Machine Learning Structural Equation Modeling and Falsificatory Data Analysis

Michael S. Truong and Ji Yeh Choi
York University

Modern Modeling Methods
June 25/2024

Conclusion

- Confirmatory and Exploratory Data Analysis are about what **is** out there
- Falsificatory Data Analysis is about what **is not** out there
- 3 Claims:
  1. FDA side-steps the problem of over-fitting
  2. ML-SEM has no equal in performing FDA
  3. FDA: Advance theories through their *Zone of Impossibility*

Today's Outline

1. What is Machine Learning? Causal Modeling? Predictive Modeling?
2. Machine Learning Structural Equation Modelling
3. Falsificatory Data Analysis
4. I-GSCA Trees and Falsificatory Data Analysis

1. What is Machine Learning? Causal Modeling? Predictive Modeling?

I. Concepts: Causal vs. Predictive Modeling
II. Archetypes of Causal Modeling
III. Pros/Cons of Causal Modeling
IV. Archetypes of Predictive Modeling
V. Pros/Cons of Predictive Modeling

I. Concepts: Causal vs. Predictive Modeling

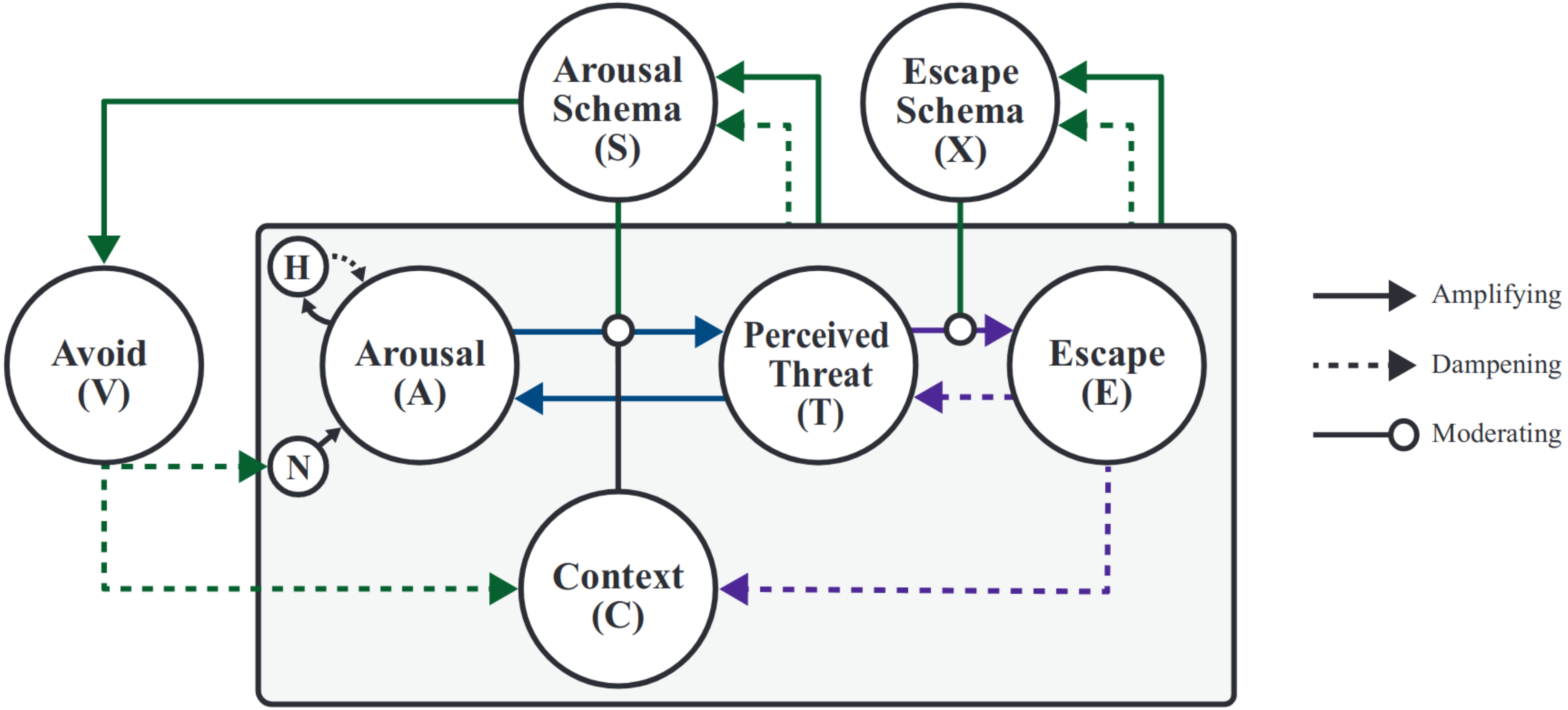
- Causal Modeling:
  - Change *X*, *Y* change?
- Predictive Modeling:
  - See *X*, *Y* is?

- Roughly, predictive modelling trades (**1**) mechanistic plausibility and interpretability for (**2**) utility and replicability
  - Neuro-genetic cognitive causal model to explain binge drinking @ 16
  - *Smoking @ 14 to predict binge drinking @ 16*

Pearl and Mackenzie, 2018; Shmueli, 2010; Yarkoni & Westfall, 2017

II. Archetypes of Causal Modeling

Causal Diagram (ABM) for Panic Stress Disorder



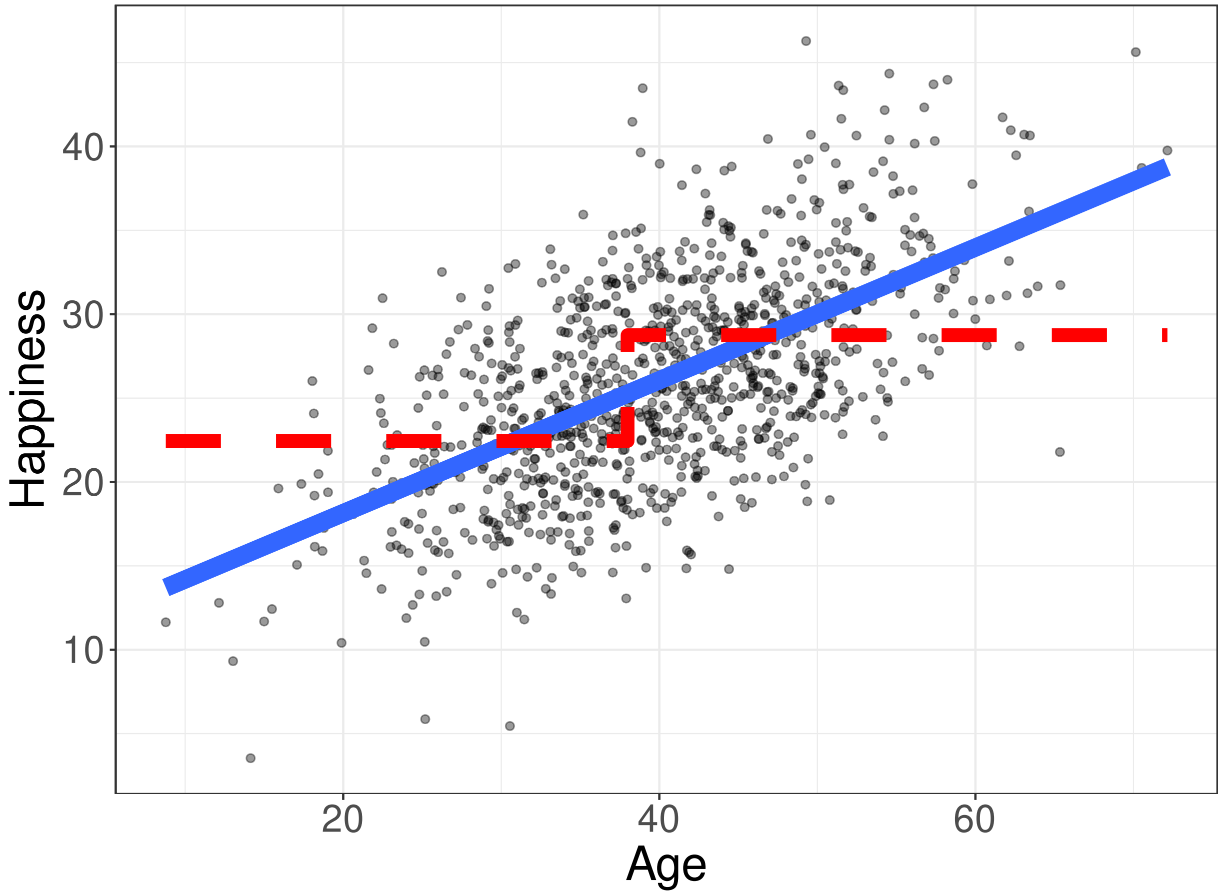
(N: Noise; H: Homeostatic Feedback)

