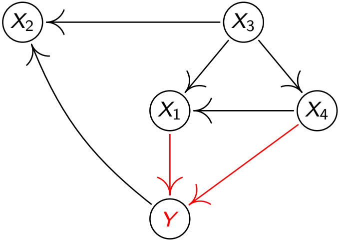
Distribution Generalization in Under-identified Causal Models

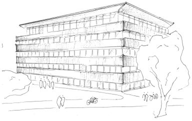


Jonas Peters, University of Copenhagen

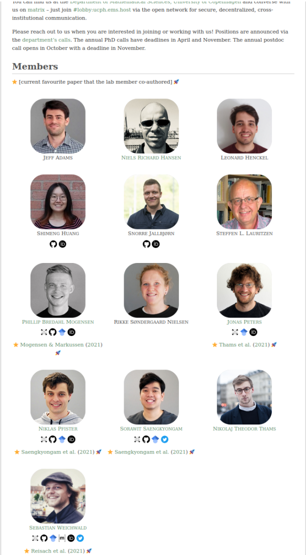
MSRI

8 March 2022

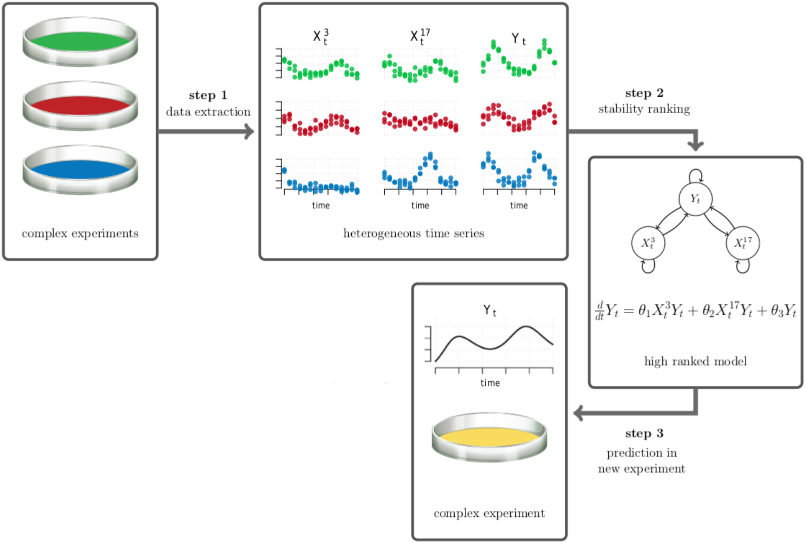
VILLUM FONDEN



joint work with...



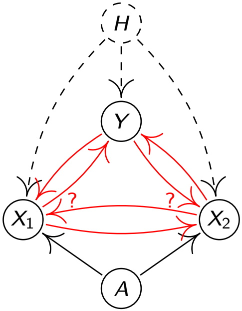
... and S. Bauer, R. Christiansen, N. Gnecco, M. Jakobsen



Real data: $(Y, X^1, \dots, X^{411})$, 11 time points, 5 exp., 3 rep.

N. Pfister, S. Bauer, JP: *Learning stable structures in kinetic systems: benefits of a causal approach*, PNAS 2019

Instrumental Variables:

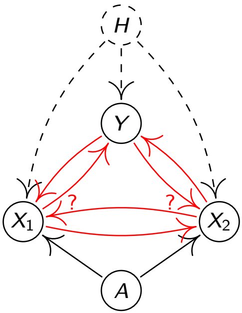


$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

A1 A and Y are d-separated when removing $X_1, X_2 \to Y$ (exclusion restriction). Then,

$$E[A(Y - \alpha_1 X_1 - \alpha_2 X_2)] = 0 \quad \Leftarrow \quad (\alpha_1, \alpha_2) = (\beta_1, \beta_2)$$

Instrumental Variables:



$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

A1 A and Y are d-separated when removing $X_1, X_2 \to Y$ (exclusion restriction). Then,

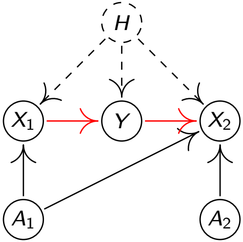
$$E[A(Y - \alpha_1 X_1 - \alpha_2 X_2)] = 0 \quad \Leftarrow \quad (\alpha_1, \alpha_2) = (\beta_1, \beta_2)$$

