December 16, 2019

Testing for QTL by environment interaction in multi-environment GWAS

(in alphabetical order) Martin Boer[1], Daniela Bustos-Korts[1], Willem Kruijer[1,*], Marcos Malosetti[1], Emilie Millet[1], Fred van Eeuwijk[1], Bart-Jan van Rossum[1]; (Maikel, Vincent, Claude, Francòis, Roberto ?)

* Corresponding author (E-mail: willem.kruijer@wur.nl)

The main focus is on multi environment GWAS; multi-trait is discussed as well, but rather briefly. This is because (1) multi-trait modeling is more complicated due to the non-diagonal error structure (2) the paper will take much longer if we try to cover both (3) in multi-trait GWAS we are in my opinion not as much competitive as with multi environment gwas, at least if it comes to implementations that are currently (or soon) available. Except maybe the contrasts in the LfN paper, but there the code is poor and difficult to generalize. (4) we will stay aways from discussions on causality (see [?]), and also from competing with very high dimensional $eQTL$ scans (which typically are rather simplistic, from a statistical point of view, but hard to beat computationally).

1   Introduction

General overview of $G \times E$, and QTL x Env mapping in experimental populations [?] and association panels ([?], [?])...

Two major challenges, that we will address:

- Deriving and implementing powerful tests for detecting $QTL \times Env$ interaction. This will be the main goal of the paper.

- Estimating the covariance structure, and avoiding (or handling) inflation. For the previous point, it will be necessary to make a choice here, but this is less important in its own right. I do think we have potentially interesting things to say bout this, but at the same time should stay away from the rather specific and refined covariance models used in limix (and used in [?], [?]).

2   Materials and Methods

2.1   Modeling and estimation of the variance-covariance

- Unstructured [?], [?]

- Unstructured, approximated pairwise [?]

- Unstructured, alternative approximation [?]

₃₃ • Factor analytic [**?**]

₃₄ • stephens (based on beta's single env gwas): [**?**]

₃₅ • limix ? Check out `http://genetics.cs.ucla.edu/pylmm/` and other software at
₃₆ `http://zarlab.cs.ucla.edu/software/`

₃₇ • low rank approximations of K

₃₈ 2.2 Testing marker effects

₃₉ Tests:

₄₀ • Single-environment GWAS; then do something heuristic to combine environments
₄₁ (e.g. average of $-log(p)$ values)

₄₂ • The 'omnibus' test, implemented in most software: test the H_0 that $\beta_1 = \ldots =$
₄₃ $\beta_p = 0$ against the alternative that at least one of the β_j is nonzero.

₄₄ • Test for a common effect: test the H_0 that $\beta_1 = \ldots = \beta_p = \beta = 0$ against the
₄₅ alternative that $\beta \neq 0$. This was proposed by [**?**], but implemented only for 2 traits
₄₆ (environments), without diagonalization.

₄₇ • Test for QTL by environment interaction: test the H_0 that $\beta_1 = \ldots = \beta_p = \beta$ (for
₄₈ some possibly nonzero β) against the alternative that not all β_j are equal. Again,
₄₉ for 2 traits (environments) this was proposed by [**?**], but with a somewhat different
₅₀ parameterization. Here we assume that $\beta_j = \beta + \gamma_j$, and test if $\gamma_1 = \ldots = \gamma_p = 0$.
₅₁ See also [**?**], [**?**]. Typically, the test for QTL by environment interaction is only
₅₂ performed when the omnibus test shows at least some significance.

₅₃ • Given one or more environmental covariates (ECs), test $QTL \times EC$. This has
₅₄ been done only for experimental populations [**?**]. For association panels it has been
₅₅ done in [**?**], but only 'a posteriori', regressing $\hat{\beta}_1, \ldots, \hat{\beta}_p$ on e.g. temperature. Also
₅₆ contrasts can be dealt with ([**?**]), in which case the EC is a vector of zeros and
₅₇ ones. The $QTL \times EC$ test is in fact an extension of the test for a common effect:
₅₈ it is now assumed that $\beta_j = \beta + EC_j \gamma_{EC}$, and we test if $\gamma_{EC} = 0$.

