Highly irregular functional generalized linear regression with electronic health records

Justin Petrovich¹ | Matthew Reimherr² | Carrie Daymont³

¹Saint Vincent College, Latrobe, Pennsylvania, USA
²Pennsylvania State University, University Park, Pennsylvania, USA
³Penn State College of Medicine, Hershey, Pennsylvania, USA

Correspondence
Justin Petrovich, Alex G. McKenna School of Business, Economics, and Government, Saint Vincent College, 300 Fraser Purchase Rd., Latrobe, PA 15650, USA. Email: justin.petrovich@stvincent.edu

Abstract
This work presents a new approach, called Multiple Imputation of Sparsely-sampled Functions at Irregular Times (MISFIT), for fitting generalized functional linear regression models with sparsely and irregularly sampled data. Current methods do not allow for consistent estimation unless one assumes that the number of observed points per curve grows sufficiently quickly with the sample size. In contrast, MISFIT is based on a multiple imputation framework, which, as we demonstrate empirically, has the potential to produce consistent estimates without such an assumption. Just as importantly, it propagates the uncertainty of not having completely observed curves, allowing for a more accurate assessment of the uncertainty of parameter estimates, something that most methods currently cannot accomplish. This work is motivated by a longitudinal study on macrocephaly, or atypically large head size, in which electronic medical records allow for the collection of a great deal of data. However, the sampling is highly variable from child to child. Using MISFIT we are able to clearly demonstrate that the development of pathologic conditions related to macrocephaly is associated with both the overall head circumference of the children as well as the velocity of their head growth.

KEYWORDS
FPCA, functional generalized linear model, macrocephaly, multiple imputation
1 | INTRODUCTION

In recent years, functional data analysis (FDA) has seen a rapid expansion into what Wang et al. (2016) called next-generation functional data analysis, as more complex applications and models are explored. An ongoing challenge in this expansion is the handling of functional data that are either very irregularly sampled, sparsely sampled or contain missing regions. Classically, FDA is concerned with the statistical analysis of data where one or more variables of interest is a function. If the functions are not completely or densely observed, then fitting certain FDA models becomes substantially more challenging. Extending FDA methods to handle such data has been a rich area of research (Brumback & Rice, 1998; James et al., 2000; Rice & Wu, 2001; Shi et al., 1996; Yao et al., 2005a, 2005b).

One of the most common approaches for handling such data is to smooth or impute what is missing (Rice & Wu, 2001). This imputation can be done on the curves themselves or on the scores in a Karhunen–Loève expansion (i.e. functional principal components) (Dauxois et al., 1982; Di et al., 2009; Greven et al., 2010; Shang, 2014; Yao et al., 2005a, 2005b). This approach is especially attractive as it allows practitioners to subsequently draw upon a wide range of methods after imputation, either multivariate in the case of score level imputation or functional in the case of curve level imputation. However, as it stands, nearly all methods presented in the literature carry out the imputation process while ignoring subsequent modelling that is to be done with the reconstructed curves or scores (a notable exception being Bayesian methods, e.g. Thompson and Rosen (2008) and Kowal et al. (2019)). Such an approach can produce substantially biased estimates as well as produce unreliable standard errors and subsequent $p$-values. Indeed, in many settings the resulting estimators need not even be consistent unless one assumes that the imputed curves converge to the truth asymptotically. This assumption is mathematically convenient, but highlights a serious concern when handling sparse functional data. For these reasons, methods such as those used by Yao et al. (2005a, 2005b), which do provide consistent estimation of functional linear models, avoid imputation and instead focus on a moment-based estimation method combined with nonparameteric smoothing of covariances and cross-covariances. However, this approach is difficult to extend to nonlinear models and does not allow for the utilization of the vast literature on dense FDA methods.

In this work we present a new approach that provides a framework for developing consistent estimation techniques for both linear and nonlinear models, while also presenting a more complete bridge between sparse and dense functional data. In particular, we attack the problem by pivoting our perspective to that of a missing data problem. In the context of the literature on missing data, the goal is to impute the missing data in a way that preserves the performance of subsequent statistical modelling. We show how combining ideas from Yao et al. (2005a, 2005b) and multiple imputation (Royston, 2004; Rubin, 1996, 2004; Schafer, 1999) results in an approach that alleviates many of the discussed issues, while also remaining quite broadly applicable. Given that our foundation is built upon multiple imputation, we call our new method Multiple Imputation of Sparsely-sampled Functions at Irregular Times (MISFIT). As a by-product of this work, in Sections 2.3 and 2.4 we also present results connecting regression, imputation and the equivalence of probability measures, which are of independent interest.

1.1 | Macrocephaly

Clinical data have long been analysed using longitudinal data methods, which enable one to account for the correlation between different measurements on the same subject. However, if
one also assumes that these repeated measurements constitute realizations of a smooth curve or data-generating process, modelling the data as functional data can be advantageous both in terms of flexibility and statistical power (Craig et al., 2018; Goldsmith & Schwartz, 2017; He et al., 2011; Szczesniak et al., 2016). Since clinical visits may occur both infrequently and irregularly, their analysis poses a challenge to the current smoothing/imputing methods discussed in Section 1. To demonstrate the effectiveness of our approach at addressing these challenges, we apply it to one such clinical data set in order to predict the presence or absence of pathologic conditions related to macrocephaly.

Head circumference is routinely measured in children between birth and 2 years of age, primarily for the purpose of detecting pathologic conditions that cause atypically large head size (macrocephaly) and atypically small head size (microcephaly). Most children with a large head are healthy (Daymont et al., 2012; Wright & Emond, 2015), but when a child does have a pathologic cause for macrocephaly, a delay in diagnosis can result in poorer long-term outcomes. For example, delayed treatment for children with hydrocephalus, or excessive fluid in the brain, may result in irreversible symptoms such as cognitive impairment (Kahle et al., 2016). A large head size can also be the result of a collection of blood near the brain, which may be caused by abusive head trauma (Zahl & Wester, 2008). Delayed identification of children with bleeding from abusive head trauma may result not only in long-term symptoms, but additional, sometimes fatal, trauma (Jenny et al., 1999).

Generally, the most appropriate test to evaluate for pathology in children with macrocephaly is brain magnetic resonance imaging (MRI). Obtaining an MRI in all children with macrocephaly is not advisable. MRIs are safe, but they can involve significant cost to the patient’s family. Additionally, infants and young children receiving an MRI generally require sedation which adds additional costs and some risk of harm. Sometimes, incidental findings are seen on an MRI that are unlikely to cause harm to the child but are worrisome enough to lead to a cascading series of tests and possibly treatment, which come with still more costs and risk (Ganguli et al., 2019).

Unfortunately, distinguishing healthy children with a large head from children with pathologic conditions causing head enlargement is challenging. Expert opinion-based methods used by clinicians have been shown to discriminate poorly between children with and without pathology (Daymont et al., 2012; Wright & Emond, 2015).

Because pathology associated with macrocephaly is rare, research in this area often requires use of existing clinical data from electronic health records, rather than prospective research with high-quality measurements performed at defined intervals. Based on the typical schedule for preventive health care visits (Workgroup, Bright Futures Periodicity Schedule, and Committee on practice and ambulatory medicine, 2017), it is rare for a child in the United States to have more than 10 head circumference measurements, and many children will have fewer. There is significant variability in the timing of appointments, and clinical measurements are affected by errors of varying type and degree (Daymont et al., 2017).

The ability to characterize trajectories of sparse irregular data has potential applicability to many other clinical questions as well. For example, it can also be difficult to distinguish quickly between patients in the emergency department who will require intensive care-level interventions, such as intubation and ventilation, or will do well with more routine levels of intervention, such as small amounts of supplemental oxygen. Although clinicians are often advised to follow the trajectory of vital signs, there is currently very little evidence to indicate which features of a trajectory are most associated with the need for increased intervention.

The growth data evaluated in this paper were extracted from the electronic health record of a large primary care network (Daymont et al., 2010). Manual chart review was used to identify
children, aged between 3 days and 3 years, with pathologic conditions that are known to be associated with macrocephaly, as described in detail previously (Daymont et al., 2012). Of 74,428 children, 85 with pathologic conditions were identified.

Although the term sparsity is somewhat subjective in the context of functional/longitudinal data, many of the subjects in the present data set have just a single measurement, while the maximum number of measurements for any subject is 23. In the left panel of Figure 1 is a histogram of the number of observations per subject in the data, clearly illustrating that the modal number of measurements is 1, while relatively few children had more than 10 clinical visits and almost none had more than 15. The right panel of Figure 1 is a cumulative histogram of the relative frequency of number of observations, allowing us to determine the proportion of subjects who had no more than a given number of observations. Specifically, about 98% of subjects received 10 or fewer measurements, 49% were measured at most five times, 24% were observed no more than twice and 14% attended a single clinical visit for measurement.

In addition to the dearth of observations for many of the subjects in the data set, subjects were not observed with any uniformity or regularity. Figure 2 illustrates this and provides a glimpse of the data. While nearly half of all subjects in the data set were observed once by the time they were a month old, notice that for both the cases and the controls, several subjects were not observed until they were at least 1 year old (represented by the red lines). Furthermore, we see again several subjects with a single observation (a single dot with no attached line), and can clearly tell that visits are not guaranteed to occur at the same ages for all subjects. Having identified the head circumference trajectories as both sparse and irregular, we proceed to introduce MISFIT as an approach that accounts for these conditions in a functional regression framework before revisiting these data in Section 4.
FIGURE 2 Spaghetti plots of head circumference trajectories for all 85 cases (left panel) and 100 randomly selected controls (right panel) from the data. Red lines and dots highlight subjects whose first visit occurred when they were at least 1 year old [Colour figure can be viewed at wileyonlinelibrary.com]

1.2 Previous work

There has been extensive work on FDA with non-densely sampled data. Maybe the most widely used perspective is the principal analysis by conditional expectation (PACE) (Yao et al., 2005a, 2005b). In the context of missing data methods, PACE is a mean imputation method, although it is traditionally carried out on the principal component scores. PACE can be carried out using an extensive software package in \texttt{Matlab}, as well as a relatively recent port to \texttt{R}, which has facilitated its implementation in a variety of applications (Chiou et al., 2014; Liu et al., 2014; Peng & Müller, 2008).

