Choix du nombre de noeuds en régression spline par l'heuristique des pentes

Marie Denis, Nicolas Molinari

Choix du nombre de nœuds en régression spline par l'heuristique des pentes

Marie Denis & Nicolas Molinari

IURC, 641 avenue du doyen Gaston GIRAUD, Montpellier 34093, France

RÉSUMÉ : Le lissage par B-splines constitue un type de régression populaire grâce à ses bonnes propriétés numériques. Cependant une difficulté réside dans le choix du nombre de noeuds intérieurs nécessaire à une bonne approximation. Nous proposons une méthode de sélection de modèle via une procédure de pénalisation pour estimer ce paramètre inconnu. Le modèle choisi est celui minimisant le critère des moindres carrés pénalisé. La fonction de pénalité est estimée à partir des données en utilisant la méthode de Birgé et Massart. Cette méthode est basée sur un mélange de théorie et d'idées heuristiques, l'heuristique des pentes. On obtient ainsi un estimateur réalisant le risque quadratique minimal. Un algorithme de calibration de pénalités reposant sur une généralisation de cette heuristique sera donné. Une étude sur des simulations ainsi que sur un vrai jeux de données est effectuée pour évaluer la performance de cette méthode.

ASTRACT : The B-spline regression is appropriate since they are numerically well-conditionned. Nevertheless, the number of interiors knots is a problem. A model selection via a penalization procedure is proposed to estimate this tuning parameter. The Birgé and Massart's method proposes to design efficient penalty functions from the data. This method relies on a mixture of theoritical and heuristics ideas, it is the so-called "slope heuristics" method. The selected estimator minimizes the quadratic risk. An data-driven calibration algorithm is proposed. An experiment on simulated and real datasets is given to access to the performance of the method.

1 Régression spline et sélection de modèles via une procédure de pénalisation

La sélection de modèles via une procedure de pénalisation est une méthode très utilisée depuis plusieurs années. Elle consiste à choisir un modèle minimisant un critère défini comme la somme d'un risque empirique et d'un terme mesurant la complexité du modèle. Nous nous intéressons dans cet exposé aux modèles de régression splines. Ces modèles sont caractérisés par le degré d des splines, par le nombre k et la position des noeuds intérieurs. Posons $r = (r_1, \ldots, r_k)'$ le vecteur des k noeuds intérieurs et $\beta = (\beta_1, \ldots, \beta_{k+d+1})'$ les coefficients spline associés. Les fonctions splines appartiennent à un espace linéaire fonctionnel de dimension $d + k + 1$. La base la plus populaire pour cet espace est

donnée par les B-splines de Shoenberg, également appelée $Basic - splines$, et notée $\{B_1^d(.,\, r),\, \ldots,\, B_{d+1+k}^d(.,\, r)\}$ pour une sequence fixée de noeuds. Le lissage par B-splines constitue un type de régression populaire grâce à ses bonnes propriétés numériques. On écrit une fonction spline de la façon suivante:

$$s(x,\, \beta,\, r) \;=\; \sum_{i=1}^{d+k+1} \beta_i\, B_i^d(x,\, r). \tag{1}$$

Pour un degré fixé et un nombre maximal de noeuds, nous avons à disposition une collection finie de modèles $\mathcal{S}_m$ definie par :

$$\mathcal{S}_m = \{s(.) = \sum_{l=1}^{m} \beta_l\, B_l^d(.,\, r),\ \beta = (\beta_1,\, \ldots,\, \beta_m)' \in \mathbb{R}^m\}.$$

Poson $\mathcal{S} = \bigcup_{m\in\mathcal{M}} \mathcal{S}_m$. Notre but est de choisir le meilleur modèle et donc le bon nombre de nœuds k afin d'obtenir un risque minimal. Notre problème se résume à un choix de modèles parmi une collection finie de modèles $\{S_m,\, m \in \mathcal{M}\}$ avec $\mathcal{M}$ défini en fonction des données. On utilise, en général, le terme D_m pour représenter la complexité du modèle S_m. Dans notre contexte la complexité associée au modèle $\mathcal{S}_m$ correspond à sa dimension, c'est à dire $D_m = k + d + 1 = m$. La méthode des moindres carrés est adaptée à ce type d'estimation. Posons $\gamma(t,\, (x,\, y)) = (t(x) - y)^2$ le contraste des moindres carrés. Soit

