Etude de cas régression non paramétrique

E. Moulines

Données

On s'intéresse dans ce travail à modéliser l'accélération de la tête d'un motard après un choc. Les données (`http://www.tsi.enst.fr/~roueff/edu/masta/non_param/Motorcycledata.txt`) contiennent des enregistrements de l'accélération en fonction du temps écoulé après l'impact. Ces données ont été analysées par [Silverman(1985)] (`http://www.tsi.enst.fr/~roueff/edu/masta/non_parama/silverman1985.pdf`). On remarquera que l'on dispose, à certains instants, de plusieurs mesures qui correspondent aux accélérations mesurées par différents capteurs. Cette difficulté (mineure) n'est pas prise en compte dans la présentation donnée ci-après. Il conviendra d'adapter les algorithmes en conséquence.

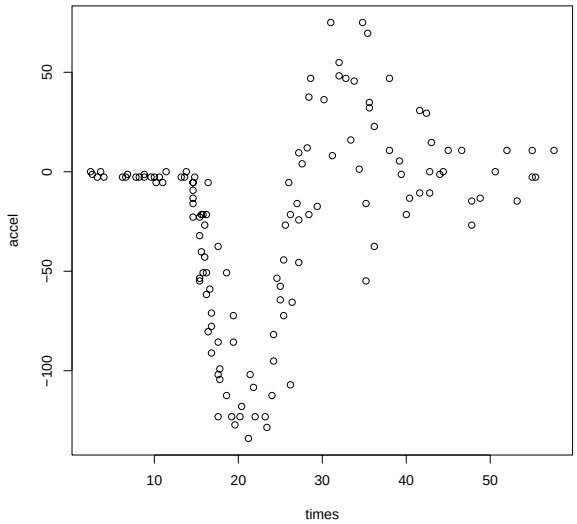

FIG. 1 – Mesure de l'accélération de la tête en fonction du temps écoulé par rapport à l'impact

Analyse exploratoire préliminaire

On modélise ces données à l'aide d'un modèle de régression non-paramétrique :

$$Y_i = g(t_i) + \epsilon_i \tag{1}$$

où Y_i sont les *observations* (l'accélération) et t_i sont les régresseurs (temps par rapport à l'impact). Dans ce travail on va chercher à évaluer différentes techniques de modélisation.

Régression polynômiale

Ajuster les coefficients d'un polynôme de degré m au sens des moindres carrés ; faire varier m et discuter.

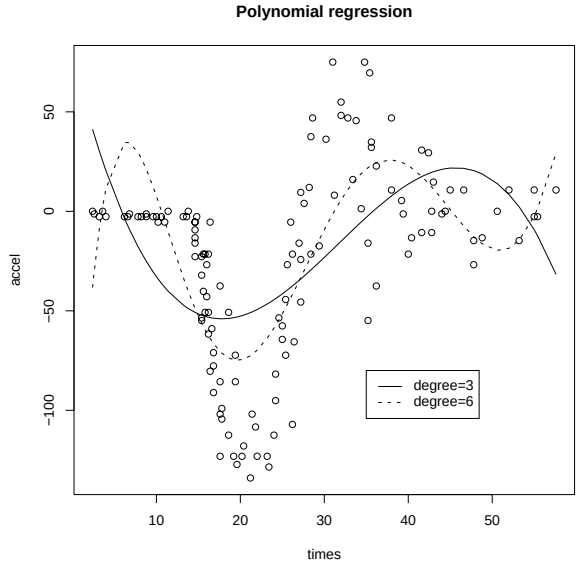

FIG. 2 – Régression non-linéaire à l'aide de polynômes de degré 3 et 6

Lissage par un méthode à noyaux

Mettre en oeuvre une méthode de lissage par noyaux. On utilisera des noyaux gaussiens et rectangulaires et on étudiera l'influence des paramètres de réglage du noyau.

