UCL CSML Seminar Series

Imperial College London

Causal Discovery through Bayesian Model Selection

Mark van der Wilk

Department of Computing
Imperial College London
https://mvdw.uk

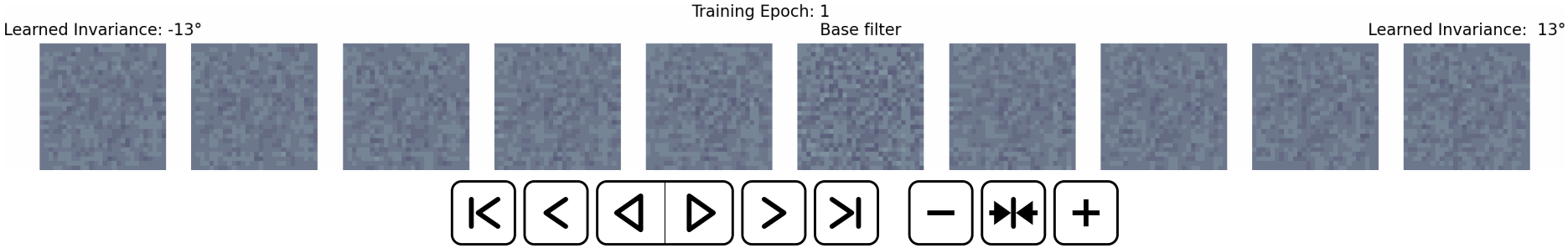
@markvanderwilk
m.vdwilk@imperial.ac.uk

Jun 2, 2023

My Research

Research interests:

- Automatically learning inductive biases (automatic + computational)
- Adaptivity & robustness: Continual learning, uncertainty
- Applications: Decision making, BayesOpt, Model-based RL
- Methods:
  - Bayesian model selection / Occam's razor / MDL
  - Meta-learning
  - Approximate inference
  - (Deep) Gaussian processes & relations to DNNs
  - Interesting architectures (capsule networks)



Bayesian Model Selection for Causality

Causal Discovery using Marginal Likelihood
CML4Impact workshop @ NeurIPS 2022
Anish Dhir · MvdW

Extended version soon to be on arxiv.

Overview

Causality: Overview and Background

Causality: Approaches and Problems

Bayesian Model Selection for Causality and its Properties

An Actual Method

Discussion & Conclusion

Causality: Motivation

- We all know that “correlation is not causation”

Causality: Motivation

- We all know that “correlation is not causation”
- Impossibility theorems of determining causation from observations only

Causality: Motivation

- We all know that "correlation is not causation"
- Impossibility theorems of determining causation from observations only
- Gold standard: Intervene! (RCT)

Causality: Motivation

- We all know that "correlation is not causation"
- Impossibility theorems of determining causation from observations only
- Gold standard: Intervene! (RCT)

One of the big questions in the causal inference community:

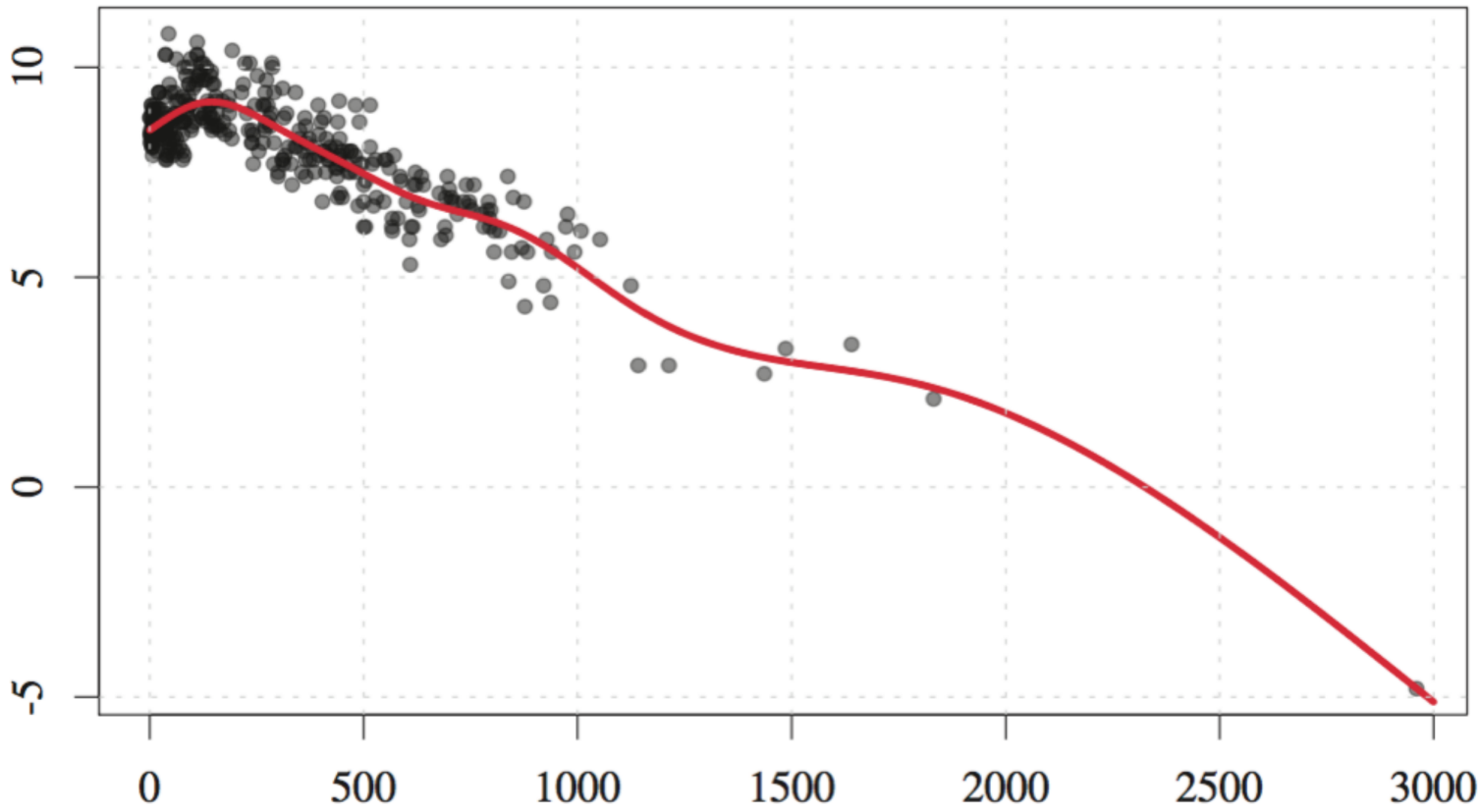
Causality: Motivation

- We all know that "correlation is not causation"
- Impossibility theorems of determining causation from observations only
- Gold standard: Intervene! (RCT)

One of the big questions in the causal inference community:

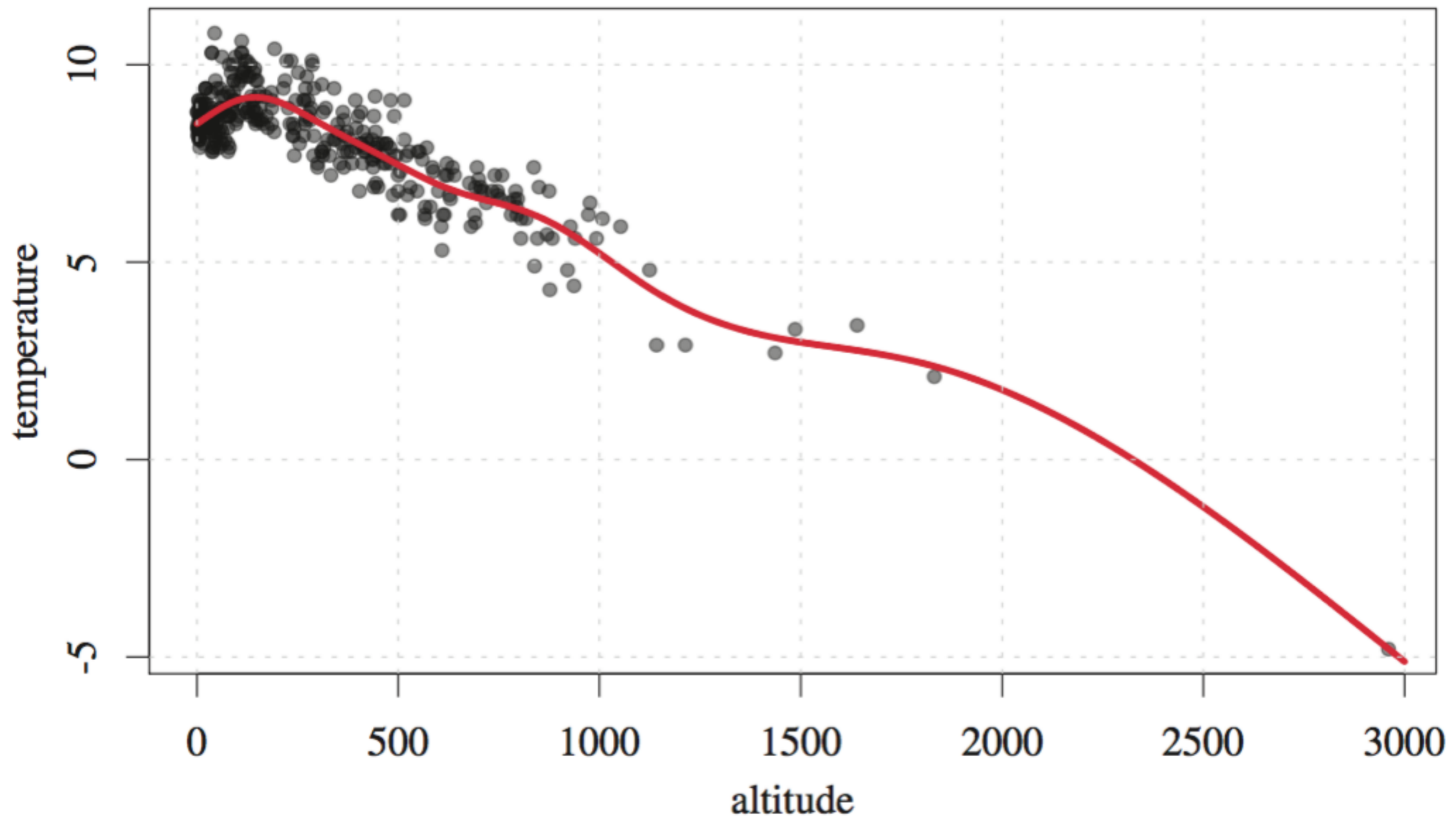
How far can we get with observational data alone?

Causality: Example



Does (X cause Y) or does (Y cause X)?

Causality: Example



Does (X cause Y) or does (Y cause X)?

Structural Causal Models

We assume data to arise from a **Structural Causal Model**

- Hierarchical order of data generation from causes to effects.

Structural Causal Models

We assume data to arise from a **Structural Causal Model**

- Hierarchical order of data generation from causes to effects.

