Leveraging the Boruta Algorithm for Variable Selection in Structural Equation Models

DGPs Congress Symposia: Methods and Applications of Modeling Technique; Hörsaal 30

Andreas M. Brandmaier

Pryanka Paul

Timothy R. Brick

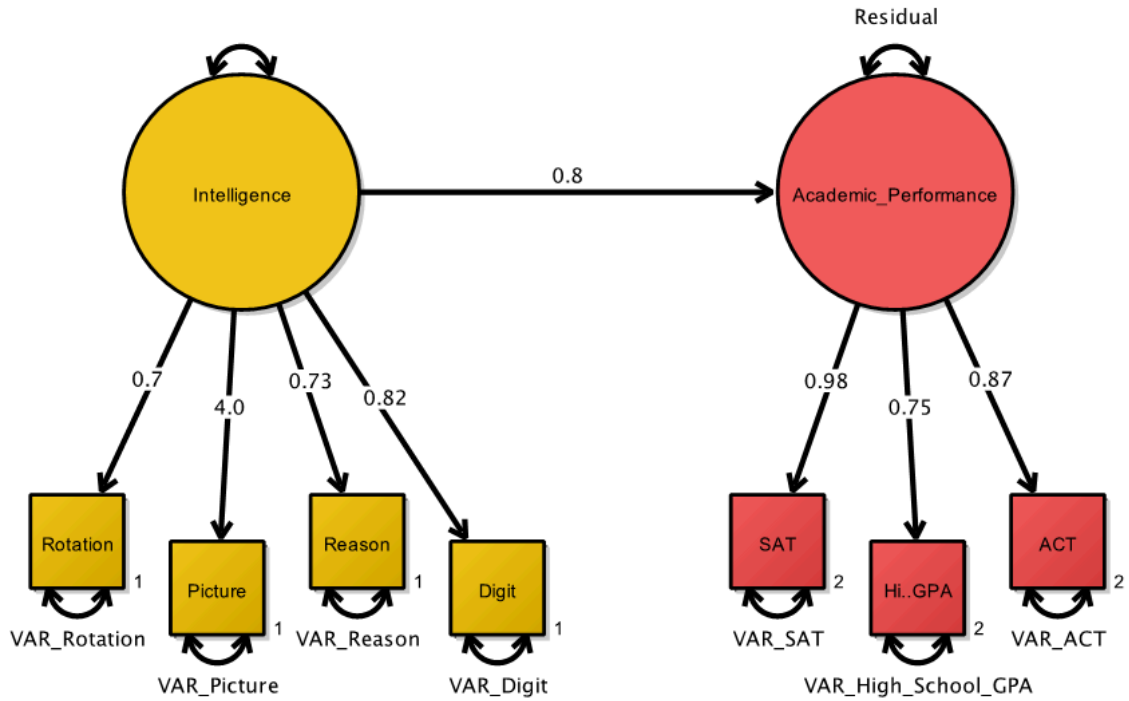
2024-09-17

Question: “Given a multivariate model, which predictors/covariates are relevant?

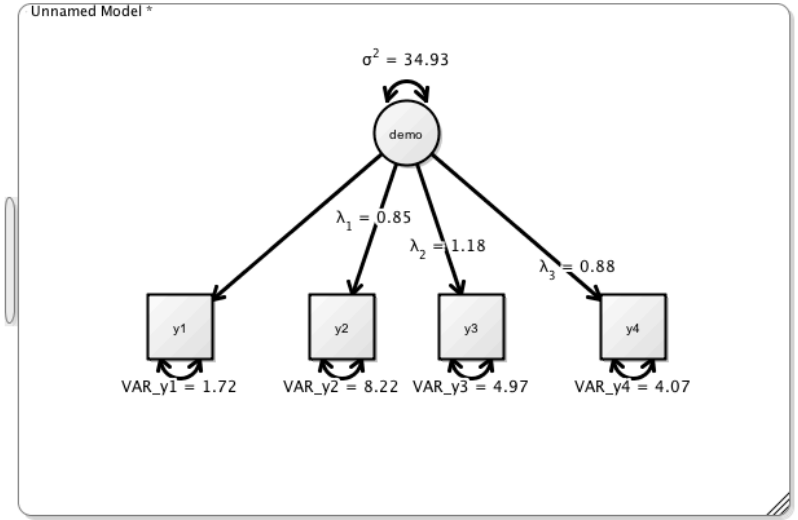
SEM + Decision Trees + Random Forests + Variable Importance + BORUTA

Roadmap

SEM


Residual
Intelligence
0.8
Academic_Performance
0.7
4.0
0.73
0.82
0.98
0.75
0.87
Rotation
Picture
Reason
Digit
SAT
Hi..GPA
ACT
1
1
1
1
2
2
2
VAR_Rotation
VAR_Picture
VAR_Reason
VAR_Digit
VAR_SAT
VAR_High_School_GPA
VAR_ACT

