Received: 5 August 2020 | Revised: 21 October 2020 | Accepted: 16 November 2020
DOI: 10.1002/sta4.331

WILEY

SPECIAL ISSUE PAPER

Modern multiple imputation with functional data

Aniruddha Rajendra Rao | Matthew Reimherr

Department of Statistics, Pennsylvania State University, State College, Pennsylvania, 16802, USA

Correspondence
Matthew Reimherr, Department of Statistics, Pennsylvania State University, State College, Pennsylvania, 16802, USA.
Email: mreimherr@psu.edu

Present Address
Matthew Reimherr, Department of Statistics, 326 Thomas Building, University Park, PA 16802.

Funding information
NSF SES-1853209

This work considers the problem of fitting functional models with sparsely and irregularly sampled functional data. It overcomes the limitations of the state-of-the-art methods, which face major challenges in the fitting of more complex non-linear models. Currently, many of these models cannot be consistently estimated unless the number of observed points per curve grows sufficiently quickly with the sample size, whereas we show numerically that a modified approach with more modern multiple imputation methods can produce better estimates in general. We also propose a new imputation approach that combines the ideas of *MissForest* with *Local Linear Forest* and compare their performance with *PACE* and several other multivariate multiple imputation methods. This work is motivated by a longitudinal study on smoking cessation, in which the electronic health records (EHR) from Penn State PaTH to Health allow for the collection of a great deal of data, with highly variable sampling. To illustrate our approach, we explore the relation between relapse and diastolic blood pressure. We also consider a variety of simulation schemes with varying levels of sparsity to validate our methods.

KEYWORDS
functional data analysis, functional regression, longitudinal data analysis, missing data, multiple imputation

1 | INTRODUCTION

Functional data analysis (FDA) is a branch of statistics that models the relationship between functions measured over a particular domain, such as time or space (Ferraty & Vieu, 2006; Ferraty & Romain, 2011; Horváth & Kokoszka, 2012; Kokoszka & Reimherr, 2018; Ramsay & Silverman, 1997). There is a rich literature on modelling functions that are densely observed but comparatively less literature on modelling functions that are sparsely observed, which introduce new challenges. Currently, there are very few imputation methods designed for functional data (He, Yucel, & Raghunathan, 2011; James, Hastie, & Sugar, 2000; Rice & Wu, 2001), with a mean imputation procedure commonly known as PACE (Yao, Müller, & Wang, 2005) being the most common.

Single imputation procedures (like mean imputation or PACE) are useful in general but can't account for the uncertainty induced from the imputation procedure; once the imputation is done, analysis then typically proceeds as if the imputed values were the truth. This leads to overly optimistic measures of uncertainty and the potential for substantial bias (Petrovich, Reimherr, & Daymont, 2018). To deal with this and other problems associated with single imputation methods, we consider multiple imputation methods. Multiple imputation involves filling in the missing values multiple times, which creates multiple "complete" data sets. The variability among these complete data sets reflects the uncertainty introduced in the imputation method. Multiple imputation procedures are very versatile and flexible, and they can be used in a wide range of settings. As multiple imputation involves creating multiple predictions for each missing value, the corresponding statistical analysis takes into account the uncertainty in the imputations and hence yields a more reliable standard error. In simple terms, if there is less information in the observed data regarding the missing values, the imputations will be more variable, leading to higher standard errors in the analysis. In contrast, if the observed data are highly predictive of the missing values, the imputations will be more consistent across the multiple imputed data sets, resulting in smaller and more reliable standard errors (Greenland & Finkle, 1995).

Longitudinal studies are amenable to FDA and often contain sparse and irregular samples. Such data can be considered as having missing values, making imputation a natural consideration. Many FDA methods analyze fully or densely observed data sets without any appreciable missing values. However, this is often not the case when dealing with large medical and biological data. Hence, in such cases, we can either apply sparse FDA methods (Kokoszka & Reimherr, 2018; Yao et al. 2005) or use imputation to apply more traditional FDA techniques. Several multiple

imputation techniques have been proposed to impute incomplete multivariate data, including multivariate imputation by chained equations (MICE) (Van Buuren, 2007) and MissForest (MF) (Stekhoven & Búhlmann, 2011). Though these methods have not been directly applied to functional data, they have worked well in general. MICE builds a separate model for each variable (that contains missing values), conditioned on the others, which can be specified based on the data type (continuous, binary, etc.). MF is similar to MICE but uses random forests (FR) for building the conditional models. In both cases, variables are sequentially imputed until convergence is reached. We use the *MICE* (van Buuren & Groothuis-Oudshoorn, 2011) and *missForest* (Stekhoven, 2013) packages in R to implement these methods. Also, local linear forest (LLF) (Friedberg, Tibshirani, Athey, & Wager, 2018), which is a modification of RF, is a powerful regression method. Functional data are naturally smooth, and LLF is equipped to model signals. Taking advantage of this interesting property of LLF, we propose another imputation method similar to that of MF and MICE using LLF.

Several other imputation methods include K-nearest-neighbor (KNN) (Acuna & Rodriguez, 2004), nonparametric imputation by data depth (Mozharovskyi, Josse, & Husson, 2017), substantive model compatible fully conditional specification (Bartlett, Seaman, White, Carpenter, & For the Alzheimer's Disease Neuroimaging Initiative*, 2015), and many more. There have also been studies comparing imputation methods (Ding & Ross, 2012; Liao et al. 2014; Ning & Cheng, 2012; Waljee et al. 2013) under different scenarios and data types, but for functional data, PACE has become the "gold standard." Unfortunately, current methods for imputation in FDA are not designed to handle complex models and do not allow for consistent estimation unless one assumes that the number of observed points per curve grows sufficiently quickly with the sample size. Though this is mathematically convenient, it highlights a serious concern when handling sparse functional data. Most of the methods impute while ignoring the response and subsequent modelling that is to be done with the reconstructed curves, a notable exception being Bayesian methods (Kowal, Matteson, & Ruppert, 2019; Thompson & Rosen, 2008). This leads to biased estimates with unreliable standard errors and misleading p values. For these reasons, PACE, which is executed using the *fdapace* package (Chen et al. 2019), uses an alternative approach to produce consistent estimates for functional linear models that do not generalize to non-linear models.

Missing data as described by Rubin (2004) can be divided into three categories: (1) Missing completely at random (MCAR), in which the missing values are independent of the observed data; (2) 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 (3) missing not at random (MNAR), also known as non-ignorable missing data or structural missing data, in which the missing data patterns depend on the observed and unobserved data. Usually, it is assumed that one is either working with MCAR or MAR to make the problem tractable. We make a similar assumption in our procedure, without formally defining the missing data mechanism. Many of the recent works in functional data imputation (Crambes & Henchiri, 2018; Ferraty, Sued, & Vieu, 2012; He et al. 2011; Preda, Saporta, & Mbarek, 2010) have built upon these ideas and adopted a missing data perspective to tackle various forms of sparsity in functional models. But all these approaches consider either a linear relationship or sparsity in the response, whereas we work with a completely observed response and sparsely observed covariates, where we can have a non-linear relation between them.

In this work, we explore the performance of several modern imputation procedures with functional data. Also, we propose another imputation method using LLF. We demonstrate how a simple modification using binning alongside careful initialization can dramatically improve the imputation and subsequent estimation for both linear and non-linear models. From a missing data perspective, the goal is to do imputation of the missing data in a way that retains the performance of subsequent statistical modelling.

1.1 | PaTH to health

Electronic health record (EHR) or clinical data often require longitudinal statistical methods, which account for the correlation between repeated measurements on the same subject. If one also assumes that these are taken from a smooth curve or data generating process, then we can exploit tools from FDA, which can produce gains in terms of flexibility and statistical power (Carnahan-Craig et al. 2018; Goldsmith & Schwartz, 2017; He et al. 2011). Since hospital visits can occur both infrequently and irregularly, we can't directly apply many FDA techniques to them. They pose a challenge to the current methods as well as for imputation. To address these challenges and illustrate the effectiveness of our approach over the current methods, we wish to apply them to an EHR data set where we predict whether a smoker will relapse or not based on their Blood Pressure (BP) recordings over a span of 18 months.

