Introduction aux méthodes de sélection de modèle

Gendre Xavier

Résumé

De façon générale, les statistiques ont pour objectif de retrouver des informations sur la loi d'un phénomène aléatoire que l'on observe. Ces informations peuvent être de nature très différentes selon le problème posé et pour les retrouver, nous sommes souvent amenés à formuler des hypothèses sur la loi elle-même bien qu'elle soit inconnue. Bien entendu, dans la pratique, ces hypothèses ne sont pas gratuites et il est donc important de pouvoir travailler avec peu d'hypothèses. Il s'agit là d'une des motivations de la sélection de modèle : fournir des méthodes dans des cadres généraux qui soient aussi "robustes" que possible. Dans ce genre d'étude, on suppose, par exemple, souvent la variance du phénomène observé connue. L'hétéroscédasticité correspond au fait que l'on ne connait pas cette variance et que l'on ne fait pas non plus l'hypothèse de sa constance au cours du temps.

1 Sélection de modèle à variance connue

1.1 Le cadre et les premiers outils

Pour présenter les grandes lignes de la sélection de modèle, nous allons commencer par voir ce qu'il en est lorsque la variance est connue et constante. On se place donc dans le cadre statistique de **régression** suivant : pour i allant de 1 à n, on observe

$$Y_i = s_i + \sigma\varepsilon_i$$

où $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$ est inconnu, $\sigma > 0$ est connu et $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ est un vecteur aléatoire dont les composantes sont indépendantes, centrées et de variance égale à 1. Nous voulons estimer le vecteur s. La notion d'estimateur est importante en statistique, on appelle **estimateur** toute variable aléatoire ne dépendant que des observations $Y_1, \ldots, Y_n$.

Munissons $\mathbb{R}^n$ d'une structure hilbertienne en prenant la norme suivante

$$\forall x \in \mathbb{R}^n,\ \|x\|_n = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{1/2} .$$

Commençons par estimer s sur un **modèle** S_D, c'est-à-dire un sous-espace vectoriel de $\mathbb{R}^n$, de dimension D. La "meilleure" approximation de s dans S_D est sa projection orthogonale s_D, c'est-à-dire le minimiseur, parmi les $t \in S_D$, de

$$\|s-t\|_n^2 = \|s\|_n^2 + \|t\|_n^2 - 2\langle s,t\rangle_n$$

ou, de façon équivalente, puisque $\|s\|_n^2$ est une constante, $\|t\|_n^2 - 2\langle s,t\rangle_n$. La quantité $\langle s,t\rangle_n$ dépend de s inconnu, cette procédure nous est donc inaccessible. Nous allons la remplacer par un estimateur **sans biais**, c'est-à-dire par un estimateur dont l'espérance vaut précisément $\langle s,t\rangle_n$. Ainsi nous sommes amenés à minimiser en $t \in S_D$ la quantité suivante

$$\gamma_n(t) = \|t\|_n^2 - 2\langle Y,t\rangle_n \ .$$

Il est simple de voir qu'il existe un unique minimiseur de γ_n dans S_D, on le notera $\hat{s}_D$ et on l'appelle **estimateur par projection**. Si l'on considère $\varphi_1,\ldots,\varphi_n$ une base orthonormale de $\mathbb{R}^n$ telle que S_D soit l'espace engendré par les $\varphi_1,\ldots,\varphi_D$, l'estimateur par projection s'écrit

$$\hat{s}_D = \sum_{j=1}^{D} \langle Y,\varphi_j\rangle_n \varphi_j \ .$$

Cette écriture est bien sûr à mettre en relation avec celle de la projection orthogonale de s sur S_D,

$$s_D = \sum_{j=1}^{D} \langle s,\varphi_j\rangle_n \varphi_j \ .$$

Ces deux écritures nous mènent aux deux égalités suivantes

$$\hat{s}_D = s_D + \sigma\sum_{j=1}^{D}\langle \varepsilon,\varphi_j\rangle_n\varphi_j \qquad \text{et} \qquad \gamma_n(\hat{s}) = -\|\hat{s}\|_n^2$$

qui nous serviront par la suite.

