Bayesian inference of genetic similarity among individuals from markers and phenotypes

Rohan Fernando and Daniel Gianola

February 13, 2018

Iowa State University and University of Wisconsin

Genetic Similarity

- Pre-genomic
  - Genotypes for quantitative traits could not be observed

Genetic Similarity

- Pre-genomic
  - Genotypes for quantitative traits could not be observed
  - Expected genetic similarity between relatives played a key role in modeling covariances

Genetic Similarity

- Pre-genomic
  - Genotypes for quantitative traits could not be observed
  - Expected genetic similarity between relatives played a key role in modeling covariances
  - Methods for prediction of breeding values and estimation of variance components relied on models for the covariance between relatives conditional on pedigrees

Genetic Similarity

- Pre-genomic
  - Genotypes for quantitative traits could not be observed
  - Expected genetic similarity between relatives played a key role in modeling covariances
  - Methods for prediction of breeding values and estimation of variance components relied on models for the covariance between relatives conditional on pedigrees
- Genomic:
  - Genotypes can be observed
  - Genetic covariance models are not necessary
  - Can estimate effects of genotypes directly
  - Inferences are based on estimated effects

Genetic Similarity

- Pre-genomic
  - Genotypes for quantitative traits could not be observed
  - Expected genetic similarity between relatives played a key role in modeling covariances
  - Methods for prediction of breeding values and estimation of variance components relied on models for the covariance between relatives conditional on pedigrees
- Genomic:
  - Genotypes can be observed
  - Genetic covariance models are not necessary
  - Can estimate effects of genotypes directly
  - Inferences are based on estimated effects
  - Observed genetic similarity matrix is **not** proportional to a genetic covariance matrix (J Anim Breed Genet. 2017, 134:213 223)
  - Should not blindly substitute **G** for **A** in genomic analyses.

Use of Observed Similarity

- Convenient
- Computational efficiency
- Control genomic inbreeding

Alterative Measures of Genomic Similarity

Let $\mathbf{X}$ denote the matrix of centered genotype covariates:

- VanRaden (2008):

$$\mathbf{G} \propto \mathbf{X}\mathbf{X}'$$

  All loci contribute equally.

- Zhang et al. (2010):

$$\mathbf{G} \propto \mathbf{X}\mathbf{D}\mathbf{X}'$$

  $\mathbf{D}$ is diagonal matrix, where d_{ii} is an estimate of the genetic variance for locus i

- Wang et al. (2012) Iterative version of Zhang et al. (2010)

Bayesian Inference of Genomic Similarity

Consider genomic model:

$$\mathbf{a} = \mathbf{X}\boldsymbol{\alpha},$$

where

$$\boldsymbol{\alpha}|\mathbf{D} \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$$

and

$$\begin{aligned} \mathrm{Var}(\mathbf{a}|\mathbf{X}, \mathbf{D}) &= \mathbf{X}\mathbf{D}\mathbf{X}' \\ &= \mathbf{G}\sigma_a^2, \end{aligned} \tag{1}$$

where $\mathbf{D}$ may not be observable, $\mathbf{G} = \frac{1}{\sigma_a^2}\mathbf{X}\mathbf{D}\mathbf{X}'$, and σ_a^2 is the genetic variance.

Bayesian Alphabet Models

- In RRBLUP (BayesC0), $\mathbf{D}$ is observable: $\mathbf{D} = \mathbf{I}\sigma^2_\alpha$, and $\sigma^2_\alpha = \frac{\sigma^2_a}{\sum_i 2p_i(1-p_i)}$

Bayesian Alphabet Models

- In RRBLUP (BayesC0), $\mathbf{D}$ is observable: $\mathbf{D} = \mathbf{I}\sigma^2_\alpha$, and $\sigma^2_\alpha = \frac{\sigma^2_a}{\sum_i 2p_i(1-p_i)}$
- In BayesA, the diagonals of $\mathbf{D}$ are unobserved, and independent $\chi^{-2}(S^2_A, \nu_A)$ priors are used for inference.

Bayesian Alphabet Models

- In RRBLUP (BayesC0), $\mathbf{D}$ is observable: $\mathbf{D} = \mathbf{I}\sigma^2_\alpha$, and $\sigma^2_\alpha = \frac{\sigma^2_a}{\sum_i 2p_i(1-p_i)}$
- In BayesA, the diagonals of $\mathbf{D}$ are unobserved, and independent $\chi^{-2}(S^2_A, \nu_A)$ priors are used for inference.
- In BayesC, a priori a diagonal is zero with probability π, and all non-null values are assigned a $\chi^{-2}(S^2_C, \nu_C)$.

