A Bayesian latent class approach to causal inference with longitudinal data

Kuan Liu[1,2], Olli Saarela[1], George Tomlinson[1,3], Eleanor Pullenayegum[1,2]

1. Dalla Lana School of Public Health, University of Toronto 2. Child Health Evaluative Sciences, The Hospital for Sick Children
3. Department of Medicine, University Health Network

Background

- Bayesian causal methods that follow a parametric specification of the joint likelihood of treatment, outcome and covariates, are analytically intractable when face with high-dimensional confounders.
- One possible approach for dimensionality reduction is to model the set of confounders as class indicators in a latent class analysis.
- This approach mimics the treatment assignment process often seen in observational studies with administrative data that contain a large number of variables which are indicative of the patient's disease and health status.
- **We aim to provide a Bayesian latent class approach to estimate causal effects.**

Objectives

- In this paper, we consider **a causal effect that is confounded by an unobserved, visit specific, latent class** in a longitudinal setting.
- We formulate the joint likelihood of the treatment, outcome and latent class models conditionally on the class indicators, which permits a **full Bayesian causal inference**.

Causal Framework

1. NOTATIONS

- n subjects indexed by i, $i = 1, \dots, n$ and k number of visits for each subject indexed by j, $j = 1, \dots, k$.
- Y_i, X_{ij}, U_{ij} and Z_{ij} are random variables representing an end-of-study outcome, time-dependent class indicators (a vector of p elements), time-dependent latent class and time-dependent treatment for i at visit j.
- There are a treatment categories available at each visit and c number of class memberships at each visit.
- History up to visit j are denoted as $\tilde{U}_{ij} = \{U_{i1}, \dots, U_{ij}\}$, $\tilde{X}_{ij} = \{X_{i1}, \dots, X_{ij}\}$ and $\tilde{Z}_{ij} = \{Z_{i1}, \dots, Z_{ij}\}$.
- Let θ, α, β and γ characterize the outcome model, the latent class model, the class indicator model and the treatment model respectively.

2. DAG

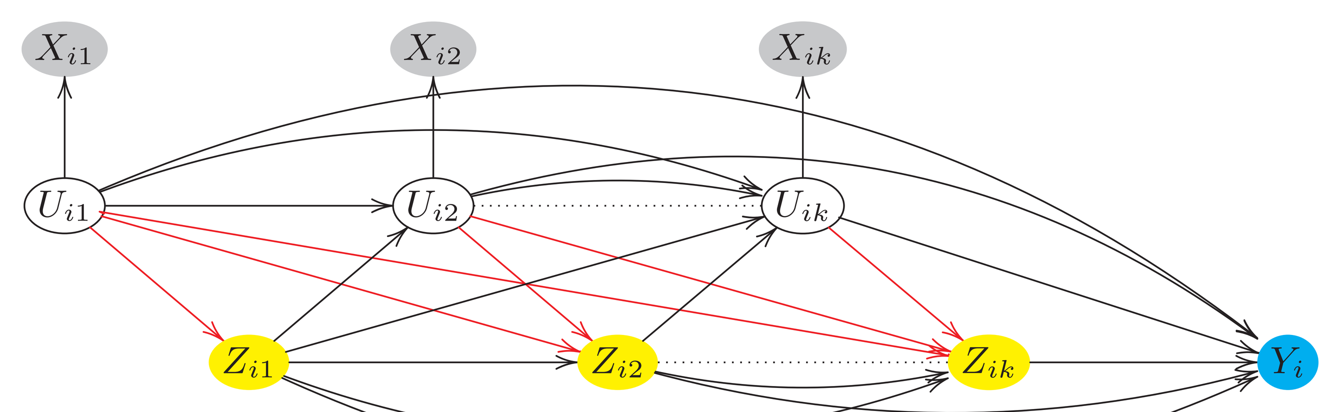

Figure 1. Longitudinal causal DAG between latent class, indicators, treatment and outcome

3. ASSUMPTIONS

- Stable unit treatment value assumption
- Consistency, $Y_i^{\tilde{a}} \mid (z_{i1} = a_1, \dots, z_{ik} = a_k) = Y_i \mid (z_{i1} = a_1 \dots, z_{ik} = a_k)$
- Positivity, at each visit, every possible treatment sequence which is compatible with treatment history up till that visit have positive probability of occurring.
- Sequential latent unconfoundedness, $Z_{ij} \perp (U_{ij}^{\tilde{a}_j}, X_{ij}^{\tilde{a}_j}, Y_i^{\tilde{a}}) \mid (U_{ij-1}^{\tilde{a}_{j-1}}, Z_{ij-1})$, for $j = 1, \dots, k$.
- Conditional independence between class indicators, $P(X_{ij} \mid U_{ij}) = \prod_{l=1}^{p} P(X_{lij} \mid U_{ij})$.

