Using people-like-me method to study the children height growth with EPIC data

An extension for predictive mean matching with Mahalanobis distrance

Randy

Department of Biostatistics & Informatics

1 Background

2 What we got so far

3 Paper

Data

The EPIC Observational Study:

- Prospective, multi-center, observational longitudinal study
- Collected through the Cystic Fibrosis Foundation Patient Registry (CFFPR)
- Of the 1772 children enrolled, we identified **1325 individuals** with usable data, total 76497 visit observations

For this study,

- 913 children are randomly assigned as training group
- The other 457 children are assigned as testing group

Data

Variable	testing, N = 457[1]	training, N = 913[1]
Genotype		
Two alleles F508del	251 (55%)	492 (54%)
One allele F508del	160 (35%)	322 (35%)
Others or Unknown	46 (10%)	99 (11%)
Gender		
Female	219 (48%)	474 (52%)
Male	238 (52%)	439 (48%)
Race		
Other	24 (5.3%)	38 (4.2%)
White	433 (95%)	875 (96%)
Ethnicity		
Hispanic	11 (2.5%)	32 (3.6%)
Non-Hispanic	425 (97%)	849 (96%)
Unknown	21	32
Visit number	46 (32, 61)	45 (31, 59)
Age mean	8.18 (6.75, 9.97)	8.06 (6.62, 9.94)
Age baseline	3.13 (3.06, 3.23)	3.13 (3.06, 3.22)
Age final	13.1 (10.5, 15.9)	12.7 (9.9, 16.0)
Follow up years	9.9 (7.3, 12.6)	9.5 (6.8, 12.8)
Height mean	125 (117, 134)	125 (117, 135)
Height baseline	93.8 (91.3, 96.6)	94.0 (91.4, 97.0)
Weight mean	27 (22, 34)	27 (22, 34)
Weight baseline	14.10 (13.00, 15.50)	14.10 (13.20, 15.20)

[1]n (%); Median (IQR)

People-like-me

The traditional predictive modeling:

- A global inference and universal modelling over all available data.
- Overlooking the cultural diversity and genetic heterogeneity for patients

People-like-me method:

- Individualized curving matching
- To use fewer but more similar samples
- To get a higher predictive performance than with more similar matches
- Information of ***the nearest neighbors of predictive mean***

Predictive mean matching

- The imputations created by predictive mean matching follow the data nicely
- With respect to **certain metrics**
- To avoid the dataset noise and model misspecification

Through exhaustive comparisons with predictive mean, for specific target and the most similar matching donor-cohort

Donor-cohort

1. "The chosen threshold"

Choose a threshold, and take all donors agreed within this threshold

2. "The nearest neighbor"

Decide the number of matching donors, and choose this number of donors with minimal metrics.

3. "Single-time" v.s "Multiple-time"
4. Which metric to use

In the past: with predictive mean matching with *single time point* and a *fixed* number of candidate donors.

