C. R. Acad. Sci. Paris, Ser. I 354 (2016) 538–542

Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistique

Modèle non paramétrique parcimonieux pour la détection des points d'impact d'une variable fonctionnelle

Nonparametric selection of impact points in functional regression

Germán Aneiros [a], Philippe Vieu [b]

[a] *Universidad de A Coruña, Spain*
[b] *Institut de Mathématiques, Toulouse, France*

I N F O A R T I C L E

Historique de l'article :
Reçu le 17 juin 2015
Accepté après révision le 25 janvier 2016
Disponible sur Internet le 16 mars 2016

Présenté par Paul Deheuvels

R É S U M É

Dans un problème de régression avec variable explicative fonctionnelle, on s'intéresse à la sélection des points les plus informatifs. Un modèle parcimonieux de type non paramétrique ainsi qu'une procédure de choix de variables basée sur une pré-sélection par dépistage sont proposés, et des résultats asymptotiques sont établis concernant à la fois la sélection des points informatifs et l'estimation des paramètres du modèle.

A B S T R A C T

A nonlinear sparse model is defined for selecting impact points in regression problems with functional predictors, and a variable selection procedure based on screening and splitting is proposed. Some asymptotics are stated both for the impact points and for the parameters of the model.

Abridged English version

Let Y be a scalar response and $\chi \sim \{\chi(t), t \in (0,T)\}$ a functional predictor. The aim is to search for the points in $(0,T)$ having most impact on Y. Hence, we consider a sparse nonparametric additive model:

$$Y = \sum_{j=1}^{p_n} f^j(X^j) + \epsilon, \text{ with } E(\epsilon) = 0 \text{ and } E(\epsilon^2) < \infty,$$

where $X^j = \chi(t_j)$ and $t_1, \ldots, t_{p_n}$ are equispaced points at which the curves have really been observed, and where the components f^j of the model are only supposed to satisfy some general smoothness condition. The idea of the method is based on a splitting of the variables into the following blocks (here $w_n q_n = p_n$):

Adresses e-mail : ganeiros@udc.es (G. Aneiros), vieu@math.univ-toulouse.fr (P. Vieu).

http://dx.doi.org/10.1016/j.crma.2016.01.019

$$\{1,\dots,p_n\}=\cup_{k=1}^{w_n}E_k \text{ for } E_k=\{(2k-1)q_n/2+m, m=-\frac{q_n}{2}+1,\dots,\frac{q_n}{2}\}.$$

Then the algorithm can be summarized in three parts:

i) choose a pilot variable selection technique;
ii) screening: use the pilot procedure to select variables only among $\{\tilde{X}^k=X^{((2k-1)q_n+1)/2}\}_{k=1}^{w_n}$;
iii) sharpening: search neighbours of the firstly selected variables having to be added to the model.

Based on the ideas of [21], we use a splitting stage for dividing the data into two independent subsamples: one being used for screening and the other one for sharpening. Asymptotics are given without need for fixing any pilot procedure along step i), and they allow one to include usual ones (such as the Adaptive Grouped Lasso of [13], which will be discussed with more details). This Note extends in a nonparametric way the earlier literature on point impact selection [17,14,19,1,2]. It is the first step of a more general work (see [3]) that will include more complete theory and applied issues.

1. Introduction

On considère un modèle de régression dans lequel la variable réponse Y est réelle et la variable explicative χ est une fonction à valeurs rélles, $\chi\sim\{\chi(t),t\in(0,T)\}$. Les développements récents en statistique fonctionnelle font état de nombreuses techniques d'estimation de l'opérateur de régression, linéaires [7] ou non paramétriques [9]. Dans le contexte non paramétrique qui sera le notre dans cette Note, les principales techniques sont basés sur les idées de noyau [9], de plus proches voisins [16,15], de réseaux de neurones [18] ou de minimax [5]. Le lecteur trouvera de plus amples détails au travers de nombreux ouvrages généraux récemment parus [11,20,12]. La question qui est abordée ici n'est pas simplement d'ordre quantitatif (prédiction de Y), mais aussi d'ordre qualitatif, en ce sens que l'objectif est aussi de déterminer les points de l'intervalle $(0,T)$ qui sont les plus informatifs pour expliquer Y.

