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

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

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

Maud Delattre
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.

Guillaume Kon Kam King
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.

Marion Naveau
Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 75005, Paris, France.
Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.

Laure Sansonnet
Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 75005, Paris, France.

June 1, 2022

Abstract

High-dimensional data, with many more covariates than observations, such as genomic data for example, are now commonly analysed. In this context, it is often desirable to be able to focus on the few most relevant covariates through a variable selection procedure. High-dimensional variable selection is widely documented in standard regression models, but there are still few tools to address it in the context of non-linear mixed-effects models. In this work, variable selection is approached from a Bayesian perspective and a selection procedure is proposed, combining the use of spike-and-slab priors and the SAEM algorithm. Similarly to LASSO regression, the set of relevant covariates is selected by exploring a grid of values for the penalisation parameter. The proposed approach is much faster than a classical MCMC algorithm and shows very good selection performances on simulated data.

Keywords Bayesian variable selection · Non-linear mixed-effects models · High-dimensional data · Spike-and-slab prior · SAEM algorithm

1 Introduction

Mixed-effects models have been introduced to analyse observations collected repeatedly on several individuals in a population of interest (Lavielle, 2014; Pinheiro and Bates, 2000). This type of data is particularly common in the fields of pharmacokinetics or when modelling biological growth for example. In this case, the intrinsic variability of the data is then attributable to different sources (intra-individual, inter-individual, residual) whose consideration is essential to characterise without bias the biological mechanisms behind the observations. Mixed-effects models allow the study of the responses of individuals with the same overall behaviour but with individual variations characterised by random individual parameters that are not observed. Thus, mixed-effects models are latent variables models. Parameter inference is therefore difficult because the likelihood and classical estimators do not have an explicit form. A widely used solution is to use an EM

(Expectation-Maximisation) algorithm, or any variant, to compute the maximum likelihood estimator or the maximum *a posteriori* estimator in a Bayesian framework (Dempster et al., 1977).

Moreover, the description of inter-individual variability may involve a number of covariates much larger than the number of individuals. In this high-dimensional context, it is often desirable to be able to focus on the few most relevant covariates through a variable selection procedure. However, in mixed-effects models, identifying the influential covariates is difficult, as the selection concerns latent variables in the model. Recent years have seen the emergence of varied contributions on high-dimensional covariate selection in mixed-effects models. The proposed tools are very different according to whether the regression function is linear or non-linear with respect to the individual parameters. More precisely, the linear case allows the development of criteria whose calculation and/or theoretical study involve explicit quantities, which is rarely the case when the model is non-linear. In linear mixed-effects models, many rely on the use of regularised methods (see Schelldorfer et al. (2011) and Fan and Li (2012) for example) and most of them include theoretical consistency results that guarantee the good properties of the proposed methods. In the more general framework of non-linear mixed-effects models (NLMEM), on the other hand, there are few results and the only published works concern computational aspects. Bertrand and Balding (2013) compare a stepwise approach using an empirical Bayes estimate, and penalised regression approaches like Ridge, LASSO and HyperLASSO penalties, and Ollier (2021) proposes a proximal gradient algorithm for computing a LASSO estimator. To our knowledge, these are the only contributions that handle high-dimensional variable selection in NLMEM.

Bayesian approaches to variable selection have not received a lot of attention in the NLMEM context. The focus for their development has been classical statistical models like linear regression or generalised linear model, for which Bayesian variable selection has been intensively developed in recent years. These methods encourage sparsity in the regression vector by using a variety of priors (see for example Tadesse and Vannucci (2021) and the references therein) which may have better properties than the double-exponential prior associated with the LASSO penalty. Very recently, Lee (2022) proposed an overview of the formulation, interpretation and implementation of Bayesian non-linear mixed-effects models. In particular, he discussed Bayesian inference methods, priors options, and model selection methods in this context. However, these Bayesian approaches are based on Markov Chain Monte-Carlo (MCMC) methods which seldom scale well enough to be usable for high-dimensional variable selection. The main objective of this paper is to propose a fast Bayesian spike-and-slab approach that can be used to identify the relevant covariates in a non-linear mixed-effects model, in a high-dimensional context. More precisely, we extend the EMVS approach of Ročková and George (2014) to the NLMEM setting. Like EMVS, the proposed approach involves two major steps. The first step is, for different values of the spike hyperparameter, to select a local version of the median probability model (Barbieri and Berger, 2004) using the Stochastic Approximation version of the EM algorithm (SAEM, see Delyon et al. (1999) and Kuhn and Lavielle (2004)). The second step consists in selecting the "best" model among those kept after the first step, using an extension of the BIC criterion (Chen and Chen, 2008). An important difference with Ročková and George (2014) is that our approach is applied to NLMEM and not to classical linear regression models. Due to the model non-linearity and to the latent nature of the model random effects, the central so-called Q-quantity of the EM algorithm often does not have a closed form expression and posterior distributions are difficult to compute. To overcome these issues, we propose an inference method using the SAEM algorithm, rather than simply the EM algorithm as in Ročková and George (2014). Another important difference is that optimal model selection among the sub-models obtained in the first step does not require the calculation of the marginal posterior of the models for a spike parameter being equal to 0, as in Ročková and George (2014), but only of the log-likelihood of the NLMEM taken at the maximum likelihood estimator.

The plan of this article is as follows. Section 2 describes the non-linear mixed-effects model to fix the notations, summarises the main objective of our procedure, and defines and motivates the hierarchical prior formulations. Section 3 details the key tools for our approach: the SAEM algorithm, to compute the maximum *a posteriori* estimator of the model parameters, and a thresholding rule to select a local version of the median probability model and put some coefficients of the regression vector to zero. Next, Section 4 describes the variable selection procedure. Section 5 evaluates the selection performance of our method through an intensive simulation study, and presents a comparison with the classical MCMC methods. Finally, Section 6 concludes with a summary discussion and prospects for future research.

2 Statistical model

2.1 Mixed-effects model and notations

The formalism used in this paper is that of Lavielle (2014) and Pinheiro and Bates (2000). Let n be the number of individuals and n_i the number of observations for individual i. Without loss of generality and to lighten the calculations, we assume that $n_i = J$ for all $1 \leq i \leq n$. Consider the following non-linear mixed-effects model: for all $1 \leq i \leq n$ and $1 \leq j \leq J$,

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

where ${}^{\text{t}}X$ denotes the transpose of a vector or matrix X. This model is described on two levels. First, at the individual level, Equation (1a) describes the intra-individual variability, where the observations y_{ij} in $\mathbb{R}$ represent the response of individual i at time t_{ij}. It is assumed that all individuals follow the same known functional form g which depends non-linearly on an individual parameter φ_i which is assumed to be real in this work. Thus, this function governs the intra-individual behaviour. The variance $\sigma^2 > 0$ of the Gaussian measurement noise is assumed unknown. Then, at the population level, Equation (1b) describes the inter-individual variability. For all $i \in \{1, \ldots, n\}$, the individual parameter φ_i is modelled as a Gaussian random variable whose mean is specified as the sum of an intercept μ in $\mathbb{R}$ and a linear combination of known covariates measured on individual i and contained in the vector $V_i = {}^{\text{t}}(V_{i1}, \ldots, V_{ip}) \in \mathbb{R}^p$. The term "covariates" refers to explanatory variables which may be relevant to explain inter-individual variability. This term is used to distinguish them from other explanatory variables such as the time variable for example. The number of covariates is denoted by p, $\beta = {}^{\text{t}}(\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ is an unknown covariate fixed effects vector, and the inter-individual variance $\Gamma^2 > 0$ is assumed unknown. Thus, the inter-individual variability is separated into two parts: on the one hand, β models the variability that can be explained by the covariates V_i, and on the other hand, ξ_i represents the part of variation that is not explained by the measured covariates. In the following, $y_i = (y_{ij})_{1 \leq j \leq J}$, $y = (y_i)_{1 \leq i \leq n}$ and $\varphi = (\varphi_i)_{1 \leq i \leq n}$ respectively denote the vector of observations for individual i, the vector of all observations and the vector of all individual parameters. Let us also note $\theta = (\mu, \beta, \Gamma^2, \sigma^2)$ the unknown parameter, also-called the population parameter.

The goal of the present work is to identify the relevant covariates, *i.e.* those that best explain the variability between individuals. This can be framed as identifying the non-zero elements in β. Indeed, for $1 \leq \ell \leq p$, the parameter β_ℓ describes the effect of the covariate ℓ on the individual parameter. More precisely, $\beta_\ell = 0$ means that the covariate ℓ has no effect on the individual parameter φ_i and $\beta_\ell \neq 0$ means that the covariate ℓ gives some information on this parameter. Identifying the relevant covariates amounts to selecting the support of β, noted S^*_β:

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

where β^* is the true covariate fixed effects vector. To solve this problem in a high-dimensional context, that is when $p >> n$, it is natural to assume that the vector β^* is sparse, which means that many β^*_ℓ are zero. An important point here is that model (1) is a model with incomplete-data. Indeed, although the first layer (1a) is observed, it is not the case for the individual parameters φ. The main difficulty here is that variable selection concerns latent variables of the model.

2.2 Prior specification

To solve this variable selection problem, it is convenient to adopt a Bayesian approach. The purpose of this section is to describe the prior distribution on $\theta = (\mu, \beta, \Gamma^2, \sigma^2)$. First, in order to find the non-zero coefficients of β, a spike-and-slab mixture prior (George and McCulloch, 1993, 1997; Ročková and George, 2014) is considered. To facilitate the formulation of this prior, a vector of binary latent variables $\delta = (\delta_\ell)_{1 \leq \ell \leq p}$ is introduced, such as:

$$\forall 1 \leq \ell \leq p \ , \ \delta_\ell = \begin{cases} 1 & \text{if covariate } \ell \text{ is to be included in the model,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Thus, $\delta_\ell = 1$ indicates that the covariate ℓ provides information on the individual parameter. In other words, δ characterises the support of β. The support S^*_β can therefore be reformulated as follows:

$$S^*_\beta = \left\{ \ell \in \{1, \ldots, p\} \middle| \delta^*_\ell = 1 \right\}, \tag{3}$$

where δ^* denotes the true support. Then, one would like to find $\widehat{\delta}$ that maximises the posterior probability $\pi(\delta|y)$. While it is possible to sample from the posterior distribution using Monte Carlo Markov Chain methods, for computational efficiency we are particularly interested in obtaining an estimator $\widehat{\delta}$ corresponding to the most promising support, that is the one which is the most compatible with the data and the prior distribution.

The prior formulations proposed here are based on the non-conjugate version of the hierarchical priors of George and McCulloch (1997), summarised as follows:

$$\pi(\beta|\delta) = \mathcal{N}_p(0, D_\delta), \text{ where } D_\delta = \text{diag}(a_1, \ldots, a_p) \text{ with } a_\ell = (1-\delta_\ell)\nu_0 + \delta_\ell\nu_1,\ 0 < \nu_0 < \nu_1, \tag{4a}$$

$$\pi(\mu) = \mathcal{N}(0, \sigma_\mu^2), \text{ with } \sigma_\mu^2 > 0, \tag{4b}$$

$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma\lambda_\sigma}{2}\right), \text{ with } \nu_\sigma, \lambda_\sigma > 0, \tag{4c}$$

$$\pi(\Gamma^2) = \mathcal{IG}\left(\frac{\nu_\Gamma}{2}, \frac{\nu_\Gamma\lambda_\Gamma}{2}\right), \text{ with } \nu_\Gamma, \lambda_\Gamma > 0, \tag{4d}$$

$$\pi(\delta|\alpha) = \alpha^{|\delta|}(1-\alpha)^{p-|\delta|}, \text{ with } \alpha \in [0,1] \text{ and } |\delta| = \sum_{\ell=1}^{p}\delta_\ell, \tag{4e}$$

$$\pi(\alpha) = Beta(a, b), \text{ with } a, b > 0. \tag{4f}$$

The key prior distribution used for variable selection in this method is the spike-and-slab Gaussian mixture prior (4a) on β. In this prior, ν_0 and ν_1 are parameters controlling the penalisation inducing sparsity in the vector β. More precisely, $(\beta_\ell)_{1\leq\ell\leq p}$ are independent conditionally on δ, with $\pi(\beta_\ell|\delta_\ell = 0) = \mathcal{N}(0, \nu_0)$ and $\pi(\beta_\ell|\delta_\ell = 1) = \mathcal{N}(0, \nu_1)$. The general recommendation for this type of prior is to set ν_0 small, to encourage the exclusion of insignificant effects, and ν_1 large enough to accommodate all plausible β values (see George and McCulloch, 1997). Indeed, when $\delta_\ell = 0$, the prior constrains β_ℓ to very small values which implies that covariate ℓ has no impact in the model. Thus, through the values δ_ℓ, the spike-and-slab prior makes it possible to distinguish the selected covariates from the rest.

