Signal identification by local functional ANOVA

remi.mahmoud@institut-agro.fr

JDS 2025

Joint work with D. Causeur

Rémi Mahmoud

2025-06-17

Introduction

What today's talk is about

1. Context

2. Functional ANOVA (fANOVA)

3. Local fANOVA

4. Simulation study

5. Application to real-world data

6. Discussion and perspectives

Functional ANOVA

Functional linear model

- A key question in statistic…
- Does X has an effect on Y / Is X linked to Y ?
- Swiss Army knife of statistic ⇒ Linear Model

The functional linear model:

- $(Y(t))_{t\in\mathcal{T}}$ functional response variable defined on a time domain $\mathcal{T}$
- $x = (x_1, \ldots, x_p)'$ stand for a p-vector of time-independent explanatory variables.

$t \in \mathcal{T}$, the following model is assumed:

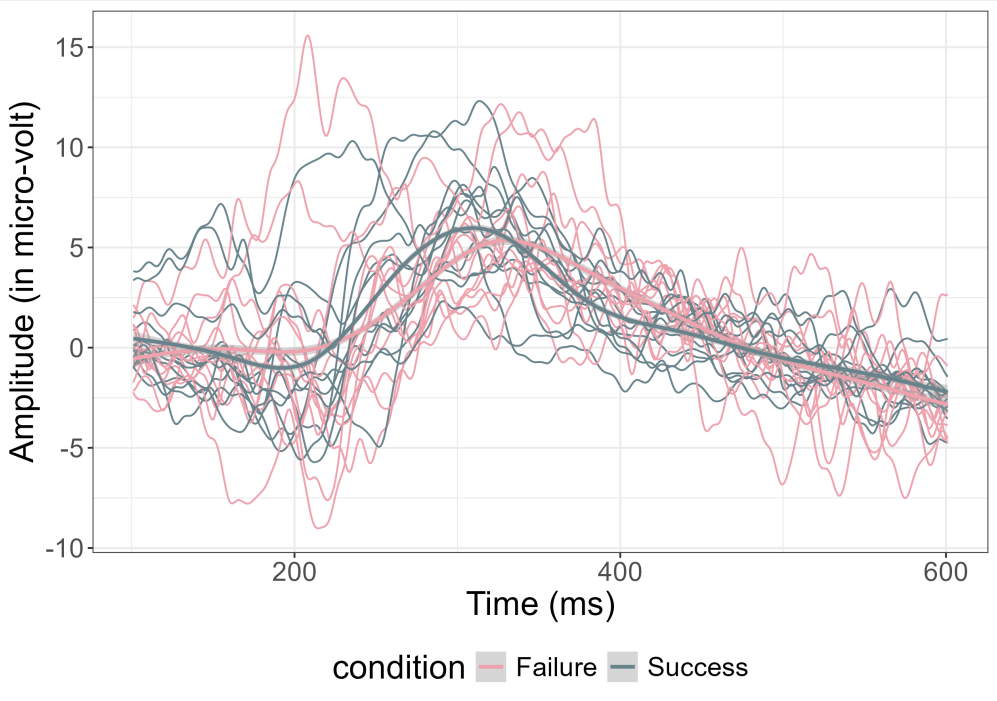
$$Y(t) = \quad \beta_0(t) + \beta_1(t)x_1 + \ldots + \beta_p(t)x_p + \varepsilon(t),$$

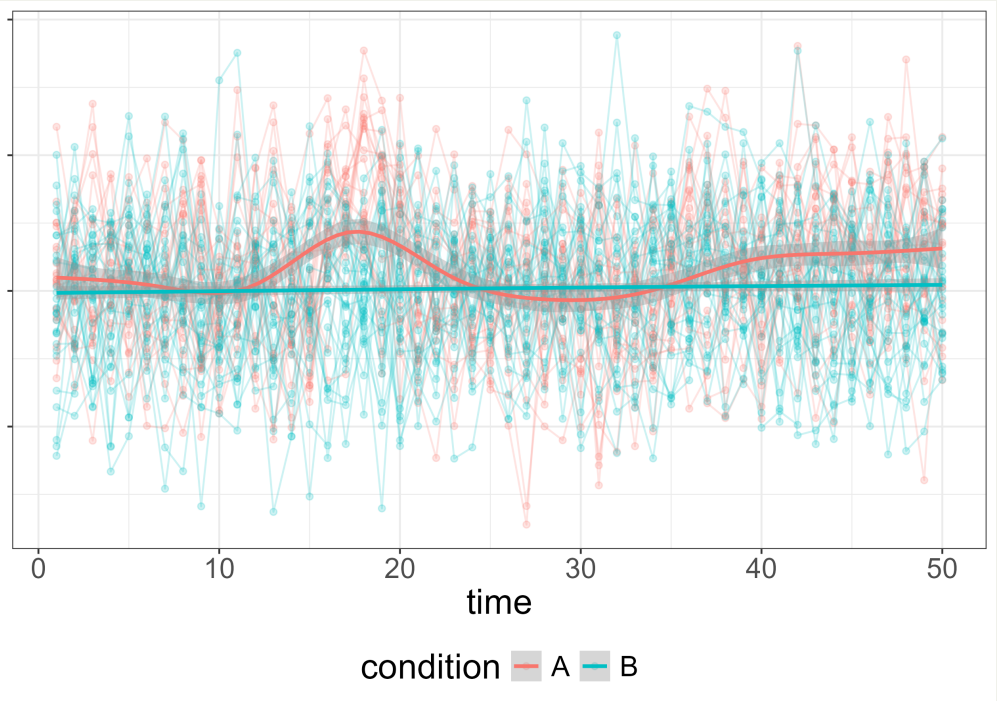
where:

- $\beta_0(t)$ pointwise intercept parameter at time t
- $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))'$ p-vector of pointwise regression parameters at time t.
- $\varepsilon(t) \sim \mathcal{N}(0, \sigma^2(t))$ & $\rho(t, t') = cor(\varepsilon(t), \varepsilon(t'))$ (potential time dependance)

Framework for today's presentation

- One-way design with a two-group covariate
- Could be applied to more complicated frameworks





What we want to test

- Is the mean function the same for each group ?
- Pointwise null hypothesis: $H_{0t} : \beta(t) = 0,\ \forall t \in \mathcal{T}$
- Global null hypothesis: $H_0 = \{H_{0t},\ t \in \mathcal{T}\}$

A decomposition to take time dependance into account

A proposal made by Sheu et al. 2016 [1] and used in Causeur et al. 2020 [2]

- *Idea*: Decorrelate pointwise test statistics and sum them across a given number of factor
- How: decompose the time-correlation matrix R into a q-factor model $\Sigma = \Lambda\Lambda' + \Psi$
  - Ψ diagonal $T \times T$ matrix
  - Λ is a $T \times q$ matrix containg factor loadings (among of "shared variance")
- Greater power

