Robustness to missing data: comparison between mixed-effects model and functional principal component analysis

Corentin Ségalas, Robin Genuer and Cécile Proust-Lima
corentin.segalas@u-bordeaux.fr

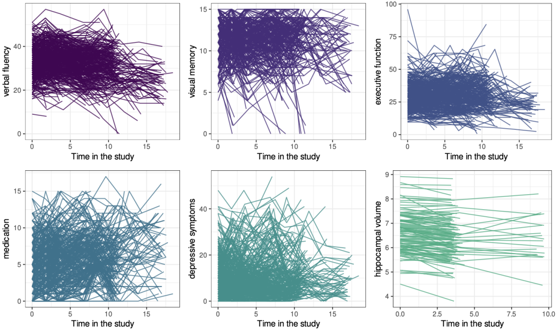
Journées de Biostatistique 2023, Toulouse

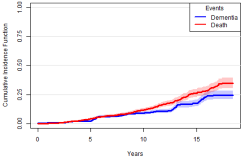
Longitudinal psychometric scores

$y_{ij} = y_i^\star(t_{ij}) + \varepsilon_{ij}$

Time-to-event outcome

$\lambda(t_{ij})$





Joint model: the shared random effect model

Two submodels linked through the random effects b_i:

1. Mixed-effect model for each longitudinal biomarker

$$y_{ij}|b_i = X_{Li}(t_{ij})^\top \beta + Z_i(t_{ij})^\top b_i + \varepsilon_{ij}$$

Joint model: the shared random effect model

Two submodels linked through the random effects b_i:

1. Mixed-effect model for each longitudinal biomarker

$$y_{ij}|b_i = X_{Li}(t_{ij})^\top \beta + Z_i(t_{ij})^\top b_i + \varepsilon_{ij}$$

2. Survival model for time-to-event outcome

$$\lambda_i(t, b_i) = \lambda_0(t) \exp(X_{Ti}(t)^\top \delta + f(t, b_i)^\top \eta)$$

Joint model: the shared random effect model

Two submodels linked through the random effects b_i:

1. Mixed-effect model for each longitudinal biomarker

$$y_{ij}|b_i = X_{Li}(t_{ij})^\top \beta + Z_i(t_{ij})^\top b_i + \varepsilon_{ij}$$

2. Survival model for time-to-event outcome

$$\lambda_i(t, b_i) = \lambda_0(t) \exp(X_{Ti}(t)^\top \delta + f(t, b_i)^\top \eta)$$

Estimation challenging with too many longitudinal predictors

- Huge numerical integration
- Too many predictors in the survival model
- Too many parameters for simultaneous estimation

DynForest: predictors into random survival forest

Predictors into RSF

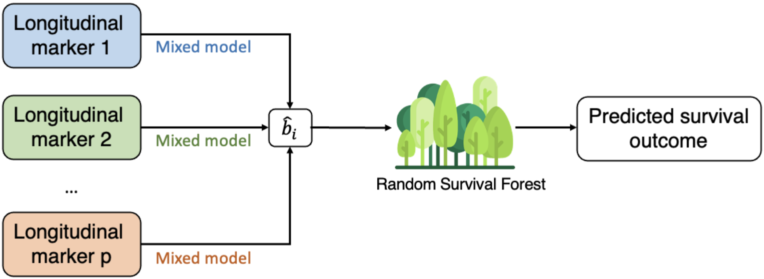
$\surd$ time-independent

$\times$ time-dependent

DynForest: predictors into random survival forest

Predictors into RSF

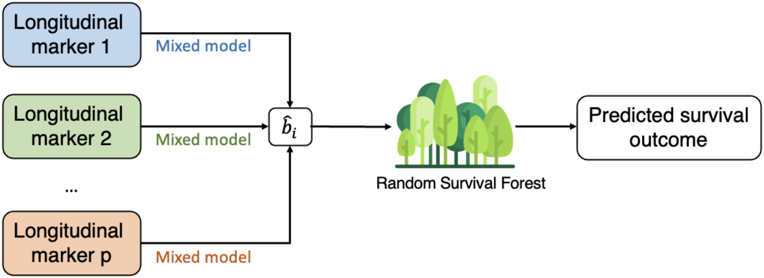
$\surd$ time-independent $\times$ time-dependent



