice, is obtained when each function $g_\ell^{(m)}$ is a piecewise constant function written as

$$g_\ell^{(m)} = \sum_{k=1}^{K} a_{m,\ell,k} \delta^{-1/2} \mathbf{1}_{((k-1)\delta,k\delta]}, \tag{21}$$

where $\delta > 0$ is the size of the bin and K the number of bins. So, we have $K\delta = A$. The $a_{m,\ell,k}$'s are the renormalized coefficients of $g_\ell^{(m)}$ on the regular partition of size K. Since $f \to \psi^{(m)}(f)$ is linear, one obtains:

$$\forall t > 0, \quad \psi_t^{(m)}(f) = \mu^{(m)} + \sum_{\ell=1}^{M} \sum_{k=1}^{K} a_{m,\ell,k} \delta^{-1/2} N^{(\ell)}([t - k\delta, t - (k-1)\delta)),$$

still for $f = \left((\mu^{(m)}, (g_\ell^{(m)})_{\ell=1,\ldots,M})_{m=1,\ldots,M} \right)$. Let us denote by $\mathbf{a}^{(m)}$ the column vector such that

$$(\mathbf{a}^{(m)})' = (\mu^{(m)}, a_{m,1,1}, \ldots, a_{m,1,K}, a_{m,2,1}, \ldots, a_{m,M,K}), \tag{22}$$

where $'$ denotes the transpose. Then, one can write

$$\forall t > 0, \quad \psi_t^{(m)}(f) = (\mathbf{Rc}_t)' \mathbf{a}^{(m)}, \tag{23}$$

with $\mathbf{Rc}_t$ being the renormalized instantaneous counts given by

$$(\mathbf{Rc}_t)' = \left(1, \delta^{-1/2}(\mathbf{c}_t^{(1)})', \ldots, \delta^{-1/2}(\mathbf{c}_t^{(M)})' \right),$$

and with $\mathbf{c}_t^{(\ell)}$ being the vector of instantaneous counts with delay of $N^{(\ell)}$ i.e.

$$(\mathbf{c}_t^{(\ell)})' = \left(N^{(\ell)}([t - \delta, t)), \ldots, N^{(\ell)}([t - K\delta, t - (K-1)\delta)) \right).$$

Hence, by (23), proposing $\psi_t^{(m)}(f)$ as a candidate for the intensity $\lambda^{(m)}$ of $N^{(m)}$ amounts to proposing a linear combination of instantaneous counts with delay to model the probability of the next occurrence of a point in $N^{(m)}$.

Now, minimizing $\gamma(f)$ over such piecewise constant functions is equivalent, by linearity, to minimizing

$$\gamma(f) = \sum_{m=1}^{M} \left(-2(\mathbf{a}^{(m)})' \mathbf{b}^{(m)} + (\mathbf{a}^{(m)})' \mathbf{G} \mathbf{a}^{(m)} \right)$$

with respect to the vectors $\mathbf{a}^{(m)}$. The vector $\mathbf{b}^{(m)}$ is observable and is given by

$$(\mathbf{b}^{(m)})' = \left(\int_{T_1}^{T_2} \mathbf{Rc}_t dN^{(m)}(t) \right)' = \left(N^{(m)}([T_1, T_2]), \delta^{-1/2} \mathbf{n}'_{m,1}, \ldots, \delta^{-1/2} \mathbf{n}'_{m,M} \right),$$

where

$$\mathbf{n}_{m,\ell} = \left(\int_{T_1}^{T_2} N^{(\ell)}([t - k\delta, t - (k-1)\delta)) dN^{(m)}(t) \right)_{k=1,\dots,K}$$

and

$$\mathbf{G} = \int_{T_1}^{T_2} \mathbf{Rc}_t (\mathbf{Rc}_t)' dt.$$

Note that the kth component of $\mathbf{n}_{m,\ell}$ is the number of couples (x, y) with $x \in N^{(m)} \cap [T_1, T_2]$, $y \in N^{(\ell)}$ and $(y - x) \in ((k-1)\delta, k\delta]$ and $\mathbf{G}$ is a symmetric matrix of size $1 + MK$ whose components are the integrated covariations of the renormalized instantaneous counts. The solution of this minimization problem is easily available: if $\mathbf{G}$ is invertible,

$$\forall m = 1, \dots, M, \quad \hat{\mathbf{a}}^{(m)} = \mathbf{G}^{-1} \mathbf{b}^{(m)}. \tag{24}$$

Heuristic arguments show that (24) is a natural expression. We can indeed informally write for any m that

$$dN^{(m)}(t) \simeq \lambda^{(m)}(t) dt + \text{ noise } \simeq \psi_t^{(m)}(f^*) dt + \text{ noise},$$

assuming that at time t, the intensity is strictly positive. By linearity of $\psi^{(m)}$, one can also write that

$$dN^{(m)}(t) \simeq (\mathbf{Rc}_t)' \mathbf{a}_*^{(m)} + \text{ noise},$$

where $\mathbf{a}_*^{(m)}$ are the coefficients corresponding to f^*, assuming that f^* can be coded in this way. Finally, we obtain

$$\int_{T_1}^{T_2} \mathbf{Rc}_t dN^{(m)}(t) = \mathbf{b}^{(m)} \simeq \int_{T_1}^{T_2} \mathbf{Rc}_t (\mathbf{Rc}_t)' \mathbf{a}_*^{(m)} dt + \text{ noise } \simeq \mathbf{Ga}_*^{(m)} + \text{ noise},$$

showing that the estimate given in (24) should be a convenient preliminary estimate.

3.3.2 Least-square estimates on several trials and connections with JPSTH and cross-correlograms

We observe now n trials and therefore we have access to $(N_i^{(1)}, \dots, N_i^{(M)})_{i=1,\dots,n}$ an i.i.d. sample of a multivariate point process on $[T_1, T_2]$. Each trial has its own history. So to each trial i, we can associate as in the previous subsection the matrix $\mathbf{G}$, the vectors $\mathbf{b}^{(m)}$ and so on. Depending on the trial i, we denote them $\mathbf{G}^{(i)}$, $\mathbf{b}^{(m,i)}$ and so on. The least-square contrast for these $n \times M$ spike trains can then be written as

$$\gamma_n(f) = \sum_{m=1}^{M} \left(-2(\mathbf{a}^{(m)})' \left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)} \right) + (\mathbf{a}^{(m)})' \left(\sum_{i=1}^{n} \mathbf{G}^{(i)} \right) \mathbf{a}^{(m)} \right) \tag{25}$$

whose solution is given by

$$\forall m = 1, \dots, M, \quad \hat{\mathbf{a}}^{(m)} = \left(\sum_{i=1}^{n} \mathbf{G}^{(i)} \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)} \right). \tag{26}$$

The quantity $\left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)}\right)$ can be reinterpreted in terms of cross-correlograms and joint-PSTH, following [59]. Indeed we can write

$$\left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)}\right)' = \left([N^{(m)}]^{a,n}([T_1, T_2]), \delta^{-1/2}\bar{\mathbf{n}}'_{m,1}, \ldots, \delta^{-1/2}\bar{\mathbf{n}}'_{m,M}\right),$$

where for any ℓ,

$$\bar{\mathbf{n}}_{m,\ell} = \sum_{i=1}^{n} \left(\int_{T_1}^{T_2} N_i^{(\ell)}([t - k\delta, t - (k-1)\delta])dN_i^{(m)}(t)\right)_{k=1,\ldots,K},$$

and $[N^{(m)}]^{a,n}$ is the aggregated process over all the n trials for the mth neuron. The quantity $\bar{\mathbf{n}}_{m,\ell}$ can be reinterpreted as a particular histogram based on the joint peristimulus time scatter diagram as the JPSTH or the cross-correlogramm (see Figure 1 of [59] and Figure 3 of the present article). More precisely as detailed in Figure 3, the counts $\bar{\mathbf{n}}_{m,\ell}$ are close to a cross-correlogram except that representations are not based on squares but on herringbones. Local features are then preserved, as for the JPSTH. Furthermore the elements of the partition have the same area and can therefore be compared more easily. Besides, for small disjoint intervals $[T_1, T_2]$ with an increasing parameter A (corresponding to the maximal size of the support of the interaction functions), and for δ close to 0, we obtain representations close to the JPSTH, except that the limits are not parallel to the axis but parallel to the diagonal. This change of orientation smooths the binning effects. Indeed the quantity that is binned for the quantities $\bar{\mathbf{n}}_{m,\ell}$ is the delay itself between two points, whereas for the JPSTH, each position of the points is first binned. Therefore, two points whose distance is less than δ are always counted as such in one of the diagonal parallelograms for $\bar{\mathbf{n}}_{m,\ell}$, whereas they may eventually not be counted in a diagonal square of the JPSTH, because one point appears in one bin and the other one in another bin (see Figures 3 **C** and 3 **D**). This problem of information loss when binning is involved has already been discussed for the coincidence counts [15].

JPSTH and cross-correlograms have been used for a long time in neuroscience, without links with any model. The formula (26), for the least-square estimate, shows the link between those descriptive statistics (more precisely the $\bar{\mathbf{n}}_{m,\ell}$'s) and the parameters of the Hawkes model. To recover the parameters, we need, in particular, to inverse the matrix $\left(\sum_{i=1}^{n} \mathbf{G}^{(i)}\right)$. This matrix quantifies for instance the following situation. Assume that $M = 3$ and that the interaction functions $h_2^{(1)}$ and $h_3^{(2)}$ are large on $[0, \delta]$ and null elsewhere. We also assume that all the other interaction functions are null. In this situation, $\bar{\mathbf{n}}_{1,3}$ (or at least its first coordinate) will be large even if there is no direct interaction from N_3 on N_1. The matrix $\left(\sum_{i=1}^{n} \mathbf{G}^{(i)}\right)$ cumulates all these features (and also the fixed effect due to the spontaneous parameter, which needs to be substracted) and inverting it enables us to find an estimate of the true interactions. See also [58] for a more visual transcription.

Note however that even if many coefficients are null as in the above described situation, due to the random noise, the estimates $\hat{\mathbf{a}}^{(m)}$ have non-zero coordinates almost surely. Therefore, it is difficult to interpret the resulting estimate in terms of functional connectivity graph [58]. Moreover if we wish to capture all the features, it is preferable to take A large and δ small. Therefore the number of parameters of the model, depending on $K = A\delta^{-1}$, increases. With a small number of trials n and a small interval $[T_1, T_2]$, the least-square estimate is doomed to be quite poor as the MLE [19].

To remedy these problems, we now consider ℓ_1 penalization to find a non parametric estimate with adaptive properties and prescribed convergence rate.

3.3.3 Lasso estimate

The Lasso method as developed by [57], is based on the following penalized least-square criterion, reformulated here in the context of n i.i.d. trials: for any $m = 1, \ldots, M$,

$$\tilde{\mathbf{a}}^{(m)} \in \arg\min_{\mathbf{a}^{(m)}} \left(-2(\mathbf{a}^{(m)})' \left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)} \right) + (\mathbf{a}^{(m)})' \left(\sum_{i=1}^{n} \mathbf{G}^{(i)} \right) \mathbf{a}^{(m)} + 2(\mathbf{d}^{(m)})'|\mathbf{a}^{(m)}| \right), \tag{27}$$

where $|\mathbf{a}^{(m)}|$ denotes the vector whose coefficients are the absolute values of the coefficients of $\mathbf{a}^{(m)}$ and where

$$(\mathbf{d}^{(m)})' = (d_{m,0}, d_{m,1,1}, \ldots, d_{m,1,K}, d_{m,2,1}, \ldots, d_{m,M,K})$$

is a vector of positive observable weights given by

$$d_{m,\ell,k} = \sqrt{2\gamma \ln(n(T_2 - T_1))\hat{V}_{m,\ell,k}} + \frac{\gamma \ln(n(T_2 - T_1))}{3}\hat{B}_{\ell,k}, \tag{28}$$

where

$$\hat{V}_{m,\ell,k} = \sum_{i=1}^{n} \int_{T_1}^{T_2} \delta^{-1} \left[N_i^{(\ell)}([t - k\delta, t - (k-1)\delta)) \right]^2 dN_i^{(m)}(t),$$

$$\hat{B}_{\ell,k} = \delta^{-1/2} \sup_{i,t\in[T_1,T_2]} N_i^{(\ell)}([t - k\delta, t - (k-1)\delta)),$$

and with

$$d_{m,0} = \sqrt{2\gamma \ln(n(T_2 - T_1))[N^{(m)}]^{a,n}([T_2, T_1])} + \frac{\gamma \ln(n(T_2 - T_1))}{3}.$$

Since the criterion (27) is convex, the minimization problem can be performed quite easily. The function $f \in \mathcal{H}$ associated with $\tilde{\mathbf{a}}$ is denoted $\hat{f}^B$, in reference to the Bernstein inequality that governs the shape of the weights (see [57]).

Because the penalty term added to the least-square criterion is a weighted ℓ_1-norm, the resulting estimate is sparse and many coordinates in $\tilde{\mathbf{a}}^{(m)}$ will be null (see [60] for the seminal paper on Lasso methods). This

27

estimate and much more general forms have been studied quite intensively in [57]. In Additional File 3, we prove an oracle inequality for a slight modification of the present estimate, whose exact form can also be found in [58].

Let us just present the result informally to highlight the main properties (the complete version can be found in Additional File 3). An oracle inequality, in the same spirit as Theorem 2, is proved. The main difference is that it holds on a event with large probability and not in expectation. We have an upper bound of

$$\sum_i \sum_m \int_{T_1}^{T_2} \left(\psi^{(m)}(\hat{f}^B)_i(t) - \lambda_i(t) \right)^2 dt, \tag{29}$$

that constitutes a compromise, as usual, between a bias term and a variance term. Minimizing the bias gives the best linear approximation of λ of the form $\psi(f)$ and this even if λ is not of the form $\psi(f)$. In this sense it applies in particular to Hawkes processes with self-inhibition (i.e. negative $h_m^{(m)}$'s), which models refractory periods [22] and for which $f \to \lambda = (\psi(f))_+$ is not linear anymore. Finally, (29) leads to a control of the left hand side of (8) adapted to the context of this section. Under further technical assumptions, we can then prove that Test 4 can be applied. We refer the reader to [57] for more details that are omitted here to avoid too tedious technical aspects.

The last point already developed in [57] is that Lasso estimates are most of the time biased in practice. To overcome this problem, a two step procedure is proposed. It consists in finding the non zero coefficients of $\hat{f}^B$ and performing a classical least-square estimate on this support. We denote this two-step estimate $\hat{f}^{BO}$.

4 Practical performance
4.1 Description of the data
4.1.1 Real data

The data used here are a small subset of already partially published data in previous experimental studies [15,22,61,62]. These data were collected on a 5-year-old male Rhesus monkey who was trained to perform a delayed multidirectional pointing task. The animal sat in a primate chair in front of a vertical panel on which seven touch-sensitive light-emitting diodes were mounted, one in the center and six placed equidistantly (60 degrees apart) on a circle around it. The monkey had to initiate a trial by touching and then holding with the left hand the central target. After a delay of 500 ms, the preparatory signal (PS) was presented by illuminating one of the six peripheral targets in green. After a delay of either 600 or 1200 ms, selected at random with various probability, it turned red, serving as the response signal and pointing target. During the first part of the delay, the probability for the response signal to occur at 500+600 ms =1.1 s was 0.3.

Once this moment passed without signal occurrence, the conditional probability for the signal to occur at 500+600+600 ms =1.7 s changed to 1. The monkey was rewarded by a drop of juice after each correct trial, i.e. a trial for which the monkey touches the correct target at the correct moment.

