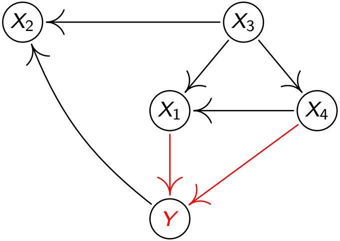
Exploiting Invariance: from Causal Discovery to Robust Decision Making



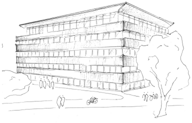
Jonas Peters, University of Copenhagen

L3S

September 2022

VILLUM FONDEN

CARLSBERGFONDET

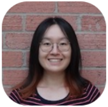


Joint work with members of the Copenhagen Causality Lab...

JEFF ADAMS

NIELS RICHARD HANSEN

LEONARD HENCKEL



SHIMENG HUANG

SNORRE JALLBJØRN

STEFFEN L. LAURITZEN

PHILLIP BREDAHL MOGENSEN

★ Mogensen & Markussen (2021)

RIKKE SØNDERGAARD NIELSEN

JONAS PETERS

★ Thams et al. (2021)

NIKLAS PFISTER

★ Saengkyongam et al. (2021)

SORAWIT SAENGKYONGAM

★ Saengkyongam et al. (2021)

NIKOLAJ THEODOR THAMS

SEBASTIAN WEICHWALD

★ Reisach et al. (2021)

... R. Christiansen and M. Jakobsen (alumni), and many other collaborators.

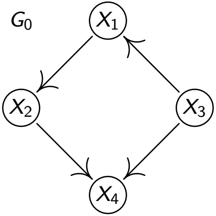
Model	Predict in i.i.d. setting	Predict under intervention	Falsification	Learnable from data
Physical model	yes	yes	obs. & int.	?
Causal model	yes	yes	obs. & int.	?
Statistical model	yes	no	obs.	yes

adapted from Peters, Janzing, Schölkopf: Elements of Causal Inference: Foundations and Learning Algorithms, 2017

SCMs (Wright 1920, Bollen 1989, Pearl 2009, Bongers et al 2021) model observational distributions over $X_1, \ldots, X_d$. Call it: P.

$$X_1 := X_3 + N_1$$
$$X_2 := 2X_1 + N_2$$
$$X_3 := N_3$$
$$X_4 := -X_2 - X_3 + N_4$$

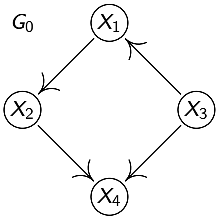
- N_i jointly independent $\mathcal{N}(0, 1)$
- G_0 has no cycles



SCMs (Wright 1920, Bollen 1989, Pearl 2009, Bongers et al 2021) model observational distributions over $X_1, \dots, X_d$. Call it: P.

$$
\begin{aligned}
X_1 &:= X_3 + N_1 \\
X_2 &:= 2X_1 + N_2 \\
X_3 &:= N_3 \\
X_4 &:= -X_2 - X_3 + N_4
\end{aligned}
$$

- N_i jointly independent $\mathcal{N}(0,1)$
- G_0 has no cycles

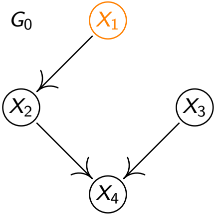


$$
\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 4 & 1 & -5 \\ 4 & 9 & 2 & -11 \\ 1 & 2 & 1 & -3 \\ -5 & -11 & -3 & 15 \end{pmatrix} \right)
$$

SCMs (Wright 1920, Bollen 1989, Pearl 2009, Bongers et al 2021) model interventions, too. Call it: $P_{do(X_1:=M)}$.

$$X_1 := M$$
$$X_2 := 2X_1 + N_2$$
$$X_3 := N_3$$
$$X_4 := -X_2 - X_3 + N_4$$

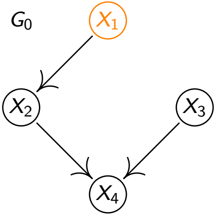
- N_i jointly independent $\mathcal{N}(0,1)$
- G_0 has no cycles



SCMs (Wright 1920, Bollen 1989, Pearl 2009, Bongers et al 2021) model interventions, too. Call it: $P_{do(X_1:=M)}$.

$$
\begin{aligned}
X_1 &:= M \\
X_2 &:= 2X_1 + N_2 \\
X_3 &:= N_3 \\
X_4 &:= -X_2 - X_3 + N_4
\end{aligned}
$$

- N_i jointly independent $\mathcal{N}(0, 1)$
- G_0 has no cycles

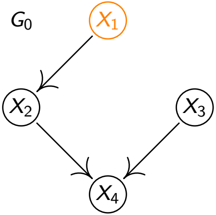


In reality, data structure may be more complex...

SCMs (Wright 1920, Bollen 1989, Pearl 2009, Bongers et al 2021) model interventions, too. Call it: $P_{do(X_1:=M)}$.

$$
\begin{aligned}
X_1 &:= M \\
X_2 &:= 2X_1 + N_2 \\
X_3 &:= N_3 \\
X_4 &:= -X_2 - X_3 + N_4
\end{aligned}
$$

- N_i jointly independent $\mathcal{N}(0,1)$
- G_0 has no cycles

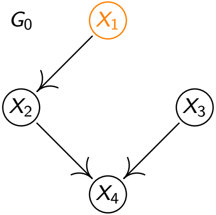


In reality, data structure may be more complex... but not today :-).

