Statistical methods in quantitative genetics

Genetic group models and marker-based regression

Stefanie Muff, CBD, Department of Mathematical Sciences, NTNU

CAS Oslo, March 17 2020

Overview

- Genetic group models
  - Differences in means
  - Differences in additive genetic variance (VA)
- Marker-based regression using genomic data

How I became an ecological statistician

- Master in mathematics
- PhD in computational structural biology
- Certificate of Advanced Studies in applied statistics
- Postdoc in medical and ecological statistics

Since Sept. 2019: Associate Professor in Statistics, CBD, Department of Mathematical Sciences, NTNU Trondheim

I'm often not sure if I am

- a mathematician?
- a biostatistician?
- an ecological statistician?
- an applied statistician?

Research interests

- *Measurement error modeling* (methods & applications), see *e.g.* Muff, Riebler, et al. (2015), Muff and Keller (2015), Muff, Puhan, and Held (2018).
- *Bayesian statistics* (ideal for taming measurement errors!).
- *Quantitative genetics*, see Ponzi et al. (2018), Ponzi, Keller, and Muff (2019), Muff, Niskanen, et al. (2019).
- *Movement ecology*, see Weinberger et al. (2016) Gehr et al. (2017) or Muff, Signer, and Fieberg (2019).

Part I: Genetic group animal models

The basic animal model

- Given phenotypic measurements y_i for individuals $1 \leq i \leq n$, the most simple form of the animal model is

  $$y_i = \mu + a_i + e_i \ ,$$

  where $e_i \sim \mathrm{N}(0, \sigma_E^2)$ and $\mathbf{a}^\top = (a_1, \ldots, a_n)^\top \sim \mathrm{N}(\mathbf{a}, \sigma_A^2 \mathbf{A})$ with additive genetic variance σ_A^2 and additive genetic relatedness matrix $\mathbf{A}$.

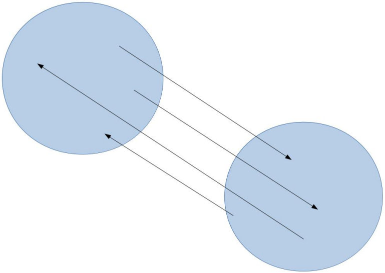
- The model can be extended by additional fixed or random effects.

- **Assumptions**:
  - All individuals derive from the same genetic population.
  - The breeding values (a_i) encode for the deviation from the mean of this population and thus have an expected value $\mathrm{E}[a_i] = 0$.

Systematic deviation from the assumptions

For example

- in cross-bred livestock.
- when genetically different wild subpopulations mix (migration).

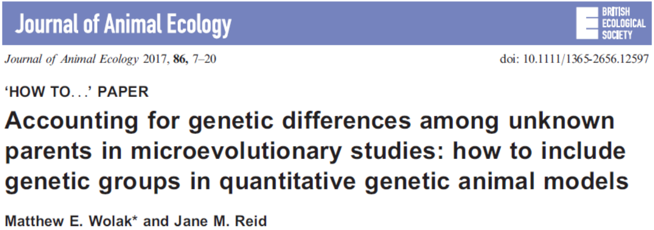


⇒ Individuals have a genetic origin that stems *partially from both populations*.

Consequence: Biased estimates of σ_A^2 and breeding values.

Genetic group models

Great overview by Wolak and Reid (2017)



Journal of Animal Ecology

BRITISH ECOLOGICAL SOCIETY

Journal of Animal Ecology 2017, **86,** 7–20 doi: 10.1111/1365-2656.12597

'HOW TO...' PAPER

Accounting for genetic differences among unknown parents in microevolutionary studies: how to include genetic groups in quantitative genetic animal models

Matthew E. Wolak* and Jane M. Reid

Idea: Allow for "founder populations" that differ in the *mean breeding value*.

Example for two isolated groups

Simple model for a phenotypic trait y_i with mean μ, breeding values a_i and environmental component e_i:

$$\begin{aligned}
\text{group 1:} \qquad & y_i = \mu + \underbrace{a_i}_{u_i} + e_i \ , \\
\text{group 2:} \qquad & y_i = \mu + \underbrace{g_2 + a_i}_{u_i} + e_i \ ,
\end{aligned}$$

thus

$$\begin{aligned}
u_i &\sim \mathrm{N}(0, \sigma_A^2 \mathbf{A}) \quad \text{in group 1.} \\
u_i &\sim \mathrm{N}(g_2, \sigma_A^2 \mathbf{A}) \quad \text{in group 2.}
\end{aligned}$$

- **Interpretation:** The total additive genetic effects u_i (thus probably also the mean phenotypic values) differ between the two groups.
- **Example:** Systematic differences in wing length, weight,...

