CSDS 452 Causality and Machine Learning

Lecture 12: Causal Effect Learning from Time-series data

Instructor: Jing Ma

Fall 2024, CDS@CWRU

Outline

- Time-varying treatments & confounders
- Assumptions
- G-formula
- Marginal structural model (MSM)
- Time-series deconfounder

Longitudinal observational studies

- All previous discussions focus on treatment at a single time (cross-section or panel settings)
- Common in real world situations, e.g. medical research, data (treatment, covariates and/or outcome) are repeatedly collected on subjects over a period – longitudinal studies
- Particularly interested in estimating the effect of a time-varying treatment on an outcome of interest measured at a later time
- Confounders can be time-varying, affected by past treatment and affecting future covariates and or outcomes
- Standard regression adjustment fails to give consistent estimators in the presence of time-varying confounders if those confounders are themselves affected by treatment

Sequentially ignorable assignment

What type of studies we consider

- Treatments with multiple time points, where those treatments assignment is ignorable conditionally on the observed history.
- If we can justify the above assumption, this is a possible template for randomized experiments or observational studies
- Example 1: patients visiting doctors at different times.
- Example 2: workers exposed to hazards at the workplace (related to health worker survivor effect)

A common goal is to estimate the accumulative (over the study period) effect of the treatment on an outcome.

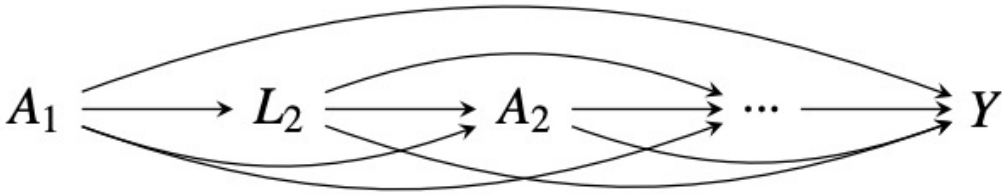
Time-varying treatments: An example

(Daniel et al. 2013, SIM)

- In many longitudinal medical studies, patients' treatment changes over time and is measured several times during the study, along with other time changing covariates.
- For example, type II diabetes patients recruited into a study comparing two antiglycaemic drugs may be followed up on several occasions, on each of which their HbA1c (a long-term measure of blood glucose level), blood pressure, cholesterol level, BMI, anaemia status, and others variables are measured
- Suppose wish to compare the effect of the two treatments on HbA1c 18 months after recruitment and on the risk (or hazard) of experiencing a cardiac event in the 18 months following recruitment.

Time-varying treatments: The challenge

(Daniel et al. 2013, SIM)



- Study allows for the dose and type of treatment to be changed according to the current (and past) values of HbA1c and other covariates.
- A high HbA1c likely lead to increasing the dose of the current drug; but high HbA1c is also thought to lead to an increased risk of a cardiac event, making HbA1c at a particular time a confounder of the relationship between subsequent treatment and the outcome
- Because HbA1c varies over time (in a way that cannot be foreseen at baseline), it is called a *time-varying confounder*

Time-varying treatments: The challenge

(Daniel et al. 2013, SIM)

- To estimate the causal effect of treatment on risk of cardiac event, it seems necessary to control for HbA1c in the analysis
- Not only does HbA1c affect treatment but also **the reverse is true**!!!
  - An effective antiglycaemic drug lowers HbA1c, and thus current value of the treatment variable has a causal effect on future values of HbA1c
- This means controlling for HbA1c is problematic, because future measurements of HbA1c lie on the causal pathway between past treatment and the outcome
  - conditioning on HbA1c blocks some of the effect of the treatment and, in addition, conditioning on a consequence of treatment risks inducing collider-stratification bias

Notation with a toy example

- We will switch notation from previous lectures to be compatible with the literature in longitudinal treatment
- Toy example: patients with cancer, visiting doctor at two time points t_1, t_2.
- $a_t (t = 1, 2)$: possible treatment at time t
- $A_{i,t}$: the observed treatment at time t ($Z_{i,t}$ in previous notation system)
- L_i^{obs}: observed cancer progression at time 2
- $Y_i(a_1, a_2)$: potential outcome at time 3
- $Y_i = Y_i(A_{i,1}, A_{i,2})$: observed outcomes

Toy example with $T = 2$

month	action	potential outcome	observed value
1	give treatment a_1 (1=high)		$A_{i,1}$
2	(i) measure cancer progression		L_i^{obs}
	(ii) give treatment a_2 (1=high)		$A_{i,2}$
3	measure cancer progression	$Y_i(a_1, a_2)$	Y_i

- For each individual, there are a total of 4 potential outcomes $\{Y(1,1), Y(1,0), Y(0,1), Y(0,0)\}$, but only one will be observed

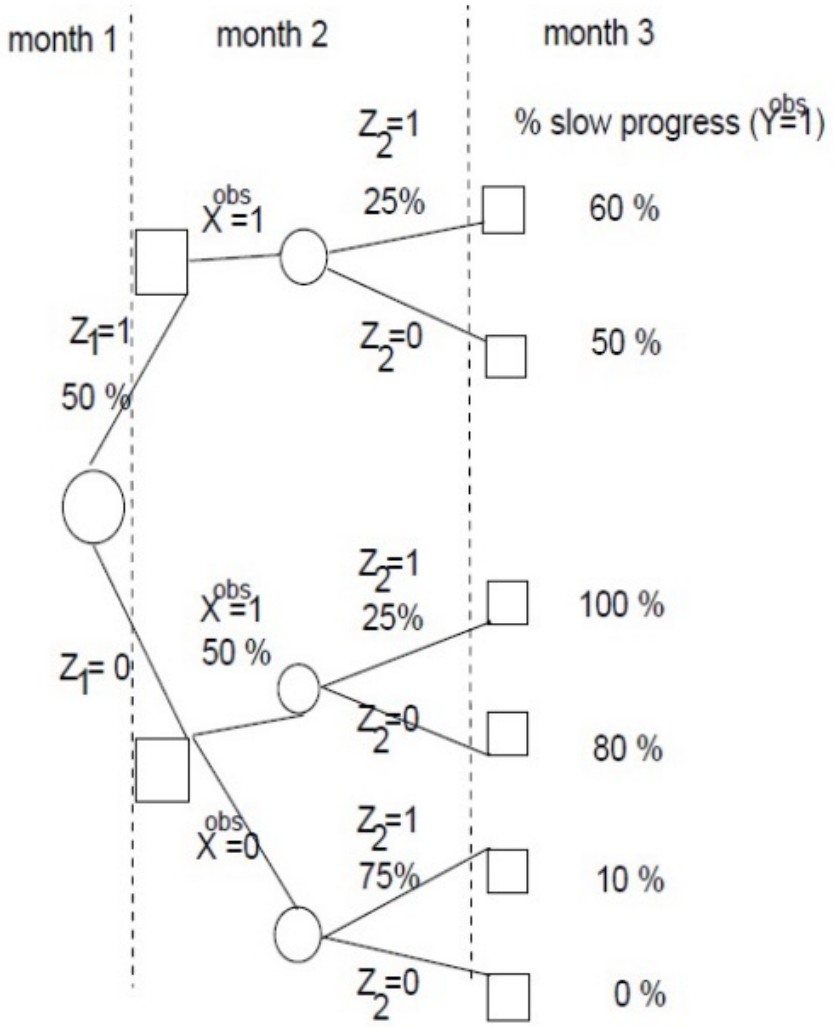


Figure: Example with treatment at two time points and sequentially ignorable assignment (Here Z is the assignment, and X^{obs} is the intermediate variable)

Causal Estimand

- Typical target estimand - marginal causal effects due to treatment sequence:

$$\tau_{a_1 a_2, a_1' a_2'} = \mathbb{E}[Y_i(a_1, a_2) - Y_i(a_1', a_2')],$$

for all $(a_1, a_2) \neq (a_1', a_2') \in \{0, 1\}^2$.

- For example, compare cancer progression Y between always taking high dose $\Pr(Y(1, 1) = 1)$ and always taking low dose $\Pr(Y(0, 0) = 1)$.