A2 In addition, $E[AX^\top]$ is full rank. Then,

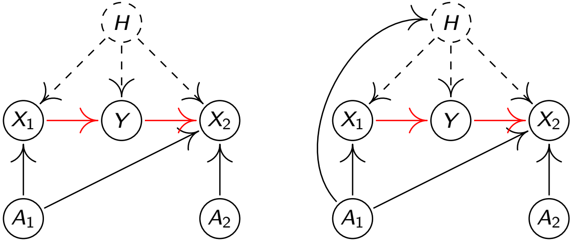
$$E[A(Y - \alpha_1 X_1 - \alpha_2 X_2)] = 0 \quad \Leftrightarrow \quad (\alpha_1, \alpha_2) = (\beta_1, \beta_2)$$

Anderson and Rubin 1949, Theil 1953, ...

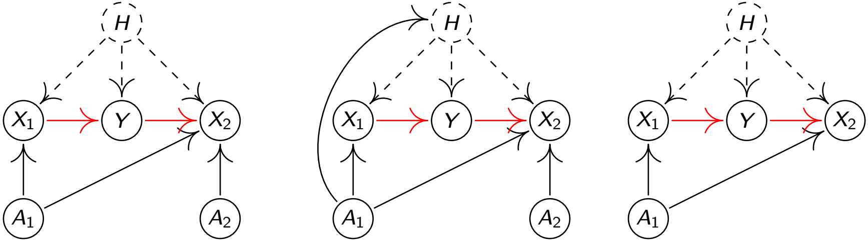
Examples:



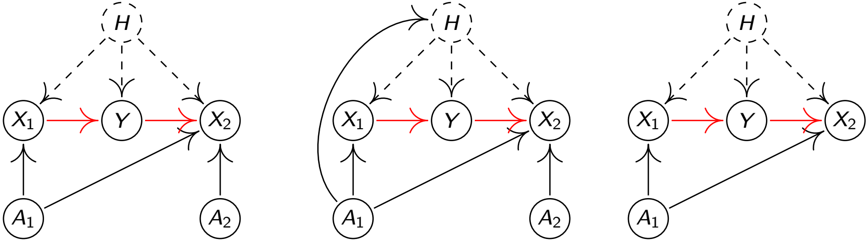
Examples:



Examples:



Examples:

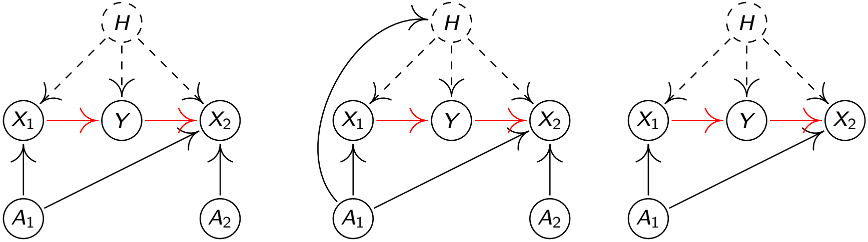


Example 3 (under-identified): solution space of

$$E[A_1(Y - \alpha_1 X_1 - \alpha_2 X_2)] = 0$$

has dimension one.

Examples:



Example 3 (under-identified): solution space of

$$E[A_1(Y - \alpha_1 X_1 - \alpha_2 X_2)] = 0$$

has dimension one.

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

How does this help for distribution generalization?

How does this help for distribution generalization?

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

Consider (unknown) model M

$$\begin{aligned} A &:= \epsilon_A & \in \mathbb{R}^q \\ H &:= \epsilon_H & \\ X &:= BX + \gamma Y + CH + GA + \epsilon_X & \in \mathbb{R}^d \\ Y &:= \beta^\top X + FH + \epsilon_Y & \end{aligned}$$

with $\mathrm{cov}(A, A)$ full rank.

How does this help for distribution generalization?

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

Consider (unknown) model M

$$
\begin{aligned}
A &:= \epsilon_A && \in \mathbb{R}^q \\
H &:= \epsilon_H && \\
X &:= BX + \gamma Y + CH + GA + \epsilon_X && \in \mathbb{R}^d \\
Y &:= \beta^\top X + FH + \epsilon_Y &&
\end{aligned}
$$

with $\mathrm{cov}(A, A)$ full rank. Then, for $\mathcal{F}_{\mathrm{inv}} := \{\alpha \in \mathbb{R}^d \mid E[A(Y - \alpha^\top X)] = 0\}$

How does this help for distribution generalization?

