Estimation and variable selection in high dimension in nonlinear mixed-effects models.

Antoine Caillebotte, Estelle Kuhn, Sarah Lemler

Estimation and variable selection in high dimension in nonlinear mixed-effects models.

Antoine Caillebotte [1,2] & Estelle Kuhn [2] & Sarah Lemler [3]

[1] *Université Paris-Saclay, INRAE, UMR GQE-Moulon, France, caillebotte.antoine@inrae.fr,*
[2] *Université Paris-Saclay, INRAE, UR MaIAGE, France, estelle.kuhn@inrae.fr,*
[3] *Université Paris-Saclay, CentraleSupélec, Laboratoire MICS, France, sarah.lemler@centralesupelec.fr*

Abstract. We consider nonlinear mixed effects models including high-dimensional covariates to model individual parameters variability. The objective is to identify relevant covariates among a large set and to estimate model parameters. To face the high dimensional setting we consider a regularized estimator namely a penalized LASSO-type estimator. We rely on the use of the eBIC model choice criterion to select an optimal reduced model. Then we estimate the parameters by maximum likelihood in the reduced model. We calculate the LASSO-type penalized estimator by a weighted proximal gradient descent algorithm with an adaptive learning rate. This choice allows us in particular to consider models that do not belong to the curved exponential family. We compare first the performance of the proposed methodology with those of the glmmLasso procedure in a linear mixed effects model in a simulation study. We then illustrate its performance in a nonlinear mixed-effects logistic growth model through simulation.

Keywords. nonlinear mixed effects model, high dimension, variable selection, LASSO penalty, stochastic gradient descent, weighted proximal.

1 Introduction

Mixed effects models are very fine and very useful statistical modeling tools for analyzing data with hierarchical structures and repeated measurements (see Pinheiro and Bates [2006]). In particular it is possible to account for several levels of variability of a phenomenon observed within a population of individuals. They are widely used in many applied fields such as agronomy, pharmacology, and even economics. Mixed effects models are composed of two nested levels of modeling: on the one hand a common structural modeling of the phenomenon of interest parameterized for each individual in the population by specific individual parameters, on the other hand a modeling of these individual parameters as random variables which accounts for their variability within the population. The first modeled the intra individual variability among the repeated measurements of each individual, the latter stands for the inter individual variability between individuals in the population.

The structural modeling of the phenomenon of interest can be linear or non-linear in the individual parameters. Non-linear type modeling can in particular account for complex phenomena modeled mechanistically by models which integrate, for example, physical or biological knowledge. The parameters of such models are often of strong practical interest because they are interpretable from an applied point of view. In such setting a central objective is to characterize and explain the variation of these individual parameters within the population. Therefore the modeling of individual parameters is done through random effects at the individual level and can also integrate descriptive covariates of the individuals.

Depending on the context, these descriptive covariates can be of high dimension and the objective is then to identify those which are the most relevant to explain the variabilities observed within the population by estimating the associated vector of regression parameters. Let us consider the field of plant ecophysiology. Mechanistic models have been proposed to describe plant development processes. These models integrate descriptive variables of the environment as covariates acting directly on the plant development process. The parameters of these models are often physical quantities such as leaf appearance speeds or light interception capacities. These parameters may vary when considering a population of plants from different varieties each characterized by its genotype (see Baey et al. [2018]). Their variations can be modeled into a mixed effects model and also integrate large genetic markers characterizing the genotypic variability within the population. In this context, identifying the relevant covariates among a set of high-dimensional covariates amounts to identifying the genetic markers that influence the phenomenon of interest (e.g. SNP, Bhatnagar et al. [2020]).

From the point of view of inference in mixed effects models including high-dimensional covariates, the objective is on the one hand to select the relevant covariates from a set of high-dimensional covariates in order to identify a parsimonious model with a reduced number of parameters and on the other hand to estimate the parameters in the reduced model. There are two main difficulties for the inference task. The first one is the high dimension of the covariates. The selection of relevant covariates can be done via a regularization approach. Furthermore, in the context of mixed effects models, an additional difficulty appears due to the presence of random effects which are not observed. This is a classic context of latent variable models. Inference is complex to carry out in this framework due to the latent structure and involves the use of efficient numerical methods. Parameter inference can be done for example by maximum likelihood via Expectation Maximization (EM) or stochastic gradient type algorithms. In the context of exponential family models, EM type algorithms are easy to implement and have good theoretical properties (see Dempster et al. [1977]; Delyon et al. [1999]). On the other hand, to our knowledge, there are no theoretical results of convergence outside the framework. In addition, the implementation of these algorithms outside the exponential family is more complex. To get around this limitation, a trick called exponentialization trick is sometimes used in practice. However, its limits have been highlighted in Debavelaere and Allassonnière [2021]. In particular, this procedure can generate significant estimation biases due to the fact that the

inference of the parameters is carried out in an extended model different from the initial model. Stochastic gradient methods can be applied in more generic models, in particular outside the exponential family. Theoretical guarantees of convergence towards an extremum of the target function have been established. However, these methods are quite in practice sensible to the tuning of the sequence of gradient steps, in particular in high-dimensional parameter spaces due to the heterogeneity of the different components of the gradient. Adaptive choices of the gradient step and procedures based on gradient preconditioning have been proposed in generic contexts to overcome this computational difficulty. More recently in the context of maximum likelihood estimation in general latent variable models, a stochastic gradient algorithm integrating a gradient preconditioning step based on an estimator of the Fisher information matrix obtained as a product A derivative of the algorithm has also been proposed, opening new possibilities for maximum likelihood inference.

In order to achieve the objective of selecting variables from a set of high-dimensional variables in mixed effects models, several regularization approaches have been developed. Jürg Schelldorfer and Bühlmann [2014] proposed a maximum likelihood estimator with a LASSO penalty (Tibshirani [1996]) in the context of linear mixed effects models with a Gaussian error term and developed a R package (see Schelldorfer et al. [2014]). Fort et al. [2019] and Ollier [2022] proposed estimators with more general penalties in the case of nonlinear mixed effects models belonging to the curved exponential family. Bertrand and Balding [2013] have also proposed an stochastic penalized versions of the EM (Delyon et al. [1999]) algorithm to take into account multiple parameters in pharmacokinetic models. Bayesian approaches based on "spike and slab" distributions have also been developed by Heuclin et al. [2020] in the case of mixed linear models and by Naveau et al. [2024] in the non linear case. On the other hand, to our knowledge, there are no high-dimensional variable selection methods for nonlinear mixed-effects models outside the exponential family.