The Penn State PaTH to Health provides patient data from multiple sources to further scientific discoveries. The data set describes patient-level data variables in a standardized manner (i.e., with the same variable name, attributes, and other metadata) along with information on demographics, encounters, diagnoses, and procedures. More information can be found on their website (https://ctsi.psu.edu/research-support/path/).

While there is a wealth of research related to smoking and blood pressure (BP) (Hansson, Hedner, & Jern, 1996; Primatesta, Falaschetti, Gupta, Marmot, & Poulter, 2001; Wang et al. 2018), our goal here is not to make deep scientific statements, but rather to illustrate the utility of our methods, which we hope will prove useful to practitioners. The highest risk of relapse for smokers is during the end of their first and second year after quitting (García-Rodríguez et al. 2013; Herd, Borland, & Hyland, 2009). We therefore focus on modelling the relapse of the patients based on monitoring their BP within the first 2 years, which may be useful to the practitioners designing interventions for patients at risk of relapse.

In general, EHR data sets vary significantly with the timing and regularity of the appointments, and clinical measurements are affected by errors of varying types and degrees (Daymont et al. 2017). Similar challenges apply to measurements in the PaTH data set where we have various kinds of clinical measurements and information recorded. The ability to characterize trajectories of sparse irregular data has potential applicability to many clinical questions. Though the term sparsity is somewhat subjective in the context of functional/longitudinal data, many of

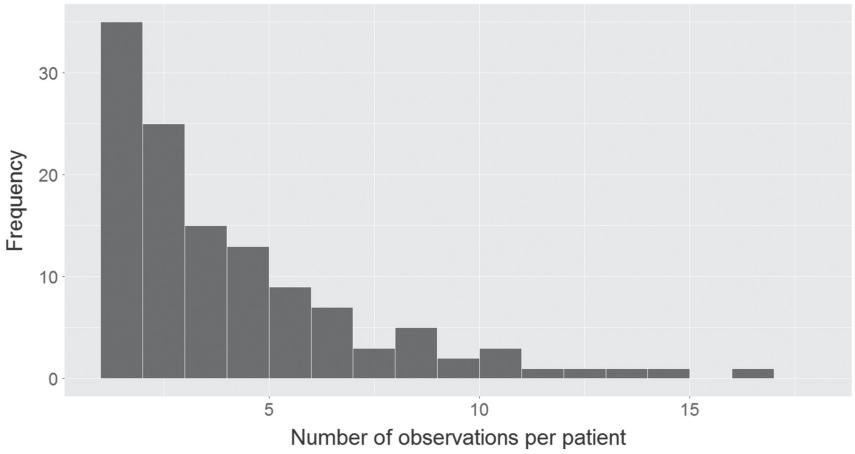

FIGURE 1 Histogram of the number of observations for BP per patient, ranging from 1 to 18

the patients in the Path data set have just two measurements, while the greatest number of measurements for any patient is 17 (after cleaning and implementing the exclusion criteria). We can see from Figure 1 the modal number of measurements is 2, while relatively few patients had more than seven clinical visits and almost none had more than 11. Also, from the cumulative observation (Figure A1), we observe that 94% of the patients had 10 or fewer measurements, 72% had at most five measurements, and 28% had no more than two measurements. On average, we have around four measurements per patient. Sparsity arises due to many reasons in an EHR setting. Some patients never come back, some patients are not observed with any uniformity or regularity, etc. Having identified the BP trajectories as both sparse and irregular, we move to introduce the imputation methods that account for these conditions in a functional framework before revisiting this data in Section 3.

1.2 | Organization

The rest of the paper is organized as follows. In Section 2, we briefly go through PACE and multivariate imputation methods (MICE and MF) before introducing our proposed method using LLF, and modifications to the multivariate imputation methods using bins and careful initialization, to better deal with functional data. We present multiple simulations for the linear and non-linear cases under different values of sparsity in Section 3. This section also includes the EHR data, where we fit a scalar-on-function regression model to determine if a patient (smoker) will relapse or not at the end of 18 months using BP as the functional predictor. These examples help us to illustrate the limitations of previous approaches and demonstrate the usefulness of our methodology, which overcomes many of the issues discussed earlier. In Section 4, we present our concluding remarks and future research directions, which pertain to better understanding of the bins and deeper statistical theory.

2 | METHODS

In this section, we give details of the current imputation methods for the scalar-on-function regression model. In Section 2.1, we define the necessary notation used in the paper. In Section 2.2, we briefly discuss scalar-on-function regression models. Section 2.3 gives an overview of PACE and the multivariate imputation method MICE in detail along with their shortcomings. In Section 2.4, we discuss LLF and the multivariate imputation method MF. We present our new imputation procedure that extends the ideas of MF to LLF, as well as discuss how to use careful binning and initialization to improve performance.

2.1 | Setup and notation

We assume the data are collected from trajectories that are independent realizations of a smooth random function, with unknown mean function $E(X(t))=\mu(t)$ and covariance function $C_X(t,s) := cov(X(s), X(t))$. We define 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. We assume that these curves are only observed at times t_{ij} $(j = 1, \ldots, m_i)$ with some error:

$$x_{ij} = X_i\left(t_{ij}\right) + \delta_{ij}.$$

Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^\top$ denote the vector of observed values of the function X_i. Let Y_i, be the outcome, which is a function of X_i and some error. Examples of such relations can be found in the next subsection.

Generally, when integration is written without limits, it is implied to be over the entire domain, usually standardized to $[0, 1]$ for simplicity. The main focus of this work is to develop tools for consistently estimating the parameters in functional regression models.

2.2 | Functional models

The scalar-on-function regression model in the linear case is defined as

$$Y_i = \alpha + \int \beta(t) X_i(t) dt + \epsilon_i. \tag{1}$$

A common way of estimating the model components is by basis or functional principal component (FPC) expansions, where we simplify the problem of estimating the parameters by projecting the functions to a finite dimension and then using multiple regression and least squares (Kokoszka & Reimherr, 2018; Ramsay & Silverman, 1997).

Non-linear modelling becomes very challenging with functional data due to *the curse of dimensionality*. One popular way of simplifying the problem is by using the generalized additive model, also known as the continuously additive model (Fan, James, & Radchenko, 2015; Ma & Zhu, 2016; McLean, Hooker, Staicu, Scheipl, & Ruppert, 2014; Müller, Wu, & Yao, 2013; Wang & Ruppert, 2015):

$$Y_i = \int f(X_i(t), t) dt + \epsilon_i, \tag{2}$$

where the bivariate function $(x, t) \to f(x, t)$ is smooth but unknown. It is commonly estimated using either basis expansions (Greven, Crainiceanu, Caffo, & Reich, 2010; Yao et al. 2005) (often with a tensor product basis) or using Reproducing Kernel Hilbert Spaces (Reimherr, Sriperumbudur, & Taoufik, 2017; Wang & Ruppert, 2015). McLean et al. (2014) developed Functional Generalized Additive Models that enabled non-scalar response mapping.

Most methods for fitting the models discussed above require densely sampled functional data. For irregular and sparsely sampled data that are observed with error, estimating the modelling parameters becomes much more challenging. Directly smoothing the x_{ij} to plug into a dense estimation framework seems like a straightforward idea but can result in substantial bias.