While PACE is based on using local polynomial smoothing to estimate the unknown parameters, a similar approach based on splines can be found in the \texttt{refund} package in \texttt{R} (Di et al., 2009; Goldsmith et al., 2013; Staniswalis & Lee, 1998). This approach, like many of the methods in \texttt{refund}, uses a more explicit mixed effects framework to impute the scores/curves. In both cases, the primary idea is to borrow information across units to help with the imputation process. However, these methods carry out the imputation without consideration of subsequent statistical analyses. We will show that such an approach can lead to biased and even inconsistent estimators.

It is worth noting that each of the above works has addressed the problem through the lens of a sparse functional data issue, but not from a missing data perspective. In Rubin (2004), three different missingness paradigms are delineated: (a) Missing Completely at Random (MCAR), in which the missing value patterns are independent of all data, (b) Missing at Random (MAR), in which the missing value patterns depend only on the observed data and are conditionally
independent of the unobserved data and (c) Missing Not at Random (MNAR), in which the missing data patterns depend on the observed and unobserved data. While there is no explicit treatment of a missing data mechanism in any of the aforementioned papers, it is implicitly assumed that the researcher is in one of the two former paradigms, either MCAR or MAR. We make the same implicit assumption, without formally defining the missing data mechanism in our procedure.

Several papers have built upon these ideas and adopted a missing data perspective to address various forms of nonresponse or sparsity in functional regression models. For example, He et al. (2011) used a Bayesian approach to handle missingness in the response variable of a longitudinal model. They employ a Gibbs sampler to draw model parameters as well as imputed values from their posterior distribution. In the spirit of multiple imputation, they propose multiple draws of imputed values upon convergence of the chain, or to simply use multiple independent chains and form one completed data set from each of the converged chains. Unlike our approach, however, they focus on the functional mixed model and address only sparsity in the functional response variable.

Focusing on a scalar-on-function linear model, Crambes and Henchiri (2019) use a functional principal component (FPC)-based estimator of the coefficient function to impute missing values of the scalar response by incorporating the missing data mechanism (assuming MAR) into the usual predictor for the response. In the same setting, Ferraty et al. (2013) study estimation of the mean of a response variable, providing one method based on averaging predicted values and another based on propensity scores. Again, these works are distinct from ours since they focus on missingness in the response variable only. Furthermore, the former considers only a linear scalar-on-function regression model (while we consider more general generalized linear regression models), and the latter is not concerned with regression, but rather (unconditional) mean estimation of the response variable.

More akin to our set-up, Preda et al. (2010) use the nonlinear iterative partial least squares (NIPALS) algorithm to impute missing data in the functional covariates of a scalar-on-function linear model, where they assume that the missingness mechanism follows a two-state, continuous-time Markov process with exponential holding times. While the PLS-based NIPALS algorithm provides a nice alternative to regression on the FPCs, it still neglects to propagate uncertainty due to incomplete functional observations. In addition, these authors also focused only on the linear scalar-on-function regression model, leaving extensions to nonlinear models unexplored. As our motivating example requires a logistic regression set-up, none of these methods are immediately applicable.

1.3 Organization

The remainder of the paper is organized as follows. We present our framework in Section 2 for both linear and generalized linear models (GLMs) with a natural exponential family. For GLMs, we consider both the case of categorical and continuous outcomes. As a by-product, we present some interesting results concerning the relationship between functional regression and the equivalence of Gaussian measures. In Section 3 we present a numerical study that highlights the limitations of previous approaches and demonstrates how MISFIT fixes many of these issues. In Section 4 we apply MISFIT to the evaluation of head circumference trajectories. We show how our approach sheds insight into the relationship between head circumference growth and the presence of a pathology related to macrocephaly. Finally, in Section 5 we finish with concluding
remarks and future research directions, especially as they pertain to more complicated models and deeper statistical theory.

2 METHODS

2.1 Set-up and notation

We denote the underlying functional covariates as \( \{X_i(t) : t \in [0,1]; 1 \leq i \leq N \} \), where \( t \) denotes the argument of the functions, usually time, and \( i \) denotes the subject or unit. However, we assume that these curves are only observed at times \( t_{ij} \) for \( j = 1, \ldots, m_i \), and with error:

\[
x_{ij} = X_i(t_{ij}) + \delta_{ij}.
\]

We let \( x_i = (x_{i1}, \ldots, x_{imi})^\top \) denote the vector of observed values on the function \( X_i \). Explicit distributional assumptions will be made later on.

We assume that we have an outcome, \( Y_i \), that is related to \( X_i \) via a known link function \( g \):

\[
E[Y_i | X_i] = g^{-1}(\eta_i) \quad \eta_i = \alpha + \int X_i(t)\beta(t) \, dt.
\]

Throughout, when integration is written without limits, it is understood to be over the entire domain, in this case \([0,1]\). The goal of this work is to develop tools for consistently estimating \( \alpha \) and \( \beta(t) \). As we will see, using PACE to first produce scores or curves and then fit the corresponding model will not, in general, result in a consistent estimate unless one can guarantee that the smoothed/imputed curve actually converges to the truth as the sample size grows.

2.2 Linear models

In this section we assume that the outcome, \( Y_i \), is continuous and related linearly to a functional predictor \( X_i(t) \):

\[
Y_i = \alpha + \int \beta(t)X_i(t) + \epsilon_i.
\]

We assume that the \( X_i(t) \in L^2[0,1] \) are iid. Gaussian processes with mean \( \mu_X(t) \equiv 0 \) and covariance function \( C_X(t,s) = E[(X(t) - \mu_X(t))(X(s) - \mu_X(s))] \). The \( X_i(t) \) can then be expressed using the well-known Karhunen–Loève expansion:

\[
X_i(t) = \mu_X(t) + \sum_{j=1}^{\infty} \xi_{ij} v_j(t).
\]

The \( v_j(t) \) are the eigenfunctions of \( C_X \) with corresponding eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \). The scores, or FPCs, are given by \( \xi_{ij} = \langle X_i - \mu_X, v_j \rangle \), which are independent (across both \( i \) and \( j \)) mean-0 normal random variables with variance \( \lambda_j \). Throughout the paper we let \( \langle \cdot, \cdot \rangle \) denote the \( L^2[0,1] \) inner product. We also assume the errors are iid \( \epsilon_i \sim N(0, \sigma^2_\epsilon) \), \( \delta_{ij} \sim N(0, \sigma^2_\delta) \) and mutually independent.
The challenge in estimating $\beta(t)$ is that the $X_i(t)$ are not densely observed, and what is observed is observed with error. Thus, directly smoothing the $x_{ij}$ to plug into a dense estimation framework can result in substantial bias. PACE solves this problem by estimating the unknown parameters via pooled nonparametric smoothing and then using them to form best linear unbiased predictors (BLUPs) of the curves/scores. One can then plug those BLUPs into a dense estimation framework. This approach is sensible and tends to work much better than direct smoothing of the $x_{ij}$, however, it still suffers from at least two major problems that our procedure addresses. First, the imputation in PACE is done without using the outcome $Y_i$ or any consideration of the subsequent models that will be fit. This can result in a biased estimate of $\beta(t)$, as we will soon show. The second problem is that the uncertainty of the imputation is not incorporated into PACE-based procedures when forming confidence or prediction intervals, or $p$-values. While there are established ways to eliminate the bias problem in the linear case, the second problem is still relatively under-explored.

To better understand the bias problem, consider a naive estimator of $\beta(t)$ based on functional principal component analysis (FPCA), given by

$$\hat{\beta}(t) = \sum_{j=1}^{J} \frac{\hat{\xi}_{ij} Y_i}{\hat{\lambda}_j N} \hat{v}_j(t),$$

where a hat denotes an arbitrary (at this point) estimate of the analogous quantity. The motivation for this estimator is the fact that $E[\xi_{ij} Y_i] = \lambda_j \langle \beta, v_j \rangle$. However, as written, even if we knew the correct values for the parameters used for imputation, PACE assumes that $\hat{\xi}_{ij} = E[\xi_{ij}|x_i]$, and thus the most we could hope for would be

$$\frac{1}{N} \sum_{i=1}^{N} \hat{\xi}_{ij} Y_i \overset{p}{\to} E[Y_i E[\xi_{ij}|x_i]] = E[Y_i E[\xi_{ij}|x_i]] 
eq E[Y_i \xi_{ij}].$$

That the last two terms are not equal implies that this estimate is biased unless either (1) $\beta(t) \equiv 0$ or (2) the average number of points per curve tends to $\infty$ as the sample size increases. In the former case, the equality holds since we would have $E[Y_i \xi_{ij}] = E[Y_i E[\xi_{ij}]]$ and $E[Y_i E[\xi_{ij}|x_i]] = E[Y_i E[\xi_{ij}|x_i]] = E[Y_i E[\xi_{ij}|x_i]]$, while in the latter case, one would have $E[\xi_{ij}|x_i] \overset{p}{\to} \xi_{ij}$ as more points are collected (this of course requires additional assumptions). Thus for very sparse, irregular designs the bias from plugging PACE BLUPs directly into subsequent estimation procedures can be meaningful.

To alleviate this problem Yao et al. (2005a, 2005b) avoided plugging the PACE BLUPs directly into Equation (2) and instead estimated $E[Y_i \hat{\xi}_{ij}]$ directly by smoothing the cross-covariance between $X_i(t)$ and $Y_i$. While this approach works well in the linear case, it is very hard to extend to other settings when the parameter cannot be explicitly written down in terms of moments to be estimated.