In this contribution, we consider a maximum likelihood estimate regularized via a LASSO-type penalty in a general mixed effects model with high dimensional covariates explaining individual parameters variability. In particular, it is not required that the model belongs to the curved exponential family. To calculate this estimator, we propose an adaptive weighted proximal stochastic gradient algorithm to simultaneously handle the latent variables and the high dimensionality of the covariates. This paper is organised as follows. The second section introduces the mixed-effects models with high-dimensional covariates and presents some classical examples. The third section is devoted to the description of the proposed estimation and variable selection procedure dedicated to the high-dimensional setting. The fourth section gathers details of the numerical methodology. Finally, we present a simulation study and discuss the potential of the proposed method.

2 Nonlinear Mixed Effects Model with high dimensional covariates

2.1 Model description

Let N be a positive integer. We consider J repeated measurements for each individual i. Therefore we have J observations per individual. Let us denoted by $Y_{i,j}$ the $j-$th observation of the $i-$th individual for $1 \leq i \leq N$ and $1 \leq j \leq J$. We assume that $Y_{i,j}$ takes value in $\mathbb{R}^d$. We model this observation with a nonlinear mixed effects model (see Pinheiro and Bates [2006] and Davidian [2017]):

$$\begin{cases} Y_{i,j} & = m(\alpha, t_{i,j}, Z_i) + \varepsilon_{i,j}, \\ Z_i & \overset{i.}{\sim} \mathcal{N}(X_i\beta, \Omega), \\ \varepsilon_{i,j} & \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma). \end{cases}$$

where m is a nonlinear function depending on population parameter α taking values in $\mathbb{R}^a$, individual covariates $(t_{i,j})$ and individual parameters modeled by the latent variable Z_i. Note that in the case of longitudinal data, the covariates $t_{i,j}$ stands for the $j-$th observation time of individual i. The term $\varepsilon_{i,j}$ is a centered additive noise with covariance matrix Σ. The latent variable Z describes the inter-individual variability of the population. Individual parameters Z_i for the $i-$th subject are $q-$dimensional random vectors, independent of $\varepsilon_{i,j}$ and assumed to be distributed as a Gaussian distribution with expectation $X_i\beta$ and covariance matrix Ω, where X_i is a matrix of size $q \times p$ and β is a vector of size p. The noise term $\varepsilon_{i,j}$ is usually assumed centered Gaussian with unknown covariance matrix Σ. The unknown parameters of the nonlinear mixed-effects model are therefore $\theta = (\alpha, \beta, \Omega, \Sigma) \in \mathbb{R}^a \times \mathbb{R}^p \times \mathbb{S}_q^{++} \times \mathbb{S}_d^{++}$ where $\mathbb{S}_d^{++}$ stands for the set of symmetric positive definite matrices of size $d \times d$.

We emphasize that our model formulation is very general and allows us to consider model function m depending on both population parameters and individual parameters. We consider the setting of high-dimensional covariates where p can be much larger than N and assume that only a small subset of covariates are relevant to be included in the model.

Remark 2.1. More general distributions than the Gaussian one can be chosen for the noise term as well as more general settings with various numbers of observations per individual can be considered.

Remark 2.2. It may be useful in practice to reparametrize the latent variables and/or the parameters. For example, consider as individual parameter $\log(Z)$ rather than Z if individual parameters are assumed to be positive. We refer for more details to the excellent reparametrization cookbook (Leger [2023]).

2.2 Practical examples

2.2.1 Logistic growth curves

We consider as first example the specific case of the logistic curve model used to model growth dynamic, which is commonly used in the nonlinear mixed-effect models' community and presented in Pinheiro and Bates [2006]. The observation $Y_{i,j}$ is the circumference of the tree i at time t_{ij} and is modeled througth the logistic model with the function m given by:

$$m(\alpha, t_{ij}, Z_i) = \frac{Z_{i1}}{1 + \exp\left(-\frac{t_{ij} - Z_{i2}}{\alpha}\right)}, \tag{1}$$

where the individual parameters Z_{1i} represents the asymptotical maximum value of the circumference, Z_{2i} represents the value of the sigmoid's midpoint, and α represents the logistic growth rate.

2.2.2 Pharmacodynamic model

We consider as a second example the two compartments pharmacodynamic model used by Pinheiro and Bates [2006]. The observation $Y_{i,j}$ is the serum concentration measured at time $t_{i,j}$. The model is given by:

$$m(t_{ij}, Z_i) = \frac{D_i k_{ai}}{V_i(k_{ai} - Cl_i/V_i)} \left(\exp(-k_{ai} t_{ij}) - \exp\left(-\frac{Cl_i}{V_i} t_{ij}\right)\right), \tag{2}$$

where the individuals parameters V_i represents the distribution volume, k_{ai} the absorption rate, Cl_i the clearance and Cl_i/V_i the elimination rate; D_i is the known dose of the drug receive by individual i.

3 Estimation and Variable Selection procedure

In this section, we propose a method for simultaneously estimating the model parameters and selecting the relevant covariates.

3.1 Estimation in Latent Variable Model

We consider the maximum likelihood estimator to infer the Non-Linear Mixed-Effects model's parameters. In the context of latent variable models, the marginal likelihood, denoted by g, is obtained by integrating the complete likelihood over the latent variables, which are not observed.

$$\begin{aligned} g(\theta; Y) &= \prod_{i=1}^{N} \int f(\theta; Y_i, Z_i) dZ_i = \prod_{i=1}^{N} \int p_\theta(Y_i|Z_i) p_\theta(Z_i) dZ_i \\ &= \prod_{i=1}^{N} \int \left\{ \prod_{j=1}^{J} p_\theta(Y_{i,j}|Z_i) \right\} p_\theta(Z_i) dZ_i \end{aligned} \tag{3}$$

where $f(\theta; Y, Z)$, $p_\theta(Y|Z)$, $p_\theta(Z)$ are respectively the density of the pair (Y, Z), the density of Y conditionally to Z, and the density of Z.

