Bias due to Genomic Selection

Clotilde Patry[1,2] and Vincent Ducrocq[1]
[1]*UR1313 Génétique Animale et Biologie Intégrative, INRA, 78352 Jouy en Josas, France*
[2]*Union Nationale des Coopératives d'Elevage et d'Insémination Animale (UNCEIA)*
149, Rue de Bercy, 75595 Paris Cédex 12
email: clotilde.patry@jouy.inra.fr

Abstract

The inclusion of a pre-selection step based on genomic selection in breeding schemes invalidates some of the assumptions necessary to get optimal BLUP properties in (inter)national genetic evaluations. In particular, the expected value of the mendelian sampling term is no longer 0 and the reduction of mendelian sampling variance is not properly accounted for. An approach is proposed to assess how large the resulting bias might be in a real life situation. This approach "uses" a fraction of the daughters of the last cohort of proven bulls to rank animals with a first estimate of their mendelian sampling modifying their parent average, hence mimicking what genomic selection does. The best candidates are then "a posteriori" selected and the other candidates and their daughters are discarded. Based on their remaining daughters, a progeny test evaluation is computed and compared to the one obtained without this pre-selection. The bias is the average difference between the two. A potential direction for investigation to correct the bias in classical genetic evaluations is proposed. It requires extra information on how pre-selection is actually performed

Introduction

National and international genetic evaluations on field data are traditionally based on Best Linear Unbiased Prediction (BLUP), which under certain conditions have optimal properties allowing a maximum efficiency of selection. For a long time, these assumptions have been roughly fulfilled but this is no longer the case with the introduction of an extra selection stage of genomic selection. This may lead to biased national and international evaluations and to less accurate rankings of bulls and cows (e.g., Banos *et al.*, 2007; van der Beek, 2007). Even the objective assessment of genomic selection efficiency through classical progeny test will be made difficult, as well as the optimal re-estimation of prediction equations. It is important to keep in mind why classical genetic evaluations may become biased. It is also essential to know whether this bias is large. If it is not, current models and practices may be kept unchanged, at least for some time. In the opposite case, statistical models have to be adapted to guarantee unbiased evaluations.

1. Why are EBV becoming biased when a genomic selection step is added?

For illustration, consider a single trait animal model. In matrix notation:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad [1]$$

where **y** is a vector of observations, **b** is the vector of fixed effects, **a** is the vector of additive genetic effects, and **e** is the vector of random residual effects. **X** and **Z** are incidence matrices. Assume that **a** is normally distributed with expected value **0** and variance $\mathbf{A}\sigma_a^2$.

The infinitesimal genetic model relates the additive genetic effect of any animal i to the ones of its parents s and d:

$$a_i = \frac{1}{2}a_s + \frac{1}{2}a_d + \varphi_i \qquad [2]$$

Here φ_i represents the random deviation caused by mendelian sampling. Kennedy *et al.* (1988) reparam-eterised the vector **a** as $\mathbf{a} = \mathbf{T\Psi}$ where $\mathbf{\Psi}$ is a vector of mendelian sampling terms for all animals with known parents and additive genetic effects in the base population for those with unknown parents. If parents are ordered before progeny, the matrix $\mathbf{T}=(\mathbf{I}-\mathbf{P})^{-1}$ is a lower triangular matrix. **P** relates progeny to parents: the row corresponding to animal i includes at most two non zero elements equal to 0.5 in the columns corresponding to the parents of i. The diagonal elements of **T** are 1 and the element (i, k) of **T** is the expected fraction of genes transmitted from ancestor k to animal i.

With these assumptions and notations, it follows that:

$$E[\mathbf{a}]=\mathbf{T}\,E[\boldsymbol{\Psi}] \qquad [3]$$

and: $\mathbf{A}^{-1}=\left(\mathrm{Var}[\mathbf{a}]\right)^{-1}=\mathbf{T}^{-\mathrm{T}}\mathrm{Var}[\boldsymbol{\Psi}]^{-1}\,\mathbf{T}^{-1}$ [4]

Note that $\mathbf{T^{-1}}=(\mathbf{I}-\mathbf{P})$, i.e. $\mathbf{T^{-1}}$ is also a lower triangular matrix with ones on the diagonal and for each row i, at most two non zero elements equal to -0.5 in the columns corresponding to the parents of i.

