Supplementary material: Equivalent models

To derive a computationally efficient implementation of a regression algorithm exploiting codata for SNP markers for $p\gg n$, we present equivalent models to the SNP-BLUP and show how the estimates of the effects, and their associated prediction error variances, can be transformed between these. Prediction error variances are important to compute since they are the basis for calculations of standard errors and effective number of parameters $p\_{eff}$ [2, 3]. Most of theory below is found in [10] but here the focus is on fitting a regression model for SNP-specific variances using codata on the SNP markers, and some minor typos of [10] have been corrected below.

Three different, but equivalent, specifications of the random effects will be used and will be referred to as the *SNP Model*, the *animal model* and the *Cholesky Model*. The SNP effects may have variances with different weights given by the diagonal matrix $D$, and the residuals $e$ are assumed iid normal with variance $σ\_{e}^{2}$:

**SNP Model**

$\begin{matrix}y&=&Xβ+Zb+e\\b&∼&N(0,σ\_{b}^{2}D)\end{matrix}$ (1)

**Animal model**

$\begin{matrix}y&=&Xβ+a+e\\a&∼&N(0,σ\_{b}^{2}G)\\G&=&ZDZ^{'}\end{matrix}$ (2)

**Cholesky Model**

$\begin{matrix}y&=&Xβ+Lv+e\\v&∼&N(0,σ\_{b}^{2}I\_{n})\\LL^{'}&=&G\end{matrix}$ (3)

The use of equivalent LMMs in the research field of animal breeding and quantitative genetics is well established. The contribution here is to present how LMM theory can be used for heteroscedastic variance components, to show how the prediction error variances can be transformed between models, to implement the theory in a computationally efficient R package **CodataGS**, and to apply it to a model including external codata on the SNP genotypes presented further below.

Different mixed model equations for the equivalent models

For LMM Henderson's mixed model equations (MME) are used to estimate both the fixed and random effects for given variance components. They can also be used iteratively to estimate variance components as implemented in the R package **hglm** [9]. Although the models above are equivalent, the MME are different.

**SNP model**

For the SNP Model we have the MME

$\left(\begin{matrix}X^{'}X&X^{'}Z\\Z^{'}X&Z^{'}Z+\frac{σ\_{e}^{2}}{σ\_{b}^{2}}D^{-1}\end{matrix}\right)\left(\begin{array}{c}β\\b\end{array}\right)=\left(\begin{array}{c}X^{'}y\\Z^{'}y\end{array}\right)$ (4)

These MMEs are of size $(k+p)×(k+p)$, where $k$ is the number of columns in $X$. Hence, the size of the equations are very large for high-dimensional data.

**Animal model**

Let the random effects $a$ be individual effects for each observation and $G=ZDZ'$ the correlation matrix between these. Then $G$ is relatively small ($n×n$) and the MME are

$\left(\begin{matrix}X^{'}X&X^{'}\\X&I+\frac{σ\_{e}^{2}}{σ\_{b}^{2}}G^{-1}\end{matrix}\right)\left(\begin{array}{c}β\\a\end{array}\right)=\left(\begin{array}{c}X^{'}y\\y\end{array}\right)$ (5)

of size $(k+n)×(k+n).$ Hence, the size of these MME is much smaller than eq. 4 for

 $p\gg n$.

**Cholesky model**

In a third equivalent model we define $LL'=G$ (where $L$ has size $n×n$ ) and the random effects $v$ are individual independent random effects. The MME are

$\left(\begin{matrix}X^{'}X&X^{'}L\\L^{'}X&L^{'}L+\frac{σ\_{e}^{2}}{σ\_{b}^{2}}I\_{n}\end{matrix}\right)\left(\begin{array}{c}β\\v\end{array}\right)=\left(\begin{array}{c}X^{'}y\\L^{'}y\end{array}\right)$ (6)

of size $(k+n)×(k+n).$

Transformation of effects between equivalent models

For $p\gg n$, the size of the MME in models (5) and (6) are much smaller than in model (4).

The random effects can be transformed between these equivalent models [6, 7] so that the estimated SNP effects $\hat{b}$ can easily be calculated from the individual effects $\hat{a}$ in model (5)

$\hat{b}=DZ^{'}G^{-1}\hat{a}$. (7)

Furthermore, we have $\hat{a}=L\hat{v}$ so that

$\hat{b}=DZ^{'}G^{-1}L\hat{v}$. (8)

The matrix $Z$ is moderately large ($n×p$) but the transformation is a simple cross-product.

Hence, the calculations can be made in parts without reading all of $Z$ into memory. They can also easily be parallelized if necessary.

Transformation of prediction error variances between equivalent models

Not only the estimates, but also the prediction error variances (*i.e.* the diagonal elements of $Var(v-\hat{v}|v)$), are important to compute to allow for model checking and inference.

In the Cholesky model (eq. 6), let $C\_{v}$ be

$C\_{v}=\frac{1}{σ\_{e}^{2}}\left(\begin{matrix}X^{'}X&X^{'}L\\L^{'}X&L^{'}L+\frac{σ\_{e}^{2}}{σ\_{b}^{2}}I\_{n}\end{matrix}\right)$. (9)

Decompose the inverse of $C\_{v} $ as

$C\_{v}^{-1}=\left(\begin{matrix}C\_{v}^{11}&C\_{v}^{12}\\C\_{v}^{21}&C\_{v}^{22}\end{matrix}\right)$. (10)

Then the prediction covariance matrix is $Var(v-\hat{v}|v)=σ\_{b}^{2}I\_{n}-C\_{v}^{22}$, see [1].

