Computational Statistics

Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model : Application to HIV dynamics model

Welschinger Lilian and Pujol Romain
MVA 2024/2025

Table of contents

Table of contents

Introduction

Figure – Two patients infected, but not at the same level.

The **viral load** is a relevant indicator for the study of drugs, in particular for the HIV.

Introduction

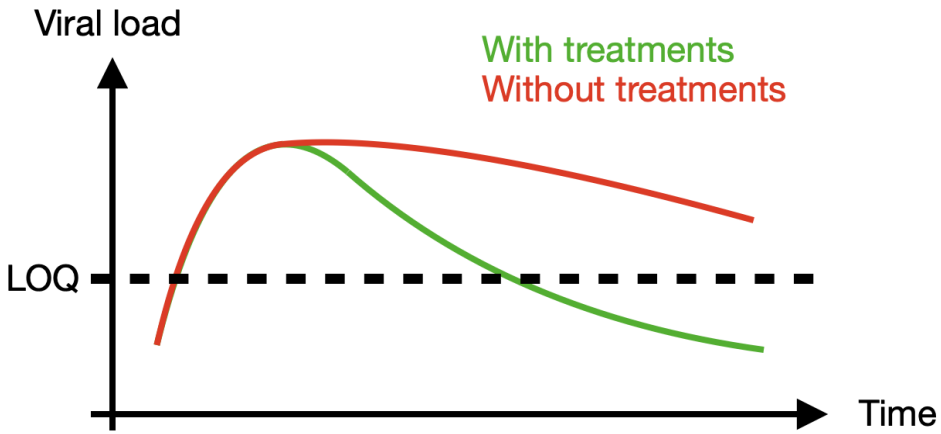


Figure – Example of what could be the evolution of the viral load with of without treatments.

Main issue : measuring instruments often have a limit below (writen LOQ).

Introduction

Because of these limits, we lose information and then there is a bias in the analysis.

Question : How can these censored data be modelled in a non-linear mixed-effects framework in the context of HIV dynamics ?

Article contribution : Extension of the *SAEM* algorithm.

Table of contents

Theoretical aspect

The model

The Nonlinear mixed effects model :

$$\begin{cases} y_{ij} & = f(\phi_i, t_{ij}) + g(\phi_i, t_{ij})\,\epsilon_{ij} \\ \epsilon_{ij} & \sim \mathcal{N}(0, \sigma^2 I_{n_i}) \\ \phi_i & = X_i\mu + b_i, \text{ with } b_i \sim \mathcal{N}(0, \Omega) \end{cases}$$

The parameters of the models are : $\theta = (\mu, \sigma^2, \Omega)$.

Question : We would like to use the $(\mathbf{EM})$ algorithm to find the optimal parameters, but how do we do this when the $\mathbf{E}$ step has no analytic form ?

Theoretical aspect

SAEM algorithm

We denote the complete data (y, z), where z are the non-observed data.

The SAEM algorithm is valid under the following main assumption :

Assumption : The log-likelihood of the complete data $L_c(y, z; \theta)$ belongs to the regular curved exponential family :

$$L_c(y, z; \theta) = -\Lambda(\theta) + \langle S(y, \phi), \Phi(\theta) \rangle$$

Description of the SAEM algorithm

The $\mathbf{E}$ step become :

- Simulation ($\mathbf{S}$) step : draw the non-observed data $z^{(m)}$ with the conditional distribution $p\left(z|y;\hat{\theta}_m\right)$.
- Stochastic approximation ($\mathbf{SA}$) step : approximate evaluation of $\mathbb{E}\left(S(y,z)|\hat{\theta}_{m-1}\right)$ recursively via s_m :

$$s_m = s_{m-1} + \gamma_m \left(S(y, z^{(m)}) - s_{m-1}\right)$$

And then $\mathbf{M}$ step, we update the estimation of θ :

$$\hat{\theta}_m = \underset{\theta\in\Theta}{\text{argmax}} \left(-\Lambda(\theta) + \langle s_m, \Phi(\theta)\rangle\right)$$

Theoretical aspect

Gibbs algorithm

In our case, $z = (\phi, y^{cens})$. We proceed to the $\mathbf{S}$ step in two steps :

- We use Metropolis-Hastings (M-H) algorithm to simulate $\phi^{(m)}$ with $p\left(\cdot|y^{(obs)}, y^{cens(m-1)}; \hat{\theta}_{m-1}\right)$.
- We simulate $y^{cens(m)}$ with the posterior right-truncated Gaussian distribution $p\left(\cdot|y^{(obs)}, \phi^{(m)}; \hat{\theta}_{m-1}\right)$.

