Causal Understanding through Bayes and Foundation Models

Mark van der Wilk
StatML Workshop, Amazon Berlin

Department of
COMPUTER
SCIENCE

https://mvdw.uk
@markvanderwilk

Causal Understanding through Bayes and Meta-Learning

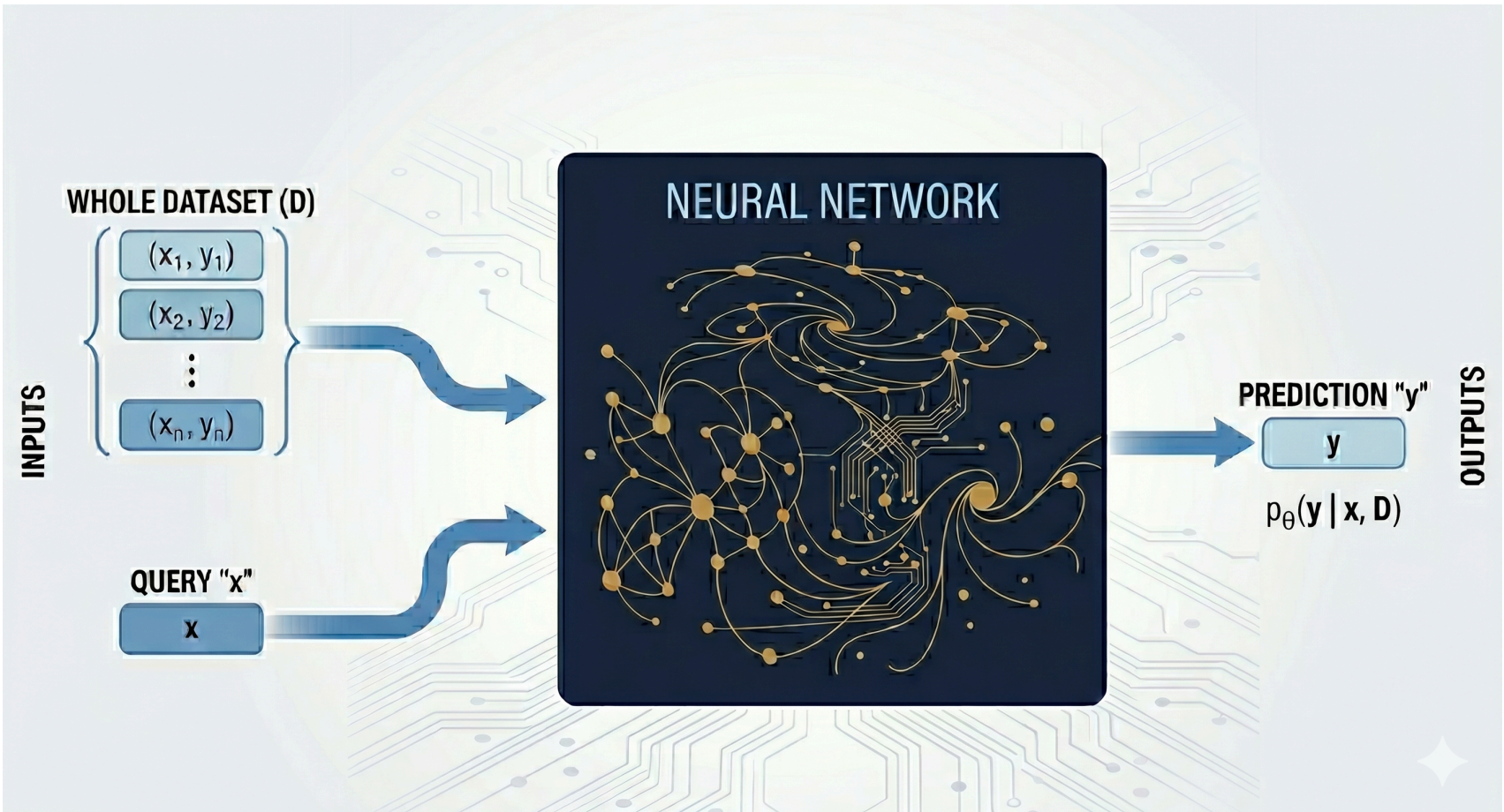
Mark van der Wilk
StatML Workshop, Amazon Berlin

Department of
COMPUTER
SCIENCE

https://mvdw.uk
@markvanderwilk

Detour:
In-Context Learning == Bayes

Pretrained In-Context Learners / Meta-Learning



Pretrained In-Context Learners / Meta-Learning

Many papers follow this structure!

- Neural Processes
  (Garnelo et al. (2018a); Garnelo et al. (2018b); Kim et al. (2019) ; etc…)
- Meta-Learning Probabilistic Inference for Prediction
  (Gordon et al., 2019)
- TabPFN models
  (Hollmann et al. (2023); Müller et al. (2022) ; etc…)
- Pretrained time-series forecasting models
  (Ansari et al., 2024; 2025)

Problem Set-Up

Data: Samples from some distribution over *datasets*

$$\pi\Big(x, y, \{x_n, y_n\}_{n=1}^N\Big) = \left[\int \pi(y|x,\eta) \prod_{n=1}^N \pi(y_n|x_n,\eta)\pi(\eta)\,\mathrm{d}\eta\right] \pi(x) \prod_{n=1}^N \pi(x_n)$$

Problem Set-Up

Data: Samples from some distribution over *datasets*

$$\pi(x, y, \mathcal{D}) = \left[\int \pi(y|x, \eta) \prod_{n=1}^{N} \pi(y_n|x_n, \eta) \pi(\eta) \, \mathrm{d}\eta \right] \pi(x) \prod_{n=1}^{N} \pi(x_n)$$

- So, your training dataset consists of many $\mathcal{D}_m$ ("dataset of datasets")
- ... with corresponding x_m, y_m, which are all sampled with the same η.

