The Annals of Statistics
2002, Vol. 30, No. 6, 1782–1810

A UNIFIED JACKKNIFE THEORY FOR EMPIRICAL BEST PREDICTION WITH M-ESTIMATION

BY JIMING JIANG,[1,2] P. LAHIRI[1,3] AND SHU-MEI WAN

University of California, Davis, University of Nebraska, Lincoln and Lunghwa University of Science and Technology, Taiwan

The paper presents a unified jackknife theory for a fairly general class of mixed models which includes some of the widely used mixed linear models and generalized linear mixed models as special cases. The paper develops jackknife theory for the important, but so far neglected, prediction problem for the general mixed model. For estimation of fixed parameters, a jackknife method is considered for a general class of M-estimators which includes the maximum likelihood, residual maximum likelihood and ANOVA estimators for mixed linear models and the recently developed method of simulated moments estimators for generalized linear mixed models. For both the prediction and estimation problems, a jackknife method is used to obtain estimators of the mean squared errors (MSE). Asymptotic unbiasedness of the MSE estimators is shown to hold essentially under certain moment conditions. Simulation studies undertaken support our theoretical results.

1. Introduction. Due to the advent of high-speed computers and powerful software, computer-oriented statistical methods, including various resampling methods, have received considerable attention in recent years as statisticians are constantly facing complex problems. The *jackknife* method is one such simple resampling method which is very popular among survey samplers, primarily due to its simplicity.

The properties of jackknife estimators, especially in the i.i.d. and regression cases, have been studied extensively in the literature [see, e.g., Efron and Tibshirani (1993), page 150; Shao and Tu (1995)]. However, the studies of jackknife for mixed models have been limited so far to simple random effects models. In this context, we refer to Arvesen (1969), Rao and Prasad (1986) and Prasad and Rao (1988) for jackknifing variance components and Lahiri (1995) and Chattopadhyay, Lahiri, Larsen and Reimnitz (1999) for jackknifing mean squared error (MSE) of empirical best predictors (EBP).

Received June 1998; revised January 2002.

[1] Supported in part by NSF Grants SES-95-11202 and SBR-97-05574 to P. Lahiri.

[2] Supported in part by NSF Grant SES-99-78101 and National Security Agency Grant MDA904-98-1-0038.

[3] Supported in part by NSF Grant SES-99-78145 and a grant from the Gallup Organization.

AMS 2000 subject classifications. 62G09, 62D05.

Key words and phrases. Empirical best predictors, mean squared errors, M-estimators, mixed linear models, mixed logistic models, small-area estimation, uniform consistency, variance components.

In this paper, we present a unified theory of the jackknife method for a general model which includes some of the widely used mixed linear models and generalized linear mixed models (GLMM) as special cases. Specifically, we propose a jackknife MSE estimator of EBP of a general mixed effect. The proposed jackknife MSE estimator can be obtained as long as it is possible to obtain an expression for the best predictor (BP) and it captures all different sources of variability. We show that, under certain mild regularity conditions, the proposed jackknife MSE estimator of EBP tends to the true MSE of EBP at a rate faster than that of a naive estimator for several important models. In fact, let m be the number of small areas in the sample. Then, it can be seen from our general results that under either a normal mixed linear model or a mixed logistic model the order of bias for our jackknife MSE estimator is $o(m^{-\lambda})$, where λ can be arbitrarily close to $3/2$, while for the naive MSE estimator the order of bias is $O(m^{-1})$. As a by-product, we establish a jackknife theory for general M-estimators which include maximum likelihood (ML), restricted maximum likelihood (REML) and method of moments (MOM) estimators. The latter method is used in inference about GLMM [Jiang (1998)]. Our regularity conditions and asymptotic theory cover many real life applications in small-area estimation, disease mapping and animal breeding problems.

The proposed jackknife method is very simple to implement and does not require the tedious derivations of various derivatives needed in the Taylor series method (see below). Thus, the method should be very attractive to practitioners.

The particular case of mixed normal linear model deserves further attention. In this case, certain MSE estimators of EBP based on Taylor series are equally efficient (in the second order asymptotic sense) to the proposed jackknife MSE estimator. See Prasad and Rao (1990) and Datta and Lahiri (2000), among others. The normality assumption is in general very crucial in deriving the Taylor series MSE estimators with the exception of Lahiri and Rao (1995) who derived Taylor series MSE estimator for nonnormal Fay–Herriot model [Fay and Herriot (1979)]. In this case, the BP can be obtained under the assumption of posterior linearity and hence the jackknife MSE estimator of EBP [also the empirical best linear unbiased predictor (EBLUP); see Section 5] continues to enjoy accurate rate of convergence even when the normality assumption does not hold true. For a discussion on posterior linearity, readers are referred to Ericson (1969). Our simulation results suggest that the jackknife estimators are more robust than the corresponding normality-based MSE estimators of Prasad and Rao (1990).

Before the paper gets technical, we first present, in Section 2, a table that lists a selection of applications of the main theorems to come. Section 3 introduces different notation used throughout the subsequent sections. Section 4 discusses jackknifing M-estimators. In Section 5, we propose a jackknife method to estimate the MSE of the proposed EBP. The asymptotic properties of our jackknife MSE estimators are also stated in these two sections. The mixed linear models and mixed logistic models which are important special cases of our general model are discussed in Sections 6 and 7, respectively. Simulation results are presented in Section 8. The proofs of these results are quite technically involved and therefore are deferred to a technical report [see Jiang (1999)].

2. Selected applications. The purpose of this section is to present some selected applications of our theorems given in later sections. The applications will be related to the following models which have been used in the literature. We begin with the simplest models. Whenever possible we motivate each model from real-life applications and then spell out formulae for a jackknife MSE estimator of EBP or variance estimator of a parameter of interest. We then indicate the relevant theorem (or proposition) which can be applied to obtain the order of bias of a particular jackknife MSE estimator.

2.1. *The James–Stein estimator.* Let $Y_i \mid \theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1)$, $i = 1, \dots, m$. In the context of simultaneous estimation of $\theta = (\theta_1, \dots, \theta_m)'$, it is well known that, for $m \geq 3$, the James–Stein estimator dominates the maximum likelihood estimator $Y = (Y_1, \dots, Y_m)'$ in terms of the frequentist risk under a sum of squared errors loss function; see Lehmann [(1983), page 302]. Efron and Morris (1973) provided an empirical Bayes justification of the James–Stein estimator. Their Bayesian model can be equivalently written as the following simple random effects model:

$$Y_i = v_i + e_i, \qquad i = 1, \dots, m,$$

where the sampling errors $\{e_i\}$ and the random effects $\{v_i\}$ are independently distributed with $v_i \sim N(0, A)$ and $e_i \sim N(0, 1)$, $i = 1, \dots, m$. Let $B = 1/(1 + A)$ be the unknown model parameter. It can be easily shown that the James–Stein estimator can be interpreted as an EBP under the above random effects model. The BP of $\theta_1 = v_1$ is given by $\breve{\theta}_1 = (1 - B)Y_1$. The model parameter B can be unbiasedly estimated by $\hat{B} = (m - 2)/\sum_{i=1}^m Y_i^2$ [Efron and Morris (1973)]. An EBP is then given by $\hat{\theta} = (1 - \hat{B})Y_1$.

Note that the MSE of the BP is given by $b(B) = 1 - B$. Thus, a naive estimator of MSE of EBP can be obtained by estimating $b(B)$ by $b(\hat{B})$. This, however, underestimates the true uncertainty of EBP since it does not incorporate the variability due to the estimation of the model parameter B. Note that, since in this case $b(\hat{B})$ is an unbiased estimator of $b(B)$, our proposed jackknife MSE estimator of θ is given by

$$\text{mse}^{\text{J}} = b(\hat{B}) + \frac{m-1}{m} \sum_{u=1}^{m} [\hat{\theta}_{-u} - \hat{\theta}]^2 = (1 - \hat{B}) + v_{\text{J}} Y_1^2,$$

where $\hat{\theta}_{-u} = (1 - \hat{B}_{-u})Y_1$, $\hat{B}_{-u} = (m - 3)/\sum_{i \neq u} Y_i^2$ and $v_{\text{J}} = [(m - 1)/m] \times \sum_{u=1}^m (\hat{B}_{-u} - \hat{B})^2$ denotes the jackknife variance estimator of $\hat{B}$.

The second term on the right-hand side of the above expression incorporates the extra variability due to the estimation of B. Proposition 5.3 states that mse^{J} has a bias of an order arbitrarily close to $O(m^{-3/2})$, which is lower than that of the naive estimator $b(\hat{B})$.

2.2. *The baseball example.* Efron and Morris (1975) considered a Bayesian model to predict the true 1970 season batting average of each of 18 major league baseball players using the data on batting averages based on first 45 official at bats. Their model can be obtained as a simple mixed linear normal model by adding an unknown μ term to the random effects model described in Section 2.1. The prediction of the true season batting average of player 1 is the same as that of the mixed effect: $\theta_1 = \mu + v_1$.

The BP of θ_1 is given by $\check{\theta}_1 = \pi_1(Y_1; \phi) = \mu + (1 - B)(Y_1 - \mu)$. Here $\phi = (\mu, B)$ can be estimated by $\hat{\phi} = (\bar{Y}, \hat{B})$, where $\bar{Y} = m^{-1}\sum_{i=1}^{m} Y_i$, and $\hat{B} = \min[(m-3)/(m-1), (m-3)/\sum_{i=1}^{m}(Y_i - \bar{Y})^2]$. Note that the cut-off point $(m-3)/(m-1)$ was suggested by Morris (1983). An EBP is given by $\hat{\theta} = \bar{Y} + (1 - \hat{B})(Y_1 - \bar{Y})$.

Note that MSE of the BP is given by $b(\phi) = (1 - B)$. Thus, a naive estimator of MSE of EBP is obtained by estimating $b(B)$ by $b(\hat{B})$. This underestimates the true uncertainty of EBP since it does not incorporate the variability due to the estimation of ϕ. Note that since the bias of $b(\hat{B})$ is of the order $o(m^{-1})$, jackknife bias correction of $b(\hat{\phi})$ is not needed and our proposed jackknife MSE estimator of $\hat{\theta}$ is given by:

$$\mathrm{mse}^{\mathrm{J}} = b(\hat{B}) + \frac{m-1}{m}\sum_{u=1}^{m}[\hat{\theta}_{-u} - \hat{\theta}]^2,$$

where $\hat{\theta}_{-u} = \bar{Y}_{-u} + (1 - \hat{B}_{-u})(Y_1 - \bar{Y}_{-u})$, $\bar{Y}_{-u} = (m-1)^{-1}\sum_{i\neq u}^{m} Y_i$ and $\hat{B}_{-u} = \min[(m-4)/(m-2), (m-4)/\sum_{i\neq u}^{m}(Y_i - \bar{Y}_{-u})^2]$.

The second term on the right-hand side of the above expression incorporates the extra variability due to the estimation of μ and B. Theorem 5.3 states that $\mathrm{mse}^{\mathrm{J}}$ has a bias of the same order as in Section 2.1.

2.3. *The baseball example continued.* Suppose in the above baseball data analysis, we are interested in comparing the true season averages of two players by taking the difference, say, $\theta = \theta_1 - \theta_2$. In this case, the BP of θ is given by $\check{\theta} = \pi_1(Y_1; \phi) - \pi_2(Y_2; \phi)$. An EBP of θ is given by $\hat{\theta} = \pi_1(Y_1; \hat{\phi}) - \pi_1(Y_1; \hat{\phi})$.