Commercial Break: Ωnyx

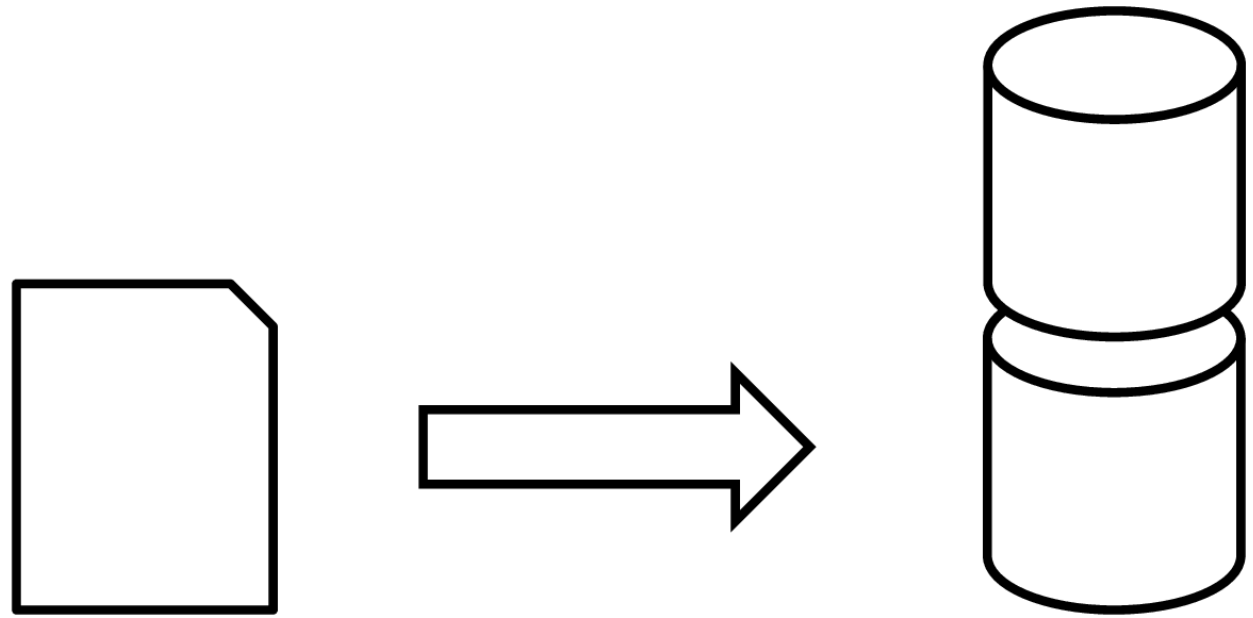


Lavaan code

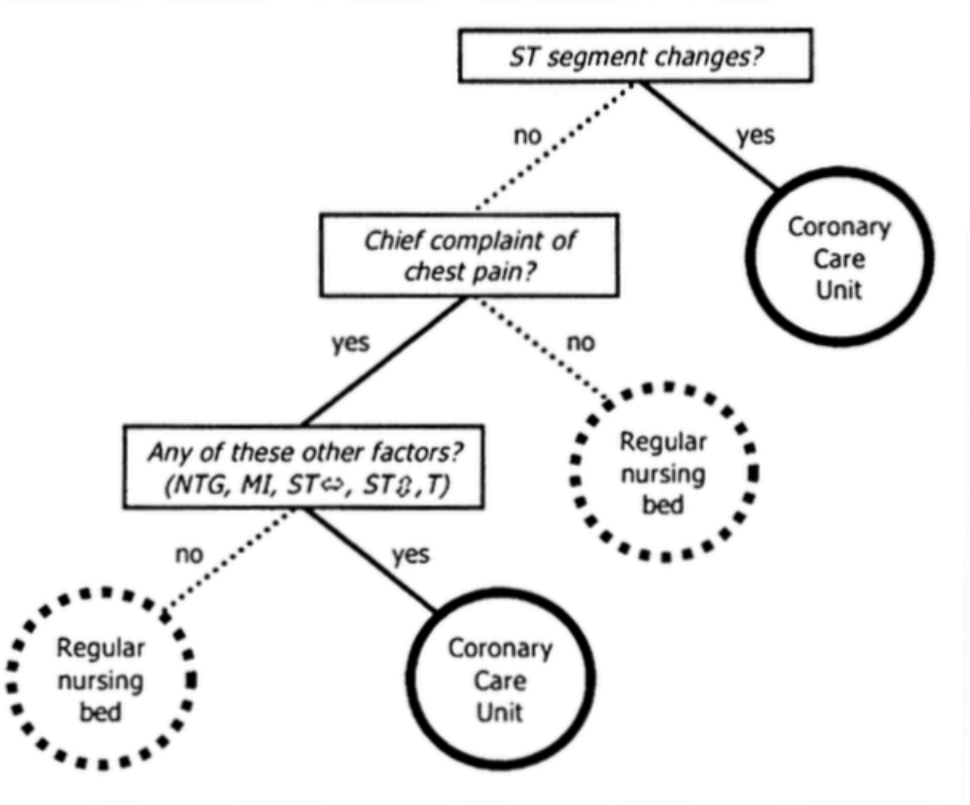
```
library(lavaan);
data<- ... ;
 onyx.model<-"
! regressions
   DEMO=~1.0*Y1
   DEMO=~Y2
   DEMO=~Y3
   DEMO=~Y4
! residuals, variances and covariances
   Y1 ~~ Y1
   Y2 ~~ Y2
   Y3 ~~ Y3
   Y4 ~~ Y4
   DEMO ~~ DEMO
! means
   Y1~0*1;
   Y2~0*1;
   Y3~0*1;
   Y4~0*1;
   DEMO~0*1;
";
fit<-lavaan(onyx.model, data=data);
summary(fit, fit.measures=TRUE);
```

