ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bootstrap estimation of uncertainty in prediction for generalized linear mixed models

Daniel Flores-Agreda [a,b,*], Eva Cantoni [a]

[a] Research Center for Statistics and Geneva School of Economics and Management, Université de Genève, Bd du Pont d'Arve 40 CH-1211 Geneva 4, Switzerland

[b] Operations Department, Faculty of Business and Economics, Université de Lausanne, Bâtiment Anthropole, CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 17 November 2017
Received in revised form 9 August 2018
Accepted 10 August 2018
Available online xxxx

Keywords:
Bootstrap
GLMM
Prediction
Random effects
MSEP
Laplace approximation

ABSTRACT

In the framework of Mixed Models, it is often of interest to provide an estimate of the uncertainty in predictions for the random effects, customarily defined by the Mean Squared Error of Prediction (MSEP). To address this computation in the Generalized Linear Mixed Model (GLMM) context, a non-parametric Bootstrap algorithm is proposed. First, a newly developed Bootstrap scheme relying on random weighting of cluster contributions to the joint likelihood function of the model and the Laplace Approximation is used to create bootstrap replicates of the parameters. Second, these replicates yield in turn bootstrap samples for the random effects and for the responses. Third, generating predictions of the random effects employing the bootstrap samples of observations produces bootstrap replicates of the random effects that, in conjunction with their respective bootstrap samples, are used in the estimation of the MSEP. To assess the validity of the proposed method, two simulation studies are presented. The first one in the framework of Gaussian LMM, contrasts the quality of the proposed approach with respect to: (i) analytical estimators of MSEP based on second-order correct approximations, (ii) Conditional Variances obtained with a Bayesian representation and (iii) other bootstrap schemes, on the grounds of relative bias, relative efficiency and the coverage ratios of resulting prediction intervals. The second simulation study serves the purpose of illustrating the properties of our proposal in a Non-Gaussian GLMM setting, namely a Mixed Logit Model, where the alternatives are scarce.

1. Introduction

In many applications of Mixed Models, it is of interest to provide an "estimate" of the value for the random effects, be it for forecasting purposes or to assess the quality of a particular fit e.g. by performing some sort of residual analysis post-estimation. The process of providing such values is customarily called Point *Prediction* of the Random Effects, a denomination used to state its difference from the *Estimation* of the Model Parameters.

The problem of Prediction of Random Effects has been widely explored in the literature of *Gaussian* GLMM or *Linear Mixed Models* (LMM), a setting for which theoretical results have led to the determination of the *Best Linear Unbiased Predictor* (BLUP) in full knowledge of the model parameters, and its *Empirical* version (EBLUP) when the parameters are estimated. Naturally, this approach has analogues in the Non-Gaussian framework in the form of the *Best Predictor* (BP) and EBP, often

* Corresponding author at: Research Center for Statistics and Geneva School of Economics and Management, Université de Genève, Bd du Pont d'Arve 40 CH-1211 Geneva 4, Switzerland.
E-mail address: daniel.flores.agreda@gmail.com (D. Flores-Agreda).

https://doi.org/10.1016/j.csda.2018.08.006
0167-9473/

Please cite this article in press as: Flores-Agreda D., Cantoni E., Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.006.

approximated by *Conditional Modes* obtained with an Empirical Bayes approach, see e.g Morris (1983) and Tierney and Kadane (1986).

Similarly to the estimation problem, where point estimates are provided alongside their *standard errors* for inferential purposes, it is useful to retrieve a measure of uncertainty of the point predictions e.g. to classify observational units according to "significant" differences in their predicted response or to construct prediction intervals for new observations drawn from a given unit. In LMM, this translates into the computation of the *Mean Squared Error* of the Prediction (MSEP), often estimated by means of second-order correct approximations that take into account the uncertainty due to the parameter estimation, such as those proposed by Kackar and Harville (1984), Prasad and Rao (1990), Datta and Lahiri (2000) and Das et al. (2004). In a more general framework, it is customary to report estimates of the *Conditional Variances* (CV) resulting from the Bayesian outlook on the GLMM, with the addition of corrections that account for the added variability of the estimation of the model parameters, see e.g. Kass and Steffey (1989), Booth and Hobert (1998) and Singh et al. (1998). The computation of these measures could also be undertaken with the use of resampling methods such as the Jackknife approach to the computation of MSEP (Jiang et al., 2002) or the more widespread *Parametric Bootstrap* (PB) method, used to produce estimates of MSEP, see for instance Butar and Lahiri (2003) or to build Prediction Intervals, as seen in Butar and Lahiri (2003), Hall and Maiti (2006), Chatterjee et al. (2008) and Li and Lahiri (2010).

To the best of our knowledge, there are very few proposals that attempt to tackle this problem by means of a non-parametric bootstrap procedure. Moreover, these methods often rely on the resampling of some sort of residuals and predictions of the random effects making their implementation in the Gaussian LMM framework straightforward and intuitive, yet harder to export to the *Generalized* i.e. non-Gaussian setting. Hence, we propose to confront the MSEP estimation by means of a non-parametric Bootstrap method resulting from the adaptation of the *Random Weighted Laplace Bootstrap* (RWLB) (Flores-Agreda, 2017), a scheme having the main advantage of being applicable in the entire class of GLMM. This proposal is compared to adaptations of other schemes such as the so-called *Random Effect Bootstrap* (REB), see e.g. Davison and Hinkley (1997), Carpenter et al. (2003), and Field et al. (2008), and the more widespread Parametric Bootstrap alternatives.

