Causality

Joris Mooij and Jonas Peters

Informatics Institute
University of Amsterdam

Department for Mathem. Sciences
University of Copenhagen

MSR, Cambridge
3 July 2018
is based on work by ...

- **UCLA:** J. Pearl
- **CMU:** P. Spirtes, C. Glymour, R. Scheines
- **Harvard University:** D. Rubin, J. Robins
- **ETH Zürich:** P. Bühlmann, N. Meinshausen, N. Pfister, D. Rothenhäusler
- **Max-Planck-Institute Tübingen:** D. Janzing, B. Schölkopf
- **University of Amsterdam:** J. Mooij
- P. Hoyer
- ... and many others
What is a causal model?
Recall Part I

- What if interested in iid prediction, i.e., observational data? Don’t worry (too much) about causality!
Recall Part I

- What if interested in iid prediction, i.e., observational data? Don’t worry (too much) about causality!
- But often, we are interested in a system’s behaviour under intervention.
Recall Part I

- What if interested in iid prediction, i.e., observational data? Don’t worry (too much) about causality!
- But often, we are interested in a system’s behaviour under intervention.
- SCMs entail graphs, obs. distr., interventions and counterfactuals.

\[
\begin{align*}
    X_1 & := f_1(X_3, N_1) \\
    X_2 & := f_2(X_1, X_3, N_2) \\
    X_3 & := f_3(N_3)
\end{align*}
\]

- \(N_i\) jointly independent
- \(G_0\) has no cycles
Recall Part I

○ What if interested in iid prediction, i.e., observational data? Don’t worry (too much) about causality!

○ But often, we are interested in a system’s behaviour under intervention.

○ SCMs entail graphs, obs. distr., interventions and counterfactuals.

\[
\begin{align*}
X_1 &:= f_1(X_3, N_1) \\
X_2 &:= f_2(X_1, X_3, N_2) \\
X_3 &:= f_3(N_3)
\end{align*}
\]

• \( N_i \) jointly independent
• \( G_0 \) has no cycles

○ adjusting: graph + observational distribution \( \rightsquigarrow \) interventions
What if interested in iid prediction, i.e., observational data? Don’t worry (too much) about causality!

But often, we are interested in a system’s behaviour under intervention.

SCMs entail graphs, obs. distr., interventions and counterfactuals.

\[
\begin{align*}
X_1 &= f_1(X_3, N_1) \\
X_2 &= f_2(X_1, X_3, N_2) \\
X_3 &= f_3(N_3)
\end{align*}
\]

- \(N_i\) jointly independent
- \(G_0\) has no cycles

- adjusting: graph + observational distribution \(\leadsto\) interventions
- instrumental variables: may help if there are hidden variables
Part II: Causal Discovery
counterfactuals

distribution $P$

causal model e.g. SCM

causal graph

interv. distr. $Q_1, Q_2, \ldots$
The Problem of Causal Discovery:

observed iid data from \( P(X_1, \ldots, X_5) \)

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>-0.3</td>
<td>5.8</td>
<td>-2.1</td>
<td>2.2</td>
</tr>
<tr>
<td>1.7</td>
<td>-0.2</td>
<td>7.0</td>
<td>-1.2</td>
<td>0.4</td>
</tr>
<tr>
<td>-2.4</td>
<td>-0.1</td>
<td>4.3</td>
<td>-0.7</td>
<td>3.5</td>
</tr>
<tr>
<td>2.3</td>
<td>-0.3</td>
<td>5.5</td>
<td>-1.1</td>
<td>-4.4</td>
</tr>
<tr>
<td>3.5</td>
<td>-0.2</td>
<td>3.9</td>
<td>-0.9</td>
<td>-3.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

causal model, e.g. DAG \( G \)
The Problem of Causal Discovery:

observed iid data from $P(X_1, \ldots, X_5)$

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>−0.3</td>
<td>5.8</td>
<td>−2.1</td>
<td>2.2</td>
</tr>
<tr>
<td>1.7</td>
<td>−0.2</td>
<td>7.0</td>
<td>−1.2</td>
<td>0.4</td>
</tr>
<tr>
<td>−2.4</td>
<td>−0.1</td>
<td>4.3</td>
<td>−0.7</td>
<td>3.5</td>
</tr>
<tr>
<td>2.3</td>
<td>−0.3</td>
<td>5.5</td>
<td>−1.1</td>
<td>−4.4</td>
</tr>
<tr>
<td>3.5</td>
<td>−0.2</td>
<td>3.9</td>
<td>−0.9</td>
<td>−3.9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Correlation (dependence) does not imply causation ... but RECALL:
Definition

$P$ satisfies the (global) Markov condition w.r.t. $G$ if

$$
\text{\underline{X and Y are d-separated by S in G}} \quad \Rightarrow \quad \text{\underline{X \perp \!\!\!\!\!\!\!\!\!\!\perp Y | S}}
$$

properties of graph

properties in $P$
Definition

\( P \) satisfies the (global) Markov condition w.r.t. \( G \) if

\[
X \text{ and } Y \text{ are } d\text{-separated by } S \text{ in } G \quad \Rightarrow \quad X \perp \!
\perp Y \mid S
\]