https://onyx-sem.com/

Theory-driven modeling

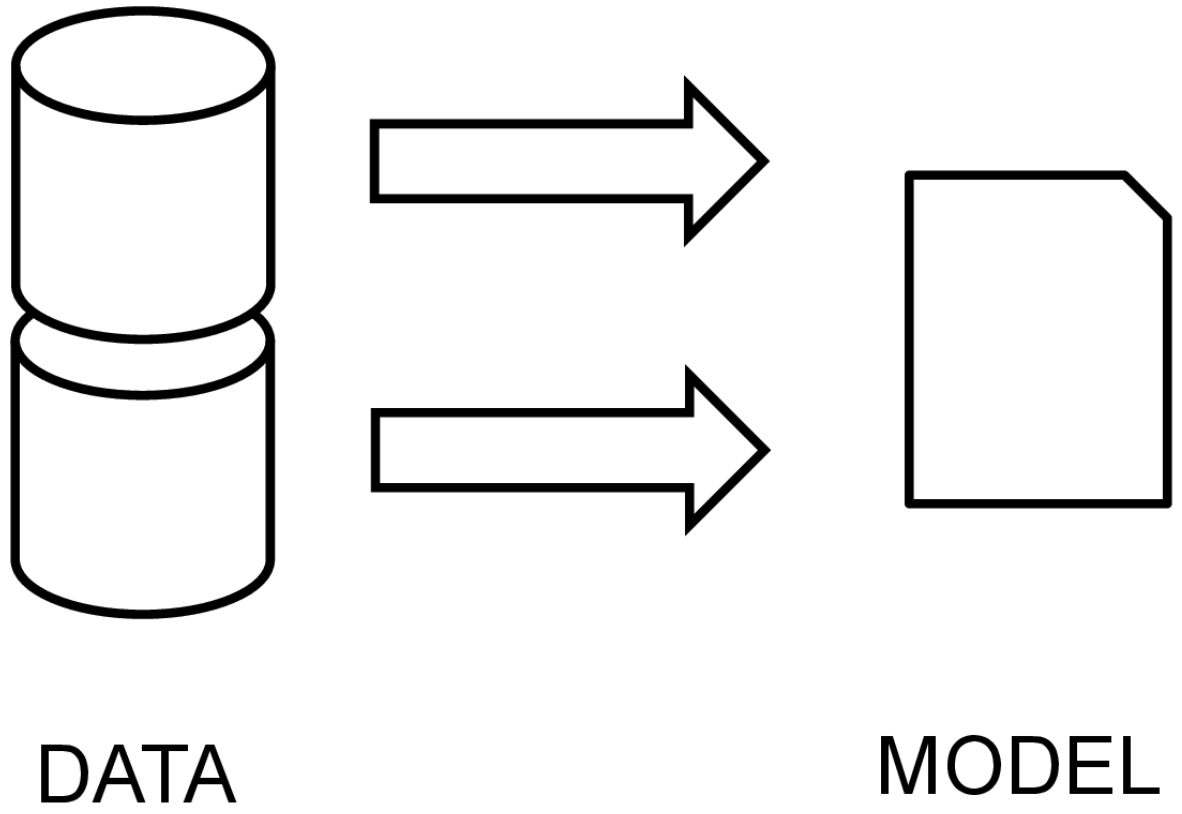


Decision Trees



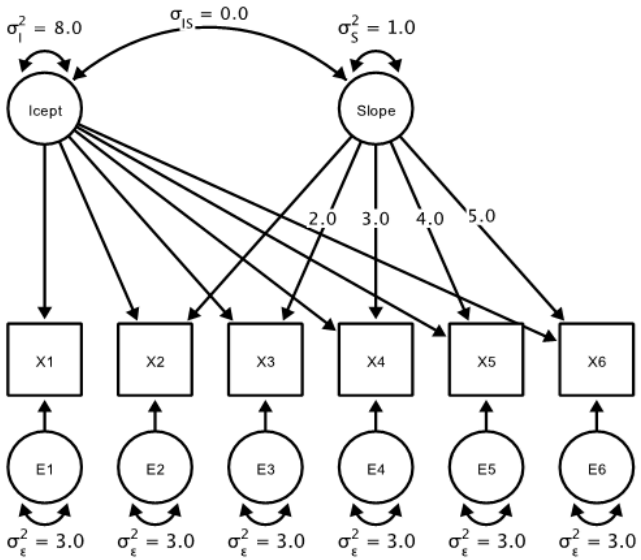
Gigerenzer and Kurzenhaeuser (2005)

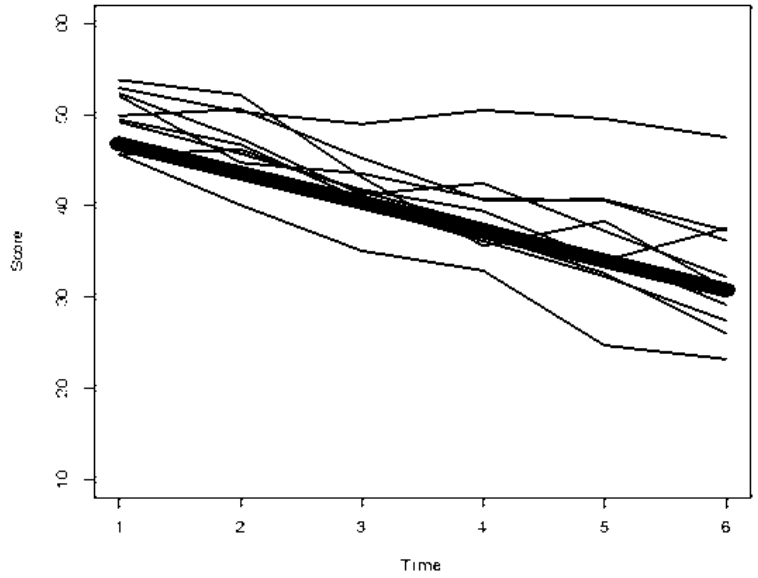
Data-driven modeling



SEM Trees

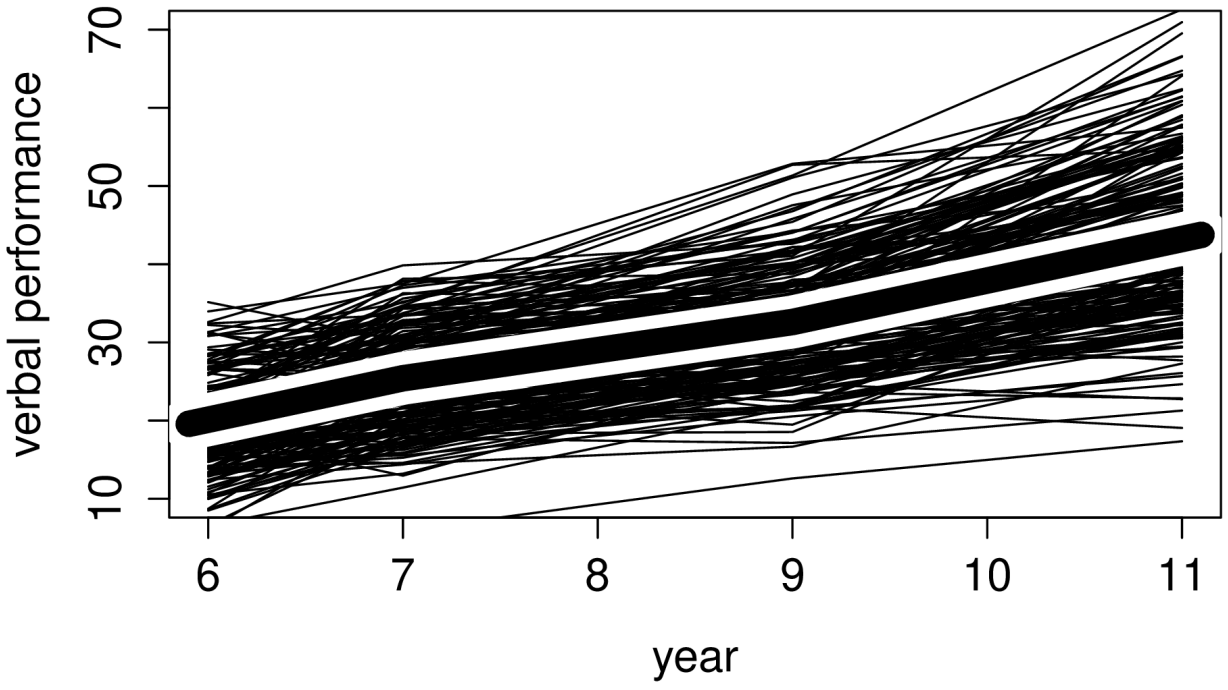
A Simple Example: Wechsler Intelligence Scale for Children





