LEARNING CAUSAL EFFECTS: BRIDGING INSTRUMENTS AND BACKDOORS

Ricardo Silva
Department of Statistical Science and Centre for Computational Statistics and Machine Learning, UCL
Fellow, Alan Turing Institute
ricardo@stats.ucl.ac.uk

Joint work with Robin Evans (Oxford) and Shohei Shimizu (Osaka)

Pittsburgh, September 2016
Goal

- To learn the causal effect of some treatment $X$ on some outcome $Y$ with observational data.

Assumptions:
- $Y$ does not precede $X$ causally
- $X$ and $Y$ do not precede any other covariates measured
- Variations of faithfulness and parameterizations
We will cover:

- **The linear case**, where all variables are continuous and all relationships are linear
  - Sets of causal effects can be discovered, sometimes.
  - The role of non-Gaussianity.

- **The nonlinear discrete case** (binary in particular)
  - The goal is to bound causal effects.
  - The faithfulness continuum.
Take-home Messages

- The results will rely on different ways of combining backdoor structures and instrumental variables.

- Discussion points:
  - How to explore redundancies and/or contradictions of assumptions?
  - How to do sensitivity analysis?
  - How to deal with weak associations, both on discovery and control?
  - Please interrupt me at any moment.
QUICK BACKGROUND
We would like to infer $P(\text{Outcome} \mid \text{Treatment})$ in a “world” (regime) like this. All we have is (lousy?) data for $P(\text{Outcome} \mid \text{Treatment})$ in a “world” (regime) like this instead.

We better make use of an indexing notation to distinguish these cases. I will adopt Pearl’s “do” operator.
The jump to causal conclusions from observational data requires assumptions linking different regimes.

Interventional Regime:
\[ P(\text{Outcome} \mid \text{do}(\text{Treatment})) \]

Observational Regime:
\[ P(\text{Outcome} \mid \text{Treatment}) \]
In what follows, we will assume we are given a treatment variable $X$, and outcome $Y$, and some covariates $Z$ that precede $X$ and $Y$ causally.

Unlike the typical graphical model structure learning problem, we are not interested in reconstructing a full graph. All we care about is $P(Y \mid \text{do}(X = x))$. 
Trick 1: “Adjust” (a.k.a., “The Backdoor Adjustment”)

- Genetic Profile
- Smoking
- Lung cancer
Why It Works

- **Estimand:** $P(Y \mid \text{do}(X = x))$, not $P(Y \mid X = x)$

- **Model:**

- **Relation to estimand:**
  \[
  P(Y \mid \text{do}(x)) = \sum_z P(Y \mid \text{do}(x), Z = z) P(Z = z \mid \text{do}(x))
  \]
Why It Works

\[ P(Y \mid \text{do}(x)) = \sum_{z} P(Y \mid \text{do}(x), Z = z) \, P(Z = z \mid \text{do}(x)) \]

\[ = \sum_{z} P(Y \mid X = x, Z = z) \, P(Z = z) \]
Note: We don’t really need “all” hidden common causes
Trick 2: Instrumental Variables

- Variables that can act as “surrogate” experiments.
- Sometimes they are surrogate experiments.
- Valuable in the presence of unmeasured confounding.
(Conditional) Instrumental Variables

- Conditionally, no direct effect, no unblocked confounding with outcome, not affected by treatment.
Why Do We Care?

- Instrumental variables **constraint** the distribution of the hidden common causes.

- It can be used to infer **bounds** on causal effects or, **under further assumptions**, the causal effects even if hidden common causes are out there.
This is work in progress
Parametric assumptions

- Assume (causal) acyclic graphical model with linear relationships

\[ X_i = \lambda_{i1} X_{p(1)} + \ldots + \lambda_{in} X_{p(n)} + \epsilon_i \]
The ultimate goal is to estimate coefficient $\lambda_{yx}$.
In practice, we will estimate sets of plausible values.
A Test for Back-Door Adjustments

- If error terms are non-Gaussian then \textbf{least-square residuals} of treatment and outcome on covariates are independent if and only if there are no unblocked hidden common causes.

\[ r_X \equiv X - (X \sim Z)_{l.s.} \]

\[ r_Y \equiv Y - (Y \sim X + Z)_{l.s.} \]

\[ r_X \perp r_Y \]

Entner et al. (AISTATS, 2012)
What If They are Dependent?

