Extending People-like-Me Methods with Mahalanobis distance

A Personalized Predictive Longitudinal Study of Growth in Children

Randy (Xin) Jin

2023-06-20

Objective

Modern clinical data analysis often deals with heterogeneous data

The traditional predictive modeling:

- Develops a model for global inference with all data
- Overlooks the diversity and heterogeneity

Personalized predictive methods

- Have the objective of addressing heterogeneity
- People-like-me method

People-like-me method:

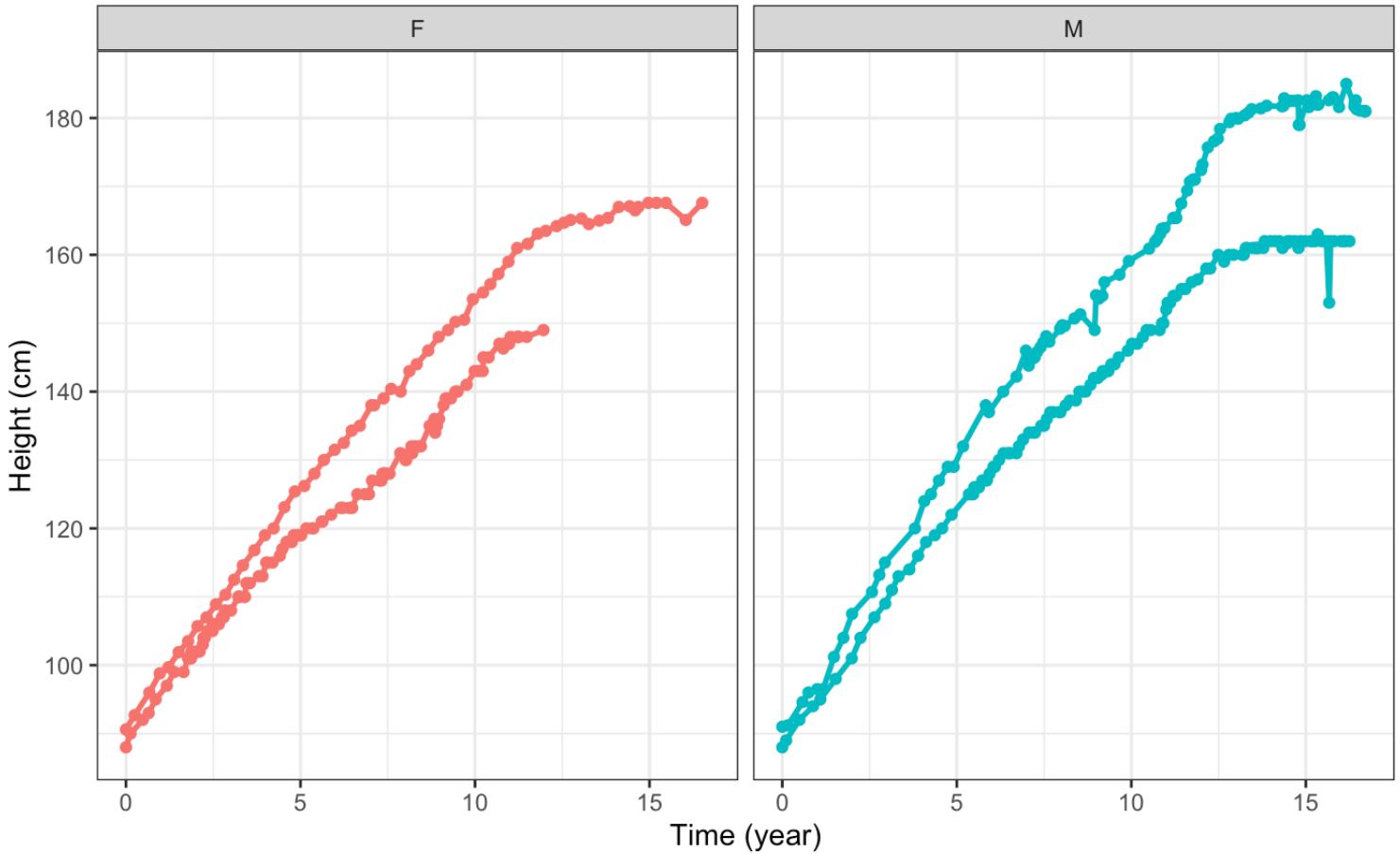
People-like-me:

- Uses a few but more similar trajectories
- Predicts based on those similar trajectories
- It can have higher predictive performance
- Previous work has focused on finding matching trajectories based on a single time point

