# Bayesian high-dimensional variable selection in non-linear mixed-effects models using the SAEM algorithm

**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

**Journées de Statistique**
13 Juin 2022

INRAE

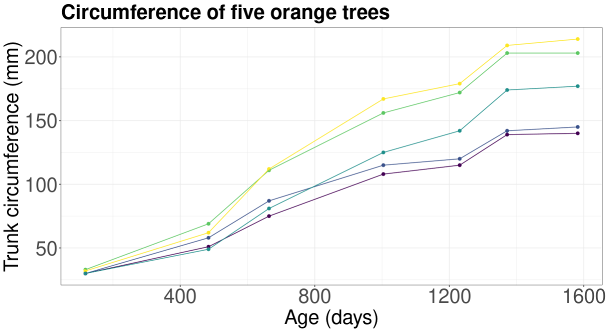
université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Stat4Plant

# Table of contents

# 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 $i \in \{1, \dots, n\}$, $j \in \{1, \dots, J\}$,

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

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

2) Description of inter-individual variability:

$$\varphi_i = \mu + {}^t\beta V_i + \xi_i,\ \ \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma^2)$$

- $\mu \in \mathbb{R}$: intercept, **unknown**.
- $V_i \in \mathbb{R}^p$: covariates for individual $i$ (known).
- $\beta = {}^t(\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ covariate fixed effects vector, **unknown**.

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

# High-dimensional covariate selection in NLMEM

- Goal: identify the non-zero components of $\beta$.
- Specificity of the problem: $p >> n$
- Main difficulties:
  - High-dimensional variable selection:
    - parsimonious estimation of $\beta$
      - regularised methods (LASSO-type, Tibshirani (1996))
      - sparsity-inducing priors (Tadesse and Vannucci, 2021)
  - Non-explicit likelihood
    - The $\varphi_i$'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.

# Spike-and-slab prior for the coefficients of $\beta$

- Introduction of latent variables $\delta_\ell$, $1 \leq \ell \leq p$:

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

- **Spike-and-slab prior** on $\beta$ 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.* $\beta_\ell$ are independent and:

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

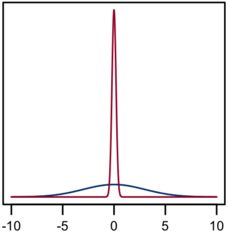
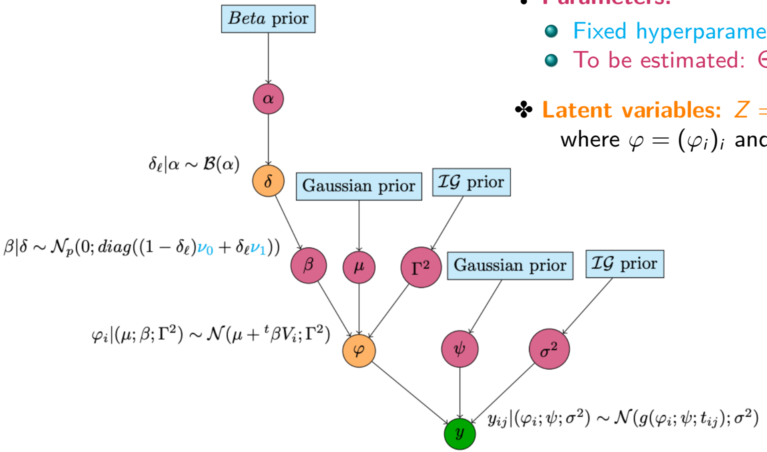



Figure: Spike-and-slab prior. Source: Deshpande et al. (2019)

# Bayesian hierarchical model




- **Observations:** $y = (y_{ij})_{i,j}$
- **Parameters:**
  - Fixed hyperparameters: $\nu_0$, $\nu_1$, ...
  - 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 $\beta$ by varying the value of $\nu_0$ in a grid $\Delta$.

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):
   $$\hat{\delta} = \underset{\delta}{\text{argmax}}\ P(\delta|\hat{\Theta}_{\nu_0}^{MAP}) \text{ such as } \hat{\delta}_\ell = 1 \iff \mathbb{P}(\delta_\ell = 1|\hat{\Theta}_{\nu_0}^{MAP}) \geq 0.5$$
   $$\iff \text{Define } \widehat{S}_{\nu_0} = \left\{ \ell \in \{1, \ldots, p\} \;\middle|\; |(\widehat{\beta}_{\nu_0}^{MAP})_\ell| \geq s_\beta(\nu_0, \nu_1, \widehat{\alpha}_{\nu_0}^{MAP}) \right\}$$

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 $B_{\nu_0}$: number of free parameters in the model $\widehat{S}_{\nu_0}$.

