GBLUP in Echidna

Notes prepared by Arthur Gilmour (Arthur.Gilmour@cargovale.com.au)

Introduction

ASReml became popular because it was efficient handling large sparse mixed models and accommodated a large variety in the models that could be fitted. A large component was the use of an additive genetic relationship matrix which is SPARSE if the parents are included (up to 4 non-zero cells per animal). However, quantitative geneticists have increasingly wanted to use genomic relationship matrices based on marker (SNP) panels. These are DENSE, typically calculated as **MM**'/*s* where **M** is the marker matrix (with values 0/1/2 but usually centred) and *s* is a scaling parameter based on average SNP value. Often **M** has many more columns (SNPs) than rows (genotypes) but may not.

This document discusses several options for fitting mixed models involving GRM matrices, loosely call GBLUP models. It was prompted by an example from Li Li of AGBU (see Large example below). Since my stated reason for creating Echidna was to improve on the performance of ASReml for GBLUP models, he provided this example which runs in GCTA (Yang, University of Queensland) in 6:29 m:s but would not run in Echidna.

The first few sections describe things previously done. Then we report on their use in a moderate example.

The ultimate goal is bivariate analysis of a genomic model involving several G matrices (on the same animals).

The performance of Echidna (and ASReml where relevant) has improved since these runs were performed.

Executive summary (as at 15 July)

The following summary pertains to Echidna 1.20

1. Echidna can now read in and invert a GRM matrix of order 32000, but fails to fit the model (in 16Gbyte ram). The SMIDENSE routine now takes 9 minutes rather than 44 hours to invert this matrix.
2. The test example had data on only 10,000 of the animals. The !TRIM facility allows the GRM to be reduced to match the data. Echidna takes 4 min to invert that GRM and 5 min per iteration to fit the model and typically takes ~ 5 iterations.
3. Preferable to supply GRM as a XX.bgrm file (dense row-wise lower triangle real*4 binary).
4. !SAVEDENSE qualifier saves the inverse as a XX.bgiv file and writes the logdet as the second value on the first line.
5. If the GRM file is specified as say XX.bgiv but XX.bgiv does not exist, Echidna will check for in order for XX.sgiv, XX.dgiv, XX.bgrm, XX.grm. If the latter are found, the inverse will be formed and saved (honouring any qual;ifiers relating to the inversion, which are otherwise ignored) and the inverse save for use in the next run. (probably won't work with !TRIM).
6. Echidna has 3 routines to invert the GRM file. The default is the MKL inversion routine. If the !ND qualifier is set, my SMIDENSE routine is used which allows the GRM to be negative definite

and/or to have null rows. If !NSD or !PSD qualifiers are set, my SMISING routine is used. It seems SMIDENSE is faster than the MKL routine which is faster than SMISING. However, SMISING allows for singularities (linear dependencies) in the GRM matrix and creates an enlarged inverse to handle them (assuming that when data is added, the singularity is resolved). If there are NULL equations and SMIDENSE is used, the user must ensure there is no data information on the effects corresponding to NULL equations.

7. In 1 example (G_r.bgrm), there was a negative pivot and the model struggled to converge using that inverse. Using the !ADD qualifier made the matrix positive definite and the model then converged.
8. Echidna has the facility to fit severable conformable GRM matrices as a composite. See MRM section for details.
9. !LDET qualifier now works as follows when the inverse GRM is supplied. If a LogDet value is present in the .[b|s|d]giv file, it is used. Otherwise, if !LDET is specified with an argument, that value is used; if !LDET is specified without an argument, the LogDet is calculated from the inverse supplied; if !LDET is not specified, an approximate log det value is calculated from the average diagonal and offdiagonals. The LogDet value does not affect the estimation of variance parameters but is reflected in the reported LogL.
10. The !GDENSE qualifier does not reduce run times for these models.
11. A client has implemented an approach based on singular value decomposition of the G matrix. The approach is demonstrated on a 10K data set for bivariate analysis involving a single GRM.

MRM variance function (April 2019)

This extension was proposed by Ricardo Pong Wong (Robin Thompson) 15? years ago and prompted now by Jim Holland of NC State

mrmk(.) specifies the relationship matrix which is a sum of other relationship matrices. The matrices must be conformable. *k* selects the components. For example '12i' would indicate the sum of GRM1, GRM2 and an Identity, and so would fit 3 components. The test job fitted equivalent models:

```
!PART 6
Ablue !WT Ywt !DISP 1 ~ mu Env !r giv1(Hyb) giv2(Hyb) ide(Hyb) +
   idv(Env).giv1(Hyb) idv(Env).giv2(Hyb) idv(Env).ide(Hyb)
!PART 66
Ablue !WT Ywt !DISP 1 ~ mu Env !r mrm12i(Hyb)
                        id(Env).mrm12i(Hyb)
```

with 65 levels of Env and 1919 Hybrids; Part 6 takes 30m per iteration, part 66 takes 7m per iteration.

TenK example

```
!WORK 14  !REN  !ARG  42   !LOG !DEBUG      !OUT
TITLE: data  !DOPART $1
id    !A  !LL20  !L ped.csv !Lskip 1  # 26IN032007000607
cg_imf   *     # 4
cg_sf5   *     # 4
imf            # 3.8252
```

```
sf5    !10.    # 22.8268

!PART 12
      a22.bgiv     !ND
      data.csv !SKIP 1
      sf5 ~ mu  !r grm1(id) !f cg_sf5

!PART 42
   a22.bgrm       !MRM
   G_r.bgrm       !MRM
   G_t.bgrm       !MRM
   GG.bgrm        !MRM
   data.csv !SKIP 1
   sf5 ~ mu  !r mrm1234(id)  !f cg_sf5

!PART 0
residual units



Echidna 1.22 13 Jul 2020 Linux     13.4 Gbyte  at Mon Jul 13 10:28:49 2020
 Licensed to arthur(arthur.gilmour@cargovale.com.au)
 TITLE: data
 Folder: /run/media/arthur/DATADRIVE1/2020/Li/TenK

 id    !A  !LL20  !L ped.csv !LSKIP 1
 9688 class names for id initialized from ped.csv
 sf5

 GRM File: a22.bgrm assuming a dense LT matrix starting
               1.0000
               0.0007      1.0000
               0.0004      0.0071      1.0000
               0.0004      0.0071      0.5010      1.0000
 MinMnMax of diagonal      1.000     1.001     1.283 Average Cov     0.004
  Note: GRM inverse not formed because !MRM flag is set.

 GRM File: G_r.bgrm assuming a dense LT matrix starting
               0.9314
              -0.0162      1.0032
              -0.0218      0.0512      0.9959
              -0.0152      0.0544      0.5228      0.9986
 MinMnMax of diagonal      0.887     0.981     1.282 Average Cov    -0.000
  Note: GRM inverse not formed because !MRM flag is set.

 GRM File: G_t.bgrm assuming a dense LT matrix starting
               0.9551
              -0.0109      1.0891
              -0.0265      0.0672      1.0595
              -0.0097      0.0654      0.5544      0.9921
```

```
 MinMnMax of diagonal      0.826     1.001     1.468 Average Cov    -0.000
  Note: GRM inverse not formed because !MRM flag is set.

 GRM File: GG.bgrm assuming a dense LT matrix starting
               0.4642
               0.0359      0.3555
               0.0260      0.2462      0.1988
               0.0260      0.2462      0.1988      0.1988
 MinMnMax of diagonal      0.107     0.378     1.000 Average Cov     0.062
  Note: GRM inverse not formed because !MRM flag is set.

 Data File: data.csv

 Summary of 9688 data records

 Variable   Levels Miss Zero      Min      Max    Distribution or Mn SD Sk Kt
 id           9688    0    0        1     9688
 cg_imf        376    0    0        1      376
 cg_sf5        376    0    0        1      376
 imf             1    0    0  1.11600  9.90950  4.23745  0.98691   0.64   1.03
 sf5             1    0    0    10.89   166.26    34.88    15.22   1.72   5.01
 Note: Using !DOPART 42
 Note: Model is fitting 10065 equations, DENSE portion has 1 equations.
 * This job may use 8 processor threads. *
    1 LogL= -27038.08   86.83          9312 DF
    2 LogL= -27004.52   90.64          9312 DF
    3 LogL= -27000.98   83.66          9312 DF
    4 LogL= -26986.47   97.38          9312 DF
    5 LogL= -26986.42   97.92          9312 DF
    6 LogL= -26986.42   98.07          9312 DF

 Akaike Information Criterion   53982.84 (assuming 5 parameters).
 Bayesian Information Criterion 54018.54

          Analysis of sf5

                        Wald F statistics
Source of Variation           NumDF     DenDF     F-inc              P-inc

 Model_Term                    Order     Gamma         Sigma     Z_ratio  %C
 mrm1234(id)_V                  9688  0.951477E-01   9.33081        3.40   0 P
 mrm1234(id)_V                  9688  0.788341E-02  0.773098        0.48   0 P
 mrm1234(id)_V                  9688  0.470544E-06  0.461446E-04    0.00   0 B
 mrm1234(id)_V                  9688  0.392120E-06  0.384538E-04    0.00   0 B
 Residual_units                 9688   1.00000       98.0665       38.64
  cg_sf5                               376 effects fitted.
  mrm1234(id)                         9688 effects fitted.
 Finished: Mon Jul 13 11:11:35 2020LogL Converged
```

Timing for various steps in the analysis

```
>> >>   Echidna Process      CPU        SumCPU    WClock     SumWC
 >> >>           Setup A     0.59 sec      0.01      0.59      0.01
 >> >>           Startup   343.84 sec      5.74     44.04      0.74
 >> >>             SSP+G     1.01 sec      5.76      1.01      0.76
 >> >>         Get Order     0.65 sec      5.77      0.66      0.77
 FILLIN  46963708 >>  46963708      1.00 FOLD  DSA     2   10064  4666
 >> >>         Reorder C     4.35 sec      5.84      4.36      0.84
        3213  Bi/Tri node calls >>
 >> >>          Absorb C   509.48 sec     14.33     64.22      1.91
 >> >>          AIAbsorb     4.69 sec     14.41      4.34      1.99
 >> >>         Ci formed  1157.73 sec     33.71    146.47      4.43
 >> >>      Ci reordered     3.35 sec     33.76      2.25      4.47
 >> >>  E71_217 MRMscore  1211.56 sec     53.95    156.44      7.07
 >> >>  Completed_2       3275.05 sec    108.54    427.30     14.19
 >> >>  Completed_3       3261.66 sec    162.90    425.08     21.28
 >> >>  Completed_4       3256.62 sec    217.18    425.97     28.38
 >> >>  Completed_5       3342.75 sec    272.89    434.65     35.62
 >> >>  Completed_6       3281.13 sec    327.57    428.59     42.77
```

Note that there are 4 expensive steps. 'Startup' involves calculating the inverse of A22 + G_r + G_2 + GG and takes 44 seconds. The 'Absorb C' and 'Ci formed' steps are typically the slowest. But here, the most expensive is the score step tr($\mathbf{G}^{-1}$ *d*$\mathbf{G}$ $\mathbf{G}^{-1}$ $\mathbf{C}^{ZZ}$), in particular forming $\mathbf{G}^{-1}$ $\mathbf{C}^{ZZ}$ $\mathbf{G}^{-1}$ (a product of 3 dense matrices of order 10,000) for use in the trace operation for each variance component.

