# Bivariate Causal Discovery using Bayesian Model Selection

Mark van der Wilk

ISBA World Meeting 2024

University of Oxford
Department of Computer Science

https://mvdw.uk
@markvanderwilk

3 July 2024

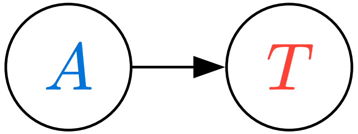
# Based on a True Story (ICML 2024)

## Bivariate Causal Discovery using Bayesian Model Selection

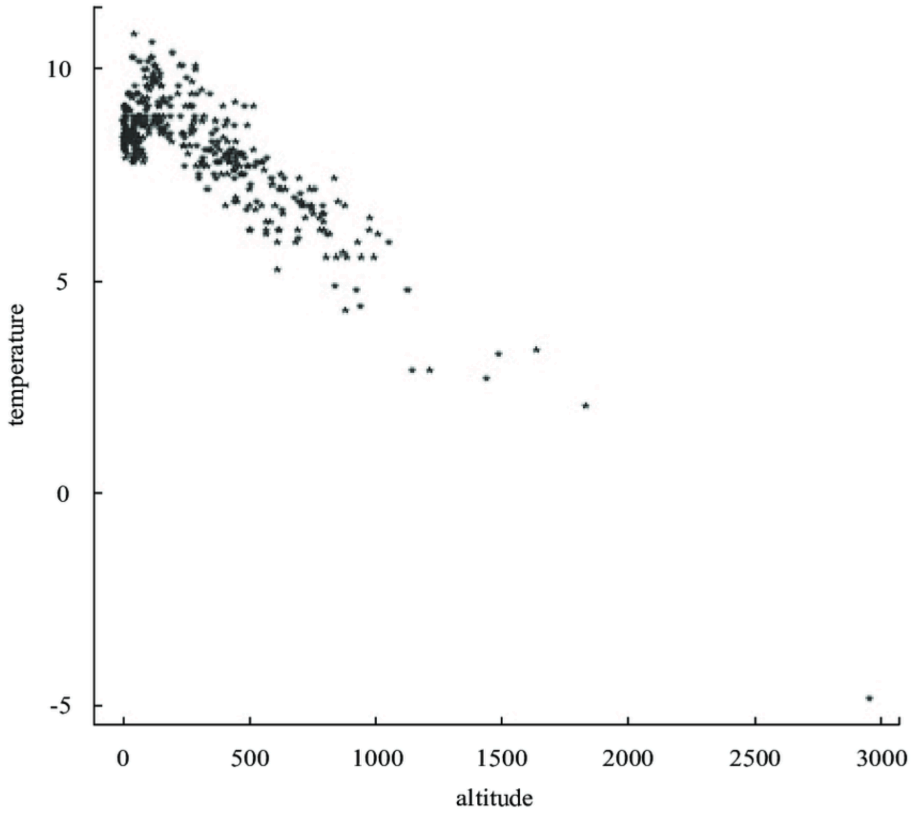
**Anish Dhir** [1] **Samuel Power** [2] **Mark van der Wilk** [3]

# Background & Problem Setting

# Why is Causality Important?




- Conditional $p(a|t)$ assumes pair $(a, t)$ sampled *jointly* from the same distribution!
- When intervening, **you cannot affect your cause**!
- Intervention breaks links to ancestors, so $p(a|\mathrm{do}(t)) = p(a)$.
- But... $p(t|\mathrm{do}(a)) = p(t|a)$.



# Two Causal Directions, Two Models

$$A \rightarrow T$$

$$p(t, a|\varphi, \theta) = p(a|t, \theta)p(t|\varphi)$$

$$p(a|t, \theta) : \text{cond. density model}$$

$$p(t|\varphi) : \text{density model}$$

$$A \leftarrow T$$

$$p(t, a|\varphi, \theta) = p(t|a, \theta)p(a|\varphi)$$

$$p(t|a, \theta) : \text{cond. density model}$$

$$p(a|\varphi) : \text{density model}$$

# Two Causal Directions, Two Models

$$A \longrightarrow T$$

$$p(t, a|\varphi, \theta) = p(a|t, \theta)p(t|\varphi)$$

$$p(a|t, \theta) : \text{cond. density model}$$

$$p(t|\varphi) : \text{density model}$$

$$A \longleftarrow T$$

$$p(t, a|\varphi, \theta) = p(t|a, \theta)p(a|\varphi)$$

$$p(t|a, \theta) : \text{cond. density model}$$

$$p(a|\varphi) : \text{density model}$$

**Causality should determine model structure**

We want sensible results if we apply intervention rules to our *model*!

