GENOME-ENABLED PREDICTION

- RIDGE REGRESSION
- GENOMIC BLUP
- BAYESIAN GENOMIC BLUP
- ROBUST METHODS:

  TMAP AND LMAP

RIDGE REGRESSION

THE ESTIMATOR (HOERL AND KENNARD, 1979) WAS DERIVED USING A CONSTRAINED MINIMIZATION ARGUMENT

$$\beta^r = \left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} = \left(\mathbf{X}'\mathbf{X}+\omega\right)^{-1}\mathbf{X}'\mathbf{y},$$

Reciprocal of Lagrange multiplier
("regularization" parameter in machine learning)

Let $\mathbf{Z} = \left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}$, so that $\beta^r = \mathbf{Z}\hat{\beta}$ is a linear transformation of the OLS estimator.

Expected value

$$E\left(\beta^r|\mathbf{X}\right) = E\left(\mathbf{Z}\hat{\beta}\right) = \mathbf{Z}\beta,$$

Bias

$$\delta\left(\beta\right) = E\left(\beta^r|\mathbf{X}\right) - \beta = \left(\mathbf{Z}-\mathbf{I}\right)\beta = -w\left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1}\beta$$

Covariance matrix

$$\begin{aligned} Var\left(\beta^r|\mathbf{X}\right) &= Var\left[\left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}\right] \\ &= \left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1}\mathbf{X}'Var\left(\mathbf{y}|\mathbf{X}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1} \\ &= \left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1}\sigma_e^2 \\ &= \left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-1}\sigma_e^2 - w\left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-2}\sigma_e^2. \end{aligned}$$

- If λ is an eigenvalue of $\mathbf{X}'\mathbf{X}$.

$$
\begin{aligned}
E\left[(\boldsymbol{\beta}^r-\boldsymbol{\beta})'(\boldsymbol{\beta}^r-\boldsymbol{\beta})\,|\mathbf{X}\right] &= \boldsymbol{\delta}(\boldsymbol{\beta})'\boldsymbol{\delta}(\boldsymbol{\beta})+ \\
&\quad \sigma_e^2\left[\sum_{i=1}^{p}(\lambda_i+\omega)^{-1}-\omega\sum_{i=1}^{p}(\lambda_i+\omega)^{-2}\right] \\
&= \boldsymbol{\beta}'w^2\left(\mathbf{X}'\mathbf{X}+\omega\mathbf{I}\right)^{-2}\boldsymbol{\beta}+\sigma_e^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+\omega)^2}. \qquad (55)
\end{aligned}
$$

The preceding shows clearly that the contribution of bias to squared error increases with ω, while the variance term decreases. Then, it is conceivable that there might be values of ω at which the squared error of the ridge regression estimator is less than that of OLS, but these depend on the unknown $\boldsymbol{\beta}$. Hoerl and Kennard (1970 a, b) suggested using as value of ω one at which regression coefficients or some related quantity, such as the length of the vector of estimates, stabilize. Alternative procedures suitable for prediction problems are

$$y = 1\beta_0 + X\beta + e$$

The value of ω affects the degree of complexity of the model: as $\omega \to \infty$ the size of the ridge regression estimates decreases, effectively reducing the number of parameter fitted. For a model including an intercept, the incidence matrix $\mathbf{X}$ has the form $\mathbf{X} = [\mathbf{1}, \mathbf{X}_c]$, where "$c$" stands for covariates. The ridge regression estimates for this model are calculated as

$$\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_c \end{bmatrix} = \begin{bmatrix} n & \mathbf{1}'\mathbf{X}_c \\ \mathbf{X}_c'\mathbf{1} & \mathbf{X}_c'\mathbf{X}_c + \mathbf{I}\omega \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1} y_i \\ \mathbf{X}_c'\mathbf{y} \end{bmatrix}. \tag{56}$$

The effective number of parameters or "effective model degrees of freedom" (Ruppert et al. 2003) is calculated as

$$df_{Model} = tr\left([\mathbf{1}, \mathbf{X}_c] \begin{bmatrix} n & \mathbf{1}'\mathbf{X}_c \\ \mathbf{X}_c'\mathbf{1} & \mathbf{X}_c'\mathbf{X}_c + \mathbf{I}\omega \end{bmatrix}^{-1} [\mathbf{1}, \mathbf{X}_c]' \right); \tag{57}$$

likewise, the effective number of residual degrees of freedom is given by $n - df_{Model}$. It can be verified that df_{Model} decreases as ω increases, approaching 1 in the limit, corresponding to a model with an intercept as sole parameter.

NOTE: THE "INTERCEPT" IS NOT REGULARIZED

WAYS IN WHICH RIDGE REGRESSION CAN BE INTERPRETED

- As shrunken estimator of regressions or marker effects (ridge)
- As predictor of random effects (BLUP) [THIS IS A CRYPTIC INTERPRETATION BECAUSE WE WISH TO LEARN GENE EFFECTS : these do not vary at random, but over a conceptual distribution]
- As mean of a conditional posterior distribution (Bayes)
- As maximum of a penalized likelihood under the L2 norm (PMLE)
- In all cases we need ω OR variance ratio (and of the individual variances for interval inference)

BLUP

A mixed linear model is one where some of the regression coefficients are regarded as fixed parameters while other regressions are assumed to be realizations from the distribution of a vector that possesses a (frequentist) probability distribution, often called a "random effects" vector. For instance in a marker-based model with an intercept β_0, the allelic substitution effects ($\boldsymbol{\beta}$) of a battery of markers may be assumed to vary at random over conceptual repeated sampling according to the distribution $\boldsymbol{\beta} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_\beta\right)$. A mixed effects linear model could have the form

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{61}$$

where β_0 is a fixed parameter, $\boldsymbol{\beta}$ follows the normal distribution indicated above, and $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_e\right)$ is an independently distributed vector of residuals with variance σ^2_e. The matrix $\mathbf{X}$ is assumed to remain constant and sampling is over the process $\left[\mathbf{y}, \boldsymbol{\beta} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2_\beta, \sigma^2_e\right]$; the model induces $\mathbf{y} \sim N\left(\mathbf{1}\beta_0, \mathbf{X}\mathbf{X}'\sigma^2_\beta + \mathbf{I}\sigma^2_e\right)$ as marginal distribution of the data. In what follows it will be assumed that the two variance components are known; let $\omega = \dfrac{\sigma^2_e}{\sigma^2_\beta}$ and $\mathbf{V} = \mathbf{X}\mathbf{X}'\sigma^2_\beta + \mathbf{I}\sigma^2_e$

$$\widetilde{\boldsymbol{\beta}} = Cov\left(\boldsymbol{\beta}, \mathbf{y}'\right) \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{1}\widetilde{\beta}_0\right) = \sigma^2_{\beta} \mathbf{X}'\mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{1}\widetilde{\beta}_0\right) \tag{62}$$

where

$$\begin{aligned} \widetilde{\beta}_0 &= \frac{\mathbf{1}'\mathbf{V}^{-1}\mathbf{y}}{\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}} = \frac{\sum_{i=1}^{n} v^{i1} y_1 + \sum_{i=1}^{n} v^{i2} y_2 + ... + \sum_{i=1}^{n} v^{in} y_n}{\sum_{i=1}^{n}\sum_{j=1}^{n} v^{ij}} \\ &= \sum_{j=1}^{n} \gamma_j y_j. \end{aligned} \tag{63}$$

Above

$$\gamma_j = \frac{\sum_{i=1}^{n} v^{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} v^{ij}}, \tag{64}$$

is a weight $0 \leq \gamma_j \leq 1$, assigned to observation $j = 1, 2, ..., n$ with the sum of the γ_j weights being equal to 1. The statistic $\widetilde{\beta}_0$ is the best linear unbiased estimator of β_0, also called generalized least-squares (GLS) estimator; under normality, it is the maximum likelihood estimator of β_0 as well. The estimator is unbiased because $E\left(\widetilde{\beta}_0\right) = \beta_0$ under conceptual repeated sampling from $\left[\mathbf{y}, \boldsymbol{\beta} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2_{\beta}, \sigma^2_e\right]$. The preceding is easy to

Note that $BLUP(\boldsymbol{\beta})$ in (62) involves inverting $\mathbf{V}$, an $n \times n$ matrix. An alternative way of computing $GLS(\beta_0)$ and $BLUP(\boldsymbol{\beta})$ involves solving the "mixed model equations" (Henderson 1984) which, for this model, are

$$\begin{bmatrix} n & \mathbf{1}'\mathbf{X} \\ \mathbf{X}'\mathbf{1} & \mathbf{X}'\mathbf{X} + \mathbf{I}\omega \end{bmatrix} \begin{bmatrix} \widetilde{\beta}_0 \\ \widetilde{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}. \tag{66}$$

Solving for $\widetilde{\boldsymbol{\beta}}$ yields

$$\widetilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\omega\right)^{-1} \mathbf{X}' \left(\mathbf{y} - \mathbf{1}\widetilde{\beta}_0\right). \tag{67}$$

If the number of columns of $\mathbf{X}$ is p, the mixed model equations are convenient provided $p + 1 < n$. If the model does not include an intercept (or any other fixed effects), (67) has exactly the same form as the ridge regression estimator given in (49). Hence, BLUP in a random effects model (an unbiased predictor) is identical to a biased estimator of $\boldsymbol{\beta}$!

It can be shown that

$$Var\left(\begin{bmatrix} \widetilde{\beta}_0 \\ \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{bmatrix}\right) = \begin{bmatrix} n & \mathbf{1}'\mathbf{X} \\ \mathbf{X}'\mathbf{1} & \mathbf{X}'\mathbf{X} + \mathbf{I}\omega \end{bmatrix}^{-1} \sigma_e^2 = \begin{bmatrix} c^{\beta_0\beta_0} & \mathbf{c}'^{\beta_0\boldsymbol{\beta}} \\ \mathbf{c}^{\boldsymbol{\beta}\beta_0} & \mathbf{C}^{\boldsymbol{\beta}\boldsymbol{\beta}} \end{bmatrix} \sigma_e^2, \tag{68}$$

where $c^{\beta_0\beta_0}\sigma_e^2$ gives the variance of $\widetilde{\beta}_0$; $\mathbf{C}^{\boldsymbol{\beta}\boldsymbol{\beta}}$ gives the variance-covariance matrix of the prediction errors $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}$, and $\mathbf{c}^{\boldsymbol{\beta}\beta_0} = Cov\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}, \widetilde{\beta}_0\right)$, as shown in Henderson (1973, 1984).

Bayes

In a Bayesian context, let the regression coefficients be assigned the joint prior distribution $N\left(\mathbf{0}, \mathbf{I}\sigma^2_\beta\right)$: the uncertainty about each of the elements of $\boldsymbol{\beta}$ is represented by the same normal distribution with a null mean and constant variance σ^2_β. Under the regression model $\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2_e \sim N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2_e\right)$ and assuming known variance components σ^2_β and σ^2_e, the joint posterior distribution (e.g., Gianola and Fernando 1986) is

$$\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2_e, \sigma^2_\beta \sim N\left(\boldsymbol{\beta}^r, \left[\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1} \sigma^2_e\right).$$

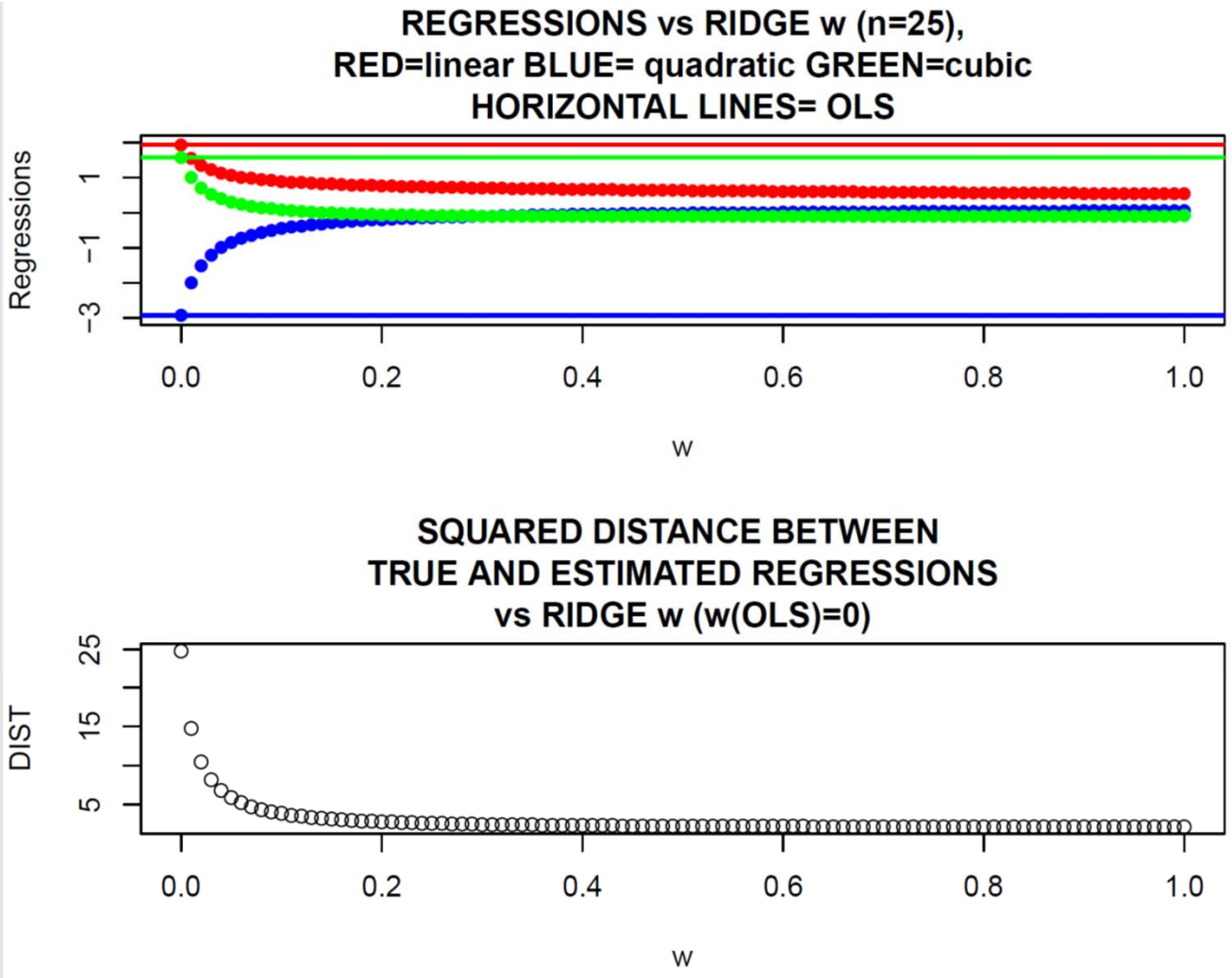
$$\boldsymbol{\beta}^r = \arg\max_{\beta}\left[N\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2_e\right) N\left(\mathbf{0}, \mathbf{I}\sigma^2_\beta\right)\right].$$

IN THE BAYESIAN INTERPRETATION, β HAS A TRUE, FIXED VALUE.

THE RANDOMNESS REPRESENTS UNCERTAINTY PRIOR AND POSTERIOR TO OBSERVING DATA

RIDGE REGRESSION EXAMPLE 1

This example illustrates how ridge regression works when the columns of $\mathbf{X}$ are strongly co-linear: a polynomial regression on a single explanatory variable. A sample of size $n = 25$ was simulated and the true model was $y = \beta_1 x + \beta_2 x^2 + \beta_2 x^3 + e$; the model does not have an intercept. The ridge regression parameter (ω) was allowed to take values between 0 (producing OLS estimates) and 1, with increments of 0.01. The script

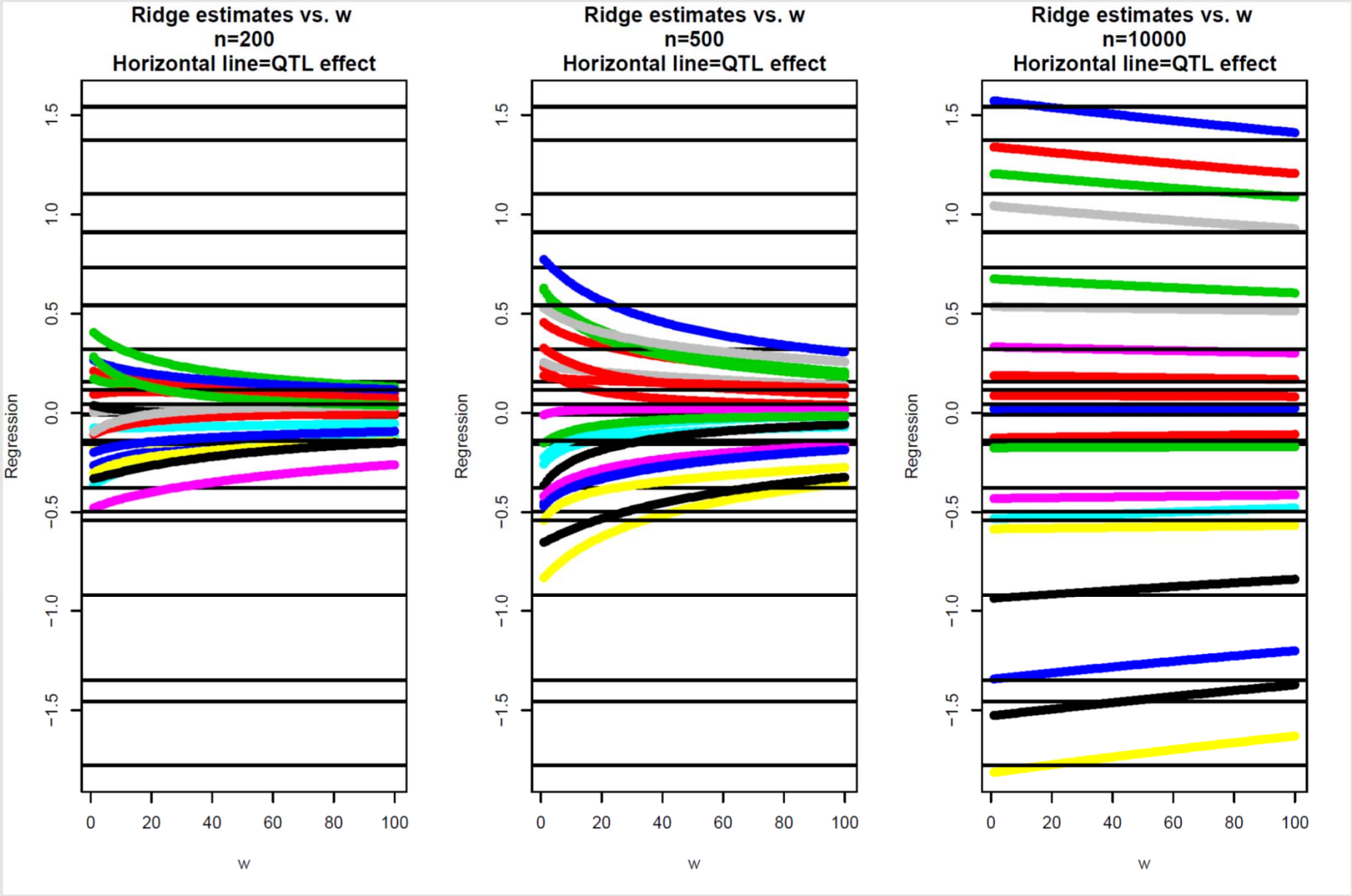


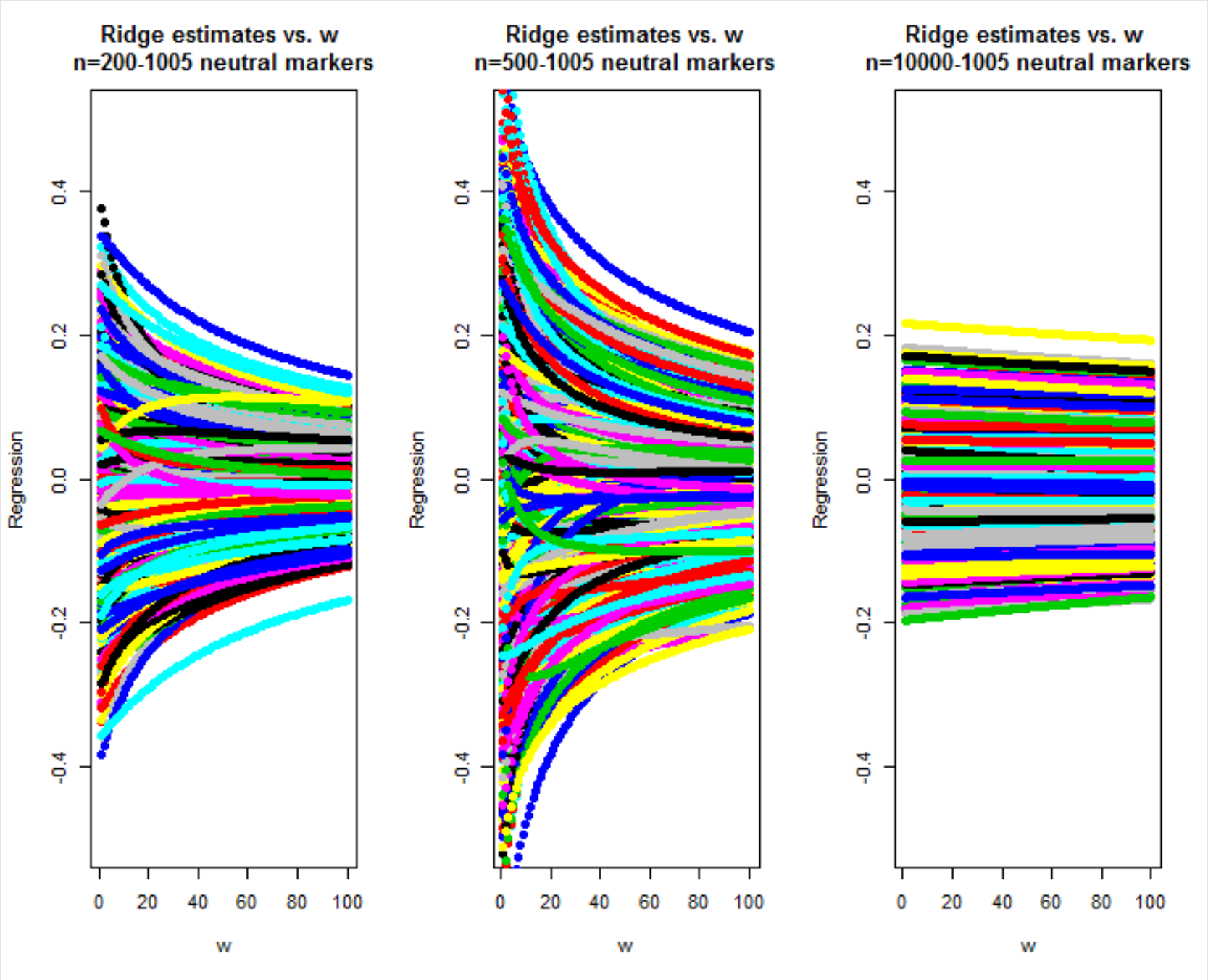
RIDGE REGRESSION EXAMPLE 2

`R-script Prediction 7` simulated biallelic 205 loci in each of 5 chromosomes. There were 4 additive biallelic QTL and 201 markers in each of the chromosomes. Linkage disequilibrium within chromosomes was generated by combined use of a Dirichlet distribution and a decay function, as in a previous example; there was no LD between chromosomes. QTL were placed in the following positions: 1-50-100-150 (Chrom. 1); 25-45-90-180 (Chrom. 2); 5-10-15-20 (Chrom. 3) and 190-195-200-205 (Chromosomes 4 and 5); the remaining 1005 loci were neutral markers, i.e., simulated as having no effect on phenotypes. The 20 QTL allelic substitution effects ($\mathbf{q}$) were independently drawn from a $N(0,1)$ distribution. Phenotypes were formed by calculating the additive genetic values of the individuals as $\mathbf{Q}_{n\times 20}\mathbf{q}$, where $\mathbf{Q}$ is the incidence matrix of QTL genotypes, and adding environmental deviates sampled independently from a $N(0,4)$ process.

