




# Model selection in linear mixed-effect models

Simona Buscemi<sup>1</sup> · Antonella Plaia<sup>1</sup> 

Received: 19 September 2018 / Accepted: 17 October 2019 / Published online: 28 October 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

## Abstract

Linear mixed-effects models are a class of models widely used for analyzing different types of data: longitudinal, clustered and panel data. Many fields, in which a statistical methodology is required, involve the employment of linear mixed models, such as biology, chemistry, medicine, finance and so forth. One of the most important processes, in a statistical analysis, is given by model selection. Hence, since there are a large number of linear mixed model selection procedures available in the literature, a pressing issue is how to identify the best approach to adopt in a specific case. We outline mainly all approaches focusing on the part of the model subject to selection (fixed and/or random), the dimensionality of models and the structure of variance and covariance matrices, and also, wherever possible, the existence of an implemented application of the methodologies set out.

**Keywords** Linear mixed model · Mixed model selection · AIC · BIC · MCP · LASSO · Shrinkage methods · MDL

## 1 Introduction

Linear mixed-effects models (LMM) represent one of the most wide instruments for modeling data in applied statistics, and increasing research on linear mixed models has been rapidly in the last 10–15 years. This is due to the wide range of its applications to different types of data (clustered data such as repeated measures, longitudinal data, panel data, and small area estimation), which involve the fields of agriculture, economics, medicine, biology, sociology etc.

Some practical issues usually encountered in statistical analysis concern the choice of an appropriate model, estimating parameters of interest and measuring the order

---

✉ Antonella Plaia  
antonella.plaia@unipa.it

Simona Buscemi  
simona.buscemi@unipa.it

<sup>1</sup> Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, Ed. 13, 90128 Palermo, Italy

or dimension of a model. This paper focuses on model selection, which is essential for making valid inference. The principle of model selection or model evaluation is to choose the “best approximating” model within a class of competing models, characterized by a different number of parameters, a suitable model selection criterion given a data set (Bozdogan 1987). The ideal selection procedure should lead to the “true” model, i.e., the unknown model behind the true process generating the observed data. In practice, one seeks, among a set of plausible candidate models, the parsimonious one that best approximates the “true” model.

The selection of only one model among a pool of candidate models is not a trivial issue in LMMs, and the different methods proposed in the literature over time are, often, not directly comparable. In fact, not only there is a different notation among papers and great confusion as regards the software (R, SAS, MATLAB, etc.) to be used, but also a lack of landmarks allowing users to prefer one method rather than others.

Hence, the main purpose of this review is to provide a view about some useful components/factors characterizing each selection criterion, so that users can identify the method to apply in a specific situation. Moreover, we will also try to tidy up the notation used in the literature, by “translating,” if necessary, the symbols and formulas found in each paper to produce a common “language.” We begin by updating the recent review by Müller et al. (2013), then add some information about each selection criteria, such as the kind of effects that each method focuses on, or the structure of variance–covariance matrix, or the model dimensionality, or even the software used for implementing each method.

When coping with LMMs, it is not a good idea to assume independence or uncorrelation among response observations. For example, in the case of repeated measures, data are collected about the same individual over time. Hence, the traditional linear regression model is not appropriate to describe the data. For a detailed description of analogies and differences between linear mixed models and linear models, see Müller et al. (2013).

An important issue associated with LMMs selection is related to the dimension of the fixed and random components. Most of the literature bases inference, selection and interpretation of models in the finite (fixed) dimensional case, which means that the number of parameters is less than the number of units. Recently, more attention has been given to the handling of high-dimensional settings, which requires more complex computational applications. The word “high-dimensional” refers to situations where the number of unknown parameters that are to be estimated is one or several orders of magnitude larger than the number of samples in the data (Bühlmann and van de Geer 2011). Furthermore, in LMMs, the number of parameters can grow exponentially with the sample size, i.e., the number of effects is strictly related to the number of units. Thus, if the sample size increases the set of effects diverges. Only recently, some authors have tried to make inference within the LMM framework, on high-dimensional settings (Fan and Li 2012; Schellldorfer et al. 2011).

Model selection is a challenge in itself when one deals with the classic linear model. It becomes more complex when mixed models are involved, because of the presence of two kinds of effects with completely different characteristics and roles. Among others, a key aspect of linear mixed model selection is how to identify the real important random effects, i.e., those whose coefficients vary among subjects.

It is important to note that the exclusion of relevant effects has a drawback on the estimation of the fixed effects: their variance–covariance matrix would be underfitted and the estimation of the variances related to the fixed part estimates would be biased. The inclusion of irrelevant random effects in a model, on the other hand, would lead to a singular variance–covariance matrix of random effects, producing instability in the model (Ahn et al. 2012). As pointed out by Müller et al. (2013), most procedures focus on the selection of fixed effects exclusively. Only Chen and Dunson (2003) and Greven and Kneib (2010) worked on random part selection before Müller et al. (2013). There are obvious difficulties due to computational issues in selecting only the random part, that is why the researchers who worked on the random effects, after Müller et al. (2013), optimize with respect to the fixed part, too, excepted for Li and Zhu (2013). In recent years, in fact, it has been easy to find procedures selecting both the effects .

It is worth noting that since the LMMs are a special case of Generalized LMMs, we obviously excluded from the current review all those methods built mainly for selecting effects in the GLMMs, such as Hui et al. (2017). Moreover, this review doesn't include works based on graphical tools for model selection if these graphical representations are referred to methods already existent in the literature. This is the case, for example, of Sciandra and Plaia (2018) who adapt an available graphical representation to the class of mixed models, in order to select the fixed effects conditioning on the random part and covariance structure, and of Singer et al. (2017) who discuss different diagnostic methods focusing on residual analysis but also addressing global and local influence, giving general guidelines for model selection.

This review mentions the available theoretical properties corresponding to the different methodologies, with comparisons among them whereas it's possible.

Müller et al. (2013) classified the proposed methods by considering four different kinds of procedures: information criteria (such as Akaike information criterion, Bayesian information criterion); shrinkage methods such as LASSO and adaptive LASSO; the Fence method; and some Bayesian methods.

In this paper, we prefer to cluster methods according to which part of the model, fixed, random or both, they focus on. The paper is organized as follows. In Sect. 2, we present the structure and notation of a linear mixed model and we discuss some problems occurring in selection models. In Sect. 3, we give an overview of model selection procedures within the LMMs framework that are useful for selecting linear mixed models, by considering the classification proposed in Müller et al. (2013). In Sects. 4 and 5, we describe the methods grouped according to the part of the model selected, i.e., fixed and both, respectively. Finally, we examine some simulations in Sect. 6 and conclude with a brief discussion and some conclusions in Sect. 7. Moreover, to help the reader decide which method to prefer, according to his own data, we include two Tables 2 and 3, that summarize the main features of each method.

## 2 LMM and the linear mixed model selection problem

Suppose data are collected from  $m$  independent groups of observations (called clusters or subjects in longitudinal data). The response variable  $Y_i$  is specified in the linear mixed models at cluster level as follows:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \quad i = 1, 2, \dots, m, \quad (1)$$

where  $Y_i$  is a  $n_i$  dimensional vector of observed responses,  $X_i$  and  $Z_i$  are the known  $n_i \times p$  and  $n_i \times q$  matrices of covariates related to the fixed effects and to the random effects, respectively,  $\beta$  is the  $p$ -vector of unknown fixed effects,  $b_i$  is the  $q$ -vector of unobserved and independent random effects and  $\epsilon_i$  is the vector of unobserved random errors. Let us assume that  $b_i$ s are independent of  $\epsilon_i$ s and that they are independent and identically distributed random variables for each group of observations in the following way:

$$b_i \sim N_q(0, \Psi), \quad \epsilon_i \sim N_{n_i}(0, \Sigma), \quad (2)$$

where  $\Psi$  is a  $q \times q$  positive definite matrix and  $\Sigma$  is a  $n_i \times n_i$  positive definite matrix. Consequently, the response vector follows a multivariate normal distribution,  $Y_i \sim N_{n_i}(X_i\beta, V_i)$ , where the variance–covariance matrix is given by  $V_i = Z_i\Psi Z_i' + \Sigma$ . The vectorized form of the model is:

$$Y = X\beta + Zb + \epsilon, \quad (3)$$

where all elements concern all macro units; therefore,  $Y$  is a  $n$ -dimensional vector ( $n = \sum n_i$ ),  $X$  and  $Z$  are the known  $n \times p$  and  $n \times q$  matrices of covariates related to the fixed effects and to the random effects, respectively,  $\beta$  is the  $p$ -vector of unknown fixed effects,  $b$  is the  $q$ -vector of unobserved and independent random effects and  $\epsilon$  represents the vector of unobserved random errors.

The selection of linear mixed-effects models implies the selection of the “true” fixed parameters and/or the “true” random effects. Even if there exists a kind of estimation for  $b$ , the Best Linear Unbiased Predictors [BLUP, see Eq. (7)], the correct investigation for identifying  $b$  requires to estimate its  $q(q+1)/2$  variance–covariance parameters. Let  $\tau$  denote the  $s$ -vector filled with all distinctive components in the variance–covariance matrices  $\Psi$  and  $\Sigma$ . A random effect is not relevant if its variance–covariance elements, for all observations, are zero (Ahn et al. 2012); hence, it suffices to identify the nonzero diagonal components in  $\Psi$  (Wu et al. 2016) correctly and, also, their related covariance terms, for avoiding the drawback of excluding random effects correlated to some explanatories.

We call  $\theta = (\beta, \tau)$  the overall set of parameters relevant in a linear mixed model. This set represents the whole group of the parameters related to the true model generating data. Let us identify the selection of linear mixed models with  $M \in \mathcal{M}$ , where  $\mathcal{M}$  is the countable set containing all candidate models involved in the selection. The number of candidate models used depends on some contextual considerations: some variance–covariance components could be known or assumed to be known; some authors could focus only on nested models; or, still, the classic null model (the one with intercept only) could not be admitted among the set of candidate models (see Sect. 7 for further details).

The conditional log-likelihood for model (3) is given by:

$$l(\theta|b; y) = \log f_y(y|b; \theta) = -\frac{1}{2} \left\{ \log |\Sigma| + (y - X\beta - Zb)' \Sigma^{-1} (y - X\beta - Zb) \right\} - \frac{n}{2} \log(2\pi), \quad (4)$$

while the marginal likelihood is:

$$l(\theta; \mathbf{y}, \mathbf{b}) = \log f_{\mathbf{y}}(\mathbf{y}; \mathbf{b}, \theta) = -\frac{1}{2} \{ \log |\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \}. \tag{5}$$

For fixed  $\boldsymbol{\tau}$ , the optimization process of the joint log-likelihood leads to an estimate of  $\boldsymbol{\beta}$  that is similar to a generalized least squares estimator:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}) = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}. \tag{6}$$

The most popular approach for predicting  $\mathbf{b}$  is an empirical Bayesian method, which uses the posterior distribution  $f(\mathbf{b}|\mathbf{y})$  yielding the following BLUP prediction:

$$\hat{\mathbf{b}}(\boldsymbol{\tau})_{\text{BLUP}} = \boldsymbol{\Psi} \mathbf{Z}' \mathbf{V}^{-1} \{ \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}) \}. \tag{7}$$

The same solutions of  $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$  and  $\hat{\mathbf{b}}(\boldsymbol{\tau})_{\text{BLUP}}$  can be obtained by solving Henderson’s linear mixed model equations (Müller et al. 2013):

$$\begin{bmatrix} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \boldsymbol{\Psi}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}) \\ \hat{\mathbf{b}}(\boldsymbol{\tau}) \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix} [\boldsymbol{\Sigma}^{-1} \mathbf{y}]. \tag{8}$$

Although consistent, the ML estimator of variance–covariance parameters is known to be biased in small samples. Hence, the restricted maximum likelihood estimators (REML) are used:

$$l_R(\boldsymbol{\tau}) = -\frac{1}{2} \{ \log |\mathbf{V}| + \log |\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}| + \mathbf{y}' \mathbf{P}^{-1} \mathbf{y} \}, \tag{9}$$

where  $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1}$  (Müller et al. 2013). Thus, the simple ML estimators for  $\boldsymbol{\beta}$  and  $\boldsymbol{\tau}$  will here forth be indicated as  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\tau}}$ , while the REML estimators as  $\hat{\boldsymbol{\beta}}_R$  and  $\hat{\boldsymbol{\tau}}_R$ .

It is important to note that in many papers dealing with LMMs some authors use the  $\sigma^2$  scaled versions of  $\boldsymbol{\Psi}$  and  $\boldsymbol{\Sigma}$ , which are  $\sigma^2 \boldsymbol{\Psi}_*$  and  $\sigma^2 \boldsymbol{\Sigma}_*$ . Then we are going to use, in the description of the methods, the symbol  $*$  for those variance–covariance matrices scaled by  $\sigma^2$ .

### 3 Introduction to model selection criteria

Within the framework of linear mixed-effect models, a large number of selection criteria are available in the literature. Model selection criteria are frequently set up by building estimators of discrepancy measures, which evaluate the distance between the “true” model and an approximating model fitted to the data.

### 3.1 AIC and its modifications

The most widely used criteria for model selection are the information criteria. Their application consists in finding the model that minimizes a function, in the form of a loss function plus a penalty, usually dependent on model complexity. The Akaike information criterion (AIC), introduced by Akaike (1992), is the most popular method. The Akaike information criterion is based on the Kullback–Leibler distance between the true density of the distribution generating the data,  $\mathbf{y}$ , and, the approximating model for fitting the data,  $g(\boldsymbol{\theta})$  (Vaida and Blanchard 2005). With his criterion, Akaike tried to combine point estimation and hypothesis testing into a single measure, thus formalizing the concept of finding a good approximation of the true model in a predictive view. In this sense, a good model is the one that is able to generate predictive values (independent of the real data) as close as possible to the observed data. AI is given by  $-2E_{f(\mathbf{y})}E_{f(\mathbf{y}^*)} \log g\{\mathbf{y}^*; \hat{\boldsymbol{\theta}}(\mathbf{y})\}$ , where  $\hat{\boldsymbol{\theta}}$  is an estimator of  $\boldsymbol{\theta}$ , while  $\mathbf{y}^*$  represents the predictive set of data obtained from the fitted model and independent of  $\mathbf{y}$ . Vaida and Blanchard (2005) defined a new version of AI by conditioning the distribution  $f(\mathbf{y}; \boldsymbol{\theta})$  to the clusters. Hence, the conditional AI (cAI) uses the conditional distribution  $f(\mathbf{y}; \boldsymbol{\theta}, \mathbf{b})$  as follows:

$$-2E_{f(\mathbf{y}, \mathbf{b})}E_{f(\mathbf{y}^* | \mathbf{b})} \log g\{\mathbf{y}^*; \hat{\boldsymbol{\theta}}(\mathbf{y}), \hat{\mathbf{b}}(\mathbf{y})\},$$

where  $\hat{\mathbf{b}}(\mathbf{y})$  is the estimator of  $\mathbf{b}$ . It should be noted that  $\mathbf{y}^*$  and  $\mathbf{y}$  have to be considered conditionally independent of  $\mathbf{b}$  and belonging to the same conditional distribution  $f(\cdot | \mathbf{b})$ . These two last assumptions imply that they have the same random effects  $\mathbf{b}$ .

The underlying reasoning of the criterion based on the Akaike information criterion is not to identify the true model generating the data, but the best approximation of it, which adapts well to the data. The estimators employed for measuring AI and cAI are known as Akaike information criterion and conditional Akaike information criterion, respectively, and they are both biased for finite samples. They approximate their own information as minus twice the relative log-likelihood function plus a penalty term,  $a_n(d_M)$ , which tries to adjust the bias. The marginal AIC, defined by Vaida and Blanchard (2005), has the following generic formula:

$$\text{mAIC} = -2l(\hat{\boldsymbol{\theta}}) + 2a_n(p + q)$$

where  $a_n = 1$  or  $a_n = n/(n - p - q - 1)$  in small samples (Vaida and Blanchard 2005; Sugiura 1978). The conditional Akaike information criterion (cAIC—Vaida and Blanchard 2005) provides a procedure for selecting variables in LMMs with the purpose of predicting specific clusters or random effects, since the mAIC is inappropriate when the focus is on clusters and not on the population. For predicting at cluster level, the likelihood needs to be computed conditionally on the clusters and the random effects  $\mathbf{b}_i$  need to be considered as parameters. Hence, for computing the cAIC, the terms to estimate are the  $p + q + s$  parameters in  $\boldsymbol{\theta}$ . If all the variance elements  $\boldsymbol{\tau}$  are known, the  $q$  random effects  $\mathbf{b}$  are predicted by the best linear unbiased predictor (BLUP) or using an estimated version of BLUP (Eq. 7). The generic formula for cAIC is:

$$cAIC = -2l(\hat{\theta}|\hat{\mathbf{b}}) + 2a_n(\rho + 1) \tag{10}$$

where  $\rho$  is connected to the effective degrees of freedom used in estimating  $\beta$  and  $\mathbf{b}$ . Many authors (Shang and Cavanaugh 2008; Kubokawa 2011; Vaida and Blanchard 2005; Greven and Kneib 2010; Srivastava and Kubokawa 2010; Liang et al. 2008) have tried to reduce the bias of mAIC and cAIC, working on the penalty term in different ways, i.e., taking into account the MLE estimator or the REML estimator for  $\theta$ , distinguishing if variance–covariance matrices are known or unknown. A clear and complete overview of all penalties used in the literature is available in Müller et al. (2013), Sects. 3.1 and 3.2.

### 3.2 Mallows’s $C_p$

Another criterion, based on a discrepancy measure (Gauss discrepancy) and used for choosing the model nearest to the true one, is given by Mallows’  $C_p$ .

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p,$$

with  $SSE_p$  and  $p$  representing, respectively, the error sum of squares and the number of parameters of the reference model and  $\hat{\sigma}^2$  an estimate of  $\sigma^2$  (Gilmour 1996). Some variants on Mallows’  $C_p$  are provided by Kubokawa (2011) and are clearly presented by Müller et al. (2013).

### 3.3 BIC

The Bayesian information criterion is based on the marginal distribution of  $\mathbf{y}$ , which requires the full prior information about all parameters  $(\beta, \theta)$  to be computed:

$$f(\mathbf{y}) = \int \int f_m(\mathbf{y}|\beta, \theta)\pi(\beta, \theta)d\beta d\theta. \tag{11}$$

BIC, proposed by Schwarz (1978), is an approximation of  $-2 \log\{f_\pi(\mathbf{y})\}$ , free of any prior distribution setup:

$$BIC = -2l(\hat{\theta}) + (p + q) \log(N). \tag{12}$$

Since BIC is a Bayesian procedure for model selection, it requires prior distributions. Kubokawa and Srivastava (2010) derived the expression of EBIC, an intermediate method between BIC and full Bayesian variable selection tools. The EBIC procedure employs partial non-subjective prior distribution only for the parameters of interest, ignoring the nuisance parameters in terms of distributional assumptions.

