Causal Inference and Invariance

Qingyuan Zhao and Charles Zheng

Stanford University

February 23, 2016

(Part 2/2)

From Last Week: Causal Graph

Causal relationships in a system represented by a graph. The graph tells you:

I. which variables are affected by an intervention.
II. what conditional independence relationships exist in the joint distribution (*d-separation*.)
III. which sets of predictors and responses will have "invariant" optimal predictive rules.

This talk is restricted to directed acyclic graph (DAG), i.e. no feedback!

From Last Week: Three *Causal* Questions

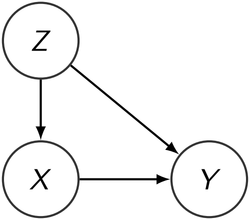
- Given a number of variables, which pairs are causally related?
  - Infer the *graph*.
- Given a number of variables and a fixed Y, which variables causally affect Y?
  - Infer the *invariance set*.
- Given a fixed X and a fixed Y, what is the causal effect of X on Y?
  - Infer the *causal effect*.

Why different languages? Convenience!

Section 1

Overview of Previous methods

Known Causal Structure



For example, suppose we want to estimate the causal effect of X on Y with known confounders Z.

- Graphical approach: the backdoor formula

$$\mathrm{P}(y|do(x)) = \sum_{z} \mathrm{P}(y|x,z)\mathrm{P}(z).$$

- Functional approach: outcome regression $Y \sim X + Z$.
- Potential outcome approach: estimate the propensity score.

Unknown Causal Structure

Conventional approach:

1. Estimate the Markov equivalence class of causal graphs via conditional independence relationships.
2. Infer or bound the identifiable causal effects.

More recent approach: impose additional functional/distributional assumptions to the structural equation model: for any variable Y,

$$Y = f(\text{parents}(Y); \epsilon_Y).$$

How should we think about the assumptions?

One thing for sure: They are no monsters!

How should we think about the assumptions?

- In statistics we make assumptions all the time: parametric, independence, function form, etc.
  - George Box: "All models are wrong but some are useful".
- To infer causation, we need to make different kinds of assumptions.
  - Problem statement: Can what we learned from this environment be generalized to another environment?
  - The ancient wisdom: "Correlation does not imply causation" (observational $\not\Rightarrow$ interventional).
  - Causal assumptions: causal graph, structural equation model, or invariant prediction.

What if we are willing to make both kinds of assumptions?

Section 2

Invariance

Assumed invariance

Focus: Given a number of variables and a fixed Y, which variables causally affect Y?
Data: i.i.d. samples of (X^e, Y^e) from different environments $e \in \mathcal{E}$.

Assumption (Invariant prediction)

There exists a vector of coefficients γ^* with support S^* such that for all $e \in \mathcal{E}$, X^e has an arbitrary distribution and

$$Y^e = \mu + X^e \gamma^* + \epsilon^e, \;\; \epsilon^e \sim F_\epsilon, \;\; \epsilon^e \perp\!\!\!\perp X^e_{S^*}.$$

Important:

- F_ϵ does not depend on e.
- ϵ is always independent of X.

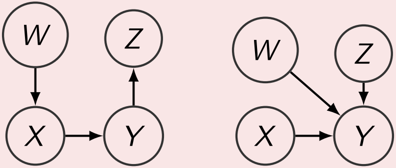
This is essentially a single structural equation with $\mathrm{parents}(Y) = S^*$.

Building block

Testing the null hypothesis that (γ, S) satisfies the assumption.

$$H_{0,\gamma,S}(\mathcal{E}): \ \gamma_k = 0 \text{ if } k \in S, \ and \ \exists F_\epsilon \text{ such that for all } e \in \mathcal{E},$$
$$Y^e = X^e \gamma + \epsilon^e, \ \epsilon^e \sim F_\epsilon, \ \epsilon^e \perp\!\!\!\perp X_S^e.$$
$$H_{0,S}(\mathcal{E}): \ \exists \gamma \text{ such that } H_{0,\gamma,S}(\mathcal{E}) \text{ is true.}$$

Difficulty



Statistically, we may end up accepting both $Y^e = X^e + \epsilon^e$ and $Y^e = X^e + 0.01W^e + 0.01Z^e + \epsilon^e$, for both causal structures.

