Journal of Multivariate Analysis 146 (2016) 313–324

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# Generalized linear model with functional predictors and their derivatives

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

[a] UCO-3 Place André Leroy, 49000 Angers, France
[b] Agrocampus Ouest-Centre d'Angers, 2 rue le Nôtre, 49000 Angers, France
[c] UMR LAREMA 6093-Université d'Angers, 2 Boulevard Lavoisier, 49045 Angers, France
[d] UPRES LARIS 7315-Université d'Angers, 62 avenue Notre Dame du Lac, 49000 Angers, France
[e] UMR IRHS-1345, 42 rue Georges Morel, 49071 Beaucouzé Cedex, France

ARTICLE INFO

*Article history:*
Received 9 January 2015
Available online 30 October 2015

*AMS subject classifications:*
62J12
62G20
65J20

*Keywords:*
Generalized functional linear model
Derivative component model

ABSTRACT

The conditional expectation $E(Y|X)$ of a generalized functional linear model with scalar response $Y$ is given by $g\{\langle X, \phi\rangle_{L^2}\}$ where $X$ and $\phi$ are functions defined in $L^2 := L^2[0, 1]$. Let us consider that $X$ belongs to the Sobolev space $W := W^{2,1}[0, 1]$ and denote $X'$ its derivative. In this paper we focus on an extension of the previous model where $E(Y|X)$ is given by $g\{\langle X, \beta\rangle_W + \langle X', \gamma\rangle_{L^2}\}$. With a similar approach to Cardot and Sarda (2005) or Stone (1986) for generalized additive models, we propose estimators for the unknown parameters $\beta$, $\gamma$ and obtain their rate of convergence. We compare numerically the prediction performance of this new model with alternative models proposed in the literature.



## 1. Introduction

Functional data analysis (FDA) refers to the area of statistics where the data are observed in the form of curves or surfaces which realizations come from an underlying random process varying over a continuum. This new field of statistics became really popular at the end of the nineties with the book of Ramsey and Silverman [32] whose first edition was published in 1997. This book presents principal exploratory and explanatory statistical methods for the analysis of functional data. Functional linear processes (FLP) or functional time series analysis was introduced by Bosq [6]. Detailed presentation of estimation and prediction problems for FLP are presented in [7,8]. Nonparametric statistical methods for functional data are considered in the book of Ferraty and Vieu [19]. The reader will find recent advances on FDA in the book of Horváth and Kokoszka [25], in various special issues in statistical journals as *Statistica Sinica* (14, 3, 2004), *Computational Statistics and Data Analysis* (51, 10, 2007), *Journal of Multivariate Analysis* (101, 2, 2010), collective book edited by Bongiorno et al. [5] and overview papers by Gonzalèz-Manteiga and Vieu [23], Cuevas [12]. Behind the statistical methodology introduced, developed and presented in books cited here above, the reader will find very interesting applications and sometimes routines written in S+ or R.

In this paper we focus our attention to a particular functional regression model with scalar response $Y$:

$$Y = r(X) + \epsilon. \tag{1}$$

* Corresponding author at: Agrocampus Ouest-Centre d'Angers, 2 rue le Nôtre, 49000 Angers, France.
*E-mail address:* Besnik.Pumo@agrocampus-ouest.fr (B. Pumo).

http://dx.doi.org/10.1016/j.jmva.2015.10.009
0047-259X/

$r$ is the regression operator, $X$ is the explanatory (functional) variable defined on an interval $T$ and $\epsilon$ is the error satisfying $E(\epsilon|X) = 0$, so $E(Y|X) = r(X)$. The aim is to estimate $r$ from a sample $(X_i, Y_i),\ i = 1, \ldots, n$. Note in particular that $X$ can be a multifunctional variable, for instance $X$ might be $(X_1, X_2)$, where $X_1$ is a univariate function and $X_2 = X_1'$ its derivative. Ferraty and Vieu [21] give a panorama of known results on estimation and convergence rate of $r$ based on Nadaraya–Watson estimator, the $k$-nearest neighbor and local polynomial estimators.

In situation when $r$ can be rewritten as

$$r(X_1, X_2) = \mu + r_1(X_1) + r_2(X_2).$$

Ferraty and Vieu [20] build estimates for $r_1, r_2$ using a two-stage procedure based on kernel additive estimate. In terms of prediction error this additive modeling is better than the nonparametric one as long as both explanatory variables are not too much correlated (Proposition 1 in [20]).

