Approximate prediction error covariances among multiple estimated breeding values for individuals

Karin Meyer and Bruce Tier

Animal Genetics and Breeding Unit, University of New England, Armidale NSW 2351, Australia

INTRODUCTION

Today's genetic evaluation schemes involve models comprising multiple, correlated additive genetic effects for each animal. These can be multi-trait (MT) models or random regression (RR) models which model trajectories in traits recorded repeatedly per animal through a set of RR coefficients. Often we are interested in linear functions of the resulting breeding value (EBV) estimates. These may be selection indexes combining EBVs for individual traits. For instance, BREEDPLAN, the Australian genetic evaluation scheme for beef cattle currently considers 22 traits (Johnston *et al.*, 1999). The companion program, BREEDOBJECT (Barwick and Henzell, 1998) provides a range of customised selection indexes from the EBVs generated by BREEDPLAN. For RR models, estimates of the genetic RR coefficients describe the complete trajectory of genetic merit for each animal. EBVs for any point on the longitudinal scale can be obtained by evaluating the regression equations. Hence, like selection indexes, such derived point EBVs are linear functions of multiple, estimated EBVs which are correlated.

When comparing EBVs, we are interested not only in their values but also in how reliable they are. The reliability or accuracy of an EBV depends on its prediction error variance (PEV) relative to the genetic variance. As such, it can be perceived as a statistic summarising the value of the information available in calculating the EBV. If the inverse of the coefficient matrix in the mixed model equations (MME) is known, PEVs can be found directly from the diagonal elements of the inverse. However, direct inversion is generally only feasible for small populations, even if sparse matrix techniques are used. Hence, a variety of methods have been developed for approximating PEVs and the resulting accuracies, which are suitable for large scale genetic evaluation schemes involving millions of animals.

Little attention has been paid to approximating accuracies of linear functions of EBVs. This requires approximation of prediction error covariances (PEC) among individual EBVs as well as PEVs. This paper describes a simple method to approximate both PEVs and PECs simultaneously, developed by Tier and Meyer (2003), extending the widely used method of equivalent number of progeny (ENP) from a single number to a matrix of values for each individual. Examples of approximate reliabilities of linear functions of EBVs for multi-trait and RR models are given and contrasted to theoretical values and simulation results.

METHOD

Prediction error (co)variances between effects in a linear mixed model are given by the corresponding elements of the inverse of the coefficient matrix in the MME, denoted by $\mathbf{C}$. Approximation methods available thus generally attempt to adjust diagonal elements of $\mathbf{C}$ for 'links' with other effects in the model, so that reciprocals of the adjusted diagonals closely resemble the diagonal elements of $\mathbf{C}^{-1}$. Early methods to approximate PEVs for single trait analyses first adjusted diagonals of animals with records for limited subclass sizes, then accumulated adjustments to parents' diagonals for limited information on their progeny, and finally adjusted diagonals of progeny for the adjusted diagonals of their parents, taking care not to double count the animal considered (e.g. Meyer, 1989). Principles involved in the more recent procedures and our method are no different. We need to account for the value of information on each animal provided by its own records, its parents and other known ancestors, its progeny and further descendants, and any collateral relatives. In contrast to previous methods, however, we are approximating the $k \times k$ diagonal block of $\mathbf{C}^{-1}$ for each animal, corresponding to all additive genetic effects fitted.

Multi-trait model. Consider k traits with single records per trait. Let

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e} \qquad (1)$$

denote the multi-trait animal model, with $\mathbf{y}$ the vector of observations, $\mathbf{b}$ the vector of fixed effects, $\mathbf{a}$ the vector of additive genetic values, $\mathbf{e}$ the vector of residual, and $\mathbf{X}$ and $\mathbf{Z}$ the incidence matrices relating observations to effects. Assume that all vectors are ordered according to traits within animal, and

$$Var(\mathbf{a}) = \mathbf{A} \otimes \mathbf{G}_0 \qquad Var(\mathbf{e}) = \sum_i {}^{+}\mathbf{R}_i$$

where $\mathbf{A}$ is the numerator relationship matrix, $\mathbf{G}_0$ and $\mathbf{R}_0$ are the $k \times k$ matrices of genetic and residual covariances among traits, '$\otimes$' denotes the Kronecker product and '$\sum^{+}$ the direct matrix sum. $\mathbf{R}_i$

is the submatrix of $\mathbf{R}_0$ for the $i-$th animal, obtained by deleting rows and columns for missing traits. Assume further that animals are ordered from oldest to youngest, i.e. that elements of **a** for parents always precede those of their progeny.

