Estimation adaptative de l'intensité de certains processus ponctuels par sélection de modèle.

Patricia Reynaud-Bouret

ORSAY
N° D'ORDRE : 6917

UNIVERSITÉ PARIS XI

UFR SCIENTIFIQUE D'ORSAY

THÈSE

Présentée

Pour obtenir

Le GRADE de DOCTEUR EN SCIENCES

DE L'UNIVERSITÉ PARIS XI ORSAY

Spécialité : Mathématiques

PAR

Patricia Reynaud-Bouret

Sujet : **ESTIMATION ADAPTATIVE DE L'INTENSITÉ DE CERTAINS PROCESSUS PONCTUELS PAR SÉLECTION DE MODÈLE.**

Rapporteurs : Mme Sara Van De GEER
M. Christian HOUDRÉ

Soutenue le **27 Juin 2002** devant la Commission d'examen :

M. Lucien BIRGÉ	Président
M. Pierre BRÉMAUD	Examinateur
M. Jean BRETAGNOLLE	Examinateur
M. Gérard GRÉGOIRE	Examinateur
M. Christian HOUDRÉ	Rapporteur
M. Pascal MASSART	Directeur de Thèse

Table des matières

Introduction 5
0.1. Quelques processus ponctuels 5
0.2. Un cas simple : estimateur en histogramme pour les processus de Poisson temporels 8
0.3. Les autres familles de modèles et les autres contrastes. 14
0.4. Inégalités de concentration 18
0.5. Sélection de modèle 22
0.6. Risque minimax et adaptation 25
Plan 26

Chapitre 1. Concentration inequalities for inhomogeneous Poisson processes and adaptive estimation of the intensity 29
1.1. Introduction 30
1.2. Concentration inequalities for Poisson processes 36
1.3. Model selection with projection estimators 42
1.4. Some lower bounds for the minimax risk 51
1.5. Comparison between the risk of p.p.e. and the minimax risk 58
1.6. Proofs 62

Chapitre 2. Exponential inequalities and martingales 79
2.1. Introduction 80
2.2. Exponential inequalities for degenerate U-statistic of order 2 with constants 83
2.3. Exponential inequalities for counting processes 92

Chapitre 3. Penalized projection estimators for the Aalen's multiplicative intensity 103
3.1. Introduction 104
3.2. Histogram quasi-least square estimators 107
3.3. Predictable models 118
3.4. Simulations 124

Annexe A. Combinatorial lemmas 139

Annexe B. Programmes Scilab 141

B.1. Simulations des variables aléatoires 141
B.2. Stratégie en histogrammes 151
B.3. Stratégie Fourier 160
B.4. Comparaison 167

Annexe. Bibliographie 181

Introduction

Le sujet de cette thèse est d'essayer d'adapter des techniques de sélection de modèle à un cadre particulier : celui des processus ponctuels. Plus précisément, on veut montrer que, pour certains processus ponctuels, les estimateurs par projection pénalisés sont adaptatifs, soit parmi une famille d'estimateurs par projection, soit pour le risque minimax. En fait, on s'est restreint à deux cas de processus ponctuels : les processus de Poisson inhomogènes et les processus de comptage à intensité multiplicative d'Aalen. Dans cette introduction, nous allons tout d'abord présenter ces deux types de processus. Puis nous exposerons la méthode dans un cadre très simple. Ensuite nous généraliserons et présenterons nos résultats principaux.

0.1. Quelques processus ponctuels

Un processus ponctuel est un ensemble aléatoire au plus dénombrable de points d'un espace $\mathbb{X}$.

0.1.1. Processus de Poisson. Les processus de Poisson sont les processus ponctuels les plus simples à étudier.

DEFINITION 1. *Soit $(\mathbb{X}, \mathcal{X})$ un espace mesurable. Soit N un processus ponctuel. On dit que N est un processus de Poisson sur $(\mathbb{X}, \mathcal{X})$ si et seulement si*

- *pour tout A de $\mathcal{X}$, le nombre de points de N apparus dans A est une variable aléatoire N_A qui suit une loi de Poisson de paramètre $\nu(A)$.*
- *pour toute famille finie d'ensemble disjoints $A_1, ..., A_n$ de $\mathcal{X}$, $N_{A_1}, ..., N_{A_n}$ sont des variables indépendantes.*

La fonction ν sur $\mathcal{X}$ définie ci-dessus est en réalité une mesure, appelée mesure moyenne du processus. Si elle est absolument continue par rapport à une mesure de référence connue μ, la dérivée de Radon-Nykodym de ν est appelée l'intensité du processus de Poisson par rapport à μ. L'intensité du processus sera notée s.

Le processus est considéré comme homogène si cette intensité est constante, inhomogène sinon.

L'étude des processus de Poisson est très florissante : les livres **[40]** et **[43]** en donnent une bonne vision d'ensemble. En effet, les processus de Poisson peuvent modéliser une grande quantité de situations. Les processus de Poisson temporels (i.e. définis sur $\mathbb{R}^+$)

peuvent modéliser les instants de pannes dans la vie d'une machine (les pannes étant réparées immédiatement après), ou les instants de réception de rayons Gamma par la Terre [**41**]. Les processus spatiaux (i.e. définis sur un ouvert de $\mathbb{R}^2$ ou $\mathbb{R}^3$) peuvent quant à eux modéliser les lieux d'émission d'appels téléphoniques dans une ville, les lieux d'émission de protons dans le cerveau au cours d'une radiographie RMN, et plein d'autres encore.

Dans chacun de ces cas, la fréquence locale des instants ou la "densité" locale de points est représentée par l'intensité du processus de Poisson qui caractérise complètement sa loi. C'est cette fonction que nous cherchons à estimer à partir de l'observation de N (ce qui sera fait dans le Chapitre 1).

0.1.2. Processus de comptage temporels. Un processus de comptage est une fonction $(N_t, t \geq 0)$ en escalier, aléatoire croissante issue de 0 et de sauts égaux à 1. Ils sont "de comptage" car généralement ils représentent, en fonction du temps t, un nombre d'événements ayant eu lieu jusqu'à t. Un processus de comptage peut être facilement associé (et donc identifié) au processus ponctuel temporel représentant l'ensemble de ses instants de sauts.

Un processus de Poisson temporel peut donc être vu comme un processus de comptage où à t fixé, N_t est le nombre de points du processus de Poisson tombés avant t, c'est-à-dire avec les notations du paragraphe précédent $N_{[0,t]}$.

De manière générale, la fonction aléatoire en escalier $(N_t, t \geq 0)$ engendre de manière classique une filtration $(\mathcal{F}_t, t \geq 0)$. Comme la fonction est croissante, on lui associe dorénavant un compensateur $(\Lambda_t, t \geq 0)$ croissant de telle sorte que $(M_t = N_t - \Lambda_t, t \geq 0)$ soit une martingale.

Pour un processus de Poisson temporel, $d\Lambda_t$ (mesure associée au sens de l'intégrale de Stieljes à $(\Lambda_t, t \geq 0)$) est la mesure moyenne du processus, $d\nu$.

Pour un processus de comptage général, si I et J sont deux intervalles disjoints, ce qui se passe dans I peut dépendre de ce qui se passe dans J si J est avant I. Les processus de comptage permettent donc de modéliser plus de situations que les processus de Poisson.

Une construction précise ainsi que les propriétés probabilistes des processus de comptage sont données dans le livre [**17**]. Une étude très poussée des nombreuses applications statistiques de ces processus a été faite dans le livre [**2**].

0.1.3. Processus à intensité multiplicative d'Aalen. Les processus de comptage que nous allons étudier ici sont dits à intensité multiplicative d'Aalen car leur compensateur vérifie

$$d\Lambda_t = Y_t s(t) dt,$$

où dt représente la mesure de Lebesgue, $(Y_t, t \geq 0)$ est un processus prévisible observé et s est une fonction déterministe inconnue, que l'on va chercher à estimer avec les observations de N et Y.

Les processus de Poisson temporels en sont un cas particulier : s est l'intensité du processus de Poisson par rapport à la mesure $d\mu = Y dt$ où Y est une constante déterministe dans ce cas.

Il y a plusieurs autres exemples classiques de ces processus (voir **[2]** pour une liste détaillée). Rappelons ici quelques uns des exemples les plus étudiés.

- **Taux de hasard**

  Il y a des processus avec un seul saut. L'exemple le plus simple est le suivant : si X est une variable positive de densité f, $(N_t, t \geq 0)$, donné par $(\mathbb{1}_{X \leq t}, t \geq 0)$, est un processus de comptage à un seul saut avec une intensité multiplicative où $Y_t = \mathbb{1}_{X \geq t}$ est un prévisible observable et où $s(t)$ est $f(t)/\mathbb{P}(X \geq t)$. On dit que s est le taux de hasard (hazard rate) de X. Cette quantité est très importante en médecine ou en fiabilité. Si X est une durée de vie, $s(t)$ représente la probabilité de rester en vie un peu après t sachant qu'on était vivant en t.

  On a souvent un n-échantillon de ces durées de vies : $X_1, ..., X_n$, chacun de taux de hasard s. A chacun correspond un processus de comptage N^i et un prévisible Y^i. Le processus N qui est la somme des N^i est lui aussi un processus de comptage à intensité multiplicative avec Y qui est la somme des Y^i et avec toujours s, le taux de hasard, comme fonction déterministe. Dans ce cas, Y_t représente le nombre d'événements qui vont avoir lieu après t. Il est borné par n, le nombre total d'événements.

- **Modèles Markoviens**

  Ces modèles sont eux aussi très fréquents : on dispose de $(X(t), t \geq 0)$ processus de Markov à espace d'états finis. Le processus de comptage $(N_t^{hj}, t \geq 0)$ est celui qui compte le nombre de transitions de l'état h vers l'état j avant l'instant t. Il a une intensité multiplicative de la forme $Y_t^h s_{hj}(t)$ où $s_{hj}(t)$ est l'intensité de transition de h vers j et où Y_t^h est $\mathbb{1}_{X(t)=h}$.

  Comme précédemment, on peut avoir un n-échantillon, à chacun d'eux associer un processus de comptage et faire la somme : le nouveau Y_t^h correspond alors au nombre de personnes dans l'état h, il est toujours majoré par le nombre total de processus individuels.

  Ces modèles Markoviens modélisent par exemple les passages d'état sain à état malade puis guérison de certains patients (processus de Markov à deux états, malade et sain).

- **Censure**

  La censure permet de modéliser le fait que certaines informations ou certains temps d'apparition exacts manquent. Cela revient à dire qu'au lieu d'observer N on observe N^c processus censuré qui ne contient qu'une partie de l'information. Il y a énormément de types de censure **[2]**.

Rappelons ici uniquement le cas de censure à droite pour un n-échantillon qui est le cas plus simple. On dispose de $X_1, ..., X_n$, n durées de vie i.i.d. de taux de hasard s. Soient $U_1, ..., U_n$ d'autres variables positives i.i.d. On observe $(\tilde{X}_i, D_i)$ où $\tilde{X}_i$ est $X_i \wedge U_i$; si $\tilde{X}_i = X_i$, D_i vaut 1, sinon D_i vaut 0 et on dit alors que X_i a été censuré par U_i. Les U_i forment la censure. Ils modélisent le fait que l'on n'observe pas toujours la fin de vie d'un composant ou d'un patient. Le patient peut quitter par exemple le suivi de l'hôpital. On n'observe donc pas le décès mais on sait que le patient était toujours en vie à l'instant de censure. Il ne faut donc pas jeter cette information mais s'en servir.

Le processus censuré individuel correspondant a au plus un saut : c'est

$$N_t^{ic} = \mathbb{1}_{\tilde{X}_i \leq t} D_i$$

qui saute quand on observe vraiment X_i et pas U_i. Il a une intensité multiplicative où Y_t^i vaut $\mathbb{1}_{\tilde{X}_i \geq t}$ et où s est le taux de hasard des X_i. On peut bien entendu passer au processus agrégé en faisant la somme en i. Le nouveau processus $(Y_t, t \geq 0)$ agrégé représente alors le nombre d'événements que l'on n'a pas encore vus en $t-$ en fonction de t. Il prend en compte entre autres les données qui ont été censurées.

- **Autres**

  Un processus de comptage n'est pas toujours une somme de processus indépendants. Prenons le processus de comptage de l'exemple III.1.10 de **[2]** qui compte les accouplements de mouches drosophiles. Le processus ne peut être vu comme une somme de processus indépendants, mais le prévisible observable Y représente toujours un nombre possible d'événements à venir.

  Il y a aussi d'autres exemples où Y n'est plus une fonction à valeur entière. Nous ne nous y intéresserons pas ici.

Dans chacun des cas précédents, on veut estimer de manière non paramétrique et si possible adaptative l'intensité s. Pour mettre en place chacune des notions, nous allons nous cantonner pour l'instant à un exemple très simple : celui des processus de Poisson temporels. Nous allons estimer leur intensité par un histogramme choisi par sélection de modèle.

0.2. Un cas simple : estimateur en histogramme pour les processus de Poisson temporels

Nous avons donc un processus de Poisson temporel inhomogène N dont on ne connaît les instants de sauts qu'entre 0 et un T fixé, temps de fin d'étude. Sa mesure moyenne est absolument continue par rapport à la mesure de Lebesgue et s est l'intensité du processus relative à cette mesure. On veut l'estimer par un histogramme. Regardons d'abord ce qui se passe quand la partition est fixée.

0.2.1. Estimateur par projection.

Avant tout, définissons ici le contraste des moindres carrés, fonctionnelle convexe ne dépendant que des observations : pour tout f de $\mathbb{L}^2([0,T],dx)$,

$$\gamma_T(f) = -\frac{2}{T}\int_0^T f(t)dN_t + \frac{1}{T}\int_0^T f^2(t)dt. \tag{0.2.1}$$

Soit m une partition fixée de $[0,T]$ et S_m l'espace vectoriel des fonctions en escalier construites sur m : S_m est appelé un **modèle**. L'**estimateur par projection** de s sur S_m est

$$\hat{s}_m = \arg\min_{f\in S_m} \gamma_T(f). \tag{0.2.2}$$

Le contraste est appelé contraste des moindres carrés car

$$\mathbb{E}(\gamma_T(f)) = \|s-f\|^2 - \|s\|^2,$$

où $\|f\|^2 = (1/T)\int_0^T f^2(t)dt$. Le contraste moyen est donc minimal en $f=s$. Quand on le minimise sur S_m seulement, on obtient donc la projection orthogonale de s sur S_m, s_m, c'est-à-dire la fonction qui minimise la distance au carré de s à S_m. La minimisation du contraste lui-même et non de sa moyenne conduit donc à l'estimateur "par projection".

En fait, on peut écrire $\hat{s}_m$ beaucoup plus simplement :

$$\hat{s}_m = \sum_{I\in m} \frac{N_I}{l(I)}\mathbb{1}_I,$$

où $l(I)$ représente la longueur de l'intervalle I. Il suffit pour cela de décomposer f sur la base $\{\sqrt{(T/l(I))}\mathbb{1}_I, I\in m\}$ qui est une base orthonormée pour la norme $\|.\|$ qui apparaît à droite dans le contraste.

De même, on peut réécrire s_m :

$$s_m = \sum_{I\in m} \frac{\alpha_I}{l(I)}\mathbb{1}_I,$$

où $\alpha_I = \int_I s(t)dt = \mathbb{E}(N_I)$. La projection s_m est alors de manière assez évidente l'espérance de $\hat{s}_m$.

Mais ce qu'on veut, c'est considérer $\hat{s}_m$ comme un estimateur potentiel de s et non de s_m. On peut alors écrire par Pythagore que

$$\begin{aligned} \|\hat{s}_m - s\|^2 &= \|s-s_m\|^2 + \|s_m-\hat{s}_m\|^2 \\ &= \|s-s_m\|^2 + \sum_{I\in m}\frac{(N_I-\alpha_I)^2}{Tl(I)}. \end{aligned} \tag{0.2.3}$$

Le terme de gauche dans (0.2.3) est déterministe. Il est appelé terme de **biais**, il mesure la capacité d'approximation de l'espace S_m vis-à-vis de s. Il a tendance à devenir très petit quand la partition devient fine.

Le terme de droite, aléatoire, est une statistique de type χ^2 : c'est une somme de variable indépendantes, au carré. Quand on étudie les estimateurs de moindres carrés dans un cadre beaucoup plus simple, celui des modèles gaussiens **[14]**, ce terme est vraiment un χ^2 gaussien. Par analogie, on appellera toujours ce terme un χ^2. On le notera $\chi^2(m)$. Son espérance est couramment appelée terme de **variance** et vaut $\sum_I(\alpha_I/(Tl(I)))$. Si s est comprise entre r et R, $\mathbb{E}(\chi^2(m))$ est comprise entre $r|m|/T$ et $R|m|/T$ où $|m|$ représente le nombre d'intervalles dans la partition m et donc la dimension du modèle S_m.

Si $\chi^2(m)$ est proche de son espérance, il va se comporter comme un multiple de la dimension du modèle et grossir à l'inverse du biais quand la partition devient fine. Or ce qu'on veut c'est trouver un bon modèle sur lequel projeter, celui qui minimise la norme 2 : il faut donc faire ce qu'on appelle couramment un compromis biais-variance.

0.2.2. Choix du modèle : estimateur par projection pénalisé. On se donne donc maintenant une famille de partitions, $\mathcal{M}_T$, qui peut aussi être vue comme une famille de modèles. En effet, par soucis de simplicité de notations, ceci n'entraînant aucune confusion dans notre contexte, nous dirons très souvent que la partition m est un modèle.

Le meilleur estimateur possible parmi les $\{\hat{s}_m, m \in \mathcal{M}_T\}$ est celui qui minimise (0.2.3), on l'appelle **l'oracle** car on ne peut le trouver sans connaître s. Il est associé au modèle $S_{\bar{m}}$ défini par

$$\begin{aligned} \bar{m} &= \arg\min_{m\in\mathcal{M}_T} \|s-\hat{s}_m\|^2 \\ &= \arg\min_{m\in\mathcal{M}_T} \left(-\|\hat{s}_m\|^2 + 2\chi^2(m) + 2(\hat{s}_m - s_m|s_m)\right). \end{aligned} \tag{0.2.4}$$

où $(.|.)$ désigne le produit scalaire associé à la norme $\|.\|$.

Ce qu'on veut c'est trouver $\hat{m}$ ne dépendant que des données de telle sorte que les deux estimateurs, $\hat{s}_{\hat{m}}$ et $\hat{s}_{\bar{m}}$, se ressemblent. On peut pour cela procéder à une heuristique qui consiste à estimer sans biais les quantités ci-dessus. Le dernier terme est nul en moyenne. Le premier est égal à $\gamma_T(\hat{s}_m)$. Quant au χ^2, il peut être estimé par $\sum_{I\in m} N_I/(Tl(I))$ qui a même moyenne.

Le bilan est donc le suivant. Si tout le monde est proche de son espérance, on peut choisir $\hat{m}$, modèle sur lequel on va projeter par la procédure suivante qui ne dépend elle que des observations :

$$\hat{m} = \arg\min_{m\in\mathcal{M}_T} \left(\gamma_T(\hat{s}_m) + 2\sum_{I\in m} \frac{N_I}{Tl(I)}\right).$$

Ce raisonnement est proche de l'heuristique de Mallows **[47]** dans le cas gaussien.

On étudiera plus généralement les critères suivants :

$$\hat{m} = \arg\min_{m\in\mathcal{M}_T} \left(\gamma_T(\hat{s}_m) + \mathrm{pen}(m)\right) \tag{0.2.5}$$

où pen est une fonction de $\mathcal{M}_T$ dans $\mathbb{R}^+$ appelée pénalité. Ce qu'on appelle l'estimateur par projection pénalisé (**p.p.e.**) est alors $\tilde{s} = \hat{s}_{\hat{m}}$.

Des critères identiques, construits pour des contrastes différents, furent étudiés dans de nombreux cas par L. Birgé et P. Massart, les cadres les plus simples étant les modèles gaussiens **[14]** et les modèles de densité **[12]**. Nous allons nous inspirer de leur méthode pour étudier $\tilde{s}$.

Précisons que la fonction de pénalité n'est pas forcément déterministe, mais ne doit pas dépendre de paramètres inconnus. Elle est sensée être de l'ordre du $\chi^2(m)$.

Si on reprend l'heuristique, la pénalité vaut donc

$$\mathrm{pen}(m) = 2\sum_{I\in m} \frac{N_I}{Tl(I)}.$$

C'est ce qu'on appellera pénalité "à la Mallows". Ce n'est pas celle proposée par Mallows dans son article, cette dernière étant connue sous le nom de C_p de Mallows.

0.2.3. Contrôle des χ^2. Pour montrer que ces pénalités conduisent à des estimateurs raisonnables, il faut entre autres pouvoir contrôler les χ^2. Or ici typiquement les χ^2 n'ont pas de transformée de Laplace car une variable de Poisson n'en a pas. Plusieurs méthodes s'offrent alors à nous.
Soit on passe aux $\chi(m)$ qui sont les racines des χ^2 et qui peuvent s'exprimer comme

$$\chi(m) = \sup_{f\in S_m, \|f\|^2=1} \int_0^T f(t)dM_t$$

où $dM_t = dN_t - s(t)dt$ est une martingale infiniment divisible.
Soit on peut se servir ici de l'indépendance dans la somme des χ^2 et utiliser l'inégalité de Bernstein sur des variables tronquées.

Imaginons qu'on dispose d'une grande partition régulière Γ de $[0,T]$ et que $\mathcal{M}_T$ soit une famille de sous-partitions de Γ (c'est-à-dire construites par réunion des intervalles de Γ.) On peut alors montrer le résultat suivant en appliquant l'inégalité de Bernstein à la somme de variables indépendantes $\sum_I (X_I/l(I))$ où

$$X_I = (N_I - \alpha_I)^2 \wedge (\varepsilon\alpha_I)^2.$$

On note $|m|$ le cardinal de m.

PROPOSITION 1. *Soit ε strictement positif. Posons*

$$\Omega(\varepsilon) = \{|N_I - \alpha_I| \le \varepsilon\alpha_I, \text{ pour tout } I \in \Gamma\}.$$

Alors pour tout x positif, sur $\Omega(\varepsilon)$ avec probabilité plus grande que $1-\exp(-x)$ on a

$$\chi^2(m) \le \frac{R_\Gamma}{T}\left[\square|m| + \square\sqrt{|m|x} + \square x\right] \tag{0.2.6}$$

où $R_\Gamma = \sup_{I\in\Gamma}(\alpha_I/l(I))$ et où les carrés sont des fonctions continues positives de ε.

Dans toute la suite de l'introduction, on notera par des carrés des fonctions positives des paramètres du problème absents du reste de l'inégalité. Ces fonctions sont calculables (voir les différents chapitres) mais l'écriture in extenso nuirait à la compréhension.

On peut remarquer plusieurs choses. Tout d'abord regardons les ordres de grandeurs. Pour pouvoir comparer cette inégalité à celles qui viendront par la suite, passons à $\chi(m)$. L'inégalité (0.2.6) devient

$$\chi(m) \leq \square\sqrt{\frac{R_\Gamma|m|}{T}} + \square\sqrt{\frac{R_\Gamma x}{T}}. \tag{0.2.7}$$

Le premier terme est déterministe de l'ordre de la racine de $\mathbb{E}(\chi^2(m))$. Il évolue donc comme un multiple de $\sqrt{|m|/T}$. Le deuxième terme est un terme quadratique, indépendant de la dimension et donc considéré comme un terme complémentaire. Ici, il n'y a pas d'autres termes complémentaires car on s'est restreint à $\Omega(\varepsilon)$. Sur $\Omega(\varepsilon)$, les $\chi(m)$ sont donc sous gaussiens, c'est-à-dire que leur queue de répartition est plus petite que $\exp(-\square x^2)$.

L'inégalité dit donc entre autres que sur $\Omega(\varepsilon)$, avec probabilité plus grande que $1 - \sum_{m\in\mathcal{M}_T} e^{-x_m}$ pour des x_m bien choisis, tous les $\chi^2(m)$ sont bornés grosso modo par un multiple de $\sqrt{|m|/T}$. En particulier $\chi^2(\hat{m})$ terme doublement aléatoire car le choix du modèle est aléatoire, est proche de $R_\Gamma|\hat{m}|/T$. C'est pour ce contrôle-ci que les inégalités exponentielles sont vraiment utiles : on ne peut contrôler $\chi^2(\hat{m})$ dont on ne sait rien (même pas son espérance) que si on contrôle tous les $\chi(m)$ et pour les contrôler finement on a besoin d'une inégalité exponentielle.

NB : Pour tout ε, $\Omega(\varepsilon)$ est un événement de grande probabilité quand T tend vers l'infini : en effet, la probabilité de son complémentaire tend vers 0 plus vite que n'importe quelle puissance négative de T si $|\Gamma|$ est inférieur à $T/\ln^2(T)$.

0.2.4. Sélection de modèle et inégalité d'oracle. Ce qu'on souhaiterait idéalement démontrer pour le p.p.e., $\tilde{s}$, serait une **inégalité d'oracle** : c'est-à-dire une inégalité du type

$$\mathbb{E}(\|s - \tilde{s}\|^2) \leq \square \inf_{m\in\mathcal{M}_T} \mathbb{E}(\|s - s_m\|^2).$$

Cette inégalité veut dire qu'à constante près, le risque de $\tilde{s}$ est de l'ordre du plus petit risque dans la famille d'estimateurs par projection considérée.(Le risque d'un estimateur $\hat{s}$ de s est $\mathbb{E}(\|s - \hat{s}\|^2)$.)

Pour cela, on veut bien entendu se servir de la majoration probabiliste du $\chi^2(\hat{m})$ qu'on a trouvé précédemment, pour donner une pénalité adéquate. Or R_Γ apparaît. Il peut facilement être ici estimer par

$$\tilde{R}_\Gamma = \sup_{I\in\Gamma} \frac{N_I}{l(I)}.$$

En utilisant cette inégalité exponentielle on n'obtient pas toujours une inégalité d'oracle mais une version plus faible appelée **inégalité de type oracle**.

PROPOSITION 2. *Soit Γ une partition régulière fixe de $[0,T]$ de cardinal inférieur à $T/\ln^2(T)$ et soit $\mathcal{M}_T$ famille de sous-partitions de Γ.*
Soit une famille de réels positifs $(L_m, m \in \mathcal{M}_T)$ appelés poids tels qu'il existe Σ,

$$\sum_{m\in\mathcal{M}_T} e^{-L_m|m|} \leq \Sigma.$$

Soit d plus grand que 1. Pour tout m dans $\mathcal{M}_T$ on pose :

$$\text{pen}(m) = d\tilde{R}_\Gamma \frac{|m|}{T}(1+\kappa(d)\sqrt{L_m})^2 \tag{0.2.8}$$

avec $\kappa(d)$ connue. Alors si s est minorée par ϱ, il existe C fonction continue positive de d telle que

$$\mathbb{E}(\|\tilde{s}-s\|^2) \leq C(d) \inf_{m\in\mathcal{M}_T} (\|s-s_m\|^2 + R_\Gamma \frac{|m|L_m}{T}) + \square\frac{1}{T}.$$

On peut tout d'abord s'interroger sur les poids L_m. Si il y a au plus un modèle par dimension, par exemple si les partitions sont emboîtées, on peut prendre L_m constant égal à 1. Si on veut prendre des familles de partitions plus complexes, par exemple toutes les partitions construites de manière exhaustive sur Γ, les poids ne sont alors plus des constantes mais des multiples de $\ln(T)$. Ils mesurent la complexité du modèle. On verra plus tard qu'on ne peut pas se passer de ces poids en un certain sens.

Nous parlerons d'asymptotique pour les processus de Poisson quand T tend vers l'infini ce qui revient à dire, quand s est minorée, que $\mathbb{E}(N_{[0,T]})$, nombre d'observations total moyen, grandit. La borne sur le cardinal de la plus grande partition est alors compréhensible : si le nombre total d'observations se comporte comme T, on a pour Γ au moins $\ln(T)^2$ points par intervalle en gros.

Dans le cas des partitions emboîtées, la proposition nous donne une inégalité d'oracle asymptotique car ce qui se situe dans l'infimum, est de l'ordre du risque de chaque estimateur par projection. Nous dirons alors que le p.p.e. est **adaptatif dans la famille d'estimateurs** $(\hat{s}_m, m \in \mathcal{M}_T)$. En effet, il a à constante près le même risque que celui qui réalise l'infimum.

La majoration donnée par le théorème ne dit rien quand d est proche de 1 ou quand d est très grand (la borne explose). De toutes façons, cette borne ne donne qu'un ordre de grandeur. Pour trouver un d adéquat, mieux vaut une étude par simulation du p.p.e. Dans le cas emboîté, généralement $d = 2$ marche bien.

On peut remarquer aussi qu'on ne valide pas la pénalité "à la Mallows", mais seulement une majoration de celle-ci : $\text{pen}(m) = 2\tilde{R}_\Gamma|m|/T$. Sous certaine hypothèses, on montrera des inégalités d'oracle pour les pénalités "à la Mallows'.

0.3. Les autres familles de modèles et les autres contrastes.

0.3.1. Processus de Poisson. On peut traiter des processus de Poisson plus généraux. Le contraste devient alors pour f dans $\mathbb{L}^2(\mathbb{X}, d\mu)$

$$\gamma_{\mathbb{X}}(f) = -\frac{2}{\mu(\mathbb{X})}\int_{\mathbb{X}} f(t)dN_t + \frac{1}{\mu(\mathbb{X})}\int_{\mathbb{X}} f^2(t)dt. \tag{0.3.1}$$

L'asymptotique se fait alors quand $\mu(\mathbb{X})$ et non plus T tend vers l'infini.

On peut aussi utiliser des modèles plus généraux que les histogrammes. Prenons $(S_m, m \in \mathcal{M}_{\mathbb{X}})$ une famille de sous-espaces vectoriels de dimension finie, qu'on appelle encore modèles. Il faut quand même qu'ils vérifient certaines conditions, les exemples classiques étant soit des polynômes trigonométriques, soit des espaces engendrés par une famille finie d'ondelettes à support compact, pour la mesure μ qui est alors de Lebesgue sur $[0, T]$.

Les estimateurs par projection sur chaque modèle se construisent comme précédemment en minimisant le contraste $\gamma_{\mathbb{X}}$ sur les S_m.

Si on connaît une base orthonormée du modèle pour la norme $\|.\|$, $\{\varphi_\lambda, 1 \leq \lambda \leq D_m\}$, (où D_m est la dimension du modèle S_m), on obtient aussi une expression plus simple de $\hat{s}_m$:

$$\hat{s}_m = \sum_{\lambda=1}^{D_m} \left(\int_{\mathbb{X}} \varphi_\lambda(x) \frac{dN_x}{\mu(\mathbb{X})} \right) \varphi_\lambda.$$

Il est alors clair qu'il a pour espérance s_m, la vraie projection de s sur le modèle S_m (toujours déterministe).

Le risque se décompose toujours en un terme de biais de la forme $\|s - s_m\|^2$ déterministe et un terme de variance qui est l'espérance d'un $\chi^2(m)$ qu'on peut écrire ici :

$$\chi^2(m) = \sum_{\lambda=1}^{D_m} \left(\int_{\mathbb{X}} \varphi_\lambda(x) \frac{dN_x - s(x)d\mu_x}{\mu(\mathbb{X})} \right)^2.$$

Chaque terme varie de la même façon que précédemment en fonction de D_m.

On peut toujours proposer une heuristique qui ressemble à celle de Mallows et qui permet de choisir le modèle sur lequel projeter. L'heuristique "à la Mallows" donne alors une pénalité en $2\sum_{\lambda=1}^{D_m} \int_{\mathbb{X}} \varphi_\lambda(x)^2 (dN_x/\mu(\mathbb{X})^2)$.

Plus généralement, on minimise le critère (0.2.5) avec $\gamma_{\mathbb{X}}$ à la place de γ_T.

La clef pour valider tout cela est toujours la concentration du $\chi^2(m)$ autour de son espérance et ici on se servira en fait de l'expression suivante

$$\chi(m) = \sup_{f \in S_m, \|f\|^2 = 1} \int_{\mathbb{X}} f(x) \frac{dN_x - s(x)dx}{\mu(\mathbb{X})}.$$

0.3.2. Processus à intensité multiplicative d'Aalen. Le contraste des moindres carrés pour les processus à intensité multiplicative est plus compliqué. Par homothétie, on va se ramener à étudier notre processus sur $[0, 1]$, intervalle sur lequel on veut estimer s

(on a éventuellement quelques observations à l'extérieur). On voit alors que le processus de Poisson temporel sur $[0,T]$ est par homothétie un processus à intensité multiplicative d'Aalen sur $[0,1]$ avec Y qui est constant et qui vaut T. On introduit donc A qui est une borne déterministe sur Y et qui jouera le rôle asymptotique de T ici. Dans le cas censure à droite, A vaut n, le nombre d'individus, par exemple.

Le contraste des moindres carrés est alors pour f dans $\mathbb{L}^2([0,1],dt)$:

$$\gamma_A(f) = -\frac{2}{A}\int_0^1 f(t)dN_t + \frac{1}{A}\int_0^1 f^2(t)Y_t dt. \tag{0.3.2}$$

L'estimateur par projection est toujours celui qui minimise ce contraste sur un modèle, espace vectoriel de dimension finie.

Le terme de droite dans le contraste est ici aléatoire, il définit une norme aléatoire $\|.\|_{\text{rand}}$: pour tout f dans $\mathbb{L}^2([0,1],dx)$,

$$\|f\|^2_{\text{rand}} = \frac{1}{A}\int_0^1 f^2(t)Y_t dt.$$

Si on prend l'espérance du terme ci-dessus, on définit une norme déterministe : pour tout f dans $\mathbb{L}^2([0,1],dt)$

$$\|f\|^2_{\text{det}} = \frac{1}{A}\int_0^1 f^2(t)\mathbb{E}(Y_t) dt.$$

Dans le cas Poisson, ces deux normes sont égales et déterministes. Elles valent $\|.\|$.

Ce contraste est intéressant car si on intègre de 0 à t au lieu de 1, on peut voir le contraste ci-dessus comme un processus de compensateur

$$-\frac{2}{A}\int_0^t f(u)s(u)Y_u du + \frac{1}{A}\int_0^t f^2(u)Y_u du.$$

En 1, c'est donc $\|f-s\|^2_{\text{rand}} - \|s\|_{\text{rand}}$ qui est bien minimal en $f=s$. Donc non seulement le contraste moyen mais aussi le compensateur du contraste sont minimaux en $f=s$.

La minimisation du contraste sur un modèle S_m, sous-espace vectoriel de $\mathbb{L}^2$ donne l'estimateur par projection de s sur S_m, $\hat{s}_m$.

Remarque : On aurait aussi pu imaginer un contraste du genre :

$$\gamma(f) = -2\int_0^1 f(t)\frac{dN_t}{Y_t} + \int_0^1 f^2(t)dt,$$

en supposant que Y ne s'annule pas. La norme à droite est alors déterministe ce qui semble plus facile à gérer. L'espérance du contraste (car Y est prévisible) est bien minimale en s et le contraste ne dépend que des observations. Cependant les estimateurs par projection qu'il fournit ont une variance qui peut devenir très grosse et cela même pour des dimensions de modèles très petites car elles sont en $1/Y$ au lieu de $1/T$. Prenons le cas censure à droite où Y décroît de n à 1, et où toutes les observations se font dans $[0,1]$ sauf la dernière. On obtient des variances qui peuvent être en D_m (où D_m est la dimension du modèle S_m) si

le modèle charge la fin de l'intervalle au lieu de D_m/n si il charge le début. En particulier, à dimension fixée si n tend vers l'infini, la variance ne tend plus vers 0 pour certains modèles. On utilisera donc dans le cas processus à intensité multiplicative le contraste défini par (0.3.2).

Le problème est maintenant de trouver des modèles sur lesquels les techniques de sélection de modèle marcheront bien. Une des conditions en particulier qui est inévitable est un contrôle de la quantité suivante :

$$\Phi(m) = \sup_{t \in S_m} \frac{\|t\|_\infty}{\|t\|_{\text{rand}}}. \tag{0.3.3}$$

Cette quantité intervient déjà dans la plupart des articles sur la sélection de modèle, en particulier [**9**]. Si cette quantité est majorée, par $\sqrt{AD_m}$ à constante près (où D_m est la dimension du modèle S_m), on pourra encore dire que la variance (et le χ^2) sont de l'ordre de D_m/A (voir Section 0.4).

Les modèles qui restent donc envisageables, sont par exemple les modèles d'histogrammes. Il faut juste trouver une base orthonormée pour cette norme aléatoire mais on a déjà une famille orthogonale : les indicatrices d'intervalles. Si m est une partition de $[0,1]$, la base orthonormée est alors $\{\mathbb{1}_I/\sqrt{b_I}, I \in m\}$ où

$$b_I = \frac{1}{A}\int_I Y_t dt.$$

La famille de modèles est alors une famille de sous-partitions d'une partition fixe Γ.

Pour un autre modèle déterministe, il faut pouvoir contrôler la quantité $\Phi(m)$: quand on ne connaît pas de base orthonormée pour la norme aléatoire, cela semble assez difficile.

Les autres modèles que nous regarderons donc dans ce cas seront des modèles aléatoires, prévisibles, définis de la manière suivante. Soit $\{\varphi_\lambda, \lambda \in \Gamma\}$ une famille finie orthonormée pour la norme classique $\|.\|$ sur $[0,1]$. Alors sur l'événement Ω "Y ne s'annule pas sur $[0,1]$", pour tout m sous-ensemble de Γ, $\{\varphi_\lambda(.)\sqrt{A}/\sqrt{Y_.}\mathbb{1}_{Y_. \neq 0}, \lambda \in m\}$, est une famille orthonormée pour $\|.\|_{\text{rand}}$. On regarde alors comme modèle S_m l'espace engendré par ces fonctions : il est aléatoire et prévisible car $Y_.$ l'est. La famille de modèles peut alors être assimilée à une famille de sous-ensembles de Γ.

Comme Y est observable, on peut très facilement voir si on est dans Ω ou non : en censure à droite, il suffit d'avoir une observation à l'extérieur de l'intervalle par exemple.

Les modèles prévisibles de ce type qui ont un $\Phi(m)$ borné corectement sont alors typiquement ceux où les φ_λ proviennent d'une base de Fourier ou d'une base d'ondelettes à support compact.

Dans les deux cas que nous étudierons donc (histogrammes et modèles prévisibles), on connaît une base orthonormée du modèle pour la norme qui apparaît dans le contraste et l'estimateur par projection a une écriture très simple.

Pour les histogrammes on obtient pour m partition de $[0,1]$

$$\hat{s}_m = \sum_{I \in m} \frac{N_I}{Ab_I} \mathbb{1}_I.$$

Si b_I nul, on a généralement N_I nul et le coefficient est nul. Même si ça ne pose pas de problème de définition, en pratique on ne garde que les coefficients pour lesquels b_I n'est pas trop petit (cf Chapitre 3). Pour ce chapitre introductif, nous oublierons ce détail. Nous supposons pour les histogrammes que le processus est agrégé, cést-à-dire que c'est une somme de n processus i.i.d.

