Causal Inference for Random Objects

Daisuke Kurisu (UTokyo),
Yidong Zhou (UCDavis),
Taisuke Otsu (LSE),
Hans-Georg Müller (UCDavis)

Fall School Time Series, Random Fields and beyond
Sep. 23-27, 2024@Ulm University

Outline

- Goal
- Examples of geodesic metric spaces
- Geodesic average treatment effect (GATE)
- Doubly robust estimator for the GATE
- Main results
  - Consistency/Rate of convergence of DR estimator
- Real data analysis
  - U.S. electricity generation data
  - New York Yellow Taxi data

Kurisu, D., Zhou, Y., Otsu, T., and Müller, H.-G. (2024) Geodesic causal inference. arXiv:2406.19604.

Outline

Goal:

- Extend the framework of causal inference for Euclidean data to general geodesic metric spaces.
  - Introduce geodesic averate treatment effect (GATE)
- Propose a doubly robust (DR) estimator for the GATE
  - Investigate asymptotic properties of the DR estimator.
  - Apply the proposed method to several real-world datasets.

Geodesic metric space

$(\mathcal{M}, d)$: a uniquely geodesic metric space.
$\forall \alpha, \beta \in \mathcal{M}$, the unique geodesic connecting α and β is a curve

$$\gamma_{\alpha,\beta} : [0,1] \mapsto \mathcal{M}$$

such that $d(\gamma_{\alpha,\beta}(s), \gamma_{\alpha,\beta}(t)) = d(\alpha,\beta)|t-s|$ for $s, t \in [0,1]$.

Extension of geodesics

- The space of interest is sometimes a subset of $(\mathcal{M}, d)$, often closed and convex.
- Assume that the geodesic $\gamma_{\alpha,\beta}$ extends to the boundary point ζ. For $\rho > 1$, the scalar multiplication is defined as

$$\rho \odot \gamma_{\alpha,\beta} = \{\gamma_{\alpha,\zeta}(t) : t \in [0, h(\rho)]\},$$
$$h(\rho) = -\left(1 - \frac{d(\alpha,\beta)}{d(\alpha,\zeta)}\right)^{\rho} + 1.$$

- Note that $\gamma_{\alpha,\zeta}(0) = \alpha$, $\gamma_{\alpha,\zeta}(h(1)) = \beta$, $\gamma_{\alpha,\zeta}(h(\infty)) = \zeta$.
- We write $\gamma_{\alpha,\zeta}(h(\rho))$ as $\gamma_{\alpha,\beta}(\rho)$.

Geodesic metric space

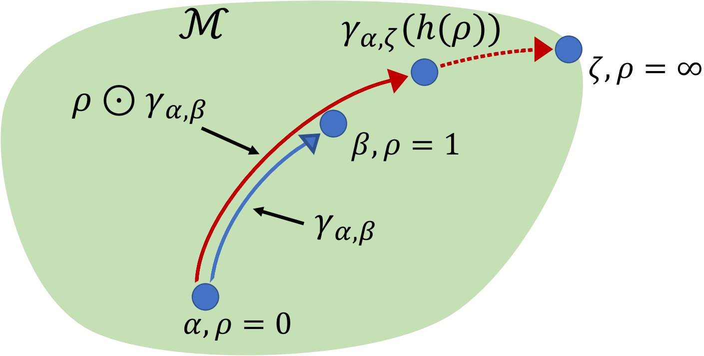


Figure: Illustration of $\rho \odot \gamma_{\alpha,\beta}$ when $\rho > 1$.

Geodesic metric space (Examples of metric spaces)

1. **Finite dimensional sphere**

   Finite dimensional case:

   - directional data
   - spherical simplex with geodesic distance $(\mathcal{S}_+^{d-1}, d_g)$
     $\rightarrow$ space for compositional data (Application 1).

     $$\Delta^{d-1} = \left\{ \boldsymbol{y} \in \mathbb{R}^d : y_j \geq 0, j = 1, \ldots, d, \text{ and } \sum_{j=1}^{d} y_j = 1 \right\}.$$

     Consider a map $\Delta^{d-1} \rightarrow \mathcal{S}_+^{d-1} = \{\boldsymbol{z} \in \mathcal{S}^{d-1} : z_j \geq 0, j = 1, \ldots, d\}$ s.t.

     $$(y_1, \ldots, y_d)' \mapsto (\sqrt{y_1}, \ldots, \sqrt{y}_d)'.$$

     For $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{S}_+^{d-1}$, $d_g(\boldsymbol{y}_1, \boldsymbol{y}_2) = \arccos(\boldsymbol{y}_1' \boldsymbol{y}_2)$.

