TEST (2023) 32:321–349
https://doi.org/10.1007/s11749-022-00836-9

ORIGINAL PAPER

On functional logistic regression: some conceptual issues

José R. Berrendero[1] · **Beatriz Bueno-Larraz**[2] · **Antonio Cuevas**[1]

Received: 1 February 2022 / Accepted: 14 October 2022 / Published online: 31 October 2022

Abstract
The main ideas behind the classic multivariate logistic regression model make sense when translated to the functional setting, where the explanatory variable X is a function and the response Y is binary. However, some important technical issues appear (or are aggravated with respect to those of the multivariate case) due to the functional nature of the explanatory variable. First, the mere definition of the model can be questioned: While most approaches so far proposed rely on the L^2-based model, we explore an alternative (in some sense, more general) approach, based on the theory of reproducing kernel Hilbert spaces (RKHS). The validity conditions of such RKHS-based model, and their relation with the L^2-based one, are investigated and made explicit in two formal results. Some relevant particular cases are considered as well. Second, we show that, under very general conditions, the maximum likelihood of the logistic model parameters fails to exist in the functional case, although some restricted versions can be considered. Third, we check (in the framework of binary classification) the practical performance of some RKHS-based procedures, well-suited to our model: They are compared to several competing methods via Monte Carlo experiments and the analysis of real data sets.

Keywords Functional data · Logistic regression · Reproducing kernel Hilbert spaces · Kernel methods in statistics

Mathematics Subject Classification 62J12 · 62R10

✉ José R. Berrendero
joser.berrendero@uam.es

Antonio Cuevas
antonio.cuevas@uam.es

[1] Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain

[2] Huesca, Spain

1 Introduction

Throughout this work, we study the situation in which a binary (0-1) response variable Y must be predicted in terms of a random explanatory variable X. We especially focus on the case where X is an infinite-dimensional random variable, typically a function that appears as a random trajectory of a stochastic process. In the classical, multivariate case, where X is finite-dimensional, taking values in $\mathbb{R}^d$, the logistic regression model is a popular approach to such problem (see, e.g., Hilbe 2009 or Cramer 2003, Ch. 9 for a historical overview). According to Hosmer et al. (2013, p. 52), one of the most appealing features of logistic regression is that the coefficients of the model are easily interpretable in terms of the values of the predictors. As an important additional motivation, the logistic model is necessarily fulfilled in the important case that the conditional distributions of X given Y are both Gaussian and homoscedastic. This finite-dimensional logistic model has been widely studied. Apart from the already mentioned references, Efron (1975) provides a comparison between logistic predictors and Fisher discriminant analysis. In addition, Munsiwamy and Wakweya (2011) gives a useful overview of asymptotic results of the estimators (firstly proved in Fahrmeir and Kaufmann (1985) and Fahrmeir and Kaufmann (1986)).

A number of papers have been devoted to the extension of the logistic regression model to the case of a functional valued explanatory variable X. The vast majority of them follow the so-called L^2-approach, assuming that X takes values in the standard $L^2[0, 1]$ of real square integrable functions defined on $[0, 1]$. The basic idea of this extension is to replace the Euclidean inner product in $\mathbb{R}^d$, that appears in the formulation of the multivariate logistic model, with the standard inner product in the function space $L^2[0, 1]$. An overview of several approaches to functional logistic regression under the L^2 point of view can be found in Mousavi and Sørensen (2018).

The purpose, contents and contributions of this work

The purpose of this paper is to explore another approach to the logistic regression functional problem. We will focus on a different model whose theoretical basis is provided by the theory of Reproducing kernel Hilbert spaces (RKHS). In fact, we will give two equivalent formulations of our model: The first one, based on purely probabilistic considerations, is easier to motivate. The second one, relying on "analytic" RKHS tools, is apparently a bit more sophisticated but provides an explicit parameterization, so it turns out to be much more convenient for inference and prediction purposes. Whatever the formulation, the RKHS model extends (in a sense to be made precise below) the L^2 one and includes as well, as particular cases, all the finite-dimensional models obtained by taking one-dimensional marginals from the functional explanatory X.

In Sect. 2, we will briefly review (in order to gain some perspective) the formulation and basic ideas of the L^2-based logistic regression model.

Section 3 is devoted to the formulation of our model. As mentioned above, we will do this in two alternative equivalent versions.

In Sect. 4, we study, in formal terms, the relation between our RKHS model and the "classical" L^2 one. We also show some other important particular cases, of practical interest. Finally, we investigate under which conditions the RKHS-based model nec-

essarily holds when both conditional distributions $X|Y = i$, for $i = 0, 1$ are Gaussian and homoscedastic.

In Sect. 5, we investigate maximum likelihood estimation in the new model and prove two results of non-existence for the maximum likelihood estimator. Such negative results can be seen as an aggravated, functional counterpart of the well-known partial non-existence results arising in finite-dimensional logistic models; see Candès and Sur (2020) and references therein.

In Sect. 6, we discuss how, still, some restricted versions of the maximum likelihood ideas can be used in practice. Section 7 is devoted to some experimental comparisons, via simulations and real data sets. Some brief concluding remarks are given in Sect. 8. Some classic auxiliary results used in the proofs are included in a final appendix, for reader's convenience.

2 Logistic regression in the functional case: the "classical" L^2-model

Here, for completeness and for comparison purposes, we will briefly recall the statement of the standard L^2-based functional logistic model, together with some basic ideas and a few references.

Let us recall that the goal is to explore the relationship between a dichotomous $0-1$ response variable Y and a functional predictor X. We will assume throughout that $X = X(t)$ is an L^2-stochastic process whose mean and covariance functions are continuous; so the trajectories of X are in $L^2[0, 1]$. Thus, the random variable Y conditional to X follows a Bernoulli distribution with parameter $p(X)$ and the prior probability of class 1 is denoted by $p = \mathbb{P}(Y = 1)$. In this setting, the most common functional logistic regression model is

$$\mathbb{P}(Y = 1|X) = \frac{1}{1 + \exp\{-\alpha_0 - \langle \alpha, X\rangle_2\}}, \tag{1}$$

where $\alpha_0 \in \mathbb{R}$, $\alpha \in L^2[0, 1]$ and $\langle \cdot, \cdot \rangle_2$ denotes the inner product in $L^2[0, 1]$. This model is the direct extension of the d-dimensional one, where the product in $\mathbb{R}^d$ is replaced by its functional counterpart. Model (1) will be referred to as the L^2-based logistic functional model.

The standard approach to this problem is to reduce the dimension of the curves using principal component analysis (PCA). That is, the curves $X = X(t)$ are projected into the subspace defined by the eigenfunctions corresponding to the d largest eigenvalues of the covariance operator of X. Then, standard logistic regression is applied to the resulting d-dimensional projections. Among others, this strategy has been explored by Escabias et al. (2004) and James (2002) from an applied perspective though, in fact, the latter reference deals with generalized linear models beyond logistic regression. These more general models are also studied by Müller and Stadtmüller (2005), but with a more mathematical focus.

3 An RKHS-based proposal for logistic regression in the functional case

We will explore here a different model for functional logistic regression problems which, as we will show, can be in fact expressed in two equivalent ways.

3.1 A first "probabilistic" formulation

We first establish our model in a preliminary version, relying on some basic tools of probability theory, with no resort to RKHS theory.

Our functional data will be trajectories in $L^2[0,1]$ of an L^2-process $X = X(t)$ with continuous covariance and mean function, denoted by $K = K(s,t)$ and $m = m(t)$, respectively. The covariance operator $\mathcal{K}$ associated with the covariance function K of the process is given by

$$\mathcal{K}(f)(\cdot) = \int_0^1 K(s,\cdot) f(s) \mathrm{d}s = \mathbb{E}\big[\langle X - m,\, f\rangle_2 \big(X(\cdot) - m(\cdot)\big)\big]. \tag{2}$$

Let $L^2(\Omega)$ be the Hilbert space of real random variables with finite second moment, endowed with the usual inner product and with associated norm $\|U\|^2 = \mathbb{E}(U^2)$. Define

$$\mathcal{L}_0(X) = \left\{ U \in L^2(\Omega) \,:\, U = \sum_{i=1}^{n} a_i \big(X(t_i) - m(t_i)\big),\ a_i \in \mathbb{R},\ t_i \in [0,1],\ n \in \mathbb{N} \right\},$$

where $m(t) = \mathbb{E}[X(t)]$, and let $\mathcal{L}(X)$ be the completion of $\mathcal{L}_0(X)$ in $L^2(\Omega)$, that is, $\mathcal{L}(X)$ is a subspace of $L^2(\Omega)$ defined as the closure of the linear span of the centered one-dimensional marginals of the process X. Given that $\mathcal{L}(X)$ contains the finite linear combinations of values $X(t_i) - m(t_i)$ or limits (in the mean square sense) of sequences of such linear combinations, it seems natural to consider the following functional logistic regression model:

$$\mathbb{P}(Y = 1|X) = \frac{1}{1 + \exp\{-\beta_0 - U_X\}}, \quad \beta_0 \in \mathbb{R},\ \ U_X \in \mathcal{L}(X). \tag{3}$$

As we will show below (see Theorem 1), model (3) includes the L^2-based model (1) as a particular case, since $\langle \alpha, X\rangle_2 = \int_0^1 \alpha(t)X(t)dt \in \mathcal{L}(X)$ for $\alpha \in L^2[0,1]$.

On the other hand, (3) is more general than the L^2-formulation (1) since, for instance, it includes finite-dimensional versions of the form

$$\mathbb{P}(Y = 1|X) = \frac{1}{1 + \exp\{-\beta_0 - \sum_{j=1}^{p} \beta_j X(t_j)\}}$$

(see Theorem 2 below) that cannot be expressed in terms of $\langle \alpha, X\rangle_2$ for any $\alpha \in L^2[0,1]$.

For linear regression models and in the context of variable selection methods, an analogous generalization of the usual L^2-model was proposed by Berrendero et al. (2019).

Some closely related ideas, aimed at the prediction problem in functional linear models, are also present in Shin and Hsing (2012), although these authors do not consider the RKHS aspects we study here.

3.2 Reproducing kernel Hilbert spaces

Since we want to reformulate model (3) in RKHS terms, we will briefly remind here, for the sake of completeness, some very basic ideas and notations about RKHS's; see Berlinet and Thomas-Agnan (2004) and Appendix F of Janson (1997) for further details and references. Furthermore, the book by Hsing and Eubank (2015) provides an excellent mathematical background on mathematical methods, including RKHS theory, for the statistical analysis of functional data. The papers by Hsing and Ren (2009) and Kneip and Liebl (2020) offer also very general perspectives and results on the applicability of RKHS methods in functional regression models, though not particularly focused on the logistic case. Some other more specific references (a few of them especially dealing with the functional logistic model) will be cited below.

