Extension of single-step ssGBLUP to many genotyped individuals

Ignacy Misztal
University of Georgia
Genomic selection and single-step

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

Aguilar et al., 2010
Christensen and Lund, 2010

• Simplicity
  – No DYD or DP
  – No index
  – No complexity

• Accuracy
  – Avoids double counting
  – Avoids fixed index
  – Accounts for preselection bias
Current implementation of SS

\[ H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A^{-1}_{22} \end{bmatrix} \]

- G and A\(_{22}\) created explicitly
- Quadratic memory and cubic computations
- Cost per 100k genotypes - 1.5 hr (Aguilar et al., 2014)
Number of genotypes and impending problem

> 2 M for Holsteins
> 400k for Angus

Genomic pre-selection issue (Patry and Ducrocq, 2011; VanRaden et al., 2013)
  – BLUP increasingly biased
  – Need all data on preselection included
Unsymmetric equations

\[
\begin{bmatrix}
X'X & X'Z \\
HZ'X & HZ'Z + \alpha I
\end{bmatrix}
\begin{bmatrix}
\hat{b} \\
\hat{u}
\end{bmatrix}
= 
\begin{bmatrix}
X'y \\
HZ'y
\end{bmatrix}
\]

Misztal et al., 2009

No convergence without good preconditioner
No convergence with large H or A
No G or $A_{22}$ inverse model

\[
\begin{bmatrix}
X'X & X'W_1 & X'W_2 & 0 & 0 \\
W_1X_1 & W_1W_1 + \alpha_u A^{11} & \alpha_u A^{12} & 0 & 0 \\
W_1X_2 & \alpha_u A^{12} & W_2W_2 + \alpha_u A^{22} & \alpha_u I & -\alpha_u I \\
0 & 0 & \alpha_u I & \alpha_u A^{22} & 0 \\
0 & 0 & \alpha_u I & 0 & \alpha_u G
\end{bmatrix}
\begin{bmatrix}
\hat{b} \\
\hat{u}_1 \\
\hat{u}_2 \\
-\varphi \\
-\hat{\gamma}
\end{bmatrix}
= \begin{bmatrix}
X'Y \\
W_1'y_1 \\
W_2'y_2
\end{bmatrix}
\]

Legarra and Ducrocq (2011)

Slow convergence with few genotypes
Divergence with many genotypes
SNP model for genotyped animals

\[
\begin{bmatrix}
X'X & X'_1W_1 \\
W'_1X_1 & W'_1W_1 + \alpha_u A^{11} \\
Z'W'_2X_2 & \alpha_u Z'A^{12} \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
X'_2W_2Z \\
\alpha_u A^{12}Z \\
Z'_2W'_2Z + \alpha_u Z'A^{22}Z + D^{-1}\sigma^2_e \\
\alpha_u Z
\end{bmatrix}
\begin{bmatrix}
0 \\
0 \\
\alpha_u Z' \\
\alpha_u A_{22}
\end{bmatrix}
= 
\begin{bmatrix}
X'y \\
W'_1y_1 \\
Z'_2W'_2y_2
\end{bmatrix}.
\]

Legarra and Ducrocq, 2011

No successful programming
SNP model for genotyped animals

Liu et al, 2014
SNP effects for all animals  
(Fernando et al., 2014)

\[ \hat{M}_1 = A_{12} A_{22}^{-1} M_2 \]

\[ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1^* \\ X_2^* \end{bmatrix} \beta^* + \begin{bmatrix} Z_1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ Z_2 \end{bmatrix} \begin{bmatrix} \hat{M}_1 \alpha + \epsilon \\ M_2 \alpha \end{bmatrix} + e \]

Cost of imputation
Requires new type of programming
Extension to complex models unclear
Can regular ssGBLUP be made more efficient?
Scaling up $A_{22}^{-1}$

$$A_{22}^{-1} = A^{22} - A^{21}(A^{22})^{-1}A^{12}$$

- $A_{22}^{-1}$ dense (Faux et al., 2014)
- For PCG iteration (Stranden et al., 2014)

$$A_{22}^{-1}q = A^{22}q - \left\{ A^{21} \left[ (A^{22})^{-1} (A^{12}q) \right] \right\}$$

- Seconds for 500k animals with good programming (Masuda et al., 2017)
Is dimensionality of genomic information limited?

