A Comparison of Linear Mixed Model Packages in R for Analysis of Plant Breeding Experiments

Sam Rogers and Julian Taylor

Biometry Hub, University of Adelaide

Introduction

The problem

We wanted to compare `ASReml-R` to other linear mixed model software used for the purposes of Plant Breeding.

- Are there other **mature** packages with the capabilities of `ASReml-R` (or close enough)?
- Are there open-source alternatives to `ASReml-R`?
- How do they compare in terms of computational performance (speed) and features?

We compared three R packages with a focus on computational aspects of large scale genomic selection LMMs.

Motivating Data

- Phenotypic data from an Australia Grains Technologies (AGT) field trial in SA.
- Full trial was 10375 varieties unreplicated + 6 check varieties with 150+ reps (see Norman, et al. (2017)).
- Genotypic data consisting of 17171 genetic markers from a 20K Affymetrix array, reduced to 3000/6000 markers for this analysis.
- We looked at rectangular subsets of the field trial containing n_g varieties, equivalent to $n_g \approx (100, 200, 500, 1000, 2000, 3000, 4000, 5000)$, and analysed zadoks (plant maturity) score trait.

The Packages

`lme4` (+ `pedigreemm`)

- Maturity: Mature (first released ~2003)
- Citations: ~25k, ~1.2k in plant breeding journals
- Availability: Open source (free)
- Latest version: CRAN: March 2019, Github: yesterday, `pedigreemm`: 2014
- Relationship matrices: `pedigreemm` package required (plus substantial hacks 👎)
- Residual correlation structures: None
- Variance structures for random effects: Limited

`ASReml-R`

- Maturity: Mature (first released ~1995)
- Citations: ~4.5k, higher proportion in plant breeding journals than `lme4`
- Availability: Closed source (licence fee 👎)
- Latest version: August 2019
- Relationship matrices: Built in
- Residual correlation structures: Extensive
- Variance structures for random effects: Extensive

`sommer`

- Maturity: New (Released 2016)
- Citations: ~90, mostly plant breeding journals
- Availability: Open source (free)
- Latest version: CRAN: November 2019, Github: October 2019
- Relationship matrices: Built in
- Residual correlation structures: Some
- Variance structures for random effects: Some

Genomic Selection LMM

For a trait response vector $\boldsymbol{y}$ of length n, consider a genetic marker matrix $\boldsymbol{M}$ (dimension $g \times r$) for varieties and the associated (mostly) additive relationship matrix $\boldsymbol{G}_a = \boldsymbol{M}\boldsymbol{M}^T$.

One specification of a LMM has the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Z}_g\boldsymbol{a} + \boldsymbol{Z}_g\boldsymbol{r} + \boldsymbol{e}$$

where

- $\boldsymbol{\tau}$ are fixed effects
- $\boldsymbol{u}$ are the random effects (None in this case)
- $\boldsymbol{a}$ is a set of additive variety effects with assumed distribution $\boldsymbol{a} \sim N(\boldsymbol{0}, \sigma_a^2\boldsymbol{G}_a)$
- $\boldsymbol{r}$ are the polygenic non-additive genetic effects with assumed distribution $\boldsymbol{r} \sim N(\boldsymbol{0}, \sigma_r^2\boldsymbol{I}_g)$.
- $\boldsymbol{e}$ are residuals with assumed distribution $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{R})$

`ASReml-R` and `sommer` use this form with the `vm()` and `vs()` functions respectively.

Genomic Selection LMM V2

An alternative specification of the LMM uses Cholesky decomposition of the relationship matrix defined as $\boldsymbol{G}_a = \boldsymbol{L}\boldsymbol{L}^T$.

