CIRJE-F-702

A Review of Linear Mixed Models and Small Area Estimation

Tatsuya Kubokawa
University of Tokyo

December 2009

A Review of Linear Mixed Models and Small Area Estimation

Tatsuya Kubokawa*

December 25, 2009

Abstract

The linear mixed models (LMM) and the empirical best linear unbiased predictor (EBLUP) induced from LMM have been well studied and extensively used for a long time in many applications. Of these, EBLUP in small area estimation has been recognized as a useful tool in various practical statistics. In this paper, we give a review on LMM and EBLUP from a aspect of small area estimation. Especially, we explain why EBLUP is likely to be reliable. The reason is that EBLUP possesses the shrinkage function and the pooling effects as desirable properties, which arise from the setup of random effects and common paramers in LMM. Such important properties of EBLUP are clarified as well as some recent results of the mean squared error estimation, the confidence interval and the variable selection procedures are summarized.

Key words and phrases: Akaike information criterion, Bartlett correction, Baysian information criterion, best linear unbiased predictor, confidence interval, empirical Bayes procedure, Fay-Herriot model, linear mixed model, maximum likelihood estimator, mean squared error, nested error regression model, restricted maximum likelihood estimator, small area estimation, Wald test.

1 Introduction

The linear mixed models (LMM) and the empirical best linear unbiased predictor (EBLUP) or the empirical Bayes estimator (EB) induced from LMM have been studied for a long time in the literature. One of important applications of LMM is the problem of small area estimation. Small area refers to a small geographical area or a group for which little information is obtained from the sample survey. When only a few observations are available from a given small area, the direct estimator based only on the data from the small area is likely to be unreliable, so that the relevant supplementary information such as data from other related small areas is used via suitable linking models to increase the precision of the estimate. The typical models used for the small area estimation are

*Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: tatsuya@e.u-tokyo.ac.jp

the Fay-Herriot model and the nested error regression model (NERM), which are special models of LMM, and the model-based estimates including EBLUP or EB are found to be very useful as illustrated by Fay and Herriot (1979) and Battese, Harter and Fuller (1988). For a good review and account on this topic, see Ghosh and Rao (1994), Rao (1999, 2003) and Pfeffermann (2002).

In this paper, we give a review on theory of the linear mixed model and applications to small area estimation unnder the normality assumption. In Sections 2 and 3, we explain the derivation of the mixed model equation and BLUP, asymptotic properties of the maximum likelihood (ML) and restricted maximum likelihood (REML) estimators of variance components, and EBLUP's features and their relation with the structure of LMM. Especially, we explain why EBLUP is likely to be reliable. As discussed there, desirabel properties of EBLUP are characterized as the shrinkage function and the pooling effect, namely, EBLUP shrinks the sample mean of the small area towards a stable quantity costructed by pooling all the data. These two feartures of EBLUP, shrinkage and pooling effects, come from the structure of LMM described as (observation) = (common parameters) + (random effects) + (error terms), namely, the function of shrinkage arises from the random effects of LMM, and the pooling effect is due to the setup of the common parameters in LMM. As seen from that fact that EBLUP is interpreted as the empirical Bayes estimator, this perspective was recognized by Efron and Morris (1975) in the context of the empirical Bayes method. While BLUP or EBLUP was proposed by Henderson (1950), EBLUP is related to the shrinkage estimator studied by Stein (1956), who established analytically that EBLUP improves on the sample means when the number of small areas is larger than or equal to three. This fact shows not only that EBLUP has a larger precision than the sample mean, but also that a similar concept came out at the same time by Henderson (1950) for practical use and Stein (1956) for theoretical interest.

When EBLUP is used to estimate a small area mean based on real data, it is important to assess how much EBLUP is reliable. Two of typical methods for measuring uncertainty of EBLUP is the estimation of the mean squared error (MSE) and the confidence interval based on EBLUP. In Section 4, we explain the results of the second-order approximation of an unbiased estimator of MSE of EBLUP and the confidence interval which satisfies the nominal confidence level with the second-order accuracy.

In Section 5, we explain the testing problem of the regression coefficients and the selection of explanatory variables.

Since the topics and results treated in this paper are limited due to shortage of page length, see Searle, Casella and McCulloch (1992) and Demidenko (2004) for LMM; Rao (2003) for small area estimation; Banerjee, Carlin and Gelfand (2004) for spatial models; Hsiao (2003) for econometric models; McCulloch and Searle (2001), McCulloch (2003), Fahrmeir and Tutz (2001) and Molenberghs and Verbeke (2006) for the generalized linear mixed models; Lawson (2006), Lawson, Browne and Vidal Rodeiro (2003), Diggle, Lian and Zeger (1994), Verbeke and Molenberghs (2000) and Fitzmaurice, Laird and Ware (2004) for disease mapping and other applications.

2 Linear mixed models and BLUP

2.1 Linear mixed models

Consider the general linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{y}$ is an $N \times 1$ observation vector of the response variable, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $N \times p$ and $N \times M$ matrices, respectively, of the explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of the regression coefficients, $\boldsymbol{v}$ is an $M \times 1$ vector of the random effects, and $\boldsymbol{\epsilon}$ is an $N \times 1$ vector of the random errors. Here, $\boldsymbol{v}$ and $\boldsymbol{\epsilon}$ are mutually independently distributed as $\boldsymbol{v} \sim \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\theta}))$ and $\boldsymbol{\epsilon} \sim \mathcal{N}_N(\boldsymbol{0}, \boldsymbol{R}(\boldsymbol{\theta}))$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ is a q dimensional vector of unknown parameters, and $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\theta})$ and $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\theta})$ are positive definite matrices. Throughout the paper, for simplicity, it is assumed that $\boldsymbol{X}$ is of full rank. Then, $\boldsymbol{y}$ has a marginal distribution

$$\mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{2.2}$$

for

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta}) + \boldsymbol{Z}\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{Z}'. \tag{2.3}$$

Three of specific models of LMM are the nested error regression model (NERM), the Fay-Herriot model and a basic area model with time series structures.

Example 2.1 (NERM) This model is described by

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + v_i + \varepsilon_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i, \tag{2.4}$$

where k is the number of small areas, $N = \sum_{i=1}^{k} n_i$, $\boldsymbol{x}_{ij}$ is a $p \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ unknown common vector of regression coefficients, and v_i's and ε_{ij}'s are mutually independently distributed as $v_i \sim \mathcal{N}(0, \sigma_v^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Here, σ_v^2 and σ^2 are referred to as, respectively, 'between' and 'within' components of variance, and both are unknown, and (2.4) is also called the *Variance Components Model*. Let $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n_i})'$, $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_k')'$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,n_i})'$, $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_k')'$ and let $\boldsymbol{\epsilon}$ be similarly defined. Let $\boldsymbol{v} = (v_1, \ldots, v_k)'$ and $\boldsymbol{Z} = \text{block diag}(\boldsymbol{j}_1, \ldots, \boldsymbol{j}_k)$ for $\boldsymbol{j}_i = (1, \ldots, 1)' \in \boldsymbol{R}^{n_i}$. Then, the model is expressed in vector notations as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\epsilon}$, where the asymptotics for large k are considered.

Battese, *et al*. (1988) used the NERM in the framework of a finite population model to predict areas under corn and soybeans for each of $k = 12$ counties in north-central Iowa. In their analysis, each county is divided into about 250 hectares segments, and n_i segments are selected from the i-th county. For the j-th segment of the i-th county, y_{ij} is the number of hectares of corn (or soybeans) in the (i, j) segment reported by interviewing farm operators, and x_{ij1} and x_{ij2} are the number of pixels (0.45 hectar) classified as corn and soybeans, respectively, by using LANDSAT satellite data. Since n_i's range from 1 to 5 with $\sum_{i=1}^{k} n_i = 37$, the sample mean $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ has large deviation for

predicting the mean crop hectare per segment $\mu_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + v_i$ for $\overline{\boldsymbol{x}}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$. The NERM enables us to construct more reliable prediction procedures not only by using the auxiliary information on the LANDSAT data, but also by combining the data of the related areas. For a further account, see Section 3.2.

Example 2.2 (Fay-Herriot model) While NERM is an individual level model, the following basic area model is useful in the small area estimation:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \ldots, k, \tag{2.5}$$

where k is the number of small areas, $\boldsymbol{x}_i$ is a $p \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ unknown common vector of regression coefficients, and v_i's and ε_i's are mutually distributed random errors such that $v_i \sim \mathcal{N}(0, \theta)$ and $\varepsilon_i \sim \mathcal{N}(0, d_i)$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)'$, $\boldsymbol{y} = (y_1, \ldots, y_k)'$, and let $\boldsymbol{v}$ and $\boldsymbol{\epsilon}$ be similarly defined. Then, the model is expressed in vector notations as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v} + \boldsymbol{\epsilon},$$

and $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\theta) = \theta\boldsymbol{I}_k + \boldsymbol{D}$ for $\boldsymbol{D} = \mathrm{diag}\,(d_1, \ldots, d_k)$ and $N = k$.

Example 2.3 (A basic area model with time series structures) The Fay-Herriot type model with time series or longitudinal structures is described by

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it} + \varepsilon_{it}, \quad i = 1, \ldots, k, \ t = 1, \ldots, T, \tag{2.6}$$

where k is the number of small areas, t is a time index, $N = kT$, $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ unknown common vector of regression coefficients, and v_{it}'s and ε_{it}'s are random errors. Let $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n_i})'$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,n_i})'$, and let $\boldsymbol{v}_i$ and $\boldsymbol{\epsilon}_i$ be similarly defined. Then, the model is expressed in vector notations as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{v}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, k.$$

Here, it is assumed that $\boldsymbol{\epsilon}_i$ and $\boldsymbol{v}_i$ are mutually distributed as $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}_i)$ for a $T \times T$ known diagonal matrix $\boldsymbol{D}_i = \mathrm{diag}\,(d_{i1}, \ldots, d_{iT})$ and $\boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_v^2\boldsymbol{\Psi}(\rho))$ for unknown scalar σ_v^2 and a $T \times T$ unknown matrix $\boldsymbol{\Psi}(\rho)$ with a parameter ρ, $|\rho| < 1$. As typical cases of $\boldsymbol{\Psi}(\rho)$, we have

$$\boldsymbol{\Psi}(\rho) = (1-\rho)\boldsymbol{I}_T + \rho\boldsymbol{j}_T\boldsymbol{j}_T' \quad \text{and} \quad \boldsymbol{\Psi}(\rho) = \mathbf{mat}_{i,j}(\rho^{|i-j|}).$$

Letting $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_k')'$, $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_k')'$ and letting $\boldsymbol{v}$ and $\boldsymbol{\epsilon}$ be defined similarly, we can express the model as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v} + \boldsymbol{\epsilon}$.

