High-dimensional variable selection in non-linear mixed-effects models using a stochastic EM spike-and-slab

Marion Naveau[1,2]

M. Delattre[2], G. Kon Kam King[2], L. Sansonnet[1]

[1]Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay

[2]Université Paris-Saclay, INRAE, MaIAGE

European Meeting of Statisticians

3rd July 2023

INRAE

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Stat4Plant

Table of contents

1. Introduction
2. Methodology
3. Summary of simulation results
4. Conclusion

1. Introduction

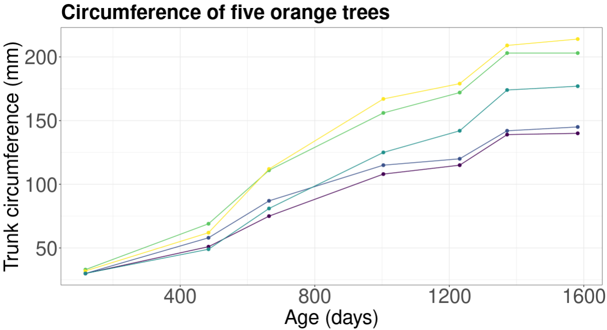
2. Methodology
   - Prior specification
   - Method

3. Summary of simulation results
   - Comparison with a two-step approach
   - Impacts of the different parameters and collinearity between covariates

4. Conclusion

Framework: repeated measurement data

✤ **Mixed-effects models:** analyse observations collected repeatedly on several individuals.



✤ Same overall behaviour but with individual variations.

✤ Non-linear growth.

✤ Are these variations due to known characteristics?

- ▶ E.g.: growing conditions, genetic markers, ...

Non-linear mixed-effects model (NLMEM)

1) Description of intra-individual variability:
   For all $i \in \{1, \dots, n\}$, $j \in \{1, \dots, n_i\}$,

$$y_{ij} = g(\varphi_i, \psi, t_{ij}) + \varepsilon_{ij}, \ \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

- $y_{ij} \in \mathbb{R}$: response of individual i at time t_{ij} (observation).
- $\varphi_i \in \mathbb{R}^q$: individual parameter, **not observed**.
- $\psi \in \mathbb{R}^r$: fixed effects, **unknown**.
- g: non-linear function with respect to φ_i (known).

2) Description of inter-individual variability:

$$\varphi_i = \mu + V_i \beta + \xi_i, \ \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(0, \Gamma)$$

- $\mu \in \mathbb{R}^q$: intercept, **unknown**.
- $V_i \in \mathbb{R}^p$: covariates for individual i (known).
- $\beta \in \mathcal{M}_{p \times q}$ covariate fixed effects matrix, **unknown**.

Population parameters: $\theta = (\mu, \beta, \psi, \sigma^2, \Gamma)$

High-dimensional covariate selection in NLMEM

✤ Specificity of the problem: $p >> n$

✤ Goal: identify the non-zero components of β.

✤ Main difficulties:

- High-dimensional variable selection:
  - parsimonious estimation of β
    - ➢ regularised methods (LASSO-type, Tibshirani (1996))
    - ➢ sparsity-inducing priors (Tadesse and Vannucci, 2021)
- Non-explicit likelihood
  - The φ_i's are not observed (latent variables model)
    - ➢ theoretical and algorithmic in LMEM (Schelldorfer et al., 2011)
  - g is non-linear
    - ➢ algorithmic only in NLMEM (Ollier, 2021)

Proposed approach

Association of a Bayesian ***spike-and-slab*** prior for variable selection with a stochastic version of the EM algorithm, called **MCMC-SAEM**, for inference.

1. Introduction

2. Methodology
   - Prior specification
   - Method

3. Summary of simulation results
   - Comparison with a two-step approach
   - Impacts of the different parameters and collinearity between covariates

4. Conclusion

One-dimensional framework

$$\begin{cases} y_{ij} = g(\varphi_i, \psi, t_{ij}) + \varepsilon_{ij} & , \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \\ \varphi_i = \mu + \beta^\top V_i + \xi_i & , \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \gamma^2). \end{cases}$$

where $\varphi_i \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\beta \in \mathbb{R}^p$, $\gamma^2 > 0$, and $\theta = (\mu, \beta, \psi, \sigma^2, \gamma^2)$.

- Goal: identify

$$S^* = \left\{ \ell \in \{1, \dots, p\} \middle| \beta^*_\ell \neq 0 \right\},$$

where $\boldsymbol{\beta}^*$ is the true fixed effects vector.

