A method for partitioning trends in genetic mean and variance

Oliveira, T.P.[1], Obšteter, J.[2], Pocrnic, I.[1], Gorjanc, G.[1]

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, UK; [2] Agricultural Institute of Slovenia, Slovenia

E-mail for correspondence: thiago.oliveira@ed.ac.uk

Abstract: Quantifying sources of genetic change is essential for identifying key breeding actions and optimising breeding programmes. However, the observed genetic change is a sum of contributions from different groups of individuals (often referred to as selection pathways), which are difficult to disentangle and quantify due to the complexity of breeding programmes. Here we extended a simple method to analyse the contributions of groups to the genetic variance. Our approach showed the importance of analysing the partition of the genetic mean and variance rather than just the genetic mean and demonstrated that the contributions are not necessarily independent.

Keywords: Partitioning method; Genetics; Mixed-Effects Models.

1 Background

We aim to genetically improve populations in animal breeding by selecting the best individuals as the next generation's parents. Ideally, we would select the parents based on their true genetic/breeding value, but we can never know that values. Alternatively, we can select parents based on i) phenotypic value, which is the expressed trait and has a medium/low accuracy; ii) estimated breeding value, which may have high accuracy since it considers the phenotypic values of the individuals and all its relatives. Thus, an important step is to understand where genetic progress comes from and which group of animals creates the most genetic gain.

Let $\boldsymbol{a}$ be a vector of breeding values sampled from a normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{A}\sigma_a^2$. We can write $\boldsymbol{a}$ as a linear combination of the individual's ancestors breeding values and individual's deviation from ancestors $\boldsymbol{a} = \boldsymbol{T}\boldsymbol{w}$. We can define $\boldsymbol{T}$ as a triangular matrix of expected

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

gene flow between ancestors and individuals, and $\boldsymbol{w} \sim N\left(\mathbf{0}, \boldsymbol{W}\sigma_a^2\right)$ as the Mendelian sampling terms representing deviations, with $\boldsymbol{W}$ being a diagonal matrix of variance coefficients and σ_a^2 the base population genetic (additive) variance. Assuming a factor with p groups and for any set $\sum_{j=1}^{p} \boldsymbol{P}_j = \boldsymbol{I}$, García-Cortés et al. [2008] partitioned the genetic mean into contributions of each level by defining $\boldsymbol{T}_j = \boldsymbol{T}\boldsymbol{P}_j$, and further partitioned the contribution of each group to breeding values *a priori* using the equality $\boldsymbol{a} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p)\, \boldsymbol{w} = \boldsymbol{a}_1 + \boldsymbol{a}_2 + \ldots + \boldsymbol{a}_p$. García-Cortés et al. (2008) further showed that these partitions can be estimated from data collected in breeding programmes (*posteriori*) by first estimating the breeding values $\widehat{\boldsymbol{a}} = E\left(\boldsymbol{a}|\boldsymbol{y}\right)$ from phenotype data ($\boldsymbol{y}$). Since $\boldsymbol{a}$ is a function of variance components in the mixed-effects model, we can estimate $\boldsymbol{a}$ and $\boldsymbol{w}$ by replacing their REML estimates $\widehat{\boldsymbol{a}} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p)\, \widehat{\boldsymbol{w}} = \widehat{\boldsymbol{a}}_1 + \widehat{\boldsymbol{a}}_2 + \ldots + \widehat{\boldsymbol{a}}_p$. By summarising these partitions, they quantified the contribution of each group (i.e. males vs. females, countries, AI centres) to the time-trend in genetic mean.

2 Methods

2.1 Partitioning of genetic trends

Here we extend the partitioning method to analyse the contribution of groups to genetic variance. Variance of breeding values is, *a priori*, $Var\left(\boldsymbol{T}\boldsymbol{w}\right) = \boldsymbol{T}\boldsymbol{W}\boldsymbol{T}^T\sigma_a^2$. Thus, we can partition genetic variance as

$$\begin{aligned} Var\left(\boldsymbol{a}\right) = Var\left[\left(\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p\right)\boldsymbol{w}\right] &= \sum_{j=1}^{p} \boldsymbol{T}_j \boldsymbol{W} \boldsymbol{T}_j^T \sigma_a^2 + 2\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} \boldsymbol{T}_j \boldsymbol{W} \boldsymbol{T}_{j'}^T \sigma_a^2 \\ &= \sum_{j=1}^{p} \sigma_{a_j}^2 + \sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} \sigma_{a_j, a_{j'}} \qquad (1) \end{aligned}$$

