Comment

David A. Harville

The publication of Robinson's article is very timely. As related by Robinson, BLUP is a statistical methodology that has been used extensively in animal breeding, with great success. Robinson has brought to our attention the similarities between BLUP and various other methodologies, like kriging and the Kalman filter, and has noted that, while BLUP was developed via a frequentist approach to statistics, it has a Bayesian interpretation. In doing so, he has performed a valuable service to those interested in the development and application of BLUP (and the related methodologies) and to the statistics community as a whole.

The main theme of my discussion, which is developed in Sections 1, 4, and 5, is that BLUP and the related methodologies should be discussed in the common framework of a general prediction problem, that BLUP has some deficiencies that could be eliminated by a more extensive use of Bayesian ideas and that a unified, yet flexible, approach to prediction is desirable and achievable. My discussion includes (in Sections 2 and 3) some comparisons between work on BLUP and related work on empirical Bayes inference and some comments about some long-standing misconceptions regarding the use of mixed-effects linear models.

1. BLUP FOR A GENERAL PREDICTION PROBLEM

It is instructive to consider BLUP in the context of the general problem of predicting the value of an unobservable random variable w based on the value of an $n \times 1$ observable random vector y, where the joint distribution of w and y has first and second moments to be denoted by $\mu_w = E(w)$, $\mu_y = E(y)$, $v_w = \text{var}(w)$, $v_{yw} = \text{cov}(y, w)$, and $V_y = \text{var}(y)$. It is assumed that μ_y belongs to a known vector space $\mathscr{M}$ and that μ_w is a known linear combination of the elements of μ_y, or equivalently, that $\mu_y = X\beta$ and $\mu_w = \lambda'\beta$, where β is a $p \times 1$ vector of unknown parameters, X is an $n \times p$ known matrix of rank p^* (any matrix whose columns span $\mathscr{M}$), and λ is a $p \times 1$ known vector that is expressible as $\lambda = X'k$ for some vector k. The quantity v_w and the elements of v_{yw} and V_v are assumed to be known functions of an unknown parameter vector $\theta = (\theta_1, \ldots, \theta_q)'$, whose value is restricted to a known set Ω, and V_y is assumed to be nonsingular (for all $\theta \in \Omega$). Note that, aside from the linearity of the mean structure, the primary limitation imposed by these assumptions is that v_w, v_{yw} and V_y are unrelated to μ_w and μ_y. Note also that, in the special case where $v_w = 0$, $w = \lambda'\beta$ (with probability one), and the problem of predicting the value of w is essentially equivalent to that of inference about $\lambda'\beta$.

Clearly, the problem of predicting a linear combination of the fixed and random effects in Robinson's model (1.1) can be formulated as a special case of the general prediction problem. Moreover, many of the problems considered by Robinson in his Section 6, which come from quality assurance, geostatistics and various other fields, can likewise be formulated as special cases of the general prediction problem. For many of these problems, it may seem more natural to formulate them (directly) in terms of the general prediction problem than to follow Robinson's approach of recasting them in terms of fixed and random effects.

In introducing the BLUP—I use BLUP as an acronym for the best linear unbiased predictor as well as for best linear unbiased prediction—and in establishing its BLUPness, I prefer an approach that differs somewhat from any of those presented by Robinson. This approach, which I now describe in the context of the general prediction problem, clearly reveals the intuitive appeal of the BLUP and takes advantage of some well-known results on statistical estimation and on the MVN distribution.

Let $\overline{w} = \mu_w + v'_{yw}V_y^{-1}(y - \mu_y) = \tau + v'_{yw}V_y^{-1}y$, where $\tau = \mu_w - v'_{yw}V_y^{-1}\mu_y = (\lambda' - v'_{yw}V_y^{-1}X)\beta$. If τ and $v'_{yw}V_y^{-1}$ were known and if the joint distribution of w and y were assumed to be MVN, then the best (minimum MSE) predictor of the value of w would be $E(w \mid y) = \overline{w}$, as is well known. Since $\overline{w}$ is linear (in y), it is clear that even if the normality assumption were dropped, $\overline{w}$ would be the best linear predictor of the value of w.

