C. R. Acad. Sci. Paris, Ser. I 352 (2014) 651–654

Contents lists available at ScienceDirect

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

www.sciencedirect.com

Statistique

Modèles linéaires généralisés fonctionnels avec dérivée

Aziza Ahmedou [a,c], Jean-Marie Marion [a], Besnik Pumo [b,c]

[a] IMA UCO, 44, rue Rabelais, 49000 Angers, France
[b] Agrocampus Ouest – Centre d'Angers, 2, rue Le Nôtre, 49000 Angers, France
[c] UMR LAREMA 6093 – Université d'Angers, 2, bd Lavoisier, 49045 Angers, France

INFO ARTICLE

Historique de l'article :
Reçu le 16 juillet 2013
Accepté après révision le 27 avril 2014
Disponible sur Internet le 19 juin 2014

Présenté par le Comité de rédaction

RÉSUMÉ

Nous considérons le modèle linéaire généralisé fonctionnel dont la fonction réponse est un opérateur linéaire dépendant d'une variable explicative X appartenant à un espace fonctionnel. Il a été étudié, entre autres, par Cardot et Sarda [4]. Nous considérons dans ce papier le modèle linéaire généralisé fonctionnel avec dérivée, noté par la suite MLGFD, dont la fonction de réponse est définie comme un opérateur linéaire dépendant de X et de sa dérivée. Nous proposons des estimateurs pour les paramètres fonctionnels de ce modèle et fournissons des vitesses de convergence.

ABSTRACT

We consider the functional generalized linear model whose response function is a linear operator depending on an explanatory variable X belonging to a functional space. It has been studied, among others, by Cardot and Sarda [4]. In this paper, we consider the functional generalized linear model with derivative component, denoted MLGFD in the following, whose response function depends on a linear operator of X and on its derivative. We propose estimators for the unknown functional parameters and provide convergence rates.

1. Introduction

Dans de nombreux domaines, on dispose d'observations correspondant à des données fonctionnelles. La littérature statistique propose différents modèles linéaires ou non linéaires pour analyser ce type de données. Nous renvoyons le lecteur aux livres de Ramsay et Silverman [14], Bosq [3] ou Ferraty et Romain [8] pour un panorama assez complet de modèles pour données fonctionnelles. Dans ce travail, nous considérons une extension du modèle linéaire généralisé pour données fonctionnelles (MLGF) étudié entre autres par [4,6,9,11,15]. Dans ces travaux, le lecteur trouvera des exemples intéressants d'application du modèle MLGF, notamment des applications de la régression logistique pour données fonctionnelles.

Soit (X, Y) un couple de variables aléatoires où $X \in \mathbb{R}$ et $Y \in S$ est un sous-ensemble quelconque de $\mathbb{R}$. Le modèle linéaire généralisé (MLG) est défini par l'intermédiaire de la loi conditionnelle $\mathcal{L}(Y|x)$ de Y, sachant que $X = x$. On suppose que $\mathcal{L}(Y|x)$ appartient à la famille exponentielle (Stone [16]) de la forme :

Adresses e-mail : ahmedouaziza@yahoo.fr (A. Ahmedou), Jean-Marie.Marion@uco.fr (J.-M. Marion), Besnik.Pumo@agrocampus-ouest.fr (B. Pumo).

http://dx.doi.org/10.1016/j.crma.2014.04.013
1631-073X/

$$\exp\{\mathbf{b}_1(\eta)y + \mathbf{b}_2(\eta)\}\nu(\mathrm{d}y) \tag{1}$$

où ν est une mesure non nulle définie sur $\mathbb{R}$ et $\eta := \eta(x;\theta)$ est une fonction linéaire en x dépendant d'un ou plusieurs paramètres inconnus θ. $\mathbf{b}_1$ et $\mathbf{b}_2$ sont deux fois continûment dérivables et $\mathbf{b}_3(\eta) = -\mathbf{b}_2'(\eta)/\mathbf{b}_1'(\eta)$ est continûment dérivable. De plus, $\mathbf{b}_1'$ et $\mathbf{b}_3'$ sont strictement positives sur $\mathbb{R}$. On montre alors que la moyenne conditionnelle $\mu = E(Y|X=x)$ est égale à $\mathbf{b}_3(\eta)$ et donc que $\eta = \mathbf{b}_3^{-1}(\mu)$. La fonction $\mathbf{b}_3^{-1}$ s'appelle *fonction de lien*.