A naive estimator of MSE of EBP is obtained by estimating the MSE of the BP and is given by $\mathrm{mse}_{\mathrm{N}} = b(\hat{B}) + b(\hat{B}) = 2b(\hat{B})$, say. Our proposed jackknife MSE estimator is given by

$$\mathrm{mse}^{\mathrm{J}} = 2b(\hat{B}) + \frac{m-1}{m}\sum_{u=1}^{m}[\hat{\theta}_{-u} - \hat{\theta}]^2,$$

where $\hat{\theta}_{-u} = \pi_1(Y_1; \hat{\phi}_{-u}) - \pi_1(Y_2; \hat{\phi}_{-u})$. As a consequence of Theorem 5.3, we obtain the same order of bias as in Section 2.1 for $\mathrm{mse}^{\mathrm{J}}$.

2.4. *A simple nested error model.* In many applications such as small-area estimation and animal breeding, the following simple random effects model has been used:

$$Y_{ij} = \mu + v_i + e_{ij}, \qquad i = 1, \dots, m,\ j = 1, \dots, n,$$

where the v_i's are independent $N(0, \sigma_v^2)$, e_{ij}'s are independent $N(0, \sigma_e^2)$ and u and e are independent. Let $\phi = (\mu, \sigma_v^2, \sigma_e^2)$. Researchers in genetics and animal breeding are interested in the estimation of nonlinear functions of the variance components such as the intraclass correlation $g(\sigma_v^2, \sigma_e^2) = \sigma_v^2/(\sigma_v^2 + \sigma_e^2)$ or a similar ratio called heritability. Such functions are also of interest in psychological and educational testing.

The MOM estimators of σ_v^2 and σ_e^2, which are identical to a solution to the REML equations [e.g., Searle, Casella and McCulloch (1992), page 253], are given by $\hat{\sigma}_e^2 = [m(n-1)]^{-1} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i\cdot})^2$ and $\hat{\sigma}_v^2 = (\mathrm{MSA} - \hat{\sigma}_e^2)/n$, where $\mathrm{MSA} = [n/(m-1)] \sum_{i=1}^m (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$. One can then estimate $g(\sigma_v^2, \sigma_e^2)$ by $g(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$. Our jackknife can then be used to reduce the bias of this estimator to an order arbitrarily close to $O(m^{-3/2})$. See Theorem 5.2.

2.5. *Estimation of a finite population mean.* Let Y_{ij} and $x_{ij} = (x_{1ij}, \dots, x_{pij})'$ denote the values of a characteristic of interest and a vector of p covariates for the jth unit in the ith stratum of known size N_i $(i = 1, \dots, m,\ j = 1, \dots, N_i)$. Suppose a simple random sample of size n_i is drawn from the ith stratum and let $\{(Y_{ij}, x_{ij}),\ i = 1, \dots, m,\ j = 1, \dots, n_i\}$ denote the sample. We assume that x_{ij} values $(i = 1, \dots, m,\ j = 1, \dots, N_i)$ are known for the entire population. Ghosh and Meeden (1986) and Ghosh and Lahiri (1987), among others, considered empirical Bayes estimation of the finite population mean, $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$. Such empirical Bayes estimators of $\bar{Y}_i$ can be motivated from an EBLUP (same as EBP in such a situation) approach under the following mixed linear normal model for Y_{ij} [see Prasad and Rao (1990)]:

$$Y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \qquad i = 1, \dots, m,\ j = 1, \dots, N_i,$$

where v_i is a random effect due to the ith stratum and e_{ij} is the pure error which accounts for any unexplained variation not taken care of by the other terms of the above mixed model. It is often assumed that $\{v_i\}$ and $\{e_{ij}\}$ are independently distributed with $v_i \sim N(0, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. Let $\phi = (\beta, \sigma_v^2, \sigma_e^2)$ denote the unknown vector of model parameters.

A real-life application of the above model can be found in Battese, Harter and Fuller (1988), who considered EBLUP of areas under corn and soybeans for $m = 12$ counties in north central Iowa. In their example, Y_{ij} denotes the number of hectares under corn (soybeans) and $x_{ij} = (1, x_{1ij}, x_{2ij})'$ with x_{1ij} and x_{2ij} representing the number of pixels of corn and soybeans for the jth segment in the ith county. The sample was obtained from the *June Enumerative Survey*. The covariates x_{1ij} and x_{2ij} were obtained for the sample segments as well as for the nonsample segments using *Landsat* satellite data. We are interested in predicting average area under corn (soybeans) for each county.

The BP of $\bar{Y}_i$ is given by $\hat{\bar{Y}}_i(\bar{y}_i;\phi) = c_i\bar{y}_i + (1-c_i)\check{\theta}_i$, where $\check{\theta}_i = \hat{\theta}_i(\bar{y}_i;\phi) = \bar{x}_i^{*\prime}\beta + (1-B_i)(\bar{y}_i - \bar{x}_i^{\prime}\beta)$ is the BP of $\theta_i = \bar{x}_i^{*\prime}\beta + v_i$, $\bar{y}_i$ and $\bar{x}_i$ are the sample means of y and x for the ith stratum, $\bar{x}_i^*$ is the mean of x for nonsampled units, $c_i = n_i/N_i$ is the finite population correction factor and $B_i = \sigma_e^2/(\sigma_e^2 + n_i\sigma_v^2)$, $i = 1,\dots,m$. In this case, usually a weighted least square estimator with estimated variance components is used to estimate β, and MOM, ML or REML (see Section 6 for details) is used to estimate the variance components $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$. Let $\hat{\phi} = (\hat{\beta}, \hat{\sigma}_v^2, \hat{\sigma}_e^2)$ be an estimator of ϕ. Plugging in these estimators in the BP, we obtain the following EBP of $\bar{Y}_i$: $\hat{\bar{Y}}_i(\bar{y}_i;\hat{\phi}) = c_i\bar{y}_i + (1-c_i)\hat{\theta}_i$, where $\hat{\theta}_i = \hat{\theta}_i(\bar{y}_i;\hat{\phi})$. In this case, our jackknife estimator of MSE of $\hat{\theta}_i$ is given by

$$\begin{aligned}\mathrm{mse}_i^{\mathrm{J}} &= b_i(\hat{\phi}) - \frac{m-1}{m}\sum_{u=1}^{m}[b_i(\hat{\phi}_{-u}) - b_i(\hat{\phi})] \\ &\quad + \frac{m-1}{m}\sum_{u=1}^{m}\bigl[\hat{\theta}_i(\bar{y}_i;\hat{\phi}_{-u}) - \hat{\theta}_i(\bar{y}_i;\hat{\phi})\bigr]^2,\end{aligned}$$

where $\hat{\phi}_{-u}$ is obtained using the formula of $\hat{\phi}$, deleting the uth observation. In the above, the second term in the right-hand side reduces the bias of $b_i(\hat{\phi})$ (see Theorem 5.2) and the third term incorporates the extra variability due to the estimation of ϕ (see Theorem 5.1). The jackknife MSE estimator of $\hat{\bar{Y}}_i(\bar{y}_i;\hat{\phi})$ is then obtained as

$$\mathrm{mse}_i^{\mathrm{J}}[\hat{\bar{Y}}_i(\bar{y}_i;\hat{\phi})] = (1-c_i)^2\bigl[\mathrm{mse}_i^{\mathrm{J}} + N_i^{-1}(1-c_i)^{-1}\hat{\sigma}_e^2\bigr].$$

As a consequence of Theorem 5.3, we obtain the same order of bias as in Section 2.1 for $\mathrm{mse}_i^{\mathrm{J}}[\hat{\bar{Y}}_i(\bar{y}_i;\hat{\phi})]$.

We are also interested in obtaining the variance estimator of $h(\hat{\phi})$, a nonlinear function of $\hat{\phi}$. For example, $h(\hat{\phi}) = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2)$. Our jackknife variance estimator is given by

$$v_{\mathrm{J}} = \frac{m-1}{m}\sum_{u=1}^{m}[h(\hat{\phi}_{-u}) - h(\hat{\phi})]^2.$$

Theorem 5.1 can be applied to claim that v_{J} has a bias of an order arbitrarily close to $O(m^{-3/2})$. Note that, since MSE = variance + (bias)^2 = variance + $O(m^{-2})$ (see Section 1, third paragraph), the variance estimator is equivalent to the MSE estimator. Also note that a special case of $\pi(Y_S,\phi)$ is $\pi(Y_S,\phi) = h(\phi)$.

2.6. *A mixed logistic model.* Suppose that, conditional on p_{ij}, Y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$, are independent Bernoulli random variables with $\mathrm{P}(Y_{ij} = 1|p_{ij}) = p_{ij}$. Furthermore, suppose that, conditional on the random effects $\alpha_1, \dots, \alpha_m$,

$$\mathrm{logit}(p_{ij}) = x_{ij}^t \beta + \alpha_i, \tag{2.1}$$

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 $\mathrm{logit}(t) = \log[t/(1-t)]$. We assume that the α's are independent and distributed as $N(0, \sigma^2)$. Then the above is a special case of the generalized linear mixed model which has received considerable attention in recent years [e.g., Breslow and Clayton (1993) and Lee and Nelder (1996)]. Malec, Sedransk, Moriarity and Le Clere (1997) used a more general version of model (2.1) to estimate the state level proportion of individuals who visited a doctor's office at least once during the past 12 months.

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 $\{x_1, \dots, x_K\}$, $\theta = \sum_{k=1}^K w_k \mathrm{logit}^{-1}(x_k^t \beta + \alpha_i)$, where w_k, $1 \le k \le K$, is a set of weights and $\mathrm{logit}^{-1}(u) = e^u/(1+e^u)$.

Jiang and Lahiri (2001) derived the BP of θ as

$$\begin{aligned} \check{\theta} &= \mathrm{E}(\theta|Y) \\ &= \frac{\mathrm{E} h_i(\beta, \sigma\xi) \exp(\psi_i(Y_i., \sigma\xi, \beta))}{\mathrm{E} \exp(\psi_i(Y_i., \sigma\xi, \beta))} = \pi_i(Y_i., \phi), \end{aligned} \tag{2.2}$$

where $\psi_i(k, u, v) = ku - \sum_{j=1}^{n_i} \log[1 + \exp(x_{ij}^t v + u)]$, $Y_i. = \sum_{j=1}^{n_i} Y_{ij}$, $\phi = (\beta^t \sigma)^t$ and the expectations are taken over $\xi \sim N(0,1)$. Let $\hat{\phi}$ be the method of simulated moments estimator of ϕ (see Section 7 for details). An EBP of θ is then given by $\hat{\theta} = \pi_i(Y_i., \hat{\phi})$.

A naive estimator of MSE is given by $b_i(\hat{\phi})$, which has an order of bias $O(m^{-1/2})$. Here $b_i(\phi) \equiv \mathrm{MSE}(\check{\theta}) = \mathrm{E} h_i^2(\beta, \sigma\xi) - \sum_{k=0}^{n_i} \pi_i^2(k, \phi) p_i(k, \phi)$ and $p_i(k, \phi) = \mathrm{P}(Y_i. = k)$. Our jackknife MSE estimator of the EBP is given by

$$\begin{aligned} \mathrm{mse}_i^{\mathrm{J}} = b_i(\hat{\phi}) &- \frac{m-1}{m} \sum_{u=1}^m [b_i(\hat{\phi}_{-u}) - b_i(\hat{\phi})] \\ &+ \frac{m-1}{m} \sum_{u=1}^m [\pi_i(Y_i., \hat{\phi}_{-u}) - \pi_i(Y_i., \hat{\phi})]^2. \end{aligned}$$

Once again, as a consequence of Theorem 5.3 we obtain the same order of bias as in Section 2.1 for $\mathrm{mse}_i^{\mathrm{J}}$.

