Statistical Science
2013, Vol. 28, No. 2, 135–167
DOI: 10.1214/12-STS410

Model Selection in Linear Mixed Models

Samuel Müller, J. L. Scealy and A. H. Welsh

Abstract. Linear mixed effects models are highly flexible in handling a broad range of data types and are therefore widely used in applications. A key part in the analysis of data is model selection, which often aims to choose a parsimonious model with other desirable properties from a possibly very large set of candidate statistical models. Over the last 5–10 years the literature on model selection in linear mixed models has grown extremely rapidly. The problem is much more complicated than in linear regression because selection on the covariance structure is not straightforward due to computational issues and boundary problems arising from positive semidefinite constraints on covariance matrices. To obtain a better understanding of the available methods, their properties and the relationships between them, we review a large body of literature on linear mixed model selection. We arrange, implement, discuss and compare model selection methods based on four major approaches: information criteria such as AIC or BIC, shrinkage methods based on penalized loss functions such as LASSO, the Fence procedure and Bayesian techniques.

Key words and phrases: AIC, Bayes factor, BIC, Cholesky decomposition, fence, information criteria, LASSO, linear mixed model, model selection, shrinkage methods.

1. INTRODUCTION

The class of linear mixed models (Henderson, 1950) provides a flexible framework for modeling a wide range of data types, including clustered, longitudinal and spatial data. This framework is increasingly widely used in Applied Statistics. It is interesting and important both in its own right and as a starting point for the development of more complicated classes of models such as generalized linear mixed models or GLMMs (e.g., McCulloch, 2003), nonlinear mixed models (e.g., Pinheiro and Bates, 2000), and various semi-parametric and nonparametric models (e.g., Ruppert, Wand and Carroll, 2003). In practical applications of statistical models (including linear mixed models), a key aspect of the analysis is often model selection, the choice of a particular model within a class of candidate models; see Claeskens and Hjort (2008) for a general review. With the increasing use of linear mixed models in practice, the need to do model selection has resulted in the implementation of a number of different methods for model selection in software packages (such as R or SAS). There are, however, other, recent methods which have not yet been implemented in standard software and there is no consensus in the statistical community on how to approach model selection for linear mixed models. This makes it very difficult for an analyst to answer the basic question: Which methods should I use and when should I use them? In this paper, as a step toward addressing these issues, we review, classify and compare a number of methods for selecting linear mixed models so that we can better understand their properties and the relationships between them.

There is a substantial literature on model selection for linear mixed models which has grown extremely rapidly in the last 5–10 years. As a consequence of

Samuel Müller is Senior Lecturer, School of Mathematics and Statistics F07, University of Sydney, NSW 2006, Australia (e-mail: samuel.mueller@sydney.edu.au). J. L. Scealy is Postdoctoral Research Associate, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia (e-mail: Janice.Scealy@anu.edu.au). A. H. Welsh is E. J. Hannan Professor of Statistics, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia (e-mail: Alan.Welsh@anu.edu.au).

this rapid growth, researchers working in parallel in the area have not had access to the developments of other researchers. The inevitable result is a lack of cross-referencing between papers using different methods for model selection, between papers using similar methods for model selection and even between papers using similar methods written by the same author. The main consequences are a limited acknowledgement of other relevant work, a limited exploration of the relationships between different methods and limited comparisons between different methods of model selection, either theoretically or through simulation. In addition, papers treating the same model use different notation and terminology; papers proposing different approaches do so for different models (including special cases of general models obtained either by imposing special structure or by treating some parameters as known) or treat different types of selection problems (such as only selecting the regression parameters), making it difficult to access and evaluate the key methods. Finally, only a few papers discuss and solve computational issues. We do not give specific examples here because our intention is not to single out any particular contributions but rather to describe the state of the literature as a whole.

Linear mixed models can be viewed as extensions of linear regression models, so many of the methods proposed for selecting mixed models can be seen as extensions of methods developed for linear regression models. However, this does not mean that model selection for linear mixed models can be subsumed within model selection for linear regression models. It is useful to exploit the similarities between the models but there are also important differences between linear mixed models and linear regression models which need to be taken into account. In linear regression models, the responses are independent, whereas, in linear mixed models, they are typically dependent. This dependence impacts on model selection by reducing the effective sample size, a quantity that affects the theoretical properties of procedures and is used explicitly in some model selection procedures such as the Bayesian Information Criteria (BIC; Schwarz, 1978) described in Section 3. The dependence also means that linear mixed models have both regression parameters (which describe the mean structure) and variance parameters (which describe the sources of variability and the dependence structure). If, as is often the case, these parameters have a different relative importance in the analysis, this should be reflected in model selection. For example, if we are evaluating a model for its predictive ability, it may be less important to get the dependence structure exactly correct than it is to get the regression structure correct. Even if we do not explicitly assign different relative importance to the parameters, it is already implicit in the model—it underlies the familiar difficulty of assigning degrees of freedom or measuring model complexity in linear mixed models. It is also often the case that regression parameters are unconstrained, whereas variance parameters are always constrained by the requirement that variance matrices must be positive semi-definite. In many problems, many of the parameters are required to be nonnegative so there are boundaries of the parameter space at zero. An important part of model selection is setting a parameter to zero which, unfortunately, means putting some of the variance parameters on the boundary. Consequently, there are boundary issues in model selection with variance parameters, either computational issues from fitting models with redundant variance parameters (as software tends not to handle this well) or statistical issues related to testing null hypotheses on the boundary of the parameter space (because selection is closely related to hypothesis testing; Claeskens and Hjort, 2008), that do not arise when selecting regression parameters. Thus, model selection for linear mixed models is different from model selection for linear regression models and it is important to acknowledge and take into account the differences between the two classes of models.

For the linear regression model there is a large and growing literature on variable selection in the high-dimensional setting (e.g., Fan and Lv, 2010; Bühlmann and van de Geer, 2011). This is very different from the fixed (finite) dimensional case because many of the fixed dimensional model selection procedures either do not work at all or, for their implementation, require some theoretical or computational adjustment. Additional assumptions such as sparsity in the true model are also needed in the high-dimensional setting in order to obtain consistent model selection. Nonetheless, sometimes similar methods can be used in both the high and fixed-dimensional cases, for example, shrinkage methods based on the LASSO (Tibshirani, 1996) are used extensively in both contexts. To date, most of the literature on model selection for the linear mixed model is for the fixed-dimensional parameter case and it is only very recently that authors have started to consider high-dimensional settings (Schelldorfer, Bühlmann and van de Geer, 2011; Fan and Li, 2012). Part of the reason for this lack of coverage is because asymptotic studies in the high-dimensional linear

mixed model case are more difficult than in the linear regression case since both the number of regression parameters and/or variance parameters can potentially grow with the sample size and at possibly different rates. There are also more complex computational and estimation issues to consider due to the presence of large, sparse covariance matrices.

In this paper we review model selection for linear mixed models focusing mostly on the fixed-dimensional parameter case. We define these models formally, distinguish different model selection problems for the models and introduce the basic notation in Section 2. We classify the different methods into four broad approaches and describe each approach in its own section. The first approach is based on choosing models to minimize information criteria such as the widely used Akaike Information Criteria (AIC; Akaike, 1973) and the Bayesian Information Criteria (BIC; Schwarz, 1978). These criteria are described in Section 3. We describe shrinkage methods like the LASSO (Tibshirani, 1996) in Section 4 and the Fence method (Jiang et al., 2008) in Section 5. We briefly discuss some Bayesian methods in Section 6. Finally, we review some published simulation results in Section 7 and conclude with discussion and conclusions in Section 8.

Although model selection can be formulated and interpreted in terms of testing, we do not review testing per se in this paper. There is a huge literature on testing, a substantial part of which could be construed to have at least some relevance to model selection, and we simply have to draw a line somewhere. We therefore focus on methods which may be motivated by and derived from tests but ultimately do not explicitly focus on tests. Second, our focus is on the ideas behind and the relationships between methods, rather than the details of the implementation of any particular method. We do identify areas of difficulty where more work is needed, including numerical and implementation issues, but these are not our main focus, and resolving them in this paper is even further from our main focus. In particular, any discussion of Bayesian methods leads quickly toward computation, but we do not review Bayesian computation.

2. THE MODEL SELECTION PROBLEM

Consider the linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\Gamma}\mathbf{u} + \boldsymbol{\Delta}\mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is a n-vector of observed responses, $\mathbf{X}$ is a known $n \times p$ matrix of covariates, $\mathbf{Z}$ is a known $n \times s$ matrix, $\mathbf{u}$ and $\mathbf{e}$ are unobserved independent s and n-vectors of independent random variables with mean zero and variance the identity matrix, $\boldsymbol{\beta}$ is a p-vector of unknown regression parameters, $\boldsymbol{\Gamma}$ is an $s \times s$ matrix which contains q_γ distinct unknown parameters and $\boldsymbol{\Delta}$ is an $n \times n$ matrix which contains q_δ distinct unknown parameters. Writing the model this way is motivated by Chen and Dunson (2003), Field, Pang and Welsh (2010), Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011). Let $\boldsymbol{\Psi} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$ and $\boldsymbol{\Sigma} = \boldsymbol{\Delta}\boldsymbol{\Delta}^T$ so we can write

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^T + \boldsymbol{\Sigma}.$$

The notation is general enough to allow the matrix square roots $\boldsymbol{\Gamma}$ and $\boldsymbol{\Delta}$ to be the symmetric matrices produced by taking the square roots of the eigenvalues in the spectral decomposition of $\boldsymbol{\Psi}$ or $\boldsymbol{\Sigma}$, the lower triangular matrices produced by the Cholesky decomposition of $\boldsymbol{\Psi}$ or $\boldsymbol{\Sigma}$, or, if $\boldsymbol{\Psi}$ is block diagonal, the block diagonal matrix of the lower triangular matrices from the Cholesky decompositions of each block. It is simpler to specify and interpret the model in terms of $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$, but it is simpler to fit and select models with $\boldsymbol{\Gamma}$ and $\boldsymbol{\Delta}$. Let $\boldsymbol{\gamma}$ denote the q_γ distinct unknown parameters in $\boldsymbol{\Gamma}$ and $\boldsymbol{\delta}$ the q_δ distinct unknown parameters in $\boldsymbol{\Delta}$. It is sometimes convenient to group the parameters into the vector of regression parameters $\boldsymbol{\beta}$, the vector of variance parameters $\boldsymbol{\tau} = (\boldsymbol{\gamma}^T, \boldsymbol{\delta}^T)^T$ of length $q = q_\gamma + q_\delta$ and the vector of all parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T)^T$ of length $p + q$.

There are other useful parametrizations for (1) which are used in the literature. One of these involves writing $\boldsymbol{\Gamma}$ as

$$\boldsymbol{\Gamma} = \mathbf{D}\boldsymbol{\Gamma}^\dagger, \tag{2}$$

where $\boldsymbol{\Gamma}^\dagger$ is lower triangular with ones on the diagonal and $\mathbf{D}$ is a diagonal matrix (Chen and Dunson, 2003). When $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$ with $\mathbf{I}_n$ the $n \times n$ identity matrix, it is sometimes convenient to write $\boldsymbol{\Gamma} = \sigma\mathbf{D}^\dagger\boldsymbol{\Gamma}^\dagger$, where $\mathbf{D}^\dagger = \mathbf{D}/\sigma$ (Bondell, Krishna and Ghosh, 2010; Saville, Herring and Kaufman, 2011). To be consistent with the terminology of Pourahmadi (2011), we will refer to these as alternative Cholesky factors. The main advantage of the alternative Cholesky parametrization is that it separates and therefore encourages different treatment of the diagonal and the off-diagonal elements of $\boldsymbol{\Gamma}$. In particular, a zero diagonal element makes the whole row zero, whereas a zero off-diagonal element affects only itself. However, it is important to keep in mind that the diagonal elements of $\boldsymbol{\Gamma}$ include

off-diagonal elements of $\boldsymbol{\Psi}$ so the order of rows and columns in $\boldsymbol{\Psi}$ can affect model selection.

An alternative to the linear mixed model (1), which is widely used in the econometric literature, can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{1/2}\boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\varepsilon}$ is an n-vector of independent random variables with mean zero and variance one. Models (1) and (3) have the same mean and variance. If all the random variables ($\mathbf{u}$, $\mathbf{e}$, $\boldsymbol{\varepsilon}$) have Gaussian distributions, the responses $\mathbf{y}$ in models (1) and (3) have the same distribution. However, the two models are not necessarily identical because they can have different parameter spaces; the parameter space for (1) requires $\boldsymbol{\Psi}$ to be positive definite, whereas that for (3) only requires $\mathbf{V}$ to be positive definite. Thus, the parameter space for (3) can be larger than and contain that for (1). If any of the random variables have non-Gaussian distributions, then the responses in the two models have the same first two moments but can have different higher order moments and different distributions (Field and Welsh, 2007), as well as different parameter spaces. We call (3) the transformation model to be consistent with Field and Welsh (2007); it is sometimes called the marginal model (e.g., Jiang, 2007). The difference between the two models is not widely appreciated, but it is important to be clear about which model each procedure is working with. Most model selection procedures have been derived for the linear mixed model (1), but some of them also apply to the transformation model (3).

It is useful to identify some special cases of the model because these give insight into the range of forms of the model and because we will refer to them specifically in what follows. We express these as special cases of the linear mixed model (1); they can also be expressed as special cases of the transformation model (3).

Variance component model (Henderson, 1950): $\boldsymbol{\Psi} = \operatorname{block\,diag}(\gamma_1^2\mathbf{I}_{r_1}, \ldots, \gamma_{q_\gamma}^2\mathbf{I}_{r_{q_\gamma}})$, where $s = \sum_{k=1}^{q_\gamma} r_k$. Write $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(q_\gamma)})$, where $\mathbf{Z}^{(k)}$ is $n \times r_k$, and $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_{q_\gamma}^T)^T$, where $\mathbf{u}_k$ is a r_k-vector, so that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \gamma_1\mathbf{Z}^{(1)}\mathbf{u}_1 + \cdots + \gamma_{q_\gamma}\mathbf{Z}^{(q_\gamma)}\mathbf{u}_{q_\gamma} + \boldsymbol{\Delta}\mathbf{e}. \tag{4}$$

Often, $\boldsymbol{\Sigma}$ is known up to an unknown constant; in this case $q_\delta = 1$ and we can write $\boldsymbol{\Sigma} = \mathbf{R}_0 + \delta^2\mathbf{R}_1$, with $\mathbf{R}_0$ and $\mathbf{R}_1$ known. It is most common to have $\mathbf{R}_0 = \mathbf{0}$ and $\mathbf{R}_1 = \mathbf{I}_n$, the $n \times n$ identity matrix, but other possibilities do occur. The parameters $\gamma_1^2, \ldots, \gamma_{q_\gamma}^2, \delta^2$ are known as variance components.

Independent cluster model: $\boldsymbol{\Psi} = \operatorname{block\,diag}(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_m)$, where $\boldsymbol{\Psi}_i$ is $s_i \times s_i$ and $s = \sum_{i=1}^m s_i$, and $\boldsymbol{\Sigma} = \operatorname{block\,diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m)$, where $\boldsymbol{\Sigma}_i$ is $n_i \times n_i$ and $n = \sum_{i=1}^m n_i$. Write $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_m^T)^T$, where $\mathbf{y}_i$ is an n_i-vector, $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_m^T)^T$, where $\mathbf{X}_i$ is an $n_i \times p$ matrix, $\mathbf{Z} = \operatorname{block\,diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$, where $\mathbf{Z}_i$ is an $n_i \times s_i$ matrix, and $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_m^T)^T$, where $\mathbf{u}_i$ is an s_i-vector, and $\mathbf{e} = (\mathbf{e}_1^T, \ldots, \mathbf{e}_m^T)^T$, where $\mathbf{e}_i$ is an n_i-vector. Then, if $\boldsymbol{\Gamma}$ and $\boldsymbol{\Delta}$ are block diagonal square roots of $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ with $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Delta}_i$ on the diagonal, respectively, we can write (1) as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\Gamma}_i\mathbf{u}_i + \boldsymbol{\Delta}_i\mathbf{e}_i, \quad i = 1, \ldots, m. \tag{5}$$

The observations $\mathbf{y}_1, \ldots, \mathbf{y}_m$ from distinct clusters are independent random vectors.

The independent cluster model is also called the Laird–Ware model, though perhaps this should be restricted to the case with constant s_i (Laird and Ware, 1982). The assumption of independence between clusters makes the model easier to work with than spatial and other models with more complete dependence structures. For this reason, much of the work on linear mixed models and model selection for linear mixed models has been carried out for the independent cluster model.

Clustered variance component model: A combination of the variance component model and the independent cluster model obtained as a special case of the independent cluster model with $\boldsymbol{\Psi}_i = \operatorname{block\,diag}(\gamma_1^2\mathbf{I}_{r_{i1}}, \ldots, \gamma_{q_\gamma}^2\mathbf{I}_{r_{iq_\gamma}})$, where $s_i = \sum_{k=1}^{q_\gamma} r_{ik}$. Write $\mathbf{Z}_i = (\mathbf{Z}_i^{(1)}, \ldots, \mathbf{Z}_i^{(q_\gamma)})$, where $\mathbf{Z}_i^{(k)}$ is $n_i \times r_{ik}$, and $\mathbf{u}_i = (\mathbf{u}_{i1}^T, \ldots, \mathbf{u}_{iq_\gamma}^T)^T$, where $\mathbf{u}_{ik}$ is a r_{ik}-vector. Then we can write (1) as

$$\begin{aligned}\mathbf{y}_i = {} & \mathbf{X}_i\boldsymbol{\beta} + \gamma_1\mathbf{Z}_i^{(1)}\mathbf{u}_{i1} + \cdots \\ & + \gamma_{q_\gamma}\mathbf{Z}_i^{(q_\gamma)}\mathbf{u}_{iq_\gamma} + \boldsymbol{\Delta}_i\mathbf{e}_i, \quad i = 1, \ldots, m.\end{aligned} \tag{6}$$

Random intercept and slope regression model: A special case of the clustered variance component model where the first column of $\mathbf{X}_i$ is $\mathbf{1}_{n_i} = (1, \ldots, 1)^T$ and the $\mathbf{Z}_i^{(k)}$ are equal to the columns of $\mathbf{X}_i$. It has $s = mp$ and $q_\gamma = p$. We also include the model in which the $\mathbf{Z}_i^{(k)}$ include the column of ones and a (nonempty) subset of the columns of $\mathbf{X}_i$. We call the model with $\mathbf{Z}_i = \mathbf{Z}_i^{(1)} = \mathbf{1}_{n_i}$ the random intercept regression model; it is also sometimes called the nested error regression model. It has $s = m$ and $q_\gamma = 1$. In the multilevel model literature (e.g., Snijders and Bosker, 1999), it is common to allow the random intercept and

slopes to be correlated, but they are usually treated as independent in the general literature.

Fay–Herriot model (Fay and Herriot, 1979): A special case of the random intercept regression model with $n_i = 1$, $\mathbf{Z} = \mathbf{I}_n$, $\mathbf{\Psi} = \gamma^2 \mathbf{I}_n$ and $\mathbf{\Sigma} = \operatorname{diag}(r_1, \ldots, r_n)$ is known, so $q = q_\gamma = 1$. Here $s = n$ and the matrix $\mathbf{\Sigma}$ is assumed known because it is not identifiable.

Longitudinal autoregression model: A special case of the independent cluster model with $s_i = n_i$, $\mathbf{Z}_i = \mathbf{I}_{n_i}$ and $\mathbf{\Psi}_i = (\psi_{ijk})$ is the $n_i \times n_i$ matrix where

$$\psi_{ijk} = \begin{cases} \sigma^2, & j = k, \\ \sigma^2 \phi^{|j-k|}, & j \neq k, \end{cases}$$

with $-1 < \phi < 1$, $1 \leq j, k \leq n_i$. Thus, $\boldsymbol{\gamma} = (\sigma, \phi)$ and $q_\gamma = 2$, $q_\delta = 1$.

Linear regression model: A special case of all the above models but a trivial linear mixed model, the linear regression model has $\mathbf{\Gamma} = \mathbf{0}$ and $\mathbf{\Delta} = \sigma^2 \mathbf{I}_n$.

We consider the selection of linear mixed models $M \in \mathcal{M}$, where $\mathcal{M} = \{M_l : l > 1\}$ is a countable set of distinct models which we call candidate models. Unlike in regression models, we cannot uniquely identify a model M by its nonzero parameter vector $\boldsymbol{\theta}_M = (\boldsymbol{\beta}_M^T, \boldsymbol{\tau}_M^T)^T$, because setting one element of $\boldsymbol{\tau}_M$ equal to zero may allow other (redundant) elements to take arbitrary values. For example, in the longitudinal autoregression model, if $\sigma^2 = 0$, then the parameter ϕ is arbitrary, although any choice of ϕ gives the same model. We adopt the convention of setting redundant parameters equal to a convenient, problem specific value (such as zero if it is part of the parameter space) so we can still distinguish models by their nonzero parameters. Some parameters are naturally grouped together (such as the coefficients for different levels of a factor) and it is useful in model selection to treat them as a group rather than as separate parameters. Also, some of the parameters such as the intercept, coefficients of particular variables, the error variance σ^2 when $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$ or specific covariance parameters can be retained in all models $M \in \mathcal{M}$. An extreme version of this occurs when the variance structure can be regarded as known from the way the data are collected (e.g., from the structure of the experiment), so is held fixed in $\mathcal{M}$. (It is generally less meaningful to select across the variance structure while retaining all the regression parameters in the model.) We will take it as understood that, depending on the context, the definition of $\mathcal{M}$ encompasses a range of possibilities. When a data generating model M_t exists we call it the true model and any model M_l that is more complex than the true model and satisfies $M_t \subseteq M_l$ (or $\boldsymbol{\theta}_{M_t} \subseteq \boldsymbol{\theta}_{M_l}$) is called a correct model. We denote the set of correct models $\mathcal{M}_c$. We assume that the complexity (sometimes called the dimensionality or cardinality) of a model, d_M, can be calculated and satisfies $d_{M_1} < d_{M_2}$ if $M_1 \subset M_2$. We will show later (see Section 3.2) that model complexity depends on the data, the model and sometimes on the estimation or model selection technique. It can be useful to identify a fixed (or full) model M_f, which has maximal model complexity and can be used as the initial model in stepwise model selection algorithms or to calculate initial parameter estimates, for example, for the Adaptive LASSO (Section 4).

We have described the model selection problem in terms of the set $\mathcal{M}$ or in terms of the parameters of the models in $\mathcal{M}$. The problem can also be described in terms of variables and, while these are similar, it turns out that they are not necessarily the same. When we describe the problem in terms of selecting variables rather than parameters, we focus on selecting columns or groups of columns in $\mathbf{X}$ and/or $\mathbf{Z}$. Selecting columns of $\mathbf{X}$ is the same as selecting nonzero parameters in $\boldsymbol{\beta}$, but selecting columns of $\mathbf{Z}$ is the same as selecting whole rows of $\mathbf{\Gamma}$ (and hence rows and columns of $\mathbf{\Psi}$) rather than selecting individual nonzero parameters in $\mathbf{\Gamma}$. This is shown neatly by our writing the relevant term in the model as $\mathbf{Z\Gamma u}$ and highlights one of the important differences between the regression and the variance parameters (which makes model selection in linear mixed or transformation models different from model selection in linear regression models). In terms of the alternative Cholesky factors, selecting columns of $\mathbf{Z}$ is equivalent to selecting the diagonal elements of $\mathbf{D}$ or $\mathbf{D}^\dagger$ while treating the terms in $\mathbf{\Gamma}^\dagger$ as nuisance parameters. Selection on $\boldsymbol{\beta}$ or $\mathbf{X}$ is sometimes called selecting fixed effects, while selection on $\mathbf{Z}$ is sometimes called selecting random effects. This is slightly misleading terminology because we are not directly selecting components of the random effects $\mathbf{u}$ and it is not really applicable to the transformation model (3) which does not include random effects. We will consider the more general problem of selecting on the parameters $\boldsymbol{\theta}$ and refer to selecting regression parameters $\boldsymbol{\beta}$ and variance parameters $\boldsymbol{\tau}$ rather than to selecting fixed or random effects.

