Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework

Chengchun Shi et al. JASA(2023)

October 25, 2023

Outline

1. Problem Formulation
2. Testing Procedure
   - Q-function
   - Test Statistics
   - Bootstrap and Online Updating
3. Real Data

Content

Potential Outcome Framework for MDP

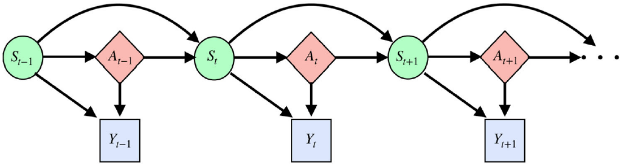

Figure 1. Causal diagram for MDP under settings where treatments depend on current states only. (S_t, A_t, Y_t) represents the state-treatment-outcome triplet. Solid lines represent causal relationships.

Challenges:

- Establishing a causal relationship between treatments and outcomes over time by taking the carryover effect into consideration;
- The testing hypothesis needs to be sequentially evaluated online as the data are being collected;
- Treatments are desired to be allocated in a manner to maximize the cumulative outcomes.

Notations

- treatment history vector up to time t:
  $\bar{a}_t = (a_0, a_1, \ldots, a_t)^\top \in \{0,1\}^{t+1}$
- potential outcome up to time t:

$$W_t^*(\bar{a}_t) = \{S_0, Y_0^*(a_0), S_1^*(a_0), \ldots, S_t^*(\bar{a}_{t-1}), Y_t^*(\bar{a}_t)\}$$

- **deterministic** policy π: a time-homogeneous function that maps the space of state variables to the set of available actions. The agent will assign actions according to π at each time.

Policy

- the goodness of a policy π is measured by its (state) value function:

$$V(\pi; s) = \sum_{t \geq 0} \gamma^t \mathbb{E} \{ Y_t^*(\pi) \mid S_0 = s \},$$

where $0 < \gamma < 1$ is a discount factor that reflects the trade-off between immediate and future outcomes.

- the Q-function:

$$Q(\pi; a, s) = \sum_{t \geq 0} \gamma^t \mathbb{E} \{ Y_t^*(\pi(a)) \mid S_0 = s \},$$

where $\pi(a)$ denotes a time-varying policy where the initial action equals a and all other actions are assigned according to π.

A/B Testing

The goal of A/B testing is to compare the two treatments.
Toward that end, we focus on two non-dynamic policies and use their value functions (denoted by $V(1;\cdot)$ and $V(0;\cdot)$) to measure their **long-term** treatment effects.

- CATE (conditional on the initial state $S_0 = s$):

$$\mathrm{CATE}(s) = V(1;s) - V(0;s).$$

- ATE:

$$\tau_0 = \int_s \{V(1;s) - V(0;s)\}\mathbb{G}(s)$$

  given a reference distribution function $\mathbb{G}$ that has a bounded density function on $\mathbb{S}$.

- Goal: testing

$$H_0 : \tau_0 = \mathrm{ATE} \leq 0 \quad \text{versus} \quad H_1 : \tau_0 = \mathrm{ATE} > 0$$

Motivating Toy Example

Table 1. Powers of t-test, DML-based test and the proposed test under Examples 1 and 2, with $T = 500$, $\delta = 0.1$. $\{A_t\}_t$ follow iid Bernoulli distribution with success probability 0.5.

Example 1			Example 2		
t-test 0.76	DML-based test 1	Our test 0.98	t-test 0.04	DML-based test 0.06	Our test 0.73

- *Example 1.* $S_t = 0.5\varepsilon_t$, $Y_t = S_t + \delta A_t$ for any $t \geq 1$ and $S_0 = 0.5\varepsilon_0$
- *Example 2.* $S_t = 0.5S_{t-1} + \delta A_t + 0.5\varepsilon_t$, $Y_t = S_t$ for any $t \geq 1$ and $S_0 = 0.5\varepsilon_0$

In both examples, the random error $\{\varepsilon_t\}_{t\geq 0}$ follows independent standard normal distributions and the parameter δ describes the degree of treatment effects.