En pratique, la variable χ n'est observée qu'en un nombre fini de points $0<t_1<\ldots<t_{p_n}<T$. Dans bon nombre de problèmes, les échantillons sont équilibrés (en ce sens que $t_{j+1}-t_j=T/p_n$), mais les résultats présentés ici sont valables sous l'hypothèse générale :

$$\exists(a,b),0<a<b<\infty,\forall j=1,\dots,p_n-1,a/p_n\le t_{j+1}-t_j\le b/p_n. \tag{1}$$

La recherche de points informatifs (parfois appelés points d'impact) se ramène, si on note

$$X^j=\chi(t_j)\text{ et }\mathbf{X}=(X^1,\dots,X^{p_n})^t, \tag{2}$$

à un problème de choix de variables dans lequel la dimension p_n est largement supérieure à la taille de l'échantillon. L'abondante littérature sur ce sujet [4] fait en général état de méthodes qui ne sont pas de nature à prendre en compte la spécificité fonctionnelle de notre problème, qui se traduit en particulier par une très forte corrélation entre les variables X^j. La nécessité d'adapter au cadre fonctionnel les techniques de choix de variables a été mise en évidence dans [8], et l'importance du problème a été soulignée par plusieurs auteurs (voir le §6 dans [6]). Quelques travaux récents ont abordé ce sujet [17,14,19,1,2] au prix de restrictions fortes (essentiellement linéaires) sur le modèle. L'objet de cette Note est de proposer une modélisation non paramétrique du problème, de développer une procédure de sélection de points d'impact et d'en étudier le comportement asymptotique.

2. Le modèle

Nous proposons un modèle additif

$$Y=\sum_{j=1}^{p_n}f^j(X^j)+\epsilon,\ \text{avec } E(\epsilon)=0\text{ et }E(\epsilon^2)<\infty. \tag{3}$$

Les points d'impact sont modélisés par une condition de parcimonie

$$s_n=\#S_0=\#(\{j=1,\dots,p_n,\ f^j\neq 0\})=o(p_n). \tag{4}$$

L'aspect non paramétrique du modèle consiste à ne faire que des hypothèses très générales sur les paramètres du modèle :

$$|f^j(x)-f^j(x')|\le C|x-x'|,\forall x,x'\in I, \tag{5}$$

$$\min_{j\in S_0}\inf_{x\in I}|f^j(x)|w_n\to\infty, \tag{6}$$

I étant l'intervalle dans lequel $\chi(t)$ prend ses valeurs et $\{w_n\}$ étant une suite entière tendant vers l'infini. Ces conditions (4), (5) et (6) apparaissent naturellement dans la littérature en matière de sélection non paramétrique de variables (voir par exemple [13]) et ne sont donc en aucun cas liées au contexte fonctionnel qui fait la spécificité de cette Note. Afin de

prendre en compte le lien fonctionnel existant entre les variables X^j, nous introduisons les trois conditions supplémentaires suivantes :

$$|\chi(t)-\chi(t')|\leq C|t-t'|, \forall t,t'\in(0,T), \tag{7}$$

$$\forall j,k\in S_0, sup_{x\in I}|f^j(x)-f^k(x)|\leq C|t_j-t_k|. \tag{8}$$

Enfin, en écrivant $\{1,\dots,p_n\}=\cup_{k=1}^{w_n}E_k$ pour $E_k=\{kq_n/2+m, m=-\frac{q_n}{2}+1,\dots,\frac{q_n}{2}\}$, et en notant $F^k(\cdot)=\sum_{j\in E_k}f^j(\cdot)$, $S_k=\{j\in E_k, f^j\neq 0\}$ et $q_n=p_n/w_n$, on suppose que

$$\forall k=1,\dots,w_n,\ F^k\neq 0\Rightarrow \exists 0<c<a_k, \#S_k\sim a_k q_n. \tag{9}$$

Alors que les conditions de régularité (7) et (8) sont très naturelles, la condition (9) est spécifique au problème abordé dans cette Note et sert à modéliser le fait que, si un point t_j est influent sur la variable Y, alors d'autres parmi ses voisins le sont aussi ; bien entendu, ce type de condition n'aurait aucun sens si le vecteur $\mathbf{X}$ n'était pas issu d'un phénomène continu tel que (2).