For bivariate problems, we simply have:

$$X_i := f_X(N_{X_i}), \qquad Y_i := f_Y(X_i, N_{Y_i}), \tag{1}$$

- N_{X_i}, N_{Y_i} are independent noise RVs with arbitrary distribution
- Procedure for $Y \to X$ is analogous

Structural Causal Models

We assume data to arise from a **Structural Causal Model**

- Hierarchical order of data generation from causes to effects.

For bivariate problems, we simply have:

$$X_i := f_X(N_{X_i}), \qquad Y_i := f_Y(X_i, N_{Y_i}), \tag{1}$$

- N_{X_i}, N_{Y_i} are independent noise RVs with arbitrary distribution
- Procedure for $Y \to X$ is analogous

We consider behaviour over many datasets, each of which is generated by different functions f_X, f_Y.

- A distribution on datasets arises from distributions $\Pi(F_X), \Pi(F_Y)$
- For dataset $\mathcal{D} = (X, Y) = \{(X_i, Y_i)\}$ we get distribution $\Pi(X, Y)$

Factorisations & Interventions

Assuming $X \to Y$, we can factorise our joint in two ways:

- $\Pi(X, Y) = \Pi(X)\Pi(Y|X)$: Causal factorisation
- $\Pi(X, Y) = \Pi(X|Y)\Pi(Y)$: Anticausal factorisation

Factorisations & Interventions

Assuming $X \to Y$, we can factorise our joint in two ways:

- $\Pi(X, Y) = \Pi(X)\Pi(Y|X)$: Causal factorisation
- $\Pi(X, Y) = \Pi(X|Y)\Pi(Y)$: Anticausal factorisation

An intervention on a variable changes its value, generation function, or noise input, **while leaving those of all other variables unchanged**. E.g. if $X \to Y$, an intervention on X

- Only changes $\Pi(X)$
- I.e. leaves $\Pi(Y|X)$ unchanged

Factorisations & Interventions

Assuming $X \to Y$, we can factorise our joint in two ways:

- $\Pi(X, Y) = \Pi(X)\Pi(Y|X)$: Causal factorisation
- $\Pi(X, Y) = \Pi(X|Y)\Pi(Y)$: Anticausal factorisation

An intervention on a variable changes its value, generation function, or noise input, **while leaving those of all other variables unchanged**. E.g. if $X \to Y$, an intervention on X

- Only changes $\Pi(X)$
- I.e. leaves $\Pi(Y|X)$ unchanged

The causal factorisation is special, because interventions leave part unchanged. For the anticausal factorisation:

- $\Pi(Y) = \sum_X \Pi(Y|X)\Pi(X)$ and $\Pi(X|Y) \propto \Pi(Y|X)\Pi(X)$
- Changing $\Pi(X)$ changes both parts.

Factorisations & Interventions

Assuming $X \to Y$, we can factorise our joint in two ways:

- $\Pi(X, Y) = \Pi(X)\Pi(Y|X)$: Causal factorisation
- $\Pi(X, Y) = \Pi(X|Y)\Pi(Y)$: Anticausal factorisation

An intervention on a variable changes its value, generation function, or noise input, **while leaving those of all other variables unchanged**. E.g. if $X \to Y$, an intervention on X

- Only changes $\Pi(X)$
- I.e. leaves $\Pi(Y|X)$ unchanged

The causal factorisation is special, because interventions leave part unchanged. For the anticausal factorisation:

- $\Pi(Y) = \sum_X \Pi(Y|X)\Pi(X)$ and $\Pi(X|Y) \propto \Pi(Y|X)\Pi(X)$
- Changing $\Pi(X)$ changes both parts.

Assumptions known as **Independent Causal Mechanism** (ICM)

Causal Models

Our model should directly parameterise the causal factorisation:

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{y}|\phi, \theta, \mathcal{M}_{\mathbf{X}\to Y}) &= p(\mathbf{x}|\phi, \mathcal{M}_{\mathbf{X}\to Y}) p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}_{\mathbf{X}\to Y}) \,, \\
p(\mathbf{x}, \mathbf{y}|\phi, \theta, \mathcal{M}_{X\leftarrow\mathbf{Y}}) &= p(\mathbf{y}|\phi, \mathcal{M}_{X\leftarrow\mathbf{Y}}) p(\mathbf{x}|\mathbf{y}, \theta, \mathcal{M}_{X\leftarrow\mathbf{Y}}) \,,
\end{aligned}
\tag{2}
$$

where the marginals are chosen from $\mathcal{R}$ and conditionals from $\mathcal{C}$

$$
\mathcal{R} = \{p(\cdot|\phi) \mid \phi \in \Phi\} \,, \qquad \mathcal{C} = \{p(\cdot|\cdot, \theta) \mid \theta \in \Theta\}. \tag{3}
$$

Causal Models

Our model should directly parameterise the causal factorisation:

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{y}|\phi, \theta, \mathcal{M}_{\mathbf{X}\to Y}) &= p(\mathbf{x}|\phi, \mathcal{M}_{\mathbf{X}\to Y}) p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}_{\mathbf{X}\to Y}) \,, \\
p(\mathbf{x}, \mathbf{y}|\phi, \theta, \mathcal{M}_{X\leftarrow\mathbf{Y}}) &= p(\mathbf{y}|\phi, \mathcal{M}_{X\leftarrow\mathbf{Y}}) p(\mathbf{x}|\mathbf{y}, \theta, \mathcal{M}_{X\leftarrow\mathbf{Y}}) \,,
\end{aligned}
\tag{2}
$$

where the marginals are chosen from $\mathcal{R}$ and conditionals from $\mathcal{C}$

$$
\mathcal{R} = \{p(\cdot|\phi) \mid \phi \in \Phi\} \,, \qquad \mathcal{C} = \{p(\cdot|\cdot, \theta) \mid \theta \in \Theta\}. \tag{3}
$$

Allows part of the model to be re-used after an intervention
without degradation of accuracy

Overview

Causality: Overview and Background

Causality: Approaches and Problems

Bayesian Model Selection for Causality and its Properties

An Actual Method

Discussion & Conclusion

Causal Discovery by Fitting Models

- We can find θ, ϕ by fitting the model (maximum likelihood)

Causal Discovery by Fitting Models

- We can find θ, ϕ by fitting the model (maximum likelihood)
- Accurate predictive models need large sets $\mathcal{R}, \mathcal{C}$

Causal Discovery by Fitting Models

- We can find θ, ϕ by fitting the model (maximum likelihood)
- Accurate predictive models need large sets $\mathcal{R}, \mathcal{C}$
- Can we determine $\mathcal{M}_{\mathbf{X} \to Y}$ or $\mathcal{M}_{X \leftarrow \mathbf{Y}}$ by fitting the model?

Causal Discovery by Fitting Models

- We can find θ, ϕ by fitting the model (maximum likelihood)
- Accurate predictive models need large sets $\mathcal{R}, \mathcal{C}$
- Can we determine $\mathcal{M}_{\mathbf{X}\to Y}$ or $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ by fitting the model?

No.

$\mathcal{M}_{\mathbf{X}\to Y}$ and $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ are in the same Markov Equivalence Class.

Causal Discovery by Fitting Models

- We can find θ, ϕ by fitting the model (maximum likelihood)
- Accurate predictive models need large sets $\mathcal{R}, \mathcal{C}$
- Can we determine $\mathcal{M}_{\mathbf{X}\to Y}$ or $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ by fitting the model?

No.

$\mathcal{M}_{\mathbf{X}\to Y}$ and $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ are in the same Markov Equivalence Class.

$$\max_{\theta,\phi} p(\mathbf{x}|\phi, \mathcal{M}_{\mathbf{X}\to Y}) p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}_{\mathbf{X}\to Y}) = \max_{\theta,\phi} p(\mathbf{y}|\phi, \mathcal{M}_{X\leftarrow\mathbf{Y}}) p(\mathbf{x}|\mathbf{y}, \theta, \mathcal{M}_{X\leftarrow\mathbf{Y}}).$$

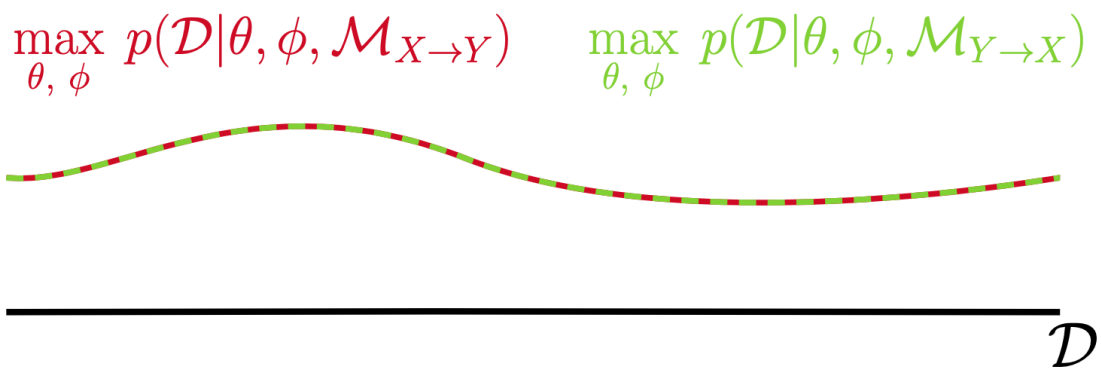
Causal Discovery by Fitting Models

- We can find θ, ϕ by fitting the model (maximum likelihood)
- Accurate predictive models need large sets $\mathcal{R}, \mathcal{C}$
- Can we determine $\mathcal{M}_{\mathbf{X}\to Y}$ or $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ by fitting the model?

No.

$\mathcal{M}_{\mathbf{X}\to Y}$ and $\mathcal{M}_{X\leftarrow\mathbf{Y}}$ are in the same Markov Equivalence Class.

$$\max_{\theta,\phi} p(\mathbf{x}|\phi, \mathcal{M}_{\mathbf{X}\to Y}) p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}_{\mathbf{X}\to Y}) = \max_{\theta,\phi} p(\mathbf{y}|\phi, \mathcal{M}_{X\leftarrow\mathbf{Y}}) p(\mathbf{x}|\mathbf{y}, \theta, \mathcal{M}_{X\leftarrow\mathbf{Y}}).$$



Related Work: Restricted Model Classes

- If we assume that data is generated from restricted $\mathcal{R}, \mathcal{C}$, we **can** identify causal direction.

