Computational Statistics and Data Analysis 117 (2018) 154–161

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A scalable and efficient covariate selection criterion for mixed effects regression models with unknown random effects structure

Radu V. Craiu [a,*], Thierry Duchesne [b]

[a] Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada
[b] Département de mathématiques et de statistique, Université Laval, Québec City, Québec G1V 0A6, Canada

ARTICLE INFO

Article history:
Received 22 March 2017
Received in revised form 27 July 2017
Accepted 28 July 2017
Available online 18 August 2017

Keywords:
Akaike information criterion
Generalized linear mixed model
h-likelihood
Random coefficient model
Two-stage estimation
Variable selection

ABSTRACT

A new model selection criterion for mixed effects regression models is introduced. The criterion is computable even when the model is fitted with a two-step method or when the structure and the distribution of the random effects are unknown. The criterion is especially useful in the early stage of the model building process when one needs to decide which covariates should be included in a mixed effects regression model, but has no knowledge of the random effect structure. This is particularly relevant in substantive fields where variable selection is guided by information criteria rather than regularization. The calculation of the criterion requires only the evaluation of cluster-level log-likelihoods and does not rely on heavy numerical integration. Theoretical and numerical arguments are used to justify the method and its usefulness is illustrated by analysing data from a youth behaviour study.

1. Introduction

Studies where a large number of observations are collected for each experimental unit, or cluster, are quite common. For instance, in behavioural ecology animals that wear GPS collars are tracked and data for each individual are collected every hour for months or years; in marketing studies banks record every credit card transaction made by a client; in some epidemiological studies data are collected on physicians who each treat a large number of patients; in criminology, data are recorded at every contact of a repeat offender with the justice system. In the social study that we use to illustrate our method, a large number of students are surveyed in a number of secondary high schools. In many instances where such data are collected, analysts will account for the dependence within each cluster by fitting a mixed effects regression model. In the construction of the latter an important and early step concerns selecting the covariates that are included in the model.

The importance of variable selection has been recognized in statistics and there is a vast body of work devoted to developing criteria for this problem (e.g., see the book of Burnham and Anderson, 2002). Traditionally, the Akaike information criterion (AIC) introduced in the foundational work of Akaike (1970) along with its small sample corrections (Hurvich and Tsai, 1989; Cavanaugh, 1997), and the Bayesian Information Criterion (BIC), introduced by Schwarz (1978), have been among the first methods used to select the covariates in regression models with fixed effects. All these are special cases of the Generalized Information Criterion (GIC) (Nishii, 1984; Shibata, 2005; Rao and Wu, 1989) where the aim is to find the model M that minimizes

$$-\mathcal{L}(M) + \lambda|M|, \tag{1}$$

* Correspondence to: University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada.
E-mail address: craiu@utstat.toronto.edu (R.V. Craiu).

http://dx.doi.org/10.1016/j.csda.2017.07.011
0167-9473/

where $\mathcal{L}(M)$ is a measure of fit and $\lambda|M|$ is the penalty incurred by a model with size $|M|$. The GIC proposed by Rao and Wu (1989) is a strongly consistent variable selection criterion with a flexible penalty function.

The introduction of mixed effects models required new strategies for selecting both the fixed and the random effects. In this context, whether the inferential focus is on marginal or conditional model parameters becomes relevant as these two scenarios require separate treatments. While in the former case one could use the traditional criteria to select the covariates in the model, the latter considers the choice of covariates conditional on random effects. In Vaida and Blanchard (2005) the authors proposed the conditional AIC (cAIC) for situations in which the inferential focus is on cluster-specific parameters. Subsequently, the cAIC for linear mixed models has been further expanded by Liang and Wu (2008); Greven and Kneib (2010) and Saefken et al. (2014) who account for the estimation of variance parameters and by Lian (2012) and Donohue et al. (2011) who have extended cAIC to generalized linear mixed models (GLMM) and survival models with random effects. Yu et al. (2013) have proposed a further adjustment for cAIC in GLMM when the variance components must be estimated. An alternative BIC suitable for mixed effects models has been proposed by Delattre et al. (2014). In a departure from classical approaches, Jiang et al. (2008) propose a method in which incorrect models are fenced off and the best model is selected from the remaining ones. An excellent review of the methods briefly discussed here and others can be found in Müller et al. (2013).

Our current contribution for a new criterion is motivated by GLMM applications in ecology and social sciences where model selection is traditionally based on information criteria and not on regularization methods. Moreover, in these fields little is known about the structure of the random effects *a priori* and numerical approximations of the marginal likelihood may be challenging due to model and data size (Craiu et al., 2011; Molenberghs et al., 2011). The new criterion is intended as a first covariate filter in the early stage of the analysis. Given this aim, it is important that the proposed criterion is computable without the need to specify the random effect structure. After this initial stage, other methods such as the cAIC can be exploited to search in the smaller model space.

