Modern Multiple Imputation With Functional Data

Aniruddha Rajendra Rao and Matthew Reimherr

Pennsylvania State University

June 05

Symposium on Data Science and Statistics 2020

1. Introduction
   - Overview of Functional Data
   - Motivational Example
2. Methods
   - PACE
   - Non-Linear Multivariate Imputation
3. Results
   - Simulation
   - EHR
4. Conclusion and Future Work

Functional Data

- Functional Data Analysis (FDA) is a branch of statistics that analyzes samples consisting of functions or smooth curves.
- The data is usually of the form:
  $x_i\left(t_{j,i}\right) \in \mathbb{R}, \quad t_{j,i} \in \left[T_1, T_2\right], \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, J_i$

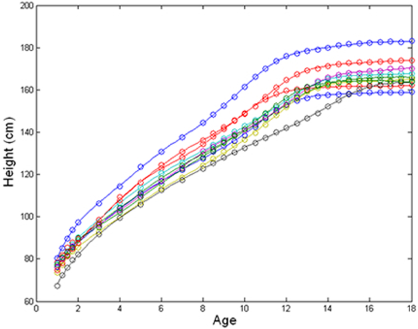

Figure: Child growth curves of boys, where height of each child was observed multiple times between ages 1 and 18 (Berkeley Growth Study).

Functional Models

General form of Linear scalar-on-function regression model is as follows:

$$Y_i = \int \beta(t) X_i(t) + \epsilon_i \qquad i = 1, 2, ..., n$$

Linear case be solved with the help of:

- Basis Expansion
  $\beta(t) \approx \sum_{k=1}^{K} b_k e_k(t) \Longrightarrow \int \beta(t) X_i(t) \approx \sum_{k=1}^{J} b_k x_{ik}$
- Functional PCA
  $X_i(t) \approx \hat{\mu}(t) + \sum_{j=1}^{p} \hat{\xi}_{ij} \hat{v}_j(t)$
  $Y_i = \alpha + \int \beta(t) \left(\hat{\mu}(t) + \sum_{j=1}^{p} \hat{\xi}_{ij} \hat{v}_j(t) \right) dt + \epsilon_n$

For Non Linear case, $Y_i = \int f(X_i(t), t) dt + \epsilon_i \qquad i = 1, 2, ..., n$

Sparse Functional Data

Sparse in Functional Data means that the curves are observed only at a small number of timepoints.

Approach:

- Sparse FDA
  Modifying the existing methods to incorporate the sparse structure (like Sparse FPCA).

- Imputation
  Imputation is a process of filling in the missing values in a reasonable manner.

Motivational Example

EHR Data:

- The Electronic Health Record Data is from Penn State Health Milton S. Hershey Medical Center, consists of n (122) patients (smokers) measured over m (18) months.
- We want to model if a patient will relapse (Y/N) at the end of 18 months as a function of Blood Pressure (BP).
- For each patient, the number of clinical visits they had between first month through the next 18 months is recorded -varies greatly across subjects.
- On an Average we have 4 out 18 timepoints observed for a patient.

PACE

PACE[1] uses Functional PCA and mean imputation. The PACE algorithm uses the Karhunen-Loéve expansion:

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

where,

$\bullet \xi_{ij} = \int (X_i(t) - \mu_X(t)) v_j(t) dt, \quad \mathrm{E}[\xi_{ij}] = 0, \quad \mathrm{Cov}(\xi_{ij}, \xi_{ik}) = \lambda_j 1_{\{j=k\}}$
$\bullet \{\lambda_j, v_j(t)\}_{j=1}^{\infty}$ are the eigenvalues, eigenfunctions of C_X
$\bullet C_X(s,t) = \sum_{j=1} \lambda_j v_j(s) v_j(t)$ by Mercer's Theorem

The BLUP for the scores given $\boldsymbol{\Sigma}_{X_i}^{-1} = c_X(t,s) + \sigma^2 I_{M_i}$ is then

$$\hat{\xi}_{ik} = \hat{E}[\xi_{ik} | \mathbf{x}_i] = \lambda_j \mathbf{v}_{ij}^T \boldsymbol{\Sigma}_{x_i}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_i)$$

[1] Fang Yao, Hans-Georg M uller, and Jane-Ling Wang, 2005. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association.

Non-Linear Multivariate Imputation

Non-Linear Multivariate Imputation methods have been proven to work well for dealing with missing data. We check the performance of the following methods under Functional Data setting:

- MICE
- MissForest
- Local Linear Forest

MICE

MICE[2] is a series of models where for each variable with missing data, it is modeled conditional upon the observed variables in the data.

Models:

- Predictive mean matching
- Regression
- MCMC
- Bootstrap
- GLM
- Random Forest

Key Feature: Multiple Imputation and Mixed Data Types.

[2] Stef van Buuren, 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research.

MissForest

Missforest[3] is a multiple imputation method, which proceeds by training a Random Forest (RF) on the observed parts of the data.

