Efficient computations for single-step genomic evaluations

I. Strandén, K. Matilainen, G.P. Aamand, E.A. Mäntysaari
Introduction

- Single-step MME:
  \[
  \begin{bmatrix}
  X'X & X'W \\
  W'X & W'W + \lambda H^{-1}
  \end{bmatrix}
  \begin{bmatrix}
  \hat{b} \\
  \hat{a}
  \end{bmatrix}
  =
  \begin{bmatrix}
  X'y \\
  Z'y
  \end{bmatrix}
  \]

  where \( \lambda = \frac{\sigma^2_e}{\sigma^2_a} \) and

  \[
  H^{-1} = \begin{bmatrix}
  A_{11} & A_{12} \\
  A_{21} & A_{22}
  \end{bmatrix}
  + \begin{bmatrix}
  0 & 0 \\
  0 & G^{-1} - (A_{22})^{-1}
  \end{bmatrix}
  \]

- Number of genotype animals increases
  \( \Rightarrow \) Making and inverting \( A_{22} \) and \( G \) matrices becomes difficult

- This study:
  - APY (Algorithm for Proven and Young animals) by Cholesky
  - Matrix vector product \( A_{22}^{-1}d_2 \) without ever making the \( A_{22} \) matrix
    - Three different approaches
  - Application: Nordic Holstein fertility model data
Original APY (Misztal et al., 2015)

APY idea:
- Divide to core (c) and young (y) animals
- Approximate $G$ inverse by making $G_{yy}$ diagonal

$$G_{APY}^{-1} = \begin{bmatrix} G_{cc}^{-1} + G_{cc}^{-1}G_{cy}M_{yy}^{-1}G_{yc}G_{cc}^{-1} & -G_{cc}^{-1}G_{cy}M_{yy}^{-1} \\ -M_{yy}^{-1}G_{yc}G_{cc}^{-1} & M_{yy}^{-1} \end{bmatrix}$$

$M_{yy} = \text{Diagonal of } G_{yy} - G_{yc}G_{cc}^{-1}G_{cy}$

Large sparse sub-matrix
→ Lower storage need
→ Lower computational load

Number of core animals low: less than number of markers
Original APY (Misztal et al., 2015)

APY idea:

- Divide to core (c) and young (y) animals
- Approximate $G$ inverse by making $G_{yy}$ diagonal

\[
G_{\text{APY}}^{-1} = 
\begin{bmatrix}
G_{cc}^{-1} + G_{cc}^{-1} G_{cy} M_{yy}^{-1} G_{yc} G_{cc}^{-1} & -G_{cc}^{-1} G_{cy} M_{yy}^{-1} \\
-M_{yy}^{-1} G_{yc} G_{cc}^{-1} & \tilde{M}_{yy}^{-1}
\end{bmatrix}
\]

\[
G_{\text{APY}} = 
\begin{bmatrix}
G_{cc} & G_{cy} \\
G_{yc} & \tilde{M}_{yy} + G_{yc} G_{cc}^{-1} G_{cy}
\end{bmatrix}
\]

Large sparse sub-matrix
→ Lower storage need
→ Lower computational load

Number of core animals low: less than number of markers
APY by Cholesky decomposition

\[
G_{APY} = \begin{bmatrix}
G_{cc} & G_{cy} \\
G_{yc} & M_{yy} + G_{yc} G_{cc}^{-1} G_{cy}
\end{bmatrix}
\]

Steps:

1. Let \( L \) in \( G_{apy} = LL' \) where \( L = \)

where:

- \( G_{cc} = L_{cc} L_{cc}' \)
- \( C = G_{yc} L_{cc}^{-1} \)
- \( M_{yy} = \) diagonal of \( G_{yy} \cdot CC' \)

2. Calculate the inverse

where \( B = M_{yy}^{-0.5} C L_{cc}^{-1} \)

\[
G_{APY}^{-1} = \begin{bmatrix}
(L_{cc}')^{-1} L_{cc}^{-1} + B' B & -B M_{yy}^{-0.5} \\
-M_{yy}^{-0.5} B & M_{yy}^{-1}
\end{bmatrix}
\]
APY original vs. Cholesky approach

Original:

\[
G_{\text{APY}}^{-1} = \begin{bmatrix}
G_{cc}^{-1} + G_{cc} G_{cy} M_{yy}^{-1} G_{yc} G_{cc}^{-1} & -G_{cc}^{-1} G_{cy} M_{yy}^{-1} \\
-M_{yy}^{-1} G_{yc} G_{cc}^{-1} & M_{yy}^{-1}
\end{bmatrix}
\]

\[M_{yy} = \text{Diagonal of } G_{yy} - G_{yc} G_{cc}^{-1} G_{cy}\]