DynForest: predictors into random survival forest

Predictors into RSF

√ time-independent × time-dependent



Semi-parametric

- mixed models
- random survival forest

Statistical context: focus on the longitudinal trajectories

We observe:

- a visit t_{ij} with $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$
- a longitudinal biomarker $y_{ij} = y_i(t_{ij}) = y_i^\star(t_{ij}) + \varepsilon_{ij}$
- a missing indicator r_{ij}, 1 if y_{ij} is observed 0 if not

Statistical context: focus on the longitudinal trajectories

We observe:

- a visit t_{ij} with $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$
- a longitudinal biomarker $y_{ij} = y_i(t_{ij}) = y_i^\star(t_{ij}) + \varepsilon_{ij}$
- a missing indicator r_{ij}, 1 if y_{ij} is observed 0 if not

Missing data

- Missing Completely At Random $p(r_{ij}|y^m, y^o) = p(r_{ij})$
- Misssing At Random $p(r_{ij}|y^m, y^o) = p(r_{ij}|y^o)$
- Missing Not At Random $p(r_{ij}|y^m, y^o) = p(r_{ij}|y^m, y^o)$

Statistical context: focus on the longitudinal trajectories

We observe:

- a visit t_{ij} with $i = 1, \dots, N$ and $j = 1, \dots, n_i$
- a longitudinal biomarker $y_{ij} = y_i(t_{ij}) = y_i^\star(t_{ij}) + \varepsilon_{ij}$
- a missing indicator r_{ij}, 1 if y_{ij} is observed 0 if not

Missing data

- Missing Completely At Random $p(r_{ij}|y^m, y^o) = p(r_{ij})$
- Misssing At Random $p(r_{ij}|y^m, y^o) = p(r_{ij}|y^o)$
- Missing Not At Random $p(r_{ij}|y^m, y^o) = p(r_{ij}|y^m, y^o)$

Dropout

When $r_{ij} = 0$ implies $r_{ik} = 0$ for all $j \leq k \leq n_i$

From functional data...

$$y_{ij} = y_i(t_{ij}) = y_i^\star(t_{ij}) + \varepsilon_{ij}$$

$y_i^\star$ realization of an unknown function f observed with noise on a dense regular grid.

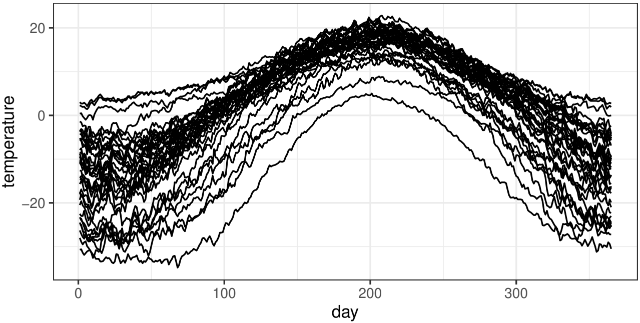


Figure: Average daily temperature from Canadian Weather data, *Functional Data Analysis, Ramsay and Silverman, Springer 2005.*

...to sparse and irregular functional data

$$y_{ij} = y_i(t_{ij}) = y_i^\star(t_{ij}) + \varepsilon_{ij}$$

$y_i^\star$ realization of an unknown random function f observed with noise on a sparse irregular grid.

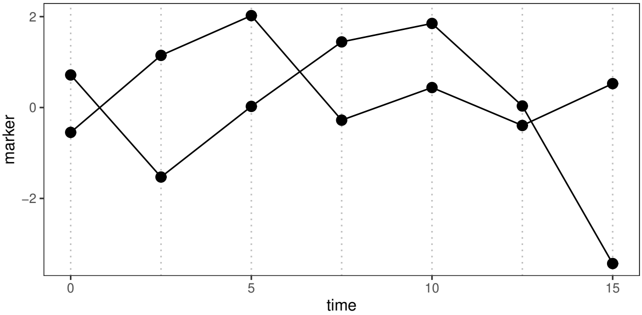


Figure: Sparse irregular functional data

...to sparse and irregular functional data

$$y_{ij} = y_i(t_{ij}) = y_i^\star(t_{ij}) + \varepsilon_{ij}$$

$y_i^\star$ realization of an unknown random function f observed with noise on a sparse irregular grid.