1.2 Risque quadratique de l'estimateur par projection

Il va maintenant nous falloir quantifier la qualité de notre estimateur. Une quantité classique pour le faire est le **risque quadratique**, c'est-à-dire l'espérance de $\|s - \hat{s}_D\|_n^2$. Un simple calcul donne

$$\mathbb{E}\left[\|s - \hat{s}_D\|_n^2\right] = \|s - s_D\|_n^2 + \frac{\sigma^2 D}{n} .$$

On appelle cette écriture une décomposition **biais-variance**. En effet, apparaissent les termes dits de biais $\|s - s_D\|_n^2$ et de variance $D\sigma^2/n$. Le premier correspond à la distance entre notre modèle et le véritable s tandis que le second traduit la complexité du modèle via la présence de la dimension D.

Lorsque D varie, ces deux termes ont des comportements opposés. En effet, si D augmente le terme de biais diminue et celui de variance augmente. Nous voudrions que le risque soit minimal et donc nous aimerions choisir un D qui donne un équilibre entre ces deux quantités.

1.3 Choix d'un modèle

Pour chercher cet équilibre biais-variance nous allons nous donner une famille finie de modèles $\{S_m\}_{m\in\mathcal{M}}$ et la famille des estimateurs par projection associée $\{\hat{s}_m\}_{m\in\mathcal{M}}$. Pour chaque S_m, on note D_m la dimension et s_m la projection orthogonale de s sur S_m.

Parmi les éléments de $\mathcal{M}$, il existe au moins un $\bar{m}$ tel que $\hat{s}_{\bar{m}}$ minimise le risque quadratique parmi les estimateurs par projection $\{\hat{s}_m\}_{m\in\mathcal{M}}$,

$$\bar{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right] = \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \|s - s_m\|_n^2 + \frac{\sigma^2 D_m}{n} \right\} .$$

Cependant, la connaissance de $\bar{m}$ nécessite celles des $\|s - s_m\|_n$ qui sont inconnues. Pour cette raison $S_{\bar{m}}$ est appelé **l'oracle**, il représente le meilleur modèle parmi tous ceux de notre famille.

On aimerait avoir une procédure, basée uniquement sur les observations, qui nous permette de choisir un $\hat{m} \in \mathcal{M}$ tel que l'estimateur $\tilde{s} = \hat{s}_{\hat{m}}$ vérifie, pour une certaine constante C, ce que l'on appelle un **inégalité oracle** :

$$\mathbb{E}\left[\|s - \tilde{s}\|_n^2\right] \leq C \inf_{m\in\mathcal{M}} \mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right] = C \inf_{m\in\mathcal{M}} \left\{ \|s - s_m\|_n^2 + \frac{\sigma^2 D_m}{n} \right\} .$$

Ce qui signifie que l'estimateur $\tilde{s}$ a des performances comparables à celles de $\hat{s}_{\bar{m}}$.

Notons que jusqu'à présent, nous n'avons pas utilisé le fait que nous connaissions la variance pour construire les différents objets. Cette hypothèse

va nous être utile maintenant, pour le choix de ce $\hat{m}$. Etant donné que l'on cherche à imiter l'oracle, une manière classique va être d'estimer le risque et de minimiser cet estimateur, c'est d'ailleurs ce qui fut fait par les premières études dues à Akaike (1) et Mallows (4). Une heuristique que l'on doit à Mallows, connue comme le "C_p de Mallows", nous guide vers la bonne façon d'estimer ce risque. Un modèle optimal est censé minimiser en m

$$\|s - s_m\|_n^2 + \frac{\sigma^2 D_m}{n} = \|s\|_n^2 - \|s_m\|_n^2 + \frac{\sigma^2 D_m}{n}$$