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

# Spike-and-slab regularisation plot

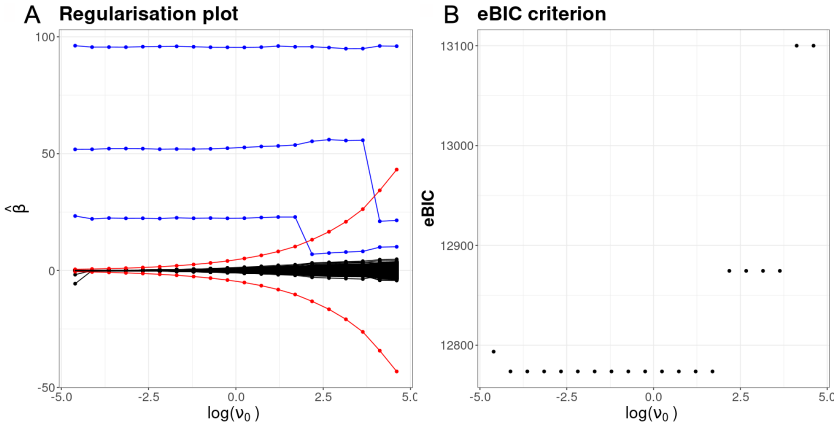



Figure: $n = 200$, $J = 10$, $p = 500$, $\Gamma^2 = 200$, $\sigma^2 = 30$, $\nu_1 = 12000$, $\mu = 1200$, $\beta = {}^t(100, 50, 20, 0, \dots, 0)$

# 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 $\Theta$
- 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

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

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

# Specifics in Spike-and-Slab-NLMEM

## Decomposition of $Q$:

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

## M-step:

- $\theta$ and $\alpha$ 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)

1. Introduction

2. Methodology
   - Prior specification
   - Method
   - Computation of the MAP

3. Simulation study

4. Conclusion

# Logistic growth model

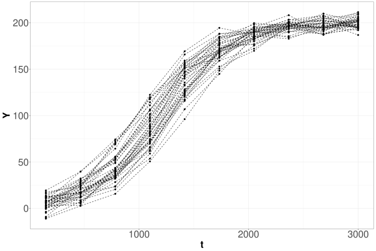



Figure: Simulated data

- 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.
- $\varphi_i$: 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:

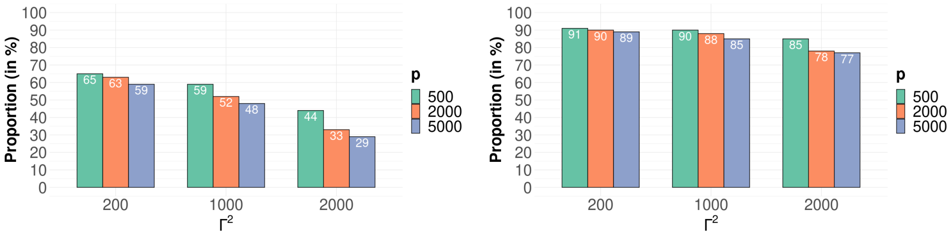
- $n \in \{100, 200\}$ individuals,
- $p \in \{500, 2000, 5000\}$ simulated covariates according to $V_i \sim \mathcal{N}(0, \Sigma)$:
  - ▶ Scenario i.i.d.: $\Sigma = Id$ ▶ Correlated scenarios: $\Sigma \neq Id$
- $\beta = {}^t(100, 50, 20, 0, \ldots, 0)$ covariate fixed effects vector,
- $\Gamma^2 \in \{200, 1000, 2000\}$ 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 $\nu_0$ values.

▶ For each combination of $(n, p, \Gamma^2)$, 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 $\Gamma^2$ increases.
- When the procedure fails, it is most often because it under-selects:
  - "Cautious" approach, few false positives!

# Conclusion and perspectives

✤ **Summary:**

- Development of an original method that combines SAEM and Bayesian variable selection.
- Very encouraging numerical results on simulated data.
- Faster method than a full MCMC implementation.

