



Effect of fitting a genotypic mean on bias and accuracy of single-step genomic prediction

Tesfaye K. Belay¹, S.Eikje², A.Gjuvsland² and T.Meuwissen¹

¹Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Norway

²GENO Breeding and A.I Association, Norway

The 70th EAAP Annual Meeting, 29th August 2019, Ghent, Belgium



Introduction

- **SS-GBLUP model** – combines all data from genotyped and ungenotyped animals.
- However, SS-GBLUP model requires the **G** and **A** matrices to refer to the **same base population**
- To handle this problem, it needs to
 - determine which **allele frequencies** to be used in the **G** matrix and
 - to adjust this matrix to the **A** matrix.



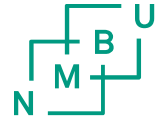
Introduction

- In theory, **allele frequencies in base animals** should be used.
 - But such frequencies rarely available in practice
- Several studies have discussed this problem and proposed solutions:
 - SS-GBLUP model (Vitezica et al. 2011, Christensen et al. 2012; Legarra et al., 2015)
 - SS-SNPBLUP model (Fernando *et al.*, 2014, 2016; Hsu et al., 2017).
- Fernando *et al.* (2014) proposed to fit an additional **fixed covariate (J)** that estimates the **intercept of the regression (μ_g)** on the genotypes.



Introduction

- Estimating this intercept **implicitly estimates the frequency** by which the marker genotypes should be centered.
- This frequency is thus **estimated from the data** by estimating the intercept.
- The J-covariate has **not been tested on empirical data in the SS-GBLUP model.**



Aims of this study

- to evaluate effect of fitting the J-covariate on bias and accuracy of SS-GBLUP evaluations of Norwegian Red cattle.

- to evaluate different ways of combining J and genetic group (\mathbf{Q}) effects and investigate biases and accuracies of the resulting breeding value estimates.

Materials and Methods

Theory

- Theoretical background for deriving the \mathbf{J} covariate is described by Fernando et al (2014) and Hsu et al (2017).

- Let \mathbf{M}_2 is matrix of genotypes for genotyped individuals

- $\widehat{\mathbf{M}}_1$ is matrix of imputed genotypes for ungenotyped individuals

$$\triangleright \text{i.e., } \widehat{\mathbf{M}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2.$$

- Model for genotypic values (\mathbf{g}) are given as (Hsu *et al.*, 2017)

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{J}\mu_g + \mathbf{M}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

- Where $\mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix}$; $\mathbf{M} = \begin{bmatrix} \widehat{\mathbf{M}}_1 \\ \mathbf{M}_2 \end{bmatrix}$; $\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 \\ \mathbf{J}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{1} \\ \mathbf{1} \end{bmatrix}$;

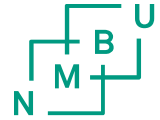


Theory...

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{J}\mu_g + \mathbf{M}\alpha + \epsilon$$

- α is a vector of marker genotype effects;
- ϵ is a vector of imputation residuals (for genotyped animals: $\epsilon_i = 0$);
- μ_g is the **intercept of the regression** of the marker genotypes
 - i.e. it is the genotypic value of an hypothetical animal i with genotypes at all markers, M_i , equal to the **mean genotype** ($E(M_i)$)

Theory...



Some special cases may explain effect of the J-covariate:

- 1) If **all animals are genotyped**, fitting $J\mu_g$ is like fitting an overall mean – thus, μ_g is confounded with the overall mean and redundant
- 2) If the genotyped animals are **unrelated** to the ungenotyped animals,

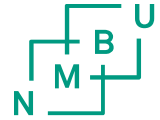
$$A_{12} = \mathbf{0} \text{ and } \mathbf{J} = \begin{bmatrix} \mathbf{J}_1 \\ \mathbf{J}_2 \end{bmatrix} = \begin{bmatrix} A_{12}A_{22}^{-1}\mathbf{1} \\ \mathbf{1} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}.$$

- Here, ϵ models the full genetic value of the ungenotyped animals using \mathbf{A} .
- $J\mu_g$ fits the **genetic difference** between the ungenotyped and genotyped
- $J\mu_g$ can account for **a difference in genetic base** between \mathbf{A} and \mathbf{G} matrices
 - hence, it is also relevant for SS-GBLUP models.



Theory...

- 3) When $A_{12} \neq \mathbf{0}$, the ungenotyped animals are modelled by
 - a combination of **marker effects** (part predicted from genotyped animals), and
 - a **pedigree-based animal effect**, ϵ .
- Here, J-covariate accounts for the fraction that can be explained by the markers using A, which is $A_{12}A_{22}^{-1}\mathbf{1}$.
- Showing that fitting of the J covariate is also relevant to the SSGBLUP
 - b/c it **corrects for differences in genetic level** due to base population differences between the G and A matrices.



