HAL
open science

Fixed effects selection in the linear mixed-effects model using adaptive ridge procedure for L0 penalty performance

Eric Adjakossa, Grégory Nuel

To cite this version:

Eric Adjakossa, Grégory Nuel. Fixed effects selection in the linear mixed-effects model using adaptive ridge procedure for L0 penalty performance. 2017. hal-01494187v2

HAL Id: hal-01494187

https://hal.science/hal-01494187v2

Preprint submitted on 4 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Fixed effects selection in the linear mixed-effects model using adaptive ridge procedure for L_0 penalty performance

Eric Houngla Adjakossa[a,b,*], Gregory Nuel[b]

[a] *International Chair in Mathematical Physics and Applications (ICMPA-UNESCO Chair) /University of Abomey-Calavi, 072 B.P. 50 Cotonou, Republic of Benin*
[b] *Laboratoire de Probabilités et Modèles Aléatoires /Université Pierre et Marie Curie, Case courrier 188 - 4, Place Jussieu 75252 Paris cedex 05 France*

Abstract

This paper is concerned with the selection of fixed effects along with the estimation of fixed effects, random effects and variance components in the linear mixed-effects model. We introduce a selection procedure based on an adaptive ridge (AR) penalty of the profiled likelihood, where the covariance matrix of the random effects is Cholesky factorized. This selection procedure is intended to both low and high-dimensional settings where the number of fixed effects is allowed to grow exponentially with the total sample size, yielding technical difficulties due to the non-convex optimization problem induced by L_0 penalties. Through extensive simulation studies, the procedure is compared to the LASSO selection and appears to enjoy the model selection consistency as well as the estimation consistency.

Keywords: linear mixed-effects model, consistent selection, iteratively weighted ridge, profiled likelihood

1. Introduction

During the last two decades, selection procedures in the linear mixed-effects model have been an active research topic due to the appealing features of the model and the advent of modern technologies facilitating the collection of many variables in scientific studies. Many of these variables are typically included in the full model at the initial stage of modeling to reduce model approximation error, and due to the complexity of the mixed-effects models, inferences and interpretations of the estimated models become challenging as the dimension of fixed or random effects increases (Fan and Li, 2012). The selection of important fixed or random effects has thus become a fundamental problem in the analysis of grouped data using mixed-effects models, especially in the high-dimensional settings where the fixed or the random effects vector dimension is allowed to grow exponentially with the sample size.

*Corresponding author
Email addresses: `ericadjakossah@gmail.com` (Eric Houngla Adjakossa), `Gregory.Nuel@math.cnrs.fr` (Gregory Nuel)

Generally, model selection procedures can be viewed as covering three main approaches: the hypothesis testing procedures, the regularization procedures and other procedures which include the Bayesian selection methods. The testing procedures include the ordinary hypothesis tests and the selection methods based on generalized information criteria. Examples of using testing hypothesis for models selection in the mixed-effects model context include Lin's works (Lin, 1997) who proposed a simple global variance component tests, which are locally asymptotically most precise and are robust in the sense that no assumption about the parametric form of the random effects is made. Despite their global form expressions which require only the fitting of conventional generalized linear models, the critical values of the global test statistics, which are based on large sample theory, are less accurate when the number of levels of each random effect is small, e.g. less than 15. Edwards and his co-workers (Edwards et al., 2008) extended the traditional coefficient of determination R^2 for the linear mixed-effects model $y \sim \mathcal{N}(X\beta, \Sigma = Z\Gamma Z^\top + \sigma^2 I_N)$, where they introduced a statistic $R^2_\beta = (q-1)\nu^{-1}F(\widehat{\beta}, \widehat{\Sigma}) / \left[1 + (q-1)\nu^{-1}F(\widehat{\beta}, \widehat{\Sigma})\right]$, with $F(\widehat{\beta}, \widehat{\Sigma}) = (C\widehat{\beta})^\top \left[C(X^\top \widehat{\Sigma}^{-1} X)^{-1} C^\top\right]^{-1} C\widehat{\beta}/\text{rank}(C)$, $\nu = N - \text{rank}(X) = N - q$, $C = \left[\mathbf{0}_{(q-1)\times 1} I_{q-1}\right]$ of rank $q-1$, in testing $H_0 : C\beta = \mathbf{0}$. R^2_β measures the multivariate association between the repeated outcomes and the fixed effects in the context of longitudinal data analysis. This R^2_β statistic arises as a $1-1$ function of an appropriate F statistic (i.e., $F(\widehat{\beta}, \widehat{\Sigma})$) for testing all the fixed effects, except the intercept. More precisely, R^2_β compares the full model with a null model having no fixed effect except typically the intercept. R^2_β is then generalized to define a partial R^2 statistic for marginal fixed effects of all sorts. Although this testing-based selection procedure of fixed effects is very useful, one of its major drawback is that the choice of the denominator of R^2_β clearly affects the rate of convergence as $N \to \infty$, and may change the parameter being estimated. Examples of R^2-based selection of fixed effects in the mixed-effects model include Snijders and Bosker (1994), Xu (2003), Kramer (2005) and references therein, where generally, too much restrictions are made on the random effects covariance matrix, and clearly may not be appropriate for a wide range of data analysis.