SCMs (Wright 1920, Bollen 1989, Pearl 2009, Bongers et al 2021) model interventions, too. Call it: $P_{do(X_1:=M)}$.

$$X_1 := M$$
$$X_2 := 2X_1 + N_2$$
$$X_3 := N_3$$
$$X_4 := -X_2 - X_3 + N_4$$

- N_i jointly independent $\mathcal{N}(0,1)$
- G_0 has no cycles

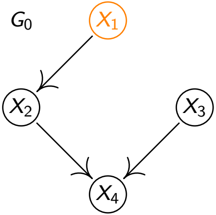


In reality, data structure may be more complex... but not today :-).
MUTE: If you intervene only on X_1, you intervene only on X_1.

SCMs (Wright 1920, Bollen 1989, Pearl 2009, Bongers et al 2021) model interventions, too. Call it: $P_{do(X_1:=M)}$.

$$X_1 := M$$
$$X_2 := 2X_1 + N_2$$
$$X_3 := N_3$$
$$X_4 := -X_2 - X_3 + N_4$$

- N_i jointly independent $\mathcal{N}(0,1)$
- G_0 has no cycles



In reality, data structure may be more complex... but not today :-).
MUTE: If you intervene only on X_1, you intervene only on X_1. Questions?

Fundamental Problem of Causal Discovery:

Fundamental Problem of Causal Discovery:

Different SCMs may induce the same observational distribution.

Fundamental Problem of Causal Discovery:

Different SCMs may induce the same observational distribution.

For example,

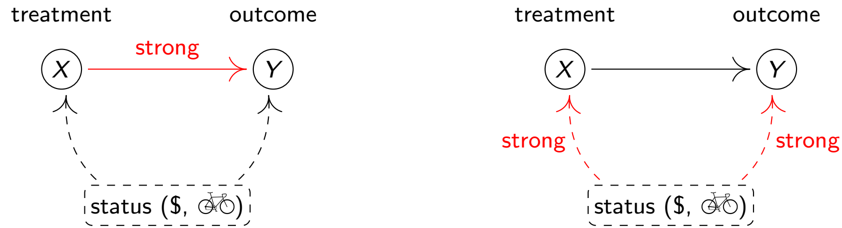
can induce the same $P_{(X,Y)}$.

Fundamental Problem of Causal Discovery:

Different SCMs may induce the same observational distribution.

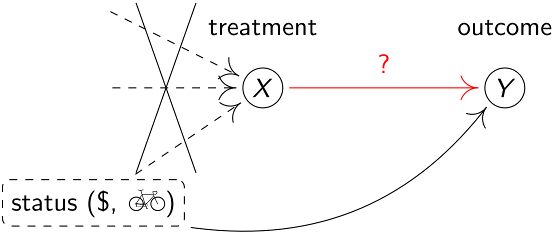
For example,

can induce the same $P_{(X,Y)}$. Similarly,

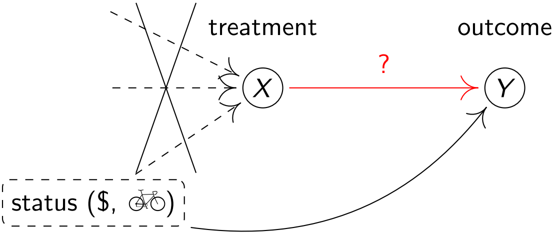


can induce the same $P_{(X,Y)}$. howtobike.info/images/CyclocrossBike.png, permission from M. Schoolfield.

Idea 1: Randomization



Idea 1: Randomization



X and Y dependent $\implies$ there is a directed causal link!

http://howtobike.info/images/CyclocrossBike.png, 14.09.2016, 3:41pm, with permission from M. Schoolfield.

Idea 2: Invariance

Pearl (2009, p. 22):

> *'that each (...) [assignment] in the network represents a stable and autonomous physical mechanism — in other words, that it is conceivable to change one such relationship without changing the others.'*

Idea 2: Invariance

Pearl (2009, p. 22):

> *'that each (...) [assignment] in the network represents a stable and autonomous physical mechanism — in other words, that it is conceivable to change one such relationship without changing the others.'*

Invariance lies at the heart of causality, ...

Idea 2: Invariance

Pearl (2009, p. 22):

> *'that each (...) [assignment] in the network represents a stable and autonomous physical mechanism — in other words, that it is conceivable to change one such relationship without changing the others.'*

Invariance lies at the heart of causality, ... and was the topic of many works and discussions in econometrics.

Haavelmo 1944, Frisch et al. 1948, Hurwicz 1962, Aldrich 1989, Hoover 2008, . . .