While this "theoretical"partitioning involves matrix products, we can also summarise partitions $\boldsymbol{a}_1 + \boldsymbol{a}_2 + \ldots + \boldsymbol{a}_p$ (calculated via $\boldsymbol{T}^{-1}$) by calculating variance of each group level contribution $f\left(\boldsymbol{a}\right) = Var\left(\boldsymbol{a}_j\right)$ and covariance of each pair of group level contributions $f\left(\boldsymbol{a}_j, \boldsymbol{a}_{j'}\right) = Cov\left(\boldsymbol{a}_j, \boldsymbol{a}_{j'}\right)$. The partitions can be summarised in many ways to quantify the contribution of different groups to change in genetic variance over time. The partitioning method can be then only applied for *a priori* or true breeding values. Although the methodology has been developed to *a priori* or true breeding values, we can use methods from Sorensen et al. [2001] to estimate partitions of genetic variance from data collected in a breeding programme (*posteriori*).

2.2 Statistical model and computational approaches

In the previous subsection, we assumed we knew the true breeding values. Consequently, the same assumption is applied to additive genetic mean and variance contributions. However, in reality, we use phenotype, pedigree, and genomic information to predict the breeding values $\hat{\boldsymbol{a}}$ and make inferences. We fitted standard animal model:

$$\begin{gathered} \boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e}, \\ \boldsymbol{a} \sim N\left(\boldsymbol{0}, \sigma_a^2 \boldsymbol{A}\right), \text{ and } \boldsymbol{e} \sim N\left(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}\right) \end{gathered} \tag{2}$$

where $\boldsymbol{y}$ is a vector of observed phenotypes, $\boldsymbol{b}$ is a vector of fixed effects with design matrix $\mathbf{X}$, $\mathbf{a}$ is a vector of random animal effects with design matrix $\mathbf{Z}$, and $\mathbf{e}$ is a vector of random residuals. It is assumed that $\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{A}\sigma_a^2\right)$ and $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$, where $\mathbf{A}$ is the pedigree-based numerator relationship matrix, while σ_a^2 and σ_e^2 are respectively known additive and residual variances.

The directed acyclic graph (DAG) representation of the model (2) considering only intercept as the fixed effect is illustrated in Figure 1, where pedigree and phenotypic records are displayed in separate plates as a generalization of the case where animals might not have phenotypic records. In addition, the dotted lines indicate a possibly missing parent in the pedigree. In the pedigree plate we have K individuals represented by founders and non-founders, where founders is a `priori` sampled from $a_k|\sigma_a^2 \sim \left(0, \sigma_a^2\right)$. Non-founders individuals given the information of their parents are then represented by $a_k = 1/2\left(a_{f(k)} + a_{m(k)}\right) + w_k$, where $a_{f(k)}$ and $a_{m(k)}$ are parent's breeding value and w_k represents the Mendelian sampling term $\left(w_k|\boldsymbol{W}_{k,k} \sim N\left(\boldsymbol{0}, \sigma_a^2\boldsymbol{W}_{k,k}\right)\right)$.

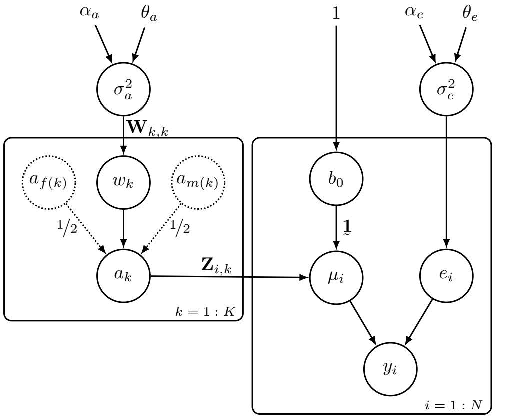

FIGURE 1. Directed acyclic graph of the animal model with nI individuals and nY phenotypic records (y_i) with explicit representation of Mendelian sampling terms (w_k) and error term (e_i), where σ_a^2 is the additive genetic variance, $a_{f(k)}$ and $a_{m(k)}$ are parent's breeding value, $\mathbf{1}$ represents a vector of ones, μ_i the linear predictor, and σ_e^2 the variance of the error term