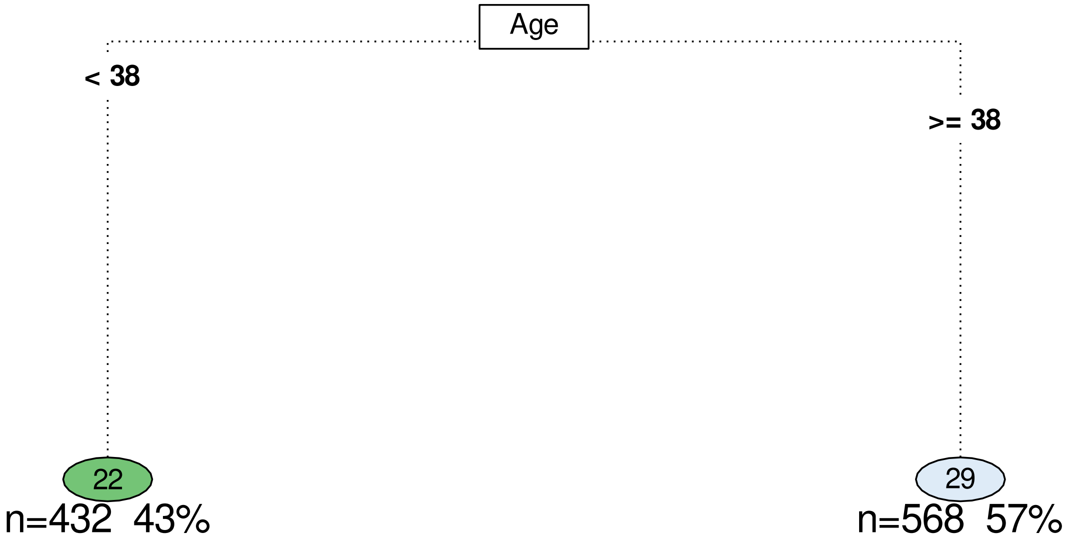
Robinaugh et al. 2019, Figure 1

III. Pros/Cons of Causal Modeling

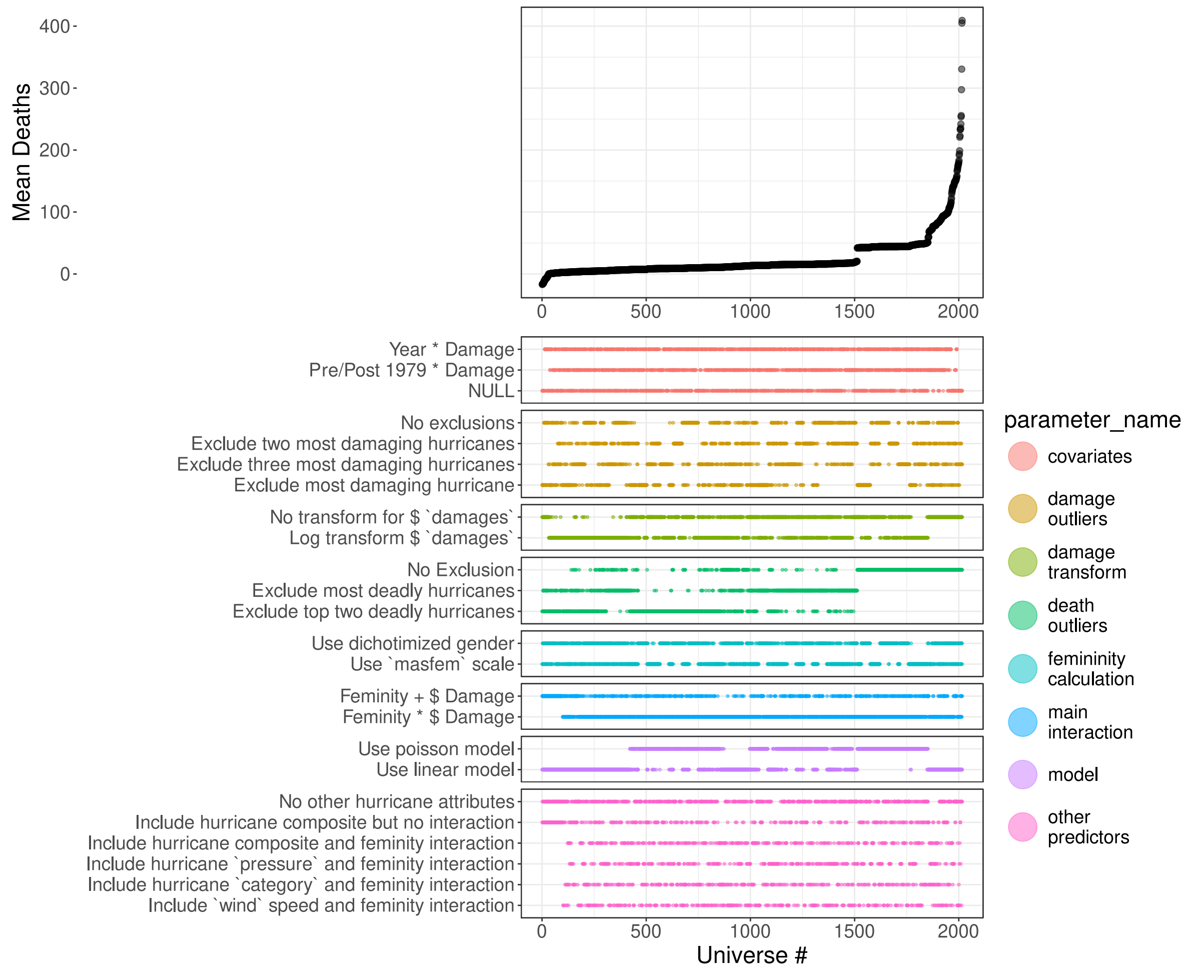
- Pros
  - Logical coherence between different datasets
  - Understanding
  - Successful Intervention
- Cons
  - Weak theory, weak model
  - N < J is tough
  - Measurability of relevant constructs?

