A SAEM Algorithm for Fused Lasso Penalized Non Linear Mixed Effect Models: Application to Group Comparison in Pharmacokinetic

Edouard Ollier, Adeline Samson, Xavier Delavenne, Vivian Viallon

A SAEM Algorithm for Fused Lasso Penalized Non Linear Mixed Effect Models: Application to Group Comparison in Pharmacokinetic

Edouard Ollier[a,b,c], Adeline Samson[b], Xavier Delavenne[c], Vivian Viallon[d]

[a] *U.M.P.A., Ecole Normale Supérieure de Lyon, CNRS UMR 5669; INRIA, Project-team NUMED. 46 Allée d'Italie, 69364 Lyon Cedex 07, France*
[b] *Université Grenoble-Alpes, Laboratoire Jean Kuntzmann, UMR CNRS 5224*
[c] *Groupe de Recherche sur la Thrombose, EA3065, Université de Saint-Etienne, Jean Monnet, F-42023, Saint-Etienne*
[d] *Université de Lyon, F-69622, Lyon, France; Université Lyon 1, UMRESTTE, F-69373 Lyon; IFSTTAR, UMRESTTE, F-69675 Bron*

Abstract

Non linear mixed effect models are classical tools to analyze non linear longitudinal data in many fields such as population Pharmacokinetic. Groups of observations are usually compared by introducing the group affiliations as binary covariates with a reference group that is stated among the groups. This approach is relatively limited as it allows only the comparison of the reference group to the others. In this work, we propose to compare the groups using a penalized likelihood approach. Groups are described by the same structural model but with parameters that are group specific. The likelihood is penalized with a fused lasso penalty that induces sparsity on the differences between groups for both fixed effects and variances of random effects. A penalized Stochastic Approximation EM algorithm is proposed that is coupled to Alternating Direction Method Multipliers to solve the maximization step. An extensive simulation study illustrates the performance of this algorithm when comparing more than two groups. Then the approach is applied to real data from two pharmacokinetic drug-drug interaction trials.

Key words: Nonlinear mixed effect model, SAEM algorithm, fused lasso, group comparison, pharmacokinetic

1. Introduction

Non Linear Mixed Effects Models (NLMEMs) are used to model and analyze longitudinal data in several fields, especially in clinical trials and population Pharmacokinetic (PK). In clinical research, observations may present a group structure corresponding to the different treatment modalities. For example, a drug-drug interaction clinical trial between two compounds includes two groups of observations, patients treated with the molecule of interest and patients treated with the two compounds. When population PK data have been collected during the trial, PK parameters are estimated through an NLMEM, and the interaction (existence and mechanism) is assessed through the variation of the PK parameters across groups.

Statistical tests are classically used to identify significant influence of the group structure on a (PK) parameter. The group affiliation is included as a categorical covariate and its influence is studied with maximum likelihood tests [Samson et al., 2007]. Note that the likelihood of

NLMEM being intractable, stochastic versions of the EM algorithm, among other methods, can be used to estimate the parameters, especially the SAEM algorithm [Delyon et al., 1999, Kuhn and Lavielle, 2005]. A stepwise procedure combined to a BIC criterion is then used to select the best model among the collection of models with the group affiliation covariate on each parameter. A drawback of this approach is that a reference group has first to be stated, and then only differences with this reference group are considered. When there are more than two groups, this does not allow to select a model with no difference between two non reference groups. In order to study the differences between non reference groups, combination of the group covariates could be used, but their number increases rapidly with the number of groups. Indeed, the number of between group differences models is equal to $(B_G)^p$ where B_G is the Bell's number [Bell, 1934] for G groups and p the number of studied parameters. Considering 5 groups and studying between group differences on 3 parameters leads to 52^3 possible models.

Nevertheless, the relevance of group differences between all the groups can be directly studied using a penalized joint modeling approach [Viallon et al., 2014, Gertheiss and Tutz, 2012, Ollier and Viallon, 2014]. The same structural model is applied to each group with a structural sparsity-inducing penalty [Bach et al., 2011] that encourages parameters to be similar in each group. In this work, group parameters are estimated by maximizing the penalized likelihood with a fused lasso penalty. This penalty was originally designed to penalize differences of coefficients corresponding to successive features [Tibshirani et al., 2005] and has been generalized to account for features with a network structure [Höfling et al., 2010].

Sparsity inducing penalties in linear mixed effects models (LMEMs) have been proposed for selecting fixed effects only [Schelldorfer et al., 2011, Rohart et al., 2014] and both fixed effects and random effects variances [Bondell et al., 2010]. Note that the joint selection of fixed effects and random effects variances is complex because the likelihood is not convex with respect to the variances. The difficulty increases with NLMEM as the likelihood is intractable (contrary to LMEMs), and only a few papers deal with penalties in NLMEM. Arribas-Gil et al. [2014] study selection of semi parametric NLMEM using a lasso penalty, the lasso selection step and the parameter estimation being realized separately. Bertrand and Balding [2013] consider l_1 penalized NLMEM for genetic variant selection. They propose a penalized version of the SAEM algorithm, in which the maximization step corresponds to an l_1 penalized weighted least square problem. The optimal sparsity parameter is set using an asymptotic estimation. The recent stochastic proximal gradient algorithm [Atchade et al., 2014] could offer an interesting solution to optimize penalized intractable likelihood but it has not been applied to NLMEMs. Up to our knowledge, no work investigates the use of a structured penalty in the context of NLMEM.

The objective of this paper is to incorporate the fused lasso penalty in the SAEM algorithm, in order to jointly estimate NLMEMs on several groups, and detect relevant differences among both fixed effects and variances of random effects. Penalties are introduced in the maximization step of the SAEM algorithm. Fixed effects and variances of random effects are penalized through a sum of absolute differences. The penalized differences correspond to edges of a graph in which the vertices correspond to the groups. Solving this penalized optimization problem is not trivial and we suggest to use an Alternating Direction Method of Multipliers (ADMM) [Boyd et al., 2011]. The direct penalization of the variances yielding to a non convex optimization problem, we propose to penalize the inverse variance-covariance matrix, assuming the matrix is diagonal. An ADMM algorithm can be used to solve the penalized optimization problem, its proximal step being explicit or not, depending on the number of groups. We also consider weighted penalties, following the ideas of the adaptive Lasso [Zou, 2006]. The selection of the two tuning parameters introduced in the two penalties is realized using the BIC criterion.

The paper is organized as follows. Section 2 introduces NLMEM and the SAEM algorithm. In Section 3 we introduce the fused lasso penalty, the penalized SAEM algorithm and the tuning parameter selection. In Section 4, the penalized-SAEM algorithm is evaluated on simulated data with 2 groups or more. Finally, it is applied to real data from a cross over clinical trials studying drug-drug interaction between dabigatran etexilate and 3 other drugs in Section 5.

2. Joint estimation of multiple non linear mixed effects models

2.1. Group structured NLMEMs

Let $y_{i,j}^g$ be the observation at time $t_{i,j}^g$ ($j \in \{1,\dots,n_i^g\}$) for the i-th patient ($i \in \{1,\dots,N_g\}$) in the g-th group ($g \in \{1,\dots,G\}$). We consider models of the form:

$$\begin{aligned} y_{i,j}^g &= f(t_{i,j}^g, \phi_i^g) + d(t_{i,j}^g, \phi_i^g)\epsilon_{i,j}^g \\ \epsilon_{i,j}^g &\sim \mathcal{N}(0,1) \text{ (iid)} \end{aligned}$$

where f and d are two given non linear functions. The function d corresponds to the error model and is assumed to be $d = af + b$ with a and b two real constants. Measurement errors $\epsilon_{i,j}^g$ are further assumed to be independent and identically distributed. Individual parameters ϕ_i^g for the i-th subject in group g is a p-dimensional random vector, independent of $\epsilon_{i,j}^g$ and assumed to be decomposable (up to a transformation h) as:

$$\begin{aligned} h(\phi_i^g) &= \mu^g + b_i^g \\ b_i^g &\sim \mathcal{N}(0, \Omega^g) \text{ (iid)} \end{aligned}$$

where $\mu^g \in \mathbb{R}^p$ is the mean parameter vector for the group g and $b_i^g \in \mathbb{R}^p$ the random effects of the i-th patient. Different transformations h can be used. Here we use the common one $h(x) = \log(x)$, that yields log-normally distributed ϕ_i^g. In this work, Ω^g is supposed diagonal as explained in section 3.1.2.

The log-likelihood then takes the form:

$$LL(\theta) = \log p(y;\theta) = \log\left(\sum_{g=1}^{G} \int p(y^g, \phi^g; \theta^g) d\phi^g\right) \tag{1}$$

where $p(y^g, \phi^g; \theta^g)$ is the likelihood of the complete data in group g:

$$\begin{aligned} \log p(y^g, \phi^g; \theta^g) = &- \sum_{i,j} \log(d(t_{i,j}^g, \phi_i^g)) - \frac{1}{2}\sum_{i,j}\left(\frac{y_{ij} - f(t_{i,j}^g, \phi_i^g)}{d(t_{i,j}^g, \phi_i^g)}\right)^2 - \frac{N_g}{2}\log(|\Omega^g|) \\ &-\frac{1}{2}\sum_i (\phi_i^g - \mu^g)^t {\Omega^g}^{-1}(\phi_i^g - \mu^g) - \frac{\sum_i n_i^g + N_g p}{2}\log(2\pi) \end{aligned}$$

with $\theta = (\theta^1, \dots, \theta^G)$ and $\theta^g = (\mu^g, \Omega^g, a, b)$ the parameters to be estimated. Note that the log-likelihood $LL(\theta)$ as defined in Equation (1) has generally no closed form expression because of the non linearity of f with respect to ϕ.

2.2. SAEM algorithm for the joint estimation problem

In this section, we present a standard version of the SAEM algorithm in the context of joint estimation that will be the base of the penalized version that we introduce later. Here, we do not account for potential similarities of the parameters across groups.

The SAEM algorithm is a classical tool for parameter estimation of NLMEMs [Delyon et al., 1999]. It iteratively maximizes the conditional expectation of the complete data log-likelihood. At iteration k given the current estimates θ_{k-1}, the problem reduces to the optimization of the following criterion:

$$Q_k(\theta) = \sum_{g=1}^{G} Q_k(\theta^g) = \sum_{g=1}^{G} \mathbb{E}\Big(\log p(y^g, \phi^g; \theta^g) \mid y^g, \theta^g_{k-1} \Big).$$

As this conditional expectation has no closed form for NLMEMs, it is approximated using a stochastic approximation scheme. The E-step of the classical EM algorithm is then divided in two parts, a simulation step where individual parameters are simulated using a Markov Chain Monte Carlo method (MCMC) and a stochastic approximation step [Kuhn and Lavielle, 2005]. At iteration k of the SAEM algorithm we have:

1. Estimation step (E-step):
   (a) Simulation step: draw ϕ^g_k using an MCMC procedure targeting $p(.|y^g, \theta^g_{k-1})$.
   (b) Stochastic approximation step of $Q_k(\theta)$: update $\tilde{Q}_k(\theta)$ using the following scheme

$$\tilde{Q}^g_k(\theta^g) = \tilde{Q}^g_{k-1}(\theta^g) + \gamma_k(\log\ p(y^g, \phi^g_k; \theta^g) - \tilde{Q}^g_{k-1}(\theta^g))$$

   where γ_k is a decreasing sequence of positive numbers. When the complete data likelihood belongs to the exponential family, this step simply reduces to the stochastic approximation of its sufficient statistics $s^g_{1,i,k}$, $s^g_{2,k}$ and $s^g_{3,k}$:

$$\begin{aligned}
s^g_{1,i,k} &= s^g_{1,i,k-1} + \gamma_k \left(\sum_{i=1}^{N_g} \phi^g_{i,k} - s^g_{1,i,k-1} \right) \\
s^g_{2,k} &= s^g_{2,k-1} + \gamma_k \left(\sum_{i=1}^{N_g} \phi^g_{i,k} \phi^{g^t}_{i,k} - s^g_{2,k-1} \right) \\
s^g_{3,k} &= \begin{cases} s^g_{3,k-1} + \gamma_k \left(\sum_{i,j} \left(y^g_{i,j} - f(t^g_{i,j}, \phi^g_{i,k}) \right)^2 - s^g_{3,k-1} \right) \text{ if } b = 0 \\ s^g_{3,k-1} + \gamma_k \left(\sum_{i,j} \left(\frac{y^g_{i,j} - f(t^g_{i,j}, \phi^g_{i,k})}{d(t^g_{i,j}, \phi^g_{i,k})} \right)^2 - s^g_{3,k-1} \right) \text{ if } a = 0 \end{cases}
\end{aligned}$$

2. Maximisation step (M-step): update of population parameters:

$$\theta_k = \underset{\theta}{\text{ArgMax}}\ \tilde{Q}_k(\theta).$$

   Within the exponential family, the solution is explicit for μ^g and Ω^g:

$$\mu^g_k = \frac{1}{N_g} \sum_{i=1}^{N_g} s^g_{1,i,k}\ , \quad \Omega^g_k = \frac{1}{N_g} \left(s^g_{2,k} - \sum_{i=1}^{N_g} \mu^g_k s^{g^t}_{1,i,k} - \sum_{i=1}^{N_g} s^g_{1,i,k} \mu^{g^t}_k \right) + \mu^g_k \mu^{g^t}_k .$$

For parameters a and b, they are updated using the whole data set because they are common to all the groups. An explicit solution exists when $a = 0$ or $b = 0$:

$$a = 0 \Rightarrow b_k = \sqrt{\frac{\sum_{g=1}^{G} s_{3,k}^{g}}{\sum_{g=1}^{G} \sum_i n_i}}$$

$$b = 0 \Rightarrow a_k = \sqrt{\frac{\sum_{g=1}^{G} s_{3,k}^{g}}{\sum_{g=1}^{G} \sum_i n_i}}.$$

When $a \neq 0$ and $b \neq 0$, the maximization problem has to be solved numerically.

Thus, except for a and b, SAEM algorithm for the joint estimation problem is implemented as if the G groups were analyzed separately.

3. Penalized joint estimation of group structured NLMEM

The previous SAEM algorithm corresponds to parameters estimated within each group. But groups can be expected to share common characteristics, so that theoretical parameters are expected to exhibit similarities. Therefore, we introduce a penalty within the SAEM algorithm that encourages parameters to be equal. We first detail the fused penalties, then the penalized SAEM algorithm and finally the selection of the two tuning parameters introduced in the penalties.

3.1. Fused lasso penalties for group comparison

The fused lasso penalty encourages parameters to have the same value between two groups. This is particularly useful when theoretical parameters of (at least some of) the groups are expected to be similar and/or when the objective of the study is to assess potential differences between groups. Depending on the context, differences between all the groups or only some specific differences might be of interest. Likewise, similarity of some parameters does not necessarily hold for all the groups. These differences and similarities of interest can be described with a graph that links groups together. Two groups are related in the graph if the comparison of these two groups is of interest, or if parameters are assumed to be similar in these two groups. Of course, any graph structure can be put forward, but some of them are naturally appealing in various contexts (see figure 1 with $G = 4$):

- Clique Graph: no assumptions on the hierarchical structure of the groups are made. All the possible differences between group parameters are penalized.
- Star Graph: a reference group is stated and only the differences between the reference group and the others are penalized. This is equivalent to the classical approach based on group affiliation covariate.
- Chain Graph: when groups can be naturally ordered.

Given a specific graph described by its edge set $\mathcal{E}$, we introduce the penalties for the fixed and the variance parameters.

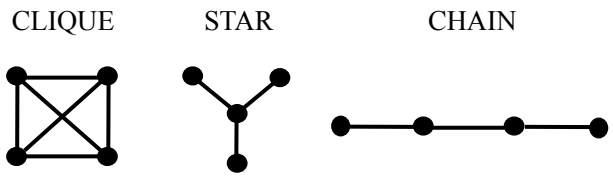

Figure 1: Examples of graphs for $G = 4$ groups

3.1.1. Fixed parameters

For fixed parameters $(\mu^1, \ldots, \mu^G)$, the fused lasso penalty corresponds to:

$$P_F(\mu^1, \ldots, \mu^G) = \sum_{(g_1,g_2)\in\mathcal{E}} \|\mu^{g_1} - \mu^{g_2}\|_1$$

where $\|x\|_1 = \sum_i |x_i|$ is the l_1-norm. The fused lasso penalty encourages the fixed parameters to be the same between two groups connected in the graph.

3.1.2. Variance parameters

Concerning random effect variances, components of the variance covariance matrix should be penalized. However, the resulting optimization problem is not convex. Some algorithms have been proposed to solve simple l_1 penalty problems [Bien and Tibshirani, 2011, Wang, 2013] but they are computationally demanding and their extension to the fused penalty context is not straightforward. However under the assumptions that Ω^g is diagonal, a simple alternative consists in penalizing the inverse variance covariance matrix. Then the penalized optimization problem becomes convex and can be solved efficiently. This corresponds to the following penalty:

$$P_V(\Omega^{1^{-1}}, \ldots \Omega^{G^{-1}}) = \sum_{(g_1,g_2)\in\mathcal{E}} \|\Omega^{g_1^{-1}} - \Omega^{g_2^{-1}}\|_1 = \sum_{i=1}^{p} \sum_{(g_1,g_2)\in\mathcal{E}} |\Omega_{ii}^{g_1^{-1}} - \Omega_{ii}^{g_2^{-1}}|.$$

Penalizing differences between $\Omega^{g^{-1}}$ is not equivalent to penalizing differences between Ω^g as $|\Omega_{ii}^{g_1^{-1}} - \Omega_{ii}^{g_2^{-1}}| \neq |\Omega_{ii}^{g_1} - \Omega_{ii}^{g_2}|$. Some issues could occur when considering parameters with very different levels of variability: the penalty rapidly discards differences for parameters with low variance. This issue is mitigated when working with log-normally distributed individual parameters, and adaptive weights can further help to prevent such a behavior (see section 4.2).

3.1.3. Matricial formulation and adaptive weights

Weights (π,ν) can be introduced in order to take into account some prior information:

$$P_F(\mu^1, \ldots, \mu^G) = \sum_{(g_1,g_2)\in\mathcal{E}} \sum_{i=1}^{p} \pi_i^{g_1 g_2} |\mu_i^{g_1} - \mu_i^{g_2}|$$

$$P_V(\Omega^{1^{-1}}, \ldots \Omega^{G^{-1}}) = \sum_{(g_1,g_2)\in\mathcal{E}} \sum_{i=1}^{p} \nu_i^{g_1 g_2} |\Omega_{ii}^{g_1^{-1}} - \Omega_{ii}^{g_2^{-1}}|.$$

These weights can be based on initial maximum likelihood estimates within each group ($\tilde{\mu}^g$,$\tilde{\Omega}^g$) following the idea of the adaptive fused lasso [Viallon et al., 2014]: $\pi_i^{g_1 g_2} = |\tilde{\mu}_i^{g_1} - \tilde{\mu}_i^{g_2}|^{-\alpha}$ and $\nu_i^{g_1 g_2} = |\tilde{\Omega}_{ii}^{g_1^{-1}} - \tilde{\Omega}_{ii}^{g_2^{-1}}|^{-\alpha}$ for some $\alpha > 0$ (typically $\alpha = 1$). These weighted penalties are particularly helpful to compute unpenalized re-estimation of the selected model (section 3.3).

Finally, the weighted penalties with weights π and ν can be written in a matricial form:

$$P_F(\mu^1, \ldots, \mu^G) = \|\pi \bullet P\mu\|_1$$
$$P_V(\Omega^{1^{-1}}, \ldots \Omega^{G^{-1}}) = \|\nu \bullet P\text{diag}(\Omega^{-1})\|_1$$

where the matrix $P \in \{-1, 0, 1\}^{|\mathcal{E}|\times Gp}$ encodes the penalized values of $\mu = (\mu^1, \ldots, \mu^G)^t$ and $\text{diag}(\Omega^{-1}) = (\text{diag}(\Omega^{1^{-1}}), \ldots, \text{diag}(\Omega^{G^{-1}}))^t$ and $\bullet$ stands for the Hadamard product.

3.2. Penalized version of SAEM algorithm

The penalized SAEM algorithm consists in iteratively maximizing the penalized stochastic approximation of the conditional expectation $Q_k(\theta)$:

$$\tilde{Q}_k(\theta) - \lambda_F P_F(\mu^1, \ldots, \mu^G) - \lambda_V P_V(\Omega^{1^{-1}}, \ldots \Omega^{G^{-1}})$$

where λ_F and λ_V are two tuning parameters that control the penalty strength and that have to be calibrated. When $\lambda_F = \lambda_V = 0$, the estimates correspond to the classical maximum likelihood estimates. For large enough values, the vector of penalized differences is set to zero ($P\mu = 0$ and/or $P\text{diag}(\Omega^{-1}) = 0$). The penalized SAEM is the standard SAEM except for the M-step: a fused lasso penalized regression problem is solved for both fixed effects and random effects variances updates, with fixed tuning parameters λ_F and λ_V. At iteration k, it corresponds to (Box 1):

Box 1: Maximization step of the penalized SAEM algorithm

1. Fixed effects update:

$$\left(\mu_k^1, \ldots, \mu_k^G\right) = \underset{\mu^1,\ldots,\mu^G}{\text{ArgMax}} \left(\sum_{g=1}^{G} \tilde{Q}_k(\mu^g, \Omega_{k-1}^g, a_{k-1}, b_{k-1}) \; - \; \lambda_F P_F(\mu^1, \ldots, \mu^G) \right)$$

2. Random effects variances update:

$$\Omega_k^1, \ldots, \Omega_k^G) = \underset{\Omega^1,\ldots\Omega^G}{\text{ArgMax}} \left(\sum_{g=1}^{G} \tilde{Q}_k(\mu_k^g, \Omega^g, a_{k-1}, b_{k-1}) \; - \; \lambda_V P_V(\Omega^{1^{-1}}, \ldots, \Omega^{G^{-1}}) \right)$$

3. Error model parameters update: usual update.

We now turn to the description of the two update steps for the fixed effects and the random effects variances.

3.2.1. Fixed effects update

For fixed effects update, the conditional expectation of the complete likelihood reduces to the following weighted least square function:

$$\tilde{Q}_k(\mu) = \sum_{g=1}^{G} \tilde{Q}_k(\mu^g, \Omega_{k-1}^g, a_{k-1}, b_{k-1}) = C - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\left(-\mu^{g^t}\Omega_{k-1}^{g^{-1}}s_{1,i,k}^g - s_{1,i,k}^{g^t}\Omega_{k-1}^{g^{-1}}\mu^g + \mu^{g^t}\Omega_{k-1}^{g^{-1}}\mu^g\right)$$

where C is a constant not depending on μ^g. The matricial form of the problem to be solved is:

$$\left(\mu_k^1, \ldots, \mu_k^G\right) \quad = \quad \underset{\mu}{\text{ArgMax}}\ \tilde{Q}_k(\mu) - \lambda_F \|P\mu\|_1. \tag{2}$$

This optimization problem corresponds to an extension of the generalized fused lasso of Höfling et al. [2010] where the least-squares is replaced by weighted least-squares. It can be solved with the Alternating Direction Method of Multipliers (ADMM) [Boyd et al., 2011], that breaks the convex optimization problem into small pieces. More precisely, problem (2) can be rewritten as an equality constraints optimization problem, where μ is split in two parts μ and z:

$$\begin{aligned} \hat{\mu} = \underset{\mu,z}{\text{ArgMin}} \quad & -\tilde{Q}_k(\mu) + \lambda_F\|z\|_1. \\ \text{s.t } & P\mu - z = 0 \end{aligned}$$

The ADMM algorithm solves (2) by iteratively solving smaller (and easier) problems for each primal (μ, z) and dual (u) variables separately using the augmented Lagrangian formulation:

$$\underset{\mu,z}{\text{ArgMin}}\ \underset{u}{\text{ArgMax}}\ -\tilde{Q}_k(\mu) + \lambda_F\|z\|_1 + \langle u, P\mu - z\rangle + \frac{\rho}{2}\|P\mu - z + u\|_2^2..$$

Here ρ is the augmented lagrangian parameter (generally set to 1) and $\|\cdot\|_2$ the l_2-norm. This corresponds to applying the steps presented in Box 2 at each iteration q until convergence. When adaptive weights are included in the penalty, the same algorithm can be used except that the tuning parameter λ_F is replaced by the vector $\lambda_F \bullet \pi$.

3.2.2. Variance Covariance matrix of random effects update

The conditional expectation of the complete likelihood for group g is:

$$\tilde{Q}_k(\mu_k^g, \Omega^g, a_{k-1}, b_{k-1}) = C - \log|\Omega^g| - \text{Trace}\left[\Omega^{g^{-1}}\tilde{\Sigma}_k^g\right]$$

where C is a constant not depending on Ω^g and $\tilde{\Sigma}_k^g$ corresponds to the solution of the non penalized problem:

$$\tilde{\Sigma}_k^g = \frac{1}{N_g}\left(s_{2,k}^g - \sum_{i=1}^{N_g}\mu_k^g s_{1,i,k}^{g^t} - \sum_{i=1}^{N_g} s_{1,i,k}^g \mu_k^{g^t}\right) + \mu_k^g\mu_k^{g^t}.$$

Then the problem to be solved is:

$$(\Omega_k^1, \ldots \Omega_k^G) = \underset{\Omega^1,\ldots\Omega^G}{\text{ArgMax}}\left(-\sum_{g=1}^{G}\left(\log|\Omega^g| + \text{Trace}\left[\Omega^{g^{-1}}\tilde{\Sigma}_k^g\right]\right) - \lambda_V P_V(\Omega^{1^{-1}}, \ldots, \Omega^{G^{-1}})\right). \tag{3}$$

Box 2: ADMM algorithm for fixed effects update

1. Initialization: $\mu_0 = \mu_{k-1}$, $z_0 = 0$, $u_0 = 0$
2. For $q = 0, 1, 2, ...$ until convergence:
   (a) μ update:

$$\mu_{q+1} = \underset{\mu}{\text{ArgMin}} \left(-\tilde{Q}_k(\mu) + \frac{\rho}{2} ||P\mu - z_q + u_q||_2^2 \right) = (\Delta + \rho P^t P)^{-1} (\Gamma + \rho P^t (z_q - u_q))$$

   with $\Gamma = \text{diag}(\sum_{i=1}^{N_1} \Omega_{k-1}^{1^{-1}} s_{1,i,k}^1, \ldots, \sum_{i=1}^{N_G} \Omega_{k-1}^{G^{-1}} s_{1,i,k}^G)$
   and $\Delta = \text{diag}(N_1 \Omega_{k-1}^{1^{-1}}, \ldots, N_G \Omega_{k-1}^{G^{-1}})$
   (b) z update:

$$z_{q+1} = \underset{z}{\text{ArgMin}} \left(\frac{\rho}{2} ||P\mu_{q+1} + u_q - z||_2^2 + \lambda_F ||z||_1 \right) = \mathcal{S}_{\frac{\lambda_F}{\rho}} (P\mu_{q+1} + u_q)$$

   with the soft thresholding operator $\mathcal{S}_\lambda(x) = sgn(x)(|x| - \lambda)_+$
   (c) dual update:

$$u_{q+1} = u_q + P\mu_{q+1} - z_{q+1}$$

Danaher et al. [2013] consider a similar optimization problem (for joint graphical models) and propose an ADMM algorithm to solve it. We apply the same methodology here: problem (3) has the following scaled augmented Lagrangian:

$$\underset{\Omega^{g^{-1}}, Z^g}{\text{ArgMin}} \; \underset{U^g}{\text{ArgMax}} \begin{cases} \sum_{g=1}^G \left(\log |\Omega^g| + \text{Trace} \left[\Omega^{g^{-1}} \tilde{\Sigma}_k^g \right] \right) + \lambda_V P_V(Z^1, \ldots, Z^G) \\ + \sum_{g=1}^G \frac{\rho}{2} ||\Omega^{g^{-1}} - Z^g + U^g||_F^2 \end{cases}$$

where $(\Omega^{g^{-1}})_{g=1,\ldots,G}$,$(Z^g)_{g=1,\ldots,G}$ are the primal variables, $(U^g)_{g=1,\ldots,G}$ the dual variables and ρ the augmented lagrangian parameter (generally set to 1). The ADMM algorithm consists in applying the steps presented in Box 3 at each iteration q until convergence. Step 1 has an explicit solution [Witten and Tibshirani, 2009]. Step 2 is the evaluation of the P_V's proximal operator. An explicit formula is available when $G = 2$ [Danaher et al., 2013], but for $G > 2$ it has to be numerically approximated. This extends the computing time significantly. As for fixed effects, when adaptive weights are included in the penalty, the same algorithm can be used except that the tuning parameter λ_V is replaced by the vector $\lambda_V \bullet \nu$.

3.3. Selection of the tuning parameters

The described SAEM algorithm is applied with a fixed value of the tuning parameters $\Lambda = (\lambda_F, \lambda_V)$. The value of these tuning parameters varying from zero to infinity, the SAEM algorithm selects a collection of models with a decreasing number of between group differences (from the full model to the model with no difference at all). The optimal Λ can be selected using the Bayesian Information Criterion (BIC): the optimal Λ is defined as returning the model with the

Box 3: ADMM algorithm for variances update

1. Initialization: $\Omega_0^g = \tilde{\Sigma}_k^g$, $Z_0^g = 0$, $U_0^g = 0$
2. For $q = 0, 1, 2, ...$ until convergence:
   (a) Ω update: for $g = 1, \ldots, G$

$$\Omega_{q+1}^{g^{-1}} = \underset{\Omega^{g^{-1}}}{\text{ArgMin}} \left(\log|\Omega^g| + \text{Trace}(\tilde{\Sigma}_k^g \Omega^{g^{-1}}) \right) + \frac{\rho}{2} \|\Omega^{g^{-1}} - Z_q + U_q\|_2^2$$

   (b) Z update:

$$Z_{q+1}^1, \ldots, Z_{q+1}^G = \underset{Z}{\text{ArgMin}} \left(\sum_{g=1}^{G} \frac{\rho}{2} \|\Omega_{q+1}^{g^{-1}} - Z^g + U_q^g\|_F^2 + \lambda_V P_V(Z^1, \ldots, Z^G) \right)$$

   (c) dual update: for $g = 1, \ldots, G$

$$U_{q+1}^g = U_q^g + \Omega_{q+1}^{g^{-1}} - Z_{q+1}^g$$

minimal BIC. In practice, we run the algorithm on a user-defined grid $(\Lambda_1, \ldots, \Lambda_M)$. Then the optimal value Λ_{BIC} is:

$$\Lambda_{BIC} = \underset{\Lambda \in \{\Lambda_1, \ldots, \Lambda_M\}}{\text{ArgMin}} \; BIC(\Lambda)$$

where $BIC(\Lambda)$ is the criterion of the model corresponding to the value Λ. For a NLMEM with random effects on all the parameters, the BIC criterion is defined as [Delattre et al., 2014]:

$$BIC = -2LL(\theta) + \log(N) \times df(\theta),$$

where $LL(\theta)$ is the log likelihood (1), $df(\theta)$, the degree of freedom, is the number of distinct fixed effects and random effects variances in the selected model. For a given Λ, the penalized SAEM algorithm estimates a model (θ_Λ) with a particular structure: some parameters have the same estimated value (their difference is set to 0). Following the Lars-OLS-Hybrid algorithm [Efron et al., 2004] (that corresponds to a relaxed lasso [Meinshausen, 2007] with relaxing parameter set to 0), an unbiased estimation $\tilde{\theta}_\Lambda$ of the parameters of this selected model is obtained by reestimating θ with a constrained SAEM algorithm:

$$\tilde{\theta}_\Lambda = \underset{\theta}{\text{ArgMin}} - 2LL(\theta)$$
$$\text{s.t } S\begin{pmatrix} P\mu \\ P \,\text{diag}\, \Omega \end{pmatrix} = S\begin{pmatrix} P\hat{\mu}_\Lambda \\ P \,\text{diag}\, \hat{\Omega}_\Lambda \end{pmatrix}$$

where $S(x)$ is the support of vector x. The constraint on the support ensures that the solution of the constrained optimization problem has the same structure as the solution of the penalized estimate. This constrained optimization problem can be solved by the penalized SAEM algorithm with appropriate choices for the adaptive weights: non-null differences are attached to null

weights (and are therefore not penalized) and null differences are attached to weights that are high enough to force them to be null in the solution $\tilde{\theta}_\Lambda$. Finally, we take:

$$BIC(\Lambda) = -2LL(\tilde{\theta}_\Lambda) + \log(N) \times df(\tilde{\theta}_\Lambda).$$

4. Simulated data analysis

Simulations are performed with the one compartment model with first order absorption:

$$f(t, k_a, Cl, V) = \frac{Dk_a}{Vk_a - Cl}(e^{-\frac{Cl}{V}t} - e^{-k_a t}) \tag{4}$$

where k_a (h^{-1}), Cl ($L.h^{-1}$) and V (L) correspond respectively to the absorption constant, the clearance and the volume of distribution parameters. The administrated dose (D) is set to 6 mg.

First, the behavior of the penalized SAEM algorithm is illustrated with one dataset simulated with 3 groups of subjects. Especially, regularization paths depending on the values of the sparsity parameter Λ is presented. Then the impact of high variances on the penalized estimation is studied on one dataset simulated with 3 groups of subjects, and the benefit of adaptive weights introduced in the penalty is shown. Next the influence of the penalty structure on selection performance is studied on 100 simulated data sets with 5 groups of subjects. Finally, joint fixed and variance parameters selection performance is compared between the penalized SAEM algorithm and the standard stepwise forward approach, on 50 simulated data sets with 2 groups of subjects.

4.1. Behavior of the penalized SAEM algorithm

One dataset of 3 groups with $N_g = 100$ subjects per group has been simulated using model (4) and fixed effects parameters:

$$\begin{aligned}
&\mu_V^1 = \mu_V^2 = 0.48 \quad \mu_V^3 = 0.58 \\
&\mu_{Cl}^1 = \mu_{Cl}^2 = 0.06 \quad \mu_{Cl}^3 = 0.042 \\
&\mu_{k_a}^1 = \mu_{k_a}^3 = 1.47 \quad \mu_{k_a}^2 = 2.18.
\end{aligned}$$

Random effects variances are:

$$\begin{aligned}
&(\omega_V^1)^2 = (\omega_V^2)^2 = (\omega_V^3)^2 = 0.1 \\
&(\omega_{Cl}^1)^2 = (\omega_{Cl}^2)^2 = 0.1 \quad (\omega_{Cl}^3)^2 = 0.2 \\
&(\omega_{k_a}^1)^2 = 0.1 \quad (\omega_{k_a}^2)^2 = 0.3 \quad (\omega_{k_a}^3)^2 = 0.2.
\end{aligned}$$

Individual parameters are log-normally distributed ($h(\phi) = \log(\phi)$). Error model parameters are set to $a = 0$ and $b = 0.1$. The penalised SAEM algorithm is implemented with 400 iterations. During the first 300 iterations, we use a constant step size equal to 1. Then, during the last 100 iterations, the stochastic approximation scheme is implemented with a step size equal to $\frac{1}{iter-300}$ at iteration *iter*. The evolution of each SAEM parameter estimate is plotted along iterations in Figure 2 for $\lambda_F = 37$ and $\lambda_V = 0.015$ using a clique graph as penalty structure. In this example, the number of iterations has been chosen such that the convergence is clearly attained for all the model parameters. For these values of λ_F and λ_V, the model selected by the algorithm corresponds to the simulated one. Figure 3 represents the regularization path of the estimates for both fixed effects and variances of random effects parameters. When increasing λ_F (or λ_V)

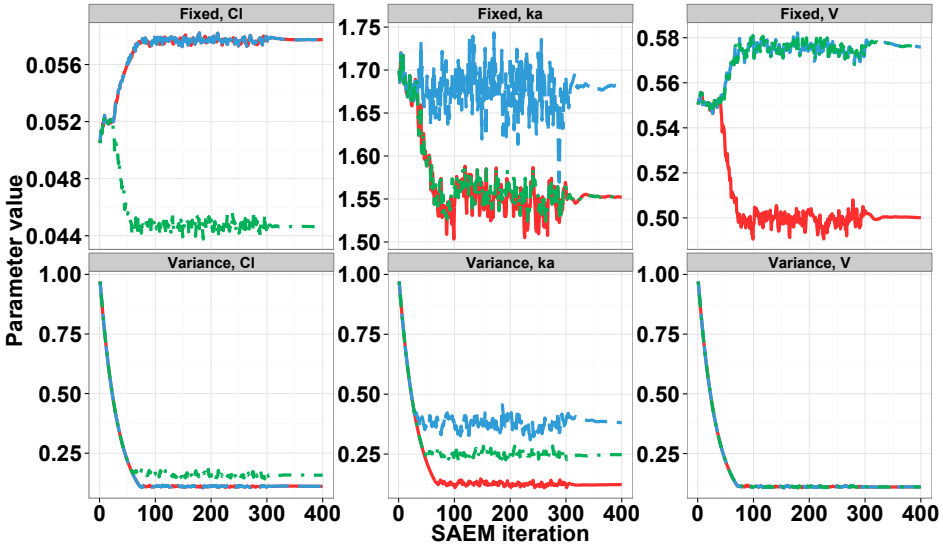

Figure 2: Simulated data, 3 groups: evolution of SAEM estimates with $\lambda_F = 25$ and $\lambda_V = 0.013$. Red, blue and green curves correspond to estimates of group 1, 2 and 3, respectively.

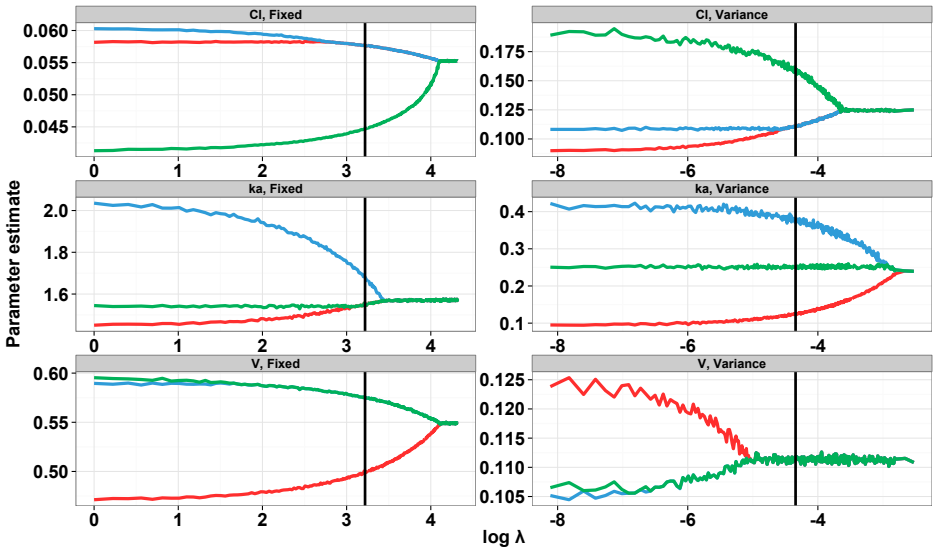

Figure 3: Simulated data, 3 groups: regularization paths of SAEM estimates for fixed effects and random effect variances. Red, blue and green curves correspond to estimates of group 1, 2 and 3, respectively. Solid black lines corresponds to the lambda values used in Figure 2 ($\lambda_F = 25, \lambda_V = 0.013$).

	$G = 2$	$G = 3$	$G = 5$
Fixed	23 s	24 s	99 s
Fixed + Variances	32 s	210 s	411 s

Table 1: Simulated data, 2, 3 or 5 groups: computational time of the penalized SAEM algorithm (400 iterations) with small tuning parameter values on a simulated data set of $N_g = 100$ subjects per group ($G = 2$, 3 or 5) with or without penalty on variance parameters. A clique graph is used for the penalty.

values, differences between estimates get smaller and smaller until being null. The number of null differences increases with the value of λ.

As we have seen in the algorithm description, when $G > 2$ the proximal operation of the variances penalty needs to be approximated numerically. The computational time is then increased when variances are penalized. Table 1 presents the computational times for different numbers of groups when variances are penalized or not. These computational times vary in function of the values of λ_F and λ_V; Table 1 corresponds to a worst-case scenario (small values of λ_F and λ_V).

4.2. Effect of variances rescaling with adaptive weights

As discussed in Section 3.1.2, penalizing the inverse of covariance matrix is not equivalent to penalizing directly the variances. It could favor differences from parameters with a high variance and then select false models. This can be attenuated by rescaling the variances with adaptive weights. We propose the following weighting strategy:

$$P_V(\Omega^{1^{-1}}, \ldots, \Omega^{G^{-1}}) = \sum_{(g_1,g_2)\in\mathcal{E}} \nu_i \sum_{i=1}^{p} |\Omega_{ii}^{g_1^{-1}} - \Omega_{ii}^{g_2^{-1}}|, \quad \nu_i = \sqrt{\sum_{g=1}^{G} (\tilde{\Omega}_{ii}^{g})^2}$$

where $\tilde{\Omega}^g$ corresponds to the unpenalized estimation of Ω^g. To illustrate this, a data set of 3 groups (100 subjects per group) is simulated using model (4) with larger ω_V and smaller ω_{ka}:

$$\begin{aligned}
&(\omega_V^1)^2 = (\omega_V^2)^2 = (\omega_V^3)^2 = 0.3\\
&(\omega_{Cl}^1)^2 = (\omega_{Cl}^2)^2 = 0.1, \quad (\omega_{Cl}^3)^2 = 0.2\\
&(\omega_{k_a}^1)^2 = 0.03, \quad (\omega_{k_a}^2)^2 = 0.075, \quad (\omega_{k_a}^2)^2 = 0.06.
\end{aligned}$$

Figure 4 presents the regularization path of estimates for $(\omega_{k_a}^g)^2, (\omega_V^g)^2, (\omega_{Cl}^g)^2$ using a clique structure for the penalty. Because the $(\omega_V^g)^2$ terms are all equal, we would hope estimates of these terms to be fused before the $(\omega_{k_a}^g)^2$ and $(\omega_{Cl}^g)^2$ terms. This is not the case without adaptive weights, and as a consequence, the optimal model is not spanned in the regularization path. Adaptive weights correct for this defect and the optimal model is spanned by the regularization path (blue shaded areas in Figure 4).

4.3. Model Selection Performances

4.3.1. Influence of penalty structure and adaptive weights

We study the selection of fixed effects differences between groups on simulated data sets and the impact of the penalty structure on the proportion of correctly selected models. One hundred

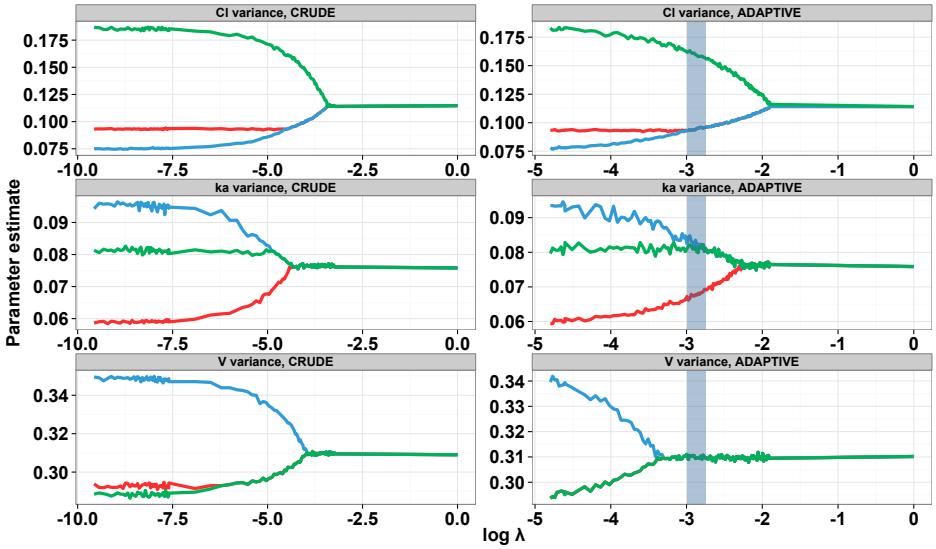

Figure 4: Simulated data, 3 groups, large ω_V, small ω_{ka}: regularisation paths of SAEM estimates for random effect variances with (ADAPTIVE) or without (CRUDE) adaptive weights. Red, blue and green curves correspond respectively to the estimates in group 1, 2 and 3. Blue shaded areas correspond to the values Λ that return the simulated differences model.

datasets are simulated with 5 groups of subjects using model (4) ($N_g = 20$ or $N_g = 100$). Fixed effects parameters are set to:

$$\mu_V^1 = 0.48 \quad \mu_V^2 = \mu_V^3 = 0.72 \quad \mu_V^4 = \mu_V^5 = 0.96$$
$$\mu_{Cl}^1 = \mu_{Cl}^2 = 0.06 \quad \mu_{Cl}^3 = \mu_{Cl}^4 = 0.03 \quad \mu_{Cl}^5 = 0.015$$
$$\mu_{k_a}^1 = \mu_{k_a}^2 = \mu_{k_a}^3 = \mu_{k_a}^4 = \mu_{k_a}^5 = 1.47.$$

Random effects variances are set equal to 0.1 for all the parameters. Individual parameters are log-normally distributed. Error model parameters are set to $a = 0$ and $b = 0.1$. For each data set, a model is selected using the fused lasso approach on a grid of 100 λ_F values with 4 different penalty structures:

- CH: chain graph (with adaptive weights or not)
- CL: clique graph (with adaptive weights or not)
- S_1: star graph with group 1 as reference (with adaptive weights or not)
- S_3: star graph with group 3 as reference (without adaptive weights)

Note that the optimal graph would penalize only the null differences that appear in the simulated model. Thus none of these graphs is optimal for all the parameters. The optimal structure for parameter μ_{k_a} is a clique structure because its value is the same for all the groups. The most appropriate structure for parameters μ_{Cl} and μ_V is the chain structure that penalizes all

	$N_g = 20$ subjects per group				$N_g = 100$ subjects per group			
	CH	CH_A	CL	CL_A	CH	CH_A	CL	CL_A
Whole	15%	39%	8%	32%	25%	59%	28%	55%
V	53%	71%	33%	56%	54%	80%	52%	81%
Cl	41%	86%	29%	69%	55%	78%	54%	70%
k_a	61%	68%	64%	81%	77%	75%	77%	87%

Table 2: Simulated data, 5 groups: Proportion of correctly selected model over 100 simulations for the whole fixed effects model and fixed effects model of V, Cl or k_a. Different penalty structures: clique (CL), adaptive clique (CL_A), chain (CH) and adaptive chain (CH_A).

theoretically null differences (unlike star graph) and less non null differences than the clique graph. As previously suggested by [Viallon et al., 2014], adaptive weights should overcome the errors due to graph misspecification. Thus the penalized SAEM algorithm is implemented with the 4 different graphs with and without adaptive weights.

The performance of each penalty structure is evaluated by comparing the selected model ($P\tilde{\mu}_s$) to the true model ($P\mu$) for each parameter, on each simulated data set ($s = 1, ..., 100$) with $\tilde{\mu}_s$ the final estimate obtained by the fused lasso procedure and P a penalty matrix that encodes the differences under study. For example, when $P = P_{CH}$, P corresponds to the $(G-1) \times G$ matrix with $P_{i,i} = 1$, $P_{i,i+1} = -1$ and 0 elsewhere. When considering the whole fixed effect model, the number of correctly selected model is:

$$\frac{1}{100} \sum_{s=1}^{100} \mathbb{1}_{P\tilde{\mu}_{V,s}=P\mu_V} \times \mathbb{1}_{P\tilde{\mu}_{Cl,s}=P\mu_{Cl}} \times \mathbb{1}_{P\tilde{\mu}_{k_a,s}=P\mu_{k_a}}.$$

Table 2 shows the results with $P = P_{CH}$ for CH and CL as they are the only two penalties that could select exactly the true model. When $N_g = 20$, the chain graph has better performance. When N_g is large, performances of chain and clique graphs are very close. In addition, using adaptive weights clearly improves performance: in particular, the clique-based approach performs similarly to the chain-based one with adaptive weighs, even for $N_g = 20$. Thus clique graph is a good candidate when there is no prior information on data structure. This results tends to confirm the asymptotic optimality result of the clique-based strategy with adaptive weights that was obtained for generalized linear models [Viallon et al., 2014].

Table 3 shows the results when P is a star graph ($P = P_{S_1}$ for S_1, $S_{1,A}$ and $P = P_{S_3}$ for S_3) that does not correspond to the real structure of the data: here only differences based on a star graph can be selected while theoretical parameters do not follow a star graph. When using a star graph penalty, the group of reference has a non negligible influence on the results. It is particularly true for μ_{Cl}. Indeed when group 1 is set as reference, theoretical values of μ^2_{Cl}, μ^3_{Cl}, μ^4_{Cl} and μ^5_{Cl} are distributed in an unbalanced way around μ^1_{Cl} (μ^3_{Cl}, μ^4_{Cl} and μ^5_{Cl} are lower than μ^1_{Cl}). The penalty unexpectedly tends first to fused μ^1_{Cl} with μ^3_{Cl}, μ^4_{Cl} and μ^5_{Cl}. The adaptive version of the penalties seems to mitigate this phenomenon when sample size is large ($N_g = 100$). This behavior is not observed when group 3 is set as reference, probably because theoretical parameters value of non reference groups are distributed in a more balanced way around μ^3_{Cl}.

	$N_g = 20$ subjects per group			$N_g = 100$ subjects per group		
	S_1	$S_{1,A}$	S_3	S_1	$S_{1,A}$	S_3
Whole	1%	6%	26%	8%	56%	49%
V	69%	72%	57%	100%	100%	90%
Cl	36%	87%	83%	29%	77%	93%
k_a	12%	23%	58%	28%	63%	60%

Table 3: Simulated data, 5 groups: Proportion of correctly selected model over 100 simulations when edges under study corresponds to a star graph. Results are given for the whole fixed effects model, fixed effects of V, Cl or k_a. Different penalty structures are considered: star with group 1 as reference (S_1), star with group 3 as reference (S_3) and adaptive star with group 1 as reference (S_1).

4.3.2. Fixed and variance parameters selection

Joint selection of fixed effects and random effects variances is evaluated through 50 simulated datasets using model (4) with only two groups for computational time reasons. Individual parameters are log-normally distributed. Error model parameters are set to $a = 0.2$ and $b = 0.02$. Fixed effects parameters are:

$$\begin{aligned}
&\mu_V^1 = 0.48 \quad \mu_V^2 = 0.58\\
&\mu_{Cl}^1 = 0.060 \quad \mu_{Cl}^2 = 0.042\\
&\mu_{k_a}^1 = \mu_{k_a}^2 = 1.47.
\end{aligned}$$

Random effects variances are:

$$\begin{aligned}
&\omega_V^{1^2} = \omega_V^{2^2} = 0.1\\
&\omega_{Cl}^{1^2} = 0.1 \quad \omega_{Cl}^{2^2} = 0.21\\
&\omega_{k_a}^{1^2} = 0.1 \quad \omega_{k_a}^{2^2} = 0.21.
\end{aligned}$$

For each data set, the best model is selected using BIC based on the penalized SAEM algorithm estimation. For comparison purpose, the selection approach based on a BIC forward stepwise method is also implemented. This stepwise method includes 2 steps: i) assuming the variances of random effects to be different between the groups, the fixed effect model is selected by BIC comparison, ii) using the selected fixed effects model, the variance model is selected by BIC comparison. The performance of the two methods is evaluated by comparing the selected model to the true model. Table 4 presents the proportion of correctly selected models for the fixed effects model, the variances model and the whole model. Table 5 presents the proportion on the 50 simulated data sets where a non null difference is detected for each parameters. On this synthetic example, our approach enjoys better selection performance than the stepwise approach. This is particularly true for variance parameters. However, Table 5 shows that the fused lasso approach tends also to select more complex models especially for small sample sizes. Indeed μ_{k_a} and $(\omega_V)^2$ are theoretically equal in the 2 groups, but the fused lasso estimates a non null difference on these two parameters. When the fused lasso approach does not select the true model, it generally includes differences that are null in the true model. This is especially true when $N_g = 25$.

	Fixed effects			Variances			Both		
N_g	25	50	100	25	50	100	25	50	100
Stepwise Forward	30%	76%	68%	10%	30%	52%	6%	24%	42%
Fused LASSO	40%	74%	76%	38%	56%	78%	14%	40%	60%

Table 4: Simulated data, 2 groups: proportion of correctly selected models on 50 simulated datasets for the fixed effects model, the variances model and the whole model. Results are given for the fused lasso and the stepwise forward approaches.

	$N_g = 25$		$N_g = 50$		$N_g = 100$	
	Fused	Forward	Fused	Forward	Fused	Forward
μ_V	76%	56%	94%	88%	100%	94%
μ_{Cl}	100%	74%	100%	96%	100%	88%
μ_{k_a}	40%	20%	20%	10%	24%	16%
ω^2_V	20%	14%	14%	6%	12%	20%
ω^2_{Cl}	64%	32%	88%	68%	96%	88%
$\omega^2_{k_a}$	72%	32%	78%	50%	92%	70%

Table 5: Simulated data, 2 groups: proportion of the 50 simulated datasets with a non null difference detected for each parameter of the model, respectively. Results are given for the fused lasso and the stepwise forward approaches.

5. Real data analysis

We now illustrate our approach on a real data example. *DE* is an orally anticoagulant drug used for the prevention of venous thromboembolism after orthopedic surgery and stroke in patients with atrial fibrillation. Its has a low bioavailability (fraction of administrated dose that reaches the systemic circulation) typically below 7%. It is mainly due to a solubility problem and to the P-glycoprotein (P-gp) efflux that has an "anti-absorption" function. P-gp inhibitors can increase Dabigatran bioavailability by improving its absorption [Delavenne et al., 2013]. Then, the addition of P-gp inhibitors could also lead to overdosing and adverse event like hemorrhage.

Data from two cross over clinical trials are considered. The two studies were conducted with two different dosing regimens for *DE*. The first trial, a two way crossover trial with 10 subjects, evaluated the interaction between *DE* (dosing regiment A) and P-Gp inhibithor 1 ($PgpI_1$). The second trial, an incomplete three way crossover trial with 9 subjects, evaluated the interaction

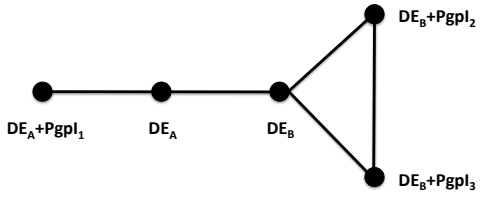

Figure 5: Graph used for the penalty of DE pooled data.

between DE (dosing regiment B), P-Gp inhibithor 2 ($PgpI_2$) and P-Gp inhibithor 3 ($PgpI_3$). Data from the two trials are pooled and five groups of subjects are defined:

- DE_A: DE with dosing regimen A alone (10 subjects).
- $DE_A + PgpI_1$: DE with dosing regimen A alone plus P-Gp inhibithor 1 (10 subjects).
- DE_B: DE with dosing regimen B (DE_B) alone (9 subjects).
- $DE_B + PgpI_2$: DE with dosing regimen B alone plus P-Gp inhibithor 2 (9 subjects).
- $DE_B + PgpI_3$: DE with dosing regimen B alone plus P-Gp inhibithor 3 (9 subjects).

In each group, dabigatran blood concentration (pharmacokinetic) is measured for each patient at 10 sampling times after oral drug administration. The following pharmacokinetic model with one compartment and an inverse gaussian absorption is used:

$$\frac{dA_c}{dt} = IG(t) - \frac{Cl}{V_c}A_c$$
$$IG(t) = Dose \times F \times \sqrt{\frac{MAT}{2\pi CV^2 t^3}} \times e^{\frac{-(t-MAT)^2}{2CV^2 MAT t}}$$

where A_c corresponds to the amount of dabigatran in the blood, and the absorption parameters F, MAT and CV correspond to bioavailability, mean absorption time and coefficient of variation of the absorption rate respectively. Finally parameters Cl and V_c are the elimination clearance and the volume of central compartment. Individual parameters are supposed log-normally distributed ($h(\phi) = \log(\phi)$).

Estimating the bioavailability with only data from orally administrated drug is an ill-posed problem. Indeed, a decreased value for F could be balanced by smaller V and Cl values. In order to regularize this problem, we add prior distributions on both V and Cl fixed parameters [Weiss et al., 2012] based on previously published results [Blech et al., 2008]. In this case, fixed parameters update is done by solving the following optimization problem:

$$\left(\mu_{k+1}^1, \ldots, \mu_{k+1}^G\right) = \underset{\mu}{\text{ArgMax}} \sum_{g=1}^{G} \tilde{Q}_k(\mu^g, \Omega_k^g, a_k, b_k) - \frac{1}{2}\sum_{g=1}^{G}(\mu^g - \mu_\star^g)^t V_\star^{g^{-1}}(\mu^g - \mu_\star^g) - \lambda_F \|P\mu\|_1$$

with $\mu^g \sim \mathcal{N}(\mu_\star^g, V_\star^g)$ as prior distribution.

Due to the small number of subjects per group, only differences between groups for the bioavailability parameter F are analyzed. The penalized SAEM algorithm is applied to this model penalizing fixed effect and random effects variance of bioavailability (F). Parameters V_c, Cl, $MAT$and CV are supposed equal between the groups. High values for the adaptive weights were used for the corresponding differences to ensure they are null across groups. This assumption seems reasonable as: i) subjects are highly comparable due to very stringent inclusion criterions and ii) P-Gp inhibitors do not seem to influence MAT and CV [Delavenne et al., 2013]. The penalized SAEM algorithm is applied using the graph structure depicted in Figure 5 and a grid composed of 400 pairs of λ_F and λ_V values.

The optimal model selected by the BIC is shown in Figure 6. Concerning fixed effects, the bioavailability is different between the two dosing regimens. It is certainly the consequence of the very low and pH-dependant solubility of DE. As the dosing regimen B was the lowest, then

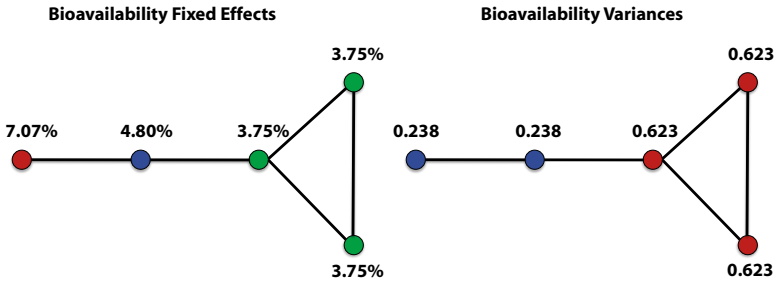

Figure 6: Model selected by the BIC and unpenalized reestimation of the bioavailability parameters from the real data. Groups with same color share equal estimates.

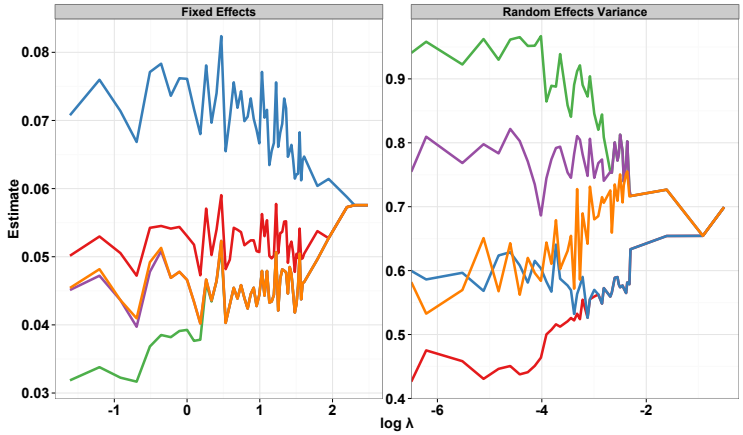

Figure 7: Regularization path for both fixed and variance bioavailability parameters from the pooled DE data set. Red, bleu, green, purple and orange lines correspond to DE_A, $DE_A + PgpI_1$, DE_B, $DE_B + PgpI_2$ and $DE_B + PgpI_3$ respectively.

the smaller the dose, the lower the DE solubility is. Among the three P-Gp inhibitors, only $PgpI_1$ is associated to an increase in DE bioavailability. It is not surprising since $PgpI_1$ is known to be a strong P-Gp inhibitor. $PgpI_2$ and $PgpI_3$ inhibit P-Gp much less in in-vitro experiment. Concerning random effects variances, a higher variance is estimated for dosing regimen B, which again is certainly attributable to solubility. Finally, Figure 7 shows the regularization path of both fixed effects and variances.

6. Discussion

In this paper, we present a fused lasso penalized version of the SAEM algorithm. It allows the introduction of sparsity in the difference between group parameters for both fixed effects and variances of random effects. This algorithm is designed to iteratively maximize the penalized conditional expectation of the complete data likelihood. Simulation results show that this algo-

rithm has good empirical convergence properties. The theoretical study of this algorithm will be the scope of future work.

Several extensions could be proposed. First, the hypothesis that the variance covariance matrix is diagonal might be too strong. For example, in pharmacokinetics the clearance Cl and the volume of distribution parameter may be strongly correlated. Neglecting this correlation could have important consequences on the model prediction properties. Moreover, the penalty used in this work does not allow for random effect selection. One way to tackle these two issues would be to directly penalize the variance-covariance matrix (instead of its inverse), which could be achieved by using the reparametrisation described by [Bondell et al., 2010].

In this work, group sizes are supposed equal or not too different, which is often the case in pharmacokinetic. The algorithm could be easily modified by introducing the group size in the sum of the group conditional expectation [Danaher et al., 2013]:

$$\sum_{g=1}^{G} N_g \tilde{Q}_k(\mu^g, \beta^g, \Omega_k^g, a_k, b_k).$$

Concerning the selection of tuning parameters, other criterions than BIC have been used for generalized linear models. The cross-validated prediction error may be particularly useful especially for high dimensional data since the unpenalized re-estimation of the log-likelihood can not always be done. For NLME, this criterion has already been studied by [Colby and Bair, 2013] and could be easily implemented.

Finally a last improvement, subject of a future work, is the extension to NLMEMs including more than one level of random effects [Panhard and Samson, 2009]. Indeed in this paper the method is applied to data from a cross-over trial, where each subject receives the two treatment modalities. This information was neglected and the five groups were considered as independent which could lead to spurious association when inter occasion variability is high.

References

A. Arribas-Gil, K. Bertin, C. Meza, and V. Rivoirard. Lasso-type estimators for semiparametric nonlinear mixed-effects models estimation. *Statistics and Computing*, 24(3):443–460, 2014.

Y. F. Atchade, G. Fort, and E. Moulines. On stochastic proximal gradient algorithms. *arXiv preprint arXiv:1402.2365*, 2014.

F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, pages 19–53, 2011.

E. T. Bell. Exponential numbers. *American Mathematical Monthly*, pages 411–419, 1934.

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

J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

S. Blech, T. Ebner, E. Ludwig-Schwellinger, J. Stangier, and W. Roth. The metabolism and disposition of the oral direct thrombin inhibitor, dabigatran, in humans. *Drug Metabolism and Disposition*, 36(2):386–399, 2008.

H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

E. Colby and E. Bair. Cross-validation for nonlinear mixed effects models. *Cross-Validation for Nonlinear Mixed Effects Models*, 2013.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, 2013.

M. Delattre, M. Lavielle, M.-A. Poursat, et al. A note on bic in mixed-effects models. *Electronic Journal of Statistics*, 8: 456–475, 2014.

X. Delavenne, E. Ollier, T. Basset, L. Bertoletti, S. Accassat, A. Garcin, S. Laporte, P. Zufferey, and P. Mismetti. A semi-mechanistic absorption model to evaluate drug–drug interaction with dabigatran: application with clarithromycin. *British journal of clinical pharmacology*, 76(1):107–113, 2013.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Annals of Statistics*, pages 94–128, 1999.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

J. Gertheiss and G. Tutz. Regularization and model selection with categorial effect modifiers. *Statistica Sinica*, 22: 957–982, 2012.

H. Höfling, H. Binder, and M. Schumacher. A coordinate-wise optimization algorithm for the fused lasso. *arXiv preprint arXiv:1011.6409*, 2010.

E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4):1020–1038, 2005.

N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

E. Ollier and V. Viallon. Joint estimation of k related regression models with simple l_1-norm penalties. *arXiv preprint arXiv:1411.1594*, 2014.

X. Panhard and A. Samson. Extension of the saem algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics*, 10(1):121–135, 2009.

F. Rohart, M. San Cristobal, and B. Laurent. Selection of fixed effects in high dimensional linear mixed models using a multicycle ecm algorithm. *computational Statistics and Data Analysis*, DOI: 10.1016/j.csda.2014.06.022, 2014.

A. Samson, M. Lavielle, and F. Mentré. The saem algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. *Statistics in medicine*, 26(27):4860–4875, 2007.

J. Schelldorfer, P. Buhlmann, and S. De Geer. Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38:197–214, 2011.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

V. Viallon, S. Lambert-Lacroix, H. Hoefling, and F. Picard. On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, pages 1–17, 2014.

H. Wang. Coordinate descent algorithm for covariance graphical lasso. *Statistics and Computing*, pages 1–9, 2013.

M. Weiss, P. Sermsappasuk, and W. Siegmund. Modeling the kinetics of digoxin absorption: Enhancement by p-glycoprotein inhibition. *The Journal of Clinical Pharmacology*, 52(3):381–387, 2012.

D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.