In this sense, matrix $\boldsymbol{A}$ can be decomposed as $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{W}\boldsymbol{T}^T$ using LDL decomposition Golub and Van Loan [1996], as described in section 2.1. The diagonal elements of $\boldsymbol{W}$ can be computed according to specific scenarios described by Mrode [2005] as i) $\boldsymbol{W}_{k,k} = \frac{1}{2} - \frac{1}{4}\left(F_{f(k)} + F_{m(k)}\right)$ when both parents are known; ii) $\boldsymbol{W}_{k,k} = \frac{3}{4} - \frac{1}{4}F_{m(k)}$ or $\boldsymbol{W}_{k,k} = \frac{3}{4} - \frac{1}{4}F_{f(k)}$ when one parent are known; and iii) $\boldsymbol{W}_{k,k} = 1$ when both parents are unknown, where $F_{f(k)}$ and $F_{m(k)}$ are the coefficients of inbreeding related to the father and mother identification of the individual k, respectively Kennedy et al. [1988], Falconer and Mackay [1996], Mrode [2005].

Accounting for inbreeding when computing $\boldsymbol{A}^{-1}$ may impact the partitioning results for genetic variance according to the inbreeding level because, for any domain D, we have $Var\left(a_k|\boldsymbol{W}_{k,k}\right) = \int_D a_k^2 \,\Pr\left(a_k\right) da_k - \left[\int_D a_k \,\Pr\left(a_k\right) da_k\right]^2 = \left(1 + F_k\right)\sigma_a^2$, where $\Pr(.)$ represents a probability density function, and F_k is the inbreeding coefficient of the kth individual. Thus, we decided to include two more scenarios i) accounting and ii) not accounting for inbreeding when constructing $\boldsymbol{A}^{-1}$. In the case of ignoring inbreeding the $\boldsymbol{W}_{k,k}$ is equal to $\frac{1}{2}$, $\frac{3}{4}$ and 1, respectively Mrode [2005].

We used the full Bayesian approach by specifying prior distribution for all model parameters, as shown in Figure 1. Thus, $\boldsymbol{b}$, σ_a^2 and σ_e^2 are assumed to have a joint prior density of the form $p\left(\boldsymbol{b},\sigma_a^2,\sigma_e^2\right) = p\left(\boldsymbol{b}\right)p\left(\sigma_a^2\right)p\left(\sigma_e^2\right)$, where $p\left(\tau_a = 1/\sigma_a^2|\alpha_a,\theta_a\right) \propto \tau_a^{\alpha_a-1}\exp\left(-\theta_a\tau_a\right)$, $p\left(\tau_e = 1/\sigma_e^2|\alpha_e,\theta_e\right) \propto \tau_e^{\alpha_e-1}\exp\left(-\theta_e\tau_e\right)$, and $p\left(\boldsymbol{b}\right) \propto 1$, with $\tau_a > 0$, $\alpha_a \geq 0$, $\theta_a \geq 0$, $\tau_e > 0$, $\alpha_e \geq 0$ and $\theta_e \geq 0$. In this case, we are assuming a flat prior distribution for β which is independent of σ_a^2 and σ_e^2. On the other hand, the inverse-gamma(α,θ) is a natural candidate for the prior distributions for variance components, and when α and θ are set to a value such as 0.1^3, it can be considered as vague prior within the conditionally conjugate family $\sigma_a^{-2},\sigma_e^{-2} \sim \text{Gamma}\left(0.1^3, 0.1^3\right)$. The posterior distribution can be obtained by applying the form of Bayes' theorem conditional on the data:

$$\begin{aligned} p\left(\boldsymbol{b},\boldsymbol{a},\sigma_a^2,\sigma_e^2|\boldsymbol{y}\right) \propto & p\left(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{a},\sigma_e^2\right)p\left(\boldsymbol{b}\right)p\left(\boldsymbol{a}|\boldsymbol{A},\sigma_a^2\right)\times \\ & p\left(\sigma_a^2|\alpha_a,\theta_a\right)p\left(\sigma_e^2|\alpha_e,\theta_e\right). \end{aligned}$$

We used Markov Chain Monte Carlo (MCMC) to generate samples from the posterior distribution using Gibbs sampler algorithm Sorensen et al. [2001]. It was considered one chain with 80,000 samples, from which 20,000 iterations are burn-in while the remaining 60,000 were stored using a thinning of length 40. Consequently, 1,500 samples of EBV's are computed observing the posterior distribution $p\left(\boldsymbol{a}|\boldsymbol{A},\sigma_a^2\right)$, which are passed as input for the `AlphaPart` package. We assessed MCMC convergence by looking at trace and autocorrelation function plots. Gibbs sampling was executed by GIBBS1F90 software Misztal et al. [2018].

