Latent Variable Forests:

Estimating Valid Latent Variable Scores From Conditionally Causal Models

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

IHF Bavarian State Institute for Higher Education Research and Planning

Franz Classe[1] and Christoph Kern[2], [1]IHF, LMU Munich and [2]University of Mannheim

Objective

We develop a ***latent variable forest*** **(LV Forest) algorithm for the estimation of latent variable scores** from conditionally causal models with one or more latent variables. LV Forest establishes **conditional causality in *structural equation models* (SEM)** with ordinal and/or numerical response variables. Through parametric model restrictions paired with a non-parametric tree-based machine learning approach, LV Forest estimates latent variables scores that fulfill the main criteria for construct validity. Latent Variable Scores are used to scale individuals on a single construct (e.g. personality trait or ability). However, a major problem is unfairness and bias of such scores, especially with respect to social minorities [4]. We tackle these problems through the estimation valid latent variable scores.

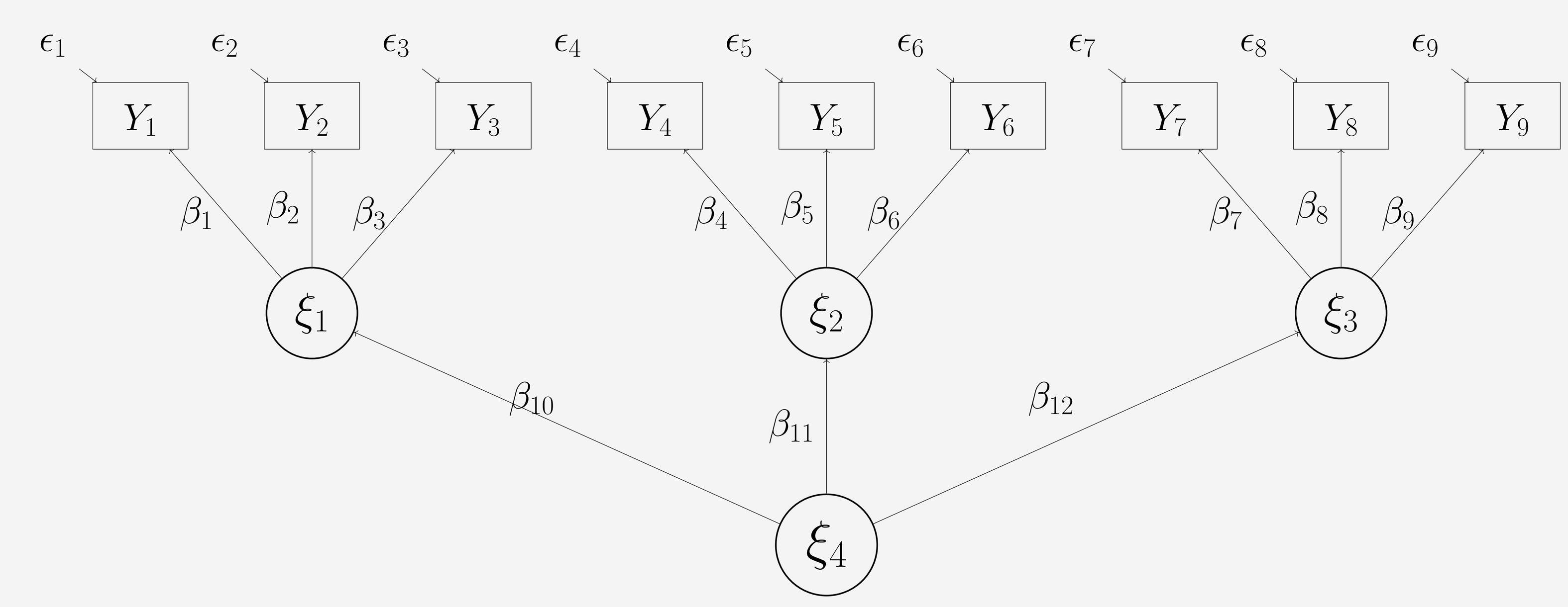
Causality in SEM

Causal relationships are expressed in the form of deterministic, structural relationships. This means that a set of latent variables $\boldsymbol{\xi}_i$ is the *only* immediate cause of Y_i.
In reality, however, there is randomness coming from disturbances due to unmodeled covariates $\boldsymbol{\zeta}_i$. In SEM, this means:

$$E(Y_i \,|\, \boldsymbol{\xi},\ \boldsymbol{\zeta}_i) = \boldsymbol{\beta}_i'\boldsymbol{\xi} + \boldsymbol{\zeta}_i, \quad \text{for numerical } Y_i$$
$$P(Y_i \geq k_i \,|\, \boldsymbol{\xi},\ \boldsymbol{\zeta}_i) = \Phi(\boldsymbol{\beta}_i'\boldsymbol{\xi} - \alpha_{ik}) + \boldsymbol{\zeta}_i, \quad \text{for categorical } Y_i$$
$$\forall\, i = 1, \ldots, m\ k = 1, \ldots, l.$$

Causality is established by three conditions [2]: Firstly, **Pseudo isolation** means that no confounding covariate outside the model is associated with the latent variables, such that $Cov(\boldsymbol{\xi}, \boldsymbol{\zeta}_i) = 0$. Secondly, **Association** means the latent variable is associated with the response variables, such that $\beta_i \neq 0\ \forall\, i = 1, \ldots, m$. Thirdly, **Direction of influence** means that the temporal sequence of the variables must be logical. We test for these causality conditions in LV Forest.

