Subgroup detection in GLMMs and GAMs

glmertrees, splinetrees and gamtrees

Marjolein Fokkema

Partitioning penalized or smoothing splines

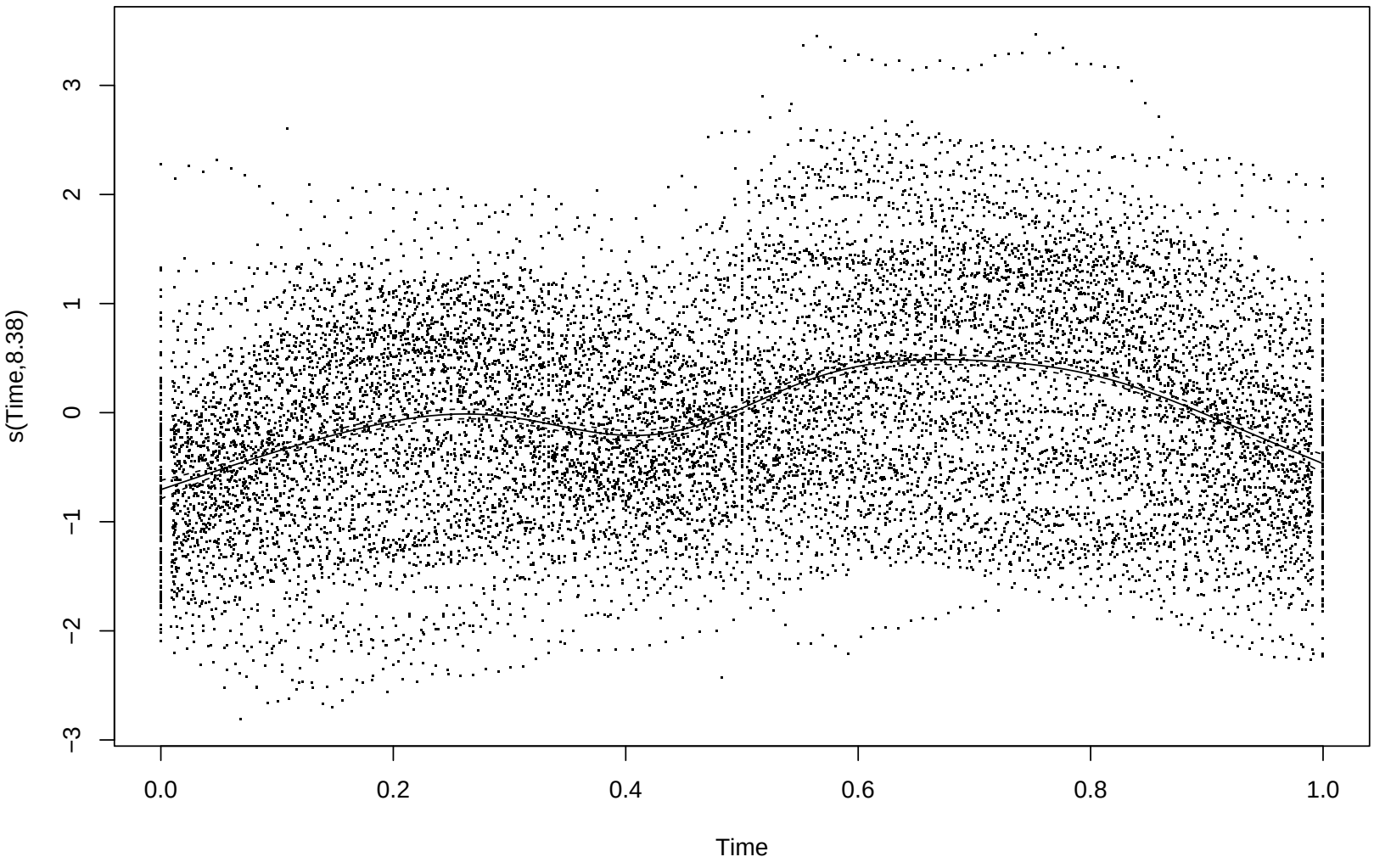
- Can we also partition **semi-parametric** splines, as fitted e.g., with `mgcv`?
  - `gamm4` (Wood & Schepl, 2020) for fitting GAMs
  - `merDeriv` (Wang & Markle, 2018) for extracting scores
    `partykit` (Hothorn & Zeileis, 2) for partitioning.
- Computational load is *very* heavy.