Related Work: Restricted Model Classes

- If we assume that data is generated from restricted $\mathcal{R}, \mathcal{C}$, we **can** identify causal direction.
- E.g. Additive Noise Models (ANM). For $\mathcal{M}_{\mathbf{X} \to Y}$ assume

$$Y_i = f(X_i) + N_{Y_i} \tag{4}$$

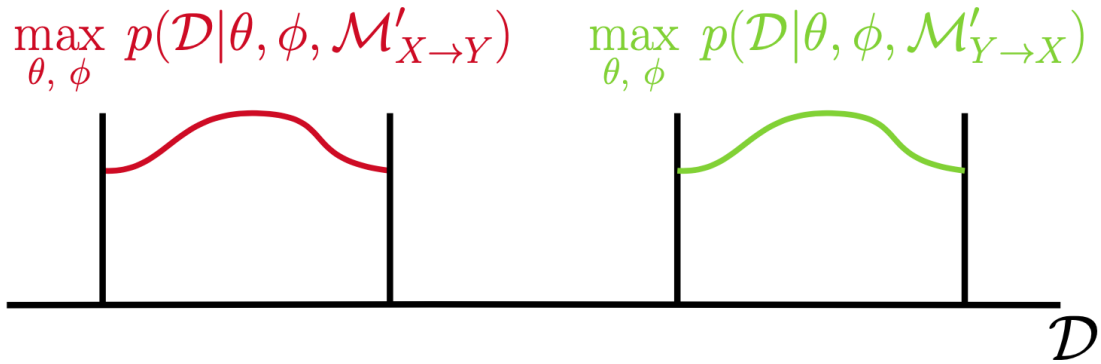
i.e. N_{Y_i} is independent from X_i and is additive.

Related Work: Restricted Model Classes

- If we assume that data is generated from restricted $\mathcal{R}, \mathcal{C}$, we **can** identify causal direction.
- E.g. Additive Noise Models (ANM). For $\mathcal{M}_{\mathbf{X} \to Y}$ assume

$$Y_i = f(X_i) + N_{Y_i} \tag{4}$$

i.e. N_{Y_i} is independent from X_i and is additive.

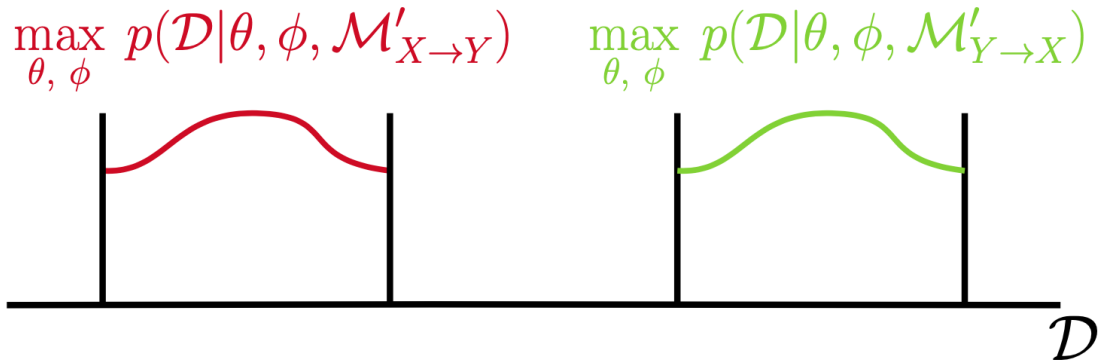


Related Work: Restricted Model Classes

- If we assume that data is generated from restricted $\mathcal{R}, \mathcal{C}$, we **can** identify causal direction.
- E.g. Additive Noise Models (ANM). For $\mathcal{M}_{\mathbf{X}\to Y}$ assume

$$Y_i = f(X_i) + N_{Y_i} \tag{4}$$

i.e. N_{Y_i} is independent from X_i and is additive.



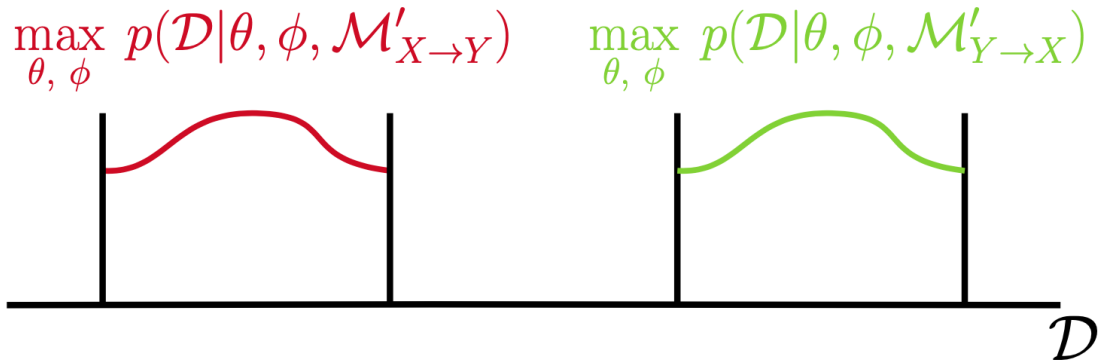
- (Example on whiteboard.)

Related Work: Restricted Model Classes

- If we assume that data is generated from restricted $\mathcal{R}, \mathcal{C}$, we **can** identify causal direction.
- E.g. Additive Noise Models (ANM). For $\mathcal{M}_{\mathbf{X}\to Y}$ assume

$$Y_i = f(X_i) + N_{Y_i} \tag{4}$$

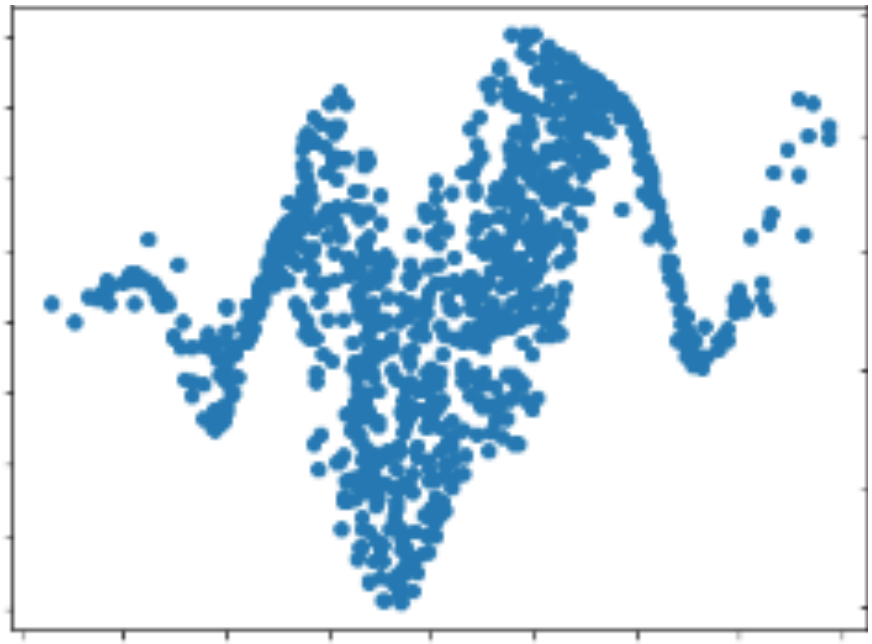
i.e. N_{Y_i} is independent from X_i and is additive.



- (Example on whiteboard.)
- Other classes of models are similarly identifiable.

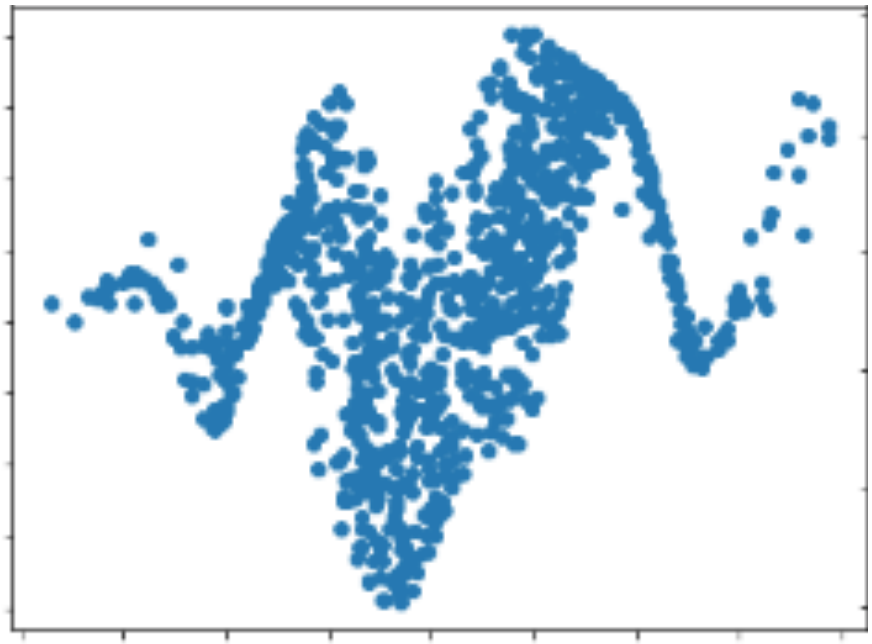
Additive Noise Model: Problem

- What to do for a dataset like this?



Additive Noise Model: Problem

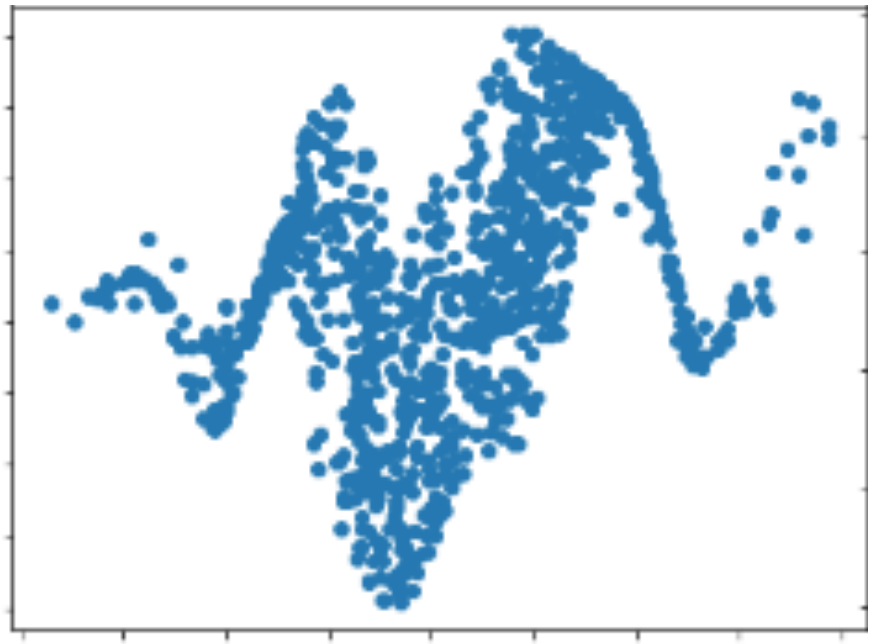
- What to do for a dataset like this?



- ANM is misspecified!