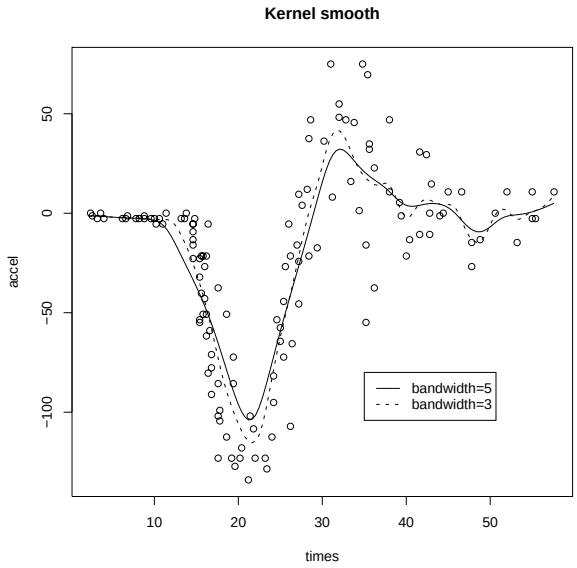

FIG. 3 – Lissage par la méthode du noyau : noyau gaussien, largeur de bande 3 et 5. Ici la largeur de bande h est définie de telle sorte que les quartiles du noyau (considéré comme une densité de probabilité) sont égaux à $\pm h/4$

Régression par splines

Rappel sur les splines cubiques

Soient $t_1 < \ldots < t_n$ n points d'un intervalle $[a, b]$. Une fonction g définie sur $[a, b]$ est une *spline cubique* si les deux conditions suivantes sont satisfaites :

1. Sur chaque intervalle (a, t_1), (t_1, t_2), $\ldots$, (t_n, b), g est un polynôme cubique ;
2. La fonction g est deux fois continûment différentiables sur $[a, b]$ (et donc g et ses dérivées d'ordre 1 et 2 sont continues aux points t_i).

Les points t_i sont appelés des *noeuds*. Il y a de nombreuses façons essentiellement équivalentes de définir des splines cubiques. La façon la plus naturelle est d'exprimer :

$$g(t) = d_i(t - t_i)^3 + c_i(t - t_i)^2 + b_i(t - t_i) + a_i, \quad t_i \leq t \leq t_{i+1} \tag{2}$$

où a_i, b_i, c_i, d_i, $i \in \{0, \ldots, n\}$ sont des constantes ; on définit dans la suite $t_0 = a$ et $t_{n+1} = b$. La continuité de g et de ses deux dérivées implique différentes relations entre les coefficients. Par exemple, la continuité de g au point t_{i+1} implique que, pour $i \in \{0, \ldots, n-1\}$,

$$d_i(t_{i+1} - t_i)^3 + c_i(t_{i+1} - t_i)^2 + b_i(t_{i+1} - t_i) + a_i = a_{i+1}$$

Une spline cubique sur l'intervalle $[a, b]$ sera dite *naturelle* si les dérivées d'ordre 2 et 3 au point a et b sont nulles.

La paramétrisation (2) n'est toutefois pas la plus facile à manipuler en pratique. Nous allons spécifier une spline cubique par ses valeurs et les valeurs de sa dérivée secondes aux noeuds t_i. Définissons

$$g_i = g(t_i) \quad \text{et} \quad \gamma_i = g''(t_i), \quad \text{pour} \quad i = 1, \ldots, n.$$

Par définition d'une spline cubique naturelle, la dérivée seconde de g est nulle aux points t_1 et t_n, de telle sorte que nous avons $\gamma_1 = \gamma_n = 0$. Notons $\mathbf{g} = (g_1, \ldots, g_n)^T$ et $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_{n-1})^T$. Les vecteurs $\mathbf{g}$ et $\boldsymbol{\gamma}$ spécifient la fonction complètement : il est possible de calculer les valeurs de la fonction et de ses dérivées en chaque point. Remarquons toutefois que des vecteurs arbitraires $\mathbf{g}$ et $\boldsymbol{\gamma}$ ne représentent pas nécessairement une splice cubique naturelle. Nous allons maintenant discuter des conditions nécessaires et suffisantes que doivent vérifier $\mathbf{g}$ et $\boldsymbol{\gamma}$ pour être associés de façon unique à une spline. Ces conditions dépendent de la donnée de deux matrices bandes Q et R définies de la façon suivante. Définissons $h_i = t_{i+1} - t_i$ pour $i = 1, \ldots, n-1$. Soit $Q = (q_{ij})$ la matrice $n \times (n-2)$ définie pour $j = 2, \ldots, n-1$ par

$$q_{j-1,j} = h_{j-1}^{-1}, q_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \quad \text{and} \quad q_{j+1,j} = h_j^{-1}$$

