Using Monte Carlo method to include polygenic effects into SNP-BLUP reliability estimation

Hafedh Ben Zaabza

Esa Mäntysaari

Ismo Strandén

Email: hafedh.benzaabza@luke.fi

Introduction

- Genomic information has become important in dairy cattle breeding (Schaeffer, 2006)
- Two equivalent genomic models have been proposed for the prediction of genetic values:
  - GBLUP

    and
  - SNP-BLUP

Introduction

- Typically genetic variation cannot be completely captured by SNP markers because of the incomplete linkage disequilibrium between QTL and genome wide markers to predict genetic values.

- The GBLUP and SNP-BLUP models can be augmented by **residual polygenic (RPG)** effects.

Introduction

- When RPG effect is included in the SNP-BLUP MME, size of the MME is increased by the number of genotyped animals.

- The model reliability calculation will be computationally more challenging (for both models) when the number of genotyped animals increases.

Introduction

- Use approximation, when computation of the exact model reliability by direct inversion of the coefficient matrix of the MME becomes unfeasible.

- Our approach:
  1. Monte Carlo (MC) based approximation is used to include RPG effect into SNP-BLUP.
  2. Use the approximate SNP-BLUP model in calculation of model reliability.

Study design

- A small dairy cattle data:
  - 19,757 genotyped animals in Finnish Red dairy cattle
  - 11,729 SNP markers
- Calculate model reliability in SNP-BLUP model with a RPG effect using a MC sampling based method
  - RPG proportion: 0.01, 0.20, 0.50, 0.80
- Compare the approximate model reliability to the correct ones
  - MC sample sizes: 1000, 5000, 10000, 20000

Monte Carlo sampling to estimate numerator relationship matrix A

Let $\mathbf{a}_i$ be the breeding value vector of q related animals from MC sample i:

$$\mathbf{a}_i \sim N(\mathbf{0}, \mathbf{A}),\ i\text{=}1,\ldots,\ n_{\mathrm{MC}}$$

where **A** is the pedigree based relationship matrix.

After ordering the animals by age, breeding value for animal j can be generated by

$$\mathbf{a}_{ij} = \frac{\mathbf{1}}{\mathbf{2}}\left(\boldsymbol{a}_{i,sj} + \boldsymbol{a}_{i,dj}\right) + \boldsymbol{\varphi}_{ij}$$

$$\boldsymbol{\varphi}_{ij} \sim \boldsymbol{N}(0, \frac{1}{2}(1\text{-}\bar{F}))$$

Monte Carlo sampling to estimate numerator relationship matrix A

- $\mathbf{A} = Var(\boldsymbol{a}) = E(\boldsymbol{a}\boldsymbol{a}') - E(\boldsymbol{a})E(\boldsymbol{a})' = E(\boldsymbol{a}\boldsymbol{a}')$, because $E(\boldsymbol{a}) = \mathbf{0}$.

- Thus, a simple Monte Carlo based estimator of $\boldsymbol{A}$ is $\widehat{\mathbf{A}} = \frac{1}{n_{\mathrm{MC}}} \sum_{i=0}^{n_{\mathrm{MC}}} \boldsymbol{a}_i \, \boldsymbol{a}_i'$

➔ an MC estimator

$$\widehat{\mathbf{A}} = \frac{1}{n_{\mathrm{MC}}} [\boldsymbol{a}_1 \quad \cdots \quad \boldsymbol{a}_{n_{\mathrm{MC}}}] \begin{bmatrix} \boldsymbol{a}_1' \\ \vdots \\ \boldsymbol{a}_{n_{\mathrm{MC}}}' \end{bmatrix} = \frac{1}{n_{\mathrm{MC}}} \mathbf{U}\mathbf{U}'$$

where $\mathbf{U} = \sqrt{\frac{1}{n_{\mathrm{MC}}}} [\boldsymbol{a}_1 \quad \cdots \quad \boldsymbol{a}_{n_{\mathrm{MC}}}]$ is an q by n_{MC} matrix.

