Probab. Theory Relat. Fields 126, 103–153 (2003)
Digital Object Identifier (DOI) 10.1007/s00440-003-0259-1

Patricia Reynaud-Bouret

Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities

Received: 24 April 2001 / Revised version: 9 October 2002 /
Published online: 15 April 2003 – © Springer-Verlag 2003

Abstract. In this paper, we establish oracle inequalities for penalized projection estimators of the intensity of an inhomogeneous Poisson process. We study consequently the adaptive properties of penalized projection estimators. At first we provide lower bounds for the minimax risk over various sets of smoothness for the intensity and then we prove that our estimators achieve these lower bounds up to some constants. The crucial tools to obtain the oracle inequalities are new concentration inequalities for suprema of integral functionals of Poisson processes which are analogous to Talagrand's inequalities for empirical processes.

1. Introduction

We consider here the problem of estimating the intensity s with respect to some measure μ of some inhomogeneous Poisson process N which is observed on the set $\mathbb{X}$. Poisson processes are known to be useful to model several random phenomena (see for instance [25]). The number of machine breakdowns can for example often be considered as a Poisson time process on some interval $[0; T]$. The phone calls in a city at some given time can also be represented by spatial Poisson processes.

There is a huge amount of papers devoted to curve estimation: in particular, the problem of estimating a density f from the observation of some n-sample $X_1, \dots , X_n$ of i.i.d. variables. This density framework is closely connected to the Poisson framework since it is well known that conditionally to the event "the number of points $N_{\mathbb{X}}$ falling into $\mathbb{X}$ is n", the points of the process obey the same law as a n-sample with density $f = s / \int_{\mathbb{X}} s \, d\mu$. This analogy has led to many works in which non parametric estimation procedures for the density framework have been transfered to the Poisson framework. For instance, M. Rudemo [37] studied in the density framework and in the Poisson framework histogram and kernel estimators. The kernel estimators for the intensity were also studied by Y.A. Kutoyants [27]: in his framework, the observation is some n-sample of Poisson processes. In analogy

P. Reynaud-Bouret: Georgia Institute of Technology, School of Mathematics, Atlanta, GA 30332, USA. e-mail: Patricia.Reynaud@dma.ens.fr

Mathematics Subject Classification (2000): 60E15, 62G05, 62G07

Key words or phrases: Inhomogeneous Poisson process – Concentration inequalities – Model selection – Penalized projection estimator – Adaptive estimation

to A.R. Barron and C.-H. Sheu [4], W.-C. Kim and J.-Y. Koo [24] studied also maximum likelihood type estimators on sieve for exponential family of wavelets.

The choice of the window in [27] or the choice of the sieve in [24] depends on the smoothness of the intensity so that the rate of convergence of the kernel estimator or of the maximum likelihood estimator respectively will be quite optimal. On the other side, M. Rudemo [37] is first to study cross-validation which is a data driven criterion to select a good window for kernel estimators or a good partition for histogram estimators. He does not use some prior assumption on the smoothness of the intensity. However no risk bounds for cross-validation are available in the Poisson framework unlike in the density framework.

Our purpose is to design adaptive estimation for the intensity, i.e. we want to design estimators which constructions require as few prior knowledge assumption on s (such as smoothness assumptions for instance) as possible. The aim is to obtain quite optimal rate of convergence for such estimators.

We want to transfer to the Poisson case, procedures which are based on model selection criterion and which were introduced by L. Birgé and P. Massart [8] in the density framework.

Let us now describe more precisely our framework and present our approach. We begin by giving the definition of a Poisson process to fix the notations.

Definition 1. *Let $(\mathbb{X}, \mathcal{X})$ be a measurable space. Let N be a random countable subset of $\mathbb{X}$. N is said to be a Poisson process on $(\mathbb{X}, \mathcal{X})$ if*

- *for all $A \in \mathcal{X}$, the number of points of N lying in A is a random variable N_A which obeys a Poisson law with parameter denoted by $\nu(A)$,*
- *for all finite family of disjoint sets $A_1, ..., A_n$ of $\mathcal{X}$, $N_{A_1}, ..., N_{A_n}$ are independent random variables.*

The so defined function $\nu\colon \mathcal{X} \to \mathbb{R}_+$ is a measure without atom (see [25]) and is called the "mean measure" of N. Here ν is assumed finite to obtain almost surely a finite set of points for N. We denote by dN the discrete random measure $\sum_{T\in N} \delta_T$.

Definition 2. *If the mean measure of a Poisson process N is absolutely continuous with respect to some measure μ, the Radon-Nikodym derivative s of the mean measure with respect to μ is called the* **intensity of the Poisson process** *N with respect to μ.*

If μ represents the Lebesgue measure and s is constant, N is called a homogeneous Poisson process. We deal with an inhomogeneous Poisson process when the intensity is a nonnegative function, but not necessarily constant. In this case, there is no assumption on μ except for its finiteness.

We are interested in estimating s knowing the almost surely finite set of points, $N(\omega)$ and assuming that s belongs to $\mathbb{L}^2 = \mathbb{L}^2(\mu/\mu(\mathbb{X}))$. We will keep the notation $\mu/\mu(\mathbb{X})$ and not deal with probability measure because we will sometimes want $\mu(\mathbb{X})$ to tend to infinity, which cannot be easily done with a probability measure notation. In this article, $\|.\|$ will always represent the $\mathbb{L}^2$ norm:

$$\|f\|^2 = \int_{\mathbb{X}} f^2(x) \frac{d\mu_x}{\mu(\mathbb{X})}.$$

At first, let us introduce the projection estimator of s on S, finite dimensional subspace of $\mathbb{L}^2$ with orthonormal basis $\{\varphi_1, \dots, \varphi_D\}$:

$$\hat{s} = \sum_{i=1}^{D} \left(\int_{\mathbb{X}} \varphi_i(x) \frac{dN_x}{\mu(\mathbb{X})} \right) \varphi_i. \tag{1.1}$$

For all i, let

$$\hat{\beta}_i = \int_{\mathbb{X}} \varphi_i(x) \frac{dN_x}{\mu(\mathbb{X})}. \tag{1.2}$$

$\hat{s}$ has to be compared with the orthogonal projection of s over S

$$\sum_{i=1}^{D} \left(\int_{\mathbb{X}} \varphi_i(x) \frac{s(x)d\mu_x}{\mu(\mathbb{X})} \right) \varphi_i.$$

From this definition, it is not clear that $\hat{s}$ depends only on S and not on the choice of some basis of S. In fact one can easily check that $\hat{s}$ is the unique minimizer over S of the following contrast:

$$\gamma_{\mathbb{X}}(f) = -\frac{2}{\mu(\mathbb{X})} \int_{\mathbb{X}} f(x) dN_x + \int_{\mathbb{X}} f^2(x) \frac{d\mu_x}{\mu(\mathbb{X})}. \tag{1.3}$$

For instance, if S is the linear subspace of all the histograms written on a given partition m, $\hat{s}$ is an histogram estimator of the form:

$$\hat{s} = \sum_{I \in m} \frac{N_I}{\mu(I)} \mathbb{1}_I.$$

This resembles M. Rudemo's ones, except that the normalization by $\mu(I)$ is replaced in his case by $N_{\mathbb{X}}$ times the length of the interval, I.

Our estimation method can be described as follows. Let $\{S_m, m \in \mathcal{M}_{\mathbb{X}}\}$ be a collection of linear models, i.e. finite dimensional subspaces of $\mathbb{L}^2$. The set $\mathcal{M}_{\mathbb{X}}$ is just a way to enumerate the linear models: for instance m can be a partition, S_m the space of all piecewise constant functions on m and $\mathcal{M}_{\mathbb{X}}$ a collection of partitions.

For each model, we denote by $\hat{s}_m$ the projection estimator of s on S_m. At last, we select among $\left\{\hat{s}_m, m \in \mathcal{M}_{\mathbb{X}}\right\}$ a good estimator through a data driven criterion which has the following form:

$$\begin{aligned} \hat{m} &= \text{argmin}_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ -\|\hat{s}_m\|^2 + \text{pen}(m) \right\} \\ &= \text{argmin}_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ \gamma_{\mathbb{X}}(\hat{s}_m) + \text{pen}(m) \right\}. \end{aligned} \tag{1.4}$$

where pen is a possibly random function: $\mathcal{M}_{\mathbb{X}} \to \mathbb{R}_+$ called the **penalty**. We denote $\tilde{s} = \hat{s}_{\hat{m}}$, the **penalized projection estimator** (p.p.e.).

For instance, let us take $\{\varphi_\lambda, \lambda \in \Lambda\}$ a finite orthonormal family of $\mathbb{L}^2$. We can look at $S_m = \text{Span}\{\varphi_\lambda, \lambda \in m\}$ where m is a subset of Λ and $\mathcal{M}_{\mathbb{X}}$ is a collection of subsets of Λ. Let $|m|$ denote the cardinal of the set m. This subset selection case leads for $\text{pen}(m) = C|m|$ and $\mathcal{M}_{\mathbb{X}} = \{m, m \subset \Lambda\}$ to a p.p.e.

which is in fact a particular hard threshold estimator. Indeed we have to minimize $-\sum_{\lambda\in m}\hat{\beta}_\lambda^2 + C|m| = -\sum_{\lambda\in m}(\hat{\beta}_\lambda^2 - C)$. Hence $\hat{m} = \{\lambda \in \Lambda/\hat{\beta}_\lambda^2 \geq C\}$ and $\tilde{s} = \sum_{\lambda\in\Lambda}\hat{\beta}_\lambda \mathbb{1}_{|\hat{\beta}_\lambda|\geq\sqrt{C}}$, i.e. a hard threshold estimator with constant level of thresholding. Threshold estimators have been introduced in the white noise framework and in the density framework by D.L. Donoho, G. Kerkyacherian and D. Picard (see for instance [18] and [23]). They are known to be adaptive and to have good approximation properties for proper threshold. Hence, in the two formulations (penalization or threshold), there is a factor to grade: the penalty or the level of thresholding. Studying low intensity image processing which is modeled by Poisson variables, D.L. Donoho [18] proposed a hard threshold: he uses the fact that the Anscombe's expression ($\sqrt{N+3/8}$ where N is a Poisson variable) [2] is asymptotically Gaussian in the Poisson parameter and he uses the level of thresholding deriving from the white noise framework. E. Kolaczyk [26] noticed that this threshold is not accurate enough in general, because the tails of the $\hat{\beta}_\lambda - \mathbb{E}(\hat{\beta}_\lambda)$'s are heavier than tails in the white noise framework and depend on the intensity s. He proposed another threshold, taking this into account, but always based on an asymptotic point of view and depending on the true intensity. It is also worth mentioning the work of L. Cavalier and J.-Y. Koo on hard threshold estimators in the tomographic data framework, where the Poisson process is observed through an inverse problem [14]. They proved that such estimators have almost optimal rate of convergence up to some factor which is a power of $\ln(\mu(\mathbb{X}))$. However, the level of thresholding depends on a prior upper bound on some smoothness norm of s.

Penalization can also generally be understood as a kind of cross-validation. Indeed, let $\{\varphi_\lambda, \lambda \in \mathcal{B}_m\}$ be an orthonormal basis of S_m; $\mathcal{B}_m$ is just a way to enumerate every members of the orthonormal basis of S_m. Let s_m be the orthogonal projection of s over S_m. We can compute the risk of a projection estimator $\hat{s}_m$ on a given model S_m:

$$\mathbb{E}(\|s - \hat{s}_m\|^2) = \|s - s_m\|^2 + \mathbb{E}(\chi_m^2), \tag{1.5}$$

where

$$\chi_m^2 = \sum_{\lambda\in\mathcal{B}_m}\left(\int_{\mathbb{X}}\varphi_\lambda(x)\frac{dN_x - s(x)d\mu_x}{\mu(\mathbb{X})}\right)^2. \tag{1.6}$$

The first term in Equation (1.5) is called the **bias term** and the second one is called **variance term**. This last term is equal to

$$\mathbb{E}(\chi_m^2) = \sum_{\lambda\in\mathcal{B}_m}\int_{\mathbb{X}}\varphi_\lambda^2(x)\frac{s(x)d\mu_x}{\mu(\mathbb{X})^2}. \tag{1.7}$$

If the models are nested, the variance term is non decreasing with the dimension of S_m and the bias term is non increasing with the dimension. More generally, the "best" model for a fixed s will be the one which makes the best compromise between these two terms. This "best" model, $\bar{m}$, is called the **oracle** and is defined as follows:

$$\bar{m} = \text{argmin}_{m\in\mathcal{M}_{\mathbb{X}}}\mathbb{E}(\|s - \hat{s}_m\|^2). \tag{1.8}$$

A way to find a good data driven criterion for model selection is to estimate without bias the risk over S_m. This heuristic is due to C.L. Mallows [30] in the Gaussian regression framework. We can adapt this heuristic to the Poisson case. However the variance depends here on s, then we have to estimate this without bias, with the same set of observations: that is the method of cross-validation developed by M. Rudemo [37] and M.M Brooks and J.S. Marron [12] for kernel estimators. More precisely, we can interpret $\bar{m}$ by:

$$\begin{aligned}\bar{m} &= \operatorname*{argmin}_{m\in\mathcal{M}_{\mathbb{X}}} \left\{-\|s_m\|^2 + \mathbb{E}(\|\hat{s}_m - s_m\|^2)\right\}\\ &= \operatorname*{argmin}_{m\in\mathcal{M}_{\mathbb{X}}} \left\{-\mathbb{E}(\|\hat{s}_m\|^2) + \mathbf{2}\,\mathbb{E}(\|\hat{s}_m - s_m\|^2)\right\}\\ &= \operatorname*{argmin}_{m\in\mathcal{M}_{\mathbb{X}}} \left\{\mathbb{E}(\gamma_{\mathbb{X}}(\hat{s}_m)) + \mathbf{2}\,\mathbb{E}(\|\hat{s}_m - s_m\|^2)\right\}. \end{aligned}\tag{1.9}$$

Hence the data driven criterion is of the form

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}_{\mathbb{X}}} \left\{\gamma_{\mathbb{X}}(\hat{s}_m) + 2\int_{\mathbb{X}} \sum_{\lambda\in m} \varphi_\lambda^2(x)\frac{dN_x}{\mu^2(\mathbb{X})}\right\}. \tag{1.10}$$

It is a penalized model selection criterion with

$$\mathrm{pen}(m) = 2\int_{\mathbb{X}} \sum_{\lambda\in m} \varphi_\lambda^2(x)\frac{dN_x}{\mu^2(\mathbb{X})}.$$

We propose in this paper penalties which either generalize or correct the previous one. These corrections are especially useful in the situation where there is exponentially many models with the same dimension in the family of models $\mathcal{M}_{\mathbb{X}}$. If the penalty is properly chosen, we shall prove that the p.p.e. performs almost as well as the "best" estimator in the family of models $\mathcal{M}_{\mathbb{X}}$., i.e.:

$$\mathbb{E}(\|s-\tilde{s}\|^2) \leq C_{\mathbb{X}} \inf_{m\in\mathcal{M}_{\mathbb{X}}} \mathbb{E}(\|s-\hat{s}_m\|^2), \tag{1.11}$$

where $C_{\mathbb{X}}$ is either some constant or some slowly varying factor of $\mu(\mathbb{X})$ depending on the complexity of the family of models. These inequalities are called **"oracle" inequalities**. L. Reboul already built some estimators of the intensity via Grenander's methods which have this property among the family of histogram estimators. But she supposed that the intensity is of the U-form, assumption which we shall not make here [34].

These oracle inequalities are the principal results of this paper since they are true in a very general setting. Moreover they do not just mean adaptivity in the family of the considered projection estimators but also they imply adaptivity properties in the minimax sense for the p.p.e. in special settings, when the family of models and the penalty are well chosen. For instance, the p.p.e. achieves (up to constants) the risk of the minimax estimator of the intensity over some collection of Besov balls for instance. It means that the p.p.e. performs as well as an estimator of the intensity where the smoothness of the intensity would be known. These results are analogous of those of L. Cavalier and J.Y. Koo [14]. For this aim we need to evaluate the minimax risk over some various classes of functions. Some asymptotic results are already available in the literature. In particular, Y.A. Kutoyants [27] computed asymptotically lower bound on minimax risks for Sobolev balls from the observations of a n-sample of Poisson processes with intensity s with respect to μ. Here we establish non asymptotic bounds for more general classes of functions, including Besov balls and also unions of finite dimensional spaces for which no lower bounds were known up to now. One can easily derive asymptotic results of the type studied by Y.A. Kutoyants from ours by noticing as L. Cavalier and J.-Y. Koo, that observing the n-sample of Poisson processes with intensity s with respect to μ is the same thing as observing the cumulative Poisson process $\mathcal{N} = \cup_{i=1}^{n} N_i$ with intensity s with respect to $n\mu$. We consider consequently in this article only one Poisson process (and if we have to give asymptotic, we do this in term of large $\mu(\mathbb{X})$).

The unbiased risk estimation is based on the idea that the risk is not very different from its expectation. In the proofs of the oracle inequalities, we need a probabilistic tool: the concentration inequalities which quantify the distance between a supremum of functions and its expectation. We apply these inequalities to χ_m remarking that

$$\chi_m = \sup_{\|a\|_2 \leq 1} \int_{\mathbb{X}} \sum_{\lambda \in \mathcal{B}_m} a_\lambda \varphi_\lambda \frac{dN_x - s(x)d\mu_x}{\mu(\mathbb{X})}. \tag{1.12}$$

These concentration phenomena are not asymptotic and lead us to non-asymptotic oracle inequalities.

A concentration inequality can be written in the following form:

$$\forall u > 0, \quad \mathbb{P}(Z \geq E(Z) + f(u)) \leq \exp[-u]$$

where Z is a random variable, and f a proper function.

Concentration inequalities were proved by B.S. Cirel'son, I.A. Ibragimov and V.N. Sudakov [15] for Z a 1-Lipschitz function of a Gaussian vector and

$$f(u) = \sqrt{2u}. \tag{1.13}$$

M. Talagrand (see [38]) proved that such inequalities can be written for

$$Z = \sup_{a \in A} (\mathbb{P}_n(\psi_a) - \mathbb{P}(\psi_a)),$$

with $\{\psi_a, a \in A\}$ countable family of functions bounded by 1 and with

$$f(u) = c_1\sqrt{v_n u} + c_2 u,$$

where $\mathbb{P}_n$ is the empirical measure for a n-sample $(X_1, ..., X_n)$ with law $d\mathbb{P} = s d\mu$ and where

$$v_n = \mathbb{E}\left(\sup_{a \in A} \sum_{i=1}^{n} (\psi_a(X_i) - \psi_a(X_i'))^2\right),$$

with $(X'_1, ..., X'_n)$ i.i.d. with $(X_1, ..., X_n)$ (for c_1, c_2 proper constants). The constants c_1 and c_2 are computed via M. Ledoux's methods in a paper of P. Massart [32].

Our main probabilistic result consists in providing some concentration inequalities for

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

with the same condition on $\{\psi_a, a \in A\}$ and with

$$f(u) = 2\sqrt{vu} + cu,$$

where

$$v = \frac{1}{2}\left[\mathbb{E}\left(\sup_{a \in A} \int_{\mathbb{X}} \psi_a^2(x) dN_x\right) + \sup_{a \in A} \int_{\mathbb{X}} \psi_a^2(x) d\nu_x\right].$$

We can remark the similarity between the two previous concentration inequalities with the correspondence $nd\mathbb{P}_n \approx dN$ and $d\mathbb{P} \approx d\nu$, which can be interpreted through the conditioning property. L. Wu [39] and C. Houdré and N. Privault [22] prove analogous results for $Z = f(N)$ where f is a 1-Lipschitz function, in some sense, of the Poisson process for [39] and of more general martingales for [22]. These results as ours can lead to concentration formula for i.i.d. vectors of Poisson variables, already proved by S.G. Bobkov and M. Ledoux [11]. Very general results about concentration inequalities for infinitely divisible vectors were also proved by C. Houdré [21]. All these results are very general but provide weaker results concerning the variance term v in this particular case of suprema. For the statistical applications, we need precisely a variance term of the form $\sup_{a \in A} \int_{\mathbb{X}} \psi_a^2(x) d\nu_x$ without any other dependence on $\mu(\mathbb{X})$: this is possible loosing some constants factors in front of each terms.

The link between concentration formula and adaptive estimation is well known. L. Birgé and P. Massart already used inequality (1.13) in the white noise framework and Talagrand concentration inequality in the density framework to get adaptive estimation by penalized model selection methods, from a non asymptotic point of view (see [7], [8], [10]). G. Castellan used concentration inequalities in the density framework for penalized maximum likelihood estimators (see [13]). Y. Baraud used it too in the regression framework (see [3]). Concentration inequalities can also be used in classification (see [33]).

The organization of this paper is as follows: in Section 2, we provide upper bounds for the risks of p.p.e, which lead to oracle inequalities. In Section 3, we give non-asymptotic lower bounds for the minimax risk on various sets of functions. In Section 4, we discuss adaptive properties of the p.p.e. Section 5 is devoted to probability and concentration inequalities for Poisson processes, tools which are at the center of our statistical demonstrations and heuristic. The last section is dedicated to the proofs of the main results.