et $q_{ij} = 0$ pour $|i-j| \geq 2$. Les colonnes de Q sont numérotées ici de la même façon non-conventionnelle que le vecteur $\boldsymbol{\gamma}$, à savoir que la première colonne est numérotée 2 (le premier élément de la première ligne est numéroté q_{12}). Soit $R = (r_{ij})_{2 \leq i,j \leq n-1}$ la matrice $(n-2) \times (n-2)$ symétrique définie par

$$r_{ii} = \frac{1}{3}(h_{i-1} + h_i) \quad \text{for} \quad i = 2, \ldots, n-1,$$
$$r_{i,i+1} = r_{i+1,i} = \frac{1}{6}h_i \quad \text{for} \quad i = 2, \ldots, n-2,$$

et $r_{ij} = 0$ pour $|i-j| \geq 2$. La matrice R est à diagonale dominante dans le sens où $r_{ii} > \sum_{j \neq i} |r_{ij}|$ pour tout i. La matrice R est donc définie positive. Nous pouvons donc définir la matrice K par

$$K := QR^{-1}Q^T. \tag{3}$$

La propriété clef des splines cubiques est donnée par le théorème suivant ([Green and Silverman(1994), Théorème 2.1].

Théorème 1. *Les vecteurs* $\mathbf{g}$ *et* $\boldsymbol{\gamma}$ *définissent une spline cubique naturelle si et seulement si la condition suivante est vérifiée :*

$$Q^T\mathbf{g} = R\boldsymbol{\gamma} \tag{4}$$

Si cette relation est satisfaite, alors nous avons

$$\int_a^b g''(t)^2 dt = \boldsymbol{\gamma}^T R \boldsymbol{\gamma} = \mathbf{g}^T K \mathbf{g}. \tag{5}$$

Existence et unicité de la spline cubique minimisante

On étudie maintenant l'estimateur de lissage par spline, associé à la minimisation du critère :

$$S(g) = \sum_{i=1}^{n} (Y_i - g(t_i))^2 + \alpha \int (g''(x))^2 dx; \tag{6}$$

où α est un paramètre de lissage. En utilisant les résultats du paragraphe précédent, il est possible de réécrire ce vecteur en termes de vecteurs et de matrices. Nous en déduirons que g la solution de ce problème est bien unique et nous serons même capable de donner une forme explicite à cette solution. Soit $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ le vecteur des observations. Nous avons

$$\sum (Y_i - g(t_i))^2 = (\mathbf{Y} - \mathbf{g})^T (\mathbf{Y} - \mathbf{g})$$

puisque les coordonnées de $\mathbf{g}$ sont précisément les valeurs de la fonction aux noeuds. En utilisant (5), nous pouvons donc réécrire (7) sous la forme

$$\begin{aligned} S(\mathbf{g}) &= (\mathbf{Y} - \mathbf{g})^T (\mathbf{Y} - \mathbf{g}) + \alpha \mathbf{g}^T K \mathbf{g} \\ &= \mathbf{g}^T (I + \alpha K)\mathbf{g} - 2\mathbf{Y}^T \mathbf{g} + \mathbf{Y}^T \mathbf{Y}. \end{aligned} \tag{7}$$

Comme αK est symétrique et semi-définie positive, la matrice $(I + \alpha K)$ est définie positive. L'équation (7) a donc un unique minimum, donné par

$$\mathbf{g} = (I + \alpha K)^{-1} \mathbf{Y}. \tag{8}$$

Le théorème 1 montre que le vecteur $\mathbf{g}$ caractérise les splines cubiques naturelles de façon unique. Nous avons donc montré que sur l'espace des fonctions splines cubiques naturelles avec des noeuds t_i, $S(\mathbf{g})$ a un minimum unique donné par (8). Il est possible de démontrer le résultat (plus fort) suivant. Soit $\mathcal{S}_2([a, b])$ l'espace des fonctions différentiables sur $[a, b]$ avec des dérivées secondes absolument continues (*i.e.* g est continue est continue et différentiable en tout point de l'intervalle $[a, b]$ et il existe une fonction g'' intégrable telle que $\int_a^x g''t(t)dt = g'(x) - g'(a)$, pour tout $x \in [a, b]$).

