gembloux
agro bio tech

Université de Liège

2010 ADSA-ASAS Joint Annual Meeting
Denver, CO, July 10-15

Working on a Method to Compute Inverse of Genomic Relationship Matrix from Sparse Matrices

P. Faux *,[1], I. Misztal[2], N. Gengler [1,3]

[1]University of Liège, Gembloux Agro-Bio Tech, Belgium
[2]University of Georgia, Animal and Dairy Science, Athens GA
[3]National Fund for Scientific Research, Brussels, Belgium

aides à la formation recherche

Fonds National de la Recherche Luxembourg

* Supported by the National Research Fund, Luxembourg (FNR)

Acknowledgements

- Supported by National Research Fund of Luxembourg (FNR), PHD-09-118
- Carried out at the Animal and Dairy Science (ADS) Department of University of Georgia (UGA): Mrs. C.Y. Chen, R. Simeone, R.A. Davis
- Animal Breeding and Genetics (ABG) Group of Animal Science Unit of Gembloux Agro-Bio Tech (GxABT) of University of Liège
- INRA: Dr. A. Legarra

Introduction

- Genomic relationship matrices(**G**) are used in GBLUP
- As the cost of genotyping decreases, the number of animals genotyped will increase
- An approximation of $\mathbf{G}^{-1}$ that is easily computed is needed

Objective

- To create an approximation of $\mathbf{G}^{-1}$: $\mathbf{G}_a^{-1}$
  - Involves 3 steps:
    1) To develop subroutines in order to provide realistic simulations of genotypes
    2) To develop a method for creating $\mathbf{G}_a^{-1}$
    3) Using simulations, to assess this method and to compare $\mathbf{G}^{-1}$ and $\mathbf{G}_a^{-1}$

Simulation: Methods

- Assumptions
    - Bi-allelic markers: alleles 0 and 1
    - Allelic frequencies are uniformly distributed (from 0.05 to 0.95)
    - Marker loci may be in linkage disequilibrium (LD)
    - Number of crossing-over (CO) follows Poisson’s distribution of average λ

- Inputs: NCHR, LCHR, λ, NBSNP, PED(nx3)

- Output: Matrix of genotypes (n x NBSNP)

Simulation: Implementation

- Performed using a package of Fortran 90 subroutines

- Steps of simulation for one given set of inputs:
  1) Spread marker loci across chromosomes
  2) Choose theoretical minor allele frequency (MAF) for each loci
  3) Simulate genotypes for founders, according to MAF
  4) Simulate genotypes for other animals, according to pedigree and λ

Estimation of G_a^{-1}: Methods

- Based on a similar decomposition to decomposition of $\mathbf{A}^{-1}$ (Henderson, 1976):
  - $\mathbf{T}^{-1}$ is equal to **I-M**, with **M** containing 0.5 for each parent known at each line (for founders, $\mathbf{T}^{-1}=\mathbf{I}$)
  - **D** is a diagonal matrix, inverted elements by elements

- $\mathbf{A}^{-1}=(\mathbf{T}^{-1})' \,.\, \mathbf{D}^{-1} \,.\, \mathbf{T}^{-1} \Longrightarrow \mathbf{G}_a^{-1}=(\mathbf{T}_a^{-1})' \,.\, \mathbf{D}_a^{-1} \,.\, \mathbf{T}_a^{-1}$

- What is the best possible way to fit $\mathbf{G}_a^{-1}$ to $\mathbf{G}^{-1}$?
  1) Fit $\mathbf{T}_a^{-1}$ to **G**
  2) Fit $\mathbf{D}_a^{-1}$ to $\mathbf{T}_a^{-1}$

Estimation of G_a^{-1}: Methods

Fitting $\mathbf{T}_a^{-1}$ to $\mathbf{G}$:

- In the case of $\mathbf{A}^{-1}$, $\mathbf{T}^{-1}$, triangular matrix, contains:
  - 1, for the animal (diagonal element)
  - -0.5, for each parent: -0.5 elements can be viewed as a reverse of tabular method:
    - $a_{i,j} = 0.5 * a_{s,j} + 0.5 * a_{d,j}$, for all animals older than *i*
    - Given the equation of decomposition of $\mathbf{A}^{-1}$

      $$\mathbf{D}=\mathbf{T}^{-1} . \mathbf{A} . (\mathbf{T}^{-1})',$$

      it returns $d_{i,j} = 1 * a_{i,j} - 0.5 * a_{s,j} - 0.5 * a_{d,j}$

      ... = 0, for all animals older than *i*
  - 0, elsewhere

Estimation of G_a^{-1}: Methods

Fitting $\mathbf{T}^{-1}_a$ to $\mathbf{G}$:

