'Single-step' genetic evaluation in WOMBAT

1 Estimating breeding values for geno- and phenotyped animals

The so-called 'single step' procedure has been proposed for genetic evaluation utilising records for all genotyped and phenotyped individuals and all pedigree information available simultaneously (e.g Misztal et al., 2009). It utilises genomic information to construct the 'realized' or genomic relationship among genotyped individuals, which is then combined with the pedigree based relationships with and among the non-genotyped animals. The resulting relationship matrix is commonly denoted as $\mathbf{H}$, and rules to obtain its inverse, $\mathbf{H}^{-1}$, directly from the pedigree information for all individuals and the inverse of the genomic relationship matrix are available (e.g. Aguilar et al., 2010).

Ordering individuals so that n_1 animals without genotypes precede n_2 which have been genotyped, $\mathbf{H}^{-1}$ is given by the inverse of the pedigree based relationship, $\mathbf{A}^{-1}$ among all $n_1 + n_2$ individuals, and a matrix of size $n_2 \times n_2$ added to the corresponding lower diagonal block of $\mathbf{H}^{-1}$, $\mathbf{H}^{22}$, which, in essence, adjusts for the difference between genomic ($\mathbf{G}$) and pedigree ($\mathbf{A}_{22}$) derived relationships. As $\mathbf{G}$ and $\mathbf{G}^{-1}$ are dense, i.e. have few non-zero elements, so is $\mathbf{H}^{22}$.

A module has been added to WOMBAT to solve a set of mixed model equations iteratively using a preconditioned conjugate gradient (PCG) algorithm which accounts for this structure. The key feature of the new module is a 'hybrid' scheme combining sparse matrix storage and processing with its dense counterparts. In addition it allows for fitting of genetic groups in an explicit manner, i.e. by fitting groups as an additional random effect, specifying the proportions of membership for each animal. In detail:

- Animals are classed as genotyped and non-genotyped.
- The diagonal block(s) for genetic effects of genotyped animals and are stored in full, using one triangle of a two-dimensional array. This reduces both memory requirements and access time for individual elements compared to sparse storage, and facilitates the use of standard library routines for matrix calculations. If fitted, explicit genetic groups are treated similarly.
- For standard multivariate analyses, genetic effects of all q traits for genotyped animals are held in a block of size $q\,n_2 \times q\,n_2$. For principal components (PC) models, fitting r components, the are stored in r separate blocks of size $n_2 \times n_2$.
- The remaining non-zero elements in the coefficient matrix are held in sparse form, using compressed row storage. WOMBAT uses INTEGER*4 addressing for this, i.e. the number of such elements cannot exceed 2 147 483 647.
  Update: a version of WOMBAT using INTEGER*8 addressing is now available.
- A partial Cholesky decomposition of the coefficient matrix is used as pre-conditioner. For the dense diagonal blocks, these are stored in the other triangle of the array used, using an additional array for diagonal elements. This can improve convergence rates as solutions for all n_2 genotyped individuals are updated simultaneously, albeit

at the expense of additional computations peer iterate.
- Dense matrix calculations are performed using BLAS and LAPACK library routines where appropriate, and execution using multiple processors is available through multi-threaded versions of these libraries.

1.1 Specifying a 'single step' run

The 'single step' module is invoked through the command line option **--s1step**.

```
wombat --s1step
```

The default for the PCG algorithm has been changed to the so-called SSOR preconditioner; see Meyer (2016). There are now five additional choices to vary the preconditioner by appending the letter A, B, C, D or E:

--s1stepA invokes a simple diagonal preconditioning scheme.

--s1stepB provides a blockdiagonal preconditioner with the blocks equal to the effects for all traits in a multivariate analysis. However, if genetic groups are fitted as an additional random effect, the diagonal block for all levels (and traits) is used.

--s1stepC is like **--s1stepB** except that the full diagonal block for genotyped animals (all levels and traits) is utilised.

