Possible implications of limited dimensionality of genomic information

Ignacy Misztal, Ivan Pocrnic and Daniela Lourenco
University of Georgia
Dimensionality of genomic information

\[ u = Z\alpha \]

Singular value decomposition
\[ Z = U\Delta V \]
\[ U'U = I, V'V = I, \Delta \]

Genomic relationship matrix
\[ G = U\Delta \Delta U' = UDU' \]
Rank(G) \( \leq \) min(#SNP, #anim)

SNP BLUP design matrix
\[ Z'Z = V'\Delta \Delta V \]
Rank(Z'Z) \( \leq \) min(#SNP, #anim)

Same dimensionality for genotypes, GRM and SNP BLUP

Dimensionality around 5-15k (VanRaden, 2008; Maciotta et al., 2013)
Origin of Haplotype blocks

Cuppen, 2005
Chromosome segments

• Theory of junctions (Fisher, 1949):
  • Heterogenetic and homogenic tracts in genome

• For randomly mating population of constant size the number of tracts:

\[ E(\text{Me}) = 4 \text{ Effective population size (Ne) } \times \text{ Genomic size (L)} \] (Stam, 1980)

• Independent chromosome segments Me (Goddard, 2009; Daetwyler et al., 2010)

• Need 12 Me SNPs to detect 90% of junctions (MacLeod et al., 2005)
Number of junctions/chromosome segments/haplotype blocks

- $\sim 4N_e L$  
  Stam (1980)  
  12,000 for Holsteins

- $2N_e L$  
  Hayes et al. (2009)  
  6,000

- $2N_e L/[\log(N_e L)]$  
  Goddard et al. (2011)  
  $\sim 500$
Fraction of $G$ variance explained

$G = UDU'$

$\frac{\text{sum}(d_{1:n})}{\text{sum}(D)}$

What fraction of variance in $G$ is information and what is noise?

Pocrnic et al., 2016a
Number of largest eigenvalues to account for a given variance

Pocrnic et al., 2016a
How to determine dimensionality in practice - APY inversion of GRM

Breeding values N chromosome segments

\[ u = Ts \]

Choose any N animals called “core”: \( u_c \)

\[ s = Qu_c + \epsilon_c \]  Segments linear function of core N animals

\[ u_n = T_n s = P_{nc} u_c + \epsilon_n \]  Noncore animals linear functions of core animals

\[
G^{-1} = \begin{bmatrix}
I & -P_{cn} \\
0 & I
\end{bmatrix}
\begin{bmatrix}
G_{cc}^{-1} & 0 \\
0 & M_{nn}^{-1}
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
-P_{nc} & I
\end{bmatrix}
\]

Sparse inverse
True accuracies as function of number of eigenvalues

Pocrnic et al., 2016a
Reliabilities – Jerseys (75k animals)

Milk
Protein
Fat

Pocrnic et al., 2016b

Assumed dimensionality

3300 ≈NeL
6100 ≈2NeL
11,500 ≈4 NeL

(number of core animals)

100% = full inverse ➔ lower accuracy
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>

Pocrnic et al. (2016b)
Impact of reduced dimensionality

- Accuracy with SNP selection
- Theoretical accuracies
- Persistency of GEBV
- GWAS
Understanding of limited dimensionality (I)

\[ \text{Genome} \]

\[ \ldots \quad \text{boxes} \quad \ldots \]

\[ \approx 4 \text{ Ne L segments} \]

Average size \( \frac{L}{4\text{NeL}} \)

With \( \text{Ne}=100, \text{L}=30, \text{Genome size 3 Gb} \Rightarrow 1 \text{ segment} \approx 250 \text{ kb} \]
Understanding of limited dimensionality (II)

Number of haplotypes: 4 Ne L
Ne within each ¼ Morgan segment

Dimensionality of ¼ Morgan case: Ne

⇒ Reduced dimensionality with weighted GRM
ssGBLUP accuracies using SNP60K and 100 QTNs – simulation study

