Clustering with variable selection for longitudinal data: application to gene expression data.

Marie Denis and Mahlet G. Tadesse

ENAR, March 21, 2023

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Outline

Biological motivations

In many domains the processes of interest are dynamic (disease progression, growth...) ↬ Need to analyze response profiles

Yeast cell cycle dataset (Spellman et al., 1998; Lee et al., 2002)

Cell cycle gene expression data over two cell cycle periods along with binding information of transcription factors (TFs) from ChIP-chip data

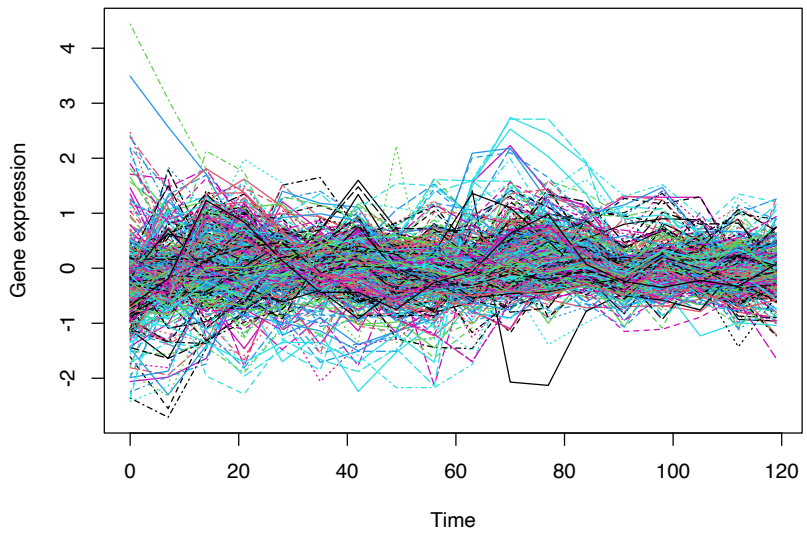

Figure 1: Cell cycle gene expression profiles over two cell cycle periods

Biological motivations

- Which cell cycle genes have similar expression profiles? Do they correspond to different biological functions?
- Which TFs are associated to the gene expression profiles? Which stage of the cell process do they influence?

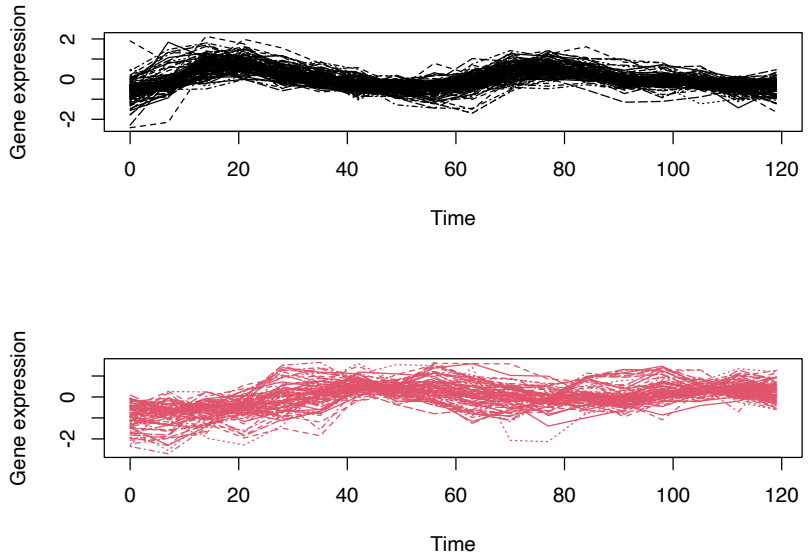

Figure 2: Gene expression profiles for two sub-groups of genes

Statistical motivations

- Which cell cycle genes have similar expression profiles? Do they correspond to different biological functions?

Statistical motivations

- Which cell cycle genes have similar expression profiles? Do they correspond to different biological functions?

To identify groups of genes with similar longitudinal response profiles

Statistical motivations

- Which cell cycle genes have similar expression profiles? Do they correspond to different biological functions?
- Which TFs are associated to the gene expression profiles? Which stage of the cell process do they influence?

