Causality, Invariance, and Distribution Generalization [1]

Sorawit Saengkyongam

ETH Zürich, Seminar for Statistics

[1] Some of the slides are provided by Jonas Peters.

Table of Contents

Table of Contents

Invariance

Setting: Multiple environments $\mathcal{E} = \{e_1, \ldots, e_k\}$

$(X^{e_1}, Y^{e_1}) \sim \mathbb{P}^{e_1}$

$(X^{e_2}, Y^{e_2}) \sim \mathbb{P}^{e_2}$

$(X^{e_3}, Y^{e_3}) \sim \mathbb{P}^{e_3}$

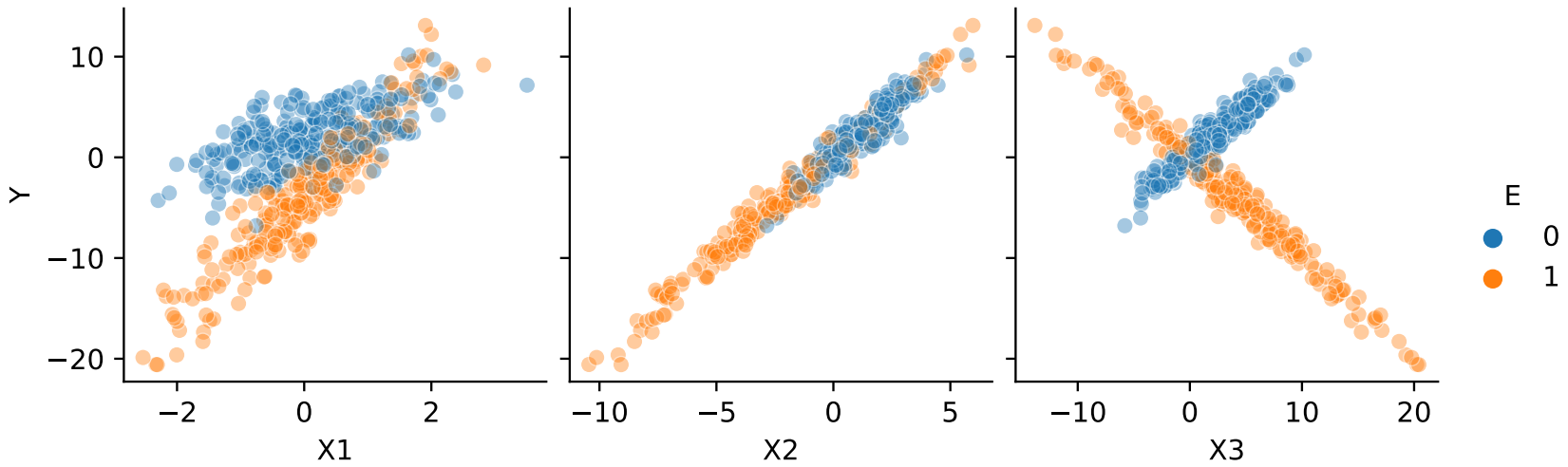
Invariance

Setting: Multiple environments $\mathcal{E} = \{e_1, \ldots, e_k\}$

$(X^{e_1}, Y^{e_1}) \sim \mathbb{P}^{e_1}$ $(X^{e_2}, Y^{e_2}) \sim \mathbb{P}^{e_2}$ $(X^{e_3}, Y^{e_3}) \sim \mathbb{P}^{e_3}$

Invariance: $\exists S \subseteq \{1, \ldots, d\} : \mathbb{P}(Y^e \mid X_S^e)$ are identical for all $e \in \mathcal{E}$

Invariance: Example



Here: $\mathbb{P}(Y^0 \mid X_2^0) = \mathbb{P}(Y^1 \mid X_2^1)$. We call X^2 is invariant.

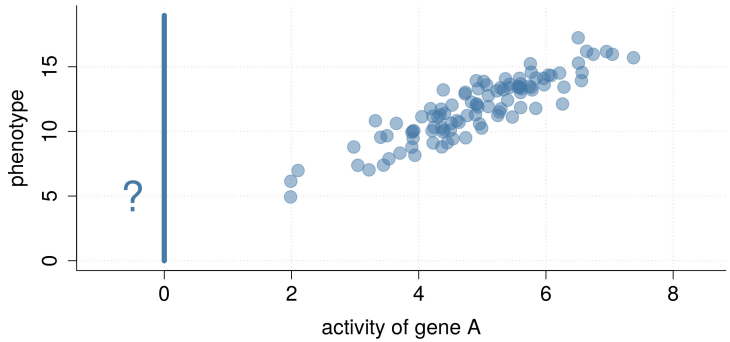
Table of Contents

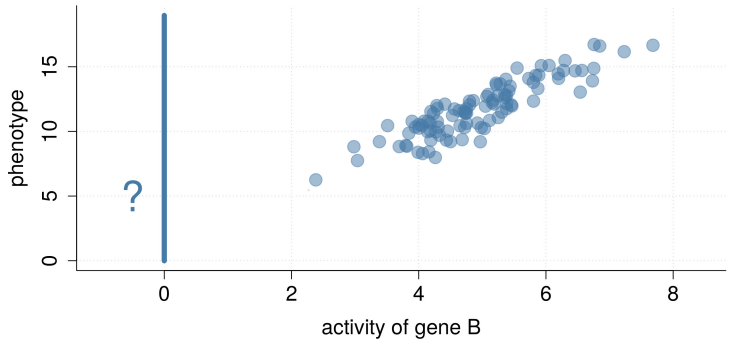
From Invariance to Causality

$$\text{Invariance} \implies \text{Causality}$$

Why does causality matter?

What is the best prediction for the phenotype if we delete gene A (or B)?