Model selection is often carried out by choosing models in $\mathcal{M}$ that minimize a specific criterion. This usually involves a trade-off between the closeness of the fit to the data and the complexity of the model. As a practical matter, since the ultimate use of a selected model may be different from that for which it is selected, it may be useful to consider several criteria (as

was done explicitly for the linear regression model in Müller and Welsh, 2010) and in fact include other considerations such as the performance in diagnostic plots.

The important problem of specifying the distributions of the random variables in a model is not usually regarded as part of model selection. Insofar as model selection is both a selection of the model and the method of estimation being used to fit the model, it can implicitly also involve a choice of underlying distributions, although it would be better if this choice were taken seriously and made more explicit, as it should also affect the choice of model selection method. Most of the papers on model selection of linear mixed models assume that all the distributions are Gaussian, although some do explore the effect of non-Gaussian distributions in simulations (e.g., Dimova, Markatou and Talal, 2011; Kubokawa, 2011). One exception is Ahn, Zhang and Lu (2012) who propose a model selection method based on moment estimation which does not require any distributional assumptions.

In addition to thinking about how we want to select a model, we also need to think about how we evaluate model selection methods. If we use the criterion which defines one of the model selection methods, then we bias the evaluation in favor of that method. This is noted by Müller and Welsh (2005, 2009) in the context of robust model selection. For this reason, we suggest using criteria which are not directly related to the definition of any specific method. These include the probability of selecting the true model, the probability of selecting a model from a subset of correct models in the neighborhood of the true model, the probability of selecting a correct model (Jiang, Nguyen and Rao, 2008, 2009) or the mean squared error of the difference between the predictions from the selected model and the predictions from the true model fitted by maximum likelihood estimation (Bondell, Krishna and Ghosh, 2010; Ibrahim et al., 2011). The performance of the model selection methods usually depends on the class of candidate models $\mathcal{M}$, the true model and the data. As with linear regression models, no single method for model selection will always perform best.

For the linear mixed model (1), the log density of $\mathbf{y}$ given $\mathbf{u}$ viewed as a function of the parameters is sometimes called the conditional log-likelihood. If $\mathbf{e}$ has a Gaussian distribution, the conditional log-likelihood is

$$\ell(\boldsymbol{\theta}|\mathbf{u}) = -\tfrac{1}{2}\{\log|\boldsymbol{\Sigma}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\Gamma}\mathbf{u})^T \cdot \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\Gamma}\mathbf{u})\} \tag{7}$$

and, for simplicity, we omit here and below the constant $-\frac{n}{2}\log 2\pi$ term. Let $\langle\mathbf{u}\rangle$ denote the density of $\mathbf{u}$. If $\mathbf{u}$ has a Gaussian distribution, the log-likelihood (sometimes called the marginal log-likelihood) is

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log\left[\int \exp\{\ell(\boldsymbol{\theta}|\mathbf{u})\}\langle\mathbf{u}\rangle\, d\mathbf{u}\right] \\ &= -\frac{1}{2}\{\log|\mathbf{V}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}.\end{aligned} \tag{8}$$

This is also the log-likelihood of the Gaussian transformation model (3). For fixed $\boldsymbol{\tau}$, the log-likelihood $\ell(\boldsymbol{\theta})$ is maximized over $\boldsymbol{\beta}$ by the generalized least squares estimator

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}. \tag{9}$$

Modifying the profile log-likelihood $\ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau}), \boldsymbol{\tau})$ by including a bias adjustment yields the useful restricted maximum likelihood (REML) criterion function

$$\ell_R(\boldsymbol{\tau}) = -\tfrac{1}{2}\{\log|\mathbf{V}| + \log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}^T\mathbf{P}^{-1}\mathbf{y}\},$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$ (Patterson and Thompson, 1971). Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\tau}}$ be maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, respectively, and let $\widehat{\boldsymbol{\tau}}_R$ be a REML estimator of $\boldsymbol{\tau}$. Put $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\tau}}_R)$.

Many of the desirable properties of maximum likelihood and REML estimators are asymptotic properties and some model selection methods use these with asymptotic expansions and approximations for their derivation or justification. There are various ways to think about asymptotics in this problem. The simplest is to let $n \to \infty$ in such a way that various matrices (such as $n^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$) converge to positive definite limits. For independent cluster models, the standard methods are to allow the number of independent groups or clusters $m \to \infty$ with either $\max(n_i)$ bounded or $\min(n_i) \to \infty$. In this model, the case of m fixed and $\min(n_i) \to \infty$ is only useful if $\boldsymbol{\Psi}$ is known because otherwise $\boldsymbol{\Psi}$ cannot be estimated consistently. Most methods also impose further restrictions on the dimensions of the model. The usual fixed parameter case has $p+q \ll n$, although some estimation methods even require $p+q+s \ll n$.

3. INFORMATION CRITERIA

Information criteria are widely used to compare and select models. In practice, they are applied by finding the model that minimizes an estimate of a criterion that is generally of the form $Q_M(\widehat{\boldsymbol{\theta}}_M) + \alpha_n(d_M)$, where Q_M is a loss function which, for candidate models M_1 and M_2 satisfying $M_1 \subset M_2$, satisfies $Q_{M_2}(\widehat{\boldsymbol{\theta}}_{M_2}) \le Q_{M_1}(\widehat{\boldsymbol{\theta}}_{M_1})$ (it is often minus twice the log-likelihood or a closely related function) and the penalty function α_n

is a function of the model complexity d_M. There are a number of approaches to obtaining information criteria such as the Akaike approach, Schwarz's Bayesian approach, etc. and within these there can be multiple possible criteria. For example, for the linear mixed model (1) to define the loss function we can use the log-likelihood, the conditional log-likelihood or the REML criterion and for the transformation model (3) we can use the log-likelihood or the REML criterion. For the linear regression model, α_n is often just a function of the number of parameters in the model (which in the present context is $p+q$; see Müller and Welsh, 2010, for a review) but for linear mixed models can be more complicated.

The Akaike Information (Akaike, 1973) is a measure of the ability of a model fitted using a particular estimator to predict an independent copy of the observed data. The particular measure used is the expectation over both the data and the independent copy of the data, of minus twice the logarithm of a density-like function representing the model which is evaluated at the independent copy of the data and the estimator of the unknown parameters based on the data. This definition is of necessity vague because we can define different versions of the Akaike Information using different log density-like functions and we can consider various estimators of $\boldsymbol{\theta}$ in these functions. In particular, if we let $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ be an estimator of $\boldsymbol{\theta}$ based on the data $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{Z}$, and let $\mathbf{y}^*$ be an independent copy of $\mathbf{y}$, then the marginal Akaike Information for a class of distributions with density-like function $g(\mathbf{y};\boldsymbol{\theta})$ is $-2\,\mathrm{E}_{\mathbf{y}}\,\mathrm{E}_{\mathbf{y}^*}\log[g\{\mathbf{y}^*;\widehat{\boldsymbol{\theta}}(\mathbf{y})\}]$ and the conditional Akaike Information for a class of distributions with conditional (i.e., $\mathbf{y}|\mathbf{u}$) density-like function $f(\mathbf{y};\boldsymbol{\theta},\mathbf{u})$ is $-2\,\mathrm{E}_{\mathbf{y},\mathbf{u}}\,\mathrm{E}_{\mathbf{y}^*|\mathbf{u}}\log[f\{\mathbf{y}^*;\widehat{\boldsymbol{\theta}}(\mathbf{y}),\widehat{\mathbf{u}}(\mathbf{y})\}]$, where $\widehat{\mathbf{u}}(\mathbf{y})$ is a predictor of $\mathbf{u}$. The expectations in the marginal case are taken with respect to either the linear mixed model (1) or the transformation model (3) and in the conditional case they are taken with respect to the linear mixed model (1). The marginal Akaike Information (based on the log-likelihood or the REML criterion) is meaningful when the independent copy of the data $\mathbf{y}^*$ is independent of $\mathbf{y}$; the conditional criterion (based on the conditional log-likelihood) is meaningful for the linear mixed model (1) when $\mathbf{y}^*$ and $\mathbf{y}$ are conditionally independent given $\mathbf{u}$ so the same random effects are common to $\mathbf{y}$ and $\mathbf{y}^*$.

The model chosen from the specified class is a model that minimizes an estimator called the Akaike Information Criterion (AIC) of the Akaike Information. Depending on how we define the Akaike Information, it is natural to consider estimating it using minus twice the corresponding log-likelihood, REML criterion or conditional log-likelihood. These functions are biased estimators of the Akaike Information because they use the same observed data $\mathbf{y}$ both to compute the parameter estimator and to evaluate the function itself. We can evaluate the bias and try to make an approximate adjustment for it: The penalty $\alpha_n(d_M)$ in an AIC can be interpreted as an adjustment to reduce bias. Much of the focus in the literature is on adjusting the bias to obtain a good estimator of the Akaike Information, although this is not the real problem in model selection.

Model selection methods like AIC which make use of the log-likelihood are closely related to likelihood ratio tests in which the models correspond to different hypotheses, a relationship which implies that hypothesis tests can be used to suggest new criteria. The important Bayesian Information Criterion or BIC (Schwarz, 1978) can be derived as an approximation to the Bayes factor for testing two hypotheses M_0 and M_1 or from asymptotic arguments to construct criteria which lead to consistent model selection.

There are a number of other information criteria in the literature. They are derived for various reasons from various considerations. Some of them are simply general criteria which could be applied in the linear mixed model, others have been applied to or developed for the linear mixed model. It is noteworthy that these are mainly marginal criteria (i.e., based on the log-likelihood) and that there are not many proposals outside the AIC framework for conditional criteria (i.e., based on the conditional log-likelihood). We discuss AIC, BIC and some of the other criteria in the subsections that follow.

3.1 Marginal AIC

The most widely used AIC criterion is what Vaida and Blanchard (2005) call the marginal AIC criterion, namely,

$$mAIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2a_n(p+q), \tag{10}$$

where $a_n = 1$ or $a_n = n/(n-p-q-1)$ in the finite sample form (Sugiura, 1978). For the independent cluster model, *mAIC* is asymptotically equivalent to leave-one-cluster-out cross-validation using a marginal generalized least squares criterion (Fang, 2011); see Section 3.4. The R function `lme()` uses *mAIC* with $a_n = 1$ and SAS `Proc Mixed` uses both the asymptotic and the finite sample forms.

The marginal AIC represents the application of a general theory to the linear mixed model (1) or the

transformation model (3) without taking into account the specific nature of these models. For the linear mixed model (1) with $\boldsymbol{\Psi} = \gamma^2 \mathbf{A}^T\mathbf{A}$, $\mathbf{A}$ known and $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$ so $q = 2$, Greven and Kneib (2010) show that *mAIC* is positively biased for the marginal Akaike Information, where the bias depends on the unknown variance parameter γ^2 and does not vanish asymptotically if $\gamma^2 = 0$. This means that there is no simple bias correction to make *mAIC* exactly unbiased and the fact that the variance parameters are constrained by the boundary of the parameter space ought to be built into the penalty. A further issue with *mAIC* is that the model complexity term in the penalty $p + q$ gives the same weight to the parameters in $\boldsymbol{\beta}$ and the parameters in $\boldsymbol{\tau}$. There is no obvious reason why this should be the case; in the variance component model, each parameter γ_k in $\boldsymbol{\tau}$ represents the variance of r_k random effects so is absorbing r_k other estimates and arguably should be given a greater weight to reflect this. The precise form of a penalty giving different weight to the parameters in $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ is not clear. However, it is possible that using a different criterion to minus twice the log-likelihood may have the effect of rescaling the parameters so that it then makes sense to give them equal weight in the penalty.

Shang and Cavanaugh (2008) propose using the bootstrap to estimate the appropriate bias-adjustment or penalty for marginal AIC. Let $\{\mathbf{y}_b^* : b = 1, \ldots, B\}$ be a set of B bootstrap samples of $\mathbf{y}$, let $\{\ell_b^*(\boldsymbol{\theta}) : b = 1, \ldots, B\}$ denote the log-likelihoods for each of the B bootstrap samples [i.e., $\ell_b^*(\boldsymbol{\theta})$ is the log-likelihood for $\boldsymbol{\theta}$ computed using $\mathbf{y}_b^*$] and let $\{\widehat{\boldsymbol{\theta}}_b^* : b = 1, \ldots, B\}$ denote the maximum likelihood estimates for each bootstrap sample. Then Shang and Cavanaugh propose the bootstrap AIC criteria

$$mAIC_{B1} = -2\ell(\widehat{\boldsymbol{\theta}}) - \frac{2}{B}\sum_{b=1}^{B}\{\ell(\widehat{\boldsymbol{\theta}}_b^*) - \ell_b^*(\widehat{\boldsymbol{\theta}}_b^*)\}$$

and

$$\begin{aligned} mAIC_{B2} &= -2\ell(\widehat{\boldsymbol{\theta}}) - \frac{4}{B}\sum_{b=1}^{B}\{\ell(\widehat{\boldsymbol{\theta}}_b^*) - \ell(\widehat{\boldsymbol{\theta}})\} \\ &= 2\ell(\widehat{\boldsymbol{\theta}}) - \frac{4}{B}\sum_{b=1}^{B}\ell(\widehat{\boldsymbol{\theta}}_b^*). \end{aligned}$$

In their simulations, Shang and Cavanaugh use the parametric bootstrap but other types of bootstrap could be used.

Rather than applying general results to the specific context, Srivastava and Kubokawa (2010) obtain a different criterion by working directly within the linear mixed model (1) with $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$. Treating $\boldsymbol{\Psi}/\sigma^2$ as known (so there are $p + 1$ unknown parameters $\boldsymbol{\beta}$ and σ^2), they obtain the criterion

$$mAIC_{SK} = -2\ell(\widehat{\boldsymbol{\theta}}) + \frac{2n(p+1)}{n-p-2}. \tag{11}$$

(They do not assume that $\mathbf{X}$ is of full rank so their expression has rank $(\mathbf{X})$ in place of p.) When $\boldsymbol{\Psi}/\sigma^2$ is unknown, they replace it by an estimator without any further adjustment for estimating these additional parameters. There seems little reason to expect the criterion to perform well in this case, unless the number of additional variance parameters q_γ is small.

The REML criterion function $\ell_R(\boldsymbol{\tau})$ is a modified profile likelihood for $\boldsymbol{\tau}$ so is not a function of $\boldsymbol{\beta}$. This seems to imply that $\ell_R(\boldsymbol{\tau})$ may not be useful for selecting regression parameters. It is, however, an implicit function of $\boldsymbol{\beta}$ in the sense that we need to specify an $\mathbf{X}$ when we do the profiling and different choices of $\mathbf{X}$ correspond to different choices of $\boldsymbol{\beta}$ which change the value of $\ell_R(\boldsymbol{\tau})$. This means that we can in fact consider using a version of marginal AIC based on the REML criterion function for model selection. SAS `Proc Mixed` uses

$$mAIC_R = -2\ell_R(\widehat{\boldsymbol{\tau}}) + 2a_n^* q, \tag{12}$$

with $a_n^* = (n-p)/(n-p-q-1)$. The form of a_n^* is related to the second order adjustment a_n of Sugiura (1978) after adjusting the sample size for implicitly having estimated $\boldsymbol{\beta}$. There is no other penalty for implicitly having estimated $\boldsymbol{\beta}$; this is the antithesis of the Srivastava–Kubokawa criterion (11) which does not adjust the complexity in the penalty $p + 1$ for estimating $\boldsymbol{\Psi}/\sigma^2$, because it does not adjust the complexity q for estimating $\boldsymbol{\beta}$ and it is unclear how well this criterion performs.

Kubokawa (2011) considers using marginal AIC with minus twice the log-likelihood evaluated at the generalized least squares estimator $\boldsymbol{\beta}(\boldsymbol{\tau}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ of $\boldsymbol{\beta}$ and a general consistent estimator $\widehat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$ which admits an expansion of the form

$$\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau} = \mathbf{t}_1(\boldsymbol{\tau}) + \mathbf{t}_2(\boldsymbol{\tau}) + O_p\big(n^{-3/2}\big), \tag{13}$$

where $\mathrm{E}\{\mathbf{t}_1(\boldsymbol{\tau})\} = \mathrm{E}\{(t_{11}(\boldsymbol{\tau}), \ldots, t_{1q}(\boldsymbol{\tau}))^T\} = \mathbf{0}$, $\mathbf{t}_1(\boldsymbol{\tau}) = O_p(n^{-1/2})$ and $\mathbf{t}_2(\boldsymbol{\tau}) = (t_{21}(\boldsymbol{\tau}), \ldots, t_{2q}(\boldsymbol{\tau}))^T = O_p(n^{-1})$. He suggests replacing the penalty in (10) by $2\{p + h_m(\widehat{\boldsymbol{\tau}})\}$, where

$$h_m(\boldsymbol{\tau}) = \frac{1}{2}\sum_{k=1}^{q}\mathrm{E}\left[\operatorname{trace}\left\{\frac{\partial \mathbf{V}}{\partial \tau_i}\frac{\partial^2 t_{1i}(\boldsymbol{\tau})}{\partial \mathbf{y}\,\partial \mathbf{y}^T}\right\}\right].$$

This shows the effect of using different estimators and confirms that the AIC approach depends both on the model and the estimator used to fit it. When maximum likelihood or REML are used to estimate $\boldsymbol{\tau}$, the penalty reduces back to $2(p+q)$ and we obtain *mAIC* defined in (10).

3.2 Conditional AIC

The conditional Akaike Information is defined only for the linear mixed model (1) and not for the transformation model (3). We need to predict $\mathbf{u}$ or, equivalently and more conveniently, $\mathbf{v} = \boldsymbol{\Gamma}\mathbf{u}$ as well as estimate the parameters $\boldsymbol{\theta}$ so there are $p+q+s$ unknown quantities to estimate. When the variance parameters $\boldsymbol{\tau}$ are known, $\mathbf{v}$ is often predicted using the best linear unbiased predictor (BLUP)

$$\widehat{\mathbf{v}}(\boldsymbol{\tau}) = \boldsymbol{\Psi}\mathbf{Z}^T\mathbf{V}^{-1}\{\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})\}, \tag{14}$$

where $\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ is the generalized least squares estimator defined in (9). When $\boldsymbol{\tau}$ is unknown, we use an estimated BLUP or EBLUP $\widehat{\mathbf{v}} = \widehat{\mathbf{v}}(\widehat{\boldsymbol{\tau}})$. Since we are working with $\mathbf{v}$ rather than $\mathbf{u}$ in this section, it is convenient to treat the conditional log-likelihood (7) as a function of $\mathbf{v}$ rather than $\mathbf{u}$.

The generalized least squares estimator $\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ and the BLUP $\widehat{\mathbf{v}}(\boldsymbol{\tau})$ of $\mathbf{v}$ can be obtained as the solution of Henderson's (1950) mixed model equations

$$\begin{pmatrix} \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \boldsymbol{\Psi}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}(\boldsymbol{\tau}) \\ \widehat{\mathbf{v}}(\boldsymbol{\tau}) \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{pmatrix} \boldsymbol{\Sigma}^{-1}\mathbf{y}.$$

These equations enable us to write $\mathbf{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau}) + \mathbf{Z}\widehat{\mathbf{v}}(\boldsymbol{\tau}) = \mathbf{H}_1(\boldsymbol{\tau})\mathbf{y}$, where

$$\mathbf{H}_1(\boldsymbol{\tau}) = (\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \boldsymbol{\Psi}^{-1} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \mathbf{X}^T\boldsymbol{\Sigma}^{-1} \\ \mathbf{Z}^T\boldsymbol{\Sigma}^{-1} \end{pmatrix},$$

and then to treat $\mathbf{H}_1(\boldsymbol{\tau})$ as a "hat" matrix. In particular, when $\boldsymbol{\tau}$ is known,

$$\begin{aligned} \rho(\boldsymbol{\tau}) &= \operatorname{trace}\{\mathbf{H}_1(\boldsymbol{\tau})\} \\ &= \operatorname{trace}\big[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\boldsymbol{\Sigma}\mathbf{V}^{-1}\mathbf{X}\big] + n \\ &\quad - \operatorname{trace}(\boldsymbol{\Sigma}\mathbf{V}^{-1}) \end{aligned} \tag{15}$$

is the *effective degrees of freedom* used in estimating $\boldsymbol{\beta}$ and $\mathbf{v}$ (Hodges and Sargent, 2001). The effective degrees of freedom satisfies $p \le \rho(\boldsymbol{\tau}) \le p+s$ so lies between the degrees of freedom of the regression model without $\mathbf{v}$ and the regression model treating $\mathbf{v}$ as fixed effects (Vaida and Blanchard, 2005). Computing $\mathbf{H}_1(\boldsymbol{\tau})$ requires both $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Psi}^{-1}$ but (15) shows that computing $\rho(\boldsymbol{\tau})$ only requires $\mathbf{V}^{-1}$, which should be more stable.

We have defined the effective degrees of freedom for the general model (1). It is worth noting that most of the literature on conditional AIC actually assumes that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$ so $\boldsymbol{\delta}$ consists of just the scalar parameter σ. In this case, it is convenient to let $\boldsymbol{\Psi}_* = \boldsymbol{\Psi}/\sigma^2$ and $\mathbf{V}_* = \mathbf{Z}\boldsymbol{\Psi}_*\mathbf{Z}^T + \mathbf{I}_n$. It follows that $\mathbf{V} = \sigma^2\mathbf{V}_*$ and hence that $\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})$, $\widehat{\mathbf{v}}(\boldsymbol{\tau})$ and $\rho(\boldsymbol{\tau})$ are functions of the parameters in $\boldsymbol{\Psi}_*$. Some conditional AIC are derived by treating $\boldsymbol{\Psi}_*$ as known and subsequently replacing it by an estimator. This case is subsumed within the general notation so we handle it by drawing specific attention to it when necessary rather than by introducing additional notation.

Conditional AIC criteria are constructed using minus twice the conditional log-likelihood as the loss function plus a penalty. All the proposed criteria are of the form

$$cAIC_{\alpha_n} = -2\ell(\widehat{\boldsymbol{\theta}}|\widehat{\mathbf{v}}) + \alpha_n(\widehat{\boldsymbol{\theta}}) \tag{16}$$

with different estimators $\widehat{\boldsymbol{\theta}}$, predictors $\widehat{\mathbf{v}}$ and different penalties $\alpha_n(\widehat{\boldsymbol{\theta}})$. A summary of the proposed penalties $\alpha_n(\widehat{\boldsymbol{\theta}})$ is given in Table 1; we discuss each of these choices in turn.