Geodesic metric space (Examples of metric spaces)

2. **Space of graph Laplacians with Frobenius metric** $(\mathcal{L}_m, d_F)$ (Application 2)

   $G = (V, E)$: an undirected weighted network.
   $V = \{v_1, \ldots, v_m\}$: a set of nodes.
   $E = \{w_{ij}, w_{ij} \geq 0, i, j = 1, \ldots, m\}$: a set of edge weights.

   $$w_{ij} = 0 \Leftrightarrow v_i \text{ and } v_j \text{ are unconnected.}$$

3. **Space of covariance/correlation matrices with Frobenius metric** $(\mathcal{S}_m, d_F)$. (Application 3)
   $\mathcal{S}_m$: symmetric and positive semidefinite matrices.

4. **Space of univariate distributions with L^2-Wasserstein metric** $(\mathcal{W}_2(I), d_{\mathcal{W}})$. For univariate distributions μ and ν, the Wasserstein metric is defined as

   $$d_{\mathcal{W}}(\mu, \nu) = \sqrt{\int_0^1 (Q_\mu(s) - Q_\nu(s))^2 ds},$$

   where Q_μ and Q_ν are quantile functions of μ and ν.

Geodesic average treatment effect (GATE)

For each unit $i = 1, \ldots, n$, we observe $(Y_i, T_i, X_i) \in \mathcal{M} \times \{0,1\} \times \mathbb{R}^p$.

- T_i: the indicator of a teatment. $T_i = 1$ if treated and $T_i = 0$ otherwise.
- Y_i: the outcome

$$Y_i = \begin{cases} Y_i(0) & \text{if } T_i = 0 \\ Y_i(1) & \text{if } T_i = 1, \end{cases}$$

  where $Y_i(0), Y_i(1) \in \mathcal{M}$ are potential outcomes
- X_i: Euclidean covariates.

Geodesic average treatment effect

The geodesic average treatment effect (GATE) of T on Y is defined as

$$\gamma_{\mathrm{E}_\oplus[Y(0)],\mathrm{E}_\oplus[Y(1)]}.$$

$\mathrm{E}_\oplus[A]$ denotes the Fréchet mean of the random object $A \in \mathcal{M}$, that is,

$$\mathrm{E}_\oplus[A] = \arg\min_{\nu \in \mathcal{M}} \mathrm{E}[d^2(\nu, A)].$$

Doubly robust estimator for the GATE

Assumption 3.1

Let $p(x) = \mathrm{P}(T_i = 1|X_i = x)$ be the propensity score.

(i) $(\mathcal{M}, d)$ is a uniquely extendable geodesic metric space.

(ii) $\{Y_i, T_i, X_i\}_i^n$ are i.i.d. samples from a super-population of $(Y, T, X) \in \mathcal{M} \times \{0,1\} \times \mathcal{X}$, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$.

(iii) There exists a positive constant $\eta_0 \in (0, 1/2)$ such that $\eta_0 \leq p(x) \leq 1 - \eta_0$ for each $x \in \mathcal{X}$.

(iv) T_i and $\{Y_i(0), Y_i(1)\}$ are conditionally independent given X_i.

Assumption 3.2

For $t \in \{0,1\}$, let $\mathcal{P}_t : \mathcal{M} \to \mathcal{M}$ be a random perturbation map and m_t be a function such that $m_t : \mathcal{X} \to \mathcal{M}$ and

(i) $Y(t) = \mathcal{P}_t(m_t(X))$,

(ii) $\mathrm{E}_\oplus[\mathcal{P}_t(m_t(X))|X] = m_t(X)$,

(iii) $\mathrm{E}_\oplus[\mathcal{P}_t(m_t(X))] = \mathrm{E}_\oplus[m_t(X)]$.

