Bayesian variable selection for structured variables: applications in genetics

Marie Denis

in collaboration with B. Heuclin, F. Mortier, M. Tadesse, S. Tisné

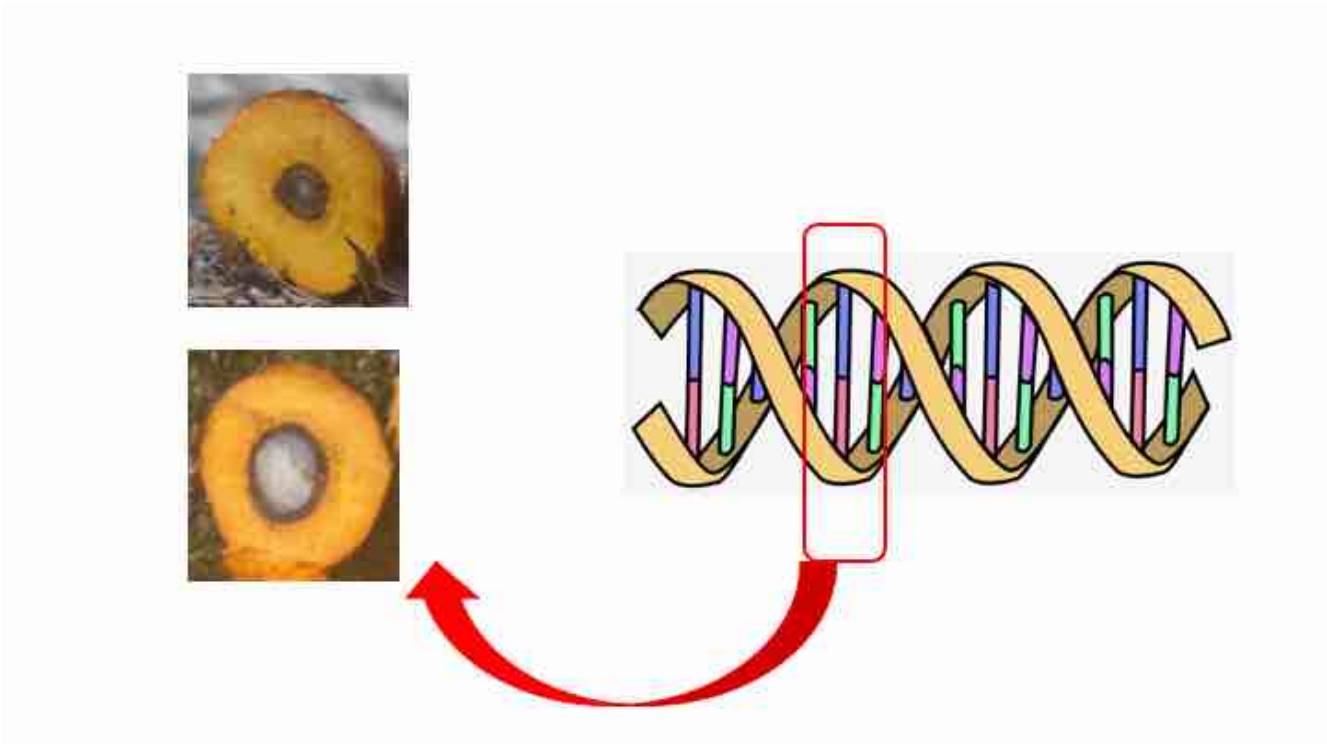
July 5th, 2023

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

Outline

Biological context

In genetics, one of the most important objective is to identify the genomic regions involved in the variability of a phenotype



Which genomic regions may explain/is related to the differences observed in oil palm fruits?

Biological context

With the advance of high-throughput genotyping and phenotyping technologies:

- ↬ **High-dimensional data**: bring opportunities for gaining insights into complex biological mechanisms involved in processes of interest
- ↬ **High-dimensional data**: bring statistical challenges

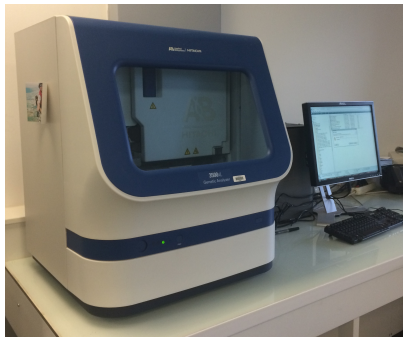


Figure 1: Sequencer Applied 3500 XL (UMR AGAP, Cirad)

Statistical challenges

↬ **A high number** of variables p measured on **a limited number** of individuals n: $p > n$.

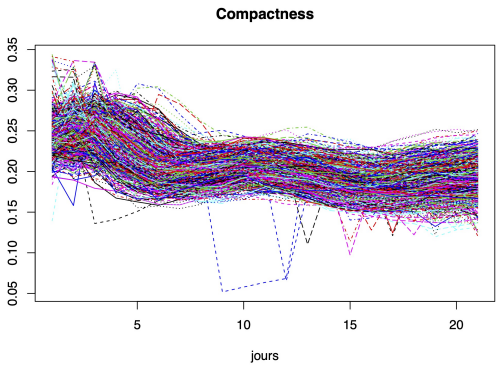
To identify the genomic regions involved in the variability of a phenotype Y

↬ $n = 100$ individuals genotyped with $p = 5,000$ Single Nucleotide Polymorphisms (SNPs) (variables, X)

How to select the most relevant subset of variables out of a large set of variables ?

