Example code: Robust Genomic Modelling Using Expert Knowledge about Additive, Dominance and Epistasis Variation

Ingeborg Gullikstad Hem (ingeborg.hem@ntnu.no), Maria Lie Selle, Gregor Gorjanc, Geir-Arne Fuglstad and Andrea Riebler

Introduction

We provide data and code for year 21 in one simulation from the wheat breeding program used in the simulation study of the paper. All the code included in this document can also be found in the R-file "run_me.R".

Setup

We must load the necessary functions and packages into R:

```
source("necessary_functions.R")

library(rstan)
library(nloptr)
library(ggplot2)
library(reshape2)
```

We also want to pre-compile the Stan-models:

```
stan_mod1 <- stan_model(file = "stan_models/wheat.stan", auto_write = TRUE)
stan_mod2 <- stan_model(file = "stan_models/wheat_ad.stan", auto_write = TRUE)
stan_mod3 <- stan_model(file = "stan_models/wheat_simple.stan", auto_write = TRUE)
```

Fitting the model

First, you choose which model you want to fit. You choose by selecting a number between 1 and 12, which corresponds to the following models:

```
##                  Model Number
## 1               A-comp      1
## 2              A-comp*      2
## 3               A-tree      3
## 4              A-tree*      4
## 5                 A-ML      5
## 6               AD-comp      6
## 7              AD-comp*      7
## 8               AD-tree      8
## 9              AD-tree*      9
## 10                AD-ML     10
## 11             ADX-comp     11
## 12            ADX-comp*     12
```

```
## 13             ADX-tree     13
## 14            ADX-tree*     14
## 15         ADX-tree-opp     15
## 16                ADX-ML     16
## 17 Phenotype selection     17
```

Then you can use the function `fit_model` to fit the model of your liking. You can add preferences on how many samples to use. We recommend "low" only for testing if the sampler works, "medium" [default] for getting an idea on how the results look (but these results may be incomplete or wrong due to the relatively low number of samples), and "high" in other cases. Note that "high" may take some time, especially for Model ADX.

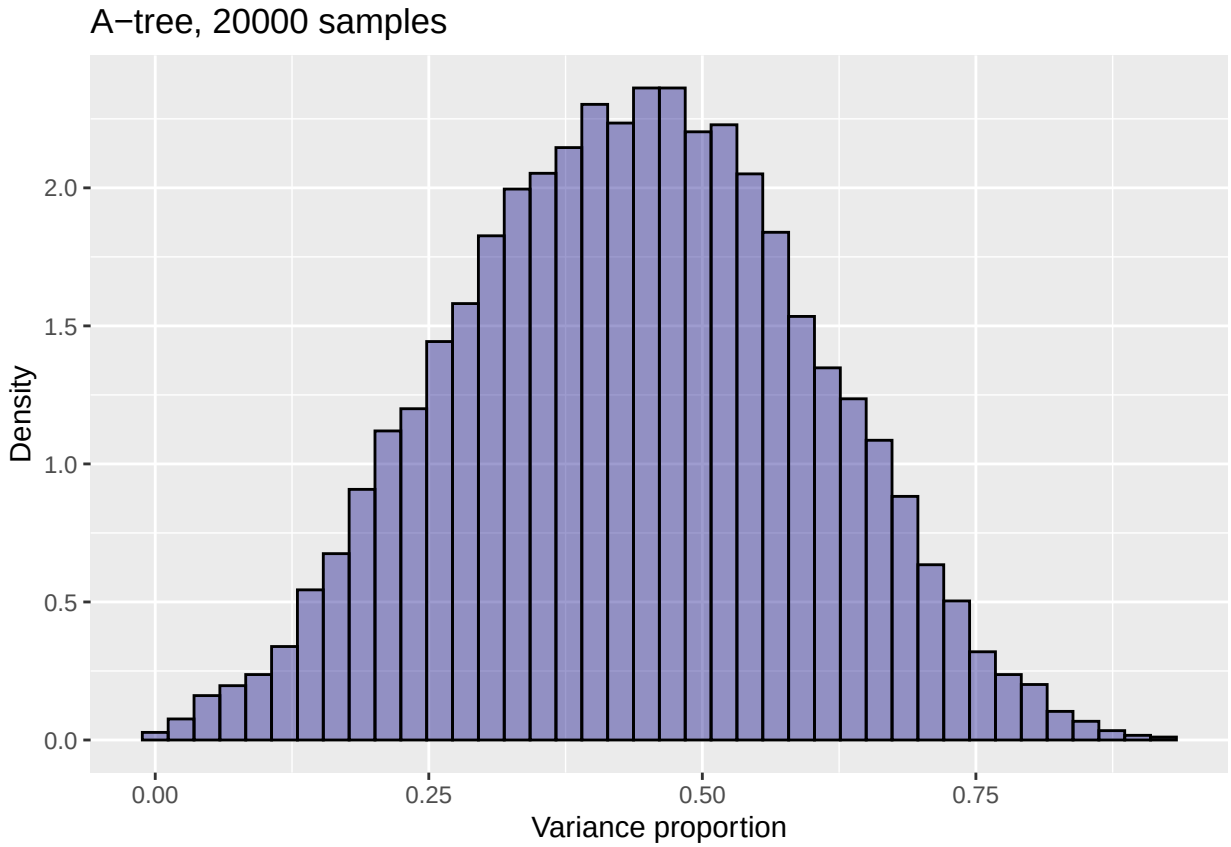
We choose the posterior of A-tree (which is model number 3) with a high number of samples as an example. The function prints info about the model fitting process, and if the sampler had any problems in terms of divergent transitions or a too low number of effective samples.

```
source("necessary_functions.R")
result <- fit_model(3, no_of_samps = "high")
```

```
##
##
## Fitting A-tree, which is model number 3
##
##
## SAMPLING FOR MODEL 'wheat_simple' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 9.9e-05 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0.99 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:     1 / 40000 [  0%]  (Warmup)
## Chain 1: Iteration:  4000 / 40000 [ 10%]  (Warmup)
## Chain 1: Iteration:  8000 / 40000 [ 20%]  (Warmup)
## Chain 1: Iteration: 12000 / 40000 [ 30%]  (Warmup)
## Chain 1: Iteration: 16000 / 40000 [ 40%]  (Warmup)
## Chain 1: Iteration: 20000 / 40000 [ 50%]  (Warmup)
## Chain 1: Iteration: 20001 / 40000 [ 50%]  (Sampling)
## Chain 1: Iteration: 24000 / 40000 [ 60%]  (Sampling)
## Chain 1: Iteration: 28000 / 40000 [ 70%]  (Sampling)
## Chain 1: Iteration: 32000 / 40000 [ 80%]  (Sampling)
## Chain 1: Iteration: 36000 / 40000 [ 90%]  (Sampling)
## Chain 1: Iteration: 40000 / 40000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 18.9046 seconds (Warm-up)
## Chain 1:                12.0542 seconds (Sampling)
## Chain 1:                30.9588 seconds (Total)
## Chain 1:
```

We want to look at the posterior of the variance proportions, and how good the model is at choosing the best individuals based on the total genetic and the additive effects. We report the number of the true 10 best individuals that are among the estimated top 10 individuals, and where the top 10 estimated individuals are ranked in the true ranking. We can only display the posterior for the Bayesian models (not for maximum likelihood). Note that these results are not necessarily reliable for inference with "low" or "medium" amount of samples, or if the sampler has experienced problems. From the results in the paper, we know that some of the models lead to unstable inference, and the results may be unreliable.

```
plot_variance_proportions(result)
```




```
find_ranking(result)
```

```
##
## A-tree
## ---------------------
## Ranking of genetic effects: 1, 4, 13, 20, 30, 31, 33, 54, 58, 69, which is 2 out of 10
## Ranking of additive effects: 2, 17, 19, 21, 25, 31, 35, 48, 63, 67, which is 1 out of 10
```