2. Model selection with projection estimators

We wish to estimate the intensity s of an inhomogeneous Poisson process N, knowing the points of N in $\mathbb{X}$. We choose as adaptive estimator of the intensity the penalized projection estimator described in Equations (1.1), (1.3) and (1.4).

We want to prove in this section several oracle inequalities of type (1.11), depending on the penalty and on the family of models. For this aim, we have to distinguish two cases.

2.1. Model selection for a polynomial collection of models

The first result deals with a not too large family of models: more precisely, it deals with polynomial collection, in the following sense.

Definition 3. *The collection of models $\mathcal{M}_{\mathbb{X}}$ is said polynomial if there exists some nonnegative absolute constants Γ and R such that for all integer D,*

$$|\{m \in \mathcal{M}_{\mathbb{X}}, D_m = D\}| \leq \Gamma D^R,$$

where D_m denotes the dimension of the model S_m.

In this case, the computations are easier and can be made in a very general context.

Theorem 1. *Let N be a Poisson process on $(\mathbb{X}, \mathcal{X})$ with intensity s with respect to μ. Assume $\rho = \int_{\mathbb{X}} s d\mu/\mu(\mathbb{X})$ positive and s in $\mathbb{L}^2$. Let $\{S_m, m \in \mathcal{M}_{\mathbb{X}}\}$ be a collection of finite dimensional linear models. For all m in $\mathcal{M}_{\mathbb{X}}$, s_m denotes the orthogonal projection of s on S_m. For a given penalty* pen *on $\mathcal{M}_{\mathbb{X}}$, let $\tilde{s}$ be the associated penalized projection estimator (see (1.4)).*
Assume that:

1. *$\mathcal{M}_{\mathbb{X}}$ is a polynomial collection (see Definition 3) with constants Γ and R.*
2. *For all m in $\mathcal{M}_{\mathbb{X}}$, $\mathbb{D}_m = \sup_{f \in S_m, \|f\|=1} \|f\|_\infty^2 \leq \mu(\mathbb{X})$.*

Then for all $c > 1$

$$\mathbb{E}(\|s - \tilde{s}\|^2) \leq C \inf_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ \|s - s_m\|^2 + \mathbb{E}(\text{pen}(m)) \right\} + \frac{C'}{\mu(\mathbb{X})},$$

if the penalty is taken such that

(a) either for all m in $\mathcal{M}_{\mathbb{X}}$: $\text{pen}(m) \geq c\dfrac{N_{\mathbb{X}}\mathbb{D}_m}{\mu(\mathbb{X})^2}$,

(b) or if we suppose that $\displaystyle\inf_{m \in \mathcal{M}_{\mathbb{X}}} \frac{\mathbb{E}(\hat{V}_m)}{\mathbb{D}_m} = \beta > 0$,

for all m in $\mathcal{M}_{\mathbb{X}}$: $\text{pen}(m) \geq c\dfrac{\hat{V}_m}{\mu(\mathbb{X})}$

with $\displaystyle\hat{V}_m = \int_{\mathbb{X}} \sum_{\lambda \in \mathcal{B}_m} \varphi_\lambda^2 \frac{dN_x}{\mu(\mathbb{X})}$ *where $\{\varphi_\lambda, \lambda \in \mathcal{B}_m\}$ is an orthonormal basis of S_m,*

(c) or for all m in $\mathcal{M}_{\mathbb{X}}$: $\text{pen}(m) \geq \dfrac{c(\hat{V}_m + \alpha(N_{\mathbb{X}}/\mu(\mathbb{X}))\mathbb{D}_m)}{\mu(\mathbb{X})}$ *with $\alpha > 0$.*

C is then a continuous positive function depending only on c (and α in case (c)) and C' is a continuous positive function depending on c, Γ, R, $\|s\|$, $\|s\|_\infty$, ρ (and β in case (b) or α in case (c)).

Remarks.

Case (a) of penalty is very useful since we do not need to know precisely a basis of each model. Furthermore, the formulation allows us to take penalties slightly different from the case of equality: in a lot of situations we can have the following upper bound $\mathbb{D}_m \leq \Phi D_m$ for Φ absolute constant and in these situations we can take $\mathrm{pen}(m) = c(\Phi N_{\mathbb{X}} D_m)/\mu(\mathbb{X})^2$ with $c > 1$. Let us remark also that the first penalty in (a) verifies $\mathbb{E}(\mathrm{pen}(m))$ slightly larger than the variance term in the quadratic risk of $\hat{s}_m$ (see Equation (1.7)).

Moreover, as $\mathbb{D}_m \geq D_m$, the term $\mathbb{D}_m$ really looks like the dimension of the model: this leads to a penalized criterion which looks like Mallows criterion [30]. There is also a simple way to compute $\mathbb{D}_m$: whatever the orthonormal basis of S_m, $\{\varphi_\lambda, \lambda \in \mathcal{B}_m\}$, is, $\mathbb{D}_m = \| \sum_{\lambda\in\mathcal{B}_m} \varphi_\lambda^2 \|_\infty$.

Let us remark also that we obtain here exactly an oracle inequality (see (1.11) with $C_{\mathbb{X}} = C$ constant) in (b) (taking $\mathrm{pen}(m) = c\hat{V}_m/\mu(\mathbb{X})$) plus a rest which tends to 0 when $\mu(\mathbb{X})$ becomes large. Moreover taking $c = 2$, we have validated the heuristic which was presented in the introduction. We can remark too that the justification of the cross-validation is made under the assumption of the existence of β. This assumption is not required if we deal with a modified cross-validation criterion. Indeed, if one take the penalty according to (c), i.e.

$$\mathrm{pen}(m) = \frac{c}{\mu(\mathbb{X})}\left(\hat{V}_m + \alpha\left(\frac{N_{\mathbb{X}}}{\mu(\mathbb{X})}\right)\mathbb{D}_m\right),$$

the corrective term ensures that $\mathrm{pen}(m)$ cannot be smaller than $\mathbb{D}_m$ which leads to the improvement mentioned above.

We can remark too that we do not make any assumption here on the relationships between the models, S_m: the assumptions are on each model but not on their sum. This makes a difference with the situation where we want to deal with a more complex family of models, as we will see later.

Now let us give some interesting applications of this theorem.

Subset selection

The subset selection case, which is mentioned in the introduction, can be described as follows. Let $\{\varphi_\lambda, \lambda \in \Lambda\}$ be a large finite orthonormal family of $\mathbb{L}^2$. The collection of models $\mathcal{M}_{\mathbb{X}}$, can be interpreted as a collection of subsets of Λ. Hence the models can be described as follows: $S_m = \mathrm{Span}\{\varphi_\lambda, \lambda \in m\}$ for all m in $\mathcal{M}_{\mathbb{X}}$. In this situation, penalization is a good way to select the position of the coefficients to be estimated in the development of the intensity s on a basis of $\mathbb{L}^2$ ($\{\varphi_\lambda, \lambda \in \Lambda\}$ being in fact only a large preliminary part of this basis). Let us give some examples of this type.

- The first example is the simplest one: the Fourier basis. We take $\mathbb{X} = [0, T]$, $d\mu = dx$ the Lebesgue measure and $\{\varphi_\lambda, \lambda \in \Lambda\}$ is the set of the functions
$$\{\exp(-2ik\pi\ (x/T)), k \in \{-n, n\}\}.$$
Hence we have $\Lambda = \{-n, n\}$. We look at the following nested family of models (hence polynomial): $m_k = \{-k, k\}$ for all k less than n. In this case, we have $\mathbb{D}_{m_k} = D_{m_k} = 2k+1$. We choose n such that $2n+1 \leq T$ to validate Assumption 2. We can choose the three forms of penalty, for example the modified cross-validation one: $\text{pen}(m_k) = 2(\hat{V}_{m_k}/\mu(\mathbb{X}) + \alpha(2k+1)(N_{\mathbb{X}}/\mu(\mathbb{X})^2))$. Hence the p.p.e. is a Fourier truncated sum whose risk is upper bounded as before.
- The second example is the polynomial one. We want to select the degree of a "good" polynomial to approach s. We take $\mathbb{X} = [0, T]$, $d\mu = dx$ the Lebesgue measure and $\{\varphi_\lambda, \lambda \in \Lambda\}$ is the set of the functions
$$\{\sqrt{2k+1}Q_k\left(\frac{2x}{T} - 1\right), k \leq r\} \cup \{\mathbb{1}_{[0,T]}\},$$
where Q_k is the k-th Legendre polynomial. The family of models is nested (hence polynomial): $m_k = \{0, ..., k\}$. We have the following upper bound $\mathbb{D}_{m_k} \leq (k+1)^2 = D_{m_k}^2$ since the infinite norm of a Legendre polynomial is equal to 1. We choose r such that $r + 1 \leq \sqrt{\mu(\mathbb{X})}$ to validate Assumption 2. Hence for example, with $\text{pen}(m_k) = 2(k+1)^2(N_{\mathbb{X}}/\mu(\mathbb{X})^2)$, we obtain an inequality for the risk of $\tilde{s}$ which is quite an oracle inequality, with an upper bound on the variance term of the form $2(k+1)^2(\rho/\mu(\mathbb{X}))$ plus a rest which tends to 0 when $\mu(\mathbb{X})$ grows.
- The third example is the additive model. We take $\mathbb{X} = [0, T]^d$ and $d\mu$ is the product Lebesgue measure. We take
$$\varphi_{k,i}(x_1, ..., x_d) = \sqrt{2k+1}Q_k\left(\frac{2x_i}{T} - 1\right),$$
where Q_k is the k-th Legendre polynomial, for $d \geq i \geq 1$ and $k \geq 1$ and $\varphi_{0,0} = 1$. Let $\{r_i, 1 \leq i \leq d\}$ be a finite family of positive integers and let Λ be $\{(k, i), 1 \leq k \leq r_i, 1 \leq i \leq d\} \cup \{(0, 0)\}$. The family $\{\varphi_{k,i}, (k, i) \in \Lambda\}$ is orthonormal, for the normalized measure. We look at the following family of additive models: $m_{\mathbf{l}} = \{(k, i), 1 \leq k \leq l_i, 1 \leq i \leq d\} \cup \{(0, 0)\}$ for all $\mathbf{l} = (l_1, ..., l_d)$ with l_i less than r_i for all i.
That is to say that we search an estimator of the intensity of the form: $f_1(x_1) + ... + f_d(x_d)$ with the f_j polynomials with degree less than r_j.
We can verify that this family is polynomial: the cardinality of $\{m \in \mathcal{M}_{\mathbb{X}}, |m| = D\}$ is less than the number of choices of d integers such that their sum is equal to $D - 1$, which is of order $C_d D^d$ with C_d depending only on d. We have an upper bound for $\mathbb{D}_{m_{\mathbf{l}}} \leq 1 + \sum_{i=1}^d (l_i^2 + 2l_i)$. Then we choose $\mathbf{r}$ such that $1 + \sum_{i=1}^d (r_i^2 + 2r_i) \leq \mu(\mathbb{X})$ to validate Assumption 2. For all the given choices of penalty, the following p.p.e. has a risk bound similar to the bound of Theorem 1 for additive models.

- We can also choose compactly support wavelet basis. As we will see Section 4, we can consequently construct a p.p.e. which verifies the assumptions of Theorem 1 and which will reach the minimax risk up to a constant on Besov balls of $B^{\alpha}_{2,2}$. But we need to look at more complex families of models, for more complex Besov spaces.

Histogram selection

Since bounded measurable functions can be approximated by piecewise constant functions, we can also imagine estimators which will be piecewise constant functions, i.e. histograms (see for instance [37]). Hence we can figure out that $\mathcal{M}_{\mathbb{X}}$ is a collection of partitions of $\mathbb{X}$ and a model S_m, for m in $\mathcal{M}_{\mathbb{X}}$, is the set of all piecewise constant functions based on the partition m. Penalization can help to find a good partition on which we can construct the histogram estimator, already mentioned in the introduction.

A good example is regular histograms. We want to estimate the intensity s on a regular partition m, i.e. all the pieces I of the partition m have the same measure $\mu(I) = \mu_m$. We want to choose consequently a good width. There is one model by dimension, hence the family is obviously polynomial (but not necessarily nested). We choose for all m in $\mathcal{M}_{\mathbb{X}}$, the basis of S_m as the renormalized indicator functions of the pieces of m, $\{\mathbb{1}_I \sqrt{(\mu(\mathbb{X})/\mu_m)}, I \in m\}$. Then we get $\mathbb{D}_m = D_m = (\mu(\mathbb{X})/\mu_m)$. Then $\mu_m \geq 1$ implies Assumption 2. The same kind of condition on μ_m is given in the density framework [13]. In this framework, this condition is obvious since, otherwise, there is less than one point in each interval. In the Poisson framework, this is the same idea, since $\mu(\mathbb{X})$ is of the same order as $\mathbb{E}(N_{\mathbb{X}})$, the expected number of observed points. For all the choices of penalty given in Theorem 1, the resulting p.p.e. has a quadratic risk bounded as in Theorem 1.

2.2. *Model selection for a more complex family of models*

We prove here a quite general bound on the risk of the p.p.e. under some assumptions on the link between each model. It explains how the complexity of the family of models can modify the penalty to obtain proper bounds on the risk. This theorem is very abstract and this is why we prefer to first give the applications of this theorem in the two previous cases: the subsets selection case and the histograms selection case.

Subset selection

We keep the notations of the previous subsection. As we do not want to make any assumptions on the complexity of the family (like polynomial assumptions), we have to make some assumptions on the largest family of coefficients Λ.

Definition 4. $\{\varphi_\lambda, \lambda \in \Lambda\}$ *is said to be localized, if and only if:*

$$\exists B > 0, \forall a \in \mathbb{R}^{\Lambda} \left\| \sum_{\lambda \in \Lambda} a_\lambda \varphi_\lambda \right\|_{\infty} \leq B\sqrt{|\Lambda|} \sup_{\lambda \in \Lambda} |a_\lambda|.$$

The Fourier basis does not verify this property with B independent of Λ which is the interesting case, as we will see later, but wavelet basis with finite support verify such a property with a constant B independent of Λ (see Section 3).

Proposition 1. *Let N be a Poisson process on $(\mathbb{X}, \mathcal{X})$ with intensity s with respect to μ; s is assumed to be in $\mathbb{L}^2$. Let $\{\varphi_\lambda, \lambda \in \Lambda\}$ be a finite orthonormal family for $\mathbb{L}^2$. Let $\mathcal{M}_\mathbb{X}$ be a collection of subsets of Λ. For every subset of indices, m, let s_m be the orthogonal projection of s on $S_m = \mathrm{Span}\{\varphi_\lambda, \lambda \in \ m\}$ and let $\hat{s}_m$ be the projection estimator on S_m (cf (1.1)). For a given penalty* pen *on $\mathcal{M}_\mathbb{X}$, let $\tilde{s}$ be the associated penalized projection estimator (see (1.4)).*
Assume that:

1. *the family $\{\varphi_\lambda, \lambda \in \Lambda\}$ is localized (cf Definition 4), with constant B independent of $\mu(\mathbb{X})$,*
2. *there exists a finite family of positive weights on $\mathcal{M}_\mathbb{X}$, $(L_m)_{m\in\mathcal{M}_\mathbb{X}}$ such that* $\sum_{m\in\mathcal{M}_\mathbb{X}} \exp(-L_m|m|) \leq \Sigma$ *with Σ independent of $\mu(\mathbb{X})$,*
3. *$|\Lambda|$ is less than $\mu(\mathbb{X})/\ln^2\mu(\mathbb{X})$.*

Then,

- *if s is supposed to be bounded by M', where M' is known and*
  (a) either if $\mathrm{pen}(m) = \frac{cM'|m|}{\mu(\mathbb{X})}\left(1+\sqrt{2\kappa L_m}\right)^2$ *with c larger than* 1,
  (b) or if (random penalty) $\mathrm{pen}(m) = \frac{c}{\mu(\mathbb{X})}\left(\sqrt{\hat{V}_m} + \sqrt{2\kappa M' L_m|m|}\right)^2$ *with c larger than* 1, *where $\hat{V}_m = \int_\mathbb{X}\sum_{\lambda\in m}\varphi_\lambda^2 dN_x/\mu(\mathbb{X})$, and furthermore, for this random penalty, if $B^2|\Lambda| \leq (3/4)\kappa M'\mu(\mathbb{X})(\sqrt{1+\varepsilon}-1)$,*
  then the risk is bounded by

$$\mathbb{E}(\|s-\tilde{s}\|^2) \leq C(c)\inf_{m\in\mathcal{M}}\left[\|s-s_m\|^2 + \frac{M'|m|}{\mu(\mathbb{X})}(1+L_m)\right] + \frac{C'(c, B, M', \Sigma)}{\mu(\mathbb{X})},$$

  where C and C' are proper positive continuous functions, and where κ is defined in Corollary 2 in Section 5,
- *otherwise, if M' is unknown or even does not exist*
  (c) replacing in the two previous formula of penalties, M' by $\|\hat{s}_\Lambda\|_\infty + K'$, where K' is an arbitrary positive constant, under the assumption that

$$M_\Lambda = \sup_{\sum_{\lambda\in\Lambda} a_\lambda^2 = 1}\int_\mathbb{X}\left(\sum_{\lambda\in\Lambda} a_\lambda\varphi_\lambda\right)^2(x)s(x)\frac{d\mu_x}{\mu(\mathbb{X})} \leq \|s_\Lambda\|_\infty + K',$$

  leads to this upper bound for the risk:

$$\mathbb{E}(\|s-\tilde{s}\|^2) \leq C_1(c)\inf_{m\in\mathcal{M}}\left[\|s-s_m\|^2 + \frac{(\|s_\Lambda\|_\infty + K')|m|}{\mu(\mathbb{X})}(1+L_m)\right] + \frac{C_1'(c, B, \|s_\Lambda\|_\infty, K', \Sigma)}{\mu(\mathbb{X})},$$

  for C_1 and C_1' proper positive continuous functions.

Here, we loose an important thing with respect to Theorem 1: we need to know an upper bound on s or at least to have an idea of a good K'. If $|\Lambda|$ is large enough (which can happen only if $\mu(\mathbb{X})$ is large enough at fix s) the assumption in (c) will be verified for a certain class of s, for example in Besov space (cf Section 4), but we are not able to say in advance if the assumption is true or not for fixed s if we do not know its Besov norm. Instead, we win the capacity to look at complex family of models: we can answer to the following problem. Let s be a function with only D coefficients different from 0 in its wavelet development, we know that they are among the N first coefficients. A possible way to estimate such a s is to look at the family $\mathcal{M}_{\mathbb{X}} = \{m \subset \Lambda = \{1, ..., N\}\}$, which is not polynomial and to construct the penalized estimator as previously in (c) with $L_m = \ln(\mu(\mathbb{X})/|m|)$ (remark: assumption of (c) is true, because $\|s\|_\infty = \|s_\Lambda\|_\infty$, and so all $K' \geq 0$ work). As we shall see in Section 4, the penalized estimator $\tilde{s}$ of this theorem is minimax for this problem.

Furthermore, we have a kind of oracle inequality (see (1.11)) with $C_{\mathbb{X}}$ depending effectively on $\mathbb{X}$ if the L_m are not constant. As we shall see in the forthcoming sections, this factor L_m is necessary and allows us to reach minimax risk in different cases.

Remark. L. Birgé and P. Massart have the same problem with this unknown bound on s in the density framework [8]. This phenomenon is called “heteroscedasticity”. In this framework, it may disappear if one chooses an other contrast, for example log-likelihood, as G. Castellan proved it in [13]. One can hope that the same improvement could be obtained in the Poisson framework but there is still some work to do to prove it.

Histogram selection

In the histograms selection case, heteroscedasticity is easier to handle.

Proposition 2. *Let N be a Poisson process on $(\mathbb{X}, \mathcal{X})$ with intensity s with respect to μ; s is assumed to be in $\mathbb{L}^2$. Let Γ be a fixed regular partition (or grid) of $\mathbb{X}$. Let $\mathcal{M}_{\mathbb{X}}$ be a family of partitions which are constructed with unions of the boxes of this grid, Γ. For any partition, m, S_m is the subspace of histograms based on the partition m, D_m denotes the number of sets in m. For a given penalty* pen *on $\mathcal{M}_{\mathbb{X}}$, let $\tilde{s}$ be the associated penalized projection estimator (see (1.4)).*
Assume that:

1. *there exists a finite family of positive weights on $\mathcal{M}_{\mathbb{X}}$, $(L_m)_{m\in\mathcal{M}_{\mathbb{X}}}$ such that* $\sum_{m\in\mathcal{M}_{\mathbb{X}}} \exp(-L_m D_m) \leq \Sigma$ *with Σ independent of $\mu(\mathbb{X})$,*
2. *D_Γ is less than $\mu(\mathbb{X})/\ln^2\mu(\mathbb{X})$.*

For all $c > 1$, if

$$\mathrm{pen}(m) = \frac{c\tilde{M}D_m}{\mu(\mathbb{X})}(1 + \sqrt{2\kappa L_m})^2,$$

where

$$\tilde{M} = \sup_{I\in\Gamma} \frac{N_I}{\mu(I)},$$

then

$$\mathbb{E}(\|s-\tilde{s}\|^2) \leq C(c) \inf_{m\in\mathcal{M}} \left[\|s-s_m\|^2 + \frac{MD_m}{\mu(\mathbb{X})}(1+L_m)\right] + \frac{C'(c,\Sigma,M)}{\mu(\mathbb{X})},$$

where C and C' are proper positive continuous functions and $M = \sup_{I\in\Gamma}\left(\int_I s d\mu / \mu(I)\right)$.