Using Cholesky:

\[
G_{\text{APY}}^{-1} = \begin{bmatrix}
(L'_{cc})^{-1} L_{cc}^{-1} + B'B & -BM_{yy}^{-0.5} \\
-M_{yy}^{-0.5} B & M_{yy}^{-1}
\end{bmatrix}
\]

where

\[B = M_{yy}^{-0.5} C L_{cc}^{-1}\]
\[C = G_{yc} (L'_{cc})^{-1}\]

\[M_{yy} = \text{Diagonal of } G_{yy} - C C'\]

Cholesky decomposition brings
- numerical stability through decomposition
- programming simplicity with (parallel) BLAS and LAPACK subroutines
A_{22} inverse part

- Matrix blocks: 
  \[ \mathbf{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \]
  \[ \mathbf{A}^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \]

- Use equality:
  \[ (A_{22})^{-1} = A_{22} - A_{21} (A_{11})^{-1} A_{12} \]

- Steps to calculate \( z_2 = (A_{22} - A_{21} (A_{11})^{-1} A_{12}) d_2 \)
  1. \[ [x_1 \ x_2]' = [A_{21} \ A_{22}]' d_2 \] Pedigree read: \( \mathbf{A}^{-1} \) rules
  2. \( y_1 = (A_{11})^{-1} x_1 \)
  3. \( z_2 = x_1 - A_{21} y_1 \)

- Step 2. alternatives:
  - **IOP**: PCG solve by \textit{Iteration on Pedigree}: pedigree reading
  - **IM**: PCG solve by \( A_{11} \) in memory
  - **CM**: Direct solving using \textit{CHOLMOD} library
Study design

Models:
- AMBLUP: Animal model BLUP, no genomics
- ssGBLUP: single-step, has genomic information

Inverse of $A_{22}$
- Calculate $(A^{11})^{-1} \rightarrow$ regular ssGBLUP
- Step 2: $y_1 = (A^{11})^{-1} x_1$
  - PCG and IOP: iteration on pedigree
  - PCG and IM: $A^{11}$ in memory
  - Direct: CHOLMOD

APY 10,000 or 20,000 in core: Highest number of progeny
Holstein fertility data and Eurogenomics genotyped animals

- 81,031 genotyped (clones removed)
  - 46,344 markers
  - APY: core was 10,000 (10K) or 20,000 (20K) animals
- 9.73 million animals in pedigree
- 7.5 million record animals
  - 11 traits, fertility model, no genetic groups
  - 48.7 million observations

File sizes:
- Text genotype marker data: 7 GB
- Binary $G^{-1}$ – $(A_{22})^{-1}$ and $G^{-1}$ file sizes: 13 GB
- Binary APY $G^{-1}$ file size: 2.9 GB (10K), 5.3 GB (20K)
## First results: ssGBLUP: APY and $A_{22}^{-1}$ by IOP

<table>
<thead>
<tr>
<th>Method</th>
<th>Preprocessor</th>
<th>Solver</th>
<th>Total</th>
<th>$N$ iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBLUP</td>
<td>6m</td>
<td>6h 15m</td>
<td>6h 21m</td>
<td>1956</td>
</tr>
<tr>
<td>ssGBLUP</td>
<td>$A_{22}$ by RelaX2</td>
<td>2h 48m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>regular</td>
<td>$G^{-1} - (A_{22})^{-1}$</td>
<td>24h 26m</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Preprocessor</td>
<td>16m</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Solver</td>
<td>38h 2m</td>
<td>65h 32m</td>
<td>2088</td>
</tr>
<tr>
<td>ssGBLUP</td>
<td>$G^{-1}$</td>
<td>10h 19m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No $A_{22}^{-1}$ build</td>
<td>Preprocessor</td>
<td>17m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOP</td>
<td>Solver</td>
<td>52h 12m</td>
<td>62h 48m</td>
<td>2108</td>
</tr>
<tr>
<td>ssGBLUP</td>
<td>$G^{-1}$</td>
<td>5h 22m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APY 10K</td>
<td>Preprocessor</td>
<td>9m</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Solver</td>
<td>28h 13m</td>
<td>34h 44m</td>
<td>1963</td>
</tr>
</tbody>
</table>

Parallel computing would allow faster computations especially for precalculating matrices like $G^{-1}$ and $G^{-1} - (A_{22})^{-1}$
## Solver CPU time and peak memory

<table>
<thead>
<tr>
<th>Solver Type</th>
<th>Mem (GB)</th>
<th>Solver time</th>
<th>Time/iter</th>
<th>N iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBLUP</td>
<td>3.61</td>
<td>8h</td>
<td>0.24m</td>
<td>1981</td>
</tr>
<tr>
<td>Regular ssGBLUP</td>
<td>3.62</td>
<td>36h</td>
<td>1.04m</td>
<td>2059</td>
</tr>
<tr>
<td>No $A_{22}^{-1}$ build, IOP</td>
<td>4.61</td>
<td>49h</td>
<td>1.44m</td>
<td>2038</td>
</tr>
<tr>
<td>No $A_{22}^{-1}$ build, IM</td>
<td>4.62</td>
<td>42h</td>
<td>1.22m</td>
<td>2046</td>
</tr>
<tr>
<td>No $A_{22}^{-1}$ build, CM</td>
<td>4.78</td>
<td>36h</td>
<td>1.06m</td>
<td>2041</td>
</tr>
<tr>
<td>APY 20K, CM</td>
<td>4.78</td>
<td>21h</td>
<td>0.60m</td>
<td>2049</td>
</tr>
<tr>
<td>APY 10K, CM</td>
<td>4.78</td>
<td>15h</td>
<td>0.46m</td>
<td>1915</td>
</tr>
</tbody>
</table>