Now, let $\tilde{\beta}$ represent any solution to the Aitken equations $X'V_y^{-1}X\tilde{\beta} = X'V_y^{-1}y$. If $(\lambda' - v'_{yw}V_y^{-1}X)(X'V_y^{-1}X)^{-}X'V_y^{-1}$ and $v'_{yw}V_y^{-1}$ were known, then the BLUE (best linear unbiased estimator) of τ would be $\tilde{\tau} = (\lambda' - v'_{yw}V_y^{-1}X)\tilde{\beta}$, and a "natural"

David A. Harville is Professor, Department of Statistics, Iowa State University, Ames, Iowa 50011.

predictor of the value of w would be the predictor

$$\tilde{w} = \tilde{\tau} + v'_{yw} V_y^{-1} y \tag{1}$$

obtained from $\bar{w}$ by substituting $\tilde{\tau}$ for τ. In fact, $\tilde{w}$ would be the BLUP.

To verify the BLUPness of $\tilde{w}$, observe that a predictor (of the value of w) is linear and unbiased if and only if it is expressible in the form $r'y + v'_{yw} V_y^{-1} y$, where $r'y$ is a (linear) unbiased estimator of τ, that is, where $r'X = \lambda' - v'_{yw} V_y^{-1} X$. Moreover, the MSE of a predictor of this form is

$$\begin{aligned} E\left[\left(r'y + v'_{yw} V_y^{-1} y - w\right)^2\right] &= \operatorname{var}\left(r'y + v'_{yw} V_y^{-1} y - w\right) \\ &= \operatorname{var}(r'y) + \operatorname{var}\left(v'_{yw} V_y^{-1} y - w\right), \end{aligned}$$

implying (since $\tilde{\tau}$ is the BLUE of τ) that $\tilde{w}$ is the BLUP. A further implication is that the MSE of $\tilde{w}$ is

$$\begin{aligned} &v_w - v'_{yw} V_y^{-1} v_{yw} + \left(\lambda' - v'_{yw} V_y^{-1} X\right) \\ &\quad \cdot \left(X' V_y^{-1} X\right)^{-} \left(\lambda - X' V_y^{-1} v_{yw}\right). \end{aligned} \tag{2}$$

Robinson (in his Section 1) has chosen to define the BLUP (and to express its MSE) in terms of the mixed-model equations. Thus, his definition is specific to mixed-model prediction. I prefer to start with the more widely applicable representations (1) and (2) and to use them (together with well-known matrix identities) to derive the representations given by Robinson. The value of the latter representations is that, in the special case of mixed-model prediction, v_w, v_{yw} and V_y have a relatively simple structure, and these representations indicate how (for computational purposes) to exploit that structure. In much the same way, the Kalman filter can be regarded as an algorithm for efficiently computing BLUPs in the special case of "time-series" prediction.

2. EMPIRICAL BLUP VERSUS EMPIRICAL BAYES

To account (in the general prediction problem) for θ being unknown, it is common practice to adopt an even, translation-invariant estimator $\hat{\theta}$ of θ and, for purposes of the (point or interval) prediction of the value of w, to act as though $\hat{\theta}$ is the true value of θ. In particular, a point predictor, say $\hat{w}$, can be obtained from the BLUP $\tilde{w}$ by substituting $\hat{\theta}$ for θ. It seems natural to refer to this predictor as the *empirical BLUP*.

Empirical BLUP, as applied to mixed-effects linear models with normally distributed random effects and errors, is equivalent to PEB (parametric empirical Bayes) inference, as applied to one very important class of problems. This equivalence is discussed by Robinson in his Section 5.7, but only briefly and in rather general terms. It may be worthwhile to examine the equivalence in one relatively simple setting.