Spike-and-slab prior for the coefficients of β

✤ Introduction of latent variables δ_ℓ, $1 \leq \ell \leq p$:

$$\delta_\ell = \begin{cases} 1 & \text{if covariate } \ell \text{ is to be included in the model,} \\ 0 & \text{otherwise.} \end{cases}$$

✤ **Spike-and-slab prior** on β George and McCulloch (1997):

$$\pi(\beta|\delta) = \mathcal{N}_p(0, \text{diag}((1-\delta_\ell)\nu_0 + \delta_\ell \nu_1)), \ 0 \leq \nu_0 < \nu_1 \text{ fixed,}$$

i.e. β_ℓ are independent and:

- $\beta_\ell|(\delta_\ell = 0) \sim \mathcal{N}(0, \nu_0)$: "spike" distribution, ν_0 small
- $\beta_\ell|(\delta_\ell = 1) \sim \mathcal{N}(0, \nu_1)$: "slab" distribution, ν_1 large

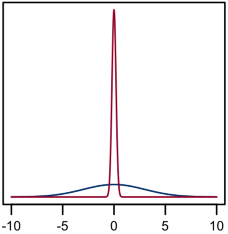
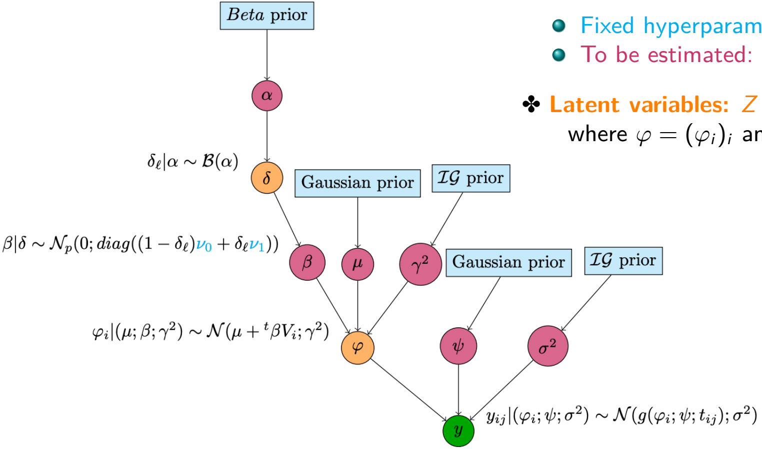


Figure: Spike-and-slab prior. Source: Deshpande et al. (2019)

Bayesian hierarchical model


$Beta$ prior
α
$\delta_\ell|\alpha \sim \mathcal{B}(\alpha)$
δ
Gaussian prior
$\mathcal{IG}$ prior
$\beta|\delta \sim \mathcal{N}_p(0; diag((1-\delta_\ell)\nu_0 + \delta_\ell\nu_1))$
β
μ
γ^2
Gaussian prior
$\mathcal{IG}$ prior
$\varphi_i|(\mu;\beta;\gamma^2) \sim \mathcal{N}(\mu + {}^t\beta V_i;\gamma^2)$
φ
ψ
σ^2
$y_{ij}|(\varphi_i;\psi;\sigma^2) \sim \mathcal{N}(g(\varphi_i;\psi;t_{ij});\sigma^2)$
y

✤ **Observations:** $y = (y_{ij})_{i,j}$

✤ **Parameters:**

- Fixed hyperparameters: ν_0, ν_1, ...
- To be estimated: $\Theta = (\theta, \alpha)$

✤ **Latent variables:** $Z = (\varphi, \delta)$
where $\varphi = (\varphi_i)_i$ and $\delta = (\delta_\ell)_\ell$

Proposed method

Idea: explore different levels of sparsity in β by varying the value of ν_0 in a grid Δ.

1. **Creation of a model collection:** for each $\nu_0 \in \Delta$,
   - Compute $\widehat{\Theta}$ by a MCMC-SAEM algorithm (Kuhn and Lavielle, 2004):
     $$\widehat{\Theta}_{\nu_0}^{MAP} = \underset{\Theta \in \Lambda}{\text{argmax}}\ \pi(\Theta|y)$$
   - Estimate $\hat{\delta}$ (Ročková and George, 2014):
     $$\hat{\delta} = \underset{\delta}{\text{argmax}}\ P(\delta|\hat{\Theta}_{\nu_0}^{MAP}) \text{ such as } \hat{\delta}_\ell = 1 \Longleftrightarrow \mathbb{P}(\delta_\ell = 1|\hat{\Theta}_{\nu_0}^{MAP}) \geq 0.5$$
     $$\Longleftrightarrow \text{Define } \widehat{S}_{\nu_0} = \left\{ \ell \in \{1, \ldots, p\} \;\middle|\; |(\widehat{\beta}_{\nu_0}^{MAP})_\ell| \geq s_\beta(\nu_0, \nu_1, \widehat{\alpha}_{\nu_0}^{MAP}) \right\}$$

2. **Select the "best" model** among $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$ by a fast criterion, eBIC (Chen and Chen, 2008):
   $$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\text{argmin}} \left\{ -2\log\left(p(y; \hat{\theta}_{\nu_0}^{MLE})\right) + B_{\nu_0} \times \log(n) + 2\log\left(\binom{p}{B_{\nu_0}}\right) \right\}$$
   with B_{ν_0}: number of free parameters in the model $\widehat{S}_{\nu_0}$.

