Obtaining estimates of marker effects and their standard errors from estimates of genomic breeding values

1 Introduction

Mixed model analyses fitting the so-called animal model have been shown to account for 'structure' in the data from relationships between individuals in genome wide association analyses (GWAS). In particular, EMMAX – standing for Efficient Mixed Model Analysis eXpedited (Kang et al., 2010) – is widely used. This involves a mixed model fitting a single marker at a time as a *fixed* covariable in addition to a random effect representing individuals' additive genetic effects (*a.k.a.* breeding values) with given relationship matrix. Estimates of variance components are usually obtained from a preliminary REML analysis omitting such covariables and potential concerns about double counting due to including the same marker in constructing a genomic relationship matrix are considered negligible (Chen et al., 2017).

Early applications executed a separate best linear unbiased prediction (BLUP) analysis for each marker which proved laborious and time consuming. Several authors developed computationally more efficient implementations. In particular, Meyer and Tier (2012) proposed a strategy dubbed "SNP Snappy", exploiting that only the values for the marker covariable differ between individual BLUP runs. This reduces the computational burden of EMMAX dramatically by carrying out the Cholesky composition of the coefficient matrix for all equations other than the marker covariable once and then processing the final equation only, considering all markers sequentially in the same analysis. This has been implemented in WOMBAT since 2012 and is available via the run option `--snap`.

More recently, it has been shown that EMMAX test statistics can be obtained by carrying out a standard BLUP analysis fitting an animal model with a genomic relationship matrix (GBLUP) (excluding marker effect covariables). Predicted marker effects are then obtained as 'back solutions', through a linear transformation of the predicted breeding values. This allows their standard errors to be determined as a function of the corresponding parts of the inverse of the coefficient matrix in the mixed model equations (MME) (Gualdrón Duarte et al., 2014; Bernal Rubio et al., 2016; Chen et al., 2017; Legarra et al., 2018). Surprisingly, while the values of predicted, random marker effects and fixed covariables and the corresponding standard errors differ, the resulting test statistics for the two approaches, i.e. the ratios of estimates divided by their standard errors, have the same values; see the appendix of Bernal Rubio et al. (2016) for an algebraic proof.

Derivations in the literature mainly considered univariate scenarios, as the extension to

multivariate cases simply involves the same calculations considering one trait at a time (Lu et al., 2018), provided the covariance matrices between traits for markers and breeding values are the same or proportional. While analyses fitting markers as fixed covariables require genotype information for all individuals with records, EMMAX via transformation of predicted breeding values is equally applicable to so-called 'single step' genomic BLUP (ssGBLUP) and thus readily accommodates the use of data from ungenotyped individuals in addition (Aguilar et al., 2019). On the other hand, it requires the inverse of the coefficient matrix of the MME, while its Cholesky factorisation suffices for calculations via "SNP Snappy".

An option to carry out 'EMMAX via linear transformation' (subsequently referred to as ssGWAS) at the end of a BLUP or REML run recently has been added to WOMBAT, and this note describes its use.

2 Background

Let $\mathbf{M}$ denote the matrix of marker counts or "gene content" (of size number of genotyped animals × number of markers) and $\mathbf{P}$ the corresponding matrix of assumed frequencies, p_i. The genomic relationship matrix (GRM) is then calculated from the centered marker counts. Popular forms are

$$\mathbf{G}_M = (\mathbf{M} - 2\mathbf{P})\,(\mathbf{M} - 2\mathbf{P})'\,/s = \mathbf{Z}\mathbf{Z}'/s \tag{1}$$

or

$$\mathbf{G}_M = (\mathbf{M} - 2\mathbf{P})\,\mathbf{W}^{-1}\,(\mathbf{M} - 2\mathbf{P})'\,/m = \mathbf{Z}\mathbf{W}^{-1}\mathbf{Z}'/m \tag{2}$$

as described by Van Raden (2008) or Yang et al. (2010), respectively, where $\mathbf{W}$ is a diagonal matrix with elements $2p_i(1-p_i)$, m denotes the number of markers and $s = 2\sum_i p_i(1-p_i)$. $\mathbf{G}_M$ is then often modified to ensure that it can 'safely' be inverted, to improve alignment between GRM and pedigree based relationships or to account for residual polygenic variation. Common types of modifications, especially for ssGBLUP, can be summarised as

$$\mathbf{G} = \lambda\left[\beta(\mathbf{G}_M + \epsilon\mathbf{I}) + \alpha\mathbf{J}\right] + (1-\lambda)\,\mathbf{A}_{22} \tag{3}$$

with $0 \leq \lambda \leq 1$ denoting the proportion of total genetic variance due to marker effects, $\mathbf{A}_{22}$ the part of the pedigree based relationship matrix for genotyped individuals, α and β the 'alignment' factors proposed by Christensen (2012) or Vitezica et al. (2011), $\mathbf{J}$ a matrix with all elements equal to unity, $\mathbf{I}$ an identity matrix and ϵ a small constant.

