Spike trains as (in)homogeneous Poisson processes or Hawkes processes: non-parametric adaptive estimation and goodness-of-fit tests

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

Spike trains as (in)homogeneous Poisson processes or Hawkes processes: non-parametric adaptive estimation and goodness-of-fit tests

Patricia Reynaud-Bouret*[1] and Christine Tuleau-Malot[1] and Vincent Rivoirard[2] and Franck Grammont[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; Christine Tuleau-Malot - malot@unice.fr; Vincent Rivoirard - rivoirard@ceremade.dauphine.fr; Franck Grammont - grammont@unice.fr;

*Corresponding author

Abstract

This article aims to propose new non-parametric adaptive estimation methods and to adapt other recent similar results to the setting of spike trains analysis. After briefly recalling main features of the homogeneous Poisson model, we focus on two main generalizations of this process : the inhomogeneous Poisson model, which is non-stationary, and the Hawkes model, which can take into account interactions. Goodness-of-fit tests are also proposed and are proved to be of prescribed asymptotical level. They enable us to test these non-parametric models. Various simulations show good performance of the estimation and test procedures. A complete analysis is also performed with these tools on single unit activity recorded on a monkey during a sensory-motor task. We can show that the homogeneous Poisson process hypothesis is always rejected and that the inhomogeneous Poisson process hypothesis is rarely accepted. The Hawkes model seems to fit most of the data.

1 Introduction

In Neurosciences, 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 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 [1] for a review). However, except in very particular tests of independence (see for instance [2, 3]), it is most of the time not sufficient to only state that the spike trains are point processes. To access more meaningful quantities, some assumptions need to be made.

The first naive model that has been used is the *homogeneous Poisson process*, which is the simplest process among point processes (see for instance [4]). In this model, parametrization is only based on the firing rate λ of the process. Due to their simplicity, homogeneous Poisson processes were extensively implicitly used in Neurosciences. In particular, the study of those firing rates leads to significative advances in the understanding of the coding of the direction of movements (see [5]). But for homogeneous Poisson processes, the firing rate λ is constant so several features of spike trains cannot be modeled and in particular among others: (i) non-stationarity along time (ii) dependence with respect to the previous spikes in the same spike train (iii) dependence with respect to spikes of other spike trains (i.e. neuronal synchronization).

The temporal behavior of the firing rate has also been extensively considered and many studies have in particular established various kinds of correlations between some motor, sensory or cognitive events in a behaving animal and the occurrences of a variation of the firing rate of specific neurons, before, during or after this event [6, 7]. In these cases, the process is clearly non-stationary. One of the easiest way to take into account non-stationarity is to allow the Poisson process for being inhomogeneous. 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 [4]).

To detect dependence between spikes constitutes another main focus of neuroscientists. Several studies have established statistical evidence of dependence between the occurrences of the spikes of several neurons (see [2, 3, 8–10]). Irrealistic but still simple models based on Poisson processes have been used such as the injection model (see [11]). But from a point process point of view, using *univariate or multivariate Hawkes processes* is the easiest and natural way to model dependence of spikes occurrences. Hawkes processes generalize homogeneous Poisson processes by using functions quantifying interactions between spikes. These functions are called *interaction functions*. These models have recently appeared in the Neurosciences literature [12–14], but they have also been mentioned earlier in the mathematical community interested by issues of Neurosciences (see for instance [15]).

In the setting of each of these three models, many statistical questions need to be answered. First of all, functions have to be estimated (the intensity for the Poisson models, the interaction functions for the Hawkes model). A very standard statistical procedure consists in assuming that these functions are parametrized by a few number of parameters, and in taking (for instance) the maximum likelihood estimator [16, 17]. 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 Neurosciences, except in the particular case of homogeneous Poisson process, there is no a priori shape for the functions to be estimated. These functions are by essence completely unknown. *Non-parametric* statistics is designed to estimate functions when no parametric model can be assumed. Note that approximating the true underlying function by a model with a large number of parameters to be estimated, is still considered as a non-parametric approach.

In Neurosciences, the most well-known non-parametric estimation approach is based on kernel rules. One goal of this article is to discuss them and in particular the classical method based on the average on a sliding window over time to estimate the firing rate which is extensively used. Indeed, the size of these windows is often fixed without clear argument for this choice. From a mathematical point of view, this amounts to estimating the function by the most basic kernel, namely the uniform kernel, with a fixed bandwidth. It is well known that the uniform kernel is the roughest choice and that much smoother kernels lead to better results. Smoother kernels, such as the Gaussian kernel, are also mentioned in the Neurosciences literature, but most of the time a fixed bandwidth is used [4]. However it is well known in the non-parametric statistical community that the choice of the bandwidth is decisive in order to perform good estimation. Indeed, as discussed in Section 4, when we look at the mean behavior of a kernel estimate, it seems that the best choice is the smallest bandwidth, but on the other hand, the kernel estimate is a random variable that randomly fluctuates around its mean and these fluctuations increase when the bandwidth tends to 0. This is the main focus of non-parametric statistics to take care of both aspects. The first classical answer in non-parametric statistics consists in choosing the bandwidth according to the smoothness of the function to be estimated (see [18]). But of course, this answer is not satisfying because the regularity of the function is typically unknown. Several works tried to find a data-driven choice of the bandwidth. For instance, inspired among others by the work of [19] in the 80's, cross-validation methods have already recently diffused in the Neurosciences community [20]. However these methods are not mathematically proved to furnish good bandwidth choices even if they seem to work well in practice. That is the role of *adaptive* statistics to find non-parametric data-driven methods that are able to do as well as the optimal choice when the regularity

is known and this without knowing this regularity. This is one of the main aims of the present article to adapt recent adaptive tools of the literature to spike trains analysis and, consequently, to furnish data-driven methods that are mathematically proved to work. In particular, the existence of some mathematical guarantee for these estimates makes the interpretation of the resulting estimate much more reliable. We propose for instance here to adapt the Goldenshluger and Lepski's method [21] for the inhomogeneous Poisson processes.

But kernels are not the only answer (see also [22] for an alternative in Neurosciences based on penalized splines). It is also not clear that such smooth estimates are always the best answer. Indeed smooth estimation cannot reproduce discontinuities, even if a rapid change in the slope of the reconstruction may give a visual identification of these discontinuities. In Neurosciences, many studies rely on visual detections of these slope changes on the average on sliding window. Hence, estimators that are able to accurately reproduce these discontinuities, such as adaptive histograms or adaptive piecewise polynomials, may provide a better answer to this problem.

Hawkes models are much more intricate to deal with than Poisson processes. Indeed, estimation for Hawkes models is a difficult task and there are no obvious estimates, unlike in the Poisson case. If parametric estimation has been available for a long time (see [23] for instance), it is only very recently that statisticians have been able to perform non-parametric adaptive estimation of the interaction functions (see [24]). We propose here to use the most up-to-date results of adaptive estimation in Hawkes models [25] for the spike trains analysis.

For all the considered models, we are consequently able to give reliable estimates of the unknown functions. However this does not mean that these models are correct. To verify that a model is adequate with respect to the data, one needs to perform a goodness-of-fit test [16]. This kind of test consists in finding a statistics whose distribution is known at least asymptotically under H_0: "the considered model is true" and to reject this hypothesis if the observed statistics seems too far from this reference behavior. It is much easier to test small models than large models. Indeed if the model just consists in one distribution, then the statistics and its distribution can usually be guessed easily. In particular, the Kolmogorov-Smirnov test is one of the oldest tests that is able to distinguish perfectly between a singleton model and any other distribution as soon as there are sufficiently many observations. However, if the model is not a singleton and even if the model is just parametric, it becomes much more difficult to find a correct test statistics with known distribution under H_0 and one usually needs to find "tricks" in order to obtain a correct test statistics. The most classical trick consists in doing as if the true unknown parameter is the one that has been estimated on the same sample,

but we will show in the sequel that such a plug-in step is wrong. In the spike trains analysis, the models that need to be tested are much more intricate since they are non-parametric and following [23] (see also in Neurosciences [26, 27]), the plug-in step has been used intensively, without mathematical justification. We show in this article how to correct this plug-in step in order to obtain non-parametric goodness-of-fit tests of Poisson or Hawkes models, that are mathematically proved to be of prescribed asymptotical level. They mainly rely on the fact that we can build non-parametric estimators in such models with good convergence properties.

The article is organized as follows. We first present the data that will illustrate our various methods, next we explain how to test that spike trains are homogeneous Poisson processes, with very classical tests. In particular, one of the presented test is an advertisement for understanding more general non-parametric goodness-of-fit tests. After showing that the considered data cannot be homogeneous Poisson processes, we propose several methods of adaptive estimation of the intensity for an inhomogeneous Poisson process and we explain how to test that an inhomogeneous Poisson process is an adequate model for the data. Finally, we introduce the Hawkes models and we also show how to estimate and test for such models.

Most of the analysis has been performed with the software `R`. We refer to [22] for a complete list of its advantages.

Notations: For Poisson processes, all neurons are considered independently. We denote by N one spike train and by $N^{(i)}$ the spike train of the ith trial for $i = 1, \dots, n$. For Hawkes processes, we denote by N_j the spike train of neuron j and by $N_j^{(i)}$ the spike train of neuron j in the ith trial for $i = 1, \dots, n$. We always assume that the trials are independent and identically distributed (i.i.d.). In the sequel, we denote by $\mathbb{P}$ the probability measure associated with the distribution of the data and by $\mathbb{E}$ the associated expectation. The point measure dN associated with a point process N is the sum of the Dirac masses in each point (spike) of N, namely for any measurable function f,

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

For any set A, $N(A)$ is the number of points (spikes) of N in A. The indicator function of A is denoted $\mathbf{1}_A$; the cardinal of A is denoted $\#A$. If A is an interval, we denote by $\ell(A)$, the Lebesgue measure (or the length) of A. We also write $u_n << v_n$ if u_n/v_n tends to 0 when n tends to infinity. Finally we denote by $\mathbf{a}'$ the transpose of any vector $\mathbf{a}$.

2 Description of the data

Our procedures are performed on real and simulated data sets. Let us first describe the real data sets and the protocol associated with these data.

2.1 Behavioral procedure

The data used here are a small subset of already partially published data in previous experimental studies [9, 10, 28, 29]. 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.

2.2 Recording technique

Signals recorded from up to seven microelectrodes (quartz insulated platinum–tungsten electrodes, impedance: 2–5MO at 1000 Hz) were amplified and band-pass filtered from 300Hz to 10 kHz. Using a window discriminator, spikes from only one single neuron per electrode were then isolated. 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.

2.3 Summary of the real data set

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, the first data set (respectively the second data set) corresponds to the pair of neurons (N1,N2) (respectively to the pair of neurons (N3,N4)). In the first data set (respectively

the second data set), 177 trials (respectively 165 trials) are considered. Figure 1 plots the rasters associated with these data. However, because 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. Therefore, distinguishing the data with respect to the direction and the trial, we obtain the following summary:

Direction	1	2	3	4	5	6	Total
First data set	31	35	28	30	28	25	177
Second data set	27	26	30	30	32	20	165

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 but the small number of trials per direction makes sometimes this distinction inappropriate and we will need to pool on the directions to obtain about 200 trials.

2.4 Simulated data

We study the performance of our procedures on the real data sets described previously but also on simulated data sets. Simulated data are carried out according to the following simulations schemes:

- (S-HomPoi) Spikes are distributed according to an homogeneous Poisson processes of intensity 20 on $[0, 2]$.
- (S-InPoi) Spikes are distributed according to an inhomogeneous Poisson processes with piecewise continuous intensity given by Figure 6.
- (S-Haw) Two spike trains are simulated according to a bivariate Hawkes process with parameters given by Figure 14 observed on $[0, 2]$. Each process is respectively denoted N_1 and N_2.

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

Mathematical descriptions of these processes are described in the subsequent sections.

3 Homogeneous Poisson Processes

Let λ be a fixed positive constant. Assuming that a spike train is an homogeneous Poisson process with intensity (or constant firing rate) λ means that (i) for two disjoint intervals I and J, $N(I)$ and $N(J)$ are independent and (ii) for any interval I, $N(I)$ is a Poisson variable with parameter $\lambda\ell(I)$. Therefore, it is

quite easy to estimate λ on one trial by $\hat{\lambda} = N([0, T_{max}])/T_{max}$, which leads to an unbiased estimate of λ since $\mathbb{E}(\hat{\lambda}) = \lambda$. Moreover, when n i.i.d. trials are available, the estimator

$$\hat{\lambda}_n = \frac{1}{nT_{max}} \sum_{i=1}^{n} N^{(i)}([0, T_{max}])$$

is unbiased but also consistent, i.e. the estimate converges to the true parameter, namely $\hat{\lambda}_n \xrightarrow[n\to\infty]{\mathbb{P}} \lambda$.

Now the question is the following: can we test that a spike train N is an homogeneous Poisson process? Many tests could answer this question but our purpose is not to review all of them. In particular, we will not consider parametric tests or tests based on asymptotic normality properties of the Poisson process whose quantiles are not always easily available (see [17]). Classical tests of reliability based on monotone alternatives will not be considered as well since we cannot assume that firing rates are monotone (see for instance [30]). So, three directions remain, each of them corresponding to one property of the homogeneous Poisson process (see [1, 31] for instance).

[1] Given the event $\{N([0, T_{max}]) = n_{tot}\}$, the spikes are n_{tot} uniform i.i.d. variables on $[0, T_{max}]$.

[2] The total number of spikes $N([0, T_{max}])$ is a Poisson variable.

[3] The delays between spikes are i.i.d. exponential variables.

[3'] For any fixed time $t \leq T_{max}$, the delay between T_t, the next spike of N after t, and t, is an exponential variable, since exponential variables are memoryless.

Note that Property **[3']** is a variant of Property **[3]**. We now derive the tests associated with respect to these three approaches.

3.1 Test of uniformity

Using Property **[1]**, the first natural idea consists in rejecting the hypothesis H_0 "the process is an homogeneous Poisson process" if the repartition of the points is not uniform enough. On the event $\{N([0, T_{max}]) = n_{tot}\}$, we consider the very classical Kolmogorov-Smirnov test (KS test in the sequel) which compares the renormalized empirical cumulative distribution function to the (theoretical) cumulative distribution function (c.d.f.) of the uniform distribution on $[0, T_{max}]$ (see [16]). More precisely, let us consider, for all $u \in [0, 1]$

$$F_{n_{tot}}(u) = \frac{1}{n_{tot}} \sum_{T \in N} \mathbf{1}_{\{\frac{T}{T_{max}} \leq u\}} \quad \text{and} \quad F(u) = u.$$

If n_{tot} is large enough, $F_{n_{tot}}(u)$ should be close to $F(u)$ for any u. The KS test is based on the statistics

$$KS_{n_{tot}} = \sup_{u\in[0,1]} |F_{n_{tot}}(u) - F(u)|$$

whose distribution is known when the spikes are uniformly distributed, which holds by Property [**1**] conditionally to the event $\{N([0,T_{max}]) = n_{tot}\}$. Let $k_{n_{tot},1-\alpha}$ be the $1-\alpha$ quantile of this distribution. The KS test consequently consists in rejecting H_0: "the process is an homogeneous Poisson process" whenever $KS_{N([0,T_{max}])} > k_{N([0,T_{max}]),1-\alpha}$ and this test is of exact level α.

We mention that when p tends to ∞, the random variable $\sqrt{p}KS_p$ tends in law to a tabulated distribution $\mathcal{K}$ (see [32]). If $\tilde{k}_{1-\alpha}$ is the $1-\alpha$ quantile of $\mathcal{K}$, $\sqrt{p}k_{p,1-\alpha}$ tends to $\tilde{k}_{1-\alpha}$ and the approximation is valid as soon as $p > 45$ [33].

Table 1 shows the behavior of the KS test on simulated and real data sets, for the level $\alpha = 5\%$. On 200 simulated homogeneous Poisson processes, 10 trials are rejected, as expected for a test of level 5%. On the real data sets, the proportion is usually larger (except for N2) but is still less than 25%, i.e. more than the three quarters of the trials accept H_0. This does not mean than three quarters of the tests are indeed homogeneous Poisson processes. To understand this behavior, Figure 2 shows the distribution of KS_p under H_0 and under a fixed known alternative hypothesis with $p = 50$. At the level 5%, the power of KS test is only 20%. It illustrates the well-known statistical principle: a test may say "no" or "maybe" to H_0 but not "yes".

For the KS test whose asymptotic power is equal to 1, the problem can be overcome if we can obtain more data. We propose to aggregate all trials to get a larger point process (see also Figure 12). Namely, we consider:

$$N^a = \cup_{i=1,\ldots,n} N^{(i)}.$$

Using the properties of Poisson processes [1,31], N^a, the aggregated process, is also an homogeneous Poisson process with intensity $n\lambda$. Hence, one can again perform the KS test and study its performance by using its p-values. Remember that the p-value of a test is the critical value of the level α such that the test passes from acceptance to rejection (see [34]). For data not in accordance with H_0, the p-values of powerful tests are small. It is the case for the KS test performed on aggregated real data sets (see Table 3). We can conclude that all these spike trains are not homogeneous Poisson processes. In particular, the ambiguity with respect to N2 is now avoided. It is due to the number of points for N^a which is much larger than for individual trials. From a multiple test point of view since all the p-values are less than 5%, all the tests will be rejected by a Benjamini and Hochberg (BH) procedure for a prescribed false discovery rate (FDR) of 5% [35] (see

the Additional File 4 for a rapid introduction to multiple testing theory, FDR and BH procedure). Note that in the case of *ex aequo* observations (which is due to the recordings protocol and is a bit troublesome from the theoretical point of view), the `R`-command `ks.test` warns the user. We could slightly modify our procedure to avoid *ex aequo* observations, but this modification would have no impact on the p-values, so we can ignore this phenomenon.

There exist other tests that are more powerful than the simple and classical KS test. For instance, [36,37] study more local approaches and provide adaptive tests of uniformity for i.i.d. variables or adaptive tests of homogeneity for Poisson processes. However, the previous KS tests lead to non-ambiguous conclusions: real data sets are not homogeneous Poisson processes.

3.2 Test of Poissonian variables

Following Property **[2]**, one alternative is to test that the number of points per trial obeys a Poisson distribution by using chi-square tests with parameter estimation [33, 38]. Under H_0, M_j, the number of trials with j spikes should be close to $\hat{m}_j = n\frac{(\hat{\lambda}_n * T_{max})^j}{j!}e^{-\hat{\lambda}_n T_{max}}$. So, we introduce the chi-square pseudo-distance between $(M_j)_j$ and $(\hat{m}_j)_j$:

$$Z_\chi = \sum_j \frac{(M_j - \hat{m}_j)^2}{\hat{m}_j}.$$

Chi-square tests are based on the fact that for classical models with finite number of states J_0, under H_0, Z_χ tends in law to $\chi^2(J_0 - 1 - d)$ where d is the number of estimated parameters. The test rejects H_0 when $Z_\chi > c_{1-\alpha}$ where $c_{1-\alpha}$ is the $1-\alpha$ quantile of $\chi^2(J_0 - 1 - d)$. In practice, and for finite n, one can perform the test only if all the $\hat{m}_j$'s are larger than 5, which ensures that the chi-square approximation based on these asymptotic arguments is valid.

