Bayesian Varying Coefficient Model with Selection

(Heuclin et al., 2020)

Marie DENIS

with B. Heuclin, F. Mortier and C. Trottier

AISC

October 8 , 2021

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

Biological context

Genetic breeding program

- Improving productivity and nutritive quality
- Minimizing impact on environment
- Adapting crops in the face of climate change

Objectives

- To understand the genetic architecture which controls part of phenotypic variations (Lynch and Walsh, 1998)
- Biological processes are dynamic (Hansen, 2006): to understand the dynamic genetic architecture across the developmental stages

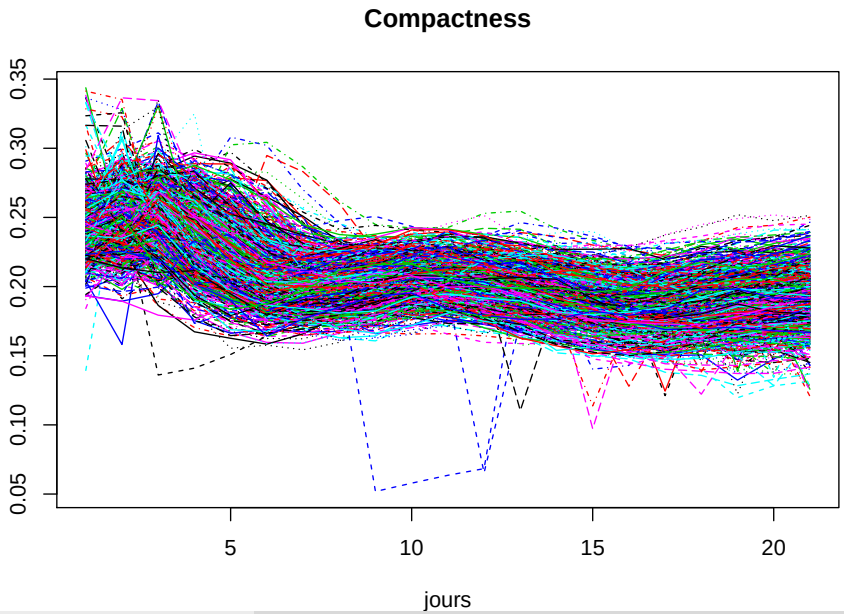
How does the genetic architecture of quantitative traits evolve over time?

Data questions

Data types

1. Phenotypic trait observed on n individuals at T times
2. Individual genotype constant over time, varying between individuals
3. Environmental conditions common to individuals, varying over time

Example in *Arabidopsis thaliana* for compactness (the ratio between the projected rosette area and the convex hull area) (Marchadier et al., 2018):



Statistical questions

Statistical challenges

- Strong correlation between successive measurements
  $\hookrightarrow$ Model such dependencies
- Genetic effects vary over time (season, age,...)
  $\hookrightarrow$ Model such time varying effects (Hastie and Tibshirani, 1993)
- Large number of genetic information
  $\hookrightarrow$ Select relevant markers
- Phenotypic variations across environmental conditions
  $\hookrightarrow$ Model smooth environmental effects (Hastie and Tibshirani, 1986)

Statistical model

- y_{it_k}, phenotype of individual $i = 1, \ldots, n$ at time t_k $(k = 1, \ldots, T)$
- $e^l = \left(e^l_{t_1}, \ldots, e^l_{t_k}, \ldots, e^l_{t_T}\right)'$, environmental conditions varying over time, $l = 1, \ldots, L$, L is low dimensional
- x_{ij}, molecular marker $j = 1, \ldots, J$ for individual i, J is large

$$y_{it_k} = \alpha + \mu(t_k) + \sum_{l=1}^{L} f_l(e^l_{t_k}) + \sum_{j=1}^{J} x_{ij}\beta_j(t_k) + \varepsilon_{it_k}. \tag{1}$$