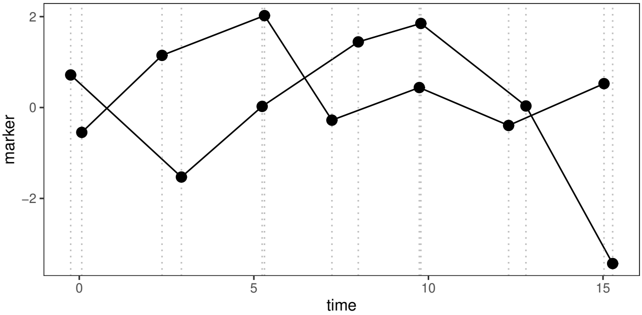


Figure: Sparse irregular functional data

...to sparse and irregular functional data

$$y_{ij} = y_i(t_{ij}) = y_i^\star(t_{ij}) + \varepsilon_{ij}$$

$y_i^\star$ realization of an unknown random function f observed with noise on a sparse irregular grid.

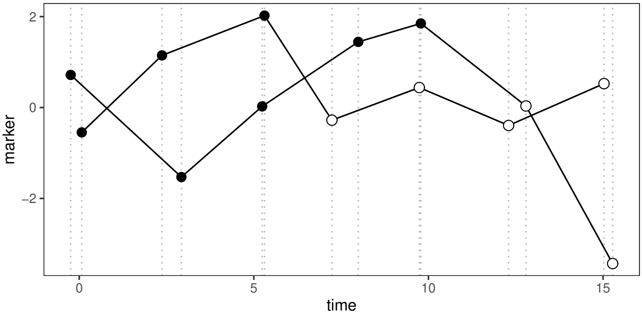


Figure: Sparse irregular functional data

Principal Component Analysis

Classic PCA

Project the scatterplot $(y_i)_{i=1,\ldots,N}$ from $\mathbb{R}^k$ to a lower dimensional space with an orthogonal basis while maximizing the variability.

Principal Component Analysis

Classic PCA

Project the scatterplot $(y_i)_{i=1,\dots,N}$ from $\mathbb{R}^k$ to a lower dimensional space with an orthogonal basis while maximizing the variability.

Noisy data

Data after PCA

Functional Principal Component Analysis

Karhunen-Loève decomposition:

$$y_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t) \quad i = 1, \ldots, N, \quad t \in \mathbb{R}$$

- μ mean function
- ψ_k orthonormal eigenfunctions of the covariance operator
- ξ_{ik} principal component scores

Functional Principal Component Analysis

Karhunen-Loève decomposition:

$$y_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t) \quad i = 1, \ldots, N, \quad t \in \mathbb{R}$$

- μ mean function
- ψ_k orthonormal eigenfunctions of the covariance operator
- ξ_{ik} principal component scores

Estimation

$\hat{\mu}$, $\hat{\xi}_{ik}$ and $\hat{\psi}_k$ for $k = 1, \ldots, K$ with PACE algorithm.

Functional Principal Component Analysis

Karhunen-Loève decomposition:

$$y_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t) \quad i = 1, \ldots, N, \quad t \in \mathbb{R}$$

- μ mean function
- ψ_k orthonormal eigenfunctions of the covariance operator
- ξ_{ik} principal component scores

Estimation

$\hat{\mu}$, $\hat{\xi}_{ik}$ and $\hat{\psi}_k$ for $k = 1, \ldots, K$ with PACE algorithm.

Prediction (for fixed K)

Plug-in $\hat{\mu}(t)$, $\hat{\xi}_{ik}$ and $\hat{\psi}_k(t)$ into the KL decomposition.

Functional Principal Component Analysis

Karhunen-Loève decomposition:

$$y_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t) \quad i = 1, \ldots, N, \quad t \in \mathbb{R}$$

- μ mean function
- ψ_k orthonormal eigenfunctions of the covariance operator
- ξ_{ik} principal component scores

Estimation

$\hat{\mu}$, $\hat{\xi}_{ik}$ and $\hat{\psi}_k$ for $k = 1, \ldots, K$ with PACE algorithm.