- Too bad! Go home empty-handed.

- Instrumental variables, maybe?
  - But how to test them?
  - What if one of my covariates could in principle be an instrumental variable?
Linear Instrumental Variables
(or: “All of Econometrics in a Single Slide”)

\[ \sigma_{wx} = \lambda_{xw} \sigma_{ww} \]
\[ \sigma_{wy} = \lambda_{xw} \lambda_{yx} \sigma_{ww} \]
\[ \lambda_{yx} = \sigma_{wy} / \sigma_{wx} \]
IV Discovery

- We would like to discover IVs in the true graph that generated the data, so we could exploit them.

- For that we will focus on a particular graphical characterization of what it means to be an IV.

- We then illustrate why this won’t be easy without further assumptions even in linear systems.
A Graphical Criteria for Defining IVs

- $W$ is an IV, conditioned on $Z$, for $X \rightarrow Y$ if
  1. $Z$ does not d-separate $W$ from $X$
  2. $Z$ d-separates $W$ from $Y$ in the graph where we remove $X \rightarrow Y$
  3. $Z$ are non-descendants of $X$ and $Y$

Notice how 1 and 3 are “easy to test”. 
Falsifying Instrumental Variables

A tetrad constraint.
The Converse Does NOT Hold!

\[ \lambda_{yx} \neq \sigma_{w1y} / \sigma_{w1x} = \sigma_{w2y} / \sigma_{w2x} \]
Say you split your set $Z$ into two: $Z_V$ and $Z_I$, where $Z_V$ are “valid IVs” given $Z_I$, the possible “invalid” ones.

sisVIVE, Kang et al. (JASA, 2015)
If in the true and unknown model we have more than half of $Z$ is valid, we are guaranteed we can use $Z_V$ as instrumental variables (given $Z_I$).
An “Equivalent” Algorithm to sisVIVE

Algorithm 1 IV-BY-MAJORITY_∞

1: **Input:** set of random variables $V \cup \{X, Y\}$
2: **Output:** the causal effect of $X$ on $Y$, or a value (NA) indicating lack of knowledge
3: for each $W_i \in V$ do
4:     $Z_i \leftarrow V \setminus \{W_i\}$
5:     $\beta_i \leftarrow \sigma_{w_i y, z_i} / \sigma_{w_i x, z_i}$
6: end for
7: if more than half of set $\{\beta_i\}$ is equal to the same value $\beta$ then
8:     return $\beta$
9: end if
10: return NA
All of $W_1, \ldots, W_{100}$ are valid IVs, if we don’t condition on $Z_{101}$

But sisVIVE requires a variable is either an IV or a conditioning variable…
Algorithm 2 TETRAD-IV∞

1: **Input:** set of random variables $V \cup \{X, Y\}$
2: **Output:** $C$, a set of candidate differential causal effects of $X$ on $Y$
3: Initialize $C \leftarrow \emptyset$
4: for each pair $\{W_i, W_j\} \subseteq V$ do
5:     for every set $Z \subseteq V \setminus \{W_i, W_j\}$ do
6:         if $\sigma_{w_i x . z} = 0$ or $\sigma_{w_j x . z} = 0$ then
7:             next
8:         end if
9:         if $\sigma_{w_i x . z} \sigma_{w_j y . z} \neq \sigma_{w_i y . z} \sigma_{w_j x . z}$ then
10:            next
11:        end if
12:        $C \leftarrow C \cup \{\sigma_{w_i y . z} / \sigma_{w_i x . z}\}$
13:    end for
14: end for
15: return $C$
What is the graphical converse of the tetrad constraint?