- In the case of $\mathbf{G}_a^{-1}$: modification of -0.5 elements to create a $\mathbf{T}_a^{-1}$ that minimize off-diagonals elements of $\mathbf{D}_a$
- The modification is estimated by OLS

If i, s and d are respectively positions of animal, sire and dam in pedigree,

$g_{i,j}$ is the genomic relationship between animals i and j,

k is the number of animals on which estimation is performed,

Then, for animal i, β_s and β_d are the coefficients estimated for sire and dam respectively, as follows:

$$\begin{pmatrix} g_{i,1} \\ .. \\ g_{i,k} \end{pmatrix} = \begin{pmatrix} g_{s,1} & g_{d,1} \\ .. & .. \\ g_{s,k} & g_{d,k} \end{pmatrix} \cdot \begin{pmatrix} \beta_s \\ \beta_d \end{pmatrix} + \begin{pmatrix} e_1 \\ .. \\ e_k \end{pmatrix}$$

where e_1 to e_k are errors of estimation.

Estimation of G_a^{-1}: Methods

Fitting $\mathbf{T}_a^{-1}$ to **G**:

- Selection of *k* animals on which estimation is performed for *i*:
  - Animals are older than *i*, to avoid offspring of *i*
  - Animals are related to *i*
  - Estimated coefficients are expected around 0.5

- If only one parent known, estimations are biased: coefficients around 0.9, where 0.5 was expected
  - These are considered founders

- -0.5 elements in $\mathbf{T}^{-1}$ move to $-\beta_s$ and $-\beta_d$ in $\mathbf{T}_a^{-1}$

Estimation of G_a^{-1}: Methods

Fitting $\mathbf{D}_a^{-1}$ to $\mathbf{T}_a^{-1}$:

- $\mathbf{D}_a^{-1}$ directly computed using $\mathbf{T}_a^{-1}$ and $\mathbf{G}$ based on $\mathbf{D}_a=\mathbf{T}_a^{-1} \,.\, \mathbf{G} \,.\, (\mathbf{T}_a^{-1})'$

- Relationships between founders ($\mathbf{G}_F$) cannot be approximated

- Sub-matrix $\mathbf{G}_F$ is inverted, sub-matrix $\mathbf{G}_D$ is computed and both added to $\mathbf{D}_a^{-1}$

- Finally,

$$\mathbf{D}_a^{-1} = \begin{pmatrix} \mathbf{G}_F^{-1} & \mathbf{0} \\ \mathbf{0} & \left[\mathbf{T}_{a,D}^{-1} \cdot \mathbf{G}_D \cdot (\mathbf{T}_{a,D}^{-1})'\right]^{-1} \end{pmatrix}$$

Simulation: Results

- Simulation Parameters:
  - NCHR=32; LCHR=6E+6; λ =1.6;
  - NBSNP=5,000,
  - 400 animals (8 generations) in pedigree: 88 founders + 312 animals with both parents known

- **G** computed as **ZZ**' product and scaled according to VanRaden(2008), by $2*\Sigma(p_i * (1-p_i))$

- $\mathbf{G}_a^{-1}$ returned is sparse:
  - 148,242 elements equal to 0, i.e. 92.65%
  - Compared with $\mathbf{A}^{-1}$, where 98.53% equal to 0

