Efficient computational strategies for multivariate single-step SNPBLUP

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

August 29, 2018

WAGENINGEN
UNIVERSITY & RESEARCH

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.

WAGENINGEN UR
For quality of life

CRV

Cobb

HENDRIX GENETICS

Topigs Norsvin
PROGRESS IN PIGS

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

$$\mathbf{y} = \mathbf{Xb} + \begin{bmatrix} \mathbf{Z}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_g & \mathbf{Z}_g\mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{u}_n \\ \mathbf{a}_g \\ \mathbf{g} \end{bmatrix} + \mathbf{e}$$

$$\mathbf{u}_g = \mathbf{a}_g + \mathbf{Mg}$$

$\mathbf{u}_n$, $\mathbf{u}_g$: aggregate GEBVs for (non-)genotyped animals

$\mathbf{a}_g$: residual polygenic effects for genotyped animals

$\mathbf{g}$: SNP effects

$\mathbf{M}$: SNP genotypes

Fernando et al., 2016; Mantysaari and Stranden, 2016

PCG in animal breeding

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

- Usually performed in two parts + matrix-free approach

$$\mathbf{C} * \mathbf{v} = \mathbf{L} * \mathbf{v} + \mathbf{R} * \mathbf{v}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} * \mathbf{v} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} * \mathbf{v} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix} * \mathbf{v}$$

Least Squares part of **C**

Random part of **C**

ssSNPBLUP – potential costs

- "Traditional" matrix-free approach

$$\mathbf{C} * \mathbf{v} = \mathbf{L} * \mathbf{v} + \mathbf{R} * \mathbf{v}$$

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 $\mathbf{A}^{xx}$)

ssSNPBLUP – solutions

1. Potential cost: 4 multiplications with **M**

➔**Solution: Rearrangement of the equations**

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

With **M**

Without **M**

➔Computation of $\mathbf{C} * \mathbf{v}$ in 4 steps

➔Only **2** multiplications with **M**

ssSNPBLUP – solutions

2. Potential cost: high memory/IO for **M**

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

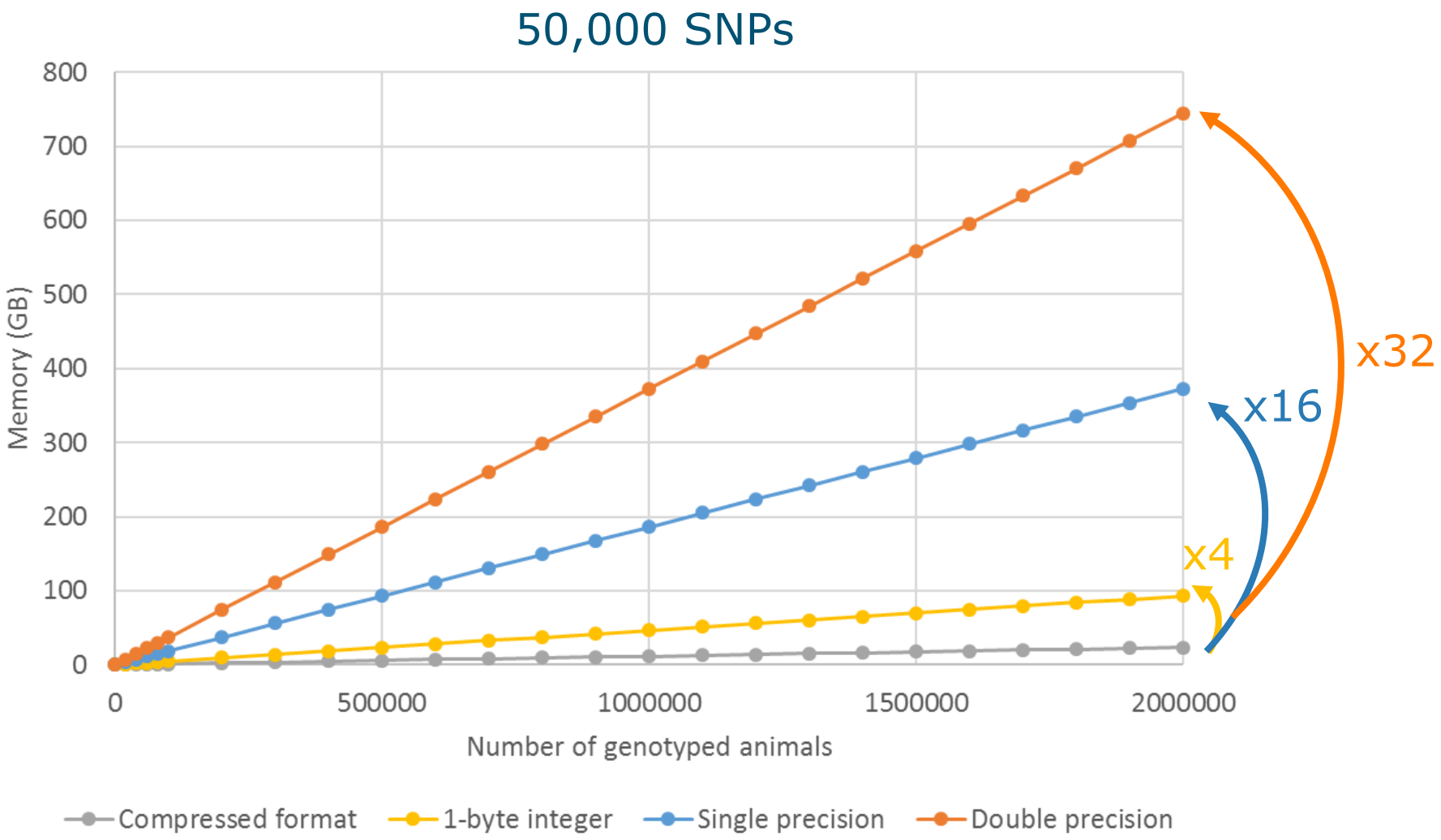
SNP genotype	Homozygous first allele	Heterozygous	Homozygous second allele	Missing
Decimal	0	1	2	3
2-bit	00	01	11	10

3210 ⇔ 00101101 ⇔ 45

4 SNP genotypes
4-32 bytes

1-byte integer

ssSNPBLUP – solutions



WAGENINGEN
UNIVERSITY & RESEARCH

ssSNPBLUP – solutions

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

➔**Solution**: **hold $\mathbf{A}^{xx}$ in memory**

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

ssSNPBLUP – solutions

4. Potential cost: "On-the-fly imputation"

➔**Solution**: **New equation**

$$\mathbf{Qv} = \left(\mathbf{A}^{gg} - \mathbf{A}_{gg}^{-1}\right)\mathbf{v}$$

$$= (\mathbf{A}^{gn}(\mathbf{A}^{nn})^{-1}\mathbf{A}^{ng})\mathbf{v}$$

$$= \left(\mathbf{A}_{anc}^{gn}(\mathbf{A}_{anc}^{nn})^{-1}\mathbf{A}_{anc}^{ng} + \Delta\right)\mathbf{v}$$

Dense $\mathbf{A}_{gg}^{-1}$

Large and sparse $\mathbf{A}^{nn}$ (~size(pedigree))

Small and sparse $\mathbf{A}_{anc}^{nn}$

$\mathbf{A}_{anc}^{nn}$: size(ancestors of genotyped animals)

Δ: sparse + depends only on non-genotyped progeny of genotyped animals

WAGENINGEN
UNIVERSITY & RESEARCH

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!

WAGENINGEN
UNIVERSITY & RESEARCH