- Note that we only control (a_1, a_2), that is why the potential outcomes $Y_i()$ are only a function of a_1, a_2 and not also of L.

Problems with standard adjustment

Two "standard" approaches to estimate

$\Pr(Y(1,1)=1) - \Pr(Y(0,0)=1)$:

- Approach 1. "Do not condition on progression L^{obs} because it is an intermediate outcome":

$$\begin{aligned} & \Pr(Y=1|A_1=1, A_2=1) - \Pr(Y=1|A_1=0, A_2=0) \\ = \quad & 60\% - 60\% = 0 \end{aligned}$$

- Approach 2. "Condition on intermediate progression L^{obs} because it was used in deciding treatment A_2":

$$\begin{aligned} & \Pr(Y=1|A_1=1, A_2=1, L^{obs}=1) - \Pr(Y=1|A_1=0, A_2=0, L^{obs}=1) \\ = \quad & 60\% - 80\% = -20\% \end{aligned}$$

Problems with standard adjustment

Two "standard" approaches to estimate

$\Pr(Y(1,1)=1) - \Pr(Y(0,0)=1)$:

- Approach 1. "Do not condition on progression L^{obs} because it is an intermediate outcome":

$$\begin{aligned} & \Pr(Y=1|A_1=1, A_2=1) - \Pr(Y=1|A_1=0, A_2=0) \\ = \quad & 60\% - 60\% = 0 \end{aligned}$$

- Approach 2. "Condition on intermediate progression L^{obs} because it was used in deciding treatment Z_2":

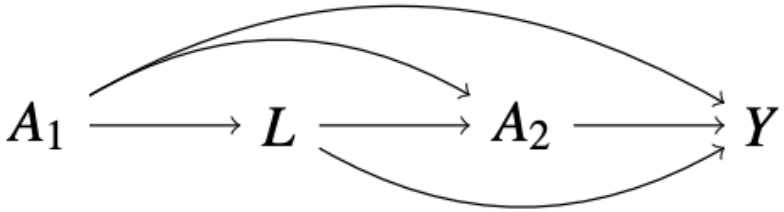
$$\begin{aligned} & \Pr(Y=1|A_1=1, A_2=1, L^{obs}=1) - \Pr(Y=1|A_1=0, A_2=0, L^{obs}=1) \\ = \quad & 60\% - 80\% = -20\% \end{aligned}$$

Both approaches are incorrect for the goal - adjusting for L^{obs} alone is not enough

Outline

- Time-varying treatments & confounders
- Assumptions
- G-formula
- Marginal structural model (MSM)
- Time-series deconfounder

Assumption 1: Positivity/Overlap



$$\tau_{a_1 a_2, a_1' a_2'} = \mathbb{E}[Y_i(a_1, a_2) - Y_i(a_1', a_2')]$$

- At every time point, units have positive probability to receive all levels of the treatment

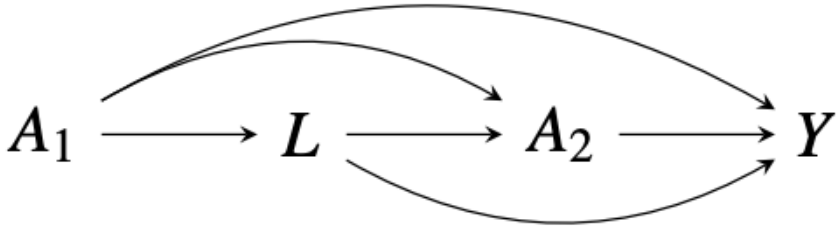
$$P(A_1 = a_1) > 0$$
$$P(A_2 = a_2 | A_1 = a_1, L = l) > 0$$

for all a_1, a_2, l.

Assumption 1: Positivity/Overlap

- Positivity states that conditional on covariate history, the probability of receiving each treatment sequence is bounded away from zero and one
- The above illustration is only for two time periods ($T = 2$), and therefore the covariate history includes only L right after treatment A_1
- Positivity is less likely to be satisfied
  - for large values of T
  - when A includes more than two values (multiple treatments)

Assumption 2: Sequential Ignorability (Robins, 1986)



Let L_t^{obs} (often shorthanded to L_t) denote the time-varying confounders, including both time-varying covariates and intermediate outcome at time t.

Let $\bar{a}_t = (a_1, a_2, ..., a_t)$, $\bar{A}_t = (A_1, \ldots, A_t)$

- Sequential ignorability: treatment at time t is randomized with probabilities depending on the observed past, *including covariates, intermediate outcomes*, that is, at any time t :

$$\{Y_i(\bar{a}_t), \forall \bar{a}_t\} \perp A_{i,t} \mid H_{i,t},$$

where $H_{i,t} = (A_1, ..., A_{t-1}; L_1, ...L_{t-1})$ is the observed history

Identifiability

- $E[Y(a_1, a_2)]$ is a function of potential outcomes
- Idenitifiability of $E[Y(a_1, a_2)]$ means that it can be written in terms of observed data

$$
\begin{aligned}
& E[Y(a_1, a_2)] \\
&= E[Y(a_1, a_2)|A_1 = a_1] && (1) \\
&= E[Y(A_1, a_2)|A_1 = a_1] && (2) \\
&= \sum_{l=0,1} E[Y(A_1, a_2)|A_1 = a_1, L^{obs} = l]P(L^{obs} = l|A_1 = a_1) && (3) \\
&= \sum_{l=0,1} E[Y(A_1, a_2)|A_1 = a_1, L^{obs} = l, A_2 = a_2]P(L^{obs} = l|A_1 = a_1) && (4) \\
&= \sum_{l=0,1} E[Y(A_1, A_2)|A_1 = a_1, L^{obs} = l, A_2 = a_2]P(L^{obs} = l|A_1 = a_1) && (5)
\end{aligned}
$$

- (1), (4) from sequential ignorability
- (2), (5) from consistency (SUTVA); (3) from law of total probability

Outline

- Time-varying treatments & confounders
- Assumptions
- G-formula
- Marginal structural model (MSM)
- Time-series deconfounder

g-computation

(Robins, 1986)

- Causal effects are identified under the assumption of sequential ignorability leading to the g-computation

$$
\begin{aligned}
\Pr(Y(0,0)=1) &= \Pr(Y(0,0)=1 \mid A_1=0) \\
&= \sum_{L^{obs}=0,1} \Pr(Y(0,0)=1 \mid A_1=0, L^{obs}) \Pr(L^{obs} \mid A_1=0) \\
&= \sum_{L^{obs}=0,1} \Pr(Y(0,0)=1 \mid A_1=0, L^{obs}, A_2=0) \Pr(L^{obs} \mid A_1=0) \\
&= \sum_{L^{obs}=0,1} \Pr(Y^{obs}=1 \mid A_1=0, L^{obs}, A_2=0) \Pr(L^{obs} \mid A_1=0) \\
&= 0\%(50\%) + 80\%(50\%) = 40\%
\end{aligned}
$$

g-computation

(Robins, 1986)

- Similarly, we can estimate $\Pr(Y_i(1,1) = 1) = 60\%$
- Therefore, the causal effect is given by

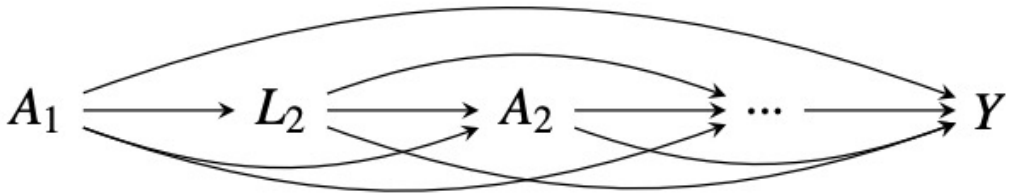
$$\Pr(Y_i(1,1) = 1) - \Pr(Y_i(0,0) = 1) = 20\%$$

- Use analogous arguments, we can also estimate $\Pr(Y_i(0,1) = 1) = 55\%$ and $\Pr(Y_i(1,0) = 1) = 50\%$.
- This procedure can be generated to longitudinal treatments with any T time points: g-computation

g-computation

(Robins, 1986)

