Time-varying Causal Inference

Klaus Frieler
Methods Consultant
Max Planck Institute for Empirical Aesthetics
Frankfurt am Main

19.01.2026

Agenda

- Introduction and basics
- Estimations methods (G-computation IPW, TMLE, SDR)
- Examples with the `lmtp` package

Introduction

Time-varying treatments

- In longitudinal studies, there is typically not only one treatment, but a sequence of time-points, where treatment is administered and confounders are measured.
- This brings about a whole new set of complication and complexity for causal inference.

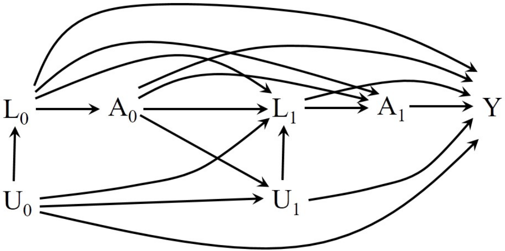
  "We have done our best to simplify those concepts [...], this is still one of the most technical chapters in the book. Unfortunately, further simplification would result in too much loss of rigor."

  (Hernan & Robins, Introduction to Ch. 19)

Time-varying treatments

- There are many different types of strategies (plans, policies, protocols, regimes), how treatments are administered.
  - Static regime (always the same)
  - Dynamic regime (treatment depending on some covariates)
  - Random strategy (treatment has a random component)
  - Optimal strategy (maximizes some counterfactual outcome, a target)
- Due to the multiple time points, the causal effect is not uniquely defined anymore, one could compare any strategy with any other.
- For binary treatments and K time points, there are 2^K possible counterfactual treatments.
- Often, the "always treated" is compared to the "never treated".

Treatment confounder feedback



- Treatment-confounder feedback: A treatment affects the confounder and vice versa
- There are arrows from L_k to A_k as well as from A_{k-1} to L_k.
- However, confounders can be time-varying without treatment confounder feedback.
- Confounding is time-varying if for the baseline covariates L_0 holds

$$E[Y^{\bar{a}}|L_0] \neq E[Y|A = \bar{a}, L_0]$$

Identifiability

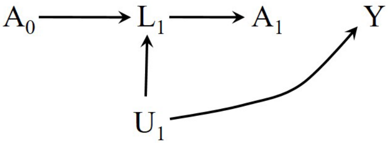
- Identifiability for time-varying system, depends on generalized versions of the usual assumption, consistency, positivity, SUTVA, and sequential exchangeability.
- Complicated definitions, I spare you that.
- Forms of sequential exchangeability:
    - Sequential conditional exchangeability ($=$ sequential exchangeability)
    - Static sequential exchangeability (weaker)
    - Full sequential exchangeability and unconditional SE (for sequential randomized experiments, all treatment assignments are random).
- Sequential randomized experiments do no have time-varying confounding.

Motivating example (from Ch. 20.2)

- Study with $N = 320$ individuals with HIV, two time points $k = 0, 1$
- Treatment $A_0 = 1$ is randomly assigned at baseline with $p = .05$ (we can omit L_0 thus)
- Treatment $A_1 = 1$ is is randomly assigned after one month, based on CD4 cell counts ($=$ covariate L_1): $p = .4$ if $L_1 = 0$ (high), and $p = .8$ if $L_1 = 1$ (low). (the higher the better). Outcome is a continuous measure of health (the higher the better, think CD4 cell counts).

Motivating example (from Ch. 20.2)

N	A_0	L_1	A_1	Y
24	0	0	0	84
16	0	0	1	84
24	0	1	0	52
96	0	1	1	52
48	1	0	0	76
32	1	0	1	76
16	1	1	0	44
64	1	1	1	44



Motivating example (from Ch. 20.2)