First, a sample of $n = 50,000$ individuals was drawn to estimate the portion of the phenotypic variability accounted for by the QTL additive effects. "Baseline" genotypic frequencies were 0.15, 0.25 and 0.60 for aa, Aa and AA genotypes, respectively. Six OLS regressions of phenotypes on QTL were fitted: one for each of the five chromosomes (4 QTL per chromosome) and a model with all 20 QTL fitted together. The R^2 values chromosome by chromosome were 0.10, 0.05, 0.01, 0.26 and 0.17, after rounding. The sum of these five R^2 values was 0.58, same as the R^2 when the regression was on the 20 QTL simultaneously. This result was expected, as QTL in different chromosomes were in linkage equilibrium, so the sum of the additive genetic variances contributed by each chromosome should be equal to the the variance generated by all QTL jointly, at least in a large sample.

In practice, QTL locations, genotypes and effects are typically unknown, and a battery of markers genotypes on a small sample, together with phenotypes, constitute the information available. In our example, all QTL were represented in the 1025 markers. As noted before, ridge regression produces a unique solution regardless of the rank of $\mathbf{X}_{n\times p} \leq \min(n,p)$. The script below together with `R-script Prediction 7` (with n modified appropriately) was used to obtain ridge regression estimates for the 1025 markers over a grid of ω values (ranging from $\frac{1}{2}$ to 100 at increments of $\frac{1}{2}$) at samples of sizes $n = 200$, 500 and 10000. We fitted the "true" model (with an OLS regression on the 20 QTL genotypes) and a marker based model with 1025 regression coefficients, apart from an intercept. Squared distances between the true substitution effects ($\mathbf{q}$) and the OLS and ridge regression estimates of the substitution effects of the 20 markers that were QTL were calculated to observe the impact of ω and n on the "closeness" of the estimates. The script applies to $n = 200$.


Ridge estimates vs. w
n=200
Horizontal line=QTL effect
Ridge estimates vs. w
n=500
Horizontal line=QTL effect
Ridge estimates vs. w
n=10000
Horizontal line=QTL effect
Regression
w
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
20
40
60
80
100


Ridge estimates vs. w
n=200-1005 neutral markers
Ridge estimates vs. w
n=500-1005 neutral markers
Ridge estimates vs. w
n=10000-1005 neutral markers
Regression
0.4
0.2
0.0
-0.2
-0.4
0
20
40
60
80
100
w

ASSIGNING A VALUE TO ω: GENERALIZED CROSS-VALIDATION

VARIANCE COMPONENTS+ BRUTE FORCE:

There are several manners in which a value of ω can be arrived at. Consider, for example, a simple training set-testing set cross-validation layout. One can adopt a Bayesian or a mixed effects model perspective and estimate the σ_e^2 and σ_β^2 variance components in each training instance, then forming estimates $\widetilde{\omega}$ together with a set of predictions. Alternatively, the model could be trained over a grid of values of ω, with prediction performance evaluated as usual. A variant of the theme is to divide the data into training-calibration-testing sets. The model would be trained over the ω grid, its predictive performance assesed with the calibration set, and $\widetilde{\omega}$ chosen as the value producing, say, the smallest predictive MSE in the calibration set. Then, $\widetilde{\omega}$ would be used to gauge predictive ability in the testing set.

GCV: MOTIVATION

Another approach is to employ all data via what is called "generalized cross-validation"; see Craven and Wahba 1979) and Golub et al. (1979) for theoretical foundations. Recall from our least-squares treatment that, for $\mathbf{H}_d = \mathbf{X}_{[d]} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'_{[d]}$, the error of prediction of d phenotypes entering into the testing set is

$$\mathbf{y}_{[d]} - \mathbf{X}_{[d]}\widehat{\boldsymbol{\beta}}_{[-d]} = \left(\mathbf{I} - \mathbf{H}_d\right)^{-1} \widehat{\mathbf{e}}_{[d]}, \tag{76}$$

and that the mean-squared error of prediction of the d observations left out (testing set) takes the form

$$PMSE(d) = \frac{1}{d}\left(\mathbf{y}_{[d]} - \mathbf{X}_{[d]}\widehat{\boldsymbol{\beta}}\right)' \left(\mathbf{I} - \mathbf{H}_d\right)^{-2} \left(\mathbf{y}_{[d]} - \mathbf{X}_{[d]}\widehat{\boldsymbol{\beta}}\right), \tag{77}$$

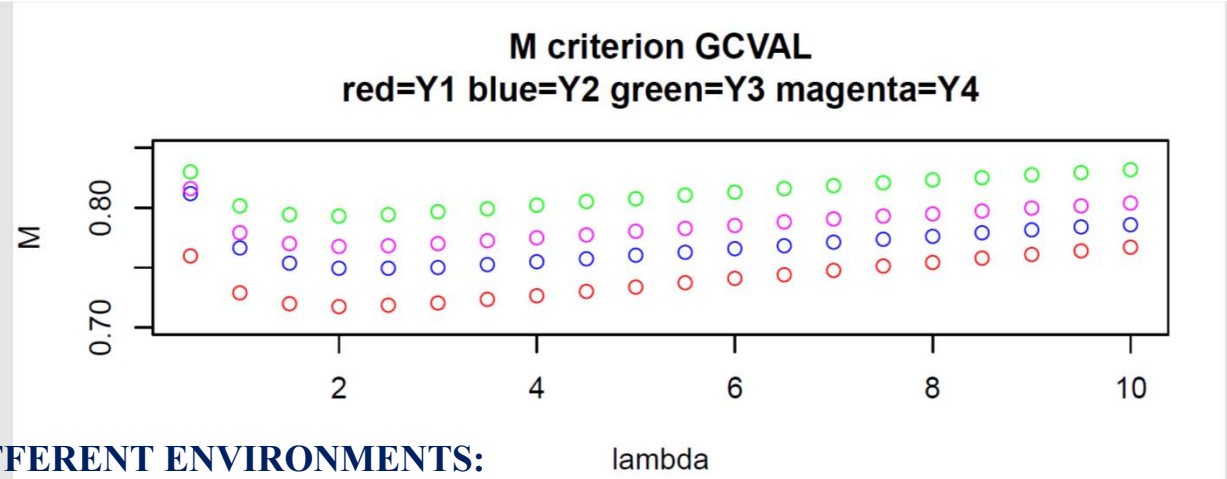
where $\widehat{\boldsymbol{\beta}}$ was the OLS estimator obtained with the entire data set and $\widehat{\mathbf{e}}_{[d]}$ = is the vector of fitted residuals $\mathbf{y}_{[d]} - \mathbf{X}_{[d]}\widehat{\boldsymbol{\beta}}$ (whole data set) for the testing set cases. The formulae above extend to ridge regression (Craven and Wahba 1979; Golub et al. 1979; Gianola and Schön, 2016). Following Golub et al. (2019), set $\omega = n\lambda$ and let $\mathbf{H}(\lambda) = \mathbf{X}\left(\mathbf{X}'\mathbf{X} + \mathbf{I}n\lambda\right)^{-1}\mathbf{X}'$; analogously to (77) write

The residual SS will be called T(λ)

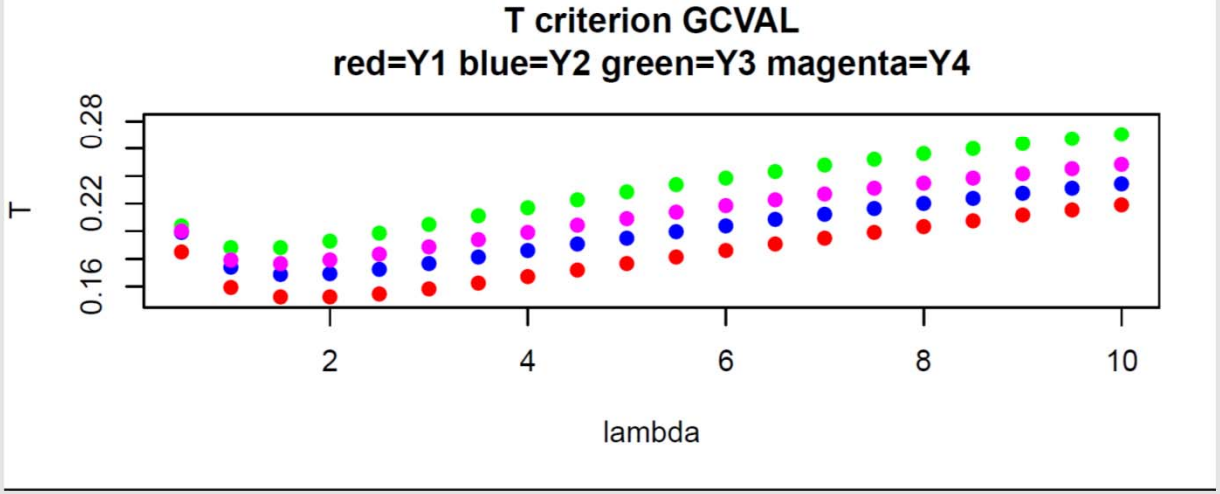
$$\begin{aligned} \mathbf{M}(\lambda) &= \frac{1}{n}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^r\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^r\right) \Big/ \left\{\frac{1}{n} tr\left(\mathbf{I} - \mathbf{H}(\lambda)\right)\right\}^2 \\ &= \frac{1}{n}\mathbf{y}'\left(\mathbf{I} - \mathbf{H}(\lambda)\right)^2 \mathbf{y} \Big/ \left\{1 - \overline{h}(\lambda)\right\}^2; \end{aligned} \tag{78}$$

note that $\frac{1}{n} tr\left(\mathbf{I} - \mathbf{H}(\lambda)\right) = 1 - \overline{h}(\lambda)$, where $\overline{h}(\lambda)$ is the average of the diagonal elements of $\mathbf{H}(\lambda)$. The value of λ that minimizes (78) is called the "generalized cross-validation" (GCV) estimate of λ. Further, using

The wheat data available in package BGLR (Pérez-Rodríguez and de los Campos 2014) was employed to illustrate how an "optimum" λ (recall that $\omega = n\lambda$) can be found numerically by evaluation of $M(\lambda)$ over a grid; $T(\lambda)$ was calculated as well. The data originated in trials conducted by the International Maize and Wheat Improvement Center (CIMMYT), Mexico. There are 599 wheat inbred lines, each genotyped with 1279 DArT (Diversity Array Technology) markers and planted in 4 environments; the target trait was grain yield in environments 1-4. Sample size was $n = 599$ with $p = 1279$ being the number of markers. These DArT markers are binary $(0, 1)$ and denote presence or absence of an allele at a marker locus in a given line. The analysis was carried out for each of the environments using the following script.



SAME TRAIT IN 4 DIFFERENT ENVIRONMENTS:
FOUR DIFFERENT BEHAVIORS!



GENOMIC BLUP

Assume that phenotypes and markers are centered, i.e., expressed as a deviation from their mean, and treat allelic substituion effects $\boldsymbol{\beta}$ as random variables following a normal $N\left(\mathbf{0}, \mathbf{I}\sigma^2_\beta\right)$ distribution, where σ^2_β can be interpreted as "variance among marker effects". If $\mathbf{X}$ is an $n \times p$, then $\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$ is an $n \times 1$ vector of marked genetic values for the n individuals and $\mathbf{g} \sim N\left(\mathbf{0}, \mathbf{X}\mathbf{X}'\sigma^2_\beta\right)$. Once marker effects are estimated as $\boldsymbol{\beta}^r$, a representation of "genomic BLUP" (GBLUP) for the n individuals is the vector $\hat{\mathbf{g}} = \mathbf{X}\boldsymbol{\beta}^r$ with its i^{th} element being $\hat{\mathbf{g}}_i = \mathbf{x}_i'\boldsymbol{\beta}^r$.

In GBLUP, "genomic relationship matrices" are used, and various genomic relationship matrices have been proposed by, e.g., Van Raden (2008), Astle and Balding (2009) and Rincent et al. (2014). These matrices are taken as proportional to $\mathbf{XX}'$ ($\mathbf{X}$ often has centered columns), i.e., some are expressible as $\mathbf{G} = \dfrac{\mathbf{XX}'}{c}$ for some constant c. Hence, $\mathbf{g} \sim N\left(\mathbf{0}, \mathbf{XX}'\sigma^2_\beta = \mathbf{G}\sigma^2_g\right)$, where $\sigma^2_g = c\sigma^2_\beta$ is called "genomic variance" or "marked additive genetic variance" whenever $\mathbf{X}$ encodes additive effects; there is no loss of generality if $c = 1$ is used, thus preserving the ω employed for BLUP (ridge regression) of marker effects. The ratio $h^2_g = \dfrac{\sigma^2_g}{\sigma^2_g + \sigma^2_e}$ has been termed "genomic heritability" (e.g., de los Campos et al. 2015). In particular, Van Raden's genomic relationship matrix takes the form

$$\mathbf{G}_{VR} = \frac{\mathbf{XX}'}{\sum_{j=1}^{p} 2q_j\left(1 - q_j\right)}, \tag{82}$$

where q_j is the frequency of the "minor" allele at marker locus j; here $c = \sum_{j=1}^{p} 2q_j\left(1 - q_j\right)$ assumes that markers are at Hardy-Weinberg equilibrium at each locos. Under some assumptions $\mathbf{G}_{VR}$ is the realization of a random matrix whose expected value is $\mathbf{A}$, a pedigree-based relationship matrix (Henderson, 1976). Another genomic relationship matrix could be constructed as

$$\mathbf{G}_{std} = \frac{\mathbf{X}_{std}\mathbf{X}'_{std}}{p}, \tag{83}$$

where $\mathbf{X}_{std}$ means that the entries of $\mathbf{X}$ have been standardized. i.e., deviated from the corresponding column means and expressed in standard deviation units.

RECALL

Genotypes (random variable W denotes genotype at a locus)

$$\begin{cases} W(aa) \rightarrow -1 \\ W(Aa) \rightarrow 0 \\ W(AA) \rightarrow 1 \end{cases}$$

$$\begin{aligned} E_{HW}(W) &= p^2 - q^2 = (p-q) = \mu \\ Var_{HW}(W) &= E(X^2) - E^2(X) \\ &= p^2 + q^2 - (p-q)^2 \\ &= 2pq \end{aligned}$$

$$\begin{cases} W(aa) \rightarrow 0 \\ W(Aa) \rightarrow 1 \\ W(AA) \rightarrow 2 \end{cases}$$

$$\begin{aligned} E_{HW}(W) &= 2p^2 + 2pq = 2p(p+q) = 2p \\ Var_{HW}(W) &= 4p^2 + 2pq - 4p^2 = 2pq \end{aligned}$$

Coding does not affect the variance of genotypes but mean shifts $2p - (p-q) = 1$

<u>Deviations from means are invariant to this type of coding</u>

$W - E(W)|Coding\ 1$

$$\begin{aligned} -1 - (p-q) &= -1 - p + q = -2p \\ 0 - (p-q) &= q - p = 1 - 2p \\ 1 - (p-q) &= 1 - p + q = 2(1-p) \end{aligned}$$

$W - E(W)|Coding\ 2$

$$\begin{aligned} 0 - 2p &= -2p \\ 1 - 2p & \\ 2 - 2p &= 2(1-p) \end{aligned}$$

A LOOK AT VAN RADEN'S GENOMIC RELATIONSHIP MATRIX

$$X_{\text{ind,marker}} = \begin{bmatrix} x_{11} & . & x_{1p} \\ x_{21} & . & x_{2p} \\ . & . & . \\ x_{n1} & . & x_{np} \end{bmatrix}$$

$$XX' = \begin{bmatrix} x_{11} & . & x_{1p} \\ x_{21} & . & x_{2p} \\ . & . & . \\ x_{n1} & . & x_{np} \end{bmatrix} \begin{bmatrix} x_{11} & x_{21} & . & x_{n1} \\ . & . & . & . \\ x_{1p} & x_{2p} & . & x_{np} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{p} x_{1_j}^2 & \sum_{j=1}^{p} x_{1_j}x_{2_j} & & \sum_{j=1}^{p} x_{1_j}x_{n_j} \\ & \sum_{j=1}^{p} x_{2_j}^2 & & \\ & & . & \\ & & & \sum_{j=1}^{p} x_{n_j}^2 \end{bmatrix}$$

In Van Raden's G-matrix :

$$E\left(\sum_{j=1}^{p} x_{ij}^2\right) = \sum_{j=1}^{p} Var(x_{ij}) + \sum_{j=1}^{p} E^2(x_{ij})$$

$$= \sum_{j=1}^{p} 2p_j q_j + \sum_{j=1}^{p} (p_j - q_j)^2$$

Term drops if $\mu = p_j - q_j = 0$

Or if x's centered

If all elements of G(VR) are divided by this factor, then scale is "consistent" with A.

$$E\left(\sum_{j=1}^{p} x_{1_j} x_{2_j}\right) = \sum_{j=1}^{p} Cov(x_{1_j} x_{2_j}) + \sum_{j=1}^{p} E(x_{1_j}) E(x_{2_j})$$

$$= \sum_{j=1}^{p} 2\phi_{ij} p_j q_j + \sum_{j=1}^{p} (p_j - q_j)^2$$

$$Cov(x_{1_j} x_{2_j}) = p_j^2 + q_j^2 - 2p_j q_j (1 - \phi) - (p_j - q_j)^2$$

$$= 2pq\phi$$

"additive relationship"

Note: LD does not enter into this form of genomic relationship matrix

UNDER HARDY-WEINBERG AND IDEALIZED CONDITIONS

After setting $\mu = p_j - q_j = 0$

$$E(XX') = \begin{bmatrix} \sum_{j=1}^{p} 2p_jq_j & a_{12}\sum_{j=1}^{p} 2p_jq_j & \cdots & a_{1n}\sum_{j=1}^{p} 2p_jq_j \\ & \sum_{j=1}^{p} 2p_jq_j & & a_{2n\mu}\sum_{j=1}^{p} 2p_jq_j \\ & & \ddots & \vdots \\ & & & a_{n,n-1}\sum_{j=1}^{p} 2p_jq \\ & & & \sum_{j=1}^{p} 2p_jq_j + \end{bmatrix}$$

Symmetric

Additive relationships

Likewise, if the x's are centered

$$E\left\{[X - E(X_{n\times p})][X - E(X_{n\times p})]'\right\} = \left(\sum_{j=1}^{p} 2p_j q_j\right)\begin{bmatrix} 1 & a_{12} & . & . & a_{1n} \\ \text{Symmetric} & 1 & & & a_{2n\mu} \\ & & . & & . \\ & & & . & a_{n,n-1} \\ & & & & 1 \end{bmatrix}$$

$$= A\left(\sum_{j=1}^{p} 2p_j q_j\right)$$

$$E\left[\frac{[X - E(X_{n\times p})][X - E(X_{n\times p})]'}{\left(\sum_{j=1}^{p} 2p_j q_j\right)}\right] = A$$

A= n x n matrix of additive relationships

Then, the “genomic” relationship matrix

$$G = \frac{(X - E(X))(X - E(X))'}{2\sum_{j=1}^{p} p_j(1-p_j)} = \frac{X^* X^{*\prime}}{V_{M,HW}}$$

Is the realization of a process. If this process is the HW process, then its expectation is

$$E\left[\frac{[X - E(X_{n\times p})][X - E(X_{n\times p})]'}{\left(\sum_{j=1}^{p} 2p_j q_j\right)}\right] = A$$

For example: parent and offspring are expected to have a relationship=0.5
but in reality it could be larger or smaller

MANY G-MATRICES
(each may provide a different variance component decomposition)
Examples

$$\mathbf{G}_{VR} = \frac{1}{\sum_{j=1}^{p} 2p_j(1-q_j)}\mathbf{X}_{cent}\mathbf{X}'_{cent}$$

$$\mathbf{G}_{ST} = \frac{1}{p}\mathbf{X}_{std}\mathbf{X}'_{std}; \mathbf{X}_{std} = \left\{ \frac{x_{ij} - \bar{x}_j}{\sqrt{Var(x_{ij}}} \right\}$$

$\overline{\mathbf{G}} = \frac{1}{2}(\mathbf{G}_{VR} + \mathbf{G}_{ST})$ followed by some scaling?

$\mathbf{G}_W = \frac{1}{p}\mathbf{X}_{std}\mathbf{W}\mathbf{X}'_{std}$ where $\mathbf{W}$ uses LD information?

$\mathbf{G}_{Blend} = \omega\mathbf{G}_{ST} + (1-\omega)\mathbf{A}$ after some re-scaling of matrices

$\mathbf{G}_{VR}$ scaled in (0,2) with mapminimax function?

$$y = ymin + \frac{(x - xmin)}{(xmax - xmin)}(ymax - ymin)$$

$$g_{ij} = 2\frac{g_{ij,VR} - \min(g_{ij,VR})}{\max(g_{ij,VR}) - \min(g_{ij,VR})}$$

FOCUS: GWAS

Statistical Science
2009, Vol. 24, No. 4, 451–471
DOI: 10.1214/09-STS307

Population Structure and Cryptic Relatedness in Genetic Association Studies

William Astle and David J. Balding[1]

FOCUS: HERITABILITY

Theoretical Population Biology 107 (2016) 26–30

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Comparing estimates of genetic variance across different relationship models

Andres Legarra
INRA, UMR 1388 GenPhySE (Génétique, Physiologie et Systèmes d'Elevage), F-31326 Castanet-Tolosan, France

FOCUS: KINSHIP

Genetics: Early Online, published on January 18, 2017 as 10.1534/genetics.116.197004
GENETICS | INVESTIGATION

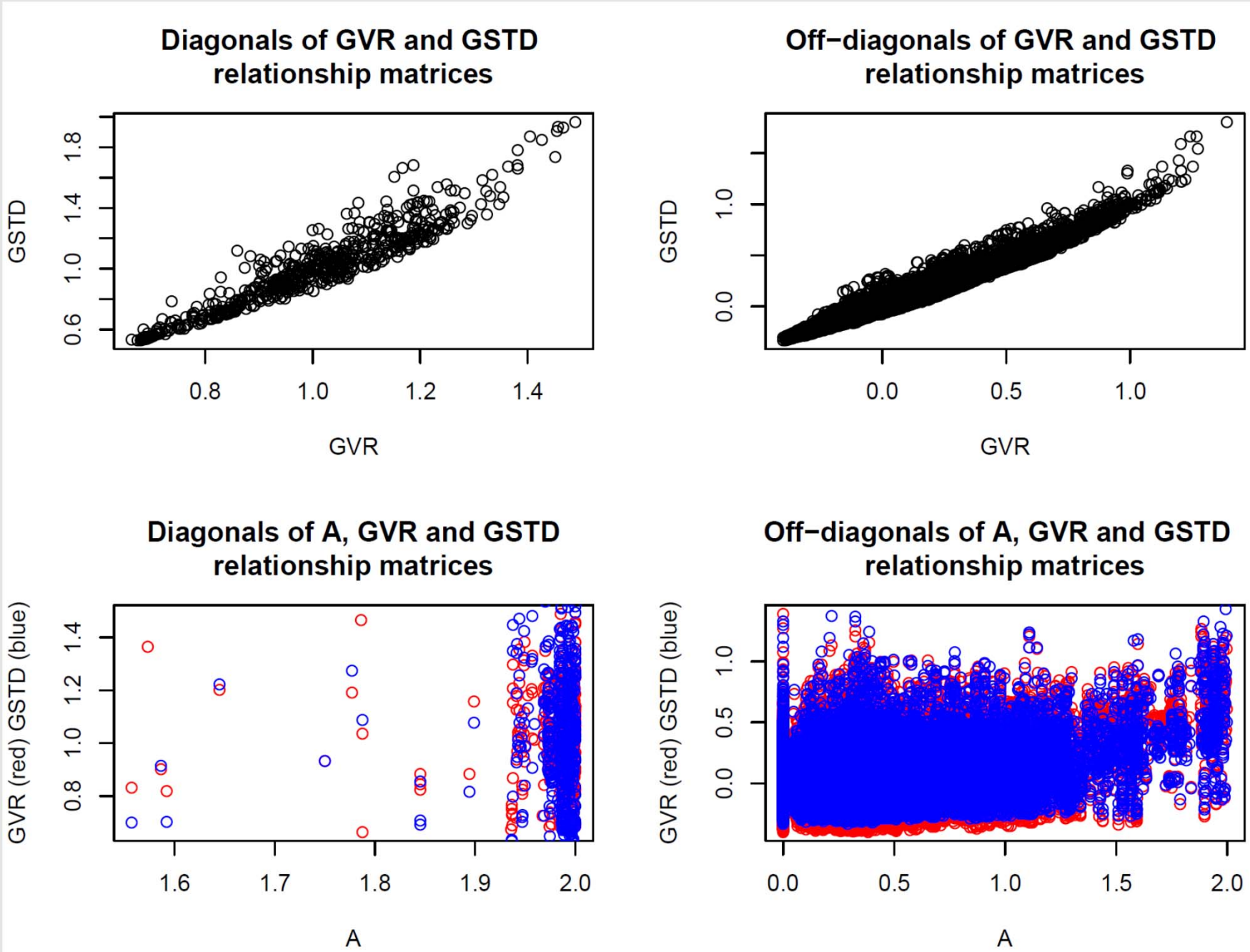
Efficient estimation of realized kinship from SNP genotypes

Bowen Wang , Serge Sverdlov and Elizabeth Thompson[1]
Department of Statistics, University of Washington, Seattle, Washington 98195-4322

EXAMPLE: TWO GENOMIC RELATIONSHIP MATRICES

The wheat data set employed in the preceding example was used to build two genomic relationship matrices from matrix $\mathbf{X}$. One was $\mathbf{G}_{VR}$ as in (82) and the other one was G_{std} as in (83).