Consider first PEB inference about the means $\mu_1, \ldots, \mu_I$ of I "groups" based on the one-way cell-mean model $y_{ij} = \mu_i + e_{ij}$ $(i = 1, \ldots, I;\ j = 1, \ldots, J_i)$, where $e_{11}, e_{12}, \ldots, e_{IJ_I}$ are normally and independently distributed random variables with mean zero and common, unknown variance σ_e^2. It is instructive to describe PEB inference by relating it to HB (hierarchical Bayes) inference. HB inference about $\mu_1, \ldots, \mu_I$ is essentially Bayesian inference based on a prior distribution that is assigned to $\mu_1, \ldots, \mu_I$, and σ_e^2 in stages, say by (1) specifying that, conditional on σ_e^2 and "hyperparameters" μ and γ, the means $\mu_1, \ldots, \mu_I$ are independently and identically distributed as $N(\mu, \gamma\sigma_e^2)$, (2) assigning μ a distribution conditional on σ_e^2 and γ, (3) assigning σ_e^2 a distribution conditional on γ, and (4) assigning γ a distribution. In PEB inference, the prior distribution might only be specified up to the values of μ, γ, and σ_e^2, and the evaluation of point or interval estimators may be based on criteria such as PEB risk that are defined in terms of the conditional distribution of $y_{11}, y_{12}, \ldots, y_{IJ_I}$ and $\mu_1, \ldots, \mu_I$ given μ, γ, and σ_e^2. The (parametric) empirical Bayesian may use frequentist methods (i.e., methods based on the conditional distribution of $y_{11}, y_{12}, \ldots, y_{IJ_I}$ given μ, γ, and σ_e^2) to obtain estimates $\hat{\mu}$, $\hat{\gamma}$, and $\hat{\sigma}_e^2$ of μ, γ, and σ_e^2, and then, acting as though $\hat{\mu}$, $\hat{\gamma}$, and $\hat{\sigma}_e^2$ are the true values of μ, γ, and σ_e^2, use what would, if the true values were known, be the posterior mean of μ_i and a credible set for μ_i as point and interval estimators. Thus, letting $\hat{\delta}_i = (1 + J_i\hat{\gamma})^{-1}$ $(i = 1, \ldots, I)$ and taking $\hat{\mu} = \Sigma_i J_i \hat{\delta}_i \bar{y}_{i\cdot} / \Sigma_i J_i \hat{\delta}_i$,

$$\hat{\mu}_i = \hat{\mu} + \left(1 - \hat{\delta}_i\right)\left(\bar{y}_{i\cdot} - \hat{\mu}\right) = \hat{\delta}_i \hat{\mu} + \left(1 - \hat{\delta}_i\right)\bar{y}_{i\cdot} \tag{3}$$

is a PEB (point) estimator of μ_i. PEB point and interval estimators may be modified to account for the estimation of μ, γ, and σ_e^2.

For purposes of comparison, consider now the empirical BLUP approach to the prediction of the values of $\mu + a_1, \ldots, \mu + a_I$, based on the one-way random-effects model

$$y_{ij} = \mu + a_i + e_{ij}\left(i = 1, \ldots, I;\ j = 1, \ldots, J_i\right), \tag{4}$$

where μ is an unknown parameter and where $a_1, \ldots, a_I$ and $e_{11}, e_{12}, \ldots, e_{IJ_I}$ are normally and independently distributed random variables with mean zero and common, unknown variances σ_a^2 and σ_e^2, respectively. Upon setting $\gamma = \sigma_a^2/\sigma_e^2$, we find that the joint distribution of $y_{11}, y_{12}, \ldots, y_{IJ_I}$ and $\mu + a_1, \ldots, \mu + a_I$ under the one-way random-effects model is the same as the conditional joint distribution of $y_{11}, y_{12}, \ldots, y_{IJ_I}$ and $\mu_1, \ldots, \mu_I$ (given μ, γ, and σ_e^2) obtained in the PEB approach to the one-way cell-mean model. Thus, the criteria adopted by the empirical BLUPer in the (point and interval) prediction of the values of $\mu + a_1, \ldots, \mu + a_I$ under the one-way random-effects model are equivalent to those adopted by an empirical Bayesian in the estimation of $\mu_1, \ldots, \mu_I$. In particular, the definitions of MSE and probability of coverage adopted by empirical BLUPers are equivalent to those adopted by empirical Bayesians. Accordingly, the (point and interval) estimators of $\mu_1, \ldots, \mu_I$ that have been proposed by empirical Bayesians tend to be similar or identical to the predictors of the values of $\mu + a_1, \ldots, \mu + a_I$ proposed by empirical BLUPers, especially the point estimators and predictors. For example, it is easy to show (say by using the mixed-model equations) that the PEB estimator $\hat{\mu}_i$ is an empirical BLUP predictor of the value of $\mu + a_i$.

While the work of empirical Bayesians has much in common with that of empirical BLUPers, there are some notable differences. In particular, in devising modifications to account for the estimation of γ and σ_e^2, empirical Bayesians tend to pay more attention to conditional (given the values of various of the quantities $y_{11}, y_{12}, \ldots, y_{IJ_I}$ and $\mu_1, \ldots, \mu_I$—or of functions of those quantities—in addition to the values of μ, γ, and σ_e^2) properties and relatedly to make greater use of HB ideas. Further, much of the work on PEB inference has been carried out by professional statisticians and has been theoretical in nature. This work has tended to focus on relatively simple models, like the balanced ($J_1 = \cdots = J_I$) one-way cell-mean model, since it is only these models that are tractable from a theoretical standpoint. Moreover, this work is often carried out under simplifying assumptions (e.g., an assumption that σ_e^2 is known). As a consequence, the impact of this work on statistical practice, while significant (Morris, 1983), has been restricted. By way of comparison, much of the work on empirical BLUP has been carried out by individuals in animal breeding and other fields of application, is applicable to relatively complex models (which may be intractable from a theoretical standpoint but which are often needed in practice), has focused on the problems (e.g., computational problems) encountered in applications and has had tremendous impact on statistical practice.

3. RANDOM EFFECTS

It is widely believed that the adoption of a mixed-effects linear model is appropriate only if interest centers on fixed effects and components of variance. A related and equally widespread belief is that, if inferences are to be confined to the effects of only those levels of a factor that are represented in the data, then those effects should be treated as fixed. As discussed by Robinson, both beliefs are incorrect. It seems ironic that Searle, who received a Ph.D. degree in animal breeding under the direction of C. R. Henderson, helped perpetuate these beliefs (through his highly influential book *Linear Models*). It may be worth mentioning that Searle partially redeemed himself by pointing out (on pages 461–462 of *Linear Models*) that the second part of the solution to the mixed-model equations ($\hat{u}$ in Robinson's notation) "is, in many situations, of interest also."

For the one-way random-effects model (4), the empirical BLUP of the value of $\mu + a_i$ is given by expression (3). If the effects $a_1, \ldots, a_I$ were treated as fixed rather than random, the empirical BLUP would be $\bar{y}_{i\cdot}$. It would seem that the ultimate goal in deciding whether to treat effects as fixed or random should be to obtain the most useful methodology. Robinson suggests that effects should be treated as random only if they "come from a probability distribution." The results of James and Stein (1961) suggest that this requirement may be overly stringent. Moreover, if strictly interpreted, it would never be satisfied. A less restrictive and perhaps more satisfactory requirement would be that the anticipated values of the effects be values that might "likely" arise in sampling a distribution.

I am not sure that the distinction between variation and uncertainty is as clear-cut as Robinson's discussion (in his Section 7.5) would seem to indicate, nor am I convinced that "classical statistics can be distinguished from Bayesian statistics by its refusal to use probability distributions to describe uncertainty." I believe that many classical statisticians are willing to use probability distributions to describe uncertainty anytime they feel their distributional assumptions will be acceptable to those to whom they wish to communicate their results.