Using mrm() in a bivariate analysis.

These univariate mrm() analyses show a small component for A22, a larger component for G_r for trait imf and the reverse for trait sf5. Also, the variance of imf is much smaller than for sf5. A subsequent bivariate analysis shows a negative covariance between the traits.

An unstructured matrix can be partitioned into an average effect, and 2 trait specific components. We therefore propose using that idea with mrm(). It will probably work better if the scales are closer and the covariance is positive, so I multiply imf by -10.

In the first attempt, Echidna hung in the equation ordering. Using the !EQN 1 ordering generated an INFILL factor of 2.25 and took 75 min per iteration. Arranging the model terms as

XX		Contemporary groups
XC	Gc	Covariance G matrix
X1	I1 G1	Trait 1 Genetic covariance
X2	I2 0 G2	Trait 2 Genetic covariance

The G matrices are on the diagonal and so requires (for 10K example) 50M cells. Absorbing G2 fills in 50M cells in I2. Absorbing G1 fills in 50M cells in I1. X1 and X2 has infill of 2M each. Xc would have infill of at least 4M. So before absorption we use about 151M cells and we add a further 108 M which suggest a minimal infill factor of 1.72. The value from !EQN1 was 2.25. !EQN5 (model order) may do better but

in my first try, it was the same as !EQN1. On restarting, it gave a INFILL factor of 2.46 so that theory failed. !EQN 1 seems best.

I am running part 3 of Bmrm.es.

```
!WORK 28  !REN  !ARG  2   !LOG !DEBUG      !OUT
TITLE: data  !DOPART $1
 # id,cg_imf,cg_sf5,imf,sf5 ...
 # 26IN032007000607,4,4,3.8252,22.8268 ...
id    !A  !LL20  !L ped.csv !Lskip 1  # 26IN032007000607
cg_imf   *     # 4
cg_sf5   *     # 4
imf   !*10.     # 3.8252
sf5        # 22.8268
Ximf !=imf !*-1

   a22.bgrm     !ND  !MRM
   G_r.bgrm     !ND !ADD  !MRM
#   G_t.bgrm  !ND  !MRM
#   GG.bgrm !ND !ADD  !MRM
!PART 0// data.csv !SKIP 1

!PART 1  //  imf sf5 ~ Trait  !r   !f at(Trait,1).cg_imf !f at(Trait,2).cg_sf5

!PART 2  //  imf sf5 ~ Trait  !r at(Tr,1).mrm12(id) at(Tr,2).mrm12(id) +
                mrm12(id)  !f at(Tr,1).cg_imf at(Tr,2).cg_sf5
!PART 3:5  //  !EQN 1 # 4 failed, 6 slower for model 3
Ximf sf5 ~ Trait  !r  +
        !f at(Tr,1).cg_imf at(Tr,2).cg_sf5 +
!PART 3 // !r mrm12(id !GU) + at(Tr,1).mrm12(id !GU) at(Tr,2).mrm12(id !GU)
!PART 4 // !r mrm12(id !GU))
!PART 5 // !r                at(Tr,1).mrm12(id !GU) at(Tr,2).mrm12(id !GU)
```

The earlier univariate mrm runs showed no variance associated with G_t and GG so I have dropped these from this model.

I fitted Ximf = -10*imf with sf5 in model 3. The 10 was to put the traits on similar scales and the - was to make them positively correlated.

Now the model you want is traditional written `us(Trait).grm1(id) + us(Trait).grm2(id)` which defines 6 parameters (2 groups of 3). Call them `V111 C112 V122 V211 C212 V222.` This model (Part 3) also fits 6 parameters (3 groups of 2). Call them `C112 C212 S111 S211 S122 S222.` mrm12(id) is the 2 covariances (C112 and C212) but since covariances can be negative, should be specified as mrm12(id !GU)

`at(Tr,1).mrm12(id)` fits the specific variances `( S111 S211)` for trait 1
`at(Tr,2).mrm12(id)` fits the specific variances `( S122 S222)` for trait 2

The two formulations should be equivalent with
V111 = C112 + S111
V122 = C112 + S122
V211 = C212 + S211
V222 = C212 + S222

```
    9 LogL= -51688.48     18624 DF
 >> >>  Completed_10       583.89 min   6021.55     77.15    953.05
   10 LogL= -51686.80     18624 DF
```

Current values are

```
# Term, Position, PType, PZUF, value
"mrm12(id);mrm12(id)_1", 1, V, P,    1.36706              C112
"mrm12(id);mrm12(id)_2", 2, V, F,   0.655532E-05          C212   !wants to be
negative
"at(Tr|1).mrm12(id);mrm12(id)_1", 3, V, F,   0.132336E-03  S111
"at(Tr|1).mrm12(id);mrm12(id)_2", 4, V, P,    2.73370      S122
"at(Tr|2).mrm12(id);mrm12(id)_1", 5, V, P,    6.04907      S211
"at(Tr|2).mrm12(id);mrm12(id)_2", 6, V, P,   0.984232      S222
"units.us(Trait);us(Trait)_1", 7, V, P,    65.7918
"units.us(Trait);us(Trait)_2", 8, G, P,    18.6025
"units.us(Trait);us(Trait)_3", 9, V, P,    99.5915
```

Li reran this model without rescaling imf

```
   17 LogL= -29669.84     18624 DF

 Akaike Information Criterion   59357.69 (assuming 9 parameters).
 Bayesian Information Criterion 59428.18

          Analysis of imf sf5

                      Wald F statistics
Source of Variation        NumDF     DenDF     F-inc              P-inc
 Trait                         2              132.29
 at(Tr,1).cg_imf             375               7.05
 at(Tr,2).cg_sf5             375              21.71

 Model_Term                  Order     Gamma         Sigma     Z_ratio  %C
 mrm12(id)_V                  9688  0.131128E-04  0.131128E-04    0.00   0 F
 mrm12(id)_V                  9688  0.582274E-04  0.582274E-04    0.00   0 F
 mrm12(id)_V                  9688  0.187611      0.187611        7.20  -1 P
 mrm12(id)_V                  9688  0.586317      0.586317       13.77   1 P
 mrm12(id)_V                  9688  0.582274E-04  0.582274E-04    0.00   0 F
 mrm12(id)_V                  9688   50.2014       50.2014       11.43   3 P
 units.us(Trait)                 19376 effects
 us(Trait)_V                     2  0.270612      0.270612       14.43   0 P
 us(Trait)_C                     2  -1.32461      -1.32461      -16.87   0 P
```

```
 us(Trait)_V                     2   85.1505       85.1505       45.65   0 P

Covariance\Variance\Correlations for us(Trait) in units.us(Trait)
  0.270612   -0.2759
 -1.324614 85.150465
```

Arthur pursued this approach as follows:

1) I analysed Ximf = -10*imf instead of imf so that the traits were on similar scales and positively correlated.

2) I removed the constraint that variance components be positive.

The analysis converged for ten iterations but then the LogL started decreasing.

After 14 iterations, the components were

```
                        Wald F statistics
Source of Variation           NumDF     DenDF     F-inc              P-inc

 Model_Term                   Order     Gamma         Sigma     Z_ratio  %C
 mrm12(id)_V                   9688   6.99416       6.99416        5.18  27 U
 mrm12(id)_V                   9688  0.625977      0.625977        0.84 100 U
 mrm12(id)_V                   9688  -7.26578      -7.26578       -5.22 -42 U
 mrm12(id)_V                   9688  -1.51056      -1.51056       -1.55-100 U
 mrm12(id)_V                   9688   7.04983       7.04983        2.84  10 U
 mrm12(id)_V                   9688  0.108958      0.108958        0.07 -78 U
 units.us(Trait)                  19376 effects
 us(Trait)_V                      2   70.5639       70.5639       50.02   2 P
 us(Trait)_C                      2   13.1033       13.1033        9.47 -13 P
 us(Trait)_V                      2   94.1260       94.1260       37.37  -2 P

and after 24 iterations (Lower LogL),

 Model_Term                   Order     Gamma         Sigma     Z_ratio  %C
 mrm12(id)_V                   9688   5.48716       5.48716        5.23  -3 U
 mrm12(id)_V                   9688 -0.318143     -0.318143       -0.60 -47 U
 mrm12(id)_V                   9688  -7.68458      -7.68458      -12.18   0 U
 mrm12(id)_V                   9688  0.465476      0.465476        1.43  25 U
 mrm12(id)_V                   9688   3.66150       3.66150        2.34  -8 U
 mrm12(id)_V                   9688  0.511877E-01  0.511877E-01    0.06 -74 U
 units.us(Trait)                  19376 effects
 us(Trait)_V                      2   71.9030       71.9030       65.34   1 P
 us(Trait)_C                      2   14.8975       14.8975       12.24   0 P
 us(Trait)_V                      2   99.2710       99.2710       39.97   0 P
```

I then tried 2 simpler models

No specific variance

```
   5 LogL= -51694.67     18624 DF

 Akaike Information Criterion   103399.34 (assuming 5 parameters).
 Bayesian Information Criterion 103438.50

          Analysis of Ximf sf5

                         Wald F statistics
Source of Variation           NumDF     DenDF     F-inc              P-inc

 Model_Term                    Order     Gamma         Sigma     Z_ratio  %C
 mrm12(id)_V                    9688   2.72183       2.72183        2.33   0 U
 mrm12(id)_V                    9688  0.674830      0.674830        0.90   0 U
 units.us(Trait)                   19376 effects
 us(Trait)_V                       2   66.3667       66.3667       49.82   0 P
 us(Trait)_C                       2   16.7351       16.7351       13.20   0 P
 us(Trait)_V                       2   104.373       104.373       57.70   0 P
```

This shows little genetic covariance.

