Genomic relationship matrix when some animals are not genotyped

(S 28)

Ole F. Christensen and Mogens S. Lund
Faculty of Agricultural Sciences, Aarhus University
E-mail: OleF.Christensen@agrsci.dk

EAAP 2009, Barcelona

Genomic prediction models

- Most approaches based on estimating effects of different SNPs (in a Bayesian setting using McMC) and obtaining EBVs by summing effects.

- Some challenges are :
  - Computational (speed, mixing and convergence of McMC),
  - Integration of all information (using EBV or DYD as response, blending pure genomic EBV with traditional EBV).

- The approach here is different:
  - Markers used to construct a relationship matrix G.
  - EBV are BLUPs in a linear mixed model.
  - Integration/blending of information by : 1) extending G to all animals, 2) including a polygenic effect.

Genomic Relationship matrix

- The Genomic relationship matrix G based on SNPs (from van Raden, 2008):

$$G = (m - p)(m - p)^T / s$$

- where

$$m_{ij} = \begin{cases} -1 & g_{ij} = 11 \\ 0 & g_{ij} = 12 \\ 1 & g_{ij} = 22 \end{cases}$$

-

$$p_j = 2\rho_j - 1, \quad s = 2\sum_j \rho_j(1 - \rho_j)$$

- ρ_j allele-frequency of allele j.

Extend G to non-genotyped animals: motivation

- Model :

$$y = X\beta + a + g + e$$

  where

$$a \sim N(0, \sigma_a^2 A), \quad g \sim N(0, \sigma_g^2 G^*)$$

- Need genomic values g for all animals.

- Therefore, need to extend G to all animals (G^*).

- A combined GEBV : $\hat{g} + \hat{a}$ for all animals.

Extend G to non-genotyped animals

- A joint model for genomic value and markers:

$$[g \mid M] \sim N(0, \sigma_g^2 (M-p)(M-p)^T/s)$$

$$E[M_j] = p_j 1, \quad \mathrm{Var}[M_j] = 2\rho_j(1-\rho_j)A$$

(based on idea by Gengler et al. 2007 to infer missing genotypes).

- Individuals with missing and observed genotypes

$$M = \begin{bmatrix} m^{obs} \\ M^{miss} \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

Extend G to non-genotyped animals

- Marginalisation (integration) gives

$$E[g \mid m^{obs}] = 0,$$

$$\mathrm{Var}[g \mid m^{obs}] = \sigma_g^2 \begin{bmatrix} G & GA_{11}^{-1}A_{12} \\ A_{21}A_{11}^{-1}G & A_{21}A_{11}^{-1}GA_{11}^{-1}A_{12} + A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

$$= \sigma_g^2 G^*.$$

- Same extension as Legarre, Augilar and Misztal (2009, to appear in J. Dairy Sci.)

Add a polygenic effect

- $$y = X\beta + a + g + e$$

  where

  $$a \sim N(0, \sigma_a^2 A), \quad g \sim N(0, \sigma_g^2 G^*)$$

- Markers may not capture all genetic differences.

- Combined genetic value : $\tilde{g} = g + a$.

- Polygenic weight $w = \sigma_a^2/(\sigma_a^2 + \sigma_g^2)$.

- Variance

  $$\mathrm{Var}[\tilde{g}] = \sigma_{\tilde{g}}^2((1-w)G^* + wA) = \sigma_{\tilde{g}}^2 G_w^*$$

Sparse inverse

- $$(G_w^*)^{-1} = \begin{bmatrix} G_w^{-1} - A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + A^{-1}.$$

  where

  $$G_w = (1-w)G + wA_{11}.$$

  This is a sparse matrix !!

- Direct computation of A^{-1} in sparse format is well-known.

- A_{11} can be computed from A^{-1} using sparse matrix computation.

- G_w is invertible (even G is often not full rank).

Inference using Sparse inverse

- Parameter estimation using AI-REML (based on solving sparse MME) implemented in software DMU.

- Prediction by solving sparse MME (implemented in DMU)

- Polygenic weight w estimated from data (but $w > 0$).