ou, de façon équivalente,

$$-\|s_m\|_n^2 + \frac{\sigma^2 D_m}{n} \ .$$

Un simple calcul donne $\mathbb{E}\left[\|\hat{s}_m\|_n^2\right] = \|s_m\|_n^2 + D_m\sigma^2/n$. L'heuristique consiste alors à remplacer $\|s_m\|_n^2$ par son estimateur sans biais et à prendre $\hat{m} \in \mathcal{M}$ qui minimise en m le critère

$$-\|\hat{s}_m\|_n^2 + \frac{2\sigma^2 D_m}{n} = \gamma_n(\hat{s}_m) + \frac{2\sigma^2 D_m}{n} \ .$$

On voit donc que pour faire ce choix, la connaissance de la variance σ^2 est indispensable.

En général, on s'intéresse plutot au problème posé dans les termes suivants : on choisit le $\hat{m}$ qui minimise le critère

$$\gamma_n(\hat{s}_m) + \mathrm{pen}(m)$$

où pen : $\mathcal{M} \longrightarrow \mathbb{R}_+$ est une fonction dite **pénalité**. L'estimateur ainsi construit $\tilde{s} = \hat{s}_{\hat{m}}$ est alors appelé **estimateur par projection pénalisé** (ou epp). La question est alors de comprendre les liens entre le choix de cette pénalité et les propriétés de l'epp, en particulier, a-t-on une inégalité oracle ?

L'heuristique de Mallows peut être validée sous certaines hypothèses sur la famille de modèles dès que ε admet un moment d'ordre p pour $p > 2$ (voir (2)). C'est par exemple le cas dans le cadre gaussien, c'est-à-dire si les ε_i sont des normales centrées réduites. Pour voir cela, faisons appel à un résultat dû à Birgé et Massart (voir (3)) valable dans le cadre gaussien :

Théorème 2 *Soit $\{x_m\}_{m\in\mathcal{M}}$ une famille de réels strictement positifs tels que*

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma < +\infty \ .$$

Supposons que la pénalité vérifie, pour une certaine constante $K > 1$,

$$pen(m) \geq \frac{K\sigma^2}{n}\left(\sqrt{D_m} + \sqrt{2x_m}\right)^2 \ .$$

L'epp $\tilde{s}$ correspondant est alors tel que

$$\mathbb{E}\left[\|s - \tilde{s}\|_n^2\right] \leq C(K)\left\{\inf_{m\in\mathcal{M}}\left(\|s - s_m\|_n^2 + pen(m)\right) + \frac{\Sigma\sigma^2}{n}\right\}$$

où $C(K)$ est une constante ne dépendant que de K.

Si notre famille de modèles $\{S_m\}_{m\in\mathcal{M}}$ est telle que chaque S_m a une dimension $D_m > 0$ et pour tout entier N compris entre 1 et n, il n'y a au plus qu'un seul S_m qui soit de dimension $D_m = N$, alors ce théorème permet de valider l'heuristique et d'obtenir une inégalité oracle. En effet, prenons $x_m = LD_m$ où $L > 0$ est une constante telle que

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \sum_{k\geq 0} e^{-Lk} \leq 1 \ .$$

Ainsi on a $\Sigma = 1$ et on peut alors considérer la pénalité

$$\text{pen}(m) = \frac{\sigma^2 D_m}{n} K\left(1 + \sqrt{2L}\right)^2 \ .$$

En choisissant $L > 0$ et $K > 1$ indépendamment de n, telles que

$$K\left(1 + \sqrt{2L}\right)^2 = 2$$

on retrouve la pénalité de Mallows. La borne du risque donnée par le théorème est donc

$$\begin{aligned}
\mathbb{E}\left[\|s - \tilde{s}\|_n^2\right] &\leq C(K)\left\{\inf_{m\in\mathcal{M}}\left(\|s - s_m\|_n^2 + \frac{2\sigma^2 D_m}{n}\right) + \frac{\sigma^2}{n}\right\} \\
&\leq C(K)\inf_{m\in\mathcal{M}}\left(\|s - s_m\|_n^2 + \frac{3\sigma^2 D_m}{n}\right) \\
&\leq 3C(K)\inf_{m\in\mathcal{M}}\mathbb{E}\left[\|s - \hat{s}_m\|_n^2\right] \ .
\end{aligned}$$