Prediction (for fixed K)

Plug-in $\hat{\mu}(t)$, $\hat{\xi}_{ik}$ and $\hat{\psi}_k(t)$ into the KL decomposition.

Robustness to dropout

What is FPCA behaviour with missing data?

A parallel between FPCA and mixed models

FPCA

- Scores ξ_{ik}
- Number of FPC
- Mean function
- FPC $\phi_k(t)$

- Non parametric
- No inference tools

- Unknown robustness to NA

Mixed models

- Random effects b_{ik}
- Number of random effects
- Marginal mean
- Covariate $X_k(t)$

- Parametric
- Inference tools

- Robustness to MAR data

Simulation study design

Aim

Evaluate robustness of FPCA to dropout.

Data Generation

- $N = 700$ subjects
- each 1 or 2 year up to 12
- dropout of 30% or 60%
- MAR and MNAR

Estimand

$\hat{y}_{ij}$ for missing observations

Methods

FPCA, LMM and JM

Simulation study design

Aim

Evaluate robustness of FPCA to dropout.

Data Generation

- $N = 700$ subjects
- each 1 or 2 year up to 12
- dropout of 30% or 60%
- MAR and MNAR

Estimand

$\hat{y}_{ij}$ for missing observations

Methods

FPCA, LMM and JM

Data Generation

- $N = 200$ subjects
- each 1 or 2 year up to 12
- dropout of 30% or 60%
- MAR and MCAR

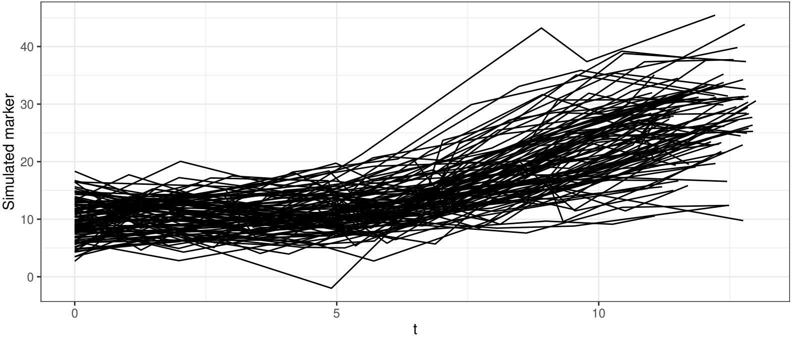
Estimand

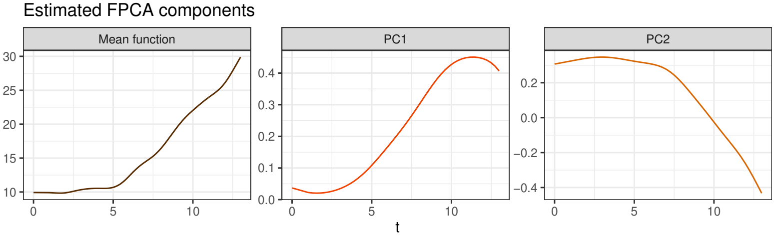
$\hat{\mu}(t)$ and $\hat{\xi}_k(t)$

Methods

FPCA

Simulation study design: data generation mechanisms





Simulation study design: missing data

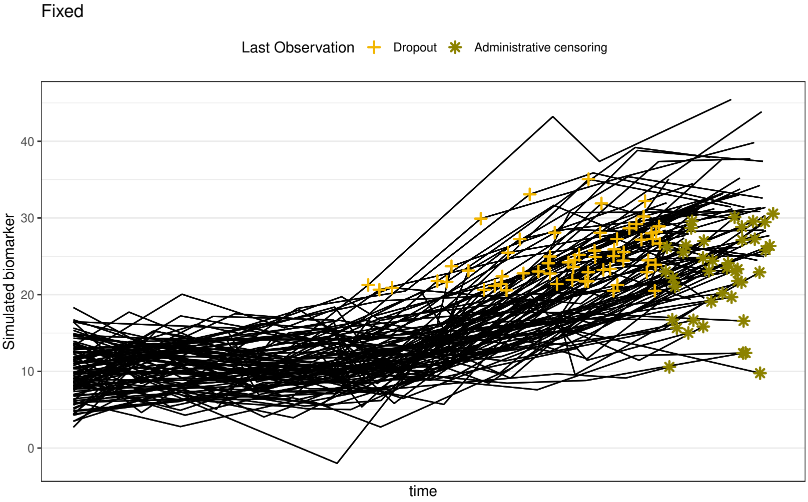


Figure: 100 simulated longitudinal trajectories: fixed threshold and increasing probability of dropout (MAR, MNAR and MCAR)