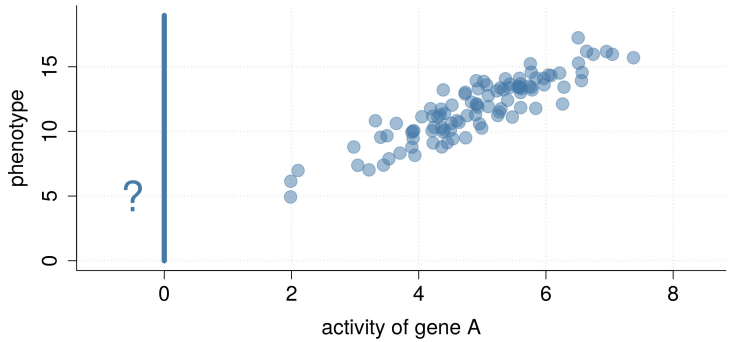


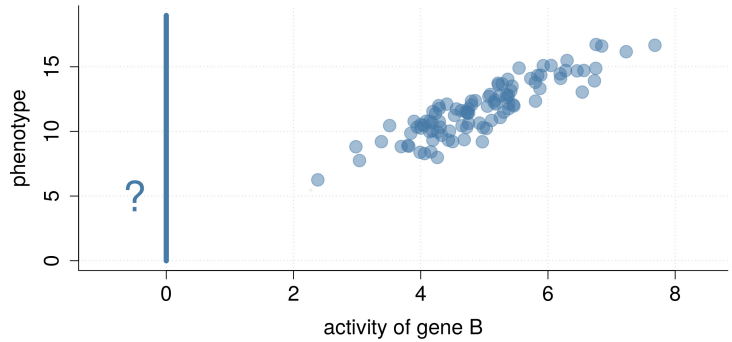


Credit: https://people.math.ethz.ch/~jopeters/

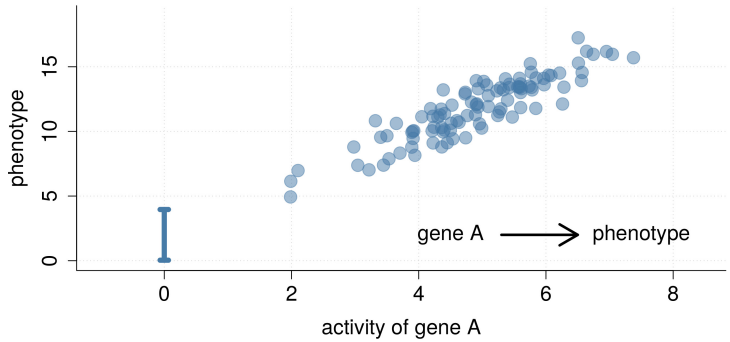
Why does causality matter?

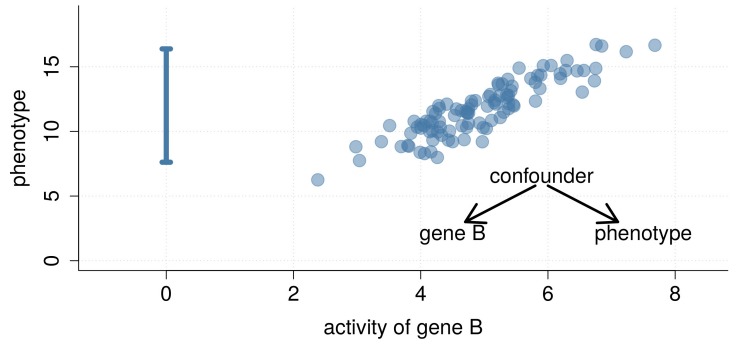
What is the best prediction for the phenotype if we delete gene A (or B)?





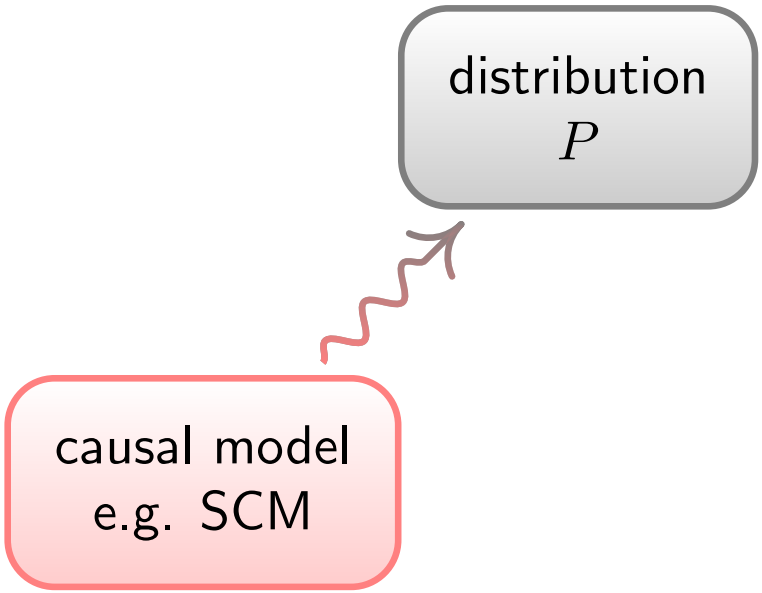
Causality matters: the optimal prediction should depend on the underlying causal structure