Signals recorded from up to seven independently moving microelectrodes (quartz insulated platinum–tungsten electrodes, impedance: 2–5MO at 1000 Hz) were amplified and band-pass filtered from 300Hz to 10 kHz. Single unit activity was obtained by performing an online discrimination of spikes on each electrode. Spikes were firstly selected by taking into account their amplitude using an online window discriminator with high-pass and low-pass filters. In cases where spikes were not discriminable due to their amplitude only, the electrode was moved until the signals were sufficiently distinct to be discriminable on this basis. Although off-line spike sorting was available, it was not used in this study. Indeed, beyond the reservations that one may have concerning the variable quality of the output of such software, the use of clean original electrophysiological signals makes safer the more specific study of precise neuronal synchronization. Neuronal data along with behavioral events (occurrences of signals and performance of the animal) were stored on a PC for off-line analysis with a time resolution of 1 kHz.

Two sets of data are here considered. They both correspond to a probability of 0.3 that the response signal occur at 1.1 s for the monkey, but only correct trials where the response signal occurs at 1.7 s are considered. On both data sets, two neurons have been recorded simultaneously over $[0, T_{max}]$ where T_{max} is approximately two seconds. In the sequel, Data Set **A** (respectively **B**) corresponds to the pair of neurons (N1A,N2A) (respectively to the pair of neurons (N1B,N2B)). In Data Set **A** (respectively **B**), 177 trials (respectively 141 trials) are considered. Figure 4 plots the rasters associated with both data sets. However, because 6 different directions of movement were proposed to the monkey, we can also consider in both data sets, 6 subsets of trials, each subset corresponding to a prescribed direction of movement (see Table 1 for a repartition of the number of trials per direction).

Therefore, n, the total number of trials will be close to 200 if one aggregates over all the directions or will belong to the interval $[20, 35]$ if one considers the trials according to the directions. Those trials are assumed to be i.i.d. This assumption is more reasonable if one considers trials for a fixed given direction.

4.1.2   Simulated data

To assess the performance of our procedure, simulated data for which the underlying model is known have also been simulated. Three different data sets have been simulated, with the thinning method [63]:

- (S-HomPoi) Spikes are distributed according to an homogeneous Poisson processes of intensity 20 Hz

on $[0,2]$s.

- (S-InPoi) Spikes are distributed according to an inhomogeneous Poisson processes with piecewise continuous intensity on $[0,2]$s given by

$$t \to \lambda(t) = \sum_{i=1}^{3} \left[g_i + h_i e^{-\frac{4*(t-c_i)^2}{r_i^2-(t-c_i)^2}} \right] \mathbf{1}_{t \in [c_i-r_i; c_i+r_i)},$$

with $g = [5, 30, 0], h = [12.5, 15, 12.5], c = [0.375, 1.25, 1.825]$ and $r = [0.375, 0.5, 0.125]$.

- (S-Haw) Two spike trains are simulated according to a bivariate Hawkes process observed on $[0,2]$. Each process is respectively denoted $N^{(1)}$ and $N^{(2)}$. Their intensities are given by (17) with spontaneous parameters $\nu^{(1)} = \nu^{(2)} = 20$ Hz and interaction functions $h_1^{(1)} = h_2^{(2)} = -20 \times \mathbf{1}_{[0,0.005]}$, $h_2^{(1)} = 60 \times \mathbf{1}_{[0,0.01]}$ and $h_1^{(2)} = 0$.

Each time a n i.i.d. sample is drawn.

The several treatments have been done in R except Willett and Nowak's estimate (WN) for which Matlab has been used.

4.2 Results
4.2.1 Checking the homogeneous Poisson assumption

One of the simplest test is to check whether the number of spikes per trial obeys a Poisson distribution with unknown parameter. Since the Poisson distribution is discrete, one can use a chi-square test with one estimated parameter, whose results are summarized in Table 2. Since the number of spikes per trial is also a Poisson variable for inhomogeneous Poisson processes, it is reasonable to have large p-values for (S-HomPoi) and (S-InPoi). Relatively smaller p-values appear for (S-Haw), but they are not small enough for a clear rejection: this test seems therefore not very powerful. Note that a very close test has been used in [8] on disjoint intervals of observations. The estimate of the parameters were computed via a prescribed parametric model and therefore the procedure was testing both the Poisson assumption and the parametric assumption. Up to our knowledge, chi-square tests cannot be adapted to non parametric plug-in. On Data Sets **A** and **B**, the most undoubtedly rejection appears for the pooled data, which can be explained by the fact that those data are not i.i.d.

The second simplest test (as prescribed by Yana et al [1] for instance) is the classical Kolmogorov-Smirnov test of uniformity performed on all the spikes, once all the trials have been aggregated, which relies on the fact that conditionally to the total number of observed spikes the points of a homogeneous Poisson process

should obey a uniform distribution. Table 3 shows the corresponding p-values. It is quite coherent to have large p-values for (S-HomPoi) and very small p-values for (S-InPoi) because of the lack of stationarity of the later. Since (S-Haw) is stationary, aggregation dilutes the dependence and it explains that this test is not powerful and that the p-values are large in this case. This test clearly rejects for Data Sets **A** and **B** the homogeneous Poisson hypothesis.

Another test of first order statistics as explained in [1] is the exponential test on the ISI (see Table 4). Here we apply our version of the exponential test, i.e. Test 1, using the subsampling scheme. As soon as there is enough trials, this test is powerful enough to detect the non stationarity (S-InPoi), but also the dependence (S-Haw), since there is no aggregation to dilute the dependence between consecutive points in one trial with respect to the previous test. On Data Sets **A** and **B**, the result is almost the same as the previous test.

4.2.2 Checking the inhomogeneous Poisson assumption

Let us first look at the reconstructions of the underlying intensities. We first program two very basic kernel estimates (10): the sliding window (i.e. $K = (1/2)\mathbf{1}_{[-1,1]}$) with length 0.1s (i.e. bandwidth 0.05), denoted $\hat{\lambda}_n^{SW_h}$, and the Gaussian kernel (used for instance in [8]) with the same bandwidth and denoted $\hat{\lambda}_n^{K_h}$. As [8], we also programmed a data-driven choice of bandwidth, called the thumb rule (here we followed [39] to construct it): it is denoted $\hat{\lambda}_n^{K_{h*}}$. Finally we programmed the three presented adaptive method. The GL estimate, $\hat{\lambda}_n^{GL}$ is programmed with the Gaussian kernel, with the choice $\eta = 0.5$ and with the bandwidths family

$$\mathcal{H} = \{D^{-1} : \; D = 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 18, 20, 22, 25, 30, 35, 40, 45, 50\},$$

which has shown a robust behavior on various simulations, for the present considered size of n ($n \approx 40$). The thresholding estimate on the Haar basis, $\hat{\lambda}_n^{Th}$, has been performed with $j_0 = 15$ and $\gamma = 1$. We use the Matlab package FreeDegree for the Willett and Nowak estimate, $\hat{\lambda}_n^{WN}$. Reconstructions for (S-HomPoi) and (S-InPoi) are given in Figure 5.

First of all, $\hat{\lambda}_n^{SW_h}$ is clearly the worst choice, as expected for such a rough kernel. Figure 5 **A** shows the reconstruction of a constant intensity. Kernel estimates with Gaussian kernels are oscillating, the thumb rule bandwidth h^* is larger than $\hat{h}$, the GL bandwidth. The fixed bandwidth $h = 0.05$ is the smallest one and is quite inadequate in this setting. The adaptive Haar thresholding rule $\hat{\lambda}_n^{Th}$ is much better in this case. The WN method is able to reconstruct perfectly the flat line. For (S-InPoi), the intensity has large jumps and smooth bumps (Figures 5 **B** and **C**). For such an irregular intensity and for a small number

of trials (Figure 5 **B**), the thresholding estimate and the WN method are both able to recover the jumps perfectly but the smooth bumps are estimated by a piecewise constant function. The Gaussian kernels are better for the estimation of the bumps but of course, they cannot detect the jumps. In this respect, the GL bandwidth is the best, whereas $\hat{\lambda}_n^{K_{h^*}}$ and $\hat{\lambda}_n^{K_h}$ are too smooth. For a large number of trials (Figure 5 **C**), the thresholding estimate is a bit refined but clearly suffers from a lack of smoothness. Unlike $\hat{\lambda}_n^{K_{h^*}}$ and $\hat{\lambda}_n^{K_h}$, $\hat{\lambda}_n^{GL}$ is reconstructing all the three bumps. The WN method is reconstructing more accurately the jumps despite some important boundary artefacts. It also gives smoother reconstructions for the bumps. In conclusion, the GL method clearly gives a bandwidth choice that adapts to high irregularity of the intensity with respect to other choices, whereas the thresholding estimate, which leads to an adaptive histogram, is more spatially adaptive despite its lack of smoothness. Up to boundary effects, the WN methodology seems to be the most accurate, since it adapts to the regularity of the underlying intensity. Note however that on an interval with a few number of points, this method provides a piecewise constant reconstruction, even if the underlying intensity is smooth, because this choice is more robust. This conclusion is also coherent with two previous and more extensive studies (see [43, 44]).

To understand more clearly the variability of WN estimates, we have drawn 20 reconstructions in Figure 6, which confirms that the variability is small and diminishes when n grows. The estimates are more likely to be piecewise constant for small n. For large n, there are some edge effects around the jumps but the amplitude around the bumps is very small thanks to a smooth estimate in those parts of the curve.

As an example of estimation on real data, Figure 7 displays an interesting case, namely N2A in direction 1, where one sees clearly that WN estimate is able to find correctly the main jump in the intensity (which is clearly seen in the raster plots of Figure 4) but also to produce a smoother estimate at some places (here a straight decreasing line on the right) to reproduce the much slower attenuation of the firing rate that can be seen in the raster plots.

Now let us apply the different proposed tests. First, Test 2 has been applied on the simulated data sets with $n = 40$ and $p = \lfloor n^{2/3} \rfloor$ and nothing was declared significant. However, the p-values corresponding to (S-Haw) are abnormally large. To take this into account we have performed also a variant of Test 2 where one rejects if the same test statistic is now smaller than $k_{n,\alpha}$. This test consists therefore in rejecting the Poisson hypothesis when both estimated distributions are too close. As Test 2, this test is also asymptotically of level α, by application of Proposition 2. Actually, all the tests presented here can be said to reject "by upper values" and have therefore a version "by lower values". On each set (simulated or not) of Test 2 p-values (by upper and lower values), one can perform a Benjamini and Hochberg procedure with FDR

5% [64]. It declares both processes in (S-Haw) as non Poissonian in the family of simulated data (p-values in $[10^{-3}.10^{-2})$). Due to the high variability of the reconstructions on Data Sets **A** and **B**, which depend on the considered direction, it was not possible to pool the data together and therefore the corresponding tests have been performed direction by direction. However Tests 2 by upper or lower values do not detect anything on Data Sets **A** and **B**.

Tests 3 (by upper or lower values) do not detect anything on the simulated data sets. However on Data Sets **A** and **B** (see Table 5), Test 3 clearly rejects the Poisson hypothesis for most of the directions of N1B and in this sense, Test 3 is more powerful than Test 2.

Test 5 is, as expected, more powerful than Tests 2 and 3 and detects the non Poissonian structure in (S-Haw) (see Table 6). More importantly on Data Sets **A** and **B**, all the p-values of Test 5 (by upper values) are smaller than 10^{-14} making clear that those data are not inhomogeneous Poisson processes.

4.2.3 Checking the Hawkes assumption

Once again, Data Sets **A** and **B** have to be treated direction by direction because of the high variability of the reconstruction. However, because of the very small number of trials per direction, it was not possible to look at very small sub-intervals $[T_1, T_2]$. Therefore we decided to look at the largest interval that one can take i.e. $[K\delta, 2]$. We choose $K = 8$, $\delta = 0.005$ and $\gamma = 1$. On simulated data (Figure 8), $\hat{f}^B$ and $\hat{f}^{BO}$ recover the support of the interaction functions and also find that $h_1^{(2)} = 0$, which could not have been possible with a classical least-square estimate (see also more comment on the functional connectivity graph in [58]). Moreover $\hat{f}^{BO}$ is less biased than $\hat{f}^B$ as expected. We provide one estimation of the interaction functions for Data Set **A** in direction 2 (Figure 9), where it clearly appears a one way excitation of N2A on N1A. This is coherent with a previous study on the same data set, which finds this pair of neurons dependent through a complete different method [22]. Note also that at short range the self-interaction functions are negative, fact which is consistent with refractory periods.

Test 4 is used to check the Hawkes assumption, with $\hat{f}$ given either by $\hat{f}^B$ or $\hat{f}^{BO}$ (see Table 7). It is coherent to find that, on simulated data, the Hawkes assumption is accepted for (S-Haw) but also for (S-HomPoi), which is a particular case of Hawkes process with null interaction functions. On the contrary, Test 4 detects that (S-InPoi) is not a Hawkes process. On Data Sets **A** and **B**, the Hawkes assumption is sometimes accepted on the whole interval of observations. Let us focus for instance on N1A in direction 2 (see also Figure 9). This process could not be a Poisson process, since the p-value of Test 5 is smaller than 10^{-14}. However the large p-value for Test 4 combined with Figure 9 means that this process can be explained

via a pure self-interaction process, which is negative at short range (refractory period) and positive at larger range. Note that 9 processes on the 24 are not explained in this way (in particular N2A in direction 2), and are not Poisson either: this can be due to a too large lack of stationarity, combined with a large dependence between points.

5 Conclusion

When using the time-rescaling theorem to assess whether an observed spike train obeys a certain probabilistic model (e.g. Poisson, Hawkes...), a plug-in step is currently performed [1–3,8,24]. If this plug-in step is done without care the resulting test may be much too conservative leading to poor detections (see Figure 1). We propose here to use the subsampling as an almost universal solution when dealing with Kolmogorov-Smirnov tests of uniformity, such a universal solution being completely new. The main requirement is to have access to an estimate of the underlying intensity, whose rate of convergence is known (see for instance (5)).

In classical previous works such as [8,23], parametric estimates such as MLE over a prescribed parametric model are used as plug-in estimates, which are converging towards the true intensity when the underlying parametric model is true. Therefore, when this parametric plug-in is used inside omnibus tests (and in particular the ones presented here), one has to be careful that we are not only testing the probabilistic assumption but also the fact that the intensity belongs to the parametric model. To overcome this problem, we advertise for the use of nonparametric estimates and more precisely to adaptive estimates, for which rates of convergence are known under lighter assumptions on the intensity than a prescribed parametric assumption. Moreover those adaptive methods have very good practical performance making them also very good in practice for estimation.

On some simulated data and on some real data sets, we have shown that our method performs very well. However there are still two main directions in which our work need to be pursued in order to provide a more complete answer on real data sets. First of all, the KS test of exponentiality on the ISI can also be performed instead of the KS test of uniformity [1]. If the method with subsampling can be adapted to this case, we have presently no guarantee that this test would have a controlled level. Indeed we have no equivalent of Proposition 2 or Theorem 1 for the ISI repartition. The second main drawback is that we are clearly able to reject (at least on the presented real data sets) both homogeneous and inhomogeneous Poisson assumptions. We are also able to test whether the processes are Hawkes or not, which in particular takes into account refractory periods and dependence between several spike trains. However the Hawkes model reflects stationary features and cannot model non stationary data. Therefore we would need a more general model

which includes the dependence as in the Hawkes model presented here but also the non stationarity as in the inhomogeneous Poisson process. At the present moment, models reflecting both are not compatible with a full agnostic approach where no assumption is made on the underlying functions, the estimation problem being not completely identifiable [16]. A first step will be therefore to provide a trade off between estimation capacity and not too restrictive assumptions on the process itself, with respect to real spike train data.

6 Acknowledgements

References

1. Yana K, Takeuchi N, Takikawa Y, Shinomura M: **A method for testing an extended Poisson hypothesis of spontaneous quantal transmitter release at neuromuscular junctions.** *Biophysical Journal* 1984, **46**(3):323–330.

2. Brown E, Barbieri R, Ventura V, Kass R, Frank L: **The time rescaling theorem and its application to neural spike train analysis.** *Neural Computation* 2002, **14**(2):325–346.

3. Pouzat C, Chaffiol A: **On goodness of fit tests for models of neuronal spike trains considered as counting processes**. Tech. rep., http://arxiv.org/abs/0909.2785 2009.