- Consistency, positivity and sequential exchangeability holds.
- $\rightarrow$ differences in means can be interpreted causally.
- The outcome is independent of A_1, identical means
- The outcome is independent of A_0, $E[Y|A_0 = 0] = 60$ and $E[Y|A_0 = 1] = 60$.
- $\rightarrow$ no causal effect. G-null theorem implies always vs. never treated to be absent: $E[Y^{\bar{1}}] - E[Y^{\bar{0}}] = 0$.
- But: $E[Y|A_0 = 0, A_1 = 0] = 68$ and $E[Y|A_0 = 1, A_1 = 1] = 54.7$, hence $E[Y^{a_0=1,a_1=1}] - E[Y^{a_0=0,a_0=0}] = -13.3 \neq 0$
- For $L_1 = 0$ and $L_1 = 1$ separately:
  $E[Y|A_0 = 1, L_1 = 0, A_1 = 1] - E[Y|A_0 = 0, L_1 = 0, A_1 = 0] = 76 - 84 = -8 \neq 0$
  $E[Y|A_0 = 1, L_1 = 1, A_1 = 1] - E[Y|A_0 = 0, L_1 = 1, A_1 = 0] = 44 - 52 = -8 \neq 0$

Estimation Methods

Time-based Estimation Methods

- For time-varying confounders and treatment confounder feedback, traditional methods cannot be used, even with sufficient longitudinal data.
- Several methods have been proposed:
  - G-estimation,
  - Sequential IPW,
  - Targeted minimum loss-based estimation (TMLE),
  - Sequential doubly robust estimators (SDR).

  All can be seen as different estimation methods of main G-estimation idea, with different advantages and disadvantages.

Notations

- Interventions (treatments and covariates at time point $t \in \{1, \dots, \tau\}$R will be written A_t and L_t, with outcome Y, measured at time $\tau + 1$
- The history of a variables X up to and includingtime t is $\bar{X}_t = (X_1, X_2, \dots, X_t)$
- The treatment history $H_t = (\bar{A}_{t-1}, \bar{L}_t)$
- Define a shift function $d(a_t, l_t, \epsilon_t)$ which encodes the shift from the observed to the counterfactual world, based on the treatment and covariates at time t as well as an optional "randomizer" ϵ_t.
- A (average) causal effect built from hypothetical treatment histories $\bar{A}^d$, depending on the shift function as $\theta = E[Y^{\bar{A}^d}]$
- *Example.* Define the constant shift functions $d_k(a_t, l_t, \epsilon_t) = k$ for $k \in \{0, 1\}$ then $\bar{A}^{d_0} = \bar{0} = (0, \dots, 0)$ (never treat) and $\bar{A}^{d_1} = \bar{1} = (1, \dots, 1)$ (always treat) and

$$\text{ATE}_{\text{always - never}} = E[Y^{\bar{1}} - Y^{\bar{0}}]$$

G-methods for time-varying treatments

Recall, that

$$E[Y^{a_1}] = \sum_{l_1} E[Y|A_1 = a_1, L_1 = l_1]Pr[L_1 = l_1],$$

the counterfactual means are a weighted sum conditional on the confounders (given identifiability).

This needs to be generalized to include treatment and confounder history.

In the simple example:

$$E[Y^{a_0,a_1}] = \sum_{l_1} E[Y|A_0 = a_0, A_1 = a_1, L_1 = l_1]Pr[L_1 = l_1|a_0],$$

In the general case, for deterministic strategies and history of length K

$$E[Y^{\bar{a}}] = \sum_{\bar{l}} E[Y|\bar{A}, \bar{L} = \bar{l}] \prod_{k=0}^{K} Pr[L_k = l_k|\bar{a}_{k-1}, \bar{l}_{k-1}],$$

However, this formula can not be calculated in many cases, particularly in high dimensions.

G-Methods for time-varying treatments

- Many possibilities to estimate the general formula.
- Following the paper and tutorial by Diaz, Williams, Hoffman & Schenck (2023).
- And using their R package `lmtp` for demonstration.
- Basic workflow: calculate counterfactual means $\theta = E[Y^{\bar{A}^d}]$ for shift functions encoding a treatment policy (using `lmtp_tlme` or `lmtp_sdr()`) and from this a contrast between policies, using `lmtp_contrast`.
- This is named Modified Treatment Policy (MTP) (lmtp = Longitudinal MTP), more general than ATE.