Additive Noise Model: Problem

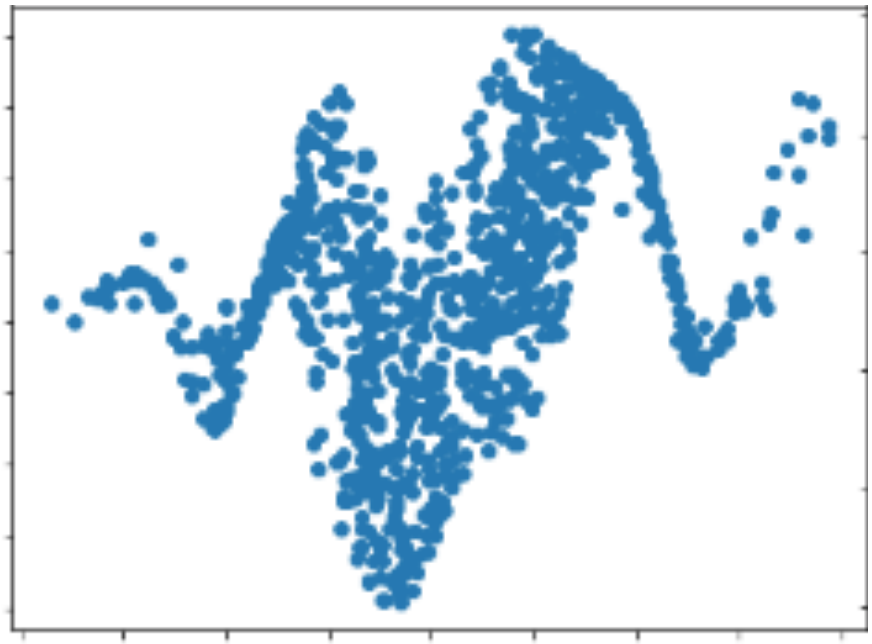
- What to do for a dataset like this?



- ANM is misspecified!
- $\implies$ won't fit properly $\implies$ bad predictive model

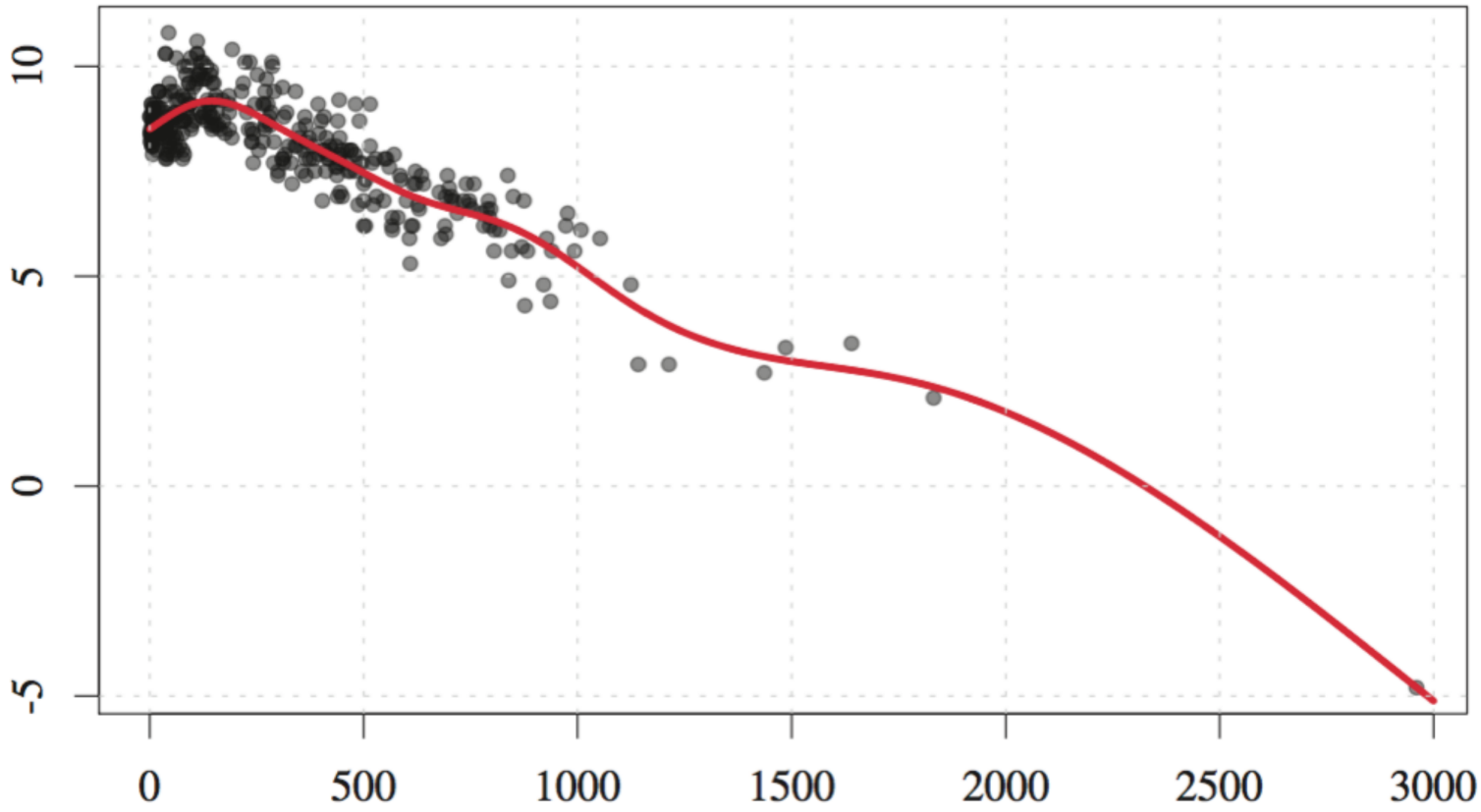
Additive Noise Model: Problem

- What to do for a dataset like this?

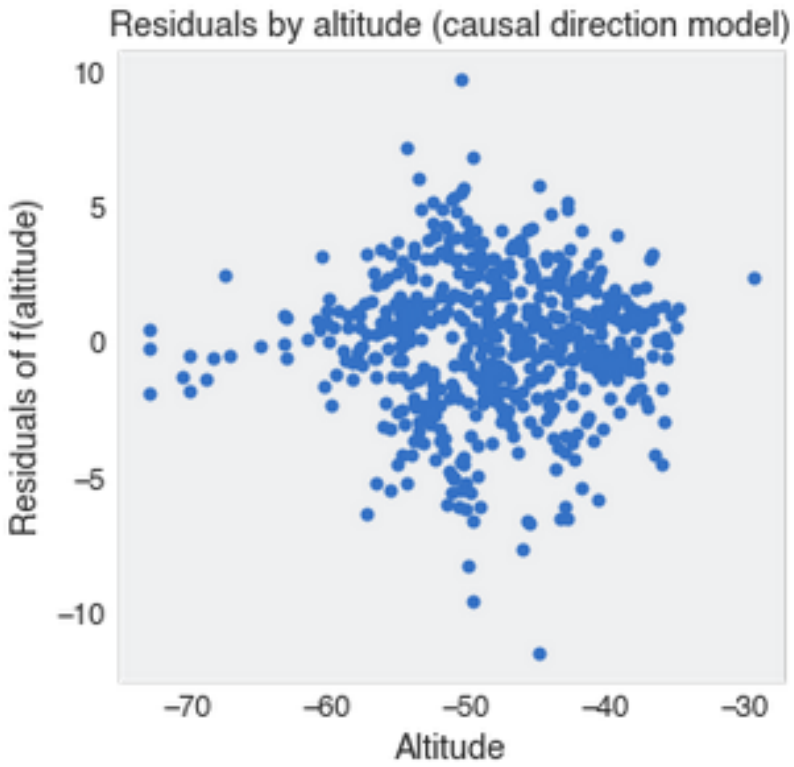


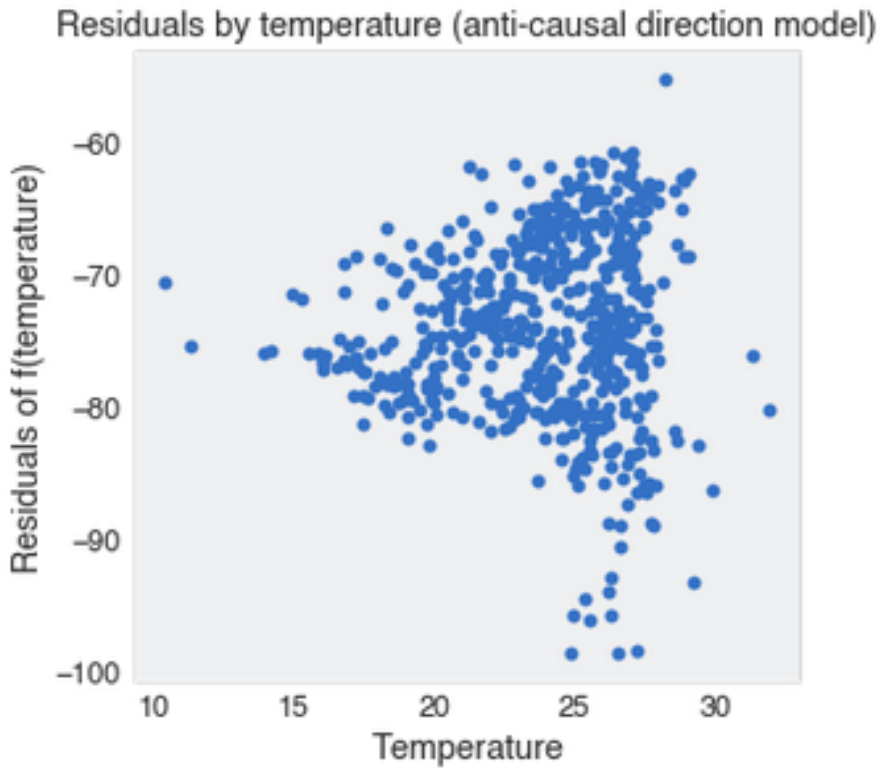
- ANM is misspecified!
- $\implies$ won't fit properly $\implies$ bad predictive model
- Guarantees don't hold anymore!

