Approximation of reliabilities for random-regression single-step genomic best linear unbiased predictor models

M. Bermann,1 I. Aguilar,2 A. Alvarez Munera,1 J. Bauer,3 J. Šplíchal,3 D. Lourenco,1 and I. Misztal1

Abstract: Random-regression models (RRM) are used in national genetic evaluations for longitudinal traits. The outputs of RRM are an index based on random regression coefficients and its reliability. The reliabilities are obtained from the inverse of the coefficient matrix of mixed model equations (MME). The reliabilities must be approximated for large data sets because it is impossible to invert the MME. There is no extensive literature on methods to approximate the reliabilities of RRM when genomic information is included by single-step GBLUP. We developed an algorithm to approximate such reliabilities. Our method combines the reliability of the index without genomic information with the reliability of a GBLUP model in terms of effective record contributions. We tested our algorithm in the 3-lactation model for milk yield from the Czech Republic. The data had 30 million test-day records, 2.5 million animals in the pedigree, and 54,000 genotyped animals. The correlation between our approximation and the reliabilities obtained from the inversion of the MME was 0.98, and the slope and intercept of the regression were 0.91 and 0.02, respectively. The elapsed time to approximate the reliabilities for the Czech data was 21 min.

Random-regression models (RRM) are used for national dairy cattle genetic evaluations in many countries (Interbull, 2022). Typically, the traits for which RRM are fitted comprise production traits (milk, fat, and protein yield), somatic cell score (SCS), conformation traits, and feed intake, among others (Schaeffer, 2004). RRM have been used for traits in other species, such as growth in beef cattle (Meyer, 2005), residual feed intake in chickens (Emamgholi Begli et al., 2018), and pigs (Wang et al., 2022).

The method of choice to include genomic information in RRM is single-step genomic best linear unbiased predictor (ssGBLUP) (Aguilar et al., 2010; Oliveira et al., 2019). Many studies successfully applied ssGBLUP with RRM in dairy cattle (e.g., Koivula et al., 2015; Kang et al., 2017; Oliveira et al., 2019; Wang et al., 2022). RRM with ssGBLUP might not be fully applied for routine genetic evaluations because no validated method exists to approximate the reliability of EBV for these models. Reliabilities for ssGBLUP based on RRM have been approximated by multiple-trait repeatability model reliabilities (Bauer et al., 2015). Nonetheless, a publication in the Interbull bulletin (Alkhoder et al., 2022) recognized the need for a method to approximate reliabilities for RRM with ssGBLUP.

Although different methods can approximate reliabilities in single- and multiple-trait ssGBLUP models (Liu et al., 2017; Edel et al., 2019; Bermann et al., 2021; Ben Zaabza et al., 2022), they are not easily extended to RRM. This is because of the need to calculate the PEC between different time points to obtain a final value of PEV and the need to process the information for the different random regression coefficients altogether. Although the methods that do not consider genomic information approximate the PEC, they cannot be extended to account for genomic information because of the loss of sparsity of the model. Thus, this study aimed to develop an efficient method to approximate reliabilities for RRM with ssGBLUP.

In matrix notation, an animal random-regression multiple-trait model is:

\[ y = Xb + Za + Qp + e, \]  

[1]
where \( y \) is the vector of phenotypes, \( b \) is a vector containing sets of fixed effects independent and dependent on the time scale, \( a \) is the vector of additive genetic random regression coefficients, \( p \) is the vector of permanent environmental effect, \( e \) is the error vector, and \( X \) is an incidence matrix, and \( Z \) and \( Q \) are matrices of random-regression coefficients. It is assumed that:

\[
E[y] = Xb
\]

\[
\text{Var}[p] = \begin{pmatrix} A \otimes G_0 & 0 & 0 \\ 0 & I \otimes P_0 & 0 \\ 0 & 0 & I \otimes R_0 \end{pmatrix},
\]

where \( A \) is the numerator relationship matrix, \( G_0 \) is the covariance matrix of the additive genetic random regression coefficients, \( P_0 \) is the covariance matrix of the permanent environmental random regression coefficients, and \( R_0 \) is the covariance of the error among traits. Given Eq. (1) and (2), the MME for the RRM are:

\[
\begin{align*}
X'X^{-1} & X'Z + A^{-1} G_0^{-1} Z'R^{-1}X' \\
Z'R^{-1}X & Z'R^{-1}Z + A^{-1} G_0^{-1} Z'R^{-1}Q \\
Q'R^{-1}X & Q'R^{-1}Z + I \otimes P_0^{-1} \end{align*}
\]

where the structure of \( H^{-1} \) is described in Aguilar et al. (2010).