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}(y_i - t(x_i))^2 \;=\; \|y - t\|_n^2, \tag{2}$$

la fonction de contraste empirique associée. $\|.\|_n$ correspond à la norme euclidienne normalisée sur $\mathbb{R}^n$. On construit la collection des estimateurs des moindres carrés $\{\hat{s}_m,\, m \in \mathcal{M}\}$ associée à la collection de modèles $\{\mathcal{S}_m,\, m \in \mathcal{M}\}$, où

$$\hat{s}_m = \arg\min_{t\in\mathcal{S}_m}\{\gamma_n(t)\}.$$

La fonction de perte associée est définie pour tous $s,\, t \in \mathcal{S}$ par :

$$l(s,\, t) = E[(t(X) - s(X))^2].$$

La qualité d'un estimateur $\hat{s}_m$ est donnée par son risque quadratique $E[l(s,\, \hat{s}_m)]$. On appelle modèle "oracle" le modèle présentant le plus petit risque. Cependant, s étant inconnue, il est impossible de le choisir. La solution donnée par la sélection de modèle consiste à déterminer un critère à partir des données pour sélectionner un estimateur $\tilde{s}$ ayant un risque aussi proche que possible du risque de l'oracle. Cet estimateur doit vérifier l'inégalité :

$$E[l(s,\, \tilde{s})] \leq C \inf_{m\in\mathcal{M}}\{E[l(s,\, \hat{s}_m)]\} \tag{3}$$

pour une constante non négative $C > 1$ indépendante de s. La sélection de modèle via une procédure de pénalisation consiste à choisir le modèle minimisant un critère pénalisé de la forme $crit(m) = \gamma_n(\hat{s}_m) + pen(m)$ avec $pen : \mathcal{M} \to R^+$. On note $\hat{m}$ le modèle sélectionné et $\hat{s}_{\hat{m}}$ l'estimateur associé. L'objectif est donc de choisir une fonction de pénalité qui sélectionne un estimateur satisfaisant l'inégalité (3). Un choix naturel pour évaluer la performance d'un estimateur est de considérer le rapport du risque quadratique de l'estimateur sélectionné sur le risque quadratique de l'oracle :

$$\mathcal{R}(\hat{s}_{\hat{m}}) = E[l(s, \hat{s}_{\hat{m}})] / \inf_{m \in \mathcal{M}} E[l(s, \hat{s}_m)]. \tag{4}$$

L'heuristique des "pentes"

La méthode proposée par Birgé et Massart (2007) repose sur un mélange de théorie et d'idées heuristiques. Elle permet d'estimer la fonction de pénalité à partir des données dans le cadre de la régression gaussienne homoscedastique. Cette méthode repose sur le concept de pénalité minimale. Supposons la pénalité proportionnelle à la dimension D_m (i.e. $pen(m) = K\, D_m$), Birgé et Massart détermine une constante minimale K_{min} tel que si $K < K_{min}$ le rapport (4) est asymptotiquement grand et est fini si $K > K_{min}$. Ils montrent également que lorsque $K = 2\, K_{min}$ on obtient une procédure de sélection de modèle efficace. Ils arrivent à la conclusion que la pénalité optimale est égale à deux fois la pénalité minimale. Cette relation caractérise l'heuristique des "pentes". Arlot et Massart (2008) ont récemment developpé, dans le cadre de la régression hétéroscedastique, un algorithme basé sur une généralisation de l'heuristique des pentes. Cet algorithme permet d'estimer K_{min} à partir des données. De nombreux résultats de concentrations sont également donnés. Bien que ces resultats soient prouvés dans le cadre particulier du régressogramme nous supposons qu'ils restent valides au moins dans le cadre de la régression des moindres carrés. En résumé, on peut considérer deux approches pour estimer la pénalité minimale à partir des données. Soit en utilisant l'algorithme développé par Arlot et Massart et mis en œuvre par Lebarbier (2005). Soit en estimant K_{min} par la pente de la partie linéaire de la fonction $-\gamma_n(t)$.

Essayons de comprendre la dernière approche. Considérons la décomposition classique du risque quadratique :

$$E[l(s, \hat{s}_m)] = l(s, \overline{s}_m) + E[l(\overline{s}_m, \hat{s}_m)], \tag{5}$$

où $l(s, \overline{s}_m)$ est le terme de biais qui correspond à la distance entre notre modèle et s, $E[l(\overline{s}_m, \hat{s}_m)]$ le terme de variance traduisant la complexité du modèle et $\overline{s}_m$ la projection orhogonale de s sur $\mathcal{S}_m$. Ces deux termes ont des comportements opposés quand D_m augmente. Dans le cadre de la régression spline, on prend $l(s, \hat{s}_m) = \|s - \hat{s}_m\|_n^2$, ainsi le risque quadratique se décompose de la façon suivante $E[\|s - \hat{s}_m\|_n^2] = \|s - \overline{s}_m\|_n^2 + \frac{D_m}{n}\, \sigma^2$. Comme les critères de Mallows (1973) ou de Akaike (1973, 1974), le critère de Birgé et

