

MATHÉMATIQUES

Un point de vue non-asymptotique pour la sélection de modèle

Pascal Massart¹

1. Introduction

Si la formation initiale d'un mathématicien ne comporte pas nécessairement un cours de probabilités ou a fortiori de statistique mathématique, notre quotidien, lui, est riche d'expressions qui empruntent au vocabulaire de la statistique. Notre mémoire est encombrée de bribes de phrase lues ou entendues ici ou là : « *les statistiques du commerce extérieur sont mauvaises* », « *le chômage ce mois-ci a baissé en données corrigées des variations saisonnières* », « *le chouchou des sondages est apparu détendu à la sortie de son quartier général* », « *les dernières estimations le donnent gagnant au deuxième tour* », « *le contrôle positif à la testostérone a été confirmé après analyse de l'échantillon témoin* », etc...

Afin de familiariser le lecteur mathématicien non averti avec les concepts et le vocabulaire de base de la statistique, il peut donc être utile (voire même judicieux) de s'appuyer sur la connaissance empirique de la statistique qu'il possède comme tout citoyen avec l'intention d'intégrer progressivement cette connaissance dans un formalisme mathématique. C'est avec ce point de vue que nous chercherons dans un premier temps à introduire la sélection de modèle au travers d'un exemple que chacun d'entre nous a eu l'occasion de rencontrer.

1.1. L'exemple des histogrammes

Les histogrammes sont communément utilisés comme outil de statistique descriptive pour représenter graphiquement des données.

1.1.1. Un outil de statistique descriptive

Supposons donc que nous disposions d'un ensemble fini de nombres réels x_1, x_2, \dots, x_n , où chaque valeur x_i correspond à une donnée. Par exemple x_i peut représenter le revenu annuel d'un individu i . Généralement on possède une bonne idée a priori d'un intervalle $[a, b]$ dans lequel varient ces données et pour simplifier on peut supposer (quitte à effectuer une transformation affine une fois pour toute) que les valeurs considérées varient dans l'intervalle $[0, 1]$. Pour définir un histogramme on choisit une partition $m = \{I_0, \dots, I_D\}$ de $[0, 1]$

¹ Université Parid-Sud.

par $D + 1$ intervalles dont les extrémités sont données par une suite croissante $y_0 = 0 < y_1 < \dots < y_D < y_{D+1} = 1$. Autrement dit, on a, pour chaque $j < D$, $I_j = [y_j, y_{j+1}[$ et $I_D = [y_D, y_{D+1}]$. Pour chaque j , on calcule le nombre n_j de données tombant dans l'intervalle I_j , à savoir

$$n_j = \sum_{i=1}^n \mathbb{1}_{I_j}(x_i)$$

et l'histogramme des données correspondant à la partition m est tout simplement défini comme étant la fonction de $[0, 1]$ dans \mathbb{R}

$$(1) \quad x \mapsto \sum_{j=0}^D \frac{n_j}{n(|I_j|)} \mathbb{1}_{I_j}(x),$$

avec pour chaque j , $|I_j| = y_{j+1} - y_j$. Cette fonction est naturellement constante sur chacun des morceaux de la partition m et c'est le graphe ou plutôt l'épigraphe de cette fonction qui est usuellement utilisé comme outil de représentation graphique des données. Noter que cette fonction est positive ou nulle et d'intégrale égale à 1, c'est donc une *densité de probabilité* sur $[0, 1]$. Si les points y_j sont équirépartis, c'est-à-dire si les intervalles I_j sont tous de même longueur $(D + 1)^{-1}$, la partition est dite régulière et l'histogramme est dit régulier. Même pour cette représentation parfaitement élémentaire des données, on voit poindre quelques questions fondamentales. La première d'entre elles est sans doute : qu'est-ce qu'une « bonne » partition m ou autrement dit comment peut-on mesurer la qualité de représentation des données par un histogramme sur une partition donnée m ? La seconde qui n'est pas comme nous le verrons sans lien avec la première est : comment en pratique choisir une partition m ? Sans formaliser davantage ce problème pour le moment, on peut aisément intuitiver qu'une partition trop pauvre, c'est-à-dire avec un trop petit nombre d'intervalles comparé à n , risque fort de conduire à une représentation non informative des données, alors qu'à l'inverse une partition trop riche qui comporterait un faible nombre de données par intervalle, fournit une représentation très erratique et donc pour le moins difficilement interprétable. Bien entendu, ces considérations sont purement qualitatives et ne permettent pas d'avancer vers un critère de qualité pour un histogramme et c'est à présent vers ce nouvel objectif que nous désirons nous diriger.

1.1.2. L'aléatoire s'en mêle

Pour progresser dans l'analyse des histogrammes, il nous faut introduire un cadre mathématique permettant de modéliser le fait qu'usuellement, les données qu'on souhaite représenter par un histogramme possèdent une certaine variabilité intrinsèque. Une façon de tenir compte de cette variabilité est d'utiliser une modélisation stochastique. C'est ce cadre que nous emploierons à présent dans lequel chaque donnée x_i est interprétée comme la réalisation $x_i = X_i(\omega)$ d'une variable aléatoire X_i , définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et prenant ses valeurs dans $[0, 1]$. Dans le cadre le plus simple que nous adopterons ici, les variables aléatoires sont supposées indépendantes et de même loi de probabilité P , ce qui correspond à l'idée que les données observées correspondent aux répétitions d'un

même phénomène aléatoire. On dit qu'alors X_1, \dots, X_n constitue un n -échantillon de loi P , ce qui mathématiquement se résume par la formule suivante :

$$\mathbb{P} \{ \omega \in \Omega \mid X_1(\omega) \in A_1, \dots, X_n(\omega) \in A_n \} = \prod_{i=1}^n P(A_i)$$

pour toute famille d'ensembles boréliens A_1, \dots, A_n de $[0, 1]$.

Dans ce nouveau cadre probabiliste, l'histogramme défini par (1) devient la réalisation d'une fonction aléatoire que nous noterons \widehat{s}_m . Plus précisément, nous avons pour tout $\omega \in \Omega$ et tout $x \in [0, 1]$

$$\widehat{s}_m(x, \omega) = \sum_{j=0}^D \frac{N_j(\omega)}{n(|I_j|)} \mathbb{1}_{I_j}(x),$$

avec pour chaque j , $N_j(\omega) = \sum_{i=1}^n \mathbb{1}_{I_j}(X_i(\omega))$. Supposons à présent que la loi P admette une densité de probabilité s par rapport à la mesure de Lebesgue, l'interprétation probabiliste de l'histogramme permet d'envisager la qualité de celui-ci pour approcher s .

1.1.3. L'histogramme vu comme un estimateur

En pratique, même s'il est raisonnable d'admettre que les données observées sont issues de la répétition d'un même phénomène aléatoire et que la loi de probabilité P commune aux variables aléatoires X_i admet une densité s par rapport à la mesure de Lebesgue, il est par contre exclu de considérer que la loi de probabilité P et par voie de conséquence cette densité s est connue. C'est le but même de la statistique dite « inférentielle » que d'obtenir des renseignements sur P à partir de l'observation d'une réalisation du n -échantillon X_1, \dots, X_n .

Rappelons que pour chaque réalisation de ce n -échantillon \widehat{s}_m est une densité de probabilité sur $[0, 1]$. Ce nouveau point de vue consiste donc à considérer à présent l'histogramme \widehat{s}_m comme une approximation aléatoire de la densité inconnue s fondée sur X_1, \dots, X_n , c'est-à-dire dans le langage de la statistique un *estimateur* de s . La distorsion entre la densité estimée \widehat{s}_m et la « vraie » densité s peut-être mesurée par exemple par le carré de la *distance de Hellinger* $\|\sqrt{s} - \sqrt{\widehat{s}_m}\|^2$, où $\|\cdot\|$ désigne la norme dans $\mathbb{L}_2([0, 1])$. Cette façon de mesurer la distorsion n'est évidemment pas la seule possible mais elle a le mérite d'une part d'être simple et d'autre part d'être invariante par un changement de mesure dominante. Évidemment l'erreur $\|\sqrt{s} - \sqrt{\widehat{s}_m}\|^2$ tout comme \widehat{s}_m est aléatoire. Afin de résumer la qualité d'estimation par un nombre plutôt que par une variable aléatoire, on a coutume d'en prendre l'espérance pour considérer le *risque de Hellinger*

$$\mathbb{E}_s \left[\|\sqrt{s} - \sqrt{\widehat{s}_m}\|^2 \right] = \int \|\sqrt{s} - \sqrt{\widehat{s}_m}\|^2 d\mathbb{P}_s(\omega).$$

Afin de rappeler explicitement leur dépendance en la densité inconnue s , on a pris soin d'utiliser ci-dessus les symboles \mathbb{P}_s et \mathbb{E}_s pour noter respectivement la probabilité ou l'espérance d'un événement ou d'une variable aléatoire fonction des variables observées X_1, \dots, X_n lorsque celles-ci sont indépendantes et de même loi

de densité s . Nous disposons à présent avec le risque de Hellinger d'une mesure objective de la qualité d'un histogramme.

1.1.4. Le « modèle histogramme »

L'histogramme \widehat{s}_m est une densité de probabilité constante par morceaux sur la partition m . On peut donc se demander quel rôle particulier joue \widehat{s}_m parmi toutes les fonctions possédant cette propriété. Autrement dit, si nous introduisons le *modèle histogramme*

$$S_m = \left\{ \sum_{j=0}^D a_j \mathbb{1}_{I_j} \mid a_0, \dots, a_D \in \mathbb{R}_+ \text{ et } \sum_{j=0}^D a_j |I_j| = 1 \right\}$$

la question est : quelle propriété spécifique possède \widehat{s}_m comme élément de ce modèle S_m ? C'est le moment d'introduire une notion clef en statistique mathématique : *la vraisemblance*. t étant une densité de probabilité donnée, la vraisemblance en t est la densité de l'observation X_1, \dots, X_n sous l'hypothèse que $s = t$, évaluée au point observé X_1, \dots, X_n , à savoir

$$\prod_{i=1}^n t(X_i).$$

On vérifie aisément que \widehat{s}_m maximise cette vraisemblance lorsque t parcourt S_m , ce qu'on peut encore synthétiser de la manière suivante

$$\widehat{s}_m = \operatorname{argmax}_{t \in S_m} \sum_{i=1}^n \ln(t(X_i)).$$

1.1.5. Le problème du choix de modèle

Nous venons de voir qu'à chaque partition m on peut faire correspondre le modèle S_m des densités constantes par morceaux sur m et associer à ce modèle l'estimateur par histogramme qui se trouve être l'estimateur par maximum de vraisemblance sur ce modèle S_m . Formaliser le problème du choix d'un « bon » modèle S_m revient donc à formaliser celui de la sélection d'un « bon » estimateur par histogramme \widehat{s}_m . Une façon naïve de procéder consiste à raisonner comme suit. Partant d'une collection finie \mathcal{M} de partitions (par exemple la collection de toutes les partitions régulières à au plus n morceaux), le meilleur estimateur par histogramme est défini par la partition $m_0(s)$ minimisant

$$m \mapsto \mathbb{E}_s \left[\left\| \sqrt{s} - \sqrt{\widehat{s}_m} \right\|^2 \right]$$

sur \mathcal{M} . Si cette définition paraît abstraitement satisfaisante, elle est inutilisable pour sélectionner effectivement une bonne partition en pratique puisque le risque de Hellinger considéré ci-dessus dépend malheureusement de la densité inconnue s . Tout le problème consiste donc à sélectionner une partition \widehat{m} construite à partir des seules observations X_1, \dots, X_n (et ne dépendant surtout pas de s), de telle façon que la performance en terme de risque de Hellinger de l'estimateur par histogramme

sélectionné $\hat{s}_{\hat{m}}$ soit comparable à celle de $\hat{s}_{m_0(s)}$. L'idée que nous allons étudier en détail ici consiste à choisir \hat{m} minimisant sur \mathcal{M} le critère suivant

$$m \mapsto - \sum_{i=1}^n \ln \hat{s}_m(X_i) + \text{pen}(m)$$

où $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}_+$ est une fonction dite de *pénalisation* convenable. Bien entendu toute la difficulté réside dans la définition judicieuse de la fonction de pénalité qui sera largement discutée dans ce qui suit. L'idée de choisir un modèle via un critère de type log-vraisemblance pénalisée remonte au début des années 70 avec les travaux précurseurs de Mallows et d'Akaike (voir [1], [13] et [24]). Il est temps à présent de quitter le strict exemple des histogrammes afin d'élargir le cadre de notre réflexion.

