Estimation of GxE in animal breeding populations and implications of GxE for breeding programs

Han Mulder

Contents

- Types of environments and size of GxE in animals
- Data structures to estimate GxE
- Dealing with GxE in breeding programs
- Statistical methods to estimate GxE in animal breeding
- Practical
  - Designs/data structures
  - Statistical analysis using ASReml
Learning outcomes

- To design experiments/datasets for estimating genetic correlations between environments
- To understand the effect of G x E on breeding programs
- To use bivariate and random regression models to analyze genotype by environment interaction

Types of environments and GxE found in animals
Environment can have many sights!

- Climate
- Housing system
- Nutrition
- Disease pressure
- Stocking density
Different types of environments

- Mega-environments
  - Different countries, different climate zones

- Macro-environments
  - Different climates within farms
  - Different farm types (organic vs conventional)

- Micro-environments
  - Each animal has a different environment
  - Some animals diseased; others not

Different types of environments

- Types of environments
  - Categorical
    - Farm types
    - Presence or absence of disease
    - → bivariate/multivariate model
  - Continuous
    - Temperature
    - Daylength
    - Rainfall
    - → Reaction norm model
How to quantify size of GxE?

- In animal breeding: aim is genetic improvement of populations by selection
  - GxE causing reranking has biggest impact

- The degree of GxE is judged by the genetic correlations between environments
  - How much is the genetic correlation deviating from 1.0?

How large is G x E in livestock?

- In dairy cattle (many studies)
  - Production: >0.8
  - Fertility/longevity: 0.5-1.0

- In pigs (fewer studies)
  - 0.5-1.0 between environments with stress and without stress (e.g. disease, heat stress)
  - 0.6-1.0 between farms with different health status

- In poultry (very few studies)
  - 0.6-1.0 between nucleus and commercial environments
How large is G x E in aquaculture?

- Extensive review
  - Different species
  - Different traits
  - Different environments: temperature, diet, location, rearing and stocking density

Sae-Lim et al., 2015. Reviews in Aquaculture 7:1-25
Data structures to estimate G x E

- Categorical environments
  - Measure genotype in different environments
    - Ideal design: animal itself or clones
    - Often animals perform in only one environment
  - In animal breeding: no clones, no experiments!
    - Extensive databases with animal phenotypes and pedigree
    - High-density SNP-genotypes
How to estimate G x E?

- Usually pedigree links
  - Use of additive genetic relationships
  - E.g. half-sisters in different environments
  - Grand-offspring in different environments
  - E.g. less related individuals

- Use of genomic relationships

What kind of design is really needed?

- How much does the design affect the standard error on estimated genetic correlation?
  - How many families do we need with offspring in both environments?
    - N
  - How large should families be?
    - Number of offspring per environment: n
  - What is the effect of the heritability?
    - \( h^2 \)
Standard error to estimate genetic correlation: Robertson (1959; Biometrics)

- Other formula
  \[ se(r_g) \approx \sqrt{\frac{[1 + nt(1-r_g^2)]^2 + r_g^2}{(N-1)n^2t^2}} \]
- \( t = \text{intraclass correlation, e.g. } t = 0.25h^2 \) for half-sibs

Bijma and Bastiaansen (2014, GSE)

- \[ se(r_g) \approx \sqrt{\frac{1}{r_{IH,X}^2r_{IH,Y}^2 + (1 + 0.5r_{IH,X}^2 + 0.5r_{IH,Y}^2 - 2r_{IH,X}^2r_{IH,Y})r_g^2 + r_{HI}^4}}{(N-1)}} \]
- \( r_{IH,X}^2 = \text{reliability of EBV in environment } x = \text{accuracy squared} \)
The effect of family size on se(rg)

- Need many more grand-offspring than half-sib offspring/clones
- Clones is most efficient, but not feasible in livestock

Bijma and Bastiaansen, 2014

The effect of number of families on se(rg)

- Need 50-100 families to get accurate estimate of genetic correlation

Bijma and Bastiaansen, 2014
Summary

- Large datasets required to estimate genetic correlations between environments
  - 50-100 families
  - Each with 50-100 offspring

- Clones slightly better than half-sibs, grand-offspring is quite a bit worse than half-sibs

