Survey Scale Forests

DAGStat Conference 2022

Franz Classe [1] Christoph Kern [2]

[1]Bavarian State Institute for Higher Education Research and Planning

[2]University of Mannheim

April 1, 2022

Latent Variable Modeling

Definition:

Latent Variables are **unobservable phenomena** like

- creativity, (Jauk et al., 2014)
- social anxiety, depression, (Prenoveau et al., 2011)
- psychopathic personality, (Drislane & Patrick, 2017)
- self-leadership. (Furtner et al., 2015)

measured as theoretical **constructs** through research tools like a questionnaire in a survey.

Latent Variable Modeling

Definition:

Latent Variables are **unobservable phenomena** like

- creativity, (Jauk et al., 2014)
- social anxiety, depression, (Prenoveau et al., 2011)
- psychopathic personality, (Drislane & Patrick, 2017)
- self-leadership. (Furtner et al., 2015)

measured as theoretical **constructs** through research tools like a questionnaire in a survey.

→ Estimation in the form of **latent variable scores**.

Latent Variable Modeling

Definition:

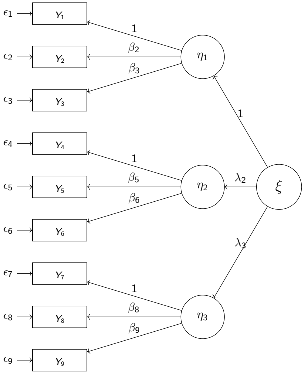
Latent Variables are **unobservable phenomena** like

- creativity, (Jauk et al., 2014)
- social anxiety, depression, (Prenoveau et al., 2011)
- psychopathic personality, (Drislane & Patrick, 2017)
- self-leadership. (Furtner et al., 2015)

measured as theoretical **constructs** through research tools like a questionnaire in a survey.

$\rightarrow$ Estimation in the form of **latent variable scores**.

Example:



Causality in Latent Variable Models

Causal relationships expressed in the form of deterministic, *structural relationships* (Bollen, 1989; Pearl, 2009):

$$Y_i = f_i(\xi_i, \zeta_i) \;\; \forall \;\; i = 1, \ldots, m. \qquad (1)$$

Causality in Latent Variable Models

Causal relationships expressed in the form of deterministic, *structural relationships* (Bollen, 1989; Pearl, 2009):

$$Y_i = f_i(\xi_i, \zeta_i) \;\; \forall \;\; i = 1, \ldots, m. \tag{1}$$

Randomness *only* because of unmodeled covariates ζ_i.

Causality in Latent Variable Models

Causal relationships expressed in the form of deterministic, *structural relationships* (Bollen, 1989; Pearl, 2009):

$$Y_i = f_i(\xi_i, \zeta_i) \ \forall \ i = 1, \ldots, m. \tag{1}$$

Randomness *only* because of unmodeled covariates ζ_i.

Why is causality in latent variable models important?

Validity is *"the **magnitude** of the direct **structural relation**"* between latent variable and observed response. (Bollen, 1989)

Causality in Latent Variable Models

Causal relationships expressed in the form of deterministic, *structural relationships* (Bollen, 1989; Pearl, 2009):

$$Y_i = f_i(\xi_i, \zeta_i) \;\; \forall \;\; i = 1, \ldots, m. \tag{1}$$

Randomness *only* because of unmodeled covariates ζ_i.

Why is causality in latent variable models important?

Validity is *"the **magnitude** of the direct **structural relation**"* between latent variable and observed response. (Bollen, 1989)

Conditions for causality $\Rightarrow$ conditions for validity of LV scores:

Causality in Latent Variable Models

Causal relationships expressed in the form of deterministic, *structural relationships* (Bollen, 1989; Pearl, 2009):

$$Y_i = f_i(\xi_i, \zeta_i) \ \forall \ i = 1, \ldots, m. \tag{1}$$

Randomness *only* because of unmodeled covariates ζ_i.

Why is causality in latent variable models important?