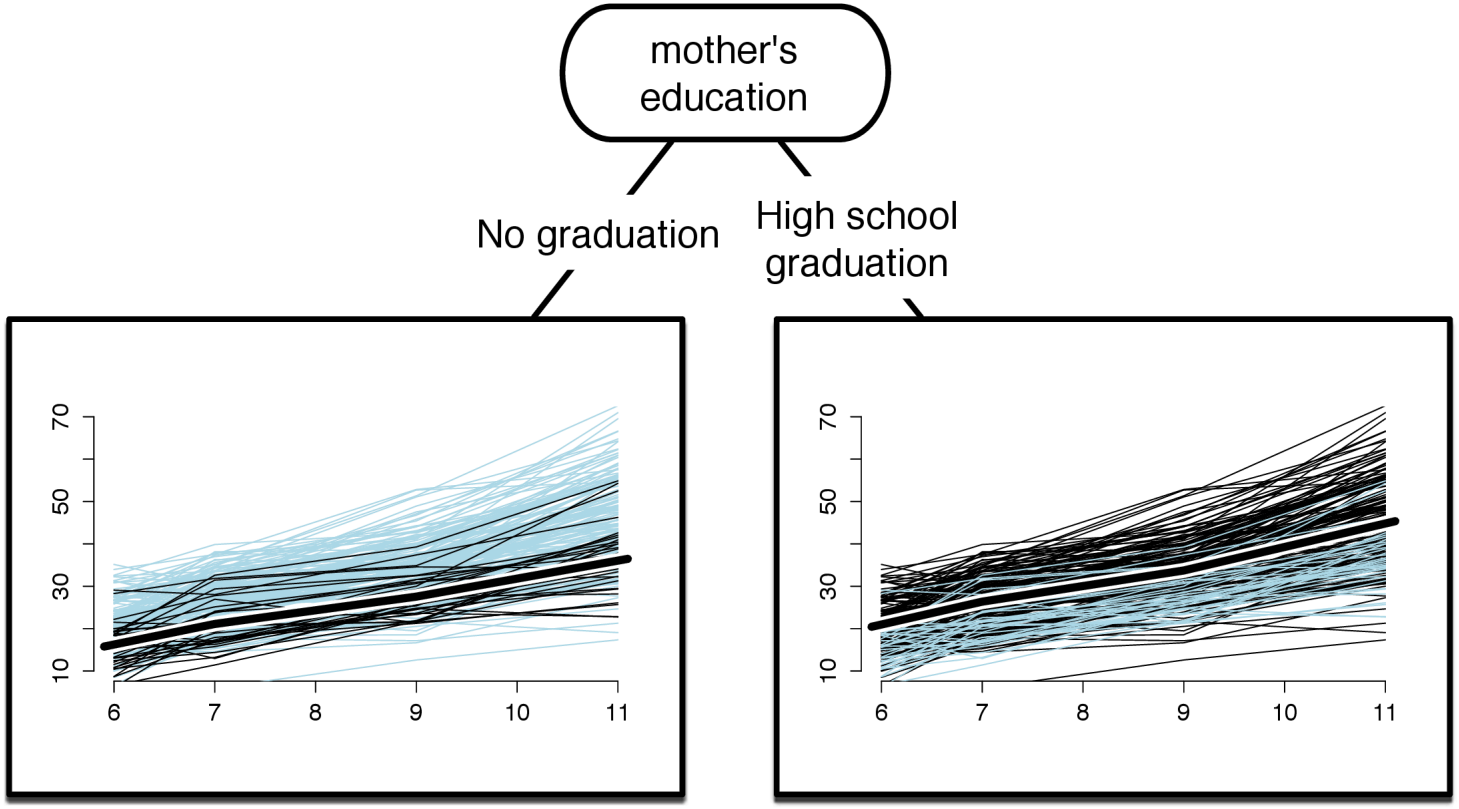
Brandmaier, von Oertzen, McArdle, and Lindenberger (2013)