2.3 | Imputation methods

PACE (Yao et al. 2005) uses functional principal components (FPC) analysis, in which the FPC scores are imputed using conditional expectations. In addition to the requirements discussed previously, PACE also relies heavily on the data being Gaussian. The Karhunen-Loéve or principal component expansion of $X_i(t)$ is given by

$$X_i(t) = \mu_X(t) + \sum_{j=1}^{\infty} \xi_{ij} v_j(t), \tag{3}$$

where $v_j(t)$ are the eigenfunctions of C_X with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. The scores are computed as

$$\xi_{ij} = \left\langle X_i - \mu_X, v_j \right\rangle. \tag{4}$$

PACE proceeds by computing the conditional expectation of the scores given the observed data. This conditioning method is straightforward and tends to work much better than direct smoothing of x_{ij}. It provides the best linear unbiased predictors (BLUPs) under Gaussian assumptions and works in the presence of both measurement errors and sparsity. We can plug the BLUPs values into a dense estimation framework to model the response.

PACE still suffers from a few major problems. One issue is that the imputation procedure of PACE does not consider the response Y_i nor does it have any consideration for subsequent models that will be fit. This results in a bias while estimating model parameters (Petrovich et al. 2018). In addition, PACE is just a single imputation method and hence the uncertainty in the imputation is not properly propagated when forming confidence intervals, prediction intervals, or p values. For this reason, the PACE software (Chen et al. 2019) uses an alternative approach for fitting linear models which does not extend to non-linear models.

After understanding PACE, we now look into some of the standard methods used for imputation in the multivariate case. All of these methods have proven to work well in the multivariate setting but have never been tested in the functional setting, where the sample sparsity can be very high.

MICE (Van Buuren, 2007), also known as ''fully conditional specification'' or ''sequential regression multiple imputation,'' has emerged in the statistical literature as one of the principal methods for addressing missing data. MICE performs multiple imputations rather than single imputation, and hence it can account for the statistical uncertainty. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or categorical) as well as complexities such as bounds or survey skip patterns. At a high level, the MICE procedure is a series of models whereby each variable with missing data is modelled conditional upon the other variables in the data. The MICE procedure is as follows: (1) We start by initialization, wherein we fill in all the missing values with mean imputation. (2) Next, we select the first variable with missing entries. A model is fit with this variable as the outcome and the other variables as predictors. (3) The missing values of the current variable are then replaced with predicted (imputed) values from the model in step 2. (4) Step 2 and step 3 are repeated while rotating through the variables with missing values sequentially. The cycling through each of the variables constitutes one iteration or cycle. (5) At the end

of one cycle, all of the missing values have been replaced with predictions from the model that reflect the relationships observed in the data. The cycles are repeated a few times and after each cycle the imputed values are updated.

The number of cycles to be performed is pre-specified and after the last cycle, the final imputations are retained, resulting in one imputed data set. Generally, 10–15 cycles are performed. The idea is that, by the end of the cycles, the distribution of the parameters governing the imputations (e.g., the coefficients in the regression models) should have converged, in the sense of becoming stable. Different MICE software packages vary somewhat in their exact implementation of this algorithm (e.g., in the order in which the variables are imputed), but the general strategy is the same. Here, we have used the *MICE* (van Buuren & Groothuis-Oudshoorn, 2011) package in R.

A key advantage of MICE is its flexibility in using different models. Generally, the modelling techniques included in MICE are predictive mean matching, linear regression, generalized linear models, Bayesian methods, RF, linear discriminant analysis, and many more. Its primary disadvantage is that it does not have the same theoretical justification as other imputation methods. In particular, fitting a series of conditional distributions, as is done using the series of regression models, may not be consistent with proper joint distribution, though some research suggests that this may not be a large issue in applied settings (Schafer & Graham, 2002).

2.4 | Our approach

2.4.1 | MF and LLF

MF (Stekhoven & Búhlmann, 2011) is a multiple imputation method, which proceeds by training an RF on the observed parts of the data. RF (Breiman, 2001) is a non-parametric method that is able to deal with mixed data types as well as allow for interactive and non-linear effects. MF addresses the missing data problem using an iterative imputation scheme by training a RF on observed values in the first step, followed by predicting the missing values in the next step and then proceeding iteratively. RF works well in high dimensional cases with good accuracy and robustness. Though the idea of MF is similar to MICE, they differ in the ordering scheme of the columns to be imputed, and MICE requires certain assumptions about the distribution of the data or subsets of the variables which may or may not be true.

For an arbitrary variable p in $X_{n\times m}$ ($p = 1, 2, \dots m$) including missing values at entries $\mathbf{i}^{(p)}\text{mis} \subseteq \{1, \dots, n\}$, we can separate the data set into four parts: The observed values for variable p, denoted by $\mathbf{y}_{obs}^{(p)}$; the missing values for variable p, denoted by $\mathbf{y}_{mis}^{(p)}$; the variables other than p with observations at $\mathbf{i}^{(p)}\text{obs} \subseteq \{1, \dots, n\}$ / $\mathbf{i}_{mis}^{(p)}$, denoted by $\mathbf{x}_{obs}^{(p)}$; and the variables other than p with observations at $\mathbf{i}_{mis}^{(p)}$, denoted by $\mathbf{x}_{mis}^{(p)}$.

The approach is as follows: We initialize the missing values in X using mean imputation or another imputation method. We then sort the variables p in X ($p = 1, \dots, m$) in ascending order of the missing values. For each variable p, the missing values are imputed by first fitting an RF with response $\mathbf{y}_{obs}^{(p)}$ and predictors $\mathbf{x}_{obs}^{(p)}$, then predicting the missing values $\mathbf{y}_{mis}^{(p)}$ by applying the trained RF to $\mathbf{x}_{mis}^{(p)}$. We sequentially do this for all variables with missing values; that is one cycle. The imputation procedure is repeated for multiple cycles until a stopping criterion is met.

The advantage of MF is that it can deal with any kind of data. Also, MF is straightforward, as it does not need any tuning of parameters nor does it require any assumption about distributional aspects of the data. The full potential of MF is deployed when the data include complex interactions or non-linear relations between variables of different types, which is not possible with PACE. Furthermore, MF can be applied to high-dimensional data sets with a low sample size and still provide excellent results. MF often outperforms other methods in terms of imputation (Stekhoven & Búhlmann, 2011), but the method has no smoothing mechanism and hence the imputed values of the curves are not smooth. To deal with this and to increase model accuracy, we integrate binning into the method as explained in the next subsection.

LLF (Friedberg et al. 2018) uses an RF to generate weights that are used as a kernel for local linear regression, i.e., LLF takes the RF weights $\alpha_i(\mathbf{x}_i)$ and uses them to solve

$$\min_{\mu,\beta} \sum_{i=1}^{n} \left(Y_i - \mu - (x - x_i)'\beta\right)^2 \alpha_i(\mathbf{x}_i). \tag{5}$$

The RF weights $\alpha_i(\mathbf{x}_i)$ are found with the help of the leaf $L_b(x_i)$ in each tree T_b in a forest of B trees as follows:

$$\alpha_i(\mathbf{x}_i) = \frac{1}{B}\sum_{b=1}^{B} \frac{1\{\mathbf{X}_i \in L_b(\mathbf{x}_i)\}}{|L_b(\mathbf{x}_i)|}, \tag{6}$$

where $\sum_{i=1}^{n} \alpha_i(\mathbf{x}_i) = 1$ and for each $i, 0 \le \alpha_i(\mathbf{x}_i) \le 1$. Athey, Tibshirani, and Wager (2016) used this perspective to harness RF for solving weighted estimating equations and gave asymptotic guarantees on the resulting predictions. With the help of the above weights, LLF solves the locally weighted least squares problem.

LLF is a modification of local linear regression with the help of RF, equipped to model signals and fix bias issues. We use this to our advantage and propose a modification to the MF method, where we replace the RF with LLF. We refer to this as miss local linear forest (MLLF). This new approach using LLF for imputation follows the same steps as MF but instead of using RF as the modelling technique to impute the missing values, we will now use LLF. Since LLF does not inherit the multiple imputation like MF, we generate multiple imputed sets and take an average like in the case of MICE. MLLF has similar benefits to MF. We update the MF code (Stekhoven, 2013) in R using the *grf* package (Tibshirani, Athey, & Wager, 2020) to implement MLLF.