- For T time points, let $\bar{a}_t = (a_1, a_2, ..., a_t)$, $\bar{A}_t = (A_1, \ldots, A_t)$



$$\Pr(Y(\bar{a}_T)) =$$

$$\sum_{L_2^{obs},\ldots,L_T^{obs}} \Pr(Y^{obs} \mid L_2^{obs}, ..., L_T^{obs}, \overline{A}_T = \overline{a}_T)$$

$$\times \Pr(L_2^{obs} | A_1 = a_1) \times \cdots \times$$

$$\times \Pr(L_T^{obs} | A_1 = a_1, L_2^{obs}, A_2 = a_2, \ldots, L_{T-1}^{obs}, A_{T-1} = a_{T-1}).$$

- This is the basic g-computation formula, or g-formula
- Can pose models for all distributions in the RHS and estimate $\Pr(\bar{a})$ – in essence, this is an outcome modeling approach

g-computation

- To operationalize the g-formula, a key component is to specify models for all components
  - outcome regression $\Pr(Y^{obs} \mid L_2^{obs}, \ldots, L_T^{obs}, \overline{A}_T = \overline{a}_T)$
  - models for time-varying confounders
    $\Pr(L_t^{obs} | A_1 = a_1, L_2^{obs}, A_2 = a_2, \ldots, L_{t-1}^{obs}, A_{t-1} = a_{t-1}) \; \forall t$
- Model for time-varying confounder can be complex
  - involves a large number of time points T
  - involves many covariates, some of which are continuous
  - may further factor
    $\Pr(L_t^{obs} | A_1 = a_1, L_2^{obs}, A_2 = a_2, \ldots, L_{t-1}^{obs}, A_{t-1} = a_{t-1}) \; \forall t$ with a series of conditional distributions
  - variable selection with longitudinal treatments still an open question

Dimension reduction and propensity score

- When all conditional distributions in g-computation are correctly specified, g-computation leads to the most efficient estimates with the smallest large sample variances
- However, dimension of variables increases exponentially with T, due to time-varying covariates
- With medium to large T, model building and model checking in g-computation can be very demanding
- Dimension reduction is crucial. Propensity score again plays a central role to achieve dimension reduction: weighting or outcome regression. Matching is less suitable.
- Positivity/overlap can be checked in terms of the propensity score, instead of directly on covariates (ignored by the g-computation estimator)

Dimension reduction and propensity score

- Define the propensity score at time t given the observed history as:

$$e_{it} = \Pr(A_{it} = 1 \mid \boldsymbol{H}_{it}), \quad i = 1, \ldots, T,$$

  where $\boldsymbol{H}_{it} = \{\bar{L}_{it}, \bar{A}_{i,t-1}\}$ is the observed history for unit i up to time t
- Under SI, easy to show SI holds for the longitudinal propensity scores: For a given t, and for all $\bar{a}_t$

$$\{Y_i(\bar{a}_t)\} \perp A_{i,t} | e_{i,1}, A_{i,1}, \ldots, e_{i,t-1}, A_{i,t-1} \tag{6}$$

- Equation (6) imply that instead of adjusting for the history of covariates, we can adjust for the history of propensity scores - substantially reduce the covariate dimension in modeling
- Two approaches: weighting (marginal structural models (MSM)) and regression on PS history

Outline

- Time-varying treatments & confounders
- Assumptions
- G-formula
- Marginal structural model (MSM)
- Time-series deconfounder

Marginal structural model

- Some simplification is necessary, for example, one can use the following GLM with inverse link function h

$$\mathbb{E}[Y(\bar{a})] = h(\bar{a}; \gamma)$$

- For example, we can posit $h(\bar{a}; \gamma) = \gamma_+ \gamma_2 \text{cum}(\bar{a})$, where $\text{cum}(\bar{a}) = \int_0^{T+1} a(t)dt = \sum_{t=0}^{T} a(t)$ is the cumulative treatment
- Can further adjust for baseline covariates (effect modifiers), for example with $V \in L_0$

$$\mathbb{E}[Y(\bar{a})] = h\left(\gamma_1 + \gamma_2 \sum_{t=0}^{T} a(t) + \gamma_3 V\right)$$

- We call such models marginal structural models (MSM)
  - models for some aspect of the conditional distribution of the counterfactuals given baseline covariates
  - always marginal with respect to post-baseline confounders

Generalizing IPW (Horvitz-Thompson) estimator

- Define the propensity score at time t given the observed history as:

$$e_t = P(A_t = 1 \mid \bar{A}_{t-1}, \bar{L}_t), \quad i = 1, \ldots, T,$$

- Obtain the stabilised inverse probability weights for each individual

$$\text{SW} = \frac{\prod_{t=0}^{T} P(A_t = A_t^{obs} \mid \bar{A}_{t-1})}{\prod_{t=0}^{T} P(A_t = A_t^{obs} \mid \bar{A}_{t-1}, \bar{L}_t)}$$

with $A_{-1} = \emptyset$

- Replace the denominator with 1 leads to the unstablized weights
- If the MSM further conditions on baseline covariates V, the stabilised inverse probability weights can be further modified as

$$\text{SW-V} = \frac{\prod_{t=0}^{T} P(A_t = A_t^{obs} \mid \bar{A}_{t-1}, V)}{\prod_{t=0}^{T} P(A_t = A_t^{obs} \mid \bar{A}_{t-1}, \bar{L}_t)}$$

IPW estimation of MSM

Robins, Hernan, Brumback, 2000, Epidemiology

- For $T = 2$
- Specify a model for the outcome, $f(y, \bar{a}, ; \gamma)$ with score function $S(y, a_0, a_1; \gamma) = \frac{\partial}{\partial \gamma} \log f(y, a_0, a_1; \gamma)$.
- For example, for a binary outcome Y with two time points, two possible models:
  - $\text{logit}\{\Pr(Y_i = 1 | A_0, A_1)\} = \gamma_1 + \gamma_2 A_0 + \gamma_3 A_1 + \gamma_4 A_0 A_1$
  - $\text{logit}\{\Pr(Y_i = 1 | A_0, A_1)\} = \gamma_1 + \gamma_2 (A_0 + A_1)$.
- Solving for the following estimating equation

$$\sum_{i=1}^{N} \text{SW}_i \times S(Y_i^{obs}, A_{i0}, A_{i1}; \gamma) = 0, \qquad (7)$$

  gives consistent estimates of the parameters γ
- Eq (7) is solved via maximizing the weighted likelihood

MSM: procedure

Step 1 Build an outcome model: $\Pr(Y(a_0, a_1)) = \Pr(Y|A_1, A_2)$ under "randomization"

Step 2 Build a propensity score model for each time: $\Pr(A_1|A_0, \bar{L}_1)$ and $\Pr(A_0|L_0)$; also build model for $\Pr(A_2|A_1)$ and $\Pr(A_1)$ (for stabilized weights) – this can be replaced with pooled models

Step 3 Estimate the propensity scores at each time, check overlap and remove units in the non-overlap region

Step 4 Calculate the stabilized weights for each unit at each time point

Step 5 Estimate the parameters of the outcome model by maximizing the weighted likelihood (weighted regression)

Case study: Hernan, Robin, Brumback (2000). outcome is survival, use the marginal structural Cox model

Doubly Robust MSM

(Bang and Robins, 2005 Biometrics)

- Simple weighting is not efficient, and lead to bias if the weights are incorrectly specified
- The ICE estimator (one form of g-formula) can be used to estimate MSM parameters as well, with a simple modification in the final sequential regression
  - recall ICE estimator imputes all potential outcomes (in an exhaustive fashion), and so in the final regression model, we can simply just fit a MSM to estimate γ
- Combine IPW and ICE to create a doubly-robust estimator for MSM
  - consistent for γ is either series of propensity score models or outcome models are correct
  - more efficient than IPW alone by exploiting a series of conditional outcome models

Strength of MSM

- Intuitive and relatively easy to explain – compared to g-formula, most closely related to standard methods
- easily extended to different types of outcome variable
- only require to specify models for the treatment assignment/propensity score and the MSM itself
  - conditional distributions of (1) outcome Y given the covariates and (2) time-varying covariates given past covariates and treatments are left unspecified (unlike g-formula)
  - less prone to model misspecification than the g-formula (high-dimensional covariates)

Misspecification of g-formula models can often lead to severe biases.