2.2 Mixed model equation and BLUP

[1] BLUP. We now consider the estimation of the regression coefficients $\boldsymbol{\beta}$ and the prediction of the random effects $\boldsymbol{v}$ in (2.1). When the convariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ are known, there exists the best unbiased predictor of $\boldsymbol{v}$ among the linear functions of $\boldsymbol{y}$. This is called the *Best Linear Unbiased Predictor* (BLUP) and denoted by $\widehat{\boldsymbol{v}}$. Also, there

exists the best linear unbiased estimator of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}$. Henderson (1950) showed that $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{v}})$ can be derived by the solution of the equation given by

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{v}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}, \tag{2.7}$$

which is called the *Mixed Model Equation*, and the solution is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, \quad \widehat{\boldsymbol{v}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}), \tag{2.8}$$

where $\widehat{\boldsymbol{\beta}}$ is the generalized least squares (GLS) estimator of $\boldsymbol{\beta}$. When we want to estimate $\mu = \boldsymbol{a}'\boldsymbol{\beta} + \boldsymbol{b}'\boldsymbol{v}$ for known vectors $\boldsymbol{a} \in \boldsymbol{R}^p$ and $\boldsymbol{b} \in \boldsymbol{R}^q$, the BLUP of μ is given by

$$\widehat{\mu}^{EB} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}} + \boldsymbol{b}'\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}), \tag{2.9}$$

where we used the notation $\widehat{\mu}^{EB}$ since it can be interpreted as an empirical Bayes procedure as discussed below.

We here confirm that $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{v}})$ is the solution of the mixed model equation (2.7). The second equation in (2.7) is written as $\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})\widehat{\boldsymbol{v}} = \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y}$, which implies that

$$\widehat{\boldsymbol{v}} = (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}). \tag{2.10}$$

It is noted that

$$\begin{aligned} &(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\ &\quad = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{R}^{-1} - \boldsymbol{G}\left\{(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}) - \boldsymbol{G}^{-1}\right\}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\ &\quad = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{R}^{-1} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1} \\ &\quad = \boldsymbol{G}\boldsymbol{Z}'\left\{\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{G}^{-1}+\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}\right\} \\ &\quad = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}, \end{aligned}$$

where at the last equality, we used the useful equality

$$\boldsymbol{\Sigma}^{-1} = (\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'+\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{G}^{-1}+\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}. \tag{2.11}$$

Thus, $\widehat{\boldsymbol{v}}$ given in (2.10) is expressed as the form given in (2.8).

We next substitute $\widehat{\boldsymbol{v}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$ into the first equation of (2.7) given by $\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{v}} = \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y}$. Then,

$$\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y},$$

which yields

$$\boldsymbol{X}'\boldsymbol{R}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}')\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{R}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}')\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

It is noted that $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'+\boldsymbol{R}$, namely, $\boldsymbol{R}^{-1}(\boldsymbol{\Sigma}-\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}') = \boldsymbol{I}$. Thus, we get the equation $\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$, which means that the solution $\widehat{\boldsymbol{\beta}}$ is described as the form in (2.8).

Example 2.4 (BLUP in NERM) As explained in Example 2.1, the mean crop hectare per segment in NERM (2.4) is described by $\mu_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + v_i$ for $i = 1, \ldots, k$. Let $\theta_1 = \sigma^2$, $\theta_2 = \sigma_v^2$, and let $\boldsymbol{\theta} = (\theta_1, \theta_2)'$. In this mode, $\boldsymbol{G}(\boldsymbol{\theta}) = \theta_2 \boldsymbol{I}_k$, $\boldsymbol{\Sigma}_i(\boldsymbol{\theta}) = \theta_1 \boldsymbol{I}_{n_i} + \theta_2 \boldsymbol{j}_i \boldsymbol{j}_i'$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{block diag}(\boldsymbol{\Sigma}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{\Sigma}_k(\boldsymbol{\theta}))$. Noting that

$$\boldsymbol{\Sigma}_i^{-1} = \frac{1}{\theta_1}\left(\boldsymbol{I}_{n_i} - \frac{\theta_2}{\theta_1 + n_i\theta_2}\boldsymbol{j}_i\boldsymbol{j}_i'\right),$$

from (2.9), it follows that the BLUP of μ_i is given by

$$\widehat{\mu}_i^{EB}(\boldsymbol{\theta}) = \overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \frac{n_i\theta_2}{\theta_1 + n_i\theta_2}\left\{\overline{y}_i - \overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right\} \tag{2.12}$$

where $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}$, and the GLS of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \Big\{\sum_{i=1}^{k}\big(\boldsymbol{x}_i\boldsymbol{x}_i' - \frac{n_i^2\theta_2}{\theta_1 + n_i\theta_2}\overline{\boldsymbol{x}}_i\overline{\boldsymbol{x}}_i'\big)\Big\}^{-1}\sum_{i=1}^{k}\big(\boldsymbol{x}_i\boldsymbol{y}_i' - \frac{n_i\theta_2}{\theta_1 + n_i\theta_2}\overline{\boldsymbol{x}}_i\overline{y}_i\big).$$

[2] Derivation of the mixed model equation. We explain how the mixed model equation (2.7) can be derived. Two of typical approaches to the derivation are the maximum likelihood (ML) method and the empirical Bayes method.

To derive (2.7) based on the ML method, it is noted that the joint probability density functon of $(\boldsymbol{y}, \boldsymbol{v})$ is written as $(2\pi)^{-N/2}|\boldsymbol{G}|^{-1/2}|\boldsymbol{R}|^{-1/2}\cdot\exp\{-h(\boldsymbol{\beta}, \boldsymbol{v})/2\}$, where $h(\boldsymbol{\beta}, \boldsymbol{v}) = \boldsymbol{v}'\boldsymbol{G}^{-1}\boldsymbol{v} + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v})$. To minimize $h(\boldsymbol{\beta}, \boldsymbol{v})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{v})$, we need to differentiate it with respect to $\boldsymbol{\beta}$ and $\boldsymbol{v}$, which yields that

$$\begin{aligned}\frac{\partial h(\boldsymbol{\beta}, \boldsymbol{v})}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v}),\\ \frac{\partial h(\boldsymbol{\beta}, \boldsymbol{v})}{\partial \boldsymbol{v}} &= 2\boldsymbol{G}^{-1}\boldsymbol{v} - 2\boldsymbol{Z}'\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v}).\end{aligned}$$

Hence, it is seen that (2.7) is a matricial expression of $\partial h(\boldsymbol{\beta}, \boldsymbol{v})/\partial\boldsymbol{\beta} = \boldsymbol{0}$ and $\partial h(\boldsymbol{\beta}, \boldsymbol{v})/\partial\boldsymbol{v} = \boldsymbol{0}$.

The other method is based on the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{y}$. Since the covariance matrix of $(\boldsymbol{y}, \boldsymbol{v})$ is given by

$$\text{Cov}\,(\boldsymbol{y}, \boldsymbol{v}) = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{Z}\boldsymbol{G} \\ \boldsymbol{G}\boldsymbol{Z}' & \boldsymbol{G} \end{pmatrix}, \tag{2.13}$$

from the well known property of multivariate normal distribution, it follows that the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{y}$ is written as

$$\boldsymbol{v}|\boldsymbol{y} \sim \mathcal{N}_q\left(\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}\right).$$

It is noted that in the Bayesian context, this conditional distribution corresponds to the posterior distribution. Using (2.11), we can see that the marginal distribution of $\boldsymbol{y}$ is given by $\boldsymbol{y} \sim \mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, whose density function is described as $(2\pi)^{-N/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-(\boldsymbol{y} -$

$\boldsymbol{X\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})/2\}$. Thus, the ML estimator of $\boldsymbol{\beta}$ based on this marginal density function is identical to the GLS estimator $\widehat{\boldsymbol{\beta}}$. Since the Bayes estimator is the mean of the posterior distribution, the expectation of the posterior distribution, givne by

$$E[\boldsymbol{v}|\boldsymbol{y}] = \boldsymbol{GZ}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}), \tag{2.14}$$

is the Bayes estimator of $\boldsymbol{v}$. Substituting $\widehat{\boldsymbol{\beta}}$ into the Bayes estimator, we get the empirical Bayes estimator $\boldsymbol{GZ}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})$, which is identical to $\widehat{\boldsymbol{v}}$ given in (2.8). Hence, the solution of the mixed model equation can be derived as the empirical Bayes estimator.

The distiction of the two methods described above is that the ML method estimates $\boldsymbol{v}$ by the mode of the posterior distribution, while the empirical Bayes method estimate $\boldsymbol{v}$ by the mean of the posterior distribution. Although both methods gives the same solution in normal distributions, their solutions are different in general. In the context of Bayesian statistics, the former method is called the *Bayesian Maximum Likelihood* method.

It is noted that the conditional expectation (2.14) means that we can predict the unobserbable variable $\boldsymbol{v}$ if $\boldsymbol{v}$ has a correlation with $\boldsymbol{y}$, namely, the structure of the covariance matrix given in (2.13) is essential for the predictability. This consideration has been widely used in various fields like finite population models and incomplete data problems.

3 Estimation of parameters and EBLUP

3.1 Estimation of the variance components

[1] ML and REML methods In the LMM given in (2.1), the covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ are, in general, functions of unknown parameters like variance components. The unknown parameters are here denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$, namely, the covariance matrix of $\boldsymbol{y}$ is described as

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta}) + \boldsymbol{ZG}(\boldsymbol{\theta})\boldsymbol{Z}.$$

The typical methods for estimating $\boldsymbol{\theta}$ are based on the *Maximum Likelihood* (ML) and *Restricted Maximum Likelihood* (REML) methods. Substituting the GLS $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ into the marginal density function whose distribution is $\mathcal{N}_N(\boldsymbol{X\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, we can see that the ML estimator of $\boldsymbol{\theta}$ is derived as a solution of minimizing the function $\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))$. On the other hand, let $\boldsymbol{K}$ be an $N \times (N-p)$ matrix satisfying $\boldsymbol{K}'\boldsymbol{X} = \boldsymbol{0}$. Then $\boldsymbol{K}'\boldsymbol{y} \sim \mathcal{N}_{N-p}(\boldsymbol{0}, \boldsymbol{K}'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{K})$, and the REML method is the ML method based on this distribution, namely, the REML estimator is derived as a solution of minimizing the function $\log|\boldsymbol{K}'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{K}| + \boldsymbol{y}'\boldsymbol{K}(\boldsymbol{K}'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{y}$. Let

$$\boldsymbol{\Pi} = \boldsymbol{\Pi}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\left\{\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\right\}^{-}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}, \tag{3.1}$$

and note that

$$(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})) = \boldsymbol{y}'\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{y}, \quad \boldsymbol{\Pi}(\boldsymbol{\theta}) = \boldsymbol{K}(\boldsymbol{K}'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{K})^{-1}\boldsymbol{K}'.$$

Also note that $\partial_i \log|\boldsymbol{\Sigma}| = \mathrm{tr}\,[\boldsymbol{\Sigma}^{-1}\partial_i\boldsymbol{\Sigma}]$, $\partial_i\boldsymbol{\Pi} = -\boldsymbol{\Pi}(\partial_i\boldsymbol{\Sigma})\boldsymbol{\Pi}$, $\partial_i \log|\boldsymbol{K}'\boldsymbol{\Sigma}\boldsymbol{K}| = \mathrm{tr}\,[\boldsymbol{\Pi}\partial_i\boldsymbol{\Sigma}]$ where ∂_i denotes the differential operator $\partial_i = \partial/\partial\theta_i$. Thus, the ML and REML estimators are sokutions of the following equations:

$$\text{(3.2)} \qquad \text{[ML]} \qquad \boldsymbol{y}'\boldsymbol{\Pi}(\boldsymbol{\theta})\{\partial_i\boldsymbol{\Sigma}(\boldsymbol{\theta})\}\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{y} = \mathrm{tr}\,\left[\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\{\partial_i\boldsymbol{\Sigma}(\boldsymbol{\theta})\}\right],$$

$$\text{(3.3)} \qquad \text{[REML]} \qquad \boldsymbol{y}'\boldsymbol{\Pi}(\boldsymbol{\theta})\{\partial_i\boldsymbol{\Sigma}(\boldsymbol{\theta})\}\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{y} = \mathrm{tr}\,[\boldsymbol{\Pi}(\boldsymbol{\theta})\{\partial_i\boldsymbol{\Sigma}(\boldsymbol{\theta})\}]\,.$$

Since the l.h.s. of the above equations can be expressed as $\boldsymbol{y}'\boldsymbol{\Pi}(\boldsymbol{\theta})\{\partial_i\boldsymbol{\Sigma}(\boldsymbol{\theta})\}\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{y} = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\{\partial_i\boldsymbol{\Sigma}(\boldsymbol{\theta})\}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})) = -(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))'\{\partial_i\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))$, we can use a convenient expression among these. For discussions about which is better, ML or REML, see Section 6.10 in McCulloch and Searle (2001). In estimation of variance components, REML seems better in that REML is closer to an unbiased estimator than ML, while both have the same covariance matrix as explained below.