Statistical challenges

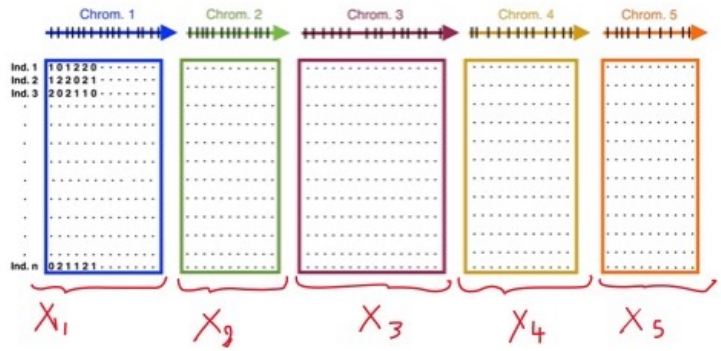
↬ Dependence structure between observations in **the response variable** (longitudinal data)



Y: Compactness over time for each plant in *Arabidopsis thaliana*

Biological context

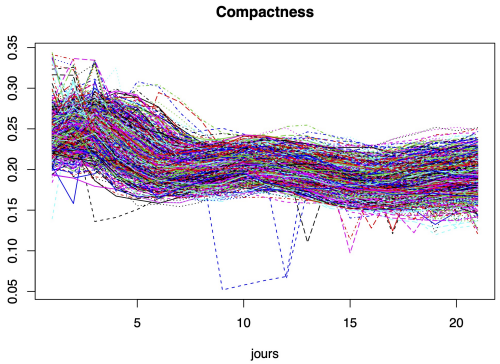
↬ Dependence structure between observations in the response variable (longitudinal data) or between **explanatory variables** that may be induced by various factors

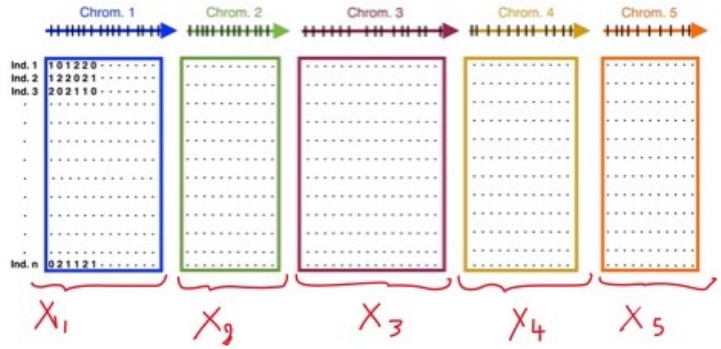


X: Genotype matrix with $X_i \perp X_j$ $(i, j = 1, \ldots, 5)$ and structure within chromosomes $X_{(i-1)1} \not\perp X_{i1}$

Biological context

↬ Dependence structure between observations in **the response variable** or between **explanatory variables** that may be induced by various factors





Y: Compactness over time for each plant in *Arabidopsis thaliana*

X: Genotype matrix with $X_i \perp X_j$ $(i, j = 1, \ldots, 5)$ and structure within chromosomes $X_{i1} \not\perp X_{(i-1)1}$ $(i = 1, \ldots, 5)$

How to analyze all time points simultaneously ? How to integrate dependence structure into variable selection approaches?

Statistical challenges

↬ Need to use appropriate statistical methods dealing with **a high number** of covariates p while considering **dependence structure** between variables

Objectives

To present two developed Bayesian variable selection approaches:

1. for identifying genomic regions related to an outcome measured over time (functional mapping analysis): Bayesian Varying Coefficient model using Group Spike-and-Slab prior (Heuclin et al., 2021)
2. for identifying genomic regions related to an outcome while considering dependence structure between SNPs (QTL mapping analysis): Graph-structured Bayesian variable selection (Denis et al., 2023 (under revision))

Next slides

- Why the "classical" approaches are not appropriate ?
- Why use the Bayesian framework ?

Linear model context

A linear model is a statistical model assuming that the response variable Y may be written as a linear combination of variables X:

$$Y = \mu + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon = \mu + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 Id_n)$$

with

- $Y = (y_1, \ldots, y_n)'$ the n-vector of outcomes,
- X the $n \times p$ matrix of predictors which may be structured and/or of high dimension,
- $\beta = (\beta_1, \ldots, \beta_p)'$ the p-vector of coefficients, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ the n-vector of residuals, σ^2 the residual variance.

↬ To estimate parameters β and σ^2

In genetics:

$$Y = \mu + SNP_1\beta_1 + \cdots + SNP_p\beta_p + \varepsilon$$

with β marker effects

Classical approaches

By analyzing one SNP by one SNP or a subset of SNPs:

- Student test with multiple testing and adjustment for multiplicity,
- Backward stepwise selection, Forward stepwise selection,
- Comparison of all models by using criteria (R^2, AIC, BIC, cross-validation, Fisher test).

But

- Approaches not optimal (consider that genetic information is carried by one genetic variant)
- Approaches not feasible when $p > n$

Classical approaches

By analyzing all SNPs simultaneously

Ordinary Least Square (OLS) regression

To minimize the loss function $L^{OLS}(\beta) = \| Y - X\beta \|^2$:

$$\hat{\beta}^{OLS} = (X'X)^{-1}(X'Y)$$

But

- When the **number of predictors** is high: $\hat{\beta}^{OLS}$ does not perform well in unseen datasets (overfitting), does not provide parsimonious models (Hadamard, 1902) and in very high dimension $(X'X)^{-1}$ not invertible
- In presence of **structures** between predictors (as collinearity): $(X'X)^{-1}$ close to singularity and so, $\hat{\beta}^{OLS}$ not accurate

$\rightsquigarrow$ Need to use **regularization methods**

Regularization methods

Consist in introducing additional information into the problem:

- By imposing constraints as in ANOVA,
- By adding a penalty term to the minimization of the loss function as in **penalized regressions** (Ridge (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996),..),
- By specifying a dependence structure for effects of variables,
- ...