3. **Return** $\widehat{S}_{\hat{\nu}_0}$.

Spike-and-slab regularisation plot

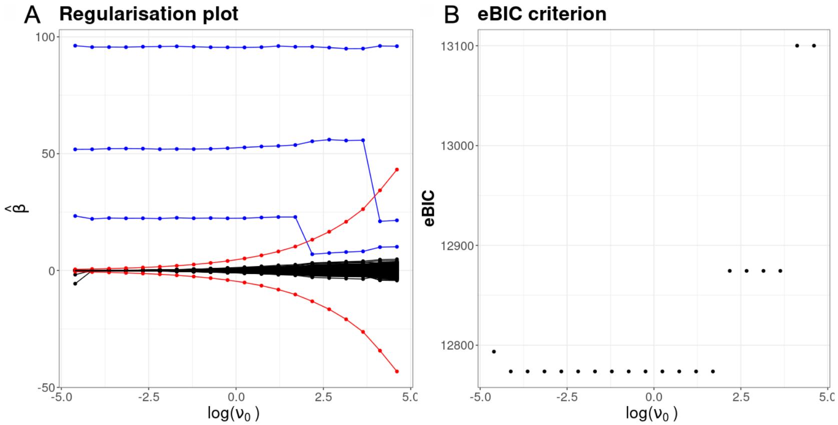


Figure: $n = 200$, $J = 10$, $p = 500$, $\gamma^2 = 200$, $\sigma^2 = 30$, $\nu_1 = 12000$, $\mu = 1200$, $\beta = {}^t(100, 50, 20, 0, \ldots, 0)$

1. Introduction

2. Methodology
   - Prior specification
   - Method

3. Summary of simulation results
   - Comparison with a two-step approach
   - Impacts of the different parameters and collinearity between covariates

4. Conclusion

Two-step approach

✤ Two-step approach:

1. Estimate the φ_i's individual-by-individual thanks to the **nlm** R function (Non-Linear Minimization),
2. Perform variable selection using the estimated parameters $\hat{\varphi}_i$ with **glmnet** R package (LASSO).

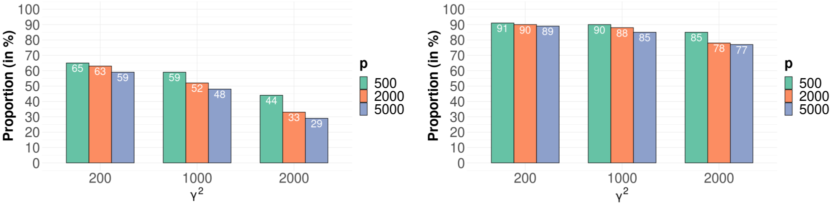
✤ Results:

- This strategy works fine in data-rich scenarios, when each parameter can be estimated very precisely, but it loses the uncertainty on the estimated parameters.
- Our procedure outperforms the two-step approach for scenarios with missing data.
- Thanks to the mixed model, individuals with missing data can benefit from the remaining fully observed individuals.

⇒ Show the interest of carrying out the selection of covariates from the data of all the individuals simultaneously thanks to the mixed effects model.

Model selection performance

✤ Independent covariates:



(a) $n = 100$

(b) $n = 200$

Figure: Empirical probability of correct model selection.

- When the procedure fails, it is most often because it under-selects:
  - ▶ "Cautious" approach, few false positives!

✤ Correlated covariates: Fairly similar good performance but with more false positives and/or false negatives in some correlation scenarios.

✤ Comparison with MCMC: The proposed inference method is about 20 times faster than a full MCMC implementation.

1. Introduction

2. Methodology
   - Prior specification
   - Method

3. Summary of simulation results
   - Comparison with a two-step approach
   - Impacts of the different parameters and collinearity between covariates

4. Conclusion

Conclusion and perspectives

✤ Summary:

- Development of an original method that combines SAEM and Bayesian variable selection.
- Very encouraging numerical results on simulated data (correlated and uncorrelated covariates).
- Faster method than a full MCMC implementation.
- More efficient than a 2-step approach.
- Relevant results on real data.

✤ Perspectives:

- Provide theoretical guarantees: **selection consistency**.

Thank you for your attention!

Naveau, M., Kon Kam King, G., Rincent, R., Sansonnet, L., and Delattre, M. (2022). **Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm.** arXiv preprint arXiv:2206.01012.

References I

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Deshpande, S. K., Ročková, V., and George, E. I. (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, 28(4):921–931.

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

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM: Probability and Statistics*, 8:115–131.

References II

Ollier, E. (2021). Fast selection of nonlinear mixed effect models using penalized likelihood. *arXiv preprint arXiv:2103.01621*.

