Frequently asked questions for the sommer package

Giovanny Covarrubias-Pazaran

2019-03-25

The sommer package was developed to provide R users a powerful and reliable multivariate mixed model solver. The package is focused in problems of the type p > n (more effects to estimate than observations) and its core algorithm is coded in C++ using the Armadillo library. This package allows the user to fit mixed models with the advantage of specifying the variance-covariance structure for the random effects, and specify heterogeneous variances, and obtain other parameters such as BLUPs, BLUEs, residuals, fitted values, variances for fixed and random effects, etc.

The purpose of this vignette is to provide answers to frequently asked questions (FAQ) related to performance and possible issues:

1) I got an error similar to:

```
# iteration    LogLik     wall    cpu(sec)   restrained
#     1      -224.676   18:11:23      3           0
# Sistem is singular. Stopping the job
# matrix multiplication: incompatible matrix dimensions: 0x0 and ...x...
```

This error indicates that your model is singular (phenotypic variance V matrix is not invertible) and therefore the model is stopped throwing the "incompatible matrix dimensions: 0x0 and ...x..." error message. Whether you can try a simpler model or just modify the argument `tolparinv` in the `mmer` function. The default is 1e-6, which means that it will try to invert V and if it fails it will try to add a small value to the diagonal of V of 1e-6 to make it invertible. If this fails then the program will stop returning that error message which should make you check the quality of your data or model attempted.

Sometimes the model becomes singular when you use variance covariance matrices (i.e. genomic relationship matrices) that are not full-rank. You can try to make it full-rank and try again.

2) My model runs very slow

Keep in mind that sommer uses direct inversion (DI) algorithm which can be very slow for large datasets. The package is focused in problems of the type p > n (more random effect levels than observations) and models with dense covariance structures. For example, for experiment with dense covariance structures with low-replication (i.e. 2000 records from 1000 individuals replicated twice with a covariance structure of 1000x1000) sommer will be faster than MME-based software. Also for genomic problems with large number of random effect levels, i.e. 300 individuals (n) with 100,000 genetic markers (p). For highly replicated trials with small covariance structures or n > p (i.e. 2000 records from 200 individuals replicated 10 times with covariance structure of 200x200) asreml or other MME-based algorithms will be much faster and we recommend you to opt for those software.

