Generalisation of the Information Source Method to Compute Reliabilities in Test Day Models

V. Ducrocq and M. del P. Schneider
UR337 Station de Génétique Quantitative et Appliquée, Département de Génétique Animale, Institut National de la Recherche Agronomique, 78352 Jouy en Josas, France
email: vincent.ducrocq@jouy.inra.fr

Abstract

The information source method of Harris and Johnson (1998a) has been shown to lead to very good approximations of reliabilities in single trait genetic evaluations and for MACE models. The method relies on a formula which combines two independent sources of information, for example information coming from progeny and own performance. A multivariate extension of this formula is proposed and is applied to the computation of reliability matrices in test day models. Its use is illustrated in a small application, which underlines the excellent performance of the approach.

1. Introduction

By definition, the reliability R_i associated to a single trait estimated breeding value (EBV) $\hat{a}_i$ of an indivi-dual i is the squared correlation between $\hat{a}_i$ and the true breeding value a_i. It is a measure of accuracy of the evaluation. $\Re_i$ is a coefficient taking a positive value between 0 and 1. Using classical results from BLUP theory, we have:

$$R_i = \rho^2(\hat{a}_i, a_i) = \text{var}(\hat{a}_i - a_i)\ \text{var}(a_i)^{-1}$$
$$= (\sigma_a^2 - PEV_i)/\sigma_a^2 = 1 - PEV_i/\sigma_a^2$$

where σ_a^2 is the additive genetic variance and PEV_i is the asymptotic prediction error variance for animal i, which under BLUP can be obtained from the mixed model equations (MME) as the diagonal element of the inverse of the MME coefficient matrix corresponding to the a_i equation. In most practical cases, the exact computation of the inverse is not feasible. Therefore, several approximate methods have been proposed in the past to estimate R_i under a single trait animal model (see Harris and Johnson, 1998a, for a short list). These usually involve the absorption into the equation of each animal i of equations corresponding to closely related animals and the equation of a contemporary group effect, excluding any other form of relationships between animals. As a result, these methods are known to almost systematically lead to upward biased estimates of R_i (e.g., Meyer, 1987; Misztal and Gianola, 1988). Multiple trait extensions have also been proposed but with even greater biases.

Harris and Johnson (1998a) proposed a different approach to approximate R_i. They suggested to partition different sources of contributions to the reliability into independent parts that can be simply combined using basic selection theory principles. This approach somewhat mimics the partitioning of mixed model equations into contributions from the animal's own performance, from progeny and from parent average to explain the construction of the final EBV (e.g., Wiggans *et al.*, 1988).

The basic equation used to combine the reliabilities R^x and R^y computed from independent sources x and y is:

$$R^{x+y} = \frac{R^x + R^y - 2R^xR^y}{1 - R^xR^y} \qquad [1]$$

The approach also requires to compute the reliability of a part y of a combined source of information x+y, where x and y are independent (for example, in order to exclude the contribution of animal i out of the contribution of all progeny of the sire of i). A simple algebraic manipulation of [1] leads to:

$$R^x = \frac{R^{x+y} - R^y}{R^{x+y}R^y + 1 - 2R^y} \qquad [2]$$

Two striking features of the information source method are its simple implementation and its excellent performance: reliabilities are nearly unbiased when compared to the ones computed from the actual inverse of the mixed model coefficient matrix. Harris and Johnson (1998b) also proposed an extension to simple multiple trait models such as MACE models, again with very good results. Their approach is used for example to compute reliabilities of EBVs from a multiple trait BLUP evaluation for type traits in France.

In the case of random regression models, the method is not directly applicable because one has to take into account the covariance between the different genetic effects simultaneously affecting a same performance. In a study aiming at computing multiple daughter yield deviations and their associated matrix of effective daughter contributions in test day models, Liu *et al.* (2002, 2004) defined a matrix $\mathfrak{R}_i$ that they called a "reliability matrix" as:

$$\begin{aligned}\mathfrak{R}_i &= \text{var}(\hat{\mathbf{a}}_i - \mathbf{a}_i)\ \text{var}(\mathbf{a}_i)^{-1} \\ &= (\mathbf{G} - \mathbf{C}_i)\mathbf{G}^{-1}\end{aligned} \qquad [3]$$

where $\mathbf{a}_i$ and $\hat{\mathbf{a}}_i$ are the vectors of additive genetic effects and EBVs for animal i; **G** is the genetic (co)variance matrix and $\mathbf{C}_i$ the diagonal block of the inverse of the MME coefficient matrix corresponding to animal i. Note that this expression is also valid for multiple trait models. Cumulating information in a manner to some extent similar to Harris and Johnson (1998a), Liu *et al.* (2004) showed that reliability matrices and EDC matrices can be computed together.