Burnham and White (2002) and Burnham and Anderson (2002) propose using the function

$$\alpha_{n,BW}(\widehat{\boldsymbol{\theta}}) = 2\{\rho(\widehat{\boldsymbol{\tau}}) + q\}.$$

Here $\rho(\widehat{\boldsymbol{\tau}})$ measures the effect of estimating $\boldsymbol{\beta}$ and $\mathbf{v}$; q is included to try to accommodate the effect of estimating $\boldsymbol{\tau}$. For the case that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, Vaida and Blanchard (2005) instead suggest using

$$\begin{aligned} &\alpha_{n,VB}(\widehat{\boldsymbol{\theta}}) \\ &\quad = \frac{2n}{n-p-2}\left\{\rho(\widehat{\boldsymbol{\tau}}) + 1 - \frac{\rho(\widehat{\boldsymbol{\tau}}) - p}{n-p}\right\}. \end{aligned} \tag{17}$$

The function $\alpha_{n,VB}(\widehat{\boldsymbol{\theta}})$ is asymptotic to $2\{\rho(\widehat{\boldsymbol{\tau}}) + 1\}$, as $n \to \infty$ with p, q fixed, which is the effective degrees of freedom for estimating $\boldsymbol{\beta}$ and $\mathbf{v}$ plus one degree of freedom for estimating σ^2. Vaida and Blanchard derive this penalty assuming that $\boldsymbol{\Psi}_*$ is known and point out that, in this case, their criterion is the same as the DIC of Spiegelhalter et al. (2002); see Section 6. For the independent cluster model, *cAIC* is asymptotically equivalent to leave-one-observation-out

TABLE 1

Penalties $\alpha_n(\widehat{\boldsymbol{\theta}})$ for conditional AIC. The entries in the table are $\alpha_n(\widehat{\boldsymbol{\theta}})/2$ so the actual penalty is obtained by multiplying each entry by 2. All the criteria other than the two asterisked criteria (Burnham and White, Kubokawa) assume $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$. The references are given in an abbreviated form using the first letters of the authors' names and the last two digits of the publication date

Notation	$\alpha_n(\widehat{\boldsymbol{\theta}})/2$	Reference
Maximum likelihood (16)		
$cAIC_{BW}$	$\rho(\widehat{\boldsymbol{\tau}}) + q$	BW02*
$cAIC_{VB}$	$\rho(\widehat{\boldsymbol{\tau}}) + 1$	VB05
	$\frac{n}{n-p-2}\{\rho(\widehat{\boldsymbol{\tau}}) + 1 - \frac{\rho(\widehat{\boldsymbol{\tau}})-p}{n-p}\}$	VB05
$cAIC_{LWZ}$	$\operatorname{trace}(\partial\widehat{\mathbf{y}}/\partial\mathbf{y})$	LWZ08
	$+\widehat{\sigma}^2(\widehat{\mathbf{y}}-\mathbf{y})^T\frac{\partial\widehat{\sigma}^{-2}}{\partial\mathbf{y}} + \frac{1}{2}\widehat{\sigma}^4\operatorname{trace}(\frac{\partial^2\widehat{\sigma}^{-2}}{\partial\mathbf{y}\,\partial\mathbf{y}^T})$	(Tech. rep.)
$cAIC_{GK}$	$\widehat{\nu}(\widehat{\sigma}^2) + 1$	GK10
REML (18)		
$cAIC_{R,VB}$	$\rho(\widehat{\boldsymbol{\tau}}_R) + 1$	VB05
	$\frac{n-p-1}{n-p-2}\{\rho(\widehat{\boldsymbol{\tau}}_R) + 1 - \frac{p+1}{n-p-1}\}$	VB05
$cAIC_{R,GK}$	$\widehat{\nu}_R(\widehat{\sigma}^2_R) + 1$	GK10
Specific $\tilde{\sigma}^2_0$ (19)		
$cAIC_{SK}$	$\frac{n[\operatorname{trace}\{(\mathbf{X},\mathbf{Z})\mathbf{C}(\widehat{\boldsymbol{\tau}})\}+1]}{n-\operatorname{rank}\{(\mathbf{X},\mathbf{Z})\}-2}$	SK10
General $\widehat{\boldsymbol{\tau}}$		
$cAIC_K$	$\rho(\widehat{\boldsymbol{\tau}}) + \widehat{h}_c(\widehat{\boldsymbol{\tau}})$	K11*

cross-validation with a conditional least squares criterion (Fang, 2011); see Section 3.4. When $\boldsymbol{\Psi}_*$ is unknown, Vaida and Blanchard suggest simply using the estimated version (17), that is, without adjustment for estimating $\boldsymbol{\Psi}_*$. Their argument is that $\ell(\boldsymbol{\theta}|\mathbf{v})$ does not depend on $\boldsymbol{\Psi}_*$. This is plausible with $\mathbf{v} = \boldsymbol{\Gamma}\mathbf{u}$ absorbing $\boldsymbol{\Gamma}$ into $\mathbf{u}$ but $\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ and $\widehat{\mathbf{v}}(\boldsymbol{\tau})$ are functions of $\boldsymbol{\Psi}_*$.

In a technical report accompanying their paper, Liang, Wu and Zou (2008) propose a different penalty to take the estimation of $\boldsymbol{\Psi}_*$ into account. When $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, they propose using

$$\alpha_{n,LWZ}(\widehat{\boldsymbol{\theta}}) = 2\biggl\{\operatorname{trace}\biggl(\frac{\partial\widehat{\mathbf{y}}}{\partial\mathbf{y}}\biggr) + \widehat{\sigma}^2(\widehat{\mathbf{y}}-\mathbf{y})^T\frac{\partial\widehat{\sigma}^{-2}}{\partial\mathbf{y}} + \frac{1}{2}\widehat{\sigma}^4\operatorname{trace}\biggl(\frac{\partial^2\widehat{\sigma}^{-2}}{\partial\mathbf{y}\,\partial\mathbf{y}^T}\biggr)\biggr\},$$

where $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{v}}$. When σ^2 is known, the penalty reduces to just the first term $\operatorname{trace}(\partial\widehat{\mathbf{y}}/\partial\mathbf{y})$, which is the *generalized degrees of freedom* of Ye (1998). Greven and Kneib (2010) derive analytic representations for these penalties. Let $\widehat{\nu}(\sigma^2)$ denote the analytic representation of the generalized degrees of freedom when σ^2 is known. On the basis of simulations, Greven and Kneib suggest using $\widehat{\nu}(\widehat{\sigma}^2) + 1$ when σ^2 is unknown. Their penalty in the general case is therefore

$$\alpha_{n,GK}(\widehat{\boldsymbol{\theta}}) = 2\{\widehat{\nu}(\widehat{\sigma}^2) + 1\}.$$

The expression for $\widehat{\nu}(\sigma^2)$ is quite complicated because it explicitly allows for the variance parameters to lie on the boundary of the parameter space. The penalty has been implemented in R and code is available from the online Supplementary Material for the paper.

Vaida and Blanchard (2005) and Greven and Kneib (2010) also consider fitting the linear mixed model with $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$ using the REML estimator $\widehat{\boldsymbol{\tau}}_R$ and then $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\tau}}_R)$ and $\widehat{\mathbf{v}}_R = \widehat{\mathbf{v}}(\widehat{\boldsymbol{\tau}}_R)$ with $\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ and $\widehat{\mathbf{v}}(\boldsymbol{\tau})$ defined in (9) and (14), respectively. Note that they use the same conditional likelihood as in the definition of *cAIC* (16) evaluated at the $\widehat{\boldsymbol{\beta}}_R$ and $\widehat{\mathbf{v}}_R$ but with a different penalty. Thus, the criteria are of the form

$$cAIC_{R,VB}(\widehat{\boldsymbol{\theta}}_R) = -2\ell(\widehat{\boldsymbol{\theta}}_R|\widehat{\mathbf{v}}_R) + \alpha_n(\widehat{\boldsymbol{\theta}}_R). \tag{18}$$

Vaida and Blanchard (2005) propose the penalty

$$\alpha_{n,R,VB}(\widehat{\boldsymbol{\theta}}_R) = \frac{2(n-p-1)}{n-p-2}\biggl\{\rho(\widehat{\boldsymbol{\tau}}_R) + 1 + \frac{p+1}{n-p-1}\biggr\}.$$

This penalty is asymptotic to $2\{\rho(\widehat{\boldsymbol{\tau}}_R) + 1\}$, as $n \to \infty$ with p, q fixed, which is like their penalty for the maximum likelihood estimator. Greven and Kneib (2010) also derive an analytic representation $\widehat{\nu}_R(\sigma^2)$ for the Liang, Wu and Zou (2008) penalty for the REML estimator. It turns out that $\widehat{\nu}_R(\sigma^2)$ is different from $\widehat{\nu}(\sigma^2)$

for the maximum likelihood estimator. The penalty in this case is therefore $\alpha_{n,R,GK}(\widehat{\boldsymbol{\theta}}_R) = 2\{\widehat{v}_R(\sigma^2) + 1\}$.

Srivastava and Kubokawa (2010) derive other conditional criteria by changing the estimators of the parameters at which minus twice the conditional log-likelihood is evaluated and then adjusting the penalty appropriately. For the model with $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$ and $\boldsymbol{\Psi}_*$ known, Srivastava and Kubokawa propose replacing the maximum likelihood estimator $\widehat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\widehat{\mathbf{V}}_*^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n$ of σ^2 by the estimator

$$\widetilde{\sigma}_0^2 = \{\mathbf{y} - (\mathbf{X}, \mathbf{Z})\widetilde{\xi}\}^T\{\mathbf{y} - (\mathbf{X}, \mathbf{Z})\widetilde{\xi}\}/n,$$

where $\widetilde{\xi} = \{(\mathbf{X}, \mathbf{Z})^T(\mathbf{X}, \mathbf{Z})\}^+(\mathbf{X}, \mathbf{Z})^T\mathbf{y}$ and $\mathbf{A}^+$ is the Moore–Penrose inverse of $\mathbf{A}$. This change in the variance estimator involves treating $\mathbf{u}$ as an unknown, fixed parameter which is to be estimated, here by ordinary least squares. The idea of treating $\mathbf{u}$ in this way is used by Jiang and Rao (2003) (Section 3.4). The use of $\widetilde{\sigma}_0^2$ changes the form of the penalty. For any estimators $\widehat{\boldsymbol{\beta}}_C$ and $\widehat{\mathbf{v}}_C$ satisfying $(\widehat{\boldsymbol{\beta}}_C^T, \widehat{\mathbf{v}}_C^T)^T = \mathbf{C}(\boldsymbol{\tau})\mathbf{y}$, they obtain the modified conditional criterion

$$cAIC_{SK} = -2\ell(\widehat{\boldsymbol{\beta}}_C, \widetilde{\sigma}_0^2|\widehat{\mathbf{v}}_C) + \frac{2n[\operatorname{trace}\{(\mathbf{X}, \mathbf{Z})\mathbf{C}(\widehat{\boldsymbol{\tau}})\} + 1]}{n - \operatorname{rank}\{(\mathbf{X}, \mathbf{Z})\} - 2}. \tag{19}$$

Note that here the parameters $\boldsymbol{\gamma}$ are absorbed into $\widehat{\mathbf{v}}_C$ so do not appear separately in the conditional log-likelihood. When $\widehat{\boldsymbol{\beta}}_C$ is either the maximum likelihood or the least squares estimator and $\widehat{\mathbf{v}}_C$ is the BLUP $\widehat{\mathbf{v}}$, $\operatorname{trace}\{(\mathbf{X}, \mathbf{Z})\mathbf{C}(\boldsymbol{\tau})\} = \rho(\boldsymbol{\tau})$; when $\widehat{\boldsymbol{\beta}}_C$ and $\widehat{\mathbf{v}}_C$ are the least squares estimators extracted from $\widetilde{\xi}$, $\operatorname{trace}\{(\mathbf{X}, \mathbf{Z})\mathbf{C}(\boldsymbol{\tau})\} = \operatorname{rank}(\mathbf{X}, \mathbf{Z})$. In the first case, the penalty is the asymptotic version of the Vaida–Blanchard penalty (17) with p replaced by the larger number $\operatorname{rank}\{(\mathbf{X}, \mathbf{Z})\}$ so the Srivastava–Kubokawa penalty is larger than the asymptotic Vaida–Blanchard penalty. When $\boldsymbol{\Psi}_*$ is unknown, for other estimators which use $\boldsymbol{\Psi}_*$, Srivastava and Kubokawa (2010) propose replacing it by an estimator $\widehat{\boldsymbol{\Psi}}_*$. For computational reasons, they consider using the truncated method of moments estimators for the special cases, but any consistent estimator can be used.

For the linear mixed model with a general $\boldsymbol{\Sigma}$, Kubokawa (2011) considers estimators $\widehat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$ which satisfy the second order expansion (13). Let $\mathbf{d}\{f(\boldsymbol{\tau})\} = (\partial f(\boldsymbol{\tau})/\partial y_j)$ denote the n-vector of derivatives of f with respect to $\mathbf{y}$ and $\mathbf{D}\{f(\boldsymbol{\tau})\} = (\partial^2 f(\boldsymbol{\tau})/\partial y_j\,\partial y_k)$ denote the $n \times n$ matrix of second derivatives of f with respect to $\mathbf{y}$. Then, under the condition that the three terms $\mathrm{E}(\operatorname{trace}[\mathbf{D}\{t_{2i}(\boldsymbol{\tau})\}])$, $\mathrm{E}(\operatorname{trace}[\mathbf{D}\{t_{1i}(\boldsymbol{\tau})\}t_{1j}(\boldsymbol{\tau})])$ and $\mathrm{E}(\operatorname{trace}[\mathbf{d}\{t_{1i}(\boldsymbol{\tau})\}\mathbf{d}\{t_{1j}(\boldsymbol{\tau})\}^T])$ are all $O(n^{-1})$, Kubokawa (2011) derives the penalty

$$\alpha_{n,K}(\widehat{\boldsymbol{\theta}}) = 2\{\rho(\widehat{\boldsymbol{\tau}}) + \widehat{h}_c(\widehat{\boldsymbol{\tau}})\},$$

where $\widehat{h}_c(\boldsymbol{\tau})$ is an estimator of

$$\begin{aligned} h_c(\boldsymbol{\tau}) = {} & -\frac{1}{2}\sum_{i=1}^{q}\operatorname{trace}\left\{\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\tau_i} - 2\boldsymbol{\Sigma}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\tau_i}\right)\right. \\ & \left.\cdot\,\mathrm{E}[\mathbf{D}\{t_{1i}(\boldsymbol{\tau})\}]\right\} \\ & - \sum_{i=1}^{q}\operatorname{trace}\left\{\frac{\partial\boldsymbol{\Sigma}}{\partial\tau_i}(\boldsymbol{\Sigma}^{-1} - \mathbf{V}^{-1})\right\}\mathrm{E}\{t_{2i}(\boldsymbol{\tau})\} \\ & - \sum_{i=1}^{q}\sum_{j=1}^{q}\operatorname{trace}\left\{\frac{1}{2}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\tau_i\,\partial\tau_j}(\boldsymbol{\Sigma}^{-1} - \mathbf{V}^{-1})\right. \\ & \left. + \frac{\partial\boldsymbol{\Sigma}}{\partial\tau_i}\left(\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\tau_j} - \frac{\partial\mathbf{V}^{-1}}{\partial\tau_j}\right)\right\} \\ & \cdot\,\mathrm{E}\{t_{1i}(\boldsymbol{\tau})t_{1j}(\boldsymbol{\tau})\}, \end{aligned}$$

which is obtained by replacing all the unknown quantities by estimators. It is a considerable task to derive the second order expansion (13) and then to derive the expressions which are needed to compute $h_c(\boldsymbol{\tau})$, but Kubokawa (2011) provides results for the maximum likelihood and REML estimators. These are still quite complicated for general use so Kubokawa (2011) specializes the expressions further to three particular models, namely, the variance component model (4), the random intercept regression model and the Fay–Herriot model.

3.3 BIC and Schwarz Criteria

The simplest and most widely used BIC for the linear mixed model (1) or the transformation model (3) is obtained by taking the marginal AIC (10) and replacing the constant 2 in the penalty by $\log(n)$ to obtain

$$BIC = -2\ell(\widehat{\boldsymbol{\theta}}) + \log(n)(p + q).$$

This is the definition used by `lme()` in R and by SAS `Proc Mixed`. This definition ensures that *BIC* bears the same relationship to *mAIC* for model (1) as BIC bears to AIC in regression and so should inherit some of its properties. Specifically, the increased weight in the penalty should encourage *BIC* to select smaller models than *mAIC*. Obviously, other *mAIC* can be converted to *BIC* in the same way by multiplying the *mAIC* penalty by $\log(n)/2$.

A more sophisticated approach is possible if we re-examine the relationship between BIC and the Bayes

factor. After reordering if necessary, partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_0^T, \boldsymbol{\theta}_1^T)^T$ into $\boldsymbol{\theta}_0 \in R^{r_0}$, $\boldsymbol{\theta}_1 \in R^{r_1}$, $p+q = r_0 + r_1$, and consider comparing the model $M_0 : \boldsymbol{\theta}_0 = \boldsymbol{\theta}_{00}$ with $M_1 : \boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_{00}$. Let h_0 be the prior density for $\boldsymbol{\theta}_1$ under M_0 and let h_1 be the prior density for $\boldsymbol{\theta}$ under M_1. Then the Bayes factor for comparing M_0 to M_1 is the ratio of the posterior odds to the prior odds for a model

$$\frac{\Pr(M_0|\mathbf{y})/\Pr(M_1|\mathbf{y})}{\Pr(M_0)/\Pr(M_1)} = \frac{\int g(\mathbf{y}|\boldsymbol{\theta}_{00}, \boldsymbol{\theta}_1) h_0(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}{\int g(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) h_1(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_0\, d\boldsymbol{\theta}_1}, \tag{20}$$

where $g(\mathbf{y}|\boldsymbol{\theta}) = \exp\{\ell(\boldsymbol{\theta})\}$ is the marginal likelihood of the model. If we hold M_0 constant at say the simplest model under consideration, this leads to choosing the model that minimizes $-2\log\{\int g(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \cdot h_1(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_0\, d\boldsymbol{\theta}_1\}$. BIC can be obtained using Laplace's method to approximate the integral in this expression.

Pauler (1998) uses this approach to derive a Schwarz criterion to select the regression parameter $\boldsymbol{\beta}$ in the independent cluster model. Partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$ into $\boldsymbol{\beta}_0 \in R^{p_0}$, $\boldsymbol{\beta}_1 \in R^{p_1}$, $p = p_0 + p_1$, and consider testing the null hypothesis $M_0 : \boldsymbol{\beta}_0 = \boldsymbol{\beta}_{00}$ against $M_1 : \boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_{00}$. Pauler required $\boldsymbol{\beta}_0$ to be null orthogonal to $(\boldsymbol{\beta}_1^T, \boldsymbol{\tau}^T)^T$ and, if the prior density for M_0 is $h_0(\boldsymbol{\beta}_1, \boldsymbol{\tau})$, the prior for M_1 to be of the form $h_1(\boldsymbol{\beta}, \boldsymbol{\tau}) = h_0(\boldsymbol{\beta}_1, \boldsymbol{\tau}) h(\boldsymbol{\beta}_0|\boldsymbol{\beta}_1, \boldsymbol{\tau})$. She notes that if $\mathbf{u}$ and $\mathbf{e}$ are Gaussianly distributed, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are orthogonal (the information matrix is block diagonal) and that $\boldsymbol{\beta}_1$ can be made null orthogonal to $\boldsymbol{\beta}_0$ by transforming $\boldsymbol{\beta}_1 \to \boldsymbol{\beta}_1 + (\mathbf{X}_1^T \mathbf{V}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}_1 \mathbf{V}^{-1} \mathbf{X}_0 \boldsymbol{\beta}_0$, where $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$ is partitioned conformably with $\boldsymbol{\beta}$. Then, using Laplace's method, she approximates the Bayes factor for comparing M_0 to M_1 by

$$\begin{aligned} S = {} & \ell\{\widehat{\boldsymbol{\theta}}(\boldsymbol{\beta}_{00})\} - \ell(\widehat{\boldsymbol{\theta}}) - \tfrac{1}{2} p_0 \log(2\pi) \\ & + \tfrac{1}{2} \log|\mathbf{X}_0^T \widehat{\mathbf{V}}^{-1} \mathbf{X}_0| - \log\{h(\widehat{\boldsymbol{\beta}}_0|\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\tau}})\}, \end{aligned} \tag{21}$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}(\boldsymbol{\beta}_{00})$ maximizes the log-likelihood under M_0. The Schwarz criterion can be made to look more familiar by dividing the $p_0 \times p_0$ matrix $\mathbf{X}_0^T \widehat{\mathbf{V}}^{-1} \mathbf{X}_0$ by n so that after taking the determinant we obtain the additional term $\frac{1}{2} p_0 \log(n)$, and then writing $p_0 = p - p_1$.

The Schwarz criterion (21) depends on the prior so, for cases when informative priors are not available, it is useful to consider using reference priors. Pauler presents Schwarz criteria using unit-information Gaussian and Cauchy reference priors. These criteria depend on what she calls the effective sample size. Write $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_0})^T$ and $\mathbf{X}_{i0}\boldsymbol{\beta}_0 = \mathbf{X}_i^{(1)}\beta_{01} + \cdots + \mathbf{X}_i^{(p_0)}\beta_{0p_0}$. Then a fixed effect parameter β_{0k} has an *associated random effect* if its covariate vector $\mathbf{X}_i^{(k)}$ is proportional to a column of $\mathbf{Z}_i$ for $i = 1, \ldots, p_0$. The *effective sample size* for β_{0k} is $E_k = m$ if β_{0k} has an associated random effect and $E_k = n$ otherwise. For the Gaussian prior

$$S_G = \ell\{\widehat{\boldsymbol{\theta}}(\boldsymbol{\beta}_{00})\} - \ell(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}\sum_{k=1}^{p_0} \log(E_k),$$

and for the Cauchy prior

$$S_C = S_G + \log(\pi^{1/2}/[2^{p_0/2}\Gamma\{(p_0+1)/2\}]).$$

The effective sample size concept seems reasonable but it is important to keep in mind that it is a result of the choice of prior which is arbitrary and is not intrinsic to the problem. For example, for the Gaussian prior, the variance is taken to be $\Delta^{1/2}(\mathbf{X}_0^T \widehat{\mathbf{V}}^{-1} \mathbf{X}_0)^{-1} \Delta^{1/2}$, where $\Delta = \operatorname{diag}(E_1, \ldots, E_{p_0})$. The log determinant of the variance is $-(1/2)\log|\Delta| + (1/2)\log|\mathbf{X}_0^T \widehat{\mathbf{V}}^{-1} \mathbf{X}_0|$ so, with this prior variance, the log determinant term in (21) is replaced by $(1/2)\log|\Delta| = (1/2) \cdot \sum_{k=1}^{p_0} \log(E_k)$. Other choices of Δ would therefore lead to other criteria.

To explore the effective sample size concept further, consider the random intercept model. Then $\mathbf{Z} = \text{block diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m})$ so any fixed effect that is constant within clusters (i.e., a cluster level covariate) has an associated random effect and any fixed effect that varies within clusters does not. Suppose we have a cluster level covariates. Then $\sum_{k=1}^{p_0} \log(E_k) = (p_0 - a)\log(n) + a\log(m) = p_0 \log(n) + a\log(m/n)$ and this reduces to $p_0 \log(n)$ if we have no cluster level covariate. Thus, if there is no cluster level covariate, the Gaussian version of the Schwarz criterion is the difference divided by -2 of two familiar terms of the form

$$BIC_G = -2\ell(\widehat{\boldsymbol{\theta}}) + \log(n)\, p.$$

The advantage of using S_G rather than BIC_G is that it can be applied to more general cluster models, but it has the disadvantage of requiring us to compare pairs of explicit hypotheses. When using the Schwarz criteria, it is a good idea to hold one of the hypotheses fixed to simplify comparison (and computation); in the example given in her paper, Pauler compares different models of interest to the null model with only an intercept.