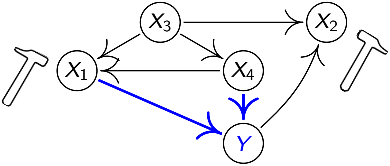
Idea 2: Invariance

Pearl (2009, p. 22):

> *'that each (...) [assignment] in the network represents a stable and autonomous physical mechanism — in other words, that it is conceivable to change one such relationship without changing the others.'*

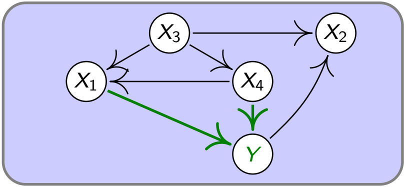
Invariance lies at the heart of causality, ... and was the topic of many works and discussions in econometrics.

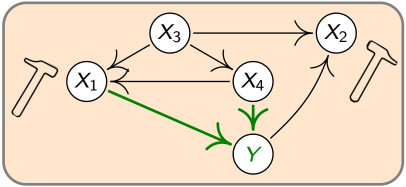
Haavelmo 1944, Frisch et al. 1948, Hurwicz 1962, Aldrich 1989, Hoover 2008, . . .



$Y \mid X_1, X_4$ is invariant

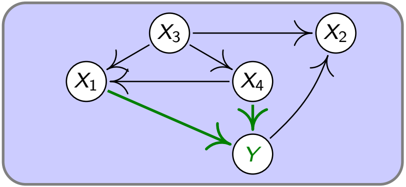
Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.

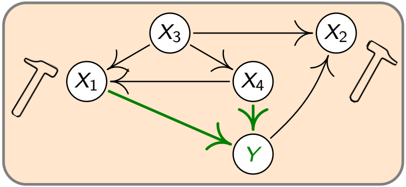




cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

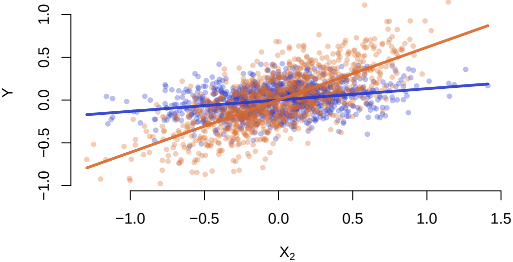
Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.



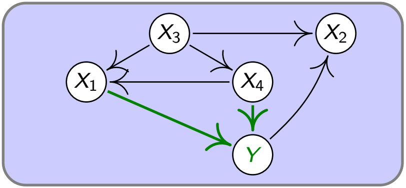


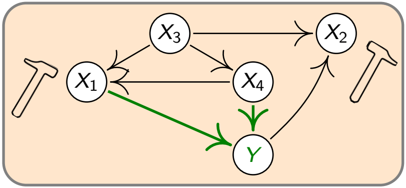
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

Not all sets of predictors yield an invariant model. Here: {2}.



Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.





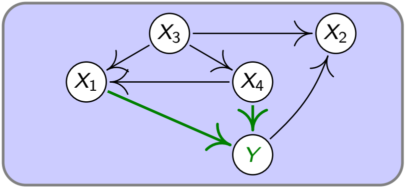
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

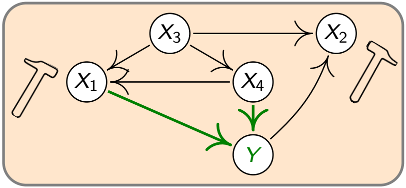
Key idea: Use and data and search for invariant models.

set	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\cdots$	$\{1,4\}$	$\{2,4\}$	$\cdots$	$\{1,3,4\}$
invariance	✗	✗	✗	✗	$\cdots$	✓	✗	$\cdots$	✓

$$\hat{S} := \bigcap_{S \text{ invariant}} S = \{1, 4\}$$

Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.





cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

Key idea: Use and data and search for invariant models.

set	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\cdots$	$\{1,4\}$	$\{2,4\}$	$\cdots$	$\{1,3,4\}$
invariance	✗	✗	✗	✗	$\cdots$	✓	✗	$\cdots$	✓

$$\hat{S} := \bigcap_{S \text{ invariant}} S = \{1,4\}$$

JP, Bühlmann, Meinshausen, JRSS-B 2016 (with discussion): $P(\hat{S} \subseteq S^*) \geq 1 - \alpha$. (ICP.ipynb)

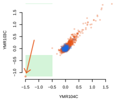
ICP (R-package `InvariantCausalPrediction`)

```
> ExpInd
```

[1]11111111111111111111111111111111...22222222222222...

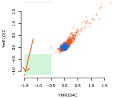
```
> icp <- ICP(X,Y,ExpInd)

            LOWER BOUND  UPPER BOUND  MAXIMIN EFFECT  P-VALUE
X1                -0.71        -0.52           -0.52  <1e-09 ***
X2                -0.46         0.00            0.00    0.55
X3                 0.58         0.70            0.58  <1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Gene data (discrete environments):



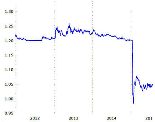
JP et al., JRSSB (with discussion) 2016

Gene data (discrete environments):



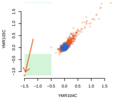
JP et al., JRSSB (with discussion) 2016

Finance data (using time):



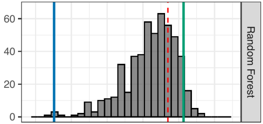
Pfister et al., JASA 2018

Gene data (discrete environments):



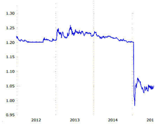
JP et al., JRSSB (with discussion) 2016

Terr. ecosystem funct. (causal GOF):



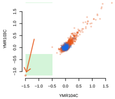
Migliavacca et al., Nature 2021

Finance data (using time):



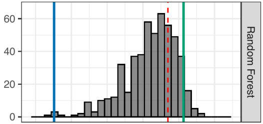
Pfister et al., JASA 2018

Gene data (discrete environments):



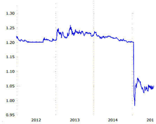
JP et al., JRSSB (with discussion) 2016

Terr. ecosystem funct. (causal GOF):



Migliavacca et al., Nature 2021

Finance data (using time):



Pfister et al., JASA 2018

Questions?

Consider the classical prediction problem...

response variable	Y
covariates	$X := X^1, \dots, X^d$
training data	i.i.d. from $P_{M_1}^{(X,Y)}, P_{M_2}^{(X,Y)}$, and $P_{M_3}^{(X,Y)}$
wanted	

Consider the classical prediction problem...

response variable	Y
covariates	$X := X^1, \dots, X^d$
training data	i.i.d. from $P_{M_1}^{(X,Y)}, P_{M_2}^{(X,Y)}$, and $P_{M_3}^{(X,Y)}$
wanted	prediction model $Y \approx \hat{f}(X)$ that performs well on test data (which may be different from training data)

domain generalization, out-of-distribution prediction, covariate shift, ...

Ben-Tal et al. 2013, Bertsimas et al. 2018, Hu and Hong 2013, Sinha et al. 2017, Magliacane et al. 2018, Lam 2019, ...

Consider the classical prediction problem...

response variable	Y
covariates	$X := X^1, \ldots, X^d$
training data	i.i.d. from $P_{M_1}^{(X,Y)}, P_{M_2}^{(X,Y)}$, and $P_{M_3}^{(X,Y)}$
wanted	prediction model $Y \approx \hat{f}(X)$ that performs well on test data (which may be different from training data)

domain generalization, out-of-distribution prediction, covariate shift, ...

Ben-Tal et al. 2013, Bertsimas et al. 2018, Hu and Hong 2013, Sinha et al. 2017, Magliacane et al. 2018, Lam 2019, ...

A causal model

Consider the classical prediction problem...

response variable	Y
covariates	$X := X^1, \ldots, X^d$
training data	i.i.d. from $P_{M_1}^{(X,Y)}, P_{M_2}^{(X,Y)}$, and $P_{M_3}^{(X,Y)}$
wanted	prediction model $Y \approx \hat{f}(X)$ that performs well on test data (which may be different from training data)