Our approach to fixing these issues is to utilize draws from the conditional distribution of $X_i(t)$ given both $Y_i$ and $x_i$, which is a form of multiple imputation. In contrast, PACE can be viewed as a form of mean imputation and does not condition on the outcome. To carry out our multiple imputation framework, we need an imputation model that is compatible with Equation (1). That is, we need to ensure that our assumed distribution for $X_i(t)|Y_i, x_i$ leads to the correct distribution for $Y_i|X_i$. However, since all terms are assumed to be jointly Gaussian, it immediately follows that $X_i(t)|Y_i, x_i$ is still a Gaussian process. To carry out the conditional draws we therefore need only
determine its mean and covariance. Assume that $E[Y_i] = \mu_Y$ and $\text{Var}[Y_i] = \sigma^2_Y$ for $i = 1, \ldots, N$, and define the cross-covariance function of $X$ and $Y$ as $C_{XY}(t) = E[(X(t) - \mu_X(t))(Y - \mu_Y)]$. Then classic results on multivariate normality imply that

$$E[X_i(t)|Y_i, x_i] = \mu_X(t) + a_i(t)^T B_i d_i$$

$$\text{Cov}(X_i(t), X_i(s)|Y_i, x_i) = C_X(t, s) - a_i(t)^T B_i a_i(s),$$

where

$$d_i = \begin{pmatrix} Y_i - \mu_Y \\ x_{i1} - \mu_X(t_{i1}) \\ \vdots \\ x_{im} - \mu_X(t_{im}) \end{pmatrix}, \quad a_i(t) = \begin{pmatrix} C_{XY}(t) \\ C_X(t, t_{i1}) \\ \vdots \\ C_X(t, t_{im}) \end{pmatrix}$$

and

$$B_i^{-1} = \begin{pmatrix} \sigma^2_Y & C_{XY}(t_{i1}) & C_{XY}(t_{i2}) & \cdots \\ C_{XY}(t_{i1}) & C_X(t_{i1}, t_{i1}) + \sigma^2_\delta & C_X(t_{i1}, t_{i2}) & \cdots \\ C_{XY}(t_{i2}) & C_X(t_{i2}, t_{i1}) & \ddots & \vdots \\ \vdots & \vdots & \cdots & C_X(t_{im}, t_{im}) + \sigma^2_\delta \end{pmatrix}.$$ 

To use these expressions we must first estimate the requisite parameters, which can be done using many well established methods (Li & Hsing, 2010; Liebl, 2019; Xiao et al., 2018; Yao et al., 2005a, 2005b; Zhang & Wang, 2016). After the parameters are estimated, the user must choose an integer $K$, and draw $K$ times from the above conditional distribution to form $K$ imputations, $X^{(k)}_i(t)$, $k = 1, \ldots, K$. Typically $K$ is around 10, although more is always better. For each of these imputed samples we can form the complete data estimates, $\hat{\beta}^{(k)}(t)$, using any of a number of estimation methods (e.g. FPCA, splines, reproducing kernel Hilbert space [RKHS], etc). In the simulations we will focus on FPC-based estimates so as to better compare against PACE.

In the case where FPCA is going to be used to estimate $\beta(t)$, it can be convenient to impute the scores directly. In this case, we are interested in the conditional distribution $\xi_{ij}|Y_i, x_i$. Nearly the same expressions can be used, except that the form for $a_i(t)$ changes into

$$A_i = \begin{pmatrix} \langle C_{XY}, v_1 \rangle & \langle C_{XY}, v_2 \rangle & \cdots & \langle C_{XY}, v_J \rangle \\ \lambda_1 v_1(t_{i1}) & \lambda_2 v_2(t_{i1}) & \cdots & \lambda_J v_J(t_{i1}) \\ \vdots \\ \lambda_1 v_1(t_{im}) & \lambda_2 v_2(t_{im}) & \cdots & \lambda_J v_J(t_{im}) \end{pmatrix},$$

and we then get that

$$E[\xi_{ij}|d_i] = A_i^T B_i d_i \quad \text{Var}(\xi_{ij}|d_i) = \text{diag}\{\lambda_1, \ldots, \lambda_J\} - A_i^T B_i A_i,$$ 

(3)

In this case, after the imputations are made, one can move directly to using the scores to estimate $\beta(t)$. 
A major advantage of using a multiple imputation approach is that we can account for the uncertainty introduced in the imputation process. This is accomplished by using Rubin’s rules (Rubin, 2004), namely, we compute the estimated within- and between-imputation covariance functions of the $\hat{\beta}^{(k)}(t)$,

$$
\hat{W}(t, s) = \frac{1}{K} \sum_{k=1}^{K} \hat{\text{Cov}}(\hat{\beta}^{k}(t), \hat{\beta}^{k}(s))
$$

and

$$
\hat{B}(t, s) = \frac{1}{K - 1} \sum_{k=1}^{K} (\hat{\beta}^{(k)}(t) - \hat{\beta}(t))(\hat{\beta}^{(k)}(s) - \hat{\beta}(s))
$$

and then use the following as our final estimated coefficient and covariance functions

$$
\hat{\beta}(t) = \hat{\beta}(t) = K^{-1} \sum_{k=1}^{K} \hat{\beta}^{(k)}(t)\ 
$$

$$
\hat{C}_{\hat{\beta}}(t, s) = \hat{W}(t, s) + (1 + 1/K)\hat{B}(t, s).
$$

Using these quantities one can carry out statistical inference for $\beta(t)$.

### 2.3 Categorical outcomes

Recall that a logistic regression model, which is the canonical link function for binary outcomes, for $Y_i|X_i$ implies that $Y_i \in \{0, 1\}$ and

$$
\logit(p_i) = \alpha + \int X_i(t)\beta(t) \ dt,
$$

where $p_i = E[Y_i|X_i] = P(Y_i = 1|X_i)$. In order to extend our methodology we have to determine a proper imputation model for the $X_i(t)$. In other words, we have to select a model for $X_i(t)|Y_i$ that implies that $Y_i|X_i(t)$ satisfies a logistic regression model. Interestingly, under certain conditions, one can still assume that $X_i(t)|Y_i$ is Gaussian. Such results come about in multivariate linear discriminant analysis when comparing to logistic regression for multivariate data (Efron, 1975). We can extend those concepts to the case of functional data and we end up with the theorem below, which is interesting in its own right.

**Theorem 1** Assume for $i = 1, \ldots, N$ that $Y_i \in \{0, 1\}$ are iid Bern($p_0$) and $X_i \in L^2[0, 1]$ with $X_i|Y_i = y \sim \mathcal{N}(\mu + y\Delta, C_X)$, where $C_X$ has full rank and $\Delta \in L^2[0, 1]$.

1. If $\|C_X^{-1/2}\Delta\|^2_{L^2} = \infty$, then the conditional distribution of $Y_i|X_i$ is degenerate.
2. If $\|C_X^{-1/2}\Delta\|^2_{L^2} < \infty$, but $\|C_X^{-1}\Delta\|^2_{L^2} = \infty$, then $Y_i|X_i$ is not degenerate, but (4) cannot hold for any $\beta \in L^2[0, 1]$.
3. If $\|C_X^{-1}\Delta\|^2_{L^2} < \infty$ then (4) holds with $\beta = C_X^{-1}\Delta \in L^2[0, 1]$.

Theorem 1 is critical for justifying our imputation procedure. It implies that if one aims to fit a logistic regression model for the response then one can use a Gaussian model to first impute the predictors and then fit the subsequent model for the response.

Before further interpreting the above theorem, we stress one point about $C_X$ and $C_X^{-1}$. Recall that $C_X$ is always a linear, self-adjoint compact operator, and its inverse will exist as long as the null space of $C_X$ only contains the zero function (i.e. all eigenvalues are positive). However, even
when \( C_X^{-1} \) does exist, it is neither compact nor even bounded. Thus, an implicit part of assuming that \( \|C_X^{-1} \Delta\|^2_{L^2} < \infty \) is the assumption that this quantity exists and is well-defined.

Another subtle point is the ‘grey area’ where \( \|C_X^{-1/2} \Delta\|^2_{L^2} < \infty \), but \( \|C_X^{-1}\|^2_{L^2} = \infty \). In this case, the logistic regression exists and is nondegenerate, but one has to replace \( \langle x, \beta \rangle \) with \( T_\beta(x) \) where \( T_\beta \) is a linear functional that is not continuous. When \( x \) is a sample path from \( X \), \( T_\beta(x) \) will be finite almost surely, but it need not be finite for a general element of \( L^2[0,1] \). Put another way, \( T_\beta \) will be an element of the algebraic dual of \( L^2[0,1] \), but not the topological dual.

Theorem 1 is based on the orthogonality/equivalence of probability measures. In particular, if the distributions for \( X_i(t) | Y_i = 0 \) and \( X_i(t) | Y_i = 1 \) are orthogonal, then it is possible to determine the value of \( Y_i \) from \( X_i(t) \) with probability 1, and thus no logistic model can exist as the probabilities would have to be 0 or 1. The quantity \( \|C_X^{-1/2} \Delta\|^2_{L^2} \) comes up in both Delaigle and Hall (2012) and Dai et al. (2017) in terms of classification for FDA. They show there that if this condition is not satisfied then perfect classification is possible. What was not discussed, however, was the connection to the orthogonality of Gaussian measures. Indeed, this same quantity was discovered at least as early as the 70s (see the historical discussions in Kuo (1975) and Bogachev (1998) for additional details). Clearly, if two measures are orthogonal then it is possible to determine, with probability 1, whether a sample came from one or the other. This issue was discussed more deeply in recent work by Berrendero et al. (2018) in the context of using RKHS methods for classification as well as Mirshani et al. (2019) in the context of functional data privacy.

With these tools in hand, we can now carry out our imputation for logistic regression. In particular, we simply impute the group for \( Y_i = 0 \) and \( Y_i = 1 \) separately (although common parameters are still estimated jointly, as discussed in Section 3):

\[
\begin{align*}
E[X_i(t)|Y_i = y,x_i] &= \mu_y(t) + a_i(t)^T B_i(x_i - \mu_i) \\
\text{Cov}(X_i(t),X_i(s)|Y_i,x_i) &= C_X(t,s) - a_i(t)^T B_i a_i(s),
\end{align*}
\]

where we now have

\[
a_i(t) = \begin{pmatrix} C_X(t,t_{i1}) \\ \vdots \\ C_X(t,t_{im_i}) \end{pmatrix}, \quad B_i^{-1} = \begin{pmatrix} C_X(t_{i1},t_{i1}) + \sigma_\delta^2 & C_X(t_{i1},t_{i2}) & \cdots \\ C_X(t_{i2},t_{i1}) & C_X(t_{i2},t_{i2}) & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ C_X(t_{im_i},t_{im_i}) + \sigma_\delta^2 \end{pmatrix},
\]

\[
\mu_i = E[x_i|Y_i = y] = \{ \mu_y(t_{ij}) \}, \quad \text{and } \mu_y(t) \text{ is the mean of the } X_i \text{ from group } y. \text{ In addition, one can do this for the scores as well. However, a caveat is that one cannot use the scores directly in fitting the subsequent logistic model as the difference between the two groups is entirely captured by the means, which are removed when computing scores. We thus recommend instead working with } \xi_{ij} = \langle \mu_y, v_j \rangle + \xi_{ij}, \text{ which does not have the mean effect removed and still allows one to estimate the coefficients of } \beta(t) \text{ in the FPCA basis. To impute the } \xi_{ij}, \text{ much the same formula can be used but now}
\]

\[
A_i = \begin{pmatrix} \lambda_1 v_1(t_{i1}) & \lambda_2 v_2(t_{i1}) & \cdots & \lambda_J v_J(t_{i1}) \\ \vdots \\ \lambda_1 v_1(t_{im_i}) & \lambda_2 v_2(t_{im_i}) & \cdots & \lambda_J v_J(t_{im_i}) \end{pmatrix},
\]
is used in place of \( a_i(t) \) in Equation (5), and \( \mu_y(t) \) and \( C_X(t, s) \) are replaced by the mean and covariance of the scores—\( \mathbf{0} \) and \( \text{diag}(\lambda_1, \ldots, \lambda_J) \) respectively. Again, after imputation one should then construct \( \tilde{\xi}_{ij} \) which have the means added back in, before fitting the subsequent logistic model.