Jones (2011) proposes using BIC with an alternative measure of the effective sample size. In the linear regression model, the coefficient of the intercept in the

normal equations for the least squares estimator is n; in the linear mixed model, the coefficient is $\mathbf{1}_n^T\mathbf{V}^{-1}\mathbf{1}_n$. Jones suggests that this coefficient be used as a measure of sample size but, since it depends on the units of measurement, $\mathbf{V}^{-1}$ be replaced by the correlation matrix. If $\mathbf{U}$ is the diagonal matrix with diagonal equal to the square root of the terms on the diagonal of the $\mathbf{V}$, the correlation matrix $\mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}$ is invariant to linear transformations of $\mathbf{y}$. Jones' measure of effective sample size is then $\mathbf{1}_n^T\mathbf{U}\mathbf{V}^{-1}\mathbf{U}\mathbf{1}_n$. Jones gives expressions for some particular cases, noting that when $\mathbf{V} = \sigma^2\mathbf{I}$, the effective sample size reduces to n, for the random intercept regression model $\sum_{i=1}^m n_i(\gamma^2+\sigma^2)/(n_i\gamma^2+\sigma^2)$, where $\gamma^2 = \mathrm{Var}(u_i)$, and the longitudinal autoregressive model $\sum_{i=1}^m\{1+(n_i-1)(1-\phi)/(1+\phi)\}$. Both measures lie between m and n, attaining these bounding values as $\gamma^2/(\gamma^2+\sigma^2)\to 1$ or $\phi\to 1$ (perfect correlation) and when $\gamma^2 = 0$ or $\phi = 0$ (zero correlation), respectively. In general, estimating the parameters in $\mathbf{U}$ and $\mathbf{V}$ leads to the criterion

$$BIC_J = -2\ell(\widehat{\boldsymbol{\theta}}) + \log\{\mathbf{1}_n^T\widehat{\mathbf{U}}\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{U}}\mathbf{1}_n\}(p+q).$$

We can compute a Bayes factor for comparing models with different variance parameters but it is then difficult to obtain simple approximations (like those given by Pauler, 1998) to the Bayes factor. In particular, it is difficult to make subsets of the parameters in $\boldsymbol{\tau}$ null orthogonal and the boundary issues need to be taken into account. Pauler, Wakefield and Kass (1999) and Saville, Herring and Kaufman (2011) ignore null orthogonality but do acknowledge and try to deal with the boundary issues.

Pauler, Wakefield and Kass (1999) approach the boundary issues in the variance component model by assuming that the parameter space $\boldsymbol{\Theta}$ can be expanded to an open set $\boldsymbol{\Theta}^o$ containing $\boldsymbol{\Theta}$ so that the boundary of $\boldsymbol{\Theta}$ is interior to $\boldsymbol{\Theta}^o$, applying the Laplace approximation on $\boldsymbol{\Theta}^o$ and then restricting it to $\boldsymbol{\Theta}$. For selecting the variance parameters $\boldsymbol{\tau}$, partition $\boldsymbol{\tau} = (\boldsymbol{\tau}_0^T, \boldsymbol{\tau}_1^T)^T$ into $\boldsymbol{\tau}_0 \in R^{q_0}$, $\boldsymbol{\tau}_1 \in R^{q_1}$, $q = q_0+q_1$, and consider testing the null hypothesis $M_0 : \boldsymbol{\tau}_0 = \mathbf{0}$ against $M_1 : \boldsymbol{\tau}_0 \neq \mathbf{0}$. Using Laplace's method, Pauler, Wakefield and Kass (1999) propose the approximation

$$\begin{aligned} S = {} & \ell\{\widehat{\boldsymbol{\theta}}^o(0)\} - \ell(\widehat{\boldsymbol{\theta}}^o) - \tfrac{1}{2}q_0\log(2\pi) \\ & + \tfrac{1}{2}\log|\mathbf{K}_{\boldsymbol{\tau}_0|\boldsymbol{\tau}_1}(\widehat{\boldsymbol{\theta}}^o)| - \log\{h(\widehat{\boldsymbol{\tau}}_0|\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\tau}}_1)\} \\ & + \log\{C_0^o/C_1^o\}, \end{aligned}$$

where $\widehat{\boldsymbol{\theta}}^o(0)$ maximizes the likelihood on $\boldsymbol{\Theta}^o$ under M_0, $\widehat{\boldsymbol{\theta}}^o$ maximizes the likelihood on $\boldsymbol{\Theta}^o$, $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (i.e., maximizes the likelihood on $\boldsymbol{\Theta}$), $\mathbf{K}_{\boldsymbol{\tau}_0|\boldsymbol{\tau}_1}(\boldsymbol{\theta}) = \mathbf{K}_{\boldsymbol{\tau}_0\boldsymbol{\tau}_0}(\boldsymbol{\theta}) - \mathbf{K}_{\boldsymbol{\tau}_0\boldsymbol{\tau}_1}(\boldsymbol{\theta})\cdot \mathbf{K}_{\boldsymbol{\tau}_1\boldsymbol{\tau}_1}(\boldsymbol{\theta})^{-1}\mathbf{K}_{\boldsymbol{\tau}_1\boldsymbol{\tau}_0}(\boldsymbol{\theta})$ is computed from the appropriate submatrices of the inverse of the observed information matrix $\mathbf{K}(\boldsymbol{\theta}) = -\ell''(\boldsymbol{\theta})^{-1}$, $h(\boldsymbol{\tau}_0|\boldsymbol{\beta},\boldsymbol{\tau}_1)$ is the conditional prior density under M_1 for $\boldsymbol{\tau}_0$ given $\boldsymbol{\beta}$ and $\boldsymbol{\tau}_1$, $C_0^o = \Pr[\mathcal{N}\{\widehat{\boldsymbol{\tau}}_1^o, \mathbf{K}_{0\boldsymbol{\tau}_1\boldsymbol{\tau}_1}(\widehat{\boldsymbol{\theta}}^o)\} \in \boldsymbol{\Theta}]$ with $\mathbf{K}_{0\boldsymbol{\tau}_1\boldsymbol{\tau}_1}$ the submatrix of the inverse observed information matrix under M_0 for $\boldsymbol{\tau}_1$, and $C_1^o = \Pr[\mathcal{N}\{\widehat{\boldsymbol{\tau}}^o, \mathbf{K}_{\boldsymbol{\tau}\boldsymbol{\tau}}(\widehat{\boldsymbol{\theta}}^o)\} \in \boldsymbol{\Theta}]$ with $\mathbf{K}_{\boldsymbol{\tau}\boldsymbol{\tau}}$ the submatrix of $\mathbf{K}$ for $\boldsymbol{\tau}$. The quantities C_0^o and C_1^o are of the same form as normalizing constants for truncated multivariate Gaussian densities. Pauler, Wakefield and Kass (1999) propose using a truncated Gaussian reference prior which leads to

$$\begin{aligned} S_{TG} = {} & \ell\{\widehat{\boldsymbol{\theta}}^o(0)\} - \ell(\widehat{\boldsymbol{\theta}}^o) + \tfrac{1}{2}q_0\log(n) \\ & + \log\{C_{TG}^o C_0^o/C_1^o\}, \end{aligned}$$

where $C_{TG}^o = \Pr[\mathcal{N}\{\mathbf{0}, n\mathbf{K}_{\boldsymbol{\tau}_0|\boldsymbol{\tau}_1}(\widehat{\boldsymbol{\theta}}^o)^{-1}\} \in \boldsymbol{\Theta}]$ is the normalizing constant for the prior density. Aside from the final boundary correction term, this is similar to the usual Schwarz criterion. Under regularity conditions, the boundary correction term is of smaller order than $\log(n)$ so, as Pauler, Wakefield and Kass (1999) note, the usual criterion can be used to select variance parameters. In contrast to Pauler (1998), Pauler, Wakefield and Kass (1999) do not attempt to make an adjustment for effective sample size.

Saville, Herring and Kaufman (2011), following on from Saville and Herring (2009), take a different approach to the boundary issue. They parametrize the linear mixed model (1) using the alternative Cholesky factorization (2) so $\boldsymbol{\Gamma} = \sigma\boldsymbol{\Phi}\boldsymbol{\Gamma}^\dagger$, where $\boldsymbol{\Gamma}^\dagger$ is a lower triangular matrix with ones on the diagonal and $\boldsymbol{\Phi} = \mathrm{diag}\{\exp(\phi_1),\ldots,\exp(\phi_s)\}$. The matrix $\boldsymbol{\Phi}$ is $\mathbf{D}^\dagger$ from (2) on the logarithmic scale. Let $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_s)^T$ and let $\boldsymbol{\gamma}^\dagger$ be the vector of free parameters in $\boldsymbol{\gamma}^\dagger$. They assume that σ^{-2} has a gamma distribution and then integrate both $\mathbf{u}$ and σ^2 from the density of $\mathbf{y}$ given $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, $\boldsymbol{\gamma}^\dagger$, $\mathbf{u}$ and σ^2 to obtain the density of $\mathbf{y}$ given $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}^\dagger$ which is a multivariate t density. They then recommend adopting weakly informative priors for the parameters and use Laplace approximations to approximate the Bayes factor for comparing M_0 to M_1. They argue that the parameters in the multivariate t density do not have boundary constraints, but in fact the boundary has been moved from zero to negative infinity and this is not necessarily more convenient for computation.

3.4 Other Criteria

There are a number of criteria of a more or less arbitrary nature which have been proposed for model selection. We describe some of these in this section.

For the linear mixed model (1) or the transformation model (3), Pu and Niu (2006) suggest a Generalized Information Criterion of the form

$$GIC_{\kappa_n} = -2\ell(\widehat{\boldsymbol{\theta}}) + \kappa_n(p+q).$$

This criterion combines both marginal AIC ($\kappa_n = 2$) and BIC ($\kappa_n = \log(n)$) and allows greater flexibility in the choice of κ_n. For example, it includes the Hannan–Quinn (1979) penalty $\kappa_n = 2\log\log(n)$ and the Bozdogan (1987) penalty $\kappa_n = \log(n) + 1$, both of which are available in SAS `Proc Mixed`. Pu and Niu also apply GIC with $\kappa_n = n^{1/2}$. For any choice of κ_n, Pu and Niu suggest implementing GIC in two stages (first fix $\boldsymbol{\tau}$ and select the model for $\boldsymbol{\beta}$ and then fix $\boldsymbol{\beta}$ and select the model for $\boldsymbol{\tau}$), but it is also possible to implement it directly. Pu and Niu explore the asymptotic properties of the procedure for selecting regression terms but not for selecting variance parameters.

The idea of treating $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ separately and differently is taken up by Jiang and Rao (2003). For any vector $\mathbf{a}$, let $\|\mathbf{a}\|^2 = \mathbf{a}^T\mathbf{a}$. Then Jiang and Rao propose selecting the regression parameter $\boldsymbol{\beta}$ using

$$\|\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\}\mathbf{y}\|^2 + a_n p,$$

where a_n is a real, positive sequence satisfying some asymptotic conditions and $\mathbf{A}^-$ is a generalized inverse of $\mathbf{A}$. Other than through the conditions on a_n, this criterion does not depend on $\boldsymbol{\tau}$ so this selection can be carried out separately. For the variance component model, partition the set of matrices $\{\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(q_\gamma)}\}$ into sets L_k of matrices which (together with $\mathbf{X}$) span the same linear space so that the matrices in L_1 have higher rank than those in L_2 and so on. Jiang and Rao give the example of a 3-factor crossed design where L_1 contains the 3-way interaction, L_2 the 2-way interaction and L_3 the main effects. Jiang and Rao suggest selecting the variance parameters $\boldsymbol{\tau}$ sequentially, starting in L_1 and progressing through the remaining sets of matrices. Let $\mathbf{B} = (\mathbf{X}, \mathbf{Z})$ and $\mathbf{B}_{-j}$ be $\mathbf{B}$ omitting $\mathbf{Z}^{(j)}$, $j \in L_1$. Then they select from L_1, the set of indices j for which, for any $1 < b < 2$,

$$\begin{aligned}
&\frac{n - \operatorname{rank}(\mathbf{B})}{\operatorname{rank}(\mathbf{B}) - \operatorname{rank}(\mathbf{B}_{-j})} \\
&\qquad \cdot \frac{\|\{\mathbf{B}(\mathbf{B}^T\mathbf{B})^-\mathbf{B} - \mathbf{B}_{-j}(\mathbf{B}_{-j}^T\mathbf{B}_{-j})^-\mathbf{B}_{-j}\}\mathbf{y}\|^2}{\|\{\mathbf{I}_n - \mathbf{B}(\mathbf{B}^T\mathbf{B})^-\mathbf{B}\}\mathbf{y}\|^2} \\
&> 1 + \{n - \operatorname{rank}(\mathbf{B})\}^{(b/2)-1} \\
&\quad + \{\operatorname{rank}(\mathbf{B}) - \operatorname{rank}(\mathbf{B}_{-j})\}^{(b/2)-1}.
\end{aligned}$$

For the second group L_2, let l_2 denote a subset of indices in L_2. Let $\mathbf{B}_1(l_2) = (\mathbf{X}, \mathbf{Z}^{(j)}, j \in l_2 \cup L_3, \cup L_4, \ldots)$ be the matrix comprised of $\mathbf{X}$ and the $\mathbf{Z}^{(j)}$, for j from $l_2, L_3, L_4, \ldots$. Then choose $l_2 \in L_2$ to minimize

$$\|[\mathbf{I}_n - \mathbf{B}_1(l_2)\{\mathbf{B}_1(l_2)^T\mathbf{B}_1(l_2)\}^-\mathbf{B}_1(l_2)]\mathbf{y}\|^2 + a_{1n}\#(l_2),$$

where a_{1n} is a real, positive sequence satisfying some asymptotic conditions and $\#(l_2)$ is the number of parameters in l_2. Jiang and Rao consider the penalties $a_{1n} \in \{2, \log(n), n/\log(n)\}$. The procedure extends naturally to the remaining groups $L_3, L_4, \ldots$. Jiang and Rao give conditions under which the procedure is consistent.

Takeuchi (1976) proposes using as a measure of model complexity $\operatorname{trace}\{\mathbf{K}(\boldsymbol{\theta})\mathbf{L}(\boldsymbol{\theta})^{-1}\}$, where $\mathbf{K}(\boldsymbol{\theta}) = \operatorname{Var}\{\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}$ is the variance of the score function and $\mathbf{L}(\boldsymbol{\theta}) = -\operatorname{E}\{\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T\}$ is the expected information. As Burnham and Anderson [(2002), page 367] note, this complexity measure can be expressed as $\operatorname{trace}\{\mathbf{L}(\boldsymbol{\theta})\mathbf{L}(\boldsymbol{\theta})^{-1}\mathbf{K}(\boldsymbol{\theta})\mathbf{L}(\boldsymbol{\theta})^{-1}\}$, which is the trace of the inverse of the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ when the model holds multiplied by the (sandwich) variance of $\widehat{\boldsymbol{\theta}}$ when the model does not hold. If the model is correct, the measure reduces to $p + q$ and the Takeuchi Information Criterion

$$TIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2\operatorname{trace}\{\mathbf{K}(\boldsymbol{\theta})\mathbf{L}(\boldsymbol{\theta})^{-1}\}$$

is the same as *mAIC*. The Neural Information Criterion (NIC) of Murata, Yoshizawa and Amari (1994) measures complexity in a similar way but uses the regularized log-likelihood $\ell(\boldsymbol{\theta}) + \log\{h(\boldsymbol{\theta})\}$ in place of $\ell(\boldsymbol{\theta})$. Let $\mathbf{K}_h(\boldsymbol{\theta}) = \operatorname{Var}(\partial[\ell(\boldsymbol{\theta}) + \log\{h(\boldsymbol{\theta})\}]/\partial\boldsymbol{\theta})$ and $\mathbf{L}_h(\boldsymbol{\theta}) = -\operatorname{E}(\partial^2[\ell(\boldsymbol{\theta}) + \log\{h(\boldsymbol{\theta})\}]/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$. Then the complexity measure in NIC, called the *effective number of parameters* by Moody (1992), is $\operatorname{trace}\{\mathbf{K}_h(\boldsymbol{\theta}) \cdot \mathbf{L}_h(\boldsymbol{\theta})^{-1}\}$. Ripley [(1996), page 140] points out that the estimation of this measure is generally not straightforward.

The minimum description length approach (MDL) developed by Rissanen in the 1980s (see Rissanen, 2007) chooses the model that achieves maximum data compression by minimizing the code length of the data and the model. There are different coding schemes which lead to different MDL criteria. The most relevant for the linear mixed model is the two-stage code which leads to a penalized likelihood and is equivalent to *BIC*, the mixture scheme which produces a criterion that is related to a Bayes factor and the normalized maximum likelihood scheme. For a geostatistical model [the linear mixed model with $\boldsymbol{\Gamma} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \sigma^2\mathbf{R}(\boldsymbol{\delta})$, where

the parameters $\boldsymbol{\delta}$ describe the spatial correlation between observations, so $\boldsymbol{\tau} = (\boldsymbol{\delta}^T, \sigma^2)^T$ and $q = q_\delta + 1$], Hoeting et al. (2006) use the two-stage code and propose the minimum description length criterion $BIC/2$. Liski and Liski (2008) consider spline smoothing by fitting the random effect model with one variance component ($q_\gamma = 1$) and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$. They use the normalized maximum likelihood coding scheme to produce the conditional criterion

$$MDL = -\ell(\widehat{\boldsymbol{\theta}}|\widehat{\mathbf{u}}) + \log\left[\int f\{\mathbf{q}|\widehat{\mathbf{u}}(\mathbf{q}); \widehat{\boldsymbol{\theta}}(\mathbf{q})\}\, d\mathbf{q}\right],$$

where $f(\mathbf{y}|\mathbf{u}; \boldsymbol{\theta}) = \exp\{\ell(\boldsymbol{\theta}|\mathbf{u})\}$ is the conditional density of $\mathbf{y}|\mathbf{u}$. The penalty term, called the *parametric complexity* of the model, is difficult to compute because the conditional density is evaluated at the estimators before being integrated.

Kubokawa (2011) introduces some prediction criteria which are variants on Mallows C_p. Let $\widetilde{\boldsymbol{\tau}}$ be an estimator of $\boldsymbol{\tau}$ from the full model which satisfies a second order expansion like (13) $\widetilde{\boldsymbol{\tau}} - \boldsymbol{\tau} = \widetilde{\mathbf{t}}_1(\boldsymbol{\tau}) + \widetilde{\mathbf{t}}_2(\boldsymbol{\tau}) + O_p(n^{-3/2})$, where $\mathrm{E}\{\widetilde{\mathbf{t}}_1(\boldsymbol{\tau})\} = \mathrm{E}\{(\widetilde{t}_{11}(\boldsymbol{\tau}), \ldots, \widetilde{t}_{1q}(\boldsymbol{\tau}))^T\} = \mathbf{0}$, $\widetilde{\mathbf{t}}_1(\boldsymbol{\tau}) = O_p(n^{-1/2})$ and $\widetilde{\mathbf{t}}_2(\boldsymbol{\tau}) = (\widetilde{t}_{21}(\boldsymbol{\tau}), \ldots, \widetilde{t}_{2q}(\boldsymbol{\tau}))^T = O_p(n^{-1})$. (He also considers estimating $\boldsymbol{\tau}$ from the current candidate model but found that it performs poorly.) Then let $\widehat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ be the generalized least squares estimator of $\boldsymbol{\beta}$ defined in (9), $\widehat{\mathbf{u}}(\boldsymbol{\tau}) = \boldsymbol{\Gamma}^{-1}\widehat{\mathbf{v}}(\boldsymbol{\tau})$ be the BLUP of $\mathbf{u}$ with $\widehat{\mathbf{v}}(\boldsymbol{\tau})$ defined in (14), and let $\widetilde{\boldsymbol{\Psi}}$, $\widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\mathbf{V}}$ be estimators of $\boldsymbol{\Psi}$, $\boldsymbol{\Sigma}$ and $\mathbf{V}$ constructed using $\widetilde{\boldsymbol{\tau}}$. Kubokawa defines

$$\begin{aligned} mPEC &= \{\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\tau}})\}^T \widetilde{\mathbf{V}}^{-1}\{\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\tau}})\} \\ &\quad + 2\{p + q_m(\widetilde{\boldsymbol{\tau}})\}, \\ cPEC &= \{\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\tau}}) - \mathbf{Z}\widetilde{\boldsymbol{\Psi}}^{1/2}\widehat{\mathbf{u}}(\widetilde{\boldsymbol{\tau}})\}^T \\ &\quad \cdot \widetilde{\boldsymbol{\Sigma}}^{-1}\{\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\tau}}) - \mathbf{Z}\widetilde{\boldsymbol{\Psi}}^{1/2}\widehat{\mathbf{u}}(\widetilde{\boldsymbol{\tau}})\} \\ &\quad + 2\{\rho(\widetilde{\boldsymbol{\tau}}) + q_c(\widetilde{\boldsymbol{\tau}})\}, \end{aligned}$$

where

$$\begin{aligned} &q_m(\boldsymbol{\tau}) \\ &= \frac{1}{2}\sum_{i=1}^{q} \operatorname{trace}\left[\frac{\partial \mathbf{V}}{\partial \tau_i}\mathrm{E}\left\{\frac{\partial^2 \widetilde{t}_{1i}(\boldsymbol{\tau})}{\partial \mathbf{y}\, \partial \mathbf{y}^T}\right\}\right] \\ &\quad + \frac{1}{2}\sum_{i=1}^{q} \operatorname{trace}\left(\frac{\partial \mathbf{V}}{\partial \tau_i}\mathbf{V}^{-1}\right)\mathrm{E}\{\widetilde{t}_{2i}(\boldsymbol{\tau})\} \\ &\quad - \frac{1}{4}\sum_{i=1}^{q}\sum_{j=1}^{q} \operatorname{trace}\left(\frac{\partial^2 \mathbf{V}^{-1}}{\partial \tau_i\, \partial \tau_j}\mathbf{V}\right)\mathrm{E}\{\widetilde{t}_{1i}(\boldsymbol{\tau})\widetilde{t}_{1j}(\boldsymbol{\tau})\}, \end{aligned}$$

$$\begin{aligned} &q_c(\boldsymbol{\tau}) \\ &= -\frac{1}{2}\sum_{i=1}^{q} \operatorname{trace}\left[\mathbf{V}\frac{\partial(\mathbf{V}^{-1}\boldsymbol{\Sigma}\mathbf{V}^{-1})}{\partial \tau_i}\mathbf{V}\,\mathrm{E}\left\{\frac{\partial^2 \widetilde{t}_{1i}(\boldsymbol{\tau})}{\partial \mathbf{y}\, \partial \mathbf{y}^T}\right\}\right] \\ &\quad + \frac{1}{2}\sum_{i=1}^{q} \operatorname{trace}\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \tau_i}\mathbf{V}^{-1}\right]\mathrm{E}\{\widetilde{t}_{2i}(\boldsymbol{\tau})\} \\ &\quad - \sum_{i=1}^{q}\sum_{j=1}^{q} \operatorname{trace}\left\{\frac{1}{4}\boldsymbol{\Sigma}^{-1}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \tau_i\, \partial \tau_j}\boldsymbol{\Sigma}\mathbf{V}^{-1}\right. \\ &\qquad\qquad \left. - \frac{\partial \boldsymbol{\Sigma}}{\partial \tau_i}\boldsymbol{\Sigma}^{-1}\frac{\partial(\boldsymbol{\Sigma}\mathbf{V}^{-1})}{\partial \tau_j}\right\} \\ &\qquad \cdot \mathrm{E}\{\widetilde{t}_{1i}(\boldsymbol{\tau})\widetilde{t}_{1j}(\boldsymbol{\tau})\}, \end{aligned}$$

and $\rho(\boldsymbol{\tau})$ is the effective degrees of freedom. The computations are quite formidable.