Generic procedure

1. For each $S \subseteq \{1, \ldots, p\}$, test $H_{0,S}(\mathcal{E})$ at level α.
2. Set $\hat{S}(\mathcal{E}) = \bigcap_{S: H_{0,S}(\mathcal{E}) \text{ not rejected}} S$.
3. For the confidence sets, set $\hat{\Gamma}(\mathcal{E}) = \bigcup_{S \subseteq \{1,\ldots,p\}} \hat{\Gamma}_S(\mathcal{E})$, where

$$\hat{\Gamma}_S(\mathcal{E}) = \begin{cases} \emptyset & H_{0,S}(\mathcal{E}) \text{ is rejected at level } \alpha, \\ \hat{S} & \text{otherwise.} \end{cases}$$

$\hat{C}(S)$ is a $(1-\alpha)$-confidence set for γ obtained by pooling the data.

Theorem (Peters et al.)

$$\mathrm{P}(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha, \ \mathrm{P}(\gamma^* \in \hat{\Gamma}(\mathcal{E})) \geq 1 - 2\alpha.$$

The Statistical Challenge

Depending on the modeling assumption, this hypothesis can be:

$$
\begin{aligned}
H_{0,S,\mathrm{lin}}(\mathcal{E}):\ & \exists\gamma \text{ s.t. } \gamma_k = 0 \text{ if } k \in S, \textit{and} \\
& \exists F_\epsilon \text{ s.t. } Y^e = X^e\gamma + \epsilon^e,\ \epsilon^e \sim F_\epsilon,\ \epsilon^e \perp\!\!\!\perp X_S^e,\ \forall e \in \mathcal{E}. \\
H_{0,S,\text{lin-gauss}}(\mathcal{E}):\ & H_{0,S,\mathrm{lin}}(\mathcal{E}) \text{ and } F_\epsilon = \mathrm{N}(0,\sigma^2). \\
H_{0,S,\mathrm{nonlin}}(\mathcal{E}):\ & \exists g(X_S,\epsilon),\ F_\epsilon \text{ s.t. } Y^e = g(X_S^e,\epsilon^e),\ \epsilon^e \ldots. \\
H_{0,S,\mathrm{additive}}(\mathcal{E}):\ & H_{0,S,\mathrm{nonlin}}(\mathcal{E}) \text{ and } g(X_S,\epsilon) \text{ is additive.} \\
H_{0,S,\mathrm{hidden}}(\mathcal{E}):\ & \epsilon^e \sim F_\epsilon,\ \forall e \in \mathcal{E},\ \text{ but } F_\epsilon \text{ can have nonzero mean.}
\end{aligned}
$$

How to test $H_{0,S}(\mathcal{E})$?

Peters et al. give concrete proposals for $H_{0,S,\text{lin-gauss}}$ and $H_{0,S,\text{lin-gauss-hidden}}$. They are implemented in their `InvariantCausalPrediction` package.

Robustness of the invariance approach

In Meinshausen's talk: we don't make false discoveries, even under a misspecified model!
Truth (at least what we believe in):

Things can go wrong	ICP's behavior
Intervene on Y (or a missing cause)	$\bigcap \emptyset$
Non-linear, non-additive	$\bigcap \emptyset$
Not enough interventions	False positives
Small sample size	$\emptyset$
Left out a confounder	$\emptyset$
Left out an unconfounding predictor	okay

Splitting purely observational data

A big bonus: we can "create" an environment by conditioning on a variable U that we know precedes Y. This is valid because

$$Y|X_{S^*} \stackrel{d}{=} Y|X_{S^*}, U = u.$$

Note: this statement is true only in the region that both conditional distributions are well defined.

Creating environment by instrumental variable

$$\begin{tikzcd} & Z \arrow[d] \arrow[dr] & \\ W \arrow[r] & X \arrow[r] & Y \end{tikzcd}$$

If there is a hidden confounder Z, we can condition on the instrumental variable W.