No genetic covariance between traits

The LogL increased to -51589.75

"at(Tr|1).mrm12(id);mrm12(id)_1", 1, V, U, -3.09116
"at(Tr|1).mrm12(id);mrm12(id)_2", 2, V, U, -1.55733
"at(Tr|2).mrm12(id);mrm12(id)_1", 3, V, U,   3.81006
"at(Tr|2).mrm12(id);mrm12(id)_2", 4, V, U,   4.75048
"units.us(Trait);us(Trait)_1", 5, V, P,   76.0781
"units.us(Trait);us(Trait)_2", 6, G, P,   20.3431
"units.us(Trait);us(Trait)_3", 7, V, P,   99.8333

which shows a negative component for Ximf and a positive components for sf5.

The genetic variation explained is much smaller than in the SVD models reported below, which suggests a problem with all the analyses I have done here .

GRM trimming (November 2019)

This extension was prompted by Li Li of AGBU.

!SAVE writes the inverse GRM to a binary file ….bgiv when a GRM is inverted, saving the LogDet on the first line of the inverse file.

!TRIM trmID ID trmID.txt assumes the GRM file specified is indexed by the levels of factor ID but many of the levels have no data; that the file trmID.txt lists the levels with data so we create a new factor trmID with just those levels and subset the GRM matrix to just those levels

The TRIM option was developed to speed up the fitting of a simple GRM model where the GRM matrix was large but many levels had no associated data.

More coding is required to predict the BLUPs for which there is no direct data.

Two examples were tested.

full LimCtrim model

This is a refitting of the Limagrain LimC dataset with 657 genotypes evaluated across 8 seasons, but only about 100 in any one season. PART 1 fits the model as previously specified. PART 11 fits the same model under the new formulation.

A2.sgrm is a binary (real) file containing the GRM matrix as a rowwise lower triangle dense matrix. The varianble Indiv is coded 1:657 in the data file. The 8 files GS1.txt to GS8.txt contain in a single column, the levels of Indiv with data in each of the 8 seasons (derived from the .etb file formed by TABULATE Yield ~ Season Indiv). So, the 8 !TRIM lines create 8 reduced GIV matrices and the 8 SUBGROUP factors selecting the of Indiv pertaining to the GIV matrices.

```
!PART 11
A2.sgrm !TRIM Ind1 Indiv GS1.txt
A2.sgrm !TRIM Ind2 Indiv GS2.txt
A2.sgrm !TRIM Ind3 Indiv GS3.txt
A2.sgrm !TRIM Ind4 Indiv GS4.txt
A2.sgrm !TRIM Ind5 Indiv GS5.txt
A2.sgrm !TRIM Ind6 Indiv GS6.txt
A2.sgrm !TRIM Ind7 Indiv GS7.txt
A2.sgrm !TRIM Ind8 Indiv GS8.txt
!PART 0
A2.sgrm !SKIP 1
!PART 1 11//Pheno1.csv !SKIP 1
!PART 1
Yield ~ mu !r rr1(Season).grm1(Indiv)  diag(Season).grm1(Indiv) Env

!PART 11
Yield ~ mu !r rr1(Season).grm9(Indiv) +
at(Seas,1).grm1(Ind1) + at(Seas,5).grm5(Ind5) +
at(Seas,2).grm2(Ind2) + at(Seas,6).grm6(Ind6) +
at(Seas,3).grm3(Ind3) + at(Seas,7).grm7(Ind7) +
at(Seas,4).grm4(Ind4) + at(Seas,8).grm8(Ind8) + Env

>> >> >>           Setup A  took     0.17 sec      0.00     Etime      0.00
 >> >> >>           Setup D  took     0.31 sec      0.01     Etime      0.00
E64_groups >>       0       0       0    1668       2    6926    6926       1
 FILLIN    318495 >>    374753      1.17 FOLD
 >> >> >>       Sparse done  took     0.17 sec      0.01     Etime      0.01
```

```
 >> >> >>          AIAbsorb  took     2.62 sec      0.05     Etime      0.02
 >> >> >>    Sparse Inverse  took     0.25 sec      0.06     Etime      0.02
 >> >> >>          Complete  took     0.34 sec      0.06     Etime      0.02
 >> >> >>       Sparse done  took     0.44 sec      0.07     Etime      0.02
 >> >> >>          AIAbsorb  took     3.00 sec      0.12     Etime      0.04
 >> >> >>    Sparse Inverse  took     0.25 sec      0.13     Etime      0.04
 >> >> >>  Sparse Inverse r  took     0.12 sec      0.13     Etime      0.04
 >> >> >>          Complete  took     0.30 sec      0.13     Etime      0.04
 Finished: Wed Nov 20 16:59:03 2019  LogL Converged      Limctrim11_11/Limctrim
 Echidna 0.097 19 Nov 2019 Windows                  Wed Nov 20 16:58:46 2019
```

15 iterations: 17 seconds

```
   15 LogL= -18269.69   11.89         10020 DF

 Akaike Information Criterion   36575.38 (assuming 18 parameters).
 Bayesian Information Criterion 36705.20

          Analysis of Yield

                         Wald F statistics
Source of Variation            NumDF     DenDF     F-inc              P-inc
 mu                                1             3218.57

 Model_Term                      Order     Gamma         Sigma     Z_ratio  %C
 grm4(Ind4)                         98  0.429835E-01  0.511232        1.84   0 P
 grm1(Ind1)                        103  0.513613E-01  0.610875        1.91   0 P
 grm5(Ind5)                        105  0.477587E-01  0.568026        2.08   0 P
 grm6(Ind6)                        108  0.561629E-01  0.667983        1.84   0 P
 grm3(Ind3)                        114  0.217698E-06  0.258923E-05    0.00   0 B
 grm2(Ind2)                        116  0.241484       2.87213        4.08   0 P
 Env                               116   16.3857       194.886        7.57   0 P
 grm8(Ind8)                        123  0.623852E-02  0.741989E-01    0.41   0 P
 grm7(Ind7)                        128  0.153256       1.82277        3.58   0 P
 rr1(Season).grm9(Indiv)              5913 effects
 rr1(Season)_L              0   1   9  0.323673       1.11626       18.07   0 P
 rr1(Season)_L              0   2   9  0.388868       1.34110       15.85   0 P
 rr1(Season)_L              0   3   9  0.471212       1.62508       36.15   0 P
 rr1(Season)_L              0   4   9  0.405416       1.39816       24.75   0 P
 rr1(Season)_L              0   5   9  0.374263       1.29073       24.00   0 P
 rr1(Season)_L              0   6   9  0.473887       1.63430       27.16   0 P
 rr1(Season)_L              0   7   9  0.397742       1.37170       20.00   0 P
 rr1(Season)_L              0   8   9  0.406401       1.40156       30.78   0 P
 Residual_units                10021   1.00000       11.8937       67.74
```

Echidna previously seems to have taken 45 sec per iteration (40 times longer).

IMF GRM

This example was provided by Li Li of AGBU. His first goal was to estimate the genomic variance component for a given GRM matrix, adjusting the data for CG (Contemporary groups). The GRM matrix has 22394 genotypes but only 9311 have data. Including the others in the analysis means we can get BLUPs for them, but does not contribute information to the analysis.

The following code has 3 parts. Imf.ped has the genotype labels in the order of the GRM file.

```
!ARG 1   !WORK 3
Top SNPs analyses    !DOPART $1
 Ani !A 22394 !LL 20 !L imf.ped
 Sir !A 2240
 Dam !A 15393
 cg !I  366
 imf

!PART 1        #  Original analysis takes over 2 hours per iteration
bin.bgrm !TRIM trmAni Ani imftrm.txt
imf.dat !MAXIT 100 !GDENSE

imf  ~ mu   grm1(Ani) 1.5 !f cg
residual idv(units)


!PART 2     # Use tabulate to get a list of genotypes with data
imf.dat !MAXIT 100 !GDENSE
tab imf ~ Ani    !LIST
imf ~ cg

!PART 3   #  Trim the GRM to just include genotypes with data.   6 min per iteration
bin.bgrm !TRIM trmAni Ani imftrm.txt
imf.dat !MAXIT 100 !GDENSE

imf  ~ mu   grm1(trmAni) 1.5 !f cg
residual idv(units)
```

Explanation:

1 In the original code (supplied by Li), imf.ped just provided the genotype names in the order of the GRM matrix but was introduced as a pedigree. A simpler way is to incorporate the list into the definition of Ani.

2 In part 2, we use TABULATE to get a list of genotypes with data. We then converted imft.etb to imftrm.txt by stripping out all fields and rows except the last row containing the list of genotypes with data.

There were 9311 of these.