For the (M-H) we use 3 differents proposal distribution :

- A certain Gaussian distribution,
- A multidimensional random walk,
- A succession of p unidimensional Gaussian random walks (where we update each components of ϕ successively.

Theoretical aspect

Model considered

In the following we will study the model where the functions are as follows :

$$f(\phi_i, t_{ij}) = \log_{10}\left(P_{1i}\mathrm{e}^{-\lambda_{1i}t_{ij}} + P_{2i}\mathrm{e}^{-\lambda_{2i}t_{ij}}\right) \text{ and } g == 1$$

Then, the log-likelihood of the complete data $L_c(y, z; \theta)$ belongs to the regular curved exponential family with parameters (to within one additive constant) :

$$\begin{cases} \Lambda(\theta) & = \frac{k}{2}\log\left(\sigma^2\right) + \frac{N}{2}\log\det\left(\Omega\right) + \frac{1}{2}\sum_{i=1}^{N}(X\mu)^T\Omega^{-1}X\mu \\ S(y,\phi) & = \left(\sum_{i=1}^{N}\phi_i, \sum_{i=1}^{N}\phi_i^2, \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(y_{ij} - f\left(\phi_i, t_{ij}\right)\right)^2\right)^T \\ \Phi(\theta) & = \left(\frac{1}{2}\Omega^{-1}X\mu, -\frac{1}{2}\tilde{\omega}, -\frac{1}{2\sigma^2}\right)^T \end{cases}$$

