CIRJE-F-702

A Review of Linear Mixed Models and Small Area Estimation

Tatsuya Kubokawa
University of Tokyo
December 2009

Discussion Papers are a series of manuscripts in their draft form. They are not intended for circulation or distribution except as indicated by the author. For that reason Discussion Papers may not be reproduced or distributed without the written consent of the author.
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 parameters 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, Bayesian 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 under 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, desirable 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 constructed by pooling all the data. These two features 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.

2 Linear mixed models and BLUP

2.1 Linear mixed models

Consider the general linear mixed model

\[ y = X\beta + Zv + \epsilon, \]

where \( y \) is an \( N \times 1 \) observation vector of the response variable, \( X \) and \( Z \) are \( N \times p \) and \( N \times M \) matrices, respectively, of the explanatory variables, \( \beta \) is a \( p \times 1 \) unknown vector of the regression coefficients, \( v \) is an \( M \times 1 \) vector of the random effects, and \( \epsilon \) is an \( N \times 1 \) vector of the random errors. Here, \( v \) and \( \epsilon \) are mutually independently distributed as \( v \sim N_M(0, G(\theta)) \) and \( \epsilon \sim N_N(0, R(\theta)) \), where \( \theta = (\theta_1, \ldots, \theta_q)' \) is a \( q \) dimensional vector of unknown parameters, and \( G = G(\theta) \) and \( R = R(\theta) \) are positive definite matrices.

Throughout the paper, for simplicity, it is assumed that \( X \) is of full rank. Then, \( y \) has a marginal distribution

\[ \mathcal{N}(X\beta, \Sigma(\theta)) \]

for

\[ \Sigma = \Sigma(\theta) = R(\theta) + ZG(\theta)Z'. \]

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

\[ y_{ij} = x_{ij}'\beta + v_i + \varepsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \]

where \( k \) is the number of small areas, \( N = \sum_{i=1}^k n_i \), \( x_{ij} \) is a \( p \times 1 \) vector of explanatory variables, \( \beta \) is a \( p \times 1 \) unknown common vector of regression coefficients, and \( v_i \)'s and \( \varepsilon_{ij} \)'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, \( \sigma_v^2 \) and \( \sigma^2 \) 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 \( X_i = (x_{i1}, \ldots, x_{in_i})' \), \( X = (X_1', \ldots, X_k')' \), \( y_i = (y_{i1}, \ldots, y_{in_i})' \), \( y = (y_1', \ldots, y_k')' \) and let \( \epsilon \) be similarly defined. Let \( v = (v_1, \ldots, v_k)' \) and \( Z = \text{block diag}(j_1, \ldots, j_k) \) for \( j_i = (1, \ldots, 1)' \in \mathbb{R}^{n_i} \). Then, the model is expressed in vector notations as \( y = X\beta + Zv + \epsilon \), 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 \( \bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i \) has large deviation for
predicting the mean crop hectare per segment $\mu_i = \bar{x}_i'\beta + v_i$ for $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$. 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:

\[(2.5) \quad y_i = x_i'\beta + v_i + \varepsilon_i, \quad i = 1, \ldots, k,\]

where $k$ is the number of small areas, $x_i$ is a $p \times 1$ vector of explanatory variables, $\beta$ is a $p \times 1$ unknown common vector of regression coefficients, and $v_i$'s and $\varepsilon_i$'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 $X = (x_1, \ldots, x_k)'$, $y = (y_1, \ldots, y_k)'$, and let $v$ and $\varepsilon$ be similarly defined. Then, the model is expressed in vector notations as

$$y = X\beta + v + \varepsilon,$$

and $y \sim \mathcal{N}(X\beta, \Sigma)$ where $\Sigma = \Sigma(\theta) = \theta I_k + D$ for $D = \text{diag}(d_1, \ldots, d_k)$ 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

\[(2.6) \quad y_{it} = x_{it}'\beta + v_{it} + \varepsilon_{it}, \quad i = 1, \ldots, k, \quad t = 1, \ldots, T,
\]

where $k$ is the number of small areas, $t$ is a time index, $N = kT$, $x_{it}$ is a $p \times 1$ vector of explanatory variables, $\beta$ is a $p \times 1$ unknown common vector of regression coefficients, and $v_{it}$'s and $\varepsilon_{it}$'s are random errors. Let $X_i = (x_{i1}, \ldots, x_{i,n})'$, $y_i = (y_{i1}, \ldots, y_{i,n})'$, and let $v_i$ and $\varepsilon_i$ be similarly defined. Then, the model is expressed in vector notations as

$$y_i = X_i\beta + v_i + \varepsilon_i, \quad i = 1, \ldots, k.$$

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

$$\Psi(\rho) = (1 - \rho)I_T + \rho j_T j_T' \quad \text{and} \quad \Psi(\rho) = \text{mat}_{ij}(\rho^{||i-j||}).$$

Letting $X = (X_1', \ldots, X_k')'$, $y = (y_1', \ldots, y_k')'$ and letting $v$ and $\varepsilon$ be defined similarly, we can express the model as $y = X\beta + v + \varepsilon$.