# Two Causal Directions, Two Models

$$A \rightarrow T$$

$$p(t, a|\varphi, \theta) = p(a|t, \theta)p(t|\varphi)$$

$$p(a|t, \theta) : \text{cond. density model}$$

$$p(t|\varphi) : \text{density model}$$

$$A \leftarrow T$$

$$p(t, a|\varphi, \theta) = p(t|a, \theta)p(a|\varphi)$$

$$p(t|a, \theta) : \text{cond. density model}$$

$$p(a|\varphi) : \text{density model}$$

**Causality should determine model structure**

We want sensible results if we apply intervention rules to our *model*!

**Goal: Predict causal structure from observational data.**

# Two Causal Directions, Two Models

$$A \rightarrow T$$

$$p(t, a|\varphi, \theta) = p(a|t, \theta)p(t|\varphi)$$

$$p(a|t, \theta) : \text{cond. density model}$$

$$p(t|\varphi) : \text{density model}$$

$$A \leftarrow T$$

$$p(t, a|\varphi, \theta) = p(t|a, \theta)p(a|\varphi)$$

$$p(t|a, \theta) : \text{cond. density model}$$

$$p(a|\varphi) : \text{density model}$$

**Could try to fit $\varphi, \theta$ with maximum likelihood..?**

# Two Causal Directions, Two Models

$$A \rightarrow T$$

$$p(t, a|\varphi, \theta) = p(a|t, \theta)p(t|\varphi)$$

$$p(a|t, \theta) : \text{cond. density model}$$

$$p(t|\varphi) : \text{density model}$$

$$A \leftarrow T$$

$$p(t, a|\varphi, \theta) = p(t|a, \theta)p(a|\varphi)$$

$$p(t|a, \theta) : \text{cond. density model}$$

$$p(a|\varphi) : \text{density model}$$

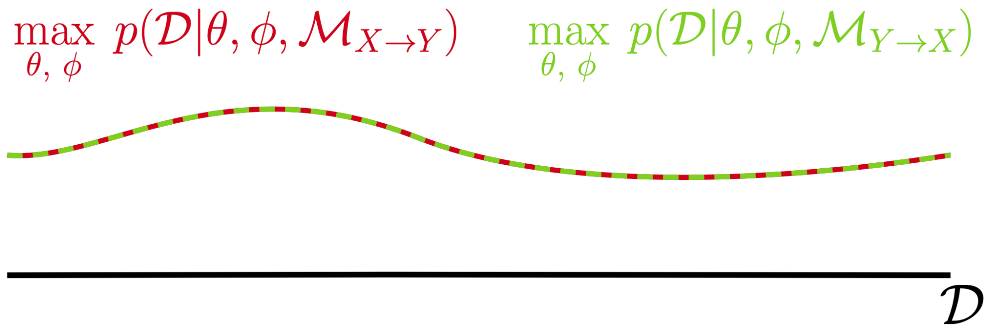
**Could try to fit $\varphi, \theta$ with maximum likelihood..?**