Table of contents

Experiments

Convergence of the parameters

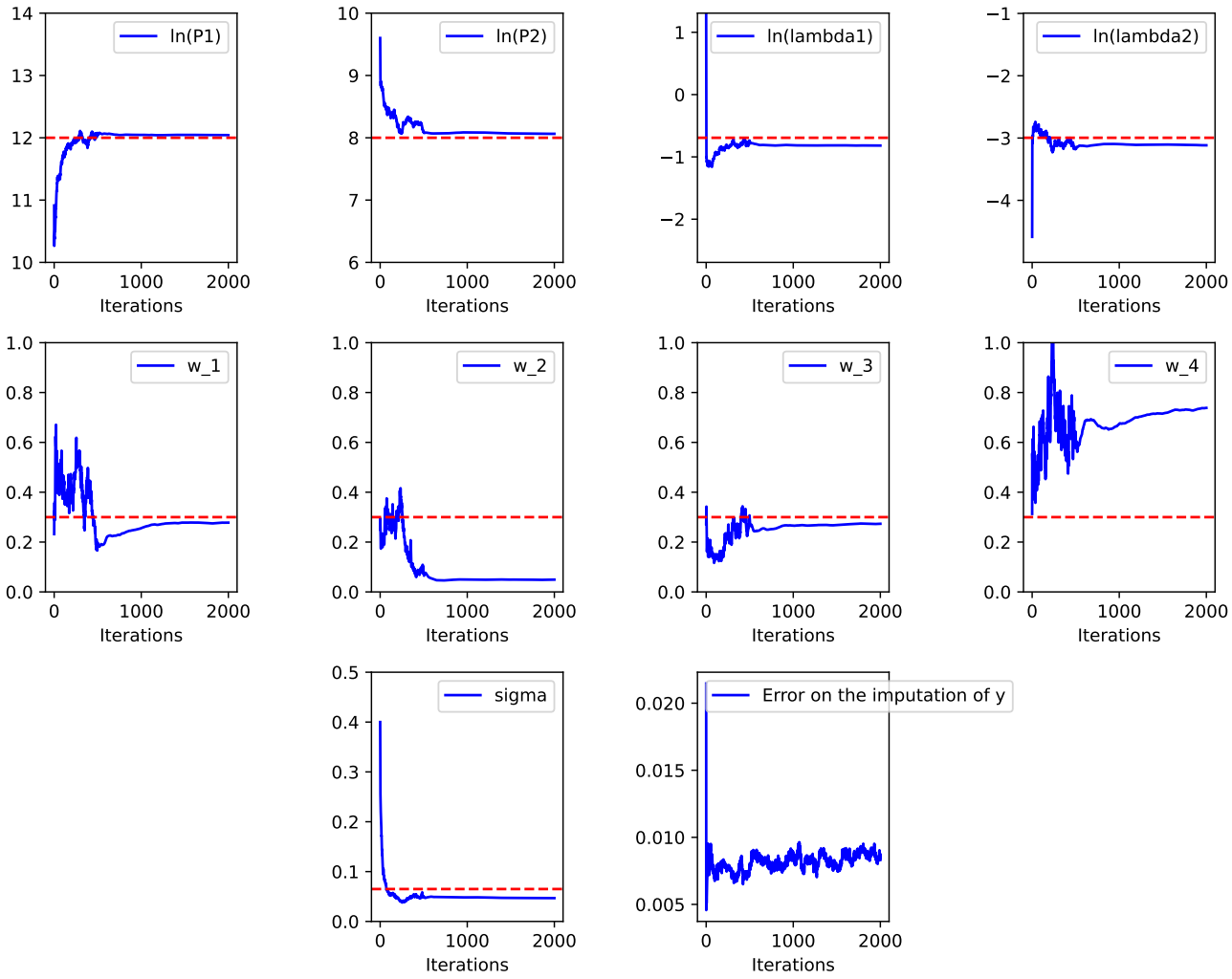


Figure – Parameters convergence

Experiments

We ran 10 times the simulation for each proposal and computed the mean biases :

Parameters	$\ln(P_1)$	$\ln(P_2)$	$\ln(\lambda_1)$	$\ln(\lambda_2)$	ω_1^2	ω_2^2	ω_3^2	ω_4^2	σ^2
Error %	0%	3%	1%	4%	30%	31%	8%	9%	38%

Table – Proposal 1 : gaussian distribution

Parameters	$\ln(P_1)$	$\ln(P_2)$	$\ln(\lambda_1)$	$\ln(\lambda_2)$	ω_1^2	ω_2^2	ω_3^2	ω_4^2	σ^2
Error %	1%	0%	7%	5%	11%	32%	13%	43%	29%

Table – Proposal 2 : multidimensional random walk

Parameters	$\ln(P_1)$	$\ln(P_2)$	$\ln(\lambda_1)$	$\ln(\lambda_2)$	ω_1^2	ω_2^2	ω_3^2	ω_4^2	σ^2
Error %	1%	1%	20%	9%	44%	78%	13%	13%	4%

Table – Proposal 3 : p unidimensional random walk

Experiments

From our simulations, we can state that proposals 1 and 2 (updating ϕ_i for each patient) are way faster than proposal 3 (updating ϕ_{ij} for each patient and each parameter) and more precise.

Experiments

The following figure presents the real trend $\log_{10}(\lambda_1 e^{-P_1 t} + \lambda_2 e^{-P_2 t})$ and the one we found with SAEM algorithm, $\mu = (\lambda_1, \lambda_2, P_1, P_2)$ is a parameter to estimate.