2.1 Predicted marker effects and their variance

Utilising the equivalence between a model which accounts for additive genetic effects by fitting marker effects directly and GBLUP (Strandén and Garrick, 2009), we can write the vector of breeding values, $\mathbf{u}$, as a function of the marker effects, $\mathbf{a}$,

$$\hat{\mathbf{u}} = \mathbf{Z}\,\hat{\mathbf{a}} \tag{4}$$

Predictions for $\mathbf{a}$ can thus be obtained from the predicted breeding values, regressing $\mathbf{a}$ on $\mathbf{u}$. Similarly, standard errors of the elements of $\hat{\mathbf{a}}$ can be determined from the matrix of prediction error variances of $\hat{\mathbf{u}}$.

Gualdrón Duarte et al. (2014) consider $\mathbf{G} = \mathbf{G}_M$ of form (Eq. 2), incorporating weighting and scaling into $\mathbf{G}$ by defining a matrix $\tilde{\mathbf{Z}}$ with elements $\tilde{z}_{ij} = (m_{ij} - 2p_i)/\sqrt{2mp_i(1-p_i)}$, so that $\mathbf{G} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}'$. This gives

$$\begin{aligned}\hat{\mathbf{a}} &= \text{Cov}(\mathbf{a}, \mathbf{u}') \,\text{Var}(\mathbf{u})^{-1} \hat{\mathbf{u}} \\ &= \sigma_a^2 \sigma_u^{-2} \tilde{\mathbf{Z}}' \mathbf{G}^{-1} \hat{\mathbf{u}} = \tilde{\mathbf{Z}}' \mathbf{G}^{-1} \hat{\mathbf{u}}\end{aligned} \tag{5}$$

for $\text{Var}(\mathbf{a}) = \sigma_a^2 \mathbf{I}$, $\text{Var}(\mathbf{u}) = \sigma_u^2 \mathbf{G}$ and assuming that, after scaling, $\sigma_a^2 = \sigma_u^2$. It follows that

$$\text{Var}(\hat{\mathbf{a}}) = \tilde{\mathbf{Z}}' \mathbf{G}^{-1} \,\text{Var}(\hat{\mathbf{u}})\, \mathbf{G}^{-1} \tilde{\mathbf{Z}} \tag{6}$$

with

$$\text{Var}(\hat{\mathbf{u}}) = \sigma_u^2 \mathbf{G} - \mathbf{C}^{uu} \tag{7}$$

and

$$\text{Var}(\hat{\mathbf{a}}) = \sigma_u^2 \tilde{\mathbf{Z}}' \mathbf{G}^{-1} \tilde{\mathbf{Z}} - \tilde{\mathbf{Z}}' \mathbf{G}^{-1} \mathbf{C}^{uu} \mathbf{G}^{-1} \tilde{\mathbf{Z}} \tag{8}$$

where $\mathbf{C}^{uu}$ is the part of the inverse of the coefficient matrix in the MME corresponding to $\hat{\mathbf{u}}$.

Recently, Aguilar et al. (2019) presented an extension to single-step analyses, considering $\mathbf{G}$ of form in (Eq. 3), with $\mathbf{G}_M = \mathbf{Z}\mathbf{Z}'/s$ (Eq. 1). Formulae given were

$$\hat{\mathbf{a}} = \lambda\beta \left[2\textstyle\sum_i p_i(1-p_i)\right]^{-1} \mathbf{Z}' \mathbf{G}^{-1} \hat{\mathbf{u}}_2 = F\, \mathbf{Z}' \mathbf{G}^{-1} \hat{\mathbf{u}}_2 \tag{9}$$

and

$$\text{Var}(\hat{\mathbf{a}}) = F^2 \mathbf{Z}' \mathbf{G}^{-1} \left[\sigma_u^2\, \mathbf{G} - \mathbf{C}_{22}^{uu}\right] \mathbf{G}^{-1} \mathbf{Z} \tag{10}$$

with $\hat{\mathbf{u}}_2$ the subvector of $\hat{\mathbf{u}}$ for genotyped individuals and $\mathbf{C}_{22}^{uu}$ the corresponding part of $\mathbf{C}^{uu}$.