To identify groups of genes with similar longitudinal response profiles

Statistical motivations

- Which cell cycle genes have similar expression profiles? Do they correspond to different biological functions?
- Which TFs are associated to the gene expression profiles? Which stage of the cell process do they influence?

To identify groups of genes with similar longitudinal response profiles

To identify subsets of TFs with time varying effects associated to each group of genes

↬ Statistical approach achieving both objectives simultaneously

Which statistical approaches ?

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

Table 1: Existing approaches wrt the type of outcome

↬A lack of (Bayesian) approaches

Outline

The proposed approach

A Bayesian stochastic partitioning method, based on the work of Monni and Tadesse (2009), which combines

1. mixture of mixed effects models for clustering taking into account temporal dependence,

and

2. variable selection for identifying relevant covariates.

The proposed approach

Data

n independent samples of T measures of an 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 $X_j = (X_{1j}, \ldots, X_{nj})'$ for $j = 1, \ldots, p$.

$$\mathcal{Y} = \begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1(T-1)} & Y_{1T} \\ Y_{21} & Y_{22} & \ldots & Y_{2(T-1)} & Y_{2T} \\ \vdots & & & & \vdots \\ Y_{n1} & Y_{n2} & \ldots & Y_{n(T-1)} & Y_{nT} \end{pmatrix} \quad \mathcal{X} = \begin{pmatrix} X_{11} & \ldots & X_{1(p-1)} & X_{1p} \\ X_{21} & \ldots & X_{2(p-1)} & X_{2p} \\ \vdots & & & \vdots \\ \vdots & \ldots & & \vdots \\ X_{n1} & \ldots & X_{n(p-1)} & X_{np} \end{pmatrix}$$

The proposed approach

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 $X_j = (X_{1j}, \ldots, X_{nj})'$ for $j = 1, \ldots, p$.

Objectives

To cluster individuals and to select their associated subsets of covariates with time varying effects

$$\mathcal{Y} = \begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1(T-1)} & Y_{1T} \\ Y_{21} & Y_{22} & \ldots & Y_{2(T-1)} & Y_{2T} \\ \vdots & & & & \vdots \\ Y_{n1} & Y_{n2} & \ldots & Y_{n(T-1)} & Y_{nT} \end{pmatrix} \quad \mathcal{X} = \begin{pmatrix} X_{11} & \ldots & X_{1(p-1)} & X_{1p} \\ X_{21} & \ldots & X_{2(p-1)} & X_{2p} \\ \vdots & & & \vdots \\ \vdots & \ldots & & \vdots \\ X_{n1} & \ldots & X_{n(p-1)} & X_{np} \end{pmatrix}$$

↬ **Mixture of mixed effects models with an unknown number of components**

Mixture of mixed effects models

Partitioning

Partition of variables into sets of pairs (X_J, Y_I) with $J \subset \{1, \ldots, p\}$ and $I \subset \{1, \ldots, n\}$ where a configuration $\mathcal{S}$ of length K is defined by:

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

$\mathcal{S}_k$ a component, $0 \leq |J_k| \leq p$, $1 \leq |I_k| \leq n$

- $\sum_{k=1}^{K} |I_k| = n \Rightarrow$ each individual is allocated to one and only one component,
- $\sum_{k=1}^{K} |J_K| \leq Kp \Rightarrow$ predictors may be allocated to several components.

An example of configuration with two components $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}})$ and $\mathcal{S}_2 = (Y_3, X_{\{1,2\}})$

$$\mathcal{Y} = \begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1(T-1)} & Y_{1T} \\ Y_{21} & Y_{22} & \ldots & Y_{2(T-1)} & Y_{2T} \\ Y_{31} & Y_{32} & \ldots & Y_{3(T-1)} & Y_{3T} \end{pmatrix} \quad \mathcal{X} = \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{pmatrix}$$

Mixture of mixed effects models

For $Y_{l_1}, \ldots, Y_{l_{nk}} \in \mathcal{S}_k = (|J_k|, |I_k|) = (m_k, n_k)$

