Recursive partitioning of longitudinal and growth-curve models

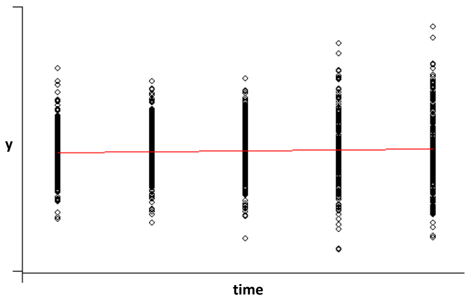
Marjolein Fokkema, PhD
Maaike Jorink, MSc

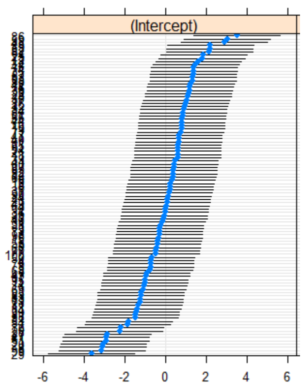
Leiden University

IFCS 2019, Thessaloniki, Greece

Linear growth curve model (LGCM)

(very basic) GLM: $\hat{y}_i = x_i^T \beta$
GLMM: $\hat{y}_i = x_i^T \beta + z_i^T b$



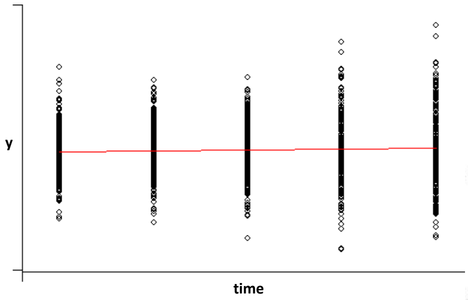


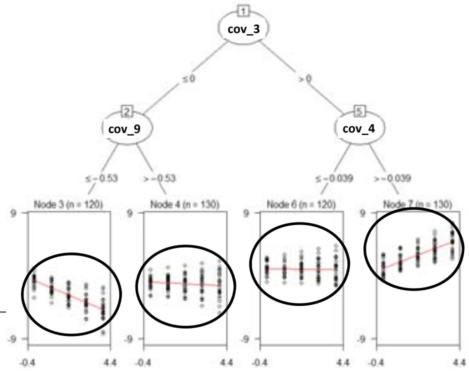
Recursive partitioning of LGCMs

(very basic) GLM: $\hat{y}_i = x_i^T \beta$

GLMM: $\hat{y}_i = x_i^T \beta + z_i^T b$

GLM tree: $\hat{y}_{ij} = x_i^T \beta_j$ (Zeileis et al., 2008)





Recursive partitioning of LGCMs

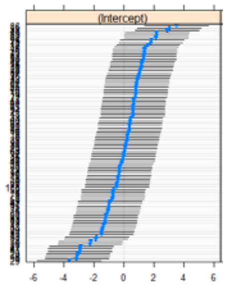
RQ1: Should variable selection tests account for level of the partitioning variables?

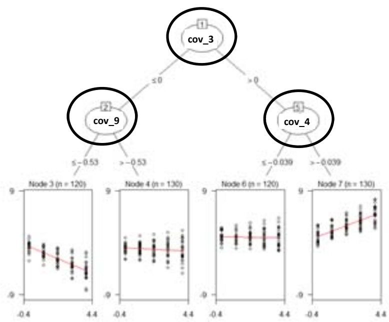
(very basic) GLM: $\hat{y}_i = x_i^T \beta$

GLMM: $\hat{y}_i = x_i^T \beta + z_i^T b$

GLM tree: $\hat{y}_{ij} = x_i^T \beta_j$ (Zeileis et al., 2008)

GLMM tree: $\hat{y}_{ij} = x_i^T \beta_j + z_i^T b$ (Fokkema et al., 2018)





Estimation of GLMM trees

RQ2: Better to initialize by estimating tree, or random-effects parameters?

GLMM tree: $\hat{y}_{ij} = x_i^T \beta_j + z_i^T b$

0) Initialize estimation assuming $\sigma_b = b = 0$

1) Estimate GLM tree, given current random

2) Estimate random effects, given current GL

3) Iterate between steps 1) and 2) until conve

RQ3: Necessary to estimate random effects? If so, random intercept and/or slopes?