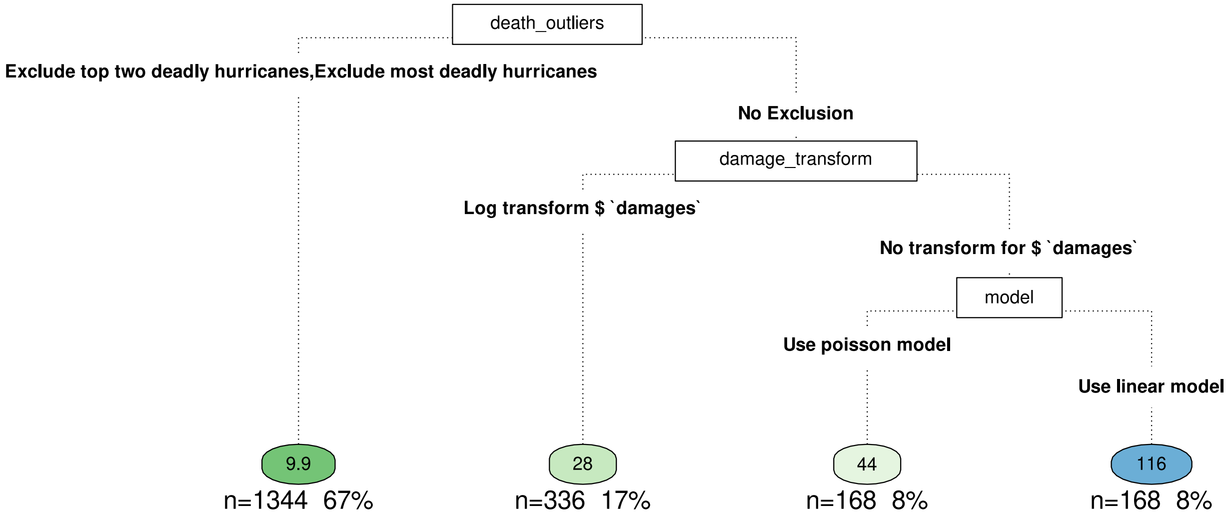
IVA. Archetypes of Predictive Modeling





IVB. Archetypes of Predictive Modeling





See our poster for more!

Data and Code Associated with Sarma et al. (2022) based on Jung et al. (2014) and Simonsohn et al. (2020)

V. Pros/Cons of Predictive Modeling

- Pros
  - Replicability
  - Utility
  - Handles N < J
  - Comparable predictive ability to true causal model (Shmueli 2010)
  - Beyond A > B, A < B… to A is *here* and B is *there*: **why**?
- Cons
  - Causally uninterpretable/incorrect (McElreath, 2020; Pearl & Mackenzie, 2018)
  - Interpretability? (c.f., Henninger et al., 2023)

Fuse ML + SEM???

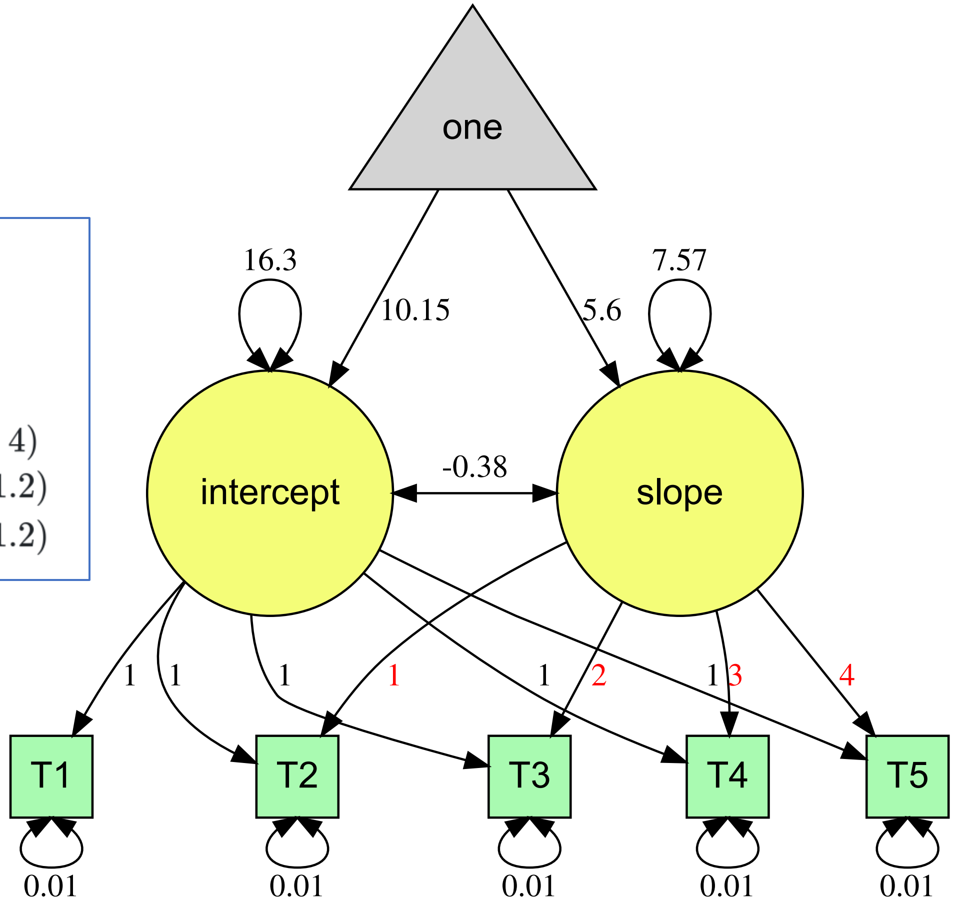
Off-set weaknesses and get best of both worlds for free?!

2. Machine Learning Structural Equation Modelling

I. The Case of SEM Trees
II. I-GSCA
III. I-GSCA Trees
IV. Pace: Capitalizing on Chance

IA. The Case of SEM Trees

$$\begin{aligned} N &= 1000 \\ N_{G1} &\sim 500 \\ N_{G2} &\sim 500 \\ Timepoints &= 5 \\ \beta_0 &\sim N(\mu = 10, \sigma = 4) \\ \beta_{slope(G1)} &\sim N(\mu = 3, \sigma = 1.2) \\ \beta_{slope(G2)} &\sim N(\mu = 8, \sigma = 1.2) \end{aligned}$$