3. Choix de variables et méthode d'estimation

Dans les questions de choix de variables en très grande dimension, il est habituel de procédér en deux étapes : on élimine dans un premier temps (de manière grossière) un grand nombre de variables avant d'affiner les résultats en utilisant des méthodes plus élaborées sur le nouvel ensemble de variables ainsi construit. Ces méthodes sont connues sous le nom de dépistage, et leur validité est assujettie au fait d'avoir préalablement divisé l'échantillon statistique en deux parties indépendantes [21]. Ainsi, nous supposerons disposer d'un échantillon indépendant ainsi divisé (prenons pour simplifier que $n_1=n/2$) :

$$\{(\chi_i,Y_i), i=1,\dots,n_1\} \text{ et } \{(\chi_i,Y_i), i=n_1+1,\dots,n\}.$$

La méthode de dépistage que nous allons présenter est adaptée au cadre fonctionnel qui nous concerne.

Préliminaire : choix d'une technique pilote. Afin d'ouvrir un maximum de possibilités d'applications, cette Note est présentée de manière générale à partir d'une procédure pilote qui volontairement ne sera pas spécifiée (voir discussion au Paragraphe 5). Supposons donc que nous disposons, pour tout modèle parcimonieux additif (Z et W^j étant des v.a.r. quelconques)

$$Z=\sum_{j\in J}\phi^j(W^j)+\epsilon \text{ et } \#(J_0)=\#(\{j\in J,\phi^j\neq 0\})=o(\#(J)),$$

d'une procédure permettant d'obtenir des estimateurs $\tilde{\phi}^j$ et de procéder à une sélection de variables :

$$\tilde{J}_0=\{j\in J,\tilde{\phi}^j\neq 0\}.$$

En général la qualité d'une telle méthode est établie au moyen de deux types de résultats, les vitesses de convergence des estimateurs des paramètres du modèle (ici U_n et V_n sont des suites positives qui dépendent de la méthode pilote employée) :

$$\|\tilde{\phi}^j-\phi^j\|_2^2=O_p(U_nV_n(\#(J))) \text{ uniformément pour } j\in\tilde{J}_0, \tag{10}$$

et la capacité de la méthode à retrouver les vraies variables informatives

$$P(\tilde{J}_0=J_0)\to 1 \text{ quand } n\to\infty. \tag{11}$$

Etape 1 : Dépistage. Partant du découpage défini ci-dessus (voir (9)), l'étape de dépistage consiste à appliquer la méthode pilote sur le modèle

$$Y=\sum_{k=1}^{w_n}g^k(\tilde{X}^k)+\epsilon, \tag{12}$$

construit à partir des variables isolées $\{\tilde{X}^k=X^{((2k-1)q_n+1)/2}\}_{k=1}^{w_n}$, et en n'utilisant que le premier sous-échantillon. À l'issue de cette première étape, nous disposons d'estimateurs $\tilde{g}^k$ des composantes du modèle (12) ainsi que d'un pemier tri des variables informatives $\tilde{S}_0=\{((2k-1)q_n+1)/2,\ k\in\tilde{K}_0\}$ où $\tilde{K}_0=\{k=1,\dots,w_n,\tilde{g}^k\neq 0\}$.

Etape 2 : Affinage. Afin d'éliminer l'effet du choix arbitraire de la décomposition (9), la deuxième étape de l'algorithme consiste à chercher celles parmi les variables voisines de celles déjà pré-sélectionnées qui pourraient être elles aussi informatives. Concrètement, la procédure pilote est utilisée une seconde fois, en n'utilisant que le deuxième sous-échantillon, sur le nouveau modèle

$$Y = \sum_{j\in E} s^j(X^j) + \epsilon \text{ où } E = \cup_{k\in\tilde{K}_0} E_k, \tag{13}$$

amenant à des estimateurs $\tilde{s}^j$ des composantes s^j et à un nouveau tri de variables $\tilde{\tilde{S}}_0 = \{j \in E, \tilde{s}^j \neq 0\}$.