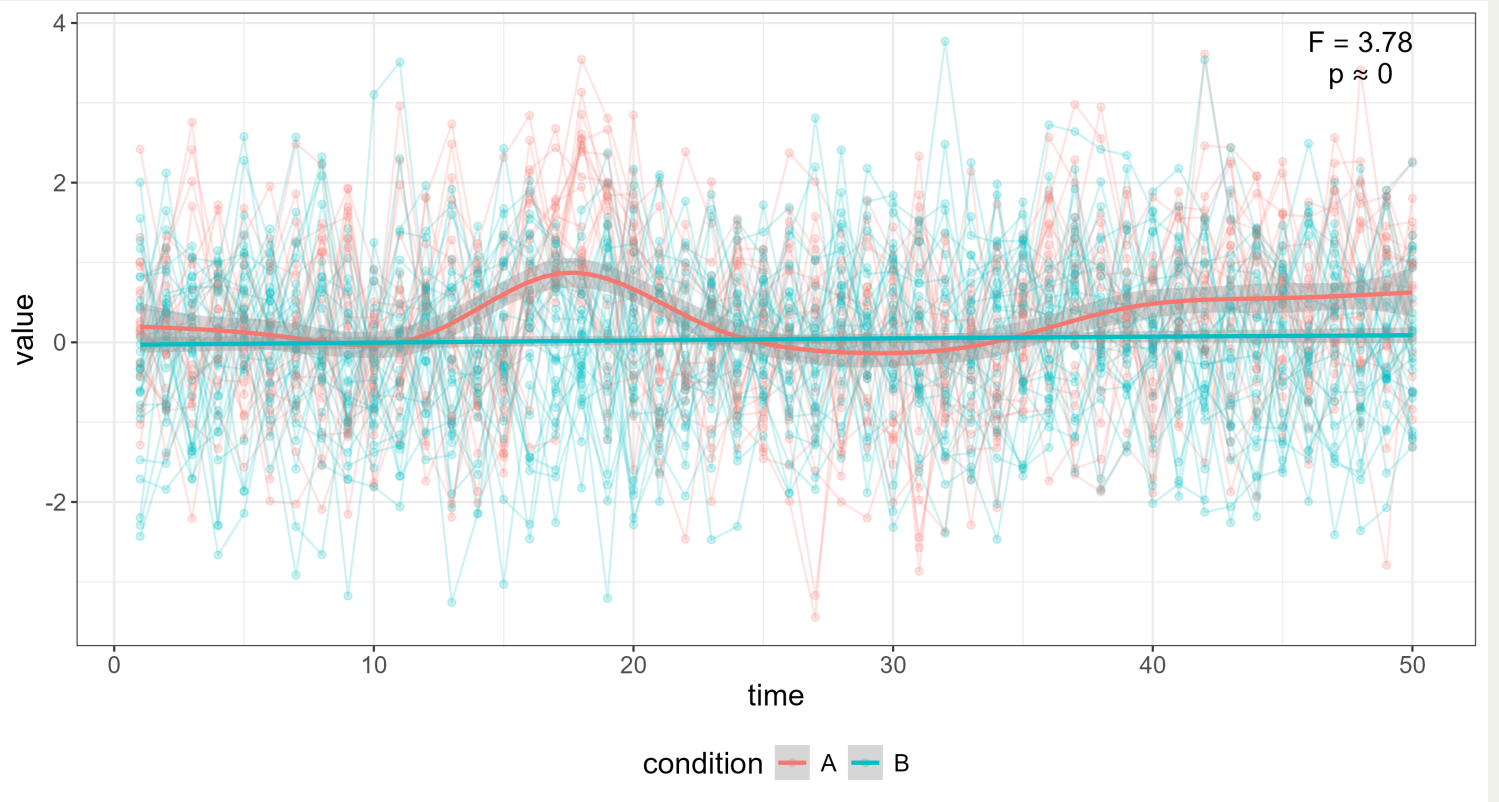
1. Pini, A., & Vantini, S. (2017). Interval-wise testing for functional data. Journal of Nonparametric Statistics, 29(2), 407–424. https://doi.org/10.1080/10485252.2017.1306627

2. David Causeur, Ching-Fan Sheu, Emeline Perthame, Flavia Rufini, A Functional Generalized F-Test for Signal Detection with Applications to Event-Related Potentials Significance Analysis,

How to gain accuracy ?

An example of significantly different curves:

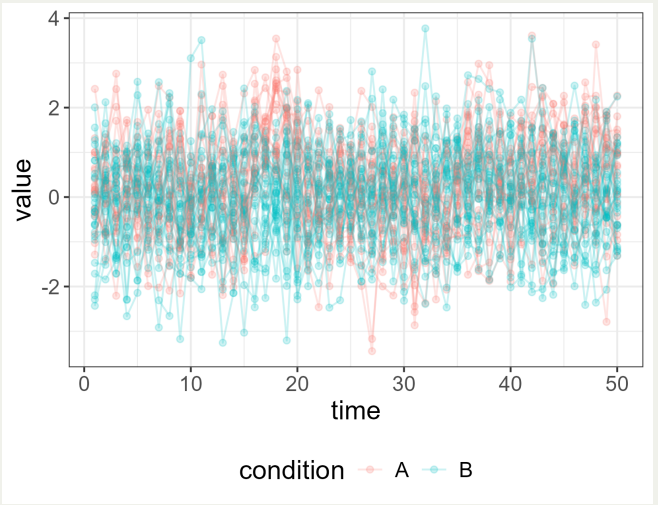
```
Error in file(filename, "r", encoding = encoding): impossible d'ouvrir la connexion
```

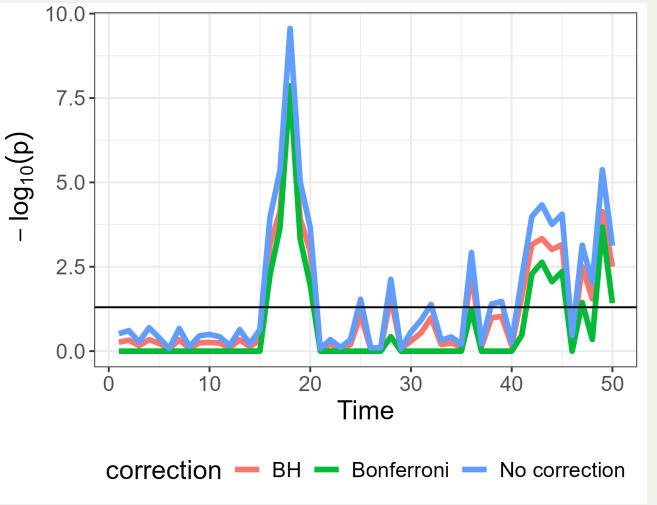


- We conclude in a difference between mean curves for each conditions…
- But another question arises: where does this difference come from ?
- Need to find a more accurate approach to conclude

Another point of view: multiple comparison procedure

1. Compute pointwise F-Statistics (F_t)
2. Compute p-values and apply a correction (Bonferonni / Benjamini Hochberg for instance)





- Some problems with these approaches:
    1. May be too conservative
    2. May be tricked by spurious time points with higher differences
    3. Functional nature of the signal (ex. time dependance) not taken into account

Local FANOVA

A compromise between multiple testing procedures and fANOVA

- *Idea*: for a given set of curves, find the largest intervals at which no significant effect can be detected with the functional ANOVA test

In a nutshell

- Screen the whole time frame partitionned into a given number of intervals, and incrementally find the largest union of intervals not significant.