La deuxième inégalité étant valable car on a exclu la possibilité qu'un modèle de dimension nulle soit dans notre famille, ainsi le risque de chaque estimateur est d'au moins σ^2/n. On obtient bien la forme d'une inégalité oracle. Ce raisonnement met surtout en relief l'importance de la connaissance de la variance pour le choix de la pénalité et donc pour la construction de l'epp $\tilde{s}$.

2 Hétéroscédasticité

La question que l'on peut alors se poser est que faire si l'on est dans un **cadre hétéroscédastique** ? C'est-à-dire si la variance est inconnue et qu'on ne la suppose pas constante. Comment construire l'epp ? A-t-on les mêmes propriétés ? Peut-on faire aussi bien ? Et surtout, a-t-on une inégalité oracle ? Le cadre hétéroscédastique trouve de nombreuses applications ; en effet, en pratique, la variance des phénomènes observés est le plus souvent inconnue et varie au fil de l'expérience. Un premier objectif de la thèse va être de s'intéresser au cadre de régression vu précédemment en l'élargissant à ce contexte hétéroscédastique. Présentons d'abord un exemple illustrant cette problématique.

2.1 Estimation par un histogramme de l'intensité d'un processus de Poisson sur $[0,1]$

Les travaux de Patricia Reynaud-Bouret ((5) et (6)), que j'ai pu étudier lors de mon mémoire de master II, donnent des résultats sur l'epp construit dans les cadres hétéroscédastiques des processus ponctuels et, en particulier, dans celui poissonnien.

Rappelons qu'un **processus de Poisson** N sur $[0,1]$ est un sous-ensemble aléatoire discret de $[0,1]$ (que l'on muni de sa tribu borélienne $\mathcal{B}$) tel que

- pour tout $A \in \mathcal{B}$, le nombre, $|A \cap N|$, de points de N tombant dans A suit une loi de Poisson de paramètre noté $\nu(A)$,
- pour tout choix de $A_1, \dots, A_k \in \mathcal{B}$ disjoints, les variables $|A_1 \cap N|$, $\dots$, $|A_k \cap N|$ sont indépendantes,

où l'application $\nu : \mathcal{B} \to \mathbb{R}_+$ est une mesure, dite **mesure principale** de N, que l'on supposera absolument continue par rapport à la mesure de Lebesgue sur $[0,1]$. On note s_{int} la dérivée de Radon-Nikodym de ν par rapport à la mesure de Lebesgue et on l'appelle **intensité** de N. On supposera que $s_{int} \in \mathbb{L}^2 = \mathbb{L}^2([0,1], dx)$ et on notera

$$\forall f \in \mathbb{L}^2,\ \|f\|^2 = \int_0^1 f(t)^2 dt\ .$$

En particulier, on a $\nu([0,1]) = \int_0^1 s_{int}(t)dt < +\infty$, ce qui signifie que le nombre de points de N est presque surement fini. Enfin, on note

$$dN = \sum_{X \in N} \delta_X$$

la mesure discrète aléatoire sur $[0,1]$ associée à N.