Fokkema et al. (2018): Works well in clustered, cross-sectional data

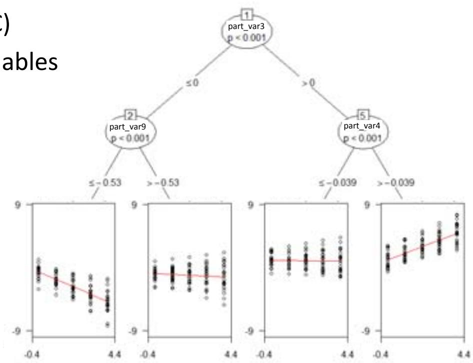
- σ_b not very large, partitioning variables measured at level I (ind. obs.)
- In partitioning LGCMs: σ_b larger, part. vars. measured at level II (cluster)

Simulation study

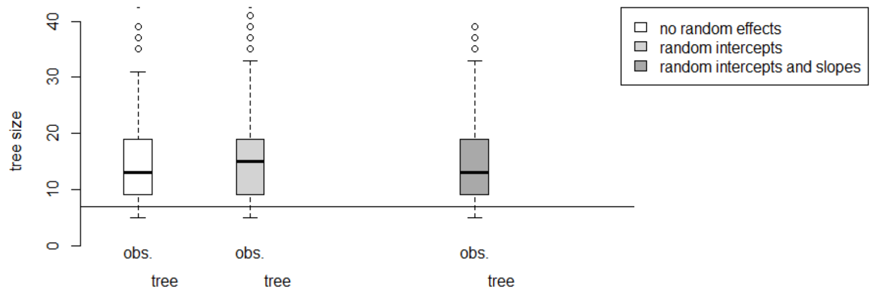
Outcome: Tree size (number of nodes)

Tree size increases with:

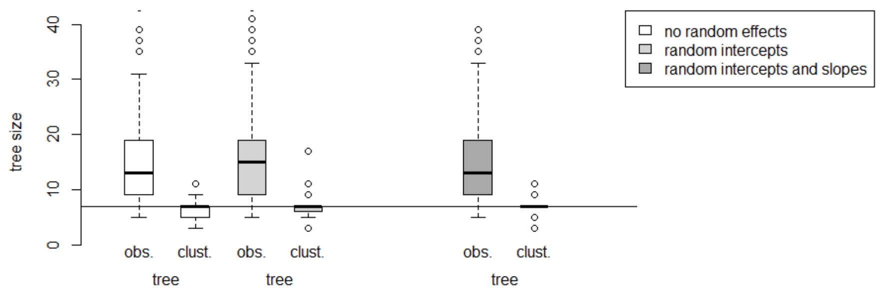
- Sample size
- Variance of random effects (ICC)
- No. of possible partitioning variables