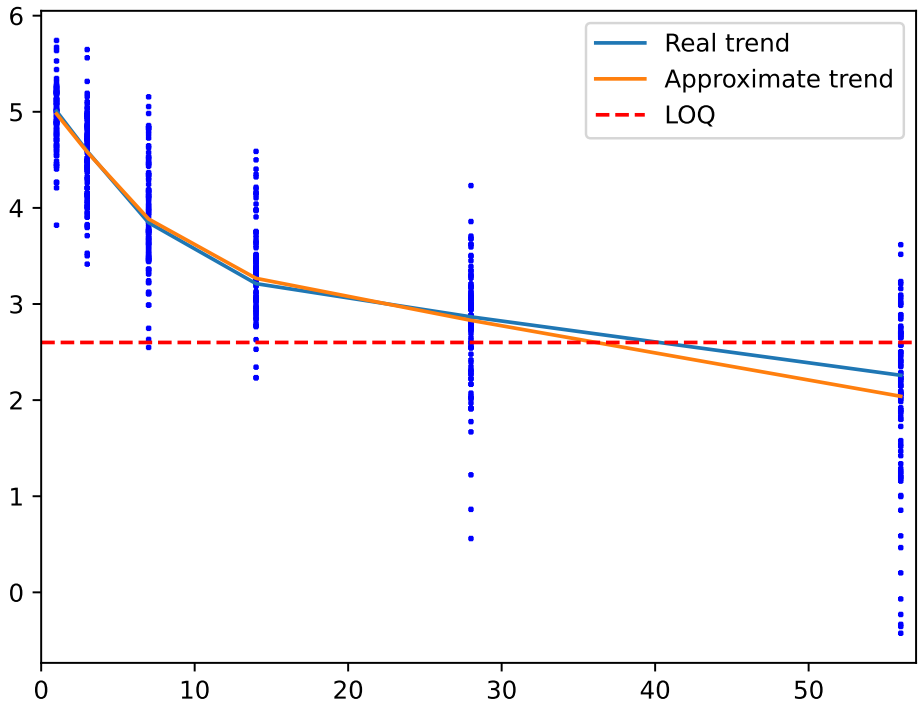


Figure – Estimated trend

Table of contents

New first guesses

First imputation of $y^{censored}$ and ϕ

SAEM relies on values that we keep updating over and over, especially $y^{censored}$ and ϕ. We thought of an idea to imput both $y^{censored}$ and ϕ. We know that :

$$f\left(\phi_i, t_j\right) = \log_{10}\left(P_{1i}e^{-\lambda_{1i}t_j} + P_{2i}e^{-\lambda_{2i}t_j}\right),$$

and that it is meant to be close to the actual y_{ij}.

New first guesses

First imputation of $y^{censored}$ and ϕ

So we want to minimize :

$$\sum_{(i,j)\in I_{obs}} \|y_{ij} - \log_{10}\left(P_{1i}e^{-\lambda_{1i}t_j} + P_{2i}e^{-\lambda_{2i}t_j}\right)\|^2,$$

this is separable for all i and we will find a starting $\bar{\phi}_i$ as :

$$\bar{\phi}_i = \arg\min_{\phi_i} \sum_{(i,j)\in I_{obs}} \|y_{ij} - \log_{10}\left(P_{1i}e^{-\lambda_{1i}t_j} + P_{2i}e^{-\lambda_{2i}t_j}\right)\|^2,$$

where i is fixed in the sum.

New first guesses

First imputation of $y^{censored}$ and ϕ

We can imput the missing values with $\bar{\phi}_i$:

$$y_{ij} = f\left(\bar{\phi}_i, t_j\right), \forall (i,j) \in I_{cens}.$$

When the imputation is not feasible, we add another control on the value (e.g. if the imputed value is above LOQ or below 0, we decide to set it respectively at LOQ and 0).

New first guesses

First imputation of $y^{censored}$ and ϕ

Used data : a denser time table, 200 generations of 40 individuals

Type of imputation	Mean $L2$ error over the 200 generations
Imputation 0	4.64
Imputation $LOQ/2$	0.89
Imputation LOQ	0.53
Fitted imputation	**0.42**

Table – What we gain on $y^{censored}$

For ϕ, we gain a legitimate first guess that we did not have before.
Limitation : we need a lot of observed data to properly fit the curve.

New first guesses

Tests

How does this first guess improve the results of the SAEM algorithm ?

Parameters	Initial start	Bias	RMSE
$\ln(P_1)$	random	0.03	0.03
	first guess	**0.02**	**0.02**
$\ln(P_2)$	random	0.08	0.09
	first guess	**0.07**	**0.08**
$\ln(\lambda_1)$	random	0.44	0.47
	first guess	**0.37**	**0.40**
$\ln(\lambda_2)$	random	**0.20**	0.21
	first guess	0.21	**0.21**

Table – Relative bias and relative root mean square error (RMSE) of the estimated parameters evaluated from 15 simulated trials