The jackknife variance estimator of $h(\hat{\phi})$, where $h(\cdot)$ is a smooth function, is given by

$$v_{\mathrm{J}} = \frac{m-1}{m} \sum_{u=1}^m [h(\hat{\phi}_{-u}) - h(\hat{\phi})]^2.$$

Again, Theorem 5.1 can be applied to obtain the same order of the bias as in Section 2.5 for v_{J}.

3. Notation. Let ϕ denote a vector of parameters associated with the distribution of the observations. Then ϕ_0, Φ and Φ^o represent the true vector of parameters, the parameter space and the interior of Φ, respectively. Also, ϕ^* denotes a given point in Φ, which may be a reasonable guess of ϕ_0 (but without using the data). If ϕ is s-dimensional ($s > 1$), then ϕ_j represents the jth component of ϕ, and similar notation applies to ϕ_0, etc.

In this paper, $\mathrm{E}(\cdot)$ means $\mathrm{E}_{\phi_0}(\cdot)$, that is, expectation taken at ϕ_0. Similarly, $\mathrm{var}(\cdot)$ means $\mathrm{var}_{\phi_0}(\cdot)$. When ϕ is a vector and $\hat{\phi}$ is an estimator of ϕ, the matrix MSE and (scale) MSE of $\hat{\phi}$ are defined, respectively, as $\mathcal{MSE}(\hat{\phi}) = \mathrm{E}(\hat{\phi} - \phi_0)(\hat{\phi} - \phi_0)'$ and $\mathrm{MSE}(\hat{\phi}) = \mathrm{E}(|\hat{\phi} - \phi_0|^2) = \mathrm{tr}[\mathcal{MSE}(\hat{\phi})]$. Let θ denote a random vector. Then, $\check{\theta}$ represents the BP of θ, that is, $\check{\theta} = \mathrm{E}(\theta|\text{data})$. If $\hat{\theta}$ is a predictor of θ, we define $\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(|\hat{\theta} - \theta|^2)$. For convenience, we also define $\mathrm{MSE}(\check{\theta}) = \mathrm{E}(|\check{\theta} - \theta|^2)$, even though $\check{\theta}$ may not be a predictor (because it may depend on unknown parameters). Similarly, we define the mean squared approximation error of $\hat{\theta}$ to $\check{\theta}$ as $\mathrm{MSAE}(\hat{\theta}) = \mathrm{E}(|\hat{\theta} - \check{\theta}|^2)$. As before, the latter has a matrix version; that is, $\mathcal{MSAE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \check{\theta})(\hat{\theta} - \check{\theta})'$, so that $\mathrm{MSAE}(\hat{\theta}) = \mathrm{tr}[\mathcal{MSAE}(\hat{\theta})]$.

Let A be a matrix. Then $A \geq 0$ means that A is nonnegative definite; and $\|A\| = \lambda_{\max}^{1/2}(A'A)$, where $\lambda_{\max}$ means largest eigenvalue. If $A = (a_{i_1,\dots,i_q})$ is a q-way array ($q \geq 3$), then $\|A\| = \max_{i_1,\dots,i_q} |a_{i_1,\dots,i_q}|$. If S is a set, then $|S|$ denotes the cardinality of S.

If $f(\phi, \xi)$ depends both on ϕ and on a random vector ξ, we simply write f for $f(\phi_0, \xi)$. If Y_j is a random variable, we shall always write

$$f_j = f_j(\phi; Y_j), \qquad g_j = (\partial/\partial\phi) f_j(\phi; Y_j), \qquad h_{j,k} = (\partial^2/\partial\phi^2) f_{j,k}(\phi; Y_j),$$

where $\phi = \phi_0$. Also,

$$f. = \sum_{j=1}^{m} f_j, \qquad f._{-i} = \sum_{j \neq i} f_j, \qquad \bar{f} = f./m,$$

$$\partial^3 f/\partial\phi^3 = (\partial^3 f/\partial\phi_i\,\partial\phi_j\,\partial\phi_k)_{i,j,k},$$

etc. For any $w > 0$, we define

$$|f|_w = \sup_{|\phi - \phi_0| \leq w} |f(\phi, \xi)|.$$

Similarly, if f and g are two functions of ϕ, we define

$$|f - g|_{w,w} = \sup_{|u - \phi_0| \leq w,\ |v - \phi_0| \leq w} |f(u) - g(v)|.$$

The definitions extend to $\|\cdot\|$ in a natural way.

Some functions a and a_{-i} will be introduced in Section 4.2, for which we define $A_{d,w,r}$, $r = 1, \dots, 4$, to be the following sets of quantities:

1. $A_{d,w,1}$: $m^{-1}|a_{-i}|, m^{-1}\|\partial a_{-i}/\partial\phi\|, m^{-1}\|\partial^2 a_{-i,k}/\partial\phi^2\|_w, 1 \le k \le s, 0 \le i \le m$;
2. $A_{d,w,2}$: $m^{-1/(2d+1)}\|\partial a/\partial\phi\|_w, m^{-1/(2d+1)}|a_{-i}|_w, m^{(d-1)/(2d+1)}|a - a_{-i}|_w$, $|a - a_{-i}|_{w,w}, 0 \le i \le m$;
3. $A_{d,w,3}$: $|\cdot|_w(\|\cdot\|_w)$ of a_{-i} and its up to third derivatives; $|\Delta_i|_w$, $m^\delta|\Delta_i|$ and $m^\delta\|\partial\Delta_i/\partial\phi\|$, $1 \le i \le m$, where $\delta = (d-2)/(2d+1)$ $(d > 2)$ and $\Delta_i = a - a_{-i}$;
4. $A_{d,w,4}$: $|\cdot|_w(\|\cdot\|_w)$ of a_{-i} and its up to third derivatives; $m^{(d-1)/(2d+1)}|\Delta_i|_w$ and $m^{(d-2)/(2d+1)}\|\partial\Delta_i/\partial\phi\|$, $1 \le i \le m$.

Finally, we use c to denote a constant whose value may be different at different places.

4. Jackknifing *M*-estimators.

For simplicity of exposition, we first state results for a simple model and then for a very general case.

4.1. *A simple case.*

Let ϕ be one-dimensional (e.g., Section 2.1). Let $Y_1, \dots, Y_m$ be independent observations with the same distribution as Y, which depends on ϕ. An M-estimator of ϕ is associated with the solution $\dot\phi$ to the equation

$$F(\phi) = \sum_{j=1}^m f(\phi, Y_j) = 0,$$

where $f(\phi, Y_j)$ is a function satisfying $\mathrm{E}[f(\phi_0, Y)] = 0$. In general, $\dot\phi$ may not always exist or, even if it does, may fall outside Φ. Of course, the $\mathrm{MSE}(\dot\phi)$ also may not exist. Therefore, we consider the following *truncated* version of $\dot\phi$. Let $\hat\phi = \dot\phi$ if $\dot\phi$ exists, lies in Φ and $|\dot\phi| \le K(\log m)^\alpha$; and let $\hat\phi = \phi^*$ otherwise, where K and α are (known) constants such that $K > 0$, $\alpha \ge 0$ and $|\phi^*| \le K(\log m)^\alpha$.

REMARK 4.1. One may wonder why $|\dot\phi|$ is truncated by a term of the order $(\log m)^\alpha$. First such a truncation will have no impact on the main theoretical results proved in this paper; that is, the order of the asymptotic bias of the jackknife estimator of the MSE remains the same irrespective of the value of α. Also, such a truncation is not restrictive. To see this, consider the simple random effects model given in Example 2.4 (although in this case ϕ is not one-dimensional). Suppose that $m \to \infty$ while n remains fixed. Then it can be shown that

$$\mathrm{P}(\hat\sigma_e^2 > \log m) \le \exp\left(-\frac{\sqrt{\log m}}{2}m\right), \qquad \mathrm{P}(\hat\sigma_v^2 > \log m) \le \exp\left(-\frac{\sqrt{\log m}}{2}m\right).$$

It is customary to truncate the ANOVA estimators of variances at 0 when they are negative. It can be shown that $\mathrm{P}(\hat\sigma_v^2 < 0) = O[\exp(-am)]$ for some constant

$a > 0$. Thus, asymptotically, the chance of $\hat{\sigma}_e^2$ or $\hat{\sigma}_v^2$ exceeding $\log m$ is much smaller than that of $\hat{\sigma}_v^2$ being negative.

We define a delete-i estimator, $\hat{\phi}_{-i}$, which is associated with a solution $\dot{\phi}_{-i}$ to the equation

$$F_{-i}(\phi) = \sum_{j \neq i} f(\phi, Y_j) = 0. \tag{4.1}$$

For convenience, we write $F_{-0}(\phi) = F(\phi)$ and $\hat{\phi}_{-0} = \hat{\phi}$.

Not surprisingly, the asymptotic behavior of the M-estimators is critical to the theoretical development in the next section, in which the following property of an estimator plays an important role.

DEFINITION 4.1. The M-estimators $\hat{\phi}_{-i}$, $0 \le i \le m$, are said to be consistent uniformly (c.u.) at the rate m^{-d} if, for any $b > 0$, there is a constant c (which may depend on b) such that

$$\mathrm{P}(A_{i,b}^c) \le cm^{-d},$$

$0 \le i \le m$, where $A_{i,b} = \{F_{-i}(\hat{\phi}_{-i}) = 0$, and $|\hat{\phi}_{-i} - \phi_0| \le b\}$, and $\mathrm{P}(\cdot)$ means $\mathrm{P}_{\phi_0}(\cdot)$.

Propositions 4.1 and 4.2 give sufficient conditions for the M-estimators to be c.u. at rate m^{-d}. First define the following.

DEFINITION 4.2. The M-estimating equations are said to be standard if $f(\phi, Y) = (\partial/\partial\phi) l(\phi, Y)$ for some function $l(\phi, Y)$ three times continuously differentiable with respect to ϕ and satisfying

$$\mathrm{E}\left(\frac{\partial^2}{\partial\phi^2} l(\phi_0, Y)\right) > 0.$$

REMARK 4.2. A standard M-estimating equation is similar to one considered by Huber [(1981), Section 3.2], while a nonstandard one may be regarded as an extension. Examples of both standard and nonstandard M-estimating equations will be considered in Sections 6 and 7.

PROPOSITION 4.1. *Suppose that the M-estimating equations are standard and that*

$$\begin{aligned} &\mathrm{E}\left(\left|\frac{\partial^r}{\partial\phi^r} l(\phi_0, Y)\right|^{2d}\right) < \infty, \qquad r = 1, 2, \quad \textit{and} \\ &\mathrm{E}\left(\sup_{|\phi - \phi_0| \le b_0} \left|\frac{\partial^3}{\partial\phi^3} l(\phi, Y)\right|^{2d}\right) < \infty \end{aligned} \tag{4.2}$$

for some $d \geq 1$ and $b_0 > 0$. Then there are M-estimators $\hat{\phi}_{-i}$, $0 \leq i \leq m$, which are c.u. at rate m^{-d}.

REMARK 4.3. Consider Section 2.1. It is easy to see that the ML equation for estimating A is equivalent to $\sum_{j=1}^{m}(\partial/\partial A)l(A, Y_j) = 0$, where $l(A, Y) = \log(A + 1) + Y^2/(A + 1)$. Clearly, the M-estimating equation is standard, and it is straightforward to show that all the conditions of Proposition 4.1 are satisfied ($d \geq 1$).