Dans la suite, nous supposerons aussi :

$$\mathbf{H1}: \quad \mathbf{b}_1''(\eta)y + \mathbf{b}_2''(\eta) < 0 \quad \forall \eta \in \mathbb{R},\ \forall y \in S \tag{2}$$

Les familles de distribution normale, binomiale, poissonnienne et gamma appartiennent à cette famille, appelée la famille des modèles linéaires généralisés.

Le MLGF est défini pour un couple de variables aléatoires (X,Y) quand $X \in L^2 := L^2[0,1]$, $Y \in S$. La loi conditionnelle $\mathcal{L}(Y|x)$ appartient à la famille exponentielle définie ci-dessus, avec $\eta(x;\alpha) = \langle \alpha, x\rangle_{L^2}$. $\alpha \in L^2$ est le paramètre inconnu du modèle.

Nous proposons d'étudier une extension du modèle précédent, que nous appelons MLGFD (modèle linéaire généralisé fonctionnel avec dérivée), en définissant :

$$\eta(x;\beta,\gamma) = \langle \beta, x\rangle_{L^2} + \langle D\beta, Dx\rangle_{L^2} + \langle \gamma, Dx\rangle_{L^2}. \tag{3}$$

x appartient à l'espace de Sobolev $W := W^{2,1}$ (voir Adams et Fournier [1]), D est l'opérateur dérivée défini sur W et β, γ sont des paramètres du modèle, définis respectivement dans les espaces de fonctions W et L^2.

L'importance qu'il y a à utiliser la dérivée première ou seconde pour augmenter la capacité explicative d'un modèle de régression pour données fonctionnelles a été soulignée par plusieurs auteurs. Nous renvoyons le lecteur à l'ouvrage de Ramsay et Silverman [14, §1.7] ou aux articles de Marx et Eilers [11] et De Belie et al. [5] pour des études qui concernent différents domaines d'application. Marion et Pumo [10] ou Mas et Pumo [12,13] ont introduit et étudié des modèles de prédiction ou de régression linéaire faisant intervenir l'opérateur de dérivation.

Après avoir introduit les conditions d'identifiabilité du MLGFD, nous proposons des estimateurs pour les paramètres inconnus β, γ du modèle et précisons leurs vitesses de convergence.

2. Définition et unicité du MLGFD

Soit (X,Y) un couple de variables aléatoires, $X \in W$, $Y \in S$. Rappelons que $W := W^{2,1}([0,1])$ est l'espace de Sobolev défini par $W = \{X \in L^2 : DX \in L^2\}$. Soit D^* l'adjoint de l'opérateur dérivée D. W est un espace de Hilbert de norme induite par le produit scalaire $\forall u, v \in W$, $\langle u, v\rangle_W = \langle u, v\rangle_{L^2} + \langle Du, Dv\rangle_{L^2}$.

La loi conditionnelle du MLGFD appartient à la famille exponentielle (1) définie ci-dessus, où η est définie par l'équation (3), qui s'écrit aussi :

$$\eta(x;\beta,\gamma) = \langle \beta, x\rangle_W + \langle \gamma, Dx\rangle_{L^2}. \tag{4}$$

Les conditions nécessaires garantissant l'identifiabilité des paramètres de ce modèle sont :

$$\mathbf{H2}: \quad \|X\|_W < \infty \quad \text{p.s.}$$

$$\mathbf{H3}: \quad \operatorname{Ker} \Gamma_X = \{0\}$$

où $\Gamma_X(\cdot) = E(\langle X, \cdot\rangle_W X)$ est l'opérateur de covariance de X, et

$$\mathbf{H4}: \quad (\beta,\gamma) \notin \mathcal{N}$$

où $\mathcal{N}$ est le sous espace de $W \times L^2$ défini par $\mathcal{N} = \{(\beta,\gamma) \in W \times L^2 :\ \beta + D^*\gamma = 0\}$.