Example

Using people-like-me method to study the children height growth with EPIC data

Randy

Background

What we got so far

Paper

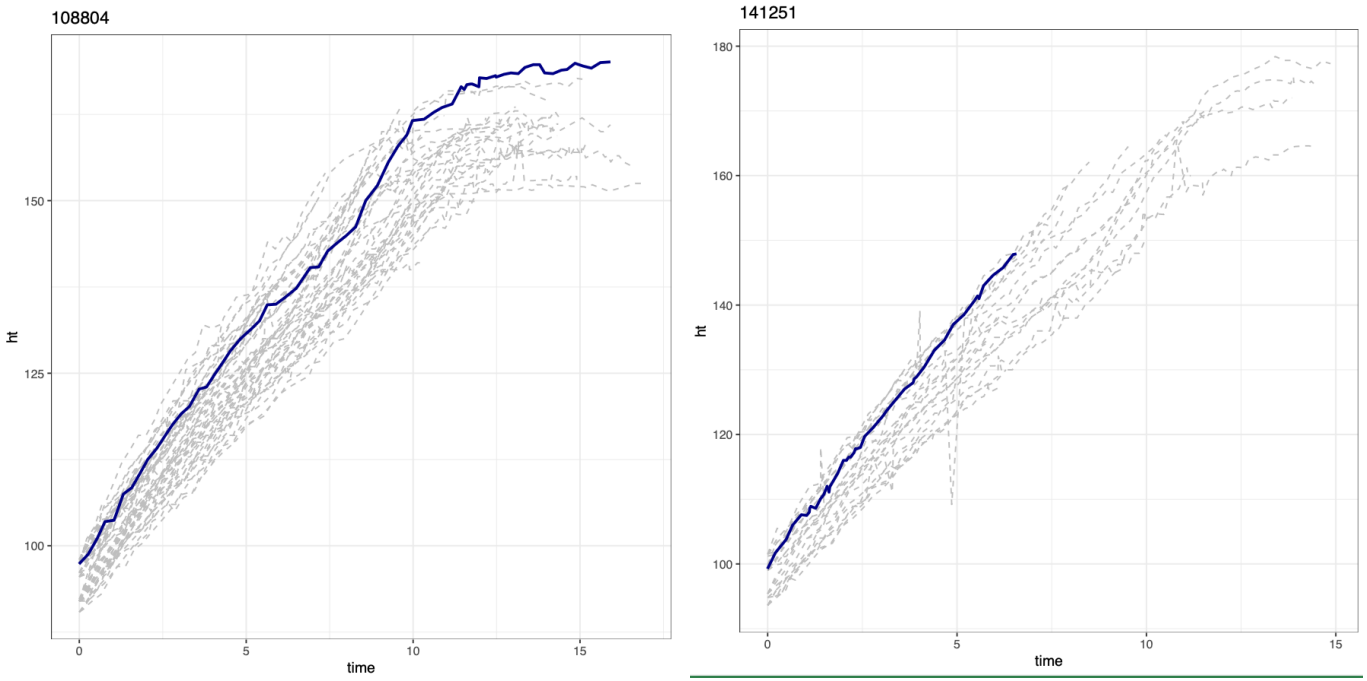


Figure 1: Individual matching trajectoreis example

Mahalanobis Distance

- Mahalanobis distance is a multivariate distance metric, measuring the distance between a point and a distribution.

$$\boldsymbol{D}^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu})$$

Where:

- $\boldsymbol{D}^2$ is the square of the Mahalanobis distance
- $\boldsymbol{x}$ is the vector of the observation (n row in a dataset)
- $\boldsymbol{\mu}$ is the vector of mean values of independent variables (mean of each column)
- $\boldsymbol{\Sigma}$ is the covariance matrix of independent variables.
- The squared Mahalanobis distance, $\boldsymbol{D}^2$, is Chi-square distributed.
- Based on the χ^2 test p-value, we can find the acceptance of given donors.

Steps

- Fitting a `brokenstick` model with training dataset and getting the predictive imputations

```
brokenstick_prediction(formula = "ht ~ time | id",
                       train_data = train,
                       knots = c(5, 10, 12),
                       pred_time = c(2, 4, 6, 8, 10, 12, 14),
                       newdata = test_baseline)
```

- Fitting a `linear model` at for all the subjects with baseline information
- Getting the predictive values at each time point from `linear models`

```
linearized_brokenstick(lm_formula = "`.pred` ~ time_factor * sex + baseline
                       bks_pred = bks_pred)
```

Steps

With each subject from testing dataset,

- Calculating the metrics, e.g. Euclidean distance or Mahalanobis distance, between the target with all the others in training dataset
- Deciding on the set and the size of `donor-cohort`
- Fitting `gamlss` model with `donor-cohort` observations
- Getting the prediction and confidence intervals

```
pred_matching(lb_data = lb_data, obs_data = test_data,
              match_methods = c("mahalanobis", "euclidean", "singletime"),
              match_alpha or match_number,
              gamlss_formula = "ht ~ cs(time, df = 3)",
              gamsigma_formula = "~ cs(time, df = 1)",
              match_plot = TRUE,  predict_plot = TRUE,)
```

Simulation of 20 datasets

	Training						Testing				
	MAD	MSE	RMSD	CI50	CI80	CI95	MAD	MSE	RMSD	CI50	
PLM_Euclidean_n10	2.14	13.65	3.69	55.12%	75.04%	88.75%	2.63	16.68	4.08	49.27%	8
PLM_Mahalanobis_n10	3.69	13.78	3.71	54.56%	76.51%	92.32%	2.82	18.46	4.29	50.28%	8
PLM_Mahalanobis_p95	2.15	12.8	3.58	63.00%	87.92%	96.08%	2.34	15.17	3.89	62.62%	8
PLM_Single_time10_n10	2.19	13.78	3.78	55.37%	81.05%	92.66%	2.77	17.64	4.2	51.01%	7