Ročková, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Schelldorfer, J., Bühlmann, P., and DE GEER, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

Tadesse, M. G. and Vannucci, M. (2021). Handbook of bayesian variable selection.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Computing the MAP in a latent variables model

✤ Let's go back to the **first step** of the proposed method:

- Compute the MAP estimator of Θ
- Goal: maximise $\pi(\Theta|y) = \int_{\mathcal{Z}} \pi(\Theta, Z|y)dZ$ with

$$\pi(\Theta, Z|y) = \frac{p(y|\Theta, Z)p(\Theta, Z)}{\int_{\mathcal{Z}} \int_{\Lambda} p(y|\Theta, Z)p(\Theta, Z)d\Theta dZ}$$

- Non-explicit integral

EM algorithm (Dempster et al., 1977)

1. Initialisation: choose $\Theta^{(0)}$.
2. Iteration $k \geq 0$:
   - **E-step (Expectation):** compute

$$Q(\Theta|\Theta^{(k)}) = \mathbb{E}_{Z|(y,\Theta^{(k)})}\left[\log(\pi(\Theta, Z|y))\middle| y, \Theta^{(k)}\right].$$

   - **M-step (Maximisation):** compute

$$\Theta^{(k+1)} = \underset{\Theta\in\Lambda}{\operatorname{argmax}}\; Q(\Theta|\Theta^{(k)}).$$

3. $\hat{\Theta} = \Theta^{(K)}$, for K large enough.

Specifics in Spike-and-Slab-NLMEM

✤ Decomposition of Q:

$$\begin{aligned} Q(\Theta|\Theta^{(k)}) &= \mathbb{E}_{(\varphi,\delta)|(y,\Theta^{(k)})}[\log(\pi(\Theta,\varphi,\delta|y))|y,\Theta^{(k)}] \\ &= C + \underbrace{\mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)})\middle|y,\Theta^{(k)}\right]}_{\text{non-explicit}} + \underbrace{\widetilde{Q}_2(\alpha,\Theta^{(k)})}_{\text{explicit}} \end{aligned}$$

✤ M-step:

- θ and α estimated separately.
- $\widehat{\alpha}$ updated as in an EM algorithm with $\widetilde{Q}_2(\alpha,\Theta^{(k)})$.
- $\widehat{\theta}$ updated via stochastic approximation of:

$$\mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)})\middle|y,\Theta^{(k)}\right].$$

➢ SAEM (Delyon et al., 1999)
➢ MCMC-SAEM (Kuhn and Lavielle, 2004)

Specifics in Spike-and-Slab-NLMEM

✤ Decomposition of Q:

$$\begin{aligned}
Q(\Theta|\Theta^{(k)}) &= \mathbb{E}_{(\varphi,\delta)|(y,\Theta^{(k)})}[\log(\pi(\Theta,\varphi,\delta|y))|y,\Theta^{(k)}] \\
&= \mathbb{E}_{\varphi|(y,\Theta^{(k)})}\left[\mathbb{E}_{\delta|(\varphi,y,\Theta^{(k)})}\left[\log(\pi(\Theta,\varphi,\delta|y))|\varphi,y,\Theta^{(k)}\right]\middle|y,\Theta^{(k)}\right] \\
&= \mathbb{E}_{\varphi|(y,\Theta^{(k)})}\left[\widetilde{Q}(y,\varphi,\Theta,\Theta^{(k)})\middle|y,\Theta^{(k)}\right] \\
&= C + \underbrace{\mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)})\middle|y,\Theta^{(k)}\right]}_{\text{non-explicit}} + \underbrace{\widetilde{Q}_2(\alpha,\Theta^{(k)})}_{\text{explicit}}
\end{aligned}$$

MCMC-SAEM algorithm in SSNLMEM

1. Initialisation: choose $\Theta^{(0)}$ and $Q_{1,0}(\theta) = 0$,
2. Iteration $k \geq 0$:
   - **S-step (Simulation):** simulate $\varphi^{(k)}$ using the result of one iteration of an MCMC procedure with $\pi(\varphi|y, \Theta^{(k)})$ for target distribution,
   - **SA-step (Stochastic Approximation):** compute

     $$Q_{1,k+1}(\theta) = Q_{1,k}(\theta) + \gamma_k(\widetilde{Q}_1(y, \varphi^{(k)}, \theta, \Theta^{(k)}) - Q_{1,k}(\theta)),$$

     and $\widetilde{Q}_2(\alpha, \Theta^{(k)})$,
   - **M-step (Maximisation):**

     $$\theta^{(k+1)} = \underset{\theta \in \Lambda_\theta}{\text{argmax}}\ Q_{1,k+1}(\theta) \text{ and } \alpha^{(k+1)} = \underset{\alpha \in [0,1]}{\text{argmax}}\ \widetilde{Q}_2(\alpha, \Theta^{(k)}),$$

3. $\hat{\Theta} = \Theta^{(K)}$, for K large enough,

where $(\gamma_k)_k$ a step sizes sequence decreasing towards 0 such that $\forall k$, $\gamma_k \in [0, 1]$, $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$.

