Selection of fixed effects in high dimensional Linear Mixed Models using a multicycle ECM algorithm

Florian Rohart[1,2], Magali San-Cristobal [2] and Béatrice Laurent [1]

[1] UMR 5219, Institut de Mathématiques de Toulouse,
INSA de Toulouse, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France

[2] UMR 444 Laboratoire de Génétique Cellulaire,
INRA Toulouse, 31320 Castanet Tolosan cedex, France

2012

Abstract

We consider linear mixed models in which observations are grouped. A ℓ^1-penalization on the fixed effects coefficients of the log-likelihood obtained by considering the random effects as missing values is proposed. A multicycle ECM algorithm, which can be combined with any variable selection method developed for linear models, was used to solve the optimization problem. The algorithm allows for a number of parameters p to be larger than the total number of observations n and is faster than the lmmLasso method of Schelldorfer *et al.* (2011) since no n x n matrix needs to be inverted. We show that the theoretical results of Schelldorfer *et al.* (2011) apply for our method when the variances of both the random effects and the residuals are known. When combined with a variable selection method of Rohart (2011), the algorithm provided good estimations for the set of relevant fixed effect coefficients as well as variances. It outperforms the lmmLasso both in common ($p < n$) and high-dimensional settings ($p \geq n$).

1 Introduction

More and more real data sets contain high-dimensional data owing to the more extensive use of new technologies such as high-thoughput DNA/RNA chips or RNA sequencing in biology. High-dimensional settings, in which the number of parameters p is greater than the number of observations n, generally means that the problem cannot be solved. In order to address this problem, various constraints are implemented. Common constraints are for example sparsity, which implies that a lot of parameters are zero, or use of a well-conditioned variance matrix for the observations. Many studies have addressed the problem of variable selection, most of which have used a linear model $Y = X\beta + \epsilon$, where X is an $n \times p$ matrix containing the observations and ϵ is a n-vector of i.i.d random variables (usually Gaussian). One of the oldest method is the Akaike Information Criterion (AIC), which is a penalization of the log-likelihood by a function of the number of parameters included in the model. More recently, the both simple and powerful Lasso (Least Absolute Shrinkage and Selection Operator) method (Tibshirani, 1996) revolutionized the field. The Lasso works by applying a ℓ^1 -penalty to the estimate of least squares which shrinks some coefficients to exactly zero. Various extensions exist for the Lasso: group Lasso (Yuan and Lin, 2007), adaptive Lasso (Huang et al., 2008) and a more stable version known as Bo-Lasso (Bach, 2009), for example. Penalizing the likelihood is not the only way to perform variable selection. Indeed recent statistical tests (Rohart, 2011) also appear to provide good results.

In all the previously described methods, observations are considered to be independent and identically distributed. These methods are therefore no longer adapted when structured information, such as family relationships or common environmental effects, becomes available. In a linear mixed model, the observations are assumed to be clustered. The variance-covariance matrix V of the observations is therefore no longer diagonal but, in some cases, can be assumed to be block diagonal. In the literature, most reports of linear mixed models relate to the estimation of variance components, using either maximum likelihood estimation (ML) (Henderson, 1973, 1953), or restricted maximum likelihood estimation (REML) which accounts for the loss in degrees of freedom due to fitting fixed effects (Patterson and Thompson, 1971; Harville, 1977; Henderson, 1984; Foulley et al., 2006). However, both methods assume that each fixed effect and each random effect is relevant. This assumption might be wrong and result in falsely estimated parameters, especially for high-dimensional analysis. Contrarily to the linear model, there are few reports on the selection of fixed effect coefficients using a linear mixed model in a high-dimensional setting.

Both Bondell et al. (2010) and Ibrahim et al. (2011) used penalized likelihoods to perform selection of both fixed and random effects. Bondell et al. (2010) introduced a constrained EM algorithm to solve the optimization problem, which becomes computationally complex in a high-dimensional context (it should be noted that their simulation studies were only designed for a low dimensional setting). Moreover, the methods of both Bondell et al. (2010) and Ibrahim et al. (2011) rely on Cholesky decompositions and, as pointed out by Müller et al. (2013), these decompositions are dependent on the order in which the random effects appear and are not permutation invariant (Pourahmadi, 2011). In the present paper, the selection of both fixed and random effects is out of the scope because the aim of the study was to analyze a real dataset with only a few random effects.

Schelldorfer et al. (2011) have studied the selection of fixed effects in a high dimensional setting. Their paper introduced an algorithm based on ℓ^1-penalization of the maximum likelihood estimator in order to select the relevant fixed effect coefficients. As highlighted in their paper, their algorithm relies on the possibly time-consuming process of inverting the variance matrix of the observations V.

We present in this paper an efficient way to select fixed effects in a linear mixed model. We propose that random effects be considered as missing data, as previously described in Bondell et al. (2010) and Foulley (1997), and to introduce a ℓ^1-penalty on the log-likelihood of the complete data . We propose a multicycle ECM algorithm with convergence properties (Foulley, 1997; McLachlan and Krishnan, 2008; Meng and Rubin, 1993) to solve the optimization problem and provide theoretical results when the variances of the observations are known. Due to its step design, the algorithm can be combined with any variable selection method built for linear models. Nevertheless, the performance of the combination depends to a great extent on the variable selection method that is used. As there is little literature on the selection of fixed effects in a high-dimensional linear mixed model, we will mainly compare our results to those of Schelldorfer et al. (2011).

The analysis is then extended to a real data set from a project in which hundreds of pigs were studied, our aim being to shed light on the relationships between some of the phenotypes of interest and metabolomic data (Rohart et al., 2012). Linear mixed models are appropriate in this case because observations are in fact repeated data collected in different environments (groups of animals reared together in the same conditions). Some individuals were also genetically related, introducing a family effect. The data set consisted of 506 individuals from 3 breeds, 8 environments and 157 families, metabolomic data contained $p = 375$ variables, and the phenotype investigated was the Daily Feed Intake (DFI).

This paper is organized as follows: first the linear mixed model and objective function are described, and then the multicycle ECM algorithm used to solve the optimization problem of the objective function is detailed. In Section 3.1, the algorithm described in Section 2 is generalized so that it can be used with any variable selection method developed for linear models. Next, the results from a simulation study are presented and show that the combination of this new algorithm with a good variable selection method performs well (Section 4). Finally, in Section 5, the method is applied to a real data set.

2 The method

Let us introduce some notations that will be used throughout the paper. $Var(a)$ denotes the variance-covariance matrix of the vector a. For all $a > 0$, set I_a the identity matrix of $\mathbb{R}^a$. For $A \in \mathbb{R}^{n \times p}$, denote I a subset of $\{1, \dots, n\}$ and J a subset of $\{1, \dots, p\}$. Let $A_{I,J}$ $A_{.,J}$ and $A_{I,.}$ denote submatrices of A respectively composed of elements of A with rows in I and columns in J, columns in J and all rows, and rows in I and all columns. Moreover, for all $a > 0, b > 0$, we denoted 0_a to be the vector of size a in which all coordinates were 0 and $0_{a \times b}$ to be the null matrix of size $a \times b$. Let us denote $|A|$ the determinant of matrix A.

2.1 Setting-up the linear mixed model

We consider the linear mixed model in which observations are grouped and we suppose that only a small subset of fixed effect coefficients are nonzero. The aim of this study is to recover this subset using the algorithm presented in the next section of the paper. In the present section we describe the linear mixed model and our objective function.

Assuming that there are q random effects, let N be the total number of groups and n the total number of observations with $n = \sum_{i=1}^{N} n_i$, where n_i is the number of observations within group i. We denoted $N_q = qN$.

The linear mixed model can be written as

$$y = X\beta + \sum_{k=1}^{q} Z_k u_k + \epsilon, \tag{1}$$

where

- y is the set of observed data of length n,
- β is an unknown vector of $\mathbb{R}^p$; $\beta = (\beta_1, \ldots, \beta_p)$,
- X is the $n \times p$ matrix of fixed effects; $X = (X_1, \ldots, X_p)$,
- For $k = 1, \ldots, q$, $u_k = (u_k^1, \ldots, u_k^N)$ is a N-vector of i.i.d. coordinates for random effect k,
- For $k = 1, \ldots, q$, Z_k is a $n \times N$ incidence matrix (each row of Z_k contains only one nonzero coefficient),
- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is a Gaussian vector with i.i.d. components $\epsilon \sim \mathcal{N}_n(0, \sigma_e^2 I_n)$, where σ_e is an unknown positive quantity. We denote by R the variance-covariance matrix of ϵ, $R = \sigma_e^2 I_n$.

An example of matrices Z_k for $n = 6$ and two random effects is provided below.

Let $Z_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ and $Z_2 = \begin{pmatrix} x_1 & 0 & 0 \\ x_2 & 0 & 0 \\ 0 & x_3 & 0 \\ 0 & x_4 & 0 \\ 0 & 0 & x_5 \\ 0 & 0 & x_6 \end{pmatrix}$. Note that Z_2 is the incidence matrix of the interaction of the variable $x = (x_1, \ldots, x_6)$ and the grouping factor.

We denote $u = (u_1', \ldots, u_q')'$ and Z the concatenation of $(Z_1, \ldots, Z_q)$, and assume that $u \sim \mathcal{N}_{N_q}(0, G)$. Let us denote by $\Psi = (\Psi_{i,j})_{1 \leq i,j \leq q}$ the matrix defined by: $\Psi_{i,j} = \begin{cases} \text{cov}(u_i^1, u_j^1) & \text{if } i \neq j \\ \text{var}(u_i^1) & \text{if } i = j \end{cases}$, then we obtain $G = \Psi \otimes I_N$, where $\otimes$ is the Kronecker product.

