Clustering with variable selection for longitudinal data.

Marie Denis and Mahlet G. Tadesse

ENAR, March 29, 2022

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Outline

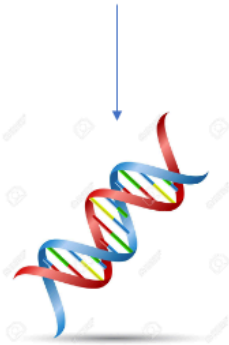
This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Biological context

In genetic, the main objective is to understand the molecular mechanisms underlying important biological processes

- to identify the best phenotypes in genetic improvement programs,
- to identify markers associated with phenotypes/diseases,
- to improve diagnostics/interventions,
- ...

Differences



$\Rightarrow$ Identification of **genetic variants** involved in the variability of phenotypic traits using **variable selection approaches**

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Biological context

In many fields, longitudinal studies are conducted to obtain a better understanding of the dynamic of biological processes

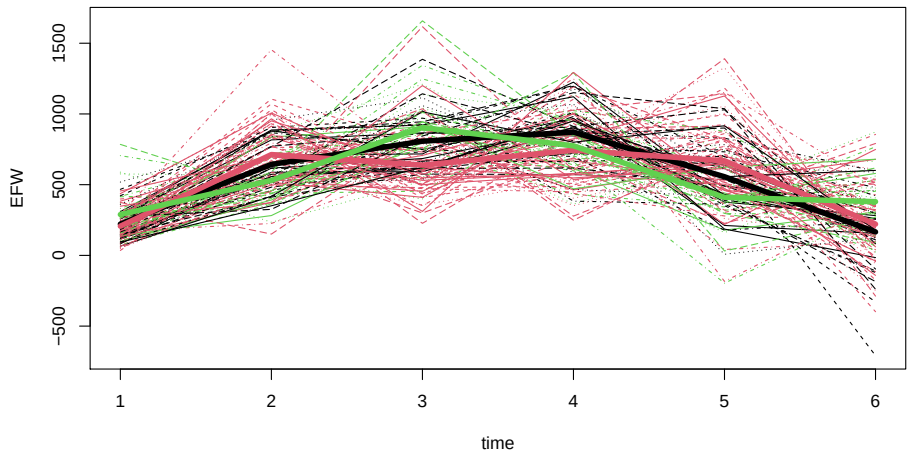


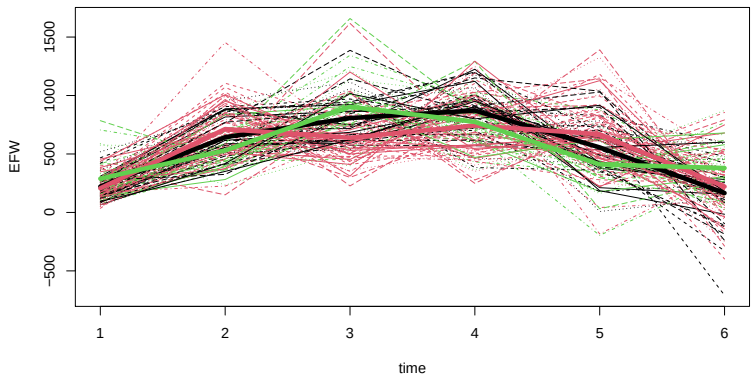
Figure 1: Fetal weights over pregnancy period.

⇒ Identification of **genetic variants** involved in the dynamic of responses and estimation of their effects over time using **variable selection for longitudinal data**

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Biological context

Limitations: selection of the **same** subset of genetic variants and estimation of the **same** dynamic effects for all individuals.



