

Supporting Material for: “Best Prediction of the Additive Genomic Variance in Random-Effects Models”

Nicholas Schreck, Hans-Peter Piepho, and Martin Schlather

9. Mai 2019

1 Estimation of the Unconditional Expectation and the Best Predictor of the Additive Genomic Variance in the “Equivalent Linear Model”

We review the formulae for the estimated unconditional expectation as well as the empirical best predictor of the additive genomic variance introduced in the paper, and provide code for an implementation using the free software *R*.

We focus on the linear model

$$y = \mu \mathbf{1}_n + g + \varepsilon,$$

for the phenotypic values y where μ denotes a fixed intercept, $\mathbf{1}_n := (1, \dots, 1)^\top$ is an n -row-vector, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}_n)$, and $\mathbf{I}_n$ is the n -dimensional identity matrix.

In the random-effects model, the mean-centered ($\bar{g} := \frac{1}{n} \mathbf{1}_n^\top g = 0$) genomic values g are distributed as

$$g \sim \mathcal{N}(0, \sigma_g^2 \mathbf{G}).$$

When using genomic marker data, the genomic relationship matrix $\mathbf{G}$ can be defined as

$$\mathbf{G} := \mathbf{P} \mathbf{X} \mathbf{X}^\top \mathbf{P} / c \quad (1)$$

where $c > 0$ is a constant, $\mathbf{X}$ the marker genotypes matrix and

$$\mathbf{P} := \mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n^\top / n \quad (2)$$

carries out the column-wise mean-centering of the marker genotypes. Let σ_β^2 be the variance component of the marker effects β . Then, the variance component in the equivalent model equals $\sigma_g^2 = c \sigma_\beta^2$ and the constant c is usually defined as

$$c := 2 \sum p_j (1 - p_j), \quad (3)$$

where p_j is the frequency of the minor allele at marker j . We use the sample variance of the phenotypic values

$$\hat{\sigma}_y^2 := \frac{1}{n-1} y^\top \mathbf{P} y \quad (4)$$

to scale the phenotypic values to unit variance by setting

$$\tilde{y} = \frac{y}{\hat{\sigma}_y}. \quad (5)$$

We fit the model using the *R*-package “sommer” which returns the estimated variance components $\hat{\sigma}_g^2$ and $\hat{\sigma}_\varepsilon^2$ (REML approach). In addition, we obtain the empirical BLUP (eBLUP)

$$\hat{\mu}_{g|y} = \hat{\sigma}_g^2 \mathbf{G} (\mathbf{G} \hat{\sigma}_g^2 + \hat{\sigma}_\varepsilon^2 \mathbf{I}_n)^{-1} (y - \hat{\mu} \mathbf{1}_n) \quad (6)$$

for the random vector g and the estimated variance-covariance matrix

$$\begin{aligned} \hat{\Sigma}_{\hat{\mu}_{g|y}} = & \hat{\sigma}_g^2 \mathbf{G} (\mathbf{G} \hat{\sigma}_g^2 + \hat{\sigma}_\varepsilon^2 \mathbf{I}_n)^{-1} \mathbf{G} \hat{\sigma}_g^2 \\ & - \frac{\hat{\sigma}_g^2 \mathbf{G} (\mathbf{G} \hat{\sigma}_g^2 + \hat{\sigma}_\varepsilon^2 \mathbf{I}_n)^{-1} \mathbf{1}_n \mathbf{1}_n^\top (\mathbf{G} \hat{\sigma}_g^2 + \hat{\sigma}_\varepsilon^2 \mathbf{I}_n)^{-1} \mathbf{G} \hat{\sigma}_g^2}{\mathbf{1}_n^\top (\mathbf{G} \hat{\sigma}_g^2 + \hat{\sigma}_\varepsilon^2 \mathbf{I}_n)^{-1} \mathbf{1}_n} \end{aligned} \quad (7)$$

of the eBLUP $\hat{\mu}_{g|y}$ of g .