Bilan. Dans le modèle (3) les points d'impact sont $\hat{S}_0 = \tilde{\tilde{S}}_0$, et les composantes estimées sont $\hat{f}^j = \tilde{s}^j 1_{j\in E}$.

4. Résultats asymptotiques

Théorème 4.1. *Sous le modèle défini par les conditions (1)–(9), et si la procédure pilote satisfait les conditions (10) et (11), alors on a :*

$$\sum_{j=1}^{p_n} \|\hat{f}^j - f^j\|_2^2 = O_p(s_n U_n V_n(s_n)) \quad \text{et} \quad P(\hat{S}_0 = S_0) \to 1. \tag{14}$$

Grandes lignes de la preuve du Théorème 4.1. On utilise la décomposition

$$\sum_{j=1}^{p_n} \|\hat{f}^j - f^j\|_2^2 = \sum_{j\in E} \|\hat{f}^j - f^j\|_2^2 + \sum_{j\in\overline{E}\cap S_0} \|\hat{f}^j - f^j\|_2^2. \tag{15}$$

On montre successivement, en utilisant (8) puis la propriété (10) de la procédure pilote, que

$$\forall \eta_0 > 0,\ P\Big(\sum_{j\in\overline{E}\cap S_0} \|\hat{f}^j - f^j\|_2^2 \geq \eta_0 s_n U_n V_n(s_n)\Big) \to 0, \tag{16}$$

$$\sum_{j\in E} \|\hat{f}^j - f^j\|_2^2 = O_p(s_n U_n V_n(s_n)). \tag{17}$$

La première partie de (14) vient de (15)–(17). Pour la seconde on montre que $P(S_0 \nsubseteq \hat{S}_0) \to 0$ et $P(\hat{S}_0 \nsubseteq S_0) \to 0$ (ce dernier point venant de (11)). Les preuves complètes sont disponibles sur demande. □

5. Commentaires

À propos de la procédure pilote et des vitesses de convergence. Les exemples de procédures satisfaisant les conditions (10) et (11) sont nombreux. Considérons le cas particulier du lasso groupé adaptatif (voir [13]) dans le cas simple où $s_n = s$ et $w_n = \sqrt{p_n}$. Les résultats de [13] ne nécessitent que deux hypothèses clés sur le modèle : une condition de régularité (du type de (5)) et une autre sur la taille du modèle, et sont donc directement utilisables dans notre contexte. Ainsi, il découle de [13] que, sous la condition (5), les conditions (10) et (11) sont satisfaites avec $U_n = n^{-2/3}$ et $V_n(\#J) = 1$, et ce dès que $\log(\#(J)) = o(n^{2/3})$. Les résultats du Théorème 4.1 se traduisent alors par

$$\sum_{j=1}^{p_n} \|\hat{f}^j - f^j\|_2^2 = O_p(n^{-2/3}),$$

sous une condition de dimension

$$p_n = (e^{o(n^{2/3})})^2.$$

À titre de comparaison (voir [13]), une application directe de la méthode du lasso groupé adaptatif nécessite une condition $p_n = e^{o(n^{2/3})}$. Ceci est loin d'être négligeable tant le nombre de variables intervenant dans ce type de problèmes est en général très important (souvent de l'ordre du millier, voire plus).

À propos des aspects calculatoires. De par sa nature, notre procédure en deux étapes va s'avérer bien moins coûteuse en temps de calcul qu'une méthode directe travaillant directement sur les p_n points de discrétisation de la courbe. Ceci sera mis en évidence dans [3], au même titre que d'autres aspects appliqués comme la question du choix automatique des paramètres (celui de la suite w_n, en particulier).

