SÉLECTION DE MODÈLE : DE LA THÉORIE À LA PRATIQUE

P. MASSART*

RÉSUMÉ

Pour choisir un modèle statistique à partir des données, une méthode devenue classique depuis les travaux précurseurs d'Akaike dans les années 70 consiste à optimiser un critère empirique pénalisé, tel que la log-vraisemblance pénalisée. Dans bon nombre de problèmes de sélection de modèle tels que la sélection de variables ou la détection de ruptures multiples par exemple, il est souhaitable de laisser croitre la taille des modèles ou encore le nombre de modèles d'une dimension donnée avec le nombre d'observations. Une théorie non asymptotique de la sélection de modèles a donc émergé durant ces dix dernières années qui vise à prendre en compte ce type de situations. L'enjeu central aussi bien sur le plan théorique que pratique est de comprendre comment pénaliser un critère de type log-vraisemblance de façon à garantir une performance de sélection optimale. La théorie non asymptotique donne des indications sur la structure des pénalités qu'il convient d'utiliser mais n'est parfois pas suffisamment précise pour arbitrer la valeur de certaines constantes qui restent donc à calibrer au moment d'implémenter effectivement ce type de critères. Ces constantes peuvent être inconnues pour des raisons diverses. Il peut s'agir d'une faiblesse de la théorie qui garantit l'existence d'une constante absolue mais sans en donner la valeur numérique. Le problème peut être également de nature plus profondément statistique lorsque cette constante dépend objectivement de la loi inconnue des observations. Notre propos est ici de promouvoir une méthode de calibration de pénalité à partir des données. Cette méthode est en partie fondée sur des résultats théoriques établis et en partie sur une heuristique permettant de l'extrapoler à d'autres cadres que le cadre strict dans lequel la théorie permet de la valider.

Mots-clés : Détection de ruptures, Inégalités de concentration, Pénalisation, Processus empiriques, Sélection de modèle, Sélection de variables.

ABSTRACT

Since the seminal work of Akaike in the early seventies, optimizing some penalized empirical criterion such as the penalized log-likelihood has become a classical solution to the problem of choosing a proper statistical model from the data. For many model selection problems such as multiple change-point detection and variable selection for instance, it is desirable to let the dimension or the number of models of a given dimension grow with the sample size. A non asymptotic theory for model selection has therefore emerged during these last ten years in order to take this type of situations into account. The main issue both from a practical and

* Laboratoire de Mathématiques, UMR 8628 et équipe-projet SELECT, Université de Paris-Sud, Centre Scientifique d'Orsay, 91405 Orsay Cedex.
E-mail : pascal.massart@math.u-psud.fr

a theoretical view point is to understand how to penalize an empirical criterion such as the log-likelihood in order to get some optimal selection procedure. Asymptotic theory provides some useful indications on the shape of the penalty but it often leaves to the user the choice of numerical constants. The optimal value for these constants is generally unknown. In some situations theory is indeed not sharp enough to lead to explicit values. In some other cases, the problem is more of a statistical nature since according to the theory, the optimal value should depend on the unknown distribution of the observations. Our purpose here is to promote some data-driven method to calibrate the penalty. This method is partly based on preliminary theoretical results that we shall recall and partly founded on some heuristics that we intend to explain.

Keywords : Change point detection, Concentration inequalities, Empirical processes, Model Selection, Penalization, Variable Selection.

1. Introduction

La sélection de modèle est un thème classique de la statistique. 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], [17] et [25]). Le critère d'Akaike par exemple propose de pénaliser la log-vraisemblance maximale sur un modèle paramétrique par le nombre de paramètres du modèle. L'heuristique de ce critère repose sur une approximation asymptotique de la log-vraisemblance. Ce résultat classique connu sous le nom de théorème de Wilks est une conséquence de la normalité asymptotique du maximum de vraisemblance. Il précise que sous des conditions de régularité convenables sur le modèle paramétrique à D paramètres considéré, l'écart entre la log-vraisemblance fondée sur n observations indépendantes et équidistribuées prise en son maximum et sa valeur au vrai paramètre θ_0 obéit à la loi limite suivante

$$2\big(l_n(\widehat{\theta}) - l_n(\theta_0)\big) \longrightarrow \chi^2(D)$$

D'autres critères proposés ultérieurement tels que le critère bayésien proposé par Schwartz, connu sous le nom de BIC (voir [29]) 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. Il se trouve que dans bon nombre de problèmes, tels que la sélection de variables ou la détection de ruptures multiples par exemple, il est souhaitable de laisser croitre la taille des modèles ou encore le nombre de modèles d'une dimension donnée avec le nombre d'observations. Une théorie non asymptotique de la sélection de modèles a donc émergé durant ces dix dernières années qui vise à prendre en compte ce type de situations. L'enjeu central aussi bien sur le plan théorique que pratique est de comprendre comment pénaliser un critère de type log-vraisemblance de façon à garantir une performance de sélection optimale. La théorie non asymptotique s'appuie sur des inégalités de concentration dont le prototype est l'inégalité de concentration de Talagrand pour les processus empiriques établie dans [30]. Elle donne des indications sur la structure des pénalités qu'il convient d'utiliser mais n'est parfois pas suffisamment précise pour arbitrer la valeur de

certaines constantes qui restent donc à calibrer au moment d'implémenter effectivement ce type de critères. Ces constantes peuvent être inconnues pour des raisons diverses. Il peut s'agir d'une faiblesse de la théorie qui garantit l'existence d'une constante absolue mais sans en donner la valeur numérique. Le problème peut être de nature statistique lorsque cette constante dépend objectivement de la loi inconnue des observations.

Notre propos est ici de promouvoir une méthode de calibration de pénalité à partir des données. Cette méthode est en partie fondée sur des résultats théoriques établis et en partie sur une heuristique permettant de l'extrapoler à d'autres cadres que le cadre strict dans lequel la théorie permet de la valider. D'un point de vue pratique cette question de la calibration de la pénalisation est essentielle. Elle reste à l'heure actuelle un écueil important qui rend le choix de modèle par critère pénalisé moins attractif que son concurrent la validation croisée dans ses différentes versions dont le « V-fold ». C'est particulièrement vrai dans le contexte où cette dernière est la plus naturelle, c'est-à-dire celui de l'apprentissage statistique où les observations sont supposées indépendantes et de même loi. L'objet de cet article est d'une part de présenter cette nouvelle méthode de calibration et ses fondements théoriques (essentiellement issus de [12]) comme une alternative crédible à la validation croisée. D'autre part nous ferons le point sur ce qui est établi et proposerons des pistes de réflexion pour des travaux futurs sur ce qui ne l'est pas encore.

2. Sélection de modèle

Le cadre statistique dans lequel nous nous plaçons est celui où l'on observe une variable aléatoire ξ dont la loi dépend d'une quantité inconnue s (généralement une fonction). Nous avons en tête la situation où $\xi = \xi^{(n)}$ dépend d'un facteur d'échelle n, typiquement $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$, où les variables $\xi_1, \ldots, \xi_n$ sont indépendantes. Les principaux exemples d'estimation fonctionnelle auxquels nous pensons sont les suivants

• Estimation de densité

On observe $\xi_1, \ldots, \xi_n$ indépendantes et de même loi, de densité inconnue s par rapport à une mesure dominante μ.