⇒ **Preprint:** Naveau and al. (2022). Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm. arXiv:2206.01012.

✤ **Perspectives:**

- Provide theoretical guarantees: selection consistency.
- Apply our method to a real dataset (in progress).
- Consider a multidimensional individual parameter.

# Thank you for your attention!

# References

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.

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.

# 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

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

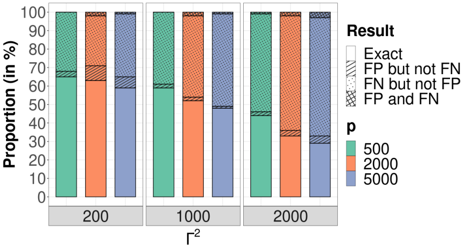
   and $\widetilde{Q}_2(\alpha, \Theta^{(k)})$,
   - **M-step (Maximisation):**

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

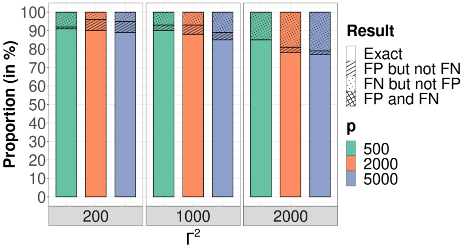
3. $\hat{\Theta} = \Theta^{(K)}$, for $K$ large enough,

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 $\sum_k \gamma_k^2 < \infty$.

# Results for uncorrelated covariates




(a) For $n = 100$

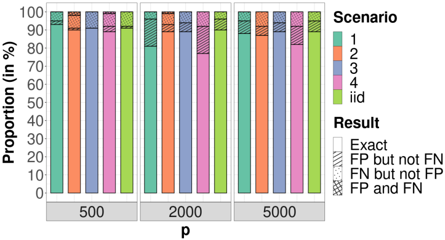



(b) For $n = 200$

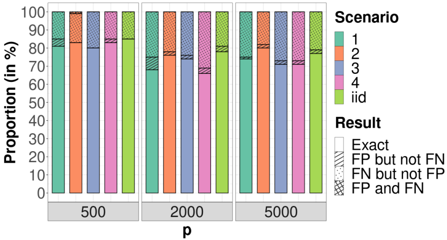
# Correlated covariates $V_i \sim \mathcal{N}(0, \Sigma)$

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

# Results for $\rho_\Sigma = 0.3$

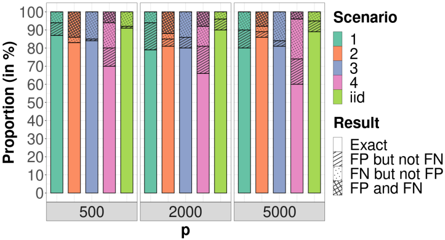



(c) For $\Gamma^2 = 200$

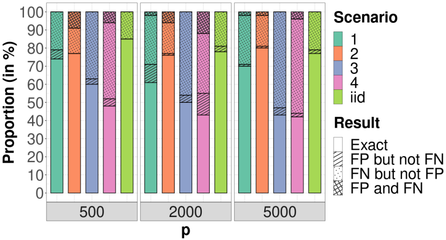



(d) For $\Gamma^2 = 2000$

# Results for $\rho_\Sigma = 0.6$

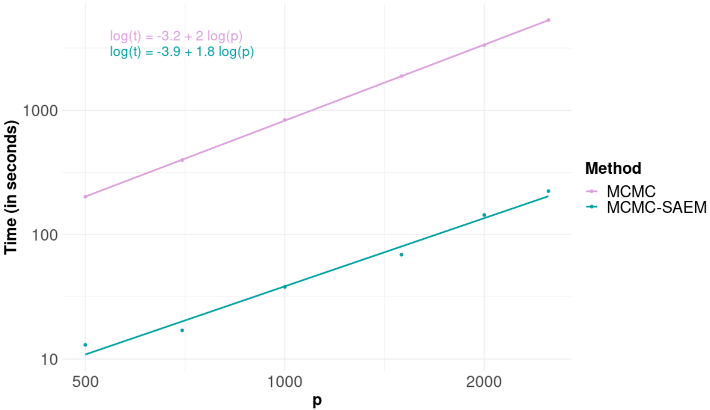



(e) For $\Gamma^2 = 200$




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

# 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 $\nu_0$ while adaptive MCMC explores a single value.