

Fitting genotype by environment models in sommer

Giovanny Covarrubias-Pazaran

2021-12-14

The sommer package was developed to provide R users a powerful and reliable multivariate mixed model solver. The package is focused on 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, specifying heterogeneous variances, and obtaining other parameters such as BLUPs, BLUEs, residuals, fitted values, variances for fixed and random effects, etc.

The purpose of this vignette is to show how to fit different genotype by environment (GxE) models using the sommer package:

- 1) Single environment model
- 2) Multienvironment model: Main effect model
- 3) Multienvironment model: Diagonal model (DG)
- 4) Multienvironment model: Compund symmetry model (CS)
- 5) Multienvironment model: Compund symmetry + diagonal model (CS+DG)
- 6) Multienvironment model: Unstructured model (US)
- 7) Multienvironment model: Random regression model (RR)
- 8) Multienvironment model: Other covariance structures for GxE

When the breeder decides to run a trial and apply selection in a single environment (whether because the amount of seed is a limitation or there's no availability for a location) the breeder takes the risk of selecting material for a target population of environments (TPEs) using an environment that is not representative of the larger TPE. Therefore, many breeding programs try to base their selection decision on multi-environment trial (MET) data. Models could be adjusted by adding additional information like spatial information, experimental design information, etc. In this tutorial we will focus mainly on the covariance structures for GxE and the incorporation of relationship matrices for the genotype effect.

1) Single environment model

A single-environment model is the one that is fitted when the breeding program can only afford one location, leaving out the possible information available from other environments. This will be used to further expand to GxE models.

```
library(sommer)
data(DT_example)
DT <- DT_example
A <- A_example

ansSingle <- mmer(Yield~1,
                 random= ~ vs(Name, Gu=A),
                 rcov= ~ units,
                 data=DT, verbose = FALSE)
summary(ansSingle)
```

```

## =====
##           Multivariate Linear Mixed Model fit by REML
## ***** sommer 4.1 *****
## =====
##           logLik      AIC      BIC Method Converge
## Value -78.80875 159.6175 162.8378      NR      TRUE
## =====
## Variance-Covariance components:
##           VarComp VarCompSE Zratio Constraint
## u:Name.Yield-Yield  6.529      2.202  2.965  Positive
## units.Yield-Yield  13.868      1.633  8.494  Positive
## =====
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept)      11.74      0.4876  24.07
## =====
## Groups and observations:
##           Yield
## u:Name      41
## =====
## Use the '$' sign to access results and parameters

```

In this model, the only term to be estimated is the one for the germplasm (here called `Name`). For the sake of example we have added a relationship matrix among the levels of the random effect `Name`. This is just a diagonal matrix with as many rows and columns as levels present in the random effect `Name`, but any other non-diagonal relationship matrix could be used.

2) MET: main effect model

A multi-environment model is the one that is fitted when the breeding program can afford more than one location. The main effect model assumes that GxE doesn't exist and that the main genotype effect plus the fixed effect for environment is enough to predict the genotype effect in all locations of interest.

```

ansMain <- mmer(Yield~Env,
               random= ~ vs(Name, Gu=A),
               rcov= ~ units,
               data=DT, verbose = FALSE)
summary(ansMain)

```

```

## =====
##           Multivariate Linear Mixed Model fit by REML
## ***** sommer 4.1 *****
## =====
##           logLik      AIC      BIC Method Converge
## Value -32.59421 71.18842 80.84949      NR      TRUE
## =====
## Variance-Covariance components:
##           VarComp VarCompSE Zratio Constraint
## u:Name.Yield-Yield  4.856      1.5233  3.188  Positive
## units.Yield-Yield  8.109      0.9615  8.434  Positive
## =====
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept)      16.385      0.5849  28.012

```

```
## 2 Yield EnvCA.2012 -5.688 0.5741 -9.908
## 3 Yield EnvCA.2013 -6.218 0.6107 -10.182
## =====
## Groups and observations:
##      Yield
## u:Name  41
## =====
## Use the '$' sign to access results and parameters
```

3) MET: diagonal model (DG)