3 Results and Discussion

The simulated cattle breeding programme illustrated the power of the partitioning method to summarise genetic trends in mean and variance, although some care is needed when using the proposed methodology. By partitioning the genetic mean and variance we showed that in a high accuracy scenario the covariance between females (F) and selected males (M) plays an important role in the contribution to the genetic variance and, consequently, in this case $Var\left(\boldsymbol{a}\right) < Var\left(\boldsymbol{a}|\mathrm{F}\right) + Var\left(\boldsymbol{a}|\mathrm{M}\right)$. In this sense, we demonstrated that the choice of groups is essential and that contributions are not necessarily independent; hence, they should not be analyzed in isolation from each other.
The advantage of combining the MCMC approach with the partition method presented here is related to drawing samples from the posterior distribution $p\left(\boldsymbol{a}|\boldsymbol{A}, \sigma_a^2\right)$ and using them to compute the point estimate partitions for genetic mean and variance and also access their uncertainty. Although the methodology presented here works fine for the extreme example proposed using medium accuracy, we again expect the reproducible inaccuracy showed in Figure 2 can be overcome with an extension of the partition method using genomic models.

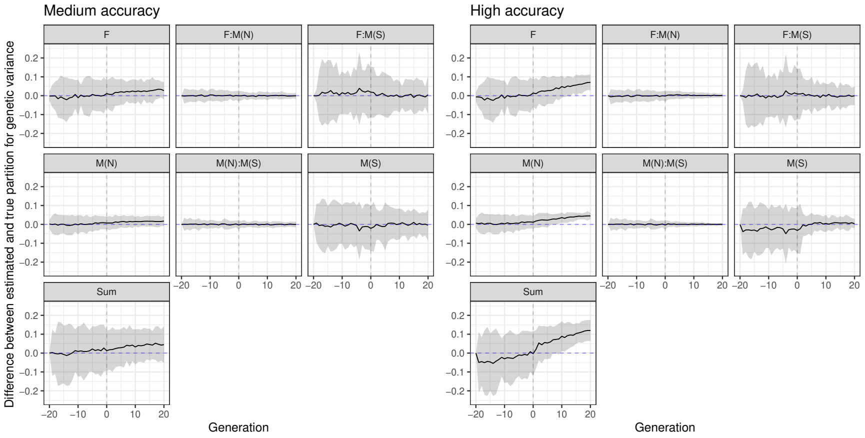

FIGURE 2. Distribution of the difference between true and estimated partitions for the total additive genetic variance (Sum) over generations by gender (male (M) and female (F)) and status (selected males (S) and non-selected males (N)) considering 30 simulations replicate

4 Conclusion

We developed a method for quantifying sources of genetic variance. This is a powerful and valuable method for understanding how different breeding groups interact within a breeding programme, and hence for optimising breeding programmes. By partitioning the genetic variance in a simulated cattle breeding programme, we showed that the covariance between paths can make a substantial contributions to the genetic variance. Hence, to comprehend and manage the genetic variance in a breeding programme, we should not consider the contribution of different groups in isolation but should perform a holistic analysis and partition of the observed genetic variance instead.

Acknowledgments: The authors acknowledge support from the BBSRC, European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801215, Roslin Institute Strategic Programme (ISP) grants, and Core financing of the Slovenian Research Agency (grant P4-0133).

5 References

D. S. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics*. Longman, Essex, 4 edition, 1996.

L.A. García-Cortés, J.C. Martínez-Ávila, and M.A. Toro. Partition of the genetic trend to validate multiple selection decisions. *Animal*, 2(6):821–824, 2008. ISSN 17517311. doi: 10.1017/S175173110800205X. URL https://linkinghub.elsevier.com/retrieve/pii/S175173110800205X.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, 1996.

B W Kennedy, L R Schaeffer, and D A Sorensen. Genetic Properties of Animal Models. *Journal of Dairy Science*, 71:17–26, 1988. ISSN 0022-0302. doi: 10.1016/S0022-0302(88)79975-0. Publisher: Elsevier.

I. Misztal, S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. Manual for BLUPF90 family programs, 2018. URL http://nce.ads.uga.edu/wiki/doku.php?id=documentation.

R. A. Mrode. *Linear models for the prediction of animal breeding values*. CAB International, Wallingford, 2 edition, 2005.

D. Sorensen, R. Fernando, and D. Gianola. Inferring the trajectory of genetic variance in the course of artificial selection. *Genetical Research*, 77(1):83–94, February 2001. ISSN 0016-6723, 1469-5073. doi: 10.1017/S0016672300004845. URL https://www.cambridge.org/core/product/identifier/S0016672300004845/type/journal_article.