Deal with GxE in breeding programs
Different situations:
1. Nucleus and commercial environment
   - Typically selection environment (SE) and production environment (PE) different
     - SE: higher health status, less diseases, optimal management
     - PE: higher disease pressure, lower management level, in pigs and poultry crossbred animals

G x E: nucleus and production environment
   - Only information from nucleus, but breeding goal is commercial environment
     - Genetic gain in commercial environment is correlated response
G x E: nucleus and production environment

- Use of sib/progeny information from commercial environment

Different situations:
2. Multiple production environments

- Breeding organization are international
- Multiple production environments
  - Different climates
  - Within countries different types of farms
    - Organic and conventional
    - Management level
    - Barn type
    - Disease status
    - Grazing and non-grazing

**G × E: How many lines/breeding programs?**

### Breeding strategies

**One breeding program**
- All 400 bulls tested in both environments: 50 daughters in each environment
- Increase average performance

**Two breeding programs**
- 200 bulls tested in one environment: 100 daughters in one environment
- Each bp: increase performance in environment of testing

---

**Average genetic gain**

- **one bp**
- **two bp**

- **break-even genetic correlation**

Mulder et al., 2006; J. Dairy Sci.
G x E and multi-trait selection

- Between environments
  - G x E per trait
  - Heterogeneity of genetic variances
  - Breeding goal differences
  - Different genetic correlations between traits

Genetic correlation between breeding goals

\[ r_{H,kl} = \frac{v_k^G v_l}{\sqrt{v_k^G v_k'} G v_l' G v_l} \]

- \( G \): full genetic variance-covariance matrix between all traits in the breeding goals of environment \( k \) and \( l \)
- \( v_k \): economic values for environment \( k \)
- \( v_l \): economic values for environment \( l \)

(Mulder, 2007)

Summary

- G x E lowers genetic gain, but more genetic diversity is conserved
- Nucleus and production environment
  - Minimize environmental difference
  - Use phenotypes of sibs or progeny in multivariate breeding value estimation
- Different production environments
  - If \( r_g > 0.6-0.7 \) then single breeding program (provided that information of sibs/progeny is collected in both environments)
  - If \( r_g < 0.6-0.7 \), then different breeding programs needed
Statistical methods to estimate GxE in animal breeding

- We use pedigree relationships and we use BLUP
- Main interest in additive genetic effects or breeding values
- Most common models to analyze G x E
  - Bivariate/multivariate models
  - Reaction norm/random regression models
BLUP

- \( y = \mu + \text{herd} + \text{animal} + e \)
- \( y \) = phenotype
- \( \mu \) = fixed mean
- \( \text{herd} \) = fixed effect for herd
- \( \text{animal} \) = random additive genetic effect = EBV
- \( e \) = residual

BLUP mixed model equations

- \( y = Xb + Za + e \)
- \( X \) = design matrix to link phenotypes to fixed effects, e.g. which cow is in which herd
- \( b \) = vector with solutions for fixed effects
- \( Z \) = design matrix to link phenotypes to EBV
- \( a \) = vector with EBV for all animals

\[
\begin{bmatrix}
X'X & X'Z \\
Z'X & Z'Z + \lambda A^{-1}
\end{bmatrix}
\begin{bmatrix}
b \\
a
\end{bmatrix}
=
\begin{bmatrix}
X'y \\
Z'y
\end{bmatrix}
\]

- \( \lambda = \frac{\sigma^2_e}{\sigma^2_a} \)
- \( A^{-1} \) = inverse of additive genetic relationship matrix
Breeding values Holstein bulls