One can remark that with these notations, Model (1) can also be written as: $y = X\beta + Zu + \epsilon$.

In the following, we assume that ϵ and u are independent. Thus $Var(u, \epsilon) = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}$.

We consider the matrices X and $\{Z_k\}_{1,\ldots,q}$ to be fixed design. Note that our model (1) and the one in Schelldorfer et al. (2011) are identical.

Let us denote by J the set of the indices of the relevant fixed effects of Model (1); $J = \{j, \beta_j \neq 0\}$. The aim of this paper is to estimate J, β, G and R. Throughout the paper, the number of fixed effects p can be greater than the total number of observations n. However, we focus on the case where only a few fixed-effects are relevant since this paper was motivated by such a case on a real data set, see Section 5. We assume $N_q + |J| < n$.

2.2 A ℓ^1 penalization of the complete log-likelihood

In the following, we consider the fixed effects coefficients β and the variance matrix G as parameters and $\{u_k\}_{k \in \{1,\ldots,q\}}$ as missing data. We denote $\Phi = (\beta, G, \sigma_e^2)$.
The log-likelihood of the complete data $x = (y, u)$ is

$$L(\Phi; x) = L_0(\beta, \sigma_e^2, G; \epsilon) + L_1(G; u), \tag{2}$$

where

$$-2L_0(\beta, \sigma_e^2, G; \epsilon) = n\log(2\pi) + n\log(\sigma_e^2) + \left|\left|y - X\beta - \sum_{k=1}^{q} Z_k u_k\right|\right|^2 / \sigma_e^2, \tag{3a}$$

$$-2L_1(G; u) = N_q \log(2\pi) + \log(|G|) + u'G^{-1}u. \tag{3b}$$

Indeed, (2) results from $p(x|\Phi) = p(y|\beta, u, \sigma_e^2)p(u|G)$; (3a) from $L_0(\beta, \sigma_e^2, G; \epsilon) = L_0(\sigma_e^2; \epsilon) = n\log(2\pi) + n\log(\sigma_e^2) + \epsilon'\epsilon/\sigma_e^2$ because $\epsilon|\sigma_e^2 \sim \mathcal{N}_n(0, \sigma_e^2 I_n)$ and (3b) from $u|G \sim \mathcal{N}_{N_q}(0, G)$.

Since we allow for a number of fixed-effects p greater than the total number of observations n, the usual maximum likelihood (ML) or restricted maximum likelihood (REML) approaches do not apply. Because we assumed that β is sparse (many coefficients are assumed to be null) and because we want to recover that sparsity, we add a ℓ^1 penalty on β to the log-likelihood of the complete data (2). Indeed ℓ^1 penalization is known to induce sparsity in the solution, as in the Lasso method (Tibshirani, 1996) or the lmmLasso method (Schelldorfer et al., 2011). Thus we consider the following objective function to be minimized:

$$g(\Phi; x) = -2L(\Phi; x) + \lambda|\beta|_1, \tag{4}$$

where λ is a positive regularization parameter. It should be noted that the function g could have been obtained in a Bayesian setting considering a Laplace prior on β.
It is interesting to note that finding a minimum of the objective function (4) is a non-linear, non-differentiable and non-convex problem. More importantly, a striking fact (especially noticeable in (3b)) is that the function g is not lower-bounded. Indeed, $L(\Phi; x)$ tends to infinity when $|G|$ tends towards 0, i.e. when a random effect should not have been included in the model. This is a well-known problem of likelihood degeneracy, especially studied in Gaussian mixture model (Biernacki and Chrétien, 2003). In linear mixed models, some authors focus on the log-likelihood of the marginal model in which the random effects are integrated in the matrix of variance of the observations Y. This is the case in Schelldorfer et al. (2011):

$$y = X\beta + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, V).$$

Note that $V = ZGZ' + R$. The degeneracy of the likelihood can also appear in the marginal model when the determinant of V tends towards zero. This phenomenon is likely to occur in a high dimensional context when the model includes too many fixed-effects, that is to say when insufficient regularization is applied by the lmmLasso penalty (Schelldorfer et al., 2011) or by λ in (4).

In the next section, a multicycle ECM algorithm is used to solve the minimization of (4) and select fixed-effects.

2.3 A multicycle ECM algorithm

The multicycle ECM algorithm (Meng and Rubin, 1993; Foulley, 1997; McLachlan and Krishnan, 2008) used to solve the minimization problem of (4) contains four steps: two E steps interlaced with two M steps. Each step is described in this section.
It should be recalled that $\Phi = (\beta, G, \sigma_e^2)$ is the vector of the parameters to estimate and that $u = (u_1', \dots, u_k')'$ is a vector of missing values. The multicycle ECM algorithm is an

iterative algorithm. Iterations are indexed by $t \in \mathbb{N}$ and $\Theta^{[t]}$ denotes the estimation of parameter Θ at iteration t.

Let $E_{u|y,\Phi=\Phi^{[t]}}$ denote the conditional expectation under the distribution of u given the vector of observations y and the current estimation of the set of parameters Φ at iteration t.

2.3.1 First E-step

Let us denote

$$Q(\Phi;\Phi^{[t]}) = E_{u|y,\Phi=\Phi^{[t]}}[g(\Phi;x)].$$

Q can be decomposed as follows:

$$Q(\Phi;\Phi^{[t]}) = Q_0(\beta, G, \sigma_e^2;\Phi^{[t]}) + Q_1(G;\Phi^{[t]}),$$

where

$$Q_0(\beta, G, \sigma_e^2;\Phi^{[t]}) = n \ \log(2\pi) + n \ \log(\sigma_e^{2[t]}) + E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon)/\sigma_e^{2[t]} + \lambda|\beta^{[t]}|_1$$

and

$$Q_1(G;\Phi^{[t]}) = N_q \log(2\pi) + \log(|G^{[t]}|) + E_{u|y,\Phi=\Phi^{[t]}}(u'G^{-1[t]}u).$$

By definition,

$$E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon) = \left|\left|E_{u|y,\Phi=\Phi^{[t]}}\left(\epsilon\right)\right|\right|^2 + tr\left(Var_{u|y,\Phi=\Phi^{(t)}}\left(\epsilon\right)\right).$$

$E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon)$ can be further detailed as:

$$E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon) = \left|\left|y - X\beta^{[t]} - ZE\left(u|y,\Phi=\Phi^{[t]}\right)\right|\right|^2 + tr\left(ZVar\left(u|y,\Phi^{[t]}\right)Z'\right). \qquad (5)$$

As designated by Henderson (1973), $E\left(u|y,\Phi=\Phi^{[t]}\right)$ is the BLUP (Best Linear Unbiased Prediction) of u for the vector of parameters Φ equal to $\Phi^{[t]}$. Let us denote $u^{[t+1/2]} = E\left(u|y,\Phi=\Phi^{[t]}\right)$, we have that

$$u^{[t+1/2]} = (Z'Z + \sigma_e^{2[t]}G^{-1[t]})^{-1}Z'\left(y - X\beta^{[t]}\right).$$

2.3.2 M-Step for β

The next step minimizes $Q_0(\beta, G, \sigma_e^2;\Phi^{[t]})$ with respect to β:

$$\beta^{[t+1]} = \underset{\beta}{Argmin}\left(\frac{1}{\sigma_e^{2[t]}}\left|\left|\left(y - Zu^{[t+1/2]}\right) - X\beta\right|\right|^2 + \lambda\left|\beta\right|_1\right). \qquad (6)$$

It can be remarked that (6) is a Lasso on β with the vector of "observed" data $\left(y - Zu^{[t+1/2]}\right)$ and the penalty $\lambda\sigma_e^{2[t]}$.

2.3.3 Second E-Step

A second E-step is performed with the update of the vector of missing values u: $u^{[t+1]} = E\left(u|y, \beta = \beta^{[t+1]}, G = G^{[t]}, \sigma_e^2 = \sigma_e^{2[t]}\right)$, thus

$$u^{[t+1]} = (Z'Z + \sigma_e^{2[t]}G^{-1[t]})^{-1}Z'\left(y - X\beta^{[t+1]}\right).$$

We define $\forall k \in \mathcal{K}$, $u_k^{[t+1]}$ to be the element of size N for the random effect k in $u^{[t+1]}$.

2.3.4 M-step for (G, σ_e^2)

Variance matrices G and R are updated based on the minimization of Q_1 and Q_0 respectively.

Let us recall that $G = \Psi \otimes I_N$. We can therefore write $Q_1(G; \Phi^{[t]}) = N_q \log(2\pi) + N\log(|\Psi^{[t]}|) + tr(\Psi^{-1[t]}\Omega^{[t]})$, where $\Omega^{[t]} = \left\{\omega_{i,j}^{[t]} = E(u_i'u_j|y, \Phi = \Phi^{[t]})\right\}$. Thanks to a lemma reported in Anderson (1984), the minimization of Q_1 with respect to Ψ gives $\Psi^{[t+1]} = \Omega^{[t]}/N$. Thus, for all $1 \leq i, j \leq q$, $\Psi_{i,j}^{[t+1]} = E\left(u_i'u_j|y, G^{[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right)/N$.

