Discussion: Causal Inference with Latent Variables

Kosuke Imai

Harvard University

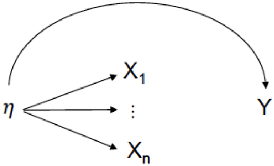
Joint Statistical Meetings

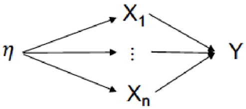
August 9, 2021

VanderWeele and Vansteelandt

- What are "latent variables"?
  - unobserved variables
  - often, they represent psycho-social constructs: depression, intelligence, well-being, socio economic status, social integration
- Structural vs. Statistical latent factor models

$$X_i = \lambda_i \eta + \epsilon_i \quad \text{for each observed variable } X_i$$





- Many researchers regress Y on $\hat{\eta}$ and give causal interpretation
- But, Factor model does not distinguish these two DAGs
- The authors derive a statistical test of the structural interpretation

What is the structural latent factor model?

- DAG implications:
  1. no arrow directly out of $(X_1, X_2, \ldots, X_n)$
  2. no arrow into $(X_1, X_2, \ldots, X_n)$ except η
- Applicable when X_i is a survey measurement of η
  - η: satisfaction with life
  - X_i: "If could live my life over, I would change almost nothing"
  - The answer to this question does not affect other variables on DAG
  - "Satisfaction with life" is the only thing that affects this question
- May not be applicable when X_i is some behavioral measurement
  - η: political ideology of legislators
  - X_i: rollcall votes
  - incoming arrows: constituency interests
  - outgoing arrows: election outcomes
- The 1st case is about measurement validity
- The 2nd case is also structural since η is causally efficacious

Key Result: Theorem 1

- Under the structural latent factor model

$$Z \perp\!\!\!\perp (X_1, \ldots, X_n) \mid \eta \quad \text{for any variable } Z \text{ on DAG}$$

- Testable implication under the linear factor model:

$$\lambda_i \mathbb{E}(X_j \mid Z = z) = \lambda_j \mathbb{E}(X_i \mid Z = z)$$

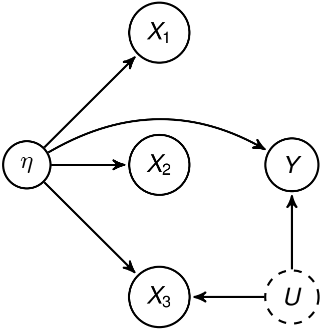
  This follows because for any *i*

$$\begin{aligned} \mathbb{E}(X_i/\lambda_i \mid Z = z) &= \mathbb{E}\{\mathbb{E}(X_i/\lambda_i \mid \eta, Z = z) \mid Z = z\} \\ &= \mathbb{E}\{\mathbb{E}(X_i/\lambda_i \mid \eta) \mid Z = z\} \\ &= \mathbb{E}(\eta \mid Z = z) \end{aligned}$$

- The authors develop the likelihood ratio test
- Extension to the case when η is multi-dimensional? Identifiability of the number of dimensions in factor model?

Questions

1. How does the identifiability of factor model affect these results?
2. What are the identification conditions for the causal effects of η on Y?
3. The role of factor model in the causal effects of X on Y $\rightsquigarrow$ stochastic intervention (Papadogeorgou et al. 2020)
4. Measurement error under the structural factor model?



Wang and Blei

- A follow-up of their influential 2019 JASA paper (2019 JSM)
- Setup:
  - multiple causes: $\mathbf{A}_i = (A_{i1}, A_{i2}, \ldots, A_{im})$
  - unobserved multi-cause confounders: $\mathbf{A}_i \perp\!\!\!\perp Y_i(\mathbf{a}) \mid \mathbf{U}_i$
- Deconfounder methodology:
  1. Factor model

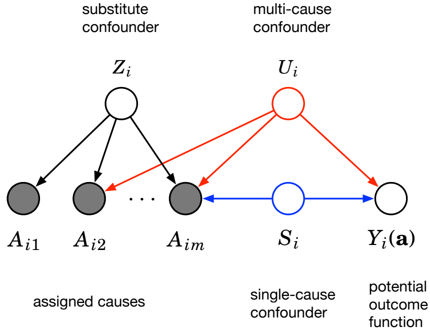
$$p(A_{i1}, A_{i2}, \ldots, A_{im}) \;=\; \int p(\mathbf{Z}_i) \prod_{j=1}^{m} p(A_{ij} \mid \mathbf{Z}_i)\, d\mathbf{Z}_i$$

  2. Substitute confounder $\mathbf{Z}_i$

$$\mathbb{E}\{Y_i(\mathbf{a}) - Y_i(\mathbf{a}')\} \;=\; \mathbb{E}\{\mathbb{E}(Y_i \mid \mathbf{A}_i = \mathbf{a}, \mathbf{Z}_i) - \mathbb{E}(Y_i \mid \mathbf{A}_i = \mathbf{a}', \mathbf{Z}_i)\}$$