domain generalization, out-of-distribution prediction, covariate shift, ...

Ben-Tal et al. 2013, Bertsimas et al. 2018, Hu and Hong 2013, Sinha et al. 2017, Magliacane et al. 2018, Lam 2019, ...

A causal model satisfies

$$f_{causal} = \operatorname*{argmin}_{f_\diamond \in \mathcal{F}} \sup_{M \in \mathcal{M}} \mathrm{E}_M \left[\left(Y - f_\diamond(X) \right)^2 \right]$$

if M: an intervention model and $\mathcal{M}$: all interventions on X.

Consider the classical prediction problem...

response variable	Y
covariates	$X := X^1, \dots, X^d$
training data	i.i.d. from $P_{M_1}^{(X,Y)}, P_{M_2}^{(X,Y)}$, and $P_{M_3}^{(X,Y)}$
wanted	prediction model $Y \approx \hat{f}(X)$ that performs well on test data (which may be different from training data)

domain generalization, out-of-distribution prediction, covariate shift, ...

Ben-Tal et al. 2013, Bertsimas et al. 2018, Hu and Hong 2013, Sinha et al. 2017, Magliacane et al. 2018, Lam 2019, ...

A causal model satisfies

$$f_{causal} = \operatorname*{argmin}_{f_\diamond \in \mathcal{F}} \sup_{M \in \mathcal{M}} \mathrm{E}_M \left[\left(Y - f_\diamond(X) \right)^2 \right]$$

if M: an intervention model and $\mathcal{M}$: all interventions on X.

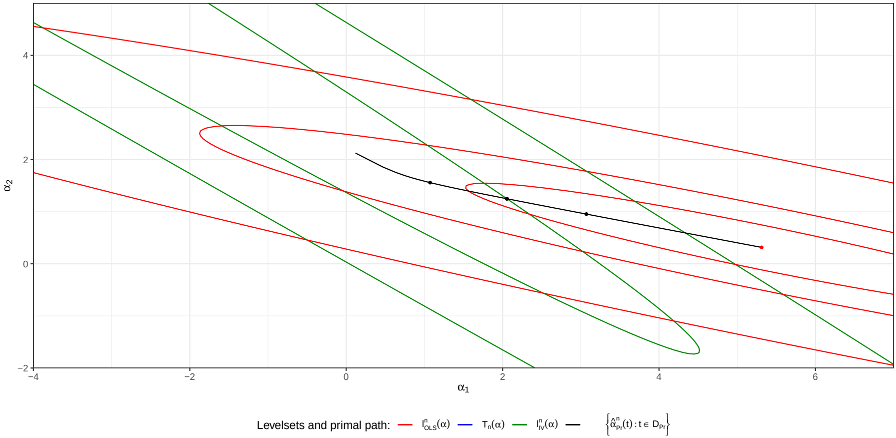
More theory: Christiansen, Pfister, Jakobsen, Gnecco, JP: TPAMI 2021

Idea: Among all invariant models, choose the most predictive one.

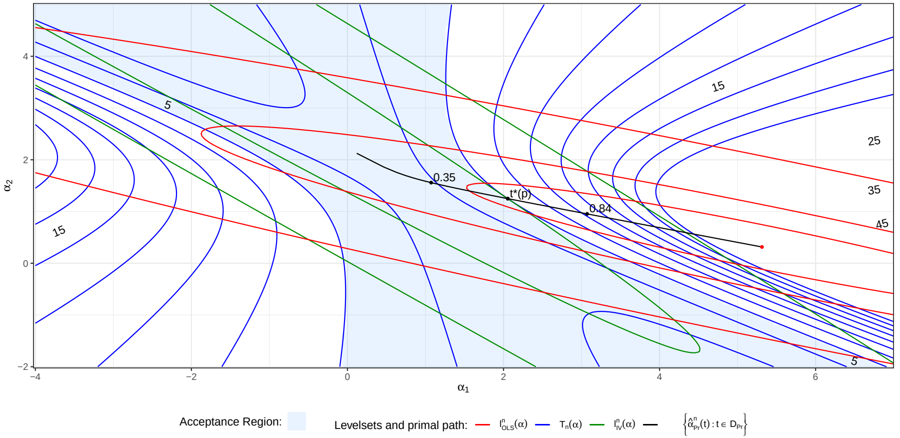
Idea: Among all invariant models, choose the most predictive one.

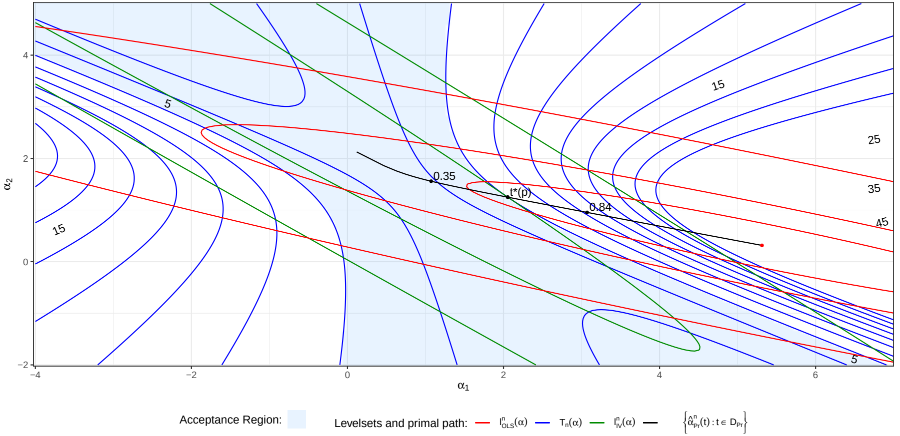
$$\alpha^\gamma := \operatorname*{argmin}_\alpha \underbrace{\mathrm{E}(Y - X\alpha)^2}_{\text{prediction}} \qquad \text{s.t.} \quad \underbrace{\|\mathrm{E}A^\top (Y - X\alpha)\|_2^2}_{\text{invariance}} \leq \gamma$$

$$\hat{\alpha}_n^\gamma := \operatorname*{argmin}_\alpha (\mathrm{Y} - \mathrm{X}\alpha)^\top (\mathrm{Y} - \mathrm{X}\alpha) \text{ s.t. } (\mathrm{Y} - \mathrm{X}\alpha)^\top \mathrm{A}(\mathrm{A}^\top \mathrm{A})^{-1} \mathrm{A}^\top (\mathrm{Y} - \mathrm{X}\alpha) \leq \gamma$$



4
2
0
−2
−4
−2
0
2
4
6
α_1
α_2
Levelsets and primal path: $l^n_{OLS}(\alpha)$ $T_n(\alpha)$ $l^n_{IV}(\alpha)$ $\left\{\hat{\alpha}^n_{Pr}(t) : t \in D_{Pr}\right\}$


α₂
α₁
0.35
t*(p)
0.84
5
15
15
25
35
45
5
Acceptance Region:
Levelsets and primal path:
$l^n_{OLS}(\alpha)$
$T_n(\alpha)$
$l^n_{IV}(\alpha)$
$\{\hat{\alpha}^n_{Pr}(t) : t \in D_{Pr}\}$



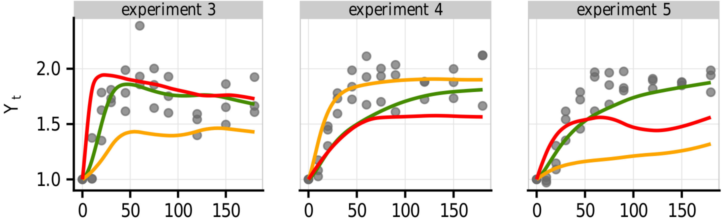
Jakobsen and JP: Distributional Robustness of K-class Estimators and the PULSE, The Econometrics Journal 2021

Rothenhäusler, Bühlmann, Meinshausen, JP, JRSSB, 2021

e.g., Anderson and Rubin 1949 and Theil 1958 and Fuller 1977

An application to metabolic networks (CausalKinetiX): Pfister, Bauer, JP, PNAS, 2019

top ranked model $\dot{Y}_t = \theta_1 Z_t X_t^{56} X_t^{122} + \theta_2 Z_t X_t^{128} X_t^{168} - \theta_3 Y_t X_t^{33} X_t^{138}$



but also

Rojas-Carulla et al. 2018, Arjovsky et al. 2019, Christiansen et al. 2021, Guo et al. 2021, Oberst et al. 2021, Pfister et al. 2021,...

Top ranked variables:

rank	held-out-experiment				
	1	2	3	4	5
1	X^{33}	X^{33}	X^{33}	X^{33}	X^{33}
2	X^{56}	X^{38}	X^{73}	X^{38}	X^{56}
3	X^{122}	X^{61}	X^{122}	X^{128}	X^{122}
4	X^{128}	X^{128}	X^{138}	X^{168}	X^{128}
5	X^{138}	X^{138}	X^{168}	X^{246}	X^{138}
6	X^{168}	X^{168}	X^{215}	X^{61}	X^{168}

Thanks to Robbie Loewith, Enric Montanana Sayas, Brendan Ryback, Uwe Sauer, and Jörg Stelling.

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \ ?$$

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \ ?$$

1. For each repetition i, smooth response trajectory. $\rightsquigarrow \hat{y}^{(i)}$

How invariant is

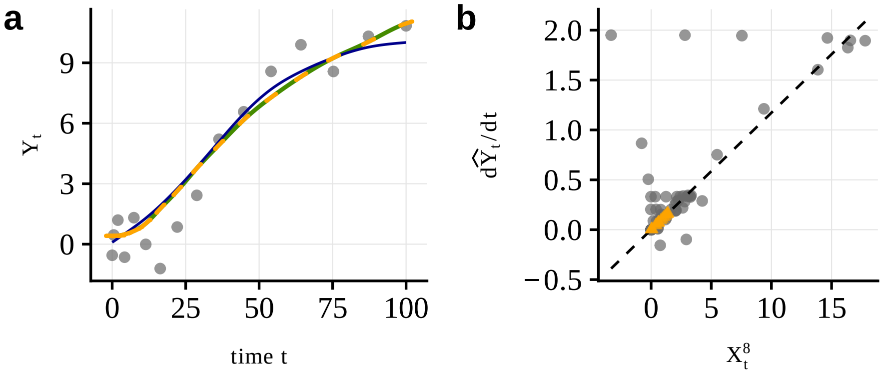
$$\dot{Y}_t = \theta_1 X_t^8 \ ?$$

1. For each repetition i, smooth response trajectory. $\rightsquigarrow \hat{y}^{(i)}$
2. Obtain fitted values for response derivatives (other exp.). $\rightsquigarrow$ $\xi_{t_1}^{(i)}, \ldots, \xi_{t_m}^{(i)}$

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \text{ ?}$$

1. For each repetition i, smooth response trajectory. $\rightsquigarrow \hat{y}^{(i)}$
2. Obtain fitted values for response derivatives (other exp.). $\rightsquigarrow$ $\xi_{t_1}^{(i)}, \ldots, \xi_{t_m}^{(i)}$
3. Smooth response trajectory, with constraints on derivatives. $\rightsquigarrow \hat{y}^{(i)}$



N. Pfister, S. Bauer, JP: *Learning stable structures in kinetic systems*, arXiv:1810.11776, PNAS 2019

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \text{ ?}$$

1. For each repetition i, smooth response trajectory. $\rightsquigarrow \hat{y}^{(i)}$
2. Obtain fitted values for response derivatives (other exp.). $\rightsquigarrow$ $\xi_{t_1}^{(i)}, \ldots, \xi_{t_m}^{(i)}$
3. Smooth response trajectory, with constraints on derivatives. $\rightsquigarrow \hat{y}^{(i)}$

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \ ?$$

1. For each repetition i, smooth response trajectory. $\rightsquigarrow \hat{y}^{(i)}$
2. Obtain fitted values for response derivatives (other exp.). $\rightsquigarrow$ $\xi_{t_1}^{(i)}, \dots, \xi_{t_m}^{(i)}$
3. Smooth response trajectory, with constraints on derivatives. $\rightsquigarrow \hat{y}^{(i)}$
4. Score for model ranking

$$\sum_{i=1}^{n} \left[\mathrm{RSS}^{(i)} - \mathrm{RSS}^{(i)}\right] / \left[\mathrm{RSS}^{(i)}\right],$$

where $\mathrm{RSS}^{(i)} := \sum_\ell \left(\hat{y}_{t_\ell}^{(i)} - Y_{t_\ell}^{(i)}\right)^2$.

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \ ?$$

1. For each repetition i, smooth response trajectory. $\rightsquigarrow \hat{y}^{(i)}$
