Pooling estimates of covariance components by penalized maximum likelihood using WOMBAT

1 Introduction

Combining estimates of covariance components from various sources or multiple analyses of subsets of traits is a common task in quantitative genetics. Specific requirements associated with this task are that the resulting overall matrices should

- be positive (semi-) definite, i.e. not have negative eigenvalues,
- have elements that are 'similar' to those from individual analyses,
- not change estimates of variance ratios substantially, and
- result in estimates of phenotypic covariances components that are little distorted.

Most methods in use combine estimates of covariances for a single source of variation at a time. Recent work has shown that this can result in substantially increased loss in estimates of the corresponding phenotypic covariance matrices, i.e. estimates that are on average further from the population values than unmodified values (Meyer, 2012). This can be attributed to the fact that pooling individual matrices does not account for sampling correlations between estimates for different sources of variation (from the same part analysis) and thus does not allow for repartitioning which would keep estimates of the total variation are more or less unchanged.

'Better' pooled matrices can be obtained by considering all matrices simultaneously. An early suggestion in this category, termed 'bending' (Hayes and Hill, 1981), has been to regress the canonical eigenvalues of the genetic and phenotypic covariance matrices towards their mean. A more flexible alternative is a likelihood based approach which implicitly treats estimates from individual analyses as if they were matrices of corrected sums of squares and cross-products due to some pseudo observations.

2 Background: The likelihood approach

Consider a vector of observations $\mathbf{y}$ for q traits from a multivariate normal distribution, $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{V})$, with $\mathbf{V}$ the covariance matrix of $\mathbf{y}$, $\mathbf{b}$ a vector of fixed effects and $\mathbf{X}$ the corresponding design matrix. If $\mathbf{y}$ can be split into independent parts, $\mathbf{y}_i$, for instance observations for independent families, $\mathbf{V}$ is block-diagonal. The corresponding log likelihood ($\log \mathcal{L}$) can then be calculated by summing over groups (Thompson, 1976)

$$-2 \log \mathcal{L} \propto \sum_i d_i \left(\log |\mathbf{V}_i| + \text{tr}\left(\mathbf{V}_i^{-1}\mathbf{M}_i\right)\right) \tag{1}$$

with $\mathbf{V}_i$ the $i-$th diagonal block of $\mathbf{V}$, $\mathbf{M}_i = \left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right)\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right)' / d_i$ the corresponding matrix of corrected mean crossproducts, and d_i denoting the degrees of freedom.

Let $\mathbf{\Sigma}_x$ denote the overall covariance for random effect x for all q traits. For example, for a simple animal model, $x = G, E$ with G and E the genetic and residual covariance matrices, respectively. Further, let $\mathbf{S}_i^x$ be the estimated covariance for random effect x from the $i-$th part analysis. $\mathbf{S}_i^x$ has expectation $\mathbf{\Sigma}_i^x$, the submatrix of $\mathbf{\Sigma}_x$, comprised of the covariance components among the traits represented in the $i-$th subset of traits. Considering a single source of variation only, estimates of $\mathbf{\Sigma}_x$ could then be obtained by replacing $\mathbf{M}_i$ with $\mathbf{S}_i^x$ and $\mathbf{V}_i$ with $\mathbf{\Sigma}_i^x$ in (Eq. 1) above, and maximising the resulting log likelihood whilst constraining $\hat{\mathbf{\Sigma}}_x$ to be positive definite.

2.1 A pseudo pedigree structure

For multiple sources of variation, strong sampling correlations between estimates can be approximated by constructing corresponding terms assuming a simple, balanced pseudo pedigree structure. Let this pseudo structure involves families with m members each, and define matrices $\mathbf{C}^x$ (of size $m \times m$) which give the coefficients for the $x-$th covariance component in the expectation of covariances between family members. This gives

$$-2 \log \mathcal{L} \propto \sum_i d_i \left(\log \left| \sum_x \mathbf{C}^x \otimes \mathbf{\Sigma}_i^x \right| + \text{tr} \left(\left[\sum_x \mathbf{C}^x \otimes \mathbf{\Sigma}_i^x \right]^{-1} \left[\sum_x \mathbf{C}^x \otimes \mathbf{S}_i^x \right] \right) \right) \tag{2}$$

with $\otimes$ denoting the direct matrix product. Maximizing (Eq. 2), subject to appropriate constraints, then yields pooled estimates of covariance matrices for all sources of variation.