Dans notre expérience, nous observons les points de n processus de Poisson sur $[0,1]$, $N^{(1)},\dots,N^{(n)}$, tous d'intensité s_{int}. Leur agrégation

$$N=\bigcup_{i=1}^{n} N^{(i)}$$

est un processus de Poisson sur $[0,1]$ d'intensité ns_{int}. Faire de la sélection de modèle pour estimer l'intensité s_{int} est possible mais cette situation ne correspond pas tout à fait au cadre de régression vu en première partie. Pour remédier à cela, fixons nous un entier $M>0$, nous allons estimer l'histogramme s qui est la projection orthogonale de s_{int} sur l'espace des histogrammes construits sur les intervalles $\left[\frac{i-1}{2^M},\ \frac{i}{2^M}\right[$ pour i allant de 1 à 2^M. Les fonctions

$$\varphi_{M,i}=2^{M/2}\mathbb{1}_{\left[\frac{i-1}{2^M},\ \frac{i}{2^M}\right[},\ i=1,\dots,2^M$$

forment une base orthonormée de cet espace dans $\mathbb{L}^2$. On a

$$s=2^{M/2}\sum_{i=1}^{2^M} s_i\varphi_{M,i}\ .$$

Pour simplifier les notations, on identifie la fonction s au vecteur $(s_1,\dots,s_{2^M})$ où

$$s_i=\int_{(i-1)/2^M}^{i/2^M} s_{int}(t)dt,\ i=1,\dots,2^M$$

et on cherche à estimer le vecteur s à partir des observations suivantes, pour i de 1 à 2^M,

$$\begin{aligned} Y_i &= \left|N\cap\left[\frac{i-1}{2^M},\ \frac{i}{2^M}\right[\right| \\ &= ns_i+\underbrace{\int_{(i-1)/2^M}^{i/2^M}(dN_t-ns_{int}(t)dt)}_{P_i}\ . \end{aligned}$$

Etant basées sur des intervalles disjoints, les P_i sont indépendantes, de plus elles sont centrées et de variance $\mathrm{Var}(P_i)=ns_i$. Rappelons que la variable de Poisson de paramètre 0 est presque surement égale à 0. Ainsi, on a l'écriture

$$Y_i=ns_i+\sqrt{ns_i}\varepsilon_i$$

où les ε_i sont des variables de Poisson indépendantes, centrées et de variance unitaire. Ecrit sous cette forme, l'hétéroscédasticité du problème est apparente. Les variances sont bien inconnues et non-égales.

On se donne maintenant la famille de modèles $\{S_m;\ m = 0, \dots, M\}$ où les S_m sont les espaces des histogrammes construits sur les intervalles $\left[\frac{i-1}{2^m},\ \frac{i}{2^m}\right[$ pour i allant de 1 à 2^m. Chaque S_m est de dimension $D_m = 2^m$, on y note s_m la projection orthogonale de s et on y considére la base orthonormée faite des fonctions

$$\varphi_{m,i} = 2^{m/2} \mathbb{1}_{\left[\frac{i-1}{2^m},\ \frac{i}{2^m}\right[},\ i = 1, \dots, 2^m\ .$$

Prenons un S_m, de même que dans la partie précédente, on peut définir l'estimateur par projection $\hat{s}_m$ comme le minimiseur sur S_m de

$$\gamma(f) = \|f\|^2 - \frac{2}{n}\langle f, Y\rangle\ .$$

Dans la base des $\varphi_{m,i}$ on a

$$\hat{s}_m = \frac{2^{m/2}}{n}\sum_{i=1}^{2^m} N_{m,i}\varphi_{m,i} \qquad \text{et} \qquad s_m = \frac{2^{m/2}}{n}\sum_{i=1}^{2^m} \alpha_{m,i}\varphi_{m,i}$$

avec

$$N_{m,i} = \left|N \cap \left[\frac{i-1}{2^m},\ \frac{i}{2^m}\right[\right| \qquad \text{et} \qquad \alpha_{m,i} = \mathbb{E}[N_{m,i}] = \int_{(i-1)/2^m}^{i/2^m} n s_{int}(t)dt\ .$$

Notons que les $N_{m,i}$ (respectivement les $\alpha_{m,i}$) sont des sommes de Y_j (respectivement de s_j). En particulier, les $N_{m,i}$ sont observables. On a la décomposition biais-variance du risque quadratique suivante

$$\begin{aligned}
\mathbb{E}\left[\|s - \hat{s}_m\|^2\right] &= \|s - s_m\|^2 + \frac{2^m}{n^2}\sum_{i=1}^{2^m}\mathbb{E}\left[(N_{m,i} - \alpha_{m,i})^2\right] \\
&= \|s - s_m\|^2 + \frac{2^m}{n^2}\sum_{i=1}^{2^m}\alpha_{m,i} \\
&= \|s - s_m\|^2 + \frac{D_m}{n^2}\underbrace{\int_0^1 n s_{int}(t)dt}_{\mathbb{E}[|N|]}\ .
\end{aligned}$$

Il nous faut donc trouver un équilibre entre ces termes parmi nos modèles et donc choisir une pénalité. On peut encore se laisser guider par une heuristique à la Mallows. En effet, on cherche à minimiser en $m \in \{0, \dots, M\}$ la quantité

$$-\|s_m\|^2 + \frac{D_m}{n^2}\mathbb{E}[|N|]\ .$$

En remarquant que $\mathbb{E}\left[\|\hat{s}_m\|^2\right] = \|s_m\|^2 + D_m\mathbb{E}[|N|]/n^2$, cette heuristique suggère de choisir $\hat{m}$ comme un minimiseur du critère

$$-\|\hat{s}_m\|^2 + \frac{2D_m}{n^2}|N| = \gamma(\hat{s}_m) + \frac{2D_m}{n^2}|N| \ .$$

On voit ainsi apparaître une pénalité à la Mallows $\text{pen}(m) = 2D_m|N|/n^2$ aléatoire. Ce caractère non-déterministe de la pénalité nous permet ici d'estimer le terme de variance puisqu'il est, pour une variable de Poisson, égal à son espérance.

Cette heuristique est valide et nous mène à une inégalité oracle, ceci via le résultat suivant dû à Patricia Reynaud-Bouret (voir (6)) :

Théorème 3 *On se place dans le cadre présenté ci-dessus. Supposons que* $\rho = \int_0^1 s_{int}(t)dt > 0$ *et que* $2^M \leq n$. *Si, pour une certaine constante* $d > 1$, *la pénalité est telle que*

$$pen(m) \geq d\frac{D_m|N|}{n^2}$$

alors l'epp $\tilde{s}$ *correspondant vérifie*

$$\mathbb{E}\left[\|s - \tilde{s}\|^2\right] \leq C \inf_{m=0,\dots,M}\left\{\|s - s_m\|^2 + \mathbb{E}[pen(m)]\right\} + \frac{C'}{n}$$

où C *est une constante ne dépendant que de* d *et* C' *est une constante ne dépendant que de* d, $\|s\|$, $\|s\|_\infty$ *et de* ρ.

En prenant $d = 2$, on a la pénalité à la Mallows

$$\text{pen}(m) = \frac{2D_m}{n^2}|N|$$

et le théorème donne la borne

$$\begin{aligned}
\mathbb{E}\left[\|s - \tilde{s}\|^2\right] &\leq C \inf_{m=0,\dots,M}\left\{\|s - s_m\|^2 + \frac{2D_m}{n^2}\mathbb{E}[|N|]\right\} + \frac{C'}{n} \\
&\leq C'' \inf_{m=0,\dots,M}\left\{\|s - s_m\|^2 + \frac{2D_m}{n^2}(\mathbb{E}[|N|] + n)\right\} \\
&\leq C''' \inf_{m=0,\dots,M}\left\{\|s - s_m\|^2 + \frac{4D_m}{n^2}\mathbb{E}[|N|]\right\} \\
&\leq 4C''' \inf_{m=0,\dots,M}\mathbb{E}\left[\|s - \hat{s}_m\|^2\right] \ .
\end{aligned}$$

On a bien la forme d'une inégalité oracle.

