
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 \( \mu \) denotes a fixed intercept, \( \mathbf{1}_n := (1, \ldots, 1)^\top \) is an \( n \)-row-vector, \( \varepsilon \sim \mathcal{N}(0, \sigma^2_\varepsilon \mathbf{I}_n) \), 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

\[ g \sim \mathcal{N}(0, \sigma^2_g \mathbf{G}). \]

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

\[ \mathbf{G} := \mathbf{PXX}^\top \mathbf{P}/c \quad (1) \]

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 \( \sigma^2_\beta \) be the variance component of the marker effects \( \beta \). Then, the variance component in the equivalent model equals \( \sigma^2_g = c \sigma^2_\beta \) 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

\[ \hat{y} = \frac{y}{\hat{\sigma}_y} \quad (5). \]
We fit the model using the R-package “sommer” which returns the estimated variance components $\hat{\sigma}^2_g$ and $\hat{\sigma}^2_\varepsilon$ (REML approach). In addition, we obtain the empirical BLUP (eBLUP) for the random vector $g$ and the estimated variance-covariance matrix $\hat{\Sigma}_{\hat{\mu}|g|y} = \hat{\sigma}^2_g G(\hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon I_n)^{-1} G \hat{\sigma}^2_g$

\[
\hat{\mu}_{g|y} = \hat{\sigma}^2_g G(\hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon I_n)^{-1} (y - \hat{\mu}_n)
\]  

(6)

and

\[
\hat{\Sigma}_{\hat{\mu}|g|y} = \hat{\sigma}^2_g G(\hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon I_n)^{-1} G \hat{\sigma}^2_g - \hat{\sigma}^2_g G(\hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon I_n)^{-1} 1_n^\top (\hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon I_n)^{-1} 1_n
\]

(7)

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

When interest is in the current population, we have to set $B_c := I_n$, and obtain the estimated unconditional expectation

\[
\hat{V} = \frac{1}{n - 1} \hat{\sigma}^2_g \text{tr}(G)
\]

(8)

and the empirical best predictor (eBP)

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

(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

\[
\hat{h}^2_V := \frac{\hat{V}}{\hat{\sigma}^2_g} = \hat{V}
\]

(10)

as well as the eBP

\[
\hat{h}^2_W := \frac{\hat{W}}{\hat{\sigma}^2_g} = \hat{W}
\]

(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:

\[
\hat{\sigma}^2_{y,V} := \hat{V} + \hat{\sigma}^2_\varepsilon
\]

(12)

and

\[
\hat{\sigma}^2_{y,W} := \hat{W} + \hat{\sigma}^2_\varepsilon.
\]

(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:

\[
\hat{h}^2_V := \frac{\hat{V}}{\hat{\sigma}^2_{y,V}}
\]

(14)
and

\[ \hat{h}_W^2 := \frac{\hat{W}}{\hat{\sigma}^2_{y,W}}. \]  

(15)

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

\[ B_b := R^{-0.5}PR^{-0.5}, \]  

(16)

and we obtain the estimated unconditional expectation

\[ \hat{V}^* = \frac{1}{n-1} \hat{\sigma}^2_g \text{tr}(R^{-0.5}PR^{-0.5}G) \]  

(17)

and the eBP

\[ \hat{W}^* = \hat{V}^* + \frac{1}{n-1} \hat{\mu}_{g|y} R^{-0.5}PR^{-0.5} \hat{\mu}_{g|y} - \frac{1}{n-1} \text{tr}(R^{-0.5}PR^{-0.5} \hat{\Sigma}_{\mu_{g|y}}) \]  

(18)

for the additive genomic variance in the base population.

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

\[ G = G^{0.5}G^{0.5}. \]

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

\[ \hat{V}^*_s = \hat{\sigma}^2_g \]  

(19)

instead of calculating

\[ \frac{1}{n-1} \hat{\sigma}^2_g \text{tr}(PG^{-0.5}GG^{-0.5}) = \frac{1}{n-1} \hat{\sigma}^2_g \text{tr}(P) = \hat{\sigma}^2_g \]

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

We proceed analogously for the eBP. Instead of calculating

\[ G^{-0.5} \hat{\mu}_{g|y} = G^{-0.5} \hat{\sigma}^2_g G \left[G \hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon I_n \right]^{-1} (y - \hat{\mu}_n) \]

and

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

in (18), we use

\[ \hat{\mu}^*_{g|y} := \hat{\sigma}^2_g G^{0.5} \left(G \hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon I_n \right)^{-1} (y - \hat{\mu}_n) \]  

(20)

and

\[ \hat{\Sigma}^*_{\mu_{g|y}} := \hat{\sigma}^2_g G^{0.5} \hat{\Sigma} G^{0.5} \hat{\sigma}^2_g - \hat{\sigma}^2_g G^{0.5} \hat{\Sigma} I_n 1_n^\top \hat{\Sigma} G^{0.5} \hat{\sigma}^2_g \]  

\[ 1_n^\top \hat{\Sigma} I_n \]  

