Sélection de modèles

Sylvain Arlot et Jean-Patrick Baudry

Exposé proposé par Yannick Baraud

21 juin 2002

Table des matières

Introduction

On a souvent besoin, en statistique, de faire des estimations, c'est-à-dire de "deviner" une quantité s dont on n'a observé qu'un nombre fini d'effets non déterministes, en général une réalisation d'une variable aléatoire y dont la loi dépend de s. Par exemple, s peut désigner la densité de cette loi par rapport à une mesure μ connue. Un des problèmes qui se pose est que l'on doit toujours faire des hypothèses sur Y et s, plus ou moins simplificatrices, si l'on veut aboutir à un résultat intéressant.

Prenons un exemple révélateur des enjeux du problème : on a observé les valeurs d'une fonction f en 1000 points de l'intervalle $[0\,;1]$, avec des erreurs de mesure (elles sont inévitables) et/ou des aléas d'expérience, et on voudrait estimer f. Il nous faut donc décider parmi quelle famille de fonctions nous cherchons une telle estimation. Ce choix est crucial. En effet, si nous voulions par exemple estimer f par une fonction mesurable quelconque, nous aurions de nombreuses candidates, dont la plupart seraient très mauvaises, pour l'estimation recherchée : ce seraient toutes les fonctions mesurables qui prennent les mêmes valeurs que celles mesurées pour f en chacun des 1000 points choisis... Il ne faut donc pas choisir une famille trop grande. Mais il faut évidemment qu'elle ne soit pas non plus trop peu fournie, pour que l'on puisse espérer y trouver une estimation correcte de f. L'idée qui vient naturellement à l'esprit est de se limiter à un espace vectoriel de dimension finie, par exemple les fonctions constantes sur chaque intervalle du type $[j/n;(j+1)/n[$ ou bien les combinaisons linéaires de $e^{ijx}, -n \leq j \leq n$. Pour chacun de ces modèles, il est facile de trouver une "bonne" estimation de f (cela revient à une projection orthogonale dans L^2). Ce qui est moins évident, c'est de choisir parmi tous ces modèles.

En effet, les erreurs de mesure n'auront pratiquement aucun impact pour le modèle où f est constante, mais une telle approximation a toutes les chances d'être assez mauvaise. A l'inverse, en cherchant f constante sur des intervalles d'amplitude $1/1000$, on est a priori beaucoup plus près de la réalité, mais les erreurs de mesure influencent considérablement le résultat. Il faut chercher un compromis entre ces deux extrêmes, c'est l'objet de la sélection de modèles. La figure 1 parle d'elle-même.

Examinons un autre exemple, la sélection de variables : on a observé une variable y (par exemple le taux de pollution à Paris) et N autres paramètres $x^1, \ldots, x^N$ (la vitesse du vent, son carré, la température, la densité du trafic automobile...) à n instants distincts (et suffisament éloignés). On suppose que y ne dépend que de ces N paramètres, et nous ne nous intéresserons ici qu'au cas d'une dépendance linéaire. Mais comme on a sans doute trop pris de variables explicatives (pour être certains de ne rien oublier), et les don-

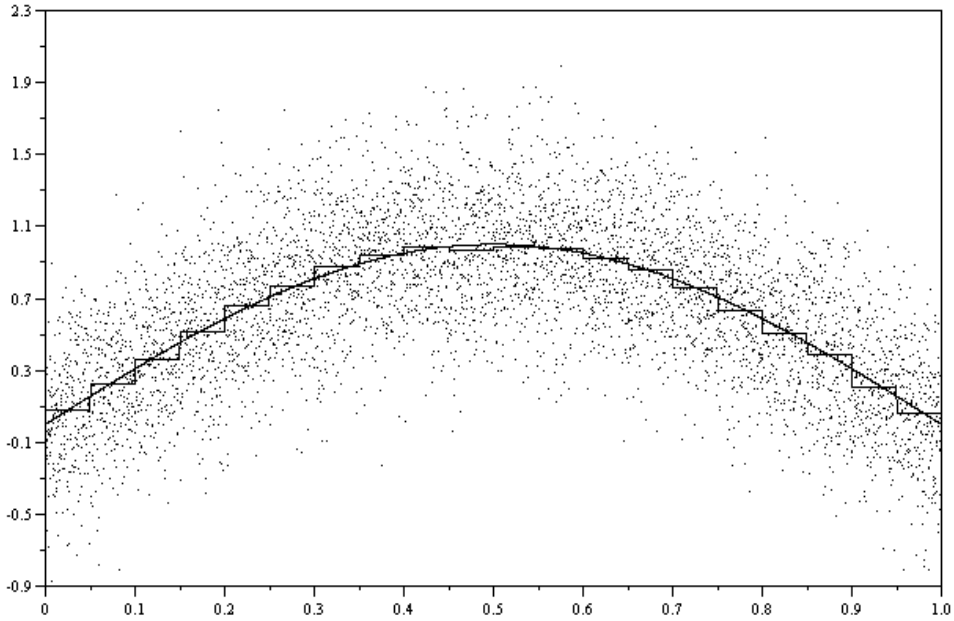

FIG. 1 – l'observation, la fonction cherchée et le "bon" modèle.

nées n'étant pas assez nombreuses, il serait désastreux de laisser y dépendre de tous les x^j. Il s'agit donc de sélectionner parmi les x^j quelles sont les variables dont l'influence est la plus significative. Et si l'on peut facilement concevoir une mesure de l'influence d'une variable sur le taux de pollution, il est beaucoup moins évident intuitivement de déterminer un seuil de sélection.

Revenons dans un cadre plus rigoureux pour constater que cette question de compromis au niveau de la dimension du modèle est au cœur du problème. Dans toute la suite, nous assimilerons les fonctions (par exemple μ) et le n-uplet correspondant, dont la $i^{ème}$ coordonnée correspond à la valeur de la fonction en question à l'instant i (pour μ, c'est $(\mu_1, \ldots, \mu_n)$).On pose aussi $Y = (y_1, \ldots, y_n)$ et $X^j = (x_1^j, \ldots, x_n^j)$. Dans le cas de la sélection de variables, pour $m \subset \{1, \ldots, N\}$, on note S_m l'espace engendré par les X^j, $j \in m$ et on suppose que l'on a $Y = \mu(X^1, \ldots, X^N) + \sigma\xi$ où ξ est un vecteur Gaussien standard. L'estimateur classique de μ est $\hat{\mu}_m$, la projection orthogonale de Y sur S_m. Pour évaluer ses performances, on regarde son risque quadratique $\mathbb{E}\Big[\|\hat{\mu}_m - \mu\|_n^2\Big]$ où $\|\cdot\|_n$ désigne la norme euclidienne usuelle divisée par un facteur $\sqrt{n}$. Celui-ci se calcule aisément :

$$\mathbb{E}\big[\|\hat{\mu}_m - \mu\|_n^2\big] = \frac{\sigma^2|m|}{n} + \|\mu_m - \mu\|_n^2,$$

μ_m désignant la projection orthogonale de μ sur S_m. Pour minimiser ce risque, il faut donc éviter de prendre $|m|$ grand (premier terme), tout en prenant un espace assez gros (deuxième terme). On retrouve ici le compromis qu'on avait repéré intuitivement et que la figure 1 permet de visualiser.

Nous allons donc nous placer dans un cadre formel assez général où nous donnerons une méthode de sélection de modèles fondée sur le principe de la pénalisation (pour favoriser les modèles de plus petite dimension). Nous verrons ensuite plus précisément ce que l'on obtient dans le cas de la sélection de variables. Enfin, à l'aide de simulations numériques, nous testerons plus précisément notre méthode, en particulier pour voir comment il faut en régler les paramètres afin d'optimiser ses performances.

1 Modèles Gaussiens

1.1 Processus gaussiens

Nous allons limiter notre étude au cas gaussien, ce qui présente l'avantage de simplifier les démonstrations tout en mettant en évidence les aspects les plus importants de la méthode décrite. Les résultats obtenus ont également l'avantage d'être plus précis.

Cependant, nous allons définir un cadre général pour notre étude, ce qui permet de traiter simultanément – entre autres – la sélection de variables et les problèmes de régression. Pour cela, il semble naturel d'introduire une généralisation en dimension infinie de la notion de vecteur gaussien standard (c'est le *processus linéaire isonormal*), puis de vecteur gaussien de matrice de covariance proportionnelle à l'identité (c'est le *processus linéaire Gaussien*).

1.1.1 Définition générale

Dans la suite, nous considérons toujours un espace de Hilbert $\mathbb{H}$ muni d'un produit scalaire $\langle \cdot, \cdot \rangle$ et de la norme $\| \cdot \|$ associée.

Définition 1 (Processus linéaire isonormal) *Soit $\mathbb{S}$ un sous-espace vectoriel de $\mathbb{H}$ et $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. Un* processus linéaire isonormal *Z indexé par $\mathbb{S}$ est un processus Gaussien centré, presque sûrement linéaire, de structure de covariance*

$$Cov(Z(t), Z(u)) = \langle t, u \rangle$$

La linéarité presque sûre signifie qu'il existe une partie Ω' de Ω telle que $\mathbb{P}(\Omega') = 1$ et

$$\forall \omega \in \Omega' \ \forall \alpha, \beta \in \mathbb{R} \ \forall t, u \in \mathbb{S}, \ on\ a\ \alpha Z(t)(\omega) + \beta Z(u)(\omega) = Z(\alpha t + \beta u)(\omega).$$

Définition 2 (Processus linéaire Gaussien) *Soit* $\mathbb{S}$ *un sous-espace vectoriel de* $\mathbb{H}$. *On appelle* processus linéaire Gaussien *sur* $\mathbb{S}$ *de moyenne* $s \in \mathbb{H}$ *et de variance* ε^2 *un processus* Y *indexé par* $\mathbb{S}$ *de la forme*

$$Y(t) = \langle s, t \rangle + \varepsilon Z(t) \quad \forall t \in \mathbb{S} \tag{1.1}$$

où Z *désigne un processus linéaire isonormal indexé par* $\mathbb{S}$.

Notre problème se résume alors ainsi : nous travaillons sur un espace abstrait Ω muni d'une tribu $\mathcal{A}$ et d'une famille $(\mathbb{P}_s)_{s\in\mathbb{H}}$ de probabilités sur cet ensemble. $\mathbb{E}_s$ représente l'espérance sous $\mathbb{P}_s$. Nous cherchons alors, disposant d'une réalisation de Y, une estimation du s correspondant à la loi $\mathbb{P}_s$ sous laquelle Y à été tirée (étant donné s, Y défini comme en (1.1) suit, sous $\mathbb{P}_s$, la loi d'un processus linéaire Gaussien de moyenne s et de variance ε^2). Nous appellerons alors estimateur toute fonction mesurable de Y à valeurs dans $\mathbb{H}$ (ou dans des espaces plus restreints, comme nous le verrons plus tard), et évaluerons l'erreur d'un estimateur $\tilde{s}$ par son risque quadratique $\mathbb{E}_s\big[\|s - \tilde{s}\|^2\big]$(ceci n'est possible que quand on connaît s et donc pour les calculs théoriques, mais pas dans la pratique).

1.1.2 Cas de la sélection de variables

Dans la pratique, on est souvent confronté au problème particulier de la sélection de variables : on dispose d'une réalisation de y et on l'observe en n instants distincts, ainsi que les valeurs correspondantes de $x^1, ..., x^N$ qui sont des variables dont on suppose que s dépend. On se propose de préciser cette dépendance, tout d'abord en sélectionnant les variables dont y dépend réellement, ou plutôt dont la dépendance est significative au vu des données. Cela conduit à la formulation abstraite suivante. La mesure à n instants distincts des quantités $x^1, \ldots, x^N$ et de la réalisation de y nous donne $N+1$ vecteurs de $\mathbb{R}^n$, $X^1, \ldots, X^N$ et Y (la valeur à l'instant i correspond à l'indice i), avec $N \leq n$ et les X^j sont supposés libres. On fait l'hypothèse que $y_i = \mu(x_i^1, \ldots, x_i^N) + \sigma\xi_i$ où μ est une fonction déterministe de $x^1, \ldots, x^N$ et les ξ_i sont des variables aléatoires i.i.d. $\mathcal{N}(0,1)$. Dans la suite, on se contentera d'une dépendance linéaire des x^j, i.e. $\mu(x_i^1, \ldots, x_i^N) = \sum_{j=1}^N \beta_j x_i^j$.