### 2.2 Mixed model equation and BLUP

[1] **BLUP.** We now consider the estimation of the regression coefficients $\beta$ and the prediction of the random effects $v$ in (2.1). When the covariance matrices $G$ and $R$ are known, there exists the best unbiased predictor of $v$ among the linear functions of $y$. This is called the **Best Linear Unbiased Predictor (BLUP)** and denoted by $\hat{v}$. Also, there
exist the best linear unbiased estimator of \( \beta \), denoted by \( \hat{\beta} \). Henderson (1950) showed that \((\hat{\beta}, \hat{v})\) can be derived by the solution of the equation given by

\[
(2.7) \quad \begin{pmatrix} X' R^{-1} X & X' R^{-1} Z \\ Z' R^{-1} X & Z' R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} X' R^{-1} y \\ Z' R^{-1} y \end{pmatrix},
\]

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

\[
(2.8) \quad \hat{\beta} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} y, \quad \hat{v} = GZ' \Sigma^{-1} (y - X \hat{\beta}),
\]

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

\[
(2.9) \quad \hat{\mu}^{EB} = a' \hat{\beta} + b' GZ' \Sigma^{-1} (y - X \hat{\beta}),
\]

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

We next substitute \( \hat{v} = GZ' \Sigma^{-1} (y - X \hat{\beta}) \) into the first equation of (2.7) given by

\[
X' R^{-1} X \hat{\beta} + X' R^{-1} Z \hat{v} = X' R^{-1} y.
\]

Then,

\[
X' R^{-1} X \hat{\beta} + X' R^{-1} Z GZ' \Sigma^{-1} (y - X \hat{\beta}) = X' R^{-1} y,
\]

which yields

\[
X' R^{-1} (\Sigma - ZGZ') \Sigma^{-1} X \hat{\beta} = X' R^{-1} (\Sigma - ZGZ') \Sigma^{-1} y.
\]

It is noted that \( \Sigma = ZGZ' + R \), namely, \( R^{-1} (\Sigma - ZGZ') = I \). Thus, we get the equation

\[
X' \Sigma^{-1} X \hat{\beta} = X' \Sigma^{-1} y,
\]

which means that the solution \( \hat{\beta} \) is described as the form in (2.8).
Hence, it is seen that (2.7) is a matricial expression of (2.12). The equation (2.7) can be derived. Two of typical approaches to the derivation are the maximum likelihood (ML) method and the empirical Bayes method.

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 \( \mu_i = \mathbf{x}_i^T \mathbf{\beta} + v_i \) for \( i = 1, \ldots, k \). Let \( \theta_1 = \sigma^2 \), \( \theta_2 = \sigma_v^2 \), and let \( \mathbf{\theta} = (\theta_1, \theta_2)' \). In this mode, \( \mathbf{G}(\mathbf{\theta}) = \theta_2 \mathbf{I}_k \), \( \Sigma_i(\mathbf{\theta}) = \theta_1 \mathbf{I}_n + \theta_2 \mathbf{j}_i \mathbf{j}_i' \) and \( \Sigma(\mathbf{\theta}) = \text{block diag}(\Sigma_1(\mathbf{\theta}), \ldots, \Sigma_k(\mathbf{\theta})) \). Noting that

\[
\Sigma_i^{-1} = \frac{1}{\theta_1} \left( \mathbf{I}_n - \frac{\theta_2}{\theta_1 + n_i \theta_2} \mathbf{j}_i \mathbf{j}_i' \right),
\]

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

\[
\hat{\mu}_i^{EB}(\mathbf{\theta}) = \mathbf{x}_i \hat{\mathbf{\beta}}(\mathbf{\theta}) + \frac{n_i \theta_2}{\theta_1 + n_i \theta_2} \left( \bar{y}_i - \mathbf{x}_i \hat{\mathbf{\beta}}(\mathbf{\theta}) \right)
\]

where \( \bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \), and the GLS of \( \mathbf{\beta} \) is

\[
\hat{\mathbf{\beta}}(\mathbf{\theta}) = \left\{ \sum_{i=1}^{k} \left( \mathbf{x}_i \mathbf{x}_i' - \frac{n_i \theta_2}{\theta_1 + n_i \theta_2} \mathbf{x}_i \mathbf{x}_i' \right) \right\}^{-1} \sum_{i=1}^{k} \left( \mathbf{x}_i \mathbf{y}_i' - \frac{n_i \theta_2}{\theta_1 + n_i \theta_2} \mathbf{x}_i \bar{y}_i \right).
\]

[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 function of \( (\mathbf{y}, \mathbf{v}) \) is written as

\[
(2\pi)^{-N/2} |G|^{-1/2} |R|^{-1/2} \exp \{-h(\mathbf{\beta}, \mathbf{v})/2\},
\]

where

\[
h(\mathbf{\beta}, \mathbf{v}) = \mathbf{v}' G^{-1} \mathbf{v} + (\mathbf{y} - \mathbf{X} \mathbf{\beta} - \mathbf{Z} \mathbf{v})' R^{-1} (\mathbf{y} - \mathbf{X} \mathbf{\beta} - \mathbf{Z} \mathbf{v}).
\]

To minimize \( h(\mathbf{\beta}, \mathbf{v}) \) with respect to \( (\mathbf{\beta}, \mathbf{v}) \), we need to differentiate it with respect to \( \mathbf{\beta} \) and \( \mathbf{v} \), which yields that

\[
\frac{\partial h(\mathbf{\beta}, \mathbf{v})}{\partial \mathbf{\beta}} = -2 \mathbf{X}' R^{-1} (\mathbf{y} - \mathbf{X} \mathbf{\beta} - \mathbf{Z} \mathbf{v}),
\]

\[
\frac{\partial h(\mathbf{\beta}, \mathbf{v})}{\partial \mathbf{v}} = 2 \mathbf{G}^{-1} \mathbf{v} - 2 \mathbf{Z}' R^{-1} (\mathbf{y} - \mathbf{X} \mathbf{\beta} - \mathbf{Z} \mathbf{v}).
\]

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

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

\[
\text{Cov} (\mathbf{y}, \mathbf{v}) = \begin{pmatrix} \Sigma & ZG \\ GZ' & G \end{pmatrix},
\]