The criterion developed here is suitable for "partitioned data" methods (sometimes referred to as "divide-and-conquer" approaches) that have been proposed to fit mixed effects models when the data are large or have a complex structure. Such methods include the two-stage approach of Korn and Whittemore (1979) and Stiratelli et al. (1984), the CREML method of Chervoneva et al. (2006), the two-step method of Craiu et al. (2011) or the pseudo-likelihood approach of Molenberghs et al. (2011). All these methods have in common that they fit separate simple models to each element of a partition of the data and then suitably unify the analyses for these simple models to produce inference for the global mixed effects model.

In this paper we focus on deriving a criterion for filtering the potential covariates for use in a standard GLMM as described, for instance, in Chapter 3 of Jiang (2007). The proposed criterion, called meanAIC, is easy to compute and it does not require the specification of the random effects structure. The two-stage estimation methods mentioned assume that none of the covariates are constant in a cluster and this is also necessary for the validity of meanAIC. We give a theoretical development of meanAIC along with heuristic arguments that justify it. Our simulation study shows that the proposed criterion exhibits good finite sample performance.

The remainder of the paper is organized as follows. Section 2 presents the data and model. The new criterion is developed and justified in Section 3. The simulation study is presented in Section 4 and a data illustration forms Section 5. The paper concludes with a discussion and ideas for future work.

2. Data and model

2.1. Population and data

We consider a population of independent clusters, each containing a number of individual observations of the form $(Y, x_1, \ldots, x_p)$ with Y being a response variable and $x_1, \ldots, x_p$ potential explanatory variables. We assume that the distribution of Y given $x_1, \ldots, x_p$ is given by a generalized linear model whose regression coefficients may vary from cluster to cluster. In order to have identifiability of all model parameters, none of the explanatory variables can be constant over a cluster. The statistical model described below will assume that all the responses in the same cluster share some commonality that makes them dependent. We assume that there are K clusters and n_i data points in each cluster, $1 \leq i \leq K$.

2.2. Generalized linear mixed model (GLMM)

Let $Y_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ be the response vector for cluster i and $X_i = (x_{0i}, x_{i1}, \ldots, x_{ir})$ be the corresponding covariate matrix, with $x_{ik} = (x_{i1k}, \ldots, x_{in_ik})^\top$, $k = 0, \ldots, r$ and x_{0i} an n_i-vector with all entries equal to 1. Throughout the paper the value of the kth covariate for the jth individual in the ith cluster will be denoted x_{ijk}. The context will clarify whether x_{ij} refers to the jth row of X_i (of length r) or x_{ik} refers to the k th column of X_i (of length n_i).

The dependence among observations in cluster i will be captured using the random vector b_i, where $\{b_i \in \mathbf{R}^q \ : \ i = 1, \ldots, K\}$ are assumed to be i.i.d. with cumulative distribution function (cdf) H and probability density function (pdf) h. Throughout the paper we suppose that $q \leq r$. Let $\mathcal{J}$ be a subset of size s of $\{0, \ldots, q\}, Z_i = \{x_{ik}, k \in \mathcal{J}\}$ and $\beta = (\beta_0, \ldots, \beta_r)^\top$. Under our population assumption and sampling scheme: (i) (Y_i, X_i), $i = 1, \ldots, K$, are independent and (ii) for all $1 \leq i \leq K$

$\{Y_{ij} : 1 \leq j \leq n_i\}$ are independent given X_i and b_i, (iii) the distribution of $Y_{ij}|b_i, X_i$ belongs to the exponential family with pdf f_{ij} and

$$\mu_{ij} = E[Y_{ij}|b_i, X_i] = g^{-1}(\beta^\top x_{ij} + b_i^\top z_{ij}), \tag{2}$$

where $x_{ij}^\top$ and $z_{ij}^\top$ denote the jth row of X_i and Z_i, respectively, and g is a known link function. This is the usual GLMM with random regression coefficients (Ch. 3 in Jiang, 2007). Let $\tilde{x}_{ij} = x_{ij} + \tilde{z}_{ij}$, where $\tilde{z}_{ijk} = z_{ijk}$ if $k \in \mathcal{J}$ and 0 otherwise. Similarly, let $\tilde{\beta}_i = \beta_i + \tilde{b}_i$ with $\tilde{b}_{ik} = b_{ik}$ if $k \in \mathcal{J}$ and 0 otherwise. Note that even though $\tilde{b}_i$ and β_i have equal dimension r, it is not assumed that all r covariates have random effects; only q of them do and $\tilde{b}_i$ is padded with zero's in its $r - q$ entries that correspond to covariates that do not have a random effect. Then (2) can be written as

$$\mu_{ij} = E[Y_{ij}|b_i, X_i] = g^{-1}(\tilde{\beta}_i^\top \tilde{x}_{ij}), \tag{3}$$

and the average conditional log-likelihood contribution in cluster i can be written as

$$\bar{\ell}_{n_i}(\tilde{\beta}_i) = n_i^{-1} \sum_{j=1}^{n_i} \log f_{ij}(Y_{ij}; \tilde{\beta}_i). \tag{4}$$