[2] Asymptotic properties of the ML and REML estimators The consistency and asymptotic normality of the ML and REML has been studied by Sweeting (1980), Mardia and Marshall (1984) and Cressie and Lahiri (1993). We here explain the asymptotic properties using the results of Kubokawa (2009b). To this end, we use the notations

$$\mathbf{col}_i(a_i) = \begin{pmatrix} a_1 \\ \vdots \\ a_q \end{pmatrix}, \quad \mathbf{mat}_{ij}(b_{ij}) = \begin{pmatrix} b_{11} & \cdots & b_{1q} \\ \vdots & \ddots & \vdots \\ b_{q1} & \cdots & b_{qq} \end{pmatrix},$$

and $\boldsymbol{C}_{(i)} = \partial\boldsymbol{C}/\partial\theta_i$ and $\boldsymbol{C}_{(ij)} = \partial^2\boldsymbol{C}/\partial\theta_i\partial\theta_j$ for matrix $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{\theta})$. Let $\lambda_1 \le \cdots \le \lambda_N$ be the eigenvalues of $\boldsymbol{\Sigma}$ and let those of $\boldsymbol{\Sigma}_{(i)}$ and $\boldsymbol{\Sigma}_{(ij)}$ be λ_a^i and λ_a^{ij} for $a = 1, \ldots, N$ respectively, where $|\lambda_1^i| \le \cdots \le |\lambda_N^i|$, $|\lambda_1^{ij}| \le \cdots \le |\lambda_N^{ij}|$. Then, we assume the following conditions for large N and $0 \le i, j \le q$:

(C1) The elements of $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{G}(\boldsymbol{\theta})$, $\boldsymbol{R}(\boldsymbol{\theta})$, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, $\boldsymbol{\Sigma}_{(i)}(\boldsymbol{\theta})$, $\boldsymbol{\Sigma}_{(ij)}(\boldsymbol{\theta})$, $\boldsymbol{a}$, $\boldsymbol{b}$, p and q are bounded, and $\boldsymbol{X}'\boldsymbol{X}$ is positive definite and $\boldsymbol{X}'\boldsymbol{X}/N$ converges to a positive definite matrix;

(C2) $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$, and $\lim_{N\to\infty}\lambda_N < \infty$, $\lim_{N\to\infty}|\lambda_N^i| < \infty$ and $\lim_{N\to\infty}|\lambda_N^{ij}| < \infty$.

(C3) The $q \times q$ matrix $\boldsymbol{A}_2 = \mathbf{mat}_{ij}(\mathrm{tr}\,[\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_{(j)}\boldsymbol{\Sigma}])$ is positive definite and $\boldsymbol{A}_2/N$ converges to a positive definite matrix.

Since the conditions of Theorem 2 in Mardia and Marshall (1984) are satisfied by (C1), (C2) and (C3), it can be seen that $\widehat{\boldsymbol{\theta}}^M - \boldsymbol{\theta} = O_p(N^{-1/2})$.

Under further appropriate assumptions, $\widehat{\boldsymbol{\theta}}^M - \boldsymbol{\theta}$ can be asymptotically expanded as

$$\text{(3.4)} \qquad \widehat{\boldsymbol{\theta}}^M - \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{M*} + \widehat{\boldsymbol{\theta}}^{M**} + O_p(N^{-3/2}),$$

where $\widehat{\boldsymbol{\theta}}^{M*} = O_p(N^{-1/2})$ and $\widehat{\boldsymbol{\theta}}^{M**} = O_p(N^{-1})$, and their terms are given by

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{M*} &= \boldsymbol{A}_2^{-1}\boldsymbol{a}_1 = \boldsymbol{A}_2^{-1}\mathbf{col}_i(-\mathrm{tr}\,[(\boldsymbol{\Sigma}^{-1})_{(i)}(\boldsymbol{y}\boldsymbol{y}' - \boldsymbol{\Sigma})]), \\ \widehat{\boldsymbol{\theta}}^{M**} &= -\boldsymbol{A}_2^{-1}\Big\{\boldsymbol{a}_0 - \frac{\boldsymbol{b}_0}{2} + \boldsymbol{A}_1\boldsymbol{A}_2^{-1}\boldsymbol{a}_1\Big\}, \end{aligned}$$

for $\boldsymbol{a}_1 = \mathbf{col}_i(-\mathrm{tr}\,[(\boldsymbol{\Sigma}^{-1})_{(i)}(\boldsymbol{y}\boldsymbol{y}'-\boldsymbol{\Sigma})])$, $\boldsymbol{a}_0 = \mathbf{col}_i(\mathrm{tr}\,[\boldsymbol{Q}_i\boldsymbol{y}\boldsymbol{y}'])$, $\boldsymbol{A}_2 = \mathbf{mat}_{ia}(-\mathrm{tr}\,[\boldsymbol{\Sigma}_{(a)}(\boldsymbol{\Sigma}^{-1})_{(i)}])$, $\boldsymbol{A}_1 = \mathbf{mat}_{ia}(\mathrm{tr}\,[(\boldsymbol{\Sigma}^{-1})_{(ia)}(\boldsymbol{y}\boldsymbol{y}'-\boldsymbol{\Sigma})])$, $\boldsymbol{b}_0 = \mathbf{col}_i(\sum_{a,b} B_{iab}\hat{\theta}_a^{M*}\hat{\theta}_b^{M*})$. Here, $\boldsymbol{Q}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{(i)}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Sigma}^{-1} - \boldsymbol{P}\boldsymbol{\Sigma}_{(i)}\boldsymbol{P}$ for $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}$, and $B_{iab} = \mathrm{tr}\,[\boldsymbol{\Sigma}_{(ab)}(\boldsymbol{\Sigma}^{-1})_{(i)}] + \mathrm{tr}\,[\boldsymbol{\Sigma}_{(a)}(\boldsymbol{\Sigma}^{-1})_{(ib)}] + \mathrm{tr}\,[\boldsymbol{\Sigma}_{(b)}(\boldsymbol{\Sigma}^{-1})_{(ia)}]$. For the details, see Datta and Lahiri (2000), Das, Jiang and Rao (2004) and Kubokawa (2009b). Das *et al.* (2004) succeeded in the derivation under the rigorous conditions, while Kubokawa (2009b) developed the third-order expansion like $\widehat{\boldsymbol{\theta}}^M - \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{M*} + \widehat{\boldsymbol{\theta}}^{M**} + \widehat{\boldsymbol{\theta}}^{M***} + O_p(N^{-2})$. Using the equality $E\big[\mathrm{tr}\,[\boldsymbol{C}_1(\boldsymbol{y}\boldsymbol{y}' - \boldsymbol{\Sigma})]\mathrm{tr}\,[\boldsymbol{C}_2(\boldsymbol{y}\boldsymbol{y}' - \boldsymbol{\Sigma})]\big] = 2\mathrm{tr}\,[\boldsymbol{C}_1\boldsymbol{\Sigma}\boldsymbol{C}_2\boldsymbol{\Sigma}]$ under the distribution $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ for matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, we can observe that

$$\begin{aligned}
E[\widehat{\boldsymbol{\theta}}^{M*}] =& \boldsymbol{0}, \quad \mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^{M*}) = 2\boldsymbol{A}_2^{-1}, \\
E[\widehat{\boldsymbol{\theta}}^{M**}] =& \boldsymbol{A}_2^{-1}\mathbf{col}_i(\mathrm{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1})_{(i)}\boldsymbol{X}]) \\
& + \boldsymbol{A}_2^{-1}\mathbf{col}_i(\mathrm{tr}\,[\boldsymbol{A}_2^{-1}\mathbf{mat}_{a,b}(\mathrm{tr}\,[\boldsymbol{\Sigma}_{(ab)}(\boldsymbol{\Sigma}^{-1})_{(i)}])]).
\end{aligned}$$

It is noted that $E[\widehat{\boldsymbol{\theta}}^{M**}] = \boldsymbol{A}_2^{-1}\mathbf{col}_i(\mathrm{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1})_{(i)}\boldsymbol{X}])$ when $\boldsymbol{\Sigma}$ or $\boldsymbol{G}$ and $\boldsymbol{R}$ are matrices of linear functions of $\boldsymbol{\theta}$.

For the REML estimator, $\widehat{\boldsymbol{\theta}}^R - \boldsymbol{\theta}$ can be asymptotically expanded as

$$\widehat{\boldsymbol{\theta}}^R - \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{R*} + \widehat{\boldsymbol{\theta}}^{R**} + O_p(N^{-2}),$$

where $\widehat{\boldsymbol{\theta}}^{R*} = \widehat{\boldsymbol{\theta}}^{M*} = \boldsymbol{A}_2^{-1}\boldsymbol{a}_1$ and

$$\widehat{\boldsymbol{\theta}}^{R**} = -\boldsymbol{A}_2^{-1}\big\{\boldsymbol{a}_0^* - \boldsymbol{b}_0/2 + \boldsymbol{A}_1\boldsymbol{A}_2^{-1}\boldsymbol{a}_1\big\},$$

where $\boldsymbol{a}_0^* = \mathbf{col}_i(\mathrm{tr}\,[\boldsymbol{Q}_i(\boldsymbol{y}\boldsymbol{y}' - \boldsymbol{\Sigma})])$. Thus, $E[\widehat{\boldsymbol{\theta}}^{R*}] = \boldsymbol{0}$, $\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^{R*}) = \mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^{M*}) = 2\boldsymbol{A}_2^{-1}$ and

$$E[\widehat{\boldsymbol{\theta}}^{R**}] = \boldsymbol{A}_2^{-1}\mathbf{col}_i(\mathrm{tr}\,[\boldsymbol{A}_2^{-1}\mathbf{mat}_{a,b}(\mathrm{tr}\,[\boldsymbol{\Sigma}_{(ab)}(\boldsymbol{\Sigma}^{-1})_{(i)}])]), \tag{3.5}$$

where $E[\widehat{\boldsymbol{\theta}}^{R**}] = \boldsymbol{0}$ when $\boldsymbol{\Sigma}$ are matrices of linear functions of $\boldsymbol{\theta}$.