Bayesian approach: a natural framework

Bayesian approaches

In Bayesian framework additional information integrating into models via prior distributions

$\rightsquigarrow$ Regularization is done by specifying **specific priors**

Selection

- To shrink towards zero small coefficients while leaving large signals large: **Shrinkage** priors

$\rightsquigarrow$ SNPs with small/medium/large effects

Structure

- Priors with a **variance-covariance matrix** related to structure information between variables

$\rightsquigarrow$ SNPs structured by genes/chromosomes

Bayesian linear model

In the Bayesian context **prior** distributions are placed on the parameters (here: μ, β, σ^2)

Bayesian linear model

$$
\begin{aligned}
Y|\mu, \beta, \sigma^2 &\sim \mathcal{N}_n(\mu + X\beta, \sigma^2 I_n) \\
\beta_j &\sim p_\beta(\beta_j), \quad j = 1, \ldots, p \\
\mu &\sim p_\mu(\mu) \\
\sigma^2 &\sim p_{\sigma^2}(\sigma^2)
\end{aligned}
$$

Usual prior distributions:

- σ^2: Inverse Gamma distribution
- β: Normal distribution (no selection)
- μ: Uniform distribution

Bayesian variable selection

Two classes of shrinkage priors $p_\beta(.)$:

- **Spike-and-slab priors**: Discrete mixture of two distributions (Mitchell and Beauchamp, 1988; George and McCulloch, 1997)
- **Continuous shrinkage priors**: Unimodal continuous distributions (Bayesian Lasso prior, Horseshoe prior, Elastic-Net prior, ...) (Kyung et al., 2010; Carvalho et al., 2008)

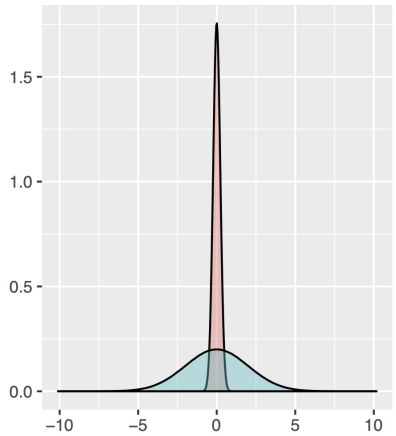


Figure 2: Spike-and-Slab prior distribution. Slab part in blue and spike part in red

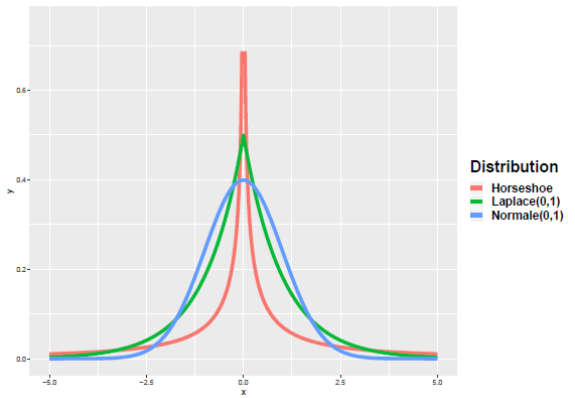


Figure 3: Continuous shrinkage prior distributions

↬ SNPs with small/medium/large effects

Integration of the dependence structure into Bayesian models

$$Y = X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon,\ \varepsilon \sim \mathcal{N}_n(0, \sigma^2)$$

- X the $n \times p$ matrix of predictors which may be structured and/or of high dimension,

$\rightsquigarrow$ $\beta = (\beta_1, \ldots, \beta_p)' \sim \mathcal{N}_p(0, \Sigma)$ with $\mathbf{\Sigma}$ related to structure between variables

$\rightsquigarrow$ **Context dependent**

Examples

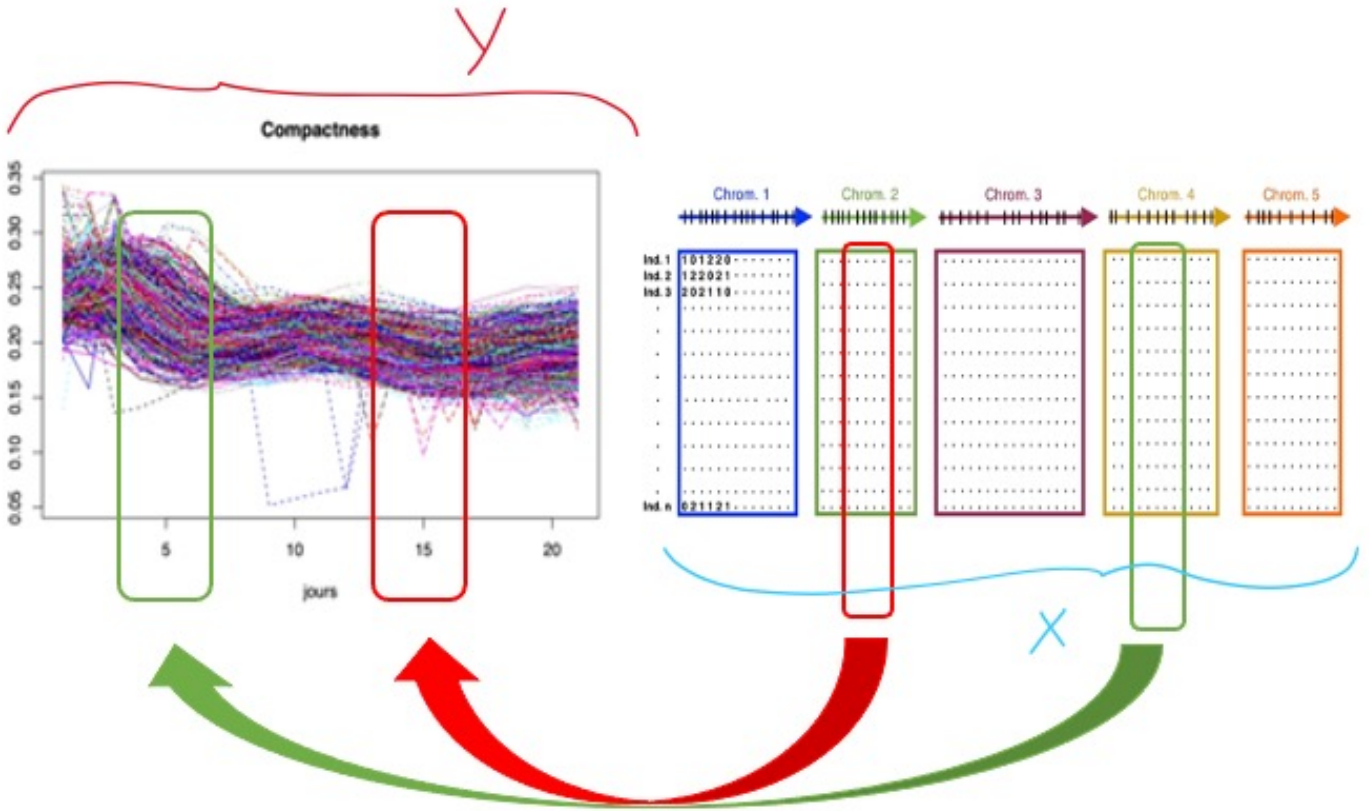
- $X_{t-1} \not\perp\!\!\!\perp X_t$: $\Sigma = AR(\rho)$ with ρ autoregressive parameter
- X_i's belong to a same group (pathways in genomic, genes in genetic, ...)

