Bagging Strong Bayesian Learners in Genomic Prediction

XAVIER*†, A; Muir**, W.M; Rainey*, K.M.

* Department of Agronomy, Purdue University
** Department of Animal Science, Purdue University
† Contact xaviera@purdue.edu

- **Problem**
  - Need for better genomic prediction [1]
  - Computational burden [4]
  - Bayesian alphabet has "ill properties" [3]
- **Solution**
  - Resampling strategy [2]
  - Bootstrap aggregating [2]
- **Innovation**
  - Ensemble Bootstrapping and MCMC
  - Bagging Markov Chains
  - Modified GSRU [4]
- **Advantages**
  - Reduce bias
  - Increase prediction ability
  - Decrease computation time

- **Dataset**
  - 599 wheat lines; 1279 markers; 4 environments;
  - Availability: BGLR package [5]
- **Analysis**
  - Prediction accuracy and computing time
  - 5 fold cross-validation (20x)
  - Bagging BayesA (no replacement)
  - Implementation: bWGR package

HOW DOES IT WORK??

In each MCMC: Z is ψ resempled fraction of X

Regression coefficient

$$\beta_j^{t+1} \propto N\left(\frac{z_j'\tilde{e}^t + \psi x_j' x_j \beta_j^t}{\psi x_j' x_j + \lambda_j}, \frac{\sigma_\varepsilon^2}{\psi x_j' x_j + \lambda_j}\right) \text{(Alphabet)}$$

$$\beta_j^{t+1} \propto N\left(\frac{z_j^{*\prime}\tilde{e}^t + \psi \beta_j^t}{\psi + \lambda/d_j}, \frac{\sigma_\varepsilon^2}{\psi + \lambda/d_j}\right) \text{(RKHS)}$$

$$\tilde{e}^{t+1} = \tilde{e}^t + z_j(\beta_j^{t+1} - \beta_j^t)$$

Variance components

$$\sigma_{\beta_j}^2 \propto \frac{\beta_j^2 + Sv}{\chi_{v+1}^2} \text{ (BayesA and B)}$$

$$\sigma_\beta^2 \propto \frac{\beta'\beta + Sv}{\chi_{v+p}^2} \text{ (BayesC and Ridge)}$$

$$\sigma_\beta^2 \propto \frac{\beta' D^{-1}\beta + Sv}{\chi_{v+q}^2} \text{ (RKHS)}$$

$$\sigma_\varepsilon^2 \propto \frac{e'e + Sv}{\chi_{v+\psi n}^2}$$

Conjugated prior

$$S \propto \gamma\left(p + \frac{v}{2} + s, \frac{1}{2\sum\sigma_{\beta_j}^2} + r\right) \text{(BayesA and B)}$$