Definition

\( P \) satisfies faithfulness w.r.t. \( G \) if

\[
X \text{ and } Y \text{ are } d\text{-separated by } S \text{ in } G \quad \iff \quad X \perp \!
\perp Y \mid S
\]
Idea 1: independence-based methods

\[ \mathbb{P}(X_1, \ldots, X_4) \]

- Independence-based methods
  - \( P(X_1, \ldots, X_4) \)
  - Independence tests
  - Faithful Markov

\[ X_1 \perp X_2 \\
X_2 \perp X_3 \\
X_1 \perp X_4 \mid \{X_3\} \\
X_1 \perp X_2 \mid \{X_3\} \\
X_2 \perp X_3 \mid \{X_1\} \]

Method: IC (Pearl 2009); PC, FCI (Spirtes et al., 2000)

1. Find all (cond.) independences from the data.
2. Select the DAG(s) that corresponds to these independences.

Jonas Peters (Københavns Universitet)
Idea 1: independence-based methods

\[ P(X_1, \ldots, X_4) \]

\[
\begin{align*}
X_1 & \independent X_2 \\
X_2 & \independent X_3 \\
X_1 & \independent X_4 | \{X_3\} \\
X_1 & \independent X_2 | \{X_3\} \\
X_2 & \independent X_3 | \{X_1\}
\end{align*}
\]

Method: IC (Pearl 2009); PC, FCI  (Spirtes et al., 2000)

1. Find all (cond.) independences from the data.
2. Select the DAG(s) that corresponds to these independences.

Jonas Peters  (Københavns Universitet)  
Causality  
3 July 2018
Idea 1: independence-based methods

\[ P(X_1, \ldots, X_4) \]

\begin{align*}
X_1 & \perp X_2 \\
X_2 & \perp X_3 \\
X_1 & \perp X_4 \mid \{X_3\} \\
X_1 & \perp X_2 \mid \{X_3\} \\
X_2 & \perp X_3 \mid \{X_1\} \\
\end{align*}

Method: IC (Pearl 2009); PC, FCI (Spirtes et al., 2000)

1. Find all (cond.) independences from the data.
2. Select the DAG(s) that corresponds to these independences.
Idea 1: independence-based methods

**Exercise 1:** Assume that in $P(A,B,C)$ we find
- $A \perp \perp C \mid B$
- $C \perp \perp A \mid B$
- no other (cond.) independence

Find all DAGs $G$ s.t. $P(A,B,C)$ is Markov and faithful wrt $G$.

**Exercise 2:** Assume that in $P(A,B,C,D)$ we find
- $A \perp B$
- $B \perp A$
- no other (cond.) independence

Find all DAGs $G$ s.t. $P(A,B,C,D)$ is Markov and faithful wrt $G$. 
**Definition: d-separation**

$X_i$ and $X_j$ are $d$-separated by $S$ if all paths between $X_i$ and $X_j$ are blocked by $S$.

Check, whether all paths blocked!!

$X_2$ and $X_5$ are $d$-sep. by $\{X_1, X_4\}$

$X_4$ and $X_1$ are $d$-sep. by $\{X_2, X_3\}$

$X_2$ and $X_4$ are $d$-sep. by $\{}$

$X_4$ and $X_1$ are NOT $d$-sep. by $\{X_3, X_5\}$
Idea 1: independence-based methods

\[ P(X_1, \ldots, X_4) \]

**Method:** IC (Pearl 2009); PC, FCI (Spirtes et al., 2000)

1. Find all (cond.) independences from the data. **Be smart.**
2. Select the DAG(s) that corresponds to these independences.
Watch out!

Watch out!

Any conditional independence test with correct level has NO POWER (... if the cond. variable is continuous)


Idea 2: restricted structural causal models

Mooij, JP, Janzing, Zscheischler, Schölkopf: *Disting. cause from effect using obs. data: methods and benchm.*, JMLR 2016
Idea 2: restricted structural causal models

Assume $P(X_1, \ldots, X_4)$ has been entailed by

\[
\begin{align*}
X_1 &= f_1(X_3, N_1) \\
X_2 &= N_2 \\
X_3 &= f_3(X_2, N_3) \\
X_4 &= f_4(X_2, X_3, N_4)
\end{align*}
\]

- $N_i$ jointly independent
- $G_0$ has no cycles

Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?
Idea 2: restricted structural causal models

Assume $P(X_1, \ldots, X_4)$ has been entailed by

\[
\begin{align*}
X_1 &= f_1(X_3, N_1) \\
X_2 &= N_2 \\
X_3 &= f_3(X_2, N_3) \\
X_4 &= f_4(X_2, X_3, N_4)
\end{align*}
\]

- $N_i$ jointly independent
- $G_0$ has no cycles

Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? No.
Idea 2: restricted structural causal models

