Forecasting binary longitudinal data by a functional PC-ARIMA model

A. M. Aguilera[1], M. Escabias[2], and M. J. Valderrama[2]

[1] Universidad de Granada
Dpto. de Estadística e I.O
Facultad de Ciencias
Campus de Fuentenueva
18071-Granada, Spain
(e-mail: aaguiler@ugr.es)

[2] Universidad de Granada
Dpto. de Estadística e I.O
Facultad de Ciencias
Campus de Fuentenueva
18071-Granada, Spain
(e-mail: escabias@ugr.es, valderra@ugr.es)

Abstract. The purpose of this paper is to forecast the time evolution of a binary response variable from an associated continuous time series observed only at discrete time points that usually are unequally spaced. In order to solve this problem we are going to use a functional logit model based on functional principal component analysis of the predictor time series that takes into account its continuous nature, close to classical ARIMA modelling of the associated discrete time series of principal components.
Keywords: Logistic Regression, Funcional Principal Components, ARIMA models.

1 Problem formulation

In this paper we propose a functional logit model based in mixed ARIMA-FPCA modelling of the functional predictor that allows to forecast the time evolution of a binary response from discrete time observations of a continuous time series. FPCA [Ramsay and Silverman, 1997] is a generalization of the classic principal component analysis (PCA) of a sample of data vectors for the reduction of dimension of a set of sample curves obtained in our case by cutting the predictor series in periods of the same amplitude. Mixed ARIMA-FPCA models [Valderrama *et al.*, 2002] allows not only to forecast a continuous time series in a whole future interval but also to reconstruct it between the discretization time points in the past.

Let us suppose that we have observations of a continuous time series $\{x(t)\}$ at discrete time points in the interval $(0, NT]$ and one observation Y_w of a related binary response Y at each period $((w-1)T, wT], w = 1, \ldots, N$. Then the purpose of this paper is to estimate a functional logit model to

forecast the binary response in future periods $((w^* - 1)T, w^*T](w^* > N)$ from the forecasting of the series $x(t)$ in such periods provided by a mixed ARIMA-FPCA model.

In order to formulate and to estimate a functional logit model based on functional principal component analysis, we propose to cut the observed series $x(t)$ in N periods of amplitude T, so that we have N sample paths of the following functional predictor (continuous time process):

$$\{X_w(s) = x((w-1)T + s) : s \in (0, T]; w = 1, \ldots, N\}, \tag{1}$$

and a sample of size N of the binary response given by $\{Y_w : w = 1, \ldots, N\}$ (see Figure 1).

Let us observe that the choice of the amplitude T is simple enough in practice when there is a well defined seasonal period as in the case of many real time series.

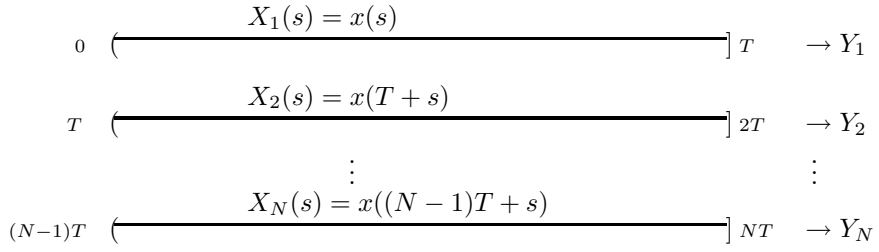

Fig. 1. Sample information obtained after cutting the original continuous time series

2 Functional logistic regression

The objective of the functional logistic regression (FLR) model is to explain a binary response variable Y in terms of a functional variable $X(s)$ whose sample information is given by a set of curves measured without error.

