

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

Florian Rohart^{a,b,*}, Magali San Cristobal^{a,b,c,d}, Béatrice Laurent^a

^aUMR 5219, Institut de Mathématiques de Toulouse, INSA de Toulouse, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France

^bINRA, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, F-31326 Castanet-Tolosan, France

^cUniversité de Toulouse INPT ENSAT, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, F-31326 Castanet-Tolosan, France

^dUniversité de Toulouse INPT ENVT, UMR1388 Génétique, Physiologie et Systèmes d'Élevage, F-31076 Toulouse, France

Abstract

Linear mixed models are especially useful when observations are grouped. In a high dimensional setting however, selecting the fixed effect coefficients in these models is mandatory as classical tools are not performing well. By considering the random effects as missing values in the linear mixed model framework, a ℓ^1 -penalization on the fixed effects coefficients of the resulting log-likelihood is proposed. The optimization problem is solved via a multicycle Expectation Conditional Maximisation (ECM) algorithm which allows for the number of parameters p to be larger than the total number of observations n and does not require the inversion of the sample $n \times n$ covariance matrix. The proposed algorithm can be combined with any variable selection method developed for linear models. A variant of the proposed approach replaces the ℓ^1 -penalization with a multiple testing procedure for the variable selection aspect and is shown to greatly improve the False Discovery Rate. Both methods are implemented in the MMS R-package, and are shown to give very satisfying results in a high-dimensional simulated setting.

Keywords: linear mixed model, lmmLasso, multiple hypothesis testing, high-dimension

1. Introduction

The more extensive use of new technologies such as high-throughput DNA/RNA chips or RNA sequencing in biology generates an increasing number of highly dimensional data sets where the number of parameters p is much greater than the number of observations n . Consequently, the high dimensional framework generally means that the problem of parameters estimation cannot be solved. In order to address this curse of dimensionality, various constraints have been proposed in linear models. Most of them aim for a parsimonious model where many parameters are set to zero (sparse constraints), or use of a well-conditioned variance matrix on the observations. Many studies have addressed the problem of variable selection by using a linear model of the form $Y = X\beta + \epsilon$, where X is an $n \times p$ matrix containing the observations and

*Corresponding author. f.rohart@uq.edu.au. Present address: Australian Institute for Bioengineering & Nanotechnology (AIBN), the University of Queensland, QLD 4072, Australia.

ϵ is a n -vector of i.i.d random and usually Gaussian variables. One of the oldest methods 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 simple and powerful Lasso (Least Absolute Shrinkage and Selection Operator) method [26] revolutionized the field. The Lasso works by applying a ℓ^1 -penalization on the least squares estimate which shrinks some coefficients to exactly zero. Various extensions exist for the Lasso, for example group Lasso [27], adaptive Lasso [14] and a more stable version known as Bo-Lasso [2]. However, penalizing the likelihood is not the only way to perform variable selection.

Indeed, statistical testing can also be used to determine the relevance of each parameter as in the False Discovery Rate [3, 6], or as in a more recent procedure that appears to provide better results in terms of variable selection [22].

In all the previously described methods, observations are considered to be independent and identically distributed. These methods are therefore no longer appropriate 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, it is block diagonal. In the literature, most reports of linear mixed models relate to the estimation of variance components, using either maximum likelihood estimation (ML) [12, 11], or restricted maximum likelihood estimation (REML) which accounts for the loss in degrees of freedom due to fitting fixed effects [19, 10, 13, 8]. 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. This might be especially the case in high-dimensional analysis. Contrary to linear models, the problem of selecting the fixed effect coefficients in a linear mixed model framework has rarely been addressed in a high dimensional setting.

Both Bondell et al. [5] and Ibrahim et al. [15] used penalized likelihoods to perform selection of both fixed and random effects. Bondell et al. [5] 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. [5] and Ibrahim et al. [15] rely on Cholesky decompositions and, as pointed out by [18], these decompositions are dependent on the order in which the random effects appear and are not permutation invariant [20]. In the present paper, we primarily focus on analyzing data sets with only a few random effects and we therefore do not address the selection of both fixed and random effects.

Schelldorfer et al. [24] 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 .

The objective of this paper is two-fold. The first is to provide a more efficient way to select fixed effects in a linear mixed model. We consider the random effects as missing data, as previously described in Bondell et al. [5] and Foulley [7], and we introduce a ℓ^1 -penalization on the log-likelihood of the complete data. A similar approach is studied in [9] in the framework of Generalized Linear Mixed Models. We propose a multicycle Expectation Conditional Maximization algorithm (ECM) with convergence properties [7, 16, 17] to solve the optimization problem and provide theoretical results when the variances of the observations are known. The second objective is to increase the performance of variable selection. Due to its step de-

sign, the ECM algorithm can be combined with any variable selection method built for linear models. We propose to use a multiple testing procedure introduced in Rohart [22] instead of the ℓ^1 -penalization of the maximum likelihood estimator. We show that this procedure exhibits a higher percentage of recovery of the exact set of variables, a lower false discovery rate and a better estimation of β , which induces a reduced mean squared error. As the selection of fixed effects in a high-dimensional linear mixed model framework has been rarely addressed before, we will mainly compare our results to those of Schelldorfer et al. [24].

The proposed approach is then applied to a real data set from a project in which hundreds of pigs were studied, the aim being to shed light on the relationships between some of the phenotypes of interest and metabolomic data [23]. 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. We first introduce the linear mixed model and its objective function to solve. We then describe the multicycle ECM algorithm used to solve the optimization problem. In Section 3, the algorithm described in Section 2 is extended to be used with any variable selection method developed for linear models. We assess the performance of the approach on a simulation study and demonstrate that the combination of this new algorithm with a multiple testing procedure for variable selection greatly improves the False Discovery Rate (Section 4). Finally, in Section 5, we illustrate the proposed approach on the metabolomic pigs data set.

2. Selection with ℓ^1 -penalization

Let us introduce some notation that will be used throughout the paper. $\text{Var}(\mathbf{a})$ denotes the variance-covariance matrix of the vector \mathbf{a} . For all $a > 0$, let \mathbf{I}_a be the identity matrix of \mathbb{R}^a . For $\mathbf{A} \in \mathbb{R}^{n \times p}$, denote I a subset of $\{1, \dots, n\}$ and J a subset of $\{1, \dots, p\}$. Let $\mathbf{A}_{I,J}$, $\mathbf{A}_{:,J}$ and $\mathbf{A}_{I,:}$ denote submatrices of \mathbf{A} respectively composed of elements of \mathbf{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 $a > 0, b > 0$, denote $\mathbf{0}_a$ to be the vector of size a in which all coordinates are 0 and $\mathbf{0}_{a \times b}$ to be the zero matrix of size $a \times b$. Let us denote $|\mathbf{A}|$ the determinant of the matrix \mathbf{A} .

2.1. Setting-up the linear mixed model

We consider the linear mixed model in which observations are grouped and we assume 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 denote $N_q = qN$.