Simulation study design: missing data

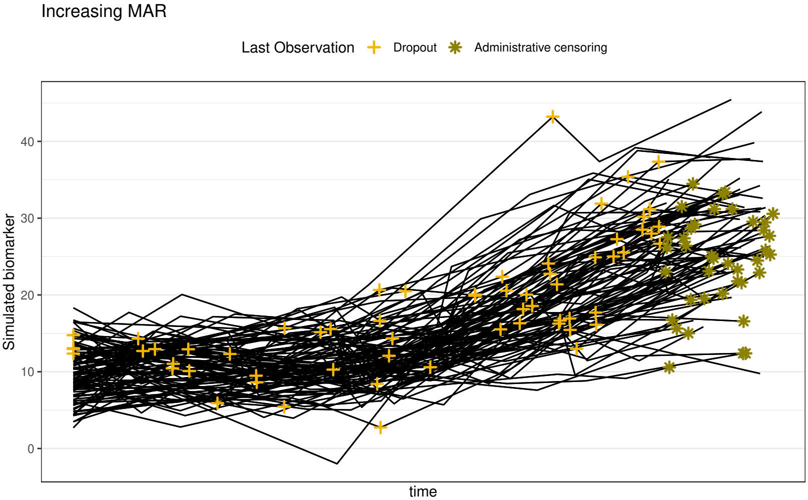


Figure: 100 simulated longitudinal trajectories: fixed threshold and increasing probability of dropout (MAR, MNAR and MCAR)

Simulation study design: missing data

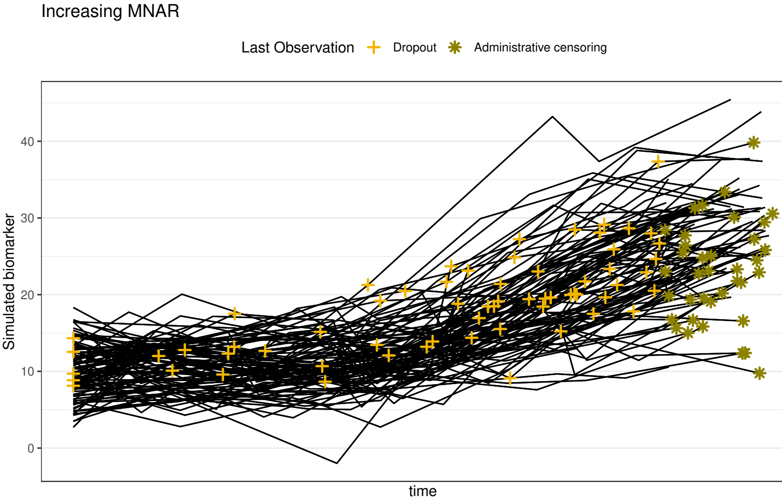


Figure: 100 simulated longitudinal trajectories: fixed threshold and increasing probability of dropout (MAR, MNAR and MCAR)