Pour les modèles prévisibles, on obtient pour m ensemble d'indices et sur Ω

$$\hat{s}_m(.) = \sum_{\lambda \in m} \left(\int_0^1 \frac{\varphi_\lambda(t)}{\sqrt{Y_t}} \frac{dN_t}{\sqrt{A}} \right) \varphi_\lambda(.) \sqrt{\frac{A}{Y_.}}.$$

Ici aussi on peut avoir des difficultés si Y s'annule pour la définition. On traite ce problème dans le Chapitre 4. Pour l'introduction, on va supposer que Y n'est jamais nul, i.e. on suppose implicitement jusqu'à la fin qu'on est sur Ω.

En outre, pour ces modèles prévisibles, on suppose que si Y n'est pas nul il est plus grand qu'un certain c qui vaut 1 dans les cas classiques car Y est entier le plus souvent.

La projection, s_m, est bien entendue prise pour la norme aléatoire.

Dans le cas des modèles prévisibles, il est assez clair (en passant par les compensateurs) que les coefficients de $\hat{s}_m$ estiment sans biais ceux de s_m. Le terme de variance est alors l'espérance de

$$\chi^2(m) = \sum_{\lambda \in m} \left(\int_0^1 \frac{\varphi_\lambda(t)}{\sqrt{Y_t}} \frac{(dN_t - Y_t s(t) dt)}{\sqrt{A}} \right)^2.$$

Il a pour compensateur pris en 1 :

$$\sum_{\lambda \in m} \int_0^1 \varphi_\lambda(t)^2 s(t) \frac{dt}{A}$$

qui est constant sur Ω. Si s est majoré par R, on a alors encore une variance de l'ordre de RD_m/A.

Le biais par contre n'est plus déterministe : on peut le voir comme la distance classique $\|.\|^2$ entre $\rho(t) = s(t)\sqrt{Y_t}$ et la projection de ρ sur le modèle déterministe $\mathrm{Vect}\{\varphi_\lambda, \lambda \in m\}$. Il a donc lui aussi tendance à diminuer quand D_m augmente.

Tout se passe donc dans les modèles prévisibles comme si on travaillait avec des Poissons (d'ailleurs dans les cas poissonniens $Y = T$ est constant et on retrouve exactement la construction décrite dans le cas Poisson).

Pour trouver un bon modèle, on minimise doncun critère pénalisé de la forme

$$\hat{m} = \arg\min_{m \in \mathcal{M}_A} -\gamma_A(\hat{s}_m) + \mathrm{pen}(m) \tag{0.3.4}$$

où la pénalité est une fonction connue de $\mathcal{M}_A$ dans $\mathbb{R}^+$. Le p.p.e. est alors $\tilde{s} = \hat{s}_{\hat{m}}$.

On peut trouver des pénalités plausibles par une heuristique à la Mallows (même si on ne pourra pas les valider).

Pour trouver de manière générale des pénalités conduisant à des inégalités de type oracle, il faut contrôler les χ^2 et pour cela on préfère passer aux χ qu'on interprète de la manière suivante :

$$\chi(m) = \sup_{\sum_{\lambda\in m} a_\lambda^2 = 1} \int_0^1 \left(\sum_{\lambda\in m} a_\lambda \frac{\varphi_\lambda}{\sqrt{Y_t}} \mathbb{1}_{Y_t\neq 0} \right) \frac{dM_t}{\sqrt{A}} \tag{0.3.5}$$

où dM_t représente la martingale $dN_t - Y_t s(t)dt$.

Pour les histogrammes, l'estimateur par projection n'a en revanche aucune raison a priori d'être un bon estimateur de la projection s_m. Pour pouvoir le dire, il faut tout d'abord dire que les b_I sont proches d'un terme déterministe par exemple leur espérance β_I et pour cela utiliser encore de la concentration. C'est le cas si Y est une somme de processus indépendants, ce qui est le cas quand Y et N sont les processus de comptage agrégés de processus de comptage individuels. Dans ces cas-là, on peut alors encore parler de biais, de variance et d'heuristique à la Mallows (voire Chapitre 3) et minimiser le même critère pénalisé que précédemment (0.3.4) pour obtenir un p.p.e.

Mais dans ces modèles, on a une expression de $\chi^2(m)$ qui n'est pas pratique (on ne connaît même pas son espérance) :

$$\chi^2(m) = \sum_{\lambda\in m} \frac{((N_I/n) - a_I)^2}{b_I}$$

où $a_I = \int_I s(t)Y_t dt/A$. La forme suivante est bien plus pratique sous réserve que les b_I soient proches de leur espérance β_I :

$$Z^2(m) = \sum_{\lambda\in m} \frac{((N_I/n) - a_I)^2}{\beta_I}.$$

Ce sont d'ailleurs les Z et non les χ qu'on sait contrôler en probabilité, car on peut interpréter la racine carré de $Z^2(m)$ par

$$Z(m) = \sup_{\sum_I \delta_I^2 \beta_I = 1} \left(\frac{1}{n} \sum_{i=1}^n \int_0^1 \left(\sum_{I\in m} \delta_I \mathbb{1}_I(t) \right) (dN_t^i - Y_t^i s(t)dt) \right)$$

si A vaut n et si les N^i et les Y^i désignent les processus individuels qui forment N et Y.

0.4. Inégalités de concentration

Dans cette partie, nous allons expliquer comment contrôler les différents χ^2

0.4.1. Variables indépendantes. L'inégalité de concentration la plus connue est due à M. Talagrand **[62]**. Elle s'applique à un supremum de processus empiriques du type de $Z(m)$. Cette inégalité a connue de très nombreuses améliorations et preuves ([**45**],[**50**]). La dernière en date et la plus simple est celle d'E. Rio **[58]**.

PROPOSITION 3 (Inégalité de Talagrand). *Soient $X_1, ..., X_n$ des variables i.i.d. à valeurs dans $(\mathbb{X}, \mathcal{X})$. Soit $\{\psi_a, a \in \mathcal{A}\}$ une famille dénombrable de fonctions mesurables réelles sur $(\mathbb{X}, \mathcal{X})$ à valeurs dans $[-b; b]$. Soit*

$$Z = \sup_{a\in\mathcal{A}} \sum_{i=1}^{n} \left[\psi_a(X_i) - \mathbb{E}(\psi_a(X))\right].$$

Si v représente $n \sup_{a\in\mathcal{A}} \mathrm{Var}(\psi_a(X_1))$, alors pour tout u positif,

$$\mathbb{P}\left(Z \geq \mathbb{E}(Z) + \sqrt{(2v + 4b\mathbb{E}(Z))u} + (b/2)u\right) \leq \exp(-u).$$

Cette inégalité implique que pour tout u positif,

$$\mathbb{P}\left(Z \geq \square\mathbb{E}(Z) + \sqrt{2vu} + \square bu\right) \leq \exp(-u).$$

On peut l'appliquer à $Z(m)$ en disant qu'il existe un sous-ensemble dense et dénombrable de δ, A, tel que

$$Z(m) = \sup_{\delta\in A} \left(\frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left(\sum_{I\in m} \delta_I \mathbb{1}_I(t) \right) (dN_t^i - Y_t^i s(t)dt) \right).$$

Il faut juste s'assurer que les variables i.i.d. sont bornées. Pour cela on va donc supposer que chaque N^i à un nombre de sauts bornés par K et que chaque Y^i est borné par B. On obtient alors si s est borné par R, pour tout x positif

$$\mathbb{P}\left(Z(m) \geq \square \sqrt{\sum_{I\in m} \frac{\alpha_I}{n\beta_I}} + \sqrt{2R_m \frac{x}{n}} + \square b \frac{K + RB}{n} x \right) \leq e^{-x}$$

où $\alpha_I = \mathbb{E}(a_I)$, où $R_m = \sup_{I\in m} \frac{\alpha_I}{\beta_I}$ et où $b = \sup_{I\in m} \frac{1}{\sqrt{\beta_I}}$.

On obtient ici le comportement typique qui permet de faire de la sélection de modèle (voir **[51]** ou **[20]**). Le premier terme est de l'ordre de la racine de l'espérance de $Z(m)^2$ et peut être majoré par $\sqrt{R_m D_m/n}$ (où D_m est la dimension du modèle S_m, i.e. $|m|$). Le deuxième terme est un terme quadratique comme dans l'inégalité qu'on avait obtenue plus haut avec l'inégalité de Bernstein. Il dépend de R_m (une sorte de majorant de s) mais pas de la dimension D_m et peut donc être vu comme un terme correctif. Le dernier terme est un terme linéaire, beaucoup plus petit que les autres en $(\sqrt{D_m})/n$.

Ce dernier terme peut d'ailleurs disparaître si on se restreint à un ensemble de grande probabilité quand n tend vers l'infini (voir Chapitre 3). On a alors comme précédemment un comportement sous-gaussien.

0.4.2. Processus de Poisson. On peut avoir exactement le même comportement pour un $\chi(m)$ provenant d'un processus de Poisson. Plus précisément, en calquant la démonstration de P. Massart **[50]** et en utilisant le caractère infiniment divisible des processus de Poisson, on obtient (voire Chapitre 1) :

PROPOSITION 4. *Soit N un processus de Poisson inhomogène sur $(\mathbb{X}, \mathcal{X})$ de mesure moyenne finie ν. Soit $\{\psi_a, a \in A\}$ une famille dénombrable de fonctions mesurables sur $(\mathbb{X}, \mathcal{X})$ à valeurs dans $[-b, b]$. On pose*

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

Alors pour tout ε et u strictement positifs :

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

avec $\kappa = 6$ et $\kappa(\varepsilon) = 1.25 + 32/\varepsilon$.

On peut l'appliquer à $\chi(m)$ en passant encore par un sous-ensemble dense. On obtient alors pour tout u positif

$$\mathbb{P}\left(\chi(m) \geq \square\sqrt{\mathbb{E}(\chi^2(m))} + \square\sqrt{\frac{2Ru}{\mu(\mathbb{X})}} + \square\frac{\Phi(m)}{\mu(\mathbb{X})}u\right) \leq \exp(-u)$$

où bien sûr ici on construit $\Phi(m)$ (cf (0.3.3))en divisant par la norme déterministe $\|.\|$ qui est dans le cas Poisson $\|.\|_{\text{rand}}$ et où R est un majorant de s.

Le premier terme est toujours en $\sqrt{RD_m/\mu(\mathbb{X})}$. Le deuxième est toujours quadratique et indépendant de la dimension. Le dernier terme est linéaire, il est bien négligeable dans les cas classiques (Fourier ou ondelette), car $\Phi(m)$ est au plus en $\sqrt{D_m}$.

On peut même si on le souhaite se restreindre à un événement de grande probabilité quand $\mu(\mathbb{X})$ tend vers l'infini et sur lequel le comportement sera comme dans le cas simple exposé plus haut sous-gaussien.

La concentration pour les processus de Poisson a été très étudiée mais dans le cadre plus général d'une fonctionnelle quelconque du processus. Que ce soient les travaux de L. Wu **[64]** ou C. Houdré et N. Privault **[36]**, ils utilisent la structure particulière de martingales du processus de Poisson temporel. Ils obtiennent alors une inégalité de concentration faisant intervenir dans le terme quadratique le gradient de la fonctionnelle (au sens où ils le définissent). Ces inégalités appliquées ici font que l'on remplace v_0 par

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

Ceci donne (à constantes près) quand on l'applique à $\chi(m)$ le même premier terme et le même terme linéaire. Par contre maintenant le terme quadratique dépend de la dimension : il est de l'ordre de $\sqrt{RD_m u/\mu(\mathbb{X})}$. Donc cette majoration est beaucoup moins bonne. Pour

des processus de comptage plus généraux, on n'aura pas le choix et on devra se contenter de ce type de concentration.

0.4.3. Processus de comptage plus généraux. On veut trouver des inégalités exponentielles pour les suprema d'intégrales contre le processus recentré. Le principe est le suivant. On utilise des inégalités exponentielles de type Bernstein ou Bennett pour des martingales du genre $\int_0^t H_s dM_s$ où H_s est prévisible (et pas seulement déterministe car dans le cas général les modèles eux-mêmes sont prévisibles).

Il suffit ensuite de trouver le compensateur du supremum et interpréter la différence comme une intégrale contre la martingale dM_s. Le compensateur va donc remplacer ici l'espérance qui apparaît d'ordinaire. Il est donc a priori plus dur de passer à l'espérance ou au compensateur du χ^2, ce qu'on peut quand même faire quitte à perdre dans les constantes. Finalement, on va obtenir la concentration suivante pour les $\chi(m)$ des modèles prévisibles.

PROPOSITION 5. *Soit $\chi(m)$ défini par (0.3.5). Alors pour tout u positif, avec probabilité plus grande que $1-2e^{-u}$,*

$$\chi(m) - \sqrt{C(m)} \leq 3\sqrt{2vu} + bu$$

où

- $C(m) = \sum_{\lambda\in m} \int_0^1 \varphi_\lambda(t)^2 \mathbb{1}_{Y_t\neq 0} s(t)\frac{dt}{A}$
- $v = \|C(m)\|_\infty$;
- *pour tout t plus petit que* 1, $\sum_{\lambda\in m}\varphi_\lambda^2(s) \leq AY_s b^2$.

C'est-à-dire qu'on obtient dans un cadre plus général, mais uniquement pour les suprema d'intégrales, ce que donnaient les inégalités de L. Wu et C.Houdré-N.Privault.

Sous des hypothèses du type "Y minoré", on a $b = \sqrt{D_m/A}$ c'est-à-dire $\Phi(m)/A$, pour les modèles prévisibles.

Donc récapitulons, le premier terme est toujours la racine de la variance en $\sqrt{RD_m/A}$. Le deuxième terme, c'est-à-dire le terme quadratique, est lui aussi en $\sqrt{RD_m u/A}$. Enfin, et c'est cela qui est vraiment spécifique aux modèles prévisibles (et aux processus non Poissonniens), le terme linéaire est lui aussi de cet ordre de grandeur en $(\sqrt{D_m/A})u$.

Ici tous les termes sont donc du même ordre. Par conséquent, même une amélioration du type de celle qu'on a obtenu pour les Poissons dans le terme de variance, ne permettra pas d'obtenir les mêmes ordres de grandeurs que les inégalités de type Talagrand.

Les liens entre concentration et martingales sont étudiés plus longuement au Chapitre 2. On y traite le cas des suprema d'intégrales contre le processus de comptage recentré et celui des U-statistiques dégénérées d'ordre 2, ce dernier point étant un travail effectué en collaboration avec C. Houdré.

0.5. Sélection de modèle

On peut maintenant essayer de prouver des inégalités de type oracle dans chacune des situations. La structure des différentes inégalités de concentration influence bien entendu nettement les différents résultats.

0.5.1. Processus de Poisson. Commençons par les processus de Poisson, cas exposé en détail dans le Chapitre 1. Ce cadre est celui qui ressemble le plus à l'estimation de densité à partir d'un n-échantillon par projection pénalisée [**12**].

On distingue deux cas.

Soit la famille de modèle est polynômiale, i.e. le nombre de modèles de même dimension D dans la famille $\mathcal{M}_{\mathbb{X}}$ sont de l'ordre de D^{η} pour un certain η positif fixé.

Dans ce cadre-là, on peut valider précisément l'heuristique de Mallows : prenons pour d plus grand que 1 une pénalité de la forme

$$\mathrm{pen}(m) = d\sum_{\lambda=1}^{D_m}\int_{\mathbb{X}}\varphi_\lambda^2\frac{dN_x}{\mu(X)^2}.$$

Pour $d=2$, le critère pénalisé est exactement un estimateur non biaisé de $\|s-\hat{s}_m\|^2-\|s\|^2$.

On peut alors montrer sous certaines hypothèses techniques qu'il existe une fonction C continue positive telle que

$$\mathbb{E}(\|s-\tilde{s}\|^2) \leq C(d)\inf_{m\in\mathcal{M}_{\mathbb{X}}}(\mathbb{E}(\|s-\hat{s}_m\|^2)) + \square\frac{1}{\mu(\mathbb{X})}.$$

Ici le carré représente une fonction positive de tout sauf de $\mu(\mathbb{X})$. Nous obtenons donc ici une inégalité d'oracle asymptotique. L'estimateur par critère pénalisé "à la Mallows" est donc adaptatif dans la famille d'estimateurs par projection.

On peut dans le cas polynômial prouver des inégalités d'oracle pour d'autres pénalités. Elles sont parfois plus pratiques, car elles permettent de s'affranchir de certaines hypothèses comme en particulier une du type "s minorée".

Si la famille de modèles n'est pas polynômiale, on se donne une famille de poids $(L_m, m\in\mathcal{M}_{\mathbb{X}})$ tels que $\sum_{m\in\mathcal{M}_{\mathbb{X}}} e^{-L_m D_m}$ soit borné par Σ, indépendant de $\mu(\mathbb{X})$. On prend comme pénalité pour d plus grand que 1

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

où $\kappa=6$ est donné par la formule de concentration et R est une borne connue sur s. Si on ne connaît pas de borne, on peut parfois l'estimer (pour les détails voir Chapitre 1).

On obtient alors sous certaines hypothèses une inégalité de type oracle (voir Proposition 2) où on remplace R_Γ par R. Cette inégalité est vraie par exemple quand on se fixe une famille orthonormée d'ondelettes jusqu'à un niveau de résolution J,$\{\varphi_\lambda, \lambda\in\Gamma\}$, et qu'on regarde comme famille $\mathcal{M}_{\mathbb{X}}$ tous les sous-ensembles de la famille précédente. On prend

alors pour poids $L_m = \log(\mu(\mathbb{X}))$. Ce poids prend en fait en compte la complexité de la famille de modèles.

Pour cette dernière stratégie, on peut facilement voir le p.p.e. comme un estimateur par seuillage. Ce phénomène est très général et existe par exemple déjà dans le cas gaussien. Ce type d'estimateur a déjà été étudié dans le cadre des processus de Poisson par E. Kolaczyk [**41**] ou par L. Cavalier et J.-Y. Koo [**22**]. Les inégalités d'oracle qu'on trouve sont donc aussi intéressantes pour les méthodes de seuillage.

Les liens entre la forme de pénalité qu'on peut valider et la concentration qu'on peut obtenir sont très importants. C'est parce que dans le cas Poisson, on a exactement la même forme de concentration que l'inégalité de Talagrand, qu'on peut obtenir exactement les mêmes résultats que dans le cadre estimation de densité [**12**].

0.5.2. Les histogrammes pour les processus agrégés. Ici le processus n'est plus un Poisson. Rappelons les hypothèses. On suppose que N est la somme de n processus individuels i.i.d. à intensité multiplicative. Les N^i ont chacun au plus K sauts et les Y^i sont bornés par 1 et donc $A = n$.

On prend une famille de partitions toutes construites sur une même partition plus fine Γ.

On prend exactement le type de pénalité proposé en (0.2.8) avec $\tilde{R}_\Gamma$ qui est ici

$$\tilde{R}_\Gamma = \sup_{I \in \Gamma} \frac{N_I}{n b_I}$$

et $\kappa(d)$ qui est ici très simple et qui vaut $\sqrt{2}$. On a toujours une famille de poids $(L_m, m \in \mathcal{M}_A)$ qui vérifient la même condition.

Pour obtenir l'inégalité de type oracle de la Proposition 2, il faut choisir la norme qu'on prend. Pour la norme aléatoire, on va avoir une inégalité de type oracle uniquement sur un gros ensemble de probabilité où entre autres les b_I ne seront pas trop loin de leur espérance, les β_I. Pour la norme déterministe (où même la projection est prise pour la norme déterministe), on peut obtenir une inégalité de type oracle sur tout l'espace probabilisé.

Le passage entre norme aléatoire et norme déterministe ressemble beaucoup à ce que fait Y. Baraud lorsqu'il construit des estimateurs par projections pénalisés pour des modèles de régression en design aléatoire [**6**].

A ce changement près, dû à la présence de différentes normes, on a donc encore validé les mêmes types de pénalités que celle du cas Poisson. En effet la concentration ici a les mêmes ordres de grandeurs.

Dans le cadre censure à droite, S. Döhler et L. Rüschendorf [**26**] ont déjà prouvé des majorations de risque pour des p.p.e. dans des modèles déterministes beaucoup plus généraux que les histogrammes. Malheureusement, la pénalité est alors très grosse (beaucoup

plus grosse que la variance) et présuppose une connaissance du majorant de s (qu'on peut facilement estimer dans le cas histogramme). Leur pénalité est très grosse car le contrôle en probabilité qu'ils ont est un peu plus grossier que le notre : ils l'ont en effet obtenus par chaînage. C'est aussi le cas pour G. Castellan et F. Letué [**21**] lorsqu'elles estiment par sélection de modèle la fonction de régression dans le modèle de Cox.

0.5.3. Les modèles prévisibles. Ici, on ne suppose plus le processus agrégé mais juste que Y est compris entre A et c s'il n'est pas nul. On utilise alors la formule de concentration la plus générale pour le χ^2. Comme tous les termes sont du même ordre cette fois-ci, on va obtenir une forme de pénalité tout à fait différente.

Soit $\{\varphi_\lambda, \lambda \in \Gamma\}$ une famille orthonormée pour $\|.\|$ de fonctions déterministe. Soit une famille de modèles $\mathcal{M}_A$, où les S_m sont de la forme $\{\varphi_\lambda \sqrt{A/Y_t}, \lambda \in m\}$ avec m inclus dans Γ. On suppose que chaque modèle vérifie que les $\Phi(m)$ correspondant sont de l'ordre de $\sqrt{A|m|}$ où $|m|$ est le cardinal de m. Typiquement, on prend comme famille orthonormée initiale une base de Fourier tronquée et on regarde tous les sous-ensembles emboîtés dans l'ordre.

On se donne une famille de poids $(L_m, m \in \mathcal{M}_A)$ tels que

$$\sum_{I \in m} |m|^2 e^{-L_m} \leq \Sigma$$

où Σ est indépendant de A.

Prenons une pénalité de la forme

$$\text{pen}(m) = d\frac{|m|}{A}\left(\sqrt{R}(1+\square\sqrt{L_m}) + \square L_m\right)^2$$

pour d plus grand que 1. (Les carrés ici représentent des fonctions connues, indépendantes de s et légèrement compliquées et R est toujours un majorant de s.)

Alors on peut prouver (voir Chapitre 3) que sur Ω, i.e. quand Y est non nul,

$$\mathbb{E}(\|s-\tilde{s}\|^2_{\text{rand}} \mathbb{1}_\Omega) \leq \square \inf_{m \in \mathcal{M}_A} \left(\mathbb{E}(\|s-s_m\|^2_{\text{rand}}) + \text{pen}(m)\right) + \square \frac{1}{A}.$$

Prenons l'exemple précédent qui part d'une base de Fourier et les poids L_m de la forme une constante fois $\log|m|$. Nous obtenons alors une pénalité qui ressemble à $RD_m \log(D_m)/A$ c'est-à-dire qu'au logarithme près on a une légère majoration du terme de variance. Le résultat est alors quasiment une inégalité d'oracle. On peut dire donc encore qu'à un facteur logarithmique près, le p.p.e. est adaptatif dans la famille d'estimateurs par projection envisagée.

Les poids L_m peuvent paraître superflus. En effet quand Y est constant, c'est-à-dire dans le cas Poisson, cela revient à regarder les familles de fonctions orthonormées emboîtées de Fourier et nous avons vu précédemment (on est dans le cas polynômial), qu'il n'y a pas de facteur logarithmique dans l'inégalité d'oracle, si on prend des pénalités plus petites.

Cependant, dans le cas de la censure à droite, au vu des simulations, il semblerait que ces poids soient fondamentaux. Si on les prend constants, la méthode va chercher des modèles de trop grosse dimension.

0.6. Risque minimax et adaptation

0.6.1. Définition. Nous choisissons donc les pénalités pour que le p.p.e. satisfasse une inégalité de type oracle c'est-à-dire pour que grosso modo le p.p.e. soit adaptatif dans sa famille d'estimateurs par projection. Maintenant, nous aimerions pouvoir comparer le p.p.e. à des estimateurs construits de manière radicalement différente, comme par exemple les estimateurs adaptatifs à noyaux construits par G. Grégoire [**32**] pour l'intensité multiplicative d'Aalen, les estimateurs par seuillage construits par E. Kolaczyk [**41**] pour les processus de Poisson ou encore les estimateurs en ondelettes construits par A. Antoniadis, G. Grégoire et G. Nason [**5**].

Pour cela, il faut comparer les risques de chaque estimateurs. En fait on peut comparer le p.p.e. non seulement à certains estimateurs connus mais aussi à tous les estimateurs possibles en introduisant la notion de risque minimax.

Si $\mathcal{F}$ est un ensemble de fonctions possibles pour s, le risque minimax sur $\mathcal{F}$ (pour la norme déterministe) est défini de la manière suivante :

$$R(\mathcal{F}) = \inf_{\hat{s}} \sup_{s \in \mathcal{F}} \mathbb{E}(\|s - \hat{s}\|^2_{\text{det}})$$

où $\hat{s}$ parcourt l'ensemble de tous les estimateurs possibles de s. Ce risque minimax représente le risque du meilleur estimateur possible pour la pire fonction à estimer dans $\mathcal{F}$.

Ce risque est minoré la plupart du temps par quelque chose de strictement positif. Imaginons que d'un côté, on trouve une minoration d'un certain type. Imaginons que de l'autre on arrive à majorer le risque du p.p.e. par quelque chose du même ordre, quand s est dans $\mathcal{F}$, en se servant des inégalités de type oracle. On pourra alors dire que le p.p.e. ainsi construit est inaméliorable (à constante près) sur $\mathcal{F}$, on dit qu'il est **minimax**.

De plus si pour un même p.p.e., on prouve ce type de résultats pour plusieurs $\mathcal{F}$ différents, le p.p.e. sera alors **adaptatif au sens du minimax** : sans connaître $\mathcal{F}$, l'ensemble dans lequel évolue s, le p.p.e. fait aussi bien à constante près que celui qui le connaît (c'est-à-dire celui qui réalise le risque minimax sur le $\mathcal{F}$ dans lequel se trouve s).

0.6.2. Résultats. Pour cette thèse, il fallait donc calculer certaines minorations de risques minimax. Le principe est toujours le même. On se sert du lemme de Fano (voire par exemple [**10**] pour une version très pratique) et du fait que les distances de Kullback-Leibler de nos problèmes sont proches sur certains ensembles $\mathcal{F}$ de s possibles (généralement s minorée), de la norme déterministe.

Voici quelques résultats possibles. Dans le cas Poisson temporel, sur des $\mathcal{F}$ du type boules d'espace de Besov pour la norme 2, de régularité α plus grande que $1/2$, on trouve une minoration du risque en $T^{-\frac{2\alpha}{2\alpha+1}}$. Cette vitesse est atteinte par le p.p.e. pour certaines stratégies de familles emboîtées construites sur des ondelettes de régularité r plus grande que α et une pénalité du type heuristique de Mallows. Le p.p.e. est donc minimax et même adaptatif au sens du minimax sur tous les α plus petits que r ainsi qu'en le rayon de la boule. On peut d'ailleurs voir que les puissances des autres paramètres du problème présents dans la minoration (comme le rayon de la boule) sont aussi atteintes quand on majore le risque du p.p.e.. On peut s'amuser à regarder des boules de Besov avec distorsion de normes et conserver pour certaines stratégies bien spécifiques encore l'adaptativité. Tous ces détails sont expliqués dans le Chapitre 1.

Pour les Poissons temporels toujours, on peut regarder le risque minimax sur des ensembles du type "s a D coefficients non nuls parmi les n premiers" dans son développement en ondelettes. Le risque minimax est alors en $D\log(n/D)/\mu(\mathbb{X})$. Le seul moyen pour que le p.p.e. atteigne ce risque minimax est de prendre la famille exhaustive de tous les sous-ensembles de cardinal D de la base d'ondelettes tronquée aux n premières fonctions. La famille n'est pas polynômiale : on prend la pénalité correspondantes avec les poids de la forme $L_m = \log(n/D)$. L'inégalité de type oracle qu'on avait nous permet alors de montrer que le p.p.e. est minimax à constante près. En particulier, on ne peut avoir une vraie inégalité d'oracle, qui donnerait une borne sur le risque inférieure au risque minimax. La présence des poids L_m en Poisson est donc absolument nécessaire.

Pour les processus de comptage à intensité multiplicative d'Aalen, on calcule dans le Chapitre 3 un risque minimax pour des ensembles du type boules d'espaces α-höldériens. Là aussi, comme les histogrammes ont de bonnes capacités d'approximation pour les espaces höldériens, on peut montrer que le p.p.e. en histogramme est adaptativement minimax pour α plus petit que 1.

Par contre pour la stratégie de modèles prévisibles, comme les modèles sont aléatoires, rien n'est montré, sauf quand Y est constant, c'est-à-dire quand on se ramène au cas Poisson. On voit alors que cette stratégie nous fait perdre un facteur logarithmique non seulement dans l'inégalité d'oracle mais aussi dans le risque de l'estimateur sur des boules d'espace de Besov pour la norme 2. Au vu des simulations, les p.p.e. proposés semblent cependant bien fonctionner pour les processus issus de la censure à droite.

Plan

Le premier Chapitre a fait l'objet d'une prépublication et a été soumis à "Probability Theory and Related Fields". Il traite le cas des processus de Poisson.

Le Chapitre 2 se consacre aux inégalités de concentration et leur lien avec les martingales.

Le Chapitre 3 traite de l'estimation de l'intensité multiplicative d'Aalen par sélection de modèles.

La première annexe est un lemme combinatoire utile dans les minorations de risque minimax et qui est prouvé de manière plus simple que dans [**14**].

La dernière annexe présente les programmes (en Scilab 2.6) qui ont permis de faire les simulations du chapitre 3.

CHAPITRE 1

Concentration inequalities for inhomogeneous Poisson processes and adaptive estimation of the intensity

Abstract

In this chapter, we prove new concentration inequalities for suprema of integral functionals of Poisson processes which are analogous to Talagrand's inequalities for empirical processes. These inequalities are used as crucial tools to 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.

AMS Classification : 60E15, 62G05, 62G07.
Keywords : Inhomogeneous Poisson process, concentration inequalities, model selection, penalized projection estimator, adaptive estimation.

Résumé

Dans ce chapitre, nous démontrons des inégalités de concentration pour les suprema d'intégrales Poissoniennes. Elles sont analogues aux inégalités de concentration pour les processus empiriques dues à M. Talagrand. Ces inégalités sont essentielles pour établir des inégalités d'oracle pour des estimateurs par projection pénalisés de l'intensité d'un processus de Poisson inhomogène. Par la suite, nous étudions les propriétés adaptatives des ces estimateurs. D'abord nous donnons des bornes inférieures pour le risque minimax sur différents ensembles de fonctions régulières. Puis nous montrons que les estimateurs par projection pénalisés atteignent ces bornes à constante près.

Classification AMS : 60E15, 62G05, 62G07.
Mots clefs : Processus de Poisson inhomogènes, inégalités de concentration, sélection de modèles, estimateur par projection pénalisé, estimation adaptative.

1.1. Introduction

We consider 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 [**40**]). 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 process.

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 [**59**] studied in density framework and in Poisson framework histogram and kernel estimators. The kernel estimators for the intensity were also studied by Y.A. Kutoyants [**43**] : in his framework, the observation is some n-sample of Poisson processes. In analogy to A.R. Barron and C.-H. Sheu [**8**], W.-C. Kim and J.-Y. Koo [**39**] studied also maximum likelihood type estimators on sieve for exponential family of wavelets.

The choice of the window in [**43**] or the choice of the sieve in [**39**] 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 [**59**] 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 [**12**] 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 2. *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 $\nu : \mathcal{X} \to \mathbb{R}_+$ is a measure without atom (see [**40**]) and is called the "mean measure"of N. This measure is supposed here to be finite to obtain almost surely a finite set of points for N. We denote by dN the random discrete measure $\sum_{T \in N} \delta_T$.

Definition 3. *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 to be finite.

We are interested in estimating s knowing the almost surely finite set of points, $N(\omega)$. At first, let us introduce the projection estimator of s on S, finite dimensional subspace of $\mathbb{L}^2(\mu/\mu(\mathbb{X}))$ with orthonormal basis $\{\varphi_1, \ldots, \varphi_D\}$:

$$\hat{s} = \sum_{i=1}^{D} \left(\int_{\mathbb{X}} \varphi_i(x) \frac{dN_x}{\mu(\mathbb{X})} \right) \varphi_i. \tag{1.1.1}$$

We denote for all i,

$$\hat{\beta}_i = \int_{\mathbb{X}} \varphi_i(x) \frac{dN_x}{\mu(\mathbb{X})}. \tag{1.1.2}$$

This definition 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 :

$$\gamma_{\mathbb{X}}(f) = -\frac{2}{\mu(\mathbb{X})} \int_{\mathbb{X}} f(x) dN_x + \int_{\mathbb{X}} f^2(x) \frac{d\mu_x}{\mu(\mathbb{X})}. \tag{1.1.3}$$

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 :

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

It resembles M. Rudemo's ones, except that in his case the normalization by $\mu(I)$ is replaced 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 $\mathbb{L}^2(\mu/\mu(\mathbb{X}))$. For each model, we denote by $\hat{s}_m$ the projection estimator of s on S_m. At last, we select among $\{\hat{s}_m, m \in \mathcal{M}_{\mathbb{X}}\}$ a good estimator through a data driven criterion which has the following form :

$$\begin{aligned} \hat{m} &= \arg\min_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ -\|\hat{s}_m\|^2 + \text{pen}(m) \right\} \\ &= \arg\min_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ \gamma_{\mathbb{X}}(\hat{s}_m) + \text{pen}(m) \right\}. \end{aligned} \tag{1.1.4}$$

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 $\mathbb{L}^2(\mu/\mu(\mathbb{X}))$. 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. 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 **[27]** and **[38]**). 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 **[27]** proposed a hard threshold : he uses the fact that the Anscombe's expression **[3]** is asymptotically Gaussian in the Poisson parameter and he uses the level of thresholding deriving from the white noise framework. E. Kolaczyk **[41]** noticed that this threshold is not accurate enough in general, because the tails of $\hat{\beta}_\lambda - \mathbb{E}(\hat{\beta}_\lambda)$'s are heavier than tails in the white noise framework and depend on the intensity s. He proposed an other 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 **[22]**. 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, and 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 :

$$\mathbb{E}(\|s - \hat{s}_m\|^2) = \|s - s_m\|^2 + \mathbb{E}(\chi_m^2) \tag{1.1.5}$$

where $\|t\|^2 = \int_{\mathbb{X}} t^2 d\mu/\mu(\mathbb{X})$ (NB : $\mathbb{L}^2$ will always denote $\mathbb{L}^2(\mu/\mu(\mathbb{X}))$ in the statistical applications of this paper) and where

$$\chi^2_m = \sum_{\lambda \in \mathcal{B}_m} \left(\int_{\mathbb{X}} \varphi_\lambda(x) \frac{dN_x - s(x)d\mu_x}{\mu(\mathbb{X})} \right)^2 . \tag{1.1.6}$$

The first term in Equation (1.1.5) is called the **bias term** and the second one is called **variance term**. This last term is equal to

$$\mathbb{E}(\chi^2_m) = \sum_{\lambda \in \mathcal{B}_m} \int_{\mathbb{X}} \varphi^2_\lambda(x) \frac{s(x)d\mu_x}{\mu(\mathbb{X})^2}. \tag{1.1.7}$$

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 :

$$\bar{m} = \arg \min_{m \in \mathcal{M}_{\mathbb{X}}} \mathbb{E}(\|s - \hat{s}_m\|^2). \tag{1.1.8}$$

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 [**47**] 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 [**59**] and M.M Brooks and J.S. Marron [**19**] for kernel estimators. More precisely, we can interpret $\bar{m}$ by :

$$\begin{aligned} \bar{m} &= \operatorname*{argmin}_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ -\|s_m\|^2 + \mathbb{E}(\|\hat{s}_m - s_m\|^2) \right\} \\ &= \operatorname*{argmin}_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ -\mathbb{E}(\|\hat{s}_m\|^2) + \mathbf{2}\ \mathbb{E}(\|\hat{s}_m - s_m\|^2) \right\} \\ &= \operatorname*{argmin}_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ \mathbb{E}(\gamma_{\mathbb{X}}(\hat{s}_m)) + \mathbf{2}\ \mathbb{E}(\|\hat{s}_m - s_m\|^2) \right\} . \end{aligned} \tag{1.1.9}$$

Hence the data driven criterion is of the form

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ \gamma_{\mathbb{X}}(\hat{s}_m) + 2 \int_{\mathbb{X}} \sum_{\lambda \in m} \varphi^2_\lambda(x) \frac{dN_x}{\mu^2(\mathbb{X})} \right\} . \tag{1.1.10}$$

It is a penalized model selection criterion with

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

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. :

$$\mathbb{E}(\|s-\tilde{s}\|^2) \leq C_{\mathbb{X}} \inf_{m\in\mathcal{M}_{\mathbb{X}}} \mathbb{E}(\|s-\hat{s}_m\|^2) \tag{1.1.11}$$

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 [**55**].

These oracle inequalities imply adaptation properties in the minimax sense for the p.p.e., when the family of models and the penalty are well chosen. 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 were known. These results are analogous of those of L. Cavalier and J.Y. Koo [**22**]. In order to prove the adaptation properties of the p.p.e. 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 [**43**] 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 $n\mu$. Hence, we consider 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

$$\chi_m = \sup_{\|a\|_2\leq 1} \int_{\mathbb{X}} \sum_{\lambda\in\mathcal{B}_m} a_\lambda \varphi_\lambda \frac{dN_x - s(x)d\mu_x}{\mu(\mathbb{X})}. \tag{1.1.12}$$

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

A concentration inequality can be written in the following form :

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

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 for Z a 1-Lipschitz function of a Gaussian vector and $f(u) = \sqrt{2u}$ (see **[23]**).

M. Talagrand (see **[62]**) 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 $d\mathbb{P} = sd\mu$ 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 **[50]**.

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 **[64]** recently proves analogous results for $Z = f(N)$ where f is a 1-Lipschitz function, in some sense, of the Poisson process. 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 **[15]**. Very general results about concentration inequalities for infinitely divisible vectors were also proved by C. Houdré **[35]**. The results of L. Wu and C. Houdré 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 $\sup_{a \in A} \int_{\mathbb{X}} \psi_a^2(x) d\nu_x$: 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 Cirel'son concentration inequality 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

[**11**], [**12**], [**14**]). G. Castellan used concentration inequalities in the density framework with maximum likelihood estimator (see [**20**]). Y. Baraud used it too in the regression framework (see [**7**]). Concentration inequalities can also be used in classification (see [**51**]).