When interest is in the current population, we have to set

$$\mathbf{B}_c := \mathbf{I}_n,$$

and obtain the estimated unconditional expectation

$$\hat{V} = \frac{1}{n-1} \hat{\sigma}_g^2 \text{tr}(\mathbf{G}) \quad (8)$$

and the empirical best predictor (eBP)

$$\widehat{W} = \hat{V} + \frac{1}{n-1} \hat{\mu}_{g|y}^\top \hat{\mu}_{g|y} - \frac{1}{n-1} \text{tr}(\hat{\Sigma}_{\hat{\mu}_{g|y}}) \quad (9)$$

for the additive genomic variance in the current population.

We define estimators for the heritabilities in the current population by dividing the estimated unconditional expectation

$$\hat{h}_V^2 := \frac{\hat{V}}{\hat{\sigma}_y^2} = \hat{V} \quad (10)$$

as well as the eBP

$$\hat{h}_W^2 := \frac{\widehat{W}}{\hat{\sigma}_y^2} = \widehat{W}. \quad (11)$$

of the additive genomic variance by the phenotypic sample variance.

Alternatively, we can assume that the phenotypic variance equals the sum of the estimated genomic samples variances in the current population and the estimated residual variance:

$$\tilde{\sigma}_{y,V}^2 := \hat{V} + \hat{\sigma}_\varepsilon^2 \quad (12)$$

and

$$\tilde{\sigma}_{y,W}^2 := \widehat{W} + \hat{\sigma}_\varepsilon^2. \quad (13)$$

These estimators for the phenotypic variance depend on the specific choice of the estimator for the genomic sample variance and lead to alternative heritability estimates:

$$\tilde{h}_V^2 := \frac{\hat{V}}{\tilde{\sigma}_{y,V}^2} \quad (14)$$

and

$$\tilde{h}_W^2 := \frac{\widehat{W}}{\tilde{\sigma}_{y,W}^2}. \quad (15)$$

When interest is in the base population, we have to set

$$\mathbf{B}_b := \mathbf{R}^{-0.5} \mathbf{P} \mathbf{R}^{-0.5}, \quad (16)$$

and we obtain the estimated unconditional expectation

$$\hat{V}^* = \frac{1}{n-1} \hat{\sigma}_g^2 \text{tr}(\mathbf{R}^{-0.5} \mathbf{P} \mathbf{R}^{-0.5} \mathbf{G}) \quad (17)$$

and the eBP

$$\widehat{W}^* = \hat{V}^* + \frac{1}{n-1} \hat{\mu}_{g|y}^\top \mathbf{R}^{-0.5} \mathbf{P} \mathbf{R}^{-0.5} \hat{\mu}_{g|y} - \frac{1}{n-1} \text{tr}(\mathbf{R}^{-0.5} \mathbf{P} \mathbf{R}^{-0.5} \hat{\Sigma}_{\hat{\mu}_{g|y}}). \quad (18)$$

for the additive genomic variance in the base population.

The GRM $\mathbf{G}$ is singular and therefore $\mathbf{G}^{-0.5}$ is not well-defined. However, it is possible to decompose $\mathbf{G}$ into

$$\mathbf{G} = \mathbf{G}^{0.5} \mathbf{G}^{0.5}.$$

Then, we directly set the unconditional expectation of the additive genomic variance to

$$\hat{V}_s^* = \hat{\sigma}_g^2 \quad (19)$$

instead of calculating

$$\frac{1}{n-1} \hat{\sigma}_g^2 \text{tr}(\mathbf{P} \mathbf{G}^{-0.5} \mathbf{G} \mathbf{G}^{-0.5}) = \frac{1}{n-1} \hat{\sigma}_g^2 \text{tr}(\mathbf{P}) = \hat{\sigma}_g^2$$

when using the GRM for the transformation to the base population in (17).

