An approach to reduce computing time in multi-trait single-step evaluations

L.-H. Maugan*[1]*, T. Tribout and V. Ducrocq

Authors' address: GABI 211, INRAE, 78322 Jouy-en-Josas
[1] *Corresponding author: laure-helene.maugan@inrae.fr*

Abstract

Multi-trait (MT) single-step (SS) evaluations can be very time-consuming, in particular because of slow convergence. An alternative consists in first running univariate SS evaluations for a limited number of iterations in order to provide reasonable starting values for GEBV or SNP effects as well as corrected phenotypes for fixed effects. This initial step can also include other features to improve the evaluation model such as a correction for heterogenous residual variances. Then a greatly simplified multi-trait SS evaluation can be implemented.

Such a strategy was successfully implemented on a Montbéliarde dataset with ~1.6 million animals with performance on a group of 8 correlated type traits.

Key words: Multiple-trait, Single-Step, convergence, genomic evaluation

Introduction

When Single-Step (SS) evaluation was mentioned for the first time, it was considered as a solution to avoid bias in evaluation due to genomic preselection (Legarra and Ducrocq, 2012). SS was proposed by Aguilar et al. (2010), Misztal et al. (2009) and Legarra et al. (2009). SS evaluations are nowadays more and more considered as the standard for genetic evaluation when selection strongly relies on genomic information of young animals (Aguilar et al., 2010). SS evaluations required new software development to be feasible on large populations (Legarra and Ducrocq, 2012). Such software now exist but nevertheless, SS often remains a complex and time-consuming evaluation, in particular due to slow convergence when there is a very large number of genotypes available, when unknown parent groups are included or when the evaluated traits have low heritability or when correlations among traits are large (Mäntysaari et al., 2020). As a consequence, univariate models (or multivariate models with traits described by the same model) are often preferred because of their simplicity. This restriction leads to drawbacks because information from correlated traits is lost and/or simplified assumptions are made (e.g., ignoring heterogeneity of residual variance). Efficient software to perform multivariate SS evaluations exist but most often used the same underlying model for all traits.

An alternative to reduce these drawbacks – loss of information from correlated traits or simplified underlying assumptions - is to perform the multivariate SS evaluation in two steps, what we will call a "combined SS evaluation", mimicking the "combined genetic evaluation" used in France since 2001 for all dairy breeds to derive the EBVs used in the Total Merit Index with Estimated Breeding Values (EBV) for all traits for all animals (Ducrocq et al., 2001). This approach is also used routinely to improve the accuracy of genetic evaluation for Functional Longevity (FL) by adding information from traits correlated to FL (Govignon-Gion et al., 2012, 2016). In practice, the first step consists of performing univariate evaluations adapted to the trait (i.e. with an accurate modelling) in order to derive adjusted phenotypes (e.g.

corrected for fixed effects and/or heterogeneity of variances) with an associated weight. The second step is a multivariate SS evaluation of corrected and weighted performances with a unique simplified model. The expected benefits are an improved accuracy for less heritable traits, a proper correction for selection on correlated traits and hopefully a decrease in overall computing time compared to a sophisticated unique Multiple trait SS evaluation. In this preliminary work, our focus will be on the number of iterations to converge per run.

Materials and Methods

Data

As a case study, we first considered a simple situation: a group of 8 correlated type traits in the Montbéliarde breed. These traits are collected together by the same technician and described by the same model (Table 1). The first 7 traits are measures (in cm) and the 8th (Rump Angle) is evaluated as a score on a scale from 1 to 9.

Performances of 1,618,843 animals collected between 1991 and 2018 were analyzed. The pedigree file included 2,062,087 animals and 70,580 of them were genotyped on 53,471 SNP.