Regular ssGBLUP: Additional time due to $A_{22}$ is about 14 hours (building and inversion).
Correlation and std for GEBV of genotyped animals: ssGBLUP vs. APY

<table>
<thead>
<tr>
<th>Trait</th>
<th>Corr</th>
<th>Std ssGBLUP</th>
<th>Std APY</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRR0</td>
<td>0.999</td>
<td>2.79</td>
<td>2.78</td>
</tr>
<tr>
<td>IFL0</td>
<td>0.999</td>
<td>3.32</td>
<td>3.31</td>
</tr>
<tr>
<td>NRR1</td>
<td>0.999</td>
<td>4.25</td>
<td>4.24</td>
</tr>
<tr>
<td>ICF1</td>
<td>0.999</td>
<td>6.56</td>
<td>6.55</td>
</tr>
<tr>
<td>IFL1</td>
<td>0.999</td>
<td>10.28</td>
<td>10.27</td>
</tr>
<tr>
<td>NRR2</td>
<td>0.999</td>
<td>4.36</td>
<td>4.33</td>
</tr>
<tr>
<td>ICF2</td>
<td>0.999</td>
<td>6.85</td>
<td>6.85</td>
</tr>
<tr>
<td>IFL2</td>
<td>0.999</td>
<td>10.05</td>
<td>10.03</td>
</tr>
<tr>
<td>NRR3</td>
<td>0.999</td>
<td>4.04</td>
<td>4.02</td>
</tr>
<tr>
<td>ICF3</td>
<td>0.999</td>
<td>6.57</td>
<td>6.56</td>
</tr>
<tr>
<td>IFL3</td>
<td>0.999</td>
<td>9.38</td>
<td>9.33</td>
</tr>
</tbody>
</table>

NRR= Non-return rate
IFL= length of service period
ICF= interval from calving to first breeding

<table>
<thead>
<tr>
<th>Trait</th>
<th>Corr</th>
<th>Std ssGBLUP</th>
<th>Std APY</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRR0</td>
<td>1.000</td>
<td>2.79</td>
<td>2.78</td>
</tr>
<tr>
<td>IFL0</td>
<td>1.000</td>
<td>3.32</td>
<td>3.32</td>
</tr>
<tr>
<td>NRR1</td>
<td>1.000</td>
<td>4.25</td>
<td>4.25</td>
</tr>
<tr>
<td>ICF1</td>
<td>1.000</td>
<td>6.56</td>
<td>6.56</td>
</tr>
<tr>
<td>IFL1</td>
<td>1.000</td>
<td>10.28</td>
<td>10.27</td>
</tr>
<tr>
<td>NRR2</td>
<td>1.000</td>
<td>4.36</td>
<td>4.34</td>
</tr>
<tr>
<td>ICF2</td>
<td>1.000</td>
<td>6.85</td>
<td>6.85</td>
</tr>
<tr>
<td>IFL2</td>
<td>1.000</td>
<td>10.05</td>
<td>10.04</td>
</tr>
<tr>
<td>NRR3</td>
<td>1.000</td>
<td>4.04</td>
<td>4.04</td>
</tr>
<tr>
<td>ICF3</td>
<td>1.000</td>
<td>6.57</td>
<td>6.57</td>
</tr>
<tr>
<td>IFL3</td>
<td>1.000</td>
<td>9.38</td>
<td>9.34</td>
</tr>
</tbody>
</table>

0= Heifer trait
1-3= cow trait 1-3 parity
IFL2, 10,000 or 20,000 core

ssGBLUP std of GEBV (genotyped) = 10.05

10,000 in core

20,000 in core

Correlations:

<table>
<thead>
<tr>
<th></th>
<th>10,000 in core</th>
<th>20,000 in core</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Young</td>
<td>0.9990</td>
<td>0.9999</td>
</tr>
</tbody>
</table>
Conclusions

• APY allowed reducing computing time substantially
  – The smaller the core the lower computing time

• Alternative computing strategies to $A_{22}^{-1}d$ gave good results:
  – CHOLMOD → Increased computing time minimally but memory need was increased and remained acceptable

→ Single-step models with many genotyped animals can be analyzed
  ▪ APY works logically: the more animals in core the better match

• Simply taking animals with highest number of progeny worked here
Thank you!
Thank you!