Bayesian Alphabet Models

- In RRBLUP (BayesC0), $\mathbf{D}$ is observable: $\mathbf{D} = \mathbf{I}\sigma^2_\alpha$, and $\sigma^2_\alpha = \frac{\sigma^2_a}{\sum_i 2p_i(1-p_i)}$
- In BayesA, the diagonals of $\mathbf{D}$ are unobserved, and independent $\chi^{-2}(S^2_A, \nu_A)$ priors are used for inference.
- In BayesC, a priori a diagonal is zero with probability π, and all non-null values are assigned a $\chi^{-2}(S^2_C, \nu_C)$.
- In BayesB, a priori a diagonal is zero with probability π, and non-null values are assigned independent $\chi^{-2}(S^2_B, \nu_B)$ priors.

Inference on G from MCMC samples

Let $\mathbf{D}_i$ denote sample i (normalized) from the MCMC procedure.

Inference on G from MCMC samples

Let $\mathbf{D}_i$ denote sample i (normalized) from the MCMC procedure.

Then,

$$\mathbf{G}_i = \mathbf{X}\mathbf{D}_i\mathbf{X}'$$

are MCMC samples of $\mathbf{G}$ that can be used for inference of genomic similarities that are specific to the trait being analyzed.

Inference on G from MCMC samples

Let $\mathbf{D}_i$ denote sample i (normalized) from the MCMC procedure.

Then,

$$\mathbf{G}_i = \mathbf{X}\mathbf{D}_i\mathbf{X}'$$

are MCMC samples of $\mathbf{G}$ that can be used for inference of genomic similarities that are specific to the trait being analyzed.

For example, the posterior mean of $\mathbf{G}$ could be estimated as:

$$\hat{\mathbf{G}} = \mathbf{X}\hat{\mathbf{D}}\mathbf{X}',$$

where $\hat{\mathbf{D}}$ is the posterior mean of $\mathbf{D}$ estimated from the MCMC samples.

Simulation

- 10 chromosomes of length 1 Morgan and 2,000 SNPs
- Random mating in a population of size 100 for 100 generations
- Population expanded to 500, 2,000 or 4,000 for training
- 100 loci randomly chosen to be QTL
- QTL effects were sampled from a standard Normal distribution
- Residual variance was chosen to get a heritability of 0.5

Frobenius Distance to True Genomic Similarity

The "true" genomic similarity matrix was defined as

$$\mathbf{G}_Q = \frac{\mathbf{Q}_c \mathbf{Q}_c'}{100},$$

Frobenius Distance to True Genomic Similarity

The "true" genomic similarity matrix was defined as

$$\mathbf{G}_Q = \frac{\mathbf{Q}_c \mathbf{Q}_c'}{100},$$

Samples of $\mathbf{G}$ were obtained as:

$$\mathbf{G}_i = \mathbf{X}\mathbf{D}_i\mathbf{X}',$$

where $\mathbf{D}_i$ was drawn from its prior or posterior.

Frobenius Distance to True Genomic Similarity

The "true" genomic similarity matrix was defined as

$$\mathbf{G}_Q = \frac{\mathbf{Q}_c \mathbf{Q}_c'}{100},$$

Samples of $\mathbf{G}$ were obtained as:

$$\mathbf{G}_i = \mathbf{X}\mathbf{D}_i\mathbf{X}',$$

where $\mathbf{D}_i$ was drawn from its prior or posterior.

The Frobenius distance between $\mathbf{G}_Q$ and $\mathbf{G}_i$ was computed as

$$\mathrm{D} = \sqrt{\mathrm{tr}(\mathbf{G}_Q - \mathbf{G}_i)^2},$$

Distributions of Frobenius Distance

Results are presented for:

- BayesCπ, where π is treated as unknown with a Uniform prior
- BayesA

BayesCπ with $n = 500$

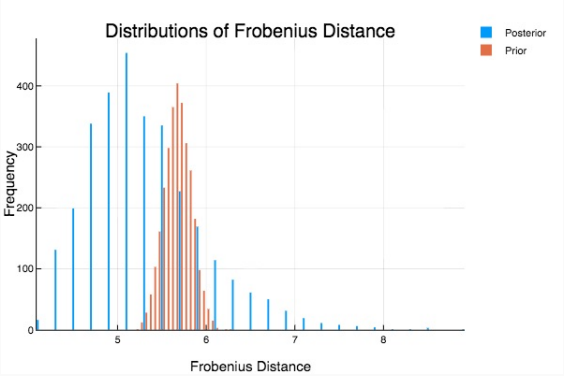

Figure 1: *Distributions of the Frobenius distance to $\mathbf{G}_Q$ from $\mathbf{G}_{D_i}$ (posterior) and $\mathbf{G}_{D_i^*}$ (prior) when training data size is 500. The mean and variance are: 5.3 and 0.42 for the posterior, and 5.7 and 0.02 for the prior.*

