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

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

▶ To cite this version:

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

Marion Naveau[1,2*], Guillaume Kon Kam King[2], Renaud Rincent[3], Laure Sansonnet[1], Maud Delattre[2]

[1*]Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France.
[2]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.
[3]Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE - Le Moulon, 91190, Gif-sur-Yvette, France.

*Corresponding author(s). E-mail(s): marion.naveau@inrae.fr;

Abstract

High-dimensional variable selection, with many more covariates than observations, is widely documented in standard regression models, but there are still few tools to address it in non-linear mixed-effects models where data are collected repeatedly on several individuals. 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 and the Stochastic Approximation version of the Expectation Maximisation (SAEM) algorithm. Similarly to Lasso regression, the set of relevant covariates is selected by exploring a grid of values for the penalisation parameter. The SAEM approach is much faster than a classical MCMC (Markov chain Monte Carlo) algorithm and our method shows very good selection performances on simulated data. Its flexibility is demonstrated by implementing it for a variety of nonlinear mixed effects models. The usefulness of the proposed method is illustrated on a problem of genetic markers identification, relevant for genomic-assisted selection in plant breeding.

Keywords: High-dimension, Non-linear mixed-effects models, SAEM algorithm, Spike-and-slab prior, Variable selection

1

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, where data is customarily analysed using a nonlinear model whose coefficients often have a biological interpretation. In this case, the intrinsic variability of the data captured by the model parameters is then attributable to different sources (intra-individual, inter-individual, and 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 variable 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 true 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 contrast, in the more general framework of non-linear mixed-effects models (NLMEM), 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 to select covariates in pharmacokinetics studies, but without calibrating the penalty parameters. Ollier (2022) proposes a proximal gradient algorithm for computing a Lasso estimator in order to perform joint selection of covariates and correlation parameters between random effects, this method being more suitable in low rather than large dimensional covariate settings. The codes available for the above-mentioned methods are not generic and

2

tuning them is difficult. So the easiest practical solution for a practitioner faced with a high-dimensional problem is to adopt a two-stage approach, *i.e.* to fit independent nonlinear models to each individual and perform variable selection using the estimated parameters, losing the uncertainty on these estimates and the beneficial shrinkage property of mixed-effect models.

Bayesian approaches to variable selection have not received much 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 (EM Variable Selection) 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.

This paper is organised as follows. Section 2 describes the non-linear mixed-effects model to introduce the notation, 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 comparisons with existing methods. To illustrate the high-dimensional variable selection method proposed in this paper, Section 6 presents an application on a real data-set composed of European elite winter wheat varieties. Finally, Section 7 concludes with a summary discussion and prospects for future research. More details about algorithms and real data are postponed in the appendices.

2 Model description

2.1 Nonlinear 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. Let $n_{tot} = \sum_{i=1}^{n} n_i$ the total number of observations. Let $X^\top$ denote the transpose of a vector or matrix X. Consider the following non-linear mixed-effects model: for all $1 \leq i \leq n$ and $1 \leq j \leq n_i$,

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

This model is described in 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 q-dimensional. Thus, this function governs intra-individual behaviour. The value of q is closely linked to the choice of the mechanistic function g. It is therefore known, and usually small. 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 q individual parameters $\varphi_i = (\varphi_{im})_{1 \leq m \leq q} \in \mathbb{R}^q$ are modelled as a multivariate Gaussian random variable whose mean is specified as the sum of an intercept μ in $\mathbb{R}^q$ and a linear combination of known covariates measured on individual i and contained in the vector $V_i = (V_{i1}, \ldots, V_{ip})^\top \in \mathbb{R}^p$. 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, $\boldsymbol{\beta} = (\beta_{\ell m})_{1 \leq \ell \leq p;\ 1 \leq m \leq q} \in \mathcal{M}_{p \times q}$ is an unknown fixed effects matrix, and the inter-individual variance-covariance matrix Γ is assumed unknown. Thus, the inter-individual variability is separated into two parts: on the one hand, $\boldsymbol{\beta}$ 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, $y_i = (y_{ij})_{1 \leq j \leq n_i}$, $\boldsymbol{y} = (y_i)_{1 \leq i \leq n}$, $\boldsymbol{\varphi} = (\varphi_{im})_{1 \leq i \leq n;\ 1 \leq m \leq q} \in \mathcal{M}_{n \times q}$ and $V = (V_{i\ell})_{1 \leq i \leq n;\ 1 \leq \ell \leq p} \in \mathcal{M}_{n \times p}$ respectively denote the vector of observations for individual i, the vector of all observations, the matrix of all individual parameters, and the matrix of all covariates. Let us also note $\theta = (\mu, \boldsymbol{\beta}, \Gamma, \sigma^2)$ 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 $\boldsymbol{\beta}$. Indeed, for each $(\ell, m) \in \{1, \dots, p\} \times \{1, \dots, q\}$, coefficient $\beta_{\ell m}$ describes the influence of covariate ℓ on the individual values of the m-th parameter of the nonlinear curves $(\varphi_{im})_{1 \leq i \leq n}$. More precisely, $\beta_{\ell m} = 0$ means that covariate ℓ has no effect on $(\varphi_{im})_{1 \leq i \leq n}$ whereas $\beta_{\ell m} \neq 0$ means that covariate ℓ gives some information on these parameters. Identifying the relevant covariates for all individual parameters amounts to selecting the support of $\boldsymbol{\beta}$, defined by S^*:

$$S^* = \left\{ (\ell, m) \in \{1, \dots, p\} \times \{1, \dots, q\} \middle| \beta_{\ell m}^* \neq 0 \right\},$$

where $\boldsymbol{\beta}^*$ is the true fixed effects matrix. To solve this problem in a high-dimensional context, that is when $p >> n$, it is natural to assume that each row of $\boldsymbol{\beta}^*$ is sparse, which means that many $\beta_{\ell m}^*$ 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 $\boldsymbol{\varphi}$. The main difficulty here is that variable selection concerns latent variables of the model.

Remark 1 As in the application presented in Section 6, we could perform variable selection on $q' < q$ components of φ_i, with the choice of the q' relevant components to be considered for variable selection driven by the initial biological question without changing the methodology presented in what follows.

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 $\theta = (\mu, \boldsymbol{\beta}, \Gamma, \sigma^2)$. First, in order to find the non-zero coefficients of $\boldsymbol{\beta}$, a spike-and-slab mixture prior (George and McCulloch, 1993, 1997; Ročková and George, 2014) is considered in a multivariate setting. To facilitate the formulation of this prior, binary latent variables $\boldsymbol{\delta} = (\delta_{\ell m})_{1 \leq \ell \leq p;\ 1 \leq m \leq q}$ are introduced, such as:

$$\forall 1 \leq \ell \leq p \ , \ \forall 1 \leq m \leq q \ , \ \delta_{\ell m} = \begin{cases} 1 \text{ if } (\ell, m) \text{ is to be included in model } S^* \ , \\ 0 \text{ otherwise.} \end{cases} \quad (2)$$

Thus, $\delta_{\ell m} = 1$ indicates that the covariate ℓ provides information on the individual parameter m. In other words, $\boldsymbol{\delta}$ characterises the support of $\boldsymbol{\beta}$. Note that contrary to many multivariate variable selection methods such as the one implemented in **glmnet** (Friedman et al., 2010), the proposed procedure is not limited to selecting the same set of covariates for all dimensions m, i.e it is not required that $\forall 1 \leq m \leq q, \quad \delta_{\ell m} = \delta_\ell$. The support S^* can therefore be reformulated as follows:

$$S^* = \left\{ (\ell, m) \in \{1, \dots, p\} \times \{1, \dots, q\} \middle| \delta_{\ell m}^* = 1 \right\}, \quad (3)$$

5

where $\boldsymbol{\delta}^*$ denotes the true support. Then, one would like to find $\widehat{\boldsymbol{\delta}}$ that maximises the posterior probability $\pi(\boldsymbol{\delta}|\boldsymbol{y})$, which corresponds to the most promising support.

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

$$\pi(\beta_{\ell m}|\delta_{\ell m}) = \mathcal{N}(0, (1 - \delta_{\ell m})\nu_0 + \delta_{\ell m}\nu_1),\ 1 \leq \ell \leq p,\ 1 \leq m \leq q,\ 0 < \nu_0 < \nu_1, \quad (4a)$$

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

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

$$\pi(\Gamma) = \mathcal{IW}(\Sigma_\Gamma, d),\ \text{with}\ \Sigma_\Gamma \in \mathcal{S}_q^{++}, d > 0, \quad (4d)$$

$$\pi(\delta_{\ell m}|\alpha_m) = \alpha_m^{\delta_{\ell m}}(1 - \alpha_m)^{1 - \delta_{\ell m}},\ \text{with}\ \alpha_m \in [0,1], 1 \leq \ell \leq p, 1 \leq m \leq q, \quad (4e)$$

$$\pi(\alpha_m) = Beta(a_m, b_m),\ \text{with}\ a_m, b_m > 0,\ 1 \leq m \leq q. \quad (4f)$$

The key prior distribution used for variable selection in this method is the spike-and-slab Gaussian mixture prior (4a) on $\boldsymbol{\beta}$. In this prior, ν_0 and ν_1 are parameters controlling the penalisation inducing sparsity in the columns of $\boldsymbol{\beta}$. More precisely, $(\beta_{\ell m})_{1 \leq \ell \leq p;\ 1 \leq m \leq q}$ are independent conditionally on $\boldsymbol{\delta}$, with $\pi(\beta_{\ell m}|\delta_{\ell m} = 0) = \mathcal{N}(0, \nu_0)$ and $\pi(\beta_{\ell m}|\delta_{\ell m} = 1) = \mathcal{N}(0, \nu_1)$. 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 $\boldsymbol{\beta}$ values (see George and McCulloch, 1997). Indeed, when $\delta_{\ell m} = 0$, the prior constrains $\beta_{\ell m}$ to very small values which implies that covariate ℓ has no impact in the individual parameter m in the model. Thus, through the values $\delta_{\ell m}$, the spike-and-slab prior makes it possible to distinguish the selected covariates from the rest.

Note that, since $\boldsymbol{\varphi}$ 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 the 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 $\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \mu^\top \\ \boldsymbol{\beta} \end{pmatrix} \in \mathcal{M}_{(p+1) \times q}$ and, for all $1 \leq i \leq n$, $\tilde{V}_i = (\tilde{V}_{i\ell'})_{1 \leq \ell' \leq p+1} = (1, V_i)^\top \in \mathbb{R}^{p+1}$, so that $\mu + \boldsymbol{\beta}^\top V_i = \tilde{\boldsymbol{\beta}}^\top \tilde{V}_i$. Let $\tilde{V} = (\tilde{V}_{i\ell'})_{1 \leq i \leq n;\ 1 \leq \ell' \leq p+1}$ such as $\boldsymbol{\varphi} = \tilde{V}\tilde{\boldsymbol{\beta}} + \xi$, where $\xi = \begin{pmatrix} \xi_1^\top \\ \vdots \\ \xi_n^\top \end{pmatrix} \in \mathcal{M}_{n \times q}$. Then, by introducing $\tilde{\boldsymbol{\delta}} = (\tilde{\delta}_{\ell' m})_{1 \leq \ell' \leq p+1;\ 1 \leq m \leq q}$ such as:

$$\forall 1 \leq \ell' \leq p + 1\ ,\ \forall 1 \leq m \leq q\ ,\ \tilde{\delta}_{\ell' m} = \begin{cases} 1 & \text{if}\ \ell' = 1, \\ \delta_{\ell m} & \text{where}\ \ell = \ell' - 1,\ \text{if}\ \ell' > 1, \end{cases}$$

to force the inclusion of the intercept in the model, Equations (4a) and (4b) can be rewritten as: for $1 \leq \ell' \leq p+1$, $1 \leq m \leq q$,

$$\pi(\tilde{\beta}_{\ell'm}|\tilde{\delta}_{\ell'm}) = \mathcal{N}(0, (1 - \tilde{\delta}_{\ell'm})\nu_0 + \tilde{\delta}_{\ell'm}(\mathbb{1}_{\ell'>1}\nu_1 + \mathbb{1}_{\ell'=1}\sigma_\mu^2)) \tag{5}$$

For the variance parameter σ^2, an inverse-gamma prior is chosen (4c), which prohibits negative values. One possibility is to set $\nu_\sigma, \lambda_\sigma$ equal to 1 for example, to make it relatively non-influential. For the inter-individual variance-covariance matrix Γ, the inverse-Wishart prior is chosen (4d), which is the multivariate extension of the inverse-Gamma density. The hyperparameter Σ_Γ is a positive matrix which can be specified as $\Sigma_\Gamma = kI_q$ with k chosen of comparable size to the maximum likely variance of $(\varphi_{im})_{1 \leq i \leq m}$ for all $m \in \{1, \ldots, q\}$. The hyperparameter d, which is a degree of freedom, can be chosen as the smallest integer value ensuring the existence of $\mathbb{E}[\Gamma]$, that is $q + 2$.