4. Daley DJ, Vere-Jones D: *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York), New York: Springer-Verlag, second edition 2003. [Elementary theory and methods].

5. Pipa G, Grün S: **Non-parametric significance estimation of joint-spike events by shuffling and resampling.** *Neurocomputing* 2003, **52-54**:31–37.

6. Pipa G, Diesmann M, Grün S: **Significance of joint-spike events based on trial-shuffling by efficient combinatorial methods.** *Complexity* 2003, **8**(4):1–8.

7. Pouzat C, Chaffiol A: **Automatic spike train analysis and report generation. An implementation with R, R2HTML and STAR**. *Journal of Neuroscience Methods* 2009, **181**:119–144.

8. Ventura V, Carta R, Kass R, Gettner S, Olson C: **Statistical analysis of temporal evolution in single-neuron firing rates**. *Biostatistics* 2002, **3**:1–20.

9. Dayan P, Abbott LF: *Theoretical neuroscience.* Computational Neuroscience.

10. Georgopoulos A, Schwartz A, Kettner R: **Neuronal population coding of movement direction.** *Science* 1986, (233):1416–1419.

11. Rizzolatti G, Craighero L: **The mirror-neuron system.** *Annual Review of Neuroscience* 2004, **27**:169–192.

12. Shinomoto S: *Analysis of Parallel Spike Trains*, Springer Series in Computational Neuroscience 2010 chap. Estimating the Firing Rate.

13. Grün S: **Unitary joint-events in multiple-neuron spiking activity: Detection, significance and interpretation.** *PhD thesis*, Thun: Verlag Harri Deutsch 1996.

14. Grün S, Diesmann M, Aertsen A: *Analysis of Parallel Spike Trains*, Springer Series in Computational Neuroscience 2010 chap. Unitary Events Analysis.

15. Riehle A, Grammont F, Diesmann M, Grün S: **Dynamical changes and temporal precision of synchronised spiking activity in monkey motor cortex during movement preparation.** *Journal of Physiology* 2000, **94**:569–582.

16. Krumin M, Reutsky I, Shoham S: **Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input.** *Frontiers in Computational Neuroscience* 2010, **4**(article 147).

17. Pernice V, Staude B, Cardanobile S, Rotter S: **How structure determines correlations in neuronal networks.** *PLoS Computational Biology* 2012, (85:031916).

18. Pernice V, Staude B, Cardanobile S, Rotter S: **Recurrent interactions in spiking networks with arbitrary topology.** *Physical review E, Statistical, nonlinear, and soft matter physics* 2011, (7:e1002059).

19. Chornoboy E, Schramm L, Karr A: **Maximum likelihood identification of Neural Point Process Systems**. *Biol. Cybernetics* 1988, **59**:265–275.

20. Pillow J, Shlens J, Paninski L, Sher A, Litke A, Chichilnisky E, Simoncelli E: **Spatio-temporal correlations and visual signalling in a complete neuronal population.** *Nature* 2008, **454**:995–999.

21. Okatan M, Wilson M, Brown E: **Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity**. *Neural Computation* 2005, **17**:1927–1961.

22. Tuleau-Malot C, Rouis A, Grammont F, Reynaud-Bouret P: **Multiple Tests based on a Gaussian Approximation of the Unitary Events method**. Tech. rep., http://hal.archives-ouvertes.fr/hal-00757323 2012.

23. Ogata Y: **Statistical models for earthquakes occurrences and residual analysis for point processes.** *Journal of the American Statistical Association* 1988, **83**(401):9–27.

24. Tuckwell H: *Introduction to theoretical neurobiology, Volume 2. Nonlinear and stochastic theories.* Cambridge Studies in Mathematical Biology. 1988.

25. Barnard G: **Time intervals between accidents - a note on Maguire, Pearson and Wynn's paper**. *Biometrika* 1953, **40**(1/2):212–213.

26. Lawless J: *Statistical Models and Methods for Lifetime Data.* Wiley-Interscience 1982.

27. Stephens M: **EDF statistics for goodness of fit and some comparisons**. *Journal of the American Statistical Association* 1974, **69**(347):730–737.

28. Bickel PJ, Doksum KA: *Mathematical statistics.* San Francisco, Calif.: Holden-Day Inc. 1976. [Basic ideas and selected topics, Holden-Day Series in Probability and Statistics].

29. Andersen PK, Borgan Ø, Gill RD, Keiding N: *Statistical models based on counting processes.* Springer Series in Statistics, New York: Springer-Verlag 1993.

30. Shimazaki H, Shinomoto S: **Kernel bandwidth optimization in spike rate estimation.** *Journal of Computational Neuroscience* 2010, **29**:171–182.

31. Brémaud P: *Point processes and queues.* New York: Springer-Verlag 1981. [Martingale dynamics, Springer Series in Statistics].

32. Papangelou F: **Integrability of expected increments of point processes and a related random change of scale**. *Trans. Amer. Math. Soc.* 1972, **165**:483–506.

33. Shorack GR, Wellner JA: *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, New York: John Wiley & Sons Inc. 1986.

34. Hogg RV, Tanis EA: *Probability and statistical inference.* Macmillan Publishing Co., New York 1977.

35. Roquain E: **Type I error rate control in multiple testing: a survey with proofs**. *Journal de la Société Francaise de Statistique* 2011, **152**:3–38.

36. Gerstein G, Perkel D: **Simultaneous recorded trains of action potentials: analysis and functional interpretation.** *Science* 1969, **164**:828–830.

37. Tsybakov AB: *Introduction to nonparametric estimation.* Springer Series in Statistics, New York: Springer 2009. [Revised and extended from the 2004 French original, Translated by Vladimir Zaiats].

38. Rudemo M: **Empirical choice of histograms and kernel density estimators**. *Scand. J. Statist.* 1982, **9**(2):65–78.

39. Silverman BW: *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability, London: Chapman & Hall 1986.

40. Goldenshluger A, Lepski O: **Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality**. *Ann. Statist.* 2011, **39**(3):1608–1632.

41. Reynaud-Bouret P: **Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities**. *Probab. Theory Related Fields* 2003, **126**:103–153.

42. Rydén T: **An EM algorithm for estimation in Markov-modulated Poisson processes**. *Comput. Statist. Data Anal.* 1996, **21**(4):431–447.

43. Reynaud-Bouret P, Rivoirard V: **Near optimal thresholding estimation of a Poisson intensity on the real line**. *Electron. J. Stat.* 2010, **4**:172–238.

44. Reynaud-Bouret P, Rivoirard V, Tuleau-Malot C: **Adaptive density estimation: a curse of support?** *J. Statist. Plann. Inference* 2011, **141**:115–139.

45. Willett RM, Nowak RD: **Multiscale Poisson intensity and density estimation**. *IEEE Trans. Inform. Theory* 2007, **53**(9):3171–3187.

46. Perkel G DH ad Gerstein, Moore G: **Neuronal Spike Trains and Stochastic Processes. II Simultaneous Spike Trains**. *Biophysical Journal* 1967, **7**:419–440.

47. Hawkes AG: **Point spectra of some mutually exciting point processes**. *J. Roy. Statist. Soc. Ser. B* 1971, **33**:438–443.

48. Gusto G, Schbath S: **FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model**. *Stat. Appl. Genet. Mol. Biol.* 2005, **4**:Art. 24, 28 pp. (electronic).

49. Carstensen L, Sandelin A, Winther O, Hansen N: **Multivariate Hawkes process models of the occurrence of regulatory elements.** *BMC Bioinformatics* 2010.

50. Brillinger D: **The identification of point process systems**. *Annals of Probability* 1975, **3**(6):909–924.

51. Brillinger D, Bryant H, Segundo J: **Identification of synaptic interactions**. *Biol. Cybernetics* 1976, **22**:213–228.

52. Brémaud P, Massoulié L: **Stability of nonlinear Hawkes processes**. *Ann. Probab.* 1996, **24**(3):1563–1588.

53. Reynaud-Bouret P, Roy E: **Some non asymptotic tail estimates for Hawkes processes**. *Bull. Belg. Math. Soc. Simon Stevin.*

54. Vere-Jones D, Ozaki T: **Some examples of statistical estimation applied to earthquake data.** *Ann. Inst. Statist. Math.* 1982, **34**(B):189–207.

55. Reynaud-Bouret P, Schbath S: **Adaptive estimation for Hawkes processes; application to genome analysis**. *Ann. Statist.* 2010, **38**(5):2781–2822.

56. Blundell C, Heller K, Beck J: **Modelling reciprocating relationships with Hawkes processes.** In *Proceedings of NIPS* 2012.

57. Hansen N, Reynaud-Bouret P, Rivoirard V: **Lasso and probabilistic inequalities for multivariate point processes**. *to appear in Bernoulli* 2013. [Http://arxiv.org/abs/1208.0570].

58. Reynaud-Bouret P, Rivoirard V, Tuleau-Malot C: **Inference of functional connectivity in Neurosciences via Hawkes processes**. In *1st IEEE Global Conference on Signal and Information Processing* 2013, Austin Texas.

59. Aertsen A, Gerstein G, Habib M, Palm G: **Dynamics of neuronal firing correlation: modulation of "effective connectivity"**. *Journal of Neurophysiology* 1989, **61**(5):900–917.

60. Tibshirani R: **Regression shrinkage and selection via the lasso**. *J. Roy. Statist. Soc. Ser. B* 1996, **58**.

61. Grammont F, Riehle A: **Spike synchronisation and firing rate in a population of motor cortical neurons in relation to movement direction and reaction time**. *Biological Cybernetics* 2003, **88**:360–373.

62. Riehle A, Grammont F, MacKay A: **Cancellation of a planned movement in monkey motor cortex.** *Neuroreport* 2006, **17**(3):281–285.

63. Ogata Y: **On Lewis' simulation method for point processes**. *IEEE Transactions on Information Theory* 1981, **27**:23–31.

64. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *J. Roy. Statist. Soc. Ser. B* 1995, **57**:289–300.

65. Doumic M, Hoffmann M, Reynaud-Bouret P, Rivoirard V: **Nonparametric estimation of the division rate of a size-structured population**. *SIAM J. Numer. Anal.* 2012, **50**(2):925–950.

66. Doumic M, Hoffmann M, Reynaud-Bouret P, Rivoirard V: **Nonparametric estimation of the division rate of a size-structured population**. Tech. rep., http://hal.archives-ouvertes.fr/hal-00578694 2012.

7 Figures
7.1 Figure 1 - Repartition of the p-values in a K.S. test of exponentiality with plug-in

Graph of the p-values as a function of their rank. A $n = 40$ i.i.d. sample of exponential variables with parameter $\lambda = 20$ has been drawn 1000 times. Each time a p-value has been computed either by estimating the parameter λ and performing the K.S. test with exactly the same sample (in black), or by estimating the parameter λ on half of the sample and performing the K.S. test on the other half (in red), or by estimating

the parameter λ on the whole sample and performing the K.S. test on a sub-sample of size $n^{2/3}$ (in green). Note that the estimated level (i.e. the number of p-values smaller than 0.05 divided by 1000) is in the first case of 0.009, of 0.12 in the second case and of 0.039 in the third case. Those levels and curves are stable with respect to the sample size: similar results are obtained for larger sample size ($n = 200$ and $n = 1000$).
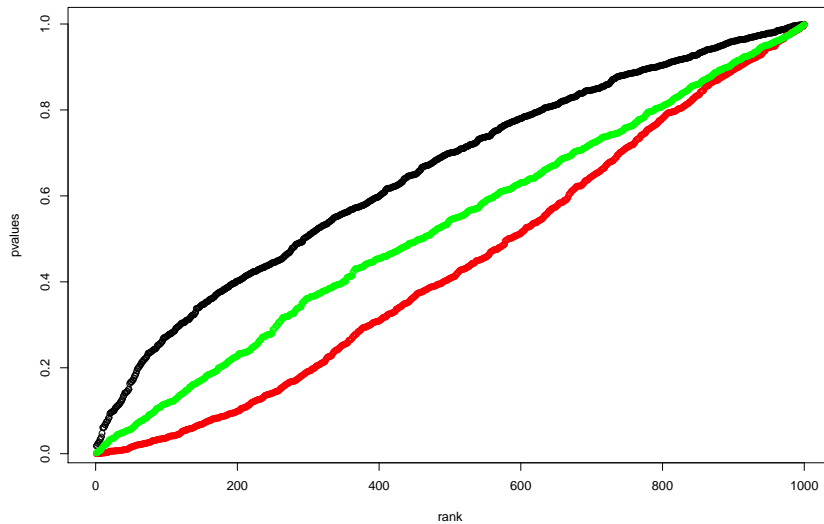

7.2   Figure 2 - Aggregation versus Cumulation

Description of the way the points are gathered together for aggregation or for cumulation. On the left hand side, the first five lines correspond to a trial, the sixth line being the aggregated process. On the right hand side, the same five lines are put together to form the cumulated process.


39
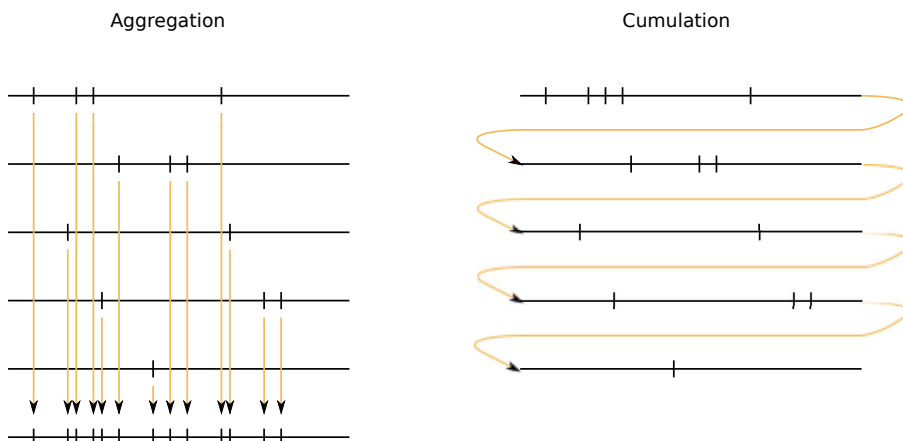
7.3 Figure 3 - Scatter diagrams and histograms