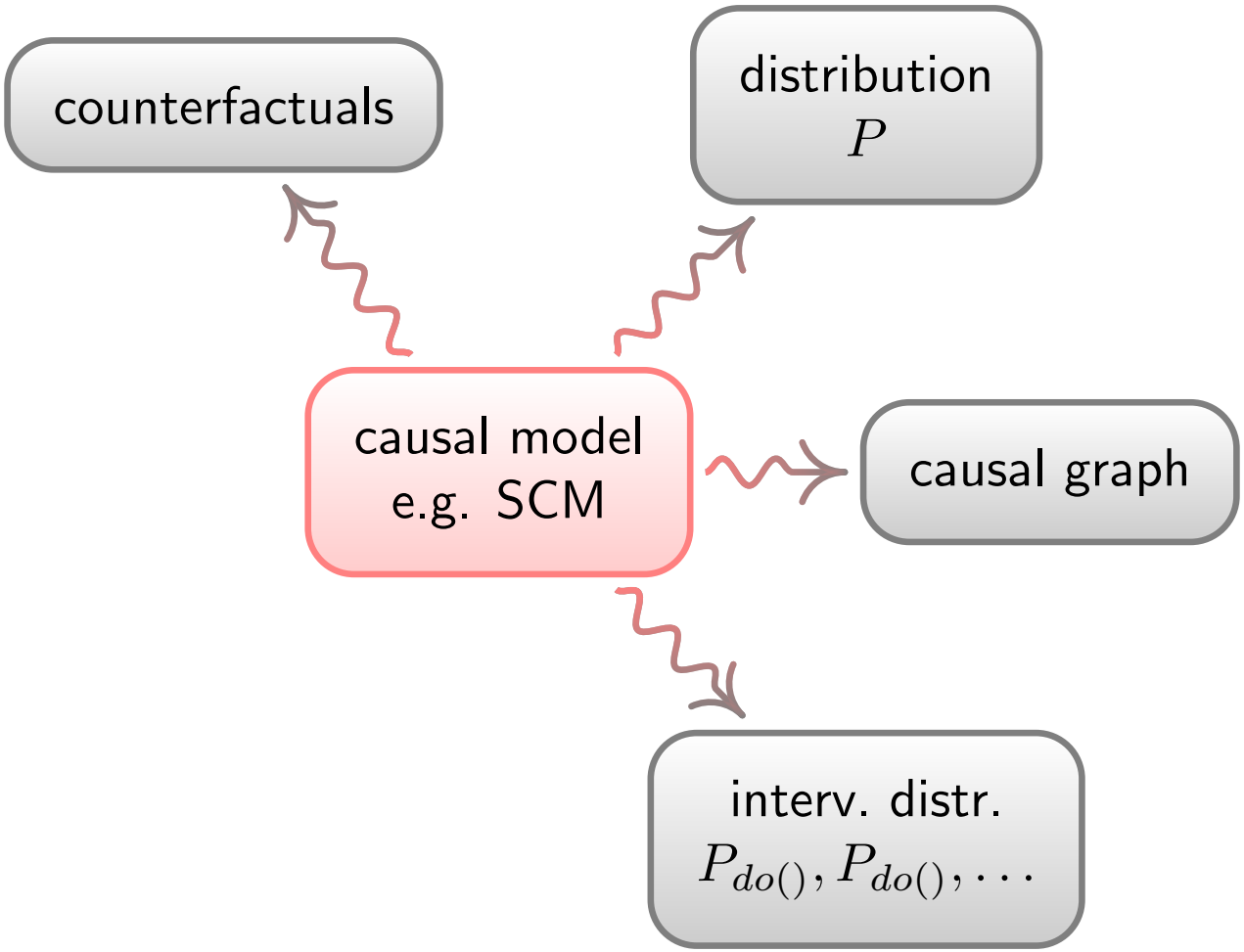


Credit: https://people.math.ethz.ch/~jopeters/

What is a causal model?



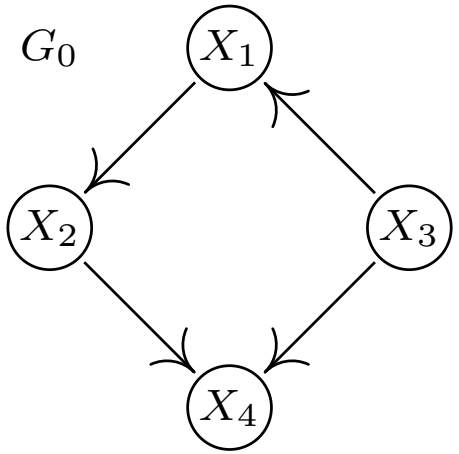
What is a causal model?



SCMs model observational distributions over $X_1, \ldots, X_d$. Call it: P.

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

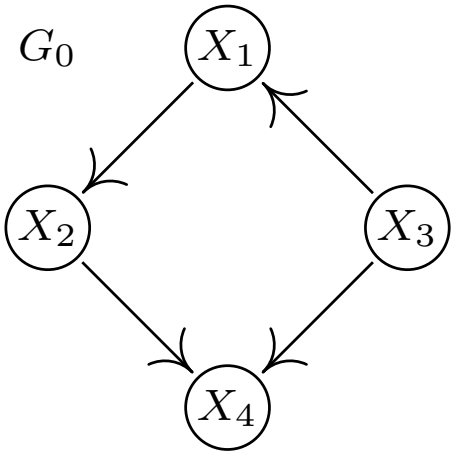
- N_i jointly independent $\mathcal{N}(0, 1)$
- G_0 has no cycles



SCMs model observational distributions over $X_1, \ldots, X_d$. 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

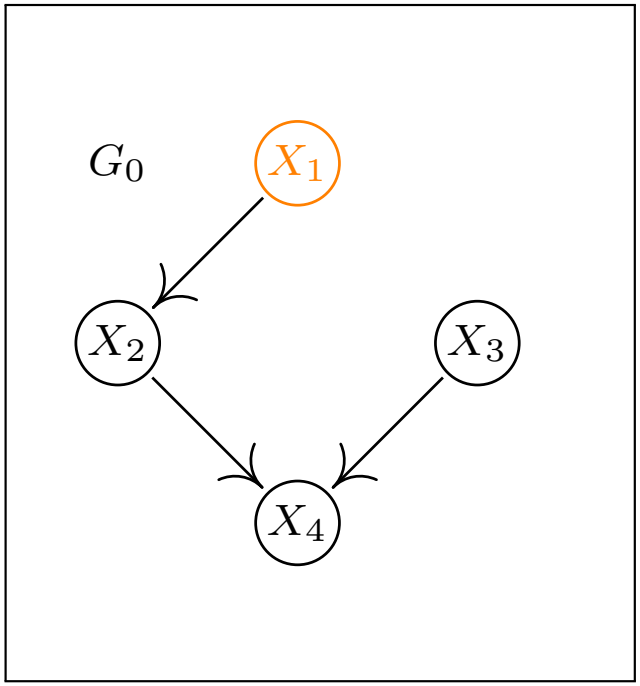


$$
\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 4 & 1 & -5 \\ 4 & 9 & 2 & -11 \\ 1 & 2 & 1 & -3 \\ -5 & -11 & -3 & 15 \end{pmatrix} \right)
$$

SCMs model interventions, too. Call it: $P_{do(X_1:=0)}$.

$$X_1 := 0$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := f_4(X_2, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles



Modularity Property

If you intervene only on X_j, you intervene only on X_j.

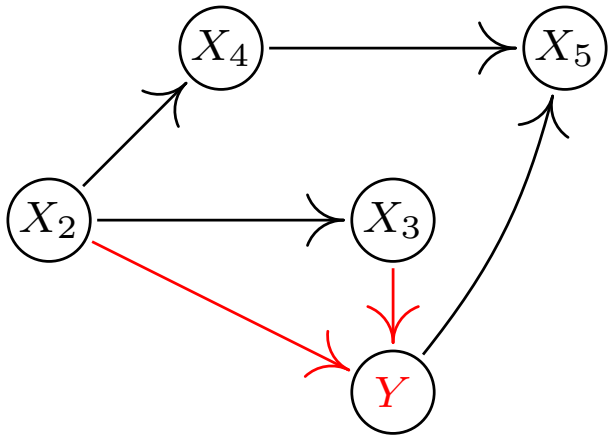
From Invariance to Causality

$$\text{Invariance} \implies \text{Causality}$$

The Problem of Causal Discovery:

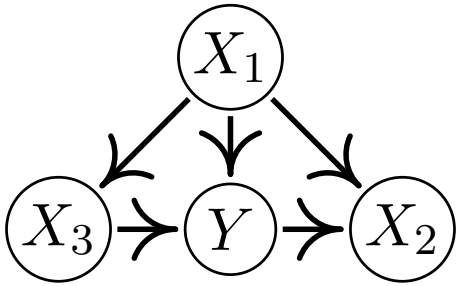
observed data $\xrightarrow{?}$ causal model