We now consider cases where the M-estimating equations may not be standard. It is more convenient to consider the following generalized M-estimator: Let $\hat{\phi}_{-i}$ be any ϕ that minimizes $|F_{-i}(\phi)|$, if such a minimizer exists and is in Φ; otherwise, define $\hat{\phi}_{-i} = \phi^*$. Similar definitions have been adapted in the literature, for example, for the generalized method of moments estimators [e.g., McFadden (1989)]. It is clear that, if the solution to (4.1) exists and lies in Φ, then $\hat{\phi}_{-i}$ satisfies (4.1).

PROPOSITION 4.2. *Suppose that $f(\phi, Y) = t(Y) - M(\phi)$ for some function $t(\cdot)$, where $M(\phi) = \mathrm{E}_\phi t(Y)$. Suppose that $M'(\phi)$ is continuous and nonzero, and that there is $d \geq 1$ such that*

$$\mathrm{E}(|t(Y)|^{2d}) < \infty.$$

Then the following hold:

(a) *For any $b > 0$ there is a constant c (which may depend on b) such that*

$$\mathrm{P}(|\hat{\phi}_{-i} - \phi_0| > b) \leq cm^{-d}, \qquad 0 \leq i \leq m.$$

(b) *Let R_t be the range of $t(Y)$. If, in addition,*

$$\forall r \in R_t, \qquad \exists \phi \in \Phi \text{ such that } M(\phi) = r,$$

then $\hat{\phi}_{-i}$, $0 \leq i \leq m$, are c.u. at rate m^{-d}.

Now define the jackknife estimator of $\mathrm{MSE}(\hat{\phi})$ as follows:

$$\widehat{\mathrm{MSE}}(\hat{\phi}) = \frac{m-1}{m}\sum_{i=1}^{m}(\hat{\phi}_{-i} - \hat{\phi})^2.$$

PROPOSITION 4.3. *Suppose that* (i)

$$\mathrm{E}\left(\frac{\partial}{\partial \phi} f(\phi_0, Y)\right) \neq 0, \tag{4.3}$$

and there are $d > 2$ and $b_0 > 0$ such that

$$\begin{aligned}
&\mathrm{E}\big(|f(\phi_0, Y)|^{2d}\big) \vee \mathrm{E}\left(\left|\frac{\partial}{\partial\phi} f(\phi_0, Y)\right|^{2d}\right) \\
&\qquad \vee \mathrm{E}\left(\sup_{|\phi-\phi_0|\le b_0}\left|\frac{\partial^2}{\partial\phi^2} f(\phi, Y)\right|^{2d}\right) < \infty, \qquad (4.4)\\
&\qquad \mathrm{E}\left(\sup_{|\phi-\phi_0|\le b_0}\left[\frac{\partial^3}{\partial\phi^3} f(\phi, Y)\right]^2\right) < \infty,
\end{aligned}$$

where $a \vee b$ is the maximum of a and b; and (ii) *$\hat{\phi}_{-i}$, $0 \le i \le m$, are c.u. at rate m^{-d}. Then*

$$\mathrm{E}\big[\widehat{\mathrm{MSE}}(\hat{\phi})\big] - \mathrm{MSE}(\hat{\phi}) = o(m^{-1-\varepsilon}),$$

for any $0 < \varepsilon < (d-2) \wedge [(d-1)/(2d+1)]$.

REMARK 4.4. Since $\mathrm{MSE}(\hat{\phi}) = O(m^{-1})$, the jackknife estimator is asymptotically unbiased.

REMARK 4.5. We now show that all the conditions of Proposition 4.3 are satisfied in the case of Section 2.1. Note that here $f(A, Y) = (\partial/\partial A)l(A, Y) = (A+1)^{-1} - Y^2/(A+1)^2$. Thus,

$$\mathrm{E}\left[\frac{\partial}{\partial A} f(A_0, Y)\right] = \frac{1}{(A_0+1)^2} > 0. \qquad (4.5)$$

Also, we have

$$\mathrm{E}|f(A_0, Y)|^{2d} \le c\left[\frac{1}{(A_0+1)^{2d}} + \frac{\mathrm{E}|Y|^{4d}}{(A_0+1)^{4d}}\right] < \infty.$$

Similarly, we have

$$\begin{aligned}
\mathrm{E}\left|\frac{\partial}{\partial A} f(A_0, Y)\right|^{2d} &\le c\left[\frac{1}{(A_0+1)^{4d}} + \frac{\mathrm{E}|Y|^{4d}}{(A_0+1)^{6d}}\right] < \infty,\\
\mathrm{E}\left[\sup_{A>0}\left|\frac{\partial^2}{\partial A^2} f(A, Y)\right|^{2d}\right] &\le c[1 + \mathrm{E}|Y|^{4d}] < \infty,\\
\mathrm{E}\left\{\sup_{A>0}\left[\frac{\partial^3}{\partial A^3} f(A, Y)\right]^2\right\} &\le c\left(1 + \mathrm{E}Y^4\right) < \infty.
\end{aligned}$$

It remains to show that $\hat{A}_{-i}$, $0 \le i \le m$, are c.u. at rate m^{-d}, but this follows immediately from Proposition 4.1, because the M-estimating equations are clearly standard [see (4.5)], and we have shown that (4.2) is satisfied. Note that each of the M-estimating equations has a unique solution.

4.2. *General results.* Now suppose that ϕ is s-dimensional ($s \geq 1$). See Examples 2.2–2.6. Let $Y_1, \dots, Y_m$ be independent (vector-valued) observations whose joint distribution depends on ϕ. We are interested in an M-estimator of ϕ, which is associated with a solution $\dot{\phi} = (\dot{\phi}_k)_{1\leq k\leq s}$ to the following equation:

$$F(\phi) = \sum_{j=1}^{m} f_j(\phi; Y_j) + a(\phi) = 0. \tag{4.6}$$

In the above, $f_j(\phi, Y_j) = (f_{j,k}(\phi, Y_j))_{1\leq k\leq s}$ are vector-valued functions such that $\mathrm{E}[f_j(\phi_0, Y_j)] = 0$, $1 \leq j \leq m$, and $a(\phi)$ is a vector-valued function which may depend on the joint distribution of $(Y_j)_{1\leq j\leq m}$. When $a(\phi) \neq 0$, it plays the role of a modifier or penalizer. We assume that $\phi_0 \in \Phi^o$. An M-estimator of ϕ, $\hat{\phi}$, is then defined based on $\dot{\phi}$ in the same way as in Section 4.1. Similarly, a delete-i estimator of ϕ, $\hat{\phi}_{-i}$, is defined based on a solution $\dot{\phi}_{-i}$ to

$$F_{-i}(\phi) = \sum_{j\neq i} f_j(\phi; Y_j) + a_{-i}(\phi) = 0.$$

Again, for convenience we write $F_{-0}(\phi) = F(\phi)$, $a_{-0}(\phi) = a(\phi)$ and $\hat{\phi}_{-0} = \hat{\phi}$.

The jackknife estimators of the $\mathcal{MSE}$ and MSE of $\hat{\phi}$ are defined as

$$\widehat{\mathcal{MSE}}(\hat{\phi}) = \frac{m-1}{m} \sum_{i=1}^{m} (\hat{\phi}_{-i} - \hat{\phi})(\hat{\phi}_{-i} - \hat{\phi})',$$

$$\widehat{\mathrm{MSE}}(\hat{\phi}) = \mathrm{tr}\bigl[\widehat{\mathcal{MSE}}(\hat{\phi})\bigr] = \frac{m-1}{m} \sum_{i=1}^{m} |\hat{\phi}_{-i} - \hat{\phi}|^2.$$

Definition 4.1 is adopted here without any change. We now give an extension of Definition 4.2.

DEFINITION 4.3. The M-estimating equations are said to be standard if $F_{-i}(\phi) = \partial l_{-i}/\partial\phi$ for some functions $l_{-i}(\phi; Y_{-i})$ three times continuously differentiable with respect to ϕ, $0 \leq i \leq m$, and $\mathrm{E}(\bar{g}) \geq 0$.

We now give analogues of Propositions 4.1, 4.2 and 4.3.

THEOREM 4.1. *Suppose that* (i) *the M-estimating equations are standard*; (ii) *there are $d \geq 1$ and $w > 0$ such that the $2d$th moments of $|f_j|$, $\|g_j\|$, $\|h_{j,k}\|_w$, $1 \leq k \leq s$, $1 \leq j \leq m$, are bounded*; (iii)

$$\limsup_{m\to\infty} \|(\mathrm{E}\bar{g})^{-1}\| < \infty; \tag{4.7}$$

and (iv) *the quantities in $A_{d,w,1}$ converge uniformly to* 0. *Then there are M-estimators $\hat{\phi}_{-i}$, $0 \leq i \leq m$, which are c.u. at rate m^{-d}.*

As in Section 4.1, we define the generalized M-estimator, $\hat{\phi}_{-i}$, as any ϕ that minimizes $|F_{-i}(\phi)|$, if such a minimizer exists and is in Φ; otherwise, define $\hat{\phi}_{-i} = \phi^*$.

THEOREM 4.2. *Suppose that* $f_j(\phi, Y_j) = t_j(Y_j) - \mathrm{E}_\phi t_j(Y_j)$ *for some function* $t_j(\cdot)$, $1 \le j \le m$, *and* $a_{-i}(\phi) = 0$, $0 \le i \le m$. *Write* $M(\phi) = m^{-1} \sum_{j=1}^m \mathrm{E}_\phi t_j(Y_j)$. *Suppose that there are* $B, \varepsilon_1, \varepsilon_2 > 0$ *such that*

$$\liminf\left(\inf_{\phi \notin \Phi_B} |M(\phi) - M(\phi_0)|\right) \ge \varepsilon_1, \tag{4.8}$$

$$\liminf\left(\inf_{\phi \in \Phi_B, \phi \ne \phi_0} \frac{|M(\phi) - M(\phi_0)|}{|\phi - \phi_0|}\right) \ge \varepsilon_2, \tag{4.9}$$

where $\Phi_B = \{\phi \in \Phi : |\phi| \le B\}$, *and that there is* $d \ge 1$ *such that*

$$\mathrm{E}(|t_j(Y_j)|^{2d}), \qquad 1 \le j \le m \text{ are bounded.} \tag{4.10}$$

Then the following hold:

(a) *For any* $b > 0$ *there is a constant* c *which may depend on* b *such that*

$$\mathrm{P}(|\hat{\phi}_{-i} - \phi_0| > b) \le cm^{-d}, \qquad 0 \le i \le m.$$

(b) *Let* R_t *be the range of* $m^{-1} \sum_{j=1}^m t_j(Y_j)$. *If, in addition,*

$$\forall r \in R_t, \qquad \exists \phi \in \Phi \text{ such that } M(\phi) = r, \tag{4.11}$$

then $\hat{\phi}_{-i}$, $0 \le i \le m$, *are c.u. at rate* m^{-d}.

THEOREM 4.3. *Suppose that* (i) (4.7) *holds, and there are* $d > 2$ *and* $w > 0$ *such that the* $2d$*th moments of* $|f_j|$, $\|g_j\|$, $\|h_{j,k}\|_w$ *and the second moments of* $\|\partial^3 f_{j,k}/\partial\phi^3\|_w$, *where* $1 \le j \le m$, $1 \le k \le s$, *are bounded*; (ii) *the quantities in* $A_{d,w,2}$ *are bounded*; *and* (iii) $\hat{\phi}_{-i}$, $0 \le i \le m$, *are c.u. at rate* m^{-d}. *Then*

$$\mathrm{E}\big[\widehat{\mathcal{MSE}}(\hat{\phi})\big] - \mathcal{MSE}(\hat{\phi}) = o(m^{-1-\varepsilon})$$

for any $0 < \varepsilon < (d-2) \wedge [(d-1)/(2d+1)]$, *and hence for the same range of* ε,

$$\mathrm{E}\big[\widehat{\mathrm{MSE}}(\hat{\phi})\big] - \mathrm{MSE}(\hat{\phi}) = o(m^{-1-\varepsilon}).$$

REMARK 4.6. Again, because $\mathrm{MSE}(\hat{\phi}) = O(m^{-1})$, the jackknife estimator is asymptotically unbiased.