- not prone to g-null paradox

Limitations of MSM

- inverse weighting can be unstable and inefficient if there are extreme weights
  - stablized weights can help, but as we see in the cross-sectional setting, not a lot
  - prone to extreme weights, can use weight trimming or truncation (Cole and Hernan, AJE 2008), but similar issue as in one time point (sensitive to cutoff and ambiguous target population
  - possible to extend to target populations, e.g. other balancing weights such as overlap weights, but remain an open question
- possible interactions between treatment and time-varying covariates cannot be explored because the MSM is marginal with respect to the latter

Hidden confounders in time-series data

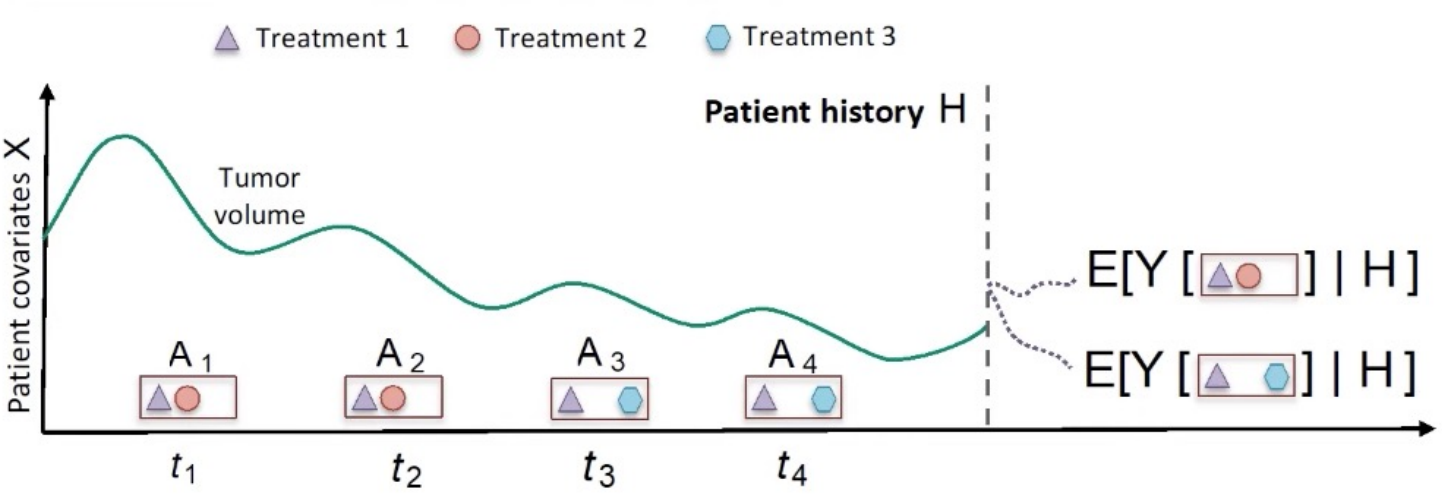
- Despite the wide applicability of these methods (g-computation, marginal structural models, and many follow-up methods) in forecasting treatment responses, they are all based on the assumption that there are no hidden confounders.
- What if hidden confounders exist?

Outline

- Time-varying treatments & confounders
- Assumptions
- G-formula
- Marginal structural model (MSM)
- Time-series deconfounder