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

Consider (unknown) model M

$$\begin{aligned} A &:= \epsilon_A && \in \mathbb{R}^q \\ H &:= \epsilon_H && \\ X &:= BX + \gamma Y + CH + GA + \epsilon_X && \in \mathbb{R}^d \\ Y &:= \beta^\top X + FH + \epsilon_Y && \end{aligned}$$

with $\mathrm{cov}(A, A)$ full rank. Then, for $\mathcal{F}_{\text{inv}} := \{\alpha \in \mathbb{R}^d \mid E[A(Y - \alpha^\top X)] = 0\}$

$$\operatorname*{argmin}_{\alpha \in \mathcal{F}_{\text{inv}}} E_M[(Y - \alpha^\top X)^2] = \operatorname*{argmin}_{\alpha} \sup_{i \in \mathbb{R}^q} E_{M(i)}[(Y - \alpha^\top X)^2],$$

where $M(i)$ corresponds to the intervention $do(A := i)$.

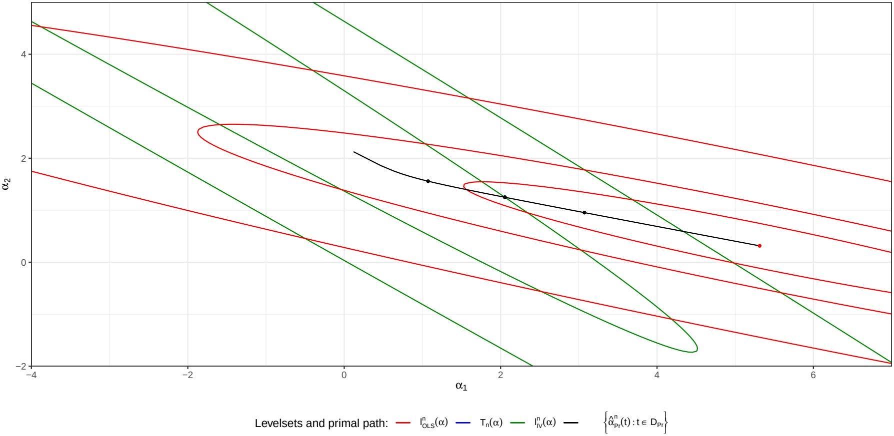
What do we do for finite sample size?

What do we do for finite sample size?

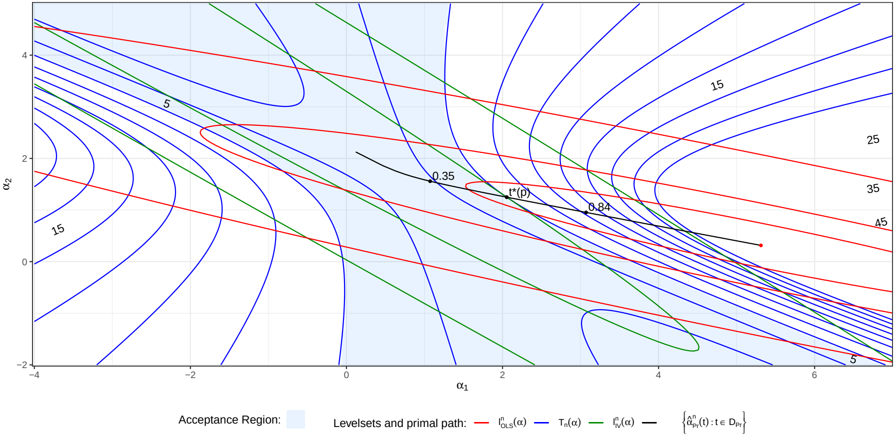
$$\alpha^\gamma := \underset{\alpha}{\operatorname{argmin}} \underbrace{\mathsf{E}(Y - X\alpha)^2}_{\text{prediction}} \qquad \text{s.t.} \underbrace{\|\mathsf{E}A^\top(Y - X\alpha)\|_2^2}_{\text{invariance}} \le \gamma$$

$$\hat{\alpha}_n^\gamma := \underset{\alpha}{\operatorname{argmin}} (\mathsf{Y} - \mathsf{X}\alpha)^\top(\mathsf{Y} - \mathsf{X}\alpha) \text{ s.t. } (\mathsf{Y} - \mathsf{X}\alpha)^\top \mathsf{A}(\mathsf{A}^\top\mathsf{A})^{-1}\mathsf{A}^\top(\mathsf{Y} - \mathsf{X}\alpha) \le \gamma$$

PULSE: Choose γ, such that $\texttt{cor.test}(A, Y - X\hat{\alpha}_n^{\gamma})\texttt{.pvalue == 0.05}$