Suitable choices for the pseudo pedigree depend on the covariance matrices to be pooled – the structure chosen should comprise a sufficient number of types of covariances between relatives so that, when equating these to their expectations, all components can be separated. For a simple animal model, a paternal half-sib design, with m paternal half-sibs per family, suffices. Coefficient matrices for this case are $\mathbf{C}^E = \mathbf{I}_m$ and $\mathbf{C}^G = \frac{1}{4}\mathbf{J}_m + \frac{3}{4}\mathbf{I}_m$ (with $\mathbf{J}_m$ a matrix of size $m \times m$ with all elements equal to unity). If there are common environmental covariances between full-sibs in addition ($x = G, C, E$ with $\mathbf{\Sigma}_C$ the common environmental covariance matrix), a hierarchical full-sib design with n offspring per dam and $k = m/n$ dams per sire would be represented by coefficient matrices $\mathbf{C}^E = \mathbf{I}_m$, $\mathbf{C}^C = \mathbf{I}_k \otimes \mathbf{J}_n$ and $\mathbf{C}^G = \frac{1}{4}\mathbf{J}_m + \mathbf{I}_k \otimes \left(\frac{1}{4}\mathbf{J}_n + \frac{1}{2}\mathbf{I}_n\right)$. For analyses with maternal genetic effects, the pseudo family should include data on at least two generations. e.g. a sire mated to two unrelated dams with two offspring per dam with records on all individuals would provides sire- and dam-offspring as well as full- and half-sib covariances.

2.2 Penalizing the likelihood

In addition, the likelihood approach provides opportunity for regularized estimation, analogous to that used in standard, multivariate analyses. This involves placing a penalty on the likelihood aimed at reducing sampling variation and thus to improve estimates. In particular, penalties designed to 'borrow strength' from estimates of the phenotypic covariance matrix by shrinking estimates of covariance matrices for individual random effects towards their sum or by reducing the spread of estimated canonical eigenvalues, have been shown to yield substantial reductions in sampling variation (Meyer and Kirkpatrick, 2010; Meyer, 2011). Corresponding penalties can by employed when pooling estimates from part analyses by maximizing

$$\log \mathcal{L}_P = \log \mathcal{L} - \tfrac{1}{2}\psi\mathcal{P} \tag{3}$$

instead of $\log \mathcal{L}$, with $\mathcal{P}$ denoting the penalty ($\mathcal{P} > 0$) and ψ a so-called tuning factor ($\psi \geq 0$), determining the emphasis to be given to $\mathcal{P}$.

3 Pooling using WOMBAT

A facility for pooling estimates of covariance components using the approach described has been implemented in WOMBAT. The main features are:

- Pooling is invoked using the *run time* option `--pool`.

- The *parameter file* should contain a block of statements, framed by POOL and END, which specify
  - the pseudo-pedigree structure to be used,
  - the type of penalty to be applied and the tuning factor(s) to be used (if any),
  - the minimum eigenvalues of the pooled covariance matrix,
  - the convergence criterion,
  - the form of input files, and
  - whether a 'minimum' or 'full' parameter file is supplied.
- Input consists of estimates of covariance components from analyses of partial, overlapping subsets of traits:
  - either individual files as generated by WOMBAT or a single, user generated file are accepted,
  - different part analyses can be assigned different weights to be given.
- The implementation allows for multiple random effects, assuming all pooled matrices have the same dimension. Estimates of residual covariances components with a value (or average) of zero are treated as covariances between traits measured on distinct subsets of animals and fixed at that value.

3.1 Pseudo pedigree structures

WOMBAT has three 'built-in' pedigree structures and, in addition, allows for other chosen structures by supplying the matrices of coefficients in the expectations of covariances between relatives (matrices $\mathbf{C}^x$ above).

The pseudo pedigree structure is given by a line (within the POOL block) starting with the directive PSEUPED, followed by one of the following options (space separated) and, depending on the option, additional numbers.

USR followed (space separated) by an integer value n and, optionally, by a value f selects a user-defined structure consisting of f families of size n. If not given, a value of $f = 2$ is used. This should be followed by the *upper* triangle of the $n \times n$ matrix of coefficients ($\mathbf{C}^x$) for all random effects fitted (i.e. excluding residual covariances). For multiple random effects, these should be in the *same* order as the corresponding VAR statements in the parameter file, and each matrix $\mathbf{C}$ should start on a new line.

Example

```
VAR   animal   4  NOSTART
VAR   residual 4  NOSTART
POOL
   PSEUPED  USR  5
   1.0  0.50  0.50  0.50  0.50
   1.0  0.25  0.25  0.25
   1.0  0.25  0.25
   1.0  0.25
   1.0
END
```

This shows the coefficients for direct additive genetic effects ('animal') for a family comprising a sire with four progeny from unrelated dams.