Besides, for all $1 \leq i, j \leq q$

$$E\left(u_i'u_j|y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right) = u_i^{[t+1]'}u_j^{[t+1]} + \sum_{k=1}^{N} cov_{u|y,\sigma_k^{2[t]},\sigma_e^{2[t]},\beta^{[t+1]}}(u_{ik}, u_{jk}).$$

Moreover, we can use the following results of Henderson (1973),

$$cov_{u|y,\sigma_k^{2[t]},\sigma_e^{2[t]},\beta^{[t+1]}}(u_i, u_j) = T_{i,j}\sigma_e^{2[t]},$$

where $T_{i,j}$ is defined as follows:

$$\begin{aligned}\left(Z'Z + \sigma_e^{2[t]}G^{-1[t]}\right)^{-1} &= \begin{pmatrix} Z_1'Z_1 + \sigma_e^{2[t]}\Psi^{1,1[t]}I_N & Z_1'Z_2 + \sigma_e^{2[t]}\Psi^{1,2[t]}I_N & \dots & Z_1'Z_q + \sigma_e^{2[t]}\Psi^{1,q[t]}I_N \\ Z_2'Z_1 + \sigma_e^{2[t]}\Psi^{2,1[t]}I_N & Z_2'Z_2 + \sigma_e^{2[t]}\Psi^{2,2[t]}I_N & \dots & Z_2'Z_q + \sigma_e^{2[t]}\Psi^{2,q[t]}I_N \\ \vdots & \vdots & \ddots & \vdots \\ Z_q'Z_1 + \sigma_e^{2[t]}\Psi^{q,1[t]}I_N & Z_q'Z_2 + \sigma_e^{2[t]}\Psi^{q,2[t]}I_N & \dots & Z_q'Z_q + \sigma_e^{2[t]}\Psi^{q,q[t]}I_N \end{pmatrix}^{-1} \\ &= \begin{pmatrix} T_{1,1} & T_{1,2} & \dots & T_{1,q} \\ T_{1,2}' & T_{2,2} & \dots & T_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ T_{1,q}' & T_{2,q}' & \dots & T_{q,q} \end{pmatrix},\end{aligned}$$

with

$$\begin{pmatrix} \Psi^{1,1} & \Psi^{1,2} & \dots & \Psi^{1,q} \\ \Psi^{2,1} & \Psi^{2,2} & \dots & \Psi^{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi^{q,1} & \Psi^{q,2} & \dots & \Psi^{q,q} \end{pmatrix} = \begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} & \dots & \Psi_{1,q} \\ \Psi_{2,1} & \Psi_{2,2} & \dots & \Psi_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{q,1} & \Psi_{q,2} & \dots & \Psi_{q,q} \end{pmatrix}^{-1}.$$

Thus:

$$\Psi_{i,j}^{[t+1]} = \frac{1}{N}\left[u_i^{[t+1]\prime} u_j^{[t+1]} + tr(T_{i,j})\sigma_e^{2[t]}\right].$$

The minimization of Q_0 with respect to σ_e^2 gives: $\sigma_e^{2[t+1]} = E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon)/n$. From (5), we have

$$\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Zu^{[t+1]}\right|\right|^2 + tr\left(Z(Z'Z + \sigma_e^{2[t]}G^{-1[t]})^{-1}Z'\right)\sigma_e^{2[t]}\right].$$

Since

$$\begin{aligned} tr\left(Z\left(Z'Z + \sigma_e^{2[t]}G^{-1[t]}\right)^{-1}Z'\right) &= tr\left(\left(Z'Z + \sigma_e^{2[t]}G^{-1[t]}\right)^{-1}Z'Z\right) \\ &= N_q - tr\left[\left(Z'Z + \sigma_e^{2[t]}G^{-1[t]}\right)^{-1}\sigma_e^{2[t]}G^{-1[t]}\right] \\ &= N_q - \sigma_e^{2[t]} tr\left(TG^{-1[t]}\right) \end{aligned}$$

we have

$$\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Zu^{[t+1]}\right|\right|^2 + \sigma_e^{2[t]}\left(N_q - \sigma_e^{2[t]} tr\left(TG^{-1[t]}\right)\right)\right].$$

In summary, the algorithm can be detailed as follows:

```
**Algorithm 2.1** (Lasso+). Initialization:
Initialize the set of parameters $\Phi^{[0]} = (G^{[0]}, \sigma_e^{2[0]}, \beta^{[0]})$.
Define $Z$ as the concatenation of $Z_1, \dots, Z_q$ and $u = (u_1', \dots, u_q')'$.
Until convergence:
1. E-step
$u^{[t+1/2]} = (Z'Z + \sigma_e^{2[t]}G^{-1[t]})^{-1}Z'\left(y - X\beta^{[t]}\right)$
2. M-step
$\beta^{[t+1]} = \underset{\beta}{Argmin}\left(\left|\left|\left(y - Zu^{[t+1/2]}\right) - X\beta\right|\right|^2 + \lambda\sigma_e^{2[t]}\left|\beta\right|_1\right)$
3. E-step
$u^{[t+1]} = (Z'Z + \sigma_e^{2[t]}G^{-1[t]})^{-1}Z'\left(y - X\beta^{[t+1]}\right)$
4. M-step
(a) Set $\Psi_{i,j}^{[t+1]} = \frac{1}{N}\left[u_i^{[t+1]\prime} u_j^{[t+1]} + tr(T_{i,j})\sigma_e^{2[t]}\right]$ and $G^{[t+1]} = \Psi^{[t+1]} \otimes I_N$
(b) Set $\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Zu^{[t+1]}\right|\right|^2 + \sigma_e^{2[t]}\left(N_q - \sigma_e^{2[t]} tr\left(TG^{-1[t]}\right)\right)\right]$
end
```

Convergence of Algorithm 2.1 is ensured because it is a multicycle ECM algorithm (Meng and Rubin, 1993).

Three stopping criteria are used to stop the convergence process of the algorithm: a first criterion on $||\beta^{[t+1]} - \beta^{[t]}||^2$, a second on $||u_k^{[t+1]} - u_k^{[t]}||^2$ for each random effect u_k and lastly a criterion on $||L(\Phi^{[t+1]}, x) - L(\Phi^{[t]}, x)||^2$ where $L(\Phi, x)$ is the log-likelihood defined by (2). Convergence occurs when all criteria are fulfilled. We implemented an additional fourth condition that limited the number of iterations. We choose to initialize the Algorithm 2.1 using the following conditions: $G^{[0]}$ is the block diagonal matrix of $\sigma_1^{2[0]}I_N, \dots, \sigma_q^{2[0]}I_N$ where for all $1 \leq k \leq q, \sigma_k^{2[0]} = \frac{0.4}{q}\ \sigma_e^{2[-1]}, \sigma_e^{2[0]} = 0.6\ \sigma_e^{2[-1]}$, and $(\sigma_e^{2[-1]}, \beta^{[0]})$ is estimated from a linear estimation (without the random effects) of the Lasso with the given penalty λ. In Section 4.4, the impact of initializing the algorithm is investigated on simulated data.

Because the estimation of the set of parameters Φ is biased (Zhang and Hunag, 2008), one last step can be added in order to address this problem once both Algorithm 2.1 has converged and the penalization parameter λ has been tuned. Indeed, it is better to use Algorithm 2.1 to estimate the support of β and then estimate the set Φ using a classic mixed model estimation based on the model:

$$y = X\beta_{\hat{J}} + \sum_{1\leq k\leq q} Z_k u_k + \epsilon,$$

where $\hat{J}$ is the estimated set of indices of the relevant fixed effects.

Proposition 2.2. *When the variances are known, minimization of the objective function* (4) *is the same as that of* $Q(\beta) = (y - X\beta)'V^{-1}(y - X\beta) + \lambda|\beta|_1$, *which is the objective function described in Schelldorfer et al. (2011) with known variances.*

Let us recall that in Schelldorfer et al. (2011), the authors provided theoretical results as regards to the consistency of their method. Based on Proposition 2.2, these results apply to our method in the case of known variances. Proof for Proposition 2.2 is provided in Appendix C.

2.4 The tuning parameter

The solution depends on a regularization parameter, included in Algorithm 2.1, that controls shrinkage. This parameter has to be tuned. We choose to use of the Bayesian Information Criterion (BIC) to do this (Schwarz, 1978):

$$\lambda_{BIC} = \underset{\lambda}{Argmin} \left\{ \log|V_\lambda| + (y - X\hat{\beta}_\lambda)'V_\lambda^{-1}(y - X\hat{\beta}_\lambda) + d_\lambda . \log(n) \right\},$$

where $V_\lambda = Z\hat{G}Z' + \hat{\sigma}_e^2 I_n$ and $\hat{G}, \hat{\sigma}_e^2, \hat{\beta}_\lambda$ are obtained from the minimization of the objective function g defined by (4). Moreover, d_λ is the sum of the number of non-zero variance-covariance parameters and the number of non-zero fixed effects coefficients included in the model selected with the regularization parameter λ.

Other methods could have been used to tune λ such as AIC or cross-validation. We opted for BIC rather than cross-validation mainly because of the gain in computational time.

In the next section, we propose a generalization of Algorithm 2.1 for use with any of the variable selection methods developed for linear models.

3 Extending the method

3.1 Generalizing the algorithm

Algorithm 2.1 provides good results, as demonstrated for the simulation study in Section 4. Nevertheless, because the aim of the second step of the algorithm is to select the relevant coefficients of β in a linear model, the Lasso method can be replaced by any variable selection method built for linear models. If the variable selection method optimizes a criterion, such as the adaptive Lasso (Zou, 2006) or the elastic net (Zou and Hastie, 2005), the resulting algorithm is a multicycle ECM algorithm and the convergence property still holds. However, the convergence property does not hold for methods that do not optimize a criterion.

