Jackknifing The Mean Squared Error of Empirical Best Predictor

Jiming Jiang, Partha Lahiri, and Shu-Mei Wan

Department of Statistics
Case Western Reserve University
Cleveland, OH 44106-7054, USA
jiang@eureka.cwru.edu

Department of Mathematics and Statistics
University of Nebraska at Lincoln
Lincoln, NE 68588-0323, USA
plahiri@mathstat.unl.edu

1. Introduction. In recent years, Empirical Best Linear Unbiased Prediction (EBLUP) approach has received considerable importance in producing small-area statistics. This is really a special case of Empirical Best Prediction (EBP) approach which can be applied even when we do not have mixed *linear* model need for EBLUP approach. The main focus of this paper is to develop a general theory for EBP approach.

We develop a suitable jackknife technique to estimate the MSE of EBP of any general mixed effect for a general model. The proposed jackknife method is very simple to implement and does not require the derivation of different derivatives needed in the Taylor series method. Thus the method should be very attractive to the practitioners. The general model we consider covers not only the mixed linear model but also many complex models like generalized linear mixed model. So long as one can get expression for EBP, our method can be applied. For example, we no longer require the assumption of normality to estimate the MSE of EBLUP given in Prasad and Rao (1990) - we just need the assumption of *posterior linearity* (see, e.g., Ghosh and Meeden 1997) which is needed anyway to justify the use of EBLUP (which is identical with linear empirical Bayes (LEB) estimator). In addition, the proposed jackknife method will work for a general M-estimator of the model parameters (which includes ML, REML and ANOVA).

The properties of the jackknife estimators have been studied extensively in the literature (see, e.g., Shao and Tu (1995)). However, the problems discussed in the paper are not currently available in the literature. First, our main interest is not in the estimation of a fixed parameter but in the prediction of a random vector which may be associated with unknown parameters. This is, of course, a more complicated problem. Secondly, even for estimating the fixed parameters, our jackknife estimator is not based on i.i.d. observations and it is not associated with the regression estimator. Furthermore, since we consider a specific class of estimators, namely, the M-estimators, the conditions under which the asymptotic results hold will be easier to verify. As will be seen, the asymptotic unbiasedness of the jackknife MSE estimator is proved essentially under some moment conditions. Thirdly, our M-estimators are more general than those considered by Reeds (1978) in the sense that ours also include the modified profile MLE (e.g., REML estimators), penalized MLE, or M-estimators not associated with a maximization process (e.g., the method of moment estimators).

Section 2 discusses the model and the proposed EBP. In section 3, we propose a jackknife method to measure the uncertainty of the proposed EBP. The asymptotic properties of our jackknife MSE estimator are also stated in this section. The mixed linear models and mixed logistic models which are important special cases of our general model are discussed in sections 2 and 4, respectively. Due to lack of space, we refer to Jiang, Lahiri and Wan (1998) for proofs of all the technical results.

2. Empirical best predictor. Let $Y_1, \dots, Y_m$ be independent (vector-valued) observations whose distributions depend on a vector $\phi = (\phi_k)_{1\le k\le s}$ of unknown parameters. We are interested in predicting an unobservable random vector $\theta = (\theta_l)_{1\le l\le t}$ based on $Y = (Y_j)_{1\le j\le m}$. Supopose that, when ϕ is known, the best predictor in terms of MSE is $\breve{\theta} = E(\theta|Y) = \pi(Y_S;\phi) = (\pi_l(Y_S;\phi))_{1\le l\le t}$, where S is a subset of $\{1,\dots,m\}$ and $Y_S = (Y_j)_{j\in S}$. For example, in small-area estimation, one is interested in predicting a mixed effect $\theta = h'\beta + \lambda' v$, where h and λ are known vectors, β is a vector of unknown fixed effects, and v is a vector of unobservable small-area specific random effects. In particular, a mixed effect associated with the ith small-area is of the form $\theta = h'\beta + v_i$. Assuming that the random effects corresponding to the small-areas are independent, the best predictor of θ is of the form $\pi(Y_i;\phi)$, where Y_i is the vector of observations associated with the ith small-area, and ϕ is the combination of β and a vector ψ of variance components.