In this case, there is no more heteroscedasticity problem. We estimate "$\|s\|_\infty$" in some sense by $\tilde{M}$ which depends only on the data.

A general model selection theorem

Theorem 2. *Let N be a Poisson process on $(\mathbb{X}, \mathcal{X})$ with intensity s with respect to μ; s is assumed to be in $\mathbb{L}^2$. Let $\{S_m, m \in \mathcal{M}_\mathbb{X}\}$ be a collection of finite dimensional linear models. For all m, m' in $\mathcal{M}_\mathbb{X}$, $(s_m, \hat{s}_m)$, respectively $(s_{m,m'}, \hat{s}_{m,m'})$, denote the orthogonal projection and the projection estimator on S_m, respectively $S_m + S_{m'}$. Let χ_m, respectively $\chi_{m,m'}$, be the norm $\|s_m - \hat{s}_m\|$, respectively $\|s_{m,m'} - \hat{s}_{m,m'}\|$. For a given penalty* pen *on $\mathcal{M}_\mathbb{X}$, let $\tilde{s}$ be the associated penalized projection estimator (see (1.4)).*
We assume the following properties.

1. *There exists S_Λ, finite dimensional linear subspace, which includes all the S_m's and there exists Φ positive such that $\mathbb{D}_\Lambda = \sup_{f\in S_\Lambda, \|f\|=1} \|f\|_\infty^2 \leq \Phi\mu(\mathbb{X})$.*

Let M be an upper bound of $\sup_{f\in S_\Lambda/\|f\|=1} \int_\mathbb{X} f^2 s d\mu/\mu(\mathbb{X})$. Let $\varepsilon > 0$ and assume the existence of some event $\Omega(\varepsilon)$ where for all m, m',

$$\|s_{m,m'} - \hat{s}_{m,m'}\|_\infty \leq \frac{2\kappa M\varepsilon}{\kappa(\varepsilon)},$$

(where κ and $\kappa(\varepsilon)$ are given in Corollary 2 in Section 5) such that the following properties hold.

2. *There exists $\Delta = \Delta(\varepsilon)$ such that $\mathbb{P}(\Omega(\varepsilon)^c) \leq \Delta/\mu(\mathbb{X})^2$.*
3. *There exists a function V: $\mathcal{M}_\mathbb{X} \to \mathbb{R}^+$ such that for all m, m' in $\mathcal{M}_\mathbb{X}$*

$$\mathbb{E}(\chi^2_{m,m'}) \leq V(m) + V(m').$$

4. *There exists an estimator $\hat{V}$: $\mathcal{M}_\mathbb{X} \to \mathbb{R}^+$, a known positive constant η and a positive constant Σ_0, such that for all m, m' in $\mathcal{M}_\mathbb{X}$, on $\Omega(\varepsilon)$, for all positive ξ, with probability larger than $1 - \Sigma_0 e^{-\xi}$, $\hat{V}(m') + \eta\xi \geq V(m')$.*
5. *There exists an estimator $\hat{M}$ such that $\hat{M} \geq M$ on $\Omega(\varepsilon)$.*
6. *There exists a constant Σ_1 and a finite family of weights, $(L_m)_{m\in\mathcal{M}_\mathbb{X}}$ such that*

$$\sum_{m\in\mathcal{M}_\mathbb{X}} e^{-L_m D_m} \leq \Sigma_1.$$

If the penalty verifies for all m in $\mathcal{M}_{\mathbb{X}}$*,*

$$\mathrm{pen}(m) \geq \frac{(1+\varepsilon)^5}{\mu(\mathbb{X})} \left(\sqrt{\hat{V}(m)} + \sqrt{2\kappa \hat{M} L_m D_m} \right)^2 ,$$

then the penalized projection estimator, $\tilde{s}$*, verifies*

$$\mathbb{E}(\|\tilde{s} - s\|^2) \leq C(\varepsilon) \inf_{m \in \mathcal{M}_{\mathbb{X}}} \big\{ \|s - s_m\|^2 + \mathbb{E}(\mathrm{pen}(m) \mathbb{1}_{\Omega(\varepsilon)}) \big\} + \frac{C'(\varepsilon, \Delta, \Sigma_0, \Sigma_1, M, \Phi, \eta)}{\mu(\mathbb{X})},$$

where C *and* C' *are proper positive continuous functions.*

This theorem is very general and does not make essential assumptions on the bases of the model. The assumptions deal mostly with the relationships between two models. In the applications of this theorem (Propositions 1 and 2), these assumptions on the relationships between the models follow from the form of the bases of each model (subsets selection case plus localization or histograms selection case).

3. Some lower bounds for the minimax risk

Now we have some kind of oracle inequalities for the p.p.e. for proper choices of penalties, that is to say that we know how to compare the p.p.e with the best estimator among the family $\{\hat{s}_m, m \in \mathcal{M}_{\mathbb{X}}\}$. But comparing it with *all* other possible estimators requires to introduce the minimax risk.

Definition 5. *Let* $\mathcal{S}$ *be a subset of possible functions of the intensity* s*. Then the minimax risk on* $\mathcal{S}$ *is*

$$R(\mathcal{S}) = \inf_{\hat{s} \in \mathcal{F}(N)} \sup_{s \in \mathcal{S}} \mathbb{E}(\|s - \hat{s}\|^2),$$

where $\mathcal{F}(N)$ *is the set of all functions of the points of* N *with values in the set of* $\mathbb{L}^2$ *intensities.*

The minimax risk on $\mathcal{S}$ represents the risk of the best estimator for the worst s to estimate in the family $\mathcal{S}$.

Remark. The minimax risk is increasing with $\mathcal{S}$ in the meaning of inclusion. Therefore comparing the risk of our estimator with the minimax risk, we answer the question: does the p.p.e. estimate as well as the best one, which knows that s belongs to $\mathcal{S}$, even if our estimator does not know this fact?
Our aim in this part is to present lower bounds for the minimax risk on some family of possible functions for s.

3.1. Minimax risk on ellipsoids

We keep the framework of the subsets selection case. Let $\{\varphi_\lambda, \lambda \in \mathbb{N}^*\}$ be an orthonormal basis of $\mathbb{L}^2$. We assume that there exists L such that for $\lambda \geq L$, φ_λ is orthogonal to the constant functions. Let $c = (c_\lambda)_{\lambda \geq 1}$ be a positive non increasing sequence and ρ a positive number. Let us denote by $\mathcal{E}(c, \rho)$ the set

$$\mathcal{E}(c, \rho) = \left\{ t = \rho + u \ : \ \int_{\mathbb{X}} u = 0,\ u = \sum_{\lambda=1}^{\infty} \beta_\lambda \varphi_\lambda \ ,\ t \geq 0,\ \sum_{\lambda \geq 1} \left(\frac{\beta_\lambda}{c_\lambda} \right)^2 \leq 1 \right\}.$$

We can find a lower bound for the minimax risk on this ellipsoid.

Proposition 3. *Assume that there exists an integer $D > L$ such that $\{\varphi_\lambda, \lambda \in \{1, ..., D\}\}$ is localized with constant B (see Definition 4). If*

$$\frac{c_D^2}{D} \leq \zeta \frac{\rho}{\mu(\mathbb{X})},$$

then

$$R(\mathcal{E}(c, \rho)) \geq \eta \left[\frac{D - L + 1}{D} \right] \left(\frac{\rho^2}{4B^2} \wedge c_D^2 \right),$$

where η and ζ are proper constants.

The term L is here to make this bound valid for some current choices of wavelet bases. For the Haar basis, $L = 2$.

3.2. Logarithmic factors in the risk

In the same framework, let n, D be two positive integers and $\mathcal{S}_{n,D}$ be $\cup_{m \subset \{1,...,n\}, |m|=D} S_m$ where $S_m = \text{Span}\{\varphi_\lambda, \lambda \in m\}$. This is the set of functions which have only D non zero coefficients in the development on $\{\varphi_\lambda, \lambda \in \mathbb{N}^*\}$ and we know that these coefficients are among the first n coefficients. Let $B_{n,D,\rho}$ be the following set:

$$B_{n,D,\rho} = \left\{ t = \rho + u \ : \ \int_{\mathbb{X}} u = 0, u \in \mathcal{S}_{n,D}, t \geq 0 \right\}. \tag{3.1}$$

We obtain the following proposition:

Proposition 4. *Let $n > L$. Assume that the family $\{\varphi_\lambda, \lambda \in \{1, ..., n\}\}$ is localized (cf Definition 4) with constant B. If $n \geq 4D$, then*

$$R(B_{n,D,\rho}) \geq \eta \left(\frac{\zeta \rho D \log \frac{n-L+1}{D}}{\mu(\mathbb{X})} \wedge \frac{\rho^2 D}{4B^2 n} \right),$$

where η and σ are proper constants.

3.3. *Besov spaces*

We now limit ourself to looking at $\mathbb{X} = [0, T]$, equipped with borelians and μ is the Lebesgue measure.

3.3.1. Wavelet expansions

We are dealing with wavelet basis on an interval (and not on $\mathbb{R}$). The best known example is the Haar basis. When we want to look at smoother wavelets, we can deal with the one constructed by A. Cohen, I. Daubechies and P. Vial [16]. More precisely, they construct a wavelet basis for $\mathbb{L}^2([0; 1])$. In practice, this basis has the following form. Let l, K be two positive integers such that $2^l \geq 2K > 0$. The family $\{p_{j,k}, j \geq l, k = 0, ..., 2^j - 1\}$ is of the following form. For $j = l$, the $p_{l,k}$'s denote "gross structure term". For $0 \leq k \leq 2^l - 2K - 1$, $p_{l,k}$ is the dilatation and translation $2^{l/2}\Phi(2^l x - k)$ of a father wavelet Φ. This father wavelet has unit integral and compact support lying in $[0, 2K - 1]$. For $2^l - 2K \leq k \leq 2^l - 1$, $p_{l,k}$ are the boundary scaling functions for edges 0 and 1. The $p_{l,k}$'s generate in particular the constant functions. For $j > l$ and $0 \leq k \leq 2^j - 2K - 1$, $p_{j,k}$ is the dilatation and translation $2^{j/2}\Psi(2^j x - k)$ of a mother wavelet Ψ. The mother is with zero integral and N vanishing moments. For $2^j - 2K \leq k \leq 2^j - 1$, $p_{j,k}$ are the scaled at level j of $2K$ functions and are the boundary wavelets at each edges. They have the same regularity and the same vanishing moments as Ψ. Then we need $4K + 2$ functions, the other one are scaled and translated from these functions.
To get a wavelet basis on $[0; T]$ for the renormalized Lebesgue measure, we set

$$\forall j \geq l, \forall k \in \{0, ..., 2^j - 1\} = \Lambda(j), \quad \varphi_{j,k}(x) = p_{j,k}(x/T).$$

In order to avoid introducing superfluous notations, we shall abusively also denote by $\{\varphi_\lambda, \lambda \in \mathbb{N}^*\}$ the previous wavelet bias ordered according to the lexicographical ordering. (For instance, for $\lambda = 1$, $\varphi_\lambda = \varphi_{l,0}$; for $\lambda = 2$, $\varphi_\lambda = \varphi_{l,2}$; for $\lambda = 2^l + 1$, $\varphi_\lambda = \varphi_{l+1,0}$.) So for t in $\mathbb{L}^2$, we have the following development

$$t(x) = \sum_{\lambda \in \mathbb{N}^*} a_\lambda \varphi_\lambda(x) = \sum_{j \geq l} \sum_{k \in \Lambda(j)} a_{j,k} \varphi_{j,k}(x). \tag{3.2}$$

We set

$$\Sigma_\infty(t) = \sum_{j \geq l} 2^{j/2} \sup_{k \in \Lambda(j)} |a_{j,k}|,$$

and note that, since the $\varphi_{j,k}$'s have almost disjoint supports for $j > l$, we have $\|t\|_\infty \leq H\Sigma_\infty(t)$ for some positive constant H. We deduce in particular from this inequality, that the family

$$\mathcal{F}_J = \left\{\varphi_{j,k}, k \in \Lambda(j), l \leq j \leq J\right\} \tag{3.3}$$

is localized in the sense of Definition 4 with constant B which depends on H and l but which is independent of J and consequently independent of the cardinality of the family. When the basis is scaled from the Haar basis (i.e. $\Phi = \mathbb{1}_{[0;1]}$; $\Psi = \mathbb{1}_{[0;1/2]} - \mathbb{1}_{]1/2;1]}$; $l = 0$; $K = 1$), we obtain for instance $1/(\sqrt{2} - 1)$.

The wavelet basis has regularity r if the functions used in the analysis are compactly supported and have r continuous derivatives. It is possible to get r large enough, by selecting $K(r)$ large enough. Such wavelet basis exists (see [16]).

Coefficients on a regular wavelet basis can be used to measure the smoothness of the function. The Besov space $B^{\alpha}_{p,p'}$ (for $\alpha > 0,\ p \geq 1,\ p' \geq 1$) is one of the space of smooth functions which is classically considered (see [17, p 55] for a definition). It can be described with wavelets (see [19, Theorem 2]): the consequence of this theorem is that we can say for all wavelet with regularity $r > \alpha$ that

$$B^{\alpha}_{p,p'} = \left\{ t \in \mathbb{L}^2[0,T],\, 2^{j(\alpha+\frac{1}{2}-\frac{1}{p})}\|a_{j,.}\|_p \in l^{p'}(\mathbb{N}) \right\}, \tag{3.4}$$

where $a_{j,k}$ are the coefficients defined in (3.2). The associated norm of this space can be taken as follows:

$$\begin{aligned} \forall p' < +\infty,\ \ \|t\|^{\alpha}_{p,p'} &= \left(\sum_{j \geq 0} 2^{jp'(\alpha+\frac{1}{2}-\frac{1}{p})}\|a_{j,.}\|_p^{p'} \right)^{1/p'}, \\ p' = +\infty,\ \ \|t\|^{\alpha}_{p,\infty} &= \sup_{j \in \mathbb{N}} \left(2^{j(\alpha+\frac{1}{2}-\frac{1}{p})}\|a_{j,.}\|_p \right). \end{aligned} \tag{3.5}$$

When $p > 2$, $B^{\alpha}_{p,p'} \subset B^{\alpha}_{2,p'}$. So, we are only interested in $p \leq 2$, since we have always supposed s in $\mathbb{L}^2$. Then we have $B^{\alpha}_{2,2} \subset B^{\alpha}_{2,\infty} \cap \mathbb{L}^2 \subset B^{\alpha}_{p,\infty} \cap \mathbb{L}^2$, provided that $\alpha > 1/p \geq 1/2$. Indeed, we remark for all t in $\mathbb{L}^2$ that

$$\|t\|^{\alpha}_{2,2} \geq \|t\|^{\alpha}_{2,\infty} \geq \|t\|^{\alpha}_{p,\infty} \tag{3.6}$$

provided that $\alpha > 1/p \geq 1/2$.

3.3.2. Minimax risk for $B^{\alpha}_{2,2}$ balls

Using wavelets approach and Proposition 3, a lower bound for the minimax risk on Besov balls can be found. Let $\rho,\ R$ and α be positive numbers. Let $\mathcal{B}(\rho, R, B^{\alpha}_{2,2})$ be the set

$$\mathcal{B}(\rho, R, B^{\alpha}_{2,2}) = \left\{ t = \rho + u\ :\ t \geq 0,\ \int_{\mathbb{X}} u\, dx = 0,\ u \in B^{\alpha}_{2,2},\ \|u\|^{\alpha}_{2,2} \leq R \right\}, \tag{3.7}$$

where the Besov norm is defined in (3.5).

Proposition 5. *We have*

$$R(\mathcal{B}(\rho, R, B^{\alpha}_{2,2})) \geq C \left(\rho^{\frac{2\alpha}{2\alpha+1}} R^{\frac{2}{2\alpha+1}} T^{\frac{-2\alpha}{2\alpha+1}} \wedge \frac{\rho^2}{4B^2} \wedge R^2 2^{-2(l+1)\alpha} \right),$$

with C and B some positive constants depending only on the wavelet basis.

Letting T go to infinity we easily derive from Proposition 5 the following asymptotic lower bound.

Corollary 1. *There exists a positive constant C depending on the wavelet basis such that*

$$\liminf_{T\to+\infty}\left(T^{\frac{2\alpha}{2\alpha+1}}R(\mathcal{B}(\rho,R,B^{\alpha}_{2,2}))\right)\geq C\rho^{\frac{2\alpha}{2\alpha+1}}R^{\frac{2}{2\alpha+1}}. \tag{3.8}$$

Note that a related asymptotic lower bound (with some explicit value of C) has already be obtained by Y.A. Kutoyants [27] for a n-sample of Poisson processes.

4. Comparison between the risk of p.p.e. and the minimax risk

We keep the notations of Section 3.3 and want to understand in this case the performance of the p.p.e. in term of minimax risk.
Let $\{\varphi_{j,k}, j\geq l, k\in\Lambda(j)\}$ be a wavelet basis (see Section 3.3.1) with regularity r.

4.1. The nested projection strategy

The first strategy is the one defined in Theorem 1 for the subsets selection case. We look at the family $\mathcal{F}_J$ defined in (3.3). The models S_h's are defined as follows: for all $h\leq J$,

$$S_h=\mathrm{Span}\{\varphi_{j,k}, h\geq j\geq l, k\in\Lambda(j)\}.$$

They are nested, hence polynomial in the sense of Definition 3. As we have seen in Section 3.3.1, the functions $\mathcal{F}_h=\{\varphi_{j,k}, h\geq j\geq l, k\in\Lambda(j)\}$ are an orthonormal localized family of functions in the sense of Definition 4. A consequence of the classical localized property (with constant B) of these wavelets is Assumption 2 of Theorem 1 since $B2^J\simeq T$. The penalty is chosen here with formula (a) of Theorem 1:

$$\mathrm{pen}(m)=c\frac{B|m|N_{\mathbb{X}}}{T^2}\text{ with }c>1.$$

Hence the quadratic risk of the resulting p.p.e. is bounded by

$$\mathbb{E}(\|s-\tilde{s}\|^2)\leq C_0\inf_{l\leq h\leq J}\left\{\|s-s_h\|^2+\mathbb{E}(\mathrm{pen}(h))\right\}+\frac{C_0'}{\mu(\mathbb{X})}.$$

<u>*Rate of convergence*</u>: The lower bound proposed in Corollary 1 for the minimax risk on the set $\mathcal{B}(\rho,R,B^{\alpha}_{2,2})$ is also true, by inclusion, for $\mathcal{B}(\rho,R,B^{\alpha}_{p,\infty})$, with $\alpha>1/p\geq 1/2$, since we have Equation (3.6). Hence (3.8) is the bound we want to compare with the risk of the p.p.e. on these different sets.

- First, what happens on $\mathcal{B}(\rho,R,B^{\alpha}_{2,2})$ ($\alpha<r$), the set where we have computed the lower bound? We denote by β the coefficients of s in the wavelet expansion. The bound of Theorem 1 makes appear the bias term, bounded as follows:

$$\forall l\leq h\leq J,\quad \|s-s_h\|^2\leq\sum_{j>h}\sum_{k\in\Lambda(j)}\beta_{j,k}^2\leq\left(\|s\|^{\alpha}_{2,2}\right)^2\sum_{j>h}2^{-2j\alpha}\leq R^2 2^{-2h\alpha}. \tag{4.1}$$

We minimize the sum of the bias term and the penalty in h. We can verify that the chosen model h is in our family of models, for T large enough. The risk of our estimator is $O\left(\rho^{\frac{2\alpha}{2\alpha+1}} R^{\frac{2}{2\alpha+1}} T^{-\frac{2\alpha}{2\alpha+1}}\right)$. Consequently for T large enough, we reach the minimax risk on $\mathcal{B}(\rho, R, B^{\alpha}_{2,2})$ up to some constant, for all $\alpha < r$.

- If we suppose s in $\mathcal{B}(\rho, R, B^{\alpha}_{2,\infty})$ $(\alpha < r)$ we have the same kind of bound on the bias term which can be found in the last section of [8]:

$$\forall l \leq h \leq J, \quad \|s - s_h\|^2 \leq B(\alpha) R^2 2^{-2\alpha h},$$

for some B continuous positive function. So up to some constant depending on α, we reach the minimax risk too, doing same computations as previously.