Finally, Wu and Zhang (2002) and Fang (2011) consider using cross-validation to select linear mixed models. For the independent cluster model with $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$, the leave-one-cluster-out criterion is

$$\begin{aligned} &m^{-1}\sum_{i=1}^{m} n_i^{-1}\big(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{[i]}\big)^T\big(\mathbf{Z}_i\widehat{\boldsymbol{\Psi}}_*^{[i]}\mathbf{Z}_i^T + \mathbf{I}_{n_i}\big)^{-1} \\ &\quad \cdot \big(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{[i]}\big), \end{aligned}$$

where $\widehat{\boldsymbol{\beta}}^{[i]}$ and $\widehat{\boldsymbol{\Psi}}_*^{[i]}$ are the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}_* = \boldsymbol{\Psi}/\sigma^2$ using the data without cluster i; the leave-one-observation-out criterion is

$$n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\big(y_{ij} - \mathbf{x}_{ij}^T\widehat{\boldsymbol{\beta}}^{[ij]} - \mathbf{z}_{ij}^T\widehat{\mathbf{v}}_i^{[ij]}\big)^2,$$

where $\mathbf{x}_{ij}^T$ is the jth row of $\mathbf{X}_i$, $\mathbf{z}_{ij}^T$ is the jth row of $\mathbf{Z}_i$, and $\widehat{\boldsymbol{\beta}}^{[ij]}$ and $\widehat{\mathbf{v}}_i^{[ij]}$ are the maximum likelihood estimators and predictors of $\boldsymbol{\beta}$ and $\mathbf{v}_i$, respectively, using the data without observation j in cluster i. The leave-one-cluster-out criterion is a marginal criterion, whereas the leave-one-observation-out criterion is a conditional criterion. Fang (2011) shows that for $m \to \infty$ with $n_i = n_1$ fixed (or $\bar{n} \to n_1$) and $\boldsymbol{\Psi}_*$ known, (i) leave-one-cluster-out cross-validation and *mAIC* of Vaida and Blanchard (2005) are asymptotically equivalent, and (ii) leave-one-observation-out cross-validation and *cAIC* of Vaida and Blanchard (2005) are asymptotically equivalent. This extends the relationship between cross-validation and AIC in the linear regression model established by Stone (1977) to the linear mixed model.

4. SHRINKAGE METHODS

One issue with the direct application of the information criteria defined in Section 3 is that they generally involve comparing 2^{p+q} different models, which is not computationally feasible when p and/or q is large. Even when $p+q \ll n$ is fixed, it is still possible for $p+q$ to be large. Shrinkage methods such as the LASSO (Tibshirani, 1996) are popular for selecting models in the linear regression setting when p is of medium or large size due to its computational feasibility and statistical accuracy (e.g., Bühlmann and van de Geer, 2011, page 20). In this section we review the shrinkage approach to model selection in the linear mixed model case. We begin by discussing the linear regression case ($\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$ and $\boldsymbol{\Psi} = \mathbf{0}$), since many of the ideas in the mixed model case are motivated by this simpler case.

For the linear regression model, Tibshirani (1996) proposes the LASSO (least absolute shrinkage and selection operator) method for simultaneous model estimation and selection. It is usual to standardize the covariates $\mathbf{X}$ and sometimes also to center $\mathbf{y}$. The selected model minimizes

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^{p} \phi_{\lambda_j}(|\beta_j|), \tag{22}$$

with respect to $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, where

$$\phi_{\lambda_j}(|\beta|) = \frac{\lambda_j |\beta|}{2n} \quad \text{and} \quad \lambda_j = \lambda, \quad j = 1, 2, \ldots, p. \tag{23}$$

When the tuning parameter $\lambda > 0$ is large enough some of the parameters in $\boldsymbol{\beta}$ are shrunk to exactly zero and, hence, minimizing this criterion does model selection automatically. The minimization problem (22) with the LASSO penalty function (24) is a convex problem and there are efficient algorithms available to compute the solution. For example, the LARS algorithm in Efron et al. (2004) or the coordinate decent algorithms defined in Friedman et al. (2007) and Meier, van de Geer and Bühlmann (2008) can be applied.

There have been various further advances in penalized least squares approaches for model selection since Tibshirani's original paper (e.g., see Fan and Lv, 2010, pages 107–117, and Tibshirani, 2011, for brief reviews). One problem with the LASSO is that it tends to shrink large $\boldsymbol{\beta}$ coefficients too much, leading to bias issues (Fan and Li, 2001). As an alternative to (24), Fan and Li (2001) suggest the SCAD (smoothly clipped absolute deviation) penalty function defined by its derivative

$$\phi'_{\lambda_j}(|\beta|) = \lambda_j \left\{ I(|\beta| \le \lambda_j) + \frac{(a\lambda_j - |\beta|)_+}{(a-1)\lambda_j} I(|\beta| > \lambda_j) \right\} \quad \text{and} \quad \lambda_j = \lambda, \quad j = 1, 2, \ldots, p,$$

with $a = 3.7$. They propose an algorithm based on local quadratic approximations and, more recently, Zou and Li (2008) propose a local linear approximation, since the SCAD penalized loss function is difficult to minimize directly due to the singularities in the penalty function. Zou (2006) introduces the ALASSO (Adaptive LASSO) which also helps overcome the bias problems associated with the LASSO. The ALASSO penalty function is

$$\phi_{\lambda_j}(|\beta|) = \frac{\lambda_j |\beta|}{2n} \quad \text{and} \quad \lambda_j = \frac{\lambda}{|\widehat{\beta}_j|^{\iota}}, \quad j = 1, 2, \ldots, p, \tag{24}$$

where $\iota > 0$ is an additional parameter often taken to be equal to 1 and $\widehat{\boldsymbol{\beta}}$ is a $n^{1/2}$-consistent estimator of $\boldsymbol{\beta}$. Zou (2006) shows that the LARS algorithm can also be used to solve the ALASSO minimization problem.

We now consider the linear mixed model case and assume $\boldsymbol{\Psi}$ has a general form. Bondell, Krishna and Ghosh (2010), Ibrahim et al. (2011) and Peng and Lu (2012) are to date the only authors to consider truly joint selection of both $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ using a shrinkage approach in the fixed parameter dimension setting. Other authors apply shrinkage methods to select on $\boldsymbol{\beta}$ only, assuming that the variance structure is not subject to selection (e.g., Foster, Verbyla and Pitchford, 2007; Ni, Zhang and Zhang, 2010; Wang, Eskridge and Crossa, 2011). We therefore focus on the methodology in Bondell, Krishna and Ghosh (2010), Ibrahim et al. (2011) and Peng and Lu (2012). All three consider model selection for the independent cluster model (5) assuming $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$ and both $s_i = s_1$ and $\boldsymbol{\Psi}_i = \boldsymbol{\Psi}_1$ are the same across clusters. Both Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) use Cholesky parametrizations and we will assume that $\boldsymbol{\Gamma}_i$ is the Cholesky factor of $\boldsymbol{\Psi}_i$ for the rest of this section. Note that Ibrahim et al. (2011) consider the more general mixed effects model setting where $\mathbf{y}_i$ given $\mathbf{u}_i$ and $\mathbf{X}_i$ belong to the exponential family, but for comparative purposes we will restrict the discussion to the Gaussian case only.

Ibrahim et al. (2011) propose maximizing a penalized marginal log-likelihood

$$\ell(\boldsymbol{\theta}) - m\sum_{j=1}^{p}\phi_{\lambda_j}(|\beta_j|) - m\sum_{k=1}^{s_1}\phi_{\lambda_{p+k}}(\|\boldsymbol{\gamma}_k\|), \tag{25}$$

with respect to $\boldsymbol{\theta}$, where $\boldsymbol{\gamma}_k$ contains the nonzero elements in the kth row of $\boldsymbol{\Gamma}_i$ and $\ell(\boldsymbol{\theta})$ is defined in (8). Either the SCAD or ALASSO penalty functions are used in (25) and there are two tuning constants which are defined by

$$\lambda_j = \lambda^{(1)}, \quad j = 1, 2, \ldots, p \quad \text{and} \quad \lambda_{p+k} = \lambda^{(2)}\sqrt{k}, \quad k = 1, 2, \ldots, s_1.$$

The ALASSO penalty functions differ slightly from (25) and are defined as

$$\phi_{\lambda_j}(|\beta|) = \lambda_j\frac{|\beta|}{|\widehat{\beta}_j|}, \quad j = 1, 2, \ldots, p \quad \text{and}$$

$$\phi_{\lambda_{p+k}}(\|\boldsymbol{\gamma}_k\|) = \lambda_{p+k}\frac{\|\boldsymbol{\gamma}_k\|}{\|\widehat{\boldsymbol{\gamma}}_k\|}, \quad k = 1, 2, \ldots, s_1,$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}_k$ are the unpenalized maximum likelihood estimators. Notice that the parameters $\boldsymbol{\gamma}$ are selected in a grouped manner similar to the LASSO for grouped variables (Yuan and Lin, 2006), and this helps preserve the positive definite constraint in $\boldsymbol{\Psi}$.

Bondell, Krishna and Ghosh (2010) use the alternative Cholesky factor parametrization $\boldsymbol{\Gamma}_i = \sigma\mathbf{D}_i^{\dagger}\boldsymbol{\Gamma}_i^{\dagger}$, where $\mathbf{D}_i^{\dagger} = \operatorname{diag}(d_1, d_2, \ldots, d_{s_1})^T$ is a diagonal matrix and $\boldsymbol{\Gamma}_i^{\dagger}$, whose (l, r)th element is $\gamma_{lr}^{\dagger}$, is a $s_1 \times s_1$ lower triangular matrix with ones on the diagonal. Setting $d_l = 0$ is equivalent to setting all the elements in the lth column and lth row to zero and, hence, a single parameter controls the inclusion/exclusion of a group of random effects. Let $\mathbf{d} = (d_1, d_2, \ldots, d_{s_1})^T$, let $\boldsymbol{\gamma}^{\dagger}$ be the vector of free parameters in $\boldsymbol{\Gamma}_i^{\dagger}$ and define $\boldsymbol{\theta}^{\dagger} = (\boldsymbol{\beta}^T, \mathbf{d}^T, \boldsymbol{\gamma}^{\dagger T})^T$. Note that σ^2 is not included in $\boldsymbol{\theta}^{\dagger}$. Bondell, Krishna and Ghosh (2010) propose maximizing an ALASSO penalized log-likelihood

$$\ell(\boldsymbol{\theta}^{\dagger}) - \lambda^{(3)}\left(\sum_{j=1}^{p}\frac{|\beta_j|}{|\widehat{\beta}_j|} + \sum_{k=1}^{s_1}\frac{|d_k|}{|\widehat{d}_k|}\right)$$

with respect to $\boldsymbol{\theta}^{\dagger}$, where $\lambda^{(3)}$ is a single tuning constant. Here $\widehat{\boldsymbol{\beta}}_j$ are the unpenalized generalized least squares estimates and the $\widehat{d}_k$ is obtained from decomposing the unpenalized restricted maximum likelihood estimate of $\boldsymbol{\Psi}_i$.

The Cholesky decompositions prove to be very helpful in estimation. The conditional expectations of $\mathbf{y}$ given $\mathbf{u}$ can be rearranged to give

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\Gamma}\mathbf{u} = \begin{pmatrix}\mathbf{X} & (\mathbf{u}^T \otimes \mathbf{Z})\mathbf{J}_{ms_1}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}, \tag{26}$$

where $\operatorname{Vec}(\boldsymbol{\Gamma}) = \mathbf{J}_{ms_1}\boldsymbol{\gamma}$ [the matrix $\mathbf{J}_{ms_1}$ transforms $\boldsymbol{\gamma}$ to $\operatorname{Vec}(\boldsymbol{\Gamma})$], or

$$\begin{aligned}&\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\Gamma}\mathbf{u}\\ &\quad= \begin{pmatrix}\mathbf{X} & \mathbf{Z}\operatorname{diag}(\boldsymbol{\Gamma}^{\dagger}(\sigma\mathbf{u}))(\mathbf{1}_m \otimes \mathbf{I}_{s_1}))\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \mathbf{d}\end{pmatrix}.\end{aligned} \tag{27}$$

The conditional expectations can therefore be written in a form which is linear in the parameters that are subject to selection. Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) both adapt the EM algorithm to estimate the parameters. They treat $\mathbf{u}$ as unobserved in the E-step and the M-step involves maximizing a penalized objective function. To incorporate grouped penalization, Ibrahim et al. (2011) use a modification of the local linear approximation algorithm proposed by Zou and Li (2008). Bondell, Krishna and Ghosh (2010) in their M-Step apply a standard quadratic programming technique. The EM penalized maximum likelihood estimators above are obtained first by assuming $\mathbf{u}$ is known, then $\boldsymbol{\Gamma}$ is estimated and then $\mathbf{u}$ estimated. This process differs subtly from the information criteria approaches in Section 3, where a different order is used when deriving the criteria there: first it is assumed that $\boldsymbol{\Gamma}$ is known, then $\mathbf{u}$ is estimated and then $\boldsymbol{\Gamma}$ is estimated.

Although the approaches of Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) share some elements in common, there are some differences between them which are important to highlight. Bondell, Krishna and Ghosh (2010) incorporate a single tuning constant which is the same for penalizing both $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, whereas Ibrahim et al. (2011) have a more flexible approach with two different tuning constants. Bondell, Krishna and Ghosh (2010) use $\mathbf{d}$ rather than $\boldsymbol{\gamma}$ in model selection and they effectively treat $\boldsymbol{\gamma}^{\dagger}$ like nuisance parameters since they do not appear in the penalty. Neither Bondell, Krishna and Ghosh (2010) nor Ibrahim et al. (2011) incorporate $\widehat{\mathbf{u}}$ into the penalized likelihood criterion and their methods are therefore more in line with the marginal information criteria of Section 3.1, rather than the conditional approach of Section 3.2.

One open issue with both Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) is that the Cholesky decompositions are dependent on the order

in which the random effects appear and are not permutation invariant (Pourahmadi, 2011). This means in the finite sample case that different model selections result from using different orders in the columns of $\mathbf{Z}_i$. We confirmed this by running the first simulation example in Bondell, Krishna and Ghosh (2010) with different orders in the columns of $\mathbf{Z}_i$. Note also that setting d_k and $\boldsymbol{\gamma}_k$ to zero is not equivalent to setting the kth diagonal element in $\boldsymbol{\Psi}_i$ to zero, which for the independent cluster model (5) is the more natural selection problem. Another issue is that both Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) use the unpenalized maximum likelihood or restricted maximum likelihood estimates as the weights in the ALASSO penalty, but in practice unpenalized maximum likelihood algorithms often fail to converge when the underlying $\boldsymbol{\tau}$ is sparse and/or p is large (e.g., Nguyen and Jiang, 2012, page 310; Jiang, Luan and Wang, 2007, page 2252). Also, some of the maximum likelihood estimates of variance parameters could be exactly on the zero boundary, implying that the ALASSO weight is infinity. Note that boundary problems do not occur in the regression case since only $\boldsymbol{\beta}$ is penalized.

Peng and Lu (2012) also apply a shrinkage method, although their approach is quite different from Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011). Instead of doing selection on $\boldsymbol{\Psi}_1$ directly, Peng and Lu (2012) select the random effects by penalizing $\mathbf{v} = \boldsymbol{\Gamma}\mathbf{u}$. Write $\boldsymbol{\Psi}_1 = \sigma^2\boldsymbol{\Psi}_1^\dagger$ and then, motivated by an asymptotic expansion, estimate $\boldsymbol{\Psi}_1^\dagger$ by

$$\widehat{\boldsymbol{\Psi}}_1^\dagger = \frac{\sum_{i=1}^m \mathbf{v}_i\mathbf{v}_i^T}{m\widehat{\sigma}^2} - \frac{\sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}}{m}. \tag{28}$$

To estimate and select the model, Peng and Lu (2012) define the following simple iterative procedure which penalizes both $\boldsymbol{\beta}$ and $\mathbf{v}$:

(1) For each i update $\mathbf{v}_i$ given $\boldsymbol{\beta}$ by minimizing with respect to $\mathbf{v}_i$ the penalized least squares criterion

$$(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{v}_i)^T(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{v}_i) + 2n\sum_{k=1}^{s_1} \phi_{\lambda^{(4)}}\Big(\sqrt{|\widehat{\boldsymbol{\psi}}_{kk}^\dagger|}\Big),$$

where $\widehat{\boldsymbol{\psi}}_{kk}^\dagger$ is the kth diagonal element of $\widehat{\boldsymbol{\Psi}}_1^\dagger$. Then update $\boldsymbol{\Psi}_1^\dagger$ using (28).

(2) Update $\boldsymbol{\beta}$ given $\boldsymbol{\Psi}_1^\dagger$ by minimizing with respect to $\boldsymbol{\beta}$ the penalized least squares criterion

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\big(\mathbf{I}_n + \mathbf{Z}\boldsymbol{\Psi}^\dagger\mathbf{Z}^T\big)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2n\sum_{k=1}^{p} \phi_{\lambda^{(5)}}(|\beta_k|),$$

where $\boldsymbol{\Psi}^\dagger = \text{blockdiag}(\boldsymbol{\Psi}_1^\dagger, \boldsymbol{\Psi}_1^\dagger, \ldots, \boldsymbol{\Psi}_1^\dagger)$ has m identical blocks on the diagonal.

In both cases the SCAD penalty function is used with tuning constants $\lambda^{(4)}$ and $\lambda^{(5)}$.

One advantage of the Peng and Lu (2012) selection method is that the random effects $\mathbf{v}$ are unconstrained and are treated like unknown regression coefficients, which make the selection and computations easy to handle. In comparison, the optimization procedures in both Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) are slow and complex and can sometimes fail to converge, especially when the underlying covariance matrices are sparse and the tuning constants are small. Another advantage of the Peng and Lu (2012) approach is that it is permutation invariant and does not depend on the order in which the random effects appear. However, the estimate of $\boldsymbol{\Psi}_1^\dagger$ is not always guaranteed to be positive semidefinite and further adjustments may be needed (Peng and Lu, 2012, page 114).

Some further insight is obtained by comparing the asymptotic results in Bondell, Krishna and Ghosh (2010), Ibrahim et al. (2011) and Peng and Lu (2012). In the linear regression setting Zou (2006) proves that the ALASSO estimators possess oracle properties asymptotically. That is, as $n \to \infty$ with $p < \infty$ fixed they (a) identify the true model and (b) achieve the optimal estimation rate (i.e., the estimator performs as well as if the true model were known in advance). Similarly, Bondell, Krishna and Ghosh (2010) show that their penalized maximum likelihood estimators possess the oracle property under some regularity conditions and

$$m \to \infty, \quad \lambda^{(3)} \to \infty \quad \text{and} \quad \frac{\lambda^{(3)}}{\sqrt{m}} \to 0$$

with finite cluster sizes $1 \le n_i \le K$, for some $K < \infty$ and $i = 1, 2, \ldots, m$. Ibrahim et al. (2011) also prove that their procedure has the oracle property under some regularity conditions. Let $\boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}_{k,t}$ be the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_k$, $k = 1, 2, \ldots, s_1$, respectively. Define

$$b_m = \min\Big[\min_{j=1,\ldots,p}\{\lambda_j : \boldsymbol{\beta}_t = 0\}, \min_{k=1,\ldots,s_1}\{\lambda_{p+k} : \|\boldsymbol{\gamma}_{k,t}\| = 0\}\Big]$$

and

$$c_m = \max\Big[\max_{j=1,\ldots,p}\{\lambda_j : \boldsymbol{\beta}_t \neq 0\}, \max_{k=1,\ldots,s_1}\{\lambda_{p+k} : \|\boldsymbol{\gamma}_{k,t}\| \neq 0\}\Big].$$

The limit conditions are

$$m \to \infty, \quad \sqrt{m} b_m \to \infty \quad \text{and} \quad c_m \to 0.$$

Peng and Lu (2012) show that their method is a consistent variable selection procedure with some oracle properties, but the extra condition $s_1 < m^{-1} \sum_{i=1}^{m} n_i$ is needed. As noted by Peng and Lu (2012), when the cluster sizes are small their method does not perform as well (and is not as efficient) as methods based on the marginal distribution. Note that both Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) use the marginal distribution when deriving their shrinkage estimators, which is an advantage in this case.

The shrinkage methods discussed above produce estimates of the model parameters and select a model conditional on the tuning constants being known. By varying the values of the tuning constants from large to small, a path through the model space is defined where more parameters get selected as $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$, $\lambda^{(4)}$ and $\lambda^{(5)}$ each approach zero. Model selection on the path is reduced to selecting the values of the tuning constants. This is one of the major advantages of shrinkage methods over direct application of information criteria: shrinkage methods do not need to consider all possible models (which is often not computationally feasible when p and s_1 are large), but only the models identified on the path. Once the path is identified, information criteria, cross-validation or other methods can then be used to select the model from the path (see Section 3 for further details). The Fence method described in Section 5 also uses a similar concept where models within a "fence" are first identified, and then the second step chooses the least complex model.

The choice of tuning constant is important because this ultimately controls which model gets selected. Bondell, Krishna and Ghosh (2010) choose the tuning constant to minimize the BIC type criterion

$$-2\ell\big(\widehat{\boldsymbol{\theta}}^{\dagger}\big) + \log(n)\#\big(\widehat{\boldsymbol{\theta}}^{\dagger}\big), \tag{29}$$

over a grid of $\lambda^{(3)}$ values, where $\#(\widehat{\boldsymbol{\theta}}^{\dagger})$ is the number of nonzero elements in $\widehat{\boldsymbol{\theta}}^{\dagger}$. Ibrahim et al. (2011) consider the broader class of generalized linear mixed models where often the marginal likelihood is not directly available. However, in the case of the linear mixed model, the marginal likelihood is available and Ibrahim et al. (2011) would apply the BIC criterion

$$-2\ell\big(\widehat{\boldsymbol{\theta}}\big) + \log(m)\#\big(\widehat{\boldsymbol{\theta}}\big) \tag{30}$$

directly. There are clearly differences between (29) and (30). The $\boldsymbol{\theta}^{\dagger}$ in Bondell, Krishna and Ghosh (2010) does not include σ^2, whereas Ibrahim et al. (2011) do include σ^2 and so $\#(\widehat{\boldsymbol{\theta}}^{\dagger})$ and $\#(\widehat{\boldsymbol{\theta}})$ are slightly different. In the linear mixed model, the definition of the effective sample size is not obvious and has long been an issue for debate. Bondell, Krishna and Ghosh (2010) use the total sample size n in (29), but (30) uses the total number of clusters m. Another alternative is to estimate the effective sample size by incorporating an estimate of the correlation matrix as suggested by Jones (2011) (see BIC_J in Section 3.3), which leads to an estimate of the effective sample size between m and n. A referee pointed out that using information criteria to choose the tuning constants here has not been rigorously justified and is somewhat ad-hoc. The issue is that the number of nonzero estimated parameters corresponding to a given tuning constant is not the same as the fixed number of independent parameters under an assumed model.

An alternative way of choosing the tuning constant is to treat it like an additional variance component in the model to be estimated directly along with $\boldsymbol{\tau}$. A similar approach is often used in the semiparametric regression literature when estimating tuning constants associated with penalized splines (Ruppert, Wand and Carroll, 2003, page 108). Tibshirani (1996) notes that $|\beta_j|$ is proportional to (minus) the log density of the double exponential distribution. Foster, Verbyla and Pitchford (2007) incorporate a LASSO penalty for $\boldsymbol{\beta}$ into a linear mixed model and for estimation of the tuning constant each β_k is assumed to have a double exponential distribution with variance $2/\lambda^2$, where λ is the tuning constant (so λ is effectively treated like a hyperparameter in a hierarchical model). Estimation of λ is then carried out by maximizing an approximate marginal log-likelihood. Ibrahim et al. (2011) also use a similar idea for estimating their two tuning constants $\lambda^{(1)}$ and $\lambda^{(2)}$, however, they note that the estimates produced from this method lead to significant overfitting.