General formulation of genetic group model with r groups

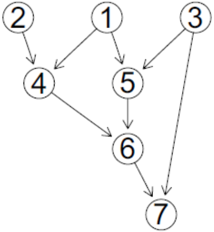
$$y_i = \mu + \underbrace{\sum_{j=1}^{r} q_{ij} g_j + a_i}_{u_i} + e_i \ , \quad \mathbf{a}^\top \sim \mathrm{N}(\mathbf{0}, \sigma_A^2 \mathbf{A}) \ ,$$

where $0 \leq q_{ij} \leq 1$ is the proportional *contribution of group j to the genome of individual i*.

How do we obtain the q_{ij}?

Example

- Group 1: Founders nodes 1 and 2
- Group 2: Founder node 3



$$\mathbf{Q} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0.5 & 0.5 \\ 0.75 & 0.25 \\ 0.375 & 0.625 \end{pmatrix} = \begin{pmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{pmatrix}$$

How to? E.g. by using the `ggcontrib()` function from the `nadive` R package (Wolak 2012).

Genetic group models for heterogeneous variances

- **Caveat:** Additive genetic variances σ_A^2 are assumed the same within the groups.

- **Idea**: Replace

$$y_i = \mu + \sum_{j=1}^{r} q_{ij} g_j + a_i + e_i \qquad \text{(homogeneous } \sigma_A^2\text{)}$$

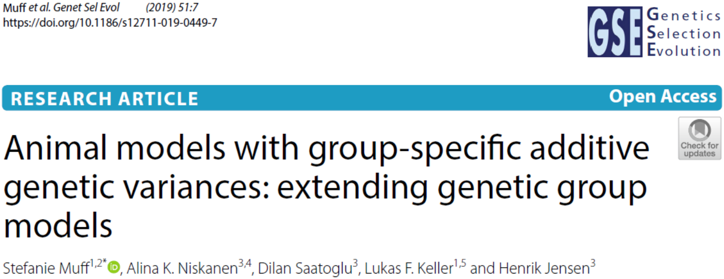
by

$$y_i = \mu + \sum_{j=1}^{r} q_{ij} g_j + \sum_{j=1}^{r} a_{ij} + e_i \ , \qquad \text{(heterogeneous } \sigma_{A_j}^2\text{)}$$

where $(a_{1j}, \ldots, a_{nj})^\top \sim \mathrm{N}(0, \sigma_{A_j}^2 \mathbf{A}_j)$.

- **Two challenges**:
  1) Segregation variances between the groups.
  2) Finding the group-specific relatedness matrics $\mathbf{A}_j$.

Addressing these challenges was the purpose of this publication:



Muff *et al. Genet Sel Evol* (2019) 51:7
https://doi.org/10.1186/s12711-019-0449-7

GSE Genetics Selection Evolution

RESEARCH ARTICLE **Open Access**

Check for updates

Animal models with group-specific additive genetic variances: extending genetic group models

Stefanie Muff[1,2*], Alina K. Niskanen[3,4], Dilan Saatoglu[3], Lukas F. Keller[1,5] and Henrik Jensen[3]

Challenge 1: Segregation variances

In principle, we would need to "blow up" our models with segregation variances (e.g., García-Cortés and Toro 2006)[1]. In our notation:

$$y_i = \mu + \sum_{j=1}^{r} q_{ij} g_j + \sum_{j=1}^{r} a_{ij} + \sum_{j<k} s_i^{(jk)} + e_i \ ,$$

$$\mathbf{s}^{(12)} \sim \mathrm{N}(\mathbf{0}, \sigma^2_{s_{12}} \mathbf{A}_{12}) \ ,$$

thus we would need to estimate $r + \binom{r}{2}$ variances.

But is this relevant here?

[1] Segregation variance refers to the increase in variance caused by differences in allele combinations, average allelic effects, and linkage disequilibrium at and between loci underlying the phenotype in the mixing breeds

- The segregation variance when crossing two genetic groups (e.g., breeds) can be computed as

$$\sigma_S^2 = \frac{1}{2}\sum_{i=1}^{m}(\alpha_i^c)^2 \ ,$$

  where α_i^c denotes the mean additive genetic difference between the groups due to locus i (Lynch and Walsh, 1998), and m is the number of loci that determine the trait.

- Why can we often safely ignore σ_S^2?

- The segregation variance when crossing two genetic groups (e.g., breeds) can be computed as

$$\sigma_S^2 = \frac{1}{2} \sum_{i=1}^{m} (\alpha_i^c)^2 \ ,$$

  where α_i^c denotes the mean additive genetic difference between the groups due to locus i (Lynch and Walsh, 1998), and m is the number of loci that determine the trait.