Traditional or baseline GAM

```
library("mgcv")
gamod <- gam(Pos ~ s(Time), data = dat) ## fit model
plot(gamod, rsiduals = TRUE) ## plot
sumary(gamod) ## print hypothesis tests
```

Output on next slides:

Traditional or baseline GAM



Traditional or baseline GAM

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Pos ~ s(Time)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.40356    0.00825   48.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Time) 8.381  8.893 170.6  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.106   Deviance explained = 10.6%
## GCV = 0.87455  Scale est. = 0.87391   n = 12839
```

Fit a smoothing spline tree

Need to reduce dataset size for feasible computation (adjust `minsize` for nodes in accordance, but results likely identical if left out).

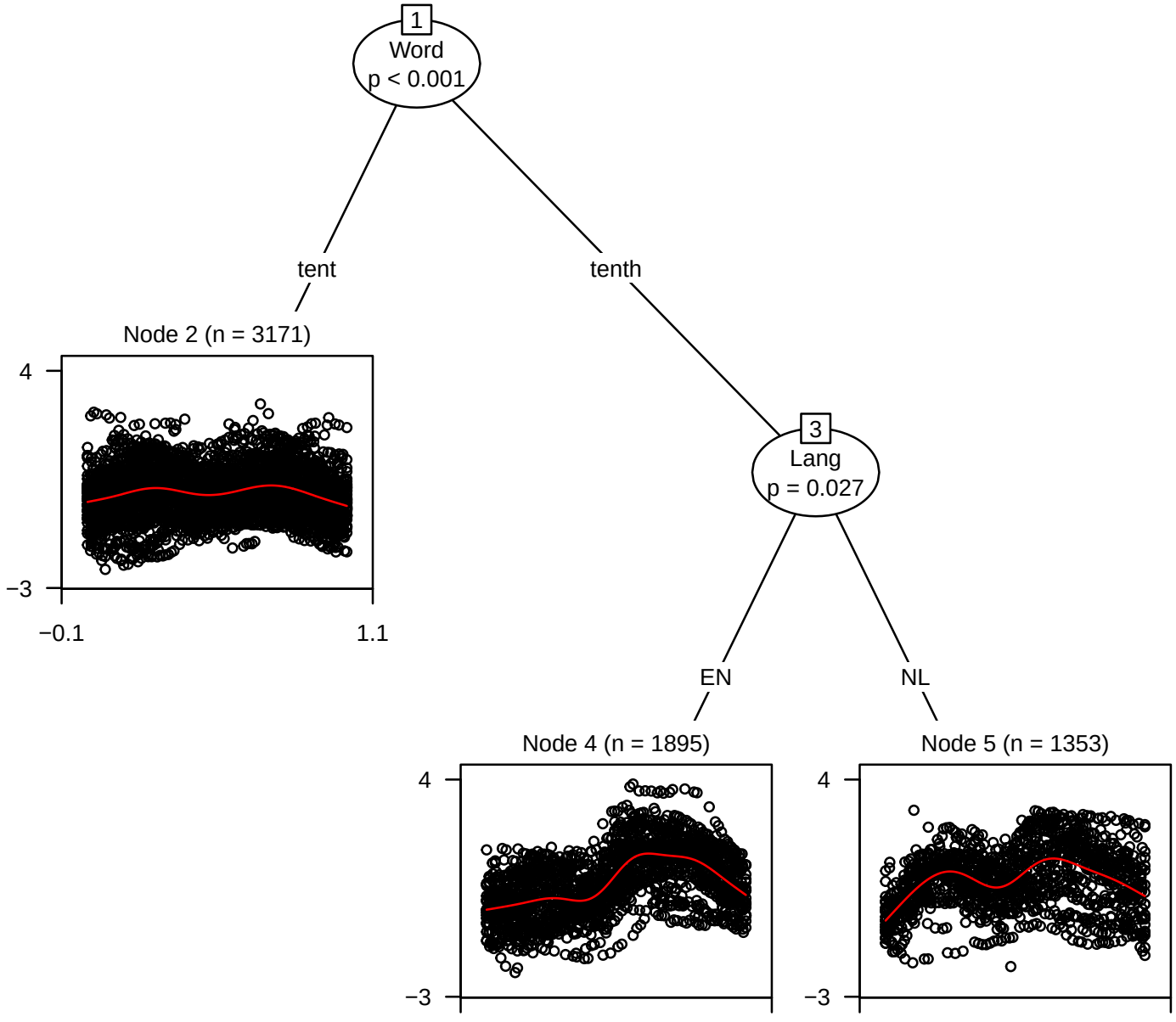
