Generalized Functional Linear Models under Choice-Based Sampling

Sophie Dabo-Niang, Mohamed-Salem Ahmed

Generalized Functional Linear Models under Choice-Based Sampling

Mohamed–Salem Ahmed
Univ. Lille, UMR 9221 - LEM, F-59000 Lille, France
mohamed-salem.ahmed@etu.univ-lille3.fr

Sophie Dabo Niang
Univ. Lille, UMR 9221 - LEM, F-59000 Lille, France
Also INRIA-MODAL Team
sophie.dabo@univ-lille3.fr

Abstract

We propose here to describe and estimate a functional binary model in a context of sampling data. This problem is known respectively in econometric and epidemiology literatures, as Choice-Based Sampling and case-control study, in discrete choice model. Unlike the random sample where all items in the population have the same probability of being chosen, the Choice-Based Sampling (CBS) in discrete choice model is a type of sampling where the classification of the population into subsets to be sampled is based on the choices or outcomes. In practice, it could be of interest to model choice of individuals using some functional covariates instead of real valued random variables. To this end, this paper introduces the Choice-Based sampling in a functional framework (functional generalized linear models). We adapt the approach of [31] to reduce the infinite dimensional of the space of the explanatory random function using a Karhunen–Loève expansion and maximize a Conditional likelihood function. Our method is based on the components of a Functional Principal Components Analysis adapted to the context of Choice-Based Sampling. Then this expansion is truncated to a finite number of terms that asymptotically increases with the sample size. Asymptotic properties of our estimate are ensured by the help of works of [4], [31]. We present some simulated experiments including genetic data, to investigate the finite sample performance of the estimation method. The proposed functional model leads to encouraging results. The potential of the functional choice-based sampling model to integrate the special non-random features of the sample, that would have been hard to see otherwise is also outlined.

Key words : Generalized linear model, Functional data analysis, Choice-based sampling, Case-control.

1 Introduction

The Case-control or Case-Referent or Retrospective Sampling study is one of the most frequently used designs in epidemiology and health research, (see [25]). It provides a powerful and useful method for identifying the impact of several factors on the occurrence of a particular event. The main idea in case-control or retrospective studies is to stratify the population according to the categorical response; we first identify the cases, a comparable control group is then sampled and we look back retrospectively to compare how frequently the exposure to a risk factors is present in each group. It is very important to ensure sampling the controls from the same population where the exposures of cases occurred. In a prospective or cohort study design, by contrast, a sample of individuals is chosen and exposed to some risk factors and then these individuals are followed through time until recording their responses. In addition to the fact that prospective studies have disadvantages in terms of time and cost, they are not good when dealing with rare events. In this case, indeed, even with very large cohort study, just a few cases (individuals with disease) may produce, but, on the other hand, prospective studies have more advantages when studying the effects of rare exposures. For a general overview of case-control studies, see [21] for a discussion in medical research. According to the method used to select controls, there are two basic types of case-control studies: unmatched and matched studies. In unmatched studies, we select a control group randomly from the population for all cases, while in matched one, each case is matched, upon some characteristics, with one or more controls. Matched case-control study design is not taken into account in this paper. One of the most serious problems in retrospective studies is that collecting exposure information rely on memory, also called recall bias. In epidemiology, people with disease are usually more motivated to recall risk factors than people in control group. [34] have described this accuracy of recall and other types of bias like interviewer bias of case-control studies. This should then be taken into account and reduced in the study design. [23] provide a good review of the bias in case-control studies.

A similar retrospective studies in econometric literature is called choice-based sampling, the choice of transportation mode is the most popular example, [27], [26], [10], [19]. Case-control study is also used in political science, [22] and sociology [36]. In many situations, there are cases in which it would be more appropriate to respect the way whose data have been selected, especially if they were collected via an endogenous stratification. The endogenous stratification is a plan of stratified sampling with random sampling in the strata, where these strata are created via the different values of the endogenous variable to which we are interested. In the literature this type of lamination had attracted the attention of the econometricians who study the behavior of choice and epidemiologists who study the origin of rare diseases. *Choice-based Sampling* will be used and studied in this paper in a context of functional data.

Functional data analysis (FDA) was widely popularized by [32]. Since its introduction, much work has been done in the representation, exploration or modelling of functional data. Many classical statistical methods have been adapted or extended to the functional framework such as distribution characterizations, principal component analysis, outlier detection or linear regression analysis. Nonparametric methods have also been developed for functional data and an overview is available in [15]. Moreover, a number of reference textbooks dealing with functional data analysis already exist, such as [2], [32], [12], and [18]. FDA is thus an active research topic with potential applications in a large number of fields.

The purpose of the present paper is to propose a generalized functional linear models (GFLM) adapted to some sampling data. Therefore, the entire available information of the sampling design can be used to obtain a finer estimation and understanding of the phenomenon of interest. Several types of functional linear models have been developed over the years, serving different purposes. When the interest is in the modelling of a functional response according to qualitative covariates, the functional analysis of variance models ([3] and [32]) were developped. When some or all covariates can be thought as continuous, we can use a functional linear model for either a scalar response or a functional response. Moreover, when the response is functional, the influence of a covariate can be either instantaneous, *i.e.* a covariate at time t only influences the response at the same time t [[17], called "point-wise" or "concurrent" model, *e.g.*,], either spread over a time interval of the response which means that the covariate at time t can influence the response at several times s possibly different from t [37].

Among all the functional linear models introduced, the most studied is perhaps the functional linear model for scalar response. This model was originally introduced by [16] in order to resolve the collinearity problem which occurs when one want to use successive measurements of a signal or a time series as covariates in a classical linear model [28]. The fitting of simple functional linear model has been studied theoretically for instance by, *e.g.*, [6]. Furthermore, related models such as the functional linear discriminant analysis are considered in [20]. These methods have already been applied in many fields, *e.g.* image processing, [5], medicine, [33], genetics, [30], ecology [1], marketing, [35]. All those applications show that there is an increasing interest in applying functional linear model and its generalization for practical purposes.

Functional linear models have also been generalized by [31], [8], [13], [14] and, more recently, functional generalized additive models, [29] have also been developed.

To the best of our knowledge, despite many potential applications there is no work on generalized functional linear models that take into account a non-random sampling of the data. However, some works exist (see [4]) on functional principal components analysis adapted to sampling data. [7] have also proposed consistent estimators of mean and variance functions based on Horvitz-Thompson estimator. Note that one work ([14]) on a functional logit model applied to case-control data exists. These

authors proposed a functional logit model to test association between a dichotomous trait and multiple genetic variants in a region using some covariates. However, they do not take into consideration the case-control feature of their data. So their model is like a classical generalized functional linear model in a case of random sampling. They are only interested to test and compare fixed and mixed effect models. They used simulated genetic case-control sequence data to evaluate type I error rates and power of the proposed statistics.
Several authors have proposed consistent methods to estimate the parameters of interest in a *Choice-Based Sampling* model, when the explanatory variables is real-valued (see [26], [27], [10], [11], [19],...).

Our goal is to generalize in a functional framework the Conditional Maximum Likelihood method suggested by [27], to estimate a generalized functional linear model in a context of Choice-Based Sampling. We adapt the approach of [31] to reduce the infinite dimensional of the space of the explanatory random function using a Karhunen–Loève expansion and maximize a Conditional likelihood function. Our method is based on the components of a Functional Principal Components Analysis adapted to the context of Choice-Based Sampling. Then this expansion is truncated to a finite number of terms that asymptotically increases with the sample size. Asymptotic properties of our estimate are ensured by the help of works of [4], [31].

It is worth saying that the expected improvements of using the functional sampling design framework are mostly in terms of performance of the constant parameter estimation rather than on that of the functional parameter. We present a way to improve the accuracy of a traditional generalized regression model applied to sampling genetic data (see [14]). We insist that our sampling generalized functional regression method is relevant since it is intuitively a well suited and theoretically justified method for survey data.

The outline of the present paper is as follows. In Section 2, we introduce the design and the model under a choice-based sampling, and we discuss the usual approach of estimation a generalized functional model in such case. We then give our proposed method of integrating the sampling design in the estimation processus. In section 3, we present asymptotic results whereas Section 4 gives some simulations to illustrate the performance of the proposed estimators. Section 4 is devoted to some conclusions. Finally, the last section gives the proofs of our main results.