3. Run part 3. The key is the extension of the GRM data line with the qualifier

!TRIM trmAni Ani imftrm.txt

trmAni is the name of a new factor created by the qualifier which is linked to the factor Ani having the levels defined in inftrm.txt

The routine creates the list of levels of Ani to be included in trmAni and then extracts those rows from the GRM matrix supplied. The new matrix has order 9311, is inverted and used as normal. So the model line links grm1() with trmAni

The timing information under Windows on HP15 is

```
N:\2019\ASR\LiLi\IMF\IMF>grep ">>" imft.esl
 >> >> >>  nrm_den_bin.bgrm  took   314.30 sec      5.24     Etime      3.08
 >> >> >>           Setup D  took     1.00 sec      5.25     Etime      3.09
 >> >> >>             SSP+G  took     0.33 sec      5.26     Etime      3.09
E64_groups >>       0       0       0    9677       2    9679    9679       1
 >> >> >>             Order  took     0.12 sec      5.26     Etime      3.09
 >> >> >>         SymbolicA  took    26.91 sec      5.71     Etime      3.54
 FILLIN  42722919 >>  42723576      1.00 FOLD
 >> >> >>         SymbolicD  took     2.31 sec      5.75     Etime      3.58
 >> >> >>       Sparse done  took  1338.55 sec     28.06     Etime      6.42
 >> >> >>          Absorb C  took     0.61 sec     28.07     Etime      6.42
 >> >> >>          AIAbsorb  took     1.22 sec     28.09     Etime      6.43
 >> >> >>    Sparse Inverse  took  1027.34 sec     45.21     Etime      8.60
 >> >> >>  Sparse Inverse r  took     3.94 sec     45.28     Etime      8.64
 >> >> >>          Complete  took     0.14 sec     45.28     Etime      8.65
 >> >> >>          Complete  took  2370.77 sec     84.79     Etime     13.79

which shows a Setup time of 3.1 minutes  (down from 34 min)
First iteration completed after 8.85 minues (down from 175 min)
Subsequent iterations just take 5 minutes.

Timings may differ slightly between Windows and Linux OS.

This run converged to

    1 LogL= -4162.02      8945 DF
    2 LogL= -2895.36      8945 DF
    3 LogL= -2818.29      8945 DF
    4 LogL= -2817.70      8945 DF
    5 LogL= -2817.70      8945 DF

 Akaike Information Criterion   5639.40 (assuming 2 parameters).
 Bayesian Information Criterion 5653.59

          Analysis of imf

                         Wald F statistics
Source of Variation           NumDF     DenDF     F-inc              P-inc

 Model_Term                  Order     Gamma         Sigma     Z_ratio  %C
 grm1(trmAni)                 9311  0.417077      0.417077       15.09   0 P
 idv(units)                   9311  0.272285      0.272285       13.17   0 P

  cg                                    366 effects fitted.
  grm1(trmAni)                         9311 effects fitted.

 Finished: Thu Nov 21 15:49:06 2019  LogL Converged      imft

 Started:  Echidna 0.097 19 Nov 2019 Windows 3360 Mbyte  at Thu Nov 21 15:19:56 2019
```

These results (5 iterations in 30 min) agree with those from ASReml (below) which took over 9 hours for 3 iterations:

```
 3 LogL=-2817.70     S2= 0.27236       8945 df

         - - - Results from analysis of imf - - -
Akaike Information Criterion     5639.40 (assuming 2 parameters).
Bayesian Information Criterion   5653.59

Model_Term                              Gamma         Sigma   Sigma/SE   % C
grm1(Ani)              GRM_V 22394   1.53175      0.417190      15.10   0 P
idv(units)                    9311 effects
Residual                SCA_V 9311   1.00000      0.272362      13.17   0 P

                                Wald F statistics
    Source of Variation          NumDF              F-inc

                   Solution       Standard Error    T-value     T-prev
  7 grm1(Ani)                   22394 effects fitted (      65 are zero)
  4 cg                            366 effects fitted
* This job used at least 5890 of the 43263 Mbyte of primary workspace. *
SLOPES FOR LOG(ABS(RES)) on LOG(PV) for Section   1   0.99
         31  possible outliers: see .res file
Finished: 07 Nov 2019 05:35:08.201   LogL Converged

Started:  06 Nov 2019 20:37:10.275  49444 Mbyte  IMFD
```

June 2020 Three examples

The stated objective is to perform bivariate genomic analysis.

Small

For this, Li provided a trimmed GRM for 1000 animals. There were actually 3 G matrices (A22, G_r and G_t), a list of genotype IDs (ped.csv) and a data file (data.csv) containing variables ID, CG_i, CG_f, imf, sf5 where the CG (Contemporary Group) factors pertain to the two traits to be analysed. This dataset is provided as a development testing example. A22 is derived from the A matrix (absorbing parents without data). G_t is based on a marker panel and G_r represents a different marker panel.

The imf data has a standard deviation of 0.97. The variance after fitting CG is 0.63 (SD 0.79).

The sf5 data has a standard deviation of 8.04. The variance after fitting CG is 41.44 (SD 6.44).

The G_t GRM is negative definite; 1 negative pivot making all diagonals of the inverse negative and large. Using !ADD (which adds 0.000001 to the diagonal) rectified the issue.

Standard univariate analysis estimates are

	LogL	Iterations	σ^2_g	σ^2_e	Time	Gmatrix	

imf	-281.62	7	0.3065	0.3223	4s	A22.bgrm	
	-277.78	7	0.3022	0.3386	5s	G_r.bgrm	
	-276.47	6	0.2193	0.4041	5s	G_r.bgrm	
	-274.74	5	0.0000 0.1496 0.1375	0.3484	10s	A22 G_r G_t	
Sf5	-2278.73	7	18.69	23.16	5s	A22.bgrm	
	-2277.38	6	13.79	27.63	5s	G_r.bgrm	
	-2274.32	5	11.92	29.12	4s	G_t.bgrm	
	-2273.81	5	5.04 0.77 9.30	26.36	10s	A22.bgrm G_r.bgrm G_t.bgrm	

These results were produced with the code:

```
!RE !ARG 1 2 3 4
Original imf !DOPART $1
ID !A
CG_i  *
CG_f  *
imf
sf5
a22.bgrm
G_r.bgrm
G_t.bgrm      !ADD
data.csv !skip 1
!part 1 //imf ~ mu CG_i
!part 2 //sf5 ~ mu CG_i
!part 3 //imf ~ CG_i !r grm1(ID)
!part 4 //sf5 ~ CG_i !r grm1(ID)
!part 5 //imf ~ CG_i !r grm2(ID)
!part 6 //sf5 ~ CG_i !r grm2(ID)
!part 7 //imf ~ CG_i !r grm3(ID)
!part 8 //sf5 ~ CG_i !r grm3(ID)
!part 9 //imf ~ CG_i !r mrm123(ID)
!part 10//sf5 ~ CG_i !r mrm123(ID)
```

Medium (10K)

This data has the same structure as the Small set but includes 9688 individuals and an extra G matrix (GG.bgrm) representing genetic groups.

Standard univariate analysis estimates (timings on HP15; 8 threads, 32Gb) are

	LogL	Iters	σ^2_g	σ^2_e	Time/itn	Form Gi	Gmatrix	Total

imf	-3517.13	5	0.0185	0.66774	230s (HP16)	152s 90s 60s 180s(HP16)	A22.bgrm	20:10
	-3286.96 -3514.07	30 5	0.0901 0.083	0.5986 0.6782	232s	 With !ADD	G_r.bgrm	153m
	-3516.96	5	0.0073	0.6885	232s		G_t.bgrm	22:50
	-	5					A22 G_r G_t	
Sf5	-26986.51	3	10.032	98.153	228s	152s	A22.bgrm	14:01
	-26855.96 -26991.79	30 4	14.047 4.2	94.616 103.674		 With !ADD	G_r.bgrm	
	-26995.47	5	1.402	106.381			G_t.bgrm	
	-						A22.bgrm G_r.bgrm G_t.bgrm	

Adding parallel processing to SmiDense Back step reduced Form Gi time from 152 to 90 on HP 15!

Adding parallel processing and 8row blocking to SmiDens Back step reduced Form Gi time from 90 to 60 on HP 15! However, 8row blocking does not appear to help with this same inversion on HP 16 (4 threads!) taking 175-180 s.

In most cases, Echidna converged in ~ 5 iterations; but with the G_r matrix, it struggles to converge, the LogL increasing by 16 over 26 iterations. The issue is that the genetic variance is negatively correlated with the error variance. This G_r matrix generates 1 negative pivot will absorbing and the result is all diagonal elements of the inverse are negative. The !ADD qualifier adds 0.000006 to the diagonal before inverting and the result is positive definite. This greatly helps convergence

I tried the !GDENSE qualifier but it was not faster (260s per iteration).

```
>> >>   Echidna Process      CPU        SumCPU    WClock     SumWC
 >> >>           Setup A     0.19 sec      0.00      0.16      0.00
 >> >>             SSP+G     0.45 sec      0.01      0.46      0.01
 >> >>         Get Order     0.62 sec      0.02      0.63      0.02
 FILLIN  46963708 >>  46963708      1.00 FOLD  DSA     2   10064  4666
 >> >>         Reorder C     5.09 sec      0.11      5.10      0.11
        3213  Bi/Tri node calls >>
 >> >>          Absorb C   528.48 sec      8.91     68.88      1.25
 >> >>          AIAbsorb     1.14 sec      8.93      0.27      1.26
 >> >>         Ci formed  1198.06 sec     28.90    155.41      3.85
 >> >>      Ci reordered     4.81 sec     28.98      3.57      3.91
 >> >>  Completed_1          0.16 sec     28.98      0.15      3.91
 >> >>  Completed_2       1735.97 sec     57.92    238.15      7.88
```

Large

For this example, Li provided the inverse G inverse of order 31572, a data file with 31572 records and a matching list of genotype names. The data file included a CG factor and a response variable present for 10580 records.

Appendix 1 shows the output from GCTA (`(C) 2010-2019, The University of Queensland, Jian Yang <jian.yang@uq.edu.au>`) which performed 15 AI iterations in 6 min 28 sec with a model also including a second G matrix (for genetic groups) and estimating the components for GG, GRM and Residual as `0.03057 0.30783 0.38199` respectively.