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

1.2. Concentration inequalities for Poisson processes

1.2.1. 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 (that is the second point of Definition 2). The second one is that N is infinitely divisible, which means that it can be written as follows for all integer n :

$$dN = \sum_{i=1}^{n} dN_i \tag{1.2.1}$$

the N_i's being mutually independent Poisson processes on $(\mathbb{X}, \mathcal{X})$ with mean measure ν/n. The first property of Definition 2 leads to the following proposition :

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

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

A proof of this proposition can be found in the book [**40**]. 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 :*

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

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

$$\forall u > 0,\quad \mathbb{P}\left(\int_{\mathbb{X}} f(x)(dN_x - d\nu_x) \geq \sqrt{2u\int_{\mathbb{X}} f^2(x)d\nu_x} + \frac{1}{3}\|f\|_\infty u\right) \leq \exp(-u) \tag{1.2.2}$$

and also

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

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

Proof. Using Markov inequality and Campbell's Theorem 6, we get the following inequality, for all $\lambda, x > 0$:

$$\mathbb{P}\left(\int_{\mathbb{X}} f(x)(dN_x - d\nu_x) \geq \xi\right) \leq \exp\left(-\lambda\xi + \int_{\mathbb{X}} \left(e^{\lambda f(x)} - \lambda f(x) - 1\right)\ d\nu_x\right).$$

Since $(e^x - x - 1)/x^2$ is nondecreasing on $\mathbb{R}$, we have for all $\lambda, x > 0$

$$\mathbb{P}\left(\int_{\mathbb{X}} f(x)(dN_x - d\nu_x) \geq \xi\right) \leq \exp\left(-\lambda\xi + \frac{e^{\lambda\|f\|_\infty} - \lambda\|f\|_\infty - 1}{\|f\|_\infty^2}\int_{\mathbb{X}} f^2(x)d\nu_x\right).$$

Taking the minimum in λ, we obtain for all $\lambda, x > 0$

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

where for all $x > 0$, $h(x) = (1+x)\log(1+x) - x$.
Let $\forall u > 0,\ g(u) = \sqrt{2u\int_{\mathbb{X}} f^2(x)d\nu_x} + \frac{1}{3}\|f\|_\infty u$. The following inequality is true : for all $x > 0$, $h(x) \geq g^{-1}(x) \geq x^2(2 + (2/3)x)^{-1}$ where g^{-1} denotes the reciprocal function of g. Hence we obtain exactly what we want to prove for the first part.
Applying all these inequalities to $-f$ leads to the second part. ■

This inequality is apparently well known and holds for more general functionals than the integrals (C. Houdré, private communication). Since we would not find (1.2.2) in the literature, we have presented here a complete proof of it (which is by the way short and simple).

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 (1.1.6) and (1.1.12)) which we mentioned in the introduction.

1.2.2. Entropy and tensorisation. Such controls are based on a fundamental property : the tensorisation of the entropy for product spaces due to M. Ledoux **[45]**. Recapturing Ledoux's method, P. Massart **[50]** 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 $\lambda \in I$

$$\lambda \mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z}) \log \mathbb{E}(e^{\lambda Z}) \leq \sum_{i=1}^n \mathbb{E}\left(e^{\lambda Z} \phi(-\lambda(Z - Z^i))\right). \tag{1.2.3}$$

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

THEOREM 1. *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

$$\log \mathbb{E}(e^{\lambda(Z - \mathbb{E}(Z))}) \leq \mathbb{E}(Z)\phi(\lambda) \tag{1.2.4}$$

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

$$\mathbb{P}(Z \geq \mathbb{E}(Z) + \xi) \leq \exp\Big(-\mathbb{E}(Z) h\big(\frac{\xi}{\mathbb{E}(Z)}\big)\Big) \tag{1.2.5}$$

and

$$\mathbb{P}(-Z \geq -\mathbb{E}(Z) + \xi) \leq \exp\Big(-\mathbb{E}(Z) h\big(\frac{-\xi}{\mathbb{E}(Z)}\big)\Big) \tag{1.2.6}$$

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 (1.1.12)), 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 **[50]**).

1.2.4. 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 **[50]**).

THEOREM 2. *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 (1.2.2). If we apply the previous theorem with only one element in A, we obtain Equation (1.2.2) up to some multiplicative constants (reasonably large).

Let us 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 **[50]** and which are proved in **[46]**. More precisely, one has (see **[50]**) :

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 $\delta > 0$*

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

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 can remark that by Jensen's inequality

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

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

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

So we obtain

$$\frac{\delta}{1+\delta}\mathbb{E}(V_1|N_{\mathbb{X}} = n) - \delta\frac{n}{\nu(\mathbb{X})}\sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2(x)d\nu_x\right) \leq \mathbb{E}_{(T_i),(T_i')}\left(\sup_{a\in A}\sum_{i=1}^{n}(\psi_a(T_i) - \psi_a(T_i'))^2\right). \tag{1.2.7}$$

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 (1.2.7) becomes

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

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

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

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

We can now update Theorem 2 :

COROLLARY 1. *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 :

$$P\big(Z \geq (1+\varepsilon)E(Z) + \sqrt{2\kappa v_0 u} + \kappa(\varepsilon)bu\big) \leq \exp(-u), \tag{1.2.8}$$

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

Proof. We apply Theorem 2 and Proposition 8. We use the additivity of the square root and the following trick

$$\forall a, b, \theta > 0, \quad 2ab \leq \theta a^2 + b^2/\theta. \tag{1.2.9}$$

Optimizing in δ leads to the result. ■

The later result is the easiest to use for the statistical applications, that are developed in Section 1.3. Comparing (1.2.8) with Cirel'son formula **[23]**, 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 1 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. 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 [**45**] since the variables are unbounded.

Note that C. Houdré [**35**] and L. Wu [**64**] 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 1 since their variance term in this situation are not bounded by v_0 like ours.

1.3. Model selection with projection estimators

Let us now give a statistical application of Corollary 1. 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), (1.1.3) and (1.1.4).

We want to prove in this section several oracle type inequalities of type (1.1.11), depending on the penalty and on the family of models. We shall in particular see that when the complexity of the family of models is not too large, the heuristic for choosing the penalty described in the introduction (see (1.1.10)) can be justified. For too rich families of models, we should take other forms of penalties. For example, we can look at the subset selection case described in the introduction : in this case, we want to choose in fact the coefficients to be estimated in the development of s on an orthonormal basis. If we choose $\mathcal{M}_{\mathbb{X}} = \{m, m \subset \Lambda\}$, i.e. a complete subsets selection, there is then exponentially many models with the same dimension : this is representative of a too rich family and in this case, we have to take a penalty larger than the penalty of the heuristic to obtain a good risk. Moreover, $C_{\mathbb{X}}$ (see Equation (1.1.11)) is not an absolute constant.

We have consequently to compute the risk of the p.p.e. to give proper forms of penalties. We need some easy computations to make appear a term which resembles a chi-square, χ_m (see (1.1.6)).
The definitions of $\tilde{s}$ and of $\gamma_{\mathbb{X}}$ (see (1.1.4) and (1.1.3)) lead, for all m in $\mathcal{M}_{\mathbb{X}}$, to :

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

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

$$\forall t \in \mathbb{L}^2, \nu_{\mathbb{X}}(t) = \int_{\mathbb{X}} t(x) \quad \frac{dN_x - s(x)d\mu_x}{\mu(\mathbb{X})}, \tag{1.3.1}$$

we have that the contrast $\gamma_{\mathbb{X}}$ defined in (1.1.3) verifies

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

Let us recall that $\mathbb{L}^2$ means here $\mathbb{L}^2(\mu/\mu(\mathbb{X}))$ and $\|.\|$ denotes the associated norm. We get consequently, for all m in $\mathcal{M}_{\mathbb{X}}$:

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

We can rewrite this :

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

We see by (1.1.12) that

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

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

$$\|s-\tilde{s}\|^2 \leq \|s-s_m\|^2 + 2\chi^2_{\hat{m}} + 2\nu_{\mathbb{X}}(s_{\hat{m}}-s_m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m). \tag{1.3.2}$$

In order to derive from (1.3.2), some oracle inequality, we see that $\mathrm{pen}(\hat{m})$ should be of the order of $\chi^2_{\hat{m}}$ 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 $\chi^2_{\hat{m}}$. 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.
The reason 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 2 of N implies that χ^2_m is a sum of independent centered quantities.

1.3.1. A linear case of concentration inequalities. As we see in (1.1.12), χ_m is a supremum of integral functionals : therefore we can use Corollary 1. If we apply Corollary 1 brutally then we set Inequality (1.3.3) 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 (1.3.4) is somehow more subtle and will replace (1.3.3) in situations where (1.3.3) 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 (1.1.1)).*
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 $B_S = \sup_{f\in S,\|f\|=1}\|f\|_\infty$. All these quantities are supposed to be finite.
Then for all ε and u positive :

$$\mathbb{P}\left(\chi(S) \geq (1+\varepsilon)\sqrt{\mathbb{E}(\chi^2(S))} + \sqrt{\frac{2\kappa M_S u}{\mu(\mathbb{X})}} + \kappa(\varepsilon)\frac{B_S}{\mu(\mathbb{X})}u\right) \leq \exp(-u) \tag{1.3.3}$$

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

$$\mathbb{P}\left(\chi(S)\mathbb{1}_{\Omega_S(\varepsilon)} \geq (1+\varepsilon)\left(\sqrt{\mathbb{E}(\chi^2(S))} + \sqrt{\frac{2\kappa M u}{\mu(\mathbb{X})}}\right)\right) \leq \exp(-u). \tag{1.3.4}$$

where κ and $\kappa(\varepsilon)$ are given in Corollary 1.

We can remark that in the first point 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 **[23]**). 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 **[48]** and can be found in the PhD Thesis of G. Castellan **[20]**, who have used it in the context of density estimation from a n-sample.

1.3.2. Model selection for a polynomial collection of models. We have to bound the risk of the p.p.e. For this aim, we have to distinguish two cases. The first result deals with a not too large family of models : more precisely, it deals with polynomial collection, in the following sense.

Definition 4. *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 3. *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 (1.1.4)).*
Assume that :

(1) *$\mathcal{M}_\mathbb{X}$ is a polynomial collection (cf Definition 4) 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 $\inf_{m\in\mathcal{M}_{\mathbb{X}}} \frac{\mathbb{E}(\hat{V}_m)}{D_m} = \beta > 0$, *for all* m *in* $\mathcal{M}_{\mathbb{X}}$ *:* $\mathrm{pen}(m) \geq c\frac{\hat{V}_m}{\mu(\mathbb{X})}$
with $\hat{V}_m = \int_{\mathbb{X}} \sum_{\lambda\in\mathcal{B}_m} \varphi_\lambda^2 \frac{dN_x}{\mu(\mathbb{X})}$ *where* $\{\varphi_\lambda, \lambda\in\mathcal{B}_m\}$ *is an orthonormal basis of* S_m,

(c) or for all m *in* $\mathcal{M}_{\mathbb{X}}$ *:* $\mathrm{pen}(m) \geq \frac{c(\hat{V}_m + \alpha(N_{\mathbb{X}}/\mu(\mathbb{X}))\mathbb{D}_m)}{\mu(\mathbb{X})}$ *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 $c, \Gamma, R, \|s\|, \|s\|_\infty, \rho$ (and β in case (b) or α in case (c)).

Remark :

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 that the first penalty in (a) verify $\mathbb{E}(\mathrm{pen}(m))$ slightly larger than the variance term in the quadratic risk of $\hat{s}_m$ (see Equation (1.1.7)).

As $\mathbb{D}_m \geq D_m$, the term $\mathbb{D}_m$ looks like the dimension of the model : this has to be compared with Mallows criterion **[47]**. (There is a simple way to compute $\mathbb{D}_m$: whenever the orthonormal basis of S_m, $\{\varphi_\lambda, \lambda\in\mathcal{B}_m\}$, $\mathbb{D}_m = \|\sum_{\lambda\in\mathcal{B}_m}\varphi_\lambda^2\|_\infty$.)

Furthermore in this polynomial case, we obtain exactly an oracle inequality (see (1.1.11) 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 the penalty cannot be smaller than the dimension of the model which leads to the improvement mentioned above.

We can remark too that we do not make any assumption here on the link between each model, 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 : $S_m = \text{Span}\{\varphi_\lambda, \lambda \in m\}$ for all m in $\mathcal{M}_\mathbb{X}$; $|m|$ will denote the cardinality of m. 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 applications of that 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 : $\text{pen}(m_k) = 2\hat{V}_{m_k} + \alpha(2k+1)(N_\mathbb{X}/\mu(\mathbb{X})^2)$. Hence the p.p.e. is a Fourier truncated sum which will have the previous upper bound on the risk.

– 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 $\mathbb{D}_{m_k} \leq (k+1)^2 = D^2_{m_k}$ 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 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 bounded like in Theorem 3 for additive models.

- We can choose also wavelets basis with compact support. As we will see Section 1.5, we can consequently construct a p.p.e. which verifies the assumptions of Theorem 3 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. As we know that all bounded measurable functions can be approximated by piecewise constant functions, we can imagine estimators which will be piecewise constant functions, i.e. histograms (see for instance [**59**]). Hence we can imagine $\mathcal{M}_{\mathbb{X}}$ as a collection of partitions of $\mathbb{X}$ and a model S_m, for m in $\mathcal{M}_{\mathbb{X}}$, will be 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 condition on μ_m is given in the density framework [**20**]. 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 $\mathbb{E}(N_{\mathbb{X}})$, i.e. the expected number of observed points. For all the choices of penalty given in Theorem 3, we get a p.p.e with bounded quadratic risk as in Theorem 3.

1.3.3. 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 seems to be very abstract and that is the reason why I would rather to give first 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 assumptions on the complexity of the family, we have to make assumptions on the largest family of coefficients Λ.

DEFINITION 5. *$\{\varphi_\lambda, \lambda \in \Lambda\}$ 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 bases with finite support verify such a property with a constant B independent of Λ (see Section 1.4).

PROPOSITION 10. *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 (1.1.1)). For a given penalty* pen *on $\mathcal{M}_\mathbb{X}$, let $\tilde{s}$ be the associated penalized projection estimator (see (1.1.4)).*
Assume that :

(1) *the family $\{\varphi_\lambda, \lambda \in \Lambda\}$ is localized (cf Definition 5), with constant B independent of $\mathbb{X}$,*
(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 $\mathbb{X}$,*
(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 $\mathrm{pen}(m) = \dfrac{cM'|m|}{\mu(\mathbb{X})}\left(1 + \sqrt{2\kappa L_m}\right)^2$ *with c larger than 1,*
  (b) or if (random penalty) $\mathrm{pen}(m) = \dfrac{c}{\mu(\mathbb{X})}\left(\sqrt{\hat{V}_m} + \sqrt{2\kappa M' L_m|m|}\right)^2$ *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,

– *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.
Remark : κ is defined in Corollary 1.

Here, we loose an important thing with respect to Theorem 3 : 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 1.5), but we do not know when to say : the assumption is true or not for fix 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 $\mathcal{M}_{\mathbb{X}} = \{m \subset \Lambda = \{1,...,N\}\}$, it is not polynomial : we 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 1.5, the penalized estimator $\tilde{s}$ of this theorem is for this problem minimax.

Furthermore, we have a kind of oracle inequality (see (1.1.11)) with $C_{\mathbb{X}}$ depending effectively on $\mathbb{X}$ if the L_m are not constant. We obtain in this case a term larger than the variance term. As we shall see in the following 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 [**12**]. This phenomenon is called "heterosedasticity". In this framework, it may disappear if one chooses an other contrast, for example log-likelihood, as G. Castellan proved it in [**20**]. 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, heterosedasticity is easier to handle. Let us give precisely the framework and the results.

PROPOSITION 11. *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 (1.1.4)).*
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* $\mathbb{X}$,

(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* $M = \sup_{I\in\Gamma}\left(\int_I sd\mu/\mu(I)\right)$.

In this case, there is no more heterosedasticity problem. We estimate "$\|s\|_\infty$" in some sense by $\tilde{M}$ which depends only on the data. Since the adaptive properties of histograms are not enough accurate, we are going to focus on subsets selection case with wavelets basis to understand the performances of our estimator (see Section 1.5).
In fact Propositions 10 and 11 are consequences of the following general result.

A general model selection theorem.

THEOREM 4. *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 (1.1.4)).*
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* $\sup_{f\in S_\Lambda/\|f\|=1}\int_{\mathbb{X}} f^2 sd\mu/\mu(\mathbb{X})$. *For all* ε *positive, we 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 1) 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* $\mathcal{M}_{\mathbb{X}}$*,*

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

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

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

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 deals mostly with the link between two models. In the applications of this theorem (Propositions 10 and 11), these assumptions on the link between the models are consequences of the form of the bases of each model (subsets selection case plus localization or histograms selection case).

We are now focusing on the subsets selection case to understand the performances of the p.p.e. in terms of minimax adaptivity properties.

1.4. 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 6. *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 intensities* $\mathbb{L}^2$*.*

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 to the question : does

the p.p.e. estimate as well as the best one, which knows that s is $\mathcal{S}$, even if our estimator does not know this fact ?
Our aim in this part is to compute lower bounds for the minimax risk on some family of possible functions for s. In order to compute lower bound on the minimax risk, there exists some recent interesting result **[10]** 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 3. *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.
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

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

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

LEMMA 4. *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 **[22]**.
Now, we have to compute lower bounds for minimax risk on some proper $\mathcal{S}$.

1.4.1. Minimax risk on ellipsoids. We keep the notations 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

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

We can find a lower bound for the minimax risk on this ellipsoid, using Lemma 3.

PROPOSITION 12. *Assume that there exists an integer $D > L$ such that $\{\varphi_\lambda, \lambda \in \{1, ..., D\}\}$ is localized with constant B (see Definition 5). 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, $L = 2$.

Proof. Let us recall a combinatorial lemma, which can be found in **[30]** :

LEMMA 5. *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 $\hat{s}'$ a minimizer of the distance in $\mathcal{C}_D$. 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 for all m and m' of $\mathcal{C}_D$ by Lemma 5

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

we have the following lower bound

$$\begin{aligned} R(\mathcal{E}(c,\rho)) &\geq \frac{\theta(D-L+1)a_D^2}{4} \inf_{\hat{s} \in \mathcal{C}_D} \sup_{s \in \mathcal{C}_D} \mathbb{P}_s(\hat{s} \neq s) \\ &\geq \frac{\theta(D-L+1)a_D^2}{4} \inf_{\hat{s} \in \mathcal{C}_D} (1 - \inf_{s \in \mathcal{C}_D} \mathbb{P}_s(\hat{s} = s)). \end{aligned} \tag{1.4.1}$$

We are going to use Lemma 3. Hence, we have to compute $\bar{K}$ and for this aim, we use Lemma 4.

$$\begin{aligned}
\forall m' \neq m \in \mathcal{M}_D, \quad K(\mathbb{P}_{t_{m'}}, \mathbb{P}_{t_m}) &= \int t_{m'} \phi(\log \frac{t_m}{t_{m'}}) d\mu_x \\
&= \int [t_m - t_{m'} - t_{m'} \log(1 + \frac{t_m - t_{m'}}{t_{m'}})] d\mu_x \\
&\leq \int \frac{(t_m - t_{m'})^2}{t_m}(x) d\mu_x \\
&\leq \frac{2}{\rho} \mu(\mathbb{X}) \|t_{m'} - t_m\|^2 \qquad (1.4.2) \\
&\leq \frac{2\mu(\mathbb{X})(D - L + 1)a_D^2}{\rho}
\end{aligned}$$

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

$$\bar{K} \leq \frac{2\mu(\mathbb{X})(D - L + 1)a_D^2}{\rho} \qquad (1.4.3)$$

Lemma 3 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 5. The result follows with $\zeta = \alpha\sigma/2$ and $\eta = (1 - \alpha)\theta/4$. ■

1.4.2. Logarithmic factors in the risk. We always keep the notations 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 n, D be two positive integers. Let $\mathcal{S}_{n,D}$ 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 n first coefficients. Let $B_{n,D,\rho}$ be the following set :

$$B_{n,D,\rho} = \left\{ t = \rho + u \Big/ \int_{\mathbb{X}} u = 0, u \in \mathcal{S}_{n,D}, t \geq 0 \right\}. \qquad (1.4.4)$$

Using Lemma 3, we obtain the following proposition :

PROPOSITION 13. *Let $n > L$. Assume that the family $\{\varphi_\lambda, \lambda \in \{1, ..., n\}\}$ is localized (cf Definition 5) 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.

Proof. We use a lemma which is due to L. Birgé and P. Massart **[13]**. Their proof being rather intricate, we have decided to present a complete and simple proof in the appendix (although our constants are slightly worse than theirs). We deduce from this lemma (Lemma 14, see the appendix) 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 $|m \triangle m'| \geq \theta' D$, 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 (1.4.1)) :

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

Using Lemma 3 and the control of the Kullback-Leibler information (1.4.2), we obtain that if

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

which is implied by

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

then

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

Choosing

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

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

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

1.4.3.1. *Wavelet expansions.* We are dealing with wavelet basis on a segment (and not on $\mathbb{R}$). The most known example is the Haar basis. When we want to look at smooth wavelets, we can deal with the one constructed by A. Cohen, I. Daubechies and P. Vial **[24]**. 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 as 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

$$t(x) = \sum_{\lambda \in \mathbb{N}^*} a_\lambda \varphi_\lambda(x) = \sum_{j \geq l} \sum_{k \in \Lambda(j)} a_{j,k} \varphi_{j,k}(x). \tag{1.4.5}$$

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

$$\mathcal{F}_J = \{\varphi_{j,k}, k \in \Lambda(j), l \leq j \leq J\} \tag{1.4.6}$$

is localized in the sense of Definition 5 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 of compact support and have r continuous derivatives. It is possible to get r large enough, by selecting $K(r)$ large enough. Such wavelet basis exists (see **[24]**).

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 **[25]** for a definition). It can be described with wavelets (see **[28**, Theorem 2]) : the consequence of this theorem is that we can say for all wavelet with regularity $r > \alpha$ that

$$B^{\alpha}_{p,p'} = \left\{ t \in \mathbb{L}^2[0,T], 2^{j(\alpha+\frac{1}{2}-\frac{1}{p})} \|a_{j,.}\|_p \in l^{p'}(\mathbb{N}) \right\} \tag{1.4.7}$$

where $a_{j,k}$ are the coefficients defined in (1.4.5). The associated norm of this space can be taken as follows :

$$\begin{aligned} \forall p' < +\infty, \quad \|t\|^{\alpha}_{p,p'} &= \left(\sum_{j \geq 0} 2^{jp'(\alpha+\frac{1}{2}-\frac{1}{p})} \|a_{j,.}\|_p^{p'} \right)^{1/p'}, \\ p' = +\infty, \quad \|t\|^{\alpha}_{p,\infty} &= \sup_{j \in \mathbb{N}} \left(2^{j(\alpha+\frac{1}{2}-\frac{1}{p})} \|a_{j,.}\|_p \right). \end{aligned} \tag{1.4.8}$$

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

$$\|t\|^{\alpha}_{2,2} \geq \|t\|^{\alpha}_{2,\infty} \geq \|t\|^{\alpha}_{p,\infty} \tag{1.4.9}$$

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

1.4.3.2. *Minimax risk for $B^{\alpha}_{2,2}$ balls.* We keep the previous notations. With Proposition 12, 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

$$\mathcal{B}(\rho, R, B^{\alpha}_{2,2}) = \left\{ t = \rho + u \,\middle/\, t \geq 0,\ \int_{\mathbb{X}} u\ dx = 0,\ u \in B^{\alpha}_{2,2},\ \|u\|^{\alpha}_{2,2} \leq R \right\}, \tag{1.4.10}$$

where the Besov norm is defined in (1.4.8).

Proposition 14. *Then 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, B some positive constants depending only on the wavelet basis.

Proof. We consider the basis with regularity $r > \alpha$ described previously. We want to apply Proposition 12. We use the wavelet basis defined in (1.4.7). 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 (1.4.6)). 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 12 leads to

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

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

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

and we obtain the result remarking that

$$\frac{2^{J+1} - 2^{l+1}}{2^{J+1} - 2^l} \geq \frac{2}{3}.$$

■

Letting T go to infinity we easily derive from Proposition 14 the following asymptotic lower bound.

Corollary 2.

$$\liminf_{T \to +\infty} \left(T^{\frac{2\alpha}{2\alpha+1}} R(\mathcal{B}(\rho, R, B^{\alpha}_{2,2})) \right) \geq C \rho^{\frac{2\alpha}{2\alpha+1}} R^{\frac{2}{2\alpha+1}} \tag{1.4.11}$$

with C positive constant depending on the wavelet basis.

Note that the some related asymptotic lower bound (with some explicit value of C) has already be obtained by Y.A. Kutoyants [**43**] for a n-sample of Poisson processes.

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

We want to analyze the performances of our penalized estimators in terms of minimax risk on various sets. In particular we shall compare the corresponding upper bounds for the minimax risk with the lower bounds computed in the previous section. We are in the case where $\mathbb{X} = [0, T]$ with the Lebesgue measure $d\mu = dx$. We keep the notations of Section 1.4.3.

Let $\{\varphi_{j,k}, j \geq l, k \in \Lambda(j)\}$ be a wavelet basis (see Section 1.4.3.1) with regularity r.

1.5.1. The nested projection strategy. The first strategy is the one defined in Theorem 3 for the subsets selection case. We look at the family $\mathcal{F}_J$ defined in (1.4.6). The models S_h's are defined as follows : for all $h \leq J$,

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

They are nested, hence polynomial in the sense of Definition 4. As we have seen in Section 1.4.3.1, 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 5. A consequence of the classical localized property

(with constant B) of these wavelets is Assumption 2 of Theorem 3 since $B2^J \simeq T$. The penalty is chosen here with formula (a) of Theorem 3 :

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

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

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

Rate of convergence : The lower bound proposed in Corollary 2 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 (1.4.9). Hence (1.4.11) is the bound that 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 compute the lower bound ? We denote by β the coefficients of s in the wavelet expansion. The bound of Theorem 3 makes appear the bias term, bounded as follows :

$$\begin{aligned} \forall l \leq h \leq J, \quad \|s - s_h\|^2 &\leq \sum_{j>h} \sum_{k \in \Lambda(j)} \beta_{j,k}^2 \\ &\leq \left(\|s\|^\alpha_{2,2}\right)^2 \sum_{j>h} 2^{-2j\alpha} \\ &\leq R^2 2^{-2h\alpha}. \end{aligned} \tag{1.5.1}$$

  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 then asymptotically bounded by :

$$C' \rho^{\frac{2\alpha}{2\alpha+1}} R^{\frac{2}{2\alpha+1}} T^{-\frac{2\alpha}{2\alpha+1}} + O(1/T)$$

  with C' positive constant depending on the wavelet basis. So, 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 **[12]** (see the last section therein) :

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

  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. Indeed, always in **[12]**, 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, we get the following upper bound for the risk of our estimator :

$$C'_p R^{\frac{2}{1+2(\alpha+\frac{1}{2}-\frac{1}{p})}} \rho^{\frac{2(\alpha+\frac{1}{2}-\frac{1}{p})}{1+2(\alpha+\frac{1}{2}-\frac{1}{p})}} T^{-\frac{2(\alpha+\frac{1}{2}-\frac{1}{p})}{1+2(\alpha+\frac{1}{2}-\frac{1}{p})}},$$

plus some asymptotically insignificant term due to the rest in the theorem. So, the simple method (using a nested family of models and Theorem 3) does not lead to the minimax risk (the other form of penalty purposed in this theorem leads to the same kind of bound). 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 $p < 2$. In the following subsections, more complex families of models will be considered which have this time good approximation properties in these spaces.

1.5.2. Thresholding. We now turn to a more complex family of models, using Proposition 10. We use once again the family $\mathcal{F}_J$ (see (1.4.6)) 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 10), 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

$$\binom{N}{D} \leq (eN/D)^D. \tag{1.5.2}$$

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 $c > 1$,

$$\text{pen}(m) = \frac{c|m|(\|\hat{s}_\Lambda\|_\infty + K')}{T}(1 + \sqrt{2\kappa L_m})^2 \tag{1.5.3}$$

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

Rate of convergence : If we want to apply Theorem 10(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 10. As previously, we want to make the compromise between the bias term 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, we will get then , the good rate of convergence on the other subsets. A proposition due to L. Birgé and P. Massart (Proposition 6 of [**12**]) 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

$$|m| \leq C2^{j'} \text{ and } \|s - s_m\|^2 \leq C'R^2 \left(2^{-2\alpha j'} + 2^{-2J(\alpha+\frac{1}{2}-\frac{1}{p})}\right). \tag{1.5.4}$$

Among the $m_{j'}$'s, we choose one such that the corresponding 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. We obtain consequently an asymptotic upper bound for the risk of our estimator :

$$C_\alpha R^{\frac{2}{1+2\alpha}} M^{\frac{2\alpha}{1+2\alpha}} \left(\frac{T}{\ln T}\right)^{-\frac{2\alpha}{2\alpha+1}}$$

plus some asymptotically insignificant term. 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.

1.5.3. Adaptive thresholding. Let Λ be the set of indices of the functions of $\mathcal{F}_J$ (see (1.4.6)). 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 10 (c). Since we have Equation (1.5.2), we can choose $L_m = \ln(n/|m|)$. Consequently we look at penalty given in (1.5.3) 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 asymptotically of the order :

$$C_\alpha R^{\frac{2}{1+2\alpha}} M^{\frac{2\alpha}{1+2\alpha}} \left(\frac{T}{\ln T}\right)^{-\frac{2\alpha}{2\alpha+1}}.$$

- Now, let us assume that s lies in $B_{n,D,\rho}$ (see (1.4.4)) for some D positive integer. Assume that $n > 2^l$ and $n > 4D$. Then we apply Proposition 10. 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 $\ln(n - L + 1) \geq \gamma \ln(n)$. The weakness comes from M which replaces ρ as previously. Furthermore it achieves precisely the good rate in n, D and T.

1.5.4. Special Besov strategy. We can improve it with a special strategy due to L.Birgé and P.Massart (see **[12]** and **[14]**). 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 **[12]**. We want always to apply Proposition 10, 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 (1.5.3).
Rate of convergence : With their proposition 6 (which is (1.5.4)), 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 bounded by a constant times

$$R^{\frac{2}{1+2\alpha}} M^{\frac{2\alpha}{1+2\alpha}} T^{-\frac{2\alpha}{2\alpha+1}}$$

asymptotically in T that 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.

1.5.5. Adaptivity. Penalized projection estimators are then adaptive : 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 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 loose of factor ρ which designed the normalized integral of s, replaced by M. Furthermore 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.

1.6. Proofs

1.6.1. Concentration Theorems.
These proofs are based on the scheme of proof of M. Ledoux and P. Massart in the n-sample framework (see **[45]** and **[50]**). The second one is inspired by the scheme of proof of E. Rio in the n-sample framework **[57]**.

1.6.1.1. *Proof of Theorem 1.*

Proof. By monotone convergence, it's sufficient to prove it for a finite family of functions. We give two proofs : the first is true for all kind of space $\mathbb{X}$ and is based on the infinitely divisible property of Poisson process, the second one gives an other proof for $\mathbb{X}$ metrizable such that the balls are precompact.

(1) N can be written as $dN = \sum_{i=1}^{n} dN^i$, as in Equation (1.2.1), with N^i independent Poisson processes with mean measure ν/n, where ν is the mean measure of N. Then we can write

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

We interpret the sigma-field of each N^i as independent random variables and 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

$$\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z}) \leq \mathbb{E}\left[e^{\lambda Z}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\right]. \tag{1.6.1}$$

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 (1.6.1) becomes by Cauchy-Schwarz :

$$\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z}) \leq \mathbb{E}\left[e^{\lambda Z}\mathbb{1}_{\Omega_n}\sum_{i=1}^{n}\phi(-\lambda(Z-Z^i))\right] + \sqrt{\frac{\nu(\mathbb{X})^2}{n}\mathbb{E}\left[e^{2\lambda Z}\Big(\sum_{i=1}^{n}\phi(-\lambda(Z-Z_i))\Big)^2\right]}. \tag{1.6.2}$$

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

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

For the second term in (1.6.2), as

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

if $\lambda>0$ and

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

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

$$\lambda\mathbb{E}(Ze^{\lambda Z})-\mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z})\leq\mathbb{E}\left[e^{\lambda Z}\sum_{T\in N}\phi(-\lambda(Z-Z_T))\right]. \tag{1.6.3}$$

(2) We suppose in this second version that $\mathbb{X}$ is metrizable, with precompact balls, for a certain distance d (typically $\mathbb{X}=\mathbb{R}^d$ works). For all n in $\mathbb{N}$, we cut $\mathbb{X}$ in n boxes, denoted by $A_0^n,A_1^n,...,A_{m_n}^n$, such that A_0^n is the complementary of a ball of center u fixed and radius n, and $A_1^n,...,A_{m_n}^n$ is a partition of this ball such that their diameter tends to 0 when n tends to infinity. To each A_i^n, we associate the set of points of N, $\mathcal{S}_i$, which appear in A_i. So we can interpret

$$Z=\sup_{a\in A}\int_{\mathbb{X}}\psi_a(x)dN_x=\sup_{a\in A}\sum_{i=0}^{m_n}\sum_{T\in\mathcal{S}_i}\psi_a(T).$$

So we can say that Z is a function of n independent variables (or n independent sigma-fields), and we can apply Lemma 1 to variables $\mathcal{S}_1,...\mathcal{S}_n$. Here, the variables $\mathcal{S'}_i$ are the empty sets. So if we note, for each $0\leq i\leq m_n$,

$$Z^i=\sup_{a\in A}\int_{\mathbb{X}\setminus A_i}\psi_a(x)dN_x$$

we obtain that for all λ,

$$\lambda\mathbb{E}(Ze^{\lambda Z})-\mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z})\leq\mathbb{E}\left[e^{\lambda Z}\sum_{i=0}^{m_n}\phi(-\lambda(Z-Z^i))\right]. \tag{1.6.4}$$

Almost everywhere, for n large enough, there is at the most one point by box (and no point in A_0^n), as ν is finite which implies that a.e. there a finite number of points of N. So we have

$$\sum_{i=0}^{m_n}\phi(-\lambda(Z-Z^i))\underset{n\to\infty}{\to}\sum_{T\in N}\phi(-\lambda(Z-Z_T))\text{ a.e. .}$$

On the other side, since $0 \leq Z - Z^i$, we have for positive λ,

$$\phi(-\lambda(Z - Z^i)) \leq \lambda(Z - Z^i) \leq \lambda \int_{A_i} \psi_{\hat{a}} dN,$$

where $\hat{a}$ is the index where Z is achieved. So, we get

$$\exp(\lambda Z) \sum_{i=1}^{n} \phi(-\lambda(Z - Z^i)) \leq \lambda \exp(\lambda Z) N_{\mathbb{X}}^2.$$

If $\lambda < 0$, then the quantity is upper bounded by $\exp(-\lambda Z) N_{\mathbb{X}}$. As $N_{\mathbb{X}}$ and Z have Laplace transform for every $\lambda \in \mathbb{R}$, we can apply a dominated convergence theorem. So Equation (1.6.4) becomes, when n tends to infinity, for every λ, Equation (1.6.3).

Remark : These two proofs are very similar. I have written the second one, because I was asked if such concentration can be proved for processes which have the property of independence by boxes but at a certain distance : I do not know but perhaps the second demonstration can help to answer. Now, we have to finish the proof, with Equation (1.6.3).

The supremum in Z is achieved in $\hat{a}$. Then we have, for all T point of N,

$$0 \leq Z - Z_T \leq \psi_{\hat{a}}(T) \leq 1.$$

Note that if x is in $[0, 1]$, we have $\phi(-\lambda x) \leq \phi(-\lambda)x$ for all $\lambda > 0$. So Equation (1.6.3) becomes, for all λ

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

Then, we only need to follow P. Massart's proof **[50]** and the result follows. ■

1.6.1.2. *Proof of Theorem 2.*

Proof. By monotone convergence, it's 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 (1.2.1) :

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

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

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

We can apply then Lemma 1 to write

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

We split the expectation in two parts :

$$\forall \lambda > 0, \quad \lambda \mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z}) \ln \mathbb{E}(e^{\lambda Z}) \leq \mathbb{E}\left(e^{\lambda Z} \sum_{i=1}^{n} \phi(-\lambda(Z - Z^i)) \mathbb{1}_{Z - Z^i \geq 0} \right) + \mathbb{E}\left(e^{\lambda Z} \sum_{i=1}^{n} \phi(-\lambda(Z - Z^i)) \mathbb{1}_{Z - Z^i < 0} \right). \tag{1.6.5}$$

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

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

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 achieved. It leads us, for all $\lambda > 0$, to :

$$\mathbb{E}\left(e^{\lambda Z} \sum_{i=1}^{n} \phi(-\lambda(Z - Z^i)) \mathbb{1}_{Z - Z^i \geq 0} \right) \leq \frac{\lambda^2}{2} \mathbb{E}\left(e^{\lambda Z} \sum_{i=1}^{n} \left(\int_{\mathbb{X}} \psi_{\hat{a}}(x)(dN_x^i - \frac{1}{n} d\nu_x) \right)_+^2 \right). \tag{1.6.6}$$

As $\mathbb{P}(N_{\mathbb{X}}^i \geq 2) \leq (\nu(\mathbb{X})/n)^2/2$, if we denote by Ω the event $\{\forall i, N_{\mathbb{X}}^i \leq 1\}$, we have that :

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

So we can split the last expectation of Equation (1.6.6) in two parts, and by Cauchy-Schwarz, we obtain :

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

The most of these terms tends to 0, and we can use a dominated convergence theorem (since Poisson law has Laplace transform) to have this upper bound :

$$\limsup_{n\to+\infty} \mathbb{E}\left(e^{\lambda Z}\sum_{i=1}^{n}\left(\int_{\mathbb{X}}\psi_{\hat{a}}(x)(dN_x^i-\frac{1}{n}d\nu_x)\right)_+^2\right)\leq \mathbb{E}\left(e^{\lambda Z}\int_{\mathbb{X}}\psi_{\hat{a}}^2 dN\right) \tag{1.6.7}$$

For the second expectation of Equation (1.6.5), we can 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

$$Z^i-Z\leq \int_{\mathbb{X}} -\psi_{\hat{a}_i}(x)(dN_x^i-\frac{1}{n}d\nu_x) \tag{1.6.8}$$

where $\hat{a}_i$ which denotes the index where the supremum in Z^i is achieved.
NB : This index $\hat{a}_i$ is independent of N^i and knowing the other processes, N^i is still a Poisson process with intensity s/n by independence. Hence we obtain :

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

Otherwise, using again Equation (1.6.8) 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 **[57]**. We obtain consequently that :

$$\mathbb{E}\left(\sum_{i=1}^{n} e^{\lambda Z}\phi(-\lambda(Z-Z^i))\mathbb{1}_{Z-Z^i<0}\right)\leq \sup_{a\in A}\left(\int_{\mathbb{X}}\psi_a^2 d\nu\right)\mathbb{E}(e^{\lambda Z}). \tag{1.6.9}$$

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 (1.6.7) and (1.6.9), that for all λ positive

$$\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z}) \leq \frac{\lambda^2}{2}\mathbb{E}\left(e^{\lambda Z}\left(\int \psi^2_{\hat{a}}(x)dN_x + \sup_{a\in A}\left(\int_{\mathbb{X}} \psi^2_a(x)d\nu_x\right)\right)\right).$$

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

(1.6.10)

$$\lambda\mathbb{E}(Ze^{\lambda Z}) - \mathbb{E}(e^{\lambda Z})\log\mathbb{E}(e^{\lambda Z}) \leq \frac{\lambda^2}{2}\mathbb{E}\left(e^{\lambda Z}\left(\sup_{a\in A}\left[\int \psi^2_a(x)dN_x\right] + \sup_{a\in A}\left[\int \psi^2_a(x)d\nu_x\right]\right)\right).$$

If we set $\tilde{Z} = Z - \mathbb{E}(Z)$, inequality (1.6.10) becomes

(1.6.11)

$$\lambda\mathbb{E}(\tilde{Z}e^{\lambda\tilde{Z}}) - \mathbb{E}(e^{\lambda\tilde{Z}})\log\mathbb{E}(e^{\lambda\tilde{Z}}) \leq \frac{\lambda^2}{2}\mathbb{E}\left(e^{\lambda\tilde{Z}}\left(\sup_{a\in A}\left[\int \psi^2_a(x)dN_x\right] + \sup_{a\in A}\left[\int \psi^2_a(x)d\nu_x\right]\right)\right).$$

We set $V_1 = \sup_{a\in A}\left[\int \psi^2_a(x)dN_x\right]$, $v_0 = \sup_{a\in A}\left[\int \psi^2_a(x)d\nu_x\right]$ and $v_1 = \mathbb{E}(V_1)$.
We can apply the following lemma obtained by P.Massart [**50**].