### 3.4 Shrinkage

Often, it is not feasible to compute information criteria in variable selection when  $p$  and/or  $q$  are large, i.e., in high-dimensional settings, when one deals with classic

linear models. Hence, in this sense, shrinkage methods such as the least absolute shrinkage and selection operator, LASSO (Tibshirani 1996), and its extensions such as the adaptive LASSO, ALASSO (Zou 2006), the elastic net (Zou and Hastie 2005) or the smooth clipped absolute deviation, SCAD (Fan and Li 2012), have been proposed in the literature. When using these techniques, thanks to a penalization system, some coefficients are shrunk toward zero, while at the same time, the once influential on response are estimated to be nonzero. The shrinkage procedures are applicable to either the least squares or the likelihood functions. For the sake of simplicity, the penalized likelihood function is readopted in the case of the classical linear model:

$$-\sum_{i=1}^n l_i(\boldsymbol{\beta}; \mathbf{y}_i) + n \sum_{j=1}^p p_\lambda(\|\boldsymbol{\beta}\|_\ell), \quad (13)$$

where  $\|\boldsymbol{\beta}\|_\ell$  is the  $\ell$ -th norm of  $\boldsymbol{\beta}$ . Taking into account that  $\ell_1$  corresponds to work with the LASSO, while  $\ell_2$  refers to ridge estimation. The adaptive LASSO is an extension of LASSO. It involves the addition of some weights depending on the  $\ell$ -th norm of  $\boldsymbol{\beta}$ , i.e.,  $p_\lambda(\|\boldsymbol{\beta}\|_\ell) = \lambda_j \|\boldsymbol{\beta}\|_\ell / 2$ , with  $\lambda_j = \lambda / \|\boldsymbol{\beta}\|_\ell$ , where  $\ell$  is an additional parameter often considered equal to 1.

The generic SCAD penalty on  $\boldsymbol{\theta}$  introduced by Fan and Li (2001) works on the first derivative of  $p_\lambda(\|\boldsymbol{\theta}\|)$ :

$$p'_{\lambda_j}(\|\boldsymbol{\theta}\|) = \lambda \left\{ I(\boldsymbol{\theta} \leq \lambda) + \frac{(a\lambda - \boldsymbol{\theta})_+}{(a-1)\lambda} I(\boldsymbol{\theta} - \lambda) \right\}. \quad (14)$$

For the solution of  $\boldsymbol{\theta}$ , Fan and Li (2001) provided an algorithm via local quadratic approximations.

### 3.5 MDL principle

The minimum description length (MDL) principle originates from data compression literature and Rissanen (1986) who developed it to “understand” the observed data; it represents a valid statistical criterion employed for selecting linear mixed models. This method aims to detect the best model approximating the observed data, among a pool of candidate models, through a data compression process based on the code length needed to describe the data. A model can be described using fewer symbols than those necessary to describe the data. Usually, this criterion is used in the presence of independent data. Li et al. (2014) propose a MDL principle for fixed effects selection when there is a correlation between observations within clusters. The principle is presented as a good trade-off between AIC, thanks to its asymptotic optimality, and BIC, because of its consistency property. The proposed criterion is a hybrid form of MDL which merges a two stage description length and the mixture MDL with the dependent data.



### 4 Fixed effects selection

AIC and its modifications consist in finding the model that minimizes a function in the form of a loss function plus a penalty, which measures model complexity. Kawakubo and Kubokawa (2014) and Kawakubo et al. (2014) propose a modified conditional AIC and a conditional AIC under covariate shift in Small Area Estimation (SAE), respectively. For linear mixed model selection, random intercept model selection in particular, in the small area estimation, Marhuenda et al. (2013) work on two variants of AIC and two variants of the Kullback symmetric divergence criterion (KIC), defined as:

$$KIC = -2 \log f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + 3(p + 1).$$

Kawakubo and Kubokawa (2014) and Kawakubo et al. (2018) provide a modified version of the exact cAIC (McAIC), because the cAIC suggested by Vaida and Blanchard (2005) is highly biased when the candidate models do not include the true model generating the data (underspecified cases). They assume that  $\boldsymbol{\Psi} = \sigma^2 \boldsymbol{\Psi}_*$ ,  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{n_i}$ , and extend cAIC to a procedure that could be valid both for the overspecified cases (situations in which the true model is included among the candidate models) and for the underspecified cases. The modified conditional AIC is given by:

$$McAIC = -2 \log f(y|\hat{\mathbf{b}}_j, \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2) + \widehat{\Delta}_{cAI}, \tag{15}$$

where  $\widehat{\Delta}_{cAI}$  is the estimate of the bias of cAIC, estimated by:

$$\widehat{\Delta}_{cAI} = B^* + \widehat{B}_1 + \widehat{B}_2 + \widehat{B}_3, \tag{16}$$

where  $B^*$  is a function of  $V^{-1}$  and  $B_1, B_2$  and  $B_3$  are functions of  $V$  and  $X$ . The authors demonstrate that  $B^*, \widehat{B}_1, \widehat{B}_2$  and  $\widehat{B}_3$  have distributions proportional to  $\chi^2$  with degrees of freedom opportunely quantified and, in the overspecified case,  $\widehat{\Delta}_{cAI}$  reduces to  $B^*$ , i.e.,  $McAIC=cAIC$  by Vaida and Blanchard (2005).

When the variable selection problem focuses on finding a set of significant variables for a good prediction, Kawakubo et al. (2014) propose a cAIC under covariate shift (CScAIC). They derive the cAIC of Vaida and Blanchard (2005) under the covariate shift for both known and unknown variances  $\sigma^2$  and  $\boldsymbol{\Psi}_*$  and with  $\boldsymbol{\Sigma}_*$  assumed to be known.

The proposed criterion replaces, in the formula of the classic cAIC, the conditional density of  $\mathbf{y}$  (the vector of the observed responses) given  $\mathbf{b}$ , with the conditional density of  $\tilde{\mathbf{y}}$  (the vector of observed responses in the “predictive model”:  $\tilde{\mathbf{y}} = \tilde{X}\boldsymbol{\beta} + \tilde{Z}\mathbf{b} + \tilde{\boldsymbol{\epsilon}}$ , a LMM with same regression coefficients  $\boldsymbol{\beta}$  and random effects  $\mathbf{b}$ , but different shifted covariates) given  $\mathbf{b}$ .

$$CScAIC = -2 \log g(\tilde{\mathbf{y}}|\hat{\mathbf{b}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + B_c^*, \tag{17}$$

when  $\sigma^2$  is unknown and estimated by its ML estimator and  $B_c^*$  is the bias correction.

Lombardía et al. (2017) introduce a mixed generalized Akaike information criterion, xGAIC, for SAE models. One typical model used in the field of SAE is the Fay–Herriot model, which is a particular type of LMMs containing only one random effect, the intercept. The clusters are represented by areas and the model in Eq. (1) for each area is reduced to:  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{b}_i + \epsilon_i$ , with  $i = 1, 2, \dots, m$ .

Instead of the usual AIC types based only on the marginal or the conditional log-likelihood, the authors propose to use a new AIC, based on a combination of both the log-likelihood functions. The quasi-log-likelihood used for deriving the new statistics is the following:

$$\log(l_x) = -\frac{1}{2}m \log(2\pi) - \frac{1}{2} \log |\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})' \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}), \tag{18}$$

where  $\boldsymbol{\mu} = E(\mathbf{Y}|\mathbf{b})$ . The generalized degrees of freedom (xGDF), linked to the quasi-log-likelihood in Eq. (18), takes into account the expectation and covariance with respect to the marginal distribution of  $\mathbf{Y}$ :

$$\text{xGDF} = \sum_{i=1}^m \frac{\partial E_y(\hat{\boldsymbol{\mu}}_i)}{\partial (\mathbf{X}_i \boldsymbol{\beta})} = \sum_{i=1}^m \sum_{j=1}^m \mathbf{V}^{ij} \text{cov}(\hat{\boldsymbol{\mu}}_i, \mathbf{y}_j), \tag{19}$$

where  $\mathbf{V}^{ij}$  is the  $ij$ -element of the matrix  $\mathbf{V}^{-1}$ . Combining the  $\log(l_x)$  with xGDF, the mixed generalized AIC is finally defined as:

$$\text{xGAIC} = -2 \log(l_x) + \text{xGDF}. \tag{20}$$

Han (2013) derives the closed form for the unbiased conditional AIC when the linear mixed model is reduced to the Fay–Herriot model. The author proposed a more suitable cAIC for three different approaches to fitting the model: the unbiased quadratic estimator (UQE), the REML estimator and the ML estimator. The unbiased cAIC for the Fay–Herriot model has the same form as for the classical LMMs, with i.i.d. errors (see Eq. 10), where the degrees of freedom are measured by  $\Phi = \sum_{i=1}^m \frac{\partial \mathbf{X}_i \hat{\boldsymbol{\beta}}}{\partial \mathbf{Y}_i} = \text{tr}(\frac{\partial \mathbf{X}' \hat{\boldsymbol{\beta}}}{\partial \mathbf{Y}})$ , which is computationally expensive, because  $\mathbf{X}_i \hat{\boldsymbol{\beta}}$  is not a linear estimator through  $\hat{\sigma}_b^2$  and the derivatives therein depend on the specific choice of estimating  $\sigma_b^2$ :

$$\text{cAIC} = -2 \log f(\mathbf{y}|\mathbf{b}, \boldsymbol{\theta}) + 2\Phi. \tag{21}$$

If  $\hat{\sigma}_b^2 = 0$ , whatever is the method used for estimating it, then  $\Phi = p$ , otherwise when  $\hat{\sigma}_b^2 > 0$  the way of measuring  $\Phi$  is different. If the unbiased quadratic estimate method is used, then:

$$\Phi = \hat{p} + 2(m - p)^{-1} \mathbf{r}' \mathbf{S} \boldsymbol{\Sigma}^{-1} \mathbf{P}^* \mathbf{r}_S. \tag{22}$$

If  $\hat{\sigma}_b^2 > 0$  is the REML or ML estimate:

$$\Phi = \hat{\rho} - 2 \left( \frac{\partial \hat{\sigma}}{\partial \sigma_b^2} \right)^{-1} r_s' \hat{\Sigma}^{-1} P^* S \Sigma^{-1} P^* r_s, \tag{23}$$

with  $\frac{\partial \hat{\sigma}}{\partial \sigma_b^2} = tr((\Sigma^{-1} P^*)^2) - 2r_s' \hat{\Sigma}^{-1} P^* r_s$  in the case of REML or  $\frac{\partial \hat{\sigma}}{\partial \sigma_b^2} = tr(\Sigma^{-2}) - 2r_s' \hat{\Sigma}^{-1} P^* r_s$  for ML estimating process,  $P^* = I - X(X' \Sigma^{-1} X)^{-1} \Sigma^{-1}$ ,  $r$  the residuals from the OLS estimation for  $\beta$  and  $r_s = \Sigma^{-1} P^* Y$  the standardized residuals obtained from the GLS estimation for  $\beta$ . The closed-form cAIC results to be an unbiased estimator for the conditional AI for the Fay–Herriot model.

It is worth mentioning (Lahiri and Suntornchost 2015) for their contribution to the selection of fixed effects in LMMs with applications in SAE models, even if their proposal doesn't concern a modification of some Information Criteria. The authors define an alternative to the usual Mean Square Error and Mean Square Total, estimating them with  $\widehat{MSE} = MSE - \overline{D}_w$  and  $\widehat{MST} = MST - \overline{D}$ , respectively, where  $\overline{D}_w = \sum_{i=1}^m ((1 - h_{ii}) D_i) / (m - p)$  and  $\overline{D}_w = \sum_{i=1}^m D_i / m$ , with  $h_{ii} = \mathbf{x}'_i (X' X)^{-1} \mathbf{x}_i$ . They suggest to use  $\widehat{MSE}$  and  $\widehat{MST}$ , because under standard regularity conditions these measures tend to the true MSE and MST with probability one, as the number of areas increases. But, since for small areas  $\widehat{MSE}$  and  $\widehat{MST}$  could be negative, the authors suggest an alternative to their estimates, through the function  $h(x, b)$  in Eq. (24) which guarantees to obtain positive values for them:

$$h(\mathbf{x}, \mathbf{b}) = \frac{2\mathbf{x}}{1 + \exp\left(\frac{2\mathbf{b}}{\mathbf{x}}\right)}. \tag{24}$$

This function allows to figure out new estimators in the following way:  $\widehat{MSE} = h(MSE, \overline{D}_w)$  and  $\widehat{MST} = h(MST, \overline{D})$ .

Kubokawa and Srivastava (2010) derive an exact expression of the empirical Bayes information criterion (EBIC) for selecting the fixed effects in a linear mixed model. Their criterion represents an intermediate solution between BIC and the full Bayes variable selection methods, because it exploits the partitioning of the vector of parameters  $(\beta, \tau_*, \sigma)$  into two sub-vectors, one for the parameters of interest  $(\beta)$  and the other one for the nuisance parameters  $(\tau_*, \sigma)$ . Specifically, it works with a partial non-subjective prior distribution for only the parameters of interest, ignoring a prior setup for the nuisance parameters and applying the Laplace approximation for this one. The full prior distribution  $\pi(\beta, \tau)$  can be written through a proper prior distribution,  $\pi_1(\beta | \tau, \lambda)$ , which is not completely subjective because of its dependence on an unknown hyperparameter  $\lambda$ :

$$\pi(\beta, \tau) = \pi_1(\beta | \tau, \lambda) \pi_2(\tau).$$

The two authors derive EBIC, starting from the BIC but they approximate the marginal distribution of  $y$ ,  $f(y)$ , with one of its two components, i.e., the conditional marginal density based on the partial prior distribution,  $m_1(y | \tau, \lambda)$ :

$$m_1(\mathbf{y}|\boldsymbol{\tau}, \lambda) = \int f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\tau})\pi_1(\boldsymbol{\beta}|\boldsymbol{\tau}, \lambda)d\boldsymbol{\beta}.$$

After estimating  $\lambda$ ,  $\hat{\lambda} = \arg \max_{\lambda} m_1(\mathbf{y}|\hat{\boldsymbol{\tau}}, \lambda)$  using a consistent estimator of  $\boldsymbol{\tau}$ , the EBIC is obtained as follows:

$$\begin{aligned} \text{EBIC} &= -2 \log\{m_1(\mathbf{y}|\hat{\boldsymbol{\tau}}, \hat{\lambda})\} + \dim(\boldsymbol{\theta}) \log(n) \\ &= -2 \log\{m_1(\mathbf{y}|\hat{\sigma}^2, \hat{\boldsymbol{\tau}}_*, \hat{\lambda})\} + (d + 1) \log(n). \end{aligned}$$

The derivation of the EBIC neglects the full prior distribution, but it uses the non-subjective prior distribution  $\pi_1(\boldsymbol{\beta}|\sigma^2, \lambda)$ , assuming that, conditioned to  $\sigma^2$ , it assumes a multivariate normal distribution:

$$\pi_1(\boldsymbol{\beta}|\sigma^2, \lambda) = N_p(\mathbf{0}, \sigma^2\lambda^{-1}W),$$

with an unknown scalar  $\lambda$  and a  $p \times p$  known matrix  $W$ . A possible choice for  $W$  could be the so called Zellner's  $q$ -prior,  $W_q = n(\mathbf{X}'\mathbf{X})^{-1}$ . The authors demonstrate that EBIC is a consistent estimator.

Wenren and Shang (2016) and Wenren et al. (2016) work on conditional conceptual predictive statistics and on marginal conceptual predictive statistics for linear mixed model selection, respectively. The conditional  $C_p$  is formalized in both cases in which  $\sigma^2$  and  $\Psi_*$  are known and unknown. The marginal  $C_p$  appears to be useful in two ways, both when the sample size is small and when there is a high correlation between the observations. Wenren et al. (2016) propose a modified variant of Mallows'  $C_p$  when there is a correlation between observations, even if not known. They work under the assumption that  $\Psi = \sigma^2\Psi_*$  and  $\Sigma = \sigma^2I_{n_i}$ . They assume that the estimator of the correlation matrix (for the candidate model) is consistent. The formalization of Modified  $C_p$  ( $\text{MC}_p$ ) is as follows:

$$\text{MC}_p = \frac{\text{SS}_{\text{RES}}}{\hat{\sigma}^2} + 2p - n, \quad (25)$$

where  $\text{SS}_{\text{RES}}$  is the residual sum of squares for the candidate model,  $\hat{\sigma}^2$  represents an asymptotically unbiased estimator for  $\sigma^2$  and it is computed for the largest candidate model.  $\text{MC}_p$  is a biased estimator for the expectation of the transformed marginal Gauss discrepancy. However, it is an unbiased estimator of  $\Delta_{C_p}(\boldsymbol{\theta})$ , if the true model is included in the pool of all candidate models. For better performance, they also provide a more accurate estimator:

$$\text{IMC}_p = \frac{(n - p_* - 2)\text{SS}_{\text{Res}}}{\text{SS}_{\text{Res}}^*} + 2p - n + 2, \quad (26)$$

using the symbol  $*$  for referring to the largest candidate model.  $\text{IMC}_p$  results to be an asymptotically unbiased estimator of the expected overall transformed Gauss discrepancy. It is preferred to  $\text{MC}_p$  because it avoids the bias introduced by  $\frac{1}{\hat{\sigma}^2}$  used for estimating  $\frac{1}{\sigma^2}$ .

