Meeting with Dom 20210217

Lena Morrill

February 17, 2021

1 Links

- Link to paper
- Link to github repo
- Script simulating data under the model

2 State of the inference

Convergence of PCAWG samples in full-RE multinomial model

- The vast majority of samples have not converged (see Figure **??**). This also happens with simulated datasets, sometimes. (`Hessian of fixed effects was not positive definite.`)
- This is especially the case for signatures (as opposed to the six nucleotide substitutions)
- signatures are, in general, more than 6, so if the problem is that there are too many categories and too few samples to converge properly, this would explain it
- The problem in inference is that the gradient is too steep. However, despite being the fixed effects covariance matrix being non-pd, the simulated data from these models (with the inferred parameters) show a very good correlation with the observed data (could it be that there is a problem with identifiability meaning that parameters are highly correlated?)
- Sometimes changing the initial conditions is enough to have a good convergence
- Another possibility is that signatures with very low abundance are the problem, and that we should remove those prior to analysing the exposures
- Another possibility is that the last category (signature), which is the one used as baseline, is very low in certain cancer types, and this doesn't allow good inference
- Using Head-SCC with signatures as an example, I am showing that using a subset of the exposures leads to correct, non-pd, inference, when using a full-RE multinomial
- Using simulated data, I see that even if convergence isn't good, the parameters are quite well recovered

3 My questions for Dom

1. What does this do exactly. My params are not well recovered unless the `opt <- do.call("optim", obj)` part is called, but `rep <- sdreport(obj)` does not take any arguments that have to do with this!

```
obj <- MakeADFun(data = TMB_data_sim, parameters = TMB_params_sim,
                 DLL="tmb_MVN_partial_ILR", random = "u_large")
opt <- do.call("optim", obj)
opt
opt$hessian ## <-- FD hessian from optim
rep <- sdreport(obj)
rep
```

2. If we're integrating over the random effects, doesn't that mean that the `log_sd`, cov from the RE should not be inferred? what exactly are the RE values that TMB spits out? are they the values from the best fit? check that way I integrate out the multivariate random effects is done correctly

3. DM vs categorical. I was wondering how individual mutations add to the likelihood. In DM or M, samples with more mutations will have a lower weight in the neighbourhood because the multinomial coefficient increases rapidly with the number of draws. However, what I had thought initially is that samples with a larger number of mutations should contribute *more* to the likelihood than samples with fewer mutations (and this scenario corresponds to the categorical distribution, as I understand it). Is there an overdispersed version of the categorical (analogous to the Dirichlet-Multinomial)? Link to the script for the multinomial fit and the categorical fit.

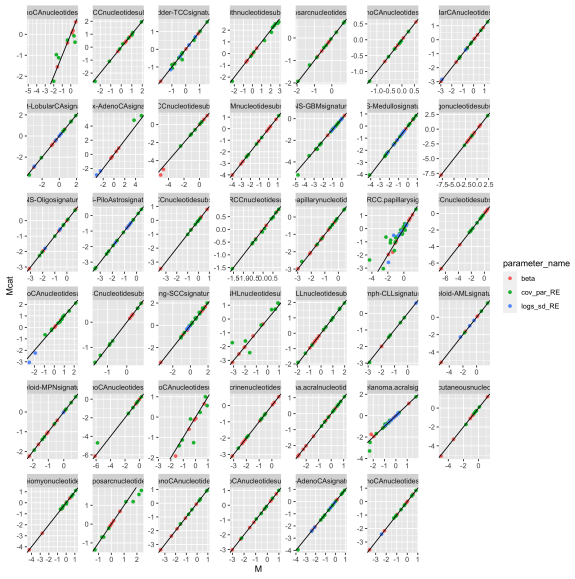

Figure 1: The estimates for the multinomial fit and the categorical fit are extremely similar, even though the number of mutations per sample varies quite a lot, which I would expect to be reflected in the estimates

