Test (2013) 22:278–292
DOI 10.1007/s11749-012-0308-0

ORIGINAL PAPER

Generalized additive models for functional data

Manuel Febrero-Bande ·
Wenceslao González-Manteiga

Received: 14 December 2011 / Accepted: 11 September 2012 / Published online: 3 October 2012

Abstract The aim of this paper is to extend the ideas of generalized additive models for multivariate data (with known or unknown link function) to functional data covariates. The proposed algorithm is a modified version of the local scoring and backfitting algorithms that allows for the nonparametric estimation of the link function. This algorithm would be applied to predict a binary response example.

Keywords Functional data · Generalized additive models · Generalized linear models

Mathematics Subject Classification 62G08 · 62J12

1 Introduction

For multivariate covariates, a Generalized Linear Model (GLM) (McCullagh and Nelder 1989) generalizes linear regression by allowing the linear model to be related with a response variable Y which is assumed to be generated from a particular distribution in the exponential family (normal, binomial, Poisson, etc.). The response is connected with the linear combination of the covariates, $\mathbf{Z} = (Z_1, \dots, Z_p)'$, through a link function. GLM models provide practitioners a great flexibility to handle with responses that arise in many fields that are far from being Gaussian. The indicator of suffering a certain disease or the number of patients are classical examples from medicine, but in every field non-Gaussian responses can be found. Generalized Additive Models (GAM) (Hastie and Tibshirani 1986) are an extension of GLMs in which the linear predictor is not restricted to be linear in the covariates but is the sum of

M. Febrero-Bande (✉) · W. González-Manteiga
Dpt. of Statistics and O.R. Faculty of Mathematics, University of Santiago de Compostela,
Campus Vida 15782, Santiago de Compostela, Spain
e-mail: manuel.febrero@usc.es

smoothing functions applied to the covariates. GAMs are a good compromise between flexibility and complexity and provide a great tool to a practitioner to decide which model could be more adequate. Some other alternatives are the Single-Index Models (SIM) (see, for example, Horowitz 1998, and references therein) and the GAM with an unknown link function (Horowitz 2001), the latter nesting all the previous models. Our aim is to extend these ideas to the functional covariates. There are some previous works in this direction specially devoted to extend GLM to functional data. As an example, the functional logit model is considered in Escabias et al. (2004, 2006) using principal components or functional PLS to represent the functional covariates. In James (2002), Cardot and Sarda (2005), Müller and StadtMüller (2005) the same is done through a representation in a basis with or without penalization. An extension of these methods using functional and nonfunctional covariates and possibly dependent responses can be found in Goia (2012). The extension to SIM models with functional predictors is studied in Ait-Saïdi et al. (2008) and more recently in Chen et al. (2011). To extend the GAM models to functional data, there are two previous papers following two different approximations. The first one, Müller and Yao (2008), implements an additive model from the projection of the functional components on the eigenbasis of the covariance operator. The second one, Ferraty and Vieu (2009), uses a two-step procedure to estimate an additive model of two functional predictors. We will see both approximations in Sect. 2. More recently, Fan and James (2012) have proposed the FAR model which extends the functional linear regression using the ideas from penalized linear squares optimization approach.

The aim of this paper is to extend the local scoring and backfitting algorithm to functional data in a nonparametric way where the response belongs to the exponential distribution family. Among all the available methods for regression in a univariate framework, the GAMs provide the flexibility to find out the contribution of every covariate against the rigidity of linear models which are more interpretable. In the comparison between GAMs and linear models, the former could be considered as a diagnostic tool to assess about when a linear model is good enough to explain the variability of the response or when a more sophisticated design is desired. This reasoning is again true in the functional context but with a main difference. A functional datum is a complex object that may contain different information depending on how we look at it or, rather, depending on the metric or semimetric employed to catch that information. So, one possibility is to explore several semimetrics for one functional covariate at the same time using a GAM model. From the results of the model, a practitioner could select the informative semimetrics, removing those with no information and/or adding new ones. This flexibility is not possible with linear models because they are restricted to work with Hilbert spaces (or transformations of Hilbert spaces that lead to a new Hilbert space). Of course, it is important to distinguish those semimetrics that can provide different sources of information. To this end, the distance correlation introduced in the paper by Szekely et al. (2007) could help. This distance correlation ($\mathcal{R}$) characterizes independence between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, p and q being arbitrary finite dimensions, and it is computed only with the distances between the elements. The same results is not proved yet for infinite spaces, but, in any case, it is an interesting empirical tool for detecting when two semimetrics provide the same information.

In Sect. 2 we describe some background on GLM and GAM. If the link is supposed to be known, the procedure could be extended to other exponential distribution families. If not, some modifications should be done. Section 3 is devoted to describe a generalized version of the local scoring algorithm that allows us (a) to estimate nonparametrically the GAM (with known or unknown link function) and thus (b) to obtain the corresponding predictive equations. In the nonparametric estimation process, kernel smoothers will be used, and the bandwidths are found automatically by generalized cross-validation. Finally, Sects. 4 and 5 are devoted to simulation studies and applications, respectively.