2.4.2 | Adapting methods to functional data

One of the key features of functional data is the smoothness of the underlying curves. MF and MLLF produce well-imputed curves but are not smooth. We overcome this problem with the help of binning and careful initialization. We improve the initialization by using PACE instead of simple mean imputation. These boosted methods using PACE are denoted as MFP for MF and MLLFP for LLF. While this leads to higher performance in general with slightly smoother imputed curves, it does not directly smooth the imputed curves or resulting model parameters. Also, this initialization comes with a computational burden as PACE itself is computationally heavy. Another restriction which all of these imputation methods have, except PACE, is that they need to pass through the observed points, which need not be optimal, especially in the presence of observation noise.

We overcome the non-smoothness issue and computational problem by the use of bins. Binning (also known as discrete binning or bucketing) is a data pre-processing method that is used to reduce the effects of minor observation errors and smooth the data. The original data values which fall in a given small interval, a bin, are replaced by a value representative for that interval, usually the mean value. Binning aggregates the values into a fixed range. We divide the desired grid of the time points into k bins and impute over the k points before interpolating back to m time points using b-splines. Here, k is a tuning parameter that acts much like a bandwidth in kernel smoothing. This not only leads to smoother imputation results but also improves the subsequent modelling. As binning helps in reducing the number of time points ($m < k$), the overall process becomes much more computation-friendly. The way bins are defined is as follows:

- The first bin is the first time point of the data.
- The last bin is the last time point of the data.
- The middle (k-2) bins are divided into equal parts and are represented by the mean of all values in that bin.

We denote the methods with the binning as MF_B for MF and MLLF_B for LLF. If we add PACE initialization to it then we denote the methods as MFP_B for MF and MLLFP_B for LLF. We can see in the next section that this not only leads to much smoother results but also improves imputations and modelling performance.

3 | SIMULATION AND RESULTS

Throughout this section, we refer to MissForest as MF, MF with PACE initialization as MFP, binning without PACE as MF_B, and binning with PACE as MFP_B. MLLF uses an analogous naming scheme. In addition to comparing all the methods in simulations, we will compare the results for the EHR data as well. In the simulation, we compare them in both linear and non-linear scalar-on-function regression settings with a scalar and binary response, investigating the imputation accuracy, model fit (prediction accuracy), and β estimates (only for the linear case). We compare across multiple simulated data sets with varying time points observed m, sample sizes n, and sparsity s.

3.1 | Simulation

Linear case:

For the linear case, we simulate n iid random curves $\{X_1(t), \dots, X_n(t)\}$ from a Gaussian process with mean 0 and covariance

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

which is the Matérn covariance function, and K_ν is the modified Bessel function of the second kind. We set $\rho = 0.5$, $\nu = 5/2$ and $\sigma^2 = 1$. These curves are evaluated at m equally-spaced time points from $[0, 1]$. We assume that each observed point contains a normal measurement error with mean zero and variance $\sigma_\delta^2 = 0.3$. We set $\beta(t) = w \times \sin(2\pi t)$, where w is a weight coefficient used to adjust the signal. The response, Y_i ($i = 1, \dots, n$), is computed using the model in Equation (1), where $\alpha = 0$ and $\sigma_\epsilon^2 = 1$. In the binary response case, we define Y_i ($i = 1, \dots, n$) using the Bernoulli and logit link function in Equation (1). Finally, for each curve, we assume a percentage (s) of the m time points is unobserved. After the scores are imputed, we fit a scalar-on-function regression model using these imputed curves.

For the linear case, we simulate the data sets of different sample sizes, $n \in \{200, 500, 1000\}$ (results for $n = 200$ and $n = 1000$ are included in the appendix); different numbers of observations per curve, $m \in \{32, 52\}$; and different sparsity levels, $s \in \{Medium, High\}$. For sparsity levels, medium means 50% of the points are missing for each curve and high means more than 85% of the points are missing for each curve. Also, the values of m are taken such that they help with the process of binning. Each of these settings is simulated 10 times. Since we are primarily interested in the accuracy of the final estimates $\hat{\beta}(t)$, $\hat{Y}(t)$, and $\hat{X}(t)$, we report the root mean square error (RMSE), or prediction error, for each of them.

Tables 1 and 2 indicate that, in general, irrespective of the number of points, all the binned methods perform better compared to other methods for prediction error or RMSE of prediction, β coefficients, and imputation when the sparsity is medium and the sample size is 500. Also, we can see that the RMSE of imputation for PACE is the same for both the tables. This is because we are using the same sample curves to generate scalar and binary response. There is no clear winner between MF and MLLF within the binned methods with or without PACE initialization. Again, for high sparsity, we notice similar behavior as before: Irrespective of the number of points, all the binned methods perform better.

TABLE 1 RMSE of prediction, β coefficients, and imputation of the curves for different methods under linear case when $n = 500$ for different time points and sparsity settings

	$n = 500$, **s = medium**						$n = 500$, **s = high**					
	$m = 32, b = 17$			$m = 52, b = 27$			$m = 32, b = 8$			$m = 52, b = 12$		
Method	**Pred**	β	**Imp**	**Pred**	β	**Imp**	**Pred**	β	**Imp**	**Pred**	β	**Imp**
MF	0.136	0.155	0.108	0.227	0.241	**0.077**	0.267	0.422	0.470	0.177	0.307	0.390
PACE	0.170	0.208	0.199	0.484	0.595	1.910	0.376	0.432	**0.369**	0.350	0.388	0.308
MLLF	0.142	**0.149**	**0.070**	0.234	0.242	**0.023**	58.010	43.721	0.601	69.24	55.43	0.508
MICE	3.611	3.612	0.090	0.237	0.253	0.089	9.840	5.950	0.910	3.162	2.390	1.073
MFP	**0.122**	**0.144**	0.105	0.228	0.250	0.130	0.141	**0.289**	**0.398**	0.232	0.310	0.328
MLLFP	0.132	0.153	0.144	0.232	0.246	0.100	0.376	0.432	**0.364**	0.384	0.386	0.302
MF_B	**0.122**	**0.136**	**0.079**	**0.173**	**0.180**	**0.052**	**0.117**	**0.264**	**0.334**	**0.152**	**0.217**	**0.261**
MLLF_B	**0.126**	**0.137**	**0.082**	**0.174**	**0.177**	**0.053**	**0.097**	**0.291**	**0.338**	**0.132**	**0.203**	**0.267**
MFP_B	**0.122**	**0.138**	**0.053**	**0.176**	**0.182**	**0.023**	**0.101**	**0.263**	**0.339**	**0.146**	**0.219**	**0.275**
MLLFP_B	**0.128**	**0.143**	**0.059**	**0.179**	**0.178**	**0.023**	**0.091**	0.817	0.614	**0.129**	**0.214**	0.307

Note: Bold values in the table indicate the best values.