Example 3.1 (NERM) In the NERM, the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ and $\boldsymbol{\Sigma}$ correspond to $\theta_1 = \sigma^2$, $\theta_2 = \sigma_v^2$ and $\boldsymbol{\Sigma} = \mathrm{blockdiag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k)$ for $\boldsymbol{\Sigma}_i = \theta_1\boldsymbol{I}_i + \theta_2\boldsymbol{j}_i\boldsymbol{j}_i'$, $\boldsymbol{I}_i$ being the $n_i \times n_i$ identity matrix. The ML estimators $\widehat{\boldsymbol{\theta}}^M = (\hat{\theta}_1^M, \hat{\theta}_2^M)'$ of $(\theta_1, \theta_2)'$ are given as the solutions of the equations $L_1(\widehat{\boldsymbol{\theta}}^M) = 0$ and $L_2(\widehat{\boldsymbol{\theta}}^M) = 0$ where

$$\begin{aligned}
L_1(\boldsymbol{\theta}) =& \frac{1}{\theta_1^2}\sum_{i=1}^k \|\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \frac{n_i\theta_2}{\theta_1 + n_i\theta_2}(\overline{y}_i - \overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))\boldsymbol{j}_i\|^2 - \sum_{i=1}^k \frac{n_i}{\theta_1}(1 - \frac{\theta_2}{\theta_1 + n_i\theta_2}), \\
L_2(\boldsymbol{\theta}) =& \sum_{i=1}^k \frac{n_i^2}{(\theta_1 + n_i\theta_2)^2}\{\overline{y}_i - \overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}^2 - \sum_{i=1}^k \frac{n_i}{\theta_1 + n_i\theta_2},
\end{aligned}$$

since $\boldsymbol{\Sigma}_{(1)} = \boldsymbol{I}$ and $\boldsymbol{\Sigma}_{(2)} = \text{block diag}(\boldsymbol{j}_1\boldsymbol{j}_1', \ldots, \boldsymbol{j}_k\boldsymbol{j}_k')$. Note that $\boldsymbol{A}_2$ and $\boldsymbol{a}_1$ can be written as

$$\begin{aligned}
\boldsymbol{A}_2 =& \mathbf{mat}_{ij}(\text{tr}\,[\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{(j)}\boldsymbol{\Sigma}^{-1}]) \\
=& \begin{pmatrix} (N-k)\theta_1^{-2} + \sum_i (\theta_1 + n_i\theta_2)^{-2} & \sum_i n_i(\theta_1 + n_i\theta_2)^{-2} \\ \sum_i n_i(\theta_1 + n_i\theta_2)^{-2} & \sum_i n_i^2(\theta_1 + n_i\theta_2)^{-2} \end{pmatrix}, \\
\boldsymbol{a}_1 =& \begin{pmatrix} \sum_i \text{tr}\,[\boldsymbol{\Sigma}_i^{-2}(\boldsymbol{y}_i\boldsymbol{y}_i' - \boldsymbol{\Sigma}_i)] \\ \sum_i \boldsymbol{j}_i'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i\boldsymbol{y}_i' - \boldsymbol{\Sigma}_i)\boldsymbol{\Sigma}_i^{-1}\boldsymbol{j}_i \end{pmatrix}.
\end{aligned}$$

Since $\widehat{\boldsymbol{\theta}}^{M*} = \boldsymbol{A}_2^{-1}\boldsymbol{a}_1$, it is observed that $E[\widehat{\boldsymbol{\theta}}^{M*}] = \mathbf{0}$ and

$$\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^{M*}) = \frac{2\theta_1^2}{d(\psi)} \begin{pmatrix} \sum_{i=1}^k n_i^2\gamma_i^2 & -\sum_{i=1}^k n_i\gamma_i^2 \\ -\sum_{i=1}^k n_i\gamma_i^2 & (N - k + \sum_{i=1}^k \gamma_i^2) \end{pmatrix},$$

where $d(\psi) = (N - k + \sum_{i=1}^k \gamma_i^2)\sum_{i=1}^k n_i^2\gamma_i^2 - (\sum_{i=1}^k n_i\gamma_i^2)^2$ and $\gamma_i = (1 + n_i\psi)^{-1}$ for $\psi = \theta_2/\theta_1$. Also,

$$E[\widehat{\boldsymbol{\theta}}^{M**}] = \frac{\theta_1}{d(\psi)} \begin{pmatrix} -p\sum_{i=1}^k n_i^2\gamma_i^2 + (\sum_{i=1}^k n_i\gamma_i)c(\psi) \\ p\sum_{i=1}^k n_i\gamma_i^2 - (N - k + \sum_{i=1}^k \gamma_i)c(\psi) \end{pmatrix},$$

where $c(\psi) = \text{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\sum_{i=1}^k n_i^2\gamma_i^2\overline{\boldsymbol{x}}_i\overline{\boldsymbol{x}}_i']$. These were obtained by Datta and Lahiri (2000).

The REML estimators $\widehat{\boldsymbol{\theta}}^R = (\hat{\theta}_1^R, \hat{\theta}_2^R)'$ of $(\theta_1, \theta_2)'$ are given as the solutions of the equations given by

$$\begin{aligned}
0 =& L_1(\boldsymbol{\theta}) + \text{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-2}\boldsymbol{X}], \\
0 =& L_2(\boldsymbol{\theta}) + \text{tr}\,[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\text{block diag}(\boldsymbol{j}_1\boldsymbol{j}_1', \ldots, \boldsymbol{j}_k\boldsymbol{j}_k')\boldsymbol{\Sigma}^{-1}\boldsymbol{X}].
\end{aligned}$$

Noting that $\widehat{\boldsymbol{\theta}}^{R*} = \boldsymbol{A}_2^{-1}\boldsymbol{a}_1 = \widehat{\boldsymbol{\theta}}^{M*}$, we can see that $E[\widehat{\boldsymbol{\theta}}^{R*}] = \mathbf{0}$, $\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^{R*}) = \mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^{M*})$ and $E[\widehat{\boldsymbol{\theta}}^{R**}] = O(N^{-2})$ as shown in Datta and Lahiri (2000).

As estimation methods other than ML and REML, Henderson's methods and Rao's MINQUE methods are well known procedures in estimation of variance components. Especially, Henderson's methods provide explicit expressions of unbiased estimators. Prasad and Rao (1990) derived estimators with explicit forms using the Henderson method (III), which is given as follows: Let $S = \boldsymbol{y}'(\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y}$ and $S_1 = \boldsymbol{y}'(\boldsymbol{E} - \boldsymbol{E}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{E}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{E})\boldsymbol{y}$ where $\boldsymbol{E} = \text{block diag}(\boldsymbol{E}_1, \ldots, \boldsymbol{E}_k)$ for $\boldsymbol{E}_i = \boldsymbol{I}_i - n_i^{-1}\boldsymbol{j}_i\boldsymbol{j}_i'$. Then, unbiased estimators of θ_1 and θ_2 are given by

$$\hat{\theta}_1^U = S_1/(N - k - p) \quad \text{and} \quad \hat{\theta}_2^U = \{S - (N - p)\hat{\theta}_1^U\}/N^*,$$

where $N_* = N - \text{tr}\,\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{i=1}^k n_i^2\overline{\boldsymbol{x}}_i\overline{\boldsymbol{x}}_i'\}$ as suggested by Prasad and Rao (1990) . In this case, $\hat{\theta}_i^U - \theta_i = \hat{\theta}_i^{U*}$ for $i = 1, 2$, and it is easy to see that $E[\hat{\theta}_1^{U*}] = 0$, $E[\hat{\theta}_2^{U*}] = 0$ and

$$\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^{U*}) = \frac{2\theta_1^2}{N-k} \begin{pmatrix} 1 & -k/N \\ -k/N & \{k^2 + (N-k)\sum_{i=1}^k (1 + n_i\theta_2/\theta_1)^2\}/N^2 \end{pmatrix} + O(N^{-2}).$$

Since $\hat{\theta}_2^U$ takes a nagative value with a positive probability, it is reasonable to use the truncated estimator $\hat{\theta}_2^{TR} = \max\{\hat{\theta}_2^U, 0\}$.

3.2 EBLUP's features and their relation with the structure of LMM

The *Estimated (or Empirical) Best Linear Unbiased Predictor* (EBLUP) is derived by substituting estimator $\widehat{\boldsymbol{\theta}}$ into BLUP given in (2.9), namely, EBLUP of $\mu = \boldsymbol{a}'\boldsymbol{\beta} + \boldsymbol{b}'\boldsymbol{v}$ is given by

$$\widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{a}'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}) + \boldsymbol{b}'\boldsymbol{G}(\widehat{\boldsymbol{\theta}})\boldsymbol{Z}'\{\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\}^{-1}\{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})\}. \tag{3.6}$$

From Example 2.4, the EBLUP of $\mu_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + v_i$ is written as

$$\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}}) = \overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}) + \frac{n_i\hat{\theta}_2}{\hat{\theta}_1 + n_i\hat{\theta}_2}\left\{\overline{y}_i - \overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})\right\}, \tag{3.7}$$

where $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ given in Example 3.1. It is note that $Var(\overline{y}_i) = \theta_1/n_i + \theta_2$. When n_i is small or $\hat{\theta}_2/\hat{\theta}_1$ is large, the sample mean $\overline{y}_i$ is not reliable because of an unaccetpable error variance, while the EBLUP $\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}})$ approaches to $\overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$, which is stable because the GLS $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$ is constructed based on all the observations. When n_i is large or $\hat{\theta}_2/\hat{\theta}_1$ is small, on the other hand, $\overline{y}_i$ is likely to be reliable, and $\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}})$ approaches to $\overline{y}_i$. The feature depending on each small area tends to appear in $\overline{y}_i$ rather than $\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}})$. This shows that $\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}})$ gives stable and reliable predicted values by appropriately adjusting the weight of $\overline{y}_i$ and $\overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$.

Such desirabel properties of EBLUP are characterized as the shrinkage function and the pooling effect, namely, $\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}})$ shrinks $\overline{y}_i$ towards $\overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$, which is costructed by pooling all the data. The two feartures of EBLUP, shrinkage and pooling effects, come from the structure of the linear mixed model described as (observation) = (common mean) + (random effect) + (error term).

[1] Shrinkage via random effects. In the case that v_i is a fixed parameter and $\boldsymbol{\beta} = \boldsymbol{0}$, the best estimator of μ_i is $\overline{y}_i$. When v_i is a random effect, however, the covariance matrix of $(\overline{y}_i, v_i)$ is

$$\mathbf{Cov}\,(\overline{y}_i, v_i) = \begin{pmatrix} \theta_2 + \theta_1/n_i & \theta_2 \\ \theta_2 & \theta_2 \end{pmatrix},$$

namely, the correlation yilds between $\overline{y}_i$ and v_i. From this correlation, it follows that the conditional expectation is written as $E[v_i|\overline{y}_i] = \theta_2 n_i(\theta_1 + \theta_2 n_i)^{-1}(\overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta})$, which means that the conditional expectation shrinks $\overline{y}_i$ towards $\overline{\boldsymbol{x}}_i'\boldsymbol{\beta}$. Thus, the random effect v_i produces the function of shrinkage in EBLUP.

[2] Pooling data via common parameters. The regression coefficients $\boldsymbol{\beta}$ is embeded as a common parameter in all the small ares. To estimate the common parameter, all the data are used, and this results in the pooling effect. Thus, the setup via the common parameters leads to the pooling effect, and we get the stable estimator $\overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$ based on the weighted least squares estimator $\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})$.

As stated above, we can obtain stable estimates via pooling data through restricting parameters to some constraints like equality or inequality, and we can shrink $\overline{y}_i$ toward

the stable estimates through incorporating random effects. This enables us to boost up the precision of the prediction. As seen from that fact that EBLUP is interpreted as the empirical Bayes estimator, this perspective was recognized by Efron and Morris (1975) in the context of the empirical Bayes method, and the usefulness of the Bayesian methods may be based on such perspective.

