MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

ISPED SCHOOL OF PUBLIC HEALTH / Université de BORDEAUX

Nonparametric Bayesian mixture models for identifying clusters from longitudinal and cross-sectional data

Anaïs Rouanet - Associate Professor

joint work with Rob Johnson, Sylvia Richardson, Brian Tom, Paul Kirk
Inserm U1219, Bordeaux Population Health, Université de Bordeaux

JSS - June 2024, 20th

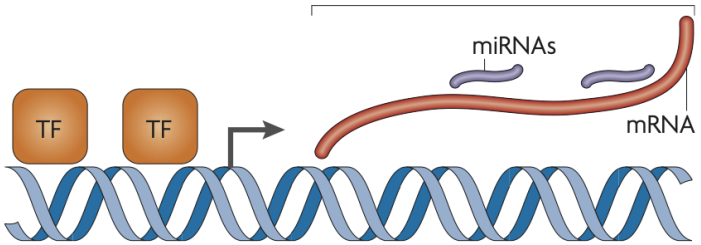
Motivation

Gene regulation

Major challenge in molecular biology

Exploring underlying regulatory mechanisms

Understanding the transcription process DNA to mRNA



(Chen & Rajewsky, 2007)

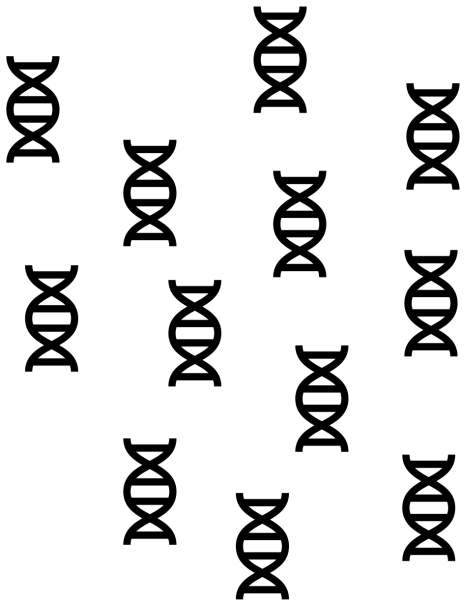
Statistical challenges

Study of model organisms (budding-yeast *Saccharomyces cerevisiae*)

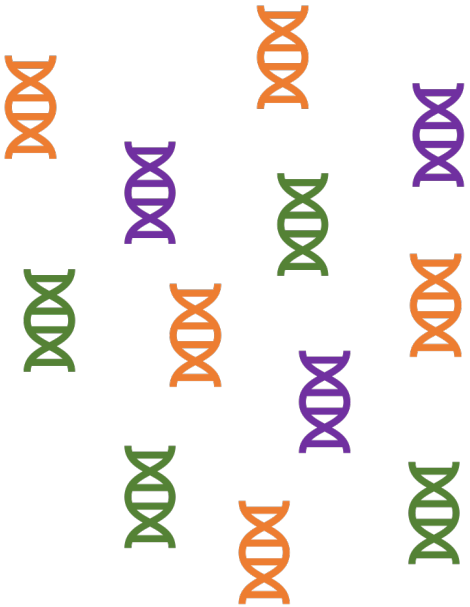
Integrating different types of omic data

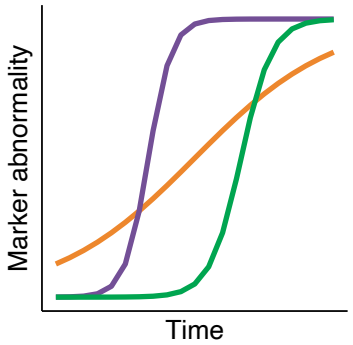
Identification of transcriptional modules

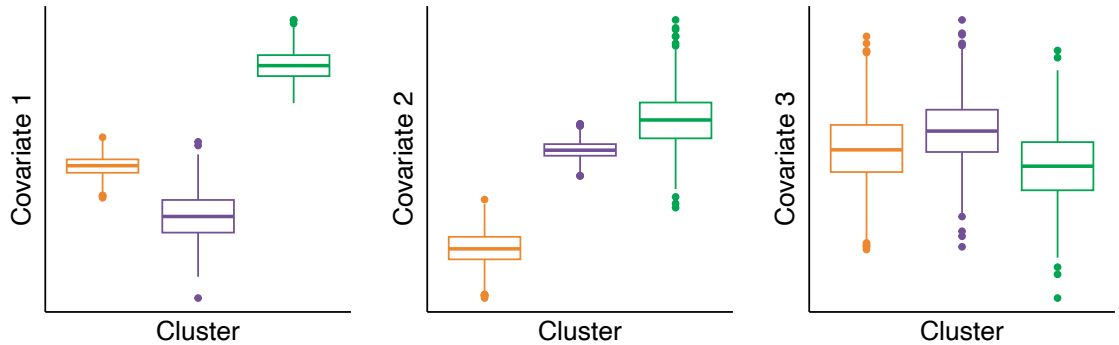
Clustering objective



Clustering objective







Methodological challenges

Outcome-driven clustering approach

- Unknown number of clusters

Integration of diverse types of data

- Longitudinal marker (gene expression)
- Cross-sectional variables (transcription factors)

Selection of key markers for discrimination

- Cross-sectional variables

Methodological challenges

Outcome-driven clustering approach

- Unknown number of clusters

→ **Bayesian infinite mixture models**

Integration of diverse types of data

- Longitudinal marker (gene expression)
- Cross-sectional variables (transcription factors)

Selection of key markers for discrimination

- Cross-sectional variables

Methodological challenges

Outcome-driven clustering approach

- Unknown number of clusters

→ **Bayesian infinite mixture models**

Integration of diverse types of data

- Longitudinal marker (gene expression)
- Cross-sectional variables (transcription factors)

→ **Joint analysis**

Selection of key markers for discrimination

- Cross-sectional variables

Methodological challenges

Outcome-driven clustering approach

- Unknown number of clusters

→ **Bayesian infinite mixture models**

Integration of diverse types of data

- Longitudinal marker (gene expression)
- Cross-sectional variables (transcription factors)

→ **Joint analysis**

Selection of key markers for discrimination

- Cross-sectional variables

→ **Variable selection**

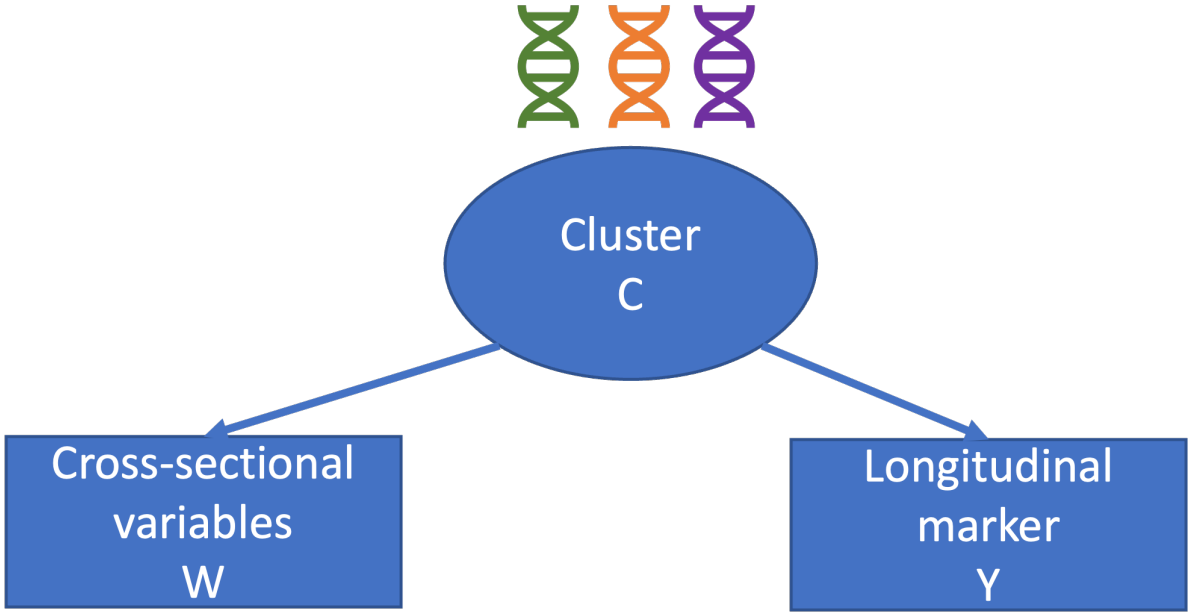
Objective

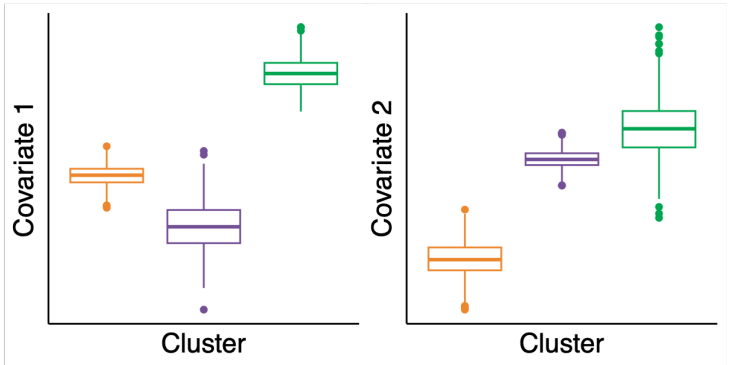
Uncovering the heterogeneity in the population

