Quick start for the sommer package

Giovanny Covarrubias-Pazaran

2017-08-23

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 random effect levels than observations). 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 quick start guide is to show the flexibility of the package under certain common scenarios:

1) Univariate homogeneous variance models
2) Univariate heterogeneous variance models
3) Multivariate homogeneous variance models
4) Multivariate heterogeneous variance models

Background

The core of the package are the `mmer2` (formula-based) and `mmer` (matrix-based) functions which solve the mixed model equations. The functions are an interface to call the `NR` Direct-Inversion Newton-Raphson (Tunnicliffe 1989; Gilmour et al. 1995; Lee et al. 2015) or the `EMMA` efficient mixed model association algorithm (Kang et al. 2008).

Since version 2.0 sommer can handle multivariate models. These have the form:

$Y = X\beta + Zu + \epsilon$

with:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_t \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} X & ... & ... \\ ... & ... & ... \\ ... & ... & X \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} Z_1 G_1 Z_1' + ... + Z_1 R_1 Z_1' & ... & Z_1 H_1 Z_t' + ... + Z_1 S_1 Z_t' \\ ... & ... & ... \\ Z_t H_1 Z_1' + ... + Z_t S_1 Z_1' & ... & Z_t G_1 Z_t' + ... + Z_t R_1 Z_t' \end{bmatrix}$$

for 't' traits, where G are H are variance and covariance matrices among random effects for the "t" trait, and R and S are variance and covariance matrices among residuals. Here R=S=Iσ_ϵ , where I is an identity matrix. We can specify the covariance matrices. BLUPs will also be corrected for such covariances usually leading to more accurate predictions.

In the following section we will go on quick examples with the same dataset of corn hybrids tested in 4 different environments.

1) Univariate homogeneous variance models

This type of models refer to single response models where a variable of interest (i.e. genotypes) needs to be analized as interacting with a 2nd random effect (i.e. environments), but you assume that across environments the genotypes have the same variance component. This is the so-called compound simmetry (CS) model.

```r
library(sommer)
data(example)
head(example)
```

```
##                    Name     Env Loc Year     Block Yield    Weight
## 33  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.1     4 -1.904711
## 65          CO02024-9W CA.2013  CA 2013 CA.2013.1     5 -1.446958
## 66  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.2     5 -1.516271
## 67            MSL007-B CA.2011  CA 2011 CA.2011.2     5 -1.435510
## 68           MSR169-8Y CA.2013  CA 2013 CA.2013.1     5 -1.469051
## 103         AC05153-1W CA.2013  CA 2013 CA.2013.1     6 -1.307167
```

```r
ans1 <- mmer2(Yield~Env,
              random= ~ Name + Env:Name,
              rcov= ~ units,
              data=example, silent = TRUE)
summary(ans1)
```

```
## ====================================================
##     Multivariate Linear Mixed Model fit by REML
## ******************  sommer 3.0  ******************
## ====================================================
##          logLik      AIC      BIC Method Converge
## Value -20.14537 46.29075 55.95182    MNR     TRUE
## ====================================================
## Variance-Covariance components:
##                       VarComp VarCompSE Zratio
## Name.Yield-Yield        3.682     1.691  2.177
## Env:Name.Yield-Yield    5.173     1.495  3.459
## units.Yield-Yield       4.366     0.647  6.749
## ====================================================
## Fixed effects:
##
## $Yield
##              Estimate Std. Error   t value
## (Intercept) 16.496351  0.6855001 24.064695
## EnvCA.2012  -5.776759  0.7558178 -7.643057
## EnvCA.2013  -6.380479  0.7960514 -8.015159
##
## ====================================================
## Groups and observations:
##          Observ Groups
## Name        185     41
## Env:Name    185    123
## ====================================================
## Use the '$' sign to access results and parameters
```