2 Generalized functional additive models

The extension of classical GLMs to functional predictor (FGLM in the following) simply consists in replacing the linear combination of the covariates by the inner product in the functional space. So, $\mathbf{Z} = \{\mathcal{X}^i\}_{i=1}^p$ being a set of functional covariates with values in the product of p infinite-dimensional Hilbert spaces $\mathbf{E} = \mathcal{E}^1 \times \cdots \times \mathcal{E}^p$, the GLM has the following expression:

$$E(Y|\mathbf{Z}) = \mu = g^{-1}(\eta_z) = g^{-1}\big(\beta_0 + \langle \mathbf{Z}, \boldsymbol{\beta}\rangle\big) \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a functional parameter taking values in $\mathbf{E}$, $\langle \mathbf{Z}, \boldsymbol{\beta}\rangle = \sum_{i=1}^p \langle \mathcal{X}^i, \beta_i\rangle$, and g is the link function, describing the functional relationship between the expected value μ of a datum y and the systematic component $\eta_z = \langle \mathbf{Z}, \boldsymbol{\beta}\rangle$. Most of the methods for estimating β on functional context differs on how the term η_z is computed or approximated. Typically, the solution is the projection of $\mathbf{Z}$ and β onto a finite number of elements of a functional basis, which can be either chosen data-adaptively such as the eigenbasis of the auto-covariance operator of the predictor (see for example, Cardot and Sarda 2005; Escabias et al. 2004, 2006, and references therein), or fixed in advance such as the Wavelet, B-spline or Fourier basis including or not, some penalization (see for example, James 2002; Ramsay and Silverman 2005, and references therein). An obvious extension of the FGLM (1) is the addition of scalar variables as is done in Goia (2012). Once selected how the term η_z will be approximated, the remaining steps of the procedure are the same as in the multivariate case and depends on the distribution family of the response: the choose of the link function and the corresponding analysis of deviance as the goodness-of-fit criterion. This method can be only applied to covariates that belongs to a Hilbert space because the use of projections (the inner product). This could be a limitation in certain situations due to, as mentioned before, a functional datum could contain different information depending on the semimetric used. Obviously, in the case of FGLM we are restricted to $\mathcal{L}_2$ spaces. For a list of common semimetrics used in Functional Data Analysis (see Ferraty and Vieu 2006, Chap. 3). Some tricks have been developed to surpass this limitation. For example, the information provided by the semimetric of the derivatives is equivalent to use the information provided by the Hilbert space of the derivatives. But, similar ideas cannot be employed with other semimetrics like, for example, the semimetric of the supremum.

The above models make the hypothesis that the link function has a known form. This fixed form is, however, rarely justified. Respect to this, the semiparametric single index model (SIM) (Horowitz 1998) generalizes the GLM (1) by allowing the link to be an arbitrary smooth function that has to be estimated from the data. The SIM can be expressed as:

$$E(Y|\mathbf{Z}) = \mu = r(\eta_z) = r\big(\langle \mathbf{Z}, \boldsymbol{\beta}\rangle\big). \tag{2}$$

This model (see, James and Silverman 2005; Ait-Saïdi et al. 2008; Chen et al. 2011) has an important drawback from the practical point of view: the procedure to select the optimal projection $\boldsymbol{\beta}$. This is solved, as in the multivariate case, computing a countable number of projections based on a truncated representation of the functional covariates in a basis or computing the optimal projection in a step-by-step way. Recently, the work by Fan and James (2012) contains two models: FAR and NL-FAR which are, respectively, modified versions of FGLM and SIM models using penalized least squares optimization techniques to find those relevant parts of the functional covariates. But again, the use of inner products restricts the application only to Hilbert spaces.

One way to extend the GLMs for multivariate data is to express the systematic component as the sum of smooth functions. This structure correspond to the so called Generalized Additive Models (GAM) and was introduced by Hastie and Tibshirani (1986). The extension to functional context maybe not so straight as in the case of GLM. The functional GAM model can be expressed as:

$$E(Y|\mathbf{Z}) = \mu = g^{-1}(\eta_z) = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} f_j\big(\mathcal{X}^j\big)\right) \tag{3}$$

where the key question here is the estimation of the partial functions f_j. An answer to this question is done in Müller and Yao (2008) through the functional principal component (FPC) scores of $\mathcal{X}^j$, $f_j(\mathcal{X}^j) := \sum_{k=1}^{K} f_j^k(\xi_j^k)$ being smooth functions of ξ_j^k, the k-principal score of variable j. We will refer to this approach as the Generalized Spectral Additive Model (GSAM) because of the use of spectral decomposition of the covariance operator of $\mathcal{X}$, although the use of another basis representation is possible or even, in certain cases, desirable. The GSAM model has an increasing flexibility while avoiding the curse of dimensionality. Indeed, the fact that the FPC scores are always uncorrelated for every functional covariate ensures that the estimation of partial functions associated with that covariate will not suffer concurvity problems. Concurvity can only occur if the scores of one functional variate are closely related with the scores of another one, but taking into account that the scores are scalars, concurvity can be detected with the usual diagnostic plots between scores of different functional variates. On the other hand, the FPC decomposition is again only valid for Hilbert spaces, and so, other sources of information depending in semimetrics are simply ignored. Also, there is no guarantee that the first K components have predictive information about the response, and the selection of the components to enter in the model is still an open problem.

The other direction to extend GAM to functional context is the work by Ferraty and Vieu (2009), where the estimation of f_j functions is done using functional kernel estimates of the partial functions and considering the response as continuous. In fact, the proposed solution is a one cycle conditional algorithm (one step for each functional covariate conditionally on previous estimation). We will refer to this model as the Functional Additive Model (FAM). Our proposal takes the FAM model as its starting point, extending to the situations with response coming from the exponential distribution family and to those situations in which there is not enough information either about the form of the link (as in the SIM) or about the shape of the partial functions (as in the GAM). Particularly, the estimation of the partial functions in a nonparametric way makes our algorithm applicable to functional covariates in Banach spaces or even in metric spaces.

