Journal of Multivariate Analysis 148 (2016) 18–33

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On conditional prediction errors in mixed models with application to small area estimation

Shonosuke Sugasawa[a,*], Tatsuya Kubokawa[b]

[a] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa-shi, Tokyo 190-8562, Japan

[b] Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

ARTICLE INFO

Article history:
Received 30 November 2014
Available online 24 February 2016

AMS 2000 subject classifications:
62F12
62J05

Keywords:
Binomial–beta mixture model
Conditional mean squared error
Fay–Herriot model
Mixed model
Natural exponential family with quadratic variance function
Poisson–gamma mixture model
Random effect
Small area estimation

ABSTRACT

The empirical Bayes estimators in mixed models are useful for small area estimation in the sense of increasing precision of prediction for small area means, and one wants to know the prediction errors of the empirical Bayes estimators based on the data. This paper is concerned with conditional prediction errors in the mixed models instead of conventional unconditional prediction errors. In the mixed models based on natural exponential families with quadratic variance functions, it is shown that the difference between the conditional and unconditional prediction errors is significant under distributions far from normality. Especially for the binomial–beta mixed and the Poisson–gamma mixed models, the leading terms in the conditional prediction errors are, respectively, a quadratic concave function and an increasing function of the direct estimate in the small area, while the corresponding leading terms in the unconditional prediction errors are constants. Second-order unbiased estimators of the conditional prediction errors are also derived and their performances are examined through simulation and empirical studies.

1. Introduction

The empirical best linear unbiased predictors (EBLUP) or empirical Bayes estimators (EB) in the Bayesian context have been used for providing reliable small-area estimates in the normal linear mixed models. The unconditional mean squared errors (MSEs) have been widely used as a measure for prediction error of EBLUP, and the asymptotic approximations of the MSEs and their approximated unbiased estimators have been studied in a lot of papers under the assumption that the number of small areas is large. For example, see Prasad and Rao [14], Ghosh and Rao [8], Rao [15], Datta, Rao and Smith [4] and Hall and Maiti [10].

When data from the small area of interest are observed, the practitioners want to know how large prediction errors the EBLUP based on the observed data have. Concerning this issue, the conventional unconditional MSEs do not give us appropriate estimation errors, since it is an integrated measure. Booth and Hobert [2] suggested the conditional MSE given the data of the small area of interest, and Datta, Kubokawa, Molina and Rao [3] and Torabi and Rao [16] derived second-order unbiased estimators of the conditional MSE in the Fay–Herriot model [5] and nested error regression model [1] which are well-known normal linear mixed models. As pointed out in both papers, the difference between the conditional and unconditional MSEs is small in the normal linear mixed models, since it appears in the second-order terms. In the generalized linear mixed models (GLMM), however, Booth and Hobert [2] showed that the difference is significant for distributions far from normality, since it appears in the first-order or leading terms.

* Corresponding author.
E-mail addresses: shonosuke622@gmail.com (S. Sugasawa), tatsuya@e.u-tokyo.ac.jp (T. Kubokawa).

http://dx.doi.org/10.1016/j.jmva.2016.02.009

Although the GLMMs are useful for analyzing count data in small area estimation, it is computationally hard to derive the EBLUP and to evaluate their conditional MSEs, because the marginal likelihood and EBLUP in the GLMM cannot be expressed in closed forms. In fact, we need relatively high dimensional numerical integration to evaluate the conditional MSEs. Another point is the assumption that sample sizes of small areas are large, under which the Laplace approximation can be used to get asymptotically unbiased estimators of the conditional MSEs. However, this assumption is against the situation in small area estimation with small samples sizes.

An alternative model is the mixed model based on the natural exponential families with quadratic variance functions (NEF-QVF) suggested in Ghosh and Maiti [6,7]. In the NEF-QVF mixed models, the BLUP or the Bayes estimator can be expressed explicitly as the weighed average of a sample mean and a prior mean. Moreover, the MSE of the empirical Bayes estimator can be approximated analytically, and their asymptotically unbiased estimator can be obtained without assuming that samples of small areas are large. The NEF-QVF mixed models include the binomial–beta mixed and the Poisson–gamma mixed models, which are practically useful for analyzing mortality data in small areas.

Thus, in this paper, we treat the NEF-QVF mixed models instead of the GLMM and focus on the conditional prediction errors or the conditional MSEs (CMSE) of the empirical Bayes estimators (EB). Assuming that the number of small areas is large, but sample sizes in small areas are bounded, we not only derive second-order approximations of the conditional MSEs and their second-order unbiased estimators in closed forms, but also show that the difference between the conditional and unconditional MSEs is significant and appears in the first-order terms under distributions far from normality.