In Figure 3 **A**, we recall how a scatter diagram is constructed for each trial (here black crosses for trial 1 and red crosses for trial 2) and then superposed. Various histograms can be built with those points, using different partitions. Figure 3 **B** gives the partition corresponding to the classical cross-correlogram. Figure 3 **C** gives the partition of the JPSTH. The diagonal squares are filled in gray: if a point of the scatter diagram lies in one of those square, its coordinates correspond to spikes that are very close to each other. Two very close spikes (one on N1, the other one on N2) and their corresponding point in the scatter diagram are added, showing that the reverse is wrong. Therefore there are some couples that are close to each other and not counted by the JPSTH as such. Figure 3 **D** gives the partition used for the computation of the vectors $\bar{\mathbf{n}}_{m,\ell}$ of dimension K ($K = 2$ here). More precisely $\bar{\mathbf{n}}_{1,2}$ corresponds to the vertical part $2 \to 1$, whereas $\bar{\mathbf{n}}_{2,1}$ corresponds to the horizontal part $1 \to 2$. The same spikes are added, they are now counted as close to each other.
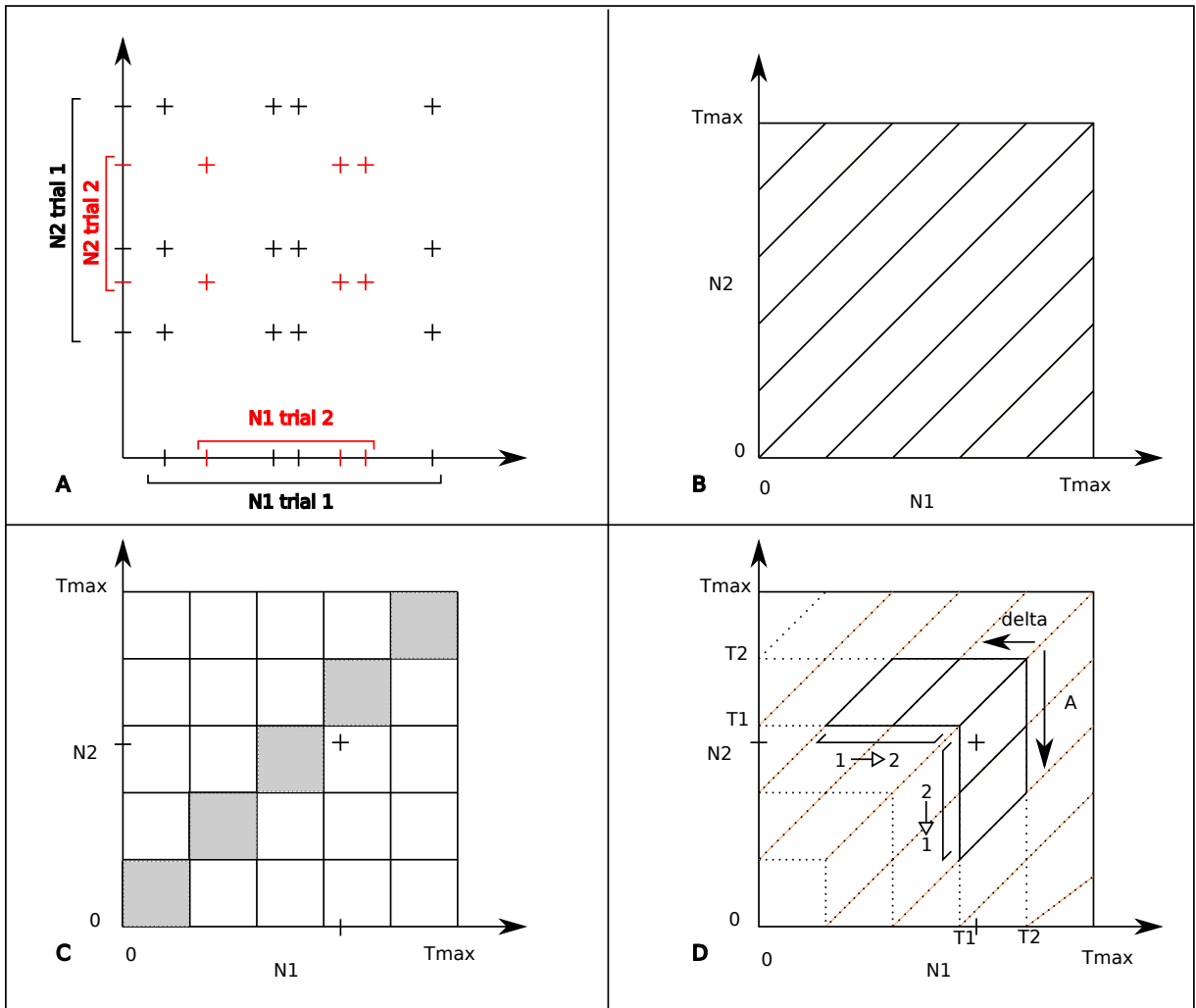
### 7.4 Figure 4 - Raster plots of the data sets

In Figure 4 **A** the rasters associated to Data Set **A** i.e. (N1A,N2A). In Figure 4 **B**, the ones associated to Data Set **B**, i.e. (N1B,N2B). Each line corresponds to a trial, each dot to a spike.
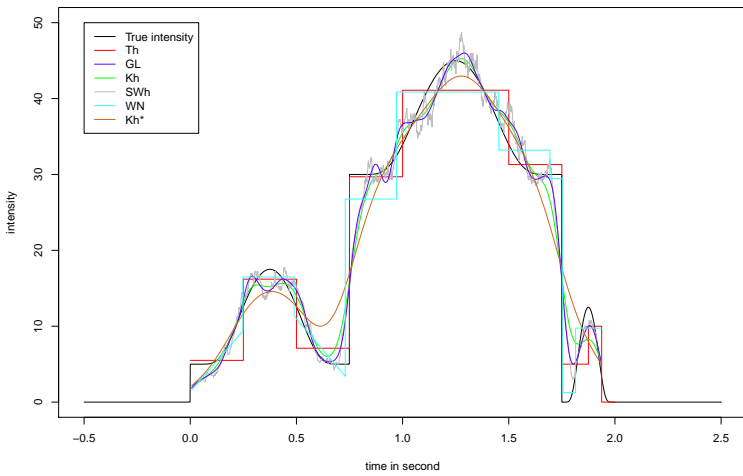
A　　　　　　　　　　　　　**B**
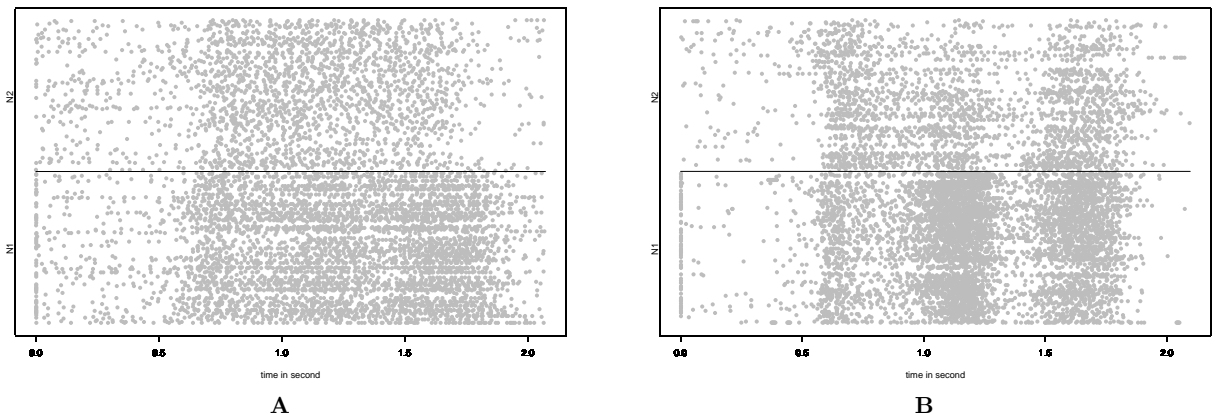
7.5 Figure 5 - Reconstructions of the intensity λ on simulated Poisson data

Reconstructions for (S-HomPoi) and (S-InPoi) over n trials, observed on $[0, 2]$. `Th` corresponds to the adaptive histogram $\hat{\lambda}_n^{Th}$, `GL` to the Goldenshluger and Lepski's method $\hat{\lambda}_n^{GL}$ with the Gaussian kernel, `Kh` to a fixed bandwidth $h = 0.05$ for the Gaussian kernel, `SWh` to the sliding window with $h = 0.05$, `WN` to the Willett and Nowak's method and `Kh*` to the rule of the thumb. Figure 5 **A** corresponds to (S-HomPoi) with $n = 40$, Figure 5 **B** to (S-InPoi) with $n = 40$ and Figure 5 **C** to (S-InPoi) with $n = 200$.
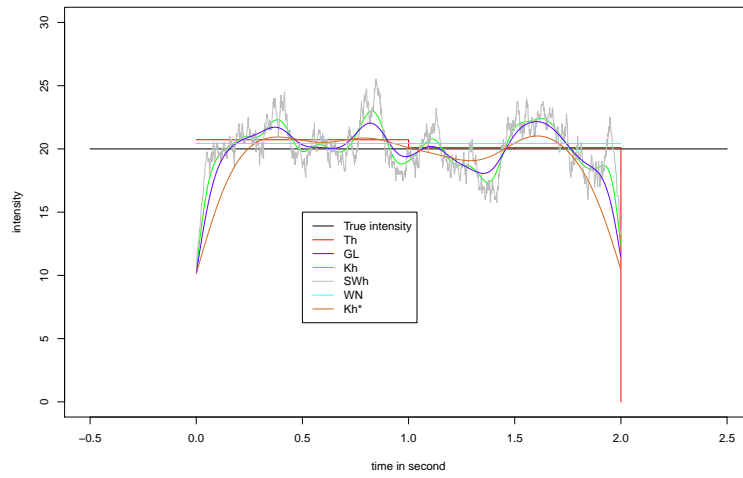
A

B
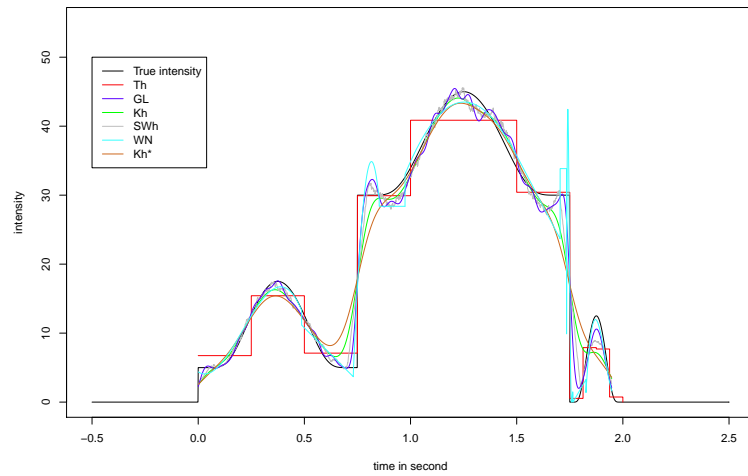
C

7.6 Figure 6 - Variability of the reconstructions of the intensity λ on simulated Poisson data

Superposition of 20 reconstructions by Willett and Nowak's method for (S-InPoi) over n trials, observed on $[0, 2]$. In Figure 6 **A**, $n = 40$. In Figure 6 **B**, $n = 200$.
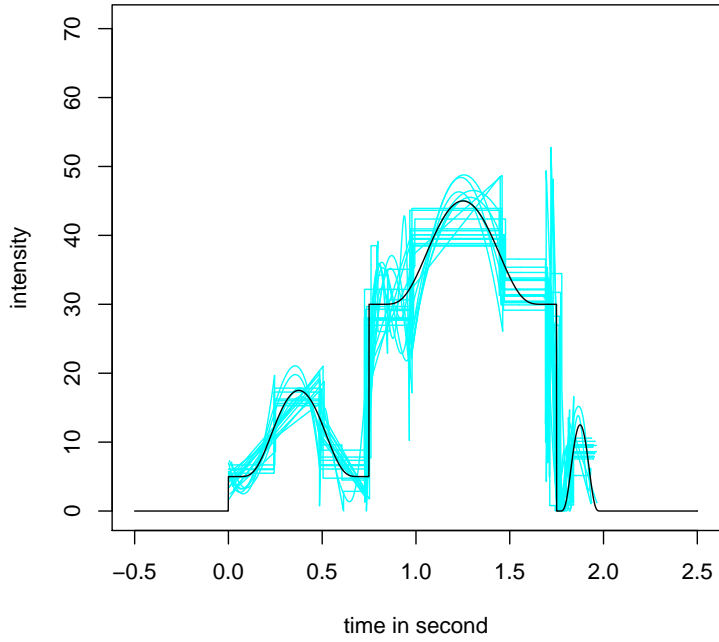
A
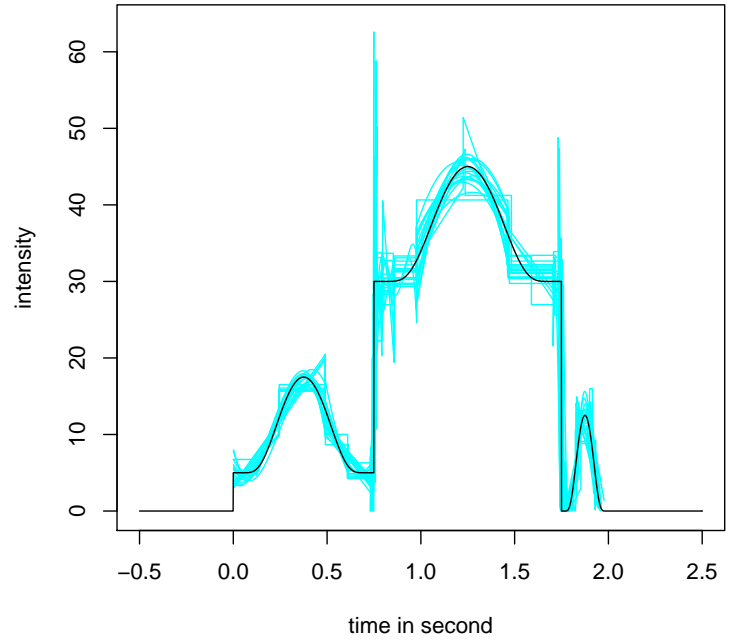
B

7.7 Figure 7 - Example of estimation of the intensity λ on real data

Estimations for neuron N2A, in direction 1. Same conventions as in Figure 5.
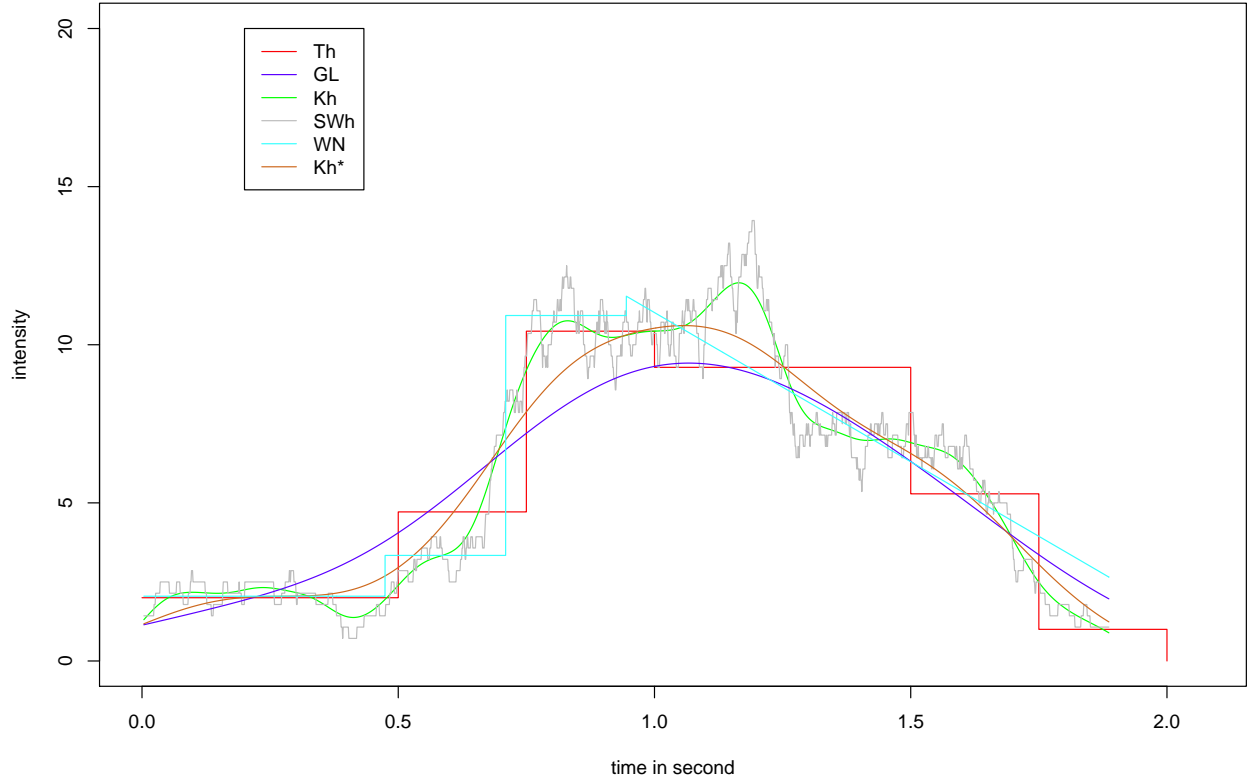
7.8 Figure 8 - Reconstruction of the interaction functions on simulated data

Reconstructions for (S-Haw) with $n = 40$ trials, with $T_1 = 0.05$ and $T_2 = 2$. On the upper left, the interaction function $h_1^{(1)}$, on the upper right, $h_2^{(1)}$, on the bottom left $h_1^{(2)}$ and on the bottom right $h_2^{(2)}$. In black the true interaction functions, in green, reconstruction by method $\hat{f}^B$, in red, reconstruction by method $\hat{f}^{BO}$. On the top of each graphics, the true spontaneous parameter $\nu^{(1)}$ on the top and the true parameter $\nu^{(2)}$ on the bottom are referred by S*. The estimated spontaneous parameters by method $\hat{f}^B$ are referred as SB and the estimated spontaneous parameters by method $\hat{f}^{BO}$ are referred as SBO. We used $K = 8$, $\delta = 0.005$ and $\gamma = 1$.