Assume \( P(X_1, \ldots, X_4) \) has been entailed by

\[
\begin{align*}
X_1 &= f_1(X_3) + N_1 \\
X_2 &= N_2 \\
X_3 &= f_3(X_2) + N_3 \\
X_4 &= f_4(X_2, X_3) + N_4
\end{align*}
\]

- \( N_i \sim \mathcal{N}(0, \sigma_i^2) \) jointly independent
- \( G_0 \) has no cycles

**Additive noise model with Gaussian noise.**

Can the DAG be recovered from \( P(X_1, \ldots, X_4) \)? **Yes iff \( f_i \) nonlinear.**

JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, JMLR 2014*

Example: altitude and temperature

![Graph showing the relationship between altitude and temperature.](chart)

- **p-value forward**: 0.024
- **p-value backward**: 0.0000000000019 (show code)

Jonas Peters (Københavns Universitet)
Example: altitude and temperature

\begin{center}
\begin{tikzpicture}
\begin{axis}[
    width=\textwidth,
    xlabel=altitude,
    ylabel=temperature,
    xmin=0, xmax=3000,
    ymin=-5, ymax=10,
    grid=both,
    grid style={line width=0.5pt, white!30!black},
    xlabel style={at={(axis description cs:1,0.5)}, anchor=north},
    ylabel style={at={(axis description cs:0.5,1)}, anchor=south},
]
\addplot[red, thick] table[x=altitude, y=temperature] {data.csv};
\end{axis}
\end{tikzpicture}
\end{center}

p-value forward: 0.024

p-value backward: 0.0000000000019 (show code)
### Idea 2: restricted structural causal models

<table>
<thead>
<tr>
<th>$d$</th>
<th>number of DAGs with $d$ nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>543</td>
</tr>
<tr>
<td>5</td>
<td>29281</td>
</tr>
<tr>
<td>6</td>
<td>3781503</td>
</tr>
<tr>
<td>7</td>
<td>1138779265</td>
</tr>
<tr>
<td>8</td>
<td>783702329343</td>
</tr>
<tr>
<td>9</td>
<td>1213442454842881</td>
</tr>
<tr>
<td>10</td>
<td>4175098976430598143</td>
</tr>
<tr>
<td>11</td>
<td>31603459396418917607425</td>
</tr>
<tr>
<td>12</td>
<td>521939651343829405020504063</td>
</tr>
<tr>
<td>13</td>
<td>18676600744432035186664816926721</td>
</tr>
<tr>
<td>14</td>
<td>1439428141044398334941790719839535103</td>
</tr>
<tr>
<td>15</td>
<td>237725265553410354992180218286376719253505</td>
</tr>
<tr>
<td>16</td>
<td>83756670773733320287699303047996412235223138303</td>
</tr>
<tr>
<td>17</td>
<td>6270792119692388999446452602494921906963551482675201</td>
</tr>
<tr>
<td>18</td>
<td>99421195322159515895228914592354524516555026878588305014783</td>
</tr>
<tr>
<td>19</td>
<td>332771901227107591736177573311261125883583076258421902583546773505</td>
</tr>
<tr>
<td>20</td>
<td>234488045105108898815255985229099188899081192234291298795803236068491263</td>
</tr>
<tr>
<td>21</td>
<td>3469876828358875002875932843018108822313944540438601719027559113446586077675521</td>
</tr>
<tr>
<td>22</td>
<td>1075822921725761493652956179327624326573727662809185218104090000500559527511693495107583</td>
</tr>
</tbody>
</table>

https://oeis.org/A003024/b003024.txt
Idea 2: restricted structural causal models

Peters et al. (2009), Bauer et al. (2016):

Theorem

Let \((X_t)_t\) be a causal\(^a\) solution of an ARMA\((p, q)\) process:

\[
X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}.
\]

Then, \(X_t\) is time reversible, i.e., a causal solution of an ARMA\((\tilde{p}, \tilde{q})\) process with reversed time, if and only if \((Z_t)_t\) is Gaussian.

\(^a\)(\(X_t)_t\) causal iff \(Z_t \perp \perp X_{t-k}, k > 0\).
Idea 2: restricted structural causal models

Pickup et al. (2014):

Can we see the arrow of time?
Algorithm can determine, with 80 percent accuracy, whether video is running forward or backward.

Larry Hardesty | MIT News Office
June 20, 2014

Einstein's theory of relativity envisions time as a spatial dimension, like height, width, and depth, and...
Idea 2: restricted structural causal models

Pickup et al. (2014):  

**Method #3: Auto-regressive model**

If object motion is linear, then the current velocity of the object should be affected only by the past. Noise on this motion will be asymmetric in the forward and backward directions, and fitting an auto-regressive model to the linear motion ought to yield independence between the noise and signal only in the forwards-time direction. This method attempts to find the forward direction by looking at the independence of AR fitting error on motion trajectories.