New first guesses

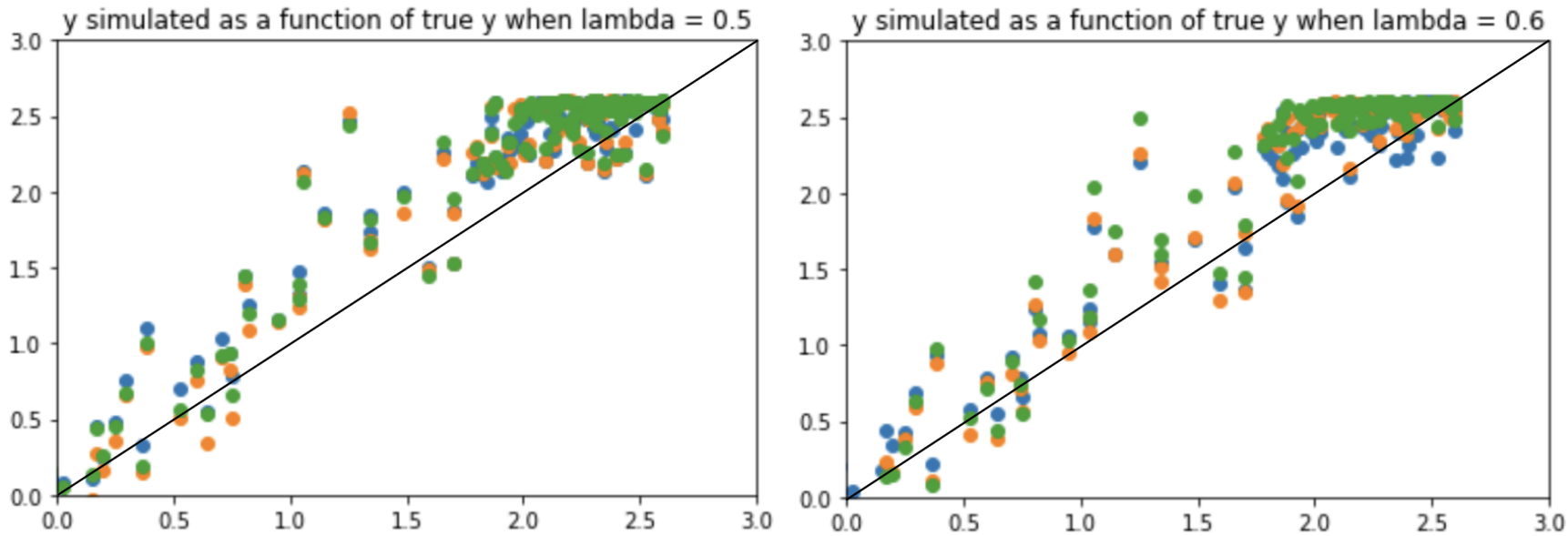
Tests

Parameters	Initial start	Bias	RMSE
ω_1^2	random	0.17	1.13
	first guess	**0.16**	**0.47**
ω_2^2	random	**0.41**	1.25
	first guess	0.67	**0.71**
ω_3^2	random	1.62	1.84
	first guess	**1.52**	**1.71**
ω_4^2	random	**0.49**	0.86
	first guess	0.55	**0.63**
σ^2	random	0.46	0.46
	first guess	**0.46**	**0.46**

Table – Relative bias and relative root mean square error (RMSE) of the estimated parameters evaluated from 15 simulated trials

Impact of the parameter λ

For the M-H procedure, when we choose the multidimensional random walk $\mathcal{N}\left(\phi^{(m-1)}, \boldsymbol{\lambda}\hat{\Omega}_{m-1}\right)$ as the proposal distribution, λ must be chosen.



(a) $\lambda = 0.5$ (b) $\lambda = 0.6$

Figure – Censored values obtained as a function of the true values. Experiment carried out on three different sets

Impact of the parameter λ

However, we quickly obtained much worse results for different λ values. We have listed these results in a table in which we have summed up the distance of each point from the line $y = x$.

λ	Sum of the distances to the line $y = x$
0.2	334
0.4	223
0.5	102
0.6	**90**
0.7	172
0.8	302

Table – Sum of the distances to the line $y = x$ depending on λ

SAEM Variant : results

The idea to impute y^{cens} before the algorithm made us think of a variant of SAEM.

Old S-step for y^{cens}	**New S-Step for y^{cens}**
$y^{\text{cens}} \sim p\left(\cdot \mid y^{\text{obs}}, \boldsymbol{\phi}^{(m)}, \hat{\boldsymbol{\theta}}_{m-1}\right)$	$y^{\text{cens}} \leftarrow f\left(t_j, \boldsymbol{\phi}_i^{(m)}\right)$

We run 20 times the whole simulation and recall the mean errors on θ.