[3] Henderson's EBLUP and Stein's shrinkage. Consider the case that $\boldsymbol{\beta} = \mathbf{0}$, $p = 0$, $n_1 = \cdots = \boldsymbol{n}_k = n$ and $N = nk$, and treat the unbiased estimator $\hat{\theta}_1^U$ and the truncated estimator $\hat{\theta}_2^{TR}$ in Example 3.1. Then $1 + n\theta_2/\theta_1$ is estimated by $\max\{1, 1 + n[(N-k)S/S_1 - N]/N\}$, which is equal to $\max\{1, (n/k)\sum_{j=1}^k \overline{y}_j^2/(S_1/(N-k)\}$ since $S = \sum_{i,j} y_{ij}^2 = S_1 + n\sum_i \overline{y}_i^2$ for $S_1 = \sum_{i,j}(y_{ij} - \overline{y}_i)^2$. Then, the EBLUP given in (3.7) can be expressed as for $\hat{\sigma}^2 = S_1/(N-k)$,

$$\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}}) = \max\Big\{0, 1 - \frac{k\hat{\sigma}^2}{n\sum_{j=1}^k \overline{y}_j^2}\Big\}\overline{y}_i,$$

which is related to the positive-part Stein estimator. The Stein problem has been developed as one of interesting topics in theoretical statistics since Stein (1956) established that the shrinkage estimator can improve on the sample means in the context of the simultaneous estimation for $k \geq 3$. This fact shows not only that EBLUP has a larger precision than the sample mean, but also that a similar concept came out at the same time by Henderson (1950) for practical use and Stein (1956) for theoretical interest.

4 Mesurements for uncertainty of EBLUP

When EBLUP is used to estimate a small area mean based on real data, it is important to assess how much EBLUP is reliable. Two methods for the purpose are to provide the estimate of the mean squared error (MSE) of EBLUP and to construct the confidence interval based on EBBLUP, and the results with second-order accuracy are explained here.

4.1 MSE estimation for EBLUP

Concerning the MLS estimation of EBLUP, asymptotically unbiased estimators of the MSE with the second-order accuracy have been derived based on the Taylor series expansion by Kackar and Harville (1984), Prasad and Rao (1990), Harville and Jeske (1992), Datta and Lahiri (2000), Datta, Rao and Smith (2005) and Das, Jiang and Rao (2004). For some recent results including jackknife and bootstrap methods, see Lahiri and Rao (1995), Hall and Maiti (2006a) and Chen and Lahiri (2008). We first approximate the MSE of EBLUP with second-order accuracy.

Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be $p \times 1$ and $M \times 1$ vectors of fixed constants, and suppose that we want to estimate the scalar quantity $\mu = \boldsymbol{a}'\boldsymbol{\beta} + \boldsymbol{b}'\boldsymbol{v}$. Since the conditional distribution of $\boldsymbol{v}$ given $\boldsymbol{y}$ is given by

$$\boldsymbol{v}|\boldsymbol{y} \sim \mathcal{N}_M(\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), (\boldsymbol{G}(\boldsymbol{\theta})^{-1} + \boldsymbol{Z}'\boldsymbol{R}(\boldsymbol{\theta})^{-1}\boldsymbol{Z})^{-1}), \tag{4.1}$$

the conditional expectation $E[\mu|\boldsymbol{y}]$ is written as

$$\begin{aligned}\widehat{\mu}^B(\boldsymbol{\beta},\boldsymbol{\theta}) =& E[\mu|\boldsymbol{y}] = \boldsymbol{a}'\boldsymbol{\beta} + \boldsymbol{b}'\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ =& \boldsymbol{a}'\boldsymbol{\beta} + \boldsymbol{s}(\boldsymbol{\theta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}),\end{aligned} \tag{4.2}$$

where $\boldsymbol{s}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{Z}\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{b}$. This can be interpreted as the Bayes estimator of μ in the Bayesian context. Substituting the GLS $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{y}$ into $\widehat{\mu}^B(\boldsymbol{\beta},\boldsymbol{\theta})$ yields the BLUP

$$\widehat{\mu}^{EB}(\boldsymbol{\theta}) = \widehat{\mu}^B(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta}) = \boldsymbol{a}'\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \boldsymbol{s}(\boldsymbol{\theta})'(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})), \tag{4.3}$$

which is also called an empirical Bayes estimator in the Bayesian context.

We first provide an accurate approximation of the mean squared error (MSE) of $\widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}})$ when N is large, where the MSE is given by

$$MSE(\boldsymbol{\theta},\widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}})) = E[\{\widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}}) - \mu\}^2].$$

For the purpose, we assume (C1), (C2) and the following conditions for large N and $1 \leq i,j \leq q$:

(C4) $\boldsymbol{a} - \boldsymbol{X}'\boldsymbol{s}(\boldsymbol{\theta}) = O(1)$, $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{s}_{(i)}(\boldsymbol{\theta}) = O_p(1)$, $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{s}_{(ij)}(\boldsymbol{\theta}) = O_p(1)$, $\boldsymbol{s}_{(i)}(\boldsymbol{\theta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{s}_{(j)}(\boldsymbol{\theta}) = O(1)$ and $\boldsymbol{s}_{(ij)}(\boldsymbol{\theta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{s}_{(k)}(\boldsymbol{\theta}) = O(1)$ for $\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$, $\partial_i\{\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\}$ and $\partial_i\partial_j\{\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\}$, $1 \leq i,j \leq q$;

(C5) $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{y}) = (\hat{\theta}_1,\ldots,\hat{\theta}_q)'$ is an estimator of $\boldsymbol{\theta}$ which satisfies that $\widehat{\boldsymbol{\theta}}(-\boldsymbol{y}) = \widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ and $\widehat{\boldsymbol{\theta}}(\boldsymbol{y}+\boldsymbol{X}\boldsymbol{\alpha}) = \widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ for any p-dimensional vector $\boldsymbol{\alpha}$.

(C6) $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is expanded as

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^* + \widehat{\boldsymbol{\theta}}^{**} + \widehat{\boldsymbol{\theta}}^{***} + O_p(N^{-2}), \tag{4.4}$$

where $\widehat{\boldsymbol{\theta}}^* = O_p(N^{-1/2})$, $\widehat{\boldsymbol{\theta}}^{**} = O_p(N^{-1})$ and $\widehat{\boldsymbol{\theta}}^{***} = O_p(N^{-3/2})$. Let $\widehat{\boldsymbol{\theta}}^* = (\hat{\theta}_1^*,\ldots,\hat{\theta}_q^*)'$, $\widehat{\boldsymbol{\theta}}^{**} = (\hat{\theta}_1^{**},\ldots,\hat{\theta}_q^{**})'$. These satisfy that $E[\hat{\theta}_i^*] = O(N^{-1})$ and $\boldsymbol{s}_{(i)}(\boldsymbol{\theta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{\nabla}_y\hat{\theta}_j^* = O_p(N^{-1})$.

Defined $g_1(\boldsymbol{\theta})$, $g_2(\boldsymbol{\theta})$ and $g_3^*(\boldsymbol{\theta})$ by

$$\begin{aligned}g_1(\boldsymbol{\theta}) =& \boldsymbol{b}'(\boldsymbol{G}(\boldsymbol{\theta})^{-1} + \boldsymbol{Z}'\boldsymbol{R}(\boldsymbol{\theta})^{-1}\boldsymbol{Z})^{-1}\boldsymbol{b}, \\ g_2(\boldsymbol{\theta}) =& (\boldsymbol{a}-\boldsymbol{X}'\boldsymbol{s}(\boldsymbol{\theta}))'(\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}(\boldsymbol{a}-\boldsymbol{X}'\boldsymbol{s}(\boldsymbol{\theta})), \\ g_3^*(\boldsymbol{\theta}) =& \mathrm{tr}\left[\Big(\frac{\partial \boldsymbol{s}(\boldsymbol{\theta})'}{\partial\boldsymbol{\theta}}\Big)\boldsymbol{\Sigma}(\boldsymbol{\theta})\Big(\frac{\partial \boldsymbol{s}(\boldsymbol{\theta})'}{\partial\boldsymbol{\theta}}\Big)'\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)\right],\end{aligned} \tag{4.5}$$

for $\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*) = E[(\widehat{\boldsymbol{\theta}}^* - E[\widehat{\boldsymbol{\theta}}^*])(\widehat{\boldsymbol{\theta}}^* - E[\widehat{\boldsymbol{\theta}}^*])']$.

Theorem 4.1 *Under the conditions (C1), (C2) and (C4)-(C6), the MSE of $\widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}})$ is approximated as*

$$MSE(\boldsymbol{\theta},\widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}})) = g_1(\boldsymbol{\theta}) + g_2(\boldsymbol{\theta}) + g_3^*(\boldsymbol{\theta}) + O(N^{-3/2}). \tag{4.6}$$

We next provide an asymptotically unbiased estimator of $MSE(\boldsymbol{\theta}, \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}}))$ with the second-order accuracy. Define $g_{11}(\boldsymbol{\theta})$ and $g_{12}(\boldsymbol{\theta})$ by

$$
\begin{aligned}
g_{11}(\boldsymbol{\theta}) =& \big(\frac{\partial g_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\big)' E[\widehat{\boldsymbol{\theta}}^{*} + \widehat{\boldsymbol{\theta}}^{**}], \\
g_{12}(\boldsymbol{\theta}) =& \frac{1}{2}\mathrm{tr}\left[\boldsymbol{B}(\boldsymbol{\theta})\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^{*})\right],
\end{aligned} \tag{4.7}
$$

where the (i,j) element of $\boldsymbol{B}(\boldsymbol{\theta})$ is given by

$$
(\boldsymbol{B}(\boldsymbol{\theta}))_{i,j} = (\boldsymbol{b} - \boldsymbol{Z}'\boldsymbol{s}(\boldsymbol{\theta}))'(\partial_{ij}\boldsymbol{G}(\boldsymbol{\theta}))(\boldsymbol{b} - \boldsymbol{Z}'\boldsymbol{s}(\boldsymbol{\theta})) + \boldsymbol{s}(\boldsymbol{\theta})'(\partial_{ij}\boldsymbol{R}(\boldsymbol{\theta}))\boldsymbol{s}(\boldsymbol{\theta}). \tag{4.8}
$$

It is noted that $g_{12}(\boldsymbol{\theta}) = 0$ when $\boldsymbol{G}$ and $\boldsymbol{R}$ are matrices of linear functions of $\boldsymbol{\theta}$. Define $mse(\widehat{\boldsymbol{\theta}}, \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}}))$ by

$$
mse(\widehat{\boldsymbol{\theta}}, \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}})) = g_1(\widehat{\boldsymbol{\theta}}) + g^{\#}(\widehat{\boldsymbol{\theta}}), \tag{4.9}
$$

where

$$
g^{\#}(\boldsymbol{\theta}) = g_2(\boldsymbol{\theta}) + 2g_3^{*}(\boldsymbol{\theta}) - g_{11}(\boldsymbol{\theta}) - g_{12}(\boldsymbol{\theta}). \tag{4.10}
$$

It is noted that $g^{\#}(\boldsymbol{\theta}) = O(N^{-1})$.