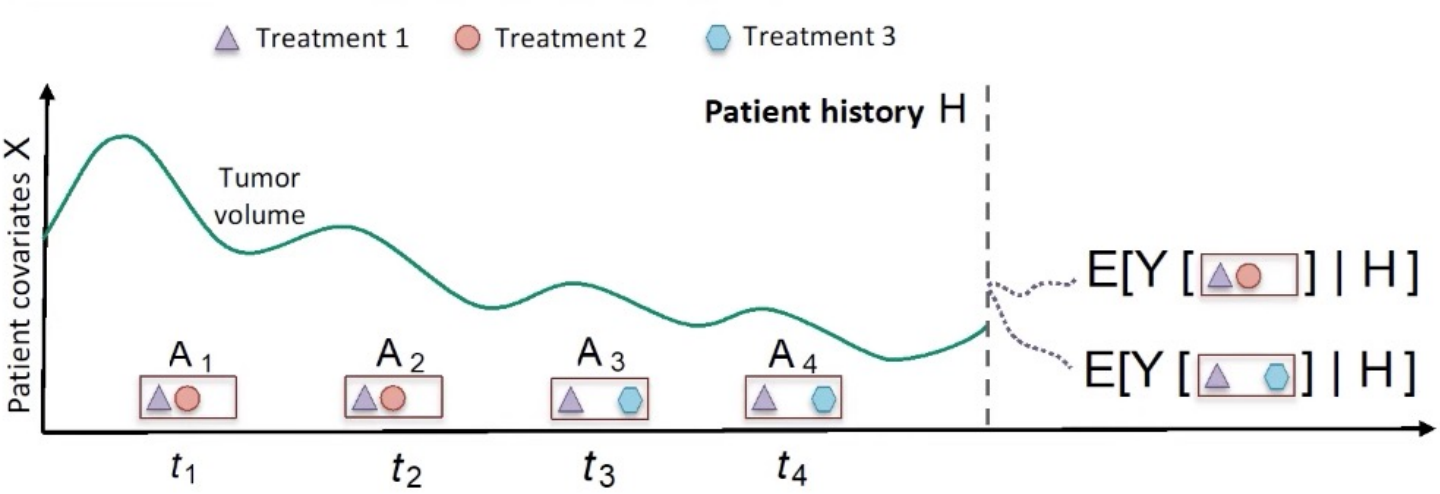
Treatment effect over time

- Aim: estimate ITE of time-dependent treatments



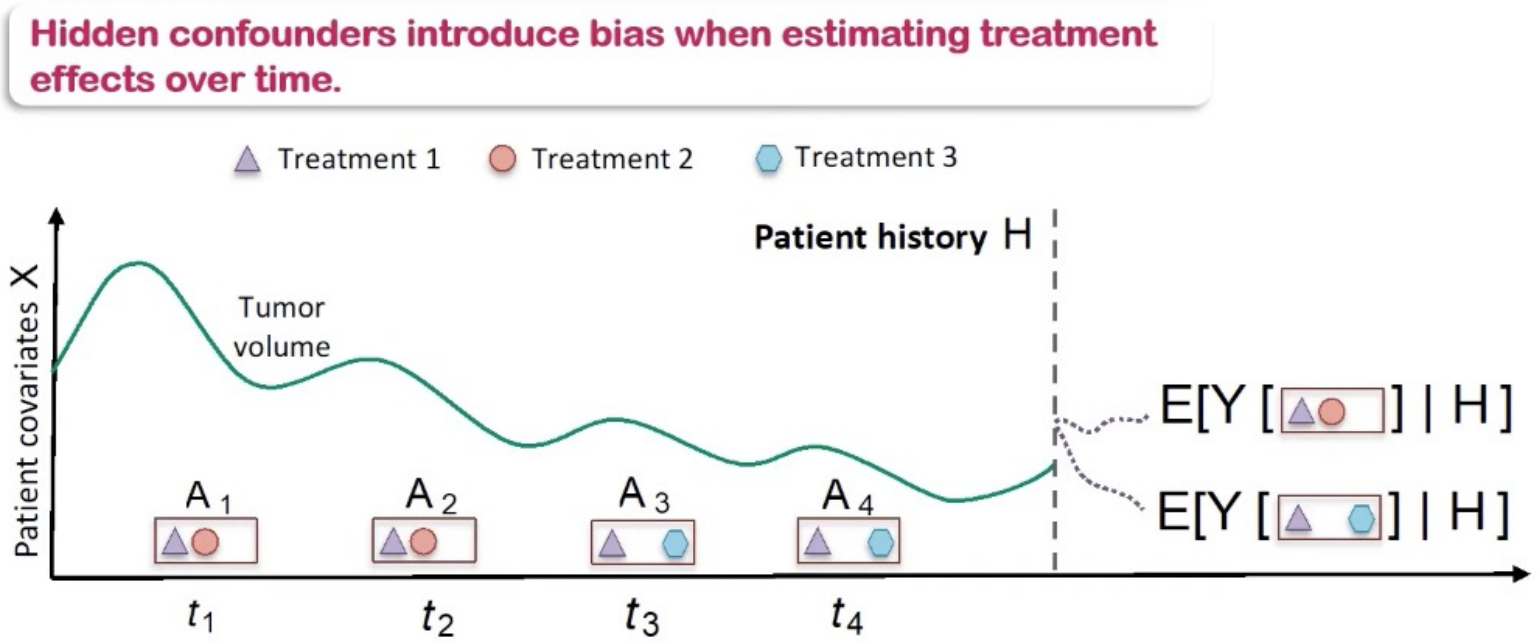
Treatment effect over time

- Aim: estimate ITE of time-dependent treatments



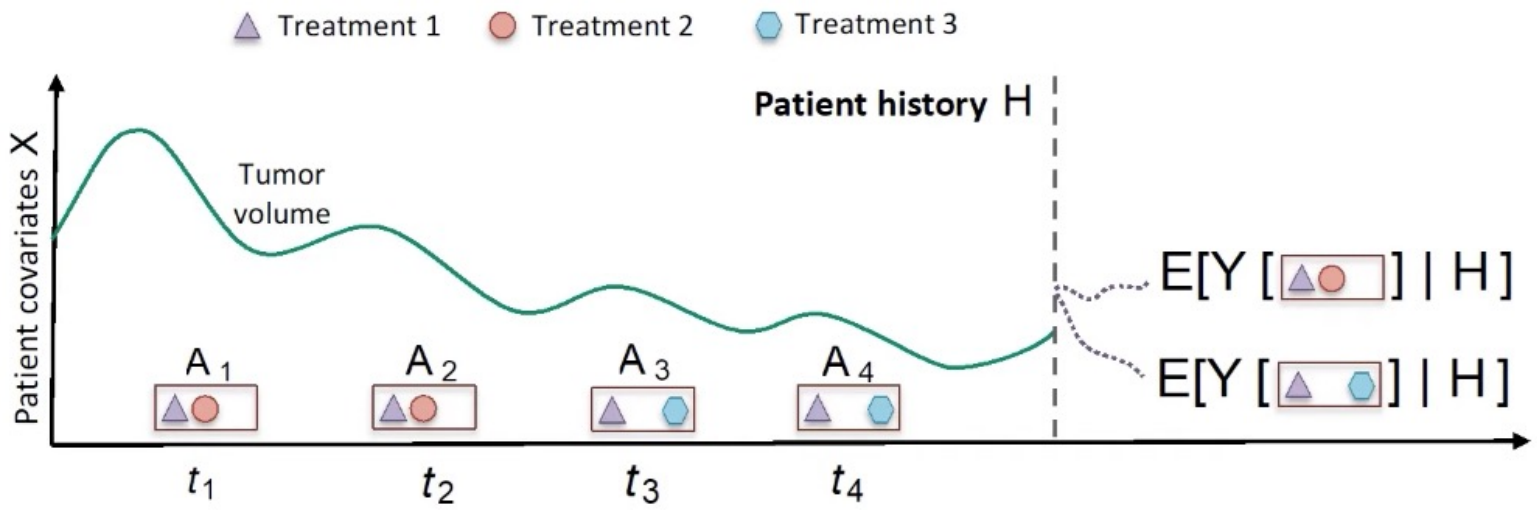
Existing methods for causal effect over time assume there are no unobserved confounders

Hidden confounders



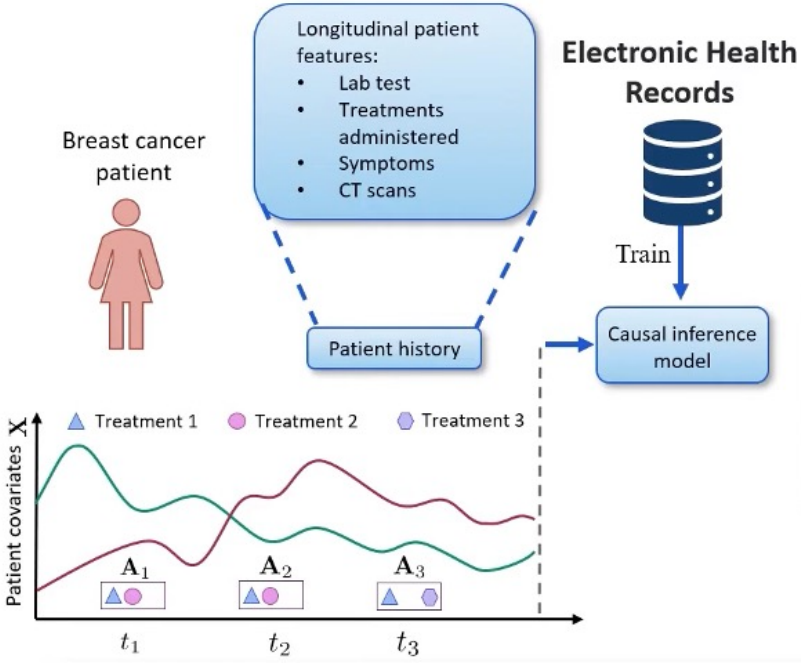
Hidden confounders

Hidden confounders introduce bias when estimating treatment effects over time.



Proposed solution: **infer latent variables that capture the dependencies in the treatment assignments over time and can be used as substitutes for the hidden confounders.**

Longitudinal patient observational data



Longitudinal patient observational data

- Time-dependent patient features: $\bar{\mathbf{X}}_t = (\mathbf{X}_1, \ldots \mathbf{X}_t)$
- Time-dependent treatments: $\bar{\mathbf{A}}_t = (\mathbf{A}_1, \ldots \mathbf{A}_t)$ where

$$\mathbf{A}_t = [A_{t1} \ldots A_{tk}]$$

Observed (factual) outcome given history of covariates $\bar{\mathbf{X}}_t$ and treatments $\bar{\mathbf{A}}_t$: $\mathbf{Y}_{t+1}$

Estimate counterfactual outcomes

Potential outcomes [Rubin (1978), Neyman (1923), Robins & Hernan (2008)] under intended sequence of future treatments:

$$\mathbb{E}[\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \mid \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{X}}_t]$$

Assumptions

→ **Sequential overlap:** $P(\mathbf{A}_t = \mathbf{a}_t \mid \bar{\mathbf{A}}_{t-1} = \bar{\mathbf{a}}_{t-1}, \bar{\mathbf{X}}_t = \bar{\mathbf{x}}_t) > 0 \ \forall \mathbf{a}_t, \forall t$

Estimate counterfactual outcomes

Potential outcomes [Rubin (1978), Neyman (1923), Robins & Hernan (2008)] under intended sequence of future treatments:

$$\mathbb{E}[\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \mid \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{X}}_t]$$

Assumptions

→ **Sequential overlap:** $P(\mathbf{A}_t = \mathbf{a}_t \mid \bar{\mathbf{A}}_{t-1} = \bar{\mathbf{a}}_{t-1}, \bar{\mathbf{X}}_t = \bar{\mathbf{x}}_t) > 0 \; \forall \mathbf{a}_t, \forall t$

Existing methods for estimating treatment effects over time assume that there are no hidden confounders

$$\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \perp\!\!\!\perp \mathbf{A}_t \mid \bar{\mathbf{H}}_t, \text{for all possible treatment plans } \bar{\mathbf{a}}_{\geq t} \text{ and } \forall t$$

Estimate counterfactual outcomes

Potential outcomes [Rubin (1978), Neyman (1923), Robins & Hernan (2008)] under intended sequence of future treatments:

$$\mathbb{E}[\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \mid \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{X}}_t]$$

Assumptions

→ **Sequential overlap:** $P(\mathbf{A}_t = \mathbf{a}_t \mid \bar{\mathbf{A}}_{t-1} = \bar{\mathbf{a}}_{t-1}, \bar{\mathbf{X}}_t = \bar{\mathbf{x}}_t) > 0 \ \forall \mathbf{a}_t, \forall t$

Existing methods for estimating treatment effects over time assume that there are no hidden confounders

$$\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \perp\!\!\!\perp \mathbf{A}_t \mid \bar{\mathbf{H}}_t, \text{for all possible treatment plans } \bar{\mathbf{a}}_{\geq t} \text{ and } \forall t$$

This assumption is untestable in practice!

