Statistical Analysis of Genetic and Phenotypic Data for Breeders: Hands on Practical Sessions (GBLUP-RR)

Paulino Pérez ¹
José Crossa ²

¹ColPos-México ²CIMMYT-México

June, 2015.
Contents

1. General comments
2. GBLUP-Ridge Regression
3. Application examples
4. Biplot from marker effects
5. Extension of BRR to include infinitesimal effect
Remember,

1. A simple model used frequently in plant breeding stands that the phenotypic value of an individual ($P$) is expressed as the summation of the genetic value ($G$) and the residual environmental effect ($E$):

$$ P = G + E, \quad (1) $$

where $G$ includes additive, dominance and epistatic effects.

2. A model that includes solely additive effects ($A$) can be easily derived from (1), and can be expressed as follows,

$$ P = A + E' \quad (2) $$

where $E'$ includes effects that are non additive.
The breeding value ($BV$) for an individual can be computed based on narrow sense heritability ($h^2$),

$$BV_i = \mu + h^2(y_i - \mu),$$

where $\mu$ is mean phenotypic value of a population and $y_i$ is the phenotypic value for individual $i$. Obviously it is necessary to have information of parents and offsprings to compute this.
In Genomic Selection (GS), genetic values are approximated using linear regression (Meuwissen et al., 2001), that is:

$$y_i = g_i + e_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + e_i$$  \hspace{1cm} (3)

Relationships between marker genotypes ($x_{1i}$: 0 and 1) and phenotypes ($y_i$) of the individuals (open circles) in a training population. If the marker genotype is correlated with the phenotype, segregation is modelled using the bold line (taken from Nakaya and Isobe, 2012).
In GS it is possible to obtain **Genomic Estimated Breeding Values** (GEBVs for short). This can be done simply by adding up marker effects (according to its marker genotypes) obtained from a training population, that is:

\[
GEBV_i = \sum_{j=1}^{p} x_{ij} \hat{\beta}_j
\]  

Next we show how to obtain the predictions \( \hat{y}_i \) (and in some cases \( \hat{\beta}_j \)) using several models.
In this presentation we will focus in Ridge Regression.

Figure 1: Graphical representation of parametric and non-parametric methods used commonly in whole-genomic prediction.
Figure 2: Prior densities of regression coefficients with Markers.
This is the most basic model used in GS. Let

\[ y_i = g_i + e_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + e_i \]

marker effects are obtained by solving the following optimization problem,

\[
\min_{\beta, \lambda} \left\{ (y - \sum X_j \beta_j)'(y - \sum X_j \beta_j) + \lambda \sum \beta_j^2 \right\},
\]

(5)

where \( \lambda > 0 \) is a regularization parameter.

Notes:

1. \( \lambda \) is unknown and can be selected by using cross-validation
2. we need to minimize a “penalized sum of squares”.

CIMMYT, México-SAGPDB  Statistical Analysis of Genetic and Phenotypic Data for Breeders: Hands (9/34)
The optimization problem has a closed solution,

\[ \hat{\beta} = (X'X + \lambda I)^{-1} X'\tilde{y}, \]

where \( \tilde{y} = y - \mu 1 \). Unfortunately, we need to know the value of \( \lambda \) to use this solution.

The problem can be solved easily using the Bayesian framework.

Let \( \beta \sim N(0, \sigma_\beta^2 I) \) and \( e \sim N(0, \sigma_e^2 I) \), and \( u = X\beta \), then model (3) can be written as:

\[ y = \mu 1 + u + e \quad (6) \]

Model (6) is known as GBLUP. Note that \( u \sim N(0, \sigma_\beta^2 XX') \)
Note also that the covariance matrix for $u$ involves the product $XX'$, which is proportional to the **Genomic Relationship Matrix** proposed by VanRaden (2008).

We will assume that $u \sim N(0, \sigma_u^2 G)$ with $G = XX' / k$. The mix-model equations for (6) are as follows:

$$
\begin{pmatrix}
1' \sigma_e^{-2} & 1' \sigma_u^{-2} \\
1' \sigma_e^{-2} & I \sigma_e^{-2} \sigma_u^{-2} + G \sigma_u^{-2}
\end{pmatrix}
\begin{pmatrix}
\hat{\mu} \\
\hat{u}
\end{pmatrix}
= 
\begin{pmatrix}
1' y \\
y
\end{pmatrix} \tag{7}
$$