5. Jackknifing MSE of EBP.

We first motivate our proposed jackknife MSE estimator of EBP using the following simple example.

5.1. *A simple case.* Suppose that one is interested in predicting a univariate random variable θ based on i.i.d. observations $Y_1, \dots, Y_m$, and it is known that the BP has the form

$$\check{\theta} = \pi(Y_1, \phi),$$

where ϕ is one-dimensional. The EBP and delete-i EBP of θ are, respectively,

$$\hat{\theta} = \pi(Y_1, \hat{\phi}),$$
$$\hat{\theta}_{-i} = \pi(Y_1, \hat{\phi}_{-i}),$$

where $\hat{\phi}$, $\hat{\phi}_{-i}$, $1 \le i \le m$, are the M-estimators given in Section 4.1.

For example, consider Section 2.1. It is easy to show that the BP is in the form $\pi(Y_1, A)$, namely, $\check{\theta} = \mathrm{E}(v_1 \mid Y_1, \dots, Y_n) = [A/(A + D)]Y_1$, and the EBP is $\hat{\theta} = \pi(Y_1, \hat{A}) = [\hat{A}/(\hat{A} + D)]Y_1$, where $\hat{A}$ is an M-estimator of A.

We now define the jackknife estimator of $\mathrm{MSE}(\hat{\theta})$. First, we have the following decomposition:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{MSAE}(\hat{\theta}) + \mathrm{MSE}(\check{\theta}). \tag{5.1}$$

A jackknife estimator of the first term on the right-hand side of (5.1) is given by

$$\widehat{\mathrm{MSAE}}(\hat{\theta}) = \frac{m-1}{m} \sum_{i=1}^{m} (\hat{\theta}_{-i} - \hat{\theta})^2.$$

As for the second term, it is often possible to obtain a closed-form expression, which is a function of ϕ. Suppose that $\mathrm{MSE}(\check{\theta}) = b(\phi)$. Then a jackknife estimator of $b(\phi)$ is given by

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

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

$$\widehat{\mathrm{MSE}}(\hat{\theta}) = \widehat{\mathrm{MSAE}}(\hat{\theta}) + \widehat{\mathrm{MSE}}(\check{\theta}). \tag{5.2}$$

REMARK 5.1. The definition of $\hat{\theta}_{-i}$ is different from that of a traditional jackknife estimator (of a parameter) in the sense that Y_1 stays the same for all i. Some might wonder why one cannot use the traditional definition, that is, let $\hat{\theta} = \tau(Y)$, where $Y = (Y_j)_{1 \le j \le m}$, and define $\hat{\theta}^*_{-i} = \tau(Y_{-i})$, where $Y_{-i} = (Y_j)_{j \ne i}$. To see the reason, consider, once again, Section 2.1. It is easy to see that $\check{\theta}_{-1} = 0$, and $\check{\theta}_{-i} = (1 - B)y_1$, if $i \ge 2$. Also, one has $\dot{A} = m^{-1} \sum_{j=1}^{m} Y_j^2 - 1$, $\dot{A}_{-i} = (m-1)^{-1} \sum_{j \ne i} Y_j^2 - 1$, $i \ge 1$. It can be shown that

$$\mathrm{MSE}(\hat{\theta}) = (1 - B) + \frac{m-1}{m^2} B^3 \operatorname{var}(Y_1^2) + O(m^{-3/2}).$$

Now, suppose that one defines the “delete-i EBP” in the “traditional way,” that is, $\hat{\theta}^*_{-1} = 0$, and

$$\hat{\theta}^*_{-i} = (1 - \hat{B}_{-i})Y_1, \qquad i \geq 2,$$

where $\hat{B}_{-i} = 1/(1 + \hat{A}_{-i})$. Then it can be shown that

$$\mathrm{E}(\hat{\theta}^*_{-1} - \hat{\theta})^2 = \mathrm{E}(\hat{\theta}^2) = A(1 - B) + o(1),$$
$$\mathrm{E}(\hat{\theta}^*_{-i} - \hat{\theta})^2 = \frac{1}{m^2} B^3 \operatorname{var}(Y_1^2) + O(m^{-5/2}), \qquad i \geq 2.$$

Thus, if one defines the “jackknife estimator” of $\mathrm{MSE}(\hat{\theta})$, say, $\widehat{\mathrm{MSE}}^*(\hat{\theta})$, the same way as $\widehat{\mathrm{MSE}}(\hat{\theta})$ [see (5.2)], one has $\mathrm{E}[\widehat{\mathrm{MSE}}^*(\hat{\theta})] - \mathrm{MSE}(\hat{\theta}) = A(1 - B) + o(1)$, which does not even go to 0 as $m \to \infty$.

In the following, we assume that

$$|\pi(Y_1, \phi)| \leq \omega(Y_1)(1 \vee |\phi|^\lambda)$$

for some constant $\lambda > 0$ and measurable function $\omega(\cdot)$ such that $\omega(\cdot) \geq 1$. Recall that, in Section 4, we define M-estimators which are c.u. at rate m^{-d}. We now generalize the concept. Let $\mathscr{B} = \sigma(Y_1, \dots, Y_m)$, the σ-fields generated by the Y_i’s. Define a measure μ_ω on $\mathscr{B}$:

$$\mu_\omega(B) = \mathrm{E}\big(\omega^2(Y_1)\mathbb{1}_B\big), \qquad B \in \mathscr{B}. \tag{5.3}$$

DEFINITION 5.1. The M-estimators $\hat{\phi}_{-i}$, $0 \leq i \leq m$, are said to be consistent uniformly with respect to μ_ω (c.u. μ_ω) at rate m^{-d}, if for any $b > 0$, there is a constant c which may depend on b such that

$$\mu_\omega(A^c_{i,b}) \leq cm^{-d},$$

$0 \leq i \leq m$, where $A_{i,b}$ is the same as in Definition 4.1.

REMARK 5.2. Because $\omega(\cdot) \geq 1$, c.u. μ_ω at rate m^{-d} implies c.u. at rate m^{-d}. On the other hand, if there is $\tau > 2$ such that $\mathrm{E}(|\omega(Y_1)|^\tau)$ is bounded, then, by Hölder’s inequality, c.u. at rate m^{-d} implies c.u. μ_ω at rate $m^{-d(1-2/\tau)}$; in particular, if $\mathrm{E}(\omega(Y_1)^4)$ is bounded, then c.u. at rate m^{-2d} implies c.u. μ_ω at rate m^{-d}. Let Y be a random variable with the same distribution as the Y_i’s.

PROPOSITION 5.1. *Suppose that the following hold*:

(i) (4.3) *holds and there are* $d > 2$, $b_0 > 0$ *such that* (4.4) *holds, and*

$$\mathrm{E}\left(\sup_{|\phi - \phi_0| \leq b_0} \left[\frac{\partial^3}{\partial \phi^3} f(\phi, Y)\right]^4\right) < \infty;$$

(ii) *there is* $\tau \geq 4$ *such that* $\mathrm{E}(|\omega(Y)|^{\tau}) < \infty$, *and*

$$\mathrm{E}\left(\left[\frac{\partial^2}{\partial\phi^2}\pi(Y,\phi_0)\right]^4\right) \vee \mathrm{E}\left(\sup_{|\phi-\phi_0|\leq b_0}\left[\frac{\partial^3}{\partial\phi^3}\pi(Y,\phi)\right]^2\right)$$
$$\vee\, \mathrm{E}\left(\sup_{|\phi-\phi_0|\leq b_0}\left|\frac{\partial}{\partial\phi}\pi(Y,\phi)\right|^{2d}\right) < \infty;$$

(iii) $\hat{\phi}_{-i}$, $0 \leq i \leq m$, *are c.u.* μ_ω *at rate* m^{-d}.

Then

$$\mathrm{E}\big[\widehat{\mathrm{MSAE}}(\hat{\theta})\big] - \mathrm{MSAE}(\hat{\theta}) = o(m^{-1-\varepsilon}) \tag{5.4}$$

for any $0 < \varepsilon < (d-2)/(2d-1)$.

REMARK 5.3. Note that ε in (5.4) approaches $1/2$ as $d \to \infty$, but ε is always less than $1/2$. One may wonder if the order in (5.4) can be improved to $O(m^{-3/2})$, as in Section 2.1 [note that in this special case it can be shown that the order is exactly $O(m^{-3/2})$; see Remark 5.1]. However, in Section 2.1 we have used the following special properties: (1) normality; (2) the BP has the form $\check{\theta} = g(A)Y_1$, where $|g'''(A)|$ is uniformly bounded; and (3) the M-estimating equation has a closed-form solution which can be expressed as $\dot{A} = m^{-1}\sum_{j=1}^{m} h(Y_j)$, where $h(Y_j)$ has mean A and bounded moments of any order. These conditions, especially (1) and (3), do not always hold in practice. Therefore, we do not make such assumptions in Proposition 5.1 (and similarly in other propositions and theorems). As a consequence, the order in (5.4) is not $O(m^{-3/2})$.

REMARK 5.4. In practice, ω may be chosen in the following way. Suppose that there is a positive number λ such that $\sup_\phi\{|\pi(y,\phi)|/(1 \vee |\phi|^\lambda)\} < \infty$ for every y. Then let

$$\omega(y) = \sup_\phi\big\{|\pi(y,\phi)|/(1 \vee |\phi|^\lambda)\big\}.$$

Once ω is chosen, μ_ω is determined by (5.3). Remark 5.2 is useful in checking whether the c.u. μ_ω property holds because, under a suitable moment condition, it reduces to checking the c.u. property.

PROPOSITION 5.2. *Suppose that* (i) (4.3) *holds*, *and there are* $d > 2$ *and* $b_0 > 0$ *such that*

$$\mathrm{E}(|f(\phi_0,Y)|^{2d}) \vee \left\{\max_{r=1,2}\mathrm{E}\left(\left|\frac{\partial^r}{\partial\phi^r}f(\phi_0,Y)\right|^{2d}\right)\right\}$$
$$\vee\, \mathrm{E}\left(\sup_{|\phi-\phi_0|\leq b_0}\left|\frac{\partial^3}{\partial\phi^3}f(\phi,Y)\right|^{2d}\right) < \infty;$$

(ii) $\sup_{|\phi-\phi_0|\le b_0}|b^{(4)}(\phi)|$ *is bounded*; *and* (iii) $\hat{\phi}_{-i}$, $0 \le i \le m$, *are c.u. at rate* m^{-d}. *Then*

$$\mathrm{E}\big[\widehat{\mathrm{MSE}}(\check{\theta})\big] - \mathrm{MSE}(\check{\theta}) = \mathrm{E}\big[\widehat{b(\phi)} - b(\phi)\big] = o(m^{-1-\varepsilon})$$

for any $0 < \varepsilon < (d-2)/(2d+1)$.

PROPOSITION 5.3. *Suppose that Proposition* 5.2 (i) *and* (ii) *and Proposition* 5.1 (ii) *and* (iii) *hold*. *Then*

$$\mathrm{E}\big[\widehat{\mathrm{MSE}}(\hat{\theta})\big] - \mathrm{MSE}(\hat{\theta}) = o(m^{-1-\varepsilon})$$

for any $0 < \varepsilon < (d-2)/(2d+1)$.

