

UNIVERSITÉ DE NICE SOPHIA-ANTIPOLIS- UFR Sciences

École Doctorale Sciences Fondamentales et Appliquées

## Habilitation à diriger les recherches

Spécialité : Mathématiques

présentée par  
**Patricia Reynaud-Bouret**

### **Adaptive statistical inference for some point processes (Poisson, Aalen, Hawkes)**

soutenue le 18 Février 2011

Après avis de :

M.	Evarist	GINÉ-MASDÉU
Mme.	Dominique	PICARD
M.	Vladimir	SPOKOINY

Devant le Jury composé de :

M.	Yannick	BARAUD	Examineur
M.	Philippe	BERTHET	Examineur
M.	Oleg	LEPSKI	Examineur
M.	Pascal	MASSART	Examineur
M.	Laurent	MASSOULIÉ	Examineur
Mme.	Dominique	PICARD	Rapporteur
M.	Stéphane	ROBIN	Examineur

Laboratoire Jean-Alexandre Dieudonné  
Université de Nice Sophia-Antipolis, Parc Valrose  
06108 NICE Cedex 2



## *Remerciements*

Je ne peux que commencer ces remerciements par un énorme "MERCI" à Pascal Mas-sart, qui fut mon directeur de thèse et qui a accepté de faire partie de mon jury. Merci pour tes conseils toujours bienveillants. Grâce à toi, les statistiques resteront toujours pour moi cet enchevêtrement si intéressant de belles théories, de mains dans le camboui et de petits "chouilles" qu'on n'explique pas encore et qui nous sépare de la vraie vie. Et tout est dans le petit "chouille". Bien sûr, je dis ça en m'excusant mille fois de reprendre le mot du maître... Oups, pardon, désolée, je retourne dans mon col roulé...

Je tiens ensuite à remercier chaleureusement Evarist Giné, Dominique Picard et Vladimir Spokoiny qui ont gentiment accepté de rapporter mon habilitation. Merci aussi à Dominique d'avoir fait le déplacement et d'être présente aujourd'hui. Je voudrais aussi exprimer ma gratitude à Philippe Berthet, Oleg Lepski, Laurent Massoulié et Stéphane Robin, qui ont accepté de faire partie de mon jury. Yannick Baraud tient dans cette liste une place à part. Nous avons partagé le même bureau à l'ENS, puis tu nous as accueillis, ma famille et moi, à Nice. Tu tiens pour moi le rôle de "grand frère" que je suis pas à pas. Je voudrais aussi mentionner certaines personnes, qui n'ont pas pu se déplacer mais qui ont durablement influencé mon travail. Je pense à Lucien Birgé et Anestis Antoniadis, mais aussi à David Vere-Jones, qui bien qu'il ne me connaissait pas, a gentiment relu officieusement mon manuscrit.

L'autre personne qui a profondément changé ma façon de faire les statistiques est Sophie Schbath. Sans toi (et les processus de Hawkes) je n'en serais probablement pas là. Merci de m'avoir fait part de ton modèle, de ta rigueur et de ta vision des applications ... Et puis bien sûr après une si longue collaboration, où nous n'avons pas encore eu l'occasion de sabrer le champagne, je te suis surtout reconnaissante de m'avoir appris le concept de "tache propre" (sans accent) qui a changé ma vie de maman.

Il paraît que nous pensons tellement de la même façon que parler à l'un revient à parler à l'autre. Je veux parler de Vincent Rivoirard bien sûr. Même si je ne pense pas que cette opinion soit complètement justifiée, elle montre bien à quel point tu as aussi influencé ma manière de faire des statistiques. Il est vrai que nous avons tendance à sauter ensemble du coq à l'âne en perdant la plupart des gens. C'est notre chère étudiante, Laure Sansonnet, qui a bien du courage de supporter nos errements sans se laisser décourager. Merci Laure de ton sang-froid !

Ensuite, il y a les filles, Magalie Fromont et Béatrice Laurent. Votre honnêteté et votre manière d'être dans la vie, qu'elle soit scientifique ou non, ont beaucoup déteint sur moi. Ma liste de coauteurs est longue. Vous m'avez tous appris quelque chose. Merci à Christian Houdré et Philippe Marchal, à Emmanuel Roy, à Marie Doumic et Marc Hoffmann, à Niels R. Hansen, à Gérard Biau. Une petite mention spéciale à Christine Tuleau-Malot. Tu ne sais pas à quel point tu es quelqu'un de précieux !

Une autre aventure qui a commencé à l'ENS-Paris et qui vient de s'achever est le projet ANR ATLAS. Un grand merci à toute l'équipe, qu'elle soit scientifique ou administrative (Zaïna et Lara). Merci surtout à Gilles Stoltz pour cette direction bicéphale qui doit beaucoup à tes énormes capacités d'organisation.

Je tiens à remercier toutes les personnes que j'ai cotoyées toutes ces années que ce soit à Orsay, Georgia Tech, l'ENS-Paris ou à Nice. Je préfère ne citer personne de peur d'en oublier...Vous vous reconnaîtrez dans ces quelques lignes.

Je voudrais finalement dédier ces quelques pages à mes petits schtroumpfs, qui se demandent bien ce que je fabrique la plupart du temps, et à mon mari si patient, qui, lui, comprend...



# Contents

<b>1</b>	<b>Introduction</b>	<b>8</b>
1.1	Statistical frameworks . . . . .	8
1.1.1	Poisson process . . . . .	8
1.1.2	Counting process . . . . .	9
1.2	Train of thought . . . . .	10
<b>2</b>	<b>Model Selection</b>	<b>12</b>
2.1	Poisson framework . . . . .	12
2.1.1	Histograms . . . . .	14
2.1.2	Concentration inequalities . . . . .	16
2.2	Other counting processes (Aalen, Hawkes) . . . . .	17
2.2.1	Exponential inequalities for counting processes . . . . .	19
2.2.2	Oracle inequalities . . . . .	20
<b>3</b>	<b>Support Assumptions</b>	<b>25</b>
3.1	Hawkes framework . . . . .	25
3.2	Poisson framework . . . . .	26
3.2.1	The method and the oracle inequality . . . . .	28
3.2.2	The curse of support . . . . .	31
<b>4</b>	<b>Target functions</b>	<b>36</b>
4.1	Estimation, Maxisets and Minimax point of view . . . . .	36
4.2	Tests . . . . .	38
<b>5</b>	<b>Calibration</b>	<b>43</b>
5.1	Ad hoc methods . . . . .	43
5.1.1	Compared methods . . . . .	44
5.1.2	Results . . . . .	46
5.2	Data-driven thresholds . . . . .	47
<b>A</b>	<b>Presented papers</b>	<b>51</b>
<b>B</b>	<b>Future work</b>	<b>52</b>



# Chapter 1

## Introduction

The present report aims at giving a survey of my work since my PhD thesis, "Estimation adaptative de l'intensité de certains processus ponctuels par sélection de modèles", which was supervised by P. Massart. Since then, I basically broaden the spectrum of statistical methods and the spectrum of point processes that are of interest for me. I especially encountered a very special process, the Hawkes process, which is at the root of most of the questions I tried to solve, and to even more questions that do not have an answer yet. Those questions are partly statistical and partly probabilistic, and most of the time are not completely solved for the Hawkes process, but partial answers appear for a simpler process : the Poisson process. It is actually the purpose of this report to emphasize the link between statistical and probabilist aspects, but also between the different statistical frameworks and questions that have arisen in my papers.

### 1.1 Statistical frameworks

Let us start with some definitions. A point process,  $N$ , is a countable set of points of some measurable space  $\mathbb{X}$ . Usually,  $N_A$  denotes the number of points of  $N$  in the set  $A$  and  $dN$  denotes the point measure *ie* the sum of the Dirac masses at each point of  $N$  (see [23] for a complete overview).

#### 1.1.1 Poisson process

The simplest point process one can encounter is the Poisson process [43].

**Definition 1.** *A Poisson process,  $N$ , on a measurable space  $\mathbb{X}$  is a random countable set of points such that*

- *for all integer  $n$ , for all  $A_1, \dots, A_n$  disjoint measurable subsets of  $\mathbb{X}$ ,  $N_{A_1}, \dots, N_{A_n}$  are independent random variables.*
- *for all measurable subset  $A$  of  $\mathbb{X}$ ,  $N_A$  obeys a Poisson law with parameter depending on  $A$  and denoted  $\nu(A)$ .*

As a consequence,  $\nu$  defines a measure on  $\mathbb{X}$ , which is called the mean measure of the process  $N$ . Usually,  $\mathbb{X}$  is a subset of  $\mathbb{R}^d$  and  $\nu$  is absolutely continuous with respect to the Lebesgue measure. In this case,  $s(x) = d\nu/dx$  is the intensity of the Poisson process  $N$ .

Poisson processes model a large variety of situations, but the outstanding examples of the present work actually come from genomic data (position of genes on the DNA sequence

for instance) [56], size of individual oil reservoirs in some petroleum field [71], stockprice changes of extraordinary magnitude [52]. All these data have something in common: there are good evidences that the underlying intensity is highly irregular, with very localized spikes with unknown position and shape and, for some of them, a very large and typically unknown support, when this support is finite.

The adaptive statistical inference aims at estimating  $s$  or at testing some hypothesis on  $s$  with as few assumptions on  $s$  (or on the alternatives) as possible. Typically  $s$  belongs to  $\mathbb{L}_p(\mathbb{X})$ , for  $1 \leq p \leq +\infty$ , but we do not want to assume that  $s$  is smooth with the precise known regularity  $\alpha$ . Since the regularity is unknown and since the rate of convergence (or rate of separation) depends on the regularity, we want the procedures to adapt to this unknown regularity and to be as precise as the methods knowing the precise regularity of the target function  $s$ .

Actually, exhibiting such kind of procedures is not just a "pure theoretical game". The procedures that are theoretically adaptive, if they are calibrated (see the final section) may provide practical methods that are performing really well and that are robust to changes in the target function  $s$ . Actually an adaptive method that does not need any information on the regularity or on the support of the target function is easier to use since we do not need to ask the practitioner to provide such information on  $s$  before proceeding to the inference.

### 1.1.2 Counting process

One can generalize the notion of Poisson process on the real (positive) line. If the point process is, say, almost surely finite in finite intervals (*ie*  $N_{[a,b]} < \infty$  a.s.), then one can "count" the points: if the positive real line represents time after 0, then there exist a first point, a second .... The random function  $N_t = N_{[0,t]}$  as a function of  $t \in \mathbb{R}$  is a piecewise constant increasing function such that  $N_0 = 0$  with jumps equal to 1. This is the definition of a *counting process*. The interested reader may find in [13] a complete review of those processes and precise definitions. Under suitable assumptions, one can informally define the (conditional) intensity  $\Lambda(t)$  of the counting process  $(N_t)_{t \geq 0}$  by

$$\mathbb{E}(dN_t | \mathcal{F}_{t-}) = \Lambda(t)dt, \tag{1.1.1}$$

where  $dN_t$  represents the infinitesimal increment of  $N_t$  at time  $t$ ,  $\mathcal{F}_{t-}$  represents the past information of the process (what happened before  $t$ ) and  $dt$  is the Lebesgue measure (see [2, Section II.4.1] for more details). Obviously this means that  $\Lambda(t)$  is random and depends on the past. So one cannot statistically infer  $\Lambda(t)$  without further assumption on the model. Note however that when we apply this definition to Poisson processes, the independence property in Definition 1 implies that  $\Lambda(t)$  cannot depend on the past, it is a deterministic function, which corresponds to the intensity  $s$  of the Poisson process defined before.

The two cases that are studied here are:

1. the Aalen multiplicative intensity, *ie*

$$\Lambda(t) = Y(t)s(t)dt, \tag{1.1.2}$$

where  $Y(t)$  is a predictable process (*ie* informally, it only depends on the past) that is observed and  $s$  is an unknown deterministic function that we want to estimate in an adaptive way. The classical examples covered by this model are right-censoring models, finite state inhomogeneous Markov processes... We refer to [2] for an extensive

list of situations. Let us just mention a classical model. For biomedical applications, a single process may correspond to the time of death or times of depression of one patient. But one usually considers several patients. The aggregated process, i.e. the sum of each individual process, is still a counting process with Aalen multiplicative intensity. In this context,  $Y(t)$  usually represents the number of people at risk at time  $t$ : people still alive and observed at time  $t$  or people not in depression at time  $t$ .

2. the Hawkes processes, which is defined in the most basic self-exciting model, by

$$\Lambda(t) = \nu + \int_{-\infty}^{t^-} h(t-u) dN_u, \quad (1.1.3)$$

where  $\nu$  is a positive parameter,  $h$  a non negative function with support on  $\mathbb{R}^+$  and  $\int h < 1$  and where  $dN_u$  is the point measure associated to the process. Since  $\Lambda(t)$  corresponds to (1.1.1), (1.1.3) basically means that there is a constant rate  $\nu$  to have a spontaneous occurrence at  $t$  but that also all the previous occurrences influence the apparition of an occurrence at  $t$ . For instance an occurrence at  $u$  increases the intensity by  $h(t-u)$ . If the distance  $d = t-u$  is favoured, it means that  $h(d)$  is really large: having an occurrence at  $u$  significantly increases the chance of having an occurrence at  $t$ . The intensity given by (1.1.3) is the most basic case, but variations of it enable us to model self interaction (*ie* also inhibition, which happens when one allows  $h$  to take negative values, see Section 2.2.2) and, in the most general case, to model interaction with another type of event.

Hawkes processes have been widely used to model the occurrences of earthquake [72] but the main focus in this report is on genomic data, where this framework was introduced in [31] to model occurrences of events such as positions of genes, promoter sites or words on the DNA sequence. The drawback is that, by definition, the Hawkes process is defined on an ordered real line (there is a past, a present and a future). But a strand of DNA itself has a direction, fact that makes this approach quite sensible.

Here the unknown quantity is the couple  $s = (\nu, h)$ . In practice very little is known on the function  $h$ , except that it should consist in localized spikes at preferred distances corresponding to biological interactions. The aim of an adaptive procedure is actually to find in practice the localisation of those spikes and their size.

## 1.2 Train of thought

Actually, the Hawkes process is probably the trickiest process of the three (Poisson, Aalen, Hawkes) from an adaptive statistical point of view. It is also a constant source of inspiration. Let us just point out some of the most basic problems that a practitioner/biologist may stress:

1. *We do not know the shape, number and position of the localized spikes.* Basically we need to propose an adaptive estimation of  $h$  at least.
2. *The interaction range cannot be larger than 10,000 bases, since after this length the DNA 3D structure may interfere, but it can be as small as 20 bases or even less.* The procedure has consequently to be adaptive to the support. Is this possible? What does this mean theoretically speaking?

3. *The functions should have not many but very localized spikes.* What does this mean from the point of view of functional analysis? What is actually the set of target functions? Can they really be considered as smooth, ie derivable, with the only unknown quantity being its regularity?
4. And finally, the question all statisticians face when they want to implement their method: *there is one tuning parameter (or more) in the method, how should we choose it?*

Actually, each question asks for a complex development depending on how sharp the answer needs to be. For instance, one cannot provide a good adaptive statistical procedure without having specific probabilistic techniques such as exponential inequalities. So we devote a part of this work to describe probabilistic tools that fit the statistical needs. Even so, the answers are not complete for Hawkes processes, sometimes because the probabilistic part is missing, sometimes because the structure itself prevents to use special adaptive methods. Even if the method we developed for Hawkes process works well, it is not the end of the story. So we tried to answer to those questions in easier set-up (basically Poisson) so that the answers become more accurate.

The report is organised as follows. In Chapter 2, the model selection is developed for the Poisson, Aalen and Hawkes framework, the link with the probabilistic inequalities is emphasized. In Chapter 3, we stress how the question of the support was handled in practice for Hawkes processes but also what this problem means for Poisson processes. In the Poisson framework, the answer is more accurate theoretically speaking. It reveals a surprising behaviour with respect to the adaptivity to the support: there is actually a curse of support in the sense that it can modify the rate of convergence. Next, we develop in Chapter 4 the understanding of "good" target functions from a minimax and maxiset point of view. The answer is again surprising: for such spiky functions it may be as difficult to test as to estimate. Chapter 5 is dedicated to ad hoc calibration methods for Hawkes processes and theoretical calibration for Poisson process. The appendices present a list of the published papers that are described here and a short presentation of what can be done after that.

Finally, let us just introduce a notation which will avoid tedious explanations. We use in the sequel the notation  $\square$  which represents a positive function of the parameters that are written in indices. Each time  $\square_\theta$  is written in some equation, one should understand that there exists a positive function of  $\theta$  such that the equation holds. Therefore the values of  $\square_\theta$  may change from line to line and even change in the same equation. When no index appears,  $\square$  represents a positive absolute constant.

# Chapter 2

## Model Selection

The model selection method in its theoretical adaptive presentation has been introduced and developed by Barron, Birgé and Massart [9]. Massart's course [51] in Saint-Flour is one of the best reference on this topic. In particular, Massart emphasizes how concentration inequalities are the fundamental tool to perform model selection. In the next sections, a brief summary of the model selection method is given for Poisson processes, and the corresponding exponential inequality is given. Then the main difficulties arising for other counting processes are emphasized.

### 2.1 Poisson framework

This section is mainly inspired by [57], but we give a simpler version here. We observe a Poisson process  $N$ , on  $\mathbb{X}$  (say a compact subset of  $\mathbb{R}^d$ ) and we want to estimate its intensity  $s$  with respect to the Lebesgue measure, denoted  $\mu$ . Let  $T$  be the Lebesgue measure of  $\mathbb{X}$ , assumed to be finite.

We work with the  $\mathbb{L}_2$ -norm defined by

$$\|f\|^2 := \frac{1}{T} \int_{\mathbb{X}} f^2(x) dx,$$

and we define the following least-square contrast

$$\gamma(f) := -\frac{2}{T} \int_{\mathbb{X}} f(x) dN_x + \frac{1}{T} \int_{\mathbb{X}} f^2(x) dx. \quad (2.1.1)$$

Note for instance that

$$\mathbb{E}(\gamma(f)) = -\frac{2}{T} \int_{\mathbb{X}} f(x) s(x) dx + \frac{1}{T} \int_{\mathbb{X}} f^2(x) dx = \|f - s\|^2 - \|s\|^2,$$

which is minimal when  $f = s$ . Hence minimizing the least-square contrast should enable us to obtain a "good" estimator.

Other contrasts may be used. Usually people in point processes theory use MLE ([74], [53], [54]) which corresponds to the minimization of

$$-\frac{2}{T} \int_{\mathbb{X}} \ln(f(x)) dN_x + \frac{1}{T} \int_{\mathbb{X}} f(x) dx,$$

but those contrasts are actually more difficult to handle than least-square contrasts for model selection (see [51, Chapter 7] for an extensive comparison of both contrasts in the density setting).

Next the contrast can be minimized on a finite vectorial subspace  $S$  of  $\mathbb{L}_2$  with orthonormal basis given by  $\{\varphi_1, \dots, \varphi_D\}$ . The minimization leads to the projection estimator of  $s$

$$\hat{s} = \sum_{i=1}^D \left( \int_{\mathbb{X}} \varphi_i(x) \frac{dN_x}{T} \right) \varphi_i. \quad (2.1.2)$$

Let us study the risk of  $\hat{s}$ ,  $\mathbb{E}(\|s - \hat{s}\|^2)$ . To do so, let us introduce  $\bar{s}$  the orthonormal projection of  $s$  on  $S$ . This gives

$$\mathbb{E}(\|s - \hat{s}\|^2) = \|s - \bar{s}\|^2 + \frac{1}{T} \sum_{i=1}^D \int \varphi_i^2(x) s(x) dx. \quad (2.1.3)$$

The first term is a bias term, it decreases when  $S$  increases whereas the second term, the variance term, increases with the dimension  $D$  of  $S$ . Obviously finding the best compromise depends on  $s$ .