To derive a general expression for the reliability of an animal’s EBV, let \( k' \) be a linear function defining the trait of interest, for example, 305 d for milk yield and \( a \) be the vector of additive genetic random regression coefficients for the animal \( i \). Then, the breeding value of the \( i \)th animal is \( u_i = k'a_i \), and its respective EBV is \( \hat{u}_i = k'\hat{a}_i \). Note that \( \hat{u}_i \) could be an index combining different traits. For instance, \( \hat{u}_i \) could be a weighted average over different lactations of the cumulative 305 d EBV for milk yield. The reliability of \( \hat{u}_i \) is:

\[
\text{rel}_i = \frac{\text{Var}(\hat{u}_i)}{\text{Var}(u_i)} = \frac{k'C'k}{k'(V \otimes G_0)_i k},
\]

where \( V \) can be either \( A \) or \( H \) and \( C_i \) is the \( i \)th diagonal block of:

\[
C = V^{-1} \otimes G_0^{-1} + Z'R^{-1}Z - Z'R^{-1}X Z'R^{-1}Q X'R^{-1}X + Q'R^{-1}Q + I \otimes P_0^{-1}.
\]

Eq. (6) can be identified as the inverse of the coefficient matrix of a multiple-trait model whose covariance structure for the additive genetic random effects is \( V \otimes G_0 \) and \( M \) is the matrix resulting from the absorption of all the effects except the additive genetic random effect. Evidently, with genomic information, one could treat Eq. (6) as a multiple-trait model and approximate its reliabilities with an existing method. However, this would be computationally demanding. Instead, we assume there exists a single-trait linear model \( y^* = X'b^* + Z'u^* + e^* \) such that \( \text{Var}(u^*) = H \) and the reliability of \( u^*_i \) is:

\[
\text{rel}_i^* = 1 - \frac{W^i}{H_i} \approx \text{rel}_i,
\]

where \( W^i \) is the \( i \)th diagonal element of \( W = \text{Var}(u^* - \hat{u}^*) \) and can be approximated as:

\[
W^i \approx (H^{-1} + D)^{-1},
\]

For ssGBLUP, the numerator relationship matrix in Eq. (2) is replaced by \( H \) as defined by Legarra et al. (2009), and the resulting MME are:

\[
\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}Q \\ Z'R^{-1}X & Z'R^{-1}Z + H^{-1} G_0^{-1} & Z'R^{-1}Q \\ Q'R^{-1}X & Q'R^{-1}Z & Q'R^{-1}Q + I \otimes P_0^{-1} \end{pmatrix} \begin{pmatrix} b \\ \hat{a} \\ \hat{p} \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ Z'R^{-1}y \\ Q'R^{-1}y \end{pmatrix}
\]

and the structure of \( H^{-1} \) is described in Aguilar et al. (2010).

1. For all the animals, calculate \( \text{rel} \), based on an RRM without genomic information using, for instance, the method of Tier & Meyer (2004).
2. Back-solve \( \text{rel} \), to obtain \( D \). Different back-solving methods will produce slightly different reliabilities. We chose to back-solve \( \text{rel} \), by reversing the method of Harris and Johnson (1998). A detailed description of the algorithm is in the Supplementary data of Bermann et al. (2021).
3. Approximate the single trait ssGBLUP reliabilities \( \text{rel}^*_i \) following a standard method for ssGBLUP reliabilities.
phenotypes. The pedigree, and 2892 genotyped animals, from which 1127 had
reduced data set had 111,494 test-day records, 44,582 animals in
inversion of the coefficient matrix of mixed model equations. The
pare the approximate reliabilities against those obtained from the
cows also had phenotypes. A reduced data set was created to com-
set. It can be noticed that the gain in reliability is larger for young
and without genomics for the genotyped animals in the large data
(see; Liu et al., 2017; Edel et al., 2019). Namely, the main
steps are:

a. Calculate reliabilities for a GBLUP or SNP-BLUP model using weights from D. The choice between
GBLUP and SNP-BLUP would depend on the number of genotyped animals and the number of SNP.
b. Obtain pedigree reliabilities for genotyped animals without considering information from non-genotyped
animals. This step can be done with step 1 or by nullifying the weights in D corresponding to non-geno-
typed animals and re-calculating reliabilities with the method of Harris & Johnson (1998) or Tier & Meyer
(2004).
c. Remove double counting in terms of effective record contributions and obtain final ssGBLUP reliabilities
for genotyped animals.
d. Propagate the information to non-genotyped animals.