The linear mixed model can be written as

$$\mathbf{y} = \mathbf{X}\beta + \sum_{k=1}^q \mathbf{Z}_k \mathbf{u}_k + \boldsymbol{\epsilon}, \quad (1)$$

where

- \mathbf{y} is the set of observed data of length n ,
- $\boldsymbol{\beta}$ is an unknown vector of \mathbb{R}^p ; $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$,
- \mathbf{X} is the $n \times p$ matrix of fixed effects; $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$,
- For $k = 1, \dots, q$, $\mathbf{u}_k = (u_k^1, \dots, u_k^N)'$ is a N -vector of i.i.d. coordinates for random effect k . Furthermore, for $i = 1, \dots, N$, the vectors $(u_1^i, u_2^i, \dots, u_q^i)$ are i.i.d. Gaussian $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi} = (\Psi_{i,j})_{1 \leq i, j \leq q}$ is the matrix defined by: $\Psi_{i,j} = \begin{cases} \text{cov}(u_i^1, u_j^1) & \text{if } i \neq j \\ \text{var}(u_i^1) & \text{if } i = j \end{cases}$.
- For $k = 1, \dots, q$, \mathbf{Z}_k is a $n \times N$ incidence matrix (each row of \mathbf{Z}_k contains only one nonzero coefficient),
- $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ is a Gaussian vector with i.i.d. components $\epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$, where σ_e is an unknown positive quantity. We denote by \mathbf{R} the variance-covariance matrix of $\boldsymbol{\epsilon}$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$.

An example of matrices \mathbf{Z}_k for $n = 6$ and two random effects is provided below.

$$\text{Let } \mathbf{Z}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{Z}_2 = \begin{pmatrix} x_1 & 0 & 0 \\ x_2 & 0 & 0 \\ 0 & x_3 & 0 \\ 0 & x_4 & 0 \\ 0 & 0 & x_5 \\ 0 & 0 & x_6 \end{pmatrix}.$$

Note that \mathbf{Z}_2 is the incidence matrix of the interaction of the variable $\mathbf{x} = (x_1, \dots, x_6)$ and the grouping factor.

We denote $\mathbf{u} = (\mathbf{u}'_1, \dots, \mathbf{u}'_q)'$ and \mathbf{Z} the concatenation of $(\mathbf{Z}_1, \dots, \mathbf{Z}_q)$. Setting $\mathbf{G} = \boldsymbol{\Psi} \otimes \mathbf{I}_N$, where \otimes is the Kronecker product, we have that $\mathbf{u} \sim \mathcal{N}_{Nq}(\mathbf{0}, \mathbf{G})$.

One can remark that with these notations, Model (1) can also be written as: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$. In the following, we assume that $\boldsymbol{\epsilon}$ and \mathbf{u} are independent. Thus $\text{Var}(\mathbf{u}, \boldsymbol{\epsilon}) = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$.

We consider the matrices \mathbf{X} and $\{\mathbf{Z}_k\}_{1, \dots, q}$ to be fixed design. Note that our model (1) and the one in [24] 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 , $\boldsymbol{\beta}$, \mathbf{G} and \mathbf{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. ℓ^1 -penalization of the complete log-likelihood

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

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

where

$$-2L_0(\boldsymbol{\beta}, \sigma_e^2; \mathbf{x}) = n \log(2\pi) + n \log(\sigma_e^2) + \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{k=1}^q \mathbf{Z}_k \mathbf{u}_k \right\|^2 / \sigma_e^2, \quad (3a)$$

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

Note that, (2) results from $p(\mathbf{x}|\boldsymbol{\Phi}) = p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_e^2)p(\mathbf{u}|\mathbf{G})$; (3a) from $\boldsymbol{\epsilon}|\sigma_e^2 \sim \mathcal{N}_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ and (3b) from $\mathbf{u}|\mathbf{G} \sim \mathcal{N}_{N_q}(\mathbf{0}, \mathbf{G})$.

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 $\boldsymbol{\beta}$ is sparse (many coefficients are assumed to be zero) and because we want to recover this sparsity, we add a ℓ^1 -penalization on $\boldsymbol{\beta}$ 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 [26] or the lmmLasso method [24]. Thus we consider the following objective function to be minimized:

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

where λ is a positive regularization parameter and $|\boldsymbol{\beta}|_1 = \sum_{j=1}^p |\beta_j|$. It should be noted that the function g could have been obtained in a Bayesian setting by considering a Laplace prior on $\boldsymbol{\beta}$. 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, $L(\boldsymbol{\Phi}; \mathbf{x})$ tends to infinity when $|\mathbf{G}|$ tends towards 0, i.e. when a random effect should not have been included in the model. This is a well-known problem of degeneracy of the likelihood, especially studied in the Gaussian mixture model [4]. Alternatively, instead of treating the random effects as missing values, 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 \mathbf{Y} . This is the case in Schelldorfer et al. [24]:

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

Note that $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. The degeneracy of the likelihood can also appear in the marginal model when the determinant of \mathbf{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 [24] or by λ in (4).

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

2.3. A multicyle ECM algorithm

The multicyle ECM algorithm [17, 7, 16] used to solve the minimization problem of (4) contains four steps: two E steps interlaced with two M steps.

Recall that $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \mathbf{G}, \sigma_e^2)$ is the vector of the parameters to estimate and that $\mathbf{u} = (\mathbf{u}'_1, \dots, \mathbf{u}'_q)'$ is a vector of missing values. The multicyle ECM algorithm is an iterative algorithm. Details can be found in Appendix A. Only a sketch is given below.

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 .

Let us define Q and decompose it as:

$$Q(\Phi; \Phi^{[t]}) = E_{u|y, \Phi = \Phi^{[t]}}[g(\Phi; \mathbf{x})] = Q_0(\beta, \sigma_e^2; \Phi^{[t]}) + Q_1(\mathbf{G}; \Phi^{[t]}),$$

where:

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

and Q_1 does not depend on β nor σ_e^2 :

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

2.3.1. First E-step

As designated by Henderson [12], $\mathbf{u}^{[t+1/2]} = E(\mathbf{u}|y, \Phi = \Phi^{[t]})$ is the BLUP (Best Linear Unbiased Prediction) of \mathbf{u} for the vector of parameters Φ equal to $\Phi^{[t]}$:

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

2.3.2. M-Step for β

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

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

Note that this can be viewed as a Lasso estimator on β with the vector of ‘‘observed’’ data $(\mathbf{y} - \mathbf{Z}\mathbf{u}^{[t+1/2]})$ 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 \mathbf{u} :

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

2.3.4. M-step for (\mathbf{G}, σ_e^2)

Variance matrices \mathbf{G} and \mathbf{R} are updated based on the minimization of Q_1 and Q_0 , respectively. Let us recall that $\mathbf{G} = \Psi \otimes \mathbf{I}_N$. Thanks to a lemma reported in [1], the minimization of Q_1 with respect to Ψ gives:

$$\Psi_{i,j}^{[t+1]} = \frac{1}{N} \left[\mathbf{u}_i^{[t+1]'} \mathbf{u}_j^{[t+1]} + \text{tr}(\mathbf{T}_{i,j}) \sigma_e^{2[t]} \right],$$

with $\mathbf{T}_{i,j}$ the (i, j) element of $\mathbf{T} = (\mathbf{Z}'\mathbf{Z} + \sigma_e^2\mathbf{G}^{-1})^{-1}$, and $\mathbf{u}_j^{[t+1]}$ the N -vector associated to the random effect j in $\mathbf{u}^{[t+1]}$.