Estimate counterfactual outcomes

Potential outcomes [Rubin (1978), Neyman (1923), Robins & Hernan (2008)] under intended sequence of future treatments:

$$\mathbb{E}[\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \mid \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{X}}_t]$$

Assumptions

→ **Sequential overlap:** $P(\mathbf{A}_t = \mathbf{a}_t \mid \bar{\mathbf{A}}_{t-1} = \bar{\mathbf{a}}_{t-1}, \bar{\mathbf{X}}_t = \bar{\mathbf{x}}_t) > 0 \ \forall \mathbf{a}_t, \forall t$

Sequential strong ignorability does not hold (hidden confounders)

$\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \not\perp\!\!\!\perp \mathbf{A}_t \mid \bar{\mathbf{H}}_t$, for all possible treatment plans $\bar{\mathbf{a}}_{\geq t}$ and $\forall t$

→ $\mathbb{E}[\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \mid \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{X}}_t]$ **is biased.**

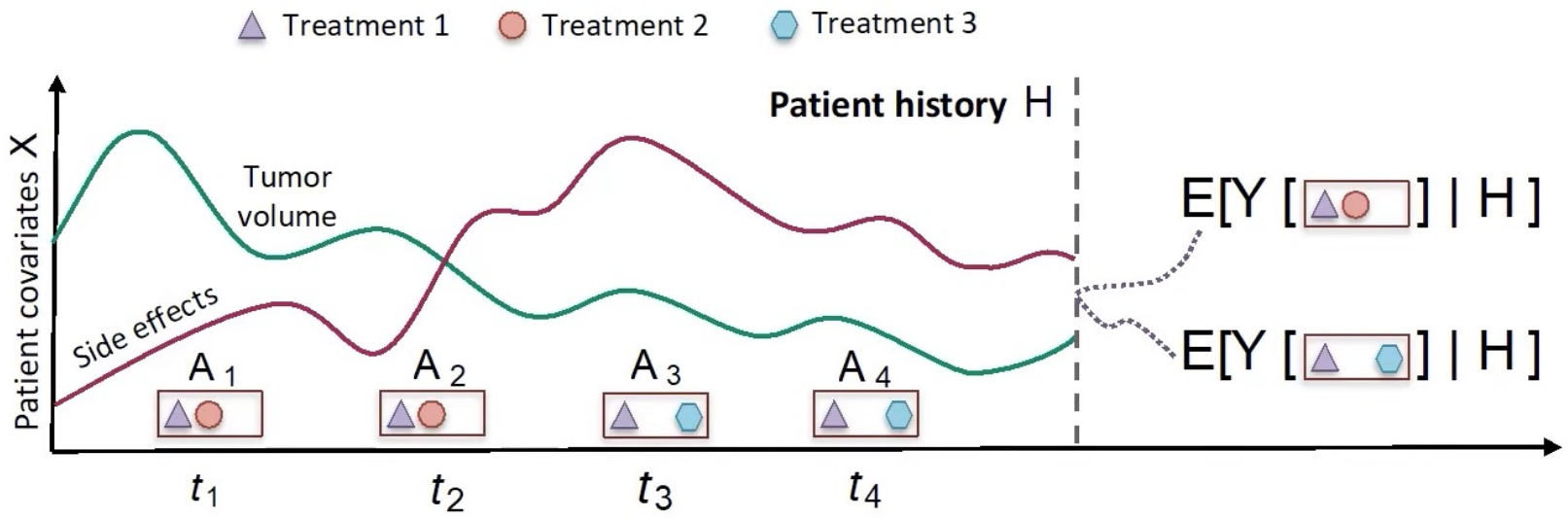
Hidden confounders – from static to temporal setting

The Blessing of Multiple Causes [Wang & Blei, 2019]:

- **Static causal inference setting.**
- **Hidden confounders introduce dependencies in the treatment assignments.**
- **Infer latent variables that capture these dependencies and render the treatments conditionally independent.**

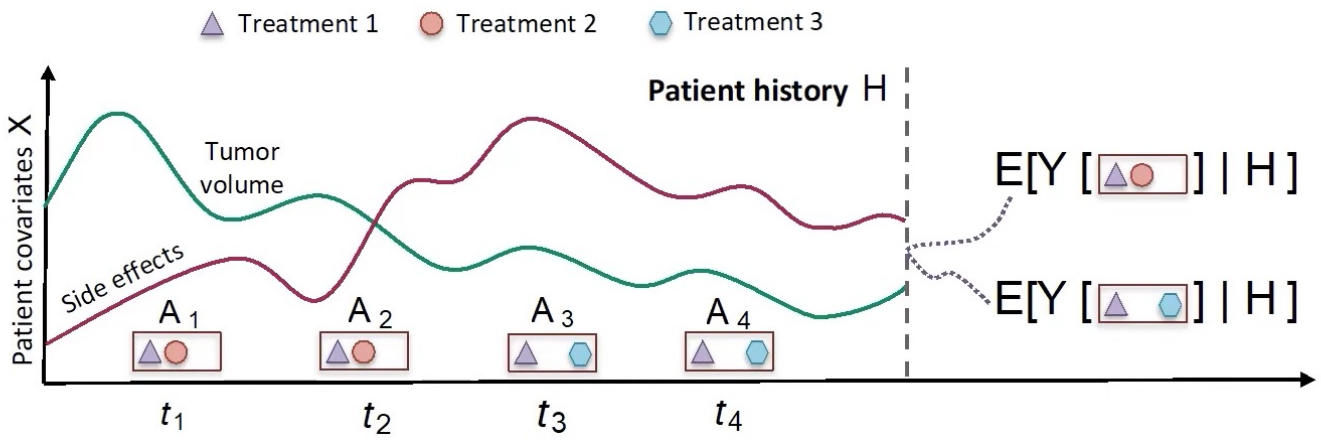
In the temporal setting, the hidden confounders may change over time and may be affected by past treatments and covariates.

Time series deconfounder



Hidden confounders may vary over time and may be affected by previous treatments and covariates.

Time series deconfounder



Hidden confounders may vary over time and may be affected by previous treatments and covariates.

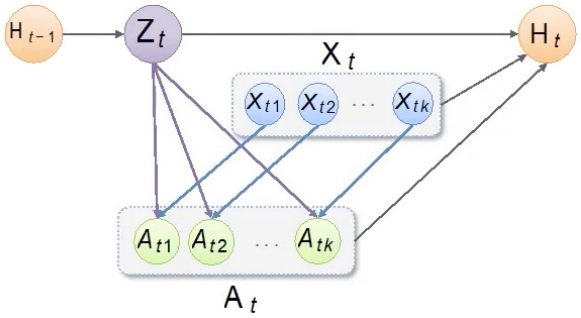
- **Take advantage of the dependencies in the way multiple treatments are assigned over time to infer latent variables:**

$$\bar{\mathbf{Z}}_t = (\mathbf{Z}_1, \ldots, \mathbf{Z}_t)$$

- **Augment the observational dataset with $\bar{\mathbf{Z}}_t$ and use an outcome model to obtain unbiased estimates of the treatment effects.**

Time series deconfounder – factor model

Step 1: Fit factor model over time to infer substitutes for the hidden confounders.