Doubly robust estimator for the GATE

Definition (DR estimator)

DR estimator for the GATE is given as $\gamma_{\hat{\Theta}_0^{(\mathrm{DR})}, \hat{\Theta}_1^{(\mathrm{DR})}}$ where

$$\hat{\Theta}_t^{(\mathrm{DR})} := \arg\min_{\nu \in \mathcal{M}} Q_{n,t}(\nu; \hat{\mu}_t, \hat{\varphi}),$$

$$Q_{n,t}(\nu; \mu, \varphi) = \frac{1}{n} \sum_{i=1}^{n} d^2 \left(\nu, \gamma_{\mu(X_i), Y_i} \left(\frac{tT_i}{e(X_i; \varphi)} + \frac{(1-t)(1-T_i)}{1 - e(X_i; \varphi)} \right) \right).$$

$e(x; \varphi)$: a parametric model of the propensity score $p(x)$.
$\hat{\varphi}$: an estimator of a (true) parameter φ_*
$\hat{\mu}_t$: an estimator of the outcome regression function m_t.

Doubly robust estimator for the GATE

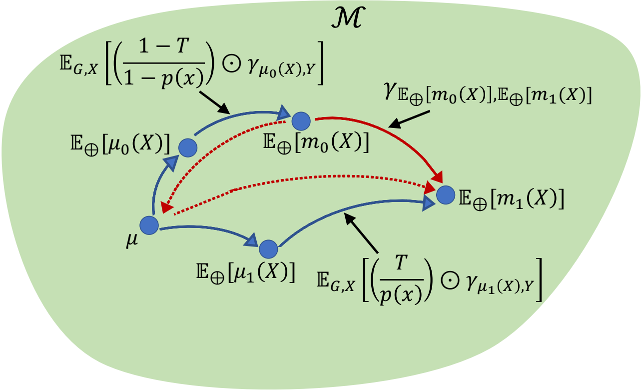


Figure: Illustration of DR representation of the GATE when $e(x) = p(x)$.

Doubly robust estimator for the GATE

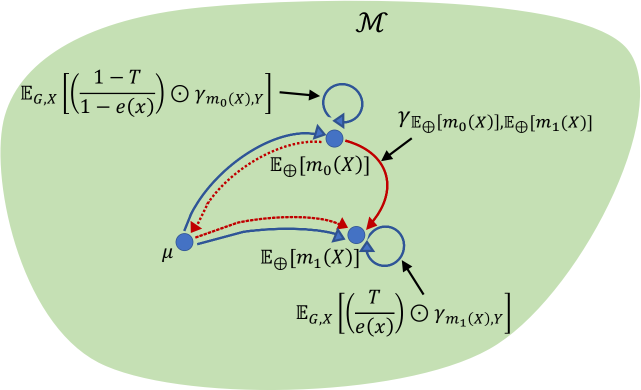


Figure: Illustration of DR representation of the GATE when $\mu_t(x) = m_t(x)$.

Doubly robust estimator for the GATE

In our real data analysis, we use

- logistic regression for $p(x)$ and
- global Fréchet regression for m_t (cf. Petersen and Müller ('19, AoS)):

$$\hat{\mu}_t(x) := \mathop{\arg\min}_{\nu \in \mathcal{M}} \frac{1}{N_t} \sum_{i \in I_t} \{1 + (X_i - \bar{X})' \hat{\Sigma}^{-1} (x - \bar{X})\} d^2(\nu, Y_i).$$

  $I_t = \{1 \leq i \leq n : T_i = t\}$.
  N_t: the sample size of I_t.
  $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$.
  $\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$.

- One can also use the local Fréchet regression for m_t (cf. Chen and Müller ('22, AoS)).

Main resutls

Assumption 4.1

Let $\Phi \subset \mathbb{R}^p$ be a compact set and let $\mathcal{M}_e = \{e(x;\varphi) : x \in \mathcal{X}, \varphi \in \Phi\}$ be a class of parametric models for propensity score $p(x)$. Additionally, let $\hat{\mu}_t(\cdot)$, $t \in \{0,1\}$ be estimators for the outcome regression functions $m_t(\cdot)$, $t \in \{0,1\}$.

