Computational Statistics and Data Analysis 56 (2012) 4259–4265

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Likelihood inference in generalized linear mixed models with two components of dispersion using data cloning

Mahmoud Torabi *

Department of Community Health Sciences, University of Manitoba, MB, R3E 0W3, Canada

ARTICLE INFO

Article history:
Received 2 June 2011
Received in revised form 29 December 2011
Accepted 14 April 2012
Available online 25 April 2012

Keywords:
Bayesian computation
Efficiency
Hierarchical models
Random effects
Variance components

ABSTRACT

This paper studies generalized linear mixed models (GLMMs) with two components of dispersion. The frequentist analysis of linear mixed model (LMM), and particularly of GLMM, is computationally difficult. On the other hand, the advent of the Markov chain Monte Carlo algorithm has made the Bayesian analysis of LMM and GLMM computationally convenient. The recent introduction of the method of data cloning has made frequentist analysis of mixed models also equally computationally convenient. We use data cloning to conduct frequentist analysis of GLMMs with two components of dispersion based on maximum likelihood estimation (MLE). The resultant estimators of the model parameters are efficient. We discuss the performance of the MLE using the well known salamander mating data, and also through simulation studies.

1. Introduction

Generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993) are obtained from generalized linear models (GLMs) (McCullagh and Nelder, 1989) by incorporating random effects into the linear predictors, and include the well-known linear mixed models (LMMs) for normal responses (Laird and Ware, 1982) as a special case. These models are useful for modeling the dependence among response variables inherent in longitudinal or repeated measures studies, for accommodating overdispersion among binomial or Poisson responses, and for producing shrinkage estimators in multiparameter problems, such as the construction of maps of small area disease rates (Breslow and Clayton, 1993; Rao, 2003).

One of the early applications of GLMMs was made by McCullagh and Nelder (1989, Section 14.5) to a salamander mating dataset with two components of dispersion, and for other applications of GLMMs, see, for example, Breslow and Clayton (1993), Lee and Nelder (1996), McCulloch (1997) and Malec and Sedransk (1997). A major difficulty in making inferences about GLMMs has been computational. In particular, obtaining consistent and efficient estimators for the regression and the variance components in GLMMs has been proven to be difficult. To overcome numerical difficulties, many authors have approximated the GLMMs to make inference. For instance, Breslow and Clayton (1993) proposed an approximation approach called penalized quasi likelihood (PQL) which may or may not yield a consistent estimator for the variance components, depending on the cluster size and the associated design matrix. Kuk (1995), Breslow and Lin (1995) and Lin and Breslow (1996), among others, provided certain asymptotic bias corrections both for the regression and the variance component estimates. Jiang (1998) proposed a simulated-moment approach that always yields consistent estimators for the parameters of the mixed model, however, the moment (MOM) estimators may be inefficient.

In many practical applications such as biological and/or biomedical studies, one may encounter discrete or continuous data with two-way correlations caused by two sources of random variation. For instance, McCullagh and Nelder (1989)

* Tel.: +1 204 272 3136; fax: +1 204 789 3905.
E-mail address: torabi@cc.umanitoba.ca.

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.04.008

reported an interesting and particularly challenging dataset on salamander mating. The dataset is challenging because the response variable for this experiment is binary and the study design is crossed rather than nested. This makes the marginal likelihood particularly difficult to evaluate. This type of correlated data has been analyzed by many authors such as Breslow and Clayton (1993), Jiang (1998), Booth and Hobert (1999), Jiang and Zhang (2001) and Sutradhar and Rao (2003) in the frequentist set up. Breslow and Clayton (1993) used a PQL approach, Jiang (1998) used a simulated-moment approach, Booth and Hobert (1999) used a Monte Carlo EM algorithm (MCEM), and Jiang and Zhang (2001) used two-step estimating equations approach. Sutradhar and Rao (2003) used a quasi-likelihood (QL) approach analogous to two-way ANOVA to analyze such two-way correlated cluster data. It is shown that the QL approach yields consistent estimators for the parameters of the GLMMs with two components of dispersion, and also efficient estimators for the parameters of the GLMMs with a single component of dispersion. In this paper, we introduce a way to compute maximum likelihood estimation (MLE) to analyze such two-way correlated cluster data. It is shown that the resulting estimators are efficient.