Let $X_1(s), \ldots, X_N(s)$ be a sample of curves of a functional variable $\{X(s) : s \in (0, T]\}$, obtained by cutting in periods of amplitude T the original predictor series $x(t)$, and let $Y_w (w = 1, \ldots, N)$ be the random observations of the binary response variable Y associated with the sample curves. Then, the FLR model is given by $Y_w = \pi_w + \varepsilon_w$, where ε_w are zero mean independent random errors with variance $\pi_w(1 - \pi_w)$, and π_w is the probability of response $Y = 1$ for a specific curve $X_w(s)$ modelled as

$\pi_w = \exp(l_w)/(1+\exp(l_w))$, with l_w being the logit transformation given by

$$l_w = \alpha + \int_0^T X_w(s)\,\beta(s)\,ds,\ w = 1,\ldots,N, \tag{2}$$

where α is a real parameter and $\beta(s)$ is a parameter function that has to be estimated. In terms of the logit transformations, the model can be equivalently seen as a functional generalized linear model [James, 2002].

As in the functional linear model [Ramsay and Silverman, 1997], it is impossible to obtain a direct estimation of the FLR model by using the usual likelihood or least squares methods. In addition functional data are usually observed only in a finite set of time points so that its true functional form has to be reconstructed from its discrete time observations by using an approximating procedure. Then, the most used solution for solving this estimation problem is based on assuming that the parameter function and the sample curves belong to a finite dimension space generated by a basis of functions $\{\phi_1(t),\ldots,\phi_p(t)\}$, so that they can be expressed in terms of the basis as

$$\beta(s) = \sum_{k=1}^{p} \beta_k \phi_k(s) \quad \text{and} \quad X_w(s) = \sum_{j=1}^{p} a_{wj}\phi_j(s). \tag{3}$$

Then, the functional model given by equation (2) is equivalent to a multiple logit model given in matrix form by $L = \mathbf{1}\alpha + A\Psi\beta$, with $L = (l_1,\ldots,l_N)'$, A the matrix that has the basis coefficients of the sample curves as rows, $\Psi = (\psi_{jk})_{p\times p}$ the one that has the L^2-usual inner products between the basic functions as entries, $\left(\psi_{jk} = \int_0^T \phi_j(s)\,\phi_k(s)\,dt\right)$, and $\beta = (\beta_1,\ldots\,\beta_p)'$ the vector of the parameter function basis coefficients.

Before estimating by likelihood the vector β, we have to compute the matrix A of sample curves basis coefficients. Let $x_w = (x_{w1},\ldots,x_{wm_w})'$ be the vector of observations of the *wth* sample curve $X_w(s)$ at m_w time points of the interval $((w-1)T, wT]$, $\forall w = 1,\ldots,N$. When discrete-time observations are considered to be measured without error, $x_{wk} = x_w(t_{wk})$, an interpolation method to estimate the basis coefficients can be used. On the other hand, if some error is considered in the observations, $x_{wk} = x_w(t_{wk}) + \varepsilon_{wk}$, least squares approximation is usually used for estimating the basis coefficients for a specific curve as $a_w = (a_{w1},\ldots,a_{wp})' = (\Phi'\Phi)^{-1}\Phi' x_w$, with $\Phi_{m_w\times p} = (\phi_j(t_{wk}))$. Let us observe that least squares approximation can be also applied when the functional variable is recorded at different time points for each individual (missing longitudinal data). On the other hand, taking into account the underlying nature of curves, different basis have been used in literature as for example, Fourier, Wavelets or Spline functions.

The problem is that likelihood estimation of the parameters of the logit model with design matrix $A\Psi$ is very unaccurate due to multicollinearity so that the estimated parameter function can not be used to stablish the true

relationship between the response and predictor variables [Escabias *et al.*, 2004].

3 Functional principal component logit model

In order to reduce dimension and to obtain better estimations of the parameter function, two different approaches based on FPCA of sample paths have been proposed in literature [Escabias *et al.*, 2004], so that the FLR model is reduced to a multiple one with a reduced number of functional principal components as covariates. In this paper we are going to perform FPCA of the sample paths $X_w(s)$ with respect to the usual inner product in $L^2((0,T])$.

