Sélection de variables en grande dimension dans les modèles non linéaires à effets mixtes

Marion Naveau, Guillaume Kon Kam King, Laure Sansonnet, Maud Delattre

Sélection de variables en grande dimension dans les modèles non linéaires à effets mixtes.

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

[1]Université Paris-Saclay, INRAE, UMR MaIAGE

[2]Université Paris-Saclay, INRAE, UMR MIA-Paris

9ème Rencontre des Jeunes Statisticien-ne-s
3 - 7 Avril 2022

INRAE

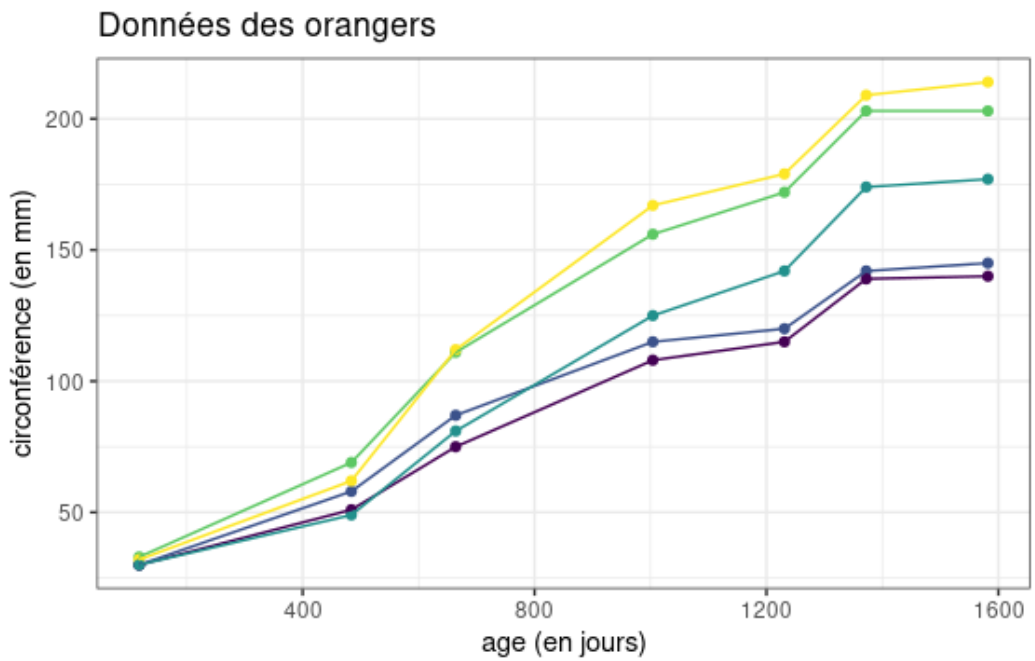
université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Stat4Plant

Sommaire

Contexte biologique : données de mesures répétées

✤ Modèles à effets mixtes : analyser des observations collectées de façon répétée sur plusieurs individus.



✤ Identifier les covariables les plus pertinentes pour caractériser la variation inter-individuelle.

- Ex : marqueurs génétiques.
  - Grande dimension des données génomiques.

Modèle non-linéaire à effets mixtes (NLMEM)

Pour tout $1 \leq i \leq n$, $1 \leq j \leq J$:

$$\begin{cases} y_{ij} = g(\varphi_i, \psi, t_{ij}) + \varepsilon_{ij} & , \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \quad (1a) \\ \varphi_i = \mu + {}^t\beta V_i + \xi_i & , \xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Gamma^2), \quad (1b) \end{cases}$$

où :

- $y_{ij} \in \mathbb{R}$: réponse de l'individu i au temps t_{ij} (observation).
- $\varphi_i \in \mathbb{R}$: paramètre individuel, **non observé**.
- $\psi \in \mathbb{R}^q$: effets fixes, **inconnus**.
- g : **fonction non linéaire** en φ_i.
- $\mu \in \mathbb{R}$: intercept, **inconnu**.
- $V_i = {}^t(V_{i1}, \ldots, V_{ip}) \in \mathbb{R}^p$, $V_{i\ell}$ valeur de la covariable ℓ de l'individu i (connue).
- $\beta = {}^t(\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ vecteur d'effets covariables fixe, **inconnu**.

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

✤ $\beta_\ell = 0 \iff$ la covariable ℓ n'a pas d'effet sur le paramètre φ_i

✤ $\beta_\ell \neq 0 \iff$ la covariable ℓ donne une information sur le paramètre φ_i

Sélection de covariables en grande dimension dans les NLMEM

✤ **Objectif :** identifier les composantes non nulles de β

✤ **Spécificité du problème :** l'ensemble des covariables est de **grande dimension** $\longrightarrow p >> n$

✤ **Principales difficultés :**

- Sélection de variables en grande dimension
  - ▶ Estimation parcimonieuse de β
- Vraisemblance non explicite
  - ▶ Les φ_i ne sont pas observés (modèle à variables latentes)
  - ▶ g est non linéaire

État de l'art

Cadre fréquentiste

✤ En linéaire : résultats théoriques de consistance en sélection pour méthodes régularisées type Lasso [Schelldorfer et al., 2011].
✤ En non linéaire : aspects computationnels uniquement [Bertrand and Balding, 2013], [Ollier, 2021].