The purpose of this paper is to propose a formal generalization of the information source method to random regression models, using the same concept of reliability matrix as Liu *et al.* (2004).

2. Methodology

2.1 Reliability matrix

First of all, an important drawback of the reliability matrix definition given in [3] is that usually $\mathfrak{R}_i$ is not symmetric even though it is the product of two symmetric matrices. Instead, one can use:

$$\mathfrak{R}_i = \mathfrak{R}_{i(\mathbf{a}_i)} = \mathbf{L}^{-1}(\mathbf{G} - \mathbf{C}_i)\mathbf{L}^{-T} = \mathbf{I} - \mathbf{L}^{-1}\mathbf{C}_i\mathbf{L}^{-T} \qquad [4]$$

where **L** is any matrix such that **LL'=G**, for example its Cholesky factor. Then $\mathfrak{R}_i$ is always symmetric. Note that when **G** is decomposed as a function of its eigenvectors and eigenvalues (**G= UΛU'** where **U** is the matrix of eigenvectors and **Λ** the diagonal matrix of eigenvalues) and if one chooses $\mathbf{L}=\mathbf{U}\mathbf{\Lambda}^{1/2}$ in [4], the reliability matrix can be interpreted as a matrix generalisation of the reliability definition to canonical "traits" (uncorrelated genetic effects, each with a variance of 1). More importantly, if one wants the reliability matrix $\mathfrak{R}_{i(\mathbf{T}\mathbf{a}_i)}$ of a (scalar or vector) function $\mathbf{T}\hat{\mathbf{a}}_i$ of the EBVs $\hat{\mathbf{a}}_i$ - for example with **T** being a matrix with two rows corresponding to the coefficients for a 305d production EBV and for a persistency EBV - it can be simply derived from:

$$\mathfrak{R}_{i(\mathbf{T}\mathbf{a}_i)} = \mathbf{I} - \mathbf{M}^{-1}(\mathbf{T}\mathbf{C}_i\mathbf{T}')\mathbf{M}^{-T}$$

where **MM'** is a decomposition of **TGT'**, or equivalently:

$$\mathfrak{R}_{i(\mathbf{T}\mathbf{a}_i)} = \mathbf{H}\mathfrak{R}_{i(\mathbf{a}_i)}\mathbf{H}' \text{ with } \mathbf{H}=\mathbf{M}^{-1}\mathbf{T}\mathbf{L} \qquad [5]$$

2.2 Basic formula

A matrix analogue of formula [1] will now be described. In the single trait case, expression [1] for two independent sources of information x and y can be rewritten as:

$$\frac{1}{R^{x+y}} = \frac{1-R^xR^y}{R^x(1-R^y)+R^y(1-R^x)}$$

or

$$\frac{1}{R^{x+y}} - 1 = \frac{(1-R^x)(1-R^y)}{R^x(1-R^y)+R^y(1-R^x)}$$

which is equivalent to:

$$\frac{1}{\frac{1}{R^{x+y}}-1} = \frac{R^{x+y}}{1-R^{x+y}} = \frac{R^x}{1-R^x} + \frac{R^y}{1-R^y} \qquad [6]$$

or: $$\frac{1}{\frac{1}{R^{x+y}}-1}=\frac{1}{\frac{1}{\Re^{x}}-1}+\frac{1}{\frac{1}{\Re^{y}}-1} \quad [7]$$

Two straightforward multivariate extensions of [6] and [7] are:

$$\Re^{x+y}\left(\mathbf{I}-\Re^{x+y}\right)^{-1}=\Re^{x}\left(\mathbf{I}-\Re^{x}\right)^{-1}+\Re^{y}\left(\mathbf{I}-\Re^{y}\right)^{-1} \quad [8]$$

which involves non symmetric matrices, and:

$$\left(\left[\Re^{x+y}\right]^{-1}-\mathbf{I}\right)^{-1}=\left(\left[\Re^{x}\right]^{-1}-\mathbf{I}\right)^{-1}+\left(\left[\Re^{y}\right]^{-1}-\mathbf{I}\right)^{-1} \quad [9]$$

where the reliability matrices are always guaranteed to be symmetric.

Using equation [9] where $\Re^{x}$ and $\Re^{y}$ are known matrices of small size, one can compute:

$$\mathbf{\Phi}=\left(\left[\Re^{x}\right]^{-1}-\mathbf{I}\right)^{-1}+\left(\left[\Re^{y}\right]^{-1}-\mathbf{I}\right)^{-1}$$

and $$\Re^{x+y}=\left(\mathbf{\Phi}^{-1}+\mathbf{I}\right)^{-1} \quad [10]$$

Similarly, the expression analogue to [2] is derived from [9] as:

$$\left(\left[\Re^{x}\right]^{-1}-\mathbf{I}\right)^{-1}=\left(\left[\Re^{x+y}\right]^{-1}-\mathbf{I}\right)^{-1}-\left(\left[\Re^{y}\right]^{-1}-\mathbf{I}\right)^{-1} \quad [11]$$

We will also need to transform the contribution $\Re^{x}$ from an own performance into a contribution to the reliability of a parent coming from a progeny. This is simply done replacing $\Re^{x}$ in [9] by:

$$\Re^{x}_{prog(*)}=\frac{1}{4}\Re^{x}. \quad [12]$$

Again, the expressions [9] and [11] are valid for multiple trait models as well as for random regression models.

2.3 Implementation of the information source method

As an example, consider the following random regression model:

$$y_{it}=\left(\sum \text{fixed effects}\right)+cg_{j}+\sum_{k=1}^{n_k}\gamma_k a_k+\sum_{m=1}^{m_k}\kappa_m p_m+e_{it} \quad [13]$$

where y_{it} is the t^{th} record of animal i, cg_j is the j^{th} contemporary group effect (e.g., a test day effect), $\left(\sum \text{fixed effects}\right)$ is a sum of environmental effects which will be ignored in the reliability computation as they are usually precisely estimated with a lot of data, a_k is the k^{th} random additive genetic effect associated with the coefficient γ_k, p_m is the m^{th} random permanent environmental effect associated with the coefficient κ_m and e_{it} is the error term with associated weight ω_{it} (assuming heterogeneous residual variances). Steps similar to Harris and Johnson (1998a) are successively applied:

1) First, the contribution $\left(\Re_{o_i}\right)$ to the reliability coming from the own performance of animal i is calculated and simultaneously cumulated into a contribution to the reliability of its sire $\left(\Re_{prog(sire)}\right)$ and dam $\left(\Re_{prog(dam)}\right)$ from progeny, using expressions [9] and [12]. This is done in sequential order, from the youngest animal to the oldest one.

To compute $\Re_{o_i}$, a matrix block $\mathbf{B_i}$ corresponding to the equations for genetic effects and permanent effects for animal i is constructed, after absorption of the contemporary group contribution:

$$\begin{bmatrix} \mathbf{Z_i'S_i^{-1}Z_i+G^{-1}} & \mathbf{Z_i'S_i^{-1}W_i} \\ \mathbf{W_i'S_i^{-1}Z_i} & \mathbf{W_i'S_i^{-1}W_i+P^{-1}} \end{bmatrix} \quad [14]$$

where $\mathbf{Z_i}$ and $\mathbf{W_i}$ are incidence matrices, $\mathbf{P}$ is the (co)variance matrix for permanent environmental effects, $\mathbf{S_i}$ is the diagonal matrix

resulting from the absorption of contemporary group effects. The elements of $\mathbf{S_i}$ are obtained by first reading the data file and cumulating $\sum\omega_{it}$ for each contemporary group, then using $\omega_{it}\left(1-\omega_{it}\Big/\sum\omega_{it}\right)\sigma_e^{-2}$ as new weight when constructing $\mathbf{B_i}$. When all the blocks $\mathbf{B_i}$ have been constructed, each expression [14] is reduced to a n_k x n_k block $\mathbf{B_i^*}$ by absorbing the permanent environment part into the genetic part and transformed into $\Re_{o_i}$ using an analogue of expression [4]:

$$\Re_{o_i} = \mathbf{I} - \mathbf{L}^{-1}\left(\mathbf{B_i^*}\right)^{-}\mathbf{L}^{-T} \qquad [15]$$