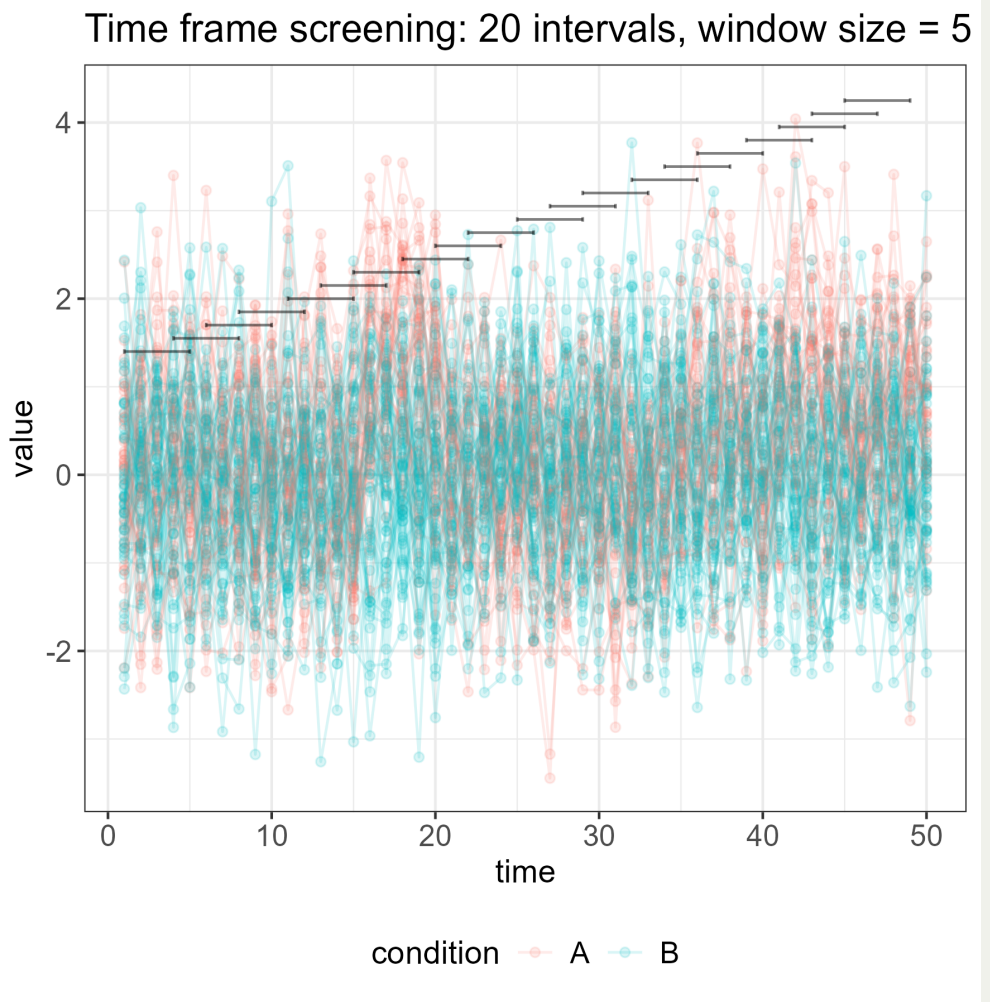
Inputs:

- k = window size
- m = number of intervals to consider

Local FANOVA: How it works

1. For each interval i in 1 to m:

- Define a time frame centered on interval i with a window size k
- Perform a fANOVA on these time frames

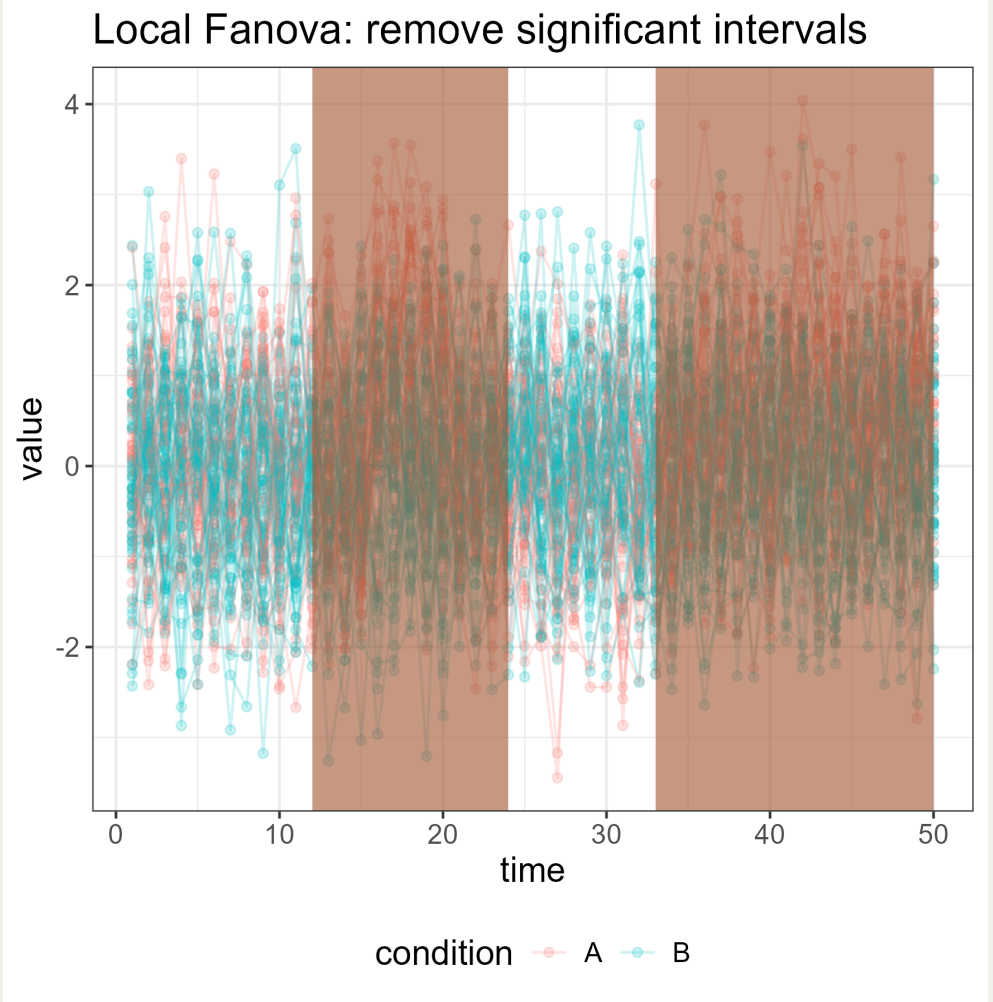


Local FANOVA: How it works

1. For each interval i in 1 to m:

- Define a time frame centered on interval i with a window size k
- Perform a fANOVA on these time frames

2. Remove all intervals from the time frame that are significant.



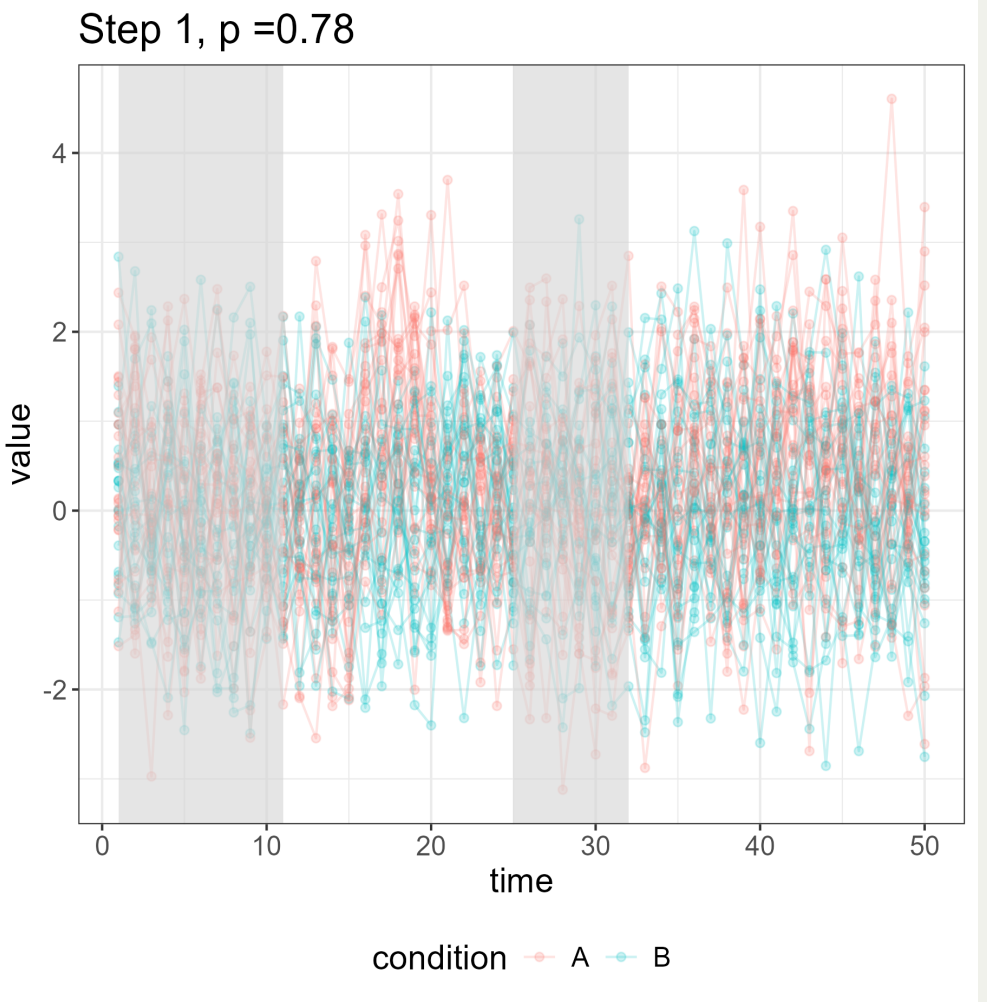
Local FANOVA: How it works

1. For each interval i in 1 to m:

- Define a time frame centered on interval i with a window size k
- Perform a fANOVA on these time frames