Bayesian hierarchichal model

$$Y_i|\boldsymbol{\beta}_k, \sigma_k^2, \rho \sim \mathcal{N}_T(\sum_{r=1}^{m_k} x_{is_r}\boldsymbol{\beta}_{ks_r}, \sigma_k^2\Omega), \quad i = l_1, \ldots, l_{n_k},$$

$$\boldsymbol{\beta}_{ks_r} = (\beta_{ks_r}^1, \ldots, \beta_{ks_r}^T)'|\tau_k^2, \sigma_k^2 \sim \mathcal{N}_T(0, \sigma_k^2\tau_k^2(D'D)^{-1}), \quad r = 1, \ldots, m_k,$$

$$\tau_k^2 \sim \mathcal{IG}(a, b), \ \sigma_k^2 \sim \mathcal{IG}(\sigma_0^2, \nu), \ \rho \sim \mathcal{U}_{(-1,1)}$$

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

Temporal dependence taking into account via:

- Ω: a $T \times T$ auto-regressive correlation matrix of order 1 with unknown parameter ρ,
- D: a matrix representation of first order finite difference operator,

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

MCMC implementation

1. Update of configuration via a reversible jump Markov chain Monte Carlo:
    1. **Type 1**: Add or delete covariate to/from a component

MCMC implementation

1. Update of configuration via a reversible jump Markov chain Monte Carlo:
   1. **Type 1**: Add or delete covariate to/from a component
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{3\}})$

MCMC implementation

1. Update of configuration via a reversible jump Markov chain Monte Carlo:
   1. **Type 1**: Add or delete covariate to/from a component
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{3\}})$
   2. **Type 2**: Reallocate observations by choosing to split/merge components (m, n) (with $n > 0$) or to reassign a single observation.

MCMC implementation

1. Update of configuration via a reversible jump Markov chain Monte Carlo:
   1. **Type 1**: Add or delete covariate to/from a component
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{3\}})$
   2. **Type 2**: Reallocate observations by choosing to split/merge components (m, n) (with $n > 0$) or to reassign a single observation.
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_{11} = (Y_{\{1\}}, X_{\{1\}}), \mathcal{S}_{12} = (Y_{\{2\}}, X_{\{1,3\}})$
      $\mathcal{S}_{11} = (Y_{\{1,2\}}, X_{\{1,3\}}), \mathcal{S}_{12} = (Y_{\{3\}}, X_{\{1,2\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2,3\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}), \mathcal{S}_2 = (Y_{\{3\}}, X_{\{1,2\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1\}}, X_{\{1,3\}}),$
      $\mathcal{S}_2 = (Y_{\{3,2\}}, X_{\{1,2\}})$

MCMC implementation

1. Update of configuration via a reversible jump Markov chain Monte Carlo:
   1. **Type 1**: Add or delete covariate to/from a component
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{3\}})$
   2. **Type 2**: Reallocate observations by choosing to split/merge components (m, n) (with $n > 0$) or to reassign a single observation.
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_{11} = (Y_{\{1\}}, X_{\{1\}}), \mathcal{S}_{12} = (Y_{\{2\}}, X_{\{1,3\}})$
      $\mathcal{S}_{11} = (Y_{\{1,2\}}, X_{\{1,3\}}), \mathcal{S}_{12} = (Y_{\{3\}}, X_{\{1,2\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2,3\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}), \mathcal{S}_2 = (Y_{\{3\}}, X_{\{1,2\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1\}}, X_{\{1,3\}}),$
      $\mathcal{S}_2 = (Y_{\{3,2\}}, X_{\{1,2\}})$

MCMC implementation

1. Update of configuration via a reversible jump Markov chain Monte Carlo:
   1. **Type 1**: Add or delete covariate to/from a component
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{3\}})$
   2. **Type 2**: Reallocate observations by choosing to split/merge components (m, n) (with $n > 0$) or to reassign a single observation.
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_{11} = (Y_{\{1\}}, X_{\{1\}}), \mathcal{S}_{12} = (Y_{\{2\}}, X_{\{1,3\}})$
      $\mathcal{S}_{11} = (Y_{\{1,2\}}, X_{\{1,3\}}), \mathcal{S}_{12} = (Y_{\{3\}}, X_{\{1,2\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2,3\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}), \mathcal{S}_2 = (Y_{\{3\}}, X_{\{1,2\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1\}}, X_{\{1,3\}}),$
      $\mathcal{S}_2 = (Y_{\{3,2\}}, X_{\{1,2\}})$
2. Update of τ_k^2 for $k = 1, \ldots, K$ via a Metropolis-Hasting algorithm,