The minimization of Q_0 with respect to σ_e^2 leads to:

$$\sigma_e^{2[t+1]} = \frac{1}{n} \left[\left\| \mathbf{y} - \mathbf{X}\beta^{[t+1]} - \mathbf{Z}\mathbf{u}^{[t+1]} \right\|^2 + \sigma_e^{2[t]} (N_q - \sigma_e^{2[t]} \text{tr}(\mathbf{T}\mathbf{G}^{-1[t]})) \right].$$

In summary, the algorithm can be detailed as follows:

Algorithm 2.1 (Lasso+). *Initialization*

Initialize the set of parameters $\Phi^{[0]} = (\beta^{[0]}, \mathbf{G}^{[0]}, \sigma_e^{2[0]})$.

Define \mathbf{Z} as the concatenation of $\mathbf{Z}_1, \dots, \mathbf{Z}_q$ and $\mathbf{u} = (\mathbf{u}'_1, \dots, \mathbf{u}'_q)'$.

Until convergence

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

2. M-step: $\beta^{[t+1]} = \underset{\beta}{\text{Argmin}} \left(\left\| (\mathbf{y} - \mathbf{Z}\mathbf{u}^{[t+1/2]}) - \mathbf{X}\beta \right\|^2 + \lambda \sigma_e^{2[t]} \|\beta\|_1 \right)$

3. E-step: $\mathbf{u}^{[t+1]} = (\mathbf{Z}'\mathbf{Z} + \sigma_e^{2[t]}\mathbf{G}^{-1[t]})^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta^{[t+1]})$

4. M-step:

(a) $\Psi_{i,j}^{[t+1]} = \frac{1}{N} \left[\mathbf{u}_i^{[t+1]'} \mathbf{u}_j^{[t+1]} + \text{tr}(\mathbf{T}_{i,j}) \sigma_e^{2[t]} \right]$ and $\mathbf{G}^{[t+1]} = \Psi^{[t+1]} \otimes \mathbf{I}_N$

(b) $\sigma_e^{2[t+1]} = \frac{1}{n} \left[\left\| \mathbf{y} - \mathbf{X}\beta^{[t+1]} - \mathbf{Z}\mathbf{u}^{[t+1]} \right\|^2 + \sigma_e^{2[t]} \left(N_q - \sigma_e^{2[t]} \text{tr}(\mathbf{T}\mathbf{G}^{-1[t]}) \right) \right]$

end

Convergence of Algorithm 2.1 is ensured as it is a multicycle ECM algorithm [17].

Three stopping criteria are used to stop the convergence process of the algorithm: a first criterion based on $\|\beta^{[t+1]} - \beta^{[t]}\|^2$, a second based on $\|\mathbf{u}_k^{[t+1]} - \mathbf{u}_k^{[t]}\|^2$ for each random effect \mathbf{u}_k and lastly a criterion based on $\|L(\Phi^{[t+1]}, \mathbf{x}) - L(\Phi^{[t]}, \mathbf{x})\|^2$ where $L(\Phi, \mathbf{x})$ 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: $\mathbf{G}^{[0]}$ is the block diagonal matrix of $\sigma_1^{2[0]}\mathbf{I}_N, \dots, \sigma_q^{2[0]}\mathbf{I}_N$ 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 [28], one last step can be added 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 parameter Φ using a classic mixed model estimation based on the model:

$$\mathbf{y} = \mathbf{X}\beta_{\hat{J}} + \sum_{1 \leq k \leq q} \mathbf{Z}_k \mathbf{u}_k + \epsilon,$$

where \hat{J} is the estimated set of indices of the relevant fixed effects. Such a final reestimation step in a Lasso-type approach for variable selection in Generalized Linear Mixed Models has also been considered in [9].

Proposition 2.2. *When the variance components σ_e^2 and \mathbf{G} are known, minimization of the objective function (4) is equivalent to the minimization of $Q(\beta) = (\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) + \lambda \|\beta\|_1$, which is the objective function described in [24] with known variances.*

In [24], the authors provided theoretical results regarding the consistency of their method. Based on Proposition 2.2, these results apply to our method in the case of known variances. The assumptions of Proposition 2.2 are strong, but there exists practical situations in genetics (for example with well known traits) where the variance components can be considered known. Proof of Proposition 2.2 is provided in Appendix A.

2.4. Tuning parameter

The solution depends on the regularization parameter λ , included in Algorithm 2.1, that controls shrinkage. This parameter has to be tuned. To that end, we choose to use the Bayesian Information Criterion (BIC) [25]:

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

where $V_\lambda = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}' + \hat{\sigma}_e^2\mathbf{I}_n$ and $\hat{\mathbf{G}}, \hat{\sigma}_e^2, \hat{\boldsymbol{\beta}}_\lambda$ 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 variable selection method developed for linear models.

3. Generalizations of the selection procedure

3.1. Generalizing the algorithm

Algorithm 2.1 provides good results, as demonstrated in the simulation study in Section 4. Nevertheless, because the aim of the second step of the algorithm is to select the relevant coefficients of $\boldsymbol{\beta}$ 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 [29] or the elastic net [30], 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 $\boldsymbol{\Phi}^{[0]} = (\boldsymbol{\beta}^{[0]}, \mathbf{G}^{[0]}, \sigma_e^{2[0]})$.

Define \mathbf{Z} *as the concatenation of* $\mathbf{Z}_1, \dots, \mathbf{Z}_q$ *and* $\mathbf{u} = (\mathbf{u}'_1, \dots, \mathbf{u}'_q)'$.

Until convergence

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

2. *Variable selection and estimation of* $\boldsymbol{\beta}$ *in the linear model* $\mathbf{y} - \mathbf{Z}\mathbf{u}^{[t+1/2]} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^{[t]}$, *where* $\boldsymbol{\epsilon}^{[t]} \sim \mathcal{N}(\mathbf{0}, \sigma_e^{2[t]}\mathbf{I}_n)$.

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

4. (a) $\boldsymbol{\Psi}_{i,j}^{[t+1]} = \frac{1}{N} \left[\mathbf{u}_i^{[t+1]'} \mathbf{u}_j^{[t+1]} + \operatorname{tr}(\mathbf{T}_{i,j})\sigma_e^{2[t]} \right]$ *and* $\mathbf{G}^{[t+1]} = \boldsymbol{\Psi}^{[t+1]} \otimes \mathbf{I}_N$

- (b) $\sigma_e^{2[t+1]} = \frac{1}{n} \left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathbf{Z}\mathbf{u}^{[t+1]}\|^2 + \sigma_e^{2[t]} \left(N_q - \sigma_e^{2[t]} \operatorname{tr}(\mathbf{T}\mathbf{G}^{-1[t]}) \right) \right]$

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 *mht.bol* method [22]. The *mht.bol* method sequentially tests multiple hypotheses and determines statistically the set of relevant variables in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

where ϵ is an i.i.d Gaussian noise. This method consists of two steps. First, variables are ordered taking into account the observations \mathbf{y} and then, in the second step, multiple hypotheses are tested to distinguish between relevant and irrelevant variables. The *mht.bol* method has proved to be powerful under certain conditions as reported in [22].