Problem Set-Up

Data: Samples from some distribution over *datasets*

$$\pi(x, y, \mathcal{D}) = \left[\int \pi(y|x, \eta) \prod_{n=1}^{N} \pi(y_n|x_n, \eta) \pi(\eta) \, \mathrm{d}\eta \right] \pi(x) \prod_{n=1}^{N} \pi(x_n)$$

- So, your training dataset consists of many $\mathcal{D}_m$ ("dataset of datasets")
- ... with corresponding x_m, y_m, which are all sampled with the same η.

Model: $p^\theta(y_m|x_m, \mathcal{D}_m)$

- ... so a generative model on y, with x and $\mathcal{D}$ as an input.

What Does the Loss Encourage?

Remember: $\operatorname{argmin}_\theta \mathcal{L}(\theta) = \operatorname{argmin}_\theta \mathcal{L}(\theta) + \text{const}$

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{E}_{\Pi_{XY\mathcal{D}}}\left[-\log p^\theta(Y|X,\mathcal{D})\right] \\ &= \mathbb{E}_{\Pi_{X\mathcal{D}}}\left[\mathbb{E}_{\Pi_{Y|\mathcal{D}X}}\left[-\log p^\theta(Y|X,\mathcal{D})\right]\right]\end{aligned}$$

$$\begin{aligned}\mathcal{L}'(\theta) &= \mathbb{E}_{\Pi_{X\mathcal{D}}}\left[\mathbb{E}_{\Pi_{Y|\mathcal{D}X}}\left[-\log p^\theta(Y|X,\mathcal{D})\right] + \mathcal{H}\left[\Pi_{Y|X,\mathcal{D}}\right]\right] \\ &= \mathbb{E}_{\Pi_{X\mathcal{D}}}\left[\mathbb{E}_{\Pi_{Y|\mathcal{D}X}}\left[\log \frac{\pi(Y|X,\mathcal{D})}{p^\theta(Y|X,\mathcal{D})}\right]\right] \\ &= \mathbb{E}_{\Pi_{X\mathcal{D}}}\left[\mathrm{KL}\left[\Pi_{Y|\mathcal{D}X} \parallel P^\theta_{Y|\mathcal{D}X}\right]\right]\end{aligned}$$

What Does the Loss Encourage?

$$\operatorname*{argmin}_{\theta} \mathcal{L}(\theta) = \operatorname*{argmin}_{\theta} \mathbb{E}_{\Pi_{X\mathcal{D}}} \left[\mathrm{KL} \left[\Pi_{Y|\mathcal{D}X} \parallel P^{\theta}_{Y|\mathcal{D}X} \right] \right]$$

This is minimised when

$$p^{\theta}(y \mid \mathcal{D}, x) = \pi(y \mid \mathcal{D}, x)$$

Training In-Context Learning is trained to mimic Bayes

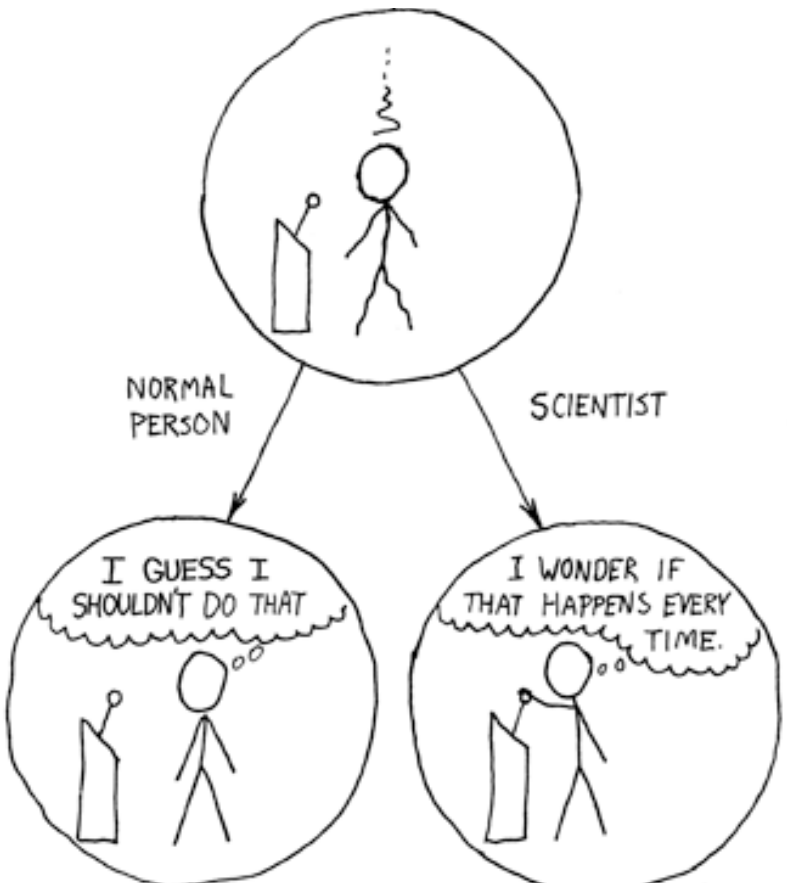
- Training "distribution on datasets" is the **prior**.
- Transformers often used to ensure permutation invariance.
- Only in the limit of large + flexible models (universal approximation).