The proposed approach to compute MLE is based on a recently introduced method called data cloning (DC) (Lele et al., 2007, 2010) for general hierarchical models. The DC is based on Bayesian ideas and uses Markov chain Monte Carlo (MCMC) methodology. Similar to the Bayesian approach, DC avoids high dimensional numerical integration and requires neither maximization nor differentiation of a function. Because these estimators are maximum likelihood (ML) estimators, unlike the Bayesian estimators, they are independent of the choice of priors, non-estimable parameters are flagged automatically and the possibility of improper posterior distribution is completely avoided.

The paper is organized as follows. In Section 2, the GLMM with two components of dispersion is described. Data cloning is then introduced to estimate the model parameters such as the regression effects and two variance components (Section 3). In Section 4, we apply the MLE approach (via DC) to reanalyze the salamander mating data. We then conduct simulation studies to examine the performance of the MLE approach in estimating regression parameters and variance components of the mixed models (Section 5). Concluding remarks are provided in Section 6.

2. GLMM with two components of dispersion

Let y_{ij} be the variable of interest for the ith level of a factor A and the jth level of a factor B ($i = 1, \ldots, m; j = 1, \ldots, n$). The y_{ij} are assumed to be conditionally independent with exponential family p.d.f.

$$f(y_{ij}|\theta_{ij}, \phi_{ij}) = \exp\{(y_{ij}\theta_{ij} - a(\theta_{ij}))/\phi_{ij} + c(y_{ij}, \phi_{ij})\}, \tag{1}$$

($i = 1, \ldots, m; j = 1, \ldots, n$). The density (1) is parameterized with respect to the canonical parameters θ_{ij}, known scale parameters ϕ_{ij} and functions $a(\cdot)$ and $c(\cdot)$. The exponential family (1) covers well-known distributions including normal, binomial and Poisson distributions. The natural parameters θ_{ij} are then modeled as

$$\theta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_i + v_j \quad (i = 1, \ldots, m; j = 1, \ldots, n), \tag{2}$$

where $\boldsymbol{x}_{ij}(p \times 1)$ are known design vectors, $\boldsymbol{\beta}(p \times 1)$ is an unknown vector regression coefficient, and u_i and v_j are the random effects. It is assumed that $u_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_u^2)$ and $v_j \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$. The model (1)–(2) which involves the regression coefficients β and two variance components σ_u^2 and σ_v^2 is refereed to as a GLMM with two variance components. It is of interest to obtain efficient estimates of the model parameters.

Note that the observations in model (1)–(2) are correlated in two ways. More precisely, at the ith level of factor A, y_{ij} and y_{ik} are independent conditional on u_i, while they are unconditionally correlated. Similarly, at the jth level of factor B, y_{ij} and y_{rj} are independent conditional on v_j, while they are unconditionally correlated.

3. Inference using data cloning

Let $\boldsymbol{y} = (y_{11}, \ldots, y_{1n}, \ldots, y_{m1}, \ldots, y_{mn})'$ be the observed data vector and, conditionally on the random effects, $\boldsymbol{b} = (u_1, \ldots, u_m, v_1, \ldots, v_n)'$, assume that the elements of $\boldsymbol{y}$ are independent and drawn from a distribution in exponential family with parameters $\boldsymbol{\alpha}_1$. It is also assumed that distribution for $\boldsymbol{b}$ depends on parameters $\boldsymbol{\alpha}_2$. The goal of the analysis is to estimate the model parameters $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)'$ and predict the random effects $\boldsymbol{b}$.

To illustrate the DC approach, we start with the standard Bayesian approach to inference for hierarchical models. Denote $L(\boldsymbol{\alpha}; \boldsymbol{y})$ as the likelihood of $\boldsymbol{\alpha}$ given data $\boldsymbol{y}$ and $\pi(\boldsymbol{\alpha})$ as prior distribution on the parameter space. The posterior distribution $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ is given by

$$\pi(\boldsymbol{\alpha}|\boldsymbol{y}) = \frac{L(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha})}{C(\boldsymbol{y})}, \tag{3}$$

where $C(\boldsymbol{y}) = \int L(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha})d\boldsymbol{\alpha}$ is the normalizing constant. There are computational tools MCMC algorithms, that facilitate generation of random variates from the posterior distribution $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ without computing the integrals in the numerator or the denominator of (3) (Gilks et al., 1996; Spiegelhalter et al., 2004).