5. FENCE METHODS

Alternative model selection methods to information criteria or shrinkage methods are rare and typically ad-hoc. A notable exception is the Fence method for selecting predictors for complex models, which was recently proposed by Jiang et al. (2008). The Fence method is computationally very demanding, particularly because it involves the estimation of the standard deviation of the difference of lack-of-fit measures, for example, the negative log-likelihood as in Section 3,

the residual sum of squares or any appropriate estimated loss, denoted by $Q_M = Q_M(\boldsymbol{\theta}_M)$, $M \in \mathcal{M}$, satisfying $Q_{M_2} \leq Q_{M_1}$ if $M_1 \subset M_2$. For example, $Q_M = [\mathbf{y} - \mathrm{E}_M(\mathbf{y})]^T[\mathbf{y} - \mathrm{E}_M(\mathbf{y})]$. The Fence procedure in Jiang et al. (2008) requires the calculation of

$$\widehat{\sigma}_{M,\widetilde{M}} = \sqrt{\widehat{\mathrm{Var}}[Q_M(\boldsymbol{\theta}_M) - Q_{\widetilde{M}}(\boldsymbol{\theta}_{\widetilde{M}})]}$$

for all models $M \in \mathcal{M}$, where $\widetilde{M}$ has the smallest loss among all considered models. Jiang, Nguyen and Rao (2009) reduce to some extent the computational burden of the Fence method in their Simplified Adaptive Fence procedure, which can be very competitive in lower-dimensional problems and where convergence of estimation procedures is not of a concern, such as when using the least squares estimator in linear regression with $\mathbf{X}^T\mathbf{X}$ of full rank.

The key idea behind the Fence method is to estimate the loss for any correct model M_l by $Q_{M_l}(\widehat{\boldsymbol{\theta}}_{M_l})$, which satisfies a range of regularity conditions and is used to construct a fence. In practice, M_l can be the full or any other sufficiently large model. The first step is to identify models $M \in \mathcal{M}$ inside the fence, that is, models satisfying

$$Q_M \leq Q_{M_l} + b_n \widehat{\sigma}_{M,M_l}, \tag{31}$$

where b_n is a sequence of tuning constants.

The second step of Fence is to identify the least complex model within the fence. If there is more than one such candidate, the model with the smallest lack-of-fit measure is selected. Conceptually, Fence shares a major advantage with shrinkage methods (see Section 4); they both consider only a small proportion of models in $\mathcal{M}$, although they choose the subset differently and select from the subset differently. For Fence, only a small number of models satisfy (31) when b_n is small and these models can be identified economically through backward or forward search algorithms. The calculation of $Q_M(\widehat{\boldsymbol{\theta}}_M)$ is often straightforward, particularly when Q_M is the negative log-likelihood or residual sum of squares. Using the residual sum of squares can be promising when focus is on the selection of regression parameters that relate to the whole population, but it could be more natural to use the conditional log-likelihood when the selection focuses mainly on parameters describing clusters.

The Simplified Adaptive Fence procedure, a computationally simpler version to Adaptive Fence introduced in Jiang et al. (2008), absorbs the difficult quantity $\widehat{\sigma}_{M,M_l}$ and the tuning constant b_n into a single constant,

$$Q_M \leq Q_{M_l} + c_n. \tag{32}$$

Thus, the model selection problem turns into optimally choosing the tuning constant c_n. Jiang, Nguyen and Rao (2009) suggest calculating for each $M \in \mathcal{M}$ the bootstrapped probability $p^*(M; c_n) = \mathbf{P}^*(M_0(c_n) = M)$, where $M_0(c_n)$ is the optimal model satisfying (32). Jiang, Nguyen and Rao (2009) calculate $p^*(M; c_n)$ with a parametric bootstrap under M_a, a large correct model with at least one redundant component. M_a can be the full or any large model which is known to be correct but not optimal. On the other hand, if the full model might be the optimal model, $\mathbf{X}$ can be extended to $(\mathbf{X}, \mathbf{x}_a)$. In our own simulations we used $\mathbf{x}_{ia} = (\frac{1}{p}\sum \mathbf{x}_j)_{l_i}$, where $(l_1, \ldots, l_n)$ is a random permutation of $\{1, \ldots, n\}$. Adding this additional explanatory variable worked well in our simulations, but there are many other possibilities. Jiang, Nguyen and Rao (2008, 2009) give an elaborate explanation of why such an adjustment is required. Essentially, it ensures that the function $p^*(c_n) = \max_M p^*(M; c_n)$ has desirable theoretical features. In particular, the model that corresponds to the first significant peak at $\widehat{c}_n$, that is, $\widehat{M}(\widehat{c}_n) = \arg\max_M p^*(M; \widehat{c}_n)$, is a consistent estimate of a correct model M_l satisfying $M_a \supset M_l \supseteq M_t$, provided the true model M_t exists and the true model is not the model used for the generation of the parametric bootstrap samples, that is, $M_t \neq M_a$. Jiang, Nguyen and Rao (2009) state a theorem, which (under some technical regularity conditions) establishes the existence of a $\widetilde{c}$ (depending on n), which is at least a local maximum and an approximate global maximum of $p^*(\widetilde{c})$, such that the corresponding $\widehat{M}(\widetilde{c})$ is consistent—in the sense that for any $\kappa_1, \kappa_2 > 0$, there exist $n_{\min}$ and $B_{\min}$ such that

$$\mathbf{P}\big(p^*(\widetilde{c}) \geq 1 - \kappa_1\big) \wedge \mathbf{P}\big(\widehat{M}(\widetilde{c}) = M_t\big) \geq 1 - \kappa_2 \quad \text{if } n \geq n_{\min} \text{ and } B \geq B_{\min}. \tag{33}$$

Jiang, Nguyen and Rao (2009) refer for the proof of (33) to the proof of Theorem 3 in Jiang et al. (2008). For specific choices of Q_M and $\mathcal{M}$ it could require some care to show that all the regularity conditions hold. Empirically, we confirmed that the first significant peak, which occurs at $\widehat{c}_n$, satisfies $p^*(\widehat{M}; \widehat{c}_n) \approx 1$ for $\widehat{M} \neq M_a$, where $\widehat{c}_n$ is the smallest possible choice of the tuning constant when the peak is a plateau. For small to moderate n and for larger true models the peaks relating to the true model tend to be smaller than one, whereas for small true models we often observed $p^*(M_t; c_n) = 1$ for $c_l < c_n < c_u$. In our own simulations we found that the following rule establishes a surprisingly successful and "simple" estimator of the true

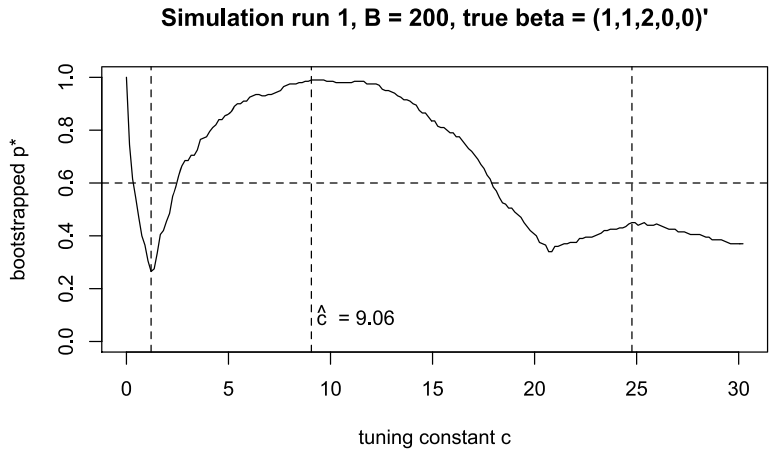

FIG. 1. *A plot of p^* based on the first simulated data set under the simple linear mixed model $y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} + \gamma u_i + \sigma\varepsilon_{ij}$, $i = 1, \ldots, 10 = m$, $j = 1, \ldots, 5$, $\gamma = \sigma = 1$ and $u_i, \varepsilon_{ij} \sim$ independent $\mathcal{N}(0, 1)$.*

model: Consider only c values with $p^*(c_n)$ attained by some $\widehat{M} \subset M_a$; choose the first $\widehat{c}_n$, which is either a peak larger than some arbitrary value τ in $(0.5, 1)$ or the smallest c_n value having maximal $p^*(c_n)$ value. In our own implementations we used $\tau = 0.6$, which was chosen before running any simulations, by a visual inspection of all published results in the series of Fence papers. (Jiang, Nguyen and Rao, 2009, suggest another adjustment, based on lower bounds of large sample 95% confidence intervals, which depend on the bootstrap sample size and p^*.)

Figure 1 shows a plot of p^* over an appropriate range of the tuning constant c_n. The data generating model is a $m = 10$ independent cluster model with group sample sizes $n_i \equiv 5$. The full model has four covariates and an intercept, and the true model has parameter vector $\boldsymbol{\beta}_t = (1, 1, 2, 0, 0)^T$. Responses were generated by $y_{ij} = \mathbf{x}_i^T \boldsymbol{\beta} + \gamma u_i + \sigma\varepsilon_{ij}$, $i = 1, \ldots, 10 = m$, $j = 1, \ldots, 5$, $\gamma = \sigma = 1$ and $u_i, \varepsilon_{ij} \sim$ independent $\mathcal{N}(0, 1)$ with $x_{i1} = 1$ and the remaining explanatory variables generated independently from $\mathcal{U}(-2, 2)$. We used the `lme()` function of the `nlme` R-library to fit a total of $2^4 + 1 = 17$ linear mixed models as described above having the same variance parameters. The Simplified Adaptive Fence procedure with $Q(\boldsymbol{\theta}) = -2\ell(\boldsymbol{\theta})$ correctly estimates the true model and the corresponding peak occurs at $\widehat{c}_n = 9.06$. Three additional models have peaks: the too large model used for the generation of the bootstrap samples at $c = 0$, the correct model having only $\beta_4 = 0$ at $c = 1.21$ (which is a local maximum difficult to detect by visual inspection) and the incorrect model with $\boldsymbol{\beta} = (\beta_0, 0, \beta_2, 0, 0)^T$ with a peak $p^*(24.77) = 0.450$.

A major attraction of the Simplified Adaptive Fence is its generality. On the other hand, since the Simplified Adaptive Fence is heavily based on bootstrapping from a too large correct model, it highlights any computational limitations in the available estimation procedures. In our simulations we noticed that fitting linear mixed models with redundant random effects can be problematic. For example, we repeatedly generated data from the same data generating model as in Bondell, Krishna and Ghosh (2010), Example 1. Using `lme()` and maximum likelihood, we found that in seven out of the first ten simulation runs the estimates failed to converge. The function `lmer()` from the R-package `lme4` never failed in the first thousand simulation runs but produced seven warnings of the type `In mer_finalize(ans): singular convergence (7)` and, more severely, calculating an auxiliary quantity such as $\widehat{\boldsymbol{\Psi}}^{-1}$ failed in five out of the first ten simulation runs returning the warning `Error in solve.default(VarCorr(M)$grp): system is computationally singular: reciprocal condition number`. This is in fact a problem for most methods, including information criteria and shrinkage methods.

We conclude that using the Simplified Adaptive Fence can be attractive when convergence is not a concern. However, it is potentially tedious to implement the Simplified Adaptive Fence in simulation studies that automatically loop through many runs of fitting mixed models with redundant random terms. This is a possible explanation for why Jiang, Nguyen and Rao (2008, 2009) focused in their simulations on the selection of $\boldsymbol{\beta}$ only, and demonstrated that the Simplified Adaptive Fence can successfully deal with linear

mixed models as long as interest focuses on selecting the regression parameters.

Recent work on the Invisible Fence (Jiang, Nguyen and Rao, 2011) and the Restricted Fence (Nguyen and Jiang, 2012) explores some ways to reduce the computational burden. Just like the Fence and Simplified Adaptive Fence, the Invisible Fence is based on the principle of selecting the model within the fence that has minimum dimension and minimum Q_M among other models within the fence of the same dimension. Jiang, Nguyen and Rao (2011) showed that the model selected by the Simplified Adaptive Fence is one of the models that minimizes Q_M at each model dimension. This means that if we can find this small set of models (one for each model dimension), the model selection problem is considerably simplified. The Invisible Fence uses the bootstrap to find the reduced set of models. Specifically, for the bth bootstrap sample, for each model dimension j, find the model M_{bj}^* of dimension j that minimizes Q_{Mb}^*, the loss Q_M computed for the bth bootstrap sample. Then, for each fixed model dimension j, find the most frequently selected model across bootstrap samples M_j^* and its bootstrap selection frequency p_j^*. The Invisible Fence selects the model M_j^* with the highest bootstrap selection frequency p_j^*. Jiang, Nguyen and Rao (2011) apply the procedure to a genetic problem (which is not a linear mixed model problem) with what they call a subtractive loss Q_M and show that, in this case, the Invisible Fence is very fast. However, in general, including for linear mixed models, it is still computationally burdensome to find the reduced set of models.

The idea of applying the Fence principle to subsets of the model space rather than to the entire space to reduce the computation is developed further in the Restricted Fence (Nguyen and Jiang, 2012). The basic idea is to partition the model space $\mathcal{M}$ into not necessarily disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_J$ and apply the Simplified Adaptive Fence to each subset $\mathcal{M}_j$. The final model is then selected by applying the Simplified Adaptive Fence again to select one of the J already selected models. In particular applications, the choice of subsets of the model space may be based on substantive considerations, but it will often involve some arbitrary choices. So just as the order of rows and columns affects the Cholesky decomposition of $\boldsymbol{\Psi}$ (see Section 2) and hence can affect model selection in shrinkage methods, the choice of subsets can, in small samples, affect model selection with the Restricted Fence. The Restricted Fence was introduced for selecting independent cluster models when interest centers on the selection of the regression parameters only and, in this case, the subsets $\mathcal{M}_j$ correspond to subsets of the columns of $\mathbf{X}$. It is then attractive to multiply both sides of the model (1) by a matrix that is orthogonal to the columns of $\mathbf{X}$ not in the current subset of interest so that these variables are removed from the model. Two further simplifications are introduced. First, instead of the generalized least squares estimator (9) of the regression parameters $\boldsymbol{\beta}$, Nguyen and Jiang (2012) use the least squares estimator so that they do not have to estimate the marginal variance matrix $\mathbf{V}$. This involves a loss of efficiency but reduces the convergence issues. Second, Nguyen and Jiang (2012) use a version of the wild bootstrap in which they bootstrap from linear regression models rather than linear mixed models. Both of these simplifications are tied to selecting regression parameters, but they suggest useful analogues for other problems and may be useful for model selection methods beyond the Restricted Fence. Generalizing and modifying the Restricted Fence to more general situations is promising and deserves further attention.

6. OTHER BAYESIAN METHODS

Bayesian model selection (also called model choice) requires us to assign a prior distribution over $\mathcal{M}$ and compute the posterior probabilities of each $M_l \in \mathcal{M}$. These computations can be difficult so are usually carried out by applying sophisticated Markov Chain Monte Carlo (MCMC) algorithms. We can actually avoid explicit model selection by working directly with the posterior distribution. If we need a single model, we can average over $\mathcal{M}$ or we can select the model with highest posterior probability. A useful way of interpreting this kind of selection (which links it conceptually to shrinkage and Fence methods) is that the MCMC algorithm reduces $\mathcal{M}$ to a small subset of models with posterior probability above a threshold and we then select one of these.

For linear mixed models, this kind of approach has been explored in a number of papers starting with Chen and Dunson (2003). They consider the problem of selecting the variance parameters $\boldsymbol{\tau}$ in the independent cluster model with $s_i = s_1$ and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$. They introduce the alternative Cholesky decomposition and define $\boldsymbol{\theta}^\dagger = (\boldsymbol{\beta}^T, \mathbf{d}^T, \boldsymbol{\gamma}^{\dagger T}, \sigma^2)^T$, where $\mathbf{d}$ contains the diagonal elements of $\mathbf{D}$ and $\boldsymbol{\gamma}^\dagger$ contains the distinct elements of $\boldsymbol{\Gamma}^\dagger$. Chen and Dunson assume that the elements of $\mathbf{d}$ are independently distributed with a point

mass at zero mixed with a Gaussian distribution truncated at zero. The assumption that the elements of $\mathbf{d}$ are independent allows each one to be treated independently and the zero-inflated truncated-Gaussian priors allow them to be exactly zero with positive probability. Selection is based on running a Gibbs sampler and computing the posterior probabilities of all possible models (of which there are at most 2^{q_γ}) by dividing the number of occurrences of each model by the number of iterations. Saville and Herring (2009) point out that these kinds of MCMC methods are generally time consuming to implement, require special software and depend on subjective choice of the hyperparameters in the priors.

As discussed in Section 3.3, the problem can also be formulated as a testing problem and the test carried out by computing Bayes factors (20). The two issues with using Bayes factors are the choice of prior, which, depending on the formulation, might need to include point mass at zero and should not be either too concentrated or too dispersed, and the computation. Han and Carlin (2001) compare a number of methods for computing Bayes factors for comparing two linear mixed models. They find that the reversible jump (Green, 1995) and marginal likelihood methods (Chib, 1995) are able to produce estimates of the Bayes factor and that the marginal likelihood methods are easier to use. The marginal likelihood here refers not to $\exp\{\ell(\boldsymbol{\theta})\}$ but, in the notation used to define the Bayes factor (20), to $\int g(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})\,d\boldsymbol{\theta}$, where $g(\mathbf{y}|\boldsymbol{\theta}) = \exp\{\ell(\boldsymbol{\theta})\}$ and h is the prior for $\boldsymbol{\theta}$. Chib (1995) and Han and Carlin (2001) comment that all the methods require substantial human intervention and computer effort for a modest payoff. These kinds of conclusions help motivate the use of approximations like BIC (Section 3.3) to the Bayes factor and also more ad hoc alternative approaches to model selection such as those of Spiegelhalter et al. (2002) and Aitkin, Liu and Chadwick (2009).

Spiegelhalter et al. (2002) propose a general Bayesian deviance criterion for model selection of the form

$$DIC = \mathrm{E}\{-2\ell(\boldsymbol{\theta})|\mathbf{y}\} + 2\log\{f(\mathbf{y})\} + 2p_D,$$

where $p_D = \mathrm{E}\{-\ell(\boldsymbol{\theta})|\mathbf{y}\} + \ell(\bar{\boldsymbol{\theta}})$, $\bar{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{y})$ is the posterior mean of the parameters and $f(\cdot)$ is a "fully specified standardizing term that is a function of the data alone." The choice of f is vague, but a natural choice is $f(\mathbf{y}) = \exp\{\ell(\widehat{\boldsymbol{\theta}})\}$ for some estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. If the estimate $\widehat{\boldsymbol{\theta}}$ is fixed for all comparisons, then we can omit the standardizing term. This is the same as just setting $f(\mathbf{y}) \equiv 1$. For selecting regression terms in the mixed model when the variance parameters $\boldsymbol{\tau}$ are known, Spiegelhalter et al. (2002) point out that p_D reduces to the effective degrees of freedom $\rho(\boldsymbol{\tau})$ defined in (15) so, as noted by Vaida and Blanchard (2005), DIC in this case is equivalent to marginal AIC with the asymptotic form of the Vaida–Blanchard penalty for conditional AIC.

Aitkin, Liu and Chadwick (2009) propose a different way of using deviances to select models from Spiegelhalter et al. (2002). They suggest comparing models M_0 and M_1 by computing the posterior distributions of the parameters $\boldsymbol{\theta}_{M_0}$ and $\boldsymbol{\theta}_{M_1}$, generating B realizations $\boldsymbol{\theta}_{M_0 b}$ and $\boldsymbol{\theta}_{M_1 b}$ from the respective posterior distributions and computing the empirical probability $\Pr\{-2\ell(\boldsymbol{\theta}_{M_0 b}) + 2\ell(\boldsymbol{\theta}_{M_1 b}) < -4.4|\mathbf{y}\}$. The value $-4.4 = -2\log(9)$ corresponds to a likelihood ratio of 9 so the event $\{-2\ell(\boldsymbol{\theta}_{M_0 b}) + 2\ell(\boldsymbol{\theta}_{M_1 b}) < -4.4\}$ represents strong evidence for M_0 over M_1. They argue that if the empirical probability of the event is 0.9 or greater, there is a high posterior probability of strong evidence in favor of M_0 over M_1. This approach has attracted criticism from some Bayesians (Gelman, Robert and Rousseau, 2010).

7. SIMULATION

Various authors have carried out simulations to compare different methods of model selection, usually with one or more similar methods and usually in problems with a small number of parameters. We review some of these simulations in this section to see what we can learn from putting the results together. Each simulation is limited but, together, they are quite informative, particularly in identifying individual problems in which particular methods work well. We think of this as like a meta-analysis which extracts more information by combining existing studies without having to repeat studies or run new studies. A summary of the settings considered is given in Table 2, which is followed by a concise overview of the most important findings. More detailed information and further comments on the simulations can be found in the online supplementary material (see Appendix following the bibliography).

