



HAL
open science

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

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

► To cite this version:

Marion Naveau, Guillaume Kon Kam King, Laure Sansonnet, Maud Delattre. High-dimensional variable selection in non-linear mixed effects models using a stochastic EM spike-and-slab. ISBA World Meeting, Jun 2022, Montréal, Canada. . hal-04248003

HAL Id: hal-04248003

<https://hal.inrae.fr/hal-04248003>

Submitted on 18 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Marion Naveau^{1,2} & Maud Delattre² & Guillaume Kon Kam King² & Laure Sansonnet¹

¹ Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, France. ² Université Paris-Saclay, INRAE, MaIAGE, France.

1 - Introduction

Mixed-effects models:

- Analyse observations collected repeatedly on several individuals.
- Individuals with the same overall behaviour but with individual variations.
- Different sources of variability: intra-individual, inter-individual, residual.

Fields of application: pharmacokinetics, biological growth, ...

Variable selection:

- Inter-individual variability may be explained by some among a very large number of covariates (*e.g.* genomic data).
- High-dimensional context: focus on the few most relevant covariates through a variable selection procedure.

2 - Non-linear mixed-effects model (NLMEM)

For $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, J\}$, denoting y_{ij} the response of individual i at time t_{ij} and V_i the p covariates measured on individual i , with $p \gg n$:

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

where

- $\varphi_i \in \mathbb{R}$: individual parameter, not observed
- g : **non-linear** function with respect to φ_i
- $\psi \in \mathbb{R}^q$: fixed effects, unknown
- $\mu \in \mathbb{R}$: intercept, unknown
- $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$: covariate fixed effects vector, unknown

Population parameter to be estimated: $\theta = (\mu, \beta, \psi, \sigma^2, \Gamma^2)$

3 - Aim and contribution

- Aim:** Identify the most relevant covariates to characterise inter-individual variability, *i.e.* identify the non-zero components of β .
- Main difficulties:** non-explicit likelihood and high-dimensional problem.
- Proposed approach:** Association of a Bayesian **spike-and-slab prior** for variable selection with **MCMC-SAEM** algorithm (stochastic version of EM) for inference [4].