BayesCπ with $n = 2000$

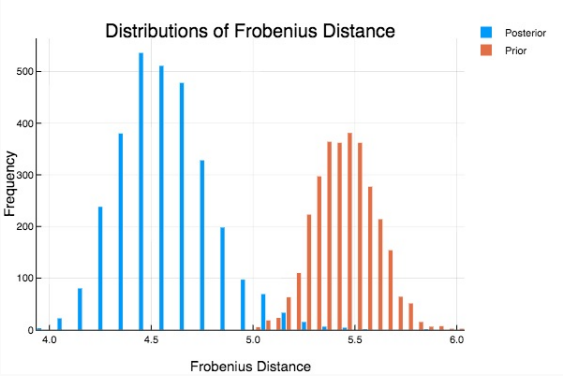

Figure 2: *Distributions of the Frobenius distance to $\mathbf{G}_Q$ from $\mathbf{G}_{D_i}$ (posterior) and $\mathbf{G}_{D_i^*}$ (prior) when training data size is 2000. The mean and variance are: 4.6 and 0.05 for the posterior, and 5.5 and 0.02 for the prior.*

BayesCπ with $n = 4000$

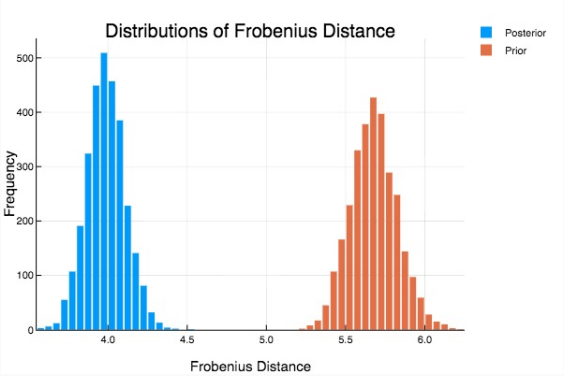

Figure 3: *Distributions of the Frobenius distance to $\mathbf{G}_Q$ from $\mathbf{G}_{D_i}$ (posterior) and $\mathbf{G}_{D_i^*}$ (prior) when training data size is 4000. The mean and variance are: 3.9 and 0.01 for the posterior, and 5.7 and 0.02 for the prior.*

BayesA with $n = 4000$

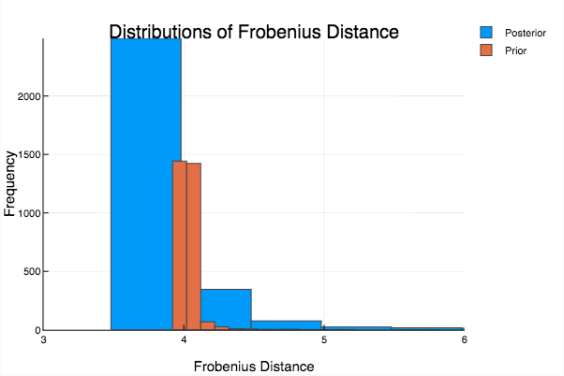

Figure 4: *Distributions of the Frobenius distance to $\mathbf{G}_Q$ from $\mathbf{G}_{D_i}$ (posterior) and $\mathbf{G}_{D_i^*}$ (prior) when training data size is 4000. The mean and variance are: 4.0 and 0.37 for the posterior, and 4.0 and 0.01 for the prior.*

Distance from Prior and Posterior Means of $\mathbf{G}_{\mathbf{D}_i}$ to $\mathbf{G}_{\mathbf{Q}}$

Mean	Frobenius Distance to $\mathbf{G}_{\mathbf{Q}}$
BayesC0: Prior/Posterior ($\mathbf{D} = \mathbf{I}\sigma^2_\alpha$)	3.96
BayesA: Prior	3.96
BayesCπ: Prior	3.96
BayesA: Posterior	3.70
BayesCπ: Posterior	2.98

Table 1: Posterior means were estimated from 3000 (thinning:20) MCMC samples with $n = 4000$

Acknowledgements

RLF is grateful for useful discussions with:

Hao Cheng

Jack Dekkers