**For *flexible* models, both directions give equally good fit!**

Both models are in the same *Markov Equivalence Class.*

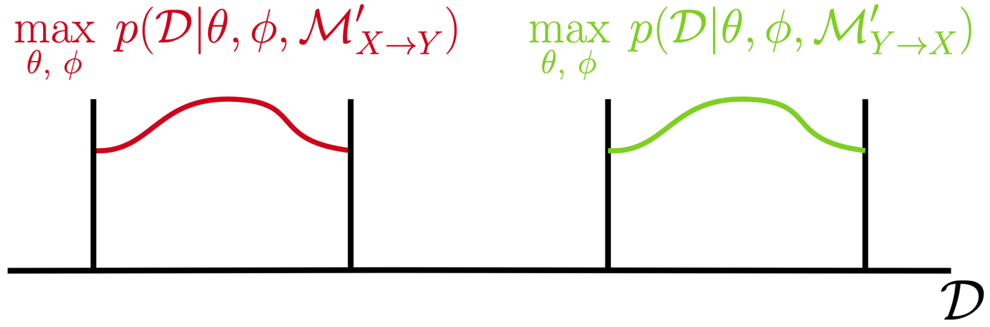
# Approach: Restricted Model Classes

For *flexible* models, both directions give equally good fit! 😢




**Add restrictions, e.g. ANM**

$$\text{effect} = f(\text{cause}) + \text{noise}$$



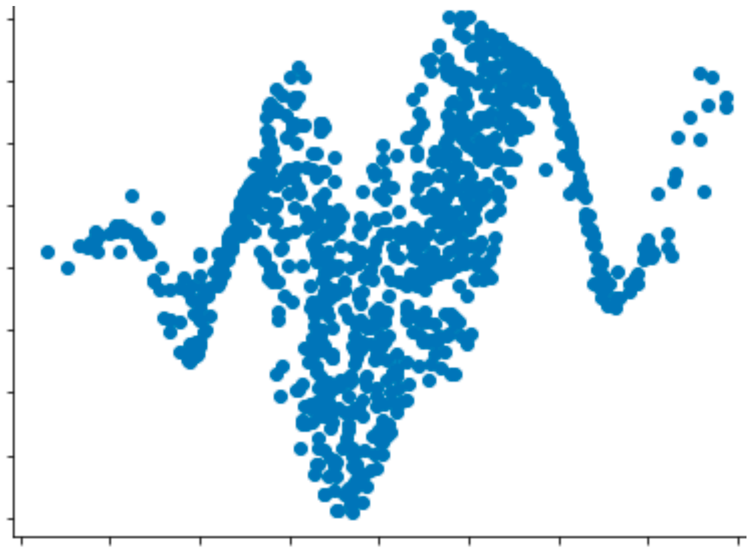

$\Rightarrow$ Non-overlapping data support

$\Rightarrow$ So... identifiable! (as $N \to \infty$)

# Problem: Restricted Model Classes

But what to do with a dataset like this one?

- Outside datasets covered by ANM!
- Poor fit $\Rightarrow$ bad predictions.
- Loss of identifiability guarantees.

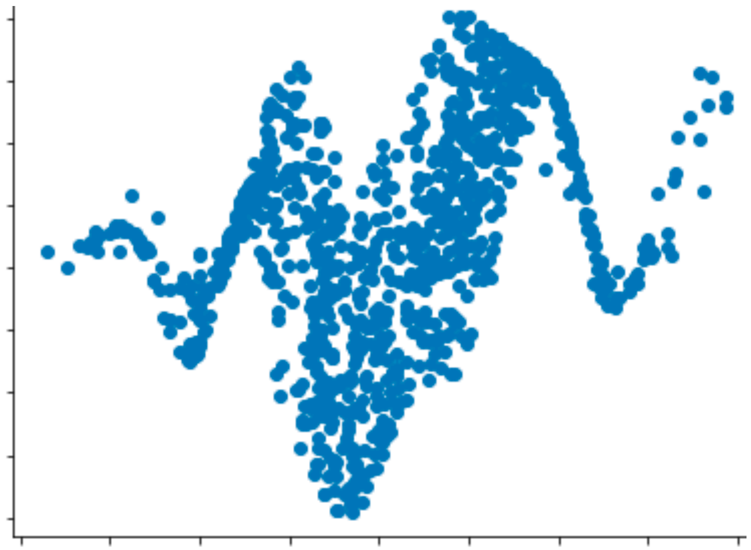


To model realistic datasets, we **want** our model to have support over all datasets!

# Problem: Restricted Model Classes

But what to do with a dataset like this one?

- Outside datasets covered by ANM!
- Poor fit $\Rightarrow$ bad predictions.
- Loss of identifiability guarantees.



