Introduction aux méthodes de sélection de modèle

Gendre Xavier

Résumé

De façon générale, les statistiques ont pour objectif de retrouver des informations sur la loi d'un phénomène aléatoire que l'on observe. Ces informations peuvent être de nature très différentes selon le problème posé et pour les retrouver, nous sommes souvent amenés à formuler des hypothèses sur la loi elle-même bien qu'elle soit inconnue. Bien entendu, dans la pratique, ces hypothèses ne sont pas gratuites et il est donc important de pouvoir travailler avec peu d'hypothèses. Il s'agit là d'une des motivations de la sélection de modèle : fournir des méthodes dans des cadres généraux qui soient aussi "robustes" que possible.

1 Sélection de modèle à variance connue

1.1 Le cadre et les premiers outils

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

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

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

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

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

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

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

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

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

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

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

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

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

Ces deux écritures nous mènent aux deux égalités suivantes

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

qui nous serviront par la suite.

1.2 Risque quadratique de l'estimateur par projection

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

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

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

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

1.3 Choix d'un modèle

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

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

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

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

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

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

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

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

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

ou, de façon équivalente,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Si notre famille de modèles $\{S_m\}_{m\in\mathcal{M}}$ est telle que chaque S_m a une dimension $D_m > 0$ et pour tout entier N compris entre 1 et n, il n'y a au plus qu'un seul S_m qui soit de dimension

$D_m = N$, alors ce théorème permet de valider l'heuristique et d'obtenir une inégalité oracle. En effet, prenons $x_m = LD_m$ où $L > 0$ est une constante telle que

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

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

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

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

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

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

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

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

2 Et ensuite ?

2.1 Schéma de la preuve

La preuve du théorème 2 suit un schéma classique en sélection de modèle. Nous n'entrerons pas dans les détails, mais il nous semble important d'en préciser la clef de voute, c'est-à-dire la façon de choisir la forme du minorant des pénalités. Revenons à la définition de notre estimateur $\tilde{s}$: on dispose d'une famille $\{\hat{s}_m\}_{m \in \mathcal{M}}$ d'estimateurs par projection et on en choisit un, noté $\tilde{s} = \hat{s}_{\hat{m}}$, tel que

$$\hat{m} = \underset{m \in \mathcal{M}}{\text{argmin}} \ (-\|\hat{s}_m\|_n^2 + \text{pen}(m)) = \underset{m \in \mathcal{M}}{\text{argmin}} \ (\gamma_n(\hat{s}_m) + \text{pen}(m)) \ .$$

Par définition, on a donc, pour tout $m \in \mathcal{M}$,

$$\gamma_n(\tilde{s}) + \text{pen}(\hat{m}) \leq \gamma_n(\hat{s}_m) + \text{pen}(m) \leq \gamma_n(s_m) + \text{pen}(m) \ .$$

En remarquant que, pour tout $t \in \mathbb{R}^n$,

$$\gamma_n(t) = -\|s\|_n^2 + \|s - t\|_n^2 - 2\sigma\langle t, \varepsilon \rangle \ ,$$

on obtient l'inégalité équivalente suivante

$$\|s - \tilde{s}\|_n^2 - 2\sigma\langle \tilde{s}, \varepsilon \rangle + \text{pen}(\hat{m}) \leq \|s - s_m\|_n^2 - 2\sigma\langle s_m, \varepsilon \rangle + \text{pen}(m) \ .$$

Nous pouvons aussi re-écrire cela sous la forme

$$\|s - \tilde{s}\|_n^2 \leq \|s - s_m\|_n^2 + 2\sigma\langle \tilde{s} - s_{\hat{m}}, \varepsilon\rangle + 2\sigma\langle s_{\hat{m}} - s_m, \varepsilon\rangle + \text{pen}(m) - \text{pen}(\hat{m}) \ .$$

En prenant l'espérance de cette inégalité, nous sommes proche d'obtenir une inégalité oracle. Cependant, il faut que $\mathbb{E}[\langle s_{\hat{m}} - s_m, \varepsilon\rangle]$ soit "proche" de zéro (ce qui serait le cas si $\hat{m}$ était déterministe) et que le terme de pénalité en $\hat{m}$ compense $2\sigma\langle \tilde{s} - s_{\hat{m}}, \varepsilon\rangle$. Donc, pour choisir la pénalité, il nous faut comprendre comment se comporte $\sigma\langle \tilde{s} - s_{\hat{m}}, \varepsilon\rangle$. Cette variable n'est pas simple à étudier car elle est doublement aléatoire : les estimateurs $\hat{s}_m$ sont issus de procédures aléatoires, puis le choix de $\hat{m}$ aussi.

Posons $\chi_m^2 = \sigma\langle \hat{s}_m - s_m, \varepsilon\rangle = \|\hat{s}_m - s_m\|_n^2$ pour tout $m \in \mathcal{M}$. Pour contrôler $\chi_{\hat{m}}^2$, nous allons contrôler tous les χ_m^2 et ainsi, nous supprimerons le double aléa. Pour contrôler un χ_m^2 quelconque, nous allons regarder la quantité suivante, pour un $m' \in \mathcal{M}$ quelconque,

$$Z = \sup_{t \in S_{m'}} \sigma \frac{\langle t - s_m, \varepsilon\rangle}{\|t - s_m\|_n} \ .$$

La variable $\sigma\langle t - s_m, \varepsilon\rangle$ est une gaussienne centrée, de variance $\sigma^2\|t - s_m\|_n^2$. L'outil ad hoc pour étudier un tel supremum est ce que l'on appelle une **inégalité de concentration**. Ici, nous invoquons l'inégalité de concentration gaussienne suivante :

$$\mathbb{P}\left(Z - \mathbb{E}[Z] \geq \sigma\sqrt{2x}\right) \leq \exp(-x), \ \forall x > 0 \ .$$

De telles inégalités ont été beaucoup étudiées dans les dernières décennies (surtout pour les variables gaussiennes) et on pourra trouver la preuve de celle-ci sous une forme plus générale dans (3). En ce qui nous concerne, elle nous fournit la forme de la pénalité.

2.2 Cadre plus général

Le problème qui m'intéresse dans le cadre de ma thèse est un peu plus général : on observe, pour i allant de 1 à n,

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

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

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

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

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

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

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

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

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

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

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

Références

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

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

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

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