Journal of Statistical Planning and Inference 141 (2011) 3391–3399

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Posterior consistency of random effects models for binary data

Yongdai Kim [a,*], Dohyun Kim [b]

[a] Department of Statistics, Seoul National University, Sillimdong, Kwanakgu, Seoul 151-878, Republic of Korea
[b] Department of Information Systems, Business Statistics and Operations Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR

ARTICLE INFO

Article history:
Received 9 May 2009
Received in revised form
10 August 2010
Accepted 29 December 2010
Available online 5 January 2011

Keywords:
Nonparametric Bayesian
Posterior consistency
Random effect model

ABSTRACT

In longitudinal studies or clustered designs, observations for each subject or cluster are dependent and exhibit intra-correlation. To account for this dependency, we consider Bayesian analysis for conditionally specified models, so-called generalized linear mixed model. In nonlinear mixed models, the maximum likelihood estimator of the regression coefficients is typically a function of the distribution of random effects, and so the misspecified choice of the distribution of random effects can cause bias in the estimator. To avoid the problem of the misspecification of the distribution of random effects, one can resort in nonparametric approaches. We give sufficient conditions for posterior consistency of the distribution of random effects as well as regression coefficients.

1. Introduction

In longitudinal studies or clustered designs, observations for each subject or cluster are dependent and exhibit intra-correlation. To account for such intra-correlation in regression problems, generalized linear mixed models (Breslow and Clayton, 1993) have been popularly used. In this paper, we consider Bayesian analysis of generalized linear mixed models, where the distribution of random effects is fully unspecified. We give sufficient conditions for posterior consistency of the distribution of random effects as well as regression coefficients.

In linear mixed models, the misspecified choice of the random effect distribution does not result in biased estimation of the regression coefficients. In nonlinear mixed models, however, the maximum likelihood estimator of the regression coefficients is typically a function of the distribution of random effects, and so the misspecified choice of the distribution of random effects can cause bias in the estimator. See, for example, Neuhaus et al. (1992), Heagerty (1999), Heagerty and Zerger (2000) and Heagerty and Kurland (2001).

To avoid the problem of the misspecification of the distribution of random effects, one can resort in nonparametric approaches. Nonparametric maximum likelihood estimation has been considered by Feinberg et al. (1985), Follman and Lambert (1989), Lindsay (1983) and Butler and Louis (1997), and Bayesian nonparametric approaches have been developed by Mukhopadhyay and Gelfand (1997), Kleinman and Ibrahim (1998), Walker and Mallick (1997) and Dunson (2005).

Posterior consistency has been one of the most important issues in the Bayesian society since the seminar paper of Doob (1949) and Diaconis and Freedman (1986a,b). Some related literature is Schwartz (1965), Le Cam (1973), Barron (1986), Barron et al. (1996), Ghosal et al. (1999), Walker (2004), Ameowou-Atisso et al. (2003), Ghosal and Roy (2006),

* Corresponding author.
E-mail address: ydkim0903@gmail.com (Y. Kim).

0378-3758/$ - see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.jspi.2010.12.021

Ge and Jiang (2006), Ghosal and van der Vaart (2007), Choi and Schervish (2007), Choi (2008) and Wu and Ghosal (2008). Even though the model considered in this paper can be considered as a special problem of the models in the aforementioned papers, however, the previous results are not directly applicable to our problem. The key difficulty in our problem is the model identifiability, which means that the regression coefficients and the distribution of random effects are identifiable. In particular, when the distribution of random effects is fully unspecified, the identifiability of the model is by no means obvious. For linear mixed models, Teicher (1961) proved the identifiability and Butler and Louis (1997) gave sufficient conditions for the identifiability of generalized linear mixed models. In this paper, we prove the identifiability by showing the existence of exponentially consistent sequence of test as is done by Ameowou-Atisso et al. (2003) for linear regression models.

It is interesting to compare our sufficient conditions with those given by Butler and Louis (1997). An important advantage of our sufficient conditions is that the minimum number of observations in each subject or cluster is equal to the dimension of random effects while it is larger than the dimension of random effects in Butler and Louis (1997). For example, in the random intercept model (i.e. the dimension of random effects is 1), the model is identifiable when each subject or cluster has only one observation based on our results. Of course, we assume that the support of random effects is compact, which is not required in Butler and Louis (1997). In this sense, the results in this paper can be considered as a useful alternative to the results of Butler and Louis (1997).

This paper is organized as follows. We describe the model in Section 2. In Section 3, we give sufficient conditions for posterior consistency and proofs are given in Section 4. To illustrate results, we performed small simulation, whose results are presented in Section. 5.