Monte Carlo SNP-BLUP (MC-SNP-BLUP)

- Define $\mathbf{G}_w = (1-w)\mathbf{Z}_c\mathbf{Z}_c' + w\mathbf{A}_{22}$ with the RPG effect → GBLUP

- MC approximation: $\mathbf{G}_w^* = (1-w)\mathbf{Z}_c\mathbf{Z}_c' + w\mathbf{U}_{22}\mathbf{U}_{22}' = \mathbf{SS}'$

  where $\mathbf{S} = [\mathbf{S}_\mathrm{Z} \quad \mathbf{S}_\mathrm{U}]$, $\mathbf{S}_\mathrm{Z} = \sqrt{1-w}\,\mathbf{Z}_c$, and $\mathbf{S}_\mathrm{U} = \sqrt{\boldsymbol{w}}\mathbf{U}_{22}$.

- An equivalent MC-SNP-BLUP model with RPG effect:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{S}\mathbf{g}_s + \mathbf{e}$$

- MME:

$$\begin{bmatrix} \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{1}_n & \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{S} \\ \mathbf{S}'\mathbf{R}^{-1}\mathbf{1}_n & \mathbf{S}'\mathbf{R}^{-1}\mathbf{S} + \mathbf{I}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\mathbf{g}_s} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{S}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Monte Carlo for reliability estimation

$$\begin{bmatrix} \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{1}_n & \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{S} \\ \mathbf{S}'\mathbf{R}^{-1}\mathbf{1}_n & \mathbf{S}'\mathbf{R}^{-1}\mathbf{S} + \mathbf{I}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\mathbf{g}_s} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{S}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Denote coefficient of the MME by $\mathbf{C}_s$ and its matrix elements as $\mathbf{C}_s^{-1} = \begin{bmatrix} \mathbf{C}_s^{\mu\mu} & \mathbf{C}_s^{\mu g} \\ \mathbf{C}_s^{g\mu} & \mathbf{C}_s^{gg} \end{bmatrix}$.

Model reliability of animal *j*: $r_j^2 = 1 - \frac{\mathbf{PEV}_j}{\mathbf{G}_{jj}^* \sigma_u^2}$ where

$\mathbf{PEV}_j$ = diagonal element *j* in the prediction error variance matrix: $\boldsymbol{s}_j \mathbf{C}_s^{gg} \boldsymbol{s}_j'$

$\mathbf{G}_{jj}^*$ = diagonal element *j* in the genomic relationship matrix $\mathbf{G}_w^* = \mathbf{SS}'$.

Data

- 19,757 genotyped animals in the Finnish Red dairy cattle

- Pedigree of the genotyped animals: 231,186 animals

- 11,729 SNP markers which are used in the joint Nordic genomic evaluations

- All genotyped animals were assumed to have 1 observation and no weights

Correlation (r) and maximum difference (max) and mean-squared error (MSE) between correct reliability from GBLUP and approximation by MC-SNP-BLUP under different number of MC samples (N_{MC}) and RPG (w)

N_{MC}	w	r	max	MSE ($\times 10^{-05}$)
1,000	0.01	1.000	0.00	0
	0.20	0.999	0.06	116
	0.50	0.992	0.24	2259
	0.80	0.974	0.47	9297
5,000	0.01	1.000	0.00	0
	0.20	1.000	0.02	7
	0.50	0.999	0.07	207
	0.80	0.997	0.17	1072
10,000	0.01	1.000	0.00	0
	0.20	1.000	0.01	2
	0.50	0.999	0.04	58
	0.80	0.999	0.09	322
20,000	0.01	1.000	0.00	0
	0.20	1.000	0.01	1
	0.50	1.000	0.02	16
	0.80	0.999	0.05	85