Simulated SEM



Algorithm

Step 1: `lvforest.train`
Result: Trained Model: List of best terminal nodes from forest
while *ntree* $\leq$ *forest size* **do**
- Build tree: Partition data set to reduce parameter heterogeneity [3];
- Exclude subgroups based on:
  - model fit
  - size of β
  - parameter stability [6];
- Save decision rules and parameter estimates;

end

Step 2: `lvforest.predict`
Result: Valid Latent Variable Scores
while *ntree* $\leq$ *forest size after training* **do**
while *nnode* $\leq$ *no. nodes after training* **do**
- Use subgroups from training to predict latent variable scores;
- Exclude subgroups based on test for *pseudo-randomization*;

end
- Compute mean of LV scores over remaining nodes;

end

Application

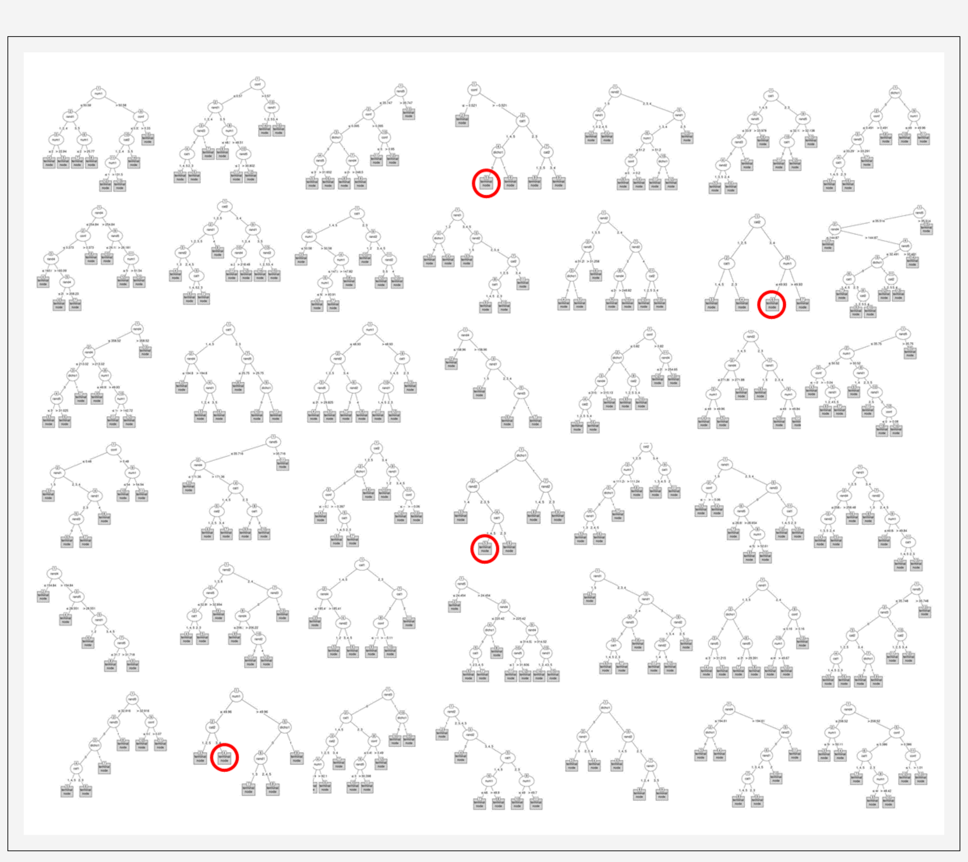
Residuals correlated with $\boldsymbol{\xi}$

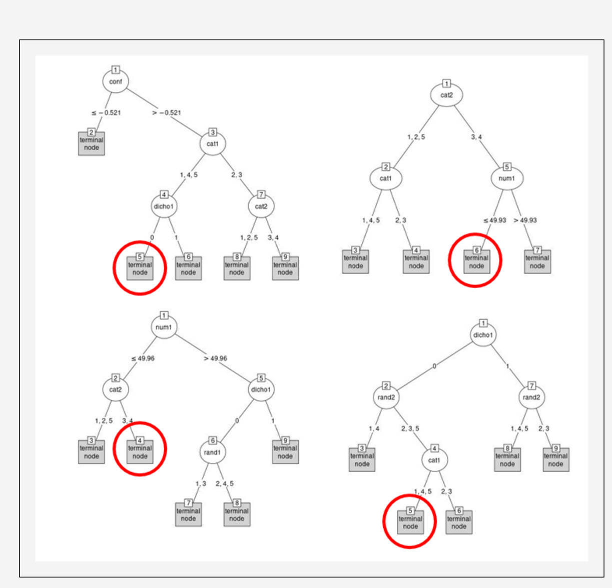
$\boldsymbol{Y}$ not sufficiently associated with $\boldsymbol{\xi}$