Since the wheat material represents 599 inbred lines (so genotypes can be either aa or AA at each marker locus) the denominator in $\mathbf{G}_{VR}$ was $\sum_{j=1}^{p} q_j (1 - q_j)$. For a bi-allelic locus in Hardy-Weinberg equilibrium with genotypes (W, say) coded as -1, 0 and 1 for aa, Aa and AA, respectively, the expected value of the distribution of codes is $E(W) = -1 \times q^2 + 0 \times 2q(1-q) + 1 \times (1-q)^2 = 1 - 2q$. Further, $E(W^2) = q^2 + (1-q)^2 = 1 - 2q + 2q^2$, so $Var(W) = E(W^2) - E^2(W) = 2q(1-q)$. Hence, the denominator in Van Raden's matrix can be interpreted as the variance of the sum of genotype scores for a set of markers in linkage equilibrium. If the only possible genotypes are aa and AA with frequencies q and $1-q$, respectively, then $Var(W) = q(1-q)$, which is the variance of a Bernoulli distribution.

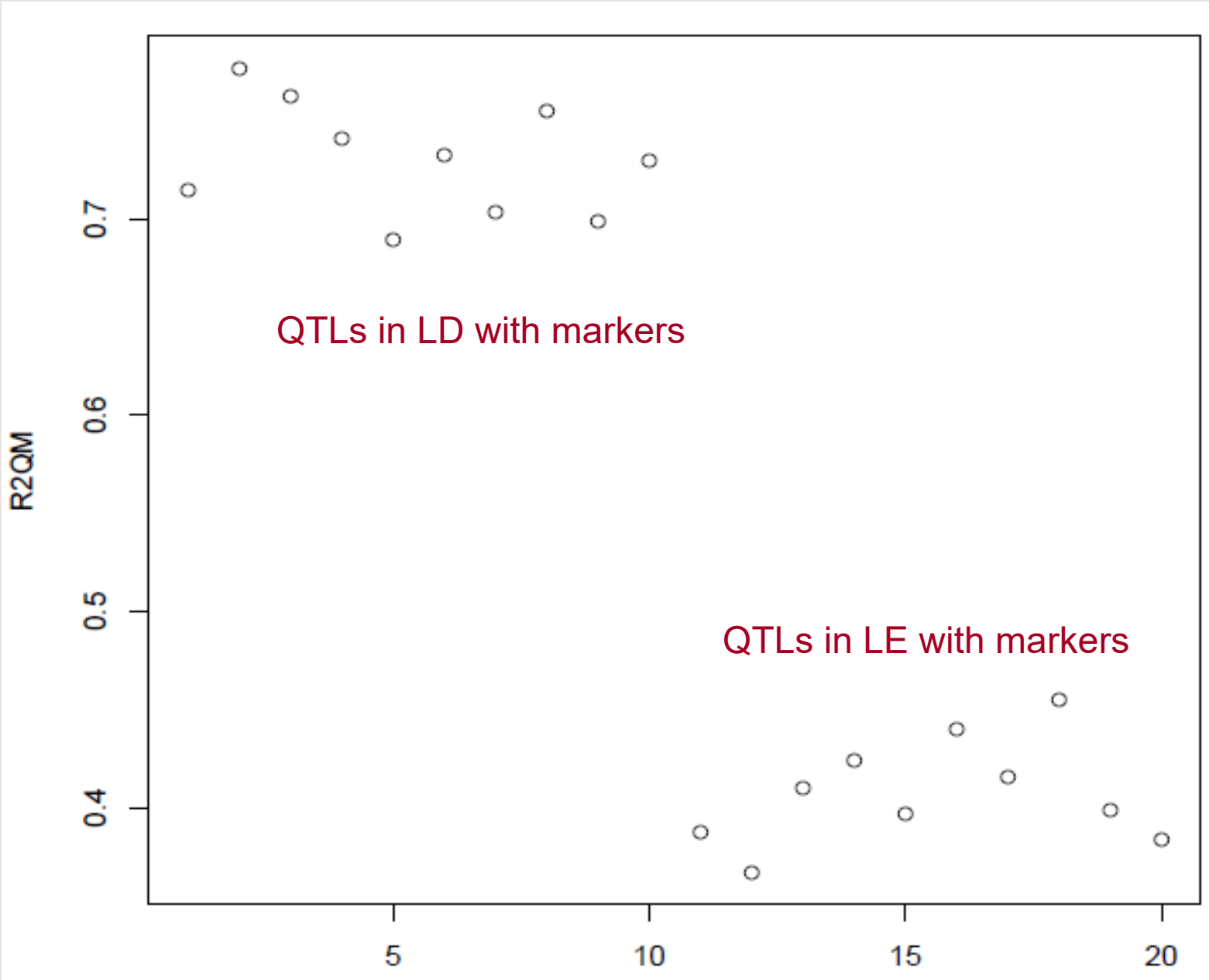

Diagonals of GVR and GSTD relationship matrices
GSTD
0.6
1.0
1.4
1.8
0.8
1.0
1.2
1.4
GVR
Off−diagonals of GVR and GSTD relationship matrices
GSTD
0.0
1.0
0.0
0.5
1.0
GVR
Diagonals of A, GVR and GSTD relationship matrices
GVR (red) GSTD (blue)
0.8
1.0
1.2
1.4
1.6
1.7
1.8
1.9
2.0
A
Off−diagonals of A, GVR and GSTD relationship matrices
GVR (red) GSTD (blue)
0.0
0.5
1.0
0.0
0.5
1.0
1.5
2.0
A

MARKERS ARE NOT QTL: a disconnect

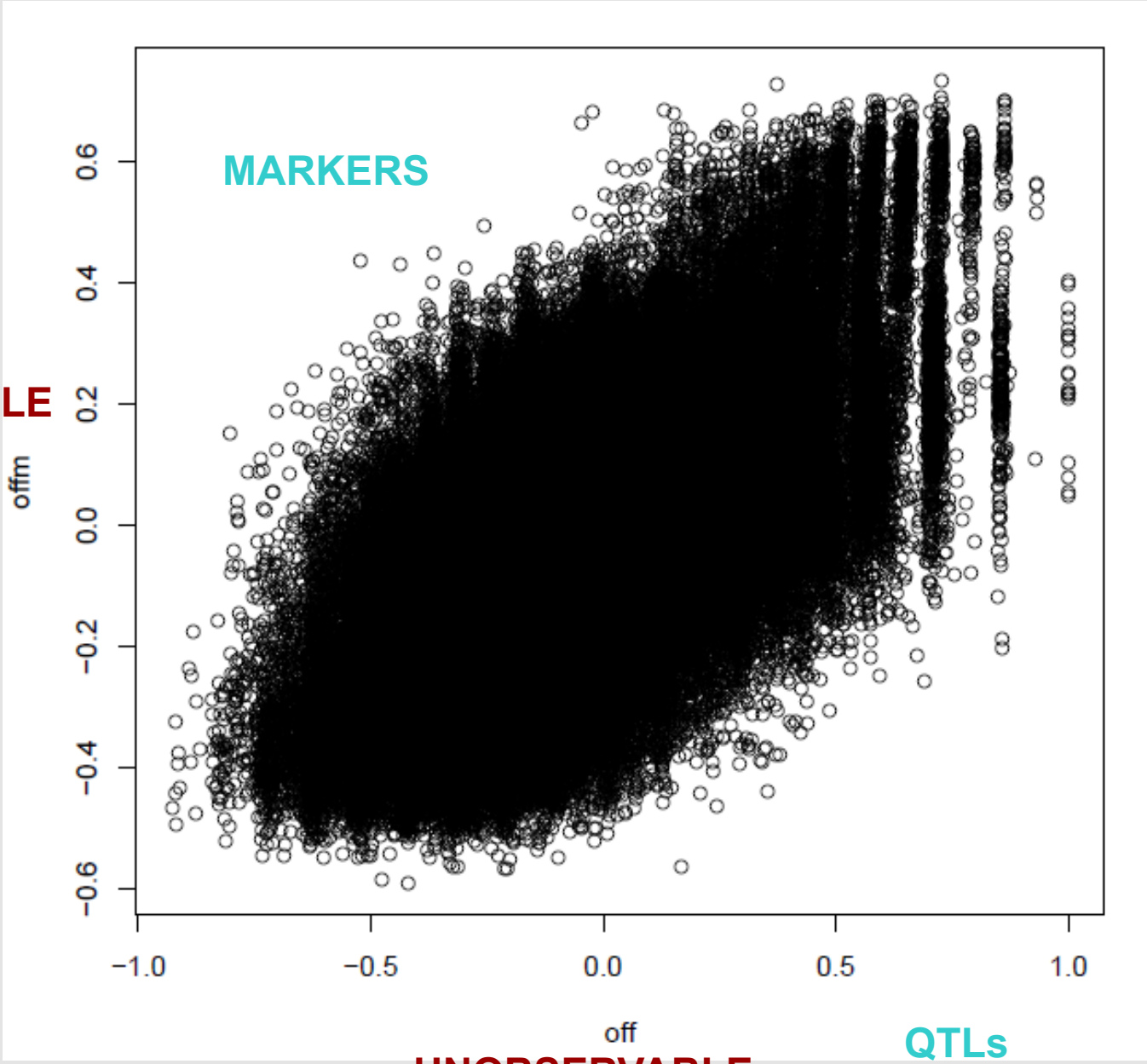
- Nqtl=20 Nmarkers=200
- $N_{Train} = 500$
- First 10 QTLs in LD and in LD with first 100 markers. QTLs 11-20 in LE with everything else
- First 100 markers in LD; other markers in LE

VARIATION IN QTL GENOTYPES EXPLAINED BY MARKERS

(R^2 of QTL genotype on all markers genotypes)



Off-diagonals of “Genomic correlations” among 500 individuals



OBSERVABLE

UNOBSERVABLE

QTLs

OPINION

- **Much "ad-hocquery" on how G ought to be constructed**
- **Impacts mostly discussion of "genetic architecture" from inferential perspectives**

[SKEPTICISM HERE]

- **It may (may not) be that form of G impacts prediction, but using CV one can see what "works" or what "does not work" but tentatively and with uncertainty!**

BACK TO BASIC SETTING: linear regression on markers

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\
\mathbf{X} &: n \times p \text{ matrix of marker genotypes} \\
& \boldsymbol{\beta}|\sigma^2_\beta \tilde{} N\left(0, \mathbf{I}\sigma^2_\beta\right) \\
\mathbf{e}|\sigma^2_e &\sim N\left(0, \mathbf{I}\sigma^2_e\right) \\
& \text{NORMALITY NOT NEEDED FOR BLUP} \\
\mathbf{y} &\sim N\left(\mathbf{0}, \mathbf{X}\mathbf{X}'\sigma^2_\beta + \mathbf{I}\sigma^2_e\right)
\end{aligned}
$$

$$BLUP(\boldsymbol{\beta}) = Cov(\boldsymbol{\beta}, \mathbf{y}')Var^{-1}(\mathbf{y})\left[\mathbf{y} - \hat{E}(\mathbf{y})\right]$$

$$\text{BRUTE FORCE 1} = \sigma^2_\beta \mathbf{X}'[\mathbf{X}\mathbf{X}'\sigma^2_\beta + \mathbf{I}\sigma^2_e]^{-1}\mathbf{y}$$

$$\text{BRUTE FORCE 2} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\left[\mathbf{I} + (\mathbf{X}\mathbf{X}')^{-1}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1}\mathbf{y}$$

$$\text{USE "MIXED MODEL EQS"} = \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma^2_e}{\sigma^2_\beta}\right)^{-1}\mathbf{X}'\mathbf{y}$$

BRUTE FORCE: invert n x n and then map onto p x n

MME: invert p x p

NO COMPELLING REASON FOR MME HERE

Equivalent model

$$\mathbf{g} = \mathbf{X}\boldsymbol{\beta}$$

MARKED GENOTYPIC VALUE

$$\mathbf{g} \sim N\left(\mathbf{0}, \mathbf{X}\mathbf{X}'\sigma^2_\beta\right) = N\left(\mathbf{0}, \frac{1}{k}\mathbf{X}\mathbf{X}'k\sigma^2_\beta\right)$$

$$= N\left(\mathbf{0}, \mathbf{G}_k\sigma^2_{g(k)}\right)$$

CLEARLY $\mathbf{G}_k$ AND $\sigma^2_{g(k)}$ DEPEND ON CHOICE OF k

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

BRUTE FORCE :

$$BLUP(\mathbf{g}) = Cov\left(\mathbf{g}, \mathbf{y}'\right) Var^{-1}\left(\mathbf{y}\right)\left(\mathbf{y} - \left(E\left(\mathbf{y}\right)\right)\right)$$

$$= \mathbf{G}_k\sigma^2_{g(k)}\left[\mathbf{G}_k\sigma^2_{g(k)} + \mathbf{I}\sigma^2_e\right]^{-1}\mathbf{y}$$

$$= \left[\mathbf{I} + \mathbf{G}_k^{-1}\frac{\sigma^2_e}{\sigma^2_{g(k)}}\right]^{-1}\mathbf{y}$$

MME :

$$\left[\mathbf{I} + \mathbf{G}_k^{-1}\frac{\sigma^2_e}{\sigma^2_{g(k)}}\right]^{-1}\mathbf{y}$$

SAME RESULT: BOTH n x n COMPUTATIONS REQUIRED

Estimate marker effects from genomic BLUP? Use standard BLUP theory under normality!

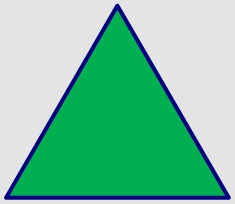
$$E(\boldsymbol{\beta}|\mathbf{y}) = E_{g|y}E_{\beta|g}(\boldsymbol{\beta})$$ "ITERATED EXPECTATIONS"

$$\hat{g} = E\left(X\beta|y, \text{variance components}\right) \text{ under normality}$$

$$\begin{aligned} E(\beta|X\beta) &= E(\beta) + Cov(\beta, \beta'X')[Var(X\beta]^{-1}[X\beta - E(X\beta)] \\ &= 0 + \sigma^2_\beta X'(XX')^{-1}\frac{1}{\sigma^2_\beta}X\beta \end{aligned}$$

$$\begin{aligned} E\left(\beta|y, \text{variance components}\right) &= E_{X\beta|y}[E(\beta|X\beta, y)] \\ &= E_{X\beta|y}\left[X'(XX')^{-1}X\beta|y\right] \\ &= X'(XX')^{-1}E_{X\beta|y}[X\beta|y] = X'(XX')^{-1}\hat{g} \end{aligned}$$

$$\hat{\beta} = E\left(\beta|y, \text{variance components}\right) = X'(XX')^{-1}\left[I + (XX')^{-1}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1}y$$



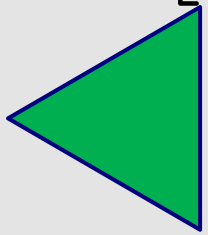
[REMEMBER THIS]

BRUTE FORCE DEFINITION: BLUP is a conditional expectation under normality

$$\hat{\beta} = E\left(\beta | y, \text{variance components}\right) = Cov(\beta, \beta' X')\left[XX'\sigma_\beta^2 + I\sigma_e^2\right]^{-1} y$$

$$= \sigma_\beta^2 X'\left[XX'\sigma_\beta^2 + I\sigma_e^2\right]^{-1} y = \sigma_\beta^2 X'(XX')^{-1}\left[\sigma_\beta^2 + (XX')^{-1}\sigma_e^2\right]^{-1} \text{y}$$

$$= X'(XX')^{-1}\left[I + (XX')^{-1}\frac{\sigma_e^2}{\sigma_\beta^2}\right]^{-1} \text{y}$$



[REMEMBER?]

CAN GO BACK AND FORTH BETWEEN GENOMIC BLUP AND RIDGE REGRESSION ESTIMATES OF MARKER EFFFECTS

$$\hat{\beta} = X'(XX')^{-1}\hat{g}$$

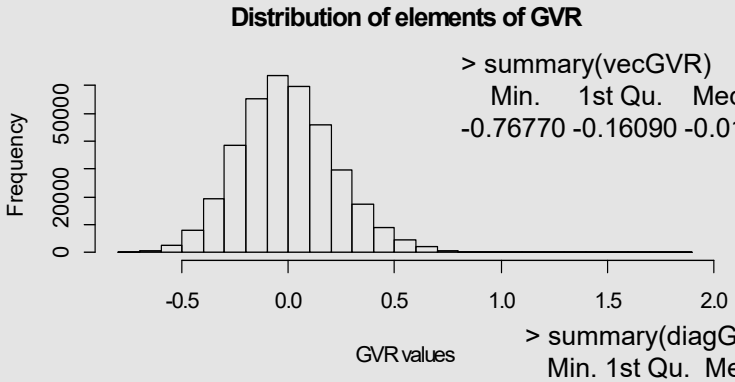
$$\hat{g} = X\hat{\beta}$$

BACK TO GENOMIC BLUP

When should a specific representation of GBLUP be used?
Suppose p<n. Then **G** has at most rank=n and the inverse of G does not exist

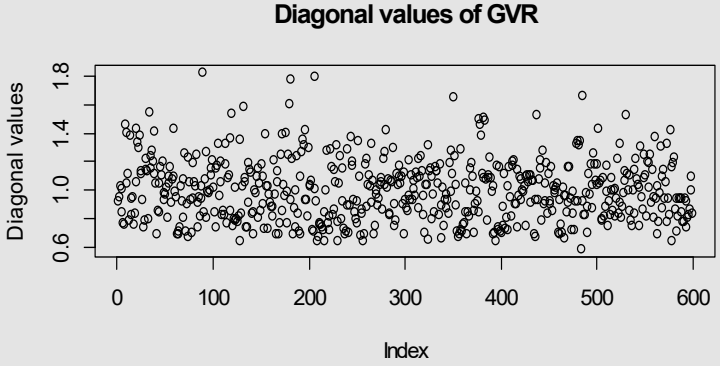
```
rm(list=ls(all=TRUE))

###LOAD LATTICE AND MATRIX
library(MASS)
library(BGLR)
library(lattice)
library(Matrix)
set.seed(1234567)

###LOAD DATA
data(wheat)
Y<-wheat.Y
X<-wheat.X
y<-Y[,1]
n<-length(y)

X<-X[,1:50]

freq<-numeric(ncol(X))
for (j in 1: ncol(X)){
freq[j]<-mean(X[,j])
}

X<-scale(X, center = TRUE, scale = FALSE)

###Markers are binary so var of marker codes is p(1-p)
###instead of 2p(1-p) per locus

varHW<-sum(freq*(1-freq))
varHW
[1] 8.438131
```

```
####GVR= genomic relationship a la Van Raden (2008)
GVR<-X%*%t(X)/varHW
par(mfrow=c(2,1))
vecGVR<-as.vector(GVR)
hist(GVR,main="Distribution of elements of GVR",xlab="GVR values")
diagGVR<-diag(GVR)
plot(diagGVR,ylab="Diagonal values",main="Diagonal values of
GVR")
par(mfrow=c(1,1))
```




```
> summary(vecGVR)
    Min.  1st Qu.   Median     Mean  3 rd Qu.     Max.
-0.76770 -0.16090 -0.01295  0.00000  0.14200  1.82600
```

```
> summary(diagGVR)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5845  0.8344  0.9899  1.0000  1.1310  1.8260
```



ISSUE HERE: SCALE DIFFERS FROM THAT OF A!