One can usually estimate the model parameters by maximizing the marginal likelihood using the maximum likelihood estimator written as follows:

$$\hat{\theta}^{\text{MLE}} = \arg\max_{\theta \in \Theta} \{ g(\theta; Y) \}, \tag{4}$$

where Θ denotes the parameter space. From a practical point of view, this estimate can often not be calculate explicitly. To deal with latent variables, classical methods used to infer the unknown parameters are Expectation Maximization like algorithms (see Ng et al. [2012]). However a limit of these procedures is that they are well adapted to models of the curved exponential family. One can use the exponentiation trick by defining certain parameters as Gaussian variables and study an augmented model that belongs to the exponential family. However, its limits have been highlighted in Debavelaere and Allassonnière [2021]. Recently Baey et al. [2023] have presented a preconditioned stochastic gradient descent for estimation in a latent variable model adapted to general latent variables models, in particular it is not required that the model belongs to the curved exponential family.

3.2 Penalized likelihood in high-dimensional setting

In our context, we must deal with the high dimension of the covariates, therefore we introduce a regularization term and consider a penalized maximum likelihood estimator. We aim to select relevant variables among the covariates and use the LASSO (Least Absolute Shrinkage and Selection Operator) procedure, which was initially developed for linear regression models in Tibshirani [1996]. This method enables us to handle high-dimensional covariates and select a subset of explanatory covariates from a large collection. We consider a LASSO penalty, which only depends on the parameter β:

$$\text{pen}_\lambda(\theta) = \lambda \|\beta\|_1 = \lambda \sum_{k=1}^{p} |\beta_k|, \tag{5}$$

where λ is a positive real called the regularization parameter. Our goal is then to maximize the penalized criterium defined as the difference of the logarithm of the marginal likelihood and of the penalty term. Let us define the penalized maximum likelihood estimator by:

$$\hat{\theta}_\lambda^{LASSO} = \arg\max_{\theta \in \Theta} \{ \log g(\theta; Y) - \text{pen}_\lambda(\theta) \}, \tag{6}$$

where Θ denotes the parameter space and where λ is a positive parameter. The larger the value of λ, the more β will be constrained to have zero components. Conversely, the smaller the value of λ, the freer the components of β will be. It is customary to determine the value of λ using cross-validation or using model criterion (see Tibshirani [1996]). We will consider the eBIC criterion well adapted to the high dimensional setting to find on optimal regularization value (see Chen and Chen [2008]). To evaluate this penalized estimate in practice, we will use a proximal algorithm due to the non-differentiability of the considered penalty as presented by Fort et al. [2019]. Therefore, we will propose finally a preconditioned stochastic proximal gradient algorithm to calculate the estimator.

3.3 Regularization Path

To select only the most explanatory covariates, it's important to choose a well-balanced value for the regularization parameter λ. We choose an optimal parameter by minimizing the extended Bayesian Information Criterion (eBIC) (see Chen and Chen [2008]). Guided by the intuition given by Delattre et al. [2014], we penalize the number of degrees of freedom by the log of the total number of observations $N \times J$. We consider a grid Λ of values for the regularization parameter λ. We conduct then the following inference methodology :

i) For all $\lambda \in \Lambda$ repeat the following steps:

- calculate $\hat{\theta}_\lambda^{LASSO} = \arg\max_{\theta\in\Theta} \{\log g(\theta; Y) - \text{pen}_\lambda(\theta)\}$.
- deduce the associated support $\hat{S}_\lambda = \{j \in \{1, \dots, p\}, \hat{\beta}_{\lambda,j}^{LASSO} \neq 0\}$
- calculate $\hat{\theta}_\lambda^{\text{MLE}} = \arg\max_{\theta\in\Theta} \{\log g_\lambda(\theta; Y)\}$, where g_λ is the marginal likelihood in the model restricted to the support $\hat{S}_\lambda$.
- calculate the eBIC criterion:

$$\text{eBIC}(\lambda) = -2 \log g_\lambda(\hat{\theta}_\lambda^{MLE}; Y) + |\hat{S}_\lambda| \log(NJ) + 2 \log \left(\binom{p}{|\hat{S}_\lambda|} \right) \tag{7}$$

ii) Select the parameter λ that minimizes the eBIC:

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \text{eBIC}(\lambda),$$

and consider the final estimator defined by $\hat{\theta}_{\hat{\lambda}}^{MLE}$.

Remark 3.1. Note that the quantity $g_\lambda(\hat{\theta}_\lambda^{MLE}; Y)$ does not have an explicit form due to the presence of the latent variables Z and the non-linearity of the model. Therefore we use a Monte Carlo procedure to calculate an approximation of $g_\lambda(\hat{\theta}_\lambda^{MLE}; Y)$.

4 Numerical methodology

Now that we have a procedure for choosing a good regularization parameter, we need a procedure to calculate the maximum likelihood estimate. Due to the integral, it is difficult to directly compute the maximum of the marginal likelihood, which does not have an analytical form in this latent variable model. Therefore, we will use numerical methods to solve this maximization problem. We propose a stochastic gradient algorithm preconditioned. We explain in the following the work on which we rely to derive our algorithm and finally the steps to implement it.

4.1 Adaptive Weighted Proximal Stochastic Gradient Descent Algorithm

Baey et al. propose a new Preconditioned Stochastic Gradient descent algorithm (PSG) for maximum likelihood estimation in models with latent variables. In particular, it does not require the joint density to belong to the curved exponential family. The authors propose to use as preconditioner a Fisher Information Matrix (FIM) estimator based on Delattre and Kuhn [2023] and provide theoretical convergence results in this setup. In our context of high-dimensional setting, it is difficult to manage a full matrix as preconditioner from a computational point of view. Therefore we consider rather an adaptive vectorial learning rate. We also add a step to take into account the non differentiable penalization term. The main idea is to combine the PSG with a Proximal Forward-Backward algorithm (Chen and Rockafellar [1997]; Tseng [2000]). The algorithm is divided into three steps; a realization of the latent variables is sampled with a first step called *Simulation*, which uses a direct sampling from the posterior distribution or a Metropolis-Hastings sampler (Geman and Geman [1984]). The second step is the preconditioned gradient descent on the approximate complete likelihood, the *Forward* step. The last step, called *Backward*, deals with the penalty term. We apply the proximal operator (Moreau [1962]; Rockafellar [1976]; Chouzenoux et al. [2014]). We present in detail in the following sections the adaptive learning rate procedure, the associated proximal operator and finally the proposed Adaptive Weighted Proximal Stochactic Gradient Descent algorithm called AWPSG.