Here, for Poisson variables, with infinite number of states, we gather different j's together in disjoint contiguous groups in order to obtain a finite number of classes. Moreover this gathering is performed to ensure that $\sum_{j\in G} \hat{m}_j \geq 5$ for all the groups G. Hence our statistics is now

$$\tilde{Z}_\chi = \sum_{G\in\mathcal{G}} \frac{(\sum_{j\in G} M_j - \sum_{j\in G} \hat{m}_j)^2}{\sum_{j\in G} \hat{m}_j},$$

and we reject H_0: "the number of spikes in each trial is a Poisson variable", whenever $\tilde{Z}_\chi > c_{1-\alpha}$ where $c_{1-\alpha}$ is the $1-\alpha$ quantile of $\chi^2(\#\mathcal{G} - 2)$, since one parameter (namely λ) has been estimated.

The results are presented in Table 4 for the simulated data and in Table 5 for the real data sets. The p-values are large for simulated homogeneous Poisson processes as expected. The p-values are also large on

the real data sets, direction per direction and a BH multiple test procedure cannot find any significant small p-values. But as explained previously, there is no guarantee that the counts are indeed Poisson variables (tests may be not powerful enough). When we pool all the directions, the p-values are very small for N1, N3 and N4, showing that they are indeed not Poisson variables. This can be explained by one of the following reasons:

- The counts are not i.i.d. and pooling the directions leads to a non i.i.d. sample.
- The counts are i.i.d. but are not Poisson variables, as detected now by the test. The latter is based on a large number of observations thanks to the pooling, so it is more powerful.

3.3 Test of Exponentiality

More generally, Kolmogorov-Smirnov tests can be applied to test H_0: "The variables have cumulative distribution function F" and therefore following Property **[3]**, if one takes $F(u) = 1 - e^{-\lambda u}$ for $u \geq 0$, it is possible to test H_0 : " The variables are exponential with parameter λ".

Several specific procedures exist to test H_0: " The variables are exponential" without specifying the parameters. These procedures are essentialy based on the fact that the exponential distributions are scale invariant (see for instance [36]). We want here to investigate more precisely the plug-in procedure, since it is at the root of the main practical tests in point processes theory [23, 26, 27]. This procedure is also implemented by default in various softwares such as `SAS`.

However, as one can see on Figure 3, it is not possible to plug directly an estimate of λ, based on exactly the same sample. It is also not possible to split the data in two equal parts and to use one half of the data for the estimation and the other half for the test. Figure 3 shows that the first plug-in procedure leads to a very conservative test whereas the second one leads to a very poor test that wrongly rejects H_0 with high probability. We propose here a quite simple approach based on sub-sampling to obtain a test of asymptotic level α, which provides a correct repartition of the p-values (see Figure 3). This method could be applied to test that the delay between spikes is exponential, following Property **[3]**. Since the previous tests of uniformity are already quite efficient, we have not implemented this new method. We have preferred to implement this method to test that the process is locally an homogeneous Poisson process, following Property **[3']**.

The test is constructed as follows.

1. Fix a time t, where one wants to check that the process is locally an homogeneous Poisson process.

2. For each trial i, find $T_t^{(i)}$ the next spike after time t and compute $\tau_i = T_t^{(i)} - t$.

3. Compute $\hat{\lambda} = n/(\tau_1 + \cdots + \tau_n)$.

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

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

$$\forall x > 0, \quad F_S(x) = \frac{1}{m(n)} \sum_{i \in S} \mathbf{1}_{\tau_i \leq x}.$$

6. Take $\tilde{k}_{1-\alpha}$ the $1-\alpha$ quantile of the asymptotic distribution $\mathcal{K}$ of $\sqrt{p} K S_p$.

7. Reject H_0: "The distribution of the τ_i's is exponential" whenever

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

where for any $x \geq 0$,

$$\hat{F}(x) = 1 - e^{-\hat{\lambda} x}.$$

Theorem 1. *The previous test is asymptotically of level α when n tends to ∞.*

A proof can be found in Additional File 2. This surprising result is due to the fact that two errors come into play, the Kolmogorov-Smirnov distance between F_S and F and the error between F and $\hat{F}$. The first one is of the order of $\tilde{k}_{1-\alpha}/\sqrt{m(n)}$, whereas the second one is of the order of $1/\sqrt{n}$. In order to make this second error negligible, it is sufficient to take $m(n) << n$.

The practical results are displayed in Table 6. The p-values for simulated homogeneous Poisson processes are obviously large. On the real data sets, we have performed our procedure for several values of t and we also found large p-values. Hence, there is no evidence that the real data sets are not locally homogeneous Poisson processes.

Let us now consider the inhomogeneous Poisson model, which can model time-dependent firing rates.

4 Inhomogeneous Poisson Processes

Assuming that a spike train is an inhomogeneous Poisson process N with intensity (or time-dependent firing rate) $\lambda(.)$ implies that (i) for two disjoint intervals I and J, $N(I)$ and $N(J)$ are independent and (ii) for any interval I, $N(I)$ is a Poisson variable with parameter $\int_I \lambda(x)dx$ (see [1,31] for a precise definition). So the process has the same properties as before except that now the law of $N(I)$ is not translation-invariant anymore,

and in this case the process is not stationary. When the intensity $\lambda(.)$ which models the time-dependent firing rate is constant, the model is equivalent to the homogeneous Poisson process. Inhomogeneous Poisson processes with intensity $\lambda(.)$ satisfy the following properties: for any square integrable function f,

$$\mathbb{E}\left(\int f(x)dN(x)\right) = \int f(x)\lambda(x)dx \qquad \text{and} \qquad \text{Var}\left(\int f(x)dN(x)\right) = \int f^2(x)\lambda(x)dx. \tag{1}$$

A classical idea to estimate $\lambda(.)$, is to assume that $\lambda(.)$ is locally constant. Several methods can then be used (see [39]) and in particular the method based on $\hat{\lambda}_n$ defined in the previous section. More precisely, for any interval I with small length, an estimate of $\lambda(.)$ on I is the constant

$$\frac{1}{n\ell(I)}\sum_{i=1}^{n} N^{(i)}(I).$$

If we want to reconstruct the whole function $\lambda(.)$, there are therefore two different strategies:

- We consider a fixed partition $\mathcal{P}$ of $[0, T_{max}]$ and we define for any x,

$$\hat{\lambda}_n^H(x) = \sum_{I\in\mathcal{P}} \left(\frac{1}{n\ell(I)}\sum_{i=1}^{n} N^{(i)}(I)\right)\mathbf{1}_I(x), \tag{2}$$

  which is an histogram, namely a piecewise constant function on the partition $\mathcal{P}$. In particular, peri stimulus time histograms (PSTH) are built in this way for a very thin, regular and fixed partition $\mathcal{P}$.

- We take $I = [x-h, x+h]$ and we consider the average on the sliding window $[x-h, x+h]$ classically used on spike trains data:

$$\hat{\lambda}_n^{SW_h}(x) = \frac{1}{2nh}\sum_{i=1}^{n} N^{(i)}([x-h, x+h]). \tag{3}$$

  On the data considered here, $h = 0.05$ s is usually taken (see [28] for instance). However one can generalize this sliding window method for other kernels K (that satisfy $\int K = 1$) and with other bandwidths h, by taking

$$\hat{\lambda}_n^{K_h}(x) = \frac{1}{n}\sum_{i=1}^{n}\int K_h(x-u)dN^{(i)}(u) = \frac{1}{n}\int K_h(x-u)dN^a(u), \tag{4}$$

  where $K_h(u) = (1/h)K(u/h)$ and where N^a is the aggregated process (see also Figure 12), which is an inhomogeneous Poisson process with intensity $n\lambda(.)$ [1, 31]. This estimate is called a kernel estimator [20]. It is sometimes referred as a linear filter [4]. The average on a sliding window (3) corresponds to the choice $K(u) = \mathbf{1}_{[-1,1]}/2$.

Histograms and kernel estimates have been used for a long time to estimate density. They both have advantages and disadvantages. Histograms provide a segmentation of the data and on each segment, the

process will be considered as homogeneous. But of course, this makes sense only if the choice of the partition $\mathcal{P}$ is data-driven. The disadvantage is of course that the global reconstruction of $\lambda(.)$ is always piecewise constant and therefore not smooth at all. Kernel estimates with regular kernels (such as the Gaussian kernel) can provide a very smooth reconstruction of the intensity and the larger the bandwidth, the smoother the estimate. Without a data-driven choice of the bandwidth, it is not clear at all that the true intensity is that smooth. The problem of bandwidth/partition selection dates back to the early eighties with [19]. So whatever the estimate, we need to perform an adaptive choice of the partition or the bandwidth. We start with the kernel estimates, which are the most classical estimates, at least in Neurosciences, next we will deal with histograms, finally we will describe more informally a procedure based on adaptive piecewise polynomials which is able to provide piecewise constant estimates if the true intensity is irregular but also smoother estimate if the underlying intensity is regular enough.

4.1 Data-driven choice of the bandwidth for kernel estimators

By the classical properties of the inhomogeneous Poisson process, the estimator $\hat{\lambda}_n^{K_h}$ is biased by (1) since

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

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 by (1)(see further details in Additional File 1):

$$\mathbb{E}\,\|\hat{\lambda}_n^{K_h} - \lambda\|_2^2 = \|K_h \star \lambda - \lambda\|_2^2 + \frac{\|\lambda\|_1}{nh}\|K\|_2^2, \tag{5}$$

which is classically interpreted as a *bias-variance* decomposition. Under mild standard regularity properties, the first term in (5) (the bias of $\hat{\lambda}_n^{K_h}$) goes to 0 when $h \to 0$ but the second term (the variance) blows up. We naturally wish to choose a bandwidth h such that the $\mathbb{L}_2$-risk of $\hat{\lambda}_n^{K_h}$ is as small as possible. The parameter $\bar{h}$ which minimizes the right hand side of (5) with respect to h is called the *oracle bandwidth* but it depends on $\lambda(.)$, so cannot be used in practice. Therefore, the (data-driven) choice of the bandwidth parameter remains a difficult task in particular when the number of observed points is not very large. To deal with this problem, several methodologies can be used:

1. The first one consists in plotting several curves; then the practitioner chooses the estimate that is most in accordance with one's prior belief about the intensity. This belief can be based, for instance, on the smoothness, the monotony or the number of modes of the intensity. This methodology is often performed in Neurosciences.

2. If we wish to deal with an automatic procedure, we can adapt to our setting the simple *rule of the thumb*, which is the most popular methodology used for practical kernel density estimation. Besides automaticity, which is useful when we have to deal with a large number of data sets, this method is very simple to perform. We refer the reader to [40] for some partial theoretical justifications. The idea consists in using approximations of the bias term and in this paragraph, we further assume that

$$\int tK(t)dt = 0, \quad k_2 := \int t^2 K(t)dt < \infty.$$

So, if $\lambda(.)$ is smooth enough, a Taylor expansion gives when $h \to 0$:

$$\begin{aligned} \mathbb{E}[\hat{\lambda}_n^{K_h}(x)] - \lambda(x) &= \int K(t)(\lambda(x-ht) - \lambda(x))dt \\ &= \frac{1}{2}h^2\lambda''(x)k_2 + o(h^2). \end{aligned}$$

Then,

$$\|K_h \star \lambda - \lambda\|_2^2 \approx \frac{1}{4}h^4 k_2^2 \int (\lambda''(x))^2 dx.$$

Setting $g := \|\lambda\|_1^{-1} \times \lambda$, the decomposition (5) gives:

$$\mathbb{E}\,\|\hat{\lambda}_n^{K_h} - \lambda\|_2^2 \approx \frac{1}{4}h^4 k_2^2 \|\lambda\|_1^2 \|g''\|_2^2 + \frac{\|\lambda\|_1}{nh}\|K\|_2^2,$$

and the right hand side is minimal when $h = h^*$, where

$$h^* := \|K\|_2^{2/5} k_2^{-2/5} \|g''\|_2^{-2/5} \|\lambda\|_1^{-1/5} n^{-1/5} \approx \|K\|_2^{2/5} k_2^{-2/5} \|g''\|_2^{-2/5} N^a([0,T_{max}])^{-1/5}. \tag{6}$$

A difficult issue remains: the estimation of $\|g''\|_2$. To overcome this problem, we assume that, conditionally to $\{N^a([0,T_{max}]) = m\}$, then the points of N^a obey the distribution of a m-sample with Gaussian density with mean μ and variance σ^2. In this case

$$\|g''\|_2^2 = \frac{3}{8}\pi^{-1/2}\sigma^{-5} \approx \frac{3}{8}\pi^{-1/2}\hat{\sigma}^{-5},$$

where $\hat{\sigma}$ is the standard estimate of the standard variation of the m observations conditionally to $\{N^a([0,T_{max}]) = m\}$. It remains to substitute this expression into (6). In particular, if K is the Gaussian kernel,

$$h^* \approx \left(\frac{4}{3}\right)^{1/5} \hat{\sigma} N^a([0,T_{max}])^{-1/5} \approx 1.06\, \hat{\sigma} N^a([0,T_{max}])^{-1/5}.$$

This last expression provides another still naive but simple choice for the bandwidth parameter. The rule of the thumb is simple but is essentially based on heuristics arguments, so it suffers from a lack of theoretical justifications. Furthermore, due to Gaussian assumptions used for estimating $\|g''\|_2$, it tends to oversmooth.

3. Cross validation is another popular methodology used in Neurosciences (see in particular [20] and the references therein). It is based on the use of an unbiased and consistent estimate of the risk of the kernel estimate $\mathbb{E}\,\|\hat{\lambda}_n^{K_h} - \lambda\|_2^2$ for a fixed bandwidth h denoted $\hat{R}_h$. The selected bandwidth is a minimizer of $h \mapsto \hat{R}_h$ among all bandwidths. Even if it is known to work very well in practice, this method suffers from two problems: its high computational time (we need to compute the risk estimate for each bandwidth) and its lack of theoretical validity, as for the rule of the thumb. Indeed, it can be proved that the method selects the best bandwidth in a fixed family of bandwidths when n tends to infinity. But this family cannot depend on n and therefore the chosen bandwidth cannot tends to zero, making the bias of the resulting estimate incompressible (see also [19]).

4. We now describe the theoretical Goldenshluger and Lepski's methodology [21], whose computational cost is slightly higher than the cross-validation method but has no theoretical gap. It also consists in estimating the risk via a direct bias estimation rather than approximating it as the thumb rule. It can be described in our setting as follows: we consider a set of bandwidths $\mathcal{H}$ and for any $h, h' \in \mathcal{H}$ we define:

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

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

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

The Additional File 1 shows that $A(h)$ constitutes a good estimate of the bias term. Finally, we select the data-driven bandwidth as follows:

$$\hat{h} := \arg\min_{h\in\mathcal{H}}\left\{A(h) + \frac{(1+\eta)(1+\|K\|_1)\|K\|_2\sqrt{N^a([0,T_{max}])}}{n\sqrt{h}}\right\}, \tag{7}$$

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

$$\hat{\lambda}_n^{GL} := \hat{\lambda}_n^{K_{\hat{h}}}. \tag{8}$$

Note that in (7), $\|K\|_2^2 N^a([0,T_{max}])/(n^2h)$ is an unbiased estimate of the variance term in (5) 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 (5) shows that our procedure mimics the performance of the oracle (restricted to $\mathcal{H}$) 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*. Take the family $\mathcal{H} = \{1, ..., \lfloor \delta n \rfloor^{-1}\}$. Thus the family of bandwidths grows with n and it is possible to select a bandwidth tending to 0 with n. If we assume that $\lambda(.)$ belongs to $\mathcal{S}_\alpha$, the standard Sobolev space of regularity α, then under standard properties of vanishing moments for the kernel K, $\|K_h \star \lambda - \lambda\|_2 = O(h^\alpha)$, so the $\mathbb{L}_2$-risk of $\hat{\lambda}_n^{GL}$ goes to 0 at the rate $n^{-2\alpha/(1+2\alpha)}$, which is the optimal minimax rate of convergence over such spaces. This rate is guaranteed, even if we do not know in advance that $\lambda(.)$ belongs to $\mathcal{S}_\alpha$. This is the *adaptive minimax property* [18].

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 straightforward computations show 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. The choice $\eta = 0.5$ with the bandwidths family

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

used in the sequel, shows a robust behavior on various simulations, for the present considered size of n (mainly $n \approx 40$).

4.2 Data-driven choice of a partition and histograms

There are several ways to select data-driven partitions for inhomogeneous Poisson processes. 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 Goldenshluger and Lepski's kernel rule. 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 a minimax

rate of convergence for regular intensities (namely intensities belonging to $\mathcal{S}_\alpha$ with $\alpha > 1$). Therefore, the choice of regular partitions is probably not the best one and one may prefer the Goldenshluger and Lepski's method. The data-driven choice of the partition becomes much more interesting when the partition is not forced to be regular. However, the method of [41] is too time consuming to be really considered in practice. Another possible direction that we will ignore is the context of Markov modulated Poisson processes [42], where the algorithms are also quite time consuming without ensuring any adaptive property (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. 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 (see (1)) and consistently estimated by

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

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

$$\begin{aligned} \mathbb{E}[\|\hat{\lambda}_n^m - \lambda\|^2] &= \sum_{(j,k) \notin m} \beta_{j,k}^2 + \sum_{(j,k) \in m} \mathrm{Var}(\hat{\beta}_{j,k}) \\ &= \sum_{j > j_0} \sum_k \beta_{j,k}^2 + \sum_{j \leq j_0} \sum_k [\beta_{j,k}^2 \mathbf{1}_{(j,k) \notin m} + v_{j,k} \mathbf{1}_{(j,k) \in m}], \end{aligned} \tag{9}$$

where (see (1))

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

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:

$$\eta_{j,k} = \sqrt{2\gamma \ln(n)\hat{v}_{j,k}} + \frac{\gamma \ln(n) 2^{j/2}}{3n} \quad \text{where} \quad \hat{v}_{j,k} = \frac{1}{n^2}\int \psi_{j,k}^2(x) dN^a(x). \tag{10}$$

Then, we obtain the following thresholding estimator:

$$\hat{\lambda}_n^{Th} = \sum_{j=-1}^{j_0} \sum_k \hat{\beta}_{j,k} \mathbf{1}_{\{|\hat{\beta}_{j,k}| > \eta_{j,k}\}} \psi_{j,k}. \tag{11}$$

In [43], it has been proved that a slight modification of this estimate satisfies an oracle inequality in the same spirit as Theorem 2. We can generalize 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 Goldeshluger and Lepski's estimate, up to some logarithmic term, as soon $\gamma > 1$. The choice $\gamma < 1$ has been shown to lead to bad rates of convergence and the choice $\gamma = 1$ has been shown to work well on extensive simulations. This method is easily implementable leading to very fast algorithms that are in particular faster than Goldenshluger and Lepski's algorithms. This method has been performed with the Haar basis, $j_0 = 15$ and $\gamma = 1$.