Note that, since φ is unobserved, one cannot simply centre the variable on which the selection is made as is usually the case in more standard models, and so inclusion of an intercept μ is necessary. Thus, a vaguely informative Gaussian prior (4b) is used for μ, with σ_μ^2 large enough. This choice of prior has the advantage of simplifying the calculations for parameter inference, thanks to a useful reformulation $\tilde{\beta} = {}^{\text{t}}(\mu, \beta) \in \mathbb{R}^{p+1}$ and, for all $1 \leq i \leq n$, $\tilde{V}_i = (\tilde{V}_{i\ell'})_{0\leq\ell'\leq p} = {}^{\text{t}}(1, V_i) \in \mathbb{R}^{p+1}$, so that $\mu + {}^{\text{t}}\beta V_i = {}^{\text{t}}\tilde{\beta}\tilde{V}_i$. Then, by introducing $\tilde{\delta} = (1, \delta) \in \{1\} \times \{0,1\}^p$ to force the inclusion of the intercept in the model, Equations (4a) and (4b) can be rewritten as:

$$\pi(\tilde{\beta}|\tilde{\delta}) = \mathcal{N}_{p+1}(0, \text{diag}(\tilde{a}_0, \ldots, \tilde{a}_p)), \text{ where } \tilde{a}_{\ell'} = (1-\tilde{\delta}_{\ell'})\nu_0 + \tilde{\delta}_{\ell'}(\mathbb{1}_{\ell'>0}\nu_1 + \mathbb{1}_{\ell'=0}\sigma_\mu^2), \text{ for } 0 \leq \ell' \leq p.$$

For variance parameters σ^2 and Γ^2, inverse-gamma priors are chosen ((4c) and (4d)), which prohibit negative values. One possibility is to set $\nu_\sigma, \lambda_\sigma, \nu_\Gamma, \lambda_\Gamma$ equal to 1 for example, to make them relatively non-influential.

Following Ročková and George (2014), the i.i.d. Bernoulli prior (4e) is used for the inclusion variable δ, where the hyperparameter α can be seen as the proportion of relevant covariates in the model, and a Beta distribution prior (4f) is chosen on α. To encourage sparsity in the model, Castillo and van der Vaart (2012) suggest choosing a small and b large, $a = 1$ and $b = p$ for example. See Ročková and George (2014) for more details on these choices of priors and the choice of hyperparameters values.

3 Maximum *a posteriori* inference and thresholding

The purpose of this section is to discuss the estimation of $\Theta = (\theta, \alpha) = (\tilde{\beta}, \Gamma^2, \sigma^2, \alpha)$ in model (1) - (4). Recall that $\Xi = (\nu_0, \nu_1, \sigma_\mu^2, \nu_\sigma, \lambda_\sigma, \nu_\Gamma, \lambda_\Gamma, a, b)$ are fixed hyperparameters. In the following, model (1) - (4) is called SSNLME (Spike-and-Slab Non-Linear Mixed-Effects) model. Note that φ, which is not observed, could be considered as a parameter to be estimated, included in Θ. However, to design a scalable inference scheme, we consider it as a latent variable which we marginalise out of the posterior. This enables us to use an EM-type approach which is considerably faster than a full MCMC approach (see details in Subsection 5.5).

The EM-type approach proposed in Section 4 requires to compute the maximum *a posteriori* (MAP) estimator for Θ:

$$\widehat{\Theta}^{MAP} = \underset{\Theta\in\Lambda}{\operatorname{argmax}}\ \pi(\Theta|y), \text{ with } \pi(\Theta|y) = \frac{p_\Theta(y)\pi(\Theta)}{\int_\Lambda p_\Theta(y)\pi(\Theta)d\Theta}\ , \tag{5}$$

where $p_\Theta(y)$ and $\pi(\Theta)$ respectively denote the probability density of y conditionally to Θ, and the prior density of Θ, and Λ denotes the parameter space. However, since the individual parameters φ are marginalised out, the $p_\Theta(y)$ distribution is not explicit. Denoting $Z = (\varphi, \delta) \in \mathcal{Z}$ the latent variables, the marginalised posterior distribution $\pi(\Theta|y)$ takes the form:

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

where $p(y|\Theta, Z)$ and $p(\Theta, Z)$ respectively designate the probability density of y conditionally to (Θ, Z), and the joint distribution of Θ and Z.

Targeting only the maximum *a posteriori* replaces a sampling problem by an optimisation problem, which turns out to be much more scalable than exploring the full posterior. Equation (5) is an optimisation problem in an incomplete data model, which is gainfully tackled using the Stochastic Approximation version of the EM algorithm (SAEM, Delyon et al. (1999)). Often in the literature, the EM algorithm and its extensions are presented in the frequentist framework for the calculation of the maximum likelihood estimator (MLE). Nevertheless, these algorithms are also very well adapted to the computation of the MAP (Dempster et al., 1977).

3.1 General description of SAEM algorithm

In this subsection, we consider the general framework of an incomplete data model with observations y and latent variables Z that characterise the distribution of observations. It it assumed that the density of the complete data (y, Z) is parameterised by Θ, which is unknown and associated with a prior $\pi(\Theta)$. The EM algorithm is iterative and allows to build a sequence $(\Theta^{(k)})_k$ of parameter estimates, which under certain regularity conditions converges to a local maximum of the observed posterior distribution

$$\pi(\Theta|y) = \int \pi(\Theta, Z|y)dZ,$$

(see Delyon et al. (1999) for more details). However, this integral is generally intractable and the idea is to maximise it by iteratively maximising an easier quantity:

$$Q(\Theta|\Theta') \ = \ \mathbb{E}_{Z|y,\Theta'}[\log(\pi(\Theta, Z|y))|y, \Theta'],$$

the conditional expectation of the complete log-posterior $\log(\pi(\Theta, Z|y))$ given the observations y and the current value of the parameter estimates Θ'. However, the quantity $Q(\Theta|\Theta')$ does not always have a closed form. This is especially the case in non-linear mixed-effects models like SSNLME model. The SAEM algorithm is an alternative to the EM algorithm when the E-step, *i.e.* the computation of the Q quantity, is intractable (Delyon et al., 1999). The idea of the SAEM algorithm is to approximate $Q(\Theta|\Theta')$ by a stochastic approximation procedure. More precisely, the E-step of the EM algorithm is replaced by two steps: a simulation step (S-step) and a stochastic approximation step (SA-step). Then the k-th iteration of the SAEM algorithm proceeds as follows:

1. **S-step (Simulation):** simulate a realisation $Z^{(k)}$ of the latent variables according to the conditional distribution $\pi(Z|y, \Theta^{(k)})$.
2. **SA-step (Stochastic Approximation):** update the approximation $Q_{k+1}(\Theta)$ of $Q(\Theta|\Theta^{(k)})$ by a stochastic approximation method, according to:

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

   where $(\gamma_k)_k$ is a sequence of step sizes decreasing towards 0 such that $\forall k$, $\gamma_k \in [0, 1]$, $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$.
3. **M-step (Maximisation):** update the parameter value by computing:

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

Remark 1. *If the model belongs to the curved exponential family, that is the complete log-posterior can be written as:*

$$\log(\pi(\Theta, Z|y)) = -\psi(\Theta) + \Big\langle S(y,Z), \phi(\Theta) \Big\rangle,$$

where ψ and ϕ denote two functions of Θ, with $\langle \cdot, \cdot \rangle$ denoting the scalar product, and $S(y,Z)$ the minimal sufficient statistics of the model, then,

$$Q(\Theta|\Theta^{(k)}) = -\psi(\Theta) + \Big\langle \mathbb{E}_{Z|y,\Theta^{(k)}}[S(y,Z)|y,\Theta^{(k)}], \phi(\Theta) \Big\rangle.$$

It is therefore sufficient to focus on the minimal sufficient statistics instead of $Q(\Theta|\Theta^{(k)})$ itself. More precisely, the SA-step and M-step of the SAEM algorithm are replaced by:

- ***SA-step:*** *update S_{k+1}, the stochastic approximation of $\mathbb{E}_{Z|y,\Theta^{(k)}}[S(y,Z)|y,\Theta^{(k)}]$, according to:*

$$S_{k+1} = S_k + \gamma_k(S(y, Z^{(k)}) - S_k).$$

- ***M-step:*** *update the parameter value by computing:*

$$\Theta^{(k+1)} = \underset{\Theta \in \Lambda}{\operatorname{argmax}} \Big\{ -\psi(\Theta) + \langle S_{k+1}, \phi(\Theta) \rangle \Big\}.$$

Note that theoretical convergence results of the SAEM algorithm are provided in Delyon et al. (1999) under the assumption that the model belongs to the curved exponential family. The SSNLME model belongs to this curved exponential family.

Remark 2. *Note that the simulation step is not always directly feasible. This is particularly true in non-linear mixed-effects models since the conditional distribution of the latent variables knowing the observations and the current value of the parameters is known only to a nearest multiplicative constant. Kuhn and Lavielle (2004) proposed an alternative which consists in coupling the SAEM method with an MCMC procedure. More precisely, at the S-step, the idea is to generate m iterations of an MCMC procedure, which consists in drawing $Z^{(k)}$ using the transition probability of a convergent Markov chain supposed to have the posterior $\pi(Z|y,\Theta^{(k)})$ as its stationary distribution. In practice, m does not need to be large and often $m = 1$ suffices. Theoretical results for the convergence of the SAEM algorithm are extended to the MCMC-SAEM algorithm by Kuhn and Lavielle (2004) and Allassonnière et al. (2010) in models belonging to the curved exponential family.*

3.2 Central decomposition of the Q quantity in spike-and-slab non-linear mixed-effects models

In the following, the notations from Section 2 are used again, and $||\cdot||$ denotes the Euclidean norm on $\mathbb{R}^n$. The SSNLME model (1) - (4) is a particular latent variables model with $y = (y_{ij})_{i,j}$ and $Z = (\varphi, \delta)$. The aim here is to decompose the Q quantity of the SAEM algorithm in the particular case of the SSNLME model, allowing then to describe an algorithm for computing the MAP estimator of Θ in the following subsection.

First, note that the quantity $Q(\Theta|\Theta^{(k)})$ in model (1) - (4) is written as:

$$Q(\Theta|\Theta^{(k)}) = \mathbb{E}_{(\varphi,\delta)|(y,\Theta^{(k)})}[\log(\pi(\Theta,\varphi,\delta|y))|y,\Theta^{(k)}] = \mathbb{E}_{\varphi|(y,\Theta^{(k)})}\left[\widetilde{Q}(y,\varphi,\Theta,\Theta^{(k)})\middle| y,\Theta^{(k)}\right], \tag{6}$$

where

$$\widetilde{Q}(y,\varphi,\Theta,\Theta^{(k)}) = \mathbb{E}_{\delta|(\varphi,y,\Theta^{(k)})}[\log(\pi(\Theta,\varphi,\delta|y))|\varphi,y,\Theta^{(k)}]. \tag{7}$$

It is interesting to write $Q(\Theta|\Theta^{(k)})$ as in Equation (6) because $\widetilde{Q}(y,\varphi,\Theta,\Theta^{(k)})$ has a closed form.