- If we suppose s in $\mathcal{B}(\rho, R, B^{\alpha}_{p,\infty})$ $(\alpha < r)$ with $p < 2$, the same strategy leads to a risk, which is too great. Actually, once again following [8], we have that for s in such a set, and $l \leq h \leq J$:

$$\|s - s_h\|^2 \leq B'(\alpha, p) R^2 2^{-2h(\alpha+\frac{1}{2}-\frac{1}{p})},$$

for some B continuous positive function. Doing as previously the compromise between the bias term and the penalty, the risk of our estimator is

$$O\left(\rho^{\frac{2(\alpha+\frac{1}{2}-\frac{1}{p})}{1+2(\alpha+\frac{1}{2}-\frac{1}{p})}} R^{\frac{2}{1+2(\alpha+\frac{1}{2}-\frac{1}{p})}} T^{-\frac{2(\alpha+\frac{1}{2}-\frac{1}{p})}{1+2(\alpha+\frac{1}{2}-\frac{1}{p})}}\right).$$

So, the simple method (using a nested family of models and Theorem 1) does not lead to the minimax risk as well as the other forms of penalty purposed in this theorem. This weakness is related to the poor approximation properties of the family of models considered here in terms of $\mathbb{L}^2$ distance in the Besov spaces $B^{\alpha}_{p,\infty}$ for $p < 2$.

4.2. Thresholding

We now turn to a more complex family of models, using Proposition 1. We use once again the family $\mathcal{F}_J$ (see (3.3)) with, this time, $2^J \simeq T/\ln^2 T$. We can remark that the localized property of $\mathcal{F}_J$ (with constant B) is exactly the assumption we need to apply the theorem. If we denote by Λ the set of indices of the functions in $\mathcal{F}_J$, and if we keep the notations of the subsets selection case (see Proposition 1), we can look at the following family of models: $\mathcal{M}_{\mathbb{X}} = \{m \subset \Lambda\}$, i.e. the collection of all the subsets of Λ. To find the weights L_m, we can remark that

$$\binom{N}{D} \leq (eN/D)^D. \tag{4.2}$$

So, in order to assure that the series converges, we can take, for all m in $\mathcal{M}_{\mathbb{X}}$, $L_m = \ln T$. We set for $c > 1$,

$$\text{pen}(m) = \frac{c|m|(\|\hat{s}_\Lambda\|_\infty + K')}{T}(1 + \sqrt{2\kappa \ln T})^2, \tag{4.3}$$

with $K' > 0$. Therefore the resulting p.p.e. is a hard threshold estimator as mentioned in the introduction.

Rate of convergence: If we want to apply Theorem 1(c), we have to choose K'. If s is in $\mathcal{B}(\rho, R, B^{\alpha}_{p,\infty})$, then $\|s - s_\Lambda\|_\infty \leq \Phi R 2^{-J(\alpha - 1/p)}$. Hence, whatever the choice of K' to construct the estimator, for T large enough and consequently for J large enough, we could apply Proposition 1. As before, we want to compromise between bias and $\frac{M|m|}{T}(1 + \sqrt{2\kappa \ln T})^2$ where M can be taken as $\|s\|_\infty + K'$ since $\|s_\Lambda\|_\infty$ is closed for T large enough to $\|s\|_\infty$.

So, we want to get the good rate of convergence, on $\mathcal{B}(\rho, R, B^{\alpha}_{p,\infty})$ with $r > \alpha > 1/p > 1/2$. By inclusion, this would imply the good rate of convergence on the other subsets. A proposition due to L. Birgé and P. Massart (Proposition 6 of [8]) allows us to find, for all $j' \leq J$ and for $\alpha > 1/p - 1/2$, one $m = m_{j'}$ in $\mathcal{M}_{\mathbb{X}}$, such that

$$|m| \leq C 2^{j'} \text{ and } \|s - s_m\|^2 \leq C' R^2 \left(2^{-2\alpha j'} + 2^{-2J(\alpha + \frac{1}{2} - \frac{1}{p})}\right). \qquad (4.4)$$

Among the $m_{j'}$'s, we choose one such that j' verifies $2^{j'} \simeq \left(\frac{R^2 T}{M \ln T}\right)^{\frac{1}{1+2\alpha}}$, where M designs $\|s\|_\infty + K'$. Moreover m, the chosen model, is in $\mathcal{M}_{\mathbb{X}}$, for T large enough. The risk of our estimator is $O\left(C_\alpha R^{\frac{2}{1+2\alpha}} M^{\frac{2\alpha}{1+2\alpha}} \left(\frac{T}{\ln T}\right)^{-\frac{2\alpha}{2\alpha+1}}\right)$. Therefore, the p.p.e. reaches up to a constant the minimax risk, asymptotically in T, except the presence of a slowly varying term, $\ln T$ and the fact that M replaces ρ.

4.3. Adaptive thresholding

Let Λ be the set of indices of the functions of $\mathcal{F}_J$ (see (3.3)). Let n be $|\Lambda|$. Assume that $n \leq T/(\ln T)^2$. We look at the family of models: $\mathcal{M}_{\mathbb{X}} = \{m \subset \Lambda\}$. We want to use the p.p.e. described in Proposition 1 (c). Since we have Equation (4.2), we can choose $L_m = \ln(n/|m|)$. Consequently we look at penalty given in (4.3) where L_m is no more a constant. The resulting p.p.e. can be viewed as a threshold estimator but with level of thresholding depending on the selected model. In fact, the procedure selects first the good dimension D and then keep the D biggest coefficients.

Rate of convergence:

- First, we assume that s lies in $\mathcal{B}(\rho, R, B^{\alpha}_{p,\infty})$ with $r > \alpha > 1/p > 1/2$. Assume that $n \simeq T/(\ln T)^2$. We get with the same computations as before, the same rate of convergence, since $\ln(n/|m|) = O(\ln(T))$, i.e. the risk is $O\left(C_\alpha R^{\frac{2}{1+2\alpha}} M^{\frac{2\alpha}{1+2\alpha}} \left(\frac{T}{\ln T}\right)^{-\frac{2\alpha}{2\alpha+1}}\right)$.
- Now, let us assume that s lies in $B_{n,D,\rho}$ (see (3.1)) for some D positive integer. Assume that $n > 2^l$ and $n > 4D$. Then we apply Proposition 1. The infimum over $\{m \subset \Lambda\}$ is less than the infimum over only $\{m \subset \Lambda / \ |m| = D\}$. The penalty is then constant and the infimum of the bias is zero. Therefore, we get for T large enough

$$\mathbb{E}(\|s - \tilde{s}\|^2) \leq CM \frac{D \ln \frac{n}{D}}{T},$$

where $M = \|s_\Lambda\|_\infty + K' = \|s\|_\infty + K'$ (we could take $K' = 0$ in this case). This is almost the minimax risk, over $B_{n,D,\rho}$ since for fixed L there exists a positive constant γ, independent of N, such that $\ln(n - L + 1) \geq \gamma \ln(n)$. Then it achieves precisely the good rate in n, D and T.

4.4. Birgé-Massart strategy

We can improve the previous strategy with a special one due to L. Birgé and P. Massart (see [8] and [10]) if we know that s is in a Besov ball without knowing the parameters. We keep as the largest family of models $\mathcal{F}_J$, with $2^J \simeq T/(\ln T)^2$. The family of models is

$$\mathcal{M}''_{\mathbb{X}} = \cup_{0\leq j'\leq J} \mathcal{M}^{j'}_{J_{\mathbb{X}}},$$

where the family is described in paragraph 4.3.2 of [8]. We want again to apply Proposition 1, since the family is not polynomial, but we can choose the weights $L_m = L$ constant independent of T, since the number of models with same cardinality is of order, exponential of a constant times $|m|$. The penalty is taken as in Equation (4.3).

<u>Rate of convergence</u>: With their proposition 6 (which is (4.4)), we can do the same type of computations for s in $\mathcal{B}(\rho, R, B^\alpha_{p,\infty})$ with $r > \alpha > 1/p \geq 1/2$. By inclusion the upper bound on the risk on these sets is also true for $\mathcal{B}(\rho, R, B^\alpha_{2,2})$. Accordingly the risk of p.p.e. is $O\left(R^{\frac{2}{1+2\alpha}} M^{\frac{2\alpha}{1+2\alpha}} T^{-\frac{2\alpha}{2\alpha+1}}\right)$ which is exactly the lower bound of the minimax risk up to some constant and the factor $M = \|s\|_\infty + K'$ which replaces ρ, the normalized integral.

4.5. Adaptivity

Consequently the oracle type inequalities of Section 2 lead us to the adaptive properties of the p.p.e.: the first class of estimator constructed and the nested family, is adaptive because without knowing the smoothness of s (not even precisely the space of regularity), the estimator reaches asymptotically the minimax risk up to some constant, on spaces like $\mathcal{B}(\rho, R, B^\alpha_{2,2})$ or $\mathcal{B}(\rho, R, B^\alpha_{2,\infty})$ $(r > \alpha > 1/2, \rho > 0, R > 0)$. Furthermore, the special Besov-strategy due to L. Birgé and P. Massart allows us to reach asymptotically the minimax risk on all $\mathcal{B}(\rho, R, B^\alpha_{p,\infty})$ $(r > \alpha > 1/p \geq 1/2, \rho > 0, R > 0)$, up to some constant with the lost of the factor ρ which represented the normalized integral of s, replaced by M. Moreover the role of the complexity of the family of models is very important since it allows us to reach the minimax risk on some special spaces.

5. Concentration inequalities for Poisson processes

Now let us show the fundamental probabilistic results which has given us Theorems 1 and 2.

5.1. First properties and a simple concentration inequality

There exist two fundamental properties for Poisson processes. Firstly, for two disjoint sets, the points of N which appear in the first one are independent of what appears in the second one (this is the second part of Definition 1). The second one is that N is infinitely divisible, which means that it can be written as follows for all integer n:

$$dN = \sum_{i=1}^{n} dN_i \tag{5.1}$$

the N_i's being mutually independent Poisson processes on $(\mathbb{X}, \mathcal{X})$ with mean measure ν/n. The first property of Definition 1 leads to the following proposition sometimes attributed to Campbell [25]:

Proposition 6. *For any function f measurable with respect to $\mathcal{X}$, one has:*

$$\mathbb{E}\left(\int_{\mathbb{X}} f(x)dN_x\right) = \int_{\mathbb{X}} f(x)d\nu_x,$$
$$\mathrm{Var}\left(\int_{\mathbb{X}} f(x)dN_x\right) = \int_{\mathbb{X}} f^2(x)d\nu_x,$$
$$\forall\lambda \in \mathbb{R},\ \ \mathbb{E}\left(\exp\left[\lambda \int_{\mathbb{X}} f(x)dN_x\right]\right) = \exp\left(\int_{\mathbb{X}} e^{\lambda f(x)} - 1\ \ d\nu_x\right).$$

A proof of this proposition can be found in [25]. We can derive from Proposition 6 an analogue of Bennett's inequality for sums of independent random variables.

Proposition 7. *For any function f measurable with respect to $\mathcal{X}$, essentially bounded, such that $\int_{\mathbb{X}} f^2(x)d\nu_x > 0$, one has:*

$$\forall\xi > 0,\ \ \mathbb{P}\left(\int_{\mathbb{X}} f(x)(dN_x - d\nu_x) \geq \xi\right)$$
$$\leq \exp\left(-\frac{\int_{\mathbb{X}}\ f^2(x)d\nu_x}{\|f\|_\infty^2} h\left(\frac{\xi\|f\|_\infty}{\int_{\mathbb{X}}\ f^2(x)d\nu_x}\right)\right),$$

where $\forall u > 0,\ h(u) = (1 + u)\ln(1 + u) - u$. *It implies*

$$\forall u > 0,\ \mathbb{P}\left(\int_{\mathbb{X}} f(x)(dN_x - d\nu_x) \geq \sqrt{2u \int_{\mathbb{X}} f^2(x)d\nu_x} + \frac{1}{3}\|f\|_\infty u\right) \leq \exp(-u) \tag{5.2}$$

and also

$$\forall\xi > 0,\ \mathbb{P}\left(\int_{\mathbb{X}} f(x)(dN_x - d\nu_x) \geq \xi\right) \leq \exp\left(-\frac{\xi^2}{2\int_{\mathbb{X}} f^2(x)d\nu_x + \frac{2}{3}\xi\|f\|_\infty}\right).$$

There exists the same upper bounds for $\mathbb{P}\Big(\int_{\mathbb{X}} f(x)(dN_x - d\nu_x) \leq -\xi\Big)$.

This inequality is known and holds for more general functionals than the integrals (see [22], Corollary 5.1). With this inequality, we can control quantities of the form $\int_{\mathbb{X}} f(x)dN_x$, for every single f. We want now to control together a family of such quantities, to control the "chi-square" type statistic (see Equations (1.6) and (1.12)) which we mentioned in the introduction.

5.2. Entropy and tensorisation

Such controls are based on a fundamental property: the tensorisation of the entropy for product spaces [28]. Recapturing Ledoux's method, P. Massart [32] deduced this lemma which allows us to control the entropy of the Laplace transform.

Lemma 1. *Let $(\Omega_1, \mathcal{A}_1), ..., (\Omega_n, \mathcal{A}_n)$ be some measurable spaces and $X_1, ..., X_n$ be independent random variables with values in $\Omega_1, ..., \Omega_n$ respectively. Let ζ be some real valued measurable function on $(\Omega, \mathcal{A}) = (\prod_{i=1}^n \Omega_i, \otimes_{i=1}^n \mathcal{A}_i)$ and $Z = \zeta(X_1, ..., X_n)$. Given some independent random variables $X'_1, ..., X'_n$ with values in $\Omega_1, ..., \Omega_n$ and independent of $X_1, ..., X_n$, let Z^i be the random variable $\zeta(X_1, ..., X_{i-1}, X'_i, X_{i+1}, ..., X_n)$ for all $1 \leq i \leq n$. Let, for any real number z, $\phi(z) = \exp(z) - z - 1$.*
If the Laplace transform $\lambda \to \mathbb{E}(\exp(\lambda Z))$ is finite on some non empty open interval I then for any $\lambda \in I$

$$\lambda \mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z}) \log \mathbb{E}(e^{\lambda Z}) \leq \sum_{i=1}^n \mathbb{E}\left(e^{\lambda Z} \phi(-\lambda(Z - Z^i))\right). \tag{5.3}$$

5.3. Concentration of nonnegative variables

The first concentration inequality, which we are able to prove, is for a supremum of positive variables.

Theorem 3. *Let N be a Poisson process on $(\mathbb{X}, \mathcal{X})$ with finite mean measure ν. Let $\{\psi_a, a \in A\}$ be a countable family of functions with values in $[0, 1]$. One considers*

$$Z = \sup_{a \in A} \int_{\mathbb{X}} \psi_a(x) dN_x.$$

Then for any λ

$$\log \mathbb{E}(e^{\lambda(Z - \mathbb{E}(Z))}) \leq \mathbb{E}(Z)\phi(\lambda), \tag{5.4}$$

where ϕ is defined in Lemma 1.
This result implies that for all $x > 0$

$$\mathbb{P}(Z \geq \mathbb{E}(Z) + \xi) \leq \exp\left(- \mathbb{E}(Z) h\big(\frac{\xi}{\mathbb{E}(Z)}\big)\right) \tag{5.5}$$

and

$$\mathbb{P}(-Z \geq -\mathbb{E}(Z) + \xi) \leq \exp\left(- \mathbb{E}(Z) h\big(\frac{-\xi}{\mathbb{E}(Z)}\big)\right), \tag{5.6}$$

where h is defined in Proposition 7.

This result is a necessary step to obtain a concentration inequality for centered processes as χ_m: when we focus on centered quantities (as appears in Equation (1.12)), a supremum of $\int_{\mathbb{X}} \psi_a^2 dN$ appears and is controlled by this first theorem. The same scheme of proof appears in the n-sample case (see [32]).

5.4. Concentration of centered processes

Hence we obtain a concentration inequality for centered processes which has exactly the same form as the result of P. Massart in the n-sample case (see [32]).

Theorem 4. *Let N be an inhomogeneous Poisson process on $(\mathbb{X}, \mathcal{X})$ with finite mean measure ν. Let $\{\psi_a, a \in A\}$ be a countable family of functions with values in $[-b, b]$. One considers*

$$Z = \sup_{a\in A} \int_{\mathbb{X}} \psi_a(x)(dN_x - d\nu_x) \text{ or } \sup_{a\in A} \left| \int_{\mathbb{X}} \psi_a(x)(dN_x - d\nu_x) \right|.$$

Then for any positive number u

$$\mathbb{P}\big(Z \geq \mathbb{E}(Z) + 2\sqrt{vu} + cbu\big) \leq \exp(-u),$$

where

$$v = \frac{1}{2}\left[\mathbb{E}\left(\sup_{a\in A} \int_{\mathbb{X}} \psi_a^2(x) dN_x\right) + \sup_{a\in A} \int_{\mathbb{X}} \psi_a^2(x) d\nu_x\right]$$

and where c can be taken equal to $5/4$.

The interest of this theorem is to control a family of fluctuations of the process around its mean without any dependence on the size of A. In particular, it allows us to control (in favorable cases) a “continuous family” of ψ_a, like finite dimensional balls of $\mathbb{L}^2$. We can also remark that the form of this inequality is very similar to Equation (5.2). If we apply the previous theorem with only one element in A, we obtain Equation (5.2) up to some multiplicative constants (reasonably large).

Let us also notice that the inequality above depends on

$$\mathbb{E}\left(\sup_{a\in A} \int_{\mathbb{X}} \psi_a^2(x) dN_x\right)$$

which we would like to compare with the supremum of the variances of the centered processes:

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

We can commute the expectation and the supremum, using the symmetrization and contraction inequalities already used in [32] and which are proved in [29]. More precisely, one has (see [32]):

Lemma 2. *Let $\{\theta_a, a \in A\}$ be a finite family of functions with values in $[-1, 1]$. Let $X_1, ..., X_n$ be independent random variables such that for all a in A, and for all $0 \leq i \leq n$, $\mathbb{E}(\theta_a(X_i)) = 0$ and such that the distribution of $\theta_a(X_i)$ is symmetric around 0.*
Then

$$\mathbb{E}\left(\sup_{a\in A}\sum_{i=1}^{n}\theta_a(X_i)^2\right) \leq \sup_{a\in A}\mathbb{E}\left(\sum_{i=1}^{n}\theta_a(X_i)^2\right) + 8\mathbb{E}\left(\sup_{a\in A}\left|\sum_{i=1}^{n}\theta_a(X_i)\right|\right).$$

From this lemma, we can derive the following proposition:

Proposition 8. *Let N be a Poisson process on $(\mathbb{X}, \mathcal{X})$ with finite mean measure ν. Let $\{\psi_a, a \in A\}$ be a countable family of functions with values in $[-b, b]$. If $b = 1/2$, one gets for all $\delta > 0$*

$$\mathbb{E}\left(\sup_{a\in A}\int_{\mathbb{X}}\psi_a^2(x)dN_x\right) \leq \frac{(1+\delta)(2+\delta)}{\delta}\sup_{a\in A}\int_{\mathbb{X}}\psi_a^2(x)d\nu_x + 16\frac{(1+\delta)}{\delta}\mathbb{E}\left(\sup_{a\in A}\left|\int_{\mathbb{X}}\psi_a(x)(dN_x - d\nu_x)\right|\right).$$

We can now update Theorem 4:

Corollary 2. *Let N be a Poisson process on $(\mathbb{X}, \mathcal{X})$ with finite mean measure ν. Let $\{\psi_a, a \in A\}$ be a countable family of functions with values in $[-b, b]$. One considers*

$$Z = \sup_{a\in A}\left|\int_{\mathbb{X}}\psi_a(x)(dN_x - d\nu_x)\right| \text{ and } v_0 = \sup_{a\in A}\int_{\mathbb{X}}\psi_a^2(x)d\nu_x.$$

Then for any positive numbers u and ε:

$$P\big(Z \geq (1+\varepsilon)E(Z) + \sqrt{2\kappa v_0 u} + \kappa(\varepsilon)bu\big) \leq \exp(-u), \tag{5.7}$$

where $\kappa = 6$ and $\kappa(\varepsilon) = 1.25 + 32/\varepsilon$.

Proof. We apply Theorem 4 and Proposition 8. We use the additivity of the square root and the following trick

$$\forall a, b, \theta > 0, \;\; 2ab \leq \theta a^2 + b^2/\theta. \tag{5.8}$$

Optimizing in δ leads to the result. □

The later result is the easiest to use for the statistical applications, that are developed in Section 2. Comparing (5.7) with (1.13), there is an extra linear term. This term is present in Talagrand's inequality too, and is a consequence of the fact that the Poisson law has heavier tail than the Gaussian law.

We can easily derive from Corollary 2 concentration inequalities for

$$\sup_{a\in A}\left\{\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{X}}\psi_a dN^i - \int_{\mathbb{X}}\psi_a s d\mu\right\}$$

i.e. for a sum of i.i.d. infinitely divisible variables. This result is interesting by itself and is not a straightforward application of Talagrand's inequalities [28] since the variables are unbounded.

Note that C. Houdré and N. Privault [22] and L. Wu [39] have also proved concentration inequalities for these Poissonian functionals (and for even more general functionals of infinitely divisible variables) but their results do not imply Corollary 2 since their variance term in this situation are not bounded by just v_0 like ours but by v_0 times some increasing function of $\mu(\mathbb{X})$.