Under the following extra assumptions:
1) The animals in the base population are unselected, unrelated and non inbred,
2) $\mathbf{A^{-1}}$ is correctly computed, in particular inbreeding is accounted for and the complete pedigree is included (i.e., **T** is complete),
3) All the data on which selection decisions were based is included in the analysis,

Kennedy *et al.* (1988) showed that Henderson's Mixed Model Equations (MME) lead to BLUP estimated breeding values (EBV) where the effects on the expected value and the variance of **a** due to selection, drift, non random matings and inbreeding are properly accounted for, via the (inverse of the) relationship matrix **A**. In particular, under these conditions, we have

$$E[\boldsymbol{\Psi}]=\mathbf{0} \Rightarrow E[\mathbf{a}]=\mathbf{0} \qquad [5]$$

And $$\mathrm{Var}(\varphi_i)=\left(\frac{1}{2}-\frac{F_s+F_d}{4}\right)\sigma_a^2 \qquad [6]$$

where F_s and F_d are the inbreeding coefficients of the sire and the dam of i.

An example where assumption 1 above is not fulfilled is when unknown parents come from different populations with different expected values for the additive genetic effects, then $E[\boldsymbol{\Psi}_{base}]\neq\mathbf{0}$ and EBV are biased if this fact is ignored and a modification of model [1] is required (Quaas, 1988).

Consider now the classical national genetic evaluations in the case when young bulls and cows are pre-selected based on their genomic evaluation. Then, assumptions 2 and 3 above are clearly violated: the (genomic) data at the origin of the selection decision is not included in the MME and the relationship matrix is incorrect: the reduction of mendelian sampling variance is not accounted for. As a result, the expected value and the variance of the mendelian sampling term for the pre-selected animals *sel* are no longer correct:

$$E[\boldsymbol{\varphi}]_{sel}\neq\mathbf{0} \text{ and } \mathrm{Var}[\boldsymbol{\varphi}_{sel}]\neq\left(\frac{1}{2}-\frac{F_s+F_d}{4}\right)\sigma_a^2 \qquad [7]$$

This violation impacts estimates of fixed effects (especially contemporary group effects) and the EBV of animals (pre-selected animals and their relatives, contemporaries,..). The more efficient the pre-selection step is, the further away the evaluation is from the usual assumptions.

2. A simulation to assess the importance of the bias based on real data

2.1 Principle

We propose a simple approach based on current real data to quantify the magnitude of the bias in national evaluations when bulls are pre-selected on the basis of their genomic evaluation. For this purpose, we need to compute the contrast δ between the average $\overline{\mathrm{EBV}}$(gs) of any cohort of animals under a scheme where pre-selection takes place, and the average $\overline{\mathrm{EBV}}$(pt) of the same cohort under a traditional progeny testing scheme. The two schemes differ through the fact that culled (or "non pre-selected") bulls under the genomic selection scheme no longer have progeny, in contrast with their situation under the progeny test scheme.

2.2 Implementation steps

The genomic pre-selection step has to be simulated. For this purpose, we will use part of the existing data and mimic an "a posteriori selection". In the real data set, progeny tested bulls without second crop daughters will be considered as the set of selection candidates. With a 2008 data set, these are bulls born in 2001, 2002 or 2003 and having between 80 and 150 daughters.

Step1: pre-selection step. BLUP evaluations are first run keeping just a random sample of N daughters per bull in the selection candidates set.

The performances of their remaining daughters are deleted but records of all other animals are kept. N is a key parameter here and is chosen so that it leads to the same extra reliability of EBV as the one permitted through genomic selection (e.g., N= 10 to 20). Assume one observation by recorded cow so animal model [1] can be applied.