Proposition 3.1. *Consider $\widetilde{Q}(y,\varphi,\Theta,\Theta^{(k)})$ defined by Equation (7) where $\Theta = (\tilde{\beta}, \Gamma^2, \sigma^2, \alpha)$. Then:*

$$\widetilde{Q}(y,\varphi,\Theta,\Theta^{(k)}) = C + \widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)}) + \widetilde{Q}_2(\alpha,\Theta^{(k)}), \tag{8}$$

where C is a normalisation constant which does not depend on Θ, and with:

$$\begin{aligned}\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)}) = &-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - g(\varphi_i,t_{ij}))^2 - \frac{1}{2\Gamma^2}||\varphi - \tilde{V}\tilde{\beta}||^2 - \frac{1}{2}\sum_{\ell'=0}^{p}\tilde{\beta}_{\ell'}^2 \widetilde{d}^*_{\ell'}(\Theta^{(k)}) \\ &- \frac{nJ+\nu_\sigma+2}{2}\log(\sigma^2) - \frac{n+\nu_\Gamma+2}{2}\log(\Gamma^2) - \frac{\nu_\Gamma\lambda_\Gamma}{2\Gamma^2} - \frac{\nu_\sigma\lambda_\sigma}{2\sigma^2}\end{aligned}$$

and

$$\widetilde{Q}_2(\alpha,\Theta^{(k)}) = \log\left(\sqrt{\frac{\nu_0}{\nu_1}}\frac{\alpha}{1-\alpha}\right)\sum_{\ell=1}^{p} p_\ell^*(\Theta^{(k)}) + (a-1)\log(\alpha) + (p+b-1)\log(1-\alpha).$$

Quantities $p_\ell^*(\Theta^{(k)})$, $1 \leq \ell \leq p$, *and* $\widetilde{d}_{\ell'}^*(\Theta^{(k)})$, $0 \leq \ell' \leq p$, *are defined as follows:*

$$p_\ell^*(\Theta^{(k)}) = \mathbb{E}[\delta_\ell|\varphi, y, \Theta^{(k)}] = \frac{\alpha^{(k)}\phi_{\nu_1}(\beta_\ell^{(k)})}{\alpha^{(k)}\phi_{\nu_1}(\beta_\ell^{(k)}) + (1-\alpha^{(k)})\phi_{\nu_0}(\beta_\ell^{(k)})} \tag{9}$$

and

$$\widetilde{d}_{\ell'}^*(\Theta^{(k)}) = \mathbb{E}\left[\frac{1}{(1-\tilde{\delta}_{\ell'})\nu_0 + \tilde{\delta}_{\ell'}(\mathbb{1}_{\ell'>0}\nu_1 + \mathbb{1}_{\ell'=0}\sigma_\mu^2)}\middle|\varphi, y, \Theta^{(k)}\right] = \begin{cases} \dfrac{1}{\sigma_\mu^2} & \text{if } \ell' = 0, \\ d_{\ell'}^*(\Theta^{(k)}) & \text{otherwise,} \end{cases} \tag{10}$$

where $\phi_\nu(\cdot)$ *is the normal density with zero mean and variance* ν *and,*

$$d_\ell^*(\Theta^{(k)}) = \frac{1-p_\ell^*(\Theta^{(k)})}{\nu_0} + \frac{p_\ell^*(\Theta^{(k)})}{\nu_1},\ 1 \leq \ell \leq p.$$

Remark 3. *Note that* $\mathbb{E}[\delta_\ell|\varphi, y, \Theta^{(k)}] = \mathbb{E}[\delta_\ell|\Theta^{(k)}]$ *because the posterior distribution of* δ *given* $(\varphi, y, \Theta^{(k)})$ *depends on* y *and* φ *only through the current estimates* $\Theta^{(k)}$.

The separability of (8) into two distinct functions, $\widetilde{Q}_1$ which depends on $(y, \varphi, \theta, \Theta^{(k)})$ and $\widetilde{Q}_2$ on $(\alpha, \Theta^{(k)})$, allows to update the estimations of θ and α independently from one another. Moreover, since $\widetilde{Q}_2$ does not depend on φ, Proposition 3.1 allows to write that:

$$Q(\Theta|\Theta^{(k)}) = C + \mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)})\middle|y,\Theta^{(k)}\right] + \widetilde{Q}_2(\alpha,\Theta^{(k)}). \tag{11}$$

However, even if $\widetilde{Q}(y,\varphi,\Theta,\Theta^{(k)})$ has a closed form, this is not the case of $Q(\Theta|\Theta^{(k)})$ because the function g is non-linear with respect to φ_i, and so $\pi(\varphi|y,\Theta^{(k)})$ is only known to a nearest multiplicative constant. Thus, it is necessary to use a stochastic approximation method to approximate $\mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)})\middle|y,\Theta^{(k)}\right]$ in Equation (11). The originality of the present extension of the MCMC-SAEM algorithm is that it combines an exact computation $\widetilde{Q}_2(\alpha,\Theta^{(k)})$ and a stochastic approximation of $\mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)})\middle|y,\Theta^{(k)}\right]$ instead of a stochastic approximation of the entire quantity $Q(\Theta|\Theta^{(k)})$. This results in the combination of an exact EM algorithm and of an MCMC-SAEM algorithm for the estimation of α and θ respectively.

Also, let us notice that $\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)})$ takes an exponential form. More precisely,

$$\widetilde{Q}_1(y,\varphi,\theta,\Theta^{(k)}) = -\psi(\theta,\Theta^{(k)}) + \Big\langle S(y,\varphi), \phi(\theta)\Big\rangle, \tag{12}$$

with:

- $S(y,\varphi) = (s_1(y,\varphi), s_2(\varphi), s_3(\varphi)) = \left(\sum_{i,j}(y_{ij} - g(\varphi_i,t_{ij}))^2\ ,\ \sum_{i=1}^n \varphi_i^2\ ,\ \varphi\right)$
- $\phi(\theta) = \left(-\dfrac{1}{2\sigma^2}\ ,\ -\dfrac{1}{2\Gamma^2}\ ,\ \dfrac{\tilde{V}\tilde{\beta}}{\Gamma^2}\right)$
- $\psi(\theta,\Theta^{(k)}) = \dfrac{||\tilde{V}\tilde{\beta}||^2}{2\Gamma^2} + \dfrac{1}{2}\sum_{\ell'=0}^p \tilde{\beta}_{\ell'}^2 \widetilde{d}_{\ell'}^*(\Theta^{(k)}) + \dfrac{nJ+\nu_\sigma+2}{2}\log(\sigma^2) + \dfrac{n+\nu_\Gamma+2}{2}\log(\Gamma^2) + \dfrac{\nu_\Gamma\lambda_\Gamma}{2\Gamma^2} + \dfrac{\nu_\sigma\lambda_\sigma}{2\sigma^2}$

Thus, following Remark 1, it suffices to approximate stochastically $\mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[S(y,\varphi)\middle|y,\Theta^{(k)}\right]$ at SA-step.

3.3 MCMC-SAEM algorithm in spike-and-slab non-linear mixed-effects models

The use of the decomposition discussed in Subsection 3.2 leads to the following extension of the MCMC-SAEM algorithm for computing the MAP estimator of Θ in the SSNLME model, where Λ_θ denotes the parameter space restricted to θ, m is small (between 1 and 5), and K is usually in the order of a few hundred.

Algorithm 1

Input: $K \in \mathbb{N}^*$, $\Theta^{(0)}$ initial parameter, hyperparameters vector $\Xi = (\nu_0, \nu_1, \sigma_\mu^2, \nu_\sigma, \lambda_\sigma, \nu_\Gamma, \lambda_\Gamma, a, b)$, $S_0 = 0$ and $(\gamma_k)_k$ a step sizes sequence decreasing towards 0 such that $\forall k$, $\gamma_k \in [0,1]$, $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$.

for $k = 0$ to $K-1$ **do**

1. **S-Step:** simulate $\varphi^{(k)}$ using the result of m iterations of an MCMC procedure with $\pi(\varphi|y, \Theta^{(k)})$ for target distribution.
2. **SA-Step:** compute $S_{k+1} = S_k + \gamma_k(S(y, \varphi^{(k)}) - S_k)$ with $S(y,\varphi)$ defined by (12), where $S_k = (s_{1,k}, s_{2,k}, s_{3,k}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$.
3. **M-Step:** update $\theta^{(k+1)} = \underset{\theta \in \Lambda_\theta}{\operatorname{argmax}} \left\{ -\psi(\theta, \Theta^{(k)}) + \langle S_{k+1}, \phi(\theta) \rangle \right\}$ and $\alpha^{(k+1)} = \underset{\alpha \in [0,1]}{\operatorname{argmax}} \widetilde{Q}_2(\alpha, \Theta^{(k)})$, which reduces to the following explicit forms in model (1)-(4):
   - $\tilde{\beta}^{(k+1)} = ({}^t\tilde{V}\tilde{V} + \Gamma^{2^{(k)}} D_k^*)^{-1}\, {}^t\tilde{V} s_{3,k+1}$ where $D_k^* = \operatorname{diag}(\widetilde{d_{\ell'}^*}(\Theta^{(k)}), 0 \le \ell' \le p)$
   - $\Gamma^{2^{(k+1)}} = \dfrac{||\tilde{V}\tilde{\beta}^{(k+1)}||^2 + \nu_\Gamma \lambda_\Gamma + s_{2,k+1} - 2\langle s_{3,k+1}, \tilde{V}\tilde{\beta}^{(k+1)} \rangle}{n + \nu_\Gamma + 2}$
   - $\sigma^{2^{(k+1)}} = \dfrac{\nu_\sigma \lambda_\sigma + s_{1,k+1}}{nJ + \nu_\sigma + 2}$
   - $\alpha^{(k+1)} = \dfrac{\sum_{\ell=1}^p p_\ell^*(\Theta^{(k)}) + a - 1}{p + b + a - 2}$

   where ψ and ϕ are defined by (12), and $\widetilde{Q}_2(\alpha, \Theta^{(k)})$, $(p_\ell^*(\Theta^{(k)}))_{1 \le \ell \le p}$ and $(\widetilde{d_{\ell'}^*}(\Theta^{(k)}))_{0 \le \ell' \le p}$ are defined in Proposition 3.1.

end for

Output: $\widehat{\Theta}^{MAP} = (\widehat{\tilde{\beta}}^{MAP}, \widehat{\Gamma}^{2^{MAP}}, \widehat{\sigma}^{2^{MAP}}, \widehat{\alpha}^{MAP}) = (\tilde{\beta}^{(K)}, \Gamma^{2^{(K)}}, \sigma^{2^{(K)}}, \alpha^{(K)})$.

Remark 4. *Note that for a linear mixed-effects model, that is when g is linear with respect to φ_i, a classical EM algorithm is applicable. Indeed, denoting $g(\varphi_i, t_{ij}) = A(t_{ij})\varphi_i$ for all $1 \le i \le n$, $1 \le j \le J$, $A(t_{ij}) \in \mathbb{R}$ and using the previous notations, the Q quantity has the following explicit form:*

$$Q(\Theta|\Theta^{(k)}) = C + \widehat{Q}_1(y, \varphi, \theta, \Theta^{(k)}) + \widetilde{Q}_2(\alpha, \Theta^{(k)}),$$

with $\widehat{Q}_1(y, \varphi, \theta, \Theta^{(k)}) = \mathbb{E}_{\varphi|y,\Theta^{(k)}}\left[\widetilde{Q}_1(y, \varphi, \theta, \Theta^{(k)}) \middle| y, \Theta^{(k)}\right]$ *and more precisely:*

$$\begin{aligned}\widehat{Q}_1(y, \varphi, \theta, \Theta^{(k)}) = &-\frac{1}{2\sigma^2}\sum_{i,j}\left[(y_{ij} - A(t_{ij})m_{i,k})^2 + A(t_{ij})^2\Sigma_{i,k}^2\right] - \frac{1}{2\Gamma^2}\sum_i\left[((\tilde{V}\tilde{\beta})_i - m_{i,k})^2 + \Sigma_{i,k}^2\right] \\ &- \frac{1}{2}\sum_{\ell'=0}^p \tilde{\beta}_{\ell'}^2 \widetilde{d_{\ell'}^*}(\Theta^{(k)}) - \frac{nJ + \nu_\sigma + 2}{2}\log(\sigma^2) - \frac{n + \nu_\Gamma + 2}{2}\log(\Gamma^2) - \frac{\nu_\Gamma\lambda_\Gamma}{2\Gamma^2} - \frac{\nu_\sigma\lambda_\sigma}{2\sigma^2}\end{aligned}$$

where $\Sigma_{i,k}^2 = \dfrac{\sigma^{2^{(k)}}\Gamma^{2^{(k)}}}{\Gamma^{2^{(k)}}\sum_j A(t_{ij}) + \sigma^{2^{(k)}}}$ *and* $m_{i,k} = \dfrac{\Gamma^{2^{(k)}}\sum_j y_{ij}A(t_{ij}) + \sigma^{2^{(k)}}(\tilde{V}\tilde{\beta}^{(k)})_i}{\Gamma^{2^{(k)}}\sum_j A(t_{ij}) + \sigma^{2^{(k)}}}$ *are defined such that* $\pi(\varphi_i|y, \Theta^{(k)}) = \mathcal{N}(m_{i,k}, \Sigma_{i,k}^2)$.