Wenren and Shang (2016) provide another conceptual predictive statistics for selecting a linear mixed model if one is interested in predicting specific clusters or random effects. Inspired by cAIC and conditional Mallows’s  $C_p$ , they construct two versions of the conditional  $C_p$  ( $CC_p$ ), according to known or unknown variance components. They work under the assumption that  $\Psi = \sigma^2\Psi_*$  and  $\Sigma = \sigma^2I_{n_i}$ , too. Assuming that  $\sigma^2$  and  $\Psi_*$  are known, they combine a goodness of fit term with a penalty term, and propose  $CC_p$  defined as:

$$CC_p = \frac{SS_{Res}}{\sigma^2} + K, \tag{27}$$

where  $K = 2\rho - n$  defines the effective degrees of freedom with  $\rho = tr(H_1)$  (Hodges and Sargent 2001). If the variance components are unknown,  $\Psi_*$  is substituted by its ML  $\hat{\Psi}_*$  or restricted MLE  $\hat{\Psi}_{*R}$  estimate. The effective degrees of freedom  $\rho$  is also estimated,  $\hat{\rho} = tr(\hat{H}_1)$  where  $\hat{H}_1 = \hat{H}_1(\hat{\Psi}_*)$  or  $\hat{H}_1 = \hat{H}_1(\hat{\Psi}_{*R})$ .  $\sigma^2$  is estimated in the largest candidate model (\*) through  $\hat{\sigma}^2 = \frac{SS_{Res}^*}{N-p_*}$ , an unbiased estimator of  $\sigma^2$ . For further details about  $\hat{H}_1$  see Hodges and Sargent (2001). By substituting the variance components by their estimators in a suitable way, the conditional  $C_p$  is:

$$CC_p = (n - p_*) \frac{SS_{Res}}{SS_{Res}^*} + \hat{K}, \tag{28}$$

with  $\hat{K} = 2\hat{\rho} - n$  indicating the (ML or REML) estimated penalty term.

Kuran and Özkale (2019) provide a conditional conceptual predictive statistic, too, in the framework of LMMs but applying a ridge estimator for overcoming multicollinearity problems. Like Wenren and Shang (2016), they work under the assumption that  $\Psi = \sigma^2\Psi_*$  and  $\Sigma = \sigma^2I_{n_i}$ . When we have to manage multicollinearity problems, usually we delete one or more variables related to the fixed effects, but this could cause some not irrelevant consequences: The fitted candidate model could be misspecified. For this reason, the two authors are motivated to require to the ridge estimator and the ridge predictor for LMMs proposed by Liu and Hu (2013) and Özkale and Can (2017):

$$\hat{\beta}_k = (X'V_*^{-1}X + kI_p)^{-1}X'V_*^{-1}y, \tag{29}$$

$$\hat{b}_k = \Psi_*Z'V_*^{-1}(y - X\hat{\beta}_k), \tag{30}$$

where  $k$ , a positive real number, represents the ridge biasing parameter. Its selection is obtained by minimizing a generalized cross-validation in the predictive step, while the same is measured through the minimization of the scalar mean square error of the ridge regression, in the estimation process (see Özkale and Can 2017). Following Wenren and Shang (2016), they propose two versions of the conditional conceptual predictive statistic, distinguishing the case in which  $\sigma^2$  and  $\Psi_*$  are known or they aren't. The proposed criteria are the same of  $CC_p$  in Eqs. (27) and in (28), substituting the effective degrees of freedom under ridge estimator for LMMs,  $\rho_k = tr(H_{1k})$ , to  $\rho$ ,  $\hat{\rho}_k = tr(\hat{H}_{1k})$  to  $\hat{\rho}$  and  $SS_{Res,k} = (y - \hat{y}_k)'(y - \hat{y}_k)$  to  $SS_{Res}$ , where  $H_{1k} = I_n - V_*^{-1}[I_n - X(X'V_*^{-1}X + kI_p)^{-1}X'V_*^{-1}]$ .

Li et al. (2014) proposed a two-stage method based on the MDL principle. When  $\beta$  is the only unknown parameter, encoding the estimated parameter represents the first stage. Then, all the sequence of data with the distribution  $f_{\hat{\theta}}$  is encoded. The resulting total length code used for transmission is equivalent to BIC:

$$L(\mathbf{y}) = L(\mathbf{y}|\hat{\theta}) + L(\hat{\theta}) = -\log f_{\hat{\theta}}(\mathbf{y}) + \frac{p}{2} \log(m).$$

The penalty term, which measures the precision used to encode each parameter, is  $\log(m)/2$  with a uniform distribution. The authors follow the idea of the mixture MDL proposed by Hansen and Yu (2003), which assumes a mixture distribution induced by the user-defined probability distribution  $w(\theta)$  on the parameter space  $\Theta$ . They assume that  $\Sigma = \sigma^2 I_{n_i}$ ,  $\beta \sim N(0, c\sigma^2(X_i' \Psi_{*i}^{-1} X_i)^{-1})$  and the hyperparameter  $c$  is a scalar constrained to be nonnegative. As regards the distribution of  $\sigma^2$ , an inverse gamma distribution is assumed with parameters  $(a, 3/2)$ . Hence, the mixture description length of  $\mathbf{y}$  is expressed as:

$$-\log m(\mathbf{y}) = -\log \int f_{\theta}(\mathbf{y})w(\theta)d\theta.$$

The code length is minimized with respect to  $c \geq 0$  and the resulting  $\hat{c}$  is plugged into the code length expression, leading to the  $\text{IMDL}_0$  criterion. The expression of the final code length, with only  $\beta$  unknown and ignoring the impact of  $\mathbf{b}$ , is:

$$\left\{ \begin{array}{ll} \frac{1}{2} \left\{ \sum_{i=1}^n \mathbf{y}'_i \Sigma_i^{-1} \mathbf{y}_i - \text{FSS}_{\sigma} + p \left[ 1 + \log \left( \frac{\text{FSS}_{\sigma}}{p} \right) \right] + \log n \right\}, & \text{if } \text{FSS}_{\sigma} > p, \\ \frac{1}{2} \sum_{i=1}^n \mathbf{y}'_i \Sigma_i^{-1} \mathbf{y}_i, & \text{otherwise,} \end{array} \right.$$

$\text{FSS}_{\sigma} = (\sum_{i=1}^n \mathbf{y}'_i \Sigma_i^{-1} X_i)(\sum_{i=1}^n X_i' \Sigma_i^{-1} X_i)^{-1}(\sum_{i=1}^n X_i' \Sigma_i^{-1} \mathbf{y}_i)$  and  $(\log n)/2$  represents the code length necessary for transmitting  $\hat{c}$ . If  $\text{FSS}_{\sigma} \leq p$ ,  $\hat{c} = 0$  and this implies that all fixed effects are null. The  $\text{IMDL}_0$  criterion has the same structure of penalized likelihoods such as AIC and BIC, but with a proper data-adaptive penalty, depending on the covariance matrices. The two-stage mixture MDL principle, in the most realistic case with  $(\sigma^2, \Psi_*)$  unknown, it consists in estimating  $\Psi_*$  and plugging it into the code length. Minimization of the code length function, with respect to  $a$  and  $c$ , leads to an even more complex IMDL structure. The authors showed that the MDL criteria possess the selection consistency of BIC for finite-dimensional models.

Marino et al. (2017) give a really important contribution to the selection of relevant covariates in the LMMS, since their proposal is aimed at mixed models with missing data. Their work deals with selection of covariates in multilevel models, hence applicable to linear mixed models being a two-level model. The authors work under the assumption that  $\Psi = \sigma^2 \Psi_*$  and  $\Sigma = \sigma^2 I_{n_i}$  and that parts of the covariates are ignorable missing, hence imputable. They propose to identify the covariates with missing data, to perform imputations producing  $m$  complete datasets (multiple imputations) and in the end to stack all these datasets into one single wide complete dataset. Before imputation, the generic linear mixed model in Eq. (1) is rewritten, taking into account for the missing values, as follows:

$$Y_i = \sum_{l=1}^L \sum_{g=1}^G (X_{ig}^{(l)} \beta_g^{(l)}) + Z_i^{(\bullet)} b_i + \epsilon_i, \quad i = 1, 2, \dots, m; \quad g = 1, \dots, G; \quad l = 1, \dots, L; \tag{31}$$

where  $X_{ig}^{(l)}$  represents the  $g$ -th predictor for the  $i$ -th cluster from the  $l$ -th imputed dataset. After grouping all datasets into one, according to group relevant variables for imputation, the model could be rewritten in a compact way:

$$Y_i = X_i^{(\bullet)} \beta^{(\bullet)} + Z_i^{(\bullet)} b_i + \epsilon_i, \tag{32}$$

where  $X_i^{(\bullet)} = (X_{i1}^{(\bullet)}, X_{i2}^{(\bullet)}, \dots, X_{iG}^{(\bullet)})'$  containing all the imputation data, and  $\beta^{(\bullet)}$  is the related  $G$ -vector of parameters. For identifying the relevant covariates, the authors suggest a shrinkage estimation process, i.e., to maximize the profile penalized REML log-likelihood built for the extended model to imputed datasets:

$$Q_R(\beta^{(\bullet)}) = l_R(\beta^{(\bullet)}, \sigma^2, \Psi_*) - \lambda \sum_{g=1}^G \sqrt{u_g} \|\beta_g^{(\bullet)}\|, \tag{33}$$

where  $\lambda$  is the positive tuning parameter,  $u_g$  is the number of covariates, belonging to the group  $g$ , with imputation data inside. In case of no missing data or only one imputation, the optimal penalized solution is obtained through the classical LASSO penalization. Instead of maximizing Eq. (33), because of some computational issues, the authors prefer to solve a different optimization problem through an iterative algorithm concerning the following penalized function:

$$Q_R^2(\beta^{(\bullet)}) = l_R(\beta^{(\bullet)}, \sigma^2, \Psi_*) - \sum_{g=1}^G \tau_g^2 - \lambda^2 \sum_{g=1}^G \frac{u_g}{4\tau_g^2} [\|\beta_g^{(\bullet)}\|]^2, \tag{34}$$

Hossain et al. (2018) propose a non-penalty Stein-like shrinkage estimator and then an adaptive version of the same estimator. This approach, first, consists in using a non-penalty shrinkage estimator (SE) and then it applies an adaptive measure related to the number of restrictions, which measures the distance between the restricted and the full model. The procedures works as follows: they propose to maximize the log-likelihood function under the postulated restricted parameter space, using the Lagrange multiplier vector, to get a restricted estimator for  $\beta$  this allows to build the profiling log-likelihood for estimating  $\tau$ . Once the RE for  $\theta = (\beta, \tau)$  are available, the likelihood ratio test statistic  $D_m = 2[l(\hat{\theta}|\theta) - l(\hat{\theta}_{RE}|\theta)]$  is introduced, and it allows to define the pretest estimator (PT) for  $\beta$ :

$$\hat{\beta}_{PT} = \hat{\beta} - I(D_m \leq \chi_{r,\alpha}^2)(\hat{\beta} - \hat{\beta}_{RE}). \tag{35}$$

Since that  $\hat{\beta}_{PT}$  is a discontinuous function of  $\hat{\beta}$  and  $\hat{\beta}_{RE}$  and it depends on the  $\alpha$ -level chosen a priori by the user, an adapted shrinkage estimator is built up, as follows:

$$\hat{\beta}_{PSE} = \hat{\beta}_{RE} + (1 - (r - 2)D_m^{-1})(\hat{\beta} - \hat{\beta}_{RE}), \quad r \geq 3, \tag{36}$$

The shrinkage estimator is, actually, a linear combination of  $\hat{\beta}$  and  $\hat{\beta}_{RE}$ :  $\lambda\hat{\beta} + (1 - \lambda)\hat{\beta}_{RE}$ , where the shrinkage parameter  $\lambda$  is an optimal value equal to  $(r - 2)D_m^{-1}$ . The final estimator proposed by the authors is the positive-part shrinkage estimator, which takes into account only the positive values of the estimator in Eq. (36) due to the not convex function of SE in  $\hat{\beta}$  and  $\hat{\beta}_{RE}$ .

Only two papers discuss the selection of fixed effects in a linear mixed model in the case of a high dimensional setting: Rohart et al. (2014) and Ghosh and Thoresen (2018).

In many fields, it happens that one has to manage quite large amount of covariates. Thus, if interest is focused on obtaining an optimal inference, then choosing only the relevant covariates is particularly important.

Ghosh and Thoresen (2018) contribute to linear mixed-effects model selection with a non-concave penalization for the selection of fixed effects. Their procedure works with a maximum penalized likelihood, where non-concave penalties are implemented, considering  $\Sigma = \sigma^2 I_{n_i}$ . A general objective function (with a general non-convex optimization):

$$Q_{n,\lambda}(\beta, \eta) = L_n(\beta, \eta) + \sum_{j=1}^p P_{n,\lambda}(|\beta_j|), \tag{37}$$

has to be minimized with respect to  $(\beta, \eta)$  for a general loss function,  $L(\beta, \eta)$ , which is assumed to be convex only in  $\beta$  and non-convex in  $\eta$ . We can distinguish two situations: the number of fixed effects is less than the number of observations ( $p < n$ ) and a high-dimensional setup where  $p$  is of non-polynomial (NP) order of sample size  $n$ .

Making some appropriate assumptions on the penalty, it is important to note that: as  $n$  increases,  $\max\{p''_{\lambda_n}(|\beta|)\} \rightarrow 0$  and  $\frac{p'_{\lambda_n}(\theta)}{\lambda_n} > 0$ . Moreover, the true parameter  $\beta_0$  is divided into two sub-vectors  $\beta_0 = (\beta_0^{(1)'}, \beta_0^{(2)'})'$ , where  $\beta_0^{(2)}$  is a null vector. If  $\lambda_n \rightarrow 0$  and  $\sqrt{n}\lambda_n \rightarrow \infty$ , as  $n$  increases, we can be sure that the local minimizer exists and satisfies that  $\hat{\beta}^{(2)}$  is equal to 0. Concerning the case of high dimensionality, when  $p$  is of non-polynomial (NP) order of sample size, one should take into account the SCAD penalty for obtaining an estimator that is simultaneously consistent and satisfies the oracle property (Fan and Li 2001) of variable selection optimality for any suitably chosen regularization sequence  $\lambda_n$ . Under some particular assumptions (extensively presented in Ghosh and Thoresen 2018) what happens is that a local minimizer is obtained, which satisfies, with a probability of reaching one as  $n$  increases, that  $\beta^{(2)} = 0$  and that the estimated active set of  $\hat{\beta}$  coincides with the true active set of the fixed effect parameters. The  $\hat{\beta}^{(1)}$  and  $\hat{\eta}$  estimators are normally distributed under both types of dimensional settings.

Rohart et al. (2014) focus on the selection of the fixed effects in a high dimensional linear mixed model, suggesting the addition of an  $\ell_1$ -penalization on  $\beta$  to the log-likelihood of the complete data. This penalization is useful in cases where the number of fixed effects is greater than the number of observations: It shrinks some coefficients to zero. They propose an iterative multicycle expectation conditional maximization (ECM) algorithm to solve the minimization problem of the objective function:



$$g(\boldsymbol{\theta}; \mathbf{x}) = -2L(\boldsymbol{\theta}; \mathbf{x}) + \lambda \|\boldsymbol{\beta}\|_1, \tag{38}$$

The algorithm consists of four steps and it converges when three stopping criteria, based, respectively, on  $\|\boldsymbol{\beta}^{[t+1]} - \boldsymbol{\beta}^{[t]}\|^2$ ,  $\|\mathbf{b}_k^{[t+1]} - \mathbf{b}_k^{[t]}\|^2$  and  $\|L(\boldsymbol{\theta}^{[t+1]}, \mathbf{x}) - L(\boldsymbol{\theta}^{[t]}, \mathbf{x})\|^2$ , are fulfilled. Since the estimation of  $\boldsymbol{\theta}$  is biased, a good choice would be to use the algorithm only for estimating the support of  $\boldsymbol{\beta}$  and, after that, to estimate  $\boldsymbol{\theta}$  using a classic mixed model estimation, based on the model that contains the only  $J$  relevant fixed effects:  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_j + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$ . The regularization parameter  $\lambda$  is tuned with the BIC,

$$\lambda_{\text{BIC}} = \min_{\lambda} \{ \log |\mathbf{V}_{\lambda}| + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\lambda})' \mathbf{V}_{\lambda}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\lambda}) + d_{\lambda} \log(n) \}, \tag{39}$$

where  $d_{\lambda}$  is the number of nonzero variance–covariance parameters plus the number of nonzero fixed effects coefficients. Substituting the LASSO method in the second step with any other variable selection method that optimizes a criterion, the algorithm becomes a multicycle ECM. All these considerations are valid assuming independence between the random effects, i.e., if there are  $q$  random effects corresponding to  $q$  grouping factors. As regards the selection of the random effects, it suffices to observe quite a small variance of a random effect to remove it at one step of the algorithm. The algorithm produces the same results and the same theoretical properties of the lmmLasso method (Schelldorfer et al. 2011) when variances are known or they are assumed to be known, but it is much faster.

### 5 Random effects selection

Testing if random effects exist is equivalent to testing the hypothesis whether their variance/covariance matrix is made by zeros. Some authors, like Zhang et al. (2016), worked on the identification of the covariance structure of random effects, and others such as Wang (2016) provided some characterizations of the response covariance matrix that cause model non-identifiability. The common perspective of these works lies in providing a preliminary analysis before the selection of the effects in a linear mixed model, without providing a tool for testing the significance of random effects. Li and Zhu (2013), instead, introduced a test for evaluating the existence of random effects in semi-parametric mixed models for longitudinal data, proposing a projection method. The two authors created a test with two estimates for the error variance, one consistent under the null hypothesis and the other consistent under both the null and the alternative. The idea was to compare the two estimates under the alternative hypothesis, leading to reject the null one in case of large values of the test. But the test showed to be not stable and powerful, because of the projection matrix of  $\mathbf{Z}$  variables onto the space spanned by the  $\mathbf{X}$  variables. Hence, the two authors propose a similar, but more powerful test, in the LMMs framework but without projections. For developing the test, no assumptions are necessary for the random effects or the random errors. The test is built using the trace of the variance/covariance matrix of random effects:

$$T_{m\Omega} = \frac{\text{tr}(\hat{A})}{\sqrt{(\hat{k} - 3\hat{\sigma}^4)\text{tr}\{\text{diag}^2(M_{0m}^{\text{tr}})\} + 2\hat{\sigma}^4\text{tr}\{(M_{0m}^{\text{tr}})^2\}}} \xrightarrow{d} N(0, 1), \quad m \rightarrow \infty. \quad (40)$$

Under the alternative, the same test converges in distribution to  $N(m_\Omega, 1)$ , where

$$m_\Omega = \frac{k_0 c_{11} - q_1 + (q_1 - 1)c_{13}\text{tr}(\Sigma_z Q_{10})}{\sqrt{(k - 3\sigma^4)C_{\text{diag}} + 2\sigma^4 C_{\text{tr}}}}, \quad (41)$$

with  $c_{11}$  and  $c_{13}$  estimates of variance/covariance matrices related to scaled  $\mathbf{Z}$ ,  $C_{\text{tr}}$  and  $C_{\text{diag}}$  two nonnegative constants such that  $\lim_{m \rightarrow \infty} [m \cdot \text{tr}\{\text{diag}^2(M_{0m}^{\text{tr}})\}] = C_{\text{diag}}$  and  $\lim_{m \rightarrow \infty} [m \cdot \text{tr}(M_{0m}^{\text{tr}})] = C_{\text{tr}}$ . The test results to be consistent, not only under the null hypothesis, but under the alternative too. Even if the rate of convergence is slower than  $m^{-1/2}$ , the test is consistent. Furthermore, the test is good even if high correlations between  $\mathbf{Z}$  and  $\mathbf{X}$  are present.