Data: 60k genotyped animals
   60k SNP + 100 QTN

98% Dimensionality:
   19k – unweighted G
   5k  - weighted G
   98  - only QTN

Fragomeni et al. (2017)
Advantage of SNP selection and size of data

Karaman et al., 2016
Accuracy as generation of core animals

Noncore animals linear functions of core animals

$$u_n = P_{nc} u_c + \varepsilon_n$$

Selection does not change segments (additive model only)

Bradford et al. (2017)
Persistence over generations with different sizes of reference populations

Very large – equivalent to 4NeL animals with 99% accuracy
Are SNP effects from Holstein national populations converging?
Accuracy approximations

- Based on equal sized segments (Daetwyler et al., 2008)
  \[ r = \sqrt{\frac{N h^2}{N h^2 + M_e}} \]
  
  \( N \) – number of animals
  
  \( M_e \) – number of segments

- Based on segments modified by QTL frequencies (Goddard, 2009)
  \[ r = \sqrt{1 - \frac{\lambda}{2N \sqrt{\alpha}} \ln \left( \frac{1 + \alpha + 2\sqrt{\alpha}}{1 + \alpha - 2\sqrt{\alpha}} \right)} \]
  
  \( M_e / (h^2 \ln(2N_e)) \)
  
  \( 1 + 2(M_e/N h^2 \ln(2N_e)) \)
Figure 1. Accuracy of genomic breeding values with 5000 phenotypic records, effective population size of 100 and increasing heritability, predicted by the deterministic formula of Goddard (2008) or Daetwyler et al. (2008).
Theory and practice

• Theoretical formulas not useful (Brard and Ricard, 2015)
  • Effect of selection?
  • Wrong numbers?
  • Segments not equal?
Reliabilities assuming different dimensionality with APY inverse – Holsteins

Are chromosome segments unequal size?

Final score

regular $G^{-1}$

Pocrnic et al., 2016b
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

Percentage of variance explained vs. Number of eigenvalues graph
Accuracies of GBLUP using GRM with largest eigenvalues only

- 4k animals with own records
- 4k bulls with 100 daughters

Accuracy

% of explained variance

Eigenvalues

% of explained variance

30 40 50 60 70 80 90 100

0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Varying amount of information

4k bulls with 50 daughters

4k animals with own records
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?
How are eigenvalues influenced by effective population size and genome length?
Largest eigenvalues do not depend on genome size - cluster haplotypes across all genome
PCA Plot

PC1 and PC2 pool segments across genome
Hypothetical accuracies as function of Ne and genome length

Simulation results depend on population parameters
Some hypothesis on GWAS

First cluster
Second cluster

QTL

BayesB1
BayesB2
GBLUP
Classical with pop. stratification
Conclusions

• Large impact of limited dimensionality of genomic info
  • Accuracies
  • Persistence
  • GWAS
  • ...

• Little data required for medium accuracy, large data for high accuracy

• Many hypotheses - potential studies with real data sets

• Collaborators welcome, funding available
Group and sponsors

Shogo Tsuruta
Ignacio Aguilar
Breno Fragomeni
Ivan Pocrnic
Daniela Lourenco
Yutaka Masuda
Andres Legarra
Heather Bradford
Accuracy and distance from markers to QTL

Fragomeni et al. (2017)
Questions - summary

• Is accuracy of GBLUP proportional to explained variance in $\mathbf{G}$?
• Can accuracy of GBLUP be expressed in the terms of variance explained by $N$ largest segments, i.e. eigenvalues of $\mathbf{G}$
  • e.g. 10% variance = 10% accuracy; 50%=50%; ... ?
• Is dimensionality of $\mathbf{G}$ related to number of core animals?
• What accuracies with $n$ core animals that have perfect BV?
• Do accuracies of GBLUP reach 0.99 with many animals?
• Are APY and SVD/EIGEN methods related?