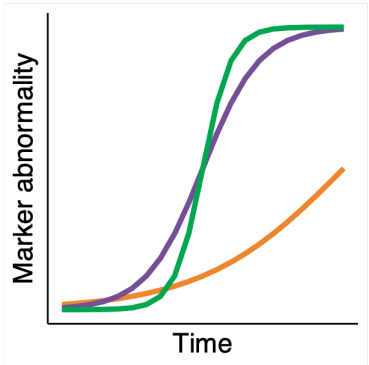
by identifying trajectories of a longitudinal marker

associated with specific profiles of cross-sectional variables

Profile regression - Rouanet et al. 2023

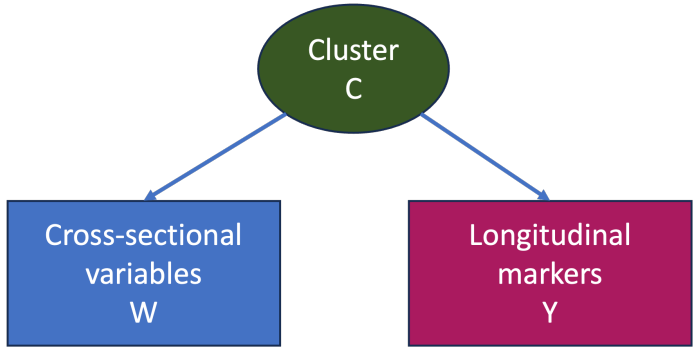






https://github.com/premium-profile-regression.

Profile regression



Notations:

c_i cluster allocation variable of subject i
W_i individual vector of cross-sectional variables
Y_i vector of individual longitudinal measurements

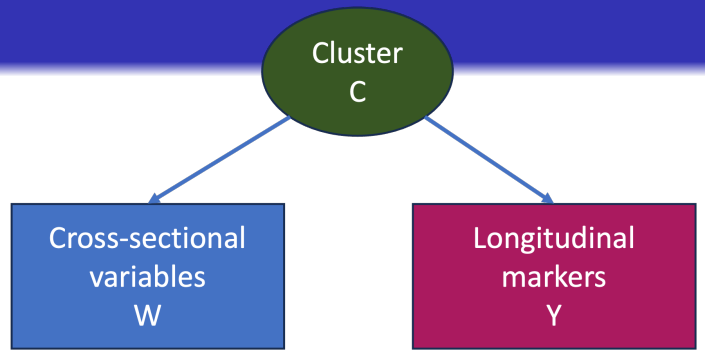
Likelihood:

$$L(\theta|Y,W) = \prod_{i=1}^{N} \sum_{g=1}^{+\infty} P(c_i = g|\theta^c)\ f(W_i|c_i = g;\theta_g^W)\ f(Y_i|c_i = g;\theta_g^Y)$$

Assumption:

Independence assumption between W and Y given cluster allocations c.

Profile regression



Notations:

c_i cluster allocation variable of subject i
W_i individual vector of cross-sectional variables
Y_i vector of individual longitudinal measurements

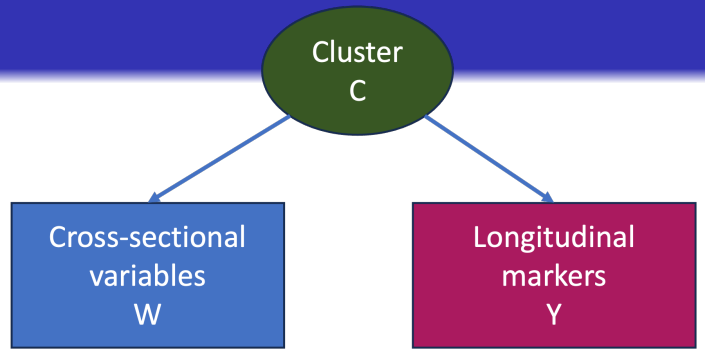
Likelihood:

$$L(\theta|Y,W) = \prod_{i=1}^{N}\sum_{g=1}^{G} \underbrace{P(c_i = g|\theta^c)}_{\text{Mixture weights}} f(W_i|c_i = g;\theta_g^W)\ f(Y_i|c_i = g;\theta_g^Y)$$

$$\mathcal{A} = \sum_{g=1}^{G} \pi_g\ \delta_{\theta_g}$$

$$\boldsymbol{\pi}|\alpha \sim Dirichlet(\alpha/G, ..., \alpha/G)$$
$$c_i|\boldsymbol{\pi} \sim Discrete(\pi_1, ..., \pi_G)$$
$$\theta_{c_i} \sim \mathcal{A}_0$$
$$Y_i, W_i|c_i \sim f(\cdot;\theta_{c_i})$$

Profile regression



Notations:

c_i cluster allocation variable of subject i
W_i individual vector of cross-sectional variables
Y_i vector of individual longitudinal measurements

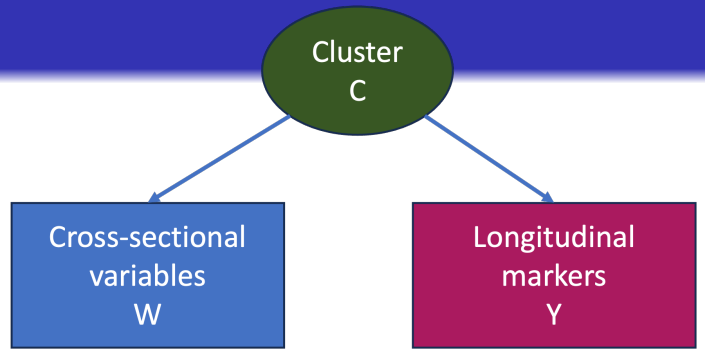
Likelihood:

$$L(\theta|Y,W) = \prod_{i=1}^{N} \sum_{g=1}^{+\infty} \underbrace{P(c_i = g|\theta^c)}_{\text{Mixture weights}} f(W_i|c_i = g;\theta_g^W)\ f(Y_i|c_i = g;\theta_g^Y)$$

Probability measure $\mathcal{A} = \sum_{g=1}^{+\infty} \pi_g\ \delta_{\theta_g}$ (Hastie et al., 2015)

$\mathcal{A} \sim DP(\alpha, \mathcal{A}_0)$ Dirichlet Process prior
$\theta_g \sim \mathcal{A}_0$ Base distribution

Profile regression



Notations:

c_i cluster allocation variable of subject i
W_i individual vector of cross-sectional variables
Y_i vector of individual longitudinal measurements

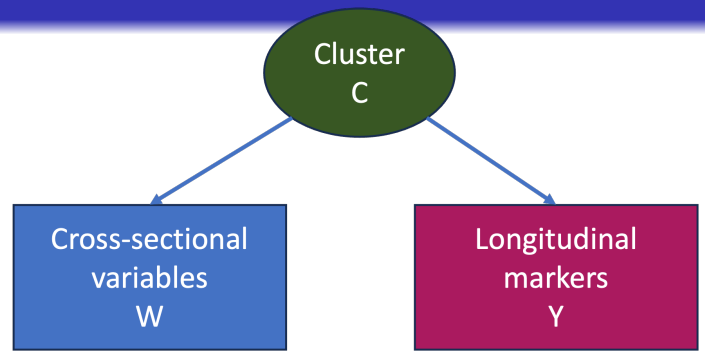
Likelihood:

$$L(\theta|Y,W) = \prod_{i=1}^{N} \sum_{g=1}^{+\infty} P(c_i = g|\theta^c) \underbrace{f(W_i|c_i = g;\theta_g^W)}_{\text{cross-sectional data}} f(Y_i|c_i = g;\theta_g^Y)$$

Model for cross-sectional variables given cluster g:

Categorical case: $W_i \sim Discrete(\phi_g)$ with $\phi_g \sim Dirichlet(a)$
Continuous case: $W_i \sim \mathcal{N}(\mu_g; V_g)$ with $\mu_g \sim \mathcal{N}(\mu_0; V_0)$
$V_g \sim InvWishart(R_0, K_0)$

Profile regression



Notations:

c_i cluster allocation variable of subject i
W_i individual vector of cross-sectional variables
Y_i vector of individual longitudinal measurements