## 6 Fixed and random effects selection

In most real cases, it is a matter of investigating the individuation of the important predictors corresponding not only to the fixed effects but, also, to the random part of the model. The joint selection of the two types of effects has drawn more attention in recent years. Most of the proposed procedures are related to shrinkage methods: It suffices to look simultaneously at Tables 2 and 3 to check this statement. The joint effect selection through penalized function can be based on a two-stage procedure, considering fixed and random effects separately, or a one-stage procedure, considering them jointly. Bondell et al. (2010) underlined that, in a separate selection, a change in the structure of one set of effects can lead to considerable different choices of variables for the other set of effects. Lin et al. (2013), on the other hand, argued that greater computation efficiency is reached if one prefer a separate selection of the effects. The number of stages employed in the shrinkage methods is reported in Table 1.

Braun et al. (2012) propose a predictive cross-validation (CV) criterion for the selection of covariates or random effects in the presence of linear mixed-effects models with serial correlation. Their approach is based on the logarithmic and the continuous ranked probability score (CRPS). Wang and Schaalje (2009) use point predictions, while Braun et al. (2012) focus on the whole predictive distribution, inspired by the proper scoring rules suggested by Gneiting and Raftery (2007), and the “mixed” cross-validation approach provided by Marshall and Spiegelhalter (2003). Going into detail, they use a very common proper score, the LS (local score), which considers the log predictive density  $f(y)$  for the observed value  $y_{\text{obs}}$  and the CRPS, which is sensitive to the distance. The CRPS considers how close a predictive value is to the observed value through a ponderation system. With the univariate Gaussian as predictive distribution, the CRPS has the following form:

$$\text{CRPS}(Y, y_{\text{obs}}) = \sigma \left[ \frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{y_{\text{obs}} - \mu}{\sigma}\right) - \frac{y_{\text{obs}} - \mu}{\sigma} \left( 2\Phi\left(\frac{y_{\text{obs}} - \mu}{\sigma}\right) - 1 \right) \right], \quad (42)$$

**Table 1** Settings of LMM selection procedures with shrinkage

Reference	Consistency	Sparsity	Asymptotic normality	Number of $\lambda_S$	Selection of $\lambda_S$	Number of stages	Penalty
<i>Inserted in Müller et al. (2013)</i>							
BKG10 (Bondell et al. 2010)	✓	✓	✓	1	BIC	1	ALASSO
IZGG11 (Ibrahim et al. 2011)	✓	✓	✓	2	BIC	1	SCAD, ALASSO
PL12 (Peng and Lu 2012)	✓	✓	✓	2	GCV AIC BIC	2	SCAD
<i>Not inserted in Müller et al. (2013)</i>							
AZL12 (Ahn et al. 2012)	✓	✓	✓	2	BIC	2	Hard, Sandwich
CLSZ15* (Chen et al. 2015)	✓	✓	✓	1	GCV	1	Orthogonality-based SCAD
FQZ14 (Fan et al. 2014)	✓	✓	✓	1	BIC	1	ALASSO
GT16* (Ghosh and Thoresen 2018)	✓	✓	✓	1	BIC	1	SCAD
HTA18* (Hossain et al. 2018)	✓	✓	✓	1	$(r - 2)LRT-1$	1	James-Stein
LPJ13 (Lin et al. 2013)	✓	✓	✓	2	BIC	2	ALASSO
LWSWZZ18 (Li et al. 2018)	✓	✓	✓	2	BIC	1	LASSO: $\ell_1, \ell_2$
MBL17 (Marino et al. 2017)	✓	✓	✓	1	BIC	1	LASSO: $\ell_1, \ell_2$
P16 (Pan 2016)	✓	✓	✓	2	BIC <sub>R</sub> BIC <sub>F</sub> GCV <sub>R</sub> GCV <sub>F</sub> AIC <sub>R</sub> AIC <sub>F</sub>	2	ALASSO
PS18 (Pan and Shang 2018b)	✓	✓	✓	2	BIC	1	ALASSO
RSL14 (Rohart et al. 2014)	✓	✓	✓	1	BIC	1	LASSO
TVCN12 (Taylor et al. 2012)	✓	✓	✓	1	BIC <sub>R</sub>	1	$L_R r < 1$
WLXZ16 (Wu et al. 2016)	✓	✓	✓	2	GCV	2	SCAD

“Reference” refers to the initials of the authors followed by the second digit of the year of publication. The second, the third and the fourth columns contain the information about the desired properties for the—fixed and/or random—estimators: consistency and the “oracle properties” (sparsity and asymptotic normality). Fan and Li (2001). The symbol \* is added to the authors that proved the oracle properties only for the fixed effects

where  $\varphi$  and  $\Phi$  indicate the p.d.f. and the distribution function of a standardized Gaussian variable, respectively. The “mixed” cross-validation approach fits a model to the whole dataset. Once the hyperparameters have been estimated through all data, one observation is left out and for this one the LS and the CRPS are computed. Finally, the cross-validation mean scores  $\overline{\text{LS}}_{\text{CV}}$  and  $\overline{\text{CRPS}}_{\text{CV}}$  are calculated from the distribution. The  $\overline{\text{LS}}_{\text{CV}}$  is asymptotically equivalent to cAIC, but it is preferable to a full cross-validation approach because only one model is fitted at the beginning instead of fitting a model for each observation left out.

Schmidt and Smith (2016) focus on model selection when the number of models involved in the process is huge. They introduce a parameter subset selection algorithm (PSS). This technique consists in ranking the parameters by their significance, to establish the influential parameters. The basic assumption regarding the variance–covariance matrices of the random effects and of the random errors is  $\Psi$  and  $\sigma^2 I_{n_i}$ , respectively. The methodology is based on the asymptotic approximation of standard errors, measured through a normalization of the estimated standard deviations for each parameter. The proposed method works as follows: at first an estimate of the error variance is measured, then using a local sensitivity matrix—containing all the derivatives with respect to all fixed and random parameters for each  $i$ -th observation—one is able to estimate the variance–covariance matrix with all variances and correlations for the fixed and for the random effects (the authors suggest to use for instance the Moore–Penrose pseudoinverse). An estimate for the standard errors for each parameter is now possible:  $\sqrt{\text{Cov}(k, k)}$ , which is used for obtaining a measure of the selection score related to each  $k$ -th parameter in the  $i$ -th individual:  $\alpha_{k_i} = |\text{st.err.}_k / \hat{\theta}_{k_i}|$ . A small selection score is equivalent to a significant parameter. A ranking of all selection scores is created assigning a selection index  $\gamma_{k_i}$  according to the position reached by each  $\alpha_{k_i}$  in the ordering. For all the parameters is calculated a global selection index  $\Gamma_k = \sum_{i=1}^m \gamma_{k_i}$ , which implies that the smallest values of this global index are related to the most significant parameters for all the clusters. If two or more parameters bring to the same  $\Gamma_k$ , then the parameter that has the smallest selection scores over all  $m$  individuals, is chosen as the most significant one. It is worth noting that since the PSS is repeated  $m$  times, the  $m$  sets of parameter rankings will be all different because the random effects parameter estimate will be different for each individual. The PSS algorithm attributes to the standard errors the role of measuring the parameter uncertainty: the parameters which obtain the smallest selection scores are those most significant and with the smallest uncertainty.

Rocha and Singer (2018) propose exploratory methods based on fitting standard regression models to the individual response profiles or to the rows of the sample within-units covariance matrix (in the case of balanced data) as supplementary tools for selecting a linear mixed-effects model. As concerns the choice of the fixed effects they examine the profile plots and suitable hypothesis tests. Assuming homoschedastic conditional independence, the model in Eq. (1) is rewritten as:

$$y_i = \mathbf{X}_i^* \boldsymbol{\beta}_i^* + \epsilon_i, \quad (43)$$

where  $\mathbf{X}_i^*$  contains the common variable between  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  and those that are unique to both the kind of variables,  $\boldsymbol{\beta}_i^*$  contains the amount of  $p + k$  parameters related to

the fixed and the random effects. To test whether the generic  $k$ -th element of  $\beta$  is null, they propose the following statistic test:

$$t = \frac{\bar{\beta}_k^*}{n^{-1}\sqrt{\hat{\sigma}^2 \text{diag}_k[(\sum_{i=1}^m X_i^* X_i^*)^{-1}]}} \sim t_v, \tag{44}$$

where the degrees of freedom  $v = \sum_{i=1}^m n_i - m(p + q)$  and the estimated  $\hat{\sigma}^2$  is given by  $\sum_{i=1}^m \frac{n_i - (p+q)}{v} \hat{\sigma}_i^2$ , with:

$$\hat{\sigma}_i^2 = \frac{1}{n_i - (p + q)} Y_i' [I_{n_i} - X_i^* (X_i^* X_i^*)^{-1} X_i] Y_i. \tag{45}$$

The variance of  $\hat{\beta}_{ik}^*$ ,  $i = 1, 2, \dots, m$ , is expected to be equal to the  $k$ -th diagonal term of  $\sigma^2 (X_i^* X_i^*)^{-1}$  when the variance of the corresponding random coefficient,  $\hat{b}_{ik}$ , is null. Otherwise, we might expect a larger variability of the  $\hat{\beta}_{ik}^*$  around its mean. The  $k$ -th element of  $\hat{\beta}_{ik}^*$ ,  $\hat{\beta}_{ik}^*$ , follows a  $\mathcal{N}(\beta_{ik}^*; v_{ik}\sigma^2)$  distribution where  $v_{ik} = \text{diag}_k\{(X_i^* X_i^*)^{-1}\}$ . Therefore,  $\hat{\beta}_{ik}^*/\sqrt{v_{ik}} \sim \mathcal{N}(\beta_{ik}^*/\sqrt{v_{ik}}; \sigma^2)$ . Letting  $\hat{w}_{ik} = \hat{\beta}_{ik}^*/\sqrt{v_{ik}}$  and  $\bar{w}_k = \sum_{i=1}^m \hat{w}_{ik}/m$ , it follows that:

$$t(\hat{w}_k) = \sqrt{n/(n - 1)}(\hat{w}_{ik} - \bar{w}_k)/\hat{\sigma} \sim t_v. \tag{46}$$

Thus, for each  $k$  we expect around  $\alpha\%$  of the values of  $t(\hat{w}_k)$  outside the corresponding global significance level  $\alpha^*\% = \alpha/(m(p + q))$  Bonferroni-corrected confidence interval, namely  $[t_v(\alpha^*/2), t_v(1 - \alpha^*/2)]$  where  $t_v(\delta)$  denotes the  $100\delta\%$  percentile of the  $t$  distribution with  $v$  degrees of freedom. A larger percentage of points outside that interval suggests that  $b_{ik}$  may be a random coefficient. Combining the two statistic tests in Eqs. (44) and (46) makes possible to detect which effects are statistically significant in the selection procedure. Another way to select the random effects requires the assumption of the homoschedastic conditional independence, i.e., when data are collected at the same time. In this case, the number of units for each  $i$ -th individual is the same and hence it is possible to estimate only one variance-covariance matrix  $V$  as  $S - \hat{\sigma}^2 I_n$ , where  $S = (m - 1)^{-1} \sum_{i=1}^m (y_i - \bar{y})(y_i - \bar{y})'$ . Fitting polynomial models, with the same degree, to the rows of  $S$  the exploratory analysis along the lines obtained becomes an additional tool for the selection of the random effects.

### 6.1 One-stage shrinkage procedures

Chen et al. (2015) propose a variable selection methodology under the ANOVA type linear mixed models, for a high-dimensional setting. They focus on the selection of the fixed effects and on testing the existence of the random effects. The authors state that  $\text{cov}(b_i) = \sigma_i^2 I_{n_i}$  and  $\Sigma = \sigma^2 I$ , without setting any distributional assumption for  $Y$ . The selection regarding the fixed effects is made through the SCAD penalty. With the main purpose of removing the heteroschedasticity and correlation of the response

variable, they modify the model in Eq. (1), through an orthogonalization applied to random variables  $\mathbf{Z}_\perp$ . Let  $\mathcal{M}(\mathbf{Z})$  be the vector space spanned by the columns of  $\mathbf{Z}$ ,  $\mathbf{Z}_\perp$  such that  $\mathbf{Z}'_\perp \mathbf{Z} = 0$ ,  $\mathcal{M}(\mathbf{Z})^\perp$  the orthogonal complementary space of  $\mathcal{M}(\mathbf{Z})$ , therefore:

$$\mathbf{Z}_\perp \mathbf{Y} = \mathbf{Z}_\perp \mathbf{X} \boldsymbol{\beta} + \mathbf{Z}_\perp \boldsymbol{\epsilon}, \tag{47}$$

A sparse estimate of  $\boldsymbol{\beta}$  can be obtained by minimizing:

$$Q(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' P_{(\mathbf{Z})_\perp} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + n \sum_{j=1}^p p_\lambda(|\beta_j|), \tag{48}$$

where  $P_{(\mathbf{Z})_\perp} = \mathbf{Z}_\perp \mathbf{Z}'_\perp$  is the orthogonal projection matrix of space  $\mathcal{M}(\mathbf{Z})^\perp$  and  $p_\lambda(\theta)$  is the SCAD penalty. Putting  $\mathbf{Y}^* = \mathbf{Z}'_\perp \mathbf{Y}$  and  $\mathbf{X}^* = \mathbf{Z}'_\perp \mathbf{X}$  the minimization algorithm  $Q(\boldsymbol{\beta})$ , the convergence test and the selection of thresholding parameters can be applied to Eq. (48) without additional effort. Once the fixed effect parameters are estimated, the authors focus on the selection of the random effects, which means to detect if some  $\sigma_i = 0$ . The formal hypothesis system is:

$$H_0 : \sigma_k^2 = 0, k \in \mathcal{D} \leftrightarrow H_a : \exists \mathcal{D}_* \subseteq \mathcal{D}, s.t., \sigma_k^2 > 0, k \in \mathcal{D}_*, \tag{49}$$

where  $\mathcal{D}$  is a subset of  $1, 2, \dots, q$ . Two estimators are proposed for  $\sigma^2$ : one,  $\hat{\sigma}^2$ , consistent even if the null hypothesis does not hold, the other one,  $\hat{\sigma}_0^2$ , consistent only under the null hypothesis. Indicating with  $\hat{l} \hat{=} \{i : \hat{\beta}_i \neq 0\}$  all the relevant fixed effects, once the fixed parameters have been estimated, with  $W_{\hat{l}} \hat{=} (\mathbf{X}_{\hat{l}}, \mathbf{Z})$  the relative covariate matrix together with the design matrix for the random effects, an estimate of  $\sigma^2$  is defined as:

$$\hat{\sigma}^2 = \frac{\mathbf{Y}' P_{(W_{\hat{l}})_\perp} \mathbf{Y}}{tr[P_{(W_{\hat{l}})_\perp}]}, \tag{50}$$

where  $P_{(W_{\hat{l}})_\perp}$  is the orthogonal projection matrix on the space of  $\mathcal{M}(W_{\hat{l}})^\perp$ :

$$\hat{\sigma}_0^2 = \frac{\mathbf{Y}' P_{(W_{\hat{l}, -\mathcal{D}})_\perp} \mathbf{Y}}{tr[P_{(W_{\hat{l}, -\mathcal{D}})_\perp}]}, \tag{51}$$

Let's assume that  $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$  with  $\mathcal{D}_1 \hat{=} \{k : k \in \mathcal{D}, m_k \rightarrow \infty \text{ when } n \rightarrow \infty\}$  and  $\mathcal{D}_2 \hat{=} \{k : k \in \mathcal{D}, m_k = O(1)\}$ . Under  $H_0$  in (49), under certain conditions and assuming that the  $\mathcal{D}_1$  is a null set, the authors built a test for assessing the existence of at least one of the random effects based on the difference between (50) and (51), which tends in distribution to  $\chi^2(g)$  where  $g$  represents the dimension of space  $\mathcal{M}(P_{(W_{\hat{l}, -\mathcal{D}})_\perp} \mathbf{Z}_{\mathcal{D}})$ . Whereas, under  $H_0$  in (49) if  $\mathcal{D}_1$  contains at least one element and knowing that  $\hat{\sigma}^2 - \hat{\sigma}_0^2 = \mathbf{Y}' M_{n, \hat{l}} \mathbf{Y}$ , with  $M_{n, \hat{l}} \hat{=} \frac{P_{(W_{\hat{l}})_\perp}}{tr(P_{(W_{\hat{l}})_\perp})} - \frac{P_{(W_{\hat{l}, -\mathcal{D}})_\perp}}{tr(P_{(W_{\hat{l}, -\mathcal{D}})_\perp})}$ , then the test to be considered is:

$$T_{nG,\hat{i}}(\gamma) = \frac{Y' M_{n,\hat{i}} Y}{\hat{\sigma}^2 \sqrt{\gamma \text{tr}\{\text{diag}^2(M_{n,\hat{i}})\} + 2\text{tr}\{M_{n,\hat{i}}\}}} \xrightarrow{d} N(0, 1) \text{ as } n \rightarrow \infty, \tag{52}$$

where  $\gamma$  indicates the kurtosis parameter that can be estimated with any consistent estimator.