4.3 More sophisticated procedures

Note that thresholding rules overcome a drawback of kernel estimates that always suffer from a lack of spatial adaptivity on the time axis. Indeed, once the bandwidth is chosen, it will be the same for all the data. If a particular spot needs more precision, whereas other spots need a larger bandwidth, then the kernel estimator cannot adapt to this heterogeneity. Several attempts have been proposed to build more local choices of the bandwidth (see [20] for instance), but up to our knowledge no mathematical proof of this spatial adaptation has been established. 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 or we could extend the Goldenshluger and Lepski's method to select the bandwidth locally and the kernel by using data, but these extensions do 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. 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 $\hat{\lambda}_n^{WN}$.

4.4 Performance of those estimators

In this section we compare $\hat{\lambda}_n^{GL}$, $\hat{\lambda}_n^{Th}$ and $\hat{\lambda}_n^{WN}$ with kernel rules with fixed bandwidth and with the rule of thumb. Namely we consider the original sliding window method, denoted $\hat{\lambda}_n^{SW_h}$, implemented with $h = 0.05$ (classically used on these data sets). The same choice for the window has been used with the smoother Gaussian kernel and the resulting estimate is denoted $\hat{\lambda}_n^{K_h}$. The rule of the thumb has also been used with the Gaussian kernel and the resulting estimate is denoted $\hat{\lambda}_n^{K_{h^*}}$. Several reconstructions are given in Figures 4, 5, 6 and 7.

First of all, all the figures show that $\hat{\lambda}_n^{SW_h}$ is clearly the worst choice, as expected for such a rough kernel. Figure 4 shows the reconstruction of a constant intensity and the adaptive Haar thresholding rule $\hat{\lambda}_n^{Th}$ is the best estimate. Kernel estimates with Gaussian kernels are oscillating, the bandwidth h^* is larger than $\hat{h}$, the Goldenshluger and Lepski's bandwidth. The fixed bandwidth $h = 0.05$ is the smallest one and is quite inadequate in this setting. Willett and Nowak's method is able to reconstruct perfectly the flat line. For a Gaussian intensity, Figure 5 shows that Gaussian kernel estimates are much better than the adaptive histogram. In this case, Willett and Nowak's methodology is able to capture the smoothness of the curve and gives the best reconstruction. A more irregular intensity is used in Figures 6 and 7. It is piecewise continuous, with large jumps and smooth bumps. For such an irregular intensity and for a small number of trials (Figure 6), the thresholding estimate and Willett and Nowak's 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 Goldenshluger and Lepski's 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 7), 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. Willett and Nowak's method is reconstructing more accurately the jumps despite some important boundary artefacts. It also gives smoother reconstructions for the bumps. In conclusion, Goldenshluger and Lepski's method gives clearly a 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, Willett and Nowak's 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]). Those studies also proved that Willett and Nowak's method can be outperformed by the thresholding estimate when the intensity becomes null or insignificant for a large fraction of the observation interval, since it is based on a log-likelihood contrast. This situation never happens on the spike trains data.

Figures 8, 9, 10 and 11 give 4 different reconstructions on some real data sets. The shapes of reconstructed intensities by Willett and Nowak's method are quite different: they can be very smooth with slow variations (Figure 8), almost piecewise constant (Figure 11) and even constant (Figure 10). Despite the latter reconstruction, p-values of Table 3 show that the process N2 in direction 3 is probably not homogeneous. Reconstructions for the thresholding estimates look like Willett and Nowak's reconstructions, but they are of course always piecewise constant. Gaussian kernel rules $\hat{\lambda}_n^{K_h}$ and $\hat{\lambda}_n^{GL}$ provide quite similar reconstructions (in particular for Figures 10 and 11) that are less regular than $\hat{\lambda}_n^{K_{h^*}}$. Finally, the average on a sliding window method is still very rough.

4.5 Goodness-of-fit tests

4.5.1 Tests of H_0: "The process is an homogeneous Poisson process"

First, the tests that have been developed for homogeneous Poisson processes can be applied on simulated inhomogeneous Poisson processes (S-InPoi). Table 2 gives the p-values for the test of uniformity and shows very small p-values. Note that even for one trial, most of the tests of uniformity reject H_0 (see Table 1). Hence, as expected, (S-InPoi) cannot be considered as an homogeneous Poisson processes.

Table 4 gives the p-values for the test of Poissonian variables. They are large as expected since for inhomogeneous Poisson processes, the total number of points per trial obeys a Poisson distribution with

parameter $\int_0^{T_{max}} \lambda(u)du$. Hence, finding large p-values for this test on the real data sets in Table 5 and small p-values on the same data for the test of uniformity (see Table 3) makes us suppose that those real data may be modelled by inhomogeneous Poisson processes.

The local test of exponentiality can also be applied (see Table 6). We note that close to a jump of the intensity associated with (S-InPoi), the p-values are small but on an interval where the intensity is very smooth, they are large. On the real data sets, we never observe such small p-values but it does not mean that the real data sets are homogeneous Poisson processes. They can be realizations of an inhomogeneous Poisson processes whose intensity does not change abruptly enough. Note that the jump detected by the Willett and Nowak's algorithm for the intensity of N3 in direction 5 (see Figure 11), just before 0.5, is not confirmed by this test.

4.5.2 Aggregated test of H_0: "The process is an inhomogeneous Poisson process."

Now let us test that the processes are indeed inhomogeneous Poisson processes. One of the main properties of inhomogeneous Poisson processes is that given the event $\{N^a([0, T_{max}]) = n_{tot}\}$, the points of the aggregated process obey a distribution with c.d.f. given by

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

But we do not know $\lambda(.)$ or F under H_0 and we need to estimate it. Following the conclusions of Theorem 1, we use a sub-sample in order to obtain tests of H_0: "The process is a Poisson process" without specifying any model on the underlying intensity $\lambda(.)$. Indeed, if we do not know $\lambda(.)$, we can estimate it (or F directly) by one of the various previous methods and plug the estimate in, as it has been done in Theorem 1. Let us describe more precisely this first procedure, which is said *aggregated*.

1. Estimate the cumulative distribution function with all the trials by $\hat{F}$, which is one of the following choices:

$$\forall t \leq T_{max}, \quad \hat{F} = F_n(t) := \frac{1}{N^a([0, T_{max}])} \sum_{i=1}^{n} N^{(i)}([0, t]) \tag{12}$$

or

$$\forall t \leq T_{max}, \quad \hat{F} = F_{\hat{\lambda}}(t) := \frac{\int_0^t \hat{\lambda}(u)du}{\int_0^{T_{max}} \hat{\lambda}(u)du}, \tag{13}$$

where $\hat{\lambda}$ is either $\hat{\lambda}_n^{GL}$ or $\hat{\lambda}_n^{Th}$ or $\hat{\lambda}_n^{WN}$.

2. Take a sub-sample $\mathcal{S}$ of size $m(n) = n^\delta$, with $\delta < 1$ ($\delta = 1/2$ or $2/3$) and denote $N^{a,\mathcal{S}}$ the aggregated point process of the point processes $N^{(i)}$, for i in $\mathcal{S}$ (see also Figure 12 for a visual representation of

the aggregation).

3. Compute
$$Z^a := \sqrt{N^{a,\mathcal{S}}([0,T_{max}])} \sup_{0\leq t\leq T_{max}} \left| \frac{1}{N^{a,\mathcal{S}}([0,T_{max}])} \sum_{T\in N^{a,\mathcal{S}}} \mathbf{1}_{T\leq t} - \hat{F}(t) \right|.$$

4. Reject if $Z^a > \tilde{k}_{1-\alpha}$ (aggregated test by upper value) or reject if $Z^a < \tilde{k}_\alpha$ (aggregated test by lower value).

We can prove the following result.

Theorem 3. *Both previous aggregated tests are of level asymptotically α in the following cases:*

- *if $\hat{F}$ is given by* (12) *and $\delta < 1$, or*
- *if $\hat{F}$ is given by* (13), $\mathbb{E}(\int_0^{T_{max}} (\hat{\lambda}(u) - \lambda(u))^2 du) = O(n^{-\eta})$ *and $\delta < \eta$.*

The proof is given in the Additional File 2. In particular, since oracle inequalities are satisfied by $\hat{\lambda}_n^{GL}$, $\hat{\lambda}_n^{Th}$ and $\hat{\lambda}_n^{WN}$, we know that if the function is regular enough, the second condition will hold. For instance with $\hat{\lambda}_n^{GL}$, if $\lambda(.)$ is in a Sobolev space of regularity strictly larger than $1/2$ then $\delta = 1/2$ is convenient. If the function $\lambda(.)$ has a finite set of non zero coefficients on the Haar basis then with $\delta < 1$, $\hat{\lambda}_n^{Th}$ will be also convenient. It is therefore quite necessary to provide estimates that are proved to achieve the best rates of convergence.

When we use (12) in the test statistics Z^a, the corresponding test by upper values is really close to a two-sample Kolmogorov-Smirnov test except that the two samples are not independent, the first one being a sub-sample of the second one. In any case, the test by upper values is, roughly speaking, nothing more than a test that can verify that the trials are i.i.d. and in this respect it is not expected to be really powerful in other situations. Indeed, Table 7 shows as expected large p-values for both tests (upper and lower values) on simulated Poisson processes. On the real data sets (see Table 8), the p-values for the test by upper values are also large direction per direction and for the pooled process. But the aggregated test by lower values can detect more relevant features. There are indeed four cases, detected by a BH multiple test, where the p-values for the aggregated test by lower values are too small, meaning that the distance is too small to be just due to the classical randomness of a Poisson process. One can also use this by forcing the sub-sample to correspond to a particular direction for the pooled process and we find once again one p-value for the aggregated test by lower values that is too small. It consequently seems that those 5 situations correspond to a behavior that seems in some sense, less random than a Poisson process.

When we use (13) in the test statistics Z^a, the main advantages is that we also verify that the estimators are consistent and that their convergence is fast. This has been theoretically shown for Poisson processes and therefore a rejection suggests that the estimates $\hat{\lambda}(.)$ are not behaving as expected by a Poisson process. In this sense, they should be more powerful than the test used with (12). As expected on simulated Poisson processes (see Table 7), the p-values are large. On the real data sets (see Table 9), most of the couples neuron/direction do not reject H_0. The remaining rejections are mainly due to $\hat{\lambda}_n^{GL}$.

4.5.3 Cumulated test of H_0: "The process is an inhomogeneous Poisson process."

The main problem with the aggregated procedure is that if there is a dependency in time between the points of the process, this dependency can be reflected in the empirical c.d.f. on one trial but can be completely attenuated by the aggregation step 2. One way to keep this dependency in time, is to cumulate the processes instead of aggregating them (see Figure 12) to keep the relative distance between the points. Then the main idea consists in using the time-rescaling theorem (see [46] but also [1, 47]), which can be stated as follows. We introduce the deterministic, continuous and non-decreasing function Λ defined by

$$\forall\, t \geq 0, \quad \Lambda(t) = \int_0^t \lambda(u)du.$$

The time-rescaling theorem for Poisson processes states that if N is a Poisson process with intensity $\lambda(.)$, then $\mathcal{N} = \{X = \Lambda(T) :\ T \in N\}$ is an homogeneous Poisson process on $[0, \Lambda(T_{max})]$ with intensity 1. If $\lambda(.)$ is known, it is then possible to use on $\mathcal{N}$ the test of uniformity. This method and many others, still based on the time-rescaling theorem, have been used to test (among many other possible uses) H_0: "the process is a Poisson process of intensity $\lambda(.)$" (see [23, 26, 27]). However, since the seminal paper [23] that also considers the test of uniformity, it has been commonly advertised that if one wants to test a parametric model of the type H_0: "there exists θ such that the process is a Poisson process of intensity $\lambda_\theta(.)$", it is sufficient to build an estimate of θ by using the same sample and to replace the unknown value θ with this estimate in the test statistics and to keep the same quantiles. As we have seen in Theorem 1, this full plug-in step may be desastrous and so, once gain, we use sub-sampling. Therefore, we introduce a second procedure, which is said *cumulated*:

1. Estimate the compensator (or cumulated intensity) with all the trials by

$$\forall t \in [0, T_{max}], \quad \hat{\Lambda}(t) = \int_0^t \hat{\lambda}(u)du, \tag{14}$$

   where $\hat{\lambda}$ is either $\hat{\lambda}_n^{GL}$ or $\hat{\lambda}_n^{Th}$ or $\hat{\lambda}_n^{WN}$.

2. Take a sub-sample $\mathcal{S}$ of size $m(n) = n^\delta$, with $\delta < 1$ ($\delta = 1/2$ or $2/3$).

3. For each trial i in $\mathcal{S}$, change time according to $\hat{\Lambda}$, that is consider

$$\hat{\mathcal{N}}^{(i)} = \{\hat{\Lambda}(T), T \in N^{(i)}\}, \tag{15}$$

point process on $[0, \hat{\Lambda}(T_{max})]$.

4. Cumulate the $\hat{\mathcal{N}}^{(i)}$'s for i in $\mathcal{S}$, as in Figure 11, to obtain a $\hat{\mathcal{N}}^{c,\mathcal{S}}$, point process on $[0, m(n)\hat{\Lambda}(T_{max})]$.

5. Fix a $\theta > 0$ and strictly smaller than $\hat{\Lambda}(T_{max})$.

6. Compute

$$Z^c := \sqrt{\hat{\mathcal{N}}^{c,\mathcal{S}}([0, m(n)\theta])} \sup_{0 \le u \le 1} \left| \frac{1}{\hat{\mathcal{N}}^{c,\mathcal{S}}([0, m(n)\theta])} \sum_{X \in \hat{\mathcal{N}}^{c,\mathcal{S}}, X \le m(n)\theta} \mathbf{1}_{X \le um(n)\theta} - u \right|.$$

7. Reject if $Z^c > \tilde{k}_{1-\alpha}$ (cumulated test by upper value) or reject if $Z^c < \tilde{k}_\alpha$ (cumulated test by lower value).

We can prove the following result.

Theorem 4. *If $\Lambda(T_{max}) > \theta$ and if $\mathbb{E}(\int_0^{T_{max}} (\hat{\lambda}(u) - \lambda(u))^2 du) = O(n^{-\eta})$ with $\delta < \eta$, then both cumulated tests are asymptotically of level α.*

The proof is given in the Additional File 2.

On simulated Poisson processes, the p-values are as expected quite large (see Table 10). If the p-values corresponding to the inhomogeneous case with 40 trials are lower then 0.05, they are still not detected by a BH multiple test procedure.

On the real data sets (see Table 11), most of the couples neuron/direction reject H_0 with small p-values, and the decision is the same for all the considered estimators. Comparing with Table 9, it is clear that for these data, cumulated tests are more powerful than aggregated tests. Moreover piecewise constant estimates combined with cumulated tests are also able to reject H_0. The main exception is N2 where most of the cumulated tests accept H_0. Indeed, if we look at Table 1, the number of rejections associated to this neuron is very small with respect to the other three neurons. Also on Tables 3 and 5, the p-values associated to this neuron are the largest ones (even if they are still small). It was already difficult to distinguish these data (N2) from homogeneous Poisson processes. It is therefore quite natural to accept the more general Poisson hypothesis on these data.

Finally, pooling all the aggregated and cumulated tests of Poisson hypothesis together, 17 couples neurons/directions over 24 clearly reject this hypothesis.

5 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 independence tests either for a given fixed model (most of the time, the homogeneous Poisson process) or via model-free independence tests based on permutations (also called trials-shuffling) [2, 3, 8–10]. It is out of the scope of the present paper to explore all the possible independence tests. We prefer to address the following question: if dependency between spike trains is detected, how to model it?

The simplest model to take the dependency into account is the Hawkes model, which generalizes the homogeneous Poisson processes. It is the point process equivalent to the auto-regressive model. It has first been introduced by Hawkes [48], 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 [49, 50]. More recently it appeared in Neurosciences [12–14].

A more complex notion of intensity is needed to describe those processes: the *conditional intensity*, which extends the classical notion of intensity function of Poisson processes. In our context, the conditional intensity can be viewed as the instantaneous firing rate and may depend on all the previous occurrences (see [47] but also the remarkably detailed paper [26]). More precisely, if a point process N has conditional intensity $\lambda(.)$, then $t \to \lambda(t)$ is a random predictable function that depends on the past before time t. Informally, given the past, the quantity $\lambda(t)dt$ gives the conditional probability to have a new point at time t. This means that if we denote by $\mathcal{F}_{t-}$, the information of the past strictly before time t, we have informally that

$$\mathbb{P}(\text{occurrence in}[t, t+dt] \ \ |\mathcal{F}_{t-}) = \lambda(t)dt. \tag{16}$$

or equivalently

$$\mathbb{E}(dN(t)|\mathcal{F}_{t-}) = \lambda(t)dt \tag{17}$$

This formalism allows us to model dependencies with respect to past occurrences of the process. Since these occurrences are by essence random, we consider conditional probabilities. To go further, if $M_t = N([0,t]) - \int_0^t \lambda(u)du$, then $(M_t)_{t\geq 0}$ is a martingale for the filtration $(\mathcal{F}_t)_{t\geq 0}$. More generally, for any predictable almost

surely bounded process $(H_t)_{t\geq 0}$ (meaning that for all t, H_t only depends on $\mathcal{F}_{t-}$)

$$Z_t = \int_0^t H_s dM_s = \int_0^t H_s[dN(s) - \lambda(s)ds],$$

defines a local martingale. In particular, if Z is integrable, $\mathbb{E}(Z_t) = 0$ but also $\forall s < t, \mathbb{E}(Z_t|\mathcal{F}_s) = Z_s$. This notion of "instantaneous expectation given the past" is very useful, because it provides more intrinsic information on the process, with respect to the classical expectation. For a more complete and classical definition of conditional intensity, see [1, 47].

The multivariate Hawkes process (see for instance [15] or [1]) 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

$$\lambda^{(m)}(t) = \left(\nu^{(m)} + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(m)}(t-u) dN_\ell(u)\right)_+ = \left(\nu^{(m)} + \sum_{\ell=1}^{M} \sum_{T_\ell \in N_\ell,\ T_\ell < t} h_\ell^{(m)}(t - T_\ell)\right)_+ . \tag{18}$$

In (18), 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, all the N_m's behave like homogeneous Poisson processes with constant intensity $\nu^{(m)}$. But as soon as an occurrence appears for a process N_ℓ, then this affects all the processes by increasing or decreasing the conditional intensity via the interaction functions $h_\ell^{(m)}$'s. For instance, if $h_\ell^{(m)}$ is positive and takes large values around the delay d, but is null elsewhere, then for any spike $T_\ell \in N_\ell$, the probability to have a new point for N_m at time $T_\ell + d$ will significantly increase: the process N_ℓ excites the process N_m. On the contrary, if $h_\ell^{(m)}$ is negative around d, then for any spike $T_\ell \in N_\ell$, the probability to have a new point for N_m at time $T_\ell + d$ will significantly decrease: the process N_ℓ inhibits the process N_m. The functions $h_m^{(m)}$ model the self-interaction. So even if $M = 1$, this model is different from a Poisson process.