Since ϕ is usually unknown, it is naturally replaced by an estimator, $\hat{\phi}$. The resulting predictor, $\hat{\theta} = \pi(Y_S;\hat{\phi})$ is called the empirical best predictor of θ. The estimator $\hat{\phi}$ of particular interest in this paper is an M-estimator (Huber (1981)), which is associated with a solution $\dot{\phi} = (\dot{\phi}_k)_{1\le k\le s}$ to the following equation:

$$F(\phi;Y) \;=\; \sum_{j=1}^{m} f_j(\phi;Y_j) + a(\phi) \;=\; 0\,. \tag{1}$$

In the above, $f_j(\phi;Y_j) = (f_{j,k}(\phi;Y_j))_{1\le k\le s}$ are vector-valued functions such that $Ef_j(\phi;Y_j) = 0$ when ϕ is the true parameter vector, and $a(\phi)$ is a vector-valued function which may depend on the joint distribution of $Y = (Y_j)_{1\le j\le m}$. When $a(\phi) \neq 0$, it plays the role of a modifier or penalizer.

Example 1. (ML estimator in mixed linear models) Consider a mixed linear model

$$Y_i \;=\; X_i\beta + Z_i v_i + e_i\,,\quad i = 1,\dots,m\,, \tag{2}$$

where X_i ($n_i \times p$) and Z_i ($n_i \times b_i$) are known matrices, v_i and e_i are independently distributed with $v_i \overset{\text{ind}}{\sim} (0, G_i)$ and $e_i \overset{\text{ind}}{\sim} (0, R_i)$, $i = 1,\dots,m$. Assume that $G_i = G_i(\psi)$ ($b_i \times b_i$) and $R_i = R_i(\psi)$ ($n_i \times n_i$) possibly depend on $\psi = (\psi_l)_{1\le l\le q}$, a $q\times 1$ vector of variance components. The ML estimator of $\phi = (\beta'\ \psi')'$ is defined as solution to the ML equations. Note that this definition does not require normality, i.e., the ML equations are used even if the data is not normal (Jiang (1996)). It is easy to show that the ML estimator of ϕ is solution to (1) where $a(\phi) = 0$, $(f_{j,k}(\phi,Y_j))_{1\le k\le p} = X_j'\Sigma_j^{-1}(\psi)(Y_j - X_j\beta)$, where $\Sigma_j(\psi) = R_j(\psi) + Z_jG_j(\psi)Z_j' = \mathrm{Var}(Y_j)$, and for $1 \le l \le q$

$$f_{j,p+l}(\phi,Y_j) = (Y_j - X_j\beta)'\Sigma_j^{-1}(\psi)\left(\frac{\partial\Sigma_j}{\partial\psi_l}\right)\Sigma_j^{-1}(\psi)(Y_j - X_j\beta) - \mathrm{tr}\left(\Sigma_j^{-1}(\psi)\frac{\partial\Sigma_j}{\partial\psi_l}\right)\,.$$

Example 2. (REML estimator in mixed linear models) Similarly, the REML estimator $\dot{\psi}$ of ψ is defined as solution to the REML equations (Jiang (1996)), and the REML estimator of β as the EBLUE $\dot{\beta} = (X'\Sigma^{-1}(\dot{\psi})X)^{-1}X'\Sigma^{-1}(\dot{\psi})Y$, where $X = \mathrm{col}_{1\le i\le m}(X_i)$, $Y = \mathrm{col}_{1\le i\le m}(Y_i)$, $\Sigma(\psi) = R + ZGZ'$, $Z = \mathrm{diag}_{1\le i\le m}(Z_i)$, $G = \mathrm{diag}_{1\le i\le m}(G_i)$, and $R = \mathrm{diag}_{1\le i\le m}(R_i)$. Again, this definition does not require normality. By the identity (e.g., Searle *et al* (1992), page 451) $\Sigma^{-1} = \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} + A(A'\Sigma A)^{-1}A'$, which holds for any $N\times(N-p)$ matrix A of full rank (N is the dimension of Y) such that $A'X = 0$, it is easy to show that the REML estimator of ϕ is solution to (1) where the f_j's are the same as in Example 1; $a(\phi) = (a_k(\phi))_{1\le k\le p+q}$ with

$a_k(\phi) = 0$, $1 \le k \le p$ and

$$a_{p+l}(\phi) \;=\; \sum_{j=1}^{m} \mathrm{tr}\left(\Sigma_j^{-1}(\psi) X_j (X'\Sigma^{-1}(\psi)X)^{-1} X_j' \Sigma_j^{-1}(\psi) \frac{\partial \Sigma_j}{\partial \psi_l} \right) , \quad 1 \le l \le q .$$