2. Obtain fitted values for response derivatives (other exp.). $\rightsquigarrow$ $\xi_{t_1}^{(i)}, \ldots, \xi_{t_m}^{(i)}$
3. Smooth response trajectory, with constraints on derivatives. $\rightsquigarrow \hat{y}^{(i)}$
4. Score for model ranking

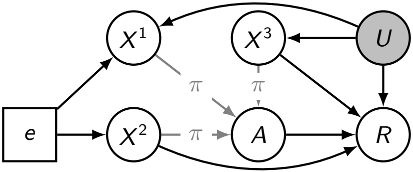
$$\sum_{i=1}^{n} \left[\mathrm{RSS}^{(i)} - \mathrm{RSS}^{(i)}\right] / \left[\mathrm{RSS}^{(i)}\right],$$

   where $\mathrm{RSS}^{(i)} := \sum_\ell \left(\hat{y}_{t_\ell}^{(i)} - Y_{t_\ell}^{(i)}\right)^2$.
5. Turn the score for models into a score for variables/complexes.

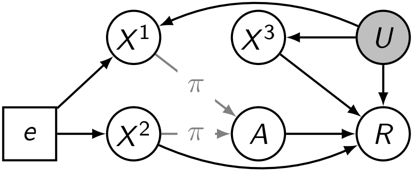
Both of the above are about prediction. What about decision making?

Both of the above are about prediction. What about decision making?

Invariant Policy Learning (cont. bandits): Saengkyongam, Thams, JP, Pfister, arXiv:2106.00808, 2021

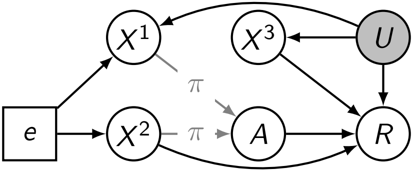


see also Song et al. 2019, Zhang et al. 2020, decision-theoretic approach by Dawid, ...



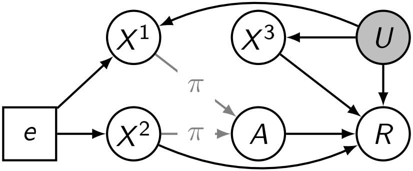
A policy π is *invariant w.r.t. S* if π depends only on S and if $\forall e, f \in \mathcal{E}^{obs}, x$

$$\mathrm{E}^{\pi,e}\left[R \mid X^S = x\right] = \mathrm{E}^{\pi,f}\left[R \mid X^S = x\right]$$



A policy π is *invariant w.r.t. S* if π depends only on S and if $\forall e, f \in \mathcal{E}^{obs}, x$

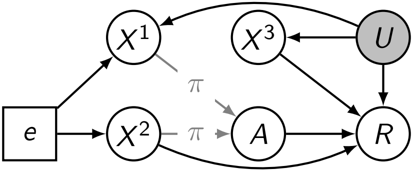
$$\mathrm{E}^{\pi,e}\left[R \mid X^S = x\right] = \mathrm{E}^{\pi,f}\left[R \mid X^S = x\right] \qquad \text{(Think: } e \perp\!\!\!\perp_{\pi} R \,|\, X^S\text{)}$$



A policy π is *invariant w.r.t.* S if π depends only on S and if $\forall e, f \in \mathcal{E}^{obs}, x$

$$\mathsf{E}^{\pi,e}\left[R \mid X^S = x\right] = \mathsf{E}^{\pi,f}\left[R \mid X^S = x\right] \qquad \text{(Think: } e \perp\!\!\!\perp_\pi R \,|\, X^S)$$

Here: $e \perp\!\!\!\perp_\pi R \,|\, X^{2,3}$



A policy π is *invariant w.r.t. S* if π depends only on S and if $\forall e, f \in \mathcal{E}^{obs}, x$

$$\mathrm{E}^{\pi,e}\left[R \mid X^S = x\right] = \mathrm{E}^{\pi,f}\left[R \mid X^S = x\right] \qquad (\text{Think: } e \perp\!\!\!\perp_{\pi} R \mid X^S)$$

Here: $e \perp\!\!\!\perp_{\pi} R \mid X^{2,3}$

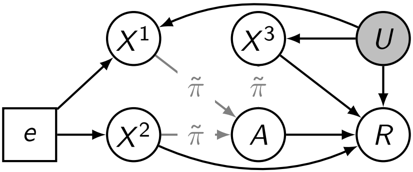
Theorem (Saengkyongam et al., 2021)

Assume 'e → hidden conf. and strong environm.'. Let $\mathcal{E}^{obs} \subset \mathcal{E}$. Then, for

$$\pi^* \in \underset{\pi \ inv}{\operatorname{argmax}} \sum_{e \in \mathcal{E}^{obs}} \mathrm{E}^{\pi,e}[R] \quad \textit{and for all } \pi$$

we have

$$\inf_{e \in \mathcal{E}} E^{\pi^*,e}[R] \geq \inf_{e \in \mathcal{E}} E^{\pi,e}[R].$$



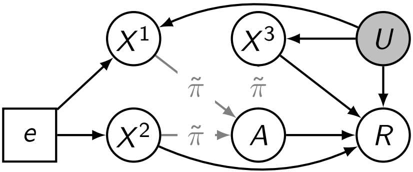
A policy π is *invariant w.r.t. S* if π depends only on S and if $\forall e, f \in \mathcal{E}^{obs}, x$