3.4 Estimator thresholding

As in Ročková and George (2014), after obtaining an estimator $\widehat{\Theta}^{MAP}$, the support S_β^*, defined in Equation (3), can be naturally estimated as the most probable model conditionally on $\widehat{\Theta}^{MAP}$. Indeed, for all $\ell \in \{1, \ldots, p\}$,

the *a posteriori* inclusion probability of the covariate ℓ can be obtained as:

$$\mathbb{P}(\delta_\ell = 1|y, \widehat{\beta}_\ell^{MAP}, \widehat{\alpha}^{MAP}) = \frac{\pi(\widehat{\beta}_\ell^{MAP}|\delta_\ell = 1)\pi(\delta_\ell = 1|\widehat{\alpha}^{MAP})}{\pi(\widehat{\beta}_\ell^{MAP}|\delta_\ell = 1)\pi(\delta_\ell = 1|\widehat{\alpha}^{MAP}) + \pi(\widehat{\beta}_\ell^{MAP}|\delta_\ell = 0)\pi(\delta_\ell = 0|\widehat{\alpha}^{MAP})}. \quad (13)$$

Then, $\widehat{\delta}$, which is the most probable δ knowing that $\Theta = \widehat{\Theta}^{MAP}$, can be computed as follows:

$$\begin{aligned}\widehat{\delta}_\ell = 1 &\iff \mathbb{P}(\delta_\ell = 1|y, \widehat{\beta}_\ell^{MAP}, \widehat{\alpha}^{MAP}) \geq 0.5 \\ &\iff |\widehat{\beta}_\ell^{MAP}| \geq \sqrt{2\frac{\nu_0\nu_1}{\nu_1 - \nu_0}\log\left(\sqrt{\frac{\nu_1}{\nu_0}}\frac{1-\widehat{\alpha}^{MAP}}{\widehat{\alpha}^{MAP}}\right)} = s_\beta(\nu_0, \nu_1, \widehat{\alpha}^{MAP}).\end{aligned}$$

Note that this estimator can be seen as a local version of the median probability model of Barbieri and Berger (2004). Thus, the following subset of covariates is selected via a thresholding operation:

$$\widehat{S} = \left\{\ell \in \{1, \dots, p\} \;\middle|\; |\widehat{\beta}_\ell^{MAP}| \geq s_\beta(\nu_0, \nu_1, \widehat{\alpha}^{MAP})\right\}. \quad (14)$$

Remark 5. *Note that threshold $s_\beta(\nu_0, \nu_1, \widehat{\alpha}_{\nu_0}^{MAP})$ is the same for all the covariates but depends on the values of the spike and slab hyperparameters ν_0 and ν_1 which act as tuning parameters for the penalty.*

Remark 6. *It is interesting to note that the thresholding rule is unchanged from the easier situation where the individual parameters φ_i's would have been directly observed, which would have fit into the framework treated in Ročková and George (2014).*

4 Covariate selection procedure

4.1 Model selection procedure

Similarly to LASSO regression (Tibshirani, 1996), it is interesting to exploit the flexibility of the spike-and-slab prior to study different levels of sparsity in the vector β, and thanks to the speed of the MCMC-SAEM algorithm, it is possible to explore a grid of values for the spike hyperparameter ν_0 rather than focusing on a single value. Indeed, mechanically, the higher ν_0 is, the less covariates are included in the estimated support of β. This is why it is more interesting to look at a grid of values and then to use a model selection criterion to choose the optimal model. Let us note Δ this grid, and $|\Delta|$ the number of grid points. Then, for all $\nu_0 \in \Delta$, the MCMC-SAEM algorithm is executed to obtain the MAP estimate of Θ, $\widehat{\Theta}_{\nu_0}^{MAP}$, which is then used to determine a subset of relevant covariates $\widehat{S}_{\nu_0}$ as explained in Subsection 3.4, Equation (14). This first step reduces the total collection of 2^p possible models to a smaller collection of $|\Delta| \ll 2^p$ promising sub-models $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$ with high posterior probability. Next, a model selection criterion can be applied to choose the "best" model from this collection.

As explained in Ročková and George (2014), a possible criterion is to maximise, along the grid, the marginal posterior of δ under the prior with $\nu_0 = 0$. This corresponds to the so-called Dirac-and-slab prior, where the spike is a Dirac distribution (Mitchell and Beauchamp, 1988). However, in our case, it is not possible to have an explicit expression for this marginal and it is also difficult to obtain it numerically, so this criterion is not convenient.

However, as the collection of models has been reduced to a small sub-collection $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$, that contains at most $|\Delta|$ models, an information criterion can be used to choose the final model. The eBIC criterion (extended Bayesian Information Criterion, Chen and Chen (2008)) is preferred to the BIC criterion (Schwarz, 1978; Delattre et al., 2014) due to the high-dimensional framework because it allows to take into account that the number of possible models with $q \leq p$ covariates increases quickly as q increases. Thus, covariate selection here consists in choosing the "best" $\nu_0 \in \Delta$, that is noted $\hat{\nu}_0$, as the one that minimises the following penalty function:

$$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\operatorname{argmin}}\Big\{\mathrm{eBIC}(\widehat{S}_{\nu_0})\Big\}, \quad (15)$$

where

$$\mathrm{eBIC}(\widehat{S}_{\nu_0}) = -2\log\left(p(y; \hat{\theta}_{\nu_0}^{MLE})\right) + \mathrm{pen}_{\mathrm{eBIC}}(\nu_0), \quad (16)$$

with:

- $\log\left(p(y;\theta)\right)$ is the log-likelihood of the model (1),
- $\hat{\theta}_{\nu_0}^{MLE} = (\hat{\tilde{\beta}}_{\nu_0}^{MLE}, \hat{\Gamma}^{2^{MLE}}_{\nu_0}, \hat{\sigma}^{2^{MLE}}_{\nu_0})$ is the maximum likelihood estimate (MLE) of the parameter $\theta = (\tilde{\beta}, \Gamma^2, \sigma^2)$ in the sub-model $\widehat{S}_{\nu_0}$ selected by thresholding process for this ν_0,
- $\text{pen}_{\text{eBIC}}(\nu_0) = B_{\nu_0} \times \log(n) + 2\log\left(\binom{p}{B_{\nu_0}}\right)$ is a penalty function, with B_{ν_0} the number of free parameters in the sub-model $\widehat{S}_{\nu_0}$.

Note that the MLE and log-likelihood do not have an explicit form here because the individual parameters are latent and the function g is non-linear with respect to φ_i. For every $\nu_0 \in \Delta$, the estimate $\hat{\theta}_{\nu_0}^{MLE}$ and the log-likelihood $\log p(y; \hat{\theta}_{\nu_0}^{MLE})$ are respectively computed with an MCMC-SAEM algorithm and importance sampling techniques (see *e.g.* Kuhn and Lavielle (2005) and Lavielle (2014) for details) to derive the corresponding value of $\text{eBIC}(\widehat{S}_{\nu_0})$.

The proposed variable selection procedure can be summarised as in Algorithm 2.

Algorithm 2

Input: Δ a grid of ν_0 values, and all required arguments for MCMC-SAEM (Algorithm 1).

▷ **Reduce the model collection:**
for $\nu_0 \in \Delta$ **do**
1. Compute the MAP estimate $\widehat{\Theta}_{\nu_0}^{MAP}$ by Algorithm 1.
2. Threshold the estimator $\widehat{\beta}_{\nu_0}^{MAP}$ to define sub-model $\widehat{S}_{\nu_0}$ according to Equation (14).

end for

▷ **Compute the eBIC criterion:**
for each unique sub-model among $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$ **do**
1. Compute the MLE estimate $\widehat{\theta}_{\nu_0}^{MLE}$ in sub-model $\widehat{S}_{\nu_0}$.
2. Compute the log-likelihood $\log p(y; \hat{\theta}_{\nu_0}^{MLE})$.
3. Compute the associated $\text{eBIC}(\widehat{S}_{\nu_0})$ criterion according to Equation (16).

end for

▷ **Identify the best level of sparsity:** compute $\hat{\nu}_0$ defined by Equation (15).

Output: $\widehat{S}_{\hat{\nu}_0}$.

Remark 7. *Note that for Algorithm 2 it is possible to parallelise the computations along the grid because the outputs of the algorithm for two given values of $\nu_0 \in \Delta$ do not depend on each other.*

4.2 Application on a detailed example

A data-set is simulated according to a logistic growth model that is model (1) with:

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

where ψ_1 and ψ_2 are known constants. This is a common and realistic model used in many fields of life sciences, such as plant growth for example. Let us consider $n = 200$ individuals, $p = 500$ covariates and $J = 10$ observations per individual. For all $i \in \{1, \dots n\}$, for all $j \in \{1, \dots, J\}$, $t_{ij} = t_j = 150 + (j-1)\dfrac{3000 - 150}{J - 1}$. For each individual, the p covariates are simulated independently according to standard normal distributions $\mathcal{N}(0, 1)$. The parameter values are set to $\sigma^2 = 30$, $\psi_1 = 200$, $\psi_2 = 300$, $\mu = 1200$, $\beta = {}^{\text{t}}(100, 50, 20, 0, \dots, 0)$ and $\Gamma^2 = 200$. Thus, only the first three covariates are influential, *i.e.* $S_\beta^* = \{1, 2, 3\}$.

4.2.1 Convergence of the MCMC-SAEM algorithm

Algorithm 1, is initialised here with: $\forall \ell \in \{1, \ldots, 10\}$, $\beta_\ell^{(0)} = 100$, $\forall \ell \in \{11, \ldots, p\}$, $\beta_\ell^{(0)} = 1$, $\mu^{(0)} = 1500$, $\sigma^{2^{(0)}} = 100$, $\Gamma^{2^{(0)}} = 5000$ and $\alpha^{(0)} = 0.5$. The hyperparameters are set to: $\nu_0 = 0.02$, $\nu_1 = 12000$, $\nu_\sigma = \lambda_\sigma = \nu_\Gamma = \lambda_\Gamma = 1$, $a = 1$, $b = p$ and $\sigma_\mu = 3000$. In practice, to allow more flexibility during the first iterations and thus to move away more quickly from the initial condition, it is usual to start the algorithm with n_{burnin} burn-in iterations, *i.e.* to use a step sizes sequence $(\gamma_k)_k$ of the form: $\gamma_k = 1$ for $0 \leq k \leq n_{\text{burnin}} - 1$ and $\gamma_k = 1/(k - n_{\text{burnin}} + 1)^\gamma$ for $n_{\text{burnin}} \leq k \leq K - 1$, where $\gamma \in]0.5, 1[$, $n_{\text{burnin}} < K$ with K the number of iterations of the SAEM algorithm (see Kuhn and Lavielle, 2005). Here, the step sizes are defined with $\gamma = 2/3$, $n_{\text{burnin}} = 350$ and $K = 500$.

