Mathematics and Computers in Simulation 225 (2024) 66–77

Contents lists available at ScienceDirect

Mathematics and Computers in Simulation

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

Original articles

# Repeated measures in functional logistic regression

Cristhian Leonardo Urbano-Leon *, Ana María Aguilera, Manuel Escabias

*Department of Statistics and Operations Research, and Institute of Mathematics, University of Granada, Granada, Spain*

ARTICLE INFO

*Keywords:*
Functional data
Functional logistic regression
Random effects
Repeated measures

ABSTRACT

We present a proposal to extend the functional logistic regression model – which models a binary scalar response variable from a functional predictor – to the case where the functional observations are not independent because the same functional variable is measured in the same individuals in different experimental conditions (repeated measures). The extension is addressed by including a random effect in the model. The functional approach of this model assumes that all functional objects are elements of the same finite-dimensional subspace of the space of square-integrable functions $L_2$ in the same compact domain allowing the functions to be treated through the basis coefficients on the basis that spans the subspace to which functional objects belong (basis expansion). This methodology usually induces a multicollinearity problem in the multivariate model that emerges, which is solved with the use of the functional principal components of the functional predictor, resulting in a new functional principal component random effects model. The proposal is contextualized through a simulation study that contains three simulation scenarios for four different functional parameters and considering the lack of independence.

## 1. Introduction

Functional data analysis (FDA) is a branch of statistics where the main studied objects are continuous functions, and not only scalar values as in classical statistics. FDA has its beginnings in the works of [28], but its popularity has increased since the works of [14,26,27] as a result of its multiple applications in a number of scientific disciplines and technological advances, allowing for increasingly precise measurements of continuous phenomena.

The developed theory for FDA has been possible thanks to the extension of concepts and methods from scalar statistics to functions, as can be seen in the works of [13,16,17,27]. Accordingly, one of the crucial techniques for scientific research is functional regression analysis, which attempts to find the relationship between a variable called dependent from one or more variables called covariates or explanatory variables. In functional regression it is possible to find different combinations between the types of variables and covariates. This is the case of functional logistic regression, which aims to model a binary random variable from a set of functional observations. This type of model is important in problems where the response can be categorized into two levels, commonly referred to as success and failure (see for example [11] in the case of peak levels of olive pollen). In this context [23] evaluates three approaches for functional logistic regression: dimension reduction using functional principal component analysis, penalized functional regression, and wavelet expansions in combination with Least Absolute Shrinking and Selection Operator penalization. Authors conclude that none of the three methods convince in their ability to reconstruct the parameter function, showing the difficulty of an accurate estimation of the functional parameter in this type of models.

* Corresponding author.
*E-mail addresses:* leonardourbano@correo.ugr.es (C.L. Urbano-Leon), aaguiler@ugr.es (A.M. Aguilera), escabias@ugr.es (M. Escabias).

https://doi.org/10.1016/j.matcom.2024.05.002
Received 15 November 2023; Received in revised form 30 April 2024; Accepted 3 May 2024
Available online 10 May 2024
0378-4754/

There are different approaches to modelling a binary response, but in this work we consider a parametric approach that assumes the existence of a functional parameter. The interpretation of this parameter explains not only the relationship between response and covariates, but also provides an explanation of how changes in parameter generate changes in the odds ratio. Thus, the main objective of this approach is to obtain a correct estimation of the functional parameter that can be carried out by different methods, where the most appropriate for the functional logistic regression problem is the maximum likelihood (ML) method. ML requires observations to be independent, however, when the observations come from the same experimental unit or subject measured repeatedly in different conditions or in different periods of time, the framework is repeated measures. In this context, to assume independence is unrealistic [7] because the phenomenon of repetition can generate a correlation structure in the design matrix of the model [8]. This problem can be addressed by adding a random effect as is done in the scalar case (see [19]), where it is considered that part of the collected variability comes from the correlation structure caused by repeated measurements. Repeated measures in FDA context have a limited development in literature, its study focuses mainly on the curves comparison problem for instance. In [22] authors study the k-sample problem when the data are from the same subjects, proposing a statistic that takes into account the variability between groups. Subsequently, in [31,32] authors consider the variability in each group for a similar problem. Additionally, in [1] a basis expansion approach is used for functional analysis of variance with repeated measures. Most of these methods treat these problems of repeated measures in FDA by adding a random effect into the model, i.e. in functional mixed models (FMM) context. FMM have been developed in literature by some authors, e.g., in [20] a linear mixed effects model is formulated from a non-parametric context. In a same way, in [30] authors address a functional additive mixed models. In [25] Bayesian perspective is used by authors to introduce a functional logistic mixed-effects model for estimating learning curves in longitudinal experiments. In the last three cases, the random effect considered is functional. Alternatively, scalar random effects are included in [21] for a functional linear mixed model in the context of scalar on function regression – functional predictor and scalar response–.

On the other hand, as a consequence of the algebraic structure of some functional spaces, it is possible to consider the functional data as elements of a finite-dimensional subspace of square integrable functions space $L_2[a,b]$. This consideration allows the use of all vector space properties, as the representation of any element in terms of a basis of fixed functions. This representation produces, for each functional datum, a unique vector of scalars that are the coefficients of the linear combination of the elements of the basis that span the subspace to which the functional data belong. This treatment of functional data allows models to be reduced to a multivariate scalar problem, as already done in works such as those seen in [1,5,10,33], among other examples. However, the treatment of functional data through their basis coefficients within the context of regression can generate a problem known as multicollinearity i.e. correlation among the explanatory variables of a model, leading to high standard errors and another problems in the model parameters estimation [15]. This concern in the context of the functional logistic regression was worked on by [9,10], where functional principal components logistic regression (FPCLR) was introduced, and an extended FPCLR model and R-package were developed. FPCLR model provides the advantage that the new vectors of coefficients no longer present the problem of multicollinearity, since no correlation is theoretically guaranteed. Additionally, FPCLR model allows the reduction of the dimensionality of the problem, through the choice of a reduced number of functional principal components.

Functional logistic models for repeated measures on basis coefficients have problems of correlation attributable to repetition, and multicollinearity caused by the same basis representation for predictor and functional parameter (see [10]). The approach proposed here consists of the combination of two methodologies to address these issues: the random effect inclusion in the model to capture some of the variability attributable to repetition of functional observations, and the use of the functional principal components to deal with the possible multicollinearity problem that may exist in the model. As far as we know, the case of repeated measures for functional logistic regression model has not been considered in literature, much less the effect of multicollinearity in the model estimation.

This paper is divided into 4 sections. The introductory section shows some background in the context of functional data analysis, repeated measures and functional mixed models. Section 2 presents the theoretical framework on functional data and functional logistic regression model for repeated measures. Section 3 develops a simulation study with three different scenarios on four different functional parameters. Finally, Section 4 contains a summary and discussion of the main results and conclusion obtained.