Top: tracked points from a sequence, and an example track. Bottom: Forward-time (left) and backward-time (right) vertical trajectory components, and the corresponding model residuals. Trajectories should be independent from model residuals (noise) in the forward-time direction only. For the example track shown, p-values for the forward and backward directions are 0.52 and 0.016 respectively, indicating that forwards time is more likely.
Suppose there is a target $Y$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.4</td>
<td>−0.3</td>
<td>5.8</td>
<td>−2.1</td>
<td>2.2</td>
</tr>
<tr>
<td>2</td>
<td>1.7</td>
<td>−0.2</td>
<td>7.0</td>
<td>−1.2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>−2.4</td>
<td>−0.1</td>
<td>4.3</td>
<td>−0.7</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>2.3</td>
<td>−0.3</td>
<td>5.5</td>
<td>−1.1</td>
<td>−4.4</td>
</tr>
<tr>
<td>5</td>
<td>3.5</td>
<td>−0.2</td>
<td>3.9</td>
<td>−0.9</td>
<td>−3.9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Regressing on \((X_1, X_2, X_3)\):

The coefficients change.
Regressing on $X_1$, $X_2$, and $X_3$:

$X_2$ yields an invariant model. Ground truth:
Regressing on $X_1$, $X_2$, and $X_3$:

$X_2$ yields an invariant model. Ground truth:


Non-invariant models can be rejected if $\sqrt{\log n / n} = o(a_n)$. 

Jonas Peters (Københavns Universitet)  
Causality  
3 July 2018
Theorem (PBM 2016)

Assume invariance is satisfied for some $S^*$. For any test level $\alpha$ we obtain

$$P(\cap_{S : S \text{ invariant}} \subseteq S^*) \geq 1 - \alpha.$$ 

Identifiability improves if we have more and stronger interventions, at better places, more heterogeneity in the data.

Theorem (PBM 2016)

Assume invariance is satisfied for some $S^*$. For any test level $\alpha$ we obtain

$$P(\cap_{S \text{ invariant}} S \subseteq S^*) \geq 1 - \alpha.$$ 

Identifiability improves if we have more and stronger interventions, at better places, more heterogeneity in the data.

How much CHF do I need to pay for buying 1 EUR?

monthly data Swiss National Bank Jan 1999 - Jan 2017

<table>
<thead>
<tr>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$ exchange rate Euro to Swiss Franks</td>
</tr>
<tr>
<td>$X^1$ change in average call money rate</td>
</tr>
<tr>
<td>$X^2$ log returns of foreign currency investments of the SNB</td>
</tr>
<tr>
<td>$X^3$ log returns of reserve positions at Intern. Monetary Fund of the SNB</td>
</tr>
<tr>
<td>$X^4$ log returns of monetary assistance loans of the SNB</td>
</tr>
<tr>
<td>$X^5$ log returns of Swiss Frank securities of the SNB</td>
</tr>
<tr>
<td>$X^6$ log returns of remaining assets of the SNB</td>
</tr>
<tr>
<td>$X^7$ log returns of Swiss GDP</td>
</tr>
<tr>
<td>$X^8$ log returns of Euro zone GDP</td>
</tr>
<tr>
<td>$X^9$ inflation rate for Switzerland</td>
</tr>
</tbody>
</table>
Figure: left plot (lowest p-value) and right plot (highest p-value)
## description

| $Y$ | exchange rate Euro to Swiss Franks |
| $X^1$ | change in average call money rate |
| $X^2$ | log returns of foreign currency investments of the SNB |
| $X^3$ | log returns of reserve positions at Intern. Monetary Fund of the SNB |
| $X^4$ | log returns of monetary assistance loans of the SNB |
| $X^5$ | log returns of Swiss Frank securities of the SNB |
| $X^6$ | log returns of remaining assets of the SNB |
| $X^7$ | log returns of Swiss GDP |
| $X^8$ | log returns of Euro zone GDP |
| $X^9$ | inflation rate for Switzerland |
variance test

log(p-value) for non-causality

number of lags

-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6

-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6

$X^1$, $X^3$, $X^4$, $X^5$, $X^6$, $X^8$, $X^9$, $X^2$, $X^7$
This works for more complicated models, too (e.g., mixture models).

Here, $X_3$ yields an invariant model. Ground truth:

Relation invariance and causality: R. Christiansen, JP (in preparation)
Here,

= time

Also possible:

= environments
Can we find causes of genes? (We have access to gene deletions.)

ICP (R-package InvariantCausalPrediction)

```r
> ExpInd
[1]11111111111111111111111111111111
...22222222222222
... 

> icp <- ICP(X,Y,ExpInd)
```

<table>
<thead>
<tr>
<th></th>
<th>LOWER</th>
<th>UPPER</th>
<th>MAXIMIN</th>
<th>EFFECT</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>-0.71</td>
<td>-0.52</td>
<td>-0.52</td>
<td>&lt;1e-09</td>
<td>***</td>
</tr>
<tr>
<td>X2</td>
<td>-0.46</td>
<td>0.00</td>
<td>0.00</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>0.58</td>
<td>0.70</td>
<td>0.58</td>
<td>&lt;1e-09</td>
<td>***</td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
What if we have no specific environment variable or time?
constructed by clustering?
= non-descendant of target
Real data: Fertility Data UN Data. World population prospects, 2013

$Y \in \mathbb{R}$: total fertility rate in a country in a given year

$X \in \mathbb{R}^9$:

- IMR – infant mortality rate
- Q5 – under-five mortality rate
- Education expenditure (% of GNI)
- Exports of goods and services (% of GDP)
- GDP per capita (constant 2005 US$)
- GDP per capita growth (annual %)
- Imports of goods and services (% of GDP)
- Primary education (% female)
- Urban population (% of total)

$E \in \mathbb{R}$: continent of the country
Given \((X_1, Y_1, E_1), \ldots, (X_n, Y_n, E_n)\).

