Computational Statistics and Data Analysis 56 (2012) 2663–2674

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian variable selection for logistic mixed model with nonparametric random effects

Mingan Yang *

Department of Mathematics, Central Michigan University, Mt. Pleasant, MI 48859, United States

ARTICLE INFO

Article history:
Received 23 May 2011
Received in revised form 19 September 2011
Accepted 21 December 2011
Available online 29 December 2011

Keywords:
Dirichlet process
Nonparametric Bayes
Variable selection
Random effects
Mixed effects model
Stochastic search

ABSTRACT

In analyzing correlated data or clustered data with linear or logistic mixed effects model, one commonly assumes that the random effects follow a normal distribution with mean zero. However, this assumption might not be appropriate in many cases. In particular, substantial violation of normality assumption might potentially impact the subset selection of variables in these models. In this article, we address the problem of joint selection of both fixed and random effects and bias control for random effects in nonparametric settings. An efficient Bayesian variable selection is implemented using a stochastic search Gibbs sampler to allow both fixed and random effects to be dropped effectively out of the model. The approach is illustrated using a simulation study and a real data example.

1. Introduction

In longitudinal studies, logistic mixed models (Drum and McCullagh, 1993; Noortgate and Boeck, 2005) are widely used for clustered binary data to study the relationship between the response and covariates. Generally the random effects are incorporated to account for subject-specific variation and are routinely assumed to follow normal distribution with mean zero. However, this assumption might not be realistic and one might question the validity of inferences of the mixed effects when it is violated. Moreover, flexible specification for random effects such as multimodal or skewness might provide insight into heterogeneity and even unveil failure to include important covariates in the model. Such concern has motivated many nonparametric approaches for the random effects. Zhang and Davidian (2001) approximated the random effects by the seminonparametric approach of Gallant and Tauchen (1987). Further, Chen et al. (2002) extended it to the generalized linear mixed models (GLMMs). There are also some other frequentist approaches proposed such as Lai and Shih (2003), and Ghidey et al. (2004). Alternatively, many Bayesian nonparametric approaches using Dirichlet process (DP) (Ferguson, 1973) and DP mixtures (DPM) are also proposed. Readers can refer to Bush and MacEachern (1996), Kleinman and Ibrahim (1998), Ishwaran and Takahara (2002), among many others. However, these methods do not address the uncertainty of predictors to be included in the mixed effects of the model.

Typically, a random variable is included when it is expected to vary among subjects. However, a practical problem is how to decide which predictors have coefficients varying among subjects. Standard approaches such as Akaike information criterion (AIC), Bayesian information criterion (BIC), generalized information criterion (GIC) and Bayes factor (BF) generally compare a few models in enumeration. However, such methods do not work well when the number of potential predictors is large. Especially, the number of possible models increases exponentially with the number of predictors. For example, with l_1

* Tel.: +1 919 541 4056; fax: +1 919 541 4311.
E-mail address: mingany@yahoo.com.

0167-9473/$ – see front matter © 2012 Published by Elsevier B.V.
doi:10.1016/j.csda.2011.12.014

fixed effects and l_2 random effects, the total number of possible models is $2^{l_1+l_2}$. When $l_1 = l_2 = 10$, the total number of model is well above one million.

Unlike the linear mixed effects (LME) model (Laird and Ware, 1982), the random effects have a rather complicated maximum likelihood form in logistic mixed models. Inference based on likelihood requires integration over the dimensions of the random effects, which is often intractable even with simple normal distribution. With this, researchers proposed Laplace and other approximation approaches, for example, Schall (1991), Breslow and Clayton (1993) etc. However, such approaches may result in biased estimates for the fixed effects (Breslow and Lin, 1995; Lin and Breslow, 1996). To resolve the difficulty, some Bayesian methods have been developed to circumvent the intense integration. Zeger and Karim (1991) used Gibbs sampling for the random effects. McCulloch (1997) and Booth and Hobert (1999) used Monte Carlo EM algorithm for posterior inference.

