Current research in unweighted and weighted ssGBLUP

Ignacy Misztal, Ivan Pocrnic and Daniela Lourenco

University of Georgia
Derivation of single step

\[ u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \]

- Ungenotyped
- Breeding values
- genotyped

\[ u_2 = Ma \]
\[ \text{var}(u_2) = M \text{var}(a) M' = MDM' \sigma_a^2 = G \sigma_u^2 \]
\[ u_1 = A_{12} A_{22}^{-1} u_2 + \varepsilon \]
\[ u_1 = A_{12} A_{22}^{-1} Ma + \varepsilon = M_g a + \varepsilon \]

Relationship matrix for ssGBLUP

\[ \text{var}(u)^{-1} = H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix} \]

Aguilar et al., 2010

SSBR equations

\[ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1^* \\ X_2^* \end{bmatrix} \beta^* + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} \hat{M}_1 \alpha + \varepsilon \\ \hat{M}_2 \alpha \end{bmatrix} + \varepsilon \]

Fernando et al., 2014
Issues in single-step

• Compatibility between pedigree and genomic information
• Reliable computations for millions of genotyped animals and complex models

• Incorporation of causative variants if known
• SNP selection/weighting with small data
Properties of G and A$_{22}$

• G - “infinite” pedigree
  • depends on gene frequencies, arbitrary scaling, genotyping accuracy & errors

• A$_{22}$ - depends on pedigree completeness, depth, errors
  • Typical heterogeneous base population

• Adapt
  • a) G to A$_{22}$?
  • b) A$_{22}$ to G?
  • c) Both?
Scaling $G$

$$G = \frac{(M - 2P)(M - 2P)'}{\sum_i 2p_i q_i}$$

M – gene content, P – gene frequencies

VanRadlen (2008)

1. Use base population gene frequencies (Gengler, 2007; VanRadlen, 2008; Christensen and Lund, 2010)
   - Assumes base scale OK
   - What if base population heterogeneous?

2. Use a constant for phenotypes of genotyped animals
   (Stranden and Christensen, 2010; Fernando et al., 2014)
3. Scale $G$ for compatibility with $A_{22}$ (VanRaden, 2008; Chen et al., 2011; Vitezica et al., 2011)

$$G = \alpha + (1 - \frac{\alpha}{2})G_c,$$

$\alpha$: $\text{avg}(a_{22,ij})=\text{avg}(g_{ij})$
Reliability ($R^2$) and bias ($b1$) for 18 traits in Holstein

Bias if $b1 < 1$

- no INB in A
- INB in A
- INB in A & UPG
- INB in A&UPG + 50%$h^2$

Reduction of $h^2$ -- Wiggans et al., 2012, 2016
Concept of Metafounders

We need to adjust the UPG theory to match $A$ to $G$ instead of viceversa

In other words, we can infer the relationships across breeds from markers

Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation


Ole F. Christensen

Genetic evaluation for three-way crossbreeding

Metafounders are related to $F_{st}$ fixation indices and reduce bias in single-step genomic evaluations

Ole F. Christensen, Andres Legarra, Mogens S. Lund and Guosheng Su

Carolina A. Garcia-Baccino, Andres Legarra, Ole F. Christensen, Ignacy Misztal, Ivan Pocrnic, Zulma G. Vitezica and Rodolfo J. C. Cantet
Crude estimates of $\Gamma$ in dairy sheep breeds (50K chip):

$$
\begin{pmatrix}
0.49 & 0.47 & 0.37 & 0.38 \\
0.47 & 0.51 & 0.37 & 0.38 \\
0.37 & 0.37 & 0.50 & 0.45 \\
0.38 & 0.38 & 0.45 & 0.49 \\
\end{pmatrix}
$$

- FR/ES Pyrenees, common origin, frequent exchanges
- Latxa Cara Rubia
- Manech Tete Rousse
- Lacaune Confederation
- Lacaune Ovitest
- Roquefort area, same breed but diverging since 1976
Validation results for ssGBLUP models

All animals in last generation confounded (genotyped and ungenotyped)

<table>
<thead>
<tr>
<th>Model</th>
<th>UPGa</th>
<th>Accuracyb</th>
<th>Accuracyc</th>
<th>b1, SD</th>
<th>b2, SD</th>
<th>Biasd</th>
<th>SD</th>
<th>Bias</th>
<th>SD</th>
<th>Rounds (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>h² = 0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssGBLUP</td>
<td>Complete</td>
<td>0.77</td>
<td>0.02</td>
<td>1.07</td>
<td>0.03</td>
<td>-0.04</td>
<td>0.03</td>
<td>181</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>ssGBLUP</td>
<td>None</td>
<td>0.76</td>
<td>0.02</td>
<td>1.10</td>
<td>0.04</td>
<td>-0.14</td>
<td>0.05</td>
<td>180</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>ssGBLUP</td>
<td>A</td>
<td>0.76</td>
<td>0.02</td>
<td>1.06</td>
<td>0.04</td>
<td>-0.21</td>
<td>0.05</td>
<td>221</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>ssGBLUP</td>
<td>H</td>
<td>0.73</td>
<td>0.03</td>
<td>1.05</td>
<td>0.05</td>
<td>-0.83</td>
<td>0.10</td>
<td>170</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>ssGBLUP</td>
<td>Meta</td>
<td>0.77</td>
<td>0.02</td>
<td>1.05</td>
<td>0.04</td>
<td>-0.03</td>
<td>0.04</td>
<td>122</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

