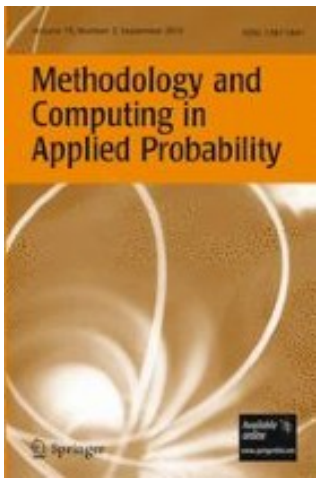
Functional Wavelet-Based Modelling of Dependence Between Lupus and Stress

- Ana M. Aguilera; Manuel Escabias; Francisco A. Ocaña; Mariano J. Valderrama

- Functional Wavelet-Based Modelling of Dependence Between Lupus and Stress

- *Methodol Comput Appl Probab (2015) 17:1015–1028*
- DOI: https://doi.org/10.1007/s11009-014-9424-5



Methodol Comput Appl Probab (2015) 17:1015–1028
DOI 10.1007/s11009-014-9424-5

Functional Wavelet-Based Modelling of Dependence Between Lupus and Stress

Ana M. Aguilera · Manuel Escabias · Francisco A. Ocaña · Mariano J. Valderrama

Received: 19 September 2013 / Revised: 13 August 2014 /
Accepted: 24 August 2014 / Published online: 5 September 2014

Abstract The power of functional linear regression to estimate a set of curves from others involved is studied in this work in the context of life sciences. The objective is to determine the relationship between the degree of lupus and the level of stress for patients suffering this autoimmune disease. Daily stress and lupus curves have a strong local behavior with missing data those days that a patient does not answer the corresponding test. Because of this, wavelet smoothing with an appropriate thresholding rule is considered. Then, functional principal component analysis of the response and predictor variables is used to reduce the dimension and solve the multicollinearity problem that affects the estimation of the functional linear regression model with functional response. Model selection is solved by using a criterion that selects those pairs of response/predictor components that explain the highest proportions of response variability. The performance of the proposed functional model is tested on simulated and real data.

Keywords Functional regression · Functional PCA · Wavelet approximation · Lupus

A. M. Aguilera
Department of Statistics and O.R. Faculty of Sciences, University of Granada, Campus de Fuentenueva, 18071 Granada, Spain
e-mail: aaguiler@ugr.es

M. Escabias (✉)
Department of Statistics and O.R. Faculty of Communication and Documentation, University of Granada, Campus de Cartuja, 18071 Granada, Spain
e-mail: escabias@ugr.es

F. A. Ocaña · M. J. Valderrama
Department of Statistics and O.R. Faculty of Pharmacy, University of Granada, Campus de Cartuja, 18071 Granada, Spain

F. A. Ocaña
e-mail: focana@ugr.es

M. J. Valderrama
e-mail: valderra@ugr.es

Mathematics Subject Classifications (2010) 60G12 · 60G17 · 62H25 · 62J05 · 62P10

1 Introduction

In many applications in life sciences is usual to have observations of vital signs of individuals over time. This is the case studied in this work in which daily stress and lupus levels of patients suffering this immunological disease have been observed.

Systemic lupus erythematosus (SLE) is an autoimmune disease that occurs when your body's immune system attacks your own tissue and organs. Inflammation caused by lupus can affect many different body systems, including your joints, skin, kidneys, blood cells, brain, heart and lungs. The course of the disease is unpredictable, with periods of illness (called flares) alternating with remissions. The specific causes of lupus are unknown and their identification is a very active area of research around the world. To date, there is some evidence that supports a number of possible factors that lead to the development of lupus. On the one hand, research indicates that SLE may have a genetic link. On the other hand, environmental factors also seem to play some role because the immune system in patients with lupus is more easily activated by external factors, such as viruses, ultraviolet light and drugs. Finally, stress has also been associated with the onset of lupus. It may be that certain genetically and hormonally susceptible persons, who have been exposed to just the right amount of environmental factors, could trigger the onset of the illness after significant life stresses (Peralta-Ramírez et al. 2006). In this line of research, the objective of this paper is to estimate continuously in time the degree of lupus of a patient from his/her level of stress and to establish the relationship between these two diseases.

The sample units associated to the variables stress and lupus in a period of time T are curves observed only those days in which the patient answers the tests provided by the doctor to measure the degree of both afflictions. Therefore, we will use functional data analysis (FDA) methodologies to model them (Ramsay and Silverman 2005). A functional linear regression model with functional response is considered in this paper to solve the problem of estimating the degree of lupus level in a period of time from daily stress level evolution in the same period. The results in this paper willshow the power of FDA to solve medical problems. The results of this paper are very important in medicine because they will highlight the importance of controlling lupus based on controlling the level of stress.

In order to estimate the model we have daily observations of stress and lupus measured from a questionnaire that patients must complete every day. On the one hand, daily stress level was measured by using the Spanish translation and adaptation of the Daily Stress Inventory (DSI) (Brantley et al. 1987) carried out by Peralta et al. (2002). This instrument measures the degree of stress produced by different stressful daily events in the last 24 hours. It contains 20 items that are categorized from 0 to 6 with 0 indicating that no stress was experienced and 6 indicating that the event caused panic. This is a well known measure of stress that presents high validity for detecting change (Peralta et al. 2002). On the other hand, daily level of lupus was measured by using the SLE Symptoms Inventory (SLESI) elaborated by the group of medical specialists in the Systemic Autoimmune Disease Unit of the Internal Medicine Service at the University Hospital Virgen de las Nieves in Granada. It refers to 8 symptoms suggestive of SLE activity, such as loss of appetite, joint pain, general malaise, fever, tiredness or fatigue, skin rash, difficulty breathing, and abdominal symptoms. These items are categorized from 1 to 10 according to the degree of intensity of symptoms on that day. The SLE Disease Activity Index (SLEDAI) was used to assess lupus activity. It consists of 24 descriptors with pre-assigned severity weights. The total SLEDAI score can

range from 0 (no activity) to 105 (maximum activity). This is a well known measured with high internal consistency, high degree of reliability that has been shown to be sensitive to changes in lupus activity measured by the treating physician.