BON selects a design comprising f families of size $n = 8$ due to Bondari et al. (1978). Again, this can be followed by an optional number of families f (default $f = 2$). For this design, expectations of covariances between relatives due to direct and maternal effects are available. For WOMBAT to 'know' which random effect has which structure, additional information is required. This should comprise one additional

line per random effect fitted, with each line consisting of a keyword specifying the type of random effect followed (space separated) by the name of the effect as specified in the VAR statements. Keywords recognized are:
DIRADD for direct additive genetic,
MATADD for maternal genetic, and
MATPE for maternal permanent environmental effects.

Example

```
VAR animal   6  NOS
VAR gmdam    6  NOS
VAR pedam    6  NOS
VAR residual 6  NOS
POOL
   PSEUPED    BON  10
      DIRADD  animal
      MATADD  gmdam
      MATPE   pedam
END
```

HFS selects a balanced hierarchical full-sib design comprised of s sires, d dams per sire and n progeny per dam. Assumed values for s, d and n can be given (space-separated) on the same line. If omitted, default values of $s = 10$, $d = 5$ and $n = 4$ are used. This design is suitable for a simple animal model or a model fitting maternal permanent environmental effects in addition. If the MINPAR option is used, codes DIRADD and MATPE need to be given in addition as described above. If not, WOMBAT uses the covariance options given in the MODEL block to identify random effects.

PHS selects a balanced, paternal half-sib design comprised of s sires with n progeny each, given by respective integer numbers. If not given, default values of 10 and 4 are used. This design is suitable for a simple animal model only. Again, if the MINPAR option is given, this must be followed by a line with the DIRADD directive.

Example

```
VAR   animal   4  NOSTART
VAR   residual 4  NOSTART
POOL
   MINPAR
   PSEUPED  PHS  100 5
     DIRADD  animal
END
```

3.2 Penalties

WOMBAT first pools estimates of covariance matrices without a penalty on the likelihood. Additional analyses, subject to a penalty are invoked by a line starting with PENALTY followed (space separated) by a code word(s) defining the type and strength of penalty to be applied. Codes for the penalty type recognised are:

CANEIG selects a penalty on the canonical eigenvalues. This has to be followed (space separated) by either ORG or LOG specifying a penalty on eigenvalues on the original scale or transformed to logarithmic scale (no default).

COVARM specifies shrinkage of each covariance matrix towards a given target.

CORREL chooses shrinkage of each correlation matrix towards a given target correlation matrix.

Either COVARM or CORREL can be followed (space separated) by the keyword MAKETAR. If this is given, WOMBAT determines the shrinkage target as the phenotypic covariance (or correlation) matrix obtained by summing estimates of covariances for all sources of variation from the preceding, unpenalized analysis. If this is not given, the upper triangle of the $q \times q$ target matrix is expected to be read from a file with the standard name PenTargetMatrix.

The last entry on the line relates to the tuning factor(s) to be used:

- If a single penalized analysis is to be carried out, the corresponding tuning factor should be given (real value).
- To specify multiple penalized analyses, specify the number of separate tuning factors multiplied by -1 as an integer value (e.g. -3 for three analyses), and list the corresponding tuning factors space separated on the next line.

Example

1. Shrink all correlation matrices towards the phenotypic correlation matrix using a single tuning factor of 0.1; calculate the shrinkage target from unpenalized results.

```
PENALTY CORREL  MAKETAR  0.1
```

2. Shrink canonical eigenvalues on the logarithm scale towards their mean, using 5 different tuning factors

```
PENALTY CANEIG LOG  -5
0.01  0.1  0.5  1.0  2.0
```

3.3 Other options

WOMBAT recognizes the following other options within the POOL block:

MINPAR The default assumption for the parameter file used is that it is a ‘full’ parameter file as for a corresponding, multivariate analysis. However, the information on pedigree and data file and their layout is not used. Hence the option MINPAR (on a line by itself) is provided which allows use of a ‘minimum’ parameter file, switching off some of the consistency checks otherwise carried out. If used, MINPAR must be the *first* entry within the POOL block.

The minimum information to be given in the parameter file must comprise:

1. The ANALysis statement
2. A VAR line for each covariance matrix, together with the NOSTART option telling WOMBAT not to expect a matrix of starting values.
3. The POOL block, including statements showing which random effect represents which type of genetic or non-genetic effect.

Example

```
ANAL MUV 14
VAR  animal    14  NOSTART
VAR  residual  14  NOSTART
POOL
   MINPAR
   SMALL  0.001d0
   PSEUPED  hfs  100  10  4
     DIRADD  animal
END
```

`SINGLE` The default form of input for results from part analysis is to read estimates from separate files, in the form of output generated by WOMBAT when carrying out multiple analyses of subsets of traits. Alternatively, all information can be given in a single file. This is selected by the option `SINGLE`, followed (space separated) by the name of the input file (same line).