When one wishes to keep all r covariates in the model, H is the multivariate normal with mean 0 and variance matrix D and the subset $\mathcal{J}$ and the structure of the matrix D are known, then inference methods for β and D, as well as predictions for b_i, are widely available in standard software. They are typically based on standard maximum likelihood, residual maximum likelihood, penalized quasi-likelihood or Bayesian methods. However, in many practical situations all r covariates are not required and the random effect structure (subset $\mathcal{J}$) and distribution (H) are not known. In the following section we will derive a criterion that can guide the selection of covariates without having to specify either $\mathcal{J}$ or H and that can be computed when inference about β is performed via a two-step method.

3. The meanAIC criterion

When fitting a mixed effects model, Vaida and Blanchard (2005) and Greven and Kneib (2010) argue that the selection of fixed effects and other marginal population parameters can be performed using the marginal Akaike information criterion (henceforth denoted mAIC) which is based on the maximized marginal log-likelihood. The latter may become numerically cumbersome when the dimension s of the random effects is large or when the data are massive, although fast and stable Laplace approximations of high dimensional integrals are currently available (e.g., see TMB package in R by Kristensen et al., 2016). A potentially more serious problem is that marginal likelihood calculation requires the specification of the random effect structure $\mathcal{J}$ and distribution H. As we shall see in the real data illustration, the covariates to be selected by mAIC may vary according to assumptions made about the random effect structure.

When considering a partitioned data approach, one realizes that the GLMM specification yields ordinary independent GLMs in each cluster. Many two-stage estimation approaches (e.g., Stiratelli et al., 1984; Renard et al., 2004; Chervoneva et al., 2006; Craiu et al., 2011; Molenberghs et al., 2011) rely on (some of) the following cluster-specific information that is usually easier to obtain than their full data counterparts:

- n_i, the number of observations in cluster i;
- $\hat{\beta}_i = \arg\max_\beta \bar{\ell}_{n_i}(\beta)$, the MLE of β in cluster i;
- $\bar{\ell}_{n_i}(\hat{\beta}_i)$, the maximized average log-likelihood in cluster i;
- $H_i = \frac{\partial^2}{\partial\beta\partial\beta^\top}\bar{\ell}_{n_i}(\beta)\Big|_{\beta=\hat{\beta}_i}$, the Hessian of the average log-likelihood evaluated at $\hat{\beta}_i$.

It is clear that a criterion that would be based only on these four elements from each cluster should be easy to compute in practice.

3.1. Derivation of meanAIC

In our derivations we work under the assumption that all cluster sizes, n_i, are large. This condition is generally satisfied when two-step inference methods are used (Molenberghs et al., 2011). The first step amounts to fitting an ordinary GLM to the data from each cluster. Under our model assumptions, the GLM in cluster i has regression parameter $\tilde{\beta}_i = \beta + \tilde{b}_i$ and because the non-zero elements in $\tilde{b}_i$ are i.i.d. from a continuous distribution, all $\tilde{\beta}_i$ have the same zero elements with probability 1. Let us assume that the true data generating model is the GLMM evaluated at $\beta = \beta^0$ and let $\tilde{\beta}_i^0 = \beta^0 + \tilde{b}_i$, i.e. the true data generating model is among the models considered. Consider the conditional cluster-level Kullback–Leibler divergence for a GLM with parameter β in cluster i,

$$\Delta_i(\beta|\tilde{b}_i) = -\int \log\left\{f_{ij}(Y_{ij}; \tilde{\beta}_i)\right\} f_{ij}(Y_{ij}; \tilde{\beta}_i^0) dY_{ij}, \tag{5}$$

and set $\Delta(\beta|\tilde{b}_1,\ldots,\tilde{b}_K) = K^{-1}\sum_i \Delta_i(\beta|\tilde{b}_i)$. Conditional on the random effects, we consider the cluster-specific empirical versions of the above divergences, respectively given by

$$\hat{\Delta}_i(\beta|\tilde{b}_i) = -n_i^{-1}\sum_j \log\left\{f_{ij}(Y_{ij};\,\tilde{\beta}_i)\right\} = -\bar{\ell}_i(\beta+\tilde{b}_i),$$

$$\hat{\Delta}(\beta|\tilde{b}_1,\ldots,\tilde{b}_K) = K^{-1}\sum_i \hat{\Delta}_i(\beta|\tilde{b}_i) = -K^{-1}\sum_i \bar{\ell}_i(\beta+\tilde{b}_i).$$

The following lemma, whose simple proof is sketched in the appendix, establishes some useful properties of $\Delta_i(\beta|\tilde{b}_1 i)$.