2 Conditional Maximum Likelihood Estimator with Functional Predictor

In a given population, we assume that we observe a binary random variable Y taking values in $\{0,1\}$ and a random fonction $\{X(t), t \in \mathcal{T}\}$ which corresponds to a square

integrable stochastic process on the interval $\mathcal{T} \subset \mathbb{R}$. Assume that the process $\{X(t), t \in \mathcal{T}\}$ takes values in some space $\mathcal{X} \subset L^2(\mathcal{T})$, where $L^2(\mathcal{T})$ is the space of square integrable functions in $\mathcal{T}$. We are interested in describing the relation between the response variable Y and the explanatory random function $X(\cdot)$. We assume that this relation is given by a generalized linear regression problem and the expectation of Y given on $X(\cdot)$ is defined by:

$$E\left(Y|X\right) \equiv P\left(Y=1|X,\alpha^*,\theta^*(\cdot)\right) = \Phi\left(\alpha^* + \int_{\mathcal{T}} X(t)\theta^*(t)dt\right) \tag{1}$$

where the link function $\Phi(\cdot)$ is some strictly increasing cumulative distribution function. The parameters of interest are the constant intercept α^*, the parameter function $\theta^*(\cdot)$ assumed to belong the the space of functions $L^2(\mathcal{T})$. Let $Q^* = P\left(Y=1\right)$ be the share of individuals such that $Y=1$ in the considered population. We assume that this population is divided according to the values of the response variable Y into two stratums $\mathcal{J}(0) = \{(0,X),\ X \in \mathcal{X}\}$ and $\mathcal{J}(1) = \{(1,X),\ X \in \mathcal{X}\}$ and let H^* be the probability according to which we drew the stratum $\mathcal{J}(1)$. We assume that we sample in this population as follows: *We take an observation by first drawing a stratum* $i \in \{0,1\}$ *with a probability* $H(i)$ *(*$H(1) = H^*$*) then we draw an observation* $(Y=i,X)$ *at random from* $\mathcal{J}(i)$.
This kind of sampling is known in econometric literature as pure Choice-Based Sampling. Choice-based sampling process allows to take into consideration the structure of the population when one of the values of the response variable Y has a small probability of being observed, compare to the random sample case where all the values have the same probability of being chosen. Under this sampling process, the conditional density of Y given $X=x$ is

$$g(i|x) = \frac{P\left(Y=i|x,\alpha^*,\theta^*(\cdot)\right)H(i)/Q(i)}{\sum_{j=0}^{1} P\left(Y=j|x,\alpha^*,\theta^*(\cdot)\right)H(j)/Q(j)}, \quad x \in \mathcal{X},\ i \in \{0,1\} \tag{2}$$

where $Q(i) = P\left(Y=i\right)$. The expectation with respect to the distribution under the CBS is defined by (see [11])

$$E_s(\cdot) = H(0)E(\cdot|Y=0) + H(1)E(\cdot|Y=1).$$

Note that $E_s(\cdot)$ is different to the expectation $E(\cdot)$ under the population distribution. Our object is to estimate using observations with same law as (Y,X), the intercept parameter α^* and the parameter function $\theta^*(\cdot)$ when the sampling process is that of the CBS defined above and when we assume that we have a prior information allowing knowledge on Q^* and H^*.
Let Γ_s denote the covariance operator of the $\mathcal{X}-$ valued random function under the

CBS:

$$\Gamma_s x(t) = \int_{\mathcal{T}} E_s(X(t)X(v))x(v)dv,\ x \in \mathcal{X},\ t \in \mathcal{T}.$$

The operator Γ_s is a linear integral operator whose integral kernel is $K(t,v) = E_s\left(X(t)X(v)\right),\ t, v \in \mathcal{T}$. It is a compact self-adjoint Hilbert-Schmidt operator because

$$\int |K(t,v)|^2\, dtdv \leq E_s\left(\int X^2(t)dt\right)^2 < \infty,$$

then it can be diagonalized (see [9], page 47).

In order to ensure identification of our model, we need the following assumptions in addition to $E\left(X(t)\right) = 0,\ \forall t \in \mathcal{T}$, concerning the inclusion of the intercept.

(H1) The eigenvalues of Γ_s are nonzero.

(H2) The link function $\Phi(\cdot)$ is monotone, invertible and has two continuous bounded derivatives with $\|\Phi'(\cdot)\| < C$, $\|\Phi''(\cdot)\| < C$ for some constant $C \geq 0$, and there exists $\delta > 0$ such that for all $x(\cdot) \in \mathcal{X}$, $\theta(\cdot) \in L^2(\mathcal{T})$ and $\alpha \in \mathbb{R}$,

$$\left(1 - \Phi\left(\alpha + \int_{\mathcal{T}} x(t)\theta(t)dt\right)\right)\Phi\left(\alpha + \int_{\mathcal{T}} x(t)\theta(t)dt\right) > \delta$$

and $0 < H(1) < 1$.

Assumptions (H1) and (H2) allow us to ensure the identification of our model (see for instance, [8], page 27). Assumption (H2) is similar to assumption $(M1)$ in [31] where it is assumed that the link function is monotone, invertible, has first and second bounded derivatives and that the conditional variance of the response variable is bounded away form 0.

2.1 Infeasible maximum likelihood estimate

We assume that we have a sample of N independent observations
$\left(Y_n = i_n, \{X_n(t),\, t \in \mathcal{T}\}\right),\ n = 1, \ldots, N$ with same law as (Y, X) and drawn through the CBS process. So, by the conditional density (2), the conditional Log likelihood function is defined by:

$$L(\alpha, \theta(\cdot)) = \sum_{n=1}^{N} \log\left(\frac{P\left(Y_n = i_n | X_n, \alpha, \theta(\cdot)\right) H(i_n)/Q(i_n)}{\sum_{j=0}^{1} P\left(Y_n = j | X_n, \alpha, \theta(\cdot)\right) H(j)/Q(j)}\right). \tag{3}$$

When the explanatory variable X is real valued, [27] have maximized (3) to find the maximum likelihood estimation of the intercept α^* and the vector of parameter θ^* in

(2). This method is usually referred to as conditional maximum likelihood estimator. In our functional context, we aim to estimate α^* and the parameter function $\theta^*(\cdot)$ by maximizing (3) on α and $\theta(\cdot)$. But this can not be done before we address the difficulty produced by the infinite dimensionality of the explanatory random function. This could be achieved by one of two very popular approaches used in generalized linear models with explanatory random function. In one hand, we have the Penalized Likelihood Method [8] that consists in projecting the parameter function in a finite-dimensional space spanned by a Spline basis and then maximizing the pseudo conditional log likelihood function obtained by replacing the parameter function $\theta(\cdot)$ in (3) by its projector, adding a penalty that controls the degree of smoothness of the parameter function. On the other hand, we have the second approach used by [31]. It is based on a truncation strategy that consists of projecting the functional explanatory variable and parameter function in a space of functions generated by a basis of functions with a dimension that increases asymptotically as the sample size tends to infinity. We shall adapt the strategy of this second approach in order to resolve infinite dimensionality problem of the functional space in the frame of the CBS. This method will be denoted *Truncated Conditional Likelihood Method*.

2.2 Truncated Conditional Likelihood Method

Analogy to [31], the truncation strategy is motivated by the following considerations. Let $\{\varphi_j,\, j = 1, 2, \ldots\}$ be an orthonormal basis of the functional space $L^2(\mathcal{T})$, usually a Fourier or a Spline basis or a basis constructed by the eigenfunctions of the covariance operator. We can rewrite $X(t)$ and $\theta^*(t)$ in the following way

$$X(t) = \sum_{j\geq 1} \varepsilon_j \varphi_j(t),\ \theta^*(t) = \sum_{j\geq 1} \theta_j^* \varphi_j(t)$$

where the real random variables ε_j and the coefficients θ_j^* are given by

$$\varepsilon_j = \int_{\mathcal{T}} X(t)\varphi_j(t)dt, \qquad \text{and } \theta_j^* = \int_{\mathcal{T}} \theta(t)^* \varphi_j(t)dt.$$

By the orthonormality of the basis $\{\varphi_j,\, j = 1, 2, \ldots\}$, we have

$$\int_{\mathcal{T}} X(t)\theta^*(t)dt = \sum_{j\geq 1} \theta_j^* \varepsilon_j.$$

Let p_N be a positive sequence of integers, increases asymptotically as $N \to \infty$, and consider the following decomposition

$$U_{p_N} = \alpha + \sum_{j=1}^{p_N} \theta_j^* \varepsilon_j, \qquad V_{p_N} = \sum_{j=p_N+1}^{\infty} \theta_j^* \varepsilon_j.$$