↬ Different subsets of genetic variants with different time varying effects may be associated to different groups of individuals

Identifying genetic markers associated with each group for

↬ a finer understanding of molecular mechanisms underlying dynamic processes,

↬ helping design improved medical and phytosanitary treatments

Need to develop a clustering approach with variable selection for longitudinal data

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Clustering approaches

To uncover groups of observations characterized by several variables

Two broad classes of approaches:

1. Based on similarity or dissimilarity distances (hierarchical methods, k-means, ...)
2. Model-based approaches which use mixture models for clustering

In the following we will focus on **model-based approaches** (Banfield and Raftery, 1993; Richardson and Green, 1997; Neal, 2000; Fraley and Raftery, 2002)

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Model-based approaches

Model-based methods without or with variable selection

↬ **Why integrate variable selection ?** Clustering information is contained into a **subset** of covariates (Friedman and Meulman, 2004; Tadesse et al., 2005; Maugis et al., 2009): including all covariates may hide group structures

⚠ Especially for high-dimensional data or data with a higher number of covariates than observations

	no outcome	non-longitudinal outcome	longitudinal outcome
Clustering without variable selection	✓	✓	✓
Clustering with variable selection	✓	✓	✗

Table 1: Existing approaches wrt the type of outcome

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Coming back to our context

Objective:

To **cluster** similar response profiles and to **select** genetic variants that help discriminate them

	no outcome	non-longitudinal outcome	**longitudinal outcome**
Clustering without variable selection	✓	✓	✓
Clustering with variable selection	✓	✓	✗

The proposed approach:

A stochastic partitioning method, based on the work of Monni and Tadesse (2009), which combines ideas of mixture models, mixed effects models, and variables selection.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Statistical model

Data

n independent samples with a repeated outcome and p covariates:

- $\mathcal{Y} = (Y_1, \ldots, Y_n)'$ with $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ for $i = 1, \ldots, n$,
- $\mathcal{X} = (X_1, \ldots, X_p)$ with $X_j = (X_{j1}, \ldots, X_{jn})'$ for $j = 1, \ldots, p$.

Partitioning approach

- Variables partitioned into sets of pairs (X_J, Y_I) with $J \subset \{1, \ldots, p\}$ and $I \subset \{1, \ldots, n\}$
- Individuals in a pair have similar dependence on the subset of covariates
- A configuration: a partition of data where the components are the pairs

For example, a configuration of length K is defined by:

$$\mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_K = (X_{J_1}, Y_{I_1}) \oplus \cdots \oplus (X_{J_K}, Y_{I_K}) = (|J_1|, |I_1|) \oplus \cdots \oplus (|J_K|, |I_K|)$$

with $0 \leq |J_k| \leq p$, $1 \leq |I_k| \leq n$ and $\sum_{k=1}^{K} |I_k| = n$ and $\sum_{k=1}^{K} |J_k| \leq Kp$.

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Hierarchical Bayesian model

↬ Mixture of multivariate Gaussian models such that the model associated with individuals $Y_{l_1}, \ldots, Y_{l_{n_k}}$ belonging to the component $\mathcal{S}_k = (|J_k|, |I_k|) = (m_k, n_k)$ is:

$$
\begin{aligned}
Y_i | \boldsymbol{\beta}_k, \sigma_k^2, \rho &\sim \mathcal{N}_T(\boldsymbol{\mu}_{ki}, \sigma_k^2 \Omega), \quad i = l_1, \ldots, l_{n_k}, \\
\boldsymbol{\beta}_{ks_r} | \tau_k^2, \sigma_k^2 &\sim \mathcal{N}_T(0, \sigma_k^2 \tau_k^2 (D'D)^{-1}), \quad r = 1, \ldots, m_k, \\
\tau_k^2 &\sim \mathcal{IG}(a, b), \ \sigma_k^2 \sim \mathcal{IG}(\sigma_0^2, \nu) \\
\rho &\sim \mathcal{U}_{(-1,1)}
\end{aligned}
$$

$$
p((m_1, n_1) \oplus \ldots \oplus ((m_K, n_K)) \propto \prod_{k=1}^{K} \pi^{m_k}
$$

