Computational Statistics and Data Analysis 89 (2015) 158–171

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Likelihood inference for small area estimation using data cloning

Mahmoud Torabi [a,*], Subhash R. Lele [b], Narasimha G.N. Prasad [b]

[a] Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, R3E 0W3, Canada

[b] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, T6G 2G1, Canada

ARTICLE INFO

Article history:
Received 19 December 2012
Received in revised form 10 March 2015
Accepted 20 March 2015
Available online 1 April 2015

Keywords:
Bayesian computation
Hierarchical model
Prediction interval and exponential family
Random effect

ABSTRACT

Policy decisions regarding allocation of resources to subgroups in a population, called small areas, are based on reliable predictors of their underlying parameters. However, in sample surveys, the information to estimate reliable predictors is often insufficient at the level of the small areas. Hence, parameters of the subgroups are often predicted based on the coarser scale data. In view of this, there is a growing demand for reliable small area predictors by borrowing information from other areas. These models are commonly based on either linear mixed models (LMMs) or generalized linear mixed models (GLMMs). The frequentist analysis of LMM, a special case 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. Recently developed data cloning method provides a frequentist approach to complex mixed models which is also computationally convenient. Data cloning which yields to maximum likelihood estimation is used to conduct frequentist analysis of small area estimation for Normal and non-Normal responses. It is shown that for the Normal and non-Normal responses, data cloning leads to predictions and prediction intervals of small area parameters that have reasonably good coverage.

1. Introduction

Sample surveys are conducted with the purpose of providing reliable predictors for the finite population parameters such as totals or means. Methods used in deriving such predictors (direct survey predictors) are based on total sample size. However in the past few decades, there has been increasing demand to use sample survey data to get predictions for sub-populations, such as counties or gender–age groups. Such sub-populations for which reliable predictions are needed are called small areas. The traditional area-specific direct predictors tend to have inadequate precision because of the small samples sizes corresponding to each small area. In health services research, policy decisions about implementing programs or projects in these small areas are made using predictions of underlying parameters. Hence, survey researchers are developing methods to provide more reliable predictions for small areas. To this end, model based estimators (Rao, 2003; Jiang and Lahiri, 2006; Jiang, 2010) have been proposed to borrow strength from stated areas by introducing random effects. Depending on the nature of the data, either the linear mixed model (LMM) (Searle et al., 1992) or the generalized linear mixed model (GLMM) (McCulloch and Searle, 2001) is most often used for small area estimation (Fay and Herriot,

* Corresponding author. Fax: +1 204 789 3905.
E-mail address: torabi@cc.umanitoba.ca (M. Torabi).

http://dx.doi.org/10.1016/j.csda.2015.03.013
0167-9473/

1979; Battese et al., 1988; Kass and Steffey, 1989; MacGibbon and Tomberlin, 1989; Prasad and Rao, 1990; Malec et al., 1997; Ghosh et al., 1998; Singh et al., 1998; Datta et al., 2000; Ghosh et al., 2009). Among other approaches, parameters of the LMM can be estimated using either the maximum likelihood (ML) or restricted ML (REML). It is straightforward to predict the small area mean under the LMM by using the best linear unbiased predictor (BLUP), however obtaining its prediction error and associated prediction interval are difficult. Both overall parameters estimation and prediction of small area parameters under the GLMM are computationally difficult under the frequentist approach. The Bayesian approach has become quite popular because of its computational convenience and the ability to provide not just the point predictors but also the associated prediction intervals. However, the Bayesian approach to prediction of small area parameters crucially depends on the specification of the prior. Non-informative or vague priors are often used to possibly get more information from the data. However, lack of unique definition of non-informative prior leads to many different suggestions for such priors. It is well known that the choice of prior affects the predicted values. Hence, implementation of the Bayesian approach requires substantial care. For example, use of an inappropriate prior distribution can lead to improper posterior distribution making the inferential statements somewhat questionable (e.g., Natarajan and McCulloch, 1995 and Hobert and Casella, 1996).

Recently, Lele et al. (2007) introduced an alternative approach, called data cloning, to compute the ML estimates and their standard errors for general hierarchical models. Similar to the Bayesian approach, data cloning avoids high dimensional numerical integration and requires neither maximization nor differentiation of a function. Because these estimators are ML estimators, unlike the Bayesian estimators, they are independent of the choice of the priors. By applying the data cloning approach, non-estimable parameters are also flagged automatically and possibility of improper posterior distribution is completely avoided. Extending this work to the GLMM situation, Lele et al. (2010) described an approach to compute prediction and prediction intervals of the random effects. Thus, the data cloning approach is well suited to address the issues in small area estimation using the frequentist paradigm.

In this paper, we use data cloning in the context of small area estimation. In the next section, we describe the small area estimation problem in general and describe how data cloning can be used to obtain prediction and prediction intervals of small area parameters. In Section 3, we use three real datasets to evaluate the performance of data cloning under cross-sectional (normal and binomial mixed models) as well as cross-sectional and time-series (normal mixed model). The data cloning is also evaluated through simulation studies (Section 4). Concluding remarks are given in Section 5.

2. Small area estimation using data cloning

The basic model in small area estimation can be described as follows. Let y_{ij} be the variable of interest for the jth unit within the ith area ($j = 1, \ldots, n_i$; $i = 1, \ldots, m$). The y_{ij} are assumed to be conditionally independent with exponential family p.d.f.

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

The density (1) is parameterized with respect to the canonical parameters θ_{ij}, known scale parameters ϕ_{ij} and functions $a(\cdot)$ and $b(\cdot)$. The natural parameters θ_{ij} are then modeled as

$$h(\theta_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i + v_{ij},$$

where h is a strictly increasing function, $\boldsymbol{x}_{ij}(p \times 1)$ are known design vectors, $\boldsymbol{\beta}(p \times 1)$ is a vector of unknown regression coefficients, and u_i and v_{ij} are random effects with $u_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_u^2)$ and $v_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$. We also assume that the population model holds for the sample. The objective in small area estimation is to make inferences on the small area parameters θ_{ij} or its variant. We now explain how data cloning can be used in the context of small area estimation.

2.1. Data cloning: a brief description

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)'$ be the observed data vector and assume that the elements of $\boldsymbol{y}$ are conditionally independent of the random effects $\boldsymbol{u} = (u_1, \ldots, u_m, v_{11}, \ldots, v_{mn_m})'$ and drawn from a distribution in the exponential family with parameters $\boldsymbol{\alpha}_1$. It is also assumed that distribution of $\boldsymbol{u}$ depends on parameters $\boldsymbol{\alpha}_2$:

$$\begin{aligned} &\boldsymbol{y}_i|\boldsymbol{u} \sim f_{\boldsymbol{y}_i|\boldsymbol{u}}(\boldsymbol{y}_i|\boldsymbol{u}, \boldsymbol{\alpha}_1), \\ &\boldsymbol{u} \sim g_{\boldsymbol{u}}(\boldsymbol{u}|\boldsymbol{\alpha}_2). \end{aligned} \tag{2}$$

The goal of the analysis is to estimate the model parameters $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)'$ and predict the random effects $\boldsymbol{u}$. The likelihood for (2) is given by

$$L(\boldsymbol{\alpha}; \boldsymbol{y}) = \int \prod_{i=1}^{m} f_{\boldsymbol{y}_i|\boldsymbol{u}}(\boldsymbol{y}_i|\boldsymbol{u}, \boldsymbol{\alpha}_1) g_{\boldsymbol{u}}(\boldsymbol{u}|\boldsymbol{\alpha}_2) d\boldsymbol{u}.$$