The article is structured as follows: In Section 2, we set up the notation of the GLMM, characterize the special case of LMM (Section 2.1) and summarize the problem of prediction of random effects (Section 2.2). Section 3, contains an overview of two methods for the evaluation and estimation of the uncertainty in prediction namely the approach via the MSEP (Section 3.1) and the Empirical Bayes approaches relying on CV (Section 3.2). We briefly review some resampling schemes for LMM in Section 3.3, highlight or propose adaptations to the estimation of uncertainty in the Non-Gaussian context and formulate our proposals based on the RWLB scheme. Finally, Section 4 contains two simulation studies as a basis of comparison of the different methods, one carried on a LMM (Section 4.1) and second one in a Mixed Logit context (Section 4.2).

2. Model and notation

Let $i = 1, \ldots, n$ denote the index of the *observational unit* and $j = 1, \ldots, n_i$ the index for an observation within this unit. Write $\boldsymbol{\theta} = [\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T]^T$ $(d \times 1)$ to denote the vector of model parameters, where $\boldsymbol{\beta}(p \times 1)$ represents the *fixed effect* parameters and $\boldsymbol{\sigma}(s \times 1)$ contains the parameters associated with the random effects sometimes referred to as *Variance Components* and $d = p + s$. Write y_{ij} to denote the observed outcomes, assumed to be independently drawn from an exponential family when conditioned on a vector of covariates $\mathbf{x}_{ij}(p \times 1)$ and a vector of random effects $\boldsymbol{\gamma}_i(q \times 1)$ following a $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Delta}_\sigma)$ distribution, endowed with a positive-definite symmetric covariance matrix $\boldsymbol{\Delta}_\sigma$. For notation simplicity, we will consider the reparametrization $\boldsymbol{\gamma}_i = \mathbf{D}_\sigma \mathbf{u}_i$ resulting from the Cholesky decomposition of $\boldsymbol{\Delta}_\sigma = \mathbf{D}_\sigma \mathbf{D}_\sigma^T$ where $\mathbf{u}_i$ are multivariate standard normal vectors. Let μ_{ij} denote the conditional expectation of the outcome, $\mathbf{z}_{ij}(q \times 1)$ a *design* vector for the random effects and $\eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{D}_\sigma \mathbf{u}_i$ the *Linear Predictor*. With g, representing a monotonic *link* function that maps the linear predictor and the conditional expectation of the outcome, the GLMM can be formulated as follows:

$$g\left(\mu_{ij}\right) = \eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{D}_\sigma \mathbf{u}_i.$$

Let f denote the Probability Density Function (PDF) or Probability Mass Function (PMF) evaluated at the observed outcomes y_{ij}, conditioned on vectors $\mathbf{x}_{ij}$, $\mathbf{u}_i$ and assumed to follow conditional exponential families :

$$f_\theta\left(y_{ij}|\mathbf{u}_i\right) := f\left(y_{ij}|\mathbf{u}_i, \mathbf{x}_{ij}; \boldsymbol{\theta}, \phi\right) = \exp\left[\frac{y_{ij}\xi_{ij} - b\left(\xi_{ij}\right)}{\phi} + c\left(y_{ij}, \phi\right)\right]$$

for ϕ a nuisance *dispersion* parameter, $\xi_{ij} = \xi(\eta_{ij})$ the so-called *canonical parameter* (when ϕ is known) and with b, the *cumulant function*, characterizing the conditional means and variances of the outcomes, e.g. $\mu_{ij} = \mathbb{E}[Y_{ij}|\mathbf{u}_i] = b'(\xi_{ij})$ and $v_{ij} = v(\mu_{ij}) = \mathbb{V}ar[Y_{ij}|\mathbf{u}_i] = \phi b''(\xi_{ij})$ and c denoting a specific function. In what follows, and without loss of generality, we consider the link g to be the *canonical link*, in other words $\mu_{ij} = b'(\eta_{ij})$, implying $\xi_{ij} = \eta_{ij}$.

The expressions of the marginal PDF/PMF $f_\theta(y_{ij})$ are obtained after integration of the random effects from the joint distribution of $[y_{ij}, \mathbf{u}_i^T]^T$. Using φ to denote the density of the standard multivariate normal random vector $\mathbf{u}_i$ and with the assumptions on the independence between y_{ij} conditional on $\mathbf{u}_i$, the Likelihood contributions are given by multivariate integrals of the form:

$$\mathcal{L}_i\left(\boldsymbol{\theta}\right) := \int_{\mathbb{R}^q} \left[\prod_{j=1}^{n_i} f_\theta\left(y_{ij}|\mathbf{u}_i\right)\right] \varphi\left(\mathbf{u}_i\right) \mathrm{d}\mathbf{u}_i = \int_{\mathbb{R}^q} L_i\left(\boldsymbol{\theta}, \mathbf{u}_i\right) \mathrm{d}\mathbf{u}_i. \tag{1}$$

Please cite this article in press as: Flores-Agreda D., Cantoni E., Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.006.