S*= 20 SB= 18.3 SBO= 20.1

S*= 20 SB= 18.3 SBO= 20.1

S*= 20 SB= 17.3 SBO= 20.6

S*= 20 SB= 17.3 SBO= 20.6
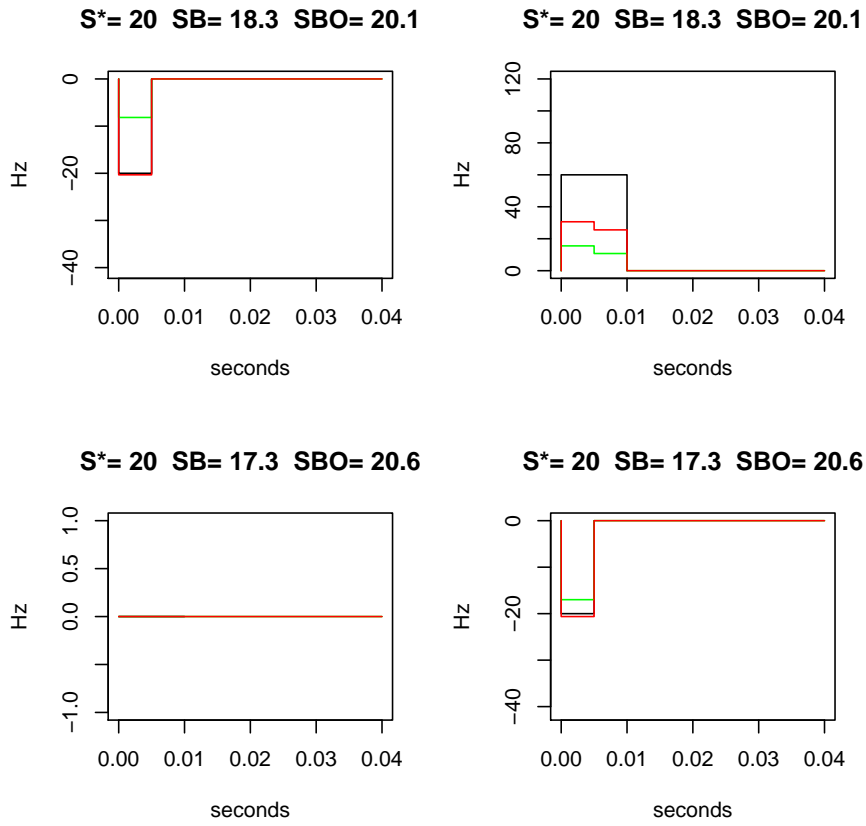
7.9   Figure 9 - Example of estimation of the interaction functions on real data

Estimations for Data Set **A** in direction 2 with $n = 40$ trials, with $T_1 = 0.05$ and $T_2 = 2$. Same conventions as in Figure 8.
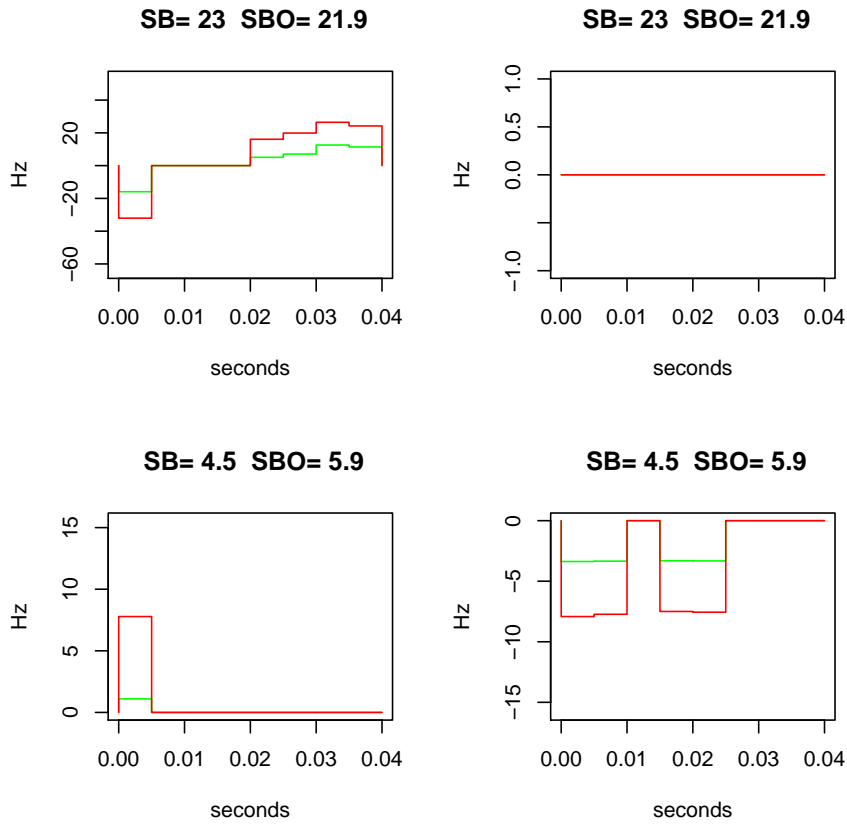
8 Tables
8.1 Table 1 - Repartition of the number of trials

Repartition of number of trials on the real data sets.

Direction	1	2	3	4	5	6	Total
Data set **A**	28	31	30	35	28	25	177
Data set **B**	23	24	26	18	30	20	141

8.2 Table 2 - Chi-square test of the Poissonian distribution for the number of spikes per trial.

P-values of the chi-square test of the Poissonian distribution for the number of spikes per trial. The following code is used: ∘ corresponds to a p-value of the test by upper values in $[10^{-2}, 10^{-1})$, △ to a p-value of the test by upper values in $[10^{-3}, 10^{-2})$, △△ to a p-value of the test by upper value in $[10^{-4}, 10^{-3})$, △△△ to a p-value of the test by upper value in $(-\infty, 10^{-4})$. The signs are filled in black if the p-values correspond to rejection of a Benjamini and Hochberg (BH) multiple test method [64] either on the simulated data (left part of the Table) or on Data sets **A** and **B** (right part of the Table).

n	40	200	Directions	1	2	3	4	5	6	Pooled
S-HomPoi			N1A	○		○	●	▲		▲▲▲
S-InPoi			N2A				▲▲	○	○	▲▲▲
S-Haw $(N^{(1)})$			N1B	○	○		○			▲▲▲
S-Haw $(N^{(2)})$	○		N2B	○	●					▲▲▲

8.3 Table 3 - K.S. test of uniformity.

P-values of the classical K.S. test of uniformity on the spikes, all trials being aggregated. Same codes as in Table 2. Note that none of the p-values were close enough to 1 to force a rejection by the test by lower values (i.e. rejection when the test statistic is smaller than $k_{n,\alpha}$, which corresponds to p-values of the test by upper values larger than $1 - \alpha$).

n	40	200	Directions	1	2	3	4	5	6	Pooled
S-HomPoi			N1A	▲▲▲	▲▲▲	▲▲▲	●	▲▲▲	▲▲	▲▲▲
S-InPoi	▲▲▲	▲▲▲	N2A	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲
S-Haw $(N^{(1)})$			N1B	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲
S-Haw $(N^{(2)})$			N2B	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲	▲▲▲

8.4 Table 4 - Test 1 on the ISI

P-values of Test 1 on the ISI, with a subsample size $\lfloor [n_{tot}^{ISI}]^{2/3} \rfloor$, where n_{tot}^{ISI} is the total number of ISI that have been observed in all the trials. Same codes as in Table 2 for the p-values. Note that none of the p-values were close enough to 1 to force a rejection by the test by lower values (i.e. rejection when the test statistic is smaller than $k_{n,\alpha}$, which corresponds to p-values of the test by upper values larger than $1 - \alpha$).

n	40	200	Directions	1	2	3	4	5	6	Pooled
S-HomPoi			N1A	▲▲▲	▲▲▲	▲	●		○	▲▲▲
S-InPoi	▲▲	▲▲▲	N2A	▲▲▲	▲▲▲	▲▲▲	▲	▲▲▲	▲▲▲	▲▲▲
S-Haw $(N^{(1)})$	○	▲	N1B	▲▲▲	▲▲▲	▲▲▲	●	▲▲▲	▲▲▲	▲▲▲
S-Haw $(N^{(2)})$	▲	▲▲▲	N2B	●	▲▲▲	▲	○	▲▲	▲	▲▲▲

8.5 Table 5 - Test 3 on the real data sets

P-values of Test 3 on Data Sets **A** and **B**, with a subsample size $\lfloor n^{2/3} \rfloor$, where n is the number of trials. Same codes as Table 2 for the p-values of the test by upper values. For the p-values of the test by lower values, same codes except that ○ becomes □ and △ becomes ▽.

Directions		1	2	3	4	5	6
	Th		□				
N1A	GL						
	WN						
	Th						
N2A	GL						
	WN						
	Th			△	○		
N1B	GL	▲▲	▲▲▲	▲▲▲	▲▲		▲
	WN			△			
	Th	△	○				□
N2B	GL	○		○	○	○	
	WN	○					

8.6 Table 6 - Test 5 on the simulated data sets

P-values of Test 5 on the simulated data sets, with $n = 40$ and a subsample size $\lfloor n^{2/3} \rfloor$. Same codes as in Tables 2 and 5 for the p-values.

	Th	GL	WN
S-HomPoi			
S-InPoi			
S-Haw ($N^{(1)}$)	▲	▲	▲
S-Haw ($N^{(2)}$)	○	○	○

8.7 Table 7 - Test 4 for the Hawkes model on both real and simulated data sets

P-values of Test 4 with a subsample size $\lfloor n^{2/3} \rfloor$. Same codes as in Tables 2 and 5 for the p-values.

		$n = 40$	Directions		1	2	3	4	5	6
S-HomPoi	B		N1A	B			○	○	○	
	BO	○		BO					▲▲▲	○
S-InPoi	B	▲	N2A	B		▲▲		▲		○
	BO			BO	○					○
S-Haw ($N^{(1)}$)	B		N1B	B		△	▲		○	
	BO			BO						
S-Haw ($N^{(2)}$)	B		N2B	B	▲▲▲	▲▲▲	○		▲	
	BO	○		BO		○	▲	▲		

9 Additional Files

9.1 Additional file 1 — Kolmogorov-Smirnov tests, plug-in and sub-sampling

9.1.1 Proof of Proposition 1

Let us introduce

$$Z_p = \sqrt{p} \sup_x |F_p(x) - \hat{F}(x)| \quad \text{and} \quad W_p = \sqrt{p} \sup_x |F_p(x) - F(x)|$$

so that $W_p \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{K}$. But, with $U_p := Z_p - W_p$, $|U_p| \le \sqrt{p}\sup_x |\hat{F}(x) - F(x)| \xrightarrow[p\to\infty]{\mathbb{P}} 0$. Therefore, by using the Slutsky Lemma,

$$Z_p = U_p + W_p \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{K}.$$

9.1.2 Proof that Test 1 is asymptotically of level α.

To prove that Test 1 is of level α, it is sufficient to prove (2) with $\hat{F}(x) = (1 - e^{-\hat{\lambda}x})1_{x>0}$ and to apply Proposition 1. But for all $x > 0$,

$$|\hat{F}(x) - F(x)| = e^{-\min(\lambda,\hat{\lambda})x}\left(1 - e^{-|\hat{\lambda}-\lambda|x}\right) \le |\hat{\lambda} - \lambda|xe^{-\min(\lambda,\hat{\lambda})x} \le \frac{|\hat{\lambda} - \lambda|}{\min(\hat{\lambda},\lambda)}e^{-1}.$$

Therefore

$$\sqrt{p(n)}\sup_x |\hat{F}(x) - F(x)| \le \frac{\sqrt{p(n)}|\hat{\lambda} - \lambda|}{\min(\lambda,\hat{\lambda})}.$$

It is well known that $\hat{\lambda}$ is the maximum likelihood estimate of λ and that $\sqrt{n}(\hat{\lambda} - \lambda)$ is asymptotically normal. Therefore $\sqrt{p(n)}|\hat{\lambda} - \lambda|$ tends in probability to 0 whereas $\min(\lambda,\hat{\lambda})$ tends to λ. Therefore (2) is satisfied.

9.1.3 Proof of Proposition 2

When we are dealing with Poisson processes, or more general counting processes, the previous asymptotic approach should be taken with care because the total number of points is random. Indeed if one observes a Poisson process $N^{a,p}$, aggregated over p trials, with constant intensity, then conditionnally to the event $\{N^{a,p}([0, T_{max}]) = n_{tot}\}$, the repartition of the points is uniform. So the following test is exactly of level α.

1. Compute $F_{N^{a,p}([0,T_{max}])}$ as in (3).

2. Compute $\sup_{t\in[0,1]}|F_{N^{a,p}([0,T_{max}])}(t) - t|$.

3. Reject when this last quantity exceeds the random quantity $k_{N^{a,p}([0,T_{max}]),1-\alpha}$, where $k_{n_{tot},1-\alpha}$ is the exact and non asymptotic quantile of KS, on the event $\{N^{a,p}([0, T_{max}]) = n_{tot}\}$.

Therefore, one can easily state that under H_0: " The process is a homogeneous Poisson process", the following upper bounds holds,

$$
\begin{aligned}
\mathbb{P}(\text{the previous test rejects } H_0) &= \sum_{n_{tot}=0}^{+\infty} \mathbb{P}(\text{the test rejects } H_0|N^{a,p}([0,T_{max}]) = n_{tot})\mathbb{P}(N^{a,p}([0,T_{max}]) = n_{tot}) \\
&\le \alpha \sum_{n_{tot}=0}^{+\infty} \mathbb{P}(N^{a,p}([0,T_{max}]) = n_{tot}) = \alpha.
\end{aligned}
$$

Now to turn this argument into an asymptotic argument and use $\sqrt{N^{a,p}([0, T_{max}]}\, \tilde{k}_{1-\alpha}$ instead of $k_{N^{a,p}([0,T_{max}],1-\alpha)}$, we need a random version of the convergence of $KS_{n_{tot}}$. Actually one can prove the following lemma, which shows that the previous replacement leads indeed to a test of asymptotic level α.