3) Can I run rrBLUP for markers and GBLUP for individuals in sommer

Both types of models can be fitted in sommer. The only thing that it changes is what is the random effect of interest; the marker matrix or the identifier for the individual.

```r
## rrBLUP for makers
data(DT_cpdata)
mix.rrblup <- mmer(fixed=cbind(color,Yield)~1,
                   random=~vs(GT,Gtc=unsm(2)) + vs(Rowf,Gtc=diag(2)),
                   rcov=~vs(units,Gtc=unsm(2)),
                   data=DT)
```

```
## iteration    LogLik     wall    cpu(sec)   restrained
##     1      -533.942   20:23:14      4           0
##     2      -373.864   20:23:18      8           0
##     3      -292.05   20:23:22      12           0
##     4      -259.206   20:23:25      15           0
##     5      -255.006   20:23:29      19           0
##     6      -254.802   20:23:33      23           0
##     7      -254.795   20:23:36      26           0
##     8      -254.794   20:23:40      30           0
```

```r
summary(mix.rrblup)
```

```
## ============================================================
##          Multivariate Linear Mixed Model fit by REML
## **********************  sommer 3.8  **********************
## ============================================================
##          logLik      AIC      BIC Method Converge
## Value -254.7943 513.5886 522.7526     NR     TRUE
## ============================================================
## Variance-Covariance components:
##                       VarComp VarCompSE  Zratio Constraint
## u:GT.color-color    4.183e-06 8.412e-07  4.9727   Positive
## u:GT.color-Yield    2.650e-04 3.458e-04  0.7663   Unconstr
## u:GT.Yield-Yield    5.904e-01 2.594e-01  2.2763   Positive
## u:Rowf.color-color  1.721e-04 1.232e-04  1.3974   Positive
## u:Rowf.Yield-Yield  8.340e+02 3.932e+02  2.1209   Positive
## u:units.color-color 2.464e-03 2.792e-04  8.8280   Positive
## u:units.color-Yield 3.812e-01 2.012e-01  1.8949   Unconstr
## u:units.Yield-Yield 3.239e+03 2.865e+02 11.3051   Positive
## ============================================================
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 color (Intercept)   0.1663   0.03875   4.291
## 2 Yield (Intercept) 132.4217  18.75134   7.062
## ============================================================
## Groups and observations:
##        color Yield
## u:GT    2889  2889
## u:Rowf    13    13
## ============================================================
## Use the '$' sign to access results and parameters
```

```r
## GBLUP for individuals
A <- A.mat(GT)
mix.gblup <- mmer(fixed=cbind(color,Yield)~1,
                  random=~vs(id,Gu=A, Gtc=unsm(2)) + vs(Rowf,Gtc=diag(2)),
                  rcov=~vs(units,Gtc=unsm(2)),
                  data=DT)
```

```
## iteration    LogLik     wall    cpu(sec)   restrained
##     1      -362.46   20:24:46      3           0
##     2      -289.256   20:24:50      7           0
##     3      -259.023   20:24:54      11           0
##     4      -254.901   20:24:57      14           0
##     5      -254.799   20:25:1      18           0
##     6      -254.794   20:25:5      22           0
##     7      -254.794   20:25:8      25           0
```

```
summary(mix.gblup)
```

```
## ============================================================
##          Multivariate Linear Mixed Model fit by REML
## **********************  sommer 3.8  **********************
## ============================================================
##          logLik      AIC      BIC Method Converge
## Value -254.7943 513.5885 522.7526     NR     TRUE
## ============================================================
## Variance-Covariance components:
##                       VarComp VarCompSE  Zratio Constraint
## u:id.color-color    4.918e-03 9.887e-04  4.9742   Positive
## u:id.color-Yield    3.120e-01 4.064e-01  0.7678   Unconstr
## u:id.Yield-Yield    6.940e+02 3.047e+02  2.2774   Positive
## u:Rowf.color-color  1.723e-04 1.235e-04  1.3954   Positive
## u:Rowf.Yield-Yield  8.339e+02 3.931e+02  2.1215   Positive
## u:units.color-color 2.464e-03 2.792e-04  8.8280   Positive
## u:units.color-Yield 3.811e-01 2.012e-01  1.8942   Unconstr
## u:units.Yield-Yield 3.239e+03 2.865e+02 11.3045   Positive
## ============================================================
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 color (Intercept)   0.1823  0.004489   40.60
## 2 Yield (Intercept) 132.3328  8.555778   15.47
## ============================================================
## Groups and observations:
##        color Yield
## u:id     363   363
## u:Rowf    13    13
## ============================================================
## Use the '$' sign to access results and parameters
```

Literature

Covarrubias-Pazaran G. 2016. Genome assisted prediction of quantitative traits using the R package sommer. PLoS ONE 11(6):1-15.

Covarrubias-Pazaran G. 2018. Software update: Moving the R package sommer to multivariate mixed models for genome-assisted prediction. doi: https://doi.org/10.1101/354639

Bernardo Rex. 2010. Breeding for quantitative traits in plants. Second edition. Stemma Press. 390 pp.

Gilmour et al. 1995. Average Information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51(4):1440-1450.

Henderson C.R. 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics vol. 31(2):423-447.

Kang et al. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723.

Lee, D.-J., Durban, M., and Eilers, P.H.C. (2013). Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. Computational Statistics and Data Analysis, 61, 22 - 37.

Lee et al. 2015. MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information. Cold Spring Harbor. doi: http://dx.doi.org/10.1101/027201.

Maier et al. 2015. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet; 96(2):283-294.

Rodriguez-Alvarez, Maria Xose, et al. Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spatial Statistics 23 (2018): 52-71.

Searle. 1993. Applying the EM algorithm to calculating ML and REML estimates of variance components. Paper invited for the 1993 American Statistical Association Meeting, San Francisco.

Yu et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Genetics 38:203-208.

Abdollahi Arpanahi R, Morota G, Valente BD, Kranis A, Rosa GJM, Gianola D. 2015. Assessment of bagging GBLUP for whole genome prediction of broiler chicken traits. Journal of Animal Breeding and Genetics 132:218-228.

Tunnicliffe W. 1989. On the use of marginal likelihood in time series model estimation. JRSS 51(1):15-27.