• Régression

On observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendantes avec

$$Y_i = s(X_i) + \varepsilon_i, \qquad 1 \leqslant i \leqslant n.$$

Les variables *explicatives* $X_1, \ldots, X_n$ ne sont pas nécessairement équidistribuées. Les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendantes, équidistribuées, et conditionnellement centrées, c'est-à-dire que $\mathbb{E}[\varepsilon_i | X_i] = 0$. s est la *fonction de régression*.

• Bruit blanc gaussien continu

Soit $s \in \mathbb{L}^2[0,1]$. On observe le processus $\xi^{(n)}$ sur $[0,1]$ défini par

$$d\xi^{(n)}(x) = s(x) + \frac{1}{\sqrt{x}} dB(x), \qquad \xi^{(n)}(0) = 0,$$

où B désigne un mouvement Brownien. Le niveau de bruit ε est écrit ici sous la forme $\varepsilon = 1/\sqrt{n}$ par pure commodité de notation et afin de permettre une comparaison aisée avec les autres cadres.

Dans tous les exemples ci-dessus on observe une variable aléatoire $\xi^{(n)}$ dont la loi dépend d'une fonction inconnue s appartenant à un certain espace fonctionnel $\mathcal{S}$ (l'ensemble de toutes les densités de probabilité par rapport à μ dans le cadre de la densité ou bien $S = \mathbb{L}^2(\mu)$ dans le cas de la régression).

2.1. Estimation dans un modèle

Considérons un critère empirique γ_n (fondé sur l'observation $\xi^{(n)}$) tel que sur l'ensemble

$$t \longrightarrow \mathbb{E}_s[\gamma_n(t)]$$

atteigne un minimum au point s. Un tel critère est appelé *contraste empirique* pour l'estimation de s. Étant donné un sous-ensemble S de $\mathcal{S}$ que nous appellerons *modèle*, un *estimateur de minimum de contraste* $\widehat{s}$ de s est un minimiseur de γ_n sur S. Cette méthode d'estimation repose sur l'idée intuitive suivante : si on substitue le contraste empirique γ_n à son espérance, la minimisation de γ_n sur un sous-ensemble S de $\mathcal{S}$ doit conduire à un estimateur raisonnable de s, tout du moins si le modèle S n'est pas trop grand et si s appartient ou est suffisamment proche du modèle S. Cette méthode d'estimation est évidemment très employée, notamment en statistique paramétrique. Elle inclut la méthode du maximum de vraisemblance comme rappelé plus bas. Voyons quels sont les principaux exemples de contraste empirique dans les différents cadres d'estimation fonctionnelle introduits ci-dessus. Dans chacun des exemples que nous étudierons, nous vérifierons que le critère introduit est bien un contraste empirique en montrant que la fonction de perte naturelle

$$\ell(s,t) = \mathbb{E}_s[\gamma_n(t)] - \mathbb{E}_s[\gamma_n(s)] \tag{1}$$

est bien positive pour tout $t \in \mathcal{S}$. Dans le cas où $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$, on peut définir un critère empirique par $\gamma_n(t) = P_n[\gamma(t,.)] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \gamma(t, \xi_i)$, si bien qu'il nous suffit dans chaque cas de préciser quelle est la fonction considérée.

• Estimation de densité

Le choix $\gamma(t,x) = -\log(t(x))$ conduit à la méthode du maximum de vraisemblance et la fonction de perte ℓ est donnée par $\ell(s,t) = K(s,t)$, où $K(s,t)$ désigne l'information de Kullback-Leibler entre les probabilités $s\mu$ et $t\mu$, i.e. $K(s,t) = \displaystyle\int s \log\left(\frac{s}{t}\right)$ si $s\mu$ est absolument continue par rapport à $t\mu$ et

$K(s,t) = +\infty$ sinon. Supposant cette fois que $\mathcal{S} = \mathbb{L}_2(\mu)$, il est également possible de définir une méthode des moindres carrés pour l'estimation de densité en posant $\gamma(t,x) = ||t||^2 - 2t(x)$, où $||.||$ note la norme dans $\mathbb{L}_2(\mu)$. La fonction de perte correspondante s'écrit dans ce cas $\ell(s,t) = ||s-t||^2$ pour tout $t \in \mathbb{L}_2(\mu)$.

• **Régression**

Soit μ la moyenne arithmétique des lois des variables explicatives $X_1, \ldots, X_n$, alors l'estimation par moindres carrés est obtenue en posant pour tout $t \in \mathbb{L}_2(\mu)$ $\gamma(t,(x,y)) = (y - t(x))^2$, la fonction de perte correspondante ℓ est donnée par $\ell(s,t) = ||s-t||^2$, où $||.||$ désigne la norme dans $\mathbb{L}_2(\mu)$.

• **Bruit blanc gaussien**

Nous définissons le critère des moindres carrés par $\gamma_n(t) = ||t||^2 - 2\int_0^1 t(x)\, d\xi^{(n)}(x)$, pour tout $t \in \mathbb{L}_2([0,1])$ et la fonction de perte correspondante s'écrit simplement $\ell(s,t) = ||s-t||^2$.

2.2. Le paradigme du choix de modèle

S'agissant de choisir au mieux un modèle S sur lequel minimiser un critère empirique donné, la difficulté est la suivante. Afin de garantir que la cible s que l'on cherche à estimer se trouve à coup sûr dans le modèle considéré et se prémunir ainsi contre toute erreur de modélisation, on pourrait avoir la tentation naïve de prendre S aussi grand que possible i.e. $S = \mathcal{S}$. Or il est bien connu qu'un tel choix est désastreux puisque qu'il conduit à des estimateurs non consistants (voir [6]) ou sous-optimaux (voir [9]). Choisir par avance un modèle présente donc les risques suivants :

- Si S est un « petit » modèle (penser à un modèle paramétrique défini par un ou deux paramètres par exemple) le comportement de l'estimateur du minimum de contraste est satisfaisant si s est proche de S mais le modèle peut aisément s'avérer faux.
- Si au contraire, $\mathcal{S}$ est un « énorme » modèle (penser par exemple à l'ensemble de toutes les fonctions continues sur [0,1] dans le cadre du bruit blanc gaussien par exemple), la minimisation du contraste empirique conduit à un estimateur de très piètre qualité de s même si en ce cas s appartient vraiment à S.