4. BAYESIAN PREDICTION

The empirical BLUP $\hat{w}$ has an unappealing feature, which is explainable in terms of the confinement of θ to Ω. Suppose, for example, that the model is the one-way random-effects model (4), and consider the empirical BLUP $\hat{\mu}_i$ of the value of

$\mu + a_{i\cdot}$. When $\hat{\gamma} = 0$, the empirical BLUP reduces to $\bar{y}_{\cdot\cdot}$, which would be the BLUP of the value of $\mu + a_i$ if it were known with certainty that $\gamma = 0$. This feature seems unappealing. The true value of γ can be larger than zero, but no smaller. Thus, when $\hat{\gamma} = 0$, we may consider it "likely" that the true value of γ is larger than the estimated value, and therefore be reluctant to act as though the estimated value is the true value.

An appealing alternative to the empirical BLUP can be obtained via the Bayesian approach. The Bayesian approach, as applied to the general prediction problem, consists of assigning β and θ a joint distribution (based on prior information), of completing the specification of the (conditional) joint distribution of w and y (given β and θ) up to the values of β and θ, and of forming the posterior distribution of w. In the event p^* equals p, β and θ are statistically independent, the density of β is proportional to a constant, and the conditional distribution of w and y given β and θ is MVN, the posterior probability density function of w is

$$(5) \quad f(w \mid y) \propto \int_{\Omega} g(w \mid y, \theta) l(\theta \mid y) \pi(\theta)\, d\theta,$$

where $g(w \mid y, \theta)$ is the probability density function of a normal distribution whose mean is the BLUP of w and whose variance is the MSE of the BLUP, $l(\theta \mid y)$ is the likelihood function employed in REML, and $\pi(\theta)$ is the (prior) density of θ (e.g., Harville, 1990). The posterior mean (i.e., the mean of distribution (5)) is a weighted "average" (over θ) of the BLUP, with weights proportional to $l(\theta \mid y)\pi(\theta)$. Point predictors of the general form of the posterior mean provide an appealing alternative to the empirical BLUP.

The Bayesian approach is computationally intensive—so much so that its use in many mixed-model applications is at present unfeasible. Recently, there has been increased interest in the computational aspects of the Bayesian approach (e.g., Kadane, 1990). More research should be directed towards the solution of the computational problems encountered in a Bayesian approach to mixed-model prediction.

5. A UNIFIED APPROACH TO PREDICTION

Methodologies have been developed for a number of special cases of the general prediction problem, including mixed-model methodology, kriging and the Kalman filter. These methodologies tend to be identified with specific types of application such as animal breeding, geostatistics and control theory, and they have tended to evolve independently of each other, with relatively little input from professional statisticians. To facilitate the exchange of ideas and to avoid duplication of effort, it would be desirable to develop a unified, but flexible, approach to prediction that accounts for the differences among the various special cases while exploiting the similarities. The terminology used in this endeavor should be kept as free as possible from highly specialized technical jargon.

Some thoughts related to the development of a unified approach to the general prediction problem are as follows.

1. It is useful to think of a model as a family of (unconditional) joint distributions for w and y obtained by (a) starting with the assumptions implicit in the general prediction problem, (b) possibly treating β and θ as random vectors having a completely specified distribution (as in the Bayesian approach) or a distribution that is specified only up to the values of unknown hyper-parameters (as in the empirical Bayes approach), and (c) possibly introducing assumptions about the form of the (conditional) joint distribution of w and y (given β and θ).
2. In formulating any particular prediction problem as a special case of the general prediction problem—and in deciding on any assumptions about the form of the (conditional) joint distribution of w and y (given β and θ)—it may be helpful to express w and y in terms of other random variables (e.g., random effects and errors) or relatedly to proceed in stages by introducing additional random variables and conditioning on their values.
3. A model may serve to suggest a methodology (consisting, e.g., of point and interval predictors) and also as a basis for evaluating that methodology. Typically, the methodology should be evaluated under more than one model. The properties of the methodology conditional on y, on various functions of y, on β and θ, or on other quantities—as well as its unconditional properties—may be of interest. The ultimate criterion for judging a methodology is whether it possesses the characteristics sought by potential users.
4. A distribution assigned to β and θ may or may not reflect a prior opinion. When prior opinion is hard to quantify or when the results of the analysis are to be communicated to individuals with different prior opinions, it may be preferable to assign to β and θ a noninformative prior distribution, in which

case prior opinion may be incorporated informally, subsequent to the analysis. In any case, the distribution assigned to β and θ should be regarded as part of the model.