For mixed effects models, it is desirable to accommodate uncertainty of predictors to be included in the model for enhanced flexibility. Bayesian methods can accommodate such flexibility and avoid cumbersome integration with MCMC algorithms. In addition, one can easily infer from the variable selections results, for example, posterior probabilities of the mixed effects inclusion and models of the Bayesian approaches. Kuo and Mallick (1998) and George and McCulloch (1993, 1997) used the approach of Bayesian variable selection for the general linear model. Chen and Dunson (2003) used the Cholesky decomposition for the random effects. Kinney and Dunson (2007) extended the approach to logistic mixed model. Bondell et al. (2010) proposed a penalized joint likelihood with an adaptive penalty in joint selection of both fixed and random effects. Ibrahim et al. (2010) used maximum penalized likelihood estimation for fixed and random effects selection. However, all these approaches do not have flexible specification for the random effects. For nonparametric specification of the uncentered random effects, the expected mean generally is not zero and thus causes identifiability with the fixed effects. Ultimately, bias is incurred. Cai and Dunson (2010) proposed a nonparametric random effect model without addressing the potential bias. Though, they might take the approach by Yang and Dunson (2010), Yang et al. (2010) and Li et al. (2011) to reduce bias. However, it is difficult for interpretation with variable selection, in particular, when the fixed effect is selected but the corresponding random effect is not. With this, Yang (2010) used the centered Dirichlet process mixture models for the random effects. To the author's best knowledge, there is no method proposed for GLMM which addresses joint selection of mixed effects, flexible prior specification and bias control simultaneously.

In this article, we address variable selection for logistic mixed model with nonparametric random effects. The article is organized as follows: Section 2 describes the logistic mixed models. Section 3 describes the approach of joint selection of fixed and random effects and the posterior inference. Sections 4 and 5 presents simulation and real data example respectively. A final discussion is provided to conclude the article.

2. Methodology

2.1. General description

Suppose there are n subjects in a study and each subject has n_i repeated observations for $i = 1, \ldots, n$. Let $\boldsymbol{X_{ij}}$ denote the predictor for subject i at observation j, a vector of dimension $l \times 1$, let y_{ij} be the corresponding binary response variable, $\boldsymbol{Z_{ij}}$ is a predictor vector of dimension $q \times 1$. Then the logistic mixed model is denoted as:

$$y_{ij} \sim \text{Bernoulli}(\wp^{-1}(\chi_{ij})), \qquad \chi_{ij} = \boldsymbol{X_{ij}}'\boldsymbol{\beta} + \boldsymbol{Z_{ij}}'\boldsymbol{\zeta_i} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_l)$ is the fixed effect coefficient vector, $\boldsymbol{\zeta_i} \sim N(0, \boldsymbol{\Omega})$ is the ith random effect, $\wp(\cdot)$ is the logistic link function with $\wp(\kappa) = \log(\kappa/(1-\kappa))$. Generally, $\boldsymbol{Z_{ij}}$ is set as a subset of $\boldsymbol{X_{ij}}$. Assume $\boldsymbol{X_{ij}}$ and $\boldsymbol{Z_{ij}}$ include all the candidate predictors, we are interested in searching for a subset of important predictors to be included in the model.

Obviously, the logistic model is nonlinear and thus we cannot get conditional conjugacy even with simple normal priors, which ultimately causes inefficiency in computation. To overcome the cumbersome nonlinear issue, we take several approaches of approximation to convert the nonlinear model to the standard linear models. First, we take the approach by Albert and Chib (1997) that the logistic distribution can be closely approximated by the t distribution. With auxiliary variables, Model (1) is equivalent to the following representation:

$$y_{ij} = 1 : y^*_{ij} > 0$$
$$y_{ij} = 0 : y^*_{ij} \le 0$$

where y^*_{ij} is an underlying value with the logistic distribution with location parameter $\boldsymbol{X_{ij}}'\boldsymbol{\beta} + \boldsymbol{Z_{ij}}'\boldsymbol{\zeta_i}$ and density function as follows:

$$f(y^*_{ij}|\boldsymbol{X_{ij}}, \boldsymbol{Z_{ij}}, \boldsymbol{\beta}_i, \boldsymbol{\zeta_i}) = \frac{\exp\{-(y^*_{ij} - \boldsymbol{X_{ij}}'\boldsymbol{\beta} - \boldsymbol{Z_{ij}}\boldsymbol{\zeta_i})\}}{\{1 + \exp[-(y^*_{ij} - \boldsymbol{X_{ij}}'\boldsymbol{\beta} - \boldsymbol{Z_{ij}}\boldsymbol{\zeta_i})]\}^2}. \tag{2}$$

Second, as noted by West (1987), the t distribution can be expressed as a scale mixture of normals. Thus, y^*_{ij} is approximated as a non-central t distribution with location parameter $\boldsymbol{X_{ij}}'\boldsymbol{\beta} + \boldsymbol{Z_{ij}}'\boldsymbol{\zeta_i}$, degree of freedom v and scale parameter σ^2. Then we can get the following model:

$$y^*_{ij} = \boldsymbol{X_{ij}}'\boldsymbol{\beta} + \boldsymbol{Z_{ij}}'\boldsymbol{\zeta_i} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2/\phi_{ij}), \phi_{ij} \sim G(v/2, v/2) \tag{3}$$