Posons $\mathbb{H} = \mathbb{R}^n$, $\mathbb{S} = ev\big\{X^j\big\}_{j\in\{1,\ldots,N\}}$. On munit $\mathbb{R}^n$ du produit scalaire $\langle t, u \rangle = n^{-1}\sum_{i=1}^n t_i u_i$ pour $t, u \in \mathbb{R}^n$ (l'intérêt du $1/n$ étant de ne pas faire intervenir le nombre n d'observations) et on pose $s = \sum_{j=1}^N \beta_j X^j$. On a alors, pour tout $t \in \mathbb{R}^n$

$$\langle t, Y \rangle = \langle s, t \rangle + \varepsilon Z(t) \quad avec \quad Z(t) = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n \xi_i t_i \quad et \quad \varepsilon = \frac{\sigma}{\sqrt{n}} \tag{1.2}$$

Avec $Y(t) = \langle t, Y \rangle$, on retrouve (1.1), et nous sommes donc dans le cadre général qui y est défini.

Nous nous intéresserons à la section 3 aux résultats précis que l'on obtient dans ce cas particulier. Nous expliquerons précisément comment effectuer cette sélection de variables pour avoir une stratégie optimale en un certain sens, et nous déterminerons les risques d'erreur correspondants.

1.1.3 Les problèmes de régression

Il s'agit d'une famille assez générale de problèmes de sélection de modèles que l'on est amené à résoudre. On observe une suite de n observations indépendantes $Y = (y_1, \ldots, y_n)$ de moyenne inconnue s, que l'on se propose de déterminer.

Par la suite, nous nous intéresserons à un cas particulier de régression, à savoir l'interpolation de fonctions par des fonctions constantes par intervalles. L'idée est la suivante : on a observé les valeurs d'une fonction f en n points $x_1, \ldots, x_n$, avec de petites erreurs $\sigma\xi_i$, et on se propose de déterminer une bonne approximation de f par une fonction constante sur tous les intervalles de la forme $[j/m; (j+1)/m[$, m étant à déterminer de façon à minimiser le risque d'erreur.

On peut formaliser ceci de la manière suivante, de façon à se ramener au cas d'un processus linéaire Gaussien : $\mathbb{H} = \mathbb{S} = \mathbb{R}^n$, muni du produit scalaire $\langle t, u \rangle = n^{-1} \sum_{i=1}^n t_i u_i$. On observe $Y = s + \sigma\xi$ où $s = (f(x_1), \ldots, f(x_n))$ et les ξ_i sont i.i.d. normales centrées réduites. Posons $Y(t) = \langle t, Y \rangle$, Y est un processus linéaire Gaussien sur $\mathbb{R}^n$ de moyenne s et de variance $\varepsilon^2 = \sigma^2/n$ avec $Z(t) = n^{-1/2} \sum_{i=1}^n \xi_i t_i$

1.2 Modèles, estimateurs par projection et oracles

Nous voulons estimer la moyenne s d'un processus linéaire Gaussien Y à partir d'une réalisation de Y. Cela revient à construire un estimateur $\tilde{s}$ de s (c'est-à-dire une fonction mesurable de Y et ε). On mesure sa qualité avec son risque quadratique $\mathbb{E}_s\big[\|\tilde{s} - s\|^2\big]$.

1.2.1 Modèles linéaires

Dans la suite, un modèle sera un sous-espace vectoriel de dimension finie de $\mathbb{H}$. Remarquons que l'on peut très bien avoir un modèle de dimension 0 (c'est $\{0\}$). Considérons une famille dénombrable $\{S_m\}_{m \in \mathcal{M}}$ de modèles. Cela ne change rien de considérer uniquement des sous-espaces vectoriels de $\mathbb{S}$. En effet, prenons pour $\mathbb{S}$ l'espace vectoriel engendré par la réunion des S_m.

Cette famille étant dénombrable, on peut construire une base orthonormale $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ de $\mathbb{S}$ par orthonormalisation. On pose alors, pour $t \in \mathbb{S}$, $Z(t) = \sum_{\lambda\in\Lambda} \langle t, \varphi_\lambda\rangle \xi_\lambda$ (un nombre fini de termes étant non-nuls) où $(\xi_\lambda)_{\lambda\in\Lambda}$ est une famille de variables aléatoires i.i.d. normales centrées réduites. On vérifie aisément que Z est un processus Gaussien isonormal indexé par $\mathbb{S}$.

On appelle estimateur à valeurs dans S_m une fonction $\tilde{s}_m$ mesurable de Y, qui prend ses valeurs dans S_m. Que peut-on espérer d'un tel estimateur, c'est-à-dire comment caractériser un estimateur optimal ? Cela n'a aucun sens de minimiser son risque en chaque point, dans la mesure où pour tout s, l'estimateur constant égal à s a un risque nul, tout en n'étant certainement pas optimal en général ! Une bonne façon de faire est de regarder pour chaque estimateur la plus grande erreur commise lorsque s varie dans une partie $\mathcal{T}$ bien choisie de $\mathbb{H}$, puis de chercher à minimiser cette quantité entre tous les estimateurs possibles. Cela conduit à la définition (classique) de risque minimax :

Définition 3 (Risque minimax) *Etant donné un processus Y défini par (1.1) et un sous-ensemble $\mathcal{T}$ de $\mathbb{H}$, le risque minimax sur $\mathcal{T}$ est donné par :*

$$R_M(\mathcal{T}, \varepsilon) = \inf_{\hat{s}} \sup_{s\in\mathcal{T}} \mathbb{E}_s\Big[\|\hat{s} - s\|^2\Big] \tag{1.3}$$

où l'infimum est pris sur tous les estimateurs possibles $\hat{s}$, i.e. les fonctions mesurables de Y.

Revenons au cas des estimateurs à valeurs dans S_m. Leur risque quadratique vaut :

$$\mathbb{E}_s\Big[\|\tilde{s}_m - s\|^2\Big] = \|s_m - s\|^2 + \mathbb{E}_s\Big[\|\tilde{s}_m - s_m\|^2\Big], \tag{1.4}$$

où s_m désigne la projection orthogonale de s sur S_m. Il s'agit donc d'optimiser la quantité $\mathbb{E}_s\Big[\|\tilde{s}_m - s_m\|^2\Big]$ parmi les estimateurs à valeurs dans S_m. Comme précédemment, il est préférable d'adopter un point de vue minimax de ce problème. On peut ainsi définir le risque minimax d'estimation de la projection s_m de s sur S_m : c'est $\inf_{\tilde{s}_m} \sup_{s\in S_m+t} \mathbb{E}_s\big[\|\tilde{s}_m - s_m\|^2\big]$, avec $t \in S_m^\perp$ (t représente la partie de s qu'on a choisi d'oublier en prenant S_m comme modèle. On le fixe pour séparer le risque lié au choix du modèle du risque d'erreur de l'estimateur sur la partie de s qu'on a choisi de garder avec S_m. Si on regardait un sup sur $\mathbb{S}$, en faisant tendre $\|t\|$ vers $+\infty$, celui-ci serait infini). Bien sûr, l'infimum est pris sur les estimateurs à valeurs dans S_m. Or, on a le résultat classique suivant (reposant sur l'indépendance des restrictions de Y à S_m et à $S_m^\perp$) :

$$\inf_{\tilde{s}_m} \sup_{s\in S_m+t} \mathbb{E}_s\Big[\|\tilde{s}_m - s_m\|^2\Big] \geq \varepsilon^2 D_m. \tag{1.5}$$

A l'aide de (1.4), on en déduit :

$$\inf_{\tilde{s}_m} \sup_{s \in S_m + t} \mathbb{E}_s\Big[\|\tilde{s}_m - s\|^2\Big] \geq \|t\|^2 + \varepsilon^2 D_m \tag{1.6}$$

1.2.2 Estimateurs par projection

Nous allons maintenant définir un estimateur qui est minimax sur le modèle S_m, puisque c'est un cas d'égalité de (1.6).

Définition 4 (Estimateur par projection) *Soit Y un processus linéaire Gaussien indexé par un sous-espace vectoriel $\mathbb{S}$ d'un espace de Hilbert $\mathbb{H}$, de moyenne $s \in \mathbb{H}$ et de variance ε^2. Soit S un sous-espace vectoriel de dimension finie de $\mathbb{S}$. Posons $\gamma(t) = \|t\|^2 - 2Y(t)$. L'*estimateur par projection sur *S est alors défini comme minimisant $\gamma(t)$ pour $t \in S$.*

Voici comment construire un tel estimateur, et pourquoi celui-ci est effectivement unique. Soit S_m de dimension D_m, $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$ une base orthonormale de S_m et écrivons $t = \sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda$. Minimiser $\gamma(t)$ revient alors à minimiser $\sum_{\lambda \in \Lambda_m} \beta_\lambda^2 - 2\beta_\lambda Y(\varphi_\lambda)$. Ainsi,

$$\hat{s}_m = \sum_{\lambda \in \Lambda_m} \hat{\beta}_\lambda \varphi_\lambda \quad avec \quad \hat{\beta}_\lambda = Y(\varphi_\lambda) = \langle s, \varphi_\lambda \rangle + \varepsilon Z(\varphi_\lambda) \tag{1.7}$$

est l'estimateur par projection sur S_m. Ceci prouve son unicité et en donne une construction explicite (que nous utiliserons par la suite). Remarquons que dans le cas où Y est un vecteur Gaussien, $\hat{s}_m$ est la projection orthogonale de Y sur S_m. Calculons le risque correspondant : en remarquant que $s_m = \sum_{\lambda \in \Lambda_m} \langle s, \varphi_\lambda \rangle \varphi_\lambda$, on a $\hat{s}_m - s_m = \varepsilon \sum_{\lambda \in \Lambda} Z(\varphi_\lambda)\varphi_\lambda$. Comme les $Z(\varphi_\lambda)$ sont i.i.d. de loi normale standard et $D_m = |\Lambda_m|$, on obtient :

$$\mathbb{E}_s\Big[\|\hat{s}_m - s\|^2\Big] = \|s_m - s\|^2 + \varepsilon^2 D_m. \tag{1.8}$$

Cela prouve que l'estimateur par projection $\hat{s}_m$ est minimax sur le modèle S_m et qu'il y a égalité dans (1.5).

Dans la suite, on s'intéressera uniquement à ce type d'estimateur, pour un modèle S_m donné. En effet, ceux-ci possèdent l'avantage d'être simples (la formule (1.7) facilite les calculs, aussi bien théoriques que pratiques), et ils minimisent le critère $\gamma(t) = \|t\|^2 - 2Y(t)$. C'est sur cette seconde propriété que repose toute la construction que nous allons décrire. Notre but sera de sélectionner parmi les estimateurs $\hat{s}_m$ le "meilleur" estimateur de s.

1.2.3 Sélection de modèles idéale et oracles

Idéalement, on voudrait, étant donnée une famille $\mathcal{F} = \{S_m\}_{m\in\mathcal{M}}$ de modèles, choisir $m = \bar{m}(s)$ qui réalise $\inf_{m\in\mathcal{M}}\Big\{\|s_m - s\|^2 + \varepsilon^2 D_m\Big\}$. Malheureusement, comme le laisse entendre la notation $\bar{m}(s)$, cela n'est possible que si l'on connaît déjà s, alors que l'on dispose seulement d'une réalisation de Y. Une telle procédure idéale de sélection est appelée *oracle*. Son intérêt est de nous donner un indicateur de la qualité statistique de la famille $\mathcal{F}$, que l'on définit ainsi :

Définition 5 (Précision d'oracle) *Soit $\mathcal{F} = \{S_m\}_{m\in\mathcal{M}}$ une famille de modèles linéaires de dimensions respectives D_m, dans un espace de Hilbert $\mathbb{H}$, une fonction $s \in \mathbb{H}$ et $\varepsilon > 0$. On note s_m la projection orthogonale de s sur S_m. La* précision d'oracle *de la famille $\mathcal{F}$ est définie par*

$$a_O(s,\mathcal{F},\varepsilon) = \inf_{m\in\mathcal{M}}\Big\{\|s_m - s\|^2 + \varepsilon^2 D_m\Big\} = \inf_{m\in\mathcal{M}}\mathbb{E}_s\Big[\|\hat{s}_m - s\|^2\Big] \tag{1.9}$$

Autrement dit, la précision d'oracle est l'erreur minimale que l'on commet en se limitant à choisir un modèle parmi la famille $\mathcal{F}$. Comme cette valeur ne peut pas raisonnablement être atteinte par la procédure de sélection de modèles que nous voulons construire, nous devrons nous contenter de chercher à obtenir un estimateur $\tilde{s}$ tel que $\mathbb{E}_s\big[\|\tilde{s} - s\|^2\big] \leq C a_O(s,\mathcal{F},\varepsilon)$.