Outline

Biological motivation: functional mapping

Objectives:

To select the genetic markers involved in the variation of the compactness of *Arabidopsis thaliana* over time and to estimate their functional effects



Statistical motivations

- To select the relevant variables and to estimate their effects over time by analyzing simultaneously all time points and considering temporal dependence between successive measures

↬ Bayesian Varying Coefficient model using Group Spike-and-Slab prior (Heuclin et al., 2021)

Varying coefficient model

We assume that the response variables is followed over T times such that for the individual i we have: $y_i = (y_i^{t_1}, \ldots, y_i^{t^T})'$.

Linear model:

$$y_i^{t_1} = \mu^{t_1} + (\beta_1^{t_1}, \ldots, \beta_q^{t_1}) \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \varepsilon_i^{t_1}, \quad \varepsilon_i^{t_1} \sim N(0, \sigma^2)$$

Varying coefficient model

We assume that the response variables is followed over T times such that for the individual i we have: $y_i = (y_i^{t_1}, \ldots, y_i^{t^T})'$.

Linear model:

$$\begin{array}{lllllllll} y_i^{t_1} & = & \mu^{t_1} & + & (\beta_1^{t_1}, & \ldots, & \beta_q^{t_1}) & + & \varepsilon_i^{t_1}, \quad \varepsilon_i^{t_1} \sim N(0, \sigma^2) \\ y_i^{t_2} & = & \mu^{t_2} & + & (\beta_1^{t_2}, & \ldots, & \beta_q^{t_2}) \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} & + & \varepsilon_i^{t_2}, \quad \varepsilon_i^{t_2} \sim N(0, \sigma^2) \end{array}$$

Varying coefficient model

We assume that the response variables is followed over T times such that for the individual i we have: $y_i = (y_i^{t_1}, \ldots, y_i^{t^T})'$.

Linear model:

$$\begin{array}{rcccccccl}
y_i^{t_1} & = & \mu^{t_1} & + & (\beta_1^{t_1}, & \ldots, & \beta_q^{t_1}) & + \varepsilon_i^{t_1}, & \varepsilon_i^{t_1} \sim N(0, \sigma^2) \\
y_i^{t_2} & = & \mu^{t_2} & + & (\beta_1^{t_2}, & \ldots, & \beta_q^{t_2}) & + \varepsilon_i^{t_2}, & \varepsilon_i^{t_2} \sim N(0, \sigma^2) \\
\vdots & & \vdots & & \vdots & & \vdots & \vdots & \\
y_i^{t_T} & = & \mu^{t_T} & + & (\beta_1^{t_T}, & \ldots, & \beta_q^{t_T}) & + \varepsilon_i^{t_T}, & \varepsilon_i^{t_T} \sim N(0, \sigma^2)
\end{array} \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix}$$

- Simple analysis at each time point does not take into account the correlations over time
  - ↪ Can lead to false positive detection and loss of statistical power

Varying coefficient model

Varying coefficient model(Hastie and Tibshirani, 1993)

$$\begin{pmatrix} y_i^{t_1} \\ \vdots \\ y_i^{t_T} \end{pmatrix} = \begin{pmatrix} \mu^{t_1} \\ \vdots \\ \mu^{t_T} \end{pmatrix} + \begin{pmatrix} \beta_1^{t_1} & \dots & \beta_q^{t_1} \\ \vdots & & \vdots \\ \beta_1^{t_T} & \dots & \beta_q^{t_T} \end{pmatrix} \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \begin{pmatrix} \varepsilon_i^{t_1} \\ \vdots \\ \varepsilon_i^{t_T} \end{pmatrix}, \quad \begin{array}{l} \varepsilon_i \sim N_T(0, \sigma^2 \Gamma) \\ \Gamma_{i,j} = \rho^{|i-j|} \\ -1 < \rho < 1 \end{array}$$

Varying coefficient model

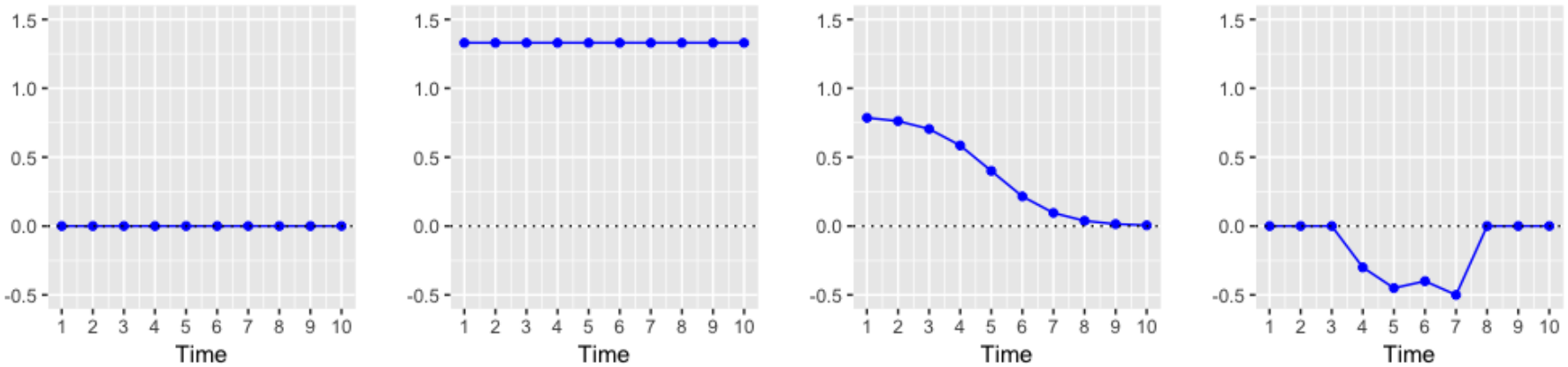
Varying coefficient model (Hastie and Tibshirani, 1993)