Algorithm 2.1 can be reshaped for a generalized algorithm as follows:

Algorithm 3.1. _Initialization:_

```
Initialize the set of parameters $\Phi^{[0]} = (G^{[0]}, \sigma_e^{2[0]}, \beta^{[0]})$.
Define $Z$ as the concatenation of $Z_1, \ldots, Z_q$ and $u = (u_1', \ldots, u_q')'$.
Until convergence:
1. $u^{[t+1/2]} = (Z'Z + \sigma_e^{2[t]} G^{-1[t]})^{-1} Z' \left(y - X\beta^{[t]}\right)$
2. Variable selection and estimation of $\beta$ in the linear model $y - Zu^{[t+1/2]} = X\beta + \epsilon^{[t]}$,
   where $\epsilon^{[t]} \sim \mathcal{N}(0, \sigma_e^{2[t]} I_n)$.
3. $u^{[t+1]} = (Z'Z + \sigma_e^{2[t]} G^{-1[t]})^{-1} Z' \left(y - X\beta^{[t+1]}\right)$
4. (a) Set $\Psi_{i,j}^{[t+1]} = \frac{1}{N}\left[u_i^{[t+1]'} u_j^{[t+1]} + tr(T_{i,j})\sigma_e^{2[t]}\right]$ and $G^{[t+1]} = \Psi^{[t+1]} \otimes I_N$
   (b) Set $\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Zu^{[t+1]}\right|\right|^2 + \sigma_e^{2[t]}\left(N_q - \sigma_e^{2[t]} tr\left(TG^{-1[t]}\right)\right)\right]$
end
```

We choose to initialize Algorithm 3.1 in the same way as Algorithm 2.1.
In the following we propose to combine Algorithm 2.1 with a method that does not require a tuning parameter, namely the procbol method (Rohart, 2011). The procbol method sequentially tests multiple hypotheses and determines statistically the set of relevant variables in the linear model $y = X\beta + \epsilon$ where ϵ is an i.i.d Gaussian noise. This method consists of two steps: first, variables are ordered taking into account the observations y and then, in the second step, multiple hypotheses are tested to distinguish between relevant and irrelevant variables. The procbol method has proved to be powerful under certain conditions as reported in Rohart (2011).

3.2 Generalizing the model to different grouping variables

Assume that there are q random effects and q grouping factors ($q \geq 1$), where some grouping factors may be identical. The levels of the factor k are denoted $\{1, 2, \ldots, N_k\}$. The i^{th}-observation belongs to the groups $(i_1, \ldots, i_q)$, where for all $l = 1, \ldots, q$, $i_l \in \{1, 2, \ldots, N_l\}$. It should be noted that two observations can belong to the same group for a given grouping factor and to different groups for another grouping factor.

In this setting, the total number of observations is $n = \sum_{i=1}^{N_k} n_{i,k}, \forall k \leq q$, where $n_{i,k}$ is the number of observations within group i from the grouping factor k. We therefore have $N = \sum_{k=1}^{q} N_k$.

The linear mixed model can be written as

$$y = X\beta + \sum_{k=1}^{q} Z_k u_k + \epsilon, \tag{7}$$

the differences with model (1) being that

- For $k = 1, \dots, q$, u_k is a N_k-vector of the random effect for grouping factor k, ,
- For $k = 1, \dots, q$, Z_k is a $n \times N_k$ incidence matrix for grouping factor k.

Both Algorithms 2.1 and 3.1 apply with Model (7) when random effects are considered to be independent. Indeed, the covariance matrix G of $(u_1, \dots, u_q)$ has to be a diagonal matrix since the two vectors have to be of the same length for the covariance matrix to be estimated. Ψ is therefore also a diagonal matrix and for all $1 \leq k \leq q$, $\Psi_{k,k} = \dfrac{1}{N_k}\left[u_k^{[t+1]\prime} u_k^{[t+1]} + tr(T_{k,k})\sigma_e^{2[t]}\right]$, where $T_{k,k}$ is defined as in Section 2.

In the particular case of independence of the random effects, a naive selection of the random effects can be performed when the variance of a random effect drops to 0. When $\Psi_{k,k}$ is too small at some step t of the ECM algorithm, the random effect u_k is removed from the model.

In Section 4, we show that the combination of Algorithm 3.1 and the procbol method performs well on simulated data.

4 Simulation study

The purpose of this section is to compare different methods that aim at selecting the correct fixed effects coefficients in a linear mixed model (1). We shall also determine whether including random effects in the model improves its performances.

4.1 Methods used

We compare several methods. Some of the methods are designed to work in a linear model: *Lasso* (Tibshirani, 1996), *adLasso* (Zou, 2006) and *procbol* (Rohart, 2011), while others are designed to work in a linear mixed model: *lmmLasso* (Schelldorfer et al., 2011), *Algorithm 2.1* (designated as *Lasso+*), *adLasso+Algorithm 3.1* (designated as *adLasso+*) and *procbol+Algorithm 3.1* (designated as *pbol+*).
The initial weights of the *adLasso* and *adLasso+* are both set to $1/|\tilde{\beta}_i|$ where for all $i \in \{1, \dots, p\}$, $\tilde{\beta}_i$ is the Ordinary Least Squares (OLS) estimate of β_i in the model $y_i = X_i\beta_i + \epsilon_i$.

The second step of the procbol method performs multiple hypothesis testing thanks to an estimation of unknown quantiles related to the matrix X. Computing these quantiles at each iteration of the convergence process would make the combination of the procbol method and Algorithm 3.1 almost impossible to run, but in this case the quantiles remain unchanged because no changes occur in the data matrix X throughout the algorithm. The procbol method could therefore be run several times on the same data set with unvarying quantiles. This results in a considerable gain in computational time. Some parameters of the procbol method are changed in order to limit the time of each iteration of the convergence process. The parameter m that denotes the number of bootstrapped samples used to sort the variables (first step of the procbol method) is set to 10. The number of variables arranged in order during the first step of the procbol method is set to 40. Note that when the procbol method is used in a linear model, we set $m = 100$ as recommended in Rohart (2011). Both the *procbol* method and the *pbol+* method are set with a user-level of $\alpha = 0.1$, which reflects for the level of the testing procedure.

For all methods requiring tuning, the tuning parameter is set using the Bayesian Information Criterion as described in Section 2.4. Particular attention is paid to tuning the regularization parameter for some methods, especially *Lasso* and *adLasso*, as it can be difficult in some cases due to the degeneracy of the likelihood (see Appendix B).

4.2 Design of our simulation study

We set X_1 to be the vector of $\mathbb{R}^n$ in which coordinates are all equal to 1 and then consider three models. For each model, the response variable y is computed via $y = \sum_{j=1}^{5} X_{i_j}\beta_{i_j} + \sum_{k=1}^{q} Z_k u_k + \epsilon$, where $J = \{i_1, \ldots, i_5\} \subset \{1, \ldots, p\}$, with q random effects being Gaussian and ϵ being a vector of independent standard Gaussian variables. We set $N = 20$ and $\forall i \in \{1, .., 20\}$ $n_i = 6$. The models used to fit the data differ in the number of parameters p, the number of random effects q, the matrix Ψ and the dependence structure of the X_i's. For each model, we have for all $j = 2, \ldots, p$: $\sum_{i=1}^{n} X_{j,i} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_{j,i}^2 = 1$. For $k = 1, \ldots, q$, the random effects regression matrix Z_k corresponds to the design matrix of the interaction between the k^{th} column of X and the grouping factor, which gives a $n \times N$ matrix. The design of the matrices Z_k's means that the q grouping variables generates both a fixed effect (for to β_k's) and a random effect (for to u_k's). As recommended in Schelldorfer et al. (2011), the variables that generate both a fixed and a random effect do not undergo feature selection to avoid shrinkage of the fixed effect coefficients for those variables towards 0. The models are defined as follows:

- M_1: $n = 120$, $p = 80$, $\beta_J = 3/4$, $q = 3$ and $\Psi = I_3$. For all $j = 2, \ldots, p, X_j \sim \mathcal{N}_n(0, I_n)$.
- M_2: $n = 120$, $p = 300$, $\beta_J = 3/4$, $q = 2$ with $\text{var}(u_1) = \text{var}(u_2) = 1$ and $\text{cov}(u_1, u_2) = 0.5$. The covariates are generated from a multivariate normal distribution with mean zero and covariance matrix Σ with the pairwise correlation $\Sigma_{kk'} = \rho^{|k-k'|}$ and $\rho = 0.5$.
- M_3: $n = 120$, $p = 600$, $\beta_J = 3/4$, $q = 2$ and $\Psi = I_2$. The covariates are generated from a multivariate normal distribution with mean zero and covariance matrix Σ with the pairwise correlation $\Sigma_{kk'} = \rho^{|k-k'|}$ and $\rho = 0.5$.

We also consider a fourth setting in order to study Section 3.2. In this setting the random effects are supposed to be independent and the grouping variables to be different:

- M_4: $n = 120$, $p = 300$, $\beta_J = 2/3$, $q = 2$ and $\Psi = I_2$. For all $j = 2, \ldots, p, X_j \sim \mathcal{N}_n(0, I_n)$. The two grouping variables are different: $N_1 = 20, \forall i \in \{1, .., 20\}$ $n_{i,1} = 6$ and $N_2 = 15, \forall i \in \{1, .., 15\}$ $n_{i,2} = 8$

For all models we set $J = \{1, 2, i_3, i_4, i_5\}$ where $\{i_3, i_4, i_5\} \subset \{3, \ldots, p\}$; in addition, $i_3 = 3$ for model M_1.