Massart (2007) repose sur une estimation non biaisée de ce risque. On peut donc se faire une "idée" de la pénalité optimale à partir du raisonnement suivant. Trouver le modèle minimisant le critère pénalisé $\gamma_n(\hat{s}_m) + pen(m)$ revient à trouver le minimum de $\hat{b}_m - \hat{V}_m + pen(m)$ où $\hat{b}_m = \gamma_n(\overline{s}_m) - \gamma_n(s)$ et $\hat{V}_m = \gamma_n(\overline{s}_m) - \gamma_n(\hat{s}_m)$. Grâce aux différents résultats de concentrations d'Arlot et Massart (2008), on est en mesure d'écrire que minimiser $\hat{b}_m - \hat{V}_m + pen(m)$ est approximativement équivalent à minimiser $\|s - \overline{s}_m\|_n^2 - E[\hat{V}_m] + pen(m)$. Sachant que notre but est de rendre minimal le risque $E[\|s - \hat{s}_m\|_n^2]$, une pénalité idéale pourrait être $pen(m) = E[\hat{V}_m] - E[\|\overline{s}_m - \hat{s}_m\|_n^2]$. Grâce à la principale hypothèse de l'heuristique qui consiste à supposer que $\hat{V}_m \approx \|\overline{s}_m - \hat{s}_m\|_n^2$ et aux arguments de concentrations, on peut écrire que $pen(m) \approx 2\,\hat{V}_m$. Le terme $\hat{V}_m$ peut se décomposer de la façon suivante :

$$\begin{aligned} \hat{V}_m = \gamma_n(\overline{s}_m) - \gamma_n(\hat{s}_m) &= \gamma_n(\overline{s}_m) - \gamma_n(s) + \gamma_n(s) - \gamma_n(\hat{s}_m) \\ &= \hat{b}_m + \gamma_n(s) - \gamma_n(\hat{s}_m). \end{aligned}$$

Ainsi pour de "grandes" valeurs de D_m le terme de biais $\hat{b}_m$ est constant et en supposant que la penalité est proportionnelle à D_m, le terme $\hat{V}_m$ dépend de la dimension qu'au travers du terme $-\gamma_n(\hat{s}_m)$. On obtient donc la pénalité finale $pen(m) = 2\,\hat{C}\,D_m$ où $\hat{C}$ est la pente de la partie linéaire de $-\gamma_n(\hat{s}_m)$.

Exemple

Nous observons un vecteur gaussien Y tel que $Y_i = s(x_i) + \epsilon_i$, pour $1 \leq i \leq n$ et $\epsilon_1, \ldots, \epsilon_n$ i.i.d. de loi Normale $\mathcal{N}(0, \sigma^2)$. s est une fonction spline de degré $d = 1$, avec $k = 3$ nœuds intérieurs. On prend $\sigma^2 = 1$ et $n = 500$. La courbe de la fonction $-\gamma_n(\hat{s}_m)$ en fonction de la dimension m est représentée dans la figure ci-dessous :

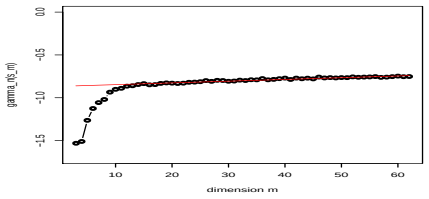

Figure 1: Graphique de $-\gamma_n(\hat{s}_m)$ en fonction de la dimension m

Par la méthode de la régression robuste (Huber, 1984), on estime la pente $\hat{C}$ de la partie linéaire. Le critère pénalisé est donc de la forme $crit(m) = \gamma_n(\hat{s}_m) + 2\,\hat{C}\,m$. On compare le critère obtenu obtenu par la méthode de l'heuristique des pentes aux critères "classiques" du BIC et du C_p de Mallow. Le ratio (4) correspondant est $\mathcal{R}_{slope} = 1.101046$,

celui obtenu avec le BIC $\mathcal{R}_{BIC} = 1.361813$ et pour le C_p de Mallow $\mathcal{R}_{C_p} = 1.851713$. Dans cet exemple, le critère défini par l'heuristique des pentes semble donc plus performant.

Bibliographie

[1] Arlot, S. and Massart, P. (2008). Data-driven calibration of penalties for least squares regression. Accepted by *Journal of Machine Learning Research*.
[2] Birgé, L., and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* 3(3), 203–268
[3] Birgé, L., and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab Theory Related Fields* 138(1-2), 33–73.
[4] Lebarbier, E. (2005). Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.* 85, 717–736.
[5] Maugis, C. and Michel, B. (2008). Slope heuristics for variable selection and clustering via Gaussian mixtures. Rapport de recherche n6550, INRIA.