$$\begin{pmatrix} y_i^{t_1} \\ \vdots \\ y_i^{t_T} \end{pmatrix} = \begin{pmatrix} \mu^{t_1} \\ \vdots \\ \mu^{t_T} \end{pmatrix} \begin{pmatrix} \beta_1^{t_1} & \dots & \beta_q^{t_1} \\ \vdots & & \vdots \\ \beta_1^{t_T} & \dots & \beta_q^{t_T} \end{pmatrix} \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \begin{pmatrix} \varepsilon_i^{t_1} \\ \vdots \\ \varepsilon_i^{t_T} \end{pmatrix}, \quad \begin{array}{l} \varepsilon_i \sim N_T(0, \sigma^2 \Gamma) \\ \Gamma_{i,j} = \rho^{|i-j|} \\ -1 < \rho < 1 \end{array}$$

Varying coefficient model

Varying coefficient model (Hastie and Tibshirani, 1993)

$$\begin{pmatrix} y_i^{t_1} \\ \vdots \\ y_i^{t_T} \end{pmatrix} = \begin{pmatrix} \mu^{t_1} \\ \vdots \\ \mu^{t_T} \end{pmatrix} \begin{pmatrix} \beta_1^{t_1} & \dots & \beta_q^{t_1} \\ \vdots & & \vdots \\ \beta_1^{t_T} & \dots & \beta_q^{t_T} \end{pmatrix} \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \begin{pmatrix} \varepsilon_i^{t_1} \\ \vdots \\ \varepsilon_i^{t_T} \end{pmatrix}, \quad \begin{array}{l} \varepsilon_i \sim N_T(0, \sigma^2 \Gamma) \\ \Gamma_{i,j} = \rho^{|i-j|} \\ -1 < \rho < 1 \end{array}$$



$(\beta_j^{t_1}, \dots, \beta_j^{t_T})'$ are assumed to be a realization of a function $\beta_j(t)$

↪ Estimation of $\beta_j(t)$ with functional or non functional methods

↪ Selection of significant variables X_j such that $(\beta_j^{t_1}, \dots, \beta_j^{t_T})' = (0, \dots, 0)'$ with a spike-and-slab prior

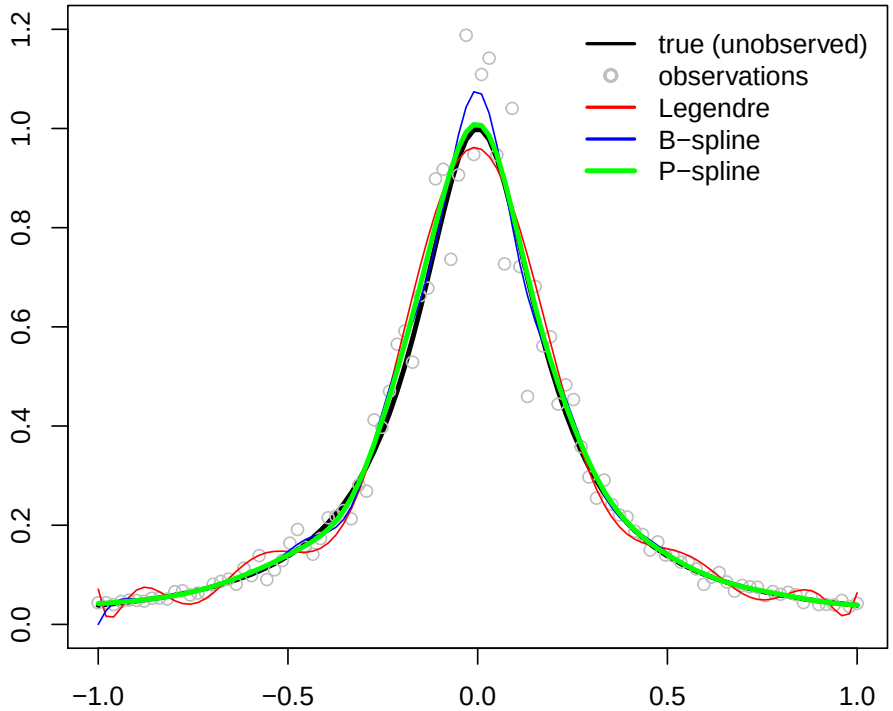
Functional method: Non-parametric interpolation

- Bayesian P-spline interpolation (Eilers and Marx, 1996; Lang and Brezger, 2004) for inducing smoothness

$$\begin{pmatrix} \beta_j^{t_1} \\ \vdots \\ \beta_j^{t_T} \end{pmatrix} = \sum_{k=1}^{v} B_k b_{k,j} = B\ b_j,$$

$$b_j \sim \mathcal{N}(0, (\lambda_j K)^{-1})$$

with K a structured matrix.



Selection in varying coefficient model

Selection of relevant variables X_j, $j = 1, \ldots, p$:

$$(b_{1,j}, \ldots, b_{v,j})' = (0, \ldots, 0)'?$$

↬ Group spike-and-slab prior on b_j (Yang and Narisetty, 2020):

$$b_j|\gamma_j, \lambda_j \sim \gamma_j N_v(0, (\lambda_j K)^{-1}) + (1 - \gamma_j)\delta_v(0)$$
$$\lambda_j \sim Gamma(s, r),$$
$$\gamma_j \sim Ber(\pi),$$

Bayesian Varying Coefficient model using Group Spike-and-Slab prior (Heuclin et al., 2021)