The paper is organized as follows: In Section 2, the CMSE of EB is addressed in the general mixed models, and the second-order approximation of the CMSE is derived under suitable conditions on estimators of model parameters and predictors. Second-order unbiased estimators of the CMSE are obtained in two ways of the analytical and parametric bootstrap methods.

In Section 3, the NEF-QVF mixed models are investigated as an application of the general results in Section 2. The second-order approximations of the CMSEs and their second-order unbiased estimators are obtained in analytical and closed forms without assuming that sample sizes of small areas tend to infinity. Ghosh and Maiti [6] derived the unconditional MSE of EB, and their estimation method and techniques for analysis are heavily used in Section 3. It is interesting to point out that the first-order term in the CMSE is an increasing function of the direct estimate in the small area for the Poisson–gamma mixed model, and it is a quadratic concave function for the binomial–beta mixed model, while the corresponding first-order terms in the unconditional MSEs are constants for both mixed models.

Simulation and empirical studies of the suggested procedures are given in Section 4. Two data sets are used for the empirical studies. One is the Stomach Cancer Mortality Data in Saitama Prefecture in Japan, and the Poisson–gamma mixed model is applied. The other is the Infant Mortality Data Before World War II in Ishikawa Prefecture in Japan, and we use the binomial–beta mixed model. Through these analyses, it is observed that the estimates of the conditional MSEs are more variable than those of the unconditional MSEs, since conditional MSE depends on the data of the area of interest. For some areas, the conditional MSE gives much higher risks than the unconditional MSE, namely, the conventional MSE seems to under-estimate the conditional MSE. Thus, we suggest providing estimates of the conditional MSE.

Finally, the concluding remarks are given in Section 5, and the technical proofs are given in the Appendix.

2. Conditional MSE of empirical Bayes estimator in general mixed models

Let $y = (y_1, \ldots, y_m)^\top$ be a vector of observable random variables, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$ be a vector of unobservable random variables. Let $\boldsymbol{\eta}$ be a q-dimensional vector of unknown parameters. In this paper, we treat continuous or discrete cases for y_i and $\boldsymbol{\theta}$. The conditional probability density (or mass) function of y_i given $(\theta_i, \boldsymbol{\eta})$ is denoted by $f(y_i|\theta_i, \boldsymbol{\eta})$, and the conditional probability density (or mass) function of θ_i given $\boldsymbol{\eta}$ is denoted by $\pi(\theta_i|\boldsymbol{\eta})$, namely,

$$\begin{aligned} y_i|(\theta_i, \boldsymbol{\eta}) &\sim f(y_i|\theta_i, \boldsymbol{\eta}) \\ \theta_i|\boldsymbol{\eta} &\sim \pi(\theta_i|\boldsymbol{\eta}) \end{aligned} \quad i = 1, \ldots, m. \tag{2.1}$$

This expresses the general parametric mixed models. Since it can be interpreted as a Bayesian model, we here use the terminology used in Bayes statistics. In the continuous case, the marginal density function of y_i for given $\boldsymbol{\eta}$ and the conditional (or posterior) density function of θ_i given $(y_i, \boldsymbol{\eta})$ are given by

$$\begin{aligned} m_\pi(y_i|\boldsymbol{\eta}) &= \int f(y_i|\theta_i, \boldsymbol{\eta})\pi(\theta_i|\boldsymbol{\eta})\mathrm{d}\theta_i \\ \pi(\theta_i|y_i, \boldsymbol{\eta}) &= f(y_i|\theta_i, \boldsymbol{\eta})\pi(\theta_i|\boldsymbol{\eta})/m_\pi(y_i|\boldsymbol{\eta}) \end{aligned} \quad i = 1, \ldots, m, \tag{2.2}$$

and we use the same notations in the discrete case. Then, for $i = 1, \ldots, m$, we consider the problem of predicting a scalar quantity $\xi_i(\theta_i, \boldsymbol{\eta})$ of each small area.

When $\xi_i(\theta_i, \boldsymbol{\eta})$ is predicted with $\widehat{\xi}_i = \widehat{\xi}_i(\boldsymbol{y})$, the predictor $\widehat{\xi}_i$ can be evaluated with the unconditional and conditional MSEs, described as

$$\mathrm{MSE}(\boldsymbol{\eta}, \widehat{\xi}_i) = \mathrm{E}\Big[\big\{\widehat{\xi}_i - \xi_i(\theta_i, \boldsymbol{\eta})\big\}^2\Big],$$

$$\mathrm{CMSE}(\boldsymbol{\eta}, \widehat{\xi}_i|y_i) = \mathrm{E}\Big[\big\{\widehat{\xi}_i - \xi_i(\theta_i, \boldsymbol{\eta})\big\}^2|y_i\Big],$$