Bayesian varying coefficient model with selection: An application to functional mapping

Benjamin Heuclin

IMAG, Université de Montpellier / AGAP, Cirad, France

benjamin.heuclin@umontpellier.fr

Supervisors: **C. Trottier** (IMAG, UPVM3), **F. Mortier** (F&S, CIRAD), **M. Denis** (AGAP, CIRAD)

June 24th 2021, AppliBUGS

Publication

Paper accepted in Journal of the Royal Statistical Society C:
https://doi.org/10.1111/rssc.12447DOI: 10.1111/rssc.12447

R package `VCGSS` is available on github: https://github.com/Heuclin/VCGSS

Outline

1. Context: plant breeding

2. Statistical model
   - Estimation of the dynamic effects
   - Bayesian variable selection

3. Application on Eucalyptus

4. Conclusions and perspective

Outline

1. Context: plant breeding

2. Statistical model
   - Estimation of the dynamic effects
   - Bayesian variable selection

3. Application on Eucalyptus

4. Conclusions and perspective

Context: plant breeding

- Agricultural objectives are to
  - produce more
  - be more resistant to disease
  - require less water
  - be more resistant to high temperature
  - be adapted to climate change
- Strategy is to:
  - identify the best individuals
  - cross them to produce subsequent generations
  - repeat this on many generations (Recurrent selection)

Context: plant breeding

High throughput genotyping tools:

provide genetic information (markers) on the whole genome

↪ better understanding of the genetic architecture which controls the phenotypic trait (statistical tools: QTL mapping (Collard et al., 2005), GWAS (Huang and Han, 2014)

↪ accelerating genetic improvement through marker-assisted selection (He et al., 2014)

Context: plant breeding

Studying the genetic architecture that controls one phenotypic trait:

- Identify the molecular markers (X_j) that control the phenotypic trait (Y)
- Estimate the effects of these markers (β_j)
- Take into account the pedigree information (A) if available

Statistical tool:

Linear mixed model

$$Y = \mu + X_1\beta_1 + \cdots + X_q\beta_q + u + \varepsilon, \quad u \sim N_n(0, \sigma_A^2 A),\ \varepsilon \sim N_n(0, \sigma^2 I_n)$$

Marker identification ⇒ Variable selection

$$\boxed{\text{Does } \beta_j = 0 \text{ or not ? For } j = 1, \ldots, q}$$

Context: plant breeding

High throughput phenotyping tools:

- regular monitoring of a phenotypic trait over time
- automated recording and screening of phenotypes
- studying the dynamic genetic architecture of one phenotypic trait across the developmental stages.

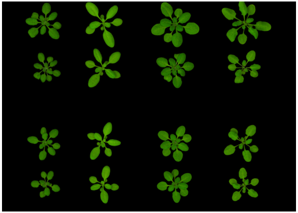


Figure: Arabidopsis thaliana (Marchadier et al., 2018)

New statistical challenges

↪ Which molecular markers (X_j) control the phenotypic trait over time $Y(t)$

↪ Estimate the dynamic effect of these markers ($\beta_j(t)$) over time

Context: plant breeding

High throughput phenotyping tools

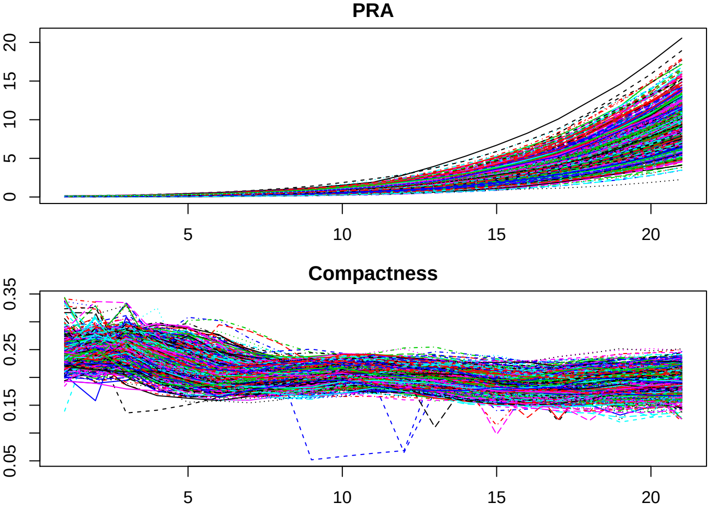


Figure: Arabidopsis thaliana (Marchadier et al., 2018)