3 GAM: estimation and prediction

The main goal of this paper is to propose an algorithm to solve this broader class of models. Such a general formulation will be presented here as G-GKAM (Generalized Kernel Additive Models with unknown link function) with the purpose of widening the assumptions regarding the link in generalized additive models. We propose to adapt the techniques shown in Roca-Pardiñas et al. (2004) in such a way that it will allow the nonparametric estimation of the partial functions f_j and, if needed, the joint nonparametric estimation of the inverse link $g^{-1} = \mathbf{H}$, when the covariates are curves. Our proposal is to extend the backfitting algorithm to this context, and so, the estimation of the partial functions at step l should be done in the following way:

$$\hat{f}_j^l(\mathcal{X}^j) = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^{-j,l}) K_j(\frac{d_j(\mathcal{X}^j,\mathcal{X}_i^j)}{h_j})}{\sum_{i=1}^n K_j(\frac{d_j(\mathcal{X}^j,\mathcal{X}_i^j)}{h_j})} \tag{4}$$

where $\hat{Y}_i^{-j,l} = \sum_{i=1}^{j-1} \hat{f}_i^l(\mathcal{X}^i) + \sum_{i=j+1}^{p} \hat{f}_i^{(l-1)}(\mathcal{X}^i)$ is the prediction without variable j, d_j is the distance (induced by the norm) in space $\mathcal{E}_j$, and K_j and h_j are an asymmetric kernel function and the bandwidth, respectively. Some advantages can be deduced from Eq. (4):

1. The estimator only uses distances between covariates, and so, it can be applied to functional covariates in general metric spaces. Of course, scalar covariates can be included. Also, it is possible to include linear terms in the algorithm.
2. The use of Nadaraya–Watson-type estimator for partial functions ensures that the algorithm converges and has a unique global solution (see, Buja et al. 1989) because the smoother matrix is always strictly shrinking. Note that, the convergence to a global solution is guaranteed, but not to the partial ones.
3. The estimator does not suffer from curse of dimensionality because at each step, the process involves just one parameter, the bandwidth.
4. Finally, additive models give us the opportunity to look at that complex object from different points of view or semimetrics and so, to extract as much as information the functional covariate contains. So, it is possible to include the same functional datum under different semimetrics in the model and, reading carefully the results, to obtain a useful insight about which semimetric is more informative.

In the case of an unknown link, the estimator of the link function must provide also its derivative in order to obtain an estimation of the variance of the linearized response.

This motivates the use of Linear Local Regression when the estimation of the link is demanded.

Before estimating the partial functions and the link, some restrictions have to be imposed in order to ensure the GKAM (G-GKAM) identification. This is a usual topic in multivariate GAM and SIM models. In the GAM context, identification is guaranteed by introducing a constant β_0 into the model and requiring a zero mean for the partial functions ($E(f_j) = 0$). In the SIM and G-GAM, however, given that the link function is not fixed, it is necessary to establish further conditions in order to avoid different combinations of $\mathbf{H}$ and $f_j s$ that could lead to the same model. In this paper, we follow the same ideas, and when estimating a GKAM (G-GKAM), we impose the following conditions:

1. (General condition) $E[f_j] = 0$ $(j = 1, \ldots, p)$.
2. (G-GKAM only) $\beta_0 = 0$ and $E[(\sum_{j=1}^{p} f_j)^2] = 1$.

These are the same two conditions as in Roca-Pardiñas et al. (2004). Note that, from these conditions, the systematic component η_z becomes standardized when a G-GKAM is estimated.

The proposed algorithm is as follows:

For a given $(\mathbf{Z}, Y)$, the local scoring maximizes an estimation of the expected log-likelihood $E[l\{\eta_z; Y\}|\mathbf{Z}]$, for example, when the response is a binary variable:

$$l\{\eta_z; Y\} = Y \log\big[\mathbf{H}(\eta_z)\big] + (1 - Y) \log\big[1 - \mathbf{H}(\eta_z)\big] \tag{5}$$

by solving iteratively a reweighted least squares problem in the following way.

In each iteration, given the current guess $\hat{\eta}_Z^0$, the linearized response $\tilde{Y}$ and the weight $\tilde{W}$ are constructed as

$$\tilde{Y} = \hat{\eta}_Z^0 + (Y - \hat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)_{\hat{\mu}_0} \quad \text{and} \quad \hat{W}_0^{-1} = \mathrm{Var}(\tilde{Y}|\mathbf{Z}) = V_0\left(\frac{d\eta}{d\mu}\right)_0^2 \tag{6}$$

where V_0 is the variance function evaluated at $\hat{\mu}_0$. Typically, the above equations can be expressed in terms of $\mathbf{H}$ and its derivatives. For example, in the case where the response is a binary variable, the above equations are reduced to the following form:

$$\tilde{Y} = \hat{\eta}_Z^0 + \frac{Y - \mathbf{H}(\hat{\eta}_Z^0)}{\mathbf{H}'(\hat{\eta}_Z^0)} \quad \text{and} \quad \hat{W}_0^{-1} = \mathrm{Var}(\tilde{Y}|\mathbf{Z}) = \frac{\mathbf{H}(\hat{\eta}_Z^0)(1 - \mathbf{H}(\hat{\eta}_Z^0))}{\mathbf{H}'(\hat{\eta}_Z^0)^2}. \tag{7}$$

To estimate the f_js, we fit an additive regression model to $\tilde{Y}$, treating it as a response variable with associated weight $\hat{W}_0$. The resulting estimation of $\hat{\eta}_Z$ is $\hat{\eta}_Z^0$ of the next iteration. This procedure must be repeated until negligible changes in the systematic component. For the estimation of the f_js and $\mathbf{H}$, the following two alternating loops must be performed.

Loop 1. Let $\hat{\eta}_Z^0$ and $\hat{\mu}_0 = \hat{\mathbf{H}}^0(\hat{\eta}_Z^0)$ (and possibly $\hat{\mathbf{H}}'^0(\hat{\eta}_Z^0)$) be the current estimates. Replacing the functions $\mathbf{H}$ (and $\mathbf{H}'$) by their current estimates, $\hat{\mathbf{H}}^0$ (and $\hat{\mathbf{H}}'^0$), in formulas given in (6), $\hat{\eta}_Z = \beta_0 + \sum_{j=1}^{p} \hat{f}_j(\mathcal{X}^j)$ is then obtained by fitting an additive model of $\tilde{Y}$ on $\mathbf{Z}$ with weights $\hat{W}$. If the link must be estimated, then the systematic component is rescaled to fulfill identifiability conditions.