After fixing a few bugs related to reading the Genotype names (because there was so many), Echidna calculated the Log Determinant of the G inverse using routine SmiDense. This routine was not optimised for speed and took 24 hours. I have now optimised that using parallel processing and nodes of size 80 to take 14 minutes, but the job ran out of memory on HP20 (Mac with 16Gb RAM and 12 threads).

Having now calculated the LogDet as 46201.55, Echidna as the option of the user supplying the LogDet or calculating an approximate one from the average variance and covariance of the inverse.

The TRIM mechanism needs to start with a GRM, which I do not have immediately. Bit I did trim the G inverse (to size 10580) and fit the model using that. It took 6 minutes per iteration (an hour to do 10 iterations – not converged) to report variances of 0.0175 (GRM) and 0.0128 (Residual)) which bare little resemblance to the GCTA values.

Calculating the determinant is a byproduct of the first part of inversion. I sought to invert this whole matrix (order 31572) and it took 20 hours. Parallel processing had not been added to the second half of the inversion processing. Adding it only reduced the time to 17 hours. But adding nodal processing, it took 9:15. These times seem inconsistent.

SVD approach

Michael Roper sought my help in 2017 to test a GBLUP analysis method he proposed which utilized the Singular Value decomposition of the G matrix.

SVD refresher

```
L = matrix(c(1,1,1,1,1 ,1,1,1,2,2,2,-1,-1,0,0,0,1,-1,0,0,0,0,0,1,-1),5,5)
M=L %*% t(L)
svdM=svd(M)
Mu=svdM$u
Md=svdM$d
Mu %*% diag(Md) %*% t(Mu)    #  is M
t(Mu) %*% Mu                 #  is I
Mu %*% diag(1/Md) %*% t(Mu)  #  is inverse(M)
```

For the case where we have a single observation for each genotype and calculate a GRM matrix as MM', and want to fit the model

y =XB +Zu where var(y) is s^2(I + gG) since Z is I, we can calculate the eigen decomposition of G as LDL' and then transform the model to L'y = L'XB + L'Zu +L'e and var(Lý) is s^2[L'L + gL'GL] = s^2 [I +gD].

That is, decompose G, premultiply y and X by L' and solve the resultant simpler equations.

In the case where rank G is less than order of G (fewer markers than genotypes) some elements of D are zero but we still need the whole matrix L'.

2017 analyses for Michael Roper

Standard approach

```
!WORKSPACE 1 !RENAME !DEBUG !LOG !OUT !ARGS 1
Title: Xy.  !DOPART $1
#meas_accuracy,aminoacid,y
#=,Z,                      2.00148000021
#=,P,                      2.56494935746
#=,P,                      3.04452243772
#=,P,                      1.79175946923
 meas_accuracy  !A      # =
 aminoacid  !A      # P
 y        # 1.79175946923
 ID * !=V0

!PART 1
G.grm         !PSD    !PRECISION
Xy.csv  !SKIP 1      !DDF !FCON      !MAXIT 21
y  ~ mu meas amino,    !r   grm1(ID) 1
residual units
```

There are 1075 data records and G is based on 462 markers (so has 613 singularities).

```
    8 LogL= -342.23  0.4954          1069 DF

                         Wald F statistics
Source of Variation           NumDF     DenDF     F-inc     F-con    P-inc
 mu                               1               10.30     10.30
 meas_accuracy                    3               90.76     90.79
 aminoacid                        2                0.29      0.29

 Model_Term                    Order     Gamma         Sigma     Z_ratio  %C
 grm1(ID)                       1688   12.7625       6.32232        7.81   0 P
 Residual_units                 1075   1.00000      0.495383       20.16
```

Michael then transformed the data and X design matrix
And we can run

```
!WORKSPACE 100 !DEBUG !LOG !RENAME !OUT !ARGS 2 4
Title: MMX-II\Ex\MRxyg\tXyG. !DOPART $1
#ID,Intercept,meas_accuracy=,meas_accuracy>,meas_accuracyNULL,aminoacidP,aminoacidZ,tG,ty
```

```
#1,-32.77608,-23.17167,-6.795595,-2.715653,-22.51105,-0.06176557,961.8945,-67.80353
#2,0.268823,-0.2023037,2.1990785,-1.7318146,-14.918853,-0.040410,13.991215,5.0663441
#3,-0.4678455,-3.3476936,2.1899236,0.692786,0.9636929,0.0127684,8.6003273,29.883288
#4,-0.2362078,3.45968686,-6.152741,2.37584257,-0.01825295,0.06537036,5.521165,-43.51705
 ID  * !M>462      # 4
 Intercept         # -0.236207834
 MEeq              # 3.459686857    #meas_accuracy=
 MEgt              # -6.152741035   #meas_accuracy>
 MEnull            # 2.375842569    #meas_accuracyNULL
 AAP               # -0.018252946   #aminoacidP
 AAZ               # 0.065370363    #aminoacidZ
 G                 # 5.521165046
 tY                # -43.51704702
 GW !=G !*12.7  !+1 !^-1

dgGi.giv
dgG.grm !skip 1   # Diag form
tXyG.csv  !SKIP 1   !MVI    !FCON !DDF

!PART 2  // tY   ~ Inter MEeq MEgt MEnull AAP AAZ !r grm1(ID)
!PART 4  // tY   ~ Inter MEeq MEgt MEnull AAP AAZ !r grm2(ID)
!PART 3  // tY !WT GW  ~ Inter MEeq MEgt MEnull AAP AAZ
```

Resulting in

```
   8 LogL=-342.225     S2= 0.49527       1069 df    12.78
 Final parameter values                            12.76

          - - - Results from analysis of tY - - -
 Akaike Information Criterion      688.45 (assuming 2 parameters).
 Bayesian Information Criterion    698.40

          Approximate stratum variance decomposition
 Stratum     Degrees-Freedom   Variance      Component Coefficients
 grm1(ID)             255.98    1.64890         0.2     1.0
 Residual Variance    813.02   0.495373         0.0     1.0

 Model_Term                              Gamma         Sigma   Sigma/SE   % C
 grm1(ID)               GRM_V  462   12.7624       6.32214       7.80   0 P
 Residual               SCA_V 1075   1.00000      0.495373      20.16   0 P

                                  Wald F statistics
     Source of Variation          NumDF    DenDF_con F-inc    F-con M P-con
   2 Intercept                        1     160.1    10.29     1.31 A 0.254
   3 MEeq                             1     878.9    44.80    15.57 A <.001

   4 MEgt                             1     887.5   221.86    40.88 A <.001
   5 MEnull                           1     875.8     5.57     5.56 A 0.019
   6 AAP                              1     135.2     0.05     0.47 A 0.494
   7 AAZ                              1     404.7     0.52     0.52 A 0.470
```

This gives the same variance components and LogL. However, the effects would need to be transformed back to the observation scale.

The actual run times were 6 seconds and 0.3 seconds (not including the time to do the eigen factorization and to transform the design matrix).

The big question is whether this can be generalised to the case of repeated observations!

2020 Gsubset GID.txt ISKIP GRM.bgrm DATA.csv DSKIP IDFLD YFLD

To develop and further test this approach, I have written a series of programs to perform individual steps. The program Gsubset replicates the Echidna GRM TRIM operation.

1. It reads the list of Genotype identifiers (GID.txt has the genotype identifiers as the first field of each line) after copying any (ISKIP) header lines to the file to contain the reduced list of genotypes: GID_yfld.txt.
2. It reads the data file (DATA.csv) first copying any (DSKIP) header lines to the file to contain the reduced data set: DATA_yfld.csv. When each record is read, field YFLD is examined. If it is a missing observation, the record is discarded. Otherwise the record is copied to DATA_yfld.csv and the presence of the ID from field IDFLD is noted against the list of genotypes from step 1.
3. Finally, the GRM file is read (bgrm means real*4 binary lower triangle rowwise). If this genotype is not in the reduced data, the record is discarded. Otherwise the cells corresponding to genotypes with data are written to GRM_yfld.bgrm.

```
IC=1
DO IA=1,nrg
   read(13) RVEC(1:IA)
   IF(IA.eq.key(Ic)) THEN
      write(14) rvec(key(1:Ic))
      Ic=Ic+1
   ENDIF
ENDDO
```

2020 Ginverse ABC.bgrm

This program reads the nominated .bgrm file (lower triangle rowwise dense) and uses Echidna's SmiDense routine to invert it. Inverse is written in the same format to a NEW file with the file extension changed to .bgiv (.bgrm if the input file is .bgiv). If the output file already exists, delete it before running the program. The determinant is written as the second element on the first line.

2020 Geigen ABC.bgrm

This program reads the nominated .bgrm file (lower triangle rowwise dense) and uses MKL to perform Singular Value Decomposition (G=UDU'), writes the dimension and D to ABC_D.bgrm and writes U to ABC_U.bgrm. The write statements are:

```
WRITE(12) NR
WRITE(12) sngl(XX(LSV:LSV+NR-1))
CLOSE (12)
LL=LU-1
DO IR=1,NR
  WRITE(13) sngl(XX(LL+1:LL+NR))
  LL=LL+NR
ENDDO
CLOSE(13)
```

This program took 3 seconds to factor a G matrix of order 1000, but 90 minutes to factor a matrix of order 9968 (on HP16 4 threads 16Gbyte all used)

2020 Gtransform ABC_U.bgrm Data.csv DSKIP Yfld IDfld CGfld

The purpose of this program is to premultiply the design for a simple model by the U matrix from the SVD (produced by Geigen) and write the result to a binary file for analysis in Echidna.

ABC_U.bgrm is file of eigen vectors produced by running Geigen ABC.bgrm

Data.csv is an ascii data file with at least 3 fields (ID, CG and Y)

Dskip is the heading lines in Data (to be ignored)

Yfld IDfld and CGfld identify fields to access in the data file. After dropping records for which the response Y is missing, the number of rows retained must match the size of the GRM matrix (retrieved from ABC_D.bgrm). It is assumed the data file is in the order of the G matrix. IDfld is not actually referenced. It is assumed the CG field is numeric coded 1:NCG where NCG is number of CG classes and the maximum value is taken as NCG.