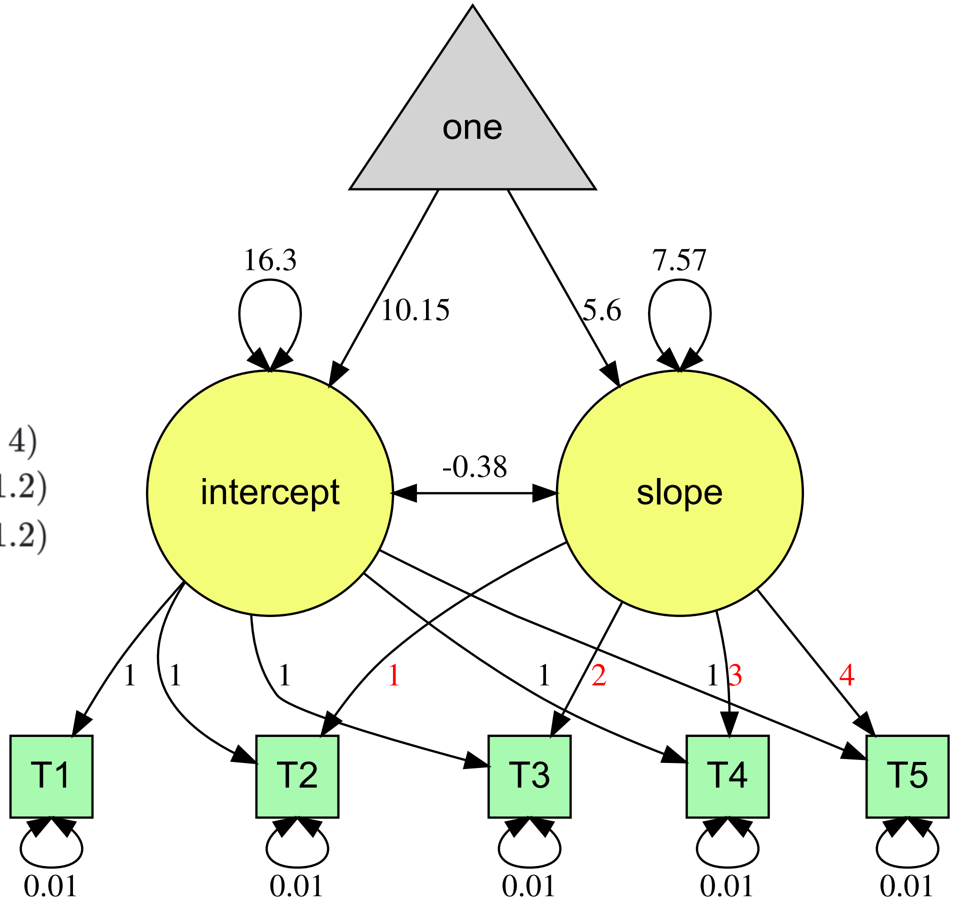


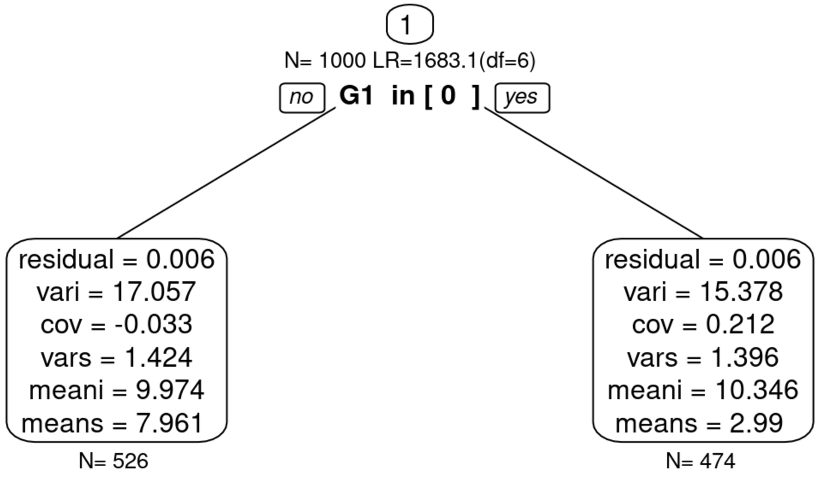
- Use DT to split data on predictor (group)
- Best fitting multi-group model?

semtree vignette: Getting Started with the semtree package

IB. The Case of SEM Trees

$$N = 1000$$
$$N_{G1} \sim 500$$
$$N_{G2} \sim 500$$
$$Timepoints = 5$$
$$\beta_0 \sim N(\mu = 10, \sigma = 4)$$
$$\beta_{slope(G1)} \sim N(\mu = 3, \sigma = 1.2)$$
$$\beta_{slope(G2)} \sim N(\mu = 8, \sigma = 1.2)$$

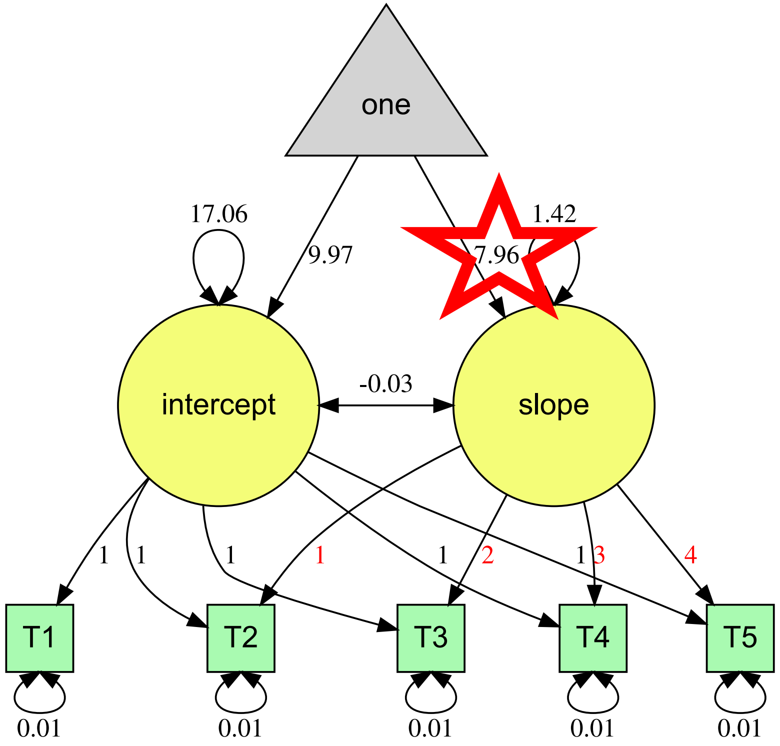


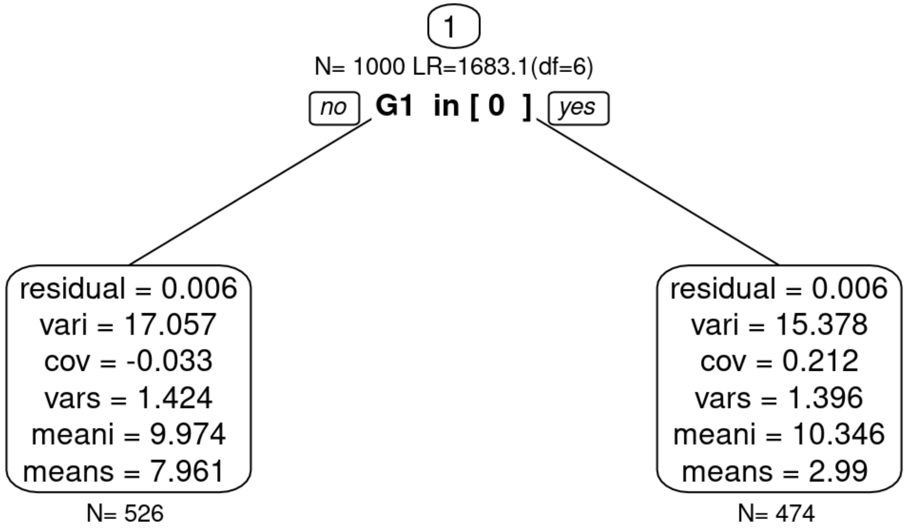


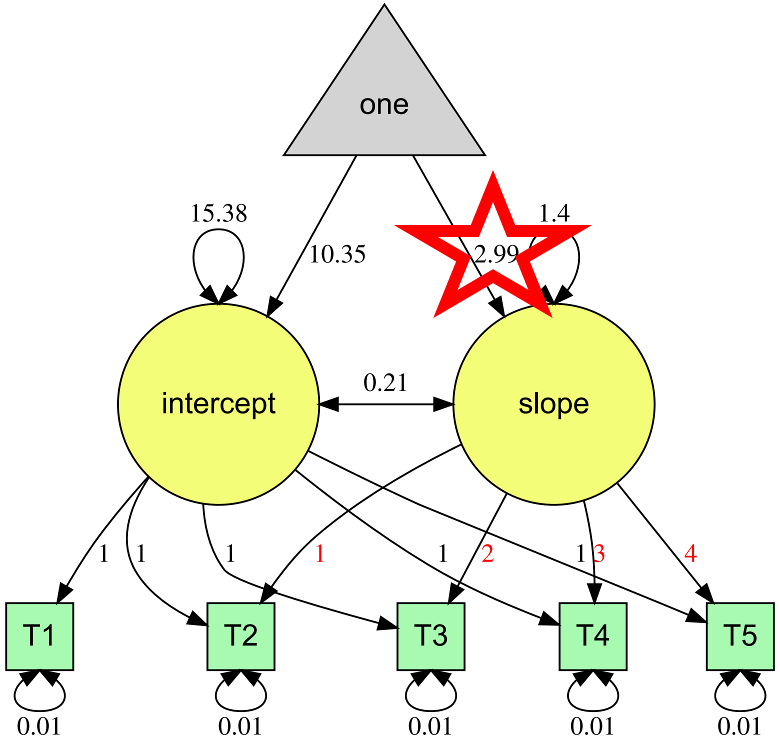
semtree vignette: Getting Started with the semtree package

semtree vignette: Getting Started with the semtree package
Brandmaier et al. (2013B, Equation 4)
Brandmaier & Jacobucci, 2023

