Statistical Learning in Mediation Analysis

_____

Chapter 6:
Direct and indirect effects for stochastic interventions

David Benkeser
Emory University

Iván Díaz
Weill Cornell Medicine

Marco Carone
University of Washington

MODULE 13

Summer Institute in Statistics for
Clinical and Epidemiological Research

July 2021

Contents of this chapter

Total effects:

1. What are stochastic interventions
2. Modified treatment policies and incremental propensity score interventions
3. When and how is the causal effect of a stochastic intervention identified
4. R packages and examples

Mediation effects:

1. Defining mediation effects for stochastic interventions
2. When and how are these effects identified
3. Estimation based on the G-computation and IPW formulas
4. Doubly-robust estimation
5. R packages and examples

- Consider a simple data structure $O = (W, A, Y)$
- Recall the counterfactual definitions

$$Y(1) \quad \text{and} \quad Y(0)$$

- Obtained by considering hypothetical worlds where A **is set to** $A = 1$ **and** $A = 0$ **with probability one**.

We will alternatively consider interventions where A **is set to some random variable** A_δ **with probability one**.

Modified treatment policies (MTP)

- Let A denote a measure of air quality given by particulate matter $PM_{2.5}$. One could be interested in an intervention that would reduce $PM_{2.5}$ by 10%:

$$d(A, W) = (1 - \delta)A,$$

where $\delta = 0.1$.

- Let A denote self-reported physical activity as measured by weekly minutes. One could be interested in an intervention that increases physical activity for people for whom it is feasible:

$$d(A, W) = \begin{cases} A + \delta & \text{if } A + \delta < u(W) \\ A & \text{if } A + \delta \geqslant u(W) \end{cases},$$

where $u(W)$ is the upper bound of physical activity for someone with covariates W (age, health status, etc.).

Incremental propensity score interventions (IPSI):

A_δ is a random draw from a Bernoulli distribution with conditional probability equal to

$$g_{A,\delta}(1 \mid w) = \frac{\delta g_A(1 \mid w)}{\delta g_A(1 \mid w) + 1 - g_A(1 \mid w)}$$

where $g_A(1 \mid w) := P(A = 1 \mid W = w)$ and δ **is a user given value** (Kennedy, 2019).

Note that:

$$\frac{\mathsf{odds}(A_\delta = 1 \mid W)}{\mathsf{odds}(A = 1 \mid W)} = \delta.$$

IPSIs can be seen as MTPs: $A_\delta = I(\varepsilon < g_{A,\delta}(1 \mid W))$

Define the counterfactual variable $Y(A_\delta)$ as the variable that **would have been observed in a hypothetical world where** $A = A_\delta$.

- We can contrast the expectation of $Y(A_\delta)$ with the expectation of Y to obtain a causal effect:
$$E[Y(A_\delta) - Y] \quad \text{or} \quad E[Y(A_\delta)]/E[Y]$$

- We know how to estimate $E(Y)$ very well (using the empirical mean)

- In what follows **we focus on identifying and estimating** $E[Y(A_\delta)]$

We need the usual two assumptions:

- **Positivity:** if $g_A(a \mid w) = 0$ then $g_{A,\delta}(a \mid w) = 0$.
- **Randomization:** $A \perp Y(a) \mid W$ for all a

We have that
$$E[Y(A_\delta)] = E[\bar{Q}_Y(A_\delta, W)],$$

where, as before, $\bar{Q}_Y(a, w) := E(Y \mid A = a, W = w)$.

Positivity assumption for IPSIs:

- Recall the definition

$$g_{A,\delta}(1 \mid w) = \frac{\delta g_A(1 \mid w)}{\delta g_A(1 \mid w) + 1 - g_A(1 \mid w)}$$

- $g_A(a \mid w) = 0$ implies $g_{A,\delta}(a \mid w) = 0$ for $a \in \{0,1\}$.