Simulation: Results

- Comparisons between **G** and **G**$_a$

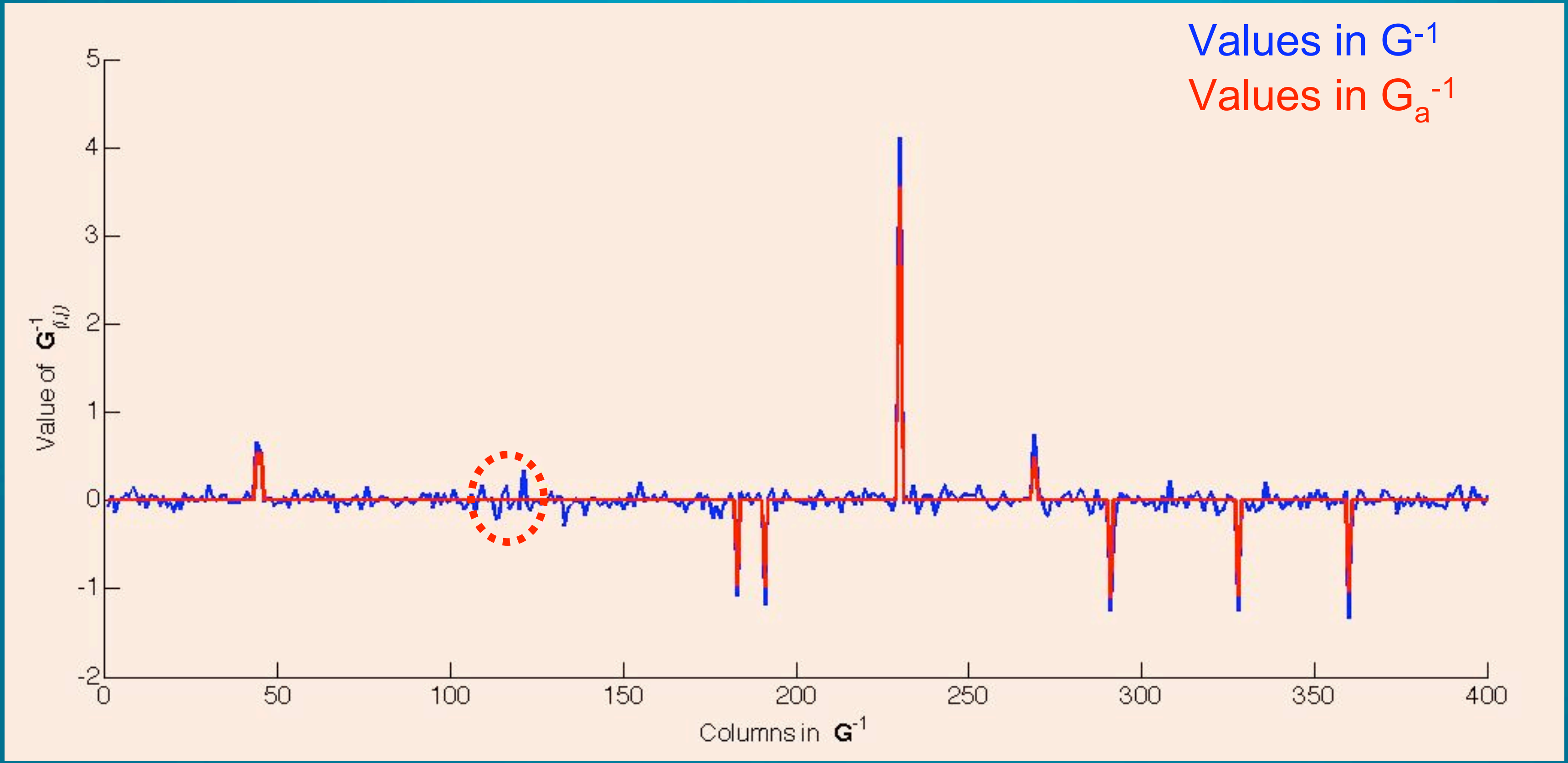
		min	mean	max
Diagonal elements	**G**	1.73	1.83	2.03
	G$_a$	1.71	1.83	2.01
Off-Diagonal elements	**G**	0.69	0.85	1.48
	G$_a$	0.72	0.85	1.51

- Correlations between diagonals
  - **G** vs. **G**$_a$: 0.91
  - **G**$^{-1}$ vs. **G**$_a^{-1}$: 0.99

- Coefficients β of **T**$_a^{-1}$ range from 0.4 to 0.6

Simulation: Results

$\mathbf{G}^{-1}$ and $\mathbf{G}_a^{-1}$ for a given animal (having both parents known)



Discussion and Conclusion

- Fewer operations are needed using this method than real inversion (if number of founders is small)
- It returns an approximation of **G** that is close to the same range of values as real **G**
- Need to be diversified for particular cases (animals with only one parent known)
- Would it be possible to estimate the greater elements of $\mathbf{G}^{-1}$ which are now considered equal to 0 in $\mathbf{G}_a^{-1}$?