Additive Noise Model: Another Example

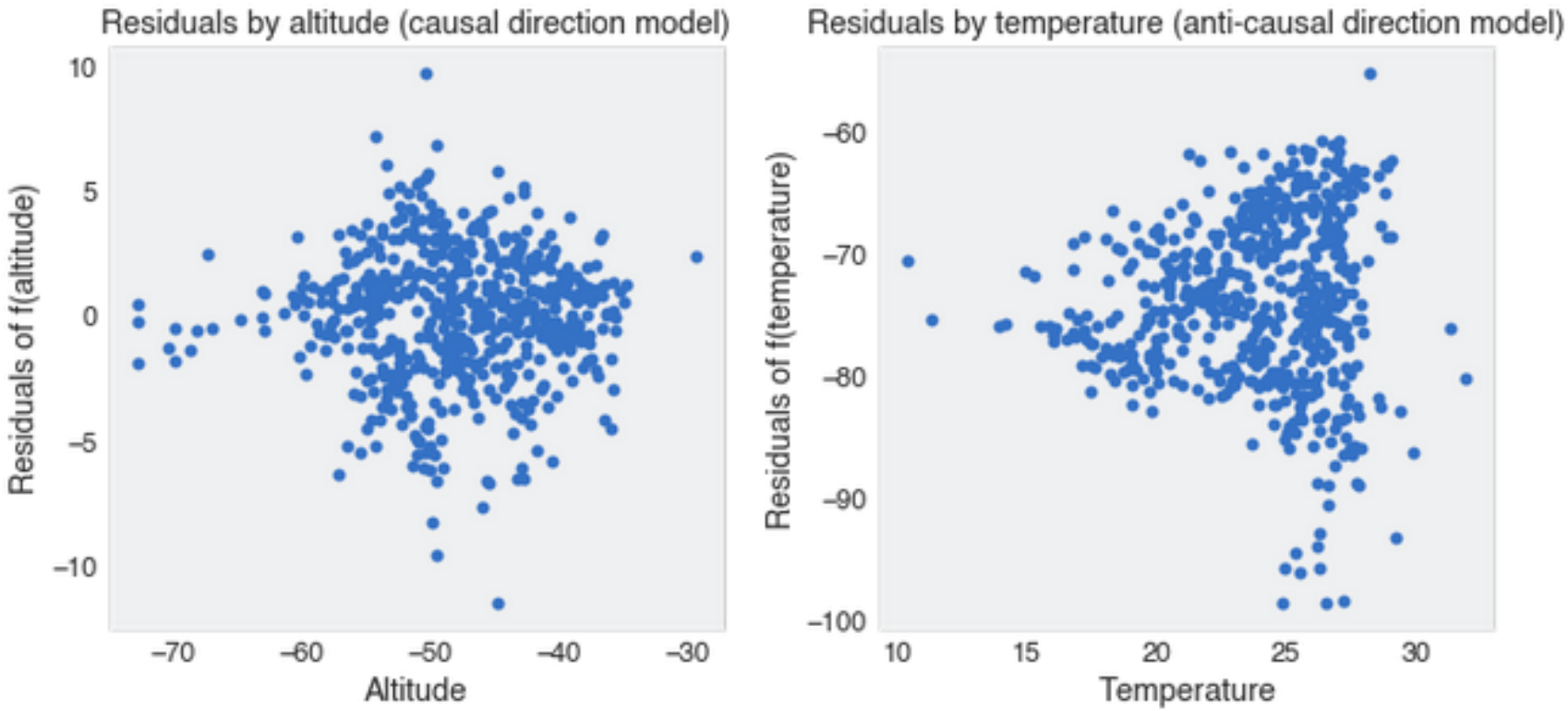


Additive Noise Model: Another Example



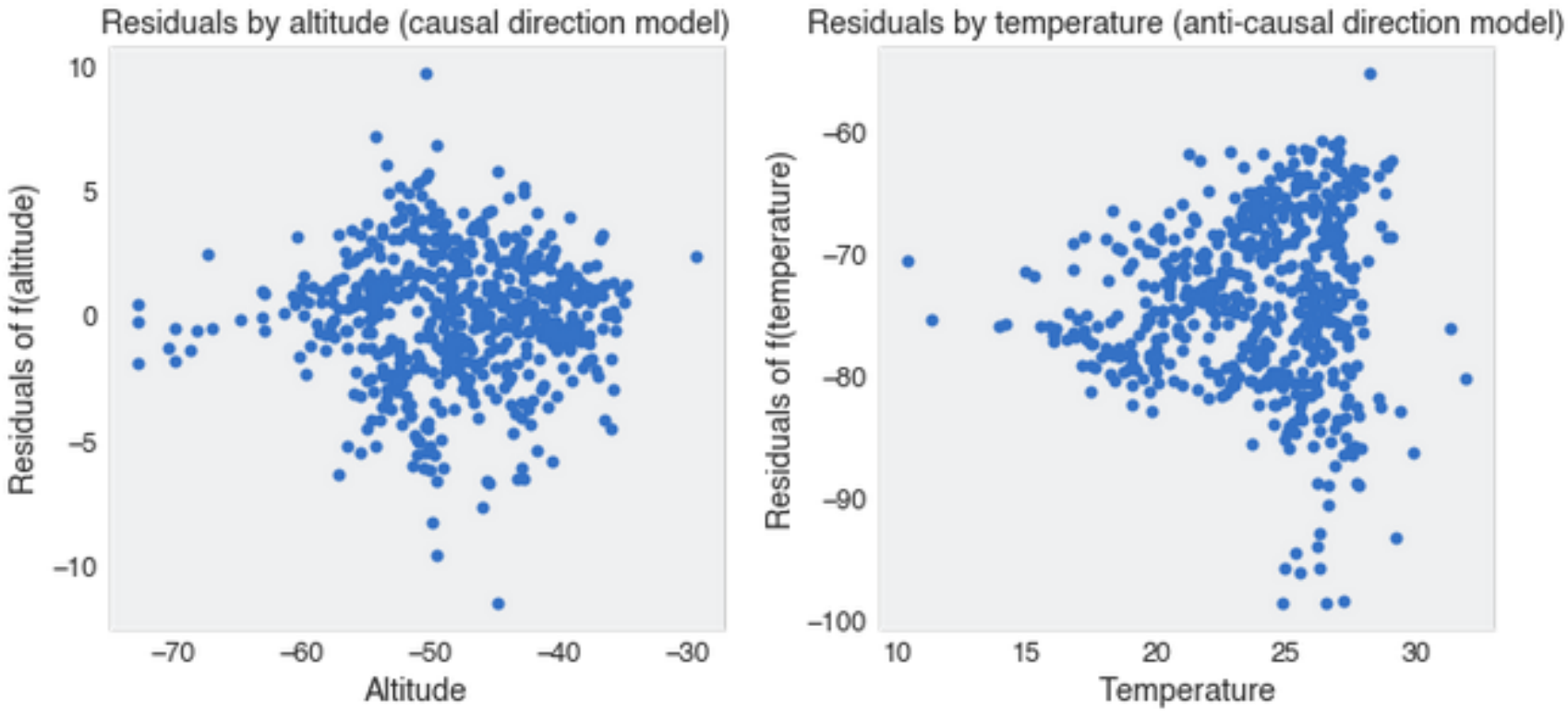


Additive Noise Model: Another Example



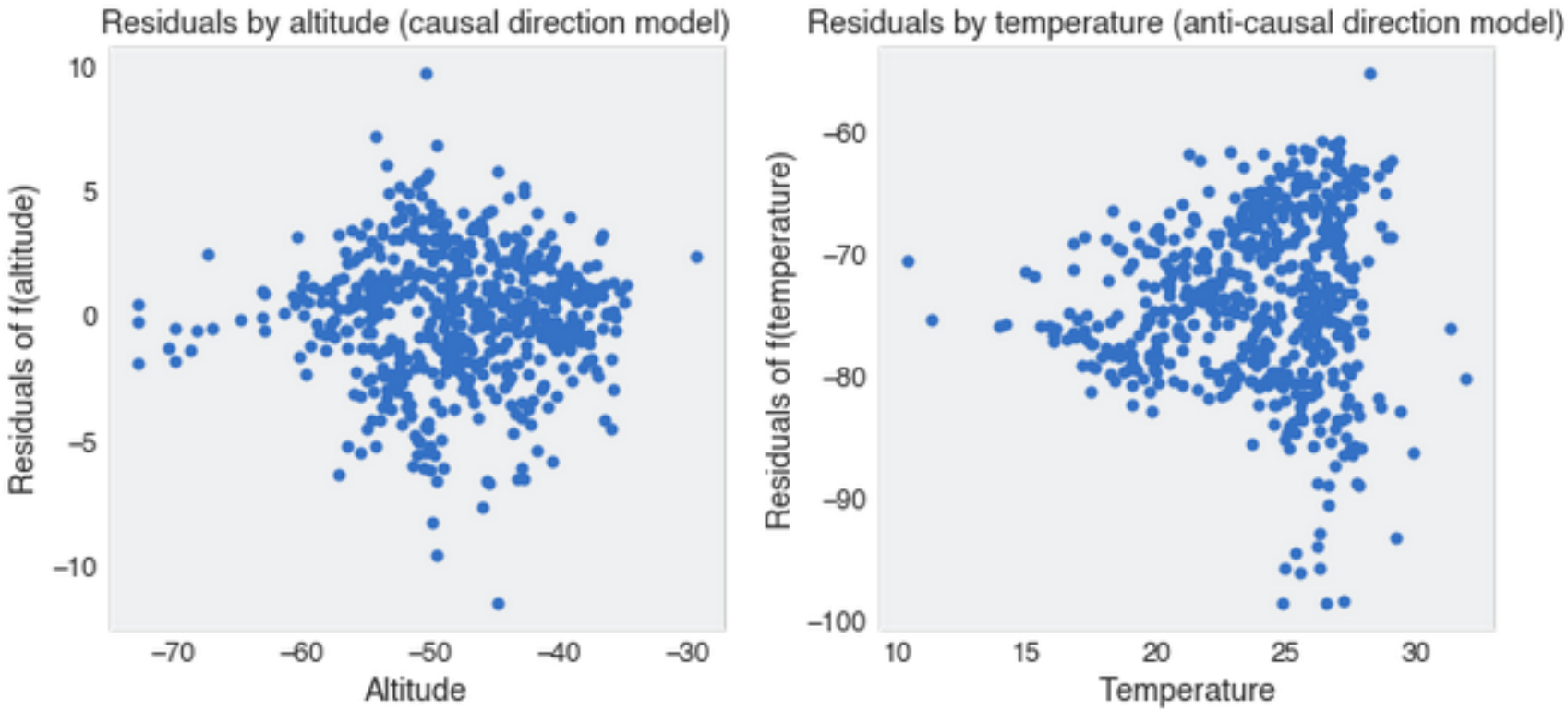
- One direction needs multimodal noise

Additive Noise Model: Another Example



- One direction needs multimodal noise
- But noise in correct causal direction not independent of cause

Additive Noise Model: Another Example



- One direction needs multimodal noise
- But noise in correct causal direction not independent of cause
- Still, one direction more complex than the other?

Complexity Metrics for Causality

The relation between simplicity and causality has been noted and used before. We have methods that

- are inspired by Kolmogorov Complexity,
- are inspired by Minimum Description Length,
- count parameters.

Complexity Metrics for Causality

The relation between simplicity and causality has been noted and used before. We have methods that

- are inspired by Kolmogorov Complexity,
- are inspired by Minimum Description Length,
- count parameters.

We show how Occam's razor from Bayesian inference can elegantly be used to discover causal direction.

Complexity Metrics for Causality

The relation between simplicity and causality has been noted and used before. We have methods that

- are inspired by Kolmogorov Complexity,
- are inspired by Minimum Description Length,
- count parameters.

We show how Occam's razor from Bayesian inference can elegantly be used to discover causal direction.

- Idea has been floated a few times before (will discuss later)

Complexity Metrics for Causality

The relation between simplicity and causality has been noted and used before. We have methods that

- are inspired by Kolmogorov Complexity,
- are inspired by Minimum Description Length,
- count parameters.

