

LECTURE 13 - CAUSAL INFERENCE, INDEPENDENT COMPONENT ANALYSIS

Q: Assume a known causal model

- Can we estimate the **causal effect** of a treatment T on an outcome Y ?
(causal inference)
- Can we estimate the **mechanism** that generates observed data X ?
(ICA, representation learning?)

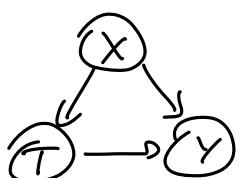
1. Causal Inference

Common scenario:

X : observed confounders / controls

T : treatment

Y : outcome



Def.: Average Treatment Effect (ATE)

For binary treatments $T \in \{0, 1\}$, we define

$$\text{ATE} = E[Y | d_1(T=1)] - E[Y | d_0(T=0)]$$

Note: Often one assumes a structural equation

$$Y := \alpha T + f(X, U_Y)$$

Then, computing ATE \Leftrightarrow estimating α

Def.: Heterogeneous Treatment Effects (HTE)

or Conditional Average Treatment Effect (CATE)
Local

$$\text{CATE}(x) = E[Y | X=x, \text{do}(T=1)] - E[Y | X=x, \text{do}(T=0)]$$

(Useful when $Y := f(T, X, U_Y)$ is non-linear)

Remarks:

- A related objective is to estimate the "value" of a policy π , i.e. $T = \pi(x)$

$$V(\pi) = E[Y | \text{do}(T := \pi(x))]$$

→ e.g. in "Contextual Bandits," we may want to find a policy π that maximizes $V(\pi)$ (policy optimization)

example: ad recommendations

T : which ad to show

Y : user clicked on the ad

π_i : how to choose the right ad
for a user with features X -

- Different framework: potential outcomes
two variables $Y(1)$ and $Y(0)$
for $T=1$ and $T=0$
We only observe the one for the applied treatment.

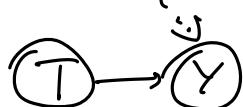
Here, we have $ATE = \bar{E}[Y(1) - Y(0)]$

- Causal Inference via randomization

Idea: set up an experiment with randomized treatment
e.g. $T \sim Ber(\frac{1}{2})$, $T \perp X$

e.g. \rightarrow Randomized Controlled Trial (RCT, in medicine)
 \rightarrow A/B testing (e.g. internet companies)

(X can be unobserved!)



We can then estimate

$$\begin{aligned}\bar{Y}_1 &= E[Y(1)] = E[Y | \text{do}(T=1)] \\ &= E[Y | T=1] \approx \frac{1}{m_1} \sum_{i:T_i=1} y_i =: \hat{Y}_1\end{aligned}$$

$$\text{and } \bar{Y}_0 = E[Y | T=0] \approx \frac{1}{m_0} \sum_{i:T_i=0} y_i =: \hat{Y}_0$$

$$\Rightarrow ATE \approx \hat{Y}_1 - \hat{Y}_0$$

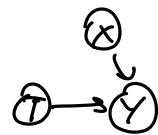
Note: → In this setup, we can make precise statistical inferences by computing confidence intervals for \hat{Y}_1 and \hat{Y}_0

- No risk of unobserved confounding if T is truly randomized
- In practice, there may still be some bias due to selection / sampling of the units (e.g. surveys, medical trials)

▪ Inverse Propensity Weighting

- Observed data from  with randomized treatment $T := f_T(x, v_T)$
- joint distribution $p(x, t, y) = p(x) p(t|x) p(y|t, x)$

• Target intervention : $do(T := t_0)$



→ distribution

$$P^{do(T=t_0)}(x, t, y) = p(x) \cdot \mathbb{1}\{t=t_0\} p(y|t, x)$$

Idea: Use importance sampling to estimate $E[Y|do(T=t_0)]$

$$P^{do(T=t_0)}(x_i, t_i, y_i) = p(x_i, t_i, y_i) \cdot \frac{\mathbb{1}\{t_i=t_0\}}{p(t_i|x_i)}$$

must be
 $\neq 0$ for $t_i=t_0$!

Observations (x_i, t_i, y_i) , $i=1, \dots, n$

$$E[Y|do(T=t_0)] = E\left[Y \cdot \frac{\mathbb{1}\{T=t_0\}}{p(T|X)}\right]$$

$$\approx \frac{1}{n} \sum_{i=1}^n y_i \cdot \frac{\mathbb{1}\{t_i=t_0\}}{p(t_i|x_i)} \quad (IPW)$$

Remarks:

- This yields an **unbiased** estimator:

$$\hat{E}_{t_i \sim p(\cdot|x_i)} \left[y_i(t_i) \cdot \frac{\mathbb{1}\{t_i=t_0\}}{p(t_i|x_i)} \right] = y_i(t_0) \cdot \frac{p(t_0|x_i)}{p(t_0|x_i)} = y_i(t_0)$$

Potential outcome notation.