Let $\mathcal{H}_0(K) := \{f \in L^2[0,1] : f(\cdot) = \sum_{i=1}^n a_i K(t_i, \cdot),\ a_i \in \mathbb{R},\ t_i \in [0,1],\ n \in \mathbb{N}\}$, be the space of all finite linear combinations of evaluations of K. This space is endowed with the inner product $\langle f, g\rangle_K = \sum_{i,j} \alpha_i \beta_j K(t_i, s_j)$, where $f(\cdot) = \sum_i \alpha_i K(t_i, \cdot)$ and $g(\cdot) = \sum_j \beta_j K(s_j, \cdot)$. Then, the RKHS associated with K is defined as the completion of $\mathcal{H}_0(K)$. In other words, $\mathcal{H}(K)$ is made of all functions obtained as pointwise limits of Cauchy sequences in $\mathcal{H}_0(K)$. The inner product is extended accordingly to the whole space $\mathcal{H}(K)$.

These spaces are named after the so-called *reproducing property*, $\langle f, K(s,\cdot)\rangle_K = f(s)$, for all $f \in \mathcal{H}(K)$, $s \in [0,1]$, which is particularly important in the applications. On account of this property it is sometimes said that RKHSs are spaces of "true functions," in the sense that the pointwise values $f(s)$, at a given s do matter, by contrast with $L^2[0,1]$ whose elements are in fact equivalence classes of functions.

3.3 The RKHS formulation of the functional logistic model

Recall that $\mathcal{L}(X)$ denotes the closure of the linear span of the centered one-dimensional marginals of the process X. A property of RKHSs especially useful in statistical applications is given by the following isometry result: The transformation Ψ_X, from $\mathcal{L}(X)$ to $\mathcal{H}(K)$, defined by

$$\Psi_X(U)(s) = \mathbb{E}[U(X(s) - m(s))] = \langle U, X(s) - m(s)\rangle \in \mathcal{H}(K), \quad \text{for } U \in \mathcal{L}(X) \tag{4}$$

is an isometry (sometimes called *Loève's isometry*) between $\mathcal{L}(X)$ and $\mathcal{H}(K)$, that is, $\Psi_X(U)$ is bijective and preserves the inner product (see Lukić and Beder 2001, Lemma 1.1). As a consequence, the Hilbert spaces $\mathcal{L}(X)$ and $\mathcal{H}(K)$ can be identified.

Note that, in informal terms, Ψ_X is the completion of the transformation from $\mathcal{L}_0(X)$ to $\mathcal{H}_0(K)$ given by $\sum_{i=1}^n a_i \big(X(t_i) - m(t_i)\big) \mapsto \sum_{i=1}^n a_i K(t_i, \cdot)$.

It is worth mentioning that while $\mathcal{H}(K)$ is, in several aspects, a natural Hilbert space associated with the process X, typically, the trajectories of the process X themselves do not belong to $\mathcal{H}(K)$ with probability one (see, e.g., Lukić and Beder 2001, Cor. 7.1; Pillai et al. 2007, Th. 11). Then, one cannot directly write $\langle x, K(s, \cdot)\rangle_K$, for x a realization of the process. However, following Parzen (1961), we will use the convenient notation $\langle x, K(s, \cdot)\rangle_K$ interpreting this expression in terms of Loève's isometry, Ψ_X; more precisely, we will identify $\langle x, f\rangle_K$ with $\Psi_x^{-1}(f) := (\Psi_X^{-1}(f))(\omega)$, for $x = X(\omega)$ and $f \in \mathcal{H}(K)$, which in particular means $\Psi_X^{-1}(\sum_{i=1}^n a_i K(t_i, \cdot)) = \sum_i a_i (X(t_i) - m(t_i))$.

The intuition behind the definition of $\langle \beta, X\rangle_K$ is reminiscent of the definition of Itô's isometry, which is used to define the stochastic integral with respect to the Wiener measure (Brownian motion), overcoming the fact that the Brownian trajectories are not of bounded variation.

The considerations above allow us to rewrite model (3) in terms of the isometric transformation $X \mapsto \langle \beta, X\rangle_K = \Psi_X^{-1}(\beta)$:

$$\mathbb{P}(Y = 1 \mid X) = \frac{1}{1 + \exp\{-\beta_0 - \langle \beta, X\rangle_K\}}, \quad \beta_0 \in \mathbb{R}, \quad \beta \in \mathcal{H}(K). \tag{5}$$

There is a one-to-one correspondence between each function in $\mathcal{H}(K)$ and each random variable in $\mathcal{L}(X)$. It is in this sense that $\mathcal{H}(K)$ is a natural parameter space for model (3). Formulation (5) turns out to be more convenient for estimation and prediction purposes.

4 Some important particular instances: the Gaussian case

4.1 Relationship between the L^2- and the RKHS-based models

In this subsection, we clarify the relation between the standard (centered on X, by convenience) L^2-logistic model

$$\mathbb{P}(Y = 1|X) = \frac{1}{1 + \exp\{-\alpha_0 - \langle \alpha, X - m\rangle_2\}}, \quad \alpha_0 \in \mathbb{R}, \quad \alpha \in L^2[0, 1], \tag{6}$$

and the RKHS version (3), or its equivalent formulation (5). We will essentially show that the RKHS-based model is more general, in the sense that, whenever the L^2-formulation (6) holds, then the RKHS counterpart (3) also holds, in a re-parametrized version (that is, the respective "slope functions" α and β will be different and with different interpretations). We will give as well a condition under which the logistic model (5) formulated in RKHS terms can be rewritten into L^2-terms as in (6).

This is made explicit in the following result:

Theorem 1 *(a) If the L^2-model (6) holds, then the RKHS model (5) holds as well with $\beta_0 = \alpha_0$ and $\beta = \mathcal{K}(\alpha)$.*

(b) If the RKHS model (5) holds and $\beta \in \mathcal{K}(L^2) = \{\mathcal{K}(f) : f \in L^2[0,1]\}$*, then the* L^2*-model (6) holds with* $\alpha_0 = \beta_0$ *and* $\mathcal{K}(\alpha) = \beta$.

Proof Let us denote $\|U\| = [\mathbb{E}(U^2)]^{1/2}$, the $L^2(\Omega)$-norm, and $\|f\|_2 = [\int_0^2 f^2(t)dt]^{1/2}$, the $L^2[0,1]$-norm.

(a) Let $U := \int_0^1 \alpha(t)(X(t) - m(t))dt$, $\alpha \in L^2[0,1]$. We have to show that $U \in \mathcal{L}(X)$. Since continuous functions are dense in $L^2[0,1]$, there exists a sequence of continuous functions α_n such that $\|\alpha_n - \alpha\|_2 \to 0$. Let $U_n := \int_0^1 \alpha_n(t)(X(t) - m(t))dt$. We have that $U_n \in \mathcal{L}(X)$ (see, e.g., Ash and Gardner 2014, page 34). Since $\mathcal{L}(X)$ is closed, it is enough to show that $\|U_n - U\| \to 0$. Indeed, using Cauchy-Schwarz inequality and Fubini's theorem

$$\|U_n - U\|^2 = \mathbb{E}\left[\left(\int_0^1 (\alpha_n - \alpha)X(t)dt\right)^2\right] \leq \|\alpha_n - \alpha\|_2^2 \int_0^1 \mathbb{E}(X(t)^2)dt \to 0,$$

taking into account $\int_0^1 \mathbb{E}(X(t)^2) = \int_0^1 (K(t,t) + m(t)^2)dt < \infty$ (recall we are assuming m and K are continuous). Now, $U \in \mathcal{L}(X)$ if and only if $U = \Psi_X^{-1}(\beta) = \langle \beta, X\rangle_K$ for $\beta \in \mathcal{H}(K)$. Moreover, by Fubini's theorem, for all $s \in [0,1]$, we have

$$\begin{aligned}\beta(s) = \Psi_X(U)(s) &= \mathbb{E}\left[\int_0^1 \alpha(t)(X(t) - m(t))dt \cdot (X(s) - m(s))\right] \\ &= \int_0^1 K(s,t)\alpha(t)dt = \mathcal{K}(\alpha)(s). \end{aligned} \tag{7}$$

(b) If $\beta = \mathcal{K}(\alpha)$, from (7), we get $\langle X, \beta\rangle_K = \Psi_X^{-1}(\beta) = U = \int_0^1 \alpha(t)(X(t) - m(t))dt$.

□

4.2 Some other important particular cases

In this subsection, we show that, in fact, the RKHS model (5) is strictly more general than the L^2 formulation (1–6), by showing (see Th. 2 (a) below) some relevant particular cases of (5) that cannot be formulated in L^2-terms. They appear in the context of dimension reduction techniques, which are often a natural alternative, motivated by criteria of interpretability of the model and classification accuracy. In fact, we will consider here two usual ways of performing dimension reduction: variable selection and linear projections.

By variable selection, we mean to replace each curve x_i by the finite-dimensional vector $(x_i(t_1), \ldots, x_i(t_p))$, for some $t_1, \ldots, t_p$ chosen in an optimal way. In this section, we analyze under which conditions it is possible to perform functional variable selection without loss. Such analysis is only feasible under the following particular RKHS model.

Whenever the slope function β has the form

$$\beta(\cdot) = \sum_{j=1}^{p} \beta_j K(t_j, \cdot), \tag{8}$$

the model in (5) is reduced to the finite-dimensional one,

$$\mathbb{P}(Y = 1|X) = \left(1 + \exp\left\{-\beta_0 - \sum_{j=1}^{p} \beta_j (X(t_j) - m(t_j))\right\}\right)^{-1}. \tag{9}$$

The main difference between the standard finite-dimensional model and this one is that now the proper choice of the points $T = (t_1, \ldots, t_p) \in [0, 1]^p$ is a part of the estimation procedure. In this sense, model (9) is truly functional since we will use the whole trajectories $x_i(t)$ to select the points. This fact leads to a critical difference between the functional and the multivariate problems. Then, our aim is to approximate the general model described by Eq. (5) with finite-dimensional models as those of Eq. (9). This amounts to get an approximation of the slope function in terms of a finite linear combination of kernel evaluations $K(t_j, \cdot)$. This model, for $p = 1$ and a particular type of Gaussian process X, is analyzed in Lindquist and McKeague (2009).

Of course, when (8–9) holds, variable selection is particularly compelling. A natural idea in this setting would be to incorporate the points $t_1, \ldots, t_p$ to the estimation procedure as additional parameters to be estimated; see Sect. 5 for details.