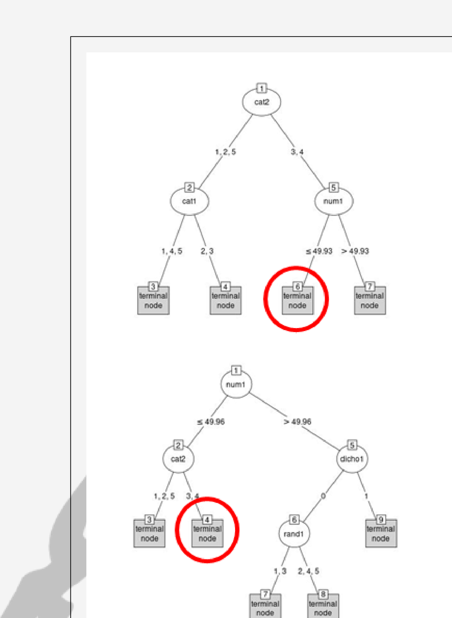
Instable parameters

Confounded relations

Data → `lvforest.train` → Trained Model → `lvforest.predict` → Valid Latent Variable Scores







	id	tree48	tree85	mean
874	1905	-2.2302828	-2.0917307	-2.1610068
3746	1697	-1.8054496	-1.6870492	-1.7462494
3881	1766	-1.7982356	-1.6667754	-1.7325055
1210	1201	-1.6308783	-1.4931119	-1.5619951
2052	1850	-1.6212129	-1.4812004	-1.5512066
602	1485	-1.5977508	-1.4430702	-1.5204105
2407	1613	-1.5548364	-1.4301529	-1.4924947
1648	1694	-1.4915598	-1.3995411	-1.4455505
3406	1764	-1.4859940	-1.3507287	-1.4183614
2846	1244	-1.4722057	-1.3507859	-1.4114958
883	1701	-1.4214872	-1.3055205	-1.3635038
2722	1928	-1.4115237	-1.2983988	-1.3549612
2948	1582	-1.3673506	-1.2803734	-1.3238620
592	1363	-1.3728592	-1.2563205	-1.3145899
567	1976	-1.3369690	-1.2245041	-1.2807365
2057	1535	-1.3092645	-1.1915843	-1.2504244
941	1922	-1.2884267	-1.1832917	-1.2358592
770	1748	-1.2842958	-1.1788328	-1.2315643

Discussion

Since tree-based methods have successfully been applied to account for parameter heterogeneity [1, 5], we utilized the machine learning perspective for handling complex subgroup structures in the population. In psychological assessment, *bias* refers to systematically under- or overestimating of personality traits or abilities [4]. Especially cultural bias has been a polarizing issue for many years. The controversy lies in the question whether differences between specific subgroups are based on real differences in ability levels or on different cognitive structures requiring different test characteristics, i.e. test bias. We argue that pseudo-isolation, and therefore **causality in SEMs, is only possible** if there are **no systematic differences in a latent ability or trait with respect to variables outside the latent variable model**. Thus, if systematic differences between groups regarding the latent variable are not part of the assumed model, they are attributable to test bias. Thus, a tool like LV Forest that measures fair and unbiased scores can serve as an relevant contribution to the discussion about bias in psychological assessment.

Contact:
- Email: classe@ihf.bayern.de
- Phone: +49 (0)89 21234-410
- Github: **github.com/chkern/lvforest**

References

[1] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7353–7360 (cit. on p. 1).

[2] Kenneth A Bollen. *Structural equations with latent variables.* Vol. 210. John Wiley & Sons, 1989 (cit. on p. 1).

[3] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3 (2006), pp. 651–674 (cit. on p. 1).

[4] Cecil R Reynolds, Robert A Altmann, and Daniel N Allen. "The problem of bias in psychological assessment". In: *Mastering Modern Psychological Testing.* Springer, 2021, pp. 573–613 (cit. on p. 1).

[5] Carolin Strobl, Julia Kopf, and Achim Zeileis. "Rasch trees: A new method for detecting differential item functioning in the Rasch model". In: *Psychometrika* 80.2 (2015), pp. 289–316 (cit. on p. 1).

[6] Achim Zeileis and Kurt Hornik. "Generalized M-fluctuation tests for parameter instability". In: *Statistica Neerlandica* 61.4 (2007), pp. 488–508 (cit. on p. 1).