We proceed analogously for the eBP. Instead of calculating

$$\mathbf{G}^{-0.5} \hat{\mu}_{g|y} = \mathbf{G}^{-0.5} \hat{\sigma}_g^2 \mathbf{G} (\mathbf{G} \hat{\sigma}_g^2 + \hat{\sigma}_\varepsilon^2 \mathbf{I}_n)^{-1} (y - \hat{\mu} \mathbf{1}_n)$$

and

$$\mathbf{G}^{-0.5} \hat{\Sigma}_{\hat{\mu}_{g|y}} = \mathbf{G}^{-0.5} \left[\hat{\sigma}_g^2 \mathbf{G} \hat{\Sigma} \mathbf{G} \hat{\sigma}_g^2 - \frac{\hat{\sigma}_g^2 \mathbf{G} \hat{\Sigma} \mathbf{1}_n \mathbf{1}_n^\top \hat{\Sigma} \mathbf{G} \hat{\sigma}_g^2}{\mathbf{1}_n^\top \hat{\Sigma} \mathbf{1}_n} \right] \mathbf{G}^{-0.5}$$

in (18), we use

$$\hat{\mu}_{g|y}^* := \hat{\sigma}_g^2 \mathbf{G}^{0.5} (\mathbf{G} \hat{\sigma}_g^2 + \hat{\sigma}_\varepsilon^2 \mathbf{I}_n)^{-1} (y - \hat{\mu} \mathbf{1}_n) \quad (20)$$

and

$$\hat{\Sigma}_{\hat{\mu}_{g|y}}^* := \hat{\sigma}_g^2 \mathbf{G}^{0.5} \hat{\Sigma} \mathbf{G}^{0.5} \hat{\sigma}_g^2 - \frac{\hat{\sigma}_g^2 \mathbf{G}^{0.5} \hat{\Sigma} \mathbf{1}_n \mathbf{1}_n^\top \hat{\Sigma} \mathbf{G}^{0.5} \hat{\sigma}_g^2}{\mathbf{1}_n^\top \hat{\Sigma} \mathbf{1}_n} \quad (21)$$

as surrogates. Then, we define

$$\widehat{W}_s^* = \hat{V}_s^* + \frac{1}{n-1} (\hat{\mu}_{g|y}^*)^\top \mathbf{P} \hat{\mu}_{g|y}^* - \frac{1}{n-1} \text{tr}(\mathbf{P} \hat{\Sigma}_{\hat{\mu}_{g|y}}^*) \quad (22)$$

for the eBP of the additive genomic variance in the base population when using the GRM for the transformation.

2 R-Script for the Estimation of the Additive Genomic Variance

We implement the quantities defined above using the following *R*-code.

```
1 #####
2 ### Estimation of the Genomic Variance in the ‘‘Equivalent Linear Model’’ #####
3 #####
4
5 # Author
6 # Nicholas Schreck, nschreck@mail.uni-mannheim.de
7
8 # Estimation of the Genomic Variance in the "Equivalent Linear Model"
9
10 # Copyright (C) 2018 -- 2019 Nicholas Schreck
11
12 # This program is free software; you can redistribute it and/or
13 # modify it under the terms of the GNU General Public License
14 # as published by the Free Software Foundation; either version 2
15 # of the License, or (at your option) any later version.
16
17 # This program is distributed in the hope that it will be useful,
18 # but WITHOUT ANY WARRANTY; without even the implied warranty of
19 # MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
20 # GNU General Public License for more details.
21
22 # You should have received a copy of the GNU General Public License
23 # along with this program; if not, write to the Free Software
24 # Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.
25
26 #####
27 # Packages
28 library(sommer)
29 library(BGLR)
30 #####
31 # Genomic data (phenotypes and marker genotypes)
32 data(mice)
33 data(wheat)
34 load("Arabidopsis1001genomes_1057lines_193697SNPs.Rdata")
35 y.all <- list(mice.pheno$Obesity.BodyLength, wheat.Y[, 1],
36              phenogeno$pheno$FT10_mean) # Phenotypic values
37 Ped.all <- list(mice.A, wheat.A, NA) # Pedigree matrices
38 X.all <- list(mice.X, wheat.X, phenogeno$geno) # marker-genotyp matrices
39 nr.datasets <- length(y.all)
40 #####
41 # Output Variables
42 Variables <- c("eps", "V_hat", "h2_V", "sum_V", "h2_V_sum", "W_hat", "h2_W",
43               "sum_W", "h2_Wsum", "Vstar_hat", "Wstar_hat", "Vstars_hat",
44               "Wstars_hat")
45 nr.variables <- length(Variables)
46 result <- matrix(NA, nr.variables, nr.datasets)
47 rownames(result) <- Variables
48 colnames(result) <- c("Mice", "Wheat", "Arabidopsis")
49 #####
50 rm(mice.A, mice.pheno, mice.X, wheat.A, wheat.X, wheat.Y, phenogeno, wheat.sets,
51     Variables, nr.variables)
52
53 #####
```

```

54 #####
55 # Model fit and estimation of genomic variances
56 for (j in 1 : nr.datasets){ # for each dataset
57   # Preliminaries
58   y <- y.all[[j]] # fix phenotype
59   y.scaled <- y / sd(y) # y scaled to variance of 1, see equation (5)
60   n <- length(y)
61   ones <- matrix(1, nrow=n, ncol=1)
62   P <- diag(1, n, n) - matrix(1, n, n) / n # Equation (2)
63   X <- X.all[[j]] # fix marker-genotypes
64   p <- dim(X)[2]
65   c <- sum( colMeans(X) * (1 - colMeans(X) / 2) ) # vanRaden c, see equation (3)
66   G <- P %*% tcrossprod(X) %*% P / c # equation (1)
67   rm(X)
68   #####
69   # Fit gBLUP model
70   lm.equi <- mmer(Y=y.scaled, X=ones, Z=list(A=list(K=G)), silent=TRUE,
71                 iters=50, tolpar=1e-7, tolparinv=1e-9, date.warning = FALSE)
72   var.g.hat <- as.numeric(lm.equi$var.comp$A) # variance component g
73   var.e.hat <- as.numeric(lm.equi$var.comp$units) # variance comp. eps
74   mu.hat <- rep(lm.equi$beta.hat, n) # Estimate of Intercept
75   g.hat <- as.vector(lm.equi$u.hat$A) # eBLUP, equation (6)
76   cov.g.hat <- as.matrix(lm.equi$Var.u.hat$A$T1) # Cov of eBLUP, equation (7)
77   rm(lm.equi)
78   # Genomic variances in the current population
79   gv_cur <- GenVarCur(GRM=G, varg=var.g.hat, eBLUP=g.hat, eBLUPcov=cov.g.hat)
80   V.hat <- gv_cur[1]
81   h2.V <- V.hat # formula (10)
82   sum.V <- V.hat + var.e.hat # formula (12)
83   h2.V.sum <- V.hat / sum.V # formula (14)
84   W.hat <- gv_cur[2]
85   h2.W <- W.hat # (11)
86   sum.W <- W.hat + var.e.hat # formula (13)
87   h2.W.sum <- W.hat / sum.W # formula (15)
88   # Genomic variances in the base population (via Pedigree)
89   Ped <- Ped.all[[j]]
90   if(!is.na(Ped)){
91     gv_base_ped <- GenVarBase_Ped(PedMat=Ped, GRM=G, varg=var.g.hat, eBLUP=g.hat,
92                                   eBLUPcov=cov.g.hat)
93     V.hat.star <- gv_base_ped[1]
94     W.hat.star <- gv_base_ped[2]
95   }else{
96     V.hat.star <- W.hat.star <- NA
97   }
98   # Genomic variances in the base population (via GRM)
99   gv_base_grm <- GenVarBase_GRM(GRM=G, phenos=y.scaled, mu=mu.hat,
100                                varg=var.g.hat, eBLUP=g.hat, eBLUPcov=cov.g.hat,
101                                vare=var.e.hat)
102   V.hat.star.s <- gv_base_grm[1]
103   W.hat.star.s <- gv_base_grm[2]
104   #####
105   result[, j] <- c(var.e.hat, V.hat, h2.V, sum.V, h2.V.sum, W.hat, h2.W,
106                   sum.W, h2.W.sum, V.hat.star, W.hat.star, V.hat.star.s,
107                   W.hat.star.s)
108   rm(var.e.hat, V.hat, h2.V, sum.V, h2.V.sum, W.hat, h2.W, sum.W, h2.W.sum,
109       V.hat.star, W.hat.star, V.hat.star.s, W.hat.star.s, y, y.scaled, var.g.hat,
110       g.hat, cov.g.hat)

```

```

111 }
112 #####
113 #####
114 # Functions for the calculation of the genomic variance in different set-ups
115 GenVarCur <- function(GRM, varg, eBLUP, eBLUPcov){
116   n <- dim(GRM)[[1]]
117   V <- varg * sum( diag(GRM) ) / (n - 1) # formula (8)
118   W <- V + sum(eBLUP ^ 2) / (n - 1) - sum( diag(eBLUPcov) ) / (n - 1)
119   return(c(V, W))
120 }
121 GenVarBase_Ped <- function(PedMat, GRM, varg, eBLUP, eBLUPcov){
122   n <- dim(GRM)[[1]]
123   P <- diag(1, n, n) - matrix(1, n, n) / n # Equation (2)
124   E <- eigen(PedMat)
125   Q <- E$vectors
126   D <- E$values
127   Ped.inv.halve <- Q %*% diag( 1 / sqrt(D), nrow(Q), nrow(Q)) %*% t(Q)
128   B_b <- Ped.inv.halve %*% P %*% Ped.inv.halve # equation (16)
129   V <- varg * sum( diag(B_b %*% GRM) ) / (n - 1) # formula (17)
130   W <- V + t(eBLUP) %*% B_b %*% eBLUP /
131     (n - 1) - sum(diag(B_b %*% eBLUPcov)) / (n - 1) # formula (18)
132   return(c(V, W))
133 }
134 GenVarBase_GRM <- function(GRM, phenos, mu, varg, eBLUP, eBLUPcov, vare){
135   n <- dim(GRM)[[1]]
136   P <- diag(1, n, n) - matrix(1, n, n) / n # Equation (2)
137   E_G <- eigen(GRM)
138   Q_G <- E_G$vectors
139   D_G <- E_G$values
140   D_G[D_G < 0] <- 0 # Last EV might be -0
141   G_half <- Q_G %*% diag(sqrt(D_G), nrow(Q_G), nrow(Q_G)) %*% t(Q_G)
142   Cov.y.minus <- chol2inv(GRM * varg + diag(vare, n, n))
143   eBLUP.s <- varg * G_half %*% Cov.y.minus %*% (phenos - mu)
144   # formula (20)
145   eBLUPcov.s <- varg ^ 2 * G_half %*% Cov.y.minus %*% G_half -
146     (varg ^ 2 / sum(Cov.y.minus)) * G_half %*% Cov.y.minus %*%
147     matrix(1, n, n) %*% Cov.y.minus %*% G_half # formula (21)
148   V <- varg # formula (19)
149   W <- V + t(eBLUP.s) %*% P %*% eBLUP.s / (n - 1) - sum(diag(P %*%
150     eBLUPcov.s)) / (n - 1) # formula (22)
151   return(c(V, W))
152 }

```

3 Results

We present the outcome of the *R*-script above in table format.

Table S1: Estimation results for the unconditional expectation V and the best predictor W for the additive genomic variance s_g^2 in the current population for the mice, wheat, and Arabidopsis datasets. We also present the corresponding heritabilities with respect to the sample variance of the phenotypic values and with respect to the sum of the additive genomic and residual variance σ_ε^2 .

In addition, we depict the estimation results for the unconditional expectation V^* (V_s^* when using the GRM for the transformation) and the best predictor W^* (W_s^* when using the GRM for the transformation) for the additive genomic variance in the base population.

Genom. Var. / Heritab.	Data	Population	Mice	Wheat	Arabidopsis
$\hat{V}$ ($= \hat{h}_V^2$)	Current		0.3737749	0.6039708	0.47333803
$\hat{V} + \hat{\sigma}_\varepsilon^2$			1.0754963	1.1449704	0.54832029
$\tilde{h}_V^2$			0.3475371	0.5274990	0.86325098
$\widehat{W}$ ($= \hat{h}_W^2$)			0.2982787	0.4590001	0.92501779
$\widehat{W} + \hat{\sigma}_\varepsilon^2$			1.0000002	0.9999998	1.00000005
$\tilde{h}_W^2$			0.2982787	0.4590002	0.92501774
$\hat{V}^*$	Base		0.3704021	3.0621134	—
$\widehat{W}^*$			0.3089758	2.0095836	—
$\hat{V}_s^*$			0.3639248	1.3158006	0.80762011
$\widehat{W}_s^*$			0.3577692	1.2300300	1.30240520