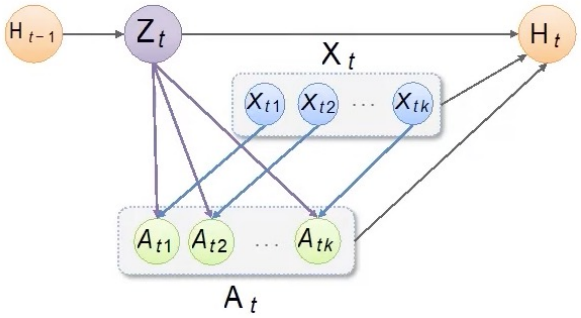


Construct the latent variable $\mathbf{Z}_t = g(\bar{\mathbf{H}}_{t-1})$ as a function of the history $\bar{\mathbf{H}}_{t-1} = (\mathbf{A}_{t-1}, \mathbf{X}_{t-1}, \mathbf{Z}_{t-1})$ such that:

$$p(A_{t1}, \ldots, A_{tk} \mid \mathbf{Z}_t, \mathbf{X}_t) = \prod_{j=1}^{k} p(A_{tj} \mid \mathbf{Z}_t, \mathbf{X}_t)$$

Time series deconfounder – factor model

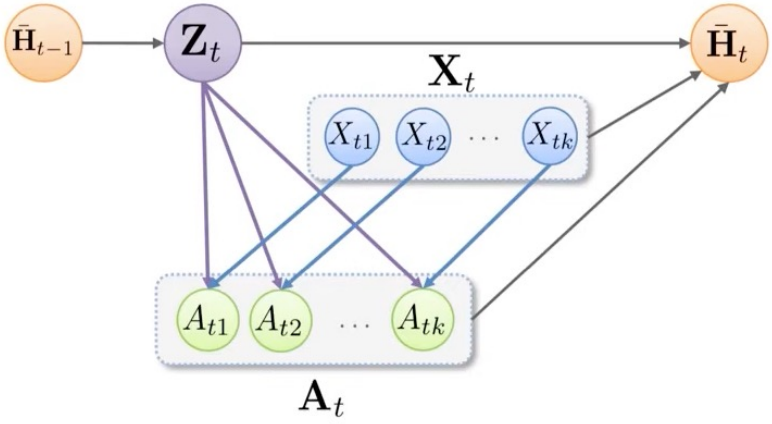
Step 1: Fit factor model over time to infer substitutes for the hidden confounders.

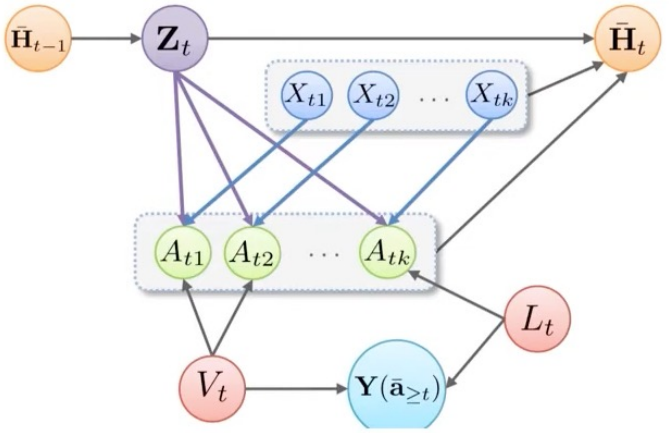


Factor model of the assigned treatments has joint distribution:

$$p(\theta_{1:k}, \bar{\mathbf{x}}_T, \bar{\mathbf{z}}_T, \bar{\mathbf{a}}_T) = p(\theta_{1:k}) p(\bar{\mathbf{x}}_T) \cdot \prod_{t=1}^{T} \left(p(\mathbf{z}_t \mid \bar{\mathbf{h}}_{t-1}) \prod_{j=1}^{k} p(a_{tj} \mid \mathbf{z}_t, \mathbf{x}_t, \theta_j) \right).$$

Time series deconfounder – factor model





Assumption (Sequential single strong ignorability)

$$\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \perp\!\!\!\perp A_{tj} \mid \mathbf{X}_t, \bar{\mathbf{H}}_{t-1},$$

$\forall \bar{\mathbf{a}}_{\geq t}$ and $\forall t \in \{0, \ldots, T\}$ and $\forall j \in \{1, \ldots, k\}$.

Time series deconfounder – Sequential strong ignorability

Theorem

If the distribution of the assigned causes $p(\bar{\mathbf{a}}_T)$ *can be written as the factor model* $p(\theta_{1:k}, \bar{\mathbf{x}}_T, \bar{\mathbf{z}}_T, \bar{\mathbf{a}}_T)$*, we obtain sequential ignorable treatment assignment:*

$$\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \perp\!\!\!\perp (A_{t1}, \ldots, A_{tk}) \mid \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{X}}_t, \bar{\mathbf{Z}}_t,$$

for all $\bar{\mathbf{a}}_{\geq t}$ *and for all* $t \in \{0, \ldots, T\}$.

Evaluate factor model

- Use predictive checks (Rubin, 1984) to assess how well the factor model captures the distribution of treatments at each timestep.
- The inferred substitutes for the hidden confounders Z_t also need to satisfy positivity, i.e.

$$P(\mathbf{A}_t = \mathbf{a}_t \mid \bar{\mathbf{A}}_{t-1} = \bar{\mathbf{a}}_{t-1}, \bar{\mathbf{Z}}_t = \bar{\mathbf{z}}_t, \bar{\mathbf{X}}_t = \bar{\mathbf{x}}_t) > 0.$$

Time series deconfounder – Outcome model

Step 2: Sample $\hat{\bar{\mathbf{Z}}}_t = (\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_t)$ from the factor model and augment observational dataset. Fit outcome model to estimate counterfactual outcomes.

Longitudinal patient observational data

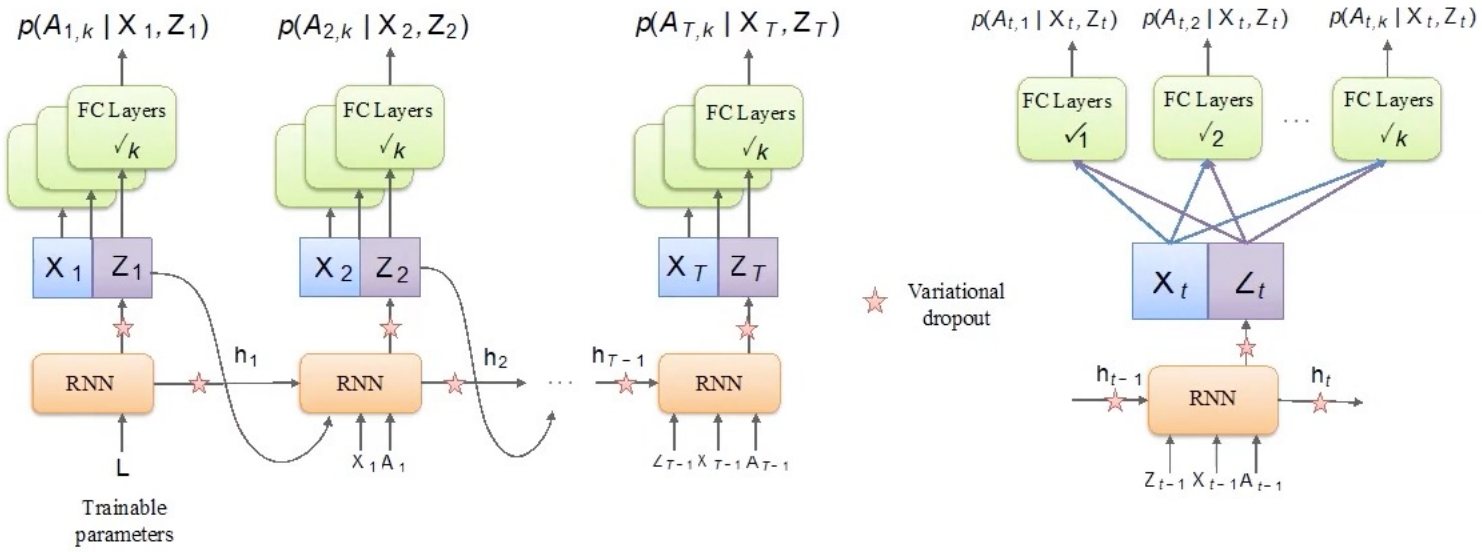
- Time-dependent patient features: $\bar{\mathbf{X}}_t = (\mathbf{X}_1, \ldots \mathbf{X}_t)$
- Time-dependent treatments: $\bar{\mathbf{A}}_t = (\mathbf{A}_1, \ldots \mathbf{A}_t)$

Latent variables sampled from factor model: $\hat{\bar{\mathbf{Z}}}_t = (\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_t)$

$\mathbb{E}[\mathbf{Y}(\bar{\mathbf{a}}_{\geq t}) \mid \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{X}}_t, \hat{\bar{\mathbf{Z}}}_t]$ is unbiased.