The binary data file produced is Data_Yfld.bin and has 2+NCG fields being ID (recoded 1:NR), Uy (transformed response) and UX (transformed CG design matrix).

With this result, the Echidna code to run the analysis is

```
Transformed Gblup analysis
ID  *    # ID * !LL 20 !L IDnames.csv  # if you have a file with the correct listof
ID names
tY       # Transformed response variable
tCG !G 51   # where 51 is the number of CG classes reported by Gtransform
ABC_D.bgrm   # has the eigen values to use in the analysis
Data_5.bin   #  Binary data
tY ~ !r tCG  grm1(ID)
   1 LogL= -27023.03   100.2          9312 DF
    2 LogL= -27003.21   103.2          9312 DF
    3 LogL= -26995.47   106.3          9312 DF
    4 LogL= -26995.47   106.4          9312 DF

 Akaike Information Criterion   53994.93 (assuming 2 parameters).
 Bayesian Information Criterion 54009.21

          Analysis of sf5

                        Wald F statistics
Source of Variation           NumDF     DenDF     F-inc              P-inc

 Model_Term                    Order     Gamma         Sigma     Z_ratio  %C
 grm1(id)                       9688  0.131835E-01   1.40248        1.71   0 P
 Residual_units                 9688   1.00000       106.381       61.66
```

```
  cg_sf5                                   376 effects fitted.
  grm1(id)                                9688 effects fitted.
```

2020 Gtransform revised: ABC Data.csv DSKIP FldKeys

This revised version (July 14) allows multiple responses to be transformed at once by use of a string variable FldKeys. This is a positional key string of up to 20 characters. An F indicates the corresponding field is a factor coded 1:n which is to be transformed. A Y indicates a (response) variate in the corresponding field is to be transformed. Any other character (e.g. I or -) means the corresponding firld is ignored. The other change is that the "_U.bgrm" file extension is assumed.

For example, the test example involved

```
grm1(id)Geigen G_t.bgrm
Gtransform G_t data.csv 1 IFFYY
```

And created a binary file data_IFFYY.bin containing the transformed data.

The screen output includes the basic code for the Echidna job to analyses the data

2020 MR (repeat Michael's analysis)

Run Gbgrm to convert G.grm (dense ASCII) to G.bgrm

Run Geigen to generate _U.bgrm ND D_bgrm

E:\MMX-II\Ex\MR>C:\Users\Arthur\Dropbox\MMX\GRM\Geigen.exe G.bgrm

```
        1075  rows read from G.bgrm Trace    1075.000
     1.000     0.939     0.912
     0.939     1.000     0.953
     0.912     0.953     1.000
   961.894    13.991     8.600

   -32.679
```

The first 3 lines show the top of the file, and that it has remapped from LT to fullstored.

The 4th line is the first the eigen values, and agrees with what Michael reported.

961.894492210141, 13.9912150352729, 8.60032731702465

The next step is to transform the data (xy.csv) which starts

```
meas_accuracy,aminoacid,y
=,Z,                    2.00148000021
=,P,                    2.56494935746
=,P,                    3.04452243772
=,P,                    1.79175946923
=,P,                    2.07944154168
=,P,                    -0.69314718056
=,Z,                    0
```

```
=,P,                    2.83321334406
=,P,                    3.43398720449
```

The data summary is

```
          y
   Count         Mean     StndDevn      Minimum      Maximum   ME     AA
       2       1.0007       1.4153       0.0000       2.0015  =       Z
     521       1.7532       1.7164      -1.6094       5.2306  =       P
     237       1.2801       1.6861      -1.6094       4.6367  =       N
     124       4.5008       0.5463       3.7136       6.2146  >       P
     101       4.2552       0.3039       3.6402       5.1671  >       N
       2      -1.6094       0.0003      -1.6094      -1.6094  <       P
       1      -1.2040       0.0000      -1.2040      -1.2040  <       N
      87       0.3062       1.0239      -1.6094       4.1431  NULL    P
```

As a table:

=	2	521	237
>		124	101
<		2	1
NULL		87	

I used Echidna to produce a binary file called xy_SAVE.bin. ME is the major fixed effect (Fc=260).

The first field has values “=”, “>” and “NULL”. The second field has values “Z” (2), “P” and “N”.

Next step is to premultiply the design matrix by U from the eigen analysis. At this time, Gtransform just accommodates 1 factor. If the input data file is binary, the DSKIP field is taken as the number of fields.

E:\MMX-II\Ex\MR>GT G_U.bgrm xy_save.bin 3 3 2 1

```
E:\MMX-II\Ex\MR>C:\Users\Arthur\Dropbox\MMX\GRM\Gtransform G_U.bgrm xy_save.bin 3 3 2 1
 Gtransform XXX_U.bgrm Datafile DSKIP Yfld IDFLD CGfld
 GRM eigen vectors read from G_U.bgrm
 Getting size from corresponding _D file
 Original data file:xy_save.bin
 Transformed data written to xy_save_3.bin
           4  levels in CG factor
   2.00   1.00   0.00   0.00   0.00
    1.0000   -1.0635   -0.3239    0.5107   -0.0289   -0.0263
    2.0000   -3.8572   -0.4599   -0.5118    0.0025    0.2352
    3.0000   -0.7129   -0.4094   -0.1420   -0.0077   -0.4601
    4.0000    2.1805   -0.4039    0.2731   -0.0014    0.3559
 transformed data written:            6  fields ID, tY, tCG
```

The transformed data supplied by Michael (with reduced accuracy) began

```
E:\MMX-II\Ex\MR>head -5 txyg.csv
ID,Intercept,meas_accuracy=,meas_accuracy>,meas_accuracyNULL,aminoacidP,aminoacidZ,tG,ty
1,-32.7761,-23.1717,-6.7956,-2.7156,-22.5111,-0.0618,961.8945,-67.8035
2,  0.2688, -0.2023, 2.1991,-1.7318,-14.9189,-0.0404, 13.9912,  5.0663
3, -0.4678, -3.3477, 2.1899, 0.6928,  0.9637, 0.0128,  8.6003, 29.8833
4, -0.2362,  3.4597,-6.1527, 2.3758, -0.0183, 0.0654,  5.5212,-43.5170
```

The values do not match.

Else with U transposed (call it Left)

```
   2.00   1.00   0.00   0.00   0.00
    1.0000  -67.8035  -23.1717   -6.7956   -0.0932   -2.7157
    2.0000   -5.0663    0.2023   -2.1991   -0.0039    1.7318
    3.0000  -29.8833    3.3477   -2.1899    0.0029   -0.6928
    4.0000  -43.5170    3.4597   -6.1527    0.0810    2.3758
 transformed data written:            6  fields ID, tY,
```

Now fitting the untransformed model:

```
XY Untransformed   !DOPART $1
ME !A
AA !A
y
ID * !=V0
G.grm  !SAVE !ND    !ADD
Xy.csv !skip 1
y ~ mu ME !r grm(ID)
```

Gave results

```
11 LogL= -341.38  0.4950          1071 DF

                        Wald F statistics
Source of Variation           NumDF     DenDF     F-inc              P-inc
 mu                               1               10.33
 ME                               3               90.92

 Model_Term                     Order     Gamma         Sigma     Z_ratio  %C
 grm(ID)                         1075   12.7194       6.29601        7.82   0 P
 Residual_units                  1075   1.00000      0.494993       20.172.3758
```

The transformed model is

```
XY transformed   !DOPART $1
ID *
tY
ME !G 4
```

```
G_D.bgrm  !SAVE   !ADD
Xy_SAVE_3.bin
y ~ mu ME !r grm(ID)
```

In ASReml 4.2 I got

```
 LogL=-341.377     S2= 0.49501       1071 df    12.72

          Approximate stratum variance decomposition
 Stratum     Degrees-Freedom   Variance      Component Coefficients
 grm(ID)              256.94    1.64955          0.2     1.0
 Residual Variance    814.06   0.494984          0.0     1.0

 Model_Term                            Gamma         Sigma   Sigma/SE   % C
 grm(ID)                GRM_V  462   12.7190       6.29571       7.82   0 P
 Residual               SCA_V 1075   1.00000      0.494984      20.18   0 P
```

Which agrees well with fit that Michael had. But I havn't managed to get Echidna to fit it correctly.

Now this GRM is based on 462 markers and ASReml is fitting just 462 effects.

Doing that in Echidna produces

```
    3 LogL= -341.38  0.4950          1071 DF


 Akaike Information Criterion   686.75 (assuming 2 parameters).

 Bayesian Information Criterion 696.71


         Analysis of tY


                        Wald F statistics

Source of Variation           NumDF     DenDF     F-inc              P-inc

 ME                               4              70.77


 Model_Term                    Order     Gamma         Sigma     Z_ratio  %C

 grm(ID)                         462   12.7194       6.29622        7.83   0 P

 Residual_units                 1075   1.00000      0.495010       20.17
```

But this is a bit messy.

Nevertheless, Now try on a TenK example.

Standard analysis of sf5 using G_t.bgrm took 150s per iteration on Mac and gave results

```
   1 LogL= -27023.03   100.2          9312 DF
    2 LogL= -27003.21   103.2          9312 DF
    3 LogL= -26995.47   106.3          9312 DF
    4 LogL= -26995.47   106.4          9312 DF

 Akaike Information Criterion   53994.93 (assuming 2 parameters).
 Bayesian Information Criterion 54009.21

          Analysis of sf5

                        Wald F statistics
Source of Variation           NumDF     DenDF     F-inc              P-inc

 Model_Term                     Order     Gamma         Sigma     Z_ratio  %C
 grm1(id)                        9688  0.131835E-01   1.40248        1.71   0 P
 Residual_units                  9688   1.00000       106.381       61.66
  cg_sf5                                 376 effects fitted.
  grm1(id)                              9688 effects fitted.
```

Geigen G_t.bgrm took 36 min on Mac

NB This took 90min (on HP16) to factorise the G_t.bgrm matrix.