Théorème 2. *Soit $n \geq 3$ et $t_1, \ldots, t_n$ des points satisfaisant $a < t_1 < \ldots < t_n < b$. Etant donné $Y_1, \ldots, Y_n$ et un paramètre de lissage α, soit $\hat{g}$ la spline cubique naturelle avec des noeuds aux points $t_1, \ldots, t_n$ telles que $\mathbf{g} = (I + \alpha K)^{-1}\mathbf{Y}$. Alors, pour toute fonction $g \in \mathcal{S}_2([a, b])$ nous avons*

$$S(\hat{g}) \leq S(g)$$

avec égalité si et seulement si g et $\hat{g}$ sont identiques.

Algorithme de Reinsch

On dit qu'une matrice à une *structure bande* si les composantes non-nulles de la matrice sont concentrées dans un "petit nombre" de diagonales ; le nombre de diagonales non nulles est appelé la "largeur de bande" de la matrice. Si B est une matrice symétrique de largeur de bande $2k + 1$, les éléments B_{ij} sont nuls dès que $|i - j| > k$. Les matrices bandes sont économiques à stocker (en matlab, voir l'aide sur *sparse matrix* et *Mathematics : Sparse Matrices : Creating Sparse Matrices*). Les matrices Q et R définies dans les paragraphes précédents ont des largeurs de bande égales à 3. La solution du problème de minimisation (7) est donnée par

$$(I + \alpha Q R^{-1} Q^T)\mathbf{g} = \mathbf{Y}. \tag{9}$$

En réarrangeant cette relation, nous pouvons écrire de façon équivalente

$$\mathbf{g} = \mathbf{Y} - \alpha Q R^{-1} Q^T \mathbf{g}.$$

En utilisant la relation (4), et en simplifiant la relation, nous obtenons ainsi une expression explicite de $\mathbf{g}$ en fonction de $\mathbf{Y}$ et de $\boldsymbol{\gamma}$,

$$\mathbf{g} = \mathbf{Y} - \alpha Q \boldsymbol{\gamma}. \tag{10}$$

En utilisant encore la relation (4), nous obtenons

$$Q^T \mathbf{Y} - \alpha Q^T Q \boldsymbol{\gamma} = R \boldsymbol{\gamma},$$

qui donne une équation pour $\boldsymbol{\gamma}$,

$$(R + \alpha Q^T Q)\boldsymbol{\gamma} = Q^T \mathbf{Y}. \tag{11}$$

Cette relation est le coeur de l'algorithme. Par rapport à la relation (9), cette relation peut-être résolue en temps linéaire en utilisant les méthodes liées aux matrices bandes.

On remarque en effet que $(R + \alpha Q^T Q)$ est une matrice symétrique définie positive de largeur de bande 5. Cette matrice a donc une décomposition de Cholesky de la forme

$$R + \alpha Q^T Q = L D L^T,$$

où D est une matrice diagonale (dont les éléments diagonaux sont strictement positifs) et L est une matrice triangulaire inférieure avec $L_{ij} = 0$ pour $j < i - 2$ et $j > i$ et $L_{ii} = 1$ pour tout i. Le calcul de cette décomposition requiert de l'ordre de $O(n)$ opérations (à comparer avec la complexité $O(n^3)$ associée habituellement à cette décomposition pour une matrice complète). Nous pouvons synthétiser l'algorithme de Reinsch [Reinsch(1967)] ci-dessous

Step 1 Calculer $Q^T \mathbf{Y}$,

Step 2 Calculer les éléments non-nuls de la matrice $R + \alpha Q^T Q$ et les facteurs de Cholesky L et D (en utilisant la représentation par matrice sparse)

Step 3 Ecrire (11) sous la forme $LDL^T \boldsymbol{\gamma} = Q^T \mathbf{Y}$ et résoudre cette équation en $\boldsymbol{\gamma}$ par substitution avant et arrière (utiliser la résolution de système linéaire de Matlab ; MATLAB *Mathematics : Sparse Matrices : Simultaneous Linear Equations*)