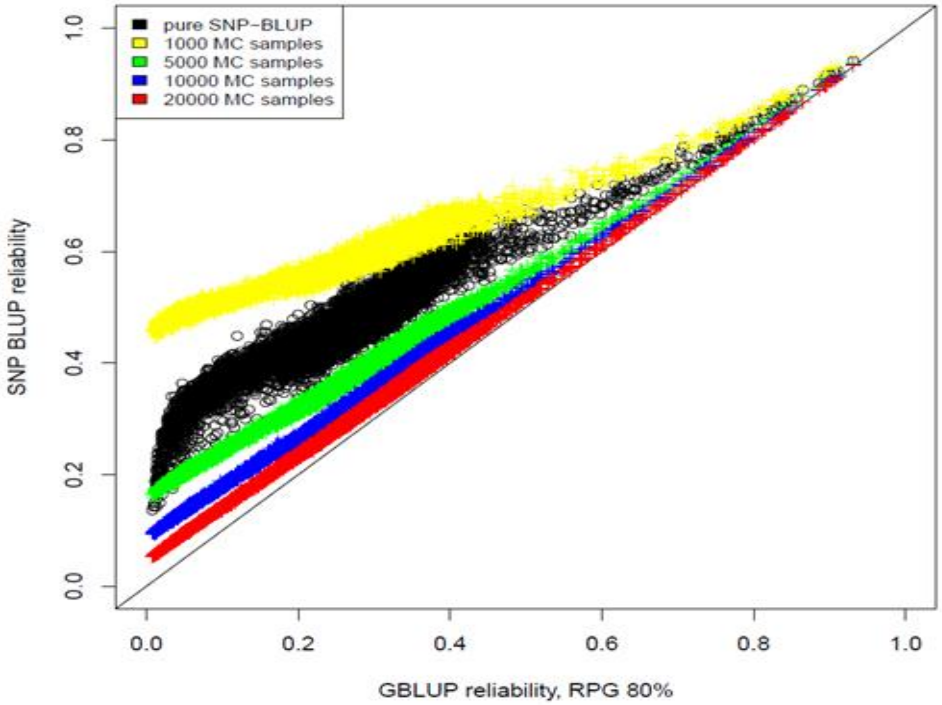
- Increasing number of MC samples, increased the correlation.
- This was clearer the larger the RPG proportion.

- The higher the MC sample size the lower the maximum error.

Exact GBLUP versus approximate MC-SNP-BLUP reliabilities for RPG of 80%

Large RPG proportion and small MC sample size (yellow):

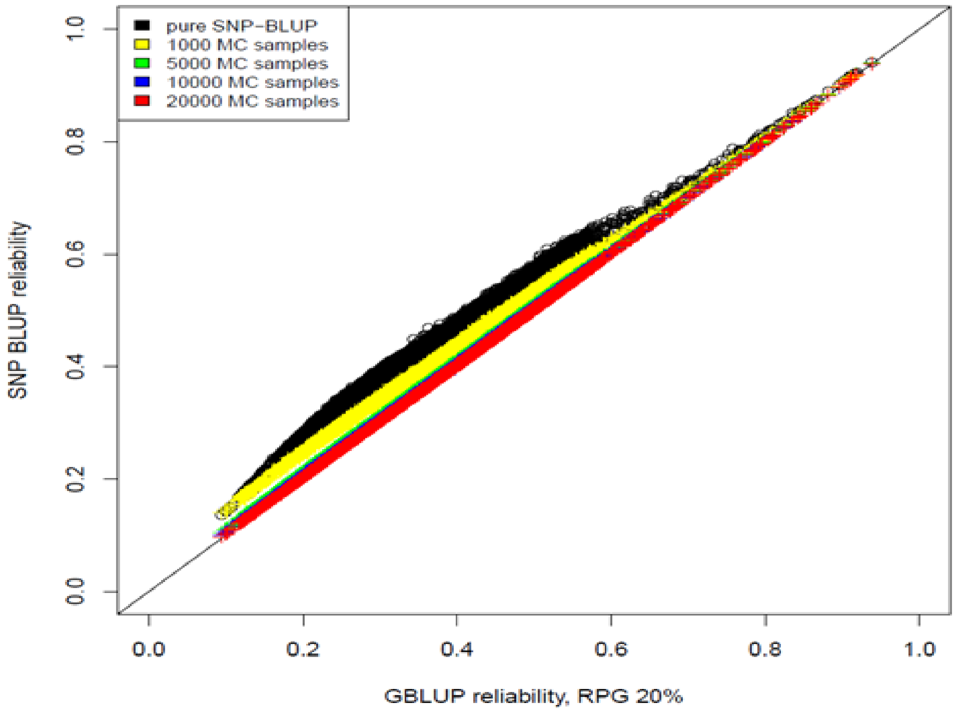
➔ reliabilities from MC-SNP-BLUP were <u>inflated markedly</u> when the reliability values were less than 0.6.