A Simple Example: WISC

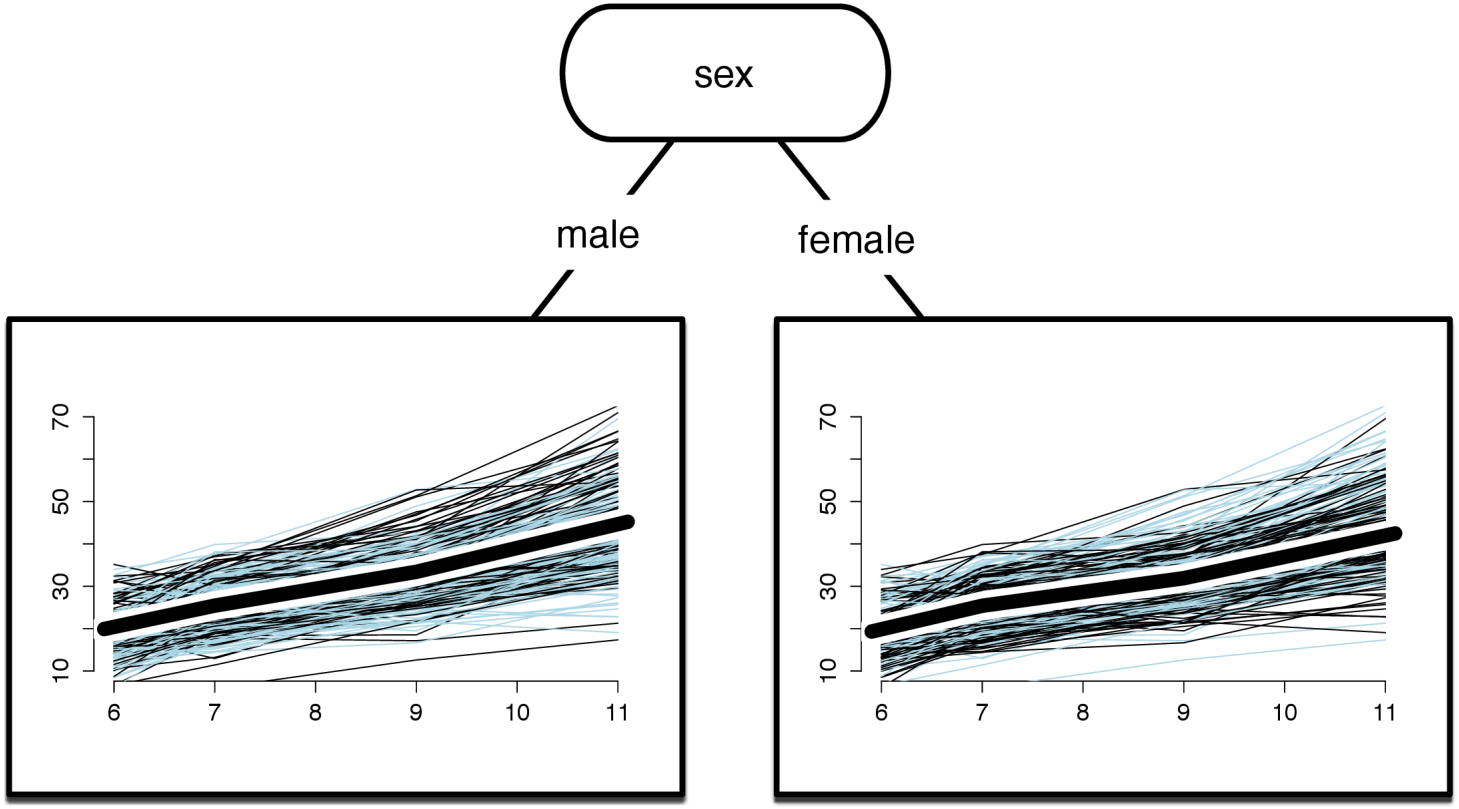


N=204 children, McArdle & Epstein, 1987

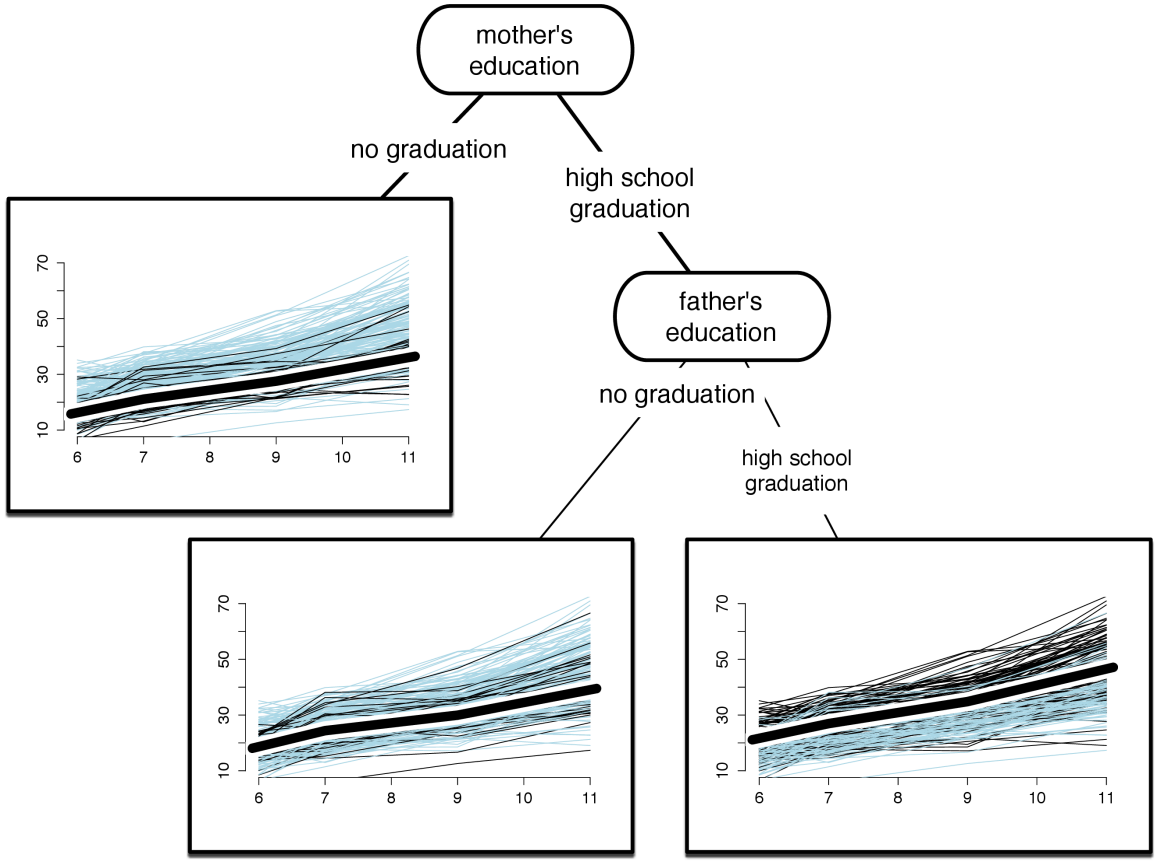
A Simple Example: WISC



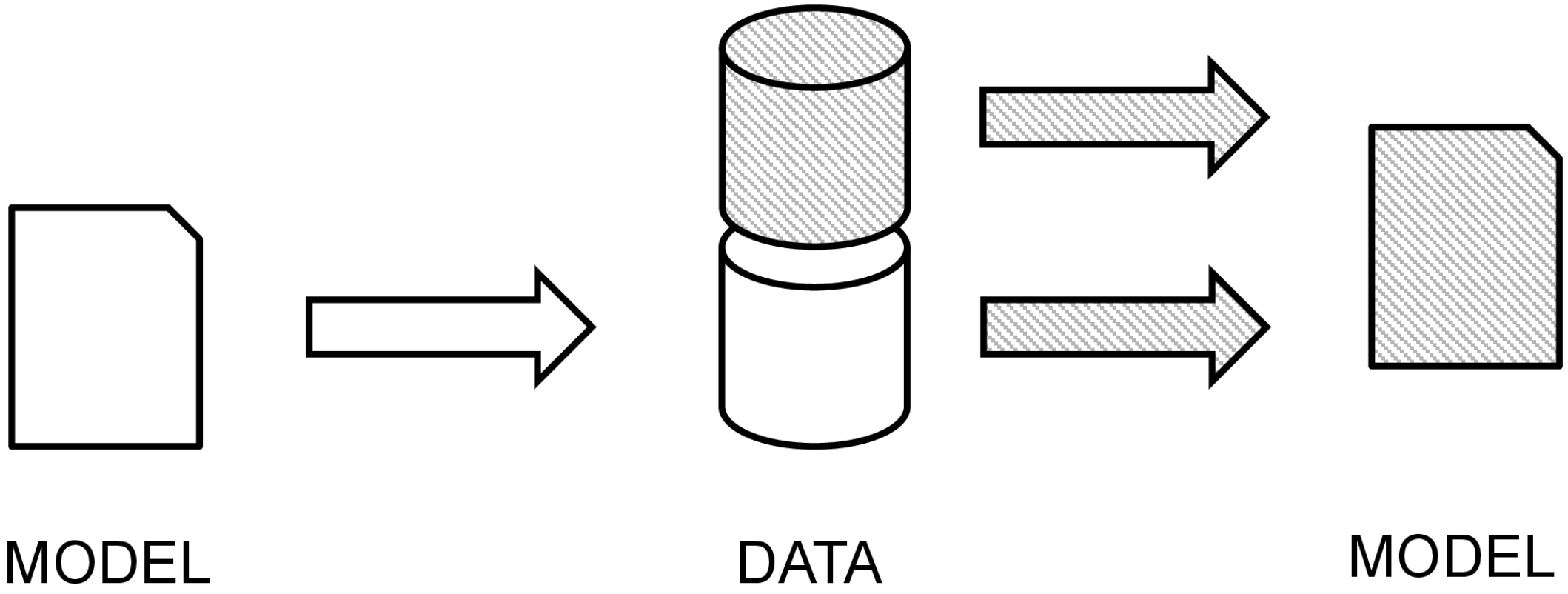
A Simple Example: WISC



A Simple Example: WISC



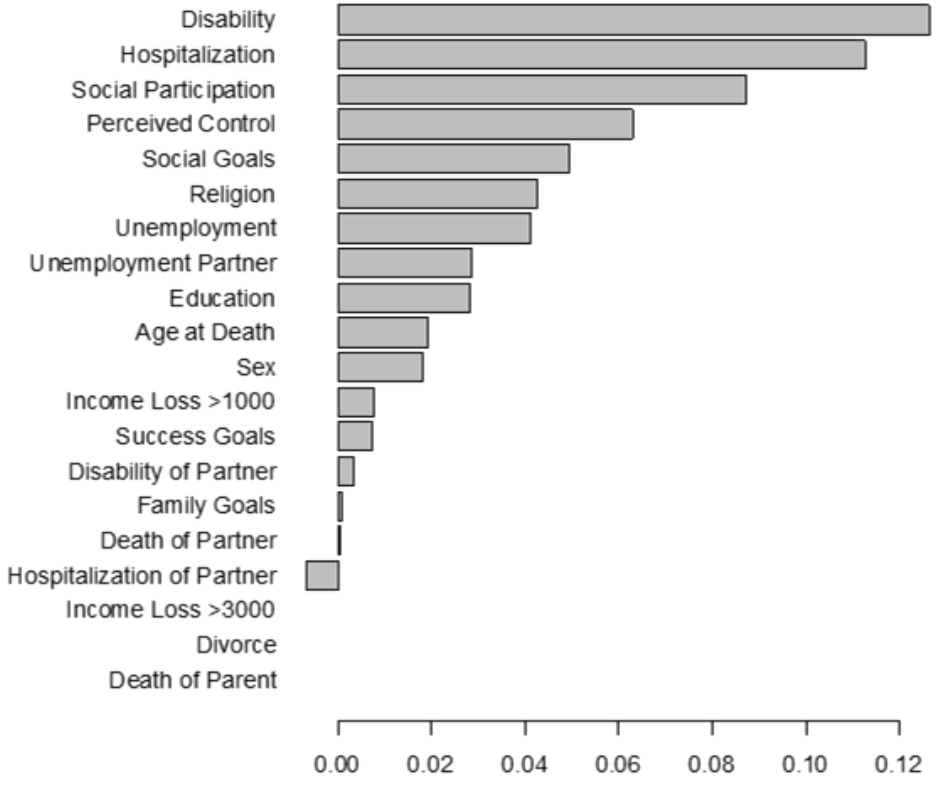
Theory-guided exploration



Brandmaier, Prindle, McArdle, and Lindenberger (2016)

Variable Importance

- single trees are unstable -> subsample data and predictors to create a forest with diverse predictor combinations
- using a permutation approach, estimate contribution of each predictor (Brandmaier, Prindle, McArdle et al., 2016)
- Example: Terminal decline of happiness from SOEP data (Brandmaier, Ram, Wagner, and Gerstorf, 2017)



BORUTA

A spirit or devil from slavic mythology, image from Wikipedia/Public Domain

Shadow features

Create random copies of all predictors (*shadow features*) features; keep predictors outperforming shadow features (Kursa and Rudnicki, 2010)