Gtransform G_t_U.bgrm data.csv 1 5 1 3

Creates the transformed data file (data_5.bin)

I then converted G_t_D.bgrm to G_t_D.giv

And ran the job

Tranformed analysis
ID *
tY
tCG 376
G_t_D.giv
dat_5.bin
tY ~ mu !r grm1(ID) !f tCG

And it took 16 seconds to do 4 iterations and estimate

LogL -52862.20 GenVar 18.26 ResVar 86.71

But this analysis reports 9690 records rather than 9688 due to a bug in the original eigen routine.

While this is a plausible result, it does not agree with the result estimated above. The LogL is not ‘correct’ because it has used an approximate LogDet.

Factorising G_r.bgrm took 11:29 on HP15 (8threads); a22.bgrm took 11:10, G_t.bgrm took 11:16 and GG.bgrm took 9:46. Presumably GG was faster because it is reduced rank. But Mac took 36 min

So, if I streamline the process: this is faster than the former approach.

Detailed results

```
!REN !ARG
 G_t_U.bgrm data.csv IFFYY     !DOPART $1
 ! id,cg_imf,cg_sf5,imf,sf5
 ID         9688
 tY4
 tY5
 tF2 !G 376
 tF3 !G 376
 G_t_D.bgrm
 data_IFFYY.bin
!PART 4
tY4 ~ tF2 !r grm1(ID)

!PART 5
tY5 ~ tF3 !r grm1(ID)

!PART 45
tY4 tY5 ~ -Trait Tr.tF3 !r us(Tr).grm1(ID)
```

Part 4 produced

```
Data File: data_IFFYY.bin
 Summary of 9688 data records

 Variable   Levels Miss Zero      Min      Max    Distribution or Mn SD Sk Kt
 ID           9688    0    0        1     9688
 tY4             1    0    0   -23.20    21.94     0.03     4.35  -0.06   2.59
 tY5             1    0    0  -211.62   179.69     0.65    38.05  -0.05   2.14
 tF2        1  376    0    0 -0.09166  0.08662  0.00037  0.02272   0.00   0.04
 tF2        2  376    0    0 -0.21321  0.13912 -0.00003  0.03520  -0.01   0.20
...
tF3      376  376    0    0 -0.91838  1.70472 -0.00007  0.07039   2.95  75.83
 Note: Using !DOPART 4
 Note: Model is fitting 10064 equations, DENSE portion has 376 equations.
 * This job may use 12 processor threads. *
    1 LogL= -23235.41  0.5760          9312 DF
    2 LogL= -23160.65  0.5475          9312 DF
    3 LogL= -23086.23  0.5023          9312 DF
    4 LogL= -23073.32  0.4787          9312 DF
    5 LogL= -23073.17  0.4759          9312 DF
    6 LogL= -23073.17  0.4760          9312 DF
```

```
 Akaike Information Criterion   46150.33 (assuming 2 parameters).
 Bayesian Information Criterion 46164.61

          Analysis of tY4

                          Wald F statistics
Source of Variation            NumDF     DenDF     F-inc              P-inc
 tF2                             376              973.52

 Model_Term                      Order     Gamma         Sigma     Z_ratio  %C
 grm1(ID)                         9688  0.411124      0.195691       17.77   0 P
 Residual_units                   9688   1.00000      0.475990       55.18
  grm1(ID)                               9688 effects fitted.
 Finished: Tue Jul 14 15:41:27 2020LogL Converged      Gt4
```

Part 5 produced

```
Note: Model is fitting 10064 equations, DENSE portion has 376 equations.
 * This job may use 12 processor threads. *
    1 LogL= -46919.42   93.24          9312 DF
    2 LogL= -46898.31   90.68          9312 DF
    3 LogL= -46885.57   87.06          9312 DF
    4 LogL= -46885.46   86.67          9312 DF
    5 LogL= -46885.46   86.70          9312 DF

 Akaike Information Criterion   93774.92 (assuming 2 parameters).
 Bayesian Information Criterion 93789.20

          Analysis of tY5

                          Wald F statistics
Source of Variation            NumDF     DenDF     F-inc              P-inc
 tF3                             376              384.36

 Model_Term                      Order     Gamma         Sigma     Z_ratio  %C
 grm1(ID)                         9688  0.211970       18.3782       13.16   0 P
 Residual_units                   9688   1.00000       86.7018       57.59
  grm1(ID)                               9688 effects fitted.
 Finished: Tue Jul 14 15:42:36 2020LogL Converged      Gt5
```

These univariate results can be compared with a summary provided by Li

Trait	Model	Vgg	Vp	Vres	Vgt	Va	Vg1	Vg2
imf	A+G1+G2	0.14	0.7	0.28	0.42	0.15	0.15	0.12
sf5	A+G1+G2	1.98	106.49	80.37	26.13	8.36	3.39	14.37

My results SVD above (fitting just one genetic term at a time)

Trait	Vres	Vgt	Va	Vg1	Vg2
imf	0.25		0.46		
sf5	74.98		33.19		

imf	0.360			0.344	
sf5	82.64			24.82	
imf	0.477				0.196
sf5	86				18

Part 45 produced

```
fit grm1(ID)                          9688  0.411124      0.195691       17.77   0 P
* This job may use 12 processor threads. *
    1 LogL= -89484.65     18624 DF
    2 LogL= -85259.08     18624 DF
    3 LogL= -78103.09     18624 DF
    4 LogL= -75132.47     18624 DF
    5 LogL= -71611.42     18624 DF
    6 LogL= -70004.81     18624 DF
    7 LogL= -69811.52     18624 DF
    8 LogL= -69791.67     18624 DF
    9 LogL= -69791.60     18624 DF
   10 LogL= -69791.60     18624 DF

 Akaike Information Criterion   139595.20 (assuming 6 parameters).
 Bayesian Information Criterion 139642.19

          Analysis of tY4 tY5

                        Wald F statistics
Source of Variation            NumDF     DenDF     F-inc              P-inc
 Tr.tF3                          752              781.99

 Model_Term                      Order     Gamma         Sigma     Z_ratio  %C
 us(Tr).grm1(ID)                     19376 effects
 us(Tr)_V                  1   1   2  0.194099      0.194099       17.70   0 P
 us(Tr)_C                  2   1   2 -0.751916     -0.751916       -8.30   0 P
 us(Tr)_V                  2   2   2   18.0471       18.0471       13.07   0 P
 units.us(Trait)                     19376 effects
 us(Trait)_V                         2  0.476706      0.476706       55.18   0 P
 us(Trait)_C                         2 -0.946800     -0.946800      -11.58   0 P
 us(Trait)_V                         2   86.8833       86.8833       57.66   0 P

Covariance\Variance\Correlations for us(Tr) in us(Tr).grm1(ID)
  0.194099   -0.4017
 -0.751916 18.047076

Covariance\Variance\Correlations for us(Trait) in units.us(Trait)
  0.476706   -0.1471
 -0.946800 86.883325
  us(Tr).grm1(ID)                     19376 effects fitted.
 Warning: An updated US matrix was not positive definite!
```

Finished: Tue Jul 14 16:27:12 2020LogL Converged Gt45

This job took 4:22 mm:ss (26s per iteration). The bivariate results are consistent with the univariate results (which took 9 s for 5 iterations). There is a negative covariance between the traits.

Summary

Trait	Vres	Vgt	Va A22	Vg1 G_r	Vg2 G_t
Imf Imf xf5	0.255 -.318 75.88		0.453 -1.63 32.11		
Imf Imf xf5	0.361 -.677 82.95			0.343 -1.215 24.37	
Imf Imf xf5	0.477 -0.947 86.88				0.194 -.752 18.05

These components are consistent with Li's results given the low correlation between these traits and that having multiple terms spreads the variance among the terms.

My current concern is that when I fitted these models in the traditional way, the genetic variances were much smaller.

My previous results for the standard and mrm models assumed that ped.csv defined the order of genotypes in the ,bgrm files which differed from the order in the data.csv file. These SVD results assumed the data order and .bgrm orders were the same. Refitting the earlier models on this assumption generates the following results!

Univariate results from MRM analyses. LogL: imf -2685.55 sf5 -26638.86

Trait	Model	Vgg	Vp	Vres	Vgt	Va	Vg1	Vg2
imf	mrm1234	0.00	0.700	0.260	0.440	0.148	0.164	0.128
sf5	mrm1234	0.95	106.49	78.23	7.37	7.44	4.47	15.46

My results fitting just one genetic term at a time

Trait	LogL	time	Vres	Va	Vg1	Vg2
imf	-3043.13	9x10 *	0.252	0.456		
imf	-2862.89	7x19 **	0.360		0.344	
imf	-2839.34	6x2.2	0.477			0.196
sf5	-26832.2	2x2.5	74.98	33.19		
sf5	-26757.5	5x10.5	82.64		24.82	
sf5	-26651.6	5x11.5	86.70			18.38

- Oddly, the imf_Va run took 2.2 min for most iterations but #5 took 41m and #6 took 31 min.

- The imf_Vg1 run took 2.2 min for most iterations but #2 took 19m and #7 took 16 min. Rerunning the job for extra timing details, #5 took 950 sec (instead of 85) as if it was not using multiple threads!

Bivariate MRM refitting

Analysis of Ximf (= -10 x imf) and sf5

Common genetic effects

LogL -50,854.26 6 iterations took 10.4 min each (1.00 FOLD)

Mrm – covariance	8.51	8.81	9.93	0.80
Residual	Va	Vr	Vt	Vg
38.65 -7.72 83.46	8.51 8.51 8.51	8.81 8.81 8.81	9.93 9.93 9.93	0.80 0.80 0.80

The Vg component is probably not significant (Z-ratio 0.75)

Independent genetic effects

LogL -50,646.72 6 iterations took 32.8 min each (3:16:00) (1.56 FOLD)

Mrm – Ximf Mrm – sf5	14.33 7.29	15.59 3.85	12.07 14.49	0.18 0.86
Residual	Va	Vr	Vt	Vg
27.14 11.22 79.42	14.33 0.00 7.29	15.59 0.00 3.85	12.07 0.00 14.49	0.18 0.00 0.86

33 min per iteration is roughly 10.4 x 2 x 1.56

The Vg components are not significant (Z-rations 0.25 and 0.68).