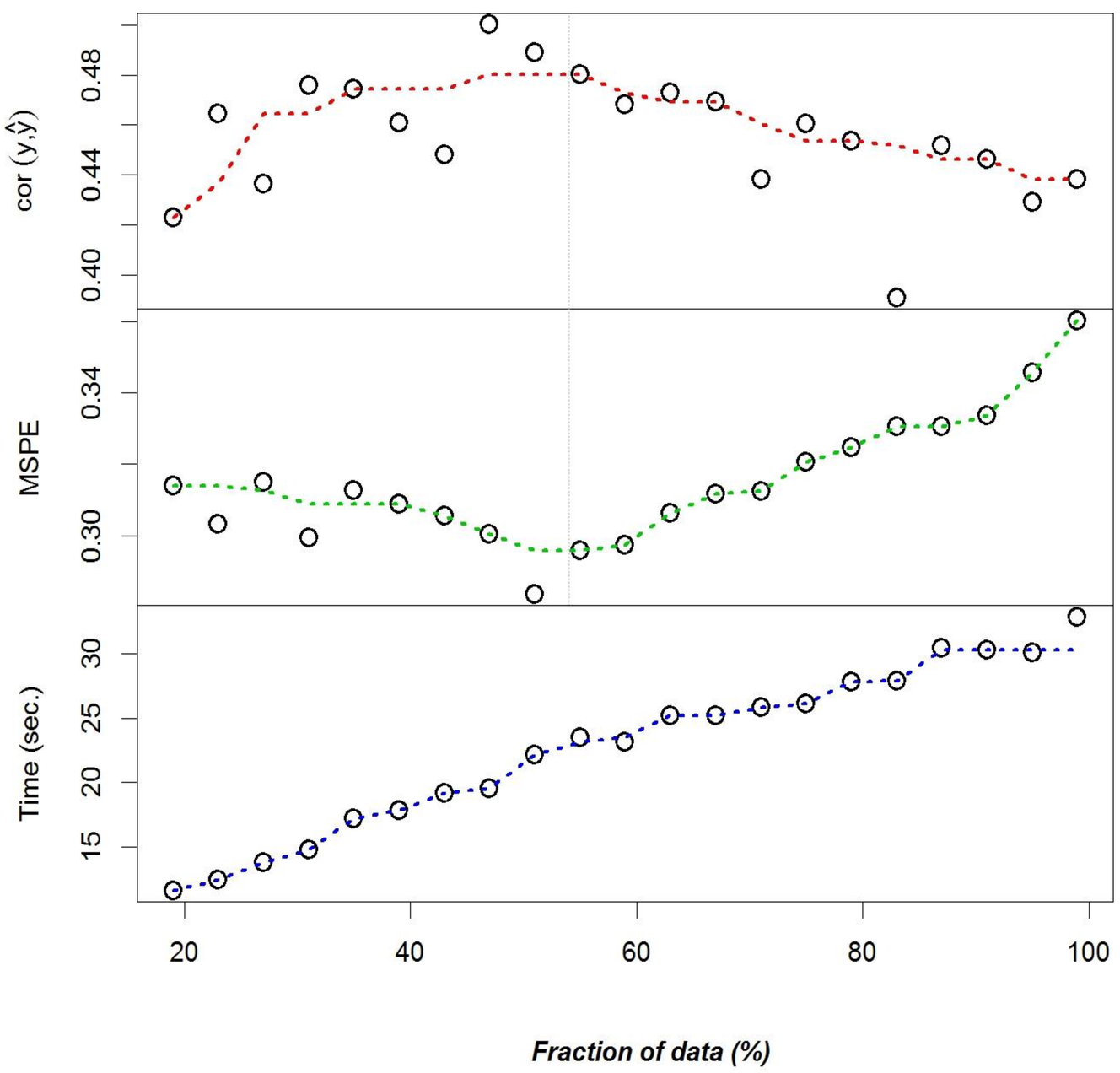

Figure 1: Effect of bootstrap aggregating (20-100%) BayesA on the correlation between predicted and observed values (*ie*. prediction accuracy), mean squared prediction error (MSPE) and computation time.

x_j - genotype of the jth marker
λ_j - regularization parameter
z_j - resampled fraction of x_j
z_j^* - resampled fraction of u_j
ψ - bagging fraction in %
u_j- jth Eigenvector of a kernel
d_j - jth Eigenvalue of a kernel
p - number of markers
n - number of observations
q - number of genotypes
e - vector of residuals
$\tilde{e}$ - e adjusted for all other markers
v - prior degrees of freedom
S - prior shape of variances
r - hyper prior rate of S
s - hyper prior shape of S

References

1. de los Campos, G., et al. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2), 327-345.
2. Gianola, D., et al. (2014). Enhancing genome-enabled prediction by bagging genomic BLUP. PloS one, 9(4). e91693.
3. Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet re-turns. Genetics, 194(3), 573-596.
4. Legarra, A., and Misztal, I. (2008). Technical note: Computing strategies in genome-wide selection. Journal of dairy science, 91(1), 360-366.
5. Pérez, P., and de los Campos, G. (2014). Genome-wide regression & prediction with the BGLR statistical package. Genetics, 198(2), 483-495.