Mais une telle borne ne peut pas, en général, être atteinte. En effet, si $\{0\} \in \mathcal{F}$ et $s = 0$, $a_O(0,\mathcal{F},\varepsilon) = 0$ et donc $\tilde{s} = 0$, $\mathbb{P}_0$ p.s. Les mesures $\mathbb{P}_s$ étant mutuellement absolument continues, cela entraîne que $\tilde{s} = 0$, $\mathbb{P}_s$ *p.s.* pour tout s, ce qui est évidemment un très mauvais estimateur... Plus généralement, on peut montrer que le même phénomène se produit si l'on exclut 0 mais que $a_O(s,\mathcal{F},\varepsilon)$ peut être arbitrairement petit. Nous devrons donc nous contenter – au mieux – d'inégalités du type

$$\forall\varepsilon > 0,\ \forall s \in \mathbb{H},\ \mathbb{E}_s\Big[\|\tilde{s} - s\|^2\Big] \leq C\Big[a_O(s,\mathcal{F},\varepsilon) + \varepsilon^2\Big] \tag{1.10}$$

Le terme en ε^2 permet d'éviter les problèmes que nous venons d'évoquer. Remarquons enfin que si l'on impose, pour un $\delta > 0$ fixé, $\|s\| \geq \delta\varepsilon$, on retrouve une inégalité du type précédent, puisque $a_O(s,\mathcal{F},\varepsilon) \geq \varepsilon^2(1\wedge\delta^2)$ et donc $a_O(s,\mathcal{F},\varepsilon) + \varepsilon^2 \leq \big[(1\wedge\delta^2)^{-1} + 1\big]a_O(s,\mathcal{F},\varepsilon)$.

2 Le Théorème principal

2.1 La méthode de sélection et ses performances

On va décrire une procédure de sélection de modèles visant à estimer l'inconnue s, à partir des estimateurs par projection associés à chacun des modèles. Comme on ne peut pas atteindre $\bar{m}(s)$, on va définir $\hat{m}(Y)$.

On veut minimiser, pour $m \in \mathcal{M}$, le risque quadratique de $\hat{s}_m$:

$$\mathbb{E}_s\Big[\|\hat{s}_m - s\|^2\Big] = \|s_m - s\|^2 + \varepsilon^2 D_m = \|s\|^2 - \|s_m\|^2 + \varepsilon^2 D_m$$

Or $\mathbb{E}_s\Big[\|\hat{s}_m\|^2\Big] = \|s_m\|^2 + \varepsilon^2 D_m$, donc $\|s\|^2 - \|\hat{s}_m\|^2 + 2\varepsilon^2 D_m$ est un estimateur sans biais du risque quadratique. Cela conduit à la méthode heuristique proposée par Mallows en 1964. L'idée de la pénalisation est de remplacer le terme $2\varepsilon^2 D_m$ par une pénalité $pen(m) \geq 0$, afin de choisir $\hat{m}$.

Définition 6 (Estimateur par projection pénalisé) *Soit $\{S_m\}_{m\in\mathcal{M}}$ une famille dénombrable de sous-espaces vectoriels de dimension finie de $\mathbb{S}$, $\hat{s}_m$ la famille correspondante d'estimateurs par projection construite à partir d'une même réalisation du processus Y, et une fonction positive* pen *définie sur $\mathcal{M}$. Un* estimateur par projection pénalisé *(associé à cette famille de modèles et à cette pénalité) est défini par $\tilde{s} = \hat{s}_{\hat{m}}$, où $\hat{m}$ (s'il existe) minimise pour $m \in \mathcal{M}$ le critère pénalisé*

$$crit(m) = -\|\hat{s}_m\|^2 + pen(m) \tag{2.1}$$

Théorème 1 *Soit Y un processus linéaire Gaussien indexé par un sous-espace vectoriel $\mathbb{S}$ d'un espace de Hilbert $\mathbb{H}$, de moyenne inconnue $s \in \mathbb{H}$ et de variance ε^2, $\{S_m\}_{m\in\mathcal{M}}$ une famille dénombrable de sous-espaces vectoriels de dimension finie de $\mathbb{S}$ de dimensions respectives D_m et $\{L_m\}_{m\in\mathcal{M}}$ une famille de poids, i.e. de réels positifs, satisfaisant la condition*

$$\Sigma = \sum_{\{m\in\mathcal{M}|D_m>0\}} \exp[-D_m L_m] < +\infty. \tag{2.2}$$

Soit une fonction de pénalité $pen(\cdot)$ définie sur $\mathcal{M}$ telle que

$$\exists K > 1, \quad \forall m \in \mathcal{M}, \quad pen(m) \geq K\varepsilon^2 D_m\Big(1 + \sqrt{2L_m}\Big)^2. \tag{2.3}$$

Alors, l'estimateur par projection pénalisé $\tilde{s}$ correspondant existe et est unique (presque sûrement). De plus, il vérifie :

$$\mathbb{E}_s\Big[\|\tilde{s}-s\|^2\Big] \leq \frac{4K(K+1)^2}{(K-1)^3}\left[\inf_{m\in\mathcal{M}}\Big\{d(s,S_m)^2+pen(m)\Big\}+(K+1)\varepsilon^2\Sigma\right], \tag{2.4}$$

où $d(s,S_m)$ désigne la distance de s à l'espace S_m.

2.2 Preuve du théorème

On a vu dans la Définition 4 que l'estimateur par projection $\hat{s}_m$ minimise $\gamma(t) = \|t\|^2 - 2[\langle s,t\rangle + \varepsilon Z(t)]$ sur S_m. Plus précisément, on déduit de (1.7) que pour tout $m' \in \mathcal{M}$,

$$\gamma(\hat{s}_{m'}) = \sum_{\lambda\in\Lambda_{m'}} -\hat{\beta}_\lambda^2 = -\|\hat{s}_{m'}\|^2. \tag{2.5}$$

Désormais, on fixe $m \in \mathcal{M}$ et on pose $\mathcal{M}' = \{m' \in \mathcal{M} \mid crit(m') \leq crit(m)\}$. Pour $m' \in \mathcal{M}'$, on a d'après (2.5) $\gamma(\hat{s}_{m'}) + pen(m') \leq \gamma(\hat{s}_m) + pen(m)$, d'où $\gamma(\hat{s}_{m'}) + pen(m') \leq \gamma(s_m) + pen(m)$ par définition de l'estimateur par projection, dans la mesure où s_m désigne la projection de s sur S_m et appartient donc à S_m. Remarquons de plus que $\forall t \in \mathbb{S},\ \gamma(t) + \|s\|^2 = \|s-t\|^2 - 2\varepsilon Z(t)$. On en tire l'inégalité suivante, pour tout $m \in \mathcal{M}'$:

$$\|s - \hat{s}_{m'}\|^2 \leq \|s - s_m\|^2 + 2\varepsilon[Z(\hat{s}_{m'}) - Z(s_m)] - pen(m') + pen(m). \tag{2.6}$$

Nous allons maintenant contrôler la quantité $Z(t) - Z(s_m)$, uniformément pour $t \in S_{m'}$ et $m' \in \mathcal{M}'$. Pour cela, rappelons tout d'abord une inégalité classique, dûe à Cirel'son, Ibragimov et Sudakov (1976), sous une forme particulière : si $X(t)$ est un processus Gaussien indexé par S, de variance identiquement égale à 1, pour tout $\lambda > 0$, on a :

$$\mathbb{P}\left[\sup_{t\in T} X(t) \geq \mathbb{E}\left[\sup_{t\in T} X(t)\right] + \lambda\right] \leq \exp\left[-\frac{\lambda^2}{2}\right]. \tag{2.7}$$

Prenons ici $X(t) = (Z(t) - Z(s_m))/\|t - s_m\|$ pour $t \in T = S_{m'}$ et $\lambda^2 = \lambda_{m'}^2 = 2(L_{m'}D_{m'} + \xi)$, $\xi > 0$. Les conditions de l'inégalité sont bien sûr vérifiées ici. On va commencer par majorer $\mathbb{E}[\sup_{t\in T} X(t)]$, en introduisant $S = S_m + S_{m'}$, de dimension D, et $\psi_1, \ldots, \psi_n$ une base orthonormale de S. Z étant linéaire sur $S \subset \mathbb{S}$ ($1^{ère}$ inégalité), et d'après l'inégalité de Cauchy-Schwartz ($2^{ème}$ inégalité), on obtient :

$$\sup_{t\in S_{m'}} \frac{Z(t) - Z(s_m)}{\|t - s_m\|} \leq \sup_{u\in S} \frac{Z(u)}{\|u\|} = \sup_{\alpha\in\mathbb{R}^D} \frac{\sum_{j=1}^D \alpha_j Z(\psi_j)}{\sqrt{\sum_{j=1}^D \alpha_j^2}} \leq \left[\sum_{j=1}^D Z^2(\psi_j)\right]^{\frac{1}{2}}.$$

Comme $D \leq D_m + D_{m'}$, on peut majorer l'espérance du sup considéré par $\sqrt{D_m + D_{m'}}$, ce qui donne, en utilisant (2.7) avec les valeurs indiquées :

$$\mathbb{P}_s\left[\sup_{t\in S_{m'}} \frac{Z(t) - Z(s_m)}{\|t - s_m\|} \geq \sqrt{D_m + D_{m'}} + \lambda_{m'}\right] \leq \exp(-L_{m'}D_{m'} - \xi)$$

En additionnant les inégalités obtenues quand m' décrit $\mathcal{M}$ et d'après la définition de Σ (2.2), en dehors d'un ensemble Ω_ξ de probabilité inférieure à $\Sigma \exp(-\xi)$, on a :

$$\begin{aligned} Z(t) - Z(s_m) &\leq \|t - s_m\| \Big[\sqrt{D_m + D_{m'}} + \sqrt{2(L_{m'} D_{m'} + \xi)} \Big] \\ &\leq \|t - s_m\| \Big[\sqrt{D_{m'}} \Big(1 + \sqrt{2 L_{m'}} \Big) + \sqrt{D_m} + \sqrt{2\xi} \Big], \end{aligned}$$

pour tout $t \in \bigcup_{m' \in \mathcal{M}} S_{m'}$ (on a utilisé que $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$). Fixons $\eta \in (0,1)$ et utilisons : (a) pour $c > 0$, $2ab \leq a^2 c + b^2 c^{-1}$, (a') $(a+b)^2 \leq a^2(1+c) + b^2(1+c^{-1})$, (a") $(a+b)^2 \leq 2(a^2+b^2)$, (b) la définition de $pen(\cdot)$, et (c) l'inégalité triangulaire. Cela donne :

$$\begin{aligned} 2\varepsilon [Z(t) - Z(s_m)] &\leq 2\varepsilon \|t - s_m\| \Big[\sqrt{D_{m'}} \Big(1 + \sqrt{2 L_{m'}} \Big) + \sqrt{D_{m'}} + \sqrt{2\xi} \Big] \\ &\leq \frac{\varepsilon^2}{1-\eta} \Big[\sqrt{D_{m'}} \Big(1 + \sqrt{2 L_{m'}} \Big) + \sqrt{D_{m'}} + \sqrt{2\xi} \Big]^2 + \|t - s_m\|^2 (1 - \eta) \qquad \text{(a)} \\ &= Q + R \end{aligned}$$

$$\begin{aligned} \textit{avec} \quad Q &\leq \frac{\varepsilon^2}{1-\eta} \Big[(1+\eta) D_{m'} \Big(1 + \sqrt{2 L_{m'}} \Big)^2 + (1 + \eta^{-1}) \Big(\sqrt{D_m} + \sqrt{2\xi} \Big)^2 \Big] \qquad \text{(a')} \\ &\leq \frac{1+\eta}{(1-\eta)K} pen(m') + \frac{2\varepsilon^2 (1+\eta^{-1})}{1-\eta} (D_m + 2\xi) \qquad \text{(b),(a")} \\ \textit{et} \quad R &\leq \Big[\|t - s\| + \|s - s_m\| \Big]^2 (1 - \eta) \qquad \text{(c)} \\ &\leq (1-\eta) \Big[\|t - s\|^2 (1+\eta) + \|s_m - s\|^2 (1 + \eta^{-1}) \Big] \qquad \text{(a')} \\ &= (1 - \eta^2) \|t - s\|^2 + (\eta^{-1} + \eta) \|s - s_m\|^2. \end{aligned}$$

Revenons au problème initial en combinant les deux inégalités que l'on vient d'obtenir avec (2.6). En dehors de Ω_ξ, pour tout $m' \in \mathcal{M}'$,

$$\begin{aligned} &\eta^2 \|s - \hat{s}_{m'}\|^2 + \frac{K - 1 - \eta(1+K)}{(1-\eta)K} pen(m') \\ &\qquad \leq (1 + \eta^{-1} - \eta) \|s - s_m\|^2 + \frac{2\varepsilon^2 (1+\eta^{-1})}{1-\eta} (D_m + 2\xi) + pen(m). \end{aligned} \tag{2.8}$$

Choisissons maintenant η assez petit pour avoir $K > 1 + \eta(1+K)$ (c'est possible avec $\eta \in (0,1)$), on a alors – sur Ω_ξ^c – $\sup \{pen(m'), m' \in \mathcal{M}'\} < \infty$

et donc il existe y tel que $\mathcal{M}' \subset \mathcal{M}(y) = \{m' \in \mathcal{M} | pen(m') \leq y\}$. Or pour tout $m' \in \mathcal{M}'$, $2K\varepsilon^2 L_{m'} D_{m'} \leq y$, d'où

$$\Sigma \geq \sum_{m' \in \mathcal{M}(y)} \exp\left[-L_{m'} D_{m'}\right] \geq |\mathcal{M}(y)| \exp\left[-\frac{y}{2K\varepsilon^2}\right]$$

Par conséquent, $\mathcal{M}(y)$ est fini, ainsi que $\mathcal{M}'$. Il existe ainsi $\hat{m}$ qui minimise $crit(m)$ sur $\mathcal{M}'$, donc sur $\mathcal{M}$. Examinons le problème de l'unicité. Si $S_{m'} = S_{m''}$ et $pen(m') = pen(m'')$, alors $crit(m') = crit(m'')$ mais $\hat{s}_{m'} = \hat{s}_{m''}$. Par contre, si $pen(m') \neq pen(m'')$, alors $crit(m) \neq crit(m')$. Si $S_{m'} \neq S_{m''}$, alors $\varepsilon^{-2}\left[\|\hat{s}_{m'}\|^2 - \|\hat{s}_{m''}\|^2\right]$ est la différence de deux variables indépendantes suivant des lois du χ^2 (comme (1.7) le montre aisément). L'événement $\left\{crit(m') = crit(m'')\right\} = \left\{\varepsilon^{-2}\left[\|\hat{s}_{m'}\|^2 - \|\hat{s}_{m''}\|^2\right] = (pen(m') - pen(m''))/\varepsilon^2\right\}$ est donc de probabilité nulle. Finalement, on a prouvé l'existence et l'unicité sur Ω_ξ^c de l'estimateur par projection pénalisé $\tilde{s} = \hat{s}_{\hat{m}}$. Comme $\mathbb{P}_s[\Omega_\xi]$ est arbitrairement petite, on a l'existence et l'unicité presque sûrement.

Pour obtenir une majoration du risque de cet estimateur, reprenons (2.8) avec $\eta = (K-1)/(K+1)$ et $m' = \hat{m}$: sur l'ensemble Ω_ξ^c, on a

$$\left(\frac{K-1}{K+1}\right)^2 \|s - \tilde{s}\|^2 \leq \frac{K^2+4K-1}{K^2-1}\|s-s_m\|^2 + pen(m) + \frac{2K(K+1)\varepsilon^2}{K-1}(D_m + 2\xi),$$

et comme $K\varepsilon^2 D_m \leq pen(m)$, en-dehors de Ω_ξ,

$$\|s-\tilde{s}\|^2 \leq \frac{(K+1)^2}{(K-1)^3}\left[\frac{K^2+4K-1}{K+1}\|s-s_m\|^2 + (3K+1)pen(m) + 4K(K+1)\varepsilon^2\xi\right].$$

Il existe donc une variable aléatoire V positive telle que $\mathbb{P}_s[V > \xi] \leq \Sigma \exp(-\xi)$ pour $\xi > 0$ telle que

$$\|s - \tilde{s}\|^2 \leq \frac{4K(K+1)^2}{(K-1)^3}\left[\|s - s_m\|^2 + pen(m) + (K+1)\varepsilon^2 V\right],$$

et comme $\mathbb{E}_s[V] \leq \Sigma$, on obtient (2.4) par intégration de cette inégalité puisque $m \in \mathcal{M}$ est arbitraire.

□

2.3 Comment utiliser ce résultat

L'inégalité (2.3) suggère de prendre

$$pen(m) = K\varepsilon^2 D_m (1 + \sqrt{2L_m})^2, \tag{2.9}$$

d'où l'intérêt de la définition suivante :

Définition 7 (Stratégie, Index de précision) *Soit $\mathbb{S}$ un sous-espace vectoriel d'un espace de Hilbert $\mathbb{H}$. On appelle* stratégie *une famille finie ou dénombrable $\{(S_m, L_m)\}_{m\in\mathcal{M}}$ où pour chaque $m \in \mathcal{M}$, S_m est un espace vectoriel de dimension D_m de $\mathbb{S}$ et L_m un réel positif vérifiant la condition (2.2). Etant donnée une stratégie $\mathcal{S}$, on définit son* index de précision *au point $s \in \mathbb{S}$ par*

$$a_I(s, \mathcal{S}, \varepsilon) = \inf_{m\in\mathcal{M}} \Big\{ d(s, S_m)^2 + \varepsilon^2 D_m(L_m + 1) \Big\} + \Sigma\varepsilon^2. \tag{2.10}$$

On peut désormais réécrire (2.4) sous la forme

$$\mathbb{E}_s\Big[\|\tilde{s} - s\|^2\Big] \leq C_0(K) a_I(s, \mathcal{S}, \varepsilon), \tag{2.11}$$

C_0 étant une fonction de K convenablement choisie (comme $(1 + \sqrt{2L_m})^2 \leq 3(1 + L_m)$, $C_0(K) = 12K(K + 1)^3(K - 1)^{-3}$ convient). L'index de précision nous suffit donc pour prouver l'optimalité d'une stratégie.

Nous pouvons d'ores-et-déjà nous poser deux questions : comment choisir K, comment choisir les poids L_m ?

2.3.1 Le rôle de K

On peut prouver (et on l'observe en réalisant des simulations) que la restriction $K > 1$ est nécessaire.

De plus, le terme $4K(K + 1)^3(K - 1)^{-3}$ suggère d'éviter de choisir K proche de 1 ou K trop grand. En revanche, si l'on cherche à minimiser cette fonction pour optimiser le risque de notre estimateur, le résulat obtenu n'est pas très satisfaisant. Une telle optimisation demande en effet plus de travail. Dans la suite, nous ne nous préoccuperons plus du choix de cette constante.

2.3.2 L'importance des poids L_m

Le choix des L_m est plus délicat, dans la mesure où il n'y a pas de solution à ce problème qui soit clairement optimale. Deux types de stratégies semblent "naturelles" : $L_m = L > 0$ ou $L_m = L(D_m)$, ce qui nécessite tout d'abord que pour tout D, il n'y ait qu'un nombre fini de modèles de dimension D.

3 Application à la sélection de variables

Nous supposons donc, selon l'expérience décrite en (1.1.2), disposer de l'observation de $x^1, \ldots, x^N$ et d'une réalisation de y à n instants distincts (ou en tout cas sous des conditions distinctes...). Nous notons désormais $Y = (y_1, \ldots, y_n)$ (y_i est la valeur de y observée à l'instant i, et nous utilisons

la même notation pour les x^j : $X^j = (x_1^j, \ldots, x_n^j)$) . Comme nous l'avions expliqué en (1.1.2), nous représentons notre expérience par $Y = s + \sigma\xi$, où s est le vecteur dont la $i^{ème}$ coordonnée est $\mu(x_i^1, \ldots, x_i^N)$.

Rappelons que le problème que nous nous posons, est, ne connaissant pas la fonction μ, donc s, d'estimer s. En effet, cela nous permettra alors d'accéder à la fonction μ, ou en tout cas à une estimation de celle-ci, et ainsi par exemple, de prévoir la valeur de μ sous de nouvelles conditions.

Nous avons déja vu que la difficulté n'est pas, étant donné un modèle, de trouver un estimateur optimal (ce que nous savons faire : il suffit de prendre l'estimateur par projection, d'après (1.8)), mais de choisir le modèle qui nous permettra d'obtenir le meilleur estimateur possible, avec les données dont nous disposons. Cela va consister, ici, à choisir quelles variables parmi $(X^1, \ldots, X^N)$ il est pertinent de garder pour calculer la fonction recherchée. Ainsi, nous cherchons un résultat sous la forme d'un sous-ensemble de $\Lambda = \{1, \ldots, N\}$.

Alors, si nous choisissons par exemple de garder exactement N observations parmi les n ($N \leq n$) dont nous disposons, il nous suffira, puisque nous avons choisi d'estimer μ par une fonction linéaire et que nous connaissons les vecteurs $(X^j)_{1 \leq j \leq N}$, de résoudre un système linéaire pour obtenir l'estimation de la fonction μ recherchée, à partir de l'estimation de s que nous aurons obtenue. Ceci est bien possible car nous avons supposé les vecteurs X^j linéairement indépendants, et la fonction que nous cherchons est donc bien unique.

Nous allons nous intéresser à deux stratégies possibles très différentes :

- La sélection de variables ordonnées, qui consiste à inclure dans la stratégie tous les ensembles de la forme $\Lambda_m = \{1, \ldots, m\}$, et alors $\mathcal{M} = \{1, \ldots, N\}$.
- La sélection complète de variables, qui consiste à s'intéresser à tous les sous-ensembles $\Lambda_m = m$ possibles de Λ ($m \in \mathcal{M} = \mathfrak{P}(\{1, \ldots, N\})$).

Nous voulons évaluer les résultats obtenus par l'estimateur pénalisé, défini à la Définition 6, lorsque la pénalité est donnée par (2.9). D'après (2.11), il suffit donc d'étudier l'index de précision.

3.1 Sélection de variables ordonnées

Nous allons d'abord nous intéresser à la stratégie la plus simple dans ce cadre, c'est-à-dire le cas des poids constants, puis nous verrons ce qu'apporte une stratégie avec des poids variables. Nous noterons S_m le sous-espace engendré par la famille $\{X^j\}_{j \in m}$

3.1.1 Stratégie à poids constants.

On choisit $L_m = L$ pour tout m et alors, (2.2) est vérifiée : $\Sigma = \sum_{m=1}^{N} \exp[-mL] < (e^L - 1)^{-1}$. Alors, l'index de précision de cette stratégie $\mathcal{S}$ est borné par

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m \in \mathcal{M}} \Big\{ d^2(s, S_m) + \varepsilon^2 m(1 + L) \Big\} + \varepsilon^2 \Big(e^L - 1 \Big)^{-1}. \tag{3.1}$$

Or $a_O(s, \mathcal{F}, \varepsilon) \geq \varepsilon^2$, donc

$$\begin{aligned} a_I(s, \mathcal{S}, \varepsilon) &\leq \inf_{m \in \mathcal{M}} \Big\{ d^2(s, S_m) + \varepsilon^2 m + L a_O(s, \mathcal{F}, \varepsilon) \Big\} + a_O(s, \mathcal{F}, \varepsilon) \Big(e^L - 1 \Big)^{-1} \\ &\leq C(L) a_O(s, \mathcal{S}, \varepsilon) \end{aligned}$$

ce qui est déjà une inégalité d'oracle, du type que l'on recherche (cf (1.10), et nous n'avons même pas besoin du terme en ε^2) : nous avons donc déjà résolu le problème que nous nous étions posé, ici dans le cas de la sélection de variables ordonnées. En effet, dans le sens que nous avions défini à la Définition 5, nous disposons d'une stratégie qui permet, avec l'estimateur par projection pénalisé, d'atteindre la meilleure précision que l'on peut espérer avec pour famille de modèles ceux de la sélection de variables ordonnées.