IC. The Case of SEM Trees







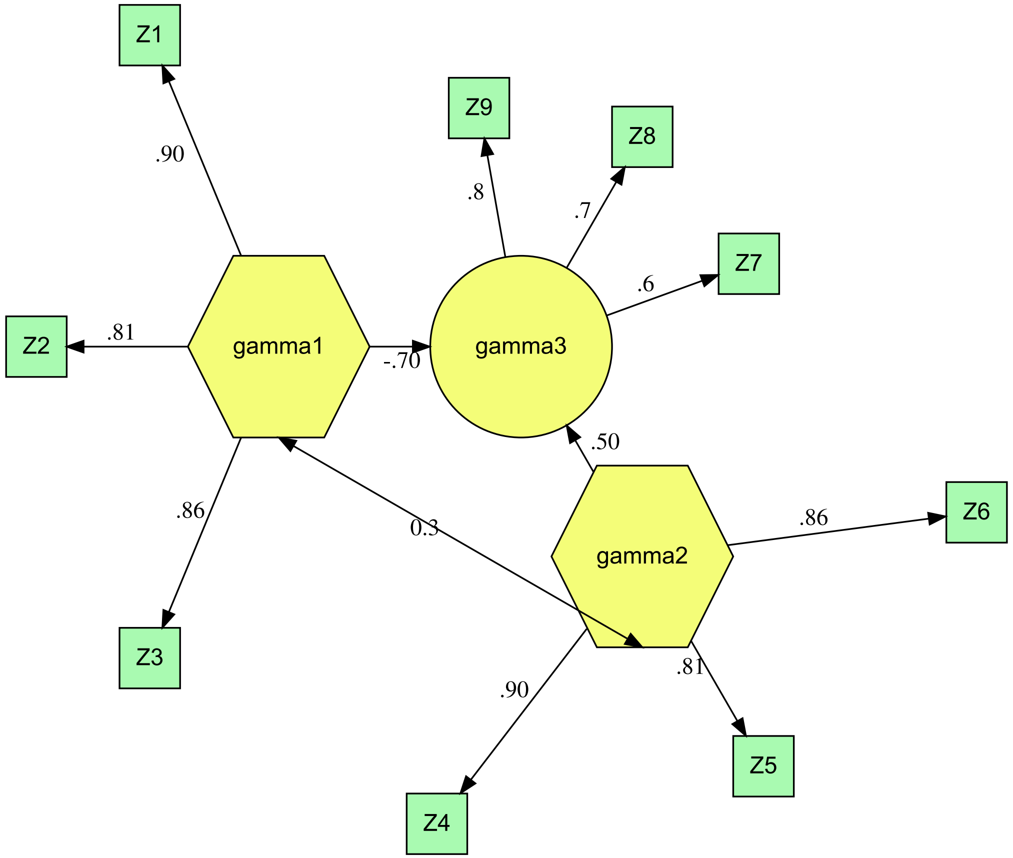
$$\Lambda = \sum_{i=0}^{k} \mathcal{LL}(\hat{\theta}_i | D_i) - \mathcal{LL}(\hat{\theta}_F | D_F)$$

$$\Lambda \sim \chi^2(df = (k-1)m)$$

II. I-GSCA: Integrated-Generalized Structured Component Analysis

- Alternative to CSA

- Combines GSCA and GSCA_m
- Unbiased loadings + paths
- No convergence problems
- Global optimization criterion + FIT statistic



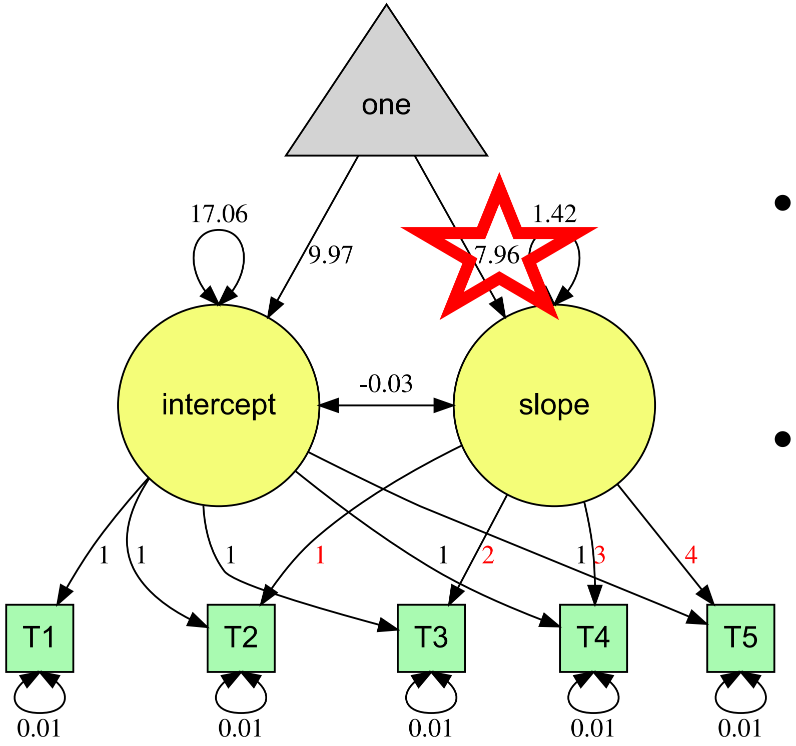
Based on Hwang et al. (2021, Figure 1)

III. I-GSCA Trees

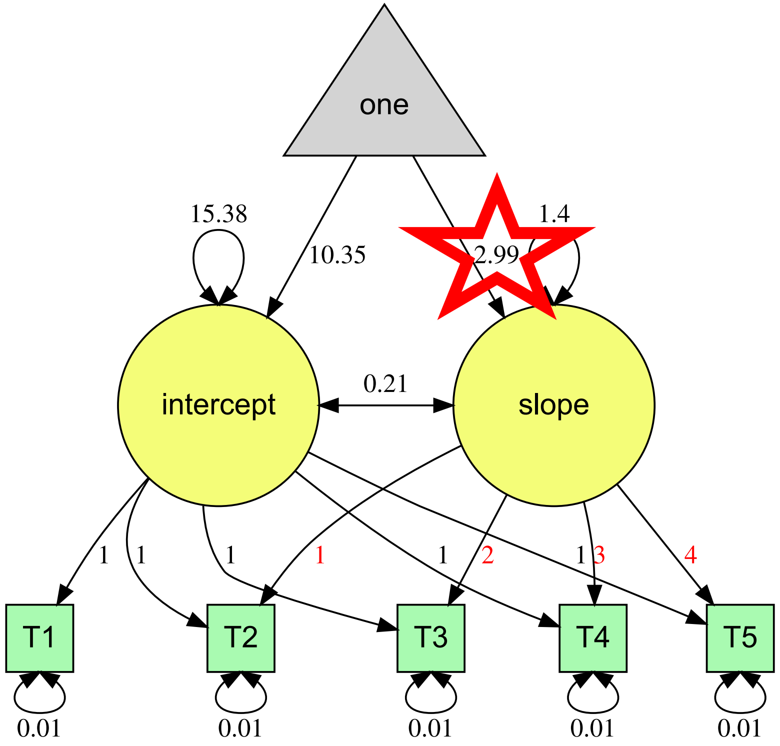
- FIT ~ Proportion of Explained Variance
- Like SEM Trees, choose multi-group models with significantly greater FIT than single group

semtree vignette: Getting Started with the semtree package
Brandmaier & Jacobucci, (2023)
MacCallum et al. (1992)

V. *Pace*: Capitalizing on Chance



- SEM may or may not vary with income, but so what? (c.f., Gelman & Carlin, 2014)
- Technology vs Theoretical purpose
- How many times has the collection of data meaningfully affected psychological theory?



