EFFICIENT BLOCK-GIBBS SAMPLING IN VARIANCE COMPONENT ESTIMATION

for predictions which combine phenotypic and genomic information

Viktor Milkevych

Per Madsen

Hongding Gao

Just Jensen

AARHUS UNIVERSITET

12 February 2018

WCGALP, Auckland

GenSAP

INTRODUCTION

- Genomic-based selection is widely applied in animal breeding.
- The data sets include non-genotyped individuals - the obvious method of choice is **single-step approach**.
- For estimation of **unknown variance** the **Gibbs sampler** is of practical importance.

MOTIVATION and OBJECTIVE

- Desirable efficiency of Gibbs sampler is not always achievable.

- It partly relies on the properties of variance-covariance matrix.

We study the **effect of amount of genomic information** in the model on **performance and efficiency of Gibbs sampler** using a consecutive and block updating schemes.

UNIVARIATE LINEAR MIXED MODEL

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e}$$

$\boldsymbol{y}$ - vector of observations;

$\boldsymbol{b}$ - vector of mean;

$\boldsymbol{a}$ - vector of random effects;

$\boldsymbol{e}$ - residual vector;

$\boldsymbol{X}$, $\boldsymbol{Z}$ - known incidence matrices.

MIXED MODEL EQUATIONS

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{Z} + {H_*}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{y} \end{bmatrix}$$

$$\boldsymbol{H}_*^{\;-1} = \boldsymbol{A}^{-1} + \boldsymbol{G}_*;$$

$$\boldsymbol{G}_* = \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - {\boldsymbol{A}_{22}}^{-1} \end{bmatrix}.$$

$\boldsymbol{A}$ - numerator relationship matrix;
$\boldsymbol{G}$ - genomic relationship matrix;
$\boldsymbol{H}$ - combined phenotypic-genomic relationship matrix;

PROPERTIES OF GIBBS SAMPLER

1. Markov chain has a transition density with **mean**:

$$E(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^{t}) = \boldsymbol{B}\boldsymbol{\theta}^{\boldsymbol{t}} + \boldsymbol{c}.$$

2. And **dispersion**: $\boldsymbol{\Sigma} - \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^{T}.$

3. The exact **convergence rate**: $\rho = \rho(\boldsymbol{B});$

ρ - spectral radius of $\boldsymbol{B} = -\boldsymbol{L}^{T}\boldsymbol{U}.$

VARIABLE $\boldsymbol{G}_*$

$$\boldsymbol{H}_*^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{G}_*$$

$\boldsymbol{G}_* \in \{\boldsymbol{0}, \boldsymbol{G}\}$: for vector of random effects

$\boldsymbol{a} \sim N(0, \boldsymbol{A}{\sigma_a}^2)$;

$\boldsymbol{a} \sim N(0, \boldsymbol{G}{\sigma_a}^2)$;

$\boldsymbol{a} \sim N(0, \boldsymbol{H}{\sigma_a}^2)$;

Markov chain transition density **mean**:

$$E(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^{t}) = \boldsymbol{B}\boldsymbol{\theta}^{t} + \boldsymbol{L}^{-1}(\boldsymbol{\mu} - \boldsymbol{d}^{t+1}); \qquad \boldsymbol{d}^{t+1} = \boldsymbol{G}_*\boldsymbol{\theta}^{t}.$$

FORMAL OBJECTIVE

We study the **effect of disturbance vector**:

$$\boldsymbol{d}^{t+1} = \boldsymbol{G}_{*}\boldsymbol{\theta}^{t}.$$

DATA

- Danish Jersey cattle population simulated using **ADAM** software (Aarhus University, QGG).
- Genome consisted of 30 chromosomes, each 100 cM in length.
- Conventional breeding scheme.
- Phenotypes: **16945**; animals in pedigree: **19701**.