Following Ročková and George (2014), the i.i.d. Bernoulli prior (4e) is used for the inclusion variable $\boldsymbol{\delta}$, where the hyperparameters $(\alpha_m)_m$ can be seen as the proportion of relevant covariates for each individual parameter, and a Beta distribution prior (4f) is chosen on each α_m, for $m \in \{1, \ldots, q\}$. To encourage sparsity in the model, Castillo and van der Vaart (2012) suggest choosing a_m small and b_m large, for example $a_m = 1$ and $b_m = p$ for all m. In the following, α, a and b respectively denote the vectors $(\alpha_m)_{1 \leq m \leq q}$, $(a_m)_{1 \leq m \leq q}$ and $(b_m)_{1 \leq m \leq q}$. See Ročková and George (2014), Liquet et al. (2017), Deshpande et al. (2019), 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 $\Theta = (\theta, \alpha) = (\tilde{\boldsymbol{\beta}}, \Gamma, \sigma^2, \alpha)$ in model (1) - (4). Recall that $\Xi = (\nu_0, \nu_1, \sigma_\mu^2, \nu_\sigma, \lambda_\sigma, \Sigma_\Gamma, d, a, b)$ are fixed hyperparameters. In the following, model (1) - (4) is called SSNLME (Spike-and-Slab Non-Linear Mixed-Effects) model. Note that $\boldsymbol{\varphi}$, 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 5.3).

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

$$\widehat{\Theta}^{MAP} = \underset{\Theta \in \Lambda}{argmax} \quad \pi(\Theta|\boldsymbol{y}), \text{ with } \pi(\Theta|\boldsymbol{y}) = \frac{p_\Theta(\boldsymbol{y})\pi(\Theta)}{\int_\Lambda p_\Theta(\boldsymbol{y})\pi(\Theta)d\Theta} , \tag{6}$$

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

$\pi(\Theta|\boldsymbol{y})$ takes the form:

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

where $p(\boldsymbol{y}|\Theta, Z)$ and $p(\Theta, Z)$ respectively designate the probability density of $\boldsymbol{y}$ 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 (6) 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 estimator (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 $\boldsymbol{y}$ and latent variables Z that characterise the distribution of observations. It it assumed that the density of the complete data $(\boldsymbol{y}, Z)$ 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

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

(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:

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

the conditional expectation of the complete log-posterior $\log(\pi(\Theta, Z|\boldsymbol{y}))$ given the observations $\boldsymbol{y}$ 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. However, even though this expectation cannot be computed in closed-form, it can be approximated by simulation. One solution proposed by Wei and Tanner (1990) is to replace the E-step, *i.e.* the computation of the Q quantity, by a Monte-Carlo approximation based on a large number of independent simulations of the latent variables Z, called the MCEM algorithm. However, this large number of simulations is computationally expensive. The SAEM algorithm is an other alternative that replaces the E-step by a stochastic approximation based on a single simulation of the latent variables, which considerably reduces the computational cost (Delyon et al., 1999). 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 $\pi(Z|\boldsymbol{y}, \Theta^{(k)})$.
2. **SA-step (Stochastic Approximation):** update the approximation $Q_{k+1}(\Theta)$ of $Q(\Theta|\Theta^{(k)})$ by a stochastic approximation method, according to:

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

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

$$\log(\pi(\Theta, Z|\boldsymbol{y})) = -\psi(\Theta) + \left\langle S(\boldsymbol{y}, Z), \phi(\Theta) \right\rangle,$$

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

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

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 $\mathbb{E}_{Z|(\boldsymbol{y},\Theta^{(k)})}[S(\boldsymbol{y}, Z)|\boldsymbol{y}, \Theta^{(k)}]$, according to:
$$S_{k+1} = S_k + \gamma_k(S(\boldsymbol{y}, Z^{(k)}) - S_k).$$

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

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

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.

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

9

an alternative which consists in coupling the SAEM method with an MCMC procedure. Interestingly, convergence of the MCMC part at each S-step is not necessary and only a few MCMC iterations are required in practice (see Kuhn and Lavielle (2004) and Kuhn and Lavielle (2005) for practical and theoretical considerations about this algorithm). In the following, this extension is called MCMC-SAEM.

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 $\boldsymbol{y} = (y_{ij})_{i,j}$ and $Z = (\boldsymbol{\varphi}, \boldsymbol{\delta})$. The aim here is to decompose the Q quantity of the SAEM algorithm in the particular case of the SSNLME model, allowing to describe an algorithm for computing the MAP estimator of Θ in the following subsection.

First, note that, by using the tower property of conditional expectation, the quantity $Q(\Theta|\Theta^{(k)})$ in model (1) - (4) is written as:

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

where

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

It is interesting to write $Q(\Theta|\Theta^{(k)})$ like this because $\widetilde{Q}(\boldsymbol{y},\boldsymbol{\varphi},\Theta,\Theta^{(k)})$ has a closed form. Indeed, Proposition 1 shows that $\widetilde{Q}(\boldsymbol{y},\boldsymbol{\varphi},\Theta,\Theta^{(k)})$ can be decomposed such as:

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

where $\widetilde{Q}_1$ and $\widetilde{Q}_2$ have a closed form given in Proposition 1. Thus, the separability of (8) into two distinct functions, $\widetilde{Q}_1$ which depends on $(\boldsymbol{y},\boldsymbol{\varphi},\theta,\Theta^{(k)})$ 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 $\boldsymbol{\varphi}$, Proposition 1 allows to write that:

$$Q(\Theta|\Theta^{(k)}) = C + \mathbb{E}_{\boldsymbol{\varphi}|(\boldsymbol{y},\Theta^{(k)})}\left[\widetilde{Q}_1(\boldsymbol{y},\boldsymbol{\varphi},\theta,\Theta^{(k)})\middle|\boldsymbol{y},\Theta^{(k)}\right] + \widetilde{Q}_2(\alpha,\Theta^{(k)}). \qquad (9)$$

However, even if $\widetilde{Q}(\boldsymbol{y},\boldsymbol{\varphi},\Theta,\Theta^{(k)})$ 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 $\pi(\boldsymbol{\varphi}|\boldsymbol{y},\Theta^{(k)})$ is only known to a nearest multiplicative constant. Thus, it is necessary to use a stochastic approximation method to approximate $\mathbb{E}_{\boldsymbol{\varphi}|(\boldsymbol{y},\Theta^{(k)})}\left[\widetilde{Q}_1(\boldsymbol{y},\boldsymbol{\varphi},\theta,\Theta^{(k)})\middle|\boldsymbol{y},\Theta^{(k)}\right]$ in

10

Equation (9). 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 $\mathbb{E}_{\boldsymbol{\varphi}|(\boldsymbol{y},\Theta^{(k)})}\left[\widetilde{Q}_1(\boldsymbol{y},\boldsymbol{\varphi},\theta,\Theta^{(k)})\Big|\boldsymbol{y},\Theta^{(k)}\right]$ 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 $\widetilde{Q}_1(\boldsymbol{y},\boldsymbol{\varphi},\theta,\Theta^{(k)})$ takes an exponential form. Thus, according to Remark 2, it suffices to approximate stochastically $\mathbb{E}_{\boldsymbol{\varphi}|(\boldsymbol{y},\Theta^{(k)})}\left[S(\boldsymbol{y},\boldsymbol{\varphi})\big|\boldsymbol{y},\Theta^{(k)}\right]$ at SA-step (see Appendix A.1 for more details about $S(\boldsymbol{y},\boldsymbol{\varphi})$ and this exponential form).

Algorithm 1 in Appendix A.1 summarises the proposed extension of the MCMC-SAEM algorithm for computing the MAP estimator of Θ in the SSNLME model.

Proposition 1 *Consider $\widetilde{Q}(\boldsymbol{y},\boldsymbol{\varphi},\Theta,\Theta^{(k)})$ defined by Equation (7) where $\Theta = (\tilde{\boldsymbol{\beta}}, \Gamma, \sigma^2, \alpha)$. Then:*

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

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

$$\widetilde{Q}_1(\boldsymbol{y},\boldsymbol{\varphi},\theta,\Theta^{(k)}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - g(\varphi_i, t_{ij}))^2 - \frac{n_{tot} + \nu_\sigma + 2}{2}\log(\sigma^2) - \frac{\nu_\sigma \lambda_\sigma}{2\sigma^2}$$

$$- \frac{1}{2}\text{Tr}\left((\boldsymbol{\varphi} - \tilde{V}\tilde{\boldsymbol{\beta}})^\top(\boldsymbol{\varphi} - \tilde{V}\tilde{\boldsymbol{\beta}})\Gamma^{-1}\right) - \frac{1}{2}\sum_{m=1}^{q}\sum_{\ell'=1}^{p+1}\tilde{\beta}_{\ell'm}^2\widetilde{d}_{\ell'm}^*(\Theta^{(k)})$$

$$- \frac{n+d+q+1}{2}\log(|\Gamma|) - \frac{1}{2}\text{Tr}(\Sigma_\Gamma \Gamma^{-1})$$

where $|A|$ and $\text{Tr}(A)$ respectively denote the determinant and the trace of a matrix A, and

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

$$(a_m - 1)\log(\alpha_m) + (p + b_m - 1)\log(1 - \alpha_m).$$

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

$$p_{\ell m}^*(\Theta^{(k)}) = \mathbb{E}[\delta_{\ell m}|\boldsymbol{\varphi},\boldsymbol{y},\Theta^{(k)}] = \frac{\alpha_m^{(k)}\phi_{\nu_1}(\beta_{\ell m}^{(k)})}{\alpha_m^{(k)}\phi_{\nu_1}(\beta_{\ell m}^{(k)}) + (1 - \alpha_m^{(k)})\phi_{\nu_0}(\beta_{\ell m}^{(k)})} \tag{10}$$

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

$$\widetilde{d}_{\ell'm}^*(\Theta^{(k)}) = \mathbb{E}\left[\frac{1}{(1 - \tilde{\delta}_{\ell'm})\nu_0 + \tilde{\delta}_{\ell'm}(\mathbb{1}_{\ell'>1}\nu_1 + \mathbb{1}_{\ell'=1}\sigma_\mu^2)}\bigg|\boldsymbol{\varphi},\boldsymbol{y},\Theta^{(k)}\right]$$

$$= \frac{1}{\sigma_\mu^2}\mathbb{1}_{\ell'=1} + \left(\frac{1 - p_{\ell m}^*(\Theta^{(k)})}{\nu_0} + \frac{p_{\ell m}^*(\Theta^{(k)})}{\nu_1}\right)\mathbb{1}_{\ell'\neq 1} \tag{11}$$

where $\ell = \ell' - 1$.

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

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.