The left Cholesky factor is incorporated into the LMM as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{Z}_g\boldsymbol{L}\boldsymbol{a}^* + \boldsymbol{Z}_g\boldsymbol{p} + \boldsymbol{e}$$

where

- $\boldsymbol{a}^*$ is a set of non-interpretable genetic effects with assumed distribution $\boldsymbol{a}^* \sim N(\boldsymbol{0}, \sigma_a^2\boldsymbol{I}_g)$ which is the same additive genetic variance as in the previous model.
  - The additive genetic effects, $\boldsymbol{a}$, can be derived through back transformation.

Although they differ in their specification, the two models produce the same likelihood and estimate of σ_a^2.

This form can be fitted in `pedigreemm()` and also in `ASReml-R` using the `mbf()` function.

Genomic Selection Models in R

To specify these models in R, we use the following code:

`LME4`/`pedigreemm`*

```
pedigreemm(matzadoks ~ 1 + (1|GenotypeA) + (1|GenotypeR),
           data = temp.dat, pedigree = list(GenotypeA =
  relmat_p))
```

* Needed substantial hacking

`ASReml-R`

```
asreml(matzadoks ~ 1, random = ~ mbf("relM") + GenotypeR,
       data = temp.dat, mbf = list(relM = list(
       key = c("GenotypeA", "GenotypeA"), cov = "lchol")))
```

```
asreml(matzadoks ~ 1, random = ~ vm(GenotypeA, relmat) +
           GenotypeR, data = temp.dat)
```

`sommer`

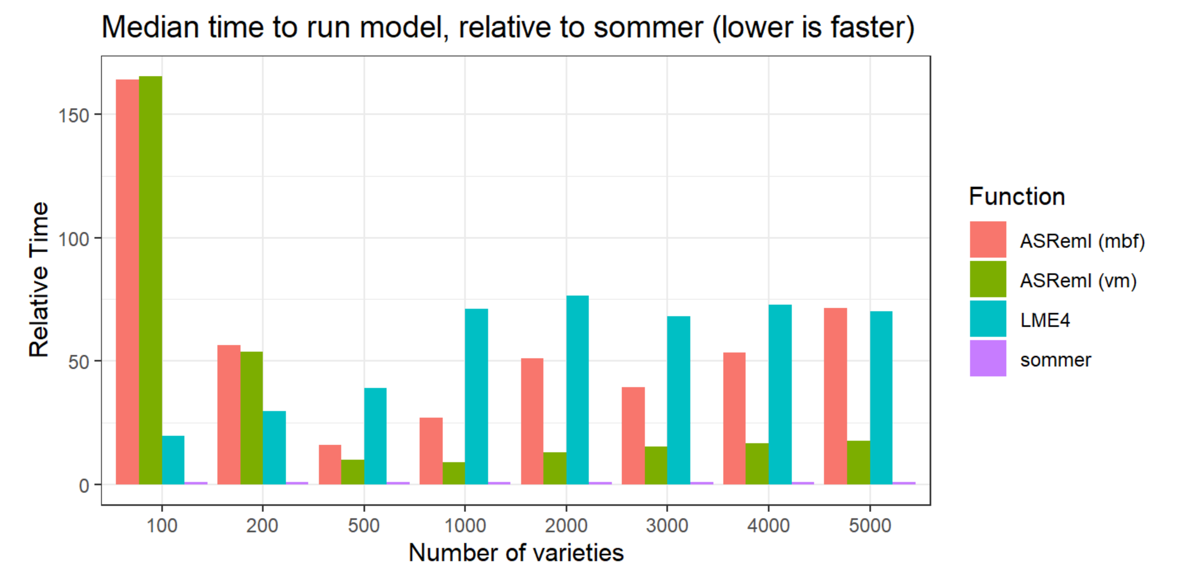
```
mmer(matzadoks ~ 1, random = ~vs(GenotypeA, Gu = relmat) +
         GenotypeR, rcov = ~units, data = temp.dat,
      na.method.X = "include")
```

Results of timings

- `Sommer` is computationally superior for all population sizes

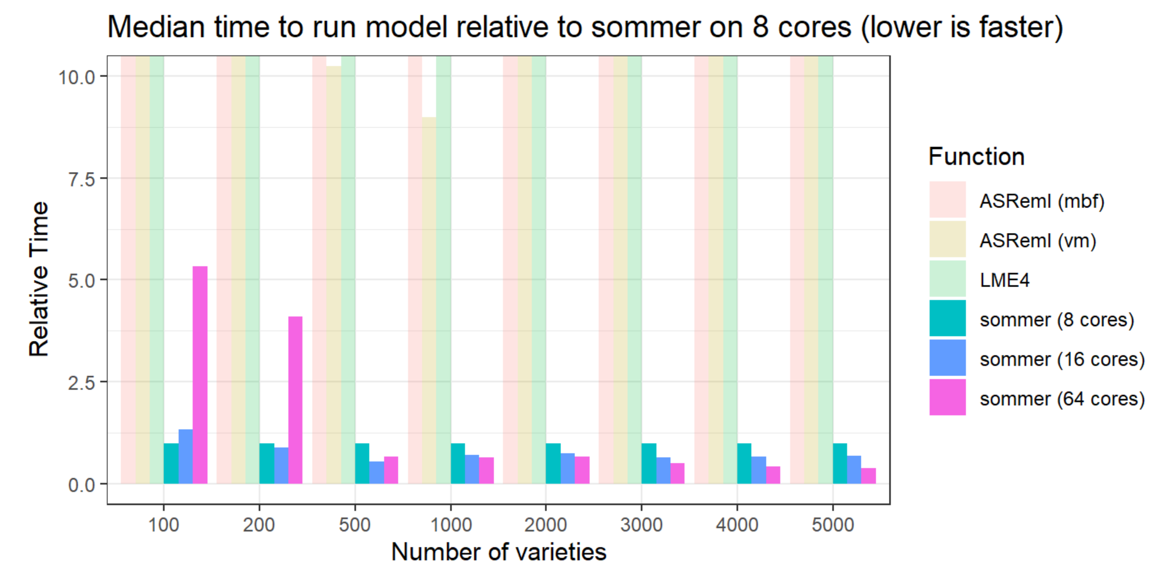
Median Sommer times (s)

100	200	500	1000	2000	3000	4000	5000
0.03	0.10	0.74	2.32	18.56	72.57	170.06	323.55



What's different about `sommer`?

It makes use of the Intel Math Kernel Libraries for parallel processing of matrix manipulations!



Summary and discoveries

`sommer` was the suprising winner by a huge margin

- At least 9x faster than `ASReml-R` and `LME4`, and up to 160x faster in some cases
- Performance increases with CPU cores available due to parallel processing
- Has much of the same capability as `ASReml-R`, though is lacking in a few areas

It is *possible* to run genomic selection models in `lme4`, but less than ideal because:

- `lme4`/`pedigreemm` is substantially slower
- Major disadvantage of no residual correlation structures available
- Relationship matrix incorporation took substantial hacking
- `pedigreemm` not updated since 2014

References

Bates, D. et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: Journal of Statistical Software 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bates, D. et al. (2007). "The lme4 package". In: R package version 2.1, p. 74.

Butler, D. G. et al. (2018). ASReml-R reference manual (version 4). School of Mathematics and Applied Statistics, University of Wollongong.

Covarrubias-Pazaran, G. (2016). "Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer". Eng. In: PloS one 11 (6).

Covarrubias-Pazaran, G. (2018). "Software update: Moving the R package sommer to multivariate mixed models for genome-assisted prediction". In: bioRxiv. DOI: 10.1101/354639. URL: https://www.biorxiv.org/content/early/2018/06/25/354639.

Norman, A. et al. (2017). "Increased genomic prediction accuracy in wheat breeding using a large Australian panel". In: Theoretical and applied genetics 130.12, pp. 2543–2555.

Vazquez, A. et al. (2010). "an R package for fitting generalized linear mixed models in animal breeding". In: Journal of animal science 88.2, pp. 497–504.