- $\boldsymbol{\mu}_{ki} = \sum_{r=1}^{m_k} x_{is_r} \boldsymbol{\beta}_{ks_r}$ with $\boldsymbol{\beta}_{ks_r} = (\beta_{ks_r}^1, \ldots, \beta_{ks_r}^T)'$ the time varying effects.
- Ω a $T \times T$ auto-regressive correlation matrix of order 1 with unknown parameter ρ.
- τ_k^2, σ_k^2 variance parameters.
- $a, b, \sigma_0^2, \nu, \pi$ fixed hyperparameters.
- D matrix representation of first order finite difference operator.

↬ Parameters $\boldsymbol{\beta_k}$ and σ_k^2 are integrated out

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

MCMC implementation

1. Update of configuration via a reversible jump Markov chain Monte Carlo algorithm:
   1. **Type 1**: Add or delete covariate to/from a component
   2. **Type 2**: Reallocate observations by choosing to split/merge components (m, n) (with $n > 0$) or to reassign a single observation.
2. Update of τ_k^2 for $k = 1, \ldots, K$ via a Metropolis-Hasting algorithm,
3. Update of ρ via a Metropolis-Hasting algorithm.

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Simulation study

Objectives

- Evaluate sensitivity to hyperparameter π
- Evaluate performance under different simulation settings:
  - varying residual variance, σ_k^2
  - varying number of covariates, p
  - varying number of relevant predictors per cluster
- Compare to a two-step approach:
  1. Clustering step: model-based clustering for longitudinal data (McNicholas and Subedi, 2012) implemented in the R package `longclust`
  2. Selection step: Bayesian varying coefficient model with selection using group spike-and-slab prior (Heuclin et al., 2021) in each identifying cluster

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Simulation parameters

- Number of individuals $n = 75$, number of time steps $T = 10$, number of clusters/groups $K = 3$
- Number of covariates $p = 150$ or $1,000$ with a varying number of relevant predictors per cluster with different varying time effects for each cluster
- Residual variance equal to 0.1 or 1
- X elements randomly sampled from the set $\{0, 1, 2\}$ to mimic SNP data

Figure 2: Simulated dynamic effects.

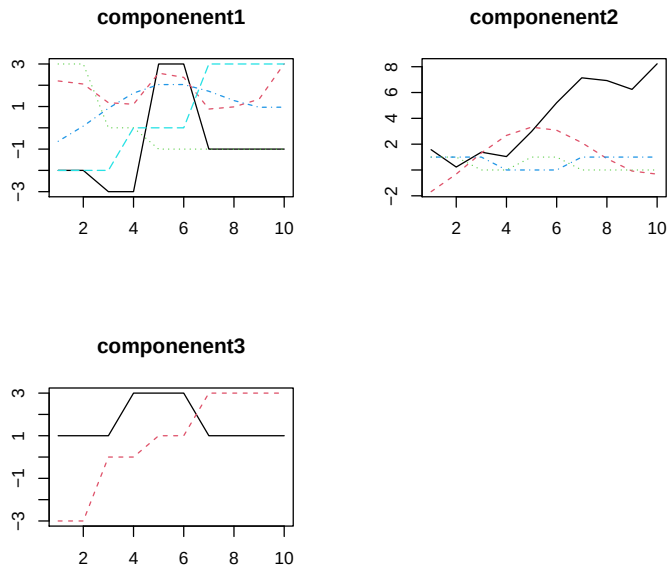
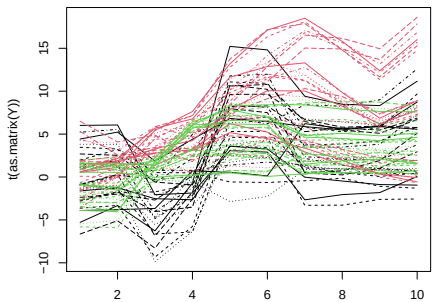


Figure 3: Simulated response profiles over time.