Lemma 3.1. *Under the GLMM assumptions made in Section 2, $\Delta_i(\beta|\tilde{b}_i)$ is a smooth function of β uniformly in $\tilde{b}_i$. Furthermore, $\Delta_i(\beta|\tilde{b}_i) \geq \Delta_i(\beta^0|\tilde{b}_i)$ for all β for any finite value of $\tilde{b}_i$.*

The most important corollary to Lemma 3.1 is that $\Delta(\beta|\tilde{b}_1,\ldots,\tilde{b}_K)$ is minimized at $\beta = \beta^0$ regardless of the values of $\tilde{b}_1,\ldots,\tilde{b}_K$, which implies that it can serve as the basis of a covariate selection criterion. Because $\Delta_i(\beta|\tilde{b}_i)$ contains unknown parameters that must be estimated, we use Remark 1 on p. 242 in Appendix of Linhart and Zucchini (1986) in the form of the following useful lemma:

Lemma 3.2.

$$\hat{E}[\hat{\Delta}_i\{\hat{\beta}(\tilde{b}_i)|\tilde{b}_i\}|\tilde{b}_i] = \hat{\Delta}_i\{\hat{\beta}(\tilde{b}_i)|\tilde{b}_i\} + \frac{r+1}{n_i}, \tag{6}$$

$$\widehat{Var}[\hat{\Delta}_i\{\hat{\beta}(\tilde{b}_i)|\tilde{b}_i\}|\tilde{b}_i] = \frac{(r+1)/2}{n_i^2}. \tag{7}$$

In Eqs. (6) and (7) $\hat{E}$ and $\widehat{Var}$ respectively denote consistent estimators of the mean and variance of the empirical divergences and the RHS of (6) can serve as a model selection criterion (Linhart and Zucchini, 1986, p. 242).

Because the AIC in cluster i is given by $2n_i\{\hat{\Delta}_i\{\hat{\beta}(\tilde{b}_i)|\tilde{b}_i\} + (r+1)/n_i\}$, Eq. (6) establishes that the best model should have smallest AIC in all clusters. Thus, when all the n_i are large we can minimize a weighted average of the cluster-level AICs to identify the model that minimizes (5). According to Eq. (7), the weighted average of the form $\sum_i w_i 2n_i\{\hat{\Delta}_i\{\hat{\beta}(\tilde{b}_i)|\tilde{b}_i\} + (r+1)/n_i\}$ with $0 \leq w_i \leq 1$ and $\sum_i w_i = 1$ with the smallest variance has $w_i = 1/K$, suggesting the following model selection criterion:

$$\text{meanAIC} = \frac{1}{K}\sum_{i=1}^{K} \text{AIC}_i, \tag{8}$$

where $\text{AIC}_i = -2\log f_i(y_i;\,\hat{\beta}_i) + 2(r+1)$. Thus, meanAIC is simply the average of all K AICs obtained when fitting an ordinary GLM separately to each of the K clusters.

The derivation of the meanAIC is based on the Kullback–Leibler distance between the cluster-specific densities defined by the stage 1 estimation procedure. Although it deviates from the canonical elicitations of AIC-type criteria, we believe it is the only way we can bypass the specification of random effects distribution or structure, e.g. identification of covariates with random effects.

The calculation of meanAIC falls within the class of embarrassingly parallel procedures, because its calculation can be performed in parallel by assigning the computation of each cluster-level set of estimators to a separate CPU. The communication between CPUs is minimal, requiring only one average of the cluster-specific AICs. Furthermore, as is the case with variable selection based on any model selection criterion, computing meanAIC for the 2^p possible submodels (or a subset thereof) can also be parallelized.

4. Simulation study

The meanAIC criterion was derived in the previous section as a potentially useful covariate selection tool when the cluster sizes n_i tend to infinity. The present section reports the results of a simulation study whose primary objective is to assess the performance of meanAIC as a covariate screening tool for finite values of n_i. A secondary objective is to compare its efficiency and robustness to that of mAIC that is computed assuming a GLMM with a random intercept. The latter choice is in line with our aim of establishing model selection criteria without having to spell out the random effects structure; under this constraint, the only specification that is common to all mixed effects submodels is the one with only a random intercept.

4.1. Study design

We considered simulation scenarios for Gaussian, logistic and Poisson mixed effects models and generated data with $K \in \{20, 200, 1000\}$ independent clusters. Our primary objective and the theoretical justification of meanAIC suggest that we consider larger values of n_i, but we also examine the performance of meanAIC when n_i is small. We ran simulations

where n_i was equal to 20, 80 or 320 in all clusters; additional simulations where n_i varied from cluster to cluster within the same dataset yielded results similar to those reported below and are summarized in Appendix A. Our simulation design is similar to other designs where model selection criteria are investigated (Yu et al., 2013). We simulate two covariates, x_{ij1} i.i.d. Bernoulli(0.5) and x_{ij2} i.i.d. uniform(0, 1). In the simulations, the link functions are canonical, i.e. log, logit and identity for Poisson, logistic and Gaussian response models, respectively. The true linear predictors used depend on x_{ij1}, a random intercept b_{0i} and a random coefficient b_{1i}. For Poisson and Gaussian we have used the linear predictor

$$\eta = 0.3 + b_{0i} + (\beta_1 + b_{1i})x_{ij1}, \tag{9}$$

while for the logistic model we have used

$$\eta = -0.3 + b_{0i} + (\beta_1 + b_{1i})x_{ij1}, \tag{10}$$

for $1 \leq i \leq K$ and $1 \leq j \leq n_i$ (we used (10) instead of (9) to avoid too large a proportion of 1's in the response). Our simulations consider two values for the fixed effects $\beta_1 \in \{0.2, 0.4\}$ and assumed the two random effects to be independently drawn from the same zero-mean normal distribution, but with possibly different variances, more precisely b_{ui} has variance σ_u^2 which can take values in the set $\{0.005, 0.15, 0.3, 0.8, 1.5\}$ for all $u \in \{0, 1\}$. We also simulate random effects from gamma and t distributions in order to investigate robustness of the criteria against asymmetric and fat-tailed distributions, respectively. These results are quite similar to those reported in Table 1 and are summarized in Appendix A.

For each sample generated, all four submodels were considered: (i) the null model; (ii) the true model with x_1 only; (iii) the model with x_2 only, and (iv) the model with both x_1 and x_2. For each sample the following model selection criteria were computed: the proposed meanAIC and the mAIC obtained by fitting a Poisson GLMM with a random intercept to the entire sample.

The meanAIC is calculated using the output of the `glm` function from R applied to each cluster separately. Maximum likelihood fitting of the random intercept GLMM was implemented with the function `glmer` from the R package `lme4`. The calculation for meanAIC was approximately three times faster than the mAIC, e.g., computation for one sample required 0.05 s of CPU time for meanAIC and 0.17 s of CPU time for the mAIC on a Lenovo X230 tablet PC with Intel Core i7-3520M CPU at 2.90 GHz, 8 GB of RAM and running on the 64 bit version of Windows 7.

4.2. Results

Each simulation scenario was replicated 500 times and the proportion of correct covariate selection decisions are reported for Poisson regression in Table 1 for $n_i = 80$ and Tables A.7–A.9 for $n_i = 20$ and K equal to 20, 200 and 1000, respectively. The simulations show that mAIC dominates meanAIC only when cluster-sizes are small and the random coefficient has a small variance. This is not surprising given that the mAIC criterion is computed assuming that only the intercept is random. Therefore, when averaging over the distribution of the random intercepts, mAIC pools all the cluster data and benefits from the resulting larger sample size, unlike meanAIC which relies on cluster-level data sizes that are not large enough to enable it to detect the small fixed effects. As soon as cluster-specific coefficients are not all small, e.g. the random coefficients have moderate variance, the accuracy of meanAIC dramatically increases. For example, when $n_i = 80$ and the random effect variance σ_1^2 increases from 0.005 to 0.15 we see that the percentage of correct decisions for meanAIC increases from about 12% to about 90% for different values of σ_0^2. This trend is not replicated by mAIC that does not seem to benefit from larger values of σ_1^2. When $n_i = 320$ both meanAIC and mAIC have a good performance, with meanAIC dominating mAIC for all values of σ_0^2 and σ_1^2 considered. It is also worth noting that for the same value of n_i, increasing the number of clusters improves the performance of meanAIC, e.g. when $n_i = 20$, $\sigma_0^2 = 0.005$ and $\sigma_1^2 = 0.3$ the percentage of correct decisions is 48.6%, 62% and 75% for K equal to 20, 200 and 1000, respectively.

To see if the performance of meanAIC with $\sigma_1^2 = 0.005$ and $n_i = 80$ improves as the fixed effect size β gets larger, we replicate the simulation scenarios involving these values of n_i and σ_1^2, but with a larger $\beta = 0.4$. The results are summarized in Table 2 and show that under these settings meanAIC outperforms mAIC.

Tables A.10 and A.11 show simulation results for the binary response (with log link changed to logit and $\beta_0 = -0.3$). Unsurprisingly, due to the reduction in information contained in the response variable, meanAIC and mAIC require larger samples sizes and random effects variances than in the Poisson response case to achieve the same levels of accuracy. We also note that the comparison between meanAIC and mAIC follows patterns similar to the Poisson response with meanAIC outperforming mAIC when $\sigma_1^2 > 0.5$.

Finally, we wanted to compare the performance of the proposed criterion with cAIC and BIC_h (Delattre et al., 2014), which have been proposed as good model selection criteria for mixed models once a random effects structure has been specified. The computation times for cAIC in generalized mixed effects models with the settings used in this simulation are prohibitively high. However, we applied cAIC to the linear mixed effects model with $\sigma_0^2 = 0.3$ $K = 20$ and $n_i \in \{40, 80, 160\}$ and the results are shown in Table A.12. We get that cAIC is comparable to mAIC, perhaps marginally better, and that meanAIC suffers when $\sigma_1^2 = 0$ but dominates the other criteria when $\sigma_1^2 \geq 0.15$. The BIC_h criterion does not perform as well as its counterparts.

In the next section we consider the analysis of a socio-economic study of youth behaviour and compare meanAIC with mAIC and cAIC.

Table 1
Poisson response: Proportion of the 500 replications where each criterion picked the true model. Regression coefficient $\beta = 0.2$. Random effects follow a normal with mean 0 and variance σ_u^2, $u = 0, 1$.

σ_0^2	σ_1^2	$n_i = 80\ \forall i$		$n_i = 320\ \forall i$	
		meanAIC	mAIC	meanAIC	mAIC
0.005	0.005	0.140	0.790	0.878	0.828
0.005	0.15	0.884	0.740	0.994	0.828
0.005	0.3	0.980	0.778	0.994	0.830
0.005	0.8	0.992	0.816	0.994	0.774
0.005	1.5	0.994	0.772	0.996	0.794
0.15	0.005	0.166	0.792	0.876	0.846
0.15	0.15	0.882	0.774	0.996	0.818
0.15	0.3	0.988	0.786	0.998	0.838
0.15	0.8	0.992	0.788	0.996	0.794
0.15	1.5	0.996	0.746	0.992	0.762
0.3	0.005	0.116	0.802	0.932	0.844
0.3	0.15	0.894	0.732	0.998	0.836
0.3	0.3	0.988	0.788	0.990	0.822
0.3	0.8	0.998	0.754	0.990	0.798
0.3	1.5	0.998	0.730	0.992	0.738

Table 2
Poisson response: Proportion of the 500 replications where each criterion picked the true model. Regression coefficient $\beta = 0.4$. Random effects follow a normal with mean 0 and variance σ_u^2, $u = 0, 1$.

σ_0^2	σ_1^2	$n_i = 80\ \forall i$	
		meanAIC	mAIC
0.005	0.005	0.866	0.806
0.15	0.005	0.910	0.856
0.3	0.005	0.928	0.854

5. Analysis of youth behaviour data

We illustrate the ability of meanAIC to identify those covariates that can have strong cluster-specific effects, but small marginal effects. These are typically covariates corresponding to important random effects variances that have modest fixed regression coefficients. The simulation studies performed in the previous section suggest that under this scenario mAIC, cAIC and meanAIC are likely to select different models. The data come from the study of Beauvais and Swaim (2015) on alcohol use among young American Indians in US schools. The individual observations are responses by students to a survey, and the clusters are the schools the students belong to. The sample we analyse consists of data from the 25 largest clusters with an average size of 170 students. The response variable is the number of times a student had more than 5 alcoholic drinks in less than two hours during the last two weeks.

All models considered will include four personal covariates (gender, age, whether the student likes school and whether the student is proud of him/herself), two covariates related to friends influence (whether some of the friends have ever been suspended from school and whether friends ask the student to get drunk some or a lot). The four candidate models differ due to presence/absence of two family-related covariates (family is a lot likely to stop the student from getting drunk, and the family members argue a lot). A more detailed description of the model covariates and their corresponding values is presented in Appendix B.

The models are fitted using a Poisson GLMM with log link, with each school as a cluster. The meanAIC is obtained by fitting an ordinary Poisson GLM separately to each cluster data using the `glm` function in R. The mAIC is computed by fitting the marginal model using the `glmer` function from package `lme4` in R (Bates et al., 2015), and cAIC is computed by applying the `cAIC` function of the `cAIC4` package (Saefken et al., 2014b) to the object produced by `glmer`. Unlike meanAIC, to fit the GLMM needed for mAIC and cAIC, one needs to specify a random effects structure. We consider two such structures: i) a random intercept only that will yield mAIC and cAIC values denoted mAIC.RI and cAIC.RI, respectively, and ii) a random intercept and a random coefficient for the "family members argue a lot" covariate that will yield mAIC and cAIC values, respectively denoted mAIC.RC and cAIC.RC. A summary of these GLMM fits is provided in Appendix B. Clearly, the random coefficient for "family members argue a lot" has a large variance and a small marginal effect. It is thus expected that meanAIC may identify more accurately than mAIC.RI and cAIC.RI the importance of this covariate for the model. The values of mAIC and cAIC for the two GLMM and of meanAIC are reported in Table 3. Simulations showed that meanAIC is better than mAIC.RI at identifying the generating model covariates that had a random coefficient. Based on these results and the magnitude of the random coefficient variance (this variance is highly significant according to the likelihood ratio test described in Section 6.3.2 of Verbeke and Molenberghs (2009)) reported in Appendix B, we believe that "family members argue a lot" should be part of the model. All criteria agree that the family is "likely to stop you from getting drunk" variable should be included in

Table 3
Data analysis: Model selection criteria for all four submodels of interest for the alcohol consumption study. mAIC.RI and cAIC.RI refer to the mAIC and cAIC criteria obtained by fitting a GLMM with random intercept only, mAIC.RC and cAIC.RC denote the mAIC and cAIC of the model with a random intercept and a random coefficient in front of the covariate "family members argue a lot", while meanAIC denotes the meanAIC criterion. The best value for each criterion appears in bold.

Family covariates in model	mAIC.RI	cAIC.RI	mAIC.RC	cAIC.RC	meanAIC
Both	7929.9	7886.5	**7898.5**	**7829.0**	**298.5**
"Lot likely to stop you" only	**7928.4**	**7885.2**	7928.4	7885.2	302.3
"Family members argue a lot" only	8166.2	8120.4	8127.0	8050.9	307.0
None	8166.1	8120.4	8166.1	8120.4	311.3

the model. It is worth pointing out that mAIC and cAIC choose different models depending on the random effects structure assumed, which can be confusing when there is no clear choice for the latter.

6. Discussion

In this paper we set out to develop a new variable selection criterion for GLMM that does not require a specification of the random effects structure. Furthermore, we wanted a criterion computable even when a two-stage estimation procedure is used to fit the model, which usually occurs when the cluster sizes are large enough to make impractical marginal likelihood inference.

We used an h-likelihood based theoretical justification to develop the meanAIC criterion. The implicit assumption is that cluster sizes are large. Simulations were performed under a number of possible combinations of response distribution (Poisson, logistic, Gaussian), cluster size (small and large), random coefficients variance (small, moderate and large), effect size (small and moderate) and different random effects distributions (normal, shifted gamma and t). We compared the ability of meanAIC and of mAIC to identify the true covariate structure. The proposed meanAIC clearly outperformed mAIC for all settings except the case where cluster size, random coefficient variance and fixed effect size were all simultaneously small. The application of these criteria to real data analysis further emphasized the importance of variable selection without specifying the random effects structure.

An important observation due to one of the paper's referees is that the quantities needed to derive meanAIC are also computable in the case of non-linear mixed effects models, as long as the cluster-specific likelihoods can be optimized. This extends considerably the range of applications beyond the GLMM class of models. The performance of the proposed criterion for non-linear regression models will be explored in future work.

Model selection based on comparing all possible submodels is not practical when the number of potential covariates is large. In future work we would like to consider the interplay between meanAIC and regularization-based methods (e.g., Ibrahim et al., 2011; Fan and Li, 2012; Lin et al., 2013), where information criteria are used to set the value of the tuning parameter in the penalty term.

Acknowledgements

We thank the Co-Editor, Ana Colubi, the Associate Editor and two anonymous referees for their comments and suggestions that have greatly improved the paper. This work has been funded by Natural Sciences and Engineering Research Council of Canada individual discovery grants 249547-2012 and 05883-2016 to RVC and TD, respectively.

Appendix A. Theoretical proofs and additional numerical results

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2017.07.011.

References

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Statist. Math. 22, 203–217.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48.