Figure 1 represents the convergence graphs of one run of the MCMC-SAEM algorithm for μ, some components of β, σ^2, Γ^2 and α. It is observed that the algorithm converges in a few iterations for any parameter. In this example, after 500 iterations, the algorithm returns $\widehat{\beta}_1^{MAP} = 95.7$, $\widehat{\beta}_2^{MAP} = 51.96$, $\widehat{\beta}_3^{MAP} = 22.46$, $\widehat{\mu}^{MAP} = 1202$, $\widehat{\sigma^2}^{MAP} = 33.86$, $\widehat{\Gamma^2}^{MAP} = 1.77$ and $\widehat{\alpha}^{MAP} = 0.003$. Moreover, the estimates of the null fixed effects of the covariates are all less than 0.07 in absolute value. Note that the parameters are all relatively correctly estimated, except for Γ^2 but this was expected because of a over-fitting situation. Indeed, the underestimation of Γ^2 can be explained by the fact that since $\nu_0 > 0$, none of the estimates of the coefficients of β is zero and therefore all the covariates are active in the model, which makes the variance estimation of the random effect tend towards 0 in Equation (1b). This illustrates the need to threshold the estimators as described in Subsection 3.4.

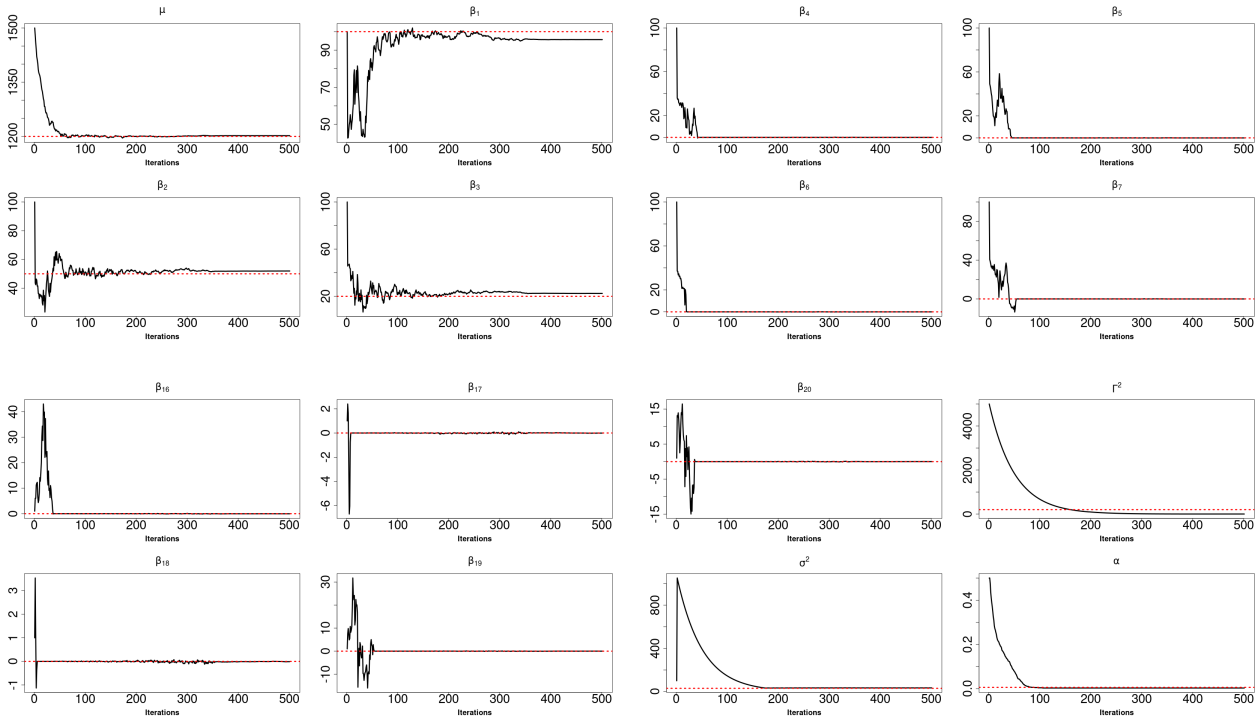


Figure 1: Convergence graphs of the MCMC-SAEM algorithm for μ, some components of β, σ^2, Γ^2 and α on one simulated data-set, for $\nu_0 = 0.02$ and $\nu_1 = 12000$. The red dashed line corresponds to the true value of the considered parameter.

4.2.2 Spike-and-slab regularisation plot and model selection

To illustrate how the variable selection works, the full procedure is applied on this simulated example on the grid of ν_0 values Δ such that $\log_{10}(\Delta) = \left\{ -2 + k \times \frac{4}{19}, k \in \{0, \ldots, 19\} \right\}$. To have a visual representation

of this procedure, a spike-and-slab regularisation plot inspired from Ročková and George (2014) is drawn. It represents the evolution of the β_ℓ estimates for all $\ell \in \{1, \ldots, p\}$ along the grid of ν_0, and the value of the selection threshold associated with $\nu_0 \in \Delta$. The regularisation plot on Figure 2(A) shows the MAP estimates of the covariate fixed effects vector β obtained for each $\nu_0 \in \Delta$. The blue lines are associated with the three relevant covariates, while the black lines are associated with the null fixed effects of the covariates. Moreover, the red lines correspond to the selection threshold of the covariates. Thus, for each $\nu_0 \in \Delta$, the selected covariates $\widehat{S}_{\nu_0}$ are those associated with a $(\widehat{\beta}^{MAP}_{\nu_0})_\ell$ located outside the two red curves in the regularisation plot. As expected, the larger ν_0 is, the smaller the support of the associated $\widehat{\beta}^{MAP}_{\nu_0}$ is. Indeed, on the one hand, the selection threshold increases with ν_0, and on the other hand, the larger ν_0 is, the more $(\widehat{\beta}^{MAP}_{\nu_0})_\ell$'s are truncated in the spike distribution. This can also be seen in Figure 3 which shows that the *a posteriori* inclusion probability (13) of covariate 3 (associated to $\beta_3 = 20$ the smallest non-zero covariate fixed effect) decreases as ν_0 increases. This illustrates the interest of going through a grid rather than focusing on a single ν_0 value.

Figure 2(B) represents the value of the eBIC criterion for all ν_0 in Δ. As desired, it is minimal for the values of ν_0 for which exactly the right model is selected. The procedure returns the second value of $\nu_0 \in \Delta$, *i.e.* $\hat{\nu}_0 \approx 0.016$, and $\widehat{S}_{\hat{\nu}_0} = \{1, 2, 3\}$. So, in this simulated example, our procedure returns exactly the right model, that is the one with only the first three covariates.

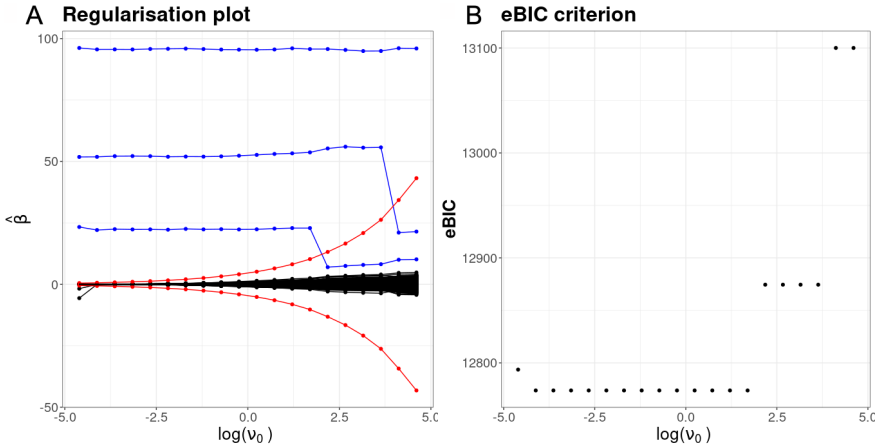


Figure 2: Example of a regularisation plot (A) with eBIC criterion graph (B) for model selection. On (A), the blue lines are associated with the three true relevant covariates, while the black lines are associated with the null fixed effects of the covariates. The red lines correspond to the selection threshold of the covariates.

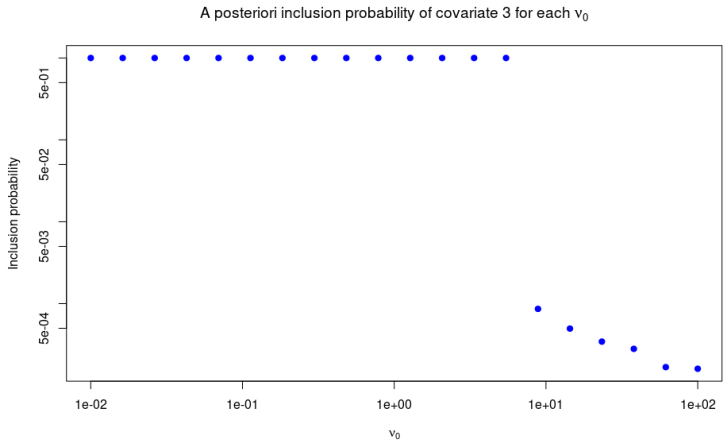


Figure 3: *A posteriori* inclusion probability of covariate 3 for each $\nu_0 \in \Delta$.

5 Simulation study and comparison with MCMC

5.1 Model for the simulation study

To test the effectiveness of the proposed variable selection procedure, an extensive simulation study was conducted. It also considers a growth model like in Subsection 4.2, but here, for a more realistic scenario, $\psi = (\psi_1, \psi_2)$ is seen as an unknown fixed effect. Parameter ψ must also be estimated and therefore the population parameters are $\theta = (\mu, \beta, \psi, \Gamma^2, \sigma^2)$. The procedure presented earlier in this paper can easily be extended to this case. Indeed, let us consider such a model:

$$\begin{cases} y_{ij} = g(\varphi_i, \psi, t_{ij}) + \varepsilon_{ij}, & \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad (17a) \\ \varphi_i = \mu + {}^{\mathrm{t}}\beta V_i + \xi_i, & \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma^2), \quad (17b) \end{cases}$$

with:

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

As the function g is not separable into φ_i and ψ, the model does not belong to the curved exponential family since it is not possible to write $\widetilde{Q}_1$ as in Equation (12). As a result, the expression for the maximum argument in ψ at M-step is not explicit. One solution would be to do numerical optimisation in ψ.

However, for ease of implementation of the MCMC-SAEM algorithm, following the idea of Kuhn and Lavielle (2005), an extended model belonging to the curved exponential family is used to estimate the parameters:

$$\begin{cases} y_{ij} \overset{\text{ind.}}{\sim} \mathcal{N}(g(\varphi_i, \psi, t_{ij}), \sigma^2) & (18a) \\ \varphi_i \overset{\text{ind.}}{\sim} \mathcal{N}(\mu + {}^{\mathrm{t}}\beta V_i, \Gamma^2) & (18b) \\ \psi \sim \mathcal{N}(\eta, \Omega) & (18c) \end{cases}$$

with φ and ψ independent, $\Omega = \mathrm{diag}(\omega_1^2, \omega_2^2)$ known and $\theta^{ext} = (\mu, \beta, \eta, \sigma^2, \Gamma^2)$ the new population parameter to be estimated. The estimation of η is then used as an estimation of ψ. As previously, the indicators δ (Equation (2)) are introduced and we consider the same priors as in (4) for $(\mu, \beta, \sigma^2, \Gamma^2, \delta, \alpha)$. For η, the following prior is chosen: for $m = 1, 2$, $\pi(\eta_m) = \mathcal{N}(0, \rho_m^2)$, with $\rho_m^2 > 0$ known. This amounts to randomising hyperparameters of the prior on ψ, implying a less informative prior than if η were fixed. Here, $\Theta = (\theta^{ext}, \alpha)$ is the population parameter and $Z = (\varphi, \psi, \delta)$ are the latent variables. The steps of the MCMC-SAEM algorithm can be derived as in Section 3 (see appendix A). The estimation method is unchanged for parameters $(\mu, \beta, \sigma^2, \Gamma^2, \alpha)$ because the quantity $\widetilde{Q}_1$ is separable into $(\mu, \beta, \sigma^2, \Gamma^2, \alpha)$ and η. The main difference is that there is another latent variable ψ, which must also be simulated at the S-step. The thresholding procedure is not modified.

