



**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, Renaud Rincent, Laure Sansonnet, Maud Delattre

## ► To cite this version:

Marion Naveau, Guillaume Kon Kam King, Renaud Rincent, Laure Sansonnet, Maud Delattre. High-dimensional variable selection in non-linear mixed-effects models using a stochastic EM spike-and-slab. European Meeting of Statisticians, Bernoulli Society, Jul 2023, Warsaw, Poland. hal-04248037

**HAL Id: hal-04248037**

**<https://hal.inrae.fr/hal-04248037>**

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 nonlinear mixed effects models using a stochastic EM spike-and-slab

Marion Naveau<sup>1</sup>, Guillaume Kon Kam King<sup>1</sup>, Renaud Rinent<sup>1</sup>,  
Laure Sansonnet<sup>1</sup>, Maud Delattre<sup>1</sup>

<sup>1</sup>Université Paris-Saclay (France)

High-dimensional variable selection is widely documented in standard regression models, but there are still few tools to address it in nonlinear mixed-effects models, where data is collected repeatedly on several individuals. For all  $1 \leq i \leq n$  and  $1 \leq j \leq n_i$ ,  $y_{ij}$  the response of individual  $i$  at time  $t_{ij}$  is modelled as follows:

$$y_{ij} = g(\varphi_i, t_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$
$$\varphi_i = \mu + \boldsymbol{\beta}^\top V_i + \xi_i, \quad \xi_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}_q(0, \Gamma),$$

where  $g$  is non-linear with respect to an individual parameter  $\varphi_i$  which is  $q$ -dimensional. Identifying the relevant covariates for all individual parameters amounts to selecting the support of  $\boldsymbol{\beta} \in \mathcal{M}_{p \times q}$ , defined by  $S^* = \left\{ (\ell, m) \in \{1, \dots, p\} \times \{1, \dots, q\} \mid \beta_{\ell m}^* \neq 0 \right\}$ , where  $\boldsymbol{\beta}^*$  is the true fixed effects matrix, and  $p$  is the number of covariates. To solve this problem in a high-dimensional context, that is when  $p \gg n$ , the assumption of sparsity is made, that is each row of  $\boldsymbol{\beta}^*$  is sparse. The main difficulty here is that variable selection concerns latent variables of the model.

In this work, variable selection is approached from a Bayesian perspective and a selection procedure is proposed, combining the use of a spike-and-slab prior [3] and the SAEM algorithm [2]. The first step of this procedure is to reduce the number of candidate models: similarly to Lasso regression, a grid of values for the spike parameter is explored to obtain a collection of promising sub-models  $(\hat{S}_{\nu_0})_{\nu_0 \in \Delta}$ . Then, an information criterion can be used to choose the final model: a pragmatic and effective choice is to use the eBIC (extended Bayesian Information Criterion, [1]) which is tailored to the high-dimensional setting.

This approach is much faster than a classical MCMC algorithm and shows very good selection performances on simulated data. The efficiency of the proposed method is illustrated on a problem of genetic markers identification, relevant for genomic assisted selection in plant breeding. The current aim is to achieve consistency in model selection for this problem, which is a work in progress.

## References

- [1] Chen J., Chen Z., *Extended Bayesian information criteria for model selection with large model spaces*, *Biometrika*, 95 (2008), 759-771.
- [2] Delyon B., Lavielle M., Moulines E., *Convergence of a stochastic approximation version of the EM algorithm*, *Annals of statistics* (1999), 94-128.
- [3] George, E. I., McCulloch R. E., *Approaches for Bayesian variable selection*, *Statistica sinica*, 7 (1997), 339-373.