• Regular G not positive definite past ~5k
  – Blending with A (VanRaden, 2008)

• Dimensionality of SNP BLUP small (Maciotta et al., 2013)

• Success of imputation

• Manhattan plots noisy until averaged by 300k-10Mb (depending on species)
Origin of Haplotype blocks

Cuppen, 2005
Heterogenetic and homogenic tracts in genome (Stam, 1980)

\[ E(\#\text{tracts}) = 4NeL \] (Stam, 1980)

- \( Ne \) – effective population size
- \( L \) – length of genome in Morgans

Holsteins: \( Ne \approx 100 \) \( L = 30 \)
\( Me = 12,000 \)
Inversion via SVD/eigenvalue decomposition

Assume 1 million animals genotyped with 60k chip

\[ \mathbf{G} = \mathbf{Z} \mathbf{Z}' = \mathbf{U} \mathbf{D} \mathbf{U}' \quad \text{Eigenvalue decomposition (1M x 1M)} \]

\[ \mathbf{G}^{-1} = \mathbf{U} \mathbf{D}^{-1} \mathbf{U}' \quad \text{Generalized inverse (1M x 1M)} \]

\[ \mathbf{Z} = \mathbf{U} \mathbf{S} \mathbf{V} = \mathbf{U} \mathbf{D}^{0.5} \mathbf{V} \quad \text{- SVD decomposition (1M x 60k)} \]

10h for 720k animals (Masuda, 2017)

t - index for non-negligible eigenvalues, say 10k

\[ \mathbf{G}^{-1} = \mathbf{U}_t \mathbf{D}_t^{-1} \mathbf{U}_t' = \mathbf{U}_t \mathbf{S}_t^{-1} \mathbf{S}_t^{-1} \mathbf{U}_t' = \mathbf{U}_* \mathbf{U}_* \]

For PCG iteration

\[ \mathbf{G}^{-1} \mathbf{q} = \mathbf{U}_* \left( \mathbf{U}_* \ \mathbf{q} \right) \quad \text{- only 1 M x 10k elements} \]
Inverse by Woodbury formula

\[ G = ZZ' + I \varepsilon, \]
\[ G^{-1} = \frac{1}{\varepsilon} I - \frac{1}{\varepsilon} Z \left( \frac{1}{\varepsilon} Z'Z + I \right)^{-1} Z' \]

Woodbury formula

\[ Z'Z \text{ 60k x 60k} \]

For PCG iteration:

\[ G^{-1}q = \frac{1}{\varepsilon} \{ I - Z(UU')^{-1}Z' \}q = \frac{1}{\varepsilon} \{ I - SS' \}q \]

\[ S = ZU'D^{-1/2} \]

With reduced rank \( S = ZU_t'(D_t)^{-1/2} \)  \( (1M \times 10k) \)

Mantysaari et al., 2017

Ostersen et al., 2017
If $G$ has limited dimensionality, can $G^{-1}$ be sparse like $A^{-1}$?
Use of relatives for $G^{-1}$
Accuracies not good enough
Theory not clear
Assumption of limited dimensionality

$s - n \times 1$ vector containing additive information of population (haplotypes, chromosome segments, LD blocks)?

Breeding value  

$$u = Ts + e$$

Very small error

If $u_c$ contains $n$ animals:

$$s \approx T_c^{-1}u_c$$

Breeding values of any $n$ animals contains all additive information
Choose core “c” and noncore “n” animals

\[ u_n = P_{nc} u_c + \varepsilon_n \]

\[ u_c = u_c \]

\[
\begin{bmatrix}
  u_c \\
  u_n \\
\end{bmatrix} =
\begin{bmatrix}
  I & 0 \\
  P_{nc} & I
\end{bmatrix}
\begin{bmatrix}
  \varepsilon_n
\end{bmatrix}
\]

\[ \text{var}(\varepsilon_n) = M_{nn} \]

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

\[
G^{-1} =
\begin{bmatrix}
  I & -P_{cn} \\
  0 & I
\end{bmatrix}
\begin{bmatrix}
  G^{-1}_{cc} & 0 \\
  0 & M_{nn}^{-1}
\end{bmatrix}
\begin{bmatrix}
  I & 0 \\
  -P_{nc} & I
\end{bmatrix}
\]
How to estimate $\mathbf{P}$ and $\text{inv}(\mathbf{G})$?

$$\text{var} \left( \begin{bmatrix} \mathbf{u}_c \\ \mathbf{u}_n \end{bmatrix} \right) = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{G}_{nn} \end{bmatrix} \sigma^2_u \quad \mathbf{G} \text{ is “true” relationship matrix}$$

$$\mathbf{u}_n \mid \mathbf{u}_c = \mathbf{G}_{nc} \mathbf{G}_{cc}^{-1} \mathbf{u}_c, \quad \mathbf{P} = \mathbf{G}_{nc} \mathbf{G}_{cc}^{-1}$$

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \mathbf{G}_{cc}^{-1} \mathbf{G}_{cn} \\ \mathbf{G}_{nc} \mathbf{G}_{cc}^{-1} \end{bmatrix} \mathbf{M}^{-1} \begin{bmatrix} \mathbf{G}_{nc} \mathbf{G}_{cc}^{-1} & \mathbf{I} \end{bmatrix}$$

APY algorithm

(Algorithm for Proven and Young)
Properties of APY algorithm

G → G^{-1}

Cost: Quadratic memory and cubic computations

G → APY G^{-1}

Cost: Almost linear memory and computations
Using recursion to compute the inverse of the genomic relationship matrix

I. Misztal,† A. Legarra,‡ and I. Aguilar§
*Department of Animal and Dairy Science, University of Georgia, Athens 30602-2771
†INRA, UR831-SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France
‡Instituto Nacional de Investigación Agropecuaria, Las Brujas 90200, Uruguay

J. Dairy Sci. 99:1–5
http://dx.doi.org/10.3168/jds.2014-9125

Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes

B. O. Fragomeni,*† D. A. L. Lourenco,* S. Tsuruta,* Y. Masuda,* I. Aguilar,† A. Legarra,‡ T. J. Lawlor,§ and I. Misztal‡
*Department of Animal and Dairy Science, University of Georgia, Athens 30602
†Instituto Nacional de Investigación Agropecuaria, Canelones, 90200, Uruguay
‡INRA, UMR1368 GenePhySE, Castanet Tolosan, 31326, France
§Holstein Association USA Inc., Betteboro, VT 05602

Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size

Ignacy Misztal†
Animal and Dairy Science, University of Georgia, Athens, Georgia 30602

The Dimensionality of Genomic Information and Its Effect on Genomic Prediction

Ivan Pocmíč,** Daniela A. L. Lourenço,† Yutaka Masuda,* Andres Legarra,*† and Ignacy Misztal‡
*Department of Animal and Dairy Science, University of Georgia, Athens, Georgia 30602, and †Institut National de la Recherche Agronomique, GenPhySE, F 31326 Castanet-Tolosan, France
Reliabilities – Holsteins (77k)

Pocrnic et al., 2016b
Distribution of segments/haplotypes/..
Costs with 720k genotyped animals

- 30 M Holsteins
- 50 M records
- 764k 60k genotypes

<table>
<thead>
<tr>
<th>Item</th>
<th>BLUP</th>
<th>ssGBLUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>APY G</td>
<td>-</td>
<td>7 h</td>
</tr>
<tr>
<td>A22-1</td>
<td>-</td>
<td>10 min</td>
</tr>
<tr>
<td>rounds</td>
<td>402</td>
<td>464</td>
</tr>
<tr>
<td>Time/round</td>
<td>51 s</td>
<td>83 s</td>
</tr>
<tr>
<td><strong>Total time</strong></td>
<td><strong>6 h</strong></td>
<td><strong>17 h</strong></td>
</tr>
</tbody>
</table>

Masuda et al., 2017
Which core animals in APY?