Remark: treatments have delayed effects on the outcomes.

Content

Procedure Overview

1. Estimating Q-function based on temporal difference learning;
2. Construct test statistics based on $\hat{\tau}_t$ with plug-in estimates;
3. Generate bootstrap samples that mimic the distribution of the test statistics and integrate the α-spreading approach to sequentially implement the test.

Assumptions

- (CA) Consistency assumption:
  $S_{t+1} = S_{t+1}^{*}\left(\bar{A}_t\right)$ and $Y_t = Y_t^{*}\left(\bar{A}_t\right)$ for all $t \geq 0$.
- (SRA) Sequential randomization assumption:
  $A_t \perp W^{*} | S_t, \{S_j, A_j, Y_j\}_{0 \leq j < t}$.
- (MA) Markov assumption: there exists a Markov transition kernel $\mathcal{P}$ such that for any $t \geq 0, \bar{a}_t \in \{0,1\}^{t+1}$ and $\mathcal{S} \subseteq \mathbb{R}^d$, we have $\Pr\left\{S_{t+1}^{*}\left(\bar{a}_t\right) \in \mathcal{S} | W_t^{*}\left(\bar{a}_t\right)\right\} = \mathcal{P}\left(\mathcal{S}; a_t, S_t^{*}\left(\bar{a}_{t-1}\right)\right)$.
- (CMIA) Conditional mean independence assumption: there exists a function r such that for any $t \geq 0, \bar{a}_t \in \{0,1\}^{t+1}$, we have $\mathbb{E}\left\{Y_t^{*}\left(\bar{a}_t\right) | S_t^{*}\left(\bar{a}_{t-1}\right), W_{t-1}^{*}\left(\bar{a}_{t-1}\right)\right\} = r\left(a_t, S_t^{*}\left(\bar{a}_{t-1}\right)\right)$.

Estimating function

Lemma 1.

Under MA, CMIA, CA, and SRA, for any $t \geq 0$, $a' \in \{0,1\}$ and any function $\varphi : \mathbb{S} \times \{0,1\} \to \mathbb{R}$, we have

$$\mathbb{E}\left[\left\{Q\left(a'; A_t, S_t\right) - Y_t - \gamma Q\left(a'; a', S_{t+1}\right)\right\}\varphi\left(S_t, A_t\right)\right] = 0.$$

Q-function can be learned by solving this estimating equation.
As a result, $\tau_0 = \mathrm{ATE}$ is estimable since $V(a;s) = Q(a;a,s)$ and τ_0 is completely determined by the value function V.

Basis Approximation

Let $\mathcal{Q} = \left\{\Psi^\top(s)\beta_a : \beta_a \in \mathbb{R}^q\right\}$ be a large approximation space for $Q(a; a, s) = V(a, s)$, where $\Psi(\cdot)$ is a vector containing q basis functions. There exists some $\boldsymbol{\beta}^* = ({\beta_0^*}^\top, {\beta_1^*}^\top)^\top$ such that

$$\mathbb{E}\left[\left\{\Psi^\top(S_t)\,\beta_a^* - Y_t - \gamma\Psi^\top(S_{t+1})\,\beta_a^*\right\}\Psi(S_t)\,\mathbb{I}(A_t = a)\right] = 0, \forall a \in \{0, 1\}.$$

Basis Approximation

The above equations can be rewritten as $\mathbb{E}(\boldsymbol{\Sigma}_t \boldsymbol{\beta}^*) = \mathbb{E}(\boldsymbol{\eta}_t)$, where

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \begin{array}{c} \Psi(S_t)\mathbb{I}(A_t=0) \\ \{\Psi(S_t) - \gamma\Psi(S_{t+1})\}^\top \end{array} & \\ & \begin{array}{c} \Psi(S_t)\mathbb{I}(A_t=1) \\ \{\Psi(S_t) - \gamma\Psi(S_{t+1})\}^\top \end{array} \end{bmatrix}$$

is a block diagonal matrix, and

$$\boldsymbol{\eta}_t = \left\{ \Psi(S_t)^\top \mathbb{I}(A_t=0) Y_t, \Psi(S_t)^\top \mathbb{I}(A_t=1) Y_t \right\}^\top.$$