Algorithm

- Create a copy of the original data set
- Create shadow features by permutation of all non-rejected features (remove association of shadow features with outcome(s))
- Run a SEM forest and variable importance
- A feature is considered relevant if it performs better than the best shadow feature (**'hit'**)
- Run a statistical test for each original predictor (H0: predictor performs like max of shadow features)
- If significant, tag predictor as **'confirmed'** or **'rejected'**, otherwise leave as **'tentative'**
- repeat until no tentative features (or too may iterations)

Us combining SEM forests and BORUTA

(according to ChatGPT)

A minimal proof of concept

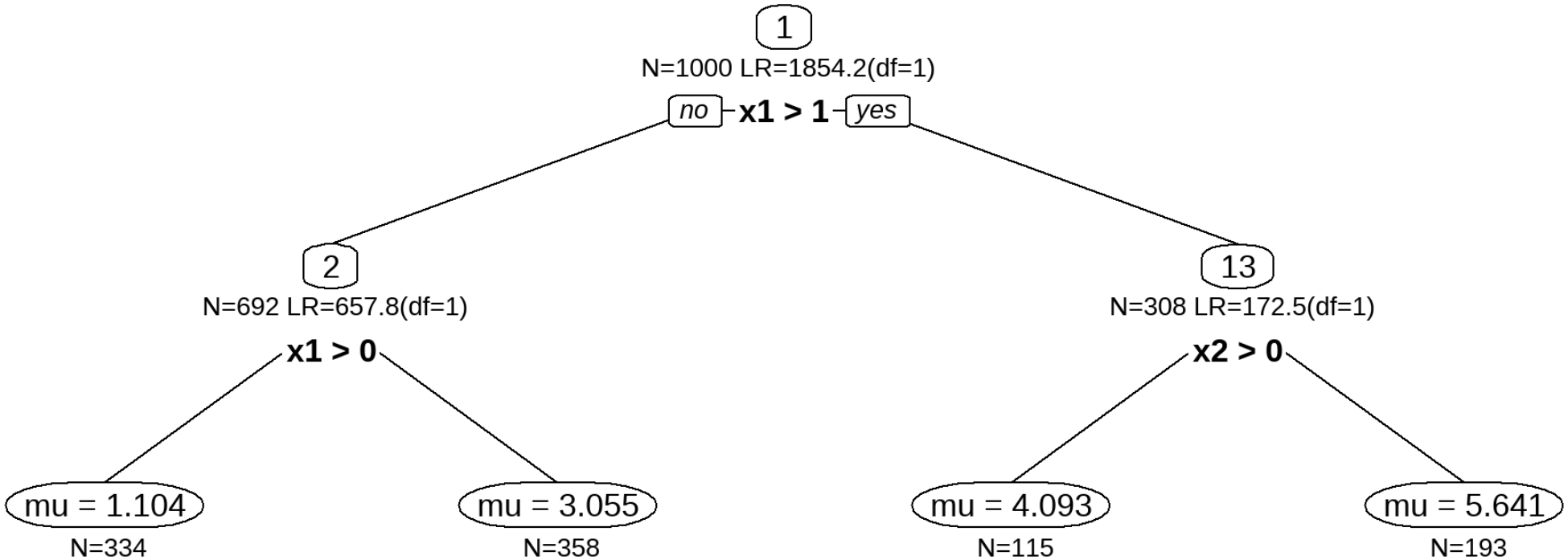
Setup

- Randomly draw ordinal $x1$ to $x6$ from $0, 1, 2$
- Generate a dataset of $N = 1000$
- Compute outcome according to:

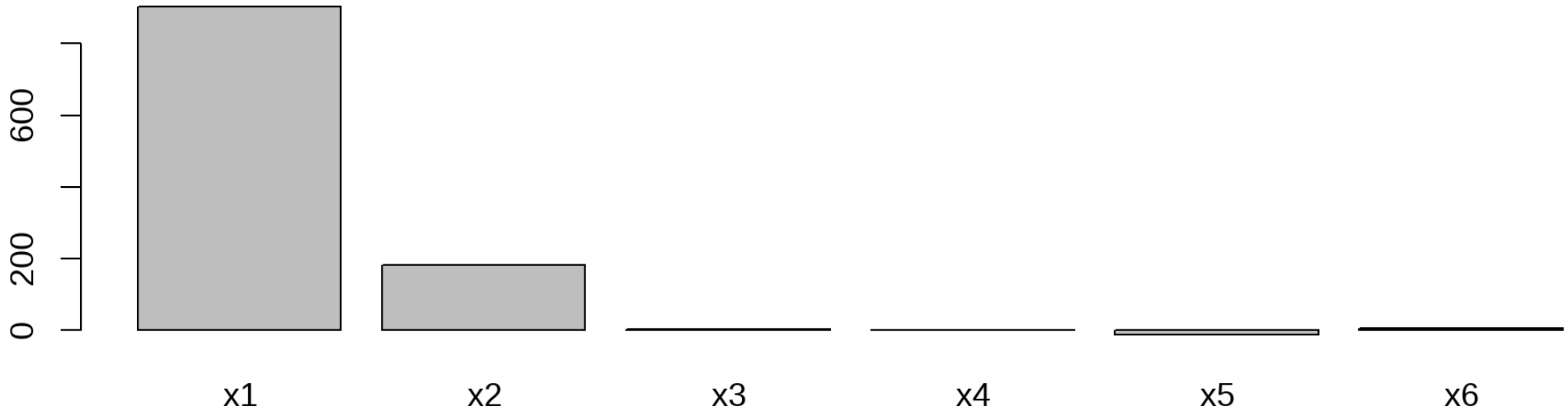
$$y = 2 \cdot x1 + 1 \cdot x2 + 0.1 \cdot x3 + \mathcal{N}(0, 0.01)$$