ICE or G-computation

Central concepts and ideas are recursion and efficient influence functions.
Set $\mathbf{m}_{\tau+1} = Y$, let $A^d = d(A_t, H_t)$. For $t = 1, \ldots, \tau$, **recursively** define

$$\mathbf{m}_t : (a_t, h_t) \mapsto E[\mathbf{m}_{t+1}(A^d_{t+1}, H_{t+1}) | A_t = a_t, H_t = h_t]$$

with

$$\theta = E[\mathbf{m}_1(A^d_1, L_1)]$$

which is equal to $E[Y^{\bar{A}^d}]$ under positivity and sequential exchangeability assumptions.
The functions $\mathbf{m}_t$ are predictive functions, e. g., regression models for the outcome over treatment and covariates (linear or logistic or whathaveyou).
This provides already a plug-in (substitution estimator), called G-computation, or iterative conditional expectation (ICE), however, it is not recommended.
Example. See Tutorial `https://beyondtheate.com/02_info_d.html`.

Sequential IPW estimator

Define the **density ratio**

$$r_t(a_t, h_t) = \frac{g_t^d(a_t, h_t)}{g_t(a_t, h_t)}$$

where $g_t^d(a_t, h_t)$ is the probability density post-intervention $d(a_t, h_t)$ and $g_t(a_t, h_t)$ is the observed density.

The ratio can be estimated using a “classification trick”, which allows to use more general and potentially better Machine Learning methods.

The **sequential IPW estimator** is then simply given by recursive weighting using the density ratios:

$$\theta = E[\left\{\prod_{t=1}^{\tau} r_i(a_t, h_t)\right\} Y]$$

Classification trick for density ratio calculation

The classification trick is based on stacking the observed (indexed by $\Lambda = 0$) with an identical copy of the data, except that the treatment a_t is substituted with the intervened treatment a_t^d (indexed by $\Lambda = 1$).
Then

$$\begin{aligned} r_t(a_t, h_t) &= \frac{g_t^d(a_t, h_t)}{g_t(a_t, h_t)} \\ &= \frac{P^\lambda(a_t, h_t|\Lambda = 1)}{P^\Lambda(a_t, h_t|\Lambda = 0)} \\ &= \frac{P^\lambda(\Lambda = 1|a_t, h_t)P^\lambda(a_t, h_t)}{P^\lambda(\Lambda = 1)} \times \frac{P^\lambda(\Lambda = 0)}{P^\lambda(\Lambda = 0|a_t, h_t)P^\lambda(a_t, h_t)} \\ &= \frac{P^\lambda(\Lambda = 1|a_t, h_t)}{P^\lambda(\Lambda = 0|a_t, h_t)} \end{aligned}$$

using Bayes' Law and because $P^\lambda(\Lambda = 0) = P^\lambda(\Lambda = 1) = \frac{1}{2}$ by construction. The last ratio can then estimated with ML predicting Λ.
Note: `lmtp` packages uses the `SuperLearner` package, which combines predictions from several user-specified methods (e. g., logistic regression, Random Forest).

Targeted minimum- loss.based estimation (TMLE)

Key to constructing the TMLE and SDR estimators is the **efficient influence function (EIF)**.

Further assumptions needed:

1. A is either discrete or A is continuous and piecewise smooth invertible
2. The shift function d does not depend on the observed distribution P.

The EIF is defined with help of a function ϕ on the observed data via outcome regression $\mathbf{m}_t$ and density ratios r_t:

$$\phi_t : o \mapsto \sum_{s=t}^{\tau} \left(\prod_{k=t}^{s} r_k(a_k, h_k) \right) \left\{ \mathbf{m}_{s+1}(a_{s+1}^d, h_{s+1}) - \mathbf{m}_s(a_s, h_s) \right\} + \mathbf{m}_t(a_t^d, h_t)$$

with

$$\theta = E[\mathbf{m}_1(A_1^d, L_1)]$$

and finally $\text{EIF}(o) = \phi_1(o) - \theta$

TMLE and SDR

Special case: only one time point:

$$r_i(a,w)\left\{(Y_i - \mathbf{m}_i(a,l)\right\} + \mathbf{m}_i(a^d,l) - \theta$$

which is similar to a doubly robust estimator or the augmented IPW estimator.
The expected value of the EIF is zero by definition, which can be used to construct an iterative algorithm for estimation.