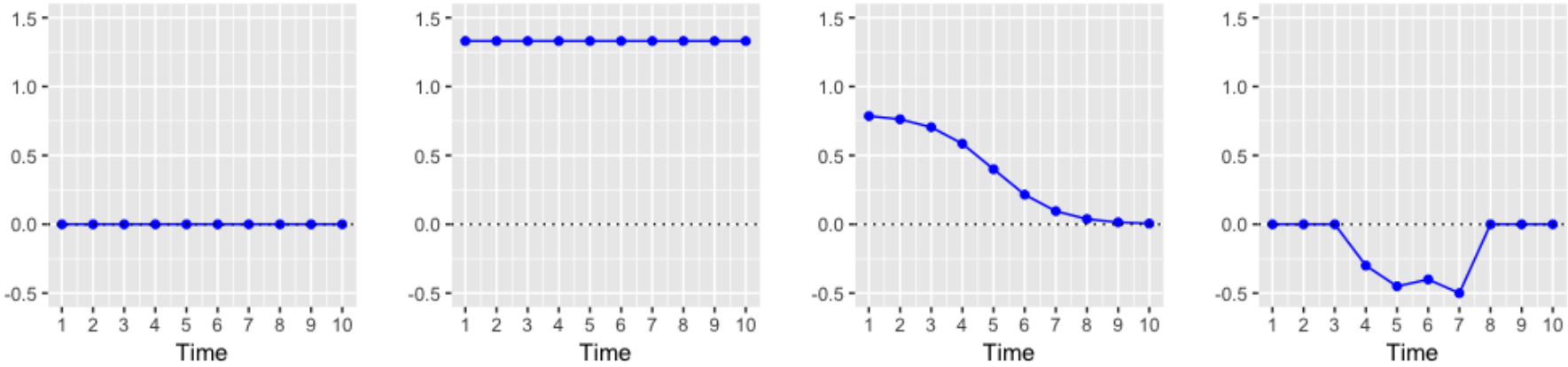
α the intercept

μ, f_l, and β_j real smooth functions of time or of environmental variables

$\varepsilon_i \sim \mathcal{N}_{t_T}(0, \sigma^2\Gamma)$, Γ a $T \times T$ correlation matrix defined by an AR(1) process

Statistical model

Estimation of the dynamic effects



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

↪ Estimation of $\mu(t)$, $f_l(e_l)$, $\beta_j(t)$ with functional or non functional methods

↪ Selection of significant variables X_j such that $(\beta_j^{t_1}, \ldots, \beta_j^{t_T})' = (0, \ldots, 0)'$

Functional approach: P-splines (Lang and Brezger, 2004; Eilers and Marx, 1996) I

B-spline approach approximates a real function h as a linear combination of ν^{th}-degree basis functions defined on K knots ($K-1$ intervals):

$$h(x) = \sum_{r=1}^{df} B_r(x, \nu)\ c_r$$

Model 1 can then be written as:

$$y_i = \alpha 1 + \widetilde{B^t}\widetilde{m} + \sum_{l=1}^{L} \widetilde{B^{e^l}}\widetilde{a}_l + \sum_{j=1}^{J} x_{ij} Z b_j + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \Gamma) \tag{2}$$

with $\widetilde{m}$, $\widetilde{a}_l, (l = 1, \ldots, L)$, $b_j\ (j = 1, \ldots, J)$ vectors of B-spline coefficients

Functional approach: P-splines (Lang and Brezger, 2004; Eilers and Marx, 1996) II

B-spline approach strongly depends on the number of knots and the choice of their positions

- ↪ A misspecification may lead to over- or under-fitting.
- ↪ Penalized B-splines (P-splines) induce smoothness,
- ↪ Penalize the first- or second-order finite differences in adjacent spline regression coefficients

Functional approach: P-splines (Lang and Brezger, 2004; Eilers and Marx, 1996) III

In a Bayesian framework

Specific prior distributions on parameters $\tilde{m}$, $\tilde{a}$ and b as a multivariate first or second order random walk prior:

$$\mathcal{N}\left(0, \tau_u^2 (K)^{-1}\right)$$

- τ_u^2 a variance parameter specific to each group of unknown parameters
- $K = D'D$ with D the matrix representation of the first or second order finite differentiating operator

Functional versus non functional methods

Functional method: P-spline interpolation

$$Z = B' \Rightarrow \sum_{j=1}^{J} x_{ij} Z b_j = \sum_{j=1}^{J} x_{ij} B' b_j$$

with first or second order difference penalty
$\hookrightarrow$ PS_1 and PS_2

Non functional method: direct estimation of time coefficient functions Li and Sillanpää (2013)

$$Z \equiv Id \Rightarrow \sum_{j=1}^{J} x_{ij} Z b_j = \sum_{j=1}^{J} x_{ij} b_j \ (b_j = \beta_j)$$

with first or second order difference penalty
$\hookrightarrow$ RW_1 and RW_2

In both methods we assume that:

$$b_j \sim \mathcal{N}\left(0, \tau_u^2 (K)^{-1}\right)$$

.

