Accuracy of genomic prediction by singular value decomposition of the genotype matrix

L. Ayres, M.P.L. Calus, J. Ødegård, T. Meuwissen

74th EAAP Annual Meeting – Lyon

August 30, 2023

Authors

- Lucas Ayres, Wageningen University & Research
- Mario Calus, Wageningen University & Research
- Jørgen Ødegård, Norwegian University of Life Sciences and AquaGen AS
- Theo Meuwissen, Norwegian University of Life Sciences

Aim

↪ Evaluate the effect of the number of components in singular value decomposition (SVD) of the genotype matrix on the accuracy of genomic prediction.

Singular Value Decomposition

Definition: Orthogonal matrix

A matrix $\mathbf{U} \in \mathbb{R}^{n \times m}$ is said to be *orthogonal* iff

$$\mathbf{U}^{\intercal}\mathbf{U} = \mathbf{I}_m$$

Singular Value Decomposition

Definition: Eigenvector and eigenvalue

A non-null vector $\mathbf{v}$ is called an *eigenvector* of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ when there is a scalar λ such that

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$$

λ is said to be the *eigenvalue* of $\mathbf{A}$ associated to the eigenvector $\mathbf{v}$.

Singular Value Decomposition

Theorem: SVD

For every matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, there exist two orthogonal matrices $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{m \times m}$ such that

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\intercal$$

where $\mathbf{S} \in \mathbb{R}^{n \times m}$ is a diagonal matrix whose non-zero values are the square roots of the eigenvalues of $\mathbf{A} = \mathbf{X}^\intercal \mathbf{X}$.

We call the diagonal values of $\mathbf{S}$ the *singular values* of $\mathbf{X}$.

Singular Value Decomposition

$$\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \begin{bmatrix} * & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 \\ 0 & 0 & * & 0 & 0 \end{bmatrix} \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix}$$

Principal Component Ridge Regression

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$[\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I}]\,\hat{\mathbf{b}} = \mathbf{X}^\intercal\mathbf{y}$$

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\intercal$$

$$\mathbf{T} = \mathbf{U}^\intercal\mathbf{S}$$

$$[\mathbf{S}^\intercal\mathbf{S} + \lambda\mathbf{I}]\,\hat{\mathbf{s}} = \mathbf{T}^\intercal\mathbf{y}$$

$$\hat{\mathbf{b}} = \mathbf{V}\hat{\mathbf{s}}$$

Truncated-SVD version of SNP-BLUP

$$\mathbf{X} \approx \mathbf{U}_k\mathbf{S}_k\mathbf{V}_k{}^\intercal$$

$$\mathbf{T} = \mathbf{U}_k{}^\intercal\mathbf{S}_k$$

$$[\mathbf{S}_k{}^\intercal\mathbf{S}_k + \lambda\mathbf{I}]\,\hat{\mathbf{s}}_k = \mathbf{T}^\intercal\mathbf{y}$$

$$\hat{\mathbf{b}} = \mathbf{V}_k\,\hat{\mathbf{s}}_k$$

Materials and methods

- Genotypes* from 1,927 Atlantic Salmon (*Salmo salar*) fish
  - 16,454 SNP markers (all on chromosome 1)

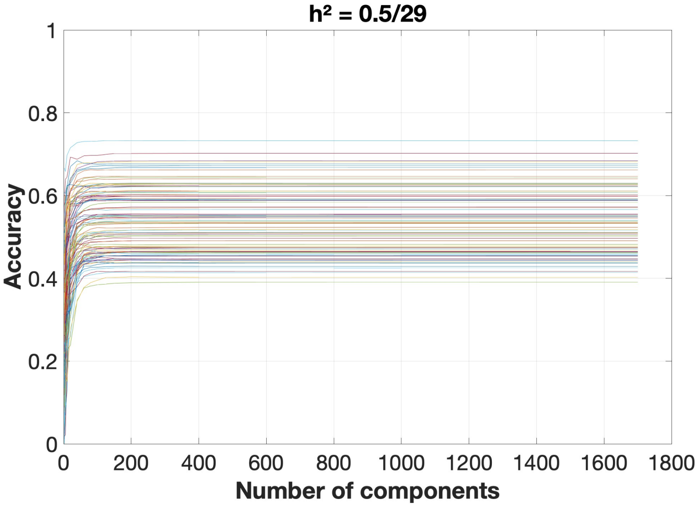
*Data provided by AquaGen AS.