Based on testing procedures, a stepwise procedure can be constructed for selecting important fixed or random effects using generalized information criteria (Pu and Niu, 2006; Fan and Li, 2012), which are a generalization of Akaike's information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz et al., 1978). Pu and Niu (2006) extended the Generalized information criterion (GIC) proposed by Rao and Wu (1989) in order to construct a procedure for selecting fixed and random effects in the linear mixed-effects model. Here, following Shao (1997), the asymptotic behavior of the extended GIC method for selecting fixed effects is studied, and the results from simulations show that if the signal-to-noise ratio is moderate or high, the percentages of choosing the correct fixed effects by the GIC procedure are close to one for finite samples. Another examples of GIC like selection procedures include those proposed by Mallows (1973),

Hannan and Quinn (1979) and Bozdogan (1987). These strategies suggest a unified approach for choosing a parameters vector β that maximizes the penalized likelihood

$$n^{-1}\ell_n(\beta|y) - \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{1}$$

which arose from the Kullback-Leibler (KL) divergence $-\ell_n(\widehat{\beta}|y) + \lambda\|\beta\|_0$ of the fitted model from the true model (Akaike, 1973), where $\ell_n(.|y)$ is the log-likelihood function, $\widehat{\beta}$ is the maximum likelihood estimator of β and p_λ is the penalty function indexed by the regularization parameter $\lambda \geqslant 0$. The L_0-norm $\|\beta\|_0$ of β counts the number of non-vanishing components ($\beta_j \neq 0$) in β, and arises naturally in many classical model selection methods. Although involving a nice interpretation of the best subset selection, and admitting good sampling properties (Barron et al., 1999), its computation is infeasible in high dimensional statistical endeavors (Fan and Lv, 2010). Other penalty functions are regularly used in the literature. A natural generalization of L_0 penalty is the so-called bridge (noted L_q) penalty in Frank and Friedman (1993), where $p_\lambda(|\beta|) = \lambda\|\beta\|^q_{L_q}$ for $0 < q \leqslant 2$. The L_q penalty encompasses L_0, L_1 - LASSO (Tibshirani, 1996) - and L_2 (ridge) penalties. Since none of the L_q penalties satisfies all the required properties (sparsity, approximate unbiasedness and continuity, see Fan and Li (2001) for more details) for their resulting parameter estimators, other penalties including SCAD (Fan, 1997; Fan and Li, 2001) and MCP (Zhang, 2007) are introduced in the literature.

Testing-based stepwise selection procedures, where λ is fixed ($\lambda = 1$ for AIC, $\lambda = \log(n)/2$ for BIC, $\lambda = \log(\log n)/2$ for HQIC (Hannan and Quinn, 1979), $\lambda = (\log n + 1)/2$ for CAIC (Bozdogan, 1987), for example) are computationally expensive for high-dimensional settings and ignore stochastic errors in the variable selection process (Fan and Li, 2001). Another severe drawback is their lack of stability (Breiman et al., 1996). Penalized likelihood approaches using data-driven choice of λ are generally preferred to handle the high-dimensional selection problem. These methods are referred to as regularization procedures. Here, the regularization parameter λ (also called the tuning parameter) is generally estimated using cross-validation methods (Breiman, 1995; Tibshirani, 1996; Fu, 1998; Fan and Li, 2001). Ibrahim et al. (2011) consider the selection of both fixed and random effects in a general class of mixed effects models by optimizing the penalized likelihood criterion $Q_\lambda(\theta|\theta_{\text{old}}) = Q(\theta|\theta_{\text{old}}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|)$ using the EM algorithm (Dempster et al., 1977), where θ is the vector of all the unknown parameters, $Q(.|\theta_{\text{old}})$ is the resulting function of the EM algorithm E-step, and p_λ is either the SCAD or the adaptive lasso - ALASSO - (Zou, 2006) penalty. They approximate the integral in $Q(\theta|\theta_{\text{old}})$ by using a Markov chain Monte Carlo and introduce the $\text{IC}_Q(\lambda) = -2Q(\widehat{\theta}_\lambda|\widehat{\theta}_0) + c_n(\widehat{\theta}_\lambda)$ statistic (Ibrahim et al., 2008) for selecting the regularization parameter, where $\widehat{\theta}_0 = \arg\max_\theta \ell(\theta)$ is the unpenalized maximum likelihood estimate and $c_n(\theta)$ a

function of the data and fitted model.

Fan and Li (2012) introduce a class of variable selection methods for fixed effects using a penalized profiled likelihood method, where the unknown covariance matrix of the random effects is replaced with a suitable proxy matrix. Here, the general idea used is as follows. After writing the joint density $f(y, \gamma)$ of the response variable y and the random effects vector γ, they consider the penalized profiled likelihood $L_n(\beta, \widehat{\gamma}(\beta)) - n \sum_{j=1}^{d_n} p_\lambda(|\beta_j|)$, where $L_n(\beta, \widehat{\gamma}(\beta)) = f(y, \widehat{\gamma}(\beta))$ with $\widehat{\gamma}(\beta)$ the empirical Bayes estimate of γ (Harville, 1977) in which the covariance matrix of γ is replaced with the proxy matrix. d_n may increase with the sample size n and p_λ is a concave penalty function (SCAD and LASSO in their simulation and application section). Although the proxy matrix may be different from the true one, it may still yield correct model selection results at the cost of some additional bias (Fan and Li, 2012). Schelldorfer et al. (2011) deal with theoretical and computational aspects for high-dimensional selection of fixed effects in the linear mixed-effects model, where the consistency of the estimator is proven along with a non-asymptotic oracle result for the adaptive LASSO estimator, under the assumption that the eigenvalues of $Z^\top Z$ are bounded. Here, Z is the random effects design matrix, and an explicit analytical expression of the regularization parameter λ is also given.