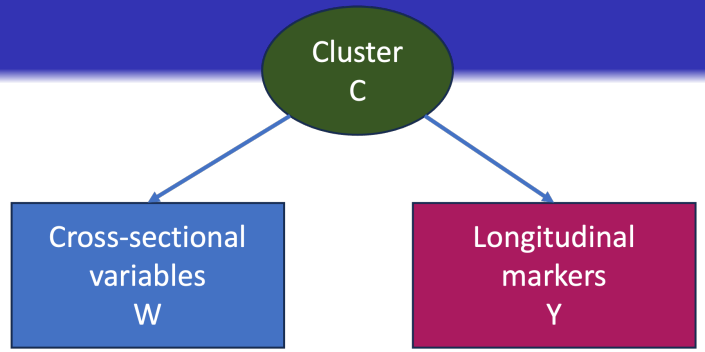
Likelihood:

$$L(\theta|Y,W) = \prod_{i=1}^{N} \sum_{g=1}^{+\infty} P(c_i = g|\theta^c)\ f(W_i|c_i = g;\theta_g^W)\ \underbrace{f(Y_i|c_i = g;\theta_g^Y)}_{\text{longitudinal data}}$$

Longitudinal outcome given cluster g:

$Y_i \sim MVN(\boldsymbol{\mu_g^L}; \boldsymbol{\Sigma_g})$ with $\boldsymbol{\mu_g^L} \sim MVN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_g/\kappa_0)$ and $\boldsymbol{\Sigma}_g \sim \mathcal{W}^{-1}(\mathrm{R}_0, \nu_0)$

Profile regression



Notations:

c_i cluster allocation variable of subject i
W_i individual vector of cross-sectional variables
Y_i vector of individual longitudinal measurements

Likelihood:

$$L(\theta|Y,W) = \prod_{i=1}^{N} \sum_{g=1}^{+\infty} P(c_i = g|\theta^c)\ f(W_i|c_i = g;\theta_g^W)$$

Longitudinal outcome given cluster g:

$Y_{ij} = f_g(t_{ij}) + \epsilon_{ijg}$, with $\epsilon_{ijg} \sim \mathcal{N}(0, L_{g,1})$ at time point t_{ij}

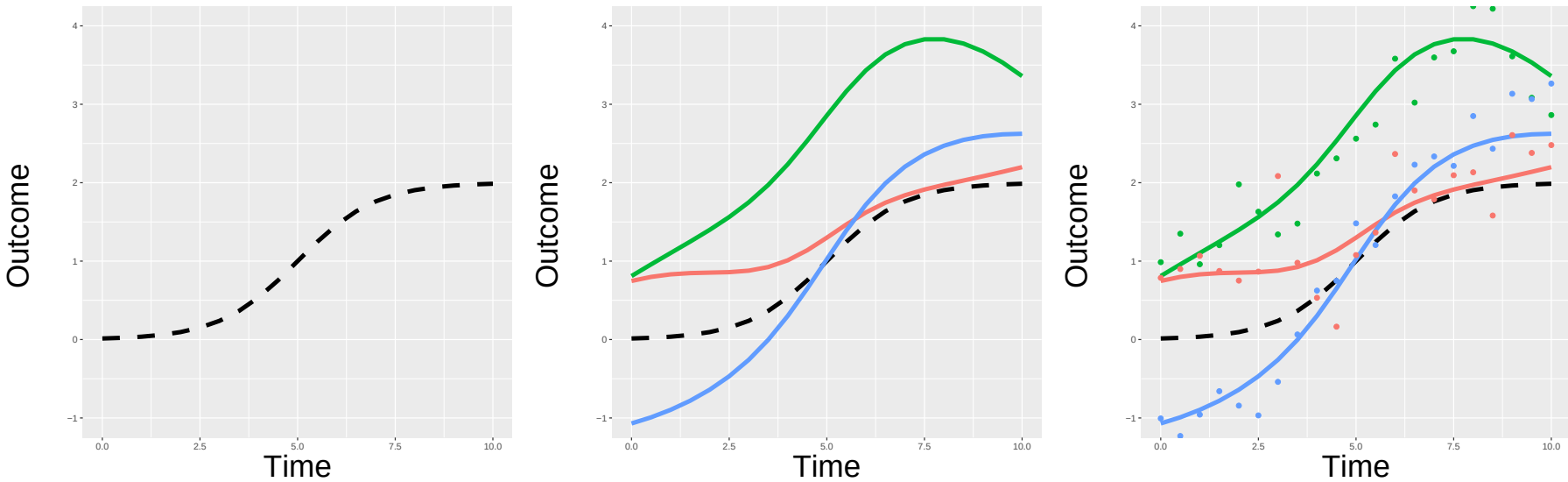
Gaussian Process on the mean evolution $f_g(\cdot)$:

$f_g(\cdot) \sim GP(\mathbf{0},\ \mathcal{K}_g(\cdot,\cdot))$ with $\mathcal{K}_g(s,t) = L_{g,2}\ \exp\left(-\frac{(s-t)^2}{L_{g,3}}\right)$

Gaussian Process prior

Given the cluster g :

$$Y_{ij} = f_g(t_{ij}) + \epsilon_{ijg} \text{ with } \epsilon_{ijg} \sim \mathcal{N}(0, L_{g,1})$$
$$f_g(\cdot) \sim GP(m_g(\cdot), \mathcal{K}_g(\cdot,\cdot))$$



Variable selection of cross-sectional variables (Liverani et al., 2015)

Cluster-specific distribution of cross-sectional variable q in cluster g:

$$W_q \sim \mathcal{N}(\mu_{g,q}, V_{g,q})$$

Selection weight for cross-sectional variable q:

$$\zeta_q \in [0, 1]$$

Composite mean parameter in the likelihood:

$$\mu^*_{g,q} = \zeta_q \ \mu_{g,q} + (1 - \zeta_q) \ \bar{w}_q$$

with $\bar{w}_q$ the empirical mean of cross-sectional variable q in the population.

NB: Sparsity inducing prior on $\zeta_q \sim 1_{\{w_q=0\}} \delta_0(\zeta_q) + 1_{\{w_q=1\}} Beta(0.5, 0.5)$, $w_q \sim Bern(0.5)$

Variable selection of cross-sectional variables (Liverani et al., 2015)

Cluster-specific distribution of cross-sectional variable q in cluster g:

$$W_q \sim \mathcal{N}(\mu_{g,q}, V_{g,q})$$

Selection weight for cross-sectional variable q:

$$\zeta_q \in [0, 1]$$

Composite mean parameter in the likelihood:

$$\mu^*_{g,q} = 0 \ \mu_{g,q} + (1 - 0) \ \bar{w}_q$$

with $\bar{w}_q$ the empirical mean of cross-sectional variable q in the population.

NB: Sparsity inducing prior on $\zeta_q \sim 1_{\{w_q=0\}} \delta_0(\zeta_q) + 1_{\{w_q=1\}} Beta(0.5, 0.5)$, $w_q \sim Bern(0.5)$

Variable selection of cross-sectional variables (Liverani et al., 2015)

Cluster-specific distribution of cross-sectional variable q in cluster g:

$$W_q \sim \mathcal{N}(\mu_{g,q}, V_{g,q})$$

Selection weight for cross-sectional variable q:

$$\zeta_q \in [0, 1]$$

Composite mean parameter in the likelihood:

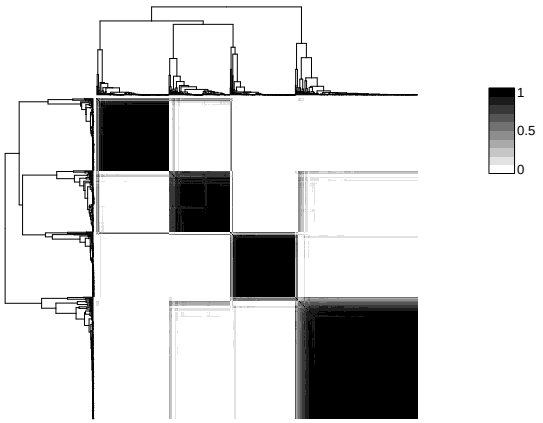
$$\mu^*_{g,q} = 1\ \mu_{g,q} + (1 - 1)\ \bar{w}_q$$

with $\bar{w}_q$ the empirical mean of cross-sectional variable q in the population.

NB: Sparsity inducing prior on $\zeta_q \sim 1_{\{w_q=0\}}\delta_0(\zeta_q) + 1_{\{w_q=1\}}Beta(0.5, 0.5)$, $w_q \sim Bern(0.5)$

Best partition via Post processing

- MCMC Gibbs sampler
- Posterior similarity matrix S: pairwise probability of co-clustering



- Selection of representative clustering P^*

  Partitioning around medoids on $1 - S$
  Maximizing average silhouette width

- Cluster-specific estimates as average over MCMC iterations:

$$\hat{\Theta}_g = \frac{1}{K} \frac{1}{N_g} \sum_{k=1}^{K} \sum_{i|c_i^*=g} \Theta_{c_i^{(k)}}$$

$\Theta_{c_i^{(k)}}$ parameter for the cluster subject i is allocated to, at iteration k
N_g the number of subjects in cluster g of P^*.

Simulations

Study 1 - Design

2 clusters of 50 individuals each.