Selection of molecular markers I

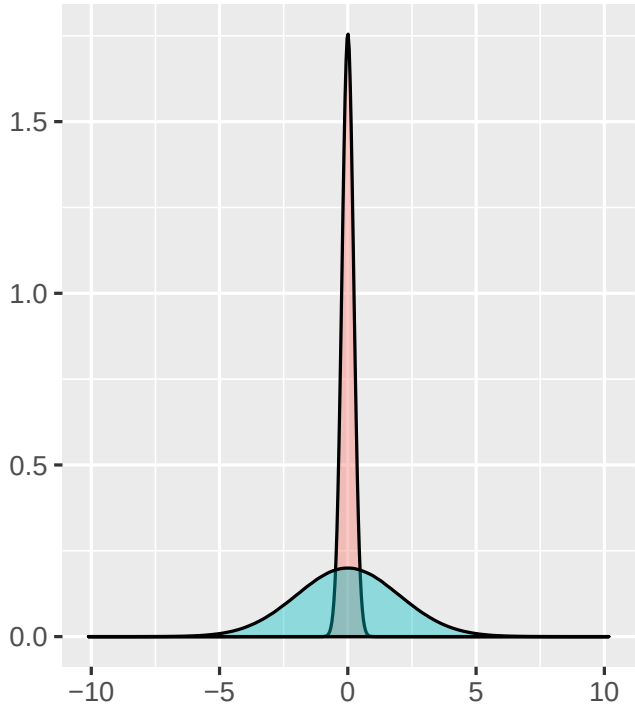
Spike-and-Slab Prior (George and McCulloch, 1993, 1997)

- Let γ such that

$$\gamma_j = \begin{cases} 1 \text{ if } j^{\text{th}} \text{ variable relevant} \\ 0 \text{ otherwise} \end{cases}$$

$$\beta_j|(\gamma_j = 1) \sim p_{Slab}(b_j) \quad \beta_j|(\gamma_j = 0) \sim p_{Spike}(b_j)$$

$$\gamma_j|\pi \sim \mathcal{B}ernoulli(\pi)$$
$$\pi \sim \mathcal{B}eta(1, 1)$$



- Posterior distribution:

$$\mathbb{P}(\gamma_j = 1|Y)$$

Selection of molecular markers II

$$y_i = \alpha 1 + \widetilde{B^t}\widetilde{m} + \sum_{l=1}^{L} \widetilde{B^{e^l}}\widetilde{a}_l + \sum_{j=1}^{J} x_{ij} Z b_j + \varepsilon_i, \;\; \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \Gamma)$$

Here b_j is a vector, not a scalar

Group spike-and-slab prior (Ghosh and Ghattas, 2015; Yang and Narisetty, 2020)

$$b_j | \gamma_j, \lambda_j^2, \sigma^2 \sim \gamma_j \, \mathcal{N}_v(0, \sigma^2 \tau_j^2 (D'D)^{-1}) + (1 - \gamma_j) \, \delta_v(0)$$

$$\gamma_j | \pi \sim \mathcal{B}ernoulli(\pi), \quad \pi \sim \mathcal{B}eta(1, 1)$$

