Longitudinal Generalizations of the Average Treatment Effect on the Treated for Multi-valued and Continuous Treatments

Herb Susmann

Division of Biostatistics
Department of Population Health
NYU Grossman School of Medicine

ENAR 2025

NYU Grossman School of Medicine

Statistics > Methodology

[Submitted on 9 May 2024 (v1), last revised 24 Oct 2024 (this version, v2)]

Longitudinal Generalizations of the Average Treatment Effect on the Treated for Multi-valued and Continuous Treatments

Herbert Susmann, Nicholas T. Williams, Kara E. Rudolph, Iván Díaz

The Average Treatment Effect on the Treated (ATT) is a common causal parameter defined as the average effect of a binary treatment among the subset of the population receiving treatment. We propose a novel family of parameters, Generalized ATTs (GATTs), that generalize the concept of the ATT to longitudinal data structures, multi-valued or continuous treatments, and conditioning on arbitrary treatment subsets. We provide a formal causal identification result that expresses the GATT in terms of sequential regressions, and derive the efficient influence function of the parameter, which defines its semi-parametric efficiency bound. Efficient semi-parametric inference of the GATT requires estimating the ratios of functions of conditional probabilities (or densities); we propose directly estimating these ratios via empirical loss minimization, drawing on the theory of Riesz representers. Simulations suggest that estimation of the density ratios using Riesz representation have better stability in finite samples. Lastly, we illustrate the use of our methods to evaluate the effect of chronic pain management strategies on the development of opioid use disorder among Medicare patients with chronic pain.

Takeaways

- We generalize the **average treatment effect on the treated** to longitudinal settings and continuous treatments.
- We show how **Riesz learning** can be used to stabilize estimation of longitudinal propensity scores.

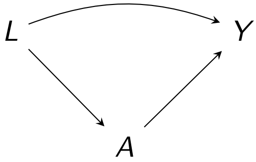
Example

- **Exposure:** measurements of human PM2.5 exposure over time.
- **Outcome:** respiratory function.
- **Intervention:** reduce PM2.5 by 10%.
- **Causal question:** What would be the average respiratory function under the intervention *among participants with high PM2.5 exposures*?

Average Treatment Effect on the Treated (ATT)

- Structural Causal Model:

$$L = f_L(U_L),$$
$$A = f_A(L, U_A),$$
$$Y = f_Y(L, A, U_Y).$$



- Causal estimand:

$$\Psi = \mathrm{E}\left[Y(1) - Y(0) \mid A = 1\right].$$

- Identification assumptions
  - No unmeasured confounding: $U_Y \perp\!\!\!\perp U_A$ or $U_Y \perp\!\!\!\perp U_L$.
  - Positivity: if $P(A = 1|X) > 0$, then $P(A = 0|X) > 0$.

NYU Grossman School of Medicine

Generalizing the ATT

- Longitudinal data structures
  - Now we have $L_1, A_1, L_2, A_2, \ldots, L_\tau, A_\tau, Y$
- Multi-valued or continuous treatments, modified treatment policies
  - Now A can live in any space $\mathcal{A}$
- Condition on arbitrary treatment status
  - Instead of conditioning on $A = 1$, we condition on $A \in \mathcal{B} \subset \mathcal{A}$

Longitudinal Notation

Notation

- H_t is the history of all variables up to t.
- Overlines indicate *history*: e.g. $\bar{A}_\tau$ is A_t from $t = 1$ to τ.
- Underlines indicate *future*: e.g. $\underline{A}_1$ is A_t from $t = 1$ to τ.

Longitudinal Structural Causal Model

For $t \in \{1, \ldots, \tau\}$,

$$\begin{aligned} L_t &= f_{L_t}(A_{t-1}, H_{t-1}, U_{L,t}), \\ A_t &= f_{A_t}(H_t, U_{A,t}), \\ Y &= f_Y(A_\tau, H_\tau, U_Y). \end{aligned}$$

Notation

H_t is the history of all variables up to right before A_t.

Longitudinal DAG

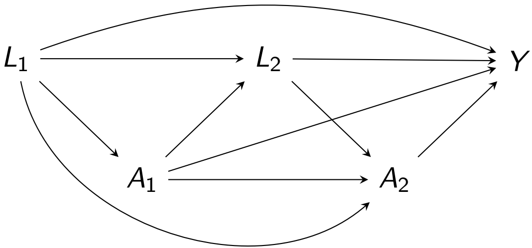


Figure: Example of the assumed Longitudinal DAG with 2 time points.