Cardot et Sarda [4] proposent les conditions **H2** et **H3** dans l'étude du MLGF. Ces deux conditions sont aussi nécessaires pour définir la régression linéaire fonctionnelle. La condition **H4** a été introduite par Mas et Pumo [13] pour montrer l'identifiabilité des paramètres de la RLFD (régression linéaire fonctionnelle avec dérivée) $Y = \langle \beta, x\rangle_W + \langle \gamma, Dx\rangle_{L^2} + e$.

Proposition 2.1. *Le couple (β,γ) est identifiable dans le modèle MLGFD où η est définie par l'équation (4) si les conditions **H2** à **H4** sont vérifiées.*

Remarque 1. Ceci découle de la bijectivité de $\mathbf{b}_3$ et du résultat analogue pour la RLFD (voir Mas et Pumo [13]).

3. Estimation des paramètres et convergence

L'estimation des paramètres d'un MLGF pour données fonctionnelles a été étudie par plusieurs auteurs – voir par exemple [4,6,7,9,11] ou [15]. Le principe d'estimation est basé sur la maximisation de la vraisemblance d'un échantillon. Notre approche est similaire à celle de Stone [16] pour un modèle additif généralisé ou de Cardot et Sarda [4] pour un modèle linéaire généralisé qui consiste à décomposer les observations de la variable explicative dans une base de fonctions splines (de Boor [2]).

Par analogie avec les travaux de Stone [16] ou de Cardot et Sarda [4], nous supposons que les paramètres β, γ satisfont les conditions :

$$\textbf{H5}: \quad \exists p' \in \mathbb{N},\ C_3 > 0,\ \text{et } r \in]0,1] \quad \begin{array}{l} \left|\beta^{(p'+1)}(t_1) - \beta^{(p'+1)}(t_2)\right| \leq C_3|t_1 - t_2|^r, \\ \left|\gamma^{(p')}(t_1) - \gamma^{(p')}(t_2)\right| \leq C_3|t_1 - t_2|^r. \end{array}$$

Cette condition est analogue à celle utilisée classiquement dans les espaces des fonctions splines. Soit $p = p' + r$ et q le plus petit entier positif $q \geq p$. On note $S_{q,k}$ l'espace des fonctions splines s définies sur $[0,1]$ de degré q et $k-1$ nœuds intérieurs équidistants. Cet espace est de dimension $k+q$ et ces fonctions satisfont deux conditions :

- s est un polynôme de degré q dans chaque intervalle $[(i-1)/k, i/k]$, $i = 1, \cdots, k$;
- s est $q-1$ continument dérivable dans $[0,1]$.

Soit $\{(X_i, Y_i); 1 \leq i \leq n\}$ un échantillon de taille n du MLGFD. La log-vraisemblance $\Lambda(\eta(x; f, g))$ calculée en un point (f, g), $f \in W$, $g \in L^2$ s'écrit :

$$\Lambda_n(f, g) = \sum_{i=1,n} \left[\mathbf{b}_1\big(\eta(X_i; f, g)\big)Y_i + \mathbf{b}_2\big(\eta(X_i; f, g)\big)\right]$$

où $\eta(X_i; f, g)$ est définie par l'équation (4).

Les estimateurs du maximum de vraisemblance sont, s'ils existent, les fonctions $\hat{\beta}_{k,q} \in S_{q+1,k}$, $\hat{\gamma}_{k,q} \in S_{q,k}$ maximisant $\Lambda_n(f, g)$. On démontre alors le résultat suivant :

Théorème 3.1. *Sous les conditions **H1**–**H5** pour le MLGFD, il existe un couple unique* $(\hat{\beta}_{k,q}, \hat{\gamma}_{k,q}) = \arg\max_{f \in S_{q+1,k}, g \in S_{q,k}} \Lambda_n(f, g)$. *De plus,*

$$E\big(\eta(X; \hat{\beta}_{k,q}, \hat{\gamma}_{k,q}) - \eta(X; \beta, \gamma)\big)^2 = O\big(k^{-2p}\big) + O\left(\frac{k}{n}\right).$$

Avec $k = n^{1/(2p+1)}$, *on obtient* $E(\eta(X; \hat{\beta}_{k,q}, \hat{\gamma}_{k,q}) - \eta(X; \beta, \gamma))^2 = O(k^{-2p/(2p+1)})$.