Fan et al. (2014) propose a robust estimator for jointly selecting the fixed and random effects. The variable selection methodology defined by the three authors is robust against outliers in both the response and the covariates. The variance–covariance matrix of the random effects is factorized using the Cholesky decomposition:  $\Psi = \Lambda \Gamma \Lambda'$ , where  $\Lambda = \text{diag}(v_1, v_2, \dots, v_q)$  and  $\Gamma$  represents a diagonal matrix and a triangular matrix with 1 on its diagonal, respectively. Hence, the random effects  $\mathbf{b}_i$  are now substituted by  $\Lambda \Gamma \mathbf{b}_i^*$ . It is worth noting that setting to zero one element of  $\Lambda$  implies that all elements of the corresponding row and column in  $\Psi$  are zero, too, i.e., the relative random effect is not significant. To obtain a robust estimator which doesn't suffer the impact of outliers in the covariates, they introduce some weights,  $w_{ij}$ , function of the Mahalanobis distance:

$$w_{ij} = \min \left\{ 1, \left\{ \frac{d_0}{(\mathbf{x}_{ij} - m_x)' S_x^{-1} (\mathbf{x}_{ij} - m_x)} \right\}^{\frac{\delta}{2}}, \left\{ \frac{b_0}{(\mathbf{z}_{ij} - m_z)' S_z^{-1} (\mathbf{z}_{ij} - m_z)} \right\}^{\frac{\delta}{2}} \right\}, \tag{53}$$

where the parameter  $\delta \geq 1$ ,  $d_0$  and  $b_0$  are the 95-th percentiles of the chi-square distributions with the dimension of  $x_{ij}$  and  $z_{ij}$  like degrees of freedom, respectively.  $S_x$  and  $S_z$  are the median absolute deviance and  $m_x$  and  $m_z$  represent the medians of the covariates and random variables, respectively. For reducing the impact of outliers in the response variable, it is modified subtracting  $v_{ij}$  to each its element in Eq. (54), considering the studentized residuals  $r_{ij} = y_{ij} - x'_{ij} \beta - z'_{ij} \Lambda \Gamma \mathbf{b}_i^*$

$$v_{ij} = \text{sign}(r_{ij})(|r_{ij}| - c)\sigma I(|r_{ij}| > c). \tag{54}$$

The robust log-likelihood is then defined as:

$$l^R(\boldsymbol{\theta}) = \log \int \sigma^{-\frac{mq+n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \|W^{\frac{1}{2}}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}I_m \otimes \Lambda I_m \otimes \Gamma \mathbf{b}^*)\|^2 \right\} \times \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{b}^{*'} \mathbf{b}^* \right\}. \tag{55}$$

To guarantee the consistency property to the estimators, a correction has to be applied to  $l^R(\boldsymbol{\theta})$ :

$$l^R_c(\boldsymbol{\theta}) = l^R(\boldsymbol{\theta}) - a_m(\boldsymbol{\theta}), \tag{56}$$

with  $a_m(\boldsymbol{\theta}) = \sum_{i=1}^m a_i(\boldsymbol{\theta})$  such that  $\frac{\partial}{\partial \boldsymbol{\theta}} a_i(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left[ \frac{\partial l^R_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]$ .

Selection and estimation of fixed and random effects are obtained maximizing:

$$Q^R(\boldsymbol{\theta}) = l^R_c(\boldsymbol{\theta}) - n \left( \sum_{j=1}^p p_{\lambda_n}(|\beta_j|) + \sum_{j=1}^q p_{\lambda_m}(|v_j|) \right), \tag{57}$$

where  $p_{\lambda m}(\cdot)$  is a shrinkage penalty with  $\lambda_n$  being the parameter which controls the amount of shrinkage, while  $\hat{\beta}_j$  and  $\bar{v}_j$  are the un-penalized maximum estimators in Eq. (55). The authors propose the ALASSO penalty to control the amount of shrinkage. For selecting  $\lambda_m$  the authors prefer to minimize the following BIC criterion:

$$\text{BIC}(\lambda) = -\frac{1}{2} \log |\hat{V}| - \frac{1}{2} \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_{\hat{V}}^2 + \log(m) \|\hat{\theta}_\lambda\|_0, \tag{58}$$

where  $\hat{\sigma}^2$ , part of  $\hat{V}$ , is the median absolute deviation estimate,  $\hat{\beta}$  and  $\hat{V}$  are obtained as robust estimators and, finally,  $\|\hat{\theta}_\lambda\|_0$  states for the zero norm, measuring the amount of nonzero elements of  $\hat{\theta}_\lambda$ .

Taylor et al. (2012) extend the two-parameter  $L_r$  penalty of Frank and Friedman (1993) and Fu (1998) in order to obtain new mixed model penalized likelihood, useful for selecting both the random and the fixed effects. The extended linear mixed model considers a set of penalized effects ( $\mathbf{a}$ ), containing a subset of some effects:

$$\mathbf{y}|\mathbf{b} \sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \mathbf{M}\mathbf{a}, \Sigma), \quad \mathbf{y} \sim N(\mathbf{X}\beta + \mathbf{M}\mathbf{a}, V(\tau)). \tag{59}$$

The authors use the scaled variance–covariance matrices  $\Sigma_* = \Sigma/\sigma^2$  and  $V(\tau)_* = V(\tau)/\sigma^2$  and identify  $\mathbf{a}$ , a potentially large vector of  $k$  effects,  $k < p + s$  and  $k < n$ , with covariates  $\mathbf{M}$ . The penalized likelihood involves the  $L_r$  class of penalties with  $0 < r < 1$ :

$$l = \log f(\mathbf{y}, \theta) - \sum_{j=1}^k p_\lambda(|\mathbf{a}_j|; r), \tag{60}$$

with the penalty term given by:  $p_\lambda(|\mathbf{a}_j|; r) = \lambda((|\mathbf{a}_j| + 1)^r - 1)/r$ ,  $\lambda > 0$ . Taking into account a simple setting with  $\sigma^2 = 1$  and  $\mathbf{M}$  as orthonormal columns, an unbiased OLS estimator for  $\mathbf{a}$  is obtained, through an iterative process:

$$\mathbf{a}_{j(s+1)} = \text{sign}(\hat{\mathbf{a}}_j)(|\mathbf{a}_j| - \lambda^*)_+. \tag{61}$$

This penalty is singular at origin, then, a local quadratic approximation is introduced to the derivative of the penalty, approximated as follows:

$$p_\lambda(|\mathbf{a}_j|; r) \approx \frac{1}{2}(\lambda(|\mathbf{a}_{js}| + 1)^{r-1}/|\mathbf{a}_{js}|)\mathbf{a}_j^2, \tag{62}$$

Thus, the introduction of a penalized term estimated iteratively, as shown is equivalent to inserting the pseudo-random effects in the linear mixed models. This it suffices to guarantee Henderson’s results for estimation (REML estimates for  $\tau$ ) and prediction of both kinds of effects. Thresholding the elements of  $|\mathbf{a}_{s+1}|$  with an optimal rule, a partitioned set of estimates into nonzero and zero components ( $\mathbf{a}_{1,s+1}, \mathbf{a}_{2,s+1}$ ) is obtained. The zero set ( $\mathbf{a}_{2,s+1}, \mathbf{M}_{2,s+1}$ ) is discarded from the set of information and the nonzero set replaces  $\mathbf{a}_2$  until the iterative penalized REML estimates converge.

Li et al. (2018) propose a doubly regularized approach for selecting both the fixed and the random effects, in two cases: a) finite dimension of fixed and/or random



effects, b) fixed and/or random effects that increase as the sample size goes to infinity. Their approach set  $\Sigma = \sigma^2 I_{n_i}$  and  $\Psi = \sigma^2 \Psi_* = \sigma^2 L L'$ , (Cholesky decomposition) with  $L$  a lower triangular matrix containing positive diagonal elements. The authors apply a double regularization (a  $\ell_1$ -norm penalty for  $\beta$  and a  $\ell_2$ -norm penalty for  $\Psi_*$  parameters) to the log-likelihood function,  $l(\beta, \sigma^2, \Psi_*)$  (equivalent to Eq. 5), as concerns the case with  $m < p$ . Hence, the objective function to maximize for estimating  $\beta, \sigma^2$  and  $\Psi_*$  is the following:

$$Q(\beta, L, \sigma^2) = \ell(\beta, \sigma^2, L) - \lambda_1 \sum_{j=1}^p |\beta_j| - \lambda_2 \sum_{k=2}^q \sqrt{L_{k1}^2 + \dots + L_{kq}^2}. \tag{63}$$

For the case  $m > p$ , they modify  $l(\cdot)$  in Eq. (63) with the following function:

$$\ell_m(\beta, \sigma^2, L) = -\frac{1}{2} \sum_{i=1}^m \log |\sigma^2 V_{*i}| - \frac{1}{2} \log \left| \sigma^{-2} \sum_{i=1}^m X_i' V_{*i}^{-1} X_i \right| - \frac{1}{2\sigma^2} (Y_i - X_i \beta)' V_{*i}^{-1} (Y_i - X_i \beta). \tag{64}$$

The authors propose an algorithm as effective as the Newton–Raphson algorithm for estimating step by step  $\beta$  and  $L$ , since the penalty function in Eq. (64) is separable.

Pan and Shang (2018b) propose a simultaneous selection procedure of fixed and random effects. Let’s assume that  $\Psi = \sigma^2 \Psi_*$ ,  $\Sigma = \sigma^2 I_{n_i}$  and  $\psi$  containing the  $\frac{q(q+1)}{2}$  unique elements in  $\Psi_*$ , and let’s indicate with  $\theta_*$  the vector related to  $(\beta, \psi)$ . The authors maximize the following penalized profile likelihood function:

$$\begin{aligned} Q(\theta_*) &= p(\theta_*) - \lambda_m \rho(|\theta_*|) \\ &= -\frac{1}{2} \sum_{i=1}^m \log |V_{i*}| - \frac{n}{2} \log \left( \sum_{i=1}^m (y_i - X_i \beta)^T V_{i*}^{-1} (y_i - X_i \beta) \right) - \lambda_m \rho(|\theta_*|), \end{aligned} \tag{65}$$

where  $\lambda_m$  is the tuning parameter controlling the amount of shrinkage and  $\rho(|\theta_*|)$  is the adaptive Lasso function:  $\rho(|\theta_*|) = |\theta_*|/|\tilde{\theta}_*|$ , with  $\tilde{\theta}_*$  the MLE estimator of  $\theta_*$  used as the initial weights vector. To maximize 65, the authors use the Newton–Raphson algorithm, considering a local quadratic approximation at each iteration step as concerns the approximation of  $|\theta_*|$ .

### 6.2 Two-stage shrinkage methods

One issue with the application of one stage shrinkage methods is that the combined dimension of both fixed and random effects is higher than the dimension of each of the two steps considered separately (Lin et al. 2013). The computational efficiency depends also on the penalized log-likelihood taken into account for the selection of the random effects: The REML is preferred by Lin et al. (2013) and Pan (2016). The reasoning behind this choice is intuitive and underlined by Lin et al. (2013): REML estimators are unbiased and seem to be more robust to outliers than ML estimators. Furthermore, REML estimators do not involve the fixed effects.

Lin et al. (2013) propose two-stage model selection by REML and pathwise coordinate optimization, inspired by the algorithm suggested by Friedman et al. (2007). The mixed model used is formulated assuming that  $\Sigma = \sigma^2 I_{n_i}$ . In detail, during the first stage, the random effects are selected by maximizing the restricted log-likelihood penalized with the adaptive LASSO penalization:

$$Q^R(\boldsymbol{\tau}) = l^R(\boldsymbol{\tau}) - \lambda_{1,m} \sum_{j=1}^s \lambda_j w_j |\Psi_j|, \tag{66}$$

where  $\Psi_j$  is the  $j$ -th diagonal element of  $\Psi$  and  $w_j$  is the known weight. Because of the non-differentiable nature of the objective function, the Newton–Raphson algorithm is used for maximizing  $Q^R(\boldsymbol{\tau})$ , after having locally approximated the penalty function by a quadratic function. Once the variance–covariance matrix is estimated, it is considered as known when the following penalized log-likelihood function is maximized to estimate the fixed effects:

$$Q^f(\boldsymbol{\beta}) = -\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})' \mathbf{v}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}) - \lambda \sum_{j=1}^p w_j |\beta_j|. \tag{67}$$

Wu et al. (2016) propose an orthogonalization-based approach, which selects separately the fixed effects, at first, and then the random effects. All the selection steps are based on the least squares and no specific distribution assumption has to be involved. This method is suggested when the dimension of fixed effects is not large. The mixed model used considers  $\Sigma = \sigma^2 I$  and the selection procedure applies, at first, a QR decomposition of the design matrices, related to the random effects, for obtaining a homogeneous linear regression model (which does not depend on the random effects). To select the fixed effects, it suffices to minimize, with respect to  $\boldsymbol{\beta}$ , the sum of residuals with SCAD penalization, thanks to possibility to find an unbiased estimate (Fan and Li 2001):

$$S_1(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' P_{z'} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + (n - ms) \sum_{j=1}^p p_{\lambda 1}(|\beta_j|), \tag{68}$$

where  $P_{z'} = I - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}$  is an idempotent matrix and  $p_{\lambda 1}(|\beta_j|)$  is a function whose first derivative depends on the tuning parameter  $\lambda$ . A ridge estimation process is computed for obtaining  $\hat{\boldsymbol{\beta}}$ , approximately:

$$\hat{\boldsymbol{\beta}}^{k+1} = (\mathbf{X}' P_{z'} \mathbf{X} + (n - ms) \sum (\lambda_{1,1}, \hat{\boldsymbol{\beta}}^k))^{-1} \mathbf{X}' P_{z'} \mathbf{Y}, \tag{69}$$

while to estimate  $\sigma^2$  they consider:

$$W_2^*(\Psi, \sigma^2) = \frac{1}{2} \sum_{i=1}^m ((\mathbf{y}_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) \otimes (\mathbf{y}_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) - \text{vec}(\mathbf{V}_i))' \tag{70}$$

$$\times ((\mathbf{y}_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) \otimes (\mathbf{y}_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) - \text{vec}(\mathbf{V}_i)), \tag{71}$$

where  $V_i$  stands for the variance–covariance matrix of  $Y_i$ ,  $\otimes$  for the Kronecker tensor product and  $\hat{\beta}$  for the estimates of the fixed effects obtained previously. Then, the objective function  $S_2(\theta)$  with the SCAD penalty becomes:

$$S_2(\tau) = \frac{1}{2} \sum_{i=1}^m (\tilde{Y} - u_i \tau)' (\hat{V}_i \otimes \hat{V}_i)^{-1} (\tilde{Y} - u_i \tau) + \sum_{i=1}^m n_i^2 \sum_{j=1}^{(q^2+q)/2+1} p_{\lambda 2}(|\tau_j|), \tag{72}$$

and even in this situation it is solved iteratively obtaining the ridge estimation for  $\tau$ :

$$\hat{\tau}^{k+1} = (U' \hat{W}^{-k} U + \sum_{i=1}^m n_i^2 \sum_{\lambda 2} (\hat{\tau}^k)^{-1} U' \hat{W}^{-k} \tilde{Y}), \tag{73}$$

knowing that  $W$  is a diagonal matrix whose elements are given by  $W_i = V_i \otimes V_i$ ,  $\tilde{Y}$  is the bias corrected  $Y$  and  $u_i$  is a function of  $z_i \otimes z_i$ .

Ahn et al. (2012) provide a class of robust thresholding and shrinkage procedures for selecting both the effects in linear mixed models. The robustness is guaranteed as they deal with non-normal correlated data and they do not assume any distribution of random effects and errors. For the estimation of the variance components, a moment-based loss function is built. For ensuring the desired sparse structure, they employ a hard thresholding estimator  $\hat{\Psi}^H = [\hat{\sigma}_{ij}^H]$ , defined as  $\hat{\sigma}_{ij}^H = \tilde{\sigma}_{ij} I(|\tilde{\sigma}_{ij}| > \nu)$ , where  $I(\cdot)$  is a typical indicator function and  $\nu \geq 0$  is the parameter which controls the thresholding criterion. Although  $\hat{\Psi}^H$  is consistent, it could not be a positive semi-definite matrix in the presence of small sample sizes. Hence, in this sense, a sandwich estimator with a shrinkage penalty is yielded, by minimizing the following function:

$$Q_R(D) = \sum_{i=1}^m \sum_{j=1}^{n_i-1} \sum_{k=j+1}^{n_i} (\tilde{y}_{ijk} - z'_{ij} D \tilde{\Psi} D z_{jk})^2 + \lambda \sum_{i=1}^q d_i, \quad \text{subject to all } d_i \geq 0, \forall i = 1, \dots, q.$$

To select the fixed effects, using  $V = Z \tilde{\Psi} Z' + \hat{\sigma}_\epsilon^2 I_n$ , a feasible generalized least square (FGLS) estimator for  $\beta$  is computed as the minimizer of the following objective function:

$$Q_F(\beta) = L_F(\beta | \hat{\Psi}, \hat{\sigma}_\epsilon^2) + \tau \sum_{j=1}^p w_j |\beta_j|,$$

where data are transformed and  $w_j$ 's are data-dependent weights.

Pan (2016) and Pan and Shang (2018a) propose a shrinkage method for selecting separately the two kinds of effects. The employment of the profile log-likelihood leads to a more efficient and stable computational procedure. Recalling the linear mixed model, let us assume that  $\Psi = \sigma^2 \Psi_*$ ,  $\Sigma = \sigma^2 I_{n_i}$  and  $\psi$  contains the  $\frac{q(q+1)}{2}$  unique elements in  $\Psi_*$ . The profile and the restricted profile log-likelihood functions are, respectively:

$$p(\boldsymbol{\beta}, \boldsymbol{\psi}) = -\frac{1}{2} \sum_{i=1}^m \log |V_i| - \frac{n}{2} \log \left( \sum_{i=1}^m (y_i - X_i \boldsymbol{\beta})^T V_i^{-1} (y_i - X_i \boldsymbol{\beta}) \right), \quad (74)$$

$$p_R(\boldsymbol{\psi}, \sigma) = -\frac{1}{2} \log \left| \sum_{i=1}^m X_i^T V_i^{-1} X_i \right| - \frac{1}{2} \sum_{i=1}^m \log |V_i| - \frac{1}{2} (n-p) \log \left[ \sum_{i=1}^m (y_i - X_i \tilde{\boldsymbol{\beta}})^T V_i^{-1} (y_i - X_i \tilde{\boldsymbol{\beta}}) \right], \quad (75)$$

The random covariance structure is selected by maximizing the penalized restricted profile log-likelihood with the adaptive LASSO, but a factorization of the vector containing the variance–covariance elements of  $\boldsymbol{\Psi}_*$  in  $(\mathbf{d}, \boldsymbol{\gamma})$  has to be carried out beforehand, with  $\mathbf{d}$  representing the vector of the diagonal elements and  $\boldsymbol{\gamma}$  the vector of parameters that can vary freely:

$$Q_R(\boldsymbol{\psi}) = p_R(\boldsymbol{\psi}) - \lambda_{1m} \sum_{j=1}^q w_{1j} d_j, \quad (76)$$

where  $\lambda_{1m}$  is the tuning parameter and  $w_1 = 1/|\tilde{\mathbf{d}}|$  are weights used for reaching the optimality of the solution, with  $\tilde{\mathbf{d}}$  computed as a root- $n$  consistent estimator vector of  $\mathbf{d}$ . The Newton–Raphson algorithm is first applied for maximizing the penalized restricted profile likelihood function leading to  $\tilde{\mathbf{V}}$  and, then, the same is applied for maximizing the penalized profile likelihood function:

$$Q_F(\boldsymbol{\beta}) = p_F(\boldsymbol{\beta}) - \lambda_{2m} \sum_{j=1}^p w_{2j} |\beta_j|, \quad (77)$$

where  $p_F(\boldsymbol{\beta})$  is the profile log-likelihood,  $\lambda_{2m}$  is the tuning parameter for fixed effect selection and  $w_{2j}$  are weights computed as the inverse of  $|\tilde{\beta}_j|$ , considering that  $\tilde{\boldsymbol{\beta}}$  is the MLE of  $\boldsymbol{\beta}$ . When the algorithm converges, the maximizer of the penalized profile log-likelihood is obtained. Hence, the set of suitable covariates is identified.