3.2. Generalizing the model to different grouping variables

Assume that there are q random effects corresponding to q grouping factors ($q \geq 1$), where some grouping factors may be identical. The levels of the factor k are denoted $\{1, 2, \dots, N_k\}$. The i^{th} -observation belongs to the groups (i_1, \dots, i_q) , where for all $l = 1, \dots, q$, $i_l \in \{1, 2, \dots, N_l\}$. 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=1}^q \mathbf{Z}_k \mathbf{u}_k + \boldsymbol{\epsilon}, \quad (5)$$

the differences with model (1) being that

- For $k = 1, \dots, q$, \mathbf{u}_k is a N_k -vector of the random effect for grouping factor k ,
- For $k = 1, \dots, q$, \mathbf{Z}_k is a $n \times N_k$ incidence matrix for grouping factor k .

Both Algorithms 2.1 and 3.1 apply for Model (5) when random effects are considered to be independent. We make the common assumption of a null covariance matrix between two random effects pertaining to two different grouping variables. The matrix \mathbf{G} is therefore block-diagonal, each block corresponding to one grouping variable. Within each block, a covariance structure $\boldsymbol{\Psi}$ can be considered as previously. A particular case is when all random effects correspond to different grouping variables, implying a diagonal \mathbf{G} matrix. This will be the case in the application to real data considered in Section 5.

A naive selection of the random effects can be performed when the variance of a random effect drops close to 0. When $\Psi_{k,k}$ is too small at some step t of the ECM algorithm, the random effect \mathbf{u}_k is removed from the model.

In Section 4, we show that the combination of Algorithm 3.1 and the *mht.bol* 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). Our methods are implemented into the R language [21] through the *MMS* package, available on CRAN (<http://cran.r-project.org/>).

4.1. Methods used

We compare four methods, the *lmmLasso* method of [24], *Algorithm 2.1* (designated as *Lasso+*), the combination of the *adLasso* method [29] and *Algorithm 3.1* (designated as *ad-Lasso+*) and the combination of the *mht.bol* procedure [22] and *Algorithm 3.1* (designated as *mht.bol+*). The *lmmLasso* method was computed using the *lmmLasso* R-package, available on CRAN (<http://www.cran.r-project.org>).

The initial weights of the *adLasso+* are 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 $\mathbf{y} = \mathbf{X}_i\beta_i + \boldsymbol{\epsilon}$.

The second step of the *mht.bol* method performs multiple hypothesis testing via an estimation of unknown quantiles related to the matrix \mathbf{X} . Computing these quantiles at each iteration of the convergence process would make the combination of the *mht.bol* method and *Algorithm 3.1* almost impossible to run. However, the quantiles remain unchanged throughout the algorithm since no changes occur in the data matrix \mathbf{X} . The *mht.bol* 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 *mht.bol* 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 *mht.bol* method) is set to 10. The number of variables arranged in order during the first step of the *mht.bol* method is set to 40. The *mht.bol+* method is 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.

4.2. Simulation study design

We set \mathbf{X}_1 to be the vector of \mathbb{R}^n in which coordinates are all equal to 1 and then consider four models. For each model, the response variable y is computed via $\mathbf{y} = \sum_{j=1}^5 \mathbf{X}_{i_j}\boldsymbol{\beta}_{i_j} + \sum_{k=1}^q \mathbf{Z}_k\mathbf{u}_k + \boldsymbol{\epsilon}$, where $J = \{i_1, \dots, i_5\} \subset \{1, \dots, p\}$, with q random effects being Gaussian and $\boldsymbol{\epsilon}$ being a vector of independent standard Gaussian variables ($\sigma_{\boldsymbol{\epsilon}}^2 = 1$). For each model, we have for all $j = 2, \dots, p$: $\sum_{i=1}^n X_{j,i} = 0$ and $\frac{1}{n} \sum_{i=1}^n X_{j,i}^2 = 1$. For $k = 1, \dots, q$, the random effects regression matrix \mathbf{Z}_k corresponds to the design matrix of the interaction between the k^{th} column of \mathbf{X} and the grouping factor, which gives a $n \times N$ matrix. The design of the matrices \mathbf{Z}_k 's means that the q grouping variables generates both a fixed effect (for the β_k 's) and a random effect (for the \mathbf{u}_k 's). As recommended in Schelldorfer et al. [24], 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 : $n = 120$, $p = 80$, $\boldsymbol{\beta}_J = 3/4$, $q = 3$ and $\boldsymbol{\Psi} = \mathbf{I}_3$. For all $j = 2, \dots, p$, \mathbf{X}_j are i.i.d. with $\mathbf{X}_j \sim \mathcal{N}_n(0, \mathbf{I}_n)$. $N = 20$ and $\forall i \in \{1, \dots, 20\} n_i = 6$
- M_2 : $n = 120$, $p = 300$, $\boldsymbol{\beta}_J = 3/4$, $q = 2$ with $\text{var}(\mathbf{u}_1) = \text{var}(\mathbf{u}_2) = 1$ and $\text{cov}(\mathbf{u}_1, \mathbf{u}_2) = 0.5$. The covariates are generated from a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$ with the pairwise correlation $\Sigma_{kk'} = \rho^{|k-k'|}$ and $\rho = 0.5$. $N = 20$ and $\forall i \in \{1, \dots, 20\} n_i = 6$
- M_3 : $n = 120$, $p = 600$, $\boldsymbol{\beta}_J = 3/4$, $q = 2$ and $\boldsymbol{\Psi} = \mathbf{I}_2$. The covariates are generated from a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$ with the pairwise correlation $\Sigma_{kk'} = \rho^{|k-k'|}$ and $\rho = 0.5$. $N = 20$ and $\forall i \in \{1, \dots, 20\} n_i = 6$

- M_4 : $n = 120$, $p = 600$, $\beta_j = 3/4$, $q = 2$ and $\Psi = I_2$. For all $j = 2, \dots, p$, X_j are i.i.d. with $X_j \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$. The two grouping variables are different: $N_1 = 20, \forall i \in \{1, \dots, 20\} n_{i,1} = 6$ and $N_2 = 15, \forall i \in \{1, \dots, 15\} n_{i,2} = 8$

For all models we set $J = \{1, 2, i_3, i_4, i_5\}$ where $\{i_3, i_4, i_5\} \subset \{3, \dots, p\}$; in addition, $i_3 = 3$ for model M_1 . In each model, the matrices \mathbf{X} and \mathbf{Z}_k are generated once and for all. A hundred replications are performed by generating different random effects \mathbf{u}_k and different random noise ϵ according to the settings defined for $M_1 - M_4$.

The designs $M_1 - M_4$ were chosen to study the influence of several settings on the performance of the variable selection procedures: effect of the dimension p of the vector β , effect of the correlation between the random effects, effect of a linear dependency between the covariates and the effect of various grouping variables for the different random effects. Model M_1 is the simplest one with a small dimension ($p < n$) and independent random effects. The covariates X_j are quasi-orthogonal: they are generated once and for all by independent $\mathcal{N}_n(0, \mathbf{I}_n)$. Model M_2 addresses the high dimension ($p > n$) and correlated random effects. Compared with M_1 , the way the covariates are generated induces linear dependences between them; these dependences make the estimation of β harder. In model M_3 , the dimension is even higher, but the random effects are independent. Model M_4 addresses the question of different grouping factors.