2. Model

Consider the following latent linear mixed model:

$$T_{ij} = X_{ij}'\beta + Z_{ij}'b_i + \varepsilon_{ij}, \quad i = 1,\ldots,n,\ j = 1,\ldots,n_i.$$

Here $X_{ij} = (X_{ij1},\ldots,X_{ijp})'$, $\beta = (\beta_1,\ldots,\beta_p)'$ are p-dimensional vector of fixed effects' covariate and corresponding parameters, and $Z_{ij} = (Z_{ij1},\ldots,Z_{ijq})'$, $b_i = (b_{i1},\ldots,b_{iq})'$ are q-dimensional vector of random effect' covariate and corresponding parameters with the distribution F. An error term, ε_{ij}, is having the known distribution G and independent with b_{ij}. Instead of observing T_{ij} directly, we observe $Y_{ij} = I(T_{ij} \geq 0)$. Let $Y_i = (Y_1,\ldots,Y_{n_i})$, $X_i = (X_{i1},\ldots,X_{in_i})$ and $Z_i = (Z_{i1},\ldots,Z_{in_i})$. Let $\Omega_i = \{0,1\}^{n_i} \times \mathbb{R}^{p\times n_i} \times \mathbb{R}^{q\times n_i}$, $i = 1,2\ldots,n$ be the sample space for D_i=(Y_i, X_i, Z_i) and $\Omega^n = \prod_{i=1}^n \Omega_i$ be the sample space for $D^n = (D_1,\ldots,D_n)$. The objective is to make inference on $\theta = (\beta,F)$ based on the observations D^n.

Let P_θ^∞ be a probability measure on $(\Omega^\infty,\mathcal{B}(\Omega^\infty))$ such that

$$P_\theta^\infty\left(D^n \in \prod_{i=1}^n C_i\right) = \prod_{i=1}^n P_\theta^{n_i}(D_i \in C_i)$$

for all n and $C_i \in \mathcal{B}(\Omega_i)$, $i = 1,\ldots,n$. Note that if $C_i = \{y_i\} \times A_i \times B_i$ where $y_i \in \{0,1\}^{n_i}$,$A_i \in \mathcal{B}(\mathbb{R}^{p\times n_i})$ and $B_i \in \mathcal{B}(\mathbb{R}^{q\times n_i})$, then

$$P_\theta^{n_i}(D_i \in C_i) = \int_{A_i\times B_i} p_{i,\theta}(y_i|x_i,z_i)p^{n_i}(x_i,z_i)\nu^{n_i}(dx_i,dz_i),$$

where $p^{ni}(x_i,z_i)$ is the joint density of (X_i,Z_i) with respect to a $\sigma-$finite measure ν^{n_i} and

$$p_{i,\theta}(y_i|x_i,z_i) = \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} G(x_{ij}'\beta + z'_{ij}b_i)^{y_{ij}}(1-G(x_{ij}'\beta + z_{ij}'b_i))^{1-y_{ij}}\, dF(b_i).$$

Note that $p^{ni}(x_i,z_i)$ does not depend on θ. Also, we assume that there exists a $\sigma-$finite measure ν on $\mathbb{R}\times\mathbb{R}$ such that $\nu^{n_i}(dx_i,dz_i) = \prod_{j=1}^{n_i} \nu(dx_{ij},dz_{ij})$.

Let $\Theta = \mathcal{G}\times\mathcal{F}(\subset \mathbb{R}^p \times \mathcal{M}(\mathbb{R}^q))$ be the parameter space, where $\mathcal{M}(\mathbb{R}^q)$ is the space of all probability measures on $\mathbb{R}^q$ equipped with the weak topology. Let $\Pi = \mu\times\Pi_1$ is a prior distribution for $\theta = (\beta,F)$ where μ is for β and Π_1 is for F. For μ, we consider a standard parametric prior such as a multivariate normal distribution and a nonparametric prior for F such as a Dirichlet process.

Finally, let $\Pi_n(\cdot|\cdot) : \mathcal{B}(\Theta)\times\Omega^n \to [0,1]$ is a posterior distribution of θ given $D^n = (D_1,\ldots,D_n)$, where $\mathcal{B}(\Theta)$ is the Borel $\sigma-$field for Θ. Note that

$$\Pi_n(d\theta|D^n) \propto \prod_{i=1}^n \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} G(X_{ij}'\beta + Z_{ij}'b_i)^{Y_{ij}}(1-G(X_{ij}'\beta + Z_{ij}'b_i))^{1-Y_{ij}}\, dF(b_i)\Pi(d\theta).$$

Remark 1. In this paper, we only consider the case of random covariates. However, all of the results proved in this paper can be modified for nonrandom covariates without much difficulty.

Remark 2. We assume that G is completely known, which may be a limitation. We may introduce unknown parameters in G (e.g. scale parameters). However, we should be careful since additional unknown parameters in G would make the model unidentifiable. For example, let $\sigma > 0$ be an unknown scale parameter such that $G(t) = G'(t/\sigma)$ for some given distribution