Since the estimation of $r$ is a difficult statistical problem some authors have considered the single index functional model (SIFM), a particular semiparametric functional regression. Assuming that $X$ acts on $Y$ only through its projection on a functional $\theta$ taking values on the same functional space as $X$, the regression operator $r$ is rewritten as

$$r(X) = g\left\{\alpha + \int_T X(t)\,\theta(t)dt\right\},$$

where $g$ is an unknown real valued function. The intercept parameter $\alpha$ is superfluous and can be replaced by zero as noted by Cai and Hall [9] for functional linear regression and Chen et al. [11] for SIFM model. So, we will omit $\alpha$ in the following of the paper.

Results on estimation and convergence concerning simple and multiple index models can be found in [27,11,17,16,22] among others. This model can be generalized by taking $g$ to be a p-variate function or p-component functional index model (see Chen et al., Section 2.3). Similarly to partial linear functional modeling [3,4] another possible extension is to incorporate a linear combination of scalar explanatory variables in the function g. Note that (under identifiability conditions) the estimation method introduced in this paper can be extended without difficulty for the estimation of parameters associated to scalar variables in partial linear modeling.

Generalized functional linear model (GFLM) is a particular SIFM where $g$ belongs to the exponential family of distributions defined in the next section. The conditional expectation is given by $E(Y|X) = g\{\int_T X(t)\theta(t)\,dt\}$ with "link function" $g^{-1}$. GFLM or particular cases of this model and the estimation problem of the unknown parameter are considered in [28,33,26, 15,30] or [10] among others.

We propose in this paper a particular GFLM that includes also the first derivative $X'$ of $X$ (see also [2]). The idea to use derivative pre-processing of predictors is not new and comes from at least two observations: firstly, differencing removes constants and sudden shifts that are not important for the regression (see [28]); secondly, in some applications the use of first or second-order derivative pre-processing may improve prediction (see for example [13,19]). We propose to consider simultaneously the predictor $X$ and its derivative $X'$ in the prediction model similarly to a precedent work by Mas and Pumo [29] by writing

$$E(Y|X) = g\left\{\int_T [X(t)\beta(t) + X'(t)\beta'(t)]\,dt + \int_T X'(t)\gamma(t)\,dt\right\}. \tag{2}$$

This new model, under assumption that the conditional distribution of $Y$ knowing $X$ belongs to the exponential family, is called generalized functional linear model with derivative and will be denoted shortly GFLMD.

Unlike the single index functional model we consider that $r$ is known and a monotone function. Consequently, we obtain better rates of convergence for the GFLMD predictor than the SIFM predictor. Nevertheless we obtain equivalent rates of convergence as GFLM model which can be explained by the fact that our results concern also the derivative of $X$ (see [10]). Following an idea on functional adaptive model (FAM) introduced by James and Silverman [27] we can define a more flexible model by choosing suitable functions $f_k$ (with the notations of Eq. (2) in [27]) associated to the first and second terms in (2) here above. Another point of view is to consider SIFM model with $g$ defined in (2) and use kernel based approach for the estimation of parameters (see also [31]).

The remainder of the paper is organized as follows. We introduce the GFLMD model and give identifiability conditions in Section 2. In Section 3 we propose estimators for the unknown parameters $\beta$ and $\gamma$ defined on some B-spline spaces and give their rates of convergence. In order to compare numerically our approach with alternative methods, we propose in Section 4 a numerical study on curves discrimination. Proofs are postponed in Section 5 and some remarks conclude the paper.

## 2. Generalized functional linear model with derivative

Following Stonec [35] or Cardot and Sarda [10] we consider the exponential family distributions

$$\exp\{\mathbf{b}_1(\eta)y + \mathbf{b}_2(\eta)\}\nu(dy) \quad \text{with } \eta \in \mathbb{R}$$

where $\nu$ is a nonzero measure on $\mathbb{R}$ which is not concentrated at a single point. The function $\mathbf{b}_1$ is twice continuously differentiable and its first derivative $\mathbf{b}_1'$ is strictly positive on $\mathbb{R}$. Consequently, $\mathbf{b}_1$ is strictly increasing and $\mathbf{b}_2$ is twice