`SMALL` followed (space separated) by a real number which gives the lower limit (≤ 1) for smallest eigenvalue allowed in the combined matrices. If this is not specified, a default value of 0.0001 is used.

`DELTAL` followed (space separated) by a real number specifying the convergence criterion used in the iterative pooling scheme: if the increase in log likelihood between iterates falls below this value, convergence is assumed to have been achieved. If not specified a stringent default of 0.00005 is used.

3.4 Input: Results from partial analyses

WOMBAT allows for two forms of input for the 'partial' results to be pooled.

3.4.1 Individual files

The first option integrates with the facility in WOMBAT to carry out analyses considering selected traits in a multivariate analysis only. This is invoked by the notation `"k"->m` when specifying the trait numbers in the `MODEL` block. This stipulates that trait number k in the data file (ranging from 1 to q) should be replaced with value m for the analysis. If this is encountered, any records with trait numbers not selected in this fashion are ignored – using the `--subset` run time option to generate parameter files for analyses of subsets uses this facility. When this notation is encountered, WOMBAT

- writes out the estimates of covariance components together with information on the trait (re)numbering to a file fit the standard name `EstimSubset`k`+`. . . `+`l`.dat` with k to l the original trait numbers (in the data file, ranging from 1 to q) for the subset of traits considered in the current analysis.
- appends the name `EstimSubset`k`+`. . . `+`l`.dat` to a file named `SubSetsList` followed by a value of 1.00. The latter is a placeholder for the weight to be given to the results in this file, which can be replaced as required.

When encountering the `--pool` option without the `SINGLE` qualifier in the `POOL` block, WOMBAT tries to acquire the names of the input files and weights for individual analyses from `SubSetsList`. The individual input files listed then must provide the following information:

(a) The number of traits in the subset (first line).
(b) The corresponding (original) trait numbers in the 'full' analysis (second line).

This is followed by the covariance matrices estimated. The first matrix given must be the matrix of residual covariances, the other covariance matrices should be given in the same order as specified in the parameter file.

(c) The first line for each covariance matrix should list the running number of the random effect (0 for 'residual'), the order of fit and the name of the effect (corresponding to a name in the `VAR` statement).
(d) The following lines should give the elements of *full-stored* covariance matrix, with each row of the matrix starting on a new line.

3.4.2 Single file

The second option assumes all partial estimates have been collected in a single file. This is specified by the `SINGLE` option described above.

For each part analysis, this file should contain the following information:

1. A line giving (space separated):
   a) The number of traits, q_i, in the part analysis ($1 \leq q_i \leq q$, with q the total number of traits).
   b) The (running) numbers of these traits in the full covariance matrix.
   c) The relative weight to be given to this part; this can be omitted and, if not given, is set to 1.
2. The elements of the upper triangle of the *residual* covariance matrix, given row-wise, i.e. $q_i(q_i + 1)/2$ elements).
3. For each random effect fitted, the elements of the upper triangle, given row-wise ($q_i(q_i + 1)/2$ elements). Each matrix must begin on a new line and the matrices must given in the same order as the corresponding `VAR` statements in the parameter file.

3.5 Output

Output form a run with option `--pool` consists of the following files:

1. `PoolEstimates.out` is the main output which summarizes characteristics of the part estimates provided, options chosen, and results for all analyses carried out.
2. `PoolBestPoint` is the equivalent to `BestPoint`, suitable for further examination or as starting values for full, multivariate analyses. The first line gives the likelihood, the number of parameters and tuning parameter used. Then the elements of the upper triangle of the pooled covariance matrices are given, starting with the residual covariance matrix and the other covariance matrices in order of the `VAR` statements in the parameter file. Each matrix begins on a new line.
   If penalized analyses are carried out, copies labelled `PoolBestPoint_unpen` and `PoolBestPoint_t`*xx*, with *xx* equal to the tuning factor, are generated so that files for all sub-analyses are available at the end of the run.

4 Example

WOMBAT example 4 gives data for four traits measured by ultra-sound scanning of beef heifers. The model of analysis is a simple animal model. Results from the 6 bivariate analyses of all possible pairs of traits are summarized in the single file:

PartAll.dat

```
2    1    2
3.53021846755401         1.95048936242497          2.08907657452187
2.36692690581371         1.34229728294400          0.977425661859041
2    1    3
3.55743271408901         2.76836382462327          23.6200689402088
2.32842498547499        0.993394132647659          7.46919029407730
2    1    4
3.50629783053394         10.6739641890506          231.756415050249
2.38789877391301         13.6673590924466          148.163889185864
2    2    3
2.12383341617844         2.09610622027606          23.7246862784793
0.929943072656729        0.401190656024861         7.33017861737474
2    2    4
2.10152459348052         9.94514189035131          240.500803642650
0.952091184635393        9.39249109396949          140.942518952472
2   3    4
23.7104763915071         21.0346940586355          243.221560720708
7.34076531794134        -3.01793177595214          144.576874008222
```