A multi-environment model is the one that is fitted when the breeding program can afford more than one location. The diagonal model assumes that GxE exists and that the genotype variation is expressed differently at each location, therefore fitting a variance component for the genotype effect at each location. The main drawback is that this model assumes no covariance among locations, as if genotypes were independent (despite the fact that is the same genotypes). The fixed effect for environment plus the location-specific BLUP is used to predict the genotype effect in each locations of interest.

```
ansDG <- mmer(Yield~Env,
              random= ~ vs(ds(Env),Name, Gu=A),
              rcov= ~ units,
              data=DT, verbose = FALSE)
summary(ansDG)
```

```
## =====
##      Multivariate Linear Mixed Model fit by REML
## ***** sommer 4.1 *****
## =====
##      logLik      AIC      BIC Method Converge
## Value -21.04157 48.08315 57.74421      NR      TRUE
## =====
## Variance-Covariance components:
##      VarComp VarCompSE Zratio Constraint
## CA.2011:Name.Yield-Yield 17.493 6.1099 2.863 Positive
## CA.2012:Name.Yield-Yield 5.337 1.7662 3.022 Positive
## CA.2013:Name.Yield-Yield 7.884 2.5526 3.089 Positive
## units.Yield-Yield 4.381 0.6493 6.747 Positive
## =====
## Fixed effects:
##      Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept) 16.621 0.948 17.532
## 2 Yield EnvCA.2012 -5.958 1.045 -5.699
## 3 Yield EnvCA.2013 -6.662 1.098 -6.067
## =====
## Groups and observations:
##      Yield
## CA.2011:Name 41
## CA.2012:Name 41
## CA.2013:Name 41
## =====
## Use the '$' sign to access results and parameters
```

4) MET: compound symmetry model (CS)

A multi-environment model is the one that is fitted when the breeding program can afford more than one location. The compound symmetry model assumes that GxE exists and that a main genotype variance-covariance component is expressed across all location. In addition, it assumes that a main genotype-by-environment variance is expressed across all locations. The main drawback is that the model assumes the same variance and covariance among locations. The fixed effect for environment plus the main effect for BLUP plus genotype-by-environment effect is used to predict the genotype effect in each location of interest.

```
E <- diag(length(unique(DT$Env)))
rownames(E) <- colnames(E) <- unique(DT$Env)
EA <- kronecker(E,A, make.dimnames = TRUE)
ansCS <- mmer(Yield~Env,
              random= ~ vs(Name, Gu=A) + vs(Env:Name, Gu=EA),
              rcov= ~ units,
              data=DT, verbose = FALSE)
summary(ansCS)
```

```
## =====
##           Multivariate Linear Mixed Model fit by REML
## ***** sommer 4.1 *****
## =====
##           logLik      AIC      BIC Method Converge
## Value -20.14538 46.29075 55.95182      NR      TRUE
## =====
## Variance-Covariance components:
##           VarComp VarCompSE Zratio Constraint
## u:Name.Yield-Yield      3.682      1.691 2.177 Positive
## u:Env:Name.Yield-Yield  5.173      1.495 3.460 Positive
## units.Yield-Yield      4.366      0.647 6.748 Positive
## =====
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept)  16.496      0.6855 24.065
## 2 Yield EnvCA.2012   -5.777      0.7558 -7.643
## 3 Yield EnvCA.2013   -6.380      0.7960 -8.015
## =====
## Groups and observations:
##           Yield
## u:Name      41
## u:Env:Name  123
## =====
## Use the '$' sign to access results and parameters
```

5) MET: compound symmetry plus diagonal (CS+DIAG)