Step 4 En utilisant (10), déterminer $\mathbf{g}$,

$$\mathbf{g} = \mathbf{Y} - \alpha Q \boldsymbol{\gamma}.$$

Choix du paramètre de lissage

Considérons l'observation (Y_i, t_i) comme une nouvelle observation en la retirant de l'ensemble des données utilisées pour déterminer la courbe. Notons par $\hat{g}^{(-i)}(t; \alpha)$ la courbe estimée à partir des autres données en utilisant α comme paramètre de lissage, *i.e* $\hat{g}^{(-i)}$ est la solution du problème de minimisation

$$\sum_{j \neq i} (Y_j - g(t_j))^2 + \alpha \int (g'')^2. \tag{12}$$

La qualité de $\hat{g}^{(-i)}$ pour "prédire" l'observation (Y_i, t_i) peut être évaluée en déterminant comment Y_i prédit Y_i. La performance de la procédure peut donc être évaluée en déterminant le score de validation croisé

$$CV(\alpha) = n^{-1} \sum_{i=1}^{n} (Y_i - \hat{g}^{(-i)}(t_i; \alpha))^2. \tag{13}$$

L'idée de base de la validation croisée est de choisir la valeur du paramètre de lissage α qui minimise le score de validation croisé. Il n'est pas possible de garantir que le score de validation croisé n'ait pas de minima locaux et la solution la plus simple consiste à évaluer ce score sur une grille de points. Quelle que soit la méthode de minimisation retenue, la minimisation de $\alpha \mapsto CV(\alpha)$ nécessite d'évaluer la fonction $CV(\alpha)$ en un grand nombre de points et il est donc très important de disposer d'une méthode efficace de calcul de $CV(\alpha)$. Il peut sembler indispensable de résoudre, pour chaque valeur du paramètre de lissage α, n problèmes de lissage pour déterminer n courbes hgi. Heureusement, comme nous allons le voir ci-dessous, une solution beaucoup plus simple s'offre à nous.

Rappelons tout d'abord que les valeurs de la spline de lissage optimale dépend linéairement des observations

$$\mathbf{g} = A(\alpha)\mathbf{Y} \tag{14}$$

où la matrice $A(\alpha)$ est donnée par :

$$A(\alpha) = (I + \alpha Q R^{-1} Q)^{-1}. \tag{15}$$

La matrice $A(\alpha)$ est appelée la *matrice chapeau*. Cette matrice permet de relier le vecteur des observations Y_i au vecteur des prédicteurs $\hat{Y}_i$. Le résultat clef est donné dans le théorème suivant :

Théorème 3. *Le score de validation croisé vérifie*

$$CV(\alpha) = n^{-1} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{g}(t_i)}{1 - A_{ii}(\alpha)} \right)^2$$

où $\hat{g}$ est la spline d'interpolation calculée à partir de l'ensemble complet $\{(t_i, Y_i)\}$ avec le paramètre de régularisation α.

Démonstration. La preuve du théorème découle du lemme suivant :

Lemme 4. *Soit α et $i \in \{1, \ldots, n\}$ donnés. Notons $\mathbf{g}^{(-i)}$ le vecteur de composants $\mathbf{g}_j^{(-i)} = \hat{g}^{(-i)}(t_j; \alpha)$. Définissons $\mathbf{Y}^*$ le vecteur*

$$Y_j^* = Y_j \quad \textit{pour} \quad j \neq i$$
$$Y_i^* = \hat{g}^{(-i)}(t_i; \alpha).$$

Alors

$$\mathbf{g}^{(-i)} = A(\alpha)\mathbf{Y}^*. \tag{16}$$

Démonstration. Pour tout $g \in \mathcal{S}_2([a,b])$ nous avons

$$\begin{aligned}\sum_{j=1}^{n}\{Y_j^*-g(t_j)\}^2+\alpha\int(g'')^2 &\geq \sum_{j\neq i}\{Y_j^*-g(t_j)\}^2+\alpha\int(g'')^2\\ &\geq \sum_{j\neq i}\{Y_j^*-\hat{g}^{(-i)}(t_j)\}^2+\alpha\int(\hat{g}^{-(i)''})^2\\ &\geq \sum_{j=1}^{n}\{Y_j^*-\hat{g}^{(-i)}(t_j)\}^2+\alpha\int(\hat{g}^{-(i)''})^2\end{aligned}$$

Par conséquent $\hat{g}^{(-i)}$ minimise $\sum_{j=1}^{n}\{Y_j^*-\hat{g}^{(-i)}(t_j)\}^2+\alpha\int(g'')^2$ et donc $\mathbf{g}^{(-i)}=A(\alpha)\mathbf{Y}^*$. □