Known: the Tetrad Representation Theorem, via the notion of “choke point”.

\[
\begin{align*}
W_1 & \quad X \quad Y \\
W_2 & \quad U \\
X & \text{ is a choke point for } \{W_1, W_2\} \times \{X, Y\}
\end{align*}
\]

\[
\begin{align*}
W_1 & \quad X \quad Y \\
W_2 & \quad U_1 \quad U_2 \\
U_1 & \text{ is a choke point for } \{W_1, W_2\} \times \{X, Y\}
\end{align*}
\]
What is the graphical converse of the conditional tetrad constraint?

- Cannot appeal to the known result anymore: DAGs are not closed under conditioning.
- Instead, *re-interpret* a more recent result by Sullivant et al. (Annals of Stats, 2010)
Cross-covariance of two sets $A$ and $B$ will drop rank if "small enough" sets "t-separate" $A$ from $B$.

Here, $V_0$ "t-separates" \{\$V_{i1}, V_{i2}, V_0\} from \{\$V_{j1}, V_{j2}, V_0\}

The rank of cross-covariance of these two sets will be (typically) 2.
Conditional Tetrad Constraint Interpretation

- If $\sigma_{i.x.z} \sigma_{j.y.z} = \sigma_{i.y.z} \sigma_{j.x.z}$, there will be a set that includes $Z$ that $t$-separates $\{W_i, W_j, Z\}$ from $\{X, Y, Z\}$.

$$|\Sigma_{\{W_i, W_j, Z\}, \{X, Y, Z\}}| = |\Sigma_{ZZ}||\sigma_{i.x.z} \sigma_{j.y.z} - \sigma_{i.y.z} \sigma_{j.x.z}|$$

- This is a necessary but not sufficient condition to guarantee Criterion 2:
  - “$Z$ $d$-separates $W$ from $Y$ in the graph where we remove $X \Rightarrow Y$”
Each TETRAD-IV output can be explained by these “choke sets”. If they differ, it is because of
- a latent element in this choke set (choke set is Z and “U_z”, instead of Z and X), which links “IVs” to Y
- a rogue non-directed path activated by conditioning
Tetrad Equivalence Class

- Size can increase linearly with the number of variables!
If there is at least one genuine pair of conditional IVs in the solution, then the output set provides upper and lower bounds on causal effect.

- This is a much weaker assumption than the one in sisVIVE.

Also: <INCLUDE FAVOURITE PET IDENTIFYING ASSUMPTION HERE>

- “Largest set wins”
- “Strongest association wins”
- “Exclude implausibly large effects”
- “Most common sign wins”
- Etc.
We can generalize the main result of Entner et al. (2012), and exclude solutions that are due to non-directed active paths by a testable condition.
Unfortunately, this also excludes some genuine IVs. Those will not be excluded if backdoors with treatment X are blocked.
Empirical Results

- This is work in progress.
- Practical implementation does not use tests of tetrad constraints: much of the signal is weak, tests perform horribly.
  - Without going in details, it clusters empirical estimates of causal effects, assumes a minimal number of IVs.
- Practical implementation does not do combinatorial search on $Z$: again too much error. Instead, an all-or-nothing is suggested: discard solutions that fail the non-Gaussianity tests.
- It does well in sample sizes relatively large, and seems to be comfortably better than sisVIVE when its assumptions fail. Non-Gaussianity tests require very large sample sizes though.
- Contact me for current manuscript (soon to be re-arXived)
THE NON-LINEAR DISCRETE (BINARY) CASE
The Problem

- Given binary X precedes binary Y causally, estimate average causal effect (ACE) using observational data

\[
\text{ACE} \equiv E[Y \mid \text{do}(X = 1)] - E[Y \mid \text{do}(X = 0)] = \\
P(Y = 1 \mid \text{do}(X = 1)) - P(Y = 1 \mid \text{do}(X = 0))
\]
Goal