PULSE: Choose γ, such that $\texttt{cor.test}(\mathsf{A}, \mathsf{Y} - \mathsf{X}\hat{\alpha}_n^{\gamma})\texttt{.pvalue == 0.05}$



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

Rothenhäusler, Bühlmann, Meinshausen, JP, JRSSB, 2021

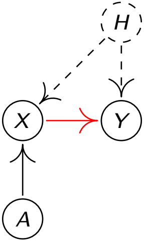
e.g., Anderson and Rubin 1949 and Theil 1958 and Fuller 1977

'Roadmap':

1. Find identifying equations.
2. Analyse identifiability conditions.
3. Among all invariant models, choose the most predictive one.

Example: HSIC-X.

$$\begin{aligned} A &:= \epsilon_A \\ H &:= \epsilon_H \\ X &:= g(A, H, \epsilon_X) \\ Y &:= \beta^\top \phi(X) + h(H, \epsilon_Y) \end{aligned}$$

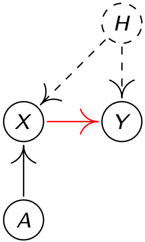


1. Identifying equation

$$A \perp\!\!\!\perp Y - \beta^\top \phi(X)$$

Example: HSIC-X.

$$
\begin{aligned}
A &:= \epsilon_A \\
H &:= \epsilon_H \\
X &:= g(A, H, \epsilon_X) \\
Y &:= \beta^\top \phi(X) + h(H, \epsilon_Y)
\end{aligned}
$$



1. Identifying equation

$$A \perp\!\!\!\perp Y - \beta^\top \phi(X)$$

2. Identifiability condition

$$A \perp\!\!\!\perp h(H, \epsilon_Y) + \tau^\top \phi(X) \quad \Rightarrow \quad \tau = 0.$$

3. Among all invariant models, choose the most predictive one (HSIC-X-pen: optimize empirical HSIC Gretton et al 2008)

S. Saengkyongam, L. Henckel, N. Pfister, J. Peters: Exploiting Indep. Instruments: Identification and Distr. Gener., arXiv 2022

$$\mathcal{F}_{\text{inv}} := \{f_\diamond \in \mathcal{F} \mid A \perp\!\!\!\perp Y - f_\diamond(X) \text{ under } \mathbb{P}_{M^0}\}\,.$$

Theorem (Invariance with respect to interventions on A)

Let $\ell : \mathbb{R} \to \mathbb{R}$ be convex and $\mathcal{I}$ be a set of interventions on A satisfying for all $i \in \mathcal{I}$ that $\mathbb{P}_{M(i)}$ is dominated by[a] $\mathbb{P}_M$.

i) *Then, for all $f \in \mathcal{F}_{inv}$ it holds that*

$$E_M\big[\ell(Y - f(X))\big] = \sup_{i \in \mathcal{I}} E_{M(i)}\big[\ell(Y - f(X))\big].$$

$$\mathcal{F}_{\text{inv}} := \{f_\diamond \in \mathcal{F} \mid A \perp\!\!\!\perp Y - f_\diamond(X) \text{ under } \mathbb{P}_{M^0}\}.$$

Theorem (Invariance with respect to interventions on A)

Let $\ell : \mathbb{R} \to \mathbb{R}$ be convex and $\mathcal{I}$ be a set of interventions on A satisfying for all $i \in \mathcal{I}$ that $\mathbb{P}_{M(i)}$ is dominated by[a] $\mathbb{P}_M$.

i) *Then, for all $f \in \mathcal{F}_{inv}$ it holds that*

$$E_M\big[\ell(Y - f(X))\big] = \sup_{i \in \mathcal{I}} E_{M(i)}\big[\ell(Y - f(X))\big].$$

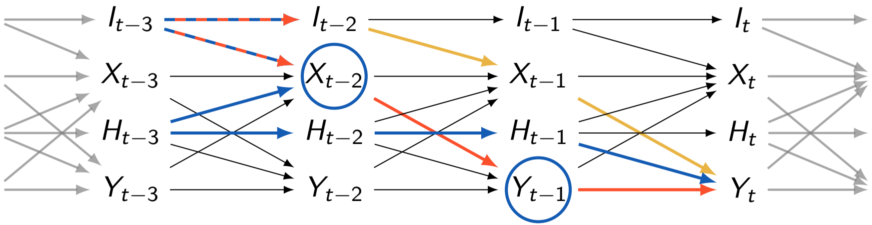
ii) *Let S be the covariates, affected by A. If there exists $i_* \in \mathcal{I}$ such that $X^S \perp\!\!\!\perp U \mid X^{S^c}$ under $\mathbb{P}_{M(i_*)}$ and $\operatorname{supp}(\mathbb{P}^X_{M(i_*)}) = \operatorname{supp}(\mathbb{P}^X_M)$, then*