6. Conclusions

Cette Note est la première étape d'un travail général [3] qui mettra, entre autres choses, l'accent sur les nombreuses conséquences directes du Théoreme 4.1. Il existe une dynamique actuelle autour des idées combinant l'analyse de données fonctionnelles et la statistique en grande dimension (voir [6,10]), deux domaines qui se sont souvent développés de manière indépendante, alors qu'ils peuvent naturellement s'enrichir mutuellement. Notre souhait le plus cher est que cette Note puisse être de nature à alimenter cette dynamique.

Remerciements

Les auteurs souhaitent chaleureusement remercier l'expert qui a rapporté sur ce projet de Note, dont les commentaires judicieux ont permis d'en améliorer significativement la présentation. Cette recherche est en partie financé par des subventions MTM 2014-52876-R et CN2012/130 de l'espagnol Ministerio de Economía y Competitividad et Xunta de Galicia, respectivement.

Références

[1] G. Aneiros, P. Vieu, Variable selection in infinite-dimensional problems, Stat. Probab. Lett. 94 (2014) 12–20.
[2] G. Aneiros, P. Vieu, Partial linear modelling with multi-functional covariates, Comput. Stat. 30 (2015) 647–671.
[3] G. Aneiros, P. Vieu, Sparse nonparametric model for regression with functional covariate, en préparation, 2015.
[4] P. Bühlmann, S. van de Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer Series in Statistics, Springer, Heidelberg, 2011.
[5] G. Chagny, A. Roche, Adaptive estimation in the functional nonparametric regression model, J. Multivar. Anal. (2015), http ://dx.doi.org/10.1016/j.jmva.2015.07.001.
[6] A. Cuevas, A partial overview of the theory of statistics with functional data, J. Stat. Plan. Inference 147 (2014) 1–23.
[7] M. Febrero-Bande, P. Galeano, W. Gonzalez-Manteiga, Functional principal components regression and functional partial least square regression: an overview and a comparative study, Rev. Int. Stat. (2015), http://dx.doi.org/10.1111/insr.12116.
[8] F. Ferraty, P. Hall, P. Vieu, Most-predictive design points for functional data predictors, Biometrika 97 (2010) 807–824.
[9] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer Series in Statistics, Springer-Verlag, New York, 2006.
[10] A. Goia, P. Vieu, An introduction to recent advances in high/infinite dimensional statistics, J. Multivar. Anal. (2015), http://dx.doi.org/10.1016/j.jmva.2015.12.001.
[11] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer Series in Statistics, Springer, New York, 2012.
[12] T. Hsing, R. Eubank, Theoretical Foundations of FDA with an Introduction to Linear Operators, Wiley and Sons, Chichester, 2015.
[13] J. Huang, J. Horowitz, F. Wei, Variable selection in nonparametric additive models, Ann. Stat. 38 (4) (2010) 2282–2313.
[14] A. Kneip, P. Sarda, Factor models and variable selection in high-dimensional regression analysis, Ann. Stat. 39 (5) (2011) 2410–2447.
[15] N. Kudraszow, P. Vieu, Uniform consistency of *k*NN regressors for functional variables, Stat. Probab. Lett. 83 (8) (2013) 1863–1870.
[16] T. Laloë, A k-nearest approach for functional regression, Stat. Probab. Lett. 10 (2008) 1189–1193.
[17] I. McKeague, B. Sen, Fractals with point impact in functional linear regression, Ann. Stat. 38 (4) (2010) 2559–2586.
[18] N. Villa, F. Rossi, Un résultat de consistance pour des SVM fonctionnels par interpolation Spline, C. R. Acad. Sci. Paris, Ser. I 343 (8) (2006) 555–560.
[19] Y. Zhao, O. Todd, P. Reiss, Wavelet-based LASSO in functional linear regression, J. Comput. Graph. Stat. 21 (3) (2012) 600–617.
[20] J. Zhang, Analysis of Variance for Functional Data, Monographs on Statistics and Applied Probability, vol. 127, CRC Press, Boca Raton, FL, USA, 2014.
[21] X. Zu, Y. Yang, Variable selection after screening: with or without splitting?, Comput. Stat. 30 (1) (2015) 191–204.