Simulation study design: missing data

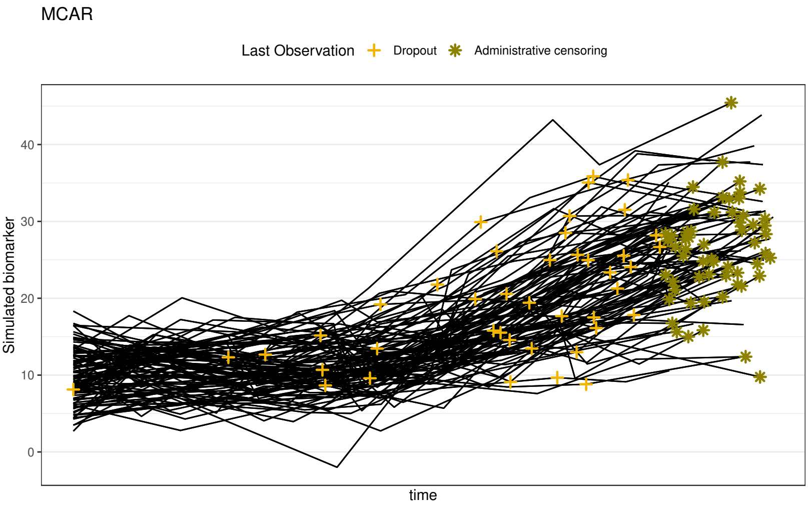


Figure: 100 simulated longitudinal trajectories: fixed threshold and increasing probability of dropout (MAR, MNAR and MCAR)

Simulation study results (1)

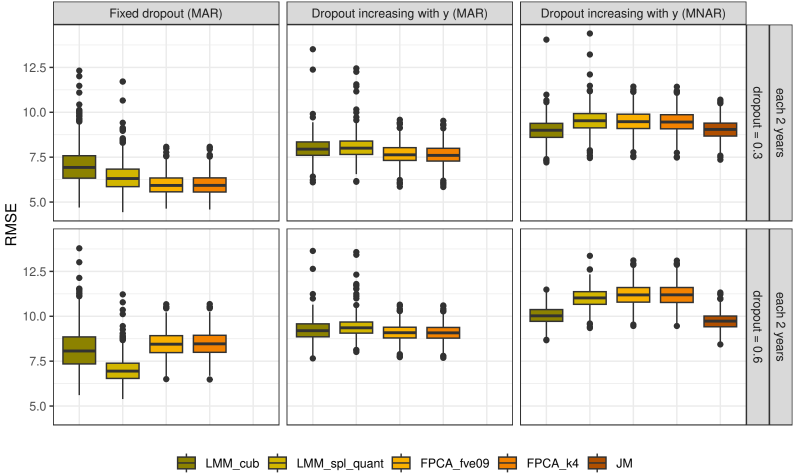


Figure: Relative Mean Square Error for the prediction of the missing y using FPCA, LMM and JM (only in the MNAR case).

Simulation study results (1)

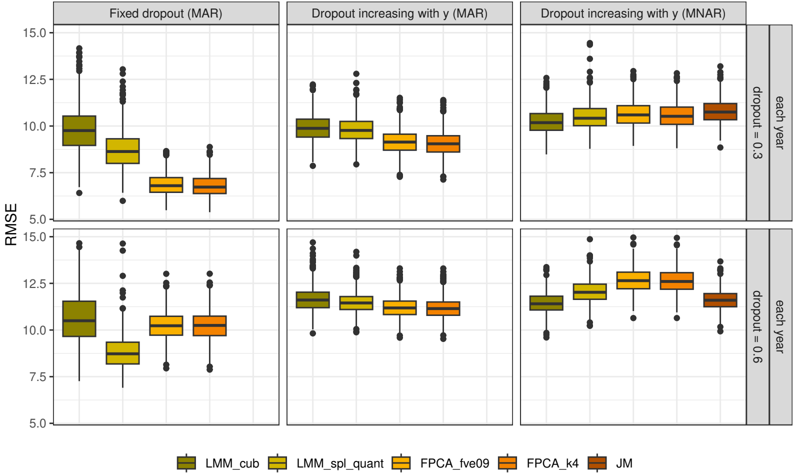


Figure: Relative Mean Square Error for the prediction of the missing y using FPCA, LMM and JM (only in the MNAR case).

