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,  \cite{Jauk2014}
- social anxiety, depression,  \cite{Prenoveau2011}
- psychopathic personality,  \cite{Drislane2017}
- self-leadership.  \cite{Furtner2015}

**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**.
**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**.
Causality in Latent Variable Models

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

\[ Y_i = f_i(\xi_i, \zeta_i) \quad \forall \quad i = 1, \ldots, m. \] (1)
Causality in Latent Variable Models

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

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

Randomness *only* because of unmodeled covariates \( \zeta_i \).
Causality in Latent Variable Models

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

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

Randomness only because of unmodeled covariates \( \zeta_i \).

Why is causality in latent variable models important?

Validity is “the \textit{magnitude} of the \textit{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):

\[ Y_i = f_i(\xi_i, \zeta_i) \quad \forall \quad i = 1, \ldots, m. \]  

(1)

Randomness only because of unmodeled covariates \( \zeta_i \).

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):

\[ Y_i = f_i(\xi_i, \zeta_i) \quad \forall \quad i = 1, \ldots, m. \]  

(1)

Randomness only because of unmodeled covariates \( \zeta_i \).

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:
    \[ \text{Cov}(\xi_i, \zeta_i) = 0 \quad \forall \quad i = 1, \ldots, m. \]
Causality in Latent Variable Models

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

\[ Y_i = f_i(\xi_i, \zeta_i) \quad \forall \quad i = 1, \ldots, m. \]  

(1)

Randomness only because of unmodeled covariates \( \zeta_i \).

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:
    \[ \text{Cov}(\xi_i, \zeta_i) = 0 \quad \forall \quad i = 1, \ldots, m. \]
Causality in Latent Variable Models

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

\[ Y_i = f_i(\xi_i, \zeta_i) \quad \forall \quad i = 1, \ldots, m. \]  

Randomness *only* because of unmodeled covariates \( \zeta_i \).

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:
    \[ \text{Cov}(\xi_i, \zeta_i) = 0 \quad \forall \quad i = 1, \ldots, m. \]

- **Association**

- **Direction of Influence**
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 \( H^r_0 : D(\psi | Z_r) = D(\psi) \) for all covariates \( r = 1, \ldots, R \), so that global hypothesis test is \( \bigcap_{r=1}^{R} H^r_0 \)
  • 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)

2. Repeat process with variation at every iteration (forest) using *random split selection* (Breiman, 2001)

3. Exclude subgroups based on:
   - model fit (randomness of errors)
   - association between $Y$ and $\xi$
   - 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.)

3. Compile results across all iterations
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)

Prediction:

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

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 (→ randomness of errors)
   - association between $Y$ and $\xi$
   - 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.)

3. Compile results across all iterations
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 (→ randomness of errors)
   - association between $Y$ and $\xi$
   - 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 (→ randomness of errors)
   - association between $Y$ and $\xi$
   - 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 (→ randomness of errors)
   - association between \( Y \) and \( \xi \)
   - 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.)
3. Compile results across all iterations
Survey Scale Forest in Practice

Survey Data → scforest.train → Trained Model → scforest.predict → Valid Latent Variable Scores

- Residuals correlated with ξ
- Instable parameters
- Confounded relations

Y not sufficiently associated with ξ
Survey Scale Forest in Practice

no pseudo-isolation

- Residuals correlated with $\xi$
- Instable parameters
- Confounded relations

Survey Data $\xrightarrow{\text{scforest.train}}$ Trained Model $\xrightarrow{\text{scforest.predict}}$ Valid Latent Variable Scores

$\gamma$ not sufficiently associated with $\xi$

no association
Conclusion

Survey Scale Forest detects two conditions for non-causality,

• lack of **pseudo-isolation** (→ confoundedness of relations in model),

• lack of **association** between $Y$ and $\xi$,

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.


Any questions?

- Contact: classe@ihf.bayern.de

- Want to try the method?:
  R-package `scforest` on GitHub
  [github.com/chkern/scforest](https://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.

→ 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 \( Z = \{Z_1, \ldots, Z_R\} \) into set of terminal nodes (leaves)
- Reduce outcome heterogeneity
- Usually built on training data and used to predict outcome in test data
<table>
<thead>
<tr>
<th>id</th>
<th>tree1R</th>
<th>tree1S</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>874</td>
<td>1905</td>
<td>-2.2302628</td>
<td>-2.0512677</td>
</tr>
<tr>
<td>1970</td>
<td>1682</td>
<td>-1.8856656</td>
<td>-1.68850199</td>
</tr>
<tr>
<td>3001</td>
<td>1766</td>
<td>-1.7982356</td>
<td>-1.6667754</td>
</tr>
<tr>
<td>1210</td>
<td>1201</td>
<td>-1.6900783</td>
<td>-1.4931119</td>
</tr>
<tr>
<td>2052</td>
<td>1810</td>
<td>-1.6221129</td>
<td>-1.4812064</td>
</tr>
<tr>
<td>602</td>
<td>1405</td>
<td>-1.6977506</td>
<td>-1.4420762</td>
</tr>
<tr>
<td>3007</td>
<td>1613</td>
<td>-1.9508186</td>
<td>-1.03015529</td>
</tr>
<tr>
<td>1640</td>
<td>1694</td>
<td>-1.4015596</td>
<td>-1.3995411</td>
</tr>
<tr>
<td>3406</td>
<td>1764</td>
<td>-1.4659346</td>
<td>-1.3567267</td>
</tr>
<tr>
<td>2846</td>
<td>1244</td>
<td>-1.4722657</td>
<td>-1.3557889</td>
</tr>
<tr>
<td>803</td>
<td>1703</td>
<td>-1.4241472</td>
<td>-1.0055205</td>
</tr>
<tr>
<td>2722</td>
<td>1928</td>
<td>-1.0118237</td>
<td>-1.2983938</td>
</tr>
<tr>
<td>2948</td>
<td>1562</td>
<td>-1.3673506</td>
<td>-1.2803734</td>
</tr>
<tr>
<td>552</td>
<td>1566</td>
<td>-1.3720552</td>
<td>-1.2565205</td>
</tr>
<tr>
<td>567</td>
<td>1976</td>
<td>-1.2389696</td>
<td>-1.2246041</td>
</tr>
<tr>
<td>2057</td>
<td>1535</td>
<td>-1.3092645</td>
<td>-1.1915643</td>
</tr>
<tr>
<td>941</td>
<td>1922</td>
<td>-1.2886267</td>
<td>-1.1832947</td>
</tr>
<tr>
<td>736</td>
<td>1720</td>
<td>-1.1806466</td>
<td>-1.1365030</td>
</tr>
</tbody>
</table>