Natural Value of Treatment

Intervening to set $A_1 \leftarrow a_1$ induces a counterfactual value for A_2, which is called its *natural value of treatment*, written $A_2(a_1)$.

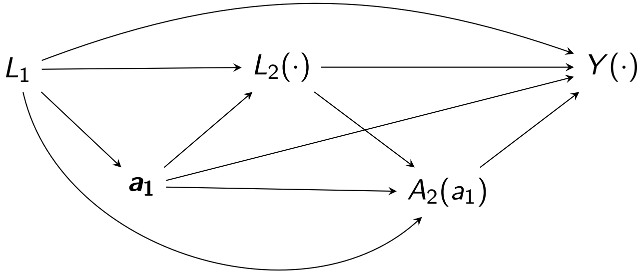


Figure: Suppose we fix $A \leftarrow a_1$, where a_1 may be a function of L_1. This induces counterfactual values of L_2, A_2, and Y.

Modified Treatment Policies

Definition (Díaz JASA 2023)

The intervention A_t^d is called a **Longitudinal Modified Treatment Policy** (LMTP) if it has a representation

$$A_t^d = d(A_t(\bar{A}_{t-1}^d), H_t(\bar{A}_{t-1}^d))$$

for an arbitrary function d.

Notation

- $\bar{A}_{t-1}^d$ is the history of the intervention up to time $t-1$.
- $A_t(\bar{A}_{t-1}^d)$ is the natural value of treatment at time t under intervention history.
- $H_t(\bar{A}_{t-1}^d)$ is counterfactual history under intervention history.

Modified Treatment Policies

Example (Shift Modified Treatment Policy)

Suppose there exists some u_t such that $P(A_t < u_t | H_t = h_t) = 1$ for all $t \in \{1, \ldots, \tau\}$. For some fixed δ, define the intervention as

$$d(a_t, h_t) = \begin{cases} a_t + \delta, & \text{if } a_t \leq u_t(h_t) - \delta, \\ a_t, & \text{if } a_t > u_t(h_t) - \delta. \end{cases}$$

Shift the natural value of treatment up by δ, as long as we stay within the support of the data. Otherwise, leave the natural value of treatment as is.

Generalized ATT Parameter

- We propose a generalized version of the ATT, which we call Generalized ATTs (GATTs).
- The GATT parameter is defined as:

$$\theta^* = \mathsf{E}\left[Y(\bar{A}^d) \mid \bar{A}(d) \in \bar{\mathcal{B}} \right].$$

- The vector $\bar{A}(d) = (A_1, A_2(d_1), ..., A_\tau(d_{\tau-1}))$ is called the *longitudinal natural value of treatment.*
- The longitudinal conditioning set $\bar{\mathcal{B}}$ is an arbitrary subset of the longitudinal treatment space.

NYU Grossman School of Medicine

Generalizing the ATT

- The GATT parameter is defined as:

$$\theta^* = \mathsf{E}\left[Y(\bar{A}^d) \mid \bar{A}(d) \in \bar{\mathcal{B}} \right].$$

- Note that we condition on the *longitudinal natural value of treatment* $\bar{A}(d) \in \bar{\mathcal{B}}$, rather than the *observed exposures* $\bar{A} \in \bar{\mathcal{B}}$:

$$\theta^{\text{bad}} = \mathsf{E}\left[Y(\bar{A}^d) \mid \bar{A} \in \bar{\mathcal{B}} \right].$$

- Intuition: conditioning on $\bar{A}$ would be conditioning on mediators.

NYU Grossman School of Medicine

Longitudinal DAG

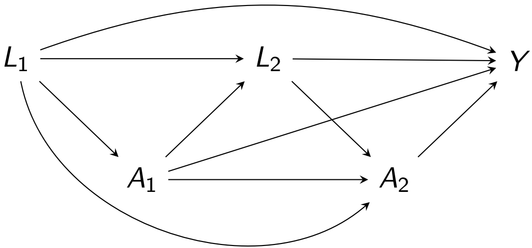


Figure: Example of the assumed Longitudinal DAG with 2 time points.