An estimation approach based on nonparametric estimation of functional principal component analysis (FPCA) was developed in Yao et al. (2005). A review of different functional regression models with applications to analyze dose-response data with functional responses from an experiment on the age-specific reproduction of medflies can be seen in Chiou et al. (2004). The case of functional regression with scalar response and functional predictor was studied by Cardot et al. (1999). Functional PLS with basis expansions of sample curves has been recently introduced with applications in the field of chemometrics Aguilera et al. (2010). As an alternative to standard longitudinal data methods used in the biological sciences, linear models where the response is a function and the predictors are vectors were first studied in Faraway (1997). Residual analysis and diagnostics for such functional models were introduced in Shen and Xu (2007). Different PCA-based estimation of the functional logit model were introduced in Escabias et al. (2004) for estimating a categorical response from a functional predictor. These principal component logit models were applied to predict the probability of lupus flare from time evolution of stress level.

The problem of forecasting a continuous-time stochastic process on an entire time-interval in terms of its recent past was also solved by using functional regression models. Principal component prediction models were first performed and adapted for predicting continuous-time series Aguilera et al. (1999). These models were successfully applied to forecast curves of pollen concentration from continuous evolution of temperatures (Valderrama et al. 2010). Different wavelet approaches for estimating autoregressive Hilbert processes were also developed with the same objective in Antoniadis and Sapatinas (2003).

A wavelet-based functional principal component estimation of functional linear regression models with functional responses is proposed in this paper. First, wavelet smoothing with thresholding is used to reconstruct each curve of stress and lupus from daily observations. Second, two functional PCAs, one for the response curves (lupus) and another for the predictor curves (stress) are computed to reduce the dimension of both functional data spaces. Third, functional linear regression is reduced to multivariate principal component regression between response and predictor principal components (PC's). Finally, several selection model procedures that take into account both, explained variability and correlation between response/predictor PC's, are developed.

Apart from this introduction, the content of the paper is divided into four sections. Basic ideas on nonlinear wavelet smoothing of curves from discrete observations are outlined in Section 2. Formulation, principal component estimation, selection of variables and interpretation of the functional linear regression model with functional response are studied in Section 3. A simulation study for analyzing the good performance of the proposed methodology for estimating the functional response and providing an accurate estimation of the functional parameter is developed in Section 4. Finally, the relationship between lupus and stress curves is established in Section 5 by selecting and fitting an optimal wavelet-based functional principal component regression model to daily observations of lupus and stress.

2 Wavelet Smoothing of Functional Data

A functional variable X is characterized because its values belong to a function space with a metric induced by an inner product. Just as in the case of an scalar variable, a sample of functional data is obtained from the observation of n identically distributed functional

variables $X_1, X_2, \ldots, X_n$. The observed values related to a functional variable can be curves, surfaces or other functions defined on a continuous argument. The argument is often time, but may also be other type as for example wavelenght in chemometric applications or spatial location in spatial data analysis.

In practice, the sampling units are functions of which only discrete observations are available in a finite set of points that may be unevenly spaced and different for the sample individuals. This means that, given a sample $\{x_w(t) : t \in T; w = 1, \ldots, n\}$ of n records of a functional variable X, we have a vector of discrete observations $\mathbf{x_w} = (x_{wk})_{k=1,\ldots,K_w}$ for each sample curve x_w at a finite set of points $(t_{wk} : k = 1, \ldots, K_w)$. Therefore, the first step in FDA is to reconstruct the true functional form of each sample function from the discrete information available. A way to solve this problem consists of assuming a basis function expansion for each observed sample curve. The dimension of the basis and the basis functions must be chosen according to the characteristic of the functional data. Most frequently used systems are Fourier basis for approximating periodic data, B-spline basis for controlling the degree of smoothness of the curve and wavelet basis for functions with sharp local features. If the observed values are errorless, interpolation procedures are used to estimate the basis coefficients. On the other hand, if the observations are affected by noise, least square smoothing or orthogonal projection is used Ramsay and Silverman (2005). When the data are smooth functions observed with error the approximated sample curves do not control the degree of smoothness. In order to improve the estimation of FDA methodologies in this case, different approaches based on penalized estimation with B-spline basis expansions of sample curves were recently introduced Aguilera et al. (2008), Aguilera and Aguilera-Morillo (2013a, b) and Aguilera-Morillo et al. (2013).

In the application developed in this paper we have daily observations of levels of stress and lupus in a period of 100 days with missing data those days that a patient does not answer the test. Daily levels of lupus and stress are measured in based to different tests provided by the doctor to each patient. Taking into account the subjectivity of such measures that can be irregularly distributed and have high variability, we consider that discrete-time observations are affected by noise. As both, lupus and stress curves, have a strong local behavior we propose to approximate the curves by using wavelet basis expansions.

By considering dilations and translations of a suitable mother wavelet, the wavelet expansion provides a decomposition of a function into orthogonal signal components at different resolution levels that it is called multiresolution analysis (MRA). The advantages of this wavelet representation derive from the ability of wavelets to represent locally non-smooth functions with only a relatively small number of coefficients. They form orthonormal basis and enable multiresolution analysis by localizing a function in different phases of both time and frequency domains simultaneously. In this paper we estimate basis coefficients by orthogonal projection of each predictor and response sample curve on basis of wavelets on bounded intervals. For simplicity, we summarize wavelet approximation for the functional predictor X representing stress level.

Let ϕ and ψ be the scaling (mother) and wavelet (father) functions for an orthogonal MRA of the space of squared integrable functions $L^2[\mathbb{R}]$. In our case we consider the adaptation of the wavelet analysis in a bounded interval (see Mallat (1998) and Daubechies (1988)). For simplicity, the space $L^2[0, 1]$ is considered in this work without loss of generality.

Let us consider the orthonormal basis systems

$$\phi^*_{j,k} = \chi_{[0,1]} \sum_{l \in \mathbb{Z}} \phi_{j,k-l2^j} \quad \text{and} \quad \psi^*_{j,k} = \chi_{[0,1]} \sum_{l \in \mathbb{Z}} \psi_{j,k-l2^j},$$

obtained, respectively, from the dilations and translations of the scaling and wavelet functions

$$\phi_{j,k}(t) \;=\; 2^{j/2}\,\phi\left(2^j t - k\right) \quad \text{and} \quad \psi_{j,k}(t) \;=\; 2^{j/2}\,\psi\left(2^j t - k\right).$$

Then, projecting each sample path $x(t)$ at resolution level J we obtain a first orthogonal approximation in terms of scaling coefficients

$$P_J x(t) = \sum_{k=0}^{2^J-1} s_{J,k}\phi^*_{J,k}(t) \qquad \forall\, t \in [0,1], \tag{1}$$

where $s_{J,k} = \int_0^1 x(t)\phi^*_{J,k}(t)dt$.