Nous pouvons déduire de ce résultat une expression pour le résidu $Y_i-\hat{g}^{(-i)}(t_i)$. Nous avons, en posant $A=A(\alpha)$,

$$\begin{aligned}\hat{g}^{(-i)}(t_i)-Y_i=\sum_{j=1}^{n}A_{ij}Y_j^*-Y_i &= \sum_{j\neq i}A_{ij}Y_j+A_{ii}\hat{g}^{(-i)}(t_i)-Y_i\\ &=\sum_{j=1}^{n}A_{ij}Y_j-Y_i+A_{ii}\{\hat{g}^{(-i)}(t_i)-Y_i\}\\ &=\hat{g}(t_i)-Y_i+A_{ii}\{\hat{g}^{(-i)}(t_i)-Y_i\}.\end{aligned}$$

Par conséquent,

$$Y_i-\hat{g}^{(-i)}(t_i)=\frac{Y_i-\hat{g}(t_i)}{1-A_{ii}(\alpha)}$$

□

Ce théorème montre que, dès que les coefficients diagonaux de la matrice $A(\alpha)$ sont connus, les scores de validation croisés peuvent être déterminés à partir des résidus de prédiction $Y_i-\hat{g}(t_i)$ obtenus à l'aide de la spline d'interpolation optimale déterminée sur l'ensemble du jeu de données. Par conséquent, il n'est pas nécessaire de résoudre d'autres problèmes d'interpolation. Remarquons si l'on s'en tient à (15), le calcul des éléments diagonaux de la matrice A est complexe, car il nécessite l'inversion d'une matrice de grande taille (que nous avions évitée en utilisant l'algorithme de Reinsch pour le calcul de la spline d'interpolation). Il est fort heureusement possible de calculer ces éléments au moyen d'un algorithme de complexité $O(n)$, en utilisant un algorithme du à [Hoog and Hutchinson(1985)]. Comme l'algorithme de Reinsch est lui aussi linéaire, l'algorithme global de détermination du score de validation croisé $CV(\alpha)$ est donc $O(n)$.

Diagonale principale de l'inverse d'une matrice bande Soit B une matrice symétrique définie positive de largeur de bande 5. Décomposons $B=LDL^T$ où L est triangulaire (bande) inférieure

à diagonale unité et D est diagonale. Notons $\bar{b}_{ij}$ les éléments de B^{-1}. Par définition nous avons

$$B^{-1} = L^{-T} D^{-1} L^{-1}$$

ce qui implique

$$L^T B^{-1} = D^{-1} L^{-1},$$

d'où nous déduisons

$$B^{-1} = D^{-1}L^{-1} + B^{-1} - L^T B^{-1} = D^{-1}L^{-1} + (I - L^T)B^{-1}.$$

La matrice L^{-1} est triangulaire inférieure à diagonale unité et donc $D^{-1}L^{-1}$ est triangulaire inférieure ; de plus, les éléments diagonaux de la matrice $D^{-1}L^{-1}$ sont donnés par d_i^{-1}. De plus, $(I - L^T)$ est triangulaire supérieure à diagonale nulle. Comme la matrice B est symétrique, B^{-1} est aussi symétrique et nous avons donc $\bar{b}_{ij} = \bar{b}_{ji}$ pour tout $j > i$, ce qui implique que pour $i = 1, \ldots, n-2$,

$$\begin{aligned}
\bar{b}_{i,i} &= d_i^{-1} - L_{i+1,i}\bar{b}_{i,i+1} - L_{i+2,i}\bar{b}_{i,i+2}, \\
\bar{b}_{i,i+1} &= -L_{i+1,i}\bar{b}_{i+1,i+1} - L_{i+2,i+1}\bar{b}_{i+1,i+2}, \\
\bar{b}_{i,i+2} &= -L_{i+1,i}\bar{b}_{i+2,i+2} - L_{i+2,i}\bar{b}_{i+2,i+2}
\end{aligned}$$

et

$$\begin{aligned}
\bar{b}_{n,n} &= d_n^{-1}, \\
\bar{b}_{n-1,n} &= -L_{n,n-1}\bar{b}_{n,n}, \\
\bar{b}_{n-1,n-1} &= d_{n-1}^{-1} - L_{n,n-1}\bar{b}_{n-1,n}.
\end{aligned}$$