₅₉ • Instead of the usual regression of phenotype on genotype, it has been proposed to
₆₀ regress markers on traits or environments; see Garin (Msc thesis), [**?**] and [**?**]. What
₆₁ about [**?**] ?

₆₂ • To do: Check for additional tests in limix

₆₃ An important contribution of this work will be that we provide an efficient implemen-
₆₄ tation for the tests for a common effect, QTL by environment and QTL by EC, and may
₆₅ be the first to do GWAS with these tests for more than 2 environments.

2.3 Software

Advertize our software here. We may also use GEMMA, GAMMA, TASSEL(?) and LIMIX

3 Results

Discuss different sources of inflation:

- in the F/Wald/score test itself [**?**]

- a strong QTL elsewhere ([**?**])

- epistasis ([**?**], [**?**])

- more generally, an inappropriate covariance model

Simulate different scenarios (type of QTLs):

- QTLs with an effect in all environments, with random effect size and sign

- QTLs with an effect in all environments, with a more consistent effect

- As in the previous two points, but only in k out of p environments

- QTL x EC

$\rightarrow$ How much power do the different tests have ?
For inspiration, see also [**?**]
Analysis of real data
Other issues: missing values; multiple testing correction; initial analysis of trials; the use of plot data to precisely estimate G x E, ...

4 Discussion

To which extent can the methodology be used for multi-trait data obtained from different plants. In any case, modelling multi-trait data measured on the same plants is more challenging.
Multi-locus models; [**?**], [**?**]

A Estimating marker effects

A.1 The general multi-trait mixed model

We consider the multi-trait mixed model of [**?**], and adopt to a large extent their notation:

$$\tilde{\mathbf{Y}} = \mathbf{B}\tilde{\mathbf{X}} + \tilde{\mathbf{G}} + \tilde{\mathbf{E}}, \quad \tilde{\mathbf{G}} \sim MN(0, V_g, K), \quad \tilde{\mathbf{E}} \sim MN(0, V_e, I_n), \tag{1}$$

where tilde superscripts distinguish the resepctive matrices from their transformed counterparts introduced below. The $p \times n$ matrix $\tilde{\mathbf{Y}}$ contains phenotypic values for p traits and n genotypes, and the matrices $\mathbf{G}$ and $\mathbf{E}$ contain the corresponding genetic and residual effects. Given an $n \times n$ genetic relatedness matrix K and $p \times p$ matrices V_g, V_e containing the genetic and residual covariances between traits, it is assumed that $\tilde{\mathbf{G}}$ follows a matrix-variate normal distribution with row covariance V_g and column covariance K; similarly, $\tilde{\mathbf{E}}$ is matrix-variate normal with row covariance V_e and column covariance I_n (the identity matrix). The $p \times c$ matrix $\mathbf{B}$ contains the trait (or environment) specific effects of c covariates, whose values are contained in the $c \times n$ matrix $\tilde{\mathbf{X}}$. Throughout this text we will always assume that $\tilde{\mathbf{X}}$ contains at least an intercept; hence $c \geq 1$. In multi-trait GWAS $\tilde{\mathbf{X}}$ will also contain marker scores; this is discussed in more detail in section A.3 below.

The sum $\tilde{\mathbf{G}} + \tilde{\mathbf{E}}$ does in general not have a matrix-variate normal distribution, but we do have that $vec(\tilde{\mathbf{G}} + \tilde{\mathbf{E}})$ is multivariate normal, where vec denotes the operation of creating a column vector by stacking the columns of a matrix:

$$\tilde{\mathbf{y}} = vec(\tilde{\mathbf{Y}}) \sim N\left(vec(\mathbf{B}\tilde{\mathbf{X}}), K \otimes V_g + I_n \otimes V_e\right), \tag{2}$$

in which $\otimes$ denotes the Kronecker product.