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

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

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 \( \mathbf{y} \) is given by

\[
\mathbf{y} \sim \mathcal{N}_N (\mathbf{X} \mathbf{\beta}, \Sigma),
\]

whose density function is described as

\[
(2\pi)^{-N/2} |\Sigma|^{-1/2} \exp \{- (\mathbf{y} - \mathbf{X} \mathbf{\beta})' \Sigma^{-1} (\mathbf{y} - \mathbf{X} \mathbf{\beta})/2\}.
\]
\(X\beta'/\Sigma^{-1}(y - X\beta)/2\). Thus, the ML estimator of \(\beta\) based on this marginal density function is identical to the GLS estimator \(\hat{\beta}\). Since the Bayes estimator is the mean of the posterior distribution, the expectation of the posterior distribution, given by

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

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

The distinction of the two methods described above is that the ML method estimates \(v\) by the mode of the posterior distribution, while the empirical Bayes method estimates \(v\) by the mean of the posterior distribution. Although both methods give 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 unobservable variable \(v\) if \(v\) has a correlation with \(y\), 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 \(G\) and \(R\) are, in general, functions of unknown parameters like variance components. The unknown parameters are here denoted by \(\theta = (\theta_1, \ldots, \theta_q)'\), namely, the covariance matrix of \(y\) is described as

\[
\Sigma = \Sigma(\theta) = R(\theta) + ZG(\theta)Z.
\]

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

\[
\Pi = \Pi(\theta) = \Sigma(\theta)^{-1} - \Sigma(\theta)^{-1}X\{X'\Sigma(\theta)^{-1}X\}^{-1}X'\Sigma(\theta)^{-1},
\]

and note that

\[
(y - X\hat{\beta}(\theta))'\Sigma(\theta)^{-1}(y - X\hat{\beta}(\theta)) = y'\Pi(\theta)y, \quad \Pi(\theta) = K(K'\Sigma(\theta)K)^{-1}K'.
\]
Also note that $\partial_i \log |\Sigma| = \text{tr}[\Sigma^{-1} \partial_i \Sigma]$, $\partial_i \Pi = -\Pi (\partial_i \Sigma) \Pi$, $\partial_i \log |K' \Sigma K| = \text{tr}[K' \Sigma K] = \text{tr}[\Pi \partial_i \Sigma]$ where $\partial_i$ denotes the differential operator $\partial_i = \partial / \partial \theta_i$. Thus, the ML and REML estimators are solutions of the following equations:

$$
\begin{align}
\text{(3.2) [ML]} & \quad y' \Pi(\theta) \{\partial_i \Sigma(\theta)\} \Pi(\theta) y = \text{tr} \left[ \Sigma(\theta)^{-1} \{\partial_i \Sigma(\theta)\} \right], \\
\text{(3.3) [REML]} & \quad y' \Pi(\theta) \{\partial_i \Sigma(\theta)\} \Pi(\theta) y = \text{tr} \left[ \Pi(\theta) \{\partial_i \Sigma(\theta)\} \right].
\end{align}
$$

Since the l.h.s. of the above equations can be expressed as $y' \Pi(\theta) \{\partial_i \Sigma(\theta)\} \Pi(\theta) y = (y - X \hat{\beta}(\theta))' \Sigma(\theta)^{-1} \{\partial_i \Sigma(\theta)\} \Sigma(\theta)^{-1} (y - X \hat{\beta}(\theta)) = -(y - X \hat{\beta}(\theta))' \{\partial_i \Sigma(\theta)^{-1}\} (y - X \hat{\beta}(\theta))$, 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

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

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

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

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

(C3) The $q \times q$ matrix $A_2 = \text{mat}_{ij}(\text{tr} [\Sigma_{(i)} \Sigma_{(j)}])$ is positive definite and $A_2/N$ 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 $\hat{\theta}^M - \theta = O_p(N^{-1/2})$.

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

$$
\hat{\theta}^M - \theta = \hat{\theta}^{M*} + \hat{\theta}^{M**} + O_p(N^{-3/2}),
$$

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

$$
\hat{\theta}^{M*} = A_2^{-1} a_1 = A_2^{-1} \text{col}(\text{tr} [(\Sigma^{-1}_{(i)}) (yy' - \Sigma)])],
\hat{\theta}^{M**} = - A_2^{-1} \left\{ a_0 - \frac{b_0}{2} + A_1 A_2^{-1} a_1 \right\},
$$
for \( a_1 = \text{col}_i(-\text{tr} [[\Sigma^{-1}(i)(yy' - \Sigma)]]), \ a_0 = \text{col}_i(\text{tr} [Q_i(yy')]), \ A_2 = \text{mat}_{ia}(-\text{tr} [\Sigma(a)(\Sigma^{-1}(i))]), \ A_1 = \text{mat}_{ia}(\text{tr} [[\Sigma^{-1}(i)(yy' - \Sigma))]), \ b_0 = \text{col}(\sum_{a,b} B_{ab} \theta_a \theta_b^M)). \) Here, \( Q_i = \Sigma^{-1}(i)P + P \Sigma(i)^{-1} - P \Sigma(i)P \) for \( P = \Sigma^{-1}(i)X(i)(X(i)(\Sigma^{-1}(i))^{-1}X(i)^T \Sigma^{-1}(i)), \) and \( B_{ab} = \text{tr} [\Sigma(ab)(\Sigma^{-1}(i)) \Sigma(i)) + \text{tr} [\Sigma(a)(\Sigma^{-1}(i))\Sigma(i)] + \text{tr} [\Sigma(b)(\Sigma^{-1}(i))\Sigma(i)]. \) 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 \( \hat{\theta}^M - \theta = \hat{\theta}^* + \hat{\theta}^* + O_p(N^{-2}). \) Using the equality \( E[\text{tr} [C_1(yy' - \Sigma)]\text{tr} [C_2(yy' - \Sigma)] = 2\text{tr} [C_1C_2\Sigma] \) under the distribution \( y \sim N(0, \Sigma) \) for matrices \( C_1 \) and \( C_2, \) we can observe that

\[
E[\hat{\theta}^M] = 0, \quad \text{Cov}(\hat{\theta}^M) = 2A_2^{-1},
\]

\[
E[\hat{\theta}^{M*}] = A_2^{-1}\text{col}_i(\text{tr} [[X(i)(\Sigma^{-1}(i))^{-1}X(i)(\Sigma^{-1}(i))X(i)])
\]

\[
+ A_2^{-1}\text{col}_i(\text{tr} [A_2^{-1}\text{mat}_{a,b}(\text{tr} [\Sigma(ab)(\Sigma^{-1}(i))])])
\]