Loop 2. (G-GKAM only) Fixing $\hat{\eta}_Z$, the two estimates $\hat{\mu}_0 = \hat{\mathbf{H}}(\hat{\eta}_Z)$ and their derivatives are then obtained by fitting a regression model of Y on $\mathbf{Z}$ weighted by V_0^{-1} using a polynomial local kernel estimators in order to also have estimations of the derivatives.

These two loops are repeated until the relative change in deviance is negligible.

At each iteration of the estimation algorithm, the partial functions are estimated by applying Nadaraya–Watson weighted kernel smoothers to the data $\{\mathcal{X}^j, R^j\}$ with weights $\hat{W}$, R^j being the residuals associated to $\mathcal{X}^j$ obtained by removing the effect of the other covariates. In this paper, for each $\hat{f}_j$, the corresponding bandwidth h_j is selected automatically by minimizing, in each of the cycles of the algorithm, the weighted GCV error criterion, whereas the bandwidth for estimating the link function (if needed) is found minimizing the cross-loglikelihood error criterion (analogous to (5)). In all cases, the computation of the optimal bandwidths are done in a suitable grid, although any other unidimensional algorithm is possible.

As initial estimates, we consider $f_j^0 := 0$, $\beta_0 = g(\bar{Y})$ for a GKAM model, and in the case of a G-GKAM, $\beta_0 = 0$, $f_1^0 = \int \mathcal{X}(t)\,dt$, and $f_j^0 = 0$, $j \neq 1$.

3.1 Practical considerations

There are several practical aspects that must be taken into account when implementing the above steps:

- The contribution of every partial function in the response can be measured in terms of the determination coefficient with respect to the linearized response η. This information, jointly with the effective number of parameters (eqPar) of the partial function, gives an idea about the importance of that functional covariate and its complexity. In our case, the effective number of parameters of partial function i is defined as $df(S_j) = \text{trace}(S_j)$ where S_j is the smoothing matrix of the partial function f_j.
- The bandwidth for every step is selected applying a GCV criterion in order to maintain the algorithm in a reasonable time consuming. Every GCV criterion uses a penalizing term which is a function of the eqPar consumed by the model. In our case, at each step, we have employed the global degree of freedom as a penalizing term (the sum of eqPar from the current estimation of partial functions).
- Stop criteria. As usual in the univariate case, the algorithm stops when the following both conditions are satisfied: (1) $|f_i^j(\mathcal{X}_i) - f_i^{j+1}(\mathcal{X}_i)| \leq \epsilon |f_i^j(\mathcal{X}_i)|$, and (2) $|\hat{\eta}^j - \hat{\eta}^{j+1}| \leq \epsilon |\hat{\eta}^j|$, ϵ being a precision constant.
- When the response is binary, the inverse of weights at each step could be arbitrary close to zero because the probability could be close to zero or one. Typically, those data are discarded. If this occurs for too many curves, the algorithm could try to estimate the partial functions without enough data. To avoid this, when the number of weights significantly distinct from zero are less than the equivalent number of parameters of the estimator, the algorithm stops.
- The known problems of additive models for univariate data have their reflection in the functional context. The boundary effect of the Nadaraya–Watson estimators also applies here, boundary effect meaning that a curve is not closely surrounded

by others. Also, additive models could suffer concurvity (some smooth terms could be approximated by one or more of the other smooth terms). In functional context, concurvity occurs when certain relationship among the distances holds, and so, avoiding those terms based on similar distances (for example, the $\mathcal{L}_2$ distance and semimetric of principal components) should be enough to ensure that the algorithm does not break down. The distance correlation mentioned above could be a useful tool providing a measure of dependence between distances discovering what information in one distance is similar to another.

4 Simulation study

We have considered three scenarios to assess the performance of the algorithm proposed (FGKAM) in comparison with the other competing methods: FGLM and GSAM. These scenarios are computed from two Gaussian processes $\mathcal{X}_1$ and $\mathcal{X}_2$ evaluated in a fine grid of $N = 101$ points $\{t_1, \ldots, t_N\} \in [0, 1]$ with covariance matrices $\Sigma_1(s,t) = \frac{1}{2}\exp(-0.8|s-t|)$ and $\Sigma_2(s,t) = \frac{2}{5}\exp(-0.6|s-t|)$, respectively. Also, we added a systematic sinusoidal trend to $\mathcal{X}_1$.

The response of the three scenarios are computed in the following way:

M1: $y = 3 + 5\langle \mathcal{X}_1, \beta_1\rangle - \langle \mathcal{X}_2, \beta_2\rangle + \epsilon$
M2: $y = 4(\int \mathcal{X}_1^2(t)\,dt)^{1/3} - 5(\int \mathcal{X}_2^2(t)\,dt)^{1/3} + \epsilon$
M3: $y = \frac{2}{5}\int \mathcal{X}_1(t)^3\,dt - 3\exp(-d_2(\mathcal{X}_2, C)) + \epsilon$

where $\beta_1(t) = t(1-t)$, $\beta_2(t) = 1$, $C(t) = \log(t+1)$, and d_2 is the distance under the $\mathcal{L}_2$ norm. An example of the functional covariates and the density of the responses for the different scenarios are shown in Fig. 1.

In every case, ϵ is a Gaussian variate with $\sigma_\epsilon^2 = snr \cdot \mathrm{Var}(S)$, where the signal-to-noise ratio $snr = 0.01, 0.1, 0.2$, and the signal $S = f_1(\mathcal{X}_1) + f_2(\mathcal{X}_2)$.

The M1 scenario corresponds to a classical functional linear model, and so, the FGLM must be optimal, although GSAM and FGKAM methods should be close. The two terms of M2 scenario are related with the $\mathcal{L}_2$ norm of $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, and so, it is a case designed for GSAM, but it will be a hard scenario for FGLM. Finally, the M3 scenario is also hard for GSAM (and, of course, for FGLM) because it contains some terms which cannot be explained successfully only with the $\mathcal{L}_2$ norm. Here, the better results can be obtained using for $\mathcal{X}_1$ a metric based on the $\mathcal{L}_3$ norm in the FGKAM procedure.

For every scenario, $N = 100, 500, 1000$ data were generated, using the first half $(50, 250, 500)$ as training sample, and the latter for validation purposes. The results for $B = 100$ replications are shown in Table 1. In all cases, the distance used was the $\mathcal{L}_2$ distance for FGKAM procedure, although in M3 the results could be improved if a metric related with $\mathcal{L}_3$ were employed. In this case, the distance correlation between distances $\mathcal{L}_2$ and distances $\mathcal{L}_3$ for the first covariate is $\mathcal{R}(\mathcal{X}_1^{\mathcal{L}_2}, \mathcal{X}_1^{\mathcal{L}_3}) = 0.9987$. For GSAM procedure, the number of eigenfunctions considered was $k = 3$, which explains about 84 % of the total variability in both functional variates. In scenario M1, the Mean Square Error (MSE) of the residuals is quite similar among the three competitors with a tendency of FGKAM method to provide smaller values especially for

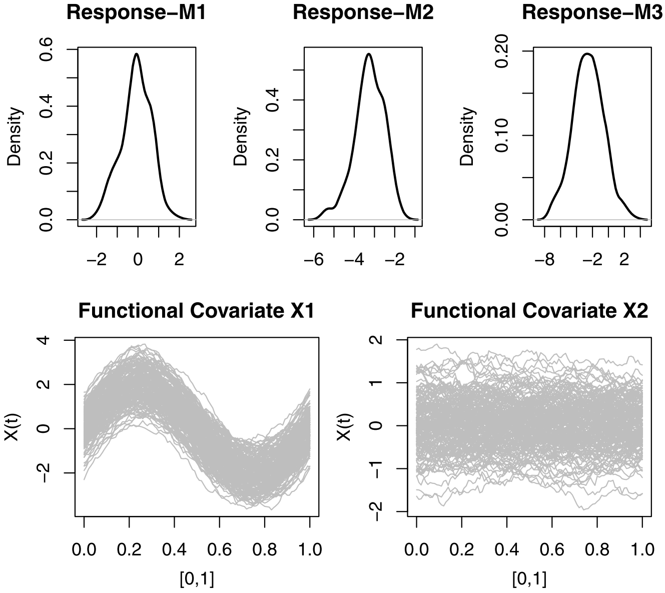

Fig. 1 Response for scenarios M1, M2 and M3 (*top row*) and functional covariates X1 and X2 (*bottom row*)

$snr \neq 0.01$. Probably, there is a tendency to overtrain for FGKAM method. The results on prediction in this scenario for FGKAM are slightly worse than for FGLM and GSAM, which basically are the same. The difference between models is reduced when the sample size or the signal-to-noise ratio increases. The loss of FGKAM is probably due to the prediction in data far from training sample (boundary effect). The scenarios M2 and M3 cannot be handled by FGLM method, and the FGLM results here are poor. The FGKAM procedure obtain the best results, closely followed by the GSAM method in the M2 scenario, although here increasing the number of eigenfunctions could improve the results of this procedure. In the M3 scenario, the FGKAM procedure clearly outperforms the other competitors, the most significative differences being when the *snr* is small and the sample size is large. The effect of sample size in the prediction is more important in the case of FGKAM method than in the others. The rows corresponding to prediction errors are quite stable for FGLM and GSAM with respect to the sample size, whereas the improvement for the MSE of the FGKAM method with respect to the sample size is notable.

Also, in order to check other distributions in the response, we have generated samples with Bernoulli response using $p = \mathbb{P}\{Y = 1|Z\} = \frac{\exp\{cS^*\}}{1+\exp\{cS^*\}}$ with $c \in \{1, 2.5\}$ and $S^* = S - \bar{S}$, the recentered signal. The results are showed in Table 2 only for the case $N = 500$ using the MSE of the residuals (with respect to p) and the percentage of good classification in the testing sample. The MSE follows the guidelines shown above for Gaussian response, but the percentage of good classification gives a not-so-clear message, especially for scenario M3. In this scenario, the FGLM obtains results in classification similar to the other two methods, GSAM and FGKAM. This can be explained taking into account that for the percentage of good classification, only the

Table 1 Mean residual square error (res.) and mean prediction square error (pred.) for scenarios M1, M2, and M3 of the three competitors: FGLM, GSAM, and FGKAM