2) After step 1), $\Re_{prog(sire)}$ and $\Re_{prog(dam)}$ only include information from daughters. To include information from grand-progeny and further generations down, the $\Re_{prog(*)}$ matrices are cumulated into their own parents' ones, again from the youngest animal to the oldest one, using [9].

3) At the same time, progeny and own performance contributions are combined together (still using [9]) into a reliability matrix $\Re_{o_i+prog(i)}$.

4) Finally, the pedigree information is added, going from the oldest animal to the youngest. Two steps are needed. First, the parent information $\left(\Re_{sire+prog(sire)}\right.$ for example for a sire) must be made independent from the progeny+own information $\left.\left(\rightarrow \Re^{(-i)}_{sire+prog(sire)}\right)\right.$. This is done using equation [11]. Secondly, the contributions (assumed independent) from the sire and dam of i are added and combined to $\Re_{o_i+prog(i)}$.

3. Numerical application

3.1. Data

The proposed approach was applied to two data sets, with the same model. Data set 1 is a small subset of data set 2, which includes 21,137,289 test-day milk yields from 1,119,201 Montbéliarde cows. Data set 1 included 221,773 TD from 12,659 cows from 30 herds (513 herd-year combinations), daughters of 295 sires with on average 42 daughters (range: 2 to 852). The pedigree files included a total of 23,410 animals for data set 1 and 1,562,349 animals for data set 2.

3.2. Model

The model considered is the one which is likely to be applied for test day evaluations in France in a near future. It is of the same form as model [13] and will not be detailed here. It includes 4 genetic effects and 4 permanent effects. The (co)variance matrices are derived from Druet *et al*. (2003, 2005). They were estimated after fitting a 5th order Legendre polynomial for both genetic and permanent environment effects to first lactation data, reducing the resulting matrix from rank 5 to rank 2. The two eigenvectors computed had a nice interpretation (corresponding to EBVs for average production and persistency) and were used as coefficients of the random effects for second and third lactations in a later analysis including 3 lactations. This led to 6x6 genetic and permanent environment (co)variance matrices each with two small eigenvalues, so a rank reduction to 4 genetic and 4 permanent genetic effects was performed. Note that the first two of these effects relate to first lactation only. In other words, the $\mathbf{B_i^*}$ matrices for cows with only first lactations have their 3rd and 4th rows and columns equal to 0.

The reliability matrix $\Re_i$ was computed for each individual. As an illustration of the use of formula [5], the reliabilities associated with 3 traits - cumulative 305d production in first lactation, and cumulative 305d production and persistency averaged over the first three lactations - were derived from each $\Re_i$.

3.3. Results

For data set 1, it was possible to invert the MME coefficient matrix (of size 147,607 when the only fixed effects considered are the test-day effects). This allowed the computation of "true" reliability matrices using expression [4], in which the Cholesky factor of the (here dense) **G** matrix was used. The information source method

took 6 seconds. Some statistics reflecting the resemblance between "true" and approximated elements of the reliability matrices are given in table 1 for the diagonal elements and in table 2 for off-diagonal ones.

The diagonal elements of the reliability matrices ranged from 0 to 0.992. The approximated "diagonal" reliabilities were nearly unbiased and very highly correlated to the true ones. Only the third term was slightly less satisfying. This may be related to the off-diagonal element (2,3) which is poorly estimated (table 2). Off-diagonal elements are on average very close to 0 and the approximated ones are again nearly unbiased. With the exception of element (2,3) which "connects" information obtained in different lactations (1 vs 2 or 3), the correlation between "true" and approximated off-diagonal elements is high.