### 2.4 Continuous outcomes

Lastly we consider the case where \( Y \) is continuous, but not necessarily Gaussian. As with the previous section we will show that a Gaussian imputation model for \( X|Y \) will be consistent with a GLM for \( Y|X \), however, we require that \( Y|X \) utilize a canonical link function. In particular, this means that the conditional density of \( Y|X = x \) is given by

\[
f_{Y|X}(y|x) = h(y) \exp\{-\frac{1}{2}T(y)^2||C_X^{-1/2}\Delta||^2\},
\]

where \( \eta_x = \alpha + \langle \beta, x \rangle \), \( h \) is some known function that determines the family of the GLM, \( T \) is a known function representing the sufficient statistic, and \( \exp\{-A(\eta_x)\} \) is the normalizing constant.

**Theorem 2** Assume \( Y_i \in \mathbb{R} \) are iid with density \( f(y) \propto h(y) \exp\{-\frac{1}{2}T(y)^2||C_X^{-1/2}\Delta||^2\} \), which is assumed to be integrable, \( X_i \in L^2[0, 1] \) with \( X_i|Y_i = y \sim \mathcal{N}(\mu + T(y)\Delta, C_X) \), and \( \Delta \in L^2[0, 1] \).

1. If \( ||C_X^{-1/2}\Delta||_{L^2}^2 = \infty \), then the conditional distribution of \( Y_i|X_i \) is degenerate.
2. If \( ||C_X^{-1/2}\Delta||_{L^2}^2 < \infty \), but \( ||C_X^{-1}\Delta||_{L^2}^2 = \infty \), then \( Y_i|X_i \) is not degenerate, but (6) cannot hold for any \( \beta \in L^2[0, 1] \).
3. If \( ||C_X^{-1}\Delta||_{L^2}^2 < \infty \) then (6) holds with \( \beta = C_X^{-1}\Delta \in L^2[0, 1] \).

Again, Theorem 2 is critical in justifying our imputation procedure. However, unlike in the linear or logistic case, one should not impute the predictors using \( Y \), but instead \( T(Y) \), which is the sufficient statistic determined by the particular parametric family. In that case, one is justified in using Gaussian imputation for \( X_i \) based on \( T(Y_i) \).

In extending to more general exponential families, the key hurdle appears to be working with a noncanonical link function. When swapping from conditioning \( Y|X \) to \( X|Y \), the latter is always in a canonical form. While on the surface this may seem minor, this appears to be a serious open mathematical question. It is desirable to assume that \( X|Y \) is Gaussian since this simplifies the imputation, however, then one is essentially ‘boxed in’ to using a canonical link function.

### 2.5 Computation

Implementing MISFIT for either linear or logistic regression requires the estimation of a number of parameters in the imputation model. While the estimation of these imputation parameters is not our focus, we wish to make clear how one can implement our imputation strategy. The results provided in Sections 3 and 4 are computed in \( \mathbb{R} \) (R Core Team, 2018), using the package \texttt{fcr} (Leroux et al., 2018) to estimate imputation parameters. Based on Leroux et al. (2017), this package is designed to fit functional concurrent regression models and allows us to regress \( X_i(t) \) on \( Y_i \) as follows:

\[
X_i(t) = f_0(t) + f_1(t)Y_i + b_i(t),
\]

(7)
where the \( b_i(t) \overset{\text{iid}}{\sim} N(0, C_b(t, s)) \) are curve-specific random effects. The curves are observed with noise such that

\[
X_i(t_{ij}) = f_0(t_{ij}) + f_1(t_{ij})Y_i + b_i(t_{ij}) + \delta_{ij},
\]

with the \( b_i(t_{ij}) \) and \( \delta_{ij} \) mutually independent. Through fitting this model, we obtain estimates of \( f_0(t), f_1(t), C_b(t, s), \) and \( \sigma^2_\delta \), from which we can in turn estimate all necessary imputation parameters.

**Linear model:** Using Equations (7) and (8), we can directly compute the mean and covariance of \( X_i(t) \) as well as the cross covariance between \( X_i(t) \) and \( Y_i \). This gives us:

\[
\mu_X(t) = f_0(t) + f_1(t)\mu_Y \quad \mu_X(t_{ij}) = f_0(t_{ij}) + f_1(t_{ij})\mu_Y
\]

\[
C_X(t, s) = f_1(t)f_1(s)\sigma^2_Y + C_b(t, s) \quad C_X(t_{ij}, s) = f_1(t_{ij})f_1(s)\sigma^2_Y + C_b(t_{ij}, s)
\]

\[
C_{XY}(t) = f_1(t)\sigma^2_Y \quad C_{XY}(t_{ij}) = f_1(t_{ij})\sigma^2_Y
\]

and

\[
C_X(t_{ij}, s_{ik}) = f_1(t_{ij})f_1(s_{ik})\sigma^2_Y + C_b(t_{ij}, s_{ik}) + 1_{(j=k)}\sigma^2_\delta.
\]

From there, one can obtain the \( \lambda_j \) and \( v_j(t) \) from a spectral decomposition of \( C_X(s, t) \).

**Logistic model:** As our logistic regression imputation described in Equation (5) is done separately for the two groups (i.e. conditioning on \( Y_i \)), the results are somewhat different than in the linear case and contain no parameters of \( Y_i \). That is, we instead compute \( E[X_i(t)|Y_i = y] = \mu_y(t) \) and \( \text{Cov}(X_i(t), X_i(s)|Y_i = y) = C_X(t, s) \). Again, from Equations (7) and (8) we get

\[
\mu_y(t) = f_0(t) + f_1(t)Y_i \quad \mu_y(t_{ij}) = f_0(t_{ij}) + f_1(t_{ij})Y_i
\]

\[
C_X(t, s) = C_b(t, s) \quad C_X(t_{ij}, s) = C_b(t_{ij}, s)
\]

and

\[
C_X(t_{ij}, s_{ik}) = C_b(t_{ij}, s_{ik}) + 1_{(j=k)}\sigma^2_\delta.
\]

Again, we can readily compute the \( \lambda_j \) and \( v_j \) from \( C_X(t, s) \), and that gives us all that we need since the logistic imputation model does not require \( C_{XY}, \mu_Y, \) or \( \sigma^2_Y \).

## 3 | SIMULATIONS

In this section, we compare the performance of our proposed method, MISFIT, with that of PACE in both a linear and logistic scalar-on-function regression setting, investigating the estimation accuracy, as well as the type 1 error rates and power, of their resulting estimators. Since we expect that the PACE imputation approach is biased for a small average number of observations per curve, \( m \), we compare across simulated data sets with varying values of \( m \), as well as varying sample sizes. In addition, since the FPCA estimator given in Equation (2) depends on the value \( J \) to truncate the sum, we must specify a fixed \( J \) for both imputation approaches with each simulated data set. For the MISFIT approach we generate \( K = 10 \) completed data sets for all of the following
Finally, while in Section 2 we detailed both a curve-level and a score-level imputation strategy, we use only the score-level imputation in all of the simulations. The code used to perform MISFIT was compiled into an R package and can be accessed through GitHub at https://github.com/justin-petrovich/sparsefreg.

### 3.1 Linear model

For a linear model, we first simulate \( N \) iid random curves, \( \{X_1(t), \ldots, X_N(t)\} \), from a mean-0 Gaussian process with the following Matérn covariance function

\[
C_X(t, s) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} \left( \frac{\sqrt{2\nu|t-s|}}{\rho} \right)^\nu K_\nu \left( \frac{\sqrt{2\nu|t-s|}}{\rho} \right),
\]

where \( K_\nu \) represents the modified Bessel function of the second kind, and we choose \( \nu = 5/2, \rho = 0.5, \) and \( \sigma^2 = 1 \). These curves are evaluated at \( M = 100 \) equally spaced times from \([0, 1]\). Since we assume that the random curves are observed with error, we add noise to the realized curves to produce the observed curves, where \( \sigma^2 = 0.5 \). We next define \( \beta(t) = w \times \sin(2\pi t) \), where \( w \) is a weight coefficient chosen to adjust the signal. The responses, \( Y_i, i = 1, \ldots, N \), are then generated according to Model 1, where \( \alpha = 0 \) and \( \sigma^2 = 1 \). Finally, for each observed curve, we randomly sample \( m \) time points from the length-100 grid to observe, so that the observed data used for imputation are \( \{Y_i, x_{i1}, \ldots, x_{im}\} \), for \( i = 1, \ldots, N \). Once the scores are imputed, a linear regression model is fit using the first \( J \) scores.

#### 3.1.1 True parameters

For the first set of simulations, we treat the true parameters as known; that is, we use the true values for \( \sigma_Y^2, C_{XY}(t, s), \sigma_{\delta_j}^2, v_j(t), \) and \( \lambda_j \), for \( j = 1, \ldots, J \). In addition, we simulate data sets of different sample sizes, \( N \in \{100, 200, 400, 800\} \); different number of observations per curve, \( m \in \{2, 5, 10, 20\} \); different numbers of FPCs, \( J \in \{1, \ldots, 6\} \); and with different signals, \( w \in \{0, 5, 10\} \). Each of these settings is simulated 1,000 times. Since we are primarily interested in the accuracy of the coefficient function estimates \( \hat{\beta}(t) \), for all simulations in Section 1 we report the median integrated squared error (MISE) of \( \hat{\beta}(t) \), defined as \( \text{median} \int (\hat{\beta}_s(t) - \beta(t))^2 dt \), where \( s = 1, \ldots, 1000 \) indexes the particular run and the median is taken over these 1,000 simulations. The median ISE is reported instead of the mean ISE because, particularly when \( m = 2 \) or \( 5 \), the imputation methods are prone to occasional large outliers.