Note that, by Propositions 4.1 and 4.2 and Remark 5.2, sufficient conditions for c.u. μ_ω can be easily obtained.

REMARK 5.5. Refer to Section 2.1. Note that $\pi(Y, A) = [1 - B]Y$ so that $|\pi(Y, A)| \le |Y|$ and $b(B) = 1 - B$. Combining with discussions in the previous section, it is straightforward to verify that all the conditions of Propositions 5.1 and 5.2 are satisfied, for any $d > 2$. Therefore, the conclusion of Proposition 5.3 holds for any $0 < \varepsilon < 1/2$.

5.2. *General results.* In general, we are interested in predicting an unobservable random vector $\theta = (\theta_l)_{1\le l\le t}$. The prediction will be based on independent (vector-valued) observations $Y_1, \dots, Y_m$, whose distributions depend on $\phi = (\phi_k)_{1\le k\le s}$. Suppose that, when ϕ is known, the BP is

$$\begin{aligned}\check{\theta} &= \mathrm{E}(\theta \mid Y_1, \dots, Y_m)\\ &= \pi(Y_S, \phi) = \big(\pi_l(Y_S, \phi)\big)_{1\le l\le t},\end{aligned}$$

where S is a subset of $\{1, \dots, m\}$ and $Y_S = (Y_j)_{j\in S}$. Let $\hat{\phi}$, $\hat{\phi}_{-i}$, $1 \le i \le m$, be the M-estimators given in Section 4.2. Then EBP and delete-i EBP are given by

$$\begin{aligned}\hat{\theta} &= \pi(Y_s, \hat{\phi}),\\ \hat{\theta}_{-i} &= \pi(Y_s, \hat{\phi}_{-i}), \qquad 1 \le i \le m.\end{aligned}$$

Note that Y_S is kept the same for all $\hat{\theta}_{-i}$'s (i.e., not affected by deleting the ith observation, see Remark 5.1). Again, we have the decomposition of the MSE:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{MSAE}(\hat{\theta}) + \mathrm{MSE}(\check{\theta}). \tag{5.5}$$

A jackknife estimator of the first term on the right-hand side of (5.5) is given by

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

As for the second term, it is often possible to obtain a closed-form expression as a function of ϕ. Suppose that $\mathrm{MSE}(\check{\theta}) = b(\phi)$. Then a jackknife estimator of $b(\phi)$ is given by

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

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

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

The jackknife estimator of $\mathcal{MSAE}(\hat{\theta})$ is defined as

$$\widehat{\mathcal{MSAE}}(\hat{\theta}) = \frac{m-1}{m}\sum_{i=1}^{m}(\hat{\theta}_{-i} - \hat{\theta})(\hat{\theta}_{-i} - \hat{\theta})'.$$

We assume that

$$|\pi(Y_S, \phi)| \le \omega(Y_S)(1 \vee |\phi|^{\lambda})$$

for some constant $\lambda > 0$ and measurable function $\omega(\cdot)$ such that $\omega(\cdot) \ge 1$. Similar to Section 5.1, we define a measure μ_ω on $\mathcal{B} = \sigma(Y_1, \dots, Y_m)$ by

$$\mu_\omega(B) = \mathrm{E}\big(\omega^2(Y_S)\mathbb{1}_B\big), \qquad B \in \mathcal{B}.$$

Definition 5.1 can be adapted here without any change. Note that $\hat{\phi}_{-i}$, $0 \le i \le m$, are now the more general M-estimators defined in Section 4.2. Similarly, we have the connection between c.u. and c.u. μ_ω (see Remark 5.2).

We now give analogues of Propositions 5.1–5.3.

THEOREM 5.1. *Suppose that the following hold*:

(i) (4.7) *holds, and there are* $d > 2$ *and* $w > 0$ *such that the* $2d$*th moments of* $|f_j|$, $\|g_j\|$, $\|h_{j,k}\|_w$, $1 \le k \le s$, $1 \le j \le m$, *and the fourth moments of* $\|\partial^3 f_{j,k}/\partial\phi^3\|_w$, *with the same range for* j *and* k, *are bounded*;

(ii) *the quantities in* $A_{d,w,2}$ *are bounded*;

(iii) $\mathrm{E}(|\omega(Y_S)|^\tau)$, $\mathrm{E}(\|\partial\pi/\partial\phi\|_w^{2d})$, $\mathrm{E}(\|\partial^2\pi_k/\partial\phi^2\|^4)$, $\mathrm{E}(\|\partial^3\pi_k/\partial\phi^3\|_w^2)$, $1 \le k \le t$, *are bounded, where* $\tau \ge 4$;

(iv) *there is* $0 \le \kappa \le [3(4d-2)^{-1}] \wedge [(\tau-4)(\tau-2)^{-1}]$ *such that* $|S| = (m^\kappa)$;

(v) $\hat{\phi}_{-i}$, $0 \le i \le m$, *are c.u.* μ_ω *at rate* m^{-d}. *Then*

$$\mathrm{E}\big[\widehat{\mathcal{MSAE}}(\hat{\theta})\big] - \mathcal{MSAE}(\hat{\theta}) = o(m^{-1-\varepsilon})$$

for any $0 < \varepsilon < (d-2)/(2d-1)$, *and hence for the same range of* ε,

$$\mathrm{E}\big[\widehat{\mathrm{MSAE}}(\hat{\theta})\big] - \mathrm{MSAE}(\hat{\theta}) = o(m^{-1-\varepsilon}).$$

THEOREM 5.2. *Suppose that the following hold*:

(i) (4.7) *holds, and there are* $d > 2$ *and* $w > 0$ *such that the* $2d$*th moments of* $|f_j|$, $\|g_j\|$, $\|h_{j,k}\|$, $\|\partial^3 f_{j,k}/\partial\phi^3\|_w$, $1 \leq k \leq s$, $1 \leq j \leq m$, *are bounded*;
(ii) *the quantities in* $A_{d,w,3}$ *are bounded*;
(iii) $\|\partial^4 b/\partial\phi^4\|_w$ *is bounded*;
(iv) $\hat{\phi}_{-i}$, $0 \leq i \leq m$, *are c.u. at rate* m^{-d}. *Then*

$$\begin{aligned}\mathrm{E}\bigl[\widehat{\mathrm{MSE}}(\check{\theta})\bigr] - \mathrm{MSE}(\check{\theta}) &= \mathrm{E}\bigl[\widehat{b(\phi)} - b(\phi)\bigr]\\ &= m^{-1}\biggl(\frac{\partial b}{\partial\phi}\biggr)'(\mathrm{E}\bar{g})^{-1}\sum_{i=1}^{m}(a_{-i} - a) + o(m^{-1-\varepsilon})\end{aligned}$$

for any $0 < \varepsilon < (d-2)/(2d+1)$. *In particular, if* $\sum_{i=1}^{m}\Delta_i = O(m^{-\nu})$ *for some* $\nu > 0$, *then*

$$\mathrm{E}\bigl[\widehat{\mathrm{MSE}}(\check{\theta})\bigr] - \mathrm{MSE}(\check{\theta}) = \mathrm{E}\bigl[\widehat{b(\phi)} - b(\phi)\bigr] = o(m^{-1-\varepsilon})$$

for any $0 < \varepsilon < [(d-2)/(2d+1)] \wedge \nu$.

REMARK 5.6. In some cases, for example, if $a = 0$ (see Example 6.1; also see Section 7), $\sum_{i=1}^{m}\Delta_i = 0$, and hence is $O(m^{\nu})$ for any $\nu > 0$. If $\sum_{i=1}^{m}\Delta_i$ does not vanish, then the value of ν depends on what a is. For example, in Example 6.2 we have $\sum_{i=1}^{m}\Delta_i = O(m^{-1})$; hence $\nu = 1$.

THEOREM 5.3. *Suppose that Theorem* 5.2(i) *and* (iii) *and Theorem* 5.1(iii)–(v) *hold and, furthermore, that the quantities in* $A_{d,w,4}$ *are bounded, and* $\sum_{i=1}^{m}\Delta_i = O(m^{-\nu})$ *for some* $\nu > 0$. *Then*

$$\mathrm{E}\bigl[\widehat{\mathrm{MSE}}(\hat{\theta})\bigr] - \mathrm{MSE}(\hat{\theta}) = o(m^{-1-\varepsilon})$$

for any $0 < \varepsilon < [(d-2)/(2d+1)] \wedge \nu$.

Note that, by Theorems 4.1 and 4.2, and Remark 5.2, sufficient conditions for c.u. μ_ω can be easily obtained.

6. Mixed linear models.

In this section, we consider a mixed linear model:

$$Y_i = X_i\beta + Z_i v_i + e_i, \qquad i = 1, \ldots, m,$$

where X_i ($n_i \times p$) and Z_i ($n_i \times b_i$) are known matrices; v_i's are independent with $\mathrm{E}(v_i) = 0$ and $\mathrm{Var}(v_i) = G_i$, where Var represents covariance matrix; e_i's are independent with $\mathrm{E}(e_i) = 0$ and $\mathrm{Var}(e_i) = R_i$; and v_i's and e_i's are independent. 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 $\psi = (\psi_l)_{1 \leq l \leq q}$, a $q \times 1$ vector of variance components. This model covers Sections 2.1–2.5.

Here we are interested in predicting a mixed effect $\theta = h'\beta + \lambda' v$, where $v = (v_i)_{1\le i\le m}$, h and λ are known vectors of order $p \times 1$ and $b \times 1$, respectively, and $b = \sum_{i=1}^{m} b_i$. Let $\phi = (\beta'\psi')'$. We assume *posterior linearity*; that is, $\mathrm{E}(\theta|Y) = c + d'Y$, where c is a constant and d a constant vector. Then, when ϕ is known, the BP of θ is the best linear predictor given by $\breve{\theta} = h'\beta + \lambda' s'(\psi)(Y - X\beta)$, where $X = \mathrm{col}_{1\le i\le m}(X_i)$, $Y = \mathrm{col}_{1\le i\le m}(Y_i)$, $s(\psi) = \Sigma^{-1}(\psi)ZG(\psi)$ with $\Sigma(\psi) = R + ZGZ'$, $Z = \mathrm{diag}_{1\le i\le m}(Z_i)$, $G = \mathrm{diag}_{1\le i\le m}(G_i)$ and $R = \mathrm{diag}_{1\le i\le m}(R_i)$. Let $\Sigma_i(\psi) = R_i + Z_i G_i Z_i' = \mathrm{Var}(Y_i)$, $1 \le i \le m$. It is interesting to note that $\breve{\theta}$ can be viewed as the linear Bayes estimator of θ under a squared error loss function. Many well-known distributions, including the normal, satisfy posterior linearity (see Section 1 for some references). It can be shown that $\mathrm{MSE}(\breve{\theta}) = b(\psi) = \lambda'[G(\psi) - G(\psi)Z'\Sigma^{-1}(\psi)ZG(\psi)]\lambda$. Note that, typically, only a portion of the components of λ are nonzero, so that $\breve{\theta}$ depends only on a subset of the Y's, say $Y_S = (Y_i)_{i\in S}$. The EBP is given by $\hat{\theta} = h'\hat{\beta} + \lambda' s'(\hat{\psi})(Y - X\hat{\beta})$.