Consider matrix $X_{n\times m}$. For an arbitrary variable t in X including missing values at entries $\mathbf{i}^{(t)}$ mis $\subseteq \{1, \dots, n\}$, we can separate the dataset into four parts: $Y^t_{mis}, Y^t_{obs}, X^t_{mis}, X^t_{obs}$.

Pseudo Algorithm:

- Initialize using mean Imputation.
- We sort the variables of X in ascending order of sparsity.
- Update missing values for all t:
  Fit a RF $Y^t_{obs} \sim X^t_{obs}$ and then predict $Y^t_{mis} \sim X^t_{mis}$.
- Repeat till convergence.

[3] Daniel J. Stekhoven and Peter Buhlmann, 2011. Missforest- Non-parametric missing value imputation for mixed-type data, Bioinformatics.

Local Linear Forest

Local Linear Forest[4] solves the problem of smoothness in the regression surface by using a RF to generate weights that are used as a kernel for local linear regression.

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

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

$$\alpha_i \left(\mathbf{x}_o \right) = \frac{1}{B} \sum_{b=1}^{B} \frac{1 \left\{ \mathbf{X_i} \in L_b \left(\mathbf{x}_o \right) \right\}}{\left| L_b \left(\mathbf{x}_o \right) \right|}$$

[4] Rina Friedberg, Julie Tibshirani, Susan Athey, and Ste-fan Wager, 2018. Local linear forests.

Adapting to Functional Data

Problem: Functional Data can be very sparse giving a very rough imputation

Solution: Use PACE to initialize and bin nearby observations.

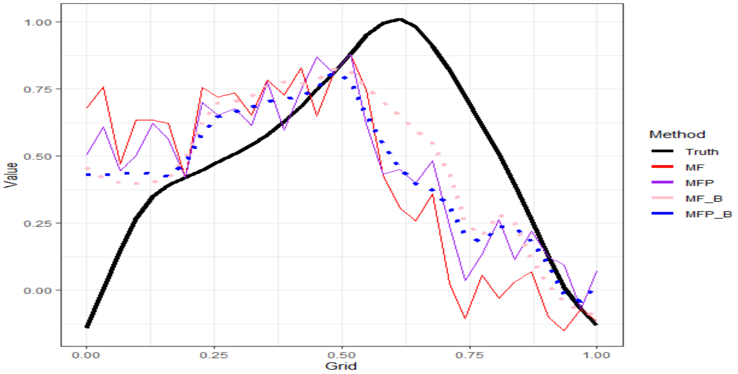

Figure: Imputed curves for different methods showing the advantage of binning leading to smoother curves.

Simulation- Linear

For the Linear case, we simulate n iid random curves $\{X_1(t), \cdots, X_N(t)\}$ from a Gaussian process with mean 0 and covariance function as Matérn covariance function and $\beta(t) = 5 \times \sin(2\pi t)$.

Method	n=500, s=50%					
	m=32, b=17			m=52, b=27		
	Pred	β	Imp	Pred	β	Imp
PACE	0.17	0.208	0.199	0.484	0.595	1.91
MICE	3.611	3.612	0.09	0.237	0.253	0.089
MF	0.136	0.155	0.108	0.227	0.241	0.077
LLF	0.142	0.149	**0.07**	0.234	0.242	**0.023**
MFP	**0.122**	**0.144**	0.105	0.228	0.250	0.13
LLFP	0.132	0.153	0.144	0.232	0.246	0.10
MF_B	**0.122**	**0.136**	**0.079**	**0.173**	**0.180**	**0.052**
LLF_B	**0.126**	**0.137**	**0.082**	**0.174**	**0.177**	**0.053**
MFP_B	**0.122**	**0.138**	**0.053**	**0.176**	**0.182**	**0.023**
LLFP_B	**0.128**	**0.143**	**0.059**	**0.179**	**0.178**	**0.023**

Table: RMSE of Prediction, β coefficients and Imputation of the curves for different methods under Linear case when sparsity (s) is 50%.

β coefficient plot

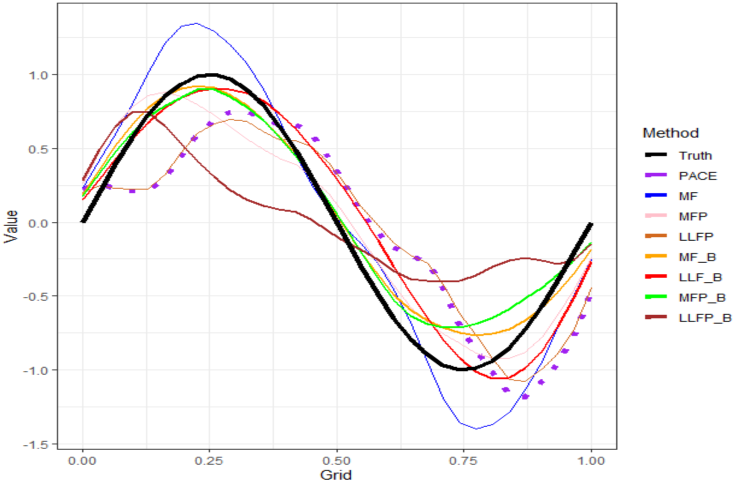

Figure: Estimated coefficient function for different methods under linear case with sample size (n)=500, time points (m)=52, sparsity (s)=50%.