Cadre bayésien

✤ En régression linéaire : développements théoriques et computationnels utilisant divers **priors *sparse*** [Tadesse and Vannucci, 2021].
✤ En NLMEM : implémentation MCMC pour l'inférence [Lee, 2022].

Notre contribution

Association d'une méthode bayésienne de type ***spike-and-slab*** pour la sélection de variables avec une version stochastique de l'algorithme EM, appelée **MCMC-SAEM**, pour l'inférence.

1. Introduction

2. Méthodologie
   - Spécification des priors bayésiens
   - Méthode
   - Calcul du MAP

3. Étude de simulations

4. Conclusion

Prior spike-and-slab

✤ Introduction des variables latentes δ_ℓ, $1 \leq \ell \leq p$:

$$\delta_\ell = \begin{cases} 1 & \text{si la covariable } \ell \text{ est à inclure dans le modèle,} \\ 0 & \text{sinon.} \end{cases}$$

✤ **Prior spike-and-slab** sur β [George and McCulloch, 1997] :

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

i.e. les β_ℓ sont indépendants et :

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

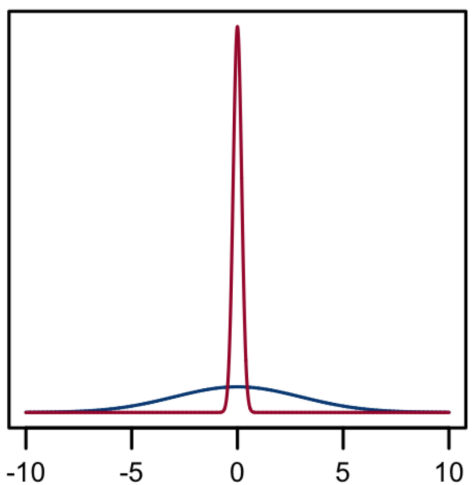
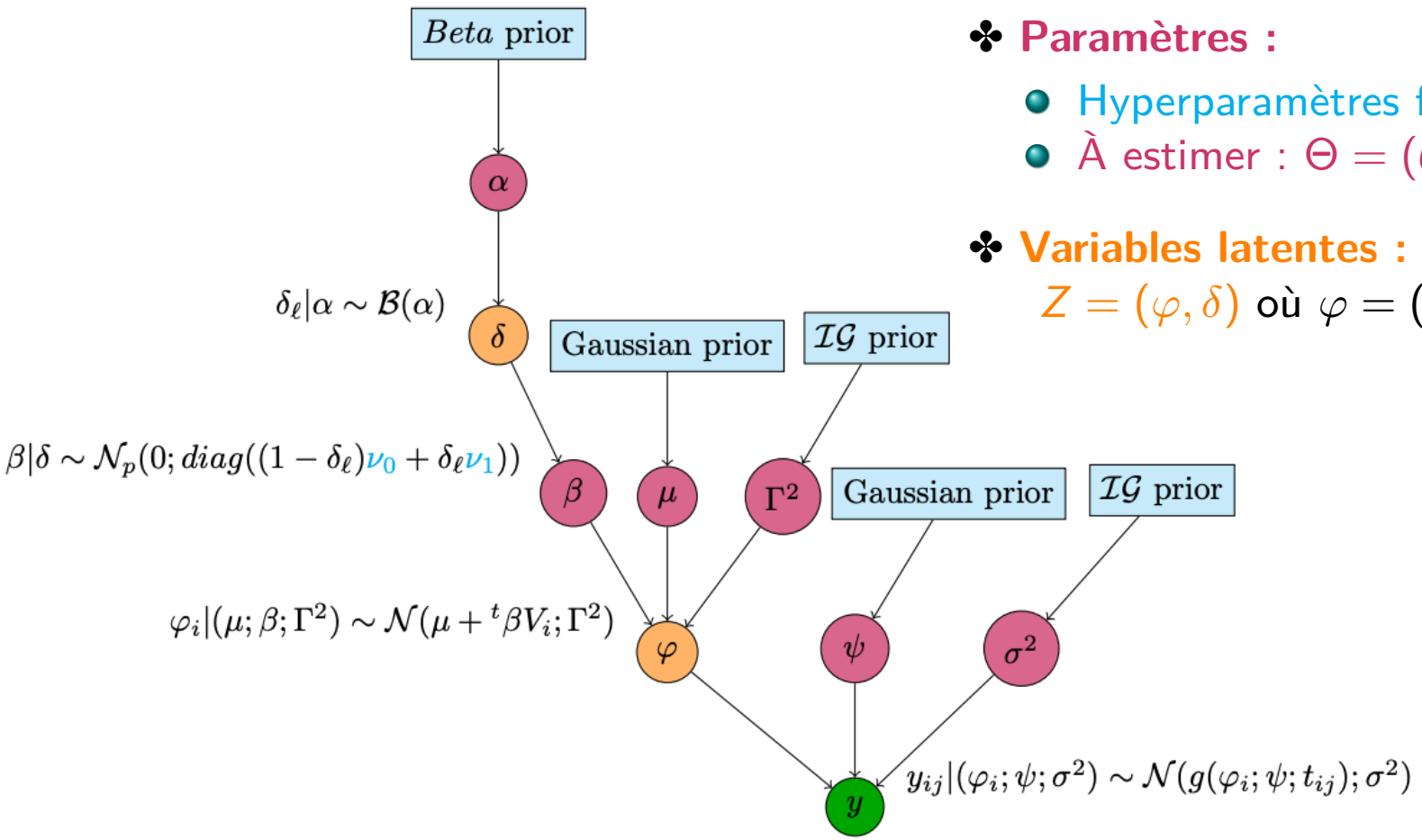


