

Estimating genomic breeding values using genomic BLUP and ridge regression BLUP

Malachy Campbell

11/13/2018

Intro

$$\mathbf{y} = \mathbf{g} + \epsilon$$

- What is **g**?

Plant ID	y	g	e
1	10	5	5
2	7	6	1
3	12	2	10

Intro

- ▶ Recall that \mathbf{g} is the cumulative additive genetic effect

$$\mathbf{y} = \mathbf{1}\mu + \sum_k \mathbf{x}_k\beta + \mathbf{e}$$

\mathbf{W} is a centered $n \times m$ marker matrix, \mathbf{a} is vector of SNP effects

Ridge regression BLUP

$$\mathbf{y} = \mathbf{1}\mu + \sum_k \mathbf{x}_k\beta + \epsilon$$

- ▶ Proposed before the 'big data' trend by Meuwissen et al (2001)
- ▶ $\hat{\beta} = \mathbf{X}'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{y}$
 - ▶ $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\beta^2}$

Genomic BLUP

$$\mathbf{y} = 1\mu + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- ▶ $\hat{\mathbf{u}} = \left[\mathbf{I} + \mathbf{G}^{-1} \frac{\sigma_e^2}{\sigma_u^2} \right] \mathbf{y}$
- ▶ How do these methods differ? How will the GEBVs differ?

Equivalence

- ▶ For gBLUP the $Var(y) = \mathbf{ZGZ}'\sigma_u^2 + \mathbf{I}\sigma_e^2$
- ▶ For rrBLUP the $Var(y) = \mathbf{XX}'\sigma_\beta^2 + \mathbf{I}\sigma_e^2$
 - ▶ What does \mathbf{XX}' represent?

Demonstration with Spindel data

- ▶ 299 elite rice lines from IRRI
- ▶ genotyped with 73,147 SNPs
 - ▶ we'll use 39,560
- ▶ phenotyped for 19 traits
 - ▶ **grain yield (GY)**
 - ▶ measured in dry and wet seasons

RESEARCH ARTICLE

Genomic Selection and Association Mapping in Rice (*Oryza sativa*): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines

Jennifer Spindel¹, Hasina Begum², Deniz Akdemir¹, Parminder Virk³, Bertrand Collard⁴, Edilberto Redona⁵, Gary Aslin¹, Jean-Luc Jannink^{1,2*}, Susan R. McCouch^{1*}



Loading data

```
##Clear all objects
```

```
rm(list = ls())
```

```
# Load the data
```

```
pheno <- read.csv("~/Downloads/Spindel/pheno_WS.csv")
```

```
dim(pheno)
```

```
## [1] 299 20
```

```
geno <- read.table("~/Downloads/Spindel/Spindel_geno.txt",  
                  sep = "\\t", header = T, row.names = 1)
```

```
dim(geno)
```

```
## [1] 39560 299
```

```
geno <- t(geno)
```

```
dim(geno)
```

```
## [1] 299 39560
```

```
sum(row.names(geno) == pheno$GHID)
```

```
## [1] 299
```


Calculate a GRM

```
head(geno[,1:5])
```

```
##           S1_189590 S1_196811 S1_204765 S1_211589 S1_212693
## A1257             2           2           2           2           2
## A1258             2           2           2           2           2
## A1302             2           2           2           2           2
## B1053             2           2           2           2           2
## A1260             2           2           2           2           2
## A1304             0           0           0           0           0
```

```
Zsc <- scale(x = geno, center = T, scale = T)
GRM <- tcrossprod(Zsc)/ncol(geno)
```

```
dim(GRM)
```

```
## [1] 299 299
```

gBLUP using rrBLUP package

```
library(rrBLUP)

gBLUP <- mixed.solve(y = pheno$YLD, K = GRM)
names(gBLUP)

## [1] "Vu"    "Ve"    "beta" "u"     "LL"

length(gBLUP$u)

## [1] 299
```

rrBLUP using rrBLUP package

```
library(rrBLUP)
```

```
rrBLUP <- mixed.solve(y = pheno$YLD, Z = Zsc)  
names(rrBLUP)
```

```
## [1] "Vu"    "Ve"    "beta" "u"     "LL"
```

```
length(rrBLUP$u)
```

```
## [1] 39560
```

- ▶ Why are the sizes rrBLUP\$u and gBLUP\$u different?
- ▶ How can we make the two comparable?