Let $\widehat{\boldsymbol{\Sigma}}(t) = t^{-1}\sum_{j<t} \boldsymbol{\Sigma}_j$ and $\widehat{\boldsymbol{\eta}}(t) = t^{-1}\sum_{j<t} \boldsymbol{\eta}_j$, it follows to have

$$\widehat{\boldsymbol{\beta}}(t) = \left\{ \widehat{\beta}_0^\top(t), \widehat{\beta}_1^\top(t) \right\}^\top = \widehat{\boldsymbol{\Sigma}}^{-1}(t)\widehat{\boldsymbol{\eta}}(t).$$

$\hat{\tau}(t)$: plug-in estimate

Let

$$\boldsymbol{U} = \left\{ -\int_{s\in\mathbb{S}} \Psi(s)^\top \mathbb{G}(ds), \int_{s\in\mathbb{S}} \Psi(s)^\top \mathbb{G}(ds) \right\}^\top,$$

It follows that

$$\widehat{\tau}(t) = \boldsymbol{U}\widehat{\boldsymbol{\beta}}(t).$$

We can prove that $\sqrt{t}\left\{\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}^*\right\}$ is multivariate normal and this implies that $\sqrt{t}\left\{\widehat{\tau}(t) - \tau_0\right\}$ is asymptotically normal.
This yields our test statistics $\sqrt{t}\widehat{\tau}(t)/\widehat{\sigma}(t)$, at time t:
for a given significance level α, we reject H_0 when $\sqrt{t}\widehat{\tau}(t)/\widehat{\sigma}(t) > z_\alpha$.
(**Theorem 1.**)

Limiting Distribution

Let $\{Z_1, \ldots, Z_K\}$ denote the sequence of our test statistics, where $Z_k = \sqrt{T_k}\widehat{\tau}(T_k)/\widehat{\sigma}(T_k)$.
The variance can be consistently estimated by

$$\widehat{\sigma}^2(t) = \boldsymbol{U}^\top \widehat{\boldsymbol{\Sigma}}^{-1}(t)\widehat{\boldsymbol{\Omega}}(t)\left\{\widehat{\boldsymbol{\Sigma}}^{-1}(t)\right\}^\top \boldsymbol{U},$$

and that

$$\widehat{\boldsymbol{\Omega}}(t) = \frac{1}{t}\sum_{j=0}^{t-1}\left\{\begin{array}{l}\Psi(S_j)(1-A_j)\widehat{\varepsilon}_{j,0}\\ \Psi(S_j)A_j\widehat{\varepsilon}_{j,1}\end{array}\right\}\left\{\begin{array}{l}\Psi(S_j)(1-A_j)\widehat{\varepsilon}_{j,0}\\ \Psi(S_j)A_j\widehat{\varepsilon}_{j,1}\end{array}\right\}^\top.$$

Remark:
$\widehat{\varepsilon}_{j,a}$ is the **temporal difference error** $Y_j + \gamma\Psi^\top(S_{j+1})\widehat{\beta}_a - \Psi^\top(S_j)\widehat{\beta}_a$ whose conditional expectation given $(A_j = a, S_j)$ is zero asymptotically.

Limiting Distribution

Theorem 1 (Limiting Distribution).