3.3 Estimator thresholding

As in Ročková and George (2014), after obtaining an estimator $\widehat{\Theta}^{MAP}$, the support S^*, defined in Equation (3), can be naturally estimated as the most probable model conditionally on $\widehat{\Theta}^{MAP}$. Indeed, for all $m \in \{1, \ldots, q\}$ and for all $\ell \in \{1, \ldots, p\}$, since $\pi(\delta_{\ell m} = 1|\widehat{\alpha}_m^{MAP}) = \widehat{\alpha}_m^{MAP}$ and $\pi(\delta_{\ell m} = 0|\widehat{\alpha}_m^{MAP}) = 1 - \widehat{\alpha}_m^{MAP}$, the *a posteriori* inclusion probability of the covariate ℓ knowing $\widehat{\Theta}^{MAP}$ for the individual parameter m can be obtained as:

$$\mathbb{P}(\delta_{\ell m} = 1|\boldsymbol{y}, \widehat{\beta}_{\ell m}^{MAP}, \widehat{\alpha}_m^{MAP}) = \frac{\pi_1(\widehat{\beta}_{\ell m}^{MAP})\widehat{\alpha}_m^{MAP}}{\pi_1(\widehat{\beta}_{\ell m}^{MAP})\widehat{\alpha}_m^{MAP} + \pi_0(\widehat{\beta}_{\ell m}^{MAP})(1 - \widehat{\alpha}_m^{MAP})},$$

where $\pi_k(\widehat{\beta}_{\ell m}^{MAP}) = \pi(\widehat{\beta}_{\ell m}^{MAP}|\delta_{\ell m} = k)$ for $k \in \{0, 1\}$. Then, $\widehat{\boldsymbol{\delta}}$, which is the most probable $\boldsymbol{\delta}$ knowing that $\Theta = \widehat{\Theta}^{MAP}$, can be computed as follows:

$$\widehat{\delta}_{\ell m} = 1 \Longleftrightarrow \mathbb{P}(\delta_{\ell m} = 1|\boldsymbol{y}, \widehat{\beta}_{\ell m}^{MAP}, \widehat{\alpha}_m^{MAP}) \geq 0.5$$

$$\Longleftrightarrow |\widehat{\beta}_{\ell m}^{MAP}| \geq \sqrt{2\frac{\nu_0 \nu_1}{\nu_1 - \nu_0} \log\left(\sqrt{\frac{\nu_1}{\nu_0}}\frac{1 - \widehat{\alpha}_m^{MAP}}{\widehat{\alpha}_m^{MAP}}\right)} = s_\beta(\nu_0, \nu_1, \widehat{\alpha}_m^{MAP}).$$

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 for the individual parameter m is selected via a thresholding operation:

$$\widehat{S} = \left\{(\ell, m) \in \{1, \ldots, p\} \times \{1, \ldots, q\} \ \middle| \ |\widehat{\beta}_{\ell m}^{MAP}| \geq s_\beta(\nu_0, \nu_1, \widehat{\alpha}_m^{MAP})\right\}. \tag{12}$$

Remark 5 Note that threshold $s_\beta(\nu_0, \nu_1, \widehat{\alpha}_m^{MAP})$ is the same for all the covariates but depends on the individual parameter m and 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

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 $\boldsymbol{\beta}$'s columns, 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 fewer covariates are included in the estimated support of $\boldsymbol{\beta}$'s columns. This is why it is more interesting to look at a grid of values and then use a model selection criterion to choose the optimal model. Let us denote 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 for all individual parameters, $\widehat{S}_{\nu_0}$, as explained by Equation (12) in Subsection 3.3. This first step reduces the total collection of 2^{pq} possible models to a smaller collection of $|\Delta| \ll 2^{pq}$ promising submodels $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$ 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 $\boldsymbol{\delta}$ 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. Thus, covariate selection would consist in choosing the "best" $\nu_0 \in \Delta$, that is noted $\hat{\nu}_0$, as:

$$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\operatorname{argmin}} \left\{ \operatorname{crit}(\widehat{S}_{\nu_0}) \right\}, \tag{13}$$

where

$$\operatorname{crit}(\widehat{S}_{\nu_0}) = -2 \log \left(p(\boldsymbol{y};\ \hat{\theta}_{\nu_0}) \right) + \operatorname{pen}(\nu_0), \tag{14}$$

with:

- $\log \left(p(\boldsymbol{y};\ \theta) \right)$ the log-likelihood of model (1),
- $\hat{\theta}_{\nu_0} = (\widehat{\tilde{\boldsymbol{\beta}}}_{\nu_0}, \widehat{\Gamma}_{\nu_0}, \widehat{\sigma}^2_{\nu_0})$ a point estimator of the parameter $\theta = (\tilde{\boldsymbol{\beta}}, \Gamma, \sigma^2)$ in sub-model $\widehat{S}_{\nu_0}$,
- $\operatorname{pen}(\nu_0)$ a penalty function which penalizes the complexity of $\widehat{S}_{\nu_0}$.

There are many ways to define the penalty in the criterion (14), e.g. AIC (Akaike, 1998), BIC (Schwarz, 1978; Delattre et al., 2014), DIC (Spiegelhalter et al., 2002), etc. Here, a pragmatic and effective choice is to use the eBIC (extended Bayesian Information Criterion, Chen and Chen (2008)) which is tailored to the high-dimensional setting. Indeed, eBIC's penalty has the following form:

$$\text{pen}_{\text{eBIC}}(\nu_0) = B_{\nu_0} \times \log(n) + 2 \log \left(\binom{pq}{|\widehat{S}_{\nu_0}|} \right), \tag{15}$$

where $|\widehat{S}_{\nu_0}|$ is the size of this support, which allows to take into account that the number of possible models with $r \leq pq$ covariates increases quickly as r increases. The eBIC uses $\hat{\theta}_{\nu_0}$ the maximum likelihood estimate (MLE). Note that this MLE and log-likelihood that are required to compute criteria with a penalty of the form (15) 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. They are calculated using an MCMC-SAEM algorithm and importance sampling techniques respectively (see *e.g.* Kuhn and Lavielle (2005) and Lavielle (2014) for details). Note that some Bayesian model selection criteria could also be appropriate instead of eBIC.

The proposed variable selection procedure can be summarised as in Algorithm 2 in Appendix A.2. In the following, this procedure is called SAEMVS (SAEM Variable Selection). A detailed example of the application of this procedure on a toy example is presented in a supporting web material given in the section "Data and code availability".

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.

5 Numerical experiments

This section studies the performance of the global selection procedure, named SAEMVS, in the non-linear mixed effects model with spike-and-slab prior, called the SSNLME model. As a reminder, this procedure is summarised in Algorithm 2 in Appendix A.2, and uses an MCMC-SAEM algorithm for inference, which is detailed in Algorithm 1 in Appendix A.1.

The numerical study is divided into three parts. The first part is a comparison with strategies that can be easily implemented from existing methods. It aims essentially at demonstrating the interest of carrying out the selection of covariates from the data of all the individuals simultaneously thanks to the mixed effects model. The second part studies in great detail the influences of the number of subjects, the number of covariates, the signal-to-variability ratio and the collinearity between covariates on the performance of SAEMVS. To both show the flexibility of our approach and simplify the presentation of this study, this second part is conducted on another nonlinear mixed effects model, this time with a one-dimensional random effect. In the third part, a comparison of SAEMVS in terms of computation time with an MCMC implementation is presented, to quantify precisely the speed improvement afforded by the SAEM algorithm.

5.1 Comparison with a two-step approach

The proposed approach is compared to a standard solution readily available to a practitioner using existing tools. The goal being to study the impact of covariates on specific parameters of the nonlinear models, a manageable approach would be to proceed in two steps and fit independently the nonlinear model to each individual, then perform variable selection using as dependent variables the estimated parameters. This second step can be carried out using for instance the popular **glmnet** package, which allows multivariate response variable selection. This strategy is expected to work fine in data-rich scenarios when each parameter can be estimated very precisely, but it loses the uncertainty on the estimated parameters and the shrinkage property of the mixed-effect model. We are not aware of alternative solutions with freely available code that a practitioner could use for a high-dimensional problem as discussed in the introduction. The interest of SAEMVS, which considers the mixed effect model and fully embraces the uncertainty in the estimations, is shown in the context of repeated measures. We show that SAEMVS performs better than the two-steps approach on two grounds:

- when the estimation of the individual parameters is more challenging, for instance in the classical scenario of lost to follow-up patients where some observation times may be missing,
- when different sets of covariates have an impact on the different dimensions of the response.

5.1.1 Model and simulation design

The following model, commonly used in pharmacokinetics, is considered:

$$\begin{cases} y_{ij} = \dfrac{D\varphi_{i1}}{\mathcal{V}}\varphi_{i1} - \varphi_{i2}\left(\mathrm{e}^{-\dfrac{\varphi_{i2}}{\mathcal{V}}t_{ij}} - \mathrm{e}^{-\varphi_{i1}t_{ij}}\right) + \varepsilon_{ij}, \ \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2), \\ \varphi_i = \mu + \boldsymbol{\beta}^\top V_i + \xi_i, \hspace{3.5cm} \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(0,\Gamma), \end{cases} \tag{16}$$

where $\varphi_i = (\varphi_{i1}, \varphi_{i2})^\top$, namely $q = 2$. The constants D and $\mathcal{V}$ are set to 100 and 30 respectively. The aim is to obtain a set of active covariates for the two individual parameters. For this study, the parameters are set to: $n = 200$ individuals, $p = 500$ covariates, $\sigma^2 = 10^{-3}$, $\Gamma = \begin{pmatrix} 0.2 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$ so that the correlation between the two individual parameters is of the order of 0.35, $\mu = (6,8)^\top$, and $\boldsymbol{\beta} = \begin{pmatrix} 3 & 2 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 3 & 2 & 1 & 0 & \dots & 0 \end{pmatrix}^\top$.

The individual covariates $V_i \in \mathbb{R}^p$, $1 \leq i \leq n$, are simulated independently according to a binomial distribution with a success probability of 0.2. Note that the support of the $\boldsymbol{\beta}$ is not the same for each dimension. It seems unrealistic that the covariates impacting φ_{i1} would be exactly the same as those impacting φ_{i2}. The covariates are standardised. Thus, 100 data-sets are simulated according to these parameter values where the number of observations by individual is fixed to $n_1 = \dots = n_n = 12$ with the following observation time points: $(t_{i1}, \dots, t_{i12}) = (0.05, 0.15, 0.25, 0.4, 0.5, 0.8, 1, 2, 7, 12, 24, 40)$.

This comparison was carried out on different scenarios, one corresponding to a data-rich situation where the two-step approach is expected to work relatively well, and a more challenging scenario where it is expected that there will be a difference between the two-step method and SAEMVS. The two scenarios correspond to different observation periods for each individual:

1. **Complete data-set.** This is the baseline scenario where all individuals are observed during the entire experiment.
2. **Partial observations.** For each $p_{\text{partial}} \in \{0.1, 0.2, 0.3, 0.4\}$, the other scenarios correspond to the case where $N_1 = p_{\text{partial}}n$ individuals are assumed to be no longer part of the experiment after the 3rd observation time, that is: for each previously simulated data-set, only the first 3 observation times are kept for the first N_1 individuals, and all observation times for the remaining $N_2 = n - N_1$ individuals.

5.1.2 Competing methods

For this comparison study, SAEMVS is compared to two other procedures. For both of these procedures, the first step consists in estimating the φ_i's individual-by-individual with the method of least squares thanks to the **nlm** R function (Non-Linear Minimization, see Schnabel et al. (1985)). Then, the Lasso method from the **glmnet** R package (Friedman et al., 2010) is applied in its multivariate and univariate versions on the second level of the model (16) using the estimated φ_i's:

- **Multivariate setting.** To take into account the correlation between the two individual parameters, the multi-response Gaussian family is used for the **glmnet** function on the estimated φ_i's. Note that this function uses a group Lasso penalty, forcing the two individual parameters to have the same support. This method is called "mgaussian" in the following.
- **Univariate setting.** To circumvent this constraint, we also compare SAEMVS to the case where selection is made on the two individual parameters independently. For this, the **glmnet** function with the Gaussian family is used for each of the individual parameters separately on the estimated φ_{i1}'s and φ_{i2}'s. This method is called "gaussian" in the following.

For both of these methods, a cross-validation procedure is performed using the **cv.glmnet** function and the largest λ at which the mean squared error (MSE) is within one standard error of the smallest MSE is chosen as recommended in Hastie et al. (2009). The algorithmic settings of SAEMVS are given in Appendix A.3 with an example of convergence graphs of the MCMC-SAEM algorithm.

5.1.3 Results

First, Table 1 compares the mean estimation errors for the two individual parameters using the first step of the two-step approach. It should be noted that 3 observation times appear sufficient to estimate the first parameter accurately, but insufficient for the second, as can be seen with the increasing estimation error for decreasing amount of data in Table 1.

Table 1 Comparison of the mean estimation errors for the first individual parameter (MEE1) and the second (MEE2) calculated on all individuals over the 100 data-sets using the first step of the two-step approach.

p_{partial}	0	0.1	0.2	0.3	0.4
MEE1[a]	0.088	0.10	0.10	0.11	0.12
MEE2[b]	0.12	0.62	1.14	1.68	2.20

[a]MEE1 is the mean of the difference in absolute value between the true φ_{i1} and its estimate over all the individuals and the 100 data-sets.
[b]MEE2 is the mean of the difference in absolute value between the true φ_{i2} and its estimate over all the individuals and the 100 data-sets.