--s1stepD uses full blocks for genotyped animals and genetic groups for preconditioning (as for **--s1stepC**), but this is combined with the diagonal preconditioner for the remaining equations

--s1stepE provides a diagonal preconditioner for all effects other than genetic groups (equivalent to B for univariate nalyses).

The convergence criterion used for the PCG algorithm is the sum of squared changes in solutions between iterates divided by the sum of squared solutions. The default threshold is 10^{-8}. A different value can be set by specifying **--s1step***k*, with *k* denoting an integer value, which sets this limit to 10^{-k} (this can be combined with modifiers 'a' or 'b' but needs to follow these, e.g. **--s1stepa6** or **--s1stepb10**).

1.2 Model and input needed

The 'single step' module expects the animal genetic effect to be specified as a random effect with a user-defined covariance matrix, denoted by the option **GIN** in the parameter file. For example:

```
MODEL
      ...
   RAN animal  GIN
      ...
END
```

WOMBAT then expects to find an input file **effectname.gin** (e.g. **animal.gin**) which contains the non-zero elements of the lower triangle of $\mathbf{H}^{-1}$. This should be set up as specified in the manual, i.e.

1. The first line should give the log determinant of the general covariance matrix. This is not needed in the 'single step' modules, i.e. any value can be given as placeholder.
2. The following lines, one per non-zero element should each contain three space-separated variables:

(a) An INTEGER code for the 'column' number
(b) An INTEGER code for the 'row' number
(c) A REAL variable specifying the element of the inverse

Here 'row' and 'column' numbers should range from 1 to N, where N is the number of levels for the random effect. Only the elements of the *lower* triangle of the inverse should be given, i.e. WOMBAT expects a 'column' number which is less than or equal to the 'row' number.

WOMBAT does not offer any options to set up $\mathbf{H}^{-1}$ – while it would be easy to set up the mechanics, quality control of genomic information and proper scaling of the genomic relationship matrix compared to $\mathbf{A}$ are crucial in this context and thus left to the individual.

In addition, WOMBAT requires a file **`effectname.codes`** (here **`animal.codes`**) which provides the mapping between 'running' numbers (1 to N) and codes in the data file. In addition, with the **`--s1step`** option, WOMBAT expects to acquire a code specifying whether animals have genotypes or not from this file. It should thus contain one line for each animal with three space-separated variables:

(a) An INTEGER code with the 'running number (1 to N)
(b) An INTEGER code with the original code in the data file
(c) An INTEGER variable equal to '1' for individuals not genotyped and '2' for individuals with genotype information.

There are no specific requirements for the data file, and a pedigree file is not needed.

2 Fitting genetic groups 'explicitly'

Genetic groups are generally taken into account by incorporating 'phantom parents' in the linear mixed model, as proposed by Westell et al. (1988). However, initial experience has found this to be problematic when used in conjunction with the 'single step' approach, as a non-zero difference between $\mathbf{G}$ and $\mathbf{A}_{22}$ may not be carried through appropriately (Misztal et al., 2013). An alternative is to revert to a model which fits genetic groups explicitly,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{ZQg} + \mathbf{e}$$

with $\mathbf{y}$, $\mathbf{b}$, $\mathbf{u}$, $\mathbf{g}$ and $\mathbf{e}$ denoting the vectors of observations, fixed effects, random effects, group effects and residuals, respectively, and $\mathbf{X}$ and $\mathbf{Z}$ the corresponding design matrices. $\mathbf{Q}$ is the matrix linking individuals to groups, with each row of $\mathbf{Q}$ potentially containing multiple non-zero elements, tracing the contribution of genetic groups through the pedigree, which sum to unity.

An option has been added to WOMBAT to allow for such group effects to be fitted as random effects.

2.1 Specifying genetic groups

Instructing WOMBAT to fit genetic groups requires these to be specified as a random effect in the model of analysis, and the corresponding (co)variance components statements need to be provided. In addition, an entry in a `SPECIAL` block at the bottom of the parameter file is needed. This should consist of a single line, with the following, space separated entries:

(a) The code `GENGROUPS` at the start of the line.

(b) The name of the random effect in the model representing genetic group effects.