Let $F^{(s)}_{V_{p_N}|U_{p_N}}$ denote the conditional distribution of V_{p_N} given U_{p_N} under the CBS. Then we can write

$$E_s\left(Y|X\right) = \mu\left(U_{p_N} + V_{p_N}\right)$$

where

$$\mu(t) = \frac{\Phi(t)H(1)/Q(1)}{\Phi(t)(H(1)/Q(1)) + (1-\Phi(t))H(0)/Q(0)}.$$

We define

$$\begin{aligned}\mu_p\left(U_{p_N}\right) &= E_s\left(Y|U_{p_N}\right) = E_s\left(E_s\left(Y|X\right)|U_{p_N}\right) \\ &= \int \mu\left(U_{p_N} + v\right) dF^{(s)}_{V_{p_N}|U_{p_N}}(v).\end{aligned}$$

By assumption (H2), we have that $\sup_{t\in\mathbb{R}} |\mu^{'}(t)| \leq C$, for some constant C, then we have

$$\begin{aligned}&\left\{\int \left(\mu\left(U_{p_N} + V_{p_N}\right) - \mu\left(U_{p_N} + v\right)\right) dF^{(s)}_{V_{p_N}|U_{p_N}}(v)\right\}^2 \\ &\qquad \leq \sup_t \left|\mu^{'}(t)\right|^2 \int \left(V_{p_N} - v\right)^2 dF^{(s)}_{V_{p_N}|U_{p_N}}(v) \\ &\qquad \leq 2C\left\{V^2_{p_N} + E_s\left(V^2_{p_N}|U_{p_N}\right)\right\}\end{aligned}$$

and therefore

$$E_s\left(\mu(U_{p_N} + V_{p_N}) - \mu_p(U_{p_N})\right)^2 \leq 4CE_s\left(V^2_{p_N}\right). \tag{4}$$

In a similar way, we can show that $E_s\left(\mu(U_{p_N}) - \mu_p(U_{p_N})\right)^2$ is bounded by the term in (4) and then $E_s\left(\mu(U_{p_N} + V_{p_N}) - \mu(U_{p_N})\right)^2$ is bounded by the same bound. The advantage of this approximation is that when we consider the eigen-basis, the approximation error would be controlled by the eigenvalues associated to this eigen-basis, that is

$$E_s\left(V^2_{p_N}\right) = E_s\left(\sum_{j=p_N+1}^{\infty} \theta^*_j \varepsilon_j\right)^2 = \sum_{j=p_N+1}^{\infty} (\theta^*_j)^2 E_s\left(\varepsilon^2_j\right) = \sum_{j=p_N+1}^{\infty} (\theta^*_j)^2 \lambda_j.$$

So this truncation strategy allows us to use the truncated distribution of $Y|X$ that is a Bernoulli of parameter $\mu\left(U_{p_N}\right)$; $\mathcal{B}\left(\mu\left(U_{p_N}\right)\right)$ in which we have p_N parameters to estime instead of the full distribution of $Y|X$ that follows $\mathcal{B}\left(\mu\left(U_{p_N} + V_{p_N}\right)\right)$ in which we have to estimate an infinite number of parameters.

Now the parameters of interest in this truncated model are the intercept α^* and the p_N first coefficients of the parameter function $\theta^*_1, \ldots, \theta^*_{p_N}$. For simplicity, let $\tilde{\theta} =$

$(\theta_0^*, \theta_1^*, \ldots, \theta_{p_N}^*)^T$ with $\alpha^* = \theta_0^*$. Then the truncated Conditional Log Likelihood function is obtained by replacing in (3), $\Phi\left(\alpha^* + \sum_{j\geq 1} \theta_j^* \varepsilon_j\right)$ by $\Phi\left(U_{p_N}\right)$. The corresponding and feasible Conditional Likelihood is

$$\begin{aligned}\tilde{L}_{p_N}(\theta) &= \sum_{n=1}^{N} i_n \log \frac{\Phi(\eta_n) H^*/Q^*}{\Lambda(\eta_n)} \\ &\quad + (1-i_n) \log \frac{\left(1-\Phi(\eta_n)\right)\left(1-H^*\right)/\left(1-Q^*\right)}{\Lambda(\eta_n)},\ \theta \in \mathbb{R}^{p_N+1} \end{aligned} \tag{5}$$

where $\eta_n = \sum_{j=0}^{p_N} \theta_j \varepsilon_j^{(n)}$ with $\varepsilon_j^{(n)} = \int X_n(t)\varphi_j(t)dt,\ \varepsilon_0^{(n)} = 1$ and

$$\Lambda(\eta_n) = \Phi(\eta_n) H^*/Q^* + \left(1-\Phi(\eta_n)\right)\left(1-H^*\right)/\left(1-Q^*\right).$$

Then $\tilde{\theta}$ is estimated by

$$\hat{\theta} = \operatorname{argmax}\left\{\tilde{L}_{p_N}(\theta),\ \theta \in \mathbb{R}^{p_N+1}\right\}$$

So the estimator of the intercept α^* is $\hat{\theta}_0$ and that of the truncated parameter function is given by

$$\hat{\theta}(t) = \sum_{j=1}^{p_N} \hat{\theta}_j \varphi_j(t).$$

We define the $(p_N+1)\times(p_N+1)$ matrix

$$\Delta_{p_N} = \left(E_s\left(\frac{\mu'^2(\eta)}{\sigma^2\left(\mu(\eta)\right)}\varepsilon_k \varepsilon_l\right)\right)_{0\leq k,l\leq p_N}$$

where ε and η are generic copies of $\varepsilon^{(n)}$ and η_n respectively and $\sigma^2(t) = t(1-t)$. This matrix is seen as an asymptotic Hessian matrix of the pseudo likelihood function (5) and will be used to establish an asymptotic normality result of the proposed estimator. In practice, this matrix can be replaced by its empirical version that is a consequence of (11). In the following section, we give the assumptions and consistency results of $\hat{\theta}$.

3 Results and Assumptions

Our results can be seen as generalization of that of [31] in the context of CBS. In addition to the previous hypotheses, we need to consider the following assumptions used by these authors:

(H3) The integer p_N satisfies $p_N \to \infty$ and $N^{-1/4} p_N \to 0$ as $N \to \infty$.

(H4) We have

$$\sum_{r_1,r_2,r_3,r_4=0}^{p_N} E_s\left(\frac{\mu'^4(\eta)}{\sigma^4\left(\mu(\eta)\right)}\varepsilon_{r_1}\varepsilon_{r_2}\varepsilon_{r_3}\varepsilon_{r_4}\right)\kappa_{r_1r_2}\kappa_{r_3r_4} = o(N/p_N^2)$$

where $\kappa_{kl},\ k,l = 0,\ldots,p_N+1$ are the elements of $\Xi_{p_N} \equiv \Delta_{p_N}^{-1}$.

(H5) We assume that

$$\sum_{r_1,\ldots,r_8=0}^{p_N} E_s\left(\frac{\mu'^4(\eta)}{\sigma^4\left(\mu(\eta)\right)}\varepsilon_{r_1}\varepsilon_{r_3}\varepsilon_{r_5}\varepsilon_{r_7}\right) \times E_s\left(\frac{\mu'^4(\eta)}{\sigma^4\left(\mu(\eta)\right)}\varepsilon_{r_2}\varepsilon_{r_4}\varepsilon_{r_6}\varepsilon_{r_8}\right)\kappa_{r_1r_2}\kappa_{r_3r_4}\kappa_{r_5r_6}\kappa_{r_7r_8} = o(N^2p_N^2)$$

Assumptions (H4) and (H5) are technical assumptions needed to establish the proof of this following asymptotic normality, they are similar to assumptions (M.3) and (M.4) in [31]. Assumption (H4) will then be used in the proof of (11) and (H5) is needed to prove (10) in the Appendix section, for more details on the utility of these assumptions, see [31].
Under these assumptions, we prove the asymptotic normality of $\hat{\theta}$.