4 - Bayesian hierarchical model

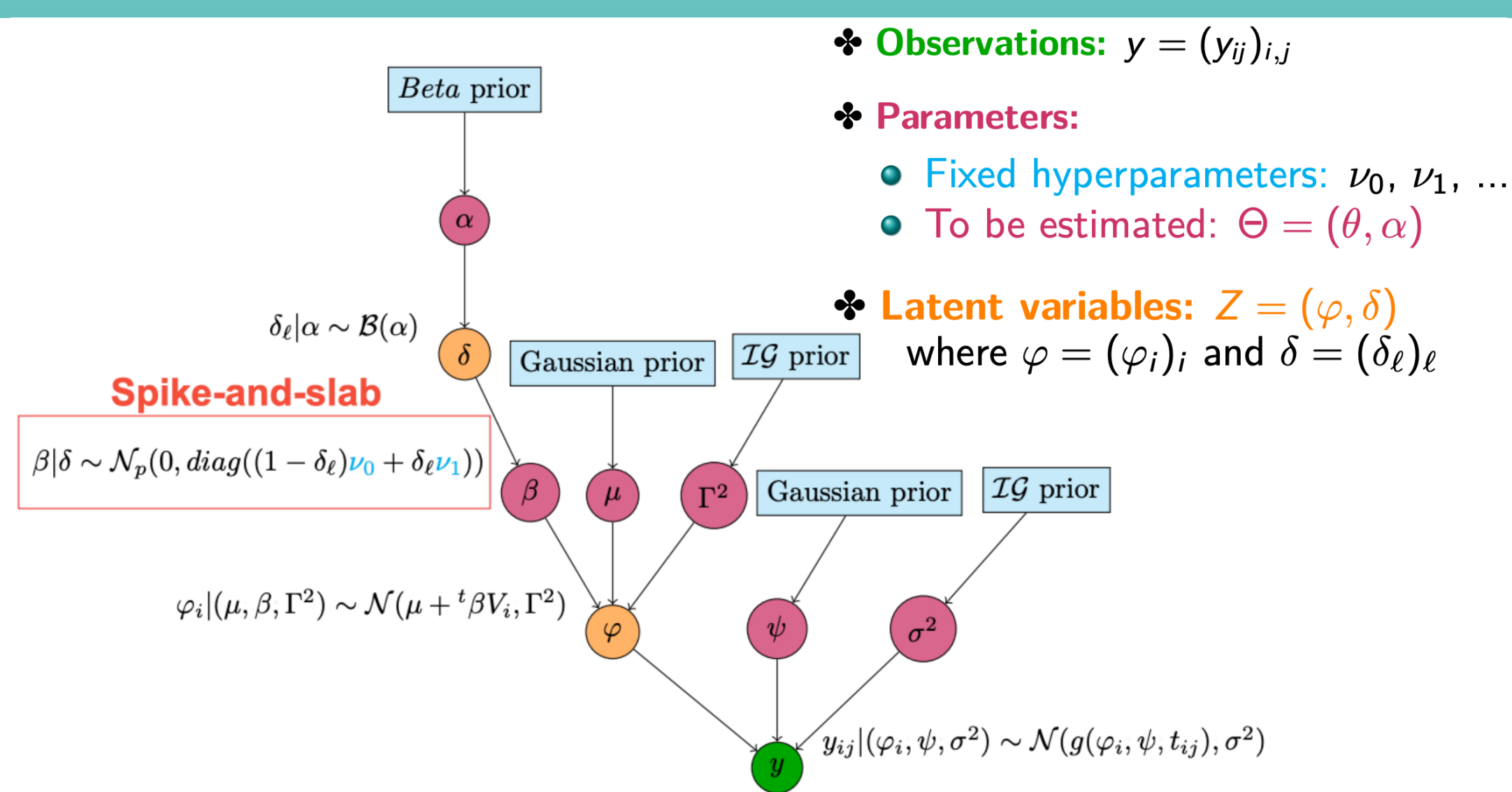


Fig. 1: Bayesian hierarchical model

5 - Method

Idea: we could choose ν_0 small, deduced from a practitioner chosen threshold for "negligible" covariate effect. However, we may be interested in exploring 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 the maximum *a posteriori* estimator with a MCMC-SAEM algorithm [1]:

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

- estimate $\hat{\delta}$ to find good models with high posterior probability [2]:

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

$$\iff \text{Define } \hat{S}_{\nu_0} = \left\{ \ell \in \{1, \dots, p\} \mid |(\hat{\beta}_{\nu_0}^{MAP})_\ell| \geq s_\beta(\nu_0, \nu_1, \hat{\alpha}_{\nu_0}^{MAP}) \right\}$$

2. **Select the "best" model** among $(\hat{S}_{\nu_0})_{\nu_0 \in \Delta}$ by a fast criterion, *e.g.* eBIC [3]:

$$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\text{argmin}} \left\{ -2 \log(p(y; \hat{\theta}_{\nu_0}^{MLE})) + B_{\nu_0} \times \log(n) + 2 \log\left(\binom{p}{B_{\nu_0}}\right) \right\}$$

with B_{ν_0} the number of free parameters in the sub-model \hat{S}_{ν_0} .

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

[1] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. ESAIM: Probability and Statistics.

[2] Rocková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. Journal of the American Statistical Association.

[3] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. Biometrika.

[4] Naveau, M. et al. (2022). Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm. arXiv:2206.01012.