$$\tau_j^2 \sim Inv - \mathcal{G}amma(s, r)$$

Full hierarchical model

$$
\begin{aligned}
y_i|\alpha, \widetilde{m}, \widetilde{a}, b, \rho, \sigma^2 &\sim \mathcal{N}(\alpha + \widetilde{B^t}\widetilde{m} + \sum_{l=1}^{L} \widetilde{B}^{e^l}\widetilde{a}_l + \sum_{j=1}^{J} x_{ij} Z b_j, \sigma^2 \Gamma) \\
\alpha &\sim \mathcal{U}_{(-\infty,\infty)} \\
\widetilde{m}|\tau_m^2 &\sim \mathcal{N}(0, \tau_m^2(\widetilde{D}_m'\widetilde{D}_m)^{-1}) \\
\widetilde{a}_l|\tau_{a_l}^2 &\sim \mathcal{N}(0, \tau_{a_l}^2(\widetilde{D}_{a_l}'\widetilde{D}_{a_l})^{-1}), \quad l = 1, \ldots, L \\
b_j|\tau_{b_j}^2, \gamma_j, \sigma^2 &\sim \gamma_j \mathcal{N}(0, \sigma^2\tau_{b_j}^2(D'D)^{-1}) + (1-\gamma_j)\delta(0), \quad j = 1, \ldots, J \\
\tau_m^2,\ \tau_{a_l}^2 \text{ and } \tau_{b_j}^2 &\sim \mathcal{IG}(0.1, 0.1), \quad l = 1, \ldots, L \text{ and } j = 1, \ldots, J \\
\gamma_j &\sim \mathcal{B}er(\pi), \quad j = 1, \ldots, J \text{ and } \pi \sim \mathcal{B}eta(1, 1) \\
\rho &\sim \mathcal{U}_{(-1,1)}, \quad \sigma^2 \sim \mathcal{IG}(0.1, 0.1)
\end{aligned} \tag{3}
$$

Simulations

Objectives

- Evaluate impact of parameters on performance
  - Number of observations (time steps and individuals)
  - Residual variance
  - Number of markers and correlation among them
- Compare with other methods
  - Estimation: Legendre polynomials (L) (Li et al., 2015), B-spline (BS), P-splines with first or second order difference penalty or a direct Gaussian process (GP)
  - Selection: Bayesian group Lasso (BGL) (Li et al., 2015) or stepwise(S-GP) (Vanhatalo et al., 2019)

Gibbs sampling used for inference - R code available at `https://github.com/Heuclin/VCGSS`

Simulation parameters

- number of individuals, $n = 100$ or 300
- Number of time steps $T = 100$
- Number of loci, $J = 500$ or 3000 and only 4 markers with significant effects