1.2. Inférence statistique

Le problème de base de l'inférence statistique consiste à prendre une décision à propos d'une quantité s liée à la loi inconnue d'une variable aléatoire observée \mathbf{X} . La nature de \mathbf{X} peut être diverse : il se peut qu'on observe un vecteur aléatoire ou un processus stochastique ou encore une image bruitée. De même s peut être un vecteur ou une fonction ou encore une image. On peut par exemple chercher à construire une zone de confiance pour s , c'est-à-dire une région aléatoire qui contient s avec une probabilité donnée. Partant d'une procédure d'estimation \hat{s} de s , (c'est-à-dire d'une fonction de l'observation \mathbf{X}) et d'une *fonction de perte* ℓ (typiquement ℓ est une distance ou le carré d'une distance comme le carré de la distance de Hellinger utilisé ci-dessus) permettant de préciser la qualité de \hat{s} , l'approche naturelle pour réaliser une telle construction consiste à analyser la répartition de $\ell(s, \hat{s})$. Il se trouve qu'en règle générale il est exclu d'évaluer de manière exacte la distribution de la procédure d'estimation. Il est alors essentiel de disposer d'outils pertinents d'approximation de cette répartition issus du Calcul des Probabilités.

1.2.1. La théorie asymptotique

Dans le cas où $\mathbf{X} = \mathbf{X}^{(n)}$ dépend d'un paramètre n (typiquement lorsque $\mathbf{X} = (X_1, \dots, X_n)$, où les variables X_1, \dots, X_n sont indépendantes et de même loi), la *théorie asymptotique* en statistique utilise les théorèmes limites (Théorème Central Limite, Principes de Grandes Déviations...) comme des outils d'approximation lorsque n tend vers l'infini. L'exemple historique le plus représentatif de cette approche est sans doute l'utilisation du Théorème Central Limite pour l'analyse du comportement lorsque n tend vers l'infini, de l'estimateur du maximum de vraisemblance sur un modèle paramétrique régulier. Si nous prenons à nouveau pour exemple le cadre de l'estimation de la densité dans lequel on observe X_1, \dots, X_n indépendantes et équidistribuées, dont la distribution commune admet une densité inconnue s par rapport à une mesure dominante μ . Un modèle S dans ce cas est simplement une partie de l'ensemble des densités de probabilité par rapport à μ . Un

modèle S étant donné, l'estimateur du maximum de vraisemblance \hat{s} (s'il existe!) est simplement défini comme minimiseur sur S du critère empirique

$$t \mapsto \sum_{i=1}^n -\ln t(X_i).$$

Lorsque le modèle S est paramétrique $S = \{s_\theta, \theta \in \Theta\}$, où Θ est un ouvert de \mathbb{R}^D , on écrit plutôt \hat{s} sous la forme $\hat{s} = s_{\hat{\theta}}$. Sous des conditions de différentiabilité convenables de s_θ par rapport au paramètre θ et à la condition que s appartienne effectivement au modèle S (et donc s'écrive $s = s_{\theta_0}$ pour un certain $\theta_0 \in \Theta$), le résultat classique de la théorie asymptotique évoqué plus haut concerne la normalité asymptotique de $\sqrt{n}(\hat{\theta} - \theta)$ lorsque n tend vers l'infini. On peut également garantir que la matrice de covariance apparaissant dans la loi gaussienne asymptotique est en un sens minimale, c'est ce qu'on appelle la propriété d'*efficacité asymptotique*. Plus récemment, avec les travaux séminaux de Dudley dans les années 70 sur les processus empiriques, la théorie des probabilités dans les espaces de Banach a profondément influencé le développement de la statistique asymptotique, conduisant à des avancées décisives dans le domaine de la théorie de l'efficacité asymptotique. Le lecteur intéressé trouvera dans les ouvrages de van der Vaart and Wellner [32] et van der Vaart [31] de nombreux résultats allant dans cette direction.

1.2.2. La sélection de modèle

Le problème majeur pour le statisticien est alors de définir un modèle S convenable. Il peut être délicat de deviner quel modèle paramétrique utiliser pour refléter un jeu de données réelles et il est clair qu'une erreur de modélisation, qui se traduit par un trop grand éloignement de s par rapport à S , peut conduire à une qualité d'estimation catastrophique. On peut être alors naïvement tenté de choisir un très grand modèle. Si on choisit S comme étant l'ensemble de toutes les densités ou comme un trop vaste sous-ensemble de celles-ci, il est bien connu que la procédure du maximum de vraisemblance devient inconsistante (voir [3]) ou sous-optimale (voir [6]). Déterminer par avance quel modèle utiliser pose donc des problèmes. Si on souhaite réaliser l'opération du choix d'un modèle convenable avec le plus d'objectivité possible, l'idée clef de la sélection de modèle consiste à s'appuyer sur les données elles-mêmes afin de construire un critère que le modèle choisi devra minimiser au sein d'une liste donnée, plutôt que de se fier au seul flair du modélisateur pour effectuer ce choix. Il s'agit donc ici de traiter un problème qui généralise celui du choix de la partition pour construire un histogramme. Plus précisément, si $(S_m)_{m \in \mathcal{M}}$ est une liste finie de modèles paramétriques réguliers où chacun des modèles S_m est défini par D_m paramètres et si $(\hat{s}_m)_{m \in \mathcal{M}}$ désigne la liste des estimateurs du maximum de vraisemblance correspondants, le critère de log-vraisemblance pénalisée d'Akaike (voir [1]) propose de sélectionner le modèle $S_{\hat{m}}$ tel que \hat{m} minimise le critère

$$m \mapsto - \sum_{i=1}^n \ln \hat{s}_m(X_i) + D_m$$

sur \mathcal{M} . La conception même de ce critère repose sur une heuristique qui s'appuie lourdement sur le comportement asymptotique de l'estimateur du maximum de

vraisemblance évoquée plus haut et notamment sur une de ses conséquences connue sous le nom de théorème de Wilks, qui assure que si $s \in S_m$ alors (sous des conditions de régularité convenables) la quantité

$$2 \left(- \sum_{i=1}^n \ln \widehat{s}_m(X_i) + \sum_{i=1}^n \ln s(X_i) \right)$$

converge en loi vers une loi du chi-deux à D_m degrés de liberté (c'est-à-dire la loi d'une somme de D_m carrés de variables aléatoires indépendantes et de même loi normale $\mathcal{N}(0, 1)$). D'autres critères proposés ultérieurement tels que le critère bayésien proposé par Schwartz, connu sous le nom de BIC (voir [28]) par exemple, possèdent exactement la même caractéristique : leur conception repose sur une approximation asymptotique qui sous-entend donc que la liste des modèles est fixée tandis que n tend vers l'infini.

1.2.3. Le point de vue non asymptotique

Il se trouve que dans plusieurs situations d'intérêt motivées par les applications, il est utile de laisser croître la taille des modèles avec n . Nous verrons d'autres exemples un peu plus loin mais il est clair que pour ce qui concerne les histogrammes réguliers par exemple, il est légitime de permettre au nombre de morceaux de varier librement entre 1 et n . Il peut même être utile de laisser croître avec n le nombre de modèles d'une dimension donnée.

Exemple : la détection de ruptures

La détection de ruptures sur la moyenne d'un signal discret fournit un cas d'école de ce type. Soit s une fonction sur $[0, 1]$ représentant un signal inconnu. Si nous observons à chaque instant j/n un signal bruité X_j , de telle sorte que le vrai signal $s(j/n)$ à l'instant j/n apparaisse comme l'espérance de X_j la question de la détection de ruptures sur la moyenne se formalise par la recherche d'une partition optimale (en un sens à préciser) de $[0, 1]$ par des intervalles dont les extrémités appartiennent à $\{j/n, 0 \leq j \leq n\}$ sur laquelle s soit une fonction constante par morceaux. Les motivations proviennent de l'analyse de signaux sismiques pour lesquels les instants de rupture (c'est-à-dire les extrémités des intervalles de la partition) correspondent à des couches géologiques différentes. Dans ce cas, à chaque partition m correspond un modèle S_m de fonctions constantes sur chacun des intervalles de la partition m . Dans cet exemple, pour chaque entier $D \leq n$, le nombre de modèles de dimension D , c'est-à-dire en fait le nombre de partitions à D morceaux, vaut $\binom{n-1}{D-1}$ et croît donc polynomialement par rapport à n .

Dans de telles circonstances l'analyse asymptotique classique n'est plus pertinente et une autre approche devient nécessaire que nous appellerons *non asymptotique*. Par non asymptotique, nous ne voulons pas dire que nous cherchons des résultats valables lorsque n est systématiquement modéré. L'idée est plutôt que quelle que soit la valeur de n , (et peut-être même surtout lorsque n est grand), il est utile d'autoriser la liste aussi bien que la taille des modèles à dépendre de n afin de garantir que l'un d'entre eux soit proche de s . Lorsque la cible s est une fonction, ceci permet d'utiliser toutes les connaissances issues de la *théorie de l'approximation* afin de définir des modèles dont les propriétés d'approximation à

des échelles variables sont bien connues (nous pensons à des polynômes par morceaux à pas et degrés variables par exemple). Dans les vingt dernières années le phénomène de concentration de la mesure a fait l'objet d'une recherche intense et féconde tout particulièrement sous l'impulsion des remarquables travaux de Michel Talagrand qui ont abouti à la découverte de nouvelles inégalités très puissantes en probabilités (voir en particulier [29] et [30]). Le principal avantage des inégalités de concentration est qu'à l'inverse des théorèmes limites, elles fournissent des outils non asymptotiques. C'est donc en un sens sans surprise que nous verrons ces inégalités jouer un rôle crucial dans l'élaboration d'une théorie non asymptotique pour la sélection de modèles telle qu'elle a émergé durant ces dix dernières années (voir en particulier [7] et [5]). Notre point de vue sera ici d'expliquer les idées et motivations centrales de cette théorie en les explicitant sur des exemples que nous espérons parlants.

