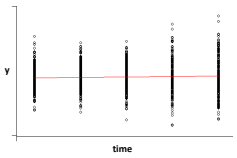
Recursive partitioning of clustered and longitudinal data with GLMM trees

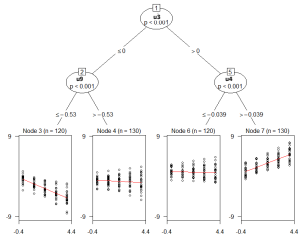
Marjolein Fokkema, PhD
Leiden University

CMstats 2019, London, UK

Recursive partitioning of clustered and longitudinal data

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

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

Estimation:

0) Set $\sigma_b = b = 0$
1) Estimate GLM tree, given current random-effects predictions
2) Estimate random effects, given current GLM-tree predictions
3) Iterate between steps 1) and 2) until convergence

- Fokkema et al. (2018): Works well in clustered, cross-sectional data (σ_b not very large, part. vars. measured at level I)
- Longitudinal data: part. vars. often at level II, σ_b often larger

Recursive partitioning methods for clustered and longitudinal data

Method	Estimation of random effects parameters	Model in terminal nodes	Possible levels of partitioning variables
REEM tree Hajjem et al., 2011 Sela & Simonoff, 2012	Global	Constant fits	Level I and II
GLMM tree Fokkema et al., 2018	Global	GLM	Level I and II
SEM tree Brandmaier et al., 2013	Local	GLMM	Level II
longRPart Abdollel et al., 2001 **longRPart2** Stegmann et al., 2018	Local	GLMM	Level II

Recursive partitioning methods for clustered and longitudinal data

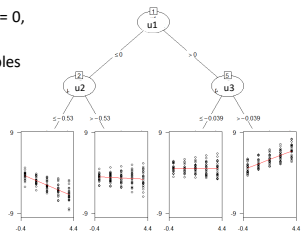
Method	Estimation of random effects parameters	Model in terminal nodes	Possible levels of partitioning variables
REEM tree Hajjem et al., 2011 Sela & Simonoff, 2012	Global	Constant fits	Level I and II
GLMM tree Fokkema et al., 2018	Global	GLM	Level I and II
SEM tree Brandmaier et al., 2013	Local	GLMM	Level II
longRPart Abdollel et al., 2001 **longRPart2** Stegmann et al., 2018	Local	GLMM	Level II

Simulation study: Design

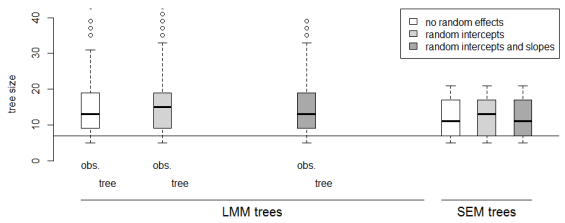
Outcome: Tree size (number of nodes)

Tree size increases with:

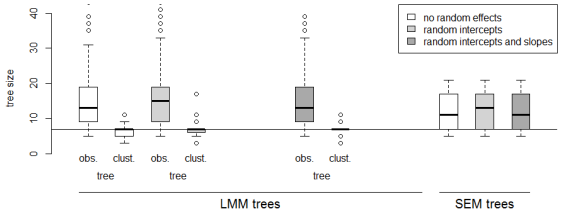
- Sample size (N = 80 or 200)
- Variance of random effects (ICC = 0, .17 or .44)
- # of potential partitioning variables (5 or 25)



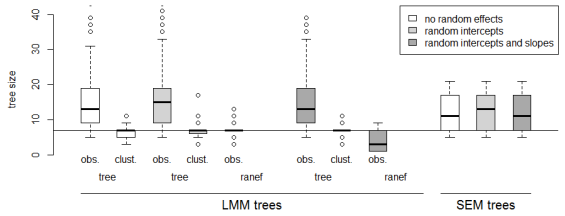
Simulation study: Results



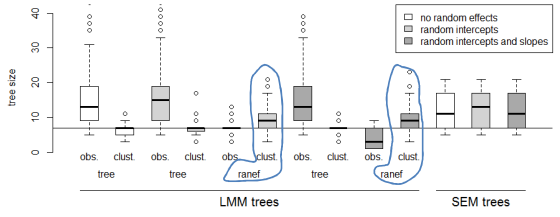
Simulation study: Results



Simulation study: Results

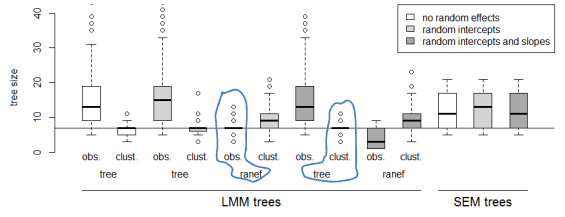


Simulation study: Results



Conclusions:
- Use **either** cluster-level tests **or** random-effects initialization (**not both**)

Simulation study: Results



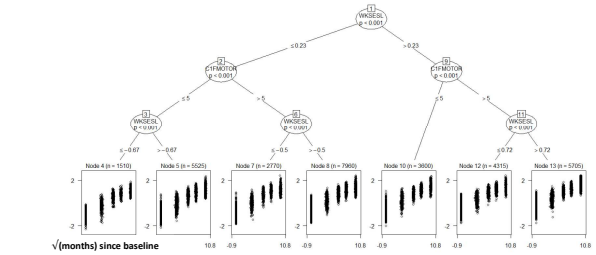
Conclusions:
- Use **either** cluster-level tests **or** random-effects initialization (**not both**)
- Initialize estimation with random effects only with simple random-effects specification

Application: Early Childhood Longitudinal Study

- Repeated assessments of reading, math and science (N ≈ 6,500; ages 5 – 12)
- 8 potential partitioning variables: gender, SES, fine and gross motor skills, psycho-social functioning, …
- ICCs ranging from .2 to .6
- SEM trees computationally infeasible with such large number of observations
- Results for reading, math and science very similar

Application: Early Childhood Longitudinal Study

First splits very similar between different fitting approaches:



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

Performance assessed with 10-fold CV with cluster-level sampling

Approach	Random intercept		Random intercept + slope	
	MSE (SE)	# nodes M (SD)	MSE (SE)	# 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)

Discussion

- GLMM trees detect subgroups in clustered and longitudinal data well
- Are computationally much less intensive than SEM trees, and can yield smaller, more accurate trees
- Default settings (observation-level tests for partitioning, initializing estimation with tree) work well if partitioning variables are measured at lowest level
- If partitioning variables measured at higher level
  - Use **either** cluster-level tests **or** random-effects initialization
  - What is best likely depends on ICC, sample size, complexity of random-effects specification

Discussion

- Implemented in R package glmertree https://CRAN.R-project.org/package=glmertree
- Future work:
  - Non-linear models in terminal nodes
  - Missing data

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T. & Kelderman, H. (2). Detecting treatment-subgroup interactions in clustered data generalized linear mixed-effects model trees. *Behavior Res Methods, 50*(5), 2016-2034.

marjolein.fokkema@gmail.com

Thank you!

Abdolell et al (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Statistics in Medicine, 21*(22), 3395-03409

Brandmaier, A.M., von Oertzen, T., McArdle, J.J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, *81*(4), 451-459.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674.

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine learning*, *86*(2), 169-207.

Stegmann, G., Jacobucci, R., Serang, S., & Grimm, K. J. (2018). Recursive Partitioning with Nonlinear Models of Change. *Multivariate Behavioral Research*, *53*(4), 559-570.

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