Logistic growth model

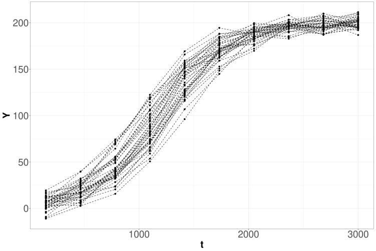


Figure: Simulated data

- Size of plant $i \in \{1, \ldots, n\}$ at time t_{ij}, $j \in \{1, \ldots, 10\}$: $y_{ij} = g(\varphi_i, \psi, t_{ij}) + \varepsilon_{ij}$, $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ where:

$$g(\varphi_i, \psi, t_{ij}) = \frac{\psi_1}{1 + \exp\left(-\frac{t_{ij} - \varphi_i}{\psi_2}\right)}$$

$\psi = (\psi_1, \psi_2)$ fixed effects.

- φ_i: characteristic time $\varphi_i = \mu + {}^t\beta V_i + \xi_i$, $\xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \gamma^2)$

$$\boxed{\theta = (\mu, \beta, \psi, \sigma^2, \gamma^2)}$$

Simulation design

✤ Parameters:

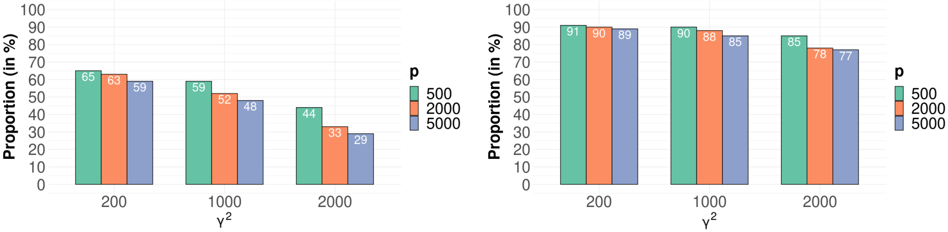
- $n \in \{100, 200\}$ individuals,
- $p \in \{500, 2000, 5000\}$ simulated covariates according to $V_i \sim \mathcal{N}(0, \Sigma)$:
  - ▶ Scenario i.i.d.: $\Sigma = Id$ ▶ Correlated scenarios: $\Sigma \neq Id$
- $\beta = {}^t(100, 50, 20, 0, \ldots, 0)$ covariate fixed effects vector,
- $\gamma^2 \in \{200, 1000, 2000\}$ inter-individual variance,
- $\mu = 1200$, $\sigma^2 = 30$, $\psi = (\psi_1, \psi_2) = (200, 300)$.

✤ Spike-and-slab hyperparameters:

- $\nu_1 = 12000$ slab variance,
- $\log_{10}(\Delta) = \left\{ -2 + k \times \frac{4}{19}, k \in \{0, \ldots, 19\} \right\}$ grid of ν_0 values.

▶ For each combination of (n, p, γ^2), the method is applied on 100 different simulated datasets.

Results for independent covariates

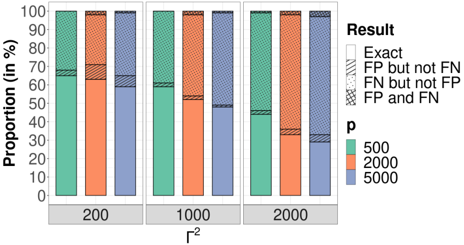


(a) $n = 100$ (b) $n = 200$

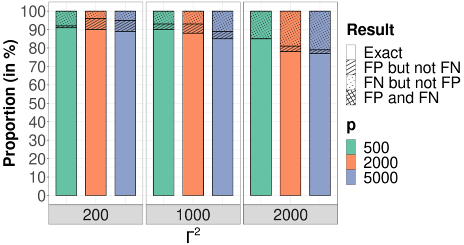
Figure: Empirical probability of correct model selection.

- Results improve as n increases.
- Degradation of results when p or γ^2 increases.
- When the procedure fails, it is most often because it under-selects:
  - "Cautious" approach, few false positives!