Correlated

LogL -50599.11, iterations took 80 min each (2.25 FOLD)

The table reports parameter values from iteration 4 (see below).

Mrm – covariance Mrm – Ximf Mrm – sf5	3.889 10.620 2.881	2.701 13.738 2.375	6.126 6.529 8.824	0.973 -3.406 -0.155
Residual	Va	Vr	Vt	Vg
26.212 4.598 78.450	14.509 3.889 6.770	16.439 2.701 5.076	12.655 6.126 14.950	-2.433 0.973 0.818

Expected time: 10.4 x 3 x 2.25 = 70.2 min per iteration. Actual time was 80min.

The LogL at iteration 7 was -50,587 but then dropped down to -50,598 at iteration 11 before starting to rise again. I stopped the iteration at iteration 13. The residual variance remained reasonably stable but the specific variances are moving between sources as shown in the following table.

		Covariance				Residual		
Iteration	LogL	Va	Vr	Vt	Vg	σ^2_{11}	σ_{12}	σ^2_{22}
1	-50740.42	1.214	0.214	0.214	0.214	33.399	10.230	81.567
2	-50666.09	2.584	0.941	1.489	0.521	30.533	8.170	80.638
3	-50601.69	3.702	2.256	4.340	0.680	26.962	5.339	78.973
4	-50599.11	3.889	2.701	6.126	0.973	26.212	4.598	78.450
5	-50603.77	2.554	4.257	6.459	-0.090	25.418	4.610	78.845
6	-50600.89	3.707	3.076	6.192	2.032	25.983	4.569	78.649
7	-50578.62	0.020	4.983	6.635	1.480	24.950	6.065	76.801
8	-50582.80	0.709	4.331	6.635	1.719	25.313	5.800	77.269
9	-50591.34	1.008	4.222	6.576	1.734	25.441	5.678	77.214
10	-50596.27	1.588	3.991	6.390	1.362	25.809	5.488	76.225
11	-50598.02	1.909	4.287	6.333	-6.206	26.249	5.189	73.654
12	-50595.81	3.133	3.397	6.314	2.654	26.014	4.741	78.822
13	-50605.74	3.376	3.684	6.134	2.251	26.042	4.424	79.605
		3.400	3.935	6.128	1.970	26.077	4.281	80.004

Iteration	Ximf specific variance				Sf5 specific variance			
	Va	Vr	Vt	Vg	Va	Vr	Vt	Vg
1	13.214	12.214	7.214	13.214	13.214	0.214	3.214	0.214
2	11.698	12.979	7.499	3.616	9.797	0.738	4.389	-0.116
3	10.687	13.600	7.139	-2.604	4.650	1.702	7.084	-0.087
4	10.620	13.738	6.529	-3.406	2.881	2.375	8.824	-0.155
5	14.530	9.882	6.097	0.568	4.888	-0.230	8.531	1.534
6	11.086	13.380	6.402	-2.872	3.064	1.818	8.996	-2.843
7	17.085	10.335	5.767	-1.637	12.305	-3.195	7.482	-3.053
8	15.568	11.527	5.838	-1.609	11.562	-2.801	7.459	-3.356
9	15.002	11.732	5.921	-1.572	11.159	-2.533	7.553	-3.505
10	13.707	12.185	6.181	-1.086	11.070	-1.456	7.696	-4.545
11	12.490	12.177	6.321	9.226	15.204	-2.419	7.089	-8.450
12	11.693	13.022	6.251	-1.857	3.345	1.762	8.766	-8.676
13	11.471	12.597	6.521	-2.174	3.476	0.470	8.592	-10.367
	11.396	12.344	6.531	-1.732	3.925	-0.835	8.267	-8.885
Zratio	4.28	6.56	5.79	-1.17	1.41	-0.39	5.85	-12.83

It is evident the LogL at iteration 3/4 was better than at iteration 13. This indicates the model is over

parameterized. Va, Vr and Vg components are unstable. If we made the usual assumption that the matrices should be positive definite, we impose that restriction, which would affect the Vg matrix.

I reset the constraint on the specific variance to P (positive) and continued with !SLOW qualifier and the model converged to the following results (but !SLOW may not have been required).

LogL -50600.17

Mrm – covariance Mrm – Ximf Mrm – sf5	3.794 10.810 3.020	3.035 13.520 1.827	6.232 6.344 8.930	0.750 0.000 0.000
Residual	Va	Vr	Vt	Vg
26.094 4.486 78.665	14.604 3.794 6.814	16.555 3.035 4.862	12.576 6.232 15.162	0.75 0.75 0.75

Comments on SVD approach

The eigen analysis is the most demanding part of the SVD approach. The MKL routine requires about $5N^2$ cells of memory. The process ran much faster on my 32Gb (12m) machine than on my 16Gb machines (90m and 45m) on a N=9688. I do not understand why.

Time for transforming the data depends on the number of columns in the design matrix but was under a minute for 10000 records and 700 columns.

The limitations of this approach are that it only permits to 1 genomic relationship and the grm matrix ids must match the data file ids in order. If there are lots of traits, we can quickly perform univariate and bivariate analyses.

The fitted effects from the SVD are correct except for the genomic factor in that we have predicted Uu rather than u. So we need a procedure to premultiply by U' to get U'Uu = u.

Comments on MRM approach

The SCORE calculation is the most demanding part of the MRM approach.

It handles multiple GRM matrices simultaneously with little penalty for adding more. It is however not easily extended to bivariate analyses.

Things to do

1 Incorporate GEigen and Gtransform into Echidna.

2 Convert SVD blups back to correlated effects.

Appendix 1: gcta.log (part)

```
**********************************************************************
* Genome-wide Complex Trait Analysis (GCTA)
* version 1.92.3 beta3 Linux
* (C) 2010-2019, The University of Queensland
* Please report bugs to Jian Yang <jian.yang@uq.edu.au>
**********************************************************************
Analysis started at 13:17:33 UTC on Tue Apr 21 2020.
Hostname: agbusheep1.une.edu.au

Accepted options:
--reml
--reml-no-lrt
--reml-pred-rand
--reml-est-fix
--mgrm-bin raneff_gcta.txt
--pheno ../pheno.dat
--covar ../cg.dat
--out gcta
--threads 28

Note: the program will be running on 28 threads.

Reading phenotypes from [../pheno.dat].
Non-missing phenotypes of 10580 individuals are included from [../pheno.dat].
Reading discrete covariate(s) from [../cg.dat].
1 discrete covariate(s) of 10580 individuals are included from [../cg.dat].

There are 2 GRM file names specified in the file [raneff_gcta.txt].
Reading the GRM from the 1th file ...
Reading IDs of the GRM from [../gg.grm.id].
31572 IDs read from [../gg.grm.id].
Reading the GRM from [../gg.grm.bin].
GRM for 31572 individuals are included from [../gg.grm.bin].
Reading the GRM from the 2th file ...
Reading IDs of the GRM from [../grm1.grm.id].
31572 IDs read from [../grm1.grm.id].
Reading the GRM from [../grm1.grm.bin].
GRM for 31572 individuals are included from [../grm1.grm.bin].
1 discrete variable(s) included as covariate(s).
10580 individuals are in common in these files.

Performing  REML analysis ... (Note: may take hours depending on sample size).
10580 observations, 395 fixed effect(s), and 3 variance component(s)(including residual
variance).
Calculating prior values of variance components by EM-REML ...
Updated prior values: 0.310566 0.323125 0.339457
logL: -3244.63
Running AI-REML algorithm ...
Iter.   logL    V(G1)   V(G2)   V(e)
1       -3207.42        0.18851 0.31844         0.35191
2       -3199.73        0.13330 0.31502         0.36096
3       -3196.02        0.10269 0.31260         0.36741
4       -3194.19        0.08377 0.31092         0.37195
5       -3193.29        0.07119 0.30978         0.37511
6       -3192.83        0.04330 0.30737         0.38203
```

```
7       -3192.33       0.03664 0.30764       0.38203
8       -3192.31       0.03374 0.30773       0.38201
9       -3192.30       0.03228 0.30778       0.38200
10      -3192.30       0.03150 0.30780       0.38200
11      -3192.30       0.03107 0.30782       0.38199
12      -3192.30       0.03083 0.30783       0.38199
13      -3192.30       0.03069 0.30783       0.38199
14      -3192.30       0.03061 0.30783       0.38199
15      -3192.30       0.03057 0.30783       0.38199
Log-likelihood ratio converged.

Summary result of REML analysis:
Source  Variance       SE
V(G1)   0.030566       0.029906
V(G2)   0.307834       0.015155
V(e)    0.381989       0.010548
Vp      0.720389       0.031397
V(G1)/Vp       0.042430       0.039789
V(G2)/Vp       0.427317       0.024964

Sum of V(G)/Vp 0.469747       0.027118

Sampling variance/covariance of the estimates of variance components:
8.943660e-04   -1.990542e-05  3.979906e-06
-1.990542e-05  2.296609e-04   -1.088222e-04
3.979906e-06   -1.088222e-04  1.112606e-04
Estimatesof fixed effects:

Source  Estimate       SE
mean    3.227613       0.322666
X_2     1.354325       0.347135
X_3     0.557837       0.354659
...
X_395   0.886731       0.416784

Summary result of REML analysis has been saved in the file [gcta.hsq].

BLUP solutions of the genetic effects for 10580 individuals has been saved in the file
[gcta.indi.blp].

Analysis finished at 13:24:01 UTC on Tue Apr 21 2020
Overall computational time: 6 minutes 28 sec.
```