<table>
<thead>
<tr>
<th>Breeding Value</th>
<th>kg milk</th>
<th>%fat</th>
<th>%protein</th>
<th>kg fat</th>
<th>kg protein</th>
<th>total merit milk</th>
<th>total merit index</th>
<th>NVI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bookem</td>
<td>+1552</td>
<td>-0.29</td>
<td>-0.11</td>
<td>+37</td>
<td>+43</td>
<td>+275</td>
<td>+299</td>
<td></td>
</tr>
<tr>
<td>G-Force</td>
<td>+771</td>
<td>+0.16</td>
<td>+0.13</td>
<td>+48</td>
<td>+39</td>
<td>+271</td>
<td>+246</td>
<td></td>
</tr>
<tr>
<td>Atlantic</td>
<td>+298</td>
<td>-0.06</td>
<td>+0.14</td>
<td>+7</td>
<td>+23</td>
<td>+110</td>
<td>+245</td>
<td></td>
</tr>
<tr>
<td>Titanium</td>
<td>+645</td>
<td>+0.21</td>
<td>+0.01</td>
<td>+47</td>
<td>+23</td>
<td>+199</td>
<td>+234</td>
<td></td>
</tr>
<tr>
<td>Snowman</td>
<td>+2576</td>
<td>-0.37</td>
<td>-0.29</td>
<td>+69</td>
<td>+57</td>
<td>+415</td>
<td>+228</td>
<td></td>
</tr>
</tbody>
</table>

Bivariate model to estimate GxE

\[ \begin{align*}
\mathbf{y}_1 &= \mathbf{X}_1 \mathbf{b}_1 + \mathbf{Z}_1 \mathbf{a}_1 + \mathbf{e}_1 \\
\mathbf{y}_2 &= \mathbf{X}_2 \mathbf{b}_2 + \mathbf{Z}_2 \mathbf{a}_2 + \mathbf{e}_2 \\
\mathbf{a}_1 &\sim N(0, \mathbf{A} \otimes \begin{bmatrix} \sigma_{a1}^2 & \sigma_{a1a2} \\ \sigma_{a1a2} & \sigma_{a2}^2 \end{bmatrix}) \\
\mathbf{e}_1 &\sim N(0, \mathbf{I}_1 \sigma_{e1}^2) \\
\mathbf{e}_2 &\sim N(0, \mathbf{I}_2 \sigma_{e2}^2)
\end{align*} \]

- no residual covariances between environments if animals are in one environment

\[ \mathbf{r}_g = \mathbf{r}_a = \frac{\sigma_{a1a2}}{\sigma_{a1}^2, \sigma_{a2}^2} \]
Reaction norm models

- \( y = \text{fixed effects} + bx + a_{int} + a_{sl}x + e \)

- Fixed reaction norm: \( bx \)

- \( [a_{int}, a_{sl}] \sim N \left( 0, A \otimes \begin{bmatrix} \sigma^2_{aint} & \sigma_{aint,asl}^2 \\ \text{symmetric} & \sigma^2_{asl} \end{bmatrix} \right) \)

- \( A = \) matrix with all additive genetic relationships

- Heterogeneity of residual variance accounted for using 3-10 groups each with their own residual variance

---

Issues with reaction norms models

1. Which covariate to use?
   - 1. External environmental factor
   - 2. Internal data derived parameter

2. Scaling and (Legendre) polynomials

3. Model comparison

4. Interpretation of results
1. Which covariate to use?

External and internal environmental factors

- **External environmental factors**
  - Temperature
  - Day length
  - Rainfall
  - Salinity, oxygen (fish)
  - ...

- **Internal derived environmental factors**
  - Finlay-Wilkinson regression
  - Mean performance
  - Herd-year-season estimated effect
Temperature

- Temperature may affect phenotypes only above a certain temperature, the so-called upper critical temperature

Upper critical temperature: Farrowing rate at first insemination

Upper critical temperature: litter size

Bloemhof et al., 2008. J. Anim. Sci. 86:3330-3337

Day length

- Day length is according to a sinus function

Internal data derived parameter: The use of herd-year-season estimates

- In many studies, it is difficult to categorize farms
  - No access to external data
  - HYS gives indication of management level, but may also contain the genetic level of the herd

- Strategy
  - Estimated herd-year-season effects on the same data using mixed model
  - Add to the data set and use random regression
  - Use of data twice = tricky

Possible solutions

- Derive environmental parameters (EP) from other traits
  - Calus et al. (2003; J. Dairy Sci. 86: 3756-3764)

- Use other animals to calculate the EP

- Use many animals to estimate EP, dependency is smaller
  - Avoid very small HYS classes
  - Include all parities
Consequences for estimation of G x E

- Reaction norm models tend to underestimate G x E
  - Underestimation of the genetic variance in slope
  - Correlations closer to 1.0 than the true value

Possible solution

- Bayesian approach
  - The x-variable is simultaneously sampled with the breeding values and the other effects in the model