(i) For $\varphi_1, \varphi_2 \in \Phi$, assume that $|e(x;\varphi_1) - e(x;\varphi_2)| \leq C_e \|\varphi_1 - \varphi_2\|$ for some positive constant $C_e > 0$, and for all $x \in \mathcal{X}$ and $\varphi \in \Phi$, $\eta_0 \leq e(x;\varphi) \leq 1 - \eta_0$.

(ii) There exist $\varphi_* \in \Phi$ and its estimator $\hat{\varphi}$ such that $\|\hat{\varphi} - \varphi_*\| = O_p(\varrho_n)$ with $\varrho_n \to 0$ as $n \to \infty$.

(iii) There exist functions $\mu_t(\cdot)$, $t \in \{0,1\}$ such that $\sup_{x \in \mathcal{X}} d(\hat{\mu}_t(x), \mu_t(x)) = O_p(r_n)$, $t \in \{0,1\}$ with $r_n \to 0$ as $n \to \infty$.

Assumption 4.2

For any $\alpha_1, \alpha_2 \in (\mathcal{M}, d)$, it holds

$$\sup_{\beta \in \mathcal{M}, \kappa \in [1/(1-\eta_0), 1/\eta_0]} d\left(\gamma_{\alpha_1,\beta}(\kappa), \gamma_{\alpha_2,\beta}(\kappa)\right) \leq C_0 d(\alpha_1, \alpha_2)$$

for some positive constant C_0 depending only on η_0.

Main resutls

Let $\Theta_t^{(\mathrm{DR})} := \underset{\nu \in \mathcal{M}}{\arg\min}\, Q_t(\nu; \mu_t, \varphi_*)$, $t \in \{0, 1\}$ where

$$Q_t(\nu; \mu, \varphi) = \mathrm{E}\left[d^2\left(\nu, \gamma_{\mu(X),Y}\left(\frac{tT}{e(X;\varphi)} + \frac{(1-t)(1-T)}{1-e(X;\varphi)}\right)\right)\right].$$

Assumption 4.3

Assume that for $t \in \{0, 1\}$,

(i) the objects $\Theta_t^{(\mathrm{DR})}$ and $\hat{\Theta}_t^{(\mathrm{DR})}$ exist and are unique, and for any $\varepsilon > 0$,

$$\inf_{d(\nu, \Theta_t^{(\mathrm{DR})}) > \varepsilon} Q_t(\nu; \mu_t, \varphi_*) > Q_t(\Theta_t^{(\mathrm{DR})}; \mu_t, \varphi_*),$$

(ii) $\Theta_t^{(\mathrm{DR})} = \mathrm{E}_\oplus[Y(t)]$.

Theorem 4.1 (Consistensy of DR estimator)

Suppose that Assumptions 3.1, 3.2, 4.1, 4.2 and 4.3 hold. Then

$$d(\hat{\Theta}_t^{(\mathrm{DR})}, \mathrm{E}_\oplus[Y(t)]) = o_p(1), t \in \{0, 1\}.$$

Main resutls

Let (Ω, d_Ω) be a metric space. For $\omega \in \Omega$, let $B_\delta(\omega)$ be the ball of radius δ centered at ω and $N(\varepsilon, B_\delta(\omega), d_\Omega)$ be its covering number using balls of size ε.

Assumption 4.4

For $t \in \{0, 1\}$,

(i) As $\delta \to 0$,

$$J_t(\delta) := \int_0^1 \sqrt{1 + \log N(\delta\varepsilon, B_\delta(\Theta_t^{(\mathrm{DR})}), d)}d\varepsilon = O(1),$$

$$J_{\mu_t}(\delta) := \int_0^1 \sqrt{1 + \log N(\delta\varepsilon, B_{\delta_1'}(\mu_t), d_\infty)}d\varepsilon = O(\delta^{-\varpi})$$

for some $\delta_1' > 0$ and $\varpi \in (0, 1)$, where for $\nu, \mu : \mathcal{X} \to \mathcal{M}$, $d_\infty(\nu, \mu) := \sup_{x \in \mathcal{X}} d(\nu(x), \mu(x))$.