Invariance assumption \(H_{0,s}\):

- \(E \perp Y \mid X_S\).
- \((X_1, Y_1, E_1), \ldots, (X_n, Y_n, E_n)\) are i.i.d.
Real data: Fertility Data UN Data. World population prospects, 2013

\( Y \in \mathbb{R} \): total fertility rate in a country in a given year

\( X \in \mathbb{R}^9 \):

- IMR – infant mortality rate
- Q5 – under-five mortality rate
- Education expenditure (% of GNI)
- Exports of goods and services (% of GDP)
- GDP per capita (constant 2005 US$)
- GDP per capita growth (annual %)
- Imports of goods and services (% of GDP)
- Primary education (% female)
- Urban population (% of total)

\( E \in \mathbb{R} \): continent of the country
Real data: Fertility Data UN Data. World population prospects, 2013

\( \mathbf{Y} \in \mathbb{R} \): total fertility rate in a country in a given year

\( \mathbf{X} \in \mathbb{R}^9 \):

- IMR – infant mortality rate
- Q5 – under-five mortality rate
- Education expenditure (% of GNI)
- Exports of goods and services (% of GDP)
- GDP per capita (constant 2005 US$)
- GDP per capita growth (annual %)
- Imports of goods and services (% of GDP)
- Primary education (% female)
- Urban population (% of total)

\( \mathbf{E} \in \mathbb{R} \): continent of the country

\( S_1 = \{ Q5 \} \)

\( S_2 = \{ \text{IMR, Imports of goods and services, Urban pop. (% of total)} \} \)

\( S_3 = \{ \text{IMR, Education expenditure (% of GNI), Exports of goods and services, GDP per capita} \} \)

Heinze-Deml, JP and Meinshausen: Predicting the effect of interventions using inv. principles for nonl. models, arxiv:1706.08576
\[ E \left[ \log(\text{TFR}) \mid \text{do}(X = x_{\text{test}}) \right] - E \left[ \log(\text{TFR}) \mid \text{do}(X = x_{\text{obs}}) \right] \]
Summary Part II:

- Idea 1: independence-based methods (single environment)

- Idea 2: additive noise (single environment)
  \[
  \begin{align*}
  X_1 &= f_1(X_3) + N_1 \\
  X_2 &= N_2 \\
  X_3 &= f_3(X_2) + N_3 \\
  X_4 &= f_4(X_2, X_3) + N_4
  \end{align*}
  \]

- Idea 3: invariant prediction (the more heterogeneity the better!)
Part III: Relations to Machine Learning
Recall:

The coefficients change.
Idea 1: anchor regression

What if there is more than one invariant model? Choose the best predictive model.
Idea 1: anchor regression

What if there is more than one invariant model? Choose the best predictive model.

anchor Regression
Idea 1: anchor regression

What if there is more than one invariant model? Choose the best predictive model.

anchor Regression

Find a trade-off between

- invariance with respect to
- predictive power
Idea 1: anchor regression

\[ b^\gamma := \arg\min_b \mathbb{E}(Y - Xb)^2 + \gamma \| \mathbb{E}A^t(Y - Xb) \|^2_2 \]
Idea 1: anchor regression

\[ b^\gamma := \arg\min_{b} \mathbb{E}(Y - Xb)^2 + \gamma \|\mathbb{E}A^t(Y - Xb)\|^2_2 \]

\( \gamma \to \infty: \text{ IV} \)
\( \gamma \to 0: \text{ OLS} \)

MSE under a shift of \( X \):
Idea 1: anchor regression

\[ b^{\gamma} := \arg \min_b E(Y - Xb)^2 + \gamma \|EA_t(Y - Xb)\|_2^2 \]

\[ \gamma \to \infty: \text{ IV} \]
\[ \gamma \to 0: \text{ OLS} \]

MSE under a shift of \( X \):

![Graph showing MSE under a shift of X]