Table 1. Mean and SE of estimates (based on posterior means) of (co)variance components over 20 replicate simulations

| Model  | $\sigma^2_\ell$   | $\sigma^2_{a_s}$ | $\sigma_{a_s|a_l}$ | $\sigma^2_\epsilon$ |
|--------|------------------|------------------|--------------------|-------------------|
| Realized | 100.4 ± 0.040 | 1.01 ± 0.002 | 5.11 ± 0.065 | 298.3 ± 0.016 |
| M1²    | 101.7 ± 1.102 | 1.02 ± 0.034 | 5.04 ± 0.101 | 297.1 ± 0.872 |
| M2⁴    | 99.3 ± 1.051 | 1.01 ± 0.013 | 5.00 ± 0.080 | 298.5 ± 0.868 |
| M3⁵    | 111.5 ± 1.440 | 0.58 ± 0.020 | 3.68 ± 0.105 | 305.5 ± 0.702 |

1$\sigma^2_\ell$ = variance of the level; $\sigma^2_{a_s}$ = variance of the slope of additive genetic reaction norm; $\sigma_{a_s|a_l}$ = covariance between the level and the slope; and $\sigma^2_\epsilon$ = residual variance.
2The variance components were calculated from the realized values of the simulation.
3Model with unknown covariate of reaction norm (the proposed approach).
4Model using true herd-year effect as covariate of reaction norm.
5Model using phenotypic mean of herd-year as covariate of reaction norm.

Su et al., J. Anim. Sci. 84:1651-1657
2. Scaling and (Legendre) polynomials

Do we need polynomials?

- Linear reaction norm
  - No need for use of polynomials
  - Would give equivalent results

- Higher order reactions norms
  - Yes, performance of REML or Gibbs much better
Linear reaction norm models without polynomials

- Scaling of covariate mean = 0, variance 1.0
  - The correlation between intercept and slope has a meaning when selection is performed in the average environment
  - Variance of 1.0 makes it feasible to compare estimates of genetic variance in slope when using different covariates

Higher reaction norms: Legendre polynomials

- They are orthogonal
  - Lower correlations between regression coefficients – faster convergence

- Scale the EP to be between -1 and 1

\[ x_l = -1 + 2 \times \left( \frac{EP_l - EP_{min}}{EP_{max} - EP_{min}} \right) \]

(Schaeffer, Random regression models:
http://www.aps.uoguelph.ca/~lrs/ABModels/NOTES/RRM14a.pdf)
Legendre polynomials

- Legendre polynomial coefficient order $n>1$, recursive equation:
  - $P_0 = 1$
  - $P_1 = x$
  - $P_{n+1}(x) = \frac{1}{n+1} ((2n + 1)xP_n(x) - n(P_{n-1}(x))$
  - $\phi_n(x) = \left(\frac{2n+1}{2}\right)^{0.5} P_n(x)$

(Schaeffer, Random regression models: http://www.aps.uoguelph.ca/~lrs/ABModels/NOTES/RRM14a.pdf)

Example polynomial coefficients

<table>
<thead>
<tr>
<th>x</th>
<th>x scaled</th>
<th>P0</th>
<th>P1</th>
<th>P2</th>
<th>$\phi_0$</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.33</td>
<td>1.00</td>
<td>0.33</td>
<td>-0.33</td>
<td>0.71</td>
<td>0.41</td>
<td>-0.53</td>
</tr>
<tr>
<td>200</td>
<td>0.67</td>
<td>1.00</td>
<td>0.67</td>
<td>0.17</td>
<td>0.71</td>
<td>0.82</td>
<td>0.26</td>
</tr>
<tr>
<td>300</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.71</td>
<td>1.22</td>
<td>1.58</td>
</tr>
<tr>
<td>-100</td>
<td>-0.33</td>
<td>1.00</td>
<td>-0.33</td>
<td>-0.33</td>
<td>0.71</td>
<td>-0.41</td>
<td>-0.53</td>
</tr>
<tr>
<td>-200</td>
<td>-0.67</td>
<td>1.00</td>
<td>-0.67</td>
<td>0.17</td>
<td>0.71</td>
<td>-0.82</td>
<td>0.26</td>
</tr>
<tr>
<td>-300</td>
<td>-1.00</td>
<td>1.00</td>
<td>-1.00</td>
<td>1.00</td>
<td>0.71</td>
<td>-1.22</td>
<td>1.58</td>
</tr>
<tr>
<td>0</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>-0.50</td>
<td>0.71</td>
<td>0.00</td>
<td>-0.79</td>
</tr>
</tbody>
</table>
3. Model comparison