Il faut appliquer ces formules dans un ordre approprié pour déterminer les éléments des 5 diagonales centrales de B^{-1}. On commence l'itération en déterminant $\bar{b}_{n,n}$ (qui dépend uniquement de l'élément diagonal d_n de la décomposition de Cholesky de B). On poursuit l'itération en calculant $\bar{b}_{n-1,n}$ puis $\bar{b}_{n-1,n-1}$ et ensuite pour $i = n-2, \ldots, 1$, on calcule $\bar{b}_{i,i+2}$, $\bar{b}_{i,i+1}$ et $\bar{b}_{i,i}$.

Expression de la matrice chapeau Rappelons que l'algorithme de Reinsch revient à évaluer $\boldsymbol{\gamma}$ et $\mathbf{g}$ à l'aide des formules suivantes

$$\boldsymbol{\gamma} = (R + \alpha Q^T Q)^{-1} Q^T \mathbf{Y}$$

et

$$\begin{aligned}
\mathbf{g} &= \mathbf{Y} - \alpha Q \boldsymbol{\gamma} = \mathbf{Y} - \alpha Q (R + \alpha Q^T Q)^{-1} Q^T \mathbf{Y} \\
&= \{I - \alpha Q (R + \alpha Q^T Q)^{-1} Q^T\} \mathbf{Y}
\end{aligned}$$

Ceci montre que la matrice chapeau $A(\alpha)$ peut s'écrire sous la forme

$$A(\alpha) = I - \alpha Q (R + \alpha Q^T Q)^{-1} Q^T$$

de telle sorte que

$$I - A(\alpha) = \alpha Q(R + \alpha Q^T Q)^{-1} Q^T.$$

Posons $B = (R + \alpha Q^T Q)$. B a une largeur de bande égale à 5. Notons $\bar{b}_{ij}$ les éléments de B^{-1}. Comme la matrice Q est tridiagonale, nous avons

$$(QB^{-1}Q^T)_{ii} = q_{i,i-1}^2 \bar{b}_{i-1,i-1} + q_{ii}^2 \bar{b}_{ii} + q_{i,i+1}^2 \bar{b}_{i+1,i+1} + 2q_{i,i-1}q_{i,i}\bar{b}_{i-1,i} + 2q_{i,i-1}q_{i,i+1}\bar{b}_{i-1,i+1} + 2q_{i,i}q_{i,i+1}\bar{b}_{i,i+1}.$$

On voit clairement que seuls les éléments $\bar{b}_{ij}$ pour $|i-j| \leq 2$ sont requis pour calculer les éléments diagonaux de $QB^{-1}Q^T$ et donc les valeurs de $1-A_{ii}(\alpha)$ requis pour calculer le score de validation croisé.

Validation croisée généralisée

La méthode de validation croisée généralisée est une forme "simplifiée" de la validation croisée, très utilisée en pratique car plus simple à mettre en oeuvre. L'idée de base de la validation croisée généralisée est de remplacer le facteur de normalisation $(I - A_{ii}(\alpha))$ par le facteur moyen $1 - n^{-1}\mathrm{tr}A(\alpha)$. Le score de validation croisée généralisé est construit, par analogie avec le score de validation ordinaire, en sommant le carré des résidus et en le pondérant par le carré de l'inverse de $1 - n^{-1}\mathrm{tr}A(\alpha)$,

$$GCV(\alpha) = n^{-1} \frac{\sum_{i=1}^n (Y_i - \hat{g}(t_i))^2}{\{1 - n^{-1}\mathrm{tr}(A(\alpha))\}^2} \tag{17}$$

Comme pour la validation croisée ordinaire, on détermine le paramètre de lissage en minimisant GCV.

Une des raisons principales de s'intéresser à la validation croisée généralisée est que cette méthode est généralement plus simple à mettre en oeuvre. En utilisant des expressions équivalentes de la trace de la matrice $A(\alpha)$ il est en fait possible de déterminer la trace de cette matrice sans avoir à évaluer les éléments diagonaux. Notons ω_ν les valeurs propres de la matrices $QR^{-1}Q^T$; comme $A(\alpha) = (I + \alpha QR^{-1}Q^T)^{-1}$ les valeurs propres de $A(\alpha)$ sont égales à $(1 + \alpha\omega_\nu)^{-1}$ et par conséquent

$$n\{1 - n^{-1}\mathrm{tr}A(\alpha)\}^2 = n\left(1 - n^{-1}\sum_{\nu=1}^n (1 + \alpha\omega_\nu)^{-1}\right)^2$$

Ainsi, une fois que les valeurs propres ω_ν sont connues, le score GCV peut être calculé pour toutes les valeurs du paramètre de lissage α en évaluant une formule très simple.