Theorem 3.1 *Under assumptions (H1)-(H5), we have that the estimator $\hat{\theta}$ of $\tilde{\theta}$ satisfies*

$$\frac{N(\hat{\theta}-\tilde{\theta})^T\Delta_{p_N}(\hat{\theta}-\tilde{\theta}) - (p_N+1)}{\sqrt{2(p_N+1)}} \to \mathcal{N}(0,1).$$

To show the convergence of $\hat{\theta}(\cdot)$ to the true parameter function $\theta^*(\cdot)$, we will proceed as follow ([31]). Let $G(\cdot,\cdot)$ denote the integral kernel defined by

$$G(t,v) = E_s\left(\frac{\mu'^2(\eta)}{\sigma^2(\mu(\eta))}X(t)X(v)\right),\ t,v \in \mathcal{T}$$

and A_G be the Hilbert-Schimdt operator associated to G. Consider φ_j^G, $j = 1,2,\ldots$, the eigen-basis of the operator A_G, and λ_j^G the eigenvalues associated to this eigen-basis. The estimated parameter function $\hat{\theta}(\cdot)$ and the parameter function $\theta^*(\cdot)$ can be expressed in this eigen-basis as

$$\theta^*(t) = \sum_{j\geq 1}\theta^*_{\varphi_j^G}\varphi_j^G(t), \quad \text{and} \quad \hat{\theta}(t) = \sum_{j=1}^{p_N}\hat{\theta}_{\phi_j^G}\varphi_j^G(t).$$

Let $d_G(\cdot,\cdot)$ denote the metric defined in the $L^2(\mathcal{T})$ through the operator A_G and defined by

$$d_G^2(f,g) = \int\int (f(t)g(t))\, G(t,v)\, (f(v)-g(v))\, dtdv,\ f,g \in L^2(\mathcal{T}).$$

Then the distance between $\hat{\theta}(\cdot)$ and $\theta^*(\cdot)$ under this metric is given by

$$\begin{aligned} d_G^2\left(\hat{\theta}(\cdot),\theta^*(\cdot)\right) &= \sum_{j=1}^{p_N} \lambda_j^G \left(\hat{\theta}_{\varphi_j^G} - \theta^*_{\varphi_j^G}\right)^2 + \sum_{j>p_N} \lambda_j^G \left(\theta^*_{\varphi_j^G}\right)^2 \\ &= \left(\hat{\theta}_{\varphi^G} - \theta^*_{\varphi^G}\right)^T \Delta_{p_N}^G \left(\hat{\theta}_{\varphi^G} - \theta^*_{\varphi^G}\right) + \sum_{j>p_N} \lambda_j^G \left(\theta^*_{\varphi_j^G}\right)^2, \end{aligned}$$

with

$$\hat{\theta}_{\varphi^G} = \left(\hat{\theta}_{\varphi_1^G},\ldots,\hat{\theta}_{\varphi_{p_N}^G}\right)^T \qquad \theta^*_{\varphi^G} = \left(\theta^*_{\varphi_1^G},\ldots,\theta^*_{\varphi_{p_N}^G}\right)^T$$

the diagonal matrix $\Delta_{p_N}^G$ is obtained by diagonalizing the $p_N \times p_N$ sub-matrix of Δ_{p_N} obtained by removing the first row/column. This latter will be denoted $\Delta_{p_N}^{(-1)}$ in the following. The asymptotic distribution of the distance between $\hat{\theta}(\cdot)$ and $\theta^*(\cdot)$ is given in the following result.

Corollary 3.1 *Under conditions of Theorem 3.1 and if the parameter function $\theta^*(\cdot)$ satisfies under the CBS the following condition*

$$\sum_{j>p_N} E_s(\varepsilon_j^2) \left(\int \theta^*(t)\varphi_j^G(t)dt\right)^2 = o(\sqrt{p_N}/N) \tag{6}$$

with $\varepsilon_j^G = \dfrac{\mu'(\eta)}{\sigma^2(\mu(\eta))} \displaystyle\int X(t)\varphi_j^G(t)dt$, then as $N \to \infty$ we have

$$\frac{N d_G^2\left(\hat{\theta}(\cdot),\theta^*(\cdot)\right) - p_N}{\sqrt{2p_N}} \to \mathcal{N}(0,1).$$

Note that (6) concerns the rate of contribution to the parameter function L^2 norm of the oscillation of the functional variable X, see [31] for more detail.
The following result permits to construct a confidence band of the parameter function.

Corollary 3.2 *Denote the eigen-elements of the matrix $\Delta_{p_N}^{(-1)}$ by $(v_1,\lambda_1),\ldots,(v_{p_N},\lambda_{p_N})$, and let*

$$v_k = (v_{k,1},\ldots,v_{k,p_N})^T, \qquad \omega_k(t) = \sum_{j=1}^{p_N} v_{k,j}\varphi_j(t),\ k = 1,\ldots,p_N$$

then for large N and p_N an approximate $(1-\rho)$ simultaneous confidence band is, under conditions of Theorem 3.1, determined by

$$\hat{\theta}(t) \pm \sqrt{c(\rho) \sum_{k=1}^{p_N} \frac{\omega_k(t)^2}{\lambda_k}}$$

where $c(\rho) = \left(p_N + t_{1-\rho}\sqrt{2p_N}\right)/N$ and $t_{1-\rho}$ is the quantile of order $(1-\rho)$ of a standard normal distribution, $0 < \rho < 1$.

Under similar assumptions as those used in [31] but adapted to our context of CBS, we show that the proposed condition maximum likelihood estimator of our generalized functional binary choice model has same asymptotic properties as the ordinary maximum likelihood estimator used in the random sampling context, see for instance [31].

To investigate the numerical performance of the proposed methodology, we conduct some numerical experiments in the following section.

4 Numerical experiments

In this section, we study the performance of the proposed model towards some simulations which highlight the importance of taking into account the way of sampling the data. We remind that our theoretical results are obtained under a choice based sampling which is taken into account in our estimation procedure. We compare our estimation method by Conditional Maximum Likelihood (CML) with the one that ignores any sampling design, that is the Ordinary Maximum Likelihood (OML) method [31]. We consider a sample of realizations of some binary random variable Y and a functional covariate X. Before studying the numerical results, we describe a useful estimation procedure of the model investigated in this work. We carry out some simulations. A genetic case study, where we consider the association between genetic variants (genotypes) and phenotypes (see [14]) is also considered. For this aim, we consider the model defined in (1) and using the first twenty functions of the Fourier basis $\{\varphi_j(t) \equiv \sqrt{2}\sin(j\pi t),\ t \in [0,1]\}$, as [31], we generate the explanatory pseudo-random function defined by

$$X(t) = \sum_{j=1}^{20} \varepsilon_j \varphi_j(t),$$

where $\varepsilon_j \sim \mathcal{N}(0, 1/j)$ for $j \geq 1$. We define the parameter function by $\theta(t) = \sum_{j=1}^{20} \theta_j \varphi_j(t)$ with $\theta_2 = 0.5,\ \theta_3 = 0.2,\ \theta_4 = 0.1$ and $\theta_j = 0$ for $j > 4$. The intercept α and the coefficient θ_1 will be chosen for each of the three following models:

- Logit Model: $\Phi(t) = \exp(t)/(1+\exp(t))$.

- Probit Model: $\Phi(\cdot)$ is the standard normal distribution function
- c–loglog Model: $\Phi(t) = \exp(-\exp(-t))$.

Then the response variable is generated as a pseudo-Bernoulli random variable with probability $\Phi\left(\alpha + \int X(t)\theta(t)dt\right)$. For each model, we generate a population (X_i, Y_i) of size 2000, with $Y_i = 0, 1$, and calculate the associated proportion of $Y_i = 1$, that is Q^*. Then we draw a stratified sample of size $N = 400$ with a same share sampling for the two strates, $H^* = 0.5$. That is, in the stratified sample, the number of individuals with response $Y_i = 1$ (named cases) is equal to the number of individuals with response $Y_i = 0$ (controls), see Figure 1.

A crucial step is to apply a FPCA (functional principal components analysis) on observations of the explanatory random function, X_i, in this selected stratified sample in order to estimate the basis of functions that would be used to reduce the dimension of the functional space. This FPCA should respect the way of selecting the stratified sample before applying the CML method.

The idea used in what follows is based on results presented by [4] on the properties of FPCA in a non-random sampling frame. In fact, a FPCA in a framework of CBS can be seen as a special case of FPCA on stratified data with a specific sampling, [4]. Then, we apply our CML method using the eigenfunctions given by this stratified FPCA that are the eigenfunctions of the integral operator associated to the integral kernel defined by the variance-covariance function, estimated by

$$\hat{K}_s(t,v) = \frac{1}{N-1}\sum_{i=1}^{N} X_i(t)X_i(v)\left(\frac{H^*}{Q^*}\mathbb{I}(Y_i=1) + \frac{1-H^*}{1-Q^*}\mathbb{I}(Y_i=0)\right),\ t,v \in [0,1]. \quad (7)$$

Note that, when we apply the OML method, the eigenfunctions will be chosen through a classic FPCA, that is equivalent to use (7) with $H^* = Q^*$.