- Why can we often safely ignore σ_S^2?

Under the *infinitesimal model* assumption, all α_i^c are very small, and thus $\sigma_S^2 \approx 0$.

Challenge 2: Finding group-specific relatedness matrices $\mathbf{A}_j$

- First we recall: $\mathbf{A}$ can be decomposed by a *generalized Cholesky decomposition* into

  $$\mathbf{A} = \mathbf{TDT}' \ ,$$

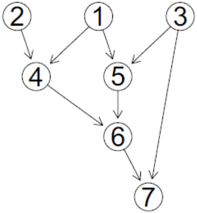
  with lower triangular matrix $\mathbf{T}$ and transposed $\mathbf{T}'$, and diagonal matrix $\mathbf{D} = \text{Diag}(d_{11}, \ldots, d_{nn})$ (Henderson 1976).
- $\mathbf{T}$ traces the flow of alleles from one generation to the other.
- The diagonal entries of $\mathbf{D}$ scale the Mendelian sampling variance.

Example

a

ID	Dam	Sire
1 (g_1)	NA	NA
2 (g_1)	NA	NA
3 (g_2)	NA	NA
4	1	2
5	1	3
6	5	4
7	6	3

b



c

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0.5 & 0.5 & 0.5 & 0.25 \\ 0 & 1 & 0 & 0.5 & 0 & 0.25 & 0.125 \\ 0 & 0 & 1 & 0 & 0.5 & 0.25 & 0.625 \\ 0.5 & 0.5 & 0 & 1 & 0.25 & 0.625 & 0.3125 \\ 0.5 & 0 & 0.5 & 0.25 & 1 & 0.625 & 0.5625 \\ 0.5 & 0.25 & 0.25 & 0.625 & 0.625 & 1.125 & 0.6875 \\ 0.25 & 0.125 & 0.625 & 0.3125 & 0.5625 & 0.6875 & 1.125 \end{pmatrix}$$

d

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 1 & 0 & 0 \\ 0.5 & 0.25 & 0.25 & 0.5 & 0.5 & 1 & 0 \\ 0.25 & 0.125 & 0.625 & 0.25 & 0.25 & 0.5 & 1 \end{pmatrix}$$

e

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.46875 \end{pmatrix}$$

Basic idea

Find $\mathbf{A}_j$ by deriving group-specific versions $\mathbf{T}_j$ and $\mathbf{D}_j$ and then use that

$$\mathbf{A}_j = \mathbf{T}_j\mathbf{D}_j\mathbf{T}_j' \ .$$

Finding $\mathbf{T}_j$

- T_j for group j represents the transmission of alleles through the generations *within each group*.

- This can be obtained when we *scale each row* of $\mathbf{T}$ by $\mathbf{q}_j$ or, equivalently, by

  $$\mathbf{T}_j = \mathbf{T} \cdot \mathrm{Diag}(\mathbf{q}_j) \ ,$$

  for diagonal matrix $\mathrm{Diag}(\mathbf{q}_j)$.

Toy-example:

Given

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 0.5 & 1 \end{pmatrix}$$

and a vector of group-membership proportions for group 1

$$\mathbf{q}_1 = \begin{pmatrix} 1 \\ 0.5 \\ 0 \end{pmatrix} \ ,$$

we obtain

$$\mathbf{T}_1 = \mathbf{T} \cdot \mathrm{Diag}(\mathbf{q}_1) = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0.25 & 0.25 & 0 \end{pmatrix}$$

Finding $\mathbf{D}_j$

- **Main finding**:

Given the entries d_{ii} in the diagonal matrix $\mathbf{D}$, we get

$$d_{ii}^{(j)} = 1 - q_{ij}(1 - d_{ii}) \ .$$

- Why? The original entries are

$$d_{ii} = \begin{cases} 1 \ , & 0 \text{ parent known,} \\ 1 - 0.25 - 0.25(F_p) \ , & 1 \text{ parent known,} \\ 1 - 0.5 - 0.25(F_s + F_d) \ , & 2 \text{ parents known.} \end{cases}$$

where F_p, F_s and F_d are the pedigree-based inbreeding coefficients of the known parent(s).

- The group-specific versions $d_{ii}^{(j)}$ are then obtained as

$$d_{ii}^{(j)} = \begin{cases} 1 \ , & 0 \text{ parent known,} \\ 1 - 0.25 \cdot q_{ij}^{(p)}(1 + F_p) \ , & 1 \text{ parent known,} \\ 1 - 0.5 \cdot q_{ij}(1 + \frac{F_s + F_d}{2}) \ , & 2 \text{ parents known.} \end{cases}$$

- Plus some algebraic rearrangements.