A more sparsely decomposition of $P_J X$ can be obtained as follows by applying the discrete wavelet transform (DWT) to the vector of scaling coefficients from an initial resolution level $J_0 < J$

$$P_J x(t) = \sum_{k=0}^{2^{J_0}-1} s_{J_0,k}\phi^*_{J_0,k}(t) + \sum_{j=J_0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi^*_{j,k}(t), \tag{2}$$

where $d_{j,k} = \displaystyle\int_0^1 x(t)\psi^*_{j,k}(t)dt$.

Unfortunately, the discrete time observation assumption makes impossible to exactly compute the scaling and thus the wavelet coordinates for the expansions given in Eqs. 1 and 2. Nevertheless, once the scaling coordinates $s_{J,k}$ are approximated, an approximation to the wavelet coordinates in Eq. 2 are directly obtained by applying the DWT to the approximated scaling coordinates. Hence, the scaling coordinates $s_{J,k}$ must be firstly approximated following one of the techniques proposed in wavelet theory. An usual and efficient way for approximating $s_{J,k}$ is

$$s_{J,k}(w) \;\approx\; \widetilde{s}_{J,k}(w) \;=\; K^{-1/2}x_w(t_k)\,, \quad \forall\, k = 0,\ldots,K\,,$$

when we have $K = 2^J$ equally spaced time points in the observed interval for each sample curve (Mallat 1998). In the case of unequally spaced points we can use first an interpolation procedure to compute the values of each sample path at these equally espaced knots. Wavelet-based estimators can be computed extremely quickly because the DWT and its inverse is computed in O(K) operations.

When the discrete observations of sample curves are subject to noise it is appropriate to use a nonlinear smoothing approach that consists of computing the DWT of noisy observations and thresholding it by deleting the small wavelet coefficients and shrinking the large ones. This provides an economical wavelet expansion with few non-zero coefficients, even if the approximated curve displays sharp local features.

The thresholded wavelet estimator of a sample curve $x(t)$ is given by Eq. 2 where the coefficients $d_{j,k}$ are replaced by $T_\theta(d_{j,k})$ with T_θ being the thresholding rule, J_0 the starting scaling level for thresholding $J_0 = \log_2(\log(K)) + 1$, with K being the number of observation knots and θ is the universal threshold $\theta = \sigma_e\sqrt{2\log(K)}$ with σ_e being the estimation of the observation noise deviation $\sigma_e = median(|d(J-1,k)| : k = 0,\ldots,2^{J-1}-1)/.6745$. The thresholding rules most frequently used in wavelet literature are hard thresholding, soft thresholding and non-negative garrote thresholding.

Thresholded wavelet estimators should adapt well to different degrees of smoothness and regularity in the function being estimated. In Fig. 1 we have the wavelet approximation with Symmlet 4 family and resolution level J=8 next to its thresholded estimators by using soft

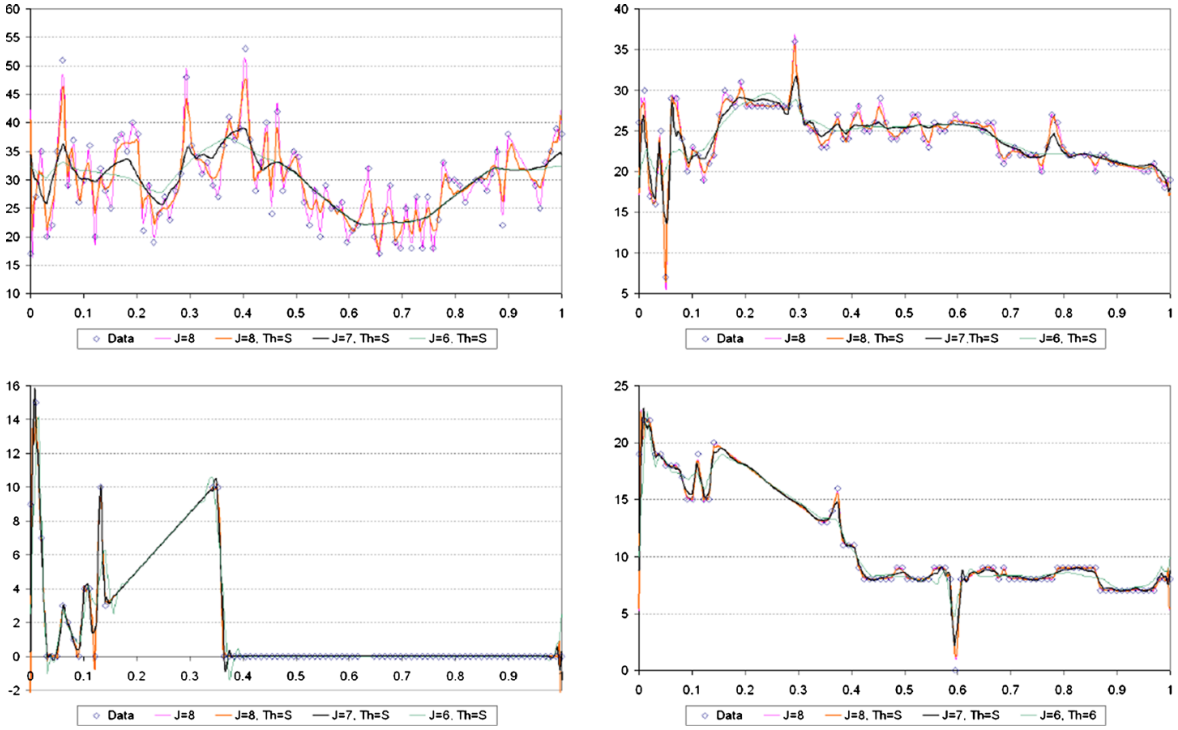


Fig. 1 Scaterplot, wavelet projection with Symmlet 4 family and resolution level J=8 (*pink*), and nonlinear wavelet smoothing with soft thresholding at resolution levels J=8 (*red*), J=7 (*black*) and J=6 (*blue*) for the curves of stress (*left*) and lupus (*right*) of different patients

thresholding rule at different resolution levels for the curves of lupus and stress of different patients.

3 Functional Regression with Functional Response

As we set in the introduction, the aim of this paper is to estimate a linear functional regression model for a random functional response variable $Y = \{Y_w(s) : s \in S, w \in \Omega\}$ in terms of the observed values of a functional predictor variable $X = \{X_w(t) : t \in T, w \in \Omega\}$ where $(\Omega, \mathcal{A}, \mathcal{P})$ is a probability space, and T and S are real intervals.

3.1 Model Formulation

The sample information we have consists of pairs of square integrable curves $\{(x_w(t), y_w(s)), w = 1, \ldots, n; t \in T, s \in S\}$ that can be seen as realizations of the functional predictor and response variables, X and Y, in the real intervals T and S, respectively.