- Run BORUTA with 100 trees per forest for 11 runs
- Outcome model with only mean of y (basically like CART)

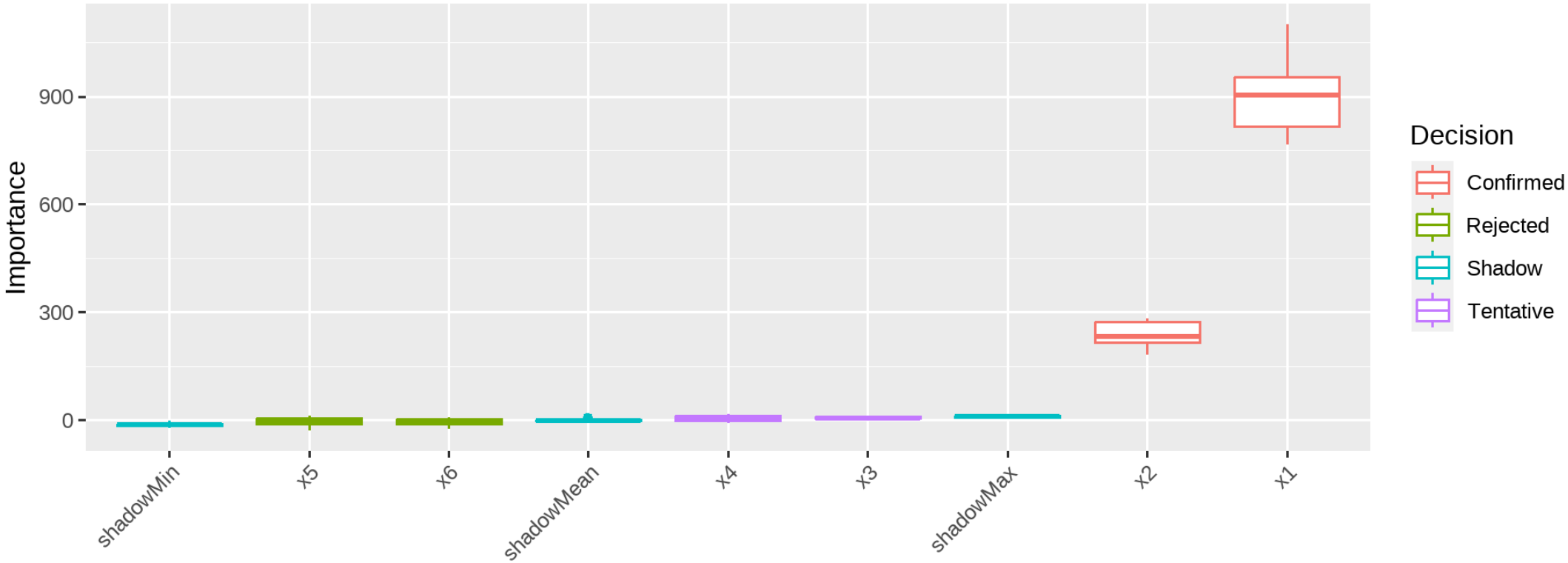
A simple SEM tree



Variable importance



BORUTA results

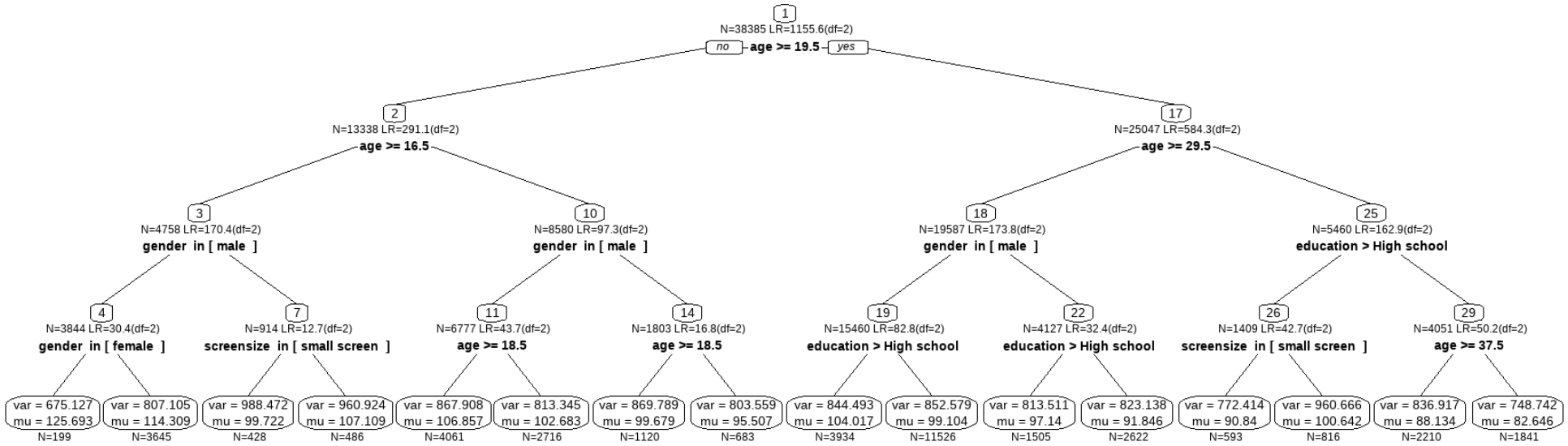


Example: Depression Anxiety Stress Scales (DASS)