4 cross-sectional data: variables 1-2 correspond to partition 1
variables 3-4 to alternative partition 2

$$\boldsymbol{\phi}_{g,q} = v\boldsymbol{\phi}^{(0)}_{g,q} + \frac{1}{R}(1-v)\mathbf{1}_R$$

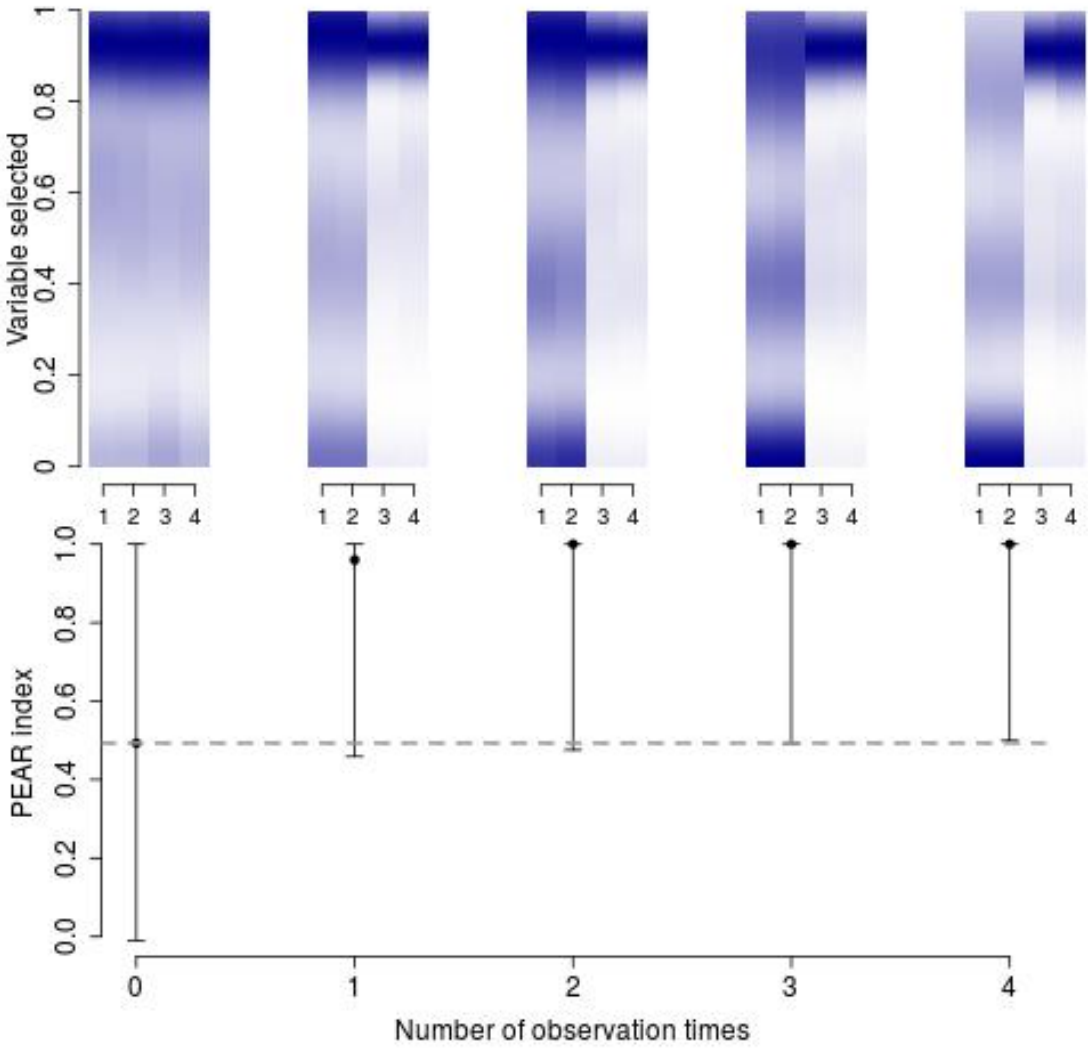
with $\boldsymbol{\phi}^{(0)}_{g,q} \sim \text{Dir}(\alpha_0 \mathbf{1}_R)$, $\alpha_0 = 0.01$
$\boldsymbol{\phi}^{(0)}_{g,q}$ and $\mathbf{1}_R$ vectors of length $R = 3$

Outcome data for the 2 clusters simulated (partition 2):

$$y_i | c_i = g \sim \left\{ \begin{array}{ll} MVN(\mathbf{1}_M, 0.5 I_M) & g = 1 \\ MVN(4 \cdot \mathbf{1}_M, 0.5 I_M) & g = 2 \end{array} \right\}.$$

In total, 5 scenari (with varying M values) repeated 1,000 times.

Study 1 - Results



Study 2 - Design

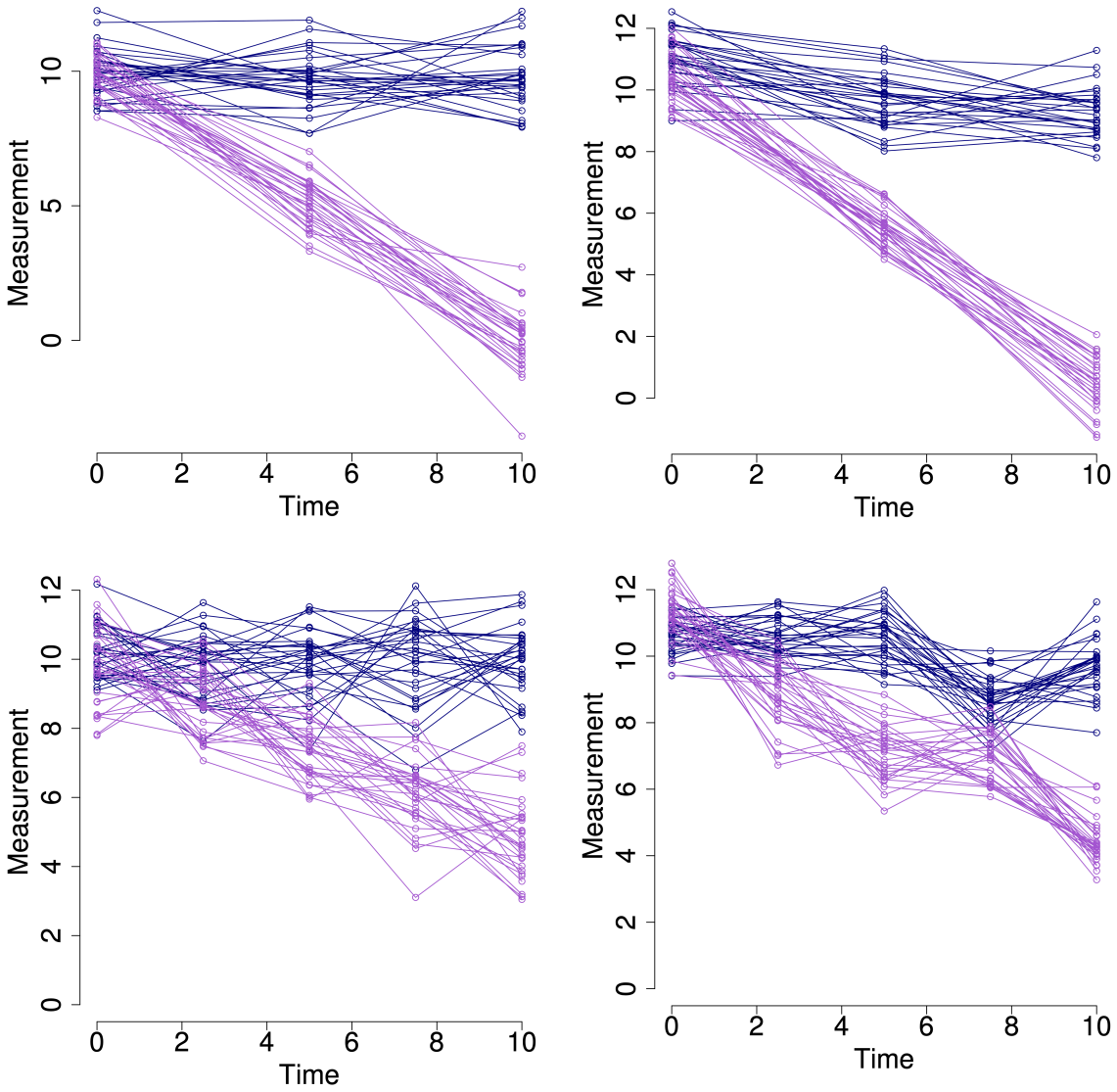
2 clusters of 30 individuals each.