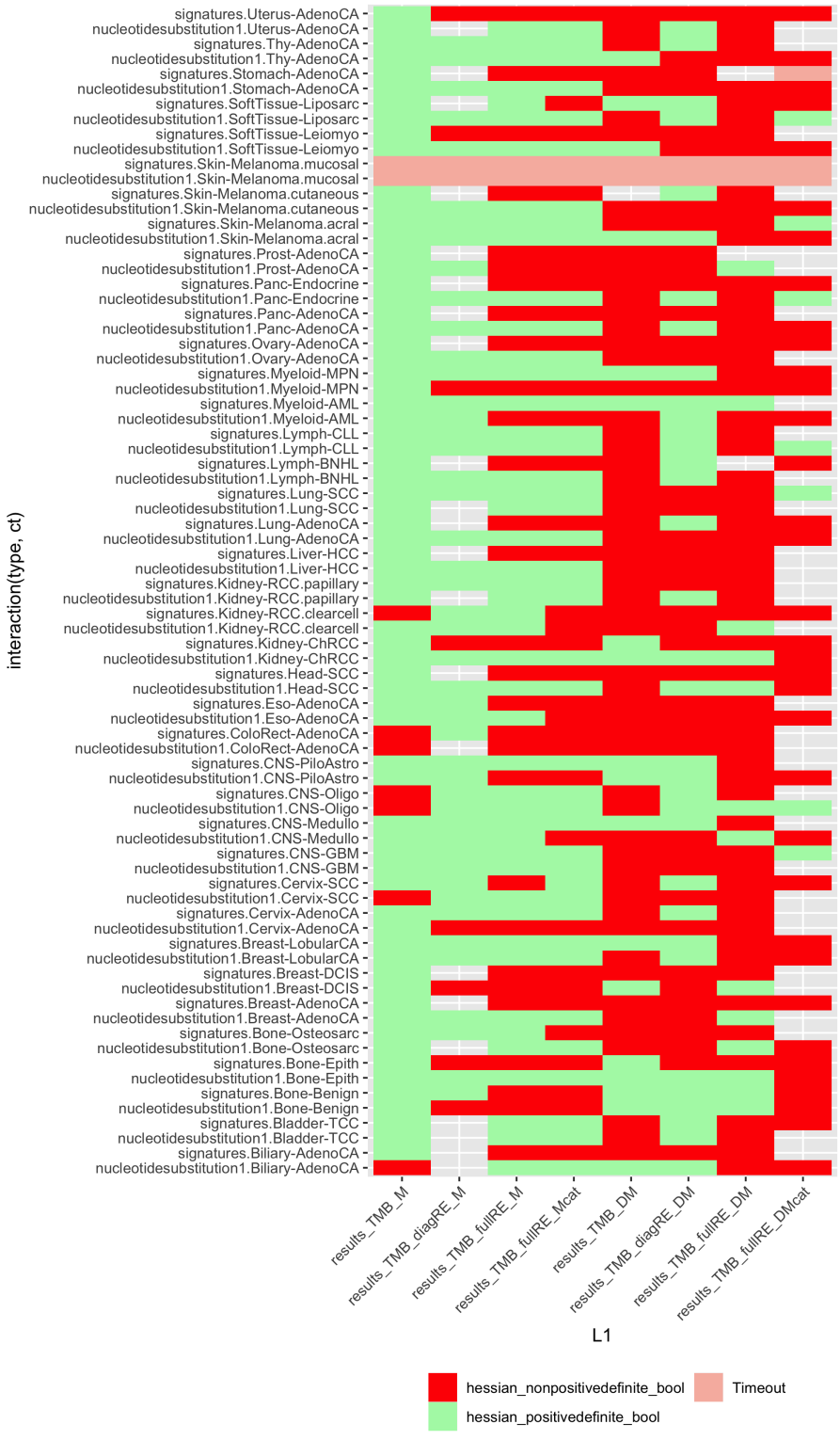

Figure 2: Convergence of PCAWG samples

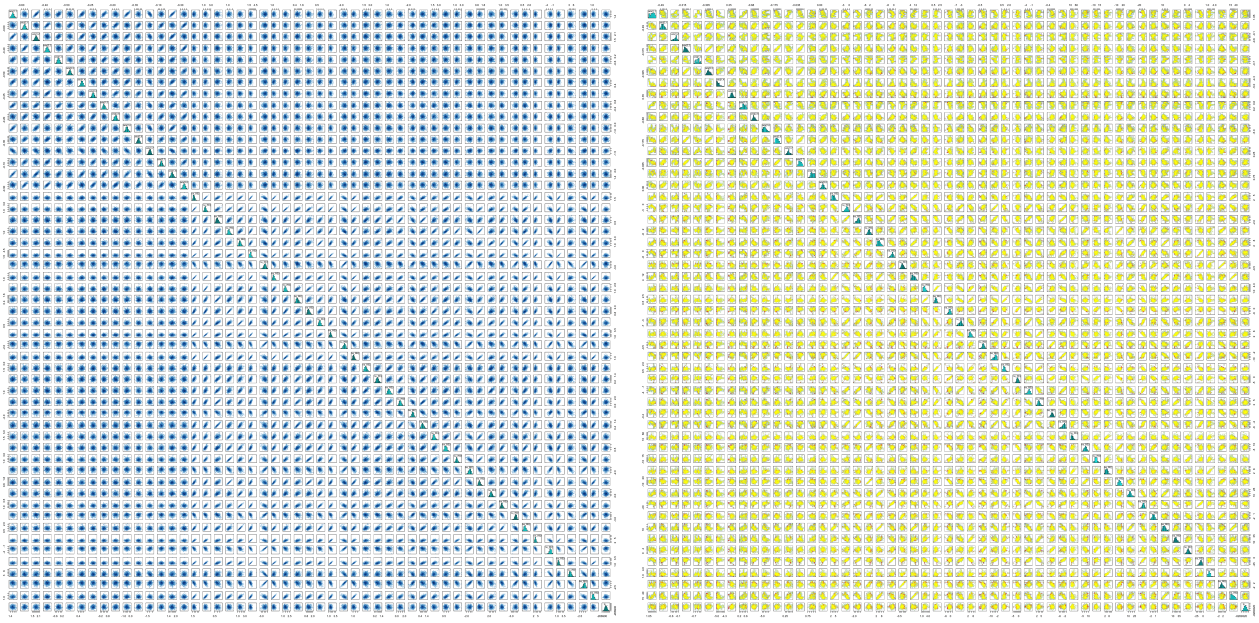

Figure 3: Pairs plot of the posteriors for the fullRE M for a sample that has converged with TMB (left; CNS-Medullo signatures) and one that hasn't (right; ColoRect-AdenoCA signatures). Yellow dots indicate transitions that hit the maximum treedepth.