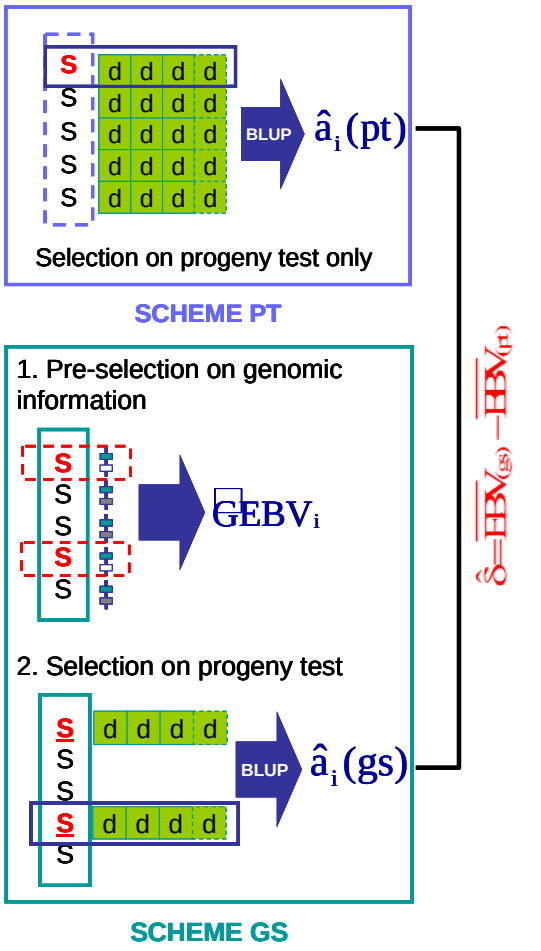

Figure 1. Estimation of the bias due to pre-selection contrasting a regular progeny testing scheme (PT) with a simulated genomic scheme (GS) where pre-selection is based on EBV.

From this classical genetic evaluation, we get an EBV $\hat{a}_i$ for each candidate i as well as an estimate of its mendelian sampling based $\hat{\varphi}_i=\hat{a}_i-\frac{1}{2}\hat{a}_s-\frac{1}{2}\hat{a}_d$, the latter with a reliability comparable to the one obtained via genomic evaluation. Various genomic selection implementations can be envisioned by selecting candidate bulls based on either $\hat{a}_i$ or $\hat{\varphi}_i$ (e.g., for within family selection). The x% top bulls according to the chosen criterion represent the "pre-selected bulls", the others are assumed to be "culled" and their daughters non existent.

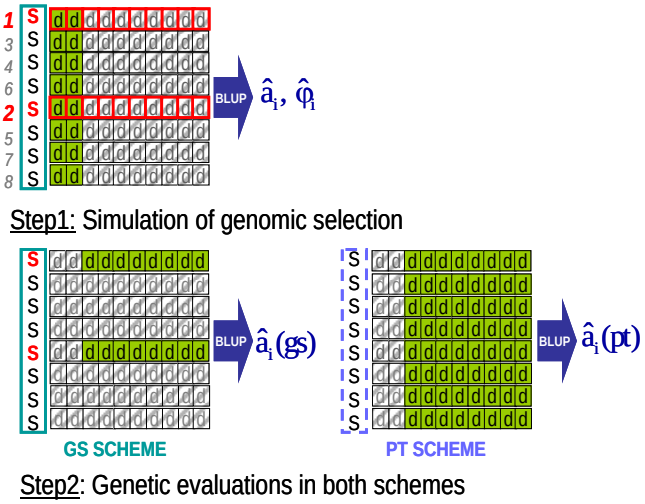

Figure 2. Mimicking pre-selection with a two step approach.

Step 2: a progeny test data set (pt) is created from the initial one by deleting the records from the N daughters chosen in step 1 for each candidate bull. This is to ensure that the result of the new progeny-test is not influenced by the pre-selection step, just as in real genomic selection. From the animal model MME based on this data set, we get $\overline{EBV}(pt)$ for any cohort of interest (candidates, their progeny, their progeny's contemporaries, etc.)

In parallel, daughters of the "culled" bulls are also deleted from the initial data set, so only daughters from pre-selected bulls remain in the data set. Again, a classical genetic evaluation is performed, leading to an estimate of $\overline{EBV}(gs)$ for exactly the same cohorts as before. Consequently, the estimate of the bias due to pre-selection is $\hat{\delta}=\overline{EBV}(gs)-\overline{EBV}(pt)$

2.3 Numerical application

Data. This approach was applied to milk production in the French Holstein breed. In step 1, more than 13 million animals were included in the pedigree file as well as 9,184,856 records. To study the importance of the bias for different types of animals, 8 distinct cohorts of animals were defined based on their sex, role (candidates, progeny test daughters, contemporaries) and year of birth.

To simplify the computations, step 1 was implemented using "pre-adjusted records" as in the French total Merit Index computations (Ducrocq *et al.,* 2001). Pre-adjusted records (**y***) are records averaged over all lactations after correction for all fixed effects (herd-year-season, calving month, calving age and length of dry period defined within year and region) and permanent environment. Each pre-adjusted record has an attached weight, measuring the amount of information included (e.g., several lactations of different lengths). In this simple case, model [1] includes only one fixed effect (year of birth effect) and the residual variance of each record is inversely proportional to the associated weight.

$$\mathbf{y^* = \mu + Z\,a + e} \qquad [8]$$

Step 2 is now being implemented. Unfortunately, no results are available yet. Note that here, pre-adjusted records can no longer be used (we learnt it the hard way!) because then any individual bias in EBV cannot be counterbalanced by an opposite bias in, for example, the herd-year-season effect. In other words, we must go back to the original data or more simply, to consider a simple repeatability model with a herd-year effect based on lactation records corrected for all fixed effects except herd-year-season.