Illustration dans le cadre du bruit blanc

Si est un sous-espace vectoriel de dimension D de $\mathbb{L}_2([0,1])$, on peut calculer explicitement l'estimateur des moindres carrés $\hat{s}$ sur S. En effet, si $(\phi_j)_{1 \leqslant j \leqslant D}$ désigne une base orthonormée de S la décomposition de $\hat{s}$ est de la forme suivante

$$\hat{s} = \sum_{j=1}^{D} \left(\int_0^1 \phi_j(x) d\xi^{(n)}(x) \right) \phi_j.$$

Par ailleurs on peut écrire les coefficients de $\hat{s}$ dans la base $(\phi_j)_{1\leqslant j\leqslant D}$ sous la forme

$$\int_0^1 \phi_j(x)d\xi^{(n)}(x) = \int_0^1 \phi_j(x)s(x)dx + \frac{1}{\sqrt{n}}\,\eta_j$$

pour tout entier $1 \leqslant j \leqslant D$, où les variables $\eta_1, \ldots, \eta_D$ sont des variables indépendantes et de même loi normale centrée réduite. La formule de Pythagore permet alors d'exprimer la perte quadratique de $\hat{s}$

$$||s - \hat{s}||^2 = d^2(s, S) + \frac{1}{n}\sum_{j=1}^{D}\eta_j^2$$

et d'obtenir le risque quadratique

$$\mathbb{E}_s\left[||s - \hat{s}||^2\right] = d^2(s, S) + \frac{D}{n}.$$

Cette formule résume bien à elle seule le paradigme du choix de modèle. En effet, choisir un modèle S garantissant un risque quadratique faible pour l'estimateur des moindres carrés sur ce modèle, implique que les deux termes antinomiques que sont le terme de biais $d^2(s, S)$ d'une part et le terme de variance D/n d'autre part doivent être simultanément petits. Plus généralement, c'est ce point de vue fondé sur l'analyse du risque que nous adoptons pour décider de la qualité d'un modèle. Si nous considérons un contraste empirique γ_n et une collection (finie ou dénombrable) de modèles $(S_m)_{m\in\mathfrak{M}}$, chaque modèle S_m est représenté par l'estimateur de minimum de contraste $\hat{s}_m$ relativement à γ_n sur S_m. Le but est de sélectionner le « meilleur » estimateur au sein de la collection $(\hat{s}_m)_{m\in\mathfrak{M}}$. Idéalement, on souhaiterait sélectionner $m(s)$ minimisant le risque $\mathbb{E}[\ell(s, \hat{s}_m)]$ par rapport à $m \in \mathfrak{M}$. L'estimateur de minimum de contraste $\hat{s}_{m(s)}$ correspondant à ce modèle idéal est usuellement appelé *oracle* selon la terminologie introduite par Donoho et Johnstone dans [18]. Bien entendu l'oracle est inaccessible, son seul intérêt est de fournir une performance étalon à laquelle il est utile de se comparer.

2.3. Choix de modèle par pénalisation

Une réponse générale à la question de la construction d'une procédure de choix de modèles visant à imiter un oracle est fournie par la méthode de minimisation de contraste empirique pénalisé. On considère une fonction de pénalité convenable pen : $\mathfrak{M} \longrightarrow \mathbb{R}_+$ et on choisit $\hat{m}$ de façon à minimiser

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

sur $\mathfrak{M}$. L'origine de cette méthode remonte au début des années 70 avec les travaux séminaux d'Akaike et de Schwartz (voir [1] et [29]) concernant la log-vraisemblance pénalisée et ceux de Mallows concernant les moindres carrés pénalisés (voir [17] et [25]). Dans chacun des deux cas ci-dessus les fonctions de pénalité sont proportionnelles au nombre de paramètres D_m du modèle S_m

- Akaike : D_m/n
- Schwartz : $\ln(n)D_m/n$
- Mallows : $2D_m/n$,

(la variance des erreurs de régression est ici supposée connue et égale à 1 pour simplifier). Comme nous l'avons rappelé dans l'introduction, l'heuristique d'Akaike repose fortement sur une approximation asymptotique de la log-vraisemblance (le théorème de Wilks) qui ne vaut que lorsque la collection de modèles est fixée et n tend vers l'infini. Il en est de même pour le critère BIC de Schwartz qui lui aussi s'appuie sur une approximation asymptotique même si celle-ci est d'une nature bayésienne plutôt que fréquentiste. Il est utile à ce stade de fournir un premier exemple naturel de problème de choix de modèle pour lequel manifestement ce cadre s'avère trop restrictif.

Un exemple d'école : la détection de ruptures multiples

La détection de ruptures multiples sur la moyenne se modélise de la manière suivante. Considérons un signal bruité ξ_i observé à chaque instant i/n de $[0,1]$ et structuré par le modèle de regression

$$\xi_i = s(i/n) + \varepsilon_i, \qquad 1 \leqslant i \leqslant n$$

dans lequel les erreurs sont des variables aléatoires centrées en espérance, indépendantes et identiquement distribuées. Détecter des ruptures sur la moyenne revient à déterminer une partition $\hat{m}$ par des intervalles dont les extrémités figurent sur la grille $\{i/n,\ 0 \leqslant i \leqslant n\}$ sur laquelle estimer le vrai signal s par un signal constant par morceaux. Autrement dit, définissant pour chaque partition m de ce type l'espace vectoriel S_m des fonctions constantes par morceaux sur m, il s'agit de sélectionner un modèle parmi la collection $(S_m)_{m\in\mathfrak{M}}$, où $\mathfrak{M}$ désigne la collection de toutes les partitions par des intervalles dont les extrémités figurent sur la grille $\{i/n,\ 0 \leqslant i \leqslant n\}$. Dans ce cas, le nombre de modèles de dimension D correspond au nombre de partitions à D intervalles dont les extrémités figurent sur la grille $\{i/n,\ 0 \leqslant i \leqslant n\}$. Il vaut donc exactement $\binom{n-1}{D-1}$ et croît par conséquent polynômialement avec le nombre d'observations n.

Dans un tel contexte, la théorie asymptotique classique ne s'applique plus et il convient de définir une approche alternative. L'approche non asymptotique pour la sélection de modèles telle qu'initiée dans [10] et [8] et vigoureusement développée depuis (voir [27] pour un panorama de résultats et une liste de références), diffère de l'approche asymptotique usuelle en ce sens que le nombre aussi bien que la dimension des modèles considérés peuvent très bien dépendre de n. Ceci ouvre la porte à l'utilisation de collections de modèles connus pour leurs propriétés d'approximation, tels que : des développements en ondelettes, des polynômes trigonométriques, des polynômes par morceaux, etc... Les fonctions de pénalité utilisées sont typiquement de la forme

$$(C_1 + C_2 L_m)\frac{D_m}{n}$$

où les poids L_m sont assujettis à vérifier la contrainte

$$\sum_{m\in\mathfrak{M}} e^{-L_m D_m} \leqslant 1$$

et les constantes C_1 et C_2 sont libres de n. Il peut parfaitement se produire que de nombreux modèles de la collection soient définis par le même nombre de paramètres. C'est le rôle des poids L_m que de tenir compte de cette situation en quantifiant la complexité de la collection de modèles.

3. Sélection de modèle gaussien

Si nous restreignons l'étude au cas gaussien, il est possible de donner une forme précise de la fonction de pénalité qu'il convient d'utiliser. Avant d'aller plus loin il est utile de préciser un cadre stochastique commode, permettant de traiter aussi bien du modèle de bruit blanc continu que du modèle discret.