Calculation of GBLUP (once one has arrived at some **G**)

$$
\begin{aligned}
\hat{\mathbf{g}} &= E(\mathbf{g}) + Cov(\mathbf{g}, \mathbf{y}')Var(\mathbf{y})^{-1}[\mathbf{y} - E(\mathbf{y})] \\
&= \mathbf{G}\sigma_g^2(\mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2)^{-1}\mathbf{y} \\
&= \mathbf{G}(\mathbf{G} + \mathbf{I}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\mathbf{y} \\
&= (\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\mathbf{y}
\end{aligned}
$$

$$
\begin{aligned}
Var(\mathbf{g}|\mathbf{y}) &= Var(\mathbf{g} - \hat{\mathbf{g}}) \\
&= \mathbf{G}\sigma_g^2 - \mathbf{G}\sigma_g^2(\mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2)^{-1}\mathbf{G}\sigma_g^2 \\
&= \mathbf{G}\sigma_g^2 - \mathbf{G}(\mathbf{G} + \mathbf{I}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\mathbf{G}\sigma_g^2 \\
&= \mathbf{G}\sigma_g^2 - (\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\mathbf{G}\sigma_g^2 \\
&= (\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\left[(\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2})\mathbf{G}\sigma_g^2 - \mathbf{G}\sigma_g^2\right] \\
&= (\mathbf{I} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\sigma_e^2
\end{aligned}
$$

```
####Does GVR have an inverse in this case? No, rank(GVR) should be 50
> GVRinv<-chol2inv(chol(GVR))
Error in chol2inv(chol(GVR)) :
  error in evaluating the argument 'x' in selecting a method for function 'chol2inv': Error in
chol.default(GVR) :
  the leading minor of order 38 is not positive definite
> rankMatrix(GVR)
Warning in rankMatrix(GVR) :
  rankMatrix(<large sparse Matrix>, method = 'tolNorm2') coerces to dense matrix.
  Probably should rather use  method = 'qrLINPACK' !?
[1] 50
attr(,"method")
[1] "tolNorm2"
attr(,"useGrad")
[1] FALSE
attr(,"tol")
[1] 5.835619e-11
```

MUST USE “STRONG ARM” FOR CALCULATING GBLUP

$$\hat{\mathbf{g}} = \mathbf{G}_{VR}(\mathbf{G}_{VR} + \mathbf{I}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\mathbf{y}$$

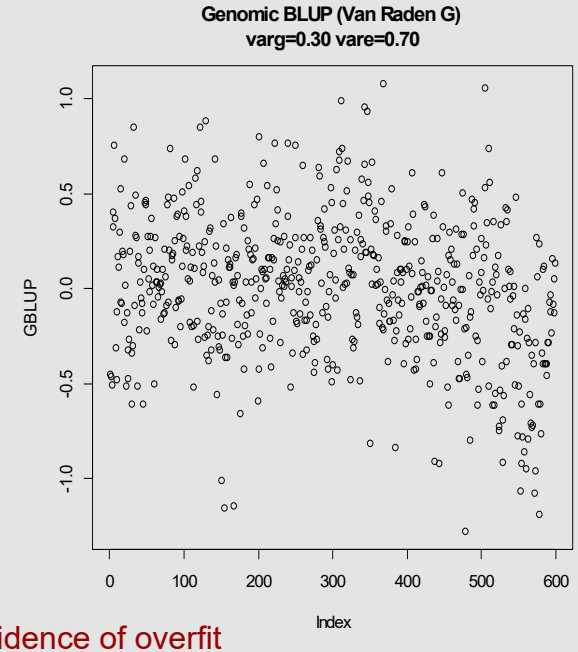
$$Var(\mathbf{g} - \hat{\mathbf{g}}) = \mathbf{G}_{VR}\sigma_g^2 - \mathbf{G}_{VR}(\mathbf{G}_{VR} + \mathbf{I}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\mathbf{G}_{VR}\sigma_g^2$$

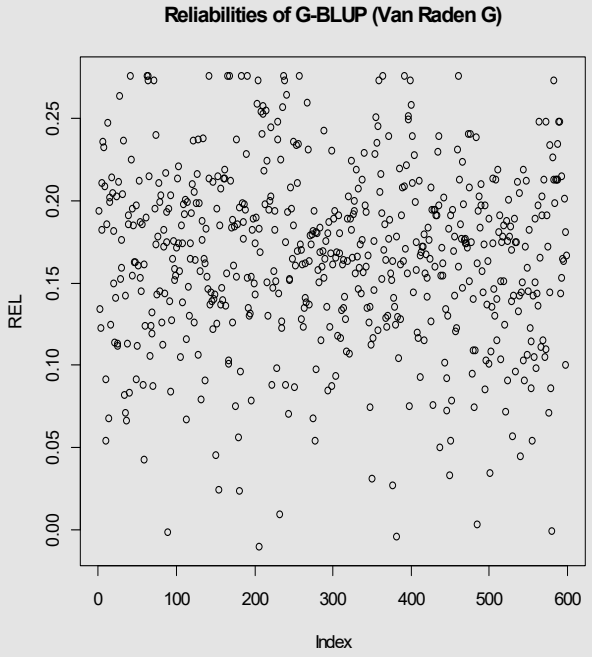
$$= \left[\mathbf{I} - \mathbf{G}_{VR}(\mathbf{G}_{VR} + \mathbf{I}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\right]\mathbf{G}_{VR}\sigma_g^2$$

Suppose

$$\sigma_g^2 = 0.30; \frac{\sigma_e^2}{\sigma_g^2} = 2.3333$$

```
####Compute GBLUP using the strong-arm method.
####varg=0.30,vare=0.70
####lambda<-2.3333
varg=0.30
vare=0.70
lambda=vare/varg
Vstar<-(GVR+lambda*diag(n))
Vstarinv<-chol2inv(chol(Vstar))
ghat<-GVR%*%Vstarinv%*%y
plot(ghat,ylab="GBLUP",main="Genomic BLUP (Van Raden G)
varg=0.30 vare=0.70")

###Compute prediction error variance covariance matrix
PEVMAT<-varg*(diag(n)-GVR%*%Vstarinv)%*%GVR

###CALCULATE MODEL DERIVED RELIABILITIES
RELS<-varg*diag(n)-diag(PEVMAT)/varg
RELGBLUPS<-diag(RELS)
plot(RELGBLUPS,ylab="REL",main="Reliabilities of G-BLUP (Van Raden G)")
```

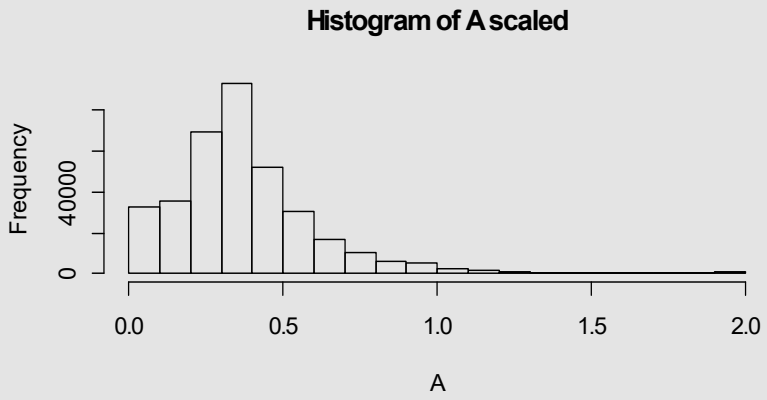


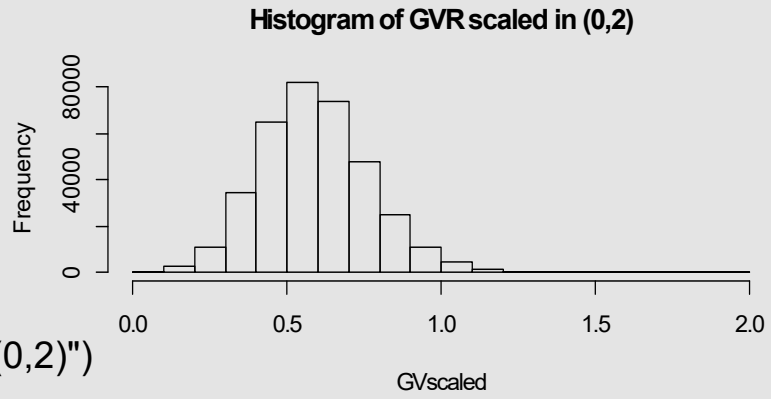


No evidence of overfit

0.4646628

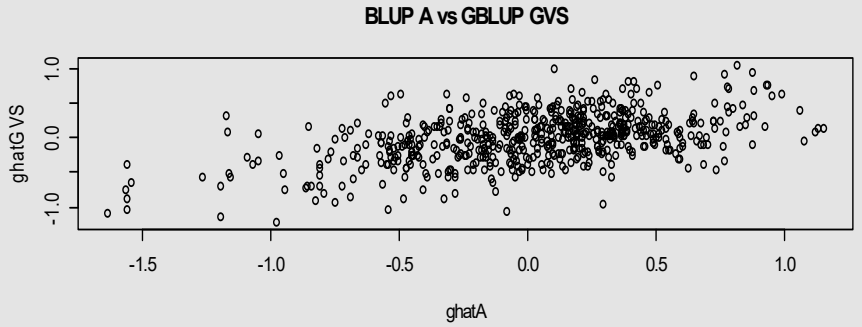
```
#####Impact of G-matrix on GBLUP
#####Assume same variance decomposition
#####Scale to be in (0,2)
GVscaled<-matrix(nrow=nrow(X),ncol=nrow(X))
VRmin<-min(GVR)
VRmax<-max(GVR)

VRmin

VRmax

for (i in 1:nrow(X)){
for (j in 1:nrow(X)){
GVscaled[i,j]<-2*(GVR[i,j]-VRmin)/(VRmax-VRmin)
}
}
```

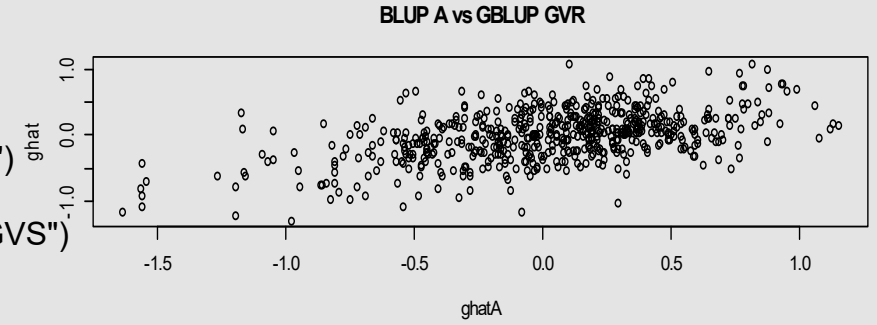


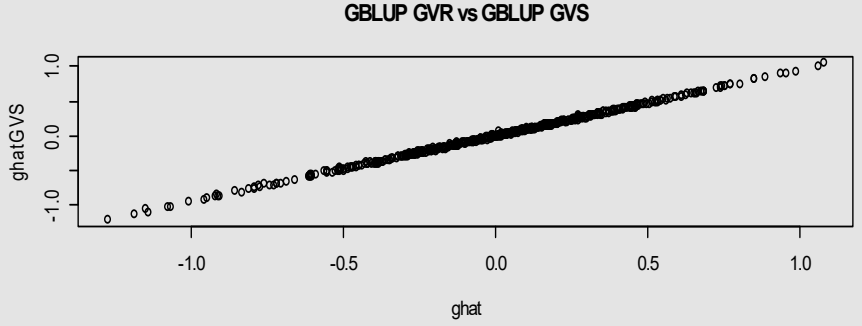



```
####How does it compare with A?
A<-wheat.A
par(mfrow=c(2,1))
hist(A,main="Histogram of A scaled")
hist(GVscaled,main="Histogram of GVR scaled in (0,2)")

par(mfrow=c(1,1))
cor(as.vector(A),as.vector(GVscaled))
[1] 0.2381
```

```
#####BLUP (assume save var decomposition)
varg=0.30
vare=0.70
lambda=vare/varg

VstarGVS<-(GVscaled+lambda*diag(n))
VstarinvGVS<-chol2inv(chol(VstarGVS))
ghatGVS<-GVscaled%*%VstarinvGVS%*%y


VstarA<-(A+lambda*diag(n))
VstarinvA<-chol2inv(chol(VstarA))
ghatA<-A%*%VstarinvA%*%y

par(mfrow=c(3,1))
plot(ghatA,ghatGVS,main="BLUP A vs GBLUP GVS")
plot(ghatA,ghat,main="BLUP A vs GBLUP GVR")
plot(ghat,ghatGVS,main="GBLUP GVR vs GBLUP GVS")
par(mfrow=c(1,1))
```








```
> cor(ghatA,ghat)
          [,1]
[1,] 0.5020861
> cor(ghatA,ghatGVS)
          [,1]
[1,] 0.500235
> cor(ghat,ghatGVS)
          [,1]
[1,] 0.9994833
```

```
mseA<-sum((y-ghatA)**2)/n
msehat<-sum((y-ghat)**2)/n
msehatGVS<-sum((y-ghatGVS)**2)/n
mseA
msehat
msehatGVS
```

```
> mseA
[1] 0.4245183
> msehat
[1] 0.7907524
> msehatGVS
[1] 0.7964134
```

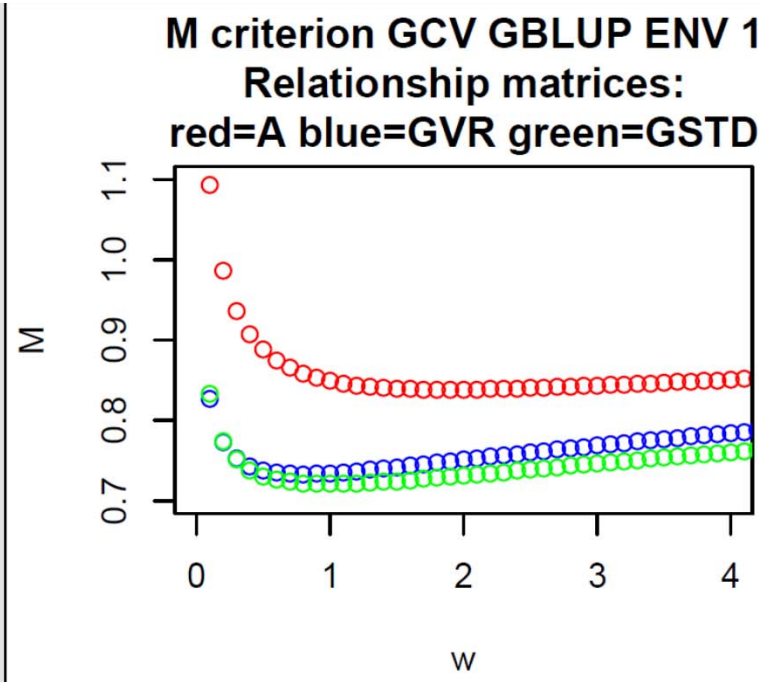
BLUP(A) FITS BETTER
(may predict worse)

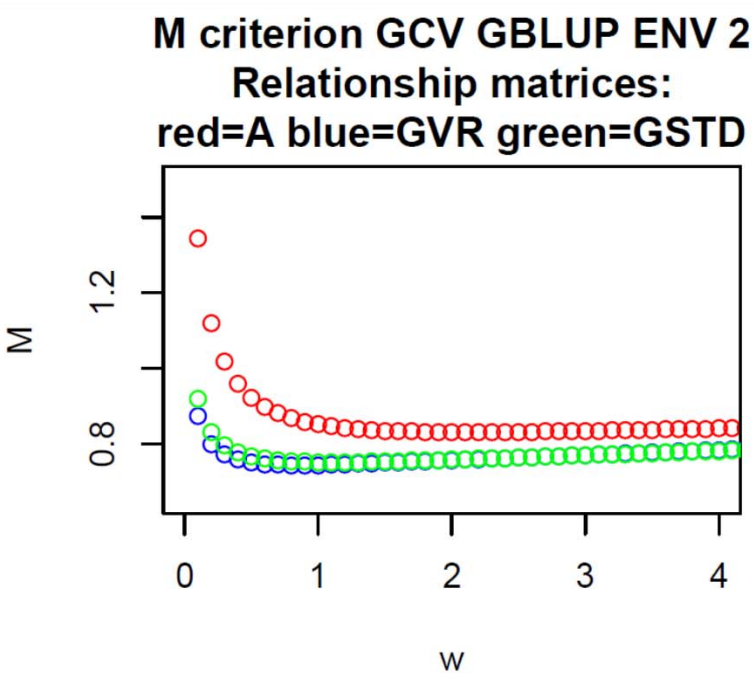
GENERALIZED CV IN GBLUP (zero-means model, wheat data)

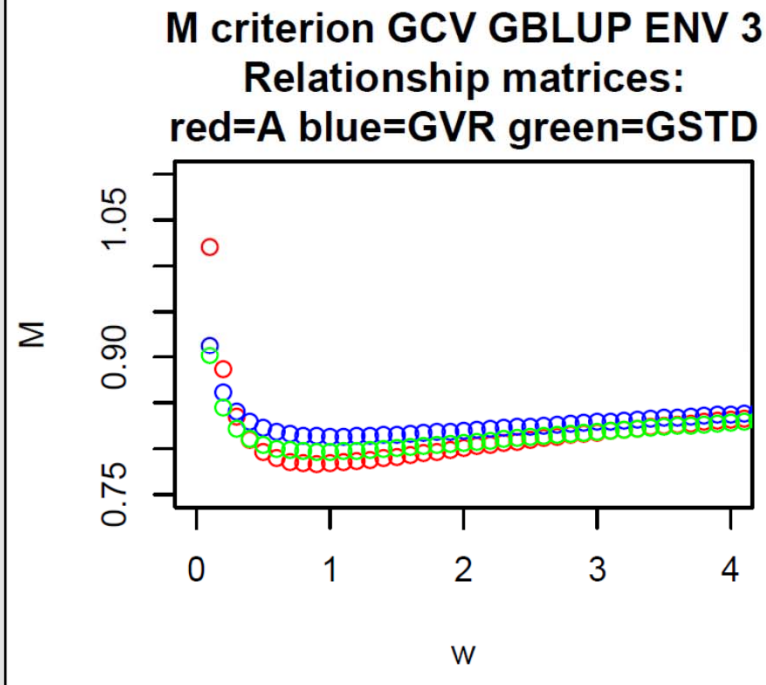
Given some genomic relationship matrix, a value of $\omega = \frac{\sigma_e^2}{\sigma_g^2}$ is needed for computing GBLUP; let the prediction be $\widehat{\mathbf{g}}(\omega)$, the prediction residual be $\mathbf{y} - \widehat{\mathbf{g}}(\omega) = \left[\mathbf{I} - \left(\mathbf{I} + \mathbf{G}^{-1}\omega\right)^{-1}\right]\mathbf{y}$ and $\mathbf{H}(\omega) = \left(\mathbf{I} + \mathbf{G}^{-1}\omega\right)^{-1}$. The $M-$criterion for generalized cross-validation given in (78) becomes

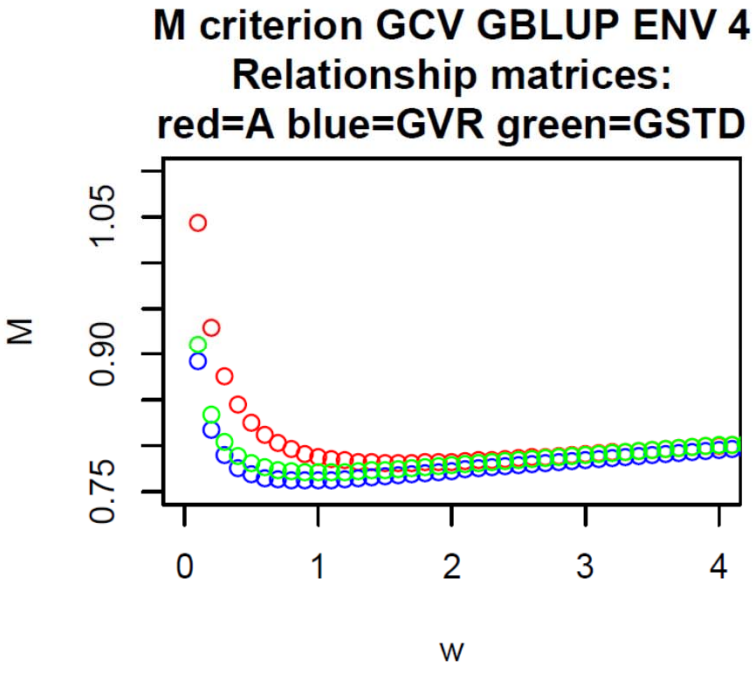
$$\begin{aligned}\mathbf{M}(\omega) &= \frac{1}{n}\left(\mathbf{y} - \widehat{\mathbf{g}}(\omega)\right)'\left(\mathbf{y} - \widehat{\mathbf{g}}(\omega)\right) \Big/ \left\{\frac{1}{n} tr\left(\mathbf{I} - \left(\mathbf{I} + \mathbf{G}^{-1}\omega\right)^{-1}\right)\right\}^2 \\ &= \frac{1}{n}\mathbf{y}'\left[\mathbf{I} - \left(\mathbf{I} + \mathbf{G}^{-1}\omega\right)^{-1}\right]^2 \mathbf{y} \Big/ \left\{1 - \overline{h}(\omega)\right\}^2; \qquad (89)\end{aligned}$$

where $\overline{h}(\omega)$ is the average of the diagonal elements of $\mathbf{H}(\omega)$.


M criterion GCV GBLUP ENV 1
Relationship matrices:
red=A blue=GVR green=GSTD
M
0.7
0.8
0.9
1.0
1.1
0
1
2
3
4
w


M criterion GCV GBLUP ENV 2
Relationship matrices:
red=A blue=GVR green=GSTD
M
0.8
1.2
0
1
2
3
4
w


M criterion GCV GBLUP ENV 3
Relationship matrices:
red=A blue=GVR green=GSTD
M
0.75
0.90
1.05
0
1
2
3
4
w


M criterion GCV GBLUP ENV 4
Relationship matrices:
red=A blue=GVR green=GSTD
M
0.75
0.90
1.05
0
1
2
3
4
w

BAYESIAN GBLUP

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{g} + \mathbf{e}$$

Priors :

$\rightarrow \mu \propto$ constant ("flat" prior)

$\rightarrow \sigma_e^2 | \nu, S^2 \sim \nu S^2 \chi_\nu^{-2}; \nu = 5, S^2 = \frac{\sigma_y^2}{2}$ Known

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

$\rightarrow \sigma_g^2 | \nu_g, S_g^2 \sim \nu_g S_g^2 \chi_{\nu_g}^{-2}; \nu_g = 4.5, S_g^2 = h^2 S^2$ (set $h^2 = 0.25$)

Arbitrary choices
used for illustration

WHAT IS THE MARGINAL PRIOR OF $\mathbf{g}$ IN BAYESIAN GBLUP?

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

$$\sigma_g^2|\nu_g, S_g^2 \sim \nu_g S_g^2 \chi_{\nu_g}^{-2}$$

$$p(\mathbf{g}|\nu_g, S_g^2) = \int p(\mathbf{g}|\sigma_g^2) p(\sigma_g^2|\nu_g, S_g^2) d\sigma_g^2$$

$$= \frac{\Gamma\left(\frac{\nu_g+n}{2}\right)}{\Gamma\left(\frac{\nu_g}{2}\right)(\nu_g\pi)^{\frac{n}{2}}|\mathbf{G}|^{\frac{1}{2}}}\left[1 + \frac{\mathbf{g}'\mathbf{G}^{-1}\mathbf{g}}{\nu_g S_g^2}\right]^{-\left(\frac{n+\nu_g}{2}\right)}$$

$$= Multivariate - t_n\left(\mathbf{0}, \frac{\nu_g}{\nu_g - 2} S_g^2 \mathbf{G}\right)$$

$\Rightarrow$ SUPPOSE THE n INDIVIDUALS ARE MOLECULARLY

$\Rightarrow$ DISSIMILAR SO $\mathbf{G}$ =diagonal.

$\Rightarrow$ DO THE INDIVIDUALS INFORM ABOUT EACH OTHER?

NO!!!!!!!!!!

$\Rightarrow$ SIGNALS NOT INDEPENDENT A PRIORI

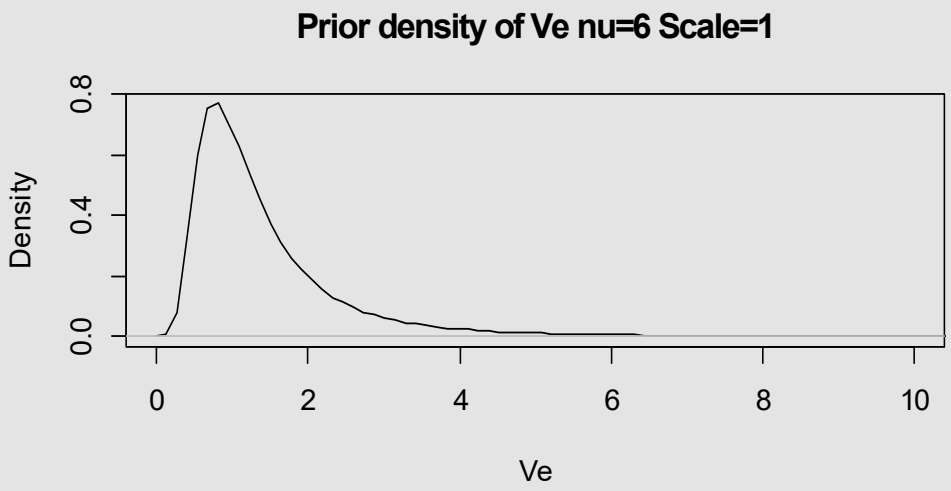
$\Rightarrow$ WHY? BECAUSE JOINT DENSITY CANNOT BE WRITTEN AS PRODUCT OF MARGINALS!!