```
set.seed(42)
train <- sample(nrow(dat), size = nrow(dat)/2)
test <- (1:nrow(dat))[-train]
gt <- gamtree(Pos ~ s(Time) | Word + Lang + xt + xs,
              data = dat[train, ],
              tree_ctrl = list(minsize = (78/2)*10),
              cluster = Trial)
```

Fit a smoothing spline trees

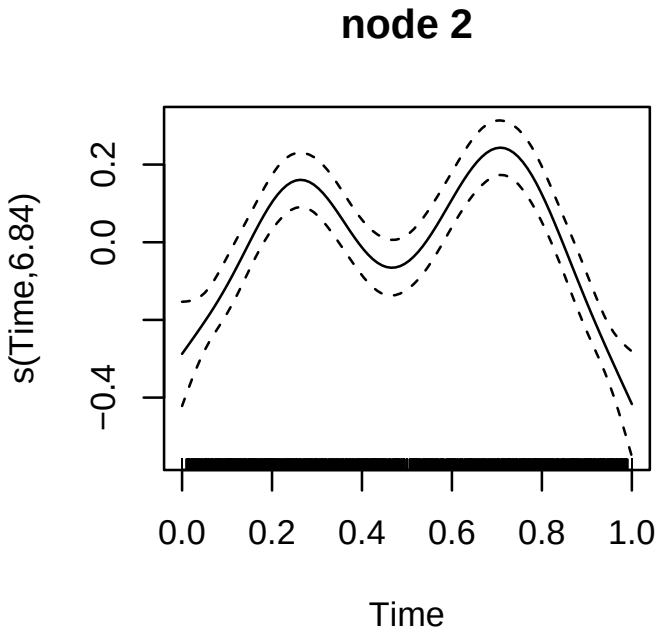
```
plot(gt, which = "tree",
     treeplot_ctrl = list(gp = gpar(cex = .7)))
par(mfrow = c(2, 2))
plot(gt, which = "terms")
```

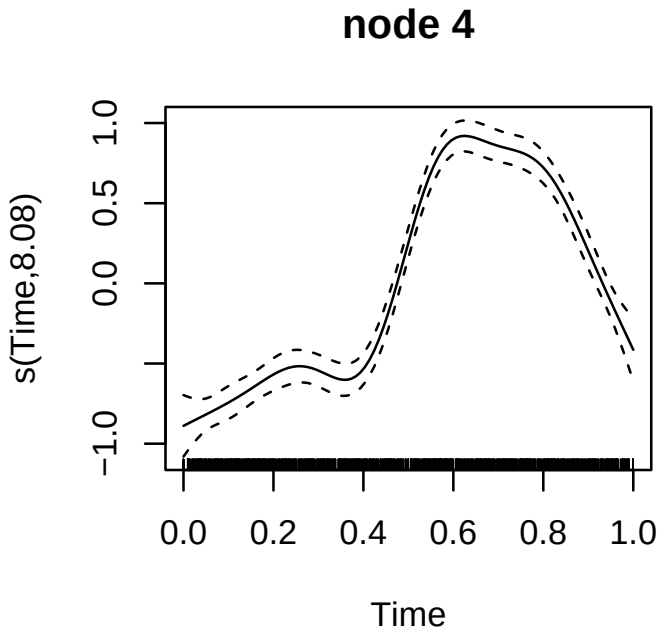
Output on next slides:

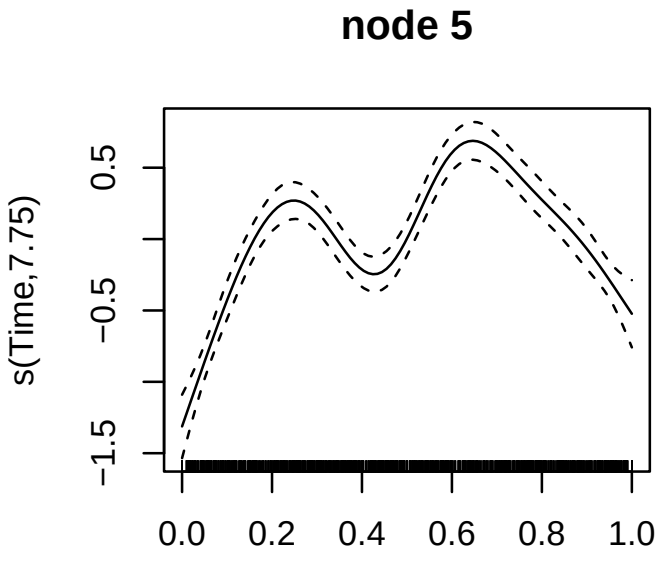
Fit a smoothing spline trees



Fit a smoothing spline trees







Getting valid tests

Sample splitting does yield a great advantage: Can obtain honest/valid hypothesis tests on data *not* used for finding the subgroups.

- Get node memberships for new data:

```
test_dat <- dat[test, ]
test_dat$nodes <- factor(predict(gt, newdata = dat[test, ],
                                 type = "node"))
```

- Fit GAM with subgroup structure to new data (though parametrization may need adjustment to test target hypotheses):

```
library("mgcv")
test_m <- bam(Pos ~ nodes + s(Time, by=nodes) +
                  s(Speaker, bs="re"),
              data = test_dat)
```

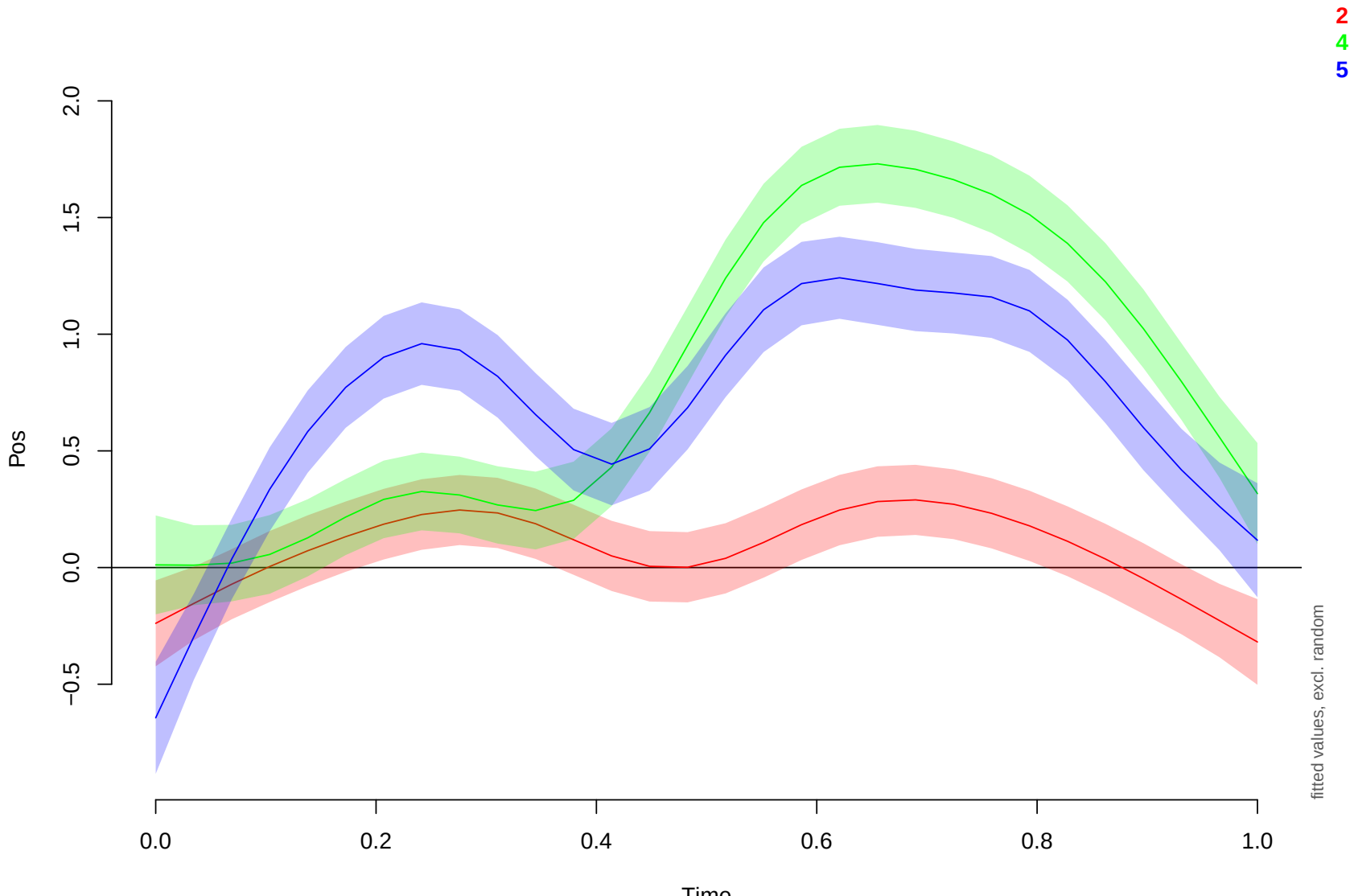
Getting valid tests

```
summary(test_m)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Pos ~ nodes + s(Time, by = nodes) + s(Speaker, bs = "re")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08354    0.06968   1.199    0.231
## nodes4       0.72409    0.02437  29.711   <2e-16 ***
## nodes5       0.62593    0.02825  22.155   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                    edf Ref.df      F p-value
## s(Time):nodes2  7.084  8.137  18.41  <2e-16 ***
## s(Time):nodes4  7.992  8.730 156.49  <2e-16 ***
## s(Time):nodes5  8.111  8.787  56.84  <2e-16 ***
## s(Speaker)     40.211 41.000  45.13  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Getting valid tests

```
library("itsadug")
plot_smooth(test_m, view = "Time", plot_all = "nodes",
            rug = FALSE, rm.ranef = TRUE)
```



Conclusion

Can we also partition smoothing or semi-parametric splines?

- Yes.
- But computational burden very heavy. Much work ahead!
- Sample splitting for reducing computational demands has a disadvantegeous side effect: Can get hypothesis tests after exploratory procedure.

Challenges to work on next:

- Improving computational speed of derivative computation and fitting of smoothing-spline trees (`gamtree`).
- Implement support for different correlation structures in `glmertree`, `splinetree` and `gamtree`.