It is noted that \( E[\hat{\theta}^{M*}] = A_2^{-1}\text{col}_i(\text{tr} [[X(i)(\Sigma^{-1}(i))^{-1}X(i)(\Sigma^{-1}(i))X(i)]) \) when \( \Sigma \) or \( G \) and \( R \) are matrices of linear functions of \( \theta. \)

For the REML estimator, \( \hat{\theta}^R - \theta \) can be asymptotically expanded as

\[
\hat{\theta}^R - \theta = \hat{\theta}^R - \theta + \hat{\theta}^{R*} + O_p(N^{-2}),
\]

where \( \hat{\theta}^{R*} = \hat{\theta}^M = A_2^{-1}a_1 \) and

\[
\hat{\theta}^{R*} = -A_2^{-1}\{a_0^* - b_0/2 + A_1^{-1}A_1a_1\},
\]

where \( a_0^* = \text{col}_i(\text{tr} [Q_i(yy' - \Sigma)]). \) Thus, \( E[\hat{\theta}^{R*}] = 0, \text{Cov}(\hat{\theta}^{R*}) = \text{Cov}(\hat{\theta}^{M*}) = 2A_2^{-1} \) and

\[
E[\hat{\theta}^{R*}] = A_2^{-1}\text{col}_i(\text{tr} [A_2^{-1}\text{mat}_{a,b}(\text{tr} [\Sigma(ab)(\Sigma^{-1}(i))])]),
\]

where \( E[\hat{\theta}^{R*}] = 0 \) when \( \Sigma \) are matrices of linear functions of \( \theta. \)

**Example 3.1 (NERM)** In the NERM, the parameters \( \theta = (\theta_1, \theta_2)' \) and \( \Sigma \) correspond to \( \theta_1 = \sigma^2, \theta_2 = \sigma_\nu^2 \) and \( \Sigma = \text{blockdiag}(\Sigma_1, \ldots, \Sigma_k) \) for \( \Sigma_i = \theta_i I_i + \theta_j j_i', \) \( I_i \) being the \( n_i \times n_i \) identity matrix. The ML estimators \( \hat{\theta}^M = (\hat{\theta}_1^M, \hat{\theta}_2^M)' \) of \( (\theta_1, \theta_2)' \) are given as the solutions of the equations \( L_1(\hat{\theta}^M) = 0 \) and \( L_2(\hat{\theta}^M) = 0 \) where

\[
L_1(\theta) = 1 \sum_{i=1}^k \|y_i - X_i\hat{\beta}(\theta) - \frac{n_i\theta_2}{\theta_1 + n_i\theta_2}(\bar{y}_i - X_i\hat{\beta}(\theta))j_i\|_2^2 - \sum_{i=1}^k \frac{n_i}{\theta_1} (1 - \frac{\theta_2}{\theta_1 + n_i\theta_2}),
\]

\[
L_2(\theta) = \sum_{i=1}^k \frac{n_i^2}{(\theta_1 + n_i\theta_2)^2} \{ \bar{y}_i - X_i\hat{\beta}(\theta) \}^2 - \sum_{i=1}^k \frac{n_i}{\theta_1 + n_i\theta_2},
\]
since $\Sigma_{(1)} = I$ and $\Sigma_{(2)} = \text{block diag}(j_1j_1', \ldots, j_kj_k')$. Note that $A_2$ and $a_1$ can be written as

$$A_2 = \text{mat}_{ij}(\text{tr} [\Sigma_{(i)} \Sigma^{-1} \Sigma_{(j)} \Sigma^{-1}])$$

$$= \left( (N-k)\theta_1^2 + \sum_i (\theta_1 + n_i\theta_2)^{-2} \sum_i n_i(\theta_1 + n_i\theta_2)^{-2} \right),$$

$$a_1 = \left( \sum_i \text{tr} [\Sigma_{(i)}^{-2}(y_i y_i' - \Sigma_{(i)})] \left/ \sum_i j_i \Sigma_{(i)}^{-1} \right) \right).$$

Since $\hat{\theta}^{M*} = A_2^{-1} a_1$, it is observed that $E[\hat{\theta}^{M*}] = 0$ and

$$\text{Cov}(\hat{\theta}^{M*}) = \frac{2\theta_1}{d(\psi)} \left( \sum_{i=1}^k n_i^2 \gamma_i^2 - \sum_{i=1}^k n_i \gamma_i \right),$$

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

$$E[\hat{\theta}^{R*}] = \frac{\theta_1}{d(\psi)} \left( -p \sum_{i=1}^k n_i^2 \gamma_i^2 + (\sum_{i=1}^k n_i \gamma_i)c(\psi) \right),$$

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

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

$$0 = L_1(\theta) + \text{tr} [(X'\Sigma^{-1}X)^{-1} X'\Sigma^{-2} X],$$

$$0 = L_2(\theta) + \text{tr} [(X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}\text{block diag}(j_1j_1', \ldots, j_kj_k')\Sigma^{-1}X].$$