A larger simulation study with 1000 datasets is going on with OGS Connect

Large Simulation Study with 1000 datasets

Using people-like-me method to study the children height growth with EPIC data

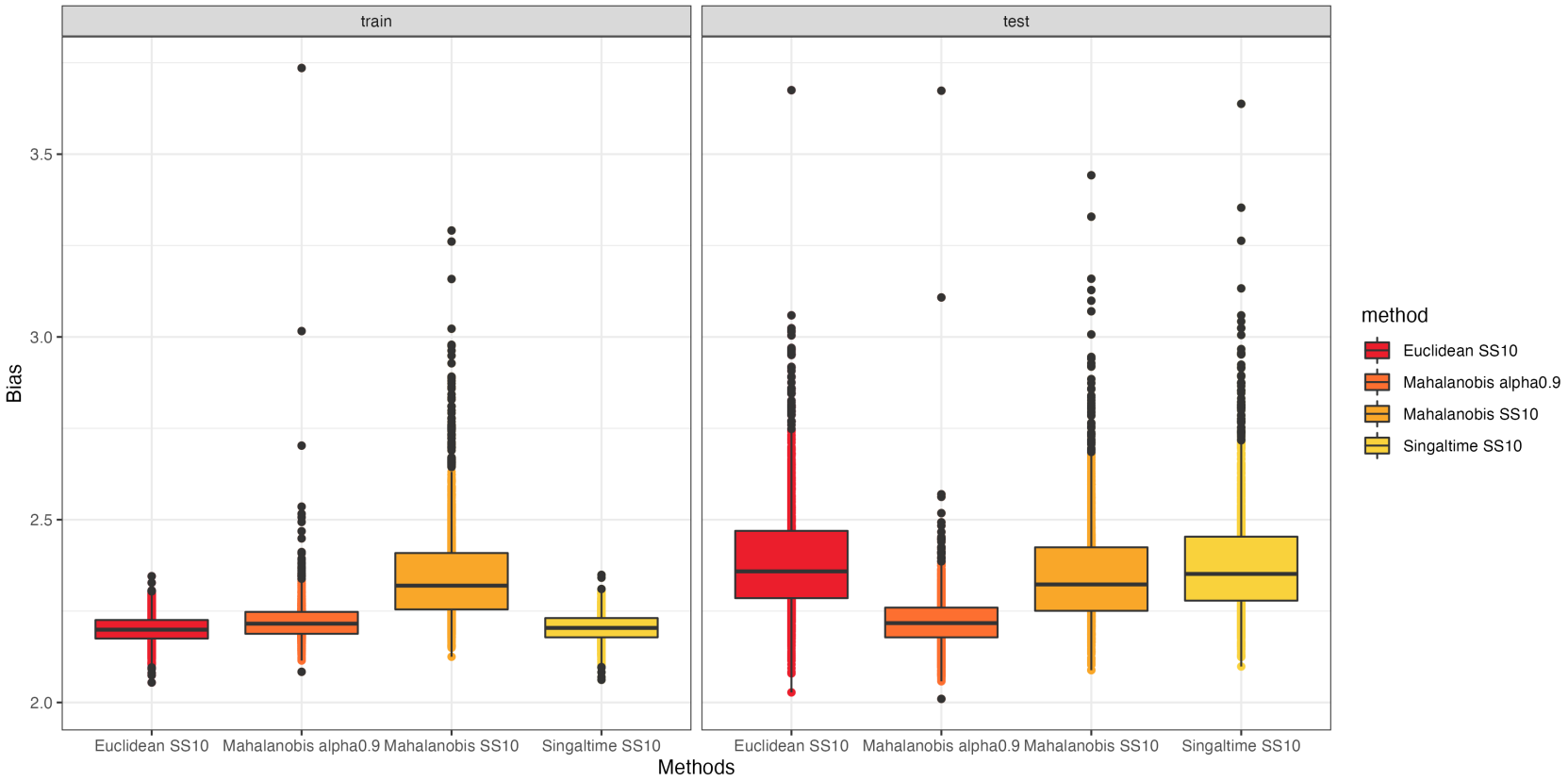
Randy

Background