In general, $\dot{\phi}$ may not always exist; or even if it does, may fall outside the parameter space. Of course, the MSE of $\dot{\phi}$ also may not Therefore, we consider the following *truncated* version of $\dot{\phi}$. Let Φ be the parameter space for ϕ. Let ϕ^* be a fixed vector in Φ. (In practice, ϕ^* may be a reasonable guess of the true ϕ). Let $\dot{\phi}$ be the solution to (1) if such a solution exists and lies in Φ; otherwise, let $\dot{\phi} = \phi^*$. Define the estimator $\hat{\phi}$ as follows: $\hat{\phi} = \dot{\phi}$ if $|\dot{\phi}| \le K(\log m)^\alpha$; and $\hat{\phi} = \phi^*$ otherwise, where K and α are positive (known) constants. It is clear that such a truncation will not affect the asymptotic properties such as consistency and efficiency of the estimator.

3. Jackknifing MSE of EBP. The main interest of this section is the estimation of the MSE of the proposed EBP, $\mathrm{MSE}(\hat{\theta}) = E(|\hat{\theta} - \theta|^2)$. We propose to do so by the *Jackknife* method. For such a purpose, we define the M-estimator $\hat{\phi}_{-i}$ after deleting the ith observation, i.e., $\hat{\phi}_{-i}$ is obtained likewise from the solution $\dot{\phi}_{-i}$ to the equation:

$$F_{-i}(\phi; Y_{-i}) \;=\; \sum_{j \ne i} f_j(\phi; Y_j) + a_{-i}(\phi) \;=\; 0 , \tag{3}$$

where $Y_{-i} = (Y_j)_{j \ne i}$. Note that $a_{-i}(\cdot)$ may not be the same function as $a(\cdot)$. Observe that

$$\mathrm{MSE}(\hat{\theta}) \;=\; E(|\hat{\theta} - \check{\theta}|^2) + E(|\check{\theta} - \theta|^2) \;=\; \mathrm{MSAE}(\hat{\theta}) + \mathrm{MSE}(\check{\theta}) , \tag{4}$$

where MSAE stands for "mean squared approximation error" (to the best predictor). A Jackknife estimator of the first term on the right side of (2) is given by

$$\widehat{\mathrm{MSAE}}(\hat{\theta}) \;=\; \frac{m-1}{m} \sum_{i=1}^{m} |\hat{\theta}_{-i} - \hat{\theta}|^2 , \tag{5}$$

where $\hat{\theta}_{-i} = \pi(Y_S; \hat{\phi}_{-i})$. Note that we keep Y_S the same (i.e., not affected by deleting the ith observation) in all $\hat{\theta}_{-i}$s. As for the second term, it is often possible to obtain a closed form expression which is a function of ϕ. Suppose $\mathrm{MSE}(\check{\theta}) = b(\phi)$. Then, a Jackknife estimator of $b(\phi)$ is given by

$$\widehat{\mathrm{MSE}}(\check{\theta}) \;=\; b(\hat{\phi}) - \frac{m-1}{m} \sum_{i=1}^{m} [b(\hat{\phi}_{-i}) - b(\hat{\phi})] . \tag{6}$$

Therefore, the Jackknife estimator of the MSE of $\hat{\theta}$ is

$$\widehat{\mathrm{MSE}}(\hat{\theta}) \;=\; \widehat{\mathrm{MSAE}}(\hat{\theta}) + \widehat{\mathrm{MSE}}(\check{\theta}) . \tag{7}$$

It can be shown that under certain regularity conditions, the bias of $\widehat{\mathrm{MSE}}(\hat{\theta})$ is of the order $o(m^{-1})$. As a by product, we also obtain the asymptotic unbiasedness of $MSE(\phi)$.

4. Mixed logistic models. Suppose that, conditional on p_{ij}, Y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$ are independent *Bernoulli* random variable with $P(Y_{ij} = 1 | p_{ij}) = p_{ij}$. Furthermore, suppose that conditional on the random effects $\alpha_1, \ldots, \alpha_m$, $\mathrm{logit}(p_{ij}) = x_{ij}^t \beta + \alpha_i$, where $x_{ij} = (x_{ijk})_{1 \le k \le p}$ is a vector of known covariates, β is a vector of unknown regression coefficients, and $\mathrm{logit}(t) =$

$\log(t/(1-t))$. Assume the α's are independent and distributed as $N(0, \sigma^2)$. Then, (5.1) is a special case of the generalized linear mixed models which have received considerable attention (e.g., Breslow and Clayton (1993), Lee and Nelder (1996)). Such models as (5.1) have been used in small-area inference with binary variables (e.g., Malec *et al* (1997)).