$u$ and $\mu$ are obtained solving the mix-model equations, assuming that the variance components $\sigma_e^2$ and $\sigma_u^2$ are known.
If we have individuals for training and testing, we can partition $G$ and $u$ as follows,

$$
G = \begin{pmatrix}
G_{11} & G_{12} \\
G_{21} & G_{22}
\end{pmatrix},
\quad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},
\quad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},
\quad 1 = \begin{pmatrix} 1_1 \\ 1_2 \end{pmatrix}
$$

1=individuals in the training set, 2=individuals in the testing set. $\hat{\mu}$ and $\hat{u}_1$ are obtained as the solution of the mix-model equations,

$$
\begin{pmatrix}
1_1' & 1_1\sigma_e^{-2} \\
1_1' & 1_1\sigma_e^{-2}
\end{pmatrix}
\begin{pmatrix}
1_1' & 1_1\sigma_e^{-2} \\
1_1\sigma_e^{-2} & I_{11}\sigma_e^{-2}\sigma_u^{-2} + G_{11}\sigma_u^{-2}
\end{pmatrix}
\begin{pmatrix}
\hat{\mu} \\
\hat{u}_1
\end{pmatrix}
= \begin{pmatrix} 1_1' y_1 \\ y_1 \end{pmatrix}
$$

The predictions for individuals in the testing set are given by

$$
\hat{y}_2 = \hat{\mu}1_2 + G_{21}G_{11}^{-1}\hat{u}_1
$$
Wheat dataset

Data for \( n = 599 \) wheat lines evaluated in 4 environments, wheat improvement program, CIMMyT. The dataset includes \( p = 1279 \) molecular markers \((x_{ij}, i = 1, ..., n, j = 1, ..., p)\) (coded as 0,1). The pedigree information is also available.

Let's load the dataset in R,

1. Load R
2. Install BGLR package (if not yet installed)
3. Load the package
4. Load the data
Figure 3: Loading the BGLR package and the wheat dataset.
You can explore the MM matrix, pedigree matrix within R,

- `fix(wheat.X)`
- `fix(wheat.A)`
Lets assume that we want to predict the grain yield for environment 1 using **ridge regression** or equivalently the GBLUP. We do not know the value for $\sigma^2_e$ and $\lambda$, so we can obtain estimates using the data.

We will use the function BGLR. R code below fit the RR model using Bayesian approach with non informative priors for $\sigma^2_e$, $\sigma^2_\beta$.

```r
rm(list=ls())
library(BGLR)
data(wheat)

X=wheat.X
Y=wheat.Y

#Linear predictor
ETA=list(list(X=X,model="BRR"))

#Or
#ETA=list(Markers=list(X=X,model="BRR"))

fmR<-BGLR(y=Y[,1],ETA=ETA,nIter=10000,burnIn=5000,thin=10)
plot(fmR$yHat,Y[,1])
```

CIMMYT, México-SAGPDB  Statistical Analysis of Genetic and Phenotypic Data for Breeders:Hands ( 16/34)
Figure shows observed vs predicted grain yield.

Predictions $\hat{y} = \hat{\mu} + X\hat{\beta}$, and estimates for $\sigma_e^2$, $\sigma_\beta^2$ can be obtained easily in R

```r
> fmR$yHat
> fmR$varE
[1] 0.5481523
> fmR$ETA[[1]]$varB
[1,] 0.002721897
```

From the output above,

$$\hat{\lambda} = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_\beta^2} = \frac{0.5482}{0.0027} = 201.38$$
A complete list of the objects attached to fmR can be obtained by typing

```r
> names(fmR)
   [1] "y"       "whichNa"  "saveAt"  "nIter"  "burnIn" 
   [8] "verbose" "response_type" "df0"     "S0"      "yHat" 
  [15] "varE"    "SD.varE"  "fit"      "ETA" 
```

A complete description of the output can be found in the BGLR reference manual.