The aim is to recover the set of relevant fixed effects coefficients J for each model as well as to estimate the variance matrix of both the random effects and residuals. To evaluate the quality of the methods, we use several criteria: the percentage of true model recovered under the label 'Truth', the cardinal of the estimated set of fixed effects coefficients $|\hat{J}|$, the number of true positives TP, the estimated variance $\hat{\sigma}_e^2$ of the residuals, the estimated variances $\hat{\Psi}$ of the random effects and the mean squared error mse calculated as an ℓ^2

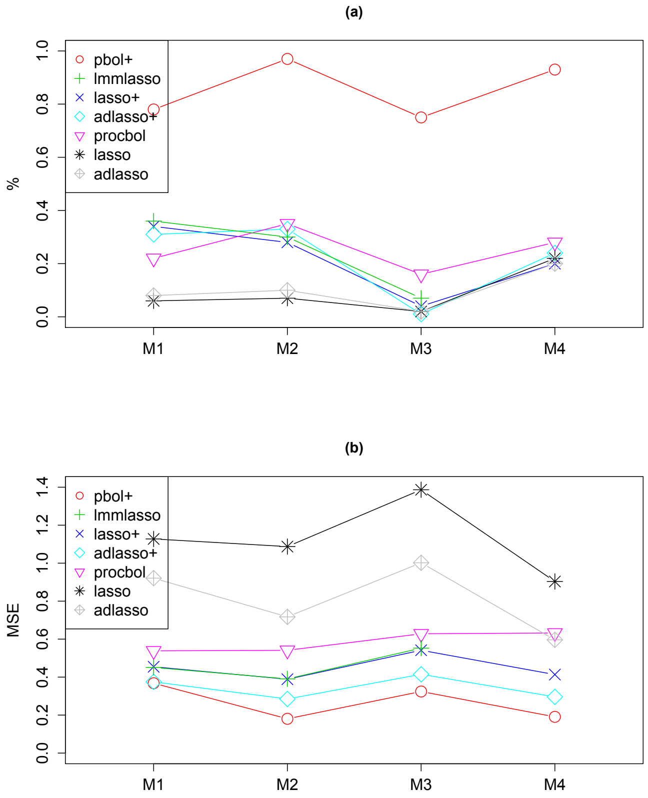


Figure 1: Summary of the results of the simulation study for models $M_1 - M_4$ (X axis). Results of '$\hat{J} = J$' (a) and Mean Squared Error (b) for each model.

error rate between the real value -$X\beta$- and the estimation -$X\hat{\beta}$-. We also determined the Signal-to-Noise Ratio (SNR) as $||X\beta||_2^2/||\sum_{k=1}^q Z_k u_k + \epsilon||_2^2$ for each of the replications.

4.3 Comments on the results

Detailed results of the simulation study are available in Appendix A. A summary of the main results is shown in Figure 1. It should be noted that the *lmmLasso* method of the R-package could not be computed for model M_4 because the function does not support different grouping variables.

In all models, results are improved by switching from a simple linear model to a linear mixed model. Indeed significant differences are observed between *Lasso* and *Lasso+* or *procbol* and *pbol+*, especially with model M_3 (high dimensional setting).

For all models, *lmmLasso* and *Lasso+* give very similar results. This is not really surprising since both methods are based on a ℓ^1-penalization of the log likelihood.

As regards to the *adLasso+* method, it provides a better *mse* result than the *Lasso+* method, but in the meantime the percentage of true positives is lower and the number of selected variables is higher. In our simulations, tuning of the regularization parameter was difficult for both of these methods. Indeed due to the degeneracy of the likelihood, the grid over which the penalty is tuned has to be chosen with care (see Appendix B).

The best results are obtained when Algorithm 3.1 is combined with the procbol method (*pbol+*). This combination provides by far the greatest percentage of true model recovery, estimated fixed effects is the closest to real values and the *mse* is the lowest among the tested methods. Nevertheless, *mse* results for both *Lasso+* and *lmmLasso* could easily be improved by using a linear mixed model estimation as described in Section 2.3 (see Table 7 in Appendix A). It is also interesting to note that the *pbol+* method always converged in our simulations.

A R-package "MMS" is available on CRAN (http://cran.r-project.org). This package contains tools for selecting fixed effects using linear mixed models, including the previously described *Lasso+*, *adLasso+*, *pbol+* methods.

All the results presented in this section were obtained following specific initialization of the algorithms. The next paragraph focuses on the impact of such initialization.

4.4 Impact of initializing our algorithms

Both Algorithm 2.1 and Algorithm 3.1 start by initializing the parameter $\Phi = (G, \sigma_e^2, \beta)$, as mentioned previously in Section 2.3.

We tested different initializations of our algorithms and found that the algorithms always converged towards the same point, whatever the initialization of Φ (not shown). However, the further $\Phi^{[0]}$ was set from the true value of Φ, the higher the number of iterations needed to converge.

5 Application on a real data-set

In this section we analyze a real data set previously described in Rohart et al. (2012). The aim of this analysis is to pinpoint metabolomic data that describes a phenotype taking into account all the available information such as the breed, the batch effect and the relationship between individuals. In the present case, we study the Daily Feed Intake phenotype (DFI). We model the data as follows:

$$y = X_B\beta_B + X_M\beta_M + Z_E u_E + Z_F u_F + \epsilon, \tag{8}$$

where y is the DFI phenotype and X_B, X_M, Z_E, Z_F are the design matrices of the breed effect, the metabolomic data, the batch effect and the family effect, respectively. We consider two random effects, the batch and the family effects, and consider that each level of these factors is a random sample drawn from a much larger population of batches and families, contrary to the breed factor. Since the grouping variables are different, we assume that the random effects are independent. We denote by G the block diagonal matrix with blocks $\sigma_E^2 I_{N_1}$ and $\sigma_F^2 I_{N_2}$, with $N_1 = 8, N_2 = 157$ and where σ_E^2 and σ_F^2 are the variances of the batch and the family effect respectively. Note that the coefficients β_B do not undergo feature selection.

We compare several methods using this model: *Lasso*, *adLasso*, *procbol*, *Lasso+*, *adLasso+* and *pbol+* (see Section 4). The model which is considered for the first three methods is $y = X_B\beta_B + X_M\beta_M + \epsilon$. Both methods *procbol* and *pbol+* were set with a user-level of $\alpha = 0.1$. The results are presented in Table 1.

	$\lvert\hat{J}\rvert$	$\hat{\sigma}_e^2$	$\hat{\sigma}_E^2$	$\hat{\sigma}_F^2$
Lasso	14	3.8×10^{-2}	-	-
adLasso	21	3.4×10^{-2}	-	-
procbol	11	4.1×10^{-2}	-	-
Lasso+	11	3.2×10^{-2}	3.2×10^{-3}	6.4×10^{-3}
adLasso+	10	3.3×10^{-2}	2.5×10^{-3}	6.5×10^{-3}
pbol+	5	3.4×10^{-2}	5.9×10^{-3}	6.5×10^{-3}

Table 1: Results for the real data set

Methods	CPU Time
Lasso+	0.80
lmmLasso	24.28

Table 2: CPU Time for a single run with the same model

When random effects are considered, we observe a decrease of both the residual variance and the number of selected metabolomic variables. This behavior is in accordance with the simulation study. The question that arises from this analysis is to determine whether the variables selected in the linear mixed models are more relevant than those in the linear model. Biological analysis will be carried out to answer that question.

Table 2 shows the computational time for one run when we only consider the batch effect is considered (in order to compute the *lmmLasso*). As can be seen, when a large number of observations are included, the *Lasso+* method is much faster than the *lmmLasso* method (due to the inversion of the matrix of variance V at each step of the convergence process). This simulation was performed on a 2.80GHz CPU with 8.00Go of RAM with a regularization parameter that selects the same model for both methods,

6 Conclusion

In this paper, we proposed to add a ℓ^1-penalization on the complete log-likelihood in order to perform selection of the fixed effects in a linear mixed model. The multicycle ECM algorithm used to minimize the objective function can also be used to select random effects. This algorithm provides the same results as the lmmLasso method described in Schelldorfer et al. (2011), but much faster. Theoretical results obtained in this paper are identical to those found in Schelldorfer et al. (2011) when the variances are known. The structure of our algorithm means that it can be combined with any variable selection method built for linear models, but in some cases this can result in loss of the convergence property. Nonetheless, the combined procbol method gives good results when tested on simulated data and outperforms other approaches.
We applied all of these methods to a real data set and demonstrated that the residual variance could be reduced, even with a smaller set of selected variables.

Acknowledgements. We thank the animal providers (BIOPORC), French ANR for funding the DéLiSus project (ANR-07-GANI-001), financial support from Région Midi-

Pyrénées, and Helen Munduteguy for the English revision of the manuscript.

References

Anderson, T. (1984). *An introduction to multivariate analysis.* Wiley Series in Probability and Statistics.

Bach, F. (2009). Model-consistent sparse estimation through the bootstrap. Technical report, hal-00354771, version 1.

Biernacki, C. and Chrétien, S. (2003). Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with em. *Statistics & Probability Letters*, 61:373–382.

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

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 94:759–771.

Foulley, J. (1997). Ecm approaches to heteroskedastic mixed models with constant variance ratios. *Genetics Selection Evolution*, 29:197–318.

Foulley, J.-L., Delmas, C., and Robert-Granié, C. (2006). Méthodes du maximum de vraisemblance en modèle linéaire mixte. *J. SFdS*, 1-2:5–52.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, 72:320–340.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9:226–252.

Henderson, C. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, pages 10–41.

Henderson, C. (1984). *Applications of linear models in Animal breeding.* University of Guelph, Ont.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptative lasso for sparse high-dimensional regression models. *Stat. Sin.*, 18(4):1603–1618.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67:495–503.

McLachlan, J. and Krishnan, T. (2008). *The EM Algorithm and Extensions, second edition.* Wiley-Interscience.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80:267–278.

Müller, S., Scealy, J., and Welsh, A. (2013). Model selection in linear mixed model. *Statist Sci.* to appear.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Pourahmadi, M. (2011). Covariance estimation: The glm and regularization perspectives. *Statist Sci.*, 26(3):369–387.

Rohart, F. (2011). Multiple hypotheses testing for variable selection. *arXiv:1106.3415v1*.

Rohart, F., Paris, A., Laurent, B., Canlet, C., Molina, J., Mercat, M. J., Tribout, T., Muller, N., Ianuccelli, N., Villa-Vialaneix, N., Liaubet, L., Milan, D., and San-Cristobal, M. (2012). Phenotypic prediction based on metabolomic data on the growing pig from three main european breeds. *Journal of Animal Science*.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scand. J. Stat.*, 38:197–214.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.*, B 58(1):267–288.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.*, B 68:46–67.

Zhang, C.-H. and Hunag, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J.R. Statist. Soc.*, B 67(2):301–320.

Appendix A - Results of the simulation study

Table 3: Results for model M_1. The recovery rate of the true model was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.60(0.12)$. Standard errors are given between parentheses, for 100 runs.

	$\hat{J} = J$	$\|\hat{J}\|$	TP	$\hat{\sigma}_e^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\sigma}_{12}^2$	$\hat{\sigma}_{23}^2$	$\hat{\sigma}_{13}^2$
Ideal	1	5	5	5	1	1	1	0	0	0
Lasso	0.06	4.43	3.43	4.68	-	-	-	-	-	-
		(2.58)	(1.44)	(1.01)	-	-	-	-	-	-
adLasso	0.08	5.25	3.78	4.15	-	-	-	-	-	-
		(2.63)	(1.18)	(1.02)	-	-	-	-	-	-
procbol	0.22	3.89	3.61	4.88		-	-	-	-	-
		(2.09)	(1.14)	(1.08)	-	-	-	-	-	-
Lasso+	0.34	6.19	4.98	1.05	0.97	1.13	0.94	-0.02	-0.00	-0.06
		(1.21)	(0.14)	(0.11)	(0.42)	(0.49)	(0.39)	(0.37)	(0.34)	(0.30)
adLasso+	0.31	6.33	4.93	1.00	0.93	1.04	0.91	-0.02	0.00	-0.06
		(1.75)	(0.26)	(0.12)	(0.41)	(0.48)	(0.39)	(0.34)	(0.32)	(0.30)
lmmLasso	0.36	6.23	4.98	1.09	0.98	1.12	0.95	0.14	0.16	0.10
		(1.52)	(0.14)	(0.22)	(0.40)	(0.47)	(0.38)	(0.24)	(0.25)	(0.20)
pbol+	0.78	4.76	4.76	1.03	0.95	1.06	0.94	0.00	-0.00	-0.07
		(0.47)	(0.47)	(0.13)	(0.40)	(0.45)	(0.37)	(0.34)	(0.35)	(0.31)

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	MSE
Ideal	0.67	0.67	0.67	0.67	0.67	0.00
Lasso	0.73	0.19	0.29	0.23	0.24	1.13
	(0.28)	(0.26)	(0.28)	(0.22)	(0.21)	(0.50)
adLasso	0.73	0.30	0.47	0.39	0.38	0.92
	(0.28)	(0.36)	(0.36)	(0.28)	(0.28)	(0.47)
procbol	0.73	0.49	0.67	0.55	0.60	0.97
	(0.28)	(0.50)	(0.45)	(0.41)	(0.41)	(0.54)
Lasso+	0.73	0.62	0.74	0.46	0.39	0.45
	(0.24)	(0.29)	(0.28)	(0.13)	(0.14)	(0.22)
adLasso+	0.74	0.66	0.75	0.57	0.52	0.37
	(0.23)	(0.29)	(0.28)	(0.17)	(0.20)	(0.23)
lmmLasso	0.73	0.62	0.74	0.46	0.40	0.45
	(0.23)	(0.29)	(0.27)	(0.13)	(0.14)	(0.21)
pbol+	0.74	0.71	0.76	0.72	0.66	0.37
	(0.24)	(0.31)	(0.28)	(0.20)	(0.34)	(0.30)

Table 4: Results for model M_2. The recovery rate of the true model was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.90(0.19)$. Standard errors are given between parentheses, for 100 runs.

Results	$\hat{J} = J$	$\|\hat{J}\|$	TP	$\hat{\sigma}_e^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_{1,2}$
Ideal	1	5	5	1	1	1	0.5
Lasso	0.07	6.86	4.16	3.64	-	-	-
		(11.81)	(1.18)	(0.95)	-	-	-
adLasso	0.10	6.56	4.45	3.05	-	-	-
		(2.67)	(0.76)	(0.76)	-	-	-
procbol	0.35	4.11	3.96	3.76		-	-
		(1.08)	(1.02)	(0.74)	-	-	-
Lasso+	0.28	6.87	5.00	1.12	0.94	0.98	0.47
		(1.89)	(0.00)	(0.16)	(0.39)	(0.38)	(0.27)
adLasso+	0.33	6.92	4.99	1.00	0.90	0.95	0.46
		(2.25)	(0.10)	(0.14)	(0.37)	(0.37)	(0.26)
lmmLasso	0.30	6.87	5.00	1.16	0.93	0.97	0.48
		(1.91)	(0.00)	(0.21)	(0.39)	(0.38)	(0.27)
pbol+	0.97	4.99	4.98	0.99	0.95	0.99	0.48
		(0.17)	(0.14)	(0.11)	(0.38)	(0.37)	(0.27)

	$\hat{\beta}_{i_1}$	$\hat{\beta}_{i_2}$	$\hat{\beta}_{i_3}$	$\hat{\beta}_{i_4}$	$\hat{\beta}_{i_5}$	MSE
Ideal	0.75	0.75	0.75	0.75	0.75	0.00
Lasso	0.81	0.25	0.32	0.25	0.28	1.09
	(0.25)	(0.26)	(0.17)	(0.18)	(0.18)	(0.51)
adLasso	0.81	0.38	0.51	0.40	0.45	0.72
	(0.25)	(0.35)	(0.19)	(0.23)	(0.19)	(0.35)
procbol	0.81	0.58	0.67	0.62	0.59	0.76
	(0.25)	(0.48)	(0.33)	(0.37)	(0.36)	(0.54)
Lasso+	0.84	0.70	0.51	0.47	0.49	0.39
	(0.23)	(0.29)	(0.12)	(0.12)	(0.11)	(0.18)
adLasso+	0.83	0.71	0.62	0.56	0.60	0.28
	(0.23)	(0.28)	(0.13)	(0.15)	(0.13)	(0.17)
lmmLasso	0.84	0.70	0.51	0.47	0.49	0.39
	(0.23)	(0.29)	(0.12)	(0.11)	(0.11)	(0.18)
pbol+	0.80	0.74	0.75	0.74	0.75	0.18
	(0.23)	(0.29)	(0.11)	(0.15)	(0.11)	(0.16)

Table 5: Results for model M_3. The recovery rate of the true model was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.92(0.20)$. Standard errors are given between parentheses, for 100 runs.

Results	$\hat{J} = J$	$\|\hat{J}\|$	TP	$\hat{\sigma}_e^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_{1,2}$
Ideal	1	5	5	5	1	1	0
Lasso	0.02	5.19	3.24	3.86	-	-	-
		(3.54)	(1.51)	(1.01)	-	-	-
adLasso	0.02	6.88	3.75	3.18	-	-	-
		(3.57)	(1.17)	(0.92)	-	-	-
procbol	0.16	3.38	3.08	4.13		-	-
		(1.32)	(1.22)	(0.76)	-	-	-
Lasso+	0.04	8.33	4.95	1.16	0.98	0.92	0.01
		(2.53)	(0.22)	(0.18)	(0.44)	(0.46)	(0.31)
adLasso+	0.01	9.31	4.88	1.01	0.93	0.89	0.01
		(3.22)	(0.36)	(0.17)	(0.40)	(0.42)	(0.31)
lmmLasso	0.07	8.23	4.96	1.23	0.97	0.92	0.13
		(2.57)	(0.20)	(0.27)	(0.42)	(0.43)	(0.19)
pbol+	0.75	4.8	4.66	1.04	0.97	0.94	0.00
		(0.68)	(0.70)	(0.20)	(0.41)	(0.44)	(0.32)

	$\hat{\beta}_{i_1}$	$\hat{\beta}_{i_2}$	$\hat{\beta}_{i_3}$	$\hat{\beta}_{i_4}$	$\hat{\beta}_{i_5}$	MSE
Ideal	0.75	0.75	0.75	0.75	0.75	0.00
Lasso	0.78	0.24	0.08	0.21	0.18	1.39
	(0.28)	(0.29)	(0.13)	(0.18)	(0.18)	(0.58)
adLasso	0.78	0.38	0.13	0.38	0.32	1.00
	(0.28)	(0.35)	(0.18)	(0.21)	(0.26)	(0.50)
procbol	0.78	0.59	0.25	0.46	0.50	1.14
	(0.28)	(0.51)	(0.38)	(0.43)	(0.43)	(0.62)
Lasso+	0.79	0.69	0.28	0.41	0.41	0.54
	(0.26)	(0.26)	(0.14)	(0.12)	(0.12)	(0.21)
adLasso+	0.78	0.69	0.35	0.53	0.51	0.41
	(0.27)	(0.24)	(0.19)	(0.13)	(0.18)	(0.21)
lmmLasso	0.78	0.69	0.28	0.40	0.40	0.55
	(0.26)	(0.25)	(0.14)	(0.12)	(0.12)	(0.21)
pbol+	0.78	0.74	0.62	0.70	0.69	0.32
	(0.27)	(0.26)	(0.30)	(0.21)	(0.26)	(0.34)

Table 6: Results for model M_4. The recovery rate of the true model was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.83(0.16)$. Standard errors are given between parentheses, for 100 runs.

Results	$\hat{J} = J$	$\lvert\hat{J}\rvert$	TP	$\hat{\sigma}_e^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	
Ideal	1	5	5	5	1	1	1
Lasso	0.22	4.96	4.13	3.32	-	-	
		(2.18)	(1.10)	(0.80)	-	-	
adLasso	0.20	6.10	4.58	2.85	-	-	
		(2.19)	(0.70)	(0.72)	-	-	
procbol	0.28	4.37	4.12	2.90	-	-	
		(1.08)	(0.77)	(0.79)	-	-	
Lasso+	0.20	7.07	4.99	1.11	0.91	0.92	
		(2.01)	(0.10)	(0.22)	(0.36)	(0.46)	
adLasso+	0.24	6.70	4.97	0.97	0.88	0.88	
		(1.51)	(0.17)	(0.19)	(0.34)	(0.45)	
lmmLasso	-	-	-	-	-	-	
		-	-	-	-	-	
pbol+	0.93	5.09	5.00	0.95	0.91	0.89	
		(0.38)	(0.00)	(0.17)	(0.33)	(0.44)	

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	MSE
Ideal	0.67	0.67	0.67	0.67	0.67	0.00
Lasso	0.69	0.69	0.18	0.20	0.27	0.90
	(0.25)	(0.32)	(0.17)	(0.17)	(0.17)	(0.40)
adLasso	0.69	0.68	0.32	0.36	0.46	0.60
	(0.25)	(0.32)	(0.21)	(0.21)	(0.22)	(0.32)
procbol	0.73	0.65	0.48	0.51	0.57	0.63
	(0.34)	(0.13)	(0.36)	(0.36)	(0.35)	(0.42)
Lasso+	0.71	0.71	0.40	0.38	0.43	0.41
	(0.24)	(0.29)	(0.12)	(0.11)	(0.11)	(0.19)
adLasso+	0.71	0.69	0.50	0.48	0.56	0.30
	(0.24)	(0.29)	(0.16)	(0.14)	(0.13)	(0.18)
lmmLasso	-	-	-	-	-	-
	-	-	-	-	-	-
pbol+	0.71	0.69	0.67	0.65	0.68	0.19
	(0.24)	(0.29)	(0.12)	(0.10)	(0.10)	(0.16)

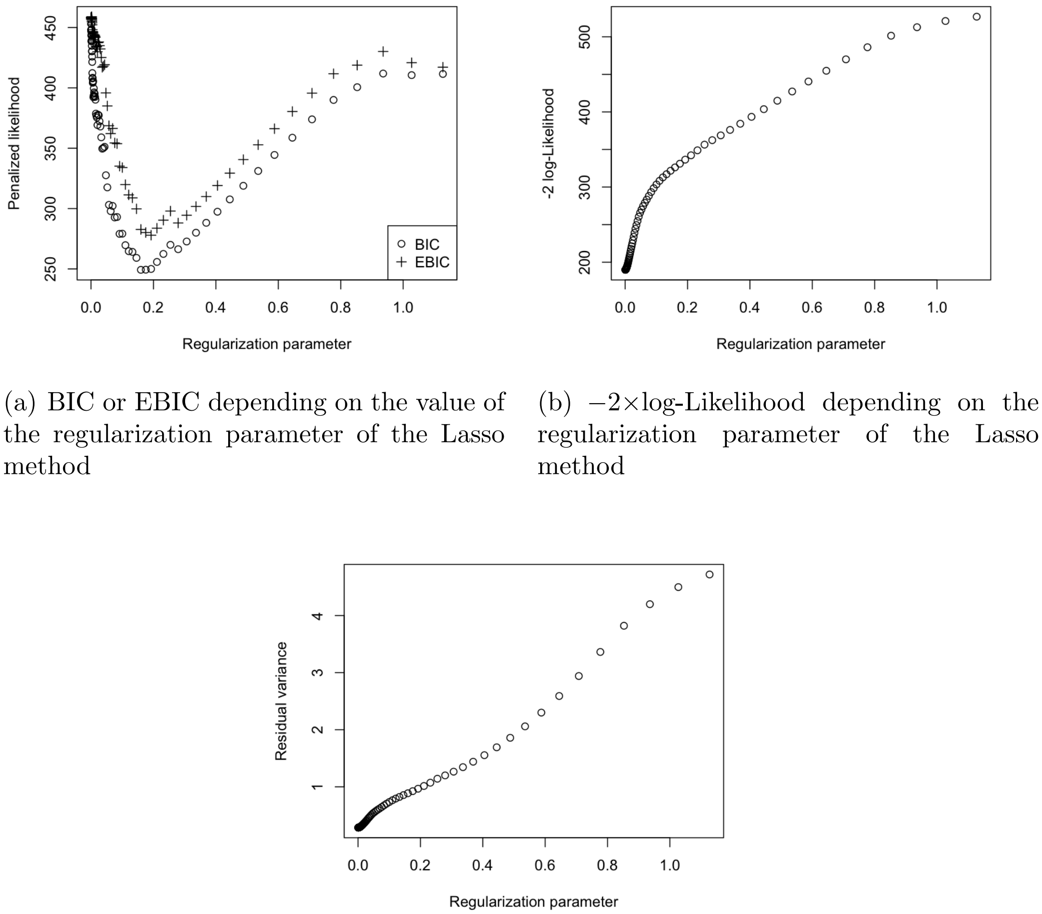
Table 7: Results for model M_2 when a ML linear regression is added after the convergence of the algorithm. The recovery rate of the true model was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.63(0.11)$. Standard errors are given between parentheses, for 100 runs.

	Ideal	lmmLasso	Lasso+
Truth	1	0.30	0.28
$\|\hat{J}\|$	5	6.87(1.91)	6.87(1.89)
TP	5	5.00(0.00)	5.00(0.00)
$\hat{\sigma}_e^2$	1	0.91(0.17)	0.90(0.13)
$\hat{\sigma}_1^2$	1	0.99(0.40)	0.92(0.38)
$\hat{\sigma}_2^2$	1	1.04(0.38)	0.97(0.36)
$\hat{\sigma}_{1,2}^2$	0.5	0.50(0.29)	0.47(0.28)
$\hat{\beta}_1$	0.75	0.81(0.23)	0.81(0.23)
$\hat{\beta}_2$	0.75	0.74(0.29)	0.74(0.29)
$\hat{\beta}_3$	0.75	0.71(0.13)	0.72(0.12)
$\hat{\beta}_4$	0.75	0.72(0.12)	0.72(0.12)
$\hat{\beta}_5$	0.75	0.72(0.13)	0.72(0.13)
mse	0	0.31(0.21)	0.31(0.20)

Appendix B - Remarks on the tuning parameter

In some cases, in particular for the *Lasso* method and the *adLasso* method, tuning of the regularization parameter could become difficult. In this section, we discuss this occurs.

To begin, we shall consider the classical linear model before moving on to the linear mixed model. Let us first examine the Lasso method when only applied in a classical linear model and compare two penalizations of the likelihood: BIC and the Extended BIC (EBIC) (Chen and Chen, 2008). The EBIC penalizes a space of dimension k with a term that depends on the number of spaces that have the same dimension, which is $\frac{p!}{k!(p-k)!}$. Thus EBIC penalizes more the complex spaces than BIC. Figure 2 shows the behavior of the BIC and EBIC criteria, the log-likelihood and the residual variance for various values of the regularization parameter of the Lasso in a low dimensional setting ($p = 80$). As can be observed, tuning the regularization parameter in this setting raises no problems.