Falsificatory Data Analysis

I. Confirm. Explore! Falsify?
II. Falsificatory Data Analysis' Gambit
III. Related Ideas & Guaranteed Returns

I. Confirm. Explore! Falsify?

- What is out there? Is ____ TRUE?
  - CDA
  - EDA

- *Instead, in FDA*:
  - What do you think is impossible?
  - What do you refuse to believe?
  - What would you need to see to change your mind?
  - When should the data be rejected?

- A theory that says that everything is possible is no theory at all

Similar to Meehl's Description of Popper's Work in 1989 Philosophical Psychology Lectures; terminologically similar, but different from, Gelman's Distinction

II. Falsificatory Data Analysis' Gambit

- Data-driven falsification of causal model: Theory Invariance
  - Predictors and Anti-Predictors
  - Height varies by country… But difference by a factor of 100X?
    - Unit conversion error? Cm to M?
    - Willingness to say that the data is incorrect and must be thrown away
    - Scientific grounds, not statistical

- Advance theories through their *Zone of Impossibility*
  - *Gambit*: Zone of Impossibility is much smaller than Zone of Possibility
  - Claim: Zone of Impossibility != Conditions for Refutation
  - Advantage: Focusing on impossible observations emphasizes link between theory and observation, not theory and statistics

These ideas are similar to those in Becker et al. (2013), but these are more extreme

III. Related Ideas & Guaranteed Returns

- Related Ideas
    - Equivalence Testing
    - Regression Diagnostics
    - Exploratory Data Analysis

- Guaranteed Minimum: Data Quality Checks
    - Number of measurements
    - Unit conversion error
    - Measurement validity
    - *Becker et al. (2013)*

I-GSCA Trees and Falsificatory Data Analysis

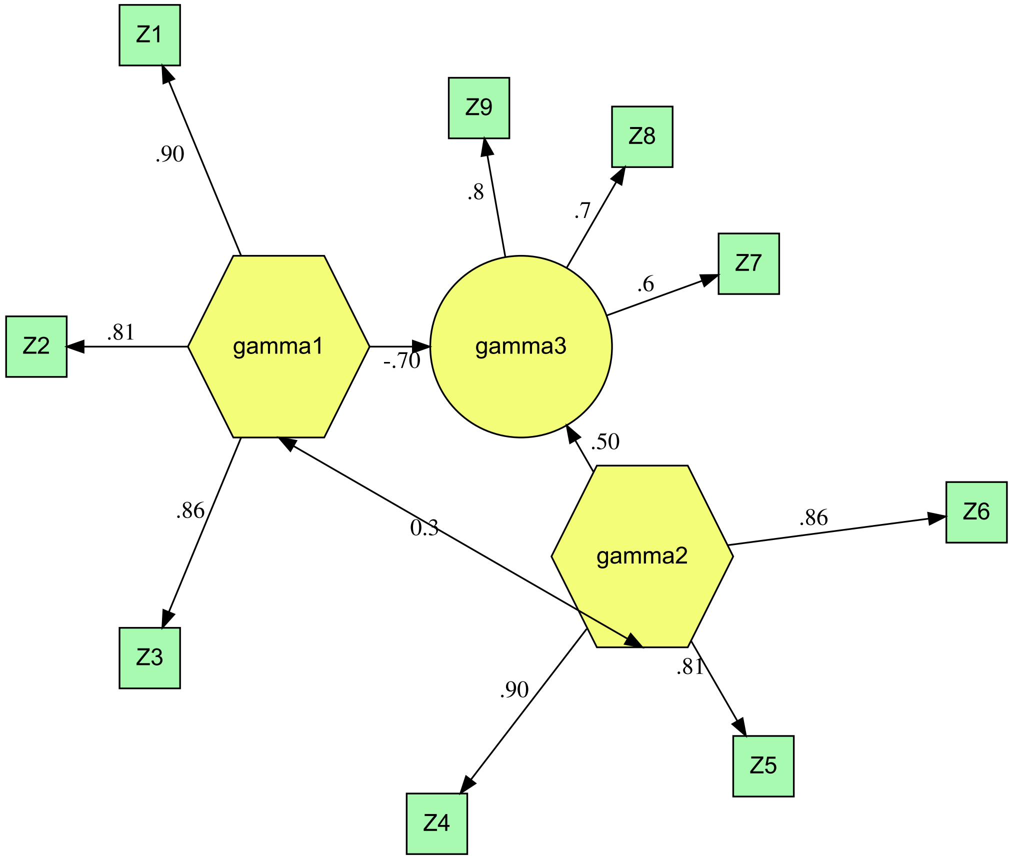
1. Monte Carlo Simulation
2. How well?
3. Future Directions

I. Monte Carlo Simulation

- Our model should not vary much based on location
  - Anti-Predictor: Location

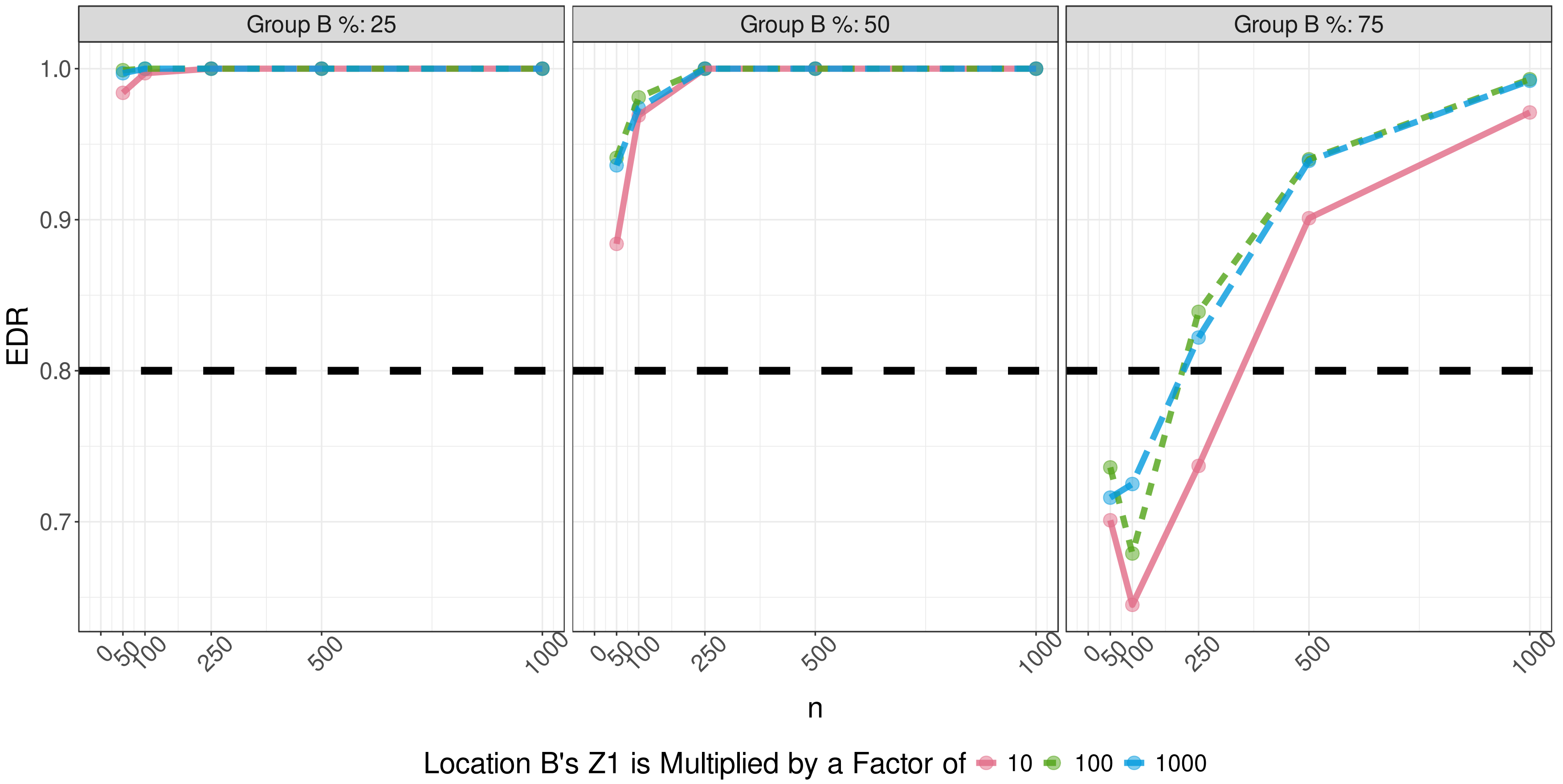
- BUT, data entry error on Z1!