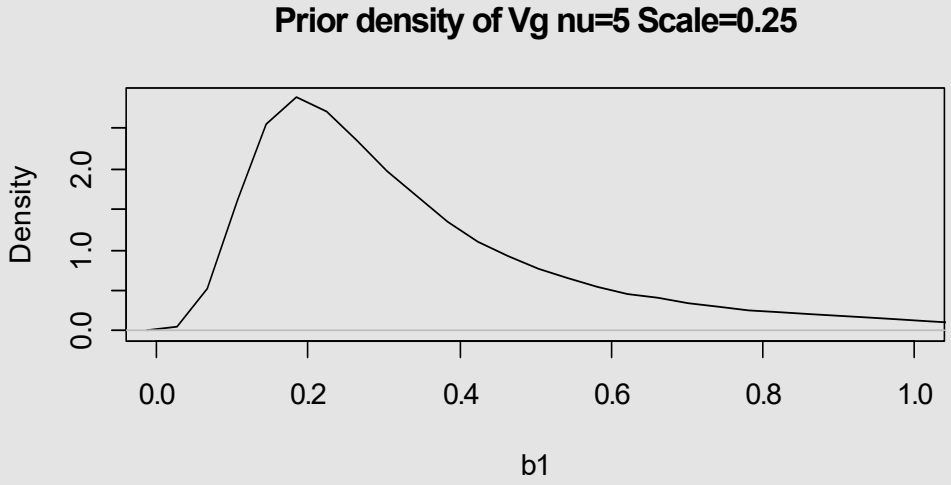
$\Rightarrow$ UNCORRELATED RANDOM VARIABLES DO NOT NECESSARILY IMPLY INDEPENDENCE

EXAMPLE OF BAYESIAN GBLUP: DATA AND PRIOR DISTRIBUTIONS

```
###USE GIBBS SAMPLER
###GENOMIC RELATIONSHIP MATRIX CONSTRUCTED WITH
###CENTERED AND SCALED MATRIX
###UNKNOWN INTERCEPT, GENETIC SIGNAL (MARKED BREEDING
VALUE)
###VG AND VE
###WHEAT DATA
rm(list=ls(all=TRUE))
library(MASS)
library(BGLR)
library(lattice)
library(Matrix)
set.seed(1234567)

###LOAD DATA
data(wheat)
Y<-wheat.Y
X<-wheat.X
y<-Y[,1]
n<-length(y)

####LOAD A MATRIX
A<-wheat.A

####USE 100 MARKERS AND SCALE X
X<-scale(X[,1:100])
p<-ncol(X)

####FORM GENOMIC RELATIONSHIP MATRIX
G<-X%*%t(X)/p
###SCALE G WITH MAPMINMAX TO BE IN (0,2)
GSC<-matrix(nrow=nrow(X),ncol=nrow(X))
Gmin<-min(G)
Gmax<-max(G)
for (i in 1:nrow(X)){
for (j in 1:nrow(X)){
GSC[i,j]<-2*(G[i,j]-Gmin)/(Gmax-Gmin)
}
}
```

```
###Prior distributions of variances
###Assign arbitrary scaled inverted chi-square to ve (Var(y),6)
###Assign arbitrary scaled inverted chi-square to veg (0.25*Var(y),5)
scale<-var(y)
nue<-6
scalg<-0.25*scale
nug<-5

####PLOT PRIOR DISTRIBUTION OF VARIANCES
####BY DRAWING 100,000 SAMPLES FROM PRIORS
priorve<-nue*scale/rchisq(100000,nue)
priorvg<-nug*scalg/rchisq(100000,nug)
####CONDITIONAL PRIOR OF g given vg is normal
####UNCONDITIONAL IS MULTIVARIATE t
par(mfrow=c(2,1))
plot(density(priorve),main="Prior density of Ve nu=6 Scale=1",xlab="Ve",
xlim=c(0,10))
plot(density(priorvg),main="Prior density of Vg nu=5 Scale=0.25",xlab="b1"
xlim=c(0,1))
par(mfrow=c(1,1))
```





These priors are informative, but not much so

Conditional posteriors

(shown without derivation)

$$\Rightarrow \mu|\mathbf{g},\sigma_e^2,\sigma_g^2,\mathbf{G},\mathbf{y},H \;=\mu|\mathbf{g},\sigma_e^2,\mathbf{y}$$

$$\sim N\left(\frac{1}{n}\mathbf{1}'(\mathbf{y}-\mathbf{g}),\frac{\sigma_e^2}{n}\right)$$

$$\Rightarrow \mathbf{g}|\mu,\sigma_e^2,\sigma_g^2,\mathbf{G},\mathbf{y},H$$

$$\sim N(\hat{\mathbf{g}},\mathbf{V}_{g.cond});\ \hat{\mathbf{g}}=\mathbf{G}\frac{\sigma_g^2}{\sigma_e^2}\left(\mathbf{G}+\frac{\sigma_e^2}{\sigma_g^2}\right)^{-1}(\mathbf{y}-\mathbf{1}\mu);$$

$$\mathbf{V}_{g.cond}=\mathbf{G}\sigma_g^2\left[\mathbf{I}-\mathbf{G}(\mathbf{G}+\mathbf{I}\frac{\sigma_e^2}{\sigma_g^2})^{-1}\right]$$

Use strong-arm method because G is singular

$$\Rightarrow \sigma_g^2|\mathbf{g},\mathbf{y},H\sim(n+\nu_g)\frac{\mathbf{g}'\mathbf{g}+\nu_g S_g^2}{(n+\nu_g)}\chi^{-2}_{(n+\nu_g)}$$

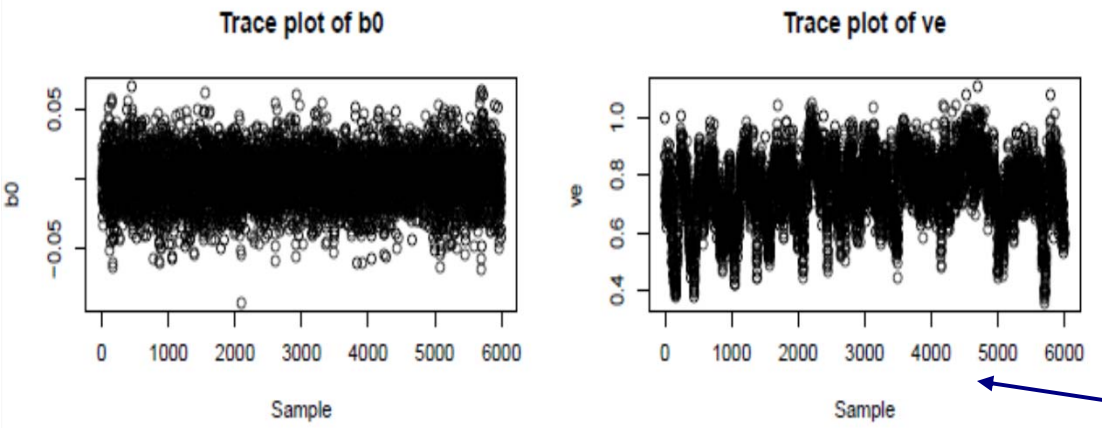
$$\Rightarrow \sigma_e^2|\mu,\mathbf{g},\mathbf{y},H$$

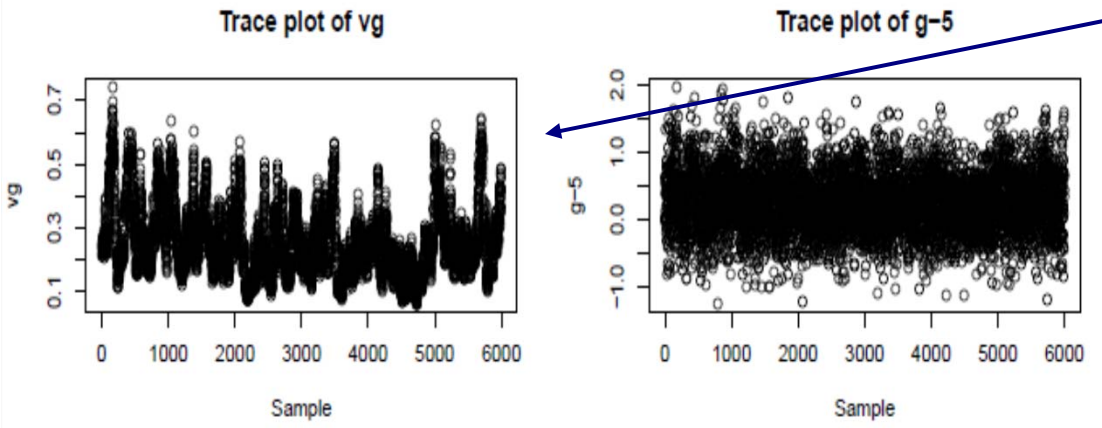
$$\sim (n+\nu)\frac{SSE+\nu S^2}{(n+\nu)}\chi^{-2}_{(n+\nu)}$$

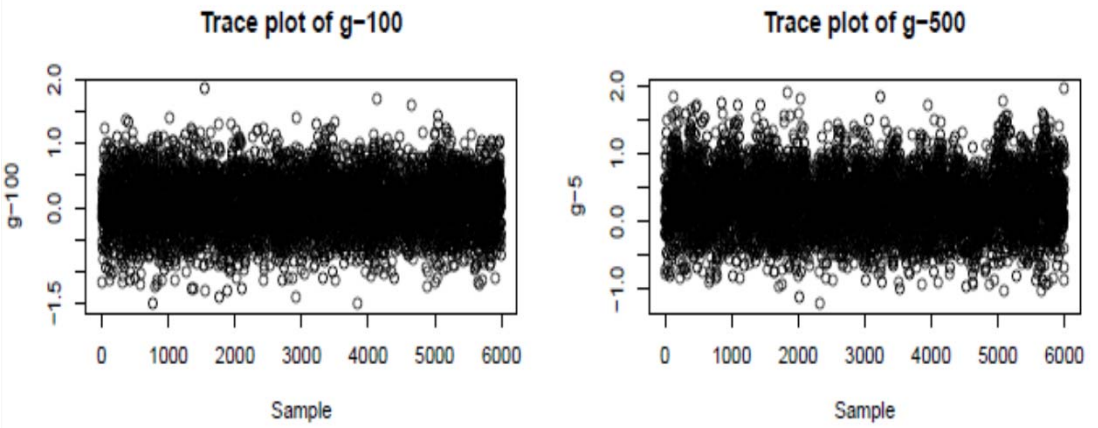
$$SSE=(\mathbf{y}-\mathbf{1}\mu-\mathbf{g})'(\mathbf{y}-\mathbf{1}\mu-\mathbf{g})$$

```
###SET NUMBER OF GIBBS SAMPLES AND BURN IN
###SHOULD USE MORE ITERATIONS AND LONGER BURN-IN IN SERIOUS ANALYSIS
NITER<-6000
BURN<-1000
POSTERIORSAMPLES<-NITER-BURN
###DEFINE CHAIN AND OBJECTS STORING SAMPLES
b0samp<-numeric(NITER)
gsamp<-matrix(nrow=NITER,ncol=n)
vesamp<-numeric(NITER)
vgsamp<-numeric(NITER)
h2genomic<-numeric(NITER)
###CHAIN STARTING VALUES (ITERATION 1)
nparams<-1+p+1+1
b0samp[1]<-0.0001
gsamp[1,]<-rep(0.0001,n)
vesamp[1]<-scale
vgsamp[1]<-scalg
###FORM MATRICES OF SUM OF SQUARES AND PRODUCTS OF
###COLUMNS IN Xmod
###FORM J VECTOR AND INCIDENCE MATRIX
J<-rep(1,n)
Xmod<-cbind(1,diag(n))
XPX0<-crossprod(J)
XPXg<-diag(n)
```

```
###GIBBS SAMPLING (CONDITIONAL POSTERIORS ARE
###PROPOSALS: all accepted). NOTE OFFSETS and
###IMMEDIATE UPDATING

for (i in 2:NITER){
###Sample b0
mean0<-sum(y-gsamp[i-1,])/XPX0
var0<-vesamp[i-1]/XPX0
b0samp[i]<-rnorm(1,mean0,var0)
###Sample g
lambda<-vesamp[i-1]/vgsamp[i-1]
Sigma1<-chol2inv(chol(GSC+diag(n)*lambda))
Sigma2<-diag(n)-GSC%*%Sigma1
Sigma3<-vgsamp[i-1]*GSC%*%Sigma2
mean1<-(1/lambda)*GSC%*%Sigma1%*%(y-J*b0samp[i])
gsamp[i,]<-mvrnorm(1,mean1,Sigma3)
###Sample ve
nuenew<-nue+n
res<-y-J*b0samp[i]-gsamp[i,]
sse<-t(res)%*%res+nue*scale
vesamp[i]<-sse/rchisq(1,nuenew)
###Sample vg
nugnew<-nug+n
ssg<-t(gsamp[i,])%*%gsamp[i,]+nug*scalg
vgsamp[i]<-ssg/rchisq(1,nugnew)
h2genomic[i]<-vgsamp[i]/(vgsamp[i]+vesamp[i])
}
```



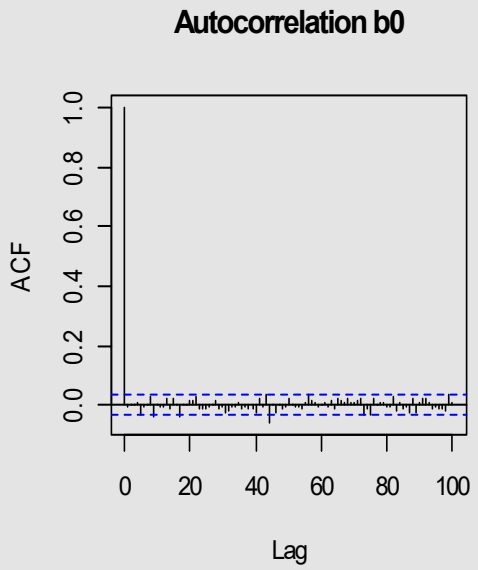


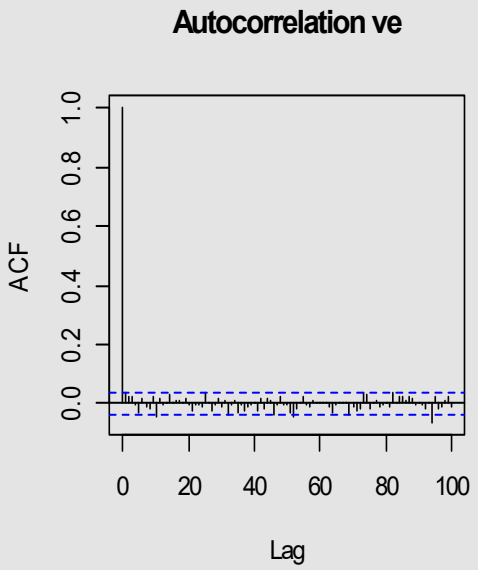


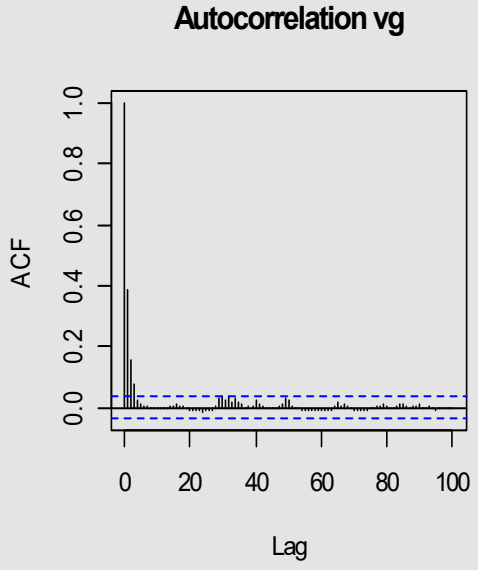

```
####Look at some trace plots

par(mfrow=c(3,2))
plot(b0samp,ylab="b0",xlab="Sample",main="Trace plot of b0")
plot(vesamp,ylab="ve",xlab="Sample",main="Trace plot of ve")
plot(vgsamp,ylab="vg",xlab="Sample",main="Trace plot of vg")
plot(gsamp[,5],ylab="g-5",xlab="Sample",main="Trace plot of g-5")
plot(gsamp[,100],ylab="g-100",xlab="Sample",main="Trace plot of g-100")
plot(gsamp[,500],ylab="g-5",xlab="Sample",main="Trace plot of g-500")
par(mfrow=c(1,1))
```

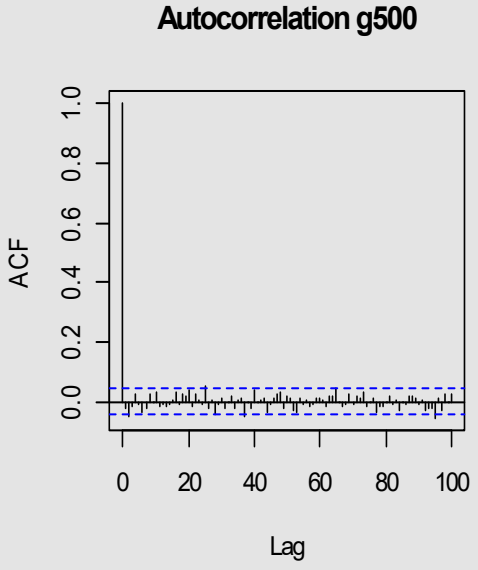
Location effects seem To mix well; variance components less so

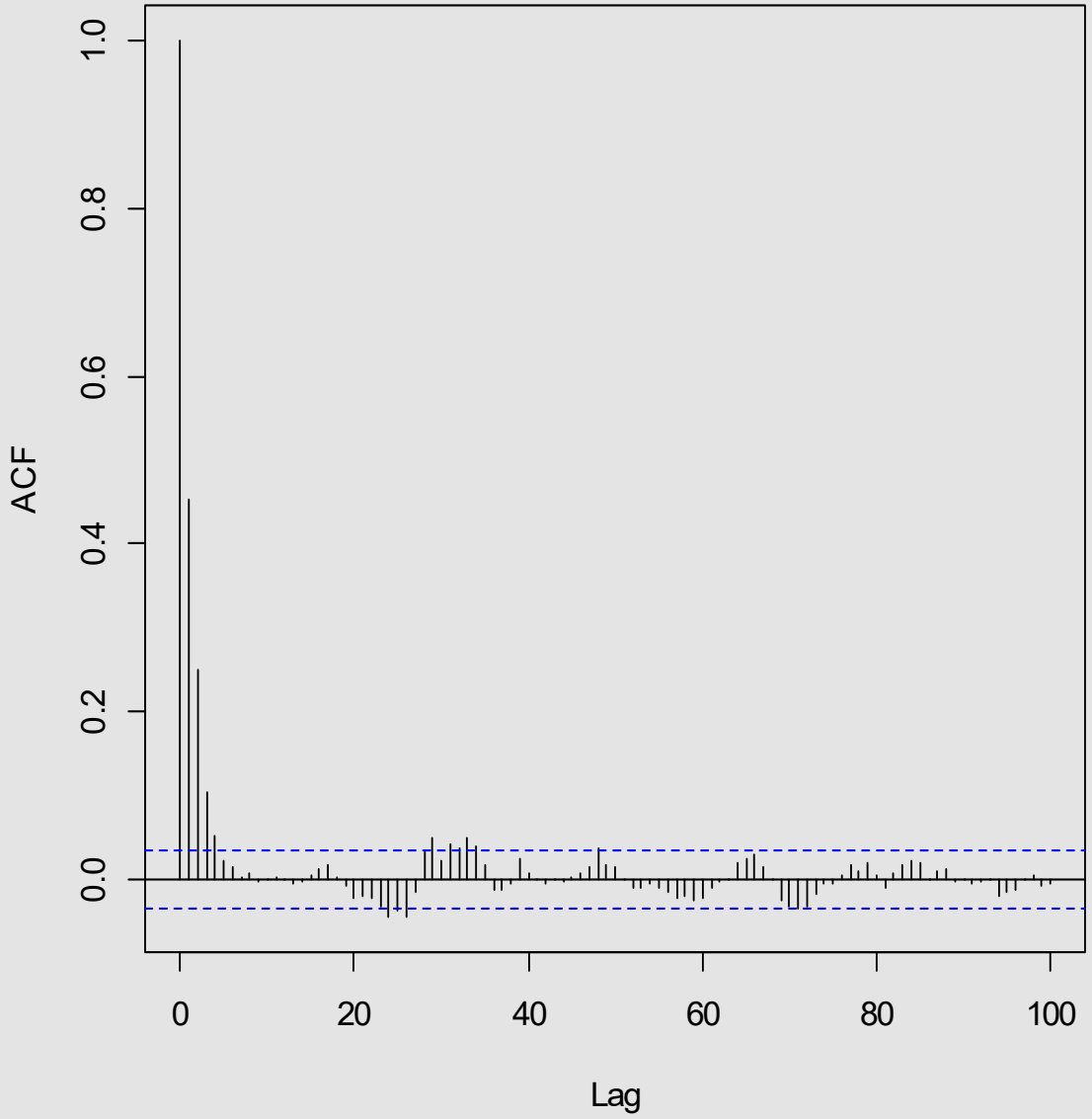
```
#######EVALUATION OF MIXING
par(mfrow=c(2,2))
acf(b0samp, lag.max = 100,
    type = c("correlation"),
    plot = TRUE, na.action = na.fail,main="Autocorrelation b0")
acf(vesamp, lag.max = 100,
    type = c("correlation"),
    plot = TRUE, na.action = na.fail,main="Autocorrelation ve")
acf(vgsamp, lag.max = 100,
    type = c("correlation"),
    plot = TRUE, na.action = na.fail,main="Autocorrelation vg")
acf(gsamp[1001:NITER,500], lag.max = 100,
    type = c("correlation"),
    plot = TRUE, na.action = na.fail,main="Autocorrelation g500")
par(mfrow=c(1,1))

acf(h2genomic, lag.max = 100,
    type = c("correlation"),
    plot = TRUE, na.action = na.fail,main="Autocorrelation h2")
```