6 - Regularisation plot and eBIC criterion

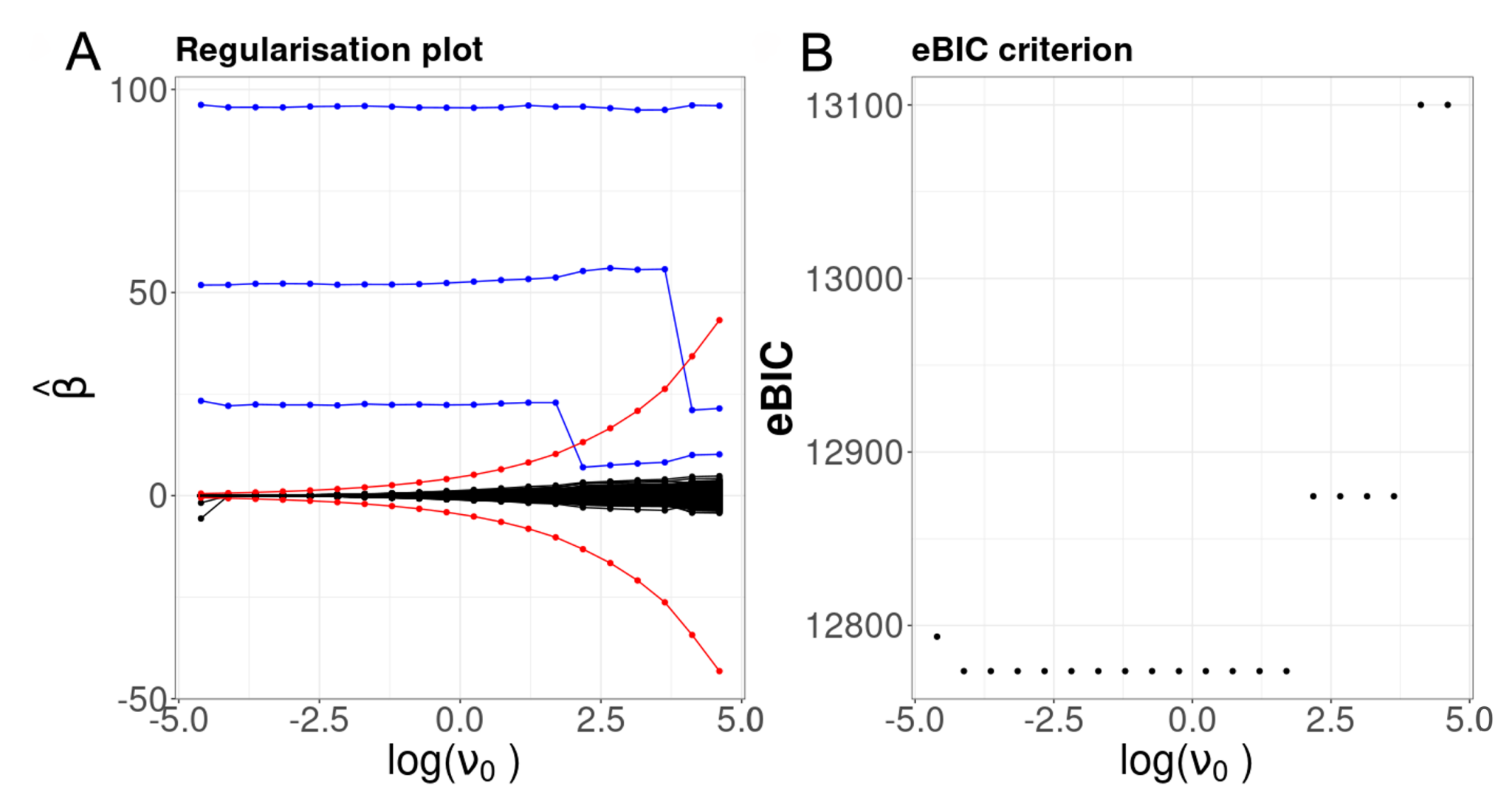


Fig. 2: Example of a regularisation plot (A) with eBIC criterion graph (B) for model selection. On (A), the red lines correspond to the selection threshold of the covariates.

