Non Linear Functional Data Imputation

Aniruddha R. Rao* and Matthew L. Reimherr*
*Department of Statistics, Pennsylvania State University

PennState Eberly College of Science

1. Introduction

- In Functional data, longitudinal studies often can be sparse and irregularly sampled. We can either apply Sparse FDA methods or use imputation to apply more traditional FDA techniques.
- The current methods for imputation fall short when dealing with complex nonlinear models.
- We have developed a multiple imputation method using Machine Learning that can overcome this issue while performing in par with **PACE** and **MICE**

2. Functional Data Analysis (FDA)

$$Y_i = \alpha + \int f(X_i(t), t)\, ds + \varepsilon_i$$

This is scalar-on-function Non-linear regression. The parameter is the constant α and function f.

- $X_i(t)$ are the functional covariates where t ∈ [0,1], i = $1, \dots, n$ and each curve is observed at m locations.
- If $f(X_i(t), t) = X_i(t)\beta(t)$ then we have a linear model

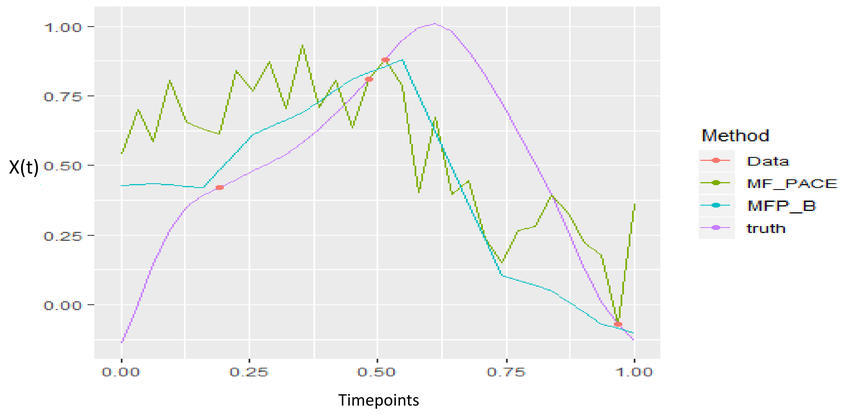
3. Methods

- **PACE:** PACE imputes by estimating the unknown parameters via pooled nonparametric smoothing, and then using them to form Best Linear Unbiased Predictors (BLUPs) of the curves/scores.
- **MICE:** A series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable can be modeled according to its distribution.
- **MISSFOREST:** It trains a Random Forest on observed values in a first step, followed by predicting the missing values and then proceeding iteratively.
- **Bins:** To overcome the issue of non smooth imputation, we divide the m timepoints in a k bins and impute over the k points to interpolate back to m timepoints.

4. Simulation

- We consider n curves and then add missingness pattern using Missing Completely at Random (MCAR) method.
- Each curve has same number of points missing.

Figure: Comparison of different Imputation methods.



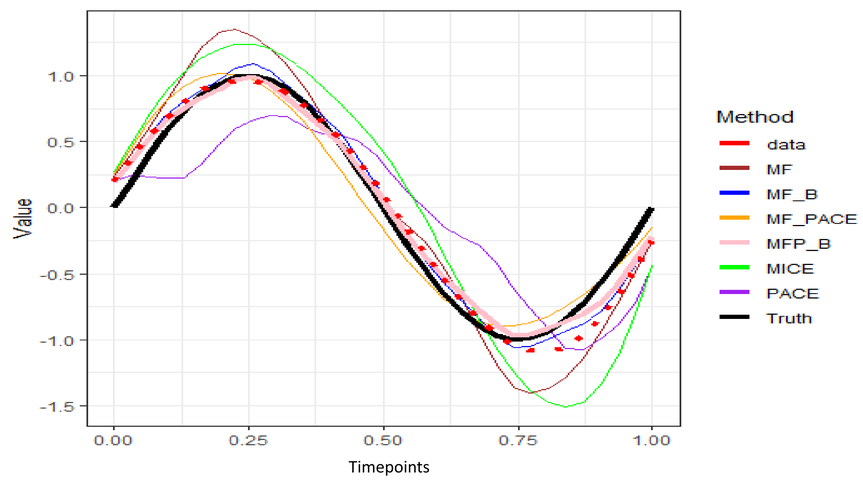
We consider a simple case where *n=200, m=102, sparsity=90%.*

Linear case: $\beta(t) = \sin(2\pi t)$ and optimum bins is 8.
We compare the different methods under ten iterations wrt RMSE for imputation, β values and $X\beta$.

Non-linear model: $f(X_i(t), t) = 5\sin(X_i(t)^2 t^2)$ and optimum bins is 12. We compare the different methods under ten iterations wrt RMSE for imputation and Prediction.

Below, we can see the bins help to smooth the imputation.

Figure: Estimated $\widehat{\beta}$ under different approaches.



5. Conclusion and Future work

Linear Case:
Our Proposed method performance better wrt modelling and estimating β.

Method	β	$X\beta$	X
MF	0.7653	0.8571	0.4533
PACE	0.3513	0.4022	**0.3584**
MF_PACE	0.32940	0.4081	0.3856
MF_B	**0.1942**	**0.2972**	0.3628
MFP_B	**0.2041**	**0.3091**	0.4433
MICE	0.3783	0.6773	0.9197

Non-Linear Case:
Our Proposed method has a better performance when it comes to modelling and estimating the regression function β.

Method	β	X
MF	0. 465	0.339
PACE	0.292	**0.2**
MF_PACE	0.431	0.328
MF_B	**0.185**	**0.181**
MFP_B	**0.166**	**0.182**
MICE	0.952	0.322

Future work:
- Relation between bins and timepoints.

6. References

1. Petrovich J, Reimherr M and Daymont C. *Functional Regression Models with Highly Irregular Designs, 2018.*
2. Stekhoven D, Bühlmann P. *MissForest—non-parametric missing value imputation for mixed-type data.* Bioinformatics, 2012.
3. F. Yao, H.-G. Müller, and J.-L. Wang. *Functional data analysis for sparse longitudinal data. JASA, 2005.*
4. Van Buuren S. *Multiple imputation of discrete and continuous data by fully conditional specification. SMMR, 2007*