Denote the *j*:th diagonal element of $Var(b-\hat{b}|b)$ as $V\_{\hat{b}\_{j}}$. Then these elements can be calculated separately as

$V\_{\hat{b}\_{j}}=σ\_{b}^{2}-M\_{j}(σ\_{b}^{2}I\_{n}-C\_{v}^{22})M\_{j}^{'}D^{-1}$ (11)

where $M\_{j}$ is the *j*:th row of the transformation matrix $M=DZ^{'}G^{-1}L$.

These computations are derived as follows. The prediction covariance matrix is

$Var(b-\hat{b}|b)=σ\_{b}^{2}D-C\_{b}^{22}$ (12)

$M$ is the matrix transforming effects $v$ to $b$ in eq. (8), then

$Var(b-\hat{b}|b)=MVar(v-\hat{v}|v)M'$ (13)

Combining these two equations, we get

$σ\_{b}^{2}D-C\_{b}^{22}=MVar(v-\hat{v}|v)M^{'}$ (14)

*i.e.*

$σ\_{b}^{2}D-C\_{b}^{22}=M(σ\_{b}^{2}I\_{n}-C\_{v}^{22})M^{'}$ (15)

So

$C\_{b}^{22}=σ\_{b}^{2}D-M(σ\_{b}^{2}I\_{n}-C\_{v}^{22})M^{'}$ (16)

A simple relationship between hat values for random effects and the prediction error variance

Henderson's MME are equivalent to the solving an augmented model using weighted least squares [4, 8]. The augmented design matrix for the SNP model is

$T=\left(\begin{matrix}X&Z\\0&I\end{matrix}\right)$ (17)

and the weight matrix is

$W=\left(\begin{matrix}\frac{1}{σ\_{e}^{2}}I&0\\0&\frac{1}{σ\_{b}^{2}}D^{-1}\end{matrix}\right)$ (18)

and the *hat matrix* for this augmented model is $H=T(T^{'}WT)^{-1}T^{'}W$.

Let $H\_{bb}$ be the lower right $p×p$ part of $H$ and $H\_{vv}$ the corresponding submatrix for the Cholesky model. Then the relationship between these submatrices and the prediction error variances are:

$H\_{vv}=C\_{v}^{22}\frac{1}{σ\_{b}^{2}}$ (19)

and

$H\_{bb}=C\_{b}^{22}D^{-1}\frac{1}{σ\_{b}^{2}}$ (20)

Consequently, the hat values for the SNP model (i.e. the *j*:th diagonal element in $H\_{bb}$, $h\_{jj}$) can be computed efficiently from the prediction error variance of the Cholesky model.

$h\_{jj}=1-M\_{j}(I\_{n}-C\_{v}^{22}/\hat{σ}\_{b}^{2})M\_{j}^{'}D\_{jj}^{-1}$ (21)

where $M\_{j} $is the *j*:th row of the transformation matrix $M$ and $D\_{jj}^{-1}$ is the j:th diagonal element of $D^{-1}.$

Let $h\_{bb}$ be the vector of the diagonal elements in $H\_{bb}$. In the R package **CodataGS**, $h\_{bb}$ is efficiently computed as $h\_{bb}=1-c\_{sum}D^{-1}$, where $c\_{sum}$ is the sum of columns of $(PM^{'})∘(PM^{'})$, $∘$ is the direct (elementwise) Hadamard product, and $P$ is a square root matrix of

 $I\_{n}-C\_{v}^{22}/\hat{σ}\_{b}^{2}$ computed using Singular Value Decomposition.

Fitting SNP variances using codata with a generalized linear model

Let $X\_{codata}$ be a design matrix with $p$ rows including covariates for the SNP markers. Then marker specific variances can be computed using a generalized linear model with a Gamma distribution and a log link function with response values $\frac{b^{2}}{1-h\_{bb}}$ and weights $\frac{1-h\_{bb}}{2}$ [4, 8], and linear predictor $X\_{codata}β\_{codata}$. An iterative procedure is applied where the diagonal elements in $X\_{codata}$ are updated using the fitted values: $exp(X\_{codata}\hat{β}\_{codata})$. This is the iterative reweighted least squares procedure for hierarchical generalized linear models derived in [3] and presented in [4] and [8].

**References**

[1] C. R. Henderson. *Applications of linear models in animal breeding*. University of Guelph, Guelph Ontario, 1984.

[2] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64(4):583-639.

[3] Y. Lee and J. A. Nelder. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society B*, 58:619-678, 1996.

[4] Y. Lee, J. A. Nelder, and Y. Pawitan. *Generalized linear models with random effects - unified analysis via h-likelihood*. Chapman & Hall/CRC, 2006.

[5] Y. Lee, L. Rönnegård, and M. Noh. *Data Analysis Using Hierarchical Generalized Linear Models with R*. Chapman and Hall/CRC, 2017.

[6] M. Lynch and B. Walsh. *Genetics and analysis of Quantitative Traits*. Sinauer Associates, Inc., 1998.

[7] Y. Nagamine. Transformation of QTL genotypic effects to allelic effects. *Genetics Selection Evolution*, 37:579-584, 2005.

[8] L. Rönnegård, M. Felleki, F. Fikse, H. A. Mulder, and E. Strandberg. Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. *Genetics Selection Evolution*, 42:8, 2010.

[9] L. Rönnegård, X. Shen, and M. Alam. hglm: A package for fitting hierarchical generalized linear models. *The R Journal,* 2(2):20-28, 2010.

[10] X. Shen, M. Alam, F. Fikse, and L. Rönnegård. A novel generalized ridge regression method for quantitative genetics. *Genetics*, 193:1255-1268, 2013.