- To get an estimate of **bounds** of the ACE
- Rely on the identification of an auxiliary variable $W$ (**witness**), an auxiliary set $Z$ (**background set**), and **assumptions about strength of dependencies** on latent variables
But where do the missing edges come from?

\[ L_{P(Y, X \mid W)} \leq \text{ACE} \leq U_{P(Y, X \mid W)} \]
Exploiting Independence Constraints

- **Faithfulness** provides a way of sometimes finding a point estimator

  - Faithfulness means independence in probability iif “structural” independence (Spirtes et al., 1993)
Faithfulness

- **W independent of Y, but not when given X:**
  conclude the following (absentia hidden common causes)

\[
X = aW + bY + e_x
\]

\[
P(W, X, Y) = P(W)P(Y)P(X \mid W, Y)
\]

\[
P(W, Y \mid X) \propto P(W)P(Y)P(X \mid W, Y)
\]
(Lack of) Faithfulness

- $W$ independent of $Y$, but not when given $X$: different structure
The Problem with Naïve Back-Door Adjustment

- It is not uncommon in applied sciences to posit that, given a large number of covariates $Z$ that are plausible common causes of $X$ and $Y$, we should adjust for all

\[
P_{\text{est}}(Y = 1 \mid \text{do}(X = x)) = \sum_{z} P(Y = 1 \mid x, z)P(z)
\]

- Even if there are remaining unmeasured confounders, a common assumption is that adding elements of $Z$ will in general decrease bias

\[
|\text{ACE}_{\text{true}} - \text{ACE}_{\text{hat}}|
\]
The Problem with Naïve Back-Door Adjustment

Example of failure:

\[ P(Y = 1 \mid \text{do}(X = x)) = P(Y = 1 \mid X = x) \neq \sum_z P(Y = 1 \mid x, z)P(z) \]

Exploiting Faithfulness: A Very Simple Example

- $W \perp\!
\!
\perp Y, W \perp\!
\!
\perp Y \mid X + \text{Faithfulness. Conclusion?}$

Naïve estimator vindicated:

$$\text{ACE} = P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0)$$

- This super-simple nugget of causal information has found some practical uses on large-scale problems

No unmeasured confounding
Entner, Hoyer and Spirtes (2013) AISTATS: two simple rules based on finding a witness $W$ for a correct admissible background set $Z$

- Generalizes “chain models” $W \rightarrow X \rightarrow Y$

R1: If there exists a variable $w \in W$ and a set $Z \subseteq W \setminus \{w\}$ such that

(i) $w \perp y \mid Z$, and
(ii) $w \perp y \mid Z \cup \{x\}$

then infer ‘±’ and give $Z$ as an admissible set.
Rule 1: Illustration

R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp y \mid \mathcal{Z}$, and

(ii) $w \perp y \mid \mathcal{Z} \cup \{x\}$

then infer ‘±’ and give $\mathcal{Z}$ as an admissible set.

- Note again the necessity of the dependence of $W$ and $Y$
What if instead of using $W$ to find $Z$ to make an adjustment by the back-door criterion, we find a $Z$ to allow $W$ to be an instrumental variable that gives bounds on the ACE?
Why do We Care?

- A way to weaken the faithfulness assumption
  - Suppose also by “independence”, we might mean “weak dependence” (and by “dependence”, we might mean “strong dependence”)

- How would interpret the properties of W in this case, given Rule 1?