Hence model selection consists in searching for the "best"  $S$  in a family of models (*ie*, here, finite vectorial subspaces)  $\{S_m, m \in \mathcal{M}_T\}$ . To each model  $S_m$  let us associate the projection estimator  $\hat{s}_m$  and the orthonormal projection of  $s$  on  $S_m$ ,  $s_m$ . In a naive approach, the best model we should use, is of course

$$\bar{m} := \operatorname{argmin}_{m \in \mathcal{M}_T} \mathbb{E}(\|s - \hat{s}_m\|^2).$$

This model,  $\bar{m}$ , is called the oracle. Of course we cannot obtain it without knowing  $s$ . Mallows [49] first noticed in the Gaussian framework that actually one can rewrite the previous formula. One can adapt the computations here to find very easily that

$$\bar{m} = \operatorname{argmin}_{m \in \mathcal{M}_T} \left\{ \mathbb{E}(\gamma(\hat{s}_m)) + 2\mathbb{E}(\|s_m - \hat{s}_m\|^2) \right\}.$$

It is possible to estimate the previous quantity without bias. Let us denote by  $(\varphi_{\lambda,m})_\lambda$  an orthonormal basis of  $S_m$ . One obtains the following choice

$$\hat{m} := \operatorname{argmin}_{m \in \mathcal{M}_T} \left\{ \gamma(\hat{s}_m) + 2 \int \sum_{\lambda} \varphi_{\lambda,m}^2(x) \frac{dN_x}{T^2} \right\}.$$

More generally, we consider minimization of the type

$$\hat{m} := \operatorname{argmin}_{m \in \mathcal{M}_T} \left\{ \gamma(\hat{s}_m) + \operatorname{pen}(m) \right\} \quad (2.1.4)$$

and  $\tilde{s} = \hat{s}_{\hat{m}}$  is the penalized projection estimator.

We would like to prove that the choice  $\hat{m}$  is "good" meaning that it can satisfy an oracle inequality in expectation, typically

$$\mathbb{E}(\|\tilde{s} - s\|^2) \leq C \mathbb{E}(\|\hat{s}_{\bar{m}} - s\|^2) = C \inf_{m \in \mathcal{M}_T} \mathbb{E}(\|s - \hat{s}_m\|^2), \quad (2.1.5)$$

with  $C$  an adequate (not too large) multiplicative factor. This would mean that we are able, without knowing  $\bar{m}$  to find a model  $\hat{m}$  that is performing in essentially the same way. However, we do not obtain (2.1.5): there is usually a small additive error, which is negligible, and, most importantly,  $C$  may grow slowly with  $T$  depending on the family of models.

### 2.1.1 Histograms

Let us illustrate this behaviour on the simplest estimators: histograms on an interval (ie  $\mathbb{X} = [0, T]$ ). This example is fundamental to understand what model selection can or cannot do. This example is also the only one that has been treated for Hawkes processes.

Let  $S_m$  be a vectorial subspace of  $\mathbb{L}_2$  defined by

$$S_m = \left\{ g \quad / \quad g = \sum_{I \in m} a_I \mathbf{1}_I, a_I \in \mathbb{R} \right\},$$

where  $m$  is a set of disjoint intervals of  $\mathbb{X}$ . For histograms, it is actually natural to identify the model  $S_m$  and the set  $m$ , which is called in the sequel "model" too for sake of simplicity. Let  $|m|$  denote the number of intervals in  $m$ .

A strategy refers to the choice of the family of models  $\mathcal{M}_T$ . To avoid any confusion, let  $\#\{\mathcal{M}_T\}$  denote the number of models  $m$  in  $\mathcal{M}_T$ . In the sequel, a partition  $\Gamma$  of  $[0, T]$  should be understood as a set of disjoint intervals of  $[0, T]$  such that their union is the whole interval  $[0, T]$ . A regular partition is such that all its intervals have the same length. We say that a model  $m$  is written on  $\Gamma$  if all the extremities of the intervals in  $m$  are also extremities of intervals in  $\Gamma$ . For instance if  $\Gamma = \{[0, 0.25], (0.25, 0.5], (0.50, 0.75], (0.75, 1]\}$  then  $\{[0, 0.25], (0.25, 1]\}$  or  $\{[0, 0.25], (0.75, 1]\}$  are models written on  $\Gamma$ . Now let us give some examples of families  $\mathcal{M}_T$ . Let  $J$  and  $N$  be two positive integers.

**Nested strategy** Take  $\Gamma$  a dyadic regular partition (i.e. such that  $|\Gamma| = 2^J$ ). Then take  $\mathcal{M}_T$  as the set of all dyadic regular partitions of  $[0, T]$  that can be written on  $\Gamma$ , including the void set. In particular, note that  $\#\{\mathcal{M}_T\} = J + 2$ . We say that this strategy is nested since for any pair of partitions in this family, one of them is always written on the other one.

**Irregular strategy** Assume now that we know that  $s$  is piecewise constant on  $[0, T]$  but that we do not know where the cuts of the resulting partition are. We can consider  $\Gamma$  a regular partition such that  $|\Gamma| = N$  and then consider  $\mathcal{M}_T$  the set of all possible partitions written on  $\Gamma$ , including the void set. In this case  $\#\{\mathcal{M}_T\} \simeq 2^N$ .

**Islands strategy** This last strategy has been especially designed to answer biological questions. We think that  $s$  has a very localized support. The interval  $[0, T]$  is really large and in fact  $s$  is non zero on a really smaller interval or a union of really smaller intervals: the resulting model is sparse. We can consider  $\Gamma$  a regular partition such that  $|\Gamma| = N$  and then consider  $\mathcal{M}_T$  the set of all the subsets of  $\Gamma$ . A typical  $m$  corresponds to a vectorial space  $S_m$  where the functions  $g$  are zero on  $[0, T]$  except on some disjoint intervals which look like several "islands". In this case  $\#\{\mathcal{M}_T\} = 2^N$ .

For all the previous strategies one can prove the following result.

**Proposition 1** (RB 2003). *Let  $\{L_m, m \in \mathcal{M}_T\}$  be a family of positive weights such that  $\sum_{m \in \mathcal{M}_T} e^{-L_m |m|} \leq \Sigma$  and assume that  $|\Gamma| \leq T(\ln T)^{-2}$  with  $T > 2$ . For any  $c > 1$ , if*

$$\text{pen}(m) = \frac{c\tilde{M}|m|}{T} (1 + \sqrt{2\kappa L_m})^2 \text{ with } \tilde{M} = \sup_{I \in \Gamma} \frac{N_I}{\mu(I)},$$

then

$$\mathbb{E}(\|s - \tilde{s}\|^2) \leq \square_c \inf_{m \in \mathcal{M}_T} \left[ \|s - s_m\|^2 + \frac{M|m|}{T} (1 + L_m) \right] + \square_{c, \Sigma, M} \frac{1}{T}, \quad (2.1.6)$$

where

$$M = \sup_{I \in \Gamma} \frac{\int_I s(x) dx}{\mu(I)}.$$

*NB* :  $\kappa$  is an absolute constant,  $\kappa = 6$  works.

This result is an adapted and simpler version of the one presented in [57]. Note that in the original paper, one can deal with any kind of vectorial subspaces, which satisfy suitable assumptions. In particular, the Fourier basis or wavelet bases may be considered.

To shorten mathematical expressions, we used  $\square_c$  and  $\square_{c,\Sigma,M}$  even if precise formulas are available. When asymptotic in  $T$  is performed, we consequently need to make  $c, \Sigma$  and  $M$  independent of  $T$ . However any dependency between  $\mathcal{M}_T$  and  $T$  is allowed. In this sense, the result of (2.1.6) (as the ones due to Barron, Birgé and Massart [9]) is non asymptotic with respect to various existing works (such as Mallows' [49]) where the family of models is held fixed whereas  $T$  tends to infinity. To obtain (2.1.6), the fundamental tool is to derive non asymptotic exponential inequalities, as we will in the sequel. Before stating these probabilistic results, let us understand the different behaviours of (2.1.6) with respect to the different strategies.

Note that for the Nested strategy there exists at most one model  $m$  in the family with dimension  $|m| = D$  and therefore choosing  $L_m = \epsilon > 0$  fixed leads to a quantity  $\Sigma$  independent of  $T$  whatever  $\Gamma$  is. We can also remark that  $M|m|/T$  is a natural upper bound for the variance term (see (2.1.3)) and that it is sufficient to assume that  $s$  is lower bounded on  $\mathbb{X}$  to lower bound the variance term by  $r|m|/T$  where  $r = \inf_{x \in \mathbb{X}} s(x)$ . Therefore the result is exactly an oracle inequality as expected in (2.1.5) with a true constant  $C$ , up to some negligible residual term. The Nested strategy is consequently adaptive in the minimax sense when Hölder<sup>1</sup> functions with unknown regularity  $0 < \alpha < 1$  are considered. Indeed,  $\mathcal{M}_T$  may be large enough to guarantee the existence of  $m$  in  $\mathcal{M}_T$  such that  $|m| \simeq T^{1/(2\alpha+1)}$ . But for this peculiar  $m$ ,  $\hat{s}_m$  achieves the minimax rate of convergence, namely  $T^{-2\alpha/(2\alpha+1)}$ . The oracle inequality directly proves that the same rate holds for our penalized projection estimator. More generally, in the same spirit, the use of wavelet bases proves the minimax adaptivity with respect to Besov spaces (see [57] for more details).

On the other hand, for the Irregular or Islands strategies, there are approximately  $(N/D)^D$  models in the family with the same dimension  $|m| = D$ , therefore one has to take  $L_m = \alpha \ln(N)$  or  $\alpha \ln(N/D)$  to ensure that  $\Sigma$  will not depend on  $T$  whatever  $\Gamma$  is (in particular when  $N = T$ ). In this case we recover an oracle inequality (see (2.1.5)) where  $C$  which is multiple of  $\ln(T)$  basically, up to some negligible residual term. This phenomenon is actually unavoidable when considering such complex families of models (ie families with complex cardinality: there are more models with the same dimension  $D$  than a power of  $D$ ). Indeed, there exists a minimax lower bound (see Proposition 4 of [57]) that proves the existence of this logarithmic factor. See also [10] and [11] for a more thorough study in the Gaussian set-up.

---

<sup>1</sup>Here, a function  $f$  is Hölder with regularity  $\alpha$  on  $[0, T]$  if there exists a fixed positive real number  $R > 0$  such that for all  $x, y \in [0, T]$ ,  $|f(x) - f(y)| \leq R(|x - y|/T)^\alpha$ .

### 2.1.2 Concentration inequalities

The fundamental probabilistic ingredient to show such oracle inequalities is to control the deviations of  $\|s_m - \hat{s}_m\|$  which can be written, in the more general set-up, as

$$\chi(m) = \sqrt{\sum_{\lambda} \left( \int_{\mathbb{X}} \varphi_{\lambda,m}(x) \frac{dN_x - s(x)dx}{T} \right)^2},$$

where  $(\varphi_{\lambda,m})_{\lambda}$  is an orthonormal basis of  $S_m$ .

In [51], Massart emphasizes the link between Gaussian concentration phenomenon (due to Cirel'son, Ibragimov and Sudakov [19]) and oracle inequalities in the Gaussian set-up, but also the link between Talagrand's inequality [70] (and the successive improvements due to Ledoux [46], Massart [50], Klein and Rio [44] or Bousquet [12]) and the density estimation or the classification problem. For Poisson processes, the existing inequalities ([75], [34]) were not sharp enough to build nice oracle inequalities. Using the infinitely divisible properties of the Poisson process and Ledoux/Massart's approach, one gets the following result

**Theorem 1** (RB 2003). *Let  $N$  be a Poisson process on  $\mathbb{X}$  with finite mean measure  $\nu$ . Let  $\{\psi_a, a \in A\}$  be a countable family of functions with values in  $[-b; b]$ . If*

$$Z = \sup_{a \in A} \int_{\mathbb{X}} \psi_a(x) (dN_x - d\nu_x),$$

then for all  $u, \varepsilon > 0$ ,

$$\mathbb{P}(Z \geq (1 + \varepsilon)\mathbb{E}(Z) + 2\sqrt{\kappa\nu u} + \kappa(\varepsilon)bu) \leq e^{-u},$$

with

$$v = \sup_{a \in A} \int_{\mathbb{X}} \psi_a^2(x) d\nu_x$$

and  $\kappa = 6$ ,  $\kappa(\varepsilon) = 1.25 + 32\varepsilon^{-1}$ .

This result, which can be found in [57], has essentially the same flavour as Talagrand's inequality. The point measure replace the empirical measure of Talagrand's inequality and the mean measure replaces the expectation. The main difficulty with respect to Talagrand's inequality is that it is not a straight application of it. Indeed, by using the infinite divisible property of the Poisson process, it is true that all the integrals can be viewed as a sum of iid variables but those variables are not bounded. Here the bounded character is placed on the functions  $\psi_a$ . One can also note that the term  $\kappa$  appearing in Theorem 1 is actually the same as the one appearing in the penalty of Proposition 1. The shape of the penalty that is required to obtain an oracle inequality is actually completely related to the shape of this concentration inequality. Indeed it is now easy to obtain an exponential inequality for  $\chi(m)$ , since

$$\chi(m) = \sup_{f \in S_m, \|f\|=1} \frac{1}{T} \int f(x) (dN_x - s(x)dx).$$

**Corollary 1** (RB 2003). *Let*

$$M_m = \sup_{f \in S_m, \|f\|=1} \frac{1}{T} \int_{\mathbb{X}} f^2(x) s(x) dx \quad \text{and} \quad B_m = \sup_{f \in S_m, \|f\|=1} \|f\|_{\infty}.$$

Then for all  $u, \varepsilon > 0$ ,

$$\mathbb{P} \left( \chi(m) \geq (1 + \varepsilon) \sqrt{\frac{1}{T} \sum_{\lambda} \int \varphi_{\lambda, m}^2(x) s(x) dx} + \sqrt{\frac{2\kappa M_m u}{T}} + \kappa(\varepsilon) \frac{B_m u}{T} \right) \leq e^{-u}. \quad (2.1.7)$$

One can see that there are actually two behaviours. When  $u$  is small, the behaviour is Gaussian with a variance of the order  $M_m/T \leq \|s\|_{\infty}/T$  which does not grow with the dimension  $|m|$  of the model. When  $u$  is large, the behaviour is sub exponential.

There are several improvements of this inequality. First it is possible by restricting oneself to a large event, depending on the model  $S_m$ , to privilege the Gaussian behaviour (see Proposition 9 of [57]). This is a classical trick due to Massart, which is easily done once one has a Talagrand like inequality. Using this trick and simplifying a little, the penalty is obtained by keeping the first two terms of (2.1.7) with  $u = L_m$ . The fact that  $c > 1$  in Proposition 1 is directly connected with the factor  $(1 + \varepsilon)$  in (2.1.7).

But it is also possible to improve, for instance in the case of histograms, the sub-exponential behaviour. This is at the root of the paper [33] which is a joint work with C. Houdré and P. Marchal. To do so, one uses a complete different technique using covariance formulas, which applies for any infinitely divisible variables. It is possible (see [33] for more details) to replace for large  $u$ , the rate  $e^{-u}$  by  $e^{-u \ln u}$ . This does not change the statistical inference, because we focus on the Gaussian behaviour for the model selection approach. However this shows that there exists more-than-exponential moments for variables as  $\chi(m)$ , when they are norms of vectors with independent coordinates. This means that there exists a positive  $\lambda$  such that  $\mathbb{E}(\exp[\lambda \chi(m) \log \chi(m)])$  is finite. This result was known since Rosiński's work [65], and we have extended it by proving that the range of such possible positive  $\lambda$  does not depend on the dimension of  $S_m$ . Actually the result holds for any euclidean norm of vectors with infinitely divisible independent coordinates whose Lévy measure has a finite support. This and other dimension-free results for infinitely divisible variables may be found in [33].

## 2.2 Other counting processes (Aalen, Hawkes)

In this section, let us present a unified approach for the Aalen multiplicative intensity and the Hawkes process. This approach is the one used in the joint work with S. Schbath [62] for the Hawkes case. A slightly different approach has been used in [59] for the Aalen case. Let us recall that the notation  $s$  represents the deterministic unknown function appearing in (1.1.2) for the Aalen set-up and that we basically assume that  $s$  in this case belongs to

$$\mathbb{L}_2 = \left\{ g \text{ with support in } [0, A] \ / \ \int_0^A g^2 < \infty \right\}.$$

Note that the natural corresponding norm is  $\|g\|^2 = \int_0^A g^2$ .

For the Hawkes process (see (1.1.3)),  $s = (\nu, h)$  represents the couple where  $\nu$  is the spontaneous rate of apparition (this is a real number) and  $h$  is the interaction function. In this case, we basically assume that  $s$  belongs to

$$\mathbb{L}_2 = \left\{ f = (\mu, g) \ / \ g \text{ with support in } (0, A] \text{ and } \int_0^A g^2 < \infty \right\}.$$

In this case, the natural corresponding norm is  $\|f\|^2 = \mu^2 + \int_0^A g^2$ .

In both cases, the intensity of the process  $\Lambda(t)$  is of the shape  $\Psi_s(t)$ , where  $\Psi$  is a linear application that transforms any  $f$  in the corresponding  $\mathbb{L}_2$  space into a predictable process. Indeed for the Aalen case,  $\Psi_f(t) = Y_t f(t)$  and in the Hawkes case,  $\Psi_f(t) = \mu + \int_{t-A}^{t-} g(t-u) dN_u$ . Note that the Poisson process is also of this type with  $\Psi_f = f$ . Actually, the preliminary forthcoming computations are true for any kind of counting process whose intensity will have this linear shape. However the corresponding theoretical result is not published anywhere, because at the end there are always difficult quantities to control and those controls are so specific to each kind of process that a general, but still meaningful, oracle inequality is difficult to write.

Let us observe the counting process  $N$  on an interval  $[0, T]$  (or  $(-A, T]$  for the Hawkes process) and let us define a least-square contrast by

$$\forall f \in \mathbb{L}_2, \quad \gamma(f) = -\frac{2}{T} \int_0^T \Psi_f(t) dN_t + \frac{1}{T} \int_0^T \Psi_f(t)^2 dt. \quad (2.2.1)$$

Indeed, because of the martingale properties, one easily sees that the compensator of the previous formula at time  $T$  is

$$-\frac{2}{T} \int_0^T \Psi_f(t) \Psi_s(t) dt + \frac{1}{T} \int_0^T \Psi_f(t)^2 dt = \frac{1}{T} \int_0^T \Psi_{f-s}(t)^2 dt - \frac{1}{T} \int_0^T \Psi_s(t)^2 dt.$$

Hence the expectation of  $\gamma(f)$  is minimal when  $\Psi_{f-s}(t) = 0$  for almost every  $t$  almost surely. The fact that this implies that  $f = s$  depends of course of the process. For the Aalen multiplicative case, this amounts to assume that  $\mathbb{E}(Y_t^2) > 0$  for all  $t \leq A$  whereas it is more difficult to prove but still true for Hawkes processes when one assumes that  $h$  has a bounded support.

We divided by  $T$  so that the contrast is exactly the one we used for Poisson process, but this division does not change the point where the minimum is reached, so it is not really necessary. Note that for the Poisson process, the division by  $T$  was a nice way to introduce asymptotic properties when  $T$  tends to infinity. Indeed remark that to derive a true oracle inequality we basically assumed the intensity to be lower bounded. Hence if  $T$  grows, the total number of points grows. This vision is still the correct one for Hawkes processes: when  $T$  grows, one observes more and more interactions so the estimation should be better. However for the Aalen case, it is not true that the estimation is better when the time  $T$  grows. Think for instance of the right-censored case where we want to estimate the hazard rate of the life time of only one patient. There is absolutely no way to think that because one observes this patient longer (after his death), we will obtain more information. On the contrary, our estimation will improve when the total number of patients is growing, and this will be true for any kind of aggregated processes. In [59], a slightly different least-square contrast was used but it heavily depends on the multiplicative shape of the intensity. Here we only need the linear transformation  $\Psi$ .