## 2. Methodology

### 2.1. Functional data

There are different approaches to extend concepts from scalar statistics to continuous functions. One of these approaches considers that a functional datum $\mathcal{X}$ is an observation of a second order stochastic process $\{\mathcal{X}(t) : t \in [a,b]\}$, i.e. it satisfies the property of Eq. (1)

$$\int_a^b \mathcal{X}^2(t)dt < \infty. \tag{1}$$

We assume that $\mathcal{X} \in \mathcal{H} : \mathcal{H} \subset L_2[a,b] \wedge dim(\mathcal{H}) = d \in \mathbb{N}$, where $\mathcal{H}$ and $L_2[a,b]$ are vector spaces over the field $\mathbb{R}$, whose elements are square integrable functions on the same domain $[a,b]$, and which has a Hilbert space structure (see [18,29]) with inner product defined as in Eq. (2):

$$\langle f,g\rangle = \int_a^b f(t)g(t)dt,\ \forall f,g \in L_2[a,b]. \tag{2}$$

This inner product induces the usual norm and distance $m$ defined by $\|f\| = \sqrt{\langle f, f\rangle}$ and $m(f,g) = \|f-g\|$ respectively. Here, $d$ represents the dimension of the subspace $\mathcal{H}$. These assumptions allow the use of functional finite basis concept of $\mathcal{H}$. Consequently, given $\{\mathcal{X}_i\}_{i=1}^n$ a set of $n$ functional data, then $\forall \mathcal{X}_i \in \mathcal{H}$, $\exists \left(a_{ij}\right)_{j=1}^d \in \mathbb{R}^d : \mathcal{X}_i(t) = \sum_{j=1}^d a_{ij}\phi_j(t)$, where the set $\Phi = \{\phi_j\}_{j=1}^d \subset \mathcal{H}$ is a basis for $\mathcal{H}$, and the $a_{ij} \in \mathbb{R}$ are called basis coefficients or coefficients of representation of the $i-th$ functional datum in basis $\Phi$.

The vector of basis coefficients for each function is unique, then it is possible to establish an isomorphism between the spaces $\mathcal{H}$ and $\mathbb{R}^d$. As a consequence, the use of a vector of basis coefficients $(a_{ij})_{j=1}^d \in \mathbb{R}^d$ instead of function $\mathcal{X}_i$ is conceptually coherent and simplifies the data treatment and the simulation process, reducing some functional problems to the multivariate scope, as can be seen for example in [2] for detecting changes in air pollution during the COVID-19 pandemic by using functional ANOVA.

It is important to note that different proposals exist for functional basis in order to obtain the representation in basis coefficients such as Fourier, B spline (see [27]), CONS basis (see [12]) or wavelets (see [4]).

### 2.2. Functional logistic model for repeated measurements

Let $L_2[a,b]$ be the vector space over $\mathbb{R}$, with Hilbert space structure, of square integrable functions defined on the interval $[a,b]$, and $\mathcal{H} \subset L_2[a,b]$ a subspace such that $dim(\mathcal{H}) = d \in \mathbb{N}$. Suppose $N$ individuals measured in different experimental conditions for the same continuous domain, where the curve $\mathcal{X}_{is}$ is given by the $s-th$ functional repetition for the $i-th$ individual. The set of functional observations $\bigcup_{i=1}^N \{\mathcal{X}_{is}\}_{s=1}^{n_i} \subset \mathcal{H}$, being $card(\bigcup_{i=1}^N \{\mathcal{X}_{is}\}_{s=1}^{n_i}) = \sum_{i=1}^N n_i = n$ the total number of functional observations. Furthermore, let us suppose that $\{y_{is}\}_{i=1,s=1}^{N,n_i}$ is a set of binary responses that represent the success or failure of a phenomenon related to the functional observations, and which are defined as in the Eq. (3)

$$y_{is} = \begin{cases} 1 & if \quad success \\ 0 & if \quad failure. \end{cases} \tag{3}$$

Each $y_{is}$, $i = 1,2,\dots,N$; $s = 1,2,\dots\dots,n_i$ is an observation of a random variable $Y$, such that $Y|\mathcal{X}_{is} \sim Be(\pi_{is})$, where $Be(\pi_{is})$ represents a Bernoulli probability distribution with parameter $\pi_{is} = P(Y = 1|\mathcal{X}_{is})$. The issue of repeated measures has been addressed in literature mainly through mixed models. These models add different random effects to the models themselves (see [8]). With this in mind, we propose the mixed functional logistic model for repeated measurements treatment and that is represented by Eq. (4)

$$l_{is} = ln\left[\frac{\pi_{is}}{1-\pi_{is}}\right] = \alpha + \int_a^b \mathcal{X}_{is}(t)\beta(t)dt + z_{is}u_i, \; i = 1,2,\dots,N; s = 1,2,\dots,n_i, \tag{4}$$

where $l_{is}$ is the logarithm of the odds of success over failure, $E\left[Y|\mathcal{X}_{is}\right] = \pi_{is}$, $\int_a^b \mathcal{X}_{is}(t)\beta(t)dt$ is a fixed effect, $u_i$ is the vector of random effects and $z_{is}$ is a repetition indicator vector. This is the classical formulation of the mixed logit model, seen as a Generalized Linear Model (GLM) with logit transformation as link function (see [3,8] for scalar case).

The $\beta \in \mathcal{H}$ is the functional parameter whose accurate estimation is the objective of the methods proposed here. As can be seen in [10] for the functional logistic regression model, this functional parameter allows an interpretation of the relationship between the binary response and the functional predictor in terms of odds ratio. Then, since $\mathcal{X}_{is}, \beta \in \mathcal{H}; i = 1,2,\dots,N, s = 1,2,\dots,n_i$, there are vectors $(b_j)_{j=1}^d$, and $(a_{is,j})_{j=1}^d$, such that

$$\mathcal{X}_{is} = \sum_{j=1}^d a_{is,j}\phi_j \qquad \wedge \qquad \beta = \sum_{j=1}^d b_j\phi_j. \tag{5}$$

Then in Eq. (4) it follows

$$\begin{aligned} \int_a^b \mathcal{X}_{is}(t)\beta(t)dt &= \int_a^b \left[\sum_{j=1}^d a_{isj}\phi_j(t)\right]\left[\sum_{j=1}^d b_j\phi_j(t)\right]dt \\ &= \left[\sum_{j=1}^d a_{isj}b_j \left\|\phi_j\right\|^2\right] + \left[\sum_{j=1,k\neq j}^d a_{isj}b_k \left\langle\phi_j,\phi_k\right\rangle\right]. \end{aligned} \tag{6}$$

Thus, the functional logit model for repeated measures can be written in matrix form as

$$L = \mathbf{1}\alpha + A\Psi\mathcal{B} + ZU, \tag{7}$$

with $\mathbf{1}$ being a vector of ones, $L$ the vector of the $n$ logit transformations $l_{is}$, $A$ the matrix of basis coefficients of curves, $\Psi$ the matrix of inner products of the elements of the basis $\Phi$, and $\mathcal{B}$ the parameter vector to be estimate that coincides with the vector of basis coefficients of functional parameter $\beta$. $U$ is the random effects vector and $Z$ the design matrix associated to $U$ that contains the repetition framework.

In this way the functional logistic regression model for repeated measures is transformed into a multivariate logistic model (for repeated measures). Assuming a spherical Gaussian distribution for the random effects, the estimation of the basis coefficients $b_j, j = 1,2,\dots,d$ of the functional parameter $\beta$ can be obtained by classic methods for repeated measures as restricted maximum likelihood (REML), and penalized iteratively re-weighted least squares (PIRLS) (see [6]). However, this formulation poses drawbacks caused by possible multicollinearity in the new explanatory variables, since it is not possible to ensure that the basis coefficients of curves are independent by columns. To deal with the multicollinearity issue, it is possible to restate the problem in terms of the functional principal components (FPCs), see [10]. Let us thus consider the functional principal components of sample curves $\{\mathcal{X}_{is}\}_{i=1,s=1}^{N,n_i}$ as

$$\xi_{is,w} = \int_a^b \left(\mathcal{X}_{is}(t) - \overline{\mathcal{X}}(t)\right) f_w(t)dt, \tag{8}$$

where the functions $f_w \in \mathcal{H}$, $(w = 1, 2, \ldots, d)$ are the solutions to the *eigenequation*

$$\int_a^b C(r,t) f_w(r) ds = \lambda f_w(t), \tag{9}$$

with $C$ being the functional sample covariance defined by

$$C(r,t) = n^{-1} \sum_{\imath=1}^{n} \left(\mathcal{X}_\imath(r) - \overline{\mathcal{X}}(r)\right)\left(\mathcal{X}_\imath(t) - \overline{\mathcal{X}}(t)\right). \tag{10}$$