Noting that $\hat{\theta}^{R*} = A_2^{-1} a_1 = \hat{\theta}^{M*}$, we can see that $E[\hat{\theta}^{R*}] = 0$, $\text{Cov}(\hat{\theta}^{R*}) = \text{Cov}(\hat{\theta}^{M*})$ and $E[\hat{\theta}^{R*}] = O(N^{-2})$ 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 $S = y' (I_N - X(X'X)^{-1}X') y$ and $S_1 = y' (E - EX(X'EX)^{-1}X'E)y$ where $E = \text{block diag}(E_1, \ldots, E_k)$ for $E_i = I_i - n_i^{-1} j_i j_i'$. Then, unbiased estimators of $\theta_1$ and $\theta_2$ 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 $N^* = N - \text{tr} \{X'X\}^{-1} \sum_{i=1}^k n_i j_i x_i x_i'$ as suggested by Prasad and Rao (1990). In this case, $\hat{\theta}_1^U - \theta_1 = \hat{\theta}_1^U$ for $i = 1, 2$, and it is easy to see that $E[\hat{\theta}_1^U] = 0$, $E[\hat{\theta}_2^U] = 0$ and

$$\text{Cov}(\hat{\theta}_1^U) = \frac{2\theta_1^2}{N-k} \left( \frac{1}{N-k} \{k^2 + (N-k) \sum_{i=1}^k (1 + n_i \theta_2/\theta_1)^2\}/N^2 \right) + O(N^{-2}).$$

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

10
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 $\hat{\theta}$ into BLUP given in (2.9), namely, EBLUP of $\mu = \alpha' \beta + b'v$ is given by

$$\hat{\mu}^{EB}(\hat{\theta}) = \alpha' \hat{\beta}(\hat{\theta}) + b' G(\hat{\theta}) Z'(\Sigma(\hat{\theta}))^{-1} \{ y - X \hat{\beta}(\hat{\theta}) \}. \quad (3.6)$$

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

$$\hat{\mu}_i^{EB}(\hat{\theta}) = x_i' \hat{\beta}(\hat{\theta}) + \frac{n_i \hat{\theta}_2}{\hat{\theta}_1 + n_i \hat{\theta}_2} \{ \bar{y}_i - x_i' \hat{\beta}(\hat{\theta}) \}, \quad (3.7)$$

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

Such desirable properties of EBLUP are characterized as the shrinkage function and the pooling effect, namely, $\hat{\mu}_i^{EB}(\hat{\theta})$ shrinks $\bar{y}_i$ towards $x_i' \hat{\beta}(\hat{\theta})$, which is constructed by pooling all the data. The two features 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 $\beta = 0$, the best estimator of $\mu_i$ is $\bar{y}_i$. When $v_i$ is a random effect, however, the covariance matrix of $(\bar{y}_i, v_i)$ is

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

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

[2] Pooling data via common parameters. The regression coefficients $\beta$ is embedded as a common parameter in all the small areas. 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 $x_i' \hat{\beta}(\hat{\theta})$ based on the weighted least squares estimator $\hat{\beta}(\hat{\theta})$.

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 $\bar{y}_i$ 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 \( \beta = 0, \)
\( p = 0, n_1 = \cdots = n_k = n \) and \( N = nk \), and treat the unbiased estimator \( \hat{\theta}_1 \)
and the truncated estimator \( \hat{\theta}_{1R} \) 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 \( \max\{1, (n/k) \sum_{j=1}^{k} \bar{y}_j^2/(S_1/(N − k))\} \) since
\( S = \sum_{i,j} y_{ij}^2 = S_1 + n \sum_i \bar{y}_i^2 \) for \( S_1 = \sum_{i,j}(y_{ij} − \bar{y}_i)^2 \). Then, the EBLUP given in (3.7) can be expressed as for \( \hat{\sigma}^2 = S_1/(N − k) \),

\[
\hat{\mu}_{EB}^E(\hat{\theta}) = \max\left\{ 0, 1 − \frac{k\hat{\sigma}^2}{n \sum_{j=1}^{k} \bar{y}_j^2} \right\} \bar{y}_i,
\]

which is related to the positive-part Stein estimator. The Stein problem has been de-
veloped 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 si-
multaneous 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 Measurements 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 expan-
sion by Kackar and Harville (1984), Prasad and Rao (1990), Harville and Jeske (1992),
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 \( a \) and \( b \) be \( p \times 1 \) and \( M \times 1 \) vectors of fixed constants, and suppose that we want
to estimate the scalar quantity \( \mu = a^T \beta + b^T v \). Since the conditional distribution of \( v \)
given \( y \) is given by

\[
v | y \sim N_M(G(\theta)Z'\Sigma(\theta)^{-1}(y − X\beta), (G(\theta)^{-1} + Z'R(\theta)^{-1}Z)^{-1}),
\]

(4.1)
the conditional expectation $E[\mu|y]$ is written as

$$
\hat{\mu}^B(\beta, \theta) = E[\mu|y] = a'\beta + b'G(\theta)Z'\Sigma(\theta)^{-1}(y - X\beta)
$$

(4.2)

$$
= a'\beta + s(\theta)'(y - X\beta),
$$

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

(4.3)

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

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 $\hat{\mu}^{EB}(\theta)$ when $N$ is large, where the MSE is given by

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

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

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

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

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

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

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

Defined $g_1(\theta)$, $g_2(\theta)$ and $g_3(\theta)$ by

$$
g_1(\theta) = b'(G(\theta)^{-1} + Z'R(\theta)^{-1}Z)^{-1}b,
g_2(\theta) = (a - X's(\theta))'(X'\Sigma(\theta)^{-1}X)^{-1}(a - X's(\theta)),
g_3(\theta) = \text{tr}\left[\left(\frac{\partial s(\theta)'}{\partial \theta}\right)\Sigma(\theta)\left(\frac{\partial s(\theta)'}{\partial \theta}\right)'\text{Cov}(\hat{\theta}^*)\right],
$$

(4.5)

for $\text{Cov}(\hat{\theta}^*) = E[(\hat{\theta}^* - E[\hat{\theta}^*])(\hat{\theta}^* - E[\hat{\theta}^*])']$.

