

A study of one and two stage analyses for genomic prediction of yield in wheat

C. You E. Tanaka A. Smith B. Cullis

School of Mathematics and Applied Statistics
University of Wollongong

GRDC
Grains
Research &
Development
Corporation



SAGI

NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA



**UNIVERSITY OF
WOLLONGONG**



Aim

Aim

To investigate the loss of efficiency in genomic prediction for a two-stage approach compared with a one-stage approach.

- One-stage approach is current practice at Australian Grain Technologies (AGT) for selection of varieties.
- In contrast, most GS papers in wheat use a two-stage approach [Scutari et al., 2013, Zhao et al., 2013, Rutkoski et al., 2014, Bentley et al., 2014] whereby:
 - a basic model or design based model is fitted in first stage with some ignoring the pedigree information
 - either the BLUPs, deregressed BLUPs, BLUEs are used in second stage
 - (in some cases, the means of the response per variety is used for second stage)

Motivating Data Set

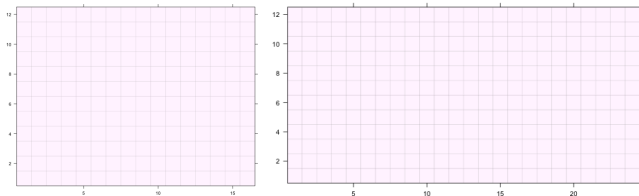


Image courtesy of AGT

An AGT breeding site testing more than 40,000 unique wheat genotypes.

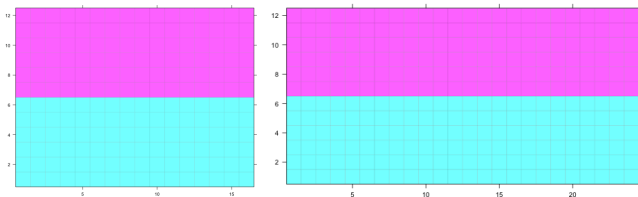
Motivating Data Set

- AGT data set contains more than 500 trials of various stages in the breeding program with a mix of Stage 1-3.
- We present results based on 48 trials that are all in stage 3 of the breeding program.
- These trials have complete marker and pedigree information for all varieties.
- Each trial is a rectangular array of 12×16 or 12×24 , i.e. a total of 192 or 288, plots.



Motivating Data Set

- Each trial consists of 138-191 varieties.
- All trials employ a design with two blocks and *either* a partial replicate [Cullis et al., 2006] or two replicates for each test lines.



- We use a total of 17,305 SNP markers.
- The missing markers were imputed using k-nearest neighbour.

Motivating Data Set

- AGT's current practice is to employ one-stage analysis using factor analytic multi-environment trial (MET) for selection of varieties with pedigree information [Smith et al., 2001, Oakey et al., 2007].
- Within trial variation is modelled using first-order separable autoregressive model denoted $AR1 \times AR1$ [Gilmour et al., 1997, Stefanova et al., 2009].
- In this talk, the results are from the analysis of single-sites.

One stage analysis

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{Z}_g((\mathbf{M}\mathbf{u}_m + \mathbf{u}_e) + \mathbf{u}_{\bar{a}}) + \mathbf{e}$$

where

- \mathbf{y} is the vector of observations
- $\boldsymbol{\tau}$ is the vector of fixed effects
- \mathbf{u}_p is the vector of random peripheral effects
- \mathbf{u}_m , \mathbf{u}_e and $\mathbf{u}_{\bar{a}}$ is the vector of marker additive, marker lack of fit, and non-additive genetic effects
- \mathbf{X} , \mathbf{Z}_p and \mathbf{Z}_g are the design matrices for fixed, random peripheral, and genetic effects respectively
- \mathbf{M} is the matrix of marker covariates
- \mathbf{e} is the residuals

One stage analysis

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{Z}_g((\mathbf{M}\mathbf{u}_m + \mathbf{u}_e) + \mathbf{u}_{\bar{a}}) + \mathbf{e}$$

and we assume

$$\begin{bmatrix} \mathbf{u}_p \\ \mathbf{u}_m \\ \mathbf{u}_e \\ \mathbf{u}_{\bar{a}} \\ \mathbf{e} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_p & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_m^2 \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_a^2 \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_a^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix} \right)$$

where

- \mathbf{A} is the pedigree numerator relationship matrix
- \mathbf{R} has the AR1 \times AR1 structure or variants of this such as AR1 \times ID