Jonas Peters (Københavns Universitet)
Causality
3 July 2018
Idea 1: anchor regression

\[
\begin{pmatrix}
X \\
Y \\
H
\end{pmatrix}
\leftarrow
B \cdot \begin{pmatrix}
X \\
Y \\
H
\end{pmatrix} + \varepsilon + MA,
\]

shifted:

\[
\begin{pmatrix}
X^v \\
Y^v \\
H^v
\end{pmatrix}
\leftarrow
B \cdot \begin{pmatrix}
X^v \\
Y^v \\
H^v
\end{pmatrix} + \varepsilon + v.
\]
Idea 1: anchor regression

\[
\begin{pmatrix}
X \\
Y \\
H
\end{pmatrix}
\leftarrow \mathbf{B} \cdot \begin{pmatrix}
X \\
Y \\
H
\end{pmatrix} + \varepsilon + \mathbf{M} \mathbf{A},
\]

shifted:

\[
\begin{pmatrix}
X^v \\
Y^v \\
H^v
\end{pmatrix}
\leftarrow \mathbf{B} \cdot \begin{pmatrix}
X^v \\
Y^v \\
H^v
\end{pmatrix} + \varepsilon + \nu.
\]

**Theorem**

For any \( b \in \mathbb{R}^d \) we have

\[
\argmin_b \mathbb{E}(Y - Xb)^2 + \gamma \| \mathbb{E} A^t (Y - Xb) \|^2 \leq \max_{\nu \in C^\gamma} \mathbb{E}[ (Y^\nu - X^\nu b)^2 ],
\]

where

\[
C^\gamma := \{ \nu = M\delta \text{ such that } \| \delta \|_2 \leq \sqrt{\gamma} \}.
\]
Idea 2: half-sibling regression
Idea 2: half-sibling regression

Milky Way Galaxy

Kepler Search Space
3,000 light years

Sagittarius Arm

Orion Spur

Perseus Arm

Sun
Idea 2: half-sibling regression
Idea 2: half-sibling regression

http://archive.stsci.edu/index.html

Schölkopf et al.: Removing systematic errors for exoplanet search via latent causes, ICML 2015
Schölkopf et al.: Modeling Confounding by Half-Sibling Regression, PNAS 2016
Idea 2: half-sibling regression

http://archive.stsci.edu/index.html

Schölkopf et al.: Removing systematic errors for exoplanet search via latent causes, ICML 2015
Schölkopf et al.: Modeling Confounding by Half-Sibling Regression, PNAS 2016
Idea 2: half-sibling regression

unobserved

true signal  sys. noise

observed

measurement  other measurements
Idea 2: half-sibling regression

Assume $Y = f(N) + Q$.

Proposed idea: Remove everything from $Y$ explained by $X$.

Or: $\hat{Q} = Y - E[Y | X]$.

Proposition: Convergence against "correct" signal $Q$ (up to reparameterization) if perfect reconstruction: $\exists \psi$ such that $f(N) = \psi(X)$.

low noise: $X = g(N) + s \cdot R$ and $s \to 0$.

many $X$'s: $X_i = g_i(N) + R_i$, $i = 1, \ldots, \infty$.
Idea 2: half-sibling regression

Assume $Y = f(N) + Q$.

Proposed idea:
Remove everything from $Y$ explained by $X$. 
Idea 2: half-sibling regression

Assume $Y = f(N) + Q$.

Proposed idea:
Remove everything from $Y$ explained by $X$.

Or: $\hat{Q} := Y - E[Y \mid X]$. 
Idea 2: half-sibling regression

Assume \( Y = f(N) + Q \).

Proposed idea:
Remove everything from \( Y \) explained by \( X \).

Or: \( \hat{Q} := Y - \mathbb{E}[Y | X] \).

Proposition

Convergence against “correct” signal \( Q \) (up to reparameterization) if

- perfect reconstruction: \( \exists \psi \) such that \( f(N) = \psi(X) \)
Idea 2: half-sibling regression

Assume $Y = f(N) + Q$.

Proposed idea:
Remove everything from $Y$ explained by $X$.

Or: $\hat{Q} := Y - E[Y | X]$.

Proposition

Convergence against “correct” signal $Q$ (up to reparameterization) if

- perfect reconstruction: $\exists \psi$ such that $f(N) = \psi(X)$
- low noise: $X = g(N) + s \cdot R$ and $s \to 0$
Idea 2: half-sibling regression

Assume $Y = f(N) + Q$.

Proposed idea:
Remove everything from $Y$ explained by $X$.

Or: $\hat{Q} := Y - E[Y|X]$.

Proposition

Convergence against “correct” signal $Q$ (up to reparameterization) if

- perfect reconstruction: $\exists \psi$ such that $f(N) = \psi(X)$
- low noise: $X = g(N) + s \cdot R$ and $s \to 0$
- many $X$’s: $X_i = g_i(N) + R_i, i = 1, \ldots, \infty$
Idea 2: half-sibling regression

Schölkopf et al.: Removing systematic errors for exoplanet search via latent causes, ICML 2015
Schölkopf et al.: Modeling Confounding by Half-Sibling Regression, PNAS 2016
Idea 3: Blackjack

(some) Rules:

- **Dealing**: player two cards, dealer one card (all face up).
- **Goal**: more points in hand. Face cards: 10, ace either 1 or 11 points.
- **Player’s moves**: *hit* (take card, but try \( \leq 21 \)), *stand*, *double down*, *split* (in case of pair).
- **Dealer’s moves**: deterministic, does not stand before \( \geq 17 \) points.
- **Blackjack**: ace and face card \( \rightarrow 1.5 \cdot \text{bet} \).
Idea 3: Blackjack

https://de.wikipedia.org/wiki/Black_Jack.JPG
Idea 3: Blackjack

When can we learn?