Remarque 2. La vitesse de convergence obtenue pour le modèle MLGFD est équivalente à celle obtenue par Cardot et Sarda [4] pour le modèle MLGF. Une discussion sur la vitesse optimale de convergence est proposée dans l'article de ces deux auteurs. Soulignons enfin que Stone [16] a obtenu une même vitesse de convergence pour les modèles additifs généralisés.

Remarque 3. La première condition de l'hypothèse **H5** impose la propriété hölderienne pour la dérivée d'ordre $p'+1$ pour β. Elle est plus forte que celle de Cardot et Sarda, qui exige cette même condition pour la dérivée d'ordre p' du paramètre α. Ceci s'explique par le fait que $E[\eta(X; \hat{\beta}_{n,q}, \hat{\gamma}_{n,q}) - \eta(X; \beta, \gamma)]^2$ dépend aussi de la dérivée de β qui satisfait la condition de Hölder d'ordre p'.

Références

[1] R.A. Adams, J.J.F. Fournier, Sobolev Spaces, Academic Press, 2003.
[2] C. de Boor, A Practical Guide to Splines, Springer, New York, 1978.
[3] D. Bosq, Linear Processes in Function Spaces, Lecture Notes in Statistics, vol. 149, Springer, 2000.
[4] H. Cardot, P. Sarda, Estimation in generalized linear model for functional data via penalized likelihood, J. Multivar. Anal. 92 (2005) 24–41.
[5] N. De Belie, D.K. Pedersen, M. Martens, R. Bro, L. Munck, J. De Baerdemaeker, The use of visible and near-infrared reflectance measurements to assess sensory changes in carrot texture and sweetness during heat treatment, Biosyst. Eng. 85 (2) (2003) 213–225.
[6] M. Escabias, A.M. Aguilera, M.J. Valderrama, Principal component estimation of functional logistic regression: discussion of two different approaches, J. Nonparametr. Stat. 16 (2004) 365–384.
[7] M. Escabias, A.M. Aguilera, M.J. Valderrama, Functional PLS logit regression model, Comput. Stat. Data Anal. 51 (2007) 4891–4902.
[8] F. Ferraty, Y. Romain, The Oxford Handbook of Functional Data Analysis, Oxford Handbooks, Oxford Univ. Press, Oxford, UK, 2010.
[9] G.M. James, Generalized linear models with functional predictors, J. R. Stat. Soc., Ser. B 64 (3) (2002) 411–432.
[10] J.-M. Marion, B. Pumo, Comparaison des modèles ARH(1) et ARHD(1) sur des données physiologiques, Ann. ISUP 48 (3) (2004) 29–38.

[11] B.D. Marx, P.H.C. Eilers, Generalized linear regression on sampled signals and curves. A p-spline approach, Technometrics 41 (1999) 1–13.
[12] A. Mas, B. Pumo, The ARHD model, J. Stat. Plan. Inference 137 (2) (2007) 538–553.
[13] A. Mas, B. Pumo, Functional linear regression with derivatives, J. Nonparametr. Stat. 21 (1) (2009) 19–40.
[14] J. Ramsay, B. Silverman, Functional Data Analysis, Springer, 1996.
[15] S.J. Ratcliffe, L.R. Leader, G.Z. Heller, Functional data analysis with application to periodically stimulated foetal heart rate data II: functional logistic regression, Stat. Med. 21 (2002) 1115–1127.
[16] C.J. Stone, The dimensionality reduction principle for generalized additive models, Ann. Stat. 14 (2) (1986) 590–606.