$$\begin{aligned}
0 = E[\phi_1(O) - \theta] &= \frac{1}{n}\sum_i r_i(a_i,l_i)\left\{Y_i - \mathbf{m}_i(a_i,l_i)\right\} + E[m_i(a^d,l)] - \theta &(1)\\
&= \frac{1}{n}\sum_i r_i(a_i,l_i)\left\{Y_i - \mathbf{m}_i(a_i,l_i)\right\} + \theta - \theta &(2)\\
\Rightarrow &\frac{1}{n}\sum_i r_i(a_i,l_i)\left\{Y_i - \mathbf{m}_i(a_i,l_i)\right\} = 0 &(3)
\end{aligned}$$

The score equation can be solved using a GLM with offset and the intercept serves as a correction for the next iteration (“targeting”).

Sequentially Doubly Robust Estimation (SDR)

SDR is based on the fact that

$$E[\phi_t(O)|A_t = a_t, H_t = h_t] \approx \mathbf{m}_t(a_t^d, h_t)$$

Iterative regression of pseudo-outcomes on covariates and treatment backwards $(t = \tau, \ldots, 1)$ with $\phi_{\tau+1}(O_i) = Y_i$.
Final estimate $\theta = \frac{1}{n}\sum_{i=1}^{n} \phi_1(O_i)$.

Properties and recommendations

	IPW	G-computation	TMLE	SDR
Simple to implement	$\star$			
Uses outcome regression		$\star$	$\star$	$\star$
Uses the propensity score	$\star$		$\star$	$\star$
Valid inference with machine learning			$\star$	$\star$
Substitution estimator		$\star$	$\star$	
$\tau + 1$ doubly robust			$\star$	$\star$
Sequentially doubly robust				$\star$
Recommendation	Don't use	Don't use	**Recommended**	Use (only dyn.)

Cons G-computation, IPW: Correctness of bootstrap requires pre-specified parametric models, consistency requires correct estimation of all regressions.

Sequential and doubly robust can be worse if both propensity and outcome models are misspecified (Kang & Schafer, 2007)!

Examples using the `lmtp` package

Examples using the `lmtp` package

- Motivating example (Ch. 20)
- Smoking cessation and weight gain (Ch. 12).
- Simulated data from Kang & Schaefer (2007)
- Flying Steps data

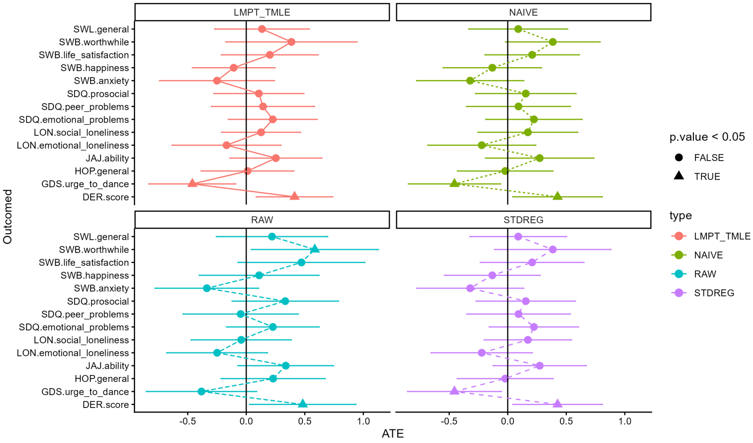
Flying Steps

- Flying Steps Education provides (street/urban) dance classes for Berlin schools since 2024 (in part substituting for regular PE lessons).
- In 2025, we had the occasion to run a pilot study to assess effects on dance classes on well-being, social emotional development and cognitive abilities (Visual Working Memory).
- Two measurement time points ($\approx$ 50 d apart) for kids with and without dance lessons.
- Battery of established measures, demographics, Big5, physical activities, dance-related questions.