Outline

1. Context: plant breeding

2. Statistical model
    - Estimation of the dynamic effects
    - Bayesian variable selection

3. Application on Eucalyptus

4. Conclusions and perspective

Statistical model

Linear model:

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

Statistical model

Linear model:

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

Statistical model

Linear model:

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

- Simple analysis at each time point does not take into account the correlations over the time

  ↪ Can lead to false positive detection and loss of statistical power

Statistical model

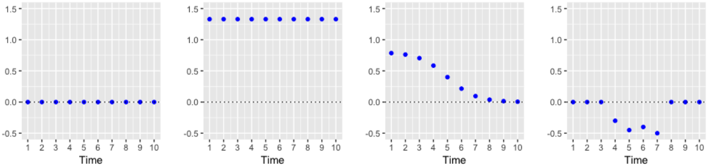
Dynamic linear model:

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

Statistical model

Dynamic linear model:

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



To understand the dynamic architecture that controls the trait:

- Estimation of coefficients β_j^t, $t = t_1, \dots t_T$, $j = 1, \dots, q$
- Selection of significant variables X_j, $j = 1, \dots, q$

  $\hookrightarrow$ Does $(\beta_j^{t_1}, \dots \beta_j^{t_T})' = (0, \dots, 0)'$?

Outline

1. Context: plant breeding

2. Statistical model
   - Estimation of the dynamic effects
   - Bayesian variable selection

3. Application on Eucalyptus

4. Conclusions and perspective

Statistical model

Estimation of the dynamic effects β_j

The qxT matrix of dynamic coefficients can be large $\begin{pmatrix} \beta_1^{t_1} & \dots & \beta_q^{t_1} \\ \vdots & & \vdots \\ \beta_1^{t_T} & \dots & \beta_q^{t_T} \end{pmatrix}$

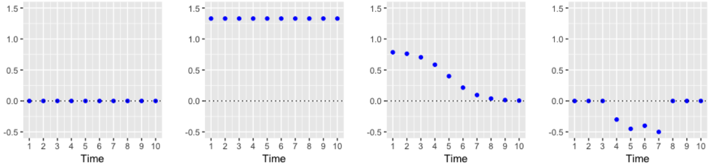
$\hookrightarrow$ Functional interpolation for each marker effects allows to reduce the number of parameters to be estimated

$\hookrightarrow$ This also has biological meaning, as we expect to see effects that change smoothly over time

Statistical model

Estimation of the dynamic effects β_j

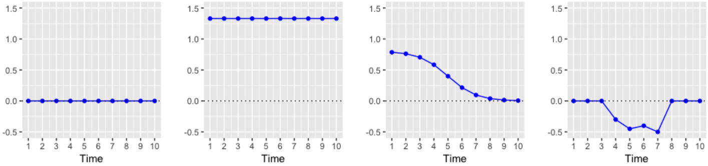
- Functional estimation of the dynamic effect



Statistical model

Estimation of the dynamic effects β_j

↬ Functional estimation of the dynamic effect



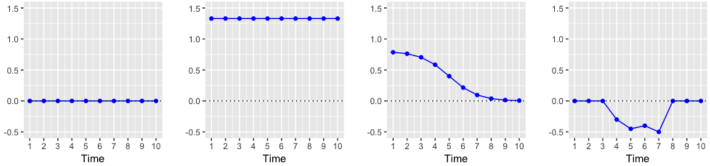
Parametric interpolation

- Linear curve (Li et al., 2014)
- Polynomial on t (Li and Sillanpää, 2015)
- Logistic curve (Wu and Lin, 2006)
- advantage: high reduction of parameters
- disadvantage: strong parametric assumption, does not correspond to complex effects