Simulation- Non Linear

For the Non Linear case, we simulate n iid random curves $\{X_1(t), \cdots, X_N(t)\}$ from a Gaussian process and $f(X_i(t), t) = 5 * sin(X(t)^2 * t^2)$.

Method	n=500, s=90%			
	m=32, b=7		m=52, b=7	
	Pred	Imp	Pred	Imp
PACE	0.431	0.659	0.386	0.551
MICE	0.652	0.892	0.644	1.02
MF	**0.303**	0.428	0.355	0.381
LLF	0.419	0.592	0.351	0.482
MFP	**0.334**	**0.379**	0.351	0.324
LLFP	0.427	**0.357**	0.342	**0.259**
MF_B	**0.293**	**0.257**	**0.318**	**0.275**
LLF_B	**0.335**	**0.364**	**0.312**	**0.282**
MFP_B	**0.290**	**0.257**	**0.311**	**0.251**
LLFP_B	**0.328**	**0.385**	**0.329**	**0.308**

Table: RMSE of Prediction and Imputation of the curves for different methods under Non Linear case when sparsity (s) is 90%.

EHR

- The Electronic Health Record Data is from Penn State Health Milton S. Hershey Medical Center, consists of n (122) patients (smokers) measured over m (18) months.
- We want to model if a patient will relapse (Y/N) at the end of 18 months as a function of Blood Pressure (BP).
- On an Average we have 4 out 18 timepoints observed for a patient.

PACE	MICE	MF	LLF	MFP	LLFP	MF_B	LLF_B	MFP_B	LLFP_B
0.42	0.39	0.39	0.38	0.36	0.37	**0.33**	**0.35**	**0.32**	**0.35**

Table: Prediction error of different methods for the EHR data

EHR- β coefficient plot

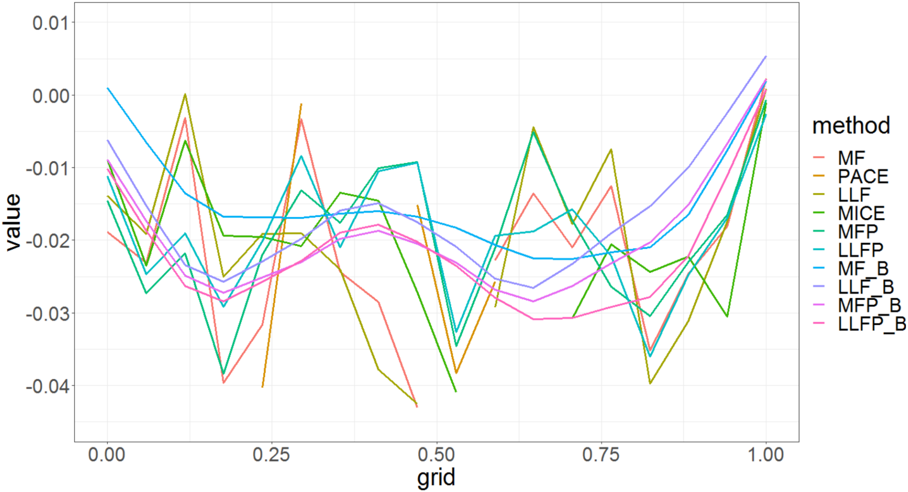

Figure: Estimated coefficient function for different methods for EHR data with **sample size (n)=122, timepoints (m)=18, bin=10.**

Conclusion and Future Work

Conclusion:

- Our method uses information of the response along with performing Multiple Imputation.
- We outperform PACE and MICE wrt Imputation and Modelling irrespective of number of timepoints or sparsity.
- Binning helps to smooth results out leading to better performance.

Future Work:

- Perform under different binning schemes.
- Define relation between number of bins and timepoints.
- Differentiate when to use MF and LLF.

References

- Fang Yao, Hans-Georg M uller, and Jane-Ling Wang, 2005. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association.
- Stef van Buuren, 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research.
- Justin Petrovich, Matthew Reimherr, and Carrie Daymont, 2018. Highly irregular functional generalizedlinear regression with electronic health records
- Daniel J. Stekhoven and Peter Buhlmann, 2011. Missforest- Non-parametric missing value imputation for mixed-type data, Bioinformatics.
- Rina Friedberg, Julie Tibshirani, Susan Athey, and Ste-fan Wager, 2018. Local linear forests.

Questions?

Thank you