Phenotypic and genotypic data

- All data (pedigree, phenotype and genotype) provided by GENO SA
- **Phenotype**: 1st lactation kg milk from 3,390,184 Norwegian Red cows
- **A pedigree** containing 4,624,098 animals.
- **Genotype data**:
 - 30,729 animals genotyped (10,989 animals had phenotype).
 - 30,300 SNP markers on 29 autosomes.

Models

Model	M.name	Description	GEBV
$y = Xb + Wh + Za + e$	SSGBLUP_N	J & Q not fitted	\hat{a}
$y = Xb + Wh + Za + ZJ\mu_g + e$	SSGBLUP_J	J-covariate fitted	$\hat{a} + J\hat{\mu}_g$
$y = Xb + Wh + Za + ZQg + e$	SSGBLUP_Q	Q fitted	$\hat{a} + Q\hat{g}$
$y = Xb + Wh + Za + ZJ\mu_g + ZQg + e$	SSGBLUP_JQ	J and Q fitted	$\hat{a} + J\hat{\mu}_g + Q\hat{g}$
$y = Xb + Wh + Za + ZQ^*g + e$	SSGBLUP_Q*	modified Q fitted	$\hat{a} + Q^*\hat{g}$
$y = Xb + Wh + Za + ZQg + ZQ^*g^* + e$	SSGBLUP_QQ*	Q & modified Q fitted	$\hat{a} + Q\hat{g} + Q^*\hat{g}^*$

- Modified Q (Q^*) = $J^{-1} \times Q$
- J, J^{-1} and Q^* computed in [Julia \(v0.64\)](#) environment
- [DMU](#) (Madsen and Jensen, 2013) used for genomic prediction

Evaluations

- Inflation, level bias and accuracy of GEBV evaluated
 - Under three scenarios:
 - phenotypes (P), or
 - genotypes (G), or
 - both phenotypes and genotypes (PG) of 675 young animals masked.
- Corrected phenotype(CP_c) = $\widehat{GEBV}_c + \hat{e}_c$
- Accuracy = $\text{cor}(\widehat{GEBV}_r, CP_c)$
- Inflation = $\text{reg}(\widehat{GEBV}_r, CP_c)$
- Level bias = $\text{mean}(\widehat{GEBV}_r - \widehat{GEBV}_c) / \text{sd}(\widehat{GEBV}_r)$

Results – Inflation

Coefficients for regression of corrected-phenotypes on breeding values

Model	Scenario		
	P-masked	G-masked	PG-masked
SSGBLUP_N	1.0636	1.9620	1.0602
SSGBLUP_J	1.0630	1.9703	1.0815
SSGBLUP_Q	1.0062	1.9419	1.0484
SSGBLUP_JQ	1.0008	1.8773	1.0218
SSGBLUP_Q*	1.0690	1.9523	1.0487
SSGBLUP_QQ*	1.0087	1.9420	1.0468

- Modified Q (Q^*) = $J^{-1}xQ$

Results - Level bias

Level bias i.e., mean difference in breeding values and scaled by SD

Model	Scenario		
	P-masked	G-masked	PG-masked
SSGBLUP_N	-0.023	-0.155	-0.245
SSGBLUP_J	-0.023	-0.131	-0.217
SSGBLUP_Q	0.036	-0.146	-0.184
SSGBLUP_JQ	0.037	-0.083	-0.094
SSGBLUP_Q*	-0.024	-0.154	-0.242
SSGBLUP_QQ*	0.035	-0.141	-0.183

- Modified Q (Q^*)= $J^{-1} \times Q$

Results - Accuracy

Correlation between corrected-phenotypes and breeding values

Model	Scenario		
	P-masked	G-masked	PG-masked
SSGBLUP_N	0.443	0.729	0.346
SSGBLUP_J	0.443	0.732	0.352
SSGBLUP_Q	0.441	0.725	0.346
SSGBLUP_JQ	0.440	0.720	0.351
SSGBLUP_Q*	0.445	0.727	0.344
SSGBLUP_QQ*	0.441	0.725	0.345

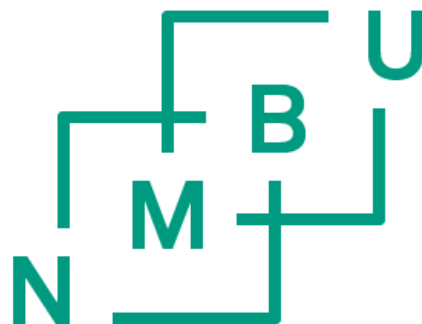
- Modified Q (Q^*)= $J^{-1} \times Q$

Conclusions

- Fitting J-covariate together with genetic group is advisable to reduce level bias and inflation of genomic prediction.
- Fitting either J, Q or both together had marginal effects on genomic prediction accuracy.



Acknowledgments



Norges miljø- og
biovitenskapelige
universitet



Thank you for your attention!!