Note that 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 [15], [51] 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 (18) cannot model non-stationary data but can model dependent data.

Therefore we fix an interval $[a, b] \subset [0, T_{max}]$, typically an interval where all the estimated mean firing rate seem constant. We want 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 $a > A$. Estimation in Hawkes models has been known for a while, in particular for parametric models, using procedures based on the likelihood [23, 50]. As already said in the introduction, it is not possible to make such an assumption in Neurosciences. Classical model selection such as AIC criteria have also been used to select the number of knots of in the spline estimate [49, 52]. However it does not adapt well to irregular functions. This is the reason why nonparametric adaptive inference has recently been developed in such models. If the univariate case ($M = 1$) has been studied in [24], a multivariate approach has been developed in [25] for more general multivariate processes. This is this last method that we want to detail and apply to the spike trains.

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.

5.1 Intensity candidates and least-square contrast on one trial

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

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

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

$$\forall t > 0, \quad \psi_t^{(m)}(f) = \mu^{(m)} + \sum_{\ell=1}^M \int_{-\infty}^{t-} g_\ell^{(m)}(t-u) dN_\ell(u). \tag{19}$$

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 overall closeness with

$$\|\psi(f) - \lambda\|^2 = \sum_{m=1}^M \int_a^b [\psi_t^{(m)}(f) - \lambda^{(m)}(t)]^2 dt. \tag{20}$$

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

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

But using (17), $\tilde{\gamma}(f) \approx \gamma(f)$ with

$$\gamma(f) = -2 \sum_{m=1}^M \int_a^b \psi_t^{(m)}(f) dN_m(t) + \sum_{m=1}^M \int_a^b [\psi_t^{(m)}(f)]^2 dt, \tag{21}$$

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

$$g_\ell^{(m)} = \sum_{k=1}^{K} a_{m,\ell,k}\delta^{-1/2}\mathbf{1}_{((k-1)\delta,k\delta]}, \tag{22}$$

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

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

Let us denote by $\mathbf{a}^{(m)}$ the vector whose transpose is:

$$(\mathbf{a}^{(m)})' = (\mu^{(m)}, a_{m,1,1}, ..., a_{m,1,K}, a_{m,2,1}, ..., a_{m,M,K}). \tag{23}$$

Then one can write

$$\forall t > 0, \quad \psi_t^{(m)}(f) = (\mathbf{Rc}_t)'\mathbf{a}^{(m)}, \tag{24}$$

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

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

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

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

Hence, by (24), proposing $\psi_t^{(m)}(f)$ as a candidate for the intensity 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

$$(\mathbf{b}^{(m)})' = \left(N_m([a,b]), \delta^{-1/2}\mathbf{n}'_{m,1}, ..., \delta^{-1/2}\mathbf{n}'_{m,M}\right),$$

where

$$\mathbf{n}_{m,\ell} = \left(\int_a^b N_\ell([t-k\delta, t-(k-1)\delta])dN_m(t)\right)_{k=1,...,K}$$

is the number of couples (x, y) with $x \in N_m \cap [a, b]$, $y \in N_\ell$ and $(y - x) \in ((k-1)\delta, k\delta]$ and where G is a symmetric matrix of size $1 + MK$ given by

$$\mathbf{G} = \int_a^b \mathbf{Rc}_t (\mathbf{Rc}_t)' dt,$$

the integrated covariation of the renormalized instantaneous count. The solution of this minimization problem is easily available: if $\mathbf{G}$ is invertible,

$$\forall m = 1, ..., M, \quad \hat{\mathbf{a}}^{(m)} = \mathbf{G}^{-1} \mathbf{b}^{(m)}. \tag{25}$$

We provide other heuristic arguments showing that (25) is a natural expression. Using again (17), we can informally write 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^*. Finally, we obtain

$$\int_a^b \mathbf{Rc}_t dN_m(t) = \mathbf{b}^{(m)} \simeq \int_a^b \mathbf{Rc}_t (\mathbf{Rc}_t)' \mathbf{a}_*^{(m)} dt + \text{ noise } \simeq \mathbf{G}\mathbf{a}_*^{(m)} + \text{ noise},$$

showing that (25) should be a good estimate.

5.2 Least-square estimates on several trials and link with joint-PSTH and cross-correlograms

We generally observe $(N_1^{(i)}, \ldots, N_M^{(i)})_{i=1,\ldots,n}$ an i.i.d. sample of a multivariate point process on $[a, b]$. So for each trial i, and each neuron m, we have a corresponding vector $\mathbf{b}^{(m,i)}$ and a corresponding matrix $G^{(i)}$. The least-square contrast for these $n \times M$ spike trains can be written as

$$\gamma_n(f) = \sum_{m=1}^{M} \left(-2(\mathbf{a}^{(m)})' \left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)} \right) + (\mathbf{a}^{(m)})' \left(\sum_{i=1}^{n} \mathbf{G}^{(i)} \right) \mathbf{a}^{(m)} \right) \tag{26}$$

whose solution is given by

$$\forall m = 1, ..., M, \quad \hat{\mathbf{a}}^{(m)} = \left(\sum_{i=1}^{n} \mathbf{G}^{(i)} \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)} \right). \tag{27}$$

The quantity $\left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)}\right)$ can be reinterpreted in terms of cross-correlograms and joint-PSTH, following [53]. Indeed we can write

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

where

$$\bar{\mathbf{n}}_{m,\ell} = \sum_{i=1}^{n} \left(\int_a^b N_\ell^{(i)}([t-k\delta, t-(k-1)\delta]) dN_m^{(i)}(t) \right)_{k=1,\dots,K},$$

and N_m^a is the aggregated process over all the trials. The quantity $\bar{\mathbf{n}}_{m,\ell}$ can be reinterpreted as a particular histogram based on the joint peristimulus time scatter diagram as the joint peristimulus time histogram (JPSTH) or the cross-correlogramm (see Figure 1 of [53] and Figure 13 of the present article). More precisely as detailed in Figure 13, the counts $\bar{\mathbf{n}}_{m,\ell}$ are close to a cross-correlogram except that it is not on a square but on a corner (or more precisely a herringbone). It is consequently more local, as the JPSTH. Also the elements of the partition have the same area and can therefore be compared more easily. On the other hand, if we take several small disjoint intervals $[a, b]$ but with an increasing parameter A, giving the maximal size of the support of the interaction functions, and a small δ, we recover something really 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 effect. Indeed the quantity that is binned for the $\bar{\mathbf{n}}_{m,\ell}$ is the delay itself between two points, whereas for the JPSTH, each position of the points is first binned. Therefore, two points that are at distance 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 Figure 13). This problem of information loss when binning is involved has also been discussed for the coincidences count [9].

JPSTH and cross-correlograms have been used for a long time in Neurosciences, without links with any models. The formula (27) 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 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 effects (and also the fixed effect due to the spontaneous parameter, which need to be substracted) and inverting it enables us to find an estimate of the true interactions.

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. Moreover if we want to capture all the effects, it is preferable to take a large A and a small δ. Therefore the number of parameters of the model increases. With a small number of trials n and a

small interval $[a,b]$, the least-square estimate is doomed to be quite poor. Indeed, if the present least-square contrast is a pure parametric one, for a fixed number K of bins, when the number of parameters becomes large with respect to the number of data, it is not possible to trust parametric estimates anymore, as this least-square estimate, but also as the maximum likelihood estimate. To perform a better estimation and to enhance sparsity, adaptive non parametric methods constitute the most efficient alternatives to proceed beyond the parametric framework.

5.3 Lasso estimate

The Lasso method as developed by [25], is based on the following penalized least-square criterion, reformulated here in the context of n i.i.d. trials:

$$\forall m=1,...,M, \quad \tilde{\mathbf{a}}^{(m)} \in \arg\min_{\mathbf{a}^{(m)}} \left(-2(\mathbf{a}^{(m)})' \left(\sum_{i=1}^{n} \mathbf{b}^{(m,i)} \right) + (\mathbf{a}^{(m)})' \left(\sum_{i=1}^{n} \mathbf{G}^{(i)} \right) \mathbf{a}^{(m)} + 2(\mathbf{d}^{(m)})' |\mathbf{a}^{(m)}| \right), \tag{28}$$

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

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

is a vector of positive observable weights given by

$$d_{m,\ell,k} = \sqrt{2\gamma \ln(n(b-a)) \hat{V}_{m,\ell,k}} + \frac{\gamma \ln(n(b-a))}{3} \hat{B}_{\ell,k}, \tag{29}$$

where

$$\hat{V}_{m,\ell,k} = \sum_{i=1}^{n} \int_a^b \delta^{-1} \left[N_\ell^{(i)}([t-k\delta, t-(k-1)\delta)) \right]^2 dN_m^{(i)}(t),$$

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

and with

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

Since the criterion (28) is convex, the minimization problem can be performed quite easily. Because the penalty term added to the least-square criterion is a weighted ℓ_1-norm, the resulting estimate is sparse and many coefficients in $\tilde{\mathbf{a}}^{(m)}$ will be null (see [54] for the seminal paper on Lasso methods). This estimate and much more general forms have been studied quite intensively in [25]. In Additional File 3, we give a precise oracle inequality for a slight modification of the present estimate. In particular, it is possible to choose the bin δ as a decreasing function of n. It is also possible to prove non-parametric rates of convergence for

similar resulting estimator as for the Goldenshluger and Lepski's method (see the remarks after the oracle inequality of Theorem 2) and Section 5.3 of [25]. There are also several theoretical and practical reasons, discussed in [25], for choosing $\gamma = 1$. However, to avoid too complicated estimation procedures, we prefer in the present article to fix the parameter δ. One can prove the following result.

Theorem 5. *Let us assume that*

- *in* (29)*,* $\hat{B}_{\ell,k}$ *is replaced by* $\kappa_1 \log(n)$ *where* κ_1 *is a fixed positive constant depending of* f^*, a, b, M, A, δ
- $\gamma \geq \kappa$ *where* κ *is an absolute positive constant*
- *the true interaction functions* $h_\ell^{(m)}$ *are piecewise constant and non negative on the partition of length* A *and bin* δ*,*
- *the spontaneous parameters* $\nu^{(m)}$ *are positive*

then there exists a constant C *depending on* f^*, M, A, δ *and* $b - a$ *such that, when* n *tends to infinity, with probability tending to* 1*,*

$$\sum_{m=1}^{M} \|\tilde{\mathbf{a}}^{(m)} - \mathbf{a}_*^{(m)}\|^2 \leq C \frac{\log(n)^3}{n}.$$

In practice, Lasso procedures behave like soft-thresholding rules and therefore there is a small bias which underestimates the non-zero coefficients. To remove this bias, it is sufficient to keep the coordinates that are non zero and perform the least-square estimate on the set of non-zero coefficients. The first procedure is referred as **B**, for the Bernstein's inequality which gives the shape of (29) (see [25]). The second procedure where we also perform an ordinary least-square step is referred as **BO**.

The reconstructions for simulated data are given in Figure 14. The non zero coefficients are adequately estimated and the method **BO** is as expected closer to the truth than **B**, which has a negative bias. On the real data sets, we also took $[a, b] = [1, 1.5]$ which coincides with a large and flat zone for $\hat{\lambda}_n^{Th}$ on both neurons. We have not found any interaction between both neurons for both couples - (N1,N2) and (N3,N4) - and we just show two typical results in Figures 15 and 16. This is coherent with the fact that both pairs have been studied in [10] and that the MTGAUE multiple test of independence of this article always accepts the independence hypothesis between both neurons. We have always found at least one and usually two negative self-interactions, at very close range. Hence it seems that both neurons in each data set are independent, do not obey the Poisson distribution but are self-inhibiting and this can be consistent with a refractory period. This is also coherent with the fact that most of the tests of the Poisson hypothesis have rejected (see Tables

8, 9 and 11). In particular the negative self interaction is really strong for N2, which compensates the strong spontaneous rate. Note that even if Theorem 5 only holds for positive interactions, the main theorem of [25] relies on a martingale approach and that up to the restriction to the event $\{\lambda^{(m)}(.) > 0\}$, all the previous results apply (see also [24] for a complete study of the potentially negative interaction case).

5.4 Goodness-of-fit Test

First, let us review the previous goodness-of-fit tests and let us perform them on the bivariate simulated Hawkes process (S-Haw) (N_1, N_2). The test of uniformity shows that when the test is performed on each trial, most of the trials are rejected (see Table 1). We draw the same conclusions by using the small p-values of the aggregated test of uniformity for N_1 (see Table 2). However, Table 2 shows a different behavior of N_2 in this case since the p-values of the aggregated test of uniformity are quite large. This can be explained by the fact that there is only one interaction function that affects N_2, whereas two functions (and one with a large support) affect N_1. Therefore interactions in N_2 play a less important role than in N_1 and are consequently not detected. In particular, the power of this test with respect to Hawkes processes probably does not increase with the number of trials and this explains the change of behavior for N_2 between Table 1 and Table 2. The test of Poisson variables clearly rejects the hypothesis that the total number of points per trial is Poissonian (see Table 4), whereas locally the test of exponentiality accepts this hypothesis (see Table 6).

If we apply now the aggregated test of Poisson hypothesis on the simulated Hawkes process (see Table 7), it rejects for N_1 but accepts for N_2. This can be explained by the different significance of the interactions for N_1 and N_2, as explained previously. When we apply the cumulated test (see Table 10), it rejects for both N_1 and N_2. Hence as expected, the cumulated test is more powerful than the aggregated test.

As for Poisson processes, we want to test now H_0: "the processes are Hawkes processes". In fact the time-rescaling theorem holds for more general point processes with conditional intensity [47] (see also the very pedagogical paper [26]). Actually Ogata already proposed several residuals analysis in the context of general point processes and in particular for Hawkes processes [23]. But, as before, the plug-in was not justified and was based on the whole sample. Since there is no meaning in aggregating Hawkes processes because that would mix different pasts, we turn towards the cumulated tests. Let us describe more precisely the procedure. For a fixed interval $[a, b]$ the test can be described as follows.

1. Estimate each $\mathbf{a}_*^{(m)}$ by $\tilde{\mathbf{a}}^{(m)}$ either by the **B** or by the **BO** method as described above.

2. Take a sub-sample $\mathcal{S}$ of size $m(n) = n^\delta$ with $\delta < 1$ ($\delta = 1/2$ or $2/3$).

3. For each trial i in $\mathcal{S}$, change time according to

$$\hat{\Lambda}^{(i)}(t) = \int_0^t (\mathbf{R}\mathbf{c}_t^{(i)})' \tilde{\mathbf{a}}^{(m)} dt,$$

and consider

$$\mathcal{N}^{(i)} = \{\hat{\Lambda}^{(i)}(T), \quad T \in \mathbb{N}^{(i)}\},$$

point process on $[0, \hat{\Lambda}^{(i)}(T_{max})]$.

4. Cumulate the $\hat{\mathcal{N}}^{(i)}$'s for i in $\mathcal{S}$, as in Figure 12, to obtain a $\hat{\mathcal{N}}^{c,\mathcal{S}}$, point process on $[0, \sum_{i\in\mathcal{S}} \hat{\Lambda}^{(i)}(T_{max})]$.

5. Fix $\theta > 0$, strictly smaller than $\sum_{i\in\mathcal{S}} \hat{\Lambda}^{(i)}(T_{max})/m(n)$.

6. Compute

$$Z^c := \sqrt{\hat{\mathcal{N}}^{c,\mathcal{S}}([0, m(n)\theta])} \sup_{0\le u\le 1} \left| \frac{1}{\hat{\mathcal{N}}^{c,\mathcal{S}}([0, m(n)\theta])} \sum_{X\in\hat{\mathcal{N}}^{c,\mathcal{S}}, X\le m(n)\theta} \mathbf{1}_{X\le um(n)\theta} - u \right|.$$

7. Reject if $Z^c > \tilde{k}_{1-\alpha}$ (cumulated test by upper values) or reject if $Z^c < \tilde{k}_\alpha$ (cumulated test by lower values).

Then one can show the following result

Theorem 6. *If* $\mathbb{E}(\int_0^{T_{max}} (\mathbf{R}\mathbf{c}_t^{(i)})' \mathbf{a}_*^{(m)} dt) > \theta$, *then under the Assumptions of Theorem 5, with estimation method* $\mathbf{B}$, *both previous tests are asymptotically of level* α.

We applied this test on simulated and real data. On Table 12 with simulated Hawkes processes, all p-values are quite large (none of them being detected by a BH multiple test method). On Table 13 with the real data sets and on the whole interval $[0, 2]$, only 5 couples neuron/direction over the 24 considered couples reject H_0 by the BH multiple test procedure. When we reduce the interval to $[1, 1.5]$ on which $\hat{\lambda}_n^{Th}$ does not vary, then there is no rejection.

This conclusion seems at first sight contradictory with the estimation performed as a Poisson process. Indeed it seemed that the real processes were usually far from being stationarity because the reconstructions were not smooth at all (see for instance reconstructions associated with N3 in direction 5 in Figure 11). We applied these reconstructions algorithms on the process N_1 of the bivariate Hawkes process defined in Figure 14. The result is given in Figure 17. We see that the reconstructions are highly irregular. Indeed, Poissonian

methods are tuned to control Poissonian randomness and not randomness due to Hawkes distributions, which can be much larger (for instance the number of points of a Hawkes process with positive interactions has exponential moment but not of every order and in this respect, the distribution has heavier tail than the Poisson distribution [51]). Note also that the small number of trials (40 in the simulation) tends to increase these fluctuations and that increasing the number of trials stabilizes this effect: the estimate for a large number of trials reproduces a constant intensity, similar to the black line in Figure 17, which is the mean stationary intensity expected for this Hawkes process.

It is consequently coherent to find reconstructions as in Figure 11 and to accept nevertheless, on the whole interval, the Hawkes hypothesis which should make it more stationary. Note however, that even if the tests are all accepted on $[1, 1.5]$, it does not mean than the Hawkes model is the correct model for this interval. This simply means that there is no contradiction between the data and this model. This can also be due to a lack of power because we restricted the interval $[a, b]$ hence taking into account a smaller number of spikes.

To summarize, 7 couples accept both Hawkes and Poisson hypotheses on $[0, T_{max}]$, 12 only accept the Hawkes hypothesis on $[0, T_{max}]$. The remaining 5 couples reject both hypotheses on this interval but accept the Hawkes hypothesis on a smaller one.