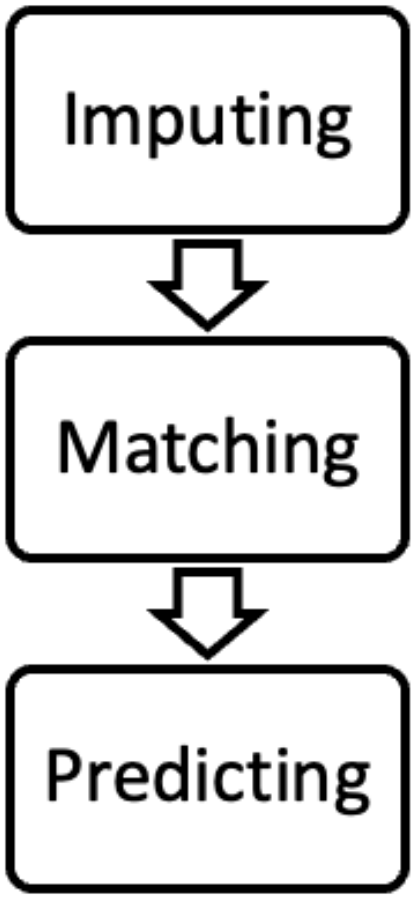
Our contribution is to:

1. extend matching step to using more time points (**anchor time**)
2. use alternative approaches to select matches

Example trajectories



Flowchart



- **Brokenstick Model**
- **Linear Model**

- **Single-time**
- **Multiple-time**
  - **Euclidean**
  - **Mahalanobis**
  - **Selection criterion**

- **GAMLSS Model**
  - **Median**
  - **Centiles**

Imputation

0. Split dataset into training and testing. For every individual in the testing set y_{test}, do the following steps.
1. Fit a brokenstick model with training set: $\boldsymbol{y}_{train} \sim BrokenStick(\boldsymbol{X}_{train})$
   - brokenstick model is a particular type of linear mixed model
   - piece-wise cubic B-splines for both fixed and random effect
2. Obtain the predictions based on brokenstick model at fixed set of anchor time points $\boldsymbol{\tau} = \{\tau_1, \ ..., \ \tau_k\}$: predict outcomes for every subjects i in training set: $\tilde{\boldsymbol{y}}_{\boldsymbol{i}} = (\tilde{y}_{i,\tau_1}, \ ..., \ \tilde{y}_{i,\tau_k})$

Imputation

3. Fit linear model with the predicted values from Step 2: $\tilde{\boldsymbol{y}} \sim LM_{\boldsymbol{\tau}}(\boldsymbol{y_{train,0}}, \boldsymbol{X_{train,0}})$
   - This model includes only the demographic information and baseline outcome, at each time point independently
4. Obtain the predictions based on the linear model for every training subject $\bar{\boldsymbol{y}}_{\boldsymbol{i}}$ and for the target individual $\bar{\boldsymbol{y}}_{\boldsymbol{test}}$

Matching

5. Calculate the distance between the target (test) and every training subject: $D_{i,test} = dist(\bar{\boldsymbol{y}}_{\boldsymbol{i}}, \bar{\boldsymbol{y}}_{\boldsymbol{test}})$.

Two types of distance will be used

- Euclidean distance: $D_E = \|\boldsymbol{y_1} - \boldsymbol{y_2}\| = \sqrt{\sum_{i=1}^{k}(y_{1i} - y_{2i})^2}$
- Mahalanobis distance: $D_M = \sqrt{(\boldsymbol{y_1} - \boldsymbol{y_2})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y_1} - \boldsymbol{y_2})}$
  - where $\boldsymbol{y_1} = (y_{11}, ...y_{1k})$ and $\boldsymbol{y_2} = (y_{21}, \ ..., \ y_{2k})$;
  - $\boldsymbol{\Sigma}$ is the variance-covariance matrix

Matching

6. Select the matching cohort as set $\boldsymbol{A}$ from the training set, based on given distance in Step 5;

- The top κ smallest distance trajectories as matching set
- Mahalanobis distance criterion
  - $D_M^2 = (\boldsymbol{y_1} - \boldsymbol{y_2})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y_1} - \boldsymbol{y_2})$, $D_M^2 \sim \chi_k^2$, where $df = k$ is the degrees of freedom
  - Then set $\boldsymbol{A}$ may be chosen such that:
    $\boldsymbol{A} = \left\{ \boldsymbol{y} : (\boldsymbol{y} - \boldsymbol{y^*})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{y^*})) \leq \chi^2_{df=k,\ \alpha} \right\}$.