3.1.2 Stratégie à poids variables.

On peut par exemple essayer $L_m = \theta^2 m^{-1/2}$ pour un certain $\theta > 0$, et alors on a bien (2.2) : $\Sigma < \Sigma_\theta = \sum_{m=1}^{\infty} \exp(-\sqrt{m}\theta^2)$, et l'index de précision de la stratégie ainsi obtenue est borné par :

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m \in M} \Big\{ d^2(s, S_m) + \varepsilon^2 m \Big(1 + \theta^2 m^{-1/2} \Big) \Big\} + \Sigma_\theta \varepsilon^2. \tag{3.2}$$

Or, pour $L = \log\left(1 + \exp(\theta^2)/2\right)$ et $\theta^2 \geq 3$, on a $\Sigma_\theta < (e^L - 1)^{-1}$, donc on peut comparer facilement (3.1) et (3.2) : si on note m_0 un élément de $\mathcal{M}$ minimisant $d^2(s, S_m) + \varepsilon^2 m(1 + \theta^2 m^{-1/2})$,(3.2) est meilleur que (3.1) dès que $L > \theta^2 m_0^{-1/2}$.

On a donc montré que dans le cas de la sélection de variables ordonnées, on ne peut pas espérer faire mieux que le résultat obtenu par l'estimateur par projection pénalisé (puisqu'il atteint la précision d'oracle). Retenons seulement pour la suite que les résultats sont semblables, parfois un peu meilleurs avec la stratégie à poids variables, bien que cela n'apparaisse pas quand on cherche à obtenir une inégalité du type (1.10).

3.2 Sélection complète de variables

Dans le cas de la sélection complète de variables, nous choisissons d'inclure dans notre stratégie tous les sous-ensembles de $\Lambda = \{1, \ldots, N\}$. On note $m \in \mathcal{M} = \mathfrak{P}(\Lambda)$ un tel sous-ensemble, et alors, $D_m = |m|$ et S_m est le sous-espace engendré par la famille $\{X^j\}_{j \in m}$. En fait, nous supposerons parfois que les X^j forment une famille orthonormale, et nous la noterons alors $\{\varphi_\lambda\}_{\lambda \in \Lambda}$. Remarquons que cette hypothèse n'est pas raisonnable dans la réalité, car il n'y a aucune raison pour que les X^j soient orthonormés... Nous la faisons quand même pour deux raisons :

- Elle va nous permettre de faire certains calculs explicites, qui nous seront bien utiles d'un point de vue théorique, comme nous le verrons par la suite.
- Il existe des situations où cette hypothèse est valable, et les résultats qui suivent s'y étendent donc. Il faut les chercher en-dehors de la sélection de variables, par exemple lorsqu'on veut estimer une fonction et que l'on choisit pour X^j une base de Fourier...

3.2.1 Stratégie à poids constants.

Le cas $L_m = L$ pour tout $m \in \mathcal{M}$ vérifie (2.2) :

$$\Sigma = \sum_{k=1}^{N} C_N^k e^{-kL} = (1 + e^{-L})^N - 1.$$

Puisque nous voulons démontrer une inégalité du type de (1.10), il faut que Σ reste borné lorsque N tend vers l'infini si nous voulons pouvoir espérer le faire disparaître dans la constante, et pouvoir choisir celle-ci indépendamment de N. Et donc, il faut choisir $L/\log N \geq 1 + o(1)$. Nous choisissons $L = \log N$. Ainsi, on a $\Sigma < e - 1$ (et on a aussi $\Sigma < 1 + \log N$). En utilisant la pénalité donnée par (2.9), on peut majorer l'index de précision de cette stratégie par :

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m \in \mathcal{M}} \Big\{ d^2(s, S_m) + \varepsilon^2 |m| (1 + \log N) \Big\} + \Sigma \varepsilon^2 \tag{3.3}$$

$$\leq (1 + \log N) \Big[a_O(s, \mathcal{F}, \varepsilon) + \varepsilon^2 \Big]. \tag{3.4}$$

Il nous manque donc un facteur $(1+\log N)$ par rapport à la précision d'oracle, et nous voyons qu'avoir cherché à garder Σ borné nous a menés à une majoration dans laquelle le facteur tend en fait vers l'infini avec N.

Nous pouvons être plus précis et calculer explicitement la précision d'oracle et l'estimateur par projection pénalisé, lorsque la famille $\{\varphi_\lambda\}$ est supposée orthonormale. Notons $s = \sum_{\lambda\in\Lambda} \beta_\lambda\varphi_\lambda$. On a :

$$a_O(s,\mathcal{F},\varepsilon) = \inf_{m\in\mathcal{M}}\Big\{\sum_{\lambda\notin m}\beta_\lambda^2 + \varepsilon^2|m|\Big\} = \|s\|^2 + \inf_{m\in\mathcal{M}}\sum_{\lambda\in m}(-\beta_\lambda^2+\varepsilon^2). \quad (3.5)$$

Le minimum est atteint pour $m^\star = \{\lambda\in\Lambda\big|\beta_\lambda^2 > \varepsilon^2\}$, et on obtient donc la formule suivante :

$$a_O(s,\mathcal{F},\varepsilon) = \sum_{\lambda\in\Lambda}(\beta_\lambda^2\wedge\varepsilon^2). \quad (3.6)$$

En reportant dans (3.4), on obtient :

$$a_I(s,\mathcal{S},\varepsilon) \le (1+\log N)[\sum_{\lambda\in\Lambda}(\beta_\lambda^2\wedge\varepsilon^2)+\varepsilon^2]. \quad (3.7)$$

Mais, en cherchant à mettre notre inégalité sous la forme de (1.10), nous avons en fait perdu beaucoup de précision : en utilisant la majoration $\Sigma < e-1$, (3.3), et en utilisant une méthode similaire à celle qui a abouti à (3.6), nous obtenons en fait :

$$a_I(s,\mathcal{S},\varepsilon) \le \sum_{\lambda\in\Lambda}\Big(\beta_\lambda^2\wedge\varepsilon^2[1+\log N]\Big)+\varepsilon^2(e-1). \quad (3.8)$$

qui est manifestement meilleure, en particulier lorsque N est grand, comme on peut le voir par exemple avec le cas $s = \varepsilon\sum_{\lambda\in m}\varphi_\lambda$, $m\in\mathcal{M}\backslash\big\{\{0\}\big\}$.

Nous allons maintenant calculer l'estimateur. Nous disposons de la formule $\hat{s}_m = \sum_{\lambda\in m}\hat{\beta}_\lambda\varphi_\lambda$ où $\hat{\beta}_\lambda$ est le coefficient $Y(\varphi_\lambda)$, qui nous donne l'estimateur par projection associé au modèle S_m. Il reste à choisir le modèle correspondant à l'estimateur pénalisé. Pour cela, il nous faut minimiser le critère pénalisé $-\|\hat{s}_m\|^2 + pen(m)$, où la pénalité est de la forme $pen(m) = T^2|m|$, donc on peut appliquer une méthode analogue à celle employée pour trouver (3.6). Cela nous amène à minimiser $\sum_{\lambda\in\Lambda_m}[-\hat{\beta}_\lambda^2+T^2]$, et donc à choisir $\hat{m} = \{\lambda\in\Lambda\big||\hat{\beta}_\lambda| > T\}$. Nous parlerons donc de sélection seuillée (le seuil étant ici T) et l'estimateur seuillé est :

$$\tilde{s}_T = \hat{s}_{\hat{m}} = \sum_{\lambda\in\Lambda}\hat{\beta}_\lambda\mathbb{1}_{\{|\hat{\beta}_\lambda|>T\}}\varphi_\lambda \quad (3.9)$$

Remarquons que pour calculer $\tilde{s}_T$, il suffit de N opérations (qui consistent à comparer un coefficient $\hat{\beta}_\lambda$ au seuil), et non, comme on aurait pu le craindre, de comparer directement les valeurs de 2^N critères de pénalisation, ce qui

aurait déjà été impossible pour $N = 100$, par exemple, ce qui est une valeur raisonnable. Finalement, nous avons pu calculer explicitement la précision d'oracle et l'estimateur par projection pénalisé, par la méthode du seuil, mais nous n'avons pas réussi à obtenir une inégalité d'oracle avec une constante indépendante de N. En fait, ceci est impossible, d'après un résultat démontré par Donoho et Johnstone (1994, Théorème 3), selon lequel :

$$\liminf_{N\longrightarrow+\infty} \frac{1}{\log N}\left[\inf_{\hat{s}} \sup_{s\in\mathbb{S}} \frac{\mathbb{E}_s[\|s-\hat{s}\|^2]}{a_O(s,\mathcal{F},\varepsilon)+\varepsilon^2}\right] \geq 2, \tag{3.10}$$

où $\hat{s}$ parcourt l'ensemble des estimateurs possibles. Rappelons que $\mathbb{S} = ev\{X^j\}_{j\in\{1,\ldots,N\}}$

Voyons un peu de quelle façon nous comparons ici les estimateurs entre eux. Définissons, sur l'ensemble des estimateurs, une relation d'ordre de la façon suivante ($\hat{s}_1$ et $\hat{s}_2$ sont deux estimateurs) :

$$\hat{s}_1 \succeq \hat{s}_2 \longleftrightarrow \sup_{s\in\mathbb{S}} \frac{\mathbb{E}_s[\|s-\hat{s}_1\|^2]}{a_O(s,\mathcal{F},\varepsilon)+\varepsilon^2} \leq \sup_{s\in\mathbb{S}} \frac{\mathbb{E}_s[\|s-\hat{s}_2\|^2]}{a_O(s,\mathcal{F},\varepsilon)+\varepsilon^2} \tag{3.11}$$

Alors, nous choisissons comme "meilleur estimateur" $\hat{s}_N$ un élément maximal pour cette relation d'ordre, et la formule ci-dessus évalue alors, asymptotiquement (c'est-à-dire lorsque $N \to \infty$), les performances de ce meilleur estimateur. Et on voit donc que le cas de la sélection complète est sensiblement plus difficile que celui de la sélection ordonnée puisqu'ici, il est impossible d'atteindre la précision d'oracle, en tout cas pas à une constante indépendante de N près. De plus, nous avons donc atteint la meilleure précision que nous pouvions espérer, d'un point de vue asymptotique, à une constante près, puisque d'après (3.4) et (2.11), en notant $\tilde{s}_T$ l'estimateur seuillé :

$$\limsup_{N\longrightarrow+\infty} \frac{1}{\log N}\left[\sup_{s\in\mathbb{S}} \frac{\mathbb{E}_s[\|s-\tilde{s}_T\|^2]}{a_O(s,\mathcal{F},\varepsilon)+\varepsilon^2}\right] \leq C_0(K). \tag{3.12}$$

La précision obtenue, pour chaque N, par l'estimateur seuillé, qui est le cas particulier de l'estimateur pénalisé dans le cas de la sélection complète de variables avec des (φ_λ) orthonormés, est donc aussi bonne asymptotiquement que celle obtenue en choisissant pour chaque N le meilleur estimateur possible.

Mais nous pouvons déjà remarquer qu'en observant ainsi de façon globale le comportement asymptotique, nous passons à côté du fait que (3.7) peut parfois être amélioré, comme nous le voyons avec (3.8). Ceci est dû au fait que nous ne regardons ici, pour chaque N et chaque estimateur, que le pire des cas, et cela cache ce qui peut se passer si on a a priori plus d'informations

sur s. Dans le cas de la diffèrence entre (3.7) et (3.8), nous avons vu que pour certaines valeurs de s, la différence de précision peut être importante, mais il n'est pas difficile d'imaginer que le pire des cas n'est pas mieux estimé dans un cas que dans l'autre (essayer par exemple avec $s = \sum_{\lambda\in\Lambda} \varepsilon\sqrt{2(1+\log N)}\varphi_\lambda$: on voit que le résultat peut facilement être aussi mauvais avec les deux inégalités, et ce sont des cas comme ceux-là qui peuvent cacher, lorsque l'on adopte un point de vue trop global, les cas où les résultats sont sensiblement différents...).