(ii) there exist constants $\eta > 0$, $\eta_1 > 0$, $C > 0$, $C' > 0$, and $\beta > 1$ such that

$$\inf_{\substack{d_\infty(\mu,\mu_t) \le \eta_1 \\ \|\varphi - \varphi_*\| \le \eta_1}} \inf_{d(\nu, \Theta_t^{(\mathrm{DR})}) < \eta} \Big\{ Q_t(\nu; \mu, \varphi) - Q_t(\Theta_t^{(\mathrm{DR})}; \mu, \varphi) - C d(\nu, \Theta_t^{(\mathrm{DR})})^\beta + C' \eta_1^{\frac{\beta}{2(\beta-1)}} d(\nu, \Theta_t^{(\mathrm{DR})})^{\frac{\beta}{2}} \Big\} \ge 0.$$

Main resutls

Theorem 4.2 (Convergence rates of DR estimator)

Suppose that Assumptions 3.1, 3.2, 4.1, 4.2, 4.3, and 4.4 hold. Then for any $\beta' \in (0,1)$, we have

$$d(\hat{\Theta}_t^{(\mathrm{DR})}, \mathrm{E}_\oplus[Y(t)]) = O_p\left(n^{-\frac{1}{2(\beta-1+\varpi)}} + (\varrho_n + r_n)^{\frac{\beta'}{(\beta-1)}}\right), \ t \in \{0,1\}.$$

- Typically, $\beta = 2$, $\varrho_n = n^{-1/2}$, $r_n = n^{-\alpha_1}$ with any $\alpha_1 > 1/2$, $\varpi, \beta' \in (0,1)$.

$$d(\hat{\Theta}_t^{(\mathrm{DR})}, \mathrm{E}_\oplus[Y(t)]) = O_p(n^{-\frac{1}{2(1+\varpi)}} + n^{-\alpha_1\beta'}), \ t \in \{0,1\}.$$

- Network, Covariance matrix, Compositional data, Distribution.

Real data analysis ($\mathcal{S}_+^2, d_g$)

U.S. electricity generation data

Description of the dataset

- Outcome : the composition of energy sources across 50 U.S. states in 2020

$$Y_i = (\sqrt{y}_{1,i}, \sqrt{y}_{2,i}, \sqrt{y}_{3,i})' \in \mathcal{S}_+^2.$$

  - $y_{1,i}$: Natural Gas
  - $y_{2,i}$: Other Fossils
    (coal, petroleum, and other gases)
  - $y_{3,i}$: Renewables and Nuclear
    (hydroelectric conventional, solar thermal and photovoltaic, geothermal, wind, wood and wood-derived fuels, other biomass, and nuclear)
- Treatment : production of coal in each state in 2020. ($T_i = 1$ if the state produced coal, $T_i = 0$ o.w.)
- Covariates : GDP per capita (the millions of chained 2012 dollars), the proportion of electricity generated from coal and petroleum in each state in 2010.
- Sample size : $n = 50$ ($n_0 = 21$, $n_1 = 29$).

Real data analysis $(\mathcal{S}^2_+, d_g)$

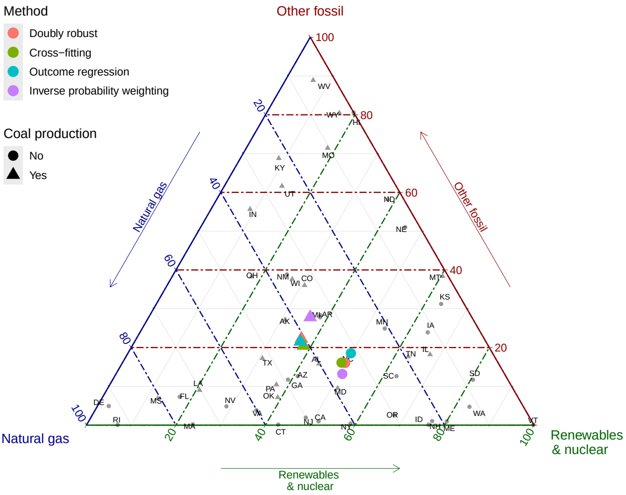


Figure: Mean potential outcomes for coal production using different methods.