The vectors $\xi_w$ and $\xi_j$ are independent for all $w \neq j$, and it is well known that each functional datum $\mathcal{X}_{is}$ can be approximated in terms of a reduced set of $p$ eigenfunctions $f_w$ and $\xi_{is,w}$ by

$$\mathcal{X}_{is} \approx \overline{\mathcal{X}} + \sum_{w=1}^{p} \xi_{is,w} f_w. \tag{11}$$

Then, from Eq. (4) we can obtain the functional principal components logit model for repeated measures (in terms of logit transformation) by

$$l_{is} = \alpha + \int_a^b \left(\overline{\mathcal{X}}(t) + \sum_{w=1}^{p} \xi_{isw} f_w(t)\right) \beta(t)dt + z_{is}u_i, \ i = 1, 2, \ldots, N; s = 1, 2, \ldots, n_i. \tag{12}$$

The model in the Eq. (12) can be expressed in matrix form as

$$L = \mathbf{1}\gamma_0 + \Gamma\gamma + ZU, \tag{13}$$

where $\Gamma$ is the matrix of the functional principal components $(\xi_{is,w})$, $\gamma_0 = \alpha + \int_a^b \overline{\mathcal{X}}(t)\beta(t)dt$ and $\gamma$ the vector of parameters with elements $\gamma_w = \int_a^b f_w(t)\beta(t)dt$. The model in Eq. (13) avoids the problem of multicollinearity since the principal components are incorrelated, so it is possible to use all the usual methods to obtain an estimate $\hat{\gamma}$ of the parameter vector $\gamma$, and through this an estimate of the original parameter vector $\mathcal{B}$, through $\hat{\mathcal{B}} = V\hat{\gamma}$, with $V$ being the matrix of the basis coefficients of eigenfunctions $f_w$ in $\mathcal{H}$.

## 3. Simulation

In order to evaluate the performance of the proposed methods we have developed a simulation study, considering three different scenarios:

- Scenario 1: Functional logit model without repeated measures.
- Scenario 2: Functional logit model with repeated measures.
- Scenario 3: Functional logit model with repeated measures, and multicollinearity

For all scenarios, four functional parameters $\beta$ in $\mathcal{H}$ (subspace spanned by finite basis in Eq. (14)) were generated from expression $s\left(sin(w_1 \cdot t)\right)\left(cos(w_2 \cdot t)\right)$, where $s$, $w_1$, and $w_2$ are scalar values that, when modified, generate changes in the scale, oscillation and roughness of the $\beta$ function. The 4 types of functional parameters considered can be seen in Fig. 1.

### 3.1. Scenario 1

The functional curves considered in this scenario belong to subspace $\mathcal{H}$ spanned by finite basis $\Phi$ with $d = 8$. The elements of the basis come from a complete orthonormal sequence (CONS), which provides multiple operational advantages thanks to its orthonormality, see [12,24]. These elements are described in Eq. (14), and shown in the left panel of Fig. 2.

$$\phi_j(x) = \begin{cases} 1 & if \quad j = 1 \\ \sqrt{2}\cos((j-1)\pi x) & if \quad 2 \le j \le d. \end{cases} \tag{14}$$

So, $n = 750$ curves $\{\mathcal{X}_\imath\}_{\imath=1}^{n}$ were considered by simulating their basis coefficients with uniform values, i.e. $(a_{\imath,j})_{j=1}^{d} = A_i \sim Unif[0.5, 3]$. A sample of 100 simulated curves can be seen in Fig. 2 (Right).

After simulating the predictor curves, we calculate the linear predictor $l_\imath$ given by Eq. (6) using the functional parameters $\beta_1$, $\beta_2$, $\beta_3$ and $\beta_4$. After adding an error term $E \sim N(0, Id_{n\times n})$ to the linear predictor the response was simulated by using a Bernoulli distribution with probabilities given by $\exp(l_\imath)/(1 + \exp(l_\imath))$. Four different models were then fitted, and the $\beta$ functional parameter estimated:

- Model 1: $L = \mathbf{1}\alpha + A\Psi\mathcal{B}$, i.e. the proposed model in Eq. (4) without random effects – called Classic Model (CL_Model) –. The estimates were obtained by $ML$.

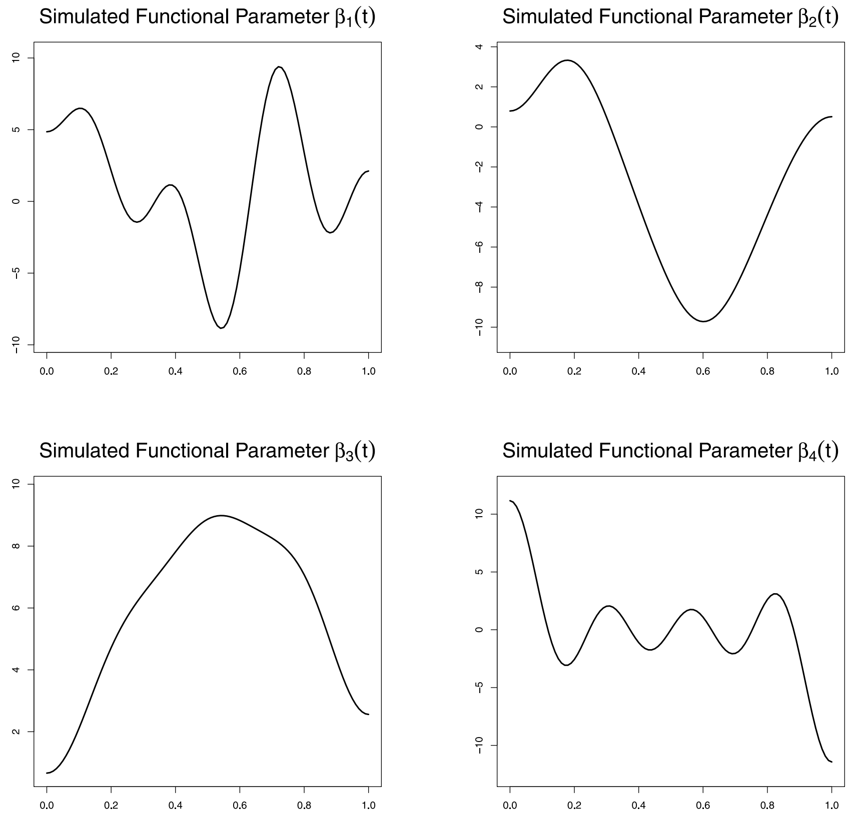


**Fig. 1.** Top left: Functional parameter $\beta_1(t)$ generated with the values $s = 10$, $w_1 = 15$ and $w_2 = 5$. Top right: Functional parameter $\beta_2(t)$ generated with the values $s = 10$, $w_1 = 3$ and $w_2 = 5$. Bottom left: Functional parameter $\beta_3(t)$ generated with the values $s = 80$, $w_1 = 0.3$, $w_2 = 1.5$. Bottom right: Functional parameter $\beta_4(t)$ generated with the values $s = 70$, $w_1 = 25$, $w_2 = 25$.

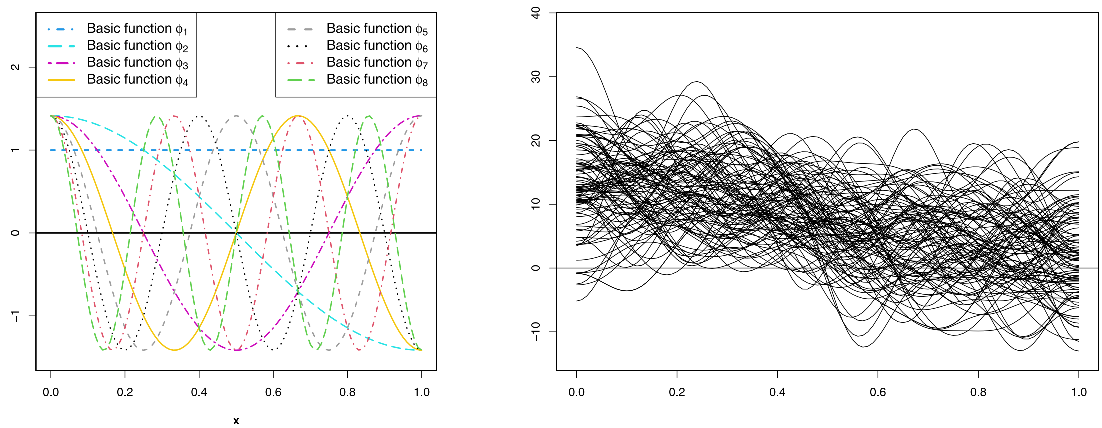


**Fig. 2.** Left: 8 functions of the basis $\Phi$ of subspace $\mathcal{H}$. Right: a sample of 100 functional data simulated as elements of subspace $H$.

- Model 2: $L = \mathbf{1}\alpha + A\Psi B + ZU$, i.e. the proposed model in Eq. (4) with random effects – called Repeated Measures Model (RM_Model) –. The estimates were obtained by $REML$.
- Model 3: $L = \mathbf{1}\gamma_0 + \Gamma\gamma$ i.e. the proposed model in Eq. (12) model without random effects – called Classic Model on the Principal Components (PC_Model) –. The estimates were obtained by $ML$.
- Model 4: $L = \mathbf{1}\gamma_0 + \Gamma\gamma + ZU$ i.e. the proposed model in Eq. (12) with random effects – called Repeated Measurements Model on the Principal Components (RMPC_Model) –. The estimates were obtained by $REML$.

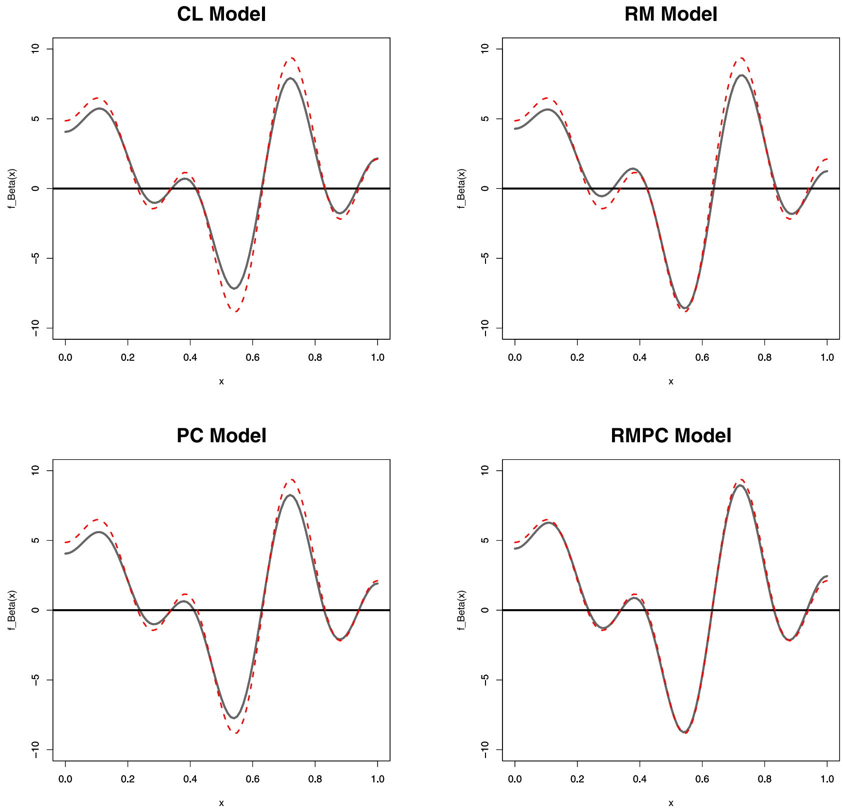


**Fig. 3.** For all figures: In red dashed line, the target functional parameter $\beta_1$, in black solid lines, the functional estimations $\hat{\beta}_1$. On the top left for the first model $CL\_Model$. On top right for the second model $RM\_Model$. Bottom left, for the third model $PC\_Model$. Bottom right, for the fourth model $RMPC\_Model$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In $PC\_Model$ and $PCRM\_Model$, the number of functional principal components used was fixed as to explain 99% of the total variability. Fig. 3 shows an example of the simulated functional parameter $\beta_1$ and estimated $\hat{\beta}_1$ with four models for Scenario 1.

The performance of the logit models was carried out by the correct classification rate ($CCR$), an indicator of the percentage of items whose prediction perfectly matches the original observation. The $CCR$ is a good indicator of model accuracy and can be calculated using the Eq. (15)

$$CCR = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i), \tag{15}$$

where $y_i$ is the binary observation for the $i-th$ individual, and $\hat{y}_i$ is the prediction made for the same individual. Thus, the proportion of model successes is obtained with all the predictions of the model, moreover, the accuracy of the estimations of the functional parameters was tested by the integrated squared error $ISE$, defined by Eq. (16)

$$ISE = \int_a^b (\beta(t) - \hat{\beta}(t))^2 dt, \tag{16}$$

This process was carried out 100 times, wherein we obtained 100 functional parameter estimations $\left\{\hat{\beta}_i\right\}_{i=1}^{100}$, $CCR$ and $ISE$ for each model, The evaluation of the 100 simulations was tested by the averages of $CCR$ and $ISE$ – referred to as $MCCR$ and $MISE$ respectively – in addition to the standard deviation of the $CCR$ ($SDCCR$). Furthermore, scalar variance for functional data ($SVFD$) developed by [33] was used to provide a scalar value of the variability of a set of curves within a single finite-dimensional subspace. This is useful for comparing the consistency of the estimates from the four models. The scalar variance for functional data is defined in Eq. (17)

$$SVFD = \frac{1}{n-1}\sum_{i=1}^{n}\int_a^b (\hat{\beta}_i(t) - \overline{\hat{\beta}}(t))^2 dt \tag{17}$$

where $\hat{\beta}_i(t)$ is the $i-th$ estimation of functional parameter $\beta_i(t)$, and $\overline{\hat{\beta}}(t)$ is the functional mean of the estimations. One of the advantages of $SVFD$ is that it can be calculated directly from the basis coefficients, offering operational advantages, even more so

**Table 1**
Accuracy measures in Scenario 1 for four target $\beta$ functions.

| Accuracy measure | Functional parameter $\beta_1$ | | | | Functional parameter $\beta_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | CL | RM | PC | RMPC | CL | RM | PC | RMPC |
| $SVFD$ | 0.24 | 0.27 | 0.24 | 0.27 | 0.29 | 0.31 | 0.29 | 0.31 |
| $ISB$ | 0.23 | 0.16 | 0.23 | 0.16 | 0.21 | 0.17 | 0.21 | 0.17 |
| $MISE$ | 0.47 | 0.43 | 0.47 | 0.43 | 0.50 | 0.48 | 0.50 | 0.48 |
| $MCCR$ | 0.90 | 0.90 | 0.90 | 0.90 | 0.91 | 0.91 | 0.91 | 0.91 |
| $SDCCR$ | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 |
| | Functional parameter $\beta_3$ | | | | Functional parameter $\beta_4$ | | | |
| | CL | RM | PC | RMPC | CL | RM | PC | RMPC |
| $SVFD$ | 0.47 | 0.58 | 0.47 | 0.58 | 0.19 | 0.20 | 0.19 | 0.20 |
| $ISB$ | 0.43 | 0.28 | 0.43 | 0.28 | 0.20 | 0.17 | 0.20 | 0.17 |
| $MISE$ | 0.90 | 0.85 | 0.90 | 0.85 | 0.39 | 0.37 | 0.39 | 0.37 |
| $MCCR$ | 0.92 | 0.93 | 0.92 | 0.93 | 0.89 | 0.89 | 0.89 | 0.89 |
| $SDCCR$ | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |

when the basis is orthonormal. Accordingly, the scalar variance for functional data in the subspace spanned by $\Phi$ orthonormal basis can be calculated as in Eq. (18)

$$SVFD = \sum_{j=1}^{d} V_j \tag{18}$$

where $V_j$ is the variance of the $j-th$ vector of basis coefficients from estimations of $\beta(t)$. The bias in the functional estimations was calculated using the integrated squared bias $ISB$ according to Eq. (19)

$$ISB = \int_a^b (\beta(t) - \overline{\hat{\beta}}(t))^2 dt, \tag{19}$$

where $\beta(t)$ is the simulated functional parameter, and $\overline{\hat{\beta}}$ is the mean of the functional estimations of $\beta(t)$. The $ISB$ provides a general scalar measure of the bias in the estimates.

Fig. 4 shows the results of the 100 estimations of simulated functional parameters, for the four models in simulation Scenario 1. Here, there is no multicollinearity, and there is no correlation structure due to repeated measurements of the same individual. As expected in this case, the four models produce similar estimates, which can be verified through accuracy measures in the top left Table 1. The graphic results are consistent in all functional parameters.

In order to compare the accuracy of the fits for the four models in simulation Scenario 1, Table 1 shows the accuracy measures for the four simulated parameters considered. It is possible to note that the results in fact show stability even with changes in the form of the functional parameter. In all cases $CCR$ are high, hanging around 90%, and there are almost negligible differences in $SVFD$, $ISB$ and $MISE$ among models and functional parameters.

### 3.2. Scenario 2

In this scenario, the predictor curves were simulated by assuming the same subspace $\mathcal{H}$ spanned by the same finite basis $\Phi$ as Scenario 1. In this case, we assumed $N = 50$ individuals and $n_i = 15$ repetitions for $i = 1, 2 \ldots, N$, $n = 750$ curves in total. For the repeated measures the random effects were simulated by using a Gaussian distribution, i.e. $U \sim N(0, 3.5)$. The covariance matrix was generated without multicollinearity but the responses had random effect because of repeated measurements. The response was also simulated in the same terms as Scenario 1. As in Scenario 1, the process was replicated 100 times, the fits and the accuracy were tested by using the same measures and techniques as in that scenario.

Fig. 5 shows the results of the 100 estimates of the functional parameters for the four models in simulation Scenario 2. Here, no multicollinearity between the columns of the design matrix was considered, but there existed a correlation structure caused by repetition and that was added through the random effects simulation. Here, it is possible to observe that the models $CL\_Model$ and $PC\_Model$ – which use ML estimation – show a bias in the functional mean of the estimates, while in the models $RM\_Model$ and $RMPC\_Model$ – which use REML estimation – the functional means of the estimates are closer. You can also observe how including functional principal components in the models, in this scenario, has no effect on the accuracy and bias of the functional parameter estimates compared to not including them. On the other hand, when comparing the results obtained for the different functional parameters, one might suspect that the shape of them could be influencing the accuracy and bias of the estimates. For example, it can be observed that in functional parameter like $\beta_2$ and $\beta_3$, the discrepancies from not using random effects in the models are greater than in $\beta_1$ and $\beta_4$. In any case, the inclusion of a random effect in the model always improves the estimates.

In Table 2 (top left), as expected, the prediction ability of the four models is very accurate with similar and high $CCR$ in all models. However, it is possible to observe an increase in bias and error in the estimates of the $CL\_Model$ and $PC\_Model$ models with respect to scenario 1, since the increase in the bias of the estimates is just a consequence of repeated measures. Despite this, the decreases from 2.65 to 0.26 in $ISB$ and from 2.81 to 0.55 in $MISE$ show the importance of including the random effect in the

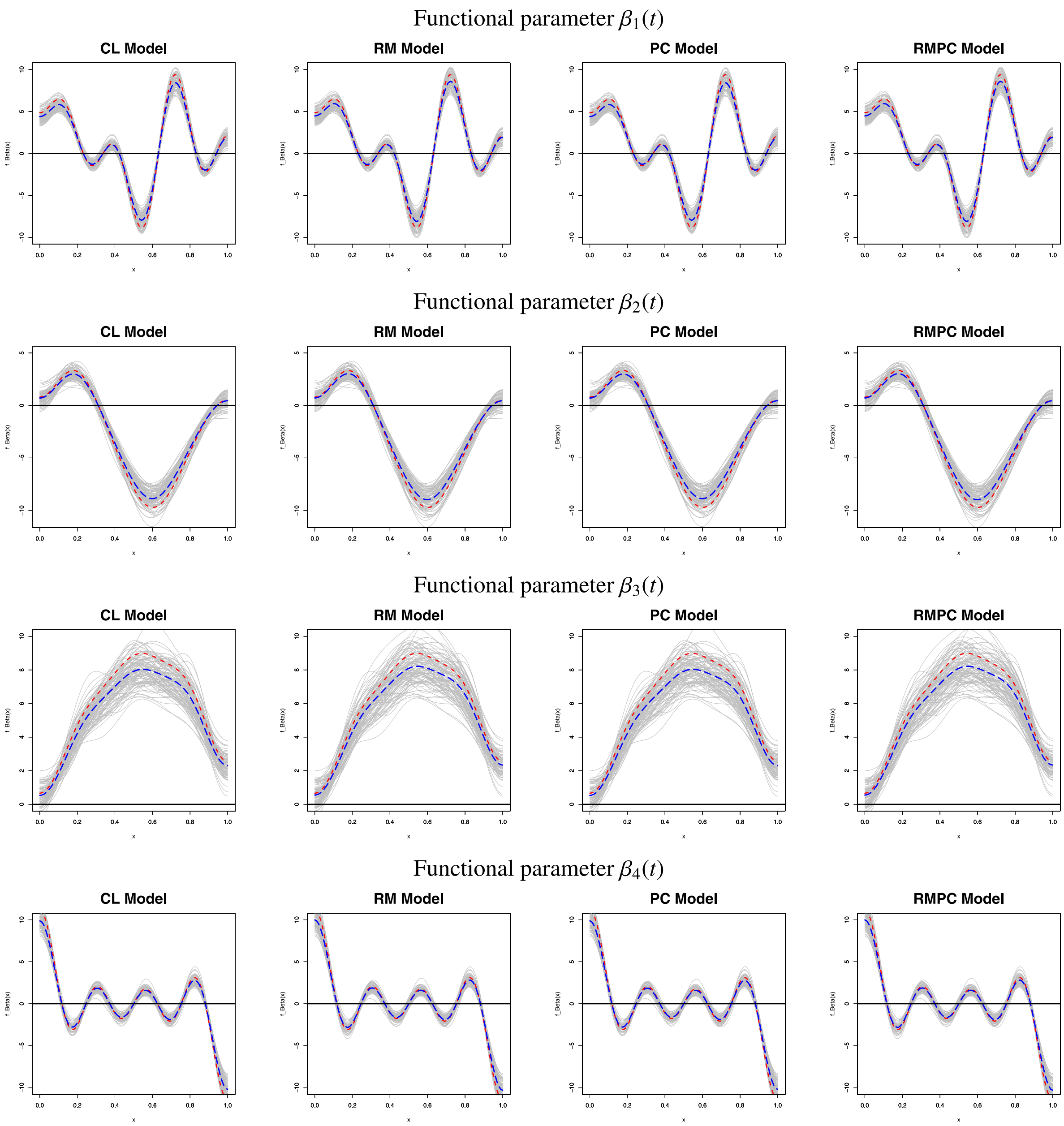


**Fig. 4.** Scenario 1: For all figures, in red dashed line the target functional parameter, in grey solid lines, the 100 functional estimations, in blue long dashed line, the functional mean of the 100 estimations. Each line shows for each functional parameter the results of fitted models $CL\_Model$, $RM\_Model$, $PC\_Model$ and $RMPC\_Model$ respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

logit model for repeated measures to improve the estimation of the functional parameter, and for obtaining a precise interpretation of the functional parameter in terms of odds ratios. As in Scenario 1, the results show stability decreasing $ISB$ and $MISE$ in models with random effect, even with changes in the form of the functional parameter given by $\beta_2$, $\beta_3$ and $\beta_4$. In terms of the $SVFD$, as in scenario 1, an increase is observed in the models with random effect, which is because the REML method can increase the variance by reducing the bias in the estimates. This can be seen in the tested functional parameters $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$, although the excessive increase in the $SVFD$ in $\beta_3$ for the $RM\_Model$ and $RMPC\_Model$ may be an indication that the it is influenced by the shape of the functional parameter.

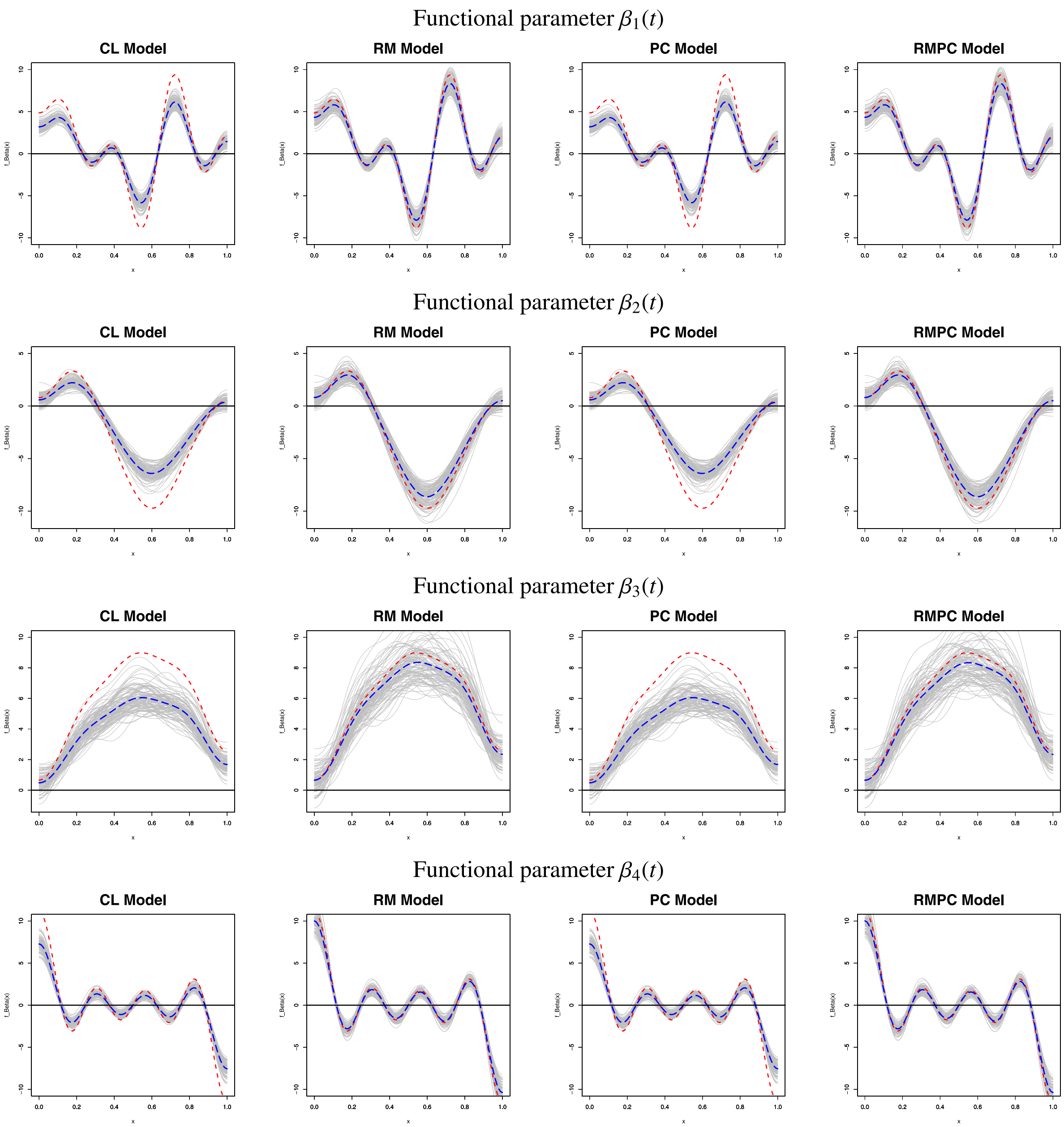


**Fig. 5.** Scenario 2: For all figures, in red dashed line the target functional parameter, in grey solid lines, the 100 functional estimations, in blue long dashed line, the functional mean of the 100 estimations. Each line shows for each functional parameter the results of fitted models $CL\_Model$, $RM\_Model$, $PC\_Model$ and $RMPC\_Model$ respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### 3.3. Scenario 3

In this scenario we deal with a more realistic case, where multicollinearity exists due to basis expansion representation of the functional objects of our models. Thus, $N = 50$ individuals and $n_i = 15$ repetitions for $i = 1, 2 \ldots, N$, $n = 750$ curves in total were now simulated with repeated measures and multicollinearity. In this scenario, a Normal distribution instead of Uniform was used for basis coefficients simulation of the functional predictors, i.e. $(a_{is,j})_{j=1}^{d} = A_j \sim N(0, \Sigma), i = 1, 2, \ldots, N$, where covariance matrix $\Sigma$ was generated with multicollinearity. The response simulation, replication, fits, and accuracy evaluation were carried out as in the two previous scenarios.

Fig. 6 shows the results of the estimations of the functional parameters for simulation of this Scenario 3, which considers multicollinearity and correlation structure because of repetition. Here it can be seen that the four models have difficulty estimating

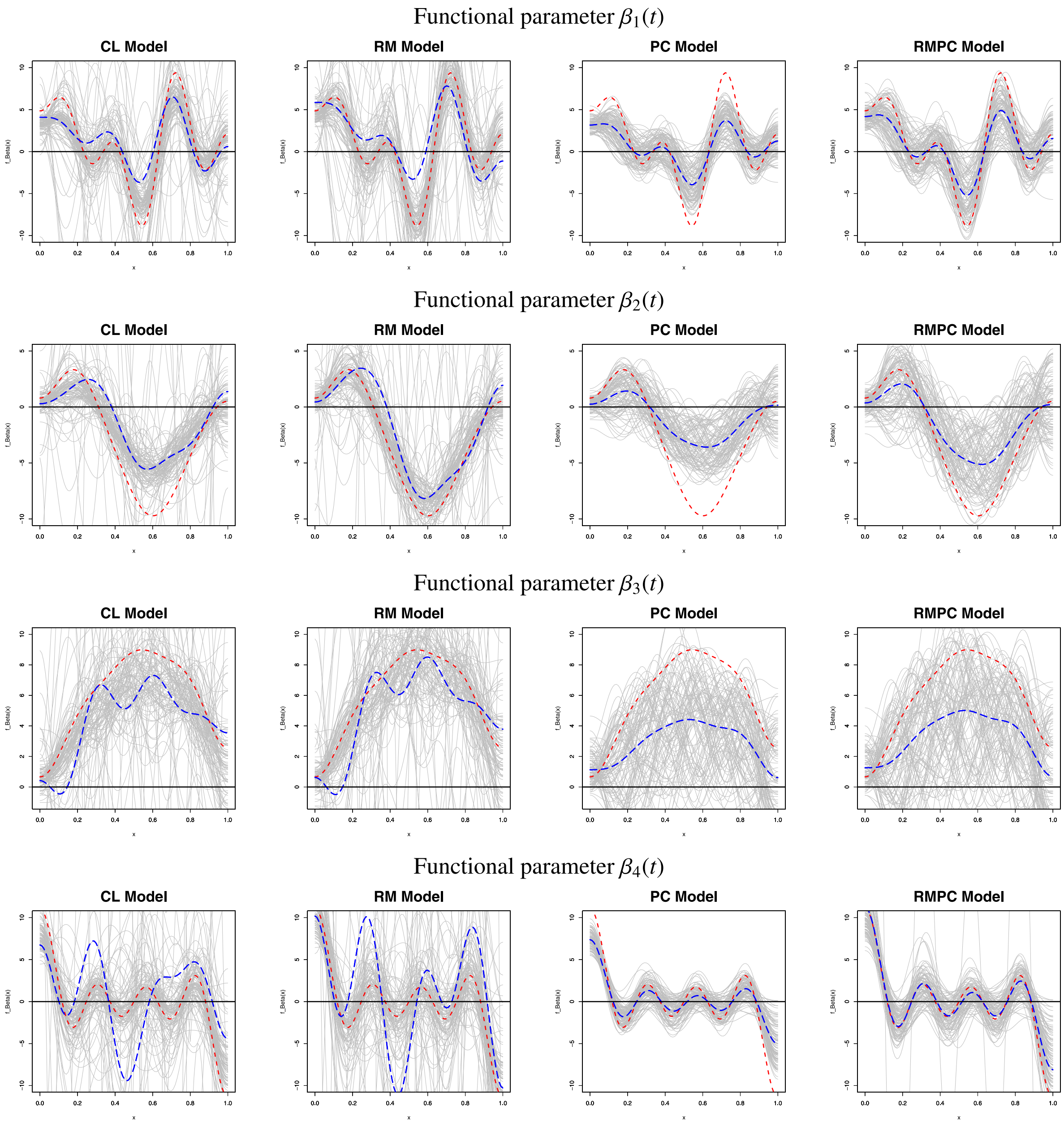


**Fig. 6.** Scenario 3: For all figures, in red dashed line the target functional parameter, in grey solid lines, the 100 functional estimations, in blue long dashed line, the functional mean of the 100 estimations. Each line shows for each functional parameter the results of fitted models $CL\_Model$, $RM\_Model$, $PC\_Model$ and $RMPC\_Model$ respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Accuracy measures in Scenario 2 for four target $\beta$ functions.

| Accuracy measure | Functional parameter $\beta_1$ | | | | Functional parameter $\beta_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | CL | RM | PC | RMPC | CL | RM | PC | RMPC |
| $SVFD$ | 0.16 | 0.29 | 0.16 | 0.29 | 0.19 | 0.34 | 0.19 | 0.34 |
| $ISB$ | **2.65** | **0.26** | 2.65 | 0.27 | 3.09 | 0.32 | 3.09 | 0.32 |
| $MISE$ | **2.81** | **0.55** | 2.81 | 0.55 | 3.28 | 0.66 | 3.28 | 0.66 |
| $MCCR$ | 0.87 | 0.92 | 0.87 | 0.92 | 0.88 | 0.93 | 0.88 | 0.93 |
| $SDCCR$ | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 |
| | Functional parameter $\beta_3$ | | | | Functional parameter $\beta_4$ | | | |
| | CL | RM | PC | RMPC | CL | RM | PC | RMPC |
| $SVFD$ | 0.35 | 0.92 | 0.35 | 0.85 | 0.12 | 0.21 | 0.12 | 0.21 |
| $ISB$ | 4.50 | 0.18 | 4.50 | 0.20 | 1.89 | 0.15 | 1.89 | 0.15 |
| $MISE$ | 4.84 | 1.09 | 4.84 | 1.04 | 2.01 | 0.36 | 2.01 | 0.36 |
| $MCCR$ | 0.90 | 0.94 | 0.90 | 0.94 | 0.85 | 0.90 | 0.85 | 0.90 |
| $SDCCR$ | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.01 |

**Table 3**
Accuracy measures in Scenario 3 for four target $\beta$ functions.

| Accuracy measure | Functional parameter $\beta_1$ | | | | Functional parameter $\beta_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | CL | RM | PC | RMPC | CL | RM | PC | RMPC |
| $SVFD$ | 334.90 | 676.02 | 2.13 | 3.83 | 41.20 | 82.05 | 2.21 | 4.56 |
| $ISB$ | 6.30 | 7.19 | 7.08 | 4.00 | 6.04 | 2.35 | 10.42 | 5.58 |
| $MISE$ | 337.82 | 676.38 | 9.19 | 7.80 | 46.82 | 83.58 | 12.61 | 10.10 |
| $MCCR$ | 0.90 | 0.94 | 0.89 | 0.93 | 0.84 | 0.91 | 0.83 | 0.90 |
| $SDCCR$ | 0.03 | 0.02 | 0.03 | 0.02 | 0.04 | 0.02 | 0.04 | 0.03 |
| | Functional parameter $\beta_3$ | | | | Functional parameter $\beta_4$ | | | |
| | CL | RM | PC | RMPC | CL | RM | PC | RMPC |
| $SVFD$ | 147.34 | 165.11 | 5.75 | 7.50 | 1853.57 | 2692.48 | 1.24 | 19.52 |
| $ISB$ | 4.18 | 2.48 | 11.51 | 8.84 | 20.05 | 22.81 | 3.56 | 0.78 |
| $MISE$ | 150.02 | 165.93 | 17.20 | 16.26 | 1854.89 | 2688.09 | 4.79 | 20.10 |
| $MCCR$ | 0.91 | 0.93 | 0.90 | 0.92 | 0.95 | 0.97 | 0.94 | 0.97 |
| $SDCCR$ | 0.03 | 0.02 | 0.04 | 0.03 | 0.01 | 0.01 | 0.01 | 0.01 |

the target functional parameter, with the cases in the $CL\_Model$ and $RM\_Model$, being notable since multicollinearity produces bad estimations with dramatic differences. Although the principal components models ($PC\_Model$ and $RMPC\_Model$) improve by decreasing the bias, error and variance of estimates, and produce stable results if compared to the other two models, this is not enough for $\beta_2$ and $\beta_3$, where no methods are able to provide suitable estimates. As in scenario 2, in this scenario, there is suspicion that the shape of the parameter function may influence the accuracy of the estimates. However, for all parameter functions considered, indicate that only the inclusion of a random effect does not improve the estimates in the presence of multicollinearity, so the use of functional principal components is necessary for a more precise estimation. On the other hand, the use of functional principal components alone does not improve the estimates in the case of repeated measures. It is the combination of both methodologies that produces a significant gain in the estimates. These conclusions can also be checked in Table 3.