- Generate MVN ~ standardized data
- Random assignment of location
- +5 all indicators
- Multiply Z1 by 1, 10, 100 or 1000



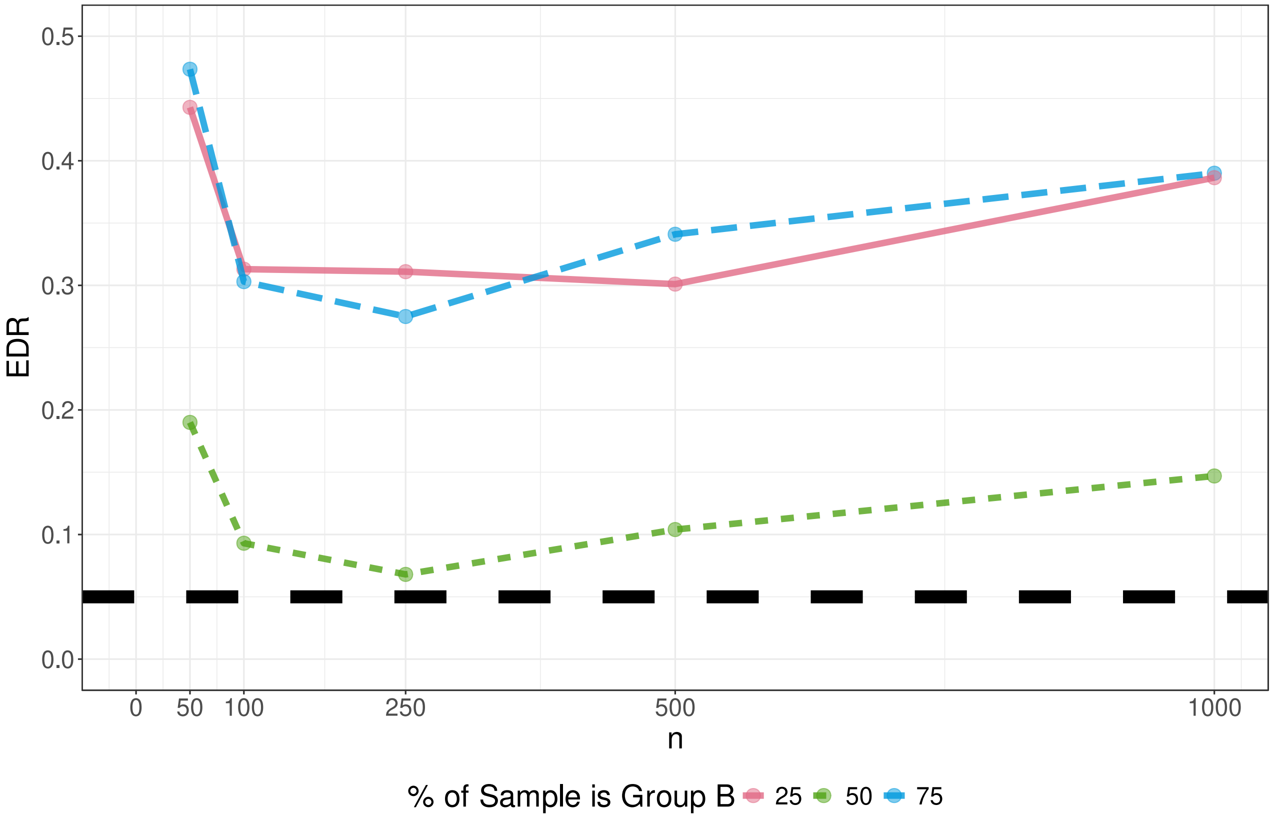
Based on Hwang et al. (2021, Figure 1)

IIA. How well? Power



- Number of digits in unstandardized measurement?
- Stratified Bootstrap?

IIB. How well? Type 1 Error



- Better testing techniques?
- Is the use of significance tests incompatible with FDA?

III. Future Directions – mtruong@yorku.ca

- Falsificatory Data Analysis?
  - Philosophy of Science justifications?
  - Advantages and disadvantages
  - Likely requires counter-induction to be useful
    - *Feyerabend, 2020*
- IGSCA-Trees?
  - Complete implementation in R cSEM Package
    - Rademaker and Schuberth, 2020
  - More extensive MCS, compare with CSA, vary number of digits in unstandardized data
  - Random Forests?
    - *Brandmaier et al., 2016*
  - Better ways of group comparison? Stratified Bootstrapping?
  - Constrained Splits?
    - *Brandmaier et al., 2013*
  - M-Fluctuation Test? Un-Biased Splits?
    - *Hothorn et al., 2006; Strobl et al., 2007; Zeileis & Hornik, 2007*
  - $N < J$: Regularization? Bayes?
    - *Choi & Hwang, 2020; Hwang & Takane, 2014*

Special Thanks

Dr. Florian Schuberth, University of Twente
Dr. Heungsun Hwang, McGill University
Dr. R. Phil Chalmers, York University

Many anonymous reviewers

Thanks does not imply endorsement

References

Becker, J.-M., Rai, A., Ringle, C. M., & Völckner, F. (2013). Discovering Unobserved Heterogeneity in Structural Equation Models to Avert Validity Threats. MIS Quarterly, 37(3), 665–694. https://doi.org/10.25300/MISQ/2013/37.3.01

Brandmaier, A. M., & Jacobucci, R. (2023). Machine-Learning Approaches to Structural Equation Modeling. In R. H. Hoyle (Ed.), Handbook of structural equation modeling (2nd Edition). Guilford Press.

Brandmaier, A. M., Oertzen, T. V., McARDLE, J. J., & Lindenberger, U. (2013). Exploratory Data Mining with Structural Equation Model Trees. In Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences. Routledge.