Another key step is the choice of the number p of eigenfunctions that will be used to describe the truncated model. [31] had discussed the consistency of the choice of this parameter using an Akaike information criterion (AIC). For a practical point of view, [13] had compared several approaches to choose this parameter, the usual Integrated Mean Square Error (IMSE), the Correct classification rate (CCR), the Sum of estimated variance of the basis coefficients associated to the estimated parameter function or the Deviance statistics. We will consider the IMSE approach. For that criteria, p is chosen such that it minimizes:

$$\text{IMSE}^{(p)} = \int_0^1 \left(\theta(t) - \hat{\theta}^{(p)}(t)\right)^2 dt \quad (8)$$

where $\hat{\theta}^{(p)}(\cdot)$ is the estimator of the parameter function $\theta(\cdot)$ using the first p eigenfunctions. We then compare the IMSE of the two methods of estimation, CML that takes into account the sampling design and OML, choosing p by the following approaches:

- Method I: we apply Akaike information criterion to choose p.
- Method II : we choose p that minimises the IMSE.

The studied models are replicated 200 times and the results are presented in Tables 1, 2 and 3. In each table, the columns named PCs, $\hat{\alpha}$ and IMSE, represent respectively the averages (with standard deviation in parenthesis) on these 200 replications of the number p of eigenfunctions (related to Method I and Method II), the intercept estimate $\hat{\alpha}$, and the associated IMSE defined in (8). The $p-$value column represents the $p-$value associated to a Student test with alternative hypothesis: *IMSE mean associated to the estimation by OML method is greater than that of the CML method.*
In IMSE column related to Method I, we add the median (values in brackets) of the 200 replications to compare the median values rather than the average values, since there are large standard deviations for this method.
In the Logit model (Table 1, panel (a) and (b) in Figure 3), one can notice that CML and OML methods give very similar results in terms of IMSE. The OML method gives a biased intercept estimate compare to CML method. This is classical to Logit models with real-valued explanatory variables, see for instance [26]. They showed that the OML estimate in the case of Logit model with real-valued covariates remains consistency in the case of choice-based sampling data. For the Probit model (Table 2, panel (c) and (d) in Figure 3), we note that a performance of CML method compare to OML method with a $p-$value of the test equals to 0.01 for Method II. A great improvement could be observed when applying the CML method on the c–loglog model (Table 3, panel (e) and (f) in Figure 3) with a very small $p-$values of the test for both methods.

4.1 Application to Simulated Genetic dataset

Epidemiologists who are interested to test the association between genetic variants (genotypes) and phenotypes (see [14], ...) found on generalized functional linear models a good tool to address these type of problems. In this part, we would like to investigate these types of problems in a framework where sampling is realized by CBS using simulated genetic dataset. We generate a population of 10.000 individuals with a sequence of 100 Single Nucleotide Polymorphisms (SNPs) by GENOME program, which was created by [24], using the initial parameters of this program.
In each 100 SNPs, we randomly select m variants that will be considered as causal variants and give genotypes $(g(t_1), \ldots, g(t_m))$ where by definition $g(t_i)$ $(= 0, 1, 2)$, is the number of minor alleles of a some individual at the ith–variant located at location t_i. We assume that each genotype $(g(t_1), \ldots, g(t_m))$ is considered as realizations of a random function $X(t)$, $t \in [0, 1]$ at locations $t_1, \ldots, t_m$. We use these genotypes to describe the

following model:

$$\pi = \Phi\left(\sum_{j=1}^{m} \theta(t_j) g(t_j)\right),$$

where the parameter function $\theta(\cdot)$ is defined by $|\theta(t_j)| = c \times |\log_{10} \mathrm{MAF_j}|$ where $\mathrm{MAF_j}$ was the minor allele frequency at location t_j and $\theta(t) = 0, \forall t \in [0,1]\backslash\{t_1, \ldots, t_m\}$. Then the phenotype ($Y$) is generated as a pseudo-Bernoulli random variable with probability π. In fact, $Y_i = 1$ represents the fact that the ith individual is affected by some disease (Case) and $Y_i = 0$ when the individual is not affected (Control). We consider two types of models, Logit, Probit and in each model, 20% and 50% of these 100 SNPs will be considered as causal variants. An example is given in Figure 2, where a simulated sample of genetic variants with 20% causal variants is given. In each case, the constant c and the signs of $\theta(t_j)$ will be chosen like a way to find proportions of Cases (Q^*), around 0.75 (around 7500 Cases among the 10000 individuals of the population). Three samples size $N = 200, 400$ and 600 will be selected through a CBS process such a way that in each sample the number of cases equals the number of controls ($H^* = 0.5$).
These simulations will be replicated 200 times and as in [14], in each replication the causal variants are the same for all the individuals, but we allow the causal variants to be different from replication to replication.
We compare the performances of our proposed CML method and that of classical OML method when we test the association between the genotypes et phenotypes generated by the previous model. We will compare the $p-$values (associated to each method) using results of Theorem 3.1, of the test with null hypothesis $H_0 : \theta_j = 0, j = 1, 2, \ldots$, where θ_j are the coefficients of the parameter function $\theta(\cdot)$ in the eigenbasis associated to each method. When we apply CML method, the eigenfunctions that construct the eigenbasis are estimated through the FPCA adapted to CBS detailed in the beginning of this section, and for OML method, we use a classical FPCA.
These two eigenbasis will be used to reduce the dimension of the space of the genotypes. The number of principal components p considered in each method will be chosen by using the approach based on AIC.
When comparing the boxplots in the case of CML method with that of the OML method in case of random sampling (RS) given in Figures 4, 5, we can notice that the $p-$values associated to CML method are generally smaller than those associated to OML method. This is particularly clear when we have a small stratified sample ($N = 200$) in both cases of causal variants (20%, 50%) and with greater performance of the CML in the case of 20% causal variants. A little difference between the logit and probit models concerning the tests can be observed when comparing CML and OML boxplots, given for instance in panel (a) and panel (d) of Figure 4 and panel (a) and panel (d) of Figure 5 respectively.

Conclusion

In this work, we propose a generalization of the functional binary choice models when one has a sample obtained from a Choice-Based Sampling process. The conditional maximum likelihood method of [27] and the truncation strategy introduced by [31] are connected to provide estimators of the intercept and the parameter function. In the truncation strategy, we use an eigenbasis provided by a Functional Principal Components Analysis adapted in the context of choice based sampling. The originality of the proposed method is to take into account both the functional nature of the covariate and the particular sampling design. It is shown that our estimator is asymptotically normal. After studying the theoretical behavior of the proposed methodology, we look at its practical use. The considered simulation study and application to genetic data show that our method performs better than the classical maximum likelihood method in presence of choice-based sampling. On the numerical results, the proposed estimation method leads to significantly more accurate estimates for the intercepts and the parameter function, particularly for a non logit model. Consequently, one can see the proposed methodology as a good alternative to the classical maximum likelihood method to estimate a generalized functional linear model under a stratified sampling.

We notice that the genetic simulated data can present some spatial dependence. This kind of data is not included within our approach and we could investigate the case of spatio-functional random covariates or space-time series of functional data.

5 Appendix

Proof of Theorem 3.1. It is ensured by adaptation of the proof of Theorem 4.1 of [31]. We consider same notations as these authors. Let $\|M\|_2^2 = (\sum_{k,l} m_{kl}^2)$ denote the matrix norm considered here.
Using $\mu(\cdot)$ and η, we rewrite the pseudo likelihood (5) as

$$\tilde{L}_{P_N}(\theta) = \sum_{n=1}^{N} Y_n \log(\mu(\eta_n)) + (1 - Y_n) \log(1 - \mu(\eta_n)).$$

Let $U(\theta)$ denote the gradient of this function, defined by

$$U(\theta) \equiv \Delta_\theta \tilde{L}_{p_N}(\theta) = \sum_{n=1}^{N} \frac{\mu^{'}(\eta_n)}{\tilde{\sigma}^2(\eta_n)} (Y_n - \mu(\eta_n)) \varepsilon^{(n)} \tag{9}$$

with $\tilde{\sigma}(\eta) = \sigma(\mu(\eta))$ and by definition $U(\hat{\theta}) = 0$. Let J_θ denote the Hessian matrix, that is

$$\begin{aligned} J_\theta &\equiv \Delta_\theta U(\theta) = \sum_{n=1}^{N} \frac{\partial}{\partial \eta_n} \left\{ \frac{\mu^{'}(\eta_n)}{\tilde{\sigma}^2(\eta_n)} (Y_n - \mu(\eta_n)) \varepsilon^{(n)} \right\} \Delta_\theta \eta_n \\ &= -\sum_{n=1}^{N} \frac{\mu^{'2}(\eta_n)}{\tilde{\sigma}^2(\eta_n)} \varepsilon^{(n)} \varepsilon^{(n)T} + \sum_{n=1}^{N} (Y_n - \mu(\eta_n)) \left\{ \frac{\mu^{''}(\eta_n)}{\tilde{\sigma}^2(\eta_n)} - \frac{\mu^{'}(\eta_n) \tilde{\sigma}^{2'}(\eta_n)}{\tilde{\sigma}^4(\eta_n)} \right\} \varepsilon^{(n)} \varepsilon^{(n)T} \\ &\equiv -D^T D + R. \end{aligned}$$

with $D = \sum_{n=1}^{N} \mu^{'}(\eta_n) \varepsilon^{(n)T} / \tilde{\sigma}(\eta_n)$. As in [31], we would like to show that the term R can be negligeable. Now applying a Taylor expansion on $U(\cdot)$, for $\tilde{\theta}$ between θ and $\hat{\theta}$ permits to get

$$\begin{aligned} U(\theta) &= U(\hat{\theta}) - J_{\tilde{\theta}}(\hat{\theta} - \theta) = -J_{\tilde{\theta}}(\hat{\theta} - \theta) \\ &= \left\{ D^T D + (J_\theta - J_{\tilde{\theta}}) - (J_\theta + D^T D) \right\} (\hat{\theta} - \theta) \end{aligned}$$