Other approches for variable selection in linear mixed-effects models include "fence" procedure and Bayesian techniques. Jiang et al. (2008) introduce a class of strategies known as a fence methods intended for variable selection in both linear and generalized linear mixed-effects models. The fence strategy is based on a measure of lack-of-fit that is a quantity $Q_M = Q_M(y, \theta_M)$, where y is the response variable, M indicating a candidate model for the selection and θ_M denotes the vector of parameters under M. Counting among the typically rare and ad-hoc selection procedures, the fence method is computationally very demanding, particularly because it involves the estimation of the standard deviation of the difference of lack-of-fit measures. For more details on these ad-hoc selection procedures, see the nice review paper of Müller et al. (2013) and references therein. Bayesian model selection requires to assign a prior distribution over the model parameters and compute the posterior probabilities of each of them. These computations can be difficult so are usually carried out by applying sophisticated Markov Chain Monte Carlo (MCMC) algorithms (Müller et al., 2013). Exemples of Bayesian model selection include Chen and Dunson (2003) and Saville and Herring (2009) who point out that these kinds of MCMC methods are generally time consuming to implement, requiring special software and depend on subjective choice of the hyperparameters in the priors.

In this paper, we discuss the selection of fixed effects in the linear mixed-effects model using an adaptive ridge (AR) penalty of the profiled log-likelihood, where the random effect covariance matrix is Cholesky factorized for solving a preliminary penalized least square problem. The profiled

likelihood is calculated by slightly modifying the approach proposed by Bates et al. (2007). The weights matrix of the AR procedure introduced here is updated in such a way that the procedure converges toward selection with L_0 penalty, as have done Frommlet and Nuel (2016). The present selection strategy is intended to both low and high-dimensional settings.

The rest of the article is organized as follows. The profiled log-likelihood is introduced in Section 2. In Section 3, we present the weighted ridge procedure, and Section 4 presents the simulation studies.

2. profiled log-likelihood for the linear mixed-effects model

In this section, we consider the classical linear mixed-effect model setting where the number of observations n is larger than the number of covariates p. By slightly modifying the approach introduced by Bates et al. (2007), we calculate the profiled likelihood function.

2.1. Model and notations

We consider the linear mixed-effects model in which the residual terms are homoscedastic and independent of the random effects as follows.

$$\mathcal{Y} = X\beta + Z\gamma + \varepsilon, \tag{2}$$

$$\gamma \sim \mathcal{N}(\mathbf{0}, \Gamma), \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n) \quad \text{and } \gamma \perp \varepsilon, \tag{3}$$

where $\mathcal{Y}$ is the random response variable whose observed value is $y_{\text{obs}} \in \mathbb{R}^n$, $\gamma \in \mathbb{R}^q$ is the unobserved random effects vector with covariance matrix Γ, ε is the residual term, $\beta \in \mathbb{R}^p$ is the fixed effects vector, and X and Z are the fixed and random effects related design matrices of dimensions $n \times p$ and $n \times q$, respectively. $\sigma^2 I_n$ is the covariance matrix of ε with $\sigma > 0$ and I_n the $n \times n$ identity matrix.

As a variance-covariance matrix, Γ must be positive semidefinite. Conveniently, the model is expressed in terms of a relative covariance factor, Λ_θ, which is a $q \times q$ matrix, depending on the variance components vector, θ, that generate the symmetric $q \times q$ variance-covariance matrix, Γ, according to

$$\Gamma = \sigma^2 \Lambda_\theta \Lambda_\theta^\top, \tag{4}$$

where σ is the same scale parameter as in Equation (3). This factorization of Γ yields the existence of a random vector $\mathcal{U}$ such that

$$\gamma = \Lambda_\theta \mathcal{U}, \tag{5}$$

with

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_q) \tag{6}$$

which is called a *spherical random effects*[1] variable. The model can therefore be re-expressed as follows

$$\mathcal{Y} = X\beta + Z\Lambda_\theta \mathcal{U} + \varepsilon, \tag{7}$$

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_q), \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n) \quad \text{and} \quad \mathcal{U} \perp \varepsilon. \tag{8}$$

The parameters of the model are the fixed effects vector β and the variance components θ and σ^2. This formulation of the linear mixed-effects model allows, not only, the use of a singular matrix Λ_θ which arises in practice, but also for a relatively compact expression for the profiled log-likelihood of β and θ, conditional on y_{obs}.

2.2. profiled likelihood

The profiled likelihood considered here is expressed through the following theorem.