Figure – Prior spike-and-slab. Source : [Deshpande et al., 2019]

Modèle hiérarchique bayésien



- **Observations :** $y = (y_{ij})_{i,j}$
- **Paramètres :**
  - Hyperparamètres fixés : ν_0, ν_1, ...
  - À estimer : $\Theta = (\theta, \alpha)$
- **Variables latentes :**
  $Z = (\varphi, \delta)$ où $\varphi = (\varphi_i)_i$ et $\delta = (\delta_\ell)_\ell$

Méthode proposée

Idée : explorer différents niveaux de parcimonie dans β

$\hookrightarrow$ Δ grille de valeurs du paramètre de pénalisation ν_0

1. **Réduire la collection de modèles :** pour chaque $\nu_0 \in \Delta$,
   - ▶ calculer l'estimateur du maximum *a posteriori* par MCMC-SAEM [Kuhn and Lavielle, 2004] :

$$\widehat{\Theta}_{\nu_0}^{MAP} = \underset{\Theta \in \Lambda}{\text{argmax}}\ \pi(\Theta|y)$$

   - ▶ Sélectionner un sous-ensemble de covariables pertinentes [Ročková and George, 2014] :

$$\widehat{S}_{\nu_0} = \left\{ \ell \in \{1, \dots, p\} \;\middle|\; |(\widehat{\beta}_{\nu_0}^{MAP})_\ell| \geq s_\beta(\nu_0, \nu_1, \widehat{\alpha}_{\nu_0}^{MAP}) \right\}$$

   Seuil ? Estimer les probabilités d'inclusion des covariables *a posteriori* :

$$\hat{\delta} = \underset{\delta}{\text{argmax}}\ P(\delta|\hat{\Theta}_{\nu_0}^{MAP}) \text{ tel que } \hat{\delta}_\ell = 1 \iff \mathbb{P}(\delta_\ell = 1|\hat{\Theta}_{\nu_0}^{MAP}) \geq 0.5$$

2. **Sélectionner le "meilleur" modèle** parmi $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$ avec un critère eBIC [Chen and Chen, 2008] :

$$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\text{argmin}} \left\{ eBIC(\widehat{S}_{\nu_0}) \right\}$$