Significance of model

- Likelihood ratio test
- \( H_0: \) model with only intercept
- \( H_1: \) model with intercept and slope

- The likelihood ratio:
  - \[ D = 2\log L(\text{full model}) - 2\log L(\text{reduced model}) \]
- If the hypothesis contains a parameter on the boundary, then \( D \) follows a mixture of Chi-square distributions
**Which degrees of freedom?**

- Suppose the model under H0 estimates:
  
  \[ G = \begin{bmatrix} \sigma_{aint}^2 & 0 \\ 0 & 0 \end{bmatrix} \]

- The model under H1:
  
  \[ G = \begin{bmatrix} \sigma_{aint}^2 & \sigma_{aint,asl} \\ \sigma_{symmetric} & \sigma_{asl}^2 \end{bmatrix} \]

- The large sample distribution is:
  - Mixture of \( \chi_1^2 \) and \( \chi_2^2 \)

Visscher, 2006; Twin research and human studies 9: 490-495
Stram and Lee, 1994; Biometrics 50:1171-1177

---

**In more general terms**

- Model H0:
  
  \[ G = \begin{bmatrix} D_0 & 0 \\ 0 & 0 \end{bmatrix} \]
  
  \( D_0 \) is a matrix with dimension \( q \times q \)

- Model H1:
  
  \[ G = D_1 = \begin{bmatrix} D_0 & d12 \\ d21 & d22 \end{bmatrix} \]
  
  \( D_1 \) is a matrix with dimension \( (q+1) \times (q+1) \)

- The large sample distribution is:
  - Mixture of \( \chi_q^2 \) and \( \chi_{q+1}^2 \)

Stram and Lee, 1994; Biometrics 50:1171-1177
Other model comparisons

- Akaike’s information criterion:
  \[ AIC = -2 \log L + 2t \]
  \( t \) = number of variance parameters in the model

- Bayesian information criterion (more conservative):
  \[ BIC = -2 \log L + 2t \log(v) \]
  \( v \) = residual degrees of freedom

- AIC/BIC are not tests for significance
- AIC/BIC favour the most parsimonious model

Other model comparisons

- Check the genetic parameters obtained from reaction norm model with a bivariate model

- Reaction norm models may lead to:
  - Extreme heritabilities in extreme environments
  - Low genetic correlation between extreme environments
Other model comparisons

- Predictive ability

- Cross-validation
  - Predict the phenotype or adjusted phenotype in the validation set

Accuracy of genomic and pedigree breeding values

4. Interpretation of results

Calculation of genetic parameters

- Genetic variance-covariance matrix between different environments

\[ H = \Phi G \Phi' \]

- \( G \) = matrix estimated (co)variances for the different orders of the polynomial

- \( \Phi \) = matrix with \( \phi \) values for the orders of the polynomial for the environments of interest.
Genetic parameters using genomic or pedigree relationship matrix (litter size)

Trait: litter size
Environments: Large White sows in 22 countries


Genetic correlations between different environments

Herrero-Medrano et al., 2015; J. Anim. Sci. 93:1494-1502
Running ASREML with reaction norms

- ASREML mean model iteration
- animal !P
- sire
- dam
- herd
- hys
- Am
- Asl
- Av
- x #!-1 # you can shift the intercept if you want
- E
- Pheno

- ped.dat !make
- cows_asreml.dat !MAXIT 100

- Pheno ~ mu !r animal animal.x
  1 1 1
  10000
  animal 2
  2 0 US 0.3 0.0 0.05
  animal

Running ASReml

- Use ASREML-W

- Or with a batch-file
Summary

- Bivariate models and reaction norm models can be used to estimate $G \times E$

- Reaction norm models are more complex
  - Heritabilities and genetic correlations for every set or pair of environments

- Different environmental parameters can be used
  - Be careful when using HYS