- Complete: full data was used with no missing pedigrees to model,
- None: did not model missing pedigrees,
- A: UPG formed for \(A^{-1}\),
- H: UPG formed for \(H^{-1}\), and
- Meta: metafounders

\[
\frac{(TBV - EBV)}{\sigma_u}, 0.01 \leq SE \leq 0.04
\]

Bradford et al., 2018
Scaling in multibreed evaluation

• Adaptations of purebred methods
• Dilemma: one breed/trait or all breeds as one trait?

• If one trait:
  • how many SNP per breed?
  • Reduction of accuracy if Holsteins and Jerseys as one trait (VanRaden, 2017)
  • Simulation studies (Stein et al., 2018)
Inversion of large genomic relationship matrix

• Limited dimensionality of genomic information
  • 3 Gb genome
  • 30 Mb SNP
  • Compresses to 4k-20k

• Inverse of GRM does not exist

• Many generalized inverses
From 1949 to 2018: R. A. Fisher's Theory of Junctions

Although this year we celebrate the centenary of Fisher's monumental foundation of quantitative genetics in 2018, I here celebrate a much later and lesser-known contribution: his Theory of Junctions. This theory was first introduced in his monograph, Theory of Inbreeding (1949, Oliver & Boyd, Edinburgh), and was developed further by Bennett (1953, Genetics 26: 392–406) and by Fisher himself (1954, Heredity 8: 187–197 and 1959, Heredity 13: 179–186). Although Fisher described this topic as important, it is only with current genomic data that it has become more broadly appreciated.

Fisher introduced his theory in the context of production under close inbreeding of recombinant inbred lines of mice, all the chromosomes in the line become IBD over their entire length. The remaining junctions internal to IBD regions are then fixed and define the segments of the inbred line that are of different ancestral origins. By considering the total number of junctions, and the formation and loss of internal junctions, Fisher was able to derive the expected number of non-IBD tracts. Together with standard formulae for the proportion of genome remaining non-IBD, this provides the expected lengths of tracts remaining non-IBD.

Interestingly, in more recent analyses of the genomes of recombinant inbred lines (e.g., Broman, 2005, Genetics 169: 1133–1146), authors have followed two-locus analy-

Moreover, on the base-pair scale, per-meiosis mutation and recombination rates are of the same order of magnitude, so that one mutation is expected in every IBD segment. For dense markers or DNA sequence data, adjustments to ROH methods are needed (Browning and Browning 2011, Am. J. Hum. Genet. 88: 173–182).

Although Fisher's theory is often not cited, Fisher's junctions are now pervasive in the Statistical Genetics literature. Ancient junctions bound segments of introgressed Denisovan and Neanderthal DNA in our genomes, cultural, livestock or human populations, consist of segments of DNA from different ancestral origins. These segments dictate the patterns of observable genetic variation in extant populations.

Elizabeth A. Thompson
Department of Statistics, University of Washington, Seattle, WA, USA
Email: eathomp@uw.edu
Estimated effective population size and the number of segments

<table>
<thead>
<tr>
<th>Specie</th>
<th>Effective population size</th>
<th>Me</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holsteins</td>
<td>149</td>
<td>18k</td>
</tr>
<tr>
<td>Jerseys</td>
<td>101</td>
<td>12k</td>
</tr>
<tr>
<td>Angus</td>
<td>113</td>
<td>13k</td>
</tr>
<tr>
<td>Pigs</td>
<td>43</td>
<td>4k</td>
</tr>
<tr>
<td>Chicken</td>
<td>44</td>
<td>4k</td>
</tr>
</tbody>
</table>
Methods to invert genomic relationship matrix

By eigenvalue decomposition

\[ G = ZZ' = UDU' \]
\[ G^{-1} = UD^{-1}U' \]

By **APY algorithm** (Misztal et al., 2014)

\[ G^{-1} = \begin{bmatrix} G_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} G_{cc}^{-1}G_{cn} \\ I \end{bmatrix} M^{-1} \begin{bmatrix} G_{nc}'G_{cc}^{-1} & I \end{bmatrix} \]

By **Woodbury formula** (Mantysaari et al., 2017)

\[ G = ZZ' + I\varepsilon, \]
\[ G^{-1} = \frac{1}{\varepsilon} I - \frac{1}{\varepsilon} Z \left( \frac{1}{\varepsilon} Z'Z + I \right)^{-1} Z' \frac{1}{\varepsilon} \]
Reliabilities assuming different dimensionality with APY inverse – Holsteins

BIF Workshop, Kansas City, Dec 6, 2018

Pocrnic et al., 2016b
Is BLUP biased? – US Holsteins with 760k genotyped animals

• ssGBLUP for 30 million US Holsteins (Masuda et al., 2017)
  • 3 traits: milk, fat & protein
  • 760K genotypes of SNP60K
  • Computing time 18 h (6.5 h by BLUP)