5. The posterior distribution of w (i.e., the conditional distribution of w given y) may suggest suitable point and interval predictors. With the possible exception of the term posterior distribution, which might be used in referring to any distribution that is conditional on y, the use of Bayesian jargon should be avoided.

Comment: The Kalman Filter and BLUP

James C. Spall

1. INTRODUCTION

Professor Robinson has given a wide-ranging account of best linear unbiased prediction with an impressive array of examples and applications. In this discussion, however, I will restrict my attention to issues regarding the Kalman filter and BLUP.

For ease of discussion, let us restate the random effects model in state-space form as given in Robinson, Section 6. The unobservable random effects (state) vector, u_t, evolves according to

$$(1.1a) \quad u_t = G_t u_{t-1} + w_t,\ u_0 = 0,\ t = 1, 2, \ldots, n,$$

where w_t is a noise term with mean 0 and covariance matrix W_t, and G_t is the state transition matrix. The second equation in the model relates the state vector to the vector of observables y_t:

$$(1.1b) \quad y_t = F_t u_t + v_t,$$

where v_t is a noise term with mean 0 and covariance matrix V_t, and F_t is the measurement matrix. Equations (1.1a, b) can be expressed in the random effects model form of Robinson by writing

$$y = Zu + e,$$

where $y = (y_1^T, y_2^T, \ldots, y_n^T)^T$, $Z = \text{block diag}[F_1, F_2, \ldots, F_n]$, $u = (u_1^T, u_2^T, \ldots, u_n^T)^T$, and $e = (v_1^T, v_2^T, \ldots, v_n^T)^T$. The covariance matrix for u, G in the notation of Robinson, is a function of G_t and W_t, $t = 1, 2, \ldots, n$. The structure of this covariance matrix allows for recursive algorithms of the Kalman filter/smoother form to be used to form BLUP estimates for the components of u. Incidentally, a slightly confusing point in Robinson, Subsection 6.4, is that it is a Kalman smoother, not filter, that produces the BLUP estimate of u based on data y. What Robinson had in mind, I presume, is the common problem where one is interested in an estimate of u_t based only on data through time t (not through some later time); the Kalman *filter*, of course, is used for this problem. For the remainder of this discussion, I will assume that the filtering problem is the one of interest (although virtually all of the ideas would also apply in the smoothing problem).

A couple of other points are worth noting here. First, Sallas and Harville (1988) address a slightly broader problem than that considered above and by Robinson: namely the estimation of random *and* fixed effects via Kalman filter techniques. Second, as noted by Robinson, the Kalman filter is not entirely due to Kalman. The filter equations were essentially derived by others prior to Kalman, but it was Kalman who crystallized much of the thinking in the area and discovered several key relationships to certain systems-theoretic concepts (see Spall, 1988, for further discussion of this).

In the next two sections, I will discuss two problems that were given fairly light treatment in the Robinson paper, but that are important from the point of view of a practitioner. Section 2 describes some problems associated with constructing uncertainty bounds for the filter estimation error $\hat{u}_n - u_n$ when the noise terms have an unknown distribution (as in the general setting of Robinson, equation 1.1). Section 3 elaborates on the brief discussion of Robinson regarding uncertainty in the model parameters θ.

2. UNCERTAINTY BOUNDS FOR $\hat{u}_n - u_n$ IN DISTRIBUTION-FREE SETTINGS

Robinson presents the formula for the covariance matrix of the BLUP estimation error in Section 1 of his paper, and it is well known that this covariance

James C. Spall is a member of the Senior Professional Staff at The Johns Hopkins University, Applied Physics Laboratory, Laurel, Maryland 20723-6099.