3.2.2 Stratégie à poids variables.

Nous allons maintenant essayer d'améliorer la stratégie précédente en introduisant des poids variables, dépendant de la dimension : $L_m = L(|m|)$. On a :

$$\Sigma = \sum_{D=1}^{N} C_N^D \exp\left[-DL(D)\right] \leq \sum_{D=1}^{N} \left(\frac{eN}{D}\right)^D \exp\left[-DL(D)\right] \tag{3.13}$$

$$\leq \sum_{D=1}^{N} \exp\left[-D\left[L(D) - 1 - \log\left(\frac{N}{D}\right)\right]\right]. \tag{3.14}$$

Ainsi, si nous posons $L(D) = 1+\theta+\log(N/D)$, on obtient $\Sigma \leq \sum_{D=1}^{\infty} e^{-D\theta} = [e^\theta - 1]^{-1}$. En posant $\theta = \log 2$, on a la majoration suivante pour l'index de précision :

$$a_I(s, \mathcal{S}, \varepsilon) \leq \inf_{m\in M} \left\{d^2(s, S_m) + \varepsilon^2 |m|[1 + \log(2N/|m|)]\right\} + \varepsilon^2. \tag{3.15}$$

Etant donnée la ressemblance entre (3.15) et (3.3), l'index de précision de cette stratégie vérifie des inégalités analogues à (3.7) et (3.8). Ceci pourrait nous faire penser que la stratégie à poids variables n'apporte pas d'amélioration. Ceci est nécessairement vrai si l'on adopte un point de vue asymptotique puisque d'après (3.10) et (3.12) la stratégie à poids constants nous fournissait déjà, de ce point de vue, les meilleurs résultats que nous pouvions espérer. Mais ce point de vue cache en fait que dans certains cas, les poids variables apportent une meilleure précision que les poids constants. Comme nous l'avions fait remarquer à la fin de la précédente section, il faudra, pour vérifier qu'il est possible d'améliorer la précision obtenue, adopter un point de vue moins global. C'est ce que nous ferons dans la prochaine partie.

Mais nous voulons d'abord expliciter l'estimateur par projection pénalisé dans le cas de cette stratégie. Encore une fois, si on suppose la famille $\{\varphi_\lambda\}_{\lambda\in\Lambda}$

orthonormale, ce calcul ne pose pas de problème : nous commençons par introduire la définition suivante, qui nous sera utile.

Définition 8 *Etant donné un ensemble fini de nombres réels $\{b_i\}_{i\in I}$, de cardinal N, on note $\{b_{(j)(I)}\}_{1\leq j\leq N}$ le même ensemble de réels rangés selon l'ordre décroissant de leurs valeurs absolues : $|b_{(1)(I)}| \geq |b_{(2)(I)}| \geq \ldots \geq |b_{(N)(I)}|$. On appelle alors $\bar{I}[D]$ pour tout $1 \leq D \leq N$ le sous-ensemble $\{b_{(j)(I)}\}_{1\leq j\leq D}$ (ce sont donc les D plus grands b_i en module).*

Nous allons maintenant calculer l'estimateur. On a :

$$\inf_{m\in\mathcal{M}}\left\{-\sum_{\lambda\in m}\hat{\beta}_\lambda^2 + K\varepsilon^2|m|\Big(1+\sqrt{2L(|m|)}\Big)^2\right\} \tag{3.16}$$

$$= \inf_{0\leq D\leq N}\left\{-\sup_{\{m\in\mathcal{M}|\,|m|=D\}}\sum_{\lambda\in m}\hat{\beta}_\lambda^2 + K\varepsilon^2 D\Big(1+\sqrt{2L(D)}\Big)^2\right\} \tag{3.17}$$

$$= \inf_{0\leq D\leq N}\left\{-\sum_{j=1}^{D}\hat{\beta}_{(j)(\Lambda)}^2 + K\varepsilon^2 D\Big(1+\sqrt{2L(D)}\Big)^2\right\} \tag{3.18}$$

On note $\hat{D}$ un D pour lequel l'infimum ci-dessus est atteint. $\hat{m} = \bar{\Lambda}[\hat{D}]$ minimise alors le critère de pénalité : l'estimateur recherché est $\sum_{\lambda\in\hat{m}}\hat{\beta}_\lambda\varphi_\lambda$.

3.2.3 Comparaison entre les deux stratégies

Comme promis dans la précédente section, nous allons étudier l'apport de la stratégie à poids constants. Comme nous l'avions vu, il faut, pour voir la différence, adopter un point de vue moins global, et donc imposer à s d'appartenir à des parties plus restreintes que $\mathbb{S}$ de $\mathbb{H}$. On s'intéresse donc ici aux ensembles $\mathbb{S}_D = \bigcup_{\{m\in\mathcal{M}||m|=D\}} S_m$. Nous supposerons que la famille $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ est orthonormale.

D'après (3.15), l'estimateur pénalisé $\tilde{s}$ dans le cas de la sélection complète à poids variables, vérifie :

$$\sup_{s\in\mathbb{S}_D}\mathbb{E}_s\Big[\|\tilde{s}-s\|^2\Big] \leq C_0(K)\varepsilon^2[D + D\log(2N/D) + 1]. \tag{3.19}$$

Alors, le résultat suivant, qui est démontré dans l'article de Birgé et Massart, montre que $\tilde{s}$ est minimax sur les espaces $\mathbb{S}_D$, $D \geq 1$:

Théorème 2 *Il existe deux constantes universelles positives κ et κ' telles que*

$$\kappa\varepsilon^2 D[1+\log(N/D)] \leq R_M(\mathbb{S}_D,\varepsilon) \leq \kappa'\varepsilon^2 D[1+\log(N/D)]. \tag{3.20}$$

pour tout $\varepsilon > 0$ et $1 \leq D \leq N$.

Mais nous allons montrer qu'en ce qui concerne le cas des poids constants, l'estimateur pénalisé – qui est obtenu par la sélection seuillé, cf.(3.9) – n'est pas minimax sur les ensembles $\mathbb{S}_D$. On peut déjà le suspecter en remarquant que, dans (3.8), le majorant de l'index de précision vérifie :

$$\sup_{s\in\mathbb{S}_D}\sum_{\lambda\in\Lambda}\Big(\beta_\lambda^2\wedge\varepsilon^2[1+\log N]\Big)=D\varepsilon^2[1+\log N]. \tag{3.21}$$

Mais nous allons même démontrer un résultat plus fort :

Proposition 1 *Soient $T>0$, $\tilde{s}_T$ l'estimateur seuillé défini par (3.9), $m\in\mathcal{M}$. On a alors :*

$$\mathbb{E}_s\Big[\|\tilde{s}_T-s\|^2\Big]\geq\frac{|\,m|\,T^2}{2}\qquad si\ s=T\sum_{\lambda\in m}\varphi_\lambda. \tag{3.22}$$

Preuve : On a, pour tout $s\in\mathbb{S}$:

$$\mathbb{E}_s\Big[\|\tilde{s}_T-s\|^2\Big]=\sum_{\lambda\in\Lambda}\mathbb{E}_s\bigg[\Big(\beta_\lambda-\hat{\beta}_\lambda\mathbb{1}_{\{|\hat{\beta}_\lambda|>T\}}\Big)^2\bigg]. \tag{3.23}$$

D'où :

$$\mathbb{E}_s\Big[\|\tilde{s}_T-s\|^2\Big]\geq\sum_{\lambda\in m}\mathbb{E}_s\bigg[\Big(T-\hat{\beta}_\lambda 1\mathrm{l}_{\{|\hat{\beta}_\lambda|>T\}}\Big)^2\bigg]=|\,m|\,\mathbb{E}_s\bigg[\Big(T-(T+\varepsilon\xi)1\mathrm{l}_{\{|T+\varepsilon\xi|>T\}}\Big)^2\bigg]. \tag{3.24}$$

où ξ est une variable aléatoire normale standard. La proposition découle alors de :

$$\Big(T-(T+\varepsilon\xi)\mathbb{1}_{\{|T+\varepsilon\xi|>T\}}\Big)^2\mathbb{1}_{\{\xi<0\}}\geq T^2\mathbb{1}_{\{\xi<0\}}. \tag{3.25}$$

□

On applique maintenant la proposition avec, par exemple, $T\geq\varepsilon\sqrt{\log N}$. Alors, d'après (3.22) :

$$\sup_{s\in\mathbb{S}_D}\mathbb{E}_s\Big[\|\tilde{s}_T-s\|^2\Big]\geq\varepsilon^2(D/2)\log N. \tag{3.26}$$

Pour $D=N$, on perd ici un facteur $\log N$ par rapport au risque minimax... En fait, toujours pour $T\geq\varepsilon\sqrt{\log N}$, on a :

$$\frac{\sup_{s\in\mathbb{S}_D}\mathbb{E}_s\big[\|\tilde{s}_T-s\|^2\big]}{\sup_{s\in\mathbb{S}_D}\mathbb{E}_s\big[\|\tilde{s}-s\|^2\big]}\geq C(K)\frac{\log N}{1+\log N/D}. \tag{3.27}$$

Ce qui devrait achever de nous convaincre que la stratégie à poids variables est sensiblement meilleure que celle à poids constants.

4 Simulations numériques

Afin de tester l'efficacité de la méthode de sélection de modèles précédemment décrite, nous l'avons expérimentée dans un cas simple (l'interpolation de fonctions), et nous nous sommes intéressés au choix de la constante K. Les résultats qui suivent ont été obtenus à l'aide de Scilab.

Nous n'avons pas cherché à résoudre un problème réel (ce qui est bien l'objectif de la sélection de modèles), mais nous sommes partis de s à trouver, nous avons simulé une réalisation aléatoire du processus Y, puis nous avons appliqué la méthode de minimisation du critère pénalisé. De cette façon, nous avons pu comparer le modèle $\hat{m}$ et le meilleur modèle $\bar{m}$ de la famille $\{S_m\}_{m\in\mathcal{M}}$ (l'oracle).

Rappelons brièvement le problème posé (décrit en 1.1.3) : on considère une fonction $f : [0;1] \to \mathbb{R}$, n points $x_1, \ldots, x_n$ de $[0;1]$, et on observe $Y = (Y_1, \ldots, Y_n)$ avec $Y_i = f(x_i) + \sigma\xi_i$, les ξ_i étant i.i.d. $\mathcal{N}(0,1)$. On voudrait estimer f par une fonction constante sur les intervalles $[j/m; (j+1)/m[$, le problème étant de choisir convenablement m. Avant d'aller plus loin, observons sur un exemple comment fonctionne en pratique la méthode.

4.1 Un exemple

Prenons $f(x) = \sin(7x) - \sin(5x)$ sur $[0;1]$, $n = 100$ points, $\sigma^2 = 0,05$, $K = 2,5$, $L_m = 0$, pour $m \in \mathcal{M} = \{1, \ldots, n\}$. Une réalisation de Y est représentée à la figure 2. C'est tout ce dont on dispose comme information, et on voudrait retrouver f.

La première étape est d'estimer la variance. En effet, on a toujours supposé qu'elle était connue, mais ce n'est pas le cas en pratique. Pour cela, nous allons utiliser un estimateur qui donne de bons résultats, sans toutefois en donner la preuve. On a vu au (2.1) que $\mathbb{E}_s\left[\|\hat{s}_m\|^2\right] = \|s_m\|^2 + \varepsilon^2 m$. Comme $\|s_m\|^2$ varie peu (par rapport à l'autre terme) pour m grand, $\mathbb{E}_s\left[\|\hat{s}_m\|^2\right]$ est asymptotiquement linéaire en m , de pente $\varepsilon^2 = \sigma^2/n$. Nous allons donc estimer σ^2 à l'aide d'une régression linéaire sur $\|\hat{s}_m\|^2$ pour m "grand" (c'est-à-dire proche de n). La figure 3 montre que ce raisonnement semble correct. On obtient ici $\widehat{\sigma^2} = 0{,}046$ en faisant la régression sur les 15 dernières valeurs de m, soit une erreur relative de 7%.