Objects of Interest:

- sample from \( p = p(X, Y, Z) \) (games),
- function of interest \( \ell = \ell(X, Y, Z) \) (money) and
- \( p^* \) replacing \( p(y \mid x) \rightarrow p^*(y \mid x) \) (strategy = decisions | game state).
Idea 3: Blackjack

When can we learn?

Objects of Interest:

- sample from $p = p(X, Y, Z)$ (games),
- function of interest $\ell = \ell(X, Y, Z)$ (money) and
- $p^*$ replacing $p(y \mid x) \rightarrow p^*(y \mid x)$ (strategy = decisions | game state).

Questions:

- What is $E_{p^*}\ell$?
Idea 3: Blackjack

When can we learn?

Objects of Interest:

- Sample from \( p = p(X, Y, Z) \) (games),
- Function of interest \( \ell = \ell(X, Y, Z) \) (money) and
- \( p^* \) replacing \( p(y | x) \rightarrow p^*(y | x) \) (strategy = decisions | game state).

Questions:

- What is \( \mathbf{E}_{p^*x} \ell \)?

Needed:

- Values of \( X_i, Y_i \) and \( \ell(X_i, Y_i, Z_i) \) (under \( p \))

<table>
<thead>
<tr>
<th>( X_i )</th>
<th>( Y_i )</th>
<th>( Z_i )</th>
<th>( \ell(X_i, Y_i, Z_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>−1.4</td>
<td>2.0</td>
<td>?</td>
<td>2.1</td>
</tr>
<tr>
<td>−0.5</td>
<td>0.7</td>
<td>?</td>
<td>2.5</td>
</tr>
<tr>
<td>−0.8</td>
<td>1.5</td>
<td>?</td>
<td>2.6</td>
</tr>
</tbody>
</table>
| : | : | : | :

\[
\begin{array}{ccc|c}
\text{\( X_i \)} & \text{\( Y_i \)} & \text{\( Z_i \)} & \text{\( \ell(X_i, Y_i, Z_i) \)} \\
\diamondsuit K, \heartsuit 9 & \text{hit} & ? & -1 \\
\spadesuit A, \spadesuit J & \text{stand} & ? & 1.5 \\
\spadesuit 10, \heartsuit 8 & \text{stand} & ? & -1 \\
\end{array}
\]
Idea 3: Blackjack
Computation: Means

Assume $p(y \mid x) \to p^*(y \mid x)$.

$$\eta := \mathbf{E}_{p^*} \ell = \int \ell(x, y, z) p^*(x, y, z) \, dx \, dy \, dz$$

$$= \int \ell(x, y, z) \frac{p^*(x, y, z)}{p(x, y, z)} p(x, y, z) \, dx \, dy \, dz$$
Idea 3: Blackjack

Computation: Means

Assume $p(y \mid x) \rightarrow p^*(y \mid x)$.

$$
\eta := E_{p^*} \ell = \int \ell(x, y, z) \ p^*(x, y, z) \ dx \ dy \ dz
$$

$$
= \int \ell(x, y, z) \ \frac{p^*(x, y, z)}{p(x, y, z)} \ p(x, y, z) \ dx \ dy \ dz
$$

$$
= \int \ell(x, y, z) \ \frac{p^*(y \mid x)}{p(y \mid x)} \ p(x, y, z) \ dx \ dy \ dz
$$
Assume $p(y \mid x) \to p^*(y \mid x)$.

$$
\eta := \mathbb{E}_{p^*} \ell = \int \ell(x, y, z) \ p^*(x, y, z) \ dx \ dy \ dz
$$

$$
= \int \ell(x, y, z) \ \frac{p^*(x, y, z)}{p(x, y, z)} \ p(x, y, z) \ dx \ dy \ dz
$$

$$
= \int \ell(x, y, z) \ \frac{p^*(y \mid x)}{p(y \mid x)} \ p(x, y, z) \ dx \ dy \ dz
$$

Estimate $\eta$ by

$$
\hat{\eta} = \frac{1}{N} \sum_{i=1}^{N} \ell(X_i, Y_i, Z_i) \ \frac{p^*(Y_i \mid X_i)}{p(Y_i \mid X_i)} = \frac{1}{N} \sum_{i=1}^{N} M_i, \quad \mathbb{E}_p \hat{\eta} = \eta
$$
Idea 3: Blackjack
Computation: Means

Assume $p(y | x) \rightarrow p^*(y | x)$.

$$
\eta := E_{p^*} \ell = \int \ell(x, y, z) \ p^*(x, y, z) \ dx \ dy \ dz
$$

$$
= \int \ell(x, y, z) \ \frac{p^*(x, y, z)}{p(x, y, z)} \ p(x, y, z) \ dx \ dy \ dz
$$

$$
= \int \ell(x, y, z) \ \frac{p^*(y | x)}{p(y | x)} \ p(x, y, z) \ dx \ dy \ dz
$$