Lemma 1. *If the p processes are homogeneous Poisson processes, then*

$$\sqrt{N^{a,p}([0, T_{max}])} \sup_{t \in [0,1]} |F_{N^{a,p}([0,T_{max}])}(t) - t| \xrightarrow[p \to \infty]{\mathcal{L}} \mathcal{K}.$$

Proof. Let W be a variable whose distribution is $\mathcal{K}$. We set

$$Z_p = \sqrt{N^{a,p}([0, T_{max}])} \sup_{t \in [0,1]} |F_{N^{a,p}([0,T_{max}])}(t) - t|.$$

Let f be a bounded continuous function and let us consider for any positive integer n_{min},

$$
\begin{aligned}
|\mathbb{E}[f(Z_p)] - \mathbb{E}[f(W)]| &= \left| \sum_{n_{tot}=0}^{+\infty} \left(\mathbb{E}[f(Z_p)|N^{a,p}([0,T_{max}]) = n_{tot}] - \mathbb{E}[f(W)] \right) \mathbb{P}(N^{a,p}([0,T_{max}]) = n_{tot}) \right| \\
&\leq \sum_{n_{tot} \geq n_{min}}^{+\infty} \left| \mathbb{E}[f(Z_p)|N^{a,p}([0,T_{max}]) = n_{tot}] - \mathbb{E}[f(W)] \right| \mathbb{P}(N^{a,p}([0,T_{max}]) = n_{tot}) \\
&\quad + 2\|f\|_\infty \mathbb{P}(N^{a,p}([0,T_{max}]) < n_{min}).
\end{aligned}
$$

On the one hand, for any $\varepsilon > 0$, there exists n_{min} such that for any $n_{tot} \geq n_{min}$,

$$\left| \mathbb{E}[f(Z_p)|N^{a,p}([0,T_{max}]) = n_{tot}] - \mathbb{E}[f(W)] \right| < \varepsilon.$$

On the other hand, $N^{a,p}([0, T_{max}]) = \sum_{i=1}^{p} N_i([0, T_{max}])$ is a sum of p i.i.d. variables and therefore tends almost surely and in probability to infinity. Therefore there exists p_{min} such that for all $p > p_{min}$, $\mathbb{P}(N^{a,p}([0,T_{max}]) < n_{min}) < \varepsilon$, which implies that

$$|\mathbb{E}[f(Z_p)] - \mathbb{E}[f(W)]| \leq (1 + 2\|f\|_\infty)\varepsilon,$$

which proves the convergence in distribution. $\qquad\square$

As already stated, if N is an inhomogeneous Poisson process with compensator Λ, $\mathcal{N} = \{\Lambda(T) : T \in N\}$ is a homogeneous Poisson process with intensity 1 by the time-rescaling theorem [2, 4]. Assume now that we observe p i.i.d. Poisson processes N_i with compensator Λ. The previous transformation on each of the N_i leads to $\mathcal{N}_i$, the $\mathcal{N}_i$'s being p homogeneous Poisson processes of intensity 1 on $[0, \Lambda(T_{max})]$. One can therefore consider the aggregated process $\mathcal{N}^{a,p}$, to which we associate the c.d.f. $F_{\mathcal{N}^{a,p}([0,\Lambda(T_{max})])}$ as in (3). We can apply the previous lemma and we have:

$$\sqrt{\mathcal{N}^{a,p}([0, \Lambda(T_{max})])} \sup_{t \in [0,1]} |F_{\mathcal{N}^{a,p}([0,\Lambda(T_{max})])}(t) - t| \xrightarrow[p \to \infty]{\mathcal{L}} \mathcal{K}. \tag{30}$$

But, using the original aggregated process $N^{a,p}$, one can also write

$$F_{\mathcal{N}^{a,p}([0,\Lambda(T_{max})])}(t) \;=\; \frac{1}{\mathcal{N}^{a,p}([0,\Lambda(T_{max})])}\sum_{T\in\mathcal{N}^{a,p}}\mathbf{1}_{\{T/\Lambda(T_{max})\leq t\}}$$

$$=\; \frac{1}{N^{a,p}([0,T_{max}])}\sum_{X\in N^{a,p}}\mathbf{1}_{\{\Lambda(X)/\Lambda(T_{max})\leq t\}}.$$

The function $\Lambda(.)/\Lambda(T_{max})$ is a continuous c.d.f. from $[0,T_{max}]$ to $[0,1]$. Therefore we obtain:

$$\sqrt{N^{a,p}([0,T_{max}])}\sup_{x\in[0,T_{max}]}\left|\frac{1}{N^{a,p}([0,T_{max}])}\sum_{X\in N^{a,p}}\mathbf{1}_{\{X\leq x\}}-\frac{\Lambda(x)}{\Lambda(T_{max})}\right|\xrightarrow[p\to\infty]{\mathcal{L}}\mathcal{K}. \tag{31}$$

The end of the proof of Proposition 2 is then similar to the proof of Proposition 1.

9.1.4 Proof that Test 2 is asymptotically of level α

Here $\hat{F}=F_{N^{a,n}([0,T_{max}])}$. Hence, since

$$\sqrt{N^{a,n}([0,T_{max}])}\sup_x|F_{N^{a,n}([0,T_{max}])}(x)-F(x)|\xrightarrow[p\to\infty]{\mathcal{L}}\mathcal{K},$$

by the Slutsky's lemma, it is sufficient to prove that

$$N^{a,p(n)}([0,T_{max}])/N^{a,n}([0,T_{max}])\xrightarrow[p\to\infty]{\mathbb{P}}0, \tag{32}$$

and use Proposition 2 to conclude the proof. But since the numerator is equivalent to $p(n)\Lambda(T_{max})$ and the denominator to $n\Lambda(T_{max})$, by the law of large number, (32) is obvious.

9.1.5 Proof that Test 3 is asymptotically of level α

Here

$$\hat{F}(t)=\frac{\int_0^t\hat{\lambda}(u)du}{\int_0^{T_{max}}\hat{\lambda}(u)du}.$$

Then for all t one can write that

$$\hat{F}(t)-F(t)\;=\;\frac{\int_0^t\hat{\lambda}(u)du}{\int_0^{T_{max}}\hat{\lambda}(u)du}-\frac{\int_0^t\lambda(u)du}{\int_0^{T_{max}}\lambda(u)du}$$

$$=\;\frac{\int_0^t\hat{\lambda}(u)du}{\int_0^{T_{max}}\hat{\lambda}(u)du}-\frac{\int_0^t\lambda(u)du}{\int_0^{T_{max}}\hat{\lambda}(u)du}+\frac{\int_0^t\lambda(u)du}{\int_0^{T_{max}}\hat{\lambda}(u)du}-\frac{\int_0^t\lambda(u)du}{\int_0^{T_{max}}\lambda(u)du}$$

$$=\;\frac{\int_0^t[\hat{\lambda}(u)-\lambda(u)]du}{\int_0^{T_{max}}\hat{\lambda}(u)du}+\frac{\int_0^t\lambda(u)du}{\int_0^{T_{max}}\lambda(u)du}\frac{\int_0^{T_{max}}[\lambda(u)-\hat{\lambda}(u)]du}{\int_0^{T_{max}}\hat{\lambda}(u)du}.$$

Therefore

$$\sup_t|\hat{F}(t)-F(t)|\leq 2\frac{\int_0^{T_{max}}|\hat{\lambda}(u)-\lambda(u)|du}{\int_0^{T_{max}}\hat{\lambda}(u)du}.$$

Therefore $\sqrt{p(n)}\sup_t|\hat{F}(t)-F(t)|\xrightarrow[p\to\infty]{\mathbb{P}}0$ which ensures (4). Applying Proposition 2 concludes the proof.

9.1.6 Explicit construction of the cumulated process and its asymptotical properties

If N, as a general point process, has compensator Λ and conditional intensity λ, the time-rescaling theorem in its general form [2, 4, 31] states that $\mathcal{N} = \{X = \Lambda(T) : T \in N\}$ is an homogeneous Poisson process with intensity 1 until the time $\Lambda(T_{max})$ which is a random predictable time and therefore a stopping time. Here we observe p i.i.d. point processes N_i with intensity λ_i and compensator Λ_i. In particular, for all t, the $\Lambda_i(T_{max})$ are i.i.d. We apply the previous transformation to all the N_i's, hence generating the $\mathcal{N}_i$'s, p homogeneous Poisson processes on $[0, X_i^{max}]$ with the random stopping time $X_i^{max} = \Lambda_i(T_{max})$. Let $(\mathcal{N}_{i,x})_{x \geq 0} = (\mathcal{N}_i([0, x]))_{x \geq 0}$ be the corresponding counting process. We cumulate the counting processes in the following way: for any $x \leq \sum_{i=1}^p X_i^{max}$, we set:

$$\mathcal{N}_x^{c,p} = \sum_{i=1}^{k_x} \mathcal{N}_{i,X_i^{max}} + \mathcal{N}_{x-\sum_{i=1}^{k_x} X_i^{max}}^{(k_x)}, \tag{33}$$

where k_x is the only index in $\{0, ..., (p-1)\}$, such that

$$\sum_{i=1}^{k_x} X_i^{max} \leq x < \sum_{i=1}^{k_x+1} X_i^{max}.$$

Because each X_i^{max} is a stopping time, due to the strong Markov property of Poisson processes, the cumulation still guarantees that the jumps of $\mathcal{N}^{c,p}$, that are identified with the points of the point process $\mathcal{N}^{c,p}$, form an homogeneous Poisson process of intensity 1 on $[0, \sum_{i=1}^p X_i^{max}]$. Let us fix some $\theta > 0$ such that $\mathbb{E}[\Lambda_i(T_{max})] > \theta$. One can prove the following result.

Lemma 2. *For all $\theta > 0$ such that $\mathbb{E}(\Lambda_i(T_{max})) > \theta$,*

$$\sqrt{\mathcal{N}^{c,p}([0, p\theta])} \sup_{u \in [0,1]} \left| \frac{1}{\mathcal{N}^{c,p}([0, p\theta])} \sum_{X \in \mathcal{N}^{c,p}, X \leq p\theta} \mathbf{1}_{\{X/(p\theta) \leq u\}} - u \right| \xrightarrow[p \to \infty]{\mathcal{L}} \mathcal{K}.$$

Proof. Using the cumulation described in (33), let us complete $\mathcal{N}^{c,p}$ with another independent homogeneous Poisson process with intensity 1 and infinite support beyond $\sum_{i=1}^p X_i^{max}$, hence obtaining $\mathcal{N}'$ an homogeneous Poisson process of intensity 1 on $\mathbb{R}_+$. Let us denote

$$Z_p = \sqrt{\mathcal{N}^{c,p}([0, p\theta])} \sup_{u \in [0,1]} \left| \frac{1}{\mathcal{N}^{c,p}([0, p\theta])} \sum_{X \in \mathcal{N}^{c,p}, X \leq p\theta} \mathbf{1}_{\{X/(p\theta) \leq u\}} - u \right|.$$

We define Z'_p with the same expression except that $\mathcal{N}^{c,p}$ is replaced by $\mathcal{N}'$. The point process defined by $\{X/(p\theta)/X \in \mathcal{N}' \cap [0, p\theta]\}$ is a Poisson process with intensity $p\theta$ on $[0, 1]$. It can therefore also be viewed as an aggregated process of p i.i.d Poisson processes with intensity θ. Therefore Z'_p can be viewed as the quantity appearing in Lemma 1 showing that Z'_p tends in distribution to $\mathcal{K}$. So following the same proof, it remains

to show that for any bounded continuous function f, $|\mathbb{E}(f(Z_p)) - \mathbb{E}(f(Z_p'))|$ tends to 0. But $\mathcal{N}^{c,p} \cap [0, p\theta] = \mathcal{N}' \cap [0, p\theta]$ on the event $\{\sum_{i=1}^p X_i^{max} > p\theta\}$. Therefore $|\mathbb{E}(f(Z_p)) - \mathbb{E}(f(Z_p'))| \leq 2\|f\|_\infty \mathbb{P}(\sum_{i=1}^p X_i^{max} \leq p\theta)$. But by the law of large numbers,

$$\frac{1}{p} \sum_{i=1}^p X_i^{max} \xrightarrow[p\to\infty]{\mathbb{P}} \mathbb{E}(\Lambda_i(T_{max})) > \theta.$$

Hence $\mathbb{P}(\sum_{i=1}^p X_i^{max} \leq p\theta)$ tends to 0, which concludes the proof. $\qquad\square$

One can go back to the classical time t by introducing the cumulated point process

$$\forall t \geq 0, \quad N_t^{c,p} = \sum_{i=1}^{j_t} N_{i,T_{max}} + N_{j_t, t - j_t T_{max}},$$

where $j_t = \lfloor t/T_{max} \rfloor$. One can also introduce

$$\forall t \geq 0, \quad \Lambda^{c,p}(t) = \sum_{i=1}^{j_t} \Lambda_i(T_{max}) + \Lambda^{(j_t)}(t - j_t T_{max}).$$

The function $\Lambda^{c,p}(.)$ is a continuous non decreasing function and therefore, one can consider its generalized inverse function $(\Lambda^{c,p})^{-1}$. Therefore one can rewrite Lemma 2 as follows:

$$\sqrt{N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)])} \sup_{t \in [0, (\Lambda^{c,p})^{-1}(p\theta))]} \left| \frac{1}{N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)])} \sum_{T \in N^{c,p}, T \leq (\Lambda^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T \leq t\}} - \frac{\Lambda^{c,p}(t)}{p\theta} \right| \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{K}.$$

9.1.7 Proof of Theorem 1

Now we want to replace $\Lambda^{c,p}$ by an estimate of the type

$$\forall t \geq 0, \quad \hat{\Lambda}^{c,p}(t) = \sum_{i=1}^{j_t} \int_0^{T_{max}} \hat{\lambda}_i du + \int_0^{t - j_t T_{max}} \hat{\lambda}^{(j_t)} du.$$

Since $\hat{\Lambda}^{c,p}$ is also continuous and non-decreasing, one has the following equality:

$$\sup_{t \in [0, (\hat{\Lambda}^{c,p})^{-1}(p\theta)]} \left| \frac{1}{N^{c,p}((\hat{\Lambda}^{c,p})^{-1}(p\theta))} \sum_{T \in N^{c,p}, T \leq (\hat{\Lambda}^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T \leq t\}} - \frac{\hat{\Lambda}^{c,p}(t)}{p\theta} \right| =$$

$$\sup_{u \in [0,1]} \left| \frac{1}{\hat{\mathcal{N}}^{c,p}([0, p\theta])} \sum_{X \in \hat{\mathcal{N}}^{c,p}, X \leq p\theta} \mathbf{1}_{\{X/(p\theta) \leq u\}} - u \right|,$$

with $\hat{\mathcal{N}}^{c,p}$ the cumulated process obtained with $\hat{\mathcal{N}}_i = \{X = \int_0^T \hat{\lambda}_i(u) du : T \in N_i\}$ instead of $\mathcal{N}_i$ in (33).

