Journal of Multivariate Analysis 126 (2014) 86–99

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Bayesian influence analysis of generalized partial linear mixed models for longitudinal data

Nian-Sheng Tang [a,*], Xing-De Duan [a,b]

[a] *Department of Statistics, Yunnan University, Kunming 650091, PR China*

[b] *Department of Mathematics, Chuxiong Normal University, Chuxiong 675000, PR China*

ARTICLE INFO

Article history:
Received 9 April 2013
Available online 21 January 2014

AMS subject classifications:
62F15
62J20

Keywords:
Bayesian case influence
Bayesian local influence
Bayesian perturbation manifold
Dirichlet process prior
Generalized partial linear mixed models

ABSTRACT

This paper develops a Bayesian local influence approach to assess the effects of minor perturbations to the prior, sampling distribution and individual observations on the statistical inference in generalized partial linear mixed models (GPLMMs) with the distribution of random effects specified by a truncated and centered Dirichlet process (TCDP) prior. A perturbation manifold is defined. The metric tensor is employed to select an appropriate perturbation vector. Several Bayesian local influence measures are proposed to quantify the degree of various perturbations to statistical models based on the first and second-order approximations to the objective functions including the ϕ-divergence, the posterior mean distance and Bayes factor. We develop two Bayesian case influence measures to detect the influential observations in GPLMMs based on the ϕ-divergence and Cook's posterior mean distance. The computationally feasible formulae for Bayesian influence analysis are given. Several simulation studies and a real example are presented to illustrate the proposed methodologies.

1. Introduction

Generalized partial linear mixed models (GPLMMs) are often used to fit longitudinal and repeated measurement or clustered data over time by considering the between-subject and within-subject variations in various settings including education, psychology, medicine, public health, epidemiology and econometrics. Various methods have been proposed to make statistical inference on GPLMMs under the assumption that the random effects follows a fully parametric distribution such as normal distribution in recent years. For example, see He, Fung and Zhu [9], Zhou, Zhu and Fung [35], Qin and Zhu [27], Liang [18] and Qin, Bai and Zhu [26]. But, the parametric assumption may be unreasonable in some practical problems (Kleinman and Ibrahim [13,14]; Lee, Lu and Song [16]; Li, Müller and Lin [17]; Yang, Dunson and Baird [34]). And violation of the parametric assumption on the random effects may lead to unreasonable or incorrect conclusions. Hence, this paper extends GPLMMs by allowing the random effects to have a nonparametric prior distribution. Particularly, this paper uses a truncated and centered Dirichlet process (TCDP) prior to specify the distribution of the random effects in GPLMMs. But, so far, we have not seen any work in the literature on Bayesian influence analysis for data from GPLMMs.

The widely used Bayesian influence analysis methods mainly include Bayesian case influence analysis and Bayesian local influence analysis. Bayesian case influence diagnostics have been developed under various statistical models based on the Kullback–Leibler (K–L) divergence and the Conditional Predictive Ordinate (CPO). For example, Pettit [22] used the K–L divergence to identify influential observations in his review of Bayesian diagnostics; Weiss and Cook [33] presented a Bayesian case influence diagnostic to detect influential observations via the K–L divergence in generalized linear

* Corresponding author.
E-mail address: nstang@ynu.edu.cn (N.-S. Tang).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.12.005

models; Peng and Dey [20] suggested a Bayesian diagnostic via some general divergence measures including the K–L divergence on the posterior distribution and examined the application of this Bayesian diagnostic to several regression models such as a nonlinear model; Weiss [32] developed a Bayesian case deletion diagnostic via model perturbations and studied its relationship to the K–L divergence and CPO; Bradlow and Zaslavsky [2] investigated application of case influence analysis to Bayesian hierarchical models; Cho, Ibrahim, Sinha and Zhu [4] proposed case influence diagnostics via the K–L divergence for complex survival models and derived a simplified formula to compute the K–L divergence between the posterior densities for the full data and the deleted data; Jackson, White and Carpenter [12] proposed two efficient ways to approximate Bayesian case influence diagnostic by using outputs from MCMC algorithm; Zhu, Ibrahim, Cho and Tang [36] investigated the statistical properties of three Bayesian case influence measures including the ϕ-divergence, Cook's posterior mode distance and Cook's posterior mean distance for identifying a set of influential observations for a variety of statistical models with missing data including models for longitudinal data and latent variable models in the absence/presence of missing data. However, very little work has been done on systematically studying Bayesian case influence measures including the ϕ-divergence and Cook's posterior mean distance in a GPLMM for longitudinal data. Hence, one of the main purposes of this paper is to develop the above mentioned two Bayesian case influence measures to identify influential observations (or sets of influential observations) for our considered GPLMMs.