6 Conclusion

Point processes are a natural framework when dealing with spike trains analysis. We have imported new or recent non parametric statistical tools to perform adaptive estimation in different models: inhomogeneous Poisson processes and Hawkes processes. We have provided goodness-of-fit tests that are able to test each model. We have proved that a subsampling procedure allows us to plug-in the various non-parametric estimates inside variants of Kolmogorov-Smirnov tests and we have proved that this procedure leads to tests for which asymptotical levels are controlled. We have also shown on different real data sets that if the homogeneous Poisson assumption is never satisfied, inhomogeneous Poisson processes are probably non-convenient alternatives. It seems indeed that models have to take at least into account self-inhibition, which is possible with Hawkes processes. It is worth noticing that using Hawkes models is not the only way to model interactions and self-interactions. Several other models exist (see for instance [55] or [56]). It is also possible to detect dependency without any assumptions on the models (see for instance [2]). However, if we want indeed to fix a model, one main question remains open. Can we find a model close in spirit to Hawkes model where we can estimate and test non-stationarity and interactions in an adaptive way?

7 Acknowledgements

Special thanks go to Alexa Riehle, leader of the Lab in which the data used in this article were previously collected. The authors wish also to thank F. Picard for fruitful discussions at several steps of this work. This research is partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration) and by the PEPS BMI 2012-2013 *Estimation of dependence graphs for thalamo-cortical neurons and multivariate Hawkes processes.*

References

1. 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].
2. Pipa G, Grün S: **Non-parametric significance estimation of joint-spike events by shuffling and resampling.** *Neurocomputing* 2003, **52-54**:31–37.
3. 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.
4. Dayan P, Abbott LF: *Theoretical neuroscience*. Computational Neuroscience.
5. Georgopoulos A, Schwartz A, Kettner R: **Neuronal population coding of movement direction.** *Science* 1986, (233):1416–1419.
6. Rizzolatti G, Craighero L: **The mirror-neuron system.** *Annual Review of Neuroscience* 2004, **27**:169–192.
7. Shinomoto S: *Analysis of Parallel Spike Trains*, Springer Series in Computational Neuroscience 2010 chap. Estimating the Firing Rate.
8. Grün S: **Unitary joint-events in multiple-neuron spiking activity: Detection, significance and interpretation.** *PhD thesis*, Thun: Verlag Harri Deutsch 1996.
9. 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.
10. Tuleau-Malot C, Rouis A, Reynaud-Bouret P, Grammont F: **Multiple Tests based on a Gaussian Approximation of the Unitary Events method**. Tech. rep., http://hal.archives-ouvertes.fr/hal-00757323 2012.
11. Grün S, Diesmann M, Aertsen A: *Analysis of Parallel Spike Trains*, Springer Series in Computational Neuroscience 2010 chap. Unitary Events Analysis.
12. 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).
13. Pernice V, Staude B, Cardanobile S, Rotter S: **How structure determines correlations in neuronal networks.** *PLoS Computational Biology* 2012, (85:031916).
14. 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).
15. Brémaud P, Massoulié L: **Stability of nonlinear Hawkes processes**. *Ann. Probab.* 1996, **24**(3):1563–1588.
16. 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].
17. Andersen PK, Borgan Ø, Gill RD, Keiding N: *Statistical models based on counting processes*. Springer Series in Statistics, New York: Springer-Verlag 1993.
18. 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].
19. Rudemo M: **Empirical choice of histograms and kernel density estimators**. *Scand. J. Statist.* 1982, **9**(2):65–78.
20. Shimazaki H, Shinomoto S: **Kernel bandwidth optimization in spike rate estimation.** *Journal of Compuational Neurosciences* 2010, **29**:171–182.
21. Goldenshluger A, Lepski O: **Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality**. *Ann. Statist.* 2011, **39**(3):1608–1632.
22. 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.
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. Reynaud-Bouret P, Schbath S: **Adaptive estimation for Hawkes processes; application to genome analysis**. *Ann. Statist.* 2010, **38**(5):2781–2822.

25. Hansen N, Reynaud-Bouret P, Rivoirard V: **Lasso and probabilistic inequalities for multivariate point processes**. Tech. rep., http://arxiv.org/abs/1208.0570 2012.

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

27. 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.

28. 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.

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

30. Cohen A, Sackrowitz HB: **Evaluating tests for increasing intensity of a Poisson process**. *Technometrics* 1993, **35**(4):446–448.

31. Kingman JFC: *Poisson processes*, *Volume 3 of* Oxford Studies in Probability. New York: The Clarendon Press Oxford University Press 1993. [Oxford Science Publications].

32. 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.

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

34. Wasserman L: *All of statistics*. Springer Texts in Statistics, New York: Springer-Verlag 2004. [A concise course in statistical inference].

35. 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.

36. Fromont M, Laurent B: **Adaptive goodness-of-fit tests in a density model**. *Ann. Statist.* 2006, **34**(2):680–720.

37. Fromont M, Laurent B, Reynaud-Bouret P: **Adaptive tests of homogeneity for a Poisson process**. *Ann. Inst. Henri Poincaré Probab. Stat.* 2011, **47**:176–213.

38. Dacunha-Castelle D, Duflo M: *Probability and statistics. Vol. II*. New York: Springer-Verlag 1986. [Translated from the French by David McHale].

39. Kutoyants YA: *Statistical inference for spatial Poisson processes*, *Volume 134 of* Lecture Notes in Statistics. New York: Springer-Verlag 1998.

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

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

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

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

49. 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).

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

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

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

53. 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.

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

55. Sansonnet L, Tuleau-Malot C: **A model of Poissonian interactions and detection of dependence**. Tech. rep., http://hal.archives-ouvertes.fr/hal-00780598 2013.

56. Pakdaman K, Perthame B, Salort D: **Dynamics of a structured neuron population**. *Nonlinearity* 2010, **23**:55–75.

57. 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.

58. 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.

59. Hochberg Y, Tamhame A: *Multiple comparison procedures*. New York: Wiley 1987.

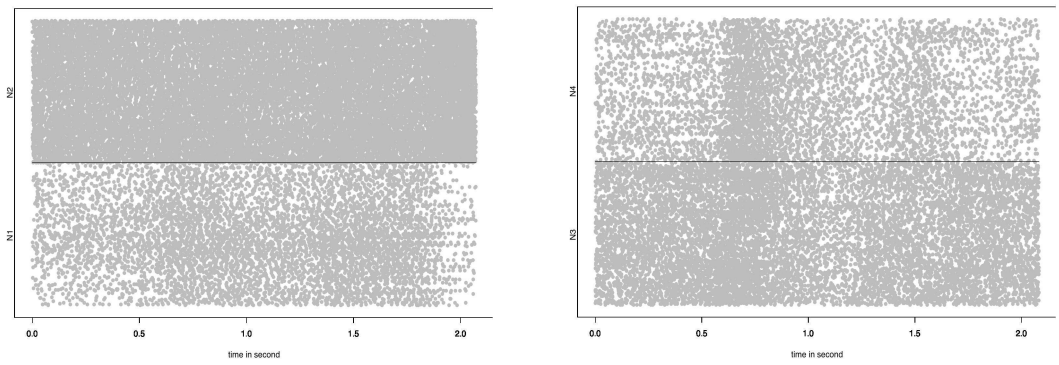
60. Holm S: **A simple sequentially rejection multiple test procedure**. *Scandinavian Journal of Statistics* 1979, **6**(2):65–70.

61. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency**. *Ann. Statist.* 2001, **29**(4):1165–1188.

8 Figures

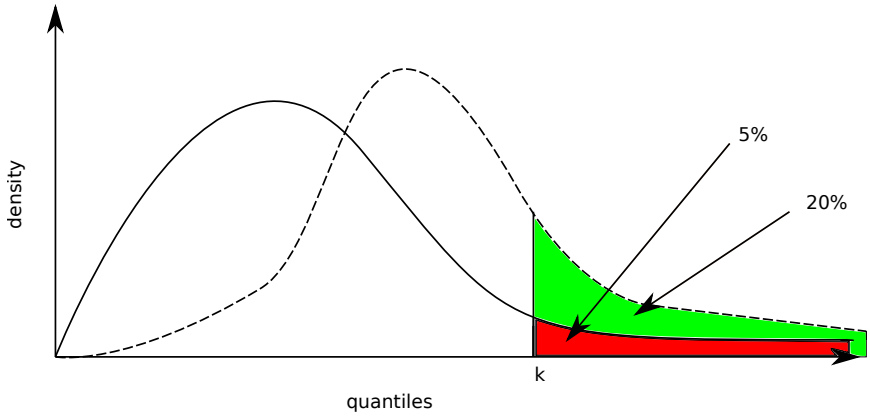
8.1 Figure 1 - Raster plots of the data sets

On the left the rasters associated to the first data set i.e. (N1,N2). On the right, the ones associated to the second data set, i.e. (N3,N4). Each line corresponds to a trial, each dot to a spike.



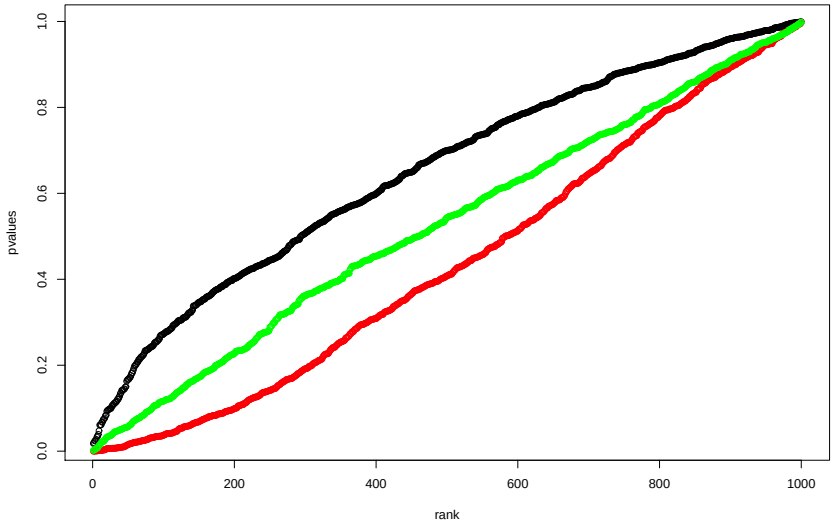
8.2 Figure 2 - Density and quantiles

Density of the KS_p statistics under H_0 in plain black. The quantile $k = k_{p,1-\alpha} = k_{p,0.95}$ is represented on the axis. It corresponds to an area under the curve in red of 5%. A possible density for the KS statistics under H_1 is plotted in dotted black. The area under the curve in green corresponds to the power of the test (here 20%). Even if the density is different from the one under H_0, the power of the test is small.



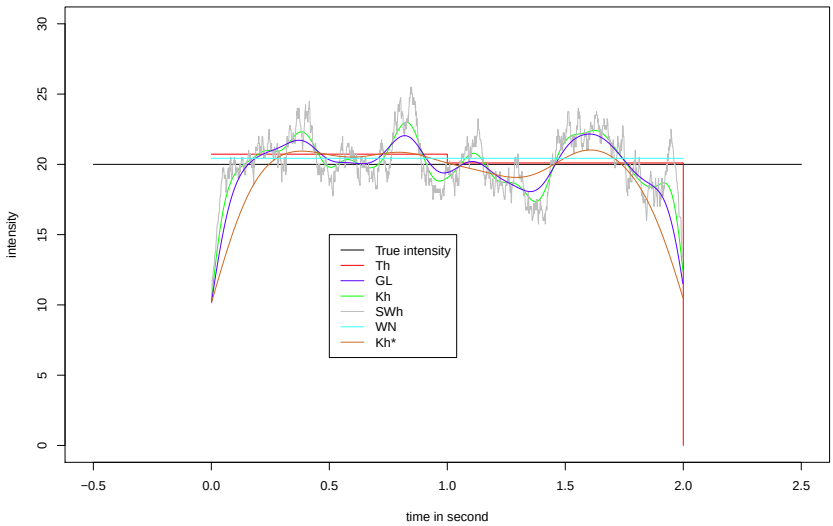
8.3 Figure 3 - Repartition of the p-values in a KS test with plug-in

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 KS test with exactly the same sample (in black), or by estimating the parameter λ on half of the sample and performing the KS test on the other half (in red), or by estimating the parameter λ on the whole sample and performing the KS 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 $n = 1000$).



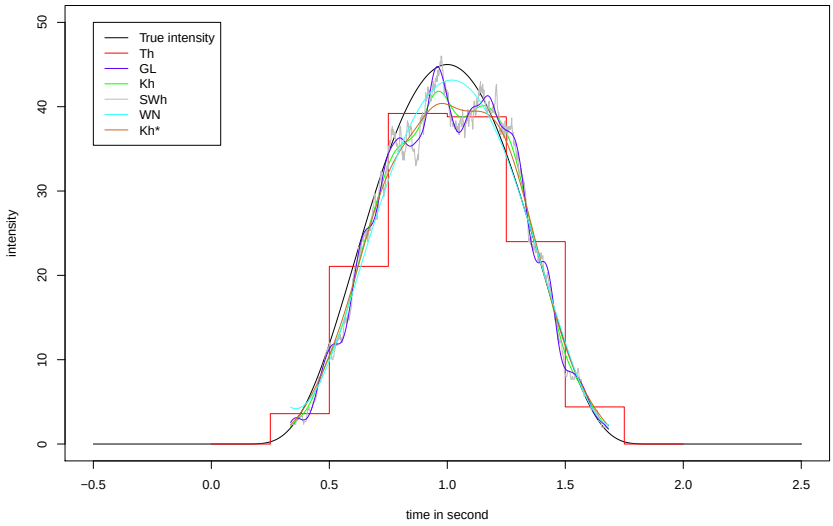
8.4 Figure 4 - Reconstructions for homogeneous simulated data

Reconstructions for a simulated Poisson process with constant intensity $\lambda = 20$ over 40 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.



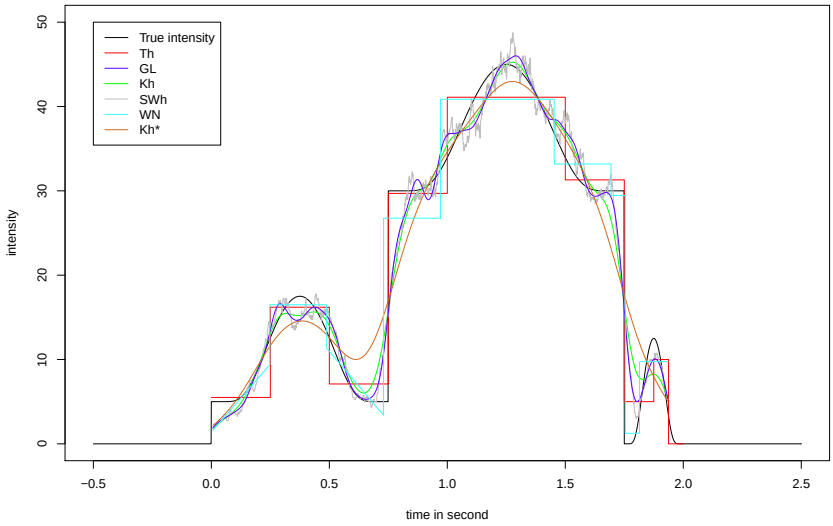
8.5 Figure 5 - Reconstructions for smooth inhomogeneous simulated data

Reconstructions for a simulated Poisson process with smooth Gaussian intensity over 30 trials, observed on $[0, 2]$. Same abbreviations as in Figure 4.



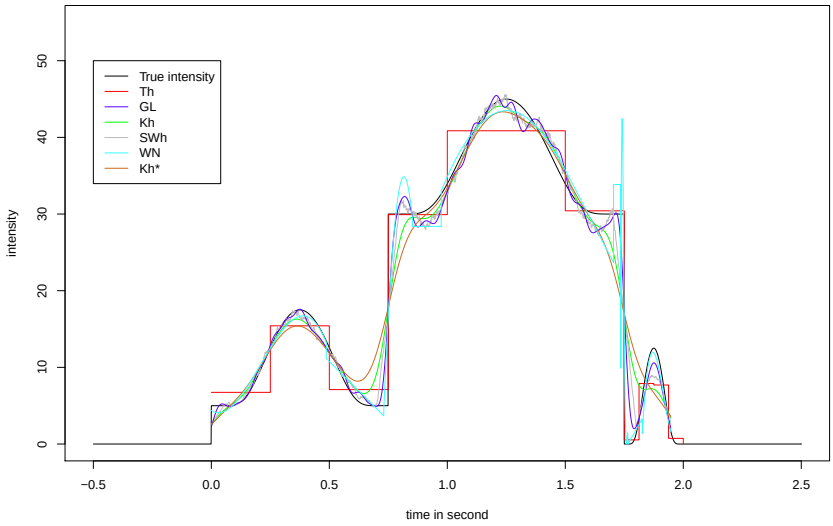
8.6 Figure 6 - Reconstructions for piecewise continuous intensity, small sample size

Reconstructions for a simulated Poisson process with piecewise continuous intensity over 40 trials, observed on $[0, 2]$. Same abbreviations as in Figure 4.



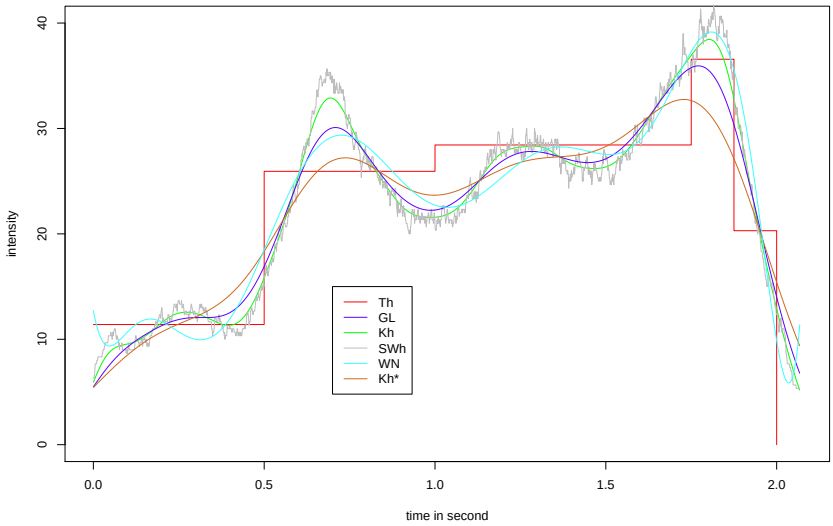
8.7 Figure 7 - Reconstructions for piecewise continuous intensity, large sample size

Reconstructions for a simulated Poisson process with the piecewise continuous intensity of Figure 5 over 200 trials, observed on $[0, 2]$. Same abbreviations as in Figure 4.