Prediction

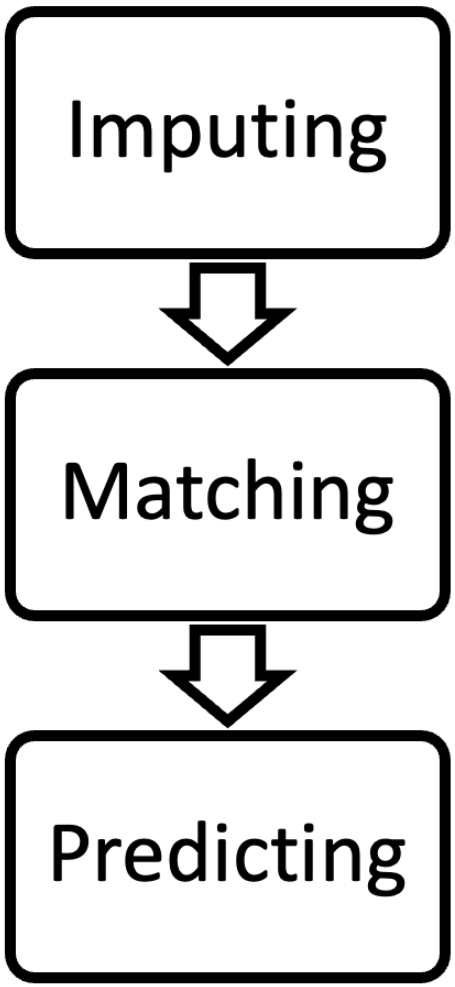
7. Use the matching trajectories of $(\boldsymbol{y_{sub}}, \boldsymbol{X_{sub}}) \in \boldsymbol{setA}$ from Step 6 to fit a flexible model, specifically GAMLSS (Generalized Additive Model for Location Scale and Shape) model: $\hat{\boldsymbol{y}} \sim GAMLSS(\boldsymbol{X_{sub}})$

- We use the predicted median as the personalized prediction and the centiles as the predictive interval for the target individual

Application

The Early Pseudomonas Infection Control (EPIC) Observational Study is a prospective, multi-center, observational longitudinal study. EPIC study aims to investigate progression of cystic fibrosis outcomes in children