Tables 1, 2 and 3 show the MISE of \( \hat{\beta}(t) \) across varying \( J, N \) and \( m \) respectively. For all of the simulations, the default settings of \( N = 200, m = 2 \) and \( J = 4 \) are used, allowing one of these to vary in each table (e.g. in Table 2 \( m = 2 \) and \( J = 4 \), while \( N \) varies). The choice of \( J = 4 \) will become obvious after consideration of Table 1. However we also include simulations with \( J = 3 \) and \( J = 5 \), the results of which can be found in the supplementary material, Section C.1. The comparisons are much the same, regardless of the choice of \( J \).

Beginning with a look at Table 1, notice that when the sample size is small (\( N = 200 \)) and curves are observed very sparsely (\( m = 2 \)), regardless of how many FPCs are used and the strength of the signal, MISFIT is more accurate than PACE. In the presence of some signal, the MISE for MISFIT can be reasonably large when too few (e.g. 1, 2 or 3) FPCs are used, driven by truncation bias. In these cases, MISFIT’s advantage over PACE is most accentuated when \( J \) is chosen as 4,
TABLE 1  Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $J$ increases, for a linear model using true imputation parameters

| $J$ | $w = 0$ | $w = 5$ | $w = 10$ |
|-----|---------|---------|----------|
|     | MISFIT  | PACE    | MISFIT   | PACE    | MISFIT   | PACE    |
| 1   | 0.00    | 0.00    | 12.38    | 12.39   | 49.52    | 49.55   |
| 2   | 0.01    | 0.05    | 3.30     | 3.40    | 13.16    | 13.44   |
| 3   | 0.02    | 0.66    | 3.31     | 4.91    | 13.17    | 17.85   |
| 4   | 0.06    | 11.67   | 0.11     | 30.94   | 0.23     | 85.16   |
| 5   | 0.19    | 184.74  | 0.23     | 471.04  | 0.34     | 1240.48 |
| 6   | 0.60    | 2026.16 | 0.58     | 5734.24 | 0.56     | 14614.36|

TABLE 2  Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $N$ increases, for a linear model using true imputation parameters

| $N$ | $w = 0$ | $w = 5$ | $w = 10$ |
|-----|---------|---------|----------|
|     | MISFIT  | PACE    | MISFIT   | PACE    | MISFIT   | PACE    |
| 100 | 0.12    | 26.03   | 0.17     | 66.26   | 0.28     | 186.51  |
| 200 | 0.06    | 11.71   | 0.11     | 30.94   | 0.23     | 85.16   |
| 400 | 0.03    | 6.11    | 0.07     | 15.83   | 0.20     | 44.40   |
| 800 | 0.01    | 3.06    | 0.06     | 7.61    | 0.19     | 19.28   |

TABLE 3  Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $m$ increases, for a linear model using true imputation parameters

| $m$ | $w = 0$ | $w = 5$ | $w = 10$ |
|-----|---------|---------|----------|
|     | MISFIT  | PACE    | MISFIT   | PACE    | MISFIT   | PACE    |
| 2   | 0.06    | 11.71   | 0.11     | 31.04   | 0.23     | 85.66   |
| 5   | 0.10    | 4.26    | 0.13     | 7.46    | 0.24     | 16.84   |
| 10  | 0.14    | 1.96    | 0.16     | 2.86    | 0.27     | 5.56    |
| 20  | 0.18    | 0.91    | 0.21     | 1.17    | 0.31     | 2.11    |

5 or 6, in which case enough FPCs are used to capture the complexity of the shape of $\beta(t)$. It is worth highlighting that, in the absence of parameter estimation error, choosing a large enough value $J$ for MISFIT results in an (approximately) unbiased estimator. However, the variance tends to increase with larger choices of $J$, so some balance is required. Based on our observations, the variance of the MISFIT estimator remains relatively low for all choices of $J$, while the variance of the mean-imputation-based PACE estimator balloons for larger $J$.

In Table 2 we see that increasing the sample size benefits both approaches. The improvement in MISE occurs through a reduction in variance rather than a reduction in bias. The MISFIT estimator noticeably has the smallest MISE across all signals and all sample sizes. Notice that the tendency of MISFIT’s MISE towards zero gives evidence that MISFIT results in a consistent estimator. Its minimal MISE at larger sample sizes is due to truncation error resulting from the use of a finite number of FPCs; as Table 1 showed, a larger choice of $J$ helps to alleviate this for MISFIT.
On the other hand, the large and persistent MISE of the PACE estimator as the sample size grows is noteworthy. PACE clearly does poorly for small sample sizes, especially when the signal is strong. And while consistency for PACE has already been established by Yao et al. (2005a, 2005b), MISFIT clearly outperforms it in such a sparse setting, even with a large sample size.

The previous two tables have showed convincing results that MISFIT is the best estimator when \( m = 2 \), as we expected. Table 3 allows us to compare MISFIT and PACE for an increasing number of observed points per curve, \( m \). Unsurprisingly, the two approaches begin to converge in MISE, regardless of the strength of the signal, as \( m \) increases and the amount of information observed prior to imputation grows. Still, MISFIT maintains its status of lowest MISE even up to 20 observations per curve.

One other interesting point borne out by Table 3 is that MISFIT results in a lower MISE the sparser the observed curves. Since MISFIT draws imputed values from the correct theoretical distribution, multiple imputation results in multiple sets of 'correct' scores. Thus, imputing m more scores actually decreases the variability of the resulting estimate; the increasing MISE as \( m \) grows is purely an artefact of this phenomenon. We will see in the next section that using estimated imputation parameters as opposed to the true parameters disturbs the distribution used by MISFIT enough to nullify this peculiarity.

3.1.2 Estimated parameters

Unlike in the previous section, in practice the imputation parameters must themselves first be estimated. Here, we mimic the above simulations replacing the true imputation parameters, \( \sigma_Y^2 \), \( C_{XY}(t) \), \( C_X(s, t) \), \( \sigma_Y^2 \), \( \lambda_j \), for \( j = 1, \ldots, J \), with their estimates from the data.

For MISFIT, we obtain estimates \( \hat{C}_{XY}(t) \), \( \hat{C}_X(s, t) \), \( \hat{\mu}(t) \), and \( \hat{\sigma}_Y^2 \) using \( \text{fcr} \) (Leroux et al., 2018) to regress \( X_i(t) \) on \( Y_i \). Estimates \( \{ \hat{\lambda}_j \}_{j=1}^J \) and \( \{ \hat{v}_j(t) \}_{j=1}^J \) are then taken to be the eigenvalues and eigenfunctions of \( \hat{C}_X(t) \), and the usual sample mean and variance are used for \( \hat{\mu}_Y \) and \( \hat{\sigma}_Y^2 \). For PACE, we ignore the \( Y_i \) and use the function \( \text{face.sparse} \) (Xiao et al., 2018) from the \textit{face} package (Xiao et al., 2019) to compute estimates \( \hat{C}_X \), \( \hat{\mu}(t) \), and \( \hat{\sigma}_s \) (where again \( \{ \hat{\lambda}_j \}_{j=1}^J \) and \( \{ \hat{v}_j(t) \}_{j=1}^J \) are obtained from a spectral decomposition of \( \hat{C}_X \)).

Simulations under each setting are performed 100 times, and when a value is not varying we take defaults of \( N = 200 \), \( m = 2 \) and \( J = 2 \). Although we only include simulations using a fixed number of points, \( m \), per curve here, Section C.2 of the supplementary materials contains additional simulations with varying number of observations per curve. Tables 4, 5 and 6, respectively, are the analogues to Tables 1, 2 and 3 above, showing the MISE of \( \hat{\beta}(t) \) as \( J \), \( N \) and \( m \) increases. In Table 4 we see that, except when \( \beta(t) \equiv 0 \) (i.e. \( w = 0 \)), in which case choosing the smallest number of FPCs clearly makes the most sense, a choice of \( J = 2 \) is otherwise the most reasonable.

Table 4 is noticeably different in contrast to Table 1, where we used the true imputation parameters. Using estimated imputation parameters instead, MISFIT has noticeable MISE across almost all values of \( J \), and in particular does not become unbiased by simply increasing \( J \). This results from bias in estimating the imputation parameters when both \( m \) and \( N \) are small. In these simulations, MISFIT’s advantage is less distinctive, although it still outperforms PACE for all values of \( J \) greater than 1.

We can see in Table 5 that when \( w = 0 \), PACE and MISFIT perform about the same. As the signal increases, however, MISFIT again displays an advantage, regardless of the value of \( N \). In particular, notice that for any sample size the MISE for PACE is about 1.5 times that of MISFIT when \( w = 5 \) or \( w = 10 \).
TABLE 4 Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $J$ increases, for a linear model using estimated imputation parameters

| $J$ | $w = 0$ | | $w = 5$ | | $w = 10$ |
|-----|---------|-------|---------|-------|---------|
|     | MISSFIT | PACE  | MISSFIT | PACE  | MISSFIT | PACE  |
| 1   | 0.01    | 0.00  | 12.37   | 12.36 | 49.47   | 49.45 |
| 2   | 0.09    | 0.11  | 6.23    | 9.83  | 21.14   | 40.55 |
| 3   | 2.50    | 49.41 | 14.53   | 188.20| 29.31   | 594.66|
| 4   | 48.30   | 8573.81| 46.57   | 23,035.97| 57.95 | 97,594.40|
| 5   | 328.17  | 360,195.65| 67.84   | 1,554,176.08| 74.35 | 4,412,004.52|

TABLE 5 Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $N$ increases, for a linear model using estimated imputation parameters

| $N$ | $w = 0$ | | $w = 5$ | | $w = 10$ |
|-----|---------|-------|---------|-------|---------|
|     | MISSFIT | PACE  | MISSFIT | PACE  | MISSFIT | PACE  |
| 100 | 0.20    | 0.21  | 6.94    | 9.35  | 21.89   | 36.72 |
| 200 | 0.09    | 0.11  | 6.26    | 10.04 | 21.20   | 43.84 |
| 400 | 0.07    | 0.06  | 5.50    | 6.47  | 17.33   | 24.78 |
| 800 | 0.02    | 0.02  | 4.67    | 5.56  | 14.69   | 22.13 |

TABLE 6 Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $m$ increases, for a linear model using estimated imputation parameters

| $m$ | $w = 0$ | | $w = 5$ | | $w = 10$ |
|-----|---------|-------|---------|-------|---------|
|     | MISSFIT | PACE  | MISSFIT | PACE  | MISSFIT | PACE  |
| 2   | 0.09    | 0.11  | 6.26    | 9.83  | 21.20   | 40.55 |
| 5   | 0.03    | 0.03  | 4.40    | 5.09  | 15.65   | 20.24 |
| 10  | 0.03    | 0.02  | 3.90    | 4.47  | 14.13   | 17.71 |
| 20  | 0.02    | 0.02  | 3.63    | 4.02  | 13.63   | 15.99 |

Table 6 shows a similar pattern, where the two methods perform similarly in the absence of a signal, but MISFIT overtakes PACE as the signal grows, and especially for smaller values of $m$. In addition, recall that in Table 3, in which the true imputation parameters were used for imputation, the MISE of the MISFIT estimator actually increased slightly as more of each curve was observed. This is no longer the case since now increasing $m$ improves the imputation parameter estimates, leading to a better approximation of the imputation distribution.