Theorem 4.2 *Under the same conditions as in Theorem 4.1, $mse(\widehat{\boldsymbol{\theta}}, \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}}))$ is a second-order unbiased estimator of MSE, namely, Then,*

$$
E[mse(\widehat{\boldsymbol{\theta}}, \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}}))] = MSE(\boldsymbol{\theta}, \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}})) + O(N^{-3/2}). \tag{4.11}
$$

4.2 Corrected confidence intervals and an example in NERM

Another method for measuring uncertainty of EBLUP is to provide a confidence interval based on EBLUP, and the confidence intervals which satisfy the nominal confidence level with the second-order accuracy have been derived based on the Taylor expansion by Datta, Ghosh, Smith and Lahiri (2002), Basu, Ghosh and Mukerjee (2003) and Kubokawa (2009a,b). Recently, Hall and Maiti (2006b) and Chatterjee, Lahiri and Li (2008) developed the method based on parametric bootstrap. We here provide a confidence interval of $\mu = \boldsymbol{a}'\boldsymbol{\beta} + \boldsymbol{b}'\boldsymbol{v}$ which satisfies the nominal confidece level with the second-order accuracy.

Let $mse(\widehat{\boldsymbol{\theta}}) = mse(\widehat{\boldsymbol{\theta}}, \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}})) = g_1(\widehat{\boldsymbol{\theta}}) + g^{\#}(\widehat{\boldsymbol{\theta}})$ for $g^{\#}$ given in (4.10). Since $mse(\widehat{\boldsymbol{\theta}})$ is an asymptotically unbiased estimator of the MSE of the empirical Bayes estimator $\widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}})$, it is reasonable to consider the confidence interval of the form

$$
I^{EB}(\widehat{\boldsymbol{\theta}}) :\ \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}}) \pm z_{\alpha/2}\sqrt{mse(\widehat{\boldsymbol{\theta}})}. \tag{4.12}
$$

However, the coverage probability $P[\mu \in I^{EB}(\widehat{\boldsymbol{\theta}})]$ cannot be guaranteed to be greater than or equal to the nominal confidence coefficient $1-\alpha$. To fix this shortcoming, we adjust

the significance point $z_{\alpha/2}$ as $z_{\alpha/2}\{1+h(\widehat{\boldsymbol{\theta}})\}$ by using an appropriate correction function $h(\widehat{\boldsymbol{\theta}})$. That is, the corrected confidence interval is described as

$$I^{CEB}(\widehat{\boldsymbol{\theta}}):\ \widehat{\mu}^{EB}(\widehat{\boldsymbol{\theta}}) \pm z_{\alpha/2}\left[1+h(\widehat{\boldsymbol{\theta}})\right]\sqrt{mse(\widehat{\boldsymbol{\theta}})}.$$

Here, we define the function $h(\boldsymbol{\theta})$ by

$$h(\boldsymbol{\theta}) = \frac{z_\alpha^2+1}{8g_1(\boldsymbol{\theta})^2}\mathrm{tr}\,[\{\frac{\partial g_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\}\{\frac{\partial g_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\}'\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)]. \tag{4.13}$$

Theorem 4.3 *Under the same conditions as in Theorem 4.1, the corrected confidence interval $I^{CEB}(\widehat{\boldsymbol{\theta}})$ satisfies the nominal confidence coefficient up to the third-order, namely,*

$$P[\mu \in I^{CEB}(\widehat{\boldsymbol{\theta}})] = 1-\alpha+O(N^{-3/2}). \tag{4.14}$$

Finally, we conclude this section with stating a remark and an example in NERM. Although Theorems 4.1, 4.2 and 4.3 provide the results of the second-order approximatoins, Kuobokawa (2009b) showed that all the results still hold with the third-order accuracy under additional appropriate conditions where the validity of the approximations are neglected in the paper and the above theorems. Das, *et al.* (2004) succeeded in the derivation of the conditions for the rigorous proofs of Theorems 4.1 and 4.2.

Example 4.1 (NERM) It is easy to see that the conditions (C1)-(C4) are satisfied in the prediction of $\mu_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta}+v_i$ in NERM. The EBLUP of μ_i is $\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}}) = \overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}) + \{n_i\hat{\theta}_2/(\hat{\theta}_1+n_i\hat{\theta}_2)\}\left\{\overline{y}_i-\overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})\right\}$ from (3.7). The MSE approximation of $\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}})$, its unbiased estimator and the confidence interval based on $\widehat{\mu}_i^{EB}(\widehat{\boldsymbol{\theta}})$ with the second-order accuracy are provided from Theorems 4.1, 4.2 and 4.3, where the functions $g_1(\boldsymbol{\theta})$, $g_2(\boldsymbol{\theta})$, $g_3^*(\boldsymbol{\theta})$, $g_{11}(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta})$ are expressed as $g_1(\boldsymbol{\theta}) = \theta_1\theta_2(\theta_1+n_i\theta_2)^{-1}$, $g_2(\boldsymbol{\theta}) = \theta_1^2(\theta_1+n_i\theta_2)^{-2}\overline{\boldsymbol{x}}_s'(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\overline{\boldsymbol{x}}_s$,

$$\begin{aligned}
g_3^*(\boldsymbol{\theta}) =& n_i(\theta_1+n_i\theta_2)^{-3}(-\theta_2,\theta_1)\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)(-\theta_2,\theta_1)',\\
g_{11}(\boldsymbol{\theta}) =& (\theta_1+n_i\theta_2)^{-2}(n_i\theta_2^2,\theta_1^2)E[\widehat{\boldsymbol{\theta}}^*+\widehat{\boldsymbol{\theta}}^{**}],\\
h(\boldsymbol{\theta}) =& \frac{z_\alpha^2+1}{8(\theta_1\theta_2)^2(\theta_1+n_i\theta_2)^2}(n_i\theta_2^2,\theta_1^2)\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)(n_i\theta_2^2,\theta_1^2)',
\end{aligned}$$

and $g_{12}(\boldsymbol{\theta}) = 0$. For estimator $\widehat{\boldsymbol{\theta}}$ satisfying the conditions (C4) and (C5), we need to obtain $\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)$ and $E[\widehat{\boldsymbol{\theta}}^*+\widehat{\boldsymbol{\theta}}^{**}]$. The ML, REML and Prasad-Rao estimators satisfy given in Example 3.1 satisfy (C4) and (C5) and their covariances and biases are given there.

It should be remarked that the corrected confidence interval $I^{CEB}(\widehat{\boldsymbol{\theta}})$ tends to be instable near $\theta_2 = \mathbf{0}$, because the corrected function $h(\boldsymbol{\theta})$ given in (4.13) includes $g_1(\boldsymbol{\theta})$ in the denominator. In NERM, $g_1(\boldsymbol{\theta})$ is $g_1(\boldsymbol{\theta}) = \theta_1\theta_2/(\theta_1+n_i\theta_2)$ and takes values near zero when θ_2 is close to zero. This causes the instability of the confidence interval. One method for fixing this problem is to use the truncation of the estimator $\hat{\theta}_2$ as $\hat{\theta}_2^{TR} = \max\{\hat{\theta}_2, N^{-2/3}\}$, which was suggested in Kubokawa (2009a), For the practical use of $I^{CEB}(\widehat{\boldsymbol{\theta}})$, we need such a modification of the estimator $\widehat{\boldsymbol{\theta}}$.

5 Testing and variable selection

In this final section, we want to address the problem of selecting significant explanatory variables. To this end, we explain the two approaches: testing hypothesis and information criteria like model selection.

5.1 Testing procedures for a linear hypothesis on regression coefficients

Consider the general linear regression model described in (2.2) without assuming the structure (2.3), namely, $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$. The hypothesis ot be tested is the linear restriction given by

$$H_0 \ : \ \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r},$$

where $\boldsymbol{R}$ is an $r \times p$ known matrix with rank r, $r \leq p$, and $\boldsymbol{r}$ is an $r \times 1$ vector. For given $\boldsymbol{\theta}$, the unrestricted and restricted estimators of $\boldsymbol{\beta}$ are given by

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} =& \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{y}, \\
\widetilde{\boldsymbol{\beta}} =& \widetilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - (\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{W}(\boldsymbol{\theta})(\boldsymbol{R}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{r}),
\end{aligned}$$

for $\boldsymbol{W}(\boldsymbol{\theta}) = [\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}$. Using these notations, we describe the unrestriced and restricted estimators of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ as $(\widehat{\boldsymbol{\beta}}_u, \widehat{\boldsymbol{\theta}})$ and $(\widetilde{\boldsymbol{\beta}}_R, \widetilde{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\beta}}_u = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$ and $\widetilde{\boldsymbol{\beta}}_R = \widetilde{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\theta}})$. We also use the notations $\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\theta}})$ and $\widetilde{\boldsymbol{\beta}}_u = \widetilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$.

[1] The Wald, likelihood ratio and Lagrange multiplier test statistics. As the general methods for testing hypotheses, the three procedures are known which are based on the Wald, likelihood ratio and Lagrange multiplier test statistics. Consider the general framework of testing $H_0 : \boldsymbol{a}(\boldsymbol{\xi}) = \boldsymbol{0}$ against $H_1 : \boldsymbol{a}(\boldsymbol{\xi}) \neq \boldsymbol{0}$, where a random variable $\boldsymbol{X}$ has a likelihood function $L(\boldsymbol{\xi}|\boldsymbol{X})$, $\boldsymbol{\xi}$ is a p-dimensional unknown vector and $\boldsymbol{a}(\boldsymbol{\xi})$ is a function from $\boldsymbol{R}^p$ to $\boldsymbol{R}^q$ for $q \leq p$. Then, the Wald, likelihood ratio and Lagrange multiplier test statistics are given by

$$\begin{aligned}
F_W =& \boldsymbol{a}(\widehat{\boldsymbol{\xi}})'[\boldsymbol{A}(\widehat{\boldsymbol{\xi}})\boldsymbol{I}(\widehat{\boldsymbol{\xi}})\boldsymbol{A}(\widehat{\boldsymbol{\xi}})']^{-1}\boldsymbol{a}(\widehat{\boldsymbol{\xi}}), \\
F_{LR} =& -2\{\log L(\widetilde{\boldsymbol{\xi}}|\boldsymbol{X}) - \log L(\widehat{\boldsymbol{\xi}}|\boldsymbol{X})\}, \\
F_{LM} =& \boldsymbol{s}(\widetilde{\boldsymbol{\xi}})'\boldsymbol{I}(\widetilde{\boldsymbol{\xi}})^{-1}\boldsymbol{s}(\widetilde{\boldsymbol{\xi}}),
\end{aligned}$$

where $\boldsymbol{A}(\boldsymbol{\xi}) = \partial \boldsymbol{a}(\boldsymbol{\xi})/\partial \boldsymbol{\xi}'$, $\boldsymbol{I}(\boldsymbol{\xi}) = E[\boldsymbol{s}(\boldsymbol{\xi})\boldsymbol{s}(\boldsymbol{\xi})']$ is the Fisher information matrix, $\boldsymbol{s}(\boldsymbol{\xi}) = \partial \log L(\boldsymbol{\xi}|\boldsymbol{X})/\partial \boldsymbol{\xi}$ is the score function, and $\widehat{\boldsymbol{\xi}}$ and $\widetilde{\boldsymbol{\xi}}$ are unrestricted and restricted estimator of $\boldsymbol{\xi}$. The Lagrange multiplier statistic is also called the score test statistic or the Rao statistic. These test statistics converge to the chi-square distribution with q degrees of freedom under H_0.