The aim is to recover the set of relevant fixed effects coefficients J for each model as well as to estimate these coefficients and the variance matrix of both the random effects and residuals. To evaluate the quality of the methods, we use several criteria: the proportion of runs where the true set of variables is recovered ($\hat{J} = J$), the proportion of runs where the true set of variables is included in the selected set ($J \subset \hat{J}$), the proportion TP of true variables that are recovered ($|\hat{J} \cap J|/|\hat{J}|$) as well as the false discovery rate FDR ($|J^c \cap \hat{J}|/|\hat{J}|$), the estimation of Φ and the mean squared error mse calculated as an ℓ^2 error rate between the real value $\mathbf{X}\beta$ and the estimation $\hat{\mathbf{X}}\hat{\beta}$. We also determined the Signal-to-Noise Ratio (SNR) as $\|\mathbf{X}\beta\|_2^2 / \|\sum_{k=1}^q \mathbf{Z}_k \mathbf{u}_k + \epsilon\|_2^2$ for each of the replications.

4.3. Comments on the results

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

For all models, *lmmLasso* and *Lasso+* gave almost identical results. This is not really surprising since both methods are based on a ℓ^1 -penalization of the log likelihood. The result in terms of true model recovered were not satisfactory for these methods because the False Discovery Rate (FDR) is too high. However, both methods exhibited a high percentage of ' $J \subset \hat{J}$ ' related to a high number of True Positive (TP).

Regarding the *adLasso+* method, it provided a better mse result than the *Lasso+* method, but in the meantime the FDR was higher and the number of TP was lower.

The best results were obtained when Algorithm 3.1 was combined with the *mht.bol* method (*mht.bol+*). This combination provided by far the greatest percentage of true model recovered, the percentage of TP was slightly lower than for the other methods but the FDR was much lower and close to 0. In addition, the estimated fixed effects were the closest to real values and the

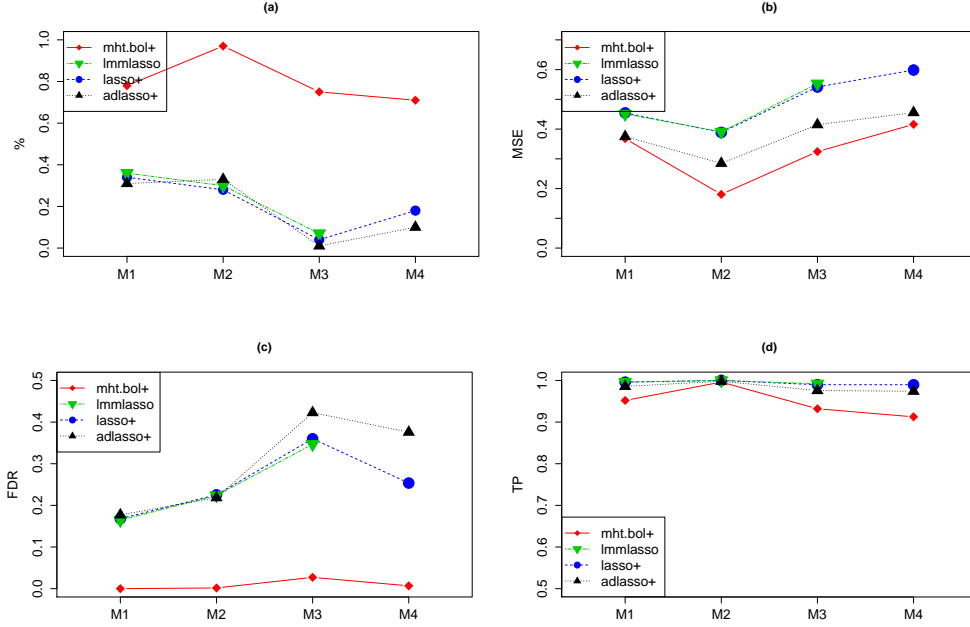


Figure 1: Summary of the results of the simulation study for models $M_1 - M_4$ (X axis). For each model: (a) Empirical probability of recovering the exact set of parameters ($\hat{J} = J$), (b) Mean Squared Error, (c) False Discovery Rate, and (d) proportion of true positives.

mse was 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 B.6 in Appendix B). Interestingly, the *mht.bol+* method always converged in our simulations.

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 $\Phi = (\beta, G, \sigma_e^2)$, as mentioned previously in Section 2.3.

We tested different initializations 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 [23]. The aim of this analysis is to pinpoint metabolomic data that describe a phenotype taking into account all 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:

$$\mathbf{y} = \mathbf{X}_B\boldsymbol{\beta}_B + \mathbf{X}_M\boldsymbol{\beta}_M + \mathbf{Z}_E\mathbf{u}_E + \mathbf{Z}_F\mathbf{u}_F + \boldsymbol{\epsilon}, \quad (6)$$

where \mathbf{y} is the DFI phenotype and $\mathbf{X}_B, \mathbf{X}_M, \mathbf{Z}_E, \mathbf{Z}_F$ are the design matrices of the breed effect, the metabolomic data, the batch effect and the family effect, respectively. As mentioned in the Introduction, \mathbf{X}_B is of size 506×3 , \mathbf{X}_M is of size 506×375 , \mathbf{Z}_E is of size 506×8 and \mathbf{Z}_F is of size 506×157 .

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 \mathbf{G} the block diagonal matrix with blocks $\sigma_E^2 \mathbf{I}_{N_1}$ and $\sigma_F^2 \mathbf{I}_{N_2}$, 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 $\boldsymbol{\beta}_B$ do not undergo feature selection.

We compare several methods using the linear mixed model defined by (6): *Lasso+*, *adLasso+* and *mht.bol+* (see Section 4). These methods are benchmarked to some feature selection procedures developed for classical linear models, such as *Lasso*, *adLasso* and *mht.bol*, to investigate the potential advantages of including random effects on the real data. In the case of linear model, the methods are fitted to $\mathbf{y} = \mathbf{X}_B\boldsymbol{\beta}_B + \mathbf{X}_M\boldsymbol{\beta}_M + \boldsymbol{\epsilon}$.

Both methods *mht.bol* and *mht.bol+* were set with a user-level of $\alpha = 0.1$. The results are presented in Table 1.

	$ \hat{J} $	$\hat{\sigma}_e^2$	$\hat{\sigma}_E^2$	$\hat{\sigma}_F^2$
Lasso	14	3.8×10^{-2}	-	-
adLasso	21	3.4×10^{-2}	-	-
mht.bol	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}
mht.bol+	5	3.4×10^{-2}	5.9×10^{-3}	6.5×10^{-3}

Table 1: Parameters estimation for the real data set

When random effects were considered, we observed a decrease in both the residual variance and the number of selected metabolomic variables.

The computational time for one run when only the batch effect is considered (in order to compute the *lmmLasso*) was 0.80 seconds for the *Lasso+* method and 24.28 seconds for the *lmmLasso* method. As can be seen, when a large number of observations were included, the *Lasso+* method was 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 8.00Gb of RAM with a regularization parameter that selects the same model for both methods.