(21)

as surrogates. Then, we define

\[ \hat{W}^* = \hat{V}^* + \frac{1}{n-1} (\hat{\mu}^*_{g|y})^\top P \hat{\mu}^*_{g|y} - \frac{1}{n-1} \text{tr}(P \hat{\Sigma}^*_{\mu_{g|y}}) \]  

(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.

```r
# Author
# Nicholas Schreck, nschreck@mail.uni-mannheim.de
# Estimation of the Genomic Variance in the "Equivalent Linear Model"
# Copyright (C) 2018 -- 2019 Nicholas Schreck

# This program is free software; you can redistribute it and/or
# modify it under the terms of the GNU General Public License
# as published by the Free Software Foundation; either version 2
# of the License, or (at your option) any later version.
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
# GNU General Public License for more details.

# You should have received a copy of the GNU General Public License
# along with this program; if not, write to the Free Software
# Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

# Packages
library(sommer)
library(BGLR)
# Genomic data (phenotypes and marker genotypes)
data(mice)
data(wheat)
load("Arabidopsis1001genomes_1057lines_193697SNPs.Rdata")
y.all <- list(mice.pheno$Obesity.BodyLength, wheat.Y[, 1],
phenogeno$pheno$FT10_mean) # Phenotypic values
Ped.all <- list(mice.A, wheat.A, NA) # Pedigree matrices
X.all <- list(mice.X, wheat.X, phenogeno$geno) # marker-genotyp matrices
nr.datasets <- length(y.all)
# Output Variables
Variables <- c("eps", "V_hat", "h2_V", "sum_V", "h2_V_sum", "W_hat", "h2_W",
"sum_W", "h2_Wsum", "Vstar_hat", "Wstar_hat", "Vstars_hat",
"Wstars_hat")
nr.variables <- length(Variables)
result <- matrix(NA, nr.variables, nr.datasets)ownames(result) <- Variables
colnames(result) <- c("Mice", "Wheat", "Arabidopsis")
Variables, nr.variables)
```

2 R-Script for the Estimation of the Additive Genomic Variance
# Model fit and estimation of genomic variances
for (j in 1 : nr.datasets){  
  # Preliminaries
  y <- y.all[[j]]  # fix phenotype
  y.scaled <- y / sd(y)  # y scaled to variance of 1, see equation (5)
  n <- length(y)
  ones <- matrix(1, nrow=n, ncol=1)
  P <- diag(1, n, n) - matrix(1, n, n) / n  # Equation (2)
  X <- X.all[[j]]  # fix marker-genotypes
  p <- dim(X)[2]
  c <- sum(colMeans(X) * (1 - colMeans(X) / 2))  # vanRaden c, see equation (3)
  G <- P %*% tcrossprod(X) %*% P / c  # equation (1)
  rm(X)
  
  # ###########################################################################
  # Fit gBLUP model
  lm.equ <- mmer(Y=y.scaled, X=ones, Z=list(A=list(K=G)), silent=TRUE,
                 iters=50, tolpar=1e-7, tolparinv=1e-9, date.warning = FALSE)
  var.g.hat <- as.numeric(lm.equ$var.comp$A)  # variance component g
  var.e.hat <- as.numeric(lm.equ$var.comp$units)  # variance comp. eps
  mu.hat <- rep(lm.equ$beta.hat, n)  # Estimate of Intercept
  g.hat <- as.vector(lm.equ$u.hat$A)  # eBLUP, equation (6)
  cov.g.hat <- as.matrix(lm.equ$Var.u.hat$A$T1)  # Cov of eBLUP, equation (7)
  rm(lm.equ)
  
  # Genomic variances in the current population
  gv_cur <- GenVarCur(GRM=G, varg=var.g.hat, eBLUP=g.hat, eBLUPcov=cov.g.hat)
  V.hat <- gv_cur[1]
  h2.V <- V.hat # formula (10)
  sum.V <- V.hat + var.e.hat # formula (12)
  h2.V.sum <- V.hat / sum.V # formula (14)
  W.hat <- gv_cur[2]
  h2.W <- W.hat # (11)
  sum.W <- W.hat + var.e.hat # formula (13)
  h2.W.sum <- W.hat / sum.W # formula (15)
  
  # Genomic variances in the base population (via Pedigree)
  Ped <- Ped.all[[j]]
  if(!is.na(Ped)){
    gv_base_ped <- GenVarBase_Ped(PedMat=Ped, GRM=G, varg=var.g.hat, eBLUP=g.hat,
                                  eBLUPcov=cov.g.hat)
    V.hat.star <- gv_base_ped[1]
    W.hat.star <- gv_base_ped[2]
  } else{
    V.hat.star <- W.hat.star <- NA
  }
  
