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
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
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)
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)
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 ($\beta_j$)
- Take into account the pedigree information ($A$) if available

Statistical tool:

**Linear mixed model**

$$Y = \mu + X_1\beta_1 + \cdots + X_q\beta_q + u + \varepsilon,$$

$$u \sim N_n(0, \sigma^2_A A), \quad \varepsilon \sim N_n(0, \sigma^2 I_n)$$

Marker identification $\Rightarrow$ Variable selection

Does $\beta_j = 0$ 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:**

\[
y_{i,t_1} = \mu_{t_1} + (\beta_{t_1}^{1}, \ldots, \beta_{t_1}^{q}) (X_{i,1}) + \epsilon_{i,t_1}, \quad \epsilon_{i,t_1} \sim N(0, \sigma^2)
\]
Statistical model

Linear model:

\[
\begin{align*}
    y_{i1}^{t1} &= \mu_{t1} + (\beta_{11}^{t1}, \ldots, \beta_{q1}^{t1})^T X_{i1} + \epsilon_{i1}^{t1}, \quad \epsilon_{i1}^{t1} \sim N(0, \sigma^2) \\
    y_{i2}^{t2} &= \mu_{t2} + (\beta_{12}^{t2}, \ldots, \beta_{q2}^{t2})^T X_{i2} + \epsilon_{i2}^{t2}, \quad \epsilon_{i2}^{t2} \sim N(0, \sigma^2) \\
    \vdots
    y_{iT}^{tT} &= \mu_{tT} + (\beta_{1T}^{tT}, \ldots, \beta_{qT}^{tT})^T X_{iT} + \epsilon_{iT}^{tT}, \quad \epsilon_{iT}^{tT} \sim N(0, \sigma^2)
\end{align*}
\]
Statistical model

**Linear model:**

\[
\begin{align*}
y_{i_1} & = \mu_{t_1} + (\beta_{t_1}^{i_1}, \ldots, \beta_{q}^{t_1})\begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \varepsilon_{i_1}^{t_1}, \quad \varepsilon_{i_1}^{t_1} \sim N(0, \sigma^2) \\
y_{i_2} & = \mu_{t_2} + (\beta_{t_2}^{i_2}, \ldots, \beta_{q}^{t_2})\begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \varepsilon_{i_2}^{t_2}, \quad \varepsilon_{i_2}^{t_2} \sim N(0, \sigma^2) \\
\vdots & \vdots \\
y_{i_T} & = \mu_{t_T} + (\beta_{t_T}^{i_T}, \ldots, \beta_{q}^{t_T})\begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \varepsilon_{i_T}^{t_T}, \quad \varepsilon_{i_T}^{t_T} \sim N(0, \sigma^2)
\end{align*}
\]

- 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:

\[
\begin{pmatrix}
y_{11}^t \\
\vdots \\
y_{i1}^t \\
y_{iT}^t
\end{pmatrix} = \begin{pmatrix}
\mu_{11}^t \\
\vdots \\
\mu_{iT}^t
\end{pmatrix} + \sum_{i=1}^p f_{ei} \begin{pmatrix}
e_{1i}^t \\
\vdots \\
e_{Ti}^t
\end{pmatrix} + \begin{pmatrix}
\beta_{11}^t & \cdots & \beta_{q1}^t \\
\vdots & \ddots & \vdots \\
\beta_{1T}^t & \cdots & \beta_{qT}^t
\end{pmatrix} \begin{pmatrix}
X_{i,1} \\
\vdots \\
X_{iT}
\end{pmatrix} + \begin{pmatrix}
\varepsilon_{1i}^t \\
\vdots \\
\varepsilon_{Ti}^t
\end{pmatrix}
\] 
\[\varepsilon_i \sim N_T(0, \sigma^2 \Gamma), \quad \Gamma_{i,j} = \rho^{\lvert i-j \rvert}, \quad -1 < \rho < 1\]
Statistical model

Dynamic linear model:

\[
\begin{pmatrix}
y_{t1}^i \\
\vdots \\
y_{tT}^i
\end{pmatrix}
= \begin{pmatrix}
\mu_{t1}^i \\
\vdots \\
\mu_{tT}^i
\end{pmatrix} + \sum_{i=1}^{p} f_{e_i} \begin{pmatrix}
e_{t1}^i \\
\vdots \\
e_{tT}^i
\end{pmatrix} + \begin{pmatrix}
\beta_{1}^{t1} & \cdots & \beta_{q}^{t1} \\
\vdots & \ddots & \vdots \\
\beta_{1}^{tT} & \cdots & \beta_{q}^{tT}
\end{pmatrix} \begin{pmatrix}
X_{i,1} \\
\vdots \\
X_{i,q}
\end{pmatrix} + \begin{pmatrix}
\varepsilon_{t1}^i \\
\vdots \\
\varepsilon_{tT}^i
\end{pmatrix},
\varepsilon_i \sim N_T(0, \sigma^2 \Gamma)
\]

\[
\Gamma_{i,j} = \rho^{|i-j|}, \quad -1 < \rho < 1
\]

To understand the dynamic architecture that controls the trait:

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

$\hookrightarrow$ Does $(\beta_{j}^{t1}, \ldots \beta_{j}^{tT})' = (0, \ldots, 0)'$?
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 $\beta_j$

The $q \times T$ matrix of dynamic coefficients can be large

$$
\begin{pmatrix}
\beta_{11} & \ldots & \beta_{1q} \\
\vdots & \ddots & \vdots \\
\beta_{T1} & \ldots & \beta_{Tq}
\end{pmatrix}
$$