The estimated reliabilities for the three illustrative traits presented are also in very good agreement with the corresponding "true" reliabilities.

Table 1. Comparison between the diagonal elements of the "true" and estimated reliability matrices for data set 1 (23410 reliability matrices).

Element or trait	Average true value	Average difference	Correlation	Regression slope
1	0.450 ± 0.226	0.0004 ± 0.0070	0.9996	1.003
2	0.389 ± 0.198	0.0021 ± 0.0063	0.9954	1.006
3	0.335 ± 0.179	-0.0154 ± 0.0252	0.9904	0.966
4	0.295 ± 0.164	-00019 ± 0.0089	0.9985	0.996
305d lact.1	0.434 ± 0.216	-0.0055 ± 0.0142	0.9978	0.990
Average 305d	0.358 ± 0.193	0.0085 ± 0.0017	0.9959	1.019
Average Persistency	0.317 ± 0.170	-0.0047 ± 0.0100	0.9983	0.989

Table 2. Comparison between the off-diagonal elements of the "true" and estimated reliability matrices for data set 1**.**

Element	Average "true" value	Average difference	Correlation	Regression slope
1,2	-0.004 ± 0.020	0.0026 ± 0.0016	0.997	0.968
1,3	-0.015 ± 0.021	-0.0002 ± 0.0126	0.811	0.765
1,4	0..005 ± 0.006	-0.0010 ± 0.0026	0.905	0.833
2,3	-0.033 ± 0.035	0.0199 ± 0.0336	0.306	0.263
2,4	-0.014 ± 0.023	-0.0021 ± 0.0095	0.915	0.880
3,4	-0.037 ± 0.026	0.0040 ± 0.0065	0.971	0.930

For the second data set, the 1,562,349 reliability matrices were computed in 9 minutes and 1 second.

4. Conclusion

Based on the numerical example, the information source method to compute reliabilities seems very attractive both in terms of accuracy and computing time to obtain reliabilities in test-day models. Other applications of the extension proposed involve multiple trait evaluations and models with direct and maternal effects. Other by-products of the method may also be improved, for example the computation of EDC matrices (Liu *et al.,* 2004).

References

Druet, T., Jaffrézic, F., Boichard, D. & Ducrocq, V. 2003. Estimation of genetic parameters for first parity lactation curves of French Holstein cows. *J. Dairy Sci. 86,* 2480 - 2490.

Druet, T., Jaffrézic, F. & Ducrocq, V. 2005. Estimation of genetic parameters for test day records of dairy traits in the first three lactations. *Genet. Sel. Evol. 37,* 257 -271.

Harris, B. & Johnson, D. 1998a. Approximate reliabilities of genetic evaluations under an animal model. *J. Dairy Sci. 81,* 2723-2728.

Harris, B. & Johnson, D. 1998. Information source reliability method applied to MACE. *Interbull Bulletin 17*, 31-36.

Liu, Z., Reinhardt, F. & Reents, R. 2002. The multiple trait effective daughter contribution method applied to approximate reliabilities of estimated breeding values of a random regression test day model for genetic evaluation in dairy cattle. Pages 553-556 in *Proc. 7WCGALP*, Communication 20-15. Montpellier, France.

Liu, Z., Reinhardt, F., Bünger, A. & Reents, R. 2004. Derivation and calculation of approximate reliabilities and daughter yield deviations of a random regression test-day model for genetic evaluation of dairy cattle. *J. Dairy Sci. 87*, 1896-1907.

Meyer, K. 1987. Approximate accuracy of genetic evaluation under an animal model. *Livest. Prod. Sci.* 21, 87-100.

Misztal, I. & Wiggans, G.R. 1988. Approximation of prediction error variance in large scale animal models. *J. Dairy Sci. 71* (Suppl. 2), 27-32.

Van Raden, P.M. 2001. Methods to combine estimated breeding values obtained from separate sources. *J. Dairy Sci. 84* (E. Suppl.), E47-E55.

Wiggans, G.R., Misztal, I. & Van Vleck, L.D. 1988. Implementation of an animal model for genetic evaluation of dairy cattle in the United State. *J. Dairy Sci*. *71* (Suppl. 2), 54-69.