Remark 8. *In order to limit the estimation error between the initial model and this extended model, the value of the covariance matrix is adapted during the iterations. Inspired by the results of Allassonnière and Debavelaere (2021) for the case of the computation of the MLE, the following process is chosen: start with a fairly large initial value* $\Omega^{(0)} = \mathrm{diag}(\omega_1^{2(0)}, \omega_2^{2(0)})$ *for a certain number* κ *of iterations, then multiply it by* $0 < \tau < 1$*, and iterate this process every* κ *iterations. Starting from a large initial value, the value of* Ω *remains large enough during the first iterations to allow a rather fast convergence speed, then it is slowly decreases towards* 0*, while remaining always strictly positive, to limit the estimation error between the initial model and the extended model.*

5.2 Simulation design

For this simulation study, individual profiles are simulated according to model (17) by considering $J = 10$ observations per individual and regular observation time points such that $t_{ij} = t_j = 150 + (j-1)\dfrac{3000 - 150}{J - 1}$, $\sigma^2 = 30$, $\psi_1 = 200$, $\psi_2 = 300$, $\mu = 1200$, $\beta = {}^{\mathrm{t}}(100, 50, 20, 0, \ldots, 0)$. Thus, only the first three covariates are assumed influential and their respective intensities are contrasted. The individual covariates $V_i \in \mathbb{R}^p$, $1 \leq i \leq n$, are simulated independently according to a centred multivariate Gaussian distribution with covariance matrix $\Sigma \in \mathcal{M}_p(\mathbb{R})$. To test the sensitivity of the proposed procedure to the correlation that may exist between covariates, different scenarios are tested corresponding to different structures for matrix Σ.

Different values of n (number of subjects) and p (number of covariates) are used according to the scenario. Several values of Γ^2 (variance of the random effects) are also used in order to evaluate the performances of the method in different "signal-to-noise" situations.

- **Scenario with uncorrelated covariates.** This is the baseline scenario where optimal performance of Algorithm 2 is expected. This corresponds to $\Sigma = I_p$, where I_p is the identity matrix of size p. The following values for n, p and Γ^2 are used : $n \in \{100, 200\}$, $p \in \{500, 2000, 5000\}$ and $\Gamma^2 \in \{200, 1000, 2000\}$.
- **Scenarios with correlations between covariates.**
  1. The first scenario leaves the three influential covariates uncorrelated with all other covariates whereas the non-influential covariates are correlated with each other. An autoregressive correlation structure is considered between the non-influential covariates. This corresponds to $\Sigma = \left(\begin{array}{c|c} I_3 & 0_{3,p-3} \\ \hline 0_{p-3,3} & (\rho_\Sigma^{|i-j|})_{i,j\in\{4,\dots,p\}} \end{array}\right)$, with $|\rho_\Sigma| < 1$.
  2. In the second scenario, the third influential covariate is assumed to be correlated to every non-influential covariate according to an autoregressive correlation structure. This corresponds to $\Sigma = \left(\begin{array}{c|c} I_3 & A \\ \hline {}^tA & I_{p-3} \end{array}\right)$, with $A = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ & & (\rho_\Sigma^{|3-j|})_{j\in\{4,\dots,p\}} & \end{pmatrix}$, $|\rho_\Sigma| < 1$.
  3. The third scenario considers correlations between the sole influential covariates. Again, an autoregressive correlation structure is used. This corresponds to $\Sigma = \left(\begin{array}{c|c} (\rho_\Sigma^{|i-j|})_{i,j\in\{1,\dots,3\}} & 0_{3,p-3} \\ \hline 0_{p-3,3} & I_{p-3} \end{array}\right)$, $|\rho_\Sigma| < 1$.
  4. In the fourth scenario, an autoregressive correlation structure is used between the covariates without making any distinction between the influential covariates and the non-influential covariates. This corresponds to $\Sigma = (\rho_\Sigma^{|i-j|})_{i,j\in\{1,\dots,p\}}$, $|\rho_\Sigma| < 1$.

  To study the impact of correlations between covariates according to the four scenarios above, the following values for n, p, Γ^2 and ρ_Σ are used: $n = 200$, $p \in \{500, 2000, 5000\}$, $\Gamma^2 \in \{200, 2000\}$ and $\rho_\Sigma \in \{0.3, 0.6\}$.

For each of the five scenarios described above and each combination (n, p, Γ^2) or $(n, p, \Gamma^2, \rho_\Sigma)$, 100 different data-sets are simulated and the support of β is estimated by applying Algorithm 2 on each data-set. Note that, in order to be able to compare covariates that do not have the same order of magnitude, the covariates are centred and reduced. The performances in terms of exact selection of the true influential covariates, over-selection and under-selection are examined (see Subsection 5.4).

5.3 Algorithmic settings

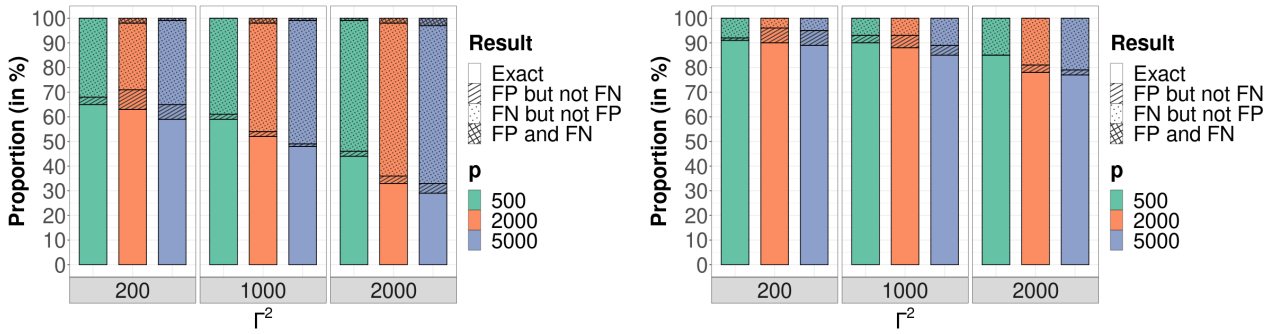
The following settings are used for Algorithm 2.

- The hyperparameter values are set to $\nu_\sigma = \lambda_\sigma = \nu_\Gamma = \lambda_\Gamma = 1$, $a = 1$, $b = p$, $\sigma_\mu = 3000$, $\rho_1^2 = \rho_2^2 = 1200$, $\nu_1 = 12000$, and the spike parameter ν_0 runs through a grid Δ defined as $\log_{10}(\Delta) = \left\{ -2 + k \times \frac{4}{19}, k \in \{0, \dots, 19\} \right\}$.
- The step sizes are defined with $\gamma = 2/3$, $n_{\text{burnin}} = 350$ and $K = 500$ as explained in Subsection 4.2.1.
- The MCMC-SAEM algorithm is initialised with: $\forall \ell \in \{1, \dots, 10\}$ $\beta_\ell^{(0)} = 100$, $\forall \ell \in \{11, \dots, p\}$ $\beta_\ell^{(0)} = 1$, $\mu^{(0)} = 1400$, $\sigma^{2^{(0)}} = 100$, $\Gamma^{2^{(0)}} = 5000$, $\alpha^{(0)} = 0.5$ and $\eta^{(0)} = {}^t(400, 400)$. Note that different initialisations have been tested and have shown similar performances.
- At the beginning of the algorithm, $\Omega = \text{diag}(20, 20)$ and it is slowly reduced during the iterations as explained in Remark 8 with $\kappa = 40$ and $\tau = 0.9$.

5.4 Results

5.4.1 Scenario with uncorrelated covariates

Figure 4 represents, for all (n, p, Γ^2) combinations, the proportion of data-sets for which Algorithm 2 selects the correct model (unpatterned bars), selects a model that contains the correct model (*i.e.* there are false positives but not false negatives, striped bars), selects a model that is included in the correct model (*i.e.* there are false negatives but not false positives, doted bars), or selects a model that contains both false positives and false negatives (crosshatched bars). The procedure selects exactly the right model in a large majority of cases for a sufficiently large number of individuals n. When n increases, the results improve, which suggests a consistency property in selection. With n and p fixed, the more the inter-individual variance Γ^2 is important, the more the results degrade. Indeed, as Γ^2 increases, the "signal-to-variability" ratio decreases, leading to difficulties in detecting the third covariate associated with the lowest non-zero coefficient in β. It could also be noted that with n and Γ^2 fixed, the results deteriorate when p increases but the effect of p seems weak when n is large. In addition, when the procedure fails, it is most often because it under-selects, that is, it selects fewer variables than there are. Indeed, in most configurations for (n, p, Γ^2), the proportion of data-sets that select a model that is included in the correct model is higher than the proportions of the other failure scenarios. It seems that the designed method tends to avoid false positives, even though this may result in not having selected all the truly influential covariates.



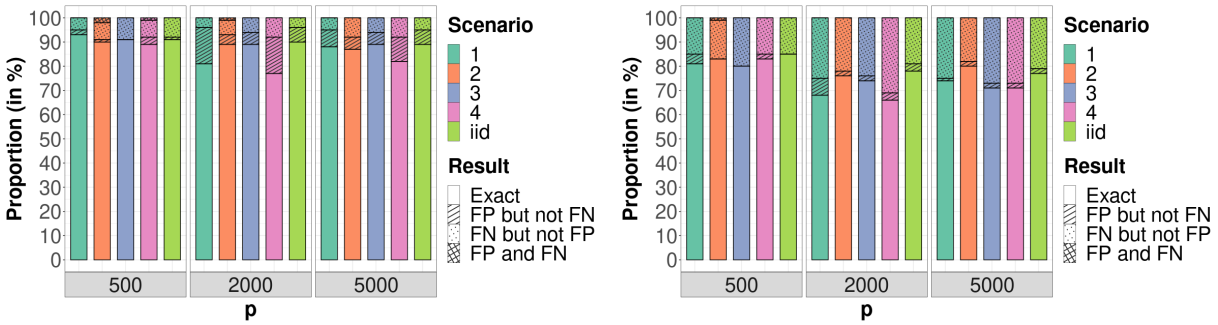
(a) For $n = 100$ (b) For $n = 200$

Figure 4: Uncorrelated covariates. Proportion of data-sets on which Algorithm 2 selects the correct model ("Exact", unpatterned bars), a model that contains false positives (FP) but not false negatives (FN) ("FP but not FN", striped bars), a model that contains false negatives but not false positives ("FN but not FP", doted bars), or a model that contains both false positives and false negatives ("FP and FN", crosshatched bars) for $n = 100$ (a) and $n = 200$ (b), and different values of p and Γ^2.

5.4.2 Scenarios with correlated covariates

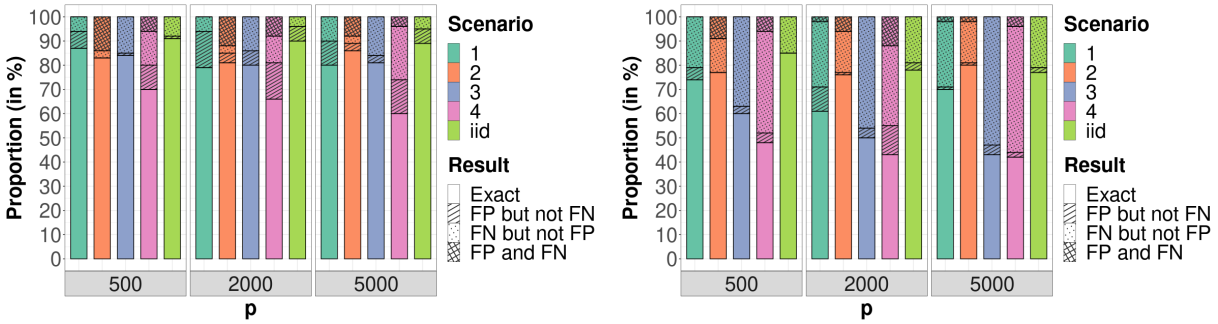
The results are presented in Figure 5 for $\rho_\Sigma = 0.3$ and in Figure 6 for $\rho_\Sigma = 0.6$. On these figures, one can compare the selection performance of our procedure in the different scenarios of correlations between covariates with the case without correlations (iid scenario). First, for scenario 1, that is when the non-active covariates are correlated, quite similar performances to the iid scenario are observed, but with more over-selection. Indeed, as a consequence of the correlation between irrelevant covariates, the latter tend to be selected more often and in small groups. Then, in scenario 2, it is assumed that the third relevant covariate is correlated to the non-active covariates. In this case, similar results to the iid scenario are observed. Indeed, the selection performances of the proposed procedure are only slightly affected by this scenario of correlations. This can be explained by the fact that, in this case, among the group of correlated covariates, the method will tend to select only one (or at least a very limited number of covariates among them): the most intense is chosen, *i.e.* the third true covariate. Next, scenario 3 describes correlations between the relevant covariates. Like the previous scenario, the procedure tends to select few covariates among the correlated covariates since they explain the response variable in a similar way. This also explains the degradation of the results when ρ_Σ increases. Thus, this scenario is inclined to under-select more than the others. Finally, scenario 4 corresponds to a full correlation matrix between all covariates. Note that the correlation matrix chosen for this scenario assumes a fairly strong correlation between the three true covariates. Thus, this scenario

also leads to much under-selection compared to the iid case. However, it over-selects more than scenario 3 because of the correlations between the true and false covariates.