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

Chemin de régularisation

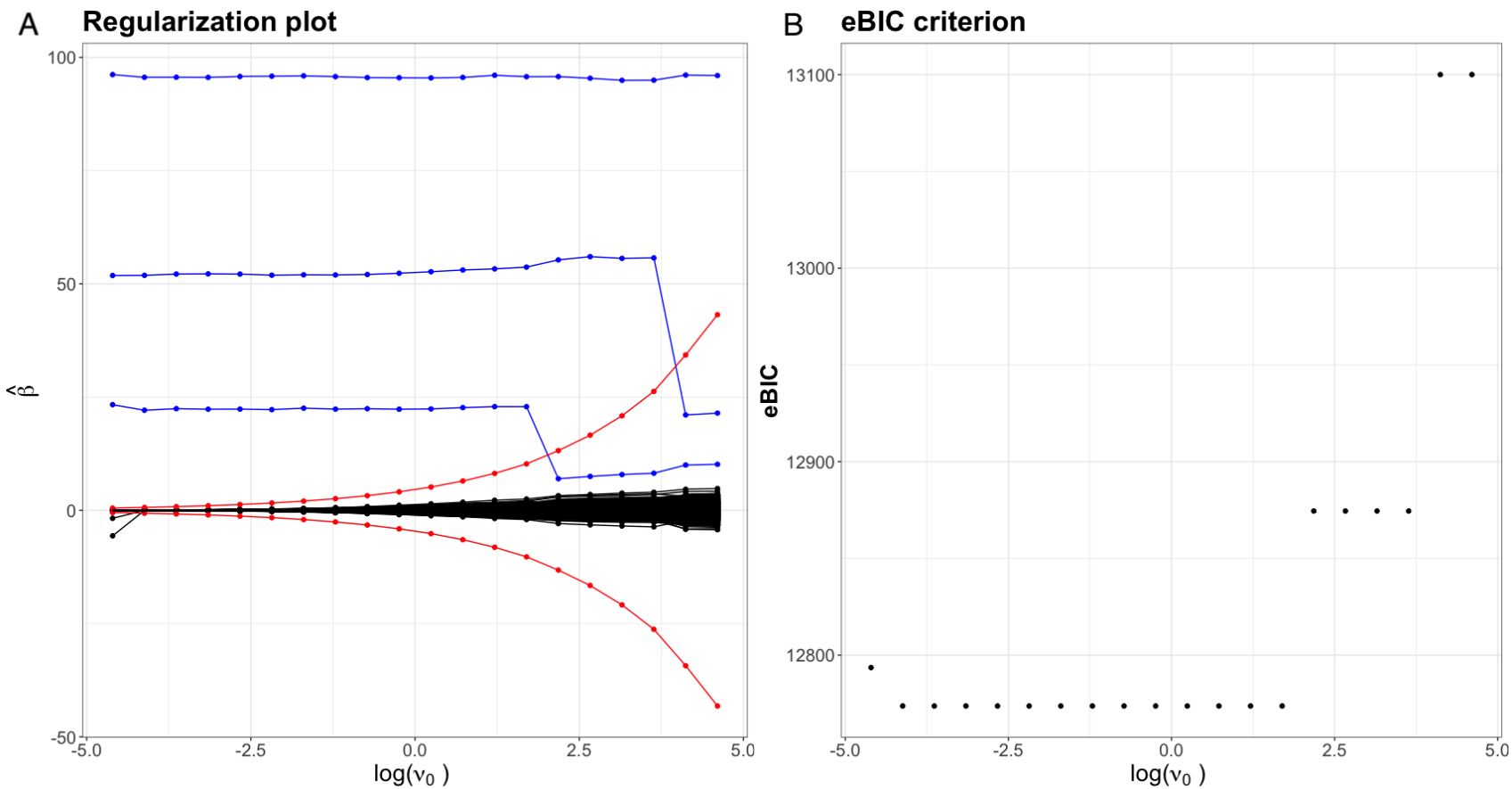


Figure – $n = 200$, $J = 10$, $p = 500$, $\Gamma^2 = 200$, $\sigma^2 = 30$, $\nu_1 = 12000$, $\mu = 1200$, $\beta = {}^t(100, 50, 20, 0, \ldots, 0)$

Calculer le MAP dans un modèle à variables latentes

✤ Objectif : maximiser $\pi(\Theta|y) = \int_{\mathcal{Z}} \pi(\Theta, Z|y)dZ$ avec

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

- Intégrale non explicite

Algorithme EM [Dempster et al., 1977]

1. Initialisation : choisir $\Theta^{(0)}$.
2. Itération $k \geq 0$:
   - **Étape E (Expectation) :** calculer

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

   - **Étape M (Maximisation) :**

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

3. $\hat{\Theta} = \Theta^{(K)}$, pour K assez grand.

Algorithme MCMC-SAEM [Delyon et al., 1999] [Kuhn and Lavielle, 2004]

1. Initialisation : choisir $\Theta^{(0)}$ et $Q_0(\Theta) = 0$.
2. Itération $k \geq 0$:
   - **Étape S (Simulation) :** simuler $Z^{(k)}$ selon $\pi(Z|y, \Theta^{(k)})$ par une itération d'une procédure MCMC.
   - **Étape SA (Stochastic Approximation) :** mettre à jour $Q_{k+1}(\Theta)$ approximation de $Q(\Theta|\Theta^{(k)})$ selon :

$$Q_{k+1}(\Theta) = Q_k(\Theta) + \gamma_k(\log \pi(\Theta, Z^{(k)}|y) - Q_k(\Theta)).$$

   - **Étape M (Maximisation) :**

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

3. $\hat{\Theta} = \Theta^{(K)}$, pour K assez grand.

où $(\gamma_k)_k$ est une suite de pas décroissante vers 0 telle que $\forall k,\ \gamma_k \in [0, 1]$, $\sum_k \gamma_k = \infty$ et $\sum_k \gamma_k^2 < \infty$.

Application à notre modèle

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

✤ Il suffit d'appliquer :

- un EM à $\widetilde{Q}_2(\alpha,\Theta^{(k)})$.
- un MCMC-SAEM à $\mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)})\middle| y,\Theta^{(k)}\right]$.

Algorithme MCMC-SAEM dans notre modèle

1. Initialisation : choisir $\Theta^{(0)}$ et $Q_0(\theta) = 0$.
2. Itération $k \geq 0$:
   - **Étape S (Simulation) :** simuler $\varphi^{(k)}$ selon $\pi(\varphi|y, \Theta^{(k)})$ par une itération d'une procédure MCMC.
   - **Étape SA (Stochastic Approximation) :** mettre à jour $Q_{k+1}(\theta)$ approximation de $\mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y, \varphi, \theta, \Theta^{(k)}) \middle| y, \Theta^{(k)}\right]$ selon :

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

   - **Étape M (Maximisation) :**

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

3. $\hat{\Theta} = \Theta^{(K)}$, pour K assez grand.

où $(\gamma_k)_k$ est une suite de pas décroissante vers 0 telle que $\forall k,\ \gamma_k \in [0, 1]$, $\sum_k \gamma_k = \infty$ et $\sum_k \gamma_k^2 < \infty$.