R1: If there exists a variable \( w \in W \) and a set \( Z \subseteq W \setminus \{w\} \) such that

(i) \( w \not\perp y \mid Z \), and

(ii) \( w \perp y \mid Z \cup \{x\} \)

then infer ‘\( \neq \)’ and give \( Z \) as an admissible set.
Modified Setup: Main Assumption Statement

- Given Rule 1, assume W is a “conditional IV for \( X \to Y \)” in the sense that given Z:
  - All active paths between W and X are into X
  - There is no “strong direct effect” of W on Y
  - There are no “strong active paths” between W and X, nor W and Y, through common ancestors of X and Y
- The definition of “strong effect/path” creates free parameters we will have to deal with, and a continuum of faithfulness-like assumptions.
Bounds on the ACE in the “standard IV model” can be quite wide even when $W \perp Y \mid X$.

This means faithfulness can be quite a strong assumption, and/or “worst-case” analysis can be quite conservative.

Upper minus lower bound = $1 - |P(X = 1 \mid W = 1) - P(X = 1 \mid W = 0)|$
Motivation

- Our analysis can be seen as a way of bridging the two extremes of point estimators of faithfulness analysis and IV bounds without effect constraints.
The High-Level Idea

- The following might be complicated, but here’s a summary:
  
  - Introduce a **redundant parameterization**, parameters for the two regimes (observational regime, and regime with intervention on $X$).
  
  - These parameters cannot be fully unconstrained if we assume “some edges are weak”.
    - Machinery behind is linear programming.
  
  - So statistical inference on the observational regime implies statistical inference on bounds of the ACE.
    - Machinery behind is Bayesian learning with MCMC.
Illustration of Result: Influenza Data

- Effect of influenza vaccination (X) on hospitalization (Y = 1 means hospitalized)
- Covariate GRP: randomized, doctor of that patient received letter to encourage vaccination
  - Bounds on **average causal effect** using standard methods: [-0.23, 0.64]
- The method we will discuss instead picked DM (diabetes history), AGE (dichotomized at 60 years) and SEX as variables that allowed for adjustment.
Influenza Data

- Our method’s estimated interval: $[-0.10, 0.17]$.
- Under some sensitivity analysis postprocessing, the estimate was $[-0.02, 0.02]$. 
Influenza Data: Full Posterior Plots
Influenza Data: Full Posterior Plots

Marginal Posterior Distribution (means: [-0.10, 0.17])

Marginal Posterior Distribution (means: [-0.07, 0.16])
The following might be complicated, but here’s a summary:

- Introduce a **redundant parameterization**, parameters for the two regimes (observational, and intervention on X).
- These parameters cannot be fully unconstrained if we assume “some edges are weak”.
  - Machinery behind is linear programming.
- So statistical inference on the observational regime implies statistical inference on bounds of the ACE.
  - Machinery behind is Bayesian learning with MCMC.
Expressing Assumptions

- Some notation first, ignoring Z for now:

\[
\zeta_{yxw}^* \equiv P(Y = y, X = x \mid W = w, U) \\
\eta_{xw}^* \equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta_w^* \equiv P(X = 1 \mid W = w, U)
\]
Stating Assumptions

\[
\begin{align*}
\zeta^*_{yxw} & \equiv P(Y = y, X = x \mid W = w, U) \\
\eta^*_{xw} & \equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta^*_w & \equiv P(X = 1 \mid W = w, U)
\end{align*}
\]

\[
|\delta^*_w - P(X = 1 \mid W = w)| \leq \epsilon_x
\]

\[
|\eta^*_{xw} - P(Y = 1 \mid X = x, W = w)| \leq \epsilon_y
\]

\[
|\eta^*_{x1} - \eta^*_{x0}| \leq \epsilon_w
\]
Stating Assumptions

\[ \beta P(U) \leq P(U | W = w) \leq \bar{\beta} P(U) \]
Let $\zeta_{yx,w}$ be the expectation of the first entry by $P(U \mid W)$: this is $P(Y = y, X = x \mid W = w)$.