Theorem 2.1. *Suppose that y is a realization of a random vector $\mathcal{Y}$ satisfying the linear mixed-effects model expressed by Equations (7) and (8), where β, θ and σ^2 are the parameters to be estimated. Denoting by L_θ the matrix such that $L_\theta^\top L_\theta = (Z\Lambda_\theta)^\top Z\Lambda_\theta + I_q$, $\tilde{u}$ the conditional mean of $\mathcal{U}$ given that $\mathcal{Y} = y$, and $g(\tilde{u}) = \|y - X\beta - Z\Lambda_\theta \tilde{u}\|^2 + \|\tilde{u}\|^2$, the profiled log-likelihood of β and θ conditional on y is*

$$\tilde{\ell}(\beta, \theta | y) = -\frac{1}{2} \log |L_\theta|^2 - \frac{n}{2} \left[1 + \log \left(\frac{2\pi g(\tilde{u})}{n} \right) \right]. \tag{9}$$

Proof. Denoting by u a realization of $\mathcal{U}$, the density of y is expressed as

$$f(y) = \int_{\mathbb{R}^q} f(y, u) du = \int_{\mathbb{R}^q} f(y|u) f(u) du.$$

$$f(y|u) f(u) = (2\pi\sigma^2)^{-(n+q)/2} \exp \left[-\frac{\|y - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2}{2\sigma^2} \right], \tag{10}$$

where $\|u\|^2 = u^\top u$, with $^\top$ denoting the transpose operator.

$$\|y - X\beta - Z\Lambda_\theta u\|^2 + \|u\|^2 = \left\| \begin{pmatrix} y - X\beta \\ 0 \end{pmatrix} - \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix} u \right\|^2 = g(u), \tag{11}$$

[1] Bates et al. (2007) argued that the term "spherical" is related to the fact that the contours of the $\mathcal{N}(0, \sigma^2 I_q)$ probability density are spheres.

and solving the penalized least squares problem that is to minimize $g(u)$ over u implies

$$\tilde{u} = \underset{u \in \mathbb{R}^q}{\arg\min}\, g(u) \iff \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix}^\top \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix} \tilde{u} = \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix}^\top \begin{pmatrix} y - X\beta \\ 0 \end{pmatrix}. \tag{12}$$

Viewed as a function of u, g is C^∞ and the first and the second differential of g are

$$\mathrm{d}g(u) = 2\,\mathrm{tr}\left\{-(y - X\beta - Z\Lambda_\theta u)^\top Z\Lambda_\theta \mathrm{d}u + u^\top \mathrm{d}u\right\}$$

and

$$\mathrm{d}^2 g(u) = 2\,\mathrm{tr}\left\{(Z\Lambda_\theta)^\top Z\Lambda_\theta \mathrm{d}u\mathrm{d}u^\top + \mathrm{d}u\mathrm{d}u^\top\right\}.$$

This implies that

$$\frac{\partial g}{\partial u}(u) = -2\left(y - X\beta - Z\Lambda_\theta u\right)^\top Z\Lambda_\theta + 2u^\top$$

and

$$\frac{\partial^2 g}{\partial u \partial u^\top}(u) = 2\left(Z\Lambda_\theta\right)^\top Z\Lambda_\theta + 2\mathrm{I}_q.$$

Then, $g(u)$ can be rewritten as

$$g(u) = g(\tilde{u}) + \frac{\partial g}{\partial u}(\tilde{u})(u - \tilde{u}) + \frac{1}{2}(u - \tilde{u})^\top \frac{\partial^2 g}{\partial u \partial u^\top}(\tilde{u})(u - \tilde{u}). \tag{13}$$

Referring to Equation (12),

$$\begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix} \left[\begin{pmatrix} y - X\beta \\ 0 \end{pmatrix} - \begin{pmatrix} Z\Lambda_\theta \\ \mathrm{I}_q \end{pmatrix} \tilde{u}\right] = 0 \implies \frac{\partial g}{\partial u}(\tilde{u}) = 0. \tag{14}$$

Then,

$$\begin{aligned} g(u) &= g(\tilde{u}) + \frac{1}{2}(u - \tilde{u})^\top \frac{\partial^2 g}{\partial u \partial u^\top}(\tilde{u})(u - \tilde{u}) \\ &= g(\tilde{u}) + (u - \tilde{u})^\top \left[\left(Z\Lambda_\theta\right)^\top Z\Lambda_\theta + \mathrm{I}_q\right](u - \tilde{u}) \\ &= g(\tilde{u}) + \|L_\theta(u - \tilde{u})\|^2, \text{ with } L_\theta^\top L_\theta = \left(Z\Lambda_\theta\right)^\top Z\Lambda_\theta + \mathrm{I}_q, \end{aligned} \tag{15}$$

and

$$\begin{aligned} f(y) &= (2\pi\sigma^2)^{-(n+q)/2} \int_{\mathbb{R}^q} \exp\left[-\frac{g(\tilde{u}) + \|L_\theta(u - \tilde{u})\|^2}{2\sigma^2}\right] du \\ &= (2\pi\sigma^2)^{-(n+q)/2} \exp\left[-\frac{g(\tilde{u})}{2\sigma^2}\right] \int_{\mathbb{R}^q} \exp\left[-\frac{\|L_\theta(u - \tilde{u})\|^2}{2\sigma^2}\right] du. \end{aligned} \tag{16}$$

$v = L_\theta(u - \tilde{u}) \implies du = \dfrac{1}{|L_\theta|}dv$, and

$$\begin{aligned} f(y) &= (2\pi\sigma^2)^{-n/2}|L_\theta|^{-1}\exp\left[-\frac{g(\tilde{u})}{2\sigma^2}\right]\int_{\mathbb{R}^q}(2\pi\sigma^2)^{-q/2}\exp\left[-\frac{\|v\|^2}{2\sigma^2}\right]dv \\ &= (2\pi\sigma^2)^{-n/2}|L_\theta|^{-1}\exp\left[-\frac{g(\tilde{u})}{2\sigma^2}\right] \end{aligned} \tag{17}$$