Bayesian hierarchical model

$$Y_i|m, b, \rho, \sigma^2 \sim N_T(Bm + BbX_i, \sigma^2\Gamma)$$

$$m|\lambda_0 \sim N_v(0, (\lambda_0 K)^{-1})$$

$$b_j|\gamma_j, \lambda_j \sim \gamma_j N_v(0, (\lambda_j^2 K)^{-1}) + (1-\gamma_j)\delta_v(0), \quad j = 1, \ldots, q$$

$$\lambda_j \sim Gamma(s, r), \quad j = 0, \ldots, q$$

$$\gamma_j \sim Ber(\pi), \quad j = 1, \ldots, q$$

$$\rho \sim U_{[-1,1]}$$

$$\sigma^2 \sim I - Gamma(s_{\sigma^2}, r_{\sigma^2})$$

To infer the distribution of $m, b_j, \lambda_j, \gamma_j, \rho, \sigma^2|Y$**:**

↪ Gibbs algorithm (Markov Chain Monte Carlo algorithm)

VCGSS package

VCGSS: an R package for implementing the sparse Bayesian **V**arying **C**oefficient model using **G**roup **S**pike-and-**S**lab prior

↬ Available on https://github.com/Heuclin/VCGSS

We will explore the two main functions:

`VCM_fct()`

To run the Bayesian Varying Coefficient model using Group Spike-and-Slab prior.

- allows to implement functional and non functional methods with penalty of order 1 or 2,
- calls an MCMC sampler implementation in C++,
- allows to run many repetitions in parallel,
- applies convergence diagnostics

`plot_functional_effects()`:

To visualize the dynamic effects

Application on *Arabidopsis thaliana*: functional mapping

Objective:

To select the genetic markers involved in the variation of the compactness of *Arabidopsis thaliana* (data publicly available)

- Individuals: $n = 357$ under well-watered environmental condition
- Markers: SNPs, $p = 532$ to reduce the collinearity between adjacents markers all markers with correlations higher than 0.95 removed $\Rightarrow$ 125 markers
- Phenotypic trait: compactness (Ratio between the projected rosette area and the convex hull area)
- Measurement frequency: daily for $T = 21$

Application on *Arabidopsis thaliana*

Marginal posterior inclusion probabilities

- 14 markers with posterior probability greater than 0.5
- Switch between some markers:

↬ identification of new genomic regions compared Marchadier et al. (2018) (with a single time point analysis) with potentially small effects

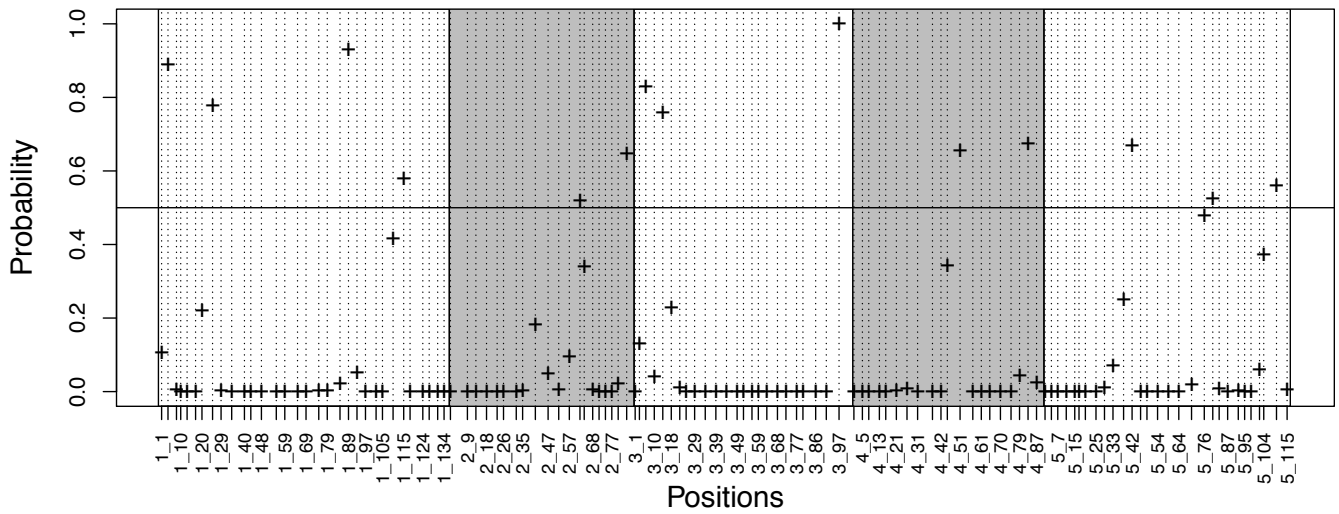
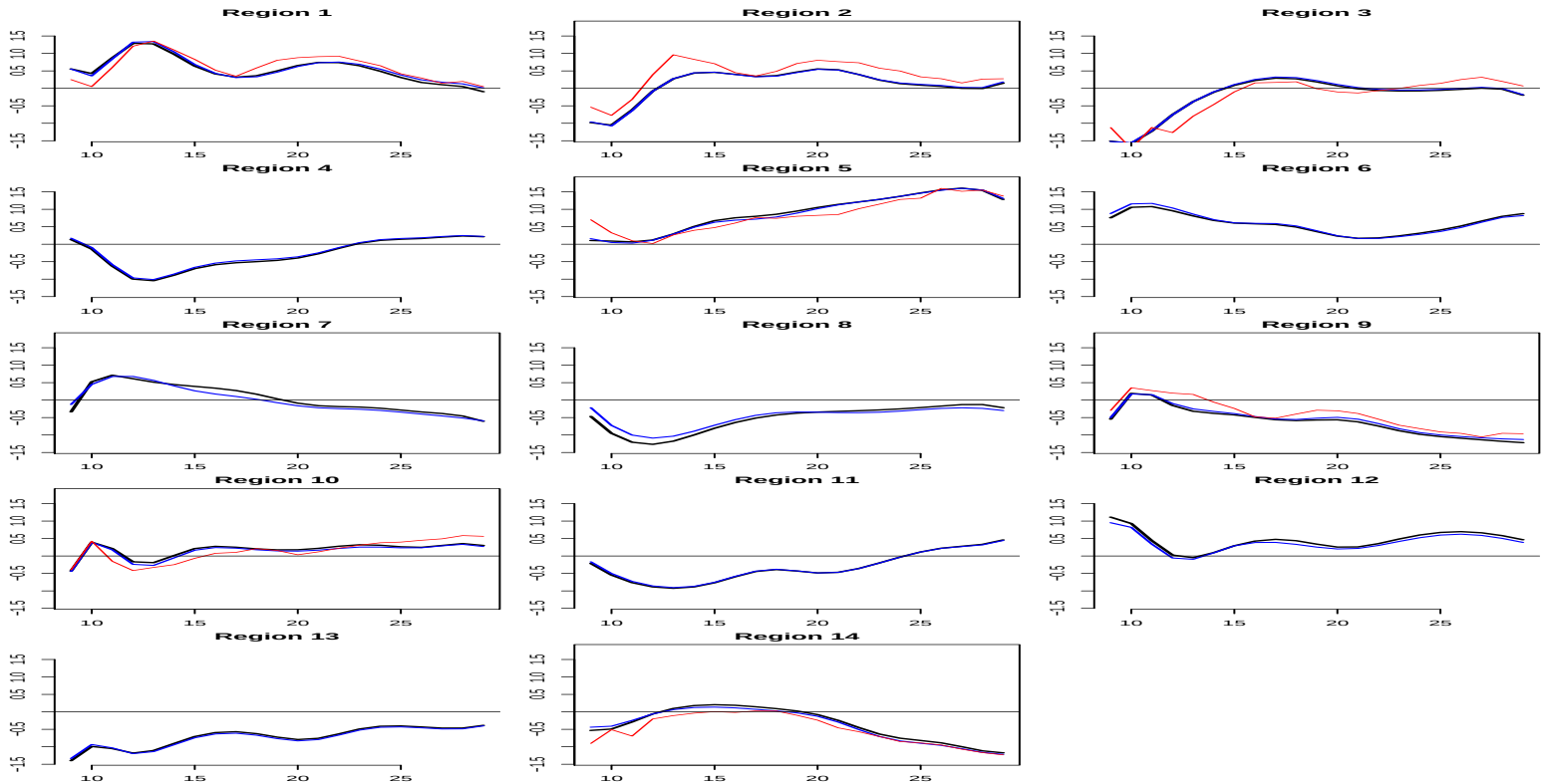