Two methods of M-estimation are the following:

EXAMPLE 6.1 (Maximum likelihood estimation). The ML estimator of ϕ is defined as a solution to the ML equations. It is easy to show that the ML estimator of ϕ is a solution to (4.6), where $a(\phi) = 0$,

$$(f_{j,k}(\phi, Y_j))_{1\le k\le p} = X_j' \Sigma_j^{-1}(\psi)(Y_j - X_j\beta),$$

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

EXAMPLE 6.2 (REML estimation). Similarly, the restricted maximum likelihood estimator of ψ is defined as a solution to the REML equations and the REML estimator of β as the EBLUE

$$\hat{\beta} = (X'\Sigma^{-1}(\hat{\psi})X)^{-1}X'\Sigma^{-1}(\hat{\psi})Y,$$

where $\hat{\psi}$ is the REML estimator. By the identity [e.g., Searle, Casella and McCulloch (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 a solution to (4.6), where the f_j's are the same as in Example 6.1; $a(\phi) = (a_k(\phi))_{1\le k\le p+q}$ with $a_k(\phi) = 0$, $1 \le k \le p$, and

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

As pointed out by Jiang (1996), these equations (ML or REML) may be regarded as M-estimating equations, and the ML and REML estimators as M-estimators. In other words, these equations may be used even if the actual data are not normal, and, under suitable conditions, the resulting estimators are consistent and asymptotically normal.

Suppose that the following regularity conditions (i)–(iv) are satisfied:

(i) $|X_i|$, $|Z_i|$, $1 \le i \le m$, are bounded;

(ii) the true parameter vector $\psi \in \Psi^o$, the interior of the parameter space for ψ;

(iii) for any compact set $B \subset \Psi^o$, the $\sup_{\psi \in B} \|\cdot\|$ of up to fourth derivatives of $R_i(\psi)$ and $G_i(\psi)$, $1 \le i \le m$, are bounded; and $\sup_{\psi \in B} \|\Sigma_i^{-1}(\psi)\|$, $1 \le i \le m$, are bounded;

(iv) $\lambda_{\min}(X_i' \Sigma_i^{-1} X_i)$ and

$$\lambda_{\min}\left[\left(\operatorname{tr}\left(\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \psi_k}\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \psi_l}\right)\right)_{1\le k,l\le q}\right]$$

are bounded away from zero, where $\lambda_{\min}$ represents the smallest eigenvalue.

Recall that $R_i(\psi)$ and $G_i(\psi)$ are the covariance matrices of e_i, the vector of errors, and v_i, the vector of random effects, respectively; and $\Sigma_i(\psi)$ is the covariance matrix of Y_i. Thus, condition (iii) may be interpreted as follows. First, the covariance matrices of the random effects and errors, as functions of ψ, are four-times differentiable, and their up to fourth derivatives are bounded over any compact subset of ψ. Second, the covariance matrix of Y_i is positive definite with its smallest eigenvalue bounded away from 0 over any compact subset of ψ. Note that here the boundedness means both in ψ and in i. We now interpret condition (iv). Under normality, the information matrix is given by [e.g., Searle, Casella and McCulloch (1992), Section 6.3]

$$I\begin{pmatrix}\beta\\ \psi\end{pmatrix} = \left\{\begin{matrix} X'\Sigma^{-1}X & 0 \\ 0 & \dfrac{1}{2}\left[\operatorname{tr}\left(\Sigma^{-1}\dfrac{\partial \Sigma}{\partial \psi_k}\Sigma^{-1}\dfrac{\partial \Sigma}{\partial \psi_l}\right)\right]_{1\le k,l\le q}\end{matrix}\right\}.$$

Note that $X'\Sigma^{-1}X = \sum_{i=1}^m X_i'\Sigma_i^{-1}X_i$ and

$$\operatorname{tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \psi_k}\Sigma^{-1}\frac{\partial \Sigma}{\partial \psi_l}\right) = \sum_{i=1}^m \operatorname{tr}\left(\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \psi_k}\Sigma_i^{-1}\frac{\partial \Sigma_i}{\partial \psi_l}\right).$$

Thus, condition (iv) is a nonsingularity condition which implies that the information matrix is nonsingular with its smallest eigenvalue bounded from below by δm, where δ is a positive constant.

In addition to (i)–(iv), we assume that the following hold:

(v) $\mathrm{E}|Y_i|^{8+\delta}$, $1 \le i \le m$, are bounded for some $\delta > 0$;
(vi) $|S| = O(m^{\kappa})$ for some $0 \le \kappa \le 3/(6+\delta)$.

A simple example may help to further illustrate these conditions.

EXAMPLE 6.3. Consider prediction of the mixed effect $\theta = \mu + v_1$ in Section 2.2. We show that conditions (i)–(vi) are satisfied.

Since $X_i = Z_i = 1$, condition (i) is obvious. Condition (ii) holds because, by assumption, $A > 0$. Also, since $R_i(A) = 1$, which does not depend on A, and $G_i(A) = A$, we have $\Sigma_i(A) = A + 1$. It follows that condition (iii) holds. Note that $\Sigma_i^{-1}(A) \le 1$, $\forall A > 0$. As for condition (iv), we have $X_i'\Sigma_i^{-1}X_i = (A+1)^{-1}$ and $[\Sigma_i^{-1}(\partial\Sigma_i/\partial A)]^2 = (A+1)^{-2}$. Thus, condition (iv) holds. Condition (v) is obvious. Here $S = \{1\}$ and $|S| = 1$. Thus condition (vi) is satisfied. This example will be revisited in a simulation study in Section 8.

If $\hat{\phi}_{-i}$, $0 \le i \le m$, are chosen as the ML estimators of Example 6.1, then $a_{-i} = 0$, $0 \le i \le m$. Therefore, it is easy to show, by Theorem 5.3, that, under conditions (i)–(vi), $\mathrm{E}[\widehat{\mathrm{MSE}}(\hat{\theta})] = \mathrm{MSE}(\hat{\theta}) + o(m^{-1})$.

If $\hat{\phi}_{-i}$, $0 \le i \le m$, are chosen as the REML estimators of Example 6.2, the a's are no longer 0. However, it can be shown that the conditions on the a's in Theorem 5.3 are satisfied [see Jiang (1999)]. It follows from Theorem 5.3 that, under conditions (i)–(vi), $\mathrm{E}[\widehat{\mathrm{MSE}}(\hat{\theta})] = \mathrm{MSE}(\hat{\theta}) + o(m^{-1})$.

It is easy to apply the above general results to the specific models in Sections 2.1–2.5.

Note that both the ML and REML equations are standard. In the following, we consider a case in which the M-estimating equations may not be standard.

7. Mixed logistic models. Consider the mixed logistic model given in Example 2.6. As for the M-estimators, we consider the MOM estimators of Jiang (1998). The MOM estimator for ϕ is the solution to the following system of equations:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ijk}Y_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ijk}\mathrm{E}_{\phi}Y_{ij}, \qquad 1 \le k \le p, \tag{7.1}$$

$$\sum_{i=1}^{m}\sum_{j\ne l} Y_{ij}Y_{il} = \sum_{i=1}^{m}\sum_{j\ne l} \mathrm{E}_{\phi}Y_{ij}Y_{il}. \tag{7.2}$$

It is clear that these equations are in the form of (4.6) with $a = 0$. Furthermore, if the x_{ijk}'s are bounded, the terms corresponding to f_j's in (4.6) and their derivatives of any order are bounded. Note that $\mathrm{E}_{\phi}Y_{ij} = \mathrm{E}\,\mathrm{logit}^{-1}(x_{ij}^t\beta + \sigma\xi)$,

$\mathrm{E}Y_{ij}Y_{il} = \mathrm{E}\,\mathrm{logit}^{-1}(x_{ij}^t\beta + \sigma\xi)\,\mathrm{logit}^{-1}(x_{il}^t\beta + \sigma\xi)$, $j \neq l$. Note that, unlike the ML and REML equations in the linear mixed model case, the MOM equations are not necessarily standard. Nevertheless, by Theorem 4.2, it is easy to give sufficient conditions, under which the MOM estimators are c.u. at rate m^{-d} for any $d > 0$. It follows, by Theorem 5.3, that $\mathrm{E}[\widehat{\mathrm{MSE}}(\hat{\theta})] = \mathrm{MSE}(\hat{\theta}) + o(m^{-1-\varepsilon})$ for any $0 < \varepsilon < 1/2$. We consider a special case in the following.

EXAMPLE 7.1. Suppose that, in (2.1), $x_{ij}^t\beta = \mu$, an unknown intercept, and that $n_i = n$, $1 \le i \le m$, where $m \to \infty$ while n is fixed. We verify the conditions of Theorem 4.2. First, note that (7.1) and (7.2) now can be written as

$$\sum_{i=1}^{m}[Y_{i\cdot} - \mathrm{E}_\phi(Y_{i\cdot})] = 0,$$

$$\sum_{i=1}^{m}[Y_{i\cdot}^2 - Y_{i\cdot} - \mathrm{E}_\phi(Y_{i\cdot}^2 - Y_{i\cdot})] = 0,$$

where $Y_{i\cdot} = \sum_{j=1}^n Y_{ij}$, $\phi = (\mu, \sigma)$. Furthermore, we have $M(\phi) = (M_1(\phi), M_2(\phi))$, where $M_1(\phi) = nh_1(\phi)$, $M_2(\phi) = n(n-1)h_2(\phi)$ with $h_r(\phi) = \mathrm{E}h^r(\mu + \sigma\xi)$, $r = 1, 2$, $h(x) = e^x/(1 + e^x)$ and $\xi \sim N(0, 1)$.

It is easy to show that $\sup_{\mu\in R}[h_1(\phi) - h_2(\phi)] \to 0$ as $\sigma \to \infty$. Thus, there are $\varepsilon > 0$ and $B_2 > 0$ such that

$$\sup_{\sigma > B_2}\sup_{\mu\in R}[h_1(\phi) - h_2(\phi)] < \varepsilon.$$

On the other hand, we have $\sup_{0\le\sigma\le B_2}[1 - h_1(\phi)] \to 0$ as $\mu \to \infty$; and $\sup_{0\le\sigma\le B_2} h_1(\phi) \to 0$ as $\mu \to -\infty$. Thus, there is $B_1 > 0$ such that

$$\left\{\sup_{\mu > B_1}\sup_{0\le\sigma\le B_2}[1 - h_1(\phi)]\right\} \vee \left[\sup_{\mu < -B_1}\sup_{0\le\sigma\le B_2} h_1(\phi)\right] < \varepsilon.$$

Let $B = B_1 \vee B_2$. It follows that $\inf_{\phi\notin\Phi_B}|h(\phi) - h(\phi_0)| \ge \varepsilon$, from which (4.8) easily follows.

Furthermore, it is easy to show, by Taylor expansion, that there is $\delta > 0$ and $\varepsilon > 0$ such that

$$|h(\phi) - h(\phi_0)| \ge \varepsilon|\phi - \phi_0|, \qquad |\phi - \phi_0| < \delta, \qquad \sigma \ge 0,$$

where $h(\phi) = (h_1(\phi), h_2(\phi))$. On the other hand, it can be shown that $h(\cdot)$ is injective [Jiang (1998)]. Thus, for any $D > \delta \vee |\phi_0|$, the continuous function $g(\phi) = |h(\phi) - h(\phi_0)|/|\phi - \phi_0|$, $\delta \le |\phi - \phi_0| \le D$, $\sigma \ge 0$, has a lower bound $\eta > 0$. Thus, $\inf_{\phi\in\Phi_B, \phi\neq\phi_0}\{|h(\phi) - h(\phi_0)|/|\phi - \phi_0|\} \ge \varepsilon \wedge \eta$, from which (4.9) easily follows.

Statements (4.10) and (4.11) are obvious. Note that, in this case, $R_t = \{(u, v) : 0 < u, v < 1, u^2 \le v < u\}$.

8. Monte Carlo simulations. Although our jackknife method is not specifically designed to produce MSE estimators for mixed linear *normal* models, it is instructive to compare its performance with a number of other measures of uncertainties of EBLUP (or EB) in the normal case. Let us first consider the prediction of $\theta = \mu + v_1$ in Section 2.2 (also Example 6.3).