3.1.3 Comparison of MISFIT, PACE, pfr and funreg

While the above simulations also compare MISFIT to PACE, another obvious benchmark is the performance of other methods used for performing functional regression. For instance, the
Comparison of median integrated squared error (MISE) when estimating $\beta(t)$ via MISFIT, Pace, pfr and funreg, using estimated imputation parameters

| $w$ | MISFIT PACE Pfr Funreg | MISFIT PACE Pfr Funreg | MISFIT PACE Pfr Funreg |
|-----|-------------------------|-------------------------|-------------------------|
| $m$ |                         |                         |                         |
| 2   | 0.09 0.02 0.50 43,275,170.75 6.26 | 8.13 9.78 38,292,105.75 21.20 32.15 38.43 73,458,117.69 | |
| 5   | 0.03 0.02 0.04 1.66 4.40 4.43 5.24 9.27 | 15.65 17.41 20.69 42.16 | |
| 10  | 0.03 0.02 0.05 0.40 3.90 3.88 5.11 5.59 | 14.13 15.52 19.31 21.54 | |
| 20  | 0.02 0.01 0.03 0.22 3.63 3.73 5.10 3.19 | 13.63 14.83 13.53 5.31 | |

fdapace package (Dai et al., 2019) contains an implementation of PACE which uses local polynomial smoothing to estimate imputation parameters. Other useful benchmarks include pfr of the package refund (Goldsmith et al., 2018), a popular and flexible tool for fitting a wide range of functional regression models, and funreg (Dziak & Shiyko, 2016), which is specifically designed to accommodate functional regression with irregularly spaced designs. The same simulation set-up as described above in Section 3.1.2 was followed and 100 simulations were performed under each setting, using different values of $m$ only.

Table 7 shows the MISE of MISFIT, PACE pfr, and funreg across the different simulation settings. Of primary importance is the comparison of PACE and MISFIT. It is again evident that MISFIT outperforms PACE in nearly every setting and remains competitive where it does not. Also worth highlighting is that the results displayed for PACE in Table 7 are slightly different than those displayed in Table 6 for PACE; these differences are due to the different approaches in estimating the imputation parameters. The performances of pfr and especially funreg are noticeably worse than both PACE and MISFIT when $m$ is as small as 2 or 5. This is unsurprising as neither are by default designed to handle sparse functional data; in fact, funreg is designed specifically for intensive longitudinal data. Also unsurprisingly, then, these two outperform PACE and MISFIT when the signal is strong and each function is observed at least as many as 20 times.

### 3.1.4 Type 1 error rates

As mentioned in Section 2.2, one of the advantages of multiple imputation is that we can incorporate missing data uncertainty into estimation and statistical inference, which is neglected by mean imputation. As such, we would expect mean imputation to produce artificially small standard errors. We substantiate these expectations by simulating data, testing the hypothesis

$H_0 : \beta(t) \equiv 0$

against the alternative

$H_1 : \beta(t) \neq 0$

and comparing rejection rates across imputation methods. If, as we expect, PACE under-represents the standard error, then we should observe higher rejection rates for PACE than for MISFIT (i.e. a gain in power, but also larger type 1 error rates).
We again use estimated parameters and follow the same simulation and estimation procedures outlined above in Section 3.1.2. Hypotheses are tested at the 0.05 nominal significance level, using the statistic

\[ T = ||\beta||^2 \sim \sum_{i} \lambda_i^* \chi^2_i(1), \]

where the \( \lambda_i^* \) are the eigenvalues of \( C_\beta(t, s) \). \( p \)-values are computed using the `imhof()` function of the `CompQuadForm` package. These simulations are run 500 times for each of three different signals, \( w \in \{0, 1, 2\} \), and across all runs \( N = 200 \) and \( J = 2 \) are fixed, while \( m \in \{2, 5, 10, 20\} \). Note that simulations for \( w = 0 \) correspond to simulations under the null hypothesis and thus we would hope for our imputation method to have an empirical rejection rate close to the nominal Type 1 error rate of 0.05. Rejection rates for \( w = 1 \) and \( w = 2 \) correspond to statistical power.

Table 8 displays the average rejection rate for each imputation method. It is interesting to see that PACE performs better under the null hypothesis, while MISFIT is better at detecting true signals. It is also noteworthy that, while PACE appears to be the most calibrated to the nominal rate under the null hypothesis, the empirical rejection rate of MISFIT tends towards the nominal rate as the curves are observed more densely. On the other hand, in the presence of a true signal the rejection rates for MISFIT are higher than for PACE, particularly when the observations are especially sparse.

In summary, even though mean imputation generally leads to higher rejection rates than multiple imputation, MISFIT tends to have higher rejection rates than PACE. This is due to the not-so-obvious result that when the covariate is observed very sparsely, imputing the covariate conditional on the response inflates the type I error rate beyond the nominal rate. This makes sense, although, as the conditional imputation approach incorporates information from the response into the imputed values of the covariate; any false signal detected is merely residue of this process. The takeaway, then, is that one should only use the conditional imputation approach if there is evidence a priori (i.e. before imputation) that the variables are related, or for larger values of \( m \).

### 3.2 Multivariate t distribution

One key assumption of MISFIT’s imputation strategy for linear models is that the functional covariate, \( X(t) \), follows a Gaussian process. To assess MISFIT’s robustness to this assumption, we instead simulate the random sample of functions from a multivariate \( t \) distribution with four degrees of freedom, mean 0, and the same Matérn covariance function used in Section 3.1. For this scenario, we use estimated imputation parameters and follow all other default simulation settings outlined in Section 3.1.2, performing the simulations 100 times. For brevity, we leave

| \( m \) | \( w = 0 \) | \( w = 1 \) | \( w = 2 \) |
|---|---|---|---|
| MISFIT | PACE | MISFIT | PACE | MISFIT | PACE |
| 2 | 0.300 | 0.044 | 0.790 | 0.470 | 0.976 | 0.912 |
| 5 | 0.086 | 0.038 | 0.896 | 0.842 | 1.000 | 1.000 |
| 10 | 0.068 | 0.046 | 0.914 | 0.912 | 1.000 | 1.000 |
| 20 | 0.052 | 0.050 | 0.950 | 0.954 | 1.000 | 1.000 |
TABLE 9  Median integrated squared error (MISE) as $m$ increases, using estimated parameters and sampling the functional covariate from a multivariate $t$ distribution

| $m$ | $w = 0$ | $w = 5$ | $w = 10$ |
|-----|---------|---------|---------|
|     | MISFIT | PACE    | MISFIT | PACE    | MISFIT | PACE    |
| 2   | 0.06    | 0.05    | 5.35   | 7.73    | 19.46  | 31.96  |
| 5   | 0.02    | 0.01    | 3.92   | 5.26    | 14.23  | 20.90  |
| 10  | 0.01    | 0.01    | 3.58   | 4.94    | 13.08  | 19.71  |
| 20  | 0.01    | 0.01    | 3.43   | 4.45    | 13.07  | 17.63  |

$N$ and $J$ fixed and only explore differing values of $m$. Table 9 summarizes the results of these simulations.

Upon inspection, the use of a multivariate $t$ distribution as opposed to a Gaussian process to generate the data yields little difference in terms of estimation performance across all four methods. The MISE results of MISFIT shown in Table 9 are strikingly similar to those shown in Table 6 and the performance of MISFIT relative to PACE seems unchanged as well. This indicates some level of robustness to the assumption of Gaussian data.

3.3 Logistic regression

Finally, a comparison of MISFIT and PACE for logistic regression is especially important because, as discussed in Section 4, we use a logistic regression model to analyse the head circumference data. For brevity, we omit the simulations using the true simulation parameters and only provide results for simulations using estimated parameters.

We begin by simulating $Y_i \overset{iid}{\sim} \text{Bern}(p)$, for $i = 1, \ldots, N$ with $p = 0.5$. Then we simulate $X_i|Y_i = y \overset{iid}{\sim} N(\mu_y, C_X(s, t))$ evaluated on an equally spaced length-100 grid in $[0, 1]$ where $\mu_0(t) \equiv 0$, $\mu_1(t) = v_1(t) + v_2(t)$, and $C_X(t, s)$ is the same Matérn covariance given in Equation (9). We again add noise to the $X_i$ using $\sigma_\epsilon^2 = 0.5$, and randomly sample $m$ of the 100 grid points to observe for each curve. According to Theorem 1, since our choices of $\mu_0$ and $\mu_1$ satisfy $\|C_X^{-1}(\mu_1 - \mu_0)\|_{L_2} < \infty$, we use $\beta(t) = C_X^{-1}(\mu_1 - \mu_0) = \frac{v_1(t)}{\dot{s}_1} + \frac{v_2(t)}{\dot{s}_2}$. Estimation of imputation parameters is performed as in Section 3.1.2, and a logistic regression model is fit using the first $J$ FPCs.

For these simulations, we choose $N = 400$, $m = 2$ and $J = 2$ by default and run 100 simulations for each setting. Note that we do not adjust the signal in these simulations as we did in the simulations for the linear model. This is partially due to brevity and partially due to the inherent instability of logistic regression. We found that simulation results are particularly sensitive to the signal such that when the signal is too large and near-perfect classification can be achieved, both methods perform quite poorly and comparisons are less interesting. Likewise, we increase the default sample size to $N = 400$ for these simulations to insert additional stability. Finally, we show additional simulations for sample sizes of 1600 and 3200 in this section to highlight when MISFIT outperforms PACE.