The corresponding 'minimum' parameter file for a run including penalized estimation for a single tuning factor is:

pool_min.par

```
RUNOP   --pool                                   1
ANAL MUV     4                                   2
VAR animal   4   NOS                             3
VAR residual 4   NOS                             4
POOL                                             5
   MINPAR                                        6
   SINGLE   PartAll.dat                          7
   PSEUPED  BON 100                              8
   DIRADD   animal                               9
   PENALTY  CORREL MAKETAR  2.5                  10
END                                              11
```

Line 1 gives the `--pool` option, alternatively this could be specified on the command line. The first line in the `POOL` block (line 6) contains the `MINPAR` option, telling the program to expect minimal input, and line 7 directs the input to be read from the file `PartAll.dat`. Line 8 selects Bondari's design as pseudo pedigree structure with 100 families. With the `MINPAR` option, line 9 is required to instruct the program that the single random effect fitted is a direct, additive genetic effect. Finally, line 10 selects penalized pooling with shrinkage of the genetic towards the phenotypic covariance matrix using a tuning factor of 2.5, and line 11 closes the block.

The first part of `PoolEstimates.out` summarizes the numeric options, statistics on the input values, and gives the simple 'average' covariance matrices constructed form the part estimates together with their eigenvalues. With all possible, pairwise analyses between the 4 traits, there are 3 estimates for each variance but only one estimate for each covariance component.

PoolEstimates.dat

```
======= Version 24-04-2012 ======================================= **KM** ====

        Program WOMBAT : Pooled estimates of covariance components
==============================================================================

 Value for "small"      =   0.00010000
 Convergence criterion  =   0.00005000

 No. of traits          =      4
                               1                trait1
                               2                trait2
                               3                trait3
                               4                trait4

 No. of part analyses   =      6
                               1   2->  1  2   "PartAll.dat"
                               2   2->  1  3   "PartAll.dat"
                               3   2->  1  4   "PartAll.dat"
                               4   2->  2  3   "PartAll.dat"
                               5   2->  2  4   "PartAll.dat"
                               6   2->  3  4   "PartAll.dat"

***** Means & ranges for residual covariances   ******************************
   1  COVS Z 1 1        3    3.53132        3.50630        3.55743
   2  COVS Z 1 2        1    1.95049
   3  COVS Z 1 3        1    2.76836
   4  COVS Z 1 4        1    10.6740
   5  COVS Z 2 2        3    2.10481        2.08908        2.12383
   6  COVS Z 2 3        1    2.09611
   7  COVS Z 2 4        1    9.94514
   8  COVS Z 3 3        3    23.6851        23.6201        23.7247
   9  COVS Z 3 4        1    21.0347
  10  COVS Z 4 4        3    238.493        231.756        243.222
 Eigenvalues of averaged covariance matrix
   241.466       21.8808       3.74837      0.719141

***** Means & ranges for RE  1   "animal"   **********************************
  11  COVS A 1 1        3    2.36108        2.32842        2.38790
  12  COVS A 1 2        1    1.34230
  13  COVS A 1 3        1   0.993394
  14  COVS A 1 4        1    13.6674
  15  COVS A 2 2        3   0.953153       0.929943       0.977426
  16  COVS A 2 3        1   0.401191
  17  COVS A 2 4         1    9.39249
  18  COVS A 3 3         3    7.38004        7.33018        7.46919
  19  COVS A 3 4         1   -3.01793
  20  COVS A 4 4         3    144.561        140.943        148.164
 Eigenvalues of averaged covariance matrix
   146.540       7.63089      0.960639      0.124311
```

The second part gives details on the assumed pseudo pedigree and the results from unpenalized pooling. As the average matrices were within the parameter space and all analyses were weighted equally, not surprisingly, results differ little from the former.