$$\mathrm{E}^{\pi,e}\left[R \mid X^S = x\right] = \mathrm{E}^{\pi,f}\left[R \mid X^S = x\right] \qquad \text{(Think: } e \perp\!\!\!\perp_{\pi} R \,|\, X^S\text{)}$$

Here: $e \perp\!\!\!\perp_{\pi} R \,|\, X^{2,3}$

Here: $e \not\perp\!\!\!\perp_{\tilde{\pi}} R \,|\, X^{1,2,3}$

Here: $e \not\perp\!\!\!\perp_{\tilde{\pi}} R \,|\, X^{3}$



A policy π is *invariant w.r.t.* S if π depends only on S and if $\forall e, f \in \mathcal{E}^{obs}, x$

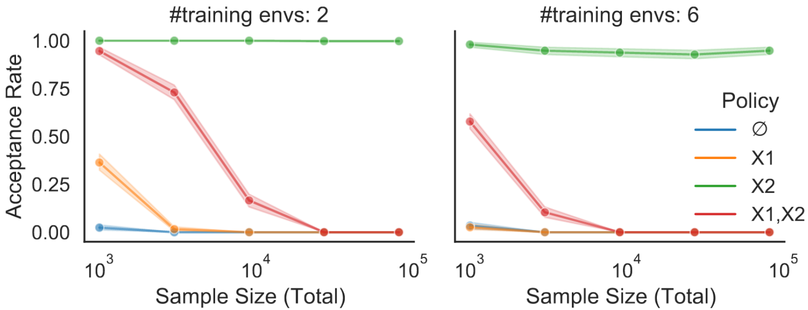
$$\mathrm{E}^{\pi,e}\left[R \mid X^S = x\right] = \mathrm{E}^{\pi,f}\left[R \mid X^S = x\right] \qquad \text{(Think: } e \perp\!\!\!\perp_{\pi} R \,|\, X^S\text{)}$$

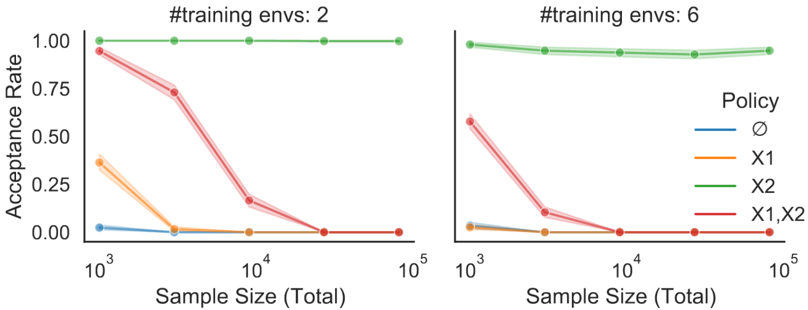
Here: $e \perp\!\!\!\perp_{\pi} R \,|\, X^{2,3}$

Here: $e \not\perp\!\!\!\perp_{\tilde{\pi}} R \,|\, X^{1,2,3}$

Here: $e \not\perp\!\!\!\perp_{\tilde{\pi}} R \,|\, X^{3}$

Requires: statistical testing under distributional shifts! (Thams et al. 2021)


#training envs: 2
#training envs: 6
Acceptance Rate
1.00
0.75
0.50
0.25
0.00
10^3
10^4
10^5
Sample Size (Total)
Policy
∅
X1
X2
X1,X2



Real data:

Warfarin (blood thinner), 5700 patients, 21 research groups

A: dosage

X: patient features

π^{tr}: based only on BMI

Summary

- Causal models are required for making good decisions.

Summary

- Causal models are required for making good decisions.
- Learning causal models from data...
  a) Idea 1: Randomization
  b) Idea 2: Invariance

Summary

- Causal models are required for making good decisions.
- Learning causal models from data...
  a) Idea 1: Randomization
  b) Idea 2: Invariance
- Invariance may help for finding distributional robust models in ...
  a) linear models (Anchor Regression, PULSE)
  b) dynamical systems (Causal KinetiX)
  c) policy adaptation (contextual bandits)

Summary

- Causal models are required for making good decisions.
- Learning causal models from data...
  a) Idea 1: Randomization
  b) Idea 2: Invariance
- Invariance may help for finding distributional robust models in ...
  a) linear models (Anchor Regression, PULSE)
  b) dynamical systems (Causal KinetiX)
  c) policy adaptation (contextual bandits)

Book: JP, D. Janzing, B. Schölkopf: *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press 2017

M. Jakobsen, JP: *Distributional Robustness of K-class Estimators and the PULSE*, The Econometrics Journal 2021

D. Rothenhäusler, P. Bühlmann, N. Meinshausen, JP: *Anchor regression: heterogeneous data meets causality*, JRSSB 2021

S. Saengkyongam, N. Thams, JP, N. Pfister: *Invariant Policy Learning: A Causal Perspective*, arXiv:2106.00808, 2021

R. Christiansen, N. Pfister, M. Jakobsen, N. Gnecco, JP: *A causal framework for distribution generalization*, IEEE TPAMI 2021

N. Thams, S. Saengkyongam, N. Pfister, JP: *Statistical Testing under Distributional Shift*, arXiv:2105.10821, 2021