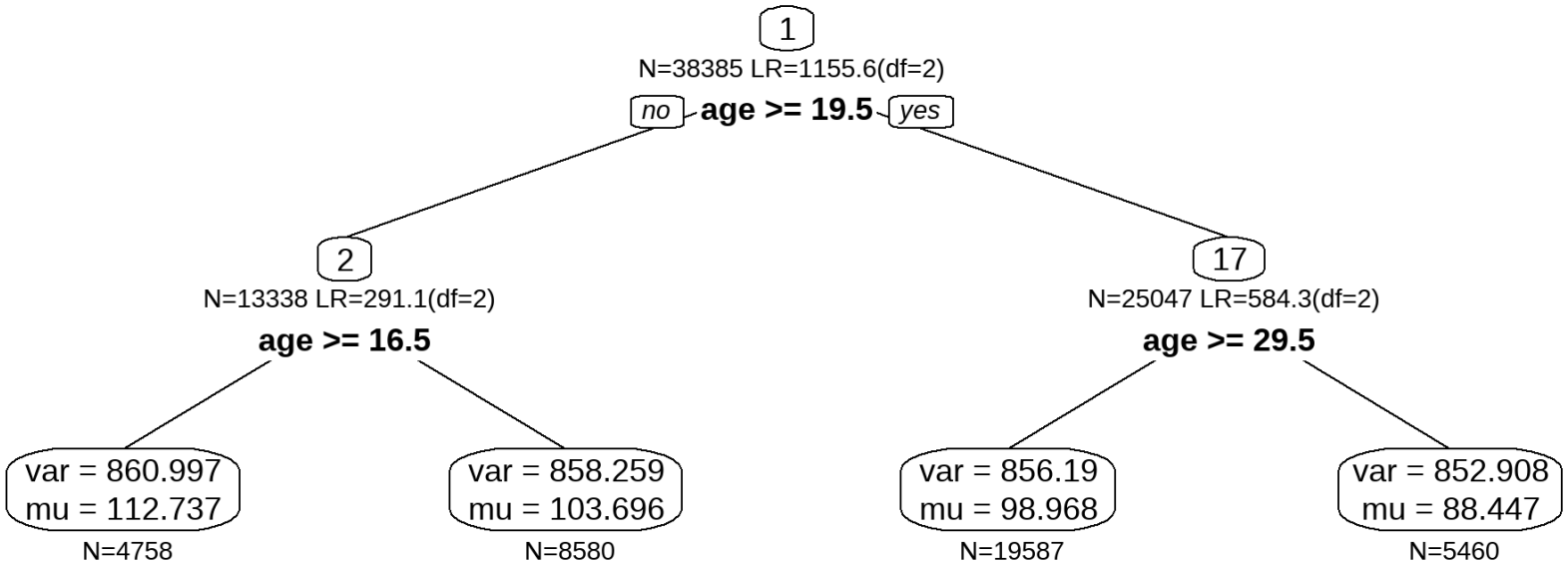
Example: DASS

- open data from openpsychometrics.org
- data was collected with an online version of the Depression Anxiety Stress Scales (DASS) (http://www2.psy.unsw.edu.au/dass/)
- 2df outcome model with mean and variance of the DASS sum score (42 items)
- predictors: age, education, gender, handedness, urban, family size, screen size, voted
- $N = 38,385$ complete cases (after some data cleaning)

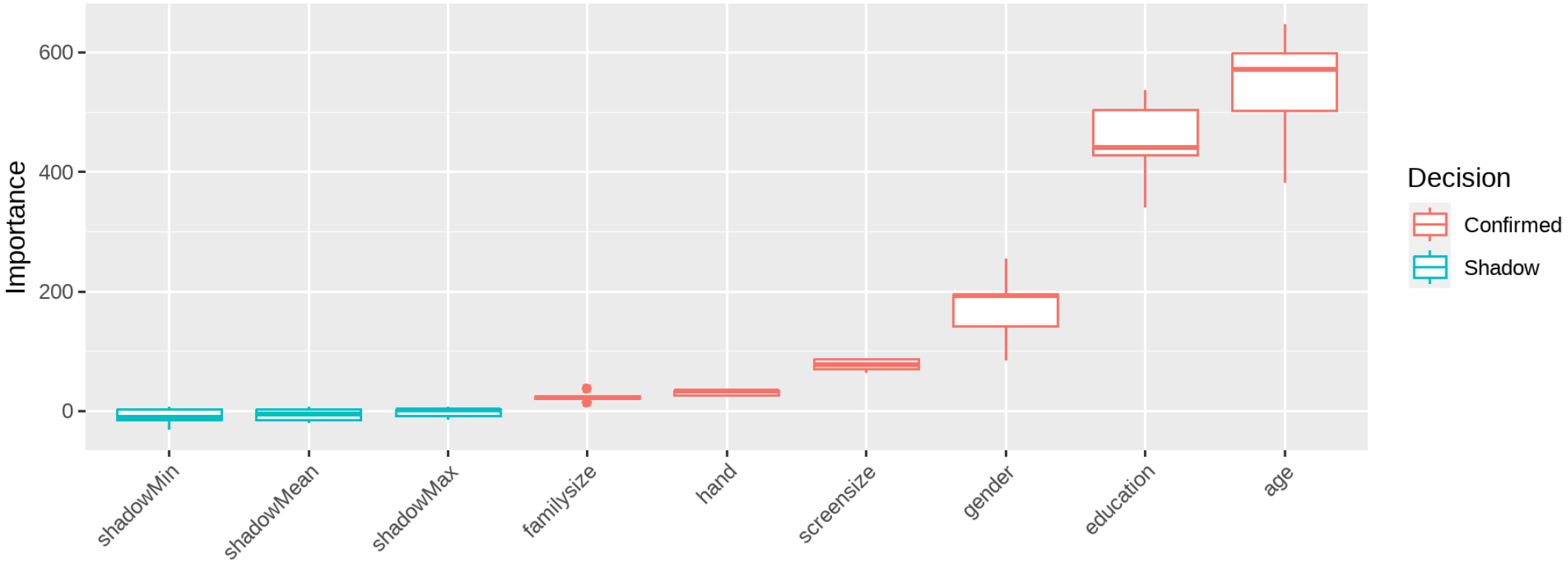
SEM Tree with DASS



SEM Tree with DASS (zoom in)



Example: DASS



Max-Operator

`max()`-operator on the shadow importance is problematic:

- the more shadow features, the larger the expected value of the maximum over all shadow features
- the threshold of relevance changes depending on the number of initial features
- as features get rejected, the threshold tends gets lower

Possible fix: quantile instead of max?

Why you should not use it

- BORUTA is a naive wrapper algorithm and computationally quite costly
- As always: type-I and type-II errors depend on sample size, effect size, multiple testing adjustments
- Difficult implications of "confirmation" and "rejection"
- Categorization hardly really helpful for theory advancement
- No short cut to a new theory

Thank You

Slides: https://github.com/brandmaier/boruta_presentation_dgps

Contact:
andreas.brandmaier@medicalschool-berlin.de or @brandmaier on X or @brandmaier.bsky.social on Bluesky or https://www.brandmaier.de

References

Brandmaier, A. M., T. von Oertzen, J. J. McArdle, et al. (2013). "Structural equation model trees." In: *Psychological methods* 18.1, pp. 71-86.

Brandmaier, A. M., J. J. Prindle, J. J. McArdle, et al. (2016). "Theory-guided exploration with structural equation model forests". In: *Psychological Methods* volume=21, pp. 66–-582.

Brandmaier, A. M., N. Ram, G. G. Wagner, et al. (2017). "Terminal decline in well-being: The role of multi-indicator constellations of physical health and psychosocial correlates." In: *Developmental Psychology* 53.5, pp. 996-1012.

Gigerenzer, G. and S. Kurzenhaeuser (2005). "Fast and frugal heuristics in medical decision making". In: *Science and medicine in dialogue: Thinking through particulars and universals* 30, pp. 3-15.

Kursa, M. B. and W. R. Rudnicki (2010). "Feature Selection with the Boruta Package". In: *Journal of Statistical Software* 36.11, pp. 1-13. URL: https://doi.org/10.18637/jss.v036.i11.