Scenario	Model\snr	$N = 50$			$N = 250$			$N = 500$		
		0.01	0.1	0.2	0.01	0.1	0.2	0.01	0.1	0.2
M1 Res.	FGLM	0.0059	0.0593	0.1146	0.0062	0.0609	0.1217	0.0062	0.0621	0.1233
	GSAM	0.0056	0.0568	0.1101	0.0061	0.0600	0.1198	0.0061	0.0616	0.1223
	FGKAM	0.0117	0.0496	0.0947	0.0110	0.0521	0.1041	0.0104	0.0557	0.1077
M1 Pred.	FGLM	0.0066	0.0676	0.1335	0.0066	0.0644	0.1285	0.0064	0.0633	0.1266
	GSAM	0.0068	0.0697	0.1365	0.0066	0.0649	0.1293	0.0064	0.0635	0.1272
	FGKAM	0.0288	0.0881	0.1602	0.0195	0.0776	0.1448	0.0155	0.0731	0.1386
M2 Res.	FGLM	0.5816	0.6429	0.6802	0.6083	0.6624	0.7339	0.6124	0.6782	0.7491
	GSAM	0.0835	0.1333	0.1925	0.0859	0.1464	0.2096	0.0859	0.1483	0.2151
	FGKAM	0.0530	0.0902	0.1355	0.0494	0.0934	0.1423	0.0467	0.0952	0.1486
M2 Pred.	FGLM	0.6913	0.7464	0.8297	0.6459	0.7136	0.7783	0.6486	0.6990	0.7724
	GSAM	0.1078	0.1726	0.2453	0.0961	0.1602	0.2286	0.0927	0.1535	0.2223
	FGKAM	0.1038	0.1691	0.2506	0.0754	0.1403	0.2104	0.0626	0.1238	0.1945
M3 Res.	FGLM	0.2438	0.3370	0.4260	0.2668	0.3605	0.4637	0.2774	0.3633	0.4700
	GSAM	0.1034	0.1861	0.2776	0.1268	0.2117	0.3149	0.1347	0.2206	0.3237
	FGKAM	0.0335	0.0933	0.1578	0.0357	0.1015	0.1783	0.0352	0.1065	0.1886
M3 Pred.	FGLM	0.3081	0.4187	0.5171	0.2935	0.3915	0.5013	0.2935	0.3873	0.4839
	GSAM	0.1608	0.2735	0.3761	0.1570	0.2522	0.3684	0.1477	0.2448	0.3413
	FGKAM	0.1327	0.2332	0.3401	0.0941	0.1891	0.3061	0.0728	0.1691	0.2691

Table 2 MSE and percentage of good classification for scenarios M1, M2, and M3 with binary response

		FGLM		GSAM		FGKAM	
		MSE	% good class.	MSE	% good class.	MSE	% good class.
$c = 1$	M1	0.0046	63.8 %	0.0048	63.8 %	0.0057	64.4 %
	M2	0.0412	49.4 %	0.0152	62.8 %	0.0149	58.2 %
	M3	0.0143	61.2 %	0.0091	61.4 %	0.0098	64.6 %
$c = 2.5$	M1	0.0051	74.0 %	0.0051	74.0 %	0.0142	73.6 %
	M2	0.1085	54.2 %	0.0177	72.8 %	0.0246	72.0 %
	M3	0.0315	71.0 %	0.0078	75.4 %	0.0152	76.0 %

behavior of the model around $p = 0.5$ is important. The constant c, when increasing, has the effect of separating the groups, making the estimate of p less important.

All simulations were done using the R-package `fda.usc` (Febrero-Bande and Oviedo de la Fuente 2012a, 2012b) where the methods in the comparison are implemented. In order to have an idea about the computing cost of each method, some of the simulations were run in a computer with Intel Core i5 CPU Processor 2.67 GHz.

Table 3 Average CPU times (in secs.) as functions of sample size

	Basis	FGLM	GSAM	Distances	FGKAM (h)
$N = 50$	0.29	0.31	0.36	0.14	1.08 (0.15)
$N = 250$	4.05	4.08	4.29	3.78	23.67 (3.15)
$N = 500$	5.73	5.77	5.89	14.86	86.9 (12.0)

The average CPU times in seconds are showed in Table 3 as functions of the sample size including some of the intermediate tasks for every method. For the FGLM and GSAM methods, the intermediate task is to represent the functional data in the chosen basis, in this case, the first three eigenfunctions for each covariate. In the case of the FGKAM method, the intermediate task is to obtain the matrix of distances between data, which is done once for every covariate and depends on the number of discretization points (done here by numerical integration). The other source of time consumption is the task of finding the optimal bandwidth for each covariate at each iteration. This is done in our implementation looking for the optimal bandwidth in a fine grid of 51 values. This time is reflected in the column FGKAM in parentheses. The FGKAM is quite high demanding, although the comparison is not fair because the FGLM and GSAM methods use standard R methods fully optimized (`stats::glm.fit` and `mgcv::gam`, respectively), and this is not done yet for FGKAM. Of course, any improvement in the computation of the distances and/or in finding the optimal bandwidth will have an important impact on the CPU times.

5 Application

In this section, we present an application of the FKGAM model (3) to the Tecator dataset. This data set was widely used in examples with functional data (see, Ferraty and Vieu 2006) to predict the content of fat content on samples of finely chopped meat. For each food sample, the spectrum of the absorbances recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850–1050 mm by the near-infrared transmission (NIT) principle is provided also with the fat, protein, and moisture contents, measured in percent and determined by analytic chemistry. We had $n = 215$ independent observations, usually divided into two data sets: the training sample with the first 165 observations and the testing sample with the others. In this study, we are trying to predict the fat content, $Y_1 = \text{Fat}$ and also, an indicator variable related with the fat content, $Y_2 = I\{\text{Fat} \geq 15\}$ where $\mathbf{Z} = (\mathcal{A}, \mathcal{A}'')$, $\mathcal{A}$ being the absorbances, and $\mathcal{A}''$ its second derivative. The use of the second derivative is justified by previous works (see, for example, Aneiros-Pérez and Vieu 2006; Ferraty and Vieu 2009, among others), where the models with information including the second derivative have better prediction results. The curves and the second derivative are shown in Fig. 2. Here, the gray group (fat over 15 %) is clearly quite well separated when considering the second derivative and quite mixing when considering the spectrum itself. This suggests that the relevant information about high percentage of fat is mainly related with the second derivative. The prediction of fat content is included only for comparison purposes with those previous works.