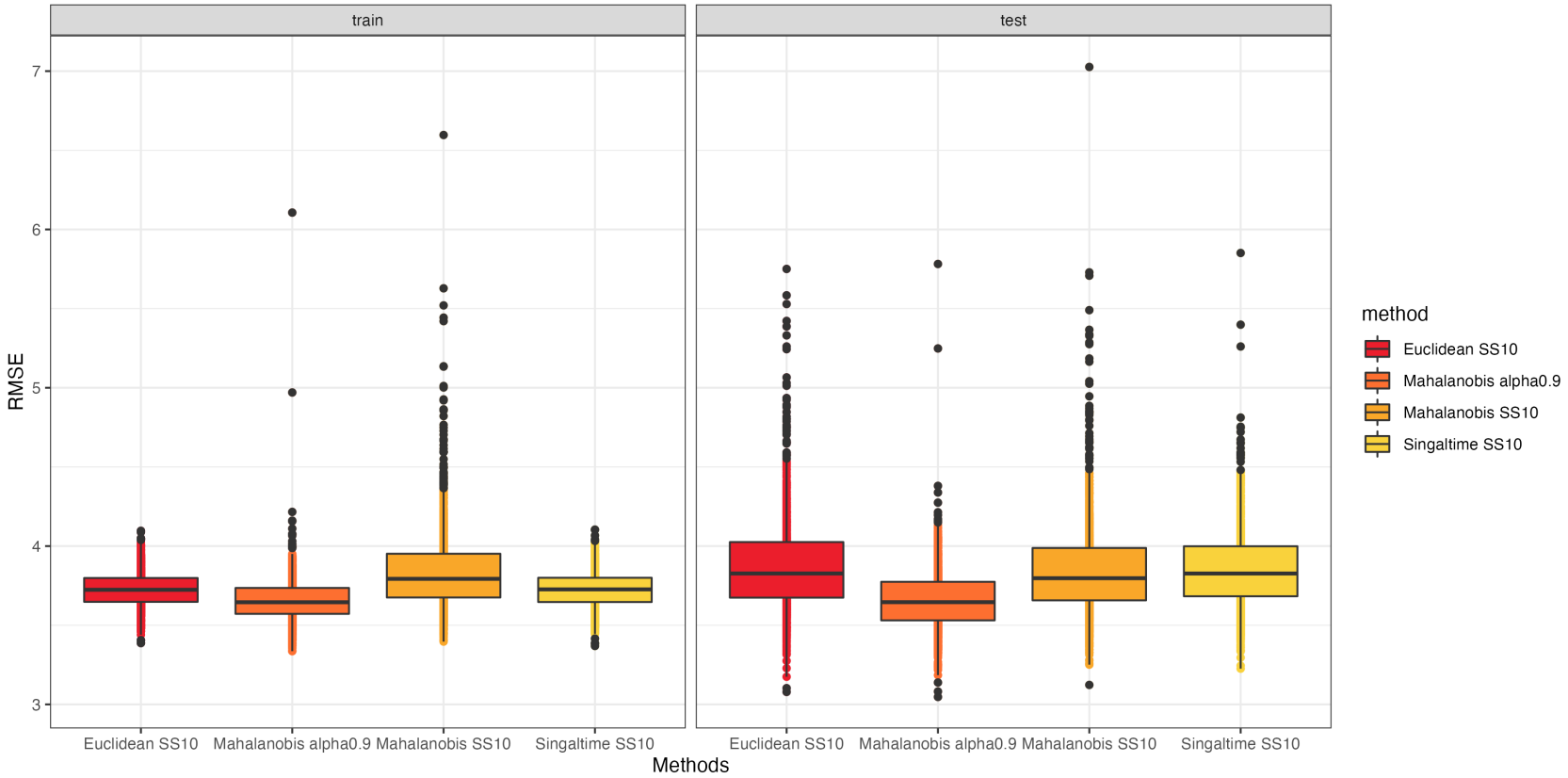
What we got so far

Paper

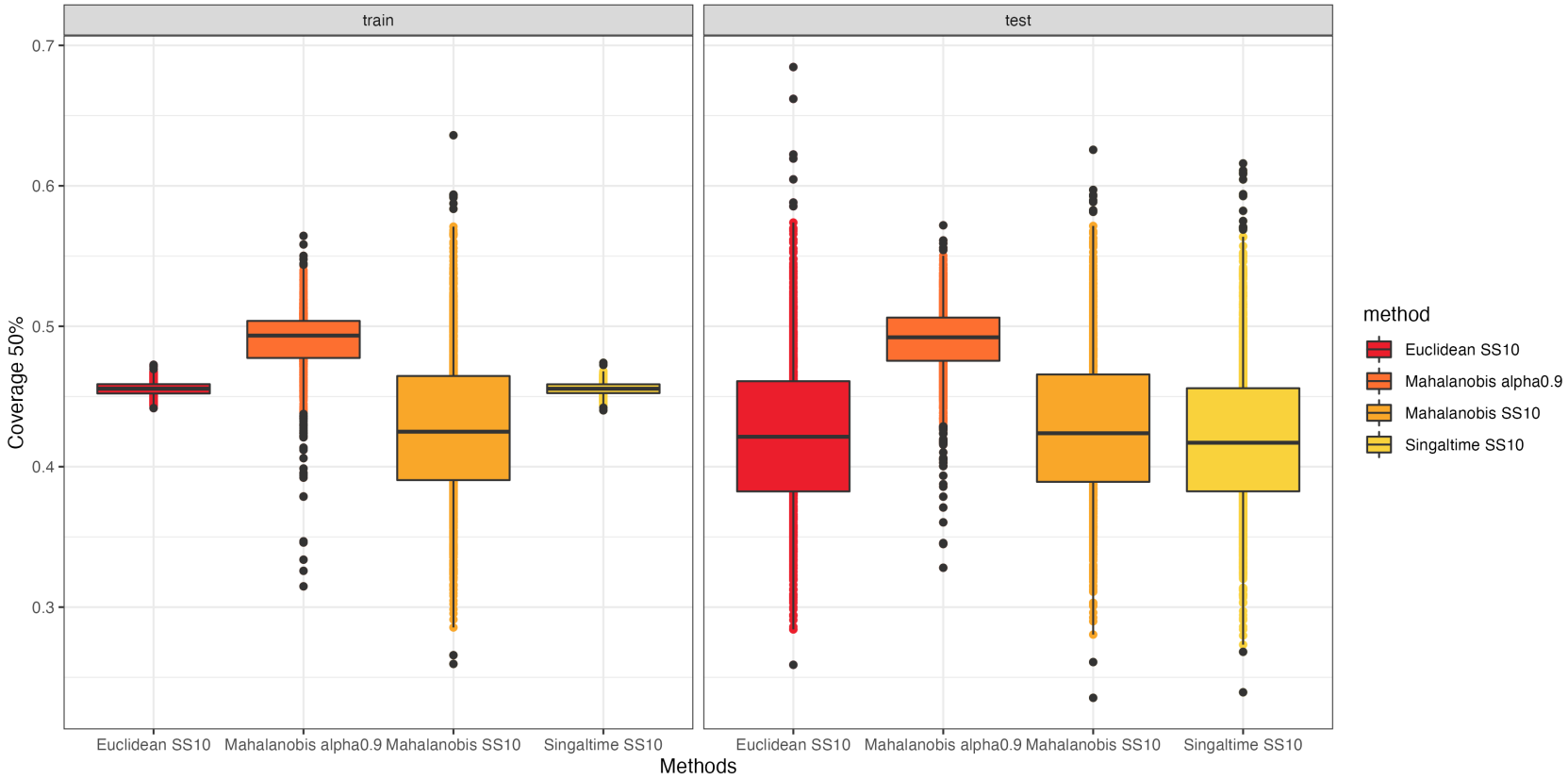
Bias



RMSE


train
test
RMSE
7
6
5
4
3
Euclidean SS10
Mahalanobis alpha0.9
Mahalanobis SS10
Singaltime SS10
Methods
method
Euclidean SS10
Mahalanobis alpha0.9
Mahalanobis SS10
Singaltime SS10

Coverage50


train
test
0.7
0.6
0.5
0.4
0.3
Coverage 50%
Euclidean SS10
Mahalanobis alpha0.9
Mahalanobis SS10
Singaltime SS10
Methods
method
Euclidean SS10
Mahalanobis alpha0.9
Mahalanobis SS10
Singaltime SS10