Modèle de croissance logistique

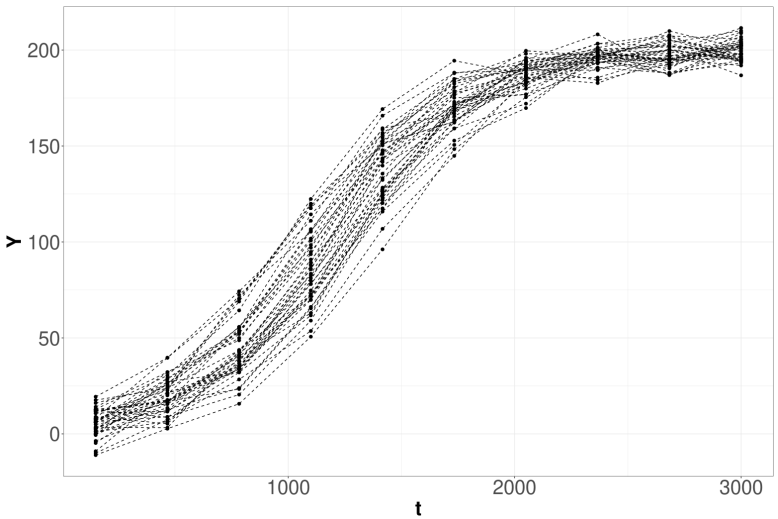


Figure – Données simulées

- Taille de la plante $i \in \{1, \dots, n\}$ au temps t_{ij}, $j \in \{1, \dots, 10\}$:

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

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

  $\psi = (\psi_1, \psi_2)$ effets fixes.

- φ_i : temps caractéristique

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

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

Plan de simulation

✤ **Paramètres :**

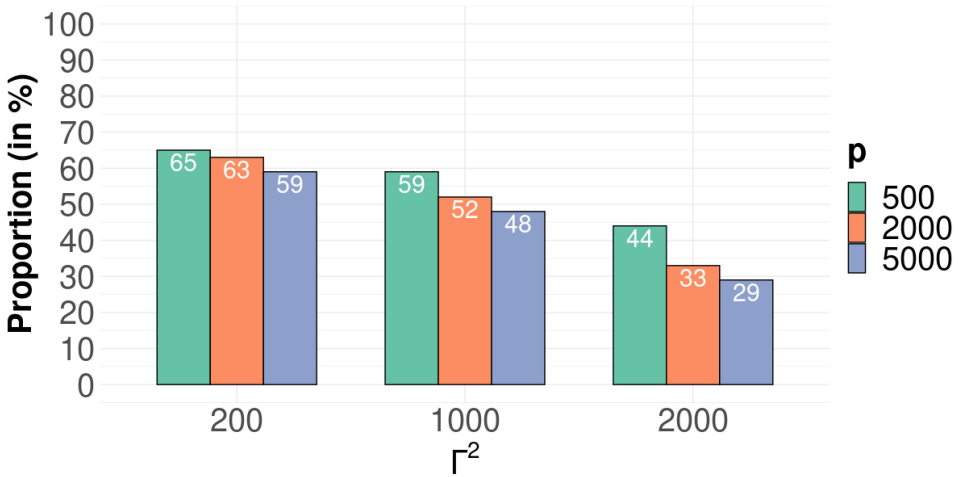
- $n \in \{100, 200\}$ individus,
- $p \in \{500, 2000, 5000\}$ covariables simulées selon $V_i \sim \mathcal{N}(0, \Sigma)$,
- $\beta = {}^t(100, 50, 20, 0, \ldots, 0)$ vecteur d'effets des covariables,
- $\Gamma^2 \in \{200, 1000, 2000\}$ variance inter-individuelle,
- $\mu = 1200$, $\sigma^2 = 30$, $\psi = (\psi_1, \psi_2) = (200, 300)$.

✤ **Hyperparamètres :**

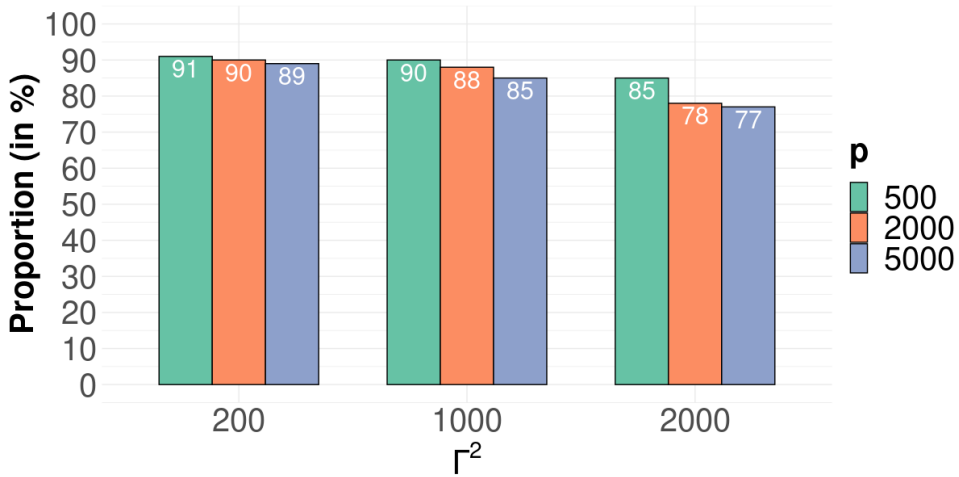
- $\nu_1 = 12000$ variance slab,
- $\log_{10}(\Delta) = \left\{ -2 + k \times \frac{4}{19}, k \in \{0, \ldots, 19\} \right\}$ grille de ν_0.