Further investigation will include biological analysis and interpretation of the variables selected by the different models.

6. Conclusion

In this paper, we proposed to add a ℓ^1 -penalization on the complete log-likelihood to allow the selection of the fixed effects in a linear mixed model. A multicycle ECM algorithm was used to minimize the objective function. This algorithm provides the same results as the ImmLasso method described in [24], but is much faster to run than the latter. Theoretical results obtained in this paper were identical to those found in [24] when the variances are known. One advantage of our algorithm is that it can be combined with any variable selection method built for linear models. The combined *mht.bol* method gave good results when tested on simulated data and outperformed other approaches. It selected most of the relevant variable and had a very small False Discovery Rate.

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 BIOPORC. We thank Helen Munduteguy, Dr Alan Huang and Dr Kim-Anh Lê-Cao for the English revision of the manuscript. We thank the referees for their constructive comments that allowed us to improve the quality paper.

- [1] Anderson, T., 1984. An introduction to multivariate analysis. Wiley Series in Probability and Statistics.
- [2] Bach, F., 2009. Model-consistent sparse estimation through the bootstrap. Technical report, hal-00354771, version 1.
- [3] Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Hypothesis Testing. *Journal of the Royal Statistical Society B* 57, 289-300.
- [4] Biernacki, C., Chrétien, S., 2003. Degeneracy in the Maximum Likelihood Estimation of Univariate Gaussian Mixtures with EM. *Statistics & Probability Letters* 61, 373-382.
- [5] Bondell, H.D., Krishna, A., Ghosh, S.K., 2010. Joint variable selection of fixed and random effects in linear mixed-effects models. *Biometrics* 66, 1069-1077.
- [6] Bunea, F., Wegkamp, M., Auguste, A., 2006. Consistent variable selection in high dimensional regression via multiple testing. *J. Statist. Plann. Inference* 136, 4349-4363.
- [7] Foulley, J., 1997. ECM approaches to heteroskedastic mixed models with constant variance ratios. *Genetics Selection Evolution* 29, 197-318.
- [8] Foulley, J.L., Delmas, C., Robert-Granié, C., 2006. Méthodes du maximum de vraisemblance en modèle linéaire mixte. *J. SFdS* 1-2, 5-52.
- [9] Groll, A., Tutz, G., 2014. Variable selection for generalized linear mixed models by ℓ^1 -penalized estimation. *Stat. Comput.* 24, 137-154.
- [10] Harville, D., 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* 72, 320-340.
- [11] Henderson, C., 1953. Estimation of Variance and Covariance Components. *Biometrics* 9, 226-252.
- [12] Henderson, C., 1973. Sire evaluation and genetic trends. *Journal of Animal Science*, 10-41.
- [13] Henderson, C., 1984. Applications of linear models in Animal breeding. University of Guelph, Ont.
- [14] Huang, J., Ma, S., Zhang, C.H., 2008. Adaptive Lasso for sparse high-dimensional regression models. *Stat. Sin.* 18, 1603-1618.
- [15] Ibrahim, J.G., Zhu, H., Garcia, R.I., Guo, R., 2011. Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics* 67, 495-503.
- [16] McLachlan, J., Krishnan, T., 2008. The EM Algorithm and Extensions, second edition. Wiley-Interscience.
- [17] Meng, X.L., Rubin, D.B., 1993. Maximum Likelihood Estimation via the ECM Algorithm: A general Framework. *Biometrika* 80, 267-278.
- [18] Müller, S., Seealy, J., Welsh, A., 2013. Model selection in linear mixed model. *Statist. Sci.* 28, 135-281.
- [19] Patterson, H., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545-554.
- [20] Pourahmadi, M., 2011. Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* 26, 369-387.

- [21] R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.
- [22] Rohart, F., 2011. Multiple Hypotheses Testing For Variable Selection. arXiv:1106.3415v1 .
- [23] 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., San-Cristobal, M., 2012. Phenotypic prediction based on metabolomic data on the growing pig from three main European breeds. *Journal of Animal Science* .
- [24] Schelldorfer, J., Bühlmann, P., van de Geer, S., 2011. Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1 -Penalization. *Scand. J. Stat.* 38, 197–214.
- [25] Schwarz, G., 1978. Estimating the dimension of a model. *Ann. Statist* 6, 461–464.
- [26] Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. B* 58, 267–288.
- [27] Yuan, M., Lin, Y., 2007. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* 68, 46–67.
- [28] Zhang, C.H., Hunag, J., 2008. The sparsity and bias of the Lasso selection in high-Dimensional linear regression. *Ann. Statist.* 36, 1567–1594.
- [29] Zou, H., 2006. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* 101 101, 1418–1429.
- [30] Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *J.R. Statist. Soc. B* 67, 301–320.

Appendix A. The ECM algorithm

Appendix A.1. First E-step

By definition,

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

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

$$E_{\mathbf{u}|\mathbf{y}, \Phi = \Phi^{[t]}}(\boldsymbol{\epsilon}' \boldsymbol{\epsilon}) = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t]} - \mathbf{Z}E(\mathbf{u}|\mathbf{y}, \Phi = \Phi^{[t]}) \right\|^2 + \text{tr}(\mathbf{Z}\text{Var}(\mathbf{u}|\mathbf{y}, \Phi^{[t]})\mathbf{Z}'). \quad (\text{A.1})$$

Let us denote $\mathbf{u}^{[t+1/2]} = E(\mathbf{u}|\mathbf{y}, \Phi = \Phi^{[t]})$. As designated by Henderson [12], $\mathbf{u}^{[t+1/2]}$ is the BLUP (Best Linear Unbiased Prediction) of \mathbf{u} for the vector of parameters Φ equal to $\Phi^{[t]}$. According to Henderson [12],

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

Appendix A.2. M-Step for $\boldsymbol{\beta}$

See main text.

Appendix A.3. Second E-Step

See main text.

Appendix A.4. M-step for (\mathbf{G}, σ_e^2)

Variance matrices \mathbf{G} and \mathbf{R} are updated based on the minimization of Q_1 and Q_0 , respectively. Let us recall that $\mathbf{G} = \boldsymbol{\Psi} \otimes \mathbf{I}_N$. We can therefore write $Q_1(\mathbf{G}; \Phi^{[t]}) = N_q \log(2\pi) + N \log(|\boldsymbol{\Psi}^{[t]}|) + \text{tr}(\boldsymbol{\Psi}^{-1[t]}\boldsymbol{\Omega}^{[t]})$, where $\boldsymbol{\Omega}^{[t]} = \{\omega_{i,j}^{[t]} = E(\mathbf{u}'_i \mathbf{u}_j | \mathbf{y}, \Phi = \Phi^{[t]})\}$. Thanks to the lemma reported in [1], the minimization of Q_1 with respect to $\boldsymbol{\Psi}$ gives $\boldsymbol{\Psi}^{[t+1]} = \boldsymbol{\Omega}^{[t]}/N$. Thus, for all $1 \leq i, j \leq q$, $\Psi_{i,j}^{[t+1]} = E(\mathbf{u}'_i \mathbf{u}_j | \mathbf{y}, \mathbf{G}^{[t]}, \sigma_e^{2[t]}, \boldsymbol{\beta}^{[t+1]})/N$.