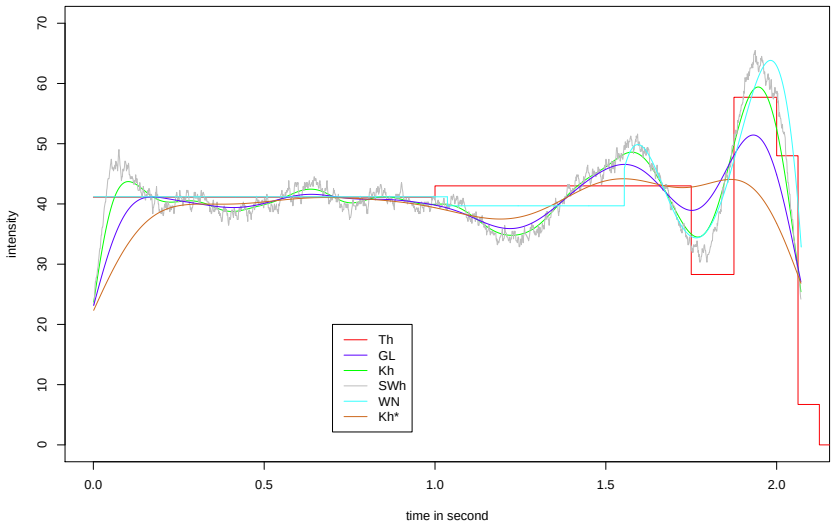
8.8 Figure 8 - Reconstructions for N1 in direction 4

Same abbreviations as in Figure 4.



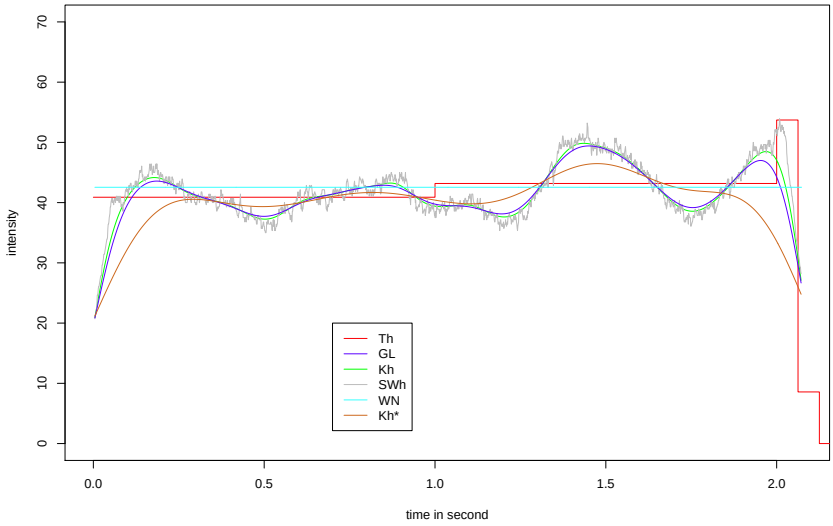
8.9 Figure 9 - Reconstructions for N2 in direction 1

Same abbreviations as in Figure 4.



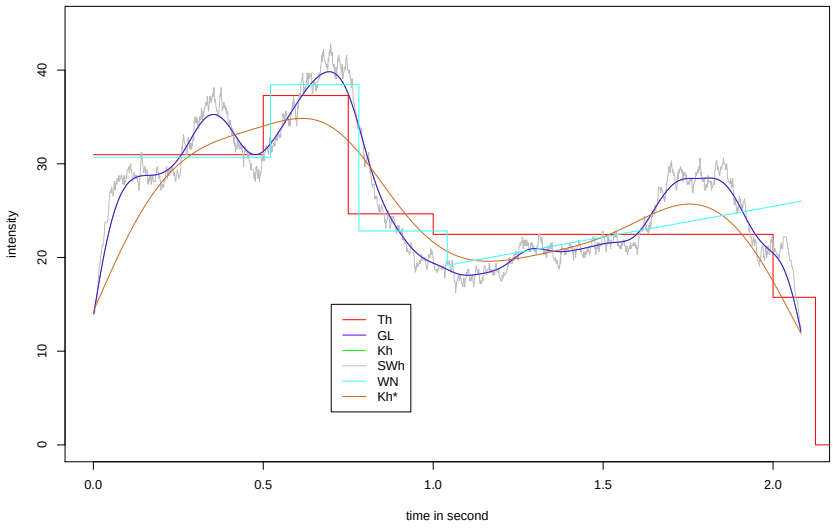
8.10 Figure 10 - Reconstructions for N2 in direction 3

Same abbreviations as in Figure 4.



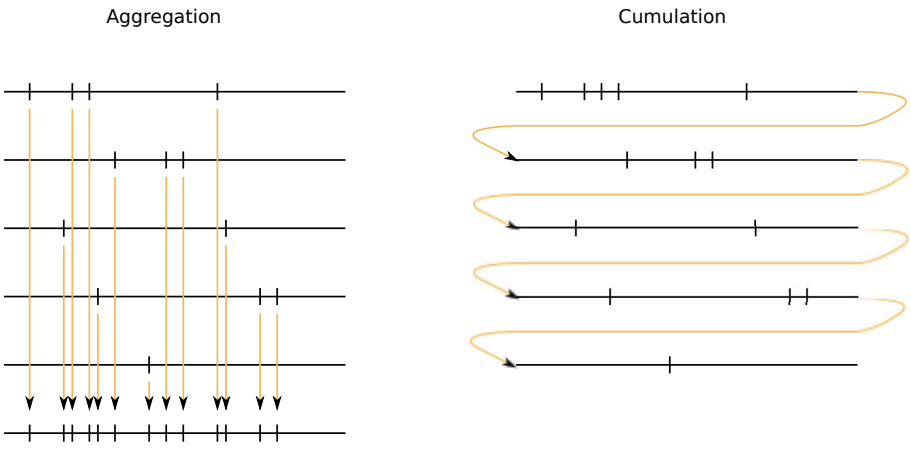
8.11 Figure 11 - Reconstructions for N3 in direction 5

Same abbreviations as in Figure 4.



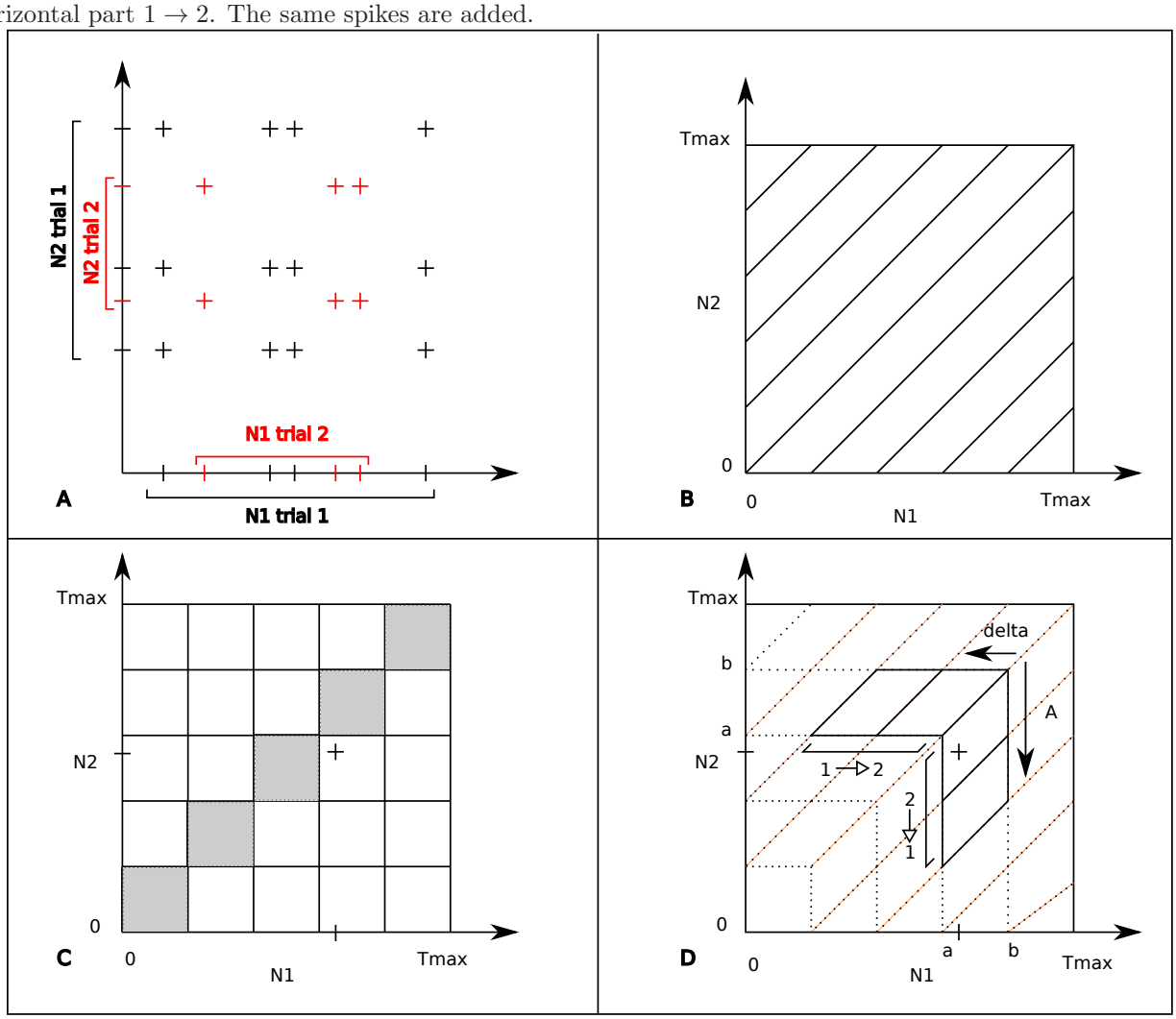
8.12 Figure 12 - Agregation versus Cumulation

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.



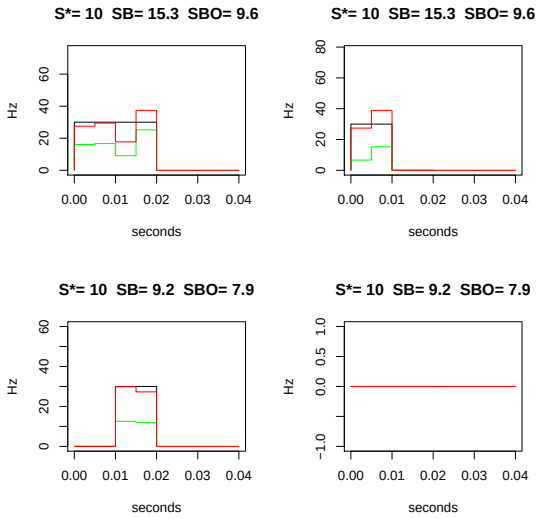
8.13 Figure 13 - Scatter diagrams and histograms

In Figure 13 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 13 B gives the partition corresponding to the classical cross-correlogram. Figure 13 C gives the partition of the joint peristimulus time histogram. The diagonal squares are filled in gray. Two very close spikes (one on N1, the other on N2) and their corresponding point in the scatter diagram are added. Figure 12 D gives the partition used for the computations 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.



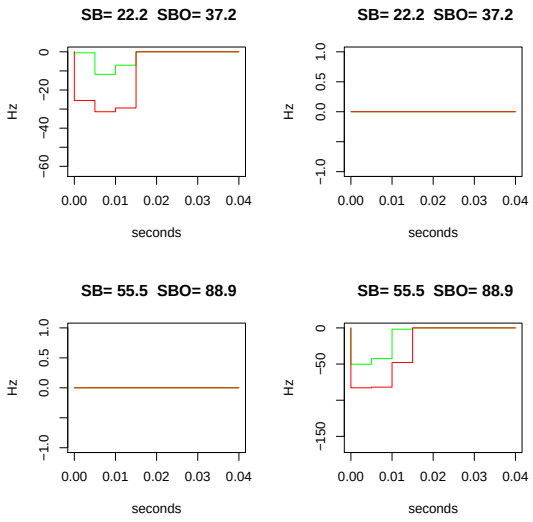
8.14 Figure 14 - Lasso Reconstructions for a simulated Hawkes process

A bivariate Hawkes process (N_1, N_2) has been simulated on $[0, 2]$ for 40 trials. The considered interval $[a, b]$ is $[1, 1.5]$. On the upper left, the interaction functions $h_1^{(1)}$, 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 **B**, in red, reconstruction by method **BO**. On the top of each graphics, the true spontaneous parameter ν_1 on the top and the true parameter ν_2 on the bottom are referred by `S*`. The estimated spontaneous parameters by method **B** are referred as `SB` and the estimated spontaneous parameters by method **BO** are referred as `SBO`.



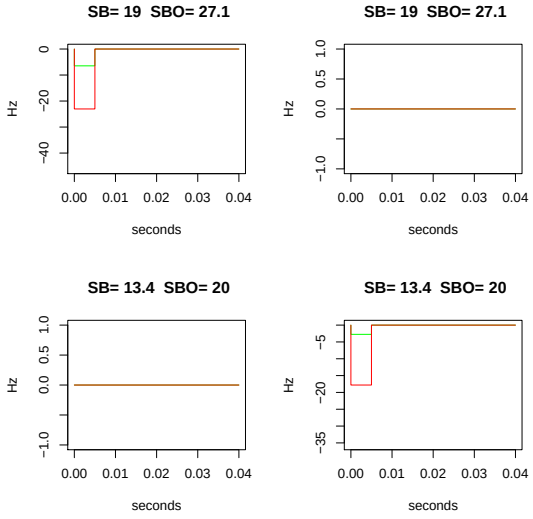
8.15 Figure 15 - Lasso Reconstructions for (N1,N2) in direction 4

The interval $[a, b] = [1, 1.5]$. Same conventions as in Figure 14.



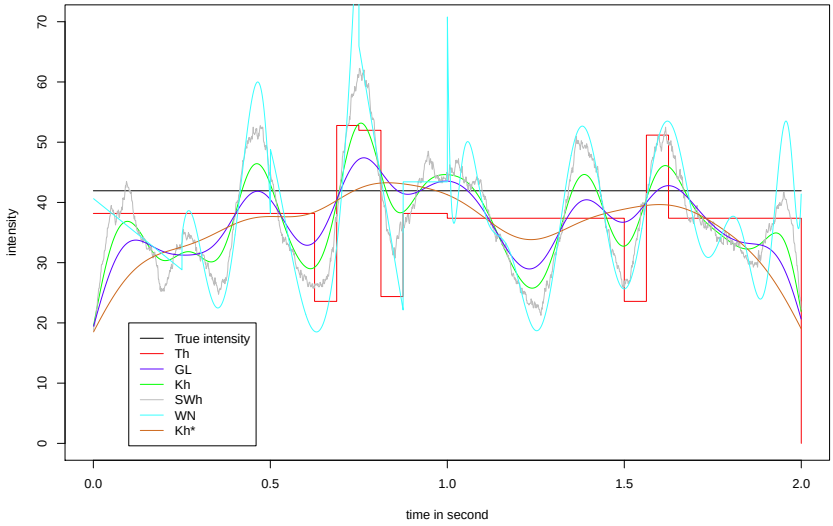
8.16 Figure 16 - Lasso Reconstructions for (N3,N4) in direction 4

The interval $[a, b] = [1, 1.5]$. Same conventions as in Figure 14.



8.17 Figure 17 - Reconstructions of the intensity as a Poisson process for a Hawkes process

Intensity estimation by the methods described in Figure 4 for (S-Haw) N_1. The process was observed on $[0, 2]$ over 40 trials. In black, the average intensity i.e. $\mathbb{E}(\lambda(t))$ if the process was stationary.



9 Tables

9.1 Table 1 - KS tests on each trial

The table gives the number of rejected trials by using the KS test of uniformity at the level $\alpha = 5\%$ on 200 simulated data and on real data sets. For real data sets, the size of the data set is given in brackets.

Simulated data with $n = 200$				Real data sets			
S-HomPoi	S-InPoi	S-Haw (N_1)	S-Haw (N_2)	N1 (177)	N2 (177)	N3 (165)	N4 (165)
10	142	198	178	20	3	25	38

9.2 Table 2 - KS tests of uniformity on simulated data

P-values of the KS test of uniformity on simulated data sets. The p-values that are detected as too small by a BH multiple test procedure with prescribed FDR of 5% are exactly the ones corresponding to (S-InPoi) and (S-Haw) N_1.

Number of trials	40	200
S-HomPoi	0.93	0.45
S-InPoi	$< 10^{-16}$	$< 10^{-16}$
S-Haw (N_1)	0.008	0.003
S-Haw (N_2)	0.46	0.79

9.3 Table 3 - KS tests on the aggregated data sets

P-values of the KS test of uniformity performed on real data sets aggregated either direction per direction or aggregated over all the directions (pooled). Since all the p-values are less than 5%, they are all detected

as significantly small by a BH multiple test procedure with prescribed FDR of 5%.

Directions	1	2	3	4	5	6	Pooled
N1	9.10^{-5}	$< 10^{-16}$	1.10^{-5}	2.10^{-6}	2.10^{-14}	1.10^{-14}	$< 10^{-16}$
N2	1.10^{-6}	4.10^{-4}	1.10^{-4}	1.10^{-2}	1.10^{-3}	7.10^{-4}	$< 10^{-16}$
N3	$< 10^{-16}$	2.10^{-7}	1.10^{-14}	6.10^{-15}	8.10^{-5}	4.10^{-5}	$< 10^{-16}$
N4	$< 10^{-16}$	1.10^{-5}	9.10^{-10}	1.10^{-14}	1.10^{-4}	$< 10^{-16}$	$< 10^{-16}$

9.4 Table 4 - Chi-square tests of Poissonian variables on simulated data

P-values of the chi-square tests of Poissonian variables on simulated data sets. The p-values that are detected as too small by a BH multiple test procedure with FDR 5% are exactly the ones corresponding to (S-Haw).

Number of trials	40	200
S-HomPoi	0.43	0.91
S-ImPoi	0.46	0.12
S-Haw (N_1)	6.10^{-6}	$< 10^{-16}$
S-Haw (N_2)	2.10^{-4}	$< 10^{-16}$

9.5 Table 5 - Chi-square tests of Poissonian variables on the aggregated data sets

P-values of the chi-square test of Poissonian variables performed on real data sets aggregated either direction per direction or aggregated over all the directions (pooled). When a Benjamini and Hochberg multiple test method is run on the 24 p-values corresponding to the 24 couples neuron/direction, none of the p-values are considered as significant at the prescribed FDR of 5%.

Directions	1	2	3	4	5	6	Pooled
N1	0.02	0.01	0.3	0.04	0.016	0.007	$< 10^{-16}$
N2	0.19	0.86	0.52	0.73	0.28	0.15	0.29
N3	0.039	0.55	0.32	0.2	0.76	0.1	3.10^{-9}
N4	0.02	0.18	0.15	0.37	0.08	0.7	6.10^{-15}

9.6 Table 6 - Local test of exponentiality

P-values of the local test of exponentiality on simulated and real data sets. We always take $m = \lfloor n^{2/3} \rfloor$.

S-HomPoi		S-InPoi				S-Haw N_1		N1		N3	
$n = 40$	$n = 200$	$n = 40$		$n = 200$		$n = 40$	$n = 200$	Direction 4		Direction 5	
$t = 0.5$		$t = 0.67$	$t = 1.2$	$t = 0.67$	$t = 1.2$	$t = 0.5$		$t = 0.42$	$t = 0.5$	$t = 0.42$	$t = 0.67$
0.48	0.54	0.02	0.93	0.01	0.41	0.30	0.22	0.70	0.27	0.18	0.70

9.7 Table 7 - Aggregated tests of Poisson hypothesis on simulated data

P-values of the aggregated test by upper values of the Poisson hypothesis for (S-HomPoi), (S-InPoi) and (S-Haw). To obtain the p-values for the test by lower values it is sufficient to take $1-p$ where p is the p-value of the test by upper values. The sub-sample is taken with $m = \lfloor n^{2/3} \rfloor$. Th , GL and WN correspond to $F_{\hat{\lambda}}$ where $\hat{\lambda}$ is equal respectively to $\hat{\lambda}_n^{Th}, \hat{\lambda}_n^{GL}$ and $\hat{\lambda}_n^{WN}$. F_n is the empirical c.d.f on all the n trials. The p-values that are detected as too small by a BH multiple test procedure with FDR 5% for both aggregated tests (lower and upper values), are exactly the ones corresponding to (S-Haw) N_1.

Estimate for $\hat{F}$	F_n	Th	GL	WN
S-HomPoi (40 trials)	0.53	0.58	0.14	0.51
S-InPoi (40 trials)	0.52	0.91	0.96	0.89
S-InPoi (200 trials)	0.53	0.59	0.47	0.55
S-Haw N_1 (40 trials)	2.10^{-7}	5.10^{-9}	3.10^{-6}	2.10^{-7}
S-Haw N_2 (40 trials)	0.55	0.15	0.66	0.13

9.8 Table 8 - Aggregated tests of Poisson hypothesis with $\hat{F} = F_n$ on the real data sets

P-values of the aggregated test of Poisson hypothesis with $\hat{F} = F_n$ on the real data sets. The sub-sample is taken with $m = \lfloor n^{2/3} \rfloor$. The following code is used: $\triangle$ corresponds to a p-value of the test by upper values in $[10^{-3}, 10^{-2})$, $\triangle\triangle$ to a p-value of the test by upper value in $[10^{-4}, 10^{-3})$, $\triangle\triangle\triangle$ to a p-value of the test by upper value in $(-\infty, 10^{-4})$. The same code is used but with $\triangledown$ for the p-value of the test by lower values. The triangles are black if the p-values corresponds to rejection of a BH multiple test method based on both p-values (upper and lower) for all couples neuron/direction (NB: because of the number of tests, no p-value that is larger than 0.01 can be rejected). PR corresponds to pooled with sub-sample $\mathcal{S}$ taken at random, Pi corresponds to pooled with sub-sample $\mathcal{S}$ taken as direction i.

Directions	1	2	3	4	5	6	PR	P1	P2	P3	P4	P5	P6
N1	▼▼	▼▼											
N2			▼▼						▼▼				
N3													
N4			▼▼▼										

9.9 Table 9 - Aggregated tests of Poisson hypothesis with $\hat{F} = F_{\hat{\lambda}}$ on the real data sets

P-values of the aggregated test of Poisson hypothesis with $\hat{F} = F_{\hat{\lambda}}$ on the real data sets. Same conventions as in Tables 7 and 8.

Directions		1	2	3	4	5	6
N1	Th GL WN				 ▽		
N2	Th GL WN						
N3	Th GL WN	 ▲▲	▲ ▲▲▲	 ▲▲			
N4	Th GL WN	 △	 △	 ▲ ▼▼	 △		

9.10 Table 10 - Cumulated tests of Poisson hypothesis on simulated data

P-values of the cumulated test by upper values of inhomogeneous Poisson hypothesis by using (S-HomPoi), (S-InPoi) and (S-Haw).The sub-sample is taken with $m = \lfloor n^{2/3} \rfloor$ each time. Same convention as in Table for the test by lower values. The p-values that are detected as too small by a BH multiple test procedure with FDR 5% for both cumulated tests (lower and upper values), are exactly the ones corresponding to (S-Haw).

Estimate for $\hat{F}$	Th	GL	WN
S-HomPoi (40 trials)	0.14	0.15	0.15
S-InPoi (40 trials)	0.03	0.03	0.03
S-InPoi (200 trials)	0.11	0.11	0.11
S-Haw N_1 (40 trials)	8.10^{-6}	4.10^{-6}	5.10^{-6}
S-Haw N_2 (40 trials)	1.10^{-5}	1.10^{-5}	1.10^{-5}

9.11 Table 11 - Cumulated tests of Poisson hypothesis on the real data sets

P-values of the cumulated test of inhomogeneous Poisson hypothesis on the real data sets. The sub-sample is taken with $m = \lfloor n^{2/3} \rfloor$ each time. Same conventions as in Tables 7 nd 8.

Directions		1	2	3	4	5	6
N1	Th	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲
	GL	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲
	WN	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲	▲▲▲
N2	Th			▲▲			
	GL						
	WN						
N3	Th		▲▲	▲			▲
	GL		▲▲	▲▲			▲
	WN		▲▲	▲▲			▲
N4	Th			▲▲▲	▲▲▲	▲▲▲	
	GL	▲	▲	▲▲▲	▲▲▲	▲▲▲	▲
	WN			▲▲▲	▲▲▲	▲▲▲	

9.12 Table 12 - Cumulated tests of Hawkes hypothesis on simulated data

P-values of the cumulated test by upper values of Hawkes hypothesis, for both Lasso methods **B** and **BO**. Same conventions as in Table 7 for the test by lower values. None of the p-values are detected significant by a BH multiple test method with prescribed FDR of 5%.

	$[a, b]$	**B**	**BO**
Hawkes N_1	$[0, 2]$	0.2	0.08
Hawkes N_2		0.02	0.59
Hawkes N_1	$[1, 1.5]$	0.12	0.27
Hawkes N_2		0.17	0.39

9.13 Table 13 - Cumulated tests of Hawkes hypothesis on real data sets

P-values of the cumulated test of Hawkes hypothesis on the real data sets. Same conventions as in Tables 7 and 8. Here $[a, b] = [0, 2]$. If $[a, b] = [1, 1.5]$, none of the p-values are detected significant by a BH multiple test method with prescribed FDR of 5%.

Directions		1	2	3	4	5	6
N1	**B**	▲▲	△				
	BO		△	▲	▲▲▲		
N2	**B**					△	
	BO						
N3	**B**						
	BO		△				
N4	**B**		▲▲				
	BO				▲		

10 Additional Files

10.1 Additional file 1 — 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, with intensity $n\lambda(.)$, exists now on $\mathbb{R}$, but there is no points outside $[0, T_{max}]$ and $N^a(\mathbb{R}) = N^a([0, T_{max}])$. The proof is inspired by Proposition 2 of [57]. We set:

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

Therefore, we have:

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

and

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

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

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

with

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

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

and

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

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

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

Therefore, by setting

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

we have:

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

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

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

For the last term, we obtain:

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

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 1. *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 $h \in \mathcal{H}$

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

with

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

We then have:

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

with

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

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

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

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

Therefore,

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

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

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

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

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

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

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

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

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

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 computations as in the proof of Lemma 1 of [58, 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 $\|\lambda\|_\infty$. □

The lemma leads to the result.

10.2 Additional file 2 — Kolmogorov-Smirnov tests, plug-in and sub-sampling

Proof of Theorem 1 Let us start with the following simple lemma.

Lemma 2. *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*

$$\sqrt{p} \sup_x |\hat{F}(x) - F(x)| \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{30}$$

Then

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

Proof. Let W be a variable whose distribution is $\mathcal{K}$ and 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 $W_p \xrightarrow[p\to\infty]{\mathcal{L}} W$. Now, it is sufficient to show that for any uniformly continuous bounded function f, $\mathbb{E}[f(Z_p)] - \mathbb{E}[f(W)]|$ tends to 0. But, for any $\delta > 0$,

$$\begin{aligned} |\,\mathbb{E}[f(Z_p)] - \mathbb{E}[f(W)]| &\leq |\,\mathbb{E}[f(Z_p) - f(W_p)]| + |\,\mathbb{E}[f(W_p)] - \mathbb{E}[f(W)]| \\ &\leq 2\|f\|_\infty \mathbb{P}(|Z_p - W_p| > \delta) + \mathbb{E}[|f(Z_p) - f(W_p)| 1_{|Z_p - W_p| \leq \delta}] + |\,\mathbb{E}[f(W_p)] - \mathbb{E}[f(W)]|. \end{aligned}$$

Since $|Z_p - W_p| \leq \sqrt{p} \sup_x |\hat{F}(x) - F(x)| \xrightarrow[p\to\infty]{\mathbb{P}} 0$ and f is uniformly continuous, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\limsup_{p\to+\infty} |\,\mathbb{E}[f(Z_p)] - \mathbb{E}[f(W)]| \leq \varepsilon,$$

which shows the result. □

To prove Theorem 1, it is sufficient to apply Lemma 2 with $p = m(n)$. To do so, we need to prove (30) with $\hat{F}(x) = (1 - e^{-\hat{\lambda}x}) 1_{x>0}$. But for all $x > 0$,

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

Therefore

$$m(n)\sup_x |\hat{F}(x) - F(x)| \leq \frac{m(n)|\hat{\lambda} - \lambda|}{\min(\lambda, \hat{\lambda})}.$$

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 $m(n)|\hat{\lambda} - \lambda|$ tends in probability to 0 whereas $\min(\lambda, \hat{\lambda})$ tends to λ. Therefore (30) is satisfied.

Proof of Theorem 3 When we are dealing with Poisson processes, or more general counting processes, one needs to be a bit more careful with this asymptotic approach because the total number of points is random. Let us precise this idea. If one observes a Poisson process N^a, aggregated over p trials, with constant intensity, then as we have stated in the section on homogeneous Poisson processes, conditionnally to the total number of points, i.e. conditionnally to the event $\{N^a([0, T_{max}]) = n_{tot}\}$, the repartition of the points is uniform. So the test of uniformity is exactly of level α. Indeed, it consists in the following steps.

1. Compute $F_{N^a([0,T_{max}])}(t) = \frac{1}{N^a([0,T_{max}])} \sum_{T \in N^a} \mathbf{1}_{\{(T/T_{max}) \leq t\}}$.
2. Compute $\sup_{t \in [0,1]} |F_{N^a([0,T_{max}])}(t) - t|$.
3. Reject when this last quantity exceeds the random quantity $k_{N^a([0,T_{max}]),1-\alpha}$, where $k_{n_{tot},1-\alpha}$ is the exact and non asymptotic quantile of KS, on the event $\{N^a([0, T_{max}]) = n_{tot}\}$.

Therefore, one can easily state that under H_0: " The process is a homogeneous Poisson process"

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

Now if we want to turn this argument into an asymptotic argument and use $\sqrt{N^a([0, T_{max}]}\, \tilde{k}_{1-\alpha}$ instead of $k_{N^a([0,T_{max}],1-\alpha}$, one needs to be a bit more careful. Actually one can prove the following lemma, which shows that the previous replacement leads indeed to a test of asymptotic level α.

Lemma 3. *If the processes are homogeneous Poisson processes, then*

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

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

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

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

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

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

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

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

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

which proves the convergence in law. □

Now if N is a Poisson process on $[0,T_{max}]$ with intensity $\lambda(.)$, then the compensator of N is given by $\Lambda(t) = \int_0^t \lambda(u)du$, a deterministic non decreasing continuous function. The main property used to build the K.S. test is that $\mathcal{N} = \{\Lambda(T), T \in N\}$ is a homogeneous Poisson process with intensity 1. 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 $\mathcal{N}^a$. We can apply the previous lemma and we have:

$$\sqrt{\mathcal{N}^a([0,\Lambda(T_{max})])} \sup_{t\in[0,1]} |F_{\mathcal{N}^a([0,\Lambda(T_{max})])}(t) - t| \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{K}. \tag{31}$$

But, using the original aggregated process N^a, one can also write

$$\begin{aligned}
F_{\mathcal{N}^a([0,\Lambda(T_{max})])}(t) &= \frac{1}{\mathcal{N}^a([0,\Lambda(T_{max})])} \sum_{T\in\mathcal{N}^a} \mathbf{1}_{\{T/\Lambda(T_{max})\leq t\}} \\
&= \frac{1}{N^a([0,T_{max}])} \sum_{X\in N^a} \mathbf{1}_{\{\Lambda(X)/\Lambda(T_{max})\leq t\}}.
\end{aligned}$$

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

$$\sqrt{N^a([0,T_{max}])} \sup_{x\in[0,T_{max}]} \left| \frac{1}{N^a([0,T_{max}])} \sum_{X\in N^a} \mathbf{1}_{\{X\leq x\}} - \frac{\Lambda(x)}{\Lambda(T_{max})} \right| \xrightarrow[p\to\infty]{\mathcal{L}} \mathcal{K}. \tag{32}$$

We state the following lemma whose proof is similar to the proof of Lemma 2.

Lemma 4. *Let $N^{(1)}, \ldots, N^{(p)}$ be p i.i.d. Poisson processes with compensator Λ, assumed to be continuous, on $[0, T_{max}]$. Let F_p be the associated empirical distribution, defined by*

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

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

$$\sqrt{N^{a,p}([0, T_{max}])} \sup_x |\hat{F}(x) - F(x)| \xrightarrow[p \to \infty]{\mathbb{P}} 0. \tag{33}$$

Then

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

Once again, let us subsample and take $p = m(n)$, with $m(n)/n \to 0$. There are at least two possible choices for $\hat{F}$ that we can describe now.

1. One can first use a global estimate of F by the empirical c.d.f. obtained by aggregating over all the n i.i.d. processes. So we distinguish between $N^{a,m(n)}$, the process aggregated over the $m(n)$ selected trials and N^a, the one obtained over all the n trials. If we use $\hat{F} \equiv F_n$ `Revoir notation`, then, since

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

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

   $$N^{a,m(n)}([0, T_{max}])/N^a([0, T_{max}]) \xrightarrow[p \to \infty]{\mathbb{P}} 0. \tag{34}$$

   But since the numerator is equivalent to $m(n)\Lambda(T_{max})$ and the denominator to $n\Lambda(T_{max})$, (34) is obvious.

2. Another possibility is to use one of the various estimates $\hat{\lambda}$ and to use

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

   Then for all t one can write that

   $$\begin{aligned}
   \hat{F}(t) - F(t) &= \frac{\int_0^t \hat{\lambda}(u)du}{\int_0^{T_{max}} \hat{\lambda}(u)du} - \frac{\int_0^t \lambda(u)du}{\int_0^{T_{max}} \lambda(u)du} \\
   &= \frac{\int_0^t \hat{\lambda}(u)du}{\int_0^{T_{max}} \hat{\lambda}(u)du} - \frac{\int_0^t \lambda(u)du}{\int_0^{T_{max}} \hat{\lambda}(u)du} + \frac{\int_0^t \lambda(u)du}{\int_0^{T_{max}} \hat{\lambda}(u)du} - \frac{\int_0^t \lambda(u)du}{\int_0^{T_{max}} \lambda(u)du} \\
   &= \frac{\int_0^t [\hat{\lambda}(u) - \lambda(u)]du}{\int_0^{T_{max}} \hat{\lambda}(u)du} + \frac{\int_0^t \lambda(u)du}{\int_0^{T_{max}} \lambda(u)du} \frac{\int_0^{T_{max}} [\lambda(u) - \hat{\lambda}(u)]du}{\int_0^{T_{max}} \hat{\lambda}(u)du}.
   \end{aligned}$$

Therefore

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

But for all the given estimates, it is usual to have $\mathbb{E}(\int_0^{T_{max}}(\hat{\lambda}(u) - \lambda(u))^2 du)$ of the order $n^{-\delta}$ for some $\delta > 0$. The same argument as before will then give the desired convergence in probability, as soon as $m(n)n^{-\delta}$ tends to 0.

For instance, the Goldenshluger and Lepski's estimate $\hat{\lambda}_n^{GL}$ has a convergence rate of the order $n^{-\frac{2\alpha}{2\alpha+1}}$ if the intensity is in S_α, the Sobolev space of regularity α. Therefore one needs $m(n) = o(n^{-\frac{2\alpha}{2\alpha+1}})$. If $\alpha > 1/2$, the choice $m(n) = \sqrt{n}$ is convenient. This easily leads to Theorem 3.

General counting processes To each finite point process on $\mathbb{R}^+$ N, we associate, by a one-to-one map, a counting process $(N_t)_{t\geq 0}$ by $N_t = N([0,t])$. Therefore $(N_t)_{t\geq 0}$ is a non decreasing piecewise constant function with jump equal to 1. The positions of the jumps are exactly the positions of the points of N. Let $(N_t)_{t\geq 0}$ be a general counting process on $[0, T_{max}]$ with conditional intensity $\lambda(.)$. To emphasize that this intensity depends on the past, we write $\lambda(t, \mathcal{F}_{t-})$. We denote by $\Lambda(t, \mathcal{F}_{t-}) = \int_0^t \lambda(u, \mathcal{F}_{u-})du$ its compensator. The classical property of point processes [1, 17] states that $\mathcal{N} = \{X = \Lambda(T, \mathcal{F}_{T-}),\ T \in N\}$ is an homogeneous Poisson process with intensity 1 up to (`c'est until non ?`) the time $\Lambda(T_{max}, \mathcal{F}_{T_{max}-})$ which is a random predictable time and therefore a stopping time. We want to use this fact to check on an i.i.d. sample of counting processes, that they have indeed an intensity given by $\lambda(.)$. Therefore we observe p i.i.d. point processes $N^{(i)}$ with intensity $\lambda(t, \mathcal{F}_{t-}^{(i)})$ and compensator $\Lambda(t, \mathcal{F}_{t-}^{(i)})$. In particular, for all t, the $\Lambda(t, \mathcal{F}_{t-}^{(i)})$ 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 $[0, X_{max}^{(i)}]$ with the random stopping time $X_{max}^{(i)} = \Lambda(T_{max}, \mathcal{F}_{T_{max}-}^{(i)})$. Let $(\mathcal{N}_x^{(i)})_{x\geq 0}$ be the corresponding counting process. Since they are not defined on the same interval, there is no sense in aggregating them. But one can cumulate the counting processes in the following way: for any $x \leq \sum_{i=1}^p X_{max}^{(i)}$, we set:

$$\mathcal{N}_x^c = \sum_{i=1}^{k_x} \mathcal{N}_{X_{max}^{(i)}}^{(i)} + \mathcal{N}_{t-\sum_{i=1}^{k_x} X_{max}^{(i)}}^{(k_x)}, \tag{35}$$

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

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

Because each $X_{max}^{(i)}$ is a stopping time, due to the strong Markov property of Poisson processes, one can concatenate them as explained above and we still obtain that the jumps of $\mathcal{N}^c$ form an homogeneous Poisson

process of intensity 1 on $[0, \sum_{i=1}^{p} X_{max}^{(i)}]$. Let us fix some $\theta > 0$ such that $\mathbb{E}[\Lambda(T_{max}, \mathcal{F}_{T_{max}-})] > \theta$. For instance, for Hawkes processes with positive interaction functions, $\theta = \mu T_{max}$ is a convenient choice. One can prove the following result.