MCMC implementation

1. Update of configuration via a reversible jump Markov chain Monte Carlo:
   1. **Type 1**: Add or delete covariate to/from a component
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{3\}})$
   2. **Type 2**: Reallocate observations by choosing to split/merge components (m, n) (with $n > 0$) or to reassign a single observation.
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}) \Rightarrow \mathcal{S}_{11} = (Y_{\{1\}}, X_{\{1\}}), \mathcal{S}_{12} = (Y_{\{2\}}, X_{\{1,3\}})$
      $\mathcal{S}_{11} = (Y_{\{1,2\}}, X_{\{1,3\}}), \mathcal{S}_{12} = (Y_{\{3\}}, X_{\{1,2\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1,2,3\}}, X_{\{1,2,3\}})$
      $\mathcal{S}_1 = (Y_{\{1,2\}}, X_{\{1,3\}}), \mathcal{S}_2 = (Y_{\{3\}}, X_{\{1,2\}}) \Rightarrow \mathcal{S}_1 = (Y_{\{1\}}, X_{\{1,3\}}),$
      $\mathcal{S}_2 = (Y_{\{3,2\}}, X_{\{1,2\}})$
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.

Outline

Simulation study

Evaluate prediction and selection performances under different simulation settings:

- different residual variances, σ_k^2
- different number of covariates, p
- different number of time points, T
- different number of relevant predictors per cluster, m_k

$n = 75, p = 150$					
$\sigma_k^2 = 0.1$			$\sigma_k^2 = 1$		
$T = 10$		$T = 50$	$T = 10$		$T = 50$
$m_k = 1$	$m_k \in \{1, \ldots, 5\}$	$m_k \in \{1, \ldots, 5\}$	$m_k = 1$	$m_k \in \{1, \ldots, 5\}$	$m_k \in \{1, \ldots, 5\}$

$n = 75, p = 1000, T = 10, m_k \in \{1, \ldots, 10\}$	
$\sigma_k^2 = 0.1$	$\sigma_k^2 = 1$

Table 2: Summary of simulated scenarios

Simulation study

No available approaches for simultaneously clustering longitudinal profiles and selecting subsets of covariates associated to each cluster

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

Compare results with a two-step approach:

1. Step 1: Cluster individuals based on response profiles ignoring covariates
2. Step 2: Fit mixed effects model with variable selection within each cluster

Simulation study

No available approaches for simultaneously clustering longitudinal profiles and selecting subsets of covariates associated to each cluster

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

Compare results with a two-step approach:

1. Step 1: Cluster individuals based on response profiles ignoring covariates
2. Step 2: Fit mixed effects model with variable selection within each cluster

Other two-step approach:

1. Step1: Fit mixed effects models with variables selection on all individual
2. Step2: Using selected covariates, cluster individual based on response profiles

Results

Good performances in terms of selection and prediction:

↬Successful inference for clustering and selection in most scenarios

- A higher number of time points impacts slightly convergence

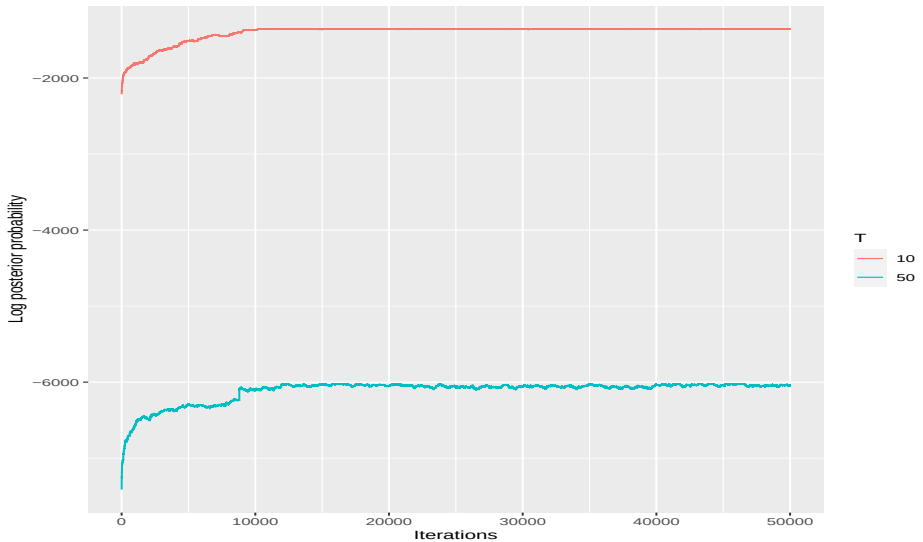

Figure 3: Log posterior probability over iterations for simulations with $T = 10$ (in red) and $T = 50$ (in blue).