This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Criterion for evaluating model performance

- Convergence evaluation:
  - Acceptance rates over iterations for type 1 and type 2 moves
  - Ratio of acceptance rate of type 1 over type 2 moves (RAR)
- Prediction evaluation:
  - Confusion matrix of cluster allocation
- Selection evaluation:
  - False positives (FP), false negatives (FN), true positives (TP), true negatives (TN) for each cluster

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Sensitivity to hyperparameter π

(n=75, p=150, $\sigma_k^2 = 0.1$, 1 to 5 relevant covariates per cluster)

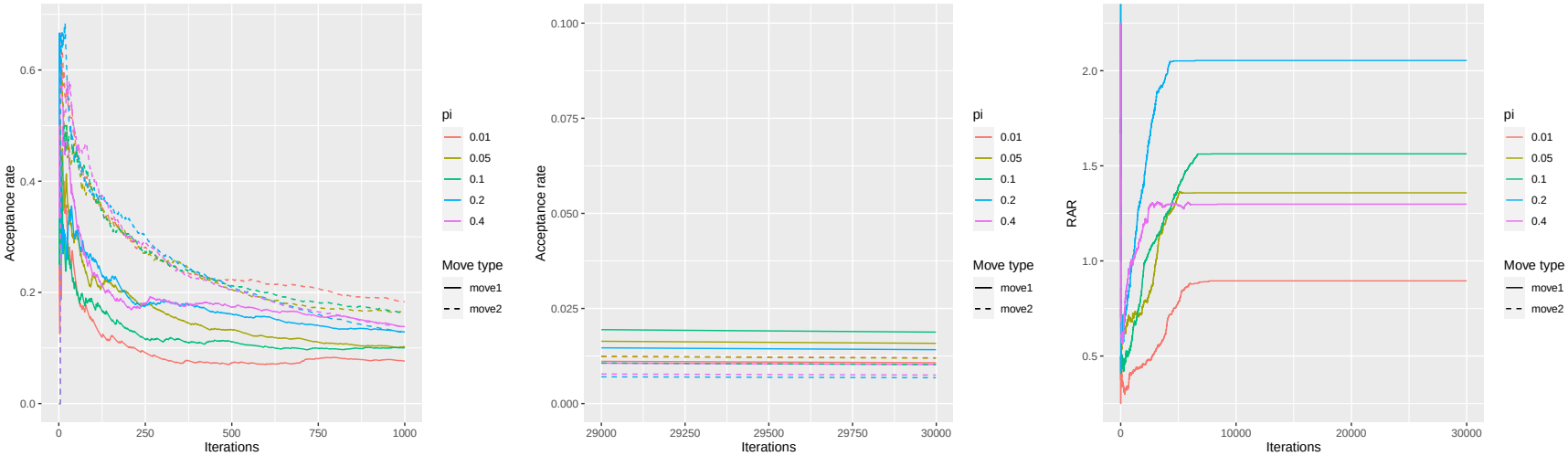


Figure 4: Acceptance rates over the first 1,000 iterations (on left), over the last 1,000 iterations (in middle), and RARs over 30,000 iterations (on right).

- Clustering: successful recovery of groups for all values of π
- Selection: successful identification of cluster-specific predictors for all values of π

↬ Weak sensitivity to hyperparameter π

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Impact of the residual variances (n=75, p=150, 1 to 5 significant covariates per cluster)

- Convergence diagnostic based on RAR:

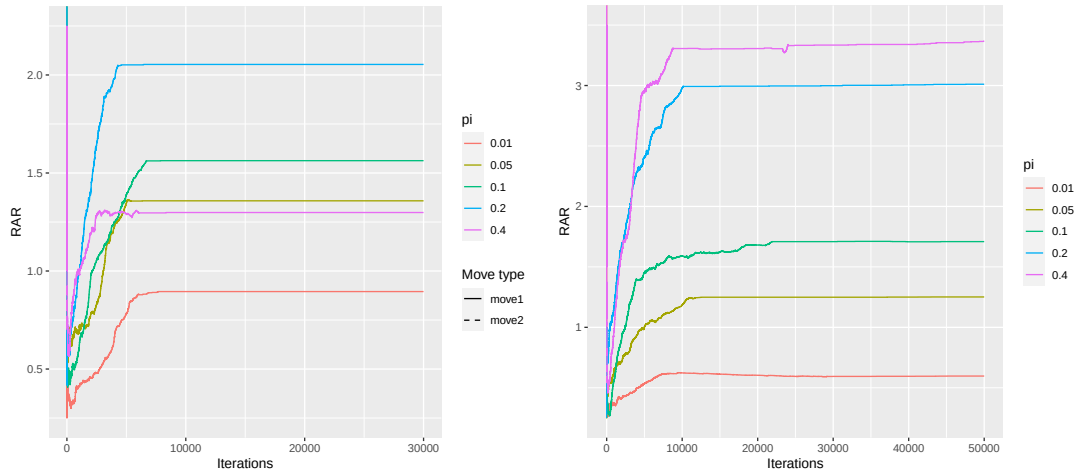


Figure 5: RARs over iterations for simulation with $\sigma_k^2 = 0.1$ (right) and $\sigma_k^2 = 1$ (left).

↬ Slower convergence for higher residual variances

- After convergence, successful inference for clustering and selection

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Impact of the number of relevant covariates per cluster

(n=75, p=150, $\sigma_k^2 = 1$)

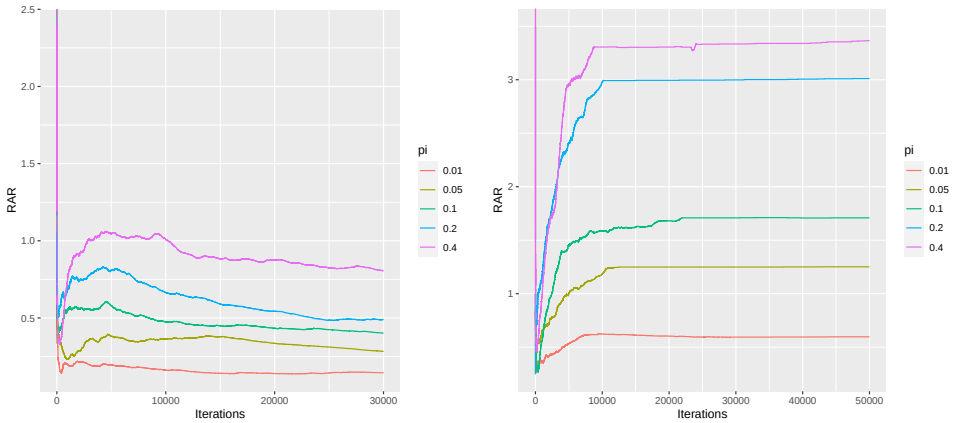


Figure 6: RARs over iterations for simulation with one relevant covariate per cluster (on right) and with 1 to 5 relevant covariates per cluster (on left).

↬ Higher number of significant covariates per cluster helps uncover groups and improves convergence

		Truth		
		1	2	3
Predicted	1	24	0	0
	2	1	25	3
	3	0	0	22

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Impact of total number of covariates (n=75, $\sigma_k^2 = 1$)