Assume C1-C3, MA, CMIA, CA, and SRA hold. Assume all immediate rewards are uniformly bounded variables, the density function of S_0 is uniformly bounded on $\mathbb{S}$ and q satisfies $q = o(\sqrt{T}/\log T)$.
Then under D1, D2 or D3, we have

- $\{Z_k\}_{1\le k\le K}$ are jointly asymptotically normal;
- their asymptotic means are nonpositive under H_0;
- their covariance matrix can be consistently estimated by some $\widehat{\boldsymbol{\Xi}}$, whose (k_1, k_2)-th element $\widehat{\boldsymbol{\Xi}}_{k_1,k_2}$ equals

$$\sqrt{\frac{T_{k_1}}{T_{k_2}}}\frac{U^\top \widehat{\boldsymbol{\Sigma}}^{-1}(T_{k_1})\widehat{\boldsymbol{\Omega}}(T_{k_1})\left\{\widehat{\boldsymbol{\Sigma}}^{-1}(T_{k_2})\right\}^\top \boldsymbol{U}}{\widehat{\sigma}(T_{k_1})\widehat{\sigma}(T_{k_2})}.$$

α-spending: control joint error rate

Suppose that the interim analyses are conducted at time points $T_1 < \ldots < T_K = T$.
To sequentially monitor the test, we need to specify the stopping boundary $\{b_k\}_{1\le k\le K}$ such that the experiment is terminated and H_0 is rejected when $Z_k > b_k$ for **some** k.
We require b_k's to satisfy

$$\Pr\left(\cup_{j=1}^{k}\{Z_j > b_j\}\right) = \alpha(T_k) + o(1), \quad \forall 1 \le k \le K. \tag{5}$$

and therefore

$$\Pr\left\{Z_k > b_k \mid \max_{1\le j<k}(Z_j - b_j) \le 0\right\} = \frac{\alpha(T_k) - \alpha(T_{k-1})}{1-\alpha(T_{k-1})} + o(1) \tag{7}$$

at each stage.

α-spending functions

The α spending function approach requires to specify a monotonically increasing function $\alpha(\cdot)$ that satisfies $\alpha(0) = 0$ and $\alpha(T) = \alpha$.
Some popular choices of the α spending function include

- $\alpha_1(t) = 2 - 2\Phi\left\{\Phi^{-1}(1 - \alpha/2)\sqrt{T/t}\right\}$,
- $\alpha_2(t) = \alpha(t/T)^{\theta}$ for $\theta > 0$.

Bootstrap: finding stopping boundaries

- The numerical integration of $\text{Cov}(Z_{k_1}, Z_{k_2})$ is not applicable.
- Bootstrap: Generate $\widehat{\beta}^{MB}$ according to the sandwich formula and recursively calculate the threshold.
- The wild bootstrap (Wu 1986) requires $O(BT_k)$ up to the k-th interim stage and can be time consuming when $\{T_k - T_{k-1}\}$ are large.
- Proposed Bootstrap: the random noise ζ_t is generated upon the arrival of each observation (T_k times). This is unnecessary as we aim to approximate the distribution of $\widehat{\beta}(\cdot)$ only at **finitely many** time points $T_1, \ldots, T_K$.

Bootstrap: sampling from covariance matrix

Key observation from Theorem 1:

$$\widehat{\Xi}_{k_1,k_2} = \frac{U^\top \widehat{\boldsymbol{\Sigma}}^{-1}\left(T_{k_1}\right)}{\sqrt{T_{k_1} T_{k_2}}\widehat{\sigma}\left(T_{k_1}\right)\widehat{\sigma}\left(T_{k_2}\right)} \left[\sum_{j=1}^{k_1} \left\{T_j \widehat{\boldsymbol{\Omega}}\left(T_j\right) - T_{j-1}\widehat{\boldsymbol{\Omega}}\left(T_{j-1}\right)\right\}\right] \left\{\widehat{\boldsymbol{\Sigma}}^{-1}\left(T_{k_2}\right)\right\}^{-1} U.$$

- Wild Bootstrap (from Covariance Matrix across **time points**):

$$\widehat{\beta}^{\mathrm{MB}}(t) = \widehat{\boldsymbol{\Sigma}}^{-1}(t) \left[\frac{1}{t}\sum_{j<t} \left\{ \begin{array}{c} \Psi\left(S_j\right)\left(1-A_j\right)\widehat{\varepsilon}_{j,0} \\ \Psi\left(S_j\right) A_j \widehat{\varepsilon}_{j,1} \end{array} \right\} \zeta_j \right].$$