Random regression model With repeated records per animal in a RR model, we need to expand (1) to include the permanent environmental (PE) effects

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wp} + \mathbf{Za} + \mathbf{e}^* \tag{2}$$

with **y**, **b** and **X** as above, and **p** and **a** the vectors of RR coefficients for animals' PE and additive genetic effects, respectively, and $\mathbf{e}^*$ the vector of residuals. Assume there are k covariables used to model the animals' genetic effects. **W** and **Z** are incidence matrices containing covariables relating regression coefficients to the functions of the continuous scale (time) along which observations have been recorded. Residuals e^*_m represent temporary environmental effects, and are assumed independently distributed with variances σ^2_m.

$$Var(\mathbf{p}) = \mathbf{I} \otimes \mathbf{P}_0 \qquad Var(\mathbf{e}) = Diag\{\sigma^2_m\}$$

Value of observations for an animal. Let $\mathbf{D}_i$, of size $k \times k$, denote the block representing the contribution of records for animal i to information on its own EBVs. This is derived from the data part of the MME.

Multi-trait model. With PE due to the animal included in the residual, $\mathbf{D}_i$ is simply the submatrix of **C** corresponding to the animal's genetic effects

$$\mathbf{D}_i = \mathbf{Z}_i' \mathbf{R}_i^{-1} \mathbf{Z}_i \tag{3}$$

with $\mathbf{Z}_i$ the submatrix of **Z** for the $i-$th animal. Genetic evaluation models generally include some 'contemporary group' effects among the fixed effects fitted, e.g. herd-test day effects for dairy cattle data. (3) does not account for limited subclass sizes. When individual i has few contemporaries, (3) should be modified to be

$$\mathbf{D}_i = \mathbf{Z}_i' \left(\mathbf{R}_i^{-1} - \mathbf{R}_i^{-1} (\mathbf{S}_i^{-1}) \mathbf{R}_i^{-1} \right) \mathbf{Z}_i \tag{4}$$

where $\mathbf{S}_i$ is the block of **C** pertaining to the contemporary groups of which animal i is a member. This discounts the value of observations to accommodate the limited number of contrasts between this animal and others, and is the multi-trait equivalent to replacing "1" by $(n-1)/n$ in a univariate, single record scenario (with n the subclass size).

Random regression model. To obtain the equivalent in the RR model, we need to 'absorb' animals' PE effects into the corresponding genetic effects

$$\mathbf{D}_i = \mathbf{Z}_i' \mathbf{R}_i^{-1} \mathbf{Z}_i - \mathbf{Z}_i' \mathbf{R}_i^{-1} \mathbf{W}_i \left(\mathbf{W}_i' \mathbf{R}_i^{-1} \mathbf{W}_i + \mathbf{P}_0^{-1} \right)^{-1} \mathbf{W}_i' \mathbf{R}_i^{-1} \mathbf{Z}_i \tag{5}$$

with $\mathbf{W}_i$ the submatrix of **W** for animal i. As above, contributions from observations can be discounted using weights $w_m = (n_m - q)/n_m \leq 1$ for the $m-$th record, with n_m the size of subclass to with the record belongs and q the number of 'repeated' records it has in that subclass, i.e. replacing $\mathbf{R}_i$ in (5) with $\mathbf{R}_i^* = Diag\{w_m \sigma^2_m\}$.