TABLE 2 Prediction error, RMSE of β coefficients, and RMSE of imputation of the curves for different methods under linear case with binary response when $n = 500$ for different time points and sparsity settings

	$n = 500$, **s = medium**						$n = 500$, **s = high**					
	$m = 32, b = 17$			$m = 52, b = 7$			$m = 32, b = 17$			$m = 52, b = 12$		
Method	**Pred**	β	**Imp**	**Pred**	β	**Imp**	**Pred**	β	**Imp**	**Pred**	β	**Imp**
MF	0.268	0.409	**0.043**	**0.280**	0.391	0.155	0.317	0.383	0.187	**0.251**	0.347	0.307
PACE	0.270	0.508	0.199	0.431	0.426	1.910	0.348	0.463	0.369	0.466	0.687	0.308
MLLF	0.364	0.453	**0.040**	0.532	1.376	0.212	0.282	0.490	0.309	0.374	2.768	1.715
MICE	0.459	1.282	0.284	0.589	0.385	0.934	0.543	4.828	1.115	0.672	1.814	1.050
MFP	0.260	0.385	0.198	**0.282**	0.340	0.198	0.258	0.356	0.185	**0.260**	0.360	**0.282**
MLLFP	0.274	0.417	**0.038**	0.364	0.379	0.213	0.360	0.462	0.307	0.366	0.372	0.651
MF_B	**0.160**	**0.262**	**0.036**	**0.276**	**0.322**	**0.132**	**0.214**	0.329	**0.174**	**0.250**	**0.290**	**0.261**
MLLF_B	**0.161**	**0.235**	**0.037**	**0.264**	0.378	**0.132**	**0.232**	0.346	**0.169**	**0.246**	**0.271**	**0.274**
MFP_B	**0.161**	**0.249**	**0.037**	**0.266**	**0.321**	**0.121**	**0.220**	**0.309**	**0.179**	**0.244**	**0.296**	**0.275**
MLLFP_B	**0.169**	**0.270**	**0.036**	**0.274**	0.357	**0.121**	**0.236**	0.467	0.250	**0.252**	**0.281**	0.304

Note: Bold values in the table indicate the best values.

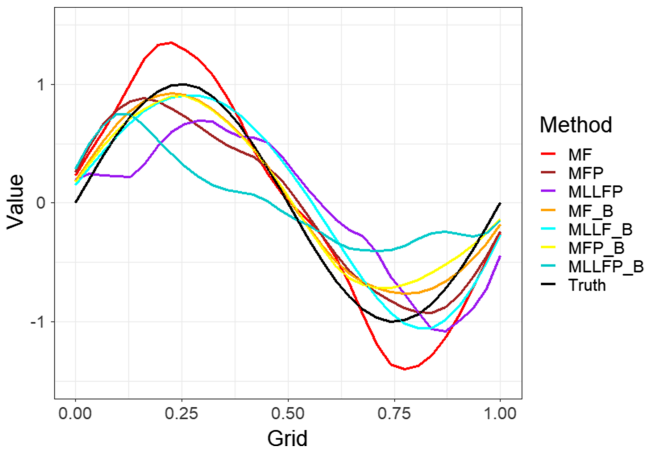

FIGURE 2 Estimated coefficient function for different methods under linear case with sample size ($n = 500$), time points ($m = 52$), sparsity (s = High), and scalar response

As we increase the sparsity, it seems like MICE and MLLF perform worse. This happens mainly because they do a poor job of imputing the curves, the effects of which get compounded when estimating the parameters and modelling. Again, binned methods are the best with no clear winner. The results for the other cases with different sample sizes and scalar response yield similar performance to these tables and can be found in the Table A1 for $n = 200$ and Table A2 for $n = 1000$.

We can see from Figure 2 how each method does in estimating the β coefficient. We can observe that most of the MF extensions are catching the right shape and doing a good job. The plot does not contain PACE, MICE, and MLLF, as their estimates were very poor, which is also reflected in the RMSE for the β coefficients from Table 1.

Figure 3 shows an example of imputed curves under different methods for one random sample curve. Here, binning helps in not only doing better imputation, as seen from Table 1, but also giving much smoother results as compared to the MF methods without the binning. The same effect can be seen with LLF methods as well. This plot is included in the Figure A2.

Non-linear case:

All simulation parameters are the same as before, except the response Y_i $(i = 1, \ldots, n)$ is computed using the model in Equation (2), where $f(X_i(t), t) = 5 * sin(X(t)^2 * t^2)$. For the non-linear case, we also simulate data sets of sample size n=500 with different numbers of observations per curve, $m \in \{32, 52\}$, and with different sparsities, $s \in \{medium, high\}$. Each of these settings is simulated 10 times. Since we are primarily interested in the accuracy of the final output $\hat{Y}(t)$ and $\hat{X}(t)$, we report the RMSE, or prediction error, for each of them. Another non-linear model result can be found in Table A3.

From Tables 3 and 4, we observe that our proposed approaches are outperforming PACE and MICE for imputation irrespective of the number of points and sparsity. When it comes to prediction, our methods are still better than PACE and MICE but the gap is not as large compared to the linear case. Overall, we see the same trend as in the linear case, with the binned methods outperforming every other method under various simulation settings.

The major takeaway from all the simulations is that our methods perform the best under various settings. This is because our methods impute smoother curves, resulting in better modelling and smoother estimates of the beta coefficients, irrespective of the relation between the response and the functional predictors.

3.2 | Electronic health records

New statistical tools are vital for data such as PATH, which are very large and have a great deal of underlying structure. We see the performance of the developed tools for imputation with this longitudinal/functional data. The electronic medical records contain information about smokers (patients) who irregularly come for a check-up at the hospital. We have the BP readings along with some other measurement values at each check-up of the patients. From previous studies, we know that the majority of the relapse among smokers occurs within the first 2 years. For cleaning the data, the exclusion criteria were based on the number of longitudinal measurements (time points). Patients who had a smoking history (smoked for at least a year) with fewer than two measurements were excluded. After cleaning, we are left with 122 patients, of whom 61 smokers relapsed and 61 smokers did not, where each smoker is under observation for 18 months. Hence, here the sample size (n) is 122 and the number of time points (m) is 18. The data is sparse naturally, as the patients don't come in for check-ups regularly, and sometimes the

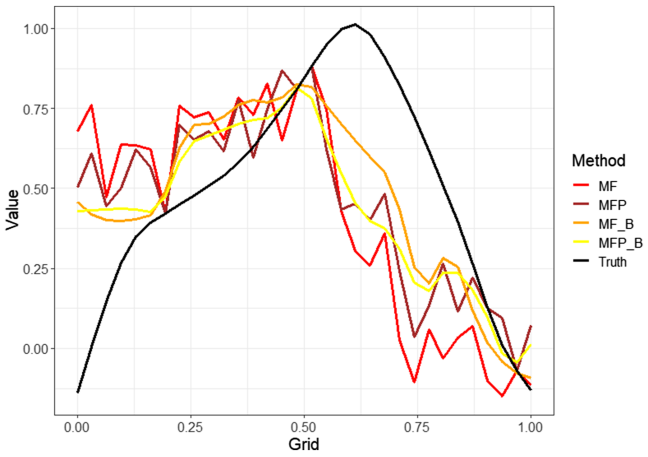

FIGURE 3 Comparing imputed curves in non-binned and binned methods of MF under the linear case for one random sample curve with time points ($m = 52$) and sparsity (s = high)

TABLE 3 RMSE of prediction and imputation of the curves for different methods under the non-linear case ($f(X_i(t), t) = 5 * sin(X(t)^2 * t^2)$) when $n = 500$ for different time points and sparsity settings

	$n = 500$, **s = medium**				$n = 500$, **s = high**			
	$m = 32, b = 17$		$m = 52, b = 27$		$m = 32, b = 8$		$m = 52, b = 12$	
Method	**Pred**	**Imp**	**Pred**	**Imp**	**Pred**	**Imp**	**Pred**	**Imp**
MF	0.236	0.102	0.223	0.078	0.303	0.428	0.355	0.381
PACE	0.328	0.238	0.214	1.253	0.431	0.659	0.386	0.551
MLLF	0.230	0.066	0.209	0.064	0.419	0.592	0.351	0.482
MICE	0.334	0.812	0.214	0.680	0.652	0.892	0.644	1.020
MFP	0.235	0.980	0.210	0.075	0.334	0.379	0.351	0.324
MLLFP	0.276	**0.044**	0.560	**0.027**	0.427	0.357	0.342	**0.259**
MF_B	**0.174**	**0.045**	**0.161**	**0.053**	**0.293**	**0.257**	**0.318**	**0.275**
MLLF_B	**0.183**	**0.045**	**0.163**	**0.054**	**0.335**	**0.364**	**0.312**	**0.282**
MFP_B	**0.176**	**0.031**	**0.160**	**0.026**	**0.290**	**0.257**	**0.311**	**0.251**
MLLFP_B	**0.174**	**0.031**	**0.163**	**0.028**	**0.328**	0.385	**0.329**	0.308

Note: Bold values in the table indicate the best values.