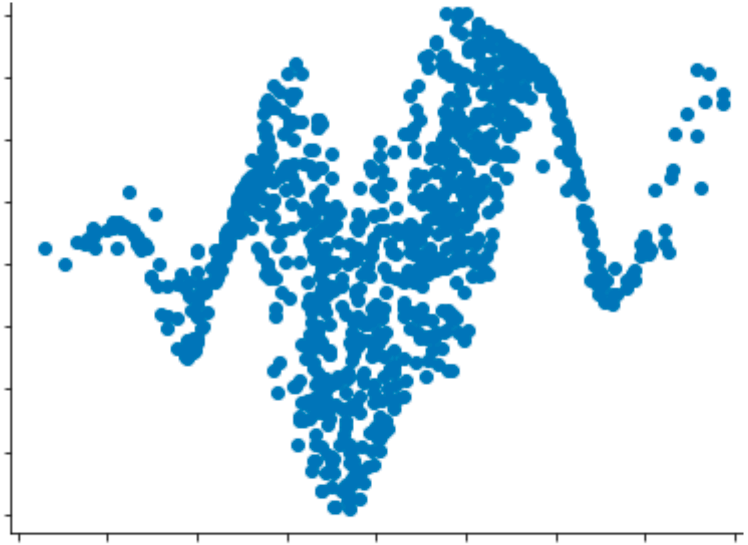
To model realistic datasets, we **want** our model to have support over all datasets!

# Problem: Restricted Model Classes

But what to do with a dataset like this one?

- Outside datasets covered by ANM!
- Poor fit $\Rightarrow$ bad predictions.
- Loss of identifiability guarantees.

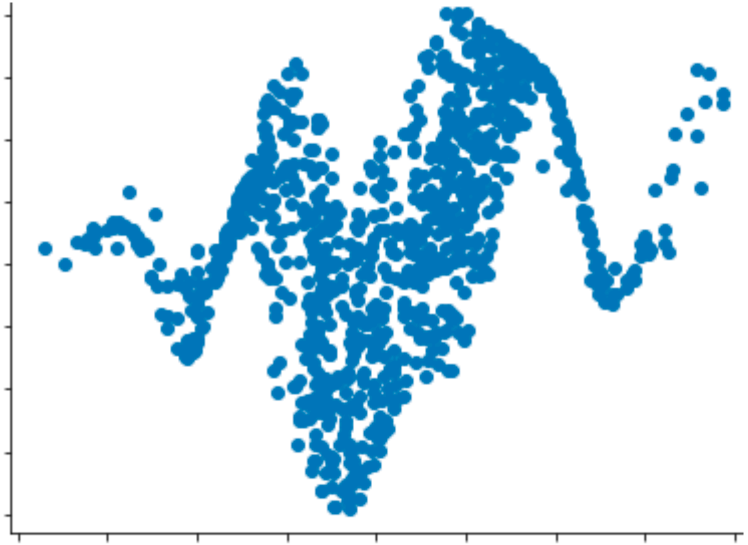
To model realistic datasets, we **want** our model to have support over all datasets!



# Problem: Restricted Model Classes

But what to do with a dataset like this one?

- Outside datasets covered by ANM!
- Poor fit $\Rightarrow$ bad predictions.
- Loss of identifiability guarantees.

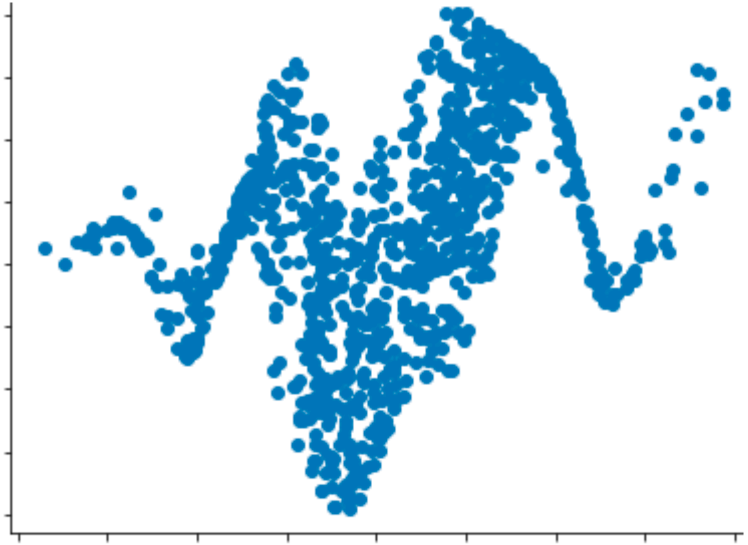


To model realistic datasets, we **want** our model to have support over all datasets!

# Problem: Restricted Model Classes

But what to do with a dataset like this one?

- Outside datasets covered by ANM!
- Poor fit $\Rightarrow$ bad predictions.
- Loss of identifiability guarantees.



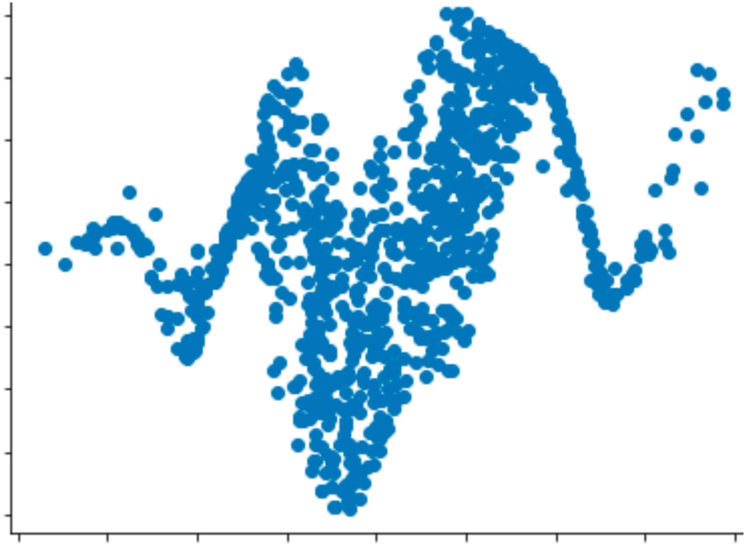
To model realistic datasets, we **want** our model to have support over all datasets!

# Problem: Restricted Model Classes

But what to do with a dataset like this one?

- Outside datasets covered by ANM!
- Poor fit $\Rightarrow$ bad predictions.
- Loss of identifiability guarantees.

To model realistic datasets, we **want** our model to have support over all datasets!



**Predict causal structure from observational data with flexible models with realistic assumptions.**

# Bayesian Perspective

# Model Selection

- We have two models, with different causal assumptions.
- Each model has its own unknown parameters.
- We want to determine which model is appropriate.

**Is this not just a hierarchical Bayesian inference problem?**

Just find the posterior over the models, using the marginal likelihood:

$$p(\mathcal{M}_{X \to Y} | \boldsymbol{x}, \boldsymbol{y}) \propto p(\boldsymbol{x}, \boldsymbol{y} | \mathcal{M}_{X \to Y}) p(\mathcal{M}_{X \to Y})$$

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

Has been investigated before, but didn't get it quite right (see paper).

# Causal Assumptions in Bayesian Models

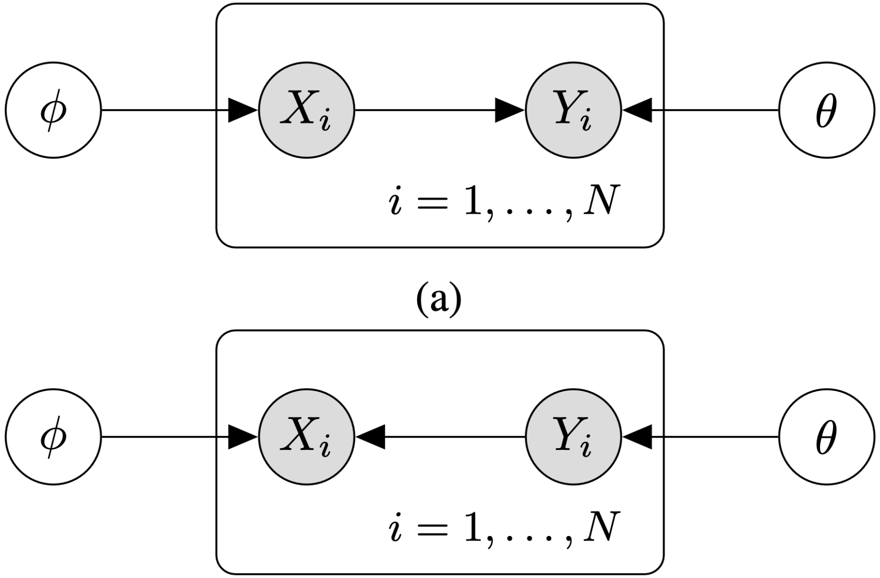
Observational data, so causality enters only through model assumptions.

