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 $1 \leq i \leq K$