Another standard way of dimension reduction is done by linear projections. This is the case, for example, of principal component analysis (PCA). As it is well known, the dimension reduction in this case is achieved by replacing the whole trajectory $x = x(t)$ by a vector in $\mathbb{R}^p$ such as $(\langle u_1, x\rangle_2, \ldots, \langle u_p, x\rangle_2)$ whose components are the projections of x along some suitable chosen "directions" $u_j \in L^2[0, 1]$. Such strategy appears also as a particular case of the RKHS-based logistic model (which, in fact, can be also formulated in L^2-terms). This is shown in Th. 2(b) below.

Theorem 2 *Assume model (5) holds. Then,*

(a) If there exists a positive integer p, $\beta_1, \ldots, \beta_p \in \mathbb{R}$, and $t_1, \ldots, t_p \in [0, 1]$ such that $\beta(\cdot) = \sum_{j=1}^{p} \beta_j K(\cdot, t_j)$, then

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{1}{1 + \exp\left\{-\beta_0 - \sum_{j=1}^{p} \beta_j (x(t_j) - m(t_j))\right\}}.$$

(b) Let $\{u_j\}$ be an orthonormal basis of $L^2[0, 1]$. If there exists a positive integer p, and $\beta_1, \ldots, \beta_p \in \mathbb{R}$ such that $\beta = \sum_{j=1}^{p} \beta_j \mathcal{K}(u_j)$, then

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{1}{1 + \exp\left\{-\beta_0 - \sum_{j=1}^{p} \beta_j \langle x - m, u_j\rangle_2\right\}},$$

Proof (a) Observe that for $j = 1, \ldots, p$, $X(t_j) - m(t_j) \in \mathcal{L}_0(K)$, and for all $s \in [0, 1]$,

$$\Psi_X(X(t_j) - m(t_j))(s) = \mathbb{E}[(X(s) - m(s))(X(t_j) - m(t_j))] = K(s, t_j).$$

Therefore $\Psi_X^{-1}(k(\cdot, t_j)) = X(t_j) - m(t_j)$, and

$$\langle X, \beta \rangle_K = \sum_{j=1}^{p} \beta_j \langle X, k(\cdot, t_j) \rangle_K = \sum_{j=1}^{p} \beta_j (X(t_j) - m(t_j)).$$

(b) Putting $\alpha(t) = u_j(t)$ in (7), we get $\Psi_X(\langle u_j, X - m \rangle_2) = \mathcal{K}(u_j)$. As a consequence,

$$\langle X, \beta \rangle_K = \sum_{j=1}^{p} \beta_j \langle u_j, X - m \rangle_2.$$

□

Part (a) of the previous result means that for some particular choices of the slope function of type $\beta(\cdot) = \sum_i^p a_i K(t_i, \cdot)$, the model (5) amounts to a finite-dimensional logistic regression model for which the explanatory variable is a p-dimensional marginal of the process X. Thus, the impact-point model studied by Lindquist and McKeague (2009) appears as a particular case of the RKHS-based model. In fact, model (5) can be seen as a true extension of the finite-dimensional logistic regression model, which is obtained when a finite-dimensional covariance matrix plays the role of the kernel. As an important by-product, this provides a mathematical ground for variable selection in logistic regression. As it turns out, functions of type $\beta(\cdot) = K(\cdot, t)$ belong to $\mathcal{H}(K)$ but do not belong to $\mathcal{K}(L^2)$. This fact implies that within the setting of the RKHS model, it is possible to regress Y on any finite-dimensional projection of X, whereas this does not make sense if we consider the L^2 model. This feature is clearly relevant if one wishes to analyze properties of variable selection methods.

Part (b) of Theorem 2 implies that model (5) also includes situations, like PCA, where the explanatory variable is replaced by p linear projections. The corresponding model is a particular case of (5), where β is in the span of $\mathcal{K}(u_1), \ldots, \mathcal{K}(u_p)$. Note that, if $\{u_j\}$ is the orthonormal basis of eigenfunctions of $\mathcal{K}$, we have $\mathcal{K}(u_j)$ is proportional to u_j, and the condition on β reduces to the fact that β belongs to the span of $u_1, \ldots, u_p$. If this is the case, there is no loss in using the first p principal components of the regressors instead of the whole trajectories.

4.3 Validity conditions in the Gaussian case

Generally speaking, the logistic regression model specifies the conditional distribution of the response Y given the regressor X. However, as in the finite-dimensional case, our model holds when the conditional distributions of the process given the two possible values of Y are Gaussian with the same covariance structure. Indeed, in this subsection,

we prove that (5) also holds when we assume that $X|Y=0$ and $X|Y=1$ are Gaussian and homoscedastic, with regular enough mean functions. Of course, (5) also may hold for other non-Gaussian assumptions on the conditional distributions $X|Y=i$.

More precisely, for $i=0,1$, assume that $\{X(t) : t \in [0,1]\}$ given $Y=i$ is a Gaussian process with continuous mean function m_i and continuous covariance function K (the same for $i=0,1$). We will assume as well throughout this subsection that all the eigenvalues λ_i of the covariance operator $\mathcal{K}$, associated with K are strictly positive (so $\mathcal{K}$ is injective). Note that, as a consequence of Spectral Theorem (see, e.g., Hsing and Eubank 2015, p.98) $\mathcal{K}x = \sum_j \lambda_i \langle x, e_i\rangle e_i$, where e_i stands for a unit eigenvector associated with λ_i; thus, the inverse $\mathcal{K}^{-1}$ is defined on the range of $\mathcal{K}$, $\mathcal{K}(L^2)$, as a linear (not continuous) transformation, by $\mathcal{K}^{-1}y = \sum_i \frac{\langle y,e_i\rangle}{\lambda_i} e_i$), for $y = \sum_i \langle y, e_i\rangle e_i \in \mathcal{K}(L^2)$.

Let P_{m_0} and P_{m_1} be the probability measures (i.e., the distributions) induced by the process X conditional to $Y=0$ and $Y=1$, respectively. Recall that when m_0 and m_1 both belong to $\mathcal{H}(K)$, we have that P_{m_0} and P_{m_1} are mutually absolutely continuous; see Theorem 7A in Parzen (1961) and the Appendix. The following theorem provides a very natural motivation for the RKHS model (5) in this Gaussian setting.

Theorem 3 *Let P_{m_0}, P_{m_1} and $\mathcal{K}$ be as in the previous lines. Then,*

(a) if $m_1 - m_0 \in \mathcal{H}(K)$, then P_{m_0} and P_{m_1} are mutually absolutely continuous and model (5) holds,

$$\mathbb{P}(Y=1 \mid X=x) = \frac{1}{1+\exp\{-\beta_0 - \langle x,\beta\rangle_K\}} \equiv \frac{1}{1+\exp\left\{-\beta_0 - \Psi_x^{-1}(\beta)\right\}},$$

with $\beta := m_1 - m_0$ and $\beta_0 := -\mathbb{E}_{m_1}[\Psi_x^{-1}(\beta)] + \|m_1 - m_0\|_K^2/2 - \log((1-p)/p)$ (where $p=\mathbb{P}(Y=1)$ and $\mathbb{E}_{m_1}(\cdot)$ stands for the expectation when the process has mean function equal to m_1).

(b) If $m_1 - m_0 \notin \mathcal{H}(K)$, then P_{m_0} and P_{m_1} are mutually singular.

(c) if $m_1 - m_0 \in \mathcal{K}(L^2) = \{\mathcal{K}(f) : f \in L^2[0,1]\}$, then P_{m_0} and P_{m_1} are mutually absolutely continuous and model (1) holds.

(d) if $m_1 - m_0 \notin \mathcal{K}(L^2)$ model (1) is never recovered, but different situations are possible, according to the condition in part (a). In particular if $m_0 = 0$, $m_1 \in \mathcal{H}(K)$ recovers scenario (a), but if $m_1 \notin \mathcal{H}(K)$, P_{m_0} and P_{m_1} are mutually singular.

Proof (a) and (b) Let P_0 be the measure induced by a Gaussian process with covariance function K but zero mean function, $m \equiv 0$. From Theorem 7A in Parzen (1961), $m_0 - m_1 \in \mathcal{H}(K)$ implies that $P_{m_0-m_1}$ and P_0 are mutually absolutely continuous, and $m_0 - m_1 \notin \mathcal{H}(K)$ implies that $P_{m_0-m_1}$ and P_0 are mutually singular. By Lemma 1.1 in Pitcher (1960), see Appendix, $P_{m_0-m_1}$ and P_0 are mutually absolutely continuous if and only if P_{m_0} and P_{m_1} are mutually absolutely continuous and, in this case, the corresponding Radon-Nikodym derivative fulfills

$$\frac{\mathrm{d}P_{m_0}}{\mathrm{d}P_{m_1}}(X) = \frac{\mathrm{d}P_{m_0-m_1}}{\mathrm{d}P_0}(X - m_1) = \exp\left\{\langle X - m_1, m_0 - m_1\rangle_K - \frac{1}{2}\|m_0 - m_1\|_K^2\right\}.$$

The last equality also follows from Theorem 7A in Parzen (1961). Notice that by the definition of Loéve's isometry, we have $\langle X - m_1, m_0 - m_1\rangle_K = \langle X, m_0 - m_1\rangle_K - \mathbb{E}_{m_1}[\langle X, m_0 - m_1\rangle_K]$.

The conditional probability of $Y = 1$ can be expressed in terms of the Radon–Nikodym derivative of P_1 with respect to P_0 (see Baíllo et al. 2011, Th.1) by

$$\mathbb{P}(Y = 1 \mid X) = \frac{p\frac{\mathrm{d}P_{m_1}}{\mathrm{d}P_{m_0}}(X)}{p\frac{\mathrm{d}P_{m_1}}{\mathrm{d}P_{m_0}}(X) + (1-p)} = \left(1 + \frac{1-p}{p}\frac{\mathrm{d}P_{m_0}}{\mathrm{d}P_{m_1}}(X)\right)^{-1}. \quad (10)$$

From the last two displayed equations, one can rewrite

$$\mathbb{P}(Y = 1 \mid X) = \left(1 + \frac{1-p}{p}\exp\left\{\langle X, m_0 - m_1\rangle_K - \mathbb{E}_{m_1}[\langle X, m_0 - m_1\rangle_K] - \frac{1}{2}\|m_0 - m_1\|_K^2\right\}\right)^{-1}.$$

Then, reordering terms in this expression, we get the logistic model in part (a).

(c) Under the assumptions, Theorem 6.1 in Rao and Varadarajan (1963), see Appendix, gives the following expression:

$$\log\left(\frac{\mathrm{d}P_{m_1}}{\mathrm{d}P_{m_0}}(x)\right) = \langle x - m_0,\, \mathcal{K}^{-1}(m_1 - m_0)\rangle_2 - \frac{1}{2}\langle m_1 - m_0,\, \mathcal{K}^{-1}(m_1 - m_0)\rangle_2,$$

for $x \in L^2[0,1]$. This entails (using the Chain Rule for Radon-Nikodym derivatives)

$$\frac{\mathrm{d}P_{m_0}}{\mathrm{d}P_{m_1}}(x) = C\exp\left(-\langle x, \alpha\rangle_2\right),$$

where $\alpha = \mathcal{K}^{-1}(m_1 - m_0)$ and $C = \exp\langle m_0 + m_1, \alpha\rangle_2/2)$. Now, replacing this expression in (10), we get the L^2-model (1) with $\alpha_0 = -\log\left(\frac{1-p}{p}C\right)$.

(d) Also, as a consequence of Theorem 6.1 in Rao and Varadarajan (1963), if $m_1 - m_0 \notin \mathcal{K}(L^2)$, it is not possible to express the Radon-Nikodym derivative in terms of inner products in L_2 or, equivalently, there is no continuous linear functional $L(x)$ and $c \in \mathbb{R}$ such that $\log(\frac{\mathrm{d}P_1}{\mathrm{d}P_0}(x)) = L(x) + c$. Finally, the last sentence of the statement is a consequence of Theorem 7A in Parzen (1961). □

Part (b) of this theorem has been recently observed by Petrovich et al. (2018, Th. 1) without reference to RKHS theory. Note that, in general, $m_1 - m_0 \notin \mathcal{K}(L^2)$ does not imply that P_{m_1} and P_{m_0} are orthogonal since the precise condition for this is $m_1 - m_0 \notin \mathcal{H}(K)$ and $\mathcal{K}(L^2) \subsetneq \mathcal{H}(K)$ (in fact $\mathcal{H}(K) = \mathcal{K}^{1/2}(L^2)$ equipped with the RKHS inner product, see, e.g., Hsing and Eubank (2015, Th. 7.6.4)). Parts (a) and (c) of the theorem above clarify this point. From part (b) of the theorem, it follows that the L^2 model is also recovered when a higher degree of "regularity" on the mean functions is imposed, since the functions in $\mathcal{K}(L^2)$ are convolutions of the functions in $L^2[0,1]$ with the covariance function of the process so that they are in a way more regular than the functions in $\mathcal{K}^{1/2}(L^2)$.

5 Maximum likelihood estimation: non-existence results

In the finite-dimensional setting, it is well known that the maximum likelihood (ML) estimator does not exist when there is an hyperplane separating the observations of the two classes; see below for details. As we will show in this section, this fact worsens dramatically for the case of functional data; more specifically, we will see that:

For a wide class of processes (including the Brownian motion), the MLE just does not exist, with probability one (see Sect. 5.1).
Under some different conditions, in the Gaussian case, the probability of non-existence of the MLE tends to one when the sample size tends to infinity (see Sect. 5.2).

A brief overview of the finite-dimensional case

Despite the fact that ML estimation of the slope function for multiple logistic regression is widely used, it has an important drawback that is sometimes overlooked. Given a sample $x_i^0 \in \mathbb{R}^d$ for $i = 1, \ldots, n_0$ drawn from population zero and another sample $x_i^1 \in \mathbb{R}^d$ for $i = 1, \ldots, n_1$ drawn from population one, the classical MLE in logistic regression is the vector $(b_0, b) \in \mathbb{R} \times \mathbb{R}^d$ that maximizes the log-likelihood

$$L_n(b, b_0) = \frac{1}{n_0} \sum_{i=1}^{n_0} \log\left(\frac{e^{-b_0 - b' x_i^0}}{1 + e^{-b_0 - b' x_i^0}}\right) + \frac{1}{n_1} \sum_{i=1}^{n_1} \log\left(\frac{1}{1 + e^{-b_0 - b' x_i^1}}\right),$$

where $a'b$ stands here for the inner product of two vectors a, b in $\mathbb{R}^d$. The existence and uniqueness of such a maximum were carefully studied by Albert and Anderson (1984) (and previously by Silvapulle (1981) and Gourieroux and Monfort (1981)). As stated in Theorem 1 of Albert and Anderson (1984), the latter expression can be made arbitrarily close to zero (note that, the log-likelihood is always negative) whenever the samples of the two populations are linearly separable. In that case, the maximum cannot be attained, and then, the MLE does not exist (the idea behind the proof is similar to the one of Theorem 4 below). There is another scenario where this estimator does not exist; the samples are linearly separable except for some points of both populations that fall into the separation hyperplane (named "quasicomplete separation"). In this case, the supremum of the log-likelihood function is strictly smaller than zero, but it is anyway unattainable.

The likelihood function in the logistic functional model

Before going on with the functional case (which is our main target here), we need to derive the likelihood function. Let assume that $\{X(s), s \in [0, 1]\}$ follows the RKHS logistic model described in Eq. (5). That is,

$$\beta_0 + \Psi_X^{-1}(\beta) \equiv \beta_0 + \langle X, \beta \rangle_K = \log\left(\frac{p_{\beta,\beta_0}(X)}{1 - p_{\beta,\beta_0}(X)}\right),$$

where $p_{\beta,\beta_0}(X) = \mathbb{P}(Y = 1 | X, \beta, \beta_0)$, $\beta_0 \in \mathbb{R}$ and $\beta \in \mathcal{H}(K)$. The random element $(X(\cdot), Y)$ takes values in the space $Z = L^2[0, 1] \times \{0, 1\}$, which is a measurable space

with measure $z = P_X \times \mu$, where P_X is the distribution induced by the process X and μ is the counting measure on $\{0, 1\}$. We can define in Z the measure $P_{(X,Y);\beta,\beta_0}$, the joint probability induced by $(X(\cdot), Y)$ for a given slope function β and an intercept β_0. Then, we define,

$$f_{\beta,\beta_0}(x, y) = \frac{\mathrm{d}P_{(X,Y);\beta,\beta_0}}{\mathrm{d}z}(x, y) = p_{\beta,\beta_0}(x)^y \left(1 - p_{\beta,\beta_0}(x)\right)^{1-y}$$
$$= \left(\frac{1}{1 + e^{-\beta_0 - \langle \beta, x\rangle_K}}\right)^y \left(\frac{e^{-\beta_0 - \langle \beta, x\rangle_K}}{1 + e^{-\beta_0 - \langle \beta, x\rangle_K}}\right)^{1-y}.$$

In view of this density function, the log-likelihood function for a given sample in $L^2[0, 1] \times \{0, 1\}$ is

$$L_n(\beta, \beta_0) = \frac{1}{n} \sum_{i=1}^{n} \log \left(p_{\beta,\beta_0}(x_i)^{y_i} \left(1 - p_{\beta,\beta_0}(x_i)\right)^{1-y_i}\right),$$

where $(x_i, y_i) \in L^2[0, 1] \times \{0, 1\}$ is a sample of the underlying random variable (X, Y).

The maximum likelihood estimator is the pair $(\widehat{\beta}, \widehat{\beta}_0)$ that maximizes this function L_n. The population counterpart of L_n is the expected log-likelihood function,

$$L(\beta, \beta_0) = \mathbb{E}_Z \left[\log f_{\beta,\beta_0}(X, Y)\right] = \mathbb{E}_Z \left[\log \left(p_{\beta,\beta_0}(X)^Y \left(1 - p_{\beta,\beta_0}(X)\right)^{1-Y}\right)\right], \tag{11}$$

where $\mathbb{E}_Z[\cdot]$ denotes the expectation with respect to the distribution of Z.

The main idea behind ML estimation stands in the infinite-dimensional situation. If our "parameter space" is $\Theta \subset \mathcal{H}(K) \times \mathbb{R}$ and the "true" value of the parameter is $(\beta^*, \beta_0^*) \in \Theta$, then a simple, standard argument based on Jensen's inequality shows that the population log-likelihood function $L(\beta, \beta_0)$ fulfills

$$L(\beta^*, \beta_0^*) \geq L(\beta, \beta_0), \quad \text{for all } (\beta, \beta_0) \in \Theta.$$

This leads to the usual, natural idea of maximizing a consistent estimator of $L(\beta^*, \beta_0^*)$ that, in our logistic model, is the log-likelihood function $L_n(\beta, \beta_0)$ defined above.

5.1 Non-existence of the MLE in functional settings

We first show that, when moving from the finite-dimensional model to the functional one, the problem of the non-existence of the MLE is drastically worsened.

This situation is quite similar to that arising in other statistical problems with infinite-dimensional parameter spaces. This is, for example, the case of density estimation where nonparametric, non-penalized, ML estimators do not exist, unless some drastic restrictions are imposed on the underlying density function; see Grenander (1981).

Since the analogous non-existence result for the case of the functional logistic regression model is not perhaps so direct, it is established in Theorem 4 below. We confine ourselves to the RKHS-based model (5), although the result can be easily extended, with a completely similar method of proof, for the standard L^2-based model of Eq. (1).

We first will need to establish a condition which plays, in the functional case, a similar role to that of the linear separability condition mentioned above in the setting of finite-dimensional logistic regression.

Assumption 1 (**SC**) The multivariate process $W(t) = (X_1(t), \ldots, X_n(t))$, $t \in [0, 1]$ satisfies the "Sign Choice" (SC) property when for all possible choice of signs $(s_1, \ldots, s_n)$, where s_j is either $+$ or $-$, we have that, with probability one, there exists some $t_0 \in [0, 1]$ such that $\text{sign}(X_1(t_0)) = s_1, \ \ldots, \ \text{sign}(X_n(t_0)) = s_n$.

Without loss of generality, we assume $\mathbb{E}(X(t)) = 0$. The non-existence result is as follows.

Theorem 4 *Let $X(s)$, $s \in [0, 1]$, be an L^2 stochastic process with $\mathbb{E}(X(s)) = 0$. Denote by K the corresponding covariance function. Consider a logistic model (5) based on $X(s)$. Let $X_1, \ldots, X_n$ be independent copies of X. Assume that the n-dimensional process $Z_n(s) = (X_1(s), \ldots, X_n(s))$ fulfills the SC property. Then, with probability one, the MLE estimator of (β, β_0) (i.e., the maximizer of the log-likelihood function $L_n(\beta, \beta_0)$) does not exist for any sample size n.*

Proof Let $x_1(s) \ldots, x_n(s)$ be a random sample drawn from $X(s)$. From the SC assumption, there is (with probability 1) one point t_0 such that $x_i(t_0) > 0$ for all i such that $y_i = 1$ and $x_i(t_0) < 0$ for those indices i with $y_i = 0$. Note that, the sample log-likelihood function can be split in two terms, as follows,

$$L_n(\beta, \beta_0) \ = \frac{1}{n} \sum_{\{i:\, y_i = 1\}} \log\Big(\frac{1}{1 + e^{-\beta_0 - \langle \beta, x_i \rangle_K}}\Big) + \frac{1}{n} \sum_{\{i:\, y_i = 0\}} \log\Big(\frac{e^{-\beta_0 - \langle \beta, x_i \rangle_K}}{1 + e^{-\beta_0 - \langle \beta, x_i \rangle_K}}\Big).$$

Note also that $L_n(\beta, \beta_0) \leq 0$ for all β. Now, take a numerical sequence $0 < c_m \uparrow \infty$ and define

$$\beta_m(\cdot) \ = \ c_m K(t_0, \cdot).$$

Then, by the definition of Loéve's isometry, if $y_i = 0$,

$$\langle \beta_m, x_i \rangle_K \ = \ c_m x_i(t_0) \ \to \ \infty, \ \text{ as } m \to \infty,$$

since we have taken t_0 such that $x_i(t_0) > 0$ for those indices i with $y_i = 1$. Likewise, $\langle \beta_m, x_i \rangle_K$ goes to $-\infty$ whenever $y_i = 0$ since we have chosen t_0 such that $x_i(t_0) < 0$ for those indices. As a consequence, $L_n(\beta_m, 0) \to 0$ as $m \to \infty$. Therefore, the likelihood function can be made arbitrarily large so that the MLE does not exist. □

Remark 1 A non-existence result for the MLE estimator, analogous to that of Theorem 4, can be also obtained with a very similar reasoning for the L^2-based logistic model of Eq. (1). The main difference in the proof would be the construction of β_m which, in the L^2 case, should be obtained as an approximation to the identity (that is, a linear "quasi Dirac delta") centered at the point t_0.

Although the SC property could seem a somewhat restrictive assumption, the following proposition shows that it applies to some important and non-trivial situations.

Proposition 1 *(a) The n-dimensional Brownian motion fulfills the SC property.*
(b) The same holds for any other n-dimensional process in $[0, 1]$ *whose independent marginals have a distribution absolutely continuous with respect to that of the Brownian motion.*

Proof (a) Given the n-dimensional Brownian motion $\mathcal{B}_n(t) = (B_1(t), \ldots, B_n(t))$, where the B_j are independent copies of the standard Brownian motion $B(t), t \in [0, 1]$, take a finite sequence of signs $(s_1, ..., s_n)$ and define the event

$$A = \{\text{for any given } t \text{ there exists } 0 < t_0 < t \text{ such that } \operatorname{sign}(B_j(t_0)) = s_j,\ j = 1, \ldots, n\} \quad (12)$$

We may express this event by

$$A = \bigcap_{t \in (0,1] \cap \mathbb{Q}} A_t, \quad (13)$$

where, for each $t \in (0, 1] \cap \mathbb{Q}$,

$$A_t = \{\text{there exists } t_0 < t \text{ such that } \operatorname{sign}(B_j(t_0)) = s_j,\ j = 1, \ldots, n\}.$$

Now, the result follows directly from Blumenthal's 0-1 Law for n-dimensional Brownian processes (see, e.g., Mörters and Peres 2010, p. 38). Such result establishes that for any event $A \in \mathcal{F}^+(0)$, we have either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$. Here, $\mathcal{F}^+(0)$ denotes the *germ* σ*-algebra* of events depending only on the values of $\mathcal{B}_n(t)$ where t lies in an arbitrarily small interval on the right of 0. More precisely,

$$\mathcal{F}^+(0) = \bigcap_{t>0} \mathcal{F}^0(t), \text{ where } \mathcal{F}^0(t) = \sigma(\mathcal{B}_n(s), 0 \le s \le t).$$

From (12) and (13), it is clear that the above defined event A belongs to the germ σ-algebra $\mathcal{F}^+(0)$. However, we cannot have $\mathbb{P}(A) = 0$ since (from the symmetry of the Brownian motion) for any given t_0 the probability of $\operatorname{sign}(B_j(t_0)) = s_j,\ j = 1, \ldots, n$ is $1/2^n$. So, we conclude $\mathbb{P}(A) = 1$ as desired.

(b) If $X(t)$ is another process whose distribution is absolutely continuous with respect to that of the n-dimensional Brownian motion $\mathcal{B}_n$, then the set A, defined by (12) and (13) in terms of $\mathcal{B}_n$, has also probability one when it is defined in terms of the process $X(t)$: Recall that, from the definition of absolute continuity, if the set A^c has probability zero under the Brownian motion, then its probability must be zero as well

when $B(t)$ is replaced with $X(t)$. Therefore, the probability of A under $X = X(t)$ must be one. □

Remark 2 As pointed out in Mörters and Peres (2010, p. 63), Blumenthal's 0-1 law can be extended to other processes, beyond Brownian Motion, including Lévy Processes (with independent stationary increments). Thus, one could think of obtaining a version of Proposition 1 valid for such cases.

The situation considered in Theorem 4 would be the functional counterpart of having a finite-dimensional problem where the supports of both classes (0 and 1) are linearly separable. However, as we have just seen, this separability issue does not only appear in degenerate problems in the functional setting. In the next section, we suggest a technique to completely avoid the problem.

From a theoretical perspective, in view of Theorem 4, it is clear that there is no hope of obtaining a general convergence result of the standard maximum likelihood estimator (MLE) defined by the maximization of the likelihood function $L_n(\beta, \beta_0)$. That is, one should define a different estimator or impose some conditions on the process X to avoid the SC property. For instance, Lindquist and McKeague (2009) prove consistency results under a model of type

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta X(\theta_0),$$

depending on a unique impact point $\theta_0 \in (0, 1)$, which must be estimated, where $X(t + \theta_0) - X(\theta_0)$ is a standard two-sided Brownian Motion.

5.2 Asymptotic non-existence for Gaussian processes

In the previous section, we have seen that the problem of non-existence of the MLE is aggravated for the case functional data. But this is not the only issue with MLE in functional logistic regression. In this section, we see that the probability that the MLE does not exist goes to one as the sample size increases, for any Gaussian process satisfying very mild assumptions.

We use the following notation: for $T = \{t_1, \ldots, t_p\} \subset [0, 1]$ and $f \in L^2[0, 1]$, let $f(T) := (f(t_1), \ldots, f(t_p))'$ and let Σ_T be the $p \times p$ matrix whose (i, j) entry is $K(t_i, t_j)$.

Theorem 5 *Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample of independent observations satisfying model (5). Assume that X is a Gaussian process such that K is continuous and Σ_T is invertible for any finite set $T \subset (0, 1)$. It holds*

$$\lim_{n\to\infty} \mathbb{P}(\textit{MLE exists}) = 0.$$

Proof Let $\beta^* \in \mathcal{H}_K, \beta_0^*$ be the true values of the parameters. Since $\|\beta^*\|_K < \infty$, we have $h(\beta_0^*, \|\beta^*\|_K) < \infty$, where h is the function defined in Candès and Sur (2020), Eq. (2.2) (see Remark 3 below). Let p_n be an increasing sequence of natural

numbers such that $\lim_{n\to\infty} p_n/n = \kappa > h(\beta_0^*, \|\beta^*\|_K)$. Consider the set of equispaced points $0 < t_1 < t_2 < \cdots < t_{p_n} < 1$ and denote $T_n = \{t_1, \ldots, t_{p_n}\}$. Define $\alpha_{T_n} = \Sigma_{T_n}^{-1}\beta^*(T_n)$. Now, consider the following sequence of finite-dimensional logistic regression models

$$\mathbb{P}(Y = 1 \mid X) = \frac{1}{1 + \exp\left\{-\beta_0^* - \alpha'_{T_n} X(T_n)\right\}},$$

and the following sequence of events

$$E_n = \{\text{There exists } \alpha \in \mathbb{R}^{p_n} : \alpha' x_i(T_n) \geq 0, \text{ if } y_i = 1; \ \ \alpha' x_i(T_n) \leq 0, \text{ if } y_i = 0\}.$$

Recall that the event E_n amounts to non-existence of MLE for finite-dimensional logistic regression models (see Albert and Anderson (1984)).

Now let us prove the validity of condition (1.3) in Candès and Sur (2020), which is required for the validity of Theorem 2.1. in that paper. In our case, such condition amounts to

$$\lim_{n\to\infty} \mathrm{Var}\big(\alpha'_{T_n} X(T_n)\big) = \lim_{n\to\infty} \alpha'_{T_n} \Sigma_{T_n} \alpha_{T_n} = \|\beta^*\|_K^2,$$

but this directly follows from Theorem 6E of Parzen (1959). Since $\lim_{n\to\infty} p_n/n = \kappa > h(\beta_0^*, \|\beta^*\|_K^2)$, we apply Theorem 2.1. in Candès and Sur (2020) to get $\lim_n \mathbb{P}(E_n) = 1$.

Now we define the auxiliary sequence of events

$$\widetilde{E}_n = \{\text{There exists } \alpha \in \mathbb{R}^{p_n} : \alpha' x_i(T_n) > 0, \text{ if } y_i = 1; \ \ \alpha' x_i(T_n) < 0, \text{ if } y_i = 0\},$$

with strict inequalities. Assume that $\widetilde{E}_n$ happens so that there exists a separating hyperplane defined by $\alpha \in \mathbb{R}^{p_n}$. Then, in the same spirit as in the proof of Theorem 4, it is possible to show that if $\hat{\beta}_{m,n} = m \sum_{j=1}^{p_n} \alpha_j K(\cdot, t_j) \in \mathcal{H}_K$, then $\lim_{m\to\infty} L_n(\hat{\beta}_{m,n}, 0) = 0$, where $L_n(\beta, \beta_0)$ is the log-likelihood function. As a consequence, for all n, if $\widetilde{E}_n$ happens, then the MLE for the RKHS functional logistic regression model does not exist. The result follows from the fact that $\mathbb{P}(E_n) = \mathbb{P}(\widetilde{E}_n)$ and the events $\alpha' x_i(T_n) = 0$ have probability zero since we are assuming that the process does not have degenerate marginals. □

Remark 3 Theorem 2.1. in Candès and Sur (2020) is a remarkable result. It applies to logistic finite-dimensional regression models with a number p of covariables, which is assumed to grow to infinity with the sample size n, in such a way that $p/n \to \kappa$. Of course, the sample is given by data (x_i, y_i), $i = 1, \ldots, n$. Essentially, the result establishes that there is a critical value such that, if κ is smaller than such critical value, one has $\lim_{n,p\to\infty} \mathbb{P}(\text{MLE exists}) = 1$; otherwise, we have $\lim_{n,p\to\infty} \mathbb{P}(\text{MLE exists}) = 0$. Such critical value is given in terms of a function h (which is mentioned in the proof of the previous result) whose definition is as follows. Let us use the notation $(\widetilde{Y}, V) \sim F_{\beta_0,\gamma_0}$ whenever $(\widetilde{Y}, V) \stackrel{d}{=} (\widetilde{Y}, \widetilde{Y}X)$, for

$\widetilde{Y} = 2Y - 1$ (note that, in the notation of Candès and Sur (2020), the model is defined for the case that the response variable is coded in $\{-1, 1\}$), $\beta_0, \gamma_0 \in \mathbb{R}$, $\gamma_0 \geq 0$ and where $X \sim \mathcal{N}(0, 1)$ and $\mathbb{P}(\widetilde{Y} = 1|X) = (1 + \exp\{-\beta_0 - \gamma_0 X\})^{-1}$. Now, define $h(\beta_0, \gamma_0) = \min_{t_0, t_1 \in \mathbb{R}} \mathbb{E}[(t_0 \widetilde{Y} + t_1 V - Z)_+^2]$, where $Z \sim \mathcal{N}(0, 1)$ independent of $(\widetilde{Y}, V)$ and $x_+ = \max\{x, 0\}$. Then, Theorem 2.1. in Candès and Sur (2020) proves that the above-mentioned critical value for κ is precisely $h(\beta_0, \gamma_0)$.

6 The estimation of β in practice

The problem of non-existence of the MLE can be circumvented if the goal is variable selection. The main idea behind the proof of Theorem 5 is that one can approximate the functional model with finite approximations as those in (9) with p increasing as fast as desired. Therefore, if we constrain p to be less than a finite fixed value, Theorem 5 does not apply.

In order to sort out the non-existence problem for a given sample (due to the SC property), it would be enough to use a finite-dimensional estimator that is always defined, even for linearly separable samples. As mentioned, an extensive study of existence and uniqueness conditions of the MLE for multiple logistic regression can be found in the paper of Albert and Anderson (1984).

A simple, RKHS-motivated alternative would be as follows. In many cases, one could assume that the "true parameter" (β^*, β_0^*) belongs to a bounded set $B_K(0, R) \times I$, I being a compact interval in the real line and $B_K(0, R)$ the closed ball centered at zero, with radius R in the RKHS associated with the covariance function K. This restriction of searching for an estimator in a ball within the parameter space resembles other regularization methods in regression such as ridge or lasso.

If K is continuous and bounded, all functions f in the RKHS space are continuous as well and, using the reproducing property $\langle f, K(\cdot, t)\rangle_K = f(t)$, we get

$$\|f\|_\infty = \sup_t |\langle f, K(\cdot, t)\rangle_K| \leq \|f\|_K \sup_t K(t, t).$$

If, for simplicity, we assume that $\sup_t K(t, t) = 1$, we have (from the definition of the RKHS $\mathcal{H}(K)$) that all functions $\beta \in B_K(0, R)$ can be approximated by functions of type

$$g(\cdot) = \sum_{j=1}^{p} \beta_j K(t_j, \cdot),$$

where β_j are real numbers with $|\beta_j| \leq R$, $p \in \mathbb{N}$, $t_j \in [0, 1]$.

Now, recall that the RKHS functional logistic model corresponding to such function g would be given by expression (9) in terms of β_i and $X(t_i)$. Then, assuming the continuity of the trajectories $X(t)$, we can ensure the existence of an approximate maximum likelihood (ML) estimator of (β^*, β_0^*) expressed in terms of $(\beta_0, \beta_1, \ldots, \beta_d, t_1, \ldots, t_p)$.

The effective calculation of such estimator could be done by a sequential "greedy" method. The idea is to exchange the direct maximization of the likelihood function by the execution of an iterative algorithm, as follows:

1. Let us fix a grid T_p of p equispaced points in $[0, 1]$. For each t on the grid, we fit the logistic model of Eq. (9) with $p = 1$ and $\hat{m}(t) = \bar{X}(t)$. The log-likelihood achieved for this t at the ML estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is stored in $\ell_1(t)$. Then, the first point $\widehat{t_1}$ is fixed as the point at which $\ell_1(t)$ achieves its maximum value.
2. Once $\widehat{t_1}$ has been selected, for each t in the grid, we fit the model

$$\mathbb{P}(Y = 1|X) = \left(1 + \exp\left\{\beta_0 + \beta_1[X(\widehat{t_1}) - \bar{X}(\hat{t_1})] + \beta_2[X(t) - \bar{X}(t)]\right\}\right)^{-1}.$$

   As in the previous step, $\ell_2(t)$ would be the log-likelihood achieved at $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$, and $\hat{t}_2$ is the point at which the maximum of $\ell_2(t)$ is attained.
3. We proceed in the same way until a suitable number of points p has been selected.

In practical problems, it is also important to determine how many points p one should retain. The common approach is to fix this value $\widehat{p}$ by cross-validation, whenever it is possible. Another reasonable approach is to increase the initial value p by repeating the whole procedure with another grid T_{p+1} of $p + 1$ equispaced points until the increase achieved in the likelihood function is smaller than a given threshold, in a similar way as in Berrendero et al. (2019).

7 Some experiments on binary classification

We focus on binary classification, a major application of logistic regression models. This empirical study comprises a group of RKHS-based methods, including the one presented previously, as well as some L^2-based proposals. Firstly, a set of simulated examples is used, in order to check the performance of the different proposals under controlled conditions. Then, a few real data sets are considered as well, for a more complete assessment. The R-codes are available from the authors.

In all cases, the classification performance, for a sample $(y_1, x_1), \ldots, (y_n, x_n)$ with $y_i \in \{0, 1\}$, is measured in terms of the misclassification rate

$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i|,$$

where $\widehat{y}_i$, $i = 1, \ldots, n$, are the predicted labels.

7.1 Methods

Below we include a brief introduction to the methods selected for the study. The corresponding nicknames are shown in boldface.

RKHS-based methods

Besides the method proposed in Sect. 6, denoted **RKHS-sq**, we will consider two further methods, also inspired (in different ways) on RKHS ideas which has been recently proposed in the literature. To be more specific, we will check the following methods:

- **RKHS-sq**: The sequential approach to the RKHS-based functional logistic model (5) presented in Sect. 6. As for the finite-dimensional estimator suitable for linearly separable samples, we use *Firth's estimator* (Firth (1993)) via the R function "brglm" of the package `brglm` (Kosmidis (2017)). The number of points included in the model is selected by fivefold cross-validation.
- **RK**: Linear Discriminant Analysis applied to a set of optimally selected variables $X(t_1), \ldots, X(t_p)$. This method has been proposed in Berrendero et al. (2018, Sec. 5.1). The idea is to select $(t_1, \ldots, t_p)$ in order to minimize a plug-in estimator of the explicit expression of the misclassification error (Izenman 2008, p. 244) available for a p-dimensional Gaussian binary homoscedastic discrimination problem based on $(X(t_1), \ldots, X(t_p))$. We use our own R translation of the original MATLAB code.
- **Mah** - The Mahalanobis-type classifier described in Berrendero et al. (2020), where the smoothing parameter α is selected by fivefold cross-validation.

For RKHS-sq and RK methods, the maximum number of selected points is limited to ten, which in practice does not turn out to be a serious limitation for the method's practical performance.

L^2-based methods

These methods are two different approaches to the standard functional logistic regression model (1).

- **Wav**: A functional adaptation of the method proposed in Zhao et al. (2012), presented in Mousavi and Sørensen (2017). It uses a wavelet representation of the curves, with the number of basis elements shrunk with LASSO. We use the modified least asymmetric version of Daubechies wavelets (as suggested in Mousavi and Sørensen (2018)), via the R function "hardThresholding" of `RFgroove` package (Gregorutti (2016)). The detail level of the basis is fixed by fivefold cross-validation.
- **PCA**: The functional logistic model considered as a particular case of the "generalized linear model" in Müller and Stadtmüller (2005). A finite logistic model is applied to the coefficients of the curves representation in the base of functional principal components of the process (see also Escabias et al. (2004)). The functional principal components are obtained via the R function "fdata2pc" of `fda.usc` package (Febrero-Bande and de la Fuente (2012)). The number of coefficients retained is fixed by fivefold cross-validation, from a maximum of 30 basis elements.

The knn benchmark

As a benchmark for the previous methods, we use functional k-nearest neighbors with $k = 5$ (**knn5**), through the function "classif.knn" of `fda.usc` R package. This is a good reference, as a simple easy-to-implement classifier whose performance is often good in functional examples.

7.2 Simulated scenarios

Hereunder, we present the data sets selected for the more theoretical half of the study, aiming at presenting a miscellaneous selection of non-trivial problems. We distinguish two ways of constructing the data sets. Samples can be seen in Fig. 1.

Choosing the conditional distributions $X|Y = i$ and the marginal of Y
Inspired by Theorem 3, these examples follow the Gaussian setting, with P_0 a standard centered Brownian motion and P_1 a standard Brownian motion plus a trend.

- **Bm fin**: Using as the trend (mean function) for class 1, a finite linear combination of the covariance function of the standard Brownian motion, $K(s,t) = \min(s,t)$. In particular, we take $m_1(s) = 2\min(0.2, s) - 3\min(0.5, s) + \min(0.7, s)$.
- **Bm log**: Using $m_1(s) = \log(s+1)$, which also belongs to the RKHS of the Brownian motion but has not a finite representation as in the previous case. Let us recall in this connection that the RKHS associated to the Brownian covariance is the so-called Dirichlet space of absolute continuous functions $F : [0,1] \to \mathbb{R}$, with $F(0) = 0$ whose derivative is in $L^2[0,1]$; see (Mörters and Peres 2010, p. 24).

Choosing the conditional distributions $Y|X = x$ and the marginal of X
We assign the class Y to a trajectory X by applying the RKHS logistic model. Note that, when the slope function β follows a finite representation as in Eq. (8), it is not necessary to know the explicit expression of K, since the finite-dimensional logistic model (9) is recovered. The following choices are considered for the distribution of X.

- **fBm**: Fractional Brownian Motion with Hurst's exponent $H = 0.9$. The responses Y are drawn from a Bernoulli random variable whose parameter is given by the functional logistic regression model presented in Theorem 3. The intercept β_0 is equal to zero and the slope function used is $\beta(s) = 2K(0.2, s) - 4K(0.5, s) - K(0.7, s)$, K being the covariance function of X. In this case, we recover the finite-dimensional model (9) with $(\beta_1, \beta_2, \beta_3) = (2, -4, -1)$ and $(t_1, t_2, t_3) = (0.2, 0.5, 0.7)$.
- **Mixt**: A mixture of a standard centered Brownian motion $B(s)$ and another independent Brownian motion $\sqrt{2}B'(s)$, being both distributions equiprobable. The response Y is generated as in the previous case using the same points t_j but with coefficients $(\beta_1, \beta_2, \beta_3) = (2, -3, 1)$.

When the slope function does not have such finite representation, one needs to know, or to approximate, the explicit expression of the inverse of the Loève's isometry.

- **Bm sin** - Covariates X are drawn from a standard centered Brownian motion. In this case, the inverse of the Loève's isometry matches Itô's stochastic integral $\int_0^1 \beta'(s)\mathrm{d}X(s)$ for $\beta \in \mathcal{H}(K)$ (see Janson 1997, Example 8.19, p. 122). The functions in this $\mathcal{H}(K)$ are a.s. derivable with respect to Lebesgue measure. Then, the responses Y are realizations of a Bernoulli variable with parameter given by Eq. (5) with slope function $\beta(s) = \sin(\pi s)$.

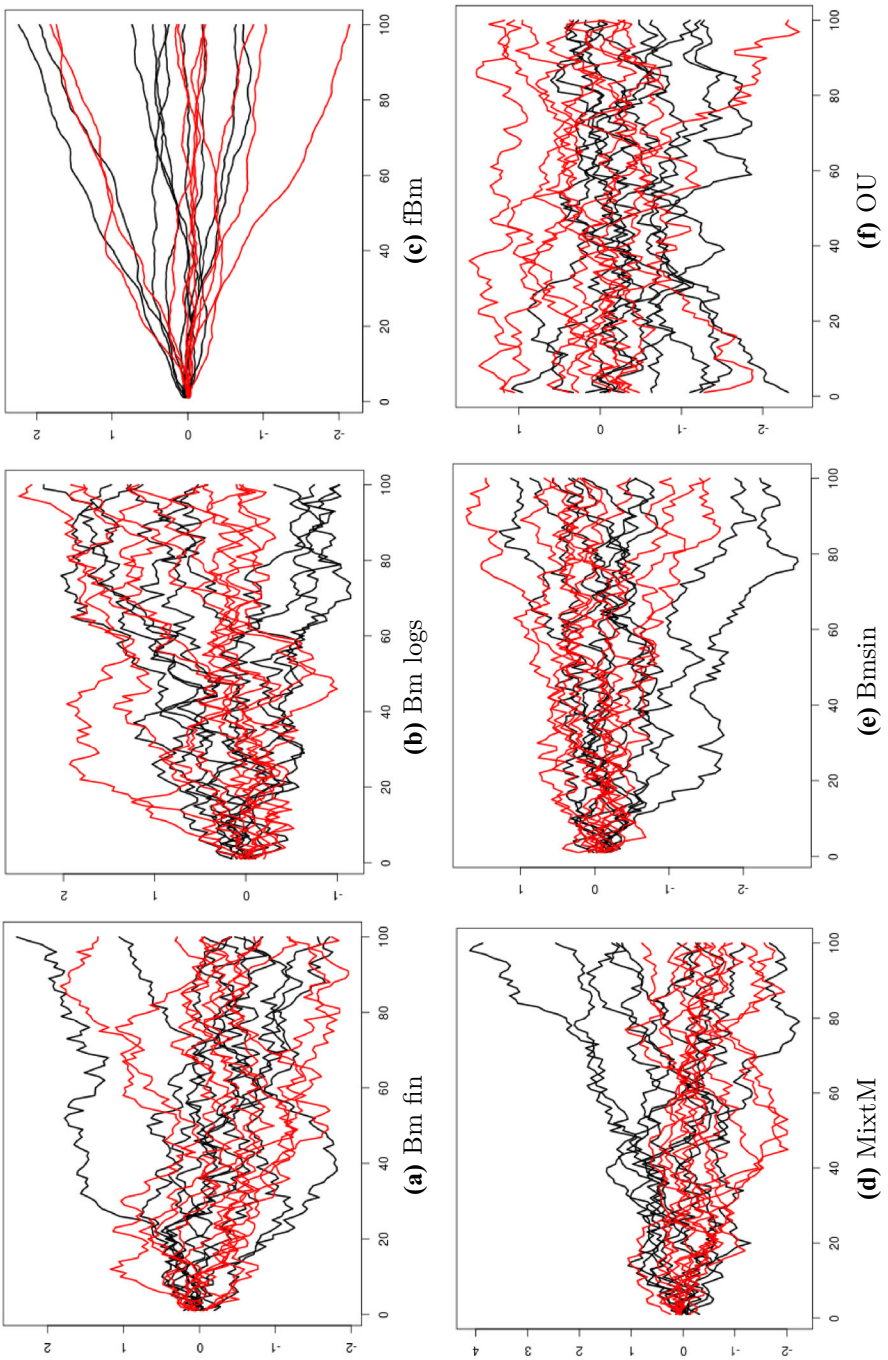

Fig. 1 Simulated data sets, 10 trajectories of each class

Table 1 Misclassification rates for simulated data sets

	Bm fin	Bm Logs	fBm
RKHS-sq	**0.319** (0.070)	0.398 (0.076)	0.214 (0.057)
RK	**0.313** (0.069)	0.391 (0.071)	**0.197** (0.047)
Mah	0.353 (0.075)	**0.383** (0.075)	0.203 (0.051)
Wav	0.338 (0.067)	0.393 (0.076)	**0.200** (0.050)
PCA	0.335 (0.073)	**0.379** (0.075)	0.205 (0.051)
knn5	0.383 (0.074)	0.417 (0.069)	0.221 (0.061)
	MixtSd	Bm sin	OU
RKHS-sq	**0.312** (0.062)	**0.235** (0.065)	**0.232** (0.055)
RK	**0.309** (0.065)	0.236 (0.068)	**0.234** (0.057)
Mah	0.357 (0.080)	0.250 (0.060)	0.276 (0.074)
Wav	0.338 (0.072)	0.246 (0.069)	0.247 (0.057)
PCA	0.314 (0.070)	**0.228** (0.062)	0.237 (0.055)
knn5	0.387 (0.071)	0.290 (0.066)	0.323 (0.060)

- **OU** - Curves are generated from a long-term (stationary) Ornstein-Uhlenbeck process, constructed as in Example 6.2 of Bosq (2000). The inverse of Loève's isometry is approximated as $\Psi_X^{-1}(\beta) \simeq \beta(S)'\Sigma_S^{-1}X(S)$, where $S = \{s_1, \ldots, s_m\}$ is an equispaced grid in $[0, 1]$ and $\beta(S)' = (\beta(s_1), \ldots, \beta(s_m))$, with $\beta(s) = \sin(\pi s)$. Equivalently for $X(S)$. By Theorem 6D of Parzen (1959) (and Theorem 6E for the convergence of the norms), we know that this expression converges to $\Psi_X^{-1}(\beta)$ when the number of points in the grid increases.

Simulation outputs

The estimated misclassification rates given in Table 1 are based on training samples of size 200 and test samples of size 50, with $\mathbb{P}(Y = 0) = 1/2$. The standard deviation of the error rates is in brackets. The two best methods are written in boldface.

The outputs in this table are based on 100 replications of each experiment.

The sequential RKHS proposal and RK method tend to be the best performing ones. It is worth mentioning that, while RK is in fact optimal for the first example, the RKHS-sq proposal obtains a similar error without using any Gaussianity assumption.

7.3 Real data sets

The different proposals are also tested with four real data sets, commonly used in the literature of functional classification and all freely available.

- **Phoneme** - Log-periodogram curves of the pronunciation of phonemes AA (695 samples) and AO (1022 samples). A total of 150 frequencies are kept per recording, sampled over a grid of 256 points. The complete data set can be found along with the online material of Ferraty and Vieu (2006).
- **MCO** - Mitochondrial calcium overload of mouse cardiac cells measured for two groups, control and treatment, of 45 samples each. Measurements are taken every

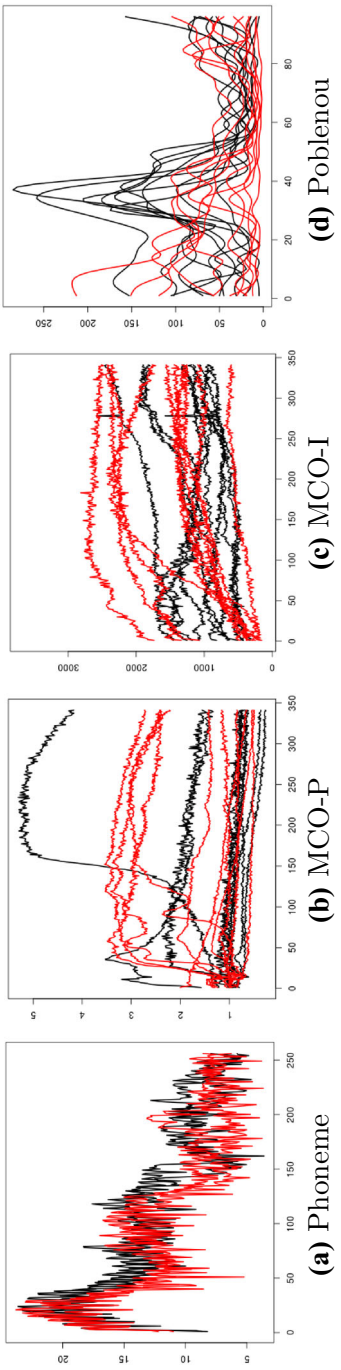

Fig. 2 Real data sets, at most 10 trajectories of each class

Table 2 Misclassification rates for real data sets

	Phoneme	MCO-P	MCP-I	Poblenou
RKHS-sq	**0.181** (0.022)	0.256 (0.150)	0.170 (0.108)	0.113 (0.050)
RK	0.187 (0.028)	0.322 (0.133)	**0.091** (0.066)	**0.052** (0.036)
Mah	0.209 (0.013)	0.389 (0.056)	0.112 (0.056)	0.252 (0.048)
Wav	**0.181** (0.027)	**0.233** (0.046)	**0.068** (0.048)	**0.096** (0.036)
PCA	0.248 (0.043)	0.411 (0.128)	0.180 (0.174)	0.174 (0.097)
knn5	0.233 (0.031)	**0.244** (0.084)	0.190 (0.133)	0.113 (0.073)

10 seconds during an hour, excluding the first three minutes. The experiment was done using the intact cells (MCO-I) and "permeabilized" cells (MCO-P). It appeared originally in Ruiz-Meana et al. (2003) and is available in `fda.usc` package.

- **Poblenou** - Daily nitrogen oxide, NO_x, measurements in Poblenou (Barcelona, Spain). We sub-sample the original hourly records to obtain measures every 15 minutes by representing the curves in a Bspline base of 50 elements and evaluating them in a thinner grid. Classification groups include weekends and bank holidays (39 samples) versus working days (76 samples).

A sample of these data sets is presented in Fig. 2. Original curves are used since under-smoothing is, in general, desirable for functional classification problems according to Carroll et al. (2013). In order to better approximate the misclassification rate, we use fivefold cross-validation. The resulting mean rates and their standard deviations (in brackets) can be found in Table 2 where, again, the boldfaced entries correspond to the best results.