Simulation Study

1. SIMULATED DATASET

- We simulated 1000 iterations of a simple two-visit ($k = 2$) longitudinal dataset with $n = 250$ and $n = 500$.
- $C_i^a \sim N(10, 3)$ and C_i^s from a Bernoulli distribution with $P(C_i^2 = 1) = 0.6$
- Z_{ij}, a binary treatment assignment. $logit(P(Z_{ij} = 1)) = -1 + u_{ij} - z_{ij-1}$
- U_{ij}, three class memberships with simulated proportion at $42\% : 29\% : 29\%$ from a Multinomial distribution with

$$log\frac{P(U_{ij}=2)}{P(U_{ij}=1)} = 0.5 - 0.1c_i^a + 0.2c_i^s + I(u_{ij-1}=2) + 0.5I(u_{ij-1}=3) - z_{ij-1}$$

$$log\frac{P(U_{ij}=3)}{P(U_{ij}=1)} = 0.5 - 0.1c_i^a + 0.2c_i^s + 0.5I(u_{ij-1}=2) + I(u_{ij-1}=3) - z_{ij-1}$$

- X_{ij}, simulated conditionally independent given U_{ij} from Bernoulli distribution.
  - High quality defined as $P(X_{ij} = 1 \mid U_{ij} = c) = 0.88$
  - Medium quality defined as $P(X_{ij} = 1 \mid U_{ij} = c) = 0.73$
  - Low quality defined as $P(X_{ij} = 1 \mid U_{ij} = c) = 0.62$
- $Y_i \sim N(\mu_{\tilde{a}}, 1)$, where $\mu_{\tilde{a}} = 0.1 + 0.5I(z_{i1} = a_1) + I(z_{i2} = a_2) - 0.2I(u_{i1} = 2) - 0.5I(u_{i1} = 3) - 0.5I(u_{i2} = 2) - I(u_{i2} = 3)$.

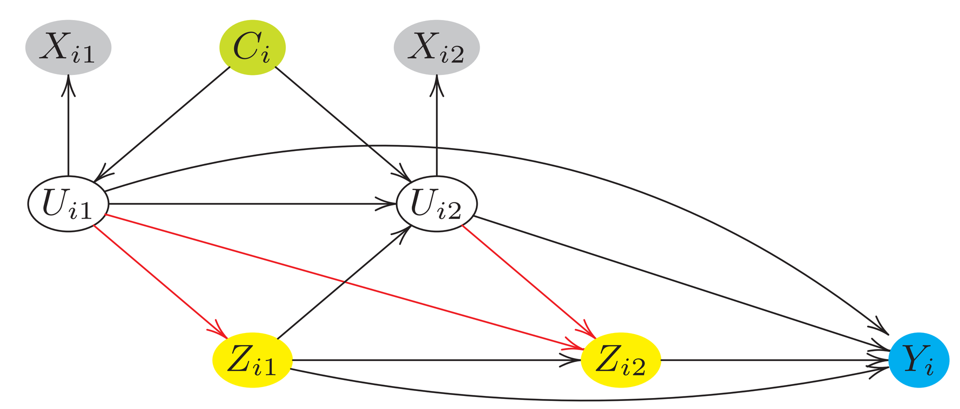

Figure 2. Causal diagram of the simulation dataset.

Table 1. Simulation results from 1000 iterations. Parameter of interest $(\mu_{11} - \mu_{00})$, treatment effect evaluated between always treated and never treated. RB: Relative Bias; ESE: Empirical Standard Error; ASE: Average Standard Error; CP: Coverage Probability; U1 and U2 represent the average proportion of correct imputation under Bayesian estimation.

Setting	Estimator	Mean	RB	ESE	ASE	CP	U1	U2
$n = 250$	Naive	1.18	-29.38	0.24	0.24	46.1		
10 indicators	Adjust	1.47	-12.21	0.22	0.23	86.4		
high quality	MSMs Sand.	1.61	-3.47	0.27	0.28	95.8		
	Bayes	1.67	0.12	0.22	0.23	96.4	0.70	0.70
$n = 250$	Naive	1.18	-29.38	0.24	0.24	46.1		
10 indicators	Adjust	1.38	-17.07	0.24	0.24	75.0		
med quality	MSMs Sand.	1.48	-11.22	0.28	0.28	87.5		
	Bayes	1.68	0.80	0.26	0.25	95.9	0.60	0.60
$n = 250$	Naive	1.18	-29.38	0.24	0.24	46.1		
10 indicators	Adjust	1.28	-23.48	0.24	0.24	61.2		
low quality	MSMs Sand.	1.32	-20.99	0.26	0.26	71.5		
	Bayes	1.19	-28.76	0.25	0.24	47.6	0.39	0.39
$n = 500$	Naive	1.18	-29.59	0.16	0.17	13.6		
10 indicators	Adjust	1.47	-12.14	0.15	0.16	78.4		
high quality	MSMs Sand.	1.62	-3.09	0.18	0.20	95.2		
	Bayes	1.68	0.45	0.15	0.16	96.6	0.78	0.78
$n = 500$	Naive	1.18	-29.59	0.16	0.17	13.6		
10 indicators	Adjust	1.38	-17.68	0.15	0.16	54.4		
med quality	MSMs Sand.	1.46	-12.31	0.17	0.19	82.2		
	Bayes	1.68	0.70	0.17	0.17	96.4	0.68	0.69
$n = 500$	Naive	1.18	-29.59	0.16	0.17	13.6		
10 indicators	Adjust	1.27	-24.12	0.16	0.17	29.8		
low quality	MSMs Sand.	1.30	-22.33	0.17	0.18	41.8		
	Bayes	1.31	-21.73	0.28	0.18	35.2	0.40	0.41