2. La sélection de modèle gaussienne

2.1. La régression linéaire gaussienne

Nous commencerons notre analyse avec le *modèle linéaire gaussien* qui est sans aucun doute l'un des modèles les plus simples et les plus utilisés en statistique. On observe dans ce cas des variables aléatoires X_1, \dots, X_n structurées par le modèle de régression linéaire suivant :

$$X_i = \sum_{j=1}^N \beta_j \varphi_j(i) + \sigma \xi_i \text{ pour } 1 \leq i \leq n,$$

où les variables aléatoires ξ_i sont indépendantes et de même loi normale $\mathcal{N}(0, 1)$ alors que les nombres $\varphi_j(i)$ sont eux connus et représentent des valeurs observées de variables explicatives φ_j . Ici, le terme variable est à considérer dans l'acception usuelle de variable « économique » ou « physique ». En pratique, X_i correspond à la valeur prise par une observation réalisée à la i^{e} expérience et le modèle ci-dessus signifie donc que cette valeur dépend linéairement des valeurs $\varphi_j(i)$ prises par les variables φ_j pour cette même expérience, plus un terme de perturbation aléatoire représenté par la variable aléatoire $\sigma \xi_i$. Les paramètres β_j sont bien entendu inconnus mais nous supposerons par contre, dans un premier temps, le paramètre σ connu. Bien qu'irréaliste en pratique cette hypothèse simplifie grandement l'analyse. Par ailleurs nous reviendrons sur le problème d'estimation de σ dans un second temps. Le cadre ci-dessus fournit bien un modèle paramétrique pour la densité du vecteur \mathbf{X} dans \mathbb{R}^n par rapport à la mesure de Lebesgue puisque les variables X_1, \dots, X_n sont indépendantes avec pour lois respectives la distribution normale de moyenne $s_i = \sum_{j=1}^N \beta_j \varphi_j(i)$ et de variance σ^2 . Pour reformuler ceci de manière équivalente, on constate que le vecteur aléatoire \mathbf{X} suit une loi gaussienne multidimensionnelle, de moyenne $s = (s_i)_{1 \leq i \leq n}$ et de matrice de covariance $\sigma^2 I_n$, où I_n désigne la matrice identité d'ordre n . Si (et c'est ce que nous supposerons dans la suite) les vecteurs φ_j sont linéairement indépendants, ils engendrent un espace de dimension N que nous noterons S_N et il devient équivalent d'estimer le vecteur de paramètres β dans \mathbb{R}^N ou le vecteur s dans S_N . C'est un problème paramétrique qui

peut se résoudre par la méthode du maximum de vraisemblance. La vraisemblance de \mathbf{X} et son logarithme valent alors respectivement

$$(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - s_i)^2\right) \text{ et } -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - s_i)^2.$$

Introduisons à présent la norme euclidienne sur \mathbb{R}^n

$$\|x\|^2 = \frac{1}{n} \sum_{i=1}^n x_i^2, \text{ pour tout } x \in \mathbb{R}^n$$

et posons $\varepsilon = \sigma/\sqrt{n}$. Les raisons pour lesquelles nous avons renormalisé la norme euclidienne canonique et introduit le paramètre ε apparaîtront plus clairement par la suite. En tout cas, nous déduisons de ce qui précède que l'estimateur du maximum de vraisemblance \widehat{s}_N de s sur S_N est tout simplement la projection orthogonale du vecteur \mathbf{X} sur l'espace S_N . De plus l'invariance par rotation de la loi gaussienne multidimensionnelle $\mathcal{N}_n(0, I_n)$ garantit que la loi de $\varepsilon^{-2} \|\widehat{s}_N - s\|^2$ est identique à celle obtenue lorsque S_N est engendré par les N premiers vecteurs de la base canonique de \mathbb{R}^n . C'est donc une loi du chi-deux à N degrés de liberté. Par conséquent le *risque quadratique* de \widehat{s}_N se calcule explicitement par la formule

$$\mathbb{E}_s \left[\|\widehat{s}_N - s\|^2 \right] = N\varepsilon^2.$$

Il est intéressant de noter que le choix de \widehat{s}_N comme estimateur de s a encore un sens même si $s \notin S_N$. La formule de Pythagore permet de corriger l'expression du risque quadratique ci-dessus qui devient donc

$$(2) \quad \mathbb{E}_s \left[\|\widehat{s}_N - s\|^2 \right] = N\varepsilon^2 + \|s - s_N\|^2,$$

où s_N désigne la projection orthogonale de s sur S_N . Ce risque quadratique apparaît comme la somme de deux termes, l'un, appelé *terme de variance*, proportionnel au nombre de paramètres à estimer N et l'autre, appelé *terme de biais*, qui mesure la qualité d'approximation de la réalité que procure le modèle S_N . Ce second terme disparaît bien entendu lorsque le modèle est exact, c'est-à-dire contient s .

2.2. La sélection de variables

Dans le modèle linéaire gaussien exposé ci-dessus, le modèle S_N est supposé exact de telle sorte que le risque quadratique s'écrit $\mathbb{E}_s \left[\|\widehat{s}_N - s\|^2 \right] = N\varepsilon^2$. Cette approche qui repose sur le choix *a priori* d'un modèle peut conduire à des soucis de natures opposées. Afin de garantir une bonne qualité d'estimation on est tenté de prendre une valeur modérée pour N , c'est-à-dire de mettre une petite partie des variables explicatives dont on dispose dans le modèle. Si on omet des variables explicatives importantes non seulement s n'appartiendra pas à S_N mais surtout le terme de biais $\|s - s_N\|^2$ peut augmenter considérablement. A contrario, si pour contourner cette difficulté on utilise beaucoup de variables explicatives pour engendrer le modèle S_N , alors, même si le modèle est exact l'estimation sera de piètre qualité. Or il peut se faire que parmi les variables $\varphi_1, \dots, \varphi_N$, seules un petit nombre D d'entre elles soient réellement influentes. Cela signifie que si S_D désigne l'espace engendré par ces D variables (disons $\varphi_1, \dots, \varphi_D$) le terme de biais

$\|s - s_D\|^2$ va rester faible de sorte que $D\varepsilon^2 + \|s - s_D\|^2$ peut être sensiblement plus petit que $N\varepsilon^2$. On voit se dessiner ici une des premières idées importantes que nous souhaitons avancer : on peut tirer bénéfice de l'utilisation d'un modèle approché (ici S_D) plutôt que d'un modèle exact (ici S_N).

La problématique intéressante qui se dégage ici est donc celle de la *sélection de variables* qui, partant d'une famille (qui peut être vaste) de variables explicatives $\varphi_1, \dots, \varphi_N$, consiste à tenter de sélectionner les plus influentes d'entre elles. La seconde idée importante qui émerge ici est la suivante : la notion de risque d'estimation permet de bien formuler mathématiquement ce problème de sélection. En effet le « meilleur » sous-ensemble $\{\varphi_j, j \in m\}$ de variables est tout simplement celui qui minimise le risque quadratique de l'estimateur par projection orthogonale \widehat{s}_m sur l'espace S_m engendré par les $\varphi_j, j \in m$. Idéalement, on souhaiterait sélectionner m minimisant

$$\mathbb{E}_s \left[\|\widehat{s}_m - s\|^2 \right] = |m| \varepsilon^2 + d^2(s, S_m),$$

où $d^2(s, S_m) = \inf_{t \in S_m} \|s - t\|^2$. Bien entendu un tel sous-ensemble idéal $m(s)$ dépend de s qui est inconnu du statisticien et non pas de la seule observation \mathbf{X} . L'enjeu statistique est alors de construire une procédure \widehat{m} de sélection d'un sous-ensemble de $\{1, \dots, N\}$ qui ne dépende que de l'observation \mathbf{X} . Le critère de qualité que nous adopterons pour une telle procédure est celui qui découle naturellement de ce qui précède, c'est-à-dire un critère de performance en terme de risque pour l'estimateur par projection correspondant $\widehat{s}_{\widehat{m}}$. Plus précisément on souhaite que ce risque quadratique $\mathbb{E}_s \left[\|\widehat{s}_{\widehat{m}} - s\|^2 \right]$ soit aussi voisin que possible du risque quadratique de $\widehat{s}_{m(s)}$, soit

$$\inf_{m \subseteq \{1, \dots, N\}} |m| \varepsilon^2 + d^2(s, S_m).$$

2.3. Le modèle linéaire gaussien généralisé

Avant d'aller plus loin il est utile de fixer le cadre stochastique général dans lequel nous allons poser le problème de la sélection de modèle gaussienne. Considérons comme dans [8] le modèle linéaire gaussien généralisé défini de la manière suivante. Étant donné un espace de Hilbert séparable \mathbb{H} , on observe le processus \mathbf{X}^ε donné par

$$(3) \quad \mathbf{X}^\varepsilon(t) = \langle s, t \rangle + \varepsilon W(t) \text{ pour tout } t \in \mathbb{H},$$

où W désigne un *processus gaussien isonormal*, c'est-à-dire que W est une isométrie de \mathbb{H} sur un sous-espace gaussien de $\mathbb{L}_2(\Omega)$, s est un paramètre inconnu dans \mathbb{H} et ε un paramètre réel positif supposé connu.

2.3.1. Exemples

Voyons en détail quelles sont les possibilités de modélisation offertes par ce nouveau cadre en commençant par vérifier qu'il généralise bien le modèle linéaire gaussien fini-dimensionnel introduit plus haut.

Le modèle linéaire gaussien fini-dimensionnel

Dans ce cas on observe, comme indiqué plus haut,

$$(4) \quad X_i = s_i + \sigma \xi_i, \quad 1 \leq i \leq n,$$

où les variables aléatoires ξ_i sont indépendantes et de même loi normale $\mathcal{N}(0, 1)$. Si nous considérons le produit scalaire normalisé sur \mathbb{R}^n

$$\langle x, y \rangle = \frac{1}{n} \sum_{i=1}^n x_i y_i$$

associé à la norme $\|\cdot\|$ et si nous posons $W(t) = \sqrt{n} \langle \xi, t \rangle$, alors W est bien un processus gaussien isonormal et

$$\mathbf{X}^\varepsilon : t \mapsto \frac{1}{n} \sum_{i=1}^n X_i t_i$$

satisfait bien à (3) avec $\varepsilon = \sigma/\sqrt{n}$.