The log-likelihood of β, θ and σ^2 conditional on y is

$$\ell(\beta, \theta, \sigma^2|y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|L_\theta|^2 - \frac{g(\tilde{u})}{2\sigma^2}. \tag{18}$$

$\dfrac{\partial\ell(\beta, \theta, \sigma^2|y)}{\partial\sigma^2} = 0 \implies \sigma^2 = \dfrac{g(\tilde{u})}{n}$. By profiling out σ^2, the profiled log-likelihood, $\tilde{\ell}(\beta, \theta|y)$, of β and θ conditional on y is expressed through

$$-2\tilde{\ell}(\beta, \theta|y) = \log|L_\theta|^2 + n\left[1 + \log\left(\frac{2\pi g(\tilde{u})}{n}\right)\right].$$

□

3. L_0 estimator of β using iteratively weighted ridge procedure

The L_0 estimator of the fixed-effects vector β using iteratively weighted ridge procedure presented here fits both the low and the high-dimensional settings. Theoretically, this selection method may enjoy great stability since it uses neither inverse of X (fixed-effects design matrix) nor inverse of Z (random effects design matrix).

3.1. Adaptive Ridge penalty for the profiled likelihood

Let us assume that the true underlying fixed-effects vector β_{true} is sparse in the sense that many of its coefficients are zero. To enforce the sparsity of the estimator of β, we add a weighted L_2 (ridge) penalty for the fixed-effects vector β to the profiled log-likelihood function. Thus, we are considering the objective function

$$\tilde{\ell}_{\lambda,w}(\beta, \theta) = -2\tilde{\ell}(\beta, \theta|y_{\text{obs}}) + \lambda\beta^\top W\beta, \tag{19}$$

where $W = \text{diag}(w_1, \ldots, w_p)$ is a $p \times p$ diagonal matrix and $\lambda \geqslant 0$ is a regularization parameter. We aim at estimating β, θ and σ^2 by

$$(\tilde{\beta}_{\lambda,w}, \tilde{\theta}_{\lambda,w}) = \arg\min_{\beta,\theta}\tilde{\ell}_{\lambda,w}(\beta, \theta) \quad \text{and} \quad \widetilde{\sigma^2}_{\lambda,w} = \frac{\|y - X\tilde{\beta}_{\lambda,w} - Z\Lambda_{\tilde{\theta}_{\lambda,w}}\tilde{u}\|^2 + \|\tilde{u}\|^2}{n}, \tag{20}$$

where $\tilde{u}$ is calculated by the penalized least squares algorithm that has been hinted at Equation (12). The criterion $\tilde{\ell}_{\lambda,w}(\beta,\theta)$ is minimized using one of the constrained optimization functions in R (R Core Team, 2015), to provide the estimators $\tilde{\beta}_{\lambda,w}$, $\tilde{\theta}_{\lambda,w}$ and $\widetilde{\sigma^2}_{\lambda,w}$.

3.2. Iteratively Weighted ridge procedure

The L_0 penalty for regularization arises naturally in many classical model selection since, indeed, it counts the number of non-vanishing parameters, giving a nice interpretation of the best subset selection and admits nice sampling properties (Barron et al., 1999). However, its computation is infeasible in high dimensional settings and clearly argued to be a combinational problem with NP-complexity (Fan and Lv, 2010). Some workarounds are reported in the literature, where the proposed procedure converge toward the L_0-penalty based selection. For example, Frommlet and Nuel (2016) introduced an adaptive ridge procedure that helps to approximate L_0-penalty performances. Here, we are using the same procedure where the weight matrix (W defined in Equation (19)) diagonal elements $w_j, 1 \leqslant j \leqslant p$ are iteratively computed and defined as

$$w_j^{(k)} = \left[\left|\beta_j^{(k)}\right|^2 + \delta^2\right]^{-1}, \tag{21}$$

as have done Frommlet and Nuel (2016), taking inspiration from Grandvalet (1998); Biihlmann and Meier (2008); Candes et al. (2008); Rippe et al. (2012) and their simulation results. In Equation (21), k identifies the iteration and β_j is the jth component of β. The general expression of $w_j^{(k)}$ is $w_j^{(k)} = \left[\left|\beta_j^{(k)}\right|^\tau + \delta^\tau\right]^{\frac{q-2}{\tau}}$, where q precises the norm $\|\cdot\|_{L_q}$ for the penalty, δ calibrates which effect sizes are considered relevant and τ determines the quality of the approximation $w_j\beta_j^2 \approx |\beta_j|^q$. In practice, $\delta = 10^{-5}$ seems to perform well. For more details, see Frommlet and Nuel (2016). More precisely, the selection procedure performed here is as follows. For a fixed λ, $\tilde{\beta}_{\lambda,w}$ is initialized at $(1,\ldots,1)^\top$, the vector *selection*, say, which identifies the selected fixed effects is initialized at $(1,\ldots,1)^\top$ and W is initialized at $\mathrm{diag}(1,\ldots,1)$. Then the steps come.