As in the previous simulations, we first fixed $m$ and $N$ and considered the MISE for different values of $J$. The results are shown in Table 10. Clearly $J = 2$ results in the lowest MISE for both MISFIT and PACE.

Table 11 shows the MISE as $N$ increases from 100 to 3200. Unlike with the linear model, PACE is more accurate than MISFIT for small sample sizes of 100 or 200, but as $N$ grows, MISFIT quickly
Table 10: Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $J$ increases, for a logistic model with estimated imputation parameters

| $J$  | MISFIT | PACE  |
|------|--------|-------|
| 1    | 22.05  | 17.46 |
| 2    | 4.11   | 2.27  |
| 3    | 136.82 | 36.55 |
| 4    | 3720.01| 3786.26|
| 5    | 95,083.42| 170,626.51|
| 6    | 715,298.79| 2,832,174.79|

Table 11: Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $N$ increases, for a logistic model with estimated imputation parameters

| $N$  | MISFIT  | PACE  |
|------|---------|-------|
| 100  | 21.87   | 5.82  |
| 200  | 22.52   | 3.92  |
| 400  | 4.45    | 2.27  |
| 800  | 3.13    | 2.35  |
| 1600 | 1.50    | 2.28  |
| 3200 | 0.76    | 2.51  |

Table 12: Median integrated squared error (MISE) of $\hat{\beta}(t)$ as $m$ increases, for a logistic model with estimated imputation parameters

| $m$  | MISFIT  | PACE  |
|------|---------|-------|
| 2    | 4.45    | 2.27  |
| 5    | 0.96    | 1.11  |
| 10   | 0.62    | 0.44  |
| 20   | 0.27    | 0.24  |

catches up to and surpasses PACE in terms of MISE. It is encouraging that despite observing each curve only twice, if one has a large enough sample size, MISFIT performs quite well. PACE, on the other hand, demonstrates inconsistency, appearing to reach a floor in MISE. Beyond a sample size of $N = 400$, PACE does not improve.

Finally, Table 12 compares the performance of MISFIT and PACE for different values of $m$. First of all, both methods noticeably improve as $m$ increases. MISFIT and PACE are generally comparable across the different values of $m$, although PACE boasts an advantage when $m = 2$. Remember, however, that the sample size is $N = 400$ for all simulations summarized in this table. As we saw in Table 11, MISFIT overtakes PACE for larger sample sizes, so MISFIT should be preferred in very sparse settings as long as one has a relatively large sample (e.g. $N = 1600$ or larger).
Equipped with a new approach to scalar-on-function regression for sparsely and irregularly sampled functional data, we are prepared to revisit the macrocephaly data. One important scientific question that we hope to answer is whether and how children’s head circumference trajectories are associated with their chance of having a pathology related to macrocephaly. We follow the approach outlined in Section 2.3 for imputing the curves in a logistic scalar-on-function regression context. Due to the association between height and head circumference, we use a ratio of head circumference to height instead of raw head circumference. Thus, according to Model 4, $Y_i$ indicates the presence or absence of pathology, $p_i$ the probability of the pathology occurring, and $X_i(t)$ the height-adjusted head circumference at age $t$ of the $i$th subject.

It is worth noting that the properties of coefficient estimators are well-studied in scalar logistic regression models applied to retrospective case–control designs. Even when the cases and controls are sampled retrospectively, as is the case in the present study, a logistic regression model can be used as if the sample was prospective (McCullagh & Nelder, 1989). In such models, the maximum likelihood estimator of the slope coefficient is asymptotically unbiased and, while the intercept is biased, a simple bias correction exists (King & Zeng, 2001; McCullagh & Nelder, 1989; Prentice & Pyke, 1979). Although the functional data analogue has been studied comparatively less, Wei et al. (2014) use a functional logistic regression model in a case–control study and show via simulations that FPC-based estimation of the regression coefficients performs similarly between cohort data and case–control data. Ahmed et al. (2018) develop an estimator for binary response scalar-on-function regression models which accounts specifically for a choice-based, or case–control, sampling design. However, their simulation results show that, specifically in the case of logistic regression, there is little difference between their proposed estimator and an estimator which ignores the sampling design. Finally, although not specifically applied in a logistic regression context, Cardot et al. (2010) also develop FPCA-based estimators which explicitly account for the sampling design.

Due to the prohibitive size of the data set compounded with the rarity of the pathology in the data, we fit a logistic regression model using a subset of $N = 800$ subjects. Specifically, we retained all 85 cases in our sub-sample, and the remaining 715 controls were randomly selected such that the distribution of measurements per child were roughly the same between cases and controls. The subjects in the resulting sub-sample had an average of 6.5 measurements.

For the sake of comparison, we use both MISFIT and PACE to impute the scores, from which estimates of the completed curves, $\{X_i(t) : i = 1, \ldots, 800\}$, are obtained. The same number of FPCs, $J = 3$ (chosen to explain at least 95% of the variance), are used for both approaches. For MISFIT, $K = 10$ imputations were used. The imputed curves are depicted below in Figure 3. The largest differences in the imputations between the two approaches appear towards the end of the age range.

After imputing the scores, a logistic regression model was subsequently fit for both approaches. The estimated coefficient functions are shown in black in Figure 4, along with their 95% pointwise confidence intervals (the black dotted lines). MISFIT’s estimate starts out at a large positive magnitude but quickly decreases to negative values before rising to large positive values again near the end of the domain. PACE, conversely, estimates a much simpler function that starts negative, increases to positive values quickly, then plateaus throughout the rest of the domain.
FIGURE 3  Imputed curves for both imputation approaches: multiple imputation of sparsely sampled functions at irregular times and principal analysis by conditional expectation. Red lines represent cases, black lines represent controls [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 4  Estimated coefficient functions. Solid black lines are the estimated coefficient functions, dotted black lines are the associated 95% pointwise confidence bands. Red lines are the estimated coefficients from the re-sampled data [Colour figure can be viewed at wileyonlinelibrary.com]
TABLE 13  

| MISFIT     | PACE     |
|------------|----------|
| 2.54 x 10^{-4} | 2.06 x 10^{-1} |

One of the main benefits of using a multiple imputation approach is the ability to better estimate uncertainty due to imputation. We expected this to manifest itself in wider confidence bands for a multiple-imputation-based approach, like MISFIT, compared to a mean-imputation-based approach, like PACE. Based on Figure 4, although, this does not appear to be the case. This is most likely because the difference between the standard errors produced by multiple imputation and mean imputation is driven by the between-imputation covariance function, $\hat{B}(t, s)$. If the within-imputation covariance function, $\hat{W}(t, s)$, dominates $\hat{B}(t, s)$, then mean imputation does not miss much uncertainty compared to multiple imputation. That is precisely the case in the macrocephaly data, which is partially why the overall uncertainty estimates are fairly similar between the mean and multiple imputation approaches. Of course, differences also exist due to the fact that MISFIT’s imputation conditions on $Y_i$ while PACE’s imputation does not.

To gauge the suitability of the confidence bands we sampled with replacement from the sub-sample, and repeated the imputation and estimation steps (basically a one-sample bootstrap). The resulting estimates are represented in Figure 4 by the red lines. Had there been more between-imputation uncertainty, we would have expected the new estimates to fall within the corresponding confidence bands for the MISFIT but to exceed the confidence bands for PACE. In both approaches, the new estimate falls entirely within its respective confidence bands, offering no obvious signs of understated uncertainty.

$p$-values for the test of a non-zero effect are presented in Table 13. We expected the $p$-values produced from PACE to be anti-conservative for the reasons given above about uncertainty estimates. In this case, MISFIT yielded a very small $p$-value, providing strong evidence for a non-null functional relationship between height-adjusted head circumference and pathology status. PACE yielded a much larger $p$-value, on the other hand. Based on Figure 4 this is not surprising and can again be explained by two factors: differences in the estimated coefficient functions themselves and the relative size of the within-imputation covariance function (as explained above).

Focusing solely on MISFIT, the left panel of Figure 5 shows a zoomed-in view of the estimated coefficient function. Notice the swift decline from large positive values to negative values early in the domain, followed by a fairly steady increase back to positive values until the end of the domain. Although it is tempting to conclude that the association between the ratio of head circumference to height and the probability of developing a pathology vacillates from largely positive, to negative, and back to positive, this misses the fact that the effect must be interpreted jointly over the entire domain.

To aid our interpretation, we turn to the right panel of Figure 5, which decomposes the coefficient function into two parts: the average effect (which is constant over time) and the total effect less its average. The latter, depicted by the solid line, suggests a contrast between the negative values occurring in much of the first half of the domain, and the positive values occurring in the second half of the domain. Such a contrast can be viewed as an indication that the velocity of head circumference-to-height growth drives the distinction between those with and those without the pathology. Taken together then, two components imply an elevated risk of developing a pathology: larger average head circumferences and unusual head circumference growth rates (both relative to height).
CONCLUSIONS

In this paper, we have introduced a multiple imputation approach, MISFIT, to performing scalar-on-function linear regression in the presence of sparse and irregular functional data. This approach captures the variation due to imputing the functional covariates, thus enabling more reliable statistical inference, and simulations indicate that it yields consistent estimates. We showed that this method can be applied more broadly to other functional generalized linear models by utilizing an appropriate imputation model, and in particular demonstrated its use in a logistic regression setting. Extensive simulations also illustrated the value of MISFIT over existing methods for fitting scalar-on-function regression models when the functional covariate is irregularly and sparsely sampled.

Beyond the general effectiveness of MISFIT, one other takeaway is worth highlighting from the simulations: the accuracy of MISFIT is quite sensitive to the estimates of the imputation parameters, especially in highly sparse and irregular designs. In practice, there are multiple justifiable approaches to estimate these parameters and our results are based on the use of fcr, which employs a spline-based approach. However, we did see some preliminary evidence in favour of the local polynomial smoothing approach used by fdapace. These differences were not explored in depth here, but could warrant further investigation given their practical implications.