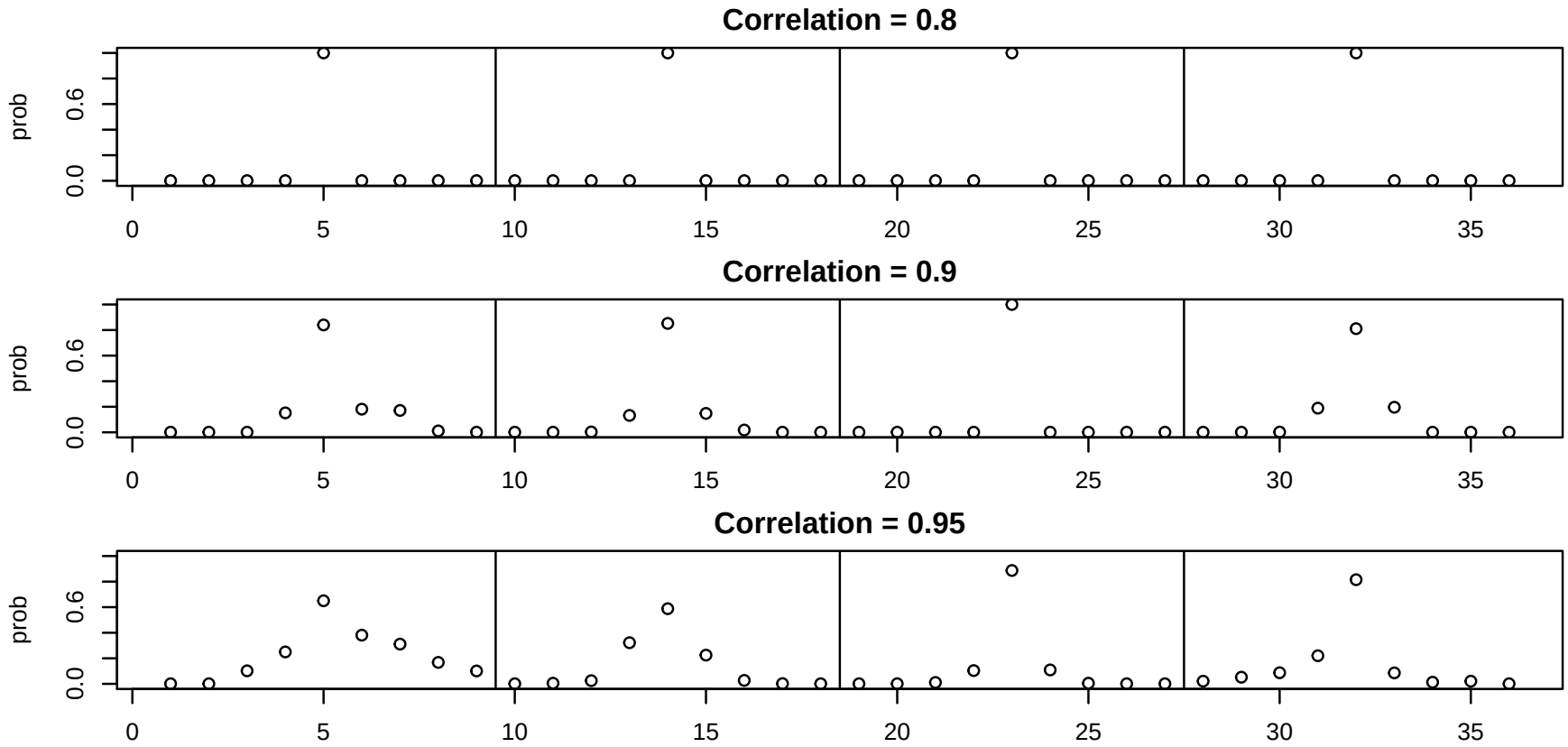
$$\begin{aligned}
\beta_1(t) &= 4 - 0.08t, \\
\beta_2(t) &= \cos\left(\frac{\pi}{15}(t-25)\right) + \frac{t}{50}, \\
\beta_3(t) &= \frac{60}{25 + (t - \frac{T}{2})^2} \\
\beta_4(t) &= 21_{t \leq \frac{T}{3}} + 01_{\frac{2T}{3} < t \leq \frac{2T}{3}} + 1_{t > \frac{2T}{3}}.
\end{aligned}$$

- Correlation between markers equal to $0.8, 0.9$ and 0.95
- Residual variance equal to 4 or 16

Impact of models and priors on variable selection

Criteria	Prior	n=300, J=500, σ^2=4	n=300, J=500, σ^2=16	n=100, J=3000, σ^2=4	n=100, J=3000, σ^2=16
MCC	BGL-PS	0.91 (0.08)	0.9 (0.082)	0.51 (0.041)	0
	BGL-BS	0.99 (0.041)	0.98 (0.046)	0.5 (0)	0
	BGL-L	0.75 (0.099)	0.7 (0.092)	0.5 (0)	0.2 (0.274)
	GSS-L	1 (1)	1 (1)	1 (1)	0.96 (0.962)
	GSS-BS	1 (0)	1 (0)	1 (0)	1 (0.019)
	GSS-PS_1	1 (0)	1 (0)	1 (0)	0.98 (0.044)
	GSS-PS_2	1 (1)	1 (1)	1 (1)	0.94 (0.941)
	GSS-RW_1	1 (0)	0.99 (0.027)	1 (0)	0.87 (0)
	GSS-RW_2	1 (0)	0.99 (0.027)	1 (0)	0.87 (0)
	S-GP	1 (0)	0.89 (0.05)	0.94 (0.063)	0.62 (0.141)
RMSE$_\beta$	BGL-PS	0.47 (0.083)	0.86 (0.17)	3.48 (0.248)	5.62 (0)
	BGL-BS	0.43 (0.042)	0.69 (0.091)	3.54 (0.065)	5.62 (0)
	BGL-L	0.75 (0.187)	1.53 (0.391)	3.56 (0.108)	4.83 (1.077)
	GSS-L	0.43 (0.429)	0.7 (0.695)	0.63 (0.628)	1.22 (1.224)
	GSS-BS	0.42 (0.022)	0.66 (0.042)	0.6 (0.04)	1.03 (0.1)
	GSS-PS_1	0.38 (0.024)	0.61 (0.041)	0.56 (0.04)	0.96 (0.176)
	GSS-PS_2	0.39 (0.39)	0.66 (0.665)	0.58 (0.578)	1.23 (1.234)
	GSS-RW_1	0.43 (0.024)	0.87 (0.106)	0.74 (0.041)	1.79 (0.054)
	GSS-RW_2	0.42 (0.04)	0.89 (0.131)	0.76(0.043)	1.81 (0.057)
	S-GP	0.44 (0.023)	1.05 (0.204)	0.76 (0.276)	2.87 (0.819)