We can pursue the construction of the projection estimators as before. If  $S_m$  is a finite vectorial subspace of  $\mathbb{L}_2$  then

$$\hat{s}_m := \operatorname{argmin}_{f \in S_m} \gamma(f). \quad (2.2.2)$$

Note however that it is not evident to find a closed-form expression for the solution of this minimization. Indeed, with respect to the Poisson case (2.1.1), on the right hand side of (2.2.1) appears a random quantity

$$D_T(f) := \frac{1}{T} \int \Psi_f(t)^2 dt$$

which is a random quadratic form on  $\mathbb{L}_2$ . It happens that in the Poisson case it is the  $\mathbb{L}_2$ -norm and that it is sufficient to write  $f$  on an  $\mathbb{L}_2$ -basis to find the solution, as we have seen before.

In the Aalen case, a closed-form expression exists in the case of histograms because the natural basis, namely  $(\mathbf{1}_I)_{I \in m}$ , is orthogonal for this quadratic form. This case was treated in [59]. The other case, also treated in [59], consists in considering random models  $S_m$  so that a -random- orthonormal basis was known for this quadratic form. Again, in this case, a closed-form expression was available.

For Hawkes processes, none of the previous tricks worked and only a computer can give the solution of the minimization. Nevertheless, this is just a quadratic minimization, so it is not time consuming.

Next we consider a family of models  $\{S_m, m \in \mathcal{M}_T\}$  and we consider again

$$\hat{m} := \operatorname{argmin}_{m \in \mathcal{M}_T} \{\gamma(\hat{s}_m) + \operatorname{pen}(m)\}, \quad (2.2.3)$$

and  $\tilde{s} = \hat{s}_{\hat{m}}$ .

### 2.2.1 Exponential inequalities for counting processes

As one can see on the Poisson case, the penalty is directly linked to the concentration inequality. Hence, before stating the corresponding oracle inequalities, let us stress the main problems and results occurring when we deal with general counting processes. Without further details, it is quite obvious to see that the main quantity to control is

$$\chi(m) = \sqrt{\sum_{\lambda} \left( \int_0^T \Psi_{\varphi_{\lambda, m}}(x) \frac{dN_x - \Lambda(x)dx}{T} \right)^2} = \sup_{f \in S_m, \|f\|=1} \int_0^T \Psi_f(x) \frac{dN_x - \Lambda(x)dx}{T},$$

where  $(\varphi_{\lambda, m})_{\lambda}$  is an orthonormal basis of  $S_m$ . In [58], the compensator of a supremum of counting processes was computed. It allows to derive the following result

**Theorem 2** (RB 2006). *Let  $(N_t)_{t \geq 0}$  be a counting process with intensity  $\Lambda(t)$  assumed to be almost surely integrable on  $[0, T]$ . Let  $\{(H_{a, t})_{t \geq 0}, a \in A\}$  be a countable family of predictable processes and let*

$$\forall t \geq 0, \quad Z_t = \sup_{a \in A} \int_0^t H_{a, s} (dN_s - \Lambda(s)ds).$$

*Then the compensator  $(A_t)_{t \geq 0}$  exists, is non negative et non decreasing and*

$$\forall 0 \leq t \leq T, \quad Z_t - A_t = \int_0^t \Delta Z(s) (dN_s - \Lambda(s)ds),$$

*for a particular predictable process  $\Delta Z(s)$  satisfying  $\Delta Z(s) \leq \sup_{a \in A} H_{a, s}$ .*

*Moreover, if the  $H_{a, s}$ 's have values in  $[-b, b]$  and if  $\int_0^T \sup_{a \in A} H_{a, s}^2 \Lambda(s)ds \leq v$  almost surely for some deterministic constants  $v$  and  $b$ , then for all  $u > 0$ ,*

$$\mathbb{P} \left( \sup_{[0, T]} (Z_t - A_t) \geq \sqrt{2vu} + \frac{bu}{3} \right) \leq e^{-u}.$$

This result is a shortened version of Proposition 1 and Theorem 1 of [58]. This result seems more general than Theorem 1 because it deals with general counting processes and as icing on the cake, we obtain an additional supremum on  $t$ . But this has a cost. Indeed we can observe that there is an exchange between the supremum and the integral in the definition of  $v$ . This cost may seem completely unavoidable in dependent setting. For instance the results developed by Wu [75] and Houdré and Privault [34], using martingales techniques, present this exchange. This exchange was also noticed in other dependent set-up (see for instance Samson's work on Markov chains [66]).

To understand more precisely what this exchange means, let us apply the previous result to  $\chi(m)$ .

**Corollary 2** (RB 2006). *Let*

$$\mathcal{C} = \sum_{\lambda} \int_0^T \frac{\Psi_{\varphi_{\lambda,m}}(x)^2}{T^2} \Lambda(x) dx,$$

*and assume that  $\mathcal{C}$  is bounded by  $v$  and  $\sum_{\lambda} \Psi_{\varphi_{\lambda,m}}(x)^2$  is bounded by  $b$  for all  $x \in [0, T]$ . Then, for all  $u > 0$ ,*

$$\mathbb{P}\left(\chi(m) \geq \sqrt{\mathcal{C}} + 3\sqrt{2vu} + bu\right) \leq 2e^{-u}.$$

Assume than in our case, one can suppose  $\int \Psi_{\varphi_{\lambda,m}}(x)^2 \Lambda(x) dx$  bounded by some fixed constant. If we denote by  $D_m$  the dimension of  $S_m$  then the Gaussian part has a variance of the order  $D_m/T$  and grows with the dimension of  $S_m$ , whereas it was a constant for the Poisson case. The oracle-type inequality that we derive using this exponential inequality cannot be as sharp as the one we obtained for Poisson processes in general.

Recently, Baraud [6] proves via chaining argument a result that supersedes Corollary 2 to some extent. His result actually states, in a more general set-up than the one of counting processes, that one can obtain a Gaussian part with dimension-free variance at the cost of a larger constant term (*ie* the concentration phenomena in his case is not around  $\sqrt{\mathcal{C}}$  but around something larger). In good cases (special choices of  $u \simeq D_m$  and "nice" counting processes), it may happen that one recovers the order of magnitude of the Poisson case instead of the present deteriorate rate.

## 2.2.2 Oracle inequalities

It is quite inappropriate to write a general oracle inequality, because it heavily depends on the norm one considers. The natural "norm" we would like to consider is  $D_T(f)$ . But  $D_T(f)$  is a random quadratic form and not strictly speaking a norm: it may eventually be null for some non zero  $f$ . Of course this function  $f$  would have to be random and very peculiar. It is easier to understand it in the Aalen case, even if the same phenomenon applies for Hawkes processes. If  $Y_t = 0$  on some subinterval of  $[0, T]$ , then a function  $f$  which is non zero on this random interval is a solution. Assuming that  $\mathbb{E}(Y_t^2) > 0$  on the whole interval  $[0, T]$  does not prevent the random variable to be null eventually. For the Aalen case, one has to restrict oneself to the event  $\{Y_t \text{ bounded from below on } [0, T]\}$ . More generally we will have to restrict oneself at least to the event

$$\mathcal{E} = \{\forall m \in \mathcal{M}_T, \quad \forall f \in S_m, \quad r^2 \|f\|^2 \leq D_T^2(f) \leq R^2 \|f\|^2\}, \quad (2.2.4)$$

for some fixed constants  $r$  and  $R$ , with  $\|f\|$  the natural norm on  $\mathbb{L}_2$ . But then of course the resulting oracle inequality (2.1.5) cannot hold in expectation on the whole probability

space. To do so, among other technicalities, one obviously needs to control  $\mathbb{P}(\mathcal{E}^c)$  and this is basically not related to the martingale structure of the counting process but to some additional properties.

### Aalen multiplicative intensity

For the Aalen case, the additional properties may come from the aggregated case. Let us just give a brief summary of the type of oracle inequalities that can be found in [59].

- If one uses histograms, and if, among other technical assumptions, one assumes that  $N$  is a bounded aggregated process, then a result strictly equivalent to Theorem 1 is available, since Talagrand's inequality can be used on the aggregated process.
- If one uses random models, with known orthonormal basis for  $D_T(f)$ , then one is forced to use the exponential inequality of Corollary 2.
  - Hence the oracle inequality is limited to not too complex families of models. One model per dimension is the basic case, for which the penalty should be  $\text{pen}(m) = cD_m/T$  for some large enough constant  $c$ .
  - the oracle inequality is stated as follows

$$\mathbb{E}(D_T(s - \tilde{s})\mathbf{1}_{\mathcal{E}}) \leq \square_{c,s} \inf_{m \in \mathcal{M}_T} \left[ \mathbb{E}(D_T(s - \hat{s}_m)) + \frac{D_m}{T} \right] + \square_{c,s} \frac{1}{T}.$$

- One can control  $\mathcal{E}$  if one assumes again the process to be aggregated.

Note that the approach based on least-square contrasts for aggregated counting processes has been developed and widened by Brunel and Comte (and co-authors) in a succession of papers, in particular under various type of censoring (see for instance [16] and [17]).

### Hawkes processes

For the Hawkes process, one cannot use that  $N$  is an aggregated process any more and that the individual processes are more or less bounded. It is true that the Hawkes process is infinitely divisible but it is typically unbounded, the number of points per interval being sensibly larger than a Poisson variable (exponential moments exist but not of any order). The concentration for infinite divisible variables developed in [33] cannot be applied directly to the resulting  $\chi(m)$ . Indeed  $\chi(m)$  can be viewed as the norm of a random vector of infinitely divisible variables, but the structure of  $\Psi$  does not allow those variables to be independent.

However, one still needs to control the event  $\mathcal{E}$ . Actually, the Hawkes process has ergodic properties that show that  $D_T(f)$  tends to a true norm on  $\mathbb{L}_2$  when  $T$  tends to infinity. Asymptotic properties are of no use for model selection, as we already said. In a joint work with E. Roy [60], we derive exponential inequalities for Hawkes processes. Two kinds of inequalities are required: first, we want to control the number of points of the Hawkes process per interval and second and most importantly, we want to refine in a non asymptotic way the ergodic theorem. This has been done using arguments such as coupling in [60]. The results of [60] imply the following result:

**Lemma 1** (RB Roy 2007). *Let  $(N_t)_{t \in \mathbb{R}}$  be a stationary Hawkes process, with intensity given by  $\Lambda(t) = \Psi_s(t)$  with  $s = (\nu, h)$  in  $\mathbb{L}_2$  and positive  $h$ . Note that the definition of  $\mathbb{L}_2$  implies that the interaction function  $h$  has a bounded support included in  $(0, A]$ . Let  $g$  be a function of the points of  $(N_t)_{t \in \mathbb{R}}$  lying in  $[-A, 0)$ , with values in  $[-B, B]$  and zero mean. Let  $(\theta_t)_{t \in \mathbb{R}}$  be the flow induced by  $(N_t)_{t \in \mathbb{R}}$  i.e.  $g \circ \theta_t$  is the same function as before, but now the points are lying in  $[-A + t, t)$ . Then there exists a positive constant  $T_0(p, A)$  depending on  $p = \int h$  and  $A$ , such that for all  $T \geq T_0(p, A)$*

$$\mathbb{P} \left( \left| \frac{1}{T} \int_0^T g \circ \theta_t dt \right| \geq 2 \sqrt{\frac{c_1 \text{Var}(g) A \log(T)^2}{T(p - \log p - 1)} + \frac{c_2 B A \log(T)^2}{T(p - \log p - 1)}} \right) \leq \frac{\square_{\nu, p}}{T^3},$$

where  $c_1$  and  $c_2$  are absolute constants.

In [60], the key tool is the behaviour of the process not in terms of martingale but in terms of branching process, when  $h$  is non negative. Indeed, one can also view the Hawkes process, as the superposition of several layers of points. The first layer consists in points that are called "ancestors", which constitutes a homogeneous Poisson process with rate  $\nu$ . Then each ancestor,  $x$ , gives birth to children according to a Poisson process with intensity  $h(t-x)$ , and each new child, gives birth to other points according to the same process. The mechanism stops almost surely if  $\int h < 1$  because the underlying number of descendants of each ancestor is just a Galton-Watson process with reproducing law, a Poisson law with parameter  $\int h$ . The Hawkes process is just the union of all the descendants and all the ancestors. This cluster representation has been discovered a long time ago by Hawkes and Oakes [32]. In [60], the main ingredient consists in understanding what happens when we keep all the ancestors before 0 and then remove all the ones appearing after 0. Certainly there is a time, called the extinction time,  $T_e$ , after which no more points can be viewed. On the one hand, we derive tail estimates on  $T_e$  depending on the function  $h$  and on the other hand, we use a coupling construction to control the total variation distance between the process  $N$  and another process, which will be piecewise independent. A more precise statement and variations of Lemma 1 can be found in [60].

With Lemma 1, it is long but not that difficult to control  $\mathcal{E}$  (see the long version of the paper [62] on arXiv), however to do so and for technical reasons, we cannot use any kind of model  $S_m$ . For the Hawkes process,  $S_m$  denotes the set of couples  $(\mu, g)$  where  $\mu$  is any real number and where  $g$  is a piecewise constant function on a set  $m$  of intervals of  $(0, A]$ . All the strategies, ie families of possible  $m$ 's, that were described as histograms strategies for Poisson processes apply here.

The only remaining problem is that we need to control  $\tilde{s}$  on  $\mathcal{E}^c$ , which can be done theoretically speaking via clipping. Let us define for all real numbers  $H > 0$ ,  $\eta > \rho > 0$ ,  $1 > P > 0$ , the following subset of  $\mathbb{L}_2$ :

$$\mathcal{L}_{H,P}^{\eta,\rho} = \left\{ f = (\mu, g) \in \mathbb{L}_2 / \mu \in [\rho, \eta], \quad g(\cdot) \in [0, H] \text{ and } \int_0^A g(u) du \leq P \right\},$$

and let us assume that we know that  $s$  belongs to this set. Recall that the penalized projection estimator  $\tilde{s} = (\tilde{\nu}, \tilde{h})$  is given by (2.2.3). Then, under the previous assumption, it is natural to consider the clipped penalized projection estimator,  $\bar{s} = (\bar{\nu}, \bar{h})$ , given, for

all positive  $t$ , by

$$\begin{cases} \bar{\nu} &= \begin{cases} \tilde{\nu} & \text{if } \rho \leq \tilde{\nu} \leq \eta, \\ \rho & \text{if } \tilde{\nu} < \rho, \\ \eta & \text{if } \tilde{\nu} > \eta, \end{cases} \\ \bar{h}(t) &= \begin{cases} \tilde{h}(t) & \text{if } 0 \leq \tilde{h}(t) \leq H, \\ 0 & \text{if } \tilde{h}(t) < 0, \\ H & \text{if } \tilde{h}(t) > H. \end{cases} \end{cases} \quad (2.2.5)$$

**Theorem 3** (RB Schbath 2010). *Let  $(N_t)_{t \in \mathbb{R}}$  be a Hawkes' process with intensity  $\Psi_s(\cdot)$ . Assume that we know that  $s$  belongs to  $\mathcal{L}_{H,P}^{\eta,\rho}$ . Moreover assume that all the models in  $\mathcal{M}_T$ , ie possible sets  $m$  of intervals, are written on  $\Gamma$ , a regular partition of  $(0, A]$  such that*

$$|\Gamma| \leq \frac{\sqrt{T}}{(\log T)^3}. \quad (2.2.6)$$

Let  $Q > 1$ . Then there exists a positive constant  $\kappa$  depending on  $\eta, \rho, P, A, H$  such that if

$$\forall m \in \mathcal{M}_T, \quad \text{pen}(m) = \kappa Q (|m| + 1) \frac{\log(T)^2}{T}, \quad (2.2.7)$$

then

$$\mathbb{E}(\|\bar{s} - s\|)^2 \leq \square_{\eta,\rho,P,A,H} \inf_{m \in \mathcal{M}_T} \left[ \|s - s_m\|^2 + (|m| + 1) \frac{\log(T)^2}{T} \right] + \square_{\eta,\rho,P,A,H} \frac{\#\{\mathcal{M}_T\}}{TQ},$$

where  $s_m$  is the orthogonal projection of  $s$  on  $S_m$ .

One can compare Proposition 1 and Theorem 3. First  $|\Gamma|$  should be smaller for the Hawkes process: this comes basically from the control of  $\mathcal{E}$ , which was unnecessary for the Poisson process since  $D_T(f)$  was deterministic in this case. Next, weights - the  $L_m$ 's - were appearing in the Poisson case: they are replaced here by the factor  $Q \log(T)$ . Actually the concentration we used (Corollary 2) is not sharp enough to use weights as precisely as in Proposition 1. Indeed since the dimension appears in the variance term in Corollary 2, one needs basically to take  $u = Q \ln(T)$  to obtain a deviation of the order  $\sqrt{D_m/T}$  up to some logarithmic term. On the contrary, the variance does not depend on the dimension in Corollary 1 for the Poisson case and one can take  $u = L_m D_m$ .

In addition, the penalty has an extra  $\log(T)$  factor which comes from the fact that the intensity  $\Lambda(t)$  is unbounded:  $\Lambda(t)$  behaves basically as the number of points in an interval of length  $A$ , quantity for which we derived tail estimates in [60]. As we want to control it on the whole interval  $[0, T]$  we lose an extra logarithmic factor.

As before it is possible to prove by using the Nested strategy, that the method is adaptive in the minimax sense for Hölder functions with unknown regularity. However with respect to the Poisson case, two restrictions appear : (i) the regularity  $\alpha$  should belong to  $(1/2, 1)$  since  $|\Gamma| \ll \sqrt{T}$ , (ii) extra logarithmic factors are appearing and we do not know if they are necessary.

For more complex strategies (Irregular or Islands), if lower bounds have been computed, there are still logarithmic factors that cannot be explained (see [62] for more details).

Actually one cannot be completely satisfied with this result. Assuming that one knows that  $s$  belongs to  $\mathcal{L}_{H,P}^{\eta,\rho}$ , is a really strong assumption that cannot be done in practice.

Moreover, for the genomic data we wanted to treat, the positivity assumption of  $h$  is probably not satisfied. Indeed a positive function  $h$  cannot model avoided distances between occurrences, biological phenomena that is known to exist: for instance, genes in *E. Coli* usually do not overlap. The reader may wonder what a possibly negative  $h$  means, mathematically speaking. The self-interaction can be modeled in a more general way by a process whose intensity is given by

$$\Lambda(t) = \left( \nu + \int_{-\infty}^{t^-} h(t-u) dN_u \right)_+ \quad (2.2.8)$$

where  $h$  may now be negative. We have taken the positive part to ensure that the intensity remains positive. Then the condition  $\int |h| < 1$  is sufficient to ensure the existence of a stationary version of the process (see [14]). When  $h(d)$  is strictly positive there is a self-excitation at distance  $d$ . When  $h(d)$  is strictly negative, then there is a self-inhibition. It is more or less the same interpretation as above (see (1.1.3)) except that now all the previous occurrences are voting whether they "like" or "dislike" to have a new occurrence at position  $t$ . The major problem is that the cluster representation does not exist for such a process. Hence, there is no oracle inequality in expectation in this case - even if Theorem 2 of [62] and the remark below tends to prove that an oracle-type inequality in probability exists. However note that our projection estimators,  $\hat{s}_m$ , and penalized projection estimators,  $\tilde{s}$ , do not take the sign of  $g$  or  $h$  into account for being computed. That is the reason why one can still use the method in practice.

In the remaining part of [62], we try to propose at least a practical method, which attempts to fulfil most of the practitioner wishes as stated in the introduction but without any strong mathematical evidence beyond Theorem 3. In the next sections, I detail more precisely what has been done in practice for the Hawkes case and what can be theoretically done for the Poisson process.

# Chapter 3

## Support Assumptions

In this chapter, the purpose is not to estimate the support of a function as in [45] for instance but to question if it is sensible to assume the support of the underlying signal to be known and compact. So we are merely considering here the size of the support as a nuisance parameter.

### 3.1 Hawkes framework

The strongest assumption in my opinion that has been done in [62] is the support assumption. We assumed that the unknown function  $h$  has a finite support  $[0, A]$  and more importantly that we know  $A$ . On the one hand, this does not seem to be a strong assumption because biologists know that after 10000 bases the 3D structure of DNA interferes and that  $A = 10000$  works. But let us take a closer look at what happens in practice.