Thus, it is expected that the number of partially observed individuals will have a negative impact on the estimated support of the second individual parameter. Indeed, this is what is observed in Figure 1 for gaussian and mgaussian methods. On this figure, for the first individual parameter (graph a), we can see that gaussian and mgaussian methods select a model that almost always includes the true support (striped bars). However, mgaussian method almost never selects the true model and the gaussian method only in one case out of 2. In contrast, SAEMVS selects exactly the right model (unpatternedbars) in a large majority of cases (about 90%) with no false positives. Note that in many applications, especially in biology, false positives are to be avoided due to the cost of the experiments, and therefore SAEMVS seems to be more efficient in this regard. For the second individual parameter (graph b), the methods gaussian and mgaussian suffer greatly from the increase in the number of partially observed individuals, whereas the mixed effect model structure in SAEMVS shows greater robustness thanks to the classical pooling of information phenomenon, whereby individuals with missing data can benefit from the remaining fully observed individuals.
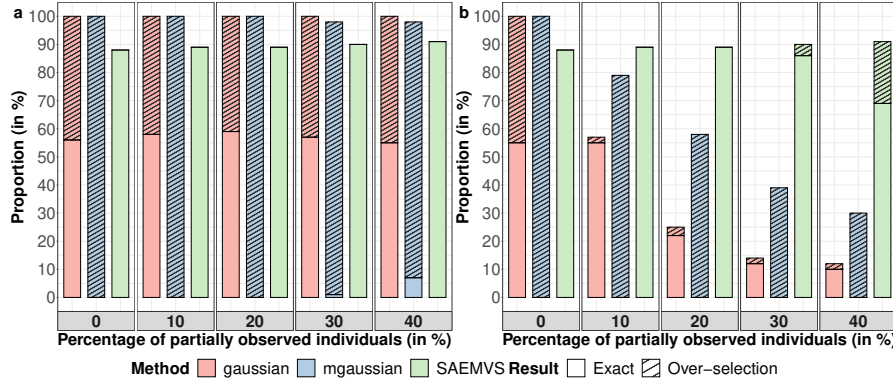

Fig. 1 Proportion of data-sets on which the three methods (in colour) select the correct model ("Exact", unpatterned bars), or a model that strictly includes the correct model ("Over-selection", striped bars) for the first individual parameter (a) and the second individual parameter (b), and different percentage of partially observed individuals (on the x-axis).

17

5.2 Impacts of the different parameters and collinearity between covariates

5.2.1 Model

Now, the performance of SAEMVS is studied under a variety of scenarios. This exploration is carried out on a different nonlinear model, highlighting our approach's flexibility. For ease of presentation, we consider variable selection in a one-dimensional setting, *i.e.* $q = 1$. The other parameters are assumed to be shared among individuals and they are estimated jointly. A data-set is simulated according to a logistic growth model such as:

$$
\begin{cases}
y_{ij} = \dfrac{\psi_1}{1 + \exp\left(-\dfrac{t_{ij} - \varphi_i}{\psi_2}\right)} + \varepsilon_{ij}, \ \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \\
\varphi_i = \mu + \beta^\top V_i + \xi_i, \hspace{2.2cm} \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma^2),
\end{cases}
\tag{17}
$$

where $\psi = (\psi_1, \psi_2)$ is seen as unknown fixed effects, $\varphi_i \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\beta \in \mathbb{R}^p$, and $\Gamma^2 > 0$. Thus, parameter ψ must also be estimated and therefore the population parameter is $\theta = (\mu, \beta, \psi, \Gamma^2, \sigma^2)$. The procedure presented earlier in this paper can easily be extended to this case.

Indeed, as the function g here 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 $\widetilde{Q}_1$ with an exponential form. 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 by considering:

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

with φ and ψ independent, $\Omega = \text{diag}(\omega_1^2, \omega_2^2)$ 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 $\delta = (\delta_\ell)_{1 \leq \ell \leq p}$ (Equation (2) with $q = 1$) are introduced, and we consider the same priors as in (4) for $(\mu, \beta, \sigma^2, \Gamma^2, \delta, \alpha)$ but in

their one-dimensional version:

$$
\begin{aligned}
\pi(\beta|\delta) &= \mathcal{N}_p(0, D_\delta), \text{ with } D_\delta = \text{diag}((\mathbf{1}-\delta)\nu_0 + \delta\nu_1), \, 0 < \nu_0 < \nu_1, \\
\pi(\mu) &= \mathcal{N}(0, \sigma_\mu^2), \text{ with } \sigma_\mu^2 > 0, \\
\pi(\sigma^2) &= \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma \lambda_\sigma}{2}\right), \text{ with } \nu_\sigma, \lambda_\sigma > 0, \\
\pi(\Gamma^2) &= \mathcal{IG}\left(\frac{\nu_\Gamma}{2}, \frac{\nu_\Gamma \lambda_\Gamma}{2}\right), \text{ with } \nu_\Gamma, \lambda_\Gamma > 0, \\
\pi(\delta|\alpha) &= \alpha^{|\delta|}(1-\alpha)^{p-|\delta|}, \text{ with } \alpha \in [0,1] \text{ and } |\delta| = \sum_{\ell=1}^p \delta_\ell, \\
\pi(\alpha) &= Beta(a,b), \text{ with } a, b > 0.
\end{aligned}
\tag{19}
$$

For η, the following prior is chosen: for $r \in \{1,2\}$, $\pi(\eta_r) = \mathcal{N}(0, \rho_r^2)$, with $\rho_r^2 > 0$ 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 adapted to this model. 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. See Appendix A.4 for more details about this extension of SAEMVS.

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 $\Omega^{(0)} = \text{diag}(\omega_1^{2(0)}, \omega_2^{2(0)})$ 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 decreased towards 0, while remaining always strictly positive, to limit the estimation error between the initial model and the extended model.

5.2.2 Simulation design

For this simulation study, individual profiles are simulated according to model (17) by considering $n_1 = \cdots = n_n = 10$ observations per individual and regular observation time points such that $t_{ij} = t_j = 150 + (j-1)\frac{3000-150}{J-1}$, $\sigma^2 = 30$, $\psi_1 = 200$, $\psi_2 = 300$, $\mu = 1200$, $\beta = (100, 50, 20, 0, \ldots, 0)^\top$. Thus, only the first three covariates are assumed to be influential and their respective intensities are contrasted. The individual covariates $V_i \in \mathbb{R}^p$, $1 \le i \le n$, 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 SAEMVS 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 SAEMVS in different "signal-to-noise ratio" 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 $\Sigma = \left(\begin{array}{c|c} I_3 & 0_{3,p-3} \\ \hline 0_{p-3,3} & (\rho_\Sigma^{|i-j|})_{i,j\in\{4,\ldots,p\}} \end{array} \right)$, 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 $\Sigma = \left(\begin{array}{c|c} I_3 & A \\ \hline A^\top & I_{p-3} \end{array} \right)$, with

  $$A = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & (\rho_\Sigma^{|3-j|})_{j\in\{4,\ldots,p\}} & \end{pmatrix}, \ |\rho_\Sigma| < 1.$$

  3. The third scenario considers correlations between the sole influential covariates. Again, an autoregressive correlation structure is used. This corresponds to $\Sigma = \left(\begin{array}{c|c} (\rho_\Sigma^{|i-j|})_{i,j\in\{1,\ldots,3\}} & 0_{3,p-3} \\ \hline 0_{p-3,3} & I_{p-3} \end{array} \right)$, $|\rho_\Sigma| < 1$.

  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 $\Sigma = (\rho_\Sigma^{|i-j|})_{i,j\in\{1,\ldots,p\}}$, $|\rho_\Sigma| < 1$.

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. The algorithmic settings are given in Appendix A.5. Note that, in order to be able to compare covariates that do not have the same order of magnitude, the covariates are standardised. The performances in terms of exact selection of the true influential covariates, over-selection and under-selection are examined (see Subsection 5.2.3).

5.2.3 Results

Scenario with uncorrelated covariates

The results are presented in Table 2 and Figure 2. SAEMVS 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 SAEMVS fails, it is most often because it under-selects, that is, it selects fewer variables than there are. Indeed, average sensitivity values are lower than those for specificity. It seems that SAEMVS tends to avoid false positives, even though this may result in not having selected all the truly influential covariates.

Table 2 Uncorrelated covariates. For each of the quantifiers (Sensitivity, Specificity, Accuracy), the mean over the 100 data sets is shown, with the empirical standard deviation divided by $\sqrt{100}$ in brackets.

Γ^2	p	$n = 100$			$n = 200$		
		Se[a]	Sp[b]	Ac[c]	Se[a]	Sp[b]	Ac[c]
200	500	0.883 (0.0180)	1 (5e-05)	0.999 (0.0001)	0.973 (0.0091)	1 (2e-05)	1 (6e-05)
200	2000	0.900 (0.0160)	1 (2e-05)	1 (3e-05)	0.987 (0.0066)	1 (2e-05)	1 (2e-05)
200	5000	0.870 (0.0189)	1 (6e-06)	1 (1e-05)	0.983 (0.0073)	1 (1e-05)	1 (1e-05)
1000	500	0.860 (0.0185)	1 (3e-05)	0.999 (0.0001)	0.977 (0.0086)	1 (3e-05)	1 (6e-05)
1000	2000	0.840 (0.0180)	1 (9e-06)	1 (3e-05)	0.977 (0.0086)	1 (1e-05)	1 (2e-05)
1000	5000	0.813 (0.0197)	1 (4e-06)	1 (1e-05)	0.963 (0.0105)	1 (6e-06)	1 (8e-06)
2000	500	0.810 (0.0185)	1 (3e-05)	0.999 (0.0001)	0.950 (0.0120)	1 (0)	1 (7e-05)
2000	2000	0.777 (0.0178)	1 (2e-05)	1 (3e-05)	0.937 (0.0131)	1 (1e-05)	1 (2e-05)
2000	5000	0.767 (0.0174)	1 (6e-06)	1 (1e-05)	0.930 (0.0137)	1 (4e-06)	1 (9e-06)

[a]Sensitivity is the proportion of true positives correctly identified.
[b]Specificity is the proportion of the true negatives correctly identified.
[c]Accuracy is the proportion of true results, either true positive or true negative.

Scenarios with correlated covariates

The results for scenarios 1 and 4 with $\Gamma^2 = 200$ are presented in Figure 3 and Table 3 for the two values of ρ_Σ. The results for the other scenarios are given in Appendix B. On
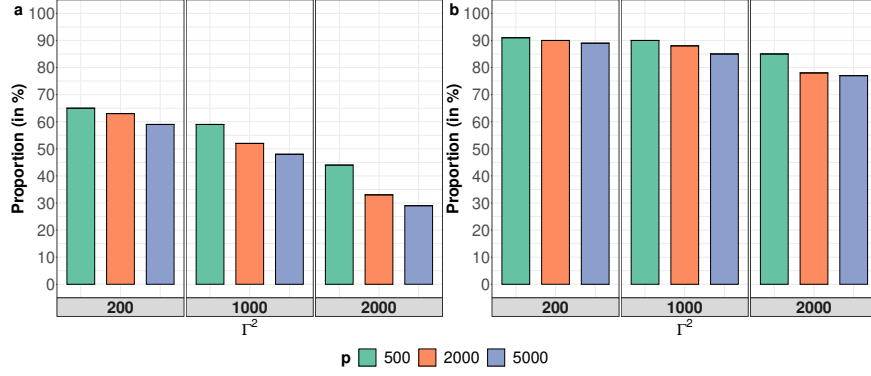
Fig. 2 Uncorrelated covariates. Proportion of data-sets on which SAEMVS selects the correct model for $n = 100$ (a) and $n = 200$ (b), and different values of p and Γ^2.

these figures, one can compare the selection performance of SAEMVS in the different scenarios of correlations between covariates with the case without correlations (i.i.d scenario). First, for scenario 1, that is when the non-active covariates are correlated, quite similar performances to the i.i.d 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, 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, as these active covariates explain the response variable in a similar way, this scenario leads to much under-selection compared to the i.i.d case. This can be seen in the sensitivity values. Moreover, it over-selects more than scenario i.i.d because of the correlations between the true and false covariates. However, by looking at the specificity values, we can see that even in cases where SAEMVS over-selects, the false positive rate (which is equivalent to $(1 - \text{specificity})$) remains very close to 0.
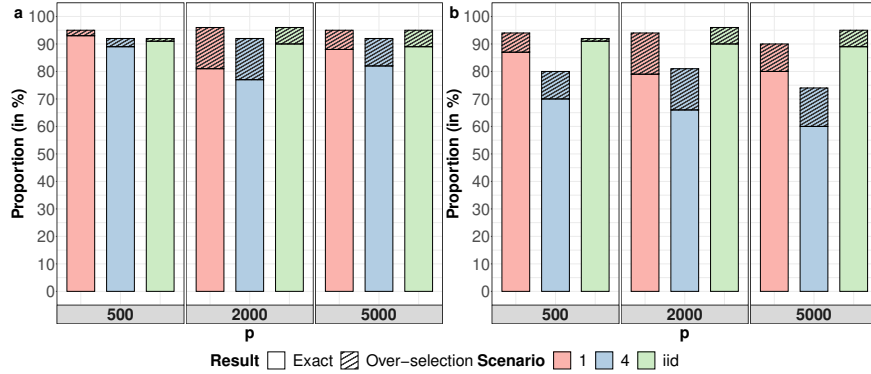

Fig. 3 Correlated covariates for $\Gamma^2 = 200$. Proportion of data-sets on which SAEMVS selects the correct model ("Exact", unpatterned bars) or a model that strictly includes the correct model ("Over-selection", striped bars) for $\rho_\Sigma = 0.3$ (a), $\rho_\Sigma = 0.6$ (b), and different values of p. Scenario "i.i.d" corresponds to the case where the covariates are not correlated and is used as a reference.

22

Table 3 Correlated covariates for $\Gamma^2 = 200$. For each of the quantifiers (Sensitivity, Specificity, Accuracy), the mean over the 100 data sets is shown, with the empirical standard deviation divided by $\sqrt{100}$ in brackets.

Scenario	p	$\rho_\Sigma = 0.3$			$\rho_\Sigma = 0.6$		
		Se[a]	Sp[b]	Ac[c]	Se[a]	Sp[b]	Ac[c]
i.i.d	500	0.973 (0.0091)	1 (2e-05)	1 (6e-05)	0.973 (0.0091)	1 (2e-05)	1 (6e-05)
1	500	0.983 (0.0073)	1 (3e-05)	1 (5e-05)	0.980 (0.0080)	1 (6e-05)	1 (8e-05)
4	500	0.973 (0.0091)	1 (4e-05)	1 (7e-05)	0.933 (0.0134)	1 (0.0001)	0.999 (0.0002)
i.i.d	2000	0.987 (0.0066)	1 (2e-05)	1 (2e-05)	0.987 (0.0066)	1 (2e-05)	1 (2e-05)
1	2000	0.987 (0.0066)	1 (3e-05)	1 (3e-05)	0.980 (0.0080)	1 (3e-05)	1 (3e-05)
4	2000	0.973 (0.0091)	1 (2e-05)	1 (3e-05)	0.937 (0.0131)	1 (3e-05)	1 (4e-05)
i.i.d	5000	0.983 (0.0073)	1 (1e-05)	1 (1e-05)	0.983 (0.0073)	1 (1e-05)	1 (1e-05)
1	5000	0.983 (0.0073)	1 (1e-05)	1 (1e-05)	0.967 (0.0101)	1 (9e-06)	1 (1e-05)
4	5000	0.973 (0.0091)	1 (8e-06)	1 (9e-06)	0.913 (0.0147)	1 (1e-05)	1 (2e-05)

[a]Sensitivity is the proportion of true positives correctly identified.
[b]Specificity is the proportion of the true negatives correctly identified.
[c]Accuracy is the proportion of true results, either true positive or true negative.

5.3 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. To understand precisely the added value of the SAEM algorithm, we 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 a 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, $n_1 = \ldots = n_n = 10$ observations per individual, $p \in \{500, 700, 1000, 1500, 2000, 2500\}$ covariates, $\sigma^2 = 30$, $\psi_1 = 200$, $\psi_2 = 300$, $\mu = 1200$, $\beta = (100, 50, 20, 0, \ldots, 0)^\top$ and $\Gamma^2 = 200$. For $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, 10\}$, $t_{ij} = t_j = 150 + (j-1)\dfrac{3000 - 150}{J - 1}$. 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 $\Theta = (\mu, \beta, \psi, \Gamma^2, \sigma^2, \alpha)$ between the MCMC-SAEM algorithm proposed in this article (Algorithm 1 adapted to model (17), see Appendix A.4) and the full MCMC procedure described above. As explained in Subsection 5.2.1, 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 (19). For η, the prior of Subsection 5.2.1 is chosen: for $r \in \{1, 2\}$, $\pi(\eta_r) = \mathcal{N}(0, \rho_r^2)$, with $\rho_r^2 > 0$ known. To stabilise the MCMC procedure, the prior on σ^2 is modified to a uniform distribution on $[0, 200]$ for both methods. This has very little consequence for SAEMVS. Indeed, the only difference lies in the updating of σ^2 at the M-step of the MCMC-SAEM algorithm, which becomes:

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

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 $\eta^{(0)} = (400, 400)^\top$. 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 SAEMVS, 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: $\nu_0 = 0.04$, $\nu_1 = 12000$, $\sigma_\mu = 3000$, $\nu_\Gamma = \lambda_\Gamma = 1$, $a = 1$, $b = p$, $\Omega = \text{diag}(20, 20)$ and $\rho_1^2 = \rho_2^2 = 1200$.

For ease of presentation, we compare the two approaches for a single value of ν_0. 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. This convergence was assessed qualitatively by looking at the difference between successive parameter estimates during iterations. 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 4. In this figure, computation times of the 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 4 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, SAEMVS 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.
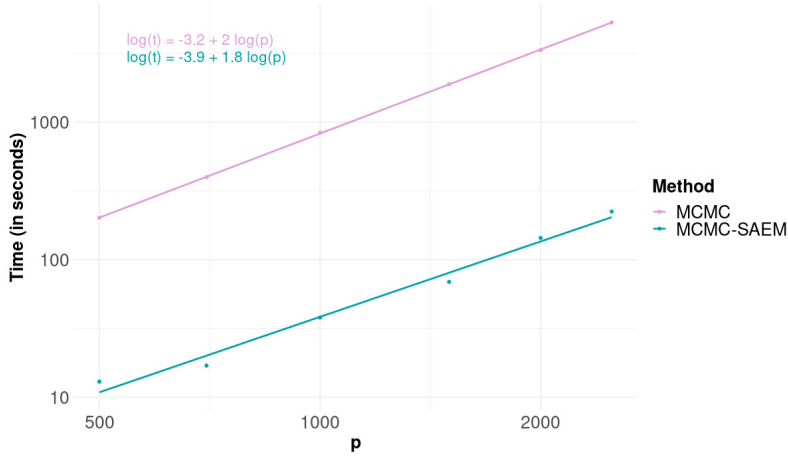

Fig. 4 Comparison of computation times between MCMC (in purple) and Algorithm 1 (MCMC-SAEM, in blue) inference methods in $\log_{10}$-$\log_{10}$ scale.

6 Application to plant senescence genetic marker identification

In this section, SAEMVS is applied to a problem of marker-assisted selection to assist breeding of winter wheat. We are interested in identifying genetic markers that impact the senescence process, *i.e.* the ageing of these plants, which is under genetic determinism. Because heading date is related to senescence, the relevance of the selected senescence markers is then compared to known flowering genes as well as to markers associated to heading data obtained by applying an association mapping model (Yu et al., 2006) to these data. In the following, these markers are called heading QTLs.

Note that the heading date can be seen as an easily measurable approximation of the flowering date.

6.1 Data description and pre-processing

The plant material used here has been previously described in Rincent et al. (2018, 2019) and Touzy et al. (2019). It is composed of $n = 220$ wheat varieties. Senescence was measured as the global proportion of senesced surface of the canopy. This proportion was observed on each variety $J = 18$ times over time.

For each of the n varieties, information from high-throughput genotyping is available on several tens of thousands of SNPs positioned along the entire genome (Rimbert et al., 2018) (see Appendix C.1). These binary variables constitute the covariates to be selected in order to explain the differences between varieties in terms of senescence. There are structurally strong correlation and collinearity between these variables, which hampers variable selection, as pointed out in Section 5.2 and discussed in many references (see e.g. Malsiner-Walli and Wagner (2018) and Heuclin et al. (2020) in spike and slab models). To permit a comparison with the genetic markers associated with flowering in the context of collinearity, SAEMVS is applied chromosome by chromosome. Note, however, that variable selection on the whole genome (≈ 30000 markers) would be computationally feasible in a matter of hours with strict pre-processing to address multicollinearity/near-multicollinearity. We still apply the following minimal pre-processing: if several SNPs in the same chromosome are strictly collinear for all varieties, only one of them is retained for analysis. This eliminates 916 SNPs across the genome. As a result, for a given chromosome C, the number p of covariates in V_i^C is always less than 2000, while remaining in a high-dimensional framework where $p >> n$ (see Table C1 in Appendix C.4 where the values of p are given for each chromosome). The wheat varieties in this data-set are structured into genetic groups, which may cause confusion between the effect of the subpopulation structure and that of SNPs. To control for subpopulation structure, we adapt our model to include covariates not subject to selection. We consider the first 5 principal components of a principal coordinates analysis performed on the available SNPs in the model. These 5 adjustment variables, capturing subpopulation structure, are denoted by $v_i \in \mathbb{R}^5$ for variety i and are guaranteed to be used in the regression.

6.2 Modelling

Some senescence curves are shown in Figure 9 in Appendix C.2. We can see that a logistic growth model (Equation (17)) with a maximum value of 100% and inter-individual variability on the two other parameters is coherent with the shape of these curves. This is a simple model for the purposes of the example, but note that the variable selection method can accommodate other models like beta regression models with minimal changes. Here, we choose to analyse the effects of SNPs from each chromosome C on a single parameter of interest: the variability of characteristic times between varieties. Denoting y_{ij} the proportion of senesced surface of the plant of the variety $1 \leq i \leq n$ at time t_{ij}, this leads to the following model when focussing on

chromosome C:

$$\begin{cases} y_{ij} = \dfrac{100}{1 + \exp\left(-\dfrac{t_{ij} - \varphi_i}{\psi_i}\right)} + \varepsilon_{ij}, \ \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \\ \varphi_i = \mu + \lambda^\top v_i + \beta^\top V_i^C + \xi_i, \qquad \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma^2), \\ \psi_i = \eta + \omega_i, \qquad\qquad\qquad\quad \omega_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Omega^2). \end{cases}$$

The parameter to be estimated is therefore: $\theta = (\mu, \lambda, \beta, \eta, \sigma^2, \Gamma^2, \Omega^2)$. As previously, the indicators δ (Equation (2) with $q = 1$) are introduced and we force the inclusion of variables v_i in the model by a useful reformulation similar to what was done for the intercept in (5). We use the following priors for η and $\Omega^2 : \pi(\eta) = \mathcal{N}(0, \sigma_\eta^2)$, with σ_η^2 known, and $\pi(\Omega^2) = \mathcal{IG}\left(\dfrac{\nu_\Omega}{2}, \dfrac{\nu_\Omega \lambda_\Omega}{2}\right)$, with $\nu_\Omega, \lambda_\Omega > 0$. The same priors as in (19) are used for the other parameters. Here, $\Theta = (\theta, \alpha)$ is the population parameter and $Z = (\varphi, \psi, \delta)$ are the latent variables, where $\varphi = (\varphi_i)_{1 \leq i \leq n}$ and $\psi = (\psi_i)_{1 \leq i \leq n}$. The SAEMVS procedure can be easily implemented with minor modifications to the version detailed in Appendix A.4. The algorithmic settings are provided in Appendix C.3.

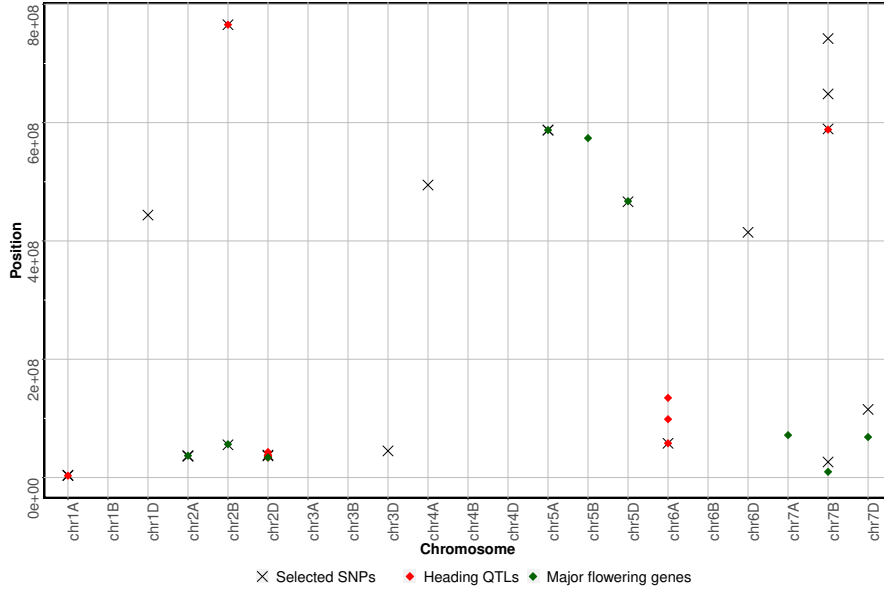## 6.3 Results and discussion

Fig. 5 Position on each chromosome of the markers selected by SAEMVS (in black cross), compared to heading QTLs (in red diamond) and major flowering genes (in green diamond).