Simulation study results (2)

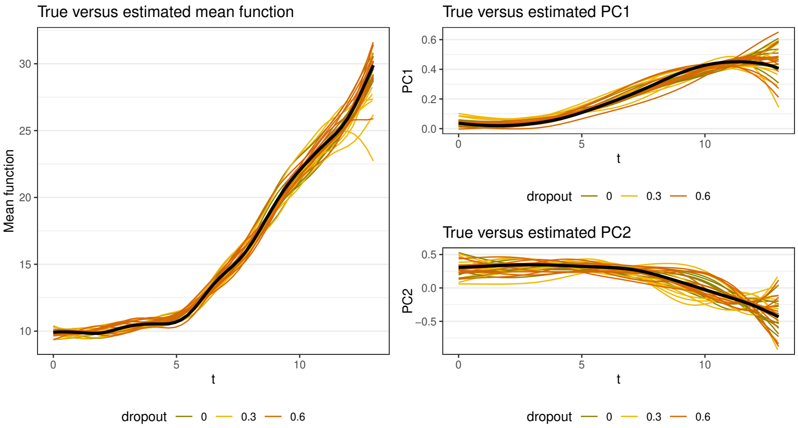


Figure: Estimated versus true mean function and functional principal components (MCAR and MAR).

Simulation study results (2)

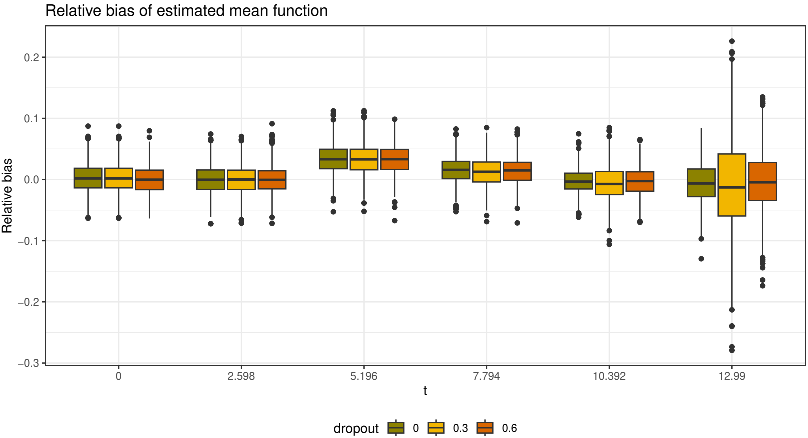


Figure: Estimated versus true mean function and functional principal components (MCAR and MAR).

Simulation study results (2)

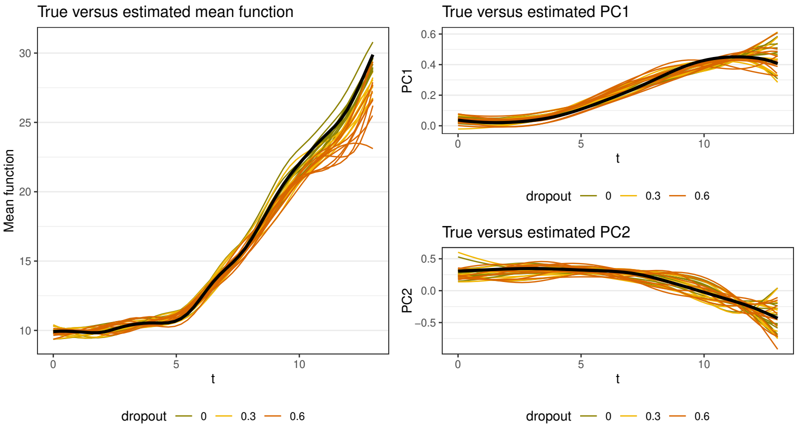


Figure: Estimated versus true mean function and functional principal components (MCAR and MAR).

Simulation study results (2)

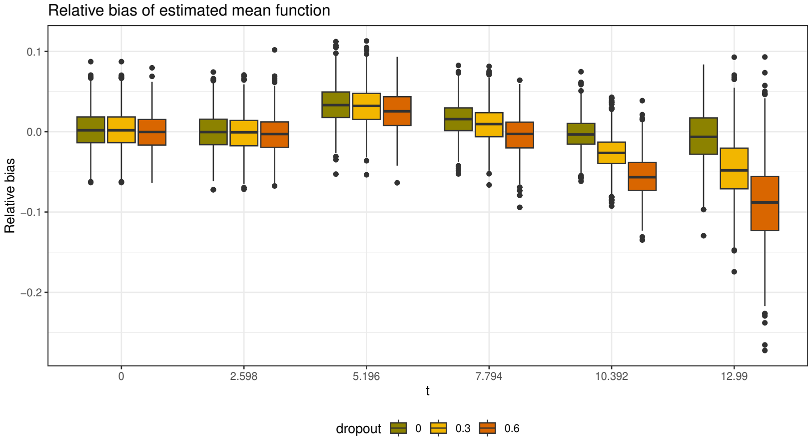


Figure: Estimated versus true mean function and functional principal components (MCAR and MAR).