On the other hand, Bayesian local influence analysis has been widely investigated under various statistical models on the basis of different objective functions. For example, Gustafson [8] used the norm of the Fréchet derivatives with respect to the prior distribution to quantify the effect of the prior perturbation; McCulloch [19] generalized Cook's [6] local influence approach to assess the effect of perturbing the prior or likelihood based on the curvature of the K–L divergence (between perturbed and unperturbed posteriors) in a Bayesian analysis; Perez, Martin and Rufo [21] used the norm of the Gateaux derivative to measure the effect of the prior perturbation; van der Linde [31] developed a local influence approach to assess the effect of multiplicative modes of perturbation on posterior distributions. Recently, Ibrahim, Zhu and Tang [10] developed a Bayesian local influence method to assess perturbations to the prior, the sampling distribution and individual observations in a survival model; Zhu, Ibrahim and Tang [37] developed a general framework of Bayesian influence analysis for assessing various perturbations to the data, the prior and the sampling distribution for a class of parametric models from a viewpoint of differential geometry. However, to the best of our knowledge, there is very little work done on proposing a general Bayesian local influence approach to simultaneously perturb the three components of a Bayesian model, assessing their effects and examining their applications in Bayesian GPLMMs analysis with random effects specified by a TCDP prior. Hence, another main purpose of this paper is to develop a Bayesian local influence approach to assess the effect of minor perturbations to the individual observations, the prior and the sampling distribution of a Bayesian model in Bayesian GPLMMs analysis, and propose a novel computationally low-cost approach to estimate the basic components in Bayesian local influence measure. Following Zhu, Ibrahim and Tang [37], we construct the Bayesian perturbation manifold for the perturbation model and the geometrical quantities for checking appropriate selection of a perturbation vector, and develop Bayesian local influence measures for identifying the most perturbations based on the objective functions such as the ϕ-divergence, the posterior mean distance and Bayes factor. Some computationally feasible formulae for Bayesian influence analysis are given by using outputs from the MCMC algorithm.

The rest of this paper is organized as follows. Section 2 describes a GPLMM and presents a TCDP with a stick-breaking prior to approximate the distribution of random effects and a P-spline approach to model nonparametric part and log of variance of spline coefficients. Also, a MCMC algorithm for simultaneously estimating the unknown parameters and random effects in GPLMMs is given in Section 2. Several Bayesian local influence measures are developed for assessing the effects of minor perturbation to the data, the priors and the sampling distribution in Section 3. Two Bayesian case influence measures including the ϕ-divergence and Cook's posterior mean distance are developed to identify the most influential observations in Section 4. Simulation studies and a real example are used to illustrate the proposed methodologies in Section 5. Some concluding remarks are given in Section 6.

2. Model and MCMC algorithm

Let $\boldsymbol{u}_i$ be a $r \times 1$ vector of random effects corresponding to the ith individual, and y_{ij} be an observation of the ith individual measured at time t_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Suppose that $y_{i1}, \ldots, y_{in_i}$ given $\boldsymbol{u}_i$ are conditionally independent and each $y_{ij}|\boldsymbol{u}_i$ is distributed as an exponential family distribution whose probability density function is given by

$$p(y_{ij}|\boldsymbol{u}_i, \phi) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \tag{1}$$

where ϕ is a scale parameter, $c(y, \phi)$ is a function only depending on y and ϕ, θ_{ij} is the (scalar) canonical parameter. In this paper, we assume that the conditional mean μ_{ij} satisfies

$$\eta_{ij} \stackrel{\Delta}{=} f(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{w}_{ij}^T\boldsymbol{u}_i + g(t_{ij}), \quad i = 1, \ldots, n,\ j = 1, \ldots, n_i, \tag{2}$$

where $f(\cdot)$ is a known monotonic "link" function, $\boldsymbol{x}_{ij}$ is a $p \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters relating to the fixed effects, $\boldsymbol{w}_{ij}$ is a $r \times 1$ vector of explanatory variables relating to the random effects, $g(\cdot)$ is