2.2 Significance testing

Test statistics for individual markers are obtained by dividing estimates $\hat{a}_i$ by their standard errors,

$$t_i = \hat{a}_i / \sqrt{\text{Var}(\hat{a}_i)} \tag{11}$$

with probability (p-value)

$$pv_i = 2\,(1 - \Phi(|t_i|)) \tag{12}$$

where $\Phi(x)$ denotes the value of cumulative density function of the standard Normal distribution at x.

Note that scaling factors used by Gualdrón Duarte et al. (2014) and Aguilar et al. (2019) differ slightly – the former authors scale $\mathbf{Z}$ by the equivalent to $\sqrt{2\sum p_i(1-p_i)}$ while the latter use $2\sum p_i(1-p_i)$. However, as the same factor enters both the numerator and denominator of the ratio t_i, the test statistic is invariant to this difference.

2.3 Equivalent EMMAX regression coefficient

Using results from Bernal Rubio et al. (2016), Aguilar et al. (2019) emphasize the possibility to convert $\hat{\mathbf{a}}$ to the estimates of fixed regression coefficients (b_i) from a corresponding EMMAX analysis. For $\mathbf{G}_M$ of form (Eq. 1),

$$b_i = F\,\sigma_u^2\,\hat{a}_i / \mathrm{Var}\,(\hat{a}_i) \tag{13}$$

(Aguilar et al. (2019) omit the factor F). For a weighted $\mathbf{G}_M$ (Eq. 2),

$$b_i = F\,\sigma_u^2\,\hat{a}_i / \left(\mathrm{Var}\,(\hat{a}_i)\ \sqrt{w_i}\right) \tag{14}$$

with $w_i = 2p_i(1-p_i)$.

3 Notes on implementation

Implementation of ssGWAS in WOMBAT follows the procedure outlined by Aguilar et al. (2019).

It is assumed that marker solutions and corresponding statistics are required for one random effect representing additive genetic effects only.

Solutions (and standard deviations) reported for marker effects are scaled so that 'reconstructing' breeding values for genotyped individuals for $\lambda = \beta = 1$ as $\tilde{\mathbf{u}}_2 = \mathbf{Z}\hat{\mathbf{a}}$ (or $\tilde{\mathbf{u}}_2 = \mathbf{Z}\mathbf{W}^{1/2}\hat{\mathbf{a}}$) yields a regression of $\hat{\mathbf{u}}_2$ on $\tilde{\mathbf{u}}_2$ close to unity.

Major computational requirements are imposed by

1. The need to invert and store the inverse of the coefficient matrix in the MME.
2. For efficiency, the complete matrix of marker counts is currently held in core. This requires an array of size $n_2 \times m$, where n_2 denotes the number of genotyped individuals. For large numbers of genotypes or markers this may require excessive RAM.
3. The current implementation also requires an additional array of size $n_2 \times n_2$, used to hold dense matrices $\mathbf{G}^{-1}$ and $\mathbf{C}_{22}^{uu}$.

The additional RAM required may be reduced by considering subsets of markers to be processed simultaneously or 'packed' storage of $\mathbf{G}^{-1}$, and may be implemented in the future.

4 Specifications

Execution of ssGWAS at the end of a BLUP or REML analysis can be invoked by adding a single line within a **SPECIAL** block in the parameter file:

```
BSOLVE-SNP rname
```

where *rname* denotes the name of the random effect in the model of analysis representing additive genetic effects which is to be used to obtain marker test statistics.

Additional information required is expected to be supplied in the same form as for the pre-analysis module to calculate the GRM, its inverse or $\mathbf{H}^{-1}$, invoked via run option **--hinv**; see the documentation for Example 20 for details. Again these consist of lines added to the **SPECIAL** block in the parameter file:

a) The number of markers to be considered is specified as

```
HINVERSE SNP m
```

where m represents the number of markers.

b) The weighting factor λ to combine genomic and pedigree based relationships in building **G** for single step analyses. This has a default value of 1. Other values can be specified by adding the line

   `HINVERSE LAMBDA` λ

   with $0 < \lambda \leq 1$.

c) The factor β used to scale **G** for single step analyses. Again, this has a default value of 1 and other values need to be specified by adding the line

   `HINVERSE BETA` β

d) The method used to build the GRM. Current values recognised are **VRADEN1** for Van Raden (2008)'s Method 1 and **YANG** for Yang et al. (2010)'s method. The line to be added is

   `HINV HOWGRM` *method*

e) A keyword to specify what frequencies were used to center marker counts when building the GRM can be given as

   `HINV CENTER` *keyword*

   Currently, keywords recognised are **FREQ** for 'observed' frequencies calculated from the marker counts used to build **G** (default) and **HALF** for all frequencies assumed to be equal to 0.5.