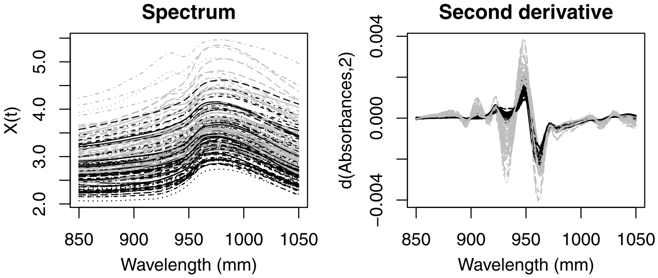

Fig. 2 Spectrum and second derivative of training sample colored by binary response (*gray* = I(Fat ≥ 15 %))

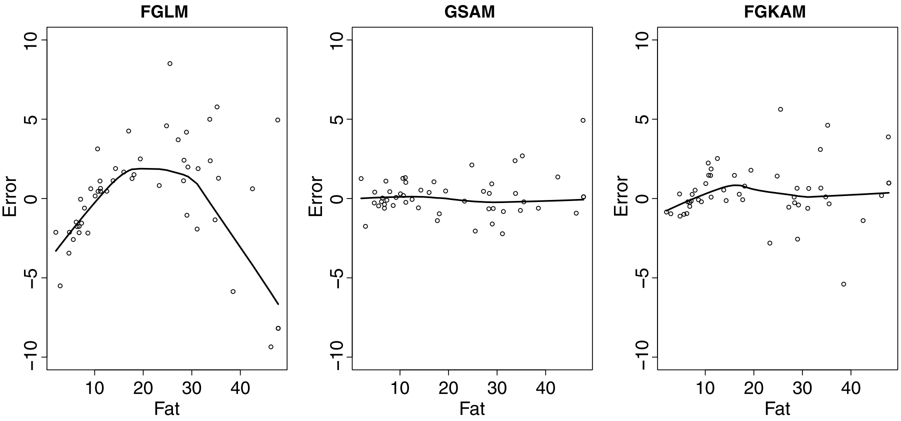

Fig. 3 Fat prediction error vs fat content in the testing sample for FGLM, GSAM, and FKGAM models

The work by Ferraty and Vieu (2009) makes the comparison between models using the Mean Square Residuals (MSR) of the testing sample obtaining 1.88 as the best result of the nonparametric additive model before the boosting stages. In Aneiros-Pérez and Vieu (2006), the functional nonparametric model using only the second derivative has an MSR of 4.31 (the original table shows the MSR divided by the variance of the fat content in the testing sample). In our case, the MSR of FGLM, GSAM, and FKGAM was 12.39, 1.53, and 3.23, respectively. The bad results obtained by the FGLM suggest that, in this example, the fat content cannot be well explained with a functional linear model. This can be checked in Fig. 3, where the prediction errors are plotted against fat content with a Lowess estimation of the trend. The Lowess line for FGLM is far for being constant, suggesting that there is something nonlinear that should be included in the model. The lines for GSAM and FKGAM are both quite flat, although the FKGAM one shows a small trend for low fat contents. Probably, this boundary effect could explain the high MSR of the FKGAM model with respect to GSAM. The distance correlation between both covariates is $\mathcal{R}(\mathcal{A}, \mathcal{A}'') = 0.522$, which clearly indicates no concurvity.

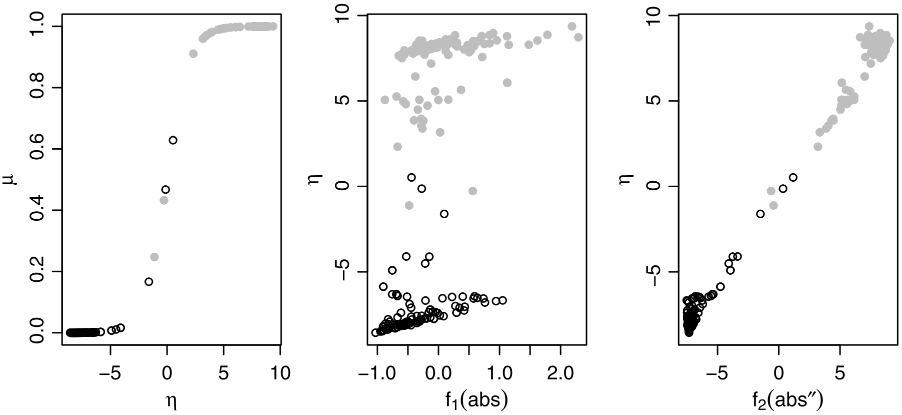

Fig. 4 Estimation of the partial effects (*gray* = I(Fat ≥ 15 %))

Table 4 Statistics for percentage of good classification in 500 replications

Method	Sample	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
FGLM	Training	100 %	100 %	100 %	100 %	100 %	100 %
	Testing	88.0 %	96.0 %	98.0 %	97.5 %	98.0 %	100 %
GSAM	Training	100.0 %	100.0 %	100.0 %	100 %	100 %	100 %
	Testing	54.0 %	92.0 %	94.0 %	93.8 %	98.0 %	100 %
FKGAM	Training	97.58 %	98.18 %	98.8 %	98.7 %	98.8 %	100 %
	Testing	90.0 %	96.0 %	98.0 %	97.9 %	100.0 %	100 %

In the second case, the model can be expressed by

$$E(Y_2 = 1|\mathbf{Z}) = p(\mathbf{Z}) = p(\mathcal{A}, \mathcal{A}'') = \mathbf{H}(\eta_z) = \mathbf{H}(\beta_0 + f_1(\mathcal{A}) + f_2(\mathcal{A}'')) \tag{8}$$

where **H** is the logit link.