For testing the hypothesis $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$ in the general linear regression model (2.2), these test statistics are written as

$$\begin{aligned}
F_W =& (\boldsymbol{R}\widehat{\boldsymbol{\beta}}_u - \boldsymbol{r})'\boldsymbol{W}(\widehat{\boldsymbol{\theta}})(\boldsymbol{R}\widehat{\boldsymbol{\beta}}_u - \boldsymbol{r}), \\
F_{LR} =& -2[\ell(\widetilde{\boldsymbol{\beta}}_R, \widetilde{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\beta}}_u, \widehat{\boldsymbol{\theta}})], \\
F_{LM} =& (\boldsymbol{R}\widetilde{\boldsymbol{\beta}}_R - \boldsymbol{r})'\boldsymbol{W}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{R}\widetilde{\boldsymbol{\beta}}_R - \boldsymbol{r}),
\end{aligned} \tag{5.1}$$

where $\boldsymbol{W}(\boldsymbol{\theta}) = [\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}$ and $-2\ell(\boldsymbol{\beta},\boldsymbol{\theta}) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$. It is known that between these test statistics, there exist the inequalities $F_W \geq F_{LR} \geq F_{LM}$.

[2] Bartlett-type corrections in LMM. The Bartlett-type corrections of the test statistcs given in (5.1) were derived by Rothenberg (1984) under the null and local alternative hypotheses. Let

$$\begin{aligned}\boldsymbol{C}_i =& (\boldsymbol{\Sigma}^{-1})_{(i)}\boldsymbol{H}(\boldsymbol{\theta}),\\ \boldsymbol{D}_{ij} =& \boldsymbol{H}(\boldsymbol{\theta})\big\{(\boldsymbol{\Sigma}^{-1})_{(ij)} - (\boldsymbol{\Sigma}^{-1})_{(i)}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1})_{(j)}\\ & - (\boldsymbol{\Sigma}^{-1})_{(j)}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{\Sigma}^{-1})_{(i)}\big\},\end{aligned}$$

for $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}'\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'$. Then, we define functions $b(\boldsymbol{\theta})$, $c(\boldsymbol{\theta})$ and $d(\boldsymbol{\theta})$ by

$$\begin{aligned}b(\boldsymbol{\theta}) =& \frac{1}{2}\mathrm{tr}\left[\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)\mathbf{mat}_{ij}\big(\mathrm{tr}\,[\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{\Sigma}_{(j)}\boldsymbol{\Sigma}^{-1}]\big)\right],\\ c(\boldsymbol{\theta}) =& \frac{1}{2}\mathrm{tr}\left[\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)\mathbf{mat}_{ij}\big(\mathrm{tr}\,[\boldsymbol{C}_i\boldsymbol{C}_j] + \frac{1}{2}\mathrm{tr}\,[\boldsymbol{C}_i]\mathrm{tr}\,[\boldsymbol{C}_j]\big)\right],\\ d(\boldsymbol{\theta}) =& E[(\widehat{\boldsymbol{\theta}}^* + \widehat{\boldsymbol{\theta}}^{**})'\mathbf{col}_i(\mathrm{tr}\,[\boldsymbol{C}_i])] + \mathrm{tr}\left[\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)\mathbf{mat}_{ij}\big(\mathrm{tr}\,[\boldsymbol{C}_i]\mathrm{tr}\,[\boldsymbol{C}_j]\big)\right]\\ & + \frac{1}{2}\mathrm{tr}\left[\mathbf{Cov}\,(\widehat{\boldsymbol{\theta}}^*)\mathbf{mat}_{ij}\big(\mathrm{tr}\,[\boldsymbol{D}_{ij}]\big)\right],\end{aligned}$$

for $\boldsymbol{\Pi}(\boldsymbol{\theta})$ defined in (3.1). Under appropriate conditions like (C1)-(C6), the Bartlett-type correction of the Wald test statistic F_W is given by

$$F_W^* = F_W/\big[1 + (\hat{d} - \hat{c} + \hat{b})/q + \hat{c}z_\alpha/\{q(q+2)\}\big],$$

where $\hat{b} = b(\widehat{\boldsymbol{\theta}})$, $\hat{c} = c(\widehat{\boldsymbol{\theta}})$, $\hat{d} = d(\widehat{\boldsymbol{\theta}})$, and z_α is the $100\alpha\%$ upper point of the χ^2_q-distribution. Rothenberg (1984) showed that F_W^* satisfies the nominal significance level up to $o(N^{-1})$, namely, $P[F_W^* \geq z_\alpha] = \alpha + o(N^{-1})$ under H_0. When $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ are the unrestricted and restricted ML estimators of $\boldsymbol{\theta}$, F_{LR} and F_{LM} are approximated as

$$\begin{aligned}F_{LR} =& F_W - \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'\boldsymbol{A}_2(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) + o_p(N^{-1}),\\ F_{LM} =& F_W - (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'\boldsymbol{A}_2(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) + o_p(N^{-1}),\end{aligned}$$

for $\boldsymbol{A}_2 = \mathbf{mat}_{ij}(\mathrm{tr}\,[\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_{(j)}\boldsymbol{\Sigma}])$. The Bartlett-type corrections for F_{LR} and F_{LM} are given by

$$\begin{aligned}F_{LR}^* =& F_{LR}/\big[1 + (\hat{d} - \hat{c})/q\big],\\ F_{LM}^* =& F_{LM}/\big[1 + (\hat{d} - \hat{c} - \hat{b})/q - \hat{c}z_\alpha/\{q(q+2)\}\big],\end{aligned}$$

which can be derived by evaluating the term $(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'\boldsymbol{A}_2(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})$. For the details of the derivations, see Rothenberg (1984).

5.2 Information criteria for variable or model selection

Related to testing the hypothesis on the regression coefficients, the variable selection procedures are useful for choosing significant explanatory variables affecting the response variables. Of these, we here treat the *Akaike Information Criterion* (AIC), the *conditional Akaike Information Criterion* (cAIC), the *Bayesian Information Criterion* (BIC) and the *Empirical Bayes Information Criterion* (EBIC). For a good account of AIC, BIC and other criteria, see Konishi and Kitagawa (2007),

For stating the concepts of these criteria, let $f(\boldsymbol{y}|\boldsymbol{v},\boldsymbol{\beta},\boldsymbol{\theta})$ and $f(\boldsymbol{v}|\boldsymbol{\theta})$ be the conditional density of $\boldsymbol{y}$ given $\boldsymbol{v}$ and the marginal density of $\boldsymbol{v}$, respectively, where $\boldsymbol{y}|\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{v}, \boldsymbol{R}(\boldsymbol{\theta}))$ and $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\theta}))$. Then, the marginal density of $\boldsymbol{y}$ is written by $f_m(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}) = \int f(\boldsymbol{y}|\boldsymbol{v},\boldsymbol{\beta},\boldsymbol{\theta})f(\boldsymbol{v}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{v}$, which has marginal distribution $\mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$.

[1] AIC and cAIC. The AIC proposed by Akaike (1973, 1974) is based on the thought of choosing a model which minimizes an unbiased estimator of the expected Kullback-Leibler information. The expected Kullback-Leibler information is defined by

$$R(\boldsymbol{\beta},\boldsymbol{\theta};\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{y}}\Big[\int\Big(\log\frac{f_m(\boldsymbol{y}^*|\boldsymbol{\beta},\boldsymbol{\theta})}{f_m(\boldsymbol{y}^*|\widehat{\boldsymbol{\beta}}(\boldsymbol{y}),\widehat{\boldsymbol{\theta}}(\boldsymbol{y}))}\Big)f_m(\boldsymbol{y}^*|\boldsymbol{\beta},\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}^*\Big],$$

which can be interpeted as a risk function for estimating $(\boldsymbol{\beta},\boldsymbol{\theta})$ by $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}})$ relative to the Kullback-Leibler distance. This quantity measures the prediction error in predicting future variable $\boldsymbol{y}^*$ based on the model $f_m(\boldsymbol{y}^*|\widehat{\boldsymbol{\beta}}(\boldsymbol{y}),\widehat{\boldsymbol{\theta}}(\boldsymbol{y}))$. In this sense, AIC is a criterion of finding a model which can provide a good prediction in light of minimizing the prediction error. $R(\boldsymbol{\beta},\boldsymbol{\theta};\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}})$ is rewritten as

$$\begin{aligned}&\iint\{\log f_m(\boldsymbol{y}^*|\boldsymbol{\beta},\boldsymbol{\theta})\}f_m(\boldsymbol{y}^*|\boldsymbol{\beta},\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}^*f_m(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}\\&\quad-\iint\{\log f_m(\boldsymbol{y}^*|\widehat{\boldsymbol{\beta}}(\boldsymbol{y}),\widehat{\boldsymbol{\theta}}(\boldsymbol{y}))\}f_m(\boldsymbol{y}^*|\boldsymbol{\beta},\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}^*f_m(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta})\mathrm{d}\boldsymbol{y},\end{aligned}$$

Since the first term is irrelevant to the model $f_m(\boldsymbol{y}^*|\widehat{\boldsymbol{\beta}}(\boldsymbol{y}),\widehat{\boldsymbol{\theta}}(\boldsymbol{y}))$, it is sufficient to estimate the second term. Thus, the *Akaike Information* (AI) is defined by

$$AI = -2\iint\{\log f_m(\boldsymbol{y}^*|\widehat{\boldsymbol{\beta}}(\boldsymbol{y}),\widehat{\boldsymbol{\theta}}(\boldsymbol{y}))\}f_m(\boldsymbol{y}^*|\boldsymbol{\beta},\boldsymbol{\theta})f_m(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta})\mathrm{d}\boldsymbol{y}^*\mathrm{d}\boldsymbol{y},$$

and AIC is derived as an asymptotically unbiased estimator of AI, namely, $E[AIC] = AI + o(1)$. When AIC is an exact unbiased estimator of AI, it is called the exact AIC, which was suggested by Sugiura (1978), but in general, it is difficult to get the exact AIC in LMM. When $\boldsymbol{\beta}$ is estimated by the GLS $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$ for a consistent estimator of $\boldsymbol{\theta}$, AIC is given as

$$AIC_c = -2\log f_m(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}}) + 2(p+q), \tag{5.2}$$

where $-2\log f_m(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}}) = N\log(2\pi) + \log|\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})| + \boldsymbol{y}'\boldsymbol{\Pi}(\widehat{\boldsymbol{\theta}})\boldsymbol{y}$ for $\boldsymbol{\Pi}(\boldsymbol{\theta})$ defined in (3.1), and p and q are dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively.