Symmetry implies that:

- $p(\mathcal{M}_{X\to Y}) = p(\mathcal{M}_{Y\to X})$
- We want the same prior on $p(y_i|x_i, \theta, \mathcal{M}_{X\to Y})$ as on $p(x_i|y_i, \varphi, \mathcal{M}_{Y\to X})$.
- And similarly for $p(x_i|\varphi, \mathcal{M}_{X\to Y})$ and $p(y_i|\theta, \mathcal{M}_{Y\to X})$.

ICM implies independent priors.

Causal direction is encoded in graph.

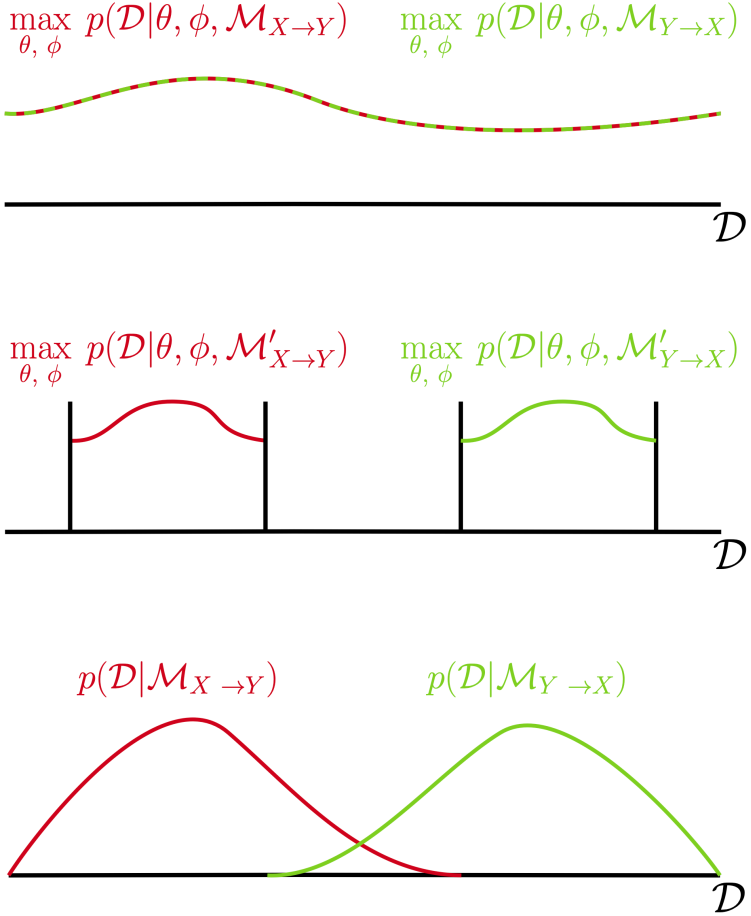


# Guarantees (or lack thereof)

- Priors **gives Bayes an opinion** on causal direction, where MaxLik does not.
- Even for flexible models with wide support!
- Price you pay: Overlap in distributions. So no perfect identifiability. Even if data sampled exactly from prior!

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

- Is this so different from existing approach?