Positivity assumption for MTPs:

- Let's look at one of the examples:

$$d(A, W) = \begin{cases} A + \delta & \text{if } A + \delta < u(W) \\ A & \text{if } A + \delta \geqslant u(W) \end{cases},$$

- Assume $P(A < u(W) \mid W) = 1$
- Then $g_A(a \mid w) = 0$ implies $g_{a,\delta}(a \mid w) = 0$

- To define **meaningful effects for non-binary exposures:**
  - In some applications (e.g., physical activity) **it may make little sense to work with counterfactuals that set $A = a$**
  - Even if defining Y_a is sensible conceptually, estimating $E(Y_a)$ non-parametrically is hard for continuous exposures

- To define and estimate causal effects in the presence of **violations of the positivity assumption:**
  - IPSIs satisfy positivity by design
  - MTPs can also be arranged to satisfy positivity by definition (but require some knowledge about the support of A)

The idea of inverse probability weighting also applies to estimation of these parameters.

Each individual needs to be reweighted to account for the number of individuals it would represent in a hypothetical population where the intervention has been performed

Specifically, define the density ratio

$$r(A, W) = \frac{g_{A,\delta}(A \mid W)}{g_A(A \mid W)}$$

The **reweighting identification formula** is given by

$$E[Y(A_\delta)] = E\left[r(A, W) \times Y\right]$$

The reweighted identification formula suggests a simple estimation strategy.

First, we construct an estimate of the density ration $r(A, W)$.

Then, we can compute an **IPTW estimator** of $E[Y(A_\delta)]$ as

$$\psi_{n,IPTW} := \frac{1}{n} \sum_{i=1}^{n} r_n(A_i, W_i) Y_i \ .$$

Simple estimation procedures

There are at least two possible strategies to estimate the density ratio

$$r(A, W) = \frac{g_{A,\delta}(A \mid W)}{g_A(A \mid W)}$$

1. Construct estimates of the densities $g_A(A, W)$ and $g_{A,\delta}(A, W)$, and plug them into the above definition of $r(A, W)$.
   - This may not be easy to do with data-adaptive estimators: the machine and statistical learning literatures have only few methods for conditional density estimation

2. **Recast the density ratio problem as a classification problem** (Details in Díaz, 2020)

Below we simulate data as follows

$$Y \mid A = a, W = w \;\sim\; \text{Bernoulli}(\text{expit}(1 - a + w))$$
$$A \mid W = w \;\sim\; \text{Normal}(1 - 0.5 \times w, 1))$$
$$W \;\sim\; \text{Normal}(0, 1)$$

```
# set a seed for reproducibility
set.seed(202)
n <- 5000
W <- rnorm(n)
A <- rnorm(n,  1 - 0.5 * W, 1)
Y <- rbinom(n, 1, plogis(1 - A + W))
```

We will estimate the effect of the intervention

$$d(A, W) = \begin{cases} A + 0.5 & \text{if } A + 0.5 < 3 \\ A & \text{if } A + 0.5 \geqslant 3 \end{cases},$$

```
## Computing the true value
d  <- function(A) (A + 0.5) * (A + 0.5 < 3) + A * (A + 0.5 >= 3)
W0  <- rnorm(1e6)
A0  <- rnorm(1e6,  1 - 0.5 * W0, 1)
Y0 <- rbinom(1e6, 1, plogis(1 - A0 + W0))
mean(plogis(1 - d(A0) + W0))

## [1] 0.42273
```

```
# Create the augmented dataset
aug_data <- rbind(data.frame(A = A,    W = W, Delta = rep(1, n)),
                  data.frame(A = d(A), W = W, Delta = rep(0, n)))
aug_data <- aug_data[order(aug_data$W), ]
head(aug_data)

##              A         W Delta
## 3165 3.272460 -3.420273     1
## 8165 3.272460 -3.420273     0
## 1509 2.660246 -3.328865     1
## 6509 2.660246 -3.328865     0
## 4672 3.300567 -3.258121     1
## 9672 3.300567 -3.258121     0
```

```
## Solve the classification problem
library(earth)
fit <- earth(Delta ~ A * W, data = aug_data, glm = list(family = binomial))
## Compute the density ratio for the real data
pred_r <- predict(fit, newdata = data.frame(A, W), type = 'response')
r    <- (1 - pred_r)/pred_r
## Compute the estimate
mean(r * Y)

## [1] 0.4149335
```