Simulation

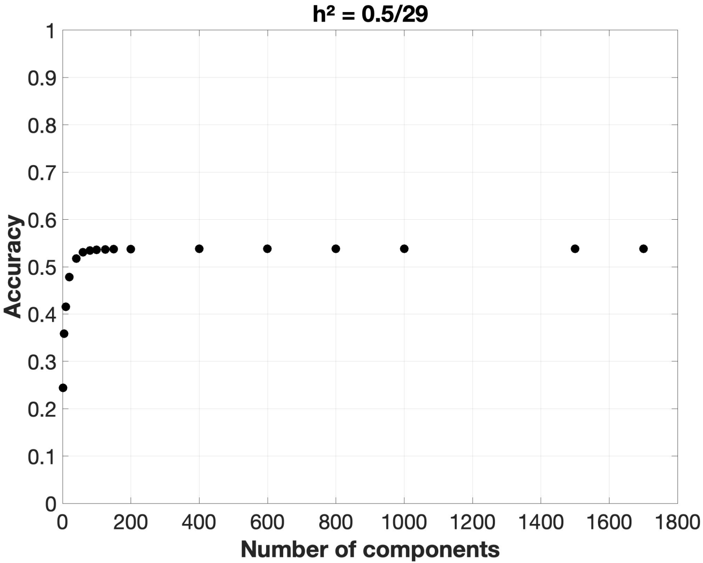
- $b_1, \ldots, b_{1000}$ i.i.d. $b_i \sim N(0, V_m)$, where $V_m = \frac{h^2}{2\sum_{i=1}^{1000} p_i(1-p_i)}$
- $h^2 = \left\{{}^{0.1}\!/_{29}, {}^{0.3}\!/_{29}, {}^{0.5}\!/_{29}\right\}$
- 1,000 QTL randomly positioned along the 16,454 SNP loci
- mask the QTL from the genotype matrix
- calculate true breeding values $\mathbf{g} = \mathbf{Xb}$
- predict breeding values with PCRR $\hat{\mathbf{g}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}\mathbf{V}_k\hat{\mathbf{s}}_k$
  - (10-fold cross-validation)
- estimate correlation coefficient $\hat{r} = \frac{Cov(g,\hat{g})}{\hat{\sigma}_g\hat{\sigma}_{\hat{g}}}$
- run 100 replicates and obtain average accuracy $\bar{r} = \sum_{j=1}^{100} \hat{r}_j/100$

$k = \{2, 5, 10, 20, 40, 60, 80, 100, 125, 150, 200, 400, 600, 800, 1000, 1500, 1700\}$

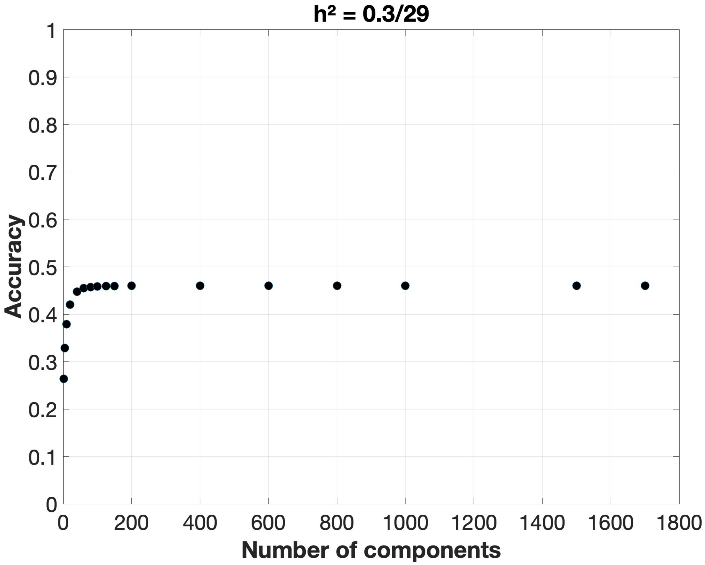
Results


h² = 0.5/29
Accuracy
0
0.2
0.4
0.6
0.8
1
Number of components
0
200
400
600
800
1000
1200
1400
1600
1800

Results

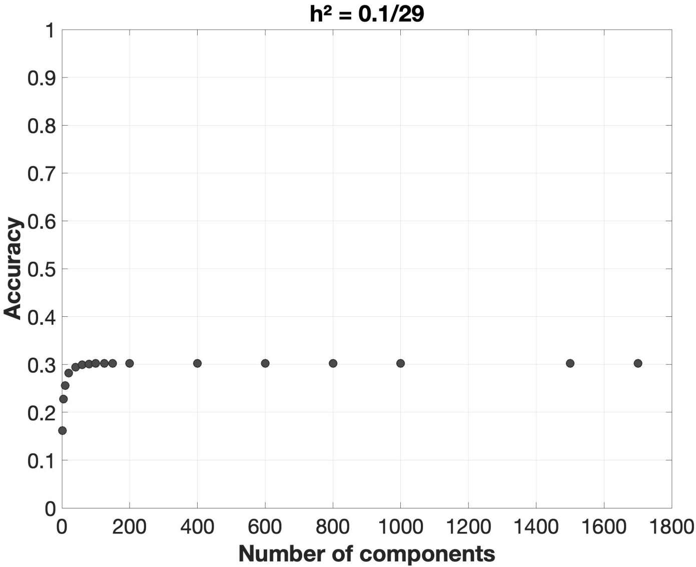

h² = 0.5/29
Accuracy
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Number of components
0
200
400
600
800
1000
1200
1400
1600
1800