## 4. Conclusions

In this work we propose functional principal components logistic regression for modelling a binary response variable from a functional covariate, when the observations are of repeated functional type. It is important to note here that the fundamental contribution sought is the appropriated estimation of the functional parameter because, as Section 1 indicates, the goal of the functional logistic model – from a parametric perspective – is the interpretation of the functional parameter, which will be realistic as long as the estimates recover the shape of the target functional parameter. From this point of view, our conclusions about the model from simulation results are the following:

- The inclusion of a random effect in the functional logistic model is effective for improving the estimation of the functional parameter in the case of functional repeated measures. As can be seen in all scenarios the inclusion of the random effect significantly improves the prediction of the response as well as the estimation of the functional parameter and, therefore, improves the interpretation.
- The use of functional principal components allows the estimation of the functional parameter to be improved, even in the random effects model in the presence of multicollinearity.
- Although the results in Section 3 show some regularity in model performance when the functional parameter is changed, the shape of the parameter could be influencing the estimates. This is more evident in Scenario 3, where it is shown that in the

presence of multicollinearity and repeated measures, no model produce suitable estimates for parameter functions with low variability along their trajectory, like $\beta_2$ and $\beta_3$. However, in the presence of multicollinearity and repeated measures, using the principal component model with random effects ($RMPC\_Model$) is appropriate for functions with higher variability, such as $\beta_1$ and $\beta_4$.

According to this last item, future model evaluation studies should examine the sensitivity of the model to changes in internal variability structures of functional parameters.

## CRediT authorship contribution statement

**Cristhian Leonardo Urbano-Leon:** Conceptualization, Formal analysis, Investigation, Methodology, Simulation, Visualization, Writing – original draft, Writing – review & editing. **Ana María Aguilera:** Supervision, Writing – review & editing. **Manuel Escabias:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision.

## Acknowledgements

The authors acknowledge the support by PID2020-113961GB-I00 project of the Spanish Ministry of Science and Innovation (also supported by the FEDER program), research group FQM-307 of the Autonomous Government of Andalusia (Spain) and the IMAG Maria de Maeztu grant CEX2020-001105-M/AEI/10.13039/501100011033. Funding for open access charge: Universidad de Granada / CBUA.

## References

[1] C. Acal, A.M. Aguilera, Basis expansion approaches for functional analysis of variance with repeated measures, Adv. Data Anal. Classif. 186 (2023) 291–321, http://dx.doi.org/10.1007/s11634-022-00500-y.

[2] C. Acal, A.M. Aguilera, A. Sarra, A. Evangelista, T. Di-Battista, S. Palermi, Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic, Stoch. Environ. Res. Risk Assess. (2022) 1083–1101, http://dx.doi.org/10.1007/s00477-021-02071-4.

[3] A. Agresti, Foundations of Linear and Generalized Linear Models, first ed., John Wiley & Sons, New Jersey, 2015.

[4] A.M. Aguilera, M. Escabias, F.A. Ocaña, M.J. Valderrama, Functional wavelet–based modelling of dependence between lupus and stress, Methodol. Comput. Appl. Probab. 17 (4) (2015) 1015–1028, http://dx.doi.org/10.1007/s11009-014-9424-5.

[5] A.M. Aguilera, M. Escabias, C. Preda, G. Saporta, Using basis expansions for estimating functional PLS regression: Applications with chemometric data, Chemometr. Intell. Lab. Syst. 104 (2) (2010) 289–305, http://dx.doi.org/10.1016/j.chemolab.2010.09.007.

[6] D. Bates, R: Computational Methods for Mixed Models, R Foundation for Statistical Computing, Vienna, Austria, 2023, URL https://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf.

[7] M. Crowder, D. Hand, Analysis of Repeated Measures, 1st ed., Chapman and Hall, 1990.

[8] C.S. Davis, Statistical Methods for the Analysis of Repeated Measurements, Springer-Verlag New York, Inc., 2002.

[9] M. Escabias, A.M. Aguilera, C. Acal, LogitFD: An R package for functional principal component logit regression, R J. 14 (3) (2022) 231–248, http://dx.doi.org/10.32614/RJ-2022-053.

[10] M. Escabias, A.M. Aguilera, M.J. Valderrama, Principal component estimation of functional logistic regression: Discussion of two different approaches, J. Nonparametr. Stat. 16 (2004) 365–384.

[11] M. Escabias, M.J. Valderrama, A.M. Aguilera, M.H. Santofimia, M.C. Aguilera-Morillo, Stepwise selection of functional covariates in forecasting peak levels of olive pollen., Stoch. Environ. Res. Risk Assess. 27 (2013) 367–376, http://dx.doi.org/10.1007/s00477-012-0655-0.

[12] R.L. Eubank, Nonparametric Regression and Spline Smoothing, Second Edition, Marcel Dekker Inc, New York, 1998.

[13] F. Ferraty, Recent Advances in Functional Data Analysis and Related Topics, Springer, 2011.

[14] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis Theory and Practice, Springer, 2006.

[15] T.B. Fomby, S.R. Johnson, R.C. Hill, Advanced Econometric Methods, Springer, New York, NY, 1984.

[16] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer, 2012.

[17] T. Hsing, R. Eubank, Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, First, Wiley, 2015.

[18] P.D. Lax, Functional Analysis, Wiley Interscience, 2002.

[19] M.J. Lindstrom, D.M. Bates, Nonlinear mixed effects models for repeated measures data, Biometrics 46 (3) (1990) 673–687.

[20] Z. Liu, W. Guo, Functional mixed effects models, WIREs Comput. Stat. 4 (6) (2012) 527–534, http://dx.doi.org/10.1002/wics.1226.

[21] W. Ma, L. Xiao, B. Liu, M.A. Lindquist, A functional mixed model for scalar on function regression with application to a functional MRI study, Biostatistics 22 (3) (2019) 439–454, http://dx.doi.org/10.1093/biostatistics/kxz046.

[22] P. Martínez-Camblor, N. Corral, Repeated measures analysis for functional data, Comput. Statist. Data Anal. 55 (12) (2011) 3244–3256.

[23] S.N. Mousavi, H. Sorensen, Functional logistic regression: A comparison of three methods, J. Stat. Comput. Simul. 88 (2) (2018) 250–268, http://dx.doi.org/10.1080/00949655.2017.1386664.

[24] J. Olaya, Metodos de Regresión no Paramétrica, Universidad del Valle, Colombia, 2012.

[25] G. Paulon, R. Reetzke, B. Chandrasekaran, A. Sarkar, Functional logistic mixed-effects models for learning curves from longitudinal binary data, J. Speech, Lang., Hear. Res. 62 (1) (2019) 543–553, http://dx.doi.org/10.1044/2018_JSLHR-S-ASTM-18-0283.

[26] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer, 2002.

[27] J.O. Ramsay, B.W. Silverman, Funcional Data Analysis, 2da., Springer, 2005.

[28] C.R. Rao, Some statistical methods for comparison of growth curves, Biometrics 14 (1958) 1–17.

[29] W. Rudin, Functional Analysis, second ed., Mc Graw Hill, 1991.

[30] F. Scheipl, A.M. Staicu, S. Greven, Functional additive mixed models, J. Comput. Graph. Statist. 24 (2) (2015) 477–501, http://dx.doi.org/10.1080/10618600.2014.901914.

[31] L. Smaga, Repeated measures analysis for functional data using box-type approximation: With applications, REVSTAT-Stat. J. 17 (4) (2019) 523–549, http://dx.doi.org/10.57805/revstat.v17i4.279.

[32] L. Smaga, A note on repeated measures analysis for functional data, AStA Adv. Stat. Anal. 104 (2) (2020) 117–139, http://dx.doi.org/10.1007/s10182-018-00348-8.

[33] C.L. Urbano-Leon, M. Escabias, D.P. Ovalle-Munoz, J. Olaya-Ochoa, Scalar variance and scalar correlation for functional data, Mathematics 11 (1317) (2023) 1–20, http://dx.doi.org/10.3390/math11061317.