2. Remove all intervals from the time frame that are significant.
3. While the Fanova on the remaining whole time frame is not significant:

- Incrementally add intervals to the time frame (decreasing p-values)
- Perform a fANOVA on this time frame



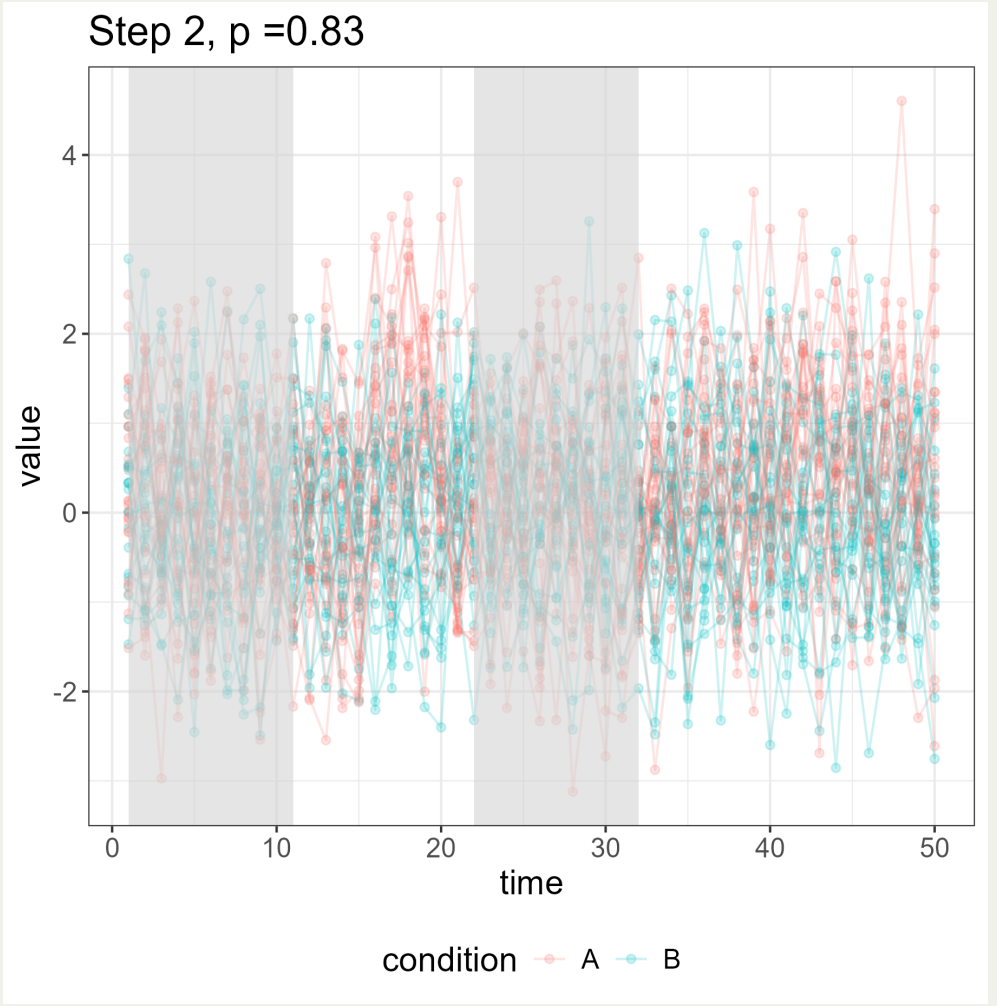
Local FANOVA: How it works

1. For each interval i in 1 to m:

- Define a time frame centered on interval i with a window size k
- Perform a fANOVA on these time frames

2. Remove all intervals from the time frame that are significant.
3. While the Fanova on the remaining whole time frame is not significant:

- Incrementally add intervals to the time frame (decreasing p-values)
- Perform a fANOVA on this time frame



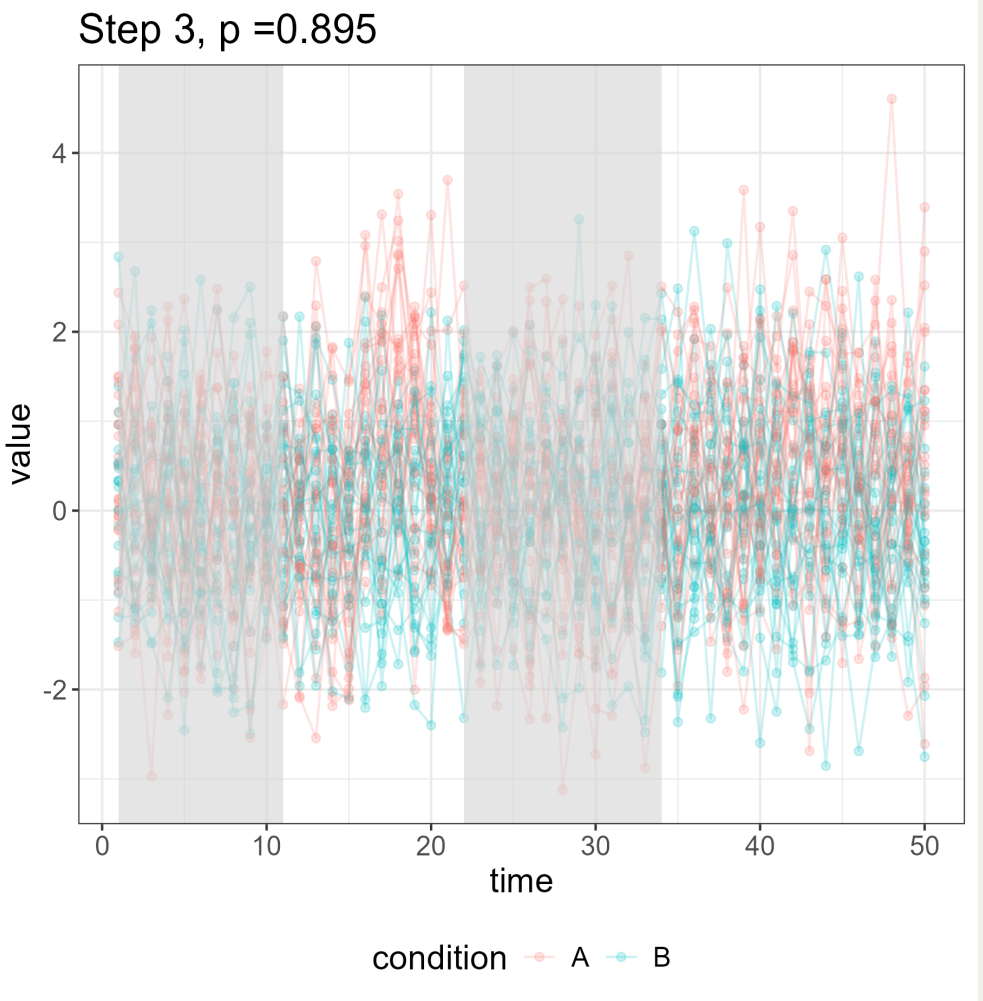
Local FANOVA: How it works

1. For each interval i in 1 to m:

- Define a time frame centered on interval i with a window size k
- Perform a fANOVA on these time frames