5.5. Consequence for χ_m

At first we need some easy computations to understand why the concentration inequalities are so fundamental for model selection.
The definitions of $\tilde{s}$ and of $\gamma_{\mathbb{X}}$ (see (1.4) and (1.3)) lead, for all m in $\mathcal{M}_{\mathbb{X}}$, to:

$$\gamma_{\mathbb{X}}(\tilde{s}) + \mathrm{pen}(\hat{m}) \le \gamma_{\mathbb{X}}(\hat{s}_m) + \mathrm{pen}(m) \le \gamma_{\mathbb{X}}(s_m) + \mathrm{pen}(m),$$

where $\tilde{s}$ is the p.p.e., $\hat{s}_m$ the projection estimator on each model S_m and s_m the orthogonal projection on S_m.
On the other hand, if we denote

$$\forall f \in \mathbb{L}^2, \nu_{\mathbb{X}}(f) = \int_{\mathbb{X}} f(x) \ \frac{dN_x - s(x)d\mu_x}{\mu(\mathbb{X})}, \tag{5.9}$$

we have that the contrast $\gamma_{\mathbb{X}}$ defined in (1.3) verifies

$$\forall f \in \mathbb{L}^2, \gamma_{\mathbb{X}}(f) = \|f\|^2 - 2 < s, f > -2\nu_{\mathbb{X}}(f) = \|s - f\|^2 - \|s\|^2 - 2\nu_{\mathbb{X}}(f).$$

We get consequently, for all m in $\mathcal{M}_{\mathbb{X}}$:

$$\begin{aligned}\|s-\tilde{s}\|^2 &\le \|s-s_m\|^2 + 2\nu_{\mathbb{X}}(\tilde{s}-s_m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m),\\ &\le \|s-s_m\|^2 + 2\nu_{\mathbb{X}}(\tilde{s}-s_{\hat{m}}) + 2\nu_{\mathbb{X}}(s_{\hat{m}}-s_m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m).\end{aligned}$$

We see by (1.12) that

$$\chi_m = \sup_{f \in S_m} \frac{\nu_{\mathbb{X}}(f)}{\|f\|} = \sqrt{\nu_{\mathbb{X}}\left(\hat{s}_m - s_m\right)} = \|\hat{s}_m - s_m\|.$$

Then we get for all m in $\mathcal{M}_{\mathbb{X}}$:

$$\|s-\tilde{s}\|^2 \le \|s-s_m\|^2 + 2\chi_{\hat{m}}^2 + 2\nu_{\mathbb{X}}(s_{\hat{m}}-s_m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m). \tag{5.10}$$

In order to derive from (5.10), some oracle inequality, we see that $\mathrm{pen}(\hat{m})$ should be of the order of $\chi_{\hat{m}}^2$ while $\mathbb{E}(\nu_{\mathbb{X}}(s_{\hat{m}} - s_m))$ should be close to 0 (which would be exact if $\hat{m}$ were deterministic). Hence we have to understand the behavior of the quantity $\chi_{\hat{m}}^2$. The difficulty comes from the fact that $\chi_{\hat{m}}$ is doubly random: for deterministic m, χ_m is random and $\hat{m}$, i.e. the choice of the model, is random. This is precisely the reason why concentration inequalities on χ_m are so useful if we want to deal also with $\chi_{\hat{m}}$.

Remark. One of the intuitive reasons for which these quantities behave like the square root of a chi-square statistics is that this is a square root of a sum of centered quantities to the square. Moreover, if the basis of S_m are functions with disjoint supports, Definition 1 of N implies that χ_m^2 is a sum of independent centered quantities.

As we see in (1.12), χ_m is a supremum of integral functionals: therefore we can use Corollary 2. If we apply Corollary 2 brutally then we set Inequality (5.11) below, which can turn to be too rough for our needs (especially for dealing with the problem of complete subset selection from an orthonormal basis). The derivation of (5.12) is somehow more subtle and will replace (5.11) in situations where (5.11) is too weak.

Proposition 9. *Let N be a Poisson process on $(\mathbb{X}, \mathcal{X})$ with intensity s in $\mathbb{L}^2$. Let S be a finite dimensional linear subspace of $\mathbb{L}^2$, $\bar{s}$ designs the orthogonal projection of s on S and $\hat{s}$ designs the projection estimator of s over S (see (1.1)).*
Let $\chi(S) = \|\hat{s} - \bar{s}\|$, $M_S = \sup_{f \in S, \|f\|=1} \int_{\mathbb{X}} f^2 s d\mu / \mu(\mathbb{X})$ and $B_S = \sup_{f \in S, \|f\|=1} \|f\|_\infty$, and assume that all these quantities are finite.
Then for all ε and u positive:

$$\mathbb{P}\left(\chi(S) \geq (1+\varepsilon)\sqrt{\mathbb{E}(\chi^2(S))} + \sqrt{\frac{2\kappa M_S u}{\mu(\mathbb{X})}} + \kappa(\varepsilon)\frac{B_S}{\mu(\mathbb{X})}u\right) \leq \exp(-u) \tag{5.11}$$

and for all $M \geq M_S$, on the event $\Omega_S(\varepsilon) = \left\{\|\hat{s} - \bar{s}\|_\infty \leq (2\kappa\varepsilon M)/\kappa(\varepsilon)\right\}$

$$\mathbb{P}\left(\chi(S)\mathbb{1}_{\Omega_S(\varepsilon)} \geq (1+\varepsilon)\left(\sqrt{\mathbb{E}(\chi^2(S))} + \sqrt{\frac{2\kappa M u}{\mu(\mathbb{X})}}\right)\right) \leq \exp(-u). \tag{5.12}$$

where κ and $\kappa(\varepsilon)$ are given in Corollary 2.

We can remark that in the first part we describe the behavior of $\chi(S)$ over all the probability space, but there is an extra linear term, when we compare it with the Gaussian concentration (see [15]). It represents the fact that Poisson variables have heavier tails than Gaussian. For a certain kind of statistic aims, this term is to large: we prefer then to restrain us to a large set of probability, on which χ behaves like a Gaussian, i.e. without the linear term. This trick is inspired by P. Massart [31] and can be found in the PhD Thesis of G. Castellan [13], who have used it in the context of density estimation from a n-sample.

6. Proofs

6.1. Concentration theorems

These proofs are based on the scheme of proof of M. Ledoux and P. Massart in the n-sample framework (see [28] and [32]). The second one is inspired by the scheme of proof of E. Rio in the n-sample framework [36].

6.1.1. Proof of Theorem 3

Proof. By monotone convergence, it is sufficient to prove it for a finite family of functions. N can be written as $dN = \sum_{i=1}^{n} dN^i$, as in Equation (5.1), with N^i independent Poisson processes with mean measure ν/n, where ν is the mean measure of N. Then we can write

$$Z = \sup_{a\in A} \int_{\mathbb{X}} \psi_a(x) \sum_{i=1}^{n} dN_x^i.$$

The sigma-fields generated by each N^i are independent. Consequently we can apply Lemma 1 where

$$Z^i = \sup_{a\in A} \int_{\mathbb{X}} \psi_a(x) \sum_{j\neq i} dN_x^j.$$

We obtain

$$\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z}) \leq \mathbb{E}\left[e^{\lambda Z}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\right]. \tag{6.1}$$

Let Ω_n be the event $\{\forall i,\, N_{\mathbb{X}}^i \leq 1\}$. We have

$$\mathbb{P}(\Omega_n^c) \leq n\mathbb{P}(N_{\mathbb{X}}^1 \geq 2) \leq \frac{\nu(\mathbb{X})^2}{n}.$$

So, Equation (6.1) becomes by Cauchy-Schwarz:

$$\begin{aligned}\lambda\mathbb{E}(Ze^{\lambda Z}) &- \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z})\\ &\leq \mathbb{E}\left[e^{\lambda Z}\mathbb{1}_{\Omega_n}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\right]\\ &+\sqrt{\frac{\nu(\mathbb{X})^2}{n}\mathbb{E}\left[e^{2\lambda Z}\Big(\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\Big)^2\right]}.\end{aligned} \tag{6.2}$$

But

$$\mathbb{E}\left[e^{\lambda Z}\mathbb{1}_{\Omega_n}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\right] = \mathbb{E}\left[e^{\lambda Z}\mathbb{1}_{\Omega_n}\sum_{T\in N}\phi(-\lambda(Z-Z^T))\right]$$

where

$$Z^T = \sup_{a\in A}\sum_{X\in N,\, X\neq T}\psi_a(X).$$

As $\mathbb{1}_{\Omega_n}$ tends, when n tends to infinity, to 1, we have by dominated convergence that

$$\mathbb{E}\left[e^{\lambda Z}\mathbb{1}_{\Omega_n}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\right] \underset{n\to\infty}{\to} \mathbb{E}\left[e^{\lambda Z}\sum_{T\in N}\phi(-\lambda(Z-Z^T))\right].$$

For the second term in (6.2), as

$$\mathbb{E}\left[e^{2\lambda Z}\Big(\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\Big)^2\right] \le \mathbb{E}\left[e^{2\lambda N_{\mathbb{X}}}\lambda^2 N_{\mathbb{X}}^4\right] < \infty,$$

if $\lambda > 0$ and

$$\mathbb{E}\left[e^{2\lambda Z}\Big(\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\Big)^2\right] \le \mathbb{E}\left[N_{\mathbb{X}}^2\right] < \infty,$$

if $\lambda < 0$, the second term tends to 0 when n tends to infinity. Hence we get

$$\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z}) \le \mathbb{E}\left[e^{\lambda Z}\sum_{T\in N}\phi(-\lambda(Z-Z^T))\right]. \qquad (6.3)$$

Since the supremum in Z is attained at $\hat{a}$, we have, for all $T \in N$,

$$0 \le Z - Z^T \le \psi_{\hat{a}}(T) \le 1.$$

Note that if $x \in [0,1]$, we have $\phi(-\lambda x) \le \phi(-\lambda)x$, for all $\lambda > 0$. So Equation (6.3) becomes, for all $\lambda > 0$

$$\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z}) \le \phi(-\lambda)\mathbb{E}(Ze^{\lambda Z}).$$

Then, we only need to follow P. Massart's proof [32] and the result follows. □

6.1.2. Proof of Theorem 4

Proof. By monotone convergence, it is again sufficient to prove it for a finite family of functions. By homogeneity, we can suppose that $b = 1$. We set

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

As N is infinitely divisible, we can write as in Equation (5.1):

$$\forall n \in \mathbb{N}^*, \quad dN = \sum_{i=1}^{n} dN^i,$$

with N^i's mutually independent Poisson processes with mean measure ν/n. We set

$$\forall i \in \{1,\dots,n\},\, Z^i = \sup_{a\in A}\int_{\mathbb{X}}\psi_a(x)\sum_{j\neq i}(dN_x^j - \frac{1}{n}d\nu_x).$$

We can apply then Lemma 1 to write

$$\forall\lambda > 0,\ \lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\ln\mathbb{E}(e^{\lambda Z}) \le \mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\right).$$

We split the expectation in two parts:

$$\begin{aligned}\forall\lambda > 0,\ \ &\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\ln\mathbb{E}(e^{\lambda Z})\\ &\leq \mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\mathbb{1}_{Z-Z^i\geq 0}\right)\\ &+\mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\mathbb{1}_{Z-Z^i<0}\right).\end{aligned}\tag{6.4}$$

For the first expectation, we have that, for all u positive, $\phi(-u)\leq u^2/2$, to obtain:

$$\begin{aligned}\forall\lambda > 0,\ \ &\mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\mathbb{1}_{Z-Z^i\geq 0}\right)\\ &\leq \frac{\lambda^2}{2}\mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}(Z-Z^i)_+^2\mathbb{1}_{Z-Z^i\geq 0}\right).\end{aligned}$$

On the event $\{Z-Z^i\geq 0\}$, we have

$$0\leq Z-Z^i\leq\int_{\mathbb{X}}\psi_{\hat{a}}(x)(dN_x^i-\frac{1}{n}d\nu_x)$$

where $\hat{a}$ is the index where the supremum in Z is attained. This leads us, for all $\lambda > 0$, to:

$$\begin{aligned}&\mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\mathbb{1}_{Z-Z^i\geq 0}\right)\\ &\leq\frac{\lambda^2}{2}\mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\left(\int_{\mathbb{X}}\psi_{\hat{a}}(x)(dN_x^i-\frac{1}{n}d\nu_x)\right)_+^2\right).\end{aligned}\tag{6.5}$$

As $\mathbb{P}(N_{\mathbb{X}}^i\geq 2)\leq(\nu(\mathbb{X})/n)^2/2$, if $\Omega_n=\{\forall i,\ N_{\mathbb{X}}^i\leq 1\}$, we have that:

$$p_n=\mathbb{P}(\Omega_n^c)\leq\frac{\nu(\mathbb{X})^2}{2n^2}.$$

So we can split the last expectation of Equation (6.5) in two parts, and by Cauchy-Schwarz, we obtain:

$$\begin{aligned}&\mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\left(\int_{\mathbb{X}}\psi_{\hat{a}}(x)(dN_x^i-\frac{1}{n}d\nu_x)\right)_+^2\right)\\ &\leq\mathbb{E}\left(e^{\lambda Z}\mathbb{1}_{\Omega_n}\left[\sum_{X\in N}\psi_{\hat{a}}^2(X)-2\sum_{X\in N}\psi_{\hat{a}}(X)_+\frac{\int\psi_{\hat{a}}d\nu}{n}+\frac{(\int\psi_{\hat{a}}d\nu)^2}{n}\right]\right)\\ &+\sqrt{p_n}\sqrt{E\left(e^{2\lambda Z}\left[\sum_{i=1}^{n}\left(\int_{\mathbb{X}}\psi_{\hat{a}}(x)(dN_x^i-\frac{1}{n}d\nu_x)\right)_+^2\right]^2\right)}.\end{aligned}$$

By a dominated convergence theorem (since Poisson law has Laplace transform), we obtain:

$$\limsup_{n\to+\infty} \mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\left(\int_{\mathbb{X}}\psi_{\hat{a}}(x)(dN_x^i - \frac{1}{n}d\nu_x)\right)_+^2\right) \leq \mathbb{E}\left(e^{\lambda Z}\int_{\mathbb{X}}\psi_{\hat{a}}^2 dN\right). \tag{6.6}$$

For the second expectation of Equation (6.4), we remark that:

$$\sum_{i=1}^{n} e^{\lambda Z}\phi(-\lambda(Z-Z^i))\mathbb{1}_{Z-Z^i<0} \leq \frac{\lambda^2}{2}\sum_{i=1}^{n} e^{\lambda Z^i}(Z^i - Z)_+^2.$$

We have that

$$Z^i - Z \leq \int_{\mathbb{X}} -\psi_{\hat{a}_i}(x)(dN_x^i - \frac{1}{n}d\nu_x) \tag{6.7}$$

where $\hat{a}_i$ which denotes the index where the supremum in Z^i is attained.
NB: This index $\hat{a}_i$ depends only on the processes N^j for $j \neq i$. Consequently $\hat{a}_i$ is independent of N^i, Poisson process with intensity s/n. Hence we obtain:

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} e^{\lambda Z^i}(Z^i - Z)_+^2\right) &\leq \mathbb{E}\left(\sum_{i=1}^{n} e^{\lambda Z^i}\left[\int_{\mathbb{X}}\psi_{\hat{a}_i}(x)(dN_x^i - \frac{1}{n}d\nu_x)\right]^2\right)\\
&\leq \sum_{i=1}^{n}\mathbb{E}\left(e^{\lambda Z^i}\mathbb{E}\left(\left[\int_{\mathbb{X}}\psi_{\hat{a}_i}(x)(dN_x^i - \frac{1}{n}d\nu_x)\right]^2 \middle| N^j, j\neq i\right)\right)\\
&\leq \sum_{i=1}^{n}\mathbb{E}\left(e^{\lambda Z^i}\int_{\mathbb{X}}\psi_{\hat{a}_i}^2(x)\frac{1}{n}d\nu_x\right)\\
&\leq \sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2(x)\frac{1}{n}d\nu_x\right)\sum_{i=1}^{n}\mathbb{E}\left(e^{\lambda Z^i}\right).
\end{aligned}$$

Moreover using again Equation (6.7) and Jensen inequality, we have that

$$\begin{aligned}
&\mathbb{E}\left(e^{\lambda Z}\middle| N^j, j\neq i\right) \geq \exp\left[\lambda\mathbb{E}(Z|N^j, j\neq i)\right]\\
&\geq \exp\left[\lambda Z^i\right]\exp\left[\lambda\mathbb{E}\left(\int_{\mathbb{X}}\psi_{\hat{a}_i}(x)(dN_x^i - \frac{1}{n}d\nu_x)\middle| N^j, j\neq i\right)\right] = \exp\left[\lambda Z^i\right].
\end{aligned}$$

This previous argument is exactly the same as in E. Rio's work [36]. We obtain consequently that:

$$\mathbb{E}\left(\sum_{i=1}^{n} e^{\lambda Z}\phi(-\lambda(Z-Z^i))\mathbb{1}_{Z-Z^i<0}\right) \leq \sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2 d\nu\right)\mathbb{E}(e^{\lambda Z}). \tag{6.8}$$

NB: We can do the same thing if Z is defined with absolute values, defining Z^i with absolute values. We obtain exactly the same result.

We obtain when n tends to infinity, with equations (6.6) and (6.8), that for all λ positive

$$\begin{aligned}&\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z})\\&\quad\le \frac{\lambda^2}{2}\mathbb{E}\left(e^{\lambda Z}\left(\int\psi_{\hat{a}}^2(x)dN_x + \sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2(x)d\nu_x\right)\right)\right).\end{aligned}$$

So we can write $\forall\lambda > 0$,

$$\begin{aligned}&\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z})\\&\quad\le \frac{\lambda^2}{2}\mathbb{E}\left(e^{\lambda Z}\left(\sup_{a\in A}\left[\int\psi_a^2(x)dN_x\right] + \sup_{a\in A}\left[\int\psi_a^2(x)d\nu_x\right]\right)\right).\end{aligned}\tag{6.9}$$

If we set $\tilde{Z} = Z - \mathbb{E}(Z)$, inequality (6.9) becomes

$$\begin{aligned}&\lambda\mathbb{E}(\tilde{Z}e^{\lambda\tilde{Z}}) - \mathbb{E}(e^{\lambda\tilde{Z}})\log\mathbb{E}(e^{\lambda\tilde{Z}})\\&\quad\le \frac{\lambda^2}{2}\mathbb{E}\left(e^{\lambda\tilde{Z}}\left(\sup_{a\in A}\left[\int\psi_a^2(x)dN_x\right] + \sup_{a\in A}\left[\int\psi_a^2(x)d\nu_x\right]\right)\right).\end{aligned}\tag{6.10}$$

We set $V_1 = \sup\limits_{a\in A}\left[\int\psi_a^2(x)dN_x\right]$, $v_0 = \sup\limits_{a\in A}\left[\int\psi_a^2(x)d\nu_x\right]$ and $v_1 = \mathbb{E}(V_1)$. We can apply the following lemma obtained by P. Massart [32].