PoolEstimates.dat continued

```
===== Pseudo pedigree structure     ==========================================
 Bondari's design
 No. of families        =   100
 No. individuals/fam.   =     8

      Coefficients for RE  1   "animal"
   1   1.0000
   2   0.5000  1.0000
   3   0.2500  0.2500  1.0000
   4   0.2500  0.2500  0.5000  1.0000
   5   0.5000  0.2500  0.1250  0.1250  1.0000
   6   0.5000  0.2500  0.1250  0.1250  0.5000  1.0000
   7   0.1250  0.1250  0.2500  0.5000  0.0625  0.0625  1.0000
   8   0.1250  0.1250  0.2500  0.5000  0.0625  0.0625  0.5000  1.0000
===============================================================================

 No. of parameters      =     20
***** Estimates of residual covariances   ************************************
 Covariance matrix
   1     3.5450
   2     1.9647       2.1023
   3     2.7673       2.0763       23.672
   4     11.131       9.8742       20.697       238.69
 Eigenvalues of covariance matrix
 Value       241.63        21.93         3.73         0.71
   (%)        90.16         8.18         1.39         0.27
 Trace       268.01
 Matrix of correlations and variance ratios
   1      0.6017
   2      0.7197    0.6870
   3      0.3021    0.2943    0.7621
   4      0.3827    0.4408    0.2753    0.6233

***** Estimates for RE  1   "animal"   ***************************************
      Covariance structure =    NRM
 Covariance matrix
   1     2.3469
   2     1.3219      0.95797
   3    0.99423      0.42503       7.3899
   4     13.338       9.5565      -2.9140       144.28
 Eigenvalues of covariance matrix
 Value       146.22         7.64         0.98         0.13
   (%)        94.35         4.93         0.64         0.08
 Trace       154.98
 Matrix of correlations and variance ratios
   1      0.3983
   2      0.8816    0.3130
   3      0.2387    0.1597    0.2379
   4      0.7248    0.8129   -0.0892    0.3767

***** Estimates of phenotypic covariances  ***********************************
 Covariance matrix
   1     5.8919
   2     3.2865       3.0602
   3     3.7615       2.5013       31.062
   4     24.469       19.431       17.783       382.97
 Eigenvalues of covariance matrix
 Value       386.48        30.51         5.13         0.87
   (%)        91.37         7.21         1.21         0.20
 Trace       422.98
 Correlation matrix
   1      1.0000
   2      0.7740    1.0000
   3      0.2780    0.2566    1.0000
   4      0.5151    0.5676    0.1630    1.0000
===============================================================================
```

Next, corresponding results when imposing a penalty are listed. For each covariance matrix, the change from the unpenalized, pooled estimates is summarized as the Frobenius norm (F-norm) of the matrix difference. If there were multiple tuning factors, such section would be given for each.

PoolEstimates.dat continued

```
*****  Pooling with penalties           ***********************************
*****  Type of penalty: "CORREL"        ***********************************

==========================================================================
===== Tuning factor =     2.50000       ==================================

***** Estimates of residual covariances   ***********************************
      F-norm diff(unpenal) =  1.6007
 Covariance matrix
   1     3.5590
   2     1.9826       2.1071
   3     2.8285       2.1021       23.670
   4     11.296       10.175       19.628       238.87
 Eigenvalues of covariance matrix
 Value       241.66        22.15         3.70         0.70
   (%)        90.10         8.26         1.38         0.26
 Trace       268.21
 Matrix of correlations and variance ratios
   1      0.6046
   2      0.7240    0.6891
   3      0.3082    0.2977    0.7622
   4      0.3874    0.4535    0.2610    0.6242

***** Estimates for RE  1   "animal"   ***************************************
      Covariance structure =    NRM
      F-norm diff(unpenal) =  2.1961
 Covariance matrix
   1     2.3273
   2     1.2979      0.95079
   3    0.90708      0.38810       7.3847
   4     13.129       9.1775      -1.4612       143.82
 Eigenvalues of covariance matrix
 Value       145.64         7.58         1.12         0.15
   (%)        94.27         4.91         0.73         0.10
 Trace       154.49
 Matrix of correlations and variance ratios
   1      0.3954
   2      0.8725    0.3109
   3      0.2188    0.1465    0.2378
   4      0.7176    0.7848   -0.0448    0.3758

***** Estimates of phenotypic covariances  **********************************
      F-norm diff(unpenal) = 0.62278
 Covariance matrix
   1     5.8863
   2     3.2805       3.0579
   3     3.7356       2.4902       31.055
   4     24.426       19.352       18.167       382.70
 Eigenvalues of covariance matrix
 Value       386.23        30.45         5.14         0.87
   (%)        91.37         7.20         1.22         0.21
 Trace       422.70
 Correlation matrix
   1      1.0000
   2      0.7732    1.0000
   3      0.2763    0.2555    1.0000
   4      0.5146    0.5657    0.1666    1.0000
```

Finally, for ease of comparison, results are listed side by side. For the combination of number of families, type of penalty and tuning factor, penalization did not affect estimates substantially, but a slight reduction in the spread of the eigenvalues of the genetic covariance is notable.