Functional principal components of $X_w(s)$ are defined as N-dimensional vectors $\xi_j (j = 1, \ldots, N-1)$ with components

$$\xi_{wj} = \int_0^T (X_w(s) - \bar{x}(s)) f_j(s)\, ds,\ w = 1, \ldots, N,$$

where $\bar{x}(s)$ is the sample mean of the sample curves and the weight functions $f_j(s)\,(j = 1, \ldots, N-1)$ that define the functional pc's are the eigenfunctions of the sample covariance function of the sample curves whose associated positive eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_{n-1} \geq 0$ are the variances of the corresponding principal components (pc's).

Then, the sample curves admit the following orthogonal representation in terms of the sample pc's:

$$X_w(s) = \sum_{j=1}^{N-1} \xi_{wj} f_j(s),\ w = 1, \ldots, N.$$

By truncating this expression we obtain a reconstruction of the sample paths in terms of a reduced number of pc's that accumulate a certain percentage of the total variance $TV = \sum_{j=1}^{N-1} \lambda_j$.

It can be shown that if the sample paths belong to a finite space of $L^2(0,T]$ generated by a basis, their functional pc's are given by the standard principal components of the matrix $A\Psi^{1/2}$. If we denote by $\Gamma = (\xi_{ij})_{N\times p}$ the matrix whose columns are the pc's of the $A\Psi^{1/2}$ matrix, and G the one whose columns are its associated eigenvectors, then $\Gamma = \left(A\Psi^{1/2}\right) G$ and the weight functions that define the functional pc's are given by

$$f_j(s) = \sum_{k=1}^{p} f_{jk}\phi_k(s),\ j = 1, \ldots, p \tag{4}$$

with $F = (f_{jk})_{p\times p} = \Psi^{-1/2} G$.

Then, FLR model (2) can be equivalently expressed in terms of the pc's as

$$l_w = \alpha + \sum_{j=1}^{p} \xi_{wj}\gamma_j,\ w = 1, \ldots, N. \tag{5}$$

[Escabias *et al.*, 2004]

The functional principal component logistic regression (FPCLR) model is obtained by truncating model (5) in terms of a subset of pc's. If we consider the matrices defined before partitioned as follows

$$\Gamma = \left(\Gamma_{(q)} \middle| \Gamma_{(r)}\right),\ F = \left(F_{(q)} \middle| F_{(r)}\right),\ r+q=p,$$

then, the FPCLR model is defined by taking as covariates the first q principal components

$$L_{(q)} = \alpha_{(q)}\mathbf{1} + \Gamma_{(q)}\gamma_{(q)},$$

where $\alpha_{(q)}$ is a real parameter and $L_{(q)} = \left(l_{1(q)}, \ldots, l_{N(q)}\right)'$ with

$$l_{w(q)} = ln\left[\frac{\pi_{w(q)}}{1-\pi_{w(q)}}\right] = \alpha_{(q)} + \sum_{j=1}^{q} \xi_{wj}\gamma_{j(q)},\ i = 1, \ldots, N. \tag{6}$$

Finally, the likelihood estimation of the parameter function given by

$$\widehat{\beta}_{(q)}(s) = \sum_{j=1}^{p} \widehat{\beta}_{j(q)}\phi_j(s), \tag{7}$$

with the coefficient vector $\widehat{\beta}_{(q)} = F_{(q)}\widehat{\gamma}_{(q)}$ is more accurate than the one obtained with the original $A\Psi$ design matrix [Escabias *et al.*, 2004].

4 Mixed ARIMA-FPCA logit model

Let us observe that functional PCA provides an orthogonal expansion of the functional predictor $\{X(s)\}$ in terms of a set of deterministic functions (the principal factors) and random variables (the principal components).

In our case, the values ξ_{wj} $(w = 1, \ldots, N)$ of each sample principal component ξ_j can be seen as observations of a discrete time series at each period $((w-1)T, wT]$ of amplitude T where the original series $x(t)$ is observed. Then, in order to forecast the binary response in future periods $((w^*-1)T, w^*T](w^* > N)$, we propose the modelization of each principal component by an ARIMA model [Box and Jenkins, 1970]. The general expression of an ARIMA(p,d,q) model for the *jth* principal component ξ_j is given by $\Phi(B)(1-B)^d\xi_{wj} = \theta(B)\epsilon_{wj}$, where B is the backward shift operator, $\Phi(B)$ is the autoregressive operator defined as $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p$, $\theta(B)$ is the moving average operator given by $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$, and ϵ_{wj} is a white noise process for each j $(j = 1, \ldots, q)$.