A multi-environment model is the one that is fitted when the breeding program can afford more than one location. The compound symmetry plus diagonal model is very similar to the CS model assuming that GxE exist and that a main genotype variance-covariance component is expressed across all location. In addition, it assumes that a location-specific genotype-by-environment variance is expressed in each location. The main drawback is that the model assumes the same covariance among locations. The fixed effect for environment plus the main effect BLUP plus environment-specific genotype-by-environment effect is used to predict the genotype effect in each location of interest.

```
ansMain <- mmer(Yield~Env,
               random= ~ vs(Name, Gu=A) + vs(ds(Env),Name, Gu=A),
               rcov= ~ units,
               data=DT, verbose = FALSE)
summary(ansMain)
```

```
## =====
##           Multivariate Linear Mixed Model fit by REML
## ***** sommer 4.1 *****
## =====
##           logLik      AIC      BIC Method Converge
## Value -18.16164 42.32327 51.98434      NR      TRUE
## =====
## Variance-Covariance components:
##           VarComp VarCompSE Zratio Constraint
## u:Name.Yield-Yield      2.965   1.5055  1.969   Positive
## CA.2011:Name.Yield-Yield 10.424   4.4544  2.340   Positive
## CA.2012:Name.Yield-Yield  2.658   1.8032  1.474   Positive
## CA.2013:Name.Yield-Yield  5.702   2.5113  2.271   Positive
## units.Yield-Yield       4.398   0.6517  6.748   Positive
## =====
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept)  16.511    0.8269  19.967
## 2 Yield EnvCA.2012   -5.809    0.8593  -6.760
## 3 Yield EnvCA.2013   -6.423    0.9358  -6.864
## =====
## Groups and observations:
##           Yield
## u:Name      41
## CA.2011:Name 41
## CA.2012:Name 41
## CA.2013:Name 41
## =====
## Use the '$' sign to access results and parameters
```

6) MET: unstructured model (US)

A multi-environment model is the one that is fitted when the breeding program can afford more than one location. The unstructured model is the most flexible model assuming that Gx_E exists and that an environment-specific variance exists in addition to as many covariances for each environment-to-environment combinations. The main drawback is that is difficult to make this models converge because of the large number of variance components, the fact that some of these variance or covariance components are zero, and the difficulty in choosing good starting values. The fixed effect for environment plus the environment specific BLUP (adjusted by covariances) is used to predict the genotype effect in each location of interest.

```
ansUS <- mmer(Yield~Env,
              random= ~ vs(us(Env),Name, Gu=A),
              rcov= ~ units,
              data=DT, verbose = FALSE)
summary(ansUS)
```

```
## =====
##           Multivariate Linear Mixed Model fit by REML
```

```
## ***** sommer 4.1 *****
## =====
##          logLik      AIC      BIC Method Converge
## Value -14.20951 34.41901 44.08008      NR      TRUE
## =====
## Variance-Covariance components:
##
##          VarComp VarCompSE Zratio Constraint
## CA.2011:Name.Yield-Yield      15.994      5.381 2.972 Positive
## CA.2012:CA.2011:Name.Yield-Yield      6.172      2.503 2.465 Unconstr
## CA.2012:Name.Yield-Yield      5.273      1.750 3.013 Positive
## CA.2013:CA.2011:Name.Yield-Yield      6.366      3.069 2.074 Unconstr
## CA.2013:CA.2012:Name.Yield-Yield      0.376      1.535 0.245 Unconstr
## CA.2013:Name.Yield-Yield      7.689      2.490 3.088 Positive
## units.Yield-Yield      4.386      0.650 6.748 Positive
## =====
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept)      16.341      0.8141 20.072
## 2 Yield EnvCA.2012      -5.696      0.7406 -7.692
## 3 Yield EnvCA.2013      -6.286      0.8202 -7.664
## =====
## Groups and observations:
##          Yield
## CA.2011:Name      41
## CA.2012:CA.2011:Name      82
## CA.2012:Name      41
## CA.2013:CA.2011:Name      82
## CA.2013:CA.2012:Name      82
## CA.2013:Name      41
## =====
## Use the '$' sign to access results and parameters
# adjust variance BLUPs by adding covariances
# ansUS$U[1:6] <- unsBLUP(ansUS$U[1:6])
```

