Efficient computational strategies for multivariate single-step SNPBLUP

Jeremie Vandenplas, Herwin Eding, Mario P.L. Calus

August 29, 2018
Breed4Food is dedicated to be the leading research consortium in animal breeding, genetics and genomics enabling the Breed4Food partners to breed better products to benefit society’s needs.
Single-step genomic evaluations

- Prediction of genomic breeding values
  - Genotyped and non-genotyped animals
  - Potentially high computational costs

- Most software not suitable for single-step SNPBLUP

- Usual iterative solver
  - Preconditioned Conjugate Gradient method
  - Main cost
    - Coefficient matrix * vector
Aim

Investigate the **main computational costs and implement solutions**

for **efficiently solving**

a multivariate **single-step SNPBLUP**

with the **PCG method**
ssSNPBLUP – model

- Hybrid model
  - Non-genotyped animals: breeding value model
  - Genotyped animals: SNP model

\[
y = Xb + \begin{bmatrix} Z_n & 0 & 0 \\ 0 & Z_g & Z_g M \end{bmatrix} \begin{bmatrix} u_n \\ a_g \\ g \end{bmatrix} + e
\]

\[
u_g = a_g + Mg
\]

- \(u_n, u_g\): aggregate GEBVs for (non-)genotyped animals
- \(a_g\): residual polygenic effects for genotyped animals
- \(g\): SNP effects
- \(M\): SNP genotypes

Fernando et al., 2016; Mantysaari and Strand, 2016
PCG in animal breeding

- **Main computational cost** for one iteration
  
  - Coefficient matrix (C) * vector (v)

- Usually performed in **two parts** + matrix-free approach

\[
C \ast v = L \ast v + R \ast v
\]

\[
\begin{bmatrix}
X'X & X'Z \\
Z'X & Z'Z + G^{-1}
\end{bmatrix} \ast v = \begin{bmatrix}
X'X & X'Z \\
Z'X & Z'Z
\end{bmatrix} \ast v + \begin{bmatrix}
0 & 0 \\
0 & G^{-1}
\end{bmatrix} \ast v
\]

- **Least Squares** part of C
- **Random** part of C
ssSNPBLUP – potential costs

- “Traditional” matrix-free approach

\[ C \ast v = L \ast v + R \ast v \]

Least Squares part

\( 2 \) multiplications with \( M \)

Random part

\( 2 \) multiplications with \( M \)

- \( 4 \) multiplications with \( M \)

- High memory/IO requirement for \( M \)
  - 1,000,000 50K genotypes (dp): 373 GB

- Difficult to parallelize (e.g., inverted pedigree relationship (sub-)matrices \( A^{xx} \))
ssSNPBLUP – solutions

1. Potential cost: 4 multiplications with \( M \)

Solution: Rearrangement of the equations

\[
\mathbf{C} \mathbf{v} = \mathbf{L} \mathbf{v} + \mathbf{R} \mathbf{v} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{M}' \end{bmatrix} (\mathbf{L}_{sub} + \mathbf{R}_{sub_1}) \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{M} \end{bmatrix} \mathbf{v} + \mathbf{R}_{sub_2} \mathbf{v}
\]

With \( M \)

Without \( M \)

Computation of \( \mathbf{C} \mathbf{v} \) in 4 steps

Only 2 multiplications with \( M \)
2. Potential cost: high memory/IO for $M$

⇒ **Solution:** Compressed genotypes (Plink bed format)

<table>
<thead>
<tr>
<th>SNP genotype</th>
<th>Homozygous first allele</th>
<th>Heterozygous</th>
<th>Homozygous second allele</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decimal</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2-bit</td>
<td>00</td>
<td>01</td>
<td>11</td>
<td>10</td>
</tr>
</tbody>
</table>

$$32_{10} \Leftrightarrow \text{00101101} \Leftrightarrow 45$$

4 SNP genotypes
4-32 bytes

1-byte integer
ssSNPBLUP – solutions

50,000 SNPs

Graph showing the memory (GB) required for different methods as the number of genotyped animals increases. The methods include compressed format, 1-byte integer, single precision, and double precision. The graph indicates a significant difference in memory requirements, with double precision requiring the most memory and compressed format requiring the least. The graph is labeled with memory sizes x4, x16, and x32 for different ranges of genotyped animals.
ssSNPBLUP – solutions

3. Potential cost: difficult to parallelize operations with $A^{xx}$

Solution: hold $A^{xx}$ in memory

- Optimized and parallelized libraries
- E.g., Sparse BLAS, Pardiso
ssSNPBLUP – solutions


Solution: New equation

\[ Qv = (A^{gg} - A^{-1}_{gg})v \]
\[ = (A^{gn}(A^{nn})^{-1}A^{ng})v \]
\[ = (A^{gn}_{anc}(A^{nn}_{anc})^{-1}A^{ng}_{anc} + \Delta)v \]

\( A^{nn}_{anc} \): size(ancestors of genotyped animals)
\( \Delta \): sparse + depends only on non-genotyped progeny of genotyped animals

Fernando et al., 2016; Legarra and Ducrocq, 2012
Example – data & hard/software

- CRV 4-trait evaluation
  - Temperament and milking speed
  - Pedigree 6,130,519
  - Phenotypes 3,882,772
  - Genotypes 90,963
  - SNPs 37,994

- Hardware: 528 GB RAM / 32 CPUs (only 5 CPUs used)

- Fortran + OpenMP program
  - Intel MKL library (BLAS, sparse BLAS, PARDISO)
Example – Time and memory

**ssSNPBLUP**: limited amount of memory and time / iteration

- Max. RAM: 7.7 GB
- Average time / iteration: 3.6 s
  - Time / imputation on-the-fly: 0.15 s
  - Time / 2 multiplications with $M$: 1.48 s
- # iterations: 10,000

**ssGBLUP + APY** (13K core animals)

- Max. RAM: 14.2 GB
- Average time / iteration: 2.5 s
- # iterations: 1258
Conclusions

- Multi-trait ssSNPBLUP
  - Feasible on current hardware
    - With limited amounts of memory and time
    - Even with >1,000,000 genotypes
      - Jan ten Napel, EAAP 2018, session 12
  - Under study
    - More complex models
    - Convergence issues
Thank you!