The prediction model proposed in this paper is based on ARIMA forecasting of each of the q principal components selected for estimating the functional logit model. After estimating in the usual form these q ARIMA

models, we will be able to obtain forecasts for each principal component in the future periods $((w^*-1)T, w^*T]$, denoted by $\widetilde{\xi}_{w^*j}$.

Finally, the original series $x(t)$ is predicted in all the interval of time $((w^*-1)T, w^*T]$, by the principal component reconstruction of the process $\{X(s)\}$ in terms of the predicted principal component values

$$\widetilde{x}((w^*-1)T+s) = \widetilde{X}^q_{w*}(s) = \bar{x}(s) + \sum_{j=1}^{q} \widetilde{\xi}_{w^*j} f_j(s) \quad s \in [0,T],$$

and the estimated probabilities of success in the future periods $((w^*-1)T, w^*T]$ are predicted from the logit transformations

$$\tilde{l}_{w*(q)} = \hat{\alpha}_{(q)} + \sum_{j=1}^{q} \tilde{\xi}_{w^*j} \hat{\gamma}_{j(q)},$$

in terms of the ARIMA forecasts of the principal components $\tilde{\xi}_{w^*j}$.

5 Predicting the risk of drought

In order to illustrate the proposed Mixed ARIMA-FPCA logit model, we are going to predict the risk of drought in the future in terms of its evolution in the past by using as predictor the past evolution of temperatures, as in [Escabias *et al.*, 2005]. With this objective, let us consider a specific zone where drought has been tested monthly for several years by classifying a month as dry or not dry according to the definition of drought based on the amount of precipitations observed in this zone. That is, if it rains less that a certain percentile during a specific month, it is considered as a dry month meanwhile in the opposite case the month is considered as not dry. Then, if we define the binary variable $Y = \{0, 1\}$ as the one that takes value one in a specific month if it is not a dry month and zero in the opposite case, we have a monthly time series of binary values.

We have daily temperatures and precipitations observed in the *Estación Meteorológica del Departamento de Botánica de la Universidad de Granada* from 01/01/1992 to 12/31/2001. In this period the precipitation have been monthly accumulated (30 days period) and each month has been classified as dry if the accumulated precipitations in this month have been lower than a specific percentile of the precipitations observed in same month over all the years. In order to test the forecasting performance of mixed ARIMA-FPCA logit models we have considered different examples by using different percentiles (0.25 and 0.50) for defining the binary time series of drought.

As predictor time series $x\,(t)$ we have considered the daily temperatures cut at 30 days periods (T=30). In order to obtain the functional form of temperatures in each month we have considered the expansion of such functions as in (3) in terms of the basis of B-splines defined from the knots

$\{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 30\}$, and we have obtained the basis coefficients of each curve by least squares approximation from the discrete time observations of daily temperatures.

Once the predictor curves have been approximated from their discrete-time observations and the response observed in each one of the considered examples (percentiles 0.25 and 0.50), we have considered the first $N = 108$ observations to fit the Mixed ARIMA-FPCA logit model and the last 12 to validate the results. ¿From the fist $N = 108$ observations of monthly temperatures (predictor) and drought (binary response) we have fitted the FPCLR model with different number of functional pc's in the model. The percentages of variance explained by the first four functional pc's can be seen in Table 1. Once the functional pc's have been computed, we have modeled them as ARIMA's obtaining that only the two first pc's have such structure. We have considered the rest as white noises. ARIMA modelling of pc's can be seen in Table 1. After modelling the pc's we have obtained 12 steps ahead forecasts (12 months) of such time series.

pc	Exp. Var.	Cum. Var.	ARIMA Model	Estimated Parameters
ξ_1	87.17%	87.17%	$SARIMA(0,0,1) \times (0,1,1)_{12}$	$\theta_1 = -0,351239$
				$\Theta_1 = 0,567944$
ξ_2	4.13%	91.30%	$SARIMA(0,0,0) \times (0,1,1)_{12}$	$\Theta_1 = 0,894991$
ξ_3	2.26%	93.56%	White Noise ($\sigma = 5.44$)	
ξ_4	1.46%	95.02%	White Noise ($\sigma = 4.37$)	

Table 1. Percentages of explained variances (Exp. Var.), cumulated variances (Cum. Var.), and ARIMA modelling for the first four pc's.

In order to test the performance of mixed ARIMA-FPCA logit models we have obtained the estimated probabilities for the response (risk of drought) from the 12 predictions provided by ARIMA modelling of the first pc's (see Table 2). These probabilities have been obtained by using the estimated parameters of the logistic models with the first 1, 2, 3 and 4 pc's in the models with each one of the responses. All adjusted logit models have high deviance statistics with low p-values what shows that the models fit well and that the logit model is a good election for estimating this response. In each case the Mean Squared Error (MSE) between predictions and observed values have been obtained. The results can be seen in Table 2. It can be observed that the MSE of the models with the components that are modelled as ARIMA are always lower than the ones that include not modelled principal components.

2002	Y defined by 0.25 precentile					Y defined by 0.50 precentile				
Months	Dry	1 cp	2 cp's	3 cp's	4 cp's	Dry	1 cp	2 cp's	3 cp's	4 cp's
Jan	1	0.684	0.684	0.779	0.670	1	0.425	0.425	0.446	0.282
Feb	1	0.678	0.679	0.522	0.537	1	0.408	0.408	0.377	0.395
Mar	1	0.690	0.674	0.627	0.581	1	0.441	0.438	0.427	0.369
Apr	1	0.713	0.698	0.616	0.719	1	0.507	0.504	0.486	0.628
May	0	0.717	0.707	0.798	0.827	0	0.520	0.518	0.539	0.591
Jun	1	0.741	0.716	0.597	0.623	0	0.593	0.588	0.562	0.598
Jul	1	0.774	0.765	0.787	0.821	1	0.691	0.689	0.692	0.743
Oct	1	0.794	0.793	0.824	0.836	1	0.748	0.748	0.753	0.768
Sep	1	0.791	0.803	0.937	0.950	1	0.742	0.744	0.785	0.823
Oct	1	0.779	0.797	0.818	0.753	0	0.705	0.709	0.711	0.596
Nov	1	0.743	0.758	0.768	0.606	1	0.600	0.603	0.603	0.363
Dec	1	0.717	0.748	0.822	0.847	1	0.519	0.525	0.543	0.586
MSE		0.108	0.107	0.130	0.142		0.248	0.247	0.247	0.267

Table 2. Observed values of the response (no drought) and estimated probabilities of no drought for the mixed ARIMA-FPCA logit model in each one of the selected responses (percentiles 0.25 and 0.50) for the models with the first 1, 2, 3 and 4 pc's.

6 Acknowledgments

This research has been supported in part by Project MTM2004-5992 of the Ministerio de Ciencia y Tecnología, Spain.

References

[Box and Jenkins, 1970]G.E.P. Box and G.M. Jenkins. *Time Series Analysis Forecasting and Control*. Holden Day, San Francisco, 1970.

[Escabias *et al.*, 2004]M. Escabias, A.M. Aguilera, and M.J. Valderrama. Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, pages 365–384, 2004.

[Escabias *et al.*, 2005]M. Escabias, A.M. Aguilera, and M.J. Valderrama. Modeling environmental data by functional principal component logistic regression. *Environmetrics*, pages 95–107, 2005.

[James, 2002]G.M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, pages 411–432, 2002.

[Ramsay and Silverman, 1997]J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer-Verlag, New York, 1997.

[Valderrama *et al.*, 2002]M. J. Valderrama, F. A. Ocaña, and A. M. Aguilera. Forecasting pc-arima models for functional data. In W. Härdle and B. Rönz, editors, *Proceedings in Computational statistics*, pages 25–36, 2002.