1) $\tilde{\beta}_{\lambda,\mathrm{old}} \leftarrow \tilde{\beta}_{\lambda,w}$
2) perform the optimization $(\tilde{\beta}_{\lambda,w}, \tilde{\theta}_{\lambda,w}) = \underset{\beta,\theta}{\arg\min}\, \tilde{\ell}_{\lambda,w}(\beta,\theta)$, initializing β by $\tilde{\beta}_{\lambda,\mathrm{old}}$ and θ by θ_0. Here, the components of θ which are variances are initialized by 1 and those which are not variances are initialized by 0. Thus, θ_0 components are 0 or 1.
3) $w_j \leftarrow \left(\tilde{\beta}^2_{\lambda,w,j} + \delta^2\right)^{-1}$, where $\tilde{\beta}_{\lambda,w,j}$ is the jth component of $\tilde{\beta}_{\lambda,w}$, and w_j is the jth element of W's diagonal.
4) $\mathit{selection}_{\mathrm{old}} \leftarrow \mathit{selection}$ and $\mathit{selection} \leftarrow W \cdot \mathrm{diag}(\tilde{\beta}_{\lambda,w}) \cdot \tilde{\beta}_{\lambda,w}$
5) if $|\mathit{selection} - \mathit{selection}_{\mathrm{old}}| < \mathrm{tol} = 10^{-5}$, then the selection can be considered as well performed for λ. Thus, we choose a new value for λ. If $|\mathit{selection} - \mathit{selection}_{\mathrm{old}}| \geqslant \mathrm{tol}$, the selection does not perform well and we go to the item 1) by choosing the current $\tilde{\beta}_{\lambda,w}$ as $\tilde{\beta}_{\lambda,\mathrm{old}}$, without changing the value of λ.

For some λ values, the *selection* vector may contain 0 or 1 as components. If it is 1, the corresponding fixed effect β_j is selected, and if it is 0 the corresponding β_j is not selected for λ.

For the choice of the regularization parameter λ, we propose to use the Bayesian Information Criterion (BIC) criterion defined as

$$c_{n,\lambda} = -2\ell(\hat{\beta}_{\lambda,\text{sel}}, \hat{\theta}_{\lambda,\text{sel}}, \widehat{\sigma^2}_{\lambda,\text{sel}} | y_{\text{obs}}) + \log(n) \cdot \hat{d}_\lambda, \tag{22}$$

where $\hat{\beta}_{\lambda,\text{sel}}, \hat{\theta}_{\lambda,\text{sel}}$ and $\widehat{\sigma^2}_{\lambda,\text{sel}}$ are the ML parameters estimators considering the selected variables. $\hat{d}_\lambda = \#\{\hat{\beta}_{\lambda,\text{sel},j} \neq 0 : 1 \leqslant j \leqslant p\} + \dim(\theta) + 1$ is the sum of the number of nonzero fixed-effects and the number of variance components. This form of $\hat{d}_\lambda$ has been suggested by Bates (2010) and empirically validated by Schelldorfer et al. (2014).

4. Simulation studies

In this section, we assess the performance of our approach using simulated data. We compare the obtained results with those coming from the Lasso implementation in a situation where the number of noise variables are excessive.

Since we use one of the optimizers available in the R software for minimizing the $\tilde{\ell}_{\lambda,w}(\beta, \theta)$ criterion, for too bigger values of p, especially when $n \ll p$, the convergence of the used algorithm ("nlminb" for example) is hardly or no more reached. Due to this convergence problem, we restrict the simulation studies to the low-dimensional setting where $p > 40$ and will focus on this problem in another coming paper with the same theoretical approach.

We restrict ourselves to the case of longitudinal data study with N observations coming from n subjects where each subject i has n_i observations. The "working" data sets are simulated under the following model.

$$y_i = X_{1i}\beta^* + \gamma_i \mathbb{1}_{n_i} + \varepsilon_i, \quad \gamma_i \sim \mathcal{N}(0, \Gamma), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i}), \quad \text{for} \quad i = 1, \ldots, n; \tag{23}$$

where $\beta^* \in \mathbb{R}^{p_1}$ is the true fixed effects vector, X_1 is a $N \times p_1$ design covariates matrix, $\mathbb{1}_{n_i} = (1, \ldots, 1) \in \mathbb{R}^{n_i}$. $\gamma \perp \varepsilon$ and $\gamma_i \perp \gamma_{i'}$ for $i \neq i'$. γ_i is an random intercept for the ith subject and ε is the residual term of the model.

We suppose that we are following up a sample of subjects where the goal is to evaluate how their weights are influenced by other variables including the age, the sex and the nutrition score "nscore", and which variables govern this influence. The covariates sex, nscore and age are staked in the model matrix X_1 and all other covariates are staked in another $N \times p_2$ model matrix X_2 such that $X = (X_1|X_2)$ with $\dim(X) = N \times p$ and $p = p_1 + p_2$. The components of the vector (variable) age are randomly sampled from a uniform distribution in $[18, 37]$ and the nscore variable

is also uniform in $[20, 50]$. We would like to fit to the data, the model

$$y_i = X_i\beta + \gamma_i \mathbb{1}_{n_i} + \varepsilon_i, \quad \gamma_i \sim \mathcal{N}(0, \Gamma), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i}), \quad \text{for} \quad i = 1, \ldots, n; \tag{24}$$

where $\beta = (\beta_1^\top, \beta_2^\top)^\top$, with $\beta_1 \in \mathbb{R}^{p_1}$ and $\beta_2 = \mathbf{0} \in \mathbb{R}^{p_2}$. We are then challenging to identify which β components are zero.