2) Univariate heterogeneous variance models

Very often in multi-environment trials, the assumption that the genetic variance or the residual variance is the same across locations may be too naive. Because of that, specifying a general genetic component and a location specific genetic variance is the way to go. This require a CS+DIAG model.

```
data(example)
head(example)
```

```
##                    Name     Env Loc Year     Block Yield    Weight
## 33  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.1     4 -1.904711
## 65          CO02024-9W CA.2013  CA 2013 CA.2013.1     5 -1.446958
## 66  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.2     5 -1.516271
## 67            MSL007-B CA.2011  CA 2011 CA.2011.2     5 -1.435510
## 68           MSR169-8Y CA.2013  CA 2013 CA.2013.1     5 -1.469051
## 103         AC05153-1W CA.2013  CA 2013 CA.2013.1     6 -1.307167
```

```
ans1 <- mmer2(Yield~Env,
              random= ~Name + at(Env):Name,
              rcov= ~ at(Env):units,
              data=example, silent = TRUE)
summary(ans1)
```

```
## ===================================================
##      Multivariate Linear Mixed Model fit by REML
## ******************  sommer 3.0  ******************
## ===================================================
##          logLik      AIC      BIC Method Converge
## Value -15.42982 36.85964 46.52071    MNR     TRUE
## ===================================================
## Variance-Covariance components:
##                           VarComp VarCompSE Zratio
## Name.Yield-Yield            2.962    1.4963  1.980
## CA.2011:Name.Yield-Yield   10.148    4.5108  2.250
## CA.2012:Name.Yield-Yield    1.879    1.8699  1.005
## CA.2013:Name.Yield-Yield    6.629    2.5027  2.649
## CA.2013:units.Yield-Yield   2.560    0.6398  4.001
## CA.2011:units.Yield-Yield   4.943    1.5246  3.242
## CA.2012:units.Yield-Yield   5.725    1.3119  4.364
## ===================================================
## Fixed effects:
##
## $Yield
##              Estimate Std. Error   t value
## (Intercept) 16.507678  0.8268665 19.964138
## EnvCA.2012  -5.816890  0.8575814 -6.782902
## EnvCA.2013  -6.412433  0.9356490 -6.853460
##
## ===================================================
## Groups and observations:
##              Observ Groups
## Name            185     41
## CA.2011:Name    185     41
## CA.2012:Name    185     41
## CA.2013:Name    185     41
```

```
## ====================================================
## Use the '$' sign to access results and parameters
```

As you can see the special function `at` or `diag` can be used to indicate that there's a different variance for the genotypes in each environment. Same was done for the residual. The difference between `at` and `diag` is that the `at` function can be used to specify the levels or specific environments where the variance is different.

3) Multivariate homogeneous variance models

Currently there's a great push for multi-response models. This is motivated by the correlation that certain variables hide and that could benefit in the prediction perspective. In sommer to specify multivariate models the response requires the use of the `cbind()` function in the response, and the `us(trait)`, `diag(trait)`, or `at(trait)` functions in the random part of the model.

```
data(example)
head(example)
```

```
##                    Name     Env Loc Year     Block Yield    Weight
## 33  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.1     4 -1.904711
## 65          CO02024-9W CA.2013  CA 2013 CA.2013.1     5 -1.446958
## 66  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.2     5 -1.516271
## 67             MSL007-B CA.2011  CA 2011 CA.2011.2     5 -1.435510
## 68            MSR169-8Y CA.2013  CA 2013 CA.2013.1     5 -1.469051
## 103          AC05153-1W CA.2013  CA 2013 CA.2013.1     6 -1.307167
```

```
ans1 <- mmer2(cbind(Yield, Weight) ~ Env,
              random= ~ us(trait):Name + us(trait):Env:Name,
              rcov= ~ us(trait):units,
              data=example, silent = TRUE)
summary(ans1)
```

```
## ===================================================
##     Multivariate Linear Mixed Model fit by REML
## ******************  sommer 3.0  ******************
## ===================================================
##         logLik       AIC       BIC Method Converge
## Value 167.0252 -322.0505 -298.5695    MNR     TRUE
## ===================================================
## Variance-Covariance components:
##                         VarComp VarCompSE Zratio
## Name.Yield-Yield         3.7091   1.68159  2.206
## Name.Yield-Weight        0.9071   0.37954  2.390
## Name.Weight-Weight       0.2244   0.08777  2.556
## Env:Name.Yield-Yield     5.0922   1.47905  3.443
## Env:Name.Yield-Weight    1.0269   0.30773  3.337
## Env:Name.Weight-Weight   0.2101   0.06663  3.153
## units.Yield-Yield        4.3838   0.64953  6.749
## units.Yield-Weight       0.9078   0.14148  6.416
## units.Weight-Weight      0.2280   0.03378  6.750
## ===================================================
## Fixed effects:
##
## $Yield
##              Estimate Std. Error   t value
## (Intercept) 14.741985  0.6783206 21.733063
```

```
## EnvCA.2012  -3.199172  0.7474097 -4.280347
## EnvCA.2013  -4.003349  0.7850509 -5.099477
##
## $Weight
##               Estimate Std. Error   t value
## (Intercept)  0.5847374  0.1497090  3.905826
## EnvCA.2012  -0.9711517  0.1592564 -6.098038
## EnvCA.2013  -1.1643244  0.1681079 -6.926052
##
## ===================================================
## Groups and observations:
##          Observ Groups
## Name        185     41
## Env:Name    185    123
## ===================================================
## Use the '$' sign to access results and parameters
```

