The Poor Man's Multi-Trait Solver

Alencar Xavier[1,2]
[1]Department of Biostatistics, Corteva Agriscience, Johnston, Iowa, United States.
[2]Department of Agronomy, Purdue University, West Lafayette, Indiana, United States.
Contact Information – alencar.xavier@corteva.com, xaviera@purdue.com

Abstract. Selections are performed on many correlated traits. However, the computation of multivariate models is prohibitive under scenarios with numerous observations, markers and traits. Pre-genomic methodologies were developed in the 80's to deal with large pedigree sets, but those have not been evaluated with genomic data. The two operations involved in the computation of multivariate genetic model are: estimation of (1) regression coefficients and (2) covariance components. This study presents an iterative solver that couples a multivariate Gauss-Seidel algorithm for estimation of marker effects with simultaneous estimation of covariance components via Tilde-Hat method. Accuracy, bias, and computation time were compared to AI-REML using simulations.

Tilde-Hat and Gauss-Seidel (THAGS)

Statistical model

- $\mathbf{y}_k = \mu_k + \mathbf{X}_k \boldsymbol{\beta}_k + \mathbf{e}_k$, for trait k
- $\boldsymbol{\beta} \sim N(0, \mathbf{I} \otimes \boldsymbol{\Sigma}_\beta)$
- $\mathbf{e} \sim N(0, \mathbf{I} \otimes \boldsymbol{\Sigma}_e)$, $\text{cov}(\mathbf{e}_i, \mathbf{e}_j) = 0$

Coefficient updates: Multivariate Gauss-Seidel

- $\boldsymbol{\beta}_j^{t+1} = \left(\mathbf{X}_j'\boldsymbol{\Sigma}_e^{-1}\mathbf{X}_j + \boldsymbol{\Sigma}_\beta^{-1}\right)^{-1}\boldsymbol{\Sigma}_e^{-1}\mathbf{X}_j'(\mathbf{X}_j\boldsymbol{\beta}_j^t + e^t)$
- $\mathbf{e}^{t+1} = \mathbf{e}^t - \mathbf{X}_j'(\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^t)$

Covariance updates: Tilde-Hat method

- $\boldsymbol{\Sigma}_{\beta(A,B)} = \frac{(\mathbf{y}_A - \hat{\mu}_A)'(\mathbf{X}_A\hat{\boldsymbol{\beta}}_B) + (\mathbf{y}_B - \hat{\mu}_B)'(\mathbf{X}_B\hat{\boldsymbol{\beta}}_A)}{n_A \sum_{j=1}^{J}\sigma^2_{x(A)_j} + n_B \sum_{j=1}^{J}\sigma^2_{x(B)_j}}$
- $\boldsymbol{\Sigma}_{e(i,i)} = \frac{(\mathbf{y}_i - \hat{\mu}_i)'\hat{\mathbf{e}}_i}{n_i - 1}$

REFERENCES

- **Tilde-Hat**: Van Raden, P. M., and Jung, Y. C. (1988). A general purpose approximation to restricted maximum likelihood: the tilde-hat approach. Journal of Dairy Science, 71(1), 187-194.
- **Gauss-Seidel**: Legarra, A., and Misztal, I. (2008). Computing strategies in genome-wide selection. Journal of Dairy Science, 91(1), 360-366.

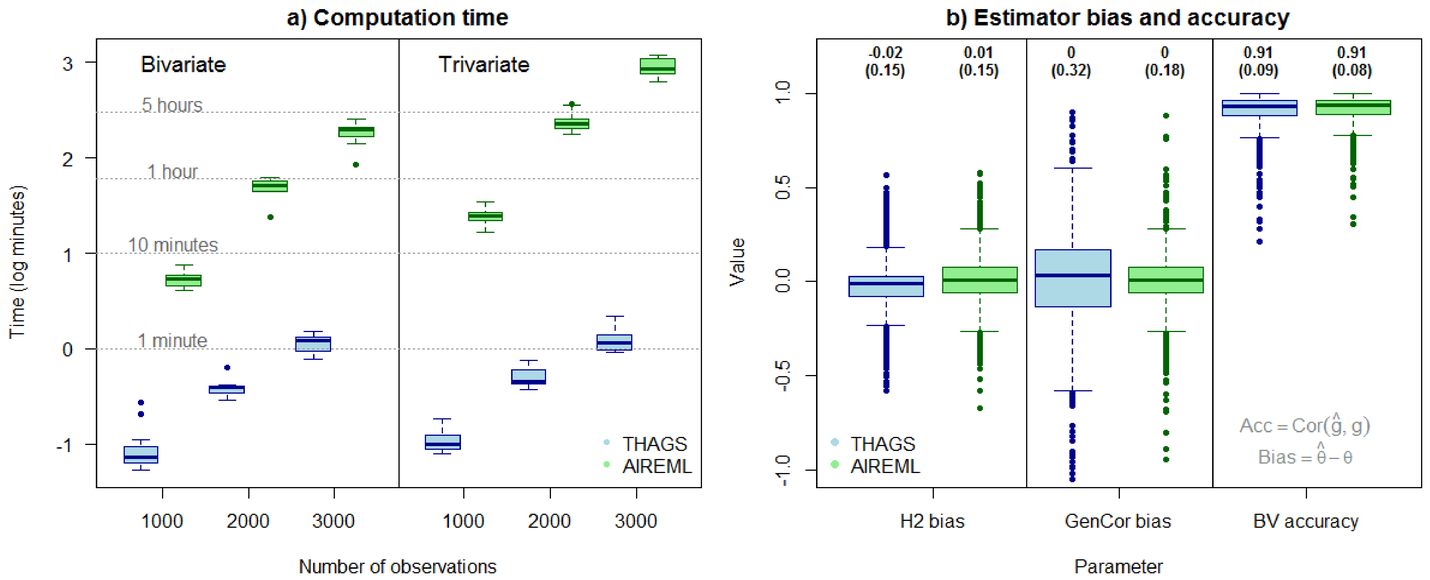

Figure 1. Box plot of computation time (**a**) estimated from multivariate GBLUP models with two or three traits and varying number of individuals comparing Tilde-Hat And Gauss-Seidel (THAGS) and Average-Information Restricted Maximum Likelihood (AIREML) fitted with AS-REML; and the estimator bias and accuracy (**b**) displaying the bias ($\hat{\theta} - \theta$) of heritability (H2) and additive genetic correlations (GenCor), and the accuracy of breeding values (BV) as $\text{Cor}(\hat{g}, g)$. Mean and standard deviation display on the top.

Figure 2. Box plot of computation time solved via Tilde-Hat And Gauss-Seidel (THAGS) from multivariate ridge regression model with varying number of traits and varying number of observations. Grain yield from the SoyNAM dataset was utilized for this graph. Computations were repeated 5x for each scenario. Lines were genotyped with 4312 SNP markers.

Estimated computation time ($R^2 = 0.99$):
$\text{Time(min)} = 0.1138 + N_{Traits} \times N_{Obs} \times 0.000136$