Two stage analysis: Step 1 - ① BLUPs with pedigree

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{Z}_g\mathbf{u}_e^* + \mathbf{e}$$

where

- \mathbf{y} is the vector of observations
- $\boldsymbol{\tau}$ is the vector of fixed effects
- \mathbf{u}_p is the vector of random peripheral effects
- \mathbf{u}_e^* is the genetic effects and $\mathbf{u}_e^* \sim N(\mathbf{0}, \sigma_a^2 \mathbf{A})$
- \mathbf{e} is the residuals and $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$
- \mathbf{X} , \mathbf{Z}_p and \mathbf{Z}_g are the design matrices for fixed, random peripheral, and genetic effects respectively

Two stage analysis: Step 1 - ② deregressed BLUPs with pedigree

- BLUP of \mathbf{u}_e^* , i.e. EBV, is often deregressed [Garrick et al., 2009].
- The deregression is applied as

$$\text{dEBV}_i = \tilde{u}_{e,i}^* \times \frac{1}{(\text{cor}(u_{e,i}^*, \tilde{u}_{e,i}^*))^2}$$

where $u_{e,i}^*/\tilde{u}_{e,i}^*$ is the BV/EBV of i -th variety

- The de-regressed EBV is used as response for next stage.

Two stage analysis: Step 1 - ③ BLUPs without pedigree

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{Z}_g\mathbf{u}_e^* + \mathbf{e}$$

where

- \mathbf{y} is the vector of observations
- $\boldsymbol{\tau}$ is the vector of fixed effects
- \mathbf{u}_p is the vector of random peripheral effects
- \mathbf{u}_e^* is the genetic effects and $\mathbf{u}_e^* \sim N(\mathbf{0}, \sigma_g^2 \mathbf{I})$
- \mathbf{e} is the residuals and $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$
- \mathbf{X} , \mathbf{Z}_p and \mathbf{Z}_g are the design matrices for fixed, random peripheral, and genetic effects respectively

Two stage analysis: Step 1 - ④ deregressed BLUPs without pedigree

- The deregression is applied as

$$\text{dEBV}_i = \tilde{u}_{e,i}^* \times \frac{1}{(\text{cor}(u_{e,i}^*, \tilde{u}_{e,i}^*))^2}$$

where $u_{e,i}^*/\tilde{u}_{e,i}^*$ is the BV/EBV of i -th variety

- The de-regressed EBV is used as response for next stage.

Two stage analysis: Step 1 - 5 BLUEs

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{Z}_g\boldsymbol{\tau}_g + \mathbf{e}$$

where

- \mathbf{y} is the vector of observations
- $\boldsymbol{\tau}$ is the vector of (non-genetic) fixed effects
- \mathbf{u}_p is the vector of random peripheral effects
- $\boldsymbol{\tau}_g$ is the fixed genetic effects
- \mathbf{e} is the residuals and $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
- \mathbf{X} , \mathbf{Z}_p and \mathbf{Z}_g are the design matrices for (non-genetic) fixed, peripheral random, and genetic effects respectively

Two stage analysis: Step 1 - ⑥ raw means

No model - simple average of response per variety.

Two stage analysis: Step 2

$$\tilde{\mathbf{y}} = \mathbf{1}\mu + \mathbf{M}\mathbf{u}_m^* + \boldsymbol{\epsilon}$$

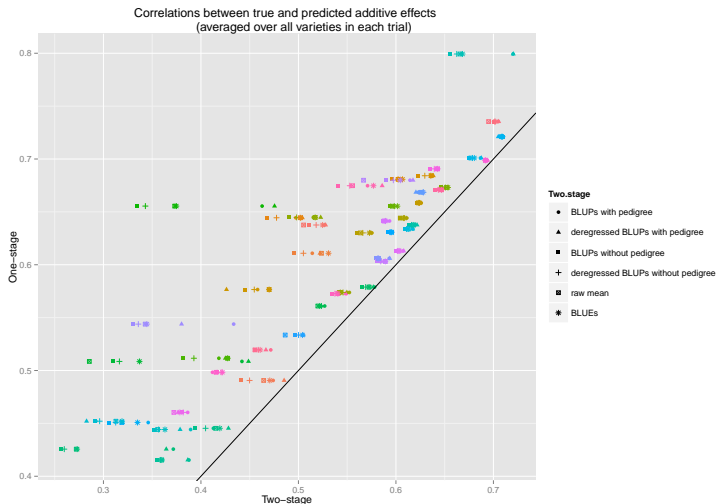
where

- $\tilde{\mathbf{y}}$ is the output from step 1
- μ is the intercept
- \mathbf{u}_m^* is the marker additive effects
- $\boldsymbol{\epsilon}$ is the residuals and we assume $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$
- \mathbf{M} is the matrix of marker covariates