You may notice that we have added the `us(trait)` behind the random effects. This is to indicate the structure that should be assume in the multivariate model. The `diag(trait)` used behind a random effect (i.e. Name) indicates that for the traits modeled (Yield and Weight) there's no a covariance component and should not be estimated, whereas `us(trait)` assumes that for such random effect, there's a covariance component to be estimated (i.e. covariance between Yield and Weight for the random effect Name). Same applies for the residual part (rcov).

4) Multivariate heterogeneous variance models

This is just an extension of the univariate heterogeneous variance models but at the multivariate level. This would be a CS+DIAG multivariate model:

```
data(example)
head(example)
```

```
##                   Name     Env Loc Year     Block Yield    Weight
## 33  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.1     4 -1.904711
## 65          CO02024-9W CA.2013  CA 2013 CA.2013.1     5 -1.446958
## 66  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.2     5 -1.516271
## 67             MSL007-B CA.2011  CA 2011 CA.2011.2     5 -1.435510
## 68            MSR169-8Y CA.2013  CA 2013 CA.2013.1     5 -1.469051
## 103          AC05153-1W CA.2013  CA 2013 CA.2013.1     6 -1.307167
```

```
ans1 <- mmer2(cbind(Yield, Weight) ~ Env,
              random= ~ us(trait):Name + us(trait):at(Env):Name,
              rcov= ~ us(trait):at(Env):units,
              data=example, silent = TRUE)
summary(ans1)
```

```
## ======================================================
##       Multivariate Linear Mixed Model fit by REML
## ********************  sommer 3.0  ********************
## ======================================================
##         logLik       AIC       BIC Method Converge
## Value 177.8154 -343.6309 -320.1498    MNR     TRUE
## ======================================================
## Variance-Covariance components:
##                             VarComp VarCompSE Zratio
```

```
## Name.Yield-Yield           3.32291   1.45386 2.2856
## Name.Yield-Weight          0.79475   0.32648 2.4343
## Name.Weight-Weight         0.19103   0.07509 2.5442
## CA.2011:Name.Yield-Yield    8.69943   4.00992 2.1695
## CA.2011:Name.Yield-Weight   1.77753   0.83835 2.1203
## CA.2011:Name.Weight-Weight  0.35939   0.17885 2.0094
## CA.2012:Name.Yield-Yield    2.57327   1.95113 1.3189
## CA.2012:Name.Yield-Weight   0.33267   0.39866 0.8345
## CA.2012:Name.Weight-Weight  0.03842   0.08600 0.4467
## CA.2013:Name.Yield-Yield    5.46657   2.16184 2.5287
## CA.2013:Name.Yield-Weight   1.34662   0.50455 2.6689
## CA.2013:Name.Weight-Weight  0.32893   0.12203 2.6954
## CA.2013:units.Yield-Yield   2.56131   0.63996 4.0023
## CA.2013:units.Yield-Weight  0.44569   0.12645 3.5246
## CA.2013:units.Weight-Weight 0.12232   0.03057 4.0009
## CA.2011:units.Yield-Yield   4.93845   1.52314 3.2423
## CA.2011:units.Yield-Weight  0.99446   0.32150 3.0932
## CA.2011:units.Weight-Weight 0.23982   0.07394 3.2433
## CA.2012:units.Yield-Yield   5.73841   1.31504 4.3637
## CA.2012:units.Yield-Weight  1.27999   0.30150 4.2454
## CA.2012:units.Weight-Weight 0.31804   0.07285 4.3657
## =======================================================
## Fixed effects:
##
## $Yield
##              Estimate Std. Error   t value
## (Intercept) 14.498157  0.7889029 18.377621
## EnvCA.2012  -3.009537  0.8264035 -3.641728
## EnvCA.2013  -3.731629  0.8754507 -4.262524
##
## $Weight
##               Estimate Std. Error   t value
## (Intercept)  0.5746062  0.1682642  3.414905
## EnvCA.2012  -0.9334404  0.1697663 -5.498384
## EnvCA.2013  -1.1375574  0.1914161 -5.942851
##
## =======================================================
## Groups and observations:
##              Observ Groups
## Name            185     41
## CA.2011:Name    185     41
## CA.2012:Name    185     41
## CA.2013:Name    185     41
## =======================================================
## Use the '$' sign to access results and parameters
```