- Advantages
  1. checkable assumption about unobserved confounders: $A_{ij} \perp\!\!\!\perp A_{ij'} \mid \mathbf{Z}_i$
  2. easy to implement

Assumptions



1. Unconfoundedness: $\mathbf{A}_i \perp\!\!\!\perp Y_i(\mathbf{a}) \mid \mathbf{U}_i$
2. $\mathbf{U}$ is a multi-cause separator:
$$P(A_1, \ldots, A_m \mid \mathbf{U}) = \prod_{j=1}^{m} P(A_j \mid \mathbf{U})$$
3. $\mathbf{U}$ does not contain a single-cause separator

- "Pinpointness" condition: All multi-cause separators $\mathbf{Z}$ ("substitute confounder") is a deterministic function of the multi causes $\mathbf{A}$
$$P(\mathbf{Z} \mid A_1, \ldots, A_m) = \delta_{f(A_1, \ldots, A_m)}$$
- Related to the propensity function of Imai & van Dyk (2004)

Mechanics of the Substitute Confounder

- Substitute confounder has the property: $\mathbf{A}_i \perp\!\!\!\perp \mathbf{U}_i \mid \mathbf{Z}_i$

$$
\begin{aligned}
& \mathbb{E}\{Y_i(\mathbf{a}, \mathbf{U}_i)\} \\
= & \int Y_i(\mathbf{a}, \mathbf{U}_i = \mathbf{u}) p(\mathbf{U}_i = \mathbf{u}) d\mathbf{u} \\
= & \int \int Y_i(\mathbf{a}, \mathbf{U}_i)\ p(\mathbf{U}_i = \mathbf{u} \mid \mathbf{Z}_i = \mathbf{z}) p(\mathbf{Z}_i = \mathbf{z}) d\mathbf{u} d\mathbf{z} \\
= & \int \int Y_i(\mathbf{a}, \mathbf{U}_i)\ p(\mathbf{U}_i = \mathbf{u} \mid \mathbf{A}_i = \mathbf{a}, \mathbf{Z}_i = \mathbf{z}) p(\mathbf{Z}_i = \mathbf{z}) d\mathbf{u} d\mathbf{z} \\
= & \int \mathbb{E}(Y_i \mid \mathbf{A}_i = \mathbf{a}, \mathbf{Z}_i = \mathbf{z}) p(\mathbf{Z}_i = \mathbf{z}) d\mathbf{z}
\end{aligned}
$$

- Implied estimator:

$$
\mathbb{E}\{\widehat{Y_i(\mathbf{a}, \mathbf{U}_i)}\} \;=\; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i \mid \widehat{\mathbf{A}_i = \mathbf{a}}, \mathbf{Z}_i = \widehat{\mathbf{Z}}_i) \text{ where } \widehat{\mathbf{Z}}_i \;=\; \hat{f}(\mathbf{A}_i)
$$

Questions

1. Why do you need the pinpointness assumption?
   - The support of $p(\mathbf{Z}_i)$ must be the same as that of $p(\mathbf{Z}_i \mid \mathbf{A}_i = \mathbf{a})$
   - Otherwise, we can't compute $\mathbb{E}(Y_i \mid \mathbf{A}_i = \mathbf{a}, \mathbf{Z}_i = \mathbf{z})$ for some $\mathbf{z}$
   - Substitute confounder $\mathbf{Z}_i$ is a deterministic function of $\mathbf{A}_i$
   - Factor model gives an estimate of this function $\rightsquigarrow$ identifiability of factor model?

2. How should one interpret the pinpointness assumption mean in scientific applications?

3. How sensitive are the results to the violation of this assumption? $\rightsquigarrow$ analogous to covariate balancing propensity score?

4. Does the assumption of no single cause confounder depend on the definition of "cause" and "confounder"?

Egleston *et al.*

- Using electronic health records to predict type 2 diabetes
- Two types of data
  1. structured fields: diagnosis code
  2. unstructured fields: notes by clinicians
- Using word2vec, the authors show how the textual data can be used to predict type 2 diabetes
- The authors also show how to quantify statistical uncertainty
- Questions:
  - Is the ultimate goal using clinician's notes to diagnose disease?
  - Probabilistic models of texts (e.g., LDA)?
- Extensions to causal inference:
  1. texts as moderator
  2. texts as treatment
- Use of latent variable modeling

Concluding Remarks

- Key roles of latent variables in many disciplines
- Causal inference with latent variables
  - VanderWeele: latent variable as treatment
  - Blei: latent variable as deconfounder
  - Egleston: latent variable as predictor

- Main difficulty: model dependent due to unobservability

- What are the roles of latent variable models in causal inference?
  - data generating process (i.e., structural)
  - summarization tool