Lines with appropriate default values are optional. If WOMBAT has been used to compute $\mathbf{G}^{-1}$ or $\mathbf{H}^{-1}$, an auxiliary file **HinvMeta** is written out which gives the settings for the many options available. If this file is found in the working directory, WOMBAT will acquire the values of the above options from it (NB: If found, values from this file will take precedence over any specification in the parameter file).

5 Additional input

5.1 Marker allele counts

ssGWAS requires the file with marker allele counts used to construct the GRM, coded as 0, 1 or 2. It is assumed to be in the same form as required for use in WOMBAT with run option **--hinv**.

File name. The file has the default name **MarkerCounts.dat**. The default extension **.dat** implies a formatted file. If this is not found, WOMBAT will attempt to open and read from **MarkerCounts.BIN** or **MarkerCounts.BI1** as an unformatted file, which can be faster.

An alternative filename can be used but requires adding the line

```
MRK filename
```

to the parameter file (where **filename** can have the same extensions as above; see the WOMBAT manual for details).

File layout. As for the single-step modules, each 'row' for an animal is expected to begin with the individual code (matching the codes in the data and .codes files) followed by the marker allele counts. This is a list-directed FORTRAN read, reading the individual code as as a standard length INTEGER and the allele counts as single precision REAL or INTEGER*1 variable. For large numbers of markers, the 'row' can be spread over several lines in the input file (but the next individual must always start with a new line).

NB: For ssGWAS and **--hinv**, animal codes are NOT used – instead it is assumed that allele counts are given in the SAME sequence as the genotyped individuals occur in the list of codes, in increasing numerical order!

Marker counts must be complete. i.e. specified for all animals marked as genotyped in the pedigree file and missing counts for individual markers are not accommodated. Note also that WOMBAT does NOT perform any checks or quality control on the contents of this file.

5.2 Type of individual

To use genomic relationships, BLUP or REML analyses in WOMBAT require the inverse of the GRM or $\mathbf{H}^{-1}$ to be supplied by the user. This needs to accompanied by a corresponding **.codes** file which lists original codes identifying individuals and their running numbers. For ssGWAS, this file is expected to have a third column with value "2" if the individual has genotype information and "1" if it doesn't.

If the **.gin** file with $\mathbf{H}^{-1}$ or $\mathbf{G}^{-1}$ has been set up using WOMBAT, this additional column is automatically added to the corresponding **.codes** file generated at the same time.

5.3 Inverse of the GRM

For analyses with all animals genotyped, the **.gin** matrix supplied is equal to the inverse of the GRM and $\mathbf{G}^{-1}$ is acquired from it.

Otherwise (i.e. when $\mathbf{G}^{-1} \neq \mathbf{H}^{-1}$), a separate file containing $\mathbf{G}^{-1}$ is required. This is expected to be a formatted file with the default name of **GRMInv.dat**. It should contain one line for each non-zero element comprised of three space-separated columns containing the row number (INTEGER), column number (INTEGER) and coefficient (REAL8), respectively, where row and column numbers are 'running numbers', i.e. 1 to number of genotyped animals.

Alternatively, for $\lambda = 1$, $\mathbf{G}_M$ can be supplied via a file with the default name of **GRM.dat**, with the same format as **GRMInv.dat**. If this is given, WOMBAT will construct $\mathbf{G}$ of form (Eq. 3) using the values of ϵ, α and β supplied or set by default. A Cholesky factorisation of $\mathbf{G}$ is then performed and $\mathbf{G}^{-1}\mathbf{Z}$ is obtained by solving $\mathbf{GX} = \mathbf{Z}$ for $\mathbf{X}$ (without inverting $\mathbf{G}$).

If WOMBAT with run option **--hinv** is used to generate the **.gin** matrix and $\mathbf{G}^{-1}$ or $\mathbf{G}$ are needed the line

```
HINV out GRMINV
```

or **HINV out GRM**, respectively, can be added to the parameter file for that preliminary run to generate these additional output files.