For the working data sets, we choose $p_1 = 4$, $p = 54$, $N = 300$, $n = 90$, $\sigma = 1$, $\Gamma = 1$ and $\beta^* = (1, -1, -1, 1)$. For the fixed-effects β, we have in fact fifty zeros components and only four components are not zeros. We simulate 100 data sets for which we perform both the lasso selection and the IWR (Iteratively Weighted Ridge) procedure. One hundred values of the regularization parameter λ are chosen in $[10^{-2}, 10^2]$. For each replication, the λ value that minimizes the BIC criterion is retained for selecting the significant fixed effects. Denoting by $\beta^{**} = (1, -1, -1, 1, 0, \ldots, 0)$, the mean squared error MSE $= \mathbb{E}\left[\|\hat{\beta} - \beta^{**}\|_2^2\right]$ (with $\hat{\beta}$ coming from lasso or IWR) is computed over the 100 replications. The cardinality of the estimated active set (i.e. $|S(\hat{\beta})|$, with $S(\hat{\beta}) = \{k : \hat{\beta}_k \neq 0\}$) is computed as well as the proportion TP of true positive (i.e. the selected set is exactly the true one). We also compute the proportion TPC in which the selected set contains the true one and the proportion ZP in which the true zeros are estimated.

Table 1: **Selection performances comparison between LASSO and IWR procedures.**

Performance criterion	LASSO	IWR
MSE	1.200	0.254
$\lvert S(\hat{\beta})\rvert$	4(1.614)	5(1.250)
TP	0%	35%
TPC	16%	90%
ZP	98%(0.028)	98%(0.023)

The selection results based on the 100 simulated data sets are indeed summarized through five performance criteria that are contained in Table 1, where the numbers between parentheses are the standard deviations related to the criteria mean values (printed just before these parentheses). Information from Table 1 show that the IWR procedure outperforms the LASSO one. IWR selects the correct model 35% of the time when LASSO has never found it (see TP values in Table 1). For information, we use the R software package lmmlasso (Schelldorfer, 2011) to perform LASSO selection. 90% of the time, the model selected by IWR contains the true one against 16% for LASSO (see TPC values from Table 1). Obviously, the estimations are of better qualities from IWR than from LASSO (MSE $= 0.254$ for IWR, and MSE $= 1.200$ for LASSO).

The true zero estimation proportions (ZP) are computed in Table 2 which contains also their number of occurrence over the simulated data sets. The LASSO seems to find more often than

Table 2: **True zero estimation proportions (ZP) with the number of occurrence over the 100 replications.**

	LASSO							IWR					
ZP	0.88	0.90	0.92	0.94	0.96	0.98	1	0.90	0.92	0.94	0.96	0.98	1
Number	2	2	3	6	14	24	49	2	2	8	16	32	40

IWR all the true zeros (49% for LASSO against 40% for IWR, in Table 2). This may explain the overfitting behavior of IWR ($|S(\hat{\beta})| = 5$ for IWR and $|S(\hat{\beta})| = 4$ for LASSO in Table 1). It therefore seems that LASSO has a stronger shrinkage capability than has IWS which shows in turn a somewhat overfitting behavior than LASSO. Through these simulations studies, it appears that the iteratively weighted aspect of the IWR procedure highers the shrinkage performance of the ordinary Ridge and results in a selection method having better performance than the LASSO.

Like in the case of LASSO, it is possible for IWR to take advantage of a warm start of the algorithm to obtain the full regularization path of the selection problem. For instance, Figure 1

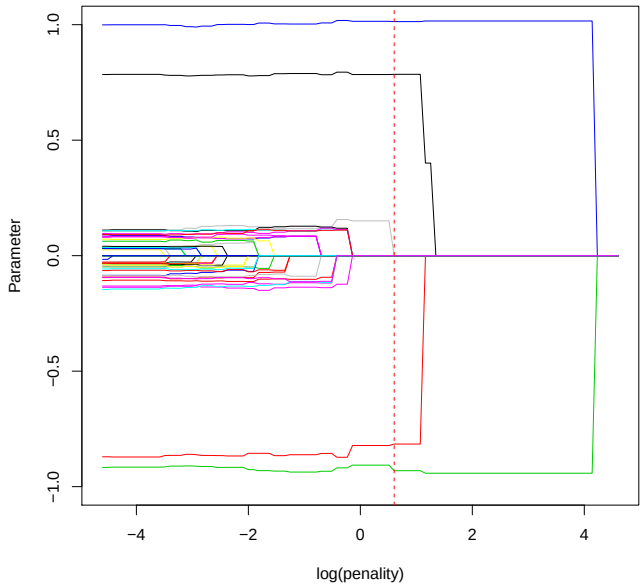

Figure 1: **Example of a full regularization path for the iteratively weighted ridge selection procedure with $N = 300$, $n = 90$, $p = 54$, $\sigma = 1 = \Gamma$. Only the first four components of β^* are nonzero. The dataset is simulated using the R software with set.seed(3). The vertical red dashed bar corresponds to the minimum of BIC and shows the selected variables.**

shows the full regularization path from one of the simulated data. The vertical red dashed bar corresponds to the minimum of BIC and shows the selected variables. On Figure 1, we clearly have $|S(\hat{\beta})| = 4$, i.e., four variables selected at the end of the procedure.