Causality

Causation, Correlation, or Coincidence?

Causation, Correlation, or Coincidence?

Causation, Correlation, or Coincidence?



Today's Journey

Typical story:

- Correlation $\neq$ causation
- $\Rightarrow$ we cannot deduce causation from passive observations
- To be certain, we have to *intervene*. Do *experiments*.

? Can we learn *something* about causation from observations?

1. Bayes is powerful for encoding causal assumptions (new!)
2. Two ways of doing this: "Classical" Bayes, and Meta-Learning (new!)

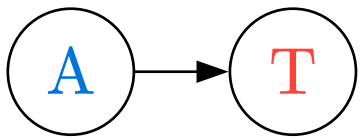
Based on a True Story (ICML 2024)

Bivariate Causal Discovery using Bayesian Model Selection

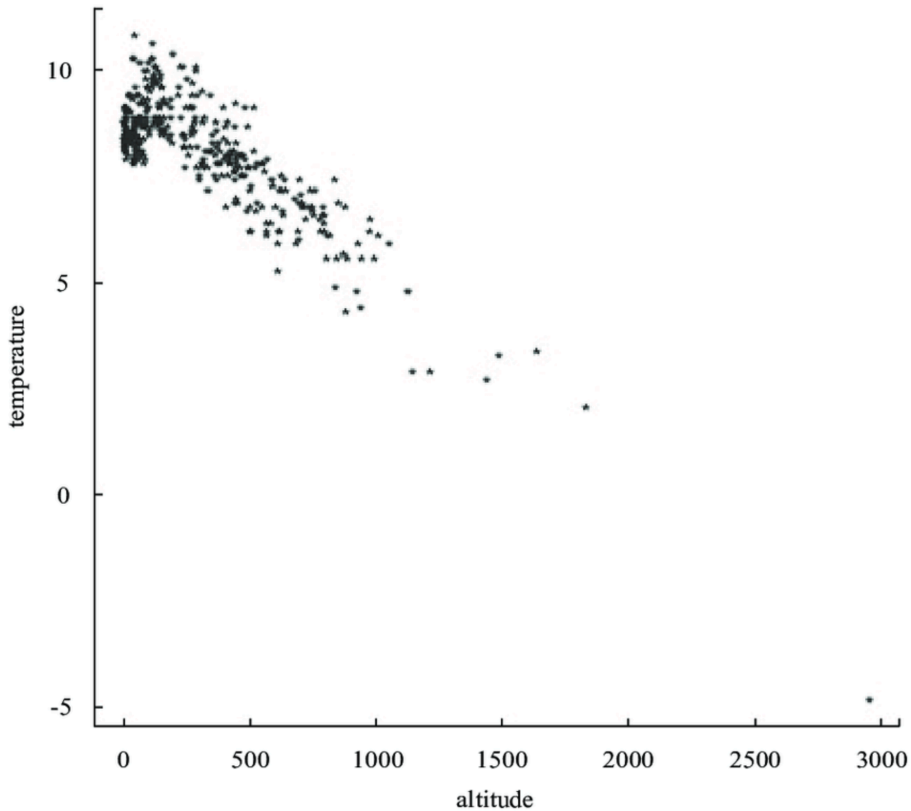
Anish Dhir [1] **Samuel Power** [2] **Mark van der Wilk** [3]

Why does Correlation $\neq$ Causation?

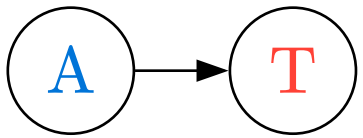
Why is Causality Important? (Pearl, 2009)



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



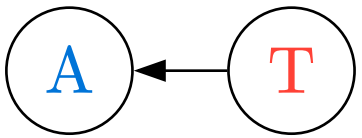
Two Causal Directions, Two Models



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

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

$p(a|\varphi)$: density model

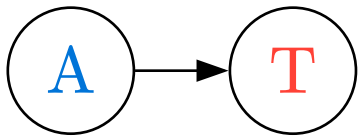


$$p(t, a|\varphi, \theta) = p(a|t, \theta)p(t|\varphi)$$

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

$p(t|\varphi)$: density model

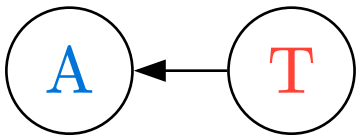
Two Causal Directions, Two Models



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

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

$p(a|\varphi)$: density model



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

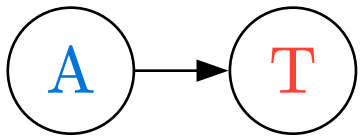
$p(a|t, \theta)$: cond. density model

$p(t|\varphi)$: density model

Model structure should match causal structure

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

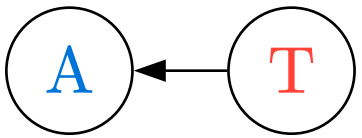
Two Causal Directions, Two Models



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

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

$p(a|\varphi)$: density model



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

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

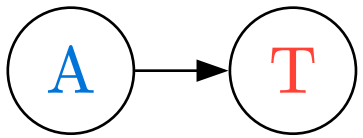
$p(t|\varphi)$: density model

Model structure should match causal structure

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

Goal: Predict causal structure from observational data.