Then, we have

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta) &= \left\{ I_{p_N+1} + \left(\frac{D^T D}{N} \right)^{-1} \left(\frac{J_\theta - J_{\tilde{\theta}}}{N} \right) - \left(\frac{D^T D}{N} \right)^{-1} \left(\frac{J_\theta + D^T D}{N} \right) \right\}^{-1} \\ &\quad \times \left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}}. \end{aligned}$$

By assumption (H2), we have $\|\mu^{(r)}(\cdot)\| < C$, $r = 1, 2$, $\tilde{\sigma}^2(\cdot) < C$ and $\tilde{\sigma}(\cdot) > \delta$, and then

$$E_s \left(\left\| \frac{J_\theta + D^T D}{N} \right\|_2^2 \right) = \mathcal{O}(\frac{p_N^2}{N}).$$

Therefore, using (H3) implies

$$\left\| \left(\frac{D^T D}{N} \right)^{-1} \frac{J_\theta + D^T D}{N} \left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}} \right\|_2 = o_p(1).$$

As above, we have

$$\left\| \left(\frac{D^T D}{N} \right)^{-1} \frac{J_\theta - J_{\tilde{\theta}}}{N} \left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}} \right\|_2 = o_p(1).$$

Then it follows that as $N \to \infty$,

$$\left\| \sqrt{N}(\hat{\theta} - \theta) - \left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}} \right\|_2 = o_p(1).$$

Now the asymptotic distribution of $\sqrt{N}(\hat{\theta} - \theta)$ is seen as that of

$$\left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}}.$$

Define the $(p_N + 1)$ vector $\mathcal{Z}_N$ and $(p_N + 1) \times (p_N + 1)$ matrix Ψ_N

$$\mathcal{Z}_N = \Xi_N^{1/2} \frac{D^T e}{\sqrt{N}}; \qquad \Psi_N = \Delta_N^{1/2} \left(\frac{D^T D}{N} \right)^{-1} \Delta_N^{1/2}$$

with $e_n = (Y_n - \mu(\eta_n)) / \tilde{\sigma}(\eta_n)$, $n = 1, \ldots, N$, $\Delta_N \equiv \Delta_{P_N}$ and $\Xi_N \equiv \Delta_{P_N}^{-1}$.
As in [31], we consider the following decomposition

$$\begin{aligned}
\left\{ \left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}} \right\}^T \Delta_N \left\{ \left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}} \right\} &= \mathcal{Z}_N^N \Psi_N^2 \mathcal{Z}_N \\
&= \mathcal{Z}_N^T \mathcal{Z}_N + 2\mathcal{Z}_N^T (\Psi_N - I_{P_N+1}) \mathcal{Z}_N \\
&\quad + \mathcal{Z}_N^T (\Psi_N - I_{P_N+1})(\Psi_N - I_{P_N+1}) \mathcal{Z}_N \\
&\equiv F_N + G_N + H_N.
\end{aligned}$$

One can easily see that

$$\left(\mathcal{Z}_N^T \mathcal{Z}_N - (p_N + 1) \right) / \sqrt{2p_N} \to \mathcal{N}(0, 1), \tag{10}$$

using for instance Proposition 7.1 in [31] where (H5) is needed. Then, we deduce that $\left| \mathcal{Z}_N^T \mathcal{Z}_N \right| = \mathcal{O}_p(p_N)$ and under assumptions (H3) and (H4), we have

$$\|\Psi_N - I_{p_N+1}\|_2 = o_p(1/\sqrt{p_N}). \tag{11}$$

Then

$$|G_N| \leq \left| \mathcal{Z}_N^T \mathcal{Z}_N \right| \|\Psi_N - I_{p_N+1}\|_2 = o_p(\sqrt{p_N})$$

and

$$|H_N| \leq \left| \mathcal{Z}_N^T \mathcal{Z}_N \right| \|\Psi_N - I_{p_N+1}\|_2^2 = o_p(1)$$

Finally, note that

$$(2p_N)^{-1/2} \left\{ \left\{ \left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}} \right\}^T \Delta_N \left\{ \left(\frac{D^T D}{N} \right)^{-1} \frac{U(\theta)}{\sqrt{N}} \right\} - (p_N + 1) \right\} \longrightarrow \mathcal{N}(0, 1).$$

This yields the Proof of Theorem 1.
Proof of Corollary 3.2. It is similar to that of Corollary 4.3 of [31] and is then omitted.

References

[1] Liliane Bel, Avner Bar-Hen, Rmy Petit, and Rachid Cheddadi. Spatio-temporal functional regression on paleoecological data. *Journal of Applied Statistics*, 38(4):695–704, 2010.

[2] Denis Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer, 2000.

[3] Babette A. Brumback and John A. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93(443):961–976, 1998.

[4] Hervé Cardot, Mohamed Chaouch, Camelia Goga, and Catherine Labruère. Properties of design-based functional principal components analysis. *Journal of statistical planning and inference*, 140(1):75–91, 2010.

[5] Hervé Cardot, Robert Faivre, and Michel Goulard. Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *Journal of Applied Statistics*, 30(10):1185–1199, 2003.

[6] Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1):11–22, 1999.

[7] Hervé Cardot and Etienne Josserand. Horvitz–thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98(1):107–118, 2011.

[8] Hervé Cardot and Pacal Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1):24–41, 2005.

[9] John B Conway. *A course in functional analysis*, volume 96. Springer Science & Business Media, 2013.

[10] Stephen R Cosslett. Maximum likelihood estimator for choice-based samples. *Econometrica: Journal of the Econometric Society*, pages 1289–1316, 1981.

[11] Stephen R Cosslett. Efficient semiparametric estimation for endogenously stratified regression via smoothed likelihood. *Journal of Econometrics*, 177(1):116–129, 2013.

[12] Sophie Dabo-Niang and Frédéric Ferraty. *Functional and operatorial statistics*. Springer, 2008.

[13] Manuel Escabias, Ana M Aguilera, and Mariano J Valderrama. Functional pls logit regression model. *Computational Statistics & Data Analysis*, 51(10):4891–4902, 2007.

[14] Ruzong Fan, Yifan Wang, James L Mills, Tonia C Carter, Iryna Lobach, Alexander F Wilson, Joan E Bailey-Wilson, Daniel E Weeks, and Momiao Xiong. Generalized functional linear models for gene-based case-control association studies. *Genetic epidemiology*, 38(7):622–637, 2014.

[15] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer, 2006.

[16] Trevor Hastie and Colin Mallows. [a statistical view of some chemometrics regression tools]: Discussion. *Technometrics*, 35(2):140–143, 1993.

[17] Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993.

[18] Lajos Horvth and Piotr Kokoszka. *Inference for functional data with applications*, volume 200. Springer, 2012.

[19] Guido W Imbens. An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica: Journal of the Econometric Society*, pages 1187–1214, 1992.

[20] Gareth M James and Trevor J Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550, 2001.

[21] Ruth H Keogh and David Roxbee Cox. *Case-control studies*, volume 4. Cambridge University Press, 2014.

[22] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[23] Jacek A Kopec and John M Esdaile. Bias in case-control studies. a review. *Journal of epidemiology and community health*, 44(3):179, 1990.

[24] Liming Liang, Sebastian Zöllner, and Gonçalo R Abecasis. Genome: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–1567, 2007.

[25] Charles F Manski. *Partial identification of probability distributions*. Springer Science & Business Media, 2003.

[26] Charles F Manski and Steven R Lerman. The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, pages 1977–1988, 1977.

[27] Charles F Manski and Daniel McFadden. Alternative estimators and sample designs for discrete choice analysis. *Structural analysis of discrete data with econometric applications*, pages 2–50, 1981.

[28] Brian D Marx and Paul HC Eilers. Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, 41(1):1–13, 1999.

[29] Mathew W. McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert. Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269, 2014.

[30] Hans-Georg Müller, Jeng-Min Chiou, and Xiaoyan Leng. Inferring gene expression dynamics via functional regression analysis. *BMC bioinformatics*, 9(1):60, 2008.

[31] Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. *Annals of Statistics*, pages 774–805, 2005.

[32] James O; Ramsay and B.W. Silverman. *Functional data analysis*. Wiley Online Library, 2nd edition, 2005.

[33] Sarah J Ratcliffe, Gillian Z Heller, and Leo R Leader. Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression. *Statistics in Medicine*, 21(8):1115–1127, 2002.

[34] Jae W Song and Kevin C Chung. Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, 126(6):2234, 2010.

[35] Ashish Sood, Gareth M James, and Gerard J Tellis. Functional regression: A new model for predicting market penetration of new products. *Marketing Science*, 28(1):36–51, 2009.

[36] Yu Xie and Charles F Manski. The logit model and response-based samples. *Sociological Methods & Research*, 17(3):283–302, 1989.

[37] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *Ann. Statist.*, 33(6):2873–2903, 2005.

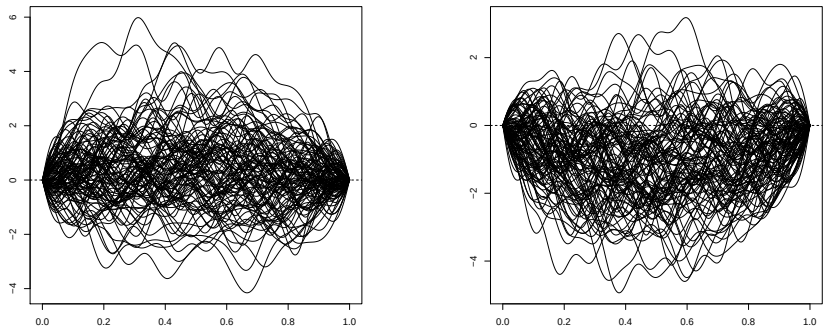


Figure 1: Simulated sample curves with response $Y = 1$ (left panel) and $Y = 0$ (right panel)

	Method I				Method II			
	PCs	$\hat{\alpha}$	IMSE	p-value	PCs	$\hat{\alpha}$	IMSE	p-value
OML	**3.20**	**0.33**	**2.39 [.19]**	0.67	**2.52**	**0.33**	**0.11**	0.51
	(2.39)	(0.07)	(10.28)		(0.63)	(0.07)	(0.08)	
CML	**3.03**	**1.35**	**2.88 [.19]**		**2.55**	**1.34**	**0.11**	
	(2.63)	(0.10)	(11.38)		(0.67)	(0.08)	(0.08)	

Table 1: Logit model with $\alpha = 1.35$ and $Q = 0.75$, $\theta_1 = 1.3$.

	Method I				Method II			
	PCs	$\hat{\alpha}$	IMSE	p-value	PCs	$\hat{\alpha}$	IMSE	p-value
OML	**3.05**	**0.63**	**0.86 [.10]**	0.29	**2.75**	**0.63**	**0.089**	0.01
	(2.08)	(0.075)	(2.98)		(0.73)	(0.071)	(0.068)	
CML	**2.99**	**1.36**	**0.72 [.09]**		**2.77**	**1.36**	**0.075**	
	(1.99)	(0.075)	(2.50)		(0.72)	(0.073)	(0.052)	

Table 2: Probit model with $\alpha = 1.35$ and $Q = 0.79$, $\theta_1 = 1.3$.

	Method I				Method II			
	PCs	$\hat{\alpha}$	IMSE	p-value	PCs	$\hat{\alpha}$	IMSE	p-value
OML	**4.23**	**-0.17**	**4.38 [.36]**	4×10^{-3}	**2.67**	**-0.19**	**0.25**	1.73×10^{-29}
	(3.95)	(0.06)	(13.19)		(0.67)	(0.06)	(0.14)	
CML	**3.16**	**1.52**	**1.63[.15]**		**2.64**	**1.49**	**0.11**	
	(2.32)	(0.10)	(5.98)		(0.72)	(0.10)	(0.09)	

Table 3: CLogLog model with $\alpha = 1.5$ and $Q = 0.70,\ \theta_1 = 1.6$.

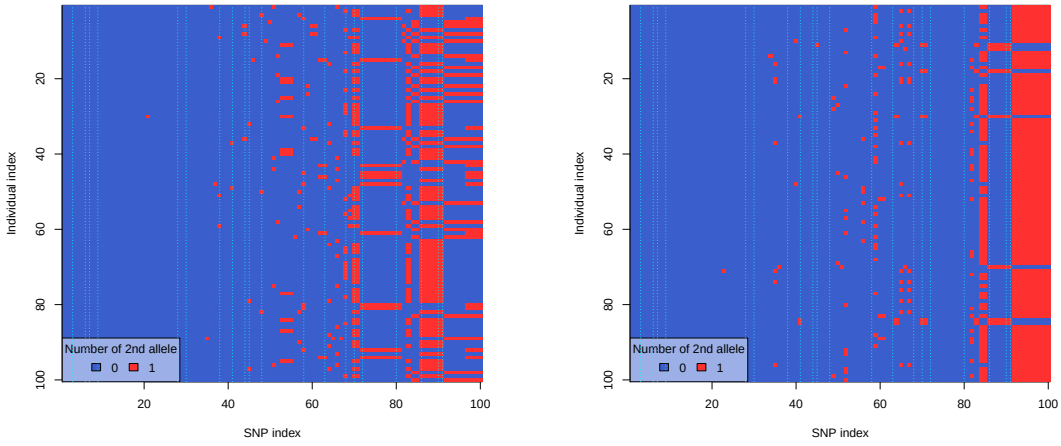


Figure 2: Simulated case-control sample of genetic variants with 20% causal variants; Cases (left panel) and Controls (right panel). Dashed lines represent causal SNPs.

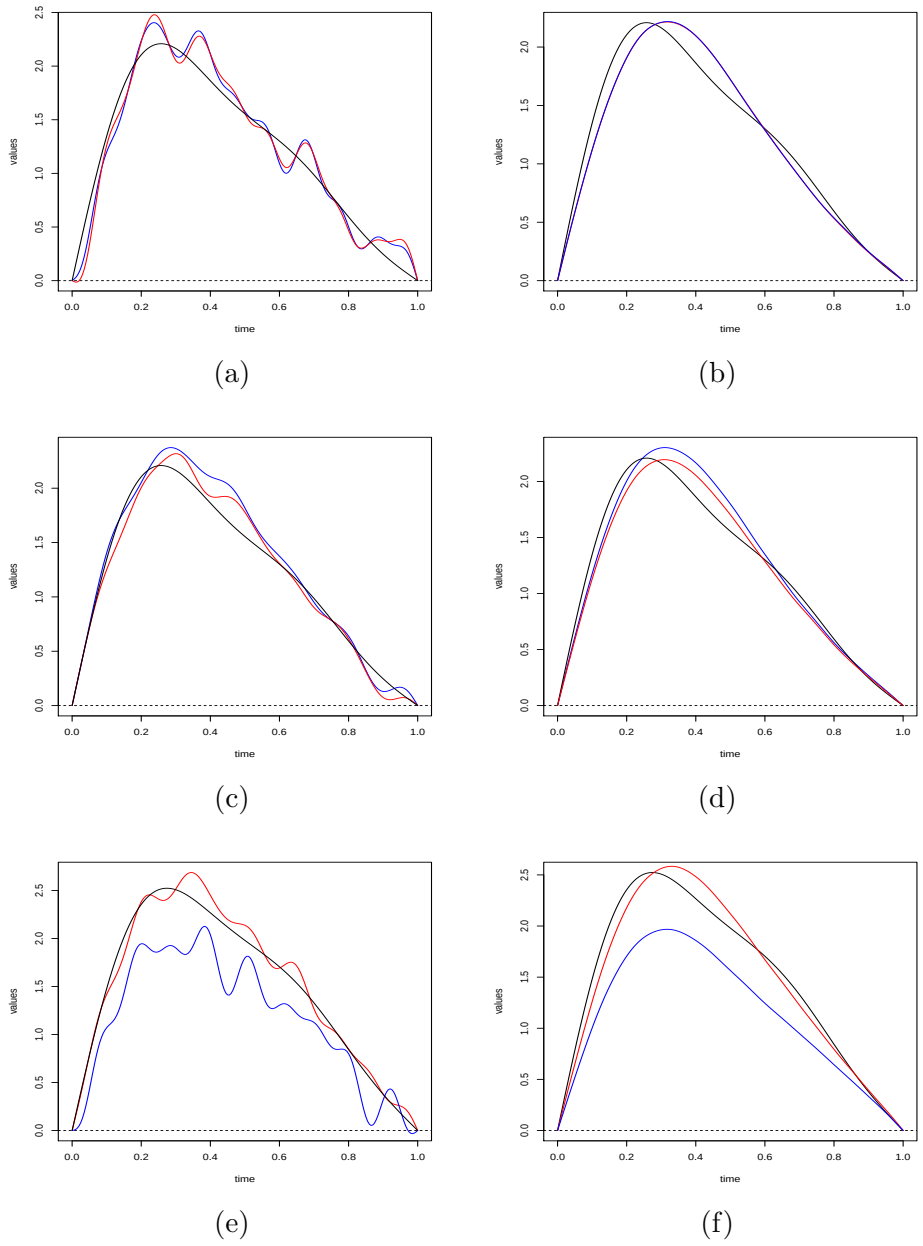
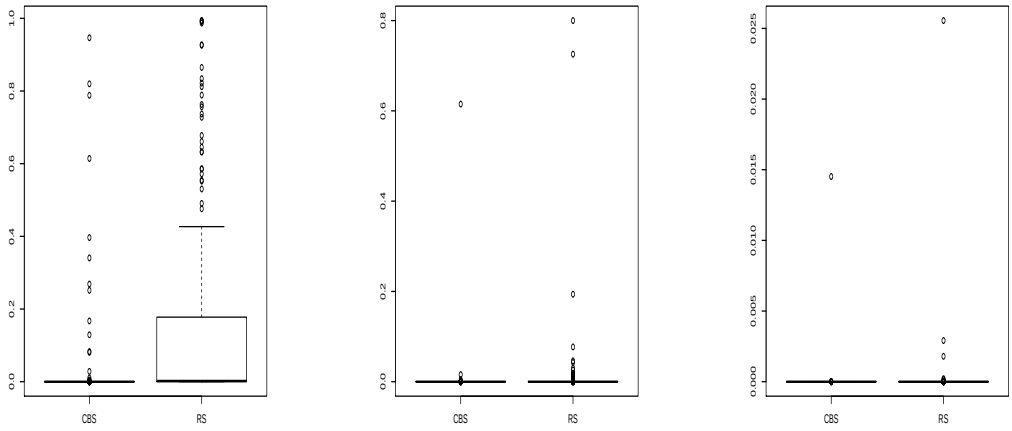