The impression that the second derivative is informative could be confirmed in Fig. 4, where the contribution of every functional covariate to η is shown in the central and right plots. The spectrum curves show a chaotic behavior with respect to η, whereas the second derivative of each curve shows a clearly increasing pattern. Indeed, the trace of the smoothing matrices S_1, S_2 associated with f_1, f_2 are respectively 2.5 and 88.6, which indicates a higher contribution of the second derivative. These values were similar to those obtained in the first case where the response is continuous. Classifying every observation according to the estimated probability, the percentage of good classification is 98.8 % and 96 % for the training and testing samples, respectively. The FGLM and GSAM methods raise to 100 % in the training sample, and to 98 % and 92 % in the testing sample, respectively.

We have also repeated this analysis 500 times changing at random which data are included in the training sample and keeping the size of the training sample in 165 observations. The results are summarized in Table 4 and are quite promising. The three methods perform well, although FGLM and GSAM methods have a tendency to overtrain. For the testing sample, the FGKAM procedure has better results with less variability, slightly better than the other methods. So, the small differences here, in contrast with the case of continuous response, could be explained pointing out that for binary responses computed as a cutpoint of a continuous one, only the goodness of the model around the cut level is relevant.

6 Conclusions

In this paper, we have proposed an algorithm to estimate a wide class of regression models for functional data with response belonging to the exponential family. This algorithm (named Generalized Kernel Additive Model or FGKAM for short) is based on a mixing of the IRLS and Backfitting algorithms adapted to the functional context. Our proposal is compared with Functional Generalized Linear Models (James 2002; Escabias et al. 2004, 2006; Cardot and Sarda 2005; Müller and StadtMüller 2005) and with Generalized Spectral Additive Models (Müller and Yao 2008) in a simulation study and in a real example using the R-package `fda.usc` (Febrero-Bande and Oviedo de la Fuente 2012a), where the three proposals are implemented in a integrated way. The FGKAM has proven to be useful in simulations and in application examples. Nevertheless, some questions arise in the application:

- The algorithm is quite high consuming, especially when the link functions have to be estimated and the convergence is slow. Due to the functional nature of the data, the usual techniques in univariate framework for speed up the computations (like for example, binning) are not available here.
- The search of an automatic optimal bandwidth is a challenging and critical task. This procedure is invoked repeatedly for every covariate at each iteration, and so, it must be a fast procedure based on GCV techniques (CV techniques must be discarded). But the type of penalizing term is an open problem, which is more complicated here with several covariates
- An obvious extension of the proposed model is to mix functional and scalar covariates, the contribution of these covariates being linear or smoothed.

Acknowledgements The authors are grateful to the editor and three anonymous referees for their useful comments. This research has been partially funded by project MTM2008-03010 from Ministerio de Ciencia e Innovación, Spain and by project 10MDS207015PR from Xunta de Galicia, Spain.

References

Ait-Saïdi A, Ferraty F, Kassa R, Vieu P (2008) Cross-validated estimations in the single-functional index model. Statistics 42(6):475–494

Aneiros-Pérez G, Vieu P (2006) Semi-functional partial linear regression. Stat Probab Lett 76:1102–1110

Buja A, Hastie TJ, Tibshirani RJ (1989) Linear smoothers and additive models. Ann Stat 17:453–455

Cardot H, Sarda P (2005) Estimation in generalized linear models for functional data via penalized likelihood. J Multivar Anal 92(1):24–41

Chen D, Hall P, Müller HG (2011) Single and multiple index functional regression models with nonparametric link. Ann Stat 39(3):1720–1747

Escabias M, Aguilera AM, Valderrama MJ (2004) Principal component estimation of functional logistic regression: discussion of two different approaches. Nonparametr Stat 16(3–4):365–384

Escabias M, Aguilera AM, Valderrama MJ (2006) Functional PLS logit regression model. Comput Stat Data Anal 51:4891–4902

Fan Y, James G (2012) Functional additive regression. Preprint, available at http://www-bcf.usc.edu/~gareth/research/FAR.pdf

Febrero-Bande M, Oviedo de la Fuente M (2012a) fda.usc: functional data analysis. Utilities for Statistical Computing. URL http://cran.r-project.org/web/packages/fda.usc/index.html. R package version 0.9.8

Febrero-Bande M, Oviedo de la Fuente M (2012b) Statistical computing in functional data analysis: the R package fda.usc. J Stat Softw, preprint

Ferraty F, Vieu P (2006) Nonparametric functional data analysis. Springer, New York

Ferraty F, Vieu P (2009) Additive prediction and boosting for functional data. Comput Stat Data Anal 53:1400–1413

Goia A (2012) A functional linear model for time series prediction with exogenous variables. Stat Probab Lett 82:1005–1011

Hastie T, Tibshirani R (1986) Generalized additive models (with discussion). Stat Sci 3:297–318

Horowitz JL (1998) Semiparametric methods in econometrics. Lecture notes in statistics, vol 131. Springer, Berlin

Horowitz JL (2001) Nonparametric estimation of a generalized additive model with an unknown link function. Econometrica 69:499–514

James GM (2002) Generalized linear models with functional predictors. J R Stat Soc B 63(3):411–432

James GM, Silverman BW (2005) Functional adaptive model estimation. J Am Stat Assoc 100(470):565–576

McCullagh P, Nelder JA (1989) Generalized linear models. Chapman & Hall, London

Müller HG, StadtMüller UF (2005) Generalized functional linear model. Ann Stat 33(2):774–805

Müller HG, Yao F (2008) Functional additive model. J Am Stat Assoc 103(484):1534–1544

Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, Berlin

Roca-Pardiñas J, González-Manteiga W, Febrero-Bande M, Prada-Sánchez JM, Cadarso-Suárez C (2004) Predicting binary time series of SO_2 using generalized additive models with unknown link function. Environmetrics 15:729–742

Szekely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35(6):2769–2794