- The **variance can be large** if $p(t_i|x_i)$ is small!!
- The "**propensities**" $p(t_i|x_i)$ are often known.
If not, we may use estimates $\hat{p}(t|x)$

■ Direct estimation and doubly robust methods

→ In some cases, we may estimate causal effects using simple regression:

$$\text{If } Y := \alpha T + \beta X + \varepsilon \text{ with } \mathbb{E}[\varepsilon | T, X] = 0$$

Then α can be estimated by regressing Y on (T, X) with OLS:

$$\min_{\hat{\alpha}, \hat{\beta}} \sum_i (y_i - \hat{\alpha} t_i - \hat{\beta} x_i)^2$$

But: The estimate $\hat{\alpha}$ will be biased if:

- T depends linearly/deterministically on X (\Rightarrow cannot identify α)
- the model is mis-specified
- some form of regularization is applied

ex: $T := m_0(X) + \varepsilon_T$ X high-dimensional
 $Y := \alpha T + g_0(X) + \varepsilon_Y$

- estimating $g_0(X)$ is difficult and benefits from regularization/ML \rightarrow bias
- "Double Machine Learning" to remove bias
 (Chernozhukov et al., 2018)

- Doubly Robust estimator :
 - First obtain $\hat{g}(t, x)$ by regression of Y on (T, x)
 - Then use the estimator : (Dudik et al., 2014)

$$(DR) \quad E[Y | do(T=t_0)] \approx \frac{1}{m} \sum_{i=1}^m \hat{g}(t_0, x_i) + \frac{1}{m} \sum_{i=1}^m (y_i - \hat{g}(t_0, x_i)) \frac{\mathbb{1}_{\{t_i=t_0\}}}{p(t_i | x_i)}$$

→ still unbiased for exact $p(t_i | x_i)$ (like IPW),
 but lower variance when \hat{g} is good estimate!

2. ICA and Representation Learning

Independent Components Analysis

Consider a causal model:

$$s = (s_1, \dots, s_m) \sim p(s) \quad (\text{components})$$

$$x := As \quad A \in \mathbb{R}^{m \times m} \quad (\text{mixing matrix})$$

Observe $x(t)$ with components $s(t)$, $t=1, \dots, T$.

Q: Can we recover both A and $s(t)$ from observations $x(t)$ alone? ("identifiability")

→ **PCA**: • captures only covariance $E[x x^T]$
 $\leadsto \frac{m^2}{2}$ parameters due to symmetry
but: need m^2 for A !!

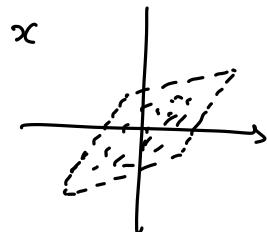
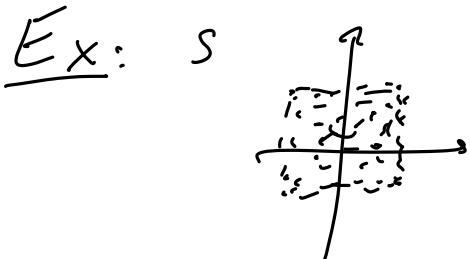
- limited to "gaussianity" of the data.

→ **ICA**: capture information beyond covariance
e.g. higher-order moments ("Kurtosis")
 \leadsto non-gaussianity

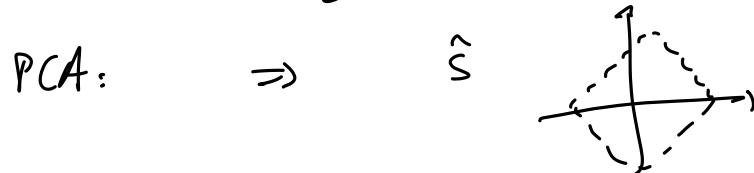
→ **Identifiability**: If

- $p(s) = p(s_1) \dots p(s_m)$ (independent components)
- $p(s_i)$ are Non-Gaussian

Then, A and $s(t)$ can be identified! (from $p(x)$)



Estimating A using:



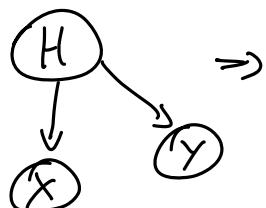
Remarks:

- Algorithms proceed by maximizing "non-Gaussianity"
 - Similar "non-gaussian" strategies can be used to identify Structural Equations in Causal models, and discover causal graphs
(e.g. $X \rightarrow Y$ vs $Y \rightarrow X$)
see Peters et.al , Chapter 5
-

- Insights for unsupervised representation learning

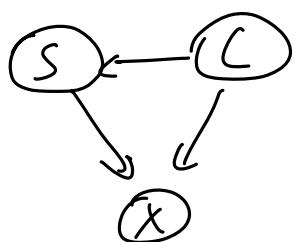
Find representation $\phi(x)$ such that learning on downstream task is easier from $\phi(x)$ than x .

Q: Can we leverage causal models and identifiability?



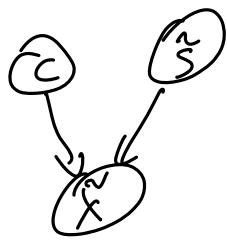
take $\phi(x) = H$

Examples:



C: content, useful representation
S: "style" e.g. texture, pose, scale

self-supervised learning: generate different styles from same image



von Kügelgen et al (2021): Contrastive learning
leads to identifiability of c.

- $x(t) = f(s(t))$
(non-linear ICA)

$s(t)$ can be identified from time information if $s(t)$ is non-stationary ("pretext" task: predict t from $x(t)$)
Time-contrastive learning, [Kyränen & Mariakka 2016]