3.1. Le modèle linéaire gaussien généralisé

Considérons comme dans [11] le modèle linéaire gaussien généralisé défini de la manière suivante. Étant donné un espace de Hilbert séparable $\mathbb{H}$, on observe le processus $\boldsymbol{\xi}^\varepsilon$ donné par

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

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. Ce cadre est conçu pour généraliser le modèle linéaire gaussien classique en dimension finie et inclut donc tout naturellement celui-ci comme exemple.

Le modèle linéaire gaussien fini-dimensionnel

Dans ce cas on observe, comme indiqué plus haut,

$$\xi_i = s_i + \sigma\,\eta_i, \qquad 1 \leqslant i \leqslant n, \tag{4}$$

où les variables aléatoires η_i sont indépendantes et de même loi normale $N(0,1)$. 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 $||.||$ et si nous posons $W(t) = \sqrt{n}\langle \eta, t\rangle$, alors W est bien un processus gaussien isonormal et le processus

$$\boldsymbol{\xi}^\varepsilon : t \longrightarrow \frac{1}{n}\sum_{i=1}^{n} \xi_i t_i$$

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

L'intérêt de (3) est de pouvoir traiter indifféremment des modèles de bruit blanc continu ou discret en choisissant un espace $\mathbb{H}$ de référence convenable.

Le modèle de bruit blanc continu

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

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

où B désigne un mouvement brownien sur $[0,1]$. Si nous définissons alors pour tout $t \in \mathbb{L}_2[0,1]$, $W(t) = \displaystyle\int_0^1 t(x)dB(x)$, W est bien un processus gaussien isonormal sur $\mathbb{L}_2[0,1]$ et $\boldsymbol{\xi}^\varepsilon(t) = \displaystyle\int_0^1 t(x)d\xi^\varepsilon(x)$ obéit bien à (3) dès lors que $\mathbb{L}_2[0,1]$ est muni de son produit scalaire usuel $\langle s,t\rangle = \displaystyle\int_0^1 s(x)t(x)dx$. Typiquement s représente un signal et $d\xi^\varepsilon(x)$ représente le signal bruité reçu à l'instant x.

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 $[0,1]$. On obtient alors un modèle de bruit blanc discret

$$\xi_i = s(i/n) + \sigma\,\eta_i, \quad 1 \leqslant i \leqslant n, \tag{6}$$

où les variables aléatoires η_i sont indépendantes et de même loi normale $N(0,1)$. Si s est un signal, ξ_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, on peut poser $\sigma = \varepsilon\sqrt{n}$ et $\eta_i = \sqrt{n}(B(i/n) - B((i-1)/n))$, pour tout $i \in [1,n]$.
Le signal bruité reçu à l'instant i/n vaut alors

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

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

Lorsque S_m désigne un sous-espace vectoriel de dimension finie D_m de $\mathbb{H}$, comme dans le cas du bruit blanc continu étudié plus haut, l'estimateur $\hat{s}_m$

des moindres carrés sur le modèle S_m est défini comme le minimiseur de

$$\gamma^{\varepsilon}(t) = ||t||^2 - 2\boldsymbol{\xi}^{\varepsilon}(t) \tag{8}$$

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

$$\hat{s}_m = \sum_{j=1}^{D_m} \boldsymbol{\xi}^{\varepsilon}(\phi_j)\phi_j.$$

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

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

on en déduit que

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

suit une loi du chi-deux à D_m degrés de liberté. La formule de Pythagore assure donc que le risque quadratique de $\hat{s}_m$ s'écrit

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

où s_m désigne la projection orthogonale de s sur S_m. S'agissant d'imiter l'oracle, c'est-à-dire l'estimateur idéal $\hat{s}_{m(s)}$ réalisant $\inf_{m\in\mathfrak{M}} \mathbb{E}_s||\hat{s}_m - s||^2$, une idée naturelle consiste à estimer sans biais le risque sur chacun des modèles, puis de minimiser le critère ainsi obtenu. C'est le principe qui préside à la construction du C_p de Mallows que nous rappelons à présent.

3.2. L'heuristique de Mallows

Sachant que $m(s)$ minimise le risque quadratique

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

l'idée la plus naturelle serait d'estimer le risque quadratique de $\hat{s}_m$ puis de minimiser cette estimation. Sa mise en œuvre bute sur la difficulté du problème de l'estimation du terme de biais $||s_m - s||^2$. Il convient donc d'aménager cette idée initiale en remarquant que, d'après la formule de Pythagore, $||s_m - s||^2 = ||s||^2 - ||s_m||^2$. Par conséquent $m(s)$ minimise également

$$-||s_m||^2 + D_m\varepsilon^2. \tag{10}$$

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

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

ce qui implique que $\mathbb{E}_S||\hat{s}_m||^2 = ||s_m||^2 + D_m\varepsilon^2$.

En substituant à $||s_m||^2$ son estimateur non biaisé naturel $||\hat{s}_m||^2 - D_m\varepsilon^2$ dans (10), on obtient le critère de Mallows

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

3.3. Un théorème général

L'heuristique ci-dessus peut-être justifiée (ou corrigée) en spécifiant comment se concentre $||\hat{s}_m||^2$ autour de son espérance $D_m\varepsilon^2$, uniformément par rapport à $m \in \mathfrak{M}$. L'inégalité de concentration gaussienne (voir [27] pour un énoncé précis de cette inégalité ainsi que sa preuve et des références) est un outil adéquat pour réaliser cela. Il faut bien noter d'ailleurs que cette inégalité fournit simultanément une forme de pénalité précise et une inégalité de comparaison à l'oracle pour l'estimateur pénalisé correspondant. C'est ce qui résulte du théorème ci-dessous extrait de [11].

THÉORÈME 1. — *Soit $(L_m)_{m\in\mathfrak{M}}$ une famille de poids positifs ou nuls vérifiant la condition*

$$\sum_{m\in\mathfrak{M}} \exp(-L_m D_m) = \Sigma < \infty.$$

Étant donné $K > 1$, supposons que pour tout $m \in \mathfrak{M}$

$$\mathrm{pen}(m) \geqslant K D_m \varepsilon^2 (1 + \sqrt{2L_m})^2.$$

Si $\hat{m}$ minimise le critère des moindres carrés pénalisé

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

l'inégalité suivante est valide

$$\mathbb{E}_s||\hat{s}_{\hat{m}} - s||^2 \leqslant C(K)\{\inf_{m\in\mathfrak{M}}(||s_m - s||^2 + \mathrm{pen}(m)) + \Sigma\varepsilon^2\}, \qquad (11)$$

où $C(K)$ 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 $\hat{s}_{\hat{m}}$ et celui de l'oracle $\inf_{m\in\mathfrak{M}} \mathbb{E}_s||\hat{s}_m - s||^2$. Pour ce faire, nous pouvons raisonner de la manière suivante. Rappelant à nouveau que le risque quadratique de $\hat{s}_m$ s'exprime sous la forme

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

considérons la situation la plus simple dans laquelle, pour un certain nombre L, le choix de $x_m = LD_m$ pour tout $m \in \mathfrak{M}$ conduit disons à $\sum_{m\in\mathfrak{M}} \exp(-x_m) \leqslant 1$ (prendre 1 comme borne supérieure n'a rien de magique ici, 2 ferait tout autant l'affaire !). Si nous choisissons $\mathrm{pen}(m) K D_m (1 + \sqrt{2L})^2 \varepsilon^2$, 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 $\inf_{m\in\mathfrak{M}} \mathbb{E}_s||\hat{s}_m - s||^2$. Dans ce cas nous obtenons bien une comparaison avec le risque idéal et l'estimateur sélectionné se comporte, à 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

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

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

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

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.

3.4 Exemples

De nombreux exemples d'applications du Théorème 1 sont développés dans [11]. Contentons-nous ici de traiter deux applications d'intérêt : la sélection de variables et la détection de ruptures.

La sélection de variables

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

La sélection de variables ordonnée

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

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

Comme α peut être choisi arbitairement 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

$$\mathbb{E}_s||\hat{s}_{\hat{m}} - s||^2 \leqslant C' \inf_{m\in\mathfrak{M}} \mathbb{E}||\hat{s}_m - s||^2,$$

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

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

et nous pouvons prendre comme pénalité

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

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

$$\mathbb{E}_s||\hat{s}_{\hat{m}} - s||^2 \leqslant C'(K) \inf_{D\geqslant 1} \left\{ \inf_{m\in\mathfrak{M}, D_m = D} (||s_m - s||^2) + D \log(N)\varepsilon^2 \right\}, \quad (12)$$

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 $\{\phi_j, j \leqslant N\}$ 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 [18]). En effet, lorsque pour un certain nombre $T > 0$

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

et

$$\hat{\beta}_j = \boldsymbol{\xi}^\varepsilon(\phi_j), \quad \text{pour } 1 \leqslant j \leqslant N,$$

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

$$\hat{s}_m = \sum_{j \in m} \hat{\beta}_j \phi_j.$$

Dans ces conditions, le critère pénalisé s'écrit

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

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

$$\hat{m} = \{j \in \Lambda, -\hat{\beta}_j^2 + T^2 \leqslant 0\}$$

En d'autres termes

$$\hat{s}_{\hat{m}} = \sum_{j=1}^{N} \hat{\beta}_j \mathbb{1}_{\{|\hat{\beta}_j| \geqslant T\}} \phi_j$$

qui est bien un estimateur par seuillage, le seuil T fourni par le Théorème 1 étant finalement de la forme $T = \sqrt{K}(1 + \sqrt{2\,\log(N)})\varepsilon$. 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 reppelons la majoration classique suivante pour le coefficient binomial

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

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

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

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

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

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

$$\mathbb{E}_s ||\hat{s}_{\hat{m}} - s||^2 \leqslant C'' \inf_{1 \leqslant D \leqslant N} \{ b_D^2(s) + D(1 + \ln(N/D))\varepsilon^2 \}, \tag{14}$$

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

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

$$\begin{aligned}
&\inf_{m\in M}\left\{-\sum_{j\in m}\hat\beta_j^2 + K\varepsilon^2|m|(1+\sqrt{2L(|m|)})^2\right\}\\
&\quad= \inf_{D\leqslant N}\left\{-\sup_{\{(m||m|=D\}}\sum_{j\in m}\hat\beta_j^2 + K\varepsilon^2 D\left(1+\sqrt{2L(D)}\right)^2\right\}\\
&\quad= \inf_{D\leqslant N}\left\{-\sum_{j=1}^{D}\hat\beta_{(j)}^2 + K\varepsilon^2 D\left(1+\sqrt{2L(D)}\right)^2\right\}
\end{aligned}$$

où $\hat\beta_{(1)}^2 \geqslant \ldots \geqslant \hat\beta_{(N)}^2$ désignent les carrés des coefficients estimés $\{\hat\beta_j, j\leqslant N\}$, rangés par ordre décroissant. Nous constatons que la minimisation du critère pénalisé revient à sélectionner une valeur $\hat D$ de D minimisant

$$-\sum_{j=1}^{D}\hat\beta_{(j)}^2 + K\varepsilon^2 D(1+\sqrt{2L(D)})^2$$

et finalement à exprimer l'estimateur pénalisé sous la forme

$$\hat s_{\hat m} = \sum_{j=1}^{\hat D}\hat\beta_{(j)}\phi_{(j)}. \tag{15}$$

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 $S_D = \bigcup_{|m|=D} S_m$, $D\leqslant N$. Autrement dit, il existe une constante absolue strictement positive κ telle que, quel soit l'estimateur $\tilde s$ de s,

$$\sup_{s\in S_D}\mathbb{E}_s||\tilde s - s||^2 \geqslant \kappa D(1+\ln(N/D))\varepsilon^2.$$

On vient de voir comment sélectionner des variables au sein d'une même base mais il est également possible bien entendu de 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 [24] et [22]).

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

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

où les erreurs ξ_j sont indépendantes et de même loi normale $N(0,1)$. 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 $\{S_m\}_{m\in\mathfrak{M}}$, où $\mathfrak{M}$ désigne la collection de toutes les partitions possibles de $[0,1]$ par des intervalles dont les extrémités se situent sur la grille $\{j/n, 0 \leqslant j \leqslant n\}$. 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 celles 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.

3.5 Conclusions

Une question naturelle se pose à lecture du Théorème 1 et de ses conséquences pour l'étude des exemples détaillée ci-dessus. Elle concerne la finesse du Théorème 1. Si on s'intéresse à la formule de pénalité suggérée par le Théorème 1

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

on peut en effet légitimement se poser la question de la valeur qu'il convient de donner à la constante K. Cette question est en fait double puisque d'une part il s'agit de savoir si la contrainte $K > 1$ est artificielle ou pas et d'autre part de comprendre s'il existe une « meilleure » valeur de K. Une réponse substantielle sinon exhaustive à cette question est fournie dans [12]. Une optimisation de $C(K)$ d'une part et des minorations sur la fonction de pénalité d'autre part permettent de mettre en évidence les phénomènes suivants

- La condition $K > 1$ dans l'énoncé du Théorème 1 n'est pas améliorable. Ceci vaut en particulier pour les exemples détaillés ci-dessus pour lesquels on peut montrer que choisir $K < 1$ conduit avec forte probabilité à sélectionner systématiquement des modèles de grande dimension.
- Si on tente d'optimiser l'inégalité (11) il en résulte que 2 est un bon choix pour la valeur de K.

Une première conséquence intéressante de la nécessité de la condition $K > 1$ est que la correction du C_p de Mallows proposée ci-dessus pour la sélection de variable complète ou la détection de ruptures multiples s'avère en fait nécessaire. Le C_p de Mallows peut donc réellement sous-pénaliser et il doit être corrigé lorsque le nombre de modèles de même dimension est trop élevé.

La seconde conséquence que nous allons détailler à présent est probablement plus importante et plus profonde.