2. Remove all intervals from the time frame that are significant.
3. While the Fanova on the remaining whole time frame is not significant:

- Incrementally add intervals to the time frame (decreasing p-values)
- Perform a fANOVA on this time frame



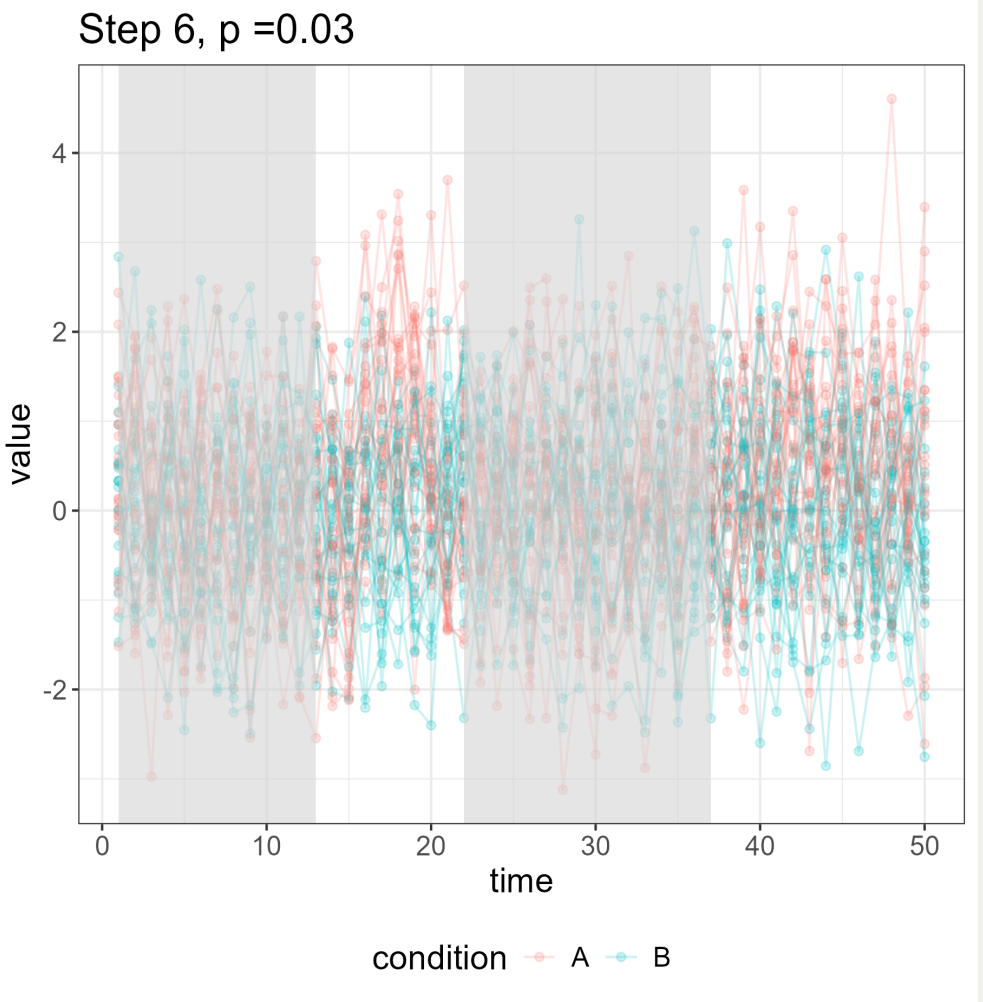
Local FANOVA: How it works

1. For each interval i in 1 to m:

- Define a time frame centered on interval i with a window size k
- Perform a fANOVA on these time frames

2. Remove all intervals from the time frame that are significant.
3. While the Fanova on the remaining whole time frame is not significant:

- Incrementally add intervals to the time frame (decreasing p-values)
- Perform a fANOVA on this time frame



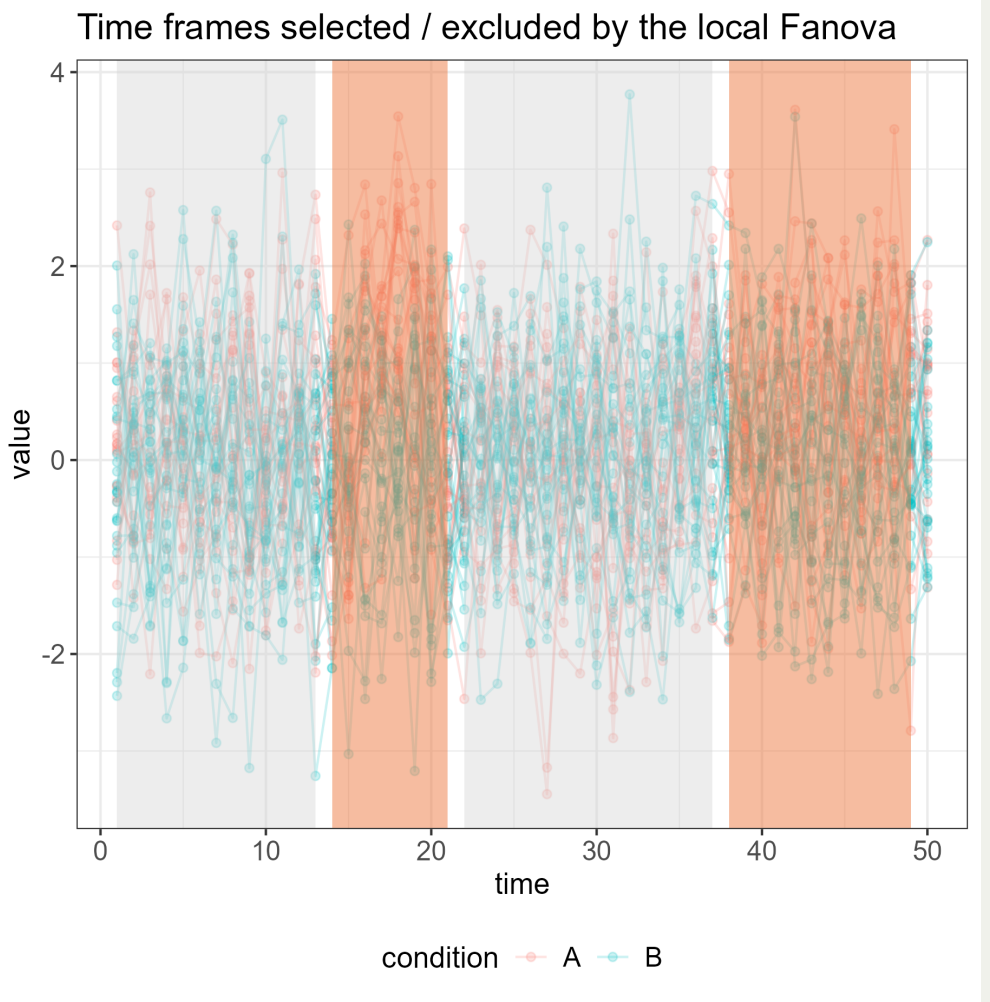
Local FANOVA: How it works

1. For each interval i in 1 to m:

- Define a time frame centered on interval i with a window size k
- Perform a fANOVA on these time frames

2. Remove all intervals from the time frame that are significant.
3. While the Fanova on the remaining whole time frame is not significant:

- Incrementally add intervals to the time frame (decreasing p-values)
- Perform a fANOVA on this time frame



Simulations

Datasets generated

Underlying observations generated by the model:

$$Y_{ij} = z_i^T \beta_j + \varepsilon_{ij}$$

with:

- Y_{ij} observation of curve i at time t_j, with $i = 1, \dots, n$ and $j = 1, \dots, T$
- $\beta_j = \beta_i(t_j)$
- $\varepsilon_{ij} = \varepsilon_i(t_j)$
- $\mathcal{T} = [0, 1]$