Bayesian causal inference with latent class

1. Causal parameter of interest - Average potential outcome (APO)

$$E[Y_i^{\tilde{a}}] = \sum_{u_{ik}=1}^{C} \dots \sum_{u_{i1}=1}^{C} E(y_i \mid \tilde{z}_i = \tilde{a}, \tilde{u}_i = \tilde{c}_k, \theta) \Big[\prod_{j=1}^{k} P(u_{ij} = c_j \mid \tilde{z}_{ij-1} = \tilde{a}_{j-1}, \tilde{u}_{ij-1} = \tilde{c}_{j-1}, \alpha_j)\Big] \quad (1)$$

2. Joint likelihood

$$P(y_i, \tilde{z}_i, \tilde{u}_i, \tilde{x}_i \mid \theta, \alpha, \beta, \gamma) = P(y_i \mid \tilde{z}_i, \tilde{u}_i = \tilde{c}_k, \theta)\Big[\prod_{j=1}^{k} P(z_{ij} \mid \tilde{z}_{ij-1}, \tilde{u}_{ij} = \tilde{c}_j, \gamma_j) \times P(u_{ij} = c_j \mid \tilde{u}_{ij-1}, \tilde{z}_{ij-1}, \alpha_j) P(x_{ij} \mid u_{ij} = c_j, \beta_j)\Big] \quad (2)$$

3. Posterior Distribution

$$P(\theta, \alpha, \beta, \gamma \mid y_i, \tilde{z}_i, \tilde{u}_i, \tilde{x}_i) = \frac{P(y_i, \tilde{z}_i, \tilde{u}_i, \tilde{x}_i \mid \theta, \alpha, \beta, \gamma) P_0(\theta, \alpha, \beta, \gamma)}{\int_{\{\theta,\alpha,\beta,\gamma\}} P(y_i, \tilde{z}_i, \tilde{u}_i, \tilde{x}_i \mid \theta, \alpha, \beta, \gamma) P_0(\theta, \alpha, \beta, \gamma) d\theta d\alpha d\beta d\gamma} \quad (3)$$

We assume a prior independence $P_0(\theta, \alpha, \beta, \gamma) = P_0(\theta)P_0(\alpha)P_0(\beta)P_0(\gamma)$.

4. Posterior Predictive Inference on APO We predict the potential outcome of a given treatment sequence for a new observation drawn from data distribution $\{Y, Z, U, X\}$,

$$P(\tilde{y}_i^{\tilde{a}} \mid y_i, \tilde{z}_i, \tilde{u}_i, \tilde{x}_i) = \int_{\{\theta,\alpha,\beta,\gamma\}} P(\tilde{y}_i^{\tilde{a}} \mid y_i, \tilde{z}_i, \tilde{u}_i, \tilde{x}_i) P(\theta, \alpha, \beta, \gamma \mid y_i, \tilde{z}_i, \tilde{u}_i, \tilde{x}_i) d\theta d\alpha d\beta d\gamma. \quad (4)$$

Conclusion & Future works

- **Based on the simulation study, Bayesian latent class approach is preferred when we have medium to high quality class indicators**.
- Full Bayesian specification permits imputation on class membership; Even when memberships are not well predicted, we can still obtain relative unbiased APO estimate.
- Our Bayesian approach can be easily implemented in common MCMC software.
- **Future works to investigate the trade off between estimation accuracy and minimum number of high quality indicators.**

Acknowledgment

- This research is funded by the Canadian Institutes of Health Research.
- Dr. Brian Feldman, The Hospital for Sick Children

References

Robins J., A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect, *Math Modelling* 7(9-12), 1986

Keil A.P., Daza E.J., Engel S.M., Buckley J.P., and Edwards J.K., A Bayesian approach to the g-formula. *Statistical methods in medical research* 27(10), 2018.

Lanza S.T., Coffman D.L., and Xu S., Causal inference in latent class analysis. *Structural equation modeling: a multidisciplinary journal* 20(3), 2013.

Liu K., Saarela O., Feldman B.M., and Pullenayegum E., Estimation of causal effects with repeatedly measured outcomes in a Bayesian framework. *Statistical Methods in Medical Research*, 2020.