Nous pouvons désormais effectuer la sélection de modèles proprement dite, comme elle a été décrite dans les sections précédentes. Tout d'abord, on calcule $crit(m)$ pour tout m entre 1 et n (figure 4), puis en cherche $\hat{m}$ qui minimise le critère pénalisé. Ici, $\hat{m} = 17$ (le modèle sélectionné) et $\bar{m} = 23$ (le modèle idéal), les erreurs correspondantes étant 0,017 et 0,013, soit un

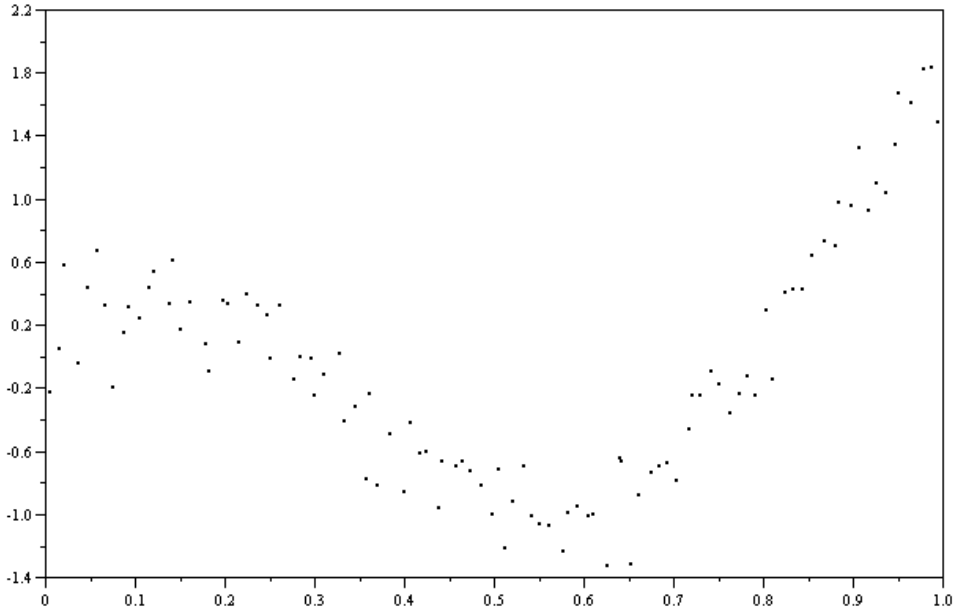

FIG. 2 – L'observation

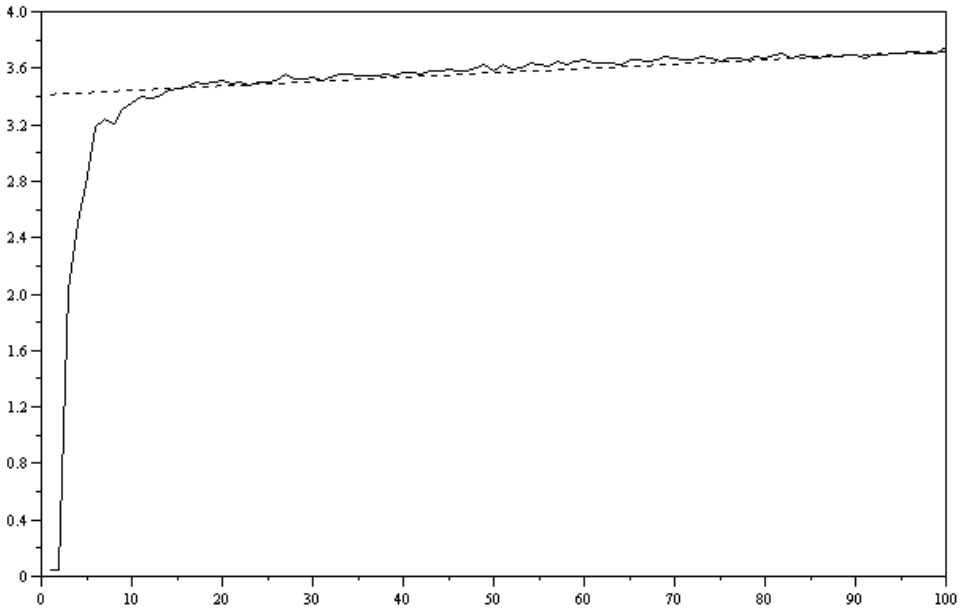

FIG. 3 – $\|\hat{s}_m\|^2$ en fonction de m ; en pointillés : l'approximation linéaire

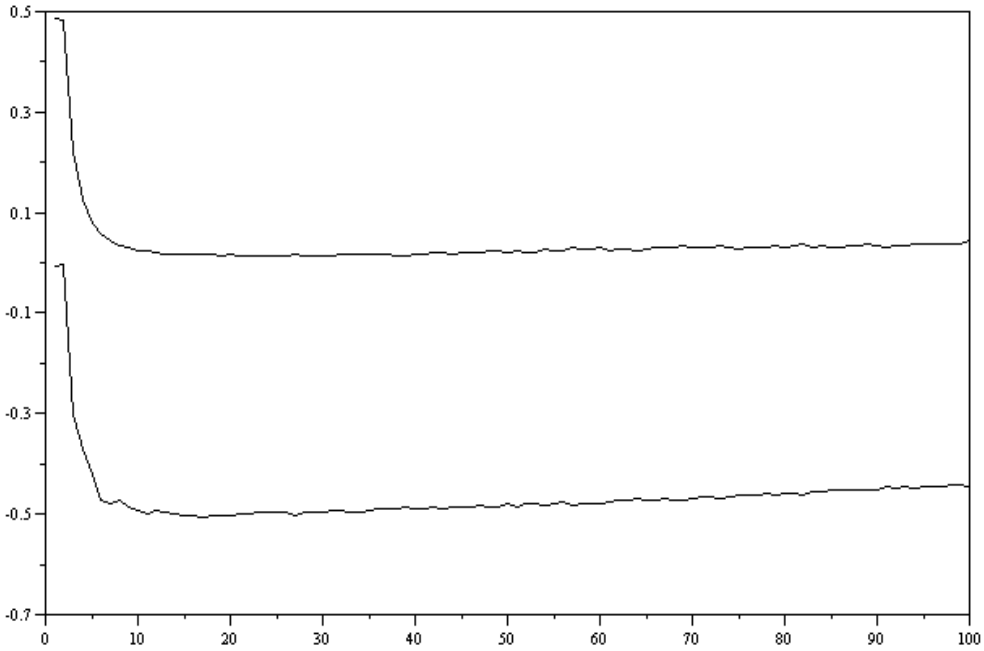

FIG. 4 – en bas : $crit(m)$ – en haut : $\|s - \hat{s}_m\|^2$

rapport de $1,3$, ce qui est un excellent résultat. En effet, on a un modèle optimal à une *petite* constante près, sans avoir a priori d'informations sur la fonction cherchée.

Pour finir, comparons les estimateurs par projection $\hat{s}_m$ pour $m = 3$, $m = \hat{m} = 17$, $m = \bar{m} = 23$ et $m = 100$, à l'aide des figures 5 et 6. Les valeurs extrêmes de m sont clairement mauvaises (il ne suffit pas de faire grandir le modèle pour améliorer le résultat). En revanche, il n'est pas évident graphiquement que $m = 23$ donne un meilleur résultat que $m = 17$.

4.2 Choix de K

Dans tout ce qui précède, on a supposé K fixée, et on ne s'est jamais préoccupé de son influence véritable sur les performances de la méthode de sélection. N'oublions pas que tous les résultats d'optimalité font intervenir une constante qui dépend de K, et qu'ils n'auraient aucun intérêt si cette constante était en réalité de l'ordre de 200 (c'est le minimum de $C_0(K)$ pour $K > 1$). Il faudrait prendre n très grand pour s'apercevoir des qualités asymptotiques de l'algorithme. Il est ainsi intéressant de chercher un ordre de grandeur de la meilleure valeur de K. Nous allons étudier ce qui se passe dans un cas particulier (avec seulement deux fonctions et n fixé). Nous essaierons

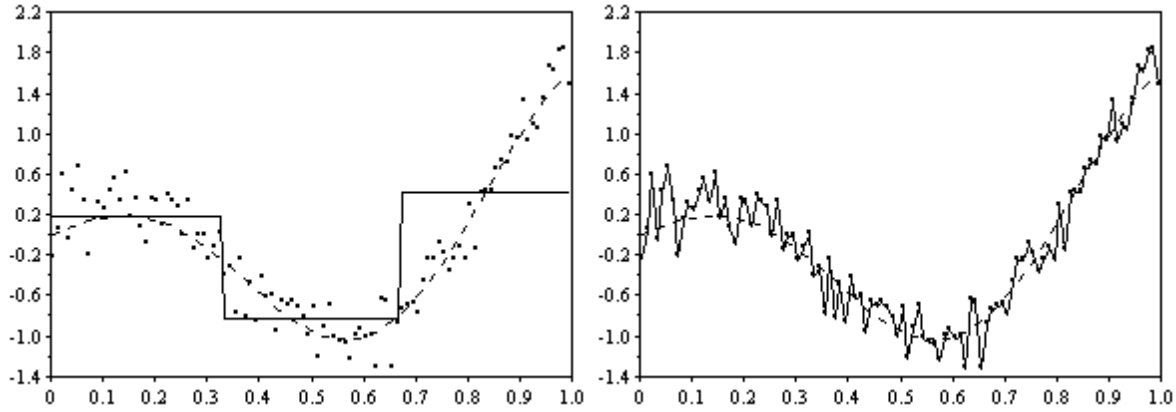

FIG. 5 – $m = 3\,;\ \|s - \hat{s}_m\|^2 = 0,21$ – $m = 100\,;\ \|s - \hat{s}_m\|^2 = 0,043$

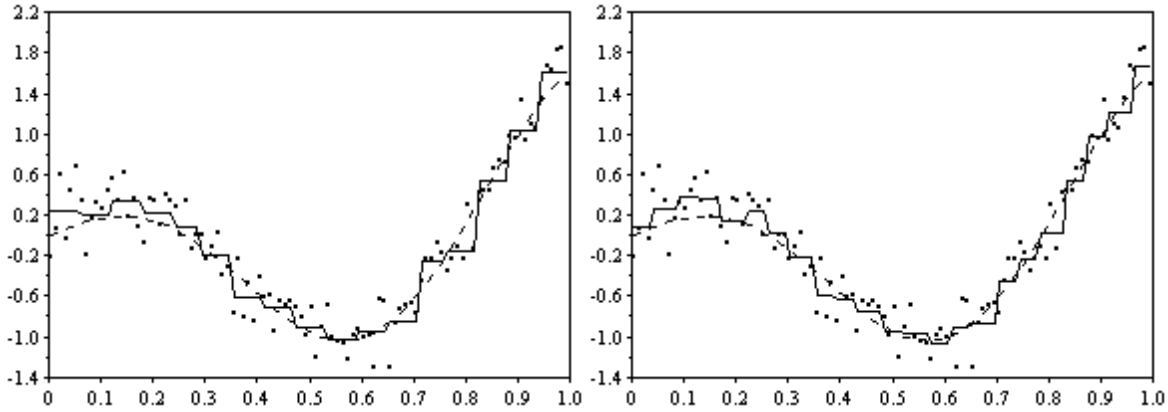

FIG. 6 – $m = 17\,;\ \|s - \hat{s}_m\|^2 = 0,017$ – $m = 23\,;\ \|s - \hat{s}_m\|^2 = 0,013$

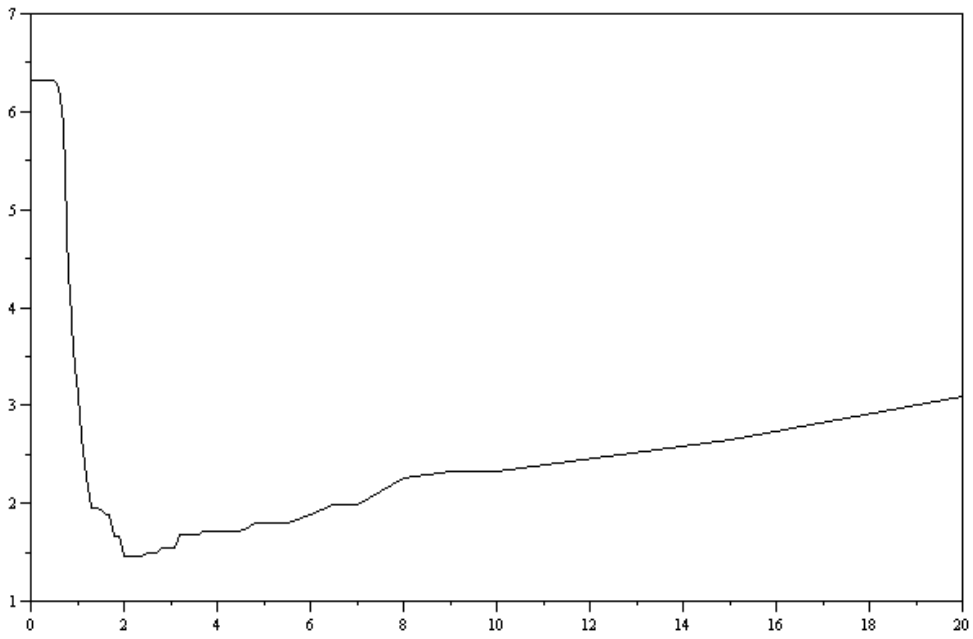

FIG. 7 – Les valeurs de K à éviter

également de comprendre la nécessité d'avoir $K > 1$, qui n'est nullement un artifice technique.