Parameters:

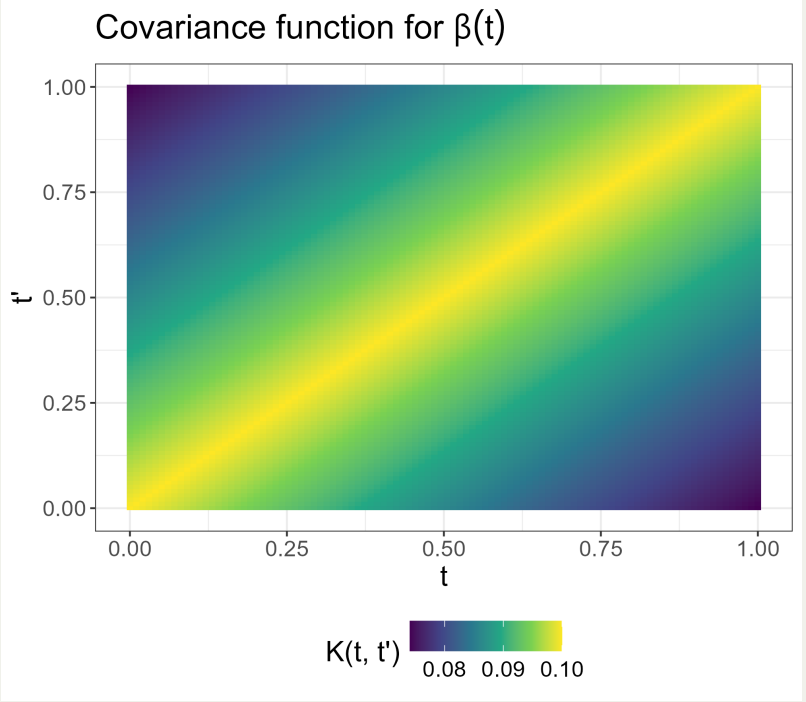
- Temporal resolution $T \in \{50, 500, 2000\}$
- Number of individuals $n \in \{100, 1000\}$
- Temporal dependancy of the residuals $\varepsilon(t) \sim AR(1),\ \rho \in \{0, 0.3, 0.8\}$
- 100 datasets for each parameter combination

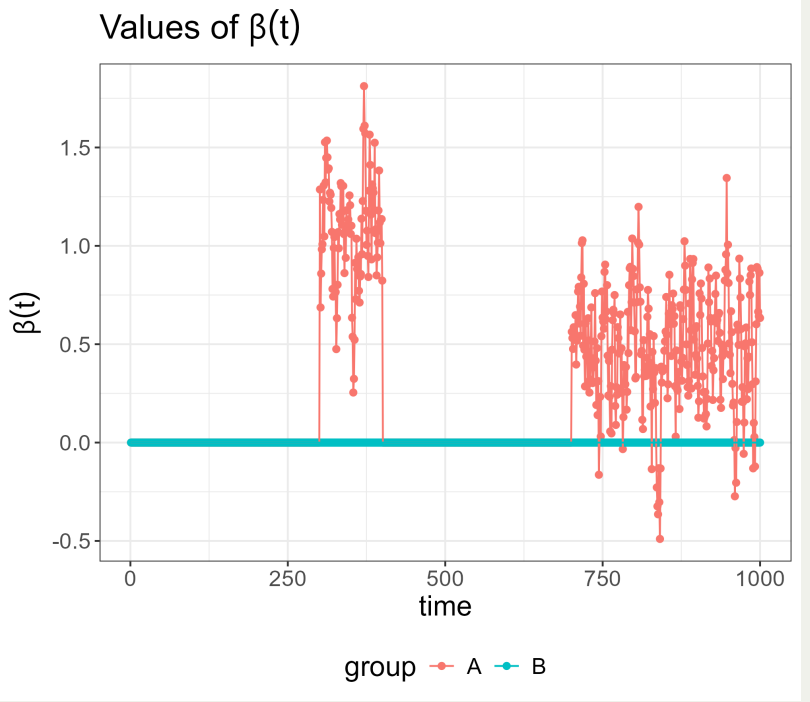
Values of $\beta(t)$

$$\beta(t) = \begin{cases} \mathcal{GP}(m(t), K(t,t')) & \text{if } t \in]0.3; 0.4] \ \cup \ ]0.7; 1] \\ 0 & \text{else} \end{cases}$$

with covariance function $K(t,t') = 0.1 * e^{-0.3|t-t'|}$ and

$$m(t) = 1_{t \in]0.3, 0.4]} + 0.5 \times 1_{t \in]0.7, 1]}$$

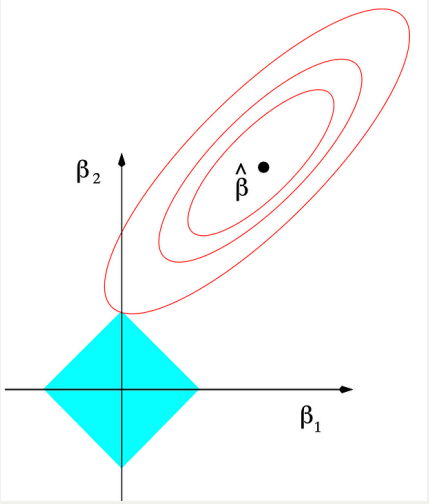




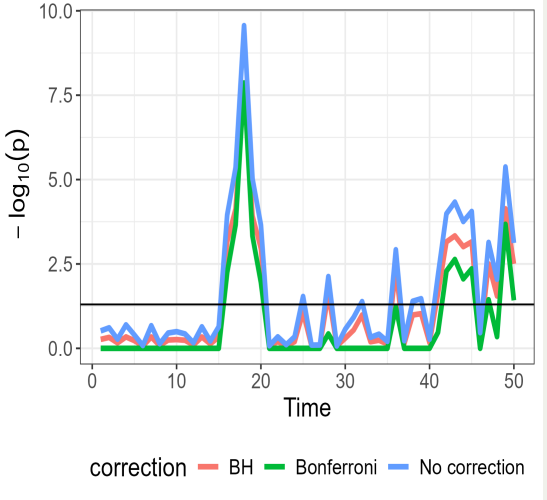
Alternatives methods compared

Lasso regression (Tibshirani):

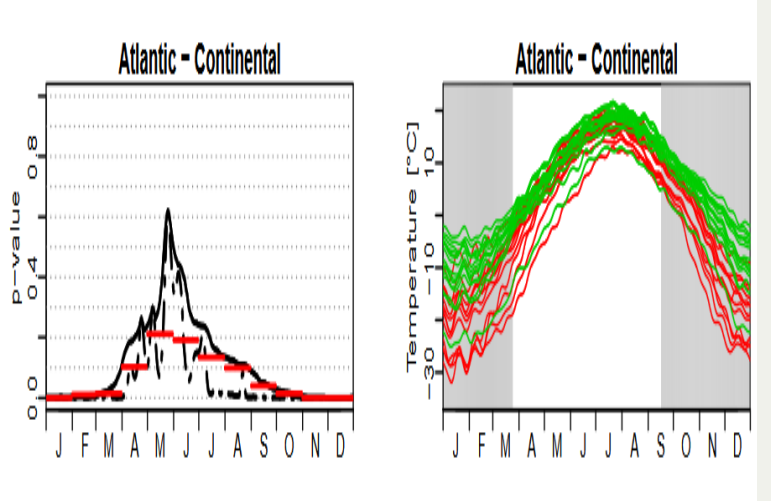
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\sum_1^n (y_i - \sum_1^p \beta_j x_{ij})^2 + \lambda \sum_1^p |\beta_k|)$$



Multiple testing procedure (BH correction)