NYU Grossman School of Medicine

Example

- **Exposure:** Let A_t denote the particulate matter PM2.5 that an individual is exposed to at time t.
- **Outcome:** Let Y be a measure of respiratory function.
- **Intervention:** reduce PM2.5 by 10%.
- *Modified Treatment Policy:*

$$d(a_t, h_t) = 0.9 \times a_t.$$

- **Conditioning set**: exposure over the EPA standard at time 1:

$$\bar{\mathcal{B}} = \{A : A > 9\}.$$

- **Generalized ATT**:

$$\theta^* = \mathsf{E}[Y(\bar{A}^d) \mid \bar{A}(d) \in \bar{\mathcal{B}}].$$

Identification Assumptions

Notation

Underlines indicate the *future* of a variable, e.g. $\underline{U}_{A,t+1}$ are all $U_{A,t}$ from $t+1$ to τ.

- **Strong sequential randomization:** For all $t \in \{1, \ldots, \tau\}$,

$$U_{A,t} \perp\!\!\!\perp (\underline{U}_{L,t+1}, \underline{U}_{A,t+1}) | H_t.$$

- **Positivity:** For all $t \in \{1, \ldots, \tau\}$, if

$$(a_t, h_t) \in \text{Support}\{A_t, H_t \mid A_t \in \mathcal{B}_t\}$$

then

$$(d(a_t, h_t), h_t) \in \text{Support}\{A_t, H_t\}.$$

If there is positive probability of $A_t = a_t$, there has to also be positive probability of seeing the shifted treatment as well.

NYU Grossman School of Medicine

Identification

Theorem (abridged)

Let $m_{\tau+1} = Y$. Recursively define for $t = \tau, \ldots, 1$ the parameters

$$m_t : (a_t, h_t) \mapsto \mathsf{E}\left[m_{t+1}(A^d_{t+1}, H_{t+1}) \mid A_t = a_t, H_t = h_t, \underline{A}_{t+1} \in \underline{\mathcal{B}}_{t+1}\right].$$

The GATT parameter is identified as

$$\theta^* = \mathsf{E}\left[m_1(A^d_1, L_1) \mid \bar{A} \in \bar{\mathcal{B}}\right].$$

The identification result is conveniently in the form of sequential regressions,

NYU Grossman School of Medicine

Semi-parametric properties

- We analyze the *von-Mises expansion* of the GATT parameter: for any P, F in the non-parametric statistical model,

$$\theta(\mathsf{P}) - \theta(\mathsf{F}) = -\mathsf{E}_\mathsf{F}\{\mathsf{D}(Z;\mathsf{P})\} + \mathsf{R}(\mathsf{P},\mathsf{F}),$$

  where D is the *efficient influence function* of the parameter and is a second-order remainder term.

Efficient Influence Function

Theorem (Efficient influence function, abridged)

θ_1 is pathwise differentiable and its EIF is given by

$$\mathsf{D}(Z;P) = \sum_{t=0}^{\tau} \alpha_{t,\mathsf{P}}(A_t, H_t) \frac{1\{\underline{A}_{t+1} \in \underline{\mathcal{B}}_{t+1}\}}{G_{t,\mathsf{P}}(A_t, H_t)} \left\{ m_{t+1,\mathsf{P}}(A^d_{t+1}, H_{t+1}) - m_{t,\mathsf{P}}(A_t, H_t) \right\},$$

where $G_{t,\mathsf{P}}(A_t, H_t) = \mathsf{P}(A_{t+1} \in \mathcal{B}_{t+1} | A_t, H_t)$ and $\alpha_{t,\mathsf{P}}$ is a reweighting term.

NYU Grossman School of Medicine

What is α_t?

- The weighting term α_t is given by

$$\alpha_t(A_t, H_t) = \prod_{k=1}^{t} r_k(A_k, H_k),$$

with the density ratio at time t defined as

$$r_t(a_t, h_t) = \frac{g^d_{t,\mathcal{B}}(a_t, h_t)}{g_t(a_t, h_t)},$$

Loosely, $g_t(a_t, h_t)$ is the conditional probability of $A_t = a_t$ conditional on $H_t = h_t$ and $g^d_{t,\mathcal{B}}(a_t, h_t)$ is the conditional probability (density) of the treatment being shifted to a_t from a treatment in the conditioning set.