Lemma 3. *Let V and Y be some random variables and $\lambda > 0$ such that $e^{\lambda V}$ and $e^{\lambda Y}$ are integrable. Then, if $\mathbb{E}(Y) = 0$, one has*

$$\mathbb{E}(Ve^{\lambda Y}) \le \mathbb{E}(Ye^{\lambda Y}) + \frac{\log\mathbb{E}(e^{\lambda V})}{\lambda}\mathbb{E}(e^{\lambda Y}).\tag{6.11}$$

Now equation (6.10) becomes

$$\begin{aligned}\lambda\mathbb{E}(\tilde{Z}e^{\lambda\tilde{Z}}) - \mathbb{E}(e^{\lambda\tilde{Z}})\log\mathbb{E}(e^{\lambda\tilde{Z}}) &\le \frac{\lambda^2}{2}v_0\mathbb{E}(e^{\lambda\tilde{Z}})\\&\quad + \frac{\lambda^2}{2}\frac{\log\mathbb{E}(e^{\lambda V_1})}{\lambda}\mathbb{E}(e^{\lambda\tilde{Z}}) + \frac{\lambda^2}{2}\mathbb{E}(\tilde{Z}e^{\lambda\tilde{Z}}).\end{aligned}$$

But Theorem 3 allows us to control the Laplace transform of V_1. So

$$\log\mathbb{E}(e^{\lambda V_1}) \le v_1(\lambda + \phi(\lambda)).$$

Hence we obtain

$$\lambda\mathbb{E}(\tilde{Z}e^{\lambda\tilde{Z}}) - \mathbb{E}(e^{\lambda\tilde{Z}})\log\mathbb{E}(e^{\lambda\tilde{Z}}) \le \lambda^2\left[v\mathbb{E}(e^{\lambda\tilde{Z}}) + v\frac{\phi(\lambda)}{\lambda}\mathbb{E}(e^{\lambda\tilde{Z}}) + \mathbb{E}(\tilde{Z}e^{\lambda\tilde{Z}})\right].$$

Now, we follow P. Massart's proof [32] to get the result. □

6.1.3. Proof of Proposition 8

Proof. Let

$$V_1 = \sup_{a \in A} \int_{\mathbb{X}} \psi_a^2(x) dN_x.$$

Conditionally to $\{N_{\mathbb{X}} = n\}$, the law of V_1 is the same as that of

$$\sup\{\sum_{i=1}^{n} \psi_a^2(T_i), a \in A\}$$

where $T_1, ..., T_n$ are independent, identically distributed random variables with density on $\mathbb{X}$, $s(x)/\int_{\mathbb{X}} s(x)d\mu_x$. If we consider some $(T_1', ..., T_n')$ i.i.d. random variables, with the same law as $(T_1, ..., T_n)$ and independent of them, we remark that by Jensen's inequality

$$\begin{aligned}
&\mathbb{E}_{(T_i),(T_i')}\left(\sup_{a\in A}\sum_{i=1}^{n}(\psi_a(T_i) - \psi_a(T_i'))^2\right)\\
&\qquad \geq \mathbb{E}_{(T_i)}\left(\sup_{a\in A}\sum_{i=1}^{n}\mathbb{E}_{(T_i')}((\psi_a(T_i) - \psi_a(T_i'))^2)\right)\\
&= \mathbb{E}_{(T_i)}\left(\sup_{a\in A}\sum_{i=1}^{n}\left(\psi_a^2(T_i) - 2\psi_a(T_i)\frac{\int_{\mathbb{X}}\psi_a(x)d\nu_x}{\nu(\mathbb{X})} + \frac{\int_{\mathbb{X}}\psi_a^2(x)d\nu_x}{\nu(\mathbb{X})}\right)\right).
\end{aligned}$$

Furthermore, we notice that if we fix some $\delta > 0$, we have

$$\begin{aligned}
2\psi_a(T_i)\frac{\int_{\mathbb{X}}\psi_a(x)d\nu_x}{\nu(\mathbb{X})} &\leq \frac{1}{1+\delta}\psi_a^2(T_i) + (1+\delta)\left(\frac{\int_{\mathbb{X}}\psi_a(x)d\nu_x}{\nu(\mathbb{X})}\right)^2\\
&\leq \frac{1}{1+\delta}\psi_a^2(T_i) + (1+\delta)\frac{\int_{\mathbb{X}}\psi_a^2(x)d\nu_x}{\nu(\mathbb{X})}.
\end{aligned}$$

So we obtain

$$\begin{aligned}
&\frac{\delta}{1+\delta}\mathbb{E}(V_1|N_{\mathbb{X}} = n) - \delta\frac{n}{\nu(\mathbb{X})}\sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2(x)d\nu_x\right)\\
&\qquad \leq \mathbb{E}_{(T_i),(T_i')}\left(\sup_{a\in A}\sum_{i=1}^{n}(\psi_a(T_i) - \psi_a(T_i'))^2\right). \qquad (6.12)
\end{aligned}$$

We can apply Lemma 2 with $X_i = (T_i, T_i')$ and $\theta_a(X_i) = \psi_a(T_i) - \psi_a(T_i')$, for all i in $\{1, ..., n\}$ and a in A as we have assumed $b = 1/2$. Then (6.12) becomes

$$\begin{aligned}\frac{\delta}{1+\delta}\mathbb{E}(V_1|N_{\mathbb{X}} = n) - \delta\frac{n}{\nu(\mathbb{X})}\sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2(x)d\nu_x\right) \\ \leq \sup_{a\in A}\mathbb{E}_{(T_i),(T_i')}\left(\sum_{i=1}^{n}(\psi_a(T_i) - \psi_a(T_i'))^2\right) \\ + 8\,\mathbb{E}_{(T_i),(T_i')}\left(\sup_{a\in A}|\sum_{i=1}^{n}\psi_a(T_i) - \psi_a(T_i')|\right).\end{aligned}$$

Finally, inserting $\int_{\mathbb{X}}\psi_a(x)d\nu_x$ in the last supremum, we get:

$$\begin{aligned}&\frac{\delta}{1+\delta}\mathbb{E}(V_1|N_{\mathbb{X}} = n) - \delta\frac{n}{\nu(\mathbb{X})}\sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2(x)d\nu_x\right) \\ &\leq \frac{2n}{\int_{\mathbb{X}}d\nu_x}\sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2(x)d\nu_x\right) + 16\,\mathbb{E}\left(\sup_{a\in A}\left|\sum_{i=1}^{n}\psi_a(T_i) - \int_{\mathbb{X}}\psi_a(x)d\nu_x\right|\right).\end{aligned}$$

It remains to integrate over $N_{\mathbb{X}}$ and the proposition follows. □

6.1.4. Proof of Proposition 9

Proof. Let $\{\varphi_1, ..., \varphi_D\}$ be an orthonormal basis of S. We are going to prove that for every finite family of measurable bounded functions $\{\varphi_1, ..., \varphi_D\}$, the quantity

$$\chi(S) = \sqrt{\sum_{i=1}^{D}\left(\int_{\mathbb{X}}\varphi_i\frac{dN - sd\mu}{\mu(\mathbb{X})}\right)^2}$$

is concentrated around its mean.

We remind that

$$M_S = \sup_{f\in S, \|f\|=1}\int_{\mathbb{X}}f^2 s\frac{d\mu}{\mu(\mathbb{X})} \text{ and } B_S = \sup_{f\in S, \|f\|=1}\|f\|_\infty.$$

We can assume that the φ_i's are bounded, otherwise B_S is infinite.

First, $\chi(S)$ can be interpreted as the following supremum:

$$\chi(S) = \sup_{a\in A}\int_{\mathbb{X}}\frac{\sum_{i=1}^{D}a_i\varphi_i}{\mu(\mathbb{X})}(dN - sd\mu)$$

where A is a dense countable subset of the unit ball for $\|.\|_2$ of $\mathbb{R}^D$, since the integral functionals in the supremum are continuous in a. Because of the same kind of continuity, the suprema in M_S and B_S can be taken on A and they can easily be interpreted in term of φ_i: they are exactly the terms which appear in concentration formula, up to the factor $\mu(\mathbb{X})$. Then we can apply Corollary 2 to obtain for all ε and u positive:

$$\mathbb{P}\left(\chi(S) \geq (1+\varepsilon)\mathbb{E}(\chi(S)) + \sqrt{\frac{2\kappa M_S u}{\mu(\mathbb{X})}} + \kappa(\varepsilon)\frac{B_S u}{\mu(\mathbb{X})}\right) \leq \exp(-u).$$

As $\mathbb{E}(\chi(S)) \leq \sqrt{\mathbb{E}(\chi(S))^2}$ by Cauchy-Schwarz inequality, we obtain exactly the first point.

Secondly, we can remark that $\chi(S)$ is attained at $\hat{a}$ which verifies, for all i, $\hat{a}_i = \nu_{\mathbb{X}}(\varphi_i)/\chi(S)$. This implies that on $\Omega_S(\varepsilon)$, $\|\sum_{i=1}^D \hat{a}_i\varphi_i\|_\infty \leq C(\varepsilon)/z$, where z is a lower bound for $\chi(S)$ and where $C(\varepsilon) = 2\kappa\varepsilon M/\kappa(\varepsilon)$.
Then we introduce

$$\chi' = \sup_{a\in\mathcal{B}} \int_{\mathbb{X}} \frac{\sum_{i=1}^D a_i\varphi_i}{\mu(\mathbb{X})}(dN - sd\mu),$$

where $\mathcal{B} = \{a \in \mathbb{R}^D/\|a\|_2 = 1 \text{ and } \|\sum_{i=1}^D a_i\varphi_i\|_\infty \leq C(\varepsilon)/z\}$.
On the event $\Omega_S(\varepsilon) \cap \{\chi(S) \geq z\}$, we have $\chi' = \chi(S)$.
We can apply Corollary 2 to χ', restricting us as in the first point to a dense countable subset of $\mathcal{B}$. The variance term which appears, can be upper bounded by $M/\mu(\mathbb{X})$. Hence we obtain the following inequality, for all ε and u positive:

$$\mathbb{P}\left(\chi' \geq (1+\varepsilon)\mathbb{E}(\chi') + \sqrt{\frac{2\kappa M u}{\mu(\mathbb{X})}} + \kappa(\varepsilon)\frac{C(\varepsilon)u}{z\mu(\mathbb{X})}\right) \leq \exp(-u).$$

As $\mathcal{B}$ is a subset of the unit ball, we have $\mathbb{E}(\chi') \leq \mathbb{E}(\chi(S))$. If we take $z = \sqrt{\frac{2\kappa M u}{\mu(\mathbb{X})}}$, we obtain that:

$$\mathbb{P}\left(\chi' \geq (1+\varepsilon)\left(\mathbb{E}(\chi(S)) + \sqrt{\frac{2\kappa M u}{\mu(\mathbb{X})}}\right)\right) \leq \exp(-u).$$

Moreover as $\chi' = \chi(S)$ on the event $\Omega_S(\varepsilon) \cap \{\chi(S) \geq z\}$ and as $\{\chi(S) \geq z\}$ is true on the event $\chi(S)\mathbb{1}_{\Omega_S(\varepsilon)} \geq (1+\varepsilon)\left(\mathbb{E}(\chi) + \sqrt{\frac{2\kappa M u}{\mu(\mathbb{X})}}\right)$, we obtain:

$$\mathbb{P}\left(\chi(S)\mathbb{1}_{\Omega_S(\varepsilon)} \geq (1+\varepsilon)\left(\mathbb{E}(\chi(S)) + \sqrt{\frac{2\kappa M u}{\mu(\mathbb{X})}}\right)\right) \leq \exp(-u).$$

As by Cauchy-Schwarz $\mathbb{E}(\chi(S)) \leq \sqrt{\mathbb{E}(\chi(S))^2}$, we obtain exactly the second point. □

6.2. *Model selection Theorems*

6.2.1. Proof of Theorem 1

Proof. Let m be a fixed index in $\mathcal{M}_{\mathbb{X}}$. We start with Equation (5.10).

1. First, we have to control $\nu_{\mathbb{X}}(s_{\hat{m}} - s_m)$. For this aim, we are going to control every $\nu_{\mathbb{X}}(s_{m'} - s_m)$, $\forall m' \in \mathcal{M}_{\mathbb{X}}$. Let $(x_{m'})_{m' \in \mathcal{M}_{\mathbb{X}}}$ be a family of positive number which we will choose later. By application of Proposition 7 and more precisely Equation (5.2), we obtain that with probability larger than $1 - \sum_{m' \in \mathcal{M}_{\mathbb{X}}} e^{-x_{m'}}$,

$$\forall m' \in \mathcal{M}_{\mathbb{X}},\ \nu_{\mathbb{X}}(s_{m'} - s_m) \le \sqrt{2x_{m'} \int_{\mathbb{X}} \frac{(s_{m'} - s_m)^2}{\mu(\mathbb{X})^2} s d\mu} + \frac{1}{3} \frac{\|s_{m'} - s_m\|_\infty x_{m'}}{\mu(\mathbb{X})}.$$

We remark that

- $\int_{\mathbb{X}} \frac{(s_{m'} - s_m)^2}{\mu(\mathbb{X})^2} s d\mu \le \|s_{m'} - s_m\|^2 \frac{\|s\|_\infty}{\mu(\mathbb{X})}$, and
- $\|s_{m'} - s_m\|_\infty \le \|s_{m'}\|_\infty + \|s_m\|_\infty$. This implies by Assumption 2 that

$$\|s_{m'} - s_m\|_\infty \le \sqrt{\mathbb{D}_{m'}} \|s_{m'}\| + \sqrt{\mathbb{D}_m} \|s_m\|,$$

 and then that $\|s_{m'} - s_m\|_\infty \le \left(\sqrt{\mathbb{D}_{m'}} + \sqrt{\mathbb{D}_m}\right) \|s\|$.

As $\hat{m}$ is in $\mathcal{M}_{\mathbb{X}}$, we have that (using (5.8)) for all positive θ, with probability larger than $1 - \sum_{m' \in \mathcal{M}_{\mathbb{X}}} e^{-x_{m'}}$

$$\nu_{\mathbb{X}}(s_{\hat{m}} - s_m) \le \theta \|s_{\hat{m}} - s_m\|^2 + \left[\frac{\|s\|_\infty}{2\theta} + \frac{1}{3}(\sqrt{\mathbb{D}_{\hat{m}}} + \sqrt{\mathbb{D}_m})\|s\|\right] \frac{x_{\hat{m}}}{\mu(\mathbb{X})}.$$

This leads for all positive η to (using (5.8))

$$\nu_{\mathbb{X}}(s_{\hat{m}} - s_m) \le \theta \|s_{\hat{m}} - s_m\|^2 + \frac{\eta\rho\mathbb{D}_m}{3\mu(\mathbb{X})} + \frac{\eta\rho\mathbb{D}_{\hat{m}}}{3\mu(\mathbb{X})} + \frac{\|s\|_\infty x_{\hat{m}}}{2\theta\mu(\mathbb{X})} + \frac{\|s\|^2 x_{\hat{m}}^2}{6\eta\rho\mu(\mathbb{X})}.$$

Let δ, ξ be positive numbers. We choose the $(x_{m'})$ as follows: for all m' in $\mathcal{M}_{\mathbb{X}}$,

$$x_{m'} = \delta\rho\sqrt{D_{m'}} \left[\frac{1}{\|s\|_\infty} \wedge \frac{1}{\|s\|}\right] + \xi.$$

Let us denote by $\mathcal{E}$, a bound on $\sum_{m' \in \mathcal{M}_{\mathbb{X}}} \exp(-\delta\rho\sqrt{D_{m'}} \left[\frac{1}{\|s\|_\infty} \wedge \frac{1}{\|s\|}\right])$. Since $\mathcal{M}_{\mathbb{X}}$ is polynomial, $\mathcal{E}$ can depend only on Γ, R, $\|s\|$, $\|s\|_\infty$, ρ, δ, η, θ but no more on the family of models or on $\mu(\mathbb{X})$. Then with probability larger than $1 - \mathcal{E}e^{-\xi}$, we have that since $\sqrt{D_m} \le D_m \le \mathbb{D}_m$

$$\nu_{\mathbb{X}}(s_{\hat{m}} - s_m) \le \theta \|s_{\hat{m}} - s_m\|^2 + \frac{\eta}{3} \frac{\rho\mathbb{D}_m}{\mu(\mathbb{X})} + \left[\frac{\eta}{3} + \frac{\delta}{2\theta} + \frac{\delta^2}{3\eta}\right] \frac{\rho\mathbb{D}_{\hat{m}}}{\mu(\mathbb{X})}$$
$$+ \frac{\|s\|_\infty}{2\theta} \frac{\xi}{\mu(\mathbb{X})} + \frac{\|s\|^2}{3\eta\rho} \frac{\xi^2}{\mu(\mathbb{X})}. \tag{6.13}$$

2. If we go back to Equation (5.10), we remark that $\chi^2_{\hat{m}} = \nu_{\mathbb{X}}(\hat{s}_{\hat{m}} - s_{\hat{m}}) = \|\hat{s}_{\hat{m}} - s_{\hat{m}}\|^2$. Let us denote by $A_m = \rho\mathbb{D}_m/\mu(\mathbb{X})$. Using (5.8), we obtain that for all $\gamma > 1$ and $\beta > 0$ (which fix δ, η, θ), with probability larger than $1 - \mathcal{E}e^{-\xi}$
$$D(\gamma,\beta)\|\tilde{s} - s\|^2 \leq D'(\gamma,\beta)\|s_m - s\|^2 + \gamma\chi^2_{\hat{m}} + D''(\gamma,\beta)A_m + \beta A_{\hat{m}} + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + \frac{f(\xi)}{\mu(\mathbb{X})} \tag{6.14}$$
where D, D', D'' are proper positive continuous functions and where f does not depend on $\mathcal{M}_{\mathbb{X}}$ and $\mu(\mathbb{X})$, depends continuously on the other parameters and is a polynomial of degree 2 of ξ.
3. Now, we have to control $\chi^2_{\hat{m}}$. In fact we control in fact all the $\chi_{m'}$ for m' in $\mathcal{M}_{\mathbb{X}}$. For this aim, we use the first part of Proposition 9. Let $(y_{m'})_{m'\in\mathcal{M}_{\mathbb{X}}}$ be a family of positive number which we will choose later. We obtain on a set of probability included in the previous one with probability larger than $1 - \mathcal{E}e^{-\xi} - \sum_{m'\in\mathcal{M}_{\mathbb{X}}} e^{-y_{m'}}$, that for all m' in $\mathcal{M}_{\mathbb{X}}$ and for all ε positive
$$\sqrt{\mu(\mathbb{X})}\chi_{m'} \leq (1+\varepsilon)\sqrt{V_{m'}} + \sqrt{2\kappa M_{m'} y_{m'}} + \kappa(\varepsilon)\frac{B_{m'}}{\sqrt{\mu(\mathbb{X})}} y_{m'},$$
where
   - $V_{m'} = \int_{\mathbb{X}} \sum_{\lambda\in\mathcal{B}_{m'}} \varphi_\lambda^2 s d\mu/\mu(\mathbb{X})$;
   - $M_{m'} = \sup_{f\in S_{m'},\|f\|=1} \int_{\mathbb{X}} f^2 s d\mu/\mu(\mathbb{X}) \leq \|s\|\sqrt{\mathbb{D}_{m'}}$ by Cauchy-Schwarz;
   - $B_{m'} = \sup_{f\in S_{m'},\|f\|=1} \|f\|_\infty = \sqrt{\mathbb{D}_{m'}} \leq \sqrt{\mu(\mathbb{X})}$ by Cauchy-Schwarz, the definition of $\mathbb{D}_m$ and Assumption 2.

   NB: On the same set of probability we always have Equation (6.14).
   Using (5.8), we obtain that on the same set of probability
$$\sqrt{\mu(\mathbb{X})}\chi_{m'} \leq (1+\varepsilon)\sqrt{V_{m'}} + \varepsilon\sqrt{\mathbb{D}_{m'}\rho} + \left(\frac{\kappa\|s\|}{2\varepsilon\sqrt{\rho}} + \kappa(\varepsilon)\right) y_{m'}.$$
   We choose $y_{m'}$ as follows:
$$\forall m' \in \mathcal{M}_{\mathbb{X}},\ y_{m'} = \varepsilon\sqrt{\rho D_{m'}} \frac{1}{\left(\frac{\kappa\|s\|}{2\varepsilon\sqrt{\rho}} + \kappa(\varepsilon)\right)} + \xi.$$
   Let $\mathcal{G}$ be an upper bound of $\sum_{m'\in\mathcal{M}_{\mathbb{X}}} \exp\left(-\varepsilon\sqrt{\rho D_{m'}}\frac{1}{\left(\frac{\kappa\|s\|}{2\varepsilon\sqrt{\rho}} + \kappa(\varepsilon)\right)}\right)$. Since $\mathcal{M}_{\mathbb{X}}$ is a polynomial family, we have that $\mathcal{G}$ can have the same dependence on the parameters as $\mathcal{E}$. In particular, it does not depend any more on $\mathcal{M}_{\mathbb{X}}$ and $\mu(\mathbb{X})$.
4. For the first choice of penalty, we can remark that $V_{m'} \leq \rho\mathbb{D}_{m'}$. We also remark that the control of all the $\chi_{m'}$ implies in particular the control of $\chi_{\hat{m}}$ on the same

event. We can take the square and use inequality (5.8). Finally, we obtain with probability larger than $1-(\mathcal{E}+\mathcal{G})e^{-\xi}$

$$\chi_{\hat{m}}^2 \leq (1+3\varepsilon)^3 A_{\hat{m}} + (1+\frac{1}{3\varepsilon})\left(\frac{\kappa\|s\|}{2\varepsilon\sqrt{\rho}}+\kappa(\varepsilon)\right)^2 \frac{\xi^2}{\mu(\mathbb{X})}, \tag{6.15}$$

and Equation (6.14).

We can then rewrite Equation (6.14): for all $d>1$ (the choice of d fixes the choices of $\alpha, \beta, \delta, \eta, \theta, \gamma$), with probability larger than $1-(\mathcal{E}+\mathcal{G})e^{-\xi}$,

$$\begin{aligned} C(d)\|\tilde{s}-s\|^2 \leq C'(d)\|s_m-s\|^2 + dA_{\hat{m}} + C''(d)A_m \\ + \frac{g(\xi)}{\mu(\mathbb{X})} + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \end{aligned} \tag{6.16}$$

where C, C', C'' are continuous positive functions and where g depends on all the parameters except $\mathcal{M}_{\mathbb{X}}$ and $\mu(\mathbb{X})$ and is a polynomial with degree 2 of ξ. We can remark that Proposition 7 leads on a subset of the previous event with probability larger than $1-(\mathcal{E}+\mathcal{G}+1)e^{-\xi}$ to $(1+\varepsilon)\left(N_{\mathbb{X}}+\left(\frac{1}{2\varepsilon}+\frac{5}{6}\right)\xi\right) \geq \int_{\mathbb{X}} s d\mu$. Then we can upper bound $A_{\hat{m}}$ by $\left((1+\varepsilon)N_{\mathbb{X}}\mathbb{D}_{\hat{m}}/\mu(\mathbb{X})^2\right)+z(\varepsilon)\xi/\mu(\mathbb{X})$ for z continuous function. We do the same for A_m. Choosing correctly the parameters ($d(1+\varepsilon)=c$), all the terms with $\hat{m}$ in the second part of inequality (6.16) disappear. We obtain

$$B(c)\|\tilde{s}-s\|^2 \leq B'(c)\|s_m-s\|^2 + B''(c)\mathrm{pen}(m) + \frac{h(\xi)}{\mu(\mathbb{X})}, \tag{6.17}$$

where B, B', B'' are continuous positive functions and h depends on all the parameters except $\mathcal{M}_{\mathbb{X}}$ and $\mu(\mathbb{X})$ and is a polynomial with degree 2 of ξ. Here we obtain in fact a trajectorial inequality and it remains to integrate in ξ to obtain the first point.