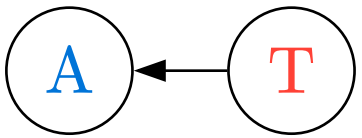
Two Causal Directions, Two Models



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

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

$p(a|\varphi)$: density model



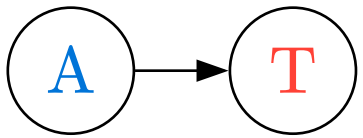
$$p(t, a|\varphi, \theta) = p(a|t, \theta)p(t|\varphi)$$

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

$p(t|\varphi)$: density model

Could try to fit φ, θ with maximum likelihood..? Board

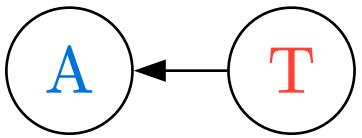
Two Causal Directions, Two Models



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

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

$p(a|\varphi)$: density model



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

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

$p(t|\varphi)$: density model

Could try to fit φ, θ with maximum likelihood..? Board

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

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

Approach: Restricted Model Classes

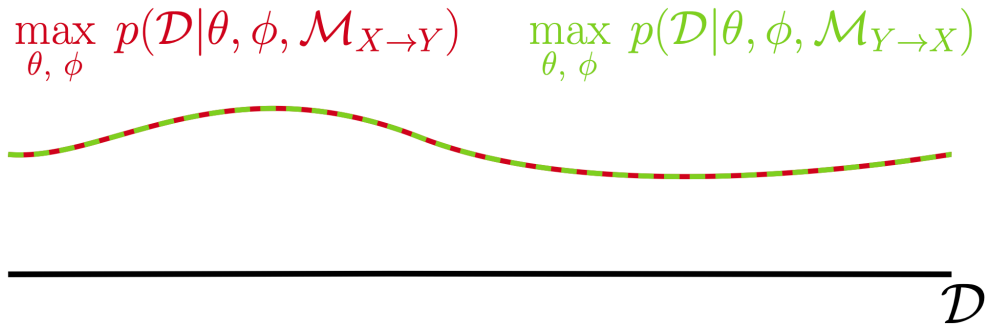
For *flexible* models, both directions give equally good fit! 😢

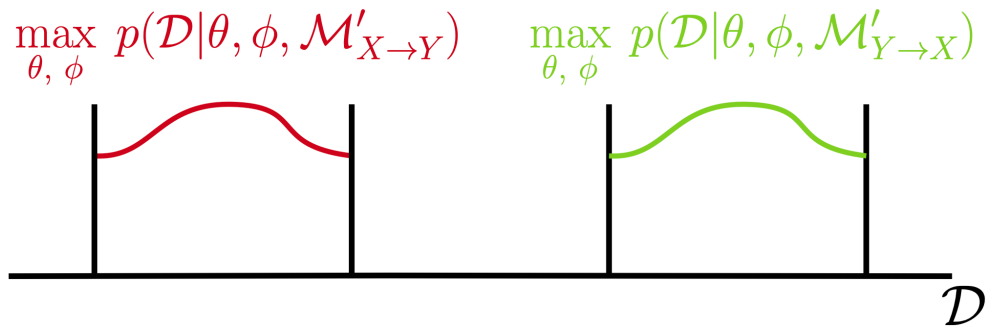
Add restrictions, e.g. ANM

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

$\Rightarrow$ Non-overlapping data support

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

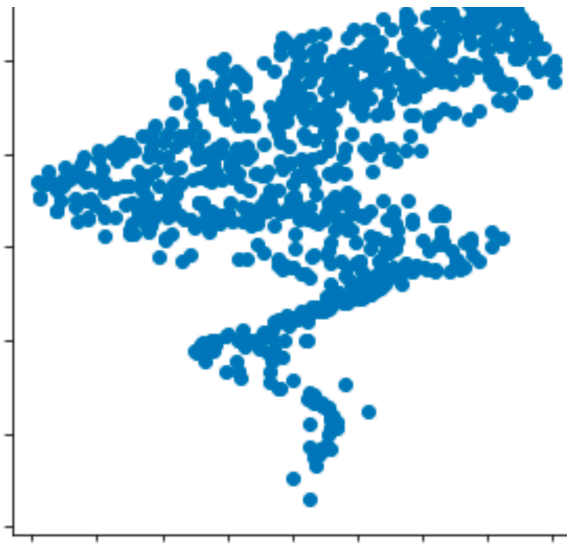




Problem: Restricted Model Classes

But what to do with a dataset like this one?

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



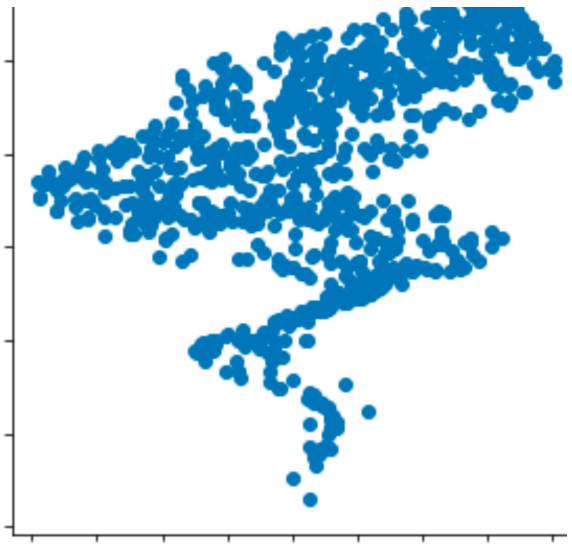
To model realistic datasets, we **want** our model to have support over all datasets!