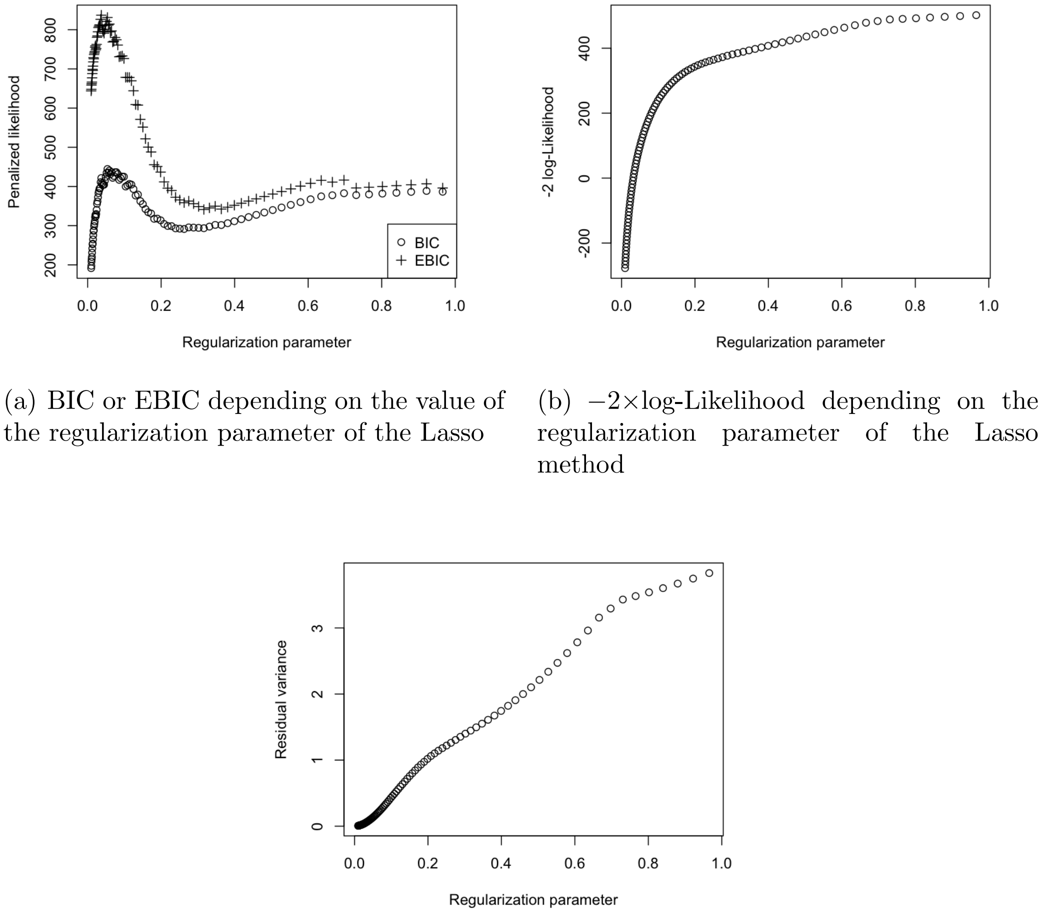
(a) BIC or EBIC depending on the value of the regularization parameter of the Lasso method

(b) $-2\times$log-Likelihood depending on the regularization parameter of the Lasso method

(c) Residual variance depending on the regularization parameter of the Lasso method

Figure 2: One simulation of linear model for the Lasso method with $n = 120, p = 80$ **and** $\beta_J = 1$**.**

Let us now consider a simulation in a high dimensional setting with $n = 120$ observations and $p = 600$ explanatory variables. Results for the regularization parameter of the Lasso are presented in Figure 3 for both methods.



(a) BIC or EBIC depending on the value of the regularization parameter of the Lasso

(b) $-2\times$log-Likelihood depending on the regularization parameter of the Lasso method

(c) Residual variance depending on the regularization parameter of the Lasso method

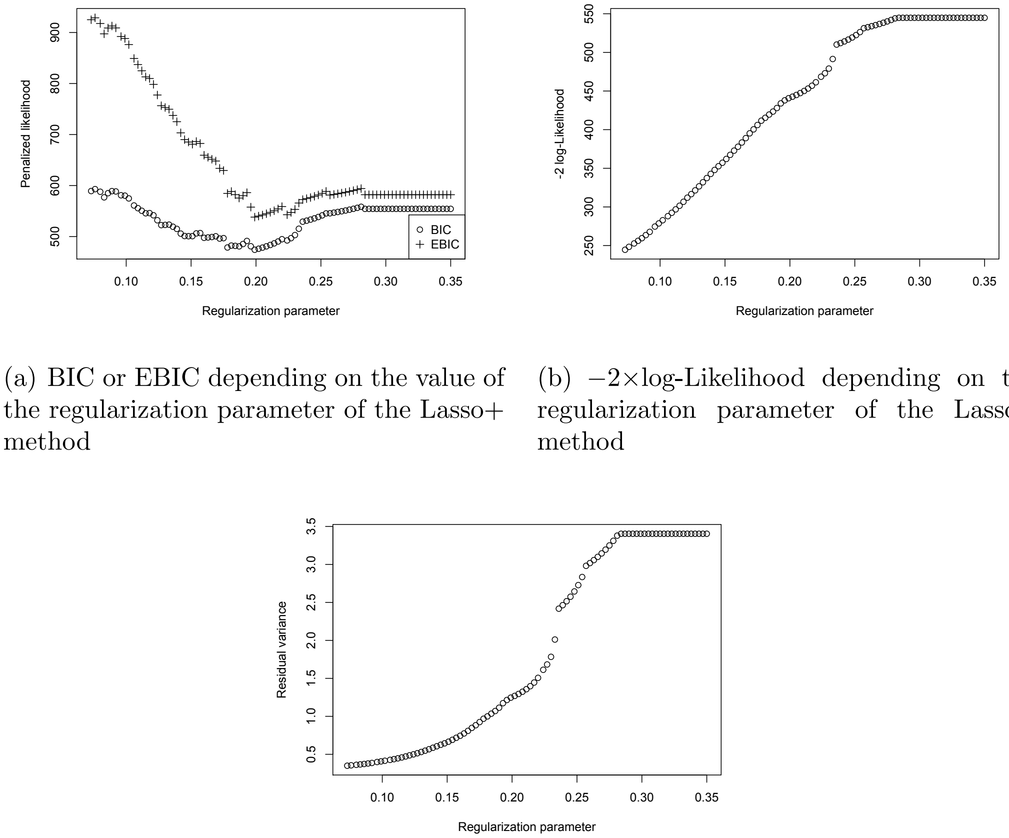
Figure 3: One simulation of linear model for the Lasso method with $n = 120, p = 600$ **and** $\beta_J = 1$**.**

Firstly, we confirm that EBIC is more conservative than BIC and penalizes complex spaces to greater extent. On the far left of Figure 3(a), we observe that both the BIC and the EBIC curves decrease when the regularization parameter is close to zero. This phenomenon is due to the degeneracy of the likelihood as seen in Figure 3(b) (stated in Section 2 for mixed models, this phenomenon also occurs for linear models). Figure 3(c) shows that degeneracy of the likelihood is due to the decrease of residual variance that drops to zero when the regularization parameter is close to zero, and thus when too many variables enter the model.

To conclude, neither BIC nor EBIC penalties are strong enough to completely balance the degeneracy of the likelihood. However, the EBIC penalty does result in selection of a more parsimonious model while BIC penalty selects a more complex model. Nonetheless, the EBIC penalty is usually too much conservative in practice, and this is why the

BIC penalty was used in our simulation study. When degeneracy occurs, when p increases for example, the regularization parameter should be optimized over an interval where the likelihood is more or less stable, i.e. not over the far left part of Figure 3(a) where the criterion decreases.

When the regularization parameter was tuned for the *Lasso+* method, degeneracy of the likelihood was never found to occur in our simulations (Figure 4). However, if it did occur, the same advice as provided above for the classical linear model should be followed.



(a) BIC or EBIC depending on the value of the regularization parameter of the Lasso+ method

(b) $-2\times$log-Likelihood depending on the regularization parameter of the Lasso+ method

(c) Residual variance depending on the regularization parameter of the Lasso+ method

Figure 4: One simulation of *Lasso+* with $n = 120, p = 600, \beta_J = 1$ and two i.i.d. random effects.

Appendix C - Proof of Proposition 2.2

G and R are supposed to be known. Thus the minimization of our objective function g reduces to the minimization of the following function in (β, u):
$h(u,\beta) = (y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u + \lambda|\beta|_1$.
Let us denote $(\hat{u},\hat{\beta}) = \underset{(u,\beta)}{\text{argmin}}\ h(u,\beta)$. Since the function h is convex, we have:

$$(\hat{u},\hat{\beta}) = \begin{cases} u(\beta) = \underset{u}{\text{argmin}}\ h(u,\beta) \\ \hat{\beta} = \underset{\beta}{\text{argmin}}\ h(u(\beta),\beta) \\ \hat{u} = u(\hat{\beta}) \end{cases}$$

Since $\dfrac{\partial h(u,\beta)}{\partial u}$ exists, we can explicit the minimum of h in u:

$$(\hat{u},\hat{\beta}) = \begin{cases} u(\beta) = (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - X\beta) \\ \hat{\beta} = \underset{\beta}{\text{argmin}}\ h(u(\beta),\beta) \\ \hat{u} = u(\hat{\beta}) \end{cases}$$

Thus, we obtain:

$$\begin{aligned} h(u(\beta),\beta) &= (y - X\beta - Zu(\beta))'R^{-1}(y - X\beta - Zu(\beta)) + u'G^{-1}u + \lambda|\beta|_1 \\ &= (y - X\beta)'R^{-1}(y - X\beta) - (y - X\beta)R^{-1}Zu(\beta) - (Zu(\beta))'R^{-1}(y - X\beta) \\ &+ (Z\hat{u})'R^{-1}Zu(\beta) + u(\beta)'G^{-1}u(\beta) + \lambda|\beta|_1 \\ &= (y - X\beta)'[R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}](y - X\beta) + \lambda|\beta|_1 \end{aligned}$$

Denote $W = R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}$. We can show that $W = (Z'GZ + R^{-1})^{-1} = V^{-1}$. This result comes from the equivalence between the resolution of Henderson's equations (Henderson, 1973) and the generalized least squares.

To conclude, we have that

$$(\hat{u},\hat{\beta}) = \left((Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - X\hat{\beta}), \underset{\beta}{\text{argmin}}\ (y - X\beta)'V^{-1}(y - X\beta) + \lambda|\beta|_1\right).$$