Outcome data for the 2 clusters simulated:

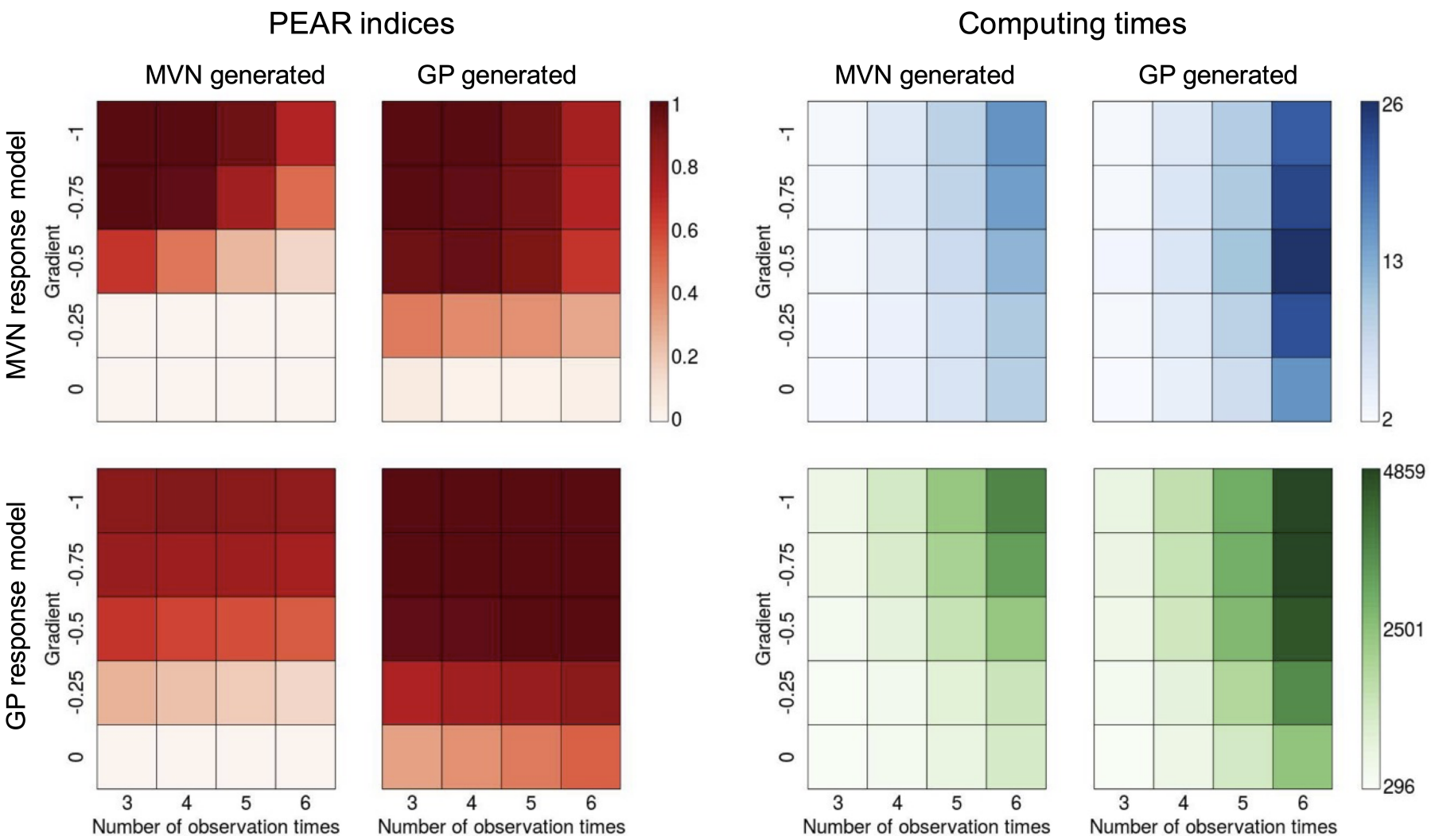
- mean vector: $10 \cdot \mathbf{1}_M$ and $10(1 - \xi \cdot j)$, with $j = 0, \cdots, M - 1$
- M the number of equally spaced time points
- ξ a gradient (or separability parameter) ranging from -1 to 0
- Same covariance matrices for both clusters:
  - MVN generation model: $M \times M$ diagonal matrix with $M_{i,i} = 1$ and $M_{i,i-1} = 0.5$
  - GP generation model: squared exponential covariance function with $\{\log(L_{g,1}), \log(L_{g,2}), \log(L_{g,3})\} = \{-0.5, -0.1, -0.5\}$

In total, 40 scenari (with varying ξ, M, MVN/GP) repeated 1,000 times.

Study 2 - Generated longitudinal data



Study 2 - Results



Application to budding yeast data

Context

ChIP data (Harbison et al., 2004)
Binary data containing binding information between transcription factors and genes.

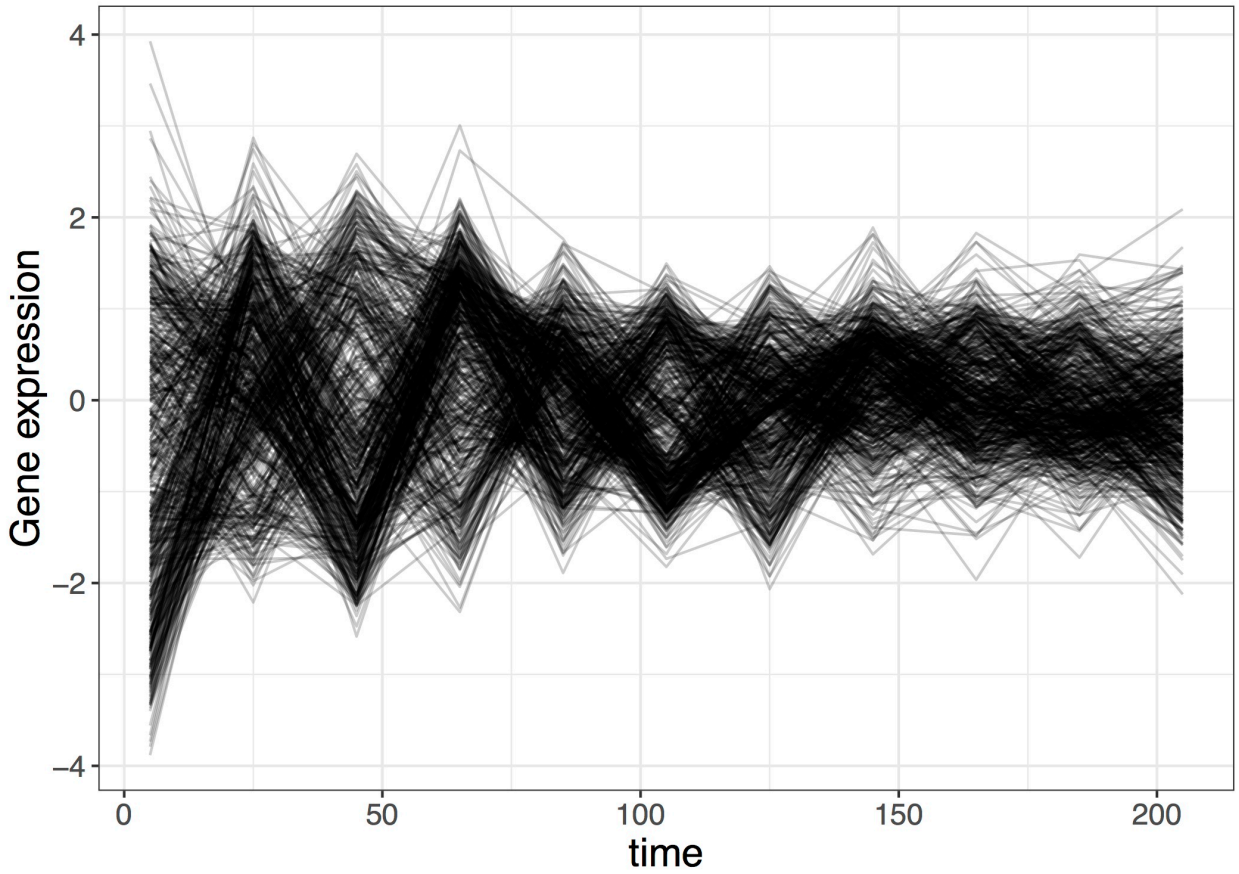
Gene expression time course (Granovskaia et al., 2010)
RNA expression collected every 5mins over 3 cell cycles.

Objective:

- Uncovering groups of genes co-regulated during budding yeast cell cycle
- Identifying transcription factors involved in this process.