→ Functional interpolation for each marker effects allows to reduce the number of parameters to be estimated

→ This also has biological meaning, as we expect to see effects that change smoothly over time
Statistical model
Estimation of the dynamic effects $\beta_j$

Functional estimation of the dynamic effect
Statistical model

Estimation of the dynamic effects $\beta_j$

raquo; 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 $\beta_j$

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

Non-parametric interpolation

---

true (unobserved) observations

Benjamin Heuclin (IMAG, U. Montpellier)
June 24th 2021, AppliBUGS
Statistical model

Non-parametric interpolation

- true (unobserved)
- observations
- Legendre
Statistical model

Non-parametric interpolation

- true (unobserved) observations
- Legendre
- B-spline

Benjamin Heuclin (IMAG, U. Montpellier)
June 24th 2021, AppliBUGS
Statistical model

Non-parametric interpolation

- true (unobserved)
- observations
- Legendre
- B-spline
- P-spline

Benjamin Heuclin (IMAG, U. Montpellier)
June 24th 2021, AppliBUGS
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
  \[ \begin{pmatrix} \beta_j(t_1) \\ \vdots \\ \beta_j(t_T) \end{pmatrix} = \sum_{k=1}^{V} B_k b_{k,j} = B b_j, \]

- Disadvantage: sensitive to the choice of knots

Disadvantage:

<table>
<thead>
<tr>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.4</td>
<td>0.8</td>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

true (unobserved)
observations
knots
B-spline

Benjamin Heuclin (IMAG, U. Montpellier)

June 24th 2021, AppliBUGS 18 / 35
Rewrite the dynamic linear model:

\[
\begin{pmatrix}
  y_{i1}^t \\
  \vdots \\
  y_{iT}^t
\end{pmatrix}
= 
\begin{pmatrix}
  \mu_{1}^t \\
  \vdots \\
  \mu_{T}^t
\end{pmatrix}
+ 
\sum_{i=1}^{p} f_{e_i}
\begin{pmatrix}
  e_{i1}^t \\
  \vdots \\
  e_{iT}^t
\end{pmatrix}
+ 
\begin{pmatrix}
  \beta_{11}^t & \cdots & \beta_{1q}^t \\
  \vdots & \ddots & \vdots \\
  \beta_{q1}^t & \cdots & \beta_{qq}^t
\end{pmatrix}
\begin{pmatrix}
  X_{i1}^t \\
  \vdots \\
  X_{iq}^t
\end{pmatrix}
+ 
\begin{pmatrix}
  \varepsilon_{i1}^t \\
  \vdots \\
  \varepsilon_{iT}^t
\end{pmatrix}, 
\varepsilon_i \sim N_T(0, \sigma^2 \Gamma)
\]

\[\Gamma_{i,j} = \rho^{|i-j|}, \quad -1 < \rho < 1\]
Penalized log-likelihood by the second order differences of adjacent B-spline coefficients (Eilers and Marx, 1996):

\[ L = \sum_{i=1}^{n} l(Y_i, m, e_1, \ldots, e_p, b_1, \ldots, b_q, \rho, \sigma^2) \]
\[ -\lambda_0 m' D' Dm - \sum_{l=1}^{p} \lambda'_l e'_l D' D e_l - \sum_{j=1}^{q} \lambda_k b'_j D' D b_j \]

- 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

\[ \uparrow \] Bayesian formulation is convenient
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

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

\[
\begin{align*}
Y_i | m, b, \rho, s & \sim N_T(Bm + \sum_{l=1}^{p} B_{env} e_l + BbX_i, \sigma^2 \Gamma) \\
m | \lambda_0 & \sim N_v(0, (\lambda_0 D'D)^{-1}) \\
e_i | \lambda_l' & \sim N_v(0, (\lambda_l'D'D)^{-1}), \quad l = 1, \ldots, p \\
b_j | \lambda_j & \sim N_v(0, (\lambda_j D'D)^{-1}), \quad j = 1, \ldots, q \\
\lambda_j & \sim \text{Gamma}(s, r), \quad j = 0, \ldots, q
\end{align*}
\]

**MAP estimator ⇔ Maximum penalized log-likelihood estimator**
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)
Bayesian Spike-and-Slab

- introduction of $\gamma$:

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

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

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

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

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

Bayesian hierarchical model

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

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

\[
e_l | \lambda'_l \sim N_v(0, (\lambda'_l D' D)^{-1}), \; l = 1, \ldots, p
\]

\[
b_j | \gamma_j, \lambda_j \sim \gamma_j N_v(0, (\lambda_j D' D)^{-1}) + (1 - \gamma_j)\delta_v(0), \; j = 1, \ldots, q
\]

\[
\lambda_j \sim Gamma(s, r), \; j = 0, \ldots, q
\]

\[
\gamma_j \sim Ber(\pi), \; j = 1, \ldots, q
\]

\[
\rho \sim U[-1,1]
\]

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

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

\( \rightarrow \) Gibbs algorithm (Markov Chain Monte Carlo algorithm)
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

```r
> 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 $\beta$

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

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

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

- Trace plot with posterior densities

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

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

```r
> plot_diagnostic_gamma(fit)
```
> 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.
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

\[ \text{for each variables } X_j, \text{ different function effects } \beta_{j}^{g}(t) \text{ for each group} \]
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:

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


Thank you for your attention!