Let us look more closely at the data set corresponding to the occurrences of the 4290 genes along both strands of the complete genome of the bacterium *Escherichia coli* ( $T = 9288442$ ). In [31], Gusto and Schbath already proposed a practical method for selecting knots for a spline estimation of the function  $h$ . They used an AIC criterion that was shown in practice to be relevant for equally spaced knots on the whole interval  $[0, A]$ . The corresponding implemented method is named FADO.

Figure 3.1 presents the results of the FADO procedure [31]. Here we have forced the estimators to be piecewise constant to make the comparison easier. If the main trend is captured, we see that FADO has some fluctuations until the end of the required interval, namely  $[0, A]$ , with  $A = 10000$ . When do we decide that the fluctuations are negligible? The model selection approach, described in the previous chapter, will produce essentially the same answer if one uses the Nested strategy.

But model selection as designed by Birgé and Massart in [10] may select a sparse signal if the model collection is complex. This is precisely the reason why we introduce the Islands strategy for histograms, in addition to the classical strategies that are Irregular and Nested. This method selects a sparse support. Of course the method has to be calibrated and tested before (see the last chapter on calibration), but the Islands strategy leads to the estimator of Figure 3.2.

Here the interpretation is easier. Figure 3.2 tells us that

- gene occurrences seem to be uncorrelated down to 2600 basepairs - this first point may not have been guessed with FADO -
- they are avoided at a short distance ( $\sim 0$ –500 bps) and

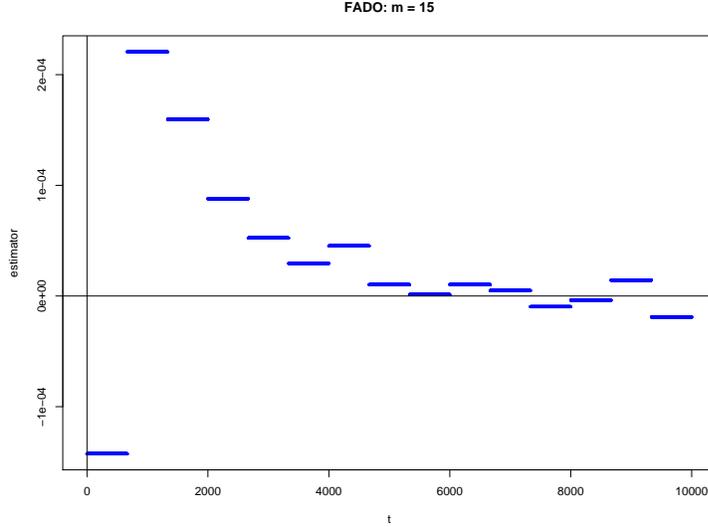


Figure 3.1: FADO estimator of the interaction function between genes.

- favored at distances  $\sim 700\text{--}2000$  bps apart.

This is completely coherent with biological observations: genes on the same strand do not usually overlap, they are about 1000 bps long in average, and there are few intergenic regions along bacterial genomes (compact genomes).

Moreover, even if no mathematical evidence encourage us to do so, we are tempted by zooming in on the first part of the plot. Indeed one can apply the method with  $A = 5000$  or  $A = 2600$  because we have a visual sign that nothing significant appears after that. We refer to [62] for the zoomed results and other biological explanations.

Hence on the one hand, not all the model selection strategies but at least the Islands strategy seems to do the job. It uses the biological knowledge ( $A = 10000$ ) and produces an answer that may have a much smaller support. On the other hand, the several steps procedures that consists in zooming in may produce something disastrous since we do not have mathematical evidence that this procedure is robust. Hence we would like a procedure that do not need  $A$  as an input. For the Hawkes model, this is completely out of reach because the fact that the support is bounded is used everywhere! For instance the fact that  $\mathbb{E}(D_T(f))$  is a norm equivalent to  $\|\cdot\|$  strongly relies on the size of  $A$  (see (2.2.4) and the proofs of [62]). But this problem has a solution for Poisson processes.

## 3.2 Poisson framework

Can we adaptively infer the intensity without any prerequisite such as a known compact support? Can we adaptively estimate a Poisson intensity on the whole real line? First let us remark that the method described in Section 2.1 relies strongly on the bounded assumption of  $\mathbb{X}$ . Actually, as it is written here, since the asymptotic is when  $T$  tends to infinity, taking  $T = \infty$  is a nonsense. However, formally speaking, we can rewrite model selection in the following way. The intensity  $f$  is defined on the whole real line, we assume that  $\int f < \infty$  and  $f$  is held fixed. But the mean measure of the Poisson process is now  $d\nu = f(x)ndx$  (with respect to Definition 1,  $s(x) = nf(x)$ ) where now  $n$  is tending to

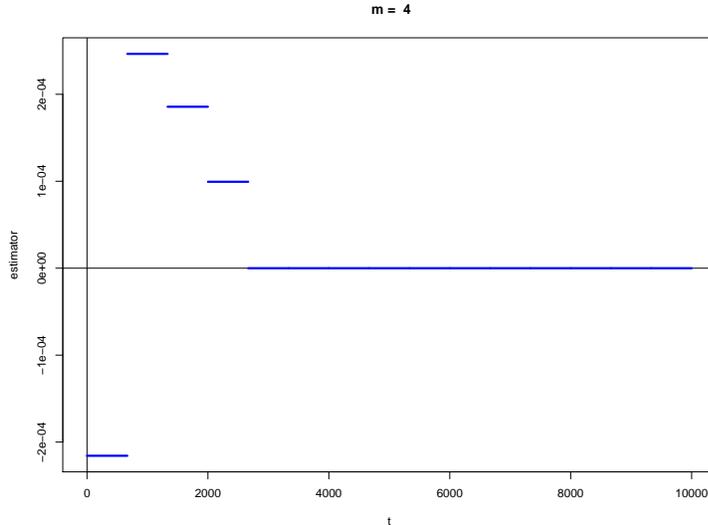


Figure 3.2: Islands estimator of the interaction function between genes.

infinity when asymptotic is considered. Indeed, the mean total number of points is  $n \int f$ , it grows when  $n$  tends to infinity. Model selection as performed by Baraud and Birgé in [7] for random measures or by Figueroa-Lopéz and Houdré in [27] for Lévy processes covers this case. They both obtain oracle inequalities under various assumptions on the family of models. In practice they need an a priori on the localisation of the support to have a reasonable family of models. In particular they cannot estimate a regular intensity that has an infinite support. Most of the existing practical methods (except kernel estimators) assume that  $f$  has a support included in, say,  $[0, 1]$ . How do we build an adaptive estimator of  $f$  without such a strong assumption?

One could think that this problem is a toy problem that has no concrete application. But this is not the case. There exist data for which one cannot assume the existence of an upper bound and that are typically heavy tailed. This situation happens for instance in the financial and geological examples mentioned previously (see [52, 71, 36, 27]) but also in a wide variety of situations (see [21]). All these Poisson processes have in common to be highly inhomogeneous with a lot of points having a small size and some points having extraordinary huge size. There are actually statistical evidence that the tail of the size of petroleum fields is Pareto distributed (see Lepez' Ph.D. thesis [47]), for instance. In general, the classical argument for applying classical methods that need support assumptions, consists in assuming that we know a constant  $M$  such that the support of  $f$  is contained in  $[0, M]$ . Then, observations are rescaled by dividing each of them by  $M$ : the new observations (that all depend on  $M$ ) belong to  $[0, 1]$ . An estimator adapted to signals supported by  $[0, 1]$  can be performed, which leads to a final estimator of  $f$  supported by  $[0, M]$  by applying the inverse rescaling. Note that such an estimator highly depends on  $M$ . For heavy tailed data as mentioned before, there is no possible  $M$ : a usual and classical trick (but not mathematically proved) is to rescale by the largest data in the sample, without any guarantee that this does not strongly interfere with the estimation procedure.

In a joint work with V. Rivoirard [61], we opt for a thresholding approach, which is the extension of the work of Juditsky and Lambert-Lacroix in the density setting [41].

Thresholding may be viewed as a particular case of model selection but actually it has two advantages. First it is simpler to implement thresholding rules than general model selection, and therefore the precision of the method for the same computation time is higher. Secondly, there is a natural way to sum the errors over the whole real line, trick that cannot be done in general by model selection. Let us describe precisely the method and the resulting oracle inequality and then emphasize what the support assumption implies.

### 3.2.1 The method and the oracle inequality

We only assume that  $f$  belongs to  $\mathbb{L}_2 \cap \mathbb{L}_1$  and that we observe  $N$  a Poisson process with intensity  $f$  with respect to  $ndx$ . For technical reason, we need to use a biorthogonal wavelet decomposition. Let us describe the method.

As classical orthonormal wavelet bases, biorthogonal wavelet bases are generated by dilations and translations of father and mother wavelets. But considering biorthogonal wavelets allows to distinguish wavelets for analysis and wavelets for reconstruction. The decomposition of  $f$  on a biorthogonal wavelet basis takes the following form:

$$f = \sum_{k \in \mathbb{Z}} \alpha_k \tilde{\phi}_k + \sum_{j \geq 0} \sum_{k \in \mathbb{Z}} \beta_{j,k} \tilde{\psi}_{j,k}, \quad (3.2.1)$$

where for any  $j \geq 0$  and any  $k \in \mathbb{Z}$ ,

$$\alpha_k = \int_{\mathbb{R}} f(x) \phi_k(x) dx, \quad \beta_{j,k} = \int_{\mathbb{R}} f(x) \psi_{j,k}(x) dx,$$

for any  $x \in \mathbb{R}$ ,

$$\begin{aligned} \phi_k(x) &= \phi(x - k), & \psi_{j,k}(x) &= 2^{\frac{j}{2}} \psi(2^j x - k), \\ \tilde{\phi}_k(x) &= \tilde{\phi}(x - k), & \tilde{\psi}_{j,k}(x) &= 2^{\frac{j}{2}} \tilde{\psi}(2^j x - k) \end{aligned}$$

and  $\Phi = \{\phi, \psi, \tilde{\phi}, \tilde{\psi}\}$  is a set of four particular functions: we consider the particular biorthogonal spline wavelet basis built by Cohen *et al.* [20] (see [61] for more details). The most important ingredient is that  $\psi$  is a compactly supported piecewise constant function. The Haar basis can be viewed as a special case of the previous system, by setting  $\tilde{\phi} = \phi = \mathbf{1}_{[0,1]}$ ,  $\tilde{\psi} = \psi = \mathbf{1}_{[0, \frac{1}{2}]} - \mathbf{1}_{(\frac{1}{2}, 1]}$ . The Haar basis is an orthonormal basis, which is not true for general biorthogonal wavelet bases. However, we have the frame property: the  $\mathbb{L}_2$ -norm of  $f$  is equivalent to the  $\ell_2$  norm of its wavelet coefficients. To shorten mathematical expressions, we set

$$\Lambda = \{\lambda = (j, k) : j \geq -1, k \in \mathbb{Z}\}$$

and for any  $\lambda \in \Lambda$ ,  $\varphi_\lambda = \phi_k$  (respectively  $\tilde{\varphi}_\lambda = \tilde{\phi}_k$ ) if  $\lambda = (-1, k)$  and  $\varphi_\lambda = \psi_{j,k}$  (respectively  $\tilde{\varphi}_\lambda = \tilde{\psi}_{j,k}$ ) if  $\lambda = (j, k)$  with  $j \geq 0$ . Similarly,  $\beta_\lambda = \alpha_k$  if  $\lambda = (-1, k)$  and  $\beta_\lambda = \beta_{j,k}$  if  $\lambda = (j, k)$  with  $j \geq 0$ . Now, (3.2.1) can be rewritten as

$$f = \sum_{\lambda \in \Lambda} \beta_\lambda \tilde{\varphi}_\lambda \quad \text{with} \quad \beta_\lambda = \int \varphi_\lambda(x) f(x) dx. \quad (3.2.2)$$

The estimate of  $f$  is based on the natural unbiased estimators of the  $\beta_\lambda$ 's defined for any  $\lambda$  by

$$\hat{\beta}_\lambda = \frac{1}{n} \int \varphi_\lambda(x) dN(x) = \frac{1}{n} \sum_{T \in N} \varphi_\lambda(T). \quad (3.2.3)$$

This shows the practical interest of using the previous wavelet system. Indeed, since the functions  $\varphi_\lambda$ 's are piecewise constant functions with an explicit mathematical expression, numerical values of these coefficients can be exactly and quickly computed. This is not the case with "usual" regular orthonormal wavelet bases for which computations of the associated coefficients are based on numerical approximations, which is not suitable. Key theoretical arguments are also based on such bases providing a convenient control of the variance of the  $\hat{\beta}_\lambda$ 's.

Now, let us specify our thresholding rule. Given some parameter  $\gamma > 0$ , we define the threshold

$$\eta_{\lambda,\gamma} = \sqrt{2\gamma\tilde{V}_\lambda \ln n} + \frac{\gamma \ln n}{3n} \|\varphi_\lambda\|_\infty, \quad (3.2.4)$$

with

$$\tilde{V}_\lambda = \hat{V}_\lambda + \sqrt{2\gamma(\ln n)\hat{V}_\lambda \frac{\|\varphi_\lambda\|_\infty^2}{n^2}} + 3\gamma \ln n \frac{\|\varphi_\lambda\|_\infty^2}{n^2}$$

where

$$\hat{V}_\lambda = \frac{1}{n^2} \int \varphi_\lambda^2(x) dN(x).$$

Note that  $\hat{V}_\lambda$  satisfies  $\mathbb{E}(\hat{V}_\lambda) = V_\lambda$ , where

$$V_\lambda = \text{Var}(\hat{\beta}_\lambda) = \frac{1}{n} \int \varphi_\lambda^2(x) f(x) dx.$$

Finally given some subset  $\Gamma_n$  of  $\Lambda$  of the form

$$\Gamma_n = \{\lambda = (j, k) \in \Lambda : j \leq j_0\}, \quad (3.2.5)$$

where  $j_0 = j_0(n)$  is an integer, we set for any  $\lambda \in \Lambda$ ,

$$\tilde{\beta}_\lambda = \hat{\beta}_\lambda \mathbf{1}_{\{|\hat{\beta}_\lambda| \geq \eta_{\lambda,\gamma}\}} \mathbf{1}_{\{\lambda \in \Gamma_n\}}$$

and we set  $\tilde{\beta}_n = (\tilde{\beta}_\lambda)_{\lambda \in \Lambda}$ . The estimator of  $f$  is

$$\tilde{f}_{n,\gamma} = \sum_{\lambda \in \Lambda} \tilde{\beta}_\lambda \tilde{\varphi}_\lambda \quad (3.2.6)$$

and only depends on the choice of  $\gamma$  and  $j_0$  fixed later. Note that when a classical model selection method requires to compute all the projection estimators, the thresholding estimator only needs to look at one coefficient at a time and states if one keeps or kills this coefficient.

When the Haar basis is used, the estimate is denoted  $\tilde{f}_{n,\gamma}^H$ .

The threshold  $\eta_{\lambda,\gamma}$  seems to be defined in a rather complicated manner but it is in fact inspired by the universal threshold proposed by Donoho and Johnstone in [25] in the Gaussian regression framework (see [61] for more details).

One can actually rewrite the method as a particular model selection method for a particular contrast which is not the one used in Section 2.1. Indeed for any  $g = \sum_{\lambda \in \Lambda} \alpha_\lambda \tilde{\varphi}_\lambda$ , one can define the following least-square contrast:

$$\gamma'(g) = -2 \sum_{\lambda \in \Lambda} \alpha_\lambda \hat{\beta}_\lambda + \sum_{\lambda \in \Lambda} \alpha_\lambda^2,$$

for which it is easy to see that  $\mathbb{E}(\gamma'(g))$  is minimal as soon as  $g = f$ . Let us define a model by

$$S_m = \left\{ g = \sum_{\lambda \in m} \alpha_\lambda \tilde{\varphi}_\lambda, \quad \alpha_\lambda \in \mathbb{R} \right\},$$

where  $m$  is any subset of indices of  $\Lambda$ . Then, obviously,

$$\hat{f}_m := \operatorname{argmin}_{g \in S_m} \gamma'(g) = \sum_{\lambda \in m} \hat{\beta}_\lambda \tilde{\varphi}_\lambda.$$

Now let us look at the following penalized criteria

$$\hat{m} = \operatorname{argmin}_{m \subset \Gamma_n} \left\{ \gamma'(\hat{f}_m) + \operatorname{pen}(m) \right\},$$

with

$$\operatorname{pen}(m) = \sum_{\lambda \in m} \eta_{\lambda, \gamma}^2.$$

Since  $\gamma'(\hat{f}_m) = -\sum_{\lambda \in m} \hat{\beta}_\lambda^2$ , very easy computations lead to  $\hat{f}_{\hat{m}} = \tilde{f}_{n, \gamma}$ . This means that our thresholding estimator is just a special case of model selection where it is not worth computing all the  $\hat{f}_m$ 's to obtain the penalized estimator.

Combining the advantages of a thresholding procedure which looks at one coefficient at a time and the model selection approach, we are able to prove the following statement, which cannot be recovered by classical model selection methods.

**Theorem 4** (RB Rivoirard 2010). *Let us fix two constants  $c \geq 1$  and  $c' \in \mathbb{R}$ , and let us define for any  $n$ ,  $j_0 = j_0(n)$  the integer such that  $2^{j_0} \leq n^c (\ln n)^{c'} < 2^{j_0+1}$ . If  $\gamma > c$ , then  $\tilde{f}_{n, \gamma}$  satisfies the following oracle inequality: for  $n$  large enough*

$$\mathbb{E} \|\tilde{f}_{n, \gamma} - f\|_2^2 \leq \square_{\gamma, c} \left[ \sum_{\lambda \in \Gamma_n} \min(\beta_\lambda^2, V_\lambda \ln n) + \sum_{\lambda \notin \Gamma_n} \beta_\lambda^2 \right] + \square_{\gamma, c, c', \|f\|_1, \Phi} \frac{1}{n}. \quad (3.2.7)$$

Note that this result holds under the very mild assumption that  $f \in \mathbb{L}_2 \cap \mathbb{L}_1$ . In particular no assumption on the support of  $f$  is done. Note also that the density may be unbounded in infinite norm, which is quite unusual (see [61] and [63] for more details and comments).

This result is actually an oracle inequality. Indeed, we easily see that  $\mathbb{E} \left[ \|\hat{f}_m - f\|_2^2 \right] \asymp R_{\ell_2}(m)$  because of the frame property of biorthogonal wavelet basis, with

$$R_{\ell_2}(m) = \sum_{\lambda \notin m} \beta_\lambda^2 + \sum_{\lambda \in m} V_\lambda.$$

Hence the best possible set of indices corresponds to  $\bar{m}$  with

$$\bar{m} = \{ \lambda \in \Gamma_n \text{ such that } \beta_\lambda^2 > V_\lambda \} \quad (3.2.8)$$

where  $\bar{m}$  minimizes  $m \mapsto R_{\ell_2}(m)$  and we have

$$R_{\ell_2}(\bar{m}) = \sum_{\lambda \in \Gamma_n} \min(\beta_\lambda^2, V_\lambda) + \sum_{\lambda \notin \Gamma_n} \beta_\lambda^2.$$

This represents the benchmark in the family of estimators that keep or kill each coefficient  $\hat{\beta}_\lambda$ . We can also associate to  $\bar{m}$  the quantity  $\hat{f}_{\bar{m}}$ , which, here, plays the role of the oracle. Note also that the corresponding family of models  $m$  (any possible subset of  $\Gamma_n$ ) has a complex cardinality, as complex as Islands. With this approach, we see that Theorem 4 provides the best possible inequality up to a logarithmic term and a residual term. As before in the model selection approach, one can see that the logarithmic term is unavoidable from a minimax point of view for such a complex family of models.