# Putting this Into Practice

# A Practical Model

A conditional GPLVM (Bayesian VAE) for the conditional density:

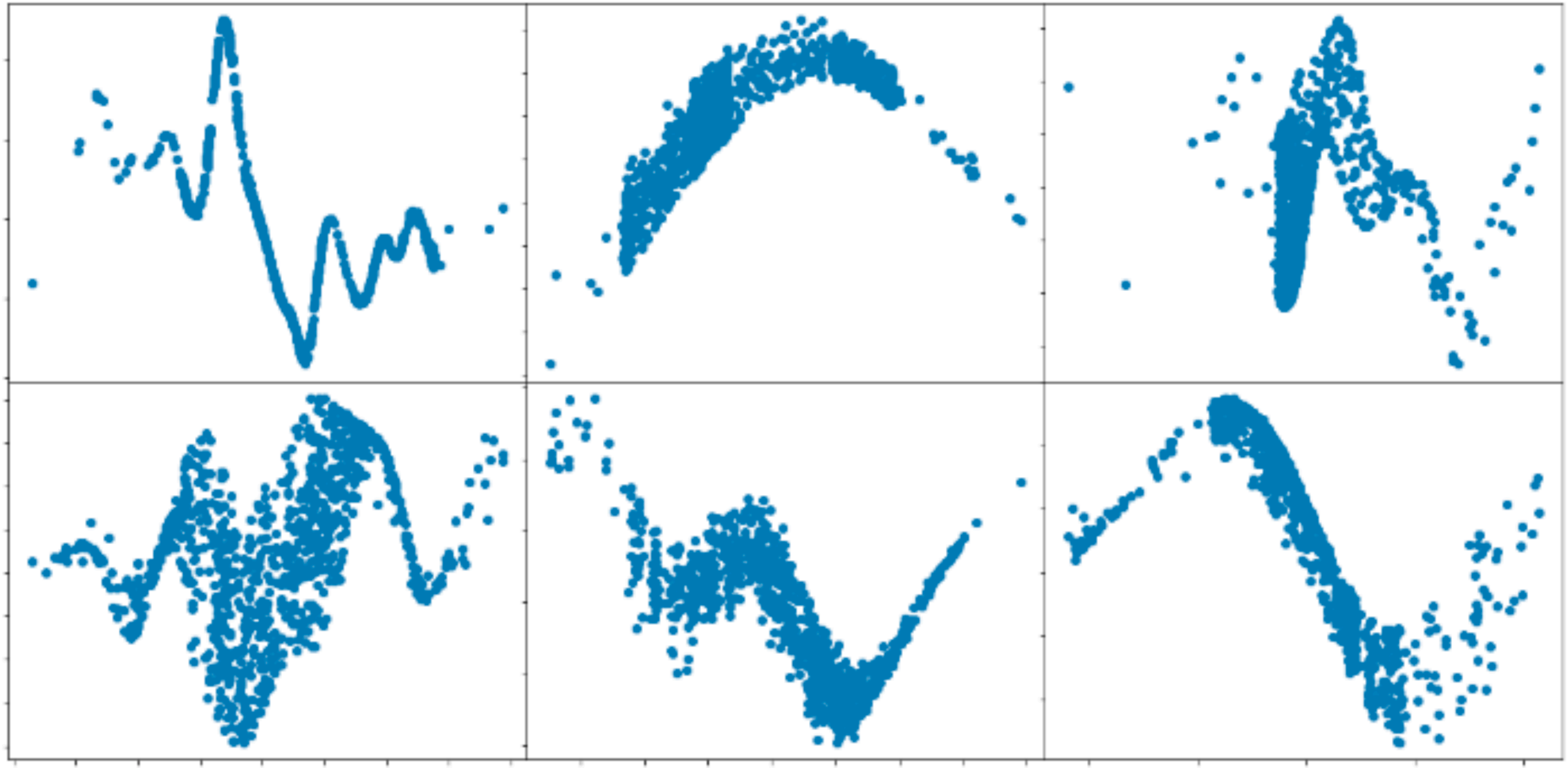
$$p(y_i|x_i, f, \mathcal{M}_{X\to Y}) = \int \mathcal{N}(y_i; f(x_i, w_i), \sigma^2)\mathcal{N}(w_i)\,\mathrm{d}w_i$$