Results for uncorrelated covariates

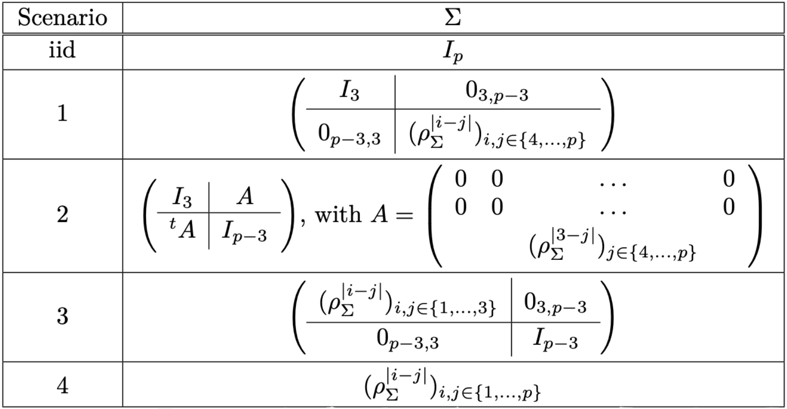


(a) For $n = 100$



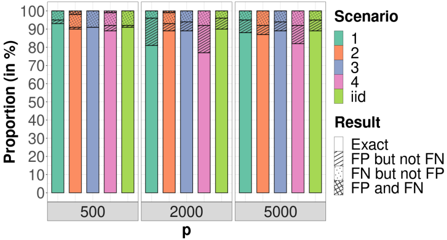
(b) For $n = 200$

Correlated covariates $V_i \sim \mathcal{N}(0, \Sigma)$

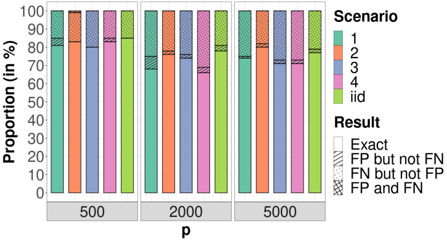


<table>
<tr><th>Scenario</th><th>Σ</th></tr>
<tr><td>iid</td><td>I_p</td></tr>
<tr><td>1</td><td>$\left(\begin{array}{c|c} I_3 & 0_{3,p-3} \\ \hline 0_{p-3,3} & (\rho_\Sigma^{|i-j|})_{i,j\in\{4,\dots,p\}} \end{array}\right)$</td></tr>
<tr><td>2</td><td>$\left(\begin{array}{c|c} I_3 & A \\ \hline {}^tA & I_{p-3} \end{array}\right)$, with $A = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ & & (\rho_\Sigma^{|3-j|})_{j\in\{4,\dots,p\}} & \end{pmatrix}$</td></tr>
<tr><td>3</td><td>$\left(\begin{array}{c|c} (\rho_\Sigma^{|i-j|})_{i,j\in\{1,\dots,3\}} & 0_{3,p-3} \\ \hline 0_{p-3,3} & I_{p-3} \end{array}\right)$</td></tr>
<tr><td>4</td><td>$(\rho_\Sigma^{|i-j|})_{i,j\in\{1,\dots,p\}}$</td></tr>
</table>

Results for $\rho_\Sigma = 0.3$

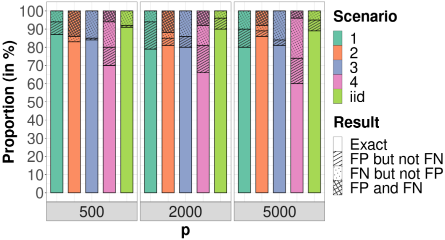


(c) For $\Gamma^2 = 200$

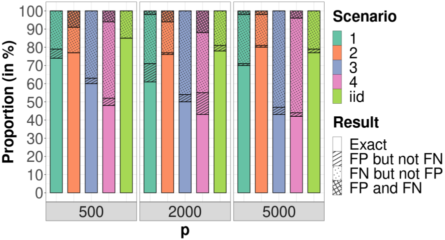


(d) For $\Gamma^2 = 2000$

Results for $\rho_{\Sigma} = 0.6$



(e) For $\Gamma^2 = 200$



(f) For $\Gamma^2 = 2000$

Summary of the results

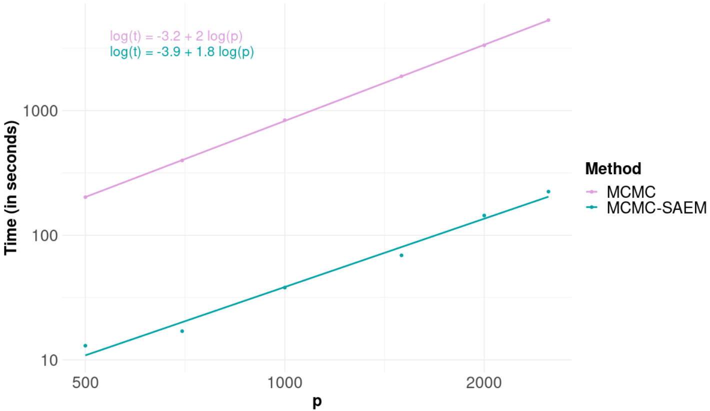
✤ Uncorrelated covariates $V_i \sim \mathcal{N}(0, I_p)$:

- Results improve as n increases.
- Degradation of results when p or Γ^2 increases.
- When the procedure fails, it is most often because it under-selects:
  - ▶ "Cautious" approach, few false positives!

✤ Correlated covariates $V_i \sim \mathcal{N}(0, \Sigma)$:

- Fairly similar good performance.
- More false positives and/or false negatives in some correlation scenarios:
  - ▶ + false positives: correlations between active and non-active covariates.
  - ▶ + false negatives: correlated active covariates.

Comparison with an MCMC implementation



NB: fast C++ adaptive MCMC (Nimble) versus R code

- Both methods have an execution time that grows **polynomially** with p.
- The proposed inference method can browse grid of about **20 values** of ν_0 while adaptive MCMC explores a single value.