Bradford et al. (2017)

- Simulated populations (QMSim; Sargolzaei and Schenkel, 2009)
- $N_e = 40$
- #genotyped animals = 50,000

- Core animals:
  - Random gen 6 || gen 7 || gen 8 || gen 9 || gen 10 (y)
  - Random all generations
Which core animals in APY?

![Diagram showing Accuracy with different percentages explained in G and NeL.](image)

Bradford et al. (2016)
Persistence over generations

Very large – equivalent to $4N_e L$ animals with 99% accuracy
Are SNP effects from Holstein national populations converging
Theory of limited dimensionality

Number of haplotypes: $4 \cdot \text{Ne} \cdot L$

$\text{Ne}$ within each $\frac{1}{4}$ Morgan segment

Fragomeni et al., 2018

Dimensionality of $\frac{1}{4}$ Morgan case: $\text{Ne}$

or number of identified QTLs

$\Rightarrow$ Reduced dimensionality with weighted GRM
ssGBLUP accuracies using SNP60K and 100 QTNs – simulation study

Fragomeni et al. (2017)

- BLUP
- ssGBLUP - unweighted SNP60k
- unweighted SNP60k + 100 QTN
- SNP60k + 100 QTN weighted by GWAS
- SNP60K + 100 QTN with "true" variance
- plus by APY
- only 100 QTN unweighted by APY

Rank

- BLUP: 0
- ssGBLUP - unweighted SNP60k: 19k
- unweighted SNP60k + 100 QTN: 5k
- SNP60k + 100 QTN weighted by GWAS: 0
- SNP60K + 100 QTN with "true" variance: 98
- plus by APY: 0
- only 100 QTN unweighted by APY: 0
Multitrait ssGBLUP or SNP selection?

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

![Graph showing $R^2$ vs. $n_R$ for GBLUP and BayesB](image.png)
Variance components

• Based on SNP
  – limitations

• REML based on relationships
  – Equations no longer sparse
  – YAMS sparse matrix package – up to 100 times speedup (Masuda et al., 2017)
  – APY for REML

• Method R (Legarra and Reverter, 2017)
Extra topics

• Matching pedigrees and genomic relationships
• Missing pedigrees
• Crossbreeding
• Causative SNP

• Haplotypes for crossbreds (Christensen et al., 2016)
• Metafounders (Legarra et al., 2016)
• Approximation of reliabilities
Conclusions

• Limited dimensionality of genomic information due to limited effective population size

• ssGBLUP suitable for any data set and model

• With large data sets for Holsteins:
  – Good persistence of predictions
  – Convergence of predictions from different countries
Acknowledgements

Tom Lawlor, Holstein Assoc
Paul VanRaden, AGIL USDA

Shogo Tsuruta
Ignacio Aguilar
Breno Fragomeni
Ivan Pocrnic
Daniela Lourenco
Yutaka Masuda
Andres Legarra
Heather Bradford
Theory for APY

• Breeding values of core animals linear functions of:
  – Independent chromosome segments (Me)
  – Independent effective SNP

• $E(Me) = 4 \times Ne \times L$ (Stam, 1980; VanRaden, 2008)
  
  \begin{align*}
  Ne & \text{ – effective population size} \\
  L & \text{ – length of genome in Morgans} \\
  Me & = 4 \times (Ne=100) \times (L=30) = 12,000
  \end{align*}
Accuracy and distance from markers to QTL

Fragomeni et al. (2017)