We consider two different values of m, $m = 30$ or 60. The data are generated with $\mu = 0.0$ and $A = 1.0$. The values of average relative bias (ARB) are reported as percentages, given by

$$\mathrm{ARB} = 100 \times \left[\frac{\mathrm{E}(\widehat{\mathrm{MSE}}) - \mathrm{MSE}}{\mathrm{MSE}}\right].$$

There are various measures of uncertainty of the above EBLUP (or EB) available in the literature, proposed from different approaches. To compare these different methods, we investigate how accurately they estimate MSE (same as the Bayes risk). The naive estimator of MSE does not capture uncertainty due to estimation of A and is given by $(1 - \hat{B})$. The measure proposed by Morris (1983) is actually an approximation to the posterior variance of θ under a flat prior on μ and A. Laird and Louis (1987) used a parametric bootstrap to mimic a hyperprior Bayesian calculation. Butar (1997) used a parametric bootstrap method to estimate MSE of EBLUP. All the measures other than our jackknife are based on normality. However, the performance of our jackknife method is quite impressive. Table 1 shows results based on one million simulations. Note that the naive estimator generally underestimates (sometimes severely) the true MSE since it does not incorporate the uncertainty due to estimation of A. The Laird–Louis measure also tends to underestimate the true MSE. This is consistent with the result of Butar (1997), who showed that the bias of the Laird–Louis measure in estimating MSE is of the order $O(m^{-1})$, same as that of the naive measure. We reiterate that the orders of the biases are $o(m^{-1})$ for Morris (1983), Prasad and Rao (1990), Butar (1997) and the proposed jackknife measure. The performances of all the measures improve as m increases. For the sake of comparison, we consider a hierarchical Bayesian method with flat priors on μ and A to produce a measure of uncertainty for $\hat{\theta}$. The posterior variance underestimates the MSE of $\hat{\theta}$. However, under a hierarchical Bayesian setup a more meaningful measure of uncertainty of $\hat{\theta}$ is $E[(\theta - \hat{\theta})^2 | y] = V(\theta | y) + [E(\theta | y) - \hat{\theta}]^2$. This measure compares quite well with the other measures given in Table 1.

TABLE 1
ARB for Example 6.3 (*Section* 2.2)

m	Naive	Jackknife	Morris	Butar	Prasad–Rao	Laird–Louis	HB
30	−8.4	0.6	0.7	0.1	0.7	−3.8	−3.0
60	−4.8	0.2	−0.1	−0.2	−0.1	−2.9	−0.4

TABLE 2
ARB for normal–normal case (*Section* 2.4)

σ_v^2	**Naive**	**Jackknife**	**Prasad–Rao**
1.5	−2.0	0.4	1.3
1.0	−3.1	0.5	1.4
0.5	−6.7	0.2	1.5

In Table 2, we report ARB (based on 4,000 simulations) of our proposed jackknife MSE estimator for the simple random effects normal model given in Section 2.4 with $\mu = 0$, $\sigma_e^2 = 1$, $m = 30$ and $n = 5$. Here the EBP (also EBLUP) of $\theta_i = \mu + v_i$ is given by $\hat{\theta}_i = \bar{Y}_{\cdot\cdot} + (1 - \hat{B})(Y_i - \bar{Y}_{\cdot\cdot})$, where $\hat{B} = \hat{\sigma}_e^2/(\hat{\sigma}_e^2 + \hat{\sigma}_v^2)$. Clearly, this table shows that our proposed jackknife method is very accurate (more accurate than the Prasad–Rao estimator developed using normality). Also, larger values of σ^2 tend to improve the ARB.

The purpose of Table 3 is to investigate robustness of our proposed jackknife MSE estimator compared to the naive and the Prasad–Rao MSE estimators when Y_{ij} is generated from a nonnormal model. For illustration, we assume that $Y_{ij}|\theta_i$ are independent point binomial with parameter θ_i and a priori θ_i's are i.i.d. beta(α, β). By choosing various values of α and β, we can get a variety of beta distributions. Note that all the formulas involved can be expressed as a function of $\bar{Y}_i$ since $\mathrm{MSW} = n \sum_{i=1}^m \bar{Y}_i(1 - \bar{Y}_i)/m(n-1)$. This simplifies the computations since we can equivalently generate $\sum_{j=1}^n Y_{ij}$ from binomial distributions with parameters n and θ_i. Here, again, we use $m = 30$ and $n = 5$. The results (based on 4,000 simulations) show that the naive estimator underestimates the true MSE considerably. Both the Prasad–Rao and our proposed jackknife MSE estimators improve on the naive estimator. Also, the jackknife estimator tends to perform better than the Prasad–Rao estimator. This simulation demonstrates that although the assumption of posterior linearity produces the same normality-based point estimator, the assumption makes a difference when one is interested in MSE estimator of the same point estimator.

TABLE 3
ARB for the beta–binomial case

(α, β)	**Naive**	**Jackknife**	**Prasad–Rao**
(2, 4)	−7.5	−1.4	−2.7
(2, 0.5)	−5.2	−2.5	−4.0
(0.5, 0.5)	−6.3	0.0	−2.3
(0.5, 2)	−3.6	0.5	−2.0
(2, 2)	−8.7	−0.2	−4.0
(1, 1)	−6.9	0.1	−1.2

TABLE 4
ARB for mixed logistic model

		Naive method		Jackknife method		Taylor series method	
m	**MSE**	**Mean**	**ARB**	**Mean**	**ARB**	**Mean**	**ARB**
10	0.996	0.729	−26.8	0.952	−4.4	0.917	−7.9
20	0.805	0.724	−10.1	0.819	1.7	0.813	1.0
40	0.767	0.721	−6.0	0.765	−0.2	0.764	−0.3

Note that our jackknife method is applicable to other kinds of nonnormal nonlinear models such as the mixed logistic models, while other rival measures considered so far in this section are not. As a final example, we consider a simulation which is related to Section 2.6.

The data are generated with $\mu = 0.5$ and $\sigma = 1.0$. By (2.2), the BP or Bayes estimator is given by

$$\tilde{\theta} = \frac{\mathrm{E}(\mu + \xi)\exp(\phi(Y_{1\cdot}, \xi, \mu))}{\mathrm{E}\exp(\phi(Y_{1\cdot}, \xi, \mu))} \equiv \psi(Y_{1\cdot}, \mu),$$

where $Y_{i\cdot} = \sum_{j=1}^{n} Y_{ij}$ and $\xi \sim N(0, 1)$. The parameter μ is estimated by the method of simulated moments [Jiang (1998)]; that is, by solving $\mathrm{E}h(\mu + \xi) = \bar{Y}_{\cdot\cdot}$, where $h(u) = e^u/(1 + e^u)$ and $\bar{Y}_{\cdot\cdot} = (mn)^{-1}\sum_{i=1}^{m} Y_{i\cdot}$. Three different sample size configurations are considered: $m = 10$, 20 and 40, and $n = 2$ in all cases. The results in Table 4 for $m = 10$, 20 and 40 are based on 1,000, 10,000 and 80,000 simulations, respectively. [There is a reason for the varying simulation sizes: It may be argued that, for the purpose of simulations in this section, the order of the simulation size should be higher than m^2. Thus, in cases of close competitor(s) such as the current example, we have chosen the simulation sizes proportional to m^3.] The table shows the true MSE; the mean of the jackknife MSE estimator based on the simulation; and the ARB. Similar results are also presented, as comparison, for the naive estimator of the MSE and for the estimator of MSE of Jiang and Lahiri (2001). The latter method is based on Taylor series expansion, which may be regarded as an extension of Prasad and Rao (1990) to mixed logistic models. As expected, the jackknife and Taylor series methods perform similarly with high accuracy for this nonnormal and nonlinear case when the sample size is relatively large ($m = 20$ or 40). For small sample size ($m = 10$) the jackknife method seems to perform better. Moreover, the jackknife method has a computational advantage over the Taylor series method, especially for complex models. This is because to compute the MSE estimator of Jiang and Lahiri (2001) one has to obtain analytic forms of the first and second derivatives involved (numerical differentiation is unstable), which is not easy when the model gets complicated, and errors are often made in derivation and programming. Note that here we only consider a very simple model.

Acknowledgments. The authors thank two referees, an Associate Editor and a Co-Editor for their constructive comments. In particular, the Associate Editor's detailed suggestions helped greatly to improve the readability of the paper. Finally, the authors thank Dr. Shijie Chen for computing support.

REFERENCES

ARVESEN, J. N. (1969). Jackknifing U-statistics. *Ann. Math. Statist.* **40** 2076–2100.

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.

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

BUTAR, F. B. (1997). Empirical Bayes methods in survey sampling. Ph.D. dissertation, Dept. Mathematics and Statistics, Univ. Nebraska–Lincoln.

CHATTOPADHYAY, M., LAHIRI, P., LARSEN, M. and REIMNITZ, J. (1999). Composite estimation of drug prevalences for sub-state areas. *Survey Methodology* **25** 81–86.

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.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors: An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

ERICSON, W. A. (1969). A note on the posterior mean of a population mean. *J. Roy. Statist. Soc. Ser. B* **31** 332–334.

FAY, R. E. and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277.

GHOSH, M. and LAHIRI, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *J. Amer. Statist. Assoc.* **82** 1153–1162.

GHOSH, M. and MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *J. Amer. Statist. Assoc.* **81** 1058–1062.

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. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93** 720–729.

JIANG, J. (1999). Jackknifing MSE of empirical best predictor: A theoretical synthesis. Technical report, Dept. Statist., Case Western Reserve Univ.

JIANG, J. and LAHIRI, P. (2001). Empirical best prediction for small area inference with binary data. *Ann. Inst. Statist. Math.* **53** 217–243.

LAHIRI, P. (1995). A jackknife measure of uncertainty of linear empirical Bayes estimators. Unpublished manuscript.

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.

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *J. Amer. Statist. Assoc.* **82** 739–757.

LEE, Y. and NELDER, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. Roy. Statist. Soc. Ser. B* **58** 619–678.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

MALEC, D., SEDRANSK, J., MORIARITY, C. L. and LE CLERE, F. B. (1997). Small area inference for binary variables in the National Health Interview Survey. *J. Amer. Statist. Assoc.* **92** 815–826.

MCFADDEN, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* **57** 995–1026.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–59.

PRASAD, N. G. N. and RAO, J. N. K. (1988). Robust tests and confidence intervals for error variance in a regression model and for functions of variance components in an unbalanced random one-way model. *Comm. Statist. Theory Methods* **17** 1111–1133.

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.

RAO, J. N. K. and PRASAD, N. G. N. (1986). Discussion of "Jackknife, bootstrap and other resampling methods in regression analysis," by C. F. J. Wu. *Ann. Statist.* **14** 1320–1322.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York.

SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

J. JIANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
ONE SHIELDS AVENUE
DAVIS, CALIFORNIA 95616
E-MAIL: jiang@wald.ucdavis.edu

P. LAHIRI
JOINT PROGRAM
IN SURVEY METHODOLOGY
UNIVERSITY OF MARYLAND
1218 LEFRAK HALL
COLLEGE PARK, MARYLAND 20742
E-MAIL: plahiri@survey.umd.edu

S.-M. WAN
DEPARTMENT OF FINANCE
LUNGHWA UNIVERSITY OF SCIENCE
AND TECHNOLOGY
300, WAN-SO ROAD, SEC. 1, KWEI-SAN SHANG
TOU-YUANG COUNTY
TAIWAN, REPUBLIC OF CHINA 333