Comparison with a two-step approach

✤ Two-step approach:

1. Estimate the φ_i's individual-by-individual thanks to the **nlm** R function (Non-Linear Minimization)
2. Perform variable selection using as dependent variables the estimated parameters with **glmnet** R package:
   a) LASSO in multivariate version (group LASSO),
   b) LASSO in univariate version.

✤ Scenarios of observations:

1. **Complete data-set:** all individuals are observed during the entire experiment.
2. **Partial observations:** For each $p_{\text{partial}} \in \{0.1, 0.2, 0.3, 0.4\}$, the other scenarios correspond to the case where $N_1 = p_{\text{partial}} n$ individuals are assumed to be no longer part of the experiment after the 3rd observation time.

Simulation design

$$\begin{cases} y_{ij} = \dfrac{D\varphi_{i1}}{V\varphi_{i1} - \varphi_{i2}} \left(\mathrm{e}^{-\dfrac{\varphi_{i2}}{V} t_{ij}} - \mathrm{e}^{-\varphi_{i1} t_{ij}} \right) + \varepsilon_{ij}, & \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \\ \varphi_i = \mu + \beta^\top V_i + \xi_i, & \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(0, \Gamma), \end{cases}$$

where $\varphi_i = (\varphi_{i1}, \varphi_{i2})^\top$, namely $q = 2$.

- $D = 100,\ V = 30,$ $(t_{i1}, \dots, t_{i12}) = (0.05, 0.15, 0.25, 0.4, 0.5, 0.8, 1, 2, 7, 12, 24, 40)$
- $n = 200,\ n_1 = \cdots = n_n = 12,\ p = 500,\ \sigma^2 = 10^{-3},$ $\Gamma = \begin{pmatrix} 0.2 & 0.05 \\ 0.05 & 0.1 \end{pmatrix},\ Cor((\varphi_{i1})_i, (\varphi_{i2})_i) = 0.35,$
- $\mu = (6, 8)^\top,\ \beta = \begin{pmatrix} 3 & 2 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 3 & 2 & 1 & 0 & \dots & 0 \end{pmatrix}^\top$
- $V_i \in \mathbb{R}^p$, $1 \leq i \leq n$, are simulated independently according to a binomial distribution with a success probability of 0.2.

Mean estimation errors

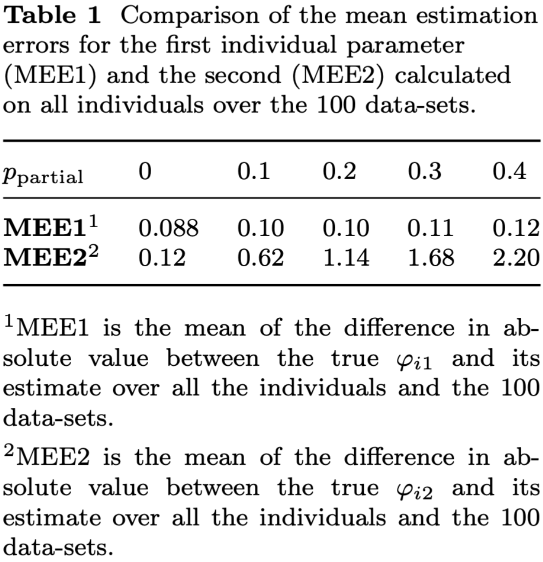


Table 1 Comparison of the mean estimation errors for the first individual parameter (MEE1) and the second (MEE2) calculated on all individuals over the 100 data-sets.

p_{partial}	0	0.1	0.2	0.3	0.4
MEE1[1]	0.088	0.10	0.10	0.11	0.12
MEE2[2]	0.12	0.62	1.14	1.68	2.20

[1] MEE1 is the mean of the difference in absolute value between the true φ_{i1} and its estimate over all the individuals and the 100 data-sets.

[2] MEE2 is the mean of the difference in absolute value between the true φ_{i2} and its estimate over all the individuals and the 100 data-sets.