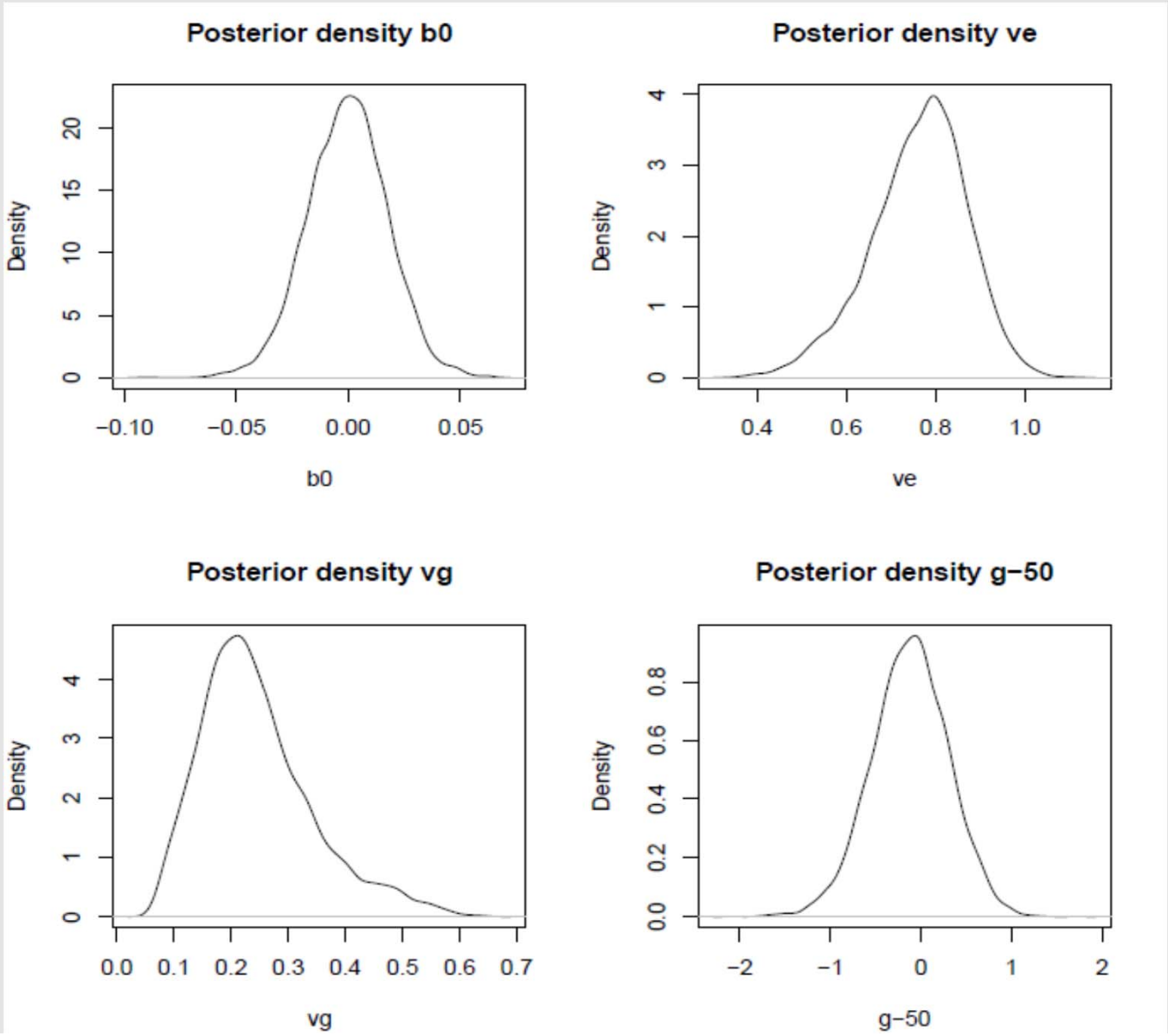




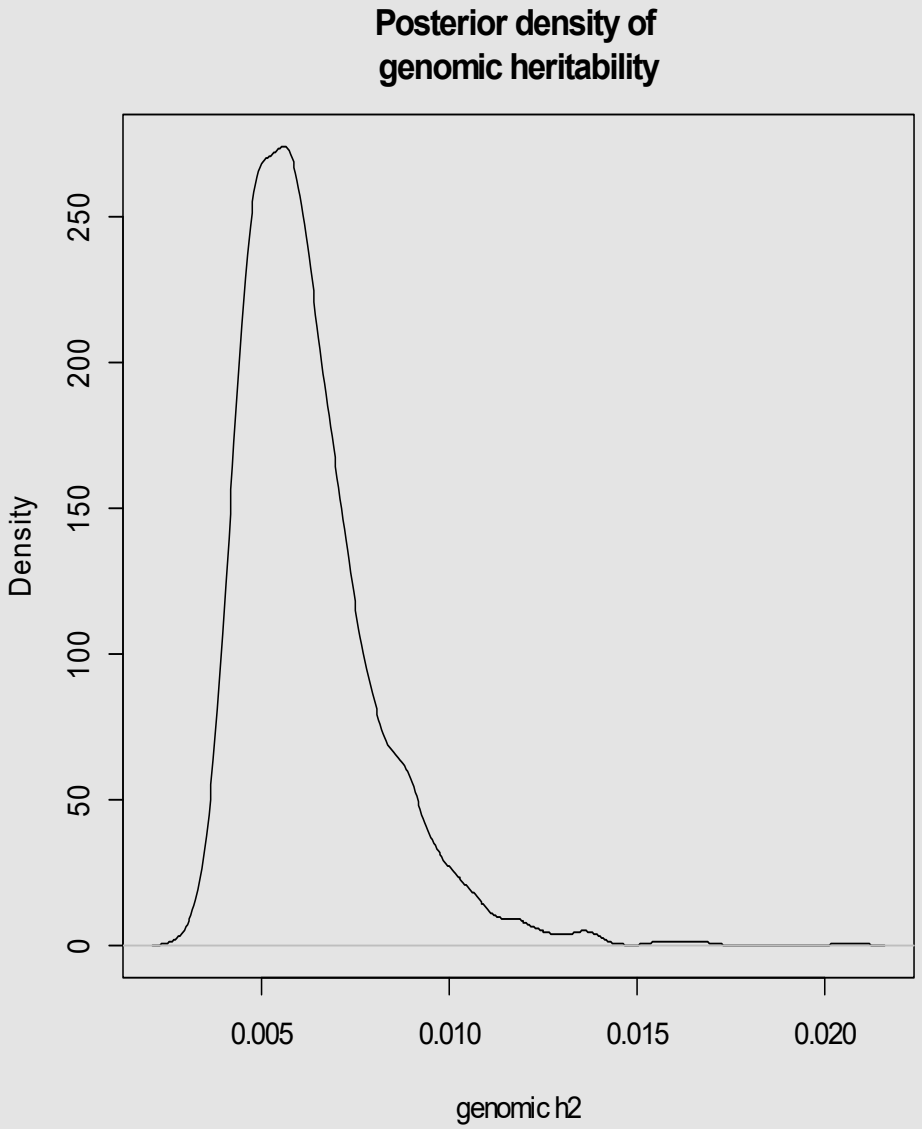



Autocorrelation h2
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Lag

```
par(mfrow=c(2,2))
plot(density(b0samp[1001:NITER]),main="Posterior density b0",xlab="b0")
plot(density(vesamp[1001:NITER]),main="Posterior density ve",xlab="ve")
plot(density(vgsamp[1001:NITER]),main="Posterior density vg",xlab="vg")
plot(density(gsamp[1001:NITER,50]),main="Posterior density g-50",xlab="g-50")
par(mfrow=c(1,1))
```



```
plot(density(h2genomic[1001,NITER]),main="Posterior density of genomic heritability",xlab="genomic h2")
summary(b0samp[1001:NITER])
summary(vesamp[1001:NITER])
summary(vgsamp[1001:NITER])
summary(gsamp[1001:NITER,500])
summary(h2genomic[1001:NITER])
```




```
> summary(b0samp[1001:NITER])
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.0970000 -0.0192000  0.0006125  0.0000910  0.0199500  0.1037000

> summary(vesamp[1001:NITER])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8495  0.9657  1.0040  1.0060  1.0450  1.2170

> summary(vgsamp[1001:NITER])
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.003056 0.005059 0.005961 0.006323 0.007106 0.019760

> summary(gsamp[1001:NITER,500])
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.2142000 -0.0423100  0.0005076  0.0015770  0.0456200  0.2126000

> summary(h2genomic[1001:NITER])
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.003022 0.004964 0.005886 0.006259 0.007074 0.020670
```

Recall: 100 markers only…

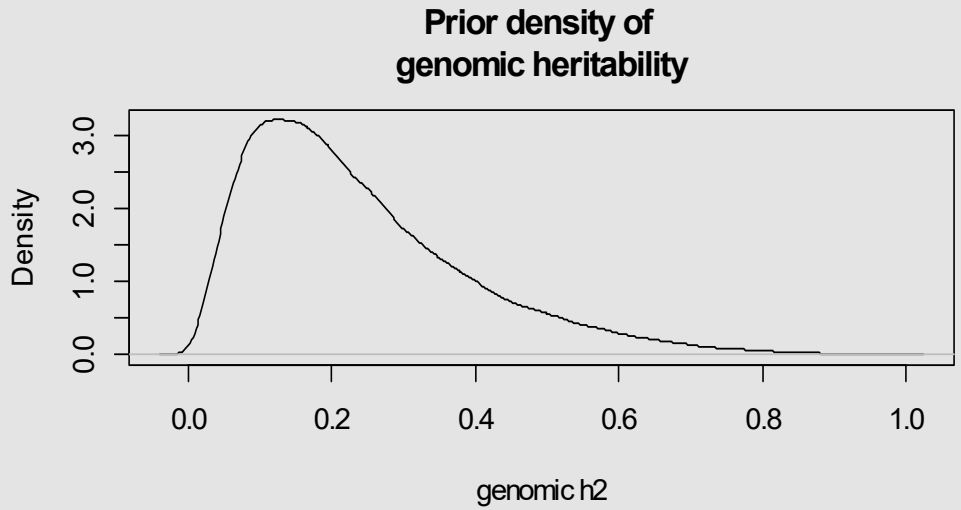
BAYESIAN LEARNING ABOUT GENOMIC HERITABILITY?

###PRIOR DISTRIBUTION OF HERITABILITY (FROM SAMPLES)

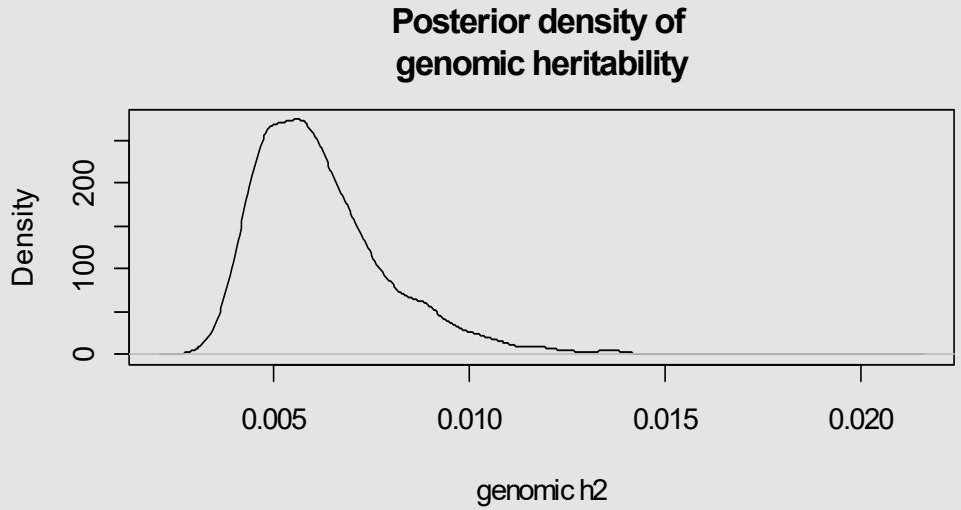
```
priorh2<-priorvg/(priorvg+priorve)
> summary(priorh2)
     Min.   1st Qu.         Median         Mean   3rd Qu.         Max.
0.0009697 0.1226000 0.2035000 0.2399000 0.3210000 0.9839000
```

###POSTERIOR DISTRIBUTION OF HERITABILITY (FROM SAMPLES)

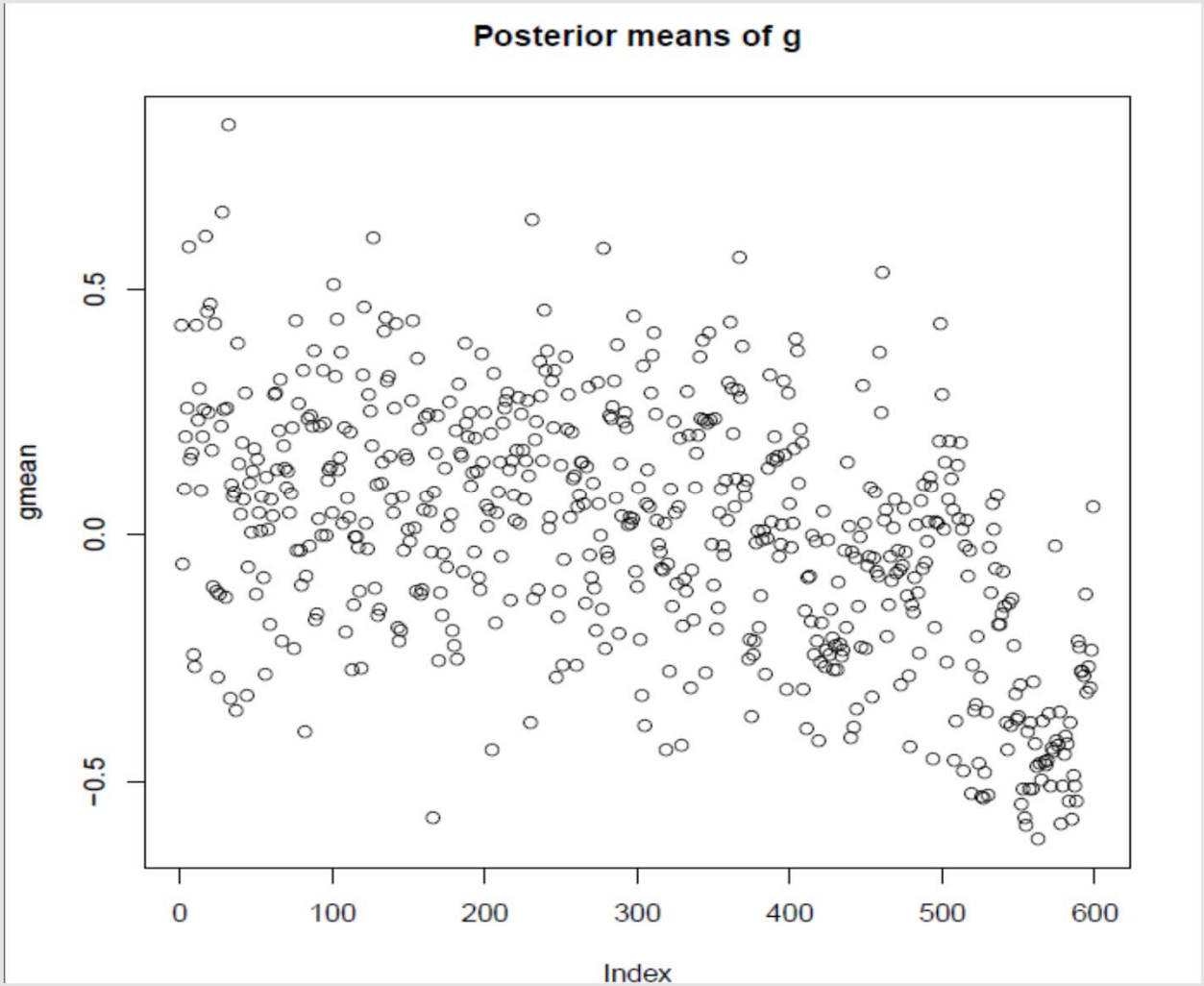
```
> summary(h2genomic[1001:NITER])
   Min.  1st Qu.         Median         Mean      3rd Qu.     Max.
0.003022 0.004964 0.005886 0.006259 0.007074 0.020670
```



LOTS OF MISSING H2 EVEN WITH 100 MARKERS (RECALL MLE with pedigree)



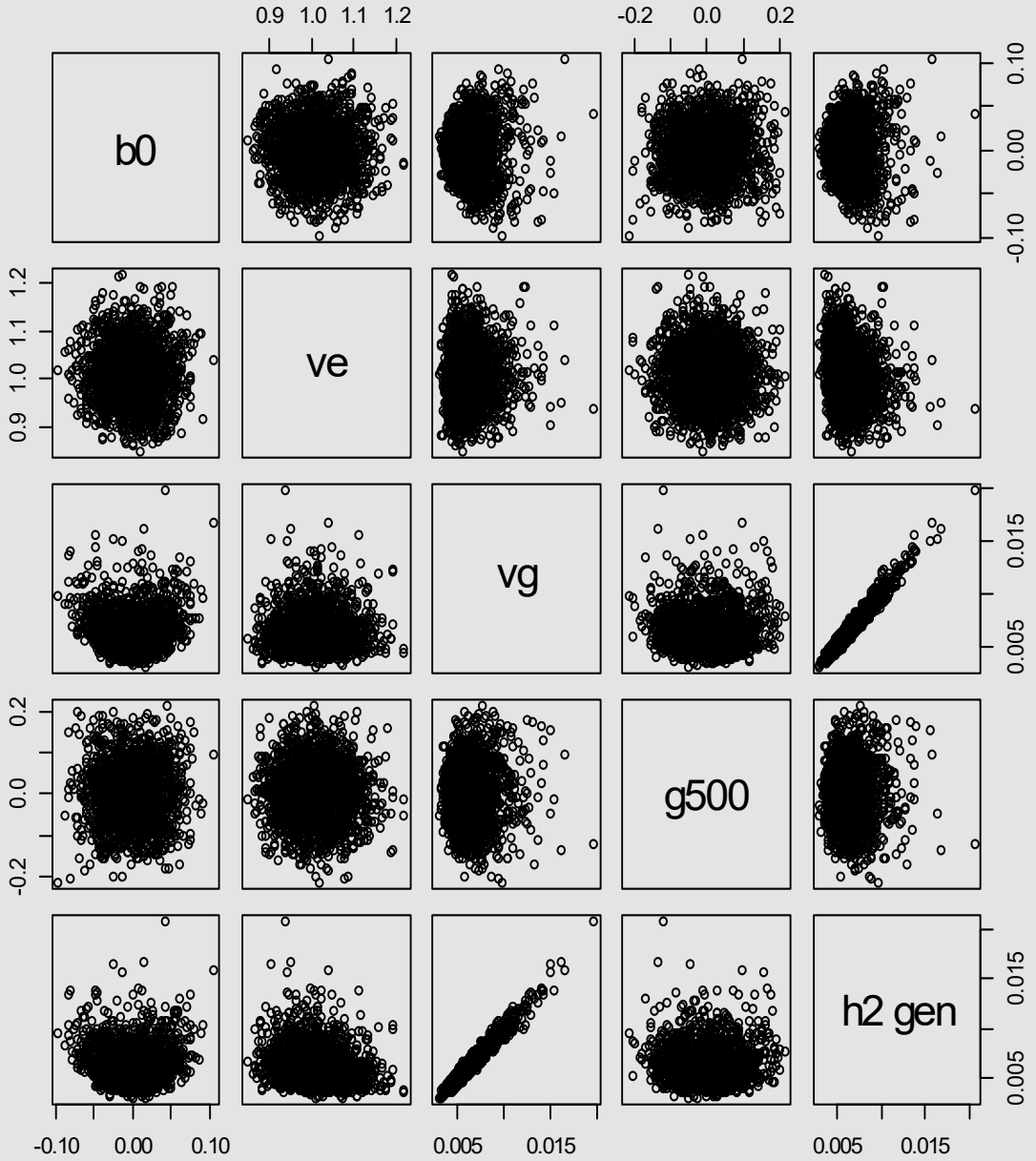
```
postmeang<-numeric(n)
for (i in 1:n){
postmeang[i]<-mean(gsamp[,i])
}
plot(postmeang,ylab="gmean",main="Posterior means of g")
```




```
> fitcor<-cor(y,postmeang)
> msefit<-crossprod(y-postmeang)/n
> fitcor
[1] 0.1038722
> msefit
          [,1]
[1,] 0.9980623
>
```

MODEL WITH 100 MARKERS FITS POORLY

POSTERIOR INTERCORRELATION STRUCTURE (ESTIMATED FROM SAMPLES)



ROBUST VARIANTS TO GBLUP: TMAP AND LMAP

ROBUST REGRESSION

- Much research devoted to robustness with respect to outliers
- Undeclared (non-random) preferential treatment
- Inadequate model specification ("population sub-structure unaccounted for)
- Gaussian distribution inadequate, thin tails
- Outliers often removed prior to analysis: ad-hoc rules. Uncertainty of exclusion not accounted for
- Robust methods: all data points used (unless clear anomaly) and down-weighted automatically

Bayesian (MCMC) in red

- Lange et al. (1989)
- Stranden and Gianola (1998, 1999)
- Rosa et al. (2003, 2004)
- Kizilkaya et al. (2003)
- Varona et al. (2006)
- Cardoso et al. (2006): 20000 post-weaning weight gains in cattle. Linear models with Gaussian vs t-residuals. Clear differences.
- ***Problem: MCMC (with one exception) is not used routinely for EBV***

RMAP: “Robust maximum a posteriori prediction”

2.1 Gaussian linear model

A fairly standard setting for a univariate mixed effects linear in model in quantitative genetics is

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{g} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of phenotypic measurements, $\boldsymbol{\alpha}$ is an $f \times 1$ vector of fixed regression coefficients and $\mathbf{W}$ is an $n \times f$ known incidence matrix with rank f ; $\mathbf{g}$ is a randomly distributed vector of genetic effects (typically additive effects) and $\mathbf{Z}$ is another known incidence matrix, and $\mathbf{e}$ is a vector of residuals. Often, it is assumed that

$$\begin{pmatrix} \mathbf{g} \\ \mathbf{e} \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}\sigma_g^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 \end{bmatrix} \right), \tag{2}$$

where $\mathbf{K}$ is a positive semi-definite symmetric similarity matrix (it can be $\mathbf{A}$ or $\mathbf{G}$ in pedigree or genome-based prediction, respectively) and σ_g^2 is a genetic or genomic variance component; $\mathbf{I}$

TMAP: “Student’s-t maximum a posteriori prediction”

Assume: t-distributed residuals and Gaussian genetic (genomic) values

Assume for $i = 1, 2, ..., n$, that the components of $\mathbf{e}$ are independent, although not identically distributed, with $e_i \sim t_\nu \left(0, \frac{\sigma_e^2}{n_i}, \nu\right)$. Here, 0 is the mean of the $t-$distribution; σ_e^2 is a scale parameter such that $Var\left(e_i\right) = \frac{\nu}{\nu - 2}\frac{\sigma_e^2}{n_i}$; n_i is a measure of the intensiveness of recording on an individual or line (e.g., number of clones) or of the degree of replication (also, n_i could be the number of daughters with records of a dairy bull if y_i is some processed average) and $\nu > 0$ is a possibly unknown positive "degrees of freedom" parameter. When $\nu \to \infty$, the t distribution converges to a normal one (e.g., Lange et at. 1989). Since $y_i = \mathbf{w}_i'\boldsymbol{\alpha} + \mathbf{z}_i'\mathbf{g} + e_i$ is a linear combination of e_i, conditionally on $\mathbf{g}$ one has that $y_i \sim t_\nu \left(\mathbf{w}_i'\boldsymbol{\alpha} + \mathbf{z}_i'\mathbf{g}, \frac{\sigma_e^2}{n_i}, \nu\right)$ (Zellner 1971; Box and Tiao 1973). In the preceding distributional statement, $\mathbf{w}_i'$ and $\mathbf{z}_i'$, are the i^{th} rows of $\mathbf{W}$ and $\mathbf{Z}$, respectively.

$$
\begin{aligned}
& p\left(\mathbf{y}, \mathbf{g} | \boldsymbol{\alpha}, \sigma_e^2, \sigma_g^2, \nu\right) \\
= & \prod_{i=1}^{n} \frac{\Gamma\left[\frac{(\nu+1)}{2}\right]}{\sqrt{\nu \frac{\sigma_e^2}{n_i} \pi} \Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{n_i}{\sigma_e^2 \nu}\left(y_i - \mathbf{w}_i' \boldsymbol{\alpha} - \mathbf{z}_i' \mathbf{g}\right)^2\right]^{-\frac{(\nu+1)}{2}} \times \\
& \left|2\pi \mathbf{K} \sigma_g^2\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_g^2} \mathbf{g}' \mathbf{K}^{-1} \mathbf{g}\right). \qquad (5)
\end{aligned}
$$

If the two variances and the degrees of freedom parameters are regarded as known, the joint density above can be used for deducing the following conditional posterior distribution

$$
\begin{aligned}
& p\left(\mathbf{g}, \boldsymbol{\alpha} | \sigma_e^2, \sigma_g^2, \nu, \mathbf{y}\right) \\
\propto & \prod_{i=1}^{n} \left[1 + \frac{n_i}{\sigma_e^2 \nu}\left(y_i - \mathbf{w}_i' \boldsymbol{\alpha} - \mathbf{z}_i' \mathbf{g}\right)^2\right]^{-\frac{(\nu+1)}{2}} \exp\left(-\frac{1}{2\sigma_g^2} \mathbf{g}' \mathbf{K}^{-1} \mathbf{g}\right). \qquad (6)
\end{aligned}
$$

The preceding is the kernel of the posterior density of $\mathbf{g}, \boldsymbol{\alpha}$ conditionally on $\sigma_e^2, \sigma_g^2, \nu$ and after adopting a flat prior for $\boldsymbol{\alpha}$ (Gianola and Fernando 1986; Sorensen and Gianola 2002).