Y	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮

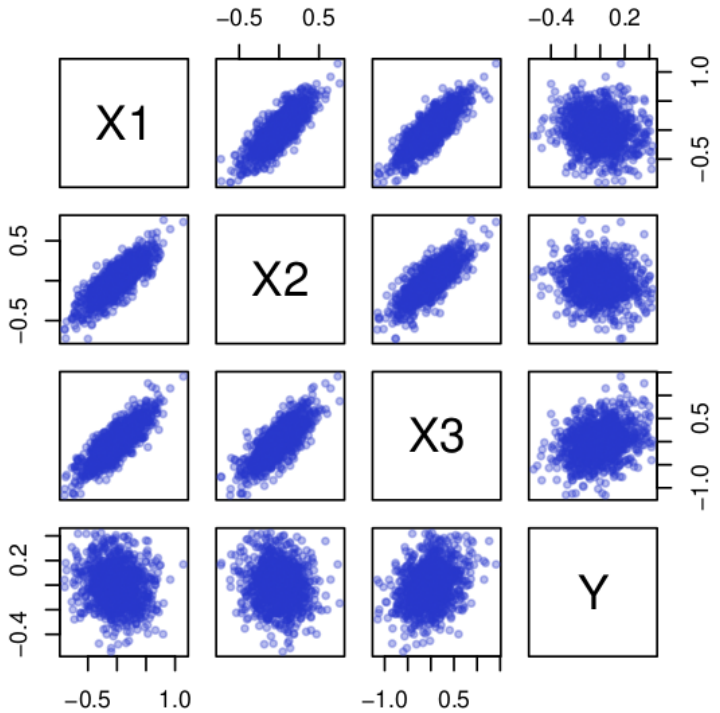


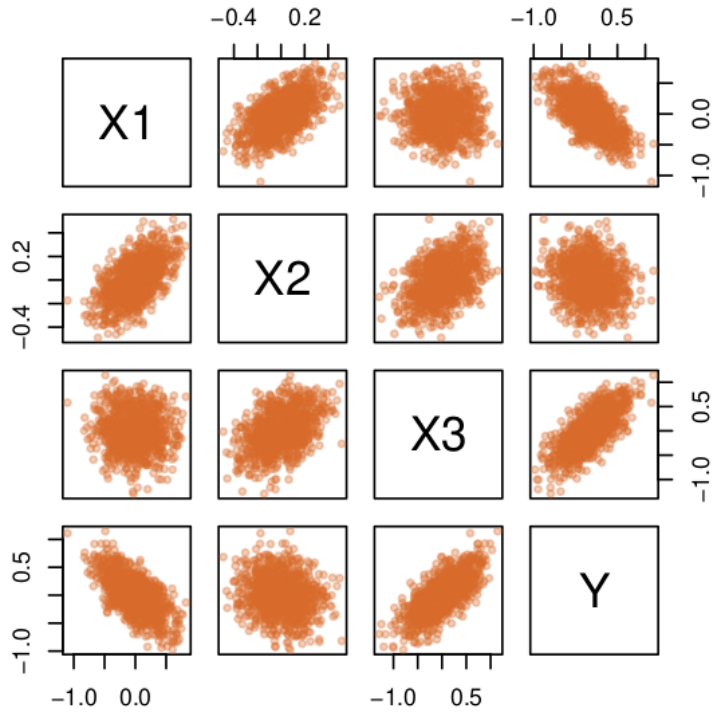
Here: Find the direct causes of Y!

unknown:



known:





linear model

```
> linmod <- lm( Y ~ X)
> summary(linmod)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.305e-05  2.067e-03    0.04    0.968
X1             -5.490e-01  9.725e-03  -56.46   <2e-16 ***
X2             -4.078e-01  1.810e-02  -22.52   <2e-16 ***
X3              6.821e-01  6.896e-03   98.91   <2e-16 ***
```

ICP (R-package InvariantCausalPrediction)

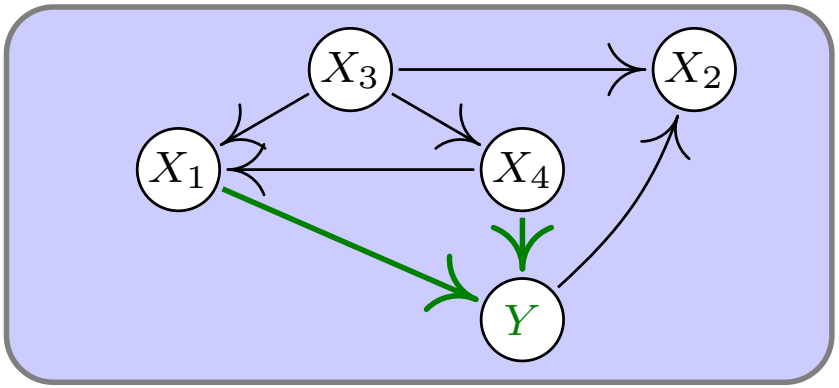
```
> ExpInd
```

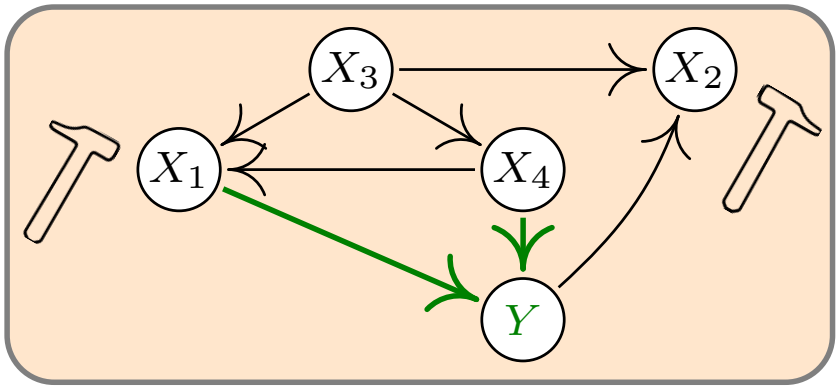
[1]111111111111111111111111111111...2222222222222...

```
> 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
```

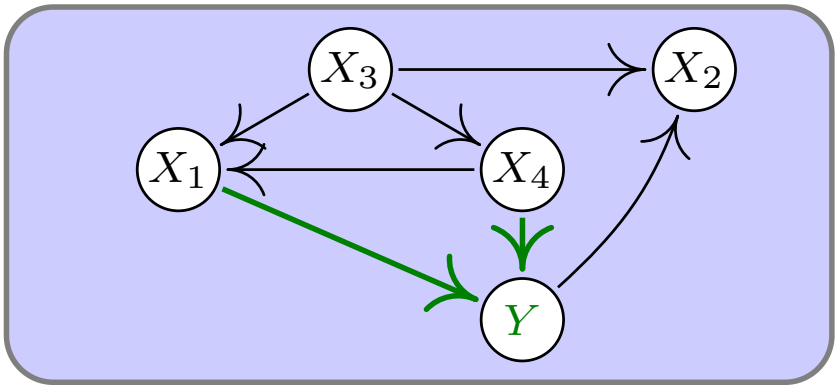
Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.