$$\inf_{f \in \mathcal{F}_{inv}} E_M\big[\ell(Y - f(X))\big] = \inf_{f \in \mathcal{F}} \sup_{i \in \mathcal{I}} E_{M(i)}\big[\ell(Y - f(X))\big].$$

[a] If A enters the system nonlinearly, this cannot be dropped (even if f is linear), see Prop. 4.9, Christiansen et al., IEEE TPAMI 2021.

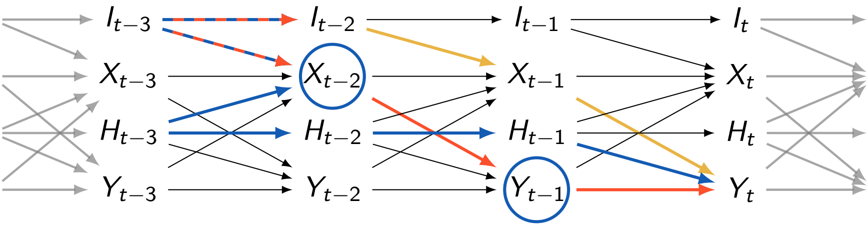
S. Saengkyongam, L. Henckel, N. Pfister, J. Peters: Exploiting Indep. Instruments: Identification and Distr. Gener., arXiv 2022

Example: VAR processes.



1. Here: $E[I_{t-2}(Y_t - \beta^\top X_{t-1})] \neq 0$ (see red path).

Example: VAR processes.

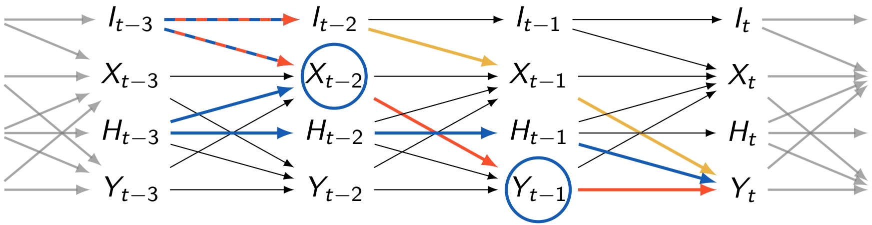


1. Here: $E[I_{t-2}(Y_t - \beta^\top X_{t-1})] \neq 0$ (see red path). Instead: e.g.,

$$E\left[\begin{pmatrix} I_{t-2} \\ I_{t-3} \end{pmatrix} (Y_t - \beta^\top X_{t-1} - \gamma Y_{t-1})\right] = 0$$

(nuisance IV).

Example: VAR processes.

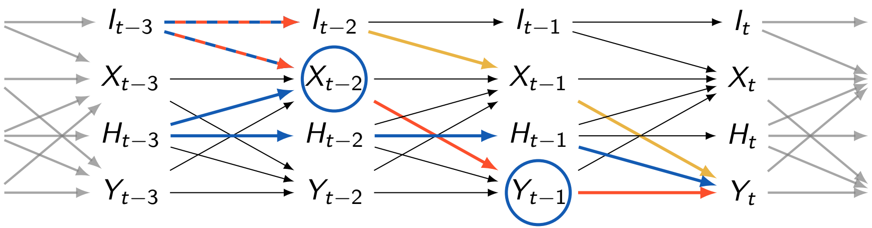


1. Here: $E[I_{t-2}(Y_t - \beta^\top X_{t-1})] \neq 0$ (see red path). Instead: e.g.,

$$E\left[\begin{pmatrix} I_{t-2} \\ I_{t-3} \end{pmatrix} (Y_t - \beta^\top X_{t-1} - \gamma Y_{t-1})\right] = 0$$

(nuisance IV). 2. Identifiability: rank condition is equivalent to conditions on the Jordan normal form of the coef matrix.

Example: VAR processes.