2.2.2 Maximum a posteriori point estimation (TMAP)

As shown in Appendix A, the $\boldsymbol{\alpha}$ and $\mathbf{g}$ components of the joint mode of the posterior distribution having density (6) can be found using the functional iteration

$$\begin{bmatrix} \boldsymbol{\alpha}^{[t+1]} \\ \mathbf{g}^{[t+1]} \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{D}^{[t]}\mathbf{W} & \mathbf{W}'\mathbf{D}^{[t]}\mathbf{Z} \\ \mathbf{Z}'\mathbf{D}^{[t]}\mathbf{Z} & \mathbf{Z}'\mathbf{D}^{[t]}\mathbf{Z}+\dfrac{\lambda\nu}{(\nu+1)}\mathbf{K}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{W}'\mathbf{D}^{[t]}\mathbf{y} \\ \mathbf{Z}'\mathbf{D}^{[t]}\mathbf{y} \end{bmatrix}, \tag{7}$$

where t denotes round of iteration, $\lambda = \dfrac{\sigma_e^2}{\sigma_g^2}$, and $\mathbf{D}$ is an $n \times n$ diagonal matrix with typical element

$$d_i = \frac{n_i}{\left[1 + \dfrac{(y_i - \mathbf{w}_i'\boldsymbol{\alpha} - \mathbf{z}_i'\mathbf{g})^2}{\dfrac{\sigma_e^2}{n_i}\nu}\right]}. \tag{8}$$

Observe that d_i decreases as y_i departs further from its conditional (given $\mathbf{g}$) expectation $\mathbf{w}_i'\boldsymbol{\alpha} + \mathbf{z}_i'\mathbf{g}$, and as ν decreases. Also, as ν goes to infinity (the t distribution approaches normality), d_i drifts towards n_i, which would be the weight assigned in a Gaussian regression model. After λ, ν and σ_e^2 are elicited in some manner (more on this in the PREDICTIVE ASSESSMENT section

2.2.3 Special case: zero-means model

Often, phenotypes are presented as pre-corrected for effects of systematic sources of variation, such as age and sex of the individual, or year-season of measurement. In such a case, there are no fixed effects in the model, which would become $\mathbf{y} = \mathbf{Zg} + \mathbf{e}$. Further, in genomed-enabled prediction, it is common to encounter data sets where the n cases have all been genotyped, so that $\mathbf{Z} = \mathbf{I}$ and $\mathbf{y} = \mathbf{g} + \mathbf{e}$. In this particular situation the functional iteration becomes

$$\mathbf{g}^{[t+1]} = \left[\mathbf{D}^{[t]} + \frac{\lambda\nu}{(\nu+1)}\mathbf{K}^{-1}\right]^{-1} \mathbf{D}^{[t]}\mathbf{y}, \tag{12}$$

where $\mathbf{D}$ is a diagonal matrix with typical element now equal to

$$d_i = \frac{n_i}{\left[1 + \dfrac{(y_i - g_i)^2}{\dfrac{\sigma_e^2}{n_i}\nu}\right]}. \tag{13}$$

LMAP: "Laplace's maximum a posteriori prediction"

It will be assumed now that observations are (conditionally) independently distributed as $y_i|\mu_i, \sigma_e^2 \sim Laplace(\mu_i, \frac{\sigma_e^2}{n_i})$, where $\mu_i = \mathbf{w}_i'\boldsymbol{\alpha} + \mathbf{z}_i'\mathbf{g}$. The mean of Laplace's (also called double-exponential) distribution is μ_i and its variance is $\frac{\sigma_e^2}{n_i}$. Often, the density of the distribution is writen as

$$p\left(y_i|\mu,\tau\right) = \frac{\sqrt{n_i}}{2\tau}\exp\left(-\frac{\sqrt{n_i}\left|y-\mu_i\right|}{\tau}\right), \tag{15}$$

where $\tau = \sqrt{\frac{\sigma_e^2}{2}}$ is a parameter that relates to spread of the distribution. The probability density function of our sampling model can be represented as

$$p\left(\mathbf{y}|\boldsymbol{\alpha},\mathbf{g}\right) = \prod_{i=1}^{n}\frac{\sqrt{n_i}}{\sqrt{2\sigma_e^2}}\exp\left(-\frac{\sqrt{n_i}\left|y_i-\mu_i\right|}{\sqrt{\frac{\sigma_e^2}{2}}}\right). \tag{16}$$

Adopting a flat prior for $\boldsymbol{\alpha}$ (as before) and the Gaussian prior in (2), the log-conditional posterior density of $\mathbf{g}$ and $\boldsymbol{\alpha}$ is (K is an additive constant) is

$$\begin{aligned} L_{DE} &= \log\left[p\left(\mathbf{g},\boldsymbol{\alpha}|\sigma_e^2,\sigma_g^2,\mathbf{y}\right)\right] \\ &= K - \frac{1}{\sqrt{\frac{\sigma_e^2}{2}}}\sum_{i=1}^{n}\sqrt{n_i}\left|y_i-\mu_i\right| - \frac{1}{2\sigma_g^2}\mathbf{g}'\mathbf{K}^{-1}\mathbf{g}. \end{aligned} \tag{17}$$

$$\begin{bmatrix} \mathbf{W}'\mathbf{M}^{[t]}\mathbf{W} & \mathbf{W}'\mathbf{M}^{[t]}\mathbf{Z} \\ \mathbf{Z}'\mathbf{M}^{[t]}\mathbf{W} & \mathbf{Z}'\mathbf{M}^{[t]}\mathbf{Z}+\omega\mathbf{K}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^{[t+1]} \\ \mathbf{g}^{[t+1]} \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{M}^{[t]}\mathbf{y} \\ \mathbf{Z}'\mathbf{M}^{[t]}\mathbf{y} \end{bmatrix}, \tag{18}$$

where $\mathbf{M} = Diag\{m_i\}$ is an $m \times m$ diagonal matrix with typical element

$$m_i = \frac{1}{|y_i - \mu_i|}; i = 1, 2, ..., n, \tag{19}$$

and $\omega = \dfrac{\sqrt{\frac{\sigma_e^2}{2}}}{\sigma_g^2}$ is a regularization parameter. If σ_g^2 is a "genomic" variance (de los Campos et al. 2015), h_g^2 is genomic heritability and σ_y^2 is the phenotypic variance,

$$\omega = \frac{\sqrt{\frac{\sigma_e^2}{2}}}{\sigma_g^2} = \frac{\sqrt{\frac{\left(1-h_g^2\right)\sigma_y^2}{2}}}{h_g^2\sigma_y^2} \tag{20}$$

2.3.3 Special case: zero-means model

Again, a special case is the zero-means model with $\mathbf{Z} = \mathbf{I}$, so that $\mathbf{y} = \mathbf{g} + \mathbf{e}$. The iteration becomes

$$\mathbf{g}^{[t+1]} = \left[\mathbf{M}^{[t]}+\omega\mathbf{K}^{-1}\right]^{-1}\mathbf{M}^{[t]}\mathbf{y}, \tag{22}$$

and phenotype i "effectively" enters into the analysis as $m_i = \sqrt{n_i}\dfrac{y_i}{|y_i - g_i|}$. Thus, phenotypic values departing markedly from their conditional expected value g_i are more heavily discounted than those that are closer to it.

Predictive assessment of regularization parameters

- EM: awkward, especially for Laplace's distribution. Gets caught in local modes with low predictive power
- Bayes MCMC: straightforward but wish to avoid it
- TMAP, vary variance ratio over a grid centered at ML estimates of heritability (pedigree or genomic). Vary degrees of freedom over grid. Typically df=4, 6, 8, 12, 16 will suffice.
- LMAP, especially if phenotypes are in SD units, evaluate the regularization parameter over a grid of heritability values
- Occasionally need residual variance: easy to estimate by any simple method (e.g., MINQUE) in each training instance
- Since re-weighted MME need to be solved, can use short-cuts in Gianola and Schoen (2016) for ad-nauseum CV.

IMPORTANT: RMAP CAN BE IMPLEMENTED IN ANY LINEAR SOLVER THAT ALLOWS FOR DIFFERENTIAL WEIGHTS

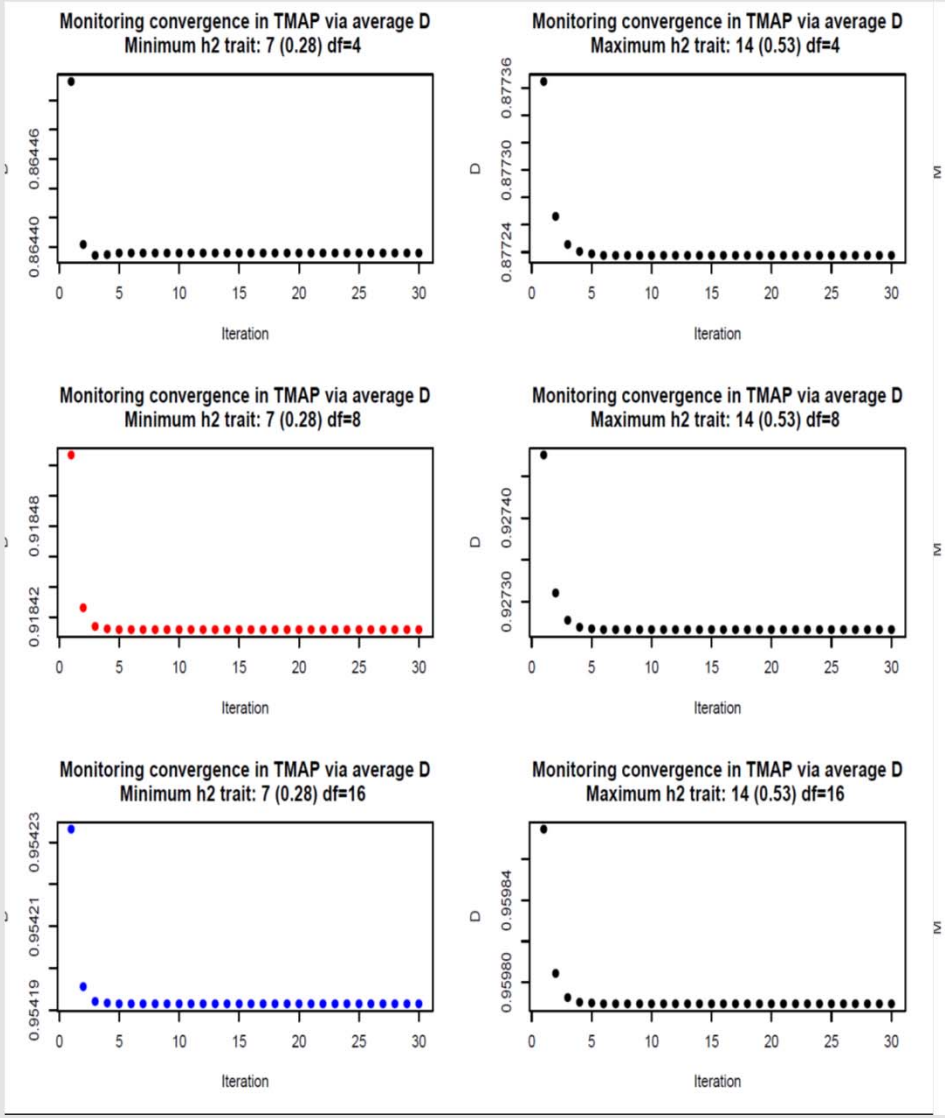
4 APPLICATION AND EVALUATION OF METHODS

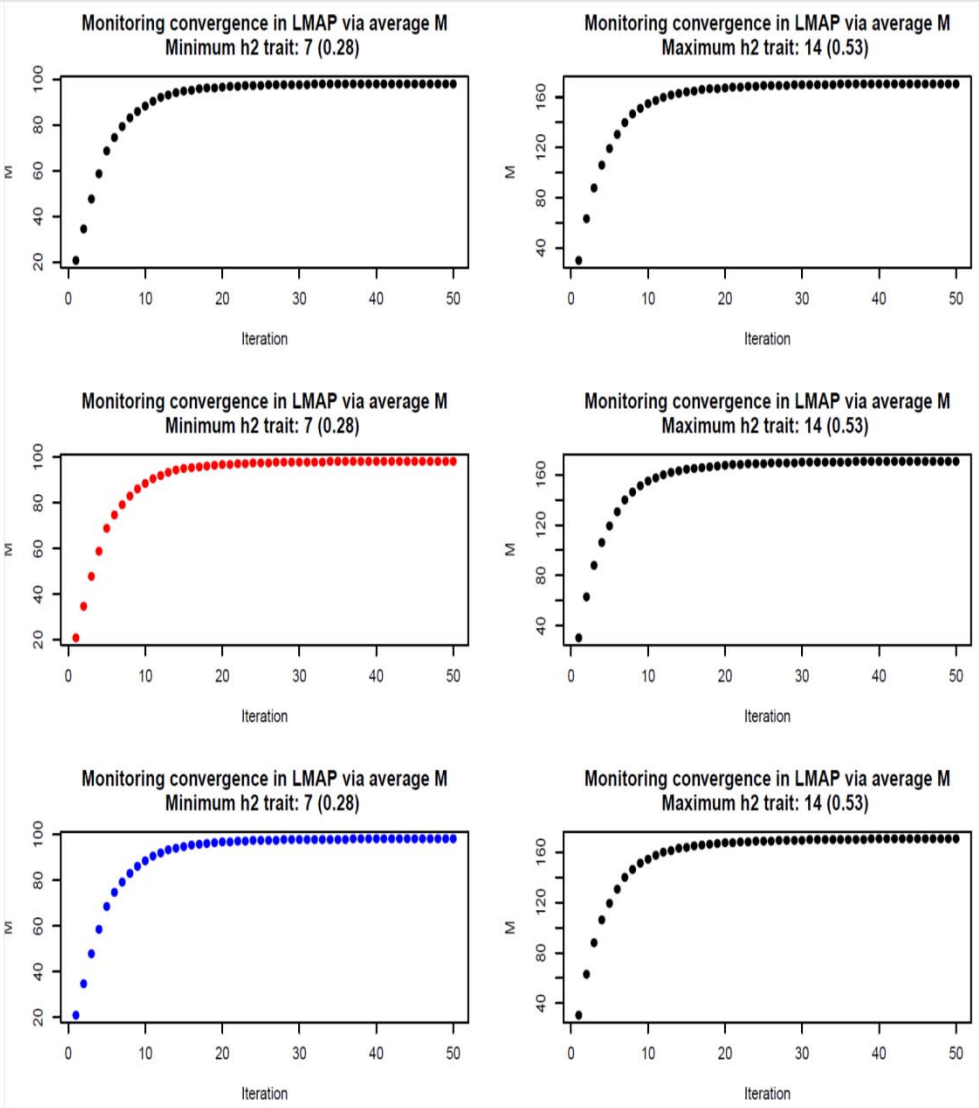
4.1 Data

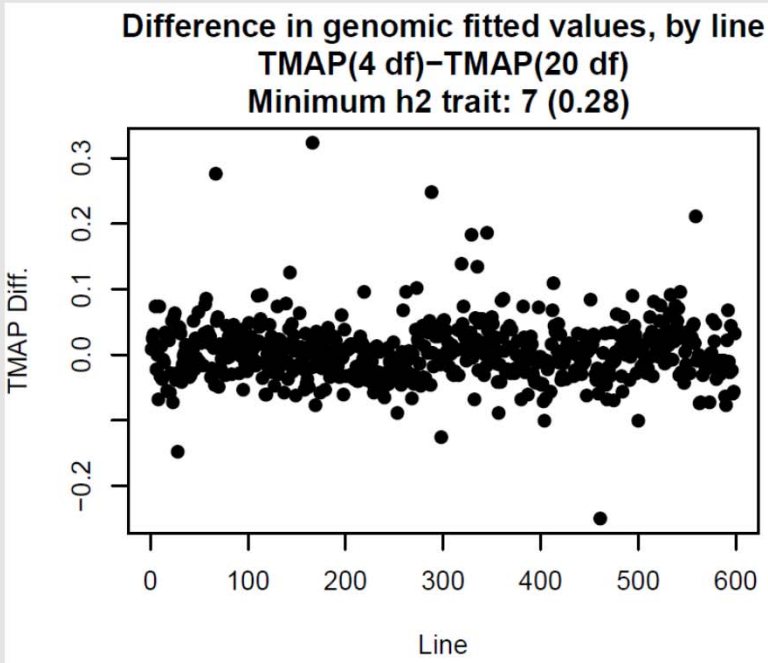
Wheat grain yield. The wheat data available in package BGLR (Pérez-Rodríguez and de los Campos 2014) was employed. This data set is well characterized and has also been used by, e.g., Crossa et al. (2010), Gianola et al. (2011), Long et al. (2011) and Gianola and Shön. The data comes from trials conducted by the International Maize and Wheat Improvement Center (CIMMYT), Mexico. There are 599 wheat inbred lines, each genotyped with 1279 DArT (Diversity Array Technology) markers and planted in 4 environments. Sample size was $n = 599$ with $p = 1279$ being the number of markers. These DArT markers are binary $(0, 1)$ and denote presence or absence of an allele at a marker locus in a given line. The data set also includes a pedigree-derived relationship matrix. The trait treated as predictand in our study was wheat grain yield measured in four distint environments, with all 599 lines represented in each environment; the corresponding vectors of phenotypes will be denoted as $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ and $\mathbf{y}_4$. We also synthesized "composite" traits" from sums of phenotypes: $\mathbf{y}_{1+2}, \mathbf{y}_{1+3}, \mathbf{y}_{1+2}, \mathbf{y}_{1+4}, \mathbf{y}_{2+3}, \mathbf{y}_{2+4}, \mathbf{y}_{1+2+3}, \mathbf{y}_{1+2+4}, \mathbf{y}_{1+3+4}, \mathbf{y}_{2+3+4},$ and $\mathbf{y}_{1+2+3+4}$, so the total number of traits was 15. Gianola et al. (2016) conducted a multi-dimensional scaling analysis of the marker matrix, which suggested strong population sub-structure. In this respect, the data set is ideal, because if a standard GBLUP model ignores concealed structure outliers may appear, with a potential impact on the ranking of lines.

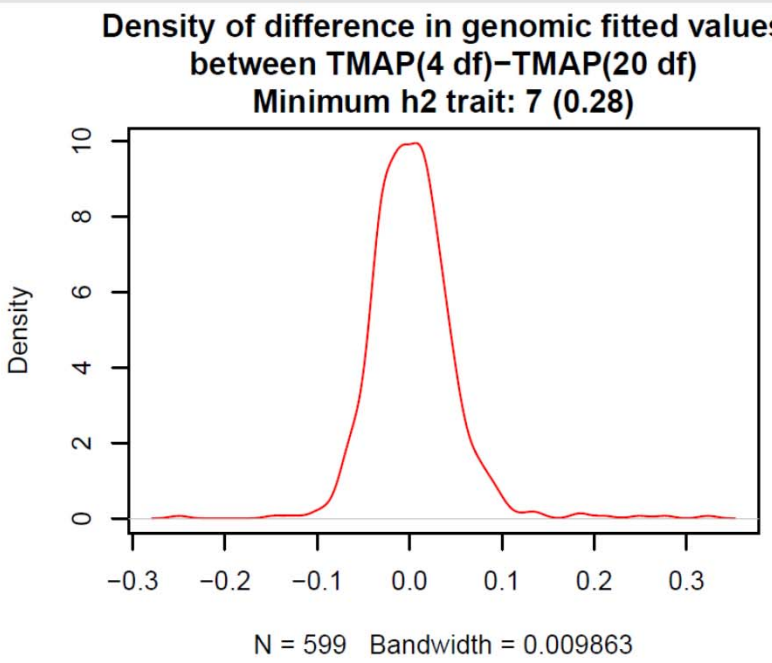
EVALUATION OF MODEL TRAINING ONLY

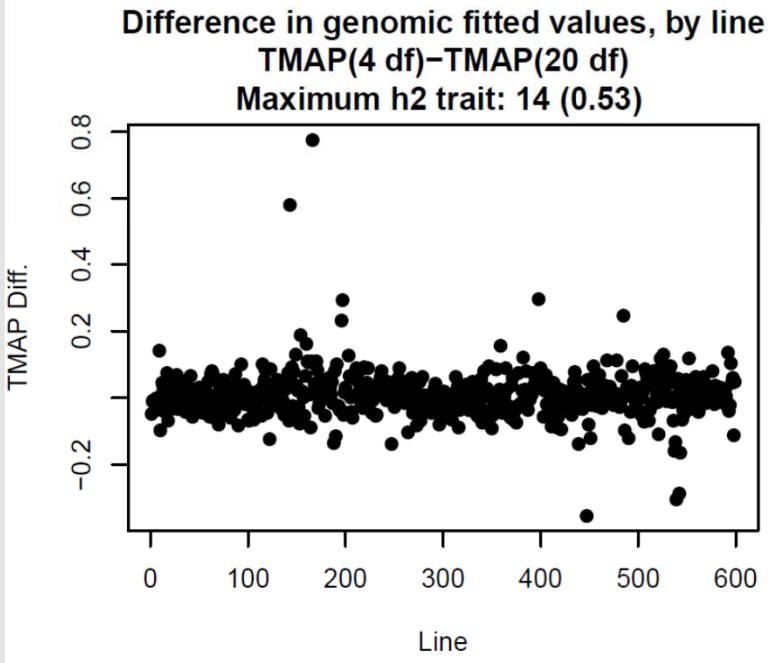
(PREDICTIVE RESULTS NOT AVAILABLE YET)

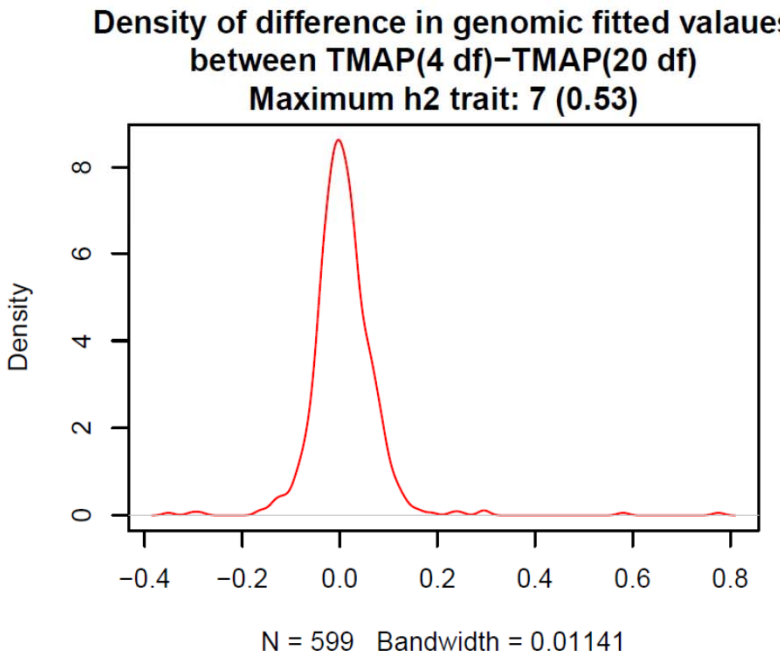

Monitoring convergence in TMAP via average D
Minimum h2 trait: 7 (0.28) df=4
Monitoring convergence in TMAP via average D
Maximum h2 trait: 14 (0.53) df=4
Monitoring convergence in TMAP via average D
Minimum h2 trait: 7 (0.28) df=8
Monitoring convergence in TMAP via average D
Maximum h2 trait: 14 (0.53) df=8
Monitoring convergence in TMAP via average D
Minimum h2 trait: 7 (0.28) df=16
Monitoring convergence in TMAP via average D
Maximum h2 trait: 14 (0.53) df=16
D
Iteration