Second-order term

- The second-order remainder term of the von-Mises expansion is given by

$$\mathsf{R}(\mathsf{P},\mathsf{F}) =$$

$$-\sum_{t=1}^{\tau} \mathsf{E}_{\mathsf{P}}[\{\alpha_{t,\mathsf{P}}(A_t, H_t) - \alpha_{t,\mathsf{F}}(A_t, H_t)\}\{m_{t,\mathsf{P},\mathsf{F}}(A_t, H_t) - m_{t,\mathsf{F}}(A_t, H_t)\}]$$

$$-\sum_{t=1}^{\tau} \mathsf{E}_{\mathsf{P}}\left[\alpha_{t,\mathsf{F}}(A_t, H_t)\left\{1 - \frac{G_{t,\mathsf{P}}(A_t, H_t)}{G_{t,\mathsf{F}}(A_t, H_t)}\right\}\{m_{t,\mathsf{P},\mathsf{F}}(A_t, H_t) - m_{t,\mathsf{F}}(A_t, H_t)\right.$$

NYU Grossman School of Medicine

How can we estimate α_t?

- The cumulative probability (density) ratios α_t have a complex form, especially for continuous treatments, involving conditional probabilities (densities) that can be difficult to estimate.
- Estimation is especially challenging for long longitudinal structures.
- We instead estimate α_t by interpreting them as *Riesz Representers*, which we can estimate using a custom loss function using techniques developed by e.g. Chernozhukov PMLR 2022.

Riesz loss function

- Empirical loss function for Generalized ATT:

$$\hat{\alpha}_t = \underset{\tilde{\alpha} \in \mathcal{A}}{\text{argmin}} \, \mathbb{E}_n \left\{ \tilde{\alpha}(A_t, H_t)^2 - \hat{\alpha}_{t-1}(A_{t-1}, H_{t-1}) \frac{1\{\underline{A}_t \in \underline{\mathcal{B}}_t\}}{\hat{G}_{t-1}(A_{t-1}, H_{t-1})} b_t(A_t, H_t; \tilde{\alpha}) \right\},$$

- We use ensembles of loss minimization algorithms to solve the above minimization problem.
- A Super Learner based approach is available in our `R` package `SuperRiesz`: `github.com/herbps10/SuperRiesz`.

NYU Grossman School of Medicine

Estimation

- Now that we have a robust way of estimating α_t (a key ingredient to the EIF), we construct an estimator using TMLE (see preprint for details)
- Estimator available as part of the `lmtp` package: `github.com/nt-williams/lmtp/tree/riesz`

TMLE Robustness

Theorem (abridged)

Assume that, for each $j \in \{1, \ldots, J\}$,

$$\sum_{t=1}^{\tau} \|\hat{\alpha}_{t,j} - \alpha_t\| \|\tilde{m}_{t,j} - m_t\| = o_P(n^{-1/2}).$$

and

$$\sum_{t=1}^{\tau} \left\|\hat{G}_{t,j} - G_t\right\| \|\tilde{m}_{t,j} - m_t\| = o_P(n^{-1/2}).$$

Assume there exists some $c < \infty$ *such that* $\mathrm{P}(\alpha_t < c) = 1$ *and* $\mathrm{P}(\hat{\alpha}_t(A_t, H_t) < c) = 1$. *Then*

$$\sqrt{n}(\hat{\theta}_{tmle} - \theta) \rightsquigarrow N(0, \sigma^2),$$

where $\sigma^2 = \mathrm{Var}_{\mathrm{P}_0}(\mathrm{D}(Z; \mathrm{P}_0))$.

Simulation results

		95% Coverage		MAE $\times$ 100		$\text{sd}(\hat{\alpha}_\tau)$	
N	τ	Riesz	Plug-in	Riesz	Plug-in	Riesz	Plug-in
1000	2	91.5%	93.5%	3.21	3.26	1.41	1.57
	4	94.5%	95.0%	4.10	4.94	2.25	3.12
	6	88.5%	95.0%	5.46	9.13	2.20	5.78
	8	92.5%	93.5%	5.40	16.35	2.49	10.32
	10	88.5%	87.0%	6.25	27.95	2.52	18.21
	12	93.5%	85.5%	6.00	40.62	2.87	28.99
	14	96.0%	58.5%	6.21	35.31	3.62	34.89

Table: Simulation results for sample size $N = 1000$ and increasing number of time points τ in the longitudinal data structure.

Takeaways

- We generalize the average treatment effect on the treated to longitudinal settings and continuous treatments.
- We demonstrate how empirical Riesz learning can be used to stabilize estimation for long longitudinal data structures.
- R packages:
  - `lmtp`: `github.com/nt-williams/lmtp/tree/riesz`
  - `SuperRiesz`: `github.com/herbps10/SuperRiesz`