1. Here: $E[I_{t-2}(Y_t - \beta^\top X_{t-1})] \neq 0$ (see red path). Instead: e.g.,

$$E\left[\begin{pmatrix} I_{t-2} \\ I_{t-3} \end{pmatrix}(Y_t - \beta^\top X_{t-1} - \gamma Y_{t-1})\right] = 0$$

(nuisance IV). 2. Identifiability: rank condition is equivalent to conditions on the Jordan normal form of the coef matrix. 3. (similar as before)

N. Thams, R. Nielsen, S. Weichwald, J. Peters:
Identifying Causal Effects using Instrumental Time Series: Nuisance IV and Correcting for the Past, arXiv 2022

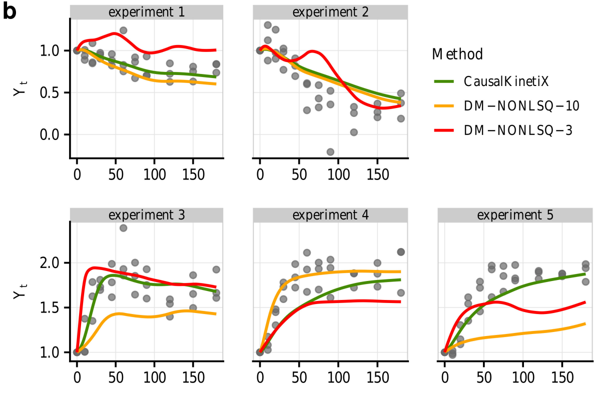
Real data: $(Y, X^1, \ldots, X^{411})$, 11 time points, 5 exp., 3 rep.; $Z_t := 2 - Y_t$

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

Real data: $(Y, X^1, \dots, X^{411})$, 11 time points, 5 exp., 3 rep.; $Z_t := 2 - Y_t$

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

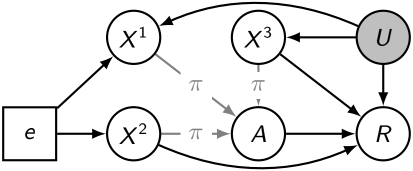
Out-of-sample plot



N. Pfister, S. Bauer, JP: *Learning stable structures in kinetic systems: benefits of a causal approach*, PNAS 2019

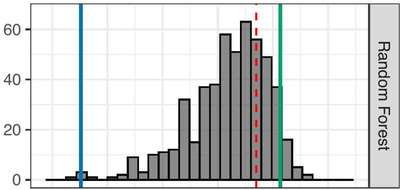
Invariant Policy Learning:

Saengkyongam, Thams, JP, Pfister, arXiv:2106.00808, 2021



Terrestrial ecosystem data:

Migliavacca et al, Nature 2021



Summary

- Invariance can be used to identify causal models ... but only if identifiability conditions hold. If not:

Summary

- Invariance can be used to identify causal models ... but only if identifiability conditions hold. If not:
- Proposal: Among all invariant models, choose the best predictive one.

Summary

- Invariance can be used to identify causal models ... but only if identifiability conditions hold. If not:
- Proposal: Among all invariant models, choose the best predictive one. This often minimises worst-case prediction errors.

Summary

- Invariance can be used to identify causal models ... but only if identifiability conditions hold. If not:
- Proposal: Among all invariant models, choose the best predictive one. This often minimises worst-case prediction errors.
  a) linear models (PULSE)
  b) exploiting independence (HSIC-X)
  c) discrete-time dynamical systems (TS-IV)
  d) chemical reaction networks (Causal KinetiX)
  e) not shown: contextual bandits (Invariant Policy Learning)
  f) not shown: Earth system science (causal GOF)

Summary

- Invariance can be used to identify causal models ... but only if identifiability conditions hold. If not:
- Proposal: Among all invariant models, choose the best predictive one. This often minimises worst-case prediction errors.
  a) linear models (PULSE)
  b) exploiting independence (HSIC-X)
  c) discrete-time dynamical systems (TS-IV)
  d) chemical reaction networks (Causal KinetiX)
  e) not shown: contextual bandits (Invariant Policy Learning)
  f) not shown: Earth system science (causal GOF)

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

N. Pfister, S. Bauer, JP: *Learning stable structures in kinetic systems: benefits of a causal approach*, PNAS 2019

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

S. Saengkyongam, L. Henckel, N. Pfister, JP: Exploiting Indep. Instruments: Identification and Distr. Gener., arXiv 2022

N. Thams, R. Nielsen, S. Weichwald, JP: Identif. Causal Effects using Instr. TS: Nuisance IV and Corr. for the Past, arXiv 2022

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

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

Migliavacca et al.: The three major axes of terrestrial ecosystem function, Nature 2021