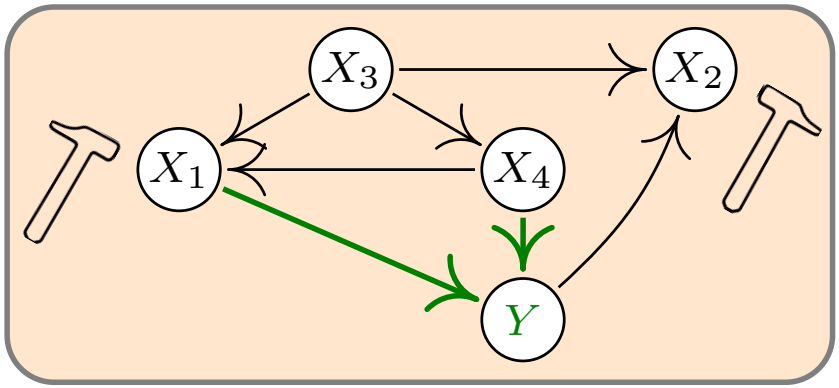




cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

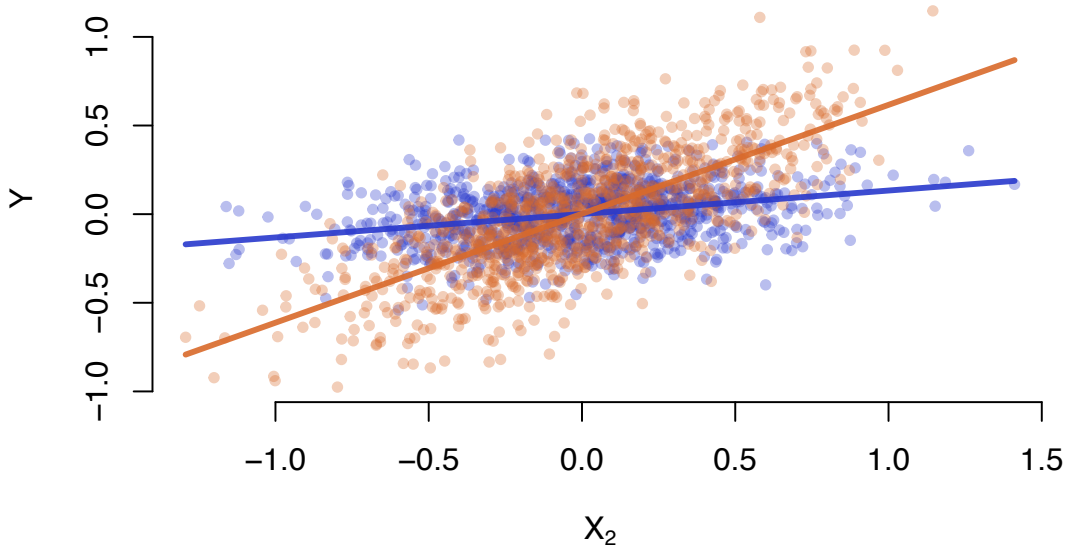
Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.



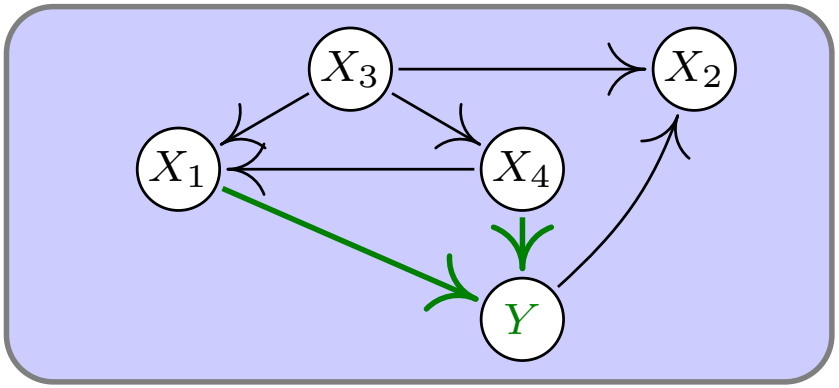


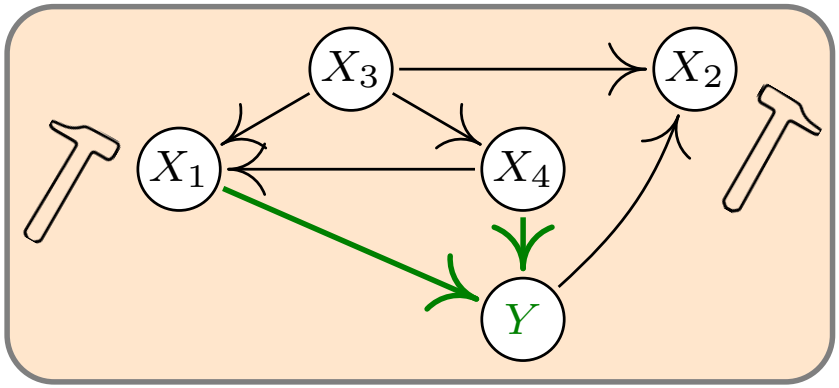
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

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



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





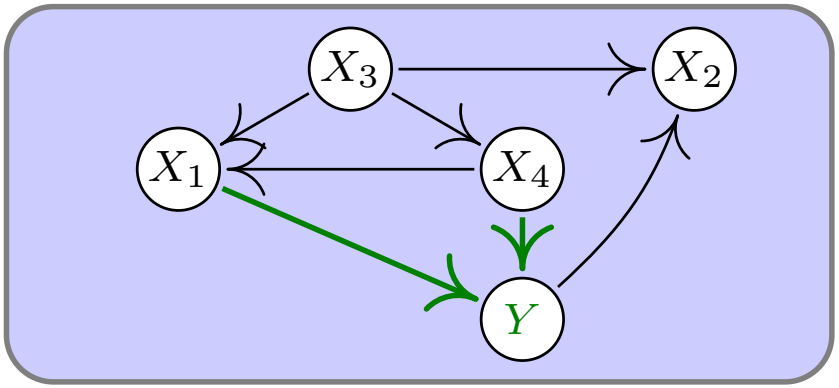
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

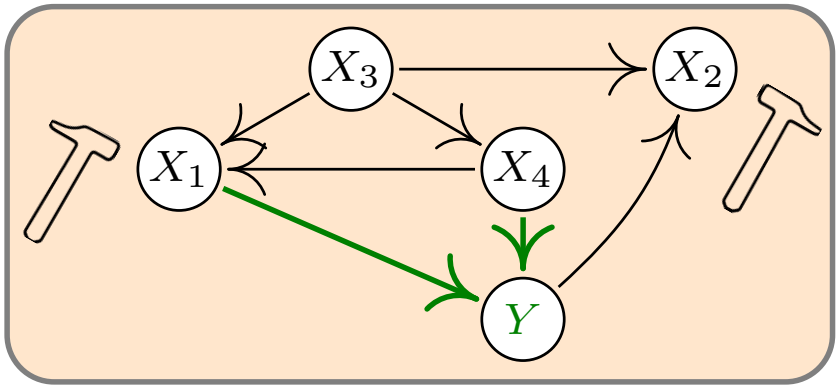
Key idea: Use ▭ and ▭ data and search for invariant models.