Similarly, let $\eta_{xw}$ be the expectation of the second entry: this is $P(Y = 1 \mid \text{do}(X = x), W = w)$. 

\[
\begin{align*}
\zeta_{yx,w}^* & \equiv P(Y = y, X = x \mid W = w, U) \\
\eta_{xw}^* & \equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta_w^* & \equiv P(X = 1 \mid W = w, U)
\end{align*}
\]
The parameterization given was originally exploited by Dawid (2000) and Ramsahai (2012). It provides an alternative to the structural equation model parameterization of Balke and Pearl (1997). Both approaches work by mapping the problem of testing the model and bounding the ACE by a linear program. We build on this strategy, with some generalizations.
Estimation

- Simpler mapping on $(\delta^*, \eta^*) \rightarrow P(W, X, Y \mid U)$, marginalized, gives constraints on $\zeta \equiv P(W, X, Y)$

- Test whether constraints hold, if not provide no bounds

- Plug-in estimates for $\zeta$ to get $(\zeta, \eta)$ polytope. Find upper bounds and lower bounds on the ACE by solving linear program and maximizing/minimizing objective function

$$f(\eta) = (\eta_{11} - \eta_{01})P(W = 1) + (\eta_{10} - \eta_{00})P(W = 0)$$
Coping with Non-linearity

- Notice that because of constraints such as
  \[ |\delta_w^* - P(X = 1 \mid W = w)| \leq \epsilon_x \]
  there will be non-linear constraints in \( \zeta \equiv P(W, X, Y) \)

- The implied constraints are still linear in \( \eta \equiv P(Y \mid \text{do}(X), W) \). So linear programming formulation still holds, treating \( \zeta \) as a constant.

  - Non-linearity on \( \zeta \) can be a problem for estimation of \( \zeta \) and derivation of confidence intervals. We will describe later a Bayesian approach that does that simply by rejection sampling.
In what follows, we assume dimensionality of $Z$ is small, $|Z| < 10$

**input**: Binary data matrix $\mathcal{D}$; set of relaxation parameters $\theta$; covariate index set $\mathcal{W}$; cause-effect indices $X$ and $Y$

**output**: A list of pairs (witness, admissible set) contained in $\mathcal{W}$

\[
\mathcal{L} \leftarrow \emptyset;
\]

for each $W \in \mathcal{W}$ do

for every admissible set $Z \subseteq \mathcal{W}\backslash\{W\}$ identified by $W$ and $\theta$ given $\mathcal{D}$ do

$\mathcal{B} \leftarrow$ posterior over upper/lowered bounds on the ACE as given by $(W, Z, X, Y, \mathcal{D}, \theta)$;

if there is no evidence in $\mathcal{B}$ to falsify the $(W, Z, \theta)$ model then

$\mathcal{L} \leftarrow \mathcal{L} \cup \{\mathcal{B}\}$;

end

end

end

return $\mathcal{L}$
Recap: So far, everything in the population

- “Rely on the identification of an auxiliary variable W (witness), an auxiliary set Z (background set), and assumptions about strength of dependencies on latent variables”
To decide on independence, we do Bayesian model selection with a contingency table model with Dirichlet priors.

For each pair $(W, Z)$, find posterior bounds for each configuration of $Z$.

- Use Dirichlet prior for $\zeta$ (for each $Z = z$), conditioned on the constraints of the model, using rejection sampling.
  - Propose from unconstrained Dirichlet
- Reject model if 95% or more of proposed parameters are rejected in the initial round of rejection sampling.
- Feed sample from the posterior of $\zeta$ into linear program to get a sample for the upper bound and lower bound.
Why not put a prior directly on the latent variable model?

- Model is unidentifiable $\Rightarrow$ results extremely sensitive to priors
- Putting priors directly into $\zeta$ produces no point estimates, but avoids prior sensibility
Finally, one is left with different posterior distributions over different bounds on the ACE.

Final step is how to summarize possibly conflicting information. Possibilities are:

- Report tightest bound
- Report widest bound
- Report combined smallest lower bound with largest upper bound
- Use “posterior of Rule 1” to pick a handful of bounds and discard others
Recap

- Invert usage of Entner’s Rules towards the instrumental variable point of view.