Analytical sample

Analytical sample: 551 genes and 80 transcription factors



Model

Specification for longitudinal and cross-sectional markers:

$$f(\mathbf{W}_i, \mathbf{Y}_i; \boldsymbol{\Theta}) = \sum_{g=1}^{\infty} \pi_g \; f(\mathbf{Y}_i|\boldsymbol{\theta}_g) \prod_{q=1}^{80} f(W_{i,q}|\boldsymbol{\phi}_g)$$

$$\boldsymbol{\pi} \sim GEM(\alpha)$$

$$W_{i,q}|_{c_i} \sim Bernoulli(\phi_{c_i,q})$$

$$Y_{ij}|_{c_i} = f^{(c_i)}(t_{ij}) + \epsilon_i(t_{ij}) \text{ with } \epsilon_i(t_{ij}) \sim \mathcal{N}(0, L_{c_i,1})$$

Model

Specification for longitudinal and cross-sectional markers:

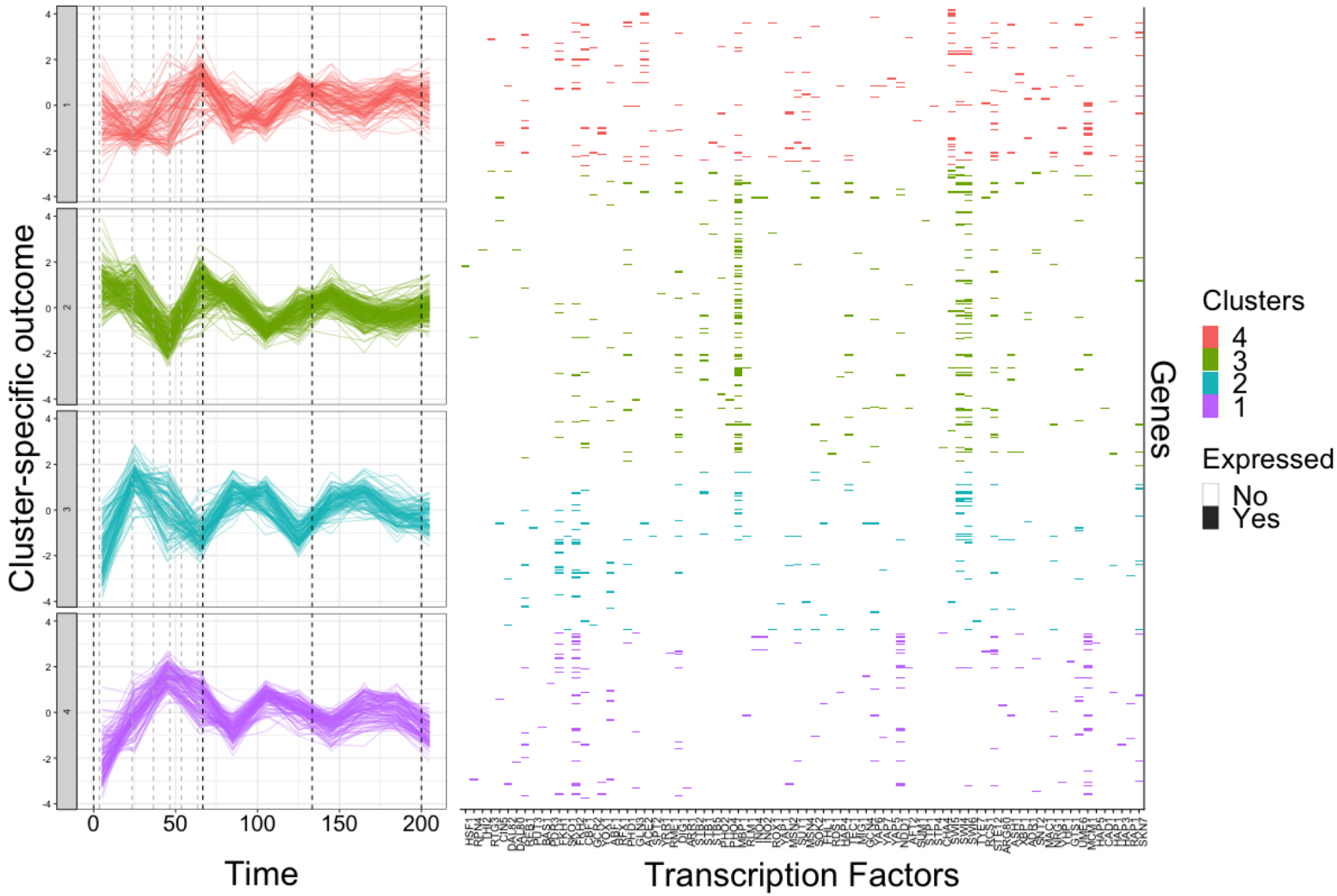
$$f(\mathbf{W}_i, \mathbf{Y}_i; \mathbf{\Theta}) = \sum_{g=1}^{\infty} \pi_g \; f(\mathbf{Y}_i|\theta_g) \prod_{q=1}^{80} f(W_{i,q}|\phi_g)$$

$$\begin{aligned} \boldsymbol{\pi} &\sim GEM(\alpha) \\ W_{i,q}|_{c_i} &\sim Bernoulli(\phi_{c_i,q}) \\ Y_{ij}|_{c_i} &= f^{(c_i)}(t_{ij}) + \epsilon_i(t_{ij}) \text{ with } \epsilon_i(t_{ij}) \sim \mathcal{N}(0, L_{c_i,1}) \end{aligned}$$

Prior distributions:

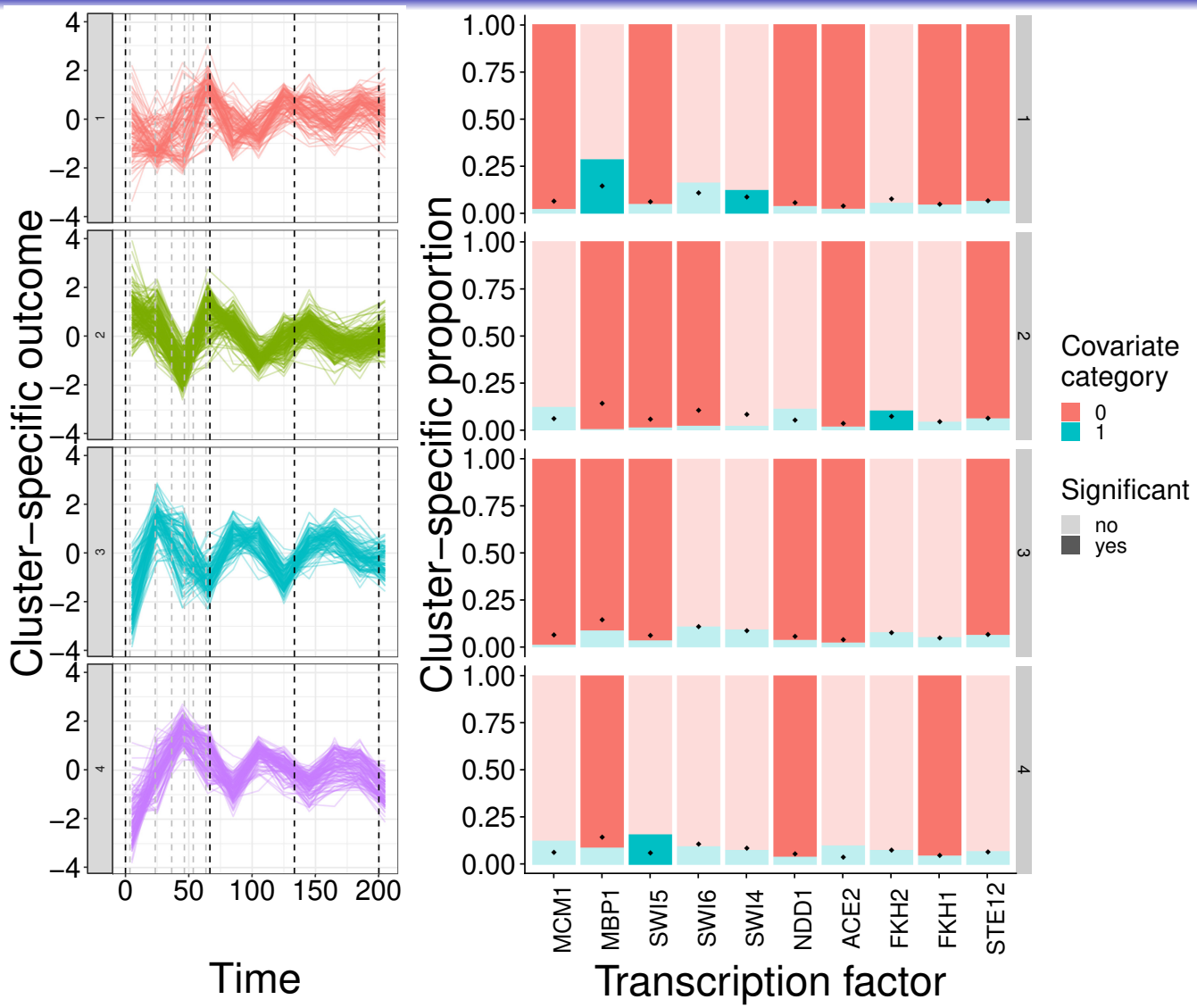
$$\begin{aligned} \alpha &\sim Gamma(2,1) \\ \phi_{g,q} &\sim Dirichlet(1) \\ f^{(g)} &\sim \mathcal{GP}(m^{(g)} \equiv 0, \mathcal{K}^{(g)}) \\ \log(L_{g,2}) &\sim \mathcal{N}(1.48, 0.5) \\ \log(L_{g,3}) &\sim \mathcal{N}(5, 0.1) \end{aligned}$$

Results



4 groups of 109, 206, 113, and 123 genes.

Results



4 groups of 109, 206, 113, and 123 genes.