7) MET: random regression model

A multi-environment model is the one that is fitted when the breeding program can afford more than one location. The random regression model assumes that the environment can be seen as a continuous variable and therefore a variance component for the intercept and a variance component for the slope can be fitted. The number of variance components will depend on the order of the Legendre polynomial fitted.

```
library(orthopolynom)
DT$EnvN <- as.numeric(as.factor(DT$Env))
ansRR <- mmer(Yield~Env,
              random= ~ vs(leg(EnvN,1),Name),
              rcov= ~ units,
              data=DT, verbose = FALSE)
summary(ansRR)
```

```
## =====
##          Multivariate Linear Mixed Model fit by REML
## ***** sommer 4.1 *****
## =====
```

```

##           logLik      AIC      BIC Method Converge
## Value -27.70318 61.40636 71.06743      NR      TRUE
## =====
## Variance-Covariance components:
##           VarComp VarCompSE Zratio Constraint
## leg0:Name.Yield-Yield 10.392  3.1473  3.302  Positive
## leg1:Name.Yield-Yield  2.079  0.9792  2.123  Positive
## units.Yield-Yield     6.297  0.8442  7.459  Positive
## =====
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept) 16.541    0.6770 24.432
## 2 Yield EnvCA.2012  -5.832    0.6425 -9.078
## 3 Yield EnvCA.2013  -6.472    0.8239 -7.854
## =====
## Groups and observations:
##           Yield
## leg0:Name  41
## leg1:Name  41
## =====
## Use the '$' sign to access results and parameters

```

In addition, an unstructured, diagonal or other variance-covariance structure can be put on top of the polynomial model:

```

library(orthopolynom)
DT$EnvN <- as.numeric(as.factor(DT$Env))
ansRR <- mmer(Yield~Env,
              random= ~ vs(us(leg(EnvN,1)),Name),
              rcov= ~ units,
              data=DT, verbose = FALSE)
summary(ansRR)

```

```

## =====
##           Multivariate Linear Mixed Model fit by REML
## ***** sommer 4.1 *****
## =====
##           logLik      AIC      BIC Method Converge
## Value -25.56967 57.13935 66.80042      NR      TRUE
## =====
## Variance-Covariance components:
##           VarComp VarCompSE Zratio Constraint
## leg0:Name.Yield-Yield 10.791  3.2745  3.295  Positive
## leg1:leg0:Name.Yield-Yield -2.428  1.3699 -1.772  Unconstr
## leg1:Name.Yield-Yield  2.286  1.0404  2.197  Positive
## units.Yield-Yield     6.260  0.8421  7.434  Positive
## =====
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept) 16.501    0.7778 21.216
## 2 Yield EnvCA.2012  -5.791    0.6704 -8.638
## 3 Yield EnvCA.2013  -6.476    0.8554 -7.570
## =====
## Groups and observations:
##           Yield

```

```
## leg0:Name      41
## leg1:leg0:Name 82
## leg1:Name      41
## =====
## Use the '$' sign to access results and parameters
```

8) Other GxE covariance structures

Although not very commonly used in GxE models, the autoregressive of order 1 (AR1) and other covariance structures could be used in the GxE modeling. Here we show how to do it (not recommending it).

```
E <- AR1(DT$Env) # can be AR1() or CS(), etc.
rownames(E) <- colnames(E) <- unique(DT$Env)
EA <- kronecker(E,A, make.dimnames = TRUE)
ansCS <- mmer(Yield~Env,
              random= ~ vs(Name, Gu=A) + vs(Env:Name, Gu=EA),
              rcov= ~ units,
              data=DT, verbose = FALSE)
summary(ansCS)
```

```
## =====
##           Multivariate Linear Mixed Model fit by REML
## ***** sommer 4.1 *****
## =====
##           logLik      AIC      BIC Method Converge
## Value -19.39067 44.78134 54.4424      NR      TRUE
## =====
## Variance-Covariance components:
##           VarComp VarCompSE Zratio Constraint
## u:Name.Yield-Yield      2.225   1.7536  1.269   Positive
## u:Env:Name.Yield-Yield  6.424   1.8293  3.512   Positive
## units.Yield-Yield      4.334   0.6418  6.752   Positive
## =====
## Fixed effects:
##   Trait      Effect Estimate Std.Error t.value
## 1 Yield (Intercept)  16.484    0.6735  24.474
## 2 Yield  EnvCA.2012   -5.780    0.7365  -7.848
## 3 Yield  EnvCA.2013   -6.372    0.7799  -8.170
## =====
## Groups and observations:
##           Yield
## u:Name      41
## u:Env:Name  123
## =====
## Use the '$' sign to access results and parameters
```

Final remarks

Keep in mind that sommer uses the direct inversion (DI) algorithm which can be very slow for large datasets. The package is focused on problems of the type $p > n$ (more random effect levels than observations) and models with dense covariance structures. For example, for experiments 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 number of individuals and 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 opt for those.

Literature

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

Covarrubias-Pazarán 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.

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