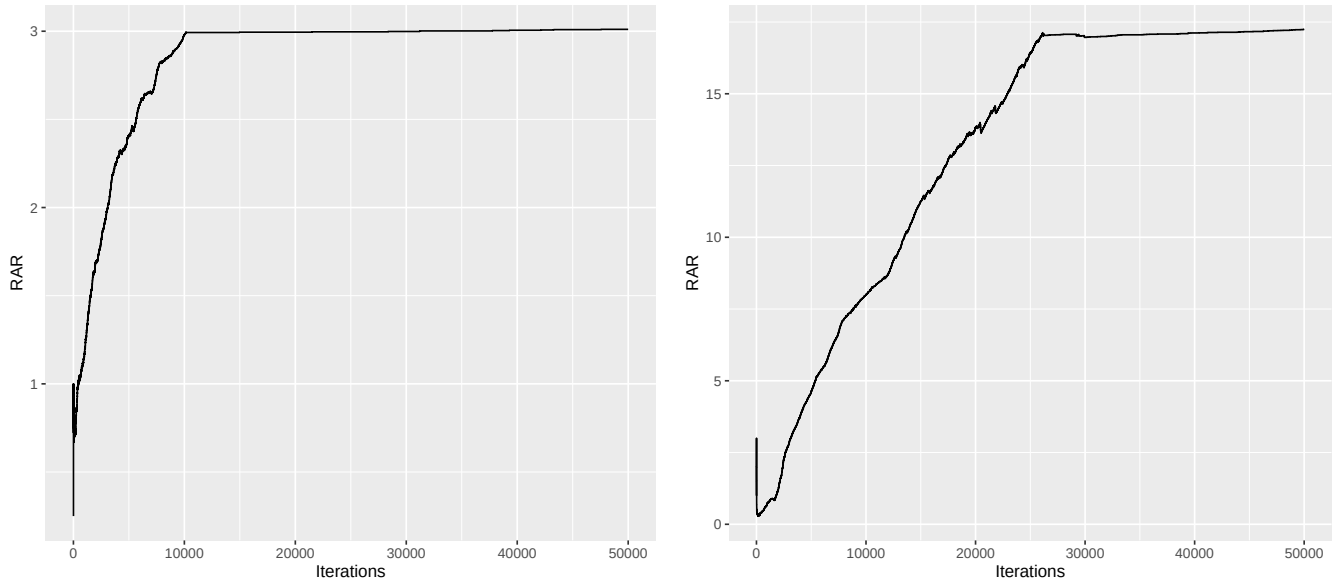


Figure 7: RARs over iterations for simulation with 150 covariates (on right) and with 1,000 (on left) using $\pi = 0.1$.

↬ Slower convergence for a higher number of covariates

- After convergence, successful recovery of groups
- Selection for $p = 1,000$: all TP identified, 2FP and 1 FN

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Comparison with two-step approach

(n=75, p=150, 1 to 5 relevant covariates per cluster)

- Step 1: Clustering with the `longclust` package:

		Truth		
		1	2	3
Predicted	1	24	0	7
	2	1	25	18

Table 2: For simulation with $\sigma_k^2 = 0.1$.

		Truth		
		1	2	3
Predicted	1	14	11	4
	2	0	0	12
	3	0	5	0
	4	9	0	0
	5	2	0	9
	6	0	9	0

Table 3: For simulation with $\sigma_k^2 = 1$.

↬ Difficulty separating some clusters

- Step 2: Variable selection in each identified cluster using Heuclin et al. (2021) fails to select the relevant covariates

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Conclusion

We proposed an innovative approach for clustering longitudinal data with variable selection

- Promising results
  - Robust to signal-to-noise ratio
  - Higher number of relevant predictors markers per cluster helps their successful recovery
  - Clustering of observations based on their profiles as well as their dependence on subsets of variables
- Perspectives
  - Need to improve computational speed
  - P-spline modeling for longitudinal effects for large number of repeated measures or high resolution outcome data
  - Extension to time varying covariates

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Thanks for your attention!

marie.denis@cirad.fr

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):815–849.

Heuclin, B., Mortier, F., Trottier, C., and Denis, M. (2021). Bayesian varying coefficient model with selection: An application to functional mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(1):24–50.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709.

McNicholas, P. D. and Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, 142(5):1114–1127.

Monni, S. and Tadesse, M. G. (2009). A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis*, 4(3):413–436.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.