Since $\Lambda^{c,p}(pT_{max})/p = \sum_i X_i^{max}/p$ tends in probability to $\mathbb{E}(\Lambda_i(T_{max})) > \theta$, with probability tending to 1, $p\theta$ belongs to $[0, \Lambda^{c,p}(pT_{max})]$. Moreover,

$$\left| \frac{\hat{\Lambda}^{c,p}(pT_{max}) - \Lambda^{c,p}(pT_{max})}{p} \right| \leq \frac{1}{p} \sum_{i=1}^p \int_0^{T_{max}} \left| \hat{\lambda}_i(u) - \lambda_i(u) \right| du,$$

54

where the right hand side tends to 0 in probability. Therefore $\hat{\Lambda}^{c,p}(pT_{max})/p$ also tends in probability to $\mathbb{E}(\Lambda_i(T_{max})) > \theta$ and with probability tending to 1, $p\theta$ belongs to $[0, \hat{\Lambda}^{c,p}(pT_{max})]$. Therefore both $(\hat{\Lambda}^{c,p})^{-1}(p\theta)$ and $(\Lambda^{c,p})^{-1}(p\theta)$ are strictly smaller than pT_{max} with probability tending to 1. Furthermore, if $(\hat{\Lambda}^{c,p})^{-1}(p\theta) < pT_{max}$ and $(\Lambda^{c,p})^{-1}(p\theta) < pT_{max}$,

$$
\begin{aligned}
\left| \Lambda^{c,p}((\hat{\Lambda}^{c,p})^{-1}(p\theta)) - \Lambda^{c,p}((\Lambda^{c,p})^{-1}(p\theta)) \right| &= \left| \Lambda^{c,p}((\hat{\Lambda}^{c,p})^{-1}(p\theta)) - p\theta \right| \\
&= \left| \Lambda^{c,p}((\hat{\Lambda}^{c,p})^{-1}(p\theta)) - \hat{\Lambda}^{c,p}((\hat{\Lambda}^{c,p})^{-1}(p\theta)) \right| \\
&\leq \sum_{i=1}^{p} \int_0^{T_{max}} \left| \hat{\lambda}_i(u) - \lambda_i(u) \right| du.
\end{aligned}
$$

Hence, by assumption, for $\delta > 0$, if

$$
\Omega_\delta = \left\{ \left| \Lambda^{c,p}((\hat{\Lambda}^{c,p})^{-1}(p\theta)) - \Lambda^{c,p}((\Lambda^{c,p})^{-1}(p\theta)) \right| \leq \delta \sqrt{p} \right\},
$$

for any $\varepsilon > 0$, there exists p_0, such that for any $p \geq p_0$, $\mathbb{P}(\Omega_\delta) \geq 1 - \epsilon$. On Ω_δ, one has therefore that

$$
p\theta - \delta\sqrt{p} \leq \Lambda^{c,p}((\hat{\Lambda}^{c,p})^{-1}(p\theta)) \leq p\theta + \delta\sqrt{p}
$$

and

$$
\begin{aligned}
\left| N^{c,p}([0, (\hat{\Lambda}^{c,p})^{-1}(p\theta)]) - N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)]) \right| &= \left| \mathcal{N}^{c,p}([0, \Lambda^{c,p}((\hat{\Lambda}^{c,p})^{-1}(p\theta))]) - \mathcal{N}^{c,p}([0, p\theta]) \right| \\
&\leq \max \left[\mathcal{N}^{c,p}([0, p\theta + \delta\sqrt{p}]) - \mathcal{N}^{c,p}([0, p\theta]) , \ \mathcal{N}^{c,p}([0, p\theta]) - \mathcal{N}^{c,p}([0, p\theta - \delta\sqrt{p}]) \right].
\end{aligned}
$$

In the previous expression, we consider the maximum of two independent Poisson variables (denoted U and V) with parameter $\delta\sqrt{p}$. For any $u > 0$,

$$
\begin{aligned}
\mathbb{P}\left[\left| N^{c,p}([0, (\hat{\Lambda}^{c})^{-1}(p\theta)]) - N^{c,p}([0, (\Lambda^{c})^{-1}(p\theta)]) \right| \geq (\delta + u)\sqrt{p} \text{ or } \Omega_\delta^c \right] &\leq \epsilon + \mathbb{P}(\max\{U, V\} \geq (\delta + u)\sqrt{p}) \\
&\leq \epsilon + 2\exp\left(-\frac{pu^2}{2\sqrt{p}\delta + \sqrt{p}u} \right).
\end{aligned}
$$

By taking $u = \delta$, the last expression shows that

$$
p^{-1/2}|N^{c,p}([0, (\hat{\Lambda}^{c,p})^{-1}(p\theta)]) - N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)])| \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{34}
$$

If we are able to show that

$$
\sqrt{\mathcal{N}^{c,p}([0, p\theta])} \sup_{u\in[0,1]} \left| \frac{1}{\hat{\mathcal{N}}^{c,p}([0, p\theta])} \sum_{X\in\hat{\mathcal{N}}^{c,p}, X\leq p\theta} \mathbf{1}_{\{X/(p\theta)\leq u\}} - u \right| \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{K},
$$

since $N^{c,p}([0,(\hat{\Lambda}^{c,p})^{-1}(p\theta)])/N^{c,p}([0,(\Lambda^{c,p})^{-1}(p\theta)])$ tends to 1 in probability, this will imply the result by using Slustky's Lemma. Now, we clip $\Lambda^{c,p}$ and $\hat{\Lambda}^{c,p}$ and we set for any t,

$$\bar{\Lambda}^{c,p}(t) = \min(\Lambda^{c,p}(t), p\theta) \quad \text{and} \quad \bar{\hat{\Lambda}}^{c,p}(t) = \min(\hat{\Lambda}^{c,p}(t), p\theta).$$

Therefore, $\bar{\Lambda}^{c,p}(\cdot)/(p\theta)$ and $\bar{\hat{\Lambda}}^{c,p}(\cdot)/(p\theta)$ are continuous c.d.f. and

$$\sup_{u\in[0,1]} \left| \frac{1}{\hat{\mathcal{N}}^{c,p}([0,p\theta])} \sum_{X\in\hat{\mathcal{N}}^{c,p}, X\le p\theta} \mathbf{1}_{\{X/(p\theta)\le u\}} - u \right| = \sup_{t>0} \left| \frac{1}{N^{c,p}([0,(\bar{\hat{\Lambda}}^{c,p})^{-1}(p\theta)])} \sum_{T\in N^{c,p}, T\le(\bar{\hat{\Lambda}}^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T\le t\}} - \frac{\bar{\hat{\Lambda}}^{c,p}(t)}{p\theta} \right|.$$

But since $p\theta$ belongs to $[0,\hat{\Lambda}^{c,p}(pT_{max})]$ with probability tending to 1, the right hand side is equal to

$$A_p = \sup_{t>0} \left| \frac{1}{N^{c,p}([0,(\hat{\Lambda}^{c,p})^{-1}(p\theta)])} \sum_{T\in N^{c,p}, T\le(\hat{\Lambda}^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T\le t\}} - \frac{\bar{\hat{\Lambda}}^{c,p}(t)}{p\theta} \right|$$

with probability tending to 1. So it is sufficient to prove that $\hat{Z}_p := \sqrt{\mathcal{N}^{c,p}([0,p\theta])} A_p$ tends in distribution to $\mathcal{K}$. Note that

$$\hat{Z}_p = \sqrt{N^{c,p}([0,(\Lambda^{c,p})^{-1}(p\theta)])} \sup_{t>0} \left| \frac{1}{N^{c,p}([0,(\hat{\Lambda}^{c,p})^{-1}(p\theta)])} \sum_{T\in N^{c,p}, T\le(\hat{\Lambda}^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T\le t\}} - \frac{\bar{\hat{\Lambda}}^{c,p}(t)}{p\theta} \right|.$$

We denote:

$$\tilde{Z}_p = \sqrt{N^{c,p}([0,(\Lambda^{c,p})^{-1}(p\theta)])} \sup_{t>0} \left| \frac{1}{N^{c,p}([0,(\Lambda^{c,p})^{-1}(p\theta)])} \sum_{T\in N^{c,p}, T\le(\Lambda^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T\le t\}} - \frac{\bar{\hat{\Lambda}}^{c,p}(t)}{p\theta} \right|$$

$$Z_p = \sqrt{N^{c,p}([0,(\Lambda^{c,p})^{-1}(p\theta)])} \sup_{t>0} \left| \frac{1}{N^{c,p}([0,(\Lambda^{c,p})^{-1}(p\theta)])} \sum_{T\in N^{c,p}, T\le(\Lambda^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T\le t\}} - \frac{\bar{\Lambda}^{c,p}(t)}{p\theta} \right|.$$

But Z_p is also equal with probability tending to 1 to

$$\sqrt{N^{c,p}([0,(\Lambda^{c,p})^{-1}(p\theta)])} \sup_{t>0} \left| \frac{1}{N^{c,p}([0,(\bar{\Lambda}^{c,p})^{-1}(p\theta)])} \sum_{T\in N^{c,p}, T\le(\bar{\Lambda}^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T\le t\}} - \frac{\bar{\Lambda}^{c,p}(t)}{p\theta} \right| =$$

$$\sqrt{N^{c,p}([0,(\Lambda^{c,p})^{-1}(p\theta)])} \sup_{u\in[0,1]} \left| \frac{1}{\mathcal{N}^{c,p}([0,p\theta])} \sum_{X\in\mathcal{N}^{c,P}, X\le p\theta} \mathbf{1}_{\{X/(p\theta)\le u\}} - u \right|,$$

because $p\theta$ belongs to $[0,\Lambda^{c,p}(pT_{max})]$ with probability tending to 1. Using Lemma 2 as before, we have that Z_p tends in distribution to $\mathcal{K}$. It is consequently sufficient to prove that $\hat{Z}_p - \tilde{Z}_p$ and $\tilde{Z}_p - Z_p$ tend both in

probability to 0. We have:

$$
\begin{aligned}
\left| \hat{Z}_p - \tilde{Z}_p \right| & \leq \sqrt{N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)])} \sup_{t>0} \left| \frac{1}{N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)])} \sum_{T \in N^{c,p}, T \leq (\Lambda^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T \leq t\}} \right. \\
& \qquad \left. - \frac{1}{N^{c,p}([0, (\hat{\Lambda}^{c,p})^{-1}(p\theta)])} \sum_{T \in N^{c,p}, T \leq (\hat{\Lambda}^{c,p})^{-1}(p\theta)} \mathbf{1}_{\{T \leq t\}} \right| \\
& \leq 2\sqrt{\mathcal{N}^{c,p}([0, p\theta])} \frac{\left| N^{c,p}([0, (\hat{\Lambda}^{c,p})^{-1}(p\theta)]) - N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)]) \right|}{\max\left(N^{c,p}([0, (\hat{\Lambda}^{c,p})^{-1}(p\theta)]), N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)]) \right)},
\end{aligned}
$$

which tends to 0 in probability by (34). Furthermore, if $(\hat{\Lambda}^{c,p})^{-1}(p\theta) < pT_{max}$ and $(\Lambda^{c,p})^{-1}(p\theta) < pT_{max}$,

$$
\begin{aligned}
\left| \tilde{Z}_p - Z_p \right| & \leq \sqrt{N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)])} \sup_{t>0} \left| \frac{\bar{\hat{\Lambda}}^{c,p}(t)}{p\theta} - \frac{\bar{\Lambda}^{c,p}(t)}{p\theta} \right| \\
& \leq \frac{\sqrt{N^{c,p}([0, (\Lambda^{c,p})^{-1}(p\theta)])}}{p\theta} \sup_{t \leq pT_{max}} \left| \hat{\Lambda}^{c,p}(t) - \Lambda^{c,p}(t) \right| \\
& \leq \frac{\sqrt{\mathcal{N}^{c,p}([0, p\theta])}}{p\theta} \sum_{i=1}^{p} \int_0^{T_{max}} \left| \hat{\lambda}_i(u) - \lambda_i(u) \right| du.
\end{aligned}
$$

Since $\mathcal{N}^{c,p}([0, p\theta])$ is also the sum of p i.i.d. Poisson variables, it behaves like a multiple of p by the law of large numbers. Therefore, by using (7), $\tilde{Z}_p - Z_p \xrightarrow[p \to \infty]{\mathbb{P}} 0$.

9.1.8 Proof that Test 4 is of level α asymptotically

We apply Theorem 1 with $\hat{\lambda}_i = ((\lambda_i)_{\hat{f}})_+$. Since the λ_i's are positive, one has that for all u,

$$
\left| \hat{\lambda}_i(u) - \lambda_i(u) \right| \leq \left| (\lambda_i)_{\hat{f}}(u) - \lambda_i(u) \right|,
$$

which gives exactly (7).

9.2 Additional file 2 — Proof of Theorem 2

In this additional file, the intensity $\lambda(.)$ should be understood as a function on the whole real line which is null outside $[0, T_{max}]$. In particular, the Poisson process $N^{a,n}$, with intensity $n\lambda(.)$, exists now on $\mathbb{R}$, but there is no points outside $[0, T_{max}]$ and $N^{a,n}(\mathbb{R}) = N^{a,n}([0, T_{max}])$. The proof is inspired by Proposition 2 of [65]. We set:

$$
\chi = (1 + \eta)(1 + \|K\|_1)\|K\|_2.
$$

Therefore, we have:

$$
A(h) = \sup_{h' \in \mathcal{H}} \left\{ \|\hat{\lambda}_n^{h,h'} - \hat{\lambda}_n^{K_{h'}}\|_2 - \frac{\chi \sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h'}} \right\}_+
$$

and

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ A(h) + \frac{\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h}} \right\}.$$

For any $h \in \mathcal{H}$,

$$\|\hat{\lambda}_n^{GL} - \lambda\|_2 \leq A_1 + A_2 + A_3,$$

with

$$A_1 := \|\hat{\lambda}_n^{GL} - \hat{\lambda}_n^{\hat{h},h}\|_2 \leq A(h) + \frac{\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{\hat{h}}},$$

$$A_2 := \|\hat{\lambda}_n^{\hat{h},h} - \hat{\lambda}_n^{K_h}\|_2 \leq A(\hat{h}) + \frac{\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h}}$$

and

$$A_3 := \|\hat{\lambda}_n^{K_h} - \lambda\|_2.$$

By definition of $\hat{h}$, we have:

$$A_1 + A_2 \leq 2A(h) + \frac{2\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h}}.$$

Therefore, by setting

$$\zeta_n(h) := \sup_{h' \in \mathcal{H}} \left\{ \|(\hat{\lambda}_n^{h,h'} - \mathbb{E}[\hat{\lambda}_n^{h,h'}]) - (\hat{\lambda}_n^{K_{h'}} - \mathbb{E}[\hat{\lambda}_n^{K_{h'}}])\|_2 - \frac{\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h'}} \right\}_+$$

we have:

$$
\begin{aligned}
A_1 + A_2 \;&\leq\; 2\zeta_n(h) + 2\sup_{h' \in \mathcal{H}} \|\mathbb{E}[\hat{\lambda}_n^{h,h'}] - \mathbb{E}[\hat{\lambda}_n^{K_{h'}}]\|_2 + \frac{2\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h}} \\
&\leq\; 2\zeta_n(h) + 2\sup_{h' \in \mathcal{H}} \|K_h \star K_{h'} \star \lambda - K_{h'} \star \lambda\|_2 + \frac{2\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h}} \\
&\leq\; 2\zeta_n(h) + 2\|K\|_1\|K_h \star \lambda - \lambda\|_2 + \frac{2\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h}}.
\end{aligned}
$$

Finally, since $(a + b + c)^2 \leq 3a^2 + 3b^2 + 3c^2$,

$$
\begin{aligned}
\mathbb{E}[(A_1 + A_2)^2] \;&\leq\; 12\,\mathbb{E}[\zeta_n^2(h)] + 12\|K\|_1^2\|K_h \star \lambda - \lambda\|_2^2 + \frac{12\chi^2\,\mathbb{E}[N^{a,n}(\mathbb{R})]}{n^2 h} \\
&\leq\; 12\,\mathbb{E}[\zeta_n^2(h)] + 12\|K\|_1^2\|K_h \star \lambda - \lambda\|_2^2 + \frac{12\chi^2\|\lambda\|_1}{nh}.
\end{aligned}
$$

For the last term, we obtain:

$$
\begin{aligned}
\mathbb{E}[A_3^2] \;&=\; \int \mathrm{Var}(\hat{\lambda}_n^{K_h}(x))dx + \|K_h \star \lambda - \lambda\|_2^2 \\
&=\; \frac{1}{n^2} \iint K_h^2(x - u)n\lambda(u)dudx + \|K_h \star \lambda - \lambda\|_2^2 \\
&=\; \frac{\|\lambda\|_1}{nh}\|K\|_2^2 + \|K_h \star \lambda - \lambda\|_2^2.
\end{aligned}
$$

58

Finally, replacing χ with its definition, we obtain: for any $h \in \mathcal{H}$,

$$
\begin{aligned}
\mathbb{E}\|\hat{\lambda}_n^{GL} - \lambda\|_2^2 &\leq 2\mathbb{E}[(A_1 + A_2)^2] + 2\mathbb{E}[A_3^2] \\
&\leq 2(1 + 12\|K\|_1^2)\|K_h \star \lambda - \lambda\|_2^2 + 2(1 + 12(1 + \eta)^2(1 + \|K\|_1)^2)\frac{\|\lambda\|_1}{nh}\|K\|_2^2 + 24\mathbb{E}[\zeta_n^2(h)].
\end{aligned}
$$

The result follows by using the next lemma.