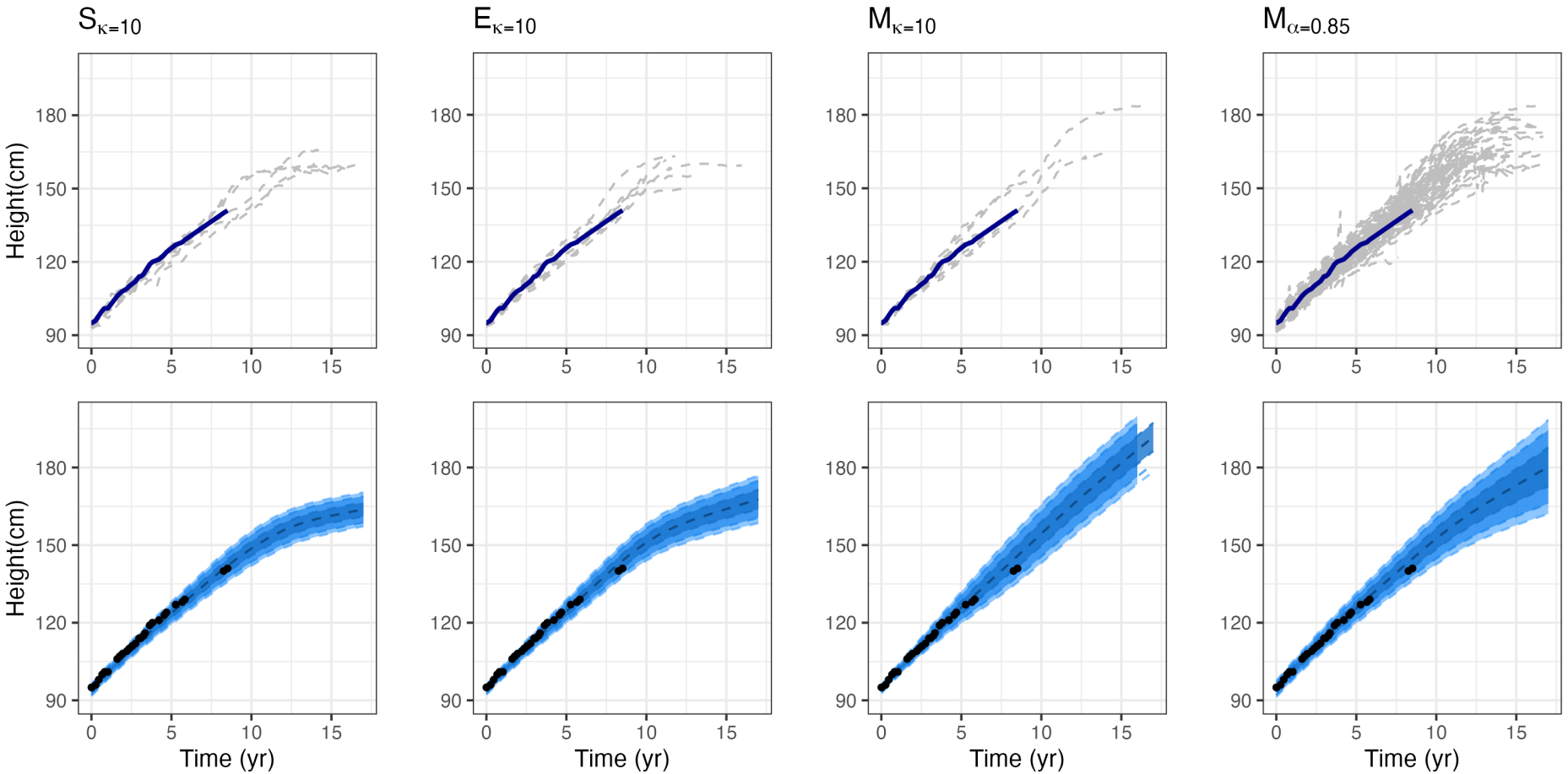
$\boldsymbol{Variable}$	$\boldsymbol{Testing}$, N = 457	$\boldsymbol{Training}$, N = 913
$\boldsymbol{Female}$	237 (52%)	456 (50%)
$\boldsymbol{Age_0}(yr)$	3.14 (3.06, 3.23)	3.13 (3.05, 3.22)
$\boldsymbol{Age_{final}}(yr)$	12.7 (10.3, 16.0)	12.8 (10.1, 16.0)
$\boldsymbol{Follow.up}(yr)$	9.5 (7.2, 12.6)	9.7 (7.0, 12.7)
$\boldsymbol{Height_0}(cm)$	94.0 (91.5, 96.7)	94.0 (91.4, 97.0)



- **Brokenstick Model**
  - **fixed : piecewise cubic B-spline**
  - **random: piecewise cubic B-spline**
  - **knots = c(5, 10, 15)**
- **Linear Model**
  - **at each anchor time**
  - **covariates: baseline and demographics**

- **Single-time:** $\kappa = 10, \tau = 10$
- **Euclidean:** $\kappa = 10$
- **Mahalanobis:** $\kappa = 10$
- **Mahalanobis:** $\alpha = 0.85$

- **GAMLSS Model**
  - **mean: smoothing spline df = 3**
  - **sigma: smoothing spline df = 1**


$S_{\kappa=10}$
$E_{\kappa=10}$
$M_{\kappa=10}$
$M_{\alpha=0.85}$
Height(cm)
Time (yr)
90
120
150
180
0
5
10
15

EPIC Data Analysis

	Anchor time	$S^{t=10}_{\kappa=10}$	$E_{\kappa=10}$	$M_{\kappa=10}$	$M_{\alpha=0.85}$
Bias	t(3, 6, 9, 12)	3.36	2.65	2.31	**2.21**
	t(4, 8, 12, 16)		2.22	2.29	**2.20**
RMSE	t(3, 6, 9, 12)	4.17	3.77	3.38	**3.30**
	t(4, 8, 12, 16)		3.43	3.49	**3.29**
50% CI	t(3, 6, 9, 12)	0.45	0.61	0.38	**0.64**
	t(4, 8, 12, 16)		0.53	0.47	**0.64**
80% CI	t(3, 6, 9, 12)	0.72	0.84	0.66	**0.89**
	t(4, 8, 12, 16)		0.79	0.74	**0.89**

Discussion

Summary:

- Improved people-like-me method for personalized prediction
- In the EPIC study, Mahalanobis distance ($\alpha = 0.85$) performed better than the other methods
- User-defined parameters: number of matches
- We also conducted a simulation study to assess performance of the methods
  - Mahalanobis distance ($\alpha = 0.85$) provide more accurate estimates, higher precision, and better coverage