6 Output

Results are written to a formatted file with standard name **BackSoln_SNPeffects.dat**. This contains one line per marker and trait comprised of 8 columns:

Column 1 gives the running number of the marker,
Column 2 gives the trait number,
Column 3 gives the solution for the marker effect,
Column 4 gives the corresponding standard error,

Column 5 gives the ratio of effect solution to its standard error,
Column 6 gives the $p-$value for the test statistic,
Column 7 gives log 10 of the $p-$value, and
Column 8 gives the solution converted to an EMMAX fixed regression coefficient.

BackSoln_SNPEffects.dat

SNPNo.	Trait	Solution	S.Error	Ratio	p-value	-log10(p)	EMMAX
1	1	-0.208986E-01	0.167423	-0.12482552	0.90066168	0.04543831	-0.648366E-01
2	1	-0.623655E-02	0.175305	-0.03557536	0.97162096	0.01250313	-0.176476E-01
3	1	0.611629E-01	0.160678	0.38065612	0.70345844	0.15276156	0.206020
4	1	0.656210E-01	0.159621	0.41110448	0.68099592	0.16685549	0.223972
5	1	-0.996457E-02	0.177951	-0.05599628	0.95534477	0.01983987	-0.273647E-01
6	1	0.397005E-01	0.168285	0.23591172	0.81350117	0.08964182	0.121909
7	1	-0.347068	0.165186	-2.10107358	0.03563451	1.44812924	-1.10611

References

Aguilar I., Legarra A., Cardoso F., Masuda Y., Lourenco D., Misztal I. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genet. Sel. Evol. 51 (2019) 28. doi: 10.1186/s12711-019-0469-3.

Bernal Rubio Y.L., Gualdron Duarte J.L., Bates R.O., Ernst C., Nonneman D., Rohrer G.A., King A., Shackelford S.D., Wheeler T.L., Cantet R.J.C., Steibel J.P. Meta-analysis of genome-wide association from genomic prediction models. Animal Genetics 47 (2016) 36–48. doi: 10.1111/age.12378.

Chen C., Steibel J.P., Tempelman R.J. Genome-wide association analyses based on broadly different specifications for prior distributions, genomic windows, and estimation methods. Genetics 206 (2017) 1791–1806. doi: 10.1534/genetics.117.202259.

Christensen O.F. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet. Sel. Evol. 44 (2012) 37. doi: 10.1186/1297-9686-44-37.

Gualdrón Duarte J.L., Cantet R.J.C., Bates R.O., Ernst C.W., Raney N.E., Steibel J.P. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics 15 (2014) 246. doi: 10.1186/1471-2105-15-246.

Kang H.M., Sul J.H., Service S.K., Zaitlen N.A., Kong S.Y., Freimer N.B., Sabatti C., Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nature Genet. 42 (2010) 348–354. doi: 10.1038/ng.548.

Legarra A., Ricard A., Varona L. GWAS by GBLUP: Single and multimarker EMMAX and Bayes factors, with an example in detection of a major gene for horse gait. G3: Genes, Genomes, Genetics 8 (2018) 2301–2308. doi: 10.1534/g3.118.200336.

Lu Y., Vandehaar M.J., Spurlock D.M., Weigel K.A., Armentano L.E., Connor E.E., Coffey M., Veerkamp R.F., de Haas Y., Staples C.R., Wang Z., Hanigan M.D., Tempelman R.J. Genome-wide association analyses based on a multiple-trait approach for modeling feed efficiency. J. Dairy Sci. 101 (2018) 3140 – 3154. doi: 10.3168/jds.2017-13364.

Meyer K., Tier B. "SNP Snappy": A strategy for fast genome wide association studies fitting a full mixed model. Genetics 190 (2012) 275–277. doi: 10.1534/genetics.111.134841.

Strandén I., Garrick D.J. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92 (2009) 2971 – 2975. doi: 10.3168/jds.2008-1929.

Van Raden P.M. Efficient methods to compute genomic predictions. J. Dairy Sci. 91 (2008) 4414–4423. doi: 10.3168/jds.2007-0980.

Vitezica Z.G., Aguilar I., Misztal I., Legarra A. Bias in genomic predictions for populations under selection. Genet. Res. 93 (2011) 357–366. doi: 10.1017/S001667231100022X.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E., Visscher P.M. Common SNPs explain a large proportion of the heritability for human height. Nature Genet. 42 (2010) 565–569. doi: 10.1038/ng.608.