Any number of random effects can be specified with different structures.

5) Including special functions

Several random effects require the use of covariance structures that specify an special relationship among the levels of such random effect. The sommer package includes the `g()` function to include such known covariance structures:

```r
data(example)
head(example)
```

```
##                    Name    Env Loc Year    Block Yield    Weight
## 33  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.1     4 -1.904711
## 65          CO02024-9W CA.2013  CA 2013 CA.2013.1     5 -1.446958
## 66  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.2     5 -1.516271
## 67            MSL007-B CA.2011  CA 2011 CA.2011.2     5 -1.435510
## 68           MSR169-8Y CA.2013  CA 2013 CA.2013.1     5 -1.469051
## 103         AC05153-1W CA.2013  CA 2013 CA.2013.1     6 -1.307167
```

```r
K[1:4,1:4]
```

```
##                    Manistee(MSL292-A) CO02024-9W MSL007-B MSR169-8Y
## Manistee(MSL292-A)                  1          0        0         0
## CO02024-9W                          0          1        0         0
## MSL007-B                            0          0        1         0
## MSR169-8Y                           0          0        0         1
```

```r
ans1 <- mmer2(Yield ~ Env,
              random= ~ g(Name) + at(Env):g(Name),
              rcov= ~ at(Env):units,
              G=list(Name=K),
              data=example, silent = TRUE)
summary(ans1)
```

```
## ====================================================
##       Multivariate Linear Mixed Model fit by REML
## ********************  sommer 3.0  ********************
## ====================================================
##          logLik      AIC      BIC Method Converge
## Value -15.42982 36.85964 46.52071    MNR     TRUE
## ====================================================
## Variance-Covariance components:
##                              VarComp VarCompSE Zratio
## g(Name).Yield-Yield            2.962    1.4963  1.980
## CA.2011:g(Name).Yield-Yield   10.148    4.5108  2.250
## CA.2012:g(Name).Yield-Yield    1.879    1.8699  1.005
## CA.2013:g(Name).Yield-Yield    6.629    2.5027  2.649
## CA.2013:units.Yield-Yield      2.560    0.6398  4.001
## CA.2011:units.Yield-Yield      4.943    1.5246  3.242
## CA.2012:units.Yield-Yield      5.725    1.3119  4.364
## ====================================================
## Fixed effects:
##
## $Yield
##              Estimate Std. Error   t value
## (Intercept) 16.507678  0.8268665 19.964138
## EnvCA.2012  -5.816890  0.8575814 -6.782902
## EnvCA.2013  -6.412433  0.9356490 -6.853460
##
## ====================================================
## Groups and observations:
##                Observ Groups
## g(Name)           185     41
```

```
## CA.2011:g(Name)    185     41
## CA.2012:g(Name)    185     41
## CA.2013:g(Name)    185     41
## =====================================================
## Use the '$' sign to access results and parameters
```