(a) For $\Gamma^2 = 200$ (b) For $\Gamma^2 = 2000$

Figure 5: Correlated covariates, $\rho_\Sigma = 0.3$. Proportion of data-sets on which Algorithm 2 selects the correct model ("Exact", unpatterned bars), a model that contains false positives but not false negatives ("FP but not FN", striped bars), a model that contains false negatives but not false positives ("FN but not FP", doted bars), or a model that contains both false positives and false negatives ("FP and FN", crosshatched bars) for $\Gamma^2 = 200$ (a) and $\Gamma^2 = 2000$ (b) and different values of p. Scenario "iid" corresponds to the case where the covariates are not correlated and is used as a reference.



(a) For $\Gamma^2 = 200$ (b) For $\Gamma^2 = 2000$

Figure 6: Correlated covariates, $\rho_\Sigma = 0.6$. Proportion of data-sets on which Algorithm 2 selects the correct model ("Exact", unpatterned bars), a model that contains false positives but not false negatives ("FP but not FN", striped bars), a model that contains false negatives but not false positives ("FN but not FP", doted bars), or a model that contains both false positives and false negatives ("FP and FN", crosshatched bars) for $\Gamma^2 = 200$ (a) and $\Gamma^2 = 2000$ (b) and different values of p. Scenario "iid" corresponds to the case where the covariates are not correlated and is used as a reference.

5.5 Comparison with an MCMC implementation

It is reasonably straightforward to implement an MCMC algorithm for full posterior inference on the spike-and-slab variable selection for non-linear mixed-effects model. This makes it relevant to compare the run time of a full MCMC approach and the MCMC-SAEM method proposed in this paper, and highlight the better scaling properties of the latter. To build the most informative comparison, the same model with a smooth spike is considered for both the MCMC and MCMC-SAEM approaches, remarking that spike-and-slab priors with a Dirac spike are known to pose challenges for MCMC (see Bai et al., 2021). For the MCMC algorithm, an efficient C++ implementation of an random walk adaptive MCMC is used through the Nimble software (de Valpine et al., 2017), which uses an adaptive scheme proposed in Shaby and Wells (2010). To make the comparison as fair as possible, we marginalise the sampler over the discrete inclusion variables δ, to mirror the marginalisation in (7). This was found to appreciably improve the mixing of the MCMC algorithm. It is possible to retrieve the δ variables from the posterior samples using their conditional posterior distribution.

Common data-sets are simulated according to model (17) with the following parameters: $n = 200$ individuals, $p \in \{500, 700, 1000, 1500, 2000, 2500\}$ covariates, $J = 10$ observations per individual, $\sigma^2 = 30$, $\psi_1 = 200$, $\psi_2 = 300$, $\mu = 1200$, $\beta = {}^{\mathrm{t}}(100, 50, 20, 0, \ldots, 0)$ and $\Gamma^2 = 200$. For $i \in \{1, \ldots n\}$ and $j \in \{1, \ldots, J\}$, $t_{ij} = t_j = 150 + (j-1)\dfrac{3000 - 150}{J - 1}$. Covariates are simulated independently and identically distributed according to $\mathcal{N}(0, 1)$. The objective is to compare the time needed to estimate the parameters $\Theta = (\mu, \beta, \psi, \Gamma^2, \sigma^2)$ between the MCMC-SAEM algorithm proposed in this article (Algorithm 1 adapted to model (17), see appendix A) and the full MCMC procedure described above. As explained in Subsection 5.2, to estimate the parameters, we consider the extended model (18). The same model structure (18) and priors are used for both approaches. For $(\mu, \beta, \Gamma^2, \delta, \alpha)$ the priors are as in (4). For η, the prior of Subsection 5.1 is chosen: for $m = 1, 2$, $\pi(\eta_m) = \mathcal{N}(0, \rho_m^2)$, with $\rho_m^2 > 0$ known. To stabilise the MCMC procedure, the prior on σ^2 is modified to an uniform distribution on $[0, 200]$ for both methods. This has very little consequence for the variable selection procedure proposed in this article. Indeed, the only difference lies in the updating of σ^2 at the M-step of the MCMC-SAEM algorithm which becomes:

$$\sigma^{2^{(k+1)}} = \begin{cases} \dfrac{s_{1,k+1}}{nJ} & \text{if } \dfrac{s_{1,k+1}}{nJ} \leq 200 \\ 200 & \text{else.} \end{cases}$$

The two methods are both initialised with: $\forall \ell \in \{1, \ldots, 10\}$, $\beta_\ell^{(0)} = 100$, $\forall \ell \in \{11, \ldots, p\}$, $\beta_\ell^{(0)} = 1$, $\mu^{(0)} = 1400$, $\sigma^{2^{(0)}} = 100$, $\alpha^{(0)} = 0.1$ and $\eta^{(0)} = {}^{\mathrm{t}}(400, 400)$. In practice, to avoid convergence toward a local maximum in the MCMC-SAEM algorithm, a simulated annealing version of SAEM (see Lavielle, 2014) is implemented. Thus, in this method, Γ^2 is initialised very large to explore the space during the first iterations, with $\Gamma^{2^{(0)}} = 5000$. For the full MCMC procedure, a more plausible value of Γ^2, $\Gamma^{2^{(0)}} = 500$, is chosen as initialisation. The hyperparameters are set in the same way for both methods as well: $\nu_0 = 0.04$, $\nu_1 = 12000$, $\sigma_\mu = 3000$, $\nu_\Gamma = \lambda_\Gamma = 1$, $a = 1$, $b = p$, $\Omega = \mathrm{diag}(20, 20)$ and $\rho_1^2 = \rho_2^2 = 1200$.

We compare the two approaches for a single value of ν_0, as it is standard practice to run an MCMC spike-and-slab model for a single value (see for instance George and McCulloch (1997) or Malsiner-Walli and Wagner (2018)). The MCMC algorithm was run for 3000 iterations, which was just enough to reach convergence (assessed by comparing multiple chains) for a variety of ν_0 and p values. The MCMC-SAEM algorithm was run for 500 iterations and showed appropriate convergence. Under these conditions, for all $p \in \{500, 700, 1000, 1500, 2000, 2500\}$, both methods were run for 50 different data-sets and the minimum time was kept for each method. The results obtained are shown in Figure 7. In this figure, computation times of full MCMC procedure (in purple) and of MCMC-SAEM (in blue) are represented by the points for the different values of p. The lines represent the regression line associated with each method. Note that a $\log_{10}$-$\log_{10}$ scale is used in this figure. This shows that both methods have an execution time that grows polynomially with p. Furthermore, the polynomial complexity of the two methods, *i.e.* the slope of the regression lines, is slightly lower for the MCMC-SAEM method. Thus, if we note respectively τ_{MCMC} and $\tau_{MCMC-SAEM}$ the execution time associated with each of the methods under the conditions previously described, empirically Figure 7 strongly suggests that $\dfrac{\tau_{MCMC}}{\tau_{MCMC-SAEM}} \approx 10^{0.7} p^{0.2}$. To sum up, the MCMC-SAEM algorithm proposed in this paper appears $10^{0.7} p^{0.2}$ times faster than the classical MCMC procedure, *i.e.* between 17 and 24 times faster for p between 500 and 2500. In other words, the proposed inference method allows to browse a grid of about 20 values of the penalisation parameter ν_0 while a classical MCMC only looked at one value of this parameter.

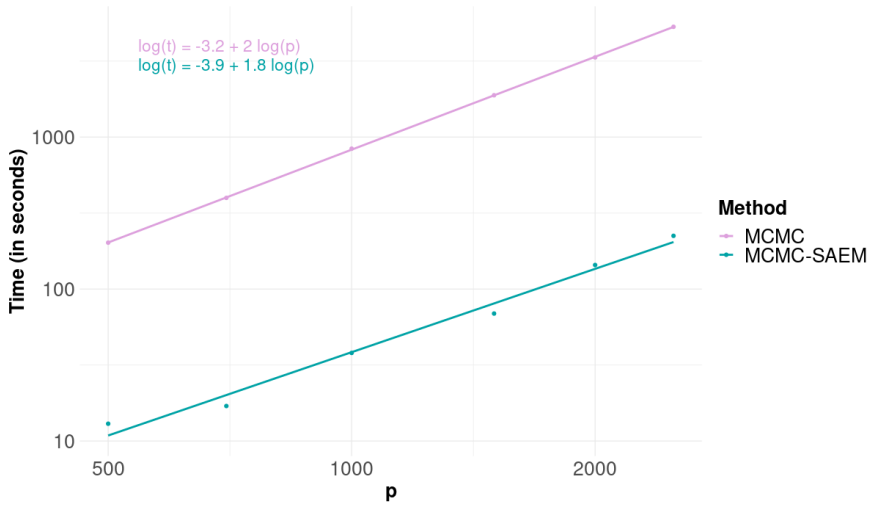


Figure 7: Comparison of computation times between MCMC (in purple) and Algorithm 1 (MCMC-SAEM, in blue) inference methods in $\log_{10}$-$\log_{10}$ scale.

6 Conclusion and perspectives

The main objective of this paper was to propose a new procedure for high-dimensional variable selection in non-linear mixed-effects models. In this work, variable selection is approached from a Bayesian perspective and a selection procedure combining the use of spike-and-slab Gaussian mixture prior and the SAEM algorithm is proposed. The spike-and-slab prior on the regression vector allows both the shrinkage towards zero of small non-significant coefficients through the spike distribution, while the largely uninformative slab distribution allows estimating influential covariates without bias from the penalisation. The speed of the SAEM algorithm allows to explore different levels of sparsity in the model through the variance of the spike distribution ν_0, with the optimal level of sparsity being selected by minimising an eBIC criterion.

The proposed methodology showed very good selection performance on simulated data. Indeed, the proposed procedure appears to select the right support in a large majority of cases. As expected, for different numbers of covariates p fixed, the right support is selected more often as the number of individuals n increases and the inter-individual variance Γ^2 decreases. Even more interesting, this method is much faster than an MCMC stochastic search alternative and can solve higher-dimensional variable selection problems.

In this work, the method has been restricted to the case where the individual parameters $(\varphi_i)_{1\leq i\leq n}$ are assumed to be real. However, in many fields of application, such as in biological growth or pharmacokinetics, the individual parameters are multiple. Thus, an important point for future research would be to adapt our procedure to be able to handle a case where the individual parameters are multivariate.

Moreover, it was observed that a reasonable correlation between covariates has little effect on the selection performances of the proposed procedure. However, when the level of correlation becomes high, the performance decreases. This could be improved if structural information on the covariates were *a priori* known. Indeed, in this article, through the i.i.d. Bernoulli prior on the indicators δ (4e), it is assumed that each covariate has the same probability of being included in the model. However, there are situations, such as genomic data, where certain covariates are *a priori* more likely to be included together in the model. This *a priori* structural information on the covariates can be taken into account in our procedure by choosing a more flexible prior on δ. In Stingo et al. (2010) and Stingo and Vannucci (2011), authors propose the independent logistic regression prior or the Markov random field prior. This could be also considered in our methodology.