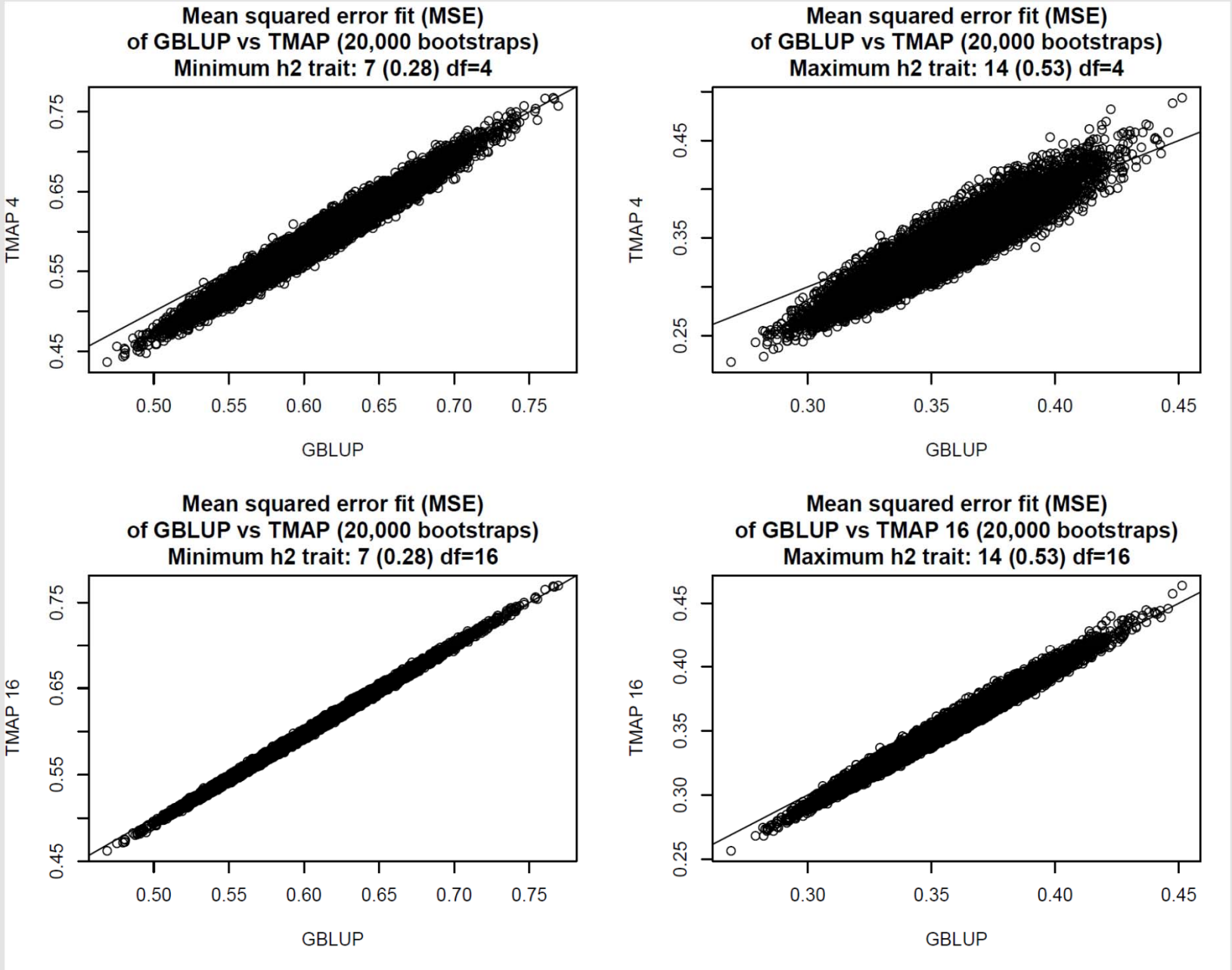

Monitoring convergence in LMAP via average M
Minimum h2 trait: 7 (0.28)
Monitoring convergence in LMAP via average M
Maximum h2 trait: 14 (0.53)
M
Iteration

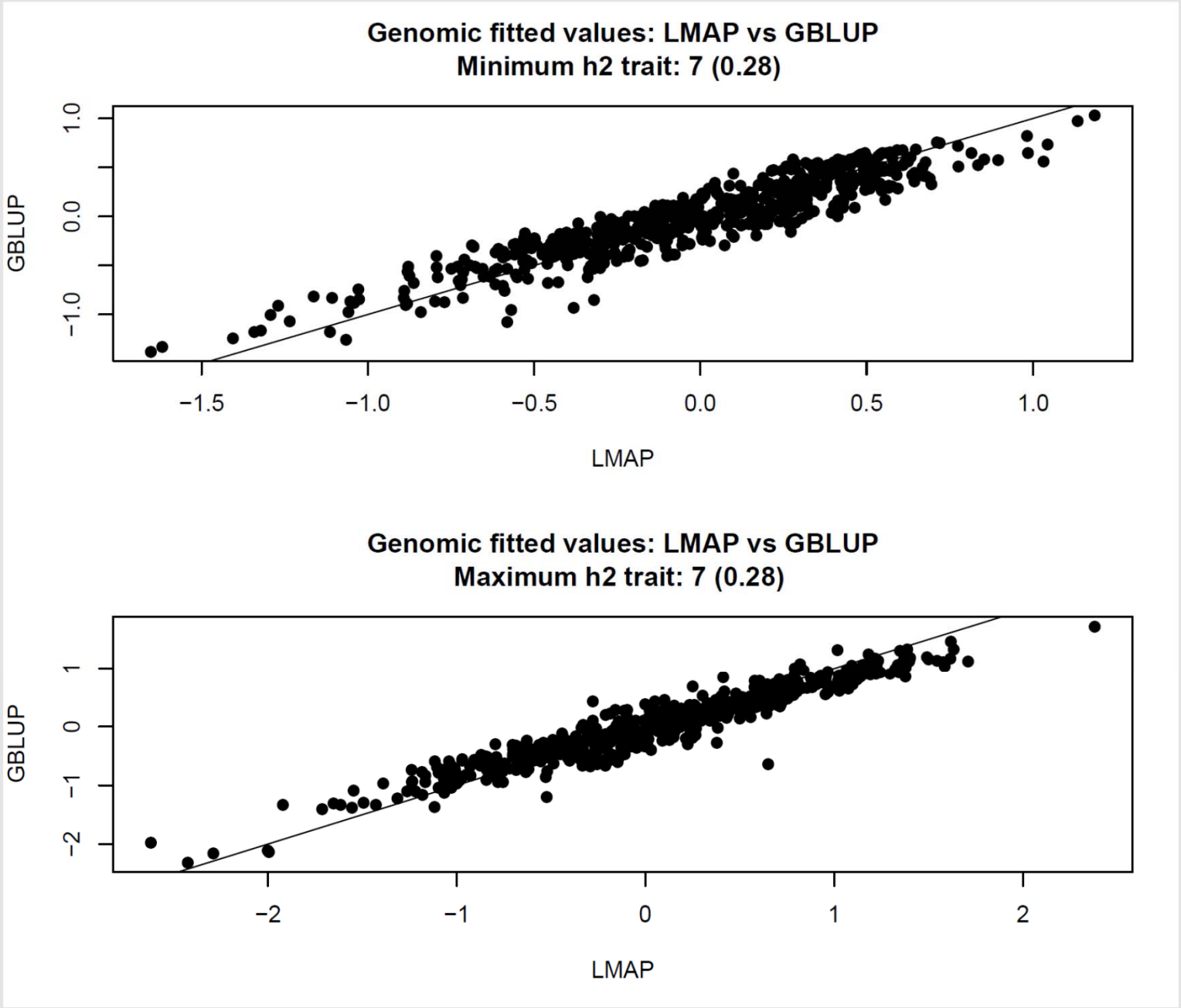

Difference in genomic fitted values, by line
TMAP(4 df)−TMAP(20 df)
Minimum h2 trait: 7 (0.28)
TMAP Diff.
−0.2
0.0
0.1
0.2
0.3
0
100
200
300
400
500
600
Line

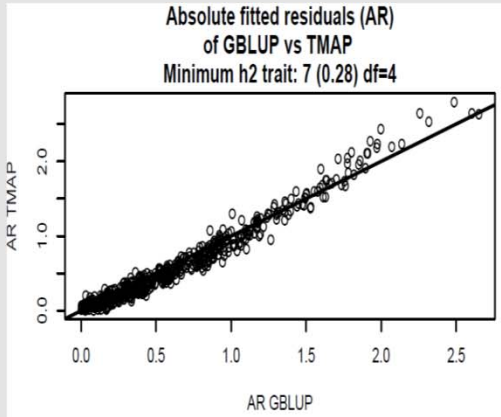

Density of difference in genomic fitted values
between TMAP(4 df)−TMAP(20 df)
Minimum h2 trait: 7 (0.28)
Density
0
2
4
6
8
10
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
N = 599 Bandwidth = 0.009863

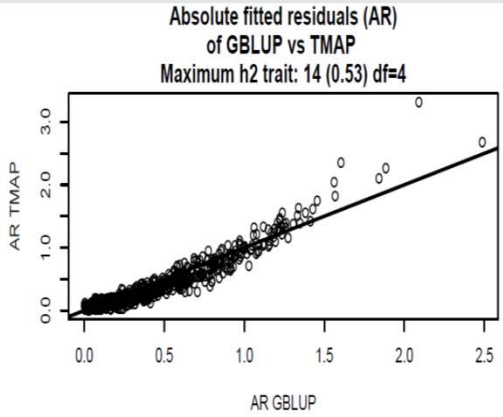

Difference in genomic fitted values, by line
TMAP(4 df)−TMAP(20 df)
Maximum h2 trait: 14 (0.53)
TMAP Diff.
−0.2
0.2
0.4
0.6
0.8
0
100
200
300
400
500
600
Line

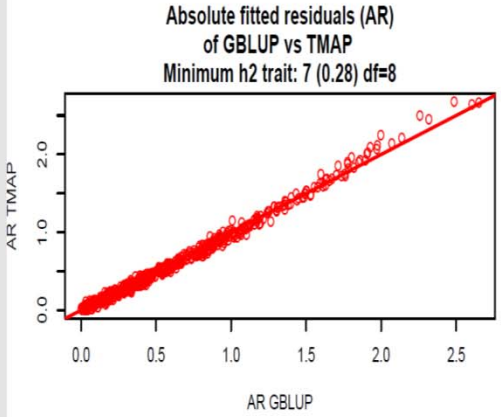

Density of difference in genomic fitted valaues
between TMAP(4 df)−TMAP(20 df)
Maximum h2 trait: 7 (0.53)
Density
0
2
4
6
8
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
N = 599 Bandwidth = 0.01141

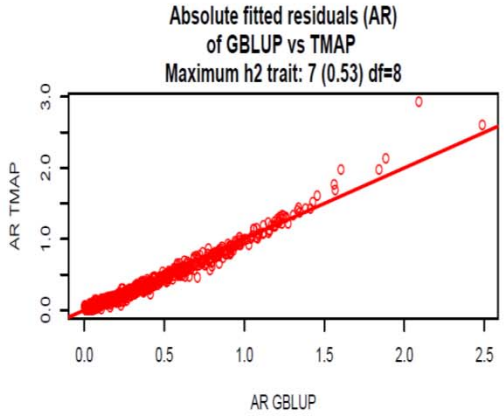

Mean squared error fit (MSE)
of GBLUP vs TMAP (20,000 bootstraps)
Minimum h2 trait: 7 (0.28) df=4
TMAP 4
GBLUP
0.45
0.55
0.65
0.75
0.50
0.55
0.60
0.65
0.70
0.75
Mean squared error fit (MSE)
of GBLUP vs TMAP (20,000 bootstraps)
Maximum h2 trait: 14 (0.53) df=4
TMAP 4
GBLUP
0.25
0.35
0.45
0.30
0.35
0.40
0.45
Mean squared error fit (MSE)
of GBLUP vs TMAP (20,000 bootstraps)
Minimum h2 trait: 7 (0.28) df=16
TMAP 16
GBLUP
0.45
0.55
0.65
0.75
0.50
0.55
0.60
0.65
0.70
0.75
Mean squared error fit (MSE)
of GBLUP vs TMAP 16 (20,000 bootstraps)
Maximum h2 trait: 14 (0.53) df=16
TMAP 16
GBLUP
0.25
0.30
0.35
0.40
0.45
0.30
0.35
0.40
0.45

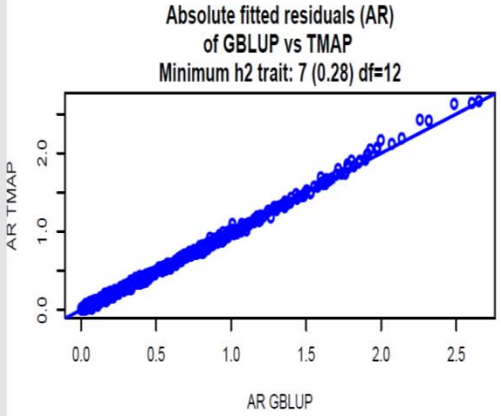

Genomic fitted values: LMAP vs GBLUP
Minimum h2 trait: 7 (0.28)
GBLUP
LMAP
−1.5
−1.0
−0.5
0.0
0.5
1.0
−1.0
0.0
1.0
Genomic fitted values: LMAP vs GBLUP
Maximum h2 trait: 7 (0.28)
GBLUP
LMAP
−2
−1
0
1
2
−2
−1
0
1

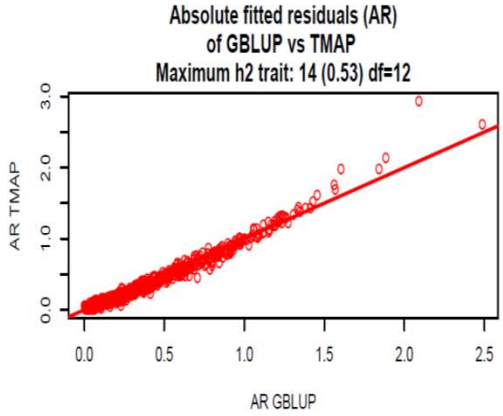

Absolute fitted residuals (AR)
of GBLUP vs TMAP
Minimum h2 trait: 7 (0.28) df=4
AR TMAP
AR GBLUP

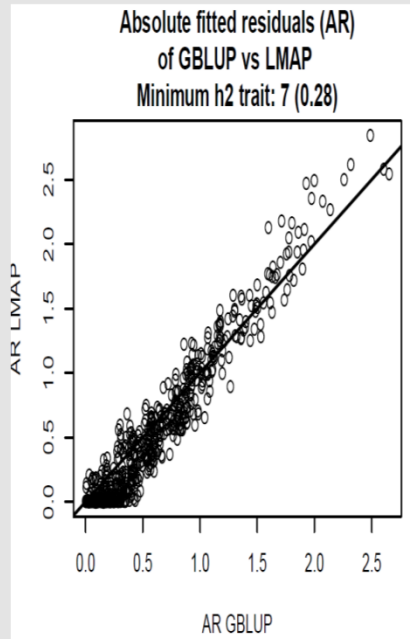

Absolute fitted residuals (AR)
of GBLUP vs TMAP
Maximum h2 trait: 14 (0.53) df=4
AR TMAP
AR GBLUP

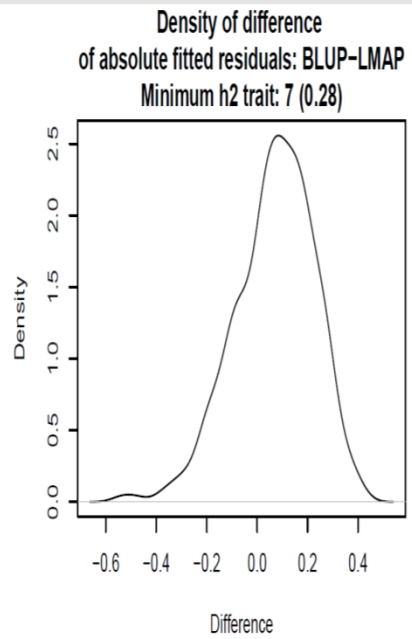

Absolute fitted residuals (AR)
of GBLUP vs TMAP
Minimum h2 trait: 7 (0.28) df=8
AR TMAP
AR GBLUP

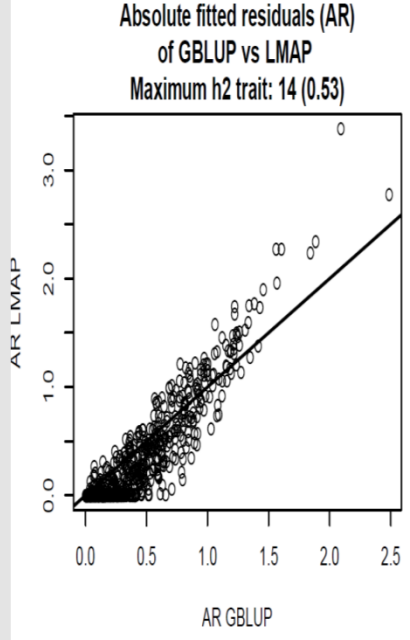

Absolute fitted residuals (AR)
of GBLUP vs TMAP
Maximum h2 trait: 7 (0.53) df=8
AR TMAP
AR GBLUP

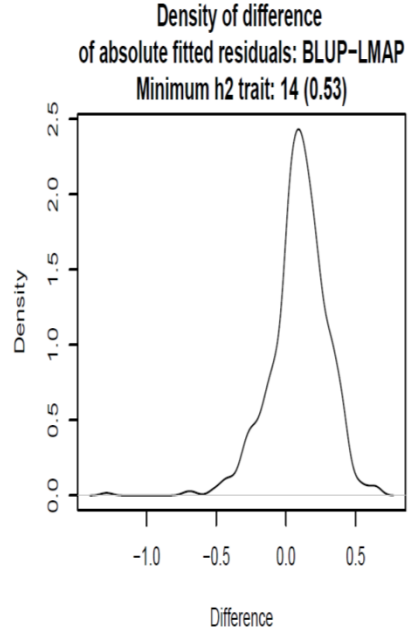

Absolute fitted residuals (AR)
of GBLUP vs TMAP
Minimum h2 trait: 7 (0.28) df=12
AR TMAP
AR GBLUP

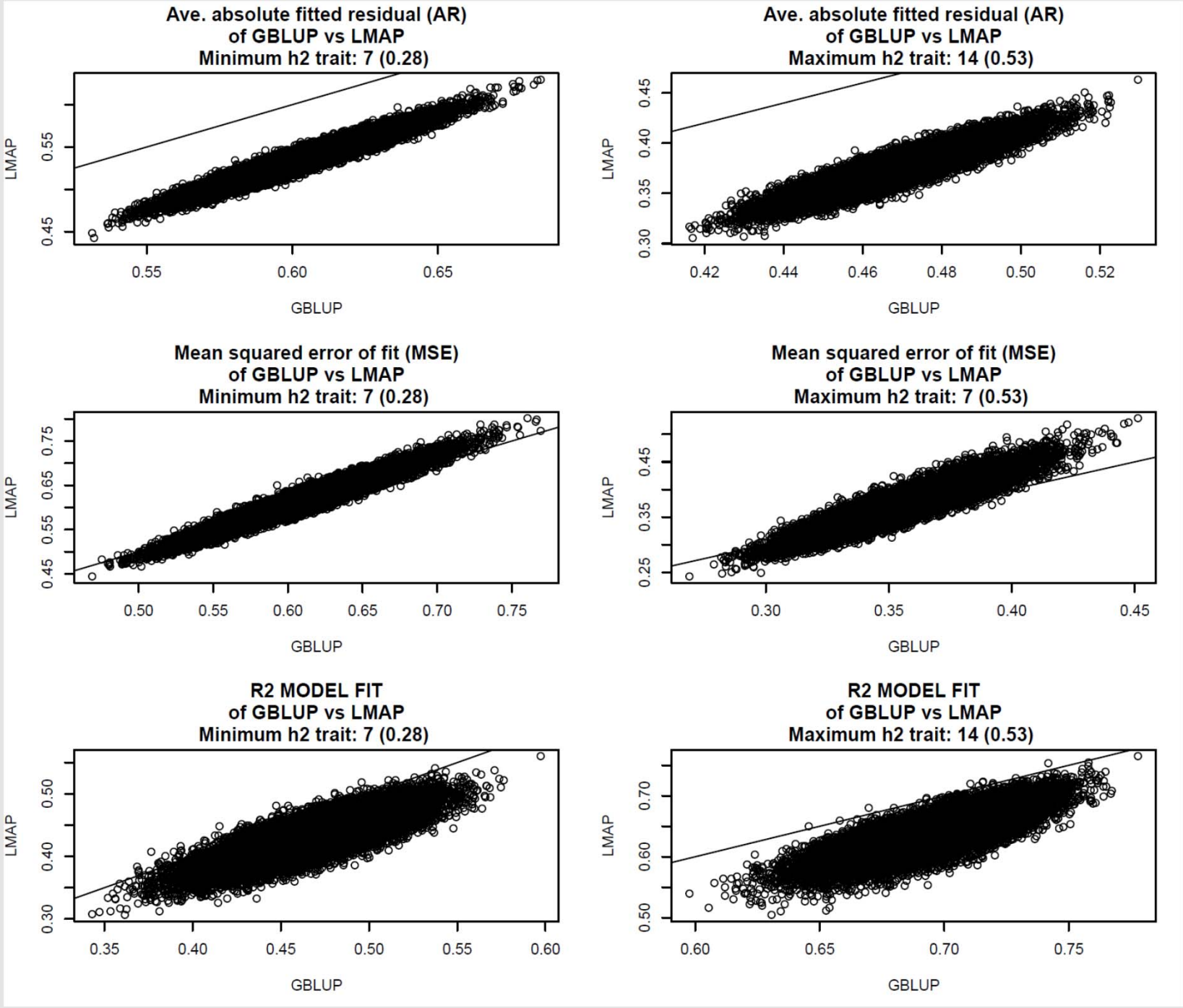

Absolute fitted residuals (AR)
of GBLUP vs TMAP
Maximum h2 trait: 14 (0.53) df=12
AR TMAP
AR GBLUP


Absolute fitted residuals (AR)
of GBLUP vs LMAP
Minimum h2 trait: 7 (0.28)
AR LMAP
AR GBLUP


Density of difference
of absolute fitted residuals: BLUP−LMAP
Minimum h2 trait: 7 (0.28)
Density
Difference


Absolute fitted residuals (AR)
of GBLUP vs LMAP
Maximum h2 trait: 14 (0.53)
AR LMAP
AR GBLUP


Density of difference
of absolute fitted residuals: BLUP−LMAP
Minimum h2 trait: 14 (0.53)
Density
Difference


Ave. absolute fitted residual (AR)
of GBLUP vs LMAP
Minimum h2 trait: 7 (0.28)
LMAP
GBLUP
Ave. absolute fitted residual (AR)
of GBLUP vs LMAP
Maximum h2 trait: 14 (0.53)
LMAP
GBLUP
Mean squared error of fit (MSE)
of GBLUP vs LMAP
Minimum h2 trait: 7 (0.28)
LMAP
GBLUP
Mean squared error of fit (MSE)
of GBLUP vs LMAP
Maximum h2 trait: 7 (0.53)
LMAP
GBLUP
R2 MODEL FIT
of GBLUP vs LMAP
Minimum h2 trait: 7 (0.28)
LMAP
GBLUP
R2 MODEL FIT
of GBLUP vs LMAP
Maximum h2 trait: 14 (0.53)
LMAP
GBLUP