A.2 Diagonalization

Following again [**?**], we use the eigen-decomposition $K = U_k D_k U_k^t$, for an $n \times n$ orthogonal matrix U_k and D_k being the diagonal matrix with eigenvalues $\delta_1, \ldots, \delta_n$. If we postmultiply all terms in (1) with U_k and stack the columns, we obtain

$$\mathbf{y} = vec(\mathbf{Y}) = vec(\tilde{\mathbf{Y}}U_k) \sim N\left(vec(\mathbf{B}\mathbf{X}), D_k \otimes V_g + I_n \otimes V_e\right). \tag{3}$$

where $\mathbf{X} = \tilde{\mathbf{X}}U_k$. Models (2) and (3) do not have the same likelihood (CORRECT ?), but are equivalent in the sense that the fixed effects matrix $\mathbf{B}$ is the same. The big advantage of (3) is that the covariance matrix is now block-diagonal with blocks $V_i = \delta_i V_g + V_e$ $(i = 1, \ldots, n)$; consequently $(D_k \otimes V_g + I_n \otimes V_e)^{-1}$ is also block diagonal, with blocks V_i^{-1}. Note that this is not the case if the observations are grouped per trait (i.e. $vec(\mathbf{Y}^t)$), instead of per genotype; this is the main reason for having genotypes as rows in (1). Summarizing, after transforming $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$, inversion of the $np \times np$ covariance matrix only requires n times an inversion of a $p \times p$ matrix.

$_{121}$ A.3 Multi-trait GWAS: estimation of marker effects

$_{122}$ We now take a closer look at the fixed effects, and will from this point onwards assume
$_{123}$ that $\mathbf{X}$ is of dimension $(c+1) \times n$, and of the form $\mathbf{X} = \begin{bmatrix} W \\ x^t \end{bmatrix}$. The $p \times 1$ vector x contains
$_{124}$ the marker scores of a particular SNP, and the $c \times p$ matrix W the remaining covariates.
$_{125}$ Recall that both W and x are the transformed versions of the original covariates; in
$_{126}$ particular $x^t = \tilde{x}^t U_k$, where $\tilde{x}$ is the original vector of marker-scores (typically containing
$_{127}$ values 0,1 and 2, but also non-integer values like probabilities or dosages are allowed).
$_{128}$ To perform the GWAS, we re-fit model (3) for each SNP in turn, i.e. the vector x
$_{129}$ changes from marker to marker, while W stays constant. We also keep the estimates of
$_{130}$ V_g and V_e constant, which have been obtained for a model without SNPs (i.e., with only
$_{131}$ W). We do however allow V_g and V_e to be refitted for each chromosome, as was proposed
$_{132}$ by [?] in a univariate setting.
$_{133}$ When $\mathbf{x}_i$ denotes the ith column of $\mathbf{X}$, the GLS estimates for the regression coefficients
$_{134}$ (not only the marker-effects) are given by

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{n} (\mathbf{x}_i \mathbf{x}_i^t) \otimes \hat{V}_i^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i \otimes (\hat{V}_i^{-1} \mathbf{y}_i) \right), \tag{4}$$

where the $p \times 1$ vector $\mathbf{y}_i$ is the ith column of $\mathbf{Y}$, i.e. the transformed phenotypic values
of individual i. The last p elements in $\hat{\beta}_{GLS}$ correspond to the marker-effects. Note that
equation (65) in [?] (supplement, p.23) seems incorrect. The variance of $\hat{\beta}_{GLS}$ is given by

$$Var(\hat{\beta}_{GLS}) = \left(\sum_{i=1}^{n} (\mathbf{x}_i \mathbf{x}_i^t) \otimes \hat{V}_i^{-1} \right)^{-1}.$$

$_{135}$ In section A.4.5 below we describe the F-test for testing if the marker-effect is zero
$_{136}$ in all environments, but first we discuss numerically efficient ways to compute the effect-
$_{137}$ estimates.