▶ Pour chaque combinaison de paramètres, on applique la méthode sur 100 jeux de données simulées différents.

Résultats cas covariables indépendantes



(a) Pour $n = 100$



(b) Pour $n = 200$

Figure – Proportion de datasets pour lesquels il y a sélection du bon modèle.

Analyse des résultats

- Covariables indépendantes $V_i \sim \mathcal{N}(0, I_p)$:
  - Amélioration des résultats quand n augmente.
  - Dégradation des résultats quand p ou Γ^2 augmente.
  - Quand la méthode se trompe c'est majoritairement à cause d'une sous-sélection.
- Covariables corrélées $V_i \sim \mathcal{N}(0, \Sigma)$:
  - Performances assez similaires.
  - Plus de faux positifs et/ou de faux négatifs dans certains scénarios de corrélations :
    - ▶ + de faux positifs : corrélations entre covariables actives et non-actives.
    - ▶ + de faux négatifs : covariables actives corrélées.

Conclusion et perspectives

✤ Conclusion :

- Développement d'une méthode originale qui mêle SAEM et sélection de variables bayésienne.
- Résultats numériques sur données simulées très encourageants.
- Méthode plus rapide qu'une implémentation MCMC complète.

✤ Perspectives :

- Apporter des garanties théoriques.
- Appliquer notre méthode sur un jeu de données réelles.
- Considérer un paramètre individuel multidimensionnel.

Remerciements

Merci de votre attention !

Références I


Bertrand, J. and Balding, D. J. (2013).
Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models.
Pharmacogenetics and genomics, 23(3) :167–174.


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.

Références II


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.


Lee, S. Y. (2022).
Bayesian Nonlinear Models for Repeated Measurement Data : An Overview, Implementation, and Applications.
Publisher : Preprints.


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.

Références III


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.

Sélection de modèle

- **Répéter SAEM seuillé** sur une grille Δ de valeurs pour ν_0
  - ▶ génère un chemin solution $\{S_{\nu_0}, \nu_0 \in \Delta\}$
- **Choisir** ν_0 tel que : ([Chen and Chen, 2008])

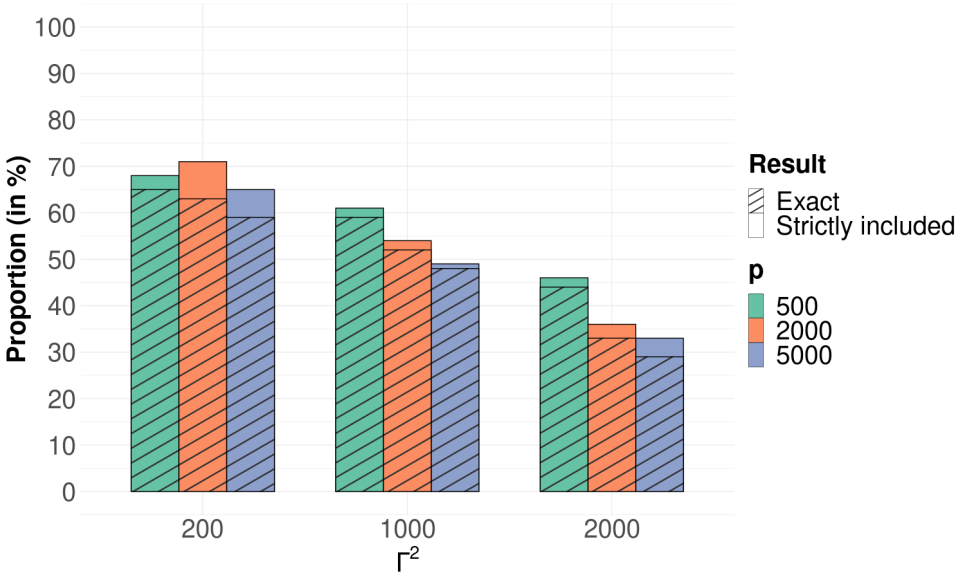
$$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\text{argmin}} \left\{ eBIC(\nu_0) \right\}$$

avec $eBIC(\nu_0) = -2\log\left(p(y; \hat{\theta}^{EMV}_{\nu_0})\right) + \text{pen}(\nu_0)$, où :