Suppose that one is interested in predicting a (possibly nonlinear) mixed effect $\theta = h_i(\beta, \alpha_i)$. For example, $\theta = \alpha_i$; or, if the covariates take values from a finite set $\{x_1, \ldots, x_K\}$, $\theta = \sum_{k=1}^K w_k \text{logit}^{-1}(x_k^t \beta + \alpha_i)$, where w_k, $1 \le k \le K$ is a set of weights, and $\text{logit}^{-1}(u) = e^u/(1+e^u)$.

Jiang and Lahiri (1998) derive the best predictor of θ as

$$\check{\theta} \;=\; E(\theta|Y) \;=\; \frac{Eh_i(\beta, \sigma\xi)\exp(\psi_i(Y_{i\cdot}, \sigma\xi, \beta))}{E\exp(\psi_i(Y_{i\cdot}, \sigma\xi, \beta))} \;=\; \pi_i(Y_{i\cdot}, \phi)\,, \tag{8}$$

where $\psi_i(k, u, v) = ku - \sum_{j=1}^{n_i} \log(1 + \exp(x_{ij}^t v + u))$, $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$, $\phi = (\beta^t\ \sigma)^t$, and the expectations are taken over $\xi \sim N(0,1)$. It is also shown that $\text{MSE}(\check{\theta}) = Eh_i^2(\beta, \sigma\xi) - \sum_{k=0}^{n_i} \pi_i^2(k, \phi) p_i(k, \phi) \equiv b_i(\phi)$, where $p_i(k, \phi) = P(Y_{i\cdot} = k)$. The empirical best predictor is given by $\hat{\theta} = \pi_i(Y_{i\cdot}, \hat{\phi})$.

As for the M-estimators, we consider the method of moments (MM) estimators of Jiang (1998). The MM estimator for ϕ is the solution to the following system of equations:

$$\sum_{i=1}^m \sum_{j=1}^{n_i} x_{ijk} Y_{ij} \;=\; \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ijk} E_\phi Y_{ij}\,, \quad 1 \le k \le p\,, \tag{9}$$

$$\sum_{i=1}^m \sum_{j \ne l} Y_{ij} Y_{il} \;=\; \sum_{i=1}^m \sum_{j \ne l} E_\phi Y_{ij} Y_{il}\,. \tag{10}$$

Note that $E_\phi Y_{ij} = E\text{logit}^{-1}(x_{ij}^t \beta + \sigma\xi)$, $EY_{ij}Y_{il} = E\text{logit}^{-1}(x_{ij}^t\beta + \sigma\xi)\text{logit}^{-1}(x_{il}^t\beta + \sigma\xi)$, $j \ne l$. Jiang and Lahiri (1998) showed that, under suitable conditions the MM estimators are consistent uniformly at rate m^{-d} for any $d > 0$. It follows can be shown that $E\widehat{\text{MSE}}(\hat{\theta}) = \text{MSE}(\hat{\theta}) + o(m^{-1-\epsilon})$ for any $0 < \epsilon < 1/2$.

REFERENCES

Breslow, N. E., and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.* 88, 9-25.

Huber, P. J. (1981), *Robust Statistics*, Wiley, New York.

Jiang, J. (1996), REML estimation: asymptotic behavior and related topics, *Ann. Statist.* 24, 255-286.

Jiang, J., and Lahiri, P. (1998), Empirical best prediction for small area inference with binary data, submitted.

Lee, Y., and Nelder, J. A. (1996), Hierarchical generalized linear models (with discussion), *J. Roy. Statist. Soc.* B 58, 619-678.

Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997), Small area inference for binary variables in the national health interview survey, *J. Amer. Statist. Assoc.* 92, 815-826.

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

Reeds, J. A. (1978), Jackknifing maximum likelihood estimates, *Ann. Statist.* 6, 727-739.

Shao, J., and Tu, D. (1995), *The Jackknife and The Bootstrap*, Springer, New York.