$\hat{\delta}$ as well as correlations between EBV_{gs} and EBV_{pt} will be computed.

3. A potential strategy to reduce/correct the EBV bias

What follows is no more than a potential direction of research that may be envisioned to correct the EBV bias due to pre-selection on genomic evaluation. Because the nature of genomic information is radically different from classical performances, it is difficult to include it directly in the MME in such a way that selection is accounted for.

3.1. Setup: We start again from the reparameter-isation $\mathbf{a}=\mathbf{T\Psi}$, considering that $E[\boldsymbol{\varphi}]_{sel} \neq \mathbf{0}$ in a way similar to the unknown parents groups situation of Quaas (1988).

Let **Q** be a matrix connecting each animal to a group of animals of a same sex going through the same pre-selection step at the same time, based on their genomic evaluation. The definition of these groups may have to be done at a very refined level. We will assume that animals i within each group j have the same expected value of the mendelian sampling term: $E[\varphi_i]=\Delta_j$ and that $Var[\varphi_i]=\gamma_j$ (considered as known for the moment). The i^{th} row of **Q** is 0 everywhere, except for a 1 in column j. For animals without genomic data, the corresponding row of **Q** is zero and the variance of the mendelian sampling term is as usual. Then we can write:

$$E[\boldsymbol{\Psi}]=\mathbf{Q\Delta} \qquad [9]$$

where $\boldsymbol{\Delta}$ is the vector of nonzero expectations of mendelian sampling terms, i.e., a vector of group biases, considered as fixed effects.

Let $\boldsymbol{\Gamma}=\boldsymbol{\Psi}-\mathbf{Q\Delta}$ such that $E[\boldsymbol{\Gamma}]=\mathbf{0}$ and define:

$$\mathbf{a}=\mathbf{T\Psi}=\mathbf{T\Gamma}+\mathbf{TQ\Delta}=\mathbf{a^*}+\mathbf{TQ\Delta} \qquad [10]$$

$var(\mathbf{a^*})= Var\,(\mathbf{a^*})=\mathbf{A^*}=\mathbf{T}Var[\boldsymbol{\Psi}]\,\mathbf{T'}$ with expression [6] replaced by the appropriate γ_j for pre-selected animals.

Consider the following model equivalent to [1]:

$$\mathbf{y= Xb + ZTQ\Delta + Za^* + e} \qquad [11]$$

Note that **T** in $\mathbf{TQ\Delta}$ ensures that the pre-selection bias for an animal is appropriately passed over all his progeny, generation after generation.

Henderson's Mixed Model Equations for this model are (with $\alpha=\sigma_e^2/\sigma_a^2$):

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'ZTQ} & \mathbf{X'Z} \\ \mathbf{Q'T'Z'X} & \mathbf{Q'T'Z'ZTQ} & \mathbf{Q'T'Z'} \\ \mathbf{Z'X} & \mathbf{Z'ZTQ} & \mathbf{Z'Z}+\alpha\mathbf{A^{*-1}} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \boldsymbol{\Delta} \\ \mathbf{a^*} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Q'T'Z'y} \\ \mathbf{Z'y} \end{bmatrix} \qquad [12]$$

This system can be simplified using a transformation proposed by Quaas (1988) where $\mathbf{I}=\mathbf{S}^{-1}\mathbf{S}$ is inserted between the coefficient matrix and the vector of unknowns and both sides are pre-multiplied by $\mathbf{S}^{-T}$ with:

$$\mathbf{S}=\begin{bmatrix}\mathbf{I} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{TQ} & \mathbf{I}\end{bmatrix}=\begin{bmatrix}\mathbf{I} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I} & \mathbf{0}\\ \mathbf{0} & -\mathbf{TQ} & \mathbf{I}\end{bmatrix}^{-1} \quad [13]$$