Application on real data

Application on real data

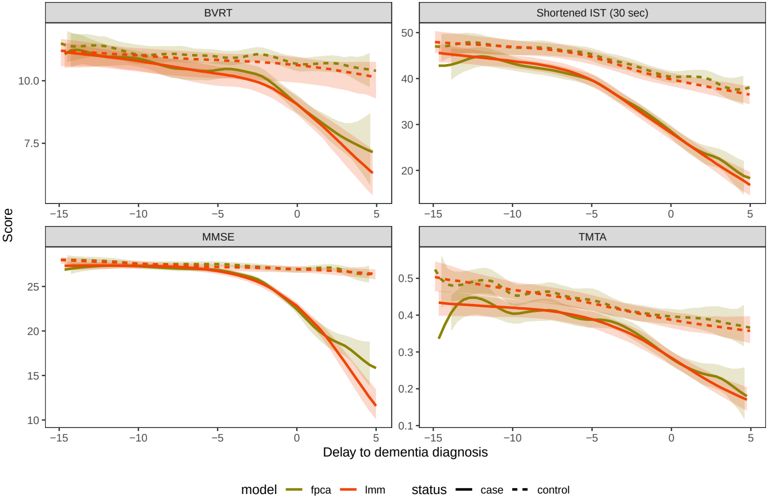


Figure: Mean function and 95%CI of LMM (spline) and FPCA ($K = 2$) on cognitive markers from a 3C nested case-control study (N=330).

Discussion and perspectives

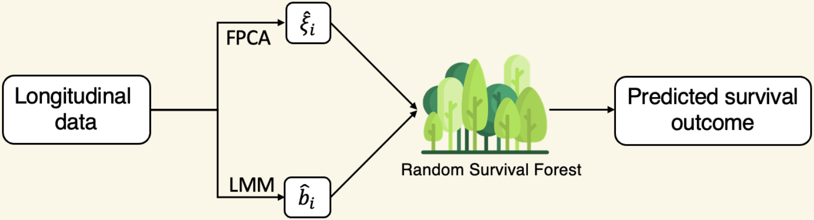
- Longitudinal data as **sparse and irregular functional data**
- FPCA a **nonparametric flexible approach** (`fdapace`, `MFPCA`)
- FPCA is a **descriptive** approach, no inference
- Robust to dropout (MAR)

Discussion and perspectives

- Longitudinal data as **sparse and irregular functional data**
- FPCA a **nonparametric flexible approach** (`fdapace`, `MFPCA`)
- FPCA is a **descriptive** approach, no inference
- Robust to dropout (MAR)

Perspectives

Use the estimated scores $\hat{\xi}_i$ as input of a predictive model.



Acknowledgement and funding

The DynForest family

Cécile Proust-Lima

Robin Genuer

Anthony Devaux

Funding

Postdoc grant from the Public Health Data Science Bordeaux Network - Université de Bordeaux

Main references and R packages

- **Ramsay J.O. and Silverman B.W.**, Functional Data Analysis, Springer (2005)
- **Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S.**, Random survival forests. *The Annals of Applied Statistics*, (2008)
- **Yao F., Müller H.-G., Wang J.-L.**, Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, (2005)
- **Devaux A., Helmer C., Dufouil C., Genuer R., Proust-Lima C.**, Random survival forests for competing risks with multivariate longitudinal endogenous covariates. *Statistical Methods in Medical Research*, (2023)

R packages: `fdapace`, `MFPCA`, `lcmm`, `JM`, `splines`, `tidyverse`.

Forest Run versus Forrest Run

```
>     rf <- randomForest(V4 ~ ., data = Ozone, na.action = na.omit)
> randomForest(V4 ~ ., data = Ozone, na.action = na.omit)

Call:
 randomForest(formula = V4 ~ ., data = Ozone, na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 21.47889
                    % Var explained: 67.82
>
```

Thank you!