Fan and Li (2001) stated that “the penalty functions have to be singular at the origin to produce sparse solutions (many estimated coefficients are zero), to satisfy certain conditions to produce continuous models (for stability of model selection), and to be bounded by a constant to produce nearly unbiased estimates for large coefficients.” The estimator obtained through the penalty functions should lead to three important properties: asymptotic unbiasedness for avoiding modeling bias; sparsity, i.e., as a thresholding rule, the estimator should shrink some estimated coefficients to zero in order to reduce model complexity; continuity in data to avoid instability in model prediction. They showed, in few, that the choice of the shrinkage parameter should guarantee the well known oracle properties in the resulting estimator: The penalized likelihood estimator is root- $n$  consistent if  $\lambda_n \rightarrow 0$ , a set of estimated parameters is set to 0 and the remaining estimators converge asymptotically to a normal distribution when  $\sqrt{n}\lambda_n \rightarrow \infty$ .

Hossain et al. (2018) show that under certain regularity conditions and for fixed alternatives  $B_{H_a} = \delta \neq 0$ , as  $n$  increases, the estimators  $\hat{\beta}_{PT}$  (see in Eq. 35),  $\hat{\beta}_{PSE}$  (see in Eq. 36) and the positive-part shrinkage estimator converge in probability to  $\hat{\beta}$  and they derive the asymptotic joint normality for the unrestricted and restricted estimators, of which the three estimators are a function. Fan et al. (2014) demonstrate that their proposed robust estimator enjoy all the properties defined by Liski and Lisk (2008). Chen et al. (2015) demonstrate only the validity of the Oracle property of only sparsity and consistency, but not the asymptotical distribution. Li et al. (2018) show the “sparsistency” property which ensures the selection consistency for the true signals of both fixed and random effects; hence, they provide analytical proofs about the validity of consistency and sparsity, but nothing about the distributional form. Pan and Shang (2018b) demonstrate that their procedure fills the consistency and the sparsity properties, without mentioning anything about the asymptotical normality. Marino et al. (2017) only refer to take a look at Rubin (2004) in which is possible to assess that “a small number of imputations can lead to high-quality inference.” As concerns Rohart et al. (2014) thus no mention about asymptotic properties fulfilled by their final estimator. Pan (2016), Pan and Shang (2018a), Ahn et al. (2012) and Lin et al. (2013) demonstrate that, if  $\lambda \rightarrow 0$  and  $\sqrt{m}\lambda \rightarrow \infty$  as  $m \rightarrow \infty$ , the estimators produced by their two stage model selection are  $\sqrt{m}$  consistent and they possess the oracle properties, i.e., sparsity and asymptotic normality (asymptotically the proposed approaches can discover the subset of significant predictors). In other words, for an oracle procedure, the covariates with nonzero coefficients will be identified with probability tending to one, and the estimates of nonzero coefficients have the same asymptotic distribution as the true model (Pan 2016). All these statements are valid if an appropriate tuning parameter is chosen.

Consistent variable selection depends on the choice of the tuning parameter. The shrinkage procedures yield estimates, assuming the tuning parameters as known, but they are not. Hence, they have to be tuned among a pool of values, from the largest to the smallest quantity, identifying a path through the model space. After constructing the path and reducing parameter space, one can apply a direct approach (information criteria, cross-validation and so forth) to better identify the important variables. For this reason, shrinkage methods are, usually, employed in the case of many variables, thanks to the fact that they do not need to focus on all possible models ( $2^{p+q}$ ). The most widely used methods in the literature for tuning the parameter, which controls regularization, are cross-validation and BIC. “A more rigorous theoretical argument justifying the use of the BIC criterion for the  $\ell_1$  penalized MLE in high-dimensional linear mixed-effects models is missing: the BIC has been empirically found to perform reasonably well” (Schelldorfer et al. 2011). This seems to be generally valid for other shrinkage methods: there is not theoretical justification for employing the BIC. Fan et al. (2014) highlight their choice to select the shrinkage parameter through the BIC criterion is due to the fact that GCV leads to over-fitting models and AIC seems not to be consistent when the true model has a sparsity structure. The BIC criterion on which the authors base their selection of  $\lambda_n$  is the following:

$$\text{BIC}(\lambda) = -\frac{1}{2} \log |\hat{V}| - \frac{1}{2} \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_{\hat{V}}^2 + \log(m) \|\hat{\theta}_\lambda\|_0, \tag{78}$$

where  $\hat{V} = \text{diag}(\hat{V}_1, \hat{V}_2, \dots, \hat{V}_m)$  and the generic  $\hat{V}_i, \hat{\beta}, \hat{\Psi}_*$  are the robust estimates contained in  $\hat{\theta}_\lambda$  upon convergence of the EM algorithm. Because of the over-fitting problems using GCV, Marino et al. (2017) choose the BIC criterion for the selection of the tuning parameter:

$$\text{BIC}(\lambda) = -2l_R(\beta^{(\bullet)}, \hat{\sigma}^2, \hat{\Psi}_*) + q \times \ln(n), \tag{79}$$

where  $l_R(\beta^{(\bullet)}, \hat{\sigma}^2, \hat{\Psi}_*)$  is the REML log-likelihood function related to the model in (32).

Li et al. (2018) select the two tuning parameter minimizing a variant of BIC, proposed by Wang (2016):

$$\text{BIC} = -2p_R(\beta, L) + \left[ d_\beta + \frac{(1 + d_{\Psi_*})d_{\Psi_*}}{2} \right] \log(n), \tag{80}$$

where  $p_R(\beta, L)$  is the profile log-likelihood in Eq. (75),  $d_\beta$  and  $d_{\Psi_*}$  are given by the amount of nonzero elements in  $\beta$  and on the diagonal of  $\Psi_*$ , respectively. Pan (2016) and Pan and Shang (2018a) propose to minimize the BIC or the AIC or the generalized CV (GCV) as possible criteria for selecting the optimal tuning parameter. The above criteria, surely, have to be computed with the corresponding profile likelihood, shown in Eqs. (74) and (75), to identify the tuning parameter for the fixed part and the random part, respectively. The degrees of freedom necessary to compute all three criteria also refer to the fixed effects in one case (the number of nonzero  $\hat{\beta}$ 's) and to the random part in the other case (the amount of nonzero parts in  $\hat{\psi}$ ). Pan and Shang (2018b) select the optimal  $\lambda$  by minimizing the BIC criterion, where the degrees of freedom takes into account the number of nonzero elements in  $\theta_*$ . The tuned parameters  $(\lambda_1, \lambda_2)$  are computed, by Wu et al. (2016), with a CV or GCV technique. Taylor et al. (2012) and Ahn et al. (2012) choose a tuning parameter that minimizes the BIC criterion; Taylor et al. (2012) focus on the value of  $r$  (from a fixed grid, see Eq. (60)), which leads to the minimum BIC, after obtaining convergence for the penalized REML estimators:

$$\text{BIC} = -2l(\hat{\beta}, \hat{a}, \hat{\tau}) + \log(m)\#df, \tag{81}$$

where  $l(\cdot)$  is the un-penalized (since it involves  $a$  as fixed effects) marginal log-likelihood over the random effects  $b$  evaluated at the REML estimates of  $\tau$  and  $\#df$  represents the number of nonzero elements in  $\hat{a}$ . Ahn et al. (2012) work on a modified version of the BIC, similar to the RSS ratio, for both the fixed effects and the random effects:

$$\text{BIC}_R(v) = \frac{L_0(\Psi_v^H)}{L_0(\Psi)} + \frac{\log(n)}{n} \times df1, \tag{82}$$

$$\text{BIC}_F(\tau) = \frac{L_F(\hat{\beta}_\tau | \hat{\Psi}, \hat{\sigma}^2)}{L_F(\hat{\beta}_G | \hat{\Psi}, \hat{\sigma}^2)} + \frac{\log(n)}{n} \times df2, \tag{83}$$

where  $\hat{\beta}_G$  is the FGLS estimator and  $df1$  and  $df2$  represent the number of nonzero components on the diagonal in  $\hat{\Psi}^H$  and in  $\hat{\beta}_\tau$ . The degrees of freedom measure the

effective model dimension. Unlike Bondell et al. (2010) and Ibrahim et al. (2011), where the degrees of freedom considered are, respectively, sample size  $n$  and cluster size  $m$ , in the methods discussed above the number of parameters that can vary freely is connected to the nonzero parameters in the working model (fixed components and variance–covariance elements of the random effects). As pointed out by Müller et al. (2013), the number of nonzero estimated components related to the tuning parameter is not equivalent to the number of independent parameters, which is instead true for the linear models.

The main characteristics associated with shrinkage procedures available in the literature are summarized in Table 1.

## 7 Review of simulations

Almost all the authors have performed at least one simulation to measure and demonstrate the reliability of their own procedure. As in a meta-analysis, we have collected the simulations but, since the results are not directly comparable, the tables synthesize the main parameters characterizing the simulations. We followed the setting of Müller et al. (2013), for continuity to purposes. Considering Table 2, the smaller the values of  $\min |\beta|/\sigma$  and  $\min\{ev(\Psi/\sigma^2)\}$  the more difficult the selection of the true model for  $\beta$  and  $\tau$ . Nevertheless, it is worth noting that these values are not useful as regards the goodness of fit of the models or the real ability of the methods, once they are applied, for identifying the true values of  $\beta$  and  $\tau$ , since they refer to initial settings of simulations and not to their results. As Müller et al. (2013) underlined, one could consider these simulations as a mere meta-analysis. The results obtained are not directly comparable, because the authors use different measures to assess the performance of their method.

It is worth noting that, all simulations are applied with a moderate number of random effects (for both the full and the true model) and of variance–covariance parameters, except for that of Li et al. (2018) and Ahn et al. (2012). A large amount of fixed effects occur in the full model of Chen et al. (2015), Ghosh and Thoresen (2018) and Rohart et al. (2014).

To determine the set of candidate models for  $\beta$ ,  $|M_\beta|$ , the authors do not follow the same criterion. Some authors focus only on covariates, and in this sense  $|M_\beta|$  is equal to  $2^{p-1}$ . (So the intercept is not included for size of  $\beta$ ). Others instead refer to  $p$  as the whole fixed regression parameters, including the intercept, and thus, the candidate models are  $2^p$ . Furthermore some authors, such as Kawakubo et al. (2014), state that they exclude from  $|M_\beta|$  the null model (i.e., the model containing only the intercept).

Kawakubo and Kubokawa (2014) found that both the McAIC and a model averaging procedure (which has more appropriate weights) depending on McAIC, work better than cAIC in terms of prediction errors. They prove empirically the same results in the case of small area prediction, which is the topic on which Kawakubo et al. (2014) and Lombardía et al. (2017) focus on. They show, therefore, a prediction error improvement of CScAIC with respect to cAIC. Compared to mAIC, cAIC and BIC, the EBIC of Kubokawa and Srivastava (2010) is the criterion which, by simulation, leads to a better selection of the true model as the number of covariates and the number of clusters

**Table 2** Summary of settings used for the simulations

Reference	Model	$m/n_i$	$p/p_f$	$q/q_f$	$s/s_f$
AZL12 (Ahn et al. 2012)	int + slope	(100, 200)/(5, 10)	3/9	(2, 4)/(5, 10)	(4, 7)/(16, 56)
CLSZ15 (Chen et al. 2015)	int	30/5, 40/6, 60/8	3/(30, 60, 100)	1/1	3/3
FQZ14 (Fan et al. 2014)	int + slope	50/5	3/8	3/4	7/11
GT16 (Ghosh and Thoresen 2018)	int + slope	25/6	5/(10, 50, 300, 500)	2/2	2/2
HTA18 (Hossain et al. 2018)	int + slope	(40, 75, 150)/5	5/(7, 10, 14, 19)	2/2	2/2
KK14 (Kwakubo et al. 2014)	int	10/50	5/7	1/1	1/1
KO18 (Kuran and Özkale 2019)	int	(10, 20, 30)/(10, 20)	(2, 3, 4)/5	1/1	2/2
KS10 (Kubokawa and Srivastava 2010)	int	(6, 15)/4	(2, 4, 6)/7	1/1	2/2
KSK14 (Kwakubo et al. 2014)	int	30/3	3/5	1/1	2/2
LLVR17 (Lombardía et al. 2017)	int	53/41	(1, 2, 3)/14	1/1	2/2
LPJ13 (Lin et al. 2013)	int + slope	(30, 60/5, 10)	2/9	3/4	7/11
LWSWZZ18 (Li et al. 2018)	int + slope	200/8, 100/5	5/101-601	4/51	$11/\frac{51 \cdot 50}{2}$
LYCZ14 (Li et al. 2014)	int + slope	(50, 80)/4	(4, 7, 10, 13)/13	3/3	7/7
MBL17 (Marino et al. 2017)	int	40/5, (60, 150)/25	3/8	1/1	2/2
LZ13 (Li and Zhu 2013)	int + slope	(40, 70, 100)/Poisson(3) + 2	2/2	2/2	3/3
P16 (Pan 2016)	int + slope	(50, 100, 200)/5	3/5	2/5	4/15
P16 (Pan 2016)	int + slope	(30, 60)/(5, 10)	2/9	3/4	7/11
P16 (Pan 2016)	int + slope	(30, 60, 90)/12	3/8	3/5	7/16
PS18 (Pan and Shang 2018b)	int + slope	(30/5), (60/10), (10, 20/10)	2/9, 3/5	3/4, 2/4	7/11, 4/11
RSCL14 (Rohart et al. 2014)	int + slope	(20, 15)/(6, 8)	5/(80, 300, 600)	(2, 3)/(2, 3)	4/11
SS16 (Schmidt and Smith 2016)	int + slope	30/5	2/9	3/4	7/11
TVCN12 (Taylor et al. 2012)	int + slope	(10, 20)/10	1/(20, 40, 80)	1/(20, 40, 80)	(2, 3)/(2, 3)



Table 2 continued

Reference	Model	$m/n_i$	$p/p_f$	$q/q_f$	$s/s_f$
WLVZ16 (Wu et al. 2016)	int + slope	(30, 60)/(5, 10)	2/9	3/4	7/11
WLVZ16 (Wu et al. 2016)	int + slope	(10, 20)/10	3/5	2/4	4/11
WLVZ16 (Wu et al. 2016)	int + slope	40/10	2/4	2/4	4/11
WS16 (Wenren and Shang 2016)	int	(5, 20, 50)/5	(2, 3, 4)/5	1/1	2/2, 0/0
WSP16 (Wenren et al. 2016)	int	5/(5, 10, 20)/5	(2, 3, 4)/5	1/1	1/1
Reference	$ M_\beta / M_\tau $	$\min  \beta_k /\sigma$	$\min\{\text{ev}(\Psi/\sigma^2)\}$	$\epsilon/b$	Method
AZL12 (Ahn et al. 2012)	511/1024	(0.8, 0.62)	(0.5, 0.3)	$N(N, t_5, \text{Exp})$	Shrinkage
CLSZ15 (Chen et al. 2015)	$2^{30}/60/100/1$	2	(0.5, 0.3)	$\sqrt{0.75}t_8/N$	Shrinkage
FQZ14 (Fan et al. 2014)	256/	1.5	0.5	$N/N$	Shrinkage
GT16 (Ghosh and Thoressen 2018)	/1	2	0.56	$N$	Shrinkage
HTA18 (Hossain et al. 2018)	/1	1.84	1.26	$N/N$	shrinkage
KK14 (Kawakubo et al. 2014)	7/1	0.35	1	$N$	McAIC
KO18 (Kuran and Özkale 2019)	15/1	2	1	$N/N$	ridge $CC_p$
KS10 (Kubokawa and Srivastava 2010)	7/1	2	0.1	$N/N$	EBIC
KSK14 (Kawakubo et al. 2014)	31/1	1	1	$N$	CScAIC
LLYR17 (Lombardía et al. 2017)	3/1	1	0.45	$N/N$	xGAIC
LPJ13 (Lin et al. 2013)	512/16	1	0.45	$N$	Shrinkage
LWSWZZ18 (Li et al. 2018)	$2^{600}/100/2^{50}$	1	0.10	$N/N$	Shrinkage
LYCZ14 (Li et al. 2014)	13/1	0.2	0.10	$N$	MDL
MBL17 (Marino et al. 2017)	256/1	1.5	0.10	$N/N$	Shrinkage
LZ13 (Li and Zhu 2013)				$t_6/\sqrt{(15)/\text{bivariate } N}$	Difference-based test

Table 2 continued

Reference	$ M_{\beta} / M_{\tau} $	$\min  \beta_k /\sigma$	$\min\{\text{ev}(\Psi/\sigma^2)\}$	$\epsilon/b$	Method
P16 (Pan 2016)	31/31	1	0.5	t, N/N	Shrinkage
P16 (Pan 2016)	511/15	1	0.45	N/N	Shrinkage
P16 (Pan 2016)	255/31	(1.5, 1.06)	0.34	N/N	Shrinkage
PS18 (Pan and Shang 2018b)					Shrinkage
RSCL14 (Rohart et al. 2014)	8/16	1	0.43	N	Shrinkage
SS16 (Schmidt and Smith 2016)	512/2	1	0.44	N/N	Subset selection
TVCN12 (Taylor et al. 2012)				N	Shrinkage
WLXZ16 (Wu et al. 2016)	512/16	1	0.45	N	Shrinkage
WLXZ16 (Wu et al. 2016)	16/16	0.5	0.43	N, t3, t copula	Shrinkage
WLXZ16 (Wu et al. 2016)	16/8	1	0.43	N	Shrinkage
WS16 (Wenren and Shang 2016)	15/1	2	1	N	$CC_p$
WSP16 (Wenren et al. 2016)	16/1	2	1	N	$MC_p$

We follow Müller et al. (2013) where “Reference” refers to the initials of the authors followed by the second digit of the year of publication; “Model” describes the model considered,  $m$  and  $n_i$  are the number of clusters and the number of units per cluster,  $p$  and  $p_f$  the number of fixed parameters in the true model and in the full one,  $q$  the number of random effects per cluster and  $s$  the dimension of  $\tau$ , with  $q_f$  and  $s_f$  the analogous quantities under the full model. The next three measures describe the difficulty involved in selecting the true model:  $|M_{\beta}|/|M_{\tau}|$  are the number of candidate models considered for  $\beta$  and  $\tau$ , respectively,  $\min |\beta_k|/\sigma$  measures the difficulty involved in selecting the smallest nonzero regression parameter when there are no random effects in the model and  $\min\{\text{ev}(\Psi/\sigma^2)\}$ , the smallest eigenvalue of  $\Psi/\sigma^2$ , measures the difficulty of selecting the smallest nonzero variance parameter. Finally,  $b$  and  $\epsilon$  describe the distributions used for these random variables and “Method” denotes the main model selection methods considered in the simulation

increase. These results constitute empirical evidence of the consistency property of the EBIC. Lombardía et al. (2017), instead, compared the extended generalized AIC they defined (20) with the conditional AIC defined by Vaida and Blanchard (2005). They discovered that the xGAIC for the Fay–Herriot model presents better performances in terms of correct classification rates of the true model. As the number of covariates increases, the xGAIC performs better and better (in a scenario with three variables it perfectly brings to the correct model), instead the vAIC selects 44% of the times a model with a fewer number of fixed effects. Wenren and Shang (2016) show that the proposed conditional criteria perform more efficiently than the classic Mallows's  $C_p$  when more significant fixed effects are added. A large number of units for each cluster is required, if one works with the random effects within clusters (for instance small area estimation) or if one could obtain a less biased estimation of the penalty term. Wenren et al. (2016) show by simulation that their two marginal  $C_p$ -types perform better, in selecting the correct model, than mAIC and mBIC in particular situations: when observations are few and highly correlated or when the true model is included in all candidate models and includes more significant fixed effect variables. Kuran and Özkale (2019) compare the performance of their conditional ridge  $C_p$  with the  $CC_p$  of Wenren and Shang (2016), in both cases of known and unknown variance–covariance matrices of the random effects and of the random errors. Furthermore, they use different values for the ridge parameters and compare various models (with different number of the explicative variables). They show that the percentages of choosing the true model by all the  $C_p$  statistics are quite optimal and comparable and they increase as the number of fixed effects increases as well. When the ridge parameter increases, the number of individuals and the number of units are quite small and the correlation between explanatory variables is not high, the  $CRC_p$  outperforms the  $CC_p$ .