Figure 4: Marginal posterior inclusion probabilities for the 125 markers

Application on *Arabidopsis thaliana*

Estimations of varying effects

Estimation of the effects for markers with the highest marginal posterior probabilities (PS_1: blue, PS_2: black, RW_2: red)



Application on *Arabidopsis thaliana*

Results

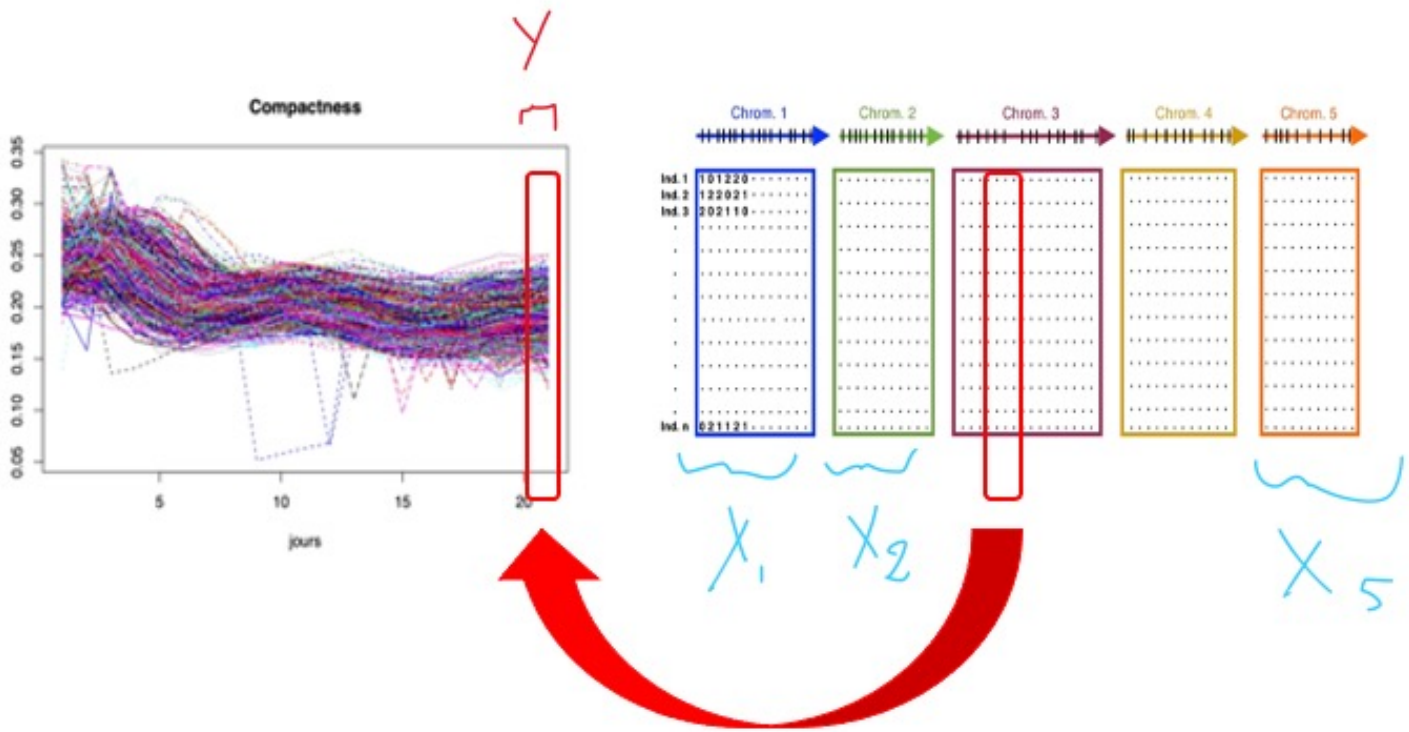
- Approach considering all SNPs and all time points: identification of new regions with potentially small effects and related to known genes,
- Dynamic effects: better biological interpretation for the identification of markers "active/inactive" at different stages of the process,
- Based on simulations: better selection and prediction accuracy.