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

7 - MCMC-SAEM algorithm for computing the MAP

At each step k of this iterative algorithm, the idea is to maximise:

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

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 $\tilde{Q}_2(\alpha, \Theta^{(k)})$ and $Q_{1,k+1}(\theta)$, approximation of $\mathbb{E}_{\varphi | y, \Theta^{(k)}} \left[\tilde{Q}_1(y, \varphi, \theta, \Theta^{(k)}) \mid y, \Theta^{(k)} \right]$, according to:

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

- M-step (Maximisation):** compute

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

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

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

8 - Simulation results in a logistic growth model

Uncorrelated covariates:

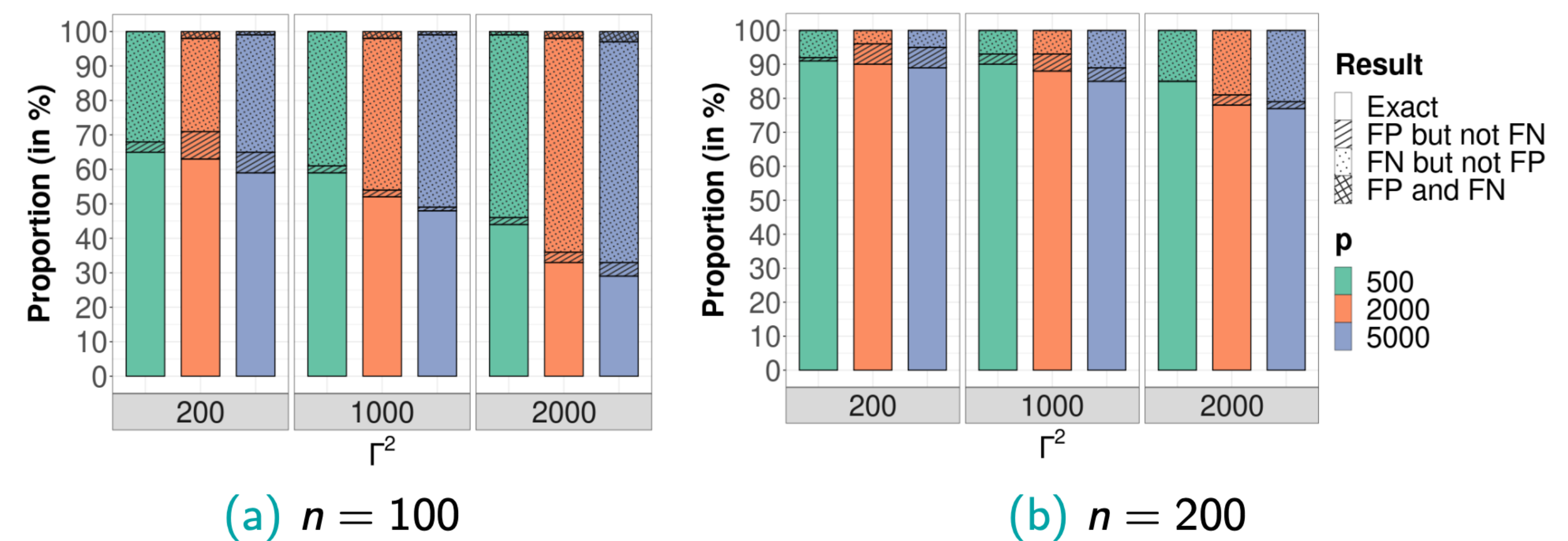


Fig. 3: Proportion of data-sets on which the proposed method selects the correct model ("Exact"), a model that contains false positives (FP) but not false negatives (FN), FN but not FP, or FP and FN.

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

- The proposed method is about 20 times faster than a full MCMC implementation.

9 - Perspectives

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