Statistical model

Estimation of the dynamic effects β_j

↬ Functional estimation of the dynamic effect



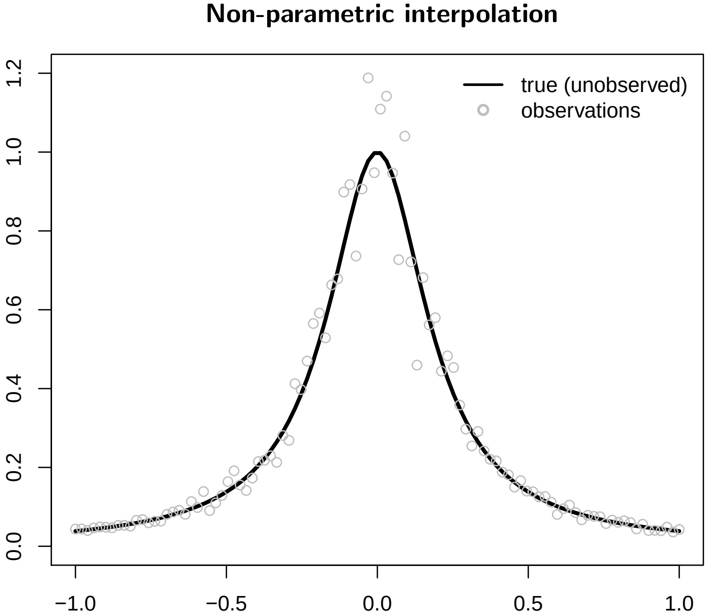
Parametric interpolation

- Linear curve (Li et al., 2014)
- Polynomial on t (Li and Sillanpää, 2015)
- Logistic curve (Wu and Lin, 2006)
- advantage: high reduction of parameters
- disadvantage: strong parametric assumption, does not correspond to complex effects

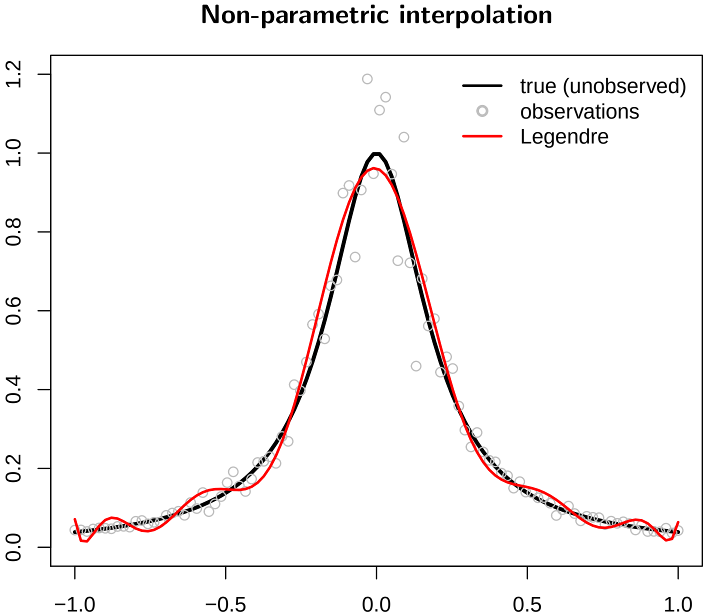
Non-parametric interpolation

- Legendre polynomial (Li et al., 2015)
- B-spline (Wang et al., 2008)
- P-spline
- advantage: more flexible
- disadvantage: more parameters than parametric curve