- Proposed Bootstrap (from Covariance Matrix across **stages**):

$$\widehat{Z}_k^* = \frac{U^\top \widehat{\boldsymbol{\Sigma}}^{-1}\left(T_k\right)}{\sqrt{T_k}\widehat{\sigma}\left(T_k\right)} \sum_{j=1}^{k} \left\{T_j \widehat{\boldsymbol{\Omega}}\left(T_j\right) - T_{j-1}\widehat{\boldsymbol{\Omega}}\left(T_{j-1}\right)\right\}^{1/2} e_j.$$

Summary: Algorithm

Algorithm 1 The testing procedure

Input: number of basis functions q, number of bootstrap samples B, an α spending function $\alpha(\cdot)$.

Initialize: $T_0 = 0$, $\mathcal{I} = \{1, 2, \ldots, B\}$. Set $\widehat{\mathbf{\Omega}}$, $\widehat{\mathbf{\Omega}}^*$, $\widehat{\mathbf{\Sigma}}_0$, $\widehat{\mathbf{\Sigma}}_1$ to zero matrices, and $\widehat{\eta}$, $\widehat{S}_1, \ldots, \widehat{S}_B$ to zero vectors.

Compute U according to (3), using either Monte Carlo methods or numerical integration, where 0_q denotes a zero vector of length q.

For $k = 1$ to K:

Step 1. Online update of ATE.

For $t = T_{k-1}$ to $T_k - 1$:

$\widehat{\mathbf{\Sigma}}_a = (1 - t^{-1})\widehat{\mathbf{\Sigma}}_a + t^{-1}\Psi(S_t)\mathbb{I}(A_t = a)\{\Psi(S_t) - \gamma\Psi(S_{t+1})\mathbb{I}(A_{t+1} = a)\}^\top$, $a = 0, 1$;

$\widehat{\eta}_a = (1 - t^{-1})\widehat{\eta}_a + t^{-1}\Psi(S_t)\mathbb{I}(A_t = a)Y_t$.

Set $\widehat{\beta}_a = \widehat{\mathbf{\Sigma}}_a^{-1}\widehat{\eta}_a$ for $a \in \{0, 1\}$ and $\widehat{\tau} = U^\top\widehat{\beta}$.

Step 2. Online update of the variance estimator.

Initialize $\widehat{\mathbf{\Omega}}^*$ to a zero matrix.

For $t = T_{k-1}$ to $T_k - 1$:

$\widehat{\varepsilon}_{t,a} = Y_t + \gamma\Psi^\top(S_{t+1})\widehat{\beta}_a - \Psi^\top(S_t)\widehat{\beta}_a$ for $a = 0, 1$;

$\widehat{\mathbf{\Omega}}^* = \widehat{\mathbf{\Omega}}^* + \{\Psi(S_t)^\top(1 - A_t)\widehat{\varepsilon}_{t,0}, \Psi(S_t)^\top A_t\widehat{\varepsilon}_{t,1}\}^\top\{\Psi(S_t)^\top(1 - A_t)\widehat{\varepsilon}_{t,0}, \Psi(S_t)^\top A_t\widehat{\varepsilon}_{t,1}\}$.

Set $\widehat{\mathbf{\Sigma}}$ to a block diagonal matrix by aligning $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{\Sigma}}_1$ along the diagonal of $\widehat{\mathbf{\Sigma}}$;

Set $\widehat{\mathbf{\Omega}} = T_k^{-1}(T_{k-1}\widehat{\mathbf{\Omega}} + \widehat{\mathbf{\Omega}}^*)$ and the variance estimator $\widehat{\sigma}^2 = U^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Omega}}\{\widehat{\mathbf{\Sigma}}^{-1}\}^\top U$.

Step 3. Bootstrap test statistic.