$_{138}$ A.4 Partitioning the design matrix

$_{139}$ Equation (4) gives the GLS-estimate of all marker-effects as well as covariate effects.
$_{140}$ Estimates of the latter will be marker-dependent, but many of the quantities required
$_{141}$ for the estimates of the (non-marker) covariate effects are independent of the markers,
$_{142}$ and need to be evaluated only once. For this reason it is numerically more efficient if we
$_{143}$ explicitly partition β; corresponding with the different components of the design matrix.
$_{144}$ To see how this follows from well known linear algebra and classical results in regression,
$_{145}$ we will consider, just in sections A.4.1-A.4.2 below, a general design matrix X with
$_{146}$ the samples (individuals) corresponding to rows and variables to columns. To make the
$_{147}$ connection with the rest of this work, we have to choose $X = \mathbf{X}^t \otimes I_p$, where $\mathbf{X}^t = [W^t \ x]$

is of dimension $n \times (c+1)$. The $(\mathbf{X}^t \otimes I_p)$ form is also used in [?] (supplement; e.g. their equation (3)). We can now derive equation (4) from the standard GLS expression $\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y$, using that V is block diagonal with blocks V_i, and that therefore $X^t V^{-1} X$ and $X^t V^{-1}$ can be expressed in terms of $p \times p$ blocks.

A.4.1   Classical theory for ordinary least squares (OLS)

When, in OLS, the $n \times p$ design matrix X consists of blocks X_1 and X_2 of dimensions $n \times p_1$ and $n \times p_2$, for some p_1 and p_2 summing to p, the corresponding least squares estimates are given by

$$\hat{\beta}_1 = X_S^{-1} X_1^t y - X_S^{-1} (X_1^t X_2)(X_2^t X_2)^{-1} X_2^t y \tag{5}$$

$$\begin{aligned} \hat{\beta}_2 = (X_2^t X_2)^{-1} X_2^t y &+ (X_2^t X_2)^{-1} X_2^t X_1 X_S^{-1} X_1^t X_2 (X_2^t X_2)^{-1} X_2^t y \\ &- (X_2^t X_2)^{-1} X_2^t X_1 X_S^{-1} X_1^t y, \end{aligned} \tag{6}$$

where

$$X_S = X_1^t X_1 - X_1^t X_2 (X_2^t X_2)^{-1} X_2^t X_1 \tag{7}$$

is the well-known Schur-complement. These expressions can be derived by working out the block inverse $(X^t X)^{-1} = ([X_1\ X_2]^t [X_1\ X_2])^{-1}$ in

$$\hat{\beta} = (X^t X)^{-1} X^t y = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}, \tag{8}$$

see e.g. `https://www.le.ac.uk/users/dsgp1/COURSES/LEIMETZ/PARTIMOD.pdf`. For the case $p_2 = 1$, see [?], or, in the matrix cookbook ([?]), section 3.2.6 ("Rank-1 update of inverse of inner product"). Also: include related references on the Frisch-Waugh-Lovell theorem.

   Another way to obtain (5)-(6) is to define the projection matrix $P_2 = I - X_2 (X_2^t X_2)^{-1} X_2^t$ and solve the least-squares problem $P_2 y = (P_2 X_1)\beta_1 + e$. Alternatively, we can solve $P_1 y = (P_1 X_2)\beta_2 + e$ (which gives a different X_S).

A.4.2   Extension to GLS

The same can be done for GLS: when $y \sim N(X\beta, V)$, then

$$\hat{\beta}_1 = X_S^{-1} X_1^t V^{-1} y - X_S^{-1} (X_1^t V^{-1} X_2)(X_2^t V^{-1} X_2)^{-1} X_2^t V^{-1} y \tag{9}$$

$$\begin{aligned} \hat{\beta}_2 = (X_2^t V^{-1} X_2)^{-1} X_2^t V^{-1} y \\ + (X_2^t V^{-1} X_2)^{-1} (X_2^t V^{-1} X_1) X_S^{-1} (X_1^t V^{-1} X_2)(X_2^t V^{-1} X_2)^{-1} X_2^t V^{-1} y \\ - (X_2^t V^{-1} X_2)^{-1} (X_2^t V^{-1} X_1) X_S^{-1} X_1^t V^{-1} y, \end{aligned} \tag{10}$$

where

$$X_S = X_1^t V^{-1} X_1 - X_1^t V^{-1} X_2 (X_2^t V^{-1} X_2)^{-1} X_2^t V^{-1} X_1. \tag{11}$$

A.4.3 Applying the general GLS partitioning to (4)