Flying Steps

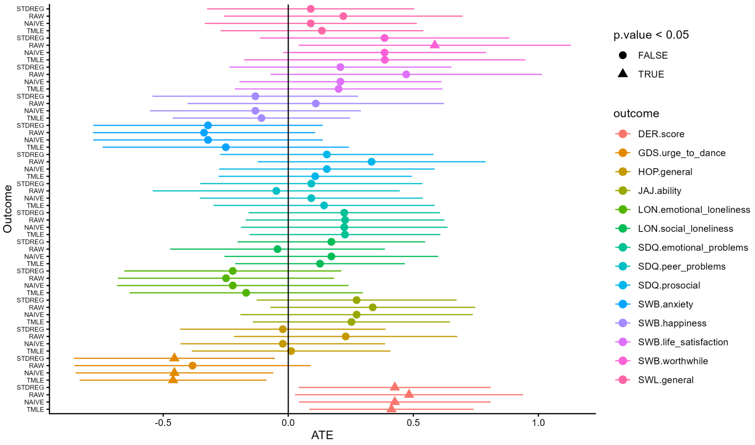
- We compared different ATE estimation methods using demographics, Big5, and physical activities (PAC) as covariates
- 14 outcomes: SDQ (school difficulties, 3 subscales), LON (Loneliness scale, 2 subscales), HOP (Hope, 1 subscale), SWB (subjective well-being, 4 subcales), SWL (Worth living, 1 subscale), JAJ (Visual Working Memory), GDS (subscale urge to dance), DER (Dance Emotion Recognition test).
- ATE for baseline, as kids had already dance lessons.
  - Raw data differences (RAW)
  - simple regression estimates (NAIVE)
  - standardization with the `stdReg` package (STDREG)
  - TMLE with the `lmtp` package (TMLE)
  - $\mathrm{N} = 94$, $N_{\text{control}} = 28.$, $N_{\text{dance}} = 68$
- ATE for both time points (static treatment), only TMLE.

Flying Steps: ATE at baseline for 14 outcomes

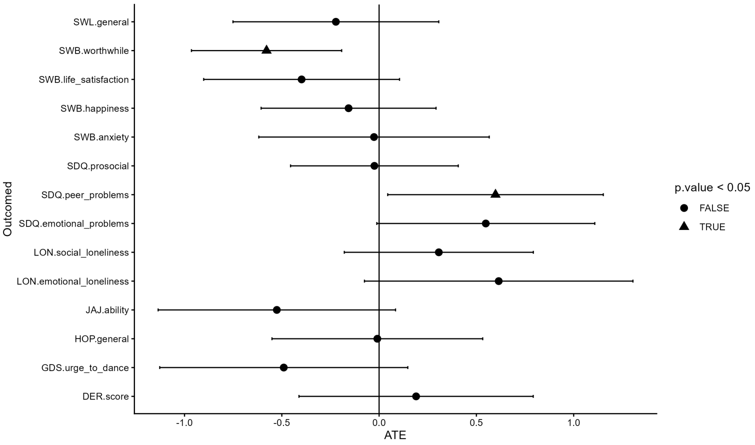


Krippendorff's α: Estimates: .9, Significance: .71

Flying Steps: ATE at baseline for 14 outcomes



Flying Steps: Full ATE for 14 outcomes



Conclusion

- Powerful methods, though complicated.
- lmtp makes life really easy (plus more general treatment policies and causal effects, allows ML).
- Troubleshooting and trustworthiness is a problem without deeper understanding of the methods.
- How do sample sizes, misspecifications of models, all those colliders, missing data, measurement errors etc. influence the misspecified?
- Flying Steps data at baseline show no improvement over simpler methods.
- However, much easier to analyze data for complete data.
- For the Kang & Schafer data, lmpt_tmle completely fails for completely misspecified models, sometimes worse bias than non-robust methods.