• Trends for sires and cows
Cows: ssGBLUP vs traditional PTA (protein)

Year of Birth

PTA (kg)

Genotyped cows

All cows

ssGBLUP

BLUP

1.1 kg/yr

0.4 kg/yr

1.2 kg/yr

*Cows with record(s)
Inclusion of causative SNP information

• In dairy, results varied
  – No improvement in Dutch (Binsbergen, et al., 2015) or German Holsteins (Erbe et al., 2016)
  – Up to 5% improvement in Nordic/French (Brøndum et al.)
  – Up to 5% improvement in US Holsteins (Vanraden et al. (2017))

• In SNP BLUP, SNP effects regressed towards zero

• To include causative SNP:
  – Need location
  – need variance
ssGBLUP accuracies using SNP60K and 100 QTNs – simulation study

Fragomeni et al. (2017)
Use of causative variants and SNP weighting in a single-step GBLUP context

Fragomeni BO¹, Lourenco DAL¹, Legarra A², Tooker ME ³, VanRaden PM³, Misztal I¹

¹University of Georgia, Athens, USA
²INRA, Castanet-Tolosan, France
³AGIL ARS-USDA, Beltsville, USA
Weighs for GWAS in weighted genomic relationship matrix

\[ G = ZDZ' \frac{\sigma_s^2}{\sigma_a^2} = \frac{ZDZ'}{\sum_i 2p_i q_i} \]

• Linear weights
  • \( \hat{d}_i = \hat{u}_i^2 \)

• Non-linear A weights
  • \( \hat{d}_i = 1.125 \frac{\|\hat{u}_i\|^2}{sd(u)^2} \)
    • Value capped at 10

• Fast-Bayes A
  • \( \hat{d}_i = \frac{SNP_{eff_i}^2 + df*S^2}{df + 1} \)
Accuracy of different methods in GWAS - simulation
US Holstein data

• 4M records for Stature
• 3M Cows
• 4.6M Animals in pedigree
• 27k Genotyped Sires

• 54k SNP
• 54k SNP + 17k Causative Variants (VanRaden et al., 2017)
Including causative variants

Unweighted ssGBLUP most accurate
Improvement with “causative SNP” partly artifact of poor modeling
Can large QTL exist despite selection?

- Genetics and genomics of mortality in US Holsteins
  - (Tokuhisa et al, 2014; Tsuruta et al., 2014)
- 6M records, SNP50k genotypes of 35k bulls
Milk – first parity

Mortality – first parity
P-values for GWAS in ssGBLUP

- P-values with classical GWAS
- BLUP biased so dergressed proofs biased
- Can we obtain p-values for ssGBLUP

\[ pval_i = 2 \left(1 - \Phi \left( \frac{snp_i}{sd(snp_i)} \right) \right) \] (Chen et al., 2017)

- Algorithm
  - Calculate PEV for GEBV
  - Convert GEBV to SNP effects
  - Convert PEV/GBEV to PEV/SNP
Is SNP selection/weighting important??

- SNP selection/weighting (BayesB, etc.)
  - Large impact with few genotypes
  - Little or no impact with many
  - Multitrait models with SNP selection hard to apply

BIF Workshop, Kansas City, Dec 6, 2018

Karaman et al., 2016
Persistence over generations with different sizes of reference populations

Very large – equivalent to 4NeL animals with 99% accuracy
Is genomic selection on chromosome segments or chromosome clusters?

- Simulation
  - 6k animals with 50 k SNP
  - $N_e \approx 50$, $L = 10M$
- GBLUP
  - Use GRM with limited number of eigenvalues (corresponding to 10 to 99% variation)
  - 4k animals in reference population, 2k in validation
Eigenvalue profile of GRM

10% should be 300 segments
Perhaps largest eigenvalue clusters 100 segments
Accuracies of GBLUP using GRM with largest eigenvalues only

4k animals with own records

4k bulls with 100 daughters

Accuracy

% of explained variance

Eigenvalues

30 40 50 60 70 80 90 100

30 40 50 60 70 80 90 100

BIF Workshop, Kansas City, Dec 6, 2018
Varying amount of information

4k bulls with 50 daughters

4k animals with own records

BIF Workshop, Kansas City, Dec 6, 2018
Does APY algorithm for inversion of GRM work on segments or eigenvalues
Selection on largest eigenvalues – important ancestors – reduced Ne

If largest eigenvalues excluded, increased diversity?
Some hypothesis on GWAS

First cluster
Second cluster

QTL

BayesB1
BayesB2
GBLUP
Classical with pop. stratification
Conclusions

• BLUP biased under genomic preselection
  • Single-step only option

• Large redundancy in genomic information – limited dimensionality
  • Computations easy for millions of genotypes

• SNP selection unimportant with large data

• Modeling missing pedigrees and multiple breeds important especially if strong selection
  • Metafounder as generalization of unknown parent groups to genomic models

• Little data required for medium accuracy, lots of data for high accuracy
  • Few SNP sufficient with small data, more SNP with lots of data
Group and sponsors