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

Invariant Causal Prediction

Setting: Multiple environments $\mathcal{E} = \{e_1, \ldots, e_k\}$
Given: For each $e \in \mathcal{E}$, an i.i.d. sample of (X^e, Y^e).

Invariant Causal Prediction

Setting: Multiple environments $\mathcal{E} = \{e_1, \ldots, e_k\}$
Given: For each $e \in \mathcal{E}$, an i.i.d. sample of (X^e, Y^e).

Invariance $H_{0,S}$:

- The conditional distributions $\mathbb{P}(Y^e \mid X_S^e)$ are identical for all $e \in \mathcal{E}$.
- (Linear setting) $Y^e = X_S^e \gamma + \epsilon^e$, where $\epsilon^e \perp\!\!\!\perp X^e$ and $\epsilon^e \sim F$
- (Non-linear setting) $Y \perp\!\!\!\perp E \mid X_S$

Invariant Causal Prediction

Setting: Multiple environments $\mathcal{E} = \{e_1, \ldots, e_k\}$
Given: For each $e \in \mathcal{E}$, an i.i.d. sample of (X^e, Y^e).

Invariance $H_{0,S}$:

- The conditional distributions $\mathbb{P}(Y^e \mid X_S^e)$ are identical for all $e \in \mathcal{E}$.
- (Linear setting) $Y^e = X_S^e \gamma + \epsilon^e$, where $\epsilon^e \perp\!\!\!\perp X^e$ and $\epsilon^e \sim F$
- (Non-linear setting) $Y \perp\!\!\!\perp E \mid X_S$

Environments: Different interventions (not on Y). Then, H_{0,PA_Y} holds.

Invariant Causal Prediction

Setting: Multiple environments $\mathcal{E} = \{e_1, \ldots, e_k\}$
Given: For each $e \in \mathcal{E}$, an i.i.d. sample of (X^e, Y^e).

Invariance $H_{0,S}$:

- The conditional distributions $\mathbb{P}(Y^e \mid X_S^e)$ are identical for all $e \in \mathcal{E}$.
- (Linear setting) $Y^e = X_S^e \gamma + \epsilon^e$, where $\epsilon^e \perp\!\!\!\perp X^e$ and $\epsilon^e \sim F$
- (Non-linear setting) $Y \perp\!\!\!\perp E \mid X_S$

Environments: Different interventions (not on Y). Then, H_{0,PA_Y} holds.

Theorem (Peters et al. 2016)

Assume H_{0,PA_Y} holds. For any test level α we obtain

$$P(\hat{S} \subseteq \mathrm{PA}_Y) \geq 1 - \alpha$$

The guarantee holds without identifiability assumptions (if PA_Y is not identifiable, then ICP returns an empty set).

Table of Contents

From Causality to Distribution Generalization

$$\text{Causality} \implies \text{Invariance} \implies \text{Generalization}$$

response variable	Y
covariates	$X := X_1, \ldots, X_d$
training model	M
training data	i.i.d. from $P_M^{(X,Y)}$
wanted	

response variable	Y
covariates	$X := X_1, \ldots, X_d$
training model	M
training data	i.i.d. from $P_M^{(X,Y)}$
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, ...

response variable	Y
covariates	$X := X_1, \ldots, X_d$
training model	M
training data	i.i.d. from $P_M^{(X,Y)}$
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, ...

minimax solution: $$\operatorname*{argmin}_{f_\diamond \in \mathcal{F}} \sup_{\tilde{M} \in \mathcal{N}(M)} E_{\tilde{M}}\big[\big(Y - f_\diamond(X)\big)^2\big].$$

response variable	Y
covariates	$X := X_1, \ldots, X_d$
training model	M
training data	i.i.d. from $P_M^{(X,Y)}$
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, ...

minimax solution: $$\operatorname*{argmin}_{f_\diamond \in \mathcal{F}} \sup_{\tilde{M} \in \mathcal{N}(M)} E_{\tilde{M}}\left[\left(Y - f_\diamond(X)\right)^2\right].$$

A causal model

response variable	Y
covariates	$X := X_1, \ldots, X_d$
training model	M
training data	i.i.d. from $P_M^{(X,Y)}$
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, ...

$$\text{minimax solution:} \quad \operatorname*{argmin}_{f_\diamond \in \mathcal{F}} \sup_{\tilde{M} \in \mathcal{N}(M)} E_{\tilde{M}}\big[\big(Y - f_\diamond(X)\big)^2\big].$$

A causal model satisfies

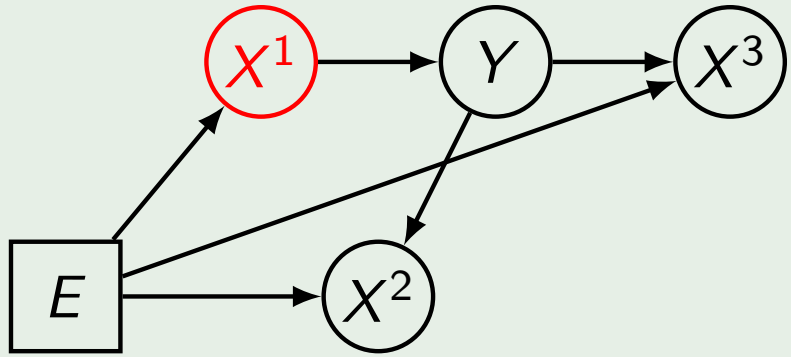
$$f_{causal} = \operatorname*{argmin}_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} E_{M(i)}\big[\big(Y - f_\diamond(X)\big)^2\big],$$

where $M(i)$: an intervention model and $\mathcal{I}$: all interventions on X.

Distribution Generalization

Example

$$\mathcal{S}(e): \begin{cases} X^1 := \mu_e + \epsilon_{X_1} \\ Y := \beta_1 X^1 + \epsilon_Y \\ X^2 := \gamma_e Y + \epsilon_{X_2} \\ X^3 := \eta_e Y + \epsilon_{X_3} \end{cases}$$

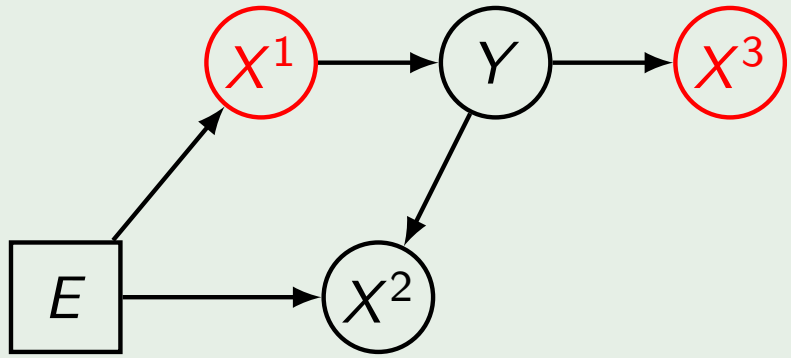


In some environments, X^2 and X^3 may be positively correlated with Y, while in others they can be negatively correlated. This poses a threat to generalization.

Distribution Generalization

Example (What if...)

$$\mathcal{S}(e): \begin{cases} X^1 := \mu_e + \epsilon_{X_1} \\ Y := \beta_1 X^1 + \epsilon_Y \\ X^2 := \gamma_e Y + \epsilon_{X_2} \\ X^3 := \eta Y + \epsilon_{X_3} \end{cases}$$



Distribution Generalization[2]

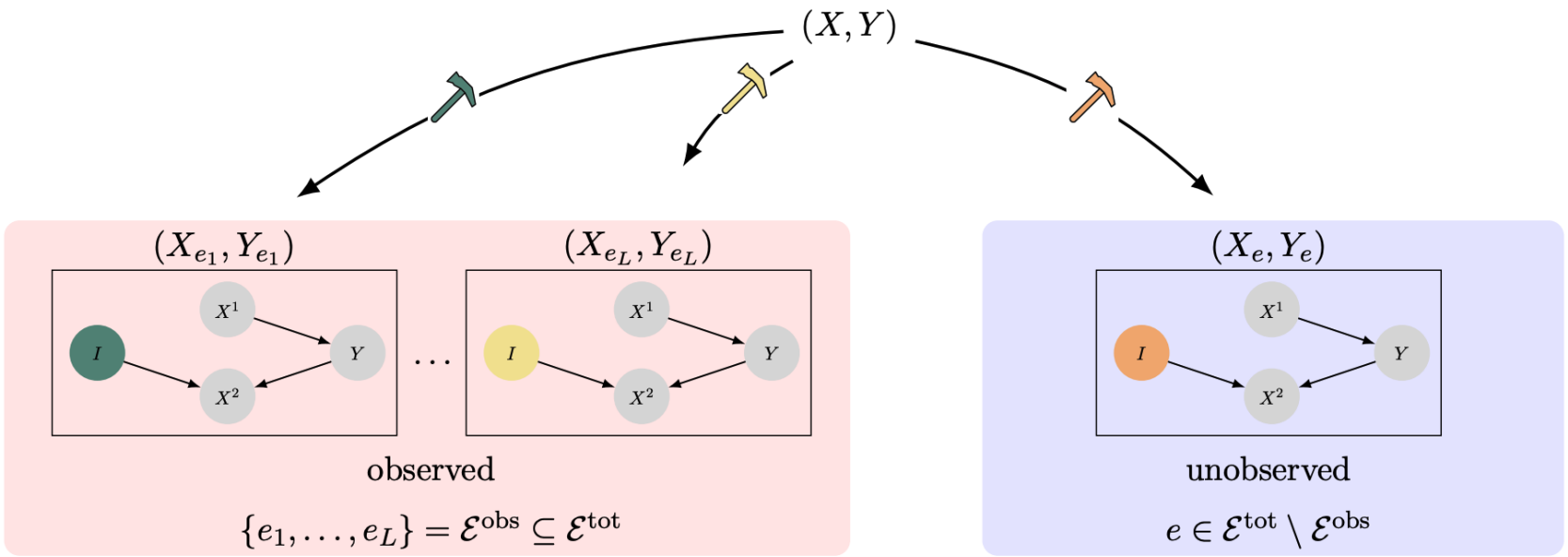


Figure 3.: Illustration of multi-environment data generation setting. Only some environments are observed, but one would like to be able to make predictions on any further potentially unobserved environment.

[2]Figure 3 is taken from Pfister et al. 2021

Distribution Generalization

Given: Training data: (Y^{tr}, X^{tr}, E^{tr}), Test data: (X^{test}, E^{test})

Distribution Generalization

Given: Training data: (Y^{tr}, X^{tr}, E^{tr}), Test data: (X^{test}, E^{test})

Invariance: $H_{0,S} : Y^{tr} \perp\!\!\!\perp E^{tr} \mid X_S^{tr}$

Invariant and most predictive: Let $\mathcal{S}^{\text{inv}}$ be a set of all invariant set. For each $S \in \mathcal{S}^{\text{inv}}$,

1. Find an estimate $\hat{\mathbb{E}}[Y^{tr} \mid X_S^{tr}]$
2. Compute the MSE $\mathbb{E}[(Y^{tr} - \hat{\mathbb{E}}[Y^{tr} \mid X_S^{tr}])^2]$

Return the subset S^* (and its corresponding estimate $\hat{\mathbb{E}}[Y^{tr} \mid X_{S^*}^{tr}]$) that yields the lowest MSE.

Distribution Generalization

Given: Training data: (Y^{tr}, X^{tr}, E^{tr}), Test data: (X^{test}, E^{test})

Invariance: $H_{0,S} : Y^{tr} \perp\!\!\!\perp E^{tr} \mid X_S^{tr}$

Invariant and most predictive: Let $\mathcal{S}^{\text{inv}}$ be a set of all invariant set. For each $S \in \mathcal{S}^{\text{inv}}$,

1. Find an estimate $\hat{\mathbb{E}}[Y^{tr} \mid X_S^{tr}]$
2. Compute the MSE $\mathbb{E}[(Y^{tr} - \hat{\mathbb{E}}[Y^{tr} \mid X_S^{tr}])^2]$

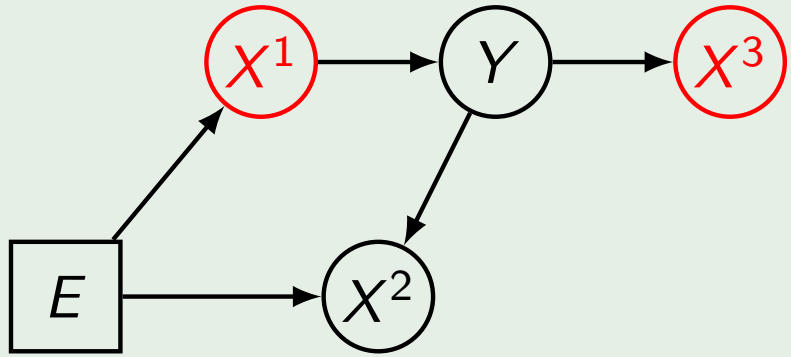
Return the subset S^* (and its corresponding estimate $\hat{\mathbb{E}}[Y^{tr} \mid X_{S^*}^{tr}]$) that yields the lowest MSE.