5. For the third choice of penalty, it is sufficient to keep $V_{\hat{m}}$ instead of bounding it by $\mu(\mathbb{X})A_{\hat{m}}$. Then we have to replace it by some estimator. Using Proposition 7, we obtain that on a subset of the previous event, with probability larger than $1-(\mathcal{E}+\mathcal{G}+1)e^{-\xi}-\sum_{m'\in\mathcal{M}_{\mathbb{X}}} e^{-z_{m'}}$, where the $z_{m'}$'s will be chosen later, for all the m''s

$$\hat{V}_{m'} \geq V_{m'} - \sqrt{2z_{m'}G_{m'}} - \frac{H_{m'}z_{m'}}{3\mu(\mathbb{X})},$$

where

- $H_{m'} = \|\sum_{\lambda\in\mathcal{B}_{m'}} \varphi_\lambda^2\|_\infty = \mathbb{D}_{m'} \leq \mu(\mathbb{X})$,
- $G_{m'} = \int_{\mathbb{X}} \frac{\left(\sum_{\lambda\in\mathcal{B}_{m'}}\varphi_\lambda^2\right)^2}{\mu(\mathbb{X})^2} s d\mu \leq V_{m'}\frac{H_{m'}}{\mu(\mathbb{X})} \leq V_{m'}$.

Since $\hat{m}$ is one m', we deduce from this (using Assumption 3) that:

$$(1+\varepsilon)\left(\hat{V}_{\hat{m}} + (\frac{5}{6}+\frac{1}{2\varepsilon})z_{\hat{m}}\right) \geq V_{\hat{m}}.$$

We choose the $z_{m'}$'s as follows:

$$z_{m'} = \varepsilon\rho D_{m'} + \xi.$$

It makes appear as previously an other $\mathcal{H}$ which is independent of $\mathcal{M}_{\mathbb{X}}$ and $\mu(\mathbb{X})$ and which is an upper bound on $\sum_{m'\in\mathcal{M}_{\mathbb{X}}}\exp(-\varepsilon\rho D_{m'})$ since the family is polynomial. Then we have an inequality which looks like Equation (6.17) with probability larger than $1-(\mathcal{E}+\mathcal{G}+\mathcal{H}+1)e^{-\xi}$. It remains to integrate in ξ to obtain the third point.

6. For the second choice of penalty, we have to change $x_{m'}$, $y_{m'}$, $z_{m'}$ such that $\beta D_{m'}$ appears instead of $\rho D_{m'}$. Through the assumption, we can upper bound then $\beta\mathbb{D}_{m'}$ by $V_{m'}$ and then by $\hat{V}_{m'}$ on a large set of probability, up to some little constants. With these choices, $\mathcal{E}$, $\mathcal{G}$, $\mathcal{H}$ depend on β and it remains to integrate as previously. □

6.2.2. Proof of Theorem 2

Proof. Let m be an index in $\mathcal{M}_{\mathbb{X}}$, Inequality (5.10) means

$$\|\tilde{s}-s\|^2 \leq \|s-s_m\|^2 + 2\nu_{\mathbb{X}}(\tilde{s}-s_m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m).$$

Using twice Equation (5.8) and the triangle inequality, we get

$$\begin{aligned}&\forall\varepsilon>0 \text{ and } m\in\mathcal{M}_{\mathbb{X}},\\ &\qquad 2\nu_{\mathbb{X}}(\tilde{s}-s_m) \leq 2\|\tilde{s}-s_m\|\chi_{m,\hat{m}}\\ &\qquad\qquad \leq \frac{2}{\varepsilon}\|s-s_m\|^2 + \frac{2}{2+\varepsilon}\|\tilde{s}-s\|^2 + (1+\varepsilon)\chi^2_{m,\hat{m}}.\end{aligned}$$

Let ε be a fixed positive real.

$$\frac{\varepsilon}{2+\varepsilon}\|\tilde{s}-s\|^2 \leq (1+\frac{2}{\varepsilon})\|s-s_m\|^2 + (1+\varepsilon)\chi^2_{m,\hat{m}} - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m). \quad (6.18)$$

We apply the concentration inequality of Proposition 9 to $\chi_{m,m'}$ for all m' in $\mathcal{M}_{\mathbb{X}}$ (with M an upper bound for the variance term) in order to control $\chi_{m,\hat{m}}$. Furthermore, $\Omega(\varepsilon)\subset\Omega_{S_m+S_{m'}}(\varepsilon)$, using the notations of Proposition 9. Let $(x_{m'}, m\in\mathcal{M}_{\mathbb{X}})$ be a family of positive numbers which we will choose later. Then on $\Omega(\varepsilon)$, for all m' in $\mathcal{M}_{\mathbb{X}}$, we have with probability larger than $1-\sum_{m'\in\mathcal{M}_{\mathbb{X}}}e^{-x_{m'}}$

$$\sqrt{\mu(\mathbb{X})}\chi_{m,m'} \leq (1+\varepsilon)\left[\sqrt{\mathbb{E}(\chi^2_{m,m'})}+\sqrt{2\kappa M x_{m'}}\right].$$

Let ξ be a positive real. We set for all m' in $\mathcal{M}_{\mathbb{X}}$, $x_{m'}=L_{m'}D_{m'}+\xi$. Using Assumption 3 and 6, we get, on $\Omega(\varepsilon)$, with probability larger than $1-\Sigma_1 e^{-\xi}$,

$$\sqrt{\mu(\mathbb{X})}\chi_{m,\hat{m}} \leq (1+\varepsilon)\left[\sqrt{V(\hat{m})}+\sqrt{2\kappa M L_{\hat{m}}D_{\hat{m}}}+\sqrt{V(m)}+\sqrt{2\kappa M\xi}\right].$$

Taking the square and using Equation (5.8), we get, on $\Omega(\varepsilon)$, with probability larger than $1-\Sigma_1 e^{-\xi}$,

$$\begin{aligned}\chi^2_{m,\hat{m}} \leq{}& \frac{(1+\varepsilon)^3}{\mu(\mathbb{X})}\left[\sqrt{V(\hat{m})}+\sqrt{2\kappa M L_{\hat{m}}D_{\hat{m}}}\right]^2\\ &+(1+1/\varepsilon)(1+\varepsilon)^3\frac{V(m)}{\mu(\mathbb{X})}+(1+1/\varepsilon)^2(1+\varepsilon)^2\frac{2\kappa M\xi}{\mu(\mathbb{X})}.\end{aligned}$$

Using Assumption 4 and 5, we get, on $\Omega(\varepsilon)$, with probability larger than $1-(\Sigma_0+\Sigma_1)e^{-\xi}$

$$\begin{aligned}\chi^2_{m,\hat{m}} \leq\ & \frac{(1+\varepsilon)^3}{\mu(\mathbb{X})}\left[\sqrt{\hat{V}(\hat{m})}+\sqrt{2\kappa\hat{M}L_{\hat{m}}D_{\hat{m}}}+\sqrt{\eta\xi}\right]^2\\ & +(1+1/\varepsilon)(1+\varepsilon)^3\frac{V(m)}{\mu(\mathbb{X})}+(1+1/\varepsilon)^2(1+\varepsilon)^2\frac{2\kappa M\xi}{\mu(\mathbb{X})}\\ \leq\ & \frac{(1+\varepsilon)^4}{\mu(\mathbb{X})}\left[\sqrt{\hat{V}(\hat{m})}+\sqrt{2\kappa\hat{M}L_{\hat{m}}D_{\hat{m}}}\right]^2+(1+1/\varepsilon)(1+\varepsilon)^3\frac{V(m)}{\mu(\mathbb{X})}\\ & +\left[(1+1/\varepsilon)^2(1+\varepsilon)^2 2\kappa M+(1+1/\varepsilon)\eta\right]\frac{\xi}{\mu(\mathbb{X})}.\end{aligned}$$

Since on the same event $\eta\xi+\hat{V}(m)\geq V(m)$, if the penalty $\mathrm{pen}(\hat{m})$ is larger than $\frac{(1+\varepsilon)^5}{\mu(\mathbb{X})}\left[\sqrt{\hat{V}(\hat{m})}+\sqrt{2\kappa\hat{M}L_{\hat{m}}D_{\hat{m}}}+\sqrt{\eta\xi}\right]^2$, Equation (6.18) becomes: on $\Omega(\varepsilon)$, with probability larger than $1-(\Sigma_0+\Sigma_1)e^{-\xi}$

$$\frac{\varepsilon}{2+\varepsilon}\|\tilde{s}-s\|^2\leq(1+\frac{2}{\varepsilon})\|s-s_m\|^2+D(\varepsilon)\mathrm{pen}(m)+D'(M,\eta,\varepsilon)\frac{\xi}{\mu(\mathbb{X})},$$

where D and D' are continuous functions. If we integrate in ξ, we get

$$\begin{aligned}\mathbb{E}\left(\|\tilde{s}-s\|^2\mathbb{1}_{\Omega(\varepsilon)}\right)\leq\ & C(\varepsilon)\left[\|s-s_m\|^2+\mathbb{E}\left(\mathrm{pen}(m)\mathbb{1}_{\Omega(\varepsilon)}\right)\right]\\ & +C''(M,\eta,\varepsilon,\Sigma_0,\Sigma_1)\frac{1}{\mu(\mathbb{X})},\end{aligned}\qquad(6.19)$$

where C and C' are continuous functions.

It remains to control $\mathbb{E}\left(\|\tilde{s}-s\|^2\mathbb{1}_{\Omega(\varepsilon)^c}\right)$. We have that, using Assumption 1,

$$\begin{aligned}\|\tilde{s}-s\|^2 &= \|\tilde{s}-s_{\hat{m}}\|^2+\|s_{\hat{m}}-s\|^2\\ &\leq \chi^2_{\hat{m}}+\|s\|^2\\ &\leq \chi^2_{\Lambda}+\|s\|^2.\end{aligned}$$

By Cauchy-Schwarz, we have (using Assumption 2)

$$\mathbb{E}(\|\tilde{s}-s\|^2\mathbb{1}_{\Omega(\varepsilon)^c})\leq p\|s\|^2+\sqrt{p\mathbb{E}(\chi^4_\Lambda)},$$

where $p=\Delta/\mu(\mathbb{X})^2$. Now we use Proposition 9 to get an upper bound for $\mathbb{E}(\chi^4_\Lambda)$. As we have done previously, with probability larger than $1-e^{-\xi}$ we have (using Assumption 1)

$$\chi_\Lambda\leq(1+\varepsilon)\sqrt{\Phi\rho}+\frac{\sqrt{2\kappa M\xi}+\kappa(\varepsilon)\sqrt{\Phi}\xi}{\sqrt{\mu(\mathbb{X})}}.$$

We integrate this in ξ to obtain:

$$\mathbb{E}(\|\tilde{s}-s\|^2\mathbb{1}_{\Omega(\varepsilon)^c})\leq C^0(M,\varepsilon,\Phi,\Delta)/\mu(\mathbb{X})$$

where C^0 is a continuous positive function. This bound and the bound in (6.19) implies exactly the bound mentioned in Theorem 2. □

6.2.3. Proof of Proposition 1

Proof. This is an application of Theorem 2. Assumption 6 of Theorem 2 is Assumption 2 of this proposition. As we are in the subsets case, we have clearly:

$$\chi^2_{m,m'} \leq \chi^2_m + \chi^2_{m'}$$

for all m and m' in $\mathcal{M}_{\mathbb{X}}$. Then Assumption 3 is verified with $m \to V(m) = \mathbb{E}(\chi^2_m)$. Assumption 1 is a trivial consequence of Assumption 3 of this proposition and the localization property. (S_Λ in Proposition 1 has the role of S_Λ in Theorem 2.)

Let M (the one of Theorem 2) be $\sup_{f\in S_\Lambda}\int_{\mathbb{X}} f^2 s d\mu/\mu(\mathbb{X})$. Assumption 2 results of the following idea.

Let $\varepsilon > 0$. For all m and m' in $\mathcal{M}_{\mathbb{X}}$, $\|\sum_{\lambda\in m\cup m'} \nu_{\mathbb{X}}(\varphi_\lambda)\varphi_\lambda\|_\infty \leq (2\kappa M\varepsilon)/\kappa(\varepsilon)$ is implied by

$$\sup_{\lambda\in\Lambda} |\nu_{\mathbb{X}}(\varphi_\lambda)| \leq \frac{2\kappa M\varepsilon}{B\sqrt{|\Lambda|}\kappa(\varepsilon)},$$

thanks to the localization property. Hence we set

$$\Omega(\varepsilon) = \left\{\sup_{\lambda\in\Lambda} |\nu_{\mathbb{X}}(\varphi_\lambda)| \leq \frac{2\kappa M\varepsilon}{B\sqrt{|\Lambda|}\kappa(\varepsilon)}\right\}.$$

This event verifies

$$\mathbb{P}(\Omega(\varepsilon)^c) \leq \sum_{\lambda\in\Lambda} \mathbb{P}\left(|\nu_{\mathbb{X}}(\varphi_\lambda)| \geq \frac{2\kappa M\varepsilon}{B\sqrt{|\Lambda|}\kappa(\varepsilon)}\right).$$

We then use Proposition 7 to obtain

$$\begin{aligned}\mathbb{P}\left[|\nu_{\mathbb{X}}(\varphi_\lambda)| \geq u\right] &\leq 2\exp\left(\frac{-\mu(\mathbb{X})u^2}{2\frac{\int_{\mathbb{X}}\varphi_\lambda^2(x)s(x)d\mu_x}{\mu(\mathbb{X})} + \frac{2}{3}\|\varphi_\lambda\|_\infty u}\right)\\ &\leq 2\exp\left(\frac{-\mu(\mathbb{X})u^2}{2M + \frac{2}{3}B\sqrt{|\Lambda|}u}\right)\\ &\leq 2\exp\left(-\eta(\varepsilon)\frac{\mu(\mathbb{X})M}{B^2|\Lambda|}\right)\end{aligned}$$

for η a positive continuous function with $u = (2\kappa M\varepsilon/(B\sqrt{|\Lambda|}\kappa(\varepsilon)))$. Assumption 3 of the proposition then implies Assumption 2 of the theorem.

Now we have different choices to valid Assumption 4 of the theorem:

- For case (a) and (b) of the proposition, $M \leq \|s\|_\infty$ and we assume that an upper bound of $\|s\|_\infty$ is known. Consequently $\hat{M} = M'$ works.
- For case (c), the assumption made in the proposition implies on $\Omega(\varepsilon)$

$$M \leq \frac{\kappa(\varepsilon)}{\kappa(\varepsilon) - 2\kappa\varepsilon}(\|\hat{s}_\Lambda\|_\infty + K') = \hat{M}$$

for $\varepsilon \leq 1.6$ which implies $\kappa(\varepsilon) - 2\kappa\varepsilon > 0$. Furthermore, for $1, 6 > \varepsilon > 0$,

$$\frac{\kappa(\varepsilon)}{\kappa(\varepsilon) - 2\kappa\varepsilon} > 1.$$

For assumption 4 of the theorem, it depends on the choice of penalty.

- For case (a), all is determinism, we bound $V(m)$ by $M'|m|/\mu(\mathbb{X})$, with $\eta = 0$.
- For case (b), the estimator of $V(m)$ is $\hat{V}_m$. We want to use Proposition 7. Let (x'_m) be a family of positive numbers which we will choose later. With probability larger than $1 - \sum_{m' \in \mathcal{M}_{\mathbb{X}}} e^{-x'_m}$, we have for all the m''s
$$\hat{V}_{m'} \geq V(m') - \sqrt{2x_{m'}G_{m'}} - \frac{H_{m'}x_{m'}}{3\mu(\mathbb{X})},$$
where
  - $H_{m'} = \| \sum_{\lambda \in m'} \varphi_\lambda^2 \|_\infty \leq B^2 |\Lambda| \leq \theta(\varepsilon)\mu(\mathbb{X})$, by Cauchy-Schwarz, the assumption in (b) and the localization property,
  - $G_{m'} = \int_{\mathbb{X}} \frac{(\sum_{\lambda \in m'} \varphi_\lambda^2)^2}{\mu(\mathbb{X})^2} s d\mu \leq V(m') \frac{H_{m'}}{\mu(\mathbb{X})}$.

  We deduce from this last fact that:
$$\sqrt{V_{\hat{m}}} \leq (\sqrt{1+\varepsilon} - 1)\sqrt{2\kappa M' x_{\hat{m}}} + \sqrt{\hat{V}_{\hat{m}}}.$$
  We choose the $x_{m'}$'s as follows:
$$x_{m'} = L_{m'} D_{m'} + \xi.$$
  It remains to take the square and (5.8) to obtain Assumption 4 with $\Sigma_0 = \Sigma$.

Then all the conclusions are consequences of Theorem 2.

NB: The result of Theorem 2 is true for all penalty larger than
$$\frac{(1+\varepsilon)^5}{\mu(\mathbb{X})} \frac{\kappa(\varepsilon)}{\kappa(\varepsilon) - 2\kappa\varepsilon} (\sqrt{\hat{V}(m)} + \sqrt{2\kappa \hat{M} L_m D_m},)$$
which for ε small enough, is less than the penalties of the proposition. Furthermore on $\Omega(\varepsilon)$, we have
$$\|\hat{s}_\Lambda\|_\infty \leq \left(1 + \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}\right)(\|s_\Lambda\|_\infty + K'),$$
which implies that $\mathbb{E}(\mathrm{pen}(m)\mathbb{1}_{\Omega(\varepsilon)}) \leq B(\varepsilon)\frac{(\|s_\Lambda\|_\infty + K')|m|}{\mu(\mathbb{X})}(1 + L_m)$ for B positive continuous function. □

6.2.4. Proof of Proposition 2

Proof. Here we are still going to apply Theorem 2. Assumption 6 of Theorem 2 is assumption 1 of the proposition. The orthonormal basis of S_m is $\{\mathbb{1}_I \sqrt{(\mu(\mathbb{X})/\mu(I))}, I \in m\}$. We remark that for all m based on the points of Γ
$$M_m = \sup_{\sum_{I \in m} a_I^2 = 1} \int_{\mathbb{X}} \left(\sum_{I \in m} a_I \mathbb{1}_I \sqrt{\frac{\mu(\mathbb{X})}{\mu(I)}} \right)^2 s \frac{d\mu}{\mu(\mathbb{X})} \leq \sup_{I \in m} \left(\frac{\int_I s d\mu}{\mu(I)} \right).$$

Since Γ is a regular partition,

$$M = \sup_{I \in \Gamma} \int_I s d\mu / \mu(I)$$

provides a bound on the variance. Indeed, for all J in m of $\mathcal{M}_{\mathbb{X}}$, there exists I_1, ..., I_k in Γ such that $J = \cup_{i=1}^k I_i$ and $\mu(I_i) = \mu(\mathbb{X})/D_\Gamma$ for all i.

To get Assumption 3 of Theorem 2, we set

$$m \to V(m) = \frac{M D_m}{\mu(X)}.$$

Indeed $S_m + S_{m'} \subset S_{m \cup m'}$ for all m and m', and $m \cup m'$, the partition constructed with the union of the points of m and m', is a partition based on some points of Γ and $D_{m \cup m'} \leq D_m + D_{m'}$.

The space S_Λ in Theorem 2 is clearly S_Γ with basis $\{\mathbb{1}_I \sqrt{D_\Gamma}, I \in \Gamma\}$. Consequently, Assumption 1 is a consequence of Assumption 2 of Proposition 2.

Assumption 2 results of the following idea. Let $\varepsilon > 0$. For all m and m' in $\mathcal{M}_{\mathbb{X}}$

$$\|s_{m,m'} - \hat{s}_{m,m'}\|_\infty = \| \sum_{I \in m \cup m'} \frac{N_I - \int_I s d\mu}{\mu(I)} \mathbb{1}_I \|_\infty \leq \frac{2\kappa M \varepsilon}{\kappa(\varepsilon)},$$

is implied by

$$|N_I - \int_I s d\mu| \leq \frac{\mu(\mathbb{X}) 2\kappa M \varepsilon}{D_\Gamma \kappa(\varepsilon)},$$

by the same reasoning as $M_m \leq M$. Hence we set

$$\Omega(\varepsilon) = \left\{ \sup_{I \in \Gamma} |N_I - \int_I s d\mu| \leq \frac{\mu(\mathbb{X}) 2\kappa M \varepsilon}{D_\Gamma \kappa(\varepsilon)} \right\}.$$

Then

$$\mathbb{P}(\Omega(\varepsilon)^c) \leq \sum_{I \in \Gamma} \mathbb{P}\left(|N_I - \int_I s d\mu| \geq \frac{\mu(\mathbb{X}) 2\kappa M \varepsilon}{D_\Gamma \kappa(\varepsilon)} \right).$$

We then use Proposition 7 to obtain

$$\begin{aligned} \mathbb{P}\left[|N_I - \int_I s d\mu| \geq \frac{\mu(\mathbb{X}) 2\kappa M \varepsilon}{D_\Gamma \kappa(\varepsilon)} \right] &\leq 2 \exp\left(- \frac{\left(\frac{\mu(\mathbb{X}) 2\kappa M \varepsilon}{D_\Gamma \kappa(\varepsilon)} \right)^2}{2 \int_I s(x) d\mu_x + \frac{2}{3} \frac{\mu(\mathbb{X}) 2\kappa M \varepsilon}{D_\Gamma \kappa(\varepsilon)}} \right) \\ &\leq 2 \exp\left(-\eta(\varepsilon) \frac{M \mu(X)}{D_\Gamma} \right), \end{aligned}$$

for η a positive continuous function. Assumption 2 of the proposition implies then Assumption 2 of the theorem.