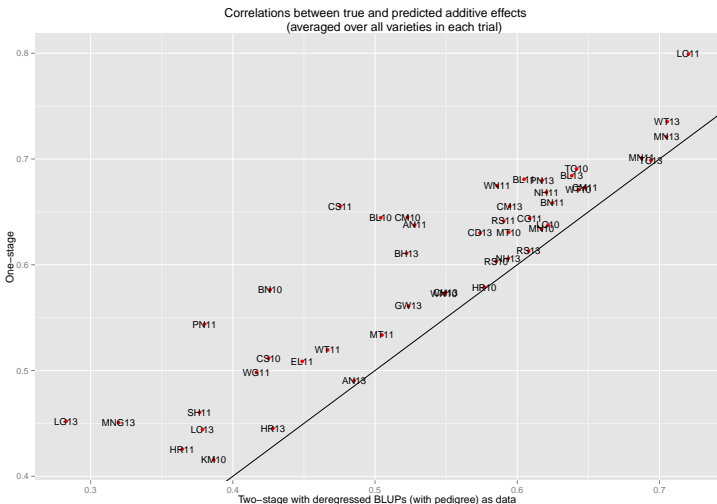
Simulation

- We simulate responses rather than apply cross validation as this way the true breeding value is known.
- For each trial we keep the original design and generate the data from the best model for that trial.
- For one stage analysis, we fit the model that matches the data generation model.
- For two stage analysis, we use the EBV from “models” proposed previously as response for next stage.
- This is repeated 200 times for each trials.
- We measure accuracy as the correlation of the predicted BV as from one-stage or two-stage analysis.