Parameters	$\ln(P_1)$	$\ln(P_2)$	$\ln(\lambda_1)$	$\ln(\lambda_2)$	ω_1^2	ω_2^2	ω_3^2	ω_4^2	σ^2
OLD S-step %	3%	**1%**	**3%**	2%	**45%**	**50%**	13%	70%	**45%**
NEW S-step	3%	3%	7%	2%	54%	69%	13%	**67%**	46%

Table – Errors on θ

It seems like we loose precision but on the other hand, on average we took 26.9 seconds and 35.5 seconds for the original SAEM. For big instances with a lot of patients, it may be an idea.

SAEM Variant : impact on y^{cens}

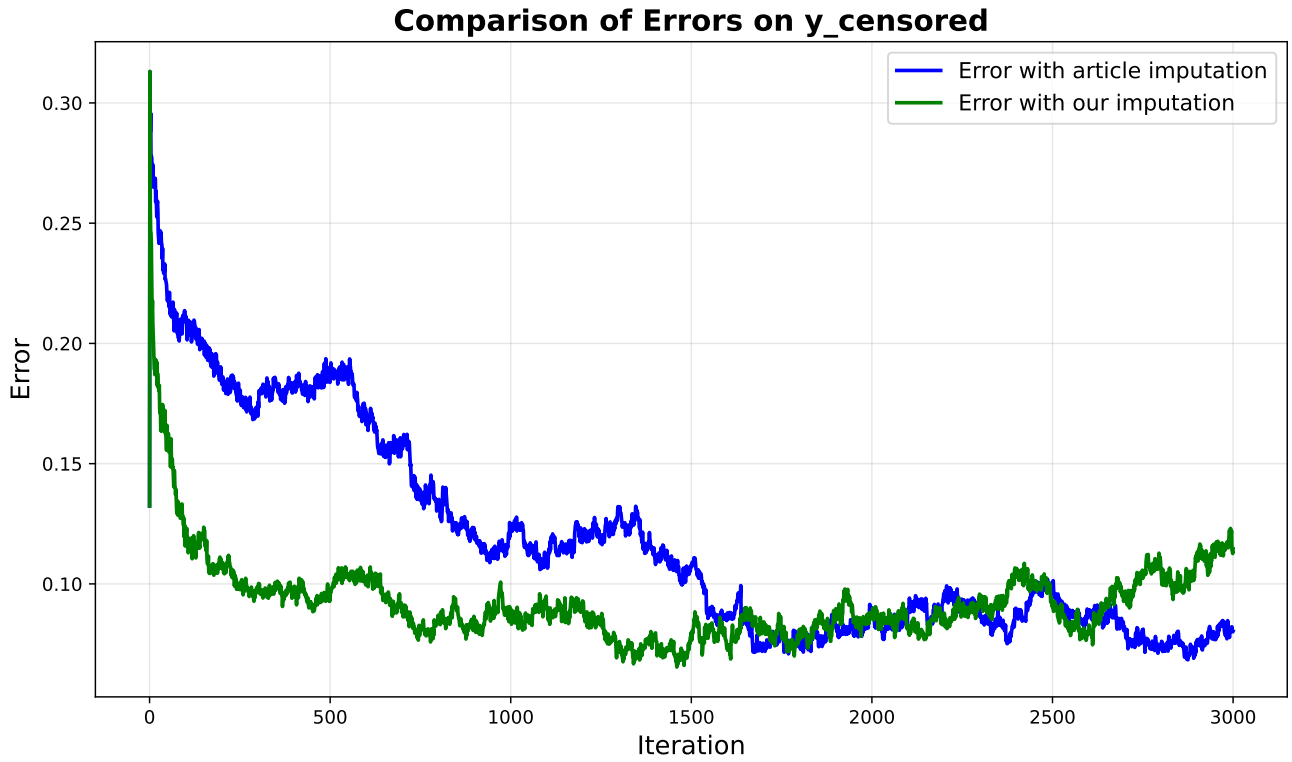


Figure – Losing precision on $y^{censored}$

On the long run, we lose precision with our imputation but it feels like using this imputation at first and maybe swapping after a certain time may be an idea.

Conclusion

Let us recap what we have done in this project :

-Implementation of SAEM

-Replication of the experiments in the article

-New experiments and ideas