On $\Omega(\varepsilon)$, we have $\hat{M} = \sup_{I \in \Gamma}(N_I/\mu(I))$ which verifies

$$M \leq \frac{\kappa(\varepsilon)}{\kappa(\varepsilon) - 2\kappa\varepsilon} \sup_{I \in \Gamma}(N_I/\mu(I)) = \hat{M}.$$

We have in fact $M \leq d\hat{M}$, with $d > 1$ for d (depending on ε) as close as we want to 1. Then Assumptions 4 and 5 are obvious, with $\Sigma_0 = 0$ and $\eta = 0$. Applying Theorem 2, we get exactly the conclusion, making the same remark between c and ε as in the proof of Proposition 1, remarking that on $\Omega(\varepsilon)$, $\hat{M} \leq \left(1 + \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}\right) M$. □

6.3. Proofs of the minimax results

In order to compute lower bound on the minimax risk, we use some recent interesting result [6] due to L. Birgé, which is a new version of Fano's Lemma and turns out to be easier to use than Fano's lemma.

Lemma 4. *Let $\{\mathbb{P}_i, i \in \{0, ..., n\}\}$ be a finite family of probability defined on the same measurable space $(\Omega, \mathcal{X})$. One sets*

$$\bar{K} = \frac{1}{n} \sum_{i=1}^{n} K(\mathbb{P}_i, \mathbb{P}_0)$$

where K is the Kullback-Leibler information between $\mathbb{P}_i$ and $\mathbb{P}_0$.
There exists an absolute constant α ($\alpha = 0.71$ works) such that if $\hat{\theta}$ is a random variable on Ω with values in $\{0, ..., n\}$, one has

$$\inf_{0 \leq i \leq n} \mathbb{P}_i(\hat{\theta} = i) \leq \alpha \vee \frac{\bar{K}}{\log(n+1)}.$$

Remark: $K(P, Q) = \mathbb{E}_P(\log(\frac{dP}{dQ}))$ if P and Q are two probability measures absolutely continuous with respect to each other.

We see through this lemma the importance of the Kullback-Leibler information. We can compute this information for Poisson processes:

Lemma 5. *Let N and N' be two Poisson processes on $\mathbb{X}$ with respectively intensity s and t. They define probabilities P (respectively Q) on the set of all countable sets of points of $\mathbb{X}$.*
Then

$$K(P, Q) = \int_{\mathbb{X}} s(x)\phi\left(\log(\frac{t}{s})\right)(x)d\mu_x$$

where $\phi(u) = \exp(u) - u - 1$.

A proof of this lemma can be found in [14].
Now, we have to compute lower bounds for minimax risk on some proper $\mathcal{S}$.

6.3.1. Proof of Proposition 3

Proof. Let us recall a combinatorial lemma, due to Gallager in information theory framework [20, Exercise 5.8, p 531 and Exercise 5.19, p 537]. A simpler proof can be found in [5, Lemma 8, p 400] which is made in the equivalent framework of algebra of sets.

Lemma 6. *Let Γ be a finite set with cardinal K. The maximal set $\mathcal{M}_\Gamma$, included in $\mathcal{P}(\Gamma)$, such that for all m, m' of $\mathcal{M}_\Gamma$, $|m \triangle m'| \geq \theta K$ verifies*

$$\log |\mathcal{M}_\Gamma| \geq \sigma K,$$

for θ and σ absolute constants.

Here we set $\Gamma = \{L, ..., D\}(K = D - L + 1)$. Let

$$\mathcal{C}_D = \left\{ t_m = \rho + a_D \sum_{\lambda \in m} \varphi_\lambda \ , \ m \in \mathcal{M}_\Gamma \right\},$$

with

$$a_D = \frac{\rho}{2B\sqrt{D}} \wedge \frac{c_D}{\sqrt{D}}.$$

This set is a subset of $\mathcal{E}(c, \rho)$ and even the t_m are bounded from below by $\rho/2$. Hence we have

$$R(\mathcal{E}(c, \rho)) \geq R(\mathcal{C}_D).$$

For all $\hat{s}$ in $\mathbb{L}^2 \cap \mathcal{F}(N)$, estimator of s, we associate $\hat{s}' = \mathrm{argmin}_{s \in \mathcal{C}_D} \|s - \hat{s}\|$. Thus we have $\|\hat{s}' - s\| \leq \|\hat{s}' - \hat{s}\| + \|\hat{s} - s\| \leq 2\|\hat{s} - s\|$. Then

$$R(\mathcal{E}(c, \rho)) \geq \frac{1}{4} \inf_{\hat{s} \in \mathcal{C}_D \cap \mathcal{F}(N)} \sup_{s \in \mathcal{C}_D} \mathbb{E}(\|s - \hat{s}\|^2).$$

Since, by Lemma 6, for all m and m' of $\mathcal{C}_D$

$$\|t_m - t_{m'}\|^2 \geq \theta(D - L + 1)a_D^2,$$

we have the following lower bound

$$\begin{aligned} R(\mathcal{E}(c, \rho)) &\geq \frac{\theta(D - L + 1)a_D^2}{4} \inf_{\hat{s} \in \mathcal{C}_D} \sup_{s \in \mathcal{C}_D} \mathbb{P}_s(\hat{s} \neq s) \\ &\geq \frac{\theta(D - L + 1)a_D^2}{4} \inf_{\hat{s} \in \mathcal{C}_D} (1 - \inf_{s \in \mathcal{C}_D} \mathbb{P}_s(\hat{s} = s)). \qquad (6.20) \end{aligned}$$

Now, we are going to use Lemma 4. Hence, we have to compute $\bar{K}$ and to do so, we use Lemma 5.

$$\begin{aligned} \forall m' \neq m \in \mathcal{M}_D, \quad K(\mathbb{P}_{t_{m'}}, \mathbb{P}_{t_m}) &= \int t_{m'} \phi(\log \frac{t_m}{t_{m'}}) d\mu_x \\ &= \int [t_m - t_{m'} - t_{m'} \log(1 + \frac{t_m - t_{m'}}{t_{m'}})] d\mu_x \\ &\leq \int \frac{(t_m - t_{m'})^2}{t_m}(x) d\mu_x \\ &\leq \frac{2}{\rho} \mu(\mathbb{X}) \|t_{m'} - t_m\|^2 \qquad (6.21) \\ &\leq \frac{2\mu(\mathbb{X})(D - L + 1)a_D^2}{\rho}, \end{aligned}$$

since $\forall x > -1, \log(1+x) \geq x/(1+x)$. Thus

$$\bar{K} \leq \frac{2\mu(\mathbb{X})(D-L+1)a_D^2}{\rho}. \tag{6.22}$$

Lemma 4 applied to the family $\{\mathbb{P}_s, s \in \mathcal{C}_D\}$ leads to

$$R(\mathcal{E}(c,\rho)) \geq \frac{(1-\alpha)\theta}{4}(D-L+1)a_D^2$$

if D is such that

$$\frac{2\mu(\mathbb{X})(D-L+1)a_D^2}{\rho} \leq \alpha\sigma(D-L+1)$$

by Lemma 6. The result follows with $\zeta = \alpha\sigma/2$ and $\eta = (1-\alpha)\theta/4$. □

6.3.2. Proof of Proposition 4

Proof. We use a lemma which is due to L. Birgé and P. Massart [9]. Their proof being rather intricate, we present a complete and simple proof in the appendix (although our constants are slightly worse than theirs). We deduce from this lemma (Lemma 8) that the maximal set $\mathcal{M}_{n,D}$, included in the set of all the parts, $\mathcal{P}(\{L,...,n\})$, such that for all m, m' of $\mathcal{M}_{n,D}$, $|m| = D$ and $|m \bigtriangleup m'| \geq \theta' D$, verifies

$$\log|\mathcal{M}_{n,D}| \geq \sigma' D \log\frac{n-L+1}{D},$$

for θ' and σ' constants. We set

$$\mathcal{C}_{n,D} = \left\{ t_m = \rho + a_{n,D}\sum_{\lambda\in m}\varphi_\lambda, m \in \mathcal{M}_{n,D} \right\}.$$

We will choose $a_{n,D}$ later. We have again the following condition to get $\mathcal{C}_{n,D} \subset B_{n,D,\rho}$:

$$a_{n,D} \leq \frac{\rho}{2B\sqrt{n}},$$

which implies that for all m in $\mathcal{C}_{n,D}$, $t_m \geq \rho/2$.
Note that $\log|\mathcal{C}_{n,D}| \geq \sigma' D \log\frac{n-L+1}{D}$ and for all $t_m \neq t_{m'} \in \mathcal{C}_{n,D}$, we have $\|t_m - t_{m'}\|^2 \geq \theta' D a_{n,D}^2$. So we have as in the previous proof (see Equation (6.20)):

$$R(B_{n,D,\rho}) \geq \frac{\theta'}{4} D a_{n,D}^2 \inf_{\hat{s}\in\mathcal{C}_{n,D}}\left(1 - \inf_{s\in\mathcal{C}_{n,D}}\mathbb{P}_s(\hat{s}=s)\right).$$

Using Lemma 4 and the control of the Kullback-Leibler information (6.21), we obtain that if

$$\frac{\bar{K}}{\log|\mathcal{C}_{n,D}|} \leq \alpha,$$

which is implied by

$$\frac{4\mu(\mathbb{X})Da_{n,D}^2}{\sigma'\rho D\log\frac{n-L+1}{D}} \leq \alpha,$$

then

$$R(B_{n,D,\rho}) \geq \frac{\theta'(1-\alpha)}{4}Da_{n,D}^2.$$

Choosing

$$a_{n,D}^2 = \frac{\alpha\sigma'\rho\log\frac{n-L+1}{D}}{4\mu(\mathbb{X})} \wedge \frac{\rho^2}{4B^2n},$$

leads to the result with $\zeta = \alpha\sigma'/4$ and $\eta = \theta'(1-\alpha)/4$. □

6.3.3. Proof of Proposition 5

Proof. We consider the basis with regularity $r > \alpha$ previously described. We want to apply Proposition 3. We use the wavelet basis defined in (3.4). The L of the proposition is here $2^l + 1$, a fixed number (depending on r and then on α). This is, when we arrange the indices by lexicographic order, exactly an ellipsoid $\mathcal{E}(c,\rho)$ with $c_{j,k} = R2^{-j\alpha}$. This sequence is piecewise constant non increasing in the lexicographic order. For all J positive, we look at $\mathcal{F}_J$ (see Equation (3.3)). The localized property is true with constant B. The cardinal of the family is equal to $2^{J+1} - 2^l$ which is larger than 2^J. Hence, Proposition 3 leads to

$$R(\mathcal{B}(\rho, R, B_{2,2}^\alpha)) \geq \eta\frac{2^{J+1}-2^{l+1}}{2^{J+1}-2^l}\left(\frac{\rho^2}{4B^2} \wedge R^2 2^{-2J\alpha}\right),$$

when

$$\frac{R^2 2^{-2J\alpha}}{2^J} \leq \zeta\frac{\rho}{T}.$$

We take $J \geq l+1$ as small as possible such that

$$2^J \geq \left(\frac{R^2T}{\zeta\rho}\right)^{\frac{1}{2\alpha+1}},$$

and we obtain the result remarking that $\frac{2^{J+1}-2^{l+1}}{2^{J+1}-2^l} \geq \frac{2}{3}$. □

A. Combinatorial lemmas

Lemma 7. *There exists a binomial variable $\mathcal{B}(D,\theta)$, N_b^*, and an hyper-geometric variable $\mathcal{H}(N,D,\theta)$, N_b, such that*

$$E(N_b^*|N_b) = N_b.$$

The lemma is proved by D.J. Aldous [1].

Lemma 8. *Let N and D be positive integers such that $N \geq AD$. Let $\mathcal{E}_{N,D}$ be the subset of* $\{0,1\}^N$ *whose elements have a number D of* 1. *We consider the distance on $\mathcal{E}_{N,D}$:*

$$\forall x, y \in \mathcal{E}_{N,D}, d(x,y) = |\{i/y_i = 1, x_i = 0\}|.$$

Then the maximal subset $\mathcal{M}_{N,D}$ such that all its elements are at distance θD, has a cardinal larger than $\exp(\sigma D \log(N/D))$ *with for instance $A = 4$, $\theta = 1/4$ and* $\sigma = 0.233$.

We recall that $|m|$ denotes the cardinality of the set m.

Proof. $\mathcal{E}_{N,D}$ is covered by the balls of radius θD and center in $\mathcal{M}_{N,D}$. We deduce from this the following inequality:

$$\binom{N}{D} \leq \sum_{x \in \mathcal{M}_{N,D}} |B(x, \theta D)|.$$

Let us look at $B(x, \theta D)$, which is the set $\{y/|\{i/y_i = 1, x_i = 1\}| \geq D - \theta D\}$. The number $N_b = |\{i/y_i = 1, x_i = 1\}|$ for x and y equally likely chosen in $\mathcal{E}_{N,D}$, is an hyper-geometric variable: if we take D balls in an urn which contains D blue balls and $N - D$ red balls, without replacement, N_b is the number of blue balls in our draw. We deduce from this comparison:

$$1 \leq |\mathcal{M}_{N,D}| \mathbb{P}(N_b \geq D - \theta D).$$

In order to understand this probability, we can apply Lemma 7: a draw without replacement is more concentrated (for convex functions) than a draw with replacement (which is here a binomial variable, $N_b^* \sim \mathcal{B}(D, D/N)$). This leads to, for all $\lambda > 0$:

$$\begin{aligned} 1 &\leq |\mathcal{M}_{N,D}| \exp(-\lambda(D - \theta D)) E(\exp(\lambda N_b)) \\ &\leq |\mathcal{M}_{N,D}| \exp(-\lambda(D - \theta D)) E(\exp(\lambda N_b^*)). \end{aligned} \quad \text{(A.1)}$$

Following the proof of Bennett's inequality ([7]), we obtain, maximizing (A.1) in λ:

$$1 \leq |\mathcal{M}_{N,D}| \exp\left(-\frac{D^2}{N} h\left(\frac{D - \theta D - D^2/N}{D^2/N}\right)\right) \quad \text{(A.2)}$$

with $\theta < 1/2$ and $\forall u > 0$, $h(u) = (1+u)\ln(1+u) - u$.
The condition on θ ensures that the deviation is greater than the expectation. We therefore deduce that:

$$\begin{aligned} |\mathcal{M}_{N,D}| &\geq \exp\left(\frac{D^2}{N}\left[\frac{D-\theta D}{D^2/N} \ln \frac{D-\theta D}{D^2/N} - \frac{D-\theta D}{D^2/N} + 1\right]\right) \\ &\geq \exp\left(\sigma D \ln \frac{N}{D}\right). \end{aligned} \quad \text{(A.3)}$$

Example : if we take $A = 4$ and $\theta = 1/4$ then $\sigma = 0.233$ works. □

References

[1] Aldous, D.J.: Exchangeability and related topics. In: Lect. Notes Math. **1117**, 1–198 (1985)

[2] Anscombe, F.J.: The transformation of Poisson, binomial and negative-binomial data. Biometrika, Cambridge **35**, 246–254 (1948)

[3] Baraud, Y.: Model selection for regression on a fixed design. Probab. Theory Relat. Fields (2000)

[4] Barron, A.R., Sheu, C.-H.: Approximation of density functions by sequences of exponential families. Ann. Statist. **19**, 1347–1369 (1991)

[5] Barron, L., Birgé, A., Massart, P.: Risk bounds for model selection via penalization. P.T.R.F. (1999)

[6] Birgé, L.: A new look at an old result : Fano's Lemma. Prépublication 632, Universités de Paris VI et Paris VII, (2001)

[7] Birgé, L., Massart, P.: Model selection from a nonasymptotic view point. Book in preparation

[8] Birgé, L., Massart, P.: From model selection to adaptive estimation. In: Festschrift for Lucien Le Cam, 55–87. Springer, New York, (1997)

[9] Birgé, L., Massart, P.: Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. Bernoulli **4**(3), 329–375 (1998)

[10] Birgé, L., Massart, P.: Gaussian model selection. J. European Math. Soc. (2001)

[11] Bobkov, S.G., Ledoux, M.: On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. J. Funct. Anal. **156**(2), 347–365 (1998)

[12] Brooks, M.M., Marron, J.S.: Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions. Stochastic Process. Appl. **38**(1), 157–165 (1991)

[13] Castellan, G.: Modified akaike's criterion for histogram density estimation. Technical report, Univ. Paris-Sud. No 99.61 (1999)

[14] Cavalier, L., Koo, J.-Y.: Poisson intensity estimation for tomographic data using a wavelet shrinkage approach. September 2000, manuscript.

[15] Cirel'son, B.S., Ibragimov, I.A., Sudakov, V.N.: Norms of gaussian sample functions. Proc. 3rd Japan-USSR Symp. Probab. Theory, Taschkent 1975, Lect. Notes Math. **550**, 20–41 (1976)

[16] Cohen, L., Daubechies, I., Vial, P.: Wavelets on the interval and fast wavelet transforms. Appl. Comput. Harmon. **1**, 54–81 (1993)

[17] DeVore, R., Lorentz, G.: Constructive approximation. Springer-Verlag, 1993

[18] Donoho, D.L.: Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In: Proc. Symp. Appl. Math. **47**, 173–205 (1993)

[19] Donoho, D.L., Johnstone, I.M.: Minimax estimation via wavelet shrinkage. Ann. Stat. **26**(3), 879–921 (1998)

[20] Gallager, R.G.: Information theory and reliable communication. New York-London-Sydney-Toronto: John Wiley and Sons, Inc. XVI, 1968

[21] Houdré, C.: Remarks on deviation inequalities for functions of infinitely divisible random vectors. Ann. Prob. (2002)

[22] Houdré, C., Privault, N.: Concentration and deviation inequalities in infinite dimensions via covariance representations. To appear in Bernoulli (2002)

[23] Kerkyacherian, G., Picard, D.: Estimation de densité par méthode de noyaux et d'ondelettes : les liens entre la géométrie du noyau et les contraintes de régularité. Comptes rendus de l'Académie des Sciences Ser. I Math **315**, 79–84 (1992)

[24] Kim, W.-C., Koo, J.-Y.: Inhomogeneous Poisson intensity via information projections onto wavelets subspaces. May 9, 2000, manuscript.

[25] Kingman, J.F.C.: Poisson processes. Oxford Studies in Probability 1993

[26] Kolaczyk, E.D.: Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. Stat. Sin. **9**(1), 119–135 (1999).

[27] Kutoyants, Yu.A.: Statistical inference for spatial Poisson processes. Lecture Notes in Statistics, **134**. Springer edition, 1998.

[28] Ledoux. M.: On Talagrand deviation inequalities for product measures. In: ESAIM:Probability and statistics 1, (1996)

[29] Ledoux, M., Talagrand, M.: Probability in Banach spaces. Springer-Verlag, Berlin, 1991. Isoperimetry and processes

[30] Mallows, C.L.: Some comments on C_p. Technometrics **15**, 661–675 (1973)

[31] Massart, P.: Some exponential bounds for the khi-square statistics with applications. To appear.

[32] Massart, P.: About the constants in Talagrand's concentration inequalities for empirical processes. Ann. Proba. (2000)

[33] Massart, P.: Some applications of concentration inequalities. Ann. de Toulouse (2000)

[34] Reboul, L.: Estimation sous restriction de forme et application à la fiabilité. Test de validation d'un modèle paramétrique pour un processus de Poisson non homogène. PhD thesis, U.P.S. (1998)

[35] Rio, E.: Inégalités exponentielles pour les processus empiriques. C.R.A.S. t.**330**(Série I): 597–600 (2000)

[36] Rio, E.: Une inégalité de Bennett pour les maxima de processus empiriques. Technical report, Université de Versailles-St Quentin en Yvelynes (2001)

[37] Rudemo, M.: Empirical choice of histograms and kernel density estimators. Scand. J. Stat. Theory Appl. **9**, 65–78 (1982)

[38] Talagrand, M.: New concentration inequalities in product spaces. Invent. Math. **126**(3), 505–563 (1996)

[39] Wu, L.: A new modified logarithmic Sobolev inequality for Poisson point process and several applications. Probability Theory and Related Fields (2000)