Table 1. – Heritability (**diagonal**) and genetic correlation (off-diagonal) between traits

	HS	CW	CD	BD	RL	WH	TW	RA
HS	**0.63**							
CW	0.38	**0.33**						
CD	0.66	0.51	**0.46**					
BD	0.62	0.53	0.82	**0.42**				
RL	0.75	0.52	0.70	0.59	**0.40**			
WH	0.59	0.59	0.64	0.56	0.67	**0.39**		
TW	0.60	0.63	0.61	0.59	0.62	0.79	**0.38**	
RA	-0.12	0.10	0.09	0.05	0.03	0.23	0.35	**0.36**

HS=Height at Sacrum; CW=Chest Width; CD=Chest Depth; BD=Body Depth; RL=Rump Length; WH=Width at Hips; TW=Width at Thurl; RA=Rump Angle;

Three different cases were considered:

1) an ideal case with all performances available,

2) a case mimicking the inclusion of new traits with time: compared with case 1, performances for traits 1 and 2 were erased for the 75% oldest cows; then performances for traits 3 and 4 were erased for the 50% oldest cows, then performances for traits 5 and 6 were erased for the 25% oldest cows. Only the 25% youngest cows had a phenotype available for the 8 traits.

3) a case mimicking the absence of some phenotypes for younger animals: compared with case 1, all cows in the oldest 25% group had records on the 8 traits, the 3 next 25% groups had records on, respectively, traits 3 to 8, traits 5 to 8 and traits 7 and 8.

In other words, case 2 simulates a situation where information is not available for the older animals (e.g., due to introduction on new type traits) while case 3 describes a hypothetic situation where animals are young and don’t have any performances yet.

Only the results for the third situation will be shown.

Model and software

Two approaches were compared: 1) a standard Multi-trait SS evaluation and 2) a combined SS evaluation.

Approach 1 used the following multivariate evaluation model:

$$\mathbf{y}_{\mathbf{kj}} = \sum_{1}^{3} \boldsymbol{b_k} + \mathbf{a}_{\mathbf{j}} + \mathbf{e}_{\mathbf{kj}}, \quad \text{(a)}$$

with $\mathbf{y}_{\mathbf{kj}}$ the animal performance of animal j for trait k, $\mathbf{b}_{\mathbf{k}}$ the fixed effects for trait k (class of age at calving x year (504 levels), class of lactation stage x age (616 levels) and herd x day x classifier (173,961 levels), $\mathbf{a}_{\mathbf{j}}$ is the additive genetic effect of animal j (2,062,087 animals) and $\mathbf{e}_{\mathbf{kj}}$ is the residual.

Approach 2 used 2 different models. In a first step, the model is the same as approach 1 (a) but for single trait evaluations.

Then, in a second step, the following simplified multivariate SS evaluation model (b) was used:

$$y_{kj\_corrected\_for\_fixed\_effects} = \mu_{year} + a_j + e_{kj}, \qquad (b)$$

where $y_{kj\_corrected\_for\_fixed\_effects}$ the animal performance corrected for fixed effects of step 1, μ_{year} a ‘year’ effect included to allow a correction of the annual genetic trend when selection on correlated traits is considered.

In each model, genetic groups of unknown parents were taken into account but are included in $\mathbf{a_j}$ for simplicity of writing.

For the combined SS evaluation, intermediate results indicated that it was not necessary to choose a very stringent convergence criterion for step 1 (data not shown). Instead, a maximum of 500 iterations was chosen for step 1 followed by 2,500 iterations for step 2. For approach 1, a maximum of 3,000 iterations was chosen.

(G)EBVs were compared to results at convergence. Convergence at iteration i was measured through 3 criteria: correlations with final results and means and standard deviations of the absolute difference between current solutions and final results

The software used in this study is HSSGBLUP developed at INRAE (Tribout et al., 2020).

Results & Discussion

Figures 1 and 2 present the mean and standard deviation of the absolute difference between GEBV at iteration i and final iteration for approach 1 and 2. Correlations between GEBV at iteration i and final iteration are shown in the Figure 3.

Figures 1 and 2 show a faster convergence to final GEBV with approach 2 compared to approach 1;

In Figure 3, all traits show a correlation between current and final GEBV larger than 0.95 only after 2800 iterations for approach 1 (Figure 3a) while this happens after only 1300 iteration for approach 2 (Figure 3b).

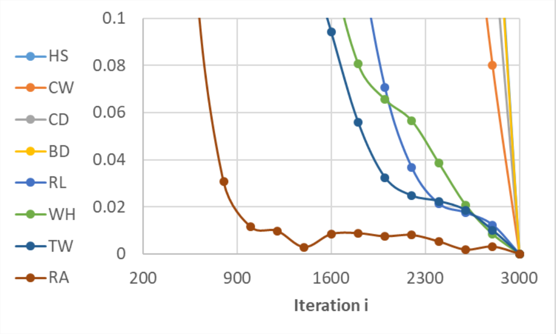

Figure 1a. Mean of the absolute difference between GEBV at iteration i and at convergence for approach 1. Values above 0.1 are not shown.

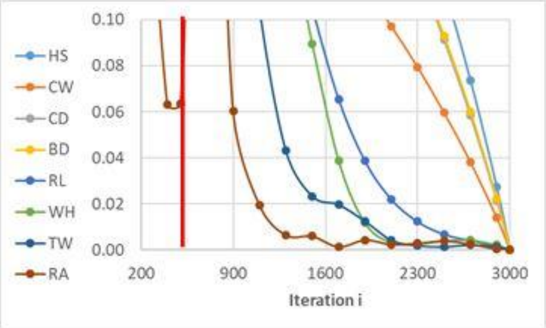

Figure 1b. Mean of the absolute difference between GEBV at iteration i and at convergence for approach 2. Values above 0.1 are not shown. The red vertical line indicates the 500th iteration when step 1 was stopped and step 2 started in approach 2.

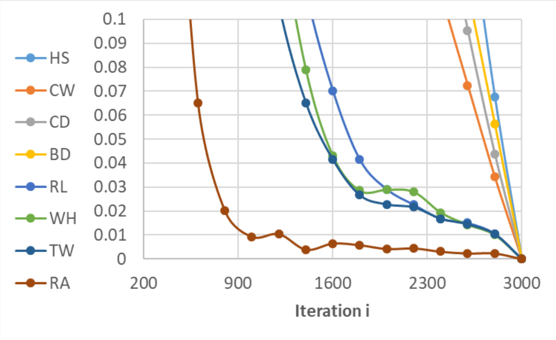

Figure 2a. Standard deviation of the absolute difference between GEBV at iteration i and final iteration for approach 1. Points above 0.1 are not shown.

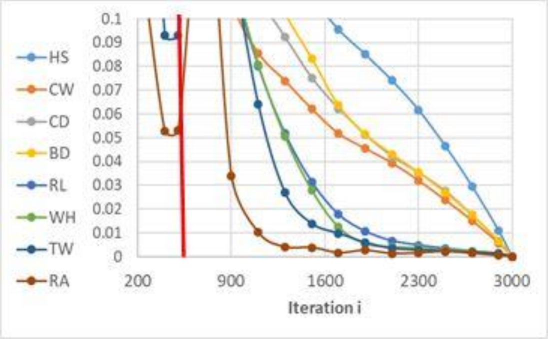

Figure 2b. Standard deviation of the absolute difference between GEBV at iteration i and final iteration for approach 2. Points above 0.1 are not shown. The red vertical line indicates the 500th iteration when step 1 was stopped and step 2 started in approach 2

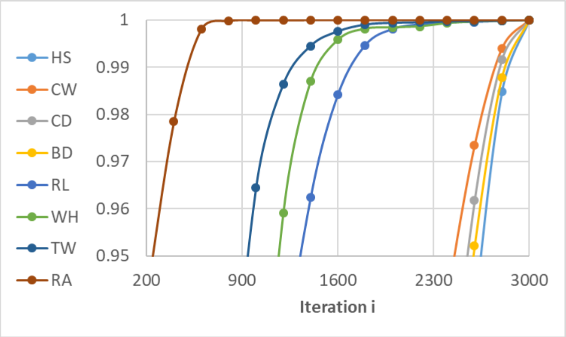

Figure 3a. Correlation between GEBV at iteration i and final iteration for approach 1. Points below 0.95 are not shown.

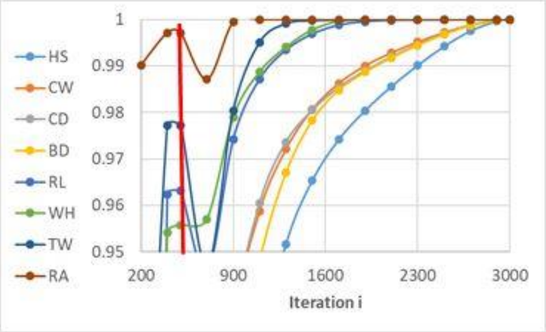

Figure 3b. Correlation between GEBV at iteration i and final iteration for approach 2. Points below 0.95 are not shown. The red vertical line indicates the 500th iteration when step 1 was stopped and step 2 started in approach 2.

In the other scenarios (missing performance for older animals and no missing performances) the difference of speed of convergence between the approaches is not so visible.

The structure and the proportion of missing records have a negative impact on convergence and is more acute for multi-trait SS models than for Combined SS models. The convergence could have been impacted also by the different values of correlations between traits and the different selection intensity applied. The traits with large proportion of missing records (HS,CW,CD,BD) converged at a slow pace with standard Multi-trait SS evaluation (Figure 1a, 2a, 3a) than with the combined SS evaluation (Figure 1b, 2b, 3b).

Moreover, in the HSSGBLUP execution, it's possible to have the correlation between GEBV and (GEBV-200) iterations. For approach 2, the last correlation showed is between iteration 2200 and 2400 (i.e. 2700-2900 for the 2 steps). For the approach 1, it's between 2800 and 3000 iterations.

The last correlations show that, in case 3-approach 1, some traits don't converge yet at the end of the iterations (correlations are going from 0.984823 to 0.999997) while the correlations are better in the approach 2 (correlations are going from 0.999034 to 1).

Conclusions

The combined SS approach converged faster than a standard Multi-trait approach (fewer iterations to reach convergence, faster iterations during the first (univariate) phase). Combined SS can decrease running time by reducing number of iterations, especially when there are missing phenotypes. Nevertheless, it seems that the number of iterations depends on the situation and 3000 iterations were barely sufficient in some cases.

Some differences between traits were found depending on their heritability and genetic correlations as in standard Multi-trait approach.

This example suffers from limitations: the traits considered are described by a rather simple model. Variances were supposed to be homogeneous. Heritabilities of traits and correlations between traits were high. However, this preliminary work suggests that single step evaluations considering groups of traits described by different models are possible and that implementing a Single-Step evaluation in two steps increases modelling flexibility.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93, 743–752. https://doi.org/10.3168/jds.2009-2730

Ducrocq, V., Boichard, D., Barbat, A., Larroque, H., 2001. Implementation of an approximate multitrait BLUP evaluation to combine production traits and functional traits into a total merit index.

Govignon-Gion, A., Dassonneville, R., Baloche, G., Ducrocq, V., 2012. Genetic Evaluation of Mastitis in Dairy Cattle in France.

Govignon-Gion, A., Dassonneville, R., Baloche, G., Ducrocq, V., 2016. Multiple trait genetic evaluation of clinical mastitis in three dairy cattle breeds. Animal 10, 558–565. https://doi.org/10.1017/S1751731115002529

Legarra, A., Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92, 4656–4663. https://doi.org/10.3168/jds.2009-2061

Legarra, A., Ducrocq, V., 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. J. Dairy Sci. 95, 4629–4645. https://doi.org/10.3168/jds.2011-4982

Mäntysaari, E.A., Koivula, M., Strandén, I., 2020. Symposium review: Single-step genomic evaluations in dairy cattle. J. Dairy Sci. 103, 5314–5326. https://doi.org/10.3168/jds.2019-17754

Misztal, I., Legarra, A., Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92, 4648–4655. https://doi.org/10.3168/jds.2009-2064

Tribout, T., Ducrocq, V., Boichard, D., 2020. HSSGBLUP: a Single-Step SNP BLUP genomic evaluation software adapted to large livestock populations.