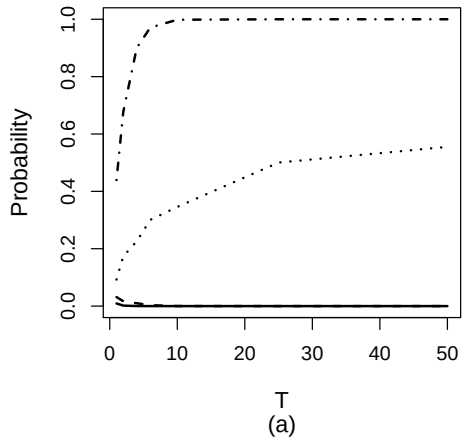
Impact of correlations between loci

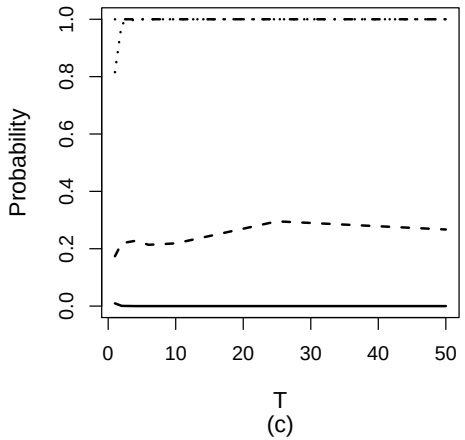


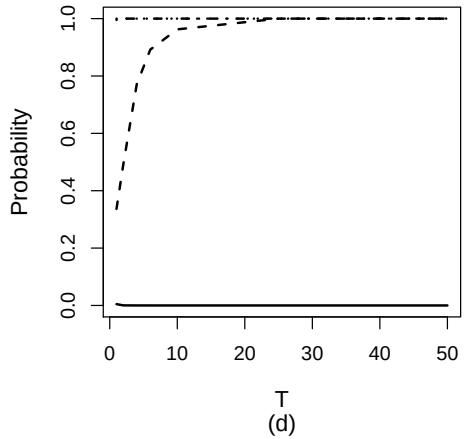
- Marginal probabilities of inclusion for each effect associated to correlated markers within four independent groups

Impact of the number of time steps and observations

- Effects are simulated as constant functions over time equal to 0.3 (Dotted-dashed line), 0.2 (dotted line), 0.1 (dashed line) and 0 (solid line).
- Marginal probabilities of inclusion as a function of T for n=100 (top left), n=500 (top right), n=1000 (bottom left).







Application on Arabidopsis thaliana

Data

- Individuals: n= 358,
- Markers: QTL, q= 125
- Measurement frequency: daily for $T = 21$ days
- Phenotypic trait: Compactness along time