Results

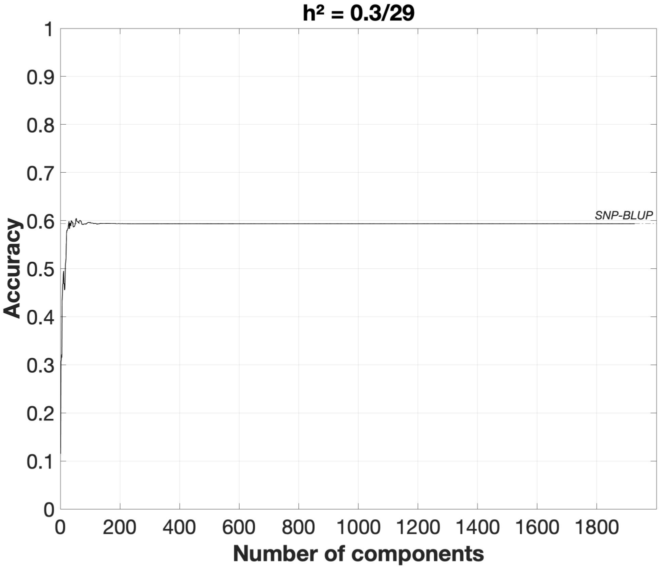


h² = 0.3/29
Accuracy
Number of components
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
200
400
600
800
1000
1200
1400
1600
1800

Results

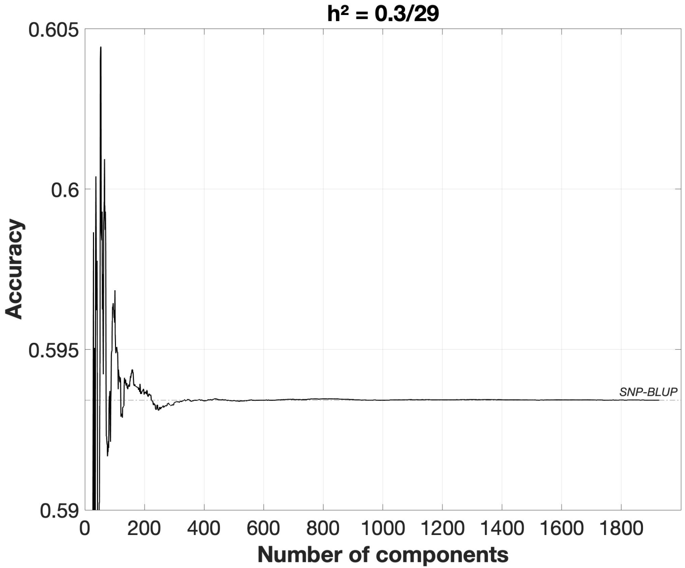

h² = 0.1/29
Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
200
400
600
800
1000
1200
1400
1600
1800
Number of components

Results



h² = 0.3/29
Accuracy
Number of components
SNP-BLUP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
200
400
600
800
1000
1200
1400
1600
1800

Results



h² = 0.3/29
Accuracy
0.605
0.6
0.595
0.59
0
200
400
600
800
1000
1200
1400
1600
1800
Number of components
SNP-BLUP

Discussion

- reduction of statistical noise
- accuracy not always an increasing function of the number of SVD components
- higher accuracy using APY (e.g., 0.5% gain)
- higher accuracy using PCRR/PCIG (e.g., 1.8% gain)

Conclusions

- SVD is useful for data reduction of the genotype matrix
- PCRR can be used for genomic prediction
  - good accuracies obtained with few components
- PCRR can provide higher accuracies than SNP-BLUP with certain numbers components
  - within replicates, maximum accuracy at 50–250 components
  - across replicates, maximum mean accuracy at 400–600 components

References

Ødegård, J., Indahl, U., Strandén, I., & Meuwissen, T. 2018. Large-scale genomic prediction using singular value decomposition of the genotype matrix. *Genetics Selection Evolution*, 50, 6.

Ayres, L. L. 2022. *The accuracy of genomic prediction by singular value decomposition of the genotype matrix*. Master's thesis. Ås: Norwegian University of Life Sciences.