GOTO scores (Mistry & Pavlidis, 2008)

Gene Ontology Term Overlap scores (GOTO):

- average number of shared GO annotations for a pair of genes from the same cluster
- validates biological function homogeneity of gene clusters

Method	GOTO (bp)	GOTO (mf)	GOTO (cc)	Number of genes
PReMiuMlongi	5.77	0.88	8.14	551
iCluster (G=2)	5.90	0.89	8.18	551

bp: biological process ontology
mf: molecular function ontology
cc: cellular component ontology

Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **73**, 314–339
https://doi.org/10.1093/jrsssc/qlad097
Advance access publication 8 November 2023
Original Article

Bayesian profile regression for clustering analysis involving a longitudinal response and explanatory variables

Anaïs Rouanet[1], Rob Johnson[1], Magdalena Strauss[1,2], Sylvia Richardson[1], Brian D. Tom[1], Simon R. White[1,3] and Paul D.W. Kirk[1,4]

[1]MRC Biostatistics Unit, Institute of Public Health, University of Cambridge, Cambridge CB2 0SR, UK
[2]EMBL-EBI, Wellcome Genome Campus, Hinxton CB10 1SD, UK
[3]Department of Psychiatry, University of Cambridge, Cambridge CB2 3EB, UK
[4]Cambridge Institute of Therapeutic Immunology & Infectious Disease, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge CB2 0AW, UK

Address for correspondence: Paul D.W. Kirk, MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, CB2 0SR, UK. Email: paul.kirk@mrc-bsu.cam.ac.uk

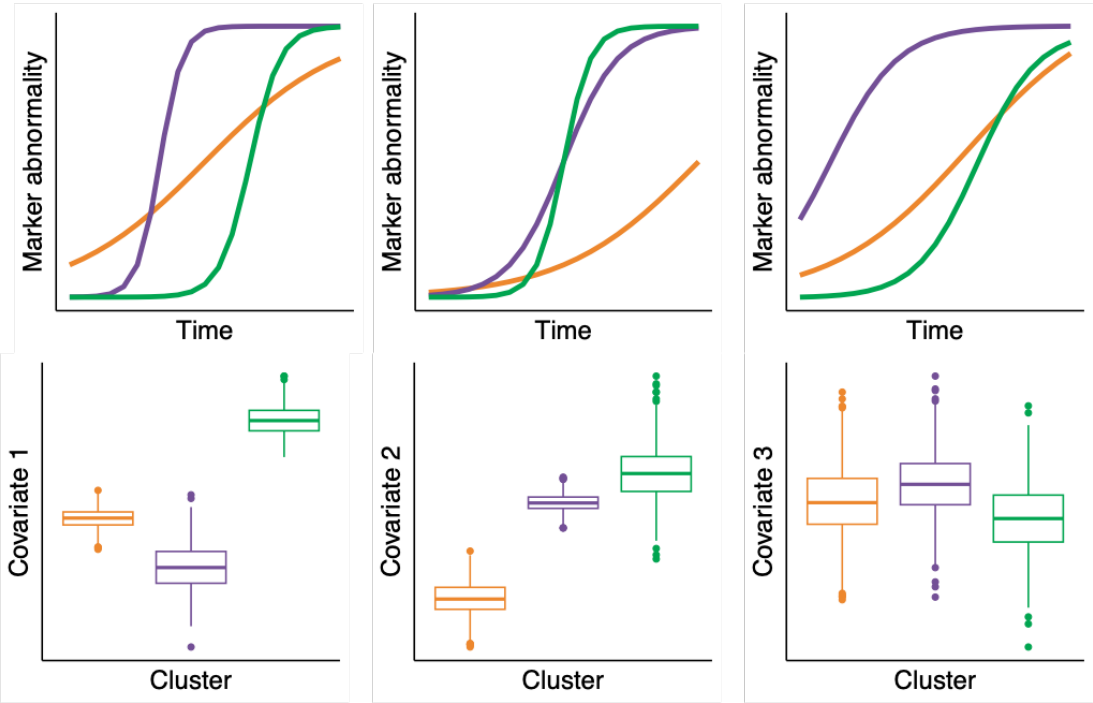
Abstract

The identification of sets of co-regulated genes that share a common function is a key question of modern genomics. Bayesian profile regression is a semi-supervised mixture modelling approach that makes use of a response to guide inference toward relevant clusterings. Previous applications of profile regression have considered univariate continuous, categorical, and count outcomes. In this work, we extend Bayesian profile regression to cases where the outcome is longitudinal (or multivariate continuous) and provide PReMiuMlongi, an updated version of PReMiuM, the R package for profile regression. We consider multivariate normal and Gaussian process regression response models and provide proof of principle applications to four simulation studies. The model is applied on budding-yeast data to identify groups of genes co-regulated during the *Saccharomyces cerevisiae* cell cycle. We identify four distinct groups of genes associated with specific patterns of gene expression trajectories, along with the bound transcriptional factors, likely involved in their co-regulation process.

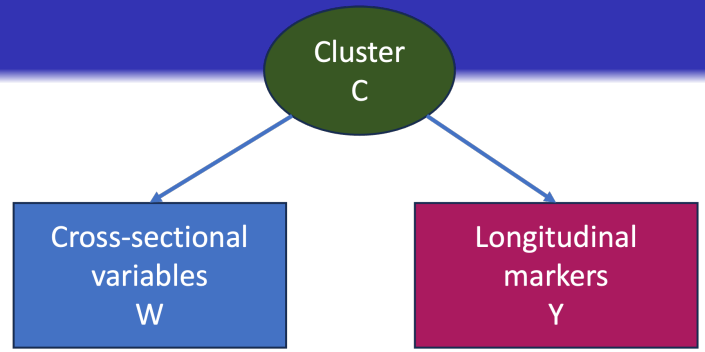
Keywords: Bayesian clustering, longitudinal outcome, PReMiuMlongi, profile regression, variable selection

Extension to multiple longitudinal markers

Clustering objective



Profile regression



Notations:

c_i cluster allocation variable of subject i
W_i individual vector of cross-sectional variables
Y_i vector of individual longitudinal measurements

Likelihood:

$$L(\theta|Y,W) = \prod_{i=1}^{N} \sum_{g=1}^{+\infty} P(c_i = g|\theta^c)\ f(W_i|c_i = g;\theta_g^W)\ \underbrace{f(Y_i|c_i = g;\theta_g^Y)}_{\text{longitudinal data}}$$

Linear mixed model for longitudinal marker m given cluster g:

$$Y_{im}(t) = \beta_{mg}\ X_{im}(t) + \alpha_{img}\ Z_{im}(t) + \epsilon_{im}(t)$$

with Gaussian random effects $\alpha_{img} \sim \mathcal{N}(0, B_{mg})$ and errors $\epsilon_{im}(t) \sim \mathcal{N}(0, \sigma_m^2)$.

Variable selection of longitudinal markers

Cluster-specific distribution of longitudinal marker m in cluster g:

$$Y_{im}(t) = \mu_{mg}(t) + \text{dev}_{igm}(t) + \epsilon_{im}(t)$$

Selection weight for marker m:

$$\zeta_m \in [0, 1]$$

Composite mean parameter in the likelihood:

$$\mu^*_{mg}(t) = \zeta_m \ \mu_{mg}(t) + (1 - \zeta_m) \times \mu_{0m}(t) \quad \forall t$$

with $\mu_{0m}(t)$ the estimated trajectory of marker m in the population.

Application to the MEMENTO cohort (Dufouil et al., 2017)

MEMENTO cohort: French clinic-based cohort of participants with isolated cognitive complaints or mild cognitive impairment

Objective: Identify Alzheimer's Disease sub-phenotypes associated with

- specific evolution patterns of cognitive decline and brain atrophy
- specific neuropathological profiles.

Application to the MEMENTO cohort (Dufouil et al., 2017)

MEMENTO cohort: French clinic-based cohort of participants with isolated cognitive complaints or mild cognitive impairment

Objective: Identify Alzheimer's Disease sub-phenotypes associated with

- specific evolution patterns of cognitive decline and brain atrophy
- specific neuropathological profiles.