2.2 Cadre plus général

Une façon de généraliser ce cas est la formulation suivante : on observe, pour i allant de 1 à n,

$$Y_i = s_i + \sigma_i \varepsilon_i$$

où $s = (s_1, \ldots, s_n) \in \mathbb{R}^n$ est inconnu, $\sigma = (\sigma_1, \ldots, \sigma_n) \in (\mathbb{R}_+^*)^n$ est inconnu et $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ est un vecteur aléatoire dont les composantes sont indépendantes, centrées et de variance égale à 1. Puis on cherche à estimer s et σ simultanément via une méthode de sélection de modèle. C'est-à-dire que l'on se donne deux familles de modèles $\{S_m\}_{m \in \mathcal{M}}$ et $\{\Sigma_{m'}\}_{m' \in \mathcal{M}'}$ qui peuvent être différentes et dans chaque $S_m \times \Sigma_{m'}$ on estime (s, σ) par un $\widehat{(s,\sigma)}_{(m,m')}$. Se pose alors la question de la forme de la pénalité pen : $\mathcal{M} \times \mathcal{M}' \to \mathbb{R}_+$ à prendre pour que l'epp $\widetilde{(s,\sigma)} = \widehat{(s,\sigma)}_{\widehat{(m,m')}}$ vérifie une inégalité oracle ?

Ces questions peuvent mener à diverses ouvertures. Par exemple, il serait possible d'affaiblir encore les hypothèses sur le bruit en ne supposant plus les ε_i indépendantes. D'autre part, le cadre présenté ci-dessus peut être vu sous l'écriture suivante

$$Y_i = s(x_i) + \sigma(x_i)\varepsilon_i$$

où les x_i sont des points connus d'un espace mesurable $(A, \mathcal{A})$. En notant

$$\mu_n = \sum_{i=1}^{n} \delta_{x_i}$$

notre problème revient à estimer les fonctions s et σ dans $\mathbb{L}^2(A, \mu_n)$ par exemple. Comment alors étendre des résultats dans ce cadre à celui de la régression sur un support aléatoire ? C'est-à-dire si on observe les couples (X_i, Y_i) où les X_i sont des variables aléatoires à valeurs dans $(A, \mathcal{A})$ et

$$Y_i = s(X_i) + \sigma(X_i)\varepsilon_i \ .$$

On peut aussi dans cette direction s'intéresser à l'auto-regression : les Y_i dépendent de leurs états précédents,

$$Y_{i+1} = s(Y_i) + \sigma(Y_i)\varepsilon_{i+1} \ .$$

L'hétéroscédasticité induit ainsi un large champ d'investigation. De plus, les applications sont nombreuses. D'une part, comme on l'a précisé précédemment, dans le cadre d'expériences n'impliquant pas la connaissance de la variance et d'autre part, lorsque celle-ci varie au fur et à mesure des expériences.

Références

[1] H. Akaike (1973) : *Information theory and extension of the maximum likelihood principle*, 2nd International Symposium on Information Theory, Akademia Kiado, Budapest, 267–281.

[2] Y. Baraud (2000) : *Model selection for regression on a fixed design*, Probab. Theory Relat. Fields **117**, 467–493.

[3] L. Birgé and P. Massart (2001) : *Gaussian model selection*, J. Eur. Math. Soc. **3**, 203–268.

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

[5] P. Reynaud-Bouret (2002), *Estimation adaptative de l'intensité de certains processus ponctuels par sélection de modèle*, Thèse de Doctorat, Université Paris XI Orsay.

[6] ——— (2003) : *Adaptative estimation of the intensity of inhomogeneous poisson processes via concentration inequalities*, Probab. Theory Relat. Fields.