Genomic models and BLUP

Ole F. Christensen

Center for Quantitative Genetics and Genomics, Aarhus University

BovReg course - September 2023

Genetic markers

- Fragments of DNA associated with certain locations in genome.
- Must be polymorphic
- Many types: Microsatelittes, insertions/deletions, Single Nucleotide Polymorphisms (SNP), etc

SNP data

- SNP's have two alleles (diallelic)

	SNP_1	SNP_2	SNP_2	
ID_1	AB	BB	BB	...
ID_2	AA	AB	BB	...
ID_3	AA	AA	AB	...
ID_4	BB	BB	AA	...
...	...	...	...	

- Chose a reference allele

- Coding: 0,1,2 copies of the reference allele

SNP data

100011122002001211101111211110111100112110002012200222011 1
120210120021112211002111200111100101101101022001100220110 1
120020110102022212112210201001110001122022122211202112012 0
201002022020000211000112020112211121110220111100002122020 0
022101202000221122011101210011121110211211002010210002200 0
220100020110000220221102211210112111012222001211212220020 0
020020202012221100222222200221211112100211112001101110112 0
020222000111201101021112121110202210021120121100111110211 1
211021112200010110111020220022111010201112111101120210210 2
121101102212200121101121101202201100222002100211000111002 1
102110111000222002022121211000222010200222212122112111200 2
011020200122222211221202121121011001211011020022000200100 2
000111101100121102121211120101012120221010101111021102112 1
211111121211121011012001111102111101111122012101212110102 2
202021211222120222002121210121210201100111222121101

SNP-BLUP

- Model: $\mathbf{y} = \mathbf{Xb} + \sum_i \mathbf{Z}_i g_i + \mathbf{e}$
- g_j's are SNP effects (genetic)
- Centering of marker codes: $\mathbf{Z}_j = \mathbf{M}_j - 2p_j\mathbf{1}$; relate SNP effects to individuals
- Compact notation: $\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$
- SNP-BLUP: assumes SNP effects are normally distributed and independent

Mixed model equations for SNP-BLUP

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \mathbf{I}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix},$$

where $\alpha = \sigma_e^2/\sigma_g^2$.

- MME provide BLUP SNP effects $\hat{\mathbf{g}}$
- GEBV on individual i is $\hat{a} = \sum_i \mathbf{Z}_{\cdot,i}\hat{g}_i$

Equivalent model - GBLUP

- Genetic effects $\mathbf{a} = \mathbf{Zg}$
- Genetic variance-covariance:

$$\mathtt{Var}(\mathbf{a}) = \mathbf{Z}\mathtt{Var}(\mathbf{g})\mathbf{Z}^T = \sigma_g^2 \mathbf{Z}\mathbf{Z}^T$$

- Model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + \mathbf{e}$$

where

$$\mathtt{Var}(\mathbf{a}) = \sigma_a^2 \mathbf{G},$$

$$\mathbf{G} = \mathbf{Z}\mathbf{Z}^T / (\sum_j 2p_j(1-p_j))$$

- Easy to compute
- Genetic variance $\sigma_a^2 = \sigma_g^2 \sum_j 2p_j(1-p_j)$

GBLUP versus Animal model (pedigree **A**)

- GBLUP is an animal model
- Genomic versus Pedigree relationships
- Realised (marker genotypes) versus expected (pedigree) relationships
- More accurate estimate of proportion of chromosome segments shared between individuals
- GBLUP is an improved version of pedigree BLUP

Mixed model equations for GBLUP

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{W} \\ \mathbf{W}^T\mathbf{X} & \mathbf{W}^T\mathbf{W} + \mathbf{G}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{y} \end{bmatrix},$$

where $\alpha = \sigma_e^2/\sigma_a^2$.