4. Pénalisation à partir des données

Les résultats d'optimalité que nous venons d'évoquer peuvent en effet être synthétisés par la formule suivante : pénalité « *optimale* » = 2 × pénalité « *minimale* ». Cette formule porte en elle une conséquence très intéressante pour calibrer la fonction de pénalité *à partir des données*.

Dès le moment qu'une formule aussi simple relie la pénalité qu'il convient d'utiliser à la pénalité minimale, il suffit d'estimer cette dernière à partir des données pour obtenir une idée de la première. Pour ce faire les indications données par la théorie sont utiles. En effet la pénalité minimale peut être devinée à partir des données grâce au phénomène d'explosion évoqué plus haut : 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. Il reste à préciser davantage la méthode d'estimation de la pénalité minimale puis d'interpréter la stratégie décrite ci-dessus d'une façon qui la rende plus aisément transposable à un contexte autre que le cadre purement gaussien.

La « recette » pour estimer la pénalité qu'il convient d'utiliser dans le contexte gaussien précédent peut se résumer ainsi.

- Calculer l'estimateur de minimum de contraste $\hat{s}_D$ sur l'union des modèles de même dimension D.
- Utiliser la théorie pour deviner la forme de la pénalité $\mathrm{pen}(D)$. Typiquement $\mathrm{pen}(D) = \alpha\, D$.
- Estimer α à partir des données en multipliant par 2 la valeur critique $\hat{\alpha}_c$ qui garantit que le critère pénalisé cesse d'exploser.

Du fait que cette valeur critique $\hat{\alpha}_c$ correspond à la pente de $\gamma^\varepsilon(\hat{s}_D)$ comme fonction approximativement affine de D pour les grandes valeurs de D, cette heuristique pour estimer la pénalité à partir des données sera dénommée « heuristique de pente » dans ce qui suit. Rechercher une pénalité proportionnelle à D est une première possibilité ; on peut, inspiré par la théorie précédente vouloir rechercher comme dans le cas de la détection de ruptures multiples des formes de pénalité un peu plus complexes du genre : $\mathrm{pen}(D) = \alpha\, D(\beta + \ln(n/D))$. C'est parfaitement possible, comme l'a montré Lebarbier dans [20] où cette méthode a été implémentée et testée.

4.1 Les heuristiques de Mallows et d'Akaike revisitées

S'agissant à présent d'interpréter les résultats obtenus dans le cas gaussien de façon à les extrapoler dans un cadre plus général, il est utile de revenir au fondement même du critère de Mallows aussi bien que celui d'Akaike :

le principe d'estimation sans biais du risque. Nous avons vu dans le cas gaussien que ce principe peut être mis en défaut en présence d'une liste pléthorique de modèles. Souvenons-nous en effet que, si le nombre de modèles par dimension devient trop élevé, des corrections logarithmiques au critère de Mallows deviennent nécessaires comme nous l'avons constaté pour la sélection de variables complète ou bien encore la détection de rupture multiples. Il convient donc d'utiliser ce principe en prenant la précaution préliminaire de regrouper ensemble les modèles comportant le même nombre de paramètres. Reprenons le cadre général pour la sélection de modèle décrit dans la Section 2. Partant d'une liste de modèles (paramétriques) $(S_m)_{m\in\mathfrak{M}}$ et d'un contraste empirique γ_n donnés, nous commençons donc par regrouper pour chaque entier D les modèles définis par D paramètres et nous notons ensuite $\hat{s}_D$ l'estimateur de minimum contraste sur la réunion de ces modèles que nous noterons désormais $\mathbb{S}_D$.

À présent analysons le comportement de $\gamma_n(\hat{s}_D)$. Si nous notons s_D une valeur pour laquelle $\mathbb{E}_s(\gamma_n(t))$ atteint un minimum lorsque t parcourt $\mathbb{S}_D$, nous pouvons décomposer le minimum du contraste empirique sur $\mathbb{S}_D$ de la façon suivante : $\gamma_n(\hat{s}_D) = \gamma_n(s_D) - [\gamma_n(s_D) - \gamma_n(\hat{s}_D)] = \gamma_n(s_D) - \hat{v}_D$. Le terme $\hat{v}_D$ défini dans cette décomposition est évidemment positif, c'est un terme de *variance empirique*. À présent remarquons que quelle que soit la pénalité $\mathrm{pen}(D)$ considérée, minimiser $\gamma_n(\hat{s}_D) + \mathrm{pen}(D)$ lorsque D varie, revient de manière équivalente à minimiser $\gamma_n(s_D) - \gamma_n(s) - \hat{v}_D + \mathrm{pen}(D)$.

Comme $\gamma_n(s_D) - \gamma_n(s)$ estime sans biais $\ell(s, s_D)$, minimiser ce critère revient (approximativement cette fois) à minimiser

$$\ell(s, s_D) - \hat{v}_D + \mathrm{pen}(D) \tag{16}$$

C'est ici qu'on peut clairement voir apparaître l'origine des concepts de *pénalité minimale* et *pénalité optimale*. En effet pour qu'une procédure de choix de modèles possède une chance de fonctionner il est nécessaire que la pénalité $\mathrm{pen}(D)$ compense à tout le moins $\hat{v}_D$. Si tel n'est pas le cas, le critère a toutes les chances « d'exploser » c'est-à-dire que son minimum sera atteint systématiquement par des modèles de très grande dimension. C'est en tout cas le phénomène constaté dans le cas gaussien. Pour en revenir à une situation simple et explicite comme la sélection de variables ordonnée pour le modèle de bruit blanc continu avec niveau de bruit $1/\sqrt{n}$ par exemple, on dispose d'un calcul explicite de $\hat{v}_D$ et l'augmentation de $\mathbb{E}(\hat{v}_D) = D/n$ avec D explique la raison pour laquelle une pénalité de la forme KD/n conduit à un critère explosif tant que $K < 1$. Autrement dit la pénalité minimale est de l'ordre de $\hat{v}_D$ ou de son espérance puisque dans cette analyse on mise sur un bonne concentration des quantités considérées autour de leur espérance (D/n donc dans le cas particulier de la sélection de variables ordonnée).

Intéressons nous maintenant à la pénalité qui conduirait à une sélection de modèle idéale. Selon le critère de performance édicté en Section 2, notre but est d'imiter l'oracle, c'est-à-dire de minimiser $\ell(s, \hat{s}_D$ (ou son espérance puisqu'encore une fois nous confondons ici l'analyse en moyenne et l'analyse trajectorielle). La pénalité idéale serait donc celle permettant à la formule (16)

de reconstituer la perte $\ell(s, \hat{s}_D)$, c'est-à-dire $\text{pen}_{\text{id}}(D) = \hat{v}_D + \ell(s_D, \hat{s}_D) = \hat{v}_D + v_D$, où nous notons (un peu abusivement) $\ell(s_D, t) = \mathbb{E}_s(\gamma_n(t)) - \mathbb{E}_s(\gamma_n(s_D))$, pour tout t de $\mathbb{S}_D$. Bien entendu cette quantité en tant que telle ne peut en aucun cas être effectivement utilisée comme pénalité puisqu'elle dépend de qui nous est par nature inconnu. C'est ici que les différentes approches pour la sélection de modèle divergent.