We show how Occam's razor from Bayesian inference can elegantly be used to discover causal direction.

- Idea has been floated a few times before (will discuss later)
- Bayes needs no more assumptions than Kolmogorov / MDL

Complexity Metrics for Causality

The relation between simplicity and causality has been noted and used before. We have methods that

- are inspired by Kolmogorov Complexity,
- are inspired by Minimum Description Length,
- count parameters.

We show how Occam's razor from Bayesian inference can elegantly be used to discover causal direction.

- Idea has been floated a few times before (will discuss later)
- Bayes needs no more assumptions than Kolmogorov / MDL
- ... and has the same (lack) of guarantees

Overview

Causality: Overview and Background

Causality: Approaches and Problems

Bayesian Model Selection for Causality and its Properties

An Actual Method

Discussion & Conclusion

Bayesian Model Selection for Causal Models

Bayes says: Just find the posterior over the causal direction

Bayesian Model Selection for Causal Models

Bayes says: Just find the posterior over the causal direction

$$p(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y})}{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y}) + p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}})} \tag{5}$$

Bayesian Model Selection for Causal Models

Bayes says: Just find the posterior over the causal direction

$$p(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y})}{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y}) + p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}})} \tag{5}$$

Since we only have two options, we can summarise with

$$\log \frac{p(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D})}{p(\mathcal{M}_{X\leftarrow\mathbf{Y}}|\mathcal{D})} = \log \frac{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y})}{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}})} \tag{6}$$

Bayesian Model Selection for Causal Models

Bayes says: Just find the posterior over the causal direction

$$p(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y})}{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y}) + p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}})} \tag{5}$$

Since we only have two options, we can summarise with

$$\log \frac{p(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D})}{p(\mathcal{M}_{X\leftarrow\mathbf{Y}}|\mathcal{D})} = \log \frac{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y})}{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}})} \tag{6}$$

Bayes says: *Must* specify a prior on which direction is more likely.

Bayesian Model Selection for Causal Models

Bayes says: Just find the posterior over the causal direction

$$p(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y})}{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y}) + p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}})} \tag{5}$$

Since we only have two options, we can summarise with

$$\log \frac{p(\mathcal{M}_{\mathbf{X}\to Y}|\mathcal{D})}{p(\mathcal{M}_{X\leftarrow\mathbf{Y}}|\mathcal{D})} = \log \frac{p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})p(\mathcal{M}_{\mathbf{X}\to Y})}{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathcal{M}_{X\leftarrow\mathbf{Y}})} \tag{6}$$

Bayes says: *Must* specify a prior on which direction is more likely.
$\implies$ We must be indifferent, so choose 0.5.

Bayesian Model Selection for Causal Models

So we just need to compute the **marginal likelihoods**

$$p(\mathbf{x},\mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = \iint p(\mathbf{x}|\phi,\mathcal{M}_{\mathbf{X}\to Y})p(\mathbf{y}|\mathbf{x},\theta,\mathcal{M}_{\mathbf{X}\to Y})p(\phi,\theta|\mathcal{M}_{\mathbf{X}\to Y})\mathrm{d}\phi\mathrm{d}\theta\,,$$

$$p(\mathbf{x},\mathbf{y}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) = \iint p(\mathbf{y}|\phi,\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathbf{x}|\mathbf{y},\theta,\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\phi,\theta|\mathcal{M}_{X\leftarrow\mathbf{Y}})\mathrm{d}\phi\mathrm{d}\theta\,.$$

Bayesian Model Selection for Causal Models

So we just need to compute the **marginal likelihoods**

$$p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = \iint p(\mathbf{x}|\phi, \mathcal{M}_{\mathbf{X}\to Y})p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}_{\mathbf{X}\to Y})p(\phi, \theta|\mathcal{M}_{\mathbf{X}\to Y})\mathrm{d}\phi\mathrm{d}\theta\,,$$

$$p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) = \iint p(\mathbf{y}|\phi, \mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathbf{x}|\mathbf{y}, \theta, \mathcal{M}_{X\leftarrow\mathbf{Y}})p(\phi, \theta|\mathcal{M}_{X\leftarrow\mathbf{Y}})\mathrm{d}\phi\mathrm{d}\theta\,.$$

- **Bayes says**: *Must* specify priors on ϕ, θ.

Bayesian Model Selection for Causal Models

So we just need to compute the **marginal likelihoods**

$$p(\mathbf{x},\mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = \iint p(\mathbf{x}|\phi,\mathcal{M}_{\mathbf{X}\to Y})p(\mathbf{y}|\mathbf{x},\theta,\mathcal{M}_{\mathbf{X}\to Y})p(\phi,\theta|\mathcal{M}_{\mathbf{X}\to Y})\mathrm{d}\phi\mathrm{d}\theta\,,$$

$$p(\mathbf{x},\mathbf{y}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) = \iint p(\mathbf{y}|\phi,\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathbf{x}|\mathbf{y},\theta,\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\phi,\theta|\mathcal{M}_{X\leftarrow\mathbf{Y}})\mathrm{d}\phi\mathrm{d}\theta\,.$$

- **Bayes says**: *Must* specify priors on ϕ,θ.
- Information on distribution on causes should not provide information on distribution of effect given cause (strict ICM).

Bayesian Model Selection for Causal Models

So we just need to compute the **marginal likelihoods**

$$p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = \iint p(\mathbf{x}|\phi, \mathcal{M}_{\mathbf{X}\to Y}) p(\mathbf{y}|\mathbf{x}, \theta, \mathcal{M}_{\mathbf{X}\to Y}) p(\phi, \theta|\mathcal{M}_{\mathbf{X}\to Y}) \mathrm{d}\phi \mathrm{d}\theta \,,$$

$$p(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) = \iint p(\mathbf{y}|\phi, \mathcal{M}_{X\leftarrow\mathbf{Y}}) p(\mathbf{x}|\mathbf{y}, \theta, \mathcal{M}_{X\leftarrow\mathbf{Y}}) p(\phi, \theta|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \mathrm{d}\phi \mathrm{d}\theta \,.$$

- **Bayes says**: *Must* specify priors on ϕ, θ.
- Information on distribution on causes should not provide information on distribution of effect given cause (strict ICM).
  $\implies \phi \perp\!\!\!\perp \theta$

Bayesian Model Selection for Causal Models

So we just need to compute the **marginal likelihoods**

$$p(\mathbf{x},\mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = \iint p(\mathbf{x}|\phi,\mathcal{M}_{\mathbf{X}\to Y})p(\mathbf{y}|\mathbf{x},\theta,\mathcal{M}_{\mathbf{X}\to Y})p(\phi,\theta|\mathcal{M}_{\mathbf{X}\to Y})\mathrm{d}\phi\mathrm{d}\theta\,,$$

$$p(\mathbf{x},\mathbf{y}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) = \iint p(\mathbf{y}|\phi,\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\mathbf{x}|\mathbf{y},\theta,\mathcal{M}_{X\leftarrow\mathbf{Y}})p(\phi,\theta|\mathcal{M}_{X\leftarrow\mathbf{Y}})\mathrm{d}\phi\mathrm{d}\theta\,.$$

- **Bayes says**: *Must* specify priors on ϕ, θ.
- Information on distribution on causes should not provide information on distribution of effect given cause (strict ICM). $\implies \phi \perp\!\!\!\perp \theta$
- Consistent with earlier constraint approaches: Zero mass in prior.

Can BMS Distinguish Causal Models?

MaxLik could not distinguish between causal models

Can BMS Distinguish Causal Models?

MaxLik could not distinguish between causal models

$\Longrightarrow$ Does BMS have the same problem?

Can BMS Distinguish Causal Models?

MaxLik could not distinguish between causal models
$\implies$ Does BMS have the same problem?

For BMS to be indifferent, we need

$$p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X\leftarrow \mathbf{Y}}) \qquad \forall \mathbf{x}, \mathbf{y} \tag{7}$$

Can BMS Distinguish Causal Models?

MaxLik could not distinguish between causal models
$\implies$ Does BMS have the same problem?

For BMS to be indifferent, we need

$$p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X\leftarrow \mathbf{Y}}) \qquad \forall \mathbf{x}, \mathbf{y} \tag{7}$$

- Theorem in the paper says that as $N \to \infty$ this **only** happens in very specific circumstances.

Can BMS Distinguish Causal Models?

MaxLik could not distinguish between causal models
$\implies$ Does BMS have the same problem?

For BMS to be indifferent, we need

$$p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X\leftarrow \mathbf{Y}}) \qquad \forall \mathbf{x}, \mathbf{y} \tag{7}$$

- Theorem in the paper says that as $N \to \infty$ this **only** happens in very specific circumstances.
- Example: Normalised linear models.

Can BMS Distinguish Causal Models?

MaxLik could not distinguish between causal models
$\implies$ Does BMS have the same problem?

For BMS to be indifferent, we need

$$p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X\leftarrow \mathbf{Y}}) \qquad \forall \mathbf{x}, \mathbf{y} \tag{7}$$

- Theorem in the paper says that as $N \to \infty$ this **only** happens in very specific circumstances.
- Example: Normalised linear models.
- So, MaxLik has no opinion, but BMS does.

Can BMS Distinguish Causal Models?

MaxLik could not distinguish between causal models
$\implies$ Does BMS have the same problem?

For BMS to be indifferent, we need

$$p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X\leftarrow \mathbf{Y}}) \qquad \forall \mathbf{x}, \mathbf{y} \tag{7}$$

- Theorem in the paper says that as $N \to \infty$ this **only** happens in very specific circumstances.
- Example: Normalised linear models.
- So, MaxLik has no opinion, but BMS does.
- Does not show that BMS finds the **correct** causal direction,

Can BMS Distinguish Causal Models?

MaxLik could not distinguish between causal models
$\implies$ Does BMS have the same problem?

For BMS to be indifferent, we need

$$p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{\mathbf{X}\to Y}) = p_{X,Y}(\mathbf{x}, \mathbf{y}|\mathcal{M}_{X\leftarrow \mathbf{Y}}) \qquad \forall \mathbf{x}, \mathbf{y} \tag{7}$$

- Theorem in the paper says that as $N \to \infty$ this **only** happens in very specific circumstances.
- Example: Normalised linear models.
- So, MaxLik has no opinion, but BMS does.
- Does not show that BMS finds the **correct** causal direction,
- but does show that BMS does not fail in the way MaxLik does.

Does BMS find the correct causal direction?

Cannot prove strict correctness.

Does BMS find the correct causal direction?

Cannot prove strict correctness. Instead, quantify probability.

Does BMS find the correct causal direction?

Cannot prove strict correctness. Instead, quantify probability.

- Must assume datasets are sampled from some $\Pi(X, Y)$.

Does BMS find the correct causal direction?

Cannot prove strict correctness. Instead, quantify probability.

- Must assume datasets are sampled from some $\Pi(X, Y)$.
- Begin by assuming that our models' priors allow one causal direction to match $\Pi(X, Y)$ (no misspecification)

Does BMS find the correct causal direction?

Cannot prove strict correctness. Instead, quantify probability.