27

The results are shown in Figure 5 and further detailed in Table C1 in Appendix C.4. For each chromosome, the SNPs selected by SAEMVS are compared to heading QTLs and major flowering genes that were identified in previous analysis (called Ppd and Vrn genes). Indeed, as mentioned above, as flowering and senescence are linked biological processes, we can expect to find SNPs that are close to these specific QTLs and genes on the genome. As an example, SAEMVS applied to chromosome 1A leads to the selection of two SNPs that are close to each other (*i.e.* within one mega-base of each other) but also close to a heading QTL on the chromosome. On a genome-wide basis, Figure 5 displays many colocalisations between flowering genes or heading QTLs and the SNPs selected by SAEMVS on the senescence data. Our procedure thus returns consistent results from a biological point of view. Conversely, SAEMVS also selects SNPs that are not close to zones associated with flowering on the genome, for example on chromosome 1D, 3D, 4A, 6D, 7B and 7D. The selected markers reveal potentially "stay-green" SNPs associated with late senescence independently of flowering time, which would be very interesting for plant breeders.

It is important to note that SAEMVS suffers from selection switch among markers that are highly correlated. This is illustrated by the chevron structure in Figure 6. On the left-hand plot, the red curves correspond to the selection threshold and the markers that are selected at least twice along the grid Δ of ν_0 values are denoted in different colours. In such situations, we expect any variable selection to face the same challenges, as one cannot hope to select markers truly associated with the phenomenon of interest. The usual approach in biology is to rather identify regions of the genome containing markers associated to the phenomenon of interest, usually via a post-processing of the variable selection results.
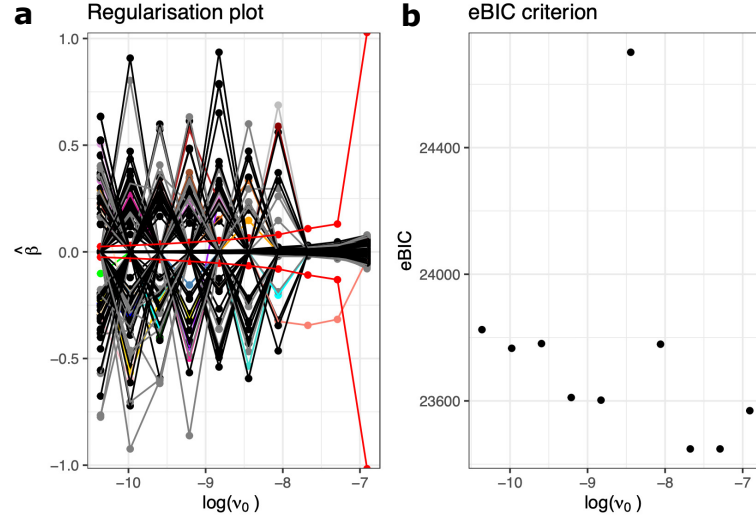

Fig. 6 Regularisation plot and eBIC criterion for chromosome 6A

7 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 was approached from a Bayesian perspective and a selection procedure combining the use of spike-and-slab Gaussian mixture prior and the SAEM algorithm was proposed. The spike-and-slab prior on the regression coefficients 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 exploring different levels of sparsity in the model through the variance of the spike distribution ν_0, which we observed to be beneficial in selecting sparse models for regression. Varying the level of sparsity provides a collection of good models among which we select the one minimising an eBIC criterion.

The SAEMVS method can do both one-dimensional and multidimensional variable selection, and is does not assume the same support for each dimension. It is very flexible and was illustrated on three different models. The proposed methodology showed very good selection performance on simulated data. Indeed, SAEMVS 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. The application of SAEMVS on a real data-set shows, on the one hand, the flexibility of the procedure, and on the other hand, convincing results from a biological point of view despite strong correlations/multicollinearity between the covariates.

Moreover, it was observed that a reasonable correlation between covariates has little effect on the selection performance 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, the i.i.d. Bernoulli prior on the indicators $\boldsymbol{\delta}$ (4e), entails the assumption that each covariate has the same probability a priori of being included in the model. However, there are situations, such as genomic data, in which 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 SAEMVS by choosing a more flexible prior on $\boldsymbol{\delta}$. 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 also be considered in our methodology.

Another important remark is that, in this article, we considered a Gaussian distribution for $p(\boldsymbol{y}|\boldsymbol{\varphi}, \sigma^2)$ 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.

Declarations

Funding

Conflict of interest

The authors declare that they have no conflict of interest.

Data and code availability

Code for running the simulations and generating exact copies of all the figures in this paper, and the real data-set are available at:
https://github.com/Marion-Naveau/Supp_Information_SAEMVS

References

Akaike, H. 1998. Information theory and an extension of the maximum likelihood principle, *Selected papers of hirotugu akaike*, 199–213. New York: Springer.

Allassonnière, S. and V. Debavelaere. 2021. On the curved exponential family in the Stochatic Approximation Expectation Maximization Algorithm. *ESAIM: Probability and Statistics* 25: 408–432 .

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

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

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

Charmet, G., L.G. Tran, J. Auzanneau, R. Rincent, and S. Bouchet. 2020. BWGS: a R package for genomic selection and its application to a wheat breeding programme. *PLoS One 15*(4): e0222733 .

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

de Valpine, P., D. Turek, C. Paciorek, C. Anderson-Bergman, D. Temple Lang, and R. Bodik. 2017. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* 26: 403–417. https://doi.org/10.1080/10618600.2016.1172487 .

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

Delyon, B., M. Lavielle, and E. Moulines. 1999. Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics* 27(1): 94–128 .

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1): 1–22 .

Deshpande, S.K., V. Ročková, and E.I. George. 2019. Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics* 28(4): 921–931 .

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

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1): 1 .

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

George, E.I. and R.E. McCulloch. 1997. Approaches for Bayesian variable selection. *Statistica sinica* 7(2): 339–373 .

Hastie, T., R. Tibshirani, J.H. Friedman, and J.H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, Volume 2. Stanford: Springer.

Heuclin, B., F. Mortier, C. Trottier, and M. Denis. 2020. Bayesian varying coefficient model with selection: An application to functional mapping. *Journal of the Royal Statistical Society: Series C Applied Statistics* 70(1): 24–50 .

International Wheat Genome Sequencing Consortium, . 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361(6403): eaar7191 .

Kuhn, E. and M. Lavielle. 2004. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics* 8: 115–131 .

Kuhn, E. and M. Lavielle. 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.* New York: CRC press.

Lee, S.Y. 2022. Bayesian Nonlinear Models for Repeated Measurement Data: An Overview, Implementation, and Applications. *Mathematics 10*(6): 898 .

Liquet, B., K. Mengersen, A.N. Pettitt, and M. Sutton. 2017. Bayesian variable selection regression of multivariate responses for group data. *Bayesian Analysis 12*(4): 1039–1067 .

Malsiner-Walli, G. and H. Wagner. 2018. Comparing spike and slab priors for Bayesian variable selection. *Austrian Journal of Statistics 40*(4): 241–264 .

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

Ollier, E. 2022. Fast selection of nonlinear mixed effect models using penalized likelihood. *Computational Statistics & Data Analysis* 167: 107373 .

Pinheiro, J.C. and D.M. Bates. 2000. *Mixed-effects Models in S and S-PLUS.* New York: Springer.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. De Bakker, M.J. Daly, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics 81*(3): 559–575 .

Rimbert, H., B. Darrier, J. Navarro, J. Kitt, F. Choulet, M. Leveugle, J. Duarte, N. Rivière, K. Eversole, I.W.G.S. Consortium, et al. 2018. High throughput SNP discovery and genotyping in hexaploid wheat. *PloS one 13*(1): e0186329 .

Rincent, R., J.P. Charpentier, P. Faivre-Rampant, E. Paux, J. Le Gouis, C. Bastien, and V. Segura. 2018. Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. *G3: Genes, Genomes, Genetics 8*(12): 3961–3972 .

Rincent, R., M. Malosetti, B. Ababaei, G. Touzy, A. Mini, M. Bogard, P. Martre, J. Le Gouis, and F. Van Eeuwijk. 2019. Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions. *Theoretical and Applied Genetics 132*(12): 3399–3411 .

Ročková, V. and E.I. George. 2014. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association 109*(506): 828–846 .

Schelldorfer, J., P. Bühlmann, and S.v. de Geer. 2011. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics 38* (2): 197–214 .

Schnabel, R.B., J.E. Koonatz, and B.E. Weiss. 1985. A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software (TOMS) 11* (4): 419–440 .

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics 6* (2): 461–464 .

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

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. Van Der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series B (statistical methodology) 64* (4): 583–639 .

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

Stingo, F.C. and M. Vannucci. 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 M. Vannucci. 2021. *Handbook of Bayesian variable selection.* Boca Raton: 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 .

Touzy, G., R. Rincent, M. Bogard, S. Lafarge, P. Dubreuil, A. Mini, J.C. Deswarte, K. Beauchêne, J. Le Gouis, and S. Praud. 2019. Using environmental clustering to identify specific drought tolerance QTLs in bread wheat (t. aestivum l.). *Theoretical and Applied Genetics 132* (10): 2859–2880 .

Wei, G.C. and M.A. Tanner. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association 85* (411): 699–704 .

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. Mc-Mullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics 38* (2): 203–208 .

Appendix A   Algorithms: synthesis and implementation details

A.1   MCMC-SAEM algorithm in SSNLME model

First, in Subsection 3.2, we notice that $\widetilde{Q}_1(\boldsymbol{y}, \boldsymbol{\varphi}, \theta, \Theta^{(k)})$ takes an exponential form. More precisely, we have:

$$\widetilde{Q}_1(\boldsymbol{y}, \boldsymbol{\varphi}, \theta, \Theta^{(k)}) = -\psi(\theta, \Theta^{(k)}) + \left\langle S(\boldsymbol{y}, \boldsymbol{\varphi}), \phi(\theta) \right\rangle, \tag{A1}$$

with:

- $S(\boldsymbol{y}, \boldsymbol{\varphi}) = \left(\sum_{i,j} (y_{ij} - g(\varphi_i, t_{ij}))^2 \,,\, vec(\boldsymbol{\varphi}^\top \boldsymbol{\varphi}) \,,\, vec(\boldsymbol{\varphi}) \right)$

- $\phi(\theta) = \left(-\dfrac{1}{2\sigma^2} \,,\, -\dfrac{1}{2} vec(\Gamma^{-1}) \,,\, vec(\tilde{V}\tilde{\boldsymbol{\beta}}\Gamma^{-1}) \right)$

- $\psi(\theta, \Theta^{(k)}) \;=\; \dfrac{1}{2} \langle (\tilde{V}\tilde{\boldsymbol{\beta}})^\top \tilde{V}\tilde{\boldsymbol{\beta}}, \Gamma^{-1} \rangle \;+\; \dfrac{1}{2} \sum_{m=1}^{q} \sum_{\ell'=1}^{p+1} \tilde{\beta}_{\ell' m}^2 \widetilde{d}_{\ell' m}^{\,*}(\Theta^{(k)}) \;+\;$
  $\dfrac{n_{tot} + \nu_\sigma + 2}{2} \log(\sigma^2) + \dfrac{n + d + q + 1}{2} \log(|\Gamma|) + \dfrac{\nu_\sigma \lambda_\sigma}{2\sigma^2} + \dfrac{1}{2} \mathrm{Tr}(\Sigma_\Gamma \Gamma^{-1})$

where $vec(A)$ denotes the vectorisation of a matrix A. To simplify the formulas and using that for two matrix A and B, $\langle A, B \rangle := \mathrm{Tr}(A^\top B) = \langle vec(A), vec(B) \rangle$, we denote by

$$(s_1(\boldsymbol{y}, \boldsymbol{\varphi}), s_2(\boldsymbol{\varphi}), s_3(\boldsymbol{\varphi})) = \left(\sum_{i,j} (y_{ij} - g(\varphi_i, t_{ij}))^2 \,,\, \boldsymbol{\varphi}^\top \boldsymbol{\varphi} \,,\, \boldsymbol{\varphi} \right). \tag{A2}$$

The use of the decomposition discussed in Subsection 3.2 leads to the following extension of the MCMC-SAEM algorithm, Algorithm 1, for computing the MAP estimator of Θ in the SSNLME model (1) - (4), where Λ_θ denotes the parameter space restricted to θ, h is small (between 1 and 5), and K is usually in the order of a few hundred.

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 $0 \le k \le n_{\text{burnin}} - 1$ and $\gamma_k = 1/(k - n_{\text{burnin}} + 1)^\gamma$ for $n_{\text{burnin}} \le k \le K - 1$, 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). Moreover, to avoid convergence toward a local maximum in the MCMC-SAEM algorithm, a simulated annealing version of SAEM (see Lavielle, 2014) is implemented.