Lemma 5. *For all $\theta > 0$ such that $\mathbb{E}(\Lambda(T_{max}, \mathcal{F}_{T_{max}-})) > \theta$,*

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

Proof. Using the concatenation scheme described in (35), let us complete $\mathcal{N}^c$ with another independent homogeneous Poisson process with intensity 1 and infinite support beyond $\sum_{i=1}^{p} X_{max}^{(i)}$, hence obtaining $\mathcal{N}'$ an homogeneous Poisson process of intensity 1 on $\mathbb{R}_+$. Let us denote

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

We define Z'_p with the same expression except that $\mathcal{N}^c$ is replaced by $\mathcal{N}'$. Lemma 3 shows 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 $\mathcal{N}^c \cap [0, p\theta] = \mathcal{N}' \cap [0, p\theta]$ on the event $\{\sum_{i=1}^{p} X_{max}^{(i)} > p\theta\}$. Therefore $|\mathbb{E}(f(Z_p)) - \mathbb{E}(f(Z'_p))| \le 2\|f\|_\infty \mathbb{P}(\sum_{i=1}^{p} X_{max}^{(i)} \le p\theta)$. But by the law of large numbers,

$$\frac{1}{p}\sum_{i=1}^{p} X_{max}^{(i)} \xrightarrow[p\to\infty]{\mathbb{P}} \mathbb{E}(\Lambda(T_{max}, \mathcal{F}_{T_{max}-})) > \theta.$$

Hence $\mathbb{P}(\sum_{i=1}^{p} X_{max}^{(i)} \le p\theta)$ tends to 0, which concludes the proof. □

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

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

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

$$\forall t \ge 0, \quad \Lambda^c(t) = \sum_{i=1}^{j_t} \Lambda(T_{max}, \mathcal{F}_{T_{max}-}^{(i)}) + \Lambda(t - j_t T_{max}, \mathcal{F}_{(t-j_t T_{max})-}^{(j_t)}).$$

The function $\Lambda^c(.)$ is a continuous non decreasing function and therefore, one can consider its generalized inverse function $(\Lambda^c)^{-1}$. Therefore one can rewrite Lemma 5 as follows:

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

Next we want to replace Λ^c by an estimate of the type

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

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

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

with $\hat{\mathcal{N}}^c$ the cumulated process obtained with $\hat{\mathcal{N}}^{(i)} = \{X = \int_0^T \hat{\lambda}(u, \mathcal{F}_{u-}^{(i)}) du, T \in N^{(i)}\}$ instead of $\mathcal{N}^{(i)}$ in (35). One can prove the following Lemma.

Lemma 6. *Assume that*

$$p^{-1/2} \left(\sum_{i=1}^{p} \int_0^{T_{max}} \left| \hat{\lambda}(u, \mathcal{F}_{u-}^{(i)}) - \lambda(u, \mathcal{F}_{u-}^{(i)}) \right| du \right) \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{36}$$

Then, for all $\theta > 0$ such that $\mathbb{E}(\Lambda(T_{max}, \mathcal{F}_{T_{max}-})) > \theta$,

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

Proof. First of all, $\Lambda^c(pT_{max})/p = \sum_i X_{max}^{(i)}/p$ tends in probability to $\mathbb{E}(\Lambda(T_{max}, \mathcal{F}_{T_{max}-})) > \theta$. Hence, with probability tending to 1, $p\theta$ belongs to $[0, \Lambda^c(pT_{max})]$. Moreover,

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

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

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

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

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

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

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

and

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

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

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

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

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

If we are able to show that

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

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

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

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

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

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

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

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

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

We denote:

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

But Z_p is also equal with probability tending to 1 to

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

because $p\theta$ belongs to $[0,\Lambda^c(pT_{max})]$ with probability tending to 1. Using Lemma 5, we have that Z_p tends in law 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:

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

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

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

Therefore, by using (36), $\tilde{Z}_p-Z_p\xrightarrow[p\to\infty]{\mathbb{P}}0$. □

Now let us verify on various examples that the $\hat{\lambda}$ given in Theorems 4 and 6 satisfies (36) by taking $p=m(n)$.

- If the processes are Poisson processes, then (36) can be rewritten as

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

This is implied by $\mathbb{E}\int_0^{T_{max}} |\hat{\lambda}(u) - \lambda(u)|^2 du = o(m(n)^{-1})$, which is true under the assumptions of Theorem 4. So, Theorem 4 is proved.

- If the processes are Hawkes processes with positive piecewise constant interaction functions on the partition of length A and bin length δ, then (36) can be rewritten for the process N_m as

$$\frac{1}{\sqrt{p}}\sum_{i=1}^{p}\int_0^{T_{max}} |\mathbf{Rc}_t^{(i)}(\hat{\mathbf{a}}^{(m)} - \mathbf{a}_*^{(m)})|dt \xrightarrow[p\to\infty]{\mathbb{P}} 0,$$

which is implied by

$$\frac{\|\hat{\mathbf{a}}^{(m)} - \mathbf{a}_*^{(m)}\|}{\sqrt{p}}\sum_{i=1}^{p}\int_0^{T_{max}} \|\mathbf{Rc}_t^{(i)}\|dt \xrightarrow[p\to\infty]{\mathbb{P}} 0.$$

But the quantities $\int_0^{T_{max}} \|\mathbf{Rc}_t^{(i)}\|dt$ are i.i.d. with finite mean, so by the law of large numbers

$$p^{-1}\sum_{i=1}^{p}\int_0^{T_{max}} \|\mathbf{Rc}_t^{(i)}\|dt,$$

has a finite limit in probability and it is sufficient to have

$$p\|\hat{\mathbf{a}}^{(m)} - \mathbf{a}_*^{(m)}\|^2 = \xrightarrow[p\to\infty]{\mathbb{P}} 0,$$

to conclude the proof. This is an easy consequence of Theorem 5.

10.3 Additional file 3 — Proof of Theorem 5

The proof of Theorem 5, that can be skipped at first reading, uses notations of [25] and transpositions of these notations. 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

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

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

$$\mathcal{H}_n = (\mathbb{R}\times\mathbb{L}_2([0,1])^{nM})^{nM} = \Big\{ f_n = \Big((\mu^{(m,i)}, (g_{\ell,j}^{(m,i)})_{\ell=1,\dots,M,\ j=1,\dots,n})_{m=1,\dots,M,\ i=1,\dots n}\Big) : \\ g_{\ell,j}^{(m,i)} \text{ with support in } (0,1] \text{ and } \|f_n\|^2 = \sum_{m,i}(\mu^{(m,i)})^2 + \sum_{m,i}\sum_{\ell,j}\int_0^1 g_{\ell,j}^{(m,i)}(t)^2 \mathrm{d}t < \infty \Big\}.$$

To reduce the complexity, it is sufficient to take a dictionary Φ, which forces $g_{\ell,j}^{(m,i)} = 0$ if $i \neq j$, $\mu^{(m,i)} = \mu^{(m)}$ and $g_{\ell,i}^{(m,i)} = g_{\ell}^{(m)}$. More precisely, we take for $m = 1, ..., M$, $\varphi_{m,0}$ such that

$$\forall i = 1, ..., n, \quad \mu_{m,0}^{(m,i)} = 1,$$

all the other components being zero. We also take for all $m, \ell = 1, \ldots, M$, for all $k = 1, \ldots, K$, $\varphi_{m,\ell,k}$ such that

$$\forall i = 1, ..., n, \quad (g_{m,\ell,k})_{\ell,i}^{(m,i)} = \delta^{-1/2} \mathbf{1}_{((k-1)\delta, k\delta]},$$

all the other components being zero. This dictionary is of size $M(1 + MK)$. To each set of coefficients a in $\mathbb{R}^{\Phi}$, one can write

$$a' = \left((\mathbf{a}^{(1)})', ..., (\mathbf{a}^{(M)})' \right),$$

and associate to the $\mathbf{a}^{(m)}$'s an element f of $\mathcal{H}$ as it has been done in (22) and (23). With this choice, it is then quite easy to see that the definition of the least square contrast $\gamma(f_n)$ of [25] coincides with the definition (26) of $\gamma_n(f)$. In the same way, the definition of b of [25], is given by

$$b' = \sum_{i=1}^{n} \left((\mathbf{b}^{(1,i)})', ..., (\mathbf{b}^{(M,i)})' \right),$$

and the definition of G is given by a matrix diagonal by blocks of M blocks

$$G = \sum_{i=1}^{n} diag(\mathbf{G}^{(i)}, ..., \mathbf{G}^{(i)}).$$

The choice of d of [25, Theorem 2] can be written as follows: for all $\mu, \varepsilon, x > 0$

$$d_{\varphi_{m,0}} = \sqrt{2(1+\varepsilon) \left[\frac{\mu}{\mu - \phi(\mu)} N_m^a([a,b]) + \frac{x}{\mu - \phi(\mu)} \right]} + \frac{x}{3},$$

and

$$d_{\varphi_{m,\ell,k}} = \sqrt{2(1+\varepsilon) \left[\frac{\mu}{\mu - \phi(\mu)} \hat{V}_{m,\ell,k} + \frac{B_{\ell,k}^2 x}{\mu - \phi(\mu)} \right]} + \frac{B_{\ell,k} x}{3},$$

where $\phi(u) = e^u - u - 1$ and where $B_{\ell,k}$ is a potential deterministic bound on the $\hat{B}_{\ell,k}$. Then Theorem 2 of [25] states that the selected estimate that is

$$\hat{f}_n = \sum_{\varphi \in \Phi} \tilde{a}_{\varphi} \varphi,$$

satisfies an oracle inequality on the event

$$\Omega_{V,B} = \{ \forall m, \ell, k, \hat{V}_{m,\ell,k} \leq V_{m,\ell,k} \text{ and } N_m^a([a,b]) \leq \mathcal{N}_m \text{ and } \hat{B}_{\ell,k} \leq B_{\ell,k} \}$$

intersected with the event $\Omega_c = \{\forall a, a'Ga \geq cI\}$, where $c, V_{m,\ell,k}, \mathcal{N}_m$ and $B_{\ell,k}$ are fixed positive constants. More precisely with probability larger than

$$1 - \left[\sum_{m,\ell,k} \frac{\log\left(1 + \frac{\mu V_{m,\ell,k}}{B_{\ell,k}^2 x}\right)}{\log(1+\varepsilon)} + \sum_m \frac{\log\left(1 + \frac{\mu \mathcal{N}_m}{x}\right)}{\log(1+\varepsilon)} \right] e^{-x} - \mathbb{P}(\Omega_c^c) - \mathbb{P}(\Omega_{V,B}^c),$$

we have that there exists an absolute positive constant C such that

$$\sum_{i=1}^n \sum_{m=1}^M \|\psi_{\hat{f}_n}^{(m,i)} - \psi_{f_n^*}^{(m,i)}\|^2 \leq C \inf_{a \in \mathbb{R}^\Phi} \left[\sum_{i=1}^n \sum_{m=1}^M \|\psi_{\sum_{\varphi \in \Phi} a_\phi \varphi}^{(m,i)} - \psi_{f_n}^{(m,i)}\|^2 + c^{-1} \sum_{\varphi \in S(a)} d_\varphi^2 \right].$$

Because the number of points of a Hawkes process in a finite interval has some finite exponential moment, we can easily see that one can choose

$$\mathcal{N}_m = \square_{f^*,a,b} n \log(M).$$

From now on, $\square_\theta$ designs a positive constant only depending on the index θ whose value may change from line to line. With this choice

$$\mathbb{P}(\exists m, N_m^a([a,b]) > \mathcal{N}_m) = O(1/n).$$

Also, one can see that

$$\mathbb{P}(\exists m, i, N_m^{(i)}([a-A,b]) > \square_{f^*,T_{max}} \log(nM)) = O(1/n).$$

Since,

$$\hat{B}_{\ell,k} \leq \delta^{-1/2} N_m^{(i)}([a-A,b]),$$

one can also take

$$B_{\ell,k} = \square_{f^*,a,b} K^{1/2} \log(nM).$$

Finally, since

$$\hat{V}_{m,\ell,k} \leq \delta^{-1} \sum_{i=1}^n N_m^{(i)}([a,b])[N_\ell^{(i)}([a-A,b])]^2,$$

one can take

$$V_{m,\ell,k} = \square_{f^*,a,b} K n \log(nM)^3.$$

Therefore with this choice, $\mathbb{P}(\Omega_{V,B}^c) = O(1/n)$. Hence in this case the probability of the oracle inequality is larger than

$$1 - \square_{M,f^*,a,b,\epsilon,\mu} K \log(n/x) e^{-x} - O(1/n) - \mathbb{P}(\Omega_c^c).$$

Therefore we can choose $K = n^\xi$ and $x = \eta \log(n)$, with $\eta > \xi$ and the oracle inequality holds with probability larger $1 - O(1/n) - \mathbb{P}(\Omega_c^c)$. Note that with this choice of x and up to the replacement of $B_{\ell,k}$ by $\hat{B}_{\ell,k}$, for

a given parameter $\gamma > \square_{a,b}$, it is always possible to find η, μ, ε such that the present weights $d_{\varphi_{m,\ell,k}}$'s are smaller than the $\mathbf{d}_{\ell,k}^{(m)}$ of (29). Moreover if γ tends to $+\infty$, it is possible to choose η tending to $+\infty$ too. In this sense, the present result also applies to the procedure described in the main part of the article for a large enough parameter γ. It remains to control Ω_c. This can be done in a non-asymptotic way following [25]. But in the present context, it is quite easy to see that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{G}^{(i)} \xrightarrow[p\to\infty]{\mathbb{P}} \mathbb{E}(\mathbf{G}^{(1)}).$$

But $\mathbf{G}^{(1)}$ being a random non negative matrix, which is not almost surely null by assumption on the $\nu^{(m)}$'s, its expectation has only strictly positive eigenvalues. Therefore if $r > 0$ is strictly smaller than the smallest eigenvalue of $\mathbb{E}(\mathbf{G}^{(1)})$, then the probability of Ω_c^c with $c = nr$, tends to 0. Since under the assumptions of Theorem 5, $f_n^* = \sum_{\varphi\in\Phi} a_\phi^*\varphi$, we obtain that with probability tending to 1,

$$\sum_{i=1}^{n}\sum_{m=1}^{M}\|\psi_{\hat{f}_n}^{(m,i)} - \psi_{f_n^*}^{(m,i)}\|^2 \leq \square_{M,f^*,a,b,K}\log(n)^3.$$

But because, this happens on Ω_c, one has that

$$\sum_{i=1}^{n}\sum_{m=1}^{M}\|\psi_{\hat{f}_n}^{(m,i)} - \psi_{f_n^*}^{(m,i)}\|^2 = (\tilde{a}-a^*)'G(\tilde{a}-a^*) \geq \square_{M,f^*,a,b,K} n\sum_{\varphi\in\Phi}[\tilde{a}_\varphi - a_\varphi^*]^2.$$

This concludes the proof of Theorem 5.

10.4 Additional file 4 — The False Discovery Rate and the Benjamini and Hochberg multiple test procedure

In most of the Tables in this article, several tests have been performed. Let us introduce some notations: a test of hypothesis H_0^i is associated to a binary random variable Δ_i, with the convention $\Delta_i = 1$ if the test rejects H_0^i and $\Delta_i = 0$ otherwise. Hence, if the K tests are performed asymptotically at level α, we are considering K random variables Δ_i such that $\mathbb{P}_{H_0^i}(\Delta_i = 1) \underset{n\to+\infty}{\to} 1-\alpha$. Thus, the number of rejections is not controlled. Indeed, assuming that those tests are independent, then if all the H_0^i are true, the probability to have no rejection is:

$$\mathbb{P}(\forall i \in \{1,\ldots,K\}, \Delta_i \text{ accepts}) \to_{n\to\infty} (1-\alpha)^K \to_{K\to\infty} 0.$$

This means that when K grows, this procedure is doomed to reject at least 1 and in average $K\alpha$ tests.

One way to control such a problem is to control the so called *Family Wise Error Rate* (FWER) [59], which consists in controlling $FWER = \mathbb{P}(\exists i \in \{1,\ldots,K\}, \Delta_i \text{ wrongly rejects})$. This can be easily done by

Bonferroni bounds:

$$\mathbb{P}(\exists i \in \{1,\dots,K\}, \Delta_i \text{ wrongly rejects}) \leq \sum_{i\in\{1,\dots,K\}} \mathbb{P}(\Delta_i \text{ wrongly rejects}) \to_{n\to\infty} K\alpha.$$

So Bonferroni method [60] consists in applying the Δ_i tests at level α/K instead of α to guarantee a FWER less than α. However, the smaller the type I error, the more difficult it is to reject. So when K is large, it potentially leads to no rejection at all, even in cases where the H_0^i are false.

Another notion, popularized by [35], has consequently been introduced in the multiple testing community leading to a large amount of publications in statistics, genomics, medicine etc in the past ten years. This is the *False Discovery Rate* (FDR). Actually, a false discovery (also named false positive) is not that bad if the ratio of the number of false discoveries divided by the total number of discoveries is small.
More formally, let us use the notations given in Table 1.

Number of i such that	Δ_i accepts	Δ_i rejects	Total
true H_0^i	U = " true negative"	V = "false positive"	K_0 = "number of true H_0^i"
false H_0^i	T = "false negative"	S = "true positive"	K_1 = "number of false H_0^i"
Total	$K-R$	R= "discoveries"	K

Table 1: Repartition of the answers in a set of K tests.

Then the *False Discovery Rate* is defined by

$$FDR = \mathbb{E}\left(\frac{V}{R}\mathbf{1}_{R>0}\right).$$

We always have that $FDR \leq FWER$. This means that when the H_0^i's do not hold, controlling the FDR is less stringent, whereas the relative confidence that we can have in the rejections is still good: if we make 100 rejections with a FDR of 5%, this means that on average only 5 of these rejections will be potentially wrong.

The question now is: how to guarantee a small FDR? To do so, Benjamini and Hochberg [35] proposed

the following procedure: for each test Δ_i, the corresponding p-value P_i is computed. They are next ordered such that:

$$P_{(1)} \leq .... \leq P_{(m)} \leq P_{(K)}.$$

Let $q \in [0, 1]$ be a fixed upper bound that we desire on the FDR and define:

$$k = \max\{m \text{ such that } P_{(m)} \leq mq/K\}.$$

Then the rejections of this BH-method are given by the tests $\Delta_{(1)}, ..., \Delta_{(k)}$ corresponding to the k smallest p-values. It can be proved under various conditions (for instance independence) that such procedure guarantees a FDR less than q (see [61]). Note that even if these conditions are not fulfilled, the FDR of this procedure is still controlled by q up to a logarithmic factor [61].