Beauvais, F., Swaim, R., 2015. Drug use among young american indians: Epidemiology and prediction, 1993-2006 and 2009-2013. Ann Arbor, MI: Inter-university Consortium for political and social research. URL http://doi.org/10.3886/ICPSR35062.v3.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Inference: A Practical Information-Theoretic Approach, second ed. Springer-Verlag, New York Inc..

Cavanaugh, J., 1997. Unifying the derivations for the Akaike and corrected Akaike information criteria. Statist. Probab. Lett. 33, 201–208.

Chervoneva, I., Iglewicz, B., Hyslop, T., 2006. A general approach for two-stage analysis of multilevel clustered non-Gaussian data. Biometrics 62, 752–759. http://dx.doi.org/10.1111/j.1541-0420.2005.00512.x.

Craiu, R.V., Duchesne, T., Fortin, D., Baillargeon, S., 2011. Conditional logistic regression with longitudinal follow-up and individual-level random coefficients: A stable and efficient two-step estimation method. J. Comput. Graph. Statist. 20, 767–784.

Delattre, M., Lavielle, M., Poursat, M.A., 2014. A note on bic in mixed-effects models. Electron. J. Statist. 8, 456–475.

Donohue, M., Overholser, R., Xu, R., Vaida, F., 2011. Conditional Akaike information under generalized linear and proportional hazards mixed models. Biometrika 98, 685–700.

Fan, Y., Li, R., 2012. Variable selection in linear mixed effects models. Ann. Statist. 40, 2043–2068.
Greven, S., Kneib, T., 2010. On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika 97, 773–789.
Hurvich, C., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.
Ibrahim, J.G., Zhu, H., Garcia, R.I., Guo, R., 2011. Fixed and random effects selection in mixed effects models. Biometrics 67, 495–503.
Jiang, J., 2007. Linear and Generalized Linear Mixed Models and their Applications. Springer.
Jiang, J., Rao, J., Gu, Z., Nguyen, T., 2008. Fence methods for mixed model selection. Ann. Statist. 36, 1669–1692.
Korn, E.L., Whittemore, A.S., 1979. Methods for analyzing panel studies of acute health effects of air pollution. Biometrics 35, 795–802.
Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., Bell, B.M., 2016. TMB: Automatic differentiation and Laplace approximation. J. Stat. Softw. 70, 1–21. http://dx.doi.org/10.18637/jss.v070.i05.
Lian, H., 2012. A note on conditional akaike information for Poisson regression with random effects. Electron. J. Stat. 6, 1–9.
Liang, H., Wu, H., 2008. A note on the conditional AIC for linear mixed-effects models. Biometrika 95, 773–778.
Lin, B., Pang, Z., Jiang, J., 2013. Fixed and random effects selection by REML and pathwise coordinate optimization. J. Comput. Graph. Statist. 22, 341–355.
Linhart, H., Zucchini, W., 1986. Model Selection. Wiley.
Molenberghs, G., Verbeke, G., Iddi, S., 2011. Pseudo-likelihood methodology for partitioned large and complex samples. Statist. Probab. Lett. 81, 892–901.
Müller, S., Scealy, J.L., Welsh, A.H., 2013. Model selection in linear mixed models. Statist. Sci. 28, 135–167.
Nishii, R., 1984. Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Statist. 12, 758–765.
Rao, C.R., Wu, Y., 1989. A strongly consistent procedure for model selection in a regression problem. Biometrika 76, 369–374.
Renard, D., Molenberghs, G., Geys, H., 2004. A pairwise likelihood approach to estimation in multilevel probit models. Comput. Statist. Data Anal. 44, 649–667.
Saefken, B., Kneib, T., van Waveren, C.S., Greven, S., 2014a. A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. Electron. J. Statist. 8, 201–225.
Saefken, B., Ruegamer, D., with contributions from Sonja Greven,, Kneib, T., 2014b. cAIC4: Conditional Akaike information criterion for lme4. URL https://CRAN.R-project.org/package=cAIC4. r package version 0.2.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.
Shibata, R., 2005. Apprximate efficienct of a selection procedure for the number of regression variables. Biometrika 92, 43–49.
Stiratelli, R., Laird, N., Ware, J.H., 1984. Random-effects models for serial observations with binary response. Biometrics 40, 961–971.
Vaida, F., Blanchard, S., 2005. Conditional Akaike information for mixed effects models. Biometrika 92, 351–370.
Verbeke, G., Molenberghs, G., 2009. Linear Mixed Models for Longitudinal Data, second ed. Springer Science & Business Media, New York.
Yu, D., Zhang, X., Yau, K.K.W., 2013. Information based model selection criteria for generalized linear mixed models with unknown variance component parameters. J. Multivariate Anal. 116, 245–262.

Further reading

Fahrmeir, L., Kaufman, H., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Statist. 13, 342–368.