### 3.2.2 The curse of support

Actually two articles refer to the title of this paragraph. In [61], we theoretically studied the previous estimator in the Poisson case, and derived strange minimax behaviours with respect to the support assumption. Rapidly people asked us why we were doing such a job for Poisson processes, since for most of our colleagues if such a phenomenon as a curse of support exists, that should be popularized via a more well known set-up, which is the one of density estimation via the observation of a  $n$ -sample. Indeed, the fact that the data are usually rescaled in practice by the largest observation is much more popular for density estimation. The method we developed for Poisson process, applies to the density setting with a lot of additional computations. In my opinion, Poisson processes are for the density setting what the white noise model is for regression: the computations are easier, the intuition is better because there are not strange constraints, such as  $\int f = 1$  for instance. So the density setting was studied in a more applied way in the joint work with V. Rivoirard and C. Tuleau-Malot [63].

In any case there are definitely two aspects for this curse of support: a minimax aspect and a practical aspect.

#### Minimax aspect

Let us study the minimax rate of convergence over Besov balls. The Besov balls we consider are classical (see [61] or [63] for a definition with respect to the biorthogonal wavelet basis) and denoted  $\mathcal{B}_{p,q}^\alpha(R)$ . Let us just point out that no restriction is made on the support of  $f$  when  $f$  belongs to  $\mathcal{B}_{p,q}^\alpha(R)$ : this support is potentially the whole real line.

Without going into details, let us say that in both settings (Poisson or density), when the functions are also bounded in infinite norms, the dichotomy of Table 3.1 applies.

	$1 \leq p \leq 2$	$2 \leq p \leq \infty$
compact support	$n^{-\frac{2\alpha}{2\alpha+1}}$	$n^{-\frac{2\alpha}{2\alpha+1}}$
non compact support	$n^{-\frac{2\alpha}{2\alpha+1}}$	$n^{-\frac{\alpha}{\alpha+1-\frac{1}{p}}}$

Table 3.1: Minimax rates on  $\mathcal{B}_{p,q}^\alpha \cap \mathbb{L}_2 \cap \mathbb{L}_\infty$  with  $1 \leq p, q \leq \infty$ ,  $\alpha > \max\left(0, \frac{1}{p} - \frac{1}{2}\right)$  under the  $\|\cdot\|_2^2$ -loss.

Those minimax results show the role played by the support of the functions to be estimated on minimax rates. When  $p \leq 2$ , the support has no influence since the rate exponent remains unchanged whatever the size of the support (finite or not). Roughly

speaking, it means that it is not harder to estimate bounded non-compactly supported functions than bounded compactly supported functions from the minimax point of view. It is not the case when  $p > 2$ . Actually, we note an elbow phenomenon at  $p = 2$  and the rate deteriorates when  $p$  increases: this illustrates the curse of support from a minimax point of view. Let us give an interpretation of this observation. Johnstone in [40] showed that when  $p < 2$ , Besov spaces  $\mathcal{B}_{p,q}^\alpha$  model sparse signals where at each level, very few wavelet coefficients are non-negligible. But these coefficients can be very large. When  $p > 2$ ,  $\mathcal{B}_{p,q}^\alpha$ -spaces typically model dense signals where the wavelet coefficients are not large but most of them can be non-negligible. This explains why the size of the support plays a role on minimax rates when  $p > 2$ : when the support is larger, the number of wavelet coefficients to be estimated increases dramatically.

The rates of Table 3.1 are adaptively achieved by our thresholding estimator with respect to  $p$ ,  $\alpha$  and the compactness of the support, up to a logarithmic term. Let us just emphasize the following point. Whatever the support of  $f$  is - bounded support with unknown localization, large bounded support with unknown size, unbounded support - our procedure does not care: we are able to estimate  $f$  for the  $\mathbb{L}_2$  loss at the correct rate of convergence, up to a logarithmic term.

### In practice

From a numerical point of view, let us just mention what is in my opinion the most striking example of this curse of support, in the density setting. Gaussian variables are so nice that one usually thinks they have an almost bounded support. Hence estimating a mixture of only two Gaussian variables should be an easy task.

We compared our method to representative methods of each main trend in density estimation, namely kernel, binning plus thresholding and model selection. The considered methods are the following. The first one is the kernel method, denoted **K**, consisting in a basic cross-validation choice of a global bandwidth with a Gaussian kernel. The second method requires a complex preprocessing of the data based on binning. The good theoretical properties of such a transformation when the support is known have been recently proved in [15]. Observations  $X_1, \dots, X_n$  are first rescaled and centred by an affine transformation denoted  $T$  such that  $T(X_1), \dots, T(X_n)$  lie in  $[0, 1]$ . We denote  $f_T$  the density of the data induced by the transformation  $T$ . We divide the interval  $[0, 1]$  into  $2^{b_n}$  small intervals of size  $2^{-b_n}$ , where  $b_n$  is an integer, and count the number of observations in each interval. We apply the root transform due to [15] and the universal hard individual thresholding rule on the coefficients computed with the DWT Coiflet-basis filter. We finally apply the unroot transform to obtain an estimate of  $f_T$  and the final estimate of the density is obtained by applying  $T^{-1}$  combined with a spline interpolation. This method is denoted **RU**. The last method is also support dependent. After rescaling as previously the data, we estimate  $f_T$  by Willett and Nowak's algorithm [74], which is a classical model selection based approach. The final estimate of the density is obtained by applying  $T^{-1}$ . This method is denoted **WN**.

Our practical method is implemented in the Haar basis (method **H**) and in a smoother biorthogonal basis (method **S**) with the choice  $\gamma = 1$  (see the last chapter on calibration). Moreover we have also implemented the choice  $\gamma = 0.5$  in the smoother basis **S\***.

The density,  $g_d$ , consists in a mixture of two standard Gaussian densities:

$$g_d = \frac{1}{2} \mathcal{N}(0, 1) + \frac{1}{2} \mathcal{N}(d, 1),$$

where  $\mathcal{N}(\mu, \sigma)$  represents the density of a Gaussian variable with mean  $\mu$  and standard deviation  $\sigma$ . The parameter  $d$  varies in  $\{10, 30, 50, 70\}$  so that we can see the curse of support on the quality of estimation.

We generate  $n$ -samples of these densities, with  $n = 1024$  and we numerically compute for each estimator  $\hat{f}$  the ISE, i.e.  $\int_{\mathbb{R}} (f - \hat{f})^2$ .

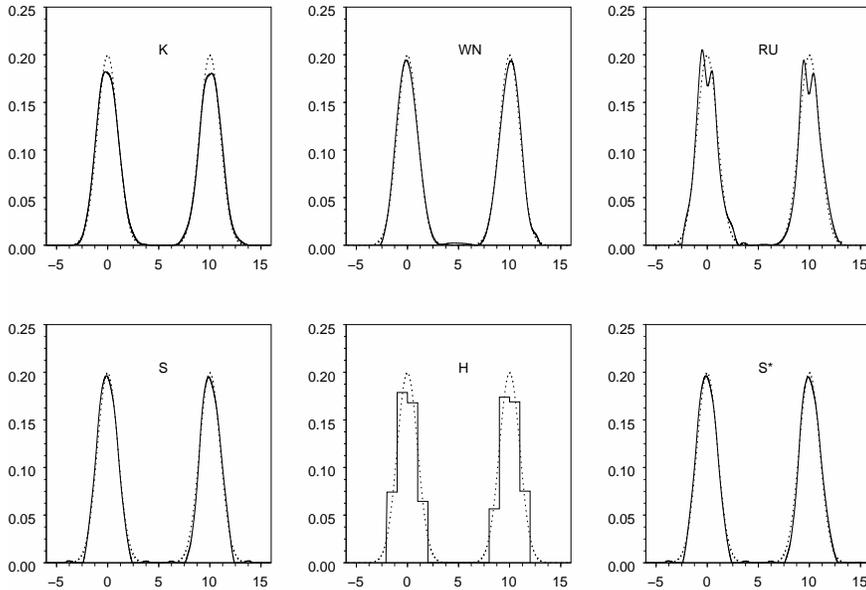


Figure 3.3: Reconstruction of  $g_d$  (true: dotted line, estimate: solid line) for the 6 different methods for  $d = 10$

Figure 3.3 shows the reconstructions for  $d = 10$  and Figure 3.4 for  $d = 70$ . In the sequel, the method **RU** is implemented with  $b_n = 5$ , which is the best choice for the reconstruction with  $d = 10$ . All the methods give satisfying results for  $d = 10$ . When  $d$  is large, the rescaling and binning preprocessing leads to a poor regression signal which makes the regression thresholding rules non convenient, as illustrated by the method **RU** with  $d = 70$ . Reconstructions for **K**, **WN**, **S** and **S\*** seem satisfying but a study of the ISE of each method (see Figure 3.5) reveals that both support dependent methods (**RU** and **WN**) have a risk that increases with  $d$ . On the contrary, methods **K** and **S** are the best ones and more interestingly their performance is remarkably stable (the boxsize is quite small) and the result does not vary with  $d$ . This robustness is also true for **H** and **S\***. **S\*** is a bit under smoothing. Finally note that, for large  $d$ , **H** is even better than **RU** despite the inappropriate choice of the Haar basis.

Of course kernel methods become poor when the signal is more spiky, as it always does (see [63] for more details). Nevertheless, one sees that the most sophisticated methods with support assumption fail when the support becomes larger. Similar simulations have been done with heavy tailed signals when the tail grows.

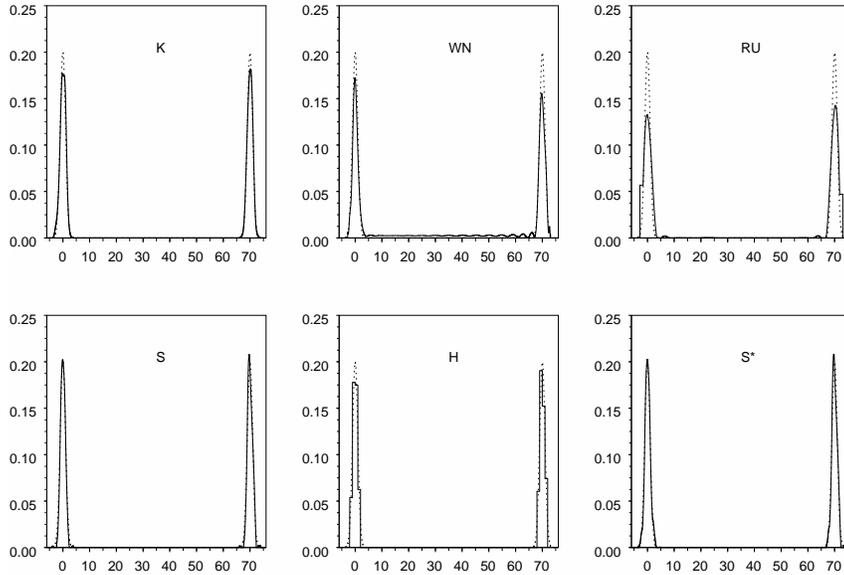


Figure 3.4: Reconstruction of  $g_d$  (true: dotted line, estimate: solid line) for the 6 different methods for  $d = 70$

To conclude this chapter let us just claim again that it is not at all armless to use support dependent methods. This may change the minimax rate of convergence drastically, and in practice the curse of support is visible too. If interesting parts of the signal are too far away from each other, any classical support dependent method will fail, since basically after rescaling, the signal becomes highly irregular.

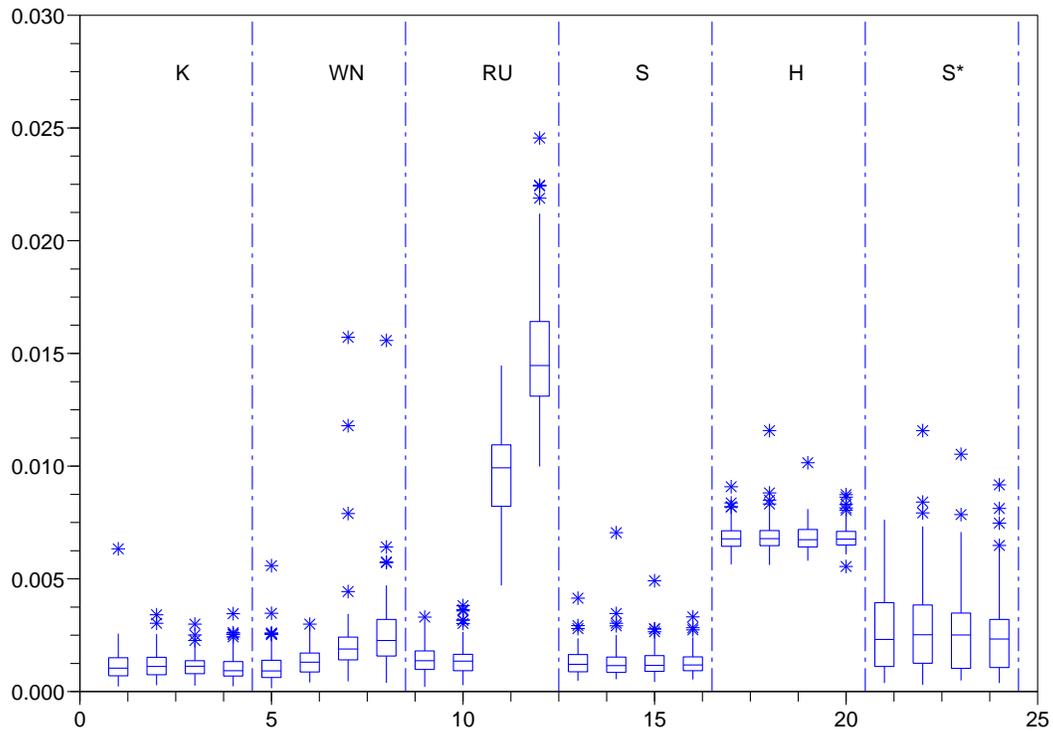


Figure 3.5: Boxplots of the ISE for  $g_d$  over 100 simulations for the 6 methods and the 4 different values of  $d$ . A column, delimited by dashed lines, corresponds to one method (respectively **K**, **WN**, **RU**, **S**, **H**, **S\***). Inside this column, from left to right, one can find for the same method the boxplots of the ISE for respectively  $d = 10, 30, 50$  and  $70$ .

# Chapter 4

## Target functions

### 4.1 Estimation, Maxisets and Minimax point of view

From a pure practical point of view, what is nice with the thresholding estimator built in [61] is the shape of the estimator itself. For easier comparison, let us focus on the estimation by the Haar basis (method **H**) in Figures 3.3 and 3.4. This is still an histogram estimator that precisely say "zero" when the density seems negligible *ie* exactly what the Islands method does as well in the Hawkes framework. However, in the Islands method, as said previously, one needs to know the maximal size of the support to implement the method, and the size of support is linked to the size of each interval in  $\Gamma$ . That means that the scale of estimation, *ie* the size of each interval in the partition, is fixed : one cannot go beyond and reach a smallest scale. If theoretically speaking, it isn't a big deal ( $\Gamma$  grows with  $T$  anyway), in practice the limitation comes from the computer. After  $|\Gamma| = 26$  it is not computationally possible to obtain the resulting estimator. The main reason is because one needs to stock all the estimators for all possible choice of  $m$ , *ie*  $2^{26}$  possibilities for Islands! On the contrary it is not a big deal to produce thresholding estimators with support  $[0, 70]$  and smallest possible scale  $2^{-10}$ ... The histogram may have until  $2^{16}$  small intervals and the computation is still fast. The outstanding advantage of thresholding over model selection method is its fast computation algorithm.

Unfortunately, in the Hawkes model, one cannot adapt the thresholding method because we do not have access to an unbiased estimate of the coefficients  $\beta_\lambda$ .

Nevertheless, assume that such a method may work for Hawkes process (see also the appendix on future work), then the shape of the estimator **H** in Figures 3.3 and 3.4 is exactly what we would like to obtain in order to furnish a good interpretation of the results: disjoint localized spikes with eventually refined scales that can be reached without zooming as for Islands. And the question now is, what are these functions ? One cannot continue to say "this estimator looks ugly, this one is nicer". There must be a good interpretation of such functions via the approximation theory.

In my opinion, the most exciting concept when one arrives here is the maxiset notion introduced in statistics by Kerkyacherian and Picard in [42]. They turn the question upside down. Basically the reasoning is as follows "If you like this estimator and what it produces, then I can tell you what kind of functions it is able to reach". Indeed, there has been a lot of works computing minimax rates of convergence over various Besov or Hölder spaces, arguing whether the  $\mathbb{L}_2$ -norm, the  $\mathbb{L}_\infty$ -norm or the Kullback-Leibler distance, is the most pertinent loss. But the regularity space was more or less fixed, only some variation in the parameters were allowed. But do we know what a "typical" Besov function looks like in

practice? More importantly have those functions usually something to do with the set of real target functions, our estimators would have to reach in practice? What Kerkyacherian and Picard introduced, changes our point of view : they describe approximation spaces via a statistical approach.

So now the question is: what is the maxiset of the thresholding procedure described previously?

Let us first describe the maxiset approach which is classical in approximation theory. For this purpose, let us assume that we are given  $f^*$  an estimation procedure. The maxiset study of  $f^*$  consists in deciding the accuracy of  $f^*$  by fixing a prescribed rate  $\rho^*$  and in pointing out all the functions  $f$  such that  $f$  can be estimated by the procedure  $f^*$  at the target rate  $\rho^*$ . The maxiset of the procedure  $f^*$  for this rate  $\rho^*$  is the set of all these functions. More precisely, we restrict our study to the signals belonging to  $\mathbb{L}_1 \cap \mathbb{L}_2$  and we set:

**Definition 2.** *Let  $\rho^* = (\rho_n^*)_n$  be a decreasing sequence of positive real numbers and let  $f^* = (f_n^*)_n$  be an estimation procedure. The maxiset of  $f^*$  associated with the rate  $\rho^*$  and the  $\mathbb{L}_2$ -loss is*

$$MS(f^*, \rho^*) = \left\{ f \in \mathbb{L}_1 \cap \mathbb{L}_2 : \sup_n \{ (\rho_n^*)^{-2} \mathbb{E} \|f_n^* - f\|^2 \} < +\infty \right\}.$$

Maxiset results have been established and extensively discussed in different settings for many classes of estimators and for various rates of convergence. Let us just mention in our framework, Autin in [4] who derives maxisets for thresholding rules with data-driven thresholds for density estimation.

One of the goals of [61] is to investigate maxisets for the thresholding estimator  $\tilde{f}_\gamma = (\tilde{f}_{n,\gamma})_n$  and we only focus on rates of the form  $\rho_s = (\rho_{n,s})_n$ , where  $0 < s < \frac{1}{2}$  and for any  $n$ ,

$$\rho_{n,s} = \left( \frac{\ln n}{n} \right)^s.$$

So, in the sequel, we investigate:

$$MS(\tilde{f}_\gamma, \rho_s) = \left\{ f \in \mathbb{L}_1 \cap \mathbb{L}_2 : \sup_n \left\{ \left( \frac{\ln n}{n} \right)^{-2s} \mathbb{E} \|\tilde{f}_{n,\gamma} - f\|^2 \right\} < \infty \right\},$$

where  $\tilde{f}_\gamma = (\tilde{f}_{n,\gamma})_n$  is the thresholding procedure described previously. To characterize maxisets of  $\tilde{f}_\gamma$ , we set for any  $\lambda \in \Lambda$ ,  $\sigma_\lambda^2 = \int \varphi_\lambda^2(x) f(x) dx$  and we introduce the following spaces.