Le modèle de bruit blanc continu

Dans ce cas, on observe le processus $\{X^\varepsilon(x), x \in [0, 1]\}$ régi par l'équation différentielle stochastique suivante

$$(5) \quad dX^\varepsilon(x) = s(x) dx + \varepsilon dB(x) \quad \text{avec } X^\varepsilon(0) = 0,$$

où B désigne un mouvement brownien sur $[0, 1]$. Si nous définissons alors pour tout $t \in \mathbb{L}_2[0, 1]$, $W(t) = \int_0^1 t(x) dB(x)$, W est bien un processus gaussien isonormal sur $\mathbb{L}_2[0, 1]$ et $\mathbf{X}^\varepsilon(t) = \int_0^1 t(x) dX^\varepsilon(x)$ obéit bien à (3) dès lors que $\mathbb{L}_2[0, 1]$ est muni de son produit scalaire usuel $\langle s, t \rangle = \int_0^1 s(x) t(x) dx$. Typiquement s représente un signal et $dX^\varepsilon(x)$ représente le signal bruité reçu à l'instant x . Ce modèle s'étend aisément au cas multivarié si l'on considère un drap brownien multivarié B sur $[0, 1]^d$ et $\mathbb{H} = \mathbb{L}_2([0, 1]^d)$.

Le modèle de bruit blanc discret

Particularisons le modèle linéaire gaussien fini-dimensionnel au cas où $s_i = s(i/n)$, où s désigne une fonction définie sur $[0, 1]$. C'est-à-dire que

$$(6) \quad X_i = s(i/n) + \sigma \xi_i, \quad 1 \leq i \leq n,$$

où les variables aléatoires ξ_i sont indépendantes et de même loi normale $\mathcal{N}(0, 1)$. Si s est un signal, X_i représente le signal bruité à l'instant i/n . Ce modèle peut être vu comme une version discrétisée du modèle de bruit blanc continu. En effet, partant du bruit blanc continu, nous pouvons poser $\sigma = \varepsilon\sqrt{n}$ et

$$\xi_i = \sqrt{n}(B(i/n) - B((i-1)/n)), \quad \text{pour tout } i \in [1, n].$$

Le signal bruité reçu à l'instant i/n vaut alors

$$X_i = n(X^\varepsilon(i/n) - X^\varepsilon((i-1)/n)) = n \int_{(i-1)/n}^{i/n} s(x) dx + \sigma \xi_i.$$

Comme les propriétés du mouvement brownien garantissent que les variables ξ_i sont bien indépendantes et de même loi normale centrée réduite, nous revenons bien au modèle de signal discret avec $s_i = s^{(n)}(i/n)$, où $s^{(n)}(x) = n \int_{(i-1)/n}^{i/n} s(y) dy$

pour tout $x \in [(i-1)/n, i/n[$. Si le signal continu s est suffisamment régulier, la fonction en escalier $s^{(n)}$ représente une approximation convenable de s , ce qui établit un lien entre les modèles de bruit blanc discret et continu.

Le modèle de suite gaussienne

Bien que le processus donné par (3) ne soit assujéti à aucune base orthonormée particulière, sitôt qu'une telle base est donnée dans \mathbb{H} , on peut filtrer le processus sur cette base et obtenir ainsi une *suite gaussienne*. Supposons donc que \mathbb{H} soit de dimension infinie (le cas de la dimension finie ayant déjà été traité plus haut) et considérons une base orthonormée $\{\varphi_j, j \geq 1\}$ de \mathbb{H} . La suite des coefficients $\widehat{\beta}_j = \mathbf{X}^\varepsilon(\varphi_j)$ est alors structurée de la manière suivante

$$(7) \quad \widehat{\beta}_j = \beta_j + \varepsilon \xi_j, \quad j \in \mathbb{N}^* \text{ et } (\beta_j)_{j \geq 1} \in \ell_2.$$

Ici $(\beta_j)_{j \geq 1}$ représente la suite des coefficients de s dans la base $\{\varphi_j, j \geq 1\}$, c'est-à-dire que $\beta_j = \langle s, \varphi_j \rangle$ pour tout $j \in \mathbb{N}^*$. De plus les variables aléatoires ξ_j sont indépendantes et de même loi $\mathcal{N}(0, 1)$. On peut donc voir le modèle de suite gaussienne comme une extension naturelle du modèle linéaire gaussien fini-dimensionnel (4). L'intérêt du modèle de suite gaussienne provient de ce que si $\mathbb{H} = \mathbb{L}_2[0, 1]$ par exemple et si la base est bien choisie (base de Fourier ou base d'ondelettes) une condition de régularité sur la fonction s du type « s appartient à une boule d'un espace de Sobolev W » se traduit par une condition de sommabilité sur les coefficients de s du type « β appartient à un ellipsoïde ».

Revenons à présent sur le problème de la sélection de variables, à la lumière du nouveau cadre que nous venons d'introduire.

2.3.2. La sélection de variables en dimension infinie

Il est intéressant de noter que le problème de la sélection de variables continue à avoir du sens lorsque les variables sont « construites » par le statisticien afin de construire des modèles approchés de la cible qu'il cherche à estimer. Pour illustrer ce propos, plaçons-nous dans le modèle de bruit blanc gaussien continu. Afin de reconstruire le signal s à partir du signal bruité, une stratégie possible est de considérer une famille de fonctions linéairement indépendantes $\{\varphi_j, j \in \Lambda\}$, où Λ est soit un ensemble fini $\Lambda = \{1, \dots, N\}$ soit $\Lambda = \mathbb{N}^*$. Si l'on songe à la situation où $\{\varphi_j, j \in \Lambda\}$ représente une vaste famille finie d'éléments d'une base d'ondelettes par exemple, rechercher une représentation « parsimonieuse » de s revient à sélectionner un sous-ensemble fini m de Λ (de cardinal sensiblement plus faible que N si Λ est fini) afin de représenter s sur la famille de fonctions $\{\varphi_j, j \in m\}$. On vient de s'intéresser à la sélection de variables *complète*, c'est-à-dire celle pour laquelle on recherche un sous-ensemble de Λ parmi toutes les parties possibles de Λ . Un autre thème d'intérêt dans ce contexte est la sélection de variables dite *ordonnée*. En effet, si la famille $\{\varphi_j, j \in \Lambda\}$ considérée est cette fois la base trigonométrique (prise dans son ordre naturel), on peut être tenté, au vu des propriétés d'approximation bien connus de cette base, de se restreindre à la recherche de sous-ensembles ordonnés, c'est-à-dire du type $[1, D]$, $D \in \mathbb{N}^*$.

On voit émerger une autre idée d'importance. Dès lors que l'on s'intéresse à des modèles approchés dont la dimension peut varier sur une échelle très vaste, on ouvre la possibilité d'approcher un élément s d'un espace de dimension infinie,

c'est-à-dire de faire de l'estimation *non paramétrique*, alors même que tous les modèles approchés que nous manipulons sont eux paramétriques (mais évidemment de dimensions très diverses).

2.4. Sélection de modèle et oracles

Notre but est à présent d'indiquer un formalisme mathématique précis dans lequel poser le problème de la sélection de modèle gaussienne que nous formulons dans le cadre du modèle linéaire gaussien généralisé (3) introduit plus haut. Considérons donc une famille au plus dénombrable $\{S_m, m \in \mathcal{M}\}$, de modèles. Bien que ce ne soit pas strictement nécessaire nous supposons que chaque modèle est un sous-espace vectoriel de dimension finie D_m de \mathbb{H} . Nous verrons que cette restriction qui peut paraître à première vue très forte couvre d'une part un nombre considérable d'exemples et d'autre part simplifie grandement la présentation. Comme dans le cas où $\mathbb{H} = \mathbb{R}^n$, l'estimateur \hat{s}_m du maximum de vraisemblance de s sur le modèle S_m est tout simplement le minimiseur

$$(8) \quad \gamma^\varepsilon(t) = \|t\|^2 - 2\mathbf{X}^\varepsilon(t)$$

sur S_m . Il est aisé de le calculer explicitement. En effet, si $\{\varphi_j, 1 \leq j \leq D_m\}$ est une base orthonormée de S_m il s'exprime sous la forme

$$\hat{s}_m = \sum_{j=1}^{D_m} \mathbf{X}^\varepsilon(\varphi_j) \varphi_j.$$

Comme la projection orthogonale s_m de s sur S_m s'écrit

$$s_m = \sum_{j=1}^{D_m} \langle s, \varphi_j \rangle \varphi_j$$

on en déduit que

$$\varepsilon^{-2} \|\hat{s}_m - s_m\|^2 = \sum_{j=1}^{D_m} W^2(\varphi_j)$$

suit une loi du chi-deux à D_m degrés de liberté. Puisque $\|s - s_m\|^2 = d^2(s, S_m)$, la formule de Pythagore assure donc que le risque quadratique de \hat{s}_m s'écrit

$$\mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right] = D_m \varepsilon^2 + d^2(s, S_m).$$

Cette formule généralise celle obtenue pour le problème de la sélection de variables. Elle reflète parfaitement le paradigme du choix de modèle puis qu'on constate que sa minimisation implique de réaliser un bon équilibre entre le terme de variance $D_m \varepsilon^2$ et le terme de biais $d^2(s, S_m)$. Autrement dit, elle illustre bien l'idée intuitive qu'un bon modèle doit être un reflet convenable, sinon parfait, de la réalité, tout en restant d'une complexité raisonnable. Nous épouserons donc définitivement le point de vue du risque pour juger de la qualité d'un modèle. Chaque modèle S_m est ainsi représenté par le minimiseur \hat{s}_m du critère des moindres carrés γ^ε défini en (8) sur S_m et le « meilleur » modèle est celui qui minimise le risque quadratique $\mathbb{E}_s \left[\|\hat{s}_m - s\|^2 \right]$ lorsque m parcourt \mathcal{M} . Bien entendu le terme de biais dépendant de s , il en est de même d'un tel modèle que nous noterons donc $S_{m(s)}$. Selon

la terminologie introduite par Donoho et Johnstone (voir [15] par exemple), le minimiseur $\widehat{s}_{m(s)}$ du critère des moindres carrés correspondant est appelé *oracle*. Ce n'est évidemment pas un estimateur puisque $m(s)$ est inconnu du statisticien mais son risque quadratique

$$\mathbb{E}_s \left[\|\widehat{s}_{m(s)} - s\|^2 \right] = \inf_{m \in \mathcal{M}} \mathbb{E}_s \left[\|\widehat{s}_m - s\|^2 \right]$$

va servir de référence et de point de comparaison pour juger de la qualité des procédures de sélection qui seront définies à partir des seules données. Il est à noter que la notion de « meilleur » modèle définie ci-dessus diffère sensiblement de celle de modèle « exact ». En effet si s appartient à S_{m_0} il se peut parfaitement que le meilleur modèle soit de dimension inférieure à D_{m_0} et que \widehat{s}_{m_0} ne soit pas un oracle. Nous avons déjà, dans le contexte de la sélection de variables, expliqué les raisons pour lesquelles cette notion de meilleur modèle (au sens du risque) correspond bien à ce qui est recherché en pratique. Il se trouve donc que l'on puisse préférer un modèle approché à un modèle exact. Il est temps à présent d'en venir au coeur du problème, c'est-à-dire à la construction de procédures de sélection \widehat{m} , fondées uniquement sur l'observation, et telles que le risque de l'estimateur $\widehat{s}_{\widehat{m}}$ correspondant soit aussi proche que possible de celui d'un oracle.