Are rrBLUP and gBLUP equivalent?

► Recall

$$\hat{g} = W\hat{a}$$

#calculate GEBVs from predicted marker effects

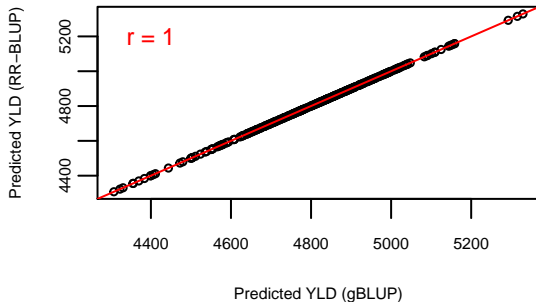
```
gBLUP_rr <- Zsc %*% rrBLUP$u
```

```
gBLUP_YLD <- gBLUP$u + as.numeric(gBLUP$beta)
```

```
gBLUP_rr_YLD <- gBLUP_rr + as.numeric(rrBLUP$beta)
```

Are rrBLUP and gBLUP equivalent?

```
par(mar=c(3,4,0.5,0.5), mgp=c(1.8,0.5,0), xpd = F, cex.lab = 0.5,  
    cex.axis = 0.5)  
plot(gBLUP_YLD, gBLUP_rr_YLD, ylab = "Predicted YLD (RR-BLUP)",  
     xlab = "Predicted YLD (gBLUP)", pch = 21, cex = 0.5)  
  
abline(lm(gBLUP_rr_YLD ~ gBLUP_YLD), col = "red")  
  
text(x = 4400, y = 5200, paste0("r = ",  
    round(cor(gBLUP_YLD, gBLUP_rr_YLD),2)), col = "red", cex = 0.75)
```



How accurate are our predictions?

- ▶ How can we estimate how accurate our predicted genomic breeding values are?

How accurate are our predictions?

- ▶ How can we estimate how accurate our predicted genomic breeding values are?
 - ▶ Compare predicted and observed breeding values for a new population
 - ▶ Partition dataset and use one for training and one for prediction

Two fold cross validation

- ▶ For some dataset
 - (1) randomly split the the individuals into two equal sized (or close to) sets
 - (2) mask the observations in one set (testing set), keep observations for other set (training set)
 - (3) fit the model using training set and predict the values for the missing individuals
 - (4) take the correlation between predicted GEBVs for test set and observed phenotypes for test set
 - (5) repeat 1 - 4

Two fold cross validation

```
pheno_train <- pheno
#define the testing and training sets
set.seed(123)
train_set <- sample(1:length(pheno$GHID), size = length(ph
test_set <- setdiff(1:length(pheno$GHID), train_set)
length(train_set)
```

```
## [1] 149
```

```
length(test_set)
```

```
## [1] 150
```

```
#Mask the phenotypes for the testing set
pheno_train[test_set,]$YLD <- NA
```

Run RRBLUP with training set

```
library(rrBLUP)
##rrBLUP
rrBLUP_train <- mixed.solve(y = pheno_train$YLD, Z = Zsc)
rrBLUP_train <- Zsc %*% rrBLUP_train$u
length(rrBLUP_train)
```

```
## [1] 299
```

Assess predictive ability from rrBLUP approaches

```
rrBLUP_test <- rrBLUP_train[test_set]
pheno_test <- pheno[test_set ,]

cor(pheno_test$YLD, rrBLUP_test)

## [1] 0.1618154
```

References

- ▶ Endelman, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255 (2011).
- ▶ Habier, D., Fernando, R. L. & Dekkers, J. C. M. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397 (2007).
- ▶ Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829 (2001).
- ▶ Spindel, J. et al. Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* 11, e1004982 (2015).