PoolEstimates.dat continued

```
===== Estimates side by side  ===============================================

Tuning factor        0.00000      2.50000
Froben. Norm-Phen    0.00000     0.622778
Sum(Froben. Norm)    0.00000      3.79674

***** Estimates of residual covariances   ***********************************
  1  1  COVS Z 1 1   3.54499      3.55898
  1  2  COVS Z 1 2   1.96468      1.98260
  1  3  COVS Z 1 3   2.76727      2.82849
  1  4  COVS Z 1 4   11.1314      11.2964
  2  2  COVS Z 2 2   2.10228      2.10706
  2  3  COVS Z 2 3   2.07627      2.10211
  2  4  COVS Z 2 4   9.87423      10.1748
  3  3  COVS Z 3 3   23.6719      23.6703
  3  4  COVS Z 3 4   20.6967      19.6281
  4  4  COVS Z 4 4   238.690      238.873

***** Estimates for RE  1   "animal"      ************************************
  1  1  COVS A 1 1   2.34691      2.32734
  1  2  COVS A 1 2   1.32185      1.29794
  1  3  COVS A 1 3  0.994227     0.907077
  1  4  COVS A 1 4   13.3376      13.1293
  2  2  COVS A 2 2  0.957967     0.950786
  2  3  COVS A 2 3  0.425032     0.388103
  2  4  COVS A 2 4   9.55651      9.17746
  3  3  COVS A 3 3   7.38991      7.38466
  3  4  COVS A 3 4  -2.91396     -1.46121
  4  4  COVS A 4 4   144.280      143.824

***** Estimates of phenotypic covariances   **********************************
  1  1  COVS T 1 1   5.89190      5.88632
  1  2  COVS T 1 2   3.28653      3.28055
  1  3  COVS T 1 3   3.76150      3.73556
  1  4  COVS T 1 4   24.4690      24.4256
  2  2  COVS T 2 2   3.06025      3.05785
  2  3  COVS T 2 3   2.50130      2.49022
  2  4  COVS T 2 4   19.4307      19.3523
  3  3  COVS T 3 3   31.0618      31.0550
  3  4  COVS T 3 4   17.7828      18.1668
  4  4  COVS T 4 4   382.971      382.697

***** Variance ratios & correlations for RE  1   "animal"  *******************
  1  1  V.ratio      0.398        0.395
  1  2  Correl.      0.882        0.873
  1  3  Correl.      0.239        0.219
  1  4  Correl.      0.725        0.718
  2  2  V.ratio      0.313        0.311
  2  3  Correl.      0.160        0.146
  2  4  Correl.      0.813        0.785
  3  3  V.ratio      0.238        0.238
  3  4  Correl.     -0.089       -0.045
  4  4  V.ratio      0.377        0.376

***** Eigenvalues for Residual      ******************************************
  1     E.value      241.633      241.657
  2     E.value      21.9283      22.1477
  3     E.value      3.73397      3.70123
  4     E.value     0.714643     0.702685

***** Eigenvalues for RE  1   "animal"    ************************************
  1     E.value      146.218      145.638
  2     E.value      7.64321      7.57981
  3     E.value     0.984761      1.12056
  4     E.value     0.129158     0.148989
======== end of file ===========================26-04-2012=========15:31====
```

5 Frequently asked questions

- **Q:** I have a collection of estimates from various sources & data sets. Can I use WOMBAT to combine these?

  A: *Yes; the* `SINGLE` *input file and* `MINPAR` *options are provided to make 'general' use easier.*

- **Q:** My estimates are based on data sets of very different sizes. Can I do a weighted analysis?

  A: *Yes; each part analysis can be assigned a different weight. The degrees of freedom used in (Eq. 2) are then equal to the product of the weight, the number of families and the number of traits in the subset.*

- **Q:** I have existing covariance matrices which I would like to modify so that the smallest eigenvalues are a bit larger and, hopefully, my genetic evaluation runs converge faster. Can I do that with the `--pool` option?

  A: *Yes;* WOMBAT *can handle a single 'part' analysis with the number of traits in the 'subset' equal to the total number of traits. Beware though that setting the limit on the smallest eigenvalue to something quite different from zero can change matrices substantially - there is thus an upper limit of 1 on the value which can be set via the* `SMALL` *option. When attempting to regularize matrices, it might be best to combine a* *smaller limit on the minimum eigenvalues with a penalty.*

- **Q:** Can I get standard errors for the pooled estimates?
  A: *No; there is no theoretical basis to derive these.*

- **Q:** What value of tuning factor should I use?
  A: *The tuning factor specifies the emphasis given to the penalty relative to the 'data'. Hence, the larger the assumed degrees of freedom (d_i in (Eq. 2) - which is proportional to the number of families in the pseudo pedigree structure, the number of traits and the weight given to a particular set of estimates), the less stringent the penalization for a given tuning factor. The general recommendation is to apply a 'mild' penalty – this may require trying a number of different values.*

- **Q:**
  A:

- **Q:** Can I use `--pool` to modify a single matrix – and what do I need to watch out for?
  A: *The main value of likelihood based pooling lies in the opportunity to allow for repartitioning of the total variance into its components. However, yes this procedure can be applied to a single matrix. The parameter file should be set up calling this a residual covariance matrix –* WOMBAT *always expects to find one of these, but not necessarily any random effects and their covariance matrices. No pseudo pedigree structure is used in this case, and the parameter file thus should not include a* `PSEUPED` *statement. Similarly, the option* `MAKETAR` *for penalized estimation is not allowed – if penalized pooling with shrinkage towards a target matrix is required, this target needs to be specified explicitly. As canonical eigenvalues for a single matrix are not defined, the eigenvalues of the matrix are substituted for these if the option* `CANEIG` *is encountered.*

References

Bondari K., Willham R.L., Freeman A.E. Estimates of direct and maternal genetic correlations for pupa weight and family size of Tribolium. J. Anim. Sci. 47 (1978) 358–365.

Hayes J.F., Hill W.G. Modifications of estimates of parameters in the construction of genetic selection indices ('bending'). Biometrics 37 (1981) 483–493.

Meyer K. Performance of penalized maximum likelihood in estimation of genetic covariances matrices. Genet. Sel. Evol. 43 (2011) 39. doi: 10.1186/1297-9686-43-39.

Meyer K. A penalized likelihood approach to pooling estimates of covariance components from analyses by parts. J. Anim. Breed. Genet. 00 (2012) 000–000 (submitted 21/3/2012).

Meyer K., Kirkpatrick M. Better estimates of genetic covariance matrices by 'bending' using penalized maximum likelihood. Genetics 185 (2010) 1097–1110. doi: 10.1534/genetics.109.113381.

Nelder J.A., Mead R. A simplex method for function minimization. Computer J. 7 (1965) 308–313.

Powell M.J.D. An efficient method for finding the minimum of a function of several variables without calculating derivatives. Computer J. 7 (1965) 155–162.

Thompson R. The estimation of maternal genetic variance. Biometrics 32 (1976) 903–917.

A Technical details

A.1 Maximization of the likelihood

As the pseudo likelihood can be very flat, maximization is carried out very 'heavy-handed' using a Method of Scoring step followed by two derivative-free search steps, using methods of Powell (1965) and Nelder and Mead (1965), respectively. This is repeated until the change in log likelihood between rounds falls below a defined value (or if 50 rounds have been reached). For a stringent convergence criterion or a large number of traits, this can take quite some time.

A.2 Calculation of penalties

Penalties are summed over all sources of variation. For a penalty on the spread of canonical eigenvalues (`CANEIG`),

$$\mathcal{P} = \begin{cases} \sum_{x \neq E} \sum_{i=1}^{q} \left(\lambda_{xi} - \bar{\lambda}_x\right)^2 & \text{for ORG with} \quad \bar{\lambda}_x = \sum_i \lambda_{xi}/q \\ \sum_{x \neq E} \sum_{i=1}^{q} \left(\log(\lambda_{xi}) - \overline{\log(\lambda_x)}\right)^2 + \left(\log(1-\lambda_{xi}) - \overline{\log(1-\lambda_x)}\right)^2 & \text{for LOG with} \quad \overline{\log \lambda}_x = \sum_i \log(\lambda_{xi})/q \end{cases}$$

and λ_{xi} the i–th canonical eigenvalue of $\hat{\mathbf{\Sigma}}_P^{-1}\hat{\mathbf{\Sigma}}_x$ and $\hat{\mathbf{\Sigma}}_P = \sum_x \hat{\mathbf{\Sigma}}_x$ the phenotypic covariance matrix. For this option, summation over x to obtain $\mathcal{P}$ includes only the covariance components due to random effects fitted, i.e. excludes the residual covariance matrix.

Shrinking estimated covariance matrices towards a target matrix $\mathbf{T}$, the penalty is

$$\mathcal{P} = \sum_x \log\left|\hat{\mathbf{\Sigma}}_x\right| + \text{tr}\left(\hat{\mathbf{\Sigma}}_x^{-1}\mathbf{T}\right)$$

with summation over x including all estimated covariance matrices. Invoking option `MAKETAR` sets $\mathbf{T}$ to $\hat{\mathbf{\Sigma}}_P$ from the unpenalized analysis. Similarly, the penalty for option `CORREL` is

$$\mathcal{P} = \sum_x \log\left|\hat{\mathbf{R}}_x\right| + \text{tr}\left(\hat{\mathbf{R}}_x^{-1}\mathbf{T}\right)$$

with $\hat{\mathbf{R}}_x$ the correlation matrix corresponding to covariance matrix $\hat{\mathbf{\Sigma}}_x$. Specifying `MAKETAR` for this case sets $\mathbf{T} = \hat{\mathbf{R}}_P$.