L'approche asymptotique

Dans le cas du C_p de Mallows, on profite du calcul explicite en espérance $\mathbb{E}(\hat{v}_D) = \mathbb{E}(v_D) = D/n$ et on évalue la pénalité idéale $\text{pen}_{\text{id}}(D)$ par son espérance $2D/n$. Pour ce qui concerne le critère d'Akaike, on utilise une évaluation asymptotique lorsque n tend vers l'infini (en liaison avec le Théorème de Wilks) $\mathbb{E}(\hat{v}_D) \approx \mathbb{E}(v_D) \approx D/(2n)$ et on évalue la pénalité idéale asymptotiquement par D/n. On a déjà dit pourquoi cette approche asymptotique pouvait s'avérer limitée et pourquoi elle pouvait s'avérer trompeuse si on l'utilise à mauvais escient, c'est-à-dire quand la liste de modèles peut dépendre de n.

L'approche non asymptotique

Si nous nous tournons vers l'approche non asymptotique telle qu'elle fut initiée dans [10] , nous voyons qu'elle repose pour sa part sur des techniques de processus empiriques (s'appuyant en particulier sur l'inégalité de concentration de Talagrand [16]). Pour situer plus précisément comment se positionne cette approche vis-à-vis de l'analyse ci-dessus, le plus simple est de revenir à la situation typique (décrite en Section 2) où le contraste empirique γ_n considéré est défini par $\gamma_n(t) = P_n(\gamma(t, .))$, où P_n désigne la mesure de probabilité empirique associée à n variables aléatoires indépendantes et de même loi P. L'expression de la « pénalité idéale » telle qu'elle est définie plus haut devient alors $\text{pen}_{\text{id}}(D) = (P_n - P)(\gamma(s_D, .) - \gamma(\hat{s}_D))$. L'idée utilisée dans [10] puis dans [8] consiste à majorer cette quantité par $\omega(s_D, \hat{s}_D) \times \sup_{t \in \mathbb{S}_D} \dfrac{(P_n - P)(\gamma(s_D, .) - \gamma(t))}{\omega(s_D, t)}$ où ω désigne un poids convenable (qui est une fonction croissante de la variance de $\gamma(s_D, .) - \gamma(t, .)$). L'introduction de la fonction de poids ω permet d'obtenir un effet de localisation et constitue l'apport essentiel de [10] et [8] par rapport à l'approche voisine initialement préconisée par Vapnik dans [31] et qui pour sa part repose sur une majoration non asymptotique globale et non pas locale de $(P_n - P)(\gamma(s_D, .) - \gamma(\hat{s}_D))$. Cette idée a été utilisée depuis dans de nombreux travaux, notamment dans [5] et [7] pour l'étude des moindres carrés en régression, dans [16] pour l'étude de la log-vraisemblance sur des log-splines en densité ou encore dans [28] pour l'estimation de l'intensité d'un processus de Poisson par sélection de modèle. On trouvera dans [27] un panorama sinon exhaustif en tout cas représentatif de résultats obtenus grâce à cette approche.

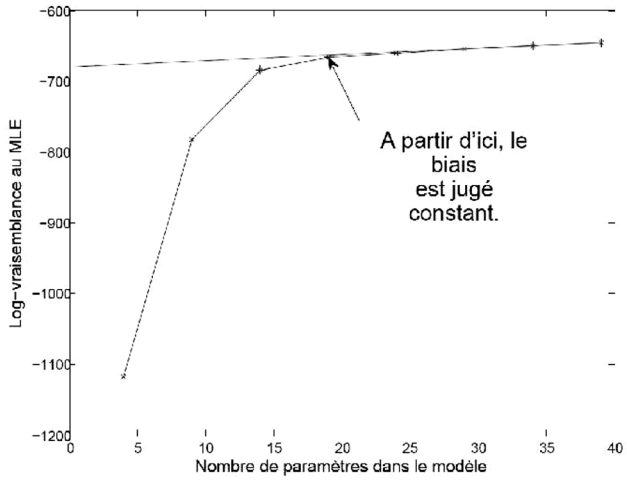
4.2 Une heuristique de pente

L'approche non asymptotique est plus flexible et permet d'envisager des listes de modèles quasiment arbitraires mais on rencontre un problème au moment

d'implémenter les critères pénalisés qui en résultent. Même si les progrès successifs effectués dans [21], [26] puis [15] ont permis de disposer de versions de l'inégalité de Talagrand avec des constantes numériques explicites et même en un certain sens optimales, il n'en reste pas moins que la plupart du temps les formules de pénalité font typiquement apparaître une constante multiplicative qui peut dépendre de la loi inconnue sous-jacente (de la variance des erreurs en régression, du maximum de la densité pour les moindres carrés en densité, du bruit de classification etc...). Autrement dit, le problème d'élimination du niveau de bruit du modèle de bruit blanc gaussien rencontré plus haut est illustratif d'un type de difficulté rencontré beaucoup plus généralement lorsqu'on utilise ce genre de méthodes.

Il est évidemment tentant d'utiliser exactement la même «recette» que dans le cas gaussien et c'est ce que nous préconisons de faire. Néanmoins, même si nous sommes conduits par l'envie d'extrapoler le cas gaussien, il est intéressant de chercher à le faire avec le plus de lucidité possible en interprétant cette procédure à la lumière de l'analyse proposée ci-dessus qui revisite les heuristiques de Mallows et d'Akaike. En effet, s'agissant du cas le plus simple où la pénalité recherchée est proportionnelle au nombre de paramètres du modèle, cela signifie qu'on s'attend à ce que $\hat{v}_D$ soit de l'ordre de $\alpha\, D$ avec α inconnu. Évidemment $\hat{v}_D$ n'est pas observable mais heureusement lorsque D devient assez grand $\gamma_n(s_D)$ a tendance à devenir constant et par conséquent (puisque rappelons le $\hat{v}_D = \gamma_n(s_D) - \gamma_n(\hat{s}_D)$), $\gamma_n(\hat{s}_D)$ est approximativement affine pour les grandes valeurs de D et on peut estimer α par la *pente* $\hat{\alpha}$ du graphe de $\gamma_n(\hat{s}_D)$ quand D est assez grand.

Voici un exemple simulé, gracieusement emprunté à Jean-Patrick Baudry, où on voit clairement apparaître ce phénomène de pente que nous venons de décrire. Il s'agit ci-dessous du graphe de la log-vraisemblance pour une centaine d'observations indépendantes et de même loi mélange de quatre lois gaussiennes bidimensionnelles.



Nous disposons donc d'une méthode graphique relativement simple pour estimer la pénalité minimale à partir des données. Il reste à comprendre sur quel présupposé repose le choix final de la pénalité optimale comme double de la pénalité minimale.