- MME provide GEBVs on individuals $\hat{\mathbf{a}}$

Polygenic effects

- If Markers are not capturing all the genetic effects, then add a residual polygenic effect

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wu} + \mathbf{Wa} + \mathbf{e}$$

- $\mathbf{u}$: residual polygenic effect, $\mathtt{Var}(\mathbf{u}) = \sigma_u^2 \mathbf{A}$

- Mixed model equations

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{W} & \mathbf{X}^T\mathbf{W} \\ \mathbf{W}^T\mathbf{X} & \mathbf{W}^T\mathbf{W} + \mathbf{A}^{-1}\alpha_1 & \mathbf{W}^T\mathbf{W} \\ \mathbf{W}^T\mathbf{X} & \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{W} + \mathbf{I}\alpha_2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{y} \end{bmatrix},$$

  where $\alpha_1 = \sigma_e^2/\sigma_u^2$ and $\alpha_2 = \sigma_e^2/(\sigma_a^2 - \sigma_u^2)$.

- Genetic variance is $\sigma_a^2 = \sigma_u^2 + (\sigma_a^2 - \sigma_u^2)$.

ssGBLUP

- A general approach to handle the situation that some animals are genotyped whereas others are not.
- Essential idea: Construct a combined relationship matrix across all animals, using both marker genotypes and pedigree, and use a model for all animals.
- Special case where no animals are genotyped: Animal model based on pedigree relationships.
- Special cased where all animals are genotyped: GBLUP

Combined relationship matrix

- Pedigree-based relationship matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

  where split is into non-genotyped and genotyped animals.

- Example

ID	FATHER	MOTHER	
1	0	0	
2	0	0	genotyped
3	1	2	
4	1	2	genotyped

Combined relationship matrix

Example:

ID	FATHER	MOTHER	
1	0	0	
2	0	0	genotyped
3	1	2	
4	1	2	genotyped

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}$$

Combined relationship matrix

- Pedigree-based relationship matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

  where split is into non-genotyped and genotyped animals (note: animals reordered).

- A first naive attempt:

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix}$$

  - Doesn't adjust relationships within non-genotyped and between genotyped and non-genotyped animals depending on $\mathbf{G}$
  - Invalid (not a positive definite matrix).

Combined relationship matrix

- Pedigree-based relationship matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

  where split is into non-genotyped and genotyped animals

- Second attempt

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}.$$

  - Modify relationships depending on $\mathbf{G}$

Combined relationship matrix

- Model: $\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + e$, where $\mathbf{a} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{H})$
- Matrix $\mathbf{H}$ has sparse inverse

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} + \mathbf{A}^{-1}$$

- Matrix $\mathbf{H}$ is valid (positive definite)
- Easy to compute, but not based on tracing of inheritance (sub-optimal)
- $\mathbf{H}^{-1}$ is provided as input to genetic evaluation software

Combined relationship matrix

- Model: $\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + e$ where $\mathbf{a} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{H})$
- Matrix $\mathbf{H}$ has sparse inverse

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} + \mathbf{A}^{-1}$$

- Special case: all animals genotyped, then $\mathbf{H} = \mathbf{G}$
- Special case: no animals genotyped, then $\mathbf{H} = \mathbf{A}$

Compatibility of **G** and **A**

- Relationships and inbreeding in **A** are relative to the base population
- Relationships and inbreeding in **G** are relative to allele frequencies
- Adjust $\mathbf{G}_a = b\mathbf{G} + a\mathbf{1}\mathbf{1}^T$ such that $\mathbf{G}_a$ and $\mathbf{A}_{22}$ are compatible (ad hoc).
- Metafounders in pedigree and $\mathbf{G}_{0.5}$ with allele frequencies 0.5 (more elegant)

Genomic models and BLUP

- SNPBLUP, GBLUP, ssGBLUP presented here
- Other genomic models:
  - Non-additive genomic relationship matrices (Dominance, G*G, G*E)
  - Breed-specific (partial) genomic relationship matrices.
- Non-genetic similarity matrices (e.g. metabolomics).
- Standard genetic evaluation software with BLUP is powerful

BLUP vs Bayes+McMC - pros/cons

- Pros (compared to Bayes + McMC)
  - Usually computationally fast
  - No Monte Carlo noise in (important for routine genetic evaluation)
  - No need to asses convergence and mixing of McMC
- Cons (compared to Bayes + McMC)
  - Parameter uncertainty not incorporated in prediction (important?)
  - Not always sufficiently flexible (important for research)