It is clear from Table 2 that only a limited set of models and limited settings have been considered. All except Srivastava and Kubokawa (2010) and Jiang, Nguyen and Rao (2009) considered the easier case with constant cluster size. The numbers of parameters and random effects are very small in both the true and the full models; the exceptions are Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011)

TABLE 2

Table summarizing the settings used in selected simulations. "Reference" shows the first letters of the surnames of the authors and the last two digits of the year of publication, "Model" describes the model considered, m the number of clusters and n_i the size of the clusters. The quantities p, s_i and q are the dimension of $\boldsymbol{\beta}$, the number of random effects per cluster and the dimension of $\boldsymbol{\tau}$ in the true model; p_f, s_{fi} and q_f are the analogous quantities under the full model. The next three measures describe the difficulty of selecting the true model: $|\mathcal{M}_{\boldsymbol{\beta}}|$ and $|\mathcal{M}_{\boldsymbol{\tau}}|$ are the number of candidate models considered for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, respectively, $\min|\beta_k|/\sigma$ measures the difficulty of selecting the smallest nonzero regression parameter when there are no random effects in the model and $\min\{ev(\boldsymbol{\Psi}/\sigma^2)\}$, the smallest eigenvalue of $\boldsymbol{\Psi}/\sigma^2$, measures the difficulty of selecting the smallest nonzero variance parameter. Finally, $\mathbf{u}$ and $\mathbf{e}$ describe the distributions used for these random variables and "Method" denotes the main model selection methods considered in the simulation

Reference	Model	m/n_i	p/p_f	s_i/s_{fi}	q/q_f
CD03	int + slope	200/8	4/4	3/4	7/11
DMT11	int + slope	10/{6, 26, 51}	2/6	2/2	4/7
PN06	int + slope	10/20	3/5	2/3	3/7
SC08	int	{15, 20, 30, 50}/3	7/12	1/1	2/2
SC08	int	{15, 20, 30, 50}/3	4/5	1/1	2/2
GK10	int	{10, 20, 40, 80}/{3, 6, 9, 12}	2/2	1/1	2/2
DMT11	int + slope	{10, 20, 50}/4	2/6	2/3	4/7
DMT11	int + slope	{10, 20, 50}/4	3/6	1/3	4/7
SK10	cluster	20/{1 + $\mathcal{B}(8, 1/2)$}	{2, 4, 6}/7	{1, 2, 3}/{1, 2, 3}	2/2
K11	Fay–Herriot	{5, 10, 30}/1	4/7	1/1	1/1
K11	int	{5, 10, 30}/4	{2, 4, 6}/{5, 7}	1/1	2/2
JR03	var comp	8000/3	2/5	{20, 40}/140	{2, 3}/8
BKG10	cluster	{30, 60}/{5, 10}	2/9	3/{4, 10}	7/{11, 56}
IZGG11	cluster	{50, 100, 200}/12	3/8	3/8	7/37
PL12	cluster	{10, 20}/{10, 20}	3/5	2/4	3/10
JRGN08	Fay–Herriot	30/1	1–5/5	1/1	1/1
JRGN08	int	100/5	{2, 4, 5}/5	1/1	2/2
JNR09	int	{10, 15}/$\mathcal{P}(3)$	{3, 6}/6	1/1	2/2
NJ12	int	{50, 100, 150}/3	7/30	1/1	2/2

Reference	$\lvert\mathcal{M}_\beta\rvert/\lvert\mathcal{M}_\tau\rvert$	$\min\lvert\beta_k\rvert/\sigma$	$\min\{\text{ev}(\Psi/\sigma^2)\}$	u/e	Method
CD03	1/16	1	0.45	$\mathcal{N}$	Post prob.
DMT11	14/3	0.35	0.01	$\mathcal{N}$	IC
PN06	31/7	0.2	0.5	$\mathcal{N}$	GIC
SC08	12/2	1	2	$\mathcal{N}$	mAIC
SC08	31/2	1	2	$\mathcal{N}$	mAIC
GK10	1/2	1	{0.1–0.8}	$\mathcal{N}$	cAIC
DMT11	14/3	1.83	0.17	$\mathcal{N}$/{$\mathcal{N}$, mixtures}	IC
DMT11	14/3	1.83	0.11	$\mathcal{N}$/{$\mathcal{N}$, mixtures}	IC
SK10	7/1	2	{0.01, 0.5, 1}	$\mathcal{N}$	AIC
K11	7/1	2	1	{$\mathcal{N}$, mixture($\mathcal{N}, \mathcal{C}$)}	AIC
K11	7/2	2	{0.1, 1}	{$\mathcal{N}, t_3$}	AIC
JR03	31/NA	1.63	0.67	$\mathcal{N}$	own
BKG10	512/16	1	0.45	$\mathcal{N}$	shrinkage
IZGG11	256/256	{1.5, 0.5}	{0.41, 0.05}	$\mathcal{N}$	shrinkage
PL12	16/16	1	0.32	$\mathcal{N}$	shrinkage
JRGN08	32/1	1	1	$\mathcal{N}$	Fence
JRGN08	32/1	1	1	$\mathcal{N}$	AFence
JNR09	64/1	1	1	$\mathcal{N}$	SAFence
NJ12	768/1	0.001	1	$\mathcal{N}$	RFence

who consider slightly larger numbers of variance parameters in the full model (q_f) and Jiang and Rao (2003) who consider large numbers of random effects in the full model. The sets of candidate models are relatively small, the largest occurring in Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011). Small values of $\min|\beta_k|/\sigma$ and $\min\{\text{ev}(\boldsymbol{\Psi}/\sigma^2)\}$ indicate that it is difficult to select the true model for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, respectively. The table shows that, with the exception of Nguyen and Jiang (2012), the settings make it relatively easy to select the true $\boldsymbol{\beta}$ and, surprisingly, often much easier than to select the true $\boldsymbol{\tau}$. This helps explain the general conclusion that selecting $\boldsymbol{\beta}$ is easier than $\boldsymbol{\tau}$. Most authors choose the true regression parameters according to their favored procedure, that is, for AIC-like criteria p is close to p_f and for BIC-like criteria and shrinkage methods p is small compared to p_f. Also, some authors apply their own variants of information criteria without any justification or explanation, and possibly with unintended effects.

For the marginal information criteria, as in linear regression models, larger penalties tend to select smaller models, while smaller penalties tend to select larger models. The bootstrap penalty is plausible ($mAIC_{B2}$ worked better than $mAIC_{B1}$) but has not been thoroughly explored. For the conditional AIC criteria, the Greven–Kneib penalty and the Srivastava–Kubokawa penalty produced promising results but need a more thorough investigation. The philosophical differences between using marginal and conditional criteria were explained by Vaida and Blanchard (2005), but the practical differences are much less clear. Dimova, Markatou and Talal (2011) found in their simulation that the conditional criteria performed worst at selecting the correct model, tending to prefer larger models. They recommended GIC with $a_n = n^{1/2}$ but noted that it does not always get the random effects right, particularly when they have small variance. On the other hand, a version of REML-based $mAIC_R$, which ignores the estimation of $\boldsymbol{\Psi}$, worked well when $\boldsymbol{\Psi}$ is close to zero. Bondell, Krishna and Ghosh (2010) and Ibrahim et al. (2011) obtained promising results for their shrinkage methods. The methods of Ibrahim et al. (2011) have the advantage of having two tuning parameters, although this makes the computations more burdensome. They found that the SCAD penalty performed best for regression parameters and ALASSO for variance parameters. The Fence methods can be difficult to implement with redundant variance parameters and have not yet been investigated in the full model selection problem. Chen and Dunson (2003) found that their approach selected the true model with high probability and the performance was robust to the choice of hyperparameters for the point mass at zero mixed with a zero-truncated Gaussian distribution prior for each d_k.

Finally, most of the studies used Gaussian distributions and those that did not found that their methods performed more poorly under the longer-tailed distributions they used.

8. DISCUSSION AND CONCLUSIONS

In this paper we have arranged, structured and reviewed a substantial body of literature on different model selection procedures for linear mixed models. A key step in achieving this is our use of a unified notation for the linear mixed model (1), which we use in particular to (i) bring together special cases of the linear mixed model such as the variance component model, the independent cluster model, the clustered variance component model, the random intercept and slope model, the Fay–Herriot model or the longitudinal autoregression model; (ii) avoid ambiguity in identifying what components are subject to selection: regression parameters $\boldsymbol{\beta}$, variance parameters $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ or $\boldsymbol{\tau} = (\boldsymbol{\gamma}^T, \boldsymbol{\delta}^T)^T$ or both simultaneously, that is, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T)^T$; and (iii) make different model selection procedures suggested by different authors easier to compare.

The performance of model selection procedures depends on how performance is measured. Much of the theoretical work on information criteria gives the right answer to a good question, such as how to estimate the Akaike Information unbiasedly (AIC) or how to approximate the Bayes factor accurately (BIC), but these criteria are not directly related to model selection. Direct performance measures, such as how often the data generating model or other correct models are detected, are more useful. Parsimony (choosing models with few parameters) is an important consideration when $p + q$ is large. It can be achieved by the choice of combinations of the measure of model complexity, the penalty function or the tuning constants and should be built into the performance measures. Procedures that are optimal under one performance measure need not be optimal under a different measure, so it may be worthwhile to consider several measures.

One of the key issues in model selection is that the set $\mathcal{M}$ of candidate linear mixed models can be very large; depending on the model, $\mathcal{M}$ can contain all 2^{p+q} possible models and, in such cases, is very large when

$p+q$ is large. Large candidate sets $\mathcal{M}$ are computationally too demanding for methods like the information criteria (Section 3) which try to compare all the models in $\mathcal{M}$. A natural alternative approach is to try to reduce $\mathcal{M}$ efficiently to a smaller subset of models and then select models from within this subset. Shrinkage methods (Section 4), Fence methods (Section 5) and implicitly some Bayesian methods (Section 6) which try to do this are better able to handle large $\mathcal{M}$. There are many open questions about how to reduce $\mathcal{M}$ in appropriate ways and we anticipate an explosion of results similar to that currently occurring in $n \ll p$ problems in linear regression.

The theoretical treatment of mixed model selectors is difficult and technical so the results that have been obtained are impressive. Generally, these results require either strong assumptions or restrictions to specific mixed models only, such as those having a single variance parameter, and more theoretical insight would be very useful.

The difficulty of developing theoretical results means that we have to rely on simulations to compare different methods. In reviewing the various simulations, we found that only a limited set of models and limited settings have been considered. In particular, the shrinkage and Fence methods have only been applied to the independent cluster model to date. As with the theory, more general and more challenging scenarios should be investigated in the future. Interesting avenues for future studies are to consider more general $\mathbf{\Sigma}$ than $\sigma^2\mathbf{I}_n$, letting n, p, q and s grow in different ways in asymptotic studies, and exploring true joint selection of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$.

With currently available software (e.g., `lmer` in R or `Proc Mixed` in SAS), it is easy to initiate a request for and, provided the problem is not too large or too sparse, to obtain a point estimate for $\boldsymbol{\theta}$. Nonetheless, there are computational issues, particularly when one or more variance parameters is zero (see Sections 4 and 5). This has implications for computer intensive selection procedures, which can fail when estimation in any one of the iterations fails. We expect that optimization routines will develop and include better methods for dealing with problems where the underlying model parameters are at or near the boundary. Similar and possibly more serious computational difficulties arise with Bayesian methods. Han and Carlin (2001) remarked that all the Bayesian methods they considered required substantial time and effort (both human and computer). They pointed out that both the boundary issues and the choice of priors have to be treated with care.

There are interesting relationships between the method of estimation, the method of selection and the definition of the possible model set $\mathcal{M}$. With clustered data, it is important to distinguish (Vaida and Blanchard, 2005) or to be conscious of the distinction (Greven and Kneib, 2010) between marginal questions regarding the underlying population from which clusters are observed and conditional questions regarding the particular clusters in the data when using information criteria (Section 3). This distinction has implications for shrinkage and Fence methods. Specifically, in order to select models to treat conditional questions, it is worthwhile developing shrinkage methods based on the conditional log-likelihood $\ell(\boldsymbol{\theta}|\widehat{\mathbf{u}})$ and measuring model complexity in Fence using one of the conditional AIC penalties in Table 1.

APPENDIX: SIMULATION SETTINGS

Vaida and Blanchard (2005) used as the full model the simple random intercept and slope model

$$y_{ij} = \beta_1 + x_j\beta_2 + \mathbf{z}_{ij}^T\mathbf{\Gamma}_i\mathbf{u}_i + \sigma e_{ij},$$
$$j = 1, \ldots, n_i \in \{6, 26, 51\}, i = 1, \ldots, 10,$$

with $\mathbf{u}_i = (u_{1i}, u_{2i})^T$, $\mathbf{\Gamma}_i$ a 2×2 matrix of parameters and $\mathbf{z}_{ij}^T = (1, x_j)$. The values of x_j were equally spaced in units of 5 from 0 to 25 ($n_i = 6$), 0 to 125 ($n_i = 26$) or 0 to 250 ($n_i = 51$). The true models had $\boldsymbol{\beta} = (-2.78, -0.186)^T$,

$$\mathbf{\Gamma}_i\mathbf{\Gamma}_i^T = \mathbf{\Psi}_i = \begin{pmatrix} 0.0367 & -0.00126 \\ -0.00126 & 0.00279 \end{pmatrix}$$

and $\sigma^2 \in \{0.0705, 0.141, 0.282\}$.

Chen and Dunson (2003) reported results from a simulation using the random intercept and slope regression model. In the part of the simulation where they considered selecting $\boldsymbol{\theta}$, the full model was

$$y_{ij} = \beta_1 + x_{2ij}\beta_2 + x_{3ij}\beta_3 + x_{4ij}\beta_4 + \mathbf{z}_{ij}^T\mathbf{D}_i\mathbf{\Gamma}_i^\dagger\mathbf{u}_i$$
$$+ \sigma e_{ij}, \quad j = 1, \ldots, 8, i = 1, \ldots, 200,$$

with $\mathbf{u}_i = (u_{1i}, u_{2i}, u_{3i}, u_{4i})^T$, $\mathbf{D}_i$ a 4×4 diagonal matrix, $\mathbf{\Gamma}_i^\dagger$ a 4×4 matrix and $\mathbf{z}_{ij}^T = (1, x_{2ij}, x_{3ij}, x_{4ij})$. The explanatory variables were generated independently from the $\mathcal{U}(-2, 2)$ distribution. The true model had $\boldsymbol{\beta} = \mathbf{1}_4$, $\mathbf{D}_i = \text{diag}(3, 1.2, 0.8, 0)$,

$$\mathbf{\Gamma}_i^\dagger = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1.33 & 1 & 0 & 0 \\ 0.25 & 0.71 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and $\sigma^2 = 1$. The set $\mathcal{M}$ of candidate models consisted of all $2^4 = 16$ possible subsets of $\{d_1, \ldots, d_4\}$. Chen and Dunson (2003) used a $\mathcal{N}_4(\mathbf{0}_4, 1000\mathbf{I}_4)$ prior for $\boldsymbol{\beta}$, a Gamma $\mathcal{G}(0.05, 0.05)$ prior for σ^{-2}, a π_0 mixture of a point mass at zero and a $\mathcal{N}(0, 30)$ distribution truncated at zero for each d_k with $\pi_0 \in \{0.2, 0.5, 0.8\}$, and independent $\mathcal{N}(0, 0.5)$ distributions for the elements of $\boldsymbol{\Gamma}_i^{\dagger}$, given that they are nonzero.

Pu and Niu (2006) carried out a simulation for the random intercept and slope model

$$\begin{aligned} y_{ij} &= \beta_1 + x_{2ij}\beta_2 + x_{3ij}\beta_3 + x_{4ij}\beta_4 \\ &\quad + x_{5ij}\beta_5 + \mathbf{z}_{ij}^T \boldsymbol{\Gamma}_i \mathbf{u}_i \\ &\quad + \sigma e_{ij}, \quad j = 1, \ldots, 20, i = 1, \ldots, 10, \end{aligned}$$

with $\mathbf{u}_i = (u_{1i}, u_{2i}, u_{3i})^T$, $\boldsymbol{\Gamma}_i$ a 3×3 matrix of parameters and $\mathbf{z}_{ij}^T = (1, x_{2ij}, x_{3ij})$. The explanatory variables were generated as independent $\mathcal{N}_4(\mathbf{0}, \mathbf{A}\mathbf{A}^T)$ random vectors with

$$\mathbf{A} = \begin{pmatrix} 2.00 & 0.66 & 0.90 & 0.02 \\ 0.66 & 2.00 & 0.68 & 0.32 \\ 0.90 & 0.68 & 2.00 & 0.94 \\ 0.02 & 0.32 & 0.94 & 2.00 \end{pmatrix}.$$

The true models had $\boldsymbol{\beta} = (\beta_1, 1.2, 0, 2.0, 0)^T$ with $\beta_1 \in \{0.5, 1.5, 0.2\}$, one of the variance matrices

$$\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T = \boldsymbol{\Psi}_i = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 0 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}$$

or

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix},$$

and $\sigma^2 = 1$. Following their suggested approach, Pu and Niu (2006) included all three random effects in the model and computed *GIC* with $a_n = \log(n)$ and $a_n = n^{1/2}$ for all 31 candidate regression models. Then, using the selected regression model, they computed the criteria over 7 candidate variance models. They then iterated the process until the selected models no longer changed.

Shang and Cavanaugh (2008) reported a simulation using the random intercept regression model with $m \in \{15, 20, 30, 50\}$ and $n_i = 3$ to compare the bootstrap AIC with *mAIC*. The full model included 12 covariates (they do not explain how these were generated). The true model had $p = 7$ with $\boldsymbol{\beta} = \mathbf{1}_7$, $\gamma^2 = \text{Var}(u_i) = 2$ and $\sigma^2 = 1$. The penalties were computed from $B = 500$ parametric bootstrap samples. Shang and Cavanaugh considered selecting the models with the first covariate, the first two covariates, etc., and with or without u_i. In a second simulation, they reduced the full model to 5 covariates and for the true model set $p = 4$ with $\boldsymbol{\beta} = \mathbf{1}_4$ and considered all possible subsets of the 5 variables and with or without u_i.

Greven and Kneib (2010) carried out a simulation for penalized spline smoothing and for the simple random intercept regression model

$$\begin{aligned} y_{ij} = \beta_1 + x_i\beta_2 + \gamma u_i + \sigma e_{ij}, \\ j = 1, \ldots, n_i \in \{3, 6, 9, 12\}, \\ i = 1, \ldots, m \in \{10, 20, 40, 80\}. \end{aligned}$$

The clusters were taken to be of equal size in each run. The covariate x was chosen equally spaced in the interval $[0, 1]$. The true models had $\boldsymbol{\beta} = (0, 1)^T$, $\gamma^2 \in \{0, 0.1, 0.2, 0.4, 0.6, 0.8\}$ and $\sigma^2 = 1$. The simulation compared the ability of *mAIC*, *cAIC* with the asymptotic version of the Vaida–Blanchard penalty, the Liang–Wu–Zhou (LWZ) penalty and the Greven–Kneib penalty to choose between the simple linear regression model and the nonlinear or mixed model. The nonlinearity in penalized spline smoothing is represented by the random vector $\mathbf{u}$, but there are only two variance parameters in $\boldsymbol{\tau}$, so both the two models considered represent cases with a small number of variance parameters.

In their simulation study, Dimova, Markatou and Talal (2011) compared a number of different versions of marginal AIC (*mAIC* with both finite sample and asymptotic penalties, *mAIC* treating $\boldsymbol{\Psi}$ as known, the REML version $mAIC_R$, $mAIC_R$ treating $\boldsymbol{\Psi}$ as known), conditional AIC (*cAIC* with both finite sample and asymptotic penalties and *cAIC* using the REML estimates with both finite sample and asymptotic penalties), BIC [which is GIC with $a_n = \log(n)$] and GIC with $a_n = n^{1/2}$. The full model was the random intercept and slope model

$$\begin{aligned} y_{ij} &= \beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{5ij}\beta_5 \\ &\quad + x_{5ij}^2\beta_6 + \mathbf{z}_{ij}^T \boldsymbol{\Gamma}_i \mathbf{u}_i \\ &\quad + \sigma e_{ij}, \quad j = 1, \ldots, 4, \\ &\qquad\qquad i = 1, \ldots, m \in \{10, 20, 50\}, \end{aligned}$$

with $\mathbf{u}_i = (u_{1i}, u_{2i}, u_{3i})^T$, $\boldsymbol{\Gamma}_i$ a 3×3 matrix of parameters and $\mathbf{z}_{ij}^T = (1, x_{5ij}, x_{5ij}^2)$. The explanatory variables $x_{2i} \sim$ independent $\mathcal{N}(0, 1)$, x_{3i} and x_{4i} were generated from the $\mathcal{N}(3, 4)$ distribution, and $x_{5i1} = 0$, $x_{5i2} = 6$, $x_{5i3} = 12$ and $x_{5i4} = 24$ so $\mathbf{x}_{5i} = (0, 6, 12, 24)^T$. The

u_{1i}'s were generated from Gaussian distributions, the e_{ij} were generated from Gaussian or Gaussian mixture distributions $\zeta\mathcal{N}(0, 1.2) + (1-\zeta)\mathcal{N}(8, 16)$ with $\zeta \in \{0.9, 0.8, 0.6\}$. The two true models considered had (a) $\boldsymbol{\beta} = (3, 2, 0, 0, 0, 0)^T$, the $(1, 1)$ entry ψ of $\boldsymbol{\Psi}_i = \boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T$ satisfying $\psi = \mathrm{Var}(u_{1i}) \in \{0.2, 0.5, 1.5, 4\}$ with all other entries zero, and $\sigma^2 = 1.2$, and (b) $\boldsymbol{\beta} = (10, 5, 0, 0, 2, 0)$,

$$\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^T = \boldsymbol{\Psi}_i = \begin{pmatrix} 4 & 0.5 & 0 \\ 0.5 & \psi & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

with $\psi \in \{0.2, 0.5, 1.5, 4\}$, and $\sigma^2 = 1.2$. Dimova et al. fitted 42 candidate models to the data. These included 6 models with u_{1i}, with u_{1i} and x_{4ij}, with u_{1i} and (x_{4ij}, x_{4ij}^2), with (u_{1i}, u_{2j}) and x_{4ij}, with (u_{1i}, u_{2j}) and (x_{4ij}, x_{4ij}^2), and with (u_{1i}, u_{2i}, u_{3i}) and (x_{4ij}, x_{4ij}^2), crossed with models for the regression structure made up of the $2^3 - 1 = 7$ subsets of (x_{1i}, x_{2i}, x_{3i}).

Srivastava and Kubokawa (2010) carried out a simulation study using the independent cluster model (5) with $m = 20$ clusters of size $n_i \sim 1 + \mathcal{B}(8, 1/2)$, where $\mathcal{B}$ denotes the binomial distribution. The full model had 7 explanatory variables with $s_i \equiv s_1 \in \{1, 2, 3\}$ random effects in each cluster. The n_i rows of $\mathbf{X}_i$ were generated independently from the $\mathcal{N}_7(\mathbf{0}_7, 0.7\mathbf{I}_7 + 0.3\mathbf{J}_7)$ and the n_i rows of $\mathbf{Z}_i$ were generated independently from $\mathcal{N}_{s_1}(\mathbf{0}_{s_1}, 0.7\mathbf{I}_{s_1} + 0.3\mathbf{J}_{s_1})$. The true models had $p \in \{2, 4, 6\}$ explanatory variables and the same random effects as the full model, as only selection on the regression parameters was considered. Srivastava and Kubokawa set $\beta_k = 2(-1)^{k+1}\{1 + \mathcal{U}(0, 1)\}$, for $k = 1, \ldots, p$, $\boldsymbol{\Psi} = \gamma^2\mathbf{I}_s$ with $\gamma^2 \in \{0.01, 0.5, 1\}$ and $\sigma^2 = 1$. The 7 candidate models had the correct variance structure and the first, first two, first three explanatory variables, etc. The simulation consisted of 10 generated values of $\mathbf{X}$ and $\mathbf{Z}$ with 30 sets of $\mathbf{y}$ for each, making 300 replications. They reported the frequency of selecting the correct model for $(p = 2, \psi = 0.01, s_1 = 1)$, $(p = 4, \psi = 0.5, s_1 = 2)$ and $(p = 6, \psi = 1, s_1 = 3)$ for both known and unknown γ^2. The Srivastava–Kubokawa conditional AIC method (19) using the different estimators of $\boldsymbol{\beta}$ and $\mathbf{v}$ performed similarly and outperformed *mAIC* and *cAIC* which were also very similar. A second simulation carried out with $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_{s_1})$ produced similar results.

Kubokawa (2011) carried out simulations under the Fay–Herriot model and the random intercept regression model, essentially comparing marginal and conditional AIC criteria with his Mallows type criteria. The Fay–Herriot model can be viewed as a special case of the random intercept regression model in which $n_i = 1$ so $n = m$ and $\mathrm{Var}(e) = \sigma^2$ is known; in the simulation, the random intercept regression model had clusters of size $n_i = 4$ so $n = 4m$. The full models had $p_f \in \{5, 7\}$ explanatory variables; the n_i rows of $\mathbf{X}_i$ were generated independently from the $\mathcal{N}_{p_f}(\mathbf{0}_{p_f}, 0.7\mathbf{I}_{p_f} + 0.3\mathbf{J}_{p_f})$ distribution. The components of $\mathbf{u}$ and $\mathbf{e}$ were generated independently from various distributions. The true models had $p \in \{2, 4, 6\}$ explanatory variables with the nonzero coefficients $\beta_k = 2(-1)^{k+1}\{1 + \mathcal{U}(0, 1)\}$, $k = 1, \ldots, p$ and various values of γ^2 and σ^2. A full list of settings is given in Table 3. The simulation was carried out by generating 20 values of $\mathbf{X}$ and 50 sets of $\mathbf{y}$ for each value of $\mathbf{X}$, making 1000 replications. In settings I and II, the 7 candidate models had the correct variance structure so only selection on the regression parameter including the first, first two, first three regressors, etc. was considered. In setting III, the models were also considered with and without the variance structure. The criteria all performed similarly in the first two settings (although, as noted in Section 3.4, the Mallows criteria estimating $\boldsymbol{\tau}$ from the candidate model performed very poorly) and *cAIC* was superior for the Fay–Herriot model but slightly inferior for the random intercept regression model. For selection on all the parameters, *mAIC* and *cAIC* worked well, but *mPEC* was poor and *cPEC* tended to select models without random effects. Kubokawa concluded that these criteria are only useful for selecting regression parameters.