The functional linear regression model to estimate the functional response $Y(s)$ in terms of the functional predictor $X(t)$ can be formulated as

$$y_w(s) = \alpha(s) + \int_T \beta(t, s)x_w(t)dt + \varepsilon_w(s) \quad s \in S, \tag{3}$$

where $\beta(t, s)$ is the functional parameter and $\{\varepsilon_w(s) : w = 1, \ldots, n; s \in S\}$ are independent and centered random errors.

As in any other functional regression model, in order to estimate model (3) we have to solve two important problems. First, the estimation of the parameter function is an ill-posed

problem due to the infinite dimension of the response and predictor function spaces (Ramsay and Silverman 2005). Second, the sample curves are only observed at a finite set of time points. This problem is solved by using wavelet basis expansions of sample curves so that the functional model becomes a multivariate linear regression model of response sample curve wavelet coefficients on predictor sample curve wavelet coefficients. But the estimation of this multivariate linear model usually presents new problems. On one hand, high correlations between the predictor basis coefficients could provide an inaccurate estimation of the parameter function. On the other hand, the number of predictor variables may be too high if many coefficients are needed to approximate the predictor curves.

Different reduction dimension approaches based on using as predictors of the functional model a set of uncorrelated variables have been considered for different functional regression models. Generalizations of PCR and PLS to the functional case have been formulated to solve this problem. In this paper we propose to use a functional PCA wavelet-based approach that transforms the functional regression problem into linear regression of a reduced set of principal components of the functional response variable on a reduced set of principal components of the functional predictor.

3.2 Functional Principal Component Estimation

In order to estimate model (3), we will consider the principal component decomposition of both predictor and response sample curves that transforms functional regression into multivariate principal component regression.

Let briefly summarized the basis ideas on PCA of a functional variable X with values in the space $L^2[T]$ of square integrable functions on the interval T. Functional principal components are defined as uncorrelated generalized linear combinations of the functional variable with maximum variance. That is, the jth principal component (PC) is given by $\xi_{wj} = \int_T (x_w(t) - \bar{x}(t))\, f_j(t)\, dt$, where the functional factor loadings $\{f_j(t) : \quad j = 1, \ldots, n-1\}$ are computed as the eigenfunctions of a second integral equation whose kernel is the sample covariance function associated with the functional variable X.

Then, we obtained for each sample curve an orthogonal principal component decomposition $x_w(t) = \bar{x}(t) + \sum_{j=1}^{n-1} \xi_{wj} f_j(t)$, where $\bar{x}(t)$ is the sample mean of the functional predictor X. The principal components are ordered in terms of their explained variability, so that truncating this expansion in terms of the first q principal components we obtain an optimal principal component reconstruction whose proportion of explained variance is given by $\left(\sum_{j=1}^{q} \lambda_j\right) \Big/ \left(\sum_{i=1}^{n-1} \lambda_i\right)$, being λ_j the explained variance of the j-th principal component.

Let us consider also the principal component decompositions of the response functional variable given by $y_w(s) = \bar{y}(s) + \sum_{j=1}^{n-1} \eta_{wj} g_j(s)$. Then, it is easy to prove that the functional linear regression becomes to linear regression for each PC of the response Y on all PC's of the predictor X

$$\eta_{wj} = \sum_{i=1}^{n-1} \xi_{wi} \nu_{ij} + \epsilon_{wj}, \tag{4}$$

and the functional parameter is given by $\beta(t,s) = \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \nu_{ij} f_i(t) g_j(s)$.

Then, by truncation and linear least squares estimation of the parameters ν_{ij} we obtain the following prediction equation for the response:

$$\hat{y}_w(s) = \bar{y}(s) + \sum_{j=1}^{J} \hat{\eta}_{wj} g_j(s) = \bar{y}(s) + \sum_{j=1}^{J} \left(\sum_{i \in I_j} \frac{\sigma_{ij}}{\sigma_i^2} \xi_{wi} \right) g_j(s), \tag{5}$$

with σ_{ij} the corresponding element of the sample cross-covariance matrix of predictor and response principal components and σ_i^2 the sample variance of the predictor PC ξ_i.

Wavelet basis expansion of each predictor and response sample curve will be considered for estimating FPCA of curves with high variability as lupus and stress. Once we have an orthonormal basis expansion for each curve, FPCA is equivalent to multivariate PCA of the matrix of basis coefficients. See Ocaña et al. (2007) for a detail study of these results in the more general context of Hilbert valued functional variables.

Let us suppose that the wavelet approximations of the response and predictor sample curves are given by $y_w = \sum_{q=1}^{Q} b_{wq}\varphi_q, \quad x_w = \sum_{p=1}^{P} a_{wp}\vartheta_p$, where $\{\varphi_q\}_{q=1}^{Q}$ and $\{\vartheta_p\}_{p=1}^{P}$ are orthonormal wavelet basis. Then, we obtain the following orthonormal wavelet estimation of the parameter function $\hat{\beta}(t,s) = \sum_{p=1}^{P}\sum_{q=1}^{Q}\hat{\beta}_{pq}\vartheta_p(t)\varphi_q(s)$, where the matrix of basis coefficients is given by $\hat{\beta} = (\hat{\beta}_{pq}) = \mathbf{V}^{\mathbf{X}}\boldsymbol{\nu}(\mathbf{V}^{\mathbf{Y}})'$, with $\boldsymbol{\nu} = (\nu_{ij}) = (\frac{\sigma_{ij}}{\sigma_i^2})$ and $\mathbf{V}^{\mathbf{X}}$ and $\mathbf{V}^{\mathbf{Y}}$ being the matrices of eigenvectors of the covariance matrices of $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$, respectively.

3.3 Model Selection

In order to select the optimal functional principal component regression (FPCR) model we have to choose an optimal set of J PC's η_j of the functional response Y and regress each response PC η_j on an optimal set of I_j PC's of the functional predictor X. In PCR it is known that principal components with smaller variances could be highly correlated with the response so that introducing PC's in a regression model by variability order could be inappropriate. Because of this in this paper we introduce a new criterion that introduces pairs of PC's in the FPCR model based on both, explained variability and correlation.

The R^2 coefficient associated to linear model (5) can be decomposed as $R^2 = \sum_{j=1}^{J}\sum_{i\in I_j} P(j,i)$, where $P(i,j) = \sum_{j=1}^{J}\sum_{i\in I_j}\frac{\alpha_j\rho^2(\eta_j,\xi_i)}{Var(Y)}$ is the proportion of response variance explained by each pair (η_j,ξ_i) of response/predictor principal components, with $\alpha_j = Var(\eta_j)$ and $Var(Y) = \sum_{j=1}^{n-1}\alpha_j$.

