Lecture 13 - Causal Inference, Independent Component Analysis

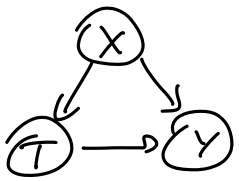
Q: Assume a known causal model

→ 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

$$ATE = \mathbb{E}[Y \mid do(T=1)] - \mathbb{E}[Y \mid do(T=0)]$$

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

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

(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. $T = \pi(x)$

$$V(\pi) = \mathbb{E}[Y \mid do(T := \pi(X))]$$

→ 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: 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.

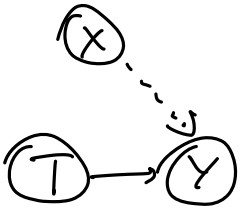
Here, we have $ATE = \mathbb{E}[Y(1) - Y(0)]$

Causal Inference via randomization

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

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

(X can be unobserved!)



We can then estimate

$$\bar{Y}_1 = \mathbb{E}[Y(1)] = \mathbb{E}[Y \mid do(T=1)]$$
$$= \mathbb{E}[Y \mid T=1] \approx \frac{1}{m_1} \sum_{i:T_i=1} y_i =: \hat{Y}_1$$

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

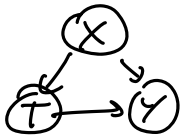
$$\Rightarrow \quad ATE \approx \hat{Y}_1 - \hat{Y}_0$$

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 

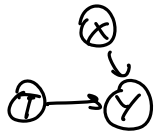
with randomized treatment

$$T := f_T(X, U_T)$$

→ joint distribution

$$p(x,t,y) = p(x)\, p(t \mid x)\, p(y \mid t, x)$$

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



→ distribution

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

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

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

← must be $\neq 0$ for $t=t_0$!

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

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

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

Remarks:

- This yields an unbiased estimator:

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

↑ 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:

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

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 is correlated with X or ε

→ 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 ⇝ bias

→ "Double Machine Learning" to remove bias (Chernozhukov et al, 2018)

- **Doubly Robust estimator**:
  - → First obtain $\hat{y}(t,x)$ by regression of Y on (T,X)
  - → then use the estimator: (Dudík et al., 2014)

(DR) $$\mathbb{E}[Y \mid do(T=t_0)] \approx \frac{1}{n}\sum_{i=1}^{n} \hat{y}(t_0, x_i) + \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}(t_0, x_i)) \frac{\mathbb{1}\{t_i = t_0\}}{p(t_i \mid x_i)}$$

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

---

2. ICA and Representation Learning

- **Independent Components Analysis**

Consider a causal model:

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

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

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 $\mathbb{E}[xx^T]$

⇝ $\frac{n^2}{2}$ parameters due to symmetry

but: need (n^2) for A !!

• limited to "Gaussianity" of the data.

→ **ICA**: capture information beyond covariance

e.g. higher-order moments ("Kurtosis")

⇝ non-gaussianity

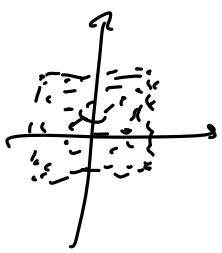
→ **Identifiability**: If

• $p(s) = p(s_1) \cdots p(s_n)$ (independent components)

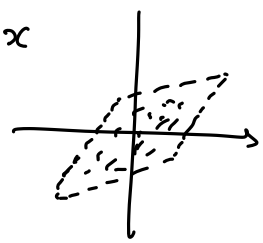
• $p(s_i)$ are Non-Gaussian

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

Ex: s

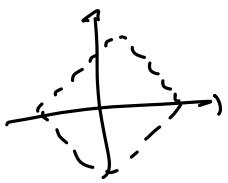


x

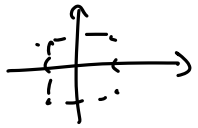


Estimating A using:

PCA: ⇒ $\hat{s}$



ICA: ⇒ $\hat{s}$



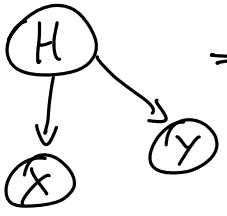
Remarks:

- 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 → Y vs Y → X)
  see Peters et.al, Chapter 4.

- Insights for unsupervised representation learning

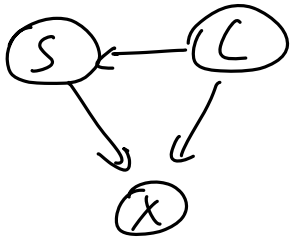
Find representation $\phi(X)$ such that learning on downstream task is easier from $\phi(X)$ than X.

Q: Can we leverage causal models and identifiability?



⇒ take $\phi(X) = H$

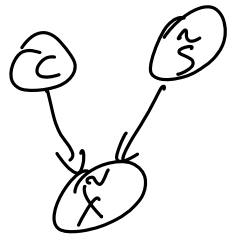
Examples: •



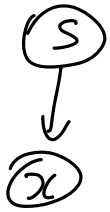
C: content, useful representation

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 $\underline{c}$.

- 

$x(t) = f(s(t))$

(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, (Hyvärinen & Morioka 2016)