Causal Inference and Invariance

Qingyuan Zhao and Charles Zheng

Stanford University

February 25, 2016

(Part 2/2)

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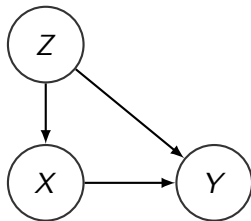
For example, suppose we want to estimate the causal effect of X on Y with known confounders Z.

- 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,

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

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.

Peters et al.: What if we are willing to make both kinds of assumptions?

One thing for sure: They are no monsters!

Section 2

Invariance

* indicates our comments.

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

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

Important:

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

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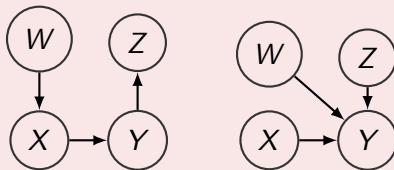
Building block

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

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

Difficulty*

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

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

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

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

A concrete proposal

$$\begin{pmatrix} Y^1 \\ Y^2 \end{pmatrix} = \begin{pmatrix} X_S^1 \\ X_S^2 \end{pmatrix} \gamma_S + \begin{pmatrix} \epsilon^1 \\ \epsilon^2 \end{pmatrix}$$

For each subset $S \subset \{1, \ldots, p\}$:

1. Estimate γ_S from the pooled data.

2. Compare distributions of the residuals $\hat{\epsilon}^1$ and $\hat{\epsilon}^2$. When Gaussian, compare their means (equal to 0 if no hidden variable) and variances.

3. If you accept that $\hat{\epsilon}^1$ and $\hat{\epsilon}^2$ have the same distribution at level α, S is accepted as an invariant set.

Finally, report $\hat{S} = \cap \{S : S \text{ accepted}\}$.

This algorithm can be easily extended to non-linear models.

Model rejection

What if *all subsets are rejected*? Then

$$\hat{S} = \bigcap_{\emptyset}$$

which is not well-defined.

In that case, the ICP() function gives the following "error" message:

```
> ICP(X, Y, ExpInd = EI)

*** 6% complete: tested 2 of 31 sets of variables
*** 13% complete: tested 4 of 31 sets of variables
*** 26% complete: tested 8 of 31 sets of variables
*** 52% complete: tested 16 of 31 sets of variables
 Invariant Linear Causal Regression at level 0.01 (in
n for the number of variables)
 Model has been rejected at the chosen level, that is
to invariance across the environments. This can be fo
 (a) non-linearities or
 (b) hidden variables or
 (c) interventions on the target variable.

 We will try to extend the functionality soon to allo
```

Robustness of the invariance approach*

From Meinshausen's talk: we don't make false discoveries, even under a misspecified model!

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

Identifiability

Peters et al. give sufficient conditions for identifiability of S^* in the linear Gaussian model.

Theorem (Peters et al.)

S^* or equivalently the parents of Y are identifiable if

1. At least one single intervention on each variable other than Y; or
2. There is only one intervention on the youngest parent of Y (say X_1), that is there is no directed path from X_1 to any other parent of Y. In this case S^* is identifiable with probability 1 (when parameters γ is drawn from some distribution).

No interventions?

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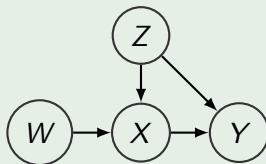
If there is a hidden confounder Z, we can condition on the instrumental variable W.

Back to the three causal questions*

Can ICP help to answer the other two questions?

Infer the graph

We can run ICP for every node with caution. Returns a partially identified graph.

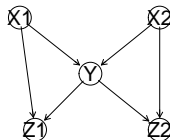## Infer the causal effect of X on Y

Two options:

- Treat X as the target variable: *propensity score*.
- Treat Y as the target variable: *outcome regression*.

Okay if $\hat{S}$ itself is invariant. Otherwise ICP may miss important causes, resulting in biased causal effect estimate.

Idea: maybe we can just use (many) "minimal" S.

Section 3

Invariance

All variables normally distributed, linear structural equation model.
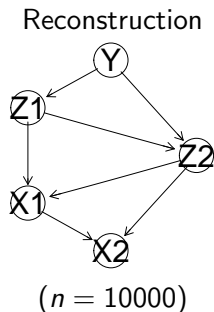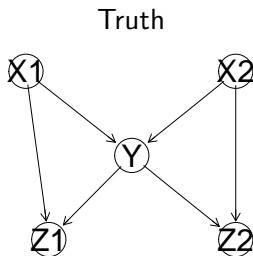
$$Y = X_1 + X_2 + \epsilon_Y.$$