Taking into account this decomposition of the multiple linear correlation coefficient, we propose in this paper a selection model criterion based on selecting pairs of response/predictor PC's in based to the priority order established by their proportions of explained variances $P(*,*)$. With respect to the pairs of PC's needed for estimating the functional parameter β, we consider two different possibilities. One one hand, all possible pairs are considered (method *a*). On the other hand, pairs with non-significant correlation (t-ratio test) are leaved out (method *s*). Finally, the number of pairs of response/predictor PC's can be selected by minimizing one of the following statistics.

- Leave-one-out cross validation error $CVMSE = \frac{1}{n}\sum_{w=1}^{n}\|y_w - \hat{y}_{(-w)}\|^2$, where $\hat{y}_{(-w)}$ is the predicted curve computed by eliminating the sample curve y_w in the sample.
- Mean square error given by $MSE = \frac{1}{n}\sum_{w=1}^{n}\|y_w - \hat{y}_w\|^2$, with $\hat{y}_w$ being the predicted curve.
- BIC statistics defined as $BIC = MSE + \frac{\log n}{n}Ps_e^2$, with P being the number of parameters, s_e^2 the residual variance given by $s_e^2 = \frac{nP_{max}ECM}{n-P_{max}}$, and P_{max} being the maximum number of parameters.
- C_p statistic given by $C_p = MSE + 2Ps_e^2$.
- Integrated error of the estimated parameter function (only for simulation studies) defined as $\text{bE} = \|\beta - \widehat{\beta}\|^2$.

3.4 Interpretation

As the principal component estimation of functional linear regression turns the functional model into a multivariate multiple linear model, we propose an interpretation of the relation between the response and the predictor variables based on the estimated scalar parameters.

From the principal component representation of the functional predictor we deduce that one unit increment in the i^{th} PC of the predictor variable produces an increase of each predictor curve according to the i^{th} weight function. That is, $\triangle\xi_i = 1 \Longrightarrow \triangle X(t) = f_i(t)$. On the other hand, from regression Eq. 4 we have that one unit increase in the i^{th} PC of the predictor variable produces also an increment of v_{ij} units in the j^{th} PC of the response variable. As a result, from Eq. 5 it follows that each response curve is increased according to the j^{th} weight curve multiplied by the parameter v_{ij}. This means that $\triangle\xi_i = 1 \Longrightarrow \triangle Y(s) = v_{ij} g_j(s)$. This interpretation allows us to establish in practice the effect of the change in the evolution of a functional predictor on the evolution of a functional response.

4 Simulation Study

In order to study the good performance of the proposed wavelet-FPCA estimation of the functional regression model, we have developed a simulation study. We randomly generated $n = 100$ sample curves of the following functional predictor (James et al. 2000): $x_w(t) = \sum_{p=1}^{14} a_{wp}\,\vartheta_p(t) + \gamma_w, \quad t \in [0, 1]$, with $a_p \rightsquigarrow \mathcal{N}(0, |10 - p|)$, $\vartheta_{2r-1}(t) = \sin(2\pi rt)$ and $\vartheta_{2r}(t) = \cos(2\pi rt), \quad r = 1, \ldots, 7$. The functional response curves were simulated by the functional model (3) on the interval [0, 1], with the functional parameter given by

$$\beta(s, t) = s\ \sin(2\pi\, t) + \cos(4\,\pi\, t) \quad s, t \in [0, 1],$$

and the errors $\epsilon(s)$ randomly generated by zero mean independent normal distributions. The error variances were fixed to control the signal to noise ratio at each time point. A contour map of the original simulated parameter function can be seen in Fig. 2.

In order to fit the functional linear model, we simulated discrete time observations of the functional predictor at time points $t_i = i/20 \quad (i = 0, \ldots, 20)$ and the functional response at $s_j = j/16 \quad (j = 0, \ldots, 16)$. The experiment was repeated 400 times.

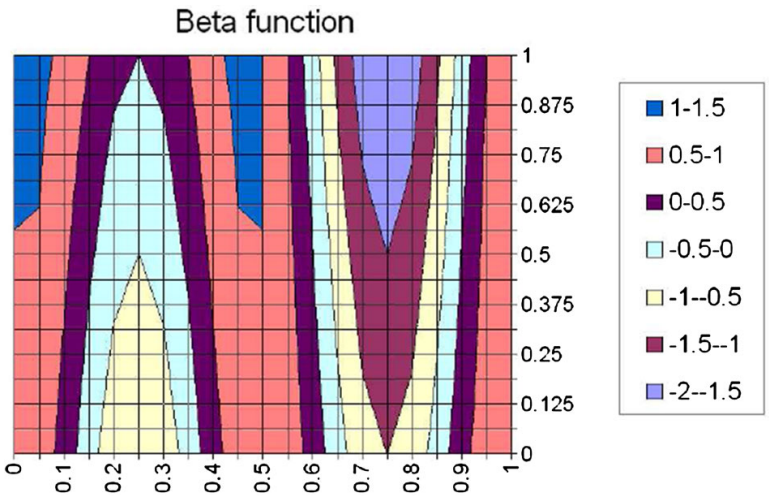


Fig. 2 Contour map of the simulated functional parameter $\beta(t, s)$

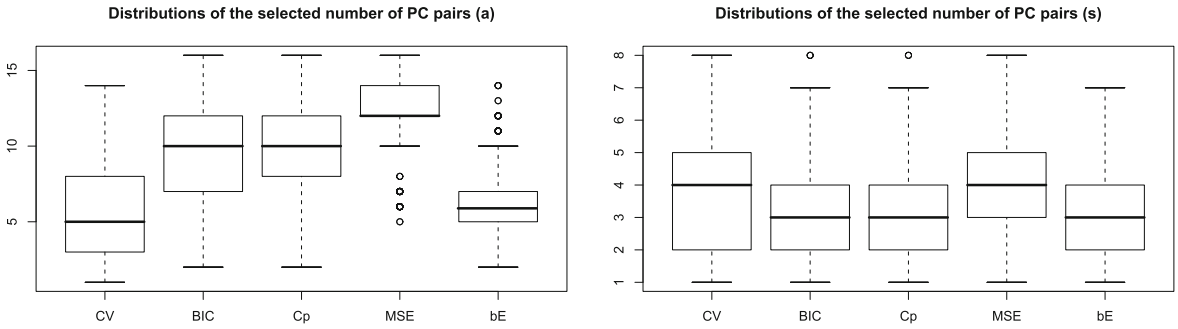


Fig. 3 Box plots for the distribution of the number of pairs of PC's selected with each model selection criterion

The wavelet-FPCA based approach introduced in this paper to fit the functional regression model with functional response had the following steps for each of the 400 trials:

- Wavelet approximation of response sample curves (resolution level 4) and predictor sample curves (resolution level 5) using D2 wavelet family.
- FPCA estimation of wavelet approximations of the predictor and the response variables.
- Model selection based on optimal pairs of response/predictor PC's. We considered two methods (it a and *s*) for selecting the pairs of PC's to be included and five criteria for selecting the number of pairs. To form pairs of response/predictor PC's, noise principal components were eliminated.
- Model estimation based on linear regression between the principal components of each selected pair.

The numerical results for the 400 simulated trials can be seen in Figs. 3, 4 and 5.
Based on the results of the simulation study we can draw the following conclusions:

1. The priority order established by $P(*, *)$ exhibits a good estimation performance with both methods *a* and *s* considered in this paper for selecting the pairs of PC's needed to get an accurate estimation of the functional parameter.
2. Considering only PC pairs with significant correlation (method *s*) provides estimations of β much more parsimonious than method a with a not excessive cost in error.
3. Taking into account the computational simplicity of BIC and Cp, they would be a good choice for method *s*.

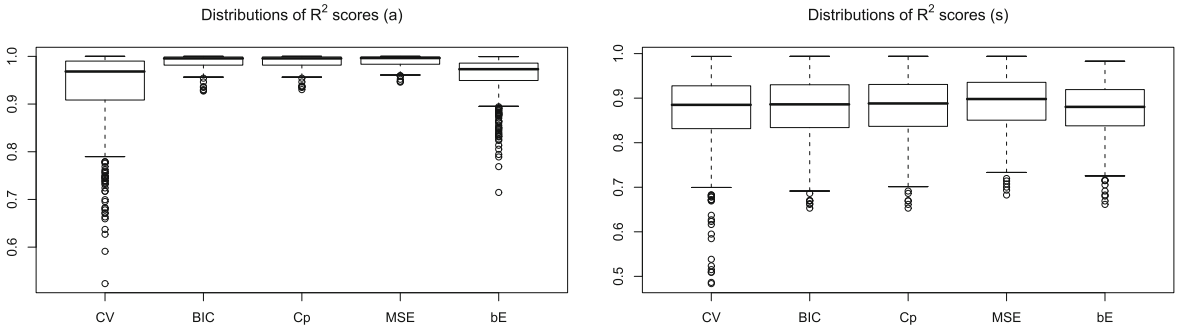


Fig. 4 Box plots for the distribution of the R^2 coefficientes of the optimal models selected with each selection criterion

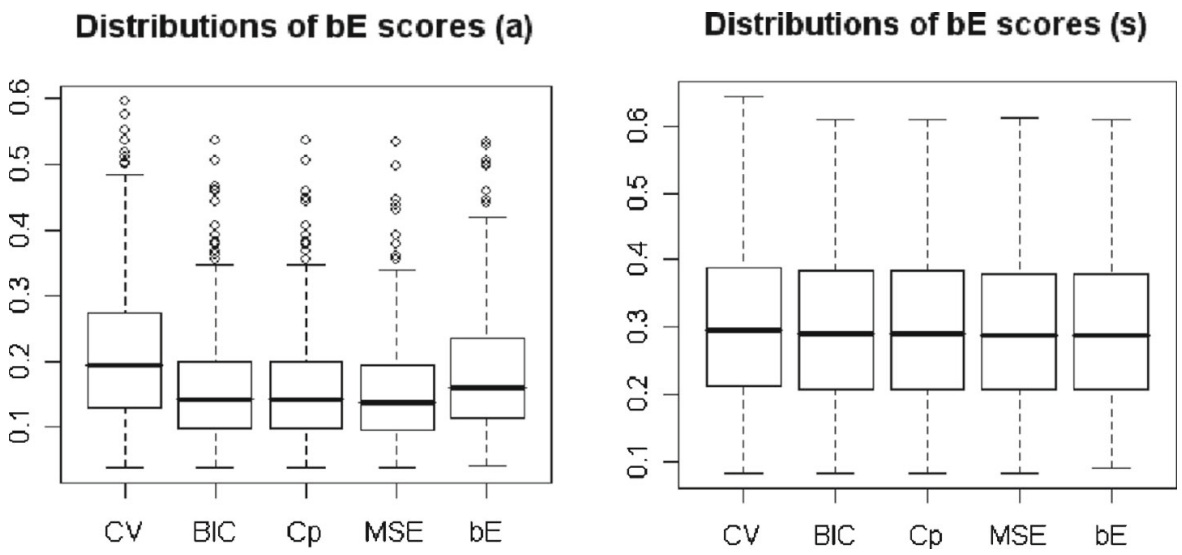


Fig. 5 Box plots for the distribution of the integrated errors bE of the parameter functions estimated by the optimal models selected with each selection criterion

5 Application with Immunology Data

In order to estimate the relation between lupus and stress we have daily observations of levels of stress and lupus in a period of 100 days for a sample of 56 patients with missing data those days that a patient does not answer the test. The sample of 56 patients has been divided in two, a training sample of size 40 to fit the model and a test sample of size 16 to evaluate the predictive ability of the estimated model.

During the first week, the patients were recruited by the internist at the outpatient clinic for autoimmune disease. When the patient attended his routine checkup, s/he was informed about this study on the effects of emotional status and lupus, and invited to participate (100% agreed). In the first session, the study was explained in detail, and subjects were asked to sign the consent form. Each subject underwent a clinical interview to find out basic data like age, education level, and diverse emotional problems occurring in his or her life. Each patient was given 31 copies of the 20-item version of the DSI and the SLESI. The subjects were told that they would have to complete the DSI and the SLESI at the end of every day for 100 days. Every month they were mailed an envelope containing the 31 questionnaires corresponding to that month and an empty stamped envelope so they could return the forms they had already completed. Furthermore, they were contacted by telephone every month, so we could find out about any problems that arose in completing the questionnaires.

The aim of this application is to estimate the functional response variable $\{Y(s) : s \in [0, 100]\}$ representing lupus evolution from the functional predictor $\{X(t) : t \in [0, 100]\}$ representing stress evolution for lupus patients. This problem is solved in this work by fitting a FPCR model based on wavelet approximation of lupus and stress curves.