Another important remark is that, in this article, we considered a Gaussian distribution for $p(y|\varphi,\sigma^2)$ in model (1). It is possible to relax this assumption and consider larger distribution classes, such as discrete distributions like the Poisson distribution for example. The proposed methodology can therefore be applied in many contexts.

7 Acknowledgements

This work was funded by the Stat4Plant project ANR-20-CE45-0012. We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources. All our experiments have been done on this platform.

References

Allassonnière, S. and Debavelaere, V. (2021). On the curved exponential family in the Stochatic Approximation Expectation Maximization Algorithm. hal-03128554.

Allassonnière, S., Kuhn, E., and Trouvé, A. (2010). Construction of Bayesian deformable models via a stochastic approximation algorithm: a convergence study. Bernoulli, 16(3):641–678.

Bai, R., Ročková, V., and George, E. I. (2021). Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO. In Handbook of Bayesian Variable Selection. Chapman and Hall/CRC.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. The annals of statistics, 32(3):870–897.

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.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. The Annals of Statistics, 40(4):2069–2101.

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

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. Journal of Computational and Graphical Statistics, 26:403–417.

Delattre, M., Lavielle, M., Poursat, M.-A., et al. (2014). A note on BIC in mixed-effects models. Electronic journal of statistics, 8(1):456–475.

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.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. Annals of statistics, 40(4):2043.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423):881–889.

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.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. Computational statistics & data analysis, 49(4):1020–1038.

Lavielle, M. (2014). Mixed effects models for the population approach: models, tasks, methods and tools. CRC press.

Lee, S. Y. (2022). Bayesian Nonlinear Models for Repeated Measurement Data: An Overview, Implementation, and Applications. Mathematics 2022, 10(898).

Malsiner-Walli, G. and Wagner, H. (2018). Comparing spike and slab priors for Bayesian variable selection. arXiv preprint arXiv:1812.07259.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. Journal of the american statistical association, 83(404):1023–1032.

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

Pinheiro, J. C. and Bates, D. M. (2000). Mixed-effects Models in S and S-PLUS. Springer.

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 l1-penalization. Scandinavian Journal of Statistics, 38(2):197–214.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, pages 461–464.

Shaby, B. and Wells, M. T. (2010). Exploring an adaptive Metropolis algorithm. Technical report, Department of Statistical Science, Duke University.

Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. The annals of applied statistics, 4(4):2024.

Stingo, F. C. and Vannucci, M. (2011). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. Bioinformatics, 27(4):495–501.

Tadesse, M. G. and Vannucci, M. (2021). Handbook of Bayesian variable selection. CRC Press.

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

Appendix

Appendix A Simulation example: differences in the procedure

As it is explained in Subsection 5.2, our procedure can easily be adapted to a model where fixed effects must be estimated. For this, within the framework of model (17), the extended model (18) is used to exponentialise the model. Indeed, by considering ψ as a latent variable with a normal distribution centred in η, an unknown parameter, and with a known covariance matrix Ω, and no longer as a parameter, it is possible to obtain an exponential form similar to Equation (12) for $\widetilde{Q}_1$. The calculation of the quantity Q of the EM algorithm becomes:

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

where:

$$\widetilde{Q}(y,\varphi,\psi,\Theta,\Theta^{(k)}) = \mathbb{E}_{\delta|(\varphi,\psi,y,\Theta^{(k)})}[\log(\pi(\Theta,\varphi,\psi,\delta|y))|\varphi,\psi,y,\Theta^{(k)}] = C + \widetilde{Q}_1(y,\varphi,\psi,\theta,\Theta^{(k)}) + \widetilde{Q}_2(\alpha,\Theta^{(k)}),$$

with

$$\begin{aligned}\widetilde{Q}_1(y,\varphi,\psi,\theta,\Theta^{(k)}) = &-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij}-g(\varphi_i,\psi,t_{ij}))^2 - \frac{1}{2\Gamma^2}||\varphi-\tilde{V}\tilde{\beta}||^2 - \frac{1}{2}\sum_{\ell'=0}^{p}\tilde{\beta}_{\ell'}^2\widetilde{d}_{\ell'}^*(\Theta^{(k)}) - \frac{nJ+\nu_\sigma+2}{2}\log(\sigma^2)\\ &-\frac{n+\nu_\Gamma+2}{2}\log(\Gamma^2) - \frac{\nu_\Gamma\lambda_\Gamma}{2\Gamma^2} - \frac{\nu_\sigma\lambda_\sigma}{2\sigma^2} - \sum_{m=1}^{2}\frac{(\psi_m-\eta_m)^2}{2\omega_m^2} - \sum_{m=1}^{2}\frac{\eta_m^2}{2\rho_m^2}\end{aligned}$$

and

$$\widetilde{Q}_2(\alpha,\Theta^{(k)}) = \log\left(\sqrt{\frac{\nu_0}{\nu_1}}\frac{\alpha}{1-\alpha}\right)\sum_{\ell=1}^{p}p_\ell^*(\Theta^{(k)}) + (a-1)\log(\alpha) + (p+b-1)\log(1-\alpha).$$

$(p_\ell^*(\Theta^{(k)}))_{1\le\ell\le p}$ and $(\widetilde{d}_{\ell'}^*(\Theta^{(k)}))_{0\le\ell'\le p}$ are defined in Proposition 3.1, Equations (9) and (10).

Note that $\widetilde{Q}_1$ is still of the exponential form. Indeed,

$$\widetilde{Q}_1(y,\varphi,\psi,\theta,\Theta^{(k)}) = -\Psi(\theta,\Theta^{(k)}) + \left\langle S(y,\varphi,\psi),\phi(\theta)\right\rangle \tag{19}$$

with:

- $S(y,\varphi,\psi) = \left(\sum_{i,j}(y_{ij}-g(\varphi_i,\psi,t_{ij}))^2\ ,\ \sum_{i=1}^n\varphi_i^2\ ,\ \varphi\ ,\ \psi^2\ ,\ \psi\right)$
- $\phi(\theta) = \left(-\frac{1}{2\sigma^2}\ ,\ -\frac{1}{2\Gamma^2}\ ,\ \frac{\tilde{V}\tilde{\beta}}{\Gamma^2}, \left(-\frac{1}{2\omega_m^2}\right)_{1\le m\le 2}, \left(\frac{\eta_m}{\omega_m^2}\right)_{1\le m\le 2}\right)$
- $\Psi(\theta,\Theta^{(k)}) = \frac{||\tilde{V}\tilde{\beta}||^2}{2\Gamma^2} + \frac{1}{2}\sum_{\ell'=0}^{p}\tilde{\beta}_{\ell'}^2\widetilde{d}_{\ell'}^*(\Theta^{(k)}) + \frac{nJ+\nu_\sigma+2}{2}\log(\sigma^2) + \frac{n+\nu_\Gamma+2}{2}\log(\Gamma^2) + \frac{\nu_\Gamma\lambda_\Gamma}{2\Gamma^2} + \frac{\nu_\sigma\lambda_\sigma}{2\sigma^2} + \sum_{m=1}^{2}\frac{\eta_m^2}{2\omega_m^2} + \sum_{m=1}^{2}\frac{\eta_m^2}{2\rho_m^2}$

The k-th iteration of the MCMC-SAEM algorithm on this model is therefore:

1. **S-Step:** simulate ($\varphi^{(k)}$,$\psi^{(k)}$) using the result of some iterations of a Metropolis-Hastings within Gibbs algorithm with $\pi(\varphi,\psi|y,\Theta^{(k)})$ for target distribution.
2. **SA-Step:** compute $S_{k+1} = S_k + \gamma_k(S(y,\varphi^{(k)},\psi^{(k)}) - S_k)$ with $S(y,\varphi,\psi)$ defined by (19), where $S_k = (s_{1,k},s_{2,k},s_{3,k},s_{4,k},s_{5,k}) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^2\times\mathbb{R}^2$.

3. **M-Step:** update $\theta^{(k+1)} = \underset{\theta\in\Lambda_\theta}{\text{argmax}} \left\{ -\psi(\theta,\Theta^{(k)}) + \langle S_{k+1}, \phi(\theta)\rangle \right\}$ and $\alpha^{(k+1)} = \underset{\alpha\in[0,1]}{\text{argmax}}\ \widetilde{Q}_2(\alpha,\Theta^{(k)})$.

   More precisely,

   - $\tilde{\beta}^{(k+1)} = ({}^{\mathrm{t}}\tilde{V}\tilde{V} + \Gamma^{2^{(k)}} D_k^*)^{-1}\, {}^{\mathrm{t}}\tilde{V} s_{3,k+1}$ where $D_k^* = \mathrm{diag}(\widetilde{d}^*_{\ell'}(\Theta^{(k)}), 0 \le \ell' \le p)$
   - $\Gamma^{2^{(k+1)}} = \dfrac{||\tilde{V}\tilde{\beta}^{(k+1)}||^2 + \nu_\Gamma\lambda_\Gamma + s_{2,k+1} - 2\langle s_{3,k+1}, \tilde{V}\tilde{\beta}^{(k+1)}\rangle}{n+\nu_\Gamma+2}$
   - $\sigma^{2^{(k+1)}} = \dfrac{\nu_\sigma\lambda_\sigma + s_{1,k+1}}{nJ+\nu_\sigma+2}$
   - $\eta_m^{(k+1)} = \dfrac{s_{5,k+1,m}}{1+\dfrac{\omega_m^2}{\rho_m^2}}$
   - $\alpha^{(k+1)} = \dfrac{\sum_{\ell=1}^p p_\ell^*(\Theta^{(k)}) + a - 1}{p+b+a-2}$

As you can see, $\widetilde{Q}_1$ is separable into $(\mu,\beta,\sigma^2,\Gamma^2,\alpha)$ and η, which means that the inference method used for the parameters $(\mu,\beta,\sigma^2,\Gamma^2,\alpha)$ is unchanged, *i.e.* the formulas to update these parameters in M-step are identical, it is only the way to simulate the sufficient statistics that has changed.

Thus, thanks to this algorithm, it is obtained an estimation $\hat{\theta}^{MAP}_{\nu_0} = (\widehat{\mu}^{MAP}_{\nu_0}, \widehat{\beta}^{MAP}_{\nu_0}, \widehat{\eta}^{MAP}_{\nu_0}, \widehat{\Gamma}^{2,MAP}_{\nu_0}, \widehat{\sigma}^{2,MAP}_{\nu_0})$, and the estimation of η is used as an estimation of ψ. Then, to finish the model collection reduction step of our procedure, Algorithm 2, the estimator $\widehat{\beta}^{MAP}_{\nu_0}$ is thresholded to obtain a promising sub-model $\widehat{S}_{\nu_0}$ given by Equation (14). The selection threshold formula is unchanged because it only depends on the second layer of the model (17).

For the model selection step, to compute the eBIC criterion, it is also necessary to go through the extended model (18). Indeed, as described in Kuhn and Lavielle (2005), the MLE in the sub-model $\widehat{S}_{\nu_0}$ is computed in the extended model by using an MCMC-SAEM algorithm and the estimation of η is used as an estimation of ψ. Then, the log-likelihood is approached by a Monte-Carlo method: for T large enough,

$$\log\left(p(y;\hat{\theta}^{MLE}_{\nu_0})\right) \approx \sum_{i=1}^n \log\left((2\pi\widehat{\sigma}^{2^{MLE}}_{\nu_0})^{-J/2} \frac{1}{T}\sum_{t=1}^T \exp\left(-\sum_{j=1}^J \frac{(y_{ij} - g(\varphi_i^{(t)}, \widehat{\psi}^{MLE}_{\nu_0}, t_{ij}))^2}{2\widehat{\sigma}^{2^{MLE}}_{\nu_0}} \right)\right)$$

where $p(y;\theta)$ denotes the likelihood of model (17), and for all $i \in \{1,\dots n\}$, $(\varphi_i^{(t)})_{t\in\{1,\dots T\}}$ are simulated i.i.d. according to $p(\varphi_i;\hat{\theta}^{MLE}_{\nu_0}) = \mathcal{N}(\widehat{\mu}^{MLE}_{\nu_0} + {}^{\mathrm{t}}\widehat{\beta}^{MLE}_{\nu_0} V_i, \widehat{\Gamma}^{2^{MLE}}_{\nu_0})$.