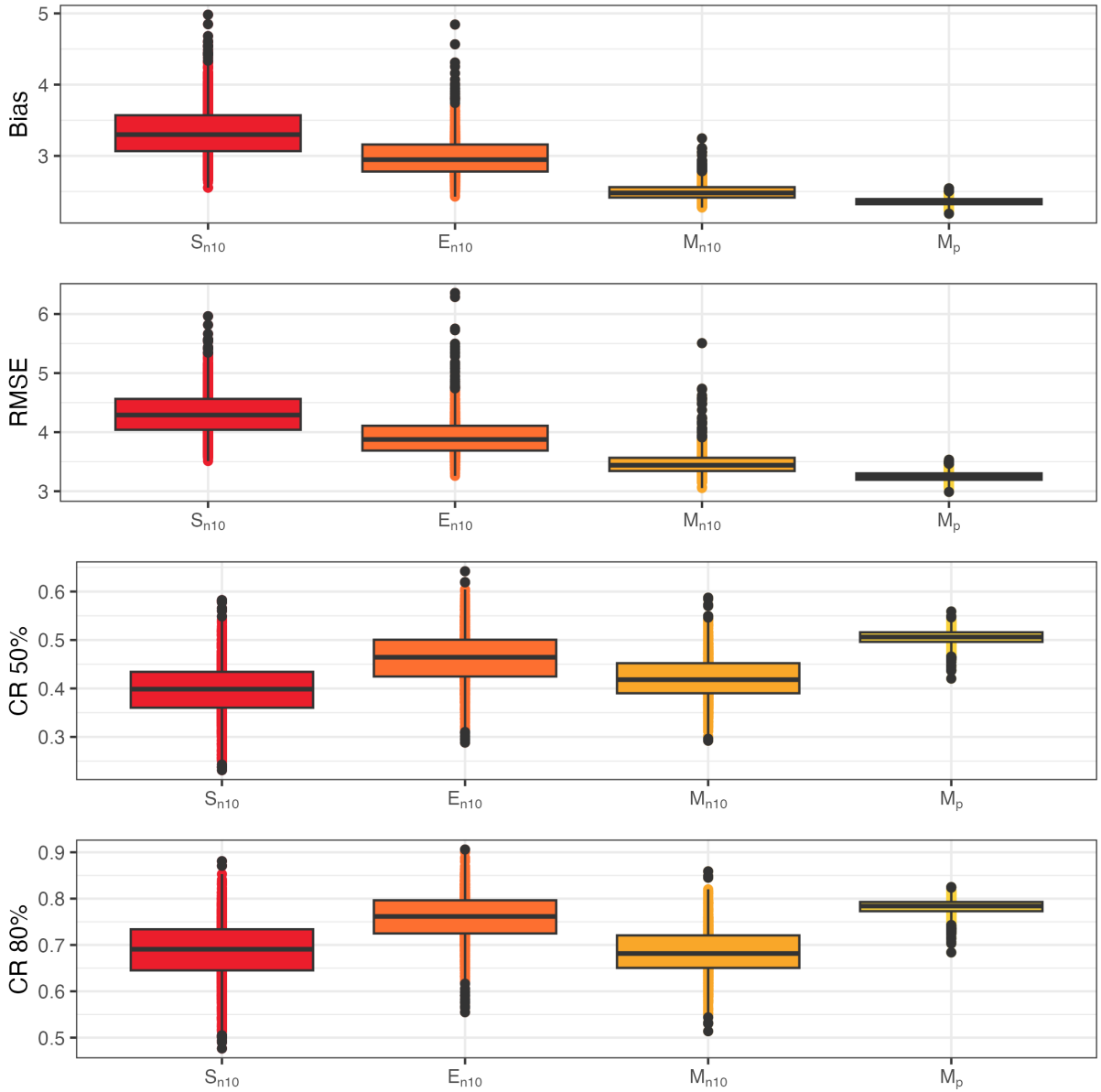
Future work

- Imputation: use other prediction models
- Prediction Modeling: weighted regression and penalization
- Packaged functions for accessibility and application

Simulation

We performed a simulation study

- 1000 simulated datasets
- data generating mechanism based on linear mixed model
  - fixed effect: piece-wise cubic B-spline
  - random effect: piece-wise quadratic B-spline
- performing PLM approach



EXTRA