Coverage80

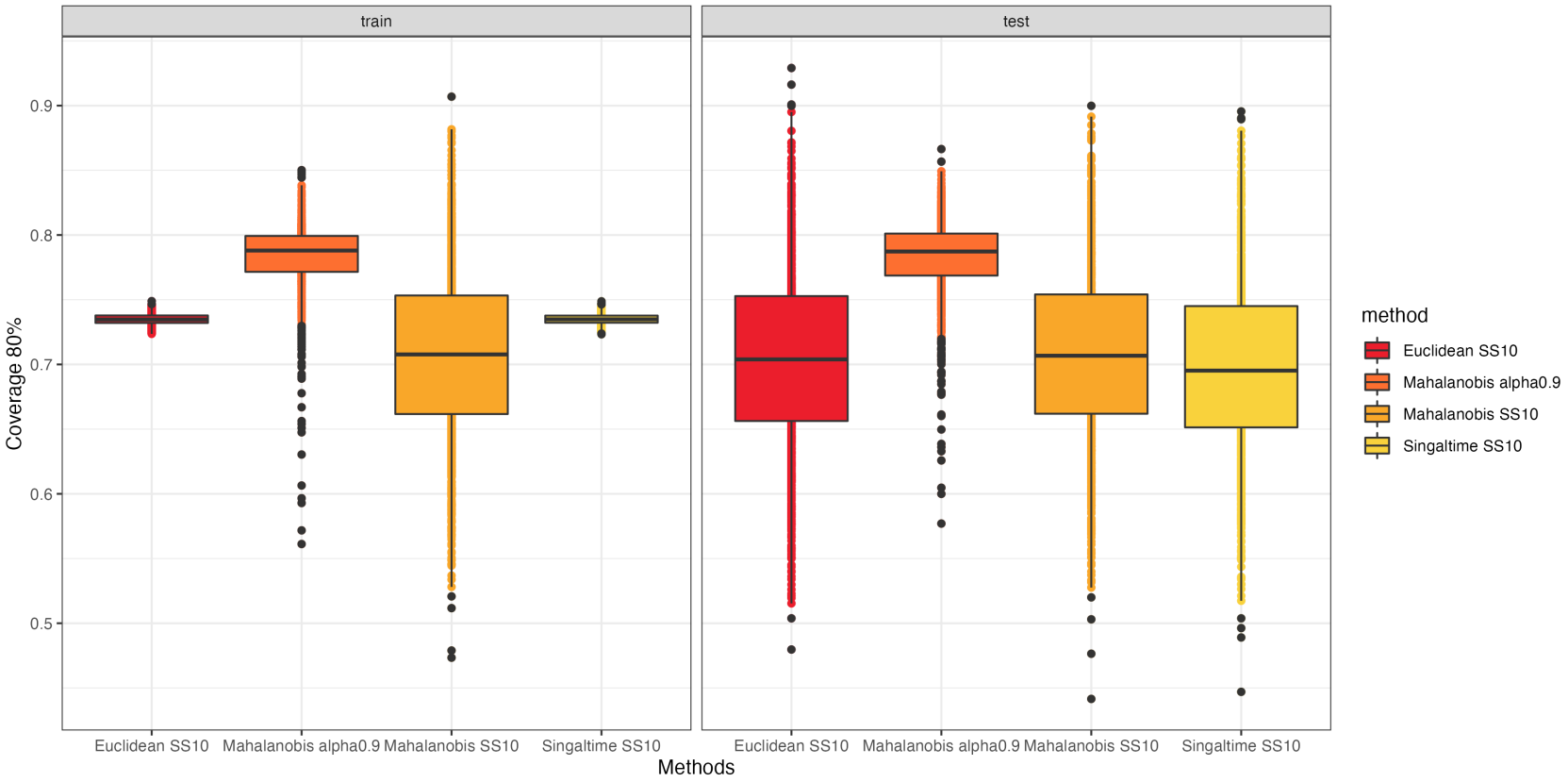
Using people-like-me method to study the children height growth with EPIC data