$$f \sim \mathcal{GP}(0, k)$$

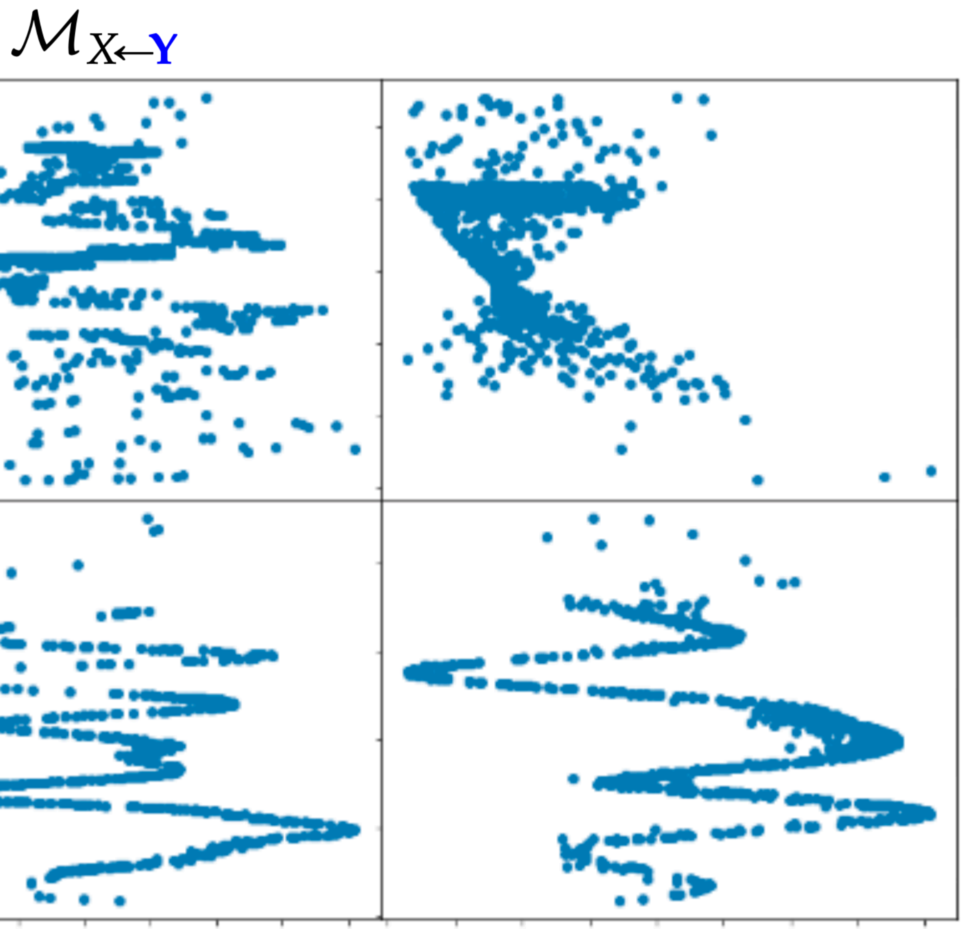
- Flexible (non-parametric) model over many conditional densities.
- Similar GPLVM prior on $p(x_i|g, \mathcal{M}_{X\to Y})$.
- Relatively standard variational approximation to perform inference.

# Overlap in Priors

$\mathcal{M}_{\mathbf{X} \to Y}$



# Overlap in Priors



# Experimental Results

Are our prior assumptions good?

- For identifiable ANM data, GPLVM gets 100% accuracy.
- For real data: Can **only** determine this experimentally, as in other approaches where theoretical assumptions are broken in practice.

| 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** |

# Summary

- Causal discovery from observational data is naturally a Bayesian Model Selection problem.
- Bayes allows specifying *realistic* assumptions, without artificial/unverifiable restrictions.

**A Bayesian method with realistic assumptions without strict guarantees**
**outperforms methods with unrealistic assumptions that do provide guarantees.**

# Future Work & Links to Deep Learning

- Can we express causal assumptions in neural network architecture, and discover them?

**Learning Layer-wise Equivariances Automatically using Gradients**

**Tycho F.A. van der Ouderaa**[1] **Alexander Immer**[2,3] **Mark van der Wilk**[1,4]

- Can we scale this to multiple variables?
- Can we use deep generative models as meta-learners to replace explicit Bayesian approximate inference? (Everything Bayes can do, meta-learning can do with simulated data.)