Then, system [12] simplifies to:

$$\begin{bmatrix}\mathbf{X'X} & \mathbf{0} & \mathbf{X'Z}\\ \mathbf{0} & \alpha\mathbf{Q'T'A^{*-1}TQ} & -\alpha\mathbf{Q'T'A^{*-1}}\\ \mathbf{Z'X} & -\alpha\mathbf{A^{*-1}TQ} & \mathbf{Z'Z}+\alpha\mathbf{A^{*-1}}\end{bmatrix}\begin{bmatrix}\mathbf{b}\\ \boldsymbol{\Delta}\\ \mathbf{a}\end{bmatrix}=\begin{bmatrix}\mathbf{X'y}\\ \mathbf{0}\\ \mathbf{Z'y}\end{bmatrix} \quad [14]$$

This form has three main advantages: it is sparser and easier to build than [13], the lower block $\begin{bmatrix}\mathbf{Q'T'A^{*-1}TQ} & -\mathbf{Q'T'A^{*-1}}\\ -\mathbf{A^{*-1}TQ} & \mathbf{A^{*-1}}\end{bmatrix}$ can be constructed using an extension of Henderson's rules for $\mathbf{A}^{-1}$ and **a** (and not **a***) is obtained directly.

3.2 Bias correction: based on this formulation, there are several potential alternatives that can be studied, depending on the available information:

Alternative 1: All direct genomic EBV (DGBV) are available for all selection candidates (not only for the animals passing the pre-selection step). This is likely to be the case for most national genetic evaluations, but only for domestic bulls. Then it is relatively easy to compute a within group selection differential $\hat{\Delta}_j$ based on actual DGBV. This selection differential can also be used to approximate selection intensity and consequently, γ_j the variance of the mendelian sampling term after selection. Then system [14] simplifies to:

$$\begin{bmatrix}\mathbf{X'X} & \mathbf{X'Z}\\ \mathbf{Z'X} & \mathbf{Z'Z}+\alpha\mathbf{A^{*-1}}\end{bmatrix}\begin{bmatrix}\mathbf{b}\\ \mathbf{a}\end{bmatrix}=\begin{bmatrix}\mathbf{X'y}\\ \mathbf{Z'y}+\alpha\mathbf{A^{*-1}TQ}\hat{\boldsymbol{\Delta}}\end{bmatrix} \quad [15]$$

Correction is limited to a different computation of $\mathbf{A}^{-1}$ and a correction term added to the right hand side.

Alternative 2: GEBV are not provided but all the elements needed to derive (approximate) selection intensities are available. Then, with some extra assumptions, for example on the correlation between true and estimated mendelian sampling term, it is possible to compute at least approximately theoretical $\hat{\Delta}_j$ and γ_j and apply [15].

Alternative 3: only **Q** is known and only for pre-selected animals. Then either a selection intensity has to be assumed, in particular to derive the effect of pre-selection on the variance of the mendelian sampling term, or this effect is ignored. In this case, $\hat{\Delta}_j$ could be estimated at the same time of **b** and **a**, using [14].

There are certainly other approaches applicable. Note that the three proposed here require at least the knowledge of **Q**, i.e., of what animals passed a pre-selection step together.

Perspectives

As indicated above, the main source of bias in genetic evaluations is likely to be initially the comparison within contemporary group (e.g., herd-year-season combination) between daughters of genomically pre-selected bulls and daughters of regular progeny-tested bulls. The fact that the relative superiority of the former due to sires with positive mendelian sampling terms is not accounted for will lead to over-estimated contemporary group effects, which then will influence the EBV of all animals in the group as well as their parents. If the proportion of daughters of pre-selected bulls is small - as it may well be the case before the adoption of genomically pre-selected young bulls is large - then the contemporary group effect will not be greatly modified and reduced biased may be expected. This may leave us one or two more years before a correction of our methods and software becomes critical…

References

Banos, G. *et al.* 2007. Interbull Scientific Advisory Committee annual report 2007, Dublin, Ireland. *www-interbull.slu.se*

Ducrocq, V. *et al.* 2001. Impletentation of an approximate multitrait BLUP to combine production traits and functional traits into a total merit index. *52th annual meting of EAAP*. Budapest, Hungary.

Kennedy, B.W. *et al.* 1998. Genetic properties of animal models. *J. Dairy Sci. 71 (suppl 2)*, 17-26.

Quaas, R.L. 1988. Additive genetic model with groups and relationships. *J. Dairy Sci. 71*, 1338-1345.

Van der Beek, S. 2007. Effect of genomic selection on national and international genetic evaluations. *Interbull Bulletin 37*, 115-118.