Focusing on the shrinkage selection procedures, Hossain et al. (2018) compare the performances, in terms of mean squared prediction errors, reached by their PT and PSE estimators against the unrestricted MLE, the restricted MLE, the LASSO and ALASSO methods. They show that their methodology, as the sample size increases and the number of active covariates decreases, brings to better performance than the other estimators except the restricted MLE. Ghosh and Thoresen (2018) try to demonstrate the great performances of the SCAD penalty over  $\ell_1$  penalization. Hence, by simulations, they point out that both in a low-dimensional setting and in a high-dimensional setting the two penalties correctly select the true fixed effects. With respect to  $\ell_1$ , SCAD focuses on a smaller activate set of  $\beta$ , especially, in the high-dimensional case. Marino et al. (2017) compare their penalized likelihood procedure for multilevel models with missing models with the LASSO method applied on data without missing values and, hence, used as benchmark reference. Therefore they also compare the performance of their method with the regularized LASSO on complete-case data. When missing data are present in the dataset the proposed methodology performs better, especially when the number of imputations increases. Taking into account only one imputation doesn't produce huge benefits. On the other hand, the methodology is quite good in identifying the correct model when the number of imputation and the number of units increases. Rohart et al. (2014) reached the same results as Schelldorfer et al. (2011) in the case of known variances, but with an algorithm much faster. It is worth noting that their method can be computationally combined with other procedures. The

orthogonal-based SCAD procedure of Wu et al. (2016) is very efficient in selecting the fixed effects as the number of total units increases, but has to be improved for the selection of the random effects. Pan (2016) compared the ability of his two-stage procedure to correctly identify the two kinds of effects with that of Ahn et al. (2012) and Bondell et al. (2010). He found that the percentage of the effects (taken both separately and together) correctly identified was higher than the others and was rose as the number of clusters increased. Only in the case of a non-normal distribution assumed for  $\epsilon$  did the method proposed by Ahn et al. (2012) perform better, since it does not need any distributional assumptions. Pan (2016) also compares the computational efficiency of his model selection with that of Bondell et al. (2010) and concludes that his algorithm takes less time to converge. There are two probable reasons:  $\sigma^2$  is not included in the profile log-likelihood used by Pan (2016) and a two-stage procedure for selecting both the effects is faster than the procedures involving only one step. Lin et al. (2013) used the same settings for their simulations as those used by Bondell et al. (2010), that is the reason why their results are missing in Table 2: They are available in Table 2 of Müller et al. (2013). The robust selection method presented by Fan et al. (2014) has been shown to lead to the same results of the equivalent non-robust method if the data do not present outliers. On the other hand, the method has no influence on the estimates if outliers are present in the data (both in the response variable and in the covariates), while the non-robust methodology brings to over-fitting with lower fit percentages and higher mean squared errors of the estimated parameters as a consequence. The robust selection method is perturbed by outliers if these are only in the response variable or in the covariates.

In the case of high-dimensional settings where the focus is on selection the fixed and the random effects, Li et al. (2018) used in their simulations two ways of controlling the tuning parameters: a non-adaptive regularization (NAR), which chooses the tuning parameter from a simple grid of values, and an adaptive regularization (AR), which attributes weights to different penalty parameters. The AR methodology leads to smaller estimation bias for the variance components and to a better control of the false discovery rate. Chen et al. (2015) obtained a good performance selection in terms of low proportion of parameters that did not shrink to zero while one expected the opposite or of parameters shrinking to zero, by mistake. Furthermore, they obtained accurate results in terms of bias and standard deviations of the estimates. They conducted some simulations excluding from the selection the fixed effects, and they discovered that in all situations the fixed effect selection never affects the power performances.

The parameter subset selection method proposed by Schmidt and Smith (2016) leads to better performances, compared to other techniques, among which LASSO, ALASSO and M-ALASSO.

As specified at the beginning of this review, our purpose is to give a clear outline of most methodologies used in linear mixed models that are available in the literature. Hence, in this sense, Table 3 summarizes all the features that easily identify all procedures: the part of the model focusing on (fixed and/or random), the dimension of the linear mixed model used and the structure of variance and covariance matrices. Dimensionality represents the level of the number of parameters ( $\theta = \beta, \tau$ ) involved in the model. We included not only the methods mentioned by this article,

**Table 3** Settings of LMM selection procedures for all the procedures analyzed in the review

Reference	Focus	Dimensionality	$\Psi$	$\Sigma$	Software
<i>Inserted in Müller et al. (2013)</i>					
BKG10 (Bondell et al. 2010)	Fixed + random	Low/medium	$\sigma^2\Psi_*$	$\sigma^2I_{n_i}$	R
CD03 (Chen and Dunson 2003)	Random	Low	$\Psi$	$\sigma^2I_{n_i}$	
DMT11 (Dimova et al. 2011)	Fixed + random	Low	$\sigma^2\Psi_*$	$\sigma^2I_{n_i}$	
GK10 (Greven and Kneib 2010)	Random	Low	$\sigma^2\Psi_*$	$\sigma^2I_{n_i}$	“cAIC4” R package
IZGG11 (Ibrahim et al. 2011)	Fixed + random	Low/medium	$\Psi$	$\sigma^2I_{n_i}$	R
JNR09 (Jiang et al. 2009)	Fixed	Low	$\sigma^2$	$\sigma^2$	“fence” R package
JR03 (Jiang and Rao 2003)	Fixed + random	Low	$\Psi$	$\Sigma$	
JRGN08 (Jiang et al. 2008)	Fixed	Medium	$\Psi$	$\Sigma$	“fence” R package
K11 (Kubokawa 2011)	Fixed + random	Low	$\Psi$	$\Sigma$	
NJ12 (Nguyen and Jiang 2012)	Fixed	High	$\sigma_b^2I$	$\sigma_\epsilon^2I$	“fence” R package
PL12 (Peng and Lu 2012)	Fixed + random	Low/medium	$\Psi$	$\sigma^2I_{n_i}$	Matlab
PN06 (Pu and Niu 2006)	Fixed + random	Low	$\Psi$	$\Sigma$	
SC08 (Shang and Cavanaugh 2008)	Fixed + random	Low	$\Psi$	$\sigma^2\Sigma_*$	
SK10 (Srivastava and Kubokawa 2010)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2I_{n_i}$	
<i>Not inserted in Müller et al. (2013)</i>					
AZL12 (Ahn et al. 2012)	Fixed + random	Low/medium	$\Psi$	$\sigma^2I_{n_i}$	
CLS15 (Chen et al. 2015)	Fixed + random	High	$\sigma_i^2I_{n_i}$	$\sigma^2I_{n_i}$	

Table 3 continued

Reference	Focus	Dimensionality	$\Psi$	$\Sigma$	Software
FQZ14 (Fan et al. 2014)	Fixed + random	Low	$\sigma^2 I_{n_i}$	$\sigma^2 I_{n_i}$	
GT16 (Ghosh and Thoresen 2018)	Fixed	Low/high	$\Psi$	$\sigma^2 I_{n_i}$	R
H13 (Han 2013)	Fixed	Low/medium	$\sigma_b^2 I_{n_i}$	$\sigma_b^2 I_{n_i}$	R
HTA18 (Hossain et al. 2018)	Fixed	Low/medium	$\Psi$	$\Sigma$	
KK14 (Kwakubo and Kubokawa 2014)	Fixed	Low	$\sigma^2 \Psi_*$	$\sigma^2 \Sigma_*$	
KO18 (Kuran and Özkale 2019)	Fixed	Low/medium	$\sigma^2 \Psi_*$	$\sigma^2 I_{n_i}$	R
KS10 (Kubokawa and Srivastava 2010)	Fixed	Low	$\sigma^2 \Psi_*$	$\sigma^2 \Sigma_*$	
KSK14 (Kwakubo et al. 2014)	Fixed	Low	$\sigma^2 \Psi_*$	$\sigma^2 \Sigma_*$	
LLVR17 (Lombardia et al. 2017)	Fixed	Low/medium	$\Psi$	$\Sigma$	R
LPJ13 (Lin et al. 2013)	Fixed + random	Medium	$\Psi$	$\sigma^2 I_{n_i}$	R
LS15 (Lahiri and Suntuorchost 2015)	Fixed	Low/medium	$\sigma_{b_i}^2$	$\sigma^2 I_{n_i}$	
LWSWZZ18 (Li et al. 2018)	Fixed + random	High	$\sigma^2 \Psi_*$	$\sigma^2 I_{n_i}$	
LYCZ14 (Li et al. 2014)	Fixed	Low	$\Psi$	$\sigma^2 I_{n_i}$	
LZ13 (Li and Zhu 2013)	Random	Low/(medium)	$\Psi$	$\sigma^2 I_{n_i}$	
MBL17 (Marino et al. 2017)	Fixed	Low (medium)	$\sigma^2 \Psi_*$	$\sigma_i^2 I_{n_i}$	R
P16 (Pan 2016)	Fixed + random	Low/medium/high	$\sigma^2 \Psi_*$	$\sigma^2 I_{n_i}$	
PS18 (Pan and Shang 2018b)	Fixed + random	Low/medium	$\sigma^2 \Psi_*$	$\sigma^2 I_{n_i}$	R
RSCL14 (Rohart et al. 2014)	Fixed(+ random)	High	$\Psi$	$\sigma^2 I_{n_i}$	"MMS" R package
SS16 (Schmidt and Smith 2016)	Fixed + random	Low/(medium)	$\Psi$	$\sigma^2 I_{n_i}$	Matlab

Table 3 continued

Reference	Focus	Dimensionality	$\Psi$	$\Sigma$	Software
TVCN12 (Taylor et al. 2012)	Fixed + random	Medium/High	$\sigma^2\Psi_*$	$\sigma^2\Sigma_*$	ASReml-R
WLXZ16 (Wu et al. 2016)	Fixed + random	Low/medium	$\Psi$	$\sigma^2I_{n_i}$	R and Matlab
WS16 (Wenren and Shang 2016)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2I_{n_i}$	R
WSP16 (Wenren et al. 2016)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2I_{n_i}$	R

“Reference” refers to the initials of the authors followed by the second digit of the year of publication (we use the same approach as (Müller et al. 2013)); “Focus” indicates the part of the model that is subject to selection (Fixed, Random or both); “Dimensionality” is inherent to the number of parameters involved in the initial model;  $\Psi$  and  $\Sigma$  describe the structure assumed for the variance-covariance matrices related to the random effects and the random component, respectively; “Software” specifies the software (when specified) used for implementation of the procedure

but also those contained in Müller et al. (2013), in order to provide a global view of all methodologies. Taking a look jointly to Table 2 of Müller et al. (2013), Tables 2 and 3, it becomes obvious that most model selection procedures, focusing on selecting both the fixed and the random part in cases of medium and/or high dimensionality, involve a shrinkage procedure. The shrinkage methods are computationally more efficient and statistically accurate (Bühlmann and van de Geer 2011; Müller et al. 2013).

## 8 Review of real examples

LMM are widely used in medical statistics and biostatistics. To enrich this review, we give a brief look at the real examples described in some of the listed papers.

Ahn et al. (2012), Pan (2016) and Hossain et al. (2018) describe the Amsterdam Growth and Health Study, widely used in literature. The Amsterdam Growth and Health Study Data were collected to explore the relationship between lifestyle and health in adolescence and young adulthood. In growing toward independence, the lifestyle habits of teenagers change substantially with respect to physical activity, food intake, tobacco smoking, etc. Accordingly, their health perspective may also change. Individual changes in growth and development can be studied by observing and measuring the same participant over a long period of time. The Amsterdam growth and health longitudinal study was designed to monitor the growth and health of teenagers and to develop future effective interventions for adolescence. A total of 147 subjects in the Netherlands participated in the study, and they were measured over 6 time points; thus, the total number of observations is 882. The continuous response variable of interest was the total serum cholesterol expressed in mmol/l. Pan (2016) in his paper analyses a second dataset, which is the colon cancer data. The goal of the analysis was to estimate the cost attributable to colon cancer after initial diagnosis by cancer stage, comorbidity, treatment regimen, and other patient characteristics. The data reported aggregate Medicare spending on a cohort of 10,109 colon cancer patients up to 5 years after initial hospitalization, and these data are considered as the response for a linear mixed model.

Taylor et al. (2012) applied their method to determine quantitative trait loci (QTL) in a wheat quality data set. The data set was obtained from a two-phase experiment conducted in 2006 involving a wheat population consisting of 180 double haploid (DH) lines from the crossing of two favored varieties. Data were collected from two phases of experimentation consisting of an initial field trial and milling laboratory experiment. A partially replicated design approach was used at both experimental phases. The field trial was designed as a randomized block design. The analysis considers a very large set of candidate variables, and matrix  $\mathbf{a}$  in Eq. (59) is a  $(390 \times 1)$  size matrix.

Jiang et al. (2008) considered a dataset from a survey conducted in Guatemala regarding the use of modern prenatal care for pregnancies where some form of care was used. They consider applying the fence method in selection of the fixed covariates in the variance component logistic model. Again, they cope with a quite large number of covariates.

Marino et al. (2017) worked on a dataset provided by the Healthy Directions–Small Business study conducted by Sorensen et al. (2005). Some recent epidemiological stud-



ies proved that there is a relationship between dietary patterns and physical inactivity with multiple cancers and chronic diseases. One of the main purposes of the study was to detect whether or not the cancer prevention (based on occupational health and health promotion) could lead to reduce significantly the red meat consumption or to improve significantly the mean consumption of fruits and vegetables, the levels of physical activity, the smoking cessation and the reduction of occupational carcinogens. The HD-SB study was a randomized, controlled trial study conducted between 1999 and 2003 as part of the Harvard Center Prevention Program Project. The study population of the study were twenty-six small manufacturing worksites that employed multi-ethnic, low-wage workers. Participating worksites were randomized to either the 18-month intervention group or minimal intervention control group. The respondents to the study were 974 but only 793 of them answered with complete information; hence, there was 18.5% of missing data. The number of variables involved in the survey was huge, and they were grouped according different areas: health behaviors, red meat consumption, physical activity and consumption of multivitamin and sociodemographic characteristics. The authors took into account 15 covariates, and they built a linear mixed model where the mean consumption of fruit and vegetables at follow-up. They proposed their methodology for missing data with 1, 3, and 5 imputations, comparing the results to the analysis made on the complete-cases data.

Fan et al. (2014) applied their robust method on a longitudinal progesterone dataset, available on Diggle P.J.'s homepage: <https://www.lancs>. The dataset contains 492 urine samples from 34 women in a menstrual, where each woman contributed from 11–28 times. The menstrual cycle length was standardized for all women to a reference 28-day cycle. A linear mixed model was analyzed by the authors with the log-transformed progesterone level as response variable, a random intercept and 7 fixed effects: age, bmi, time, the squared values of time and the three first-level interactions among age, bmi and time.

Li et al. (2018) in their paper analyze two datasets. The first is related to a longitudinal randomized controlled trial, involving 423 adolescent children from an Hispanic population in New York City had their parents affected by HIV+. The main purpose was to investigate about a negative state of mind (measured by a Basic Symptoms Inventory, a score well described by Weiss 2005), over six years (each person has been visited about 11.5 times). Six variables were involved in the original dataset, i.e., treatment (or control group), age, gender, Hispanic (1 = Yes, 0 = No), visit time (expressed in logarithm of year) and visit season. The authors, worked on a linear mixed model containing the six covariates plus the two-way interactions between treatment and time, gender and Hispanic, counting so 10 predictors, which were included in all the two types of effects. Their regularization procedure was applied both with the non-adaptive version and with the adaptive version (through the inverse of the estimated from the ridge-penalization procedure). Their second dataset is related to a clinical study that investigated on a possible relationship of some protein signatures with post-transplant renal functions for people with a kidney transplant. The study involved 95 renal transplant patients. The main purpose of the study was to analyze which proteins had a significant influence on the longitudinal trajectory of renal function measured by glomerular filtration rate (GFR) of the patients.