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

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

(4.6)
We next provide an asymptotically unbiased estimator of $MSE(\theta, \hat{\mu}_{EB}(\hat{\theta}))$ with the second-order accuracy. Define $g_{11}(\theta)$ and $g_{12}(\theta)$ by

$$g_{11}(\theta) = \left(\frac{\partial g_1(\theta)}{\partial \theta}\right)' E[\hat{\theta}^* + \hat{\theta}^{**}],$$

$$g_{12}(\theta) = \frac{1}{2} \text{tr} \left[ B(\theta) \text{Cov}(\hat{\theta}^*) \right],$$

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

$$B_{i,j}(\theta) = (b - Z's(\theta))'(\partial_{ij} G(\theta))(b - Z's(\theta)) + s(\theta)'(\partial_{ij} R(\theta))s(\theta).$$

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

$$\text{mse}(\hat{\theta}, \hat{\mu}_{EB}(\hat{\theta})) = g_1(\hat{\theta}) + g^#(\hat{\theta}),$$

where

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

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

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

$$E[\text{mse}(\hat{\theta}, \hat{\mu}_{EB}(\hat{\theta}))] = MSE(\theta, \hat{\mu}_{EB}(\hat{\theta})) + O(N^{-3/2}).$$

### 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 $\mu = a'\beta + b'v$ which satisfies the nominal confidence level with the second-order accuracy.

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

$$I_{EB}(\hat{\theta}) : \hat{\mu}_{EB}(\hat{\theta}) \pm z_{\alpha/2} \sqrt{\text{mse}(\hat{\theta})}.$$
the significance point $z_{\alpha/2}$ as $z_{\alpha/2}\{1 + h(\hat{\theta})\}$ by using an appropriate correction function $h(\theta)$. That is, the corrected confidence interval is described as

$$I_{CEB}(\theta) : \hat{\mu}_{EB}(\theta) \pm z_{\alpha/2} \left[ 1 + h(\theta) \right] \sqrt{mse(\theta)}.$$ 

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

(4.13) \[ h(\theta) = \frac{z_{\alpha}^2 + 1}{8g_1(\theta)^2} \text{tr} \left[ \left( \frac{\partial g_1(\theta)}{\partial \theta} \right) \left( \frac{\partial g_1(\theta)}{\partial \theta} \right)' \text{Cov} \left( \hat{\theta}^* \right) \right]. \]

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

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

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 approximations, 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 $\mu_i = \bar{\mathbf{x}}_i \hat{\beta} + v_i$ in NERM. The EBLUP of $\mu_i$ is $\hat{\mu}_{iEB}(\theta) = \bar{\mathbf{x}}_i \hat{\beta}(\theta) + \{n_i\hat{\theta}_2/(\theta_1 + n_i\hat{\theta}_2)\} \left\{ \bar{y}_i - \bar{\mathbf{x}}_i \hat{\beta}(\theta) \right\}$ from (3.7). The MSE approximation of $\hat{\mu}_{iEB}(\theta)$, its unbiased estimator and the confidence interval based on $\hat{\mu}_{iEB}(\theta)$ with the second-order accuracy are provided from Theorems 4.1, 4.2 and 4.3, where the functions $g_1(\theta)$, $g_2(\theta)$, $g_3(\theta)$, $g_{11}(\theta)$ and $h(\theta)$ are expressed as $g_1(\theta) = \theta_1\theta_2(\theta_1 + n_i\hat{\theta}_2)^{-1}$, $g_2(\theta) = \theta_1^2(\theta_1 + n_i\hat{\theta}_2)^{-2} \mathbf{X}_s' \mathbf{X} \mathbf{X}_s^{-1} \mathbf{X}^{-1} \mathbf{X}_s$, $g_3(\theta) = n_i(\theta_1 + n_i\hat{\theta}_2)^{-3}(\theta_2, \theta_1) \text{Cov} \left( \hat{\theta}^* \right) (\theta_2, \theta_1)'$, $g_{11}(\theta) = (\theta_1 + n_i\hat{\theta}_2)^{-2}(n_i\hat{\theta}_2^3, \theta_1^2) E[\hat{\theta}^* + \hat{\theta}^{**}']$, $h(\theta) = \frac{z_{\alpha}^2 + 1}{8(\theta_1 + n_i\hat{\theta}_2)^2(n_i\hat{\theta}_2^2, \theta_1^2) \text{Cov} \left( \hat{\theta}^* \right) (n_i\hat{\theta}_2^2, \theta_1^2)'}$.

and $g_{12}(\theta) = 0$. For estimator $\hat{\theta}$ satisfying the conditions (C4) and (C5), we need to obtain $\text{Cov} \left( \hat{\theta}^* \right)$ and $E[\hat{\theta}^* + \hat{\theta}^{**}]$. 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 $I_{CEB}(\theta)$ tends to be unstable near $\theta_2 = 0$, because the corrected function $h(\theta)$ given in (4.13) includes $g_1(\theta)$ in the denominator. In NERM, $g_1(\theta)$ is $g_1(\theta) = \theta_1\theta_2/(\theta_1 + n_i\hat{\theta}_2)$ and takes values near zero when $\theta_2$ 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 $\hat{\theta}_2^{IR} = \max\{\hat{\theta}_2, N^{-2/3}\}$, which was suggested in Kubokawa (2009a). For the practical use of $I_{CEB}(\theta)$, we need such a modification of the estimator $\hat{\theta}$. 