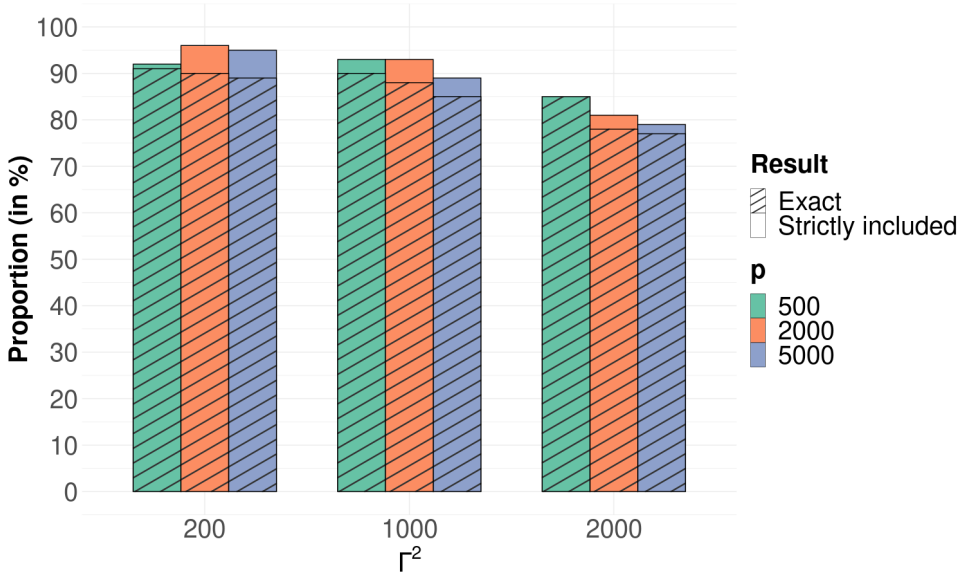
- $\log\left(p(y;\theta)\right)$: log-vraisemblance du modèle
- $\hat{\theta}^{EMV}_{\nu_0} = (\hat{\mu}^{EMV}_{\nu_0}, \hat{\beta}^{EMV}_{\nu_0}, \hat{\sigma}^{2^{EMV}}_{\nu_0}, \hat{\Gamma}^{2^{EMV}}_{\nu_0})$: EMV de θ dans le sous-modèle S_{ν_0} sélectionné par SAEM seuillé pour ce ν_0 (MCMC-SAEM)
- $\text{pen}(\nu_0) = B_{\nu_0} \times \log(n) + 2\log\left(\binom{p}{B_{\nu_0}}\right)$: fonction de pénalité, avec B_{ν_0} le nombre de paramètres libres de S_{ν_0}

$$S_{\hat{\nu}_0} = \left\{ V_\ell \mid |\hat{\beta}^{MAP,\hat{\nu}_0}_\ell| \geq s_\beta(\hat{\nu}_0, \nu_1, \hat{\alpha}^{MAP}_{\hat{\nu}_0}) \right\}$$