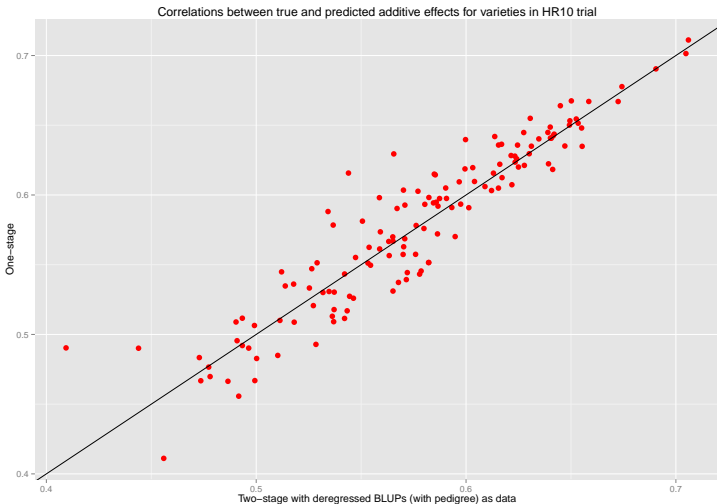
Result



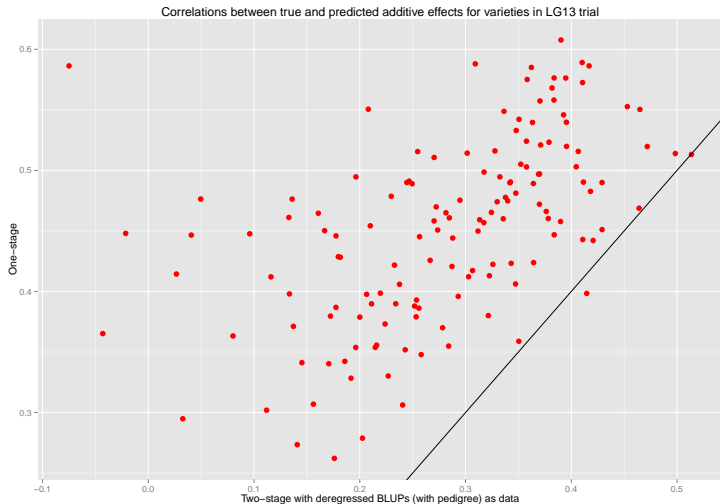
Result



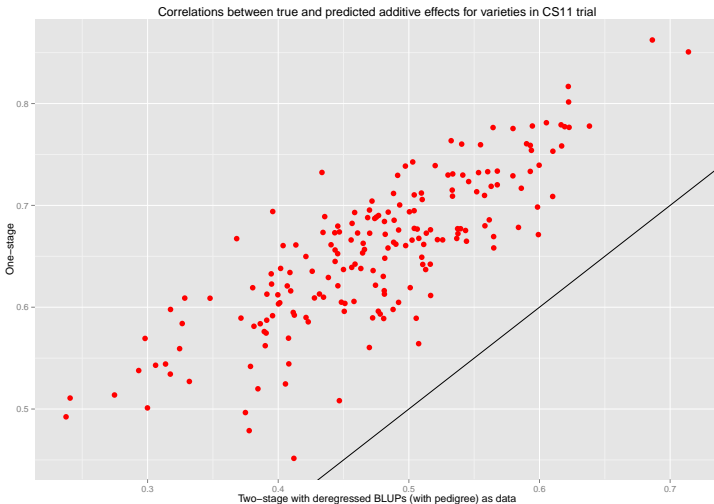
Result



Result



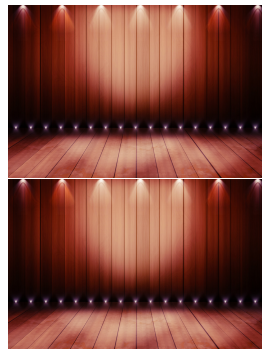
Result



Conclusion



>



- Our simulations show that the one-stage analysis has a clear advantage over two-stage analysis for prediction accuracies.

Future Research

- Incorporate a model selection process for one-stage analysis.
- Extend these results to MET to account for Genotype \times Environment.


Acknowledgement

- Dr. Julian Taylor, Dr. Haydn Kuchel and Adam Norman
- Australian Grain Technologies
- Australian Research Council for funding



Image courtesy of AGT

References I

-  Bentley, A. R., Scutari, M., Gosman, N., Faure, S., Bedford, F., Howell, P., Cockram, J., Rose, G. a., Barber, T., Irigoyen, J., Horsnell, R., Pumfrey, C., Winnie, E., Schacht, J., Beauchêne, K., Praud, S., Greenland, A., Balding, D. J., and Mackay, I. J. (2014).

Applying association mapping and genomic selection to the dissection of key traits in elite European wheat.



Theoretical and applied genetics.

-  Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006).

On the design of early generation variety trials with correlated data.

Journal of Agricultural, Biological, and Environmental Statistics, 11(4):381–393.

References II

-  Garrick, D. J., Taylor, J. F., and Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, 41(55).
-  Gilmour, A. R., Cullis, B. R., and Verbyla, A. P. (1997). Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(3):269–293.

References III



Oakey, H., Verbyla, A. P., Cullis, B. R., Wei, X., and Pitchford, W. S. (2007).

Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials.

Theoretical and Applied Genetics, 114(8):1319–1332.






Rutkoski, J., Poland, J. a., Singh, R. P., Huerta-Espino, J., Bhavani, S., Barbier, H., Rouse, M. N., Jannink, J.-L., and Sorrells, M. E. (2014).

Genomic selection for quantitative adult plant stem rust resistance in wheat.

The Plant Genome, (november):1–10.

References IV

-  Scutari, M., Mackay, I., and Balding, D. (2013). Improving the efficiency of genomic selection. *Statistical applications in genetics and molecular biology*, 12(4):517–527.
-  Smith, A. B., Cullis, B. R., and Thompson, R. (2001). Analyzing Variety by Environment Mixed Models and Adjustments Data Using Multiplicative for Spatial Field Trend. *Biometrics*, 57(4):1138–1147.
-  Stefanova, K. T., Smith, A. B., and Cullis, B. R. (2009). Enhanced diagnostics for the spatial analysis of field trials. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(4):392–410.

References V



Zhao, Y., Gowda, M., Würschum, T., Longin, C. F. H., Korzun, V., Kollers, S., Schachschneider, R., Zeng, J., Fernando, R., Dubcovsky, J., and Reif, J. C. (2013).
Dissecting the genetic architecture of frost tolerance in Central European winter wheat.
Journal of Experimental Botany, 64(14):4453–4460.