4.2 Adaptive vectorial learning rate in Stochastic Gradient Descent Algorithm

Adaptive algorithms such as AdaGrad [Duchi et al., 2011], RMSPprop [Tieleman, 2012] and Adam [Kingma and Ba, 2017] have proved their worth. We choose to benefit from the advantages offered by the Adam algorithm. Given random effect realizations we calculate and store the values of the gradient of the log-likelihood, and define the preconditioning diagonal matrix as follows in order to scale the different components of the gradient, homogenizing the evolution of the algorithm. On the other hand, we introduce a momentum in the gradient, guided by Adam, using an exponentially weighted moving average of past gradients.

4.3 Description of the Weighted Proximal Operator

First of all, for any symmetric positive-definite matrix $A \in \mathbb{S}_d^{++}$, let's denote by $\|.\|_A$ the A-weighted norm defined with the A-weighted scalar product by $\forall (x,y) \in \mathbb{R}^d, \langle x,y \rangle_A = \langle x, Ax \rangle$ and $\|x\|_A^2 = \langle x,x \rangle_A$.

Definition 4.1. Weighted Proximal Operator Assume that $A \in \mathbb{S}_d^{++}$. Let $\psi : \Theta \to \mathbb{R}$ be a proper, lower semicontinuous, convex function and $\theta \in \Theta$. The weighted proximal operator of ψ in θ relative to the metric induced by A is the unique minimizer of $\psi(\theta') + \frac{1}{2}\|. - \theta\|_A^2$:

$$\text{Prox}_{A,\psi}(\theta) = \underset{\theta' \in \mathbb{R}^d}{\arg\min} \left(\psi(\theta') + \frac{1}{2} \left\|\theta' - \theta\right\|_A^2 \right) \tag{8}$$

Remark 4.2. If $A = \gamma^{-1} Id$, where $\gamma > 0$, then $\text{Prox}_{\gamma^{-1} Id, f} := \text{Prox}_{\gamma f}$ is the proximal operator defines by Moreau [1962].

The θ update step of the AWPSG, the pairing of PSDG and a proximal Forward-Backward procedure, can then be written :

$$\theta_{k+1} = \text{Prox}_{A_k, \text{pen}}(\theta_k - A_k^{-1} \nabla_\theta f(\theta_k))$$

where $(\gamma_k)_{k \geq 1}$ is a sequence of step-size.

Remark 4.3. If $A_k = \gamma_k^{-1} Id$, we obtain the standard proximal stochastic gradient descent. Where $(\gamma_k)_{k \geq 1}$ is a step size such that $\forall k \in \mathbb{N}, \gamma_k \in [0,1], \sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.

In general, $\text{Prox}_{A,\text{pen}}$ is well defined but may not be quickly computable. Becker and Fadili state several assumptions on A and the function pen that allow the proximal operator to be computable. For example, rank restriction and pen is separable, i.e. $\text{pen}(x) = \sum_{i=1}^{p} \text{pen}_i(x_i)$. If no assumptions are made, the proximal operator is often not explicit and will require additional steps. To limit computational costs, we propose to use an adaptative preconditioning diagonal matrix and LASSO penalization. From now on will be denoted $s = diag(A)$ for the sake of clarity. In this configuration the proximal operator has an explicit form :

$$(\text{Prox}_{s,\text{pen}_\lambda}(\beta)) = \begin{cases} 0 & \text{if } |\beta_i| < \lambda s_i \\ \beta_i - \lambda s_i & \text{if } \beta_i \geq \lambda s_i \\ \beta_i + \lambda s_i & \text{if } \beta_i \leq -\lambda s_i \end{cases} ; \forall i \in \{1, ..., p\}. \tag{9}$$

We provide in detail the steps of the algorithm.

Algorithm 1: Adaptive Weighted Proximal Stochastic Gradient (AWPSG)

Require: γ_0 learning rate
Require: γ_1, γ_2 exponential decat rates for the moment estimate
Initialize Starting point $\theta_0 \in \mathbb{R}^d$,
$m_0 \leftarrow 0$ Initialize moment vector,
$s_0 \leftarrow 0$ Initialize Adaptive stepsize vector.
while θ_k *not converged* **do**
  $k \leftarrow k + 1$
  • **Simulation step :**
    Draw $Z^{(k)}$from the posterior distribution $p(.|\theta_k)$ or using a single step of a Hastings Metropolis procedure having as stationary distribution the posterior.
  • **Gradient computation :** $v_k \leftarrow \frac{1}{N} \sum_{i=1}^{N} \nabla \log f_{\theta_k}(Y_i, Z_i^{(k)})$
  • **Moment estimate :** $m_k \leftarrow \gamma_1 m_{k-1} + (1 - \gamma_1) v_k$
  • **Adaptive stepsize :** $s_{k,l}^2 \leftarrow \gamma_2 s_{k-1,l}^2 + (1 - \gamma_2) v_{k,l}^2$ for $l \in \{1, \ldots, p\}$
  • **Update parameters :**
    • **Forward step :** $\omega_{k,l} \leftarrow \theta_{k-1,l} + \gamma_0 m_{k,l} / s_{k,l}$ for $l \in \{1, \ldots, p\}$
    • **Backward step :** $\theta_{k,l} \leftarrow \text{Prox}_{s_{k,l}^2, \text{pen}_\lambda}(\omega_{k,l})$ for $l \in \{1, \ldots, p\}$
end
Return $\hat{\theta} = \theta_k$

4.4 Regularization Path

In this section, we give more detail about the procedure used to choose a regularization parameter λ (see section 3.3). The minimization of the eBIC criterion (7) calculated on a proposal grid Λ can be represented by the regularization path which can be plotted to visualize the evolution of the support of β with respect to the regularization parameter λ (e.g. figure 2). This type of graph shows that the larger the regularization parameter, the more the component of the vector β is restricted to zero. The smaller λ is, the more β components are free. In practice, the proposal grid Λ is built on a logarithmic scale. In fact, it's preferable to explore the smaller values a little more, in order to capture rapid changes in the support of β. This way, we don't spend too much time on ranges of regularization values that would have returned the same support.

5 Numerical experiments

In this section, we study the performance of the procedure presented. The numerical study is divided into two parts. First, we present results in a linear mixed-effect model. In the

second part we test the robustness of our method for fitting a nonlinear mixed-effect model in different configurations. For each scenario, we generate n_{run} independent data sets, and we fit the corresponding model using the routine described in section 3.3 and get an estimate $\hat{\theta}_i^{\text{MLE}}$ for each run $i \in \{1, ..., n_{\text{run}}\}$. To compare the results across different scenarios, we evaluate several metrics: the Relative Root Mean Square Errors (RRMSE) to measure the estimation quality of the method defined as:

$$\text{RRMSE}(\hat{\theta}_k^{\text{MLE}}) = \sqrt{\frac{1}{n_{\text{run}}} \sum_{i=1}^{n_{\text{run}}} \frac{(\hat{\theta}_{i,k}^{\text{MLE}} - \theta_k)^2}{\theta_k^2}} \quad \text{mse}(\hat{\theta}_k^{\text{MLE}}) = \frac{1}{n_{\text{run}}} \sum_{i=1}^{n_{\text{run}}} (\hat{\theta}_{i,k}^{\text{MLE}} - \theta_k)^2$$

where $\hat{\theta}_{i,k}^{\text{MLE}}$ is the k-th component of $\hat{\theta}_i^{\text{MLE}}$. We also evaluate the sensitivity, specificity, and accuracy to study the selection capacity of the method. Sensitivity (Se) measures the proportion of true positives (TP) correctly identified, while Specificity (Sp) quantifies the proportion of true negatives (TN) correctly identified. Accuracy (Ac) represents the overall proportion of correctly classified instances, including both TP and TN. We abbreviated false negatives and false positives respectively by FN and FP.

$$\text{Se} = \frac{TP}{TP + FN} \quad \text{Sp} = \frac{TN}{TN + FP} \quad \text{Ac} = \frac{TP + TN}{P + N}$$

5.1 Simulation study in a Linear Mixed Effects Model in high dimension

In this section, we compare our method AWPSG to the R package GLMMLASSO (Groll and Tutz [2014]). We use the same procedure to select the regularization parameter. We consider the following linear model :

$$\begin{cases} Y_{i,j} & = Z_{i,j1} + Z_{i2} \times t_j + \varepsilon_{i,j} \\ Z_{i,j1} & \overset{i.}{\sim} \mathcal{N}(\mu_1 + X_{i,j}\beta, \gamma_1^2) \\ Z_2 & \overset{i.i.d}{\sim} \mathcal{N}(\mu_2, \gamma_2^2) \\ \varepsilon_{i,j} & \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \end{cases} \tag{10}$$

The model parameters are $\beta \in \mathbb{R}^p$, $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$, and $\gamma_1^2, \gamma_2^2, \sigma^2 \in \left(\mathbb{R}_+^*\right)^3$. In this model, high-dimensional covariates model variability between individuals. We generated 30 data sets independently according to equation (10) for several scenarios. It is assumed that each individual was observed $J = 10$ times, at the same instants spread over a range between 0 and 1, at equal intervals. We use the following values for the parameters : $\mu = (2, 5)$, $\gamma_1 = 1, \gamma_2 = 2$ and $\sigma = 1$. For each different value of p, we choose the vector β such that the first three components are equal to $(\beta_1, \beta_2, \beta_3) = (8, -10, 20)$ and the rest are equal to zero. Additionally, we generate the matrix of covariates X with N rows and p columns, following a uniform distribution $X_{i,k} \sim \mathcal{U}([-1, 1])$; $\forall i \in \{1, ..., N\}, \forall k \in \{1, ..., p\}$.

By testing scenarios with increasing numbers of individuals $N \in \{100, 200\}$, we highlight that our method estimates more accurately when the sample size increases. In parallel we show different scenarios where the number of covariates increases $p \in \{200, 500\}$. We demonstrate how our method, despite the high-dimensional context, manages to select the most explanatory variables. Tables 1 and 5 (in Appendix) show the mean value over the 30 data sets of the Relative Root Mean Square Errors (RRMSE) for each model parameter, along with their estimates. To assess the selection variable capacity of our method, we present selection scores in Table 2. Whatever the configuration, our method AWPSG and GLMMLASSO perform very similar in selecting the right covariates. However, we emphasize that AWPSG also adequately estimates the variances of the random effects in this model, unlike GLMMLASSO.

Table 1: Average Relative Root Mean Square Errors (RRMSE) and parameter estimates over 30 repetitions for the Linear Mixed Effects Model (LMEM) with $p = 500$, where $\hat{\theta}_{\text{AWPSG}}$ and $\hat{\theta}_{\text{GLMMLASSO}}$ are the estimates using the AWPSG and GLMMLASSO methods, respectively.

		P = 500							
		N = 100				**N = 200**			
		AWPSG		GLMMLASSO		AWPSG		GLMMLASSO	
	θ^*	$\hat{\theta}_{\text{AWPSG}}$	RRMSE	$\hat{\theta}_{\text{GLMMLASSO}}$	RRMSE	$\hat{\theta}_{\text{AWPSG}}$	RRMSE	$\hat{\theta}_{\text{GLMMLASSO}}$	RRMSE
μ_1	2.00	1.98	6.36	1.98	6.15	1.98	3.97	2.00	3.64
μ_2	5.00	5.03	4.03	5.03	3.69	5.01	2.70	5.01	3.49
γ_1^2	1.00	1.00	20.55	0.61	39.38	1.01	13.92	0.61	39.53
γ_2^2	4.00	3.93	17.32	0.88	78.11	4.03	12.72	0.88	77.92
σ^2	1.00	0.99	4.99	0.98	5.00	1.01	3.91	0.98	3.97
β_1	8.00	8.14	7.33	8.13	7.45	8.08	5.35	7.98	5.45
β_2	-10.00	-9.98	6.05	-10.00	6.63	-9.99	3.91	-9.96	4.22
β_3	20.00	19.98	3.30	19.97	3.42	19.98	1.91	19.97	2.14

Table 2: Average of sensitivity (Se), specificity (Sp), accuracy (Ac), mean square error (mse) over 30 repetitions within the LMEM.

	P = 200				**P = 500**			
	N = 100		**N = 200**		**N = 100**		**N = 200**	
	AWPSG	GLMMLASSO	AWPSG	GLMMLASSO	AWPSG	GLMMLASSO	AWPSG	GLMMLASSO
Ac	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Se	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sp	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
mse	4.25e-03	3.44e-03	1.96e-03	1.87e-03	1.90e-03	1.94e-03	5.99e-04	8.17e-04

5.2 Simulation study in a Non-Linear Mixed-Effects Model with high dimensional covariates

We study this section the model presented in (1), where we explain a part of the variability of the logistic's midpoint by the high-dimensional covariates. The model can be written as follows:

$$\begin{cases} Y_{i,j} & = \dfrac{Z_{i1}}{1+\exp\left(-\frac{t_{ij}-Z_{i2}}{\alpha}\right)} + \varepsilon_{i,j} \\ Z_{i,1} & \overset{i.i.d}{\sim} \mathcal{N}(\mu_1, \gamma_1^2) \\ Z_{i,2} & \overset{i.}{\sim} \mathcal{N}(\mu_2 + X_i\beta, \gamma_2^2) \\ \varepsilon_{i,j} & \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \end{cases} \tag{11}$$

The model parameters are $\beta \in \mathbb{R}^p$, $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$, and $\gamma_1^2, \gamma_2^2, \sigma^2 \in \left(\mathbb{R}_+^*\right)^3$. Z_{i1} represents the asymptotic maximum value of the curve, Z_{i2} represents the value of the logistic's midpoint, and α represents the logistic growth rate. Note that due to the presence of the fixed effect α, the joint density of (Y_i, Z_i) does not belong to the curved exponential family as defined in Delyon et al. [1999]. We generated 100 data sets independently according to equation (11) for several scenarios. It is assumed that each individual was observed $J = 15$ times, at the same instants equally spaced over a range between 150 and 3000. We use the following values for the parameters: $\mu = (200, 1200)$, $\gamma_1^2 = 7^2$, $\gamma_2^2 = 30^2$, $\alpha = 300$ and $\sigma^2 = 30$. For each different value of p, we choose the vector β such that the first three components are equal to $(100, 200, -300)$ and

the rest equal to zero. Furthermore, we generate the matrix of covariates X_i with N rows and p columns, following a uniform distribution $X_{i,k} \sim \mathcal{U}([-1,1])$; $\forall i \in \{1,...,N\}, \forall k \in \{1,...,p\}$.

As in the linear model, we present scenarios with an increasing number of individuals $N \in \{100, 200\}$ and different covariates size $p \in \{200, 500, 1000\}$. Tables 3 and 6 summarize the results of the estimation of the model parameters for $N = 100$ and $N = 200$, respectively,y, and table 4 presents selection scores of the covariates. We observe that the estimation of the parameters is accurate in all configurations. As expected, the Relative Root mean square errors decrease with N for a given covariate size p. For $N = 100$, we observe that the relative root mean square errors generally increase with p, which can be explained by the fact that increasing the size of the covariates increases the difficulty of the selection procedure. The phenomenon is less pronounced as the sample size increases, reflecting the asymptotically expected behavior.

Table 3: Average Relative Root Mean Square Errors (RRMSE) and parameter estimates over 100 repetitions for the Non Linear Mixed Effects Model (NLMEM) with $N = 100$.

		N = 100					
		P = 200		**P = 500**		**P = 1000**	
	θ^*	$\hat{\theta}$	RRMSE	$\hat{\theta}$	RRMSE	$\hat{\theta}$	RRMSE
μ_1	200.00	200.08	0.51	199.95	0.51	199.98	0.48
μ_2	1200.00	1200.87	0.28	1200.69	0.28	1200.47	0.30
γ_1^2	49.00	48.61	14.91	49.01	15.33	48.68	15.16
γ_2^2	900.00	847.08	19.46	832.47	20.33	967.13	98.62
α	300.00	300.24	0.58	300.03	0.65	300.23	0.64
σ^2	30.00	30.11	3.99	30.10	3.96	30.05	4.15
β_0	100.00	100.45	4.33	99.90	3.96	97.99	14.75
β_1	200.00	200.13	1.83	200.06	1.95	199.63	2.02
β_2	-300.00	-300.50	1.19	-300.40	1.34	-299.67	1.58

Table 4: Average sensitivity (Se), specificity (Sp), accuracy (Ac), and mean square error (mse) over 100 repetitions for the Non-Linear Mixed Effects Model (NLMEM)

	N = 100			**N = 200**		
	P = 200	**P = 500**	**P = 1000**	**P = 200**	**P = 500**	**P = 1000**
Ac	1.000	1.000	1.000	1.000	1.000	1.000
Se	1.000	1.000	0.993	1.000	1.000	1.000
Sp	0.999	1.000	1.000	1.000	1.000	1.000
mse	2.61e-01	1.10e-01	2.72e-01	1.22e-01	4.98e-02	2.54e-02

As an illustration of the behavior of the algorithm, figure 1 displays the estimated parameter as a function of iterations during the execution of the AWPSG algorithm. We can observe the convergence of the algorithm to the true parameter values used in the simulation.

6 Conclusion and perspectives

In this work, we jointly address variable selection in high-dimension and parameter estimation in general mixed-effect model. We emphasize that the proposed procedure can handle linear and non-linear mixed effects models, without requiring them to belong to the curved exponential family. We perform an adaptive weighted proximal stochastic gradient algorithm to deal simultaneously with the latent variables and the LASSO penalty when maximizing the

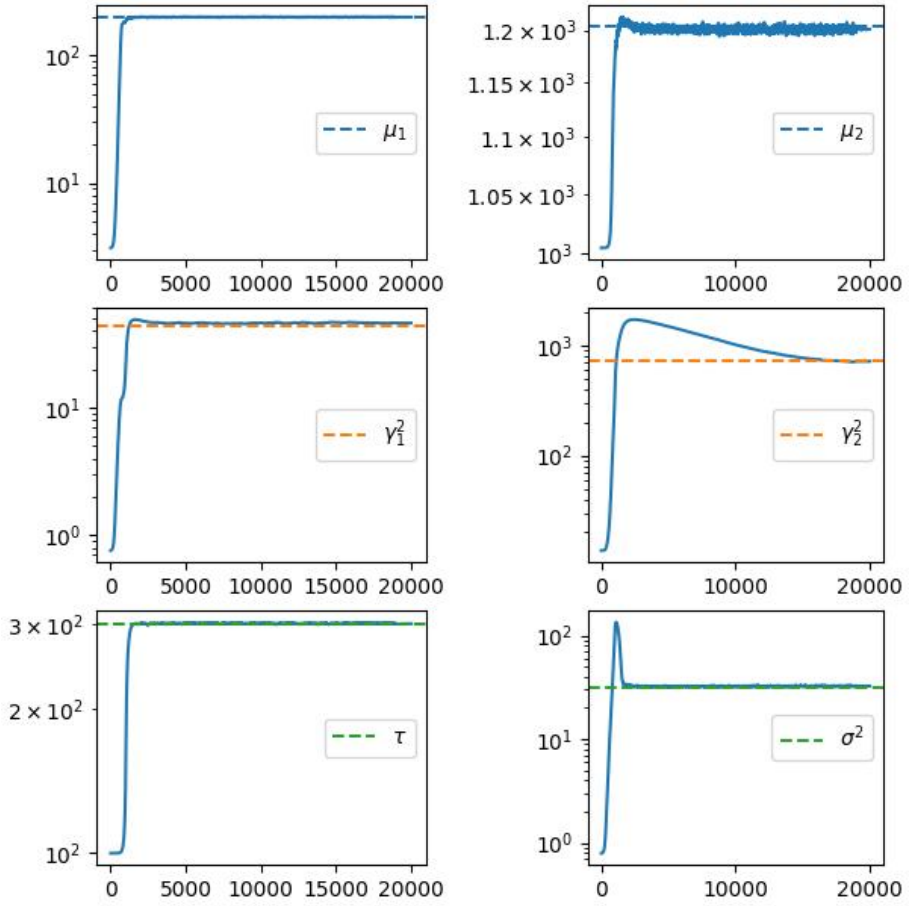

Figure 1: Parameter estimates across the AWPSG iterations within the logistic NLMEM. Dotted lines : simulated value

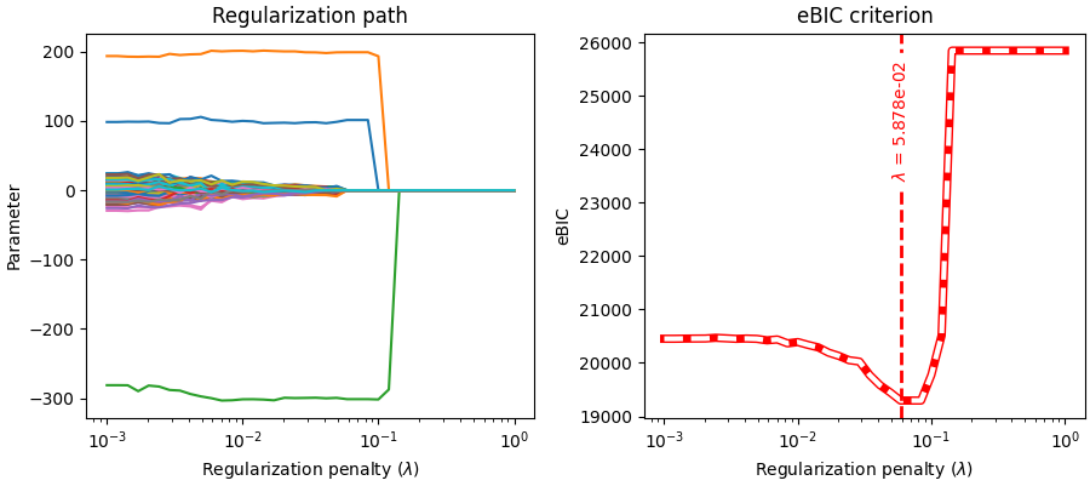

Figure 2: Regularization path, i.e. values of β, in solid line and the eBIC criterion in dotted line; the dotted vertical lines represent the chosen regularization values

penalized likelihood. We use the eBIC criterion to perform the model selection procedure. Our methodology has been thoroughly evaluated in simulation study to demonstrate its performance. Moreover, we emphasize that it is possible to consider a wide range of modeling choices for the error term. An interesting perspective of this work consists of studying theoretical properties of the estimator and prediction after the model selection step.

Funding and Acknowledgements

This work was funded by the https://stat4plant.mathnum.inrae.fr/(Stat4Plant) project ANR-20-CE45-0012.

References

Charlotte Baey, Amélie Mathieu, Alexandra Jullien, Samis Trevezas, and Paul-Henry Cournède. Mixed-effects estimation in dynamic models of plant growth for the assessment of inter-individual variability. *Journal of agricultural, biological and environmental statistics*, 23: 208–232, 2018.

Charlotte Baey, Maud Delattre, Estelle Kuhn, Jean-Benoist Leger, and Sarah Lemler. Efficient preconditioned stochastic gradient descent for estimation in latent variable models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Stephen Becker and Jalal Fadili. A quasi-newton proximal splitting method. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper_files/paper/2012/file/e034fb6b66aacc1d48f445ddfb08da98-Paper.pdf`.

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

Sahir R Bhatnagar, Yi Yang, Tianyuan Lu, Erwin Schurr, JC Loredo-Osti, Marie Forest, Karim Oualkacha, and Celia MT Greenwood. Simultaneous snp selection and adjustment for population structure in high dimensional prediction models. *PLoS genetics*, 16(5):e1008766, 2020.

George H.-G. Chen and R. Tyrrell Rockafellar. Convergence rates in forward-backward splitting. *SIAM J. Optim.*, 7:421–444, 1997. URL `https://api.semanticscholar.org/CorpusID:7104716`.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008. ISSN 00063444, 14643510. URL `http://www.jstor.org/stable/20441500`.

Emilie Chouzenoux, Jean-Christophe Pesquet, and Audrey Repetti. Variable metric forward—backward algorithm for minimizing the sum of a differentiable function and a convex function. *J. Optim. Theory Appl.*, 162(1):107–132, July 2014. ISSN 0022-3239. doi: 10.1007/s10957-013-0465-7. URL `https://doi.org/10.1007/s10957-013-0465-7`.

Marie Davidian. *Nonlinear models for repeated measurement data*. Routledge, 2017.

Vianney Debavelaere and Stéphanie Allassonnière. On the curved exponential family in the stochastic approximation expectation maximization algorithm. *ESAIM: Probability & Statistics*, 25, 2021.

Maud Delattre and Estelle Kuhn. Computing an empirical Fisher information matrix estimate in latent variable models through stochastic approximation. *Computo*, 2023. ISSN 2824-7795. doi: 10.57750/r5gx-jk62. URL `https://computo.sfds.asso.fr/published-202311-delattre-fim/`.

Maud Delattre, Marc Lavielle, and Marie-Anne Poursat. A note on BIC in mixed-effects models. *Electronic Journal of Statistics*, 8(1):456 – 475, 2014. doi: 10.1214/14-EJS890. URL https://doi.org/10.1214/14-EJS890.

Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999. ISSN 00905364. URL http://www.jstor.org/stable/120120. Publisher: Institute of Mathematical Statistics.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Gersende Fort, Edouard Ollier, and Adeline Samson. Stochastic proximal-gradient algorithms for penalized mixed models. *Statistics and Computing*, 29:231–253, 2019.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984. ISSN 0162-8828. doi: 10.1109/TPAMI.1984.4767596. URL http://ieeexplore.ieee.org/document/4767596/.

Andreas Groll and Gerhard Tutz. Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and Computing*, 24(2):137–154, 2014. ISSN 1573-1375. doi: 10.1007/s11222-012-9359-z. URL https://doi.org/10.1007/s11222-012-9359-z.

Benjamin Heuclin, Frédéric Mortier, Catherine Trottier, and Marie Denis. Bayesian varying coefficient model with selection: An application to functional mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70, 11 2020. doi: 10.1111/rssc.12447.

Lukas Meier Jürg Schelldorfer and Peter Bühlmann. Glmmlasso: An algorithm for high-dimensional generalized linear mixed models using l1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477, 2014. doi: 10.1080/10618600.2013.773239. URL https://doi.org/10.1080/10618600.2013.773239.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Jean-Benoist Leger. Parametrization cookbook: A set of bijective parametrizations for using machine learning methods in statistical inference. *arXiv preprint arXiv:2301.08297*, 2023.

Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 255:2897–2899, 1962.

Marion Naveau, Guillaume Kon Kam King, Renaud Rincent, Laure Sansonnet, and Maud Delattre. Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the saem algorithm. *Statistics and Computing*, 34(1):53, 2024.

Shu Kay Ng, Thriyambakam Krishnan, and Geoffrey J McLachlan. The em algorithm. *Handbook of computational statistics: concepts and methods*, pages 139–172, 2012.

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

José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer science & business media, 2006.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976. ISSN 0363-0129, 1095-7138. doi: 10.1137/0314056. URL http://epubs.siam.org/doi/10.1137/0314056.

Jürg Schelldorfer, Lukas Meier, and Peter Bühlmann. Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x.

T. Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, 2012. URL https://cir.nii.ac.jp/crid/1370017282431050757.

P. Tseng. A modified forward-backward splitting method for maximal monotone mapping. *Siam Journal on Control and Optimization - SIAM*, 38, 01 2000.

Appendix

Appendix A. Simulation study results for LMEM and NLMEM

Table 5: Average Relative Root Mean Square Errors (RRMSE) and parameter estimates over 30 repetitions for the Linear Mixed Effects Model (LMEM) with $P = 200$. Where $\hat{\theta}_{\text{AWPSG}}$ and $\hat{\theta}_{\text{GLMMLASSO}}$ are the estimates using the AWPSG and GLMMLASSO methods, respectively.

		P = 200							
		N = 100				**N = 200**			
		AWPSG		GLMMLASSO		AWPSG		GLMMLASSO	
	θ^*	$\hat{\theta}_{\text{AWPSG}}$	RRMSE	$\hat{\theta}_{\text{GLMMLASSO}}$	RRMSE	$\hat{\theta}_{\text{AWPSG}}$	RRMSE	$\hat{\theta}_{\text{GLMMLASSO}}$	RRMSE
μ_1	2.00	1.97	6.85	1.98	6.13	2.00	4.35	2.00	3.90
μ_2	5.00	5.03	3.71	5.03	3.70	5.01	3.00	5.01	3.54
γ_1^2	1.00	0.99	19.59	0.61	39.74	0.99	15.24	0.61	39.24
γ_2^2	4.00	3.94	17.35	0.88	78.10	4.09	12.94	0.89	77.86
σ^2	1.00	0.99	5.09	0.98	5.07	0.99	3.52	0.98	4.08
β_1	8.00	8.04	7.22	8.05	7.20	8.05	4.40	8.03	4.61
β_2	-10.00	-9.98	5.49	-9.97	5.81	-10.04	3.88	-9.97	4.05
β_3	20.00	19.99	2.79	19.98	2.89	19.95	2.17	19.99	2.15

Table 6: Average Relative Root Mean Square Errors (RRMSE) and parameter estimates over 100 repetitions for the Non-Linear Mixed Effects Model (NLMEM) with $N = 200$.

		N = 200					
		P = 200		**P = 500**		**P = 1000**	
	θ^*	$\hat{\theta}$	RRMSE	$\hat{\theta}$	RRMSE	$\hat{\theta}$	RRMSE
μ_1	200.00	200.01	0.33	199.96	0.35	199.89	0.36
μ_2	1200.00	1199.95	0.23	1200.04	0.26	1200.07	0.24
γ_1^2	49.00	47.72	11.90	48.38	11.89	48.05	11.55
γ_2^2	900.00	903.22	12.96	904.94	14.41	902.09	13.90
α	300.00	299.63	0.46	299.84	0.45	299.86	0.42
σ^2	30.00	30.08	3.06	30.13	3.02	30.14	3.11
β_0	100.00	100.16	2.74	100.11	2.71	100.09	3.11
β_1	200.00	200.10	1.34	200.27	1.42	199.57	1.28
β_2	-300.00	-299.96	0.90	-299.67	1.02	-300.27	1.00