**Definition 3.** *We define for all  $R > 0$  and for all  $0 < s < \frac{1}{2}$ ,*

$$W_s = \left\{ f = \sum_{\lambda \in \Lambda} \beta_\lambda \tilde{\varphi}_\lambda : \sup_{t>0} \left\{ t^{-4s} \sum_{\lambda \in \Lambda} \beta_\lambda^2 \mathbf{1}_{|\beta_\lambda| \leq \sigma_\lambda t} \right\} < \infty \right\},$$

and for any sequence of spaces  $\mathcal{G} = (\Gamma_n)_n$  included in  $\Lambda$ , we also define

$$B_{2,\mathcal{G}}^s = \left\{ f = \sum_{\lambda \in \Lambda} \beta_\lambda \tilde{\varphi}_\lambda : \sup_n \left\{ \left( \frac{\ln n}{n} \right)^{-2s} \sum_{\lambda \notin \Gamma_n} \beta_\lambda^2 \right\} < \infty \right\}.$$

These spaces just depend on the coefficients of the biorthogonal wavelet expansion. The spaces  $W_s$  can be viewed as weak versions of classical Besov spaces, hence they are denoted in the sequel Weak Besov spaces. An interested reader may look at [64] for "pictures" of typical functions in Weak Besov spaces. Note that if for all  $n$ ,

$$\Gamma_n = \{\lambda = (j, k) \in \Lambda : j \leq j_0\}$$

with

$$2^{j_0} \leq \left(\frac{n}{\ln n}\right)^c < 2^{j_0+1}, \quad c > 0$$

then,  $B_{2,\mathcal{G}}^s$  is the classical Besov space  $\mathcal{B}_{2,\infty}^{c-1s}$  if the reconstruction wavelets are regular enough. We have the following result.

**Theorem 5** (RB Rivoirard 2010). *Let us fix two constants  $c \geq 1$  and  $c' \in \mathbb{R}$ , and let us define for any  $n$ ,  $j_0 = j_0(n)$  the integer such that  $2^{j_0} \leq n^c(\ln n)^{c'} < 2^{j_0+1}$ . Let  $\gamma > c$ . Then, the procedure defined in (3.2.6) with the sequence  $\mathcal{G} = (\Gamma_n)_n$  such that*

$$\Gamma_n = \{\lambda = (j, k) \in \Lambda : j \leq j_0\}$$

*achieves the following maxiset performance: for all  $0 < s < \frac{1}{2}$ ,*

$$MS(\tilde{f}_\gamma, \rho_s) = B_{2,\mathcal{G}}^s \cap W_s.$$

The maxiset of  $\tilde{f}_\gamma$  is characterized by two spaces: a Weak Besov space that is directly connected to the thresholding nature of  $\tilde{f}_\gamma$  and the space  $B_{2,\mathcal{G}}^s$  that handles the coefficients that are not estimated, which corresponds to the indices  $j > j_0$ .

Hence the targets function are the intersection of Weak Besov bodies and classical Besov bodies. Remark that the previous result enables us to affirm for instance that one can estimate unbounded functions at convenient rate, because they belong to the maxiset (see [61] for more details). One can add that actually the minimax rate of convergence on the maxiset  $MS(\tilde{f}_\gamma, \rho_s)$  is exactly  $\rho_s$  (at least when the Haar basis is considered) and that the logarithmic term is indeed unavoidable when one uses thresholding rules. Informally speaking, if one likes thresholding estimators because of their "shape", one naturally wants to estimate target functions in their maxisets. As they are adaptive minimax over their maxiset (see [61] for more details), one cannot really improve them, hence let us use when possible thresholding estimators!

## 4.2 Tests

Let us just quickly go back again to Section 3.1. We have seen that when the Islands strategy is performed in practice on the genomic data with the Hawkes model, we are tempted by using the method to zoom in. Of course this leads to questions such as, can we effectively test that  $h$  is null and/or test that  $h$  is null after 5000 bases? There is no answer yet, because actually the question, even for the simpler Poisson process, was not that easy to solve. There exists a parametric local approach due to Dachian and Kutoyants in [22] which is quite far from an adaptive procedure where the alternatives in terms of  $h$  may be quite intricate. Let us just focus on the first question. If  $h$  is null, this amounts to test if the process  $N$  is an homogeneous Poisson process, the alternative being that its intensity is given by

$$\Lambda(t) = \nu + \sum_{x \in N, x < t} h(t - x),$$

for a non zero function  $h$ .

Now let us simplify the problem a bit by considering this intensity for the alternative:

$$\Lambda(t) = \nu + h(t),$$

where  $h$  is non zero and belongs to the set of possible target functions we described above. Can we provide an adaptive test that is able to detect such spiky alternatives? This is the subject of the joint work with M. Fromont and B. Laurent in [29].

We still observe a Poisson process  $N$  with unknown intensity  $f(x)$  wrt  $ndx$  on  $[0, 1]$ . For testing, the bounded support assumption is not a problem since it is part of our null hypothesis. The fact that the interval is  $[0, 1]$  just simplifies some notations. Note that a homogeneous Poisson process on the whole real line would have an infinite number of points, hence it cannot be observed in its totality.

We assume that  $\|f\|_\infty < \infty$  and that one can decompose  $f$  on the Haar basis :

$$f = \alpha_0 \varphi_0 + \sum_{j \in \mathbb{N}} \sum_{k=0}^{2^j-1} \alpha_{(j,k)} \varphi_{(j,k)},$$

with  $\varphi_0(x) = \mathbf{1}_{[0,1]}(x)$  and  $\varphi_{(j,k)}(x) = 2^{j/2} \psi(2^j x - k)$  where  $\psi(x) = \mathbf{1}_{[0,1/2]}(x) - \mathbf{1}_{[1/2,1]}(x)$ . Let us state our test more precisely: we want to test  $H_0$ : " $f$  is constant" (ie  $N$  is homogeneous) against  $H_1$ : " $f$  is not constant". Adaptive testing procedures consist in designing a test that will be powerful for a wide class of possible spaces as alternatives. Since testing is not that usual in adaptive problems, let us point out the main differences with (adaptive) estimation.

First let us understand what happens on one finite vectorial subspace. Let  $m \subset \{(j, k), j \geq 0, k = 0, \dots, 2^j - 1\}$  and  $S_m = \text{Span}(\varphi_0, \varphi_\lambda, \lambda \in m)$ . The dimension of  $S_m$  is denoted  $D_m$ . The least-square or projection estimator (see (2.1.2)) can be rewritten with this formalism as

$$\hat{f}_m = \hat{\beta}_0 \varphi_0 + \sum_{\lambda \in m} \hat{\beta}_\lambda \varphi_\lambda,$$

with as usual  $\hat{\beta}_\lambda = \frac{1}{n} \int_{[0,1]} \varphi_\lambda(x) dN_x$ . Let  $f_m$  be the orthogonal projection of  $f$  on  $S_m$ , then the risk of this estimator satisfies

$$\mathbb{E}(\|f - \hat{f}_m\|^2) \leq \|f - f_m\|^2 + \frac{D_m \|f\|_\infty}{n}, \quad (4.2.1)$$

where  $\|f\|$  represents the classical  $\mathbb{L}_2$  norm on  $[0, 1]$ . When we want to test the homogeneity, we actually want to reject when the distance between  $f$  and  $S_0 = \text{Span}(\varphi_0)$  is too large. This distance can be estimated and the estimate may be used as test statistic. This idea is very old. It has been introduced in the Poisson setting by Watson [73]. The procedure is consequently decomposed as follows:

1. We approximate  $d(f, S_0)^2$  by  $\sum_{\lambda \in m} \alpha_\lambda^2$ .
2. We estimate it in an unbiased way by  $T_m = \sum_{\lambda \in m} T_\lambda$  with

$$T_\lambda = \hat{\beta}_\lambda^2 - \frac{1}{n^2} \int \varphi_\lambda^2 dN.$$

3. Under  $H_0$  the law of  $T_m$  given that  $N_{[0,1]} = K$  is free of  $f$ , so there exists  $t_{m,\alpha}^{(K)}$  such that

$$\mathbb{P}(T_m > t_{m,\alpha}^{(K)} | N_{[0,1]} = K) \leq \alpha.$$

4. We consequently reject when  $T_m > t_{m,\alpha}^{(N)} := t_{m,\alpha}^{(N_{[0,1]})}$ .
5. One possible choice is  $t_{m,\alpha}^{(K)} = q_{m,\alpha}^{(K)}$  the  $1 - \alpha$  quantile of the conditional distribution of  $T_m$ .

The performance of the test is measured in term of separation distance, i.e. the question is: under  $H_1$ , how far from  $S_0$  should  $f$  be to obtain  $\mathbb{P}(\text{accept } H_0) \leq \beta$ ?

If  $\mathbb{P}(t_{m,\alpha}^{(N)} \geq A_{m,\alpha,\beta}) \leq \beta/3$ , and if

$$d^2(f, S_0) \geq \|f - f_m\|^2 + \square_{\beta, \|f\|_\infty} \frac{\sqrt{D_m}}{n} + A_{m,\alpha,\beta},$$

then, under suitable assumptions, the error of second kind is less than  $\beta$  (see the precise version with slightly different notations in Theorem 4 of [29]). Remark that with respect to the estimation part (see (4.2.1)),  $\sqrt{D_m}$  is replacing  $D_m$ . Tests are consequently usually thought to be easier than the corresponding estimation: the separation distance seems smaller than the risk. However the presence of  $A_{m,\alpha,\beta}$  is crucial.

If  $t_{m,\alpha}^{(N)} = q_{m,\alpha}^{(N)}$ , one can prove, using exponential inequalities for degenerate U-statistics of order 2 - which is the case for  $T_m$  - that

$$A_{m,\alpha,\beta} = \square_{\beta, \|f\|_\infty} \left[ \frac{\sqrt{D_m \log(\alpha^{-1})}}{n} + \frac{\log(\alpha^{-1})}{n} + \frac{E_m \log^2(\alpha^{-1})}{n^2} \right],$$

where  $E_m = \sum_{j/(j,k) \in m} 2^j$  may be much larger than  $D_m$ . Once again, as for adaptive estimation, concentration inequalities are the fundamental tool to build adaptive test. Indeed, the dependency of  $A_{m,\alpha,\beta}$  in  $\alpha$  will be crucial, once we want to combine those tests (ie we apply several tests at once and we accept or reject depending on the outcome of all the tests). The smaller the dependency in  $\alpha$  is, the larger the number of tests one can combine is. In particular, it is fundamental to use exponential deviations since the dependency in  $\alpha$  will consequently be logarithmic. Exponential inequalities for U-statistics of independent variables are described in the book of de la Peña and Giné [24]. These upper bounds have been improved but still with unknown constants by Giné, Latala and Zinn [30]. In a joint work with C. Houdré in [35], we derive precise constants in those formula by combining Talagrand's inequality and martingale properties for degenerate U-statistics of order 2. For our precise set-up here, the ingredients also apply to Poisson processes since one can replace Talagrand's inequality by Theorem 1 (see [35] for more details). Let us just mention one of the existing extension of this work due to Adamczak [1], which involves degenerate U-statistics of any order for independent variables but also for processes with independent increments.

Actually a powerful test consists in having the best separation distance within a certain subclass of alternatives. If one is given a whole collection of possible subclasses, adaptive testing procedures should achieve the same separation distance (up to some minor losses) without knowing to which subclass the alternative actually belongs.

When one considers, as subclasses, the models  $S_m$ , an adaptive test procedure, as described above, should satisfy an inequality which has essentially the same flavour as the

one obtained in Proposition 1 by adaptive estimators. This is achieved by combining tests, which is quite easy. The most natural approach consists in a model selection approach, but one can actually also consider a thresholding approach. Both model selection and thresholding approaches have already been used to construct adaptive tests by Spokoiny ([68] and [69]) in Gaussian white noise models. Note also the work of Baraud, Huet and Laurent [8] in a Gaussian regression framework. As for the density framework, adaptive tests were proposed by Ingster [38] or Fromont and Laurent [28], using model selection type methods and by Butucea and Tribouley [18] using thresholding type methods.

In our present framework, let us start with the model selection approach. Let  $\mathcal{M}$  be a collection of possible  $m$ 's. Then one rejects  $H_0$  when there exists one  $m \in \mathcal{M}$  such that  $T_m > t_{m, \alpha_m}^{(N)} = q_{m, \alpha_m}^{(N)}$ , where under  $H_0$ ,  $\mathbb{P}(\exists m \in \mathcal{M}, T_m > t_{m, \alpha_m}^{(N)}) \leq \alpha$ . The basic choice for  $\alpha_m$  is the Bonferroni choice ie  $\alpha_m = \alpha/|\mathcal{M}|$ . This allow us to define a *nested test* ie  $\mathcal{M} = \{\Lambda_1, \dots, \Lambda_{\bar{J}}\}$ , with  $\Lambda_J = \{(j, k), j \leq J, k = 0, \dots, 2^j - 1\}$  and  $2^{\bar{J}} \simeq n^2/(\log(n)^2)$  whose separation distance is at least

$$\inf_{J \leq \bar{J}} \left\{ \|f - f_{\Lambda_J}\|^2 + \square_{\alpha, \beta, \|f\|_\infty} \left[ \frac{\sqrt{2^J \bar{J}}}{n} + \frac{\bar{J}}{n} + \frac{2^J \bar{J}^2}{n^2} \right] \right\}.$$

In the same way, the thresholding estimation procedure has a test version which is defined by the following *thresholding test*: one rejects  $H_0$  when there exists  $\lambda \in \Lambda_{\bar{J}}$ , with  $2^{\bar{J}} \simeq n$  such that

$$T_\lambda > q_{\lambda, \alpha/(2^j \bar{J})}^{(N)}.$$

This is equivalent to ask if there exists  $m \subset \Lambda_{\bar{J}}$  such that

$$\sum_{\lambda \in m} T_\lambda = T_m > \sum_{\lambda \in m} q_{\lambda, \alpha/(2^j \bar{J})}^{(N)} = t_{m, \alpha}^{(N)}.$$

Then the separation distance of this test is at least

$$\inf_{\Lambda \subset \Lambda_{\bar{J}}} \left\{ \|f - f_m\|^2 + \square_{\alpha, \beta, \|f\|_\infty} \left[ \frac{D_m \tilde{J}}{n} + \frac{D_m \tilde{J}^2 2^{\bar{J}}}{n^2} \right] \right\}.$$

One can see the difference between both tests by looking at the minimax separation rate over various functional subclasses of alternatives. Without going into details and tedious definitions, let us just cite for the interested reader two fundamental papers on minimax separation distances for the Gaussian set-up: [37] and [5]. In the Poisson setting, the minimax separation rate for classical Besov spaces has been computed by Ingster in [39]. However minimax separation rate for Weak Besov spaces was not known, whatever the set-up (Gaussian, Poisson, density ...), until the present joint work [29]. More precisely, we compute this rate for alternatives in  $\mathcal{B}_{2, \infty}^\delta(R) \cap \mathcal{W}_\gamma(R')$  where the classical Besov body is defined by

$$\mathcal{B}_{2, \infty}^\delta(R) = \left\{ s \geq 0 \mid \forall j \in \mathbb{N}, \sum_{k=0}^{2^j-1} \alpha_{(j,k)}^2 \leq R^2 2^{-2j\delta} \right\}$$

and the weak Besov body is defined by

$$\mathcal{W}_\gamma(R') = \left\{ s \geq 0 \mid \forall t > 0, \sum_{j \in \mathbb{N}} \sum_{k=0}^{2^j-1} \alpha_{(j,k)}^2 \mathbf{1}_{\alpha_{(j,k)}^2 \leq t} \leq R'^2 t^{\frac{2\gamma}{1+2\gamma}} \right\}.$$

Note that  $\mathcal{W}_\gamma$  is a simplified version of  $W_s$  (see Definition 3) with  $s = \gamma/(1 + 2\gamma)$ : one assumes the variance term  $\sigma_\lambda$  to be constant which does not change the rates if one assumes that the intensity has a finite support and that the intensity is lower bounded on its support. Hence the minimax separation rate over  $\mathcal{B}_{2,\infty}^\delta \cap \mathcal{W}_\gamma$  for various  $\delta$  and  $\gamma$  should exhibit a particularly interesting behaviour when  $\delta = \gamma/(1 + 2\gamma)$  which corresponds to the maxiset of the thresholding estimation procedure when  $c = 1$  (see Theorem 5).

**Theorem 6** (Fromont Laurent RB 2010).

- If  $\delta \geq \max(\gamma/2, \gamma/(1 + 2\gamma))$ , then the minimax separation rate for alternatives in  $\mathcal{B}_{2,\infty}^\delta(R) \cap \mathcal{W}_\gamma(R') \cap \mathbb{L}_\infty(R'')$  is

$$\square_{\delta,\gamma,R,R',R'',\alpha,\beta} n^{-\frac{4\delta}{1+4\delta}},$$

which is achieved by the nested test up to a  $\ln \ln n$  term.

- If  $\delta < \gamma/2$  and  $\gamma > 1/2$ , then the minimax separation rate for alternatives in  $\mathcal{B}_{2,\infty}^\delta(R) \cap \mathcal{W}_\gamma(R') \cap \mathbb{L}_\infty(R'')$  is larger than

$$\square_{\delta,\gamma,R,R',R'',\alpha,\beta} \left( \frac{\ln n}{n} \right)^{\frac{2\gamma}{1+2\gamma}},$$

rate which is achieved by the thresholding test when  $\delta \geq \frac{\gamma}{1+2\gamma}$ .

Note that the second rate is also the minimax rate of estimation over those spaces. That means that when Weak Besov bodies are involved and when their regularity is large, it is not easier to test than to estimate. More details and more precise statements may be found in [29].

To conclude this chapter, let us just state that a convenient space of target functions may be the intersection of a classical large Besov space with small regularity and a Weak Besov space whose regularity may be much larger. In the Poisson framework, those kinds of sets are the maxiset of the thresholding estimation procedure, procedure which is adaptive minimax over those sets. The minimax rate of convergence for estimation is actually also the minimax separation distance for testing in certain cases. This means that, if we are interested by those functions, it will be as difficult to test as to estimate. This phenomena was not known, up to our knowledge and is certainly true for other settings such as Gaussian set-up. Finally a thresholding approach allows us to achieve this separation rate of testing, whereas those tests may be completely useless when only classical Besov spaces are considered as alternatives. Indeed, those tests do not only lose a logarithmic factor, the rate is definitely worse, whereas this phenomenon does not appear for thresholding estimators whose loss is only due to a logarithmic factor.

Of course, it is always possible to combine *thresholding tests* and *nested tests*. This combined test is adaptive minimax for  $\delta \geq \gamma/(1 + 2\gamma)$ . These tests also perform really well in practice (see [29] for more details).

# Chapter 5

## Calibration

The final question deals with calibration. Let us start with the Hawkes model. In Theorem 3, a penalty proportional to the dimension of the model divided by  $T$  appears. However the multiplicative constant depends on parameter that cannot be guessed in practice. So what should we do? One can think that Theorem 3 is a theoretical result that guides our intuition but that the right multiplicative constant should depend in practice on the data. In [11], Birgé and Massart investigate the resulting theoretical problem in a homoscedastic Gaussian set-up. In this regression set-up, the theoretical optimal penalty, for, say, the Nested strategy, is  $2\sigma^2 D_m/n$  where  $\sigma^2$  is the variance of the Gaussian noise and  $n$  the total number of observations. One could wonder if an unbiased estimation of  $\sigma^2$  may work. Birgé and Massart actually prove that there exists a minimal penalty  $\sigma^2 D_m/n$ . If one of the models with high dimension in the Nested strategy is less penalized than this benchmark, then the whole model selection method will select a model with too high dimension. This phenomenon can be detected in practice: hence one can guess what the minimal penalty is and then, multiplying by 2, one can obtain the optimal penalty. This study has been reinforced by the recent theoretical and practical results of Arlot in the heteroscedastic set-up [3]. If other frameworks (see [48] in density or [67] for more general set-up) have been very recently studied, such studies heavily rely on very tight exponential inequalities. In particular, it is completely out of reach at this point to obtain such kind of theoretical results for the Hawkes process: actually even the precise shape of the optimal penalty is not known. This Chapter is divided in a numerical study to obtain at least an ad-hoc calibration in the Hawkes model that will work well in practice [62] and in a more theoretical study of the calibration of the thresholding rule in the Poisson case [61], and in the density case [63]. Note that the theoretical calibration results of thresholding rules started with Donoho and Johnstone [25] in their seminal work in the Gaussian set-up. Further attempts usually involve pure model selection and thresholding as an appended result. Here we wanted to calibrate a full data-driven threshold, which corresponds to a penalty that, as is, is usually not used in classical model selection and that cannot be recovered by classical model selection techniques.