It is noted that the AIC stated above is based on the marginal distribution of $\boldsymbol{y}$, namely, it measures the prediction error of the predictor based on the marginal distribution $\mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$. This means that the marginal AIC is not appropriate for the focus on the prediciton of spcific areas or random effects as explained in the context of the small area estimation. Taking this point into account, Vaida and Blanchard (2005) proposed the conditional AIC as an asymptotically unbiased estimator of AI, where AI is the conditional Akaike information defined by

(5.3)
$$cAI = -2\int\int\int \log\{f(\boldsymbol{y}^*|\widehat{\boldsymbol{v}}(\boldsymbol{y}), \widehat{\boldsymbol{\beta}}(\boldsymbol{y}), \widehat{\boldsymbol{\theta}}(\boldsymbol{y}))\} f(\boldsymbol{y}^*|\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{v}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{y}^* \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{v},$$

where $\widehat{\boldsymbol{v}}(\boldsymbol{y}) = \widehat{\boldsymbol{v}}$ is the empirical Bayes estimator of $\boldsymbol{v}$ given in (2.8). When $\boldsymbol{\theta}$ is known, Vaida and Blanchard (2005) derived an exact unbiased estimator of cAI, and it gives the same value as DIC, the deviance information criterion proposed by Spiegelhalter, Best, Carlin and van der Linde (2002) for Bayesian inference. Although an exact unbiased estimator of cAI is hard to get in LMM, we can derive an asymptotically unbiased estimator of cAI, given by

(5.4)
$$cAIC_c = -2\log f(\boldsymbol{y}|\widehat{\boldsymbol{v}}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}) - \Delta_c,$$

where $-2\log f(\boldsymbol{y}|\widehat{\boldsymbol{v}}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}) = N\log(2\pi) + \log|\widehat{\boldsymbol{R}}| + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{G}}\boldsymbol{Z}')\widehat{\boldsymbol{R}}^{-1}(\boldsymbol{I} - \boldsymbol{Z}\widehat{\boldsymbol{G}}\boldsymbol{Z}'\widehat{\boldsymbol{\Sigma}}^{-1})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ for $\widehat{\boldsymbol{R}} = \boldsymbol{R}(\widehat{\boldsymbol{\theta}})$, $\widehat{\boldsymbol{G}} = \boldsymbol{G}(\widehat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$.

[2] BIC and EBIC. The Bayesian information criterion (BIC) proposed by Schwarz (1978) assumes a proper prior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\theta})$ formally and evaluate asymptotically the marginal distribution

$$f_\pi(\boldsymbol{y}) = \int\int f_m(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}\mathrm{d}\boldsymbol{\theta}.$$

The Laplace approximation can be used to get the approximation as $-2\log\{f_\pi(\boldsymbol{y})\} = BIC + o_p(\log(N))$, where

(5.5)
$$BIC = -2\log\{f_m(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}})\} + (p+q)\log(N),$$

where $-2\log\{f_m(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}})\}$ is given below (5.2). The distinction between AIC and BIC appears in the penalty terms as seen from (5.2) and (5.5).

The Bayesian criteria like Bayes factors use all the prior information on $(\boldsymbol{\beta}, \boldsymbol{\theta})$, while all the prior information is neglected in BIC, because the prior information comes into neglected terms asymptotically. Thus, we can consider the intermediate case, that is, the parameter are decomposed into two parts of interest and nuisance, and we want to use only the prior information on the interest parameters. For example, we consider the case that $\boldsymbol{\beta}$ is the parameters of interest and $\boldsymbol{\theta}$ is the nuisance parameters in LMM. Assume that $(\boldsymbol{\beta}, \boldsymbol{\theta})$ has the prior distirbution

$$(\boldsymbol{\beta}, \boldsymbol{\theta}) \sim \pi_1(\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{\lambda})\pi_2(\boldsymbol{\theta}),$$

where given $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ conditionally has $\pi_1(\boldsymbol{\beta}|\boldsymbol{\theta},\boldsymbol{\lambda})$ with hyperparameter $\boldsymbol{\lambda}$. Let $f_{\pi,1}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\lambda})$ be the conditional marginal density given by

$$f_{\pi,1}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\lambda}) = \int f_m(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta})\pi_1(\boldsymbol{\beta}|\boldsymbol{\theta},\boldsymbol{\lambda})\mathrm{d}\boldsymbol{\beta},$$

and $\widehat{\boldsymbol{\lambda}}$ is the ML estimator of $\boldsymbol{\lambda}$ based on this distribution, namely,

$$\widehat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}}\{f_{\pi,1}(\boldsymbol{y}|\widehat{\boldsymbol{\theta}},\boldsymbol{\lambda})\}.$$

Then, Kubokawa and Srivastava (2009) proposed the empirical Bayes information criterion (EBIC) as

$$EBIC = -2\log\{f_{\pi,1}(\boldsymbol{y}|\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\lambda}})\} + q\log(N). \tag{5.6}$$

Concerning the prior distribution of $\boldsymbol{\beta}$, the common prior used in the ordinary linear regression model is

$$\pi_1(\boldsymbol{\beta}|\lambda) = \mathcal{N}_p(\boldsymbol{0}, \lambda^{-1}\boldsymbol{W})$$

for an unknown scalar λ and a $p\times p$ known matrix $\boldsymbol{W}$. The prior with $\boldsymbol{W} = N(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is called Zellner's g-prior, and other choices of $\boldsymbol{W}$ are $\boldsymbol{W} = \mathrm{diag}\,(N/\boldsymbol{x}'_{(1)}\boldsymbol{x}_{(1)},\ldots,N/\boldsymbol{x}'_{(p)}\boldsymbol{x}_{(p)})$ where $\boldsymbol{X} = (\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(p)})$ and $\boldsymbol{W} = \boldsymbol{I}_p$. Then, the marginal density $f_{\pi,1}(\boldsymbol{y}|\boldsymbol{\theta},\lambda)$ is expressed as

$$f_{\pi,1}(\boldsymbol{y}|\boldsymbol{\theta},\lambda) = \frac{1}{(2\pi)^{N/2}}\frac{1}{|\boldsymbol{\Sigma}(\boldsymbol{\theta}) + \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'/\lambda|^{1/2}}\exp\Big\{-\frac{1}{2}\boldsymbol{y}'\boldsymbol{\Pi}^*(\boldsymbol{\theta},\lambda)\boldsymbol{y}\Big\},$$

where

$$\boldsymbol{\Pi}^*(\boldsymbol{\psi},\lambda) = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\{\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{X} + \lambda\boldsymbol{W}^{-1}\}^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}.$$

The hyper-parameter λ is estimated by $\hat{\lambda}$ through the maximization of $f_{\pi,1}(\boldsymbol{y}|\widehat{\boldsymbol{\theta}},\lambda)$ with respect to λ, namely, it is given by $\hat{\lambda} = \max(\lambda_0, 0)$ where λ_0 is the solution of the equation

$$\begin{aligned}&\boldsymbol{y}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X} + \lambda_0\boldsymbol{W}^{-1})^{-1}\boldsymbol{W}^{-1}(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X} + \lambda_0\boldsymbol{W}^{-1})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}\\ &\quad = \mathrm{tr}\,[(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X} + \lambda_0\boldsymbol{W}^{-1})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}]/\lambda_0.\end{aligned}$$

Then the EBIC is given by

$$EBIC = N\log(2\pi) + \log(|\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}) + \hat{\lambda}^{-1}\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'|) + \boldsymbol{y}'\boldsymbol{\Pi}^*(\widehat{\boldsymbol{\theta}},\hat{\lambda})\boldsymbol{y} + q\log(N). \tag{5.7}$$

Finally, we should note that AIC and BIC are derived through different thoughts, which results in a different asymptotic properties, namely, BIC has consistency for selecting the true model, while AIC is not consistent. In general, BIC, EBIC and Bayesian procedures based on proper priors are consistent. However, AIC and cAIC choose models which give smaller prediction errors, while those Bayesian procedures do not guarantee such a property.

Acknowledgments. The research of the author was supported in part by a grant from the Ministry of Education, Japan, Nos. 19200020 and 21540114.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (B.N. Petrov and Csaki, F, eds.), 267-281, Akademia Kiado, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. System identification and time-series analysis. *IEEE Trans. Autom. Contr.*, **AC-19**, 716-723.

[3] Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, New York.

[4] Basu, R., Ghosh, J.K., and Mukerjee, R. (2003). Empirical Bayes prediction intervals in a normal regression model: higher order asymptotics. *Statist. Prob. Letters*, **63**, 197-203.

[5] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.

[6] Chatterjee, S., Lahiri, P., and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Ann. Statist.*, **36**, 1221-1245.

[7] Chen, S. and Lahiri, P. (2008). On mean squared prediction error estimation in small area estimation problems. *Commun. Statist.-Theory Methods*, **37**, 1792-1798.

[8] Cressie, N. and Lahiri, S.N. (1993). The asymptotic distribution of REML estimators. it J. Multivariate Analysis, **45**. 217-233.

[9] Das, K., Jiang, J. and Rao, J.N.K. (2004). Mean squared error of empirical predictor. *Ann. Statist.*, **32**, 818-840.

[10] Datta, G.S., Ghosh, M., Smith, D.D. and Lahiri, P. (2002). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence Intervals. *Scandinavian J. Statist.*, **29**, 139-152.

[11] Datta, G.S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica*, **10**, 613-627.

[12] Datta, G.S., Rao, J.N.K. and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92**, 183-196.

[13] Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Wiley.

[14] Diggle, P., Liang, K.-Y., and Zeger, S.L. (1994). *Longitudinal Data Analysis*. Oxford Univ. Press.

[15] Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, **70**, 311-319.

[16] Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2nd ed. Springer, New York.

[17] Fay, R.E. and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

[18] Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2004). *Applied Longitudinal Analysis*. Wiley.

[19] Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statist. Science*, **9**, 55-93.

[20] Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Ann. Statist.*, **34**, 1733-1750.

[21] Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *J. Royal Statist. Soc.*, **68**, 221-238.

[22] Harville, D.A. and Jeske, D.R. (1992). Mean squared error of estimation or prediction under a general linear model. *J. Amer. Statist. Assoc.*, **87**, 724-731.

[23] Henderson, C.R. (1950). Estimation of genetic parameters. *Ann. Math. Statist.*, **21**, 309-310.

[24] Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge University Press.

[25] Kackar, R.N. and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.*, **79**, 853-862.

[26] Konishi, S. and Kitagawa, G. (2007). *Information Criteria and Statistical Modeling.* Springer.

[27] Kubokawa, T. (2009a). Corrected empirical Bayes confidence intervals in nested error regression models. Discussion Paper Series, CIRJE-F-632. Journal of the Korean Statistical Society, to appear.

[28] Kubokawa, T. (2009b) Higher order corrections in MSE estimation and confidence intervals in linear mixed models. Discussion Paper Series, CIRJE-F-666.

[29] Kubokawa, T. and Srivastava, M.S. (2009) Consistency of the empirical Bayes information criterion in linear mixed models. Discussion Paper Series, CIRJE-F-614.

[30] Lahiri, P. and Rao, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *J. Amer. Statist. Assoc.*, **90**, 758-766.

[31] Lawson, A.B. (2006). *Statistical Methods in Spacial Epidemiology*. 2nd ed. Wiley, England.

[32] Lawson, A.B., Browne, W.J. and Vidal Rodeiro, C.L. (2003). *Disease Mapping with WinBUGS and MLwiN*. Wiley, England.

[33] Mardia K.V. and Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135-146.

[34] McCulloch, C.E. (2003). *Generalized Linear Mixed Models*. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 7. IMS, USA.

[35] McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. Wiley, New York.

[36] Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. Springer.

[37] Pfeffermann, D. (2002). Small area estimation - new developments and directions. *Int. Statist. Rev.*, **70**, 125-143.

[38] Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.

[39] Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, **25**, 175-186.

[40] Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, New Jersey.

[41] Rothenberg, T. (1984). Hypothesis testing in linear models when the error covariance matrix is nonscalar. **52**, 827-842.

[42] Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components*, Wiley, New York.

[43] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

[44] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *J. Royal Statist. Soc.*, **B 64**, 583-639.

[45] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135-1151.

[46] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. - Theory Methods*, **1**, 13-26.

[47] Sweeting, T.J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.*, **8**, 1375-1381.

[48] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Probab.*, **1**, 197-206. University of California University, Berkeley.

[49] Vaida, F., and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351-370.

[50] Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.