2.5. Sélection de modèle par pénalisation

Décrivons tout d'abord formellement la méthode. Il s'agit d'une procédure de moindres carrés pénalisée. On se donne une fonction dite de *pénalité* $\text{pen} : \mathcal{M} \mapsto \mathbb{R}_+$ et on considère \widehat{m} minimisant

$$(9) \quad \gamma^\varepsilon(\widehat{s}_m) + \text{pen}(m)$$

lorsque m parcourt \mathcal{M} . Le modèle et l'estimateur sélectionnés sont alors respectivement définis par $S_{\widehat{m}}$ et $\widehat{s}_{\widehat{m}}$.

Cette méthode remonte au début des années 70 avec les critères dit du C_p de Mallows et d'Akaike (communément appelé AIC). Le problème essentiel est de comprendre quelle fonction de pénalité il convient de choisir. La proposition formulée par Mallows (voir [13] et [24]) dans le contexte du modèle linéaire gaussien fini-dimensionnel (qui, rappelons-le, correspond dans notre formalisme au cas où $\mathbb{H} = \mathbb{R}^n$) est de prendre comme fonction de pénalité $\text{pen}(m) = 2D_m\sigma^2/n$, ou encore $\text{pen}(m) = 2D_m\varepsilon^2$ puisque l'interprétation du modèle linéaire gaussien classique comme un modèle de type (3) passe par le changement de variable $\varepsilon = \sigma/\sqrt{n}$. Il se trouve que cette proposition est strictement identique à celle d'Akaike dans ce contexte de sélection de modèle linéaire au sein du modèle linéaire gaussien à variance σ^2 connue. Voyons sur quelle idée repose cette proposition et surtout en quoi elle est reliée aux notions de meilleur modèle et d'oracle.

2.6. L'heuristique de Mallows

L'idée de base est la suivante. Rappelons que notre souhait est que l'estimateur sélectionné $\widehat{s}_{\widehat{m}}$ imite l'oracle $\widehat{s}_{m(s)}$ et que $m(s)$ minimise le risque quadratique

$$\mathbb{E}_s \left[\|\widehat{s}_m - s\|^2 \right] = D_m\varepsilon^2 + \|s_m - s\|^2.$$

L'idée la plus naturelle serait d'estimer le risque quadratique de \widehat{s}_m puis de minimiser cette estimation. Sa mise en œuvre butte sur la difficulté du problème de l'estimation du terme de biais $\|s_m - s\|^2$. C'est ici qu'il convient d'amender légèrement (mais subtilement) l'idée initiale en remarquant que, d'après la formule de Pythagore, $\|s_m - s\|^2 = \|s\|^2 - \|s_m\|^2$, donc que $m(s)$ minimise également

$$(10) \quad -\|s_m\|^2 + D_m \varepsilon^2.$$

Contrairement au terme de biais, la quantité $\|s_m\|^2$ est facile à estimer. En effet, notons que

$$\|\widehat{s}_m\|^2 = \|s_m\|^2 + \|\widehat{s}_m - s_m\|^2 + 2 \langle s_m, \widehat{s}_m - s_m \rangle,$$

ce qui implique que

$$\mathbb{E}_s \left[\|\widehat{s}_m\|^2 \right] = \|s_m\|^2 + D_m \varepsilon^2.$$

$\|\widehat{s}_m\|^2 - D_m \varepsilon^2$ est donc un estimateur sans biais de $\|s_m\|^2$. Si nous substituons à $\|s_m\|^2$ son estimateur sans biais $\|\widehat{s}_m\|^2 - D_m \varepsilon^2$ dans (10) nous obtenons le critère du C_p de Mallows :

$$-\|\widehat{s}_m\|^2 + 2D_m \varepsilon^2.$$

2.7. Un théorème non asymptotique

L'heuristique de Mallows peut être justifiée (ou corrigée) en contrôlant l'écart entre $\|\widehat{s}_m\|^2$ et son espérance $\|s_m\|^2 + D_m \varepsilon^2$, uniformément en $m \in \mathcal{M}$. L'inégalité de concentration gaussienne est précisément un outil adapté à cet usage. Elle constitue la pierre angulaire de la preuve du théorème suivant (voir [8]) dans lequel nous obtenons simultanément une proposition de forme pour la pénalité et une borne non asymptotique pour le risque de l'estimateur sélectionné dont nous verrons, dans un second temps, qu'elle permet une comparaison effective avec le risque de l'oracle.

Théorème 1. Soit $(x_m)_{m \in \mathcal{M}}$ une famille de nombres positifs ou nuls tels que

$$\sum_{m \in \mathcal{M}} \exp(-x_m) = \Sigma < \infty.$$

Soit $K > 1$. Supposons que pour tout $m \in \mathcal{M}$

$$\text{pen}(m) \geq K \varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m} \right)^2.$$

Si \widehat{m} minimise le critère pénalisé

$$-\|\widehat{s}_m\|^2 + \text{pen}(m),$$

alors l'inégalité suivante est valide

$$(11) \quad \mathbb{E}_s \|\widehat{s}_{\widehat{m}} - s\|^2 \leq C(K) \left\{ \inf_{m \in \mathcal{M}} \left(\|s_m - s\|^2 + \text{pen}(m) \right) + \Sigma \varepsilon^2 \right\},$$

où la constante $C(K)$ ne dépend que de K .

Il est important de comprendre dans quelle mesure le Théorème 1 permet une comparaison effective entre le risque de l'estimateur pénalisé \widehat{s}_m et celui de l'oracle $\inf_{m \in \mathcal{M}} \mathbb{E}_s \|\widehat{s}_m - s\|^2$. Pour ce faire, nous pouvons raisonner de la manière suivante. Rappelant à nouveau que le risque quadratique de \widehat{s}_m s'exprime sous la forme

$$\mathbb{E}_s \|\widehat{s}_m - s\|^2 = \|s_m - s\|^2 + D_m \varepsilon^2,$$

considérons la situation la plus simple dans laquelle pour un certain nombre L , le choix de $x_m = LD_m$ pour tout $m \in \mathcal{M}$ conduit d'après à $\sum_{m \in \mathcal{M}} \exp(-x_m) \leq 1$ (prendre 1 comme borne supérieure n'a rien de magique ici, 2 ferait tout autant l'affaire!). Si nous choisissons $\text{pen}(m) = KD_m (1 + \sqrt{2L})^2 \varepsilon^2$, nous voyons que le membre de droite de la borne de risque (11) est majoré (à un facteur près dépendant de K et de L) par $\inf_{m \in \mathcal{M}} \mathbb{E} \|\widehat{s}_m - s\|^2$. Dans ce cas nous obtenons bien une comparaison avec le risque idéal et l'estimateur sélectionné se comporte, à une constante près, comme un oracle.

Il est également intéressant de noter le lien qu'établit le Théorème 1 entre Statistique et Théorie de l'Approximation. Pour ce faire supposons que le nombre de modèles d'une dimension donnée soit fini et considérons une façon raisonnable de choisir les poids x_m comme fonction de la dimension de chacun des modèles, c'est-à-dire de la forme $x_m = x(D_m)$ avec

$$x(D) = \alpha D + \ln \# \{m \in \mathcal{M}; D_m = D\} \text{ et } \alpha > 0.$$

La pénalité peut alors être choisie de la manière suivante

$$\text{pen}(m) = \text{pen}(D_m) = K\varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x(D_m)} \right)^2$$

et (11) devient

$$\mathbb{E}_s \|\widehat{s}_m - s\|^2 \leq C' \inf_{D \geq 1} \left\{ \inf_{m \in \mathcal{M}, D_m = D} \left(\|s_m - s\|^2 \right) + D\varepsilon^2 \left(1 + \sqrt{2x(D)} \right)^2 \right\},$$

où la constante positive C' ne dépend que de K et de α . À la lecture de cette inégalité on constate que les propriétés d'approximation de $\bigcup_{D_m = D} S_m$ sont absolument cruciales. On peut en particulier espérer un gain substantiel dans le terme de biais grâce à la redondance de modèles de dimension D pour un prix $x(D)$ relativement modeste puisque la dépendance de $x(D)$ en le nombre de modèles de dimension D est logarithmique. C'est typiquement l'effet constaté lorsqu'on utilise une base d'ondelettes pour débruiter un signal.

2.8. Exemples

De nombreux exemples d'applications du Théorème 1 sont développés dans [8]. Contentons-nous de reprendre ici les deux applications mentionnées plus haut : la sélection de variables et la détection de ruptures.

2.8.1. Sélection de variables

Soit $\{\varphi_j, j \in \Lambda\}$ une famille d'éléments linéairement indépendants de \mathbb{H} avec soit $\Lambda = \{1, \dots, N\}$, soit $\Lambda = \mathbb{N}^*$. Pour chaque sous-ensemble m de Λ , nous définissons le sous-espace S_m engendré par $\{\varphi_j, j \in m\}$ et nous considérons une collection \mathcal{M} de parties finies de Λ .

La sélection de variables ordonnées

Nous choisissons dans ce cas pour \mathcal{M} la collection de tous les sous-ensembles de Λ de la forme $\{1, \dots, D\}$. Puisque cette collection ne comporte qu'un seul modèle de dimension donnée D , on peut choisir comme poids $x_m = \alpha D_m$, ce qui conduit à

$$\Sigma = \sum_{m \in \mathcal{M}} e^{-x_m} \leq \sum_{D=1}^{\infty} e^{-\alpha D} = (e^\alpha - 1)^{-1}.$$

Comme α peut être choisi arbitrairement petit, le Théorème 1 autorise de prendre une pénalité de la forme $\text{pen}(m) = K' |m| \varepsilon^2$ avec $K' > 1$. Ce choix conduit en utilisant (11) à une inégalité de comparaison avec le risque de l'oracle de la forme

$$\mathbb{E}_s \|\widehat{S}_m - s\|^2 \leq C' \inf_{m \in \mathcal{M}} \mathbb{E}_s \|\widehat{S}_m - s\|^2,$$

où la constante C' ne dépend que de K' . Par conséquent l'estimateur sélectionné se comporte (à constante près) comme un oracle. De plus, il est possible de prouver que la contrainte $K' > 1$ est optimale au sens suivant. Si $K' < 1$, on peut démontrer que même si $s = 0$, le critère de choix de modèle explose, c'est-à-dire qu'avec une grande probabilité, le modèle sélectionné est systématiquement de grande dimension. Ce comportement a pour corollaire que le risque de l'estimateur sélectionné est d'ordre $N\varepsilon^2$, où N est arbitrairement grand si Λ est infini et $N = |\Lambda|$ sinon, ce qui prouve qu'en aucun cas l'estimateur sélectionné ne peut se comporter comme un oracle.

La sélection de variables complète

Nous considérons le cas où $\Lambda = \{1, \dots, N\}$. Dans le contexte de la sélection de variables complète, \mathcal{M} désigne la collection de tous les sous-ensembles de $\{1, \dots, N\}$. Si nous choisissons comme poids $x_m = |m| \log(N)$, alors

$$\Sigma = \sum_{m \in \mathcal{M}} \exp(-x_m) = \sum_{D \leq N} \binom{N}{D} \exp(-D \log(N)) \leq e$$

et nous pouvons prendre comme pénalité

$$\text{pen}(m) = K |m| \left(1 + \sqrt{2 \log(N)}\right)^2 \varepsilon^2$$

avec $K > 1$. Dans ces conditions (11) devient

$$(12) \quad \mathbb{E}_s \|\widehat{S}_m - s\|^2 \leq C'(K) \inf_{D \geq 1} \left\{ \inf_{m \in \mathcal{M}, D_m=D} \left(\|s_m - s\|^2 \right) + D \log(N) \varepsilon^2 \right\},$$

où $C'(K)$ ne dépend que de K . Nous constatons que le facteur supplémentaire $\log(N)$ est un prix relativement modeste à payer comparé au gain potentiel dans le terme de biais que procure la redondance de modèles de dimension identique. Il est intéressant de noter qu'aucune hypothèse d'orthogonalité entre les éléments $\{\varphi_j, j \leq N\}$ n'est nécessaire pour obtenir ce résultat. Si toutefois le système est

orthonormé, l'estimateur par pénalisation ci-dessus peut être explicitement calculé et l'on retrouve l'estimateur par seuillage introduit par Donoho et Johnstone dans le cadre du modèle de bruit blanc (voir [15]). En effet, lorsque pour un certain nombre $T > 0$

$$\text{pen}(m) = T^2 |m|,$$

et

$$\widehat{\beta}_j = \mathbf{X}^\varepsilon(\varphi_j), \text{ pour } 1 \leq j \leq N,$$

l'estimateur des moindres carrés sur le modèle S_m engendré par φ_j , $j \in m$ a pour expression

$$\widehat{s}_m = \sum_{j \in m} \widehat{\beta}_j \varphi_j.$$

Dans ces conditions, le critère pénalisé s'écrit

$$\text{crit}(m) = -\|\widehat{s}_m\|^2 + \text{pen}(m) = \sum_{j \in m} \left(-\widehat{\beta}_j^2 + T^2 \right).$$

Par conséquent l'ensemble \widehat{m} minimisant le critère $\text{crit}(m)$ lorsque m parcourt la collection de tous les sous-ensembles de Λ vaut exactement

$$\widehat{m} = \left\{ j \in \Lambda, -\widehat{\beta}_j^2 + T^2 \leq 0 \right\}.$$

En d'autres termes

$$\widehat{s}_{\widehat{m}} = \sum_{j=1}^N \widehat{\beta}_j \mathbb{1}_{|\widehat{\beta}_j| \geq T} \varphi_j$$

qui est bien un estimateur par seuillage, le seuil T fourni par notre Théorème étant finalement de la forme $T = \sqrt{K} \left(1 + \sqrt{2 \log(N)} \right) \varepsilon$. On peut à nouveau prouver que la contrainte $K > 1$ est fine.

Notons que les calculs précédents sur les poids peuvent être légèrement améliorés. Plus précisément il est possible de remplacer le facteur logarithmique $\log(N)$ ci-dessus par $\log(N/|m|)$. En effet rappelons la majoration classique suivante pour le coefficient binomial

$$(13) \quad \ln \binom{N}{D} \leq D \ln \left(\frac{eN}{D} \right).$$

Un choix de x_m de la forme $x_m = |m| L(|m|)$ implique que

$$\begin{aligned} \Sigma &= \sum_{D \leq N} \binom{N}{D} \exp[-DL(D)] \leq \sum_{D \leq N} \left(\frac{eN}{D} \right)^D \exp[-DL(D)] \\ &\leq \sum_{D \leq N} \exp \left[-D \left(L(D) - 1 - \ln \left(\frac{N}{D} \right) \right) \right]. \end{aligned}$$

Si nous fixons $L(D) = 1 + \theta + \ln(N/D)$ avec $\theta > 0$ nous obtenons que $\Sigma \leq \sum_{D=0}^{\infty} e^{-D\theta} = [1 - e^{-\theta}]^{-1}$. Avec $\theta = \ln 2$, le Théorème 1 nous autorise à prendre comme pénalité

$$\text{pen}(m) = K\varepsilon^2 |m| \left(1 + \sqrt{2(1 + \ln(2N/|m|))} \right)^2$$

avec $K > 1$ et nous en déduisons la borne de risque suivante pour l'estimateur pénalisé correspondant

$$(14) \quad \mathbb{E}_s \left[\|\widehat{s}_m - s\|^2 \right] \leq C'' \inf_{1 \leq D \leq N} \{ b_D^2(s) + D(1 + \ln(N/D)) \varepsilon^2 \},$$

où $b_D^2(s) = \inf_{m \in \mathcal{M}, |m|=D} (\|s_m - s\|^2)$. Cette inégalité améliore légèrement (12).

Par ailleurs, l'estimateur pénalisé reste facilement calculable lorsque le système $\{\varphi_j\}_{j \leq N}$ est orthonormé. En effet

$$\begin{aligned} & \inf_{m \in \mathcal{M}} \left\{ - \sum_{j \in m} \widehat{\beta}_j^2 + K\varepsilon^2 |m| \left(1 + \sqrt{2L(|m|)} \right)^2 \right\} \\ &= \inf_{D \leq N} \left\{ - \sup_{\{m \mid |m|=D\}} \sum_{j \in m} \widehat{\beta}_j^2 + K\varepsilon^2 D \left(1 + \sqrt{2L(D)} \right)^2 \right\} \\ &= \inf_{D \leq N} \left\{ - \sum_{j=1}^D \widehat{\beta}_{(j)}^2 + K\varepsilon^2 D \left(1 + \sqrt{2L(D)} \right)^2 \right\} \end{aligned}$$

où $\widehat{\beta}_{(1)}^2 \geq \dots \geq \widehat{\beta}_{(N)}^2$ désignent les carrés des coefficients estimés $\{\widehat{\beta}_j, j \leq N\}$, rangés par ordre décroissant. Nous constatons que la minimisation du critère pénalisé revient à sélectionner une valeur \widehat{D} de D minimisant

$$- \sum_{j=1}^D \widehat{\beta}_{(j)}^2 + K\varepsilon^2 D \left(1 + \sqrt{2L(D)} \right)^2$$

et finalement à exprimer l'estimateur pénalisé sous la forme

$$(15) \quad \widehat{s}_m = \sum_{j=1}^{\widehat{D}} \widehat{\beta}_{(j)} \varphi_{(j)}.$$

La performance de cet estimateur est en un sens optimale. On peut démontrer en effet que la borne de risque (14) est optimale au sens dit *du minimax* sur l'ensemble $\mathbb{S}_D = \bigcup_{|m|=D} S_m$, $D \leq N$. Autrement dit, il existe une constante absolue strictement positive κ telle que quel soit l'estimateur \widetilde{s} de s

$$\sup_{s \in \mathbb{S}_D} \mathbb{E}_s \|\widetilde{s} - s\|^2 \geq \kappa D (1 + \ln(N/D)) \varepsilon^2.$$

2.8.2. Détection de ruptures multiples

Considérons le problème de détection de ruptures sur la moyenne décrit ci-dessus dans le cadre du modèle de bruit blanc discret. Le signal bruité observé est donc de la forme

$$X_j = s(j/n) + \sigma \xi_j, \quad 1 \leq j \leq n,$$

où les erreurs ξ_j sont indépendantes et de même loi normale $\mathcal{N}(0, 1)$. Définissons l'espace vectoriel S_m des fonctions constantes par morceaux sur la partition m . Détecter les ruptures revient à sélectionner un modèle au sein de la famille

$\{S_m\}_{m \in \mathcal{M}}$, où \mathcal{M} désigne la collection de toutes les partitions possibles de $[0, 1]$ par des intervalles dont les extrémités se situent sur la grille $\{j/n, 0 \leq j \leq n\}$. Puisque le nombre de modèles de dimension D , c'est-à-dire le nombre de partitions à D morceaux est égal à $\binom{n-1}{D-1}$, cette collection de modèles possède des propriétés combinatoires analogues à celle de la collection de modèles correspondant à la sélection de variables complète au sein de $N = n - 1$ variables. Concernant le choix de la pénalité et les bornes de risque qui en résultent, les mêmes considérations que dans le cas de la sélection de variables complète étudié ci-dessus restent donc valides.

2.9. Estimation adaptative et Théorie de l'Approximation

Le principal avantage de la borne de risque fournie par le Théorème 1 est qu'elle vaut pour toute valeur de s . Son principal désavantage est qu'elle ne compare le risque de l'estimateur sélectionné qu'avec celui des estimateurs appartenant à la collection $\{\hat{s}_m\}_{m \in \mathcal{M}}$ dont il est issu mais pas avec celui d'une procédure d'estimation quelconque. Exprimé en des termes familiers on peut donc craindre d'avoir sélectionné un estimateur « borgne au royaume des aveugles ». Bien entendu si nous souhaitons effectuer une comparaison de risque avec un estimateur quelconque, il nous faudra abandonner l'idée de la réaliser ponctuellement (c'est-à-dire pour toute valeur de s) car il est clair qu'un estimateur constamment égal à s_0 est parfait si $s = s_0$ puisque de risque nul même s'il est par ailleurs stupide. Il convient donc pour donner du sens à une telle comparaison de s'intéresser aux performances des estimateurs en plusieurs points simultanément. Une approche classique est de considérer le risque maximal sur certains sous-ensembles. C'est le point de vue dit *minimax*. Le risque minimax sur un sous-ensemble \mathcal{T} de \mathbb{H} est ainsi défini par

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

où l'infimum porte sur l'ensemble de tous les estimateurs possibles de s . La performance d'un estimateur donné \hat{s} , peut alors être mesurée par le rapport

$$\frac{\sup_{s \in \mathcal{T}} \mathbb{E}_s \left[\|\hat{s} - s\|^2 \right]}{R_M(\mathcal{T}, \varepsilon)}.$$

Si ce rapport est borné indépendamment des valeurs de ε , \hat{s} sera dit *approximativement minimax* sur \mathcal{T} . Pour illustrer notre propos, revenons au cas du modèle de bruit blanc continu en dimension 1, pour lequel $\mathbb{H} = \mathbb{L}_2[0, 1]$. Un exemple typique de choix pour l'ensemble \mathcal{T} est une boule d'un certain espace de Banach de fonctions régulières comme par exemple, une boule de rayon R d'un espace de Sobolev de régularité α . Si nous notons $W^\alpha(R)$ une telle boule, un sérieux désavantage de l'approche minimax est qu'un estimateur approximativement minimax peut très bien dépendre de α et de R , quantités inconnues en pratique. Il est donc préférable d'exiger d'une procédure d'estimation donnée, qu'elle soit approximativement minimax sur toute une famille d'ensembles \mathcal{T} simultanément (dans notre exemple toutes les boules $W^\alpha(R)$ lorsque α et R varient). C'est précisément le point de vue de l'adaptation au sens du minimax qui a fait l'objet de très nombreux travaux en statistique depuis le début des années 90. Mentionnons les nombreuses contributions de Donoho, Johnstone, Kerkycharian et Picard qui ont étudié les propriétés

d'adaptation des estimateurs par seuillage de coefficients d'ondelettes sur des familles de boules d'espaces de Besov (voir [14] pour un panorama). D'une manière ou d'une autre, toutes les constructions d'estimateurs adaptatifs reposent sur une procédure de sélection (ou d'agrégation) d'une famille d'estimateurs préliminaires qui peuvent bien entendu différer de la sélection par pénalisation comme c'est le cas par exemple de la méthode de Lepskii (voir [21] et [22]). Néanmoins, le principe est toujours le même. Dès lors qu'une procédure de sélection d'estimateurs se comporte comme un oracle, il suffit de vérifier grâce à des arguments de Théorie de l'Approximation que l'oracle lui-même est approximativement minimax sur la famille d'ensembles $\{\mathcal{T}_\theta\}_{\theta \in \Theta}$ d'intérêt pour conclure à l'adaptativité de l'estimateur sélectionné. Pour en revenir à la sélection de modèle proprement dite, il convient que la famille de modèles $\{S_m\}_{m \in \mathcal{M}}$ possède de bonnes qualités d'approximation vis-à-vis de la famille d'ensembles $\{\mathcal{T}_\theta\}_{\theta \in \Theta}$. Les performances de l'estimateur pénalisé \tilde{s} , sont ensuite évaluées pour chaque $\theta \in \Theta$ par

$$\sup_{s \in \mathcal{T}_\theta} \inf_{m \in \mathcal{M}} \left(\|s_m - s\|^2 + \text{pen}(m) \right).$$

Illustrons à présent ce principe.