Value of observations on descendants. In the second step, we accumulate the values of progeny and other descendants for each animal, processing the pedigree 'upwards', i.e. from youngest to oldest. Conceptually, this is obtained by assuming each progeny has only one parent known and that this parent has no further information, building the MME for the animal and the parent, and then 'absorbing' the animal equations into those of the parent. Let $\mathbf{E}_i$ denote the $k \times k$ block of contributions for animal i and p_i the number of progeny it has.

$$\mathbf{E}_i = \frac{1}{3}\mathbf{G}_0^{-1} - \frac{4}{9}\mathbf{G}_0^{-1} \left(\mathbf{D}_i + \sum_{l=1}^{p_i} \mathbf{E}_l + \frac{4}{3}\mathbf{G}_0^{-1} \right)^{-1} \mathbf{G}_0^{-1} \tag{6}$$

This block is accumulated for both sire and dam of animal i. As the pedigree is processed 'upwards' any blocks $\mathbf{E}_l$ required for progeny of animal i have already been fully determined. (6) is adequate if animal i has been directly contrasted to relatively few half-sibs. If the animal's records were in contemporary groups which included many of its half-sibs, however, $\mathbf{D}_i$ in (6) would give an overestimate of the individual's contribution to its parents. As above, we can discount the information required by weighing contributions with a factor determined by the proportion of sibs in a subclass; see Tier and Meyer (2003) for details.

Value of observations on ancestors. Finally, we accumulate the values of parents, ancestors and collateral relatives for each animal by processing the pedigree from oldest to youngest. However, in the previous step, the value of descendants for all animals was accumulated. Hence the block $\mathbf{E}_j$ for parent j of animal i includes the contribution for i. This has to be removed first to avoid double counting. The adjusted block is

$$\mathbf{E}_j^* = \frac{1}{3}\mathbf{G}_0^{-1} - \frac{4}{9}\mathbf{G}_0^{-1} \left(-\mathbf{E}_i + \mathbf{F}_j + \frac{4}{3}\mathbf{G}_0^{-1} \right)^{-1} \mathbf{G}_0^{-1}$$

where $\mathbf{F}_j$ is the $k \times k$ block for parent j in which contributions from all sources of information has been accumulated. As the pedigree is processed 'downwards', blocks $\mathbf{F}_j$ have always been finalised when the contribution of parent j for animal i to be calculated. The 'final' block $\mathbf{F}_i$ for an animal is

then the sum of blocks for its parents, 'unadjusted' for the animal, the block for the contribution from its own records, and the blocks for its progeny

$$\mathbf{F}_i = \sum_{j=1}^{t_i} \mathbf{E}_j^* + \mathbf{D}_i + \sum_{l=1}^{p_i} \mathbf{E}_l \quad (7)$$

where $t_i = 0, 1,$ or 2 denotes the number of known parents for animal *i*.

Prediction error covariances. Matrices $\mathbf{T}_i$ of approximate PEV and PEC for the *k* genetic values estimated for animal *i* are then obtained as

$$\mathbf{T}_i = (\mathbf{F}_i + \mathbf{G}_0)^{-1} \quad (8)$$

The approximate reliability of a linear function of estimated breeding values for animal *i* is then

$$\rho_i^2 = 1 - \mathbf{k}'\mathbf{T}_i\mathbf{k}/\mathbf{k}'\mathbf{G}_0\mathbf{k} \quad (9)$$

where $\mathbf{k}$ is the vector of index weights or covariables evaluated for a given point along the longitudinal scale.

APPLICATION

Data. Data for RR analyses consisted of weight records for beef cattle from birth to 730 days of age. Data set I comprised records from an experimental herd, weighing animals at monthly intervals. Data set II were all records available for Australian Murray Grey cattle. Data set III comprised all records for 600 day weight (W600), P8 fat depth for heifers/steers (P8-H) and bulls (P8-B), and eye muscle area for heifers/steers (EMA-H) and bulls (EMA-B) for this breed. Table 1 summarises characteristics of the data structure.

Analyses. RR analyses fitted a cubic regression on Legendre polynomials of age at recording for direct genetic, maternal genetic, direct permanent environmental and maternal permanent environmental effects. Variances among RR coefficients and heterogeneous measurement error variances were assumed to be those estimated for Hereford cattle (Meyer, 2002). Fixed effects fitted were contemporary groups (CG) and a quartic regression on LP of age, with CG defined as herd-sex-management group-year/month of weighing subclasses for birth weights, and herd-sex-management group-date of weighing subclasses otherwise, dividing CGs further by applying an "age slicing" of 45 days up to 300 days, and 60 days for higher ages. The multitrait analysis (Data set III) fitted a simple animal model with CGs as fixed effects. All 5 traits were assumed to have moderate heritabilities (0.2-0.3), with the same traits measured on different sexes assumed genetically highly correlated (0.8), P8 measures assumed to have virtually no genetic association with the other traits, and EMA assumed have

Table 1. Data structure.

	Set I	Set II	Set III
No. records	75,829	227,219	47,655
Animals in data	7,305	117,977	28,768
... 1 obs.	800	58,396	17,838
... 2 obs.	545	26,840	2,973
... 3 obs.	158	19,795	7,957
... 4 obs.	271	10,776	0
... $\geq$ 5 obs.	5,531	2,170	0
Ancestors	1,138	55,149	21,659
Groups (CG)	11,417	54,263	7,407

Table 2. Reliabilities (in %) of RR coefficients.

	interc.	linear	quadr.	cubic
Data set I				
Simulation	38.2	34.4	23.8	21.7
'True'	38.6	34.6	23.8	21.8
Approximation	37.8	33.9	24.4	22.8
$\beta_{T.A}$	0.999	0.982	0.954	0.949
R^2(%)	95.3	95.0	94.6	94.6
Data set II				
Simulation	26.5	21.5	9.3	8.7
'True'	28.0	23.0	11.0	10.5
Approximation	27.8	22.2	9.7	9.5
$\beta_{T.A}$	0.956	0.950	0.939	0.906
R^2(%)	88.6	85.1	69.4	71.5

have a low genetic correlation (0.4) with W600. Covariance matrices used are given in Tier and Meyer (2003). The index used assumed equal emphasis for all traits, i.e. $\mathbf{k}' = (1\ 1\ 1\ 1\ 1)$.

Measures of reliability. Approximate PECs of estimated genetic RR were calculated for all animals as described above, amalgamating maternal covariances with permanent environmental components. From these, approximate reliabilities of RR coefficients and EBVs for weights at birth, 200, 400 and 600 days were determined. Results were contrasted to approximate reliabilities computed using the procedure of Jamrozik *et al.* (2000), and 'true' reliabilities obtained from the inverse of the coefficient matrix in the MME using a Gibbs sampling algorithm as described by Harville (1999) to estimate the diagonal blocks for $\mathbf{C}$ required, drawing 400,000 Gibbs samples and discarding the first 20,000 samples as burn in. In addition, empirical reliabilities for the data sets considered were available from a simulation study (Meyer, 2003), which calculated reliability as the square of the correlation between true and estimated breeding value across all animals.

Criteria. Method were compared by contrasting means across all animals and by linear regression of theoretical values obtained from $\mathbf{C}^{-1}$ on their counterparts from approximations.

Table 3. Reliabilities (in %) of EBVs.

	0 d	200 d	400 d	600 d
Data set I				
Simulation	40.8	30.8	36.1	37.6
'True'	41.5	34.8	36.7	38.1
Approximation	43.6	34.2	36.0	37.2
$\beta_{T.A}$	0.936	1.017	1.004	0.997
R^2(%)	96.0	94.6	95.9	95.8
Jamrozik	45.3	37.7	38.6	39.7
$\beta_{T.J}$	0.900	0.912	0.952	0.962
R^2(%)	96.8	95.9	97.0	96.9
Data set II				
Simulation	28.7	24.7	26.0	26.4
'True'	30.3	26.2	27.5	27.8
Approximation	31.6	26.5	27.3	27.5
$\beta_{T.A}$	0.911	0.955	0.954	0.954
R^2(%)	90.2	88.5	89.5	89.0
Jamrozik	34.9	31.0	31.1	31.5
$\beta_{T.J}$	0.894	0.885	0.925	0.923
R^2(%)	88.8	86.9	88.4	87.8

Table 4. Reliabilities (in %) of EBVs and index.