Rappelons qu'idéalement nous devrions pénaliser par $\mathrm{pen}_{\mathrm{id}}(D) = \hat{v}_D + v_D$. La pertinence du choix de la pénalité sous la forme $2\hat{\alpha}D$ repose donc sur l'espoir que $\hat{v}_D$ et v_D sont du même ordre de grandeur. C'est évidemment très exactement ce qui se passe dans le cas gaussien pour la sélection de variables ordonnée par exemple, puisque dans ce cas $\hat{v}_D$ et v_D sont égaux en espérance. Dans le cas général, cette question reste très largement ouverte et la théorie justifiant l'heuristique que nous venons de décrire reste donc à écrire pour l'essentiel. Il existe cependant une situation distincte du cas gaussien pour laquelle ce que nous venons d'avancer est justifiable de bout en bout (y compris l'optimalité du choix de pénalité qui en résulte), c'est la sélection d'histogrammes en régression bornée (voir le travail récent en collaboration avec Sylvain Arlot dans [4]).

4.3 Approche alternative et perspectives

Dans ce même contexte Sylvain Arlot est par ailleurs en mesure d'étudier finement le comportement de différentes methodes de rééchantillonage permettant d'estimer la pénalité idéale, l'idée principale étant là que la pénalité idéale $(P_n - P)(\gamma(s_D, .) - \gamma(\hat{s}_D))$ peut être estimée par $(P_n^* - P_n)(\gamma(\hat{s}_D, .) - \gamma(s_D^*))$, où P_n^* désigne la mesure de probabilité empirique randomisée et s_D^* le minimiseur empirique correspondant sur le modèle $\mathbb{S}_D$. Il est à même de comparer différentes méthodes de rééchantillonnage (dont le bootstrap) et d'inclure dans son analyse la validation croisée « V-fold » dont il propose des corrections (voir [2] et [3]). Pour aller plus loin dans ce type d'analyse, des résultats généraux de concentration de $\hat{v}_D$ autour de son espérance seront sans doute utiles. C'est précisément l'objet de [14] qui constitue un travail en cours en collaboration avec Stéphane Boucheron dans lequel nous établissons des analogues non asymptotiques du théorème de Wilks en nous appuyant sur les nouvelles inégalités de concentration établies dans [13]. Une autre question intéressante et directement orientée vers les utilisateurs de ces méthodes de choix de modèle par critère pénalisé, concerne la calibration du Lasso. Lorsque le nombre de variables présentes dans un modèle de régression linéaire devient trop élevé, il devient irréaliste d'optimiser un critère pénalisé sur tous les sous-ensembles de variables possibles comme préconisé précédemment. C'est la raison pour laquelle les méthodes de type Lasso ou LARS, fondées sur des pénalisations en norme ℓ_1 ont connu un essor considérable ces dernières années (voir en particulier [19]). L'intérêt de ces nouvelles méthodes réside dans leur capacité à traiter des problèmes de sélection en très grande dimension car l'optimisation d'un critère convexe régularisé de ce genre est extrêmement rapide. Indépendamment des performances statistiques de ces méthodes qui font aujourd'hui débat et attirent l'attention des théoriciens, on peut se demander (voir la discussion [23]) s'il est possible d'adapter les heuristiques que nous venons de développer ici afin de calibrer ces critères de régularisation ℓ_1. Un des enjeux théoriques serait de mettre en évidence des notions de pénalisation

ℓ_1 minimale et optimale analogues à celles dont nous venons de débattre avec les conséquences pratiques espérées sur la calibration à partir des données.

Références

[1] Akaike H. (1973). 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.

[2] Arlot S. (2007). Model selection by resampling penalization. arXiv :math/0701542v2

[3] Arlot S. (2008). V-fold cross-validation improved : V-fold penalization. arXiv :0802.0566v2

[4] Arlot S. and Massart P. (2008). Data-driven calibration of penalties for least-squares regression. arXiv :0802.0837v2.

[5] Baraud Y. (2000). Model selection for regression on a fixed design. *Probability Theory and Related Fields* **117**, n° 4 467-493.

[6] Bahadur R.R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhya Ser.A* **20**, 207-210.

[7] Baraud Y., Comte F. and Viennet G. (2001). Model selection for (auto-) regression with dependent data. *ESAIM : Probability and Statistics* **5**, 3349. http ://www.emath.fr/ps/.

[8] Barron A.R., Birgé L., Massart P. (1999). Risk bounds for model selection via penalization. *Probab. Th. Rel. Fields.* **113**, 301-415 .

[9] Birgé L. and Massart P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Th. Relat. Fields* **97**, 113-150.

[10] Birgé L. and Massart P. (1997). 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, Springer-Verlag, New-York.

[11] Birgé L. and Massart P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, n° 3 , 203-268.

[12] Birgé L., Massart P. (2007). Minimal penalties for Gaussian model selection. *Probab. Th. Rel. Fields* **138**, n° 1–2, 33–73.

[13] Boucheron S., Bousquet O., Lugosi G., Massart P. (2005). Moment inequalities for functions of independent random variables. *Ann. of Probability* **33**, n° 2, 514-560.

[14] Boucheron S. and Massart P. (en préparation). A poor man's Wilks phenomenon.

[15] Bousquet O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C.R. Math. Acad. Sci. Paris* **334**, n° 6, 495-500.

[16] Castellan G. (2003). Density estimation via exponential model selection. *IEEE Trans. Inform. Theory* **49**, n° 8, 2052-2060.

[17] Daniel C. and Wood F.S. (1971). Fitting Equations to Data. Wiley, New York.

[18] Donoho D.L. and Johnstone I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

[19] Efron B., Hastie T., Johnstone I. and Tibshirani R. (2004). Least angle regression. *Ann. Statist.* **32** n° 2, 407-499.
[20] Lebarbier E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing* **85**, n° 4, 717-736.
[21] Ledoux M. (1996). On Talagrand deviation inequalities for product measures. *ESAIM : Probability and Statistics* **1**, 63-87.
http ://www.emath.fr/ps/.
[22] Le pennec E. and Mallat S. (2005). Sparse Geometric Image Representation with Bandelets. *IEEE Trans. on Image Processing* **14**, n° 4, 423-438.
[23] Loubes J.M., Massart P. (2004). Discussion to Least Angle Regression. *Ann. of Statistics* **32**, n°2, 476-482.
[24] Mallat S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press.
[25] Mallows C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.
[26] Massart P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. of Probability* **28**, n°2, 863-884.
[27] Massart P. (2007). *Concentration inequalities and model selection.* In *Lectures on Probability Theory and Statistics, École d'Été de Probabilités de St-Flour XXXIII-2003* (J. Picard, ed.). Lecture notes in Mathematics n° 1896, Springer, Berlin.
[28] Reynaud-Bouret P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Relat. Fields* **126**, n° 1, 103-153.
[29] Schwartz G. (1978). Estimating the dimension of a model. *Ann. of Statistics* **6**, 461-464.
[30] Talagrand M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126**, 505-563.
[31] Vapnik V.N. (1982). *Estimation of dependencies based on empirical data.* Springer, New York.