Caveat: $\mathbf{D}_j$ is an approximation

- This idea to find $\mathbf{D}_j$ is an *approximation*.
- Why? We assumed that parental inbreeding can be scaled by the genetic group proportions q_{ij}.
- The correct way would be to use the actual *partial (i.e. group-specific) parental inbreeding coefficients* $F_s^{(j)}$ and $F_d^{(j)}$, which capture inbreeding emerging *within group* j.
- These can be obtained with some extra work, but we showed that approximations are *typically not critical*.

Computations

- Deriving $\mathbf{T}_j$ and $\mathbf{D}_j$ from $\mathbf{T}$ and $\mathbf{D}$ is cheap (simple algebraic transformations).

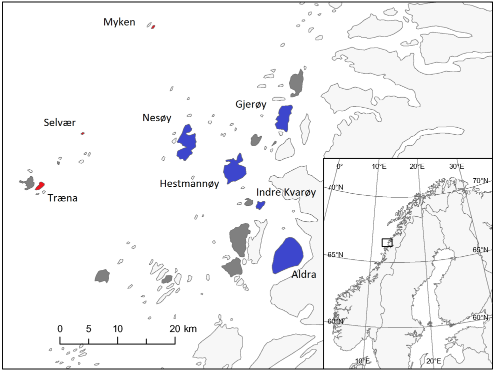
- The resulting $\mathbf{A}_j$ are *singular*. Replace 0's on the diagonal with a small value like 10^{-6}.

- All operations can also directly be carried out using $\mathbf{A}^{-1}$.

- Models can be fitted with ingegrated nested Laplace approximations (INLA, Rue, Martino, and Chopin 2009) or MCMC (using e.g., MCMCglmm, Hadfield 2010).

Example: House sparrow system

- House sparrow metapoplation at the Helgeland coast.
- Study running since 1993.
- Island-system can be broken up into three groups: *inner*, *outer* and *other* islands.



Part II: Marker-based regression

Acknowledgements

- The Centre for Biodiversity Dynamics (CBD) and the research council of Norway (projects 221956 and 223257)
- All my collaborators, in particular Henrik Jensen and Lukas F. Keller
- All the people that are collecting fantastic field data.

References

García-Cortés, Luis Alberto, and Miguel Ángel Toro. 2006. “Multibreed Analysis by Splitting the Breeding Values.” *Genetics Selection Evolution* 38: 601–15.

Gehr, B., E. Hofer, S. Muff, A. Ryser, E. Vimercati, K. Vogt, and L. F. Keller. 2017. “Spatial Scale and Behavioral State Interact in Shaping Temporal Dynamics of Habitat Selection in Eurasian Lynx.” *Oikos* 126: 1389–99.

Hadfield, J.D. 2010. “MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package.” *Journal of Statistical Software* 33: 1–22.

Henderson, C.R. 1976. “Simple Method for Computing Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values.” *Biometrics* 32: 69–83.

Muff, S., and L. F. Keller. 2015. “Reverse Attenuation in Interaction Terms Due to Covariate Error.” *Biometrical Journal* 57: 1068–83.

Muff, S., A.K. Niskanen, D. Saatoglu, L.F. Keller, and H. Jensen. 2019. “Animal Models with Group-Specific Additive Genetic Variances: Extending Genetic Group Models.” *Genetics, Selection, Evolution* 51:7.

Muff, S., M. A. Puhan, and L. Held. 2018. “Bias Away from the Null Due to Miscounted Outcomes?” *Statistical Methods in Medical Research* 27: 3151–66.

Muff, S., A. Riebler, L. Held, H. Rue, and P. Saner. 2015. “Bayesian Analysis of Measurement Error Models Using Integrated Nested Laplace Approximations.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 64: 231–52.

Muff, S., J. Signer, and J. Fieberg. 2019. “Accounting for Individual-Specific Variation in Habitat-Selection Studies: Efficient Estimation of Mixed-Effects Models Using Bayesian or Frequentist Computation.” *Journal of Animal Ecology* 89: 80–92.

Ponzi, E., L. F. Keller, T. Bonnet, and S. Muff. 2018. “Heritability, Selection, and the Response to Selection in the Presence of Phenotypic Measurement Error: Effects, Cures, and the Role of Repeated Measurements.” *Evolution* 72: 1992–2004.

Ponzi, E., L. F. Keller, and S. Muff. 2019. “The Simulation Extrapolation Technique Meets Ecology and Evolution: A General and Intuitive Method to Account for Measurement Error.” *Methods in Ecology and Evolution* 10: 1734–48.

Rue, H., S Martino, and N Chopin. 2009. “Approximate Bayesian Inference for Latent Gaussian