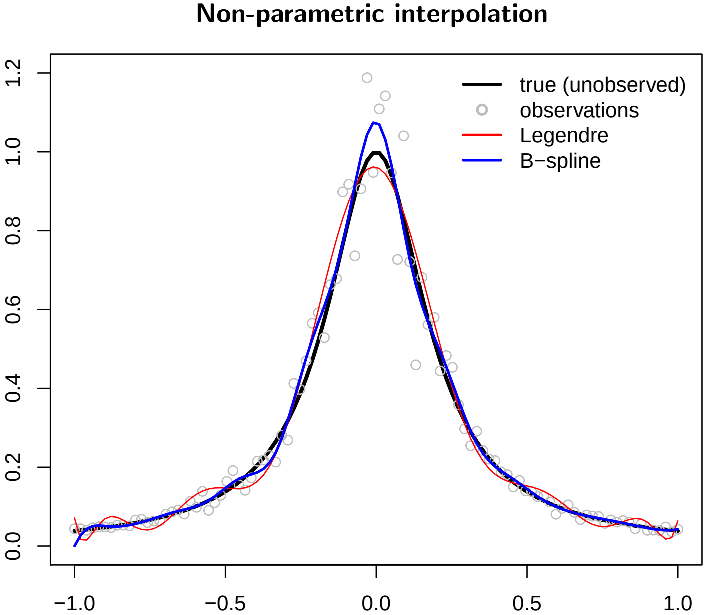
Statistical model



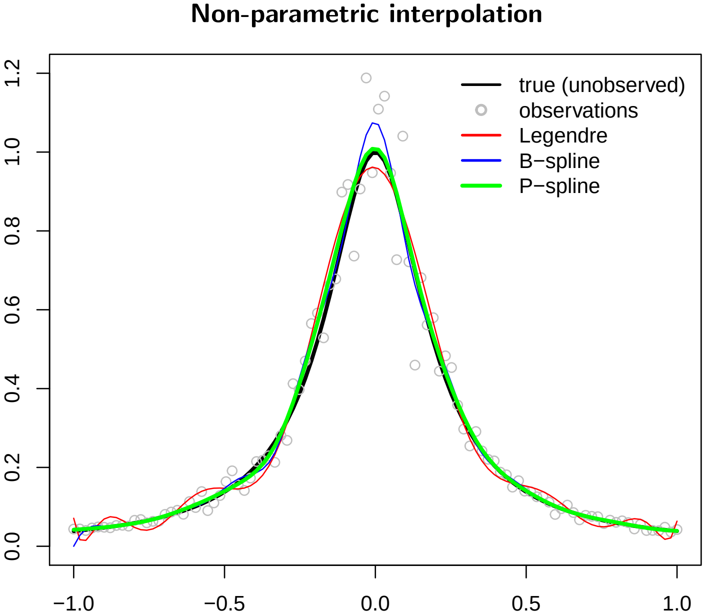
Statistical model



Statistical model



Statistical model



Statistical model

What is P-spline ?

P-spline = B-spline + Penalisation

Statistical model

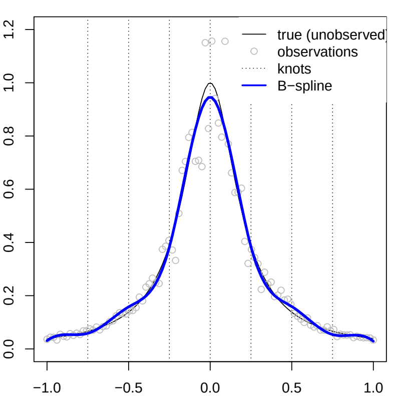
B-spline (Eubank, 1999)

- define knots
- interpolate third degree polynomials on each piece
- conditions: C^0, C^1, C^2
- can be formulated as linear combination

  $\hookrightarrow$ define a new basis: B

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

- Disadvantage: sensitive to the choice of knots



Statistical model

Rewrite the dynamic linear model:

$$\begin{pmatrix} y_i^{t_1} \\ \vdots \\ y_i^{t_T} \end{pmatrix} = \underbrace{\begin{pmatrix} \mu^{t_1} \\ \vdots \\ \mu^{t_T} \end{pmatrix}}_{B\ \ m} + \underbrace{\sum_{i=1}^{p} f_{e_i}\left(\begin{pmatrix} e_i^{t_1} \\ \vdots \\ e_i^{t_T} \end{pmatrix}\right)}_{\sum_{l=1}^{p} B_{env_l} e_l} + \underbrace{\begin{pmatrix} \beta_1^{t_1} & \dots & \beta_q^{t_1} \\ \vdots & & \vdots \\ \beta_1^{t_T} & \dots & \beta_q^{t_T} \end{pmatrix}}_{B\ \ b} \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \begin{pmatrix} \varepsilon_i^{t_1} \\ \vdots \\ \varepsilon_i^{t_T} \end{pmatrix}, \quad \begin{array}{l} \varepsilon_i \sim N_T(0, \sigma^2\Gamma) \\ \Gamma_{i,j} = \rho^{|i-j|} \\ -1 < \rho < 1 \end{array}$$

$$Y_i = B\ \ m \ + \ \sum_{l=1}^{p} B_{env_l} e_l \ \ + \ \ B\ \ b\ X_i + \varepsilon_i, \quad \begin{array}{l} \varepsilon_i \sim N_T(0, \sigma^2\Gamma) \\ \Gamma_{i,j} = \rho^{|i-j|} \\ -1 < \rho < 1 \end{array}$$

Statistical model

Penalized log-likelihood by the second order differences of adjacent B-spline coefficients (Eilers and Marx, 1996):

$$
\begin{aligned}
L = \textstyle\sum_{i=1}^{n} l(Y_i, m, e_1, \ldots, e_p, b_1, \ldots, b_q, \rho, \sigma^2) \\
-\lambda_0 m' D' Dm - \textstyle\sum_{l=1}^{p} \lambda_l' e_l' D' De_l - \textstyle\sum_{j=1}^{q} \lambda_k b_j' D' Db_j
\end{aligned}
$$

- smooth curves
- not sensitive to the knot positions
- disadvantage: choice of $\lambda_0, \lambda_1', \ldots, \lambda_p', \lambda_1, \ldots, \lambda_q$ via cross-validation is computationally intensive

↬ Bayesian formulation is convenient

Statistical model

Penalized log-likelihood on the second order derivative of each curve:

Bayesian point of view (Lang and Brezger, 2004):

- replace the penalties by their stochastic analogues
  - ↪ second-order random walk process

$$b_j|\lambda_j \sim N_v(0, (\lambda_j D'D)^{-1}),$$

$$\begin{aligned}
&Y_i|m, b, \rho, s \sim N_T(Bm + \textstyle\sum_{l=1}^{p} B_{env_l} e_l + BbX_i, \sigma^2\Gamma)\\
&m|\lambda_0 \sim N_v(0, (\lambda_0 D'D)^{-1})\\
&e_l|\lambda'_l \sim N_v(0, (\lambda'_l D'D)^{-1}),\ l = 1, \dots, p\\
&b_j|\lambda_j \sim N_v(0, (\lambda_j D'D)^{-1}),\ j = 1, \dots, q\\
&\lambda_j \sim Gamma(s, r),\ j = 0, \dots, q
\end{aligned}$$

MAP estimator ⇔ Maximum penalized log-likelihood estimator

Outline

1 Context: plant breeding

2 Statistical model

- Estimation of the dynamic effects
- Bayesian variable selection

3 Application on Eucalyptus

4 Conclusions and perspective

Statistical model

Bayesian variable selection

Bayesian prior for variable selection

- Shrinkage prior
  - Lasso prior (Park and Casella, 2008)
  - Group Lasso (Kyung et al., 2010)
  - Elastic-net prior (Kyung et al., 2010)
  - Horseshoe prior (Carvalho et al., 2008)
- Spike-and-slab prior (George and McCulloch, 1997)

Statistical model

Bayesian variable selection

Bayesian Spike-and-Slab

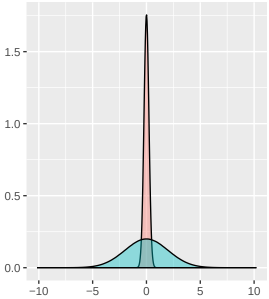
- introduction of γ:

$$\gamma_j = \begin{cases} 1 \text{ if variable } j \text{ is selected} \\ 0 \quad \text{otherwise} \end{cases}$$