Brandmaier, A. M., Prindle, J. J., Arnold, M., & Lissa, C. J. V. (2022). semtree: Recursive Partitioning for Structural Equation Models (0.9.18) [Computer software]. https://CRAN.R-project.org/package=semtree

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. Psychological Methods, 21(4), 566–582. https://doi.org/10.1037/met0000090

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. Psychological Methods, 18(1), 71–86. https://doi.org/10.1037/a0030001

Cho, G., & Choi, J. Y. (2020). An empirical comparison of generalized structured component analysis and partial least squares path modeling under variance-based structural equation models. Behaviormetrika, 47(1), 243–272. https://doi.org/10.1007/s41237-019-00098-0

Choi, J. Y., & Hwang, H. (2020). Bayesian generalized structured component analysis. British Journal of Mathematical and Statistical Psychology, 73(2), 347–373. https://doi.org/10.1111/bmsp.12166

Feyerabend, P. (2020). Against Method: Outline of an Anarchistic Theory of Knowledge. Verso Books.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science, 9(6), 641–651. https://doi.org/10.1177/1745691614551642

Getting Started with the semtree package. (n.d.). Retrieved June 11, 2023, from https://brandmaier.github.io/semtree/articles/getting-started.html

Henninger, M., Debelak, R., Rothacher, Y., & Strobl, C. (2023). Interpretable machine learning for psychological research: Opportunities and pitfalls. Psychological Methods, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000560

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651–674. https://doi.org/10.1198/106186006X133933

Hwang, H., Cho, G., Jung, K., Falk, C. F., Flake, J. K., Jin, M. J., & Lee, S. H. (2021). An approach to structural equation modeling with both factors and components: Integrated generalized structured component analysis. Psychological Methods, 26, 273–294. https://doi.org/10.1037/met0000336

Hwang, H., & Takane, Y. (2014). Generalized structured component analysis: A component-based approach to structural equation modeling. CRC Press, Taylor & Francis Group.

Hwang, H., Takane, Y., & Jung, K. (2017). Generalized Structured Component Analysis with Uniqueness Terms for Accommodating Measurement Error. Frontiers in Psychology, 8. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02137

Jones, P. J., Mair, P., Simon, T., & Zeileis, A. (2020). Network Trees: A Method for Recursively Partitioning Covariance Structures. Psychometrika, 85(4), 926–945. https://doi.org/10.1007/s11336-020-09731-4

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. Proceedings of the National Academy of Sciences, 111(24), 8782–8787. https://doi.org/10.1073/pnas.1402786111

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. Psychological Bulletin, 111(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

McElreath, R. (2020). Statistical rethinking: A Bayesian course with examples in R and Stan (2nd ed.). Taylor and Francis, CRC Press.

Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books.

Rademaker, M. E., & Schuberth, F. (2020). cSEM: Composite-based structural equation modeling [Computer software]. https://floschuberth.github.io/cSEM/

Robinaugh, D., Haslbeck, J., Waldorp, L., Kossakowski, J., Fried, E. I., Millner, A., McNally, R. J., Ryan, O., Ron, J. de, Maas, H. van der, Nes, E. H. van, Scheffer, M., Kendler, K. S., & Borsboom, D. (2019). Advancing the Network Theory of Mental Disorders: A Computational Model of Panic Disorder. OSF. https://doi.org/10.31234/osf.io/km37w

Sarma, A., Kale, A., Moon, M., Taback, N., Chevalier, F., Hullman, J., & Kay, M. (2021). multiverse: Multiplexing alternative data analyses in R notebooks (version 0.6.1). OSF Preprints. https://github.com/MUCollective/multiverse

Shmueli, G. (2010). To Explain or to Predict? Statistical Science, 25(3), 289–310. https://doi.org/10.1214/10-STS330

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. Nature Human Behaviour, 4(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8(1), 25. https://doi.org/10.1186/1471-2105-8-25

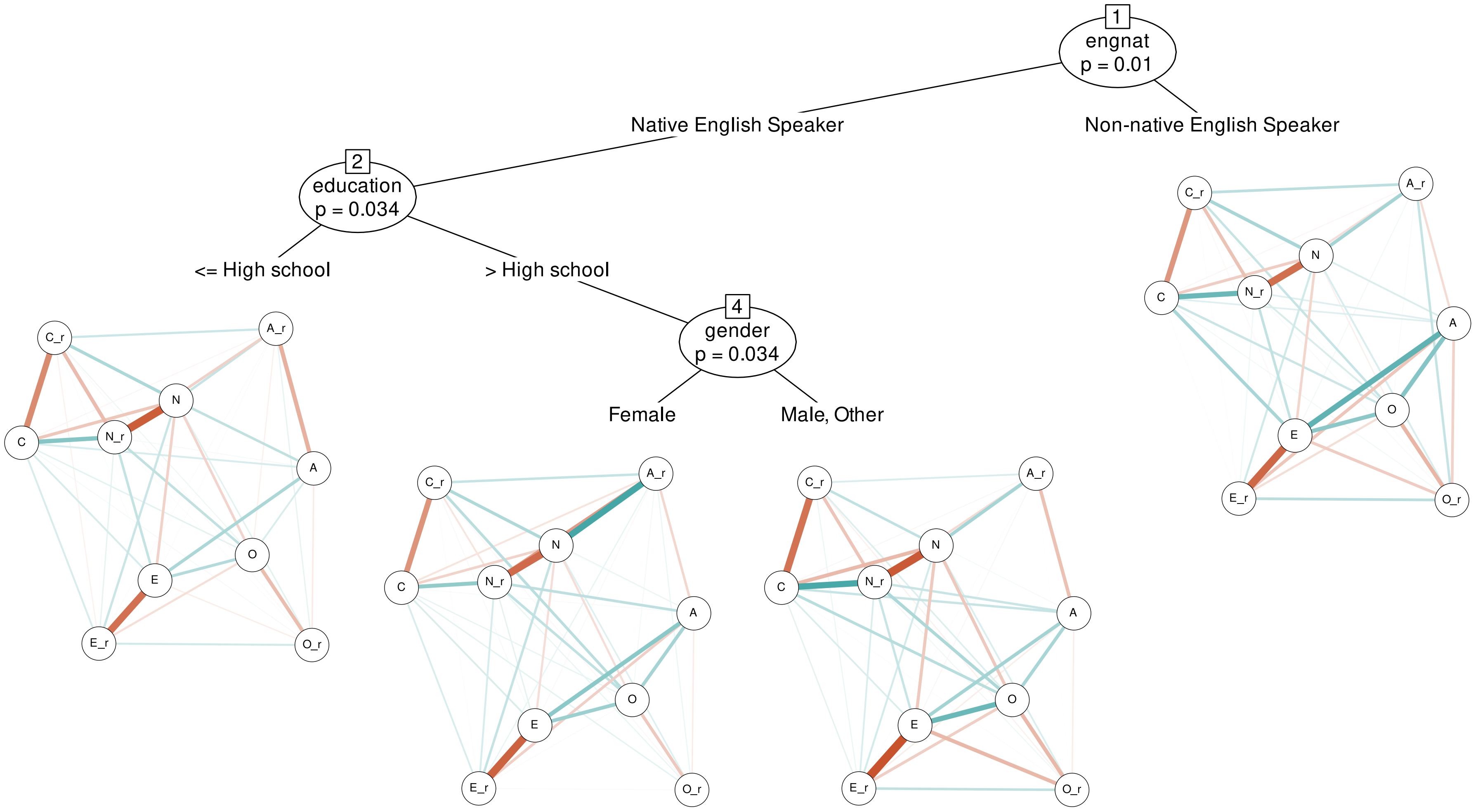
Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science, 12(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. Statistica Neerlandica, 61(4), 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x

Data Generation Procedure

- Please see Cho and Choi (2020) and Hwang et al. (2021, Appendix B)
- Composite
  - Specify Var-Cov Mx of Indicators
  - Use both largest eigenvalue and parts of Var-Cov Mx to get Weights
  - Use Weights and Var-Cov Mx to get Loadings
- Factor
  - Use specified loadings matrix to get variance of residuals
- Construct Var-Cov Mx
  - Use path-coefficients, and construct covariances to derive Var-Cov Mx
- Population Var-Cov Mx for Indicators
  - Use block-diagonalized loadings Mx, Construct Var-Cov Mx and residual Mxs to get pop var-cov Mx

Results Depend on... Research Assistant???



Jones et al. (2020) and Ten Item Personality Questionnaire Data from Open Psychometrics