Jiang and Rao (2003) reported results from a simulation using a crossed three factor regression model. In the part of the simulation where they considered selecting $\boldsymbol{\theta}$, the full model with $m_j = 20$, $n_i = 3$ (so the sample size $n = 3 \times 20^3 = 24{,}000$) was

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{j=1}^{7} \gamma_j \mathbf{Z}^{(j)}\mathbf{u}_j + \sigma\mathbf{e},$$

with $\mathbf{X}$ a $n \times 5$ matrix, $\boldsymbol{\beta}$ a 5-vector, $\mathbf{Z}^{(j)}$ the $n \times 20$ matrices defined in Table 4 and $\mathbf{u}_j$ independent random 20-vectors. The explanatory variables were generated as standard Gaussian random variables. They do not explain how the $\mathbf{u}_j$'s and $\mathbf{e}$ were generated, but they are most likely standard Gaussian. The two true models considered both had $\boldsymbol{\beta} = (2, 0, 0, 4, 0)^T$ and either $\boldsymbol{\gamma} = (1, \mathbf{0}_6^T)^T$ or $\boldsymbol{\gamma} = (1, 0, 1.5, \mathbf{0}_4^T)^T$. In both cases $\sigma^2 = 1.5$. Jiang and Rao did not specify the set of candidate models; implicitly it is the set of $2^5 - 1 = 31$

TABLE 3

Simulation settings for the simulation reported by Kubokawa (2011). The first two cases are from the Fay–Herriot model in which $n_i = 1$ so $n = m$; the last eight are from the random intercept regression model with $n_i = 4$ so $n = 4m$. In I-1 *and* I-2, σ^2 *is treated as known.* II-1 *and* II-2 *use the same settings, but in* II-1 *the variance parameters are treated as known. Also, p_f is the dimension of the regression parameter $\boldsymbol{\beta}$ in the largest candidate model and p is the dimension of the regression parameter in the true model. In the mixture models, $\mathcal{C}$ denotes the Cauchy distribution*

Code	p_f	p	m	γ^2	σ^2	u	e
I-1	7	4	10	0.25	0.25	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$
I-2	7	4	50	0.25	0.25	$0.9\mathcal{N}(0,1)+0.1\mathcal{C}$	$0.9\mathcal{N}(0,1)+0.1\mathcal{C}$
II-1	7	4	10	0.1	1	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$
II-2	7	4	10	0.1	1	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$
II-3	7	6	5	1	1	t_3	t_3
II-4	7	2	30	1	1	t_3	$\mathcal{N}(0,1)$
III-1	5	2	5	0	1	–	$\mathcal{N}(0,1)$
III-2	5	4	30	0	1	–	t_3
III-3	5	2	5	1	1	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$
III-4	5	4	30	1	1	t_3	t_3

all possible regression models multiplied by the number of choices in each of L_1, L_2 and L_3. However, L_2 and L_1 contain the two-way and three-way interactions of the terms in L_3, so it would be more usual to select from L_1 and, only if the model in L_1 is not selected, select from L_2, allowing the results of this selection to determine what we consider for selection in L_3. Jiang and Rao found that the penalty $a_n = a_{jn} = n/\log(n)$ worked well.

Bondell, Krishna and Ghosh (2010) undertook a simulation study to examine the properties of their LASSO procedure in finite sample size settings. The full model was the independent cluster model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\Gamma}_i\mathbf{u}_i + \sigma\mathbf{e}_i, \quad j = 1,\ldots,n_i \in \{5,10\}, \quad i = 1,\ldots,m \in \{30,60\},$$

with equal size clusters in each run, $\mathbf{X}_i$ a $n_i \times 9$ matrix of independent $\mathcal{U}(-2,2)$ random variables, $\mathbf{Z}_i$ either a $n_i \times 4$ matrix independent of $\mathbf{X}$ with first column $\mathbf{1}_{n_i}$ and the remaining 3 columns generated from the $\mathcal{U}(-2,-2)$ distribution [when $(n_i,m) \in \{(5,30),(10,60)\}$] or $\mathbf{Z}_i = (\mathbf{1}_{n_i},\mathbf{X}_i)$ a $n_i \times 10$ matrix [when $(n_i,m) = (5,60)$], $\boldsymbol{\Gamma}_i$ either a 4×4 or a 10×10 matrix, $\mathbf{u}_i$ either a 4- or a 10-vector, σ a scalar and $\mathbf{e}_i$ an n_i-vector. For the true model, for the first two scenarios $(n_i,m) \in \{(5,30),(10,60)\}$, they set $\boldsymbol{\beta} = (1,1,\mathbf{0}_7^T)^T$ and for the third $(n_i,m) = (5,60)$, they set $\boldsymbol{\beta} = (1,0,1,\mathbf{0}_6^T)^T$. In all three scenarios,

$$\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i^T = \boldsymbol{\Psi}_i = \begin{pmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{pmatrix}$$

and $\sigma = 1$. Bondell, Krishna and Ghosh (2010) compared their model selection procedure with the earlier approach suggested in the literature which first selects either the regression or variance structure using AIC and/or BIC while fixing the other at the full model (e.g., Pu and Niu, 2006). As a further comparison they also applied the ALASSO, LASSO and a stepwise procedure for selecting $\boldsymbol{\beta}$ given $\boldsymbol{\tau}$ after first selecting $\boldsymbol{\tau}$ by fixing $\boldsymbol{\beta}$ at the full model. The new procedure was shown to be closest to "oracle" and to correctly identify the true model most often.

TABLE 4

The $n \times 20$ matrices $\mathbf{Z}^{(j)}$ used by Jiang and Rao (2003) in their simulation. Here $\mathbf{I}_m$ is the $m \times m$ identity matrix, $\mathbf{1}_m$ is the m-vector of ones and $\otimes$ is the Kronecker product

L_3	L_2	L_1
$\mathbf{Z}^{(1)} = \mathbf{I}_{20} \otimes \mathbf{1}_{20} \otimes \mathbf{1}_{20} \otimes \mathbf{1}_3$	$\mathbf{Z}^{(4)} = \mathbf{I}_{20} \otimes \mathbf{I}_{20} \otimes \mathbf{1}_{20} \otimes \mathbf{1}_3$	$\mathbf{Z}^{(7)} = \mathbf{I}_{20} \otimes \mathbf{I}_{20} \otimes \mathbf{I}_{20} \otimes \mathbf{1}_3$
$\mathbf{Z}^{(2)} = \mathbf{1}_{20} \otimes \mathbf{I}_{20} \otimes \mathbf{1}_{20} \otimes \mathbf{1}_3$	$\mathbf{Z}^{(5)} = \mathbf{I}_{20} \otimes \mathbf{1}_{20} \otimes \mathbf{I}_{20} \otimes \mathbf{1}_3$	
$\mathbf{Z}^{(3)} = \mathbf{1}_{20} \otimes \mathbf{1}_{20} \otimes \mathbf{I}_{20} \otimes \mathbf{1}_3$	$\mathbf{Z}^{(6)} = \mathbf{1}_{20} \otimes \mathbf{I}_{20} \otimes \mathbf{I}_{20} \otimes \mathbf{1}_3$	

Ibrahim et al. (2011) also undertook a simulation study and considered six different scenarios for the independent cluster model. They considered the model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\Gamma}_i\mathbf{u}_i + \sigma\mathbf{e}_i, \quad j = 1, \ldots, 12, \quad i = 1, \ldots, m \in \{50, 100, 200\},$$

where $\mathbf{X}_i$ is a 12×8 matrix with independent rows $\mathbf{x}_{ij}^T$ and $\mathbf{x}_{ij} \sim \mathcal{N}_8(\mathbf{0}_8, \boldsymbol{\Sigma}_X)$, $\boldsymbol{\Sigma}_X = (0.5^{|r-s|})$, $\mathbf{Z}_i = \mathbf{X}_i$, $\boldsymbol{\Gamma}_i$ is a 8×8 matrix, $\mathbf{u}_i$ is a 8-vector, σ is a nonnegative scalar and $\mathbf{e}_i$ a 12-vector. For the true model they set $\boldsymbol{\beta} = (3, 2, 1.5, \mathbf{0}_5)^T$,

$$\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i^T = \boldsymbol{\Psi}_i = \begin{pmatrix} \boldsymbol{\Psi}_i^* & \mathbf{0}_{3\times5} \\ \mathbf{0}_{5\times3} & \mathbf{0}_{5\times5} \end{pmatrix} \quad \text{with } \boldsymbol{\Psi}_i^* = \begin{pmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix},$$

and $\sigma \in \{1, 3\}$. The full model contains 5 unnecessary sets of random effects in each cluster. The simulation study concluded that for selecting the regression parameters, the SCAD penalty performed best in terms of estimation error and minimizing overfit. For the variance parameters, the ALASSO penalty performed best. In all cases the penalized maximum likelihood estimates performed better than the maximum likelihood estimates from the full model.

Peng and Lu (2012) carried out a simulation using the same setting as Bondell, Krishna and Ghosh (2010) and then their own setting to examine the properties of their shrinkage method in finite sample size settings. The full model was the independent cluster model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\Gamma}_i\mathbf{u}_i + \sigma\mathbf{e}_i, \quad j = 1, \ldots, n_i \in \{10, 20\}, \quad i = 1, \ldots, m \in \{10, 20\},$$

with equal size clusters in each run, $\mathbf{X}_i$ a $n_i \times 5$ matrix of independent standard Gaussian random variables, $\mathbf{Z}_i$ a $n_i \times 4$ matrix with columns equal to the first 4 columns of $\mathbf{X}_i$, $\boldsymbol{\Gamma}_i$ a 4×4 matrix, $\mathbf{u}_i$ a 4-vector, σ a scalar and $\mathbf{e}_i$ an n_i-vector. For the true model, they set $\boldsymbol{\beta} = (1, 0, 1.5, 1, 0)^T$,

$$\boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i^T = \boldsymbol{\Psi}_i = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.354 \\ 0 & 0 & 0 & 0 \\ 0 & 0.354 & 0 & 1 \end{pmatrix}$$

and $\sigma = 1$. In their first simulation using the same setting as Bondell, Krishna and Ghosh (2010), Peng and Lu (2012) compared the effect of using different methods to select the tuning parameters in their method. Their conclusion is that BIC is the best method of selecting their tuning parameters. This conclusion is based on the average percentage of coefficients that are incorrectly estimated to be nonzero, the average percentage of coefficients that are incorrectly estimated to be zero, the average size of the selected model and the probability of identifying the correct model. The first 3 measures are all marginal measures which are less stringent criteria than the probability of identifying the correct model. They use the simulation probability of identifying the correct model to compare their results with the reported results of Bondell, Krishna and Ghosh (2010), that is, without recalculating these estimates for their data. Their method performs very poorly for the smaller sample sizes but well for the larger sample sizes. They used the second simulation setting to compare their parameter estimates in the selected model with the maximum likelihood estimators for the true model and showed that their performance is comparable.

Jiang et al. (2008) illustrated the use of the Adaptive Fence method in two scenarios. The first is the Fay–Herriot model and they showed that if the data generating model is

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \gamma u_i + e_i, \quad i = 1, \ldots, 30,$$

where u_i, $e_i \sim$ independent $\mathcal{N}(0, 1)$, then, for $M_c \subset M_f$, the quantity σ_{M_c,M_f} is completely known. They considered only selection on $\boldsymbol{\beta}$ and did not compare their results with other selection procedures. The true models had $\boldsymbol{\beta}^T = (1, 0, 0, 0, 0)$, $(1, 2, 0, 0, 0)$, $(1, 2, 3, 0, 0)$, $(1, 2, 3, 2, 0)$, $(1, 2, 3, 2, 3)$ and $\gamma = 1$. Jiang et al. (2008) reported simulation results based on 100 runs for a range of choices of the tuning parameter b_n in equation (31). The second scenario is the random intercept regression model

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \gamma u_i + \sigma e_{ij}, \quad j = 1, \ldots, 5, \quad i = 1, \ldots, 100,$$

where $\boldsymbol{\beta}$ is a 5-vector and γ and σ are scalar. They generated $u_i \sim$ independent $\mathcal{N}(0, 1)$, $\mathbf{e}_i = (e_{i1}, \ldots, e_{i5})^T \sim$ independent $\mathcal{N}(\mathbf{0}_5, (1 - \zeta)\mathbf{I}_5 + \zeta\mathbf{J}_5)$, $\zeta \in \{0, 0.2, 0.5, 0.8\}$ and $x_{ij2}, \ldots, x_{ij5} \sim$ independent $\mathcal{N}(0, 1)$ so $\mathbf{x}_{ij}^T = (1, x_{ij2}, \ldots, x_{ij5})$. The true models had $\boldsymbol{\beta}^T = (2, 0, 0, 4, 0)$, $(2, 9, 0, 4, 8)$, $(1, 2, 3, 2, 3)$ and $\gamma = \sigma = 1$. As a lack-of-fit measure they choose the residual sum of squares and showed that the Adaptive Fence chooses the true model in all 100 simulation runs. In comparison, the performance of two GIC type criteria as introduced in Jiang and

Rao (2003) is less impressive, particularly when the true model is the full model and ζ is large.

Jiang, Nguyen and Rao (2009) reported limited simulation results using the Simplified Adaptive Fence for a different random intercept regression model, but again selection only focused on the regression parameters $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_6)$. The model was

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \gamma u_i + \sigma e_{ij}, \quad j = 1, \ldots, n_i \sim \mathcal{P}(3), \\ i = 1, \ldots, m \in \{10, 15\},$$

where γ and σ are scalar, $u_i, e_{ij}, x_{ij1}, x_{ij2} \sim$ independent $\mathcal{N}(0, 1)$ and $\mathbf{x}_{ij}^T = (1, x_{ij1}, x_{ij2}, x_{ij1}^2, x_{ij2}^2, x_{ij1}x_{ij2})$. A total of 100 simulation runs were run with two true models with $\boldsymbol{\beta}^T = (1, 1, 1, 0, 0, 0)$ and $\boldsymbol{\beta} = \mathbf{1}_6$ (i.e., the full model) and $\gamma = \sigma = 1$. As a performance measure the number of times the true model was selected was used and the reported results only showed the selection probabilities, which seem to be good, without comparing them to other selection criteria.

Nguyen and Jiang (2012) evaluated the Restricted Fence method in a simulation based on data from a bone turnover study. The setting is the random intercept regression model

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \gamma u_i + \sigma e_{ij}, \\ j = 1, \ldots, 3, i = 1, \ldots, m = \{50, 100, 150\},$$

where $\boldsymbol{\beta}$ is a 30-vector and γ and σ are scalar. They generated $u_i \sim$ independent $\mathcal{N}(0, 1)$, $\mathbf{e}_i = (e_{i1}, \ldots, e_{i3})^T \sim$ independent $\mathcal{N}(\mathbf{0}_3, \mathbf{I}_3)$, one explanatory variable corresponding to dietary group as binary and the remaining explanatory variables as independent Gaussian variables with means and variances the same as those for the variables in the bone turnover study. The true models had 7 variables in the mean and $\gamma = \sigma = 1$. For the Restricted Fence, the variables were divided into 4 subsets of 7 or 8 variables using biological considerations and 100 bootstrap samples used in each selection. Nguyen and Jiang (2012) compared the Restricted Fence with particular backward and forward search implementations of information criteria. They showed that the Restricted Fence underfits when the sample size is small but performs well when the sample size increases. They found that the information criteria tend to overfit and that BIC performed best of the information criteria.

ACKNOWLEDGMENTS

This research was supported by an Australian Research Council discovery project grant. We thank two referees and an Associate Editor for their reviews which have lead to an improved paper.

REFERENCES

AHN, M., ZHANG, H. H. and LU, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statist. Sinica* **22** 1539–1562.

AITKIN, M., LIU, C. C. and CHADWICK, T. (2009). Bayesian model comparison and model averaging for small-area estimation. *Ann. Appl. Stat.* **3** 199–221. MR2668705

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) (B. N. Petrov and F. Csaki, eds.) 267–281. Akadémiai Kiadó, Budapest. MR0483125

BONDELL, H. D., KRISHNA, A. and GHOSH, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66** 1069–1077. MR2758494

BOZDOGAN, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52** 345–370. MR0914460

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference*: *A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York. MR1919620

BURNHAM, K. P. and WHITE, G. C. (2002). Evaluation of some random effects methodology applicable to bird ringing data. *J. Appl. Stat.* **29** 245–266. MR1881055

CHEN, Z. and DUNSON, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59** 762–769. MR2025100

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321. MR1379473

CLAESKENS, G. and HJORT, N. L. (2008). *Model Selection and Model Averaging*. Cambridge Univ. Press, Cambridge. MR2431297

DIMOVA, R. B., MARKATOU, M. and TALAL, A. H. (2011). Information methods for model selection in linear mixed effects models with application to HCV data. *Comput. Statist. Data Anal.* **55** 2677–2697. MR2802345

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, Y. and LI, R. (2012). Variable selection in linear mixed effects models. *Ann. Statist.* **40** 2043–2068.

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

FANG, Y. (2011). Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models. *J. Data Sci.* **9** 15–21. MR2828946

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019

FIELD, C. A., PANG, Z. and WELSH, A. H. (2010). Bootstrapping robust estimates for clustered data. *J. Amer. Statist. Assoc.* **105** 1606–1616. MR2796575

FIELD, C. A. and WELSH, A. H. (2007). Bootstrapping clustered data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 369–390. MR2323758

FOSTER, S. D., VERBYLA, A. P. and PITCHFORD, W. S. (2007). Incorporating LASSO effects into a mixed model for quantitative trait loci detection. *J. Agric. Biol. Environ. Stat.* **12** 300–314. MR2370788

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

GELMAN, A., ROBERT, C. P. and ROUSSEAU, J. (2010). Do we need an integrated Bayesian/likelihood inference? Available at arXiv:1012.2184v1.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

GREVEN, S. and KNEIB, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97** 773–789. MR2746151

HAN, C. and CARLIN, B. P. (2001). Markov Chain Monte Carlo methods for computing Bayes factors: A comparative review. *J. Amer. Statist. Assoc.* **96** 1122–1132.

HANNAN, E. J. and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195. MR0547244

HENDERSON, C. R. (1950). Estimation of genetic parameters. *Ann. Math. Statist.* **21** 309–310.

HODGES, J. S. and SARGENT, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **88** 367–379. MR1844837

HOETING, J. A., DAVIS, R. A., MERTON, A. A. and THOMPSON, S. E. (2006). Model selection for geostatistical models. *Ecolog. Appl.* **16** 87–98.

IBRAHIM, J. G., ZHU, H., GARCIA, R. I. and GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67** 495–503. MR2829018

JIANG, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York. MR2308058

JIANG, J., LUAN, Y. and WANG, Y.-G. (2007). Iterative estimating equations: Linear convergence and asymptotic properties. *Ann. Statist.* **35** 2233–2260. MR2363970

JIANG, J., NGUYEN, T. and RAO, J. S. (2009). A simplified adaptive fence procedure. *Statist. Probab. Lett.* **79** 625–629. MR2662277

JIANG, J., NGUYEN, T. and RAO, J. S. (2011). Invisible fence methods and the identification of differentially expressed gene sets. *Stat. Interface* **4** 403–415. MR2834974

JIANG, J. and RAO, J. S. (2003). Consistent procedures for mixed linear model selection. *Sankhyā* **65** 23–42. MR2016775

JIANG, J., RAO, J. S., GU, Z. and NGUYEN, T. (2008). Fence methods for mixed model selection. *Ann. Statist.* **36** 1669–1692. MR2435452

JONES, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Stat. Med.* **30** 3050–3056. MR2851399

KUBOKAWA, T. (2011). Conditional and unconditional methods for selecting variables in linear mixed models. *J. Multivariate Anal.* **102** 641–660. MR2755021

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LIANG, H., WU, H. and ZOU, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95** 773–778. MR2443190

LISKI, E. P. and 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,* 18–20 *August,* 2008 (J. Heikkonen et al., eds.).

MCCULLOCH, C. E. (2003). *Generalized Linear Mixed Models. NSF-CBMS Regional Conference Series in Probability and Statistics* **7**. IMS, Beachwood, OH. MR1993816

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

MOODY, J. E. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems* 4 (J. E. Moody, S. J. Hanson and R. P. Lippmann, eds.) 847–854. Morgan Kaufmann, San Mateo.

MÜLLER, S. and WELSH, A. H. (2005). Outlier robust model selection in linear regression. *J. Amer. Statist. Assoc.* **100** 1297–1310. MR2236443

MÜLLER, S. and WELSH, A. H. (2009). Robust model selection in generalized linear models. *Statist. Sinica* **19** 1155–1170. MR2536149

MÜLLER, S. and WELSH, A. H. (2010). On model selection curves. *Intnl. Statist. Reviews* **78** 240–256.

MURATA, N., YOSHIZAWA, S. and AMARI, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Trans. Neural. Netw.* **5** 865–872.

NGUYEN, T. and JIANG, J. (2012). Restricted fence method for covariate selection in longitudinal studies. *Biostatistics* **13** 303–314.

NI, X., ZHANG, D. and ZHANG, H. H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics* **66** 79–88. MR2756693

PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58** 545–554. MR0319325

PAULER, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85** 13–27. MR1627254

PAULER, D. K., WAKEFIELD, J. C. and KASS, R. E. (1999). Bayes factors and approximations for variance component models. *J. Amer. Statist. Assoc.* **94** 1242–1253. MR1731486

PENG, H. and LU, Y. (2012). Model selection in linear mixed effect models. *J. Multivariate Anal.* **109** 109–129. MR2922858

PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and Splus*. Springer, New York.

POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. MR2917961

PU, W. and NIU, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. *J. Multivariate Anal.* **97** 733–758. MR2236499

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge. MR1438788

RISSANEN, J. (2007). *Information and Complexity in Statistical Modeling*. Springer, New York. MR2287233

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. *Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720

SAVILLE, B. R. and HERRING, A. H. (2009). Testing random effects in the linear mixed model using approximate Bayes factors. *Biometrics* **65** 369–376. MR2751460

SAVILLE, B. R., HERRING, A. H. and KAUFMAN, J. S. (2011). Assessing variance components in multilevel linear models using approximate Bayes factors: A case-study of ethnic disparities in birth weight. *J. Roy. Statist. Soc. Ser. A* **174** 785–804. MR2829038

SCHELLDORFER, J., BÜHLMANN, P. and VAN DE GEER, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scand. J. Stat.* **38** 197–214. MR2829596

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHANG, J. and CAVANAUGH, J. E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Comput. Statist. Data Anal.* **52** 2004–2021. MR2418486

SNIJDERS, T. A. B. and BOSKER, R. J. (1999). *Multilevel Analysis*: *An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications, London.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 583–639. MR1979380

SRIVASTAVA, M. S. and KUBOKAWA, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *J. Multivariate Anal.* **101** 1970–1980. MR2671195

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39** 44–47. MR0501454

SUGIURA, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. A* **7** 13–26.

TAKEUCHI, K. (1976). Distribution of informational statistics and a criterion for model fitting. *Suri-Kagaku* **153** 12–18. (in Japanese.)

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 273–282. MR2815776

VAIDA, F. and BLANCHARD, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92** 351–370. MR2201364

WANG, D., ESKRIDGE, K. M. and CROSSA, J. (2011). Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. *J. Agric. Biol. Environ. Stat.* **16** 170–184. MR2818543

WU, H. and ZHANG, J.-T. (2002). Local polynomial mixed-effects models for longitudinal data. *J. Amer. Statist. Assoc.* **97** 883–897. MR1941417

YE, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131. MR1614596

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443