Results

- A higher signal-to-noise ratio slows convergence (higher number of covariates or/and higher residual variance)

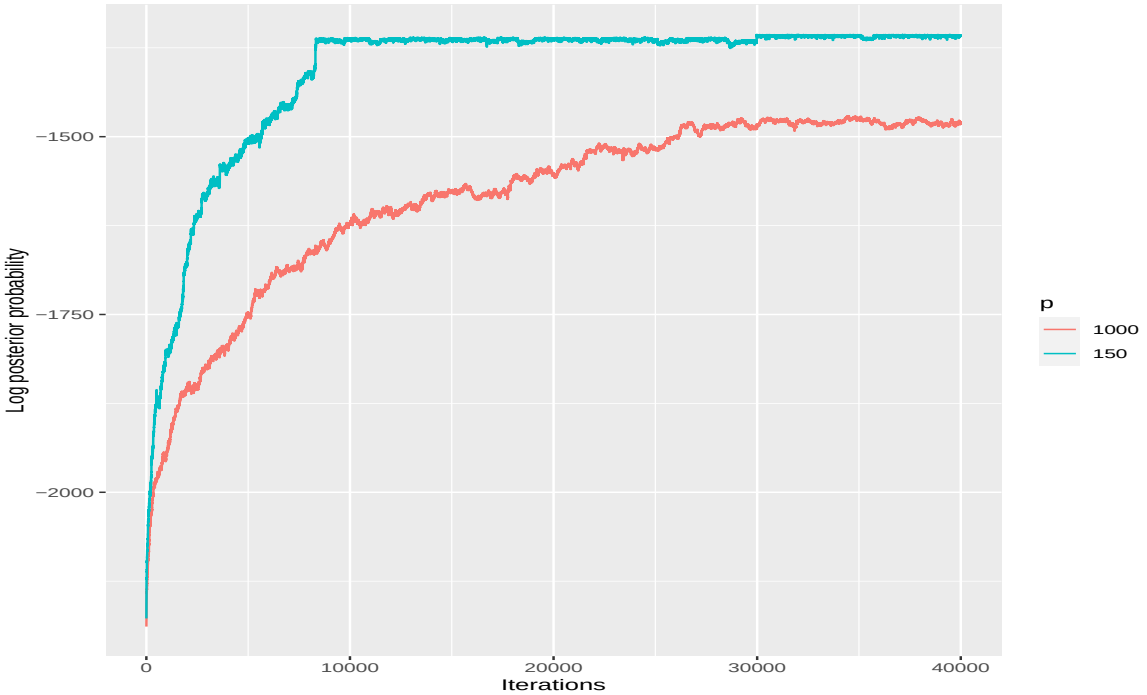

Figure 4: Log posterior probability over iterations for simulations with $p = 150$ (in blue) and $p = 1000$ (in red).

- A higher number of significant covariates per cluster helps uncover groups

Comparison with two-step approach

- Step 1: Clustering with the `longclust` package (McNicholas and Subedi, 2012):

↬ Difficulty separating some clusters

		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: Confusion matrix for simulation with $\sigma_k = 1$.

- Step 2: Variable selection in each identified cluster using the approach developed by Heuclin et al. (2021) fails to identify the relevant variables

Outline

Biological context

Yeast Cell Cycle Dataset

- Y: Expression levels of 542 cell cycle genes measured every 7 minutes during 119 minutes (18 time points) (Spellman et al., 1998),
- X: Binding information from ChIP-chip data for 106 transcription factors (Lee et al., 2002),

Objectives:

Identification of TFs involved in gene regulation during the cell cycle process

- ↬ Need to uncover groups of gene expression profiles
- ↬ Need to identify TFs associated to cell cycle gene profiles in each group

Results

- Identification of 4 clusters of cell cycle genes enriched in different biological processes