Simulation study

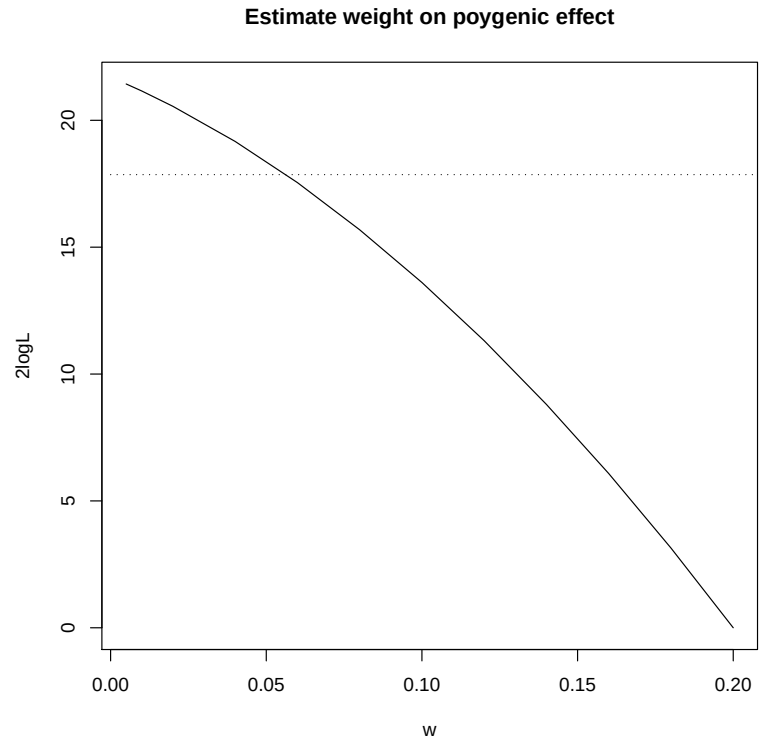
Inspired by a nucleous pig breeding scheme (simplified).

- 150 boars and 1500 sows produce 15000 offspring (50 % males).
- For next generation: 150 males selected based on phenotype, 1500 females selected randomly.
- 5 generations with phenotype on all males (7500*5).
- SNP panel of size 5000, and 500 true QTLs.
- The 150*3 selected males in generation 3, 4, and 5 are genotyped.
- The 150 males in generation 0 (no phenotype) are genotyped.
- 300 males in generation 6 (no phenotype) are genotyped.

Simulation study

- 46950 animals in pedigree
- 37500 animals with phenotype
- 900 animals with genotype
- 450 animals with both genotype and phenotype.
- Genotypes of 150 base animals used for allele freqeuncies. But also contain information about unknown genotype of other animals.
- Genotypes of 300 selection candidates contain information about the (unknown) genotypes of their mothers.

Simulation study - results



- Estimated polygenic effect is about 0.

- For computational reasons, we use $\hat{w} = 0.01$.

Simulation study - results

- Estimated parameters (with $w = 0.01$):

$$\sigma^2_{\tilde{g}} = 4.16, \quad \sigma^2_e = 16.22$$

- Predictions,

$$Cor(GEBV, trueBV) = 0.660$$

- For comparison, an alternative approach : Analyse 600 animals (generation 0, 3,4,5) where EBV (computed in an animal model) is response, and predict 300 genotyped animals.
  - Estimated parameters (with $w = 0.01$):

$$\sigma^2_{\tilde{g}} = 7.56, \quad \sigma^2_e = 0.069$$

  - Predictions,

$$Cor(GEBV, trueBV) = 0.587$$

Discussion

- Using the extended G :
  - No preprocessing of phenotypes to EBVs or DYDs for a genomic selection model (causing possible bias).
  - Improves prediction.
- Equal weight on markers (alternative: some areas high weight).
- Computation of allele-frequencies in founders is an issue.
- The computational bottlenecks for the methods seem to be :
  - The computation of G, (computing time: $O(n_{obs}^2 n_{snp})$).
  - The computation $(G_w)^{-1}$, (computing time: $O(n_{obs}^3)$).
  - The storage of G, ($O(n_{obs}^2)$).