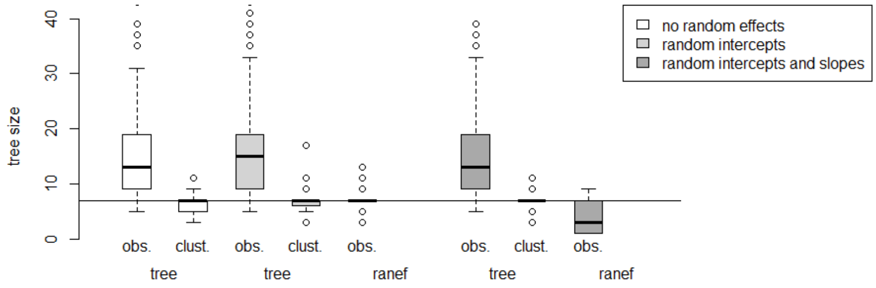
Simulation study



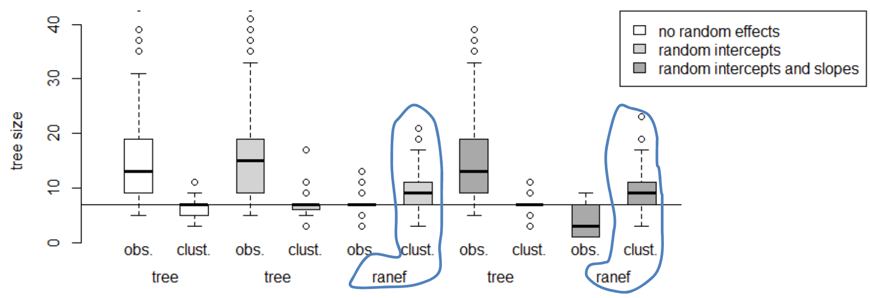
Simulation study



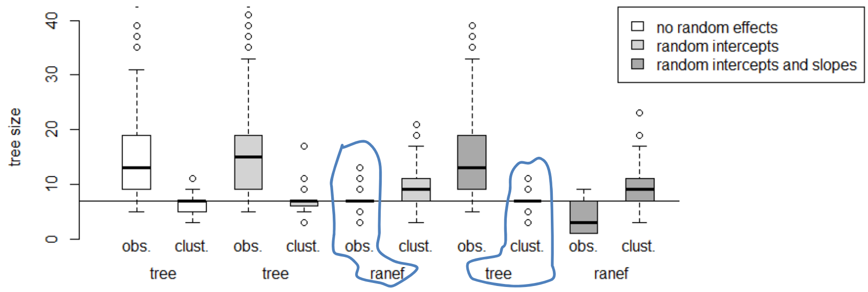
Simulation study



Simulation study



Simulation study



Application:
Early Childhood Longitudinal Study

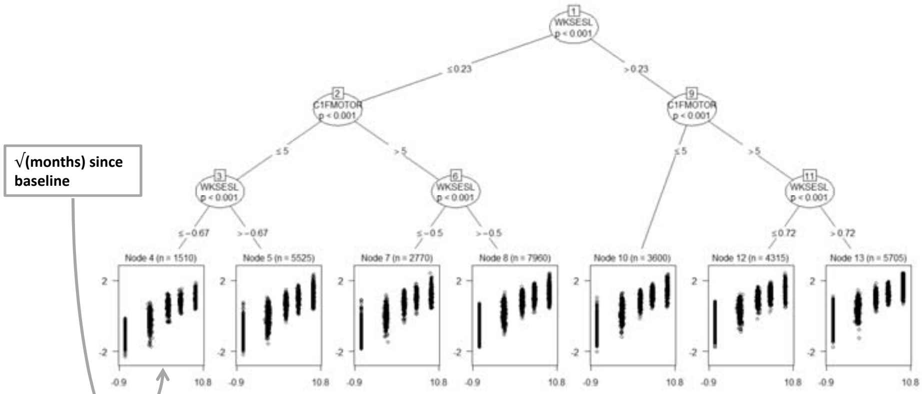
Repeated assessments of reading, math and science (N ≈ 6,500; ages 5 – 12)

8 potential partitioning variables:

- gender, SES, motor skills, psycho-social functioning…

Application:
Early Childhood Longitudinal Study

First splits very similar between different approaches (and responses):



Fixed intercept	**-1.00**	**-0.78**	**-0.66**	**-0.51**	**-0.56**	**-0.41**	**-0.32**
Fixed slope	**0.22**	**0.22**	**0.21**	**0.21**	**0.22**	**0.21**	**0.21**

Application:
Early Childhood Longitudinal Study

Results based on 10-fold CV with cluster-level sampling:

Approach	Random intercept only		Random intercept + slope	
	MSE (SE)	No. of nodes M (SD)	MSE (SE)	No. of nodes M (SD)
Observation-level tests				
Tree initialization	.157 (.002)	304.4 (31.71)	.164 (.002)	333.4 (27.93)
Random-effects initialization	.157 (.002)	304.4 (31.71)	.117 (.002)	19.6 (2.12)
Cluster-level tests				
Tree initialization	.145 (.002)	104.2 (18.38)	.149 (.002)	107.4 (16.38)
Random-effects initialization	.145 (.002)	104.2 (18.38)	.150 (.002)	245.8 (23.04)

Preliminary conclusions

- GLMM trees can detect subgroups in LGCMs (well)
- Should use **either** cluster-level tests or random-effects initialization to account for dependence of observations (not both)
  - Which is best? Different conclusion in simulation and real data:
    - Both indicate: initializing estimation with random effects (random intercept + slope) yields smallest trees
      - Likely also avoids overfitting
    - Real data analyses: smallest trees also most accurate
    - What is best likely depends on:
      - strength of random effects in data and
      - complexity of random-effects specification
    - Needs further study

Thank you!

R package glmertree: https://CRAN.R-project.org/package=glmertree

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T. & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods, 50*(5), 2016-2034.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514.