A 3-lactation model from the Czech Republic was used for test-
ing the proposed method. The full data set had 30,366,184 test-day
records for milk yield across 3 lactations. The number of records
for each lactation was 13,991,590, 10,116,514, and 6,258,080,
respectively. The pedigree included 2,518,681 animals from which
54,221 had available genotypes for 38,386 SNP. 8054 genotyped
cows also had phenotypes. A reduced data set was created to com-
pare the approximate reliabilities against those obtained from the
inversion of the coefficient matrix of mixed model equations. The
reduced data set had 111,494 test-day records, 44,582 animals in
the pedigree, and 2892 genotyped animals, from which 1127 had
phenotypes.

The proposed method was tested for the average of the 3 lacta-
tions of cumulative 305 d EBV based on comparing the approxi-
mated and the exact reliabilities (i.e., calculated from the inverse
of the MME). The exact reliabilities were calculated based on the
PEV of the additive random regression coefficients obtained with
the OPTION store_pev_pec from BLUPF90+ (Lourenco et al.,
2022). The approximated reliabilities without genomic informa-
tion (step 1 from P1) were obtained with the method of Tier &
Meyer (2004) using ACCF90RR. Then, those reliabilities were
given to ACCF90GS2 (Bermann et al., 2022) to perform steps 2
and 3 from P1.

Figure 1 shows the comparison between the reliabilities ob-
tained by the inversion of the coefficient matrix of the mixed
model equations and the approximated ones for the reduced data
set. The correlation between the exact and the estimated value was
0.98, whereas the slope and intercept of the regression of exact on
estimated reliabilities were 0.91 and 0.02, respectively. For young-
genotyped animals without records and progeny with records, the
correlation was 0.93, and the slope and intercept were 0.94 and
0.06, respectively. These values are acceptable for routine evalua-
tions and matched previous studies on approximation of reliabilities
(Ben Zaabza et al., 2023). The observed underestimation in Figure
1 was because of young and inbred genotyped animals without
records, for which the reliability without genomic information was
underestimated. Figure 2 compares approximated reliabilities with
and without genomics for the genotyped animals in the large data
set. It can be noticed that the gain in reliability is larger for young
animals, which have low pedigree reliability. The animals that did
not show an increase when adding genomics are non-genotyped
animals, animals from commercial herds that are poorly connected
with the population and have missing parents, or with diagonals of
the genomic relationship matrix lower than the numerator relation-
ship matrix. The approximation of all the single-step reliabilities
for the large data set took 21 min.

The member countries of Interbull include MACE proofs in
their official dairy evaluations to account for foreign information.
Although the Czech Republic is a member of Interbull, neither
Bauer et al. (2015) nor we included MACE proofs in our study.
The reason for this is that MACE proofs and test-day models are
on a different scale, which adds another layer of complexity to
the approximation of reliabilities. Boerner et al. (2023) presented
a method to include MACE proofs directly into test-day models
using de-regression and weights in terms of effective daughter
contributions. However, they did not provide details on the ap-
proximation of reliabilities. Future research with the Czech data
will include approximating reliabilities with foreign data.

Bauer et al. (2015) approximated the reliability of a single-step
RRM using a multiple-trait repeatability model for Czech Holstein
cows. The authors considered the first 3 lactations as separate traits
and reported EBV obtained by averaging the values of 300 d of
lactation over the 3 lactations. For each lactation, they approxi-
mated reliabilities for a single-trait model with a permanent envi-
ronmental effect without genomic information using the method
of Misztal et al. (1993). Then, they calculated ERC and obtained
reliabilities for genotyped animals utilizing the method of Misztal
(2013). After that, they adjusted reliabilities for non-genotyped
animals. Finally, they calculated multiple-trait reliabilities with

Figure 1. Comparison between the reliability obtained from the inverse of
the mixed model equations and the approximated reliability for the small
data set. The correlation (cor), intercept (b0), and slope (b1) of the regres-
sion are reported. The red-dashed line shows a line of slope equal to one and
intercept equal to zero.

Bermann et al. | Approximation of reliabilities for random-regression genomic models

JDS Communications TBC; TBC
the method of Stráb et al. (2001) and used them to approximate the reliability of EBV. The authors did not compare the approximated reliabilities with those obtained by inverting of the MME. However, they observed that the reliability of young bulls without progeny increased on average by 0.02 when including genotypes for 2,236 animals. In our study, we observed an average increase of 0.42 for young bulls with a reference population of 54,221 animals. The main difference between the method of Bauer et al. (2015) and ours is that we do not approximate the RRM as a multiple-trait repeatability model. Instead, we aim to find a single-trait model with similar EBV reliabilities as the EBV calculated from an RRM.