The G-computation formula suggests another natural estimation strategy

$$E[Y(A_\delta)] = E[\bar{Q}_Y(A_\delta, W)],$$

1. **STEP 1:** Regress Y on A and $W \longrightarrow \bar{Q}_{Y,n}$
2. **STEP 2:** Predict under the intervention $\longrightarrow \bar{Q}_{Y,n}(A_\delta, W)$
3. **STEP 3:** Average the predictions across the sample

$$\psi_{n,gcomp} := \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_{Y,n}(A_{\delta,i}, W_i)$$

```
## Estimate the outcome regression with your favorite prediction method
fitY <- glm(Y ~ A + W, family = binomial)
## Compute the predictions under the intervention
pred_d <- predict(fitY, newdata = data.frame(A = d(A), W), type = 'response')
## Compute the estimate
gcomp <- mean(pred_d)
gcomp

## [1] 0.4192603
```

Improved estimation procedures (for MTPs)

The following assumes $d(A, W)$ is known and does not need to be estimated (precludes IPSIs).

As before, a hybrid between the G-computation and IPTW estimator can be constructed, and it enjoys improved properties

The **one-step estimator** (akin to AIPW) is given by

$$\psi_{n,os} := \psi_{n,gcomp} + \frac{1}{n} \sum_{i=1}^{n} r_n(A_i, W_i)[Y - \bar{Q}_{Y,n}(A_i, W_i)]$$

```
predY <- predict(fitY, type = 'response')
one_step <- gcomp + mean(r * (Y - predY))
one_step
## [1] 0.4186593
```

For MTPs, the one-step estimator enjoys the following improved properties:

- It hits the right target if **either** r **or** $\bar{Q}_Y$ is estimated well

- Valid confidence intervals (even when flexible regression is used) can be constructed if **both** r **and** $\bar{Q}_Y$ are estimated well

As before, a TMLE can also be computed (and it is implemented in the R packages given below). TMLE has the advantage that it is always in the parameter space

These estimators are implemented in several R packages:

- `lmtp` (on CRAN)
    - Implements cross-fitting for improved properties
    - Longitudinal data
- `txshift`
- `tmle3`

```
library(lmtp)
d_lmtp <- function(data, x) d(data[[x]])
tmle1  <- lmtp_tmle(data.frame(A, W, Y), "A", "Y", "W", shift = d_lmtp,
                    outcome_type = "binomial", folds = 5)
tmle1

LMTP Estimator: TMLE
   Trt. Policy: (d_lmtp)

Population intervention effect
      Estimate: 0.4199
    Std. error: 0.0073
       95% CI: (0.4055, 0.4342)
```
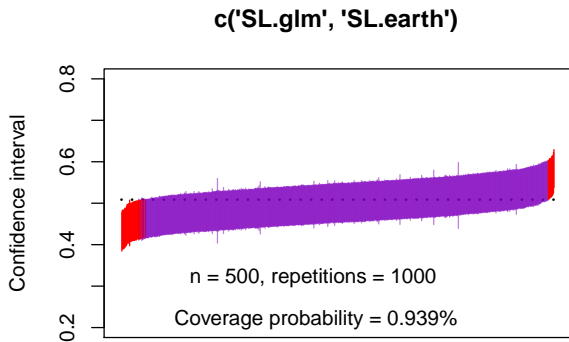
We simulate data as follows

$$
\begin{aligned}
Y \mid A = a, W = w &\sim \text{Bernoulli}(\text{expit}(1 - a + w^2)) \\
A \mid W = w &\sim \text{Normal}(1 - 0.5 \times \log(|w|)), 1)) \\
W &\sim \text{Normal}(0, 1)
\end{aligned}
$$

- Sample size $n = 500$
- Simulate $m = 1000$ datasets
- Run lmtp using a Super Learner. The library contains:
    - MARS and logistic regression, or
    - only logistic regression
- Look at the bias and coverage of confidence intervals across the $m = 1000$ datasets

c('SL.glm', 'SL.earth')

n = 500, repetitions = 1000

Coverage probability = 0.939%

SL.glm

n = 500, repetitions = 1000

Coverage probability = 0.767%

Below we simulate data as follows

$$Y \mid A = a, W = w \ \sim \ \text{Bernoulli}(\text{expit}(1 - 2 * a + w))$$
$$A \mid W = w \ \sim \ \text{Bernoulli}(\text{expit}(-1 + 2 * w))$$
$$W \ \sim \ \text{Normal}(0, 1)$$

```r
# set a seed for reproducibility
set.seed(202)
n <- 5000
W <- rnorm(n)
A <- rbinom(n, 1, plogis(- 1 + 2 * W))
Y <- rbinom(n, 1, plogis(1 - 2 * A + W))
```

Notice that in our example positivity is violated:
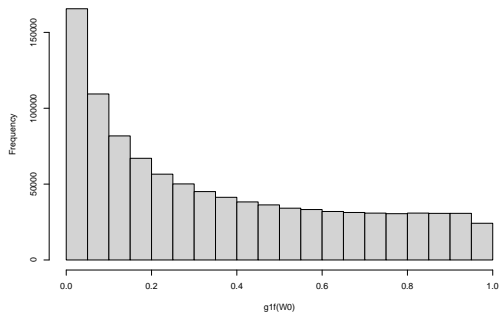
Figure 1: Propensity score distribution

This data generating mechanism is a perfect candidate for IPSIs

First, let us compute the true value of the parameter

```r
g1f <- function(W) plogis(- 1 + 2 * W)
delta <- 2
gdelta1f <- function(delta, W) {
    delta * g1f(W) / (delta * g1f(W) + 1 - g1f(W))
}
W0 <- rnorm(1e6)
A0 <- rbinom(1e6, 1, gdelta1f(delta, W0))
Y0 <- rbinom(1e6, 1, plogis(1 - 2 * A0 + W0))
mean(Y0)
```

```
## [1] 0.519703
```

To illustrate IPSIs, we use the npcausal package:

```
library(devtools)
install_github("ehkennedy/npcausal")
library(npcausal)
result <- ipsi(Y, A, W, W, time = rep(1, n),
               id = 1:n, 2, progress_bar = FALSE)
```

```
result$res
##      increment      est       se     ci.ll     ci.ul
## 95%          2 0.5070921 0.5497941 0.4919289 0.5222552
```

For IPSIs, the one-step estimator enjoys the following improved properties:

- It **requires that the propensity score is well estimated** to hit the target

- The above is a consequence of the fact that the intervention variable

$$A_\delta = I(\varepsilon < g_{A,\delta}(1 \mid W))$$

  depends on g_A, and thus it also needs to be estimated

- Valid confidence intervals (even when flexible regression is used) can be constructed if **both r and $\bar{Q}_Y$** are estimated well

References:

Kennedy, Edward H. "Nonparametric causal effects based on incremental propensity score interventions." Journal of the American Statistical Association 114.526 (2019): 645-656.

Young, Jessica G., Miguel A. Hernán, and James M. Robins. "Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data." Epidemiologic methods 3.1 (2014): 1-19.

Díaz, Iván, and Mark van Der Laan. "Population intervention causal effects based on stochastic interventions." Biometrics 68.2 (2012): 541-549.

Díaz, Iván, et al. "Non-parametric causal effects based on longitudinal modified treatment policies." arXiv preprint arXiv:2006.01366 (2020).
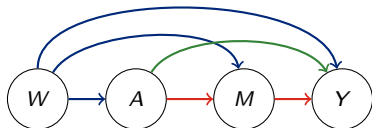
Assume we want to decompose the total effect

$$E[Y(A_\delta)] - E[Y]$$

into:

- A direct effect operating through $A \to Y$
- An in direct effect operating through $A \to M \to Y$

As before, we have the following definitions:

- $Y(a)$ is the counterfactual obtained by setting $A = a$

- $Y(a, m)$ is the counterfactual obtained by setting $A = a$ and $M = m$
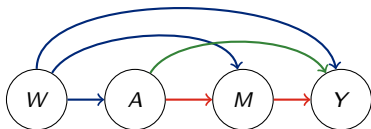
Notice that an intervention setting A equal to A_δ induces two counterfactual variables:

- $M(A_\delta)$

- $Y(A_\delta) = Y(A_\delta, M(A_\delta))$

$$\psi(\delta) = E[Y(A_\delta) - Y]$$
$$= \underbrace{E[Y(A_\delta, M(A_\delta)) - Y(A_\delta, M)]}_{\text{indirect effect}} + \underbrace{E[Y(A_\delta, M) - Y(A, M)]}_{\text{direct effect}}.$$

- We have discussed estimation of $E[Y(A_\delta)] = E[Y(A_\delta, M_\delta)]$

- As before, estimation of $E(Y) = E[Y(A, M)]$ is straightforward

- In what follows we focus on **identification and estimation of** $E[Y(A_\delta, M)]$

- In what follows we focus on identification and estimation for MTPs
- Identification and estimation methods for mediational effects under incremental propensity score interventions have also been developed (Díaz and Hejazi, 2020)

- The R package `medshift` provides an implementation of the relevant estimators for IPSIs

We need a modified version of the usual two assumptions:

- **Positivity:**
    - If $g_{A,\delta}(a \mid w) > 0$ then $g_A(a \mid w) > 0$
    - If $g_M(m \mid w) > 0$ then $g_M(m \mid a, w) > 0$
- **Randomization:**
    - $Y(a, m) \perp A \mid W$
    - $Y(a, m) \perp M \mid (A, W)$

Under these assumptions we have:

$$E[Y(A_\delta, M)] = E\left[\bar{Q}_Y(M, d(A, W), W)\right]$$

$$
\begin{aligned}
E[Y(A_\delta, M) \mid M = m, A = a, W = w] &= E[Y(d(a, w), m) \mid M = m, A = a, W = w] \\
&= E[Y(d(a, w), m) \mid M = m, A = d(a, w), W = w] \\
&= E[Y \mid M = m, A = d(a, w), W = w] \\
&= \bar{Q}_Y(m, d(a, w), w),
\end{aligned}
$$

where we have defined

$$
E(Y \mid M = m, A = a, W = w) := \bar{Q}_Y(m, a, w)
$$

Averaging with respect to the distribution of (M, A, W) yields

$$
E[Y(A_\delta, M)] = E\left[\bar{Q}_Y(M, d(A, W), W)\right]
$$

- Let $\tilde{W} = (W, M)$
- The identification formula becomes

$$E\left[\bar{Q}_Y(d(A, \tilde{W}), \tilde{W})\right],$$

where

$$\bar{Q}_Y(a, \tilde{w}) = E[Y \mid A = a, \tilde{W} = \tilde{w}]$$

- This is identical to the formula that we studied in the previous chapter
- Thus, for purposes of estimation, we can forget about the nature of M as a mediator and W as a confounder and **proceed using the methods discussed for estimating $E[\bar{Q}_Y(d(A, W), W)]$.**

The **indirect effect** is thus identified as

$$E\left[\bar{Q}_Y(d(A, W), W) - \bar{Q}_Y(M, d(A, W), W)\right]$$

The **direct effect** is identified as

$$E\left[\bar{Q}_Y(M, d(A, W), W) - Y\right]$$

Let us revisit the **randomization assumption:**

- $Y(a, m) \perp A \mid W$
- $Y(a, m) \perp M \mid (A, W)$

This assumption precludes intermediate confounders Z.

However, unlike the cross-world assumption necessary for the direct and indirect effects, **this assumption can be satisfied by design** if the study randomizes the mediator and treatment.

$$
\begin{aligned}
Y \mid A = a, W = w &\sim \text{Bernoulli}(\text{expit}(1 - a + w)) \\
M \mid A = a, W = w &\sim \text{Normal}(-1 - 2 \times a - w, 1)) \\
A \mid W = w &\sim \text{Normal}(1 - 0.5 \times w, 1)) \\
W &\sim \text{Normal}(0, 1)
\end{aligned}
$$

And consider the modified treatment policy

$$
d(A, W) = \begin{cases} A + 0.5 & \text{if } A + 0.5 < 3 \\ A & \text{if } A + 0.5 \geqslant 3 \end{cases},
$$

```
# set a seed for reproducibility
set.seed(202)
n <- 5000
W <- rnorm(n)
A <- rnorm(n,  1 - 0.5 * W, 1)
M <- rnorm(n,  - 1 - 2 * A - W, 1)
Y <- rbinom(n, 1, plogis(1 - A + W))
```

We can use all the estimators discussed in the previous chapter to estimate both

$$E\left[\bar{Q}_Y(d(A, W), W)\right] \text{ and } E\left[\bar{Q}_Y(M, d(A, W), W)\right]$$

We will next see an example of how to use the `lmtp` package to implement a TMLE for this

```r
library(lmtp)
d   <- function(A) (A + 0.5) * (A + 0.5 < 3) + A * (A + 0.5 >= 3)
d_lmtp <- function(data, x) d(data[[x]])


tmle1  <- lmtp_tmle(data.frame(A, W, Y), "A", "Y", "W", shift = d_lmtp,
                    outcome_type = "binomial", folds = 5)


tmle2  <- lmtp_tmle(data.frame(A, M, W, Y), "A", "Y", c("M", "W"),
                    shift = d_lmtp,
                    outcome_type = "binomial", folds = 5)
```
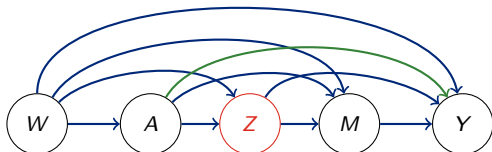
```
## Indirect effect estimate
ie <- tmle1$theta - tmle2$theta
se_ie <- sd(tmle1$eif - tmle2$eif) / sqrt(n)
c(ie, se_ie)

## [1] 0.009670311 0.006782272

## Direct effect estimate
de <- tmle2$theta - mean(Y)
se_de <- sd(tmle1$eif - Y) / sqrt(n)
c(de, se_de)

## [1] -0.090372218  0.002492682
```

Stochastic effect decomposition in the presence of intermediate confounders

- Let G denote a random draw from the distribution of M conditional on (A, W)

- Let G_δ denote a random draw from the distribution of $M(A_\delta)$ conditional on (A_δ, W)

Then we have the following decomposition of the total effect:

$$E[Y(A_\delta, G_\delta) - Y(A, G)] = \underbrace{E[Y(A_\delta, G_\delta) - Y(A_\delta, G)]}_{\text{indirect effect}} + \underbrace{E[Y(A_\delta, G) - Y(A, G)]}_{\text{direct effect}}.$$

Identification requires the usual two assumptions:

- **Positivity:**
  - If $g_{A,\delta}(a \mid w) > 0$ then $g_A(a \mid w) > 0$
  - If $g_M(m \mid w) > 0$ then $g_M(m \mid a, w) > 0$
- **Randomization:**
  - $Y(a, m) \perp A \mid W$
  - $Y(a, m) \perp M \mid (A, W, Z)$

Identificaiton formulas, estimators, and software, may be found in the research paper (Hejazi et al., 2020)

References:

Hejazi, Nima S., et al. "Nonparametric causal mediation analysis for stochastic interventional (in) direct effects." arXiv preprint arXiv:2009.06203 (2020).

Díaz, Iván, and Nima S. Hejazi. "Causal mediation analysis for stochastic interventions." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82.3 (2020): 661-683.