	$n = 500$, s = medium				$n = 500$, s = high			
	$m = 32, b = 17$		$m = 52, b = 12$		$m = 32, b = 17$		$m = 52, b = 12$	
Method	**Pred**	**Imp**	**Pred**	**Imp**	**Pred**	**Imp**	**Pred**	**Imp**
MF	0.388	0.157	0.390	0.296	0.478	0.522	0.306	0.402
PACE	0.405	0.238	0.356	1.253	0.589	0.659	0.411	0.551
MLLF	0.386	0.144	0.392	0.274	0.486	2.790	0.382	1.490
MICE	0.451	0.154	0.388	0.282	0.641	1.262	0.712	2.139
MFP	0.388	0.213	0.292	0.238	0.492	0.573	**0.300**	0.708
MLLFP	0.382	0.268	0.288	0.276	**0.290**	0.413	0.402	1.401
MF_B	**0.296**	**0.112**	**0.262**	**0.232**	**0.308**	**0.374**	**0.288**	**0.372**
MLLF_B	**0.294**	**0.113**	**0.262**	**0.232**	**0.294**	**0.369**	**0.292**	**0.372**
MFP_B	**0.294**	**0.091**	**0.262**	**0.221**	**0.302**	**0.379**	**0.290**	**0.366**
MLLFP_B	**0.294**	**0.090**	**0.262**	**0.221**	**0.280**	0.550	**0.292**	**0.370**

TABLE 4 Prediction error and RMSE of imputation of the curves for different methods under the non-linear case with binary response ($f(X_i(t), t) = 5 * sin(X(t)^2 * t^2)$) when $n = 500$ for different time points and sparsity settings

Note: Bold values in the table indicate the best values.

Method	**MF**	**PACE**	**MLLF**	**MICE**	**MFP**	**MLLFP**	**MF_B**	**MLLF_B**	**MFP_B**	**MLLFP_B**
Linear model	0.39	0.42	0.39	0.39	0.36	0.37	**0.33**	0.35	**0.32**	0.35
CAM	0.36	0.39	0.36	0.38	0.35	0.36	**0.28**	0.32	**0.28**	0.31

TABLE 5 Prediction error of different methods for the EHR data

Note: Bold values in the table indicate the best values.

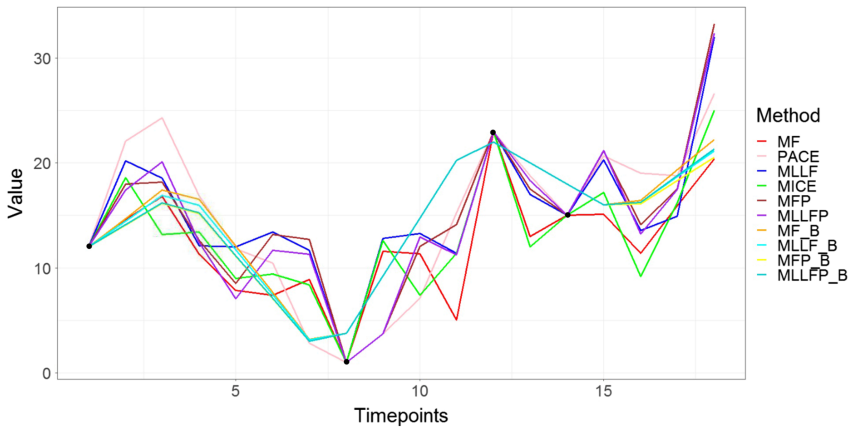

FIGURE 4 Imputation results for one curve from the EHR data

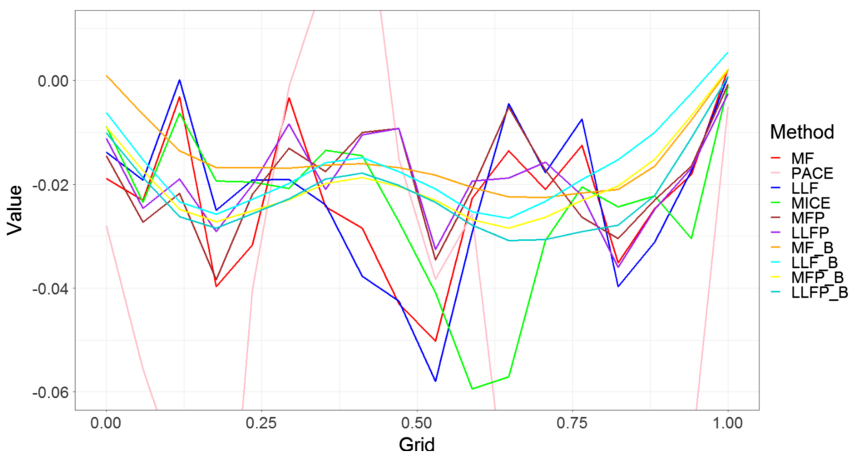

FIGURE 5 Estimated coefficient function of all the methods for the EHR data using linear scalar-on-function regression

measurements are missing even though there was a visit, due to unknown reasons. We build a model to predict whether a patient will relapse or not using the BP measurement over an 18 month period.

We can infer from Table 5 that the binned methods again are outperforming the other methods in prediction, irrespective of linear or non-linear modelling. Also, since both the model results are so close, the true relation looks linear. This is further supported by Figure 4, where we see only the binned methods have smooth imputed curves, with the black points denoting the observed values for that patient. Although all the estimated β coefficients seem to follow the same trend in Figure 5, the methods with binning have much smoother results, leading to better interpretability. Smoothness is inherent to functional data and that is why binned methods are able to perform so well.

Also, Figure 5 suggests that patients with low BP or sudden changes in their BP have a higher risk of relapse. Also, the curve isn't constant, suggesting that acceleration/velocity of the BP curve is important. We did check and found out that the average BP was higher in the control group (no relapse) than for the cases (relapse). We feel there might be confounding variables and further analysis is needed, which is outside the scope of the project as we are only interested in demonstrating the efficacy of our methods for imputation and training the model, which results in better analysis and interpretation.

4 | DISCUSSION AND CONCLUSIONS

In this project, we explored different multivariate imputation methods under sparse and irregular functional data settings. We have proposed a new imputation method, MLLF, which is a mixture of MF and MICE. Also, we modified this method along with MF to deal with functional data in a systematic fashion by careful initialization using PACE and smoothing out the results using bins. Our proposed approaches overcome a lot of the challenges faced by the current methods (like PACE and MICE) to give consistent estimates. They incorporate the response and deal with complex non-linear relations with multiple imputations. Results under multiple simulation settings also illustrate the value of our approach over existing methods for fitting scalar-on-function regression models when the functional predictors are irregularly and sparsely sampled irrespective of the sparsity level, number of points in the curve, and sample size. All the binned methods work equally well with slight variations in some cases; though there is no clear winner, MF with binning (MF_B) was the most consistent performer.

Our approach is sensitive to the subdivision of the time points into the bins. Different binning strategies were not explored in depth but are one of the directions for further investigation. Another interesting avenue is defining a relationship between the number of time points (m) and the number of bins (k), to ease the search for the optimum bin number. Also, even though it looks like the extension of MF performs better than LLF imputation (MLLF), further analysis is required to differentiate between the methods. Deep learning has become a major research area in multiple fields and the application of neural networks to the imputation setting might be very interesting, though our initial efforts did not bear strong results.