- $d_g(\hat{\Theta}_0^{(\mathrm{DR})}, \hat{\Theta}_1^{(\mathrm{DR})}) = 0.133$, 95% adaptive HulC : $(0.112, 0.269)$.

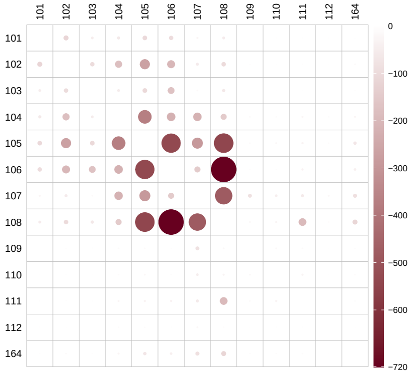
Real data analysis ($\mathcal{L}_{13}, d_F$)

New York Taxi system after COVID-19 outbreak

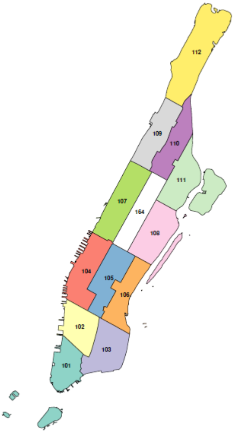
Description of the dataset

- Outcome : daily undirected network $G_i = (V_i, E_i)$
  - Nodes corresponding to the 13 regions in Manhattan.
  - Edge weights representing the number of people who traveled between the regions.
  - Period: April 12, 2020 $\sim$ Sep. 30, 2020 (172 days).
- Treatment : number of COVID-19 new cases in Manhattan area ($T_i = 1$ if > 60 and $T_i = 0$ if ≤ 60)
- Covariates : weekend indicator, temperature.
- Sample size : $n = 172$ ($n_0 = 79$, $n_1 = 93$).

Real data analysis ($\mathcal{L}_{13}, d_F$)



(A) Doubly robust



(B) 13 regions in Manhattan

Figure: Left: Average treatment effects (differences between adjacency matrices) represented as heatmaps using different methods.

Real data analysis ($\mathcal{L}_{13}, d_F$)

- Regions with the largest differences: (105, 106, 108)
  - 105: Penn Station, Times Square, The Museum of Modern Art.
  - 106: Grand Central Station, The United Nations HQs.
  - 108: residential area.
- $d_F(\hat{\Theta}_0^{(\mathrm{DR})}, \hat{\Theta}_1^{(\mathrm{DR})}) = 5216$,
- 95% adaptive HulC : $(2362, 10979)$.

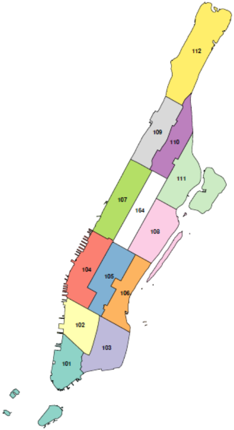


Figure: 13 regions in Manhattan

Conclusion

In this project, we

- introduced the geodesic average treatment effect (GATE) for the causal analysis of random objects;
- proposed four estimators for estimating the GATE
  - doubly robust
  - (cross-fitting)
  - (outcome regression)
  - (inverse probability weighting)
- established consistency and convergence rates of the estimators;
- applied the proposed methods to three datasets:
  - U.S. electricity generation data
  - New York Yellow Taxi data
  - (Alzheimer's disease data).

Kurisu, D., Zhou, Y., Otsu, T., and Müller, H.-G. (2024) Geodesic causal inference. arXiv:2406.19604.

References

- Chen, Y. and Müller, H.-G. (2022). Uniform convergence of local Fréchet regression with applications to locating extrema and time warping for metric space valued trajectries. *Annals of Statistics* **50**, 1573-1592.
- Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics* **47**, 691-719.