Other future work involves extending to more complicated models. It would be beneficial, for instance, to apply the same approach to generalized additive models (GAMs). While the same ideas can be carried over easily enough, establishing a compatible imputation model could prove challenging. Even including multiple functional covariates in the linear or logistic models requires more careful thought than one would hope for such a seemingly direct extension. Lastly, there are many interesting questions surrounding the asymptotic theory, which are unfortunately linked to the methodology used to estimate the imputation parameters. Theoretical work is needed concerning the accuracy of Rubin’s rules and the asymptotic distribution of the
resulting estimators, which is a notoriously difficult question in functional regression, even the dense setting.

ACKNOWLEDGEMENTS
The authors would like to acknowledge the helpful comments from two reviewers and the associate editor, which improved the quality of the paper.

ORCID
Justin Petrovich https://orcid.org/0000-0001-9034-0102
Matthew Reimherr https://orcid.org/0000-0002-7149-0591

REFERENCES
Ahmed, M.-S., Attouch, M.K. & Dabo-Niang, S. (2018) Binary functional linear models under choice-based sampling. Econometrics and Statistics, 7, 134–152.
Berrendero, J.R., Cuevas, A. & Torrecilla, J.L. (2018) On the use of reproducing kernel Hilbert spaces in functional classification. Journal of the American Statistical Association, 113(523), 1210–1218.
Bogachev, V.I. (1998) Gaussian measures, Number 62. Providence, RI: American Mathematical Society.
Brumback, B.A. & Rice, J.A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. Journal of the American Statistical Association, 93(443), 961–976.
Cardot, H., Chaouch, M., Goga, C. & Labruère, C. (2010) Properties of design-based functional principal components analysis. Journal of Statistical Planning and Inference, 140(1), 75–91.
Chiou, J.-M., Zhang, Y.-C., Chen, W.-H. & Chang, C.-W. (2014) A functional data approach to missing value imputation and outlier detection for traffic flow data. Transportmetricsa B: Transport Dynamics, 2(2), 106–129.
Craig, S., Blankenberg, D., Parodi, A., Paul, I., Birch, L., Savage, J., et al. (2018) Child weight gain trajectories linked to oral microbiome composition. Scientific Reports, 8(1), 1–14.
Crambes, C. & Hencrili, Y. (2019) Regression imputation in the functional linear model with missing values in the response. Journal of Statistical Planning and Inference, 201, 103–119.
Dai, X., Müller, H.-G. & Yao, F. (2017) Optimal bayes classifiers for functional data and density ratios. Biometrika, 104(3), 545–560.
Dai, X., Hadjipantelis, P.Z., Han, K. & Ji, H. (2019) fdapace: functional data analysis and empirical dynamics. R package version 0.4.1.
Dauxois, J., Pousse, A. & Romain, Y. (1982) Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. Journal of Multivariate Analysis, 12(1), 136–154.
Daymont, C., Hwang, W.-T., Feudtner, C. & Rubin, D. (2010) Head-circumference distribution in a large primary care network differs from CDC and WHO curves. Pediatrics, 126(4), e836–e842.
Daymont, C., Zabel, M., Feudtner, C. & Rubin, D.M. (2012) The test characteristics of head circumference measurements for pathology associated with head enlargement: a retrospective cohort study. BMC Pediatrics, 12(1), 9.
Daymont, C., Ross, M.E., Russell Localio, A., Fiks, A.G., Wasserman, R.C. & Grundmeier, R.W. (2017) Automated identification of implausible values in growth data from pediatric electronic health records. Journal of the American Medical Informatics Association, 24(6), 1080–1087.
Delaigle, A. & Hall, P. (2012) Achieving near perfect classification for functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(2), 267–286.
Di, C.-Z., Crainiceanu, C.M., Caffo, B.S. & Punjabi, N.M. (2009) Multilevel functional principal component analysis. The Annals of Applied Statistics, 3(1), 458.
Dziak, J. & Shiiko, M. (2016) funreg: functional regression for irregularly timed data. R package version 1.2.
Efron, B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association, 70(352), 892–898.
Ferraty, F., Sued, M. & Vieu, P. (2013) Mean estimation with data missing at random for functional covariables. Statistics, 47(4), 688–706.
Ganguli, I., Simpkin, A.L., Lupo, C., Weissman, A., Mainor, A.J., Orav, E.J. et al. (2019) Cascades of care after incidental findings in a US national survey of physicians. *JAMA Network Open*, 2(10), e1913325.

Goldsmith, J. & Schwartz, J.E. (2017) Variable selection in the functional linear concurrent model. *Statistics in Medicine*, 36(14), 2237–2250.

Goldsmith, J., Greven, S. & Crainiceanu, C. (2013) Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1), 41–51.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J. et al. (2018) *refund*: regression with functional data. R package version 0.1-17.

Greven, S., Crainiceanu, C., Caffo, B. & Reich, D. (2010) Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, 4, 1022–1054.

He, Y., Yucel, R. & Raghunathan, T.E. (2011) A functional multiple imputation approach to incomplete longitudinal data. *Statistics in Medicine*, 30(10), 1137–1156.

James, G.M., Hastie, T.J. & Sugar, C.A. (2000) Principal component models for sparse functional data. *Biometrika*, 87(3), 587–602.

Jenny, C., Hymel, K.P., Ritzen, A., Reinert, S.E. & Hay, T.C. (1999) Analysis of missed cases of abusive head trauma. *JAMA*, 281(7), 621–626.

Kahle, K.T., Kulkarni, A.V., Limbrick Jr., D.D. & Warf, B.C. (2016) Hydrocephalus in children. *The Lancet*, 387(10020), 788–799.

King, G. & Zeng, L. (2001) Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163.

Kowal, D.R., Matteson, D.S. & Ruppert, D. (2019) Functional autoregression for sparsely sampled data. *Journal of Business & Economic Statistics*, 37(1), 97–109.

Kuo, H.-H. (1975) Gaussian measures in banach spaces. In: Dold, A. & Eckmann, B. (Eds.) *Gaussian measures in banach spaces*. Berlin: Springer, pp. 1–109.

Leroux, A., Xiao, L., Crainiceanu, C. & Checkley, W. (2017) Dynamic prediction in functional concurrent regression with an application to child growth. *Statistics in Medicine*, 37(8), 1376–1388.

Leroux, A., Xiao, L., Crainiceanu, C. & Checkly, W. (2018) *fcr*: functional concurrent regression for sparse data. R package version 1.0.

Li, Y. & Hsing, T. (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6), 3321–3351.

Liebl, D. (2019) Inference for sparse and dense functional data with covariate adjustments. *Journal of Multivariate Analysis*, 170, 315–335.

Liu, Y., Nie, Z., Zhou, J., Farnum, M., Narayan, V.A., Wittenberg, G. et al. (2014) Sparse generalized functional linear model for predicting remission status of depression patients. In: Altman, R.B., Hunter, L., Klein, T.E., Murray, T.A., Dunker, A.K. & Ritchie, M.D. (Eds.) *Pacific Symposium on Biocomputing*. Hawaii: World Scientific, pp. 364–375.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*, 2nd edition. London: Chapman & Hall.

Mirshani, A., Reimherr, M. & Slavković, A. (2019) Formal privacy for functional data with Gaussian perturbations. In: Chaudhuri, K. & Salakhutdinov, R. (Eds.) *International Conference on Machine Learning*. PMLR, pp. 4595–4604.

Peng, J. & Müller, H.-G. (2008) Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, 2(3), 1056–1077.

Preda, C., Saporta, G. & Mbarek, M.H.B.H. (2010) The NIPALS algorithm for missing functional data. *Revue Roumaine de Mathématiques Pures et Appliquées*, 55(4), 315–326.

Prentice, R.L. & Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*, 66(3), 403–411.

R Core Team. (2018) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rice, J.A. & Wu, C.O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57(1), 253–259.

Royston, P. (2004) Multiple imputation of missing values. *Stata Journal*, 4(3), 227–241.

Rubin, D.B. (1996) Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434), 473–489.

Rubin, D.B. (2004) *Multiple imputation for nonresponse in surveys*, Volume 81. Boca Raton, FL: John Wiley & Sons.
Schafer, J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), 3–15.
Shang, H.L. (2014) A survey of functional principal component analysis. *ASTA Advances in Statistical Analysis*, 98(2), 121–142.
Shi, M., Weiss, R.E. & Taylor, J.M. (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics*, 45(2), 151–163.
Shang, H.L. (2014) A survey of functional principal component analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444), 1403–1418.
Szczesniak, R.D., Li, D., Duan, L.L., Altaye, M., Miodovnik, M. & Khoury, J.C. (2016) Longitudinal patterns of glycemic control and blood pressure in pregnant women with type 1 diabetes mellitus: phenotypes from functional data analysis. *American Journal of Perinatology*, 33(13), 1282.
Thompson, W.K. & Rosen, O. (2008) A Bayesian model for sparse functional data. *Biometrics*, 64(1), 54–63.
Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and Its Application*, 3, 257–295.
Wei, P., Tang, H. & Li, D. (2014) Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study. *Genetic Epidemiology*, 38(7), 638–651.
Workgroup, Bright Futures Periodicity Schedule, and Committee on Practice and Ambulatory Medicine. (2017) 2017 recommendations for preventive pediatric health care. *Pediatrics*, e20170254.
Wright, C.M. & Emond, A. (2015) Head growth and neurocognitive outcomes. *Pediatrics*, 135(6), e1393–e1398.
Xiao, L., Li, C., Checkley, W. & Crainiceanu, C. (2018) Fast covariance estimation for sparse functional data. *Statistics and Computing*, 28(3), 511–522.
Xiao, L., Li, C., Checkley, W. & Crainiceanu, C. (2019) face: fast covariance estimation for sparse functional data. R package version 0.1-5.
Yao, F., Müller, H.-G. & Wang, J.-L. (2005a) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470), 577–590.
Yao, F., Müller, H.-G. & Wang, J.-L. (2005b) Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6), 2873–2903.
Zahl, S.M. & Wester, K. (2008) Routine measurement of head circumference as a tool for detecting intracranial expansion in infants: what is the gain? A nationwide survey. *Pediatrics*, 121(3), e416–e420.
Zhang, X. & Wang, J.-L. (2016) From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5), 2281–2321.

**SUPPORTING INFORMATION**
Additional supporting information may be found in the online version of the article at the publisher’s website.

---

**How to cite this article:** Petrovich, J., Reimherr, M. & Daymont, C. (2022) Highly irregular functional generalized linear regression with electronic health records. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(4), 806–833. Available from: https://doi.org/10.1111/rssc.12556