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

$$E(\mathbf{u}'_i \mathbf{u}_j | \mathbf{y}, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \boldsymbol{\beta}^{[t+1]}) = \mathbf{u}_i^{[t+1]'} \mathbf{u}_j^{[t+1]} + \sum_{k=1}^N \text{cov}_{\mathbf{u}|\mathbf{y}, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \boldsymbol{\beta}^{[t+1]}}(\mathbf{u}_i^k, \mathbf{u}_j^k).$$

Moreover, we can use the following results of [12] that gives us the expression of the covariance matrix $\text{cov}_{\mathbf{u}|\mathbf{y}, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \boldsymbol{\beta}^{[t+1]}}(\mathbf{u}_i, \mathbf{u}_j)$ from which the coefficient $\text{cov}_{\mathbf{u}|\mathbf{y}, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \boldsymbol{\beta}^{[t+1]}}(\mathbf{u}_i^k, \mathbf{u}_j^k)$ can be extracted,

$$\text{cov}_{\mathbf{u}|\mathbf{y}, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \boldsymbol{\beta}^{[t+1]}}(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{T}_{i,j} \sigma_e^{2[t]},$$

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

$$\begin{aligned} (\mathbf{Z}'\mathbf{Z} + \sigma_e^{2[t]}\mathbf{G}^{-1[t]})^{-1} &= \begin{pmatrix} \mathbf{Z}'_1\mathbf{Z}_1 + \sigma_e^{2[t]}\Psi^{1,1[t]}\mathbf{I}_N & \mathbf{Z}'_1\mathbf{Z}_2 + \sigma_e^{2[t]}\Psi^{1,2[t]}\mathbf{I}_N & \dots & \mathbf{Z}'_1\mathbf{Z}_q + \sigma_e^{2[t]}\Psi^{1,q[t]}\mathbf{I}_N \\ \mathbf{Z}'_2\mathbf{Z}_1 + \sigma_e^{2[t]}\Psi^{2,1[t]}\mathbf{I}_N & \mathbf{Z}'_2\mathbf{Z}_2 + \sigma_e^{2[t]}\Psi^{2,2[t]}\mathbf{I}_N & \dots & \mathbf{Z}'_2\mathbf{Z}_q + \sigma_e^{2[t]}\Psi^{2,q[t]}\mathbf{I}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}'_q\mathbf{Z}_1 + \sigma_e^{2[t]}\Psi^{q,1[t]}\mathbf{I}_N & \mathbf{Z}'_q\mathbf{Z}_2 + \sigma_e^{2[t]}\Psi^{q,2[t]}\mathbf{I}_N & \dots & \mathbf{Z}'_q\mathbf{Z}_q + \sigma_e^{2[t]}\Psi^{q,q[t]}\mathbf{I}_N \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \mathbf{T}_{1,1} & \mathbf{T}_{1,2} & \dots & \mathbf{T}_{1,q} \\ \mathbf{T}'_{1,2} & \mathbf{T}_{2,2} & \dots & \mathbf{T}_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{T}'_{1,q} & \mathbf{T}'_{2,q} & \dots & \mathbf{T}_{q,q} \end{pmatrix}, \end{aligned}$$

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,

$$\Psi_{i,j}^{[t+1]} = \frac{1}{N} \left[\mathbf{u}_i^{[t+1]'} \mathbf{u}_j^{[t+1]} + \text{tr}(\mathbf{T}_{i,j}) \sigma_e^{2[t]} \right].$$

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

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

Let us denote $\mathbf{T} = (\mathbf{Z}'\mathbf{Z} + \sigma_e^{2[t]}\mathbf{G}^{-1[t]})^{-1}$. Since

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

we have

$$\sigma_e^{2[t+1]} = \frac{1}{n} \left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{[t+1]} - \mathbf{Z}\mathbf{u}^{[t+1]}\|^2 + \sigma_e^{2[t]} (N_q - \sigma_e^{2[t]}\text{tr}(\mathbf{T}\mathbf{G}^{-1[t]})) \right].$$

Appendix B. Results of the simulation study

In the following tables, we present the results obtained from 100 runs. The first row gives the proportion of runs where the true model is recovered ($\hat{J} = J$). The proportion of runs where the true set of variables is included in the selected set is given in the second row ($J \subset \hat{J}$). TP denotes the proportion of true variables that are recovered: $|\hat{J} \cap J|/|\hat{J}|$. FDR is the false discovery rate: $|\hat{J}^c \cap \hat{J}|/|\hat{J}|$. Averages over the 100 runs of TP , FDR and of the parameter estimates are given, as well as the mean squared error (MSE). Standard errors are given between parentheses.

Table B.2: Results for model M_1 . The signal to noise ratio is equal to $SNR = 0.60(0.12)$.

	Ideal	Lasso+	ImmLasso	adLasso+	mht.bol+
$\hat{J} = J$	1	0.34	0.36	0.31	0.78
$J \subset \hat{J}$	1	0.98	0.98	0.93	0.78
TP	1	0.99(0.03)	0.99(0.03)	0.99(0.05)	0.95(0.09)
FDR	0	0.17(0.14)	0.17(0.16)	0.18(0.17)	0.00(0.00)
$\hat{\sigma}_e^2$	1	1.05(0.11)	1.09(0.22)	1.00(0.12)	1.03(0.13)
$\hat{\sigma}_1^2$	1	0.97(0.42)	0.98(0.40)	0.93(0.41)	0.95(0.40)
$\hat{\sigma}_2^2$	1	1.13(0.49)	1.12(0.47)	1.04(0.48)	1.06(0.45)
$\hat{\sigma}_3^2$	1	0.94(0.39)	0.95(0.38)	0.91(0.39)	0.94(0.37)
$\hat{\sigma}_{12}^2$	0	-0.02(0.37)	0.14(0.24)	-0.02(0.34)	0.00(0.34)
$\hat{\sigma}_{23}^2$	0	-0.00(0.34)	0.16(0.25)	0.00(0.32)	-0.00(0.35)
$\hat{\sigma}_{13}^2$	0	-0.06(0.30)	0.10(0.20)	-0.06(0.30)	-0.07(0.31)
$\hat{\beta}_1$	0.75	0.73(0.24)	0.73(0.23)	0.74(0.23)	0.74(0.24)
$\hat{\beta}_2$	0.75	0.62(0.29)	0.62(0.29)	0.66(0.29)	0.71(0.31)
$\hat{\beta}_3$	0.75	0.74(0.28)	0.74(0.27)	0.75(0.28)	0.76(0.28)
$\hat{\beta}_4$	0.75	0.46(0.13)	0.46(0.13)	0.57(0.17)	0.72(0.20)
$\hat{\beta}_5$	0.75	0.39(0.14)	0.40(0.14)	0.52(0.20)	0.66(0.34)
MSE	0	0.45(0.22)	0.45(0.21)	0.37(0.23)	0.37(0.30)

Table B.3: Results for model M_2 . The signal to noise ratio is equal to $SNR = 0.90(0.19)$.