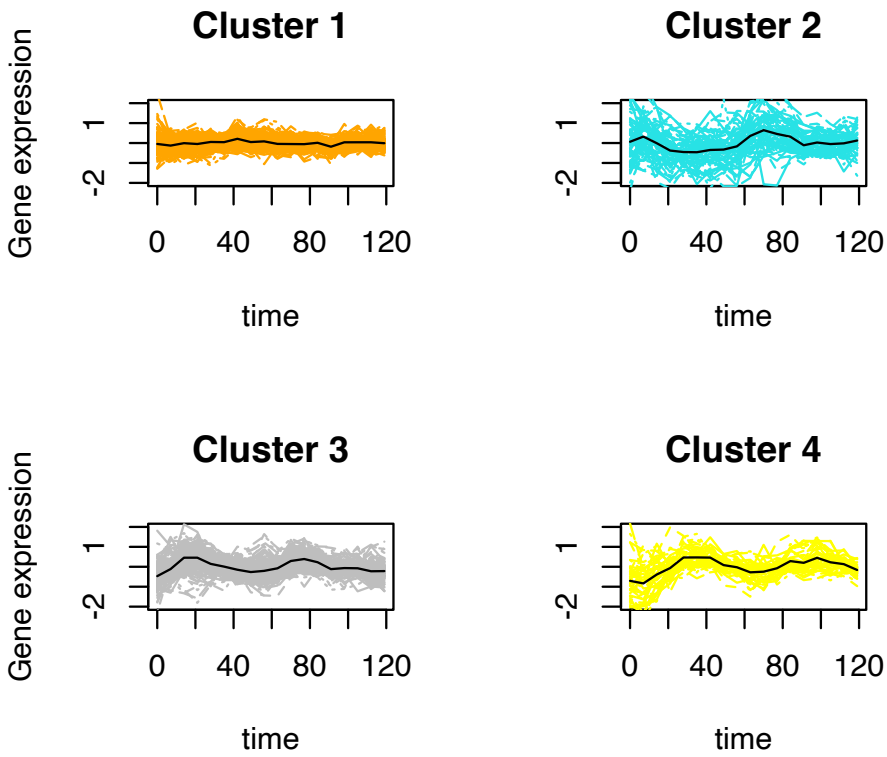

Figure 5: Gene expression profiles for each cluster

Results

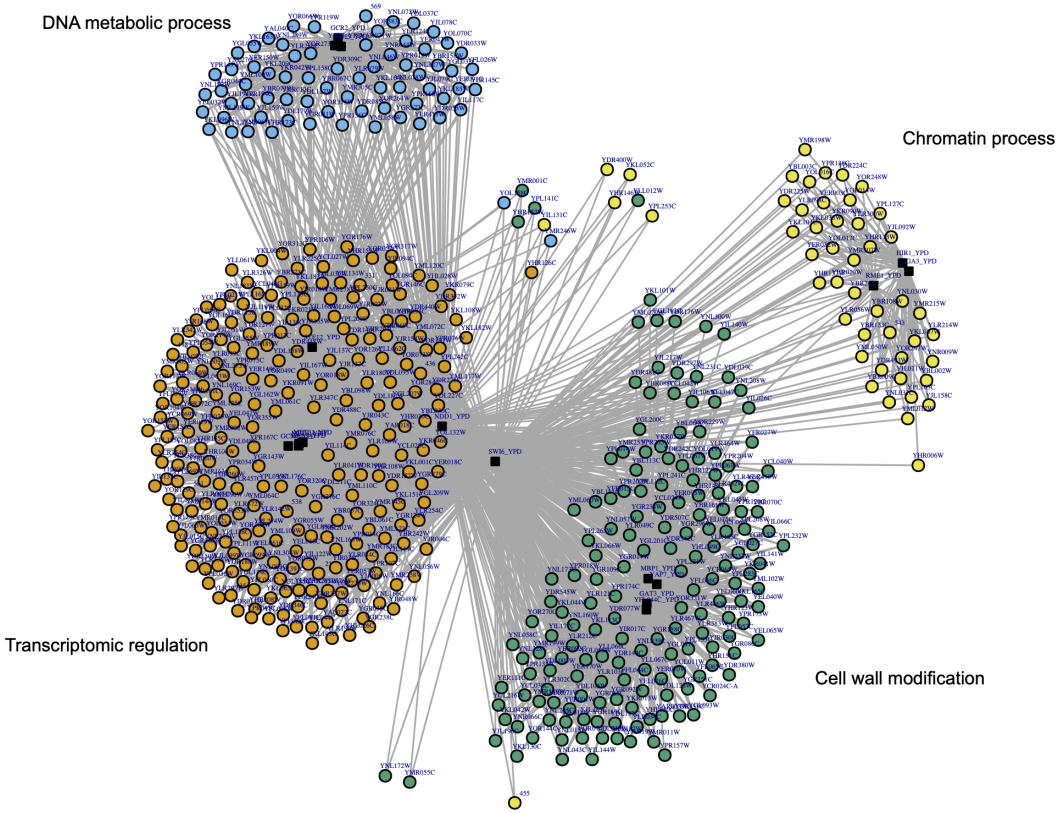

Figure 6: Network of association between TFs (black squares) and genes (circles) focusing on posterior probabilities > 0.5. Genes are colored according to their enrichment.

Results

- Selection of 17 TFs: 7 of these are experimentally verified
- Post-hoc estimation of time varying effects associated to the selected TFs: identification of stages of the cell process that are influenced by TFs

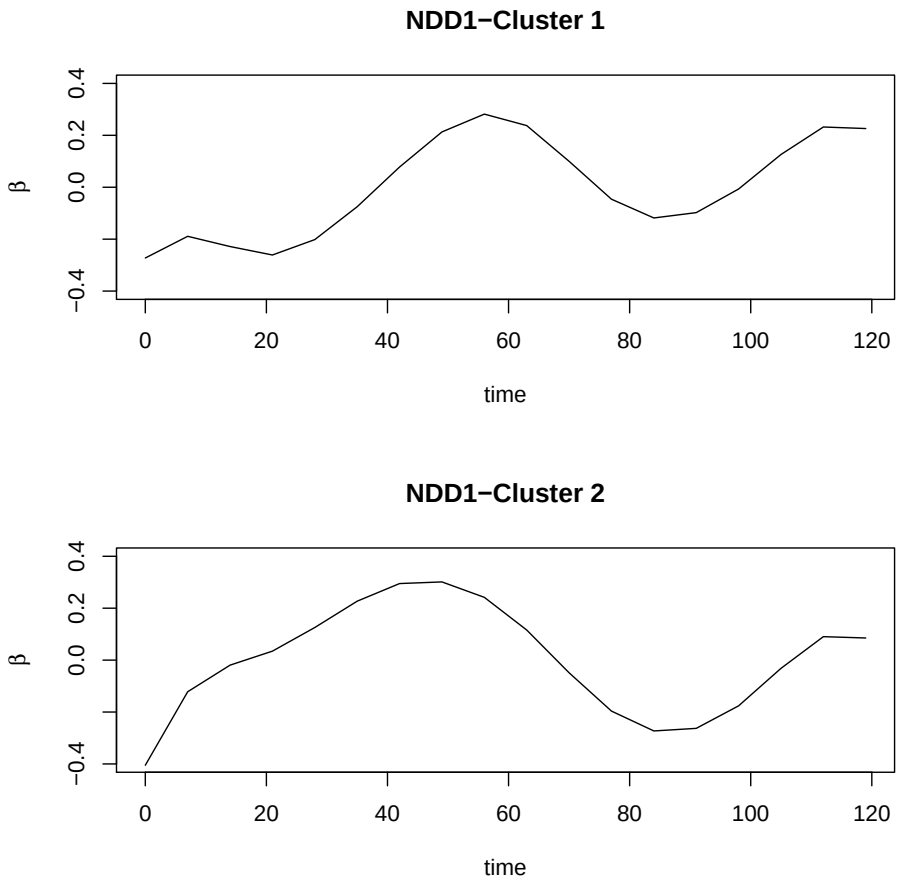

Figure 7: Varying coefficients for transcription factor NDD1 in clusters 1 and 2.

Outline

Conclusion

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

- Promising results
  - Robust to signal-to-noise ratio
  - Combination of clustering and selection helps successful recovery of clusters
  - Relevant biological results
- 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

Thanks for your attention!

marie.denis@cirad.fr

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.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *science*, 298(5594):799–804.

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.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297.

Yang, L. and Wu, T. T. (2022). Model-based clustering of high-dimensional longitudinal data via regularization. *Biometrics*.