The wavelets L^2-based model seems to outperform the others in general. The RKHS-sq proposal is competitive and is among the best options for the phoneme set, which is the largest one. Regarding variable selection, let us note that this method selects 8.5 points on average for these data sets.

As an overall, tentative conclusion of our experiments, we could say that the method RKHS-sq, compatible with our RKHS-based model, seems to be competitive. Then, the gain of interpretability associated with the use of finite-dimensional marginals does not seem to come at a high cost in efficiency. We do not see an obvious explanation for the relative good behavior of the **Wav** in the real data examples; it might be associated with a particular flexibility of this method against non-structured data. In any case, of course, more detailed experiments should be done to get a broader perspective.

8 Some concluding remarks

Our results suggest that there is a case for considering alternative formulations to the more popular L^2-based model for the problem of functional logistic regression.

In particular, an RKHS-based formulation allows us to encompass all the finite-dimensional models based on marginals $(X(t_1), \ldots, X(t_p))$ from the explanatory

process $\{X(t),\ t \in [0,1]\}$ as particular cases. This provides a unified framework to consider all this models when facing, for example, the problem of optimal variable selection.

The RKHS formulation seems to be flexible and general enough. Still, the estimation issues remain challenging. Different adapted versions of the maximum likelihood paradigm, as well as Bayesian methodologies, could be considered here.

While the empirical results shown in Sect. 7 are just preliminary, they are generally encouraging on the possibility of using RKHS-based approaches in functional logistic regression.

Acknowledgements This work has been partially supported by Spanish Grant PID2019-109387GB-I00. The constructive comments from an Associate Editor and two reviewers are gratefully acknowledged.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Appendix

In order to make reading easier, we review here some of the results we have used in our proofs.

In this section, P_m denotes the probability measure induced on the space of sample paths by a separable Gaussian stochastic process $\{X(t) : t \in [0,1]\}$ with continuous covariance kernel K and mean function m, and P_0 stands for the probability measure corresponding to the Gaussian process with covariance kernel K and mean function $m \equiv 0$. Moreover, $\mathcal{K}$ denotes the covariance operator defined by K.

The following theorem, that can be found in Parzen (1961), p. 979, provides an expression for the Radon-Nikodym of probability measures induced by homoscedastic Gaussian processes:

Theorem (Parzen, 1961) *The measures P_m and P_0 are mutually absolutely continuous or orthogonal depending on whether m does or does not belong to $\mathcal{H}(K)$. If $m \in \mathcal{H}(K)$, then the Radon-Nikodym density of P_m with respect to P_0 is given by*

$$\frac{\mathrm{d}P_m}{\mathrm{d}P_0}(X) = \exp\left\{\langle X, m\rangle_K - \frac{1}{2}\|m\|_K^2\right\}.$$

In Parzen's result, one of the measures must correspond to a zero mean process. To remove this condition, Lemma 1.1 in Pitcher (1960) can be applied:

Lemma (Pitcher, 1960) *The measures P_{m_1} and P_{m_0} are orthogonal if and only if $P_{m_1-m_0}$ and P_0 are. The measure P_{m_1} is absolutely continuous with respect to P_{m_0} if*

and only if $P_{m_1-m_0}$ is absolutely continuous with respect to P_0, and in this case

$$\frac{\mathrm{d}P_{m_1}}{\mathrm{d}P_{m_0}}(X) = \frac{\mathrm{d}P_{m_1-m_0}}{\mathrm{d}P_0}(X - m_0).$$

Theorem 6.1, p. 317, in Rao and Varadarajan (1963) provides an alternative expression for the Radon-Nikodym density of P_{m_1} with respect to P_{m_0} under the more restrictive condition $m_1 - m_0 \in \mathcal{K}(L^2)$. Moreover, the density can be expressed in terms of a continuous linear functional on $L^2[0, 1]$ if and only if $m_1 - m_0 \in \mathcal{K}(L^2)$:

Theorem (Rao and Varadarajan, 1963) *Suppose that P_{m_1} and P_{m_0} are mutually absolutely continuous. Then, there exists a continuous linear functional $T : L^2[0, 1] \to \mathbb{R}$ and a constant c such that*

$$\frac{\mathrm{d}P_{m_1}}{\mathrm{d}P_{m_0}}(X) = \exp\{T(X) + c\}$$

if and only if $m_1 - m_0 \in \mathcal{K}(L^2)$. In that case,

$$\frac{\mathrm{d}P_{m_1}}{\mathrm{d}P_{m_0}}(X) = \exp\left\{\langle X - m_0, \mathcal{K}^{-1}(m_1 - m_0)\rangle_2 - \frac{1}{2}\langle m_1 - m_0, \mathcal{K}^{-1}(m_1 - m_0)\rangle_2\right\}.$$

It can be checked that the expressions of the Radon-Nikodym densities given by Parzen (1961) and Rao and Varadarajan (1963) coincide when the condition $m_1 - m_0 \in \mathcal{K}(L^2)$ holds. When $m_1 - m_0 \in \mathcal{H}(\mathcal{K}) = \mathcal{K}^{1/2}(L^2)$ but $m_1 - m_0 \notin \mathcal{K}(L^2)$, the use of Loéve's isometry to write the density is unavoidable.

References

Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1):1–10

Ash RB, Gardner MF (2014) Topics in stochastic processes. Academic Press, Cambridge

Baíllo A, Cuevas A, Cuesta-Albertos JA (2011) Supervised classification for a family of Gaussian functional models. Scand J Stat 38(3):480–498

Berlinet A, Thomas-Agnan C (2004) Reproducing kernel Hilbert spaces in probability and statistics. Kluwer Academic, Boston

Berrendero JR, Bueno-Larraz B, Cuevas A (2019) An RKHS model for variable selection in functional linear regression. J Multivar Anal 170:25–45

Berrendero JR, Bueno-Larraz B, Cuevas A (2020) On Mahalanobis distance in functional settings. J Mach Learn Res 21:1–33

Berrendero JR, Cuevas A, Torrecilla JL (2018) On the use of reproducing kernel Hilbert spaces in functional classification. J Am Stat Assoc 113:1210–1218

Bosq D (2000) Linear processes in function spaces: theory and applications. Springer, New York

Candès EJ, Sur P (2020) The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. Ann Stat 48(1):27–42

Carroll RJ, Delaigle A, Hall P (2013) Unexpected properties of bandwidth choice when smoothing discrete data for constructing a functional data classifier. Ann Stat 41(6):2739–2767

Cramer JS (2003) Logit models from economics and other fields. Cambridge University Press, Cambridge

Efron B (1975) The efficiency of logistic regression compared to normal discriminant analysis. J Am Stat Assoc 70(352):892–898

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
Fahrmeir L, Kaufmann H (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann Stat 13(1):342–368
Fahrmeir L, Kaufmann H (1986) Asymptotic inference in discrete response models. Stat Hefte 27(1):179–205
Febrero-Bande M, de la Fuente M (2012) Statistical computing in functional data analysis: The r package fda.usc. J Stat Softw 51(4):1–28
Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer, Berlin
Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80(1):27–38
Gourieroux C, Monfort A (1981) Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. J Econom 17(1):83–97
Gregorutti B (2016) RFgroove: Importance Measure and Selection for Groups of Variables with Random Forests. R Package 1:1
Grenander U (1981) Abstract inference. Wiley, Hoboken
Hilbe JM (2009) Logistic regression models. CRC Press, Boca Raton
Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, 3rd edn. Wiley, Hoboken
Hsing T, Eubank R (2015) Theoretical foundations of functional data analysis, with an introduction to linear operators. Wiley, Hoboken
Hsing T, Ren H (2009) An RKHS formulation of the inverse regression dimension-reduction problem. Ann Stat 37(2):726–755
Izenman AJ (2008) Modern multivariate statistical techniques: regression, classification and manifold learning. Springer, Berlin
James GM (2002) Generalized linear models with functional predictors. J R Stat Soc Ser B (Stat Methodol) 64(3):411–432
Janson S (1997) Gaussian Hilbert spaces. Cambridge University Press, Cambridge
Kneip A, Liebl D (2020) On the optimal reconstruction of partially observed functional data. Ann Stat 48(3):1692–1717
Kosmidis I (2017) brglm: Bias reduction in binary-response generalized linear models. R package, v. 0.6.1
Lindquist MA, McKeague IW (2009) Logistic regression with Brownian-like predictors. J Am Stat Assoc 104(488):1575–1585
Lukić MN, Beder JH (2001) Stochastic processes with sample paths in reproducing kernel Hilbert spaces. Trans Am Math Soc 353(10):3945–3969
Mörters P, Peres Y (2010) Brownian motion. Cambridge University Press, Cambridge
Mousavi SN, Sørensen H (2017) Multinomial functional regression with wavelets and LASSO penalization. Econom Stat 1:150–166
Mousavi SN, Sørensen H (2018) Functional logistic regression: a comparison of three methods. J Stat Comput Simul 88(2):250–268
Müller H-G, Stadtmüller U (2005) Generalized functional linear models. Ann Stat 33(2):774–805
Munsiwamy B, Wakweya ST (2011) Asymptotic properties of estimates for the parameters in the logistic regression model. Asian-Afr J Econ Econom 11(1):165–174
Parzen E (1959) Statistical inference on time series by Hilbert space methods, I. Technical Report 23, Stanford University
Parzen E (1961) An approach to time series analysis. Ann Math Stat 32(4):951–989
Petrovich J, Reimherr M, Daymont C (2018) Highly irregular functional generalized linear regression with electronic health records. arXiv preprint arXiv:1805.08518
Pillai NS, Wu Q, Liang F, Mukherjee S, Wolpert RL (2007) Characterizing the function space for Bayesian kernel models. J Mach Learn Res 8:1769–1797
Pitcher TS (1960) Likelihood ratios of Gaussian processes. Ark Mat 4:35–44
Rao CR, Varadarajan V (1963) Discrimination of Gaussian processes. Sankhyā: Indian J Stat, Ser A 25(3):303–330
Ruiz-Meana M, Garcia-Dorado D, Pina P, Inserte J, Agulló L, Soler-Soler J (2003) Cariporide preserves mitochondrial proton gradient and delays ATP depletion in cardiomyocytes during ischemic conditions. Am J Physiol-Heart Circ Physiol 285(3):H999–H1006
Shin H, Hsing T (2012) Linear prediction in functional data analysis. Stoch Process Iheir Appl 122(11):3680–3700

Silvapulle MJ (1981) On the existence of maximum likelihood estimators for the binomial response models. J R Stat Soc B 43(3):310–313
Zhao Y, Ogden RT, Reiss PT (2012) Wavelet-based LASSO in functional linear regression. J Comput Graph Stat 21(3):600–617