In other words, while Bauer et al. (2015) incorporate the genomic information while calculating the EBV reliability, we include the genomic information as a post-processing step after obtaining the reliability of the index without genomic information.

In light of Eq. (6), RRM are similar to multiple-trait models but with denser incidence matrices. Therefore, one could use a method to approximate reliabilities for ssGBLUP multiple-trait models in RRM (Bermann et al., 2021; Ben Zaabza et al., 2022). However, this would lead to a high computational cost, and how to approximate prediction error covariances between effects with genomics is unclear. The method of Bermann et al. (2021) could be implemented by treating each regressor as a single trait, approximating their reliabilities, and adjusting them for multiple traits as a post-processing step. This would require calculating \( n_{\text{trait}} n_{\text{coef}} \) single-trait model reliabilities for all the animals, where \( n_{\text{trait}} \) is the number of traits, and \( n_{\text{coef}} \) is the number of additive genetic random regression coefficients per animal. Since the most expensive task is to invert a single-trait GBLUP model, the computational complexity of this approach would be dominated by \( O \left( n_{\text{trait}} n_{\text{coef}} n_{\text{gen}}^3 \right) \), being \( n_{\text{gen}} \) the number of genotyped animals. For example, for the Czech Holstein dairy cattle evaluation, this implies calculating 12 single-trait reliabilities, which translates into inverting 12 times the genomic relationship matrix. The method of Ben Zaabza et al. (2022) can approximate the reliabilities for the additive random regressors altogether. However, this implies inverting a multiple-trait GBLUP model of size \( n_{\text{trait}} n_{\text{coef}} n_{\text{gen}}^3 \), which translates into a complexity dominated by \( O \left( n_{\text{trait}} n_{\text{coef}} n_{\text{gen}}^3 \right) \). For the data used in this manuscript, this implies inverting a matrix of size equal to 650,652 rows and columns. If a model does not include a residual polygenic effect, the multiple-trait GBLUP might be replaced with an equivalent SNP-BLUP model, leading to a complexity of \( O \left( n_{\text{trait}} n_{\text{coef}} n_{\text{gen}}^3 \max \left( n_{\text{snp}}^2 n_{\text{gen}} n_{\text{gen}}^2 \right) \right) \), where \( n_{\text{snp}} \) is the number of SNP. Since the running time for approximating reliabilities without genomic information goes from seconds to minutes, the difference in running times between the method of Bermann et al. (2021) and Ben Zaabza et al. (2022) would depend on how each method deals with genomic information. In other words, the difference between both methods would depend on whether the inversion of a \( n_{\text{trait}} n_{\text{coef}} n_{\text{gen}}^3 \times n_{\text{trait}} n_{\text{coef}} n_{\text{gen}}^3 \) matrix is faster than inverting a \( n_{\text{trait}} n_{\text{coef}} \) times a matrix of size \( n_{\text{gen}} \times n_{\text{gen}} \). This depends on the machine and level of parallelism allowed; however, the latter is usually faster. The method proposed in this manuscript only requires calculating one single-trait reliability. The asymptotical complexity of the method is \( O \left( n_{\text{gen}}^3 \right) \). If the Algorithm for Proven and Young (APY; Misztal 2016) is used, that complexity will drop up to \( O \left( \max \left( n_{\text{proven}}^2 n_{\text{young}}^2 n_{\text{proven}}^2 \right) \right) \), where \( n_{\text{proven}} \) is the number of proven animals, which is usually between 5000 and 35,000 animals (Pocrnic et al., 2016), and \( n_{\text{young}} \) is the number of young animals, which is \( n_{\text{young}} = n_{\text{gen}} - n_{\text{proven}} \).

We developed a procedure for approximating the reliabilities of EBV for RRM with ssGBLUP. The proposed method is computationally efficient and accurate compared with the reliabilities obtained from the inversion of the MME. Thus, our approach can be implemented for routine genetic evaluations with RRM based on ssGBLUP.

**References**


Bermann, M., D. Lourenco, A. Cesaranì, and I. Misztal. 2022. ACCF09GS2: software for fast approximation of reliabilities of estimated breeding values
in single-step GBLUP. In: Proceedings of the world congress on genetics applied to livestock production: 3–8 July; Rotterdam 2022.


Notes

Funding information This study was partially funded by Agriculture and Food Research Initiative Competitive Grant no. 2020-67015-31030 from the US Department of Agriculture's National Institute of Food and Agriculture.

Institutional Review Board or an Animal Care and Use Statement This article does not contain any studies with human or animal subjects and did not require [IACUC/IRB] approval.

Conflicts of interest The authors declare that they have no conflicts of interest.