	Ideal	Lasso+	ImmLasso	adLasso+	mht.bol+
$\hat{J} = J$	1	0.28	0.30	0.33	0.97
$J \subset \hat{J}$	1	1.00	1.00	0.99	0.98
TP	1	1.00(0.00)	1.00(0.00)	1.00(0.02)	1.00(0.03)
FDR	0	0.23(0.18)	0.22(0.18)	0.22(0.20)	0.00(0.02)
$\hat{\sigma}_e^2$	1	1.12(0.16)	1.16(0.21)	1.00(0.14)	0.99(0.11)
$\hat{\sigma}_1^2$	1	0.94(0.39)	0.93(0.39)	0.90(0.37)	0.95(0.38)
$\hat{\sigma}_2^2$	1	0.98(0.38)	0.97(0.38)	0.95(0.37)	0.99(0.37)
$\hat{\sigma}_{12}^2$	0	0.47(0.27)	0.48(0.27)	0.46(0.26)	0.48(0.27)
$\hat{\beta}_{i_1}$	0.75	0.84(0.23)	0.84(0.23)	0.83(0.23)	0.80(0.23)
$\hat{\beta}_{i_2}$	0.75	0.70(0.29)	0.70(0.29)	0.71(0.28)	0.74(0.29)
$\hat{\beta}_{i_3}$	0.75	0.51(0.12)	0.51(0.12)	0.62(0.13)	0.75(0.11)
$\hat{\beta}_{i_4}$	0.75	0.47(0.12)	0.47(0.11)	0.56(0.15)	0.74(0.15)
$\hat{\beta}_{i_5}$	0.75	0.49(0.11)	0.49(0.11)	0.60(0.13)	0.75(0.11)
MSE	0	0.39(0.18)	0.39(0.18)	0.28(0.17)	0.18(0.16)

Table B.4: Results for model M_3 . The signal to noise ratio is equal to $SNR = 0.92(0.20)$.

	Ideal	Lasso+	ImmLasso	adLasso+	mht.bol+
$\hat{J} = J$	1	0.04	0.07	0.01	0.75
$J \subset \hat{J}$	1	0.95	0.96	0.89	0.78
TP	1	0.99(0.04)	0.99(0.04)	0.98(0.07)	0.93(0.14)
FDR	0	0.36(0.16)	0.35(0.17)	0.42(0.17)	0.03(0.08)
$\hat{\sigma}_e^2$	1	1.16(0.18)	1.23(0.27)	1.01(0.17)	1.04(0.20)
$\hat{\sigma}_1^2$	1	0.98(0.44)	0.97(0.42)	0.93(0.40)	0.97(0.41)
$\hat{\sigma}_2^2$	1	0.92(0.46)	0.92(0.43)	0.89(0.42)	0.94(0.44)
$\hat{\sigma}_{12}^2$	0	0.01(0.31)	0.13(0.19)	0.01(0.31)	0.00(0.32)
$\hat{\beta}_{i_1}$	0.75	0.79(0.26)	0.78(0.26)	0.78(0.27)	0.78(0.27)
$\hat{\beta}_{i_2}$	0.75	0.69(0.26)	0.69(0.25)	0.69(0.24)	0.74(0.26)
$\hat{\beta}_{i_3}$	0.75	0.28(0.14)	0.28(0.14)	0.35(0.19)	0.62(0.30)
$\hat{\beta}_{i_4}$	0.75	0.41(0.12)	0.40(0.12)	0.53(0.13)	0.70(0.21)
$\hat{\beta}_{i_5}$	0.75	0.41(0.12)	0.40(0.12)	0.51(0.18)	0.69(0.26)
MSE	0	0.54(0.21)	0.55(0.21)	0.41(0.21)	0.32(0.34)

Table B.5: Results for model M_4 . The signal to noise ratio is equal to $SNR = 0.86(0.16)$.

	Ideal	Lasso+	ImmLasso	adLasso+	mht.bol+
$\hat{J} = J$	1	0.18	-	0.24	0.71
$J \subset \hat{J}$	1	0.97	-	0.90	0.72
TP	1	0.99(0.07)	-	0.97(0.09)	0.91(0.16)
FDR	0	0.25(0.17)	-	0.38(0.20)	0.01(0.03)
$\hat{\sigma}_e^2$	1	1.29(0.29)	-	1.06(0.29)	1.32(0.41)
$\hat{\sigma}_1^2$	1	0.90(0.46)	-	0.88(0.41)	0.88(0.43)
$\hat{\sigma}_2^2$	1	0.95(0.53)	-	0.90(0.48)	0.76(0.39)
$\hat{\beta}_1$	0.67	0.80(0.25)	-	0.78(0.24)	0.77(0.24)
$\hat{\beta}_2$	0.67	0.69(0.29)	-	0.68(0.28)	0.73(0.29)
$\hat{\beta}_3$	0.67	0.36(0.13)	-	0.48(0.18)	0.68(0.27)
$\hat{\beta}_4$	0.67	0.36(0.11)	-	0.44(0.18)	0.66(0.27)
$\hat{\beta}_5$	0.67	0.34(0.14)	-	0.48(0.19)	0.62(0.31)
MSE	0	0.60(0.25)	-	0.46(0.26)	0.42(0.42)

Table B.6: Results for model M_2 when a linear mixed model estimation is added after the convergence of the algorithm. The signal to noise ratio is equal to $SNR = 0.90(0.19)$.

	Ideal	Lasso+	lmmLasso	adLasso+
$\hat{J} = J$	1	0.28	0.30	0.33
$J \subset \hat{J}$	1	1.00	1.00	0.99
TP	1	1.00(0.00)	1.00(0.00)	1.00(0.02)
FDR	0	0.23(0.18)	0.22(0.18)	0.22(0.20)
$\hat{\sigma}_e^2$	1	0.90(0.13)	0.91(0.17)	0.92(0.13)
$\hat{\sigma}_1^2$	1	0.92(0.38)	0.99(0.40)	0.87(0.36)
$\hat{\sigma}_2^2$	1	0.97(0.36)	1.04(0.38)	0.92(0.36)
$\hat{\sigma}_{12}^2$	0	0.47(0.28)	0.50(0.29)	0.45(0.26)
$\hat{\beta}_{i_1}$	0.75	0.81(0.23)	0.81(0.23)	0.81(0.23)
$\hat{\beta}_{i_2}$	0.75	0.74(0.29)	0.74(0.29)	0.72(0.28)
$\hat{\beta}_{i_3}$	0.75	0.72(0.12)	0.71(0.13)	0.71(0.12)
$\hat{\beta}_{i_4}$	0.75	0.72(0.12)	0.72(0.12)	0.72(0.14)
$\hat{\beta}_{i_5}$	0.75	0.72(0.13)	0.72(0.13)	0.70(0.11)
MSE	0	0.31(0.20)	0.31(0.21)	0.31(0.24)

Appendix C. Proof of Proposition 2.2

G and R are assumed 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 $(\hat{u}, \hat{\beta}) = \underset{(u, \beta)}{\operatorname{argmin}} h(u, \beta)$. Since the function h is convex, we have:

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

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

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

Thus, we obtain:

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

We denote by $W = R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}$. 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 [12] and the generalized least squares. To conclude, we obtain

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