2.9.1. Exemple : adaptation sur des ellipsoïdes

En guise d'illustration, supposons à nouveau que \mathbb{H} soit de dimension infinie et considérons une base orthonormée $\{\varphi_j, j \geq 1\}$ de \mathbb{H} . Considérons pour chaque suite $(\theta_j)_{j \geq 1}$ décroissant vers 0, l'ellipsoïde de \mathbb{H} défini par

$$\mathcal{E}_2(\theta) = \left\{ s \in \mathbb{H}, \sum_{j \geq 1} \left(\frac{\langle s, \varphi_j \rangle}{\theta_j} \right)^2 \leq 1 \right\}.$$

Étudions les propriétés d'adaptation à la famille des ellipsoïdes ci-dessus de l'estimateur associé à la procédure de sélection de variable ordonnée décrite au paragraphe 2.8.1. Rappelons qu'en pareil cas

$$\mathcal{M} = \{[1, D], D \geq 1\}.$$

De plus si $m = [1, D]$, S_m désigne le sous-espace engendré par $\{\varphi_j, 1 \leq j \leq D\}$ et la pénalité en m s'écrit $\text{pen}(m) = K'D\varepsilon^2$ avec $K' > 1$. Ce choix conduit en utilisant (11) à une inégalité de comparaison avec le risque de l'oracle de la forme

$$\mathbb{E}_s[\|\hat{s}_m - s\|^2] \leq C' \inf_{m \in \mathcal{M}} \mathbb{E}[\|\hat{s}_m - s\|^2] = C' \inf_{D \geq 1} \left(\sum_{j > D} \langle s, \varphi_j \rangle^2 + D\varepsilon^2 \right).$$

À présent si $s \in \mathcal{E}_2(\theta)$, le terme de biais dans l'inégalité ci-dessus se contrôle aisément

$$\sum_{j > D} \langle s, \varphi_j \rangle^2 = \sum_{j > D} \frac{\langle s, \varphi_j \rangle^2}{\theta_j^2} \theta_j^2 \leq \theta_{D+1}^2,$$

d'où

$$\sup_{s \in \mathcal{E}_2(\theta)} \mathbb{E}_s[\|\hat{s}_m - s\|^2] \leq C' \inf_{D \geq 1} (\theta_{D+1}^2 + D\varepsilon^2).$$

Or, pourvu que $\theta_1 \geq \varepsilon$, il est possible de prouver (voir par exemple [26]) que le risque minimax sur l'ellipsoïde $\mathcal{E}_2(\theta)$ est effectivement minoré, à une constante absolue multiplicative près par $\inf_{D \geq 1} (\theta_{D+1}^2 + D\varepsilon^2)$, ce qui démontre que \widehat{m} est approximativement minimax sur chacun des ellipsoïdes $\mathcal{E}_2(\theta)$, et ce quelle que soit la suite $\theta = (\theta_j)_{j \geq 1}$ décroissant vers 0, telle que $\theta_1 \geq \varepsilon$. Si $\mathbb{H} = \mathbb{L}_2[0, 1]$ et si $\{\varphi_j, j \geq 1\}$ désigne la base de Fourier, la propriété d'adaptation sur les ellipsoïdes ci-dessus implique que l'estimateur pénalisé est adaptatif sur la collection de toutes les boules de Sobolev $W^\alpha(R)$, avec $\alpha > 0$ et $R \geq \varepsilon$.

Cet exemple est représentatif de nombreux autres du même type (voir [26]) qui tendent en définitive à analyser les performances de diverses stratégies de sélection de variables lorsque celles-ci sont élaborées avec pour objectif d'approcher au mieux l'objet à estimer. On peut pour ce faire sélectionner des variables au sein d'une même base mais aussi profiter de la souplesse offerte par un résultat comme le Théorème 1 pour utiliser des variables provenant de différentes bases. Autrement dit rien n'oblige a priori à travailler avec une seule et même base. Afin de résoudre un problème de traitement du signal ou d'image donné, on peut donc tenter de choisir la meilleure représentation possible parmi plusieurs disponibles. On peut le faire globalement ou même à chaque niveau de résolution si on travaille avec des représentations multi-échelles. On trouvera dans les travaux de Stéphane Mallat et de ses collaborateurs plusieurs méthodes et résultats allant dans cette direction (voir en particulier [23] et [20]).

2.10. Conclusions

Les points saillants suivants émergent de l'étude de la sélection de modèle gaussienne.

- Le C_p de Mallows peut sous-pénaliser et il doit être corrigé lorsque le nombre de modèles de même dimension est trop élevé.
- On peut utiliser la sélection de modèle comme outil d'estimation non paramétrique et choisir des listes de modèles inspirées par la Théorie de l'Approximation afin de produire des estimateurs adaptatifs.
- La condition $K > 1$ apparaissant dans l'énoncé du Théorème 1 est fine.
- Quelle pénalité doit être au bout du compte recommandée ? On peut tenter d'optimiser la borne de risque fournie par le Théorème 1. C'est le travail effectué dans [9], dont la conclusion est que $K = 2$ est en général un bon choix.
- En pratique, le niveau de bruit est inconnu mais on peut finalement retenir de la théorie la formule suivante : pénalité "*optimale*" = $2 \times$ pénalité "*minimale*". Le point clef pour tirer profit de cette remarque est que la pénalité minimale peut être devinée à partir des données grâce au phénomène d'explosion : tant que la pénalité n'est pas assez lourde, le critère pénalisé choisit des modèles de très grande dimension. Une fois la pénalité minimale estimée, il reste à la multiplier par 2 pour obtenir la pénalité (présumée optimale) désirée. Cette stratégie fournit donc une pénalité dépendant des données qui ne nécessite pas la connaissance a priori du niveau de bruit ε (voir [17] pour les détails d'implémentation de cette méthode).

3. Extension au cas non gaussien

Revenons au problème plus général d'estimation d'une quantité $s \in \mathcal{S}$ liée à la loi de probabilité inconnue d'une observation X . Outre le cadre du modèle linéaire gaussien généralisé décrit précédemment, les cadres typiques auxquels on peut penser sont les suivants :

– estimation de la densité où $X = (X_1, \dots, X_n)$ les X_i , $1 \leq i \leq n$ étant des variables aléatoires indépendantes et de même loi admettant la densité s par rapport à une mesure dominante μ .

– modèle de régression avec apprentissage, où les variables aléatoires $X_i = (\xi_i, Y_i)$ sont des copies indépendantes d'un couple (ξ, Y) . La réponse Y à la variable explicative ξ est supposée de carré intégrable. La fonction de régression s est définie par $s(x) = \mathbb{E}_s [Y \mid \xi = x]$.

La méthode du maximum de vraisemblance possède une généralisation naturelle appelée estimation par minimum de contraste.

3.1. Sélection par minimum de contraste pénalisé

Considérons un critère $\gamma(\mathbf{X}, \cdot)$ tel que

$$t \mapsto \mathbb{E}_s [\gamma(\mathbf{X}, t)]$$

atteigne un minimum au point s sur \mathcal{S} . Un tel critère est appelé *contraste*. On peut associer à ce contraste la fonction de perte naturelle ℓ définie par

$$(16) \quad \ell(s, t) = \mathbb{E}_s [\gamma(\mathbf{X}, t)] - \mathbb{E}_s [\gamma(\mathbf{X}, s)] \geq 0.$$

Les exemples les plus connus de contraste sont l'opposé de la log-vraisemblance d'une part et le contraste des moindres carrés d'autre part. Voici comment les définir dans les cadres de l'estimation de la densité et de la régression.

– Densité

On observe $X = (X_1, \dots, X_n)$, où X_1, \dots, X_n sont des variables aléatoires indépendantes de même loi admettant pour densité s par rapport à une mesure dominante μ . Le choix de

$$\gamma(\mathbf{X}, t) = -\frac{1}{n} \sum_{i=1}^n \log(t(X_i))$$

conduit à la fonction de perte

$$\ell(s, t) = K(s, t).$$

$K(s, t)$ désigne l'information de Kullback-Leibler de la probabilité $s\mu$ relativement à $t\mu$, définie par

$$K(s, t) = \int s \log\left(\frac{s}{t}\right) d\mu$$

si $s\mu$ est absolument continue par rapport à $t\mu$ et $K(s, t) = +\infty$ sinon. Supposant cette fois que $s \in \mathbb{L}_2(\mu)$, il est possible de définir le critère des moindres carrés par

$$\gamma(\mathbf{X}, t) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^n t(X_i),$$

où $\|\cdot\|$ note la norme dans $\mathbb{L}_2(\mu)$. Le fonction de perte correspondante vaut dans ce cas

$$\ell(s, t) = \|s - t\|^2,$$

pour tout $t \in \mathbb{L}_2(\mu)$.

– Régression

On observe $\mathbf{X} = (\xi_1, Y_1), \dots, (\xi_n, Y_n)$ indépendantes et de même loi. Soit μ la loi commune aux variables ξ . Le contraste des moindres carrés se définit alors par

$$\gamma(\mathbf{X}, t) = \frac{1}{n} \sum_{i=1}^n (Y_i - t(\xi_i))^2$$

et la fonction de perte qui lui est associée vaut tout simplement

$$\ell(s, t) = \|s - t\|^2,$$

pour tout $t \in \mathbb{L}_2(\mu)$.

Si nous considérons donc un contraste $\gamma(\mathbf{X}, \cdot)$ et un modèle $S \subseteq \mathcal{S}$, un *estimateur de minimum de contraste* de s sur S est un minimiseur de $t \mapsto \gamma(\mathbf{X}, t)$ sur S . Étant donné une collection au plus dénombrable de modèles $(S_m)_{m \in \mathcal{M}}$, on peut associer à chaque modèle S_m un estimateur du minimum de contraste \hat{s}_m sur S_m . Le problème de sélection de modèle devient alors un problème de sélection d'estimateurs. La notion d'oracle introduite plus haut pour la sélection de modèle gaussienne s'étend aisément. Plus précisément, si nous considérons $m(s)$ minimisant $m \rightarrow \mathbb{E}_s[\ell(s, \hat{s}_m)]$ sur \mathcal{M} , $\hat{s}_{m(s)}$ est appelé oracle. $\hat{s}_{m(s)}$ représente donc une sélection idéale puisque son risque relativement à la fonction de perte naturelle ℓ est minimum.