	W600	P8-H	P8-B	EMA-H	EMA-B	Index
'True'	31.5	18.7	17.1	19.1	17.9	30.2
Approx.	28.2	16.4	16.2	15.5	14.9	27.2
$\beta_{T.A}$	0.957	0.951	0.951	0.960	0.978	0.970
R^2(%)	91.4	84.6	84.0	90.3	88.7	90.2

Results. Reliabilities for estimates of RR coefficients for data set I and II are summarised in Table 2, together with the regression of 'true' on approximate values ($\beta_{T.A}$) and the corresponding coefficient of determination (R^2). On the whole, there was good agreements between approximate and theoretical values. Whilst mean approximate values were slightly lower than 'true' values, regression coefficients were less than unity, as the approximation procedure tended to overestimate reliabilities for high accuracy animals. Simulation results tended to be lower than their expectations, especially for data set II, but with empirical standard deviations between 2.0 and 2.7% differences were not significant. R^2 values were around 95% for all RR coefficients for data set I with an average of 10.4 records per animal, but were dramatically lower for data set II, in particular the quadratic and cubic coefficients. With an average of 1.9 records per animal and few animals with 4 or more records for this data set, this was not surprising. Doubling the amount of data by adding a fictitious record about 100 days after each actual one increased R^2 for the quadratic and cubic coefficients to just over 80%.

Corresponding statistics for EBVs at individual ages are given in Table 3. With the intercept and linear coefficients dominating the linear function, differences between individual ages were small. Again there was good agreement between approximate reliabilities and their 'true' values. Approximations using Jamrozik's method were generally higher than those from our method, resulting in lower regression coefficients of 'true' on approximate. R^2 values for the two approximations were comparable, however, with the former performing slightly better for data set I and copying somewhat less well with the bad structure of data set II.

Results for the multi-trait analysis are given in Table 4. With most animals having W600 records, but only about 25% of animals having both scan traits recorded as well, reliability of the index was largely determined by that of W600. Other indexes examinded (not shown) yielded similar results.

CONCLUSIONS

Reliabilities of linear functions of estimated breeding values and hence prediction error covariances can be satisfactorily approximated for data structures typical for beef cattle. The approximation procedure described is computationally undemanding and applicable to large scale problems.

ACKNOWLEDGMENTS

This work was supported by Meat and Livestock Australia Ltd under grant BFGEN.100.

REFERENCES

Barwick S. and Henzell A., 1998. Breedobject: Breeding objective and indexing software for beef breeding. Proc. Sixth World Congr. Genet. Appl. Livest. Prod., Vol. 27: 445–446

Harville D.A., 1999. Use of the Gibbs sampler to invert large, possibly sparse, positive definite matrices. Lin. Alg. Appl. 289:203–224

Jamrozik J., Schaeffer L. and Jansen G.B., 2000. Approximate accuracies of prediction from random regression models. Livest. Prod. Sci. 66:85–92

Johnston D.J., Tier B., Graser H.U. and Girard C., 1999. Presenting BREEDPLAN version 4.1. Proc. Ass. Advan. Anim. Breed. Genet. 13:193–196

Meyer K., 1989. Approximate accuracy of genetic evaluation under an individual animal model. Livest. Prod. Sci. 21:87–100

Meyer K., 2002. Estimates of covariance functions for growth of Australian beef cattle from a large set of field data. CD-ROM Seventh World Congr. Genet. Appl. Livest. Prod. Communication No. 11–01

Meyer K., 2003. Scope for a random regression model in genetic evaluation of beef cattle for growth. Livest. Prod. Sci. 00:000–000. (in press)

Tier B. and Meyer K., 2003. Approximating prediction error covariances in multiple-trait and random regression models. J. Anim. Breed. Genet. 00:000–000. (submitted)