- Must assume datasets are sampled from some $\Pi(X, Y)$.
- Begin by assuming that our models' priors allow one causal direction to match $\Pi(X, Y)$ (no misspecification)

Decision rule (optimal for symmetric loss):

$$\mathcal{M}^* = \begin{cases} \mathcal{M}_{\mathbf{X}\to Y} \text{ if } p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) > p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \\ \mathcal{M}_{X\leftarrow\mathbf{Y}} \text{ if } p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) < p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \end{cases}, \tag{8}$$

Does BMS find the correct causal direction?

Cannot prove strict correctness. Instead, quantify probability.

- Must assume datasets are sampled from some $\Pi(X, Y)$.
- Begin by assuming that our models' priors allow one causal direction to match $\Pi(X, Y)$ (no misspecification)

Decision rule (optimal for symmetric loss):

$$\mathcal{M}^* = \begin{cases} \mathcal{M}_{\mathbf{X}\to Y} \text{ if } p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) > p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \\ \mathcal{M}_{X\leftarrow\mathbf{Y}} \text{ if } p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) < p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \end{cases}, \tag{8}$$

$$P(E|\mathcal{M}_{\mathbf{X}\to Y}) = \int_{\mathcal{R}_Y} p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) \mathbb{1}\{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) > p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\} \mathrm{d}\mathcal{D}$$

Does BMS find the correct causal direction?

Cannot prove strict correctness. Instead, quantify probability.

- Must assume datasets are sampled from some $\Pi(X, Y)$.
- Begin by assuming that our models' priors allow one causal direction to match $\Pi(X, Y)$ (no misspecification)

Decision rule (optimal for symmetric loss):

$$\mathcal{M}^* = \begin{cases} \mathcal{M}_{\mathbf{X}\to Y} \text{ if } p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) > p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \\ \mathcal{M}_{X\leftarrow\mathbf{Y}} \text{ if } p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) < p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) \end{cases}, \tag{8}$$

$$P(E|\mathcal{M}_{\mathbf{X}\to Y}) = \int_{\mathcal{R}_Y} p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) \mathbb{1}\{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) > p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\} \mathrm{d}\mathcal{D}$$

Probability of error in both causal directions are equal:

$$\begin{aligned} P(E) &= P(E|\mathcal{M}_{\mathbf{X}\to Y})P(\mathcal{M}_{\mathbf{X}\to Y}) + P(E|\mathcal{M}_{X\leftarrow\mathbf{Y}})P(\mathcal{M}_{X\leftarrow\mathbf{Y}}) = P(E|\mathcal{M}_{\mathbf{X}\to Y}) \\ &= \frac{1}{2}(1 - \mathrm{TV}[P_{\mathcal{D}}(\cdot|\mathcal{M}_{\mathbf{X}\to Y}), P_{\mathcal{D}}(\cdot|\mathcal{M}_{X\leftarrow\mathbf{Y}})]) \end{aligned}$$

BMS under Model Misspecification

What if we don't have the "true" priors?

BMS under Model Misspecification

What if we don't have the "true" priors?

$$\Pi(E|\mathcal{M}_{\mathbf{X}\to Y}) = \int_{\mathcal{R}_Y} \pi(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) \mathbb{1}\{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) > p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\} \mathrm{d}\mathcal{D}$$

BMS under Model Misspecification

What if we don't have the "true" priors?

$$\Pi(E|\mathcal{M}_{\mathbf{X}\to Y}) = \int_{\mathcal{R}_Y} \pi(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) \mathbb{1}\{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) > p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\} \mathrm{d}\mathcal{D}$$

$$|\Pi(\text{Error}) - P(\text{Error})| \leqslant \text{TV}[\Pi_{\mathcal{D}}(\cdot|\mathbf{X}\to Y), P_{\mathcal{D}}(\cdot|\mathcal{M}_{\mathbf{X}\to Y})] \tag{9}$$

$$= \frac{1}{2}\int |\pi(\mathcal{D}|\mathbf{X}\to Y) - p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})| \mathrm{d}\mathcal{D}, \tag{10}$$

BMS under Model Misspecification

What if we don't have the "true" priors?

$$\Pi(E|\mathcal{M}_{\mathbf{X}\to Y}) = \int_{\mathcal{R}_Y} \pi(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\mathbb{1}\{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) > p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\}\mathrm{d}\mathcal{D}$$

$$|\Pi(\text{Error}) - P(\text{Error})| \leqslant \text{TV}[\Pi_{\mathcal{D}}(\cdot|\mathbf{X}\to Y), P_{\mathcal{D}}(\cdot|\mathcal{M}_{\mathbf{X}\to Y})] \tag{9}$$

$$= \frac{1}{2}\int |\pi(\mathcal{D}|\mathbf{X}\to Y) - p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})|\mathrm{d}\mathcal{D}, \tag{10}$$

- Hard to verify that priors are right.

BMS under Model Misspecification

What if we don't have the "true" priors?

$$\Pi(E|\mathcal{M}_{\mathbf{X}\to Y}) = \int_{\mathcal{R}_Y} \pi(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y}) \mathbb{1}\{p(\mathcal{D}|\mathcal{M}_{X\leftarrow\mathbf{Y}}) > p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})\} \mathrm{d}\mathcal{D}$$

$$|\Pi(\text{Error}) - P(\text{Error})| \leqslant \text{TV}[\Pi_{\mathcal{D}}(\cdot|\mathbf{X}\to Y), P_{\mathcal{D}}(\cdot|\mathcal{M}_{\mathbf{X}\to Y})] \tag{9}$$

$$= \frac{1}{2}\int |\pi(\mathcal{D}|\mathbf{X}\to Y) - p(\mathcal{D}|\mathcal{M}_{\mathbf{X}\to Y})| \mathrm{d}\mathcal{D}, \tag{10}$$

- Hard to verify that priors are right.
- But at least it shows that there is a limit to the brittleness.

Overview

Causality: Overview and Background

Causality: Approaches and Problems

Bayesian Model Selection for Causality and its Properties

An Actual Method

Discussion & Conclusion

A More Realistic Model

Want to specify a flexible model with large $\mathcal{R}, \mathcal{C}$ to minimise model misspecification $\implies$ we want good predictive ability.

1

A More Realistic Model

Want to specify a flexible model with large $\mathcal{R}, \mathcal{C}$ to minimise model misspecification $\implies$ we want good predictive ability.

- Conditional GPLVM for $\mathcal{C}$, normal GPLVM for $\mathcal{R}$:

1

A More Realistic Model

Want to specify a flexible model with large $\mathcal{R}, \mathcal{C}$ to minimise model misspecification $\implies$ we want good predictive ability.

- Conditional GPLVM for $\mathcal{C}$, normal GPLVM for $\mathcal{R}$:

$$p(y_i|x_i, f, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(y_i|f(x_i, w_i), x_i, w_i, \mathcal{M}_{\mathbf{X}\to Y}) p(w_i) \mathrm{d}w_i \,,$$

$$p(x_i|g, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(x_i|g(v_i), v_i, \mathcal{M}_{\mathbf{X}\to Y}) p(v_i) \mathrm{d}v_i \,,$$

1

A More Realistic Model

Want to specify a flexible model with large $\mathcal{R}, \mathcal{C}$ to minimise model misspecification $\implies$ we want good predictive ability.

- Conditional GPLVM for $\mathcal{C}$, normal GPLVM for $\mathcal{R}$:

$$p(y_i|x_i, f, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(y_i|f(x_i, w_i), x_i, w_i, \mathcal{M}_{\mathbf{X}\to Y}) p(w_i) \mathrm{d}w_i \,,$$

$$p(x_i|g, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(x_i|g(v_i), v_i, \mathcal{M}_{\mathbf{X}\to Y}) p(v_i) \mathrm{d}v_i \,,$$

- These are basically Bayesian VAEs[1], but with $f \sim \mathcal{GP}$, $g \sim \mathcal{GP}$

[1] Normal VAEs have a point estimate over the density that is to be estimated! A Bayesian VAE has a full posterior.

A More Realistic Model

Want to specify a flexible model with large $\mathcal{R}, \mathcal{C}$ to minimise model misspecification $\implies$ we want good predictive ability.

- Conditional GPLVM for $\mathcal{C}$, normal GPLVM for $\mathcal{R}$:

$$p(y_i|x_i, f, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(y_i|f(x_i, w_i), x_i, w_i, \mathcal{M}_{\mathbf{X}\to Y}) p(w_i) \mathrm{d}w_i \,,$$

$$p(x_i|g, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(x_i|g(v_i), v_i, \mathcal{M}_{\mathbf{X}\to Y}) p(v_i) \mathrm{d}v_i \,,$$

- These are basically Bayesian VAEs[1], but with $f \sim \mathcal{GP}$, $g \sim \mathcal{GP}$
- Very flexible density estimators! There exists a f, g that can perfectly model any joint in both causal directions!

[1] Normal VAEs have a point estimate over the density that is to be estimated! A Bayesian VAE has a full posterior.

A More Realistic Model

Want to specify a flexible model with large $\mathcal{R}, \mathcal{C}$ to minimise model misspecification $\implies$ we want good predictive ability.

- Conditional GPLVM for $\mathcal{C}$, normal GPLVM for $\mathcal{R}$:

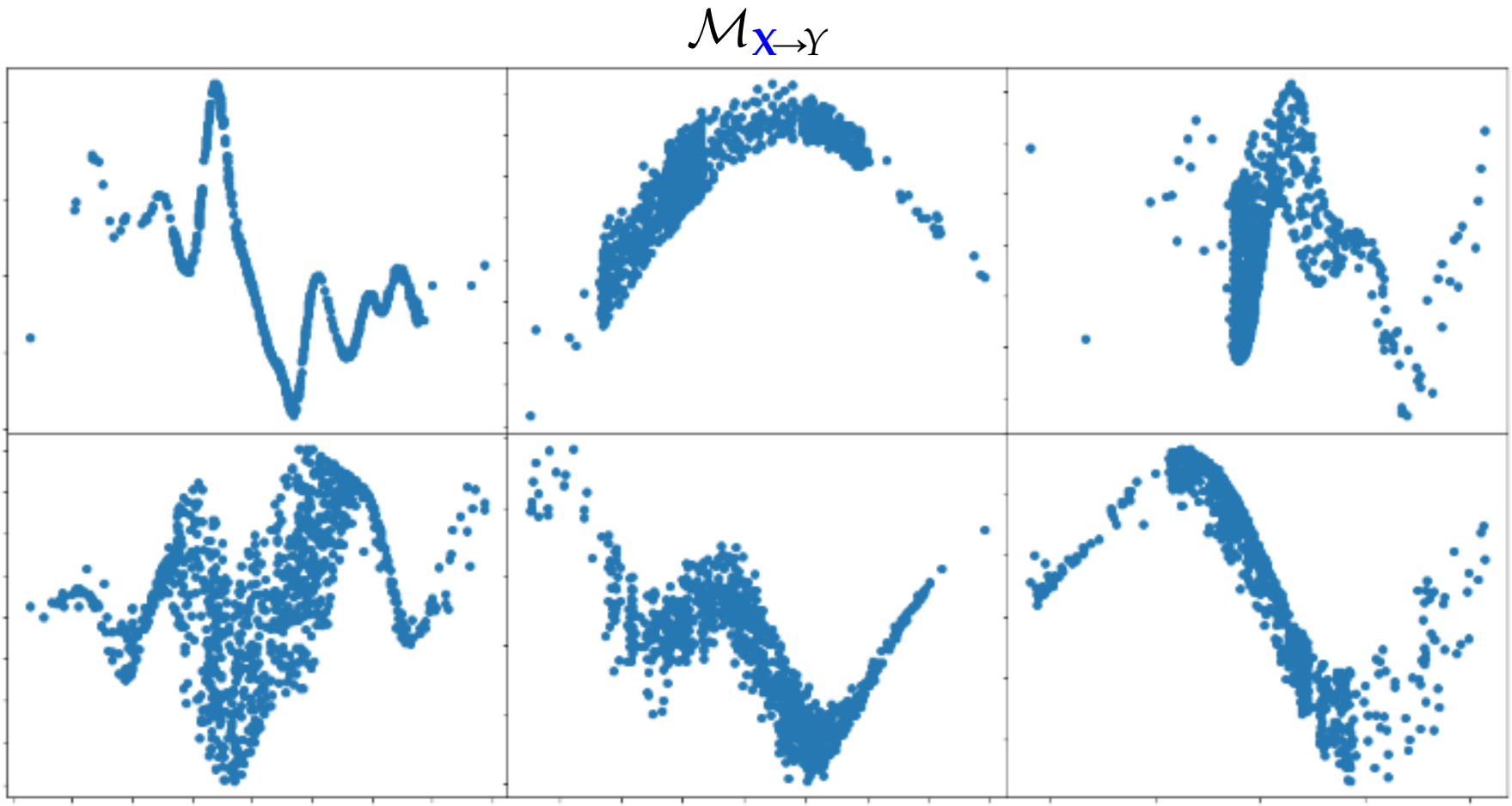
$$p(y_i|x_i, f, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(y_i|f(x_i, w_i), x_i, w_i, \mathcal{M}_{\mathbf{X}\to Y}) p(w_i) \mathrm{d}w_i \,,$$

$$p(x_i|g, \mathcal{M}_{\mathbf{X}\to Y}) = \int p(x_i|g(v_i), v_i, \mathcal{M}_{\mathbf{X}\to Y}) p(v_i) \mathrm{d}v_i \,,$$

- These are basically Bayesian VAEs[1], but with $f \sim \mathcal{GP}$, $g \sim \mathcal{GP}$
- Very flexible density estimators! There exists a f, g that can perfectly model any joint in both causal directions!
- We use (existing) Variational Inference to approximate the marginal likelihoods (ELBO)

[1] Normal VAEs have a point estimate over the density that is to be estimated! A Bayesian VAE has a full posterior.

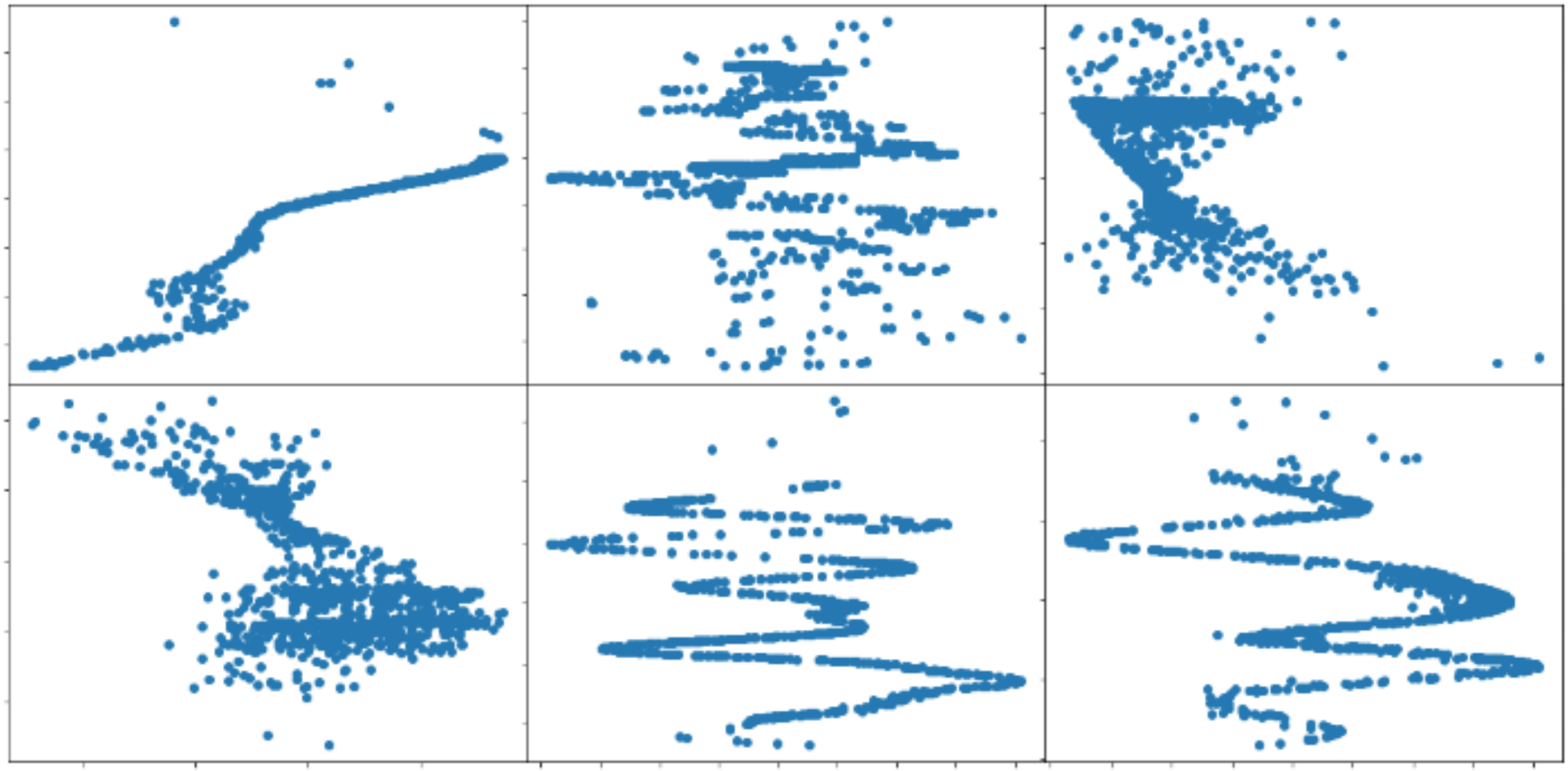
Probability of Error



Can use Monte Carlo to estimate probability of error (MC est = 0%)

Probability of Error

$\mathcal{M}_{X \leftarrow Y}$



Can use Monte Carlo to estimate probability of error (MC est = 0%)

Experiments: ANM data

- How does the non-fully-identifiable model work on fully-identifiable data?
- How does this compare to a fully-identifiable model?

Methods	ANM
Gaussian Process	100.0
GPLVM	100.0

Table: ROC AUC scores for identifying causal direction of datastes generated by an ANM (higher is better).

$\Longrightarrow$ Addition of flexibility does not (detectably) hurt performance on "easy" datasets.

Experiments: Real & Synthetic Data

Are our priors (i.e. assumptions) good?

Experiments: Real & Synthetic Data

Are our priors (i.e. assumptions) good?

- Can **only** determine based on empirical evaluation.

Experiments: Real & Synthetic Data

Are our priors (i.e. assumptions) good?

- Can **only** determine based on empirical evaluation.
- Not that different to existing methods, which are applied in situations where assumptions are broken!

Experiments: Real & Synthetic Data

Are our priors (i.e. assumptions) good?

- Can **only** determine based on empirical evaluation.
- Not that different to existing methods, which are applied in situations where assumptions are broken!

Methods	CE-Cha	CE-Multi	CE-Net	CE-Gauss	CE-Tueb
CGNN	76.2	94.7	86.3	89.3	76.6
GPI	71.5	73.8	88.1	90.2	70.6
PNL	78.6	51.7	75.6	84.7	73.8
ANM	43.7	25.5	87.8	90.7	63.9
IGCI	55.6	77.8	57.4	16.0	63.1
LiNGAM	57.8	62.3	3.3	72.2	31.1
RECI	59.0	94.7	66.0	71.0	70.5
CCS	69.3	96.0	89.7	90.5	N/A
CHD	72.0	97.6	90.5	91.4	N/A
CKL	69.8	95.5	89.3	91.0	N/A
CKM	69.7	90.6	94.3	91.6	N/A
CTV	72.2	95.8	91.9	**91.8**	N/A
GPLVM	**82.1**	**97.7**	**98.8**	90.2	**78.3**

Overview

Causality: Overview and Background

Causality: Approaches and Problems

Bayesian Model Selection for Causality and its Properties

An Actual Method

Discussion & Conclusion

Discussion

Prior says 0% error. But we get higher. Why?

Discussion

Prior says 0% error. But we get higher. Why?

- Prior could be misspecified (but still better than other methods!)

Discussion

Prior says 0% error. But we get higher. Why?

- Prior could be misspecified (but still better than other methods!)
- E.g. more ambiguous datasets than expected by prior

Discussion

Prior says 0% error. But we get higher. Why?

- Prior could be misspecified (but still better than other methods!)
- E.g. more ambiguous datasets than expected by prior
- Variational inference could be inaccurate in some cases

Origins of Bayes for Causality

Discuss. References in paper.

Conclusions

- Causal discovery requires assumptions.

Conclusions

- Causal discovery requires assumptions.
- Bayes offers a way of encoding assumptions.

Conclusions

- Causal discovery requires assumptions.
- Bayes offers a way of encoding assumptions.
- In fact, **tells** you where to make assumptions!

Conclusions

- Causal discovery requires assumptions.
- Bayes offers a way of encoding assumptions.
- In fact, **tells** you where to make assumptions!
- Naturally guides you to a method.

Conclusions

- Causal discovery requires assumptions.
- Bayes offers a way of encoding assumptions.
- In fact, **tells** you where to make assumptions!
- Naturally guides you to a method.
- Allows specification of realistic assumptions.

Conclusions

- Causal discovery requires assumptions.
- Bayes offers a way of encoding assumptions.
- In fact, **tells** you where to make assumptions!
- Naturally guides you to a method.
- Allows specification of realistic assumptions.
- Realistic assumptions lead to good performance.

Join us!


Anish Dhir


Artem Artemev


Jose Pablo Folch


Ruby Sedgwick


Seth Nabarro


Tycho van der Ouderaa

- I'm looking for PhD candidates (in Oxford from September)
- Check my website (`https://mvdw.uk/`) for tips on applying, and how to get in touch.
- And do find me to chat if you have questions.