Number of non-zero elements in variance structure

Genotyped individuals, $\times 10^3$	**Number of elements,** $\times 10^6$	gi
0	0.06	0
3.2	10.17	0.19
6.5	41.90	0.38
8.4	70.04	0.50
10.7	114.04	0.63
12.8	163.57	0.76
14.9	221.57	0.88
16.6	276.30	0.98
16.9	287.13	1

MODEL

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e}$$

$\boldsymbol{y}$ - vector of observations (stature) ;

$\boldsymbol{b}$ - vector of mean (herd-year-season, HYS: 4 seasons, 5 years, 25 herds);

$\boldsymbol{a}$ - vector of animal effects;

$\boldsymbol{e}$ - residual vector.

UPDATING SCHEME

- Target vector $\boldsymbol{\theta} = (\boldsymbol{b}, \boldsymbol{a}, {\sigma_a}^2, {\sigma_e}^2)^T$ with a density $P(\boldsymbol{\theta})$.

- Conventional update:

  Gibbs sampler generates transition states $\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t+1}$ consecutively.

- Block update:

  The *m*-dimensional random effect vector $\boldsymbol{\theta_a}$ is grouped into one block

  $\boldsymbol{\theta}_a = (\theta_{a_1}, \theta_{a_2}, \dots, \theta_{a_m})^T$, the rest

  $\boldsymbol{\theta}_{-a} = (\boldsymbol{b}, {\sigma_a}^2, {\sigma_e}^2)^T$ - not blocked.

COMPUTATIONAL DETAILS

Sampling algorithm:

- **input:** precision matrix $\boldsymbol{M}$
- **output:** $\boldsymbol{\theta} \sim N(\boldsymbol{0}, \boldsymbol{M}^{-1})$

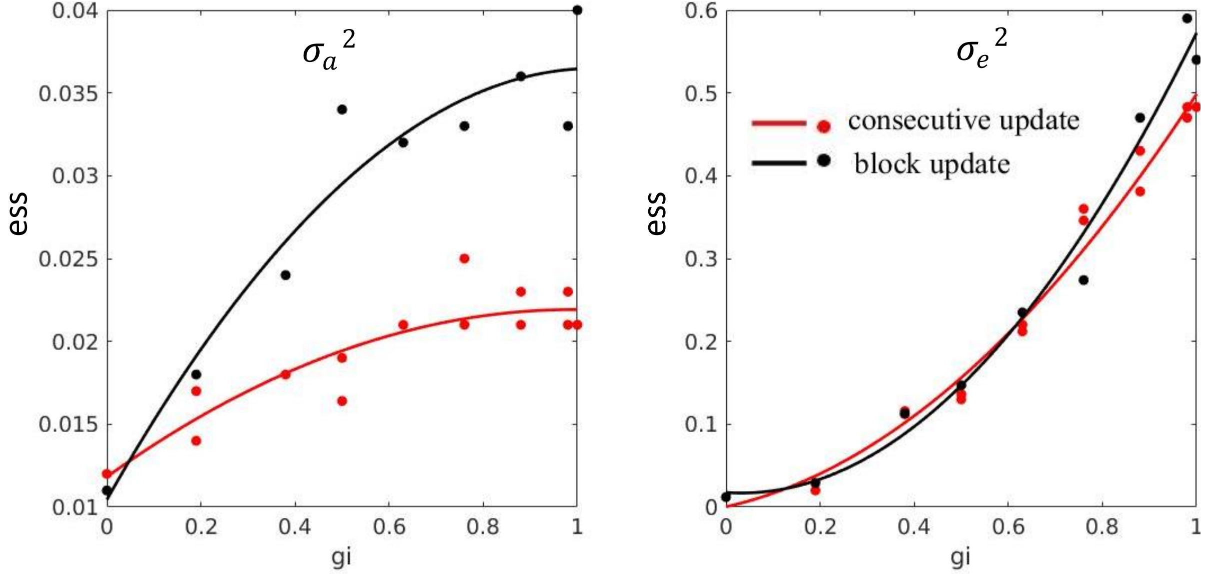
1. Cholesky decomposition: $\boldsymbol{M} = \boldsymbol{C}\boldsymbol{C}^T$
2. Sampling: $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I})$
3. Solving: $\boldsymbol{C}^T\boldsymbol{\theta} = \boldsymbol{z}$

COMPUTATIONAL DETAILS

Implementation:

- **MCMC** package of **DMU** software (Aarhus University, QGG).
- **DMU** is software for analysis of multivariate mixed models.