Factor model implementation

Proposed architecture for the factor model: recurrent neural network(RNN) with multitask output and variational dropout.



$$\mathbf{Z}_1 = \mathrm{RNN}(\mathbf{L}) \qquad \mathbf{Z}_t = \mathrm{RNN}(\bar{\mathbf{Z}}_{t-1}, \bar{\mathbf{X}}_{t-1}, \bar{\mathbf{A}}_{t-1}, \mathbf{L})$$

$$A_{tj} = \mathrm{FC}(\mathbf{X}_t, \mathbf{Z}_t; \theta_j), \text{ for all } j = 1, \ldots k$$

Experiments on synthetic data

- **Validate method on synthetic data where we have ground truth knowledge about hidden confounders and where we can vary their effect on the treatment assignment and patient outcomes.**
- **Build synthetic dataset using p−order autoregressive processes:**

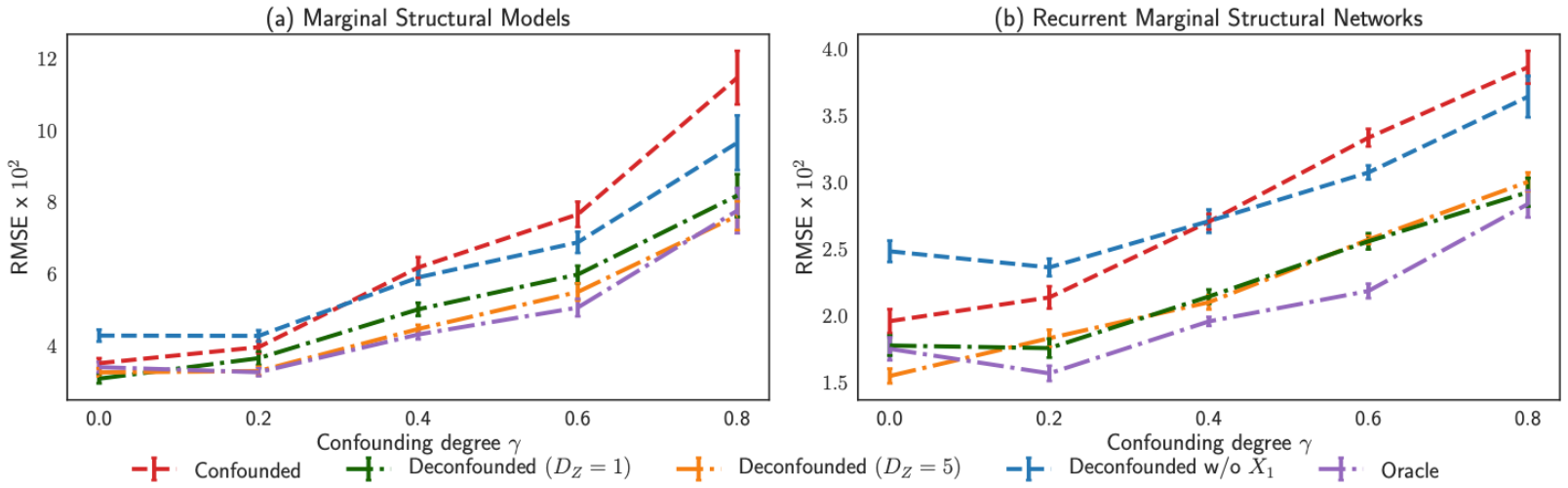
$$X_{t,j} = \frac{1}{p}\sum_{i=1}^{p}(\alpha_{i,j}X_{t-i,j} + \omega_{i,j}A_{t-i,j}) + \eta_t,$$

$$Z_t = \frac{1}{p}\sum_{i=1}^{p}(\beta_i Z_{t-i} + \sum_{j=1}^{k}\lambda_{i,j}A_{t-i,j}) + \epsilon_t,$$

$$\pi_{tj} = \gamma_A \hat{Z}_t + (1-\gamma_A)\hat{X}_{tj}, \qquad A_{tj} \mid \pi_{tj} \sim \text{Bernoulli}(\sigma(\lambda\pi_{tj})),$$

$$\mathbf{Y}_{t+1} = \gamma_Y Z_{t+1} + (1-\gamma_Y)\left(\frac{1}{k}\sum_{j=1}^{k}X_{t+1,j}\right).$$

Experiments on synthetic data



- **Outcome models:**
  Marginal Structural Models [Robins et al.] and Recurrent Marginal Structural Networks [Lim et al.].
- **Root mean squared error (RMSE) obtained for one-step ahead estimation of treatment effects.**
- **The parameters $\gamma = \gamma_A = \gamma_Y$ control the amount of hidden confounding.**

Experiments on ICU dataset

- **6256 patients, with 25 covariates (lab tests and vital signs) per person and trajectories up to 50 days.**
- **Estimate the effect of antibiotics, vassopressors and mechanical ventilator on patient covariates.**
- **Hidden confounding is present since patient comorbidities and several lab tests were not included in the data.**

Outcome model	White blood cell count		Blood pressure		Oxygen saturation	
	MSM	R-MSN	MSM	R-MSN	MSM	R-MSN
Confounded	3.90 ± 0.00	2.91 ± 0.05	12.04 ± 0.00	10.29 ± 0.05	2.92 ± 0.00	1.74 ± 0.03
$D_Z = 1$	3.55 ± 0.05	2.62 ± 0.07	11.69 ± 0.14	9.35 ± 0.11	2.42 ± 0.02	1.24 ± 0.05
$D_Z = 5$	3.56 ± 0.04	2.41 ± 0.04	11.63 ± 0.10	9.45 ± 0.10	2.43 ± 0.02	1.21 ± 0.07
$D_Z = 10$	3.58 ± 0.03	2.48 ± 0.06	11.66 ± 0.14	9.20 ± 0.12	2.42 ± 0.01	1.17 ± 0.06
$D_Z = 20$	3.54 ± 0.04	2.55 ± 0.05	11.57 ± 0.12	9.63 ± 0.14	2.40 ± 0.01	1.28 ± 0.08

Discussion and limitations

- **The Time Series Deconfounder enables the estimation of treatment effects over time using weaker assumptions than existing methods.**

- **Identifiability of the potential outcomes using the deconfounder framework may represent an issue:**

  → **non-identifiability will be indicated by the high variance of the estimated outcomes.**

Take-aways

- The availability of longitudinal observational data about patients prompted the development of methods for modeling the effects of treatments on the disease progression in patients.
- Hidden confounders introduce bias when estimating treatment effects over time.
- Time Series Deconfounder takes advantage of patterns in the multiple treatment assignments over time to infer latent variables that can be used as substitutes for the hidden confounders.

https://github.com/vanderschaarlab/mlforhealthlabpub

https://github.com/ioanabica/Time-Series-Deconfounder

Bica, I., Alaa, A.M. & van der Schaar, M. "Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders", ICML 2020

References

- Robins, J. M. (1997). Causal inference from complex longitudinal data. In Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics, M. Berkane (eds), vol 120, New York, NY: Springer
- Bang, H., Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4), 962-973.
- Imai, K., Ratkovic, M. (2015). Robust estimation of inverse probability weights for marginal structural models. Journal of the American Statistical Association, 110(511), 1013-1023.
- Some slides are from:
  - Ioana Bica, et al. Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders. ICML 2020.
  - STA 640 — Causal Inference Chapter 9 - Sequential/Longitudinal Treatments. Department of Statistical Science. Duke University.

Reading Materials

- Hernan, M.A. and Robins, J.M. (2020) Causal Inference: What If. Chapter 19 & 21.
  - https://www.hsph.harvard.edu/miguel-hernan/wp-content/uploads/sites/1268/2023/07/hernanrobins_WhatIf_19jul23.pdf
- Bica I, Alaa A, Van Der Schaar M. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders[C]. ICML. PMLR, 2020: 884-895.
  - https://proceedings.mlr.press/v119/bica20a/bica20a.pdf

Thank you!
Q&A