La méthode de sélection par minimum de contraste pénalisé consiste à se donner une fonction de pénalité $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}_+$ et à sélectionner \hat{m} minimisant le critère

$$\gamma(\mathbf{X}, \hat{s}_m) + \text{pen}(m)$$

sur \mathcal{M} . L'estimateur puis le modèle sélectionnés sont respectivement définis par $\hat{s}_{\hat{m}}$ et $S_{\hat{m}}$.

3.2. Le rôle des inégalités de concentration

Indiquons à présent quelques éléments sur la façon dont peuvent se démontrer les résultats non asymptotiques sur les critères de choix de modèle par minimisation d'un critère pénalisé dont le Théorème 1 peut être considéré comme un prototype. Introduisons donc le processus centré

$$\bar{\gamma}(\mathbf{X}, t) = \gamma(\mathbf{X}, t) - \mathbb{E}_s[\gamma(\mathbf{X}, t)], \quad t \in \mathcal{S}.$$

Par définition un estimateur pénalisé $\hat{s}_{\hat{m}}$ satisfait pour tout $m \in \mathcal{M}$ et tout point $s_m \in S_m$

$$\gamma(\mathbf{X}, \hat{s}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma(\mathbf{X}, \hat{s}_m) + \text{pen}(m) \leq \gamma(\mathbf{X}, s_m) + \text{pen}(m),$$

ou de façon équivalente en substituant $\bar{\gamma}(\mathbf{X}, \cdot) + \mathbb{E}_s[\gamma(\mathbf{X}, \cdot)]$ à $\gamma(\mathbf{X}, \cdot)$,

$$\bar{\gamma}(\mathbf{X}, \hat{s}_{\hat{m}}) + \text{pen}(\hat{m}) + \mathbb{E}_s[\gamma(\mathbf{X}, \hat{s}_{\hat{m}})] \leq \bar{\gamma}(\mathbf{X}, s_m) + \text{pen}(m) + \mathbb{E}_s[\gamma(\mathbf{X}, s_m)].$$

Soustrayant $\mathbb{E}_s [\gamma(\mathbf{X}, s)]$ à chaque membre de l'inégalité ci-dessus, nous obtenons l'importante relation suivante

$$\begin{aligned} \ell(s, \widehat{s}_m) &\leq \ell(s, s_m) + \text{pen}(m) \\ &\quad + \overline{\gamma}(\mathbf{X}, s_m) - \overline{\gamma}(\mathbf{X}, \widehat{s}_m) - \text{pen}(\widehat{m}) \end{aligned}$$

On voit donc que la fonction de pénalité doit simultanément être choisie

- suffisamment grande pour contrer les fluctuations de $\overline{\gamma}(\mathbf{X}, s_m) - \overline{\gamma}(\mathbf{X}, \widehat{s}_m)$,
- mais pas trop non plus car idéalement on souhaiterait que $\ell(s, s_m) + \text{pen}(m) \leq \mathbb{E}_s [\ell(s, \widehat{s}_m)]$.

Par conséquent le secret d'une calibration convenable de la pénalité réside dans notre capacité à évaluer finement les fluctuations de $\overline{\gamma}(\mathbf{X}, s_m) - \overline{\gamma}(\mathbf{X}, \widehat{s}_m)$. C'est précisément ce que procure l'utilisation des inégalités de concentration combinée avec un procédé de localisation. Il s'agit en effet d'obtenir des estimées qui sont sensibles au fait que les fluctuations de $\overline{\gamma}(\mathbf{X}, t) - \overline{\gamma}(\mathbf{X}, u)$ sont d'autant plus faibles que t est proche de u . Il convient donc d'obtenir pour chaque $m' \in \mathcal{M}$, un bon contrôle de

$$\sup_{t \in S_{m'}} \frac{\overline{\gamma}(\mathbf{X}, s_{m'}) - \overline{\gamma}(\mathbf{X}, t)}{\omega(s_{m'}, t)},$$

pour une pondération convenable $\omega(s_{m'}, t)$ permettant de facturer la proximité entre $s_{m'}$ et t .

L'inégalité de concentration gaussienne (voir [12]) et l'inégalité de Talagrand pour les processus empiriques (voir [30]) constituent les prototypes des inégalités utiles pour réaliser le contrôle ci-dessus. Rappelons un énoncé de l'inégalité de Talagrand. Étant donné X_1, \dots, X_n indépendantes et de même loi, et une classe \mathcal{F} (au plus dénombrable) de fonctions centrées en espérance et uniformément bornées par 1, on définit $Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)$ et $v = \mathbb{E} [\sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2(X_i)]$. Alors, pour tout x positif, excepté sur un événement de probabilité moindre que $K \exp(-x)$ l'inégalité suivante est valide

$$Z \leq \mathbb{E}[Z] + \sqrt{2v\kappa x} + cx$$

où K , κ et c sont des constantes universelles. Si on suit l'approche de Ledoux fondée sur des inégalités de type Sobolev logarithmiques (voir [18] et [19]), on peut expliciter la valeur des constantes et prendre $K = 1$, $\kappa = 4$ et $c = 2$ (voir [25]). Si on modifie la définition du facteur de variance v en posant $v = 2\mathbb{E}[Z] + n \sup_{f \in \mathcal{F}} \mathbb{E}[f^2(X_1)]$, il est même possible d'obtenir les valeurs optimales $\kappa = 1$ et $c = 1/3$ (voir [10]) pour les constantes. Comme indiqué pour la première fois dans [7] dans le contexte de l'estimation de densité par moindres carrés pénalisés, l'inégalité de Talagrand pour les processus empiriques permet de prouver des théorèmes de sélection de modèle, analogues au Théorème 1 pour la sélection de modèle gaussienne. Parmi les travaux s'appuyant sur cette même idée, citons [11] pour la log-vraisemblance pénalisée sur des log-splines, [2] pour des critères de type Mallows dans le contexte de la régression avec des erreurs non gaussiennes (voir également [4] si les erreurs sont faiblement dépendantes) et enfin [27] pour l'estimation de l'intensité d'un processus de Poisson inhomogène par sélection de modèle.

4. Références

- [1] AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In P.N. Petrov and F. Csaki, editors, *Proceedings 2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973.
- [2] BARAUD, Y. Model selection for regression on a fixed design. *Probability Theory and Related Fields* **117**, n°4 467-493 (2000).
- [3] BAHADUR, R.R. Examples of inconsistency of maximum likelihood estimates. *Sankhya Ser.A* **20**, 207-210 (1958).
- [4] BARAUD, Y., COMTE, F. and VIENNET, G. Model selection for (auto-)regression with dependent data. *ESAIM : Probability and Statistics* **5** 33–49 (2001) <http://www.emath.fr/ps/>.
- [5] BARRON, A.R., BIRGÉ, L., MASSART, P. Risk bounds for model selection via penalization. *Probab. Th. Rel. Fields.* **113**, 301-415 (1999).
- [6] BIRGÉ, L. and MASSART, P. Rates of convergence for minimum contrast estimators. *Probab. Th. Relat. Fields* **97**, 113-150 (1993).
- [7] BIRGÉ, L. and MASSART, P. From model selection to adaptive estimation. In *Festschrift for Lucien Lecam : Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.), 55-87 (1997) Springer-Verlag, New-York.
- [8] BIRGÉ, L. and MASSART, P. Gaussian model selection. *Journal of the European Mathematical Society*, n°3, 203-268 (2001).
- [9] BIRGÉ, L., MASSART, P. Minimal penalties for Gaussian model selection. *Probab. Th. Relat. Fields* **138**, 33-73 (2007).
- [10] BOUSQUET, O. A Bennett concentration inequality and its application to suprema of empirical processes. *C.R. Math. Acad. Sci. Paris* **334** n°6, 495-500 (2002).
- [11] CASTELLAN, G. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory* **49** n°8, 2052-2060 (2003).
- [12] CIREL'SON, B.S., IBRAGIMOV, I.A. and SUDAKOV, V.N. Norm of Gaussian sample function. In *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*, Lecture Notes in Mathematics **550** 20-41 (1976) Springer-Verlag, Berlin.
- [13] DANIEL, C. and WOOD, F.S. *Fitting Equations to Data*. Wiley, New York (1971).
- [14] DONOHO, D.L. and JOHNSTONE, I.M., KERKYACHARIAN, G. and PICARD, D. Wavelet shrinkage :Asymptopia? *J. R. Statist. Soc. B* **57**, 301-369 (1995).
- [15] DONOHO, D.L. and JOHNSTONE, I.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455 (1994).
- [16] DUDLEY, R.M. *Uniform Central Limit Theorems*. Cambridge Studies in advanced mathematics **63**, Cambridge University Press (1999).
- [17] LEBARBIER, E. Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, **85** n°4, 717-736 (2005).
- [18] LEDOUX, M. On Talagrand deviation inequalities for product measures. *ESAIM : Probability and Statistics* **1**, 63-87 (1996) <http://www.emath.fr/ps/>.
- [19] LEDOUX, M. *The concentration of measure phenomenon*. Mathematical Surveys and Monographs **89**, American Mathematical Society.
- [20] LE PENNEC, E. et MALLAT, S. Sparse Geometric Image Representation with Bandelets. *IEEE Trans. on Image Processing*, **14**, n°4, 423-438, (2005).
- [21] LEPSKII, O.V. On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **36**, 454-466 (1990).
- [22] LEPSKII, O.V. Asymptotically minimax adaptive estimation I : Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36**, 682-697 (1991).
- [23] MALLAT, S. *A Wavelet Tour of Signal Processing*. Academic Press, 1999
- [24] MALLOWS, C.L. Some comments on C_p . *Technometrics* **15**, 661-675 (1973).
- [25] MASSART, P. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. of Probability*. **28**, n°2, 863-884 (2000).
- [26] MASSART, P. *Concentration inequalities and model selection*. In *Lectures on Probability Theory and Statistics, École d'Eté de Probabilités de St-Flour XXXIII-2003* (J. Picard, ed.). Lecture notes in Mathematics n° 1896 (2007) Springer, Berlin.

- [27] REYNAUD-BOURET, P. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Relat. Fields* **126**, n°1, 103-153 (2003).
- [28] SCHWARTZ, G. Estimating the dimension of a model. *Ann. of Statistics* **6**, 461-464 (1978).
- [29] TALAGRAND, M. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.* **81** 73-205 (1995).
- [30] TALAGRAND, M. New concentration inequalities in product spaces. *Invent. Math.* **126**, 505-563 (1996).
- [31] VAN DER VAART, A. *Asymptotic statistics*. Cambridge University Press (1998).
- [32] VAN DER VAART, A. and WELLNER J. *Weak Convergence and Empirical Processes*. Springer, New York (1996).