Lemma 3. *If $\mathcal{H} \subset \{D^{-1} : D = 1, ..., D_{\max}\}$ with $D_{\max} = \delta n$ for some $\delta > 0$, and if $\|\lambda\|_\infty < \infty$, then there exists a constant C depending on δ, η, $\|K\|_2$, $\|K\|_1$, $\|\lambda\|_1$ and $\|\lambda\|_\infty$ such that for any $h \in \mathcal{H}$*

$$
\mathbb{E}[\zeta_n^2(h)] \leq Cn^{-1}.
$$

Proof. First, we notice that for any $h \in \mathcal{H}$

$$
\begin{aligned}
\zeta_n(h) &\leq \sup_{h' \in \mathcal{H}} \left\{ \|\hat{\lambda}_n^{h,h'} - \mathbb{E}[\hat{\lambda}_n^{h,h'}]\|_2 + \|\hat{\lambda}_n^{K_{h'}} - \mathbb{E}[\hat{\lambda}_n^{K_{h'}}]\|_2 - \frac{\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h'}} \right\}_+ \\
&\leq \sup_{h' \in \mathcal{H}} \left\{ (\|K\|_1 + 1)\|\hat{\lambda}_n^{K_{h'}} - \mathbb{E}[\hat{\lambda}_n^{K_{h'}}]\|_2 - \frac{\chi\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h'}} \right\}_+ \\
&\leq (\|K\|_1 + 1)S_n,
\end{aligned}
$$

with

$$
S_n := \sup_{h \in \mathcal{H}} \left\{ \|\hat{\lambda}_n^{K_h} - \mathbb{E}[\hat{\lambda}_n^{K_h}]\|_2 - \frac{\|K\|_2(1 + \eta)\sqrt{N^{a,n}(\mathbb{R})}}{n\sqrt{h}} \right\}_+.
$$

We then have:

$$
\mathbb{E}[\zeta_n^2(h)] \leq (\|K\|_1 + 1)^2(A + B)
$$

with

$$
A := \mathbb{E}[S_n^2 \mathbf{1}_{\{N^{a,n}(\mathbb{R}) \leq (1-\alpha)^2 n\|\lambda\|_1\}}],
$$

$$
B := \mathbb{E}[S_n^2 \mathbf{1}_{\{N^{a,n}(\mathbb{R}) > (1-\alpha)^2 n\|\lambda\|_1\}}]
$$

for all $\alpha \in (0, 1)$. We have:

$$
\begin{aligned}
S_n^2 &\leq 2\sup_{h \in \mathcal{H}} \|\hat{\lambda}_n^{K_h}\|_2^2 + 2\sup_{h \in \mathcal{H}} \|\mathbb{E}[\hat{\lambda}_n^{K_h}]\|_2^2 \\
&\leq 2n^{-2}\sup_{h \in \mathcal{H}} \int dx \left(\int K_h(x - u)dN^{a,n}(u) \right)^2 + 2\sup_{h \in \mathcal{H}} \|K_h \star \lambda\|_2^2 \\
&\leq 2n^{-2}\sup_{h \in \mathcal{H}} \iint K_h^2(x - u)dN^{a,n}(u))dx \times N^{a,n}(\mathbb{R}) + 2\sup_{h \in \mathcal{H}} \|K_h \star \lambda\|_2^2 \\
&\leq 2n^{-2}\sup_{h \in \mathcal{H}} \frac{\|K\|_2^2}{h} \times N^{a,n}(\mathbb{R})^2 + 2\sup_{h \in \mathcal{H}} \frac{\|K\|_2^2}{h}\|\lambda\|_1^2 \\
&\leq 2\delta n\|K\|_2^2 \left(\frac{N^{a,n}(\mathbb{R})^2}{n^2} + \|\lambda\|_1^2 \right).
\end{aligned}
$$

Therefore,

$$A \leq 4\delta n \|K\|_2^2 \|\lambda\|_1^2 \times \mathbb{P}(N^{a,n}(\mathbb{R}) \leq (1-\alpha)^2 n \|\lambda\|_1).$$

To bound the last term, we use, for instance, Inequality (5.2) of [41] (with $\xi = (2\alpha - \alpha^2)n\|\lambda\|_1$ and with the function $f \equiv -1$), which shows that there exists $\alpha' > 0$ only depending on α such that

$$\mathbb{P}(N^{a,n}(\mathbb{R}) \leq (1-\alpha)^2 n\|\lambda\|_1) \leq \exp(-\alpha'\|\lambda\|_1 \times n).$$

This shows that

$$A \leq C_A n^{-1},$$

where C_A depends on α, δ, $\|K\|_2$ and $\|\lambda\|_1$. Now, we deal with the term B by fixing the previous value α: we set $\alpha = \min(\eta/2, 1/4)$. This implies

$$(1+\eta)(1-\alpha) \geq 1 + \frac{\eta}{4}$$

and

$$
\begin{aligned}
B &= \mathbb{E}[S_n^2 1_{\{N^{a,n}(\mathbb{R}) > (1-\alpha)^2 n\|\lambda\|_1\}}] \\
&\leq \mathbb{E}\left[\sup_{h \in \mathcal{H}} \left\{\|\hat{\lambda}_n^{K_h} - \mathbb{E}[\hat{\lambda}_n^{K_h}]\|_2 - \frac{(1+\frac{\eta}{4})\|K\|_2\sqrt{\|\lambda\|_1}}{\sqrt{nh}}\right\}_+^2\right] \\
&= \int_0^{+\infty} \mathbb{P}\left(\sup_{h \in \mathcal{H}} \left\{\|\hat{\lambda}_n^{K_h} - \mathbb{E}[\hat{\lambda}_n^{K_h}]\|_2 - \frac{(1+\frac{\eta}{4})\|K\|_2\sqrt{\|\lambda\|_1}}{\sqrt{nh}}\right\}_+^2 \geq x\right) dx \\
&\leq \sum_{h \in \mathcal{H}} \int_0^{+\infty} \mathbb{P}\left(\left\{\|\hat{\lambda}_n^{K_h} - \mathbb{E}[\hat{\lambda}_n^{K_h}]\|_2 - \frac{(1+\frac{\eta}{4})\|K\|_2\sqrt{\|\lambda\|_1}}{\sqrt{nh}}\right\}_+^2 \geq x\right) dx.
\end{aligned}
$$

To conclude, it remains to control for any $x \geq 0$, the probability inside the integral. For this purpose, we apply Corollary 2 of [41] and we set

$$U(x) = \hat{\lambda}_n^{K_h}(x) - \mathbb{E}[\hat{\lambda}_n^{K_h}(x)] = \frac{1}{n}\int K_h(x-u)dN^{a,n}(u) - (K_h \star \lambda)(x).$$

If $\mathcal{A}$ is a countable dense subset of the unit ball of $\mathbb{L}_2(\mathbb{R})$, we have:

$$
\begin{aligned}
\|U\|_2 &= \sup_{a \in \mathcal{A}} \int a(t)U(t)dt \\
&= \sup_{a \in \mathcal{A}} \int a(t)\left(\frac{1}{n}\int K_h(t-u)dN^{a,n}(u) - \int K_h(t-u)\lambda(u)du\right)dt \\
&= \sup_{a \in \mathcal{A}} \int \psi_a(u)(dN^{a,n}(u) - n\lambda(u))du,
\end{aligned}
$$

with for any $a \in \mathcal{A}$ and any $u \in \mathbb{R}$,

$$\psi_a(u) = \frac{1}{n} \int a(t) K_h(t-u) dt.$$

We have for any u,

$$\psi_a^2(u) \leq \frac{1}{n^2} \int a^2(t) dt \int K_h^2(t-u) dt = \frac{\|K\|_2^2}{n^2 h}.$$

So, if $b = \frac{\|K\|_2}{n\sqrt{h}}$, $\|\psi_a\|_\infty \leq b$. We have

$$\mathbb{E}[\|U\|_2^2] = \int \mathbb{E}[U^2(t)] dt = \int \mathrm{var}(\hat{\lambda}_n^{K_h}(t)) dt = \frac{\|\lambda\|_1}{nh} \|K\|_2^2,$$

which implies

$$\mathbb{E}[\|U\|_2] \leq \frac{\sqrt{\|\lambda\|_1}}{\sqrt{nh}} \|K\|_2.$$

Finally,

$$
\begin{aligned}
v &:= \sup_{a \in \mathcal{A}} \int \psi_a^2(x) n\lambda(x) dx \\
&= \sup_{a \in \mathcal{A}} \int \left(\int \frac{a(t)}{n} K_h(t-x) dt \right)^2 n\lambda(x) dx \\
&\leq \frac{1}{n} \int \lambda(x) dx \int a^2(t) |K_h(t-x)| dt \int |K_h(t-x)| dt \\
&\leq \frac{\|K\|_1^2 \|\lambda\|_\infty}{n}.
\end{aligned}
$$

Corollary 2 of [41] gives: for any $\epsilon > 0$, for any $u > 0$,

$$\mathbb{P}\left(\|\hat{\lambda}_n^{K_h} - \mathbb{E}[\hat{\lambda}_n^{K_h}]\|_2 \geq (1+\epsilon) \frac{\sqrt{\|\lambda\|_1}}{\sqrt{nh}} \|K\|_2 + \sqrt{\frac{12}{n} \|K\|_1^2 \|\lambda\|_\infty u} + \left(\frac{5}{4} + \frac{32}{\epsilon} \right) \frac{\|K\|_2}{n\sqrt{h}} u \right) \leq \exp(-u).$$

Then, we conclude by using exactly the same computation as in the proof of Lemma 1 of [66, p 32-34] that gives:

$$B \leq C_B n^{-1},$$

where C_B is a constant depending on δ, η, $\|K\|_2$, $\|K\|_1$ and $\|\lambda\|_\infty$. $\qquad\square$

The lemma leads to the result.

9.3 Additional file 3 — Adaptive properties of the Lasso estimate for Hawkes processes

Theorem 1 of [58] does not explicit particular basis on which the interaction functions are expanded. It is stated in its general form as follows. Note that [58] is a proceedings and therefore no proof has been given of the result in [58]. If n i.i.d. trials are recorded, each trial i corresponds to the observation of $N_i =$

$(N_i^{(1)}, ..., N_i^{(M)})$, the multivariate Hawkes process whose intensity is given by the predictable transformation denoted ψ_i. Furthermore, to each trial i, we can associate an intensity λ_i and a contrast $\gamma^{(i)}$. The global least-squares contrast over the n trials can also be seen as

$$\gamma_n(f) = \sum_{i=1}^{n} \gamma^{(i)}(f). \tag{35}$$

We use the following notation: for any predictable processes $H = (H_i^{(1)}, ..., H_i^{(M)})_{i=1,...n}$, $K = (K_i^{(1)}, ..., K_i^{(M)})_{i=1,...,n}$, set

$$H \bullet N = \sum_{i=1}^{n} \sum_{m=1}^{M} \int_{T_1}^{T_2} H_{i,t}^{(m)} dN_i^{(m)}(t), \tag{36}$$

$$H \diamond K = \sum_{i=1}^{n} \sum_{m=1}^{M} \int_{T_1}^{T_2} H_{i,t}^{(m)} K_{i,t}^{(m)} dt, \tag{37}$$

and $H^{\diamond 2} = H \diamond H$.

In general, we use a dictionary Φ of known functions of $\mathcal{H}$ and we only consider linear combinations of functions of Φ for estimating f^*:

$$f_a = \sum_{\varphi \in \Phi} a_\varphi \varphi, \quad \text{for} \quad a \in \mathbb{R}^\Phi. \tag{38}$$

Then, by linearity of ψ, one can rewrite (35) as

$$\gamma_n(f_a) = -2a'b_n + a'G_n a, \tag{39}$$

where for any φ and $\tilde{\varphi}$ in Φ,

$$(b_n)_\varphi = \psi(\varphi) \bullet N \quad \text{and} \quad (G_n)_{\varphi, \tilde{\varphi}} = \psi(\varphi) \diamond \psi(\tilde{\varphi}).$$

Given a vector of positive weights d, the *Lasso estimate* of f^* is $\tilde{f}_n := f_{\tilde{a}_n}$ where $\tilde{a}_n$ is a minimizer of the following ℓ_1-penalized least-square contrast:

$$\tilde{a}_n \in \arg\min_{a \in \mathbb{R}^\Phi} \{-2a'b_n + a'G_n a + 2d'|a|\}. \tag{40}$$

Then Theorem 1 of [58] is stated as follows:

Theorem 3. *We introduce the following two events:*

$$\Omega_{V,B} = \{\forall \varphi \in \Phi, \sup_{t \in [T_1, T_2], m, i} |\psi_{i,t}^{(m)}(\varphi)| \leq B_\varphi \text{ and } (\psi(\varphi))^2 \bullet N \leq V_\varphi\},$$

for positive deterministic constants B_φ and V_φ and

$$\Omega_c = \left\{\forall a \in \mathbb{R}^\Phi, \quad a'G_n a \geq c\, a'a\right\}, \tag{41}$$

for a positive constant c. Let x and ε be strictly positive constants and for all $\varphi \in \Phi$,

$$d_\varphi = \sqrt{2(1+\varepsilon)\hat{V}_\varphi^\mu x} + \frac{B_\varphi x}{3}, \tag{42}$$

with

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

for a real number μ such that $\mu > \phi(\mu)$, where $\phi(\mu) = \exp(\mu) - \mu - 1$. Then, with probability larger than

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

the following inequality holds

$$[\psi(\tilde{f}_n) - \lambda]^{\diamond 2} \leq C \inf_{a \in \mathbb{R}^\Phi} \left\{ [\psi(f_a) - \lambda]^{\diamond 2} + \frac{1}{c} \sum_{\varphi \in S(a)} d_\varphi^2 \right\},$$

where C is an absolute positive constant and where $S(a)$ is the support of a, i.e. its coordinates with non-zero coefficients.

Proof. We use the notation of [57] and transposition of this notation. First by scaling the data, it is always possible to assume that $A = 1$. We have at hand $n \times M$ point processes $N_m^{(i)}$. In the more general case, we need to model each $\lambda^{(m,i)}$, intensity of $N_m^{(i)}$, by a

$$\psi_{f_n}^{(m,i)}(t) = \mu^{(m,i)} + \sum_{\ell,j} \int_{-\infty}^{t-} g_{\ell,j}^{(m,i)}(t-u) dN_j^{(\ell)}(u),$$

where f_n belongs to $\mathcal{H}_n$ which replaces the space $\mathcal{H}$:

$$\mathcal{H}_n = (\mathbb{R} \times \mathbb{L}_2((0,1])^{nM})^{nM} = \left\{ f_n = \left((\mu^{(m,i)}, (g_{\ell,j}^{(m,i)})_{\ell=1,\ldots,M, \ j=1,\ldots,n})_{m=1,\ldots,M, \ i=1,\ldots n} \right) : \right.$$

$$\left. g_{\ell,j}^{(m,i)} \text{ with support in } (0,1] \text{ and } \|f_n\|^2 = \sum_{m,i}(\mu^{(m,i)})^2 + \sum_{m,i}\sum_{\ell,j} \int_0^1 g_{\ell,j}^{(m,i)}(t)^2 dt < \infty \right\}.$$

For every $f_n = \left((\mu^{(m,i)}, (g_{\ell,j}^{(m,i)})_{\ell=1,\ldots,M, \ j=1,\ldots,n})_{m=1,\ldots,M, \ i=1,\ldots n} \right)$ in $\mathcal{H}_n$, we denote for each m and i,

$$f_n^{(m,i)} = (\mu^{(m,i)}, (g_{\ell,j}^{(m,i)})_{\ell=1,\ldots,M, \ j=1,\ldots,n}).$$

In the same way for every $f = \left((\mu^{(m)}, (g_{\ell,j}^{(m)})_{\ell=1,\ldots,M, \ j=1,\ldots,n}) \right)_{m=1,\ldots,M}$ in $\mathcal{H}$, we denote for each m and i,

$$f^{(m)} = (\mu^{(m)}, (g_{\ell,j}^{(m)})_{\ell=1,\ldots,M, \ j=1,\ldots,n}).$$

Now our dictionary Φ of $\mathcal{H}$ can be transformed into a dictionary Φ_n of $\mathcal{H}_n$ by stating that for any φ in Φ we associate a φ_n in $\mathcal{H}_n$ such that for all i,m, $\varphi_n^{(m,i)} = \varphi^{(m)}$. Therefore it is easy to see that the vector b

of [57] associated to f_n is actually our vector b_n and that the matrix G of [57] is our matrix G_n. The V and B are in the same way translated and the present result is a pure application of Theorem 2 of [57]   □