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

- zero-inflated group spike-and-slab prior for P-spline coefficients $b_j = (b_{1,j}, \ldots, b_{v,j})'$:

$$\begin{aligned} & b_j|\gamma_j, \lambda_j \sim \gamma_j N_v(0, (\lambda_j D' D)^{-1}) + (1-\gamma_j)\delta_v(0) \\ & \lambda_j \sim Gamma(s, r), \\ & \gamma_j \sim Ber(\pi), \end{aligned}$$



- The estimation of $\mathbb{P}(\gamma_j = 1|Y)$ gives access to the a posteriori probability of variable selection

Statistical model

Bayesian hierarchical model

$$Y_i|m, b, \rho, \sigma^2 \sim N_T(Bm + \sum_{l=1}^{p} B_{env_l} e_l + BbX_i, \sigma^2\Gamma)$$

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

$$e_l|\lambda_l' \sim N_v(0, (\lambda_l' D'D)^{-1}),\ \ l = 1, \dots, p$$

$$b_j|\gamma_j, \lambda_j \sim \gamma_j N_v(0, (\lambda_j D'D)^{-1}) + (1-\gamma_j)\delta_v(0), \quad j = 1, \dots, q$$

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

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

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

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

To infer the distribution of $m, e, b, \lambda, \gamma, \rho, \sigma^2|Y$**:**

↪ Gibbs algorithm (Markov Chain Monte Carlo algorithm)

Outline

1 Context: plant breeding

2 Statistical model
- Estimation of the dynamic effects
- Bayesian variable selection

3 Application on Eucalyptus

4 Conclusions and perspective

Application on Eucalyptus

The data:

- specie: E.urophylla (n = 201 individuals)
- phenotypic trait: daily amplitude of radial shrinkage (DA)
- month: June-2013 (T = 31)
- application on all chromosomes (11)
  (q = 85 markers after removed markers with a correlation upper than 0.8)
- one environmental variable: VPD

Settings:

- P-spline with difference penalty order = 1
- repetitions: 30
- iterations: 20000
- burnin: 10000
- thin: 10

```
> fit <- VCM_fct(Y, X, ENV, interpolation = "P-spline", order_diff = 1,
+                    rep = 30, niter = 10000, burnin = 7000, thin = 10)
```

Application on Eucalyptus

Convergence diagnostics

- Gelman-Rubin's potential scale reduction factor for all parameters except β

```
> fit$gelman.diag
```

- Gelman-Rubin's potential scale reduction factor for beta

```
> fit$gelman.diag.b.psrf.median
```

- Trace plot with posterior densities

```
> plot(fit$mcmc_list)
```

- Visual diagnostic of convergence of marginal posterior probabilities of variable inclusion (gamma parameters)

```
> plot_diagnostic_gamma(fit)
```

Application on Eucalyptus

```
> plot(fit$estimation$mean.marginal.probabilities)
> abline(0.5, 0, lty = 2, col = "red")
```

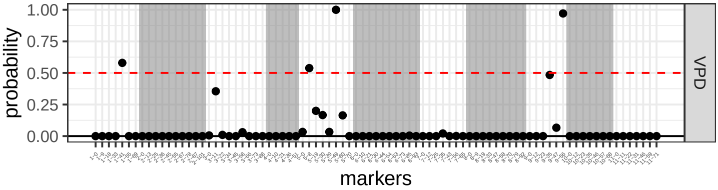


Figure: Posterior marginal probabilities of inclusion of markers. White and gray areas delimit the different chromosomes.