### 5.1 Ad hoc methods

Let us start with the numerical study of [62].

### 5.1.1 Compared methods

We implement 3 strategies : Regular, Irregular and Islands. Regular is a slightly different version of Nested: the family  $\mathcal{M}_T$  consists in all the regular partitions of  $[0, A]$  up to a certain level  $N$ , but they are not forced to be dyadic. Indeed, for numerical reason, we are forced to consider  $|\Gamma| \leq 15$  and the Nested family will be in this case much too small.

Since we are looking for a penalty that is inspired by (2.2.7), we compare our penalized methods to the most naive approach, namely the hold-out procedure described below. Moreover, the truncated estimators are designed for minimax theoretical purposes, but of course they depend on parameters ( $H, \dots$  - see (2.2.5)) that cannot be guessed in practice. They also force the estimate of  $h$  to be nonnegative. Therefore in this section we only use non truncated estimators (see (2.2.3)).

**Hold-out** The naive approach is based on the following fact (which can be made completely and theoretically explicit in the self-exciting case). We know that  $\gamma_T$  is a contrast. We would like to select a model  $\hat{m}$  such that  $\hat{s}_{\hat{m}}$  is as good as the best possible  $\hat{s}_m$ . So one way to select a good model  $m$  should be to observe a second independent Hawkes process with the same  $s$  and to compute the minimizer of  $\gamma_{T,2}(\hat{s}_m)$  over  $\mathcal{M}_T$  (where  $\hat{s}_m$  is computed with the first process and  $\gamma_{T,2}$  is our contrast but computed with the second process). However we do not have in practice two independent Hawkes processes at our disposal. But one can cut  $[-A, T]$  in two almost independent pieces. Indeed the points of the process in  $[-A, T/2 - A]$  and in  $[T/2, T]$  can be equal to those of independent stationary Hawkes processes and this with high probability (see [60]). Hence in the sequel whenever the Hold-out estimator is mentioned, and whatever the family  $\mathcal{M}_T$  is, it is referring to the following procedure.

1. Cut  $[-A, T]$  into two pieces:  $H_1$  refers to the points of the process on  $[-A, T/2 - A]$ ,  $H_2$  refers to the points of the process on  $[T/2, T]$ .
2. Compute  $\hat{s}_m$  for all the  $m$  in  $\mathcal{M}_T$  by minimizing the least-square contrast  $\gamma_{T,1}$  on  $S_m$  computed with only the points of  $H_1$ , ie

$$\forall f \in \mathbb{L}_2, \quad \gamma_{T,1}(f) = -\frac{2}{T} \int_0^{T/2-A} \Psi_f(t) dN_t + \frac{1}{T} \int_0^{T/2-A} \Psi_f(t)^2 dt,$$

3. Compute  $\gamma_{T,2}(\hat{s}_m)$  where  $\gamma_{T,2}$  is computed with  $H_2$ , i.e.,

$$\forall f \in \mathbb{L}_2, \quad \gamma_{T,2}(f) = -\frac{2}{T} \int_{T/2+A}^T \Psi_f(t) dN_t + \frac{1}{T} \int_{T/2+A}^T \Psi_f(t)^2 dt,$$

and find  $\hat{m} = \arg \min_{m \in \mathcal{M}_T} \gamma_{T,2}(\hat{s}_m)$ .

4. The Hold-out estimator is defined by  $\tilde{s}^{HO} := \hat{s}_{\hat{m}}$ .

Note that we used all the observed points since the computation of  $\Psi_f(t)$  requires to have the points in  $(t - A, t]$ . Theoretically speaking, the gap should have been a bit larger to ensure almost independence.

**Penalized** Theorem 3 shows that theoretically speaking a penalty of the type  $K(|m| + 1)$  should work. However the theoretical multiplicative constant is not only not computable, it is also too large for practical purpose. So one needs to consider Theorem 3 as a result that guides our intuition towards the right shape of penalty. Therefore we investigate two ways of calibrating the multiplicative constants.

1. The first one follows the conclusions of [11]. In the Regular strategy, there exists at most one model per dimension. If there exists a true model  $m_0$ , then for  $|m|$  large (larger than  $|m_0|$ )  $\gamma_T(\hat{s}_m)$  should behave like  $-k(|m| + 1)$ . So there is a minimal penalty as defined by Birgé and Massart of the form  $\text{pen}_{\min} = k(|m| + 1)$ . In this situation their rule is to take  $\text{pen}(m) = 2 * \text{pen}_{\min}(m)$ .

We find a  $\hat{k}$  by doing a least-square regression for large values of  $|m|$  so that

$$\gamma_T(\hat{s}_m) \simeq -\hat{k}(|m| + 1).$$

Then we take

$$\hat{m} = \operatorname{argmin}_{m \in \mathcal{M}_T} \left\{ \gamma_T(\hat{s}_m) + 2\hat{k}(|m| + 1) \right\},$$

and we define  $\tilde{s}^{\min} := \hat{s}_{\hat{m}}$ .

For the Irregular and Islands strategy, as a preliminary step, we need to find the best data-driven model per dimension i.e.

$$\hat{m}_D = \operatorname{argmin}_{m \in \mathcal{M}_T, |m|=D} \left\{ \gamma_T(\hat{s}_m) \right\}.$$

Then one can plot as a function of  $D$ ,  $\gamma_T(\hat{s}_{\hat{m}_D})$ . In [11], they also obtain another kind of minimal penalty of the form  $\text{pen}_{\min} = k(D + 1)(\log(|\Gamma|/D) + 5)$  when the Irregular strategy is used. But for very small values of  $|\Gamma|$  (as here) we would not see the difference between this form of penalty and the linear form. Moreover theoretically speaking we are not able to justify, even heuristically, such a form of penalty for large values of  $|\Gamma|$ . So we have decided that we will use the same penalty as before even in the Irregular and Islands strategies. That is to say that we find a  $\hat{k}$  by doing a least-square regression for large value of  $D$  so that

$$\gamma_T(\hat{s}_{\hat{m}_D}) \simeq -\hat{k}(D + 1).$$

Then we take

$$\hat{m} = \operatorname{argmin}_{m \in \mathcal{M}_T} \left\{ \gamma_T(\hat{s}_m) + 2\hat{k}(|m| + 1) \right\},$$

and we define  $\tilde{s}^{\min} := \hat{s}_{\hat{m}}$  even for the Irregular and Islands strategies.

2. On the other hand, the choice of  $\hat{m}$  by  $\tilde{s}^{\min}$  was not completely satisfying when using the Islands or Irregular strategies (see the comments on the simulations hereafter). But on the contrast curve:  $D \rightarrow \gamma_T(\hat{s}_{\hat{m}_D})$ , we could see a perfectly clear angle at the true dimension. So we have decided to compute  $-\bar{k} = \frac{\gamma_T(\hat{s}_\Gamma) - \gamma_T(\hat{s}_{\hat{m}_1})}{|\Gamma| - 1}$  and to choose

$$\hat{m} = \operatorname{arg} \min_{m \in \mathcal{M}_T} \left\{ \gamma_T(\hat{s}_m) + \bar{k}(|m| + 1) \right\}.$$

We define  $\tilde{s}^{\text{angle}} := \hat{s}_{\hat{m}}$ . This seems to be a proper automatic way to obtain this angle without having to look at the contrast curve. It is still based on the fact that a multiple of the dimension should work. This has only been implemented for the Irregular and Islands strategies.

Table 5.1 summarizes our 8 different estimators.

We have simulated Hawkes processes with parameters  $(\nu, h)$ , with  $\nu$  in  $\{0.001, 0.002, 0.003, 0.004, 0.005\}$ ,  $h = 0.002\mathbf{1}_{[200,400]}$  having a bounded support in  $(0, 1000]$  (i.e.  $A = 1000$ ) and on a sequence of length  $[-A, T]$  with  $T = 100000$  or  $T = 500000$ .

Methods	Strategy	Selection
1	Regular $N = 15$	minimal penalty $\tilde{s}^{min}$
2	Irregular $ \Gamma  = 15$	angle method $\tilde{s}^{angle}$
3	Irregular $ \Gamma  = 15$	minimal penalty $\tilde{s}^{min}$
4	Islands $ \Gamma  = 15$	angle method $\tilde{s}^{angle}$
5	Islands $ \Gamma  = 15$	minimal penalty $\tilde{s}^{min}$
6	Regular $N = 15$	Hold-Out $\tilde{s}^{HO}$
7	Irregular $ \Gamma  = 15$	Hold-Out $\tilde{s}^{HO}$
8	Islands $ \Gamma  = 15$	Hold-Out $\tilde{s}^{HO}$

Table 5.1: Table of the different methods.

### 5.1.2 Results

We call *Risk* of an estimator the Mean Square Error of this estimator over 100 simulations, i.e. we compute for each simulation  $\|s - \hat{s}\|^2$  and next we compute the average over 100 simulations. Note that with the range of our parameters, the error of estimation of  $\nu$  will be really negligible with respect to the error of estimation for  $h$ , so that  $\|s - \hat{s}\|^2 \simeq \int_0^A (h - \hat{h})^2$ .

Figure 5.1 gives the *Risk* of our estimators for  $h = 0.5 * f_1$  for various  $\nu$  and  $T$ . We first clearly see that the risk decreases when  $T$  increases whatever the method. There seems also to be a slight improvement when  $\nu$  becomes larger, tending to prove that, if the mean total number of points  $\mathbb{E}(N[0, T]) = \nu T / (1 - p)$  grows, the estimation is improved – at least in our range of parameters. A study in terms of  $p = \int h$  can be found in [62].

We see that the "best methods" are Methods 1, 2 and 4, i.e. the Regular strategy with minimal penalty and the Irregular and Islands Strategies with the angle method. For the Irregular and Islands Strategies, the minimal penalty seems to behave like the Hold-out Strategies. This fact is confirmed in terms of oracle ratio too (see [62] for more details).

Finally Figure 5.2 shows the resulting estimators of Methods 1, 2 and 4 on one simulation. In particular, before penalizing, note that one clearly sees an angle on the contrast curve at the true dimension for the Irregular and Islands Strategies and that penalizing by the angle method (Methods 2 and 4) gives an automatic way to find the position of this angle.

Hence it seems that the model selection method for Hawkes processes clearly behaves differently when a complex family is involved, with respect to the Gaussian case. The contrast curve (see Figure 5.2) is so flat after the true dimension that any "slope" heuristic will fail to detect the angle whereas the angle method does. Note also that even for Gaussian model selection the problem of optimal constant for complex families has not been solved, up to our knowledge. For further illustrations, see [62]. The basic conclusion is that Method 4 (Islands + Angle method) provides a nice way to select a sparse support and to estimate spikes of various localisation and size. This is the method used to derive Figure 3.2. In the next section, we will further investigate one special complex model selection, which is the one associated to the threshold method implemented in [61] and [63].

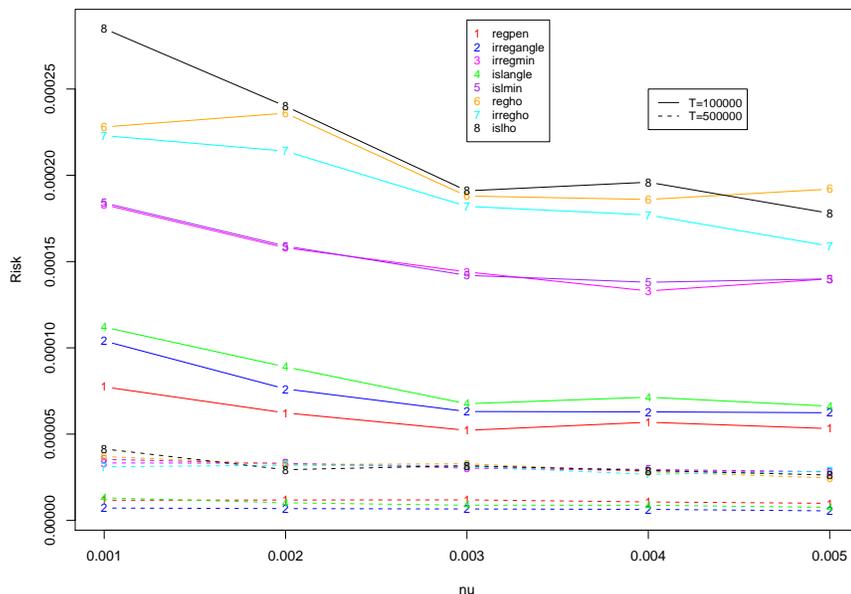


Figure 5.1: Risk of the 8 different methods for  $h = 0.5 * f_1$  for different values of  $\nu$  and  $T$ .

## 5.2 Data-driven thresholds

Now let us go back to section 3.2 and let us rewrite the threshold here :

$$\eta_{\lambda,\gamma} = \sqrt{2\gamma\tilde{V}_\lambda \ln n} + \frac{\gamma \ln n}{3n} \|\varphi_\lambda\|_\infty.$$

First one can notice that there is no dependency in unknown parameters depending on the signal with respect to for instance the penalty of Theorem 3. This is because we have been able to estimate very accurately the variance of  $\hat{\beta}_\lambda$  by  $\tilde{V}_\lambda$  (see [61] for more details). This cannot usually be done by standard model selection methods which will make appear at the very least an estimator of the supremum of  $f$  as in Proposition 1, when one uses least-square contrast. If one uses log-likelihood criteria, one can also obtain a penalty that does not depend on this supremum, but the price is to assume the existence of a lower bound on the intensity (see [74]).

Next, one can ask how should we choose  $\gamma$ ? Theorem 4 tells us that if  $2^{j_0} = n$  then one should take  $\gamma > 1$ . Is there a minimal threshold as well? Here "minimal" should be understood in the sense of Birgé and Massart in [11].

We are able to prove the following result.

**Theorem 7** (RB Rivoirard, 2010). *Let  $f = \mathbf{1}_{[0,1]}$ . If  $\gamma < 1$  then there exists  $\delta < 1$  not dependent of  $n$  such that*

$$\mathbb{E} \|\tilde{f}_{n,\gamma}^H - f\|_2^2 \geq \frac{c}{n^\delta},$$

where  $c$  is a constant.

A similar lower bound exists in the density setting [63]. Since the result is proved for the Haar basis, for which  $\mathbf{1}_{[0,1]}$  is the simplest signal one can imagine, one is forced to

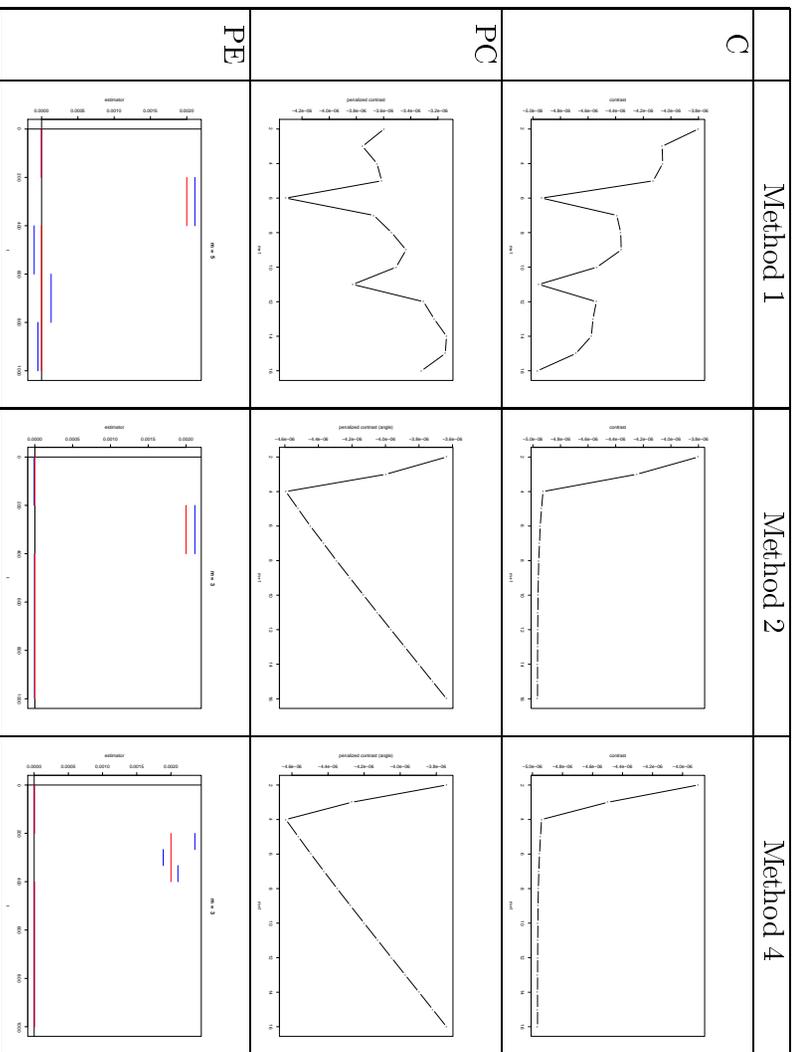


Figure 5.2: Contrast (C) and penalized contrast (PC) as a function of the dimension for the three favourite methods on one simulation with  $T = 500000$ ,  $\nu = 0.001$  and  $h = 0.5 * f_1$ . The chosen estimators (PE) are in blue whereas the function  $h = 0.5 * f_1$  is in red.

conclude that one cannot use  $\gamma < 1$ . We were not able to prove theoretically that 1 is optimal, however in the Poisson case, we know that one should take  $\gamma \in [1, 12]$  (see [61] for more details).

There exists one remaining gap : how do we choose  $\gamma$  in  $[1, 12]$  ? Once again, we can only conclude via a simulation study. I reproduce here the simulation study of [63] in the density setting for a slightly different version of the threshold (3.2.4), that is essentially equivalent. A more thorough study has been done for the Poisson process in [61].