Exact GBLUP versus approximate MC-SNP-BLUP reliabilities for RPG of 20%

For RPG proportion of 20%:

MC-SNP-BLUP reliability deviation was considerably less compared with those for RPG of 80%.



Computing time (wall clock in seconds) for calculating model reliability in MC-SNP-BLUP

Computing step	1,000	5,000	10,000	20,000
MC-Sampling	4	18	33	71
Making MME	18	28	45	97
Inversion of MME	15	33	66	239
Total	78	131	220	536

Computing time (wall clock) for calculating model reliability in GBLUP

Computing step	GBLUP (sec)
Making G matrix	53
Inverting G matrix	50
Making MME	2
Inversion MME	52
Total	365

Conclusions

- The approximation gave high correlations with GBLUP model reliability even in the scenario with a low number of MC samples.

- More MC samples were needed to give small maximum absolute difference when the RPG was high.

- The promising results suggest that the MC approach is useful in the calculation of the genomic reliability from SNP-BLUP model with residual polygenic effect, especially when $n > m + n_{MC}$.

Thank you!

NATURAL RESOURCES
INSTITUTE FINLAND

Data 2 for testing the approach

- Irish beef cattle carcass conformation evaluation having several breeds
- 13.35 millions pedigree animals
- 222,619 genotyped animals
- 50,240 SNP markers
- Heritability of the trait was 0.27
- Effective record number was used as weight

Correlation (r) and maximum difference (max) and mean-squared error (MSE) between correct reliability from GBLUP and approximation by MC-SNP-BLUP under different number of MC samples (N_{MC}) and RPG (w)

More MC samples were needed to attain as high correlation and small MSE

N_{MC}	w	r	max	MSE ($\times 10^{-5}$)
5,000	0.20	0.999	0.07	242
	0.50	0.994	0.30	4323
	0.80	0.983	0.54	16945
10,000	0.20	1.000	0.05	81
	0.50	0.998	0.19	1822
	0.80	0.993	0.39	8280
20,000	0.20	1.000	0.03	24
	0.50	0.999	0.11	622
	0.80	0.998	0.24	3142
40,000	0.20	1.000	0.01	6
	0.50	1.000	0.06	184
	0.80	0.999	0.13	995
60,000	0.20	1.000	0.01	3
	0.50	1.000	0.04	81
	0.80	1.000	0.09	481

222,619 genotyped animals by 50,240 markers

Signed absolute maximum difference (max) and mean-squared error (MSE) between the true A_{22}diagonal elements and the Monte Carlo (MC) approximated A_{22} diagonal elements under different number of MC samples (N_{MC}) in analysis of Data 1* and Data 2**

Increasing the number of MC samples decreased both the maximum difference and MSE

	Data 1			Data 2	
N_{MC}	max	MSE ($\times 10^{-5}$)		max	MSE ($\times 10^{-5}$)
1,000	-0.21	216		-	-
5,000	-0.10	43		-0.10	41
10,000	-0.06	21		-0.07	20
20,000	0.04	11		0.05	10
40,000	-	-		0.04	5
60,000	-	-		-0.03	3

*19757 animals by 11,729 SNP markers

**222,619 animals by 50,240 SNP markers

Computing time (wall clock in minutes) for calculating model reliability in MC-SNP-BLUP under different Monte Carlo sample sizes during various computing steps. Data 2 (222,619 genotyped animals by 50,240 SNP markers)

Inversion of the MME was often computationally less expensive than making MME

Computing step	5,000	10,000	20,000	40,000	60,000
MC-sampling	11	19	38	78	141
Making MME	40	47	69	106	220
Inversion of MME	16	21	33	74	206
Total	140	170	252	432	861

Computing time (wall clock) for calculating model reliability in GBLUP for Data 1(19,757 genotyped animals by 11,729 SNP markers and Data 2(222,619 genotyped animals by 50,240 SNP markers)

Computing step	Data 1(sec)	Data 2(min)
Making G matrix	53	276
Inverting G matrix	50	1116
Making MME	2	216
Inversion MME	52	1092
Total	365	3090