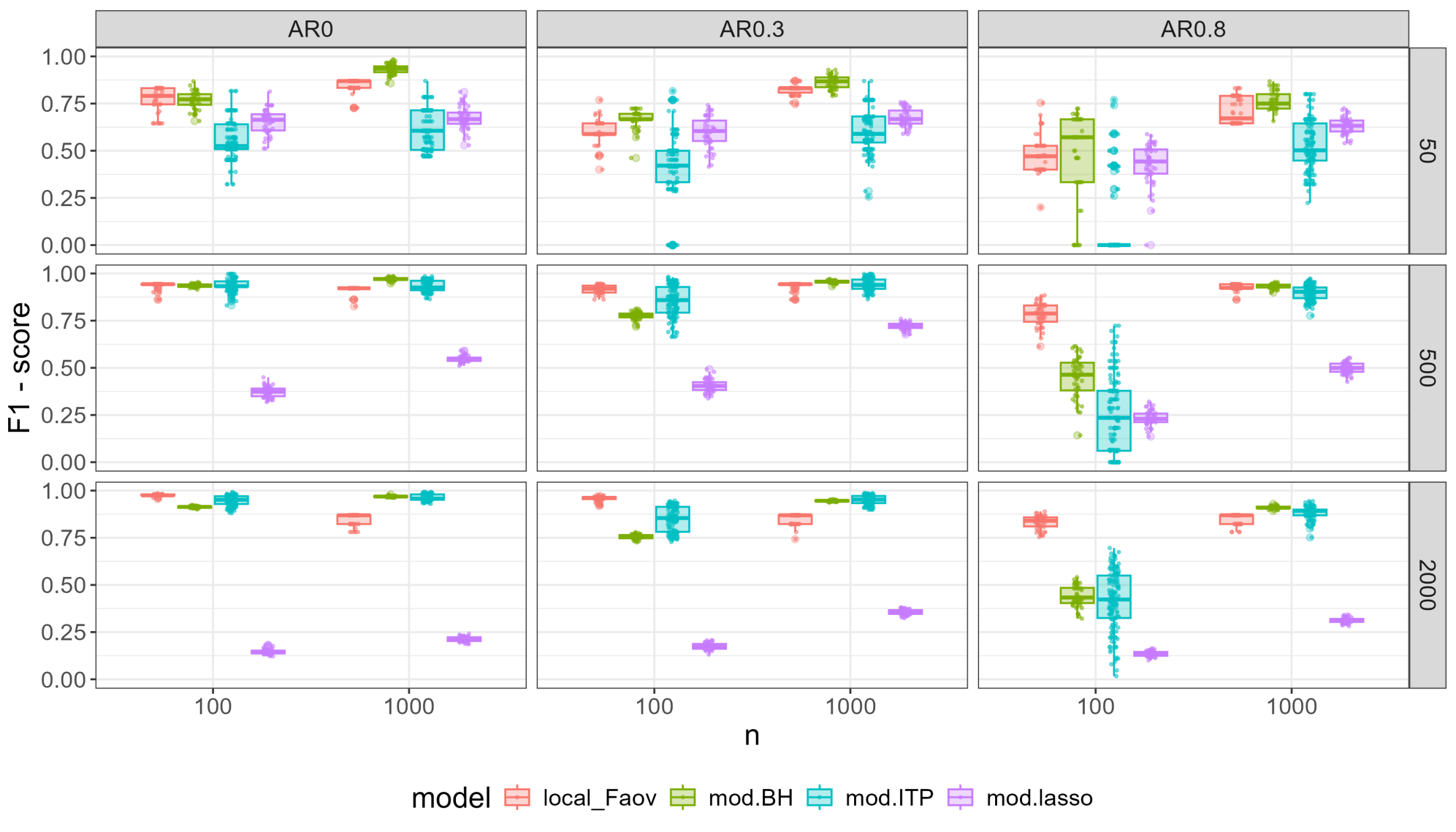


Interval wise testing procedure (Pini et al. 2018[1])



Simulation results

The moment that every statistician waits for / fears



Real world data

Context

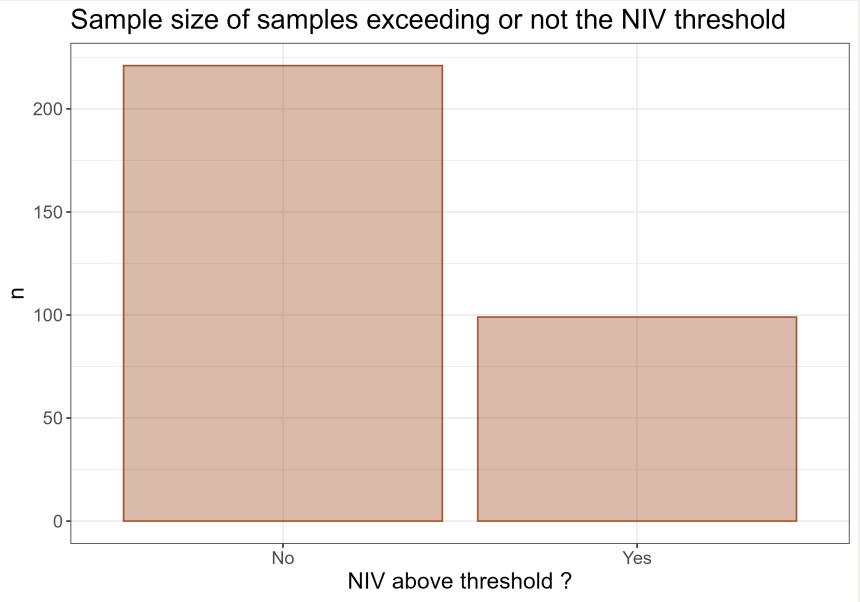
- Fusarium head blight: Fungal disease $\Rightarrow$ lot of damages (yield / food safety / added value)
- Mycotoxins emitted by Fusarium (ex. Nivalénol (NIV))

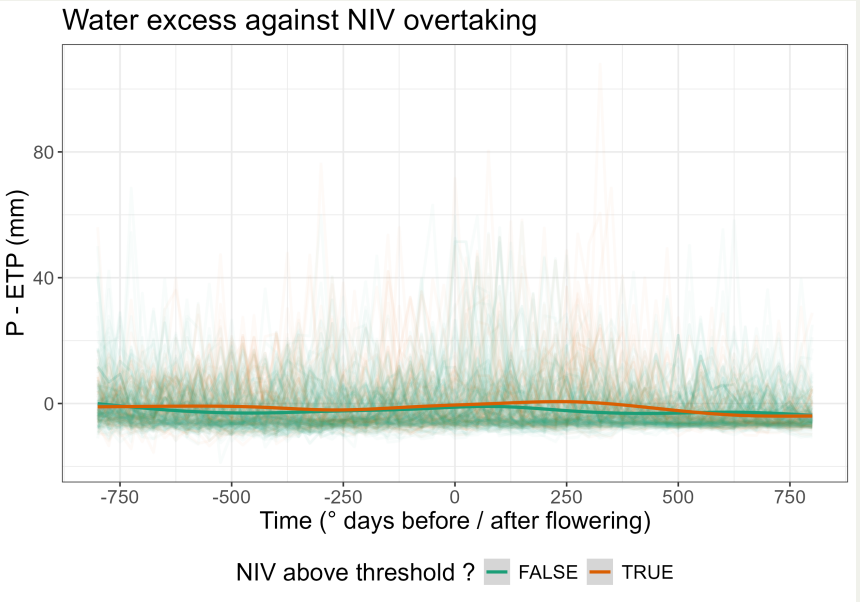
- How is the overtaking of a given toxin related to a given climate variable ?

Dataset (Arvalis)

- $\approx$ 1500 farms: climate timeseries and toxins concentrations
- Different agronomic practices
- **Binarization of the concentrations**: above/below of a legal threshold

Example with 320 farms sharing the same agronomic practices

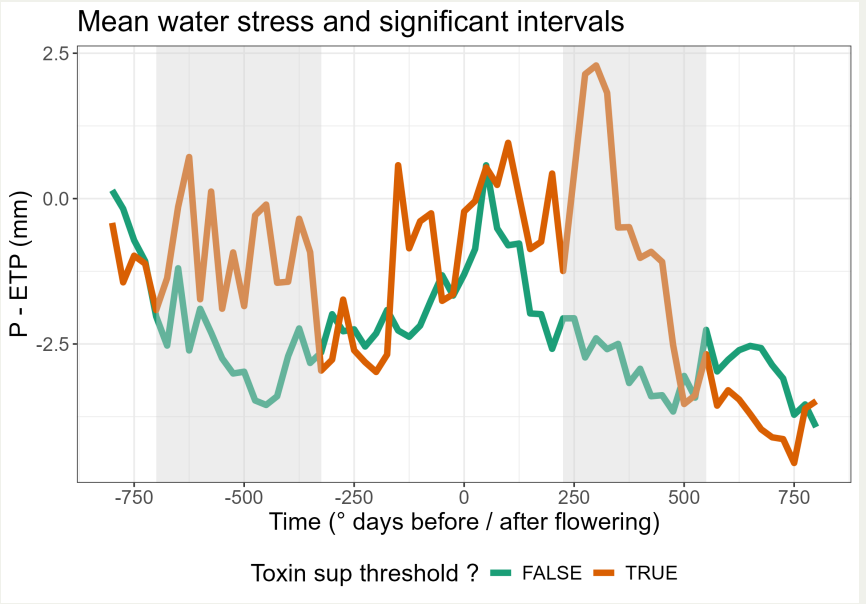




Application

Application of local fANOVA in our case:

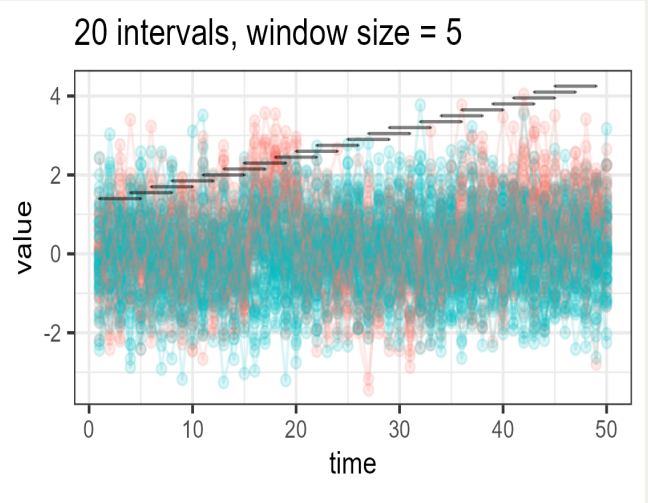
Y is a time serie of a climate variable (ex. water excess) and $x_1 = 1_{\text{NIV} > \text{legal threshold}}$

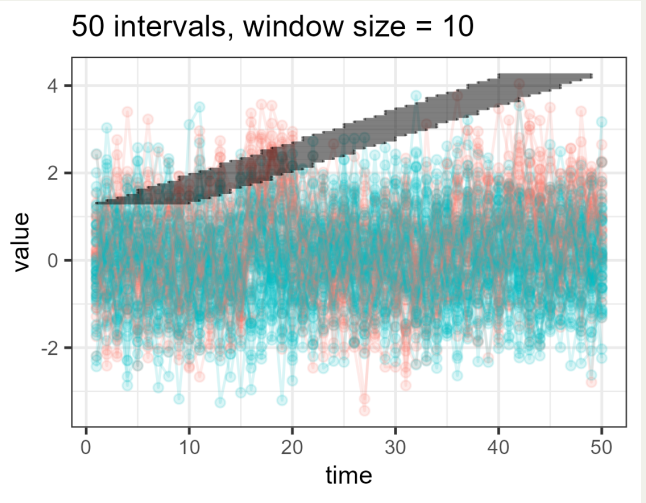


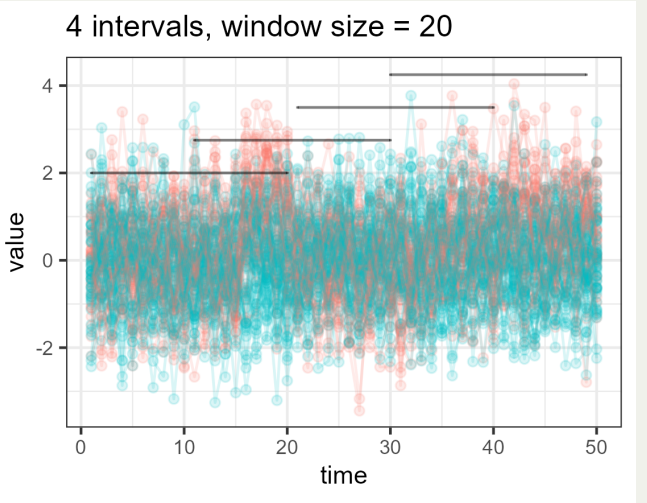
Discussion and perspectives

On-going thoughts on the choice of hyperparameters

- In our case, two key parameters:
  - k = window size
  - m = number of intervals to consider







- Choice of hyperparameters, different points of view:
  - statistical: find the best combination of hyperparameters that optimize a given metric
  - practical: find the best compromise between “good results” and acceptable computational costs
  - field-related: when possible, set hyperparameters that have a mean

Perspectives and conclusion

Development version available on Github/RemiMahmoud/localFaov

```
devtools::install_github("RemiMahmoud/localFaov")
```

This is an ongoing work

- Promising but contrasting simulation results
- Convincing application on some real world cases
- Some issues still need to be fixed (choice of hyperparameters, situations of bad results etc.).

Functional data

- Observations represented by curves or functions
- Old but new: original studies by Grenander & Rao (1948 / 1952 resp) but huge works from Ramsay and Silverman 2005[1]
- Similarly to NN, has arosen in the last years because of higher data availability and computing abilities
- Challenges:
  - temporal dependency
  - specific questions (adaptation of classical methods to FD etc.)

Metrics: how good / bad are the results

Recall of the goal (in our case):

- Detect curve segments responsible for a significant difference between modalities
- Need for a metric designed for this goal

- Let $(I_i)_{i=1,\cdots,m}$ the collection of intervals linked to this difference
- (*Reminder*) In our simulation study, $m = 2,\ I_1 =]0.3, 0.4]$ and $I_2 =]0.7, 1]$

Metrics used

- Mean Overlap: $\text{Mean Overlap} = \frac{1}{m} \sum_{i=1}^{m} \frac{\#\text{ points selected in } I_i}{|I_i|}$
- $\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
- $F1 = \frac{2 \times \text{Mean Overlap} \times \text{Sensitivity}}{\text{Mean Overlap} + \text{Sensitivity}}$
- Other metrics exist (mean distance to closest interval, FDR etc.)

What we want to test

- Is the mean function the same for each group ?

- Pointwise null hypothesis: $H_{0t} : \beta(t) = 0,\ \forall t \in \mathcal{T}$
- Global null hypothesis: $H_0 = \{H_{0t},\ t \in \mathcal{T}\}$

A common approach:

1. Compute pointwise F-Statistics (F_t, Ramsay and Silverman 2005)
2. Aggregate the F-Statistics (F_t)
    - $\int$ (Zhang and Liang 2014[1])
    - $F_{\max} = \sup_{t \in \mathcal{T}} \mathcal{F}(t)$ (Zhang et al. 2019[2])
3. Distribution of F under H_0 known or derived by permutation

But generally time dependance not taken into account (Shen et al. 2016) $\Rightarrow$ Increase risk of type-I error !

1. Zhang, J. T., & Liang, X. (2014). One-way ANOVA for functional data via globalizing the pointwise F-test. Scandinavian Journal of Statistics, 41(1), 51-71.
2. Zhang, J. T., Cheng, M. Y., Wu, H. T., & Zhou, B. (2019). A new test for functional one-way ANOVA with applications to ischemic heart screening. Computational Statistics & Data Analysis,