Settings

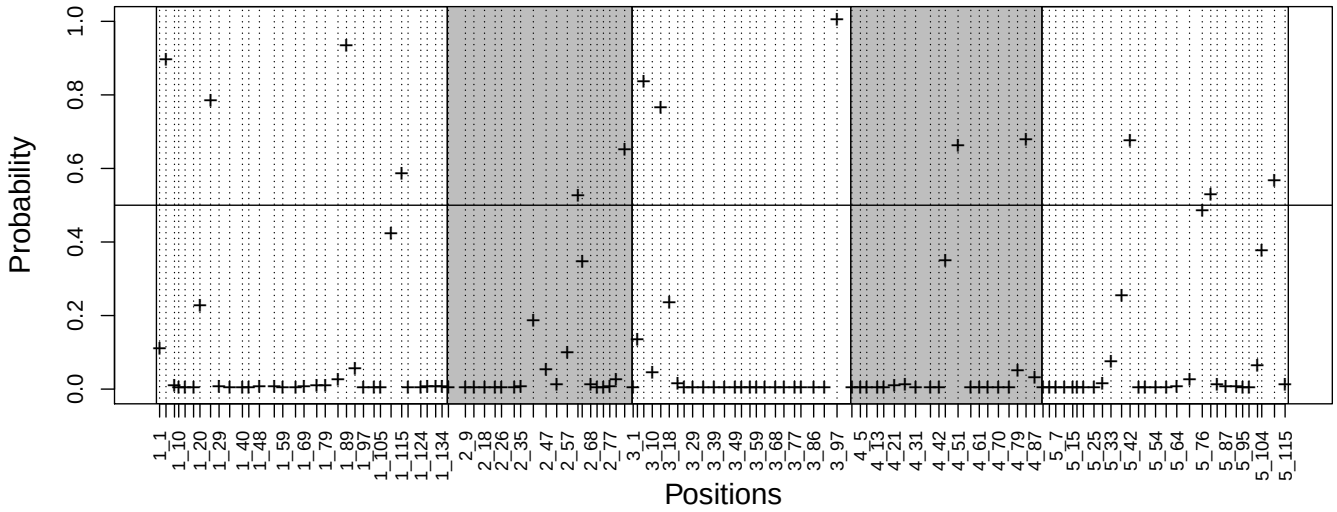
- 100 MCMC chains, 1,000,000 iterations
- Interpolation method: P-spline with second order difference penalty

Application on Arabidopsis thaliana

Results

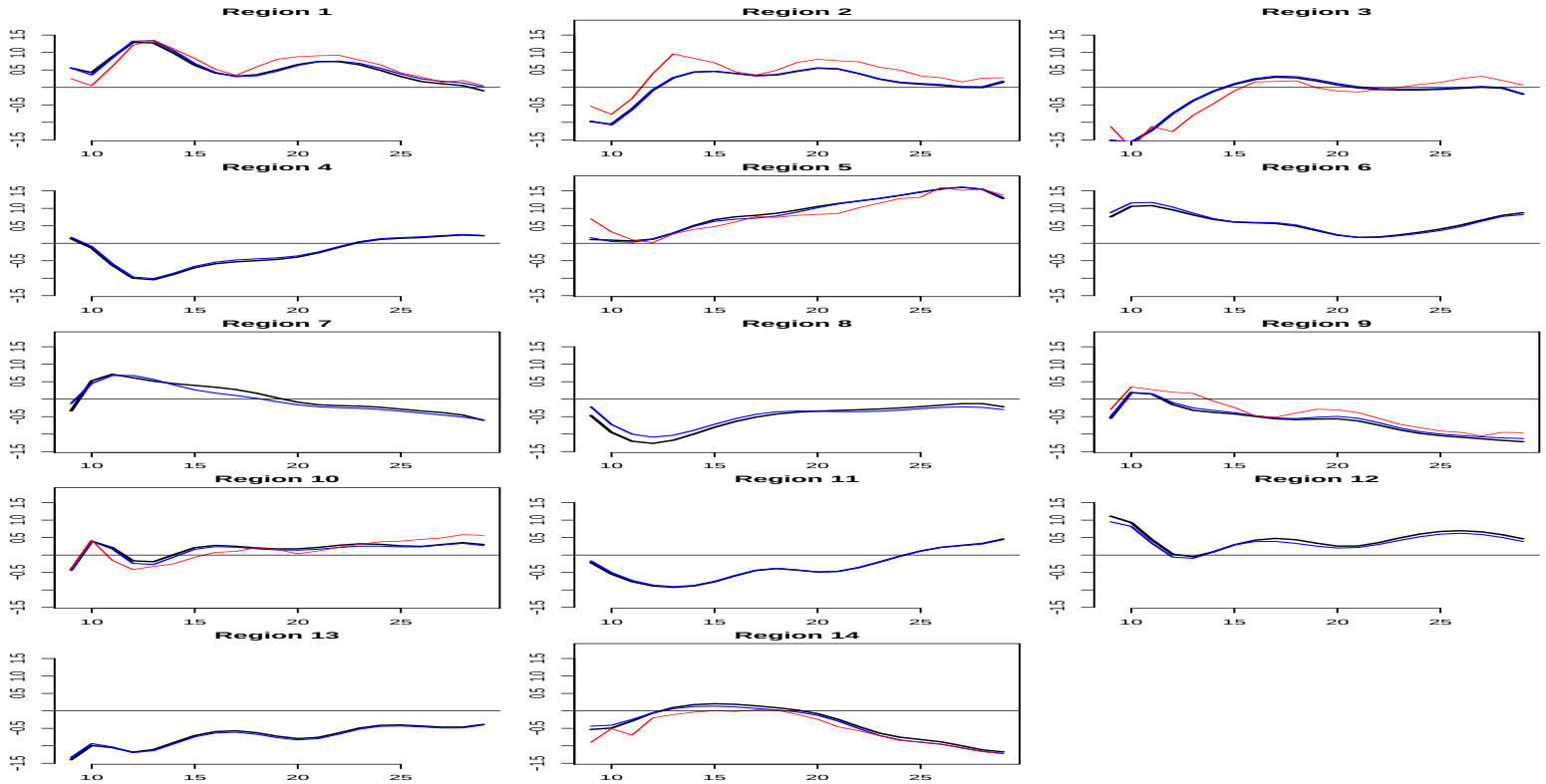
- 14 markers with posterior probability greater than 0.5
- Switch between some markers:

↬ Identification of new genomic regions compared to Marchadier et al. (2018)



Application on Arabidopsis thaliana

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



Conclusions and perspectives

Conclusions

- Estimation:
  - Functional approach allows reduction of the number of parameters
  - Non-parametric interpolation does not restrict the form of the effect curves
- Selection:
  - Bayesian group Lasso leads to biased estimation which can affect the selection
  - Group spike-and-slab performs well
- Various applications: Arabidopsis, Eucalyptus, Human, ...

Perspectives

Group spike-and-slab can have poor mixing when T increases
$\hookrightarrow$ extend continuous shrinkage prior: group horseshoe prior.

Bibliography

P. Eilers and B. Marx. Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2):89–121, May 1996. ISSN 0883-4237. doi: 10.1214/ss/1038425655.

E. George and R. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

E. George and R. McCulloch. Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

J. Ghosh and A. Ghattas. Bayesian Variable Selection Under Collinearity. *The American Statistician*, 69(3):165–173, 2015.

T. Hansen. The Evolution of Genetic Architecture. *Annual Review of Ecology, Evolution, and Systematics*, 37:123–157, 2006. doi: 10.1146/annurev.ecolsys.37.091305.110224.

T. Hastie and R. Tibshirani. *Generalized Additive Models*, volume 1. The Institute of Mathematical Statistics, 08 1986.

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

B. Heuclin, F. Mortier, C. Trottier, and M. Denis. Bayesian varying coefficient model with selection: An application to functional mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2020.

S. Lang and A. Brezger. Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, Mar. 2004. ISSN 1061-8600, 1537-2715. doi: 10.1198/1061860043010.

J. Li, Z. Wang, R. Li, and R. Wu. Bayesian group Lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *The Annals of Applied Statistics*, 9:640–664, June 2015. ISSN 1932-6157. doi: 10.1214/15-AOAS808.

Z. Li and M. Sillanpää. A Bayesian Nonparametric Approach for Mapping Dynamic Quantitative Traits. *Genetics*, 194(4):997–1016, Aug. 2013. doi: 10.1534/genetics.113.152736.

M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer, 1998.

E. Marchadier, M. Hanemian, S. Tisne, L. Bach, C. Bazakos, E. Gilbault, P. Haddadi, L. Virlouvet, and O. Loudet. The complex genetic architecture of shoot growth natural variation in Arabidopsis thaliana. July 2018. doi: 10.1101/354738.

J. Vanhatalo, Z. Li, and M. Sillanpää. A Gaussian process model and Bayesian variable selection for mapping function-valued quantitative traits with incomplete phenotypic data. *Bioinformatics*, 2019.

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