Lombardía et al. (2017) analyzed a dataset about surveys conducted from the behavioral risk factors information system in Galicia (2010–2011). The sample design applied in the survey was a stratified random sampling, allocating with equal proportions by sex and age group. Forty-one areas from the 53 counties in Galicia were involved in the survey. The authors tried to estimate the prevalence of smokers (at least 16 years old) distinguished by sex. The minimum sample size in the domain was 44 for men and 48 for women. The response variable, employed in the Fay–Herriot model used, was the logarithmic transformation of smokers' numbers. The covariates were globally 14, classified in four groups: age, degree of urbanization, activity and educational level.

Han (2013) analyzed a public health dataset about obesity released by the U.S. Centers for Disease Control and Prevention, which realized a large health study (6971 people) in the United States (51 counties of California) in the years between 2006 and 2010. The information obtained by the surveys. The purpose of the author was to estimate county level obesity rates for the female Hispanic population within working ages of 18–64.

Bondell et al. (2010) consider a recent study of the association between total nitrate concentration in the atmosphere ( $\text{TNO}_3$ ,  $\mu\text{g}/\text{m}^3$ ) and a set of measured predictors. Nitrate is one of the major components of fine particulate matter ( $\text{PM}_{2.5}$ ) across the USA. However, it is one of the most difficult components to simulate accurately using numerical air quality models. Identifying the empirical relationships that exist between nitrate concentrations and a set of observed variables that can act as surrogates for the different nitrate formation and loss pathways can help the research and can allow for more accurate simulation of air quality. To formulate these relationships, data obtained from the U.S. EPA Clean Air Status and Trends Network (CASTNet) sites are used. The CASTNet dataset consists of multiple sites with repeated measurements of pollution and meteorological variables on each site, i.e., the mean ambient particulate ammonium concentration ( $\text{NH}_4$ ,  $\mu\text{g}/\text{m}^3$ ), the mean ambient particulate sulfate concentration ( $\text{SO}_4$ ,  $\mu\text{g}/\text{m}^3$ ), relative humidity (RH, %), ozone ( $\text{O}_3$ , ppb), precipitation (P, mm/h), solar radiation (SR,  $\text{W}/\text{m}^2$ ), temperature (T, °C), temperature difference between 9 m and 2 m probes (TD, °) and scalar wind speed (WS, m/s). The same data were used by Li et al. (2014) to apply their proposed MDL procedure. A subset of the CASTnet dataset was, instead, implied by Chen et al. (2015), who focused only on five sites across the eastern USA, (2001–2009) and took as original variables  $\text{TNO}_3$ ,  $\text{NH}_4$  and  $\text{SO}_4$ , instead the others variables were transformed from ours to seasonal, substituting the maximum value for  $\text{O}_3$  and the mean value for the others. The total number of observations were 175, and in the two-way random effect model the variable time and sites were included as main random effect.

Ghosh and Thoresen (2018) investigated the effects of intake of oxidized and non-oxidized fish oil on inflammatory markers in a randomized study of 52 subjects (dataset already studied in literature). Inflammatory markers were measured at baseline and after three and seven weeks. They use the data to investigate whether there are any associations between gene expressions measured at baseline and level of the inflammatory marker ICAM-1 throughout the study. From a vast set of genes, they initially selected  $p = 506$  genes having absolute correlation greater than or equal to 0.2 with the response at any time point, so that the total number of fixed effects considered

becomes  $p = 512$ . On the other hand, removing the missing observations in the response variable they obtain  $n = 150$  observations, making it a high-dimensional selection problem. Further, due to the longitudinal structure of the data, they additionally considered random effect components in the model: they included random intercept and a random slope.

Finally, Rohart et al. (2014) apply their approach 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 metabolic data. 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 three breeds, eight environments and 157 families, metabolic data contained  $p = 375$  variables, and the phenotype investigated was the daily feed intake (DFI).

Li and Zhu (2013) applied their new covariance-based test on a famous pig weight dataset, containing the weights of 48 pigs, measured in nine successive weeks.

## 9 Discussion and conclusion

In this paper, we have discussed most of the model selection procedures for linear mixed models available to date. The purpose of our review is to allow users to easily identify the type of method they need, according to certain characteristics, such as the number of clusters and/or the number of units per cluster, the part of the model to be selected (fixed and/or random), the dimension of the model and the structure of the variance–covariance matrices. For all the methods, a description of the simulations, if available, is reported in Table 2: the purpose is to give an idea of the model settings and not to provide evidence of the best methods. We used more or less the same notation as Müller et al. (2013) for alignment with the previous review and, hence, facilitating the comparisons of the various methods over time. But this review is not only an update of Müller's review (Müller et al. 2013, but an attempt to cluster the procedures from a different point of view: the part of the model to be selected, fixed and/or random. As a matter of fact, this is one of the main issues when looking for an appropriate method to choose. Moreover, particular attention is given to the SW used, together with the implementation and the availability of the code.

This review mentions the available theoretical properties corresponding to the different methodologies, with comparisons among them whereas it's possible. A relevant importance is given here to the shrinkage methods (focused on the selection of fixed and/or random effects), since these procedures need for the oracle properties established by Fan and Li (2001).

By simulation the authors considered in this review try to achieve the best result, i.e., to identify the optimal model among a pool of candidate models and not the true model. Many issues are related to the choice of the optimal model, one of which is determined by the dimension of the pool of candidate models ( $2^{p+s}$ ). The larger this set  $\mathcal{M}$ , the lower computational efficiency is. This has been proven by Fence methods and a number of Bayesian methods reported in Müller et al. (2013) as well as the two-

stage procedures of Sect. 6.2, which select the two effects separately, thus reducing the overall dimension of models.

Over time, greater attention has been given to the generalization of  $\Sigma$  in Eq. (2): the scaled version  $\sigma^2 \Sigma_*$  replaced  $\sigma^2 I_{n_i}$ , but except for Shang and Cavanaugh (2008) the scaled version  $\sigma^2 \Psi_*$  is assumed for  $\Psi$ . There is still poor theoretical support for a generalized scenario of the variance–covariance matrices for both the effects.

Most of the methods were implemented in R, using different packages or through their own codes (not published in any package). Some authors, however, do not even specify the software used (see Table 3). As in a meta-analysis, we gathered the simulations presented in the papers described, since the results are not directly comparable, the tables synthesize the main parameters characterizing the simulations.

Hence, the main purpose of this review was to provide an overview of some useful components/factors characterizing each selection criterion, so that users can identify which method to apply in a specific situation also. In addition, an effort was made to tidy up the notation used in the literature, by “translating,” if necessary, symbols and formulas in each paper into a common “language.”

## References

- Ahn, M., Zhang, H.H., Lu, W.: Moment-based method for random effects selection in linear mixed models. *Stat. Sin.* **22**(4), 1539 (2012)
- Akaike, H.: Information theory and an extension of the maximum likelihood principle. *Breakthroughs in Statistics*, pp. 610–624. Springer, Berlin (1992)
- Bondell, H.D., Krishna, A., Ghosh, S.K.: Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**(4), 1069–1077 (2010)
- Bozdogan, H.: Model selection and Akaike’s information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**, 345–370 (1987)
- Braun, J., Held, L., Ledergerber, B.: Predictive cross-validation for the choice of linear mixed-effects models with application to data from the Swiss HIV Cohort Study. *Biometrics* **68**(1), 53–61 (2012)
- Bühlmann, P., van de Geer, S.: *Statistics for High-Dimensional Data*. Springer, Berlin (2011)
- Chen, Z., Dunson, D.B.: Random effects selection in linear mixed models. *Biometrics* **59**(4), 762–769 (2003)
- Chen, F., Li, Z., Shi, L., Zhu, L.: Inference for mixed models of anova type with high-dimensional data. *J. Multivar. Anal.* **133**, 382–401 (2015)
- Dimova, R.B., Markatou, M., Talal, A.H.: Information methods for model selection in linear mixed effects models with application to HCV data. *Comput. Stat. Data Anal.* **55**(9), 2677–2697 (2011)
- Fan, Y., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**(456), 1348–1360 (2001)
- Fan, Y., Li, R.: Variable selection in linear mixed effects models. *Ann. Stat.* **40**(4), 2043–2068 (2012)
- Fan, Y., Qin, G., Zhu, Z.Y.: Robust variable selection in linear mixed models. *Commun. Stat. Theory Methods* **43**(21), 4566–4581 (2014)
- Frank, I.E., Friedman, J.H.: A statistical view of some chemometric regression tools. *Technometrics* **35**, 109–148 (1993)
- Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. *Ann. Appl. Stat.* **1**, 302–332 (2007)
- Fu, W.J.: Penalized regressions: the bridge versus the LASSO. *J. Comput. Graph. Stat.* **7**, 397–416 (1998)
- Ghosh, A., Thoresen, M.: Non-concave penalization in linear mixed-effects models and regularized selection of fixed effects. *ASTA Adv. Stat. Anal.* **102**(2), 179–210 (2018)
- Gilmour, S.G.: The interpretation of mallow’s cp statistic. *The Statistician* **45**, 49–56 (1996)
- Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007)

- Greven, S., Kneib, T.: On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika* **97**, 773–789 (2010)
- Han, B.: Conditional akaike information criterion in the Fay-Herriot model. *Stat. Methodol.* **11**, 53–67 (2013)
- Hansen, M.H., Yu, B.: Minimum description length model selection criteria for generalized linear models. *Stat. Sci. A Festschrift Terry Speed* **40**, 145–163 (2003)
- Hodges, J.S., Sargent, D.J.: Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **88**, 367–379 (2001)
- Hossain, S., Thomson, T., Ahmed, E.: Shrinkage estimation in linear mixed models for longitudinal data. *Metrika* **81**(5), 569–586 (2018)
- Hui, F.K., Müller, S., Welsh, A.: Joint selection in mixed models using regularized PQL. *J. Am. Stat. Assoc.* **112**(519), 1323–1333 (2017)
- Ibrahim, J.G., Zhu, H., Garcia, R.I., Guo, R.: Fixed and random effects selection in mixed effects models. *Biometrics* **67**(2), 495–503 (2011)
- Jiang, J., Rao, J.S.: Consistent procedures for mixed linear model selection. *Sankhya Ser A* **65**(1), 23–42 (2003)
- Jiang, J., Rao, J.S., Gu, Z., Nguyen, T., et al.: Fence methods for mixed model selection. *Ann. Stat.* **36**(4), 1669–1692 (2008)
- Jiang, J., Nguyen, T., Rao, J.S.: A simplified adaptive fence procedure. *Stat. Probab. Lett.* **79**, 625–629 (2009)
- Kawakubo, Y., Kubokawa, T.: Modified conditional AIC in linear mixed models. *J. Multivar. Anal.* **129**, 44–56 (2014)
- Kawakubo Y, Sugasawa S, Kubokawa T, et al. (2014) Conditional AIC under covariate shift with application to small area prediction. Technical report, CIRJE, Faculty of Economics, University of Tokyo
- Kawakubo, Y., Sugasawa, S., Kubokawa, T.: Conditional akaike information under covariate shift with application to small area estimation. *Can. J. Stat.* **46**(2), 316–335 (2018)
- Kubokawa, T.: Conditional and unconditional methods for selecting variables in linear mixed models. *J. Multivar. Anal.* **102**(3), 641–660 (2011)
- Kubokawa, T., Srivastava, M.S.: An empirical Bayes information criterion for selecting variables in linear mixed models. *J. Jpn. Stat. Soc.* **40**(1), 111–131 (2010)
- Kuran, Ö., Özkale, M.R.: Model selection via conditional conceptual predictive statistic under ridge regression in linear mixed models. *J. Stat. Comput. Simul.* **89**(1), 155–187 (2019)
- Lahiri, P., Suntorchost, J.: Variable selection for linear mixed models with applications in small area estimation. *Sankhya B* **77**(2), 312–320 (2015)
- Li, Z., Zhu, L.: A new test for random effects in linear mixed models with longitudinal data. *J. Stat. Plan. Inference* **143**(1), 82–95 (2013)
- Li, L., Yao, F., Craiu, R.V., Zou, J.: Minimum description length principle for linear mixed effects models. *Stat. Sin.* **24**, 1161–1178 (2014)
- Li, Y., Wang, S., Song, P.X.K., Wang, N., Zhou, L., Zhu, J.: Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Stat. Interface* **11**(4), 721 (2018)
- Liang, H., Wu, H., Zou, G.: A note on conditional aic for linear mixed-effects models. *Biometrika* **95**, 773–778 (2008)
- Lin, B., Pang, Z., Jiang, J.: Fixed and random effects selection by reml and pathwise coordinate optimization. *J. Comput. Graph. Stat.* **22**(2), 341–355 (2013)
- Liski EP, Liski A (2008) Model selection in linear mixed models using mdl criterion with an application to spline smoothing. In: Proceedings of the First Workshop on Information Theoretic Methods in Science and Engineering, Tampere, Finland, pp. 18–20
- Liu, X.Q., Hu, P.: General ridge predictors in a mixed linear model. *Statistics* **47**(2), 363–378 (2013)
- Lombardía, M.J., López-Vizcaíno, E., Rueda, C.: Mixed generalized Akaike information criterion for small area models. *J. R. Stat. Soc. Ser. A Stat. Soc.* **180**:1229–1252 (2017)
- Marhuenda, Y., Molina, I., Morales, D.: Small area estimation with spatio-temporal Fay-Herriot models. *Comput. Stat. Data Anal.* **58**, 308–325 (2013)
- Marino, M., Buxton, O.M., Li, Y.: Covariate selection for multilevel models with missing data. *Stat* **6**(1), 31–46 (2017)
- Marshall, E.C., Spiegelhalter, D.J.: Approximate cross-validators predictive checks in disease mapping models. *Stat. Med.* **22**, 1649–1660 (2003)

- Müller, S., Scealy, J.L., Welsh, A.H., et al.: Model selection in linear mixed models. *Stat. Sci.* **28**(2), 135–167 (2013)
- Nguyen, T., Jiang, J.: Restricted fence method for covariate selection in longitudinal data analysis. *Biostatistics* **13**(2), 303–314 (2012)
- Özkale, M.R., Can, F.: An evaluation of ridge estimator in linear mixed models: an example from kidney failure data. *J. Appl. Stat.* **44**(12), 2251–2269 (2017)
- Pan J (2016) Adaptive LASSO for mixed model selection via profile log-likelihood. Ph.D. thesis, Bowling Green State University
- Pan, J., Shang, J.: Adaptive lasso for linear mixed model selection via profile log-likelihood. *Commun. Stat. Theory Methods* **47**(8), 1882–1900 (2018a)
- Pan, J., Shang, J.: A simultaneous variable selection methodology for linear mixed models. *J. Stat. Comput. Simul.* **88**(17), 3323–3337 (2018b)
- Peng, H., Lu, Y.: Model selection in linear mixed effect models. *J. Multivar. Anal.* **109**, 109–129 (2012)
- Pu, W., Niu, X.F.: Selecting mixed-effects models based on a generalized information criterion. *J. Multivar. Anal.* **97**(3), 733–758 (2006)
- Rissanen, J.: Stochastic complexity and modeling. *Ann. Stat.* **14**(3), 1080–1100 (1986)
- Rocha, F.M., Singer, J.M.: Selection of terms in random coefficient regression models. *J. Appl. Stat.* **45**(2), 225–242 (2018)
- Rohart, F., San Cristobal, M., Laurent, B.: Selection of fixed effects in high dimensional linear mixed models using a multicycle ecm algorithm. *Comput. Stat. Data Anal.* **80**, 209–222 (2014)
- Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys, vol. 81. Wiley, Hoboken (2004)
- Schellhdorfer, J., Bühlmann, P., De Geer, S.V.: Estimation for high-dimensional linear mixed-effects models using  $l_1$ -penalization. *Scand. J. Stat.* **38**(2), 197–214 (2011)
- Schmidt, K., Smith, R.C.: A parameter subset selection algorithm for mixed-effects models. *Int. J. Uncertain. Quantif.* **6**(5), 405–416 (2016)
- Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
- Sciandra, M., Plaia, A.: A graphical model selection tool for mixed models. *Commun. Stat. Simul. Comput.* **47**(9), 2624–2638 (2018)
- Shang, J., Cavanaugh, J.E.: Bootstrap variants of the akaike information criterion for mixed model selection. *Comput. Stat. Data Anal.* **52**(4), 2004–2021 (2008)
- Singer, J.M., Rocha, F.M., Nobre, J.S.: Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *Int. Stat. Rev.* **85**(2), 290–324 (2017)
- Sorensen, G., Barbeau, E., Stoddard, A.M., Hunt, M.K., Kaphingst, K., Wallace, L.: Promoting behavior change among working-class, multiethnic workers: results of the healthy directions-small business study. *Am. J. Public Health* **95**(8), 1389–1395 (2005)
- Srivastava, M.S., Kubokawa, T.: Conditional information criteria for selecting variables in linear mixed models. *J. Multivar. Anal.* **101**(9), 1970–1980 (2010)
- Sugiura, N.: Further analysis of the data by akaike's information criterion and the finite corrections. *Commun. Stat. A* **7**, 13–26 (1978)
- Taylor, J.D., Verbyla, A.P., Cavanaugh, C., Newberry, M.: Variable selection in linear mixed models using an extended class of penalties. *Aust. N. Z. J. Stat.* **54**(4), 427–449 (2012)
- Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996)
- Vaida, F., Blanchard, S.: Conditional akaike information for mixed-effects models. *Biometrika* **92**(2), 351–370 (2005)
- Wang, W.: Identifiability of covariance parameters in linear mixed effects models. *Linear Algebra Appl.* **506**, 603–613 (2016)
- Wang, J., Schaalje, G.B.: Model selection for linear mixed models using predictive criteria. *Commun. Stat. Simul. Comput.* **38**(4), 788–801 (2009)
- Weiss, R.E.: *Modeling Longitudinal Data*. Springer, Berlin (2005)
- Wenren, C., Shang, J.: Conditional conceptual predictive statistic for mixed model selection. *J. Appl. Stat.* **43**(4), 585–603 (2016)
- Wenren, C., Shang, J., Pan, J.: Marginal conceptual predictive statistic for mixed model selection. *Open J. Stat.* **6**(02), 239 (2016)
- Wu, P., Luo, X., Xu, P., Zhu, L.: New variable selection for linear mixed-effects models. *Ann. Inst. Stat. Math.* **69**, 627–646 (2016)

- Zhang, X., Liang, H., Liu, A., Ruppert, D., Zou, G.: Selection strategy for covariance structure of random effects in linear mixed-effects models. *Scand. J. Stat.* **43**(1), 275–291 (2016)
- Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429 (2006)
- Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**(2), 301–320 (2005)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.