(c) An INTEGER variable giving the number of genetic groups to be fitted (NB: This must be the actual number, not just some number at least as big).

(d) A variable giving a scale factor or special option:

- When proportions of membership to genetic groups (i.e. elements of the so-called **Q**-matrix) are supplied, the variable is the scale factor by which the proportions have been multiplied.
- If the variable has a value of −9, it signifies that the genetic group codes for sires and dams are to be read instead. These are used by WOMBAT to build the **Q**-matrix in core.

NEW

(e) The name of the random effect in the model representing individuals' additive genetic effect corresponding to the genetic groups. If this is omitted, it is assumed to be the first random effect with the **GIN** covariance option found in the model specified.

For example,

```
MODEL
       ...
# specify animal genetic effects first
  RAN animal  GIN
  RAN gggrps  IDE
       ...
END
       ...
SPECIAL
# specify which random effect represents groups,
# no. of levels and scale factor for proportions
  GENGROUPS gggrps 27 1  animal
       ...
END
```

2.2 Input files

2.2.1 Q matrix

The elements of the rows of **Q** are expected to be read as a string of g space separated INTEGER values (with g the number of groups as given in the **SPECIAL** block) for each individual. These integer values are assumed to have been obtained by multiplying the proportions with the scale factor specified. While such values are expected for all individuals, only those for individuals represented in the data (i.e. with phenotypes) are used. For analyses with a 'general' covariance matrix with is inverse supplied in a ***.gin** file, these values need to be given in the ***.codes** file, following (space separated) the running and original identities and the type code, i.e. from column 4 onwards.

For example, for $g = 5$ and a scale factor of 1000, the ***.codes** file might be:

```
1 2301   1   900     0   100     0     0
2 2302   1     0   900     0     0   100
3 2303   2   100     0   800   100     0
4 2304   2     0     0     0     0  1000
            ...
```

with columns 1 and 2 the running number and original animal code, column 3 the type of animal (1 = non genotyped, 2 = genotyped; for compatibility, this code needs to be included for all run options if genetic groups are fitted, including those that do not treat genotyped animals differently). Columns 4 to 8 then give the proportions of group membership, multiplied by the scale of 1000 for each of the 5 groups.

N.B.: WOMBAT does not perform any checks on the elements of **Q** given! The groups proportions given are treated as unique for each individual, i.e. are assumed to be the same for all traits in multivariate analyses.

Similarly, for analyses fitting the pedigree based relationship matrix, proportions can be supplied in the pedigree file, from column 5 onwards. If $\mathbf{H}^{-1}$ for a single-step analysis is built using WOMBAT (Run option **`--hinv`**; see Example 20), the program will transfer the group proportions from the pedigree to the **`*.codes`** file generated.

2.2.2 Unknown parent group codes

If it is chosen to supply unknown parent group codes instead, these are expected to be given as columns 5 (sire) and 6 (dam) in the pedigree or columns 4 (sire) and 5 (dam) in the **`*.codes`** file. NB: For this option the genetic group codes are assumed to be coded from 1 to the number of groups. Again, when generating $\mathbf{H}^{-1}$ using WOMBAT, these codes can be transferred from the pedigree to the **`*.codes`** file.

References

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S., Lawlor T.J. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93 (2010) 743–752. doi: 10.3168/jds.2009-2730.

Meyer K. Technical note: A successive over-relaxation pre-conditioner to solve mixed model equations for genetic evaluation. J. Anim. Sci. 94 (2016) 4530–4535. doi: 10.2527/jas.2016-0665.

Misztal I., Legarra A., Aguilar I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92 (2009) 4648–4655. doi: 10.3168/jds.2009-2064.

Misztal I., Vitezica Z.G., Legarra A., Aguilar I., Swan A.A. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130 (2013) 252–258. doi: 10.1111/jbg.12025.

Westell R.A., Quaas R.L., Van Vleck L.D. Genetic groups in an animal model. J. Dairy Sci. 71 (1988) 1310–1318. doi: 10.3168/jds.S0022-0302(88)79688-5.