Estimate $\eta$ by

$$
\hat{\eta} = \frac{1}{N} \sum_{i=1}^{N} \ell(X_i, Y_i, Z_i) \ \underbrace{\frac{p^*(Y_i | X_i)}{p(Y_i | X_i)}}_{w_i} = \frac{1}{N} \sum_{i=1}^{N} M_i, \quad E_p \hat{\eta} = \eta
$$

Confidence intervals available!
Idea 3: Blackjack

\[ p(y \mid x) \rightarrow p^*(y \mid x) \]

Which \( p^* \) is best?
Idea 3: Blackjack

\[ p(y \mid x) \rightarrow p^*(y \mid x) \]

Which \( p^* \) is best? Parameterize and estimate

\[ \nabla_\theta E_{p_\theta} \bigg|_{\theta = \tilde{\theta}} \]
Ideas 3: Blackjack

\[ p(y \mid x) \to p^*(y \mid x) \]

Which \( p^* \) is best? Parameterize and estimate

\[ \nabla_\theta E_{p_\theta} \big|_{\theta=\tilde{\theta}} \]

Goal: Optimize \( E_{p_\theta} \ell \)
Idea: Use gradient \( \nabla_\theta E_{p_\theta} \ell \) and optimize step-by-step.
Issues: confidence intervals, step size, . . . .
Idea 3: Blackjack

How to exploit causal structure:
Idea 3: Blackjack

How to exploit causal structure:

- Deck
  - Open cards
    - Open cards (w/out suit)
      - Decision
        - Lost money
  - Hidden cards
    - Hidden cards (w/out suit)
Idea 3: Blackjack

How to exploit causal structure:

- Deck
  - Open cards
    - Open cards (w/out suit)
    - Decision
      - Lost money
  - Hidden cards
    - Hidden cards (w/out suit)
Idea 3: Blackjack

Learning Blackjack.

![Graph showing performance of different strategies in Blackjack.](image)
Idea 3: Blackjack

Learning Blackjack.

Idea 3: Blackjack

What can we do with 100,000 samples?

<table>
<thead>
<tr>
<th></th>
<th>Online</th>
<th>Offline</th>
</tr>
</thead>
<tbody>
<tr>
<td>reached strategy</td>
<td>$E_{p^* \ell} \approx -5.1Ct$</td>
<td>$E_{p^* \ell} \approx -5.8Ct$</td>
</tr>
<tr>
<td>irrelevant games</td>
<td>33,653</td>
<td>61,048</td>
</tr>
<tr>
<td>costs</td>
<td>$29,300$</td>
<td>$51,500$</td>
</tr>
<tr>
<td>speed</td>
<td>slow: probabilities</td>
<td>even slower: gradients</td>
</tr>
</tbody>
</table>
Idea 3: advertisement

```
user data

main line reserve

# ads in main line

user intention

click
```
Idea 3: advertisement

Average clicks per page vs. Mainline reserve variation

Jonas Peters (Københavns Universitet)
Idea 3: advertisement

![Diagram]

- User data
- Main line reserve
- User intention
- # ads in main line
- Click

Jonas Peters (Københavns Universitet)
Idea 3: advertisement

Old:

Average clicks per page

Mainline reserve variation

Jonas Peters (Københavns Universitet)
Idea 3: advertisement

Using discrete variable (ads shown in mainline):

Average clicks per page

Mainline reserve variation

Bottou et al.: Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, JMLR 2013

Jonas Peters (Københavns Universitet)
Optimization under constraints is possible, too
Summary Part III:

- Idea 1: anchor regression trades off prediction and invariance
- Idea 2: half-sibling regression
- Idea 3: reformulate reinforcement learning, use causal structure

Tusind tak!


Check out our jupyter notebooks: [http://web.math.ku.dk/~peters/elements.html](http://web.math.ku.dk/~peters/elements.html)
Idea 1: semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

\[ p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i | x_{pa(i)}) \]
Idea 1: semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

\[ p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)}) \]

Modularity suggests:

\[ p(x_1 \mid x_{pa(1)}), \ldots, p(x_d \mid x_{pa(d)}) \] are “independent”
Idea 1: semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

\[ p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i | x_{pa(i)}) \]

Modularity suggests:

\[ p(x_1 | x_{pa(1)}), \ldots, p(x_d | x_{pa(d)}) \text{ are "independent"} \]

Special case:

\[ p(cause), p(effect | cause) \text{ are "independent"} \]
Idea 1: semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

\[
p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i | x_{pa(i)})
\]

Modularity suggests:

\[
p(x_1 | x_{pa(1)}), \ldots, p(x_d | x_{pa(d)}) \text{ are “independent”}
\]

Special case:

\[
p(cause), p(effect | cause) \text{ are “independent”}
\]

But then: Semi-supervised Learning does not work from cause to effect.
Idea 1: semi-supervised learning

Schölkopf et al.: *On causal and anticausal learning*, ICML 2012

comparison to base classifier