The GEBVs can be obtained easily in R,

```r
#GEVBs

#option 1
X%*%fmR$ETA[[1]]$b

#option 2
fmR$yHat-fmR$mu

#Or fmR$ETA$Markers$b
```
Let's assume that we want to predict the grain yield for some wheat lines. Assume that we have only the genotypic information for those lines,

```r
# Training and testing set
sets<- wheat.sets
y<- Y[,1]
yNa=y
whichNa=(sets==2)
yNa[whichNa]=NA
fmR<- BGLR(y=yNa, ETA=ETA, nIter=10000,
          burnIn=5000, thin=10)
```

```r
plot(fmR$yHat,y,xlab="Phenotype",
     ylab="Pred. Gen. Value", cex=.8, bty="L")
points(x=y[whichNa], y=fmR$yHat[whichNa], col=2, cex=.8, pch=19)
legend("topleft", legend=c("training","testing"), bty="n",
        pch=c(1,19), col=c("black","red"))
```
> MSE.tst <- mean((fmR$yHat[whichNa] - y[whichNa])^2)
> MSE.tst
[1] 0.8110028
> MSE.trn <- mean((fmR$yHat[-whichNa] - y[-whichNa])^2)
> MSE.trn
[1] 0.4364856
> COR.tst <- cor(fmR$yHat[whichNa], y[whichNa])
> COR.tst
[1] 0.4338218
> COR.trn <- cor(fmR$yHat[-whichNa], y[-whichNa])
> COR.trn
[1] 0.7839615
Questions?
Exercise

Suppose that we want to predict the grain Yields for individuals in set 2 and environment 4.

- Write an R program to solve the problem described above
- Obtain the correlations in the training set
- Obtain the correlations in the testing set
- Write the predictions to a csv (comma separated values) so that you can read the file in Excel.
Suppose that we are interested in studying the predictive power of GBLUP. We can perform a simulation study to that end. In this exercise you will perform 10 fold cross validation.

- Write an R program to perform a 10 fold cross-validation, use the use object `sets` to allocate observations to folds.
- Report your results.
A biplot is a two-dimensional representation of a data matrix $C$ showing a point for each of the $n$ observation vectors (rows of $C$) along with a point for each of the $p$ variables (columns of $C$), see Gabriel, 1971.

Perform the SVD of $C$, that is $C = UDV'$

Let $U_{p \times q} = [\alpha_1, ..., \alpha_q]$ and $V_{q \times q} = [\gamma_1, ..., \gamma_q]$.

Plot $\alpha_1$ vs $\alpha_2$

Draw arrows, the coordinates of the end of the arrow are given in $\gamma_1$ and $\gamma_2$. 
Figure 4: An example of a Biplot derived from marker effects.
Why is the biplot useful?

1. Points in the biplot are the marker effects projected in the first two components.

2. The “environmental effects” are displayed as vectors whose coordinates are given by $\gamma_1$ and $\gamma_2$.

3. The length of the vectors approximates the variance accounted for by the specific molecular marker and “environmental effect”.

4. The cosine of the angle between two environments, approximates the correlation of the two environments with an angle of zero indicating a correlation of $+1$, an angle of $90^\circ$ (or $-90^\circ$) a correlation of $0$, and an angle of $180^\circ$ a correlation of $-1$. 
```r
rm(list=ls())

# Set the working directory
setwd("C:/Users/P.P.RODRIGUEZ/Desktop/Slides Paulino/3. GBLUP-RR/examples")

# Function for biplots
source("biplot.R")

# Import the data
data=read.csv("mean_betas_PMBL.csv",header=TRUE)
data=data[,2:5]

# Principal component analysis
pca.betas= princomp(data,cor=TRUE)

# Default biplot in R
biplot(pca.betas)

# Modified function
my.biplot.princomp(pca.betas)
```
Exercise

1. Use the wheat dataset described in the previous slides and fit a GBLUP-BRR model and save the marker effects for the 4 environments to an Excel file (csv).

2. Use the marker effects obtained in Step 1, and create a biplot.
de los Campos et al. (2009) extended the basic BRR model to include an infinitesimal effect, that is:

\[ y_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + u_i + e_i, \]  

\[ (8) \]

where \( u \sim N(0, \sigma_u^2 A) \) and \( A \) is the pedigree matrix.

The model can be fitted using the standard linear mixed model theory or using Bayesian methods.
Extension of BRR to include infinitesimal effect

Example

```r
rm(list=ls())
library(BGLR)
data(wheat)

X=wheat.X
Y=wheat.Y
A=wheat.A

#Linear predictor
ETA=list(list(X=X,model="BRR"),
          list(K=A,model="RKHS"))

fmR<-BGLR(y=Y[,1],ETA=ETA,nIter=10000,burnIn=5000,thin=10)
plot(fmR$yHat,Y[,1])
```
Exercise

Suppose that we want to predict the grain Yields for individuals in set 2 and environment 1 using the marker and pedigree information jointly.

- Write an R program to solve the problem described above
- Obtain the correlations in the training set
- Obtain the correlations in the testing set
- Write the predictions to a csv (comma separated values) so that you can read the file in Excel.
How can we design a simulation to study prediction ability of a model with markers only and a model that includes Markers + Pedigree?
Concluding remarks
References