and for multivariate models:

```
data(example)
head(example)
```

```
##                     Name     Env Loc Year     Block Yield    Weight
## 33  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.1     4 -1.904711
## 65          CO02024-9W CA.2013  CA 2013 CA.2013.1     5 -1.446958
## 66  Manistee(MSL292-A) CA.2013  CA 2013 CA.2013.2     5 -1.516271
## 67            MSL007-B CA.2011  CA 2011 CA.2011.2     5 -1.435510
## 68           MSR169-8Y CA.2013  CA 2013 CA.2013.1     5 -1.469051
## 103         AC05153-1W CA.2013  CA 2013 CA.2013.1     6 -1.307167
```

```
K[1:4,1:4]
```

```
##                      Manistee(MSL292-A) CO02024-9W MSL007-B MSR169-8Y
## Manistee(MSL292-A)                    1          0        0         0
## CO02024-9W                            0          1        0         0
## MSL007-B                              0          0        1         0
## MSR169-8Y                             0          0        0         1
```

```
ans1 <- mmer2(cbind(Yield, Weight) ~ Env,
              random= ~ us(trait):g(Name) + us(trait):at(Env):g(Name),
              rcov= ~ us(trait):at(Env):units,
              G=list(Name=K),
              data=example, silent = TRUE)
summary(ans1)
```

```
## =======================================================
##        Multivariate Linear Mixed Model fit by REML
## ********************  sommer 3.0  ********************
## =======================================================
##          logLik       AIC       BIC Method Converge
## Value 177.8154 -343.6309 -320.1498    MNR     TRUE
## =======================================================
## Variance-Covariance components:
##                               VarComp VarCompSE Zratio
## g(Name).Yield-Yield           3.32291   1.45386 2.2856
## g(Name).Yield-Weight          0.79475   0.32648 2.4343
## g(Name).Weight-Weight         0.19103   0.07509 2.5442
## CA.2011:g(Name).Yield-Yield   8.69943   4.00992 2.1695
## CA.2011:g(Name).Yield-Weight  1.77753   0.83835 2.1203
## CA.2011:g(Name).Weight-Weight 0.35939   0.17885 2.0094
## CA.2012:g(Name).Yield-Yield   2.57327   1.95113 1.3189
## CA.2012:g(Name).Yield-Weight  0.33267   0.39866 0.8345
## CA.2012:g(Name).Weight-Weight 0.03842   0.08600 0.4467
## CA.2013:g(Name).Yield-Yield   5.46657   2.16184 2.5287
## CA.2013:g(Name).Yield-Weight  1.34662   0.50455 2.6689
## CA.2013:g(Name).Weight-Weight 0.32893   0.12203 2.6954
## CA.2013:units.Yield-Yield     2.56131   0.63996 4.0023
## CA.2013:units.Yield-Weight    0.44569   0.12645 3.5246
```

```
## CA.2013:units.Weight-Weight   0.12232   0.03057 4.0009
## CA.2011:units.Yield-Yield     4.93845   1.52314 3.2423
## CA.2011:units.Yield-Weight    0.99446   0.32150 3.0932
## CA.2011:units.Weight-Weight   0.23982   0.07394 3.2433
## CA.2012:units.Yield-Yield     5.73841   1.31504 4.3637
## CA.2012:units.Yield-Weight    1.27999   0.30150 4.2454
## CA.2012:units.Weight-Weight   0.31804   0.07285 4.3657
## =========================================================
## Fixed effects:
##
## $Yield
##              Estimate Std. Error   t value
## (Intercept) 14.498157  0.7889029 18.377621
## EnvCA.2012  -3.009537  0.8264035 -3.641728
## EnvCA.2013  -3.731629  0.8754507 -4.262524
##
## $Weight
##               Estimate Std. Error   t value
## (Intercept)  0.5746062  0.1682642  3.414905
## EnvCA.2012  -0.9334404  0.1697663 -5.498384
## EnvCA.2013  -1.1375574  0.1914161 -5.942851
##
## =========================================================
## Groups and observations:
##                 Observ Groups
## g(Name)            185     41
## CA.2011:g(Name)    185     41
## CA.2012:g(Name)    185     41
## CA.2013:g(Name)    185     41
## =========================================================
## Use the '$' sign to access results and parameters
```

Notice that the `g()` function is applied at the random effect called "Name", and the covariance structure is provided in the argument "G". In the example, we used a diagonal covariance structure for demonstration purposes but any dense covariance matrix can be used.

Other special functions such as `and()` for overlay models, `eig()` for an eigen decomposition of the covariance matrix, `grp()` for customized random effects providing an incidence matrix.

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.

Literature

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

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 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.

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.