On prend $n = 150$ points, $\sigma^2 = 0,1$ (pour optimiser K, on ne va pas chercher à estimer la variance dans notre sélection de modèles), des poids $L_m = 0$, et la fonction $f : x \mapsto \sin(7x) - \sin(5x)$. On fait varier K de 0 à 20, et on observe la moyenne du rapport $\|s - \tilde{s}\|^2 / \|s - \hat{s}_{\bar{m}}\|^2$ (pour 10 observations différentes) en fonction de K (figure 7). La courbe obtenue permet déjà d'éliminer certaines valeurs de K : $K \leq 1$ et K grand (i.e. $K \geq 6$). On avait déjà eu cette intuition avec la constante donnée par le théorème 1, elle se précise désormais.

Pour chercher quelle valeur de K est optimale, nous avons effectué d'autres simulations (30 observations au total, avec $f(x) = \sin(7x) - \sin(5x)$ et $f(x) = \sin(\pi x)$) en nous limitant à $1 \leq K \leq 4$. La figure 8 représente la courbe obtenue, dont on déduit $K_{opt} \approx 2,4$ avec $\|s - \tilde{s}\|^2 / \|s - \hat{s}_{\bar{m}}\|^2 = 1,39$ pour cette valeur de K. Cependant, l'important n'est pas d'optimiser K mais plutôt d'avoir $pen(m) = K\hat{\varepsilon^2}m$ rendant la sélection de modèles optimale. Comme on n'a qu'une estimation de la variance, la valeur de $K\hat{\varepsilon^2}$ est proche de $K\varepsilon^2$, et il est donc intéressant de choisir K au centre du plateau inférieur de la courbe 8, afin de minimiser l'influence d'une erreur d'estimation de la variance. Dans ces conditions, on ferait sans doute mieux de choisir $K = 2,5$

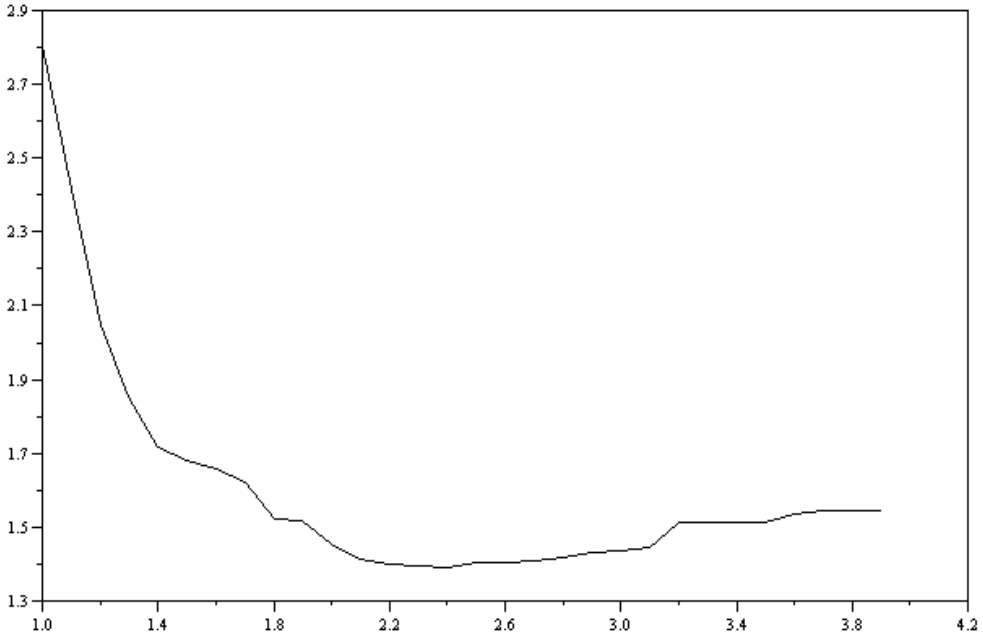

FIG. 8 – Optimisation de K

ou $K = 2,6$ plutôt que K_{opt}, un peu trop proche de l'extérieur de ce plateau.

Revenons pour finir sur les valeurs K à éviter. Pourquoi est-ce mauvais de prendre $K < 1$ ou K grand ? La réponse est donnée par la comparaison des courbes $crit(m)$ pour ces différentes valeurs de K. L'idéal est d'atteindre un minimum au même point que $\|s - \hat{s}_m\|^2$, et c'est à peu près ce qui se passe avec $K = 2,5$ (figure 4).

Lorsque $K < 1$, le critère pénalisé est décroissant pour m grand (figure 9), ce qui conduit toujours à sélectionner $m = n$. La figure 5 montre bien que c'est un mauvais choix. En remarquant que $\mathbb{E}_s[crit(m)] = -\|s_m\|^2 + \varepsilon^2 m(K-1)$, comme $\|s_m\|^2$ varie peu pour m grand, on voit que la condition $K > 1$ s'interprète effectivement comme une condition de croissance pour les grandes valeurs de m.

Pour de grandes valeurs de K, bien que la remontée de la courbe 7 soit moins prononcée qu'au voisinage de 1, on est tout de même conduits à de mauvais choix. En effet, comme le montre la figure 10, le critère pénalisé commence à croître trop tôt par rapport à l'erreur réelle, et donc on choisit un modèle de dimension trop petite. C'est également désastreux en termes de risque comme on l'a vu avec la figure 5.

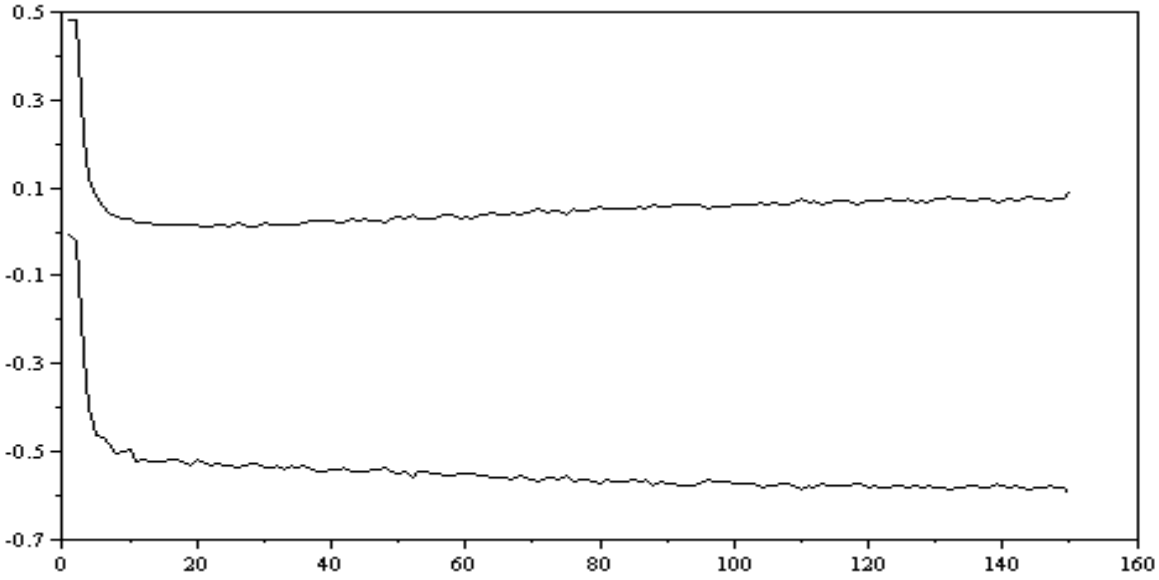

FIG. 9 – $crit(m)$ et $\|s - \hat{s}_m\|^2$ avec $K = 0,2$

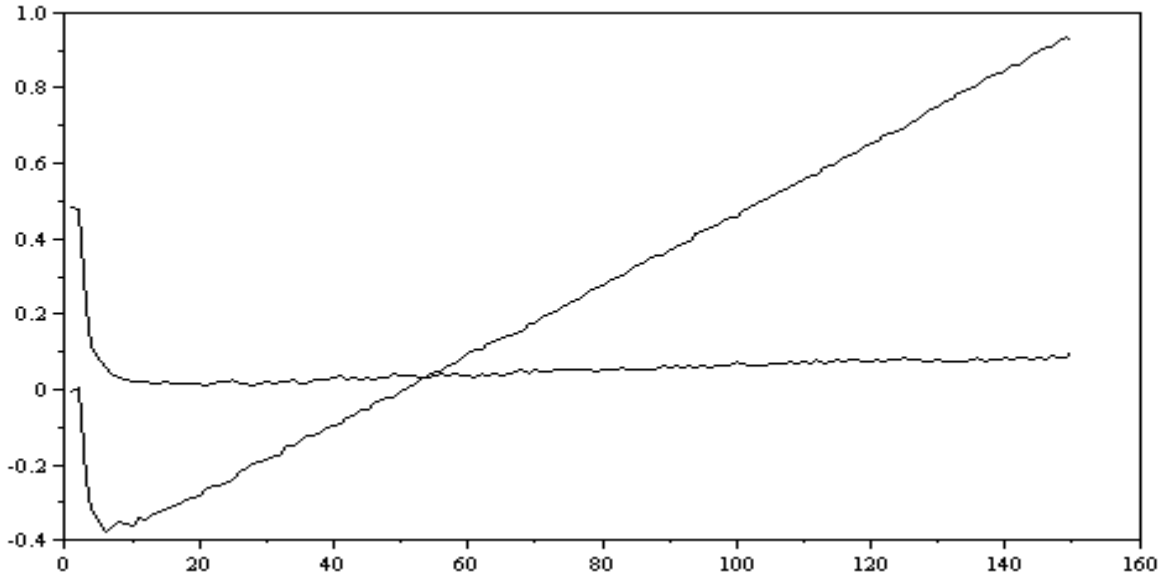

FIG. 10 – $crit(m)$ et $\|s - \hat{s}_m\|^2$ avec $K = 15$

Bien sûr, les simulations que nous avons réalisées ne sont pas suffisantes pour affirmer que $K = 2,4$ est effectivement optimale. Il faudrait faire varier σ^2, n et f (pour vérifier que K_{opt} ne varie pas), tester d'autres types de modèles (découpages irréguliers de $[0\,;1]$, polynômes, polynômes trigonométriques, etc.), avec un plus grand nombre de simulations dans chaque cas. On trouve alors $K_{opt} \approx 2$. L'intérêt de cette expérimentation est surtout de voir que les constantes obtenues avec le théorème sont loin d'être optimales, et que la sélection de modèles par pénalisation marche assez bien en pratique.

Bibliographie

BIRGÉ, L. et MASSART, P. (2000). Gaussian model selection, 1-26.

DONOHO, D.L. et JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.