Algorithm 1 MCMC-SAEM

Input: $K \in \mathbb{N}^*$, $\Theta^{(0)}$ initial parameter, hyperparameters vector $\Xi = (\nu_0, \nu_1, \sigma_\mu^2, \nu_\sigma, \lambda_\sigma, \Sigma_\Gamma, d, a, b)$, $S_0 = 0$ 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 $\boldsymbol{\varphi}^{(k)}$ using the result of h iterations of an MCMC procedure with $\pi(\boldsymbol{\varphi}|\boldsymbol{y}, \Theta^{(k)})$ for target distribution.

2. **SA-Step:** for $u \in \{1, 2, 3\}$, compute $s_{u,k+1} = s_{u,k} + \gamma_k(s_u(\boldsymbol{y}, \boldsymbol{\varphi}^{(k)}) - s_{u,k})$ with $s_u(\boldsymbol{y}, \boldsymbol{\varphi}^{(k)})$ defined by (A2).

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

   $\alpha^{(k+1)} = \underset{\alpha \in [0,1]^q}{\operatorname{argmax}} \widetilde{Q}_2(\alpha, \Theta^{(k)})$, which reduces to the following explicit forms in model (1)-(4):

   • $vec(\tilde{\boldsymbol{\beta}}^{(k+1)}) = \left(I_q \otimes \tilde{V}^\top \tilde{V} + (\Gamma^{(k)} \otimes I_{p+1})\operatorname{diag}\left(vec(\tilde{D}^*) \right) \right)^{-1} vec(\tilde{V}^\top s_{3,k+1})$

   • $\Gamma^{(k+1)} = \dfrac{\Sigma_\Gamma + s_{2,k+1} - (\tilde{V}\tilde{\boldsymbol{\beta}}^{(k+1)})^\top s_{3,k+1} - s_{3,k+1}^\top \tilde{V}\tilde{\boldsymbol{\beta}}^{(k+1)} + (\tilde{V}\tilde{\boldsymbol{\beta}}^{(k+1)})^\top \tilde{V}\tilde{\boldsymbol{\beta}}^{(k+1)}}{n + d + q + 1}$

   • $\sigma^{2(k+1)} = \dfrac{\nu_\sigma \lambda_\sigma + s_{1,k+1}}{n_{tot} + \nu_\sigma + 2}$

   • $\alpha_m^{(k+1)} = \dfrac{\sum_{\ell=1}^p p_{\ell m}^*(\Theta^{(k)}) + a_m - 1}{p + b_m + a_m - 2}$ for $1 \le m \le q$

   where ψ and ϕ are defined by (A1), and $\widetilde{Q}_2(\alpha, \Theta^{(k)})$, $(p_{\ell m}^*(\Theta^{(k)}))_{1 \le \ell \le p;\, 1 \le m \le q}$ and $(\widetilde{d}_{\ell' m}^*(\Theta^{(k)}))_{1 \le \ell' \le p+1;\, 1 \le m \le q}$ are defined in Proposition 1, with $\tilde{D}^* = (\widetilde{d}_{\ell' m}^*(\Theta^{(k)}))_{1 \le \ell' \le p+1;\, 1 \le m \le q} \in \mathcal{M}_{(p+1) \times q}$.

end for

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

A.2 Proposed variable selection procedure: SAEMVS

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

Algorithm 2 SAEMVS procedure

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 $\widehat{\boldsymbol{\beta}}_{\nu_0}^{MAP}$ to define sub-model $\widehat{S}_{\nu_0}$ according to Equation (12).

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 $\widehat{S}_{\nu_0}$ with an MCMC-SAEM algorithm.
2. Compute the log-likelihood $\log p(\boldsymbol{y}; \ \hat{\theta}_{\nu_0}^{MLE})$ with importance sampling techniques.
3. Compute the associated eBIC($\widehat{S}_{\nu_0}$) according to Equation (15).

end for

▷ **Identify the best level of sparsity:** compute $\hat{\nu}_0$ defined by Equation (13).

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

A.3 Algorithmic settings of SAEMVS for the comparison study

For the comparison study, the following settings are used for Algorithm 2.

- The hyperparameter values are set to $\nu_\sigma = \lambda_\sigma = 1$, $d = 4$, $\Sigma_\Gamma = 0.2I_2$, $a = (1,1)^\top$, $b = (p,p)^\top$, $\sigma_\mu = 5$, $\nu_1 = 1000$, and the spike parameter ν_0 runs through a grid Δ defined as $\log_{10}(\Delta) = \left\{ -3 + k \times \dfrac{1}{3}, k \in \{0, \dots, 9\} \right\}$.
- The step sizes are defined with $\gamma = 2/3$, $n_{\text{burnin}} = 150$ and $K = 300$ as explained in Appendix A.1.
- The MCMC-SAEM algorithm is initialised with: $\forall m \in \{1,2\}, \forall \ell \in \{1, \dots, 10\}$ $\boldsymbol{\beta}_{\ell m}^{(0)} = 1$, and $\forall m \in \{1,2\}, \forall \ell \in \{11, \dots, p\}$ $\boldsymbol{\beta}_{\ell m}^{(0)} = 0.1$, $\mu^{(0)} = (10,10)^\top$, $\sigma^{2(0)} = 10^{-2}$, $\Gamma^{(0)} = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}$ and $\alpha^{(0)} = 0.5$. Note that different initialisations have been tested and have shown similar performances.

Figure 7 represents the convergence graphs of one run of the MCMC-SAEM algorithm for μ, some components of $\boldsymbol{\beta}$, σ^2, Γ and α. It is observed that the algorithm

converges in a few iterations for any parameter. Note that the parameters are all relatively correctly estimated, except for Γ but this was expected because of a over-fitting situation. Indeed, the underestimation of Γ can be explained by the fact that since $\nu_0 > 0$, none of the estimates of the coefficients of $\boldsymbol{\beta}$ is zero and therefore all the covariates are active in the model, which makes the variance estimation of the random effect tend towards 0.

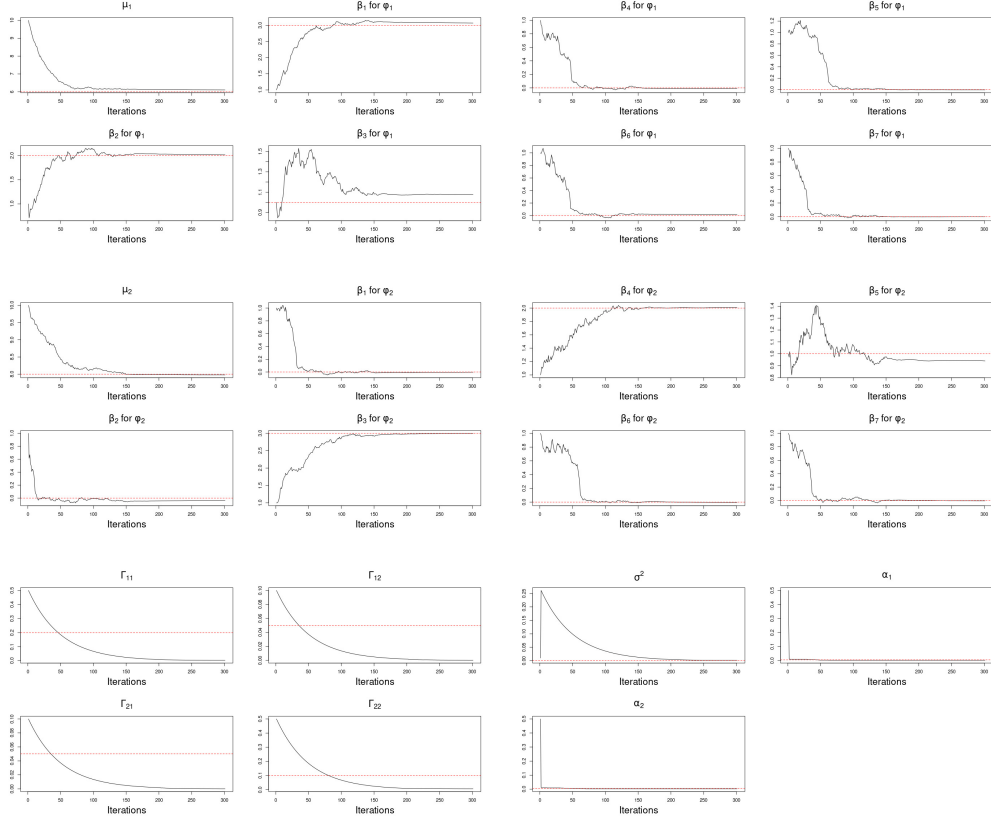**Fig. 7** Convergence graphs of the MCMC-SAEM algorithm for μ, some components of $\boldsymbol{\beta}$, σ^2, Γ and α on one simulated data-set, for $\nu_0 = 0.005$ and $\nu_1 = 1000$. The red dashed line corresponds to the true value of the considered parameter.

A.4 Extension in one-dimensional setting with estimation of fixed effects

As it is explained in Subsection 5.2.1, SAEMVS can easily be adapted to a model where fixed effects must be estimated. For this, the following general one-dimensional

model is considered:
$$\begin{cases} y_{ij} \overset{\text{ind.}}{\sim} \mathcal{N}(g(\varphi_i, \psi, t_{ij}), \sigma^2), \\ \varphi_i \overset{\text{ind.}}{\sim} \mathcal{N}(\mu + \beta^\top V_i, \Gamma^2), \end{cases}$$
where $\psi \in \mathbb{R}^s$ are fixed effects to be estimated. Thus, 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 (A1) for $\widetilde{Q}_1$. The priors described in Subsection 5.2.1 in the particular case of the logistic growth model are used here, and we have $\Theta = (\mu, \beta, \eta, \sigma^2, \Gamma^2, \alpha)$ and $Z = (\varphi, \psi, \delta)$. The calculation of the quantity Q of the EM algorithm becomes:

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

where:

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

with

$$\widetilde{Q}_1(\boldsymbol{y}, \varphi, \psi, \theta, \Theta^{(k)}) = -\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - g(\varphi_i, \psi, t_{ij}))^2 - \frac{1}{2\Gamma^2}||\varphi - \tilde{V}\tilde{\beta}||^2 -$$
$$\frac{1}{2}\sum_{\ell'=1}^{p+1}\tilde{\beta}_{\ell'}^2\widetilde{d}_{\ell'}^*(\Theta^{(k)}) - \frac{n_{tot} + \nu_\sigma + 2}{2}\log(\sigma^2) - \frac{n + \nu_\Gamma + 2}{2}\log(\Gamma^2) -$$
$$\frac{\nu_\Gamma\lambda_\Gamma}{2\Gamma^2} - \frac{\nu_\sigma\lambda_\sigma}{2\sigma^2} - \sum_{r=1}^s\frac{(\psi_r - \eta_r)^2}{2\omega_r^2} - \sum_{r=1}^s\frac{\eta_r^2}{2\rho_r^2}$$

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 $(\widetilde{d}_{\ell'}^*(\Theta^{(k)}))_{1\le\ell'\le p+1}$ are defined in Proposition 1, Equations (10) and (11) with $q = 1$.

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

$$\widetilde{Q}_1(\boldsymbol{y}, \varphi, \psi, \theta, \Theta^{(k)}) = -\Psi(\theta, \Theta^{(k)}) + \left\langle S(\boldsymbol{y}, \varphi, \psi), \phi(\theta)\right\rangle \qquad \text{(A3)}$$

with:

- $S(\boldsymbol{y}, \varphi, \psi) = \left(\sum_{i,j} (y_{ij} - g(\varphi_i, \psi, t_{ij}))^2 \ , \ \sum_{i=1}^n \varphi_i^2 \ , \ \varphi \ , \ \psi^2 \ , \ \psi \right)$

- $\phi(\theta) = \left(-\dfrac{1}{2\sigma^2} \ , \ -\dfrac{1}{2\Gamma^2} \ , \ \dfrac{\tilde{V}\tilde{\beta}}{\Gamma^2}, \left(-\dfrac{1}{2\omega_r^2} \right)_{1 \leq r \leq s} , \left(\dfrac{\eta_r}{\omega_r^2} \right)_{1 \leq r \leq s} \right)$

- $\Psi(\theta, \Theta^{(k)}) = \dfrac{||\tilde{V}\tilde{\beta}||^2}{2\Gamma^2} + \dfrac{1}{2} \sum_{\ell'=1}^{p+1} \tilde{\beta}_{\ell'}^2 \widetilde{d_{\ell'}^*}(\Theta^{(k)}) + \dfrac{n_{tot} + \nu_\sigma + 2}{2} \log(\sigma^2) +$

  $\dfrac{n + \nu_\Gamma + 2}{2} \log(\Gamma^2) + \dfrac{\nu_\Gamma \lambda_\Gamma}{2\Gamma^2} + \dfrac{\nu_\sigma \lambda_\sigma}{2\sigma^2} + \sum_{r=1}^s \dfrac{\eta_r^2}{2\omega_r^2} + \sum_{r=1}^s \dfrac{\eta_r^2}{2\rho_r^2}$

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 $\pi(\varphi, \psi | \boldsymbol{y}, \Theta^{(k)})$ for target distribution.