Validity is *"the **magnitude** of the direct **structural relation**"* between latent variable and observed response. (Bollen, 1989)

Conditions for causality $\Rightarrow$ conditions for validity of LV scores:

- Isolation:
    - Impossible to attain $\rightarrow$ pseudo-isolation: $Cov(\xi_i, \zeta_i) = 0 \ \forall \ i = 1, \ldots, m.$

Causality in Latent Variable Models

Causal relationships expressed in the form of deterministic, *structural relationships* (Bollen, 1989; Pearl, 2009):

$$Y_i = f_i(\xi_i, \zeta_i) \ \forall \ i = 1, \dots, m. \tag{1}$$

Randomness *only* because of unmodeled covariates ζ_i.

Why is causality in latent variable models important?

Validity is *"the **magnitude** of the direct **structural relation**"* between latent variable and observed response. (Bollen, 1989)

Conditions for causality $\Rightarrow$ conditions for validity of LV scores:

- Isolation:
  - Impossible to attain $\rightarrow$ pseudo-isolation: $Cov(\xi_i, \zeta_i) = 0 \ \forall \ i = 1, \dots, m.$
- Association

Causality in Latent Variable Models

Causal relationships expressed in the form of deterministic, *structural relationships* (Bollen, 1989; Pearl, 2009):

$$Y_i = f_i(\xi_i, \zeta_i) \;\; \forall \;\; i = 1, \ldots, m. \tag{1}$$

Randomness *only* because of unmodeled covariates ζ_i.

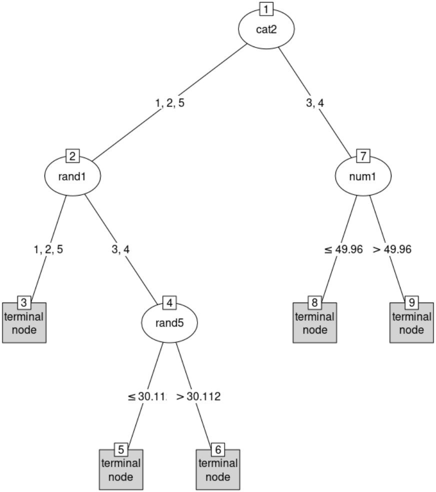
Why is causality in latent variable models important?
Validity is *"the **magnitude** of the direct **structural relation**"* between latent variable and observed response. (Bollen, 1989)

Conditions for causality $\Rightarrow$ conditions for validity of LV scores:

- Isolation:
  - Impossible to attain $\rightarrow$ pseudo-isolation: $Cov(\xi_i, \zeta_i) = 0 \;\; \forall \;\; i = 1, \ldots, m.$
- Association
- Direction of Influence

Goal

How can we use latent variable modeling together with **machine learning techniques** to **detect subgroups** in which the assumed model is **not conditionally causal** and exclude them to be able to **estimate valid latent variable scores**?



Conditional Inference Tree in SC Forest

- Reducing parameter heterogeneity by using *fitted model scores* as outcome:

$$\psi(y_j, \theta) = \left(\frac{\partial F_{ML}(y_j, \theta)}{\partial \theta_1}, \ldots, \frac{\partial F_{ML}(y_j, \theta)}{\partial \theta_k} \right), \ \forall\, j = 1, \ldots, n. \quad (2)$$

- Unbiased selection of covariate used for splitting Z_r^*:
  - permutation-based association test between covariates and outcome $\rightarrow$ scale of outcome & covariate irrelevant for the test result!
  - Test $H_0^r : D(\psi | Z_r) = D(\psi)$ for all covariates $r = 1, \ldots, R$, so that global hypothesis test is $\bigcap_{r=1}^{R} H_0^r$
  - If the global hypothesis not rejected $\rightarrow$ algorithm stops splitting

Survey Scale Forest

Train model:

1. Partition data set to reduce parameter heterogeneity (*tree*) using *double sampling* (Athey & Imbens, 2016)

Survey Scale Forest

Train model:

1. Partition data set to reduce parameter heterogeneity (*tree*) using *double sampling* (Athey & Imbens, 2016)
2. Repeat process with variation at every iteration (*forest*) using *random split selection* (Breiman, 2001)

Survey Scale Forest

Train model:

1. Partition data set to reduce parameter heterogeneity (*tree*) using *double sampling* (Athey & Imbens, 2016)
2. Repeat process with variation at every iteration (*forest*) using *random split selection* (Breiman, 2001)
3. Exclude subgroups based on:
   - model fit ($\rightarrow$ randomness of errors)
   - association between $\boldsymbol{Y}$ and ξ
   - parameter stability with respect to covariates (Zeileis & Hornik, 2007)

Survey Scale Forest

Train model:

1. Partition data set to reduce parameter heterogeneity (*tree*) using *double sampling* (Athey & Imbens, 2016)
2. Repeat process with variation at every iteration (*forest*) using *random split selection* (Breiman, 2001)
3. Exclude subgroups based on:
   - model fit ($\rightarrow$ randomness of errors)
   - association between $\boldsymbol{Y}$ and $\boldsymbol{\xi}$
   - parameter stability with respect to covariates (Zeileis & Hornik, 2007)
4. Save decision rules and parameter estimates for remaining models.

Predict scores:

1. Use subgroups from training to predict latent variable scores

Survey Scale Forest

Train model:

1. Partition data set to reduce parameter heterogeneity (*tree*) using *double sampling* (Athey & Imbens, 2016)
2. Repeat process with variation at every iteration (*forest*) using *random split selection* (Breiman, 2001)
3. Exclude subgroups based on:
   - model fit ($\rightarrow$ randomness of errors)
   - association between $\boldsymbol{Y}$ and $\boldsymbol{\xi}$
   - parameter stability with respect to covariates (Zeileis & Hornik, 2007)
4. Save decision rules and parameter estimates for remaining models.

Predict scores:

1. Use subgroups from training to predict latent variable scores
2. Exclude subgroups based on test for *confoundedness* of relations in model (Steyer & Nagel, n.d.)

Survey Scale Forest

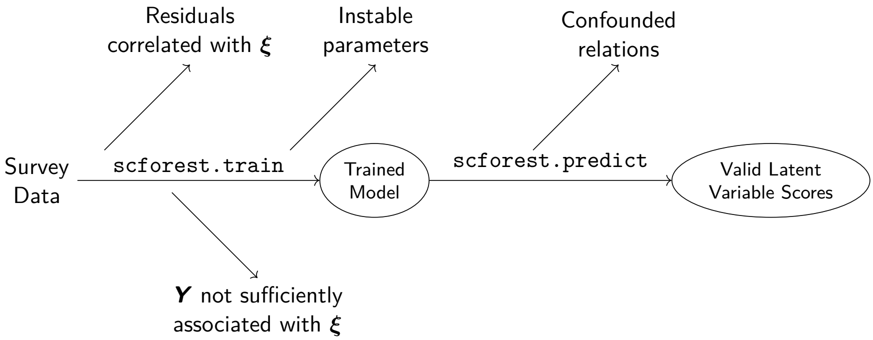
Train model:

1. Partition data set to reduce parameter heterogeneity (*tree*) using *double sampling* (Athey & Imbens, 2016)
2. Repeat process with variation at every iteration (*forest*) using *random split selection* (Breiman, 2001)
3. Exclude subgroups based on:
   - model fit ($\rightarrow$ randomness of errors)
   - association between $\boldsymbol{Y}$ and ξ
   - parameter stability with respect to covariates (Zeileis & Hornik, 2007)
4. Save decision rules and parameter estimates for remaining models.

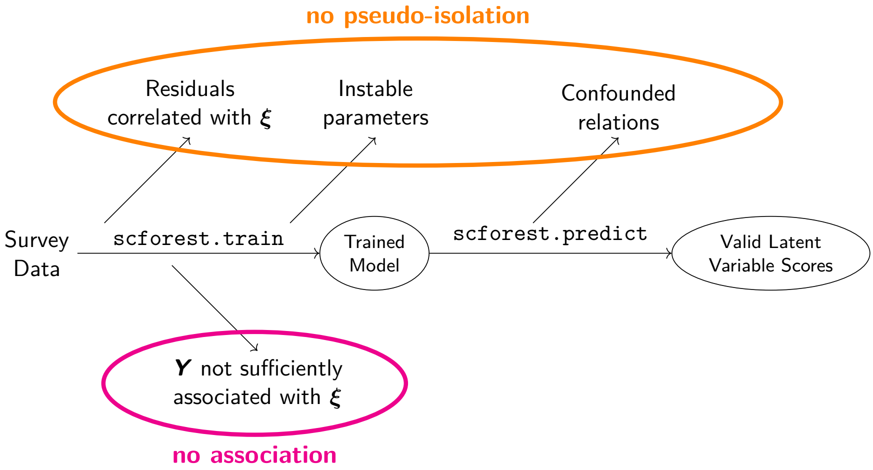
Predict scores:

1. Use subgroups from training to predict latent variable scores
2. Exclude subgroups based on test for *confoundedness* of relations in model (Steyer & Nagel, n.d.)
3. Compile results across all iterations

Survey Scale Forest in Practice



Survey Scale Forest in Practice



Conclusion

Survey Scale Forest detects two conditions for non-causality,

- lack of pseudo-isolation ($\rightarrow$ confoundedness of relations in model),
- lack of association between $\boldsymbol{Y}$ and $\boldsymbol{\xi}$,

and excludes all subgroups that fulfill these conditions. This way, predicted latent variable scores fulfill criteria for validity although construct may not generally be valid.

References

American Psychological Association. (2014). Standards for psychological and educational testing. *New York: American Educational Research Association*.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Drislane, L. E., & Patrick, C. J. (2017). Integrating alternative conceptions of psychopathic personality: A latent variable model of triarchic psychopathy constructs. *Journal of personality disorders*, *31*(1), 110–132.

Furtner, M. R., Rauthmann, J. F., & Sachse, P. (2015). Unique self-leadership: A bifactor model approach. *Leadership*, *11*(1), 105–125.

Jauk, E., Benedek, M., & Neubauer, A. C. (2014). The road to creative achievement: A latent variable model of ability and personality predictors. *European journal of personality*, *28*(1), 95–105.

Pearl, J. (2009). *Causality*. Cambridge university press.

Prenoveau, J. M., Craske, M. G., Zinbarg, R. E., Mineka, S., Rose, R. D., & Griffith, J. W. (2011). Are anxiety and depression just as stable as personality during late adolescence? results from a three-year longitudinal latent variable study.. *Journal of Abnormal Psychology*, *120*(4), 832.

Steyer, R., & Nagel, W. (n.d.). *Probability and causality: Volume i: Causal total effects* [Book in preparation.].

Zeileis, A., & Hornik, K. (2007). Generalized m-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508.

Any questions?

- Contact: classe@ihf.bayern.de
- Want to try the method?:
  R-package `scforest` on GitHub
  github.com/chkern/scforest

Validity

Validity is...

> *"...the degree to which evidence and theory support the interpretations of test scores".* (APA, 2014)

Need for researchers to find **evidence to support proposed interpretation** of item responses.

Four sources of evidence for construct validity (APA, 2014):

Appropriate

- test content,
- internal structure,
- response processes,
- relation to other variables.

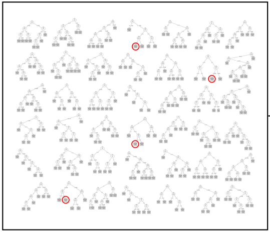
$\Big\}\longrightarrow$ Causal Model

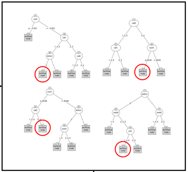
Validity is *"the magnitude of the direct structural relation"* between latent variable and observed response. (Bollen, 1989)

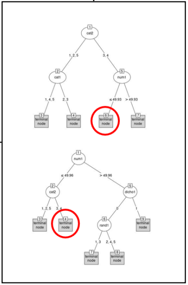
Decision Trees

- Non-parametric machine-learning method
- Recursively partitions covariate space over $\boldsymbol{Z} = \{Z_1, \ldots, Z_R\}$ into set of terminal nodes (*leaves*)
- Reduce outcome heterogeneity
- Usually built on training data and used to predict outcome in test data

Simulation







	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