Figure 3: Graphs of the simulated parameter function θ (black curve), the means of its estimates by the OML method (blue curve) and the CML method (red curve) associated with the Method I (left panels) and method II (right panels), Logit model (panels (a) and (b)), Probit model (panels (c) and (d)) and C–loglog model (panels (e) and (f)), using 200 replications.

(a) Causal= 20%, $N = 200$ (b) Causal= 20%, $N = 400$ (c) Causal= 20%, $N = 600$



(d)Causal= 50%, $N = 200$ (e)Causal= 50%, $N = 400$ (f) Causal= 50%, $N = 600$

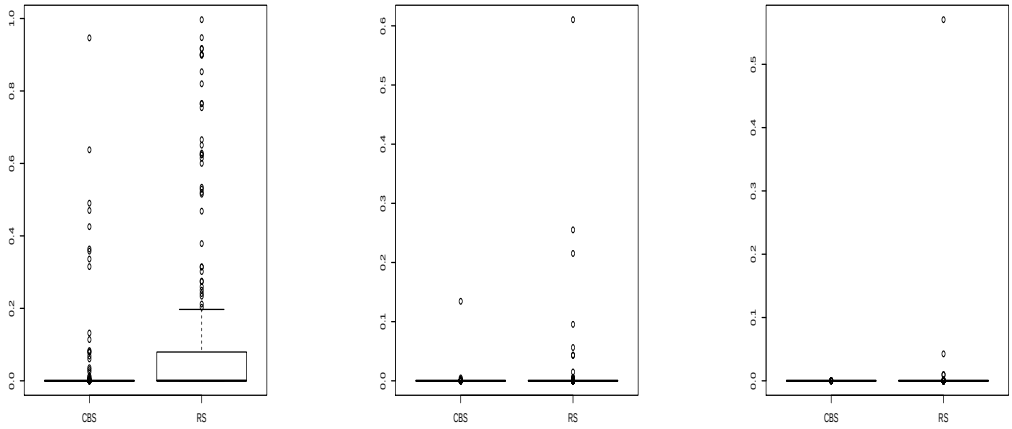
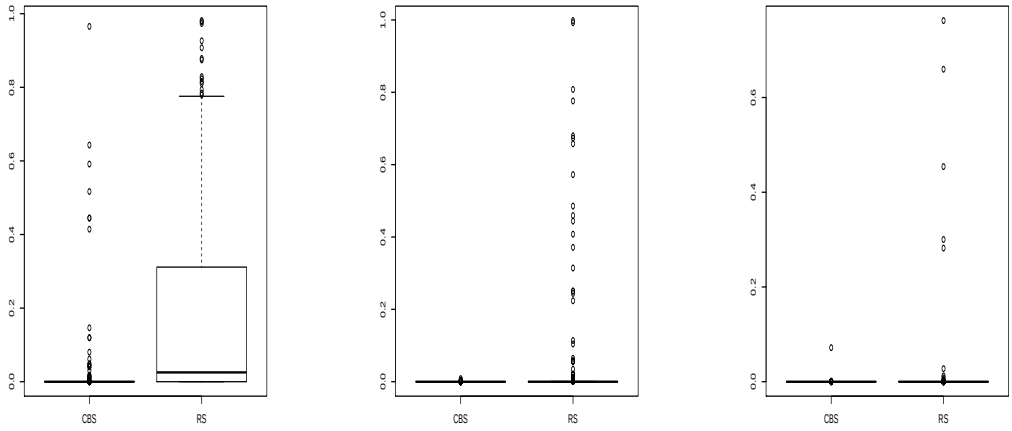


Figure 4: The boxplots of the $p-$values on the 200 replications for logit model with 20% causl variants, $c = 1.3$, $Q^* \approx 0.74$ and 10%/90% of $\theta(t_j)$ are negatives/positives (panels (a), (b), (c)). For logit model with 50% causal variants, $c = 1.5$, $Q^* \approx 0.78$ and 20%/80% of $\theta(t_j)$ are negatives/positives (panels (d), (e), (f)).

(a) Causal= 20%, $N = 200$ (b) Causal= 20%, $N = 400$ (c) Causal= 20%, $N = 600$



(d)Causal= 50%, $N = 200$ (e)Causal= 50%, $N = 400$ (f) Causal= 50%, $N = 600$

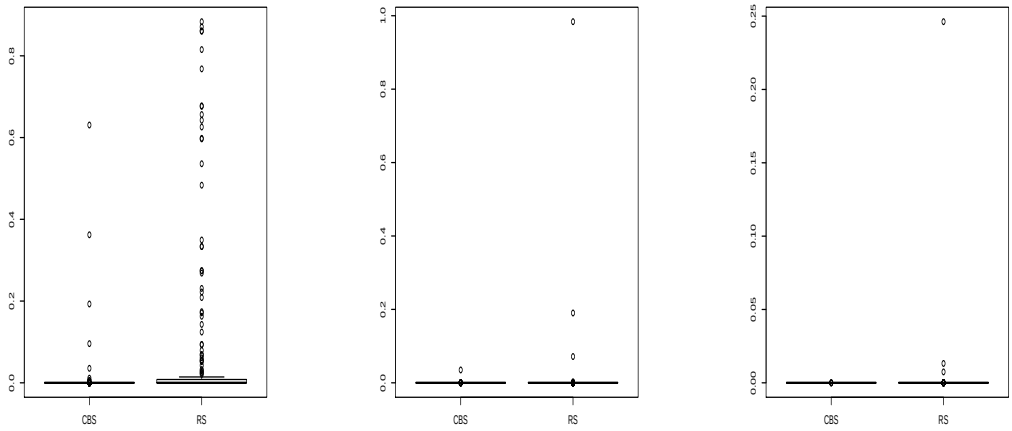


Figure 5: The boxplots of the $p-$values on these 200 replications for probit model, with 20% causal variants, $c = 0.97$, $Q^* \approx 0.74$ and 10%/90% of $\theta(t_j)$ are negatives/positives (panels (a), (b), (c)). For probit model, with 50% causal variants, $c = 0.93$, $Q^* \approx 0.78$ and 20%/80% of $\theta(t_j)$ are negatives/positives (panels (d), (e), (f)).