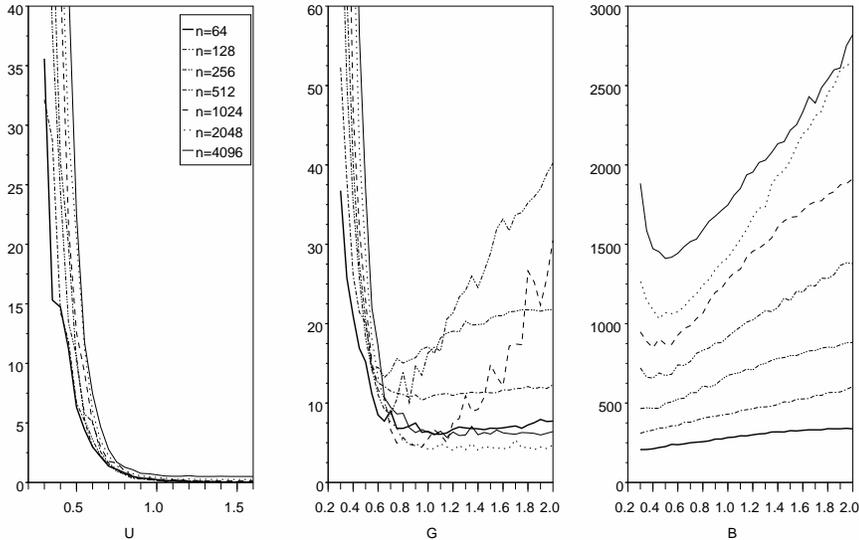


Figure 5.3:  $n \times MISE_n(\gamma)$  for **(U)**  $f = \mathbf{1}_{[0,1]}$  (the Haar basis is used) ; **(G)**  $f$  is the Gaussian density with mean 0.5 and standard deviation 0.25 (the Spline basis is used) ; **(B)**  $f$  is the renormalized Bumps signal (the Spline basis is used)

First we simulate 1000  $n$ -samples of density  $f = \mathbf{1}_{[0,1]}$ . We estimate  $f$  by our method using the Haar basis. For any  $\gamma$ , we have computed  $MISE_n(\gamma)$  i.e. the average over the 1000 simulations of  $\|\tilde{f}_{n,\gamma} - f\|_2^2$ . On the left part of Figure 5.3 **(U)**,  $MISE_n(\gamma) \times n$  is plotted as a function of  $\gamma$  for different values of  $n$ . Note that when  $\gamma > 1$ ,  $MISE_n(\gamma)$  is null meaning that our procedure selects just one wavelet coefficient, the one associated to  $\psi_{-1,0} = \mathbf{1}_{[0,1]}$ ; all others are equal to zero. This fact remains true for a very large range of values of  $\gamma$ . This plateau phenomenon is also noticed in the Poisson framework (see [61]). However as soon as  $\gamma < 1$ ,  $MISE_n(\gamma) \times n$  is positive and increases when  $\gamma$  decreases. It also increases with  $n$  tending to prove that  $MISE_n(\gamma) \gg 1/n$  for  $\gamma < 1$ . This is completely coherent with Theorem 7. We consider two other density functions  $f$ . The first one is the density of a Gaussian variable whose results appear in the middle part of Figure 5.3 **(G)** and the second one is the renormalized Bumps signal<sup>1</sup> whose results appear in the right part of Figure 5.3 **(B)**. In both cases we computed  $\tilde{f}_{n,\gamma}$  with a smoother

<sup>1</sup> The renormalized Bumps signal is a very irregular signal that is classically used in wavelet analysis. It is here renormalized so that the integral equals 1 and it can be defined by  $\left( \sum_j g_j \left( 1 + \frac{|x - p_j|}{w_j} \right)^{-4} \right) \frac{\mathbf{1}_{[0,1]}}{0.284}$  with

basis than the Haar basis, whose precise description is available in [61]. We computed the associate  $MISE_n(\gamma)$  over 100 simulations. Note that for the Bumps signal, there is no plateau phenomenon and that the best choice for  $\gamma$  is  $\gamma = 0.5$  as soon as the highest level of resolution,  $j_0(n)$  is high enough to capture the irregularity of the signal. If  $n$  is too small, the best choice is to keep all the coefficients. Recall that  $\bar{m}$  is defined by (3.2.8). One can actually identify two behaviours: either the oracle  $\hat{f}_{\bar{m}}$  is close to  $f$  and the best possible choice is  $\gamma \simeq 1$  with a plateau phenomenon, or the oracle  $\hat{f}_{\bar{m}}$  is far from  $f$  and it is better to take a smaller  $\gamma$  (for instance  $\gamma = 0.5$ ). The Gaussian density (**G**) exhibits both behaviors. For large  $n$  ( $n \geq 1024$ ), there is a plateau phenomenon around  $\gamma = 1$ . But for smaller  $n$ , the oracle  $\hat{f}_{\bar{m}}$  is not accurate enough and taking  $\gamma = 0.5$  is better. Note finally that the choice  $\gamma = 1$ , is the more robust with respect to both situations.

Let us just conclude by saying that after a calibration step that is partly theoretical and partly a simulation study, we are able to propose convenient methods to estimate, on the one hand, Hawkes interaction function with a given bound on the support and, on the other hand, Poisson intensity (or density) on the whole real line.

Note that testing procedures are by nature data-driven: before even asking what the power of the test is, one has to propose a test that only depends on the data and that has a level  $\alpha$  ! Hence the calibration of test mainly relies on how we can provide a test of precise size  $\alpha$ . Extensive Monte-Carlo methods actually do the job even if the accuracy of the Monte-Carlo method is not taken into account yet (see [29] for more details).

---

p	=	[	0.1	0.13	0.15	0.23	0.25	0.4	0.44	0.65	0.76	0.78	0.81	]
g	=	[	4	5	3	4	5	4.2	2.1	4.3	3.1	5.1	4.2	]
w	=	[	0.005	0.005	0.006	0.01	0.01	0.03	0.01	0.01	0.005	0.008	0.005	]

# Appendix A

## Presented papers

- Reynaud-Bouret, Patricia *Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities*. Probab. Theory Related Fields **126** (1), 103–153 (2003).
- Houdré, Christian ; Reynaud-Bouret, Patricia *Exponential inequalities, with constants, for U-statistics of order two*. Stochastic inequalities and applications, Progr. Probab., 56 Birkhäuser, Basel, 55–69 (2003).
- Reynaud-Bouret, Patricia *Compensator and exponential inequalities for some suprema of counting processes*. Statistics and Probability Letters, **76**(14), 1514–1521 (2006).
- Reynaud-Bouret, Patricia *Penalized projection estimators of the Aalen multiplicative intensity*. Bernoulli, **12**(4), 633–661 (2006).
- Reynaud-Bouret, Patricia ; Roy, Emmanuel *Some non asymptotic tail estimates for Hawkes processes*. Bulletin of the Belgian Mathematical Society-Simon Stevin, **13**(5), 883–896 (2007).
- Houdré, Christian ; Marchal, Philippe ; Reynaud-Bouret, Patricia *Concentration for norms of infinitely divisible vectors with independent components*. Bernoulli, **14**(4), 926–948 (2008).
- Reynaud-Bouret, Patricia ; Rivoirard, Vincent *Near optimal thresholding estimation of a Poisson intensity on the real line*. Electronic Journal of Statistics, **4**, 172–238 (2010).
- Reynaud-Bouret, Patricia ; Schbath, Sophie *Adaptive estimation for Hawkes processes; application to genome analysis*. Ann. Statist., **38**(5), 2781–2822 (2010).
- Fromont, Magalie ; Laurent, Béatrice ; Reynaud-Bouret, Patricia *Adaptive test of homogeneity for a Poisson process*. to appear in Annals of IHP.
- Reynaud-Bouret, Patricia ; Rivoirard, Vincent ; Tuleau-Malot, Christine *Adaptive density estimation: a curse of support?* J. Statist. Plann. Inference, **141**, 115–139 (2011).

# Appendix B

## Future work

There are a lot of questions that have not been solved yet. I give just here a non exhaustive list of my current work and open questions.

- Actually, the Hawkes model as initiated in genomic data by [31] is multivariate. If theoretically speaking, the model selection may work to some extent, in practice the computer limitations make this method useless. One way to go further is to try to use threshold estimators in a close, simplified framework. This is the Ph.D. subject, that V. Rivoirard (Dauphine) and I proposed to our student, Laure Sansonnet. Another approach will be to use the Lasso method. This is a joint work with N. R. Hansen (Copenhagen) and V. Rivoirard. We hope to be able to calibrate the method at least in practice.
- Another assumption, which is quite delicate to remove is the stationary assumption of the Hawkes process. Indeed a non stationary version may model neuronal activity. We would like to tackle this problem with C. Tuleau-Malot, Franck Grammont and Yann Bouret (Nice).
- The Poisson process structure makes things easier for testing than even the classical density iid framework. M. Fromont (Rennes), B. Laurent (Toulouse) and I are building an adaptive test of  $H_0$ : "the Poisson processes,  $N_1$  and  $N_2$ , have same intensity" versus "they have not", test whose power only depends on the difference between both intensities.
- Adaptive tests for Hawkes processes are still an open question.
- The construction of maxisets for testing procedures is also a natural question that we would like to solve with V. Rivoirard (Dauphine).
- Calibration in a theoretical way is a very vast subject with many open questions and not only in point processes theory. One aspect that I'm trying to better understand, is the link between practical adaptive statistical procedures and practical analytical procedures, both of them having to face this calibration problem. A current work with M. Doumic (INRIA Rocquencourt), M. Hoffmann (ENSAE), V. Rivoirard (Dauphine) focuses on the adaptive statistical answer to an inverse problem that has already been solved in an analytical way on size structured populations (see [55, 26]).

# Bibliography

- [1] Adamczak, R. *Moment Inequalities for U-statistics*. Ann. Probab. **34** (6), 2288–2314 (2006).
- [2] Andersen, P. K., Borgan, Ø., Gill, R. D., Keiding, N. *Statistical models based on counting processes*. Springer Series in Statistics (1993).
- [3] Arlot, S. *Model selection by resampling penalization*. E.J.S. **3**, 557–624 (2009).
- [4] Autin, F. *Maxiset for density estimation on  $\mathbb{R}$* . Mathematical Methods of Statistics, **15**(2), 123–145 (2006).
- [5] Baraud, Y. *Non asymptotic minimax rates of testing in signal detection*. Bernoulli, **8**, 577–606 (2002).
- [6] Baraud, Y. *A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression*. to appear in Bernoulli.
- [7] Baraud, Y., Birgé, L. *Estimating the intensity of a random measure by histogram type estimators*. Probab. Theory Related Fields **143**(1-2), 239–284 (2009).
- [8] Baraud, Y., Huet, S., and Laurent, B. *Adaptive tests of linear hypotheses by model selection*. Ann. Statist., **31**, no. 1, 225–251 (2003).
- [9] Barron, A., Birgé, L., Massart, M. *Risk bounds for model selection via penalization*. Probab. Theory Related Fields, **113**(3), 301–413 (1999).
- [10] Birgé, L., Massart, P. *Gaussian model selection*. J. Eur. Math. Soc. **3**(3), 203–268 (2001).
- [11] Birgé, L. Massart, P. *Minimal penalties for Gaussian model selection*. P.T.R.F. **138**(1-2), 33–73 (2007).
- [12] Bousquet, O. *A Bennett concentration inequality and its application to suprema of empirical processes*. C. R. Acad. Sci. Paris, Ser. I **334**, 495–500 (2002).
- [13] Brémaud, P. *Point processes and queues*. Springer Series in Statistics (1981).
- [14] Brémaud, P., Massoulié, L. *Stability of nonlinear Hawkes processes*. Ann. Prob. **24**(3), 1563–1588 (1996).
- [15] Brown, L., Cai, T., Zhang, R., Zhao, L. and Zhou, H. *The root-unroot algorithm for density estimation as implemented via wavelet block thresholding*. Probability Theory and Related Fields, **146**(3-4), 401–433 (2010).

- [16] Brunel, E., Comte, F. *Penalized contrast estimation of density and hazard rate with censored data.* Sankhya **67**(3), 441–475 (2005).
- [17] Brunel, E., Comte, F. *Adaptive estimation of hazard rate with censored data.* Communications in Statistics, Theory and methods **37**(8), 1284–1305 (2008).
- [18] Butucea, C., and Tribouley, K. *Nonparametric homogeneity tests.* J. Statist. Plann. Inference, **136**, 597–639 (2006) .
- [19] Cirel’son, B.S., Ibragimov, I.A., Sudakov, V.N. *Norms of gaussian sample functions.* Proc. 3rd Japan-USSR Symp. Probab. Theory, Taschkent 1975, LNM **550**, 20–41 (1976).
- [20] Cohen, A., Daubechies, I. and Feauveau, J.C. *Biorthogonal bases of compactly supported wavelets.* Comm. Pure Appl. Math., **45**(5), 485–560 (1992).
- [21] Coronel-Brizio, H.F. and Hernandez-Montoya, A.R. *On fitting the Pareto-Lévy distribution to stock market index data: Selecting a suitable cutoff value.* Physica A: Statistical Mechanics and its Applications, **354**, 437–449 (2005).
- [22] Dachian, S., and Kutoyants, Yu.A. *Hypotheses testing: Poisson versus self-exciting.* Scand. J. Statist., **33**, 391–408 (2006).
- [23] Daley, D.J., Vere-Jones, D. *An introduction to the theory of point processes.* Springer series in statistics Volume I (2005).
- [24] de la Peña, V, Giné, E. *Decoupling : from dependence to independence.* Springer series in statistics (1999).
- [25] Donoho, D.L. and Johnstone, I.M. *Ideal spatial adaptation by wavelet shrinkage.* Biometrika, **81**(3), 425–455 (1994).
- [26] Doumic, M., Perthame, B. and Zubelli, J. *Numerical Solution of an Inverse Problem in Size-Structured Population Dynamics.* Inverse Problems, **25**, 045008, 25pp (2009).
- [27] Figueroa-López, J.E. and Houdré, C. *Risk bounds for the non-parametric estimation of Lévy processes.* IMS Lecture Notes-Monograph series High Dimensional Probability, **51**, 96–116 (2006).
- [28] Fromont, M., and Laurent, B. *Adaptive goodness-of-fit tests in a density model.* Ann. Statist., **34**(2), 680–720 (2006).
- [29] Fromont, M., Laurent, B., Reynaud-Bouret, P. *Adaptive test of homogeneity for a Poisson process.* to appear in Annals of IHP.
- [30] Giné, E., Latala, R., Zinn, J. *Exponential and Moment Inequalities for U-statistics.* High Dimensional Probability II - Progress in Probability, Birkhäuser, 13–38 (2000).
- [31] Gusto, G., Schbath, S. *FADO: a statistical method to detect favored or avoided distances between motif occurrences using the Hawkes’ model.* Statistical Applications in Genetics and Molecular Biology, **4**(1), Article 24 (2005).
- [32] Hawkes, A. G., Oakes, D. *A cluster process representation of a self-exciting process.* J. Appl. Prob. **11**(3), 493–503 (1974).

- [33] Houdré, C., Marchal, P., Reynaud-Bouret, P. *Concentration for norms of infinitely divisible vectors with independent components*. Bernoulli, **14**(4), 926–948 (2008).
- [34] Houdré, C., Privault, N. *Concentration and deviation inequalities in infinite dimensions via covariance representations*. Bernoulli **8**(6), 697–720 (2002).
- [35] Houdré, C., Reynaud-Bouret, P. *Exponential inequalities, with constants, for U-statistics of order two*. Stochastic inequalities and applications, Progr. Probab., 56 Birkhäuser, Basel, 55–69 (2003).
- [36] Houghton, J.C. *Use of the truncated shifted Pareto distribution in assessing size distribution of oil and gas fields*. Mathematical geology, **20**(8), 907–937(1988).
- [37] Ingster, Yu.I. *Asymptotically minimax testing for nonparametric alternatives I-II-III*. Math. Methods Statist., **2**, 85–114, 171–189, 249–268 (1993).
- [38] Ingster, Yu.I. *Adaptive chi-square tests*. J. Math. Sci., **99** (2), 1110–1119 (2000).
- [39] Ingster, Yu.I., and Kutoyants, Yu.A. *Nonparametric hypothesis testing for intensity of the Poisson process*. Math. Methods Statist., **16** (3), 217–245 (2007).
- [40] Johnstone, I.M. *Minimax Bayes, asymptotic minimax and sparse wavelet priors*. Statistical decision theory and related topics, V (West Lafayette, IN, 1992), New York: Springer, 303–326 (1994).
- [41] Juditsky, A. and Lambert-Lacroix S. *On minimax density estimation on  $\mathbb{R}$* . Bernoulli, **10**(2), 187–220 (2004).
- [42] Kerkycharian, G. and Picard, D. *Thresholding algorithms, maxisets and well-concentrated bases*. Test, **9**, 283–344 (2000).
- [43] Kingman, J.F.C. *Poisson processes*. Oxford studies in Probability (1993).
- [44] Klein, T. ,Rio, E. *Concentration around the mean for maxima of empirical processes*. Ann. Proba., **33**(3), 1060–1077 (2005).
- [45] Korostelev, A.P., Simar, L., Tsybakov, A.B. *Efficient Estimation of Monotone Boundaries*. Ann. Stat., **23**(2), 476–489 (1995).
- [46] Ledoux, M. *On Talagrand inequalities for product measures*. ESAIM PS, **1**, 95-144 (1996).
- [47] Lepez, V. *Some estimation problems related to oil reserves*. PhD manuscript of Paris Sud University (2002).
- [48] Lerasle, M. *Optimal model selection in density estimation*. Unpublished manuscript, hal-00422655 (2009).
- [49] Mallows, C.L. *Some comments on  $C_p$* . Technometrics, **15**, 661–675 (1973).
- [50] Massart, P. *About the constants in Talagrand’s concentration inequalities for empirical processes*. Ann. Proba., **28**(2), 863–884 (2000).

- [51] Massart, P. *Concentration inequalities and model selection*. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Springer, Berlin (2007).
- [52] Merton, R.C. *Option pricing when underlying stock returns are discontinuous*. Working paper Sloan School of Management, 787–795 (1975).
- [53] Ogata, Y., Akaike, H. *On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes*. Journal of the Royal Statistical Society, Series B, **44**(1), 102–107 (1982).
- [54] Ozaki, T. *Maximum likelihood estimation of Hawkes’ self-exciting point processes*. Ann. Inst. Statist. Math., **31**(B), 145–155 (1979).
- [55] Perthame, B., Zubelli, J. P. *On the inverse problem for a size-structured population model*. Inverse Problems, **23**(3), 1037–1052 (2007).
- [56] Reinert, G., Schbath, S., Waterman, M.S. *Probabilistic and Statistical Properties of Words: An Overview*. Journal of Computational Biology, **7**(1–2), 1–46 (2000).
- [57] Reynaud-Bouret, P. *Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities*. Probab. Theory Related Fields, **126** (1), 103–153 (2003).
- [58] Reynaud-Bouret, P. *Compensator and exponential inequalities for some suprema of counting processes*. Statistics and Probability Letters, **76**(14), 1514–1521 (2006).
- [59] Reynaud-Bouret, P. *Penalized projection estimators of the Aalen multiplicative intensity*. Bernoulli, **12**(4), 633–661 (2006).
- [60] Reynaud-Bouret, P., Roy, E. *Some non asymptotic tail estimates for Hawkes processes*. Bulletin of the Belgian Mathematical Society-Simon Stevin, **13**(5), 883–896 (2007).
- [61] Reynaud-Bouret, P., Rivoirard, V. *Near optimal thresholding estimation of a Poisson intensity on the real line*. Electronic Journal of Statistics, **4**, 172–238 (2010).
- [62] Reynaud-Bouret, P., Schbath, S. *Adaptive estimation for Hawkes processes; application to genome analysis*. Ann. Statist., **38**(5), 2781–2822 (2010).
- [63] Reynaud-Bouret, P., Rivoirard, V., Tuleau-Malot, C. *Adaptive density estimation: a curse of support?* J. Statist. Plann. Inference, **141**, 115–139 (2011).
- [64] Rivoirard, V. *Nonlinear estimation over weak Besov spaces and minimax Bayes method*. Bernoulli, **12**(4), 609–632 (2006).
- [65] Rosiński, J. *Remarks on strong exponential integrability of vector-valued random series and triangular arrays*. Ann. Probab., **23**, 464–473 (1996).
- [66] Samson, P.-M., *Concentration of measure inequalities for Markov chains and  $\phi$ -mixing processes*. Ann. Probab., **28**(1), 416–461 (2000).
- [67] Saumard, A. *Sélection de modèles optimale par pénalités de rééchantillonnage pour des  $M$ -estimateurs à contraste régulier*. Working paper, inria-00386782 (2009).

- [68] Spokoiny, V. G. *Adaptive hypothesis testing using wavelets*. Ann. Statist., **24**(6), 2477–2498 (1996).
- [69] Spokoiny, V. G. *Adaptive and spatially hypothesis testing of a nonparametric hypothesis*. Math. Methods Statist., **7**(3), 245–273 (1998).
- [70] Talagrand, M. *New concentration inequalities in product spaces*. Invent. Math., **126**(3), 505–563 (1996).
- [71] Uhler, R.S. and Bradley, P. G. *A Stochastic Model for Determining the Economic Prospects of Petroleum Exploration Over Large Regions*. Journal of the American Statistical Association, **65**(330), 623–630 (1970).
- [72] Vere-Jones, D., Ozaki, T. *Some examples of statistical estimation applied to earthquake data*. Ann. Inst. Statist. Math., **34**(B), 189–207 (1982).
- [73] Watson, G.S. *Estimating the intensity of a Poisson process*. Applied time series analysis, 1st proceeding, Tulsa, 1976, 325–345 (1978).
- [74] Willett, R.M. and Nowak, R.D. *Multiscale Poisson Intensity and Density Estimation*. IEEE Transactions on Information Theory, **53**(9), 3171–3187 (2007).
- [75] Wu, L. *A new modified logarithmic Sobolev inequality for Poisson point process and several applications*. Probab. Theory Related Fields, **118**(3), 427–438 (2000).