- Cognition: Semantic verbal fluency (Animals), Executive functions (TMT-A), Episodic memory (FCSRT)
- Brain atrophy: middle temporal, entorhinal, fusiform volumes and hippocampal volume (relative to intracranial volume), glucose metabolism (SUVR)
- Small Vessel Disease: White matter hyperintensities (WMH)
- Neuropathology: Amyloid-β42 peptide, total tau (t-tau), and phosphorylated tau (p-tau181) from CSF, standardized at baseline

Description and model specification

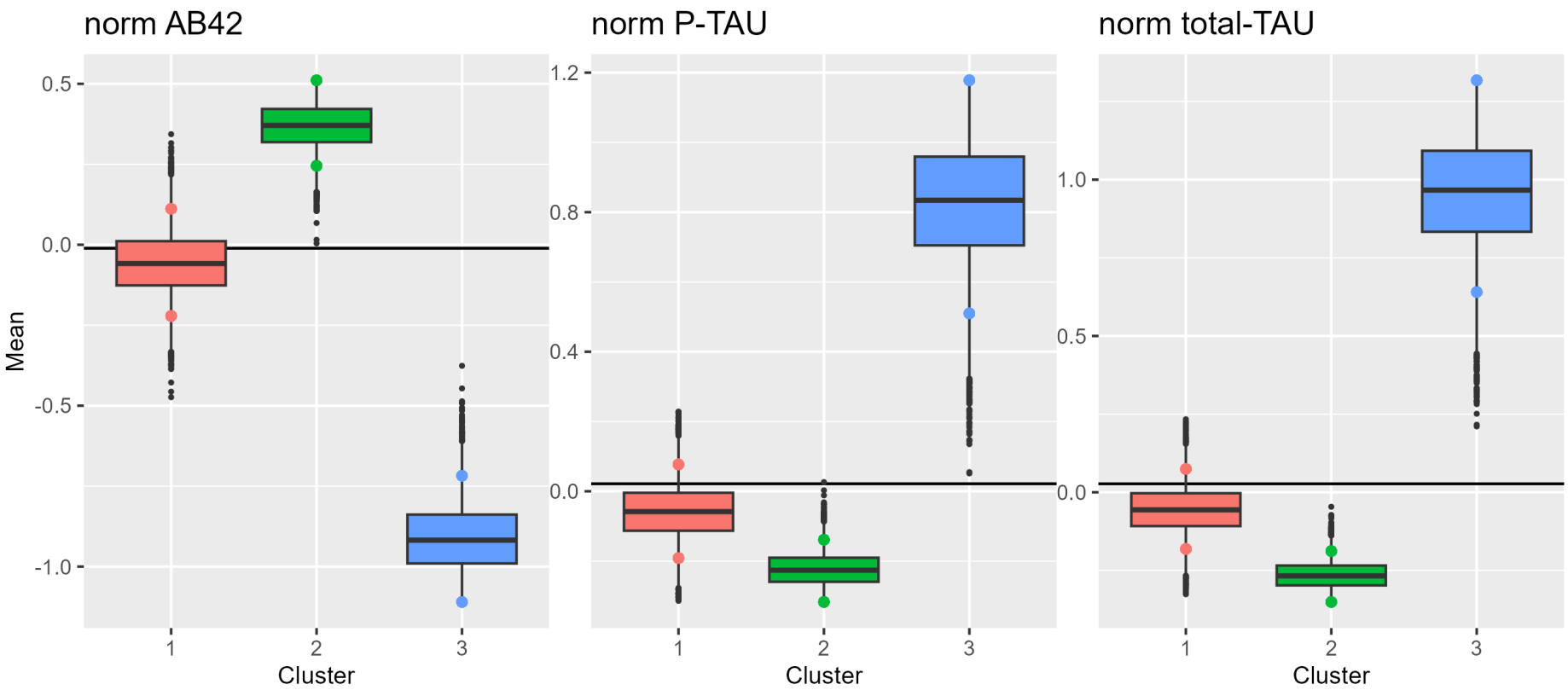
Study sample

- 293 individuals aged 68.5 years old at baseline on average (sd=8.3), 47% male, 61% high education, 40% Apoe carriers
- Mean number of repeated measurements per individual: 4.9 for cognition, 1.7 for brain imaging and 1.7 for WMH

Linear mixed effects models for longitudinal markers

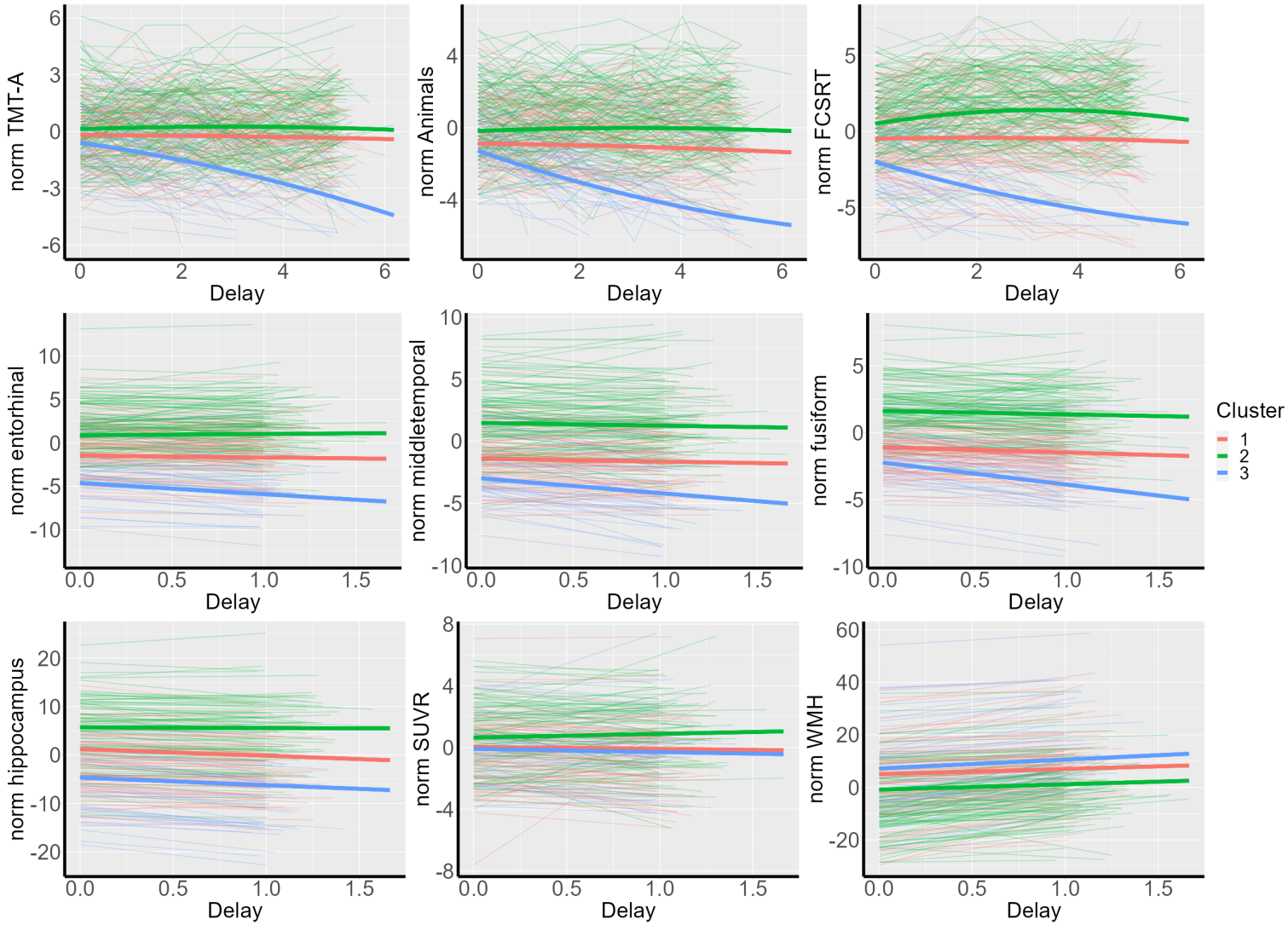
- Repeated observations normalized using spline transformations
- Quadratic time trends for cognitive tests (3 random effects of time)
- Linear time trends for imaging markers (2 random effects of time)
- Adjustment for baseline age, Male, High education, Apoe $\epsilon 4$ status

Cluster-specific neuropathological profiles



Cluster sizes : 109, 133, 51.

Cluster-specific longitudinal evolutions



Discussion

Conclusion

- Mixture model with unconstrained number of clusters
- Integration of both multivariate longitudinal and cross-sectional data, with variable selection
- Extension of the R package **PReMiuMlongi** (Rouanet et al., 2023)
  `https://github.com/premium-profile-regression/PReMiuMlongi`

Future application

- Select the key longitudinal and cross-sectional biomarkers for dementia sub-phenotyping

Methodological perspectives

- Introduce a latent biological timescale
- Accommodate various types of markers

References

Rouanet A, Richardson S, Tom B. Benefit of Bayesian clustering of longitudinal data: study of cognitive decline for precision medicine. In Bayesian Methods in Pharmaceutical Research (eds: Lesaffre E, Baio G, and Boulanger B), Chapman & Hall/CRC Biostatistics Series, forthcoming, ISBN: 978-1-138-74848-4.

Rouanet A, Johnson R, Strauss M, Richardson S, Tom B, White S, Kirk P (2024). Bayesian profile regression for clustering analysis involving a longitudinal response and explanatory variables, Journal of the Royal Statistical Society Series C: Applied Statistics, 73, 314–339.

Molitor J, Papathomas M, Jerrett M, Richardson S (2010). Bayesian Profile Regression with an Application to the National Survey of Children's Health. Biostatistics, 11(3), 484-498.

Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S (2015). PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. Journal of Statistical Software, 64(7),1-30.

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005). Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers and Dementia, 1(1):55-66.

Neal RM (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics, 9(2), 249.