The first step was to approximate lupus and stress curves by using Symmlet 4 wavelet family with resolution level J=7. Soft thresholding with universal threshold was used for denoising and smoothing.

The second step was to perform functional PCA of both set of smoothed curves in the observed interval [0, 100] from multivariate PCA of the associated wavelet coefficients. To facilitate interpretation, Fig. 6 shows plots of the overall mean functions of stress and lupus curves and the functions obtained by adding and substracting a suitable multiple of each of the first two PC curves that define the main modes of variation in the stress and lupus curves.

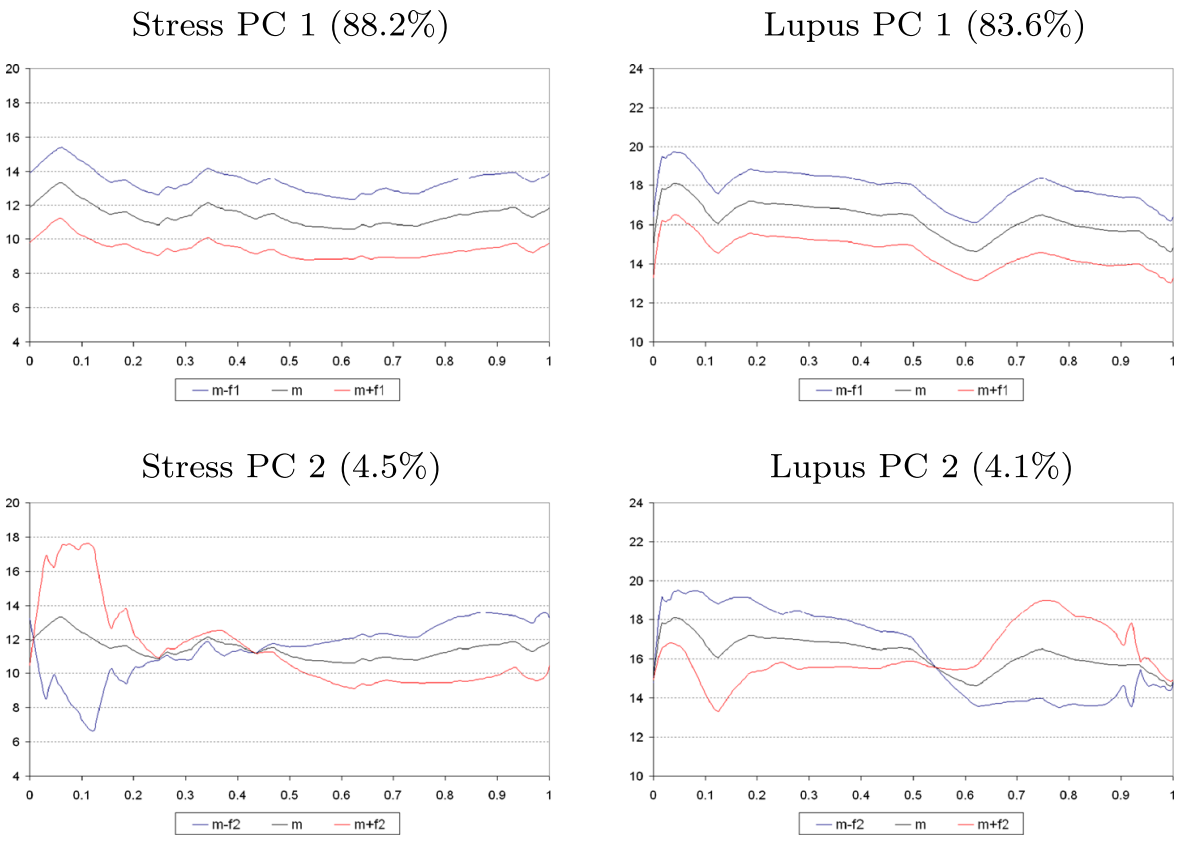


Fig. 6 The mean curves (*black*) and the effects of adding (*blue*) and subtracting (*red*) a suitable multiple of each PC curve for stress (*left*) and lupus (*right*) sample curves

We can see that in both samples the effect of the first principal component of variation is approximately to add or subtract a constant to the degree of stress or lupus throughout the observed period of time. That means that the greatest variability between lupus patients is due to the effect of size. The second PC of stress accounts approximately a 4 % of the total variability and consists of a positive contribution for the first period and a negative contribution for the second one. On the other hand, the second PC curve of lupus is similar but with opposite sign. These second PC curves can be seen as a measure of uniformity of stress and lupus throughout the observed period.

The third step was to select the optimal pairs of lupus/stress PC's to estimate the functional linear regression model. After eliminating noise components we considered the first 8 response PC's that explained a 97.5 % of variability and the first 7 predictor PC's explaining a 97.7 %. The pairs of response/predictor PC's were ordered in based to the proportions $P(*,*)$ of response variance explained by their associated linear regression models. The linear regression model with all these pairs of PC's explains a 58 % of the total variability of the response and the model with only the significant pairs explains a 56.8 %.

The number of PC pairs was selected with BIC and C_P criteria. In both cases the optimal FPCR model includes only the first PC of the response. The prediction equation is given by $\hat{y}(s) = \bar{y}(s) + \hat{\eta}_1 g_1(s) \quad s \in [0, 100]$, where $\hat{\eta}_1$ is the estimation of the first principal component given by

- BIC: $\hat{\eta}_1 = 0.586\xi_1 - 1.769\xi_4$. In this case 2 pairs of response/predictor PC's were selected for estimating a FPCR model that explains a 50 % of the total variability of lupus and a 57% of the variability of its reconstruction with the first PC. The multiple correlation coefficient associated to this model is $R^2 = 0.586$.