Environments: If $Y^{test} \perp\!\!\!\perp E^{test} \mid X_{S^*}^{test}$ holds then $\hat{\mathbb{E}}[Y^{tr} \mid X_{S^*}^{tr}]$ is expected to perform well on the test data.

Distribution Generalization

Example

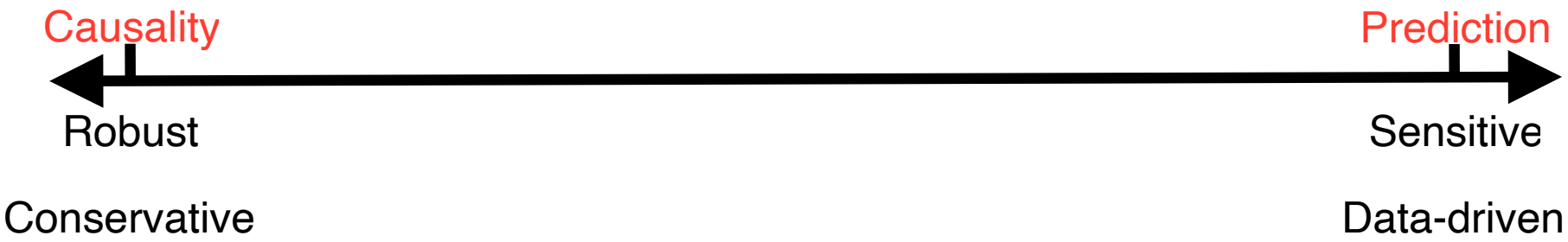
$$\mathcal{S}(e): \begin{cases} X^1 := \mu_e + \epsilon_{X_1} \\ Y := \beta_1 X^1 + \epsilon_Y \\ X^2 := \gamma_e Y + \epsilon_{X_2} \\ X^3 := \eta Y + \epsilon_{X_3} \end{cases}$$



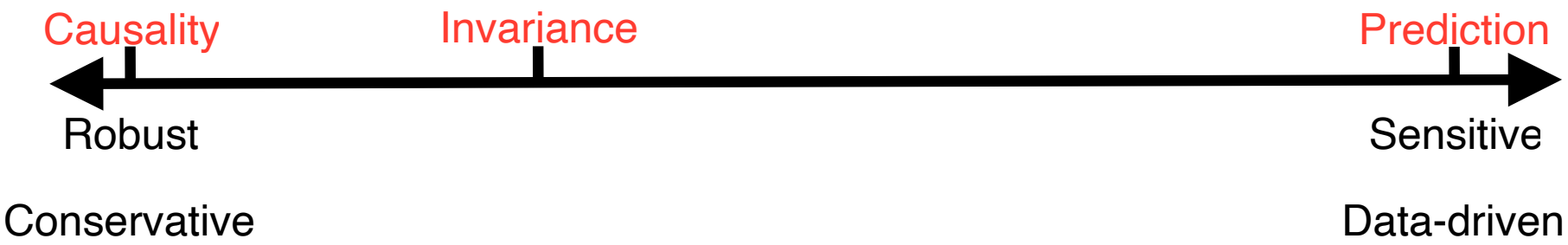
We have that $\{X^1\}$ and $\{X^1, X^3\}$ are the invariant sets. Although the causal parent is $\{X^1\}$, we may want to use both X^1, X^3 for predicting Y since the conditional expectation $\mathbb{E}[Y \mid X^1, X^3]$ is stable across environments and more predictive.

Conclusion and Open Questions

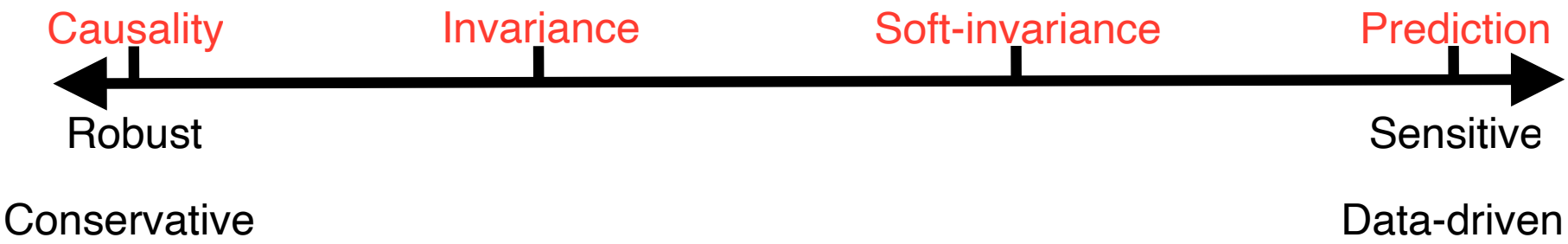
Conclusion and Open Questions



Conclusion and Open Questions



Conclusion and Open Questions



$$b^{\gamma}_{AR} := \operatorname*{argmin}_{b} \underbrace{\mathbf{E}_M (Y - Xb)^2}_{\text{prediction}} + \gamma \underbrace{\| \mathbf{E}_M A^t (Y - Xb) \|_2^2}_{\text{invariance}}$$

References I

- Bühlmann, Peter (2020). "Invariance, causality and robustness". In.
- Heinze-Deml, Christina, Jonas Peters, and Nicolai Meinshausen (2018). "Invariant Causal Prediction for Nonlinear Models". In: *Journal of Causal Inference* 6.2.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd. New York, USA: Cambridge University Press.
- Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen (2016). "Causal inference by using invariant prediction: identification and confidence intervals". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012.
- Pfister, N., E. G. William, J. Peters, R. Aebersold, and P. Bühlmann (2021). "Stabilizing Variable Selection and Regression". In: *Annals of Applied Statistics* 15.3, pp. 1220–1246.

References II


Rothenhäusler, Dominik, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters (2021). "Anchor regression: Heterogeneous data meet causality". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.2, pp. 215–246.


Saengkyongam, Sorawit, Nikolaj Thams, Jonas Peters, and Niklas Pfister (2023). "Invariant policy learning: A causal perspective". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.