Résultats



(a) Pour $n = 100$

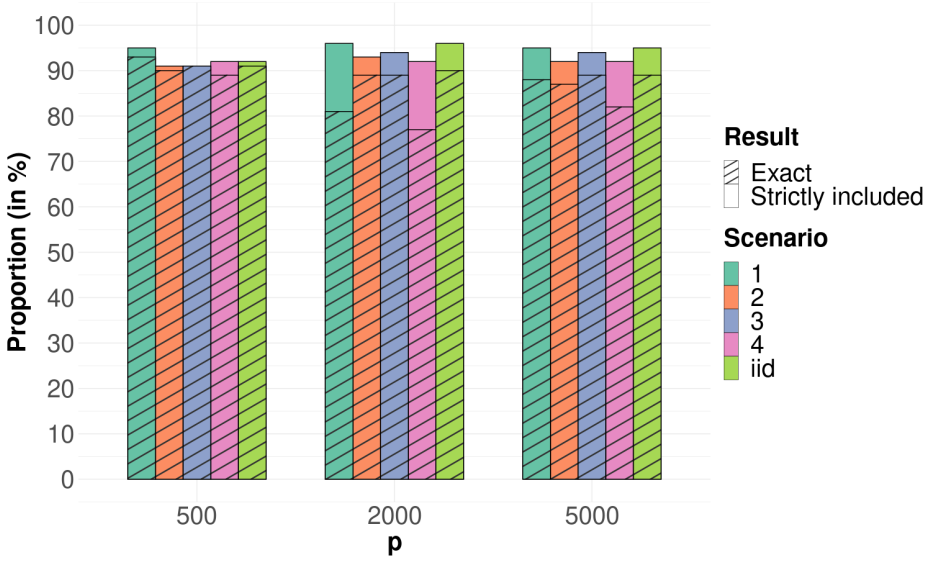


(b) Pour $n = 200$

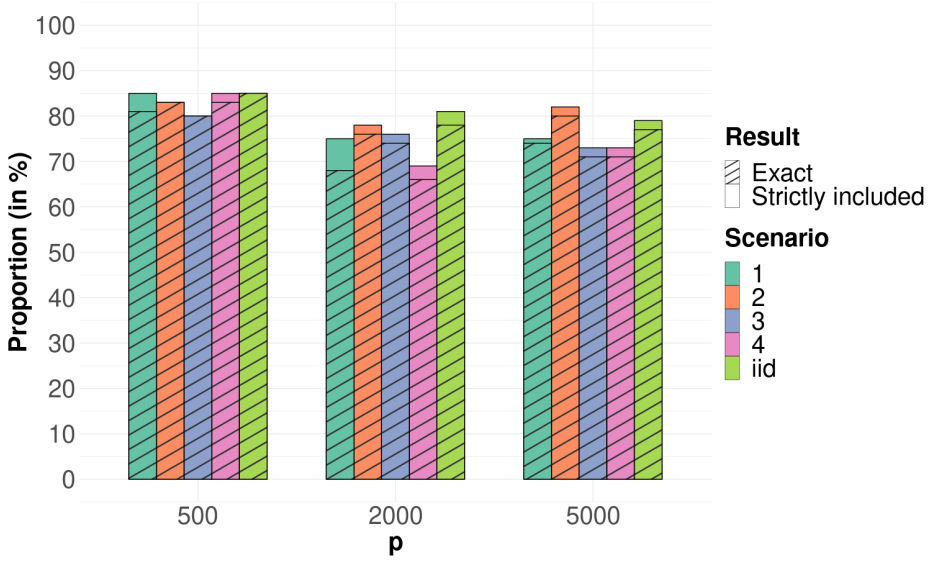
Covariables corrélées $V_i \sim \mathcal{N}(0, \Sigma)$

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

Résultats pour $\rho_\Sigma = 0.3$

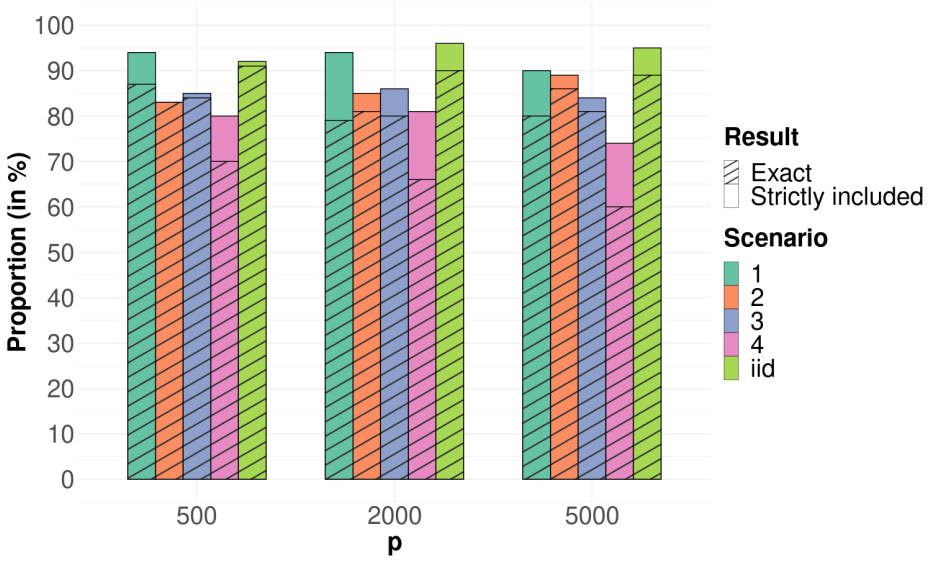


(c) Pour $\Gamma^2 = 200$

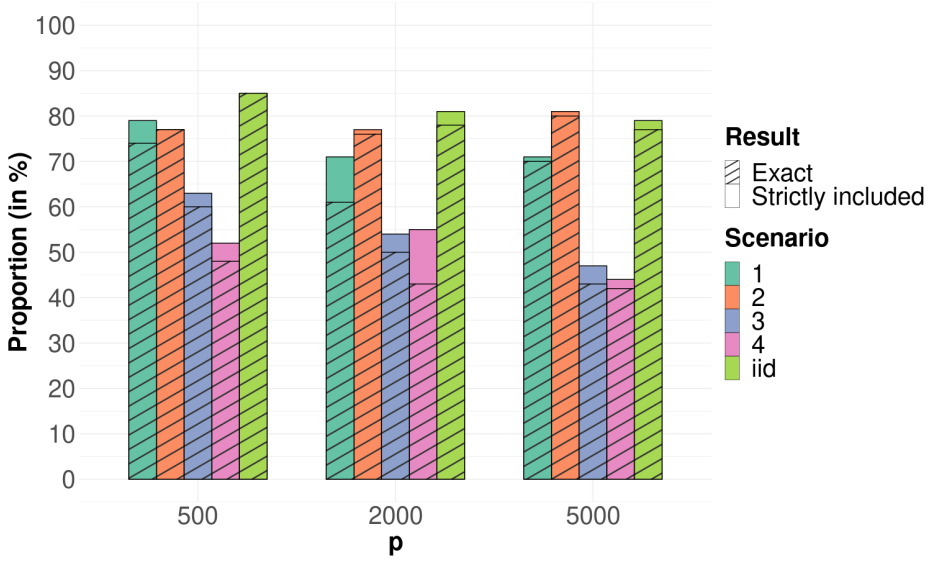


(d) Pour $\Gamma^2 = 2000$

Résultats pour $\rho_\Sigma = 0.6$



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



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

Comparaison avec une implémentation MCMC