Outline

Biological motivation: QTL mapping

Objectives:

To select the genetic markers involved in the variation of the compactness of *Arabidopsis thaliana* at one time point while considering the structure along the genome.



Statistical motivations

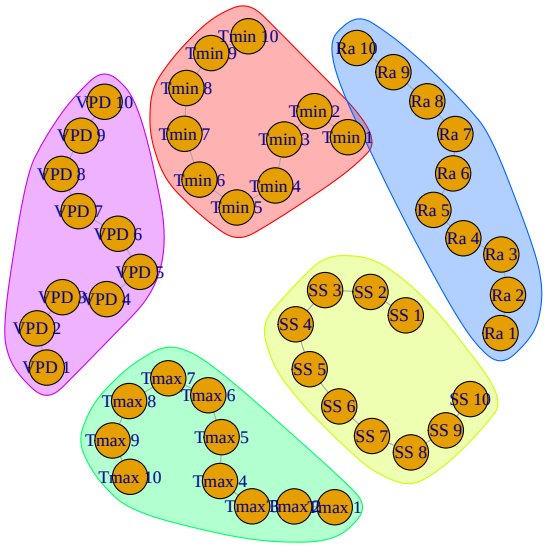
- To select the relevant variables (X_j, $j = 1, \ldots, p$) while considering the dependence structure between them,

↬ Graph-structured Bayesian variable selection (Denis et al., 2023 (under revision))

Graph-structured Bayesian variable selection

Combination of

1. a **continuous shrinkage prior** (horseshoe prior (HS)) for the selection with
2. a **Gaussian Markov random field** (GMRF) for the integration of a dependence structure between explanatory variables characterized by a graph



Bayesian hierarchical model

We assume that $\mathcal{G} = \bigcup_{i=1}^{I} \mathcal{G}_i = \bigcup_{i=1}^{I} (V_i, E_i)$ a disjoint union of I subgraphs and $\mathcal{S}$ the set of indices associated to one representative of each of the I subgraphs.

HS-GMRF model

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathsf{I}_n) \\
\beta|\tau^2, \lambda^2 &\sim \mathcal{N}_p(0, \lambda^2 \mathsf{Q}^{-1}), \\
\tau_{jj'} &\sim \mathcal{C}^+(0,1) \text{ for } (j,j') \in \bigcup_{i=1}^{I} E_i; \tau_j \sim \mathcal{C}^+(0,1) \text{ for } j \in \mathcal{S} \\
\lambda|\sigma &\sim \mathcal{C}^+(0,\sigma);\ \sigma^2 \sim \mathcal{IG}(a_0, b_0)
\end{aligned}
$$

- $s_{jj'} = \text{sign}\{\text{cor}(X_j, X_{j'})\}$: to encourage regression coefficients of negatively correlated variables to take opposite signs,
- Q: the precision matrix integrating the graph information

$$
Q_{jj} = \begin{cases} \frac{1}{\tau_j^2} + \sum_{j' \in \mathcal{N}(j)} s_{jj'} \frac{1}{\tau_{jj'}^2} & \text{if } \ j \in \mathcal{S} \\ \sum_{j' \in \mathcal{N}(j)} s_{jj'} \frac{1}{\tau_{jj'}^2} & \text{otherwise} \end{cases}
$$

Application on *Arabidopsis thaliana*: QTL mapping

Objective:

To select SNPs involved in the variation of the compactness at $T = 21$

- Methods picks contiguous markers: selection of genomic regions
- Overlapping with markers/regions found in Marchadier et al. (2018) and Heuclin et al. (2021) and new ones that are related to genes involved in compactness

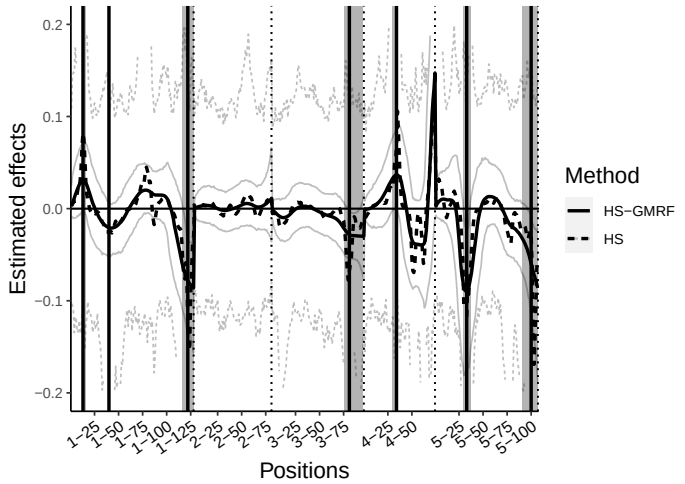


Figure 5: Estimated coefficients along the genome.

Outline

Conclusions/perspectives

↬ Variable selection approaches considering dependence structure between variables give better results and improve biological understanding.

Simulation study (results not shown)

- Increase power to detect associations: encourages the identification of groups of dependent variables acting jointly on the response, especially those with subtle individual effects,
- Improves the predictive power,
- Helps the model building process by reducing the complexity of models and by circumventing the problem of high collinearity through identifiability.

Conclusions/perspectives

Biological interpretation

- Identification of regions versus single molecular marker make more sense for identifying QTLs,
- Identification of markers/regions with small effects (rare allele),
- Better understanding of the dynamic genetic architecture via estimations of marker effects over time.

Perspectives

- More complex structures between variables may be considered (as spectrum, gene networks, ...),
- To combine both approaches for analyzing outcomes measured over time as well as structured covariates.

Thanks for your attention !

marie.denis@cirad.fr

Outline

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2008). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2):89–121.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Hastie, T. and Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796.

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.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Marchadier, E., Hanemian, M., Tisne, S., Bach, L., Bazakos, C., Gilbault, E., Haddadi, P., Virlouvet, L., and Loudet, O. (2018). The complex genetic architecture of shoot growth natural variation in Arabidopsis thaliana.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Yang, X. and Narisetty, N. N. (2020). Consistent group selection with bayesian high dimensional modeling. *Bayesian Analysis*.