2. **SA-Step:** compute $S_{k+1} = S_k + \gamma_k(S(\boldsymbol{y}, \varphi^{(k)}, \psi^{(k)}) - S_k)$ with $S(\boldsymbol{y}, \varphi, \psi)$ defined by (A3), where $S_k = (s_{1,k}, s_{2,k}, s_{3,k}, s_{4,k}, s_{5,k}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^q$.

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

   $\alpha^{(k+1)} = \underset{\alpha \in [0,1]}{\operatorname{argmax}} \ \widetilde{Q}_2(\alpha, \Theta^{(k)})$. More precisely,

   - $\tilde{\beta}^{(k+1)} = (\tilde{V}^\top \tilde{V} + \Gamma^{2^{(k)}} \operatorname{diag}((\widetilde{d_{\ell'}^*}(\Theta^{(k)}))_{1 \leq \ell' \leq p+1}))^{-1} \tilde{V}^\top s_{3,k+1},$

   - $\Gamma^{2^{(k+1)}} = \dfrac{||\tilde{V}\tilde{\beta}^{(k+1)}||^2 + \nu_\Gamma \lambda_\Gamma + s_{2,k+1} - 2\langle s_{3,k+1}, \tilde{V}\tilde{\beta}^{(k+1)} \rangle}{n + \nu_\Gamma + 2},$

   - $\sigma^{2^{(k+1)}} = \dfrac{\nu_\sigma \lambda_\sigma + s_{1,k+1}}{n_{tot} + \nu_\sigma + 2},$

   - $\eta_r^{(k+1)} = \dfrac{(s_{5,k+1})_r}{1 + \dfrac{\omega_r^2}{\rho_r^2}}$ for $1 \leq r \leq s,$

   - $\alpha^{(k+1)} = \dfrac{\sum_{\ell=1}^p p_\ell^*(\Theta^{(k)}) + a - 1}{p + b + a - 2}.$

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 $\hat{\theta}_{\nu_0}^{MAP}$, where $\hat{\theta}_{\nu_0}^{MAP} = (\hat{\mu}_{\nu_0}^{MAP}, \hat{\beta}_{\nu_0}^{MAP}, \hat{\eta}_{\nu_0}^{MAP}, \hat{\Gamma}_{\nu_0}^{2,MAP}, \hat{\sigma}_{\nu_0}^{2,MAP})$, and the estimation of η is used as an estimation of ψ. Then, to finish the model collection reduction step of SAEMVS, Algorithm 2, the estimator $\hat{\beta}_{\nu_0}^{MAP}$ is thresholded to obtain a promising sub-model $\widehat{S}_{\nu_0}$ given by Equation (12). 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,

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

where $p(\boldsymbol{y};\ \theta)$ 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 $p(\varphi_i;\,\hat{\theta}_{\nu_0}^{MLE}) = \mathcal{N}(\,\hat{\mu}_{\nu_0}^{MLE} + (\hat{\beta}_{\nu_0}^{MLE})^{\top}V_i,\hat{\Gamma}_{\nu_0}^{2\,MLE})$.

A.5 Algorithmic settings in the simulation study

For the simulation study, 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 Appendix A.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 $\eta^{(0)} = (400, 400)^{\top}$. 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 $\tau = 0.9$.

Appendix B Results for scenarios with correlated covariates

In this appendix, we give the results for scenarios not discussed in Section 5.2.3.

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 i.i.d scenario are observed. Indeed, the selection performances of SAEMVS 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.
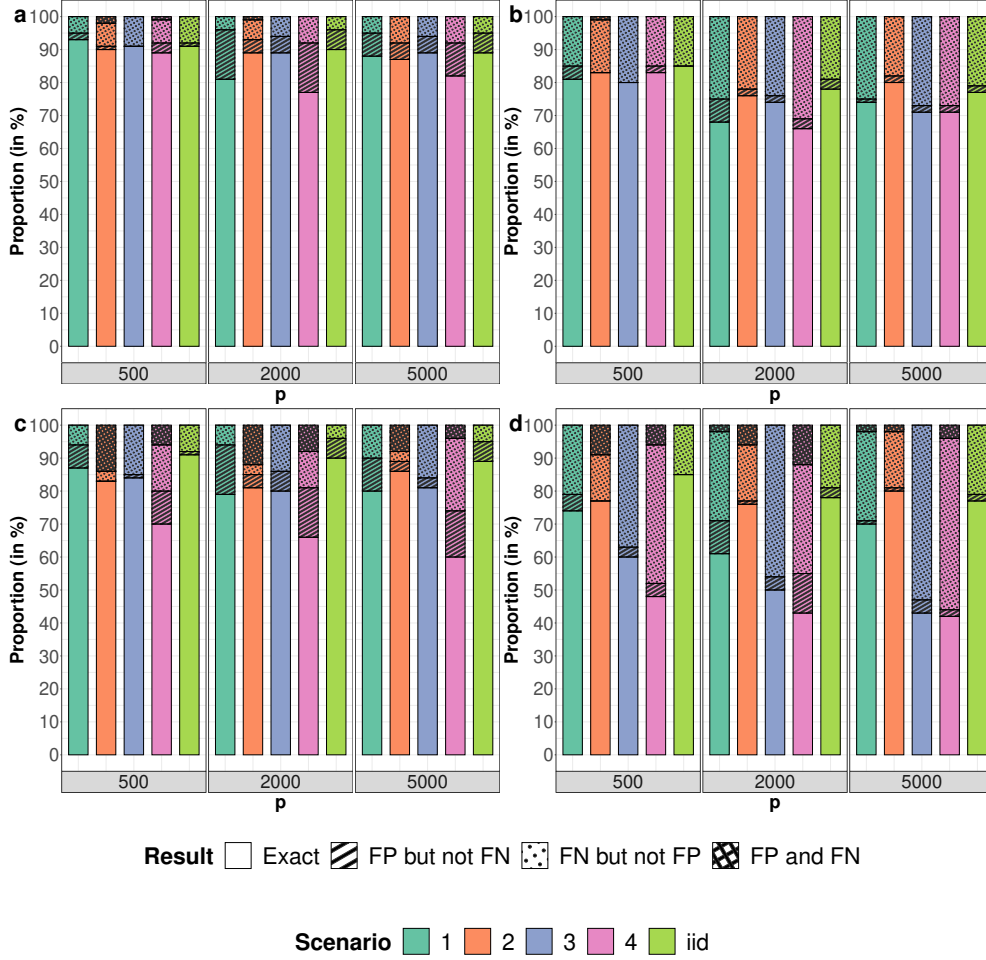
Fig. 8 Correlated covariates. Proportion of data-sets on which Algorithm 2 selects the correct model ("Exact", unpatterned bars), a model that contains false positives but not false negatives ("FP but not FN", striped bars), a model that contains false negatives but not false positives ("FN but not FP", dotted bars), or a model that contains both false positives and false negatives ("FP and FN", crosshatched bars) for $\rho_\Sigma = 0.3$ and $\Gamma^2 = 200$ (a), $\rho_\Sigma = 0.3$ and $\Gamma^2 = 2000$ (b), $\rho_\Sigma = 0.6$ and $\Gamma^2 = 200$ (c), and $\rho_\Sigma = 0.6$ and $\Gamma^2 = 2000$ (d), and different values of p. Scenario "i.i.d" corresponds to the case where the covariates are not correlated and is used as a reference.

Appendix C   Further details on real data

C.1   More details on the variety genotyping process

The varieties of the data-set used in this application was genotyped with the TaBW280K high-throughput genotyping array (Rimbert et al., 2018). This array was designed to cover both genic and intergenic regions of the three bread wheat subgenomes. Markers in strong Linkage Disequilibrium (LD) were filtered out using

the pruning function of PLINK (Purcell et al., 2007) with a window of size 100 SNPs (Single Nucleotide Polymorphism, also called "molecular markers" in the following), a step of 5 SNPs and a LD threshold of 0.8, as proposed in Charmet et al. (2020). Missing values were imputed as the marker observed frequency, and then these imputed values are replaced by 0 or 1 using a threshold of 0.5. Monomorphic and unmapped markers were removed from the data-set. Eventually, we obtained $p = 26{,}189$ polymorphic high-resolution SNPs with a physical position on the v1 reference genome (International Wheat Genome Sequencing Consortium, 2018).

C.2 Representation of the data-set

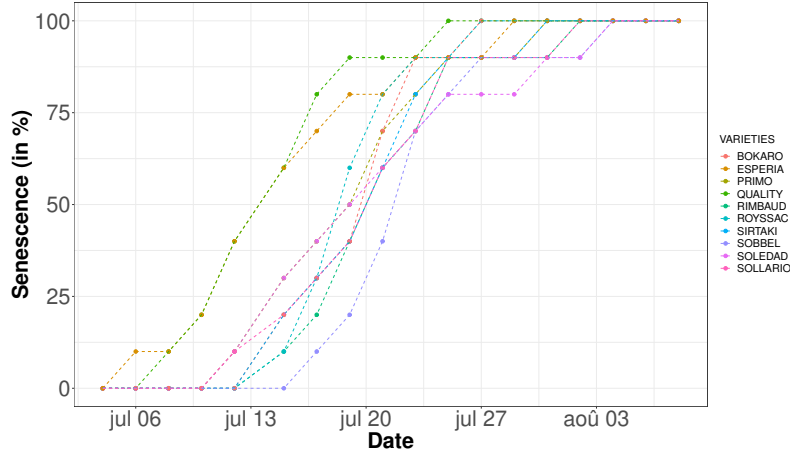Figure 9 shows part of the real data-set.

Fig. 9 Representation of the data-set for 11 different varieties

C.3 Settings

For application on real data, the following settings are used for Algorithm 2.

- The hyperparameter values are set to $\nu_\sigma = \lambda_\sigma = \nu_\Gamma = \lambda_\Gamma = \nu_\Omega = \lambda_\Omega = 1$, $a = 1$, $b = p$, $\sigma_\mu^2 = \sigma_\lambda^2 = \sigma_\eta^2 = 100$, $\nu_1 = 10$, and the spike parameter ν_0 runs through a grid Δ defined as $\log_{10}(\Delta) = \left\{ -4.5 + k \times \dfrac{1.5}{9}, k \in \{0, \ldots, 9\} \right\}$.
- The step sizes are defined with $\gamma = 2/3$, $n_{\text{burnin}} = 250$ and $K = 400$ as explained in Appendix A.1.
- The MCMC-SAEM algorithm is initialised with: $(\beta^{(0)})_\ell = 0.25$ if ℓ is a marker less than 1 mega base distance from a heading QTL or a major flowering gene, and otherwise $(\beta^{(0)})_\ell = 0.1$, $\mu^{(0)} = 20$, $\lambda^{(0)} = (0.25, \ldots, 0.25)^\top$, $\sigma^{2^{(0)}} = 80$, $\Gamma^{2^{(0)}} = 50$, $\eta^{(0)} = 5$, $\Omega^{2^{(0)}} = 50$, and $\alpha^{(0)} = 0.5$.

C.4 Table of results of the application on real data

Table C1 summarises for each chromosome, the number of covariates, the number of heading QTLs, the number of major flowering genes present in that chromosome, and the number of SNPs selected by SAEMVS.

Table C1 Summary of data and number of SNPs selected by SAEMVS for each chromosome

Chromosome	p	Number of heading QTLs	Number of flowering genes	Number of selected SNPs
1A	1473	1	0	2
1B	1604	0	0	0
1D	497	0	0	1
2A	1416	0	1	3
2B	1672	1	1	2
2D	696	7	1	2
3A	1477	0	0	0
3B	1888	0	0	0
3D	722	0	0	1
4A	1259	0	0	1
4B	961	0	0	0
4D	571	0	0	0
5A	1598	0	1	2
5B	1535	0	1	0
5D	772	0	1	1
6A	1119	3	0	1
6B	1317	0	0	0
6D	584	0	0	1
7A	1819	0	1	0
7B	1515	1	1	4
7D	778	0	1	1