We will now apply (9)-(11) to multi-trait GWAS, by setting $X_1 = W^t \otimes I_p$ and $X_2 = x \otimes I_p$. In each of the following equations, the left-hand side describes a matrix or vector in the notation in the R-code; after the first equality sign we express it in the generic notation of sections A.4.1-A.4.2 with X_1, X_2, and $y = \mathbf{y} = vec(\mathbf{Y})$, and on the right-hand-side we work that out in terms of kronecker products involving W and x, as in [**?**]. Let the $c \times 1$ column vector w_i denote the ith column of W and x_i the ith element of x, and recall that the covariance matrix $V = (D_k \otimes V_g + I_n \otimes V_e)$ is block diagonal with blocks $V_i = \delta_i V_g + V_e$ ($i = 1, \ldots, n$). We have:

$$\mathtt{v} = X_1^t V^{-1} \mathbf{y} = \sum_{i=1}^{n} w_i \otimes (V_i^{-1} \mathbf{y_i}) \tag{12}$$

$$\mathtt{Vbeta} = X_1^t V^{-1} X_1 = \sum_{i=1}^{n} (w_i w_i^t) \otimes V_i^{-1} \tag{13}$$

$$\mathtt{v.snp} = X_2^t V^{-1} \mathbf{y} = \sum_{i=1}^{n} x_i (V_i^{-1} \mathbf{y_i}) \tag{14}$$

$$\mathtt{Vbeta.snp} = X_2^t V^{-1} X_2 = \sum_{i=1}^{n} x_i^2 \, V_i^{-1} \tag{15}$$

$$\mathtt{X2VinvX1} = X_2^t V^{-1} X_1 = \sum_{i=1}^{n} (x_i w_i^t) \otimes V_i^{-1} \tag{16}$$

In a multi-trait GWAS the last 3 quantities need to be evaluated for each SNP (change the name X2VinvX1 to X2VinvX1.snp ?), while v and Vbeta need to be computed only once.

A.4.4 Evaluating terms per individual

In the computation of v.snp, Vbeta.snp and X2VinvX1.snp we want to avoid a loop over all SNPs. Instead we loop over the individuals, and for each i, the corresponding terms in equations (14)-(16) are evaluated for all SNPs simultaneously, using the tcrossprod function. When the number of individuals and/or SNPs is large, this may only be feasible for blocks of SNPs (because of memory contraints), but even then this strategy still seems faster.

Once all quantities in equations (12)-(16) have been computed, the remaining calculations are those in equations (9)-(11), and these are done in a loop over the SNPs. For each SNP, the effect-estimates for the p environments are given by $\hat{\beta}_2$ in (10), and the covariate effects by $\hat{\beta}_1$ in (9). The latter are only needed if fitted-values and p-values are

required; note that by itself, $\hat{\beta}_2$ can be evaulated without evalualating $\hat{\beta}_1$. The SNP and covariate effects are stored in matrices `EFFECTS` and `EFFECTS.cov`, respectively.

For the standard errors and p-values, we also require for each SNP the matrix

$$
\texttt{Q.matrix.inv} = \left(\begin{array}{cc} \texttt{Vbeta} & \texttt{X2VinvX1} \\ (\texttt{X2VinvX1})^t & \texttt{Vbeta.snp} \end{array} \right)^{-1} = \left(\begin{array}{cc} X_1^t V^{-1} X_1 & X_2^t V^{-1} X_1 \\ X_1^t V^{-1} X_2 & X_2^t V^{-1} X_2 \end{array} \right)^{-1}.
\tag{17}
$$

The standard errors of covariate and SNP effects (in that order) are given by the square root of the diagonal of this matrix. During the compuation of the SNP-effects, we store the standard errors of the SNP-effects in `EFFECTS.se`. For later use we also store, in vectorized form, the matrix `Q.matrix.inv`. A $((c+1)p)^2 \times n_s$ matrix `Q.matrix.inv.all` contains `Q.matrix.inv` for all SNPs $j = 1, \ldots, n_s$.

A.4.5   The likelihood ratio test (LRT), the F-test and p-values

Apart from the standard errors for all effects, it is usually of interest to test various statistical hypotheses for each SNP. The most common null-hypothesis is that all effects are zero ($\beta_1 = \ldots = \beta_p = 0$) which is tested against the alternative that at least on of the effects is nonzero. As discussed in [?], this hypothesis can be tested using the likelihood ratio test (LRT), the Wald test and the score test. Since [?] observed that the latter two usually give inflated respectively deflated p-values, we restrict our attention here to the LRT.

The LRT is defined as twice the difference between the log-likelihood of model (3) with $\mathbf{BX}$ containing both the SNPs and the covariates, and the model with only the covariates. Since we use the same V_g and V_e for both models (and also for all SNPs), the (log) determinants of the corresponding Gaussian distributions are the same and cancel out, and the LRT is the difference between the (generalized) sums of squares of the two models, which are quadratic forms of the form

$$
(\mathbf{y} - \hat{\mathbf{y}})^t \left(D_k \otimes V_g + I_n \otimes V_e \right)^{-1} (\mathbf{y} - \hat{\mathbf{y}}),
\tag{18}
$$

where we recall that $\mathbf{y} = vec(\mathbf{Y})$, and $\hat{\mathbf{y}}$ is the vector of fitted values. For the full model, the latter depends on the SNP as well as covariate effects. Using the classical result of ... (add reference), [?] (supplement, p. 19, equ. (26)-(29)) re-expressed (18) as

$$
\mathbf{y}^t P \mathbf{y},
\tag{19}
$$

where P is defined in their equation (29). See also [?].

Note that we use the ML-based LRT; for the REML-based LRT we get the extra terms of equation (27) in [?].

Instead of the (ML) LRT, we use the related F-test

$$
F = \frac{df_{full}}{df_{red} - df_{full}} \left(\frac{SS_{red} - SS_{full}}{SS_{full}} \right) = \frac{(n-c-1)p}{p} \left(\frac{SS_{red} - SS_{full}}{SS_{full}} \right),
\tag{20}
$$

₂₂₇ which, under the null-hypothesis, has an F-distribution with degrees of freedom $df1 = $
₂₂₈ $df_{red} - df_{full} = p$ and $df2 = df_{full} = (n - c - 1)p$.

₂₂₉ To compute F we need SS_{red} and SS_{full}, i.e. equations (18) and (19) for the reduced
₂₃₀ and the full model. The former is the same for all SNPs, and we evaluate it with our
₂₃₁ `LL.quad.form.diag` function. It would be convenient to use the same for SS_{full}, but this
₂₃₂ would require storing all np residuals for all n_s SNPs. To avoid this we use equation (50)
₂₃₃ of [?], on p.21: $\mathbf{y}^t P \mathbf{y} = q - \mathbf{q}^t \mathbf{Q}^{-1} \mathbf{q}$, where the terms on the right-hand-side are define
₂₃₄ at the bottom of p.20. This formula is also used in the `LL.quad.form.diag` function.
₂₃₅ However, instead of directly looping over all SNPs and calling `LL.quad.form.diag` each
₂₃₆ time, we first compute the terms $\mathbf{Q}^{-1}$, $\mathbf{q}$ and q for all SNPs simultaneously, looping over
₂₃₇ the individuals.

B   Simulation

₂₃₉ Suppose we have variance-covariance matrices $S \in \mathbb{R}^{n \times n}$ and $T \in \mathbb{R}^{p \times p}$. When Z is an $n \times p$
₂₄₀ matrix whose elements follow independent standard normal distributions, $Y = S^{\frac{1}{2}} Z T^{\frac{1}{2}}$ fol-
₂₄₁ lows the matrix-variate normal distribution $MN(M; S, T)$ (`https://stat.ethz.ch/pipermail/r-help`
₂₄₂ find a reference).

B.1   R-code

```
# http://realizationsinbiostatistics.blogspot.com/2008/08/matrix-square-roots-in-r_18.
MatrixRoot <- function(x) { # assumes that x is symmetric
x.eig <- eigen(x,symmetric=TRUE)
x.sqrt <- x.eig$vectors %*% diag(sqrt(x.eig$values)) %*% solve(x.eig$vectors)
return(x.sqrt)
}
```

References