Questions

Q-1 Etablir la preuve du théorème 1 ;

Q-2 Démontrer (8) ;

Q-3 Supposons que l'on dispose, pour la valeur t_i de l'instant de régression, de m_i observations, Y_{ij}, $j \in \{1, \ldots, m_i\}$. Notons $\bar{Y}_i = m_i^{-1} \sum_{j=1}^{m_i} Y_{ij}$. Soit $S(g)$ la somme pénalisée associée à l'ensemble initial

$$S(g) = \sum_i \sum_j \{Y_{ij} - g(t_i)\}^2 + \alpha \int (g'')^2;$$

Montrer que ce problème est équivalent à minimiser la somme pondérée

$$S(g) = \sum_i m_i \{\bar{Y}_i - g(t_i)\}^2 + \alpha \int (g'')^2.$$

Q-4 Soit $w_1, \ldots, w_n$ des poids positifs ; considérons le critère pondéré

$$\sum_{i=1}^{n} w_i \{Y_i - g(t_i)\}^2 + \alpha \int (g'')^2$$

Enoncer et démontrer un analogue du théorème 2 ; Proposer un algorithme analogue à l'algorithme de Reinsch pour résoudre ce problème ;

Q-5 On définit le risque de validation croisé par

$$CV(\alpha) = \sum_{i=1}^{n} w_i \{Y_i - \hat{g}^{(-i)}(t_i, \alpha)\}^2$$

Montrer que

$$CV(\alpha) = \sum_{i=1}^{n} w_i \left(\frac{Y_i - \hat{g}(t_i)}{\{I - A_W(\alpha)\}_{ii}} \right)^2$$

où $A_W(\alpha) = (W + \alpha Q R^{-1} Q)^{-1} W$. Proposer une adaptation de l'algorithme de [Hoog and Hutchinson(1985)] pour le calcul du coût de validation croisé ;

Q-6 Le score naturel de validation croisé en présence d'observations multiples s'écrit

$$CV_T(\alpha) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \{Y_{i,j} - \hat{g}^{(-ij)}(t_i)\}^2$$

où $N = \sum_{i=1}^{n} m_i$ est le nombre total d'observations et $\hat{g}^{(-ij)}$ est la spline de d'interpolation optimale pour les observations dont on a omis (Y_{ij}, t_i) ; montrer que

$$Y_{ij} - \hat{g}^{(-ij)}(t_i) = \frac{Y_{ij} - \hat{g}(t_i)}{1 - m_i^{-1}(A_W)_{ii}}$$

et que le score de validation croisé peut s'écrire :

$$CV_T(\alpha) = N^{-1} \sum_{i=1}^{n} \frac{m_i \{\bar{Y}_i - \hat{g}(t_i)\}^2 + S_i^2}{\{1 - m_i^{-1}(A_W)_{ii}\}^2}$$

où $S_i^2 = \sum_{j=1}^{m_i} \{Y_{ij} - \bar{Y}_i\}^2$.

Q-7 Implémenter la méthode de validation croisée généralisée ; comparer ;

Q-8 Analyser les résidus de prédiction en s'inspirant de la discussion donnée dans [Silverman(1985)].

Références

[Green and Silverman(1994)] GREEN, P. and SILVERMAN, B. (1994). *Non parametric regression and generalized linear models*. Monographs on Statistics and Applied Probability, Chapman and Hall, London. Au département.

[Hoog and Hutchinson(1985)] HOOG, F. D. and HUTCHINSON, M. (1985). Smoothing noisy data with spline functions. *Numer. Math.* **47** 99–106.

[Reinsch(1967)] REINSCH, C. (1967). Smoothing by spline function. *Numer. Math.* **10** 177–183.

[Silverman(1985)] SILVERMAN, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression fitting. *Journal of the Royal Statistical Society, Series B* **47** 1–52.