Randy

Background

What we got so far

Paper



Coverage90

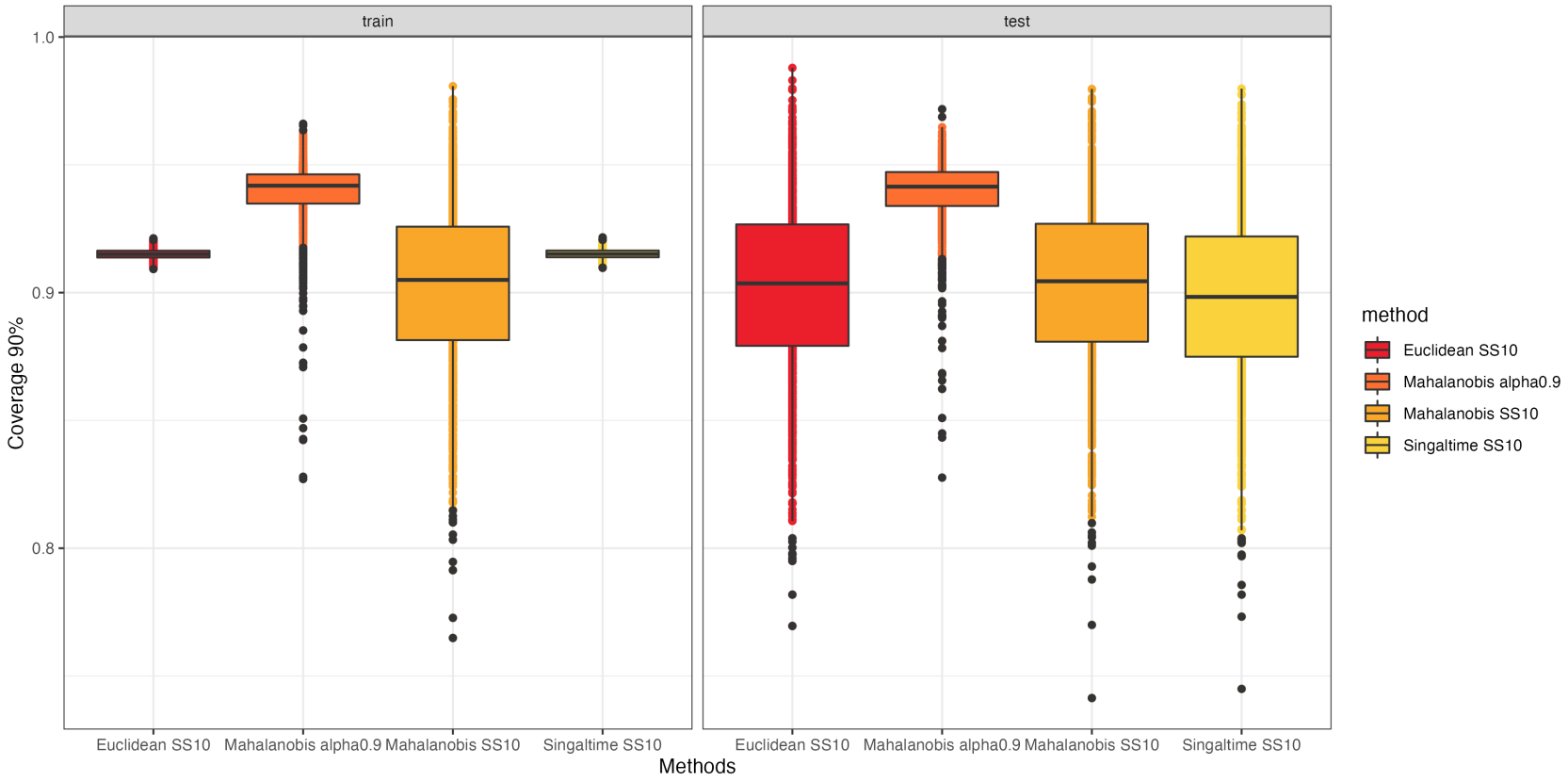
Using people-like-me method to study the children height growth with EPIC data

Randy

Background

What we got so far

Paper



Package and Shinyapp


```
library(plmlmm)

plmlmm::run_shiny()
```

Paper

- Finish the data analysis and simulation study
- Abstract
- Introduction (**)
- Dataset Epic (in detals)
  - table1 (**)
  - sample trajectory (**)
- Methods
  - People-like-me methods
    - Euclidean distance v.s. Mahalanobis distance
    - Single time v.s. Multiple time
    - fixed sample size v.s. unfixed sample size
  - Dynamic prediction
- Results
  - Analysis for Epic data
    - different setting
    - results summary
  - Simulation study 1000 dataset
- Discussion and Extension
  - use a Guassian kernel for Mahalanobis distance
  - parameter tuning why use $\alpha = 0.9$