15
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, \( y \sim N(X\beta, \Sigma) \) for \( \theta = (\theta_1, \ldots, \theta_q)' \). The hypothesis to be tested is the linear restriction given by

\[
H_0 : R\beta = r,
\]

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

\[
\hat{\beta} = \hat{\beta}(\theta) = (X'\Sigma(\theta)^{-1}X)^{-1}X'\Sigma(\theta)^{-1}y,
\]

\[
\tilde{\beta} = \tilde{\beta}(\theta) = \hat{\beta}(\theta) - (X'\Sigma(\theta)^{-1}X)^{-1}R'W(\theta)(R\hat{\beta}(\theta) - r),
\]

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

[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 \( H_0 : a(\xi) = 0 \) against \( H_1 : a(\xi) \neq 0 \), where a random variable \( X \) has a likelihood function \( L(\xi|X) \), \( \xi \) is a \( p \)-dimensional unknown vector and \( a(\xi) \) is a function from \( R^p \) to \( R^q \) for \( q \leq p \). Then, the Wald, likelihood ratio and Lagrange multiplier test statistics are given by

\[
F_W = a(\xi)'[A(\xi)I(\xi)A(\xi)']^{-1}a(\xi),
\]

\[
F_{LR} = -2\{\log L(\xi|X) - \log L(\tilde{\xi}|X)\},
\]

\[
F_{LM} = s(\xi)'[I(\tilde{\xi})^{-1}s(\xi)],
\]

where \( A(\xi) = \partial a(\xi)/\partial \xi \), \( I(\xi) = E[s(\xi)s(\xi)'] \) is the Fisher information matrix, \( s(\xi) = \partial \log L(\xi|X)/\partial \xi \) is the score function, and \( \xi \) and \( \tilde{\xi} \) are unrestricted and restricted estimator of \( \xi \). 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 \( H_0 : R\beta = r \) in the general linear regression model (2.2), these test statistics are written as

\[
F_W = (R\hat{\beta}_u - r)'W(\hat{\theta})(R\hat{\beta}_u - r),
\]

\[
F_{LR} = -2[\ell(\hat{\beta}_R, \hat{\theta}) - \ell(\hat{\beta}_u, \hat{\theta})],
\]

\[
F_{LM} = (R\tilde{\beta}_R - r)'W(\tilde{\theta})(R\tilde{\beta}_R - r),
\]

(5.1)
where \( W(\theta) = [R(X'\Sigma^{-1}X)^{-1}R']^{-1} \) and \(-2\ell(\beta, \theta) = \log |\Sigma(\theta)| + (y - X\beta)'\Sigma(\theta)^{-1}(y - X\beta)\). 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 statistics given in (5.1) were derived by Rothenberg (1984) under the null and local alternative hypotheses. Let

\[
C_i = (\Sigma^{-1})_{(i)} H(\theta),
\]

\[
D_{ij} = H(\theta)\{ (\Sigma^{-1})_{(ij)} - (\Sigma^{-1})_{(i)}X(X'\Sigma^{-1}X)^{-1}X'(\Sigma^{-1})_{(j)} 
- (\Sigma^{-1})_{(j)}X(X'\Sigma^{-1}X)^{-1}X'(\Sigma^{-1})_{(i)} \},
\]

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

\[
b(\theta) = \frac{1}{2} \text{tr} \left[ \text{Cov}(\hat{\theta}) \text{mat}_{ij} \left( \text{tr} \left[ H(\theta)\Sigma^{-1}\Sigma_{(i)}\Pi(\theta)\Sigma_{(j)}\Sigma^{-1} \right] \right) \right],
\]

\[
c(\theta) = \frac{1}{2} \text{tr} \left[ \text{Cov}(\hat{\theta}) \text{mat}_{ij} \left( \text{tr} \left[ C_i C_j \right] + \frac{1}{2} \text{tr} \left[ C_i \text{tr} \left[ C_j \right] \right] \right) \right],
\]

\[
d(\theta) = \text{E}[ (\hat{\theta} + \hat{\theta}^\ast) \text{col}_{i}(\text{tr} \left[ C_i \right]) + \text{tr} \left[ \text{Cov}(\hat{\theta}) \text{mat}_{ij} \left( \text{tr} \left[ C_i \text{tr} \left[ C_j \right] \right] \right) \right] + \frac{1}{2} \text{tr} \left[ \text{Cov}(\hat{\theta}) \text{mat}_{ij} \left( \text{tr} \left[ D_{ij} \right] \right) \right],
\]

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

\[
F^*_W = F_W / \left[ 1 + (\hat{\theta} + \hat{\theta}) / q + \hat{c}z_{\alpha} / \{q(q + 2)\} \right],
\]

where \( \hat{b} = b(\theta) \), \( \hat{c} = c(\theta) \), \( \hat{d} = d(\theta) \), and \( z_{\alpha} \) is the 100\( \alpha \)% upper point of the \( \chi^2_q \)-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 \( \hat{\theta} \) and \( \hat{\theta}^\ast \) are the unrestricted and restricted ML estimators of \( \theta \), \( F_{LR} \) and \( F_{LM} \) are approximated as

\[
F_{LR} = F_W - \frac{1}{2} (\hat{\theta} - \hat{\theta}^\ast)'A_2(\hat{\theta} - \hat{\theta}^\ast) + o_p(N^{-1}),
\]

\[
F_{LM} = F_W - (\hat{\theta} - \hat{\theta}^\ast)'A_2(\hat{\theta} - \hat{\theta}^\ast) + o_p(N^{-1}),
\]

for \( A_2 = \text{mat}_{ij}(\text{tr} \left[ \Sigma_{(i)}\Sigma\Sigma_{(j)}\Sigma \right]) \). The Bartlett-type corrections for \( F_{LR} \) and \( F_{LM} \) are given by

\[
F^*_{LR} = F_{LR} / \left[ 1 + (\hat{d} - \hat{c}) / q \right],
\]

\[
F^*_{LM} = F_{LM} / \left[ 1 + (\hat{d} - \hat{c} - \hat{b}) / q - \hat{c}z_{\alpha} / \{q(q + 2)\} \right],
\]

which can be derived by evaluating the term \( (\hat{\theta} - \hat{\theta}^\ast)'A_2(\hat{\theta} - \hat{\theta}^\ast) \). 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 \( f(y|v, \beta, \theta) \) and \( f(v|\theta) \) be the conditional density of \( y \) given \( v \) and the marginal density of \( v \), respectively, where \( y|v \sim N(X\beta + Zv, R(\theta)) \) and \( v \sim N(0, G(\theta)) \). Then, the marginal density of \( y \) is written by \( f_m(y|\beta, \theta) = \int f(y|v, \beta, \theta)f(v|\theta)dv \), which has marginal distribution \( N(X\beta, \Sigma(\theta)) \).

[A1] 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 distance. This quantity measures the prediction error in predicting future variable \( y^* \) based on the model \( f_m(y^*|\hat{\beta}(y), \hat{\theta}(y)) \). In this sense, AIC is a criterion of finding a model which can provide a good prediction in light of minimizing the prediction error. \( R(\beta, \theta; \hat{\beta}, \hat{\theta}) \) is rewritten as

\[
R(\beta, \theta; \hat{\beta}, \hat{\theta}) = E_y \left[ \int \left( \log \frac{f_m(y^*|\beta, \theta)}{f_m(y^*|\hat{\beta}(y), \hat{\theta}(y))} \right) f_m(y|\beta, \theta)dy^* \right],
\]

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

\[
\int \int \{ \log f_m(y^*|\beta, \theta) \} f_m(y|\beta, \theta)dy^* f_m(y|\beta, \theta)dy \\
- \int \int \{ \log f_m(y^*|\hat{\beta}(y), \hat{\theta}(y)) \} f_m(y|\beta, \theta)dy^* f_m(y|\beta, \theta)dy,
\]

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

\[
AI = -2 \int \int \{ \log f_m(y^*|\hat{\beta}(y), \hat{\theta}(y)) \} f_m(y^*|\beta, \theta)f_m(y|\beta, \theta)dy^*dy,
\]

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 \( \beta \) is estimated by the GLS \( \hat{\beta}(\hat{\theta}) \) for a consistent estimator of \( \theta \), AIC is given as

\[
(5.2) \quad AIC_c = -2\log f_m(y|\hat{\beta}(\hat{\theta}), \hat{\theta}) + 2(p + q),
\]

where \(-2\log f_m(y|\hat{\beta}(\hat{\theta}), \hat{\theta}) = N\log(2\pi) + \log |\Sigma(\hat{\theta})| + y'\Sigma(\hat{\theta})y \) for \( \Pi(\theta) \) defined in (3.1), and \( p \) and \( q \) are dimensions of \( \beta \) and \( \theta \), respectively.
It is noted that the AIC stated above is based on the marginal distribution of \( y \), namely, it measures the prediction error of the predictor based on the marginal distribution \( N(X\beta, \Sigma) \). This means that the marginal AIC is not appropriate for the focus on the prediction of specific 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 area estimation. Taking this point into account, Vaida and Blanchard (2005) proposed the conditional AIC as an asymptotically unbiased estimator of area estimation, where \( AI \) is the conditional Akaike information defined by

\[
(5.3) cAI = -2\int\int\int \log\{f(y^*|\hat{\nu}(y), \hat{\beta}(y), \hat{\theta}(y))\} f(y^*|v, \beta, \theta) f(y|v, \beta, \theta) f(v|\theta) dy^* dy dv,
\]

where \( \hat{\nu}(y) = \hat{\nu} \) is the empirical Bayes estimator of \( v \) given in (2.8). When \( \theta \) 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

\[
(5.4) cAIC_c = -2\log f(y|\hat{\nu}(\theta), \hat{\beta}(\theta), \hat{\theta}) - \Delta_c,
\]

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

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

\[
f_\pi(y) = \int\int f_m(y|\beta, \theta)\pi(\beta, \theta) d\beta d\theta.
\]

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

\[
(5.5) BIC = -2\log\{f_m(y|\hat{\beta}(\theta), \hat{\theta})\} + (p + q) \log(N),
\]

where \(-2\log\{f_m(y|\hat{\beta}(\theta), \hat{\theta})\}\) 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 \((\beta, \theta)\), 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 \( \beta \) is the parameters of interest and \( \theta \) is the nuisance parameters in LMM. Assume that \((\beta, \theta)\) has the prior distribution

\[
(\beta, \theta) \sim \pi_1(\beta|\theta, \lambda)\pi_2(\theta),
\]

19
where given $\theta, \beta$ conditionally has $\pi_1(\beta|\theta, \lambda)$ with hyperparameter $\lambda$. Let $f_{x,1}(y|\theta, \lambda)$ be the conditional marginal density given by

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

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

$$
\hat{\lambda} = \arg \max_{\lambda} \{f_{x,1}(y|\hat{\theta}, \lambda)\}.
$$

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

$$
(5.6) \quad EBIC = -2 \log \{f_{x,1}(y|\hat{\theta}, \hat{\lambda})\} + q \log(N).
$$

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

$$
\pi_1(\beta|\lambda) = N_p(0, \lambda^{-1}W)
$$

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

$$
f_{x,1}(y|\theta, \lambda) = \frac{1}{(2\pi)^{N/2}} |\Sigma(\theta)|^{-1/2} \exp\left\{-\frac{1}{2} y' \Pi^*(\theta, \lambda)y \right\},
$$

where

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

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

$$
y'\hat{\Sigma}^{-1}X(X'\hat{\Sigma}^{-1}X + \lambda_0 W^{-1})^{-1}W^{-1}(X'\hat{\Sigma}^{-1}X + \lambda_0 W^{-1})^{-1}X'\hat{\Sigma}^{-1}y
= \text{tr} [(X'\hat{\Sigma}^{-1}X + \lambda_0 W^{-1})^{-1}X'\hat{\Sigma}^{-1}X]/\lambda_0.
$$

Then the EBIC is given by

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

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