LEMMA 6. *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*

$$\mathbb{E}(Ve^{\lambda Y}) \leq \mathbb{E}(Ye^{\lambda Y}) + \frac{\log\mathbb{E}(e^{\lambda V})}{\lambda}\mathbb{E}(e^{\lambda Y}). \tag{1.6.12}$$

Now equation (1.6.11) becomes

$$\lambda\mathbb{E}(\tilde{Z}e^{\lambda\tilde{Z}}) - \mathbb{E}(e^{\lambda\tilde{Z}})\log\mathbb{E}(e^{\lambda\tilde{Z}}) \leq \frac{\lambda^2}{2}v_0\mathbb{E}(e^{\lambda\tilde{Z}}) + \frac{\lambda^2}{2}\frac{\log\mathbb{E}(e^{\lambda V_1})}{\lambda}\mathbb{E}(e^{\lambda\tilde{Z}}) + \frac{\lambda^2}{2}\mathbb{E}(\tilde{Z}e^{\lambda\tilde{Z}}).$$

But Theorem 1 allows us to control the Laplace transform of V_1. So

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

Hence we obtain

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

Now, we follow P. Massart's proof [**50**] to get the result. ■

1.6.2. Model selection Theorems.

1.6.2.1. *Proof of Proposition 9.*

Proof. Let $\{\varphi_1, ..., \varphi_D\}$ be an orthonormal basis of S. We are going to prove in fact 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.

NB : we can suppose the φ_i's 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 1 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 achieved in $\hat{a}$ which verifies, for all i, $\hat{a}_i = \nu_{\mathbb{X}}(\varphi_i)/\chi(S)$. It implies that on $\Omega_S(\varepsilon)$, $\|\sum_{i=1}^{D} \hat{a}_i \varphi_i\|_\infty \leq C(\varepsilon)/z$ where $C(\varepsilon) = 2\kappa\varepsilon M/\kappa(\varepsilon)$ and where z is a lower bound for $\chi(S)$.

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 $\mathcal{B}$ is $\{a \in \mathbb{R}^D / \|a\|_2 = 1 \text{ and } \|\sum_{i=1}^{D} a_i \varphi_i\|_\infty \leq C(\varepsilon)/z\}$.

On the event $\Omega_S(\varepsilon) \cap \{\chi(S) \geq z\}$, we have $\chi' = \chi(S)$.

We can apply Corollary 1 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. ■

1.6.2.2. *Proof of Theorem 3.*

Proof. Let m be a fixed index in $\mathcal{M}_{\mathbb{X}}$. We start with Equation (1.3.2).

(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 (1.2.2), we obtain that with probability larger than $1 - \sum_{m' \in \mathcal{M}_{\mathbb{X}}} e^{-x_{m'}}$,

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

We can remark that
- $\int_{\mathbb{X}} \frac{(s_{m'} - s_m)^2}{\mu(\mathbb{X})^2} s d\mu \leq \|s_{m'} - s_m\|^2 \frac{\|s\|_\infty}{\mu(\mathbb{X})}$ and
- $\|s_{m'} - s_m\|_\infty \leq \|s_{m'}\|_\infty + \|s_m\|_\infty$. It implies by Assumption 2 that

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

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

As $\hat{m}$ is in $\mathcal{M}_{\mathbb{X}}$, we have that (using (1.2.9)) for all positive θ, with probability larger than $1 - \sum_{m' \in \mathcal{M}_{\mathbb{X}}} e^{-x_{m'}}$

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

It leads for all positive η to (using (1.2.9))

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

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

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

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

$$\begin{aligned} \nu_{\mathbb{X}}(s_{\hat{m}} - s_m) \quad &\leq \quad \theta \|s_{\hat{m}} - s_m\|^2 + \frac{\eta}{3} \frac{\rho \mathbb{D}_m}{\mu(\mathbb{X})} + \left[\frac{\eta}{3} + \frac{\delta}{2\theta} + \frac{\delta^2}{3\eta} \right] \frac{\rho \mathbb{D}_{\hat{m}}}{\mu(\mathbb{X})} + \\ &\frac{\|s\|_\infty}{2\theta} \frac{\xi}{\mu(\mathbb{X})} + \frac{\|s\|^2}{3\eta\rho} \frac{\xi^2}{\mu(\mathbb{X})}. \end{aligned} \tag{1.6.13}$$

(2) If we go back to Equation (1.3.2), we can 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 (1.2.9), we obtain that for

all $\gamma > 1$ and $\beta > 0$ (which fix δ, η, θ), with probability larger than $1 - \mathcal{E}e^{-\xi}$

$$\begin{aligned} D(\gamma,\beta)\|\tilde{s} - s\|^2 &\leq D'(\gamma,\beta)\|s_m - s\|^2 + \gamma\chi^2_{\hat{m}} + D''(\gamma,\beta)A_m + \beta A_{\hat{m}} + \\ &\quad + \text{pen}(m) - \text{pen}(\hat{m}) + \frac{f(\xi)}{\mu(\mathbb{X})} \end{aligned} \tag{1.6.14}$$

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 $\chi^2_{\hat{m}}$. 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

- $V_{m'} = \int_{\mathbb{X}} \sum_{\lambda \in \mathcal{B}_{m'}} \varphi^2_\lambda s d\mu / \mu(\mathbb{X})$;
- $M_{m'} = \sup_{f \in S_{m'}, \|f\|=1} \int_{\mathbb{X}} f^2 s d\mu / \mu(\mathbb{X}) \leq \|s\| \sqrt{\mathbb{D}_{m'}}$ 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 have always Equation (1.6.14).

Using (1.2.9), 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 remark too that the control of all the $\chi_{m'}$ implies in particular the control of $\chi_{\hat{m}}$ on the same set of probability. We can take the square and use inequality (1.2.9). Finally, we obtain with probability larger than $1 - (\mathcal{E} + \mathcal{G})e^{-\xi}$

$$\chi^2_{\hat{m}} \leq (1+3\varepsilon)^3 A_{\hat{m}} + (1 + \frac{1}{3\varepsilon})\left(\frac{\kappa\|s\|}{2\varepsilon\sqrt{\rho}} + \kappa(\varepsilon)\right)^2 \frac{\xi^2}{\mu(\mathbb{X})} \tag{1.6.15}$$

and Equation (1.6.14).

We can then rewrite Equation (1.6.14) : for all $d > 1$ (the choice of d fixes the choices of $\alpha, \beta, \delta, \eta, \theta, \gamma$), with probability larger than $1 - (\mathcal{E} + \mathcal{G})e^{-\xi}$,

$$C(d)\|\tilde{s} - s\|^2 \leq C'(d)\|s_m - s\|^2 + dA_{\hat{m}} + C''(d)A_m + \frac{g(\xi)}{\mu(\mathbb{X})} + \text{pen}(m) - \text{pen}(\hat{m}) \tag{1.6.16}$$

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 (1.6.16) disappear. It remains

$$B(c)\|\tilde{s} - s\|^2 \leq B'(c)\|s_m - s\|^2 + B''(c)\text{pen}(m) + \frac{h(\xi)}{\mu(\mathbb{X})} \tag{1.6.17}$$

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's 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 (1.6.17) but on a

set of probability with measure $1-(\mathcal{E}+\mathcal{G}+\mathcal{H}+1)e^{-\xi}$. 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 $\rho D_{m'}$. We do not make the increase $D_{m'} \leq \mathbb{D}_{m'}$. Through the assumption, we can upper bound then $\beta D_{m'}$ 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.

■

1.6.2.3. *Proof of Theorem 4.*

Proof. Let m be an index in $\mathcal{M}_{\mathbb{X}}$, Inequality (1.3.2) 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 (1.2.9) 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 number.

$$\frac{\varepsilon}{2+\varepsilon}\|\tilde{s}-s\|^2 \leq (1+\frac{2}{\varepsilon})\|s-s_m\|^2 + (1+\varepsilon)\chi^2_{m,\hat{m}} - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m). \tag{1.6.18}$$

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 number. 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 (1.2.9), 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 $1-(\Sigma_0+\Sigma_1)e^{-\xi}$

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

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 (1.6.18) 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

(1.6.19)

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

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 a large bound $\mathbb{E}(\chi^4_\Lambda)$. 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 this bound

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

where C^0 is a continuous positive function. This bound and the bound in (1.6.19) implies exactly the bound mentioned in Theorem 4. ■

1.6.2.4. *Proof of Proposition 10.* **Proof.** This is an application of Theorem 4. Assumption 6 of Theorem 4 is Assumption 2 of this proposition. As we are in the subsets case, we have obviously :

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

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 10 has the role of S_Λ in Theorem 4.)

Let M (the one of Theorem 4) 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 use then Proposition 7 to obtain

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

for η continuous positive function with $u = (2\kappa M\varepsilon/(B\sqrt{|\Lambda|}\kappa(\varepsilon)))$. Assumption 3 of the proposition implies then 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 suppose an upper bound of $\|s\|_\infty$ to be known. This upper bound is for the theorem $\hat{M} = M'$.

– 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,
  - $G_{m'} = \int_{\mathbb{X}} \frac{\left(\sum_{\lambda \in m'} \varphi_\lambda^2\right)^2}{\mu(\mathbb{X})^2} s d\mu \leq V(m') \frac{H_{m'}}{\mu(\mathbb{X})}$.

  We deduce from this 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 (1.2.9) to obtain Assumption 4 with $\Sigma_0 = \Sigma$.

Then all the conclusions are consequences of Theorem 4.

NB : The result of Theorem 4 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 $\mathbb{E}(\text{pen}(m)\mathbb{1}_{\Omega(\varepsilon)}) \leq B(\varepsilon)\frac{(\|s_\Lambda\|_\infty + K')|m|}{\mu(\mathbb{X})}(1 + L_m)$ for B positive continuous function. ■

1.6.2.5. *Proof of Proposition 11.* **Proof.** We are going to apply here still Theorem 4. Assumption 6 of Theorem 4 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 can 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, we can take the bound on the variances like this

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

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 4, we set

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

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 4 is obviously S_Γ with basis $\{\mathbb{1}_I\sqrt{D_\Gamma}, I \in \Gamma\}$. Consequently, Assumption 1 is a consequence of Assumption 2 of Proposition 11.

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

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

for the same reasons 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 use then 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 continuous positive 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 as close as we want to 1, it depends on ε. Then Assumptions 4 and 5 are obvious, with $\Sigma_0 = 0$ and $\eta = 0$. Applying Theorem 4, we get exactly the conclusion, making the same remark between c and ε as in the proof of Proposition 10, remarking that on $\Omega(\varepsilon)$,

$$\hat{M} \leq \left(1 + \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}\right) M.$$

■

CHAPITRE 2

Exponential inequalities and martingales

Abstract

In this chapter, we study exponential inequalities using martingale approach. We focus on degenerate U-statistics of order 2 and on suprema of centred integrals with respect to a counting process. Indeed these quantities appear naturally in the study of some non parametric statistical problems.

AMS Classification : 60E15, 60G42, 60G55.
Keywords : Exponential inequalities, degenerate U-statistics of order 2, counting process, supremum.

Résumé

Dans ce chapitre, nous étudions des inégalités exponentielles utilisant des techniques de martingales. Nous nous intéressons principalement sur des U-statistiques dégénérées d'ordre 2 et sur des suprema d'intégrales centrées contre la mesure de comptage. Ces quantités apparaissent en effet naturellement dans l'étude de certains problèmes de statistiques non paramétriques.

Classification AMS : 60E15, 60G42, 60G55.
Mots clefs : Inégalités exponentielles, U-statistiques dégénérées d'ordre 2, processus de comptage, supremum.

2.1. Introduction

The aim of this chapter is to find exponential inequalities for degenerate U-statistics of order 2 and for suprema of integrals with respect to a counting process. These two problems are closely connected since they both use martingale approach.

At first, let us fix the notations by recalling what an exponential concentration inequality is.

Let X be a real random variable. We say that one has an exponential inequality for the deviations of X above m if there exists some positive constant C and some positive function f such that for all positive x

$$\mathbb{P}(X \geq m + x) \leq Ce^{-f(x)}. \tag{2.1.1}$$

This is equivalent to say that there exists some positive constant C and some positive function h such that for all positive x

$$\mathbb{P}(X \geq m + h(x)) \leq Ce^{-x}. \tag{2.1.2}$$

We speak in this case of an inverse formula.

One of the most known concentration inequalities is the one due to B.S. Cirel'son, I.A. Ibragimov and V.N. Sudakov. Let X be a 1-Lipschitz function of a standard Gaussian vector. Then one has a concentration inverse formula with $C = 1$, $m = \mathbb{E}(X)$ and $h(x) = \sqrt{2x}$ or, for the direct formula, with $f(x) = x^2/2$ (see **[23]**).

We see here the typical Gaussian behavior : $h(x)$ is a purely quadratic term of the type $\sqrt{2vx}$ where v is called the variance term. If X is a standard Gaussian variable, one has again the previous formula with $h(x) = \sqrt{2vx}$ and with $v = 1$ which is the true variance of X .

There exists also very classical concentration inequalities for a sum of independent variables. Let $Y_1, ..., Y_n$ be n independent real variables and $X_n = \sum_{i=1}^n Y_i$.

If we assume that each Y_i takes its values in $[a_i, b_i]$, then we can obtain Hoeffding's inequality. This is an inverse concentration formula with $m = \mathbb{E}(X_n)$, $C = 1$ and

$$h(x) = \sqrt{\frac{(\sum_{i=1}^n (a_i - b_i)^2)\, x}{2}}.$$

Consequently, $h(x)$ is also here a purely quadratic term and X_n is sub-Gaussian with a variance term equal to $\sum_{i=1}^n (a_i - b_i)^2/4$, which is an upper bound of the true variance of X_n.

Under the same assumptions, one can improve this inequality to get the true variance v of X_n in the variance term and not just an upper bound. If we assume that the Y_i's are centred and bounded in absolute value by b, we can obtain a concentration inequality with $m = \mathbb{E}(X_n) = 0$, $C = 1$ and $h(x) = \sqrt{2vx} + bx/3$. In this case, $h(x)$ is no more a purely quadratic term : we add to the quadratic term with variance term v a linear

term, depending on a bound on the Y_i's. In this section, we call this inequality a "weak Bernstein's" inequality.

In fact this last exponential inequality is a consequence of two different inequalities.

For the first one, called Bennett's inequality, we must assume the same things on the Y_i's. Then we obtain a concentration inequality (see (2.1.1)) with $C = 1$, $m = 0$ and

$$f(x) = \frac{v}{b^2}\left(\left(1 + \frac{bx}{v}\right)\log\left(1 + \frac{bx}{v}\right) - \frac{bx}{v}\right).$$

When x is closed to 0, we recover a sub-Gaussian behavior with variance v as for the "weak Bernstein's" inequality. For large x, it is a little better than the "weak Bernstein's" inequality but we cannot yet recover sub-Gaussian behavior with variance term equal to v for all the possible x.

The second inequality, called Bernstein's inequality, allows us to deal with unbounded Y_i's. If the Y_i's are centred and verify that there exists c and w positive numbers such that for all integer k larger than 2,

$$\sum_{i=1}^{n} \mathbb{E}(|Y_i|^k) \leq \frac{k!}{2} w c^{k-2},$$

then one has an inverse concentration formula (2.1.2), with $C = 1$, $m = 0 = \mathbb{E}(X_n)$ and $h(x) = \sqrt{2wx} + cx$.

We can remark that in all the previous inequalities $(X_n)_{n\in\mathbb{N}}$ is in fact a martingale.

More generally, let us assume that $(X_n)_{n\in\mathbb{N}}$ is just a martingale with $X_0 = 0$. Let n be a fixed integer. One has exactly an Hoeffding's inequality for X_n with $m = 0$ and variance term equal to $\sum_{i=1}^{n} d_i^2/4$ if one assumes that each $|X_i - X_{i-1}|$ is bounded by d_i. This exponential inequality is known as Azuma-Hoeffding's inequality.

I. Pinelis generalizes also a lot of others exponential inequalities, such as Bennett's or Bernstein's ones, to the discrete time martingale setup in [**53**].

One can also consider continuous time martingale as S. van de Geer does in [**63**] or as O. Kallenberg does in [**37**] and recover the same type of results.

The previous inequalities are not just useful to consider more general setup of dependency, but also to get exponential inequalities for $f(Y_1, ..., Y_n)$ where the Y_i's are independent variables.

For instance, if we assume that for each i and for all x and y_1^n,

$$|f(y_1, ..., y_n) - f(y_1, ..., y_{i-1}, x, y_{i+1}, ..y_n)| \leq d_i,$$

we can prove an exponential inequality for the deviations of $X_n = f(Y_1, ..., Y_n)$ above its mean ($m = \mathbb{E}(X_n)$) with $C = 1$ and $h(x) = \sqrt{\left(\sum_{i=1}^{n} d_i^2\right) x/2}$. For this aim, it is sufficient to apply correctly the Azuma-Hoeffding's inequality.

This inequality is known as the McDiarmid inequality. It can be viewed as the generalization of Hoeffding's inequality to a general function of the Y_i's.

By the same trick, one can also derived exponential inequalities for specific time processes using continuous time martingale approach. Indeed, even if the study of Poisson processes can be made using only its infinitely divisible properties, the use of martingale technics gives exponential inequalities for a general functional of all the Poisson process. This is the work of L. Wu [**64**]. C. Houdré and N. Privault [**36**] prove a similar result with similar technics for particular martingales called normal martingales which Poisson processes are part of.

In both cases, the resulting exponential inequality is an intermediate formula between the weak Bernstein's inequality and the Hoeffding's inequality. One has $C = 1$, m is the mean and $h(x) = \sqrt{2wx} + cx$ where w is an upper bound of the true variance but not as tall as the one which would be given by an Hoeffding's type inequality.

Let us return to the i.i.d. case. We can improve the result of McDiarmid for specific function of $Y_1, ..., Y_n$, i.i.d. variables. M. Talagrand proves in [**62**] an exponential inequality for

$$X_n = \sup_{a \in A} \sum_{i=1}^{n} [\psi_a(Y_i) - \mathbb{E}(\psi_a(Y_i))],$$

where $\{\psi_a, a \in A\}$ is a countable family of bounded measurable functions.

This inequality has known a lot of improvements due to M. Ledoux [**45**] (for a simpler proof), to P. Massart [**50**] (for the knowledge of the constants)... The last and simplest version is due to E. Rio [**58**]. He proves an exponential inequality of type (2.1.1) with $m = \mathbb{E}(X_n)$, $C = 1$ and

$$h(x) = \sqrt{(2v + 4b\mathbb{E}(X_n))x} + (b/2)x, \tag{2.1.3}$$

where $v = n \sup_{a \in \mathcal{A}} \mathrm{Var}(\psi_a(X_1))$ and $b = \sup_{a \in A} \|\psi_a\|_\infty$.

This inequality is better in the variance term than McDiarmid's inequality, even if we do not recover the true variance of X_n. However, its form is closer to the "weak Bernstein's" inequality than to the Hoeffding's inequality.

The proof of Talagrand's inequality does not consist in a martingale approach but in the use of entropy tensorisation which is a specific technics for the independent case.

Using the infinitely divisible property and the entropy tensorisation, we improve likewise in Chapter 1 Wu's inequality for the Poisson process for a particular functional of the process, the supremum of centred integrals.

There exists also inequalities for more general functions of the Y_i's derived by entropy tensorisation in [**16**].

Degenerate U-statistics of order 2 are also particular functionals of independent variables. We will give exponential concentration inequalities for these functions in Section 2.2. We use for this aim a hybrid method combining Talagrand's inequality and martingale approach. These results can in particular not be recovered easily by purely entropy tensorisation method.

As the Poisson case is very similar, we derive also in the same section exponential inequalities for double integrals of Poisson processes.

There exists also cases where there are no independent structures and where we can only use martingale approach. This is the case for suprema of integrals with respect to a counting process, treated in Section 2.3.

2.2. Exponential inequalities for degenerate U-statistic of order 2 with constants

The degenerate U-statistics of order 2 are very interesting since they naturally appear in the problem of adaptively estimating $\int s^2 dx$ where s is a density of an observed n-sample [**44**]. Consequently, they appear also in the problem of testing s in a non parametric way [**29**]. To construct those estimators or these tests, B. Laurent and M. Fromont need precise exponential formula with precise constants.

In this section, we establish these exponential inequalities and we treat also the Poissonian case since the two frameworks are very similar. We do not give in this PhD thesis precise statistical applications of these results to estimation or tests.

W. Hoeffding has already studied degenerate U-statistics of order 2. He has proved convergence theorems in [**33**] and also exponential inequalities in [**34**]. These inequalities are not sharp enough to be used in the previous statistical problems.

The recent work of E. Giné, R. Latala and J. Zinn [**31**] provides exponential bound for general U-statistics with proper orders of magnitude but with unknown constants.

Some precise constants are already derived in a very special case by J. Bretagnolle [**18**].

The work in this section is not achieved and carries on with C. Houdré to treat more general U-statistics than only degenerate U-statistics of order 2.

2.2.1. The n-sample framework. Let us recall some known facts about degenerate U-statistics of order 2. Let $T_1, ..., T_n, ...$ be a sequence of i.i.d. variables. We set for all integer n,

$$U_n = \sum_{i=1}^{n} \sum_{j=1}^{i-1} g(T_i, T_j), \tag{2.2.1}$$

where g is symmetric with $\mathbb{E}(g(T_1, T_2)|T_2) = 0$.

U_n represents a degenerate U-statistics of order 2 since we have :

LEMMA 7. *Let $\mathcal{U}$ be a general degenerate U-statistics of order 2, i.e. for some measurable f, $\mathcal{U}$ is defined by*

$$\mathcal{U} = \sum_{i \neq j} \left(f(T_i, T_j) - \mathbb{E}(f(T_i, T_j)|T_j) - \mathbb{E}(f(T_i, T_j)|T_i) + \mathbb{E}(f(T_i, T_j)) \right).$$

- *If f is symmetric, then one has $2U_n = \mathcal{U}$ with g defined by*

$$g(T_1, T_2) = f(T_1, T_2) - 2\mathbb{E}(f(T_1, T_2)|T_2) + \mathbb{E}(f(T_1, T_2)).$$

- *If f is not symmetric, then one has $U_n = \mathcal{U}$ with g defined by*

$$g(T_1, T_2) = f(T_1, T_2) - \mathbb{E}(f(T_1, T_2)|T_2) - \mathbb{E}(f(T_1, T_2)|T_2) + \mathbb{E}(f(T_1, T_2)) + f(T_2, T_1) - \mathbb{E}(f(T_2, T_1)|T_1) - \mathbb{E}(f(T_2, T_1)|T_1) + \mathbb{E}(f(T_2, T_1)).$$

For all n, we denote by $\mathcal{F}_n$ the σ-field generated by $\{T_1, ..., T_n\}$ and

$$X_n = \sum_{j=1}^{n-1} g(T_n, T_j).$$

There exists another easy but very important lemma :

LEMMA 8. *$(U_n, n \in \mathbb{N})$ is a discrete time martingale with respect to the filtration $(\mathcal{F}_n, n \in \mathbb{N})$ and for all n, $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = 0$.*

Proof. Let n be a positive integer. X_n is $\mathcal{F}_n$-measurable. Then one has

$$\mathbb{E}(X_n|\mathcal{F}_{n-1}) = \sum_{j=1}^{n-1} \mathbb{E}(g(T_n, T_j)|\mathcal{F}_{n-1}) = \sum_{j=1}^{n-1} \mathbb{E}(g(T_n, T_j)|T_j)$$

by the independence property. The assumptions on g imply that this quantity is null. Moreover, $U_n = \sum_{i=1}^{n} X_i$. Then one has the martingale property :

$$\mathbb{E}(U_n|\mathcal{F}_{n-1}) = U_{n-1} + \mathbb{E}(X_n|\mathcal{F}_{n-1}) = U_{n-1}.$$

■

We give now two different inequalities using either a "Bennett's" type inequality or a "Bernstein's" type inequality for discrete time martingale.

2.2.1.1. *Using "Bennett's" inequality for martingales.*

We denote by V_n the angle bracket of U_n, i.e. (see **[52]**,p 148) : $V_n = \sum_{i=1}^{n} \mathbb{E}(X_i^2|\mathcal{F}_{i-1})$. We denote also by B_n, $\sup_{i \leq n} |X_i|$. Finally, $\mathbb{E}_x(f(x, y))$ means that we integrate only with respect to x with x independent from y.

THEOREM 5. *Let u and ε be positive real numbers. If g is bounded by A, then*

$$\mathbb{P}\left[U_n \geq (1+\varepsilon)C\sqrt{2u} + \left(2\sqrt{\kappa}D + \frac{1+\varepsilon}{3}F\right)u + \left(\sqrt{2}\kappa(\varepsilon) + \frac{2\sqrt{\kappa}}{3}\right)Bu^{3/2} + \frac{\kappa(\varepsilon)}{3}Au^2\right] \leq 3e^{-u} \wedge 1,$$

where

$$C^2 = \frac{n(n-1)}{2}\mathbb{E}(g(T_1,T_2)^2),$$

$$D = \sup\left\{\mathbb{E}(\sum_{i=1}^{n}\sum_{j=1}^{i-1} g(u,v)a_i(u)b_j(v)), \mathbb{E}(\sum_{i=1}^{n} a_i(u)^2) \leq 1, \mathbb{E}(\sum_{j=1}^{n} b_j(v)^2) \leq 1\right\},$$

$$F = \mathbb{E}\left(\sup_{i,u}\left|\sum_{j=1}^{i-1} g(u,T_j)\right|\right),$$

and

$$B^2 = n\sup_u \mathbb{E}_v(g(u,v)^2).$$

(The values of $\kappa = 4$ and $\kappa(\varepsilon) = (2.5 + 32\varepsilon^{-1})$ are given by P. Massart in **[50]***.)*

Proof. Let u be a positive number. At first, let us get some upper bounds for V_n and B_n.

LEMMA 9. *With probability larger than $1 - 2e^{-u}$, one has*

$$\sqrt{V_n} \leq (1+\varepsilon)C + D\sqrt{2\kappa u} + \kappa(\varepsilon)Bu = \sqrt{v}$$

and

$$B_n \leq (1+\varepsilon)F + B\sqrt{2\kappa u} + \kappa(\varepsilon)Au = b.$$

This lemma consists in applying properly Talagrand's inequality of Massart's version **[50]** to get a version for independent variables but not necessarily with the same law. Indeed, we can remark that

$$\sqrt{V_n} = \sup_{\sum_{i=1}^n \mathbb{E}_u(a_i(u)^2)=1} |\sum_{j=1}^{n}\sum_{i=j+1}^{n} \mathbb{E}_u(a_i(u)g(u,T_j))|$$

and

$$B_n = \sup_i |X_i| \leq \sup_{i,u} |\sum_{j=1}^{i-1} g(u,T_j)|.$$

In both cases, by density of a countable subset of indices, they are suprema of the form $\sup_{t\in\mathcal{T}}\sum_{i=1}^n f_{i,t}(\xi_i)$ where $\mathcal{T}$ is countable, the $f_{i,t}$'s are centered and bounded and the ξ_i's are independent variables.

Now we go back to U_n. More precisely we define the stopping time T by

$$T + 1 = \inf\{k \in \mathbb{N}, V_k > v \text{ or } B_k > b\}.$$

Then U_n^T, the martingale stopped in T, is always a martingale. As V_k and B_k are nondecreasing, the angle bracket and the jumps of this new martingale are bounded respectively by v and b. Consequently, (see **[52]**, Lemma VII-2-8, p 154), we have that for all λ positive

$$\left(e^{\lambda U_n^T - \phi_c(\lambda)v}, n \in \mathbb{N}\right)$$

is a super-martingale where $\phi_c(\lambda) = (e^{\lambda c} - \lambda c - 1)/c^2$. Then we make the same computations on the Laplace transform of U_n^T as the ones we do to prove the Bennett's inequality in the i.i.d. case. We get consequently by Bienaymee-Tchebicheff's inequality a "weak Bernstein's" inequality for U_n^T :

$$\mathbb{P}\left(U_n^T \geq \sqrt{2vu} + \frac{b}{3}u\right) \leq e^{-u}.$$

Thus

$$\mathbb{P}\left(U_n \geq \sqrt{2vu} + \frac{b}{3}u\right) \leq \mathbb{P}\left(U_n^T \geq \sqrt{2vu} + \frac{b}{3}u\right) + \mathbb{P}(T+1 \leq n)$$

is bounded by $3e^{-u}$ using Lemma 9. ■

2.2.1.2. *Using "Bernstein's" inequality for martingales.*

Let us at first derive the equivalent of the previous super-martingale when there exists no bounded increments.

LEMMA 10. *Let $(Y_n, n \in \mathbb{N})$ be a martingale with mean 0. For all $k \geq 2$, let*

$$A_n^k = \sum_{i=1}^{n} \mathbb{E}\left((Y_i - Y_{i-1})^k | \mathcal{F}_{i-1}\right).$$

Then for all integer n and for all λ,

$$\mathcal{E}_n = \exp\left(\lambda Y_n - \sum_{k\geq 2} \frac{\lambda^k}{k!} A_n^k\right)$$

is a super-martingale.

Proof. For all integer n,

$$\mathbb{E}(\mathcal{E}_n|\mathcal{F}_{n-1}) = \mathcal{E}_{n-1}\mathbb{E}(e^{\lambda(Y_n - Y_{n-1})}|\mathcal{F}_{n-1}) \exp\left(-\sum_{k\geq 2} \frac{\lambda^k}{k!}\mathbb{E}\left((Y_n - Y_{n-1})^k|\mathcal{F}_{n-1}\right)\right).$$

But

$$\begin{aligned}\mathbb{E}\left(e^{\lambda(Y_n - Y_{n-1})}|\mathcal{F}_{n-1}\right) &\leq 1 + \sum_{k\geq 2} \frac{\lambda^k}{k!}\mathbb{E}\left((Y_n - Y_{n-1})^k|\mathcal{F}_{n-1}\right) \\ &\leq \exp\left(\sum_{k\geq 2} \frac{\lambda^k}{k!}\mathbb{E}\left((Y_n - Y_{n-1})^k|\mathcal{F}_{n-1}\right)\right).\end{aligned}$$

The result is then obvious. ■

NB : A_n^2 is the classical angle bracket.

If the A_n^k's are bounded by some a_n^k's positive, we have for all λ positive

$$\mathbb{E}(e^{\lambda Y_n}) \leq \exp\left(\sum_{k\geq 2} \frac{\lambda^k}{k!} a_n^k\right). \tag{2.2.2}$$

I. Pinelis proves theses results in Theorem 8.5 of [**53**].

Now let us give the exponential inequality for degenerate U-statistics of order 2 using a "Bernstein's" type inequality.

THEOREM 6. *Let us keep the notations of Theorem 5. Then for all ε and x positive,*

$$\mathbb{P}(U_n \geq 2(1+\varepsilon)^{3/2} C\sqrt{x} + 2\eta(\varepsilon)Dx + \beta(\varepsilon)Bx^{3/2} + \gamma(\varepsilon)Ax^2) \leq 2.77e^{-x}$$

where $\eta(\varepsilon) = 1.42\sqrt{\kappa}(2+\varepsilon+\varepsilon^{-1})$, $\beta(\varepsilon) = e(1+\varepsilon^{-1})^2\kappa(\varepsilon) + (1.42\sqrt{\kappa}(2+\varepsilon+\varepsilon^{-1})) \wedge \frac{(1+\varepsilon)^2}{\sqrt{2}}$ *and* $\gamma(\varepsilon) = (e(1+\varepsilon^{-1})^2\kappa(\varepsilon)) \wedge \frac{(1+\varepsilon)^2}{3}$ *with* $\kappa = 4$ *and* $\kappa(\varepsilon) = 2.5 + 32\varepsilon^{-1}$.

Proof. The A_n^k's of Lemma 10 corresponding to U_n are

$$A_n^k = \sum_{i=1}^n \mathbb{E}_u\left[\left(\sum_{j=1}^{i-1} g(u,T_j)\right)^k\right] \leq V_n^k = \sum_{i=1}^n \mathbb{E}_u\left[\left|\sum_{j=1}^{i-1} g(u,T_j)\right|^k\right].$$

We can as previously bound these V_n^k in probability using Talagrand's inequality.

LEMMA 11. *With probability larger than $1 - 1.77e^{-x}$, one has for all $k \geq 2$*

$$(V_n^k)^{1/k} \leq (1+\varepsilon)(\mathbb{E}(V_n^k))^{1/k} + \sigma_k\sqrt{2\kappa k x} + \kappa(\varepsilon) b_k k x,$$

where

$$\sigma_k^2 = n \sup_{\sum_{i=1}^n \mathbb{E}_u(|a_{i,u}|^{k/(k-1)})=1} \left\{\sum_{j=1}^n \mathbb{E}_v\left(\mathbb{E}_u\left[\sum_{i=j+1}^n a_{i,u} g(u,v)\right]\right)^2\right\}$$

and

$$b_k = \sup_{\sum_{i=1}^n \mathbb{E}_u(|a_{i,u}|^{k/(k-1)})=1, j\leq n} \|\mathbb{E}_u\left[\sum_{i=j+1}^n g(u,T_j)a_{i,u}\right]\|_\infty.$$

Proof. We can wrote by Hölder's inequality :

$$(V_n^k)^{1/k} = \sup_{\sum_{i=1}^n \mathbb{E}_u(|a_{i,u}|^{k/(k-1)})=1} \left\{\sum_{j=1}^n \mathbb{E}_u\left(\sum_{i=j+1}^n g(u,T_j)a_{i,u}\right)\right\}.$$

By density of a countable subset of indices, the V_n^k's are suprema of the form

$$\sup_{t\in\mathcal{T}} \sum_{i=1}^n f_{i,t}(\xi_i)$$

where $\mathcal{T}$ is countable, the $f_{i,t}$'s are centered and bounded and the ξ_i's are i.i.d. variables. We can consequently apply Talagrand's inequality ([**50**]) : for all $k \geq 2$, for all $z, \varepsilon > 0$

$$\mathbb{P}\left((V_n^k)^{1/k} \geq (1+\varepsilon)(\mathbb{E}(V_n^k))^{1/k} + \sigma_k\sqrt{2\kappa z} + \kappa(\varepsilon)b_k z\right) \leq e^{-z}.$$

We apply it to $z = kx$ and we sum to obtain :

$$\mathbb{P}\left(\forall k \geq 2, \quad (V_n^k)^{1/k} \geq (1+\varepsilon)(\mathbb{E}(V_n^k))^{1/k} + \sigma_k\sqrt{2\kappa k x} + \kappa(\varepsilon)b_k kx\right) \leq \sum_{k\geq 2} e^{-kx}.$$

In fact the right part is precisely

$$1 \wedge \sum_{k\geq 2} e^{-kx} \leq 1 \wedge e^{-x}/x \leq 1.77e^{-x}.$$

■

Now we have to bound these quantities. The easiest is b_k : by Hölder one has

$$b_k \leq \sup_{j\leq n, v}\left((n-j)\mathbb{E}_u(|g(u,v)|^k)\right)^{1/k} \leq (B^2A^{k-2})^{1/k}.$$

For the variance term, this is a little more intricate :

$$\begin{aligned}
\sigma_k^2 &= \sup_{\substack{\sum_{i=1}^n \mathbb{E}_u(|a_{i,u}|^{k/(k-1)}) = 1 \\ \sum_{j=1}^n \mathbb{E}_v(|b_{j,v}|^2) = 1}} \sum_{j=1}^n \mathbb{E}_v\left(\sum_{i=j+1}^n \mathbb{E}_u(g(u,v)a_{i,u}b_{j,v})\right) \\
&= \sup_{\substack{\sum_{i=1}^n \mathbb{E}_u(|a_{i,u}|^{k/(k-1)}) = 1 \\ \sum_{j=1}^n \mathbb{E}_v(|b_{j,v}|^2) = 1}} \sum_{i=1}^n \mathbb{E}_u\left(\sum_{j=1}^{i-1} \mathbb{E}_v(g(u,v)b_{j,v})a_{i,u}\right) \\
&= \sup_{\sum_{j=1}^n \mathbb{E}_v(|b_{j,v}|^2)=1} \left[\sum_{i=1}^n \mathbb{E}_u\left[\mathbb{E}_v(\sum_{j=1}^{i-1} g(u,v)b_{j,v})\right]^k\right]^{1/k} \\
&\leq (B^{k-2}D^2)^{1/k}.
\end{aligned}$$

We keep for the moment the expectation of V_n^k. By derivation, one has the following upper bound :

$$\forall k > 1, x, \varepsilon > 0, (1+x)^k \leq (1+\varepsilon)^{k-1} + (1+\varepsilon^{-1})^{k-1}x^k. \tag{2.2.3}$$

Thus we get that with probability larger than $1 - 1.77e^{-x}$, for all $k \geq 2$, V_n^k is bounded by a_n^k given by

$$a_n^k = (1+\varepsilon)^{2k-1}\mathbb{E}(V_n^k) + (2+\varepsilon+\varepsilon^{-1})^{k-1}D^2B^{k-2}\sqrt{2\kappa k x}^k + (1+\varepsilon^{-1})^{2k-2}B^2A^{k-2}\kappa(\varepsilon)^k(kx)^k.$$

Like in the previous proof, we call

$$T + 1 = \inf\{p \in \mathbb{N}, \exists k, V_p^k \geq a_n^k\}.$$

We have since the V_n^k are nondecreasing, $\mathbb{P}(T < n) \leq 1.77e^{-x}$.

Now we stop U_n in T. The new martingale U_n^T has, for new "A_n^k"'s, the A_n^k stopped in T : consequently they are bounded by a_n^k. We get for the Laplace transform of U_n^T, that for all $\lambda > 0$,

$$\mathbb{E}(e^{\lambda U_n^T}) \leq \exp\left(\sum_{k\geq 2} \frac{\lambda^k}{k!} a_n^k\right).$$

It remains to simplify this bound and to use Bienaymee-Tchebicheff inequality.

$$\begin{aligned} a_n &= \sum_{k\geq 2} \frac{\lambda^k}{k!} a_n^k \\ &\leq \sum_{k\geq 2} \frac{\lambda^k}{k!} (1+\varepsilon)^{2k-1} \mathbb{E}(V_n^k) + \\ &\quad + \sum_{k\geq 2} \frac{\lambda^k}{k!} (2+\varepsilon+\varepsilon^{-1})^{k-1} D^2 B^{k-2} \sqrt{2\kappa k x}^k + \\ &\quad + \sum_{k\geq 2} \frac{\lambda^k}{k!} (1+\varepsilon^{-1})^{2k-2} B^2 A^{k-2} \kappa(\varepsilon)^k (kx)^k. \end{aligned}$$

Let us call respectively α, β and γ, the three previous sums. For the last term, if we call $\delta(\varepsilon) = e(1+\varepsilon^{-1})^2 \kappa(\varepsilon)$ we get

$$\gamma \leq \sum_{k\geq 2} (\delta(\varepsilon))^k B^2 A^{k-2} (\lambda x)^k = \frac{\lambda^2 (B\delta(\varepsilon)x)^2}{1 - (A\delta(\varepsilon)x)\lambda}$$

for $\lambda < (A\delta(\varepsilon)x)^{-1}$.

For the second term, if we call $\eta(\varepsilon) = 1.0007\sqrt{2\kappa}(2+\varepsilon+\varepsilon^{-1})$, we get with similar computations

$$\beta \leq \frac{\lambda^2 (D\eta(\varepsilon)\sqrt{x})^2}{1 - (B\eta(\varepsilon)\sqrt{x})\lambda}$$

for $\lambda < (B\eta(\varepsilon)\sqrt{x})^{-1}$.

For the first term, this is more intricate :

$$\alpha = \frac{1}{1+\varepsilon} \sum_{i=1}^{n} \mathbb{E}_u \left(\mathbb{E}_{(T_j)_{j\in\mathbb{N}}}(\exp(\mu|C_i|)) - \mu \mathbb{E}_{(T_j)_{j\in\mathbb{N}}}(|C_i|) - 1 \right)$$

where $C_i = \sum_{j=1}^{i-1} g(u, T_j)$ and $\mu = \lambda(1+\varepsilon)^2$.

As $e^x - x - 1 > 0$ for all x, we can add $\mathbb{E}_{(T_j)_{j\in\mathbb{N}}}(\exp(-\mu|C_i|)) + \mu \mathbb{E}_{(T_j)_{j\in\mathbb{N}}}(|C_i|) - 1$. We get

$$\alpha \leq \frac{1}{1+\varepsilon} \sum_{i=1}^{n} \mathbb{E}_u \left(\mathbb{E}_{(T_j)_{j\in\mathbb{N}}}(\exp(\mu C_i)) - 1 + \mathbb{E}_{(T_j)_{j\in\mathbb{N}}}(\exp(-\mu C_i)) - 1 \right).$$

As C_i is a sum of i.i.d. centred and bounded quantities, we get using Bernstein's inequality that

$$\alpha \leq \frac{2}{1+\varepsilon} \sum_{i=1}^{n} \mathbb{E}_u \left(e^{\frac{\mu^2 v_{i,u}}{2-2\mu\frac{A}{3}}} - 1 \right)$$

where $v_{i,u} = (i-1)\mathbb{E}_v(g(u,v)^2)$. But $v_{i,u} \leq B^2$, thus $\sum_{i=1}^{n} \mathbb{E}_u(v_{i,u}^k) \leq C^2 B^{2(k-1)}$. After computations we get

$$\alpha \leq \frac{(1+\varepsilon)^3 C^2 \lambda^2}{1 - \lambda(1+\varepsilon)^2 A/3 - \lambda^2 (1+\varepsilon)^4 B^2/2}.$$

We can upper bound it by :

$$\alpha \leq \frac{(1+\varepsilon)^3 C^2 \lambda^2}{1 - (1+\varepsilon)^2 \lambda (A/3 + B/\sqrt{2})}$$

for $\lambda \leq [(1+\varepsilon)^2(A/3 + B/\sqrt{2})]^{-1}$. Finally one has,

$$\mathbb{E}(e^{\lambda U_n^T}) \leq \exp\left(\frac{\lambda^2 W}{1 - \lambda c} \right),$$

where

$$W = (1+\varepsilon)^{3/2} C + \eta(\varepsilon) D \sqrt{x} + \delta(\varepsilon) B x$$

and

$$c = \max\left((1+\varepsilon)^2 (A/3 + B/\sqrt{2}), \eta(\varepsilon) B \sqrt{x}, \delta(\varepsilon) A x \right).$$

This implies :

$$\mathbb{P}(U_n^T \geq 2W\sqrt{x} + cx) \leq e^{-x}.$$

Then with the same trick as previously one gets the bound

$$\mathbb{P}(U_n \geq 2W\sqrt{x} + cx) \leq 2.77 e^{-x}.$$

Moreover if $x \leq 1$, $2.77 \exp(-x) > 1$. Hence we get the result. ■

2.2.1.3. *Comments.*

The results of Theorem 5 and of Theorem 6 are both interesting. The quadratic term in the first one is of the form (when ε tends to 0) $\sqrt{2Cu}$ which is the optimal rate for the Central Limit Theorem since the variance term C represents the true variance of the process.

The quadratic term in the second theorem is larger : it is of the form $2\sqrt{Cu}$, the factor $\sqrt{2}$ coming from the fact that we have done some symmetrization in the proof. This theorem gives precise constants which are unknown in the result of **[31]**. Moreover this last Theorem has better order of magnitude than Theorem 5 as we see in the following example.

Let us look at an easy example, coming from the statistical applications (see **[44]**). The T_i's are uniformly distributed on $[0;1]$. Let m be a regular partition of $[0;1]$ ($|m| = d$).

We set

$$\forall (u,v)\in[0;1]^2, g(u,v)=d\sum_{I\in m}(\mathbb{1}_I(u)-1/d)(\mathbb{1}_I(v)-1/d).$$

Let U_n be the corresponding U-statistics. We refer to the computations of [**44**] in the appendix. One has

$$A\leq 4d,\quad B^2\leq 2nd,\quad C^2\leq \frac{n(n-1)}{2}d,\quad D\leq\frac{(n-1)}{2}.$$

Moreover F can be compute with the use of the Laplace transform : we get F of the order of $d\ln n+n$.

One has consequently the two following concentration inequalities : for all ε and u positive,

– by applying Theorem 5 : with probability smaller than $3e^{-u}$ one has

$$\frac{1}{n(n-1)}\sum_{i\neq j}g(T_i,T_j)=\frac{2U_n}{n(n-1)}\leq$$
$$2(1+\varepsilon)\sqrt{\frac{d}{n(n-1)}u}+\square\left(\frac{1}{n}+\frac{d\ln n}{n^2}\right)u+\square\frac{\sqrt{d/n}}{n-1}u^{3/2}+\square\frac{d}{n(n-1)}u^2.$$

– by applying Theorem 6 : with probability smaller than $2.77e^{-u}$ one has

$$\frac{2U_n}{n(n-1)}\leq 2(1+\varepsilon)^3\sqrt{\frac{2d}{n(n-1)}u}+\square\frac{1}{n}u+\square\frac{\sqrt{d/n}}{n-1}u^{3/2}+\square\frac{d}{n(n-1)}u^2.$$

(The squares are for known but intricate constants.) The second inequality is sharper in the second term. In particular if d is of the order of n^2, the second one remains bounded when the first one tends to infinity when n grows.

2.2.2. The Poisson framework. One can do the same thing for double integral of Poisson processes. Let N be a time Poisson process with compensator Λ. We denote by $(M_t=N_t-\Lambda_t, t\geq 0)$ the corresponding martingale.

The U-statistic or the double integral for Poisson process is defined by

$$Z_t=\int_0^t\int_{(0;y)}f(x,y)dM_xdM_y$$

for f symmetric deterministic function.

Then we can easily obtain the equivalent of Theorem 5.

THEOREM 7. *Let $u,\varepsilon>0$. If f is bounded by A, then*

$$\mathbb{P}\left[Z_t\geq(1+\varepsilon)C\sqrt{2u}+\left(2\sqrt{\kappa}D+\frac{1+\varepsilon}{3}F\right)u+\left(\sqrt{2}\kappa(\varepsilon)+\frac{2\sqrt{\kappa}}{3}\right)Bu^{3/2}+\frac{\kappa(\varepsilon)}{3}Au^2\right]$$
$$\leq 3e^{-u},$$

where

$$C^2 = \frac{1}{2}\int_0^t \int_0^t f(x,y)^2 d\Lambda_x \; d\Lambda_y,$$

$$D = \sup_{\int a_y^2 d\Lambda_y = 1, \int b_x^2 d\Lambda_x = 1} \int_0^t b_x \int_{(x,t)} a_y f(x,y) d\Lambda_y d\Lambda_x,$$

$$F = \mathbb{E}\left(\sup_{y \le t} \left| \int_0^t \mathbb{1}_{x<y} f(x,y) dM_x \right| \right),$$

and

$$B^2 = \sup_{y \le t} \int_0^t f(x,y)^2 d\Lambda_x.$$

where κ and $\kappa(\varepsilon)$ are given by Corollary 1 of [**56**].

Proof. We do the same computations with a continuous time. It's sufficient to replace Talagrand's inequality by Corollary (1.2.8) of Chapter 1 and Neveu's super-martingale by the corresponding Lemma derived by S. van de Geer in [**63**] or O. Kallenberg in [**37**]. ■

We can also have the equivalent of Theorem 6.

THEOREM 8. *For all $\varepsilon, x > 0$,*

$$\mathbb{P}(Z_t \ge 2(1+\varepsilon)^{3/2} C\sqrt{x} + 2\eta(\varepsilon)Dx + \beta(\varepsilon)Bx^{3/2} + \gamma(\varepsilon)x^2) \le 2.77 e^{-x}$$

where $\eta(\varepsilon) = 1.42\sqrt{\kappa}(2+\varepsilon+\varepsilon^{-1})$, $\beta(\varepsilon) = e(1+\varepsilon^{-1})^2\kappa(\varepsilon) + (1.42\sqrt{\kappa}(2+\varepsilon+\varepsilon^{-1})) \wedge \frac{(1+\varepsilon)^2}{\sqrt{2}}$ and $\gamma(\varepsilon) = (e(1+\varepsilon^{-1})^2\kappa(\varepsilon)) \wedge \frac{(1+\varepsilon)^2}{3}$ with κ and $\kappa(\varepsilon)$ given by Corollary 1.

The obvious applications of these inequalities would be to construct test for the Poisson intensity for instance. This would be a future possible work.

2.3. Exponential inequalities for counting processes

In this Section, we want to derive exponential inequalities for suprema of integral with respect to a counting process. We wish to find the same kind of concentration formula as in Theorem 2 but now for counting processes instead of Poisson processes. The aim is to be able to construct adaptive estimators of the intensity with these inequalities like in the Poisson framework. We explain this statistical application in Chapter 3.

Let $(N_t)_{t\ge 0}$ be a counting process, i.e. a random increasing piecewise constant function with $N_0 = 0$. Let $(\mathcal{F}_t)_{t\ge 0}$ be the filtration generated by N and $(\Lambda_t)_{t\ge 0}$ be its compensator i.e. the nondecreasing function such that $(M_t = N_t - \Lambda_t)_{t\ge 0}$ is a martingale.

One can find a very efficient exponential inequality in the book of O. Kallenberg [**37**]. Theorem 23.17 of this book says :

THEOREM 9. *Let $(Z_t)_{t\geq 0}$ be a local martingale with $Z_0 = 0$ and with jumps bounded by b smaller than 1. Suppose that a.s. $< Z >$ is bounded by 1. Then there exists some constant C such that for all r positive,*

$$\mathbb{P}(\sup_{t\geq 0} Z_t \geq r) \leq C \exp\{-\frac{1}{2} r \log(1 + rb)/b\}.$$

This result exists also for a supremum on $[0, T]$ for fixed positive T by stopping the martingale. It is also proved by S. van de Geer in **[63]** with precise constants.

In fact this result can be proved very easily in the case of counting process when Z_t denotes the martingale $\int_0^t H_s dM_s$ where H is a locally bounded predictable process. Theorem 9 is then very close in its form and also in the proof in this special case to the Bennett's inequality. We need some assumptions on the process to do this simple proof.

ASSUMPTION 1. *The compensator $(\Lambda_t)_{t\geq 0}$ is absolutely continuous and finite a.s. on $[0, T]$.*

The first part means that $d\Lambda_s$ is absolutely continuous with respect to the Lebesgue measure ds and the second part implies that N has a.s. a finite number of jumps. All the processes we consider in this PhD thesis verify these assumptions. It means in particular that there is no accumulation point for the jumps of the process on $[0, T]$.

The other classical exponential inequality in the i.i.d. framework is Bernstein's one. We can find a similar result for the integral martingale.

PROPOSITION 15. *Let Z be the process defined by :*

$$\forall t \geq 0, \quad Z_t = \int_0^t H_s dM_s$$

where H is a predictable process. Let T be a positive real number. If there exists c and v deterministic positive, such that

$$\forall k \geq 2, \left(\int_0^T H_s^k d\Lambda_s \right) \leq c^{k-2} v \frac{k!}{2}$$

then under Assumption 1, for all positive u,

$$\mathbb{P}\left(\sup_{[0,T]} Z_t \geq \sqrt{2vu} + cu \right) \leq \exp(-u).$$

One has also the following corollary.

COROLLARY 3. *Let Z be the process defined by :*

$$\forall t \geq 0, \quad Z_t = \int_0^t H_s dM_s$$

where H is a predictable process. Let T be a positive real number. If H and $< Z >$ are bounded respectively by b and v on $[0, T]$ with b, v deterministic positive, then under Assumption

1, for all positive u,

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

This last result is equivalent to the "weak Bernstein's" inequality. This is a corollary of Theorem 9 and also an obvious corollary of Proposition 15. Simple proofs of these results can be found in Paragraph 2.3.1, they are based on the existence of an exponential supermartingale. These results are proved in a more general framework by S. van de Geer **[63]**.

It is also possible to obtain some results equivalent to Talagrand's inequality (2.1.3) for counting processes.

PROPOSITION 16. *Let $\{(H_{a,t})_{t\geq 0}, a \in \mathcal{A}\}$ be a countable family of predictable processes. Let Z be the process defined by*

$$\forall t \geq 0, Z_t = \sup_{a\in\mathcal{A}} \left[\int_0^t H_{a,s} dM_s\right].$$

Let T be a positive real number. Under Assumption 1, the process $(Z_{t\wedge T})_{t\geq 0}$ has a positive nondecreasing compensator $(A_t)_{t\geq 0}$ and one has :

(a) if $\sup_{a\in A, t\leq T} |H_{a,t}|$ and $\int_0^T \sup_{a\in\mathcal{A}}[H_{a,s}^2]d\Lambda_s$ are bounded respectively by b and v, both deterministic positive numbers then for all u positive,

$$\mathbb{P}\left(\sup_{[0,T]} (Z_t - A_t) \geq \sqrt{2vu} + \frac{1}{3}bu\right) \leq \exp(-u);$$

(b) if there exists c and v deterministic positive, such that

$$\forall k \geq 2, \left(\int_0^T \sup_{a\in\mathcal{A}} |H_{a,s}|^k d\Lambda_s\right) \leq c^{k-2} v \frac{k!}{2}$$

then for all u positive,

$$\mathbb{P}\left(\sup_{[0,T]} (Z_t - A_t) \geq \sqrt{2vu} + cu\right) \leq \exp(-u).$$

We can compare these two inequalities (Talagrand's inequality (2.1.3) and Proposition 16) using the following correspondence :

- the empirical measure becomes the counting measure dN,
- expectation and compensator are equivalent,
- $n\mathrm{Var}(\psi(X))$ becomes consequently $\int \psi(s)^2 d\Lambda_s$.

In the case of Poisson process (see Theorem 2), expectation and compensator are the same if we integrate deterministic functions since the compensator of the process is deterministic. The correspondence is then more obvious.

If we assume this correspondence, it seems at first look that this new inequality is in some sense stronger than Talagrand's inequality (2.1.3) in the i.i.d. framework or Theorem 2 in the Poisson framework (which are up to constants equivalent) : we can manage random

(predictable) functions and one has also a "moment" version (see (b)), which does not assume that the family of functions to integrate (or to sum in the i.i.d. framework) is bounded.

The presence of the $\sup_{[0,T]}$ is just a refinement due to the martingale structure but does not affect the orders of magnitude.

However we loose some important fact with respect to Talagrand's inequality : let us compare the variance term v for each case. In (2.1.3), v can be seen as

$$v = \sup_{a\in\mathcal{A}} \mathbb{E}(\sum_{i=1}^{n} (\psi_a(X_i) - \mathbb{E}(\psi_a(X_i))^2).$$

The supremum is outside the sum but in Proposition 16(a), it lies inside the integral and is consequently of bigger order.

This phenomenon is underlined by P.-M. Samson **[61]**. He recovers Talagrand's inequality for Φ-mixing up to this exchange between the supremum and the sum.

For Poisson processes, L. Wu **[64]** and C. Houdré and N. Privault **[36]** use martingale approach to derive exponential inequalities for very general functionals of the process. When we apply these inequalities to supremum, it appears also this exchange in the variance term. To have supremum on the left hand side in the Poisson case **[56]**, we need some technics using the infinitely divisible property of the Poisson process.

The exchange between supremum and sum (or integral) seems consequently to be possible only when there exists some independence property in the problem.

The study of this supremum and the proof of Proposition 16 is made in Paragraph 2.3.2.

In Paragraph 2.3.3, we give some explanations about the statistical interest of such suprema and an useful application of these inequalities. We give also the orders of magnitude in a simple case in order to better explain the difference between the variance terms.

2.3.1. Basic exponential inequalities. We start like for the U-statistics by the existence of some super-martingale.

PROPOSITION 17.
Let $(H_t)_{t\geq 0}$ be a locally bounded predictable process and $(Z_t)_{t\geq 0}$ be defined by for all positive t, $Z_t = \int_0^t H_s dM_s$. Let $\phi(u) = e^u - u - 1$ for all u and finally

$$\forall t \geq 0, \quad E_t = \exp\left(\lambda Z_t - \int_0^t \phi(\lambda H_s) d\Lambda_s\right).$$

Let T be a positive real number. Let I be an interval such that for all λ in I, $\int_0^T e^{\lambda H_s} d\Lambda_s$ is a.s. finite, then under Assumption 1 $(E_{t\wedge T})_{t\geq 0}$ is a super-martingale and for all τ stopping time less than T, $\mathbb{E}(E_\tau)$is less than 1.

Proof. Let us fix some $\lambda \in I$. $(E_{t\wedge T})_{t\geq 0}$ is exactly the process $(L_t)_{t\geq 0}$ defined in Theorem VI-2 of **[17]** for $\mu_s = \exp(\lambda H_s)$ applied to the counting process $(N_{t\wedge T})_{t\geq 0}$ with compensator $\Lambda_{t\wedge T}$. As we verify all the assumptions of this theorem (with the notations of this theorem, μ_s is positive), we apply it and we obtain that $E_{t\wedge T}$ is a super-martingale ■

If N is a time Poisson process (i.e. Λ is deterministic), we can prove that $E_{t\wedge T}$ is a real martingale using Theorem II-8 of **[17]**, with $H_s = f(s)$ deterministic. Therefore we recover the classical expression for the Laplace transform :

$$\forall t \geq 0, \mathbb{E}\left(e^{\lambda \int_0^t f(s)dN_s - d\Lambda_s}\right) = \exp\left(\int_0^t \phi(\lambda f(s))d\Lambda_s\right).$$

For the other cases, the compensator is not deterministic, hence we cannot derive precise formula. This formula can be compared to its discrete time versions in Lemma VII-2-8 of **[52]** or in the article **[53]**. These two versions can be applied to the discrete time martingale $\sum_{i=1}^n f(X_i)$ where the X_i's are i.i.d. where f has zero mean : we recover Bernstein's and Bennett's inequalities. We can do the same thing here.

COROLLARY 4. *For all positive λ such that $\int_0^T e^{\lambda H_s} d\Lambda_s$ is a.s. finite and for all positive ε*

$$\mathbb{P}\left(\sup_{[0,T]} Z_t \geq \varepsilon\right) \leq e^{-\lambda\varepsilon} \exp\left|\left|\int_0^T \phi(\lambda H_s)d\Lambda_s\right|\right|_\infty .$$

Proof. Let $\tau = \inf\{t \leq T / Z_t > \varepsilon\}$: it is a stopping time. We get by positivity of ϕ that

$$\begin{aligned}
\mathbb{P}\left(\sup_{[0,T]} Z_t \geq \varepsilon\right) &= \mathbb{P}\left(Z_\tau \geq \varepsilon\right) \\
&= \mathbb{P}\left(\exp\left(\lambda Z_\tau - \int_0^\tau \phi(\lambda H_s)d\Lambda_s\right) \geq \exp\left(\lambda\varepsilon - \int_0^\tau \phi(\lambda H_s)d\Lambda_s\right)\right) \\
&\leq \mathbb{P}\left(\exp\left(\lambda Z_\tau - \int_0^\tau \phi(\lambda H_s)d\Lambda_s\right) \geq \exp\left(\lambda\varepsilon - \left|\left|\int_0^T \phi(\lambda H_s)d\Lambda_s\right|\right|_\infty\right)\right).
\end{aligned}$$

Then Markov inequality and Proposition 17 leads to the result. ■

We only have to apply this result and optimizing it in λ to recover Theorem 9 and Proposition 15. In the both cases, the assumption "$\int_0^T e^{\lambda H_s} d\Lambda_s$ a.s. finite" is an obvious consequence of the assumptions of boundedness or moment existence for H.

2.3.2. Suprema. Let $\{(H_{a,t})_{t\geq 0}, a \in \mathcal{A}\}$ be a countable family of locally bounded predictable processes. Let $(Z_t)_{t\geq 0}$ be defined by

$$\forall t \geq 0, Z_t = \sup_{a\in\mathcal{A}} \left[\int_0^t H_{a,s} dM_s\right].$$

Hence $(Z_t)_{t\geq 0}$ is an adapted process with bounded variations.

Under Assumption 1, the jumps of Z happen only when N jumps. Let T be a fixed positive number. For all t less than T, let us denote by $(T_i, 1 \leq i \leq n_t)$ the ordered jumps of N before t : there exists a.s. a finite number of these jumps as consequence of Theorem II-8 (α) of **[17]** and Assumption 1. Consequently we can write :

$$\forall t \leq T, \quad Z_t = \sum_{T_i \leq t} [Z_{T_i} - Z_{T_i -}] + Z_{t-} - Z_{T_{n_t}} + \sum_{T_i \leq t} [Z_{T_i -} - Z_{T_{i-1}}] \text{ a.e.} \tag{2.3.1}$$

where $Z_{T_0} = Z_0 = 0$.

LEMMA 12. *Assume $\mathcal{A} = \{1...k\}$ finite. Let $i \geq 1$ be some integer. Let v be a real number in $]T_{i-1}; T_i[$. Then under Assumption 1, $Z_v - Z_{T_{i-1}} = \int_{T_{i-1}}^{v} -H_{\hat{a}_{s-},s} d\Lambda_s$. where $\hat{a}_{s-}$ is the first index where Z_{s-} is achieved.*

Proof. Let us denote by f the Radon-Nikodym derivative $d\Lambda_t/dt$ which exists by Assumption 1. Then we can write for all v in $]T_{i-1}; T_i[$, $Z_v = \sup_{a \in \mathcal{A}} g_a(v)$ where $g_a(v) = \left[b_a + \int \mathbb{1}_{]T_{i-1},v]} f_a(s) ds\right]$ and where $f_a(s) = -H_{a,s} f(s)$. As the g_a's are absolutely continuous and with normalized bounded variations and $\mathcal{A}$ is finite, Z is also absolutely continuous and with normalized bounded variations. Consequently **[60]**, Z is almost surely differentiable and $Z_v = \int \mathbb{1}_{]T_{i-1},v]} Z'_s ds$. So we have to compute its derivative to conclude. Let us restrict ourselves to the set of v's (with plain measure) where $\forall a \in \mathcal{A}, g_a(v)' = f_a(v)$. For u also in $]T_{i-1}; T_i[$, we have the following inequalities :

$$\int_{]u,v]} f_{\hat{a}_u}(s) ds \leq Z_v - Z_u \leq \int_{]u,v]} f_{\hat{a}_v}(s) ds.$$

Let us divide by $v - u$. We do the proof for the left derivative. The same proof can be done for the right derivative. We take the limit when $u \uparrow v$ and we obtain :

$$f_{\hat{a}_{v-}}(v) \leq \liminf_{u \uparrow v} \frac{Z_v - Z_u}{v - u} \leq \limsup_{u \uparrow v} \frac{Z_v - Z_u}{v - u} \leq f_{\hat{a}_v}(v). \tag{2.3.2}$$

If the inequality is strict in (2.3.2) for v_0, it means that "$f_{\hat{a}_{v_0-}}(v_0) < f_{\hat{a}_{v_0}}(v_0)$". Hence we are in the following case : "$g_{\hat{a}_{v_0-}}(v_0) = g_{\hat{a}_{v_0}}(v_0)$ (else there is no reason to change the index where the supremum is achieved) and on a neighborhood of the form $]v_0 - \varepsilon, v_0]$ for ε positive well chosen, we have "$g_{\hat{a}_{v_0-}}(u) > g_{\hat{a}_{v_0}}(u)$" because the functions are differentiable in v_0 : v_0 is then isolated. (In order to be absolutely rigorous, we have to make a recurrence on the cardinality of $\mathcal{A}$.) Consequently, the set of v such that the inequality is strict is with zero measure (it is countable). Then we get that almost everywhere "$Z'_v = f_{\hat{a}_{v-}}(v)$". This concludes the proof. ■

Using this lemma we get the following proposition.

PROPOSITION 18. *Let T be a fixed positive number. Under Assumptions 1, one has if $\mathcal{A}$ is finite :*

$$\forall t \leq T, \quad Z_t = \int_0^t \Delta Z(s) dN_s - \int_0^t H_{\hat{a}_{s-},s} d\Lambda_s \text{ a.s.}$$

where $\Delta Z(s) = \sup_{a\in\mathcal{A}}\left[H_{a,s} + \int_0^{s-} H_{a,u}dM_u\right] - \sup_{a\in\mathcal{A}}\left[\int_0^{s-} H_{a,u}dM_u\right]$. *Consequently, the compensator of* $Z_{t\wedge T}$ *is*

$$A_t = \int_0^{t\wedge T} \left[\Delta Z(s) - H_{\hat{a}_{s-},s}\right] d\Lambda_s.$$

If $\mathcal{A}$ *is just countable, the compensator of* $Z_{t\wedge T}$, A_t *exists, is positive and nondecreasing and*

$$\forall t \leq T, \quad Z_t - A_t = \int_0^t \Delta Z(s) dM_s.$$

Proof. Assume $\mathcal{A}$ finite. The first integral in Z_t is exactly the first part in (2.3.1). For the second part, all the differences are between two consecutive jumps and we use the previous Lemma. Moreover $\Delta Z(s)$ introduced in the proposition is predictable. The compensator is then obvious. As $\Delta Z(s) - H_{\hat{a}_{s-},s}$ is positive and Λ nondecreasing, A is positive nondecreasing.

If $\mathcal{A}$ is just countable, $\mathcal{A}$ is an increasing union of finite sets B_n. Let us denote by Z^n the supremum over B_n instead of $\mathcal{A}$. As for all n B_n is finite, Z^n verifies the first part of the proposition. But, for all t less than T, $Z_t^n - \int_0^t \Delta Z^n(s) dN_s$ which is predictable, converges almost surely to $Z_t - \int_0^t \Delta Z(s) dN_s$ which is X_t : X_t is consequently also predictable. Then obviously, $A_t = \int_0^t \Delta Z(s) d\Lambda_s + X_t$ is the compensator of $Z_{t\wedge T}$ and it stays positive nondecreasing as a limit of positive nondecreasing functions.

■

To derive Proposition 16, it is sufficient to apply Proposition 15 or Corollary 3 to $Z - A$ noticing that ΔZ is bounded by $\sup_{a\in\mathcal{A}} |H_a|$.

2.3.3. Statistical applications.

These exponential inequalities are useful to provide exponential deviations for the following χ^2-type statistics. Let T be a fixed positive real number and let $\{h_\lambda, \lambda \in m\}$ be a finite family of predictable processes. We set

$$\chi_T^2 = \sum_{\lambda\in m} \left(\int_0^T h_\lambda(s) dM_s\right)^2. \tag{2.3.3}$$

The typical case is Gaussian model selection in the white noise framework [**14**]. One has a model i.e. a linear finite dimensional subspace with orthonormal basis $\{\varphi_\lambda, \lambda \in m\}$ for the classical scalar product on $[0, T]$. The classical projection estimator on this subspace has a variance term of the previous form with φ_λ instead of h_λ i.e. deterministic functions and dW the white noise instead of dM. In this case, this variance term is a real χ^2-statistics. We must understand and control the deviation of this variance term to manage a lot of models and compare them. In the Gaussian framework, we can use classical exponential inequalities on the χ^2-statistics.

When we want to estimate density from a n-sample, we obtain the same form of χ^2-statistics with the $h_\lambda = \varphi_\lambda$'s deterministic functions, orthonormal basis of a model for the classical scalar product. In this context, L. Birgé and P. Massart use Talagrand's inequality to provide control on the χ^2-type statistics [**12**].

In the Poisson case, the $h_\lambda = \varphi_\lambda$ are always deterministic, forming an orthonormal family of $\mathbb{L}^2([0,1],dt)$ and we can use Theorem 2, which give the same order as in the n-sample framework.

More generally, we can look at the Aalen multiplicative intensity model where the compensator of N verifies $d\Lambda = Ys(t)dt$ with Y predictable and known. For instance the censorship framework verifies this model. We want to estimate the deterministic function s using the observations of the process N. We can also do it by model selection in Chapter 3. In this case we are using a random scalar product $\int_0^t fgYdt$ instead of the classical one for the Poisson process when Y is constant. In this context the h_λ's become (in the good cases) predictable.

Indeed, if one has $\{\varphi_\lambda, \lambda \in m\}$ orthonormal family of $\mathbb{L}^2([0,1],dt)$ (typically histograms or Fourier basis), $\left\{h_\lambda = \varphi_\lambda/\sqrt{Y}, \lambda \in m\right\}$ becomes an orthonormal family for the random product (when Y is positive) and the h_λ's are predictable. The model is no more deterministic but predictable.

To provide concentration inequalities for χ^2_T, we can remark that

$$\forall t \geq 0, \quad \chi_t = \sup_{\sum_{\lambda\in m} a_\lambda^2 = 1} \int_0^t \left(\sum_{\lambda\in m} a_\lambda h_\lambda(s)\right) dM_s \tag{2.3.4}$$

Consequently we can use Proposition 16 on a countable dense subset of the unit ball of $\mathbb{R}^m$. But as we do not know in practice the compensator of $(\chi_t)_{t\geq 0}$ we want to compare it to $(\sqrt{C_t})_{t\geq 0}$ where

$$\forall t \geq 0, \quad C_t = \sum_{\lambda\in m} \int_0^t h_\lambda(s)^2 d\Lambda_s \tag{2.3.5}$$

is the compensator of χ^2_t. Finally we can obtain the following result.

PROPOSITION 19 (An inequality which is ready for immediate application). *Let T be a fixed positive real number. Let χ_T be defined by (2.3.3). Then, for all u positive, with probability larger than $1 - 2e^{-u}$,*

$$\chi_T - \sqrt{C_T} \leq 3\sqrt{2vu} + bu$$

where

- *C_T is defined by (2.3.5);*
- *$v = \|C_T\|_\infty$;*
- *b is a positive deterministic number verifying for all s less than T, $\sum_{\lambda\in m} h_\lambda^2(s) \leq b^2$.*

Proof. Let u be positive. At first we can interpret χ_t by a supremum (see (2.3.4)). But moreover, we can take B countable dense subset of the unit ball of $\mathbb{R}^m$ and say that

$$\chi_t = \sup_{a\in B} \int_0^t \left(\sum_{\lambda\in m} a_\lambda h_\lambda(s) \right) dM_s. \tag{2.3.6}$$

Then we can apply Proposition 16(a) with $H_a = \sum_{\lambda\in m} a_\lambda h_\lambda$. We obtain that χ_t has a compensator A_t and

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

We can replace the H_a by $-H_a$ to obtain

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

Let $B_T = \sup_{[0,T]} |\chi_t - A_t|$. Now we must compare A and C. One has for all t less than T :

$$\begin{aligned} \chi_t^2 - A_t^2 &= (\chi_t - A_t)^2 + 2A_t(\chi_t - A_t) \\ &= (\chi_t - A_t)^2 + 2\int_0^t (\chi_{s-} - A_{s-}) dA_s + 2\int_0^t A_s d(\chi_s - A_s). \end{aligned}$$

But the first term has for compensator $\int_0^t (\Delta\chi)^2(s) d\Lambda_s$ and the last term is a martingale. Moreover A is predictable, thus we can take the compensator of the previous expression to obtain :

$$C_t - A_t^2 = \int_0^t (\Delta\chi)^2(s) d\Lambda_s + 2\int_0^t (\chi_{s-} - A_{s-}) dA_s.$$

Consequently, one has :

$$\begin{aligned} \chi_t - \sqrt{C_t} &= \chi_t - A_t + A_t - \sqrt{C_t} \\ &= \chi_t - A_t - \frac{C_t - A_t^2}{A_t + \sqrt{C_t}} \\ &= \chi_t - A_t - \frac{\int_0^t (\Delta\chi)^2(s) d\Lambda_s + 2\int_0^t (\chi_{s-} - A_{s-}) dA_s}{A_t + \sqrt{C_t}}. \end{aligned}$$

As A is positive and nondecreasing, one gets $\chi_t - C_t \leq 3B_T$, for all t less than T which implies the result. ■

Orders of magnitude

The inequality available for Poisson processes (Corollary 1) is exactly the same orders of magnitude as in Talagrand's inequality in the i.i.d. framework.

Let us compare it to Proposition 19 applied to Poisson process. We are in the case where s is a constant equal to 1, Y is a constant equal to n and $T = 1$ (i.e. $d\nu = nds$ is the mean measure of the Poisson process). This is the case for the aggregated process build from the sum of n i.i.d. homogeneous Poisson processes on $[0, 1]$ with intensity 1.

Assume that the model is the set of histograms constructed on a regular partition m of $[0,1]$. Then the basis is deterministic and it is of the form $\sqrt{(D/n)}\mathbb{1}_I$ for I in m where D is the number of intervals in m.

If we apply Corollary 1 on $[0,1]$, to

$$\chi = \sup_{\sum_{I\in m} a_I^2=1} \int_0^1 \sum_{I\in m} a_I \sqrt{\frac{D}{n}} \mathbb{1}_I (dN_s - nds) = \sqrt{\sum_{I\in m} \frac{D}{n} \left(N_I - \frac{n}{D}\right)^2},$$

we obtain, for all u and ε positive numbers,

$$\mathbb{P}\left(\chi \geq (1+\varepsilon)\sqrt{D} + \sqrt{2\kappa u} + \kappa(\varepsilon)\sqrt{\frac{D}{n}}u\right) \leq \exp -u. \tag{2.3.7}$$

But, if we apply Proposition 19, we obtain, for all u positive :

$$\mathbb{P}\left(\chi \geq \sqrt{D} + 3\sqrt{2Du} + \sqrt{\frac{D}{n}}u\right) \leq 2\exp -u. \tag{2.3.8}$$

The variance term in (2.3.8) is bigger than the corresponding term in (2.3.7). It is of the same order than the expectation ($\sqrt{D}$) and no more a complementary term. We must mention that D can become very big : it can be as tall as n for some model and to have good estimation properties n grows to infinity. In this sense, Inequality (2.3.7) is better than (2.3.8).

But for more general processes, Y is no more constant : it often decreases and can become very small. When t tends to 1, Y_t is equal to 1 in the right censorship framework for instance and when in 0, Y_0 is n, the number of observations. In this case, the third linear term "bx" becomes of order $\sqrt{D}$ too. Consequently these two types of inequalities lead to the same order Du^2 for the χ^2-type statistics, in this framework and no more D as in the Poisson framework (2.3.7). This leads us to manage only a little family of possible models (see Chapter 3) when we do some model selection in the Aalen multiplicative intensity framework.

CHAPITRE 3

Penalized projection estimators for the Aalen's multiplicative intensity

Abstract

We want to study the problem of non parametric estimation of the intensity of counting processes satisfying the Aalen's multiplicative intensity model. We use for this aim model selection technics and more precisely penalized projection estimators for a random scalar product. For histogram estimators, under some assumptions on the process, we obtain adaptive results for the minimax risk. Generally, for more intricate (predictable) models, we have only oracle inequalities.

AMS Classification : 62G07, 62M09.
Keywords : penalized projection estimators, model selection, counting processes, multiplicative intensity model.

Résumé

Dans ce chapitre, nous nous intéressons à l'estimation non paramétrique de l'intensité multiplicative d'Aalen de certains processus de comptage. Nous utilisons pour cela des techniques de sélection de modèles et plus précisément des estimateurs par projection pénalisés pour un produit scalaire aléatoire. Pour des estimateurs en histogrammes, sous certaines hypothèses sur le processus, on obtient des résultats adaptatifs pour le risque minimax. De manière générale, quand les modèles sont plus compliqués et en particulier prévisibles, nous obtenons seulement des inégalités d'oracle.

Classification AMS : 62G07, 62M09.
Mots Clefs : estimateurs par projection pénalisés, sélection de modèles, processus de comptage, intensité multiplicative d'Aalen.

3.1. Introduction

We want to estimate the intensity of a counting process when it verifies the Aalen's multiplicative intensity model. Moreover we want to obtain a non parametric and adaptive estimation procedure.

At first, let us present these processes. Let $(N_t)_{t\geq 0}$ be a counting process i.e. a nondecreasing random piecewise constant function with $N_0 = 0$. This process generates classically a filtration $(\mathcal{F}_t)_{t\geq 0}$ and with respect to this filtration, $(N_t)_{t\geq 0}$ has a compensator $(\Lambda_t)_{t\geq 0}$ i.e. a nondecreasing random function such that $(M_t = N_t - \Lambda_t)_{t\geq 0}$ is a martingale.

The counting process $(N_t)_{t\geq 0}$ verifies the Aalen's multiplicative intensity model if we can write :

$$d\Lambda_t = Y_t s(t) dt, \tag{3.1.1}$$

where $(Y_t)_{t\geq 0}$ is a nonnegative predictable process and where s is a deterministic function.

We want to estimate the intensity s on $[0, \tau]$ using the observations of $(N_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$. By rescaling, we can restrict ourselves to the problem of estimating s on $[0, 1]$.

Let us give some examples.

The easiest case is the time Poisson process. It corresponds to the case where Y (3.1.1) is constant. Then the quantity N_t may for instance represent the number of machine breakdowns before time t. Many people are interested in the non parametric estimation of the intensity of a Poisson process : let us mention the work of M. Rudemo **[59]** for histogram and kernel estimators, the work of W.-C. Kim and J.-Y. Koo **[39]** for wavelet estimators and the work of L. Cavalier and J.-Y. Koo **[22]** for thresholding procedures. We have treat this case in Chapter 1.

Another simple example is the process defined by $(N_t = \mathbb{1}_{X\leq t}, t \geq 0)$ where X is a positive random variable with density f. This process has only one jump and verifies (3.1.1) with $Y_t = \mathbb{1}_{X\geq t}$ and with $s(t) = f(t)/\mathbb{P}(X \geq t)$. One calls s the **hazard rate** of X. If X represents the life time of some patient, $s(t)$ represents the probability to stay alive after t if one was alive in t.

The observations of life times may be sometimes censored. This is the case when the patient goes out the hospital study. We cannot observe the death time : we just know that he was alive when he leaves the study. This situation is modeled by some other positive variable U independent of X and we just observe the variables $T = X \wedge U$ and $D = \mathbb{1}_{T=X}$. This model is known as the **right-censoring model** with independent censorship.

If we dispose of a n-sample of counting processes, $N^1, ..., N^n$ with corresponding predictable processes $Y^1, ..., Y^n$ and with same intensity s, we can study the **aggregated process** N with predictable process Y defined by

$$N_t = \sum_{i=1}^{n} N_t^i \text{ and } Y_t = \sum_{i=1}^{n} Y_t^i \text{ for all } t \geq 0. \tag{3.1.2}$$

For instance in the right-censoring model, the aggregated Y is a nonincreasing process with integer values and with $Y_0 = n$ the number of observations. At time t, Y_t is the number of events which happen after t.

The problem of estimating the hazard rate in this model is well known. For instance, let us mention the work of A. Antoniadis, G. Grégoire and G. Nason [**5**] for wavelets estimators on sieves and the reference therein and the work of C. Kooperberg, C.J. Stone and Y.K. Truong [**42**] for splines estimators. Moreover there exists some adaptive procedures due to S. Döhler and L. Rüschendorf in [**26**] by model selection methods. The same method is used by G. Castellan and F. Letué [**21**] for the Cox model with right-censorship.

There exists a lot of other examples of processes with multiplicative intensity in the book of P.K. Andersen, O. Borgan, R. Gill and N. Keiding [**2**]. For instance, if $(X_t)_{t\geq 0}$ is a Markov process with finite state space, the counting process $(N_t^{hj}, t \geq 0)$, where N_t^{hj} represents the number of transitions from h to j before t, has a multiplicative intensity of the form (3.1.1) where s is the transition intensity from h to j and where Y is defined by $(Y_t = \mathbb{1}_{X(t)=h}, t \geq 0)$. We can also dispose of an i.i.d. n-sample of counting processes corresponding to each individual Markov processes and aggregate them by (3.1.2). The aggregated Y is still an integer bounded by n. At time t, Y_t represents the number of individuals in state h. This situation models for instance the healthy-diseased transitions (see Example I.3.10 in [**2**] and many others in this book.)

There exists also cases where the process can not be divided in individual processes. Consequently, it cannot be written as in (3.1.2). This is the case for the model of matings for Drosophila flies proposed in Example III.1.10 of [**2**]. However it verifies the multiplicative intensity model (3.1.1) with a Y, which in some sense corresponds always to a number of events which may arrive after time t and which is bounded by a known number.

A lot of papers consider the problem of estimating s in the general Aalen's multiplicative intensity model. Let us mention some of them. H. Ramlau-Hansen [**54**] proves consistency and asymptotic normality results for some kernel estimators. G. Grégoire [**32**] chooses the bandwidth of the Ramlau-Hansen estimators by a kind of cross-validation and proves also consistency and asymptotic normality results. A. Antoniadis [**4**] proves the same kind of results for penalized maximum likelihood estimators, the penalization depending on the regularity of the functions. In [**63**], S. Van de Geer studies the rates of convergence of maximum likelihood estimators.

We want to construct estimators of the intensity with as few assumptions on the process and on the intensity as possible. In fact, we would like to give some generalization of the results of S. Döhler and L. Rüschendorf in [**26**] to the general Aalen's multiplicative intensity model. We would like to give also non asymptotic results which prove the adaptivity of the resulting estimators.

The method is inspired by the work of L. Birgé and P. Massart **[12]** on penalized projection estimators (**p.p.e.**) for density estimation. Let us describe it.

At first, we need however some classical assumptions on the process :

ASSUMPTION 2. *Y is bounded by a known A.*

In the right-censoring model or in the Markovian model, $A = n$ represents the number of observations.

Now let us define the least-square contrast in our context : for all f in $\mathbb{L}^2([0,1], dt)$,

$$\gamma_A(f) = -2\int_0^1 f(t)\frac{dN_t}{A} + \int f^2(t)Y'_t dt, \tag{3.1.3}$$

where $Y'_t = Y_t/A$. This is not the contrast used by S. Döhler and L. Rüschendorf. They use a log-likelihood contrast which is much more intricate to deal with than the least-square one, even if it is famous and gives good results.

The projection estimator of s on a finite dimensional linear subspace S is defined by

$$\hat{s} = \operatorname*{argmin}_{f\in S} \gamma_A(f). \tag{3.1.4}$$

The associated random norm is defined for all f in $\mathbb{L}^2([0,1], dt)$ by

$$\|f\|^2_{\text{rand}} = \int_0^1 f^2(t)Y'_t dt. \tag{3.1.5}$$

If $\{h_\lambda, \lambda \in \Gamma\}$ is an orthonormal basis of S for the random norm, we can simplify the projection estimator and write :

$$\hat{s} = \sum_{\lambda\in\Gamma} \left(\int_0^1 h_\lambda(t)\frac{dN_t}{A}\right) h_\lambda. \tag{3.1.6}$$

We want to estimate s by a projection estimator. Thus we have to choose correctly the space S. If we want to do some adaptive estimation, this space or **model** must be chosen via a data driven criterion. For this aim we introduce a family of models (finite dimensional linear subspaces) $\{S_m, m \in \mathcal{M}_A\}$ and we associate to each S_m the projection estimator $\hat{s}_m$ of s on it. We find a good model via a penalty pen, which is a positive function on $\mathcal{M}_A$, if necessary random. Then it is sufficient to minimize the following data driven criterion :

$$\hat{m} = \arg\min_{m\in\mathcal{M}_A} (\gamma_A(\hat{s}_m) + \text{pen}(m)) \tag{3.1.7}$$

and the penalized projection estimator (p.p.e.), $\tilde{s}$ is defined by $\hat{s}_{\hat{m}}$.

For the density estimation **[12]**, the models to handle verify that the infinite norm of the functions f in the model S_m is well controlled by their $\mathbb{L}^2$ norm. Here we have the same problem and we need the same control between the infinite norm and the random norm.

Consequently, we have restrict ourselves to the cases where we know an orthonormal basis of the models for the random norm to check on this basis that the coefficients in the

orthonormal decomposition have good links with the infinite norm of the functions . Hence we deal with two cases.

The first one is the histogram case. The basis is obvious but we must restrict ourselves to aggregated processes (3.1.2) to be able to control the variance of the estimator. We treat this case in Section 3.2.

Other deterministic models do not vouch for a good link between their orthonormal basis for the random norm and the infinite norm in the model.

Consequently, the second case deals with random predictable models. They are built as follows. If $\{\varphi_\lambda, \lambda \in m\}$ is a classical deterministic orthonormal basis of $\mathbb{L}^2([0,1], dt)$ typically a part of a Fourier basis, then $\{h_\lambda = \varphi_\lambda/\sqrt{Y'}, \lambda \in m\}$ becomes, when Y' is positive, an orthonormal basis for the random product. Hence the model $S_m = \text{Span}\{h_\lambda, \lambda \in m\}$ is a random predictable subspace where we can ensure good links between infinite norm and random norm if the $\{\varphi_\lambda, \lambda \in m\}$ are well chosen (Fourier or wavelet basis). One of the advantage of these models is to allow us to remove the aggregation assumption among many other technical assumptions. We manage this example in Section 3.3.

In both cases, we want at least to prove oracle type inequalities, i.e. some inequality of the form :

$$\|s - \tilde{s}\|^2 \leq C \inf_{m \in \mathcal{M}_A} \left(\|s - s_m\|^2 + \text{pen}(m) \right), \tag{3.1.8}$$

for some positive constant C where s_m is the projection of s on S_m for $\|.\|$.

The norm $\|.\|$ can be either the random norm $\|.\|_{\text{rand}}$ or the deterministic norm defined by : for all f in $\mathbb{L}^2([0,1], dt)$

$$\|f\|^2_{\text{det}} = \int_0^1 f^2(t) \mathbb{E}(Y'_t) dt. \tag{3.1.9}$$

As for (3.1.8), we may have this inequality either in probability or in expectation, with if necessary the addition of some negligible term.

In both cases, if

$$\|s - s_m\|^2 + \text{pen}(m) \simeq \|s - \hat{s}_m\|,$$

we say that one has a true oracle inequality. This means that up to some constant the p.p.e. $\tilde{s}$ does as well as the best possible estimator in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$ without knowing s. This proves the adaptivity of the p.p.e. in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$.

The last section is devoted to simulation studies of those two strategies in the right-censoring model.

3.2. Histogram quasi-least square estimators

In this section, the purpose is to manage deterministic histogram models. For this aim, we assume that N is an aggregated process (see (3.1.2)), with predictable process Y

and with individual processes $N^1, ..., N^n$ and $Y^1, ..., Y^n$. This is the case for censoring or Markovian models for instance.

Furthermore, we assume those two facts.

ASSUMPTION 3.

Each Y^i is bounded by 1.

The number of individual jumps of the N^i is bounded by a known positive number K.

In the right-censoring model, one has $K = 1$ for instance. However Markovian models do not verify the last assumption. Under these assumptions, A defined by Assumption 2 is equal to n.

We will also assume that there exists R unknown such that s is bounded by R.

Remark : If the Y^i's are just bounded by a known B, it is sufficient to divide the Y^i's by B and to estimate Bs to recover this case.

We can compare these assumptions with those of G. Grégoire **[32]**. He assumes N to be aggregated and Y to be bounded too. He does not assume a bound on N. Thus he can manage Markovian models but he assumes that n/Y is bounded by something which does not depend on n. We do not need this assumption here.

3.2.1. Study on one model. Under these assumptions, the least-square contrast becomes

$$\gamma_A(f) = -\frac{2}{n}\int_0^1 f(t)dN_t + \frac{1}{n}\int_0^1 f^2(t)Y'_t dt.$$

Now, let us construct the projection histogram estimator.

Let m be a partition of $[0, 1]$. For all I in m, we denote

$$b_I = (1/n)\int_0^1 \mathbb{1}_I Y_t dt.$$

Then the family $\{\mathbb{1}_I/\sqrt{b_I}, I \in m\}$ is an orthonormal family for the random norm $\|.\|_{\text{rand}}$.

We can notice that b_I depends only on the observations. Let β_I be $\mathbb{E}(b_I)$ and let N_I be the number of points of N lying in I.

Let $\mathcal{I}_m$ be the set of intervals I in m such that the b_I's are larger than $1/n^2$.

The **quasi-least square** histogram estimator on S_m, the space of the piecewise constant functions on m, is the projection estimator of s defined by (3.1.4)on S'_m the set of piecewise constant functions on m, null outside $\mathcal{I}_m$. We can rewrite it using (3.1.6) :

$$\hat{s}_m = \sum_{I\in\mathcal{I}_m} \frac{N_I}{nb_I}\mathbb{1}_I. \tag{3.2.1}$$

It is more convenient to deal with this quasi-least square estimator (i.e. the projection estimator of s on S'_m) than with the projection estimator of s on S_m, because $\hat{s}_m$ is bounded.

Risk of the quasi-least square estimator :

Let us denote by s_m the projection of s on S_m for the random scalar product :

$$s_m = \sum_{I \in m} \frac{a_I}{b_I} \mathbb{1}_I$$

where

$$a_I = (1/n) \int_0^1 \mathbb{1}_I Y_t s(t) dt.$$

Note that if $b_I = 0$ then $a_I = 0$ and the corresponding coefficient of s_m is null. We denote by α_I, $\mathbb{E}(a_I)$ and by s'_m the projection of s on S'_m :

$$s'_m = \sum_{I \in \mathcal{I}_m} \frac{a_I}{b_I} \mathbb{1}_I.$$

Finally we denote by s_m^{det} the projection of s on S_m for the deterministic scalar product :

$$s_m^{\text{det}} = \sum_{I \in m} \frac{\alpha_I}{\beta_I} \mathbb{1}_I.$$

The distance between s and $\hat{s}_m$ can be divided in the following way :

$$\|s - \hat{s}_m\|^2_{\text{rand}} = \|s - s'_m\|^2_{\text{rand}} + \|s'_m - \hat{s}_m\|^2_{\text{rand}}. \tag{3.2.2}$$

The first term is a bias term. This is here a random term, unlike the density estimation (see **[12]**). We can bound it by :

$$\begin{aligned} \|s - s'_m\|^2_{\text{rand}} &= \|s - s_m\|^2_{\text{rand}} + \|s_m - s'_m\|^2_{\text{rand}} \\ &\leq \inf_{t \in S_m} \|s - t\|^2_{\text{rand}} + \frac{R^2}{n}, \end{aligned} \tag{3.2.3}$$

assuming that there is less intervals than A i.e. $|m| \leq n$. Consequently the expectation of the bias term is less than

$$\mathbb{E}(\|s - s'_m\|^2_{\text{rand}}) \leq \inf_{s \in S_m} \mathbb{E}(\|s - t\|^2_{\text{rand}}) + \frac{R^2}{n} = \|s - s_m^{\text{det}}\|^2_{\text{det}} + \frac{R^2}{n},$$

which diminishes when the intervals of the partition become small.

The behavior is very different for the second term in (3.2.2) whose expectation is called the variance term. Let us denote for a set of intervals J :

$$\chi^2_J = \sum_{I \in J} \frac{(\frac{N_I}{n} - a_I)^2}{b_I}. \tag{3.2.4}$$

Then the second term in (3.2.2) is exactly $\chi^2_{\mathcal{I}_m}$. If we assume that the b_I's are close to their expectation denoted by β_I and supposed to be non zero, we can identify $\chi^2_{\mathcal{I}_m}$, χ^2_m and finally Z^2_m where

$$Z^2_m = \sum_{I \in m} \frac{(\frac{N_I}{n} - a_I)^2}{\beta_I}. \tag{3.2.5}$$

Let us assume also that Z_m^2 is close to its expectation which is

$$\mathbb{E}(Z_m^2) = \sum_{I \in m} \frac{\alpha_I}{n\beta_I}.$$

Let $|m|$ be the number of intervals in m. Then this expectation lies between $r|m|/n$ and $R|m|/n$ if s is upper bounded by R and lower bounded by r.

If all the previous identifications are licit, the variance term must grow like the dimension of the model S_m when the bias term decreases.

3.2.2. Penalized least square histograms. In order to find a good partition, which minimizes this distance, we have to make some compromise between the bias term and the variance term.

Let $\{S'_m, m \in \mathcal{M}_A\}$ be the family of models corresponding to the family of partitions $\mathcal{M}_A$ of $[0,1]$.

The best partition or the best model, the one which we would choose if we know s is called the **oracle** and is defined by :

$$\begin{aligned} \bar{m} &= \arg\min_{m \in \mathcal{M}_A} \mathbb{E}(\|s - \hat{s}_m\|_{\text{rand}}^2) \\ &= \arg\min_{m \in \mathcal{M}_A} \mathbb{E}(-\|s'_m\|_{\text{rand}}^2 + \|s'_m - \hat{s}_m\|_{\text{rand}}^2) \\ &\simeq \arg\min_{m \in \mathcal{M}_A} \mathbb{E}(-\|\hat{s}_m\|_{\text{rand}}^2 + 2\|s'_m - \hat{s}_m\|_{\text{rand}}^2) \\ &\simeq \arg\min_{m \in \mathcal{M}_A} \mathbb{E}(-\|\hat{s}_m\|_{\text{rand}}^2 + 2\chi_m^2). \end{aligned} \tag{3.2.6}$$

The symbol $\simeq$ means that the expectations are not equal but if the coefficients of $\hat{s}_m - s'_m$ are close to zero, the expectations are close to each other.

Moreover we has,

$$\|\hat{s}_m\|_{\text{rand}}^2 = -\gamma_A(\hat{s}_m).$$

Consequently, if we do some estimation of the previous quantities, we are going to choose the model $\hat{m}$ given by (3.1.7) with the penalty that verifies "$\text{pen}(m)/2$ is an estimate of the variance term".

The equation (3.2.6) corresponds to the ISE minimization done by G. Grégoire for kernel estimators. He does an unbias estimation of the risk by a leave-one out procedure.

Here we choose the partition by the general penalized data-driven criterion given in (3.1.7). We prove in the next paragraphs oracle type inequality for penalties which overestimate the variance term.

3.2.3. Control of the chi-square statistic. Consequently, to prove that a penalty is well chosen, we need to understand the behavior of the $\chi^2_{\mathcal{I}_m}$'s. But they are very difficult to control. Thus we bound them by

$$\chi^2_{\mathcal{I}_m} \leq Z_m^2 V_m \text{ for all } m \in \mathcal{M}_A$$

where Z_m^2 is given by (3.2.5) and where $V_m = \sup_{I\in m} \frac{\beta_I}{b_I}$.

Moreover, the square root of Z_m^2 can be seen as

$$Z_m = \sup_{\delta/\sum_{I\in m}\delta_I^2\beta_I=1} \left(\frac{1}{n}\sum_{i=1}^{n} \int_0^1 \left(\sum_{I\in m} \delta_I \mathbb{1}_I(t) \right) (dN_t^i - Y_t^i s(t)dt) \right).$$

Consequently we can apply Talagrand's inequality. We use the recent version of this result due to E. Rio [**58**].

PROPOSITION 20. *Under Assumptions 3.1.2 and 3, for all $\varepsilon, x > 0$,*

$$\mathbb{P}\left(Z_m \geq (1+\varepsilon)\sqrt{\sum_{I\in m}\frac{\alpha_I}{n\beta_I}} + \sqrt{2v_m\frac{x}{n}} + (1/2+\varepsilon^{-1})b\frac{K+R}{n}x \right) \leq e^{-x}$$

where

$$b = \sup_{I\in m}\frac{1}{\sqrt{\beta_I}}, \qquad R \geq \|s\|_\infty$$

and where

$$v_m = \sup_{\delta/\sum_{I\in m}\delta_I^2\beta_I=1} \int_0^1 \sum_{I\in m} \delta_I^2 \mathbb{1}_I(t)\mathbb{E}(Y_t^1)s(t)dt$$

is bounded by

$$R_m = \sup_{I\in m}\frac{\alpha_I}{\beta_I}.$$

Proof. We apply Theorem 1.4 of [**50**] to

$$''X''_{i,\delta} = \frac{1}{n}\int_0^1 (\sum_{I\in m}\delta_I \mathbb{1}_I(t))(dN_t^i - Y_t^i s(t)dt)$$

which are centered variables. It is very easy to derive the previous bound, knowing that the number of jumps of the N^i is bounded by K and that Y^i is bounded by 1.
NB : We can restrict the supremum on a countable dense family of δ in order to apply very carefully the result of Rio. But by density, we recover this result. ■

We can also find a large event on which the behavior of Z_m is sub-Gaussian, as P. Massart does for estimating the density of an i.i.d. n-sample in [**48**].

PROPOSITION 21. *Let ε be a positive number and let $\Omega_m(\varepsilon)$ be the event*

$$\Omega_m(\varepsilon) = \left\{ \forall I \in m, |(N_I/n) - a_I| \leq \left(\frac{2\varepsilon}{(K+R)(1/2+\varepsilon^{-1})} \right)\beta_I \right\}.$$

Then under Assumptions 3.1.2 and 3, for all positive x,

$$\mathbb{P}\left(Z_m \mathbb{1}_{\Omega_m(\varepsilon)} \leq (1+\varepsilon)\left(\sqrt{\sum_{I\in m}\frac{\alpha_I}{n\beta_I}} + \sqrt{\frac{2R_m x}{n}} \right) \right) \leq e^{-x},$$

where

$$R_m = \sup_{I\in m} \frac{\alpha_I}{\beta_I}.$$

Proof. We know that Z_m is achieved in $\hat{\delta}$ such that for all I,

$$\hat{\delta}_I = \frac{(N_I/n) - a_I}{\beta_I Z_m}.$$

Consequently, Z_m can be seen on $\Omega_m(\varepsilon) \cap \{Z_m \geq z\}$ as

$$\sup_{\substack{\delta / \sum_{I\in m} \delta_I^2 \beta_I = 1, \\ \sup_{I\in m} \delta_I \leq \frac{2\varepsilon}{(K+R)(1/2+\varepsilon^{-1})z}}} \frac{1}{n}\int_0^1 (\sum_{I\in m} \delta_I 1\!\!1_I(t))(dN_t^i - Y_t^i s(t)dt).$$

If we apply Talagrand's inequality for this last supremum with $z = \sqrt{(2R_m x)/n}$, we get precisely the previous result. ■

NB : We can obtain the same kind of result replacing R_m by every proper upper bound.

Assumption 3 is a technical assumption precisely here to derive Propositions 20 and 21. If we have a Bernstein's type inequality instead of Talagrand's inequality for the suprema, we would be able to remove maybe the assumption of the existence of K and treat Markovian models too by the same method.

3.2.4. Oracle inequalities. Now we can construct oracle inequalities. The first one is a bound in probability on a large event, for the random norm. The second one is an expectation bound for the deterministic norm.

THEOREM 10. *Let N be a counting process with multiplicative intensity $Y_t s(t)$ (see (3.1.1)) satisfying Assumptions 3.1.2 and 3. Assume that s is bounded by an unknown positive R.*
Let Γ be a fixed regular partition of $[0;1]$. Let $\mathcal{M}_A$ be a family of partitions which are constructed with unions of intervals of Γ. For a given penalty pen *on $\mathcal{M}_A$, let $\tilde{s}$ be the associated penalized projection estimator (see (3.1.4)).*
Assume that :

(1) *there exists μ and ρ strictly positive such that* $\inf_{I\in\Gamma}(|\Gamma|\alpha_I) \geq \mu$ *and* $\inf_{I\in\Gamma}(|\Gamma|\beta_I) \geq \rho$,
(2) *there exists a finite family of positive weights on $\mathcal{M}_A$, $(L_m)_{m\in\mathcal{M}_A}$ such that* $\sum_{m\in\mathcal{M}_A} \exp(-L_m|m|) \leq \Sigma$ *for some Σ independent of n,*
(3) *$|\Gamma|$ is less than $n/\ln^2 n$.*

Let d larger than 1. One sets for all m in $\mathcal{M}_A$,

$$\text{pen}(m) = d\tilde{R}_\Gamma \frac{|m|}{n}\left(1+\sqrt{2L_m}\right)^2$$

where

$$\tilde{R}_\Gamma = \sup_{I\in\Gamma} \frac{N_I}{nb_I}.$$

Then there exists an event $\Omega(d)$ *such that for all* η *positive, there exists* C, C', C'' *continuous functions such that*

$$\mathbb{P}(\Omega(d)^c) \leq C''(d, K, R, \rho, \mu)/n^\eta$$

and such that on $\Omega(d)$ *for all* ξ *positive with probability larger than* $1 - \Sigma e^{-\xi}$,

$$\|s - \tilde{s}\|^2_{\text{rand}} \leq C(d) \inf_{m\in\mathcal{M}_A} \left\{ \|s - s'_m\|^2_{\text{rand}} + \frac{|m|L_m}{n} R_\Gamma \right\} + C'(d) R_\Gamma \frac{\xi}{n},$$

where $R_\Gamma = \sup_{I\in\Gamma} \frac{\alpha_I}{\beta_I}$.

Corollary 5. *Under the previous assumptions and notations, the penalized least square histogram estimator described in Theorem 10 verifies that there exists* H *and* H' *continuous positive functions such that*

$$\mathbb{E}(\|s - \tilde{s}\|^2_{\text{det}}) \leq H(d) \inf_{m\in\mathcal{M}_A} \left(\|s - s_m^{\text{det}}\|^2_{\text{det}} + R_\Gamma \frac{|m|L_m}{n} \right) + \frac{H(d, R, K, \rho, \mu, \Sigma)}{n}.$$

The weights L_m can be constant if the family of partitions is nested for instance. Then these oracle type inequalities become true oracle inequalities and the p.p.e. is consequently adaptive in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$.

The oracle type inequality of Theorem 10 is a probability bound. It is a stronger result than the one of Corollary 5 since the inequality is true for each event. But for the minimax risk, it is better to have oracle inequality for deterministic loss function (here $\|.\|^2_{\text{det}}$).

We can compare it to the model selection built in **[26]** for the right-censoring case. In this article, the penalty is very big (in $\exp(\exp(R))$) and depends on the knowledge of a bound on s. Here the penalty is linear in R and as we deal with histogram estimators, we can estimate the bound on s by $\tilde{R}_\Gamma$. We see in the simulations (see Section 3.4), that when the penalty is too large, the estimator is very bad and $C(d)$ becomes very large. The advantage of their estimator is that they can build those estimators for very different kind of deterministic models.

The weights L_m are here to take into account the complexity of the family of models. We refer to **[14]** for an extensive list of applications of these weights. Here we can make so complex applications, but we cannot see their performances since we restrict ourselves to the histogram models.

Proof.[Theorem 10] Let d be a positive number larger than 1 and let ε be a positive continuous function of d which we will choose later. Let $\Omega(d)$ be the following event :

$$\left\{\forall I \in \Gamma, |(N_I/n) - a_I| \leq \frac{2\varepsilon}{(K+R)(1+\varepsilon^{-1})}\beta_I, \right. \\ \left. |(N_I/n) - \alpha_I| \leq \frac{\varepsilon}{1+\varepsilon}\alpha_I, |b_I - \beta_I| \leq \frac{\varepsilon}{1+\varepsilon}\beta_I \right\}.$$

Let us bound the probability of $\Omega(d)^c$.

$$\begin{aligned}\mathbb{P}(\Omega(d)^c) \quad \leq \sum_{I\in\Gamma} \quad & \left[\mathbb{P}\left(|(N_I/n) - a_I| \geq \frac{2\varepsilon}{(K+R)(1+\varepsilon^{-1})}\beta_I\right) + \right. \\ & \left. + \mathbb{P}\left(|(N_I/n) - \alpha_I| \geq \frac{\varepsilon}{1+\varepsilon}\alpha_I\right) + \mathbb{P}\left(|b_I - \beta_I| \geq \frac{\varepsilon}{1+\varepsilon}\beta_I\right)\right].\end{aligned}$$

For each of these quantities one can use Bernstein's inequality, using the individual counting processes. All the quantities are sum of n independent and centered quantities. For the first probability, this is the sum of $(1/n)\int_0^1 \mathbb{1}_I dN^i - Y^i s dt$ with variance $(1/n^2)\alpha_I$. For the second probability, that is the sum of $(1/n)\int_0^1 \mathbb{1}_I dN^i - \mathbb{E}(Y^i)s dt$ with variance less than $(1/n^2)\alpha_I$. The both are bounded by $M = K + R$ divided by n. For the third probability, that is the sum of $(1/n)\int_0^1 \mathbb{1}_I(Y^i - \mathbb{E}(Y_i))dt$ with variance bounded by $(1/n^2)\beta_I$ and it is bounded by $1/n$. Consequently we get

$$\mathbb{P}(\Omega(d)^c) \leq 2\sum_{I\in\Gamma}\left[\exp(-n\beta_I h(\varepsilon, M, R_\Gamma)) + \exp(-n\alpha_I h'(\varepsilon, K, M)) + \exp(-n\beta_I h''(\varepsilon))\right]$$

where h, h', h'' are continuous functions. Finally we get for some f continuous positive :

$$\mathbb{P}(\Omega(d)^c) \leq 6\frac{n}{\ln^2 n}\exp(-(\ln n)^2 f(\varepsilon, \rho, \mu, K, R))$$

which is for fixed positive η less than some $C''(d, \rho, \mu, K, \|s\|_\infty)/n^\eta$.

Now let us look at $\Omega(d)$. Let m be some fixed partition in $\mathcal{M}_A$. We know that by construction

$$\gamma_A(\tilde{s}) + \text{pen}(\hat{m}) \leq \gamma_A(\hat{s}_m) + \text{pen}(m) \leq \gamma_A(s'_m) + \text{pen}(m).$$

Let us denote for all g in $\mathbb{L}^2([0,1], dt)$

$$\nu_n(g) = \int_0^1 g(t)\frac{dN_t - Y_t s(t)dt}{n}.$$

Using the fact that $\gamma_A(g)$ is $\|s-g\|^2_{\text{rand}} - \|s\|^2_{\text{rand}} - 2\nu_n(g)$, we obtain :

$$\|s - \tilde{s}\|^2_{\text{rand}} \leq \|s - s'_m\|^2_{\text{rand}} + 2\nu_n(\tilde{s} - s'_m) - \text{pen}(\hat{m}) + \text{pen}(m).$$

For a partition m', we denote by $m \cup m'$ the partition built on the extremities of the intervals of both m and m' and by χ_J the square root of χ^2_J defined in (3.2.4) for a set of intervals J.

Then for all $m' \in \mathcal{M}_A$, one has

$$\sup_{f \in S'_m + S'_{m'}} (\nu_n(f)/\|f\|_{\rm rand}) \leq \sup_{f \in S_{m\cup m'}} (\nu_n(f)/\|f\|_{\rm rand}) = \chi_{m\cup m'}.$$

Consequently we can write that

$$\begin{aligned} 2\nu_n(\tilde{s} - s'_m) &\leq 2\|\tilde{s} - s'_m\|_{\rm rand}\chi_{m\cup\hat{m}} \\ &\leq \frac{2}{\varepsilon}\|s - s'_m\|^2_{\rm rand} + \frac{2}{2+\varepsilon}\|s - \tilde{s}\|^2_{\rm rand} + (1+\varepsilon)\chi_{m\cup\hat{m}}, \end{aligned}$$

using twice the fact that for all a, b, θ positive numbers,

$$2ab \leq \theta a^2 + b^2/\theta.$$

Then we obtain

$$\frac{\varepsilon}{2+\varepsilon}\|s - \tilde{s}\|^2_{\rm rand} \leq \left(1 + \frac{2}{\varepsilon}\right)\|s - s'_m\|^2_{\rm rand} + (1+\varepsilon)\chi^2_{m\cup\hat{m}} - {\rm pen}(\hat{m}) + {\rm pen}(m). \tag{3.2.7}$$

In order to control $\chi^2_{m\cup\hat{m}}$, we have to control all the $\chi^2_{m\cup m'}$ for m' in $\mathcal{M}_A$. First we bound $\chi^2_{m\cup m'}$ by $Z^2_{m\cup m'}V_\Gamma$ since $S_{m\cup m'} \subset S_\Gamma$. We control all the $Z^2_{m\cup m'}$'s using Proposition 21 with an upper bound on $R_{m\cup m'}$, R_Γ (this is an upper bound by additivity). As we are on $\Omega(d)$, by additivity we are on $\Omega_{m\cup m'}(\varepsilon)$ defined in Proposition 21, and we can write that for all $x_{m'}$ positive, with probability larger than $1 - \exp(-x_{m'})$,

$$Z_{m\cup m'} \leq (1+\varepsilon)\left(\sqrt{\sum_{I\in m\cup m'}\frac{\alpha_I}{n\beta_I}} + \sqrt{\frac{2R_\Gamma x_{m'}}{n}}\right).$$

We choose $x_{m'} = L_{m'}|m'| + \xi$. With probability larger than $1 - \Sigma e^{-\xi}$, we control all the $Z_{m\cup m'}$ and also $Z_{m\cup\hat{m}}$. After some easy computations we get on $\Omega(d)$ with probability larger than $1 - \Sigma e^{-\xi}$:

$$Z^2_{m\cup\hat{m}} \leq (1+\varepsilon)^3 R_\Gamma \frac{|\hat{m}|}{n}(1 + \sqrt{2L_m})^2 + (1+\varepsilon)^3(1+\varepsilon^{-1})R_\Gamma\frac{|m|}{n} + (1+\varepsilon)^2(1+\varepsilon^{-1})^2\frac{2R_\Gamma\xi}{n}.$$

Now we can remark that we have build $\Omega(d)$ such that on $\Omega(d)$,

$$V_\Gamma \leq (1+\varepsilon) \text{ and } R_\Gamma \leq (1+2\varepsilon)\tilde{R}_\Gamma$$

We take ε such that $(1+\varepsilon)^5(1+2\varepsilon) = d$ which fixes ε and we obtain the result. ■

Proof.[Corollary 5] Let us return to the proof of Theorem 10. One has

$$\|s - \tilde{s}\|^2_{\rm det} = \|s - s^{\rm det}_\Gamma\|^2_{\rm det} + \|s^{\rm det}_\Gamma - \tilde{s}\|^2_{\rm det}.$$

On $\Omega(d)$, the random norm and the deterministic norms are equivalent for functions in S_Γ. Thus one has :

$$\|s_\Gamma^{\det} - \tilde{s}\|_{\det}^2 \leq (1+\varepsilon)\|s_\Gamma^{\det} - \tilde{s}\|_{\text{rand}}^2.$$

Then on $\Omega(d)$, we get

$$\|s - \tilde{s}\|_{\det}^2 \leq \|s - s_\Gamma^{\det}\|_{\det}^2 + 2(1+\varepsilon)\|s - s_\Gamma^{\det}\|_{\text{rand}}^2 + 2(1+\varepsilon)\|s - \tilde{s}\|_{\text{rand}}^2.$$

We apply Theorem 10 to the last term and we integrate on $\Omega(d)$ in ξ. We obtain after some easy computations

$$\mathbb{E}(\|s - \tilde{s}\|_{\det}^2 \mathbb{1}_{\Omega(d)}) \leq (3+2\varepsilon)\|s - s_\Gamma^{\det}\|_{\det}^2 + C(d)\mathbb{E}\left(\inf_{m\in\mathcal{M}_A}\left\{\|s - s'_m\|_{\text{rand}}^2 + \frac{|m|L_m}{n}R_\Gamma\right\}\right) + C'(d)R_\Gamma\frac{\Sigma}{n}.$$

Using (3.2.3) and exchanging the expectations and the infimum, there exists D and D' continuous positive functions such that

$$\mathbb{E}(\|s - \tilde{s}\|_{\det}^2 \mathbb{1}_{\Omega(d)}) \leq D(d)\left(\inf_{m\in\mathcal{M}_A}\left\{\mathbb{E}(\|s - s_m^{\det}\|_{\det}^2) + \frac{|m|L_m}{n}R_\Gamma\right\}\right) + \frac{D'(d,\Sigma,R)}{n}.$$

On $\Omega(d)^c$, we use the fact that $|s - \tilde{s}|$ is bounded by $R + Kn^2$ and the bound on its probability with $\eta = 3$ to obtain the result. ■

3.2.5. Minimax risk. The oracle inequalities imply that the p.p.e. is adaptive in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$: he finds without knowing s the best possible estimator in the family, up to some multiplicative constant for the risk. But we may also want to compare it with all the other possible estimators. This is the aim of this minimax study.

We know that the histograms have good approximation properties for a-hölderian functions with a positive less than 1.
Let L, r be positive numbers and

$$\mathcal{H}_{L,a,r} = \{f \in L^2([0,1],dt)/\forall x,y \in [0,1], |f(x)-f(y)| \leq L|x-y|^a \text{ and } r+L \geq f(x) \geq r)\}.$$

Then we can define the minimax risk on $\mathcal{H}_{L,a,r}$ by

$$R(\mathcal{H}_{L,a,r}) = \inf_{\hat{s}} \sup_{s\in\mathcal{H}_{L,a,r}} \mathbb{E}_s(\|s - \hat{s}\|_{\det}^2),$$

where $\hat{s}$ describes all the possible estimators in $\mathbb{L}^2([0,1],dt)$. The minimax risk on $\mathcal{H}_{L,a,r}$ represents the risk of the best estimator for the worst s to estimate in the family $\mathcal{H}_{L,a,r}$.

PROPOSITION 22. *If there exists μ and M such that for all s in $\mathcal{H}_{L,a,r}$, $\mu \leq \mathbb{E}_s(Y_t^1) \leq M$, then there exists c continuous positive function such that*

$$R(\mathcal{H}_{L,a,r}) \geq c(a)n^{-\frac{2a}{2a+1}}L^{\frac{2}{2a+1}}r^{\frac{2a}{2a+1}}\mu M^{-\frac{2a}{2a+1}}.$$

These assumptions are true in a lot of problems. For instance, in the right-censoring model, $\mathbb{E}_s(Y_t^1)$ is less than 1 and larger than $\exp -(r+L)$.

Proof. Let ψ be a symmetric positive function on $[0,1]$, in $\mathcal{H}_{1,a,0}$ with $\psi(0)=0$. Then for all D positive integer, $\psi_D(x) = LD^{-a}\psi(Dx)$ belongs to $\mathcal{H}_{L,a,0}$. Let us fix the regular partition Γ of $[0,1]$ with D intervals. Let m be a set of intervals of Γ and let for all I in Γ, u_I be the left extremity. Then

$$s_m = r + \sum_{I\in m} \psi_D(x-u_I)$$

belongs to $\mathcal{H}_{L,a,r}$. Let $\mathcal{C}$ be a set such that for all m, m' of $\mathcal{C}$, $|m \triangle m'| \geq \theta D$ and

$$\log|\mathcal{C}| \geq \sigma D$$

for θ and σ absolute constants. A such set exists by application of a combinatorial Lemma (see **[30]**). This lemma is already used a lot in Chapter 1. Let $\mathcal{A} = \{s_m, m\in\mathcal{C}\}$.
One has obviously that

$$R(\mathcal{H}_{L,a,r}) \geq \frac{1}{4}\inf_{\hat{s}\in\mathcal{A}}\sup_{s\in\mathcal{A}}\mathbb{E}(\|s-\hat{s}\|_{\text{det}}^2).$$

But for all $m\neq m'$ in $\mathcal{C}$,

$$\begin{aligned}\|s_m - s_{m'}\|_{\text{det}}^2 &= \int_0^1 \sum_{I\in m\triangle m'} \psi_D(t-a_I)^2\mathbb{E}(Y_t^1)dt \\ &\geq \mu|m\triangle m'|\int_0^1\psi_D(t)^2dt \\ &\geq \mu\theta L^2D^{-2a}P\end{aligned}$$

where $P=\int_0^1\psi^2$ depends only on a. Consequently, one has

$$\begin{aligned}R(\mathcal{H}_{L,a,r}) &\geq \frac{1}{4}\mu\theta PL^2D^{-2a}\inf_{\hat{s}\in\mathcal{A}}\sup_{s\in\mathcal{A}}\mathbb{P}_s(\hat{s}\neq s) \\ &\geq \frac{1}{4}\mu\theta PL^2D^{-2a}\inf_{\hat{s}\in\mathcal{A}}(1-\inf_{s\in\mathcal{A}}\mathbb{P}_s(\hat{s}=s)).\end{aligned}$$

We use a new version of Fano's lemma due to L. Birgé **[10]** : the infimum of the probabilities on the right hand side is bounded by an absolute constant α if the Kullback-Leibler distance is bounded by $\alpha\log|\mathcal{C}|$. By the combinatorial lemma, it is sufficient to bound it by $\alpha\sigma D$. But one has (taking the expectation of the classical formula of log-likelihood for counting

processes [**2**]) :

$$\begin{aligned}\forall m' \neq m \in \mathcal{C}, \quad K(\mathbb{P}_{s_{m'}}, \mathbb{P}_{s_m}) &= \int s_{m'} \phi(\log \frac{s_m}{s_{m'}}) \mathbb{E}_{s_m}(Y_t) dt \\ &\leq \int \frac{(s_m - s_{m'})^2}{s_m}(x) \mathbb{E}_{s_m}(Y_t) dt \\ &\leq \frac{1}{r} n M P L^2 D^{-2a}.\end{aligned}$$

One fixes D such that

$$\frac{1}{r} n M P L^2 D^{-2a} \simeq \alpha \sigma D.$$

This leads us to the result. ■

We can remark that we recover for the power of n the rate of convergence of the classical regression problem.

Now we want to compare the risk of $\tilde{s}$ built in Theorem 10 with the minimax risk. Let us look at the following classical strategy : $|\Gamma| = 2^J$ is of the order of $n/\ln^2 n$ and we take the sub-partitions, m, of Γ which are also regular with 2^j intervals, for j less than J. There is at the most one model by dimension, so we can take constant weights ($L_m = 1$, for instance) to build the penalty and consequently the p.p.e. We call this strategy the **nested histogram strategy**. Now let us apply Corollary 5.

If s is in $\mathcal{H}_{L,a,r}$, the bias $\|s - s_m^{\det}\|_{\det}^2$ is bounded by $L^2 |m|^{-2a} \varsigma$ where ς represents $\int_0^1 \mathbb{E}(Y_t^1) dt$. When n tends to infinity, we obtain taking m such that $|m|$ is of the order of $(n\varsigma L^2/R)^{1/(2a+1)}$ (which is less than $|\Gamma|$ for n large enough)

$$\mathbb{E}(\|s - \tilde{s}\|_{\det}^2) = \mathcal{O}\left(n^{-\frac{2a}{2a+1}} L^{\frac{2}{2a+1}} R^{\frac{2a}{2a+1}} \varsigma^{\frac{1}{2a+1}}\right).$$

This is quite the lower bound found in Proposition 22. One has the good power of n and of L. The bound R on s replaces r the infimum of s. As for ς, it replaces $\mu^{2a+1} M^{-2a}$ and represents the order of magnitude of $\mathbb{E}(Y_t^1)$.

This means that $\tilde{s}$ without knowing a and L (depending on s) does quite as well as the best possible estimator which knows this fact (they have quite the same rate of convergence). In this sense, $\tilde{s}$ is an **adaptive estimator** for the a-hölderian functions with a less than 1.

3.3. Predictable models

We have seen what can be done easily for aggregated processes. Let us remove this assumption and let us manage the predictable models. We keep the notations defined in (3.1.3) and (3.1.5).

We assume Assumption 2 and the following fact.

Assumption 4. *There exists c positive such that if there exists a positive t such that $Y_t < c$ then $Y_t = 0$.*

For a Poisson process, one has $A = c$. For the other examples, Y is an integer-valued function and$c = 1$ is convenient.

The aggregated case lead us to think that A plays the same role as n and would tend to infinity for an asymptotical point of view in the general framework.

We denote by J_t, $\mathbb{1}_{Y_t \neq 0}$.

3.3.1. Construction and risk for one model. The family of models is built as follows : let $\{\varphi_\lambda, \lambda \in \Gamma\}$ be a classical orthonormal basis for $\mathbb{L}^2([0,1], dt)$; let $\mathcal{M}_A$ be a family of subsets of Γ. Then for m in $\mathcal{M}_A$, we set $S_m = \text{Span}\{h_\lambda(.) = \left(\varphi_\lambda(.)/\sqrt{Y'_.}\right) J_., \lambda \in m\}$. We associate to S_m, $\hat{s}_m$ the projection estimator defined by (3.1.4). We denote by $|m|$ the cardinality of m.

Let us define the following observable event :

$$\Omega = \{\forall t \geq 0, Y_t \neq 0\}. \tag{3.3.1}$$

We will see later that in a lot of situations, Ω has a very large probability to happen when A is large enough.

On Ω, the h_λ's form an orthonormal basis of S_m for the random scalar product and consequently $\hat{s}_m$ is of the form (3.1.6).

Risk of the projection estimator :

On Ω, the projection s_m of s on S_m for the random product is

$$s_m(.) = \sum_{\lambda \in m} \left(\int_0^1 \varphi_\lambda(t) s(t) \sqrt{Y'_t} dt \right) \frac{\varphi_\lambda(.)}{\sqrt{Y'_.}}.$$

Then we can write

$$\|s - \hat{s}_m\|^2_{\text{rand}} = \|s - s_m\|^2_{\text{rand}} + \|s_m - \hat{s}_m\|^2_{\text{rand}}.$$

The first term corresponds to a bias term. Here it is also a random term as for the histogram case. We can write

$$\|s - s_m\|^2_{\text{rand}} = \int_0^1 \left(s(t)\sqrt{Y'_t} - \sum_{\lambda \in m} \left(\int_0^1 \varphi_\lambda(t) s(t) \sqrt{Y'_t} dt \right) \varphi_\lambda(t) \right)^2 dt.$$

Consequently the bias term corresponds to the classical $\mathbb{L}^2([0,1], dt)$ error when one projects $s\sqrt{Y'}$ on $\text{Span}\{\varphi_\lambda, \lambda \in m\}$. If m grows, this term generally diminishes.

The second term corresponds to a χ^2 type statistics of the form (3.2.4) in the histogram case : on Ω, it is $\chi(m)^2_1$ where the process $(\chi(m)^2_t)_{t \geq 0}$ is defined by

$$\chi(m)^2_t = \sum_{\lambda \in m} \left(\int_0^t \frac{\varphi_\lambda(u)}{\sqrt{Y'_u}} J_u \frac{dM_u}{A} \right)^2, \text{ for all } t \geq 0. \tag{3.3.2}$$

Its compensator $(C(m)_t)_{t\geq 0}$ is defined by

$$C(m)_t = \sum_{\lambda\in m} \int_0^t \varphi_\lambda^2(u)s(u)J_u\frac{du}{A}, \text{ for all } t \geq 0.$$

But, on Ω, $C(m)_1$ is a constant between $r|m|/A$ and $R|m|/A$ if s is upper bounded by R and lower bounded by r. Consequently, if $\chi(m)_1^2$ is close to $C(m)_1$, it grows like the dimension of the model.

3.3.2. Penalized projection estimator. Hence, if we want to find a good model, we must balance the bias term and the variance term, but we must do this through a data driven criteria, without knowing s. Consequently we use (3.1.7) and we obtain $\tilde{s}$, the p.p.e. for the family of models $\{S_m, m \in \mathcal{M}_A\}$.

There exists here also an heuristic argument. We can also defined an oracle, the best model which we could choose if we know s :

$$\begin{aligned} \bar{m} &= \arg\min_{m\in\mathcal{M}_A} \|s - \hat{s}_m\|_{\text{rand}}^2 \\ &= \arg\min_{m\in\mathcal{M}_A} -\|s_m\|_{\text{rand}}^2 + \|s_m - \hat{s}_m\|_{\text{rand}}^2 \\ &\simeq \arg\min_{m\in\mathcal{M}_A} -\|\hat{s}_m\|_{\text{rand}}^2 + 2\|s_m - \hat{s}_m\|_{\text{rand}}^2 \\ &\simeq \arg\min_{m\in\mathcal{M}_A} -\|\hat{s}_m\|_{\text{rand}}^2 + 2\chi(m)_1^2. \end{aligned} \tag{3.3.3}$$

The approximations are licit if the coefficients (as in the histogram case) of $s_m - \hat{s}_m$ are close to their expectation 0. If $\chi(m)_1^2$ is close to $C(m)_1$, a penalty of the form $2c|m|/A$ would be convenient (where c is of the order of s). We found again the factor 2 which always appears when we do a Mallows heuristic.

The study of the probabilistic behavior of $\chi(m)_t$ around its compensator has been made in Chapter 2.

3.3.3. Oracle inequalities. Now we can derive oracle type inequalities for predictable models.

THEOREM 11. *Let N be a counting process with multiplicative intensity $Y_t s(t)$ (see (3.1.1)) satisfying Assumptions 2 and 4.*

Let $\{S_m, m \in \mathcal{M}_A\}$ be a family of predictable models built as previously from the deterministic classically orthonormal family $\{\varphi_\lambda, \lambda \in \Gamma\}$. For a given penalty pen *on $\mathcal{M}_A$, let $\tilde{s}$ be the associated penalized projection estimator (see (3.1.7)).*
Assume that :

(1) *there exists Φ positive, such that for all m in $\mathcal{M}_A$,*

$$\|\sum_{\lambda\in m} \varphi_\lambda^2\|_\infty \leq \Phi|m|.$$

(2) *there exists a finite family of positive weights on* $\mathcal{M}_A$, $(L_m)_{m\in\mathcal{M}_A}$ *such that* $\sum_{m\in\mathcal{M}_A} |m|^2 \exp(-L_m) \leq \Sigma$,

Moreover assume that we know a bound on s *denoted by* R.

Let d *be larger than* 1. *One sets for all* m *in* $\mathcal{M}_A$ *:*

$$\mathrm{pen}(m) = d\frac{|m|}{A}\left(\sqrt{R}(1+3\sqrt{2L_m}) + \sqrt{\frac{\Phi}{c}}L_m\right)^2.$$

Then there exists C, C' *continuous positive functions such that on* Ω *defined by (3.3.1), one has*

$$\mathbb{E}(\|s-\tilde{s}\|^2_{\mathrm{rand}}\mathbb{1}_\Omega) \leq C(d)\inf_{m\in\mathcal{M}_A}\left(\mathbb{E}(\|s-s_m\|^2_{\mathrm{rand}}) + \mathrm{pen}(m)\right) + \frac{C'(d,R,\Phi,c,\Sigma)}{A}.$$

As the models are random, we can only derive oracle inequalities for the random norm. Probability bounds exist but are much more intricate than in Theorem 10.

The classical case is when $\{\varphi_\lambda, \lambda\in\Gamma\}$ is a Fourier basis $\{\exp(-2ik\pi x), k\in\mathbb{Z}\}$ with $\mathcal{M}_A = \{m_k = \{-k,k\}, k\geq 0\}$, then $|m_k| = 2k+1$ and $L_{m_k} = 4\ln k$. The constant Φ in the theorem equals 1. Obviously in practice we must take a finite family of models i.e. take $k\leq A$ for instance.

We can also consider a wavelet basis $\{\varphi_{j,k}, j\geq 0, k\geq 0\}$ with regularity h and $\mathcal{M}_A = \{m_l, l\geq 0\}$ where $m_j = \{(l,k), l\leq j\}$. If the wavelet has finite support, Φ defined in Theorem 11 depends only on the basis.

As the family of models is nested in both previous cases, the penalty is of the order of $|m|R\log(|m|)/A$. Thus we recover an oracle inequality up to some logarithmic factor, since the variance term is of the order of $|m|R/A$. We can imagine more complex family of models (i.e. more models with same dimension). If the number of models with dimension D in the family is of the order of a power of D, we can have the same kind of penalty and we recover also oracle inequalities up to the logarithmic factor. If the number of models with same dimension D is of the order of e^D, the penalty must be in $R|m|^\gamma/A$ for $\gamma>1$. It is really larger than the variance term and there is no more oracle inequality.

When one has an oracle inequality, we can also say that the p.p.e. is adaptive in the family $\{\hat{s}_m, m\in\mathcal{M}_A\}$. But as we do not know the approximation properties of the random spaces S_m, we cannot generally consider adaptivity in the minimax sense.

If N is a Poisson process (it implies that all the norms are determinist), let us assume that s belongs to

$$\mathcal{B}(\rho,L,B^\alpha_{2,2}) = \left\{ t=\rho+u \,\middle/\, t\geq 0,\ \int_O^1 u\,dx = 0,\ u\in B^\alpha_{2,2},\ \|u\|^\alpha_{2,2}\leq L \right\}$$

where $B^\alpha_{2,2}$ is the classical Besov space with regularity α between h and $1/2$ and with $\mathbb{L}^2$ norm. Consider the last strategy with the wavelet of regularity h. Then making a

compromise between the penalty and the bias in the oracle inequality, we obtain when A tends to infinity :

$$\mathbb{E}(\|s-\tilde{s}\|^2) = \mathcal{O}\left(L^{\frac{2}{2\alpha+1}} R^{\frac{2\alpha}{2\alpha+1}} \left(\frac{A}{\ln^2 A}\right)^{-\frac{2\alpha}{2\alpha+1}}\right).$$

This is the minimax rate (see Chapter 1) up to the logarithmic factor and the replacement of $\int_0^1 s$ by R. Consequently, the resulting p.p.e. is adaptive in the minimax sense for all the Besov balls with regularity less than h.

We can loose this logarithmic factor in the Poisson case : we prove in Chapter that penalties of the type $R|m|/A$ with the same previous families of models gives oracle inequalities without logarithmic factor and consequently minimax rate without logarithmic factor. If we apply Theorem 11 which is valid for more general processes, the weights L_m are null and the last term explodes with Σ for large family of models : there is no more oracle inequality.

The same kind of remark can be made if we want to use more complex family of models (i.e. more models with same dimension in the family of models). In the Poisson framework, there exists penalties of the type $R|m|(\log A)/A$ which gives up to some logarithmic factor oracle inequalities. Here the same type of strategies give an explosive last term.

However, the counting processes are very well adapted to biomedical data. In these cases, the number of observations $n \simeq A$ is not very large and if we take also a small number of models, there is no more explosive phenomenon. This is the reason why having non-asymptotical results is interesting.

Proof.[Theorem 11] Let d be a positive real number larger than 1 and let ε be a positive continuous function of d which we will choose later. On Ω, we can do the same computations as in the histogram case to obtain :

$$\|s-\tilde{s}\|^2_{\text{rand}} \leq \|s-s'_m\|^2_{\text{rand}} + 2\nu_A(\tilde{s}-s'_m) - \text{pen}(\hat{m}) + \text{pen}(m)$$

where for all g in $\mathbb{L}^2([0,1],dt)$,

$$\nu_A(g) = \int_0^1 g(t)\frac{dM_t}{A}.$$

On Ω, one can see that

$$\chi(m\cup\hat{m})_1 = \sup\{\nu_A(f)/f \in S_{m\cup\hat{m}}, \|f\|_{\text{rand}} = 1\}.$$

Consequently, one can use the same trick as for the histograms. We obtain

$$\frac{\varepsilon}{2+\varepsilon}\|s-\tilde{s}\|^2_{\text{rand}} \leq \left(1+\frac{2}{\varepsilon}\right)\|s-s'_m\|^2_{\text{rand}} + (1+\varepsilon)\chi(m\cup\hat{m})^2_1 - \text{pen}(\hat{m}) + \text{pen}(m). \quad (3.3.4)$$

Moreover one has $\chi(m\cup\hat{m})^2_1 \leq \chi(m)^2_1 + \chi(\hat{m})^2_1$.

But for all m' in $\mathcal{M}_A$, we can apply the exponential formula derived in Chapter 2 : for all $x_{m'}$ positive with probability larger than $1-2\exp(-x_{m'})$

$$\chi(m')_1 \leq \sqrt{C(m')_1} + 3\sqrt{2v_{m'}x_{m'}} + b_{m'}x_{m'}$$

where

– $v_{m'}$ is a deterministic bound on $C(m')_1$,

– $b^2_{m'}$ is a deterministic bound on $\sum_{\lambda\in m'} \varphi^2_\lambda/(Y'A^2)$.

Under the assumptions of the theorem, we obtain that : for all $x_{m'} > 0$ with probability larger than $1-2\exp(-x_{m'})$

$$\chi(m')_1 \leq \sqrt{\frac{|m'|}{A}}\left[\sqrt{R} + 3\sqrt{2Rx_{m'}} + \sqrt{\frac{\Phi}{c}}x_{m'}\right].$$

Let $\xi > 0$ and let us take $x_{m'} = L_{m'} + \xi/|m'|$. Then we can bound $\chi(m')^2_1$ by

$$(1+\varepsilon)\frac{|m'|}{A}\left(\sqrt{R}(1+3\sqrt{2L_{m'}}) + \sqrt{\frac{\Phi}{c}}L_{m'}\right) + (1+\varepsilon^{-1})(1+\varepsilon)\frac{18\xi}{A} + (1+\varepsilon^{-1})^2\frac{\xi^2}{A}.$$

If we take $d = (1+\varepsilon)^2$ which fixes ε with probability larger than

$$1 - 2\sum_{m'\in\mathcal{M}_A} \exp(-L_{m'} - \frac{\xi}{|m'|})$$

one has that

$$\frac{\varepsilon}{2+\varepsilon}\|s-\tilde{s}\|^2_n \leq \left(1+\frac{2}{\varepsilon}\right)\|s-s'_m\|^2_n + 2\mathrm{pen}(m) + 2\left[(1+\varepsilon^{-1})(1+\varepsilon)\frac{18\xi}{A} + (1+\varepsilon^{-1})^2\frac{\xi^2}{A}\right].$$

It remains to integrate in ξ. We obtain by change of variables and Beppo-Levy, the result. ■

3.3.4. Improvement.

Estimation of R : The fact that the penalty depends on the knowledge of a bound on s can be a nuisance. But we can in some cases estimate it.

Let Γ be a regular partition of $[0,1]$. Suppose that s is L,α-hölderian, and let s_Γ be the projection of s for the random norm on the space of histograms with partition Γ then

$$\|s - s_\Gamma\|_\infty \leq L|\Gamma|^{-\alpha}.$$

Take $|\Gamma|$ of order $A/\ln^2 A$. Then $\|s\|_\infty \leq \|s_\Gamma\| + o(1)$ when A goes to infinity. But

$$\|s_\Gamma\|_\infty = \sup_{I\in\Gamma} \frac{\int_I s(t)Y'_t dt}{\int_I Y'_t dt}.$$

So we can replace R by $(1+\varepsilon)\tilde{R}_\Gamma$ where

$$\tilde{R}_\Gamma = \sup_{I\in\Gamma} \frac{N_I}{A\int_I Y_t' dt}$$

if we are on

$$\Omega(\varepsilon) = \left\{ |\int_I dM_t/A| \leq \frac{\varepsilon}{1+\varepsilon} \int_I s(t) Y_t' dt \right\}.$$

The complementary of this event is very small (with probability of order $o(A^{-\eta})$, for all $\eta > 0$) if we assume the process to be aggregated and Assumption 3 (or moment assumptions). Then we can use Bernstein's inequality to $\int_I dM/A$ and to $\int_I s(t)Y_t' dt$. On $\Omega \cap \Omega(d)^c$, the estimator is bounded and one can conclude as in the proof of Corollary 5.

Magnitude of Ω : In the aggregated cases, Ω is a very large event and we can also give an oracle type inequality for $\mathbb{E}(\|s - \tilde{s}\|^2_{\text{rand}})$.

Let us look more precisely at the right-censoring model. In this case $A = n$ and $Y_t' = \sum_{i=1}^n \mathbb{1}_{X_i \wedge U_i \geq t}$ where the X_i's are the life times and the U_i's form the censorship. It can be seen as $1 - \hat{F}_n(t)$ where $\hat{F}_n(t)$ is the empirical repartition function associated with the $X_i \wedge U_i$'s. One has

$$\forall \lambda > 0, \mathbb{P}\left(\sqrt{n} \sup_{\mathbb{R}} \left| \hat{F}_n(t) - F(t) \right| \geq \lambda \right) \leq 2e^{-2\lambda^2},$$

where F is the true repartition function (see **[49]**).

Thus if we assume that there exists a positive μ such that $\mathbb{E}(Y_t^1) \geq \mu > 0$ on $[0,1]$, then

$$\Omega^c \subset \{\sup \left| Y_t' - \mathbb{E}(Y_t^1) \right| \geq \mu/2\}$$

and has a probability less than $2\exp(-n\mu^2/2)$.

Hence we can defined the estimators on all the probability space by :

$$\hat{s}_m(.) = \sum_{\lambda \in m} \left(\int_0^1 \frac{\varphi_\lambda(t)}{\sqrt{Y_t'}} J_t \frac{dN_t}{A} \right) \frac{\varphi_\lambda(.)}{\sqrt{Y_.'}} J_.,$$

even if we are not in Ω. This estimator is a projection estimator only on Ω. We do the model selection as in Theorem 11. As these estimators are always bounded, we do like in Corollary 5 and we can bound $\mathbb{E}(\|s - \tilde{s}\|^2_{\text{rand}})$ (on all the probability space) by the same kind of bound as in Theorem 11.

3.4. Simulations

The aim of this paragraph is not an extensive simulations study but just an illustration of the previous methods. We restrict ourself to the right-censoring case.

The life times $X_1, ..., X_n$ are generated for a given hazard rate s on $[0,1]$. The censorship $U_1, ...U_n$ are generated as uniform variables on $[0,2]$ or $[0,1.2]$ in order to give some different censoring rate. Some of the observations will be outside $[0,1]$: it is a good case since it

ensures that we are in the set Ω of the previous section. We denote $T_i = X_i \wedge U_i$ for all i less than n.

Three different methods are compared here for just some examples. I do not pretend to give precise formula for the penalty but just some penalties which work quite well.

The **nested histogram strategy** (N.H.S.) was already described in Section 3.2.4 : the family of models consists of all the regular dyadic histograms with at the most 2^J intervals where J is the integer part of $[\log_2(n) - 2\log_2(\log(n))]$. Theorem 10 allows us to have constant weights. The penalty is consequently of the form

$$\text{pen}(m) = \frac{d\tilde{R}_\Gamma |m|}{n}$$

where $\tilde{R}_\Gamma$ is given by this theorem .

The **exhaustive histogram strategy** (E.H.S.) look at all the sub-partitions of Γ where Γ is a regular partition with d intervals where d is the minimum of 8 and the integer part of $[n/\log(n)^2]$. The factor 8 is here to have small computing times. The penalty is of the form

$$\text{pen}(m) = \frac{d\tilde{R}_\Gamma |m|}{n}\left(1 + \sqrt{2\log\left(\frac{n}{|m|}\right)}\right)^2$$

i.e. the weights L_m of Theorem 10 are of the form $\log(n/|m|)$ to ensure the convergence of Σ.

The **Fourier strategy** (F.S.) is the strategy described in the last section. The φ_λ's are the Fourier basis and we consider the nested models described in this part. According to Theorem 11, the penalty is of the form

$$\text{pen}(m) = \frac{d|m|}{n}(\sqrt{\tilde{R}_\Gamma} + \log|m|)^2$$

(in order to have simple formula, we delete the second term of the penalty which is smaller than the other terms).

At the end of the interval, there are sometimes few points to see, so it is well known that all estimations of the hazard rate become bad. To take this into account, the random norm $\|s - \tilde{s}\|^2_{\text{rand}}$ is a good quantity since it multiplies the functions by Y' , decreasing function. Moreover this norm is not just convenient, it is very close to the Kullback-Leibler distance as we have seen when we have done minimax computations in Section 3.2.5. This random norm is denoted by "Risk" on the figures.

3.4.1. Influence of d. To illustrate the influence of d, let us look at the easiest strategy : the nested histogram strategy.

In Figure 1, the unknown hazard rate is already an histogram. If the penalty is equal to $2\tilde{R}_\Gamma|m|/n$ (i.e. $d = 2$), the p.p.e. recovers the good model with two intervals. If $d = 100$, the procedure gives a very high data-driven criteria for large dimension and the p.p.e. finds

FIG. 1 – Example of the influence of d for the nested histogram strategy

the model with only one interval. If $d = 0.1$, the penalty is not large enough. The p.p.e. finds the model with two much intervals (here 4). The best case (or the oracle, i.e. the model chosen if we know s) is the one found by the procedure for $d = 2$.

If we compare with the classical model selection technics **[14]**, we must have for $d < 1$ very bad estimators. But here the frontier in $d = 1$ is not clear in the simulations. This is probably due to the presence of $\tilde{R}_\Gamma$ which must overestimate s.

However a type Mallows type heuristics (i.e. with $d = 2$) seems to work very well for the nested histogram strategy, on a lot of examples and in fact there exists a large interval of possible d which work well.

There exists methods to estimate a proper d by a data-driven criterion in the Gaussian framework. This is the work of E. Lebarbier in her PhD Thesis. We do not try here to do this kind of work which is quite long and hard.

FIG. 2 – Example of small sample

For the F.S. and the E.H.S., we have found good d by the same study as before. In these cases, there exists also intervals of possible d on which the estimator is close to the oracle. However in both cases, $d = 2$ is too large because of the presence of the logarithmic factor, but we cannot remove it since it is then under-penalized (the p.p.e. gives models with large dimension). Consequently, we set $d = 0.4$ for the E.H.S. and $d = 1$ for the F.S.

3.4.2. Influence of the number of data. In Figures 2 and 3, we can compare the performances between the three strategies for small and large samples. We take an hazard rate which does not belong to any model.

For small and large sample, the three strategies seem to perform quite well. The fact that the N.H.S. is better is only due to the form of s, but the three risk have the same order. The large sample gives smaller distance. But even for only 30 uncensored data, the estimators seem to be not so bad.

FIG. 3 – Example of large sample

FIG. 4 – Histogram performances

3.4.3. Comparison of the three strategies.

N.H.S. versus E.H.S.

Obviously, the exhaustive strategy performs very well when the hazard rate s is an irregular histogram.

In Figure 4, we see that the E.H.S. gives the good model (the oracle too), and as the partition for s is not regular, it performs better than the N.H.S.

In Figure 5, the E.H.S. is always better than the N.H.S. but he does not find the model of s. In fact he cannot find it properly since the last interval in s is not in the partitions of the family of models. The E.H.S. seems to find some good points of the partition but not all. At the opposite, the N.H.S. finds more points than the partition of s. Moreover this problem happens at the end of the interval when all the estimates becomes bad.

N.H.S. versus F.S.

FIG. 5 – Histogram performances

The first figure (Figure 6) shows the relative performances of the N.H.S. and the F.S. The E.H.S. has here the same result as the N.H.S. We see that the F.S. strategy is better. It seems to explode at 1, but this is not a problem for the random norm. Moreover this capacity to explode at the end, can be viewed as an advantage. Let us assume that the X_i's have quite support in $[0, 1]$, thus their hazard rate becomes very large in 1. This is the situation in Figure 7. We see that the N.H.S as the E.H.S do not grow in 1. But as Y' is near zero at the end, the F.S. explodes too, and has a really smaller random distance.

Another possible remark is that even if the F.S. is not continuous and smooth, it seems to approximate better smooth functions (except from the constants). Indeed, it seems to perform as well and sometimes even better than the histogram strategies if the true hazard rate s does not belong to any models and is not in particular an histogram.

FIG. 6 – Fourier performances

FIG. 7 – Fourier performances for explosive hazard rate

3.4.4. Comparison with another adaptive estimator. In this paragraph, we want to compare our estimators with the one given by A. Antoniadis, G. Grégoire and G. Nason in [**5**]. Their estimator is a wavelet estimator and they choose the coefficients to keep by a cross-validation criterion in their simulation. Consequently, their estimator has the same quality as ours : this is a completely data-driven non parametric estimator.

As their estimator is built on $[0, \tau]$ where τ is the last observation, we do the following rescaling. We divide the observations by τ to obtain a new set of observations in $[0, 1]$ and as the last point is always 1, we are always in Ω. This new set of observations has (if τ was deterministic) an intensity of the form $\bar{s}(t) = \tau s(\tau t)$. We estimate it on $[0, 1]$ by $\tilde{s}$ coming either from the N.H.S.($d = 2$) or the F.S. ($d = 1$). Then the resulting estimator for s on $[0, \tau]$ is $\hat{s}(x) = \tilde{s}(x/\tau)/\tau$.

In the first set of simulations, the X_i's follow a Gamma distribution with shape parameter 5 and scale 1 and the U_i's follow an exponential distribution with mean 6. We can observe the result in Figure 8.

In the second set of simulations, the X_i's have a bimodal density defined by

$$f = 0.8g + 0.2h$$

where g is the density of $\exp(Z/2)$ with Z having a standard normal distribution and where h is the density of $0.17Z + 2$. The U_i's follow an exponential distribution with mean 2.5. We can observe the result in Figure 9.

In both cases, we see that the estimator is very bad at the end of the interval since one has few observations by construction at the end.

FIG. 8 – Estimation by the N.H.S. and the F.S for a Gamma distribution

FIG. 9 – Estimation by the N.H.S. and the F.S for a bimodal distribution

We can compare our estimators with theirs by computing the same error on a lot of simulations. If one takes in $[0,\tau]$ K points regularly spaced denoted by t_k, the AMSE error is defined by :

$$\text{AMSE} = \frac{1}{K}\sum_{k=1}^{K}(\hat{s}(t_k) - s(t_k))^2.$$

The AMSE2 error is defined for the first simulation by the same kind of mean squared error but only for the t_k's less than 6. This is done in order to remove the effect of scarcity of the observations (one has $\mathbb{P}(X > 6) = 25\%$).

For the second simulations, the AMSE2 is done for the t_k's less than 2 in Figure 9 (one has here $\mathbb{P}(X > 2) = 16\%$.)

We can give also an average of these errors over 200 simulations as they do in **[5]**. In their results, they compute the AMSE2 in the second case for t_k's less than 2.5 which seems strange since $\mathbb{P}(X > 2.5) = 2\%$. However we keep the same level in the following arrays, to be absolutely rigorous.

We see in Figure 10 the performances of their estimator, in Figure 11 the performances of the N.H.S. with $d = 2$ and in Figure 12 the performances of the F.S. with $d = 1$. All the errors in these figures are averaged over 200 simulations.

We see that our estimators are better on the whole interval $[0,\tau]$ than theirs : we have built them to have good behavior as far as possible. However since the N.H.S. is not very smooth, it cannot give as good errors as theirs for a smaller interval. The F.S. is smoother than the N.H.S. when Y_t is large enough and consequently it has better results than the N.H.S. and results of the same order as their estimator at least for the Gamma distribution on $[0, 6]$. The difference between AMSE and AMSE2 is not clear for our estimators when we compute the AMSE2 in 2.5 for the bimodal distribution. However if we compute it in 2, we have an error of order 238.10^{-3} for the N.H.S. and 95.10^{-3} for the F.S. for 200 observations and, 151.10^{-3} for the N.H.S.and 47.10^{-3} for the F.S. for 500 observations. Consequently, one has the same phenomenon of reduction of the error but for a smaller interval than theirs.

We can also remark that there is no obvious differences in the errors for different K for our estimators since they do not depend on K unlike theirs. For the AMSE, there is no difference also for 200 or 500 intervals : this is due to the fact that for large sample size, the last observation τ becomes larger and the scarcity of observations at the end of the interval of observations is consequently the same. For the AMSE2, there is some amelioration taking more observations.

Distributions		Gamma		Bimodal	
Number of observations		200	500	200	500
	K				
AMSE	16	64.4	55.4	3050	3090
	32	78.6	55.4	4060	1820
	64	112.0	99.5	2080	1970
AMSE2	16	5.8	5.9	182	295
	32	2.6	2.1	152	66
	64	2.5	1.6	48	32

FIG. 10 – Results of A. Antoniadis, G. Grégoire and G. Nason. (Errors $\times 10^{-3}$)

Distributions		Gamma		Bimodal	
Number of observations		200	500	200	500
	K				
AMSE	16	32.9	31.7	810.5	711.7
	32	33.9	37.6	899.4	746.3
	64	34.2	33.6	767.7	850.2
AMSE2	16	5.8	4.3	677.8	449.1
	32	6.0	4.1	735.9	470.6
	64	6.0	4.1	650.6	478.9

FIG. 11 – Results of the N.H.S.(d=2) (Errors $\times 10^{-3}$)

Distributions		Gamma		Bimodal	
Number of observations		200	500	200	500
	K				
AMSE	16	42.1	43.8	769.3	769.2
	32	47.3	54.0	1057.1	996.2
	64	55.3	53.1	884.3	1009.4
AMSE2	16	2.3	1.16	470.9	358.0
	32	2.2	1.16	679.1	428.1
	64	2.2	1.18	563.7	483.0

FIG. 12 – Results of the F.S.(d=1) (Errors $\times 10^{-3}$)

ANNEXE A

Combinatorial lemmas

LEMMA 13. *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*

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

This lemma has been proved by Aldous [**1**].

LEMMA 14. *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 with 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 fixed x and y chosen random uniformly 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 :

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

In order to understand this probability, we can apply Lemma 13 : a draw without replacement is more concentrated (for convex functions) than a draw with replacement (which is here a binomial variable, $N_b^* \sim \mathcal{B}(D, D/N)$). It leads to, for all $\lambda > 0$:

$$\begin{aligned} 1 &\leq |\mathcal{M}_{N,D}| \exp(-\lambda(D - \theta D)) E(\exp(\lambda N_b)) \\ &\leq |\mathcal{M}_{N,D}| \exp(-\lambda(D - \theta D)) E(\exp(\lambda N_b^*)). \end{aligned} \tag{A.0.1}$$

Following the proof of Bennett's inequality ([**11**]), we obtain, maximizing (A.0.1) in λ :

$$1 \leq |\mathcal{M}_{N,D}| \exp\left(-\frac{D^2}{N} h\left(\frac{D-\theta D-D^2/N}{D^2/N}\right)\right) \tag{A.0.2}$$

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

$$\begin{aligned} |\mathcal{M}_{N,D}| &\geq \exp\left(\frac{D^2}{N}\left[\frac{D-\theta D}{D^2/N}\ln\frac{D-\theta D}{D^2/N} - \frac{D-\theta D}{D^2/N} + 1\right]\right) \\ &\geq \exp\left(\sigma D \ln\frac{N}{D}\right) \end{aligned} \tag{A.0.3}$$

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

ANNEXE B

Programmes Scilab

Cette annexe présente les programmes en Scilab 2.6 qui ont permis de faire les simulations du chapitre 3. Nous n'avons simulé que le cas censure à droite. Je tiens à remercier tout particulièrement Yann pour son aide précieuse dans ces programmes.