Finally, at a high level, there are still many remaining challenges with the imputation of functional data. When evaluating the performance of future methods, we suggest considering at least three critical points: (1) Do the imputations improve subsequent modelling? (2) Can the imputations incorporate the assumed underlying smoothness of the curves or at least domain information? and (3) Can the imputations handle measurement noise in the observed points? A multiple imputation approach seems to be critical for the first point, while the latter two are still quite open. Our binning approach, while simple, helped a great deal with the second point. However, the third point was basically untouched in this work. When using methods such as PACE, incorporating observation error is straightforward, but it is unclear how to incorporate it into more complicated imputation procedures.

ACKNOWLEDGEMENT

This research was supported in part by the following grant to Pennsylvania State University: NSF SES-1853209.

ORCID

Aniruddha Rajendra Rao https://orcid.org/0000-0003-3541-7095

REFERENCES

Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy, *Classification, Clustering, and Data Mining Applications*, pp. 639–647.

Athey, S., Tibshirani, J., & Wager, S. (2016). Generalized random forests.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., & For the Alzheimer's Disease Neuroimaging Initiative* (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487. https://doi.org/10.1177/0962280214521348, PMID: 24525487.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carnahan-Craig, S., Blankenberg, D., Parodi, A., Paul, I., Birch, L., Savage, J.,..., & Makova, K.(2018). Child weight gain trajectories linked to oral microbiota composition. *Scientific Reports*, *8*, 4–14.

Chen, Y., Carroll, C., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K.,..., & Wang, J.-L. (2019). Fdapace: Functional data analysis and empirical dynamics. https://CRAN.R-project.org/package=fdapace, R package version 0.5.1.

Crambes, C., & Henchiri, Y. (2018). Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference*, *201*, 103–119.

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. https://doi.org/10.1093/jamia/ocx037

Ding, Y., & Ross, A. (2012). A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognition*, *45*(3), 919–933. http://www.sciencedirect.com/science/article/pii/S0031320311003153

Fan, Y., James, G. M., & Radchenko, P. (2015). Functional additive regression. *Annals of Statistics*, *43*(5), 2296–2325. https://doi.org/10.1214/15-AOS1346

Ferraty, F., & Romain, Y. (2011). The oxford handbook of functional data analysis.

Ferraty, F., Sued, M., & Vieu, P. (2012). Mean estimation with data missing at random for functional covariables. *Statistics*, *47*, 688–706.iFirst.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice (springer series in statistics)*. Berlin, Heidelberg: Springer-Verlag.

Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2018). Local linear forests.

García-Rodríguez, O., Secades-Villa, R., Florez-Salamanca, L., Okuda, M., Liu, S.-M., & Blanco, C.(2013). Probability and predictors of relapse to smoking: Results of the national epidemiologic survey on alcohol and related conditions (nesarc). *Drug and Alcohol Dependence*, *132*, 479–485.

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

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, *142*(12), 1255–1264. https://doi.org/10.1093/oxfordjournals.aje.a117592

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

Hansson, L., Hedner, T., & Jern, S. (1996). Smoking affects blood pressure. *Blood Pressure*, *5*, 68.

He, Y., Yucel, R. M., & Raghunathan, T. E. (2011). A functional multiple imputation approach to incomplete longitudinal data. *Statistics in medicine*, *30*(10), 1137–56.

Herd, N., Borland, R., & Hyland, A. (2009). Predictors of smoking relapse by duration of abstinence: Findings from the international tobacco control (itc) four country survey. *Addiction (Abingdon, England)*, *104*, 2088–99.

Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*, Springer Series in Statistics. New York: Springer. https://books.google.com/books?id=OVezLB__ZpYC

James, G. M., Hastie, T. J., & Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*, *87*(3), 587–602. https://doi.org/10.1093/biomet/87.3.587

Kokoszka, P., & Reimherr, M. (2018). *Introduction to functional data analysis*. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781315117416

Kowal, D. R., Matteson, D. S., & Ruppert, D. (2019). Functional autoregression for sparsely sampled data. *Journal of Business & Economic Statistics*, *37*(1), 97–109. https://doi.org/10.1080/07350015.2017.1279058

Liao, S., Lin, Y., Kang, D., Chandra, D., Bon, J., Kaminski, N.,..., & Tseng, G. (2014). Missing value imputation in high-dimensional phenomic data: Imputable or not, and how? *BMC Bioinformatics*, *15*, 346.

Ma, H., & Zhu, Z. (2016). Continuously dynamic additive models for functional data. *Journal of Multivariate Analysis*, *150*, 1–13. http://www.sciencedirect.com/science/article/pii/S0047259X16300240

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., & Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, *23*(1), 249–269. https://doi.org/10.1080/10618600.2012.729985

Mozharovskyi, P., Josse, J., & Husson, F. (2017). Nonparametric imputation by data depth.

Müller, H.-G., Wu, Y., & Yao, F. (2013). Continuously additive models for nonlinear functional regression. *Biometrika*, *100*(3), 607–622. https://doi.org/10.1093/biomet/ast004

Ning, J., & Cheng, P. E. (2012). A comparison study of nonparametric imputation methods. *Statistics and Computing*, *22*, 273–285.

Petrovich, J., Reimherr, M., & Daymont, C. (2018). Highly irregular functional generalized linear regression with electronic health records.

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

Primatesta, P., Falaschetti, E., Gupta, S., Marmot, M., & Poulter, N. R. (2001). Association between smoking and blood pressure: Evidence from the health survey for england. *Hypertension*, *37*, 187–93.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional data analysis*, Springer series in statistics. New York, NY USA: Springer. https://books.google.com/books?id=vYXCsgEACAAJ

Reimherr, M., Sriperumbudur, B., & Taoufik, B. (2017). Optimal prediction for additive function-on-function regression.

Rice, J. A., & Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, *57*(1), 253–259. https://doi.org/10.1111/j.0006-341x.2001.00253.x

Rubin, D. (2004). *Multiple imputation for nonresponse in surveys*, Wiley classics library edition. Hoboken, NJ. [u.a.]: Wiley.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Stekhoven, D. J. (2013). missforest: Nonparametric missing value imputation using random forest. R package version 1.4.

Stekhoven, D. J., & Búhlmann, P. (2011). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Thompson, W., & Rosen, O. (2008). A Bayesian model for sparse functional data. *Biometrics*, *64*, 54–63.

Tibshirani, J., Athey, S., & Wager, S. (2020). GRF: Generalized random forests. https://github.com/grf-labs/grf, R package version 1.1.0.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242. https://doi.org/10.1177/0962280206074463, PMID: 17621469.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. https://www.jstatsoft.org/v45/i03/

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., … , & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*,(8) 1–7, e002847. https://doi.org/10.1136/bmjopen-2013-002847

Wang X., & Ruppert D. (2015). Optimal Prediction in an Additive Functional Model. *Statistica Sinica*, *25*,(2) 567–589. https://doi.org/10.5705/ss.2013.074

Wang, Y., Zheng, X., Zhang, C., Yang, Y., Liu, L., Qi, Y., & Bu, P. (2018). A12426 association between smoking and blood pressure in elderly male patients with essential hypertension. *Journal of Hypertension*, *36*, e321.

Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, *100*(470), 577–590. https://doi.org/10.1198/016214504000001745

How to cite this article: Rao AR, Reimherr M. Modern multiple imputation with functional data. *Stat*. 2021;10:e331. https://doi.org/10.1002/sta4.331

APPENDIX A

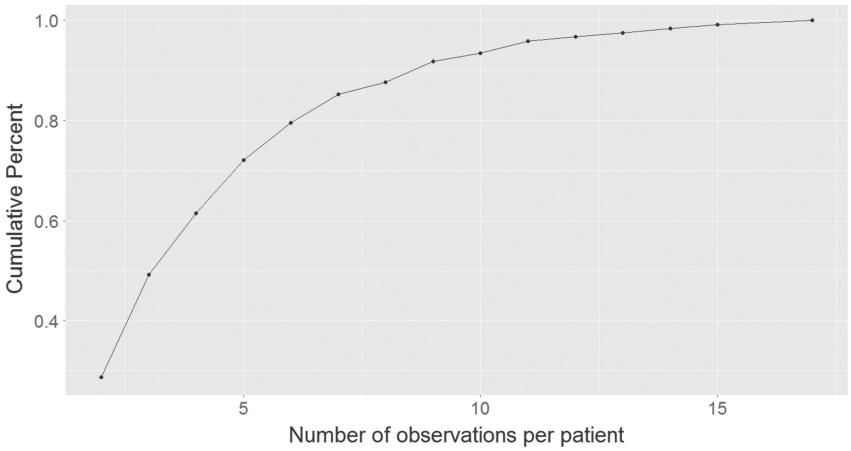

FIGURE A1 Cumulative percentage of observations for BP per patient

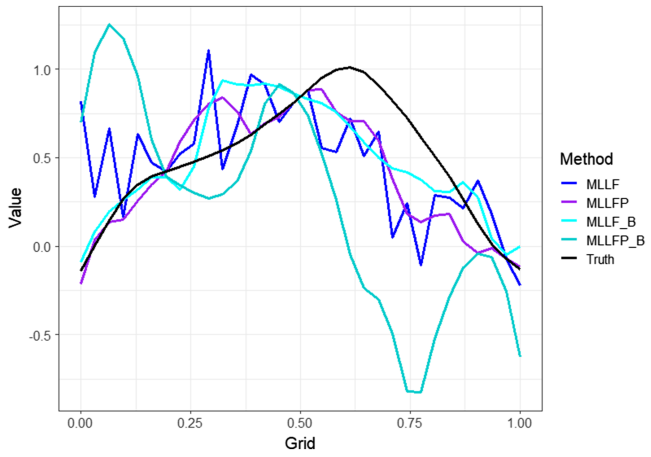

FIGURE A2 Comparing imputed curves in non-binned and binned methods of MLLF under linear case for one random sample curve with time points (m) equal to 52 and sparsity (s) High

TABLE A1 RMSE of prediction, β coefficients, and imputation of the curves for different methods under the linear case when $n = 200$ under different time points and sparsity settings

	$n = 200$, **s = medium**						$n = 200$, **s = high**					
	$m = 32, b = 17$			$m = 52, b = 27$			$m = 32, b = 8$			$m = 52, b = 12$		
Method	**Pred**	β	**Imp**	**Pred**	β	**Imp**	**Pred**	β	**Imp**	**Pred**	β	**Imp**
MF	0.149	0.1656	0.141	0.476	0.459	0.117	1.290	1.272	0.573	0.349	0.476	0.529
PACE	0.221	0.332	0.393	0.700	0.667	1.863	0.433	0.461	0.434	0.449	0.456	0.362
MLLF	0.140	0.144	**0.085**	0.515	0.491	**0.025**	30.871	20.224	0.663	26.841	20.301	0.583
MICE	0.148	0.191	0.204	0.462	0.452	0.172	0.400	0.704	0.939	0.440	0.649	0.961
MFP	0.140	0.158	0.165	0.468	0.455	0.225	0.340	0.483	0.509	0.220	0.354	0.451
MLLFP	0.170	0.188	0.186	0.493	0.472	0.196	0.434	0.461	0.427	0.454	0.459	0.354
MF_B	**0.124**	**0.135**	**0.89**	**0.264**	**0.258**	0.078	**0.285**	**0.340**	**0.395**	**0.191**	**0.255**	**0.312**
MLLF_B	**0.124**	**0.132**	**0.091**	**0.266**	**0.260**	0.077	**0.286**	**0.313**	0.434	**0.178**	**0.283**	**0.301**
MFP_B	**0.126**	**0.133**	**0.071**	**0.266**	**0.269**	**0.037**	0.411	0.434	0.450	0.211	**0.262**	**0.293**
MLLFP_B	**0.124**	**0.135**	**0.088**	**0.268**	**0.265**	**0.030**	0.423	0.662	0.652	**0.197**	**0.273**	0.345

Note: Bold values in the table indicate the best values.

TABLE A2 RMSE of prediction, β coefficients, and imputation of the curves for different methods under the linear case when $n = 1000$ under different time points and sparsity settings

	$n = 1000$, **s = medium**						$n = 1000$, **s = high**					
	$m = 32, b = 17$			$m = 52, b = 27$			$m = 32, b = 7$			$m = 52, b = 27$		
Method	**Pred**	β	**Imp**	**Pred**	β	**Imp**	**Pred**	β	**Imp**	**Pred**	β	**Imp**
MF	0.155	**0.163**	**0.085**	0.259	0.265	**0.061**	3.137	3.070	0.404	0.395	0.455	0.328
PACE	0.176	0.242	0.192	**0.125**	0.583	0.868	0.412	0.441	**0.389**	0.326	0.354	0.281
MLLF	**0.140**	**0.144**	**0.061**	0.265	0.272	**0.027**	51.947	41.243	0.545	61.500	51.082	0.476
MICE	0.152	0.168	0.118	0.254	0.262	0.092	0.365	0.660	0.873	0.911	0.838	0.856
MFP	0.151	**0.164**	**0.085**	0.251	0.278	0.110	0.198	**0.296**	**0.353**	0.288	0.338	**0.274**
MLLFP	0.167	0.198	0.134	0.0.255	0.284	0.099	0.412	0.441	**0.339**	0.324	0.357	**0.244**
MF_B	**0.143**	**0.168**	**0.071**	**0.116**	**0.220**	**0.046**	**0.149**	**0.267**	**0.357**	**0.104**	**0.194**	**0.263**
MLLF_B	**0.147**	**0.153**	**0.072**	**0.113**	**0.208**	**0.046**	**0.123**	**0.293**	**0.373**	**0.111**	**0.265**	**0.295**
MFP_B	**0.144**	**0.165**	**0.048**	**0.116**	**0.221**	**0.026**	**0.145**	**0.273**	**0.348**	**0.128**	**0.219**	0.366
MLLFP_B	**0.143**	**0.160**	**0.051**	**0.113**	**0.211**	**0.031**	**0.127**	0.802	0.644	**0.138**	0.354	0.475

Note: Bold values in the table indicate the best values.

TABLE A3 RMSE of prediction and imputation of the curves for different methods under the non-linear case $(f(X_i(t), t) = cos(X(t)^3 * t) + 5 * t)$ when $n = 500$ under different time points and sparsity settings

	$n = 500$, **s = medium**				$n = 500$, **s = high**			
	$m = 32, b = 17$		$m = 52, b = 12$		$m = 32, b = 7$		$m = 52, b = 27$	
Method	**Pred**	**Imp**	**Pred**	**Imp**	**Pred**	**Imp**	**Pred**	**Imp**
MF	0.180	0.189	0.292	0.112	0.682	0.541	0.832	0.468
PACE	0.147	1.912	0.253	0.549	0.598	0.393	0.557	0.292
MLLF	0.138	0.104	0.239	0.093	25.376	2.586	18.231	1.581
MICE	0.199	0.275	0.448	0.267	0.712	0.837	0.802	1.007
MFP	0.178	0.177	0.237	0.102	0.630	0.434	0.842	0.431
MLLFP	0.320	0.060	0.212	0.091	0.483	0.320	0.570	**0.160**
MF_B	**0.155**	**0.054**	**0.189**	**0.032**	**0.458**	**0.178**	**0.502**	**0.155**
MLLF_B	**0.156**	**0.053**	**0.184**	**0.032**	**0.459**	**0.179**	**0.524**	**0.159**
MFP_B	**0.154**	**0.055**	**0.191**	**0.030**	**0.455**	**0.179**	**0.502**	**0.156**
MLLFP_B	**0.155**	**0.055**	**0.183**	**0.031**	**0.468**	**0.206**	0.550	0.207

Note: Bold values in the table indicate the best values.