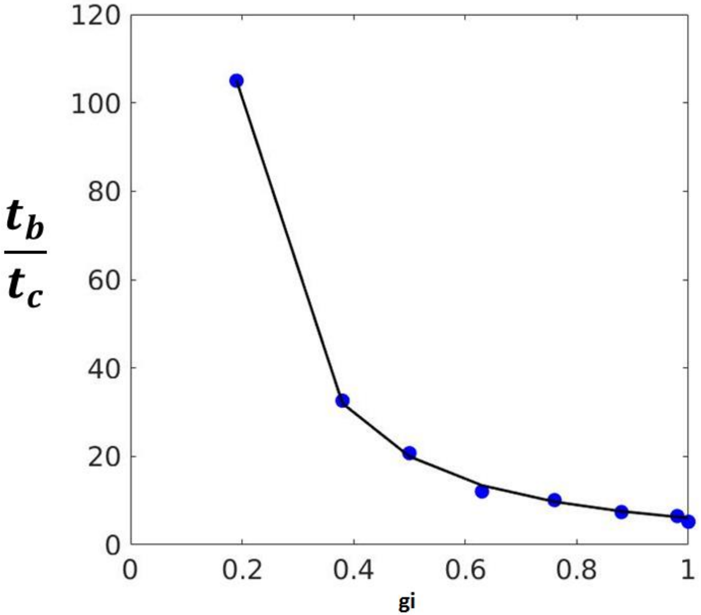
RESULTS: RELATIVE EFFICIENCY OF SAMPLING

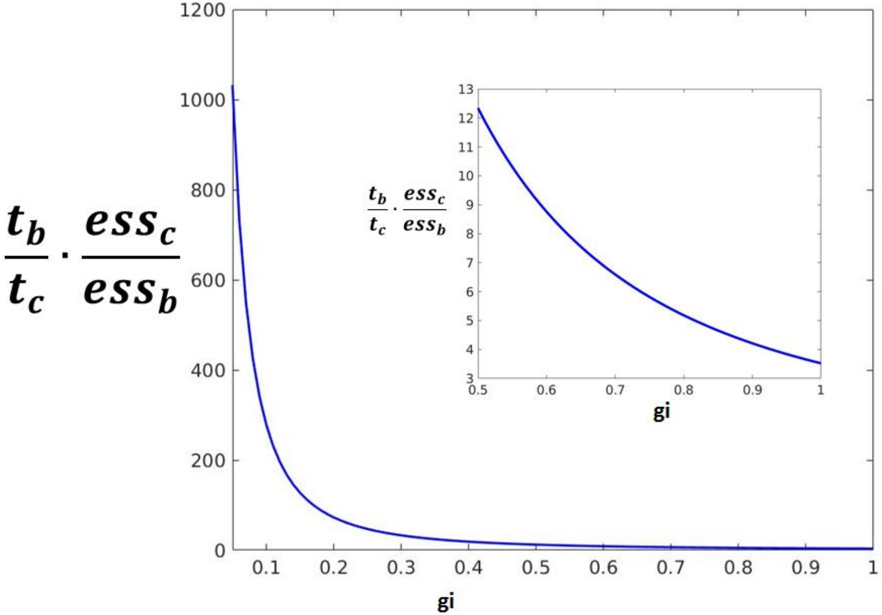


ess - effective sample size normalized by the chain size;

gi - relative amount of genomic information in variance-covariance matrix

RESULTS: COMPUTATIONAL EFFICIENCY





CONCLUSIONS

1. Sampling efficiency increases proportionally to amount of genomic information.
2. Computational efficiency is low for block update.
3. Sampling standard error decrease proportionally to increase of amount of genomic information in a model.