Selected markers with posterior inclusion probabilities upper than 0.45:

↪ 1-41, 5-8, 5-59, 9-35, 9-56

Application on Eucalyptus

plot functional effects

```
> plot_functional_effects(fit, plot=c("Y", "mu", "env","beta"), mfrow = c(2, 4),
+             id = which(prob>0.45), add = c("matplot", "quantile"))
```

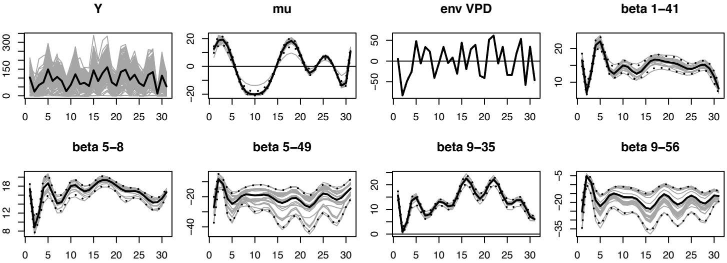


Figure: Estimated effects of the intercept (mu), environmental variable VPD and selected markers. Gray lines are the estimation for each repetition, black lines are the mean of the estimation over the repetitions and dotted lines are the credible intervals.

Outline

1 Context: plant breeding

2 Statistical model

- Estimation of the dynamic effects
- Bayesian variable selection

3 Application on Eucalyptus

4 Conclusions and perspective

Conclusions

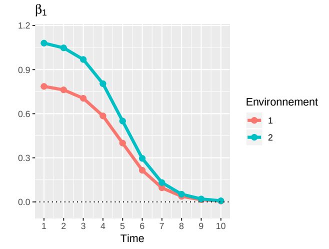
We propose a Bayesian approach combining P-spline interpolation and spike-and-slab selection

- **Estimation:**
  - functional approach allows reduction of the number of parameters
  - non-parametric interpolation does not restrict the form of the effect curves
  - P-spline allows fitting smooth or rather complicated curve
- **Selection:**
  - Bayesian group Lasso leads to biased estimation which can affect the selection
  - Spike-and-slab does not give biased the estimation
  - Spike-and-slab presents a good selection performance

Perspective: take into account the environment

Constant environment over time, different groups of individuals

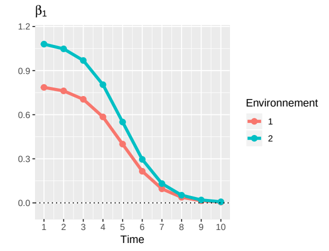
↪ for each variables X_j, different function effects $\beta_j^g(t)$ for each group



Perspective: take into account the environment

Constant environment over time, different groups of individuals

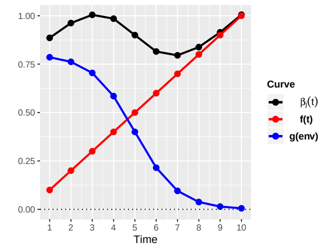
↪ for each variables X_j, different function effects $\beta_j^g(t)$ for each group



Variable environment over time, one group of individuals

↪ for each variables X_j, decomposition of function effect $\beta_j(t)$ as a sum of different function effects:

$$\beta_j(t) = g(env) + f(t)$$



Bibliography

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

Collard, B., Jahufer, M., Brouwer, J., and Pang, E. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142(1-2):169–196.

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

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. CRC press.

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

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science*, 5.

Huang, X. and Han, B. (2014). Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annual Review of Plant Biology*, 65(1):531–551.

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.

Li, J., Wang, Z., Li, R., and Wu, R. (2015). Bayesian group Lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *The Annals of Applied Statistics*, 9(2):640–664.

Li, Z., Hallingbäck, H. R., Abrahamsson, S., Fries, A., Gull, B. A., Sillanpää, M. J., and García-Gil, M. R. (2014). Functional Multi-Locus QTL Mapping of Temporal Trends in Scots Pine Wood Traits. *G3&amp;#58; Genes|Genomes|Genetics*, 4(12):2365–2379.

Li, Z. and Sillanpää, M. J. (2015). Dynamic Quantitative Trait Locus Analysis of Plant Phenomic Data. *Trends in Plant Science*, 20(12):822–833.

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.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements. *Journal of the American Statistical Association*, 103(484):1556–1569.

Wu, R. and Lin, M. (2006). Functional mapping — how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics*, 7(3):229–237.

Thank you for your attention!