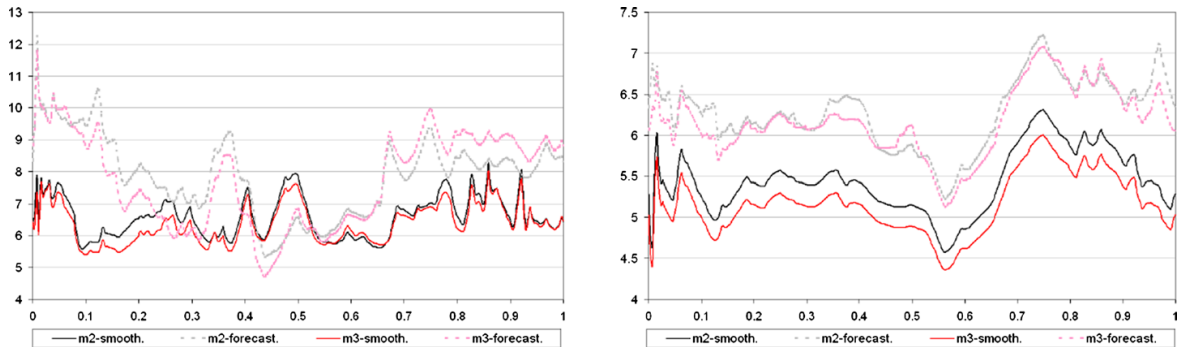


Fig. 7 Prediction error curve with respect to the wavelet approximation of lupus (*left*) and with respect to the approximation of lupus in terms of the first PC (*right*) provided by the BIC model with 2 PC pairs (in *black*) and the C_P model with 3 PC pairs (in *red*) for the training sample (*solid line*) and the test sample (*broken line*)

– C_P: $\hat{\eta}_1 = 0.586\xi_1 - 1.769\xi_4 - 1.575\xi_5$. In this case 3 pairs of response/predictor PC's were selected for estimating a FPCR model that explains a 52 % of the total variability of lupus and a 63.8 % of the variability of its reconstruction with the first PC. $R^2 = 0.626$.

Taking into account the values of the multiple correlation coefficients associated to the fitted model, we can not expect the forecasting ability of the model is very high. Two kinds of prediction errors were computed for stress and lupus sample curves of both, training and test samples. On one hand, Fig. 7 (left) shows the prediction errors with respect to the wavelet approximation of sample curves at resolution level 7. On the other hand, Fig. 7 (right) shows the prediction errors with respect to the representation of the curves in terms of the first PC.

But the main purpose of this application was not predict but to estimate the relationship between lupus and stress. This relation is stablished in terms of the estimated parameters of the linear models. Each increase in the overall level of stress (stress first PC) in an unit produces an increase in the overall level of lupus (lupus first PC) by half an unit. Then, an increase of the stress curve according to the the first PC curve of stress produces an increase of the associated lupus curve according to the half of the first weight curve of lupus.

6 Conclusions

This paper presents a FDA methodology for interpreting the relationship between two functional variables whose sample observations are two set of related curves. A wavelet-based approach for FPCA has been proposed to estimate the functional parameter of the functional lineal model with functional response. This approach is appropriate when sample data are curves with a strong local behavior (high peaks with great variability). This is the case of the stress and lupus curves analyzed in the life sciences application developed in this paper.

Different criteria for selecting the optimal PC's included in the regression models were proposed. The novelty of these criteria is based on including pairs of response/predictor PC's in terms of the percentage of response variance explained by each pair. Methods to select the number of PC pairs based on minimizing cross validation, BIC, C_P and MSE errors were also considered an compared on a simulation study. All criteria provided an accurate estimation of the functional parameter and a good forecasting performance. Considering only those PC pairs with significant correlation provided an estimation of the

functional parameter much more parsimonious with a not excessive cost in error. On the other hand, BIC and C_P criteria are preferred because give similar results to CV and are computationally simpler.

With respect to the application to establish the relation between lupus and stress, we can conclude that stress can explain only a fifty percent of the total variability of lupus. The other fifty percent of variability must be explained by others variables related with lupus illness. On the other hand, for every unit increase in the overall level of stress of a patient its overall level of lupus increases half unit.

Acknowledgments The authors would like to thank two anonymous referees for their interesting comments that helped to improve the quality of the paper. Thanks also to professor Miguel Pérez (Department of Personality, Testing and Psychological Treatment of the University of Granada, Spain) for providing us with the data. This research has been funded by project P11-FQM-8068 from *Consejería de Innovación, Ciencia y Empresa. Junta de Andalucía* and project MTM2013-47929-P from *Ministerio de Economa y Competitividad*, Spain.

References

Aguilera AM, Ocaña FA, Valderrama MJ (1999) Forecasting with unequally spaced data by a functional principal component approach. Test 8(1):233–254

Aguilera AM, Escabias M, Valderrama MJ (2008) Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus. Comput Stat Data Anal 53(1):151–163

Aguilera AM, Escabias M, Preda C, Saporta G (2010) Using basis expansions for estimating functional PLS regression: applications. Chemometr Intell Lab Syst 104(2):289–305

Aguilera AM, Aguilera-Morillo MC (2013a) Comparative study of different B-spline approaches for functional data. Math Comput Model 58(7–8):1568–1579

Aguilera AM, Aguilera-Morillo MC (2013b) Penalized PCA approaches for B-spline expansions of smooth functional data. Appl Math Comput 219(14):7805–7819

Aguilera-Morillo MC, Aguilera AM, Escabias M, Valderrama MJ (2013) Penalized spline approaches for functional logit regression. Test 22(2):251–277

Antoniadis A, Sapatinas T (2003) Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. J Multivar Anal 87(1):133–158

Brantley PJ, Waggoner CD, Jones GN, Rappaport NB (1987) A daily stress inventory: development, reliability, and validity. J Behav Med 10(1):61–74

Cardot H, Ferraty F, Sarda P (1999) Functional linear model. Stat Probab Lett 45:11–22

Chiou JM, Müller H-G, Wang J-L (2004) Functional response models. Stat Sin 14:659–677

Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Commun Pur Appl Math 41:909–996

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

Faraway JJ (1997) Regression analysis for a functional response. Technometrics 39(3):254–262

James G, Hastie T, Sugar C (2000) Principal component models for sparse functional data. Biometrika 87:587–602

Mallat S (1998) A wavelet tour of signal processing. Academic Press, San Diego

Ocaña FA, Aguilera AM, Escabias M (2007) Computational considerations in functional principal component analysis. Comput Stat 22:449–466

Peralta MI, López F, Godoy JF, Godoy D, Sánchez MB, Pérez M (2002) Validación de la detección de cambio del inventario de estrés cotidiano. Psicol Conduct 10:343–354

Peralta-Ramírez MI, Coín-Mejías MA, Jiménez-Alonso J, Ortego-Centeno N, Callejas-Rubio JL, Caracuel-Romero A, Pérez-García M (2006) Lupus 15:858–864

Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edition. Springer-Verlag, New York

Shen Q, Xu H (2007) Diagnostics for linear models with functional response. Technometrics 49(1):26–33

Valderrama MJ, Ocaña FA, Aguilera AM, Ocaña-Peinado FM (2010) Forecasting pollen concentration by a twostep functional model. Biometrics 66:578–585

Yao F, Müller H-G, Wang J-L (2005) Functional linear regression analysis for longitudinal data. Ann Stat 33(6):2873–2903