5. Conclusion

In this paper, we have focused on the fixed-effects selection problem in the linear mixed-effects model. We have introduced an iteratively weighted ridge procedure which enhances the shrinkage performance of the ordinary ridge in order to approximate the performance of the L_0 based penalty selection. This selection method is based on an adaptive ridge penalty of the profiled likelihood, where the covariance matrix of the random effects is Cholesky factorized. The procedure fits both the low and the high-dimensional settings and may enjoy a great numeric stability since it needs no use of the inverse of the fixed-effects design matrix or the design matrix of the random-effects. Due the problems of the lack of convergence in the high-dimensional settings using available optimizers, we restrict the simulations studies to the low-dimensional case where we find out that our selection procedure outperforms the LASSO selection. In another forcoming paper, we will focus on this convergence problem in the high-dimensional cases with the same theoretical approach.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory (Edited by B. N. Petrov and F. Csaki), 267–281.

Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19, 716–723.

Barron, A., Birgé, L., Massart, P., 1999. Risk bounds for model selection via penalization. Probability theory and related fields 113, 301–413.

Bates, D., Sarkar, D., Bates, M.D., Matrix, L., 2007. The lme4 package. R package version 2, 74.

Bates, D.M., 2010. lme4: Mixed-effects modeling with r. URL http://lme4. r-forge. r-project. org/book .

Biihlmann, R., Meier, L., 2008. Discussion of" one-step sparse estimates in nonconcave penalized likelihood models," by h. zou and r. li. Ann. Statist 36, 1534–1541.

Bozdogan, H., 1987. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. Psychometrika 52, 345–370.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.

Breiman, L., et al., 1996. Heuristics of instability and stabilization in model selection. The annals of statistics 24, 2350–2383.

Candes, E.J., Wakin, M.B., Boyd, S.P., 2008. Enhancing sparsity by reweighted ? 1 minimization. Journal of Fourier analysis and applications 14, 877–905.

Chen, Z., Dunson, D.B., 2003. Random effects selection in linear mixed models. Biometrics 59, 762–769.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) , 1–38.

Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F., Schabenberger, O., 2008. An r2 statistic for fixed effects in the linear mixed model. Statistics in medicine 27, 6137–6157.

Fan, J., 1997. Comments on wavelets in statistics: A review by a. antoniadis. Journal of the Italian Statistical Society 6, 131–138.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96, 1348–1360.

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20, 101.

Fan, Y., Li, R., 2012. Variable selection in linear mixed effects models. Annals of statistics 40, 2043.

Frank, L.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35, 109–135.

Frommlet, F., Nuel, G., 2016. An adaptive ridge procedure for l 0 regularization. PloS one 11, e0148620.

Fu, W.J., 1998. Penalized regressions: the bridge versus the lasso. Journal of computational and graphical statistics 7, 397–416.

Grandvalet, Y., 1998. Least absolute shrinkage is equivalent to quadratic penalization, in: ICANN 98. Springer, pp. 201–206.

Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society. Series B (Methodological) , 190–195.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72, 320–338.

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

Ibrahim, J.G., Zhu, H., Tang, N., 2008. Model selection criteria for missing-data problems using the em algorithm. Journal of the American Statistical Association .

Jiang, J., Rao, J.S., Gu, Z., Nguyen, T., et al., 2008. Fence methods for mixed model selection. The Annals of Statistics 36, 1669–1692.

Kramer, M., 2005. R 2 statistics for mixed models. Statistics in medicine .

Lin, X., 1997. Variance component testing in generalised linear models with random effects. Biometrika 84, 309–326.

Mallows, C.L., 1973. Some comments on c p. Technometrics 15, 661–675.

Müller, S., Scealy, J.L., Welsh, A.H., et al., 2013. Model selection in linear mixed models. Statistical Science 28, 135–167.

Pu, W., Niu, X.F., 2006. Selecting mixed-effects models based on a generalized information criterion. Journal of multivariate analysis 97, 733–758.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Rao, R., Wu, Y., 1989. A strongly consistent procedure for model selection in a regression problem. Biometrika 76, 369–374.

Rippe, R.C., Meulman, J.J., Eilers, P.H., 2012. Visualization of genomic changes by segmented smoothing using an l 0 penalty. PloS one 7, e38230.

Saville, B.R., Herring, A.H., 2009. Testing random effects in the linear mixed model using approximate bayes factors. Biometrics 65, 369–376.

Schelldorfer, J., 2011. lmmlasso: Linear mixed-effects models with lasso. R package version 0.1-2 .

Schelldorfer, J., Bühlmann, P., DE, G., VAN, S., 2011. Estimation for high-dimensional linear mixed-effects models using l1-penalization. Scandinavian Journal of Statistics 38, 197–214.

Schelldorfer, J., Meier, L., Bühlmann, P., 2014. Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using ?1-penalization. Journal of Computational and Graphical Statistics 23, 460–477.

Schwarz, G., et al., 1978. Estimating the dimension of a model. The annals of statistics 6, 461–464.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica , 221–242.

Snijders, T.A., Bosker, R.J., 1994. Modeled variance in two-level models. Sociological methods & research 22, 342–363.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) , 267–288.

Xu, R., 2003. Measuring explained variation in linear mixed effects models. Statistics in medicine 22, 3527–3541.

Zhang, C.H., 2007. Penalized linear unbiased selection. Department of Statistics and Bioinformatics, Rutgers University , 2007–003.

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