  # Genomic variances in the base population (via GRM)
  gv_base_grm <- GenVarBase_GRM(GRM=G, phenos=y.scaled, mu=mu.hat,
                               varg=var.g.hat, eBLUP=g.hat, eBLUPcov=cov.g.hat,
                               vare=var.e.hat)
  V.hat.star.s <- gv_base_grm[1]
  W.hat.star.s <- gv_base_grm[2]
  
  result[, j] <- c(var.e.hat, V.hat, h2.V, sum.V, h2.V.sum, W.hat, h2.W,
                   sum.W, h2.W.sum, V.hat.star, W.hat.star, V.hat.star.s,
                   W.hat.star.s)
  rm(var.e.hat, V.hat, h2.V, sum.V, h2.V.sum, W.hat, h2.W, sum.W, h2.W.sum,
      V.hat.star, W.hat.star, V.hat.star.s, W.hat.star.s, y, y.scaled, var.g.hat,
      g.hat, cov.g.hat)
# Functions for the calculation of the genomic variance in different set-ups

```r
def GenVarCur(GRM, varg, eBLUP, eBLUPcov)
    n <- dim(GRM)[1]
    V <- varg * sum(diag(GRM)) / (n - 1)  # formula (8)
    W <- V + sum(eBLUP^2) / (n - 1) - sum(diag(eBLUPcov)) / (n - 1)
    return(c(V, W))

def GenVarBase_Ped(PedMat, GRM, varg, eBLUP, eBLUPcov)
    n <- dim(GRM)[1]
    P <- diag(1, n, n) - matrix(1, n, n) / n  # Equation (2)
    E <- eigen(PedMat)
    Q <- E$ vectors
    D <- E$ values
    Ped.inv.halve <- Q %*% diag(1 / sqrt(D), nrow(Q), nrow(Q)) %*% t(Q)
    B_b <- Ped.inv.halve %*% P %*% Ped.inv.halve  # equation (16)
    V <- varg * sum(diag(B_b %*% GRM)) / (n - 1)  # formula (17)
    W <- V + t(eBLUP) %*% B_b %*% eBLUP / (n - 1) - sum(diag(B_b %*% eBLUPcov)) / (n - 1)  # formula (18)
    return(c(V, W))

def GenVarBase_GRM(GRM, phenos, mu, varg, eBLUP, eBLUPcov, vare)
    n <- dim(GRM)[1]
    P <- diag(1, n, n) - matrix(1, n, n) / n  # Equation (2)
    E_G <- eigen(GRM)
    Q_G <- E_G$ vectors
    D_G <- E_G$ values
    G_halve <- Q_G %*% diag(sqrt(D_G), nrow(Q_G), nrow(Q_G)) %*% t(Q_G)
    Cov.y.minus <- chol2inv(GRM * varg + diag(vare, n, n))
    eBLUP.s <- varg * G_halve %*% Cov.y.minus %*% (phenos - mu)
    # formula (20)
    eBLUPcov.s <- varg^2 * G_halve %*% Cov.y.minus %*% G_halve -
    (varg^2 / sum(Cov.y.minus)) * G_halve %*% Cov.y.minus %*%
    matrix(1, n, n) %*% Cov.y.minus %*% G_halve # formula (21)
    V <- varg  # formula (19)
    W <- V + t(eBLUP.s) %*% P %*% eBLUP.s / (n - 1) - sum(diag(P %*%
    eBLUPcov.s)) / (n - 1)  # formula (22)
    return(c(V, W))
```

6
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 $s^2_g$ 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 $\sigma^2_e$.

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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$V (= \hat{h}^2_V)$</td>
<td>Current</td>
<td>0.3737749</td>
<td>0.6039708</td>
<td>0.47333803</td>
<td></td>
</tr>
<tr>
<td>$\hat{V} + \hat{\sigma}^2_e$</td>
<td>Current</td>
<td>1.0754963</td>
<td>1.1449704</td>
<td>0.54832029</td>
<td></td>
</tr>
<tr>
<td>$\hat{h}^2_V$</td>
<td></td>
<td>0.3475371</td>
<td>0.5274990</td>
<td>0.86325098</td>
<td></td>
</tr>
<tr>
<td>$\hat{W} (= \hat{h}^2_W)$</td>
<td></td>
<td>0.2982787</td>
<td>0.4590001</td>
<td>0.92501779</td>
<td></td>
</tr>
<tr>
<td>$\hat{W} + \hat{\sigma}^2_e$</td>
<td></td>
<td>1.0000002</td>
<td>0.9999998</td>
<td>1.00000005</td>
<td></td>
</tr>
<tr>
<td>$\tilde{h}^2_W$</td>
<td></td>
<td>0.2982787</td>
<td>0.4590002</td>
<td>0.92501774</td>
<td></td>
</tr>
<tr>
<td>$V^*$</td>
<td>Base</td>
<td>0.3704021</td>
<td>3.0621134</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>$\hat{W}^*$</td>
<td></td>
<td>0.3089758</td>
<td>2.0095836</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>$\hat{V}^*_s$</td>
<td></td>
<td>0.3639248</td>
<td>1.3158006</td>
<td>0.80762011</td>
<td></td>
</tr>
<tr>
<td>$\hat{W}^*_s$</td>
<td></td>
<td>0.3577692</td>
<td>1.2300300</td>
<td>1.30240520</td>
<td></td>
</tr>
</tbody>
</table>