For $b = 1$ to B:

Generate $e_k^{(b)} \sim N(0, I_{4q})$;

$\widehat{S}_b = \widehat{S}_b + \widehat{\mathbf{\Omega}}^{*1/2}e_k^{(b)}$;

$\widehat{Z}_b^* = T_k^{-1/2}\widehat{\sigma}^{-1}U^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{S}_b$;

Set z to be the upper $\{\alpha(t) - |\mathcal{I}^c|/B\}/(1 - |\mathcal{I}^c|/B)$-th percentile of $\{\widehat{Z}_b^*\}_{b\in\mathcal{I}}$.

Update $\mathcal{I}$ as $\mathcal{I} \leftarrow \{b \in \mathcal{I} : \widehat{Z}_b^* \le z\}$;

Step 4. Reject or not?

Reject the null if $\sqrt{T_k}\widehat{\sigma}^{-1}\widehat{\tau} > z$.

Type-I error & Power

Theorem 2 (Type-I error).

Suppose that the conditions of Theorem 1 hold and $\alpha(\cdot)$ is continuous. Then the proposed thresholds satisfy

$$\Pr\left(\bigcup_{j=1}^{k}\left\{Z_j > \widehat{b}_j\right\}\right) \leq \alpha\left(T_k\right) + o(1),$$

for all $1 \leq k \leq K$ under H_0. The equality holds when $\tau_0 = 0$.

Theorem 3 (Power).

Suppose that the conditions of Theorem 2 hold. Assume $\tau_0 \gg T^{-1/2}$, then

$$\Pr\left(\bigcup_{j=1}^{k}\left\{Z_j > \widehat{b}_j\right\}\right) \to 1.$$

Assume $\tau_0 = T^{-1/2}h$ for some $h > 0$. Then

$$\lim_{T\to\infty}\left[\Pr\left(\cup_{j=1}^{k}\left\{Z_j > \widehat{b}_j\right\}\right) - \alpha\left(T_k\right)\right] > 0.$$

Content

Background

The proposed test is applied to a large-scale ride-sharing platform.

- new strategy: dispatch a given order to a nearby driver that has not yet finished their previous ride request but almost.
- standard control: assign orders to drivers that have completed their ride requests.
- The new strategy is expected to reduce the chance that the customer will cancel an order in regions with only a few available drivers. It is expected to meet more call orders and increase drivers' income on average.

Data

- A/A experiment: conducted from November 12 to November 25.
- A/B experiment: conducted from December 3 to December 16.

Both experiments last for two weeks. Thirty minutes is defined as a one-time unit. We set $K = 8$ and $T_k = 48(k+6)$ for $k = 1, \ldots, 8$. That is, the first interim analysis is performed at the end of the first week, followed by seven more at the end of each day during the second week.

- response Y_t: the overall drivers' income in each time unit
- state S_t: 1)the number of requests (demand) and 2)drivers' online time (supply) during each 30-minute time interval, 3)the supply and demand equilibrium metric.
- $\Psi(\cdot)$: fourth-degree polynomial basis
- $\gamma = 0.6$, $B = 1000$
- α spending function: $\alpha_1(t)$

Result

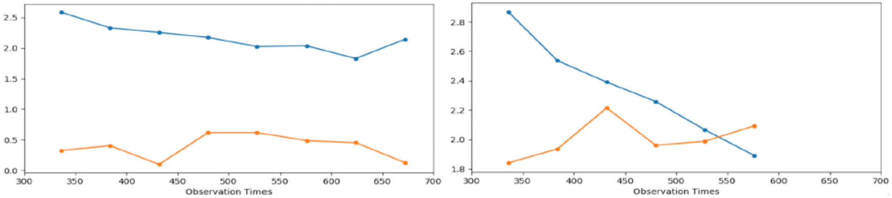

Figure 4. Our test statistic (the orange line) and the rejection boundary (the blue line) in the A/A (left plot) and A/B (right plot) experiments.

Remark:

- The p-value of applying the two-sample t-test to the data collected from the A/B experiment is 0.18. This result is consistent with the previous findings: the t-test cannot detect such carryover effects, leading to a low power.

Reference

- Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song (2023) Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework, Journal of the American Statistical Association, 118:543, 2059-2071, DOI: 10.1080/01621459.2022.2027776