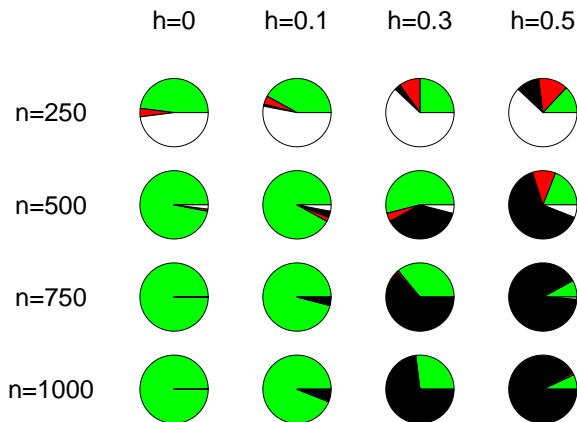Adjustable heterosketasticity:

$$\epsilon_Y \sim N(0, \sigma^2(1 + h(X_1 + X_2))^2).$$

Homoskedastic if $h = 0$.

Causal discovery approach

Correct model is not even identifiable if given purely observational data!

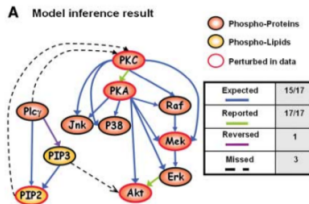Truth

Reconstruction

$(n = 10000)$

How does invariant prediction fare?

Invariant prediction ICP()

Protein Signalling Data

A Model inference result

- Data from Sachs et. al.: 9 interventions.
- 11 variables, 8319 total observations.
- First step: convert all data to log-scale.

Protein Signalling Data

Intervention	Proteins affected	Sample size
1	PIP2, PIP3, Raf	853
2	PIP2, PIP3, Raf	902
3	PIP2, PIP3, Raf	911
4	PIP2, PIP3, Raf, PKC	723
5	PIP2, PIP3, Raf, PIP2	810
6	PIP2, PIP3, Raf, Erk	799
7	PIP2, PIP3, Raf	848
8	PKC	913
9	PKA	707

- For each variable, apply invariant causal prediction algorithm to find its parents.
- Use on interventions which don't affect the variable, e.g. for PKC, consider data from interventions 1, 2, 3, 5, 6, 7, 9.

Protein Signalling Data ($\alpha = 0.1$)

Protein	Ground truth	ICP output
Raf	PKC	$\cap_\emptyset$
Mek	Raf, PKC	$\cap_\emptyset$
PLCg	PIP3	$\cap_\emptyset$
PIP2	PIP3	PIP3 ($p = 0.11$)
PIP3	PLCg	Mek ($p = 0.18$), Jnk ($p = 0.19$)
Erk	Mek, PKA	$\cap_\emptyset$
Akt	PKA, Erk(?)	$\cap_\emptyset$
PKA	PKC(?)	$\cap_\emptyset$
PKC	PLCg, PIP2	$\cap_\emptyset$
p38	PKA, PKC	$\cap_\emptyset$
Jnk	PKA, PKC	$\cap_\emptyset$

hiddenICP() gave many false positives, but it may be due to lack of model reject ($\cap_\emptyset$) functionality.

Protein Signalling Data ($\alpha = 0.1$)

A closer look at the the PIP3 case: a correct invariant set includes PLCg(3) but not Akt(5) or PIP2(4). Did any correct set get accepted??

accepted set of variables 4,10	X
accepted set of variables 1,2,4	X
accepted set of variables 2,3,10	correct
accepted set of variables 1,4,10	X
accepted set of variables 2,4,10	X
accepted set of variables 3,4,10	correct
accepted set of variables 4,5,10	X
accepted set of variables 3,6,10	correct
accepted set of variables 4,6,10	X
accepted set of variables 3,7,10	correct
accepted set of variables 4,7,10	X
$\vdots$	$\vdots$

Section 4

Discussion

Discussion

- Traditional experiments tend to have extremely precise interventions (e.g. set X to 5.3...) The 'interventions' considered by Peters et al. can be much more general (increase X by 2, add noise to X). Does this extend the applicability of the method?

- Combination of interventional $+$ observational data seems to be much more promising for causal inference than pure observational...

- The method can be easily extended to nonlinear models, but it is not as straightforward to relax additive errors. Could one test invariance of prediction rule rather than invariance of errors?

- The method returns a lower bound of the invariant set, $\hat{S} \subset S^*$. How could one obtain an upper bound of S^* instead?

- It is important to extend the method for hidden variables, but the method supplied in the R package does not seem satisfactory.

References

- Sachs, Karen, et al. "Causal protein-signaling networks derived from multiparameter single-cell data." *Science* 308.5721 (2005): 523-529.
- Nagarajan, Radhakrishnan, Marco Scutari, and Sophie Lèbre. "Bayesian networks in R." Springer 122 (2013): 125-127.
- Pearl, Judea. *Causality.* Cambridge university press, 2009.
- Morgan, Stephen L., and Christopher Winship. *Counterfactuals and causal inference.* Cambridge University Press, 2014.
- Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference using invariant prediction: identification and confidence intervals." arXiv preprint arXiv:1501.01332 (2015).