- Obtain bounds, not point estimates.

- Use Bayesian inference, set up a rule to combine possibly conflicting information.
Because the framework relies on using a linear program to protect a witness variable against violations of faithfulness, we call this the **Witness Protection Program** (WPP) algorithm.
Illustration: Synthetic Studies

- 4 observable nodes, “basic set”, form a pool that can generate a possible (witness, background set) pair
- 4 observable nodes form a “decoy set”: none of them should be included in the background set
- Graph structures over “basic set” + \{X, Y\} are chosen randomly
- Observable parents of “decoy set” are sampled from “basic set”
- Each decoy has another four latent parents, \{L_1, L_2, L_3, L_4\}
- Latents are mutually independent
- Each latent variable \(L_i\) uniformly chooses either X or Y as a child
- Conditional distributions are logistic regression models with pairwise interactions
Illustration: Synthetic Studies

- Relaxations

- Estimators:
  - Posterior expected bounds
  - Naïve 1: back-door adjustment conditioning on everybody
  - Naïve 2: plain \( P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0) \)
  - Backdoor by faithfulness
Example

- Note: no theoretical witness solution
Evaluation

- **Bias definition:**
  - For point estimators, just absolute value of difference between true ACE and estimate.
  - For bounds, Euclidean distance between true ACE and nearest point in the bound.

- **Summaries (over 100 simulations):**
  - Bias average
  - Bias tail mass at 0.1
    - Proportion of cases where bias exceeds 0.1

- **Notice difficulty of direct comparisons**
### Summary

#### Hard, Solvable: \( \text{NE1} = (0.18, 1.00), \text{NE2} = (0.19, 0.63) \)

<table>
<thead>
<tr>
<th>( k_\epsilon )</th>
<th>( \text{Found} )</th>
<th>( \text{Faith.1} )</th>
<th>( \text{WPP1} )</th>
<th>( \text{Width1} )</th>
<th>( \text{WPP2} )</th>
<th>( \text{Width2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.75</td>
<td>0.15</td>
<td>0.57</td>
<td>0.12</td>
<td>0.43</td>
<td>0.06</td>
</tr>
<tr>
<td>0.10</td>
<td>0.95</td>
<td>0.16</td>
<td>0.58</td>
<td>0.10</td>
<td>0.37</td>
<td>0.14</td>
</tr>
<tr>
<td>0.15</td>
<td>0.99</td>
<td>0.15</td>
<td>0.60</td>
<td>0.07</td>
<td>0.27</td>
<td>0.21</td>
</tr>
<tr>
<td>0.20</td>
<td>0.99</td>
<td>0.15</td>
<td>0.61</td>
<td>0.05</td>
<td>0.17</td>
<td>0.28</td>
</tr>
<tr>
<td>0.25</td>
<td>1.00</td>
<td>0.15</td>
<td>0.60</td>
<td>0.03</td>
<td>0.13</td>
<td>0.36</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.16</td>
<td>0.61</td>
<td>0.02</td>
<td>0.09</td>
<td>0.43</td>
</tr>
</tbody>
</table>

#### Easy, Solvable: \( \text{NE1} = (0.04, 0.13), \text{NE2} = (0.08, 0.29) \)

<table>
<thead>
<tr>
<th>( k_\epsilon )</th>
<th>( \text{Found} )</th>
<th>( \text{Faith.1} )</th>
<th>( \text{WPP1} )</th>
<th>( \text{Width1} )</th>
<th>( \text{WPP2} )</th>
<th>( \text{Width2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.78</td>
<td>0.04</td>
<td>0.09</td>
<td>0.03</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>0.10</td>
<td>0.97</td>
<td>0.04</td>
<td>0.09</td>
<td>0.02</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td>0.15</td>
<td>0.99</td>
<td>0.05</td>
<td>0.10</td>
<td>0.01</td>
<td>0.04</td>
<td>0.21</td>
</tr>
<tr>
<td>0.20</td>
<td>0.99</td>
<td>0.05</td>
<td>0.10</td>
<td>0.01</td>
<td>0.04</td>
<td>0.29</td>
</tr>
<tr>
<td>0.25</td>
<td>0.99</td>
<td>0.05</td>
<td>0.09</td>
<td>0.00</td>
<td>0.01</td>
<td>0.37</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.05</td>
<td>0.09</td>
<td>0.00</td>
<td>0.01</td>
<td>0.44</td>
</tr>
</tbody>
</table>

- Bias average
- Bias tail mass at 0.1
### Summary

**Hard, Not Solvable:** $NE_1 = (0.16, 1.00), NE_2 = (0.20, 0.88)$

<table>
<thead>
<tr>
<th>$k_c$</th>
<th>Found</th>
<th>Faith.1</th>
<th>WPP1</th>
<th>Width1</th>
<th>WPP2</th>
<th>Width2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.67</td>
<td>0.20</td>
<td>0.17</td>
<td>0.76</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>0.10</td>
<td>0.91</td>
<td>0.19</td>
<td>0.13</td>
<td>0.63</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>0.15</td>
<td>0.97</td>
<td>0.19</td>
<td>0.10</td>
<td>0.41</td>
<td>0.18</td>
<td>0.01</td>
</tr>
<tr>
<td>0.20</td>
<td>0.99</td>
<td>0.19</td>
<td>0.07</td>
<td>0.25</td>
<td>0.24</td>
<td>0.01</td>
</tr>
<tr>
<td>0.25</td>
<td>1.00</td>
<td>0.19</td>
<td>0.03</td>
<td>0.13</td>
<td>0.31</td>
<td>0.00</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.19</td>
<td>0.02</td>
<td>0.06</td>
<td>0.39</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Easy, Not Solvable:** $NE_1 = (0.09, 0.32), NE_2 = (0.14, 0.56)$

<table>
<thead>
<tr>
<th>$k_c$</th>
<th>Found</th>
<th>Faith.1</th>
<th>WPP1</th>
<th>Width1</th>
<th>WPP2</th>
<th>Width2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.68</td>
<td>0.13</td>
<td>0.10</td>
<td>0.37</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>0.10</td>
<td>0.97</td>
<td>0.12</td>
<td>0.08</td>
<td>0.28</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>0.15</td>
<td>1.00</td>
<td>0.12</td>
<td>0.05</td>
<td>0.17</td>
<td>0.16</td>
<td>0.01</td>
</tr>
<tr>
<td>0.20</td>
<td>1.00</td>
<td>0.12</td>
<td>0.03</td>
<td>0.08</td>
<td>0.23</td>
<td>0.01</td>
</tr>
<tr>
<td>0.25</td>
<td>1.00</td>
<td>0.12</td>
<td>0.02</td>
<td>0.05</td>
<td>0.31</td>
<td>0.00</td>
</tr>
<tr>
<td>0.30</td>
<td>1.00</td>
<td>0.12</td>
<td>0.01</td>
<td>0.04</td>
<td>0.39</td>
<td>0.00</td>
</tr>
</tbody>
</table>
On-going Work

- Finding a more primitive default set of assumptions where assumptions about the relaxations can be derived from
- Doing without a given causal ordering
- Large scale experiments
- Scaling up for a large number of covariates
- Continuous data
- More real data experiments
- R package available at CRAN/GitHub: “CausalFX”
Thank You, and Shameless Ad

What If? Inference and Learning of Hypothetical and Counterfactual Interventions in Complex Systems

A NIPS 2016 Workshop
Centre Convencions Internacional Barcelona, Barcelona, Spain
December 10th 2016

Deadline: October 31st

https://sites.google.com/site/whatif2016nips/call-for-papers