B.1. Simulations des variables aléatoires

Le fichier **mes_fonctions.sci** est un fichier de fonctions nécessaires pour créer à partir d'un taux de hasard s sur $[0,1]$, la fonction de répartition correspondante F. Une fois qu'on a la fonction F, on a besoin de sa réciproque pour simuler les variables de taux de hasard s. Si le taux de hasard est borné, il correspond à une fonction de répartition qui a un support beaucoup plus gros que $[0,1]$. Comme on a juste besoin de connaître les emplacements de ces points sur $[0,1]$, les points qui peuvent tomber à l'extérieur de l'intervalle sont mis en 1 par le programme.

```
// Fonction F

function [y] = F(t,s)
 y = 1-exp(-intg( 0, t, s));
endfunction

// Fonctions d'essai

function [y] = un( x )
 y = (x*0)+1.0;
endfunction

function [y] = deux(x)
 y = (x*0)+2.0;
endfunction

function [y] = up_linear(x)
 y = x;
endfunction
```

```
function [y] = down_linear(x)
 y = 1-x;
endfunction
function [y] = escalier(x)
 n=size(x,'*');
 y=x;
 for i=1:n
 if ((0<=x(i)) & (x(i)<0.5))
   y(i)=1;
 end
 if ((0.5<=x(i)) & (x(i)<=1))
   y(i)=5;
 end
end
endfunction

function [y] = escalier1(x)
 n=size(x,'*');
 y=x;
 for i=1:n
  if ((0<=x(i)) & (x(i)<0.3))
    y(i)=4;
  end
  if ((0.3<=x(i)) & (x(i)<=1))
    y(i)=1;
  end
end
endfunction

function [y] = escalier2(x)
n=size(x,'*');
y=x;
for i=1:n
  if ((0<=x(i)) & (x(i)<0.25))
    y(i)=4;
  else
   if ((0.25<=x(i)) & (x(i)<0.75))
```

```
     y(i)=1;
   else
     y(i)=5;
   end
 end
end
endfunction

function [y] = escalier3(x)
n=size(x,'*');
y=x;
for i=1:n
  if ((0<=x(i)) & (x(i)<0.25))
    y(i)=4;
  else
   if ((0.25<=x(i)) & (x(i)<0.9))
     y(i)=2;
   else
     y(i)=1;
   end
 end
end
endfunction

function [y] = mon_sinus(x)
y=sin(2*%pi*x)+1;
endfunction

function [y] = mon_sinus2(x)
y=sin(4*%pi*x)+1;
endfunction

function [y] =mon_exponentielle(x)
y=9*x.*exp(-9*x)+1;
endfunction

function [y] =mon_expsinus(x)
y= exp(3*x).*(sin(4*%pi*x)+1)+1;
```

```
endfunction

// Pour appliquer F à un vecteur de valeurs de t

function [y] = appliqueF(vec,s)
y = vec;

// a: Calcul intégrale de s

y(1) = 0.0;
for i=2:size(vec,'*')
  y(i) = y(i-1) + intg( vec(i-1), vec(i), s);
end

// b: calcul de F

y = 1-exp(-y);
endfunction


// Réciproque de Y = f(X) pour Y = y0

function [x] = reciproque( vecX, vecY, y0 )

// Trouver l'intervalle dans lequel est compris y0, vecY croissant

n = size(vecY,'*');
x=y0;
for j = 1:size(y0,'*')

  // Test la valeur de y0

  if ( y0(j) >= vecY(n))
    x(j) = 1.0;
  else

    // Cas général
```

```
    if ( y0(j) <= vecY(1))
      x(j) = 0.0;
    else
      for i=2:n
        if (y0(j) < vecY(i) )
          break // sortie de boucle
        end
      end

      // Interpolation lineaire

      x(j)=vecX(i-1)+(vecX(i)-vecX(i-1))/(vecY(i)-vecY(i-1))*(y0(j)-vecY(i-1));

    end
  end
end
endfunction

function [x] = reciproque_ana(f,y_0)

 // Pour faire la reciproque d'une fonction et non plus d'un tableau

 deff('[z]=g(x)','z=f(x)-y_0');
 x=fsolve(0,g);

endfunction
```

Avec le script **genf.sce**, on génère F associé à s et notée dans le programme Fbase, en tant que tableau de 1000 valeurs.

```
// Génération de F

getf("mes_fonctions.sci");
Nbase = 1000;
xbase = [1:Nbase]'/Nbase;

//s=un;
//s=deux;
//s = escalier;
```

```
//s = escalier1;
//s=escalier2;
//s= escalier3;
//s=up_linear;
//s=down_linear;
//s=mon_sinus;
//s= mon_sinus2;
//s = mon_exponentielle;

s=mon_expsinus;

Fbase = appliqueF( xbase, s);
```

Les outils de base pour créer à partir de X et U, $\tilde{X}$ (notée T dans le programme) et D sont dans **ma_base.sci**, ainsi que les histogrammes qui vont avec (c'est à dire des coefficients, N_I, b_I et le coefficient correspondant suivant la taille de b_I). Une partition est repérée par ses points, et chaque intervalle par son extrémité gauche.

```
// A est la duree de vie, B est la censure

function [y,ind] = mon_minimum( A, B)
y = A;
ind = y;
n =size( A, '*');

for i = 1:n

  y(i)    = A(i);
  ind(i)  = 1.0;

  if ( B(i) < A(i) )
    y(i)   = B(i);
    ind(i) = 0.0;
  end

end

endfunction
```

```
//Construction de l'histogramme

function [NI,bI,hst] = mon_hist(partition,observation,indicateur)

// initialisation

hst=0*partition;
NI=hst;
nobs = size(observation,'*');
npart = size(partition,'*');
bI = hst;

// Passage en force

for j=1:npart-1

  for i=1:nobs

    tmp = observation(i);


    // cas 1
    //if ( tmp < partition(j) )
    // rien du tout
    //end

    // cas 2

    if ((tmp>=partition(j))&(tmp < partition(j+1)))//observation dans intervalle j
      bI(j) = bI(j) + (tmp - partition(j));

      if (indicateur(i) <> 0)
        NI(j) = NI(j) +1;
      end

   end

    // cas 3
```

```
    if ( tmp >= partition(j+1))
      bI(j) = bI(j) + (partition(j+1) - partition(j));
    end

  end

end

tmp = 1/nobs;
for j=1:npart-1

  if ( bI(j) >= tmp)
    hst(j) = NI(j) / bI(j);
  end

end

endfunction
```

Maintenant on peut générer les variables par le script **genobs.sce** : il suffit de changer $Nvar$ pour modifier le nombre de variables tirées. Ici la censure est prise uniforme sur $[0, 2]$ mais on peu très facilement changer cela. Par exemples si on prend $U = ones$, cela revient à ne pas mettre de censure du tout. Tant qu'on y est et puisqu'on va en avoir besoin ensuite, on construit la fonction Y appelée ici "decroit" et les intégrales contre Y et on compte le nombre d'événements entre 0 et 1 censurés ou non.

```
//Vecteur iid

Nvar = 1000;
Xtilde = rand(1,Nvar,'uniform')';

//Génération  vecteur réparti selon F

X = reciproque( xbase, Fbase, Xtilde);

// Génération des censures

//U = ones(X);
Utilde = rand(1,Nvar,'uniform');
```

```
U=reciproque(xbase,xbase/2,Utilde);

//Génération des observations

getf("ma_base.sci");
[T,D] = mon_minimum( X, U);

// On batît la fonction Yprevisible (decroit)
// et la façon de calculer integral( f*Y )

Tord = gsort(T,'g','i'); // On range

x_ord=list();
y_ord=list();

x_ord(1) = 0;
y_ord(1) = Nvar;


//Détermination des classes
idx=1;
for i=1:size(Tord,'*')

  if ( Tord(i) <> x_ord(idx) )
    idx=idx+1;
    x_ord(idx) = Tord(i);
    y_ord(idx) = y_ord(idx-1) -1;

  else
    y_ord(idx) = y_ord(idx)-1;

  end

end


n_ord = size(x_ord);
X_ord = [1:n_ord]';
```

```
Y_ord = X_ord;

for i=1:n_ord

  X_ord(i) = x_ord(i);
  Y_ord(i) = y_ord(i);

end

n_ordm1 = n_ord-1;

// Obtention de la fonction decroit
// Attention !!! fonction discontinue, à intégrer par morceaux

function [y] = decroit(t)

res = find( X_ord(1:n_ordm1) <= t & t <= X_ord(2:n_ord) );
y = Y_ord( res(1) );

endfunction

// Integration par morceaux de Yprevisible(t) * f(t) entre a et b
// Attention !!! a et b dans [0:1] (large)

function [y] = integre_previsible( a,b, f)

y = 0.0;
idx_a = find( X_ord(1:n_ordm1) <= a & a <= X_ord(2:n_ord) );
idx_b = find( X_ord(1:n_ordm1) <= b & b <= X_ord(2:n_ord) );

//Premier terme
y = intg( a, X_ord( idx_a+1), f)*Y_ord(idx_a);

for i= (idx_a+1) : (idx_b-1)
  y = y + intg( X_ord(i), X_ord(i+1), f)*Y_ord(i);
end

//Dernier terme
```

```
y = y+intg( X_ord(idx_b), b, f)*Y_ord(idx_b);

endfunction

//Calcul du nombre de vus et du nombre de non censures
Nvrai=0;
Nnoncens=0;

for i=1:Nvar

  if T(i)<1

    Nvrai=Nvrai+1;

    if D(i)==1
      Nnoncens=Nnoncens+1;

    end

  end

end
```

B.2. Stratégie en histogrammes

Les fonctions qui sont dans **parti.sci** permettent de construire les listes de partitions emboîtées ou exhaustives sur une grosse partition fixée.

```
function [y] = factorielle(n)

if (n > 1)

  y = n*factorielle( n -1);

else

  y = 1;
```

```
end

function [y] = Comb(n,p)

y = factorielle(n)/(factorielle(p)*factorielle(n-p));

endfunction

// Génération de partition

//fonction annexe evolution


// idx: index source, k=valeur maxi

function [ind] = y_evolution(idx,k)

ind = idx;
n = size(ind,'*');

//Première étape
ind(n) = ind(n) + 1;

if (ind(n) >= k) // il faut faire bouger le reste

  tmp = y_evolution( idx(1:n-1), k-1);

  // Boucle qui marche
  //for j=1:n-1
  //  ind(j) = tmp(j);
  //end

  ind(1:n-1) = tmp;
  ind(n) = ind(n-1)+1;

end
```

```
endfunction

//Les partitions exhaustives

function [list_partition] = gen_partitions_exhaust(d)

pas = 1.0/d;


// Partition triviale
list_partition = list();
n=1;

//Principe
list_partition(n) = [0:1]';

//Tableau d'indice

// Boucle sur le nombre de points centraux

for i=1:d-1

  //Création du tableau d'indices initialisés

  idx=[1:i]';

 // -> i+2 points pour définir la partition

  vec = [1:i+2]';

// Valeurs de cette partition

  vec(1)  = 0.0;

  //for j = 1:i
  //  vec( j+1 ) = (idx(j))/d;
  //end
```

```
  vec(2:1+i) = idx *pas;

  vec(i+2) = 1.0;

//Ajouter à la liste
  n=n+1;
  list_partition(n) = vec;

//Boucle d'évolution des indices: il doit
  // y avoir Comb( (d-1), i)
  for k=1: (Comb( (d-1), i)-1)

    tmp = y_evolution( idx, d );
    idx = tmp;
    n=n+1;


    //Création de la sous partition
    //for j=1:i
    //  vec(j+1) = (idx(j))/d;
    //end

    vec(2:1+i) = idx *pas;

    // Addition de la partition

    list_partition(n) = vec;

  end

end

endfunction

//Les partitions emboîtées

function [emb]=gen_partitions_emboitees(J)
```

```
emb=list();

for i=1:(J+1)

  emb(i)=(0:2^(i-1))'/(2^(i-1));

end

endfunction
```

Dans **critere.sci**, on a les fonctions qui permettent de calculer le critère pénalisé.

```
// Fonction qui donne le critere penalisee, connaissant la partition
//les observations, le c de la penalite et le poids Lm de la partition
//et R l'estimateur du max sur la plus grosse partition

function [crit] = crit_pena(partition,observation, indicateur ,c,R,lm)

y=0;
npart=size(partition,'*');
nobs=size(observation,'*');
[N,b,hist]=mon_hist(partition,observation,indicateur);

for i=1:(npart-1)

  y=y+hist(i)^2*b(i)/(nobs);

end

//R=max(hist);
crit=-y+c*(npart-1)*R*(1+sqrt(2*lm))^2/nobs;

endfunction
```

Dans **risk2.sci**, on a les fonctions qui permettent de calculer le risque.

```
// Fonction de calcul du risque
// calcul des a(i)=int_i s Y  dt pour une partition fixee m
```

```
function [a]= coeffproj(s,m)

a=m;
tmp = size(m,'*');

for i=1:tmp-1

  a(i)=integre_previsible(m(i),m(i+1),s);

end

a(tmp) = 0.0;

endfunction


//calcul du risque connaissant la partition l'histogramme,
// les b correspondants et la vraie s

function [rsk] = calcul_risk( the_partition, the_hst, the_b, s)

deff( '[z]=s_sq(x)', 'z=s(x)^2');
rsk =  integre_previsible(0.0,1.0,s_sq)/Nvar;
n  = size(the_partition,'*');
a  = coeffproj(s,the_partition);

for i=1:n-1

  rsk = rsk -2*the_hst(i)*a(i)/Nvar+the_hst(i)^2*the_b(i)/Nvar;

end

endfunction
```

Le script **strategyhist_emboit.sce** trace au final le p.p.e. choisi par stratégie emboîtée avec une pénalité en $2\tilde{R}|m|/n$ et le vrai taux de hasard s. Il donne aussi le risque de l'estimateur.

```
// Chargement parti.sci

getf("parti.sci");
J= int( log( (Nvar/log(Nvar)^2) )/log(2) );

emb=gen_partitions_emboitees(J);


// Chargement du modèle
getf("critere.sci");

// Paramètres du modèle
c  = 2;
lm = 0.0;

nc = size(emb);
critere=zeros(1,nc)';

//calcul de R le majorant estimé de s

[N_maj, b_maj, hist_maj] = mon_hist( emb(nc), T, D);
R=0.0;

for i=1:size(N_maj,'*')

  if ( b_maj(i) > 0.0)
    tmp = N_maj(i) / b_maj(i);

    if (tmp > R)
      R = tmp;

    end

  end

end

for i=1:nc
```

```
  critere(i) = crit_pena(emb(i), T, D, c, R, lm);

end

[crit_min, idx_min] = min(critere);

[N_min, b_min, hst_min] = mon_hist(emb(idx_min),T,D);

// Affichage du modèle sélectionné

xbasc();xselect();
plot2d2(emb(idx_min),hst_min,5);
plot2d(xbase,s(xbase),2);

getf("risk2.sci");
//Calcul du risque

Risque_aleatoire=calcul_risk(emb(idx_min),hst_min,b_min,s);


//affichage de Nvrai, Nnoncens, Risque

xtitle('bleu = la vraie, noir = le p.p.e.histogramme emboites, c=2, lm=0');
xnumb([0.3 0.4 0.5],[0.1 0.1 0.1],[Nvrai Nnoncens Risque_aleatoire]);
```

Dans **strategyhist_exhaust.sce**, on fait la même chose pour une stratégie exhaustive avec la pénalité adaptée.

```
// Chargement parti.sci

getf("parti.sci");

//J= int( log( (Nvar/log(Nvar)^2) )/log(2) );

d=min(8, floor(Nvar/log(Nvar)^2));

emb=gen_partitions_exhaust(d);
```

```
//emb=gen_partitions_exhaust(2^J);

// Chargement du modèle

getf("critere.sci");

// Paramètres du modèle
c  = 0.4;

//lm = 0;

nc = size(emb);

critere=zeros(1,nc)';

[N_maj, b_maj, hist_maj] = mon_hist( emb(nc), T, D);

R=0.0;

for i=1:size(N_maj,'*')

  if ( b_maj(i) > 0.0)
    tmp = N_maj(i) / b_maj(i);

    if (tmp > R)
      R = tmp;

    end

  end

end

// les poids

for i=1:nc

  lm=log(Nvar/(size(emb(i),'*')-1));
```

```
  critere(i) = crit_pena(emb(i), T, D, c, R, lm);

end

[crit_min, idx_min] = min(critere);

[N_min, b_min, hst_min] = mon_hist(emb(idx_min),T,D);

// Affichage du modèle sélectionné

//xbasc();xselect();
plot2d2(emb(idx_min),hst_min);
//plot2d(xbase,s(xbase),2);

getf("risk2.sci");
//Calcul du risque

Risque_aleatoire=calcul_risk(emb(idx_min),hst_min,b_min,s);


//affichage de Nvrai, Nnoncens, Risque
xtitle('bleu = la vraie, noir = le p.p.e.histogramme exhaust, c=0.4, lm=ln(N/|m|)');
xnumb([0.3 0.4 0.5],[0.2 0.2 0.2],[Nvrai Nnoncens Risque_aleatoire]);
```

B.3. Stratégie Fourier

Dans **y_fourier.sci**, on calcule les coefficients dans le modèle prévisible associé à une base de Fourier ainsi que le critère.

```
//Calcul des coefficients de fourier
// associé à decroit(t)

function [a] = gen_coeff_fourier( K )
dim = 2*K + 1;

a=zeros(dim,1);

//Observations qui comptent
res = find( (T < 1) & (D>0) );
```

```
T_tmp = T(res);
decT = T_tmp;

for i=1:size(T_tmp,'*')

  decT(i) = 1/sqrt(decroit( T_tmp(i)));

end
// Calcul du premier terme
a(1) = sum( decT );


for j = 1:K
  trig_arg = 2*%pi*j;

  // Cas 2j: sinus
  a( 2*j ) = sum( decT.* sin(trig_arg*T_tmp));

  //Cas 2j+1: cosinus
  a( 2*j+1 ) = sum( decT.* cos(trig_arg*T_tmp) );

end

a(2:dim) = sqrt(2.0) * a(2:dim);

// Normalisation
a = a / sqrt( Nvar );

endfunction

function [est] = get_fourier_estim( a, x)
est=zeros(x);

ncoeff = size( a, '*');
sq2 = sqrt( 2.0);

for i=1:size(x,'*')
```

```
  y = decroit( x(i));

 if ( y > 0.0 )
    tmp = 1.0 / sqrt( y );
    est(i) = a(1);
    for j=1:floor( (ncoeff-1)/2)

      trig_arg = 2*j*%pi;
      est(i)=est(i)+sq2*(a(2*j)*sin(trig_arg*x(i))+a(2*j+1)*cos(trig_arg*x(i)));

    end

    est(i) = est(i) * tmp;

  end

end

est = est * sqrt( Nvar );

endfunction

//calcule le critere en fourier avec la mauvaise concentration demontree

function [crit] = criteref (a,c,R,lm)

n=size(a,'*');
a2 = a^2;
Ktmp=ceil((n-1)/2);
disp(Ktmp)
crit =[0:Ktmp]';

for i=1:(Ktmp+1)

 crit(i)=-sum(a2(1:(2*(i-1)+1)))+
              c*(2*(i-1)+1)/Nvar*(sqrt(R)*(1+sqrt(2*lm(i)))+lm(i)/3)^2;

end
```

```
endfunction

//calcule le critere en fourier avec la concentration utopique

function [crit] = criteref2 (a,c,R,lm)

n=size(a,'*');
a2 = a^2;
Ktmp=ceil((n-1)/2);
disp(Ktmp)
crit =[0:Ktmp]';

for i=1:(Ktmp+1)

  crit(i)=-sum( a2(1:(2*(i-1)+1)) )+c*(2*(i-1)+1)/ Nvar*(sqrt(R)+lm(i)/3)^2;

end

endfunction
```

Ensuite on trace comme précédemment les résultats pour une stratégie emboîtée dans le script **strategyfourier.sce**. On peut aussi tracer $\rho = s\sqrt{Y}$ et son vrai estimateur en base de Fourier.

```
// chargement y_fourier.sci
getf('y_fourier.sci');

//plus gros modele
K = int((Nvar-1)/2);

c=1;
lm=(0:K);
for i=1:(K+1)
 lm(i)= log ((2*i-1));
end

//calcul du critere
a = gen_coeff_fourier( K );
```

```
crit_pena_fourier= criteref2(a,c,R,lm);

//minimisation du critere

[crit_min_f, idx_min_f] = min(crit_pena_fourier);

// le k choisi
k_min_f = idx_min_f-1;
support_fourier = a(1:(2*k_min_f+1));

function [chaps] = chapeau_s (x)
chaps = get_fourier_estim(support_fourier, x);
endfunction

//xbasc();xselect();
fplot2d(xbase,chapeau_s);
// plot2d(xbase,s(xbase),2);

function [diff] = carrediff(x)
 diff= (s(x)-chapeau_s(x))^2;
endfunction

Risque_Fourier= integre_previsible (0,1,carrediff) /Nvar;

xtitle('bleu = la vraie, noir = le p.p.e.Fourier emboite, c=1, lm = ln D');
xnumb([0.3 0.4 0.5],[0.3 0.3 0.3],[Nvrai Nnoncens Risque_Fourier]);

function [chapssrt] = chapeau_s_sqrt(x)
 chapssrt=chapeau_s(x)*sqrt(decroit(x))/sqrt(Nvar);
endfunction

function [ssrt] = s_sqrt(x)
 ssrt = s(x)*sqrt(decroit(x))/sqrt(Nvar);
endfunction

//fplot2d(xbase,chapeau_s_sqrt,3);
//fplot2d(xbase,s_sqrt,4);
```

Enfin, comme il était difficile au début de trouver l'oracle à l'oeil nu, on a fait un script qui l'affiche ainsi que son risque : c'est **oracle.sce**. Malheureusement il est très lent pour 1000 observations au départ par exemple.

```
// coeff de la projection

function [y] = integre_previsible_sqr( a,b, f)
y = 0.0;
idx_a = find( X_ord(1:n_ordm1) <= a & a <= X_ord(2:n_ord) );
idx_b = find( X_ord(1:n_ordm1) <= b & b <= X_ord(2:n_ord) );

//Premier terme
y = intg( a, X_ord( idx_a+1), f)*sqrt(Y_ord(idx_a));


for i= (idx_a+1) : (idx_b-1)

  y = y + intg( X_ord(i), X_ord(i+1), f)*sqrt(Y_ord(i));

end

//Dernier terme
y = y+intg( X_ord(idx_b), b, f)*sqrt(Y_ord(idx_b));

endfunction

// calculs des coeffs de la vrai projection

function [b]= coeffproj_fourier(s,K)

b=zeros(a);
b(1)= integre_previsible_sqr(0,1,s);

for i=1:K

  arg_tmp=2*i*%pi
  deff('[z]=fbis(x)','z=sin(arg_tmp*x)*s(x)');
  b(2*i)=integre_previsible_sqr(0,1,fbis);
```

```
  deff('[z]=fter(x)','z=cos(arg_tmp*x)*s(x)');
  b(2*i+1)=integre_previsible_sqr(0,1,fter);

end

b(2:(2*K+1))=b(2:(2*K+1))*sqrt(2);
b=b/sqrt(Nvar);

endfunction

b=coeffproj_fourier(s,K);

//calcul des differents risques

risk = zeros(crit_pena_fourier);
deff( '[z]=s_sq(x)', 'z=s(x)^2');

// le 1er terme

risk_tmp(1)=  integre_previsible(0.0,1.0,s_sq)/Nvar-2*a(1)*b(1)+a(1)^2;

//risk(1)=integre_previsible(0.0,1.0,s_sq)/Nvar-2*a(1)*b(1)+a(1)^2;

//les autres

risk_tmp(2:K+1)=-2.*a(2:K+1).*b(2:K+1)+a(2:K+1).^2;
risk = cumsum(risk_tmp);

//for i=2:(K+1)
  //risk(i) =risk(i-1)-2*a(i)*b(i)+a(i)^2;
//end

[riskoracle, idx_oracle]=min(risk);

k_or_f = idx_oracle-1;
support_fourier_or = a(1:(2*k_or_f+1));

function [chaps] = oracle_s (x)
```

```
chaps = get_fourier_estim(support_fourier_or, x);
endfunction

fplot2d(xbase,oracle_s,5);
```

B.4. Comparaison

Voici la manière dont ont été simulées les comparaisons avec l'estimateur en ondelettes de [**5**].

Tout d'abord voici le script pour la loi Gamma **ma_boucle.sce** :

```
NESSAI = 1;
Kanton = 64;
Nvar = 200;

Finrisk=6;

// la vraie fonction

function [y] = ma_gamma(x)
 u = x^4;
 v= 24+24*x+12*x^2+4*x^3+x^4;
 y = u./v;
endfunction

s = ma_gamma ;

// initialisation

tableau_riskhis=[1:NESSAI]';
tableau_riskhispartout=[1:NESSAI]';
tableau_riskfour = [1:NESSAI]';
tableau_riskfourpartout = [1:NESSAI]';

// la boucle des tirages successifs

for index_boucle =1:NESSAI
  tableau_risk(index_boucle)=-1.0;
  exec vraiebouc.sce;
```

```
  exec genobsbouc.sce;
  exec hist_embbouc.sce;
  exec fourierbouc.sce;
  xbasc();xselect();
  plot2d(Xnouveau,s(Xnouveau),1);
  plot2d2(mnouv,his_nouv,2);
  fplot2d(Xnouveau,chapeau_s_nouv,3);
  xnumb([1 2 3 4 5 6],[0.1 0.1 0.1 0.1 0.1 0.1],[Nvrai Nnoncens Riskanton Riskantonfour Ris
  tableau_riskhis(index_boucle)= Riskanton;
  tableau_riskhispartout(index_boucle)= Riskantonpartout;
  tableau_riskfour(index_boucle)= Riskantonfour;
  tableau_riskfourpartout(index_boucle)= Riskantonfourpartout;
  index_boucle
end

AMSEhis = sum(tableau_riskhis)/NESSAI
AMSEfour = sum(tableau_riskfour)/NESSAI

AMSEhispart = sum(tableau_riskhispartout)/NESSAI
AMSEfourpart = sum(tableau_riskfourpartout)/NESSAI
```

Voici les programmes appelés par la boucle.

Tout d'abord on simule les variables aléatoires avec **vraiebouc.sce** pour la loi Gamma :

```
// Generation des vraies variables
// Ucensure exponentielle de moyenne 6
// X gamma(5,1)

Utilde = rand(1,Nvar,'uniform');
Utmp = -6*log(1-Utilde);

//y=gsort(Utmp,'g','i');

//xbasc();
//xselect();

//plot2d2(y,(1:Nvar)/Nvar,1);
```

```
//x=(0:0.01:30);

//plot2d(x,(1-exp(-x/6)),2);


Xtilde = rand(1,Nvar,'uniform');
Xtmp = Xtilde;
for i=1:Nvar
  Xtmp(i) = cdfgam("X",5,1,Xtilde(i),1-Xtilde(i));
end

//z = gsort(Xtmp,'g','i');

//plot2d2(z,(1:Nvar)/Nvar,3);

//w = 1+x+x^2./2+x^3./6+x^4./24 ;
//p = 1-exp(-x).*w;

//plot2d(x,p,5);
```

Puis on a les vraies observations censurées ramenées dans $[0,1]$ par **genobsbouc.sce** :

```
//Les observations
getf("ma_base.sci");
[Ttmp,D] = mon_minimum( Xtmp, Utmp);

// On se remet sur [0,1]
Fin = max(Ttmp) ;

T = Ttmp/Fin ;

// On batît la fonction Yprevisible (decroit)
// et la façon de calculer integral( f*Y )

Tord = gsort(T,'g','i'); // On range

x_ord=list();
y_ord=list();
```

```
x_ord(1) = 0;
y_ord(1) = Nvar;


//Détermination des classes
idx=1;
for i=1:size(Tord,'*')

  if ( Tord(i) <> x_ord(idx) )
    idx=idx+1;
    x_ord(idx) = Tord(i);
    y_ord(idx) = y_ord(idx-1) -1;
  else
    y_ord(idx) = y_ord(idx)-1;
  end

end


n_ord = size(x_ord);
X_ord = [1:n_ord]';
Y_ord = X_ord;

for i=1:n_ord
  X_ord(i) = x_ord(i);
  Y_ord(i) = y_ord(i);
end
n_ordm1 = n_ord-1;

// Obtention de la fonction decroit
// Attention !!! fonction discontinue, à intégrer par morceaux

function [y] = decroit(t)
res = find( X_ord(1:n_ordm1) <= t & t <= X_ord(2:n_ord) );
y = Y_ord( res(1) );
endfunction

// Integration par morceaux de Yprevisible(t) * f(t) entre a et b
```

```
// Attention !!! a et b dans [0:1] (large)
function [y] = integre_previsible( a,b, f)
y = 0.0;
idx_a = find( X_ord(1:n_ordm1) <= a & a <= X_ord(2:n_ord) );
idx_b = find( X_ord(1:n_ordm1) <= b & b <= X_ord(2:n_ord) );

//Premier terme
y = intg( a, X_ord( idx_a+1), f)*Y_ord(idx_a);

for i= (idx_a+1) : (idx_b-1)
  y = y + intg( X_ord(i), X_ord(i+1), f)*Y_ord(i);
end

//Dernier terme
y = y+intg( X_ord(idx_b), b, f)*Y_ord(idx_b);

endfunction

//Calcul du nombre de vus et du nombre de non censures
Nvrai=0;
Nnoncens=0;
for i=1:Nvar
  if T(i)<1
    Nvrai=Nvrai+1;
    if D(i)==1
      Nnoncens=Nnoncens+1;
    end
  end
end
```

Ensuite on reprend l'estimateur par histogrammes emboités en modifiant juste le calcul du risque à la fin pour que ca donne le critère AMSE et AMSE2. C'est le fichier **hist_embbouc.sce**.

```
// Chargement parti.sci
getf("parti.sci");
// Nombre de possibilités jusqu'a 16 intervalles
J = int( log( (Nvar/log(Nvar)^2) )/log(2) );
```

```
//generation des partitions

emb=gen_partitions_emboitees(J);


// Chargement du modèle
getf("critere.sci");

// Paramètres du modèle
c  = 2;
lm = 0.0;

nc = size(emb);
critere=zeros(1,nc)';

//calcul de R le majorant estimé de s

[N_maj, b_maj, hist_maj] = mon_hist( emb(nc), T, D);
R=0.0;
for i=1:size(N_maj,'*')
  if ( b_maj(i) > 0.0)
    tmp = N_maj(i) / b_maj(i);
    if (tmp > R)
      R = tmp;
    end
  end
end

for i=1:nc
  critere(i) = crit_pena(emb(i), T, D, c, R, lm);
end

[crit_min, idx_min] = min(critere);

[N_min, b_min, hst_min] = mon_hist(emb(idx_min),T,D);
// Affichage du modèle sélectionné
//xbasc();xselect();
//plot2d2(emb(idx_min),hst_min,5);
```

```
//plot2d(xbase,s(xbase),2);


// on remet à l'échelle

Xnouveau = (0:0.001:Fin);


//plot2d(Xnouveau,s(Xnouveau),1);

mnouv = Fin*emb(idx_min);
his_nouv = hst_min./Fin;

//plot2d2(mnouv,his_nouv,2);

// Calcul du risque version antoniadis

// Nombre intervalles


tAnton = (0:Fin/Kanton:Fin);

// calcul de l'estimateur en tk

// initialisation

snouv_chap = (1:Kanton)';

for i=1:Kanton
  numero = find(mnouv>tAnton(i));
  ntemp = numero(1);
  snouv_chap(i) = his_nouv(ntemp-1);
end

// et la vraie

snouv_vraie = s(tAnton(1:Kanton));
```

```
// et le risque

// on ne somme que sur les tAnton <Finrisk

//initialisation

Riskanton=0;
nanton=0;

for i=1:Kanton
  if tAnton(i)<=Finrisk
    Riskanton = Riskanton + (snouv_chap(i)-snouv_vraie(i))^2;
    nanton = nanton+1;
  end

end


Riskanton = Riskanton / nanton;

Riskantonpartout=0;
for i=1:Kanton
Riskantonpartout = Riskantonpartout + (snouv_chap(i)-snouv_vraie(i))^2;
end
Riskantonpartout=Riskantonpartout/Kanton;

//utmp= (snouv_chap-snouv_vraie)^2;

//Riskanton = sum(utmp)/Kanton;

//legends('la vraie', 'histo emb');
//xnumb([1 2 3],[0.1 0.1 0.1],[Nvrai Nnoncens Riskanton]);
```

On fait pareil pour l'estimateur en Fourier : c'est **fourierbouc.sce**.

```
// chargement y_fourier.sci
getf('y_fourier.sci');

//plus gros modele
```

```
K = int((Nvar-1)/2);

c=1;
lm=(0:K);
for i=1:(K+1)
 lm(i)= log ((2*i-1));
end

//calcul du critere
a = gen_coeff_fourier( K );

crit_pena_fourier= criteref2(a,c,R,lm);

//minimisation du critere

[crit_min_f, idx_min_f] = min(crit_pena_fourier);

// le k choisi
k_min_f = idx_min_f-1;
support_fourier = a(1:(2*k_min_f+1));

function [chaps] = chapeau_s (x)
chaps = get_fourier_estim(support_fourier, x);
endfunction

//xbasc();xselect();
//fplot2d(xbase,chapeau_s);
// plot2d(xbase,s(xbase),2);


function [chaps] = chapeau_s_nouv (x)
chaps = chapeau_s(x/Fin)/Fin;
endfunction

snouv_four =(1:Kanton);

// on remet a l'echelle
for i=1:Kanton
```

```
  snouv_four(i) = chapeau_s_nouv(tAnton(i));
end
//fplot2d(Xnouveau,chapeau_s_nouv,3);
Riskantonfour=0;

for i=1:nanton
  Riskantonfour = Riskantonfour +(snouv_four(i)-snouv_vraie(i))^2;
end

Riskantonfour = Riskantonfour / nanton;

Riskantonfourpartout=0;

for i=1:Kanton
  Riskantonfourpartout = Riskantonfourpartout +(snouv_four(i)-snouv_vraie(i))^2;
end

Riskantonfourpartout = Riskantonfourpartout / Kanton;
//legends('la vraie', 'histo emb');
//xnumb([1 2 3],[0.2 0.2 0.2],[Nvrai Nnoncens Riskantonfour]);
```

Pour la loi bimodale, il suffit d'enlever **vraiebouc.sce** et de le remplacer par **simbouc2.sce** : ca donne le programme **ma_boucle2.sce**.

```
NESSAI = 1;
Kanton = 64;
Nvar = 5000;
//Fin = 2.5;
Finrisk=2;
// la vraie fonction


function [y] = ma_stone(x)
y=x;
for i=1:size(x,'*')
  if x(i)>0
    u = sqrt(2/%pi)*x(i)^{-2*log(x(i))-1};
    tmp = 2*log(x(i));
```

```
    [P1,Q1]=cdfnor("PQ",tmp,0,1);
  else
    u = 0;
    Q1 = 1;
  end
  v = 1/(0.17*sqrt(2*%pi))*exp(-1/2*((x(i)-2)/0.17)^2);
  z = 0.8*u + 0.2*v;
  [P,Q]=cdfnor("PQ",x(i),2,0.17);
  w = 0.8*Q1 + 0.2*Q;
  y(i) = z/w;
end
endfunction
s = ma_stone ;

// intialisation

tableau_riskhis=[1:NESSAI]';
tableau_riskhispartout=[1:NESSAI]';
tableau_riskfour = [1:NESSAI]';
tableau_riskfourpartout = [1:NESSAI]';
MAX=(1:NESSAI)';
MAXT=(1:NESSAI)';

//boucle

for index_boucle =1:NESSAI
  tableau_risk(index_boucle)=-1.0;
  exec simbouc2.sce;
  exec genobsbouc.sce;
  MAX(index_boucle)=max(Xtmp);  MAXT(index_boucle) = max(Ttmp);
  exec hist_embbouc.sce;
  exec fourierbouc.sce;
  xbasc();xselect();
  plot2d(Xnouveau,s(Xnouveau),1);
  plot2d2(mnouv,his_nouv,2);
  fplot2d(Xnouveau,chapeau_s_nouv,3);
  xnumb([0.4 0.8 1.2 1.6 2 2.4],[0.1 0.1 0.1 0.1 0.1 0.1],
           [Nvrai Nnoncens Riskanton Riskantonfour Riskantonpartout
```

```
                                                    Riskantonfourpartout]);

  tableau_riskhis(index_boucle)= Riskanton;
  tableau_riskhispartout(index_boucle)= Riskantonpartout;
  tableau_riskfour(index_boucle)= Riskantonfour;
  tableau_riskfourpartout(index_boucle)= Riskantonfourpartout;
  index_boucle
end

AMSEhis = sum(tableau_riskhis)/NESSAI
AMSEfour = sum(tableau_riskfour)/NESSAI

AMSEhispart = sum(tableau_riskhispartout)/NESSAI
AMSEfourpart = sum(tableau_riskfourpartout)/NESSAI

moymax = sum(MAX)/NESSAI
moymaxT =  sum(MAXT)/NESSAI
```

Le programme **simbouc2.sce** est le suivant :

```
// Simulation 2 de  Antoniadis-Gregoire
// simulation de la loi normale (2,0.17)

//Nvar=200;



Vtilde = rand(1,Nvar,'normal');
V = 0.17*Vtilde+2;

//xbase = (0.001:0.001:3.5);

//n=size(xbase,'*');
//y=xbase;
//for i=1:n
//  [P,Q]=cdfnor("PQ",xbase(i),2,0.17);
//  y(i)= P;

//end
```

```
//z = gsort(V,'g','i');
//xbasc();xselect();
//plot2d2(z,(1:Nvar)/Nvar,1);

//plot2d(xbase,y,2);


//simulation de la loi log normale

Wtilde = rand(1,Nvar,'normal');
W=exp(Wtilde./2);

//for i=1:n
//  tmp = 2*log(xbase(i));
//  [P,Q]=cdfnor("PQ",tmp,0,1);
//  x(i)= P;
//end

//d= gsort(W,'g','i');

//xbasc();xselect();
//plot2d2(d,(1:Nvar)/Nvar,3);

//plot2d(xbase,x,4);


B = rand(1,Nvar,'uniform');

Xtmp = W;

for i=1:Nvar
  if B(i)<=0.2
    Xtmp(i) = V(i);
  end

end
```

```
//h = gsort(Xtmp,'g','i');

//xbasc(); xselect();

//plot2d2(h,(1:Nvar)/Nvar,1);

//plot2d(xbase,0.8*x+0.2*y,2);

//Simulation de la censure

Utilde = rand(1,Nvar,'uniform');
Utmp = -2.5*log(1-Utilde);


//ju=gsort(Utmp,'g','i');

//xbasc();
//xselect();

//plot2d2(y,(1:Nvar)/Nvar,3);

//x=(0:0.01:30);

//plot2d(xbase,(1-exp(-xbase/2.5)),4);
```

Bibliographie

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

[2] P.K. Andersen, O. Borgan, R. Gill, and N. Keiding. *Statistical Models Based on Counting Processes.* Springer Series in Statistics, 1993.

[3] F.J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika, Cambridge 35, 246-254* ., 1948.

[4] A. Antoniadis. A penalty method for nonparametric estimation of the intensity function of a counting process. *Ann. Inst. Statist. Math.*, 41(4) :781–807, 1989.

[5] A. Antoniadis, G. Grégoire, and G Nason. Density and hazard rate estimation for right-censored data by using wavelet methods. *J. R. Statist. Soc.*, 61(Part 1) :63–84, 1999.

[6] Y. Baraud. Model selection on a random design. 2001.

[7] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 2000.

[8] A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3) :1347–1369, 1991.

[9] L. Barron, A. ; Birgé and P. Massart. Risk bounds for model selection via penalization. *P.T.R.F.*, 1999.

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

[11] L. Birgé and P. Massart. Model selection from a nonasymptotic view point. Book in preparation.

[12] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[13] L. Birgé and P Massart. Minimum contrast estimators on sieves : Exponential bounds and rates of convergence. *Bernoulli 4, No.3, 329-375. [ISSN 1350-7265]*, 1998.

[14] L. Birgé and P. Massart. Gaussian model selection. *To appear in Journal of the European Mathematical Society*, 2001.

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

[16] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. 2001.

[17] P. Brémaud. *Point processes and queues.* Springer-Verlag, 1981.

[18] J. Bretagnolle. A new large deviation inequality for U-statistics of order 2. *ESAIM : Probability and Statistics*, 1999.

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

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

[21] G. Castellan and F. Letué. Estimation of the cox regression function via model selection. in F. Letué's PhD Thesis, UPS, 2001.

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

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

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

[25] R. DeVore and G. Lorentz. *Constructive approximation*. Springer-Verlag, 1993.

[26] S. Döhler and L. Rüschendorf. Adaptive estimation of hazard functions. 2000.

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

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

[29] M. Fromont and B. Laurent. Test d'adéquation dans un modèle de densité. Paris XI.

[30] R.G. Gallager. *Information theory and reliable communication*. New York-London-Sydney-Toronto : John Wiley and Sons, Inc. XVI, 588 p., 1968.

[31] E. Giné, R. Latala, and J. Zinn. Exponential and moment inequalities for U-statistics. *High Dimensional Prob. II*, 2000.

[32] G. Grégoire. Lest squares cross-validation for counting process intensities. *Scand. J. of Statist.*, 1993.

[33] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.*, XIX(3), 1948.

[34] W. Hoeffding. Probability inequalities for sums of bounded random variables . *Journal of the American Statistical Association*, 58, 1963.

[35] C. Houdré. Remarks on deviation inequalities for functions of infinitely divisible random vectors. *Ann. Prob.*, 2001.

[36] C. Houdré and N. Privault. Concentration and deviation inequalities in infinite dimensions via covariance representations. To appear in Bernoulli.

[37] O. Kallenberg. *Foundations of modern probability*. Springer, 1997.

[38] G. Kerkyacherian and D. Picard. 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.

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

[40] J.F.C. Kingman. *Poisson processes*. Oxford Studies in Probability., 1993.

[41] E. D. Kolaczyk. Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Stat. Sin. 9, No.1, 119-135 (1999). [ISSN 1017-0405]*, 1999.

[42] C. Kooperberg, C.J. Stone, and Y.K. Truong. The L_2 rate of convergence for hazard regression. *Scand. J. Statist.*, 22 :143–157, 1995.

[43] Yu.A. Kutoyants. *Statistical inference for spatial Poisson processes*, volume 134. Lecture Notes in Statistics, springer edition, 1998.

[44] B. Laurent. Adaptive estimation of a quadratic functional of a density by model selection. Paris XI.

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

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

[47] C.L. Mallows. Some comments on C_p. *Technometrics 15, 661-675*, 1973.

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

[49] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Prob.*, 18(3) :1269–1283, 1990.

[50] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Proba.*, 2000.

[51] P. Massart. Some applications of concentration inequalities. *Ann. de Toulouse*, 2000.

[52] J. Neveu. *Discrete-Parameter Martingales. Translated by T. P. Speed.* American Elsevier Publishing Company, Inc., 1975.

[53] I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Prob.*, 22 :1679–1706, 1994.

[54] H. Ramlau-Hansen. Smoothing counting process intensity by means of kernel functions. *Ann. Stat.*, 11 :453–466, 1983.

[55] L. Reboul. *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.

[56] P. Reynaud-Bouret. Concentration inequalities for inhomogeneous Poisson processes and adaptive estimation of the intensity. Technical Report 18, Université de Paris-Sud, 2001.

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

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

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

[60] W. Rudin. *Analyse réelle et complexe*. MASSON, 1987.

[61] P.-M. Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *Ann. Prob.*, 2000.

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

[63] Sara van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Stat.*, 23(5) :1779–1801, 1995.

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