Results

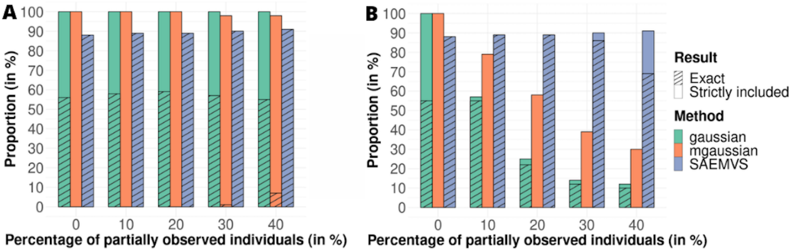
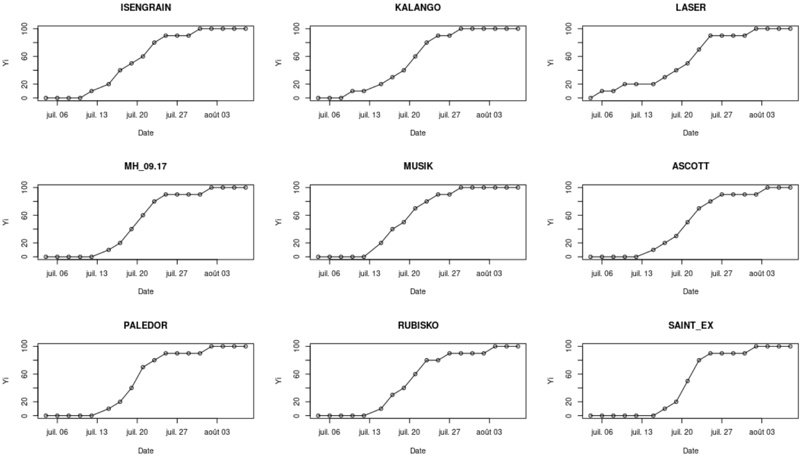


Figure: Proportion of data-sets on which the three methods (in colour) select the correct model ("Exact", striped bars), or a model that strictly includes the correct model ("Strictly included", unpatterned bars) for the first individual parameter (a) and the second individual parameter (b), and different percentage of partially observed individuals (on the *x*-axis).

Presentation of the dataset

- Wheat leaf senescence data.
- **Panel:** $n = 220$ soft wheat varieties subjected to nitrogen stress, observed $J = 18$ times.
- Varieties respond differently to stress: for example, some of them tolerate stress better and senescence is delayed.
- For each variety: genotyping information on several tens of thousands of SNPs.
- **Aim:** select molecular markers, from among $p = 34838$ markers, which could be associated with the senescence process.

Data representation: percentage of desiccated leaves



$\Longrightarrow$ Logistic growth

Modelling for one chromosome

$$\begin{cases} y_{ij} = \dfrac{100}{1+\exp\left(-\dfrac{t_{ij}-\varphi_i}{\psi_i}\right)} + \varepsilon_{ij} & , \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2) \\ \varphi_i = \mu + \lambda^\top v_i + \beta^\top V_i^C + \xi_i & , \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Gamma^2) \\ \psi_i = \eta + \omega_i & , \omega_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Omega^2) \end{cases}$$

where:

- $v_i \in \mathbb{R}^5$: covariates not subject to selection, allows the inclusion of sub-populations in the model,
- $V_i^C \in \mathbb{R}^p$: molecular markers, subject to selection, which contains heading QTLs and flowering genes.

$$\theta = (\mu, \lambda, \beta, \eta, \sigma^2, \Gamma^2, \Omega^2)$$

Results

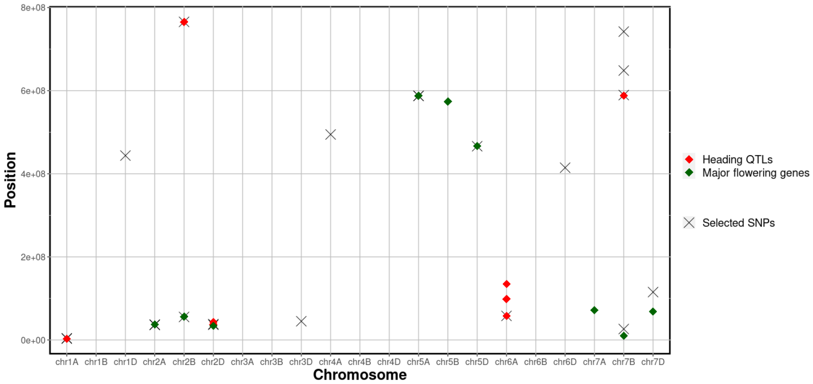


Figure: Position on each chromosome of the markers selected by SAEMVS (in black cross), compared to heading QTLs (in red diamond) and major flowering genes (in green diamond).

Markers highly correlated

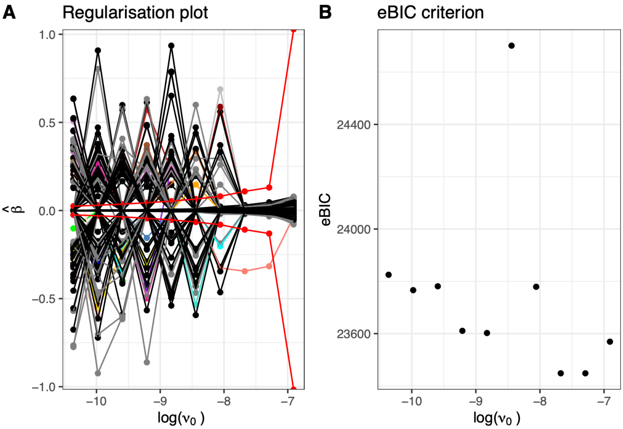


Figure: Regularisation plot and eBIC criterion for chromosome 6A