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

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.



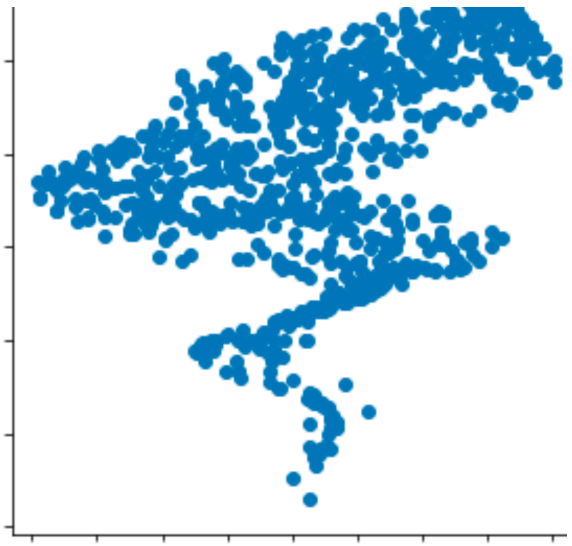
To model realistic datasets, we **want** our model to have support over all datasets!

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

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.



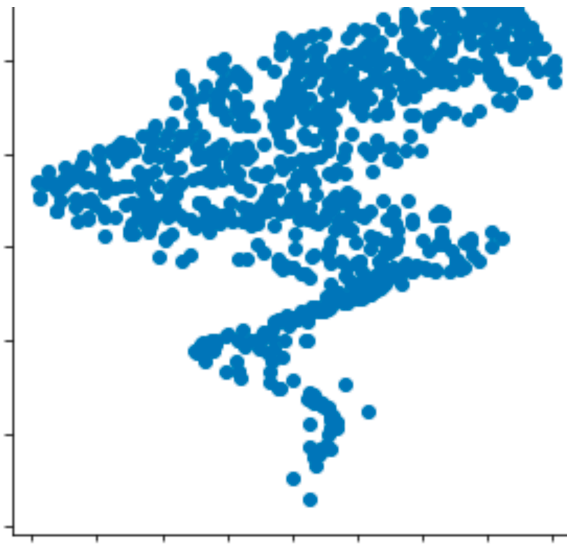
To model realistic datasets, we **want** our model to have support over all datasets!

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

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.



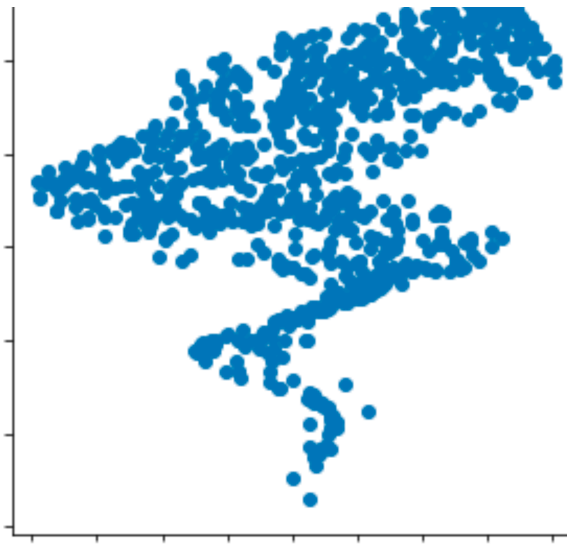
To model realistic datasets, we **want** our model to have support over all datasets!

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

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!

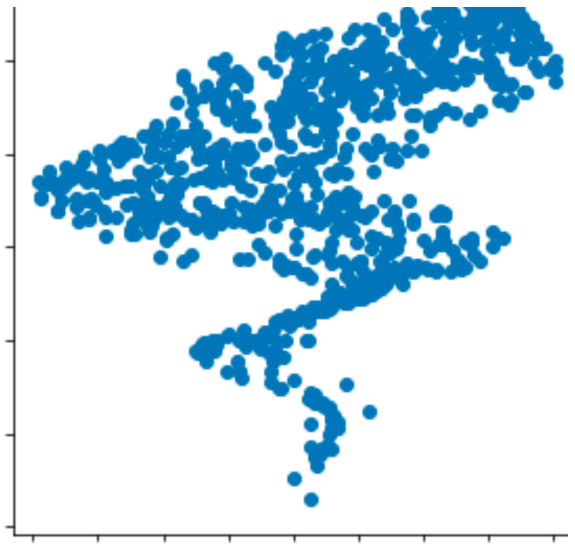
🎯 **Predict causal structure from observational data with flexible models with realistic assumptions.**

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!



❓ **But which axis do *you* think is the cause? X or Y?**

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

Restricted Model Classes: Summary

Assumptions about the *causal mechanism* can make *causal direction* identifiable.

⚠ Restrictive assumptions are a problem

- *If* your restrictive assumptions hold, you can *guarantee* recovering casual direction.
- If they don't: Your model won't fit, and causal discovery won't work.

? Can we still get information about causal direction, without imposing hard restrictions on datasets?

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

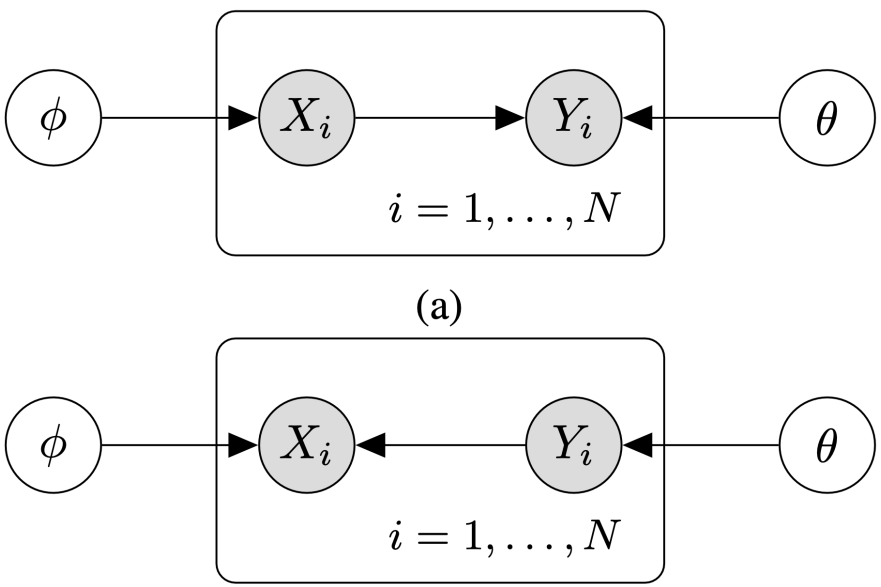
Observational data, so causality enters only through model assumptions.

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.



Assumptions (prior) is only placed on the *causal mechanism*

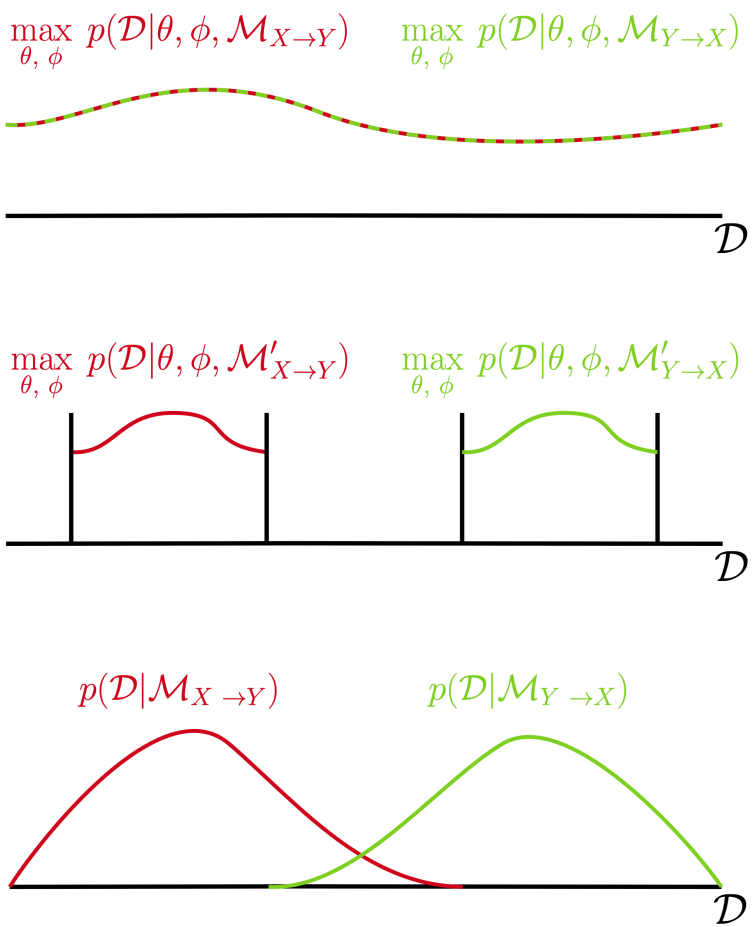
- Previous "Restricted Model Class" methods do the same!
- … just with hard restrictions.
- We use Bayesian priors to encode more realistic "softer" restrictions.
- Not so different!
- Hard restrictions are a special case: priors with limited support.

Guarantees (or lack thereof: the price to pay)

- 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

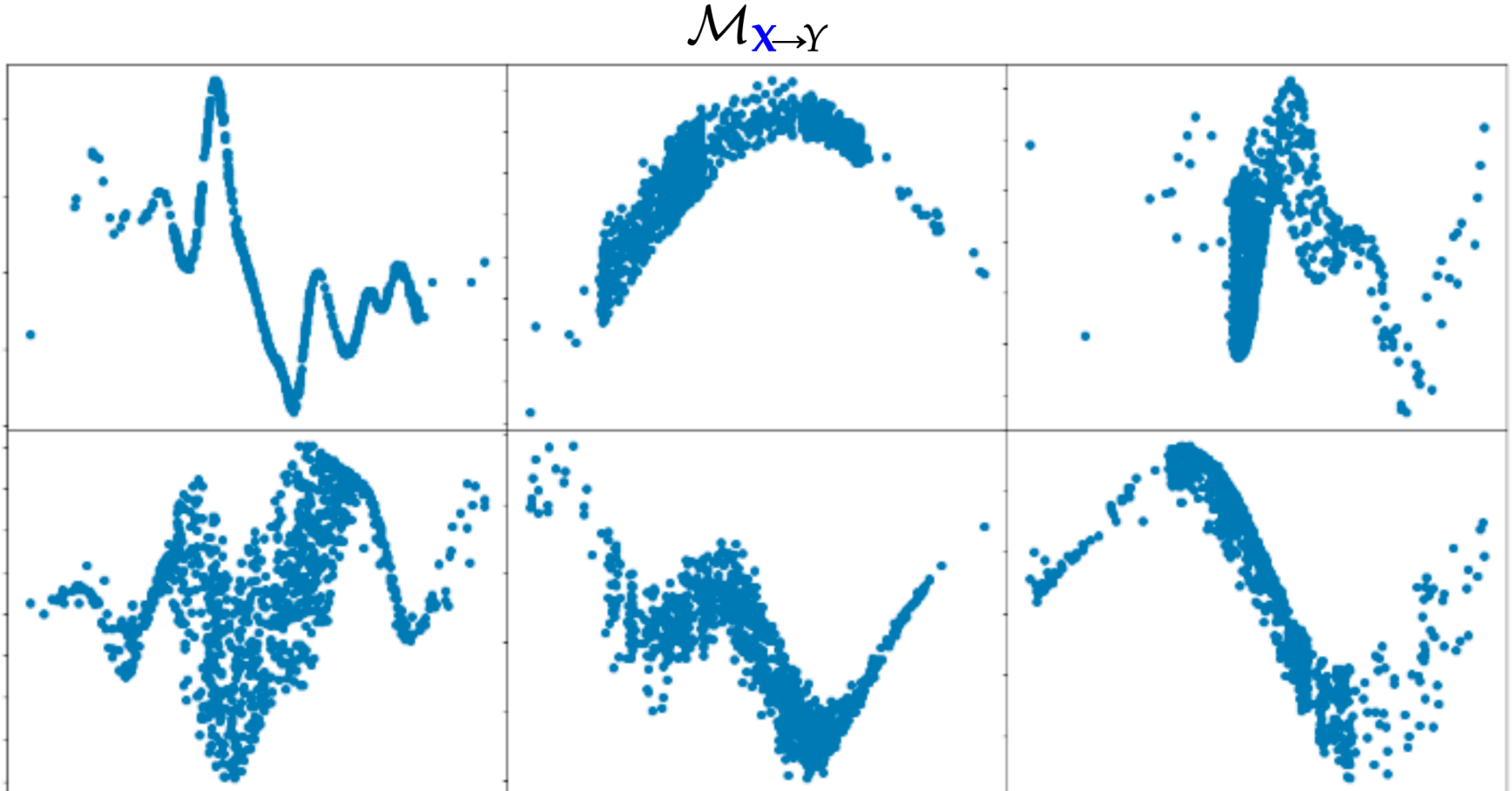
A conditional GPLVM (Bayesian VAE) for the conditional density:

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

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

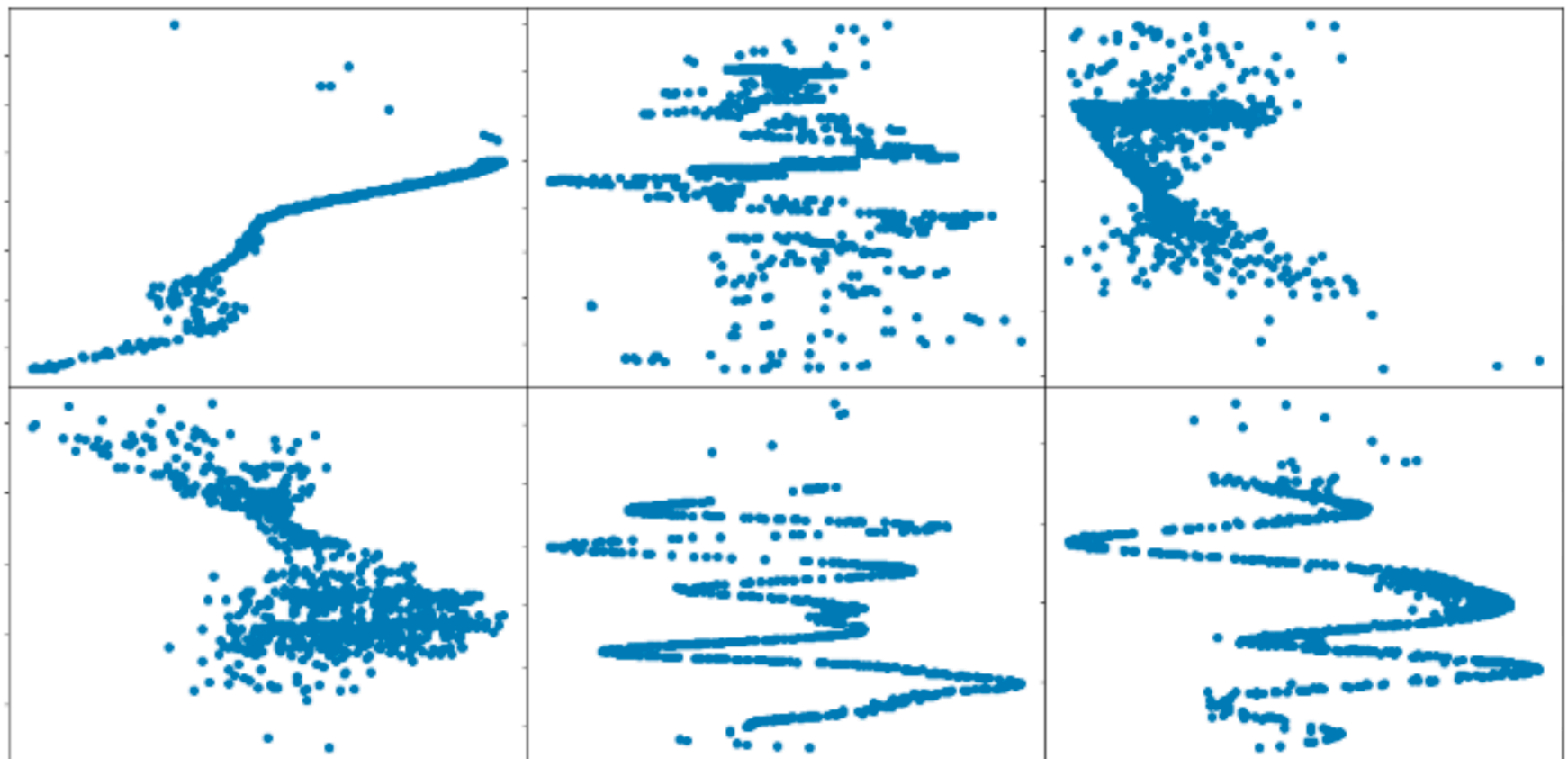
- Flexible (non-parametric) model over many conditional densities.
- Similar GPLVM prior on $p(x_i|g, \mathcal{M}_{X\to Y})$.
- Relatively standard variational approximation for inference. board

Overlap in Priors



Overlap in Priors

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



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**

Objections?

What happens when data is inherently ambiguous?

E.g. when the relationship is linear? (board)

Bayes will automatically quantify uncertainty!

$$p(\mathcal{M}_{X \to Y} | \boldsymbol{x}, \boldsymbol{y}) = 0.5$$

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.

How do we deal with more than two variables?

Bayes is difficult! How to handle inaccurate approximations?

Neural Foundation Models to Emulate Bayes

Learning Posteriors on Graphs

Published as a conference paper at ICLR 2025

A META-LEARNING APPROACH TO BAYESIAN CAUSAL DISCOVERY

Anish Dhir
Imperial College London
anish.dhir13@imperial.ac.uk

Matthew Ashman
University of Cambridge

James Requeima
University of Toronto

Mark van der Wilk
University of Oxford

Learning Posteriors on Graphs

Everything Bayes can do, a Neural Network can as well.

… but at the cost of needing data.

How would this work?

- Input would need to be an *entire dataset*.
- Output should be a distribution on causal graphs. (board)
- $\Rightarrow$ Need a "dataset of datasets", with each dataset "labelled" with its ground truth causal graph.

Training Objective

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log p(\mathcal{G}_n | \mathcal{D}_n, \theta) \qquad \mathcal{D}_n, \mathcal{G}_n \overset{\text{iid}}{\sim} \pi(\mathcal{D}, \mathcal{G})$$

In the limit as $N \to \infty$, with an expressive enough model, the minimiser of this loss satisfies:

$$p(\mathcal{G}_n | \mathcal{D}_n, \theta) = \pi(\mathcal{G} | \mathcal{D})$$

- Construct a synthetic data distribution π
- Neural network *replicates* Bayesian inference!

Learning Inteventional distributions

- Graphs are usually just an intermediate! More interested effects!
- Can we directly learn *interventional* distributions in the same way?

Estimating Interventional Distributions with Uncertain Causal Graphs through Meta-Learning

---

Anish Dhir*
Imperial College London

Cristiana Diaconu*
University of Cambridge

Valentinian Mihai Lungu
University of Cambridge

James Requeima
University of Toronto
Vector Institute

Richard E. Turner
University of Cambridge
Alan Turing Institute

Mark van der Wilk
University of Oxford

Conclusion

Inferring causal direction requires assumptions. How to do this?

- Bayesian inference is a good way to specify *realistic* assumptions!
- Strict identifiability is lost, but we can still have a high **probability** of success!

How to do the Bayesian inference?

- Can do classical Bayesian inference. Can be slow/difficult!
- Can also do meta-learning with large neural networks!
- Can predict various quantities: graphs, interventions!

Bibliography

Ansari, A. F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S. S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D. C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P. M., Wang, H., … Bohlke-Schneider, M. (2025,). *Chronos-2: From Univariate to Universal Forecasting*. **https://arxiv.org/abs/2510.15821**

Ansari, A. F., Stella, L., Turkmen, A. C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, B. (2024). Chronos: Learning the

Language of Time Series. *Transactions on Machine Learning Research*. **https://openreview.net/forum?id=gerNCVqqtR**

Dhir, A., Ashman, M., Requeima, J., & Wilk, M. van der. (2025,). A Meta-Learning Approach to Bayesian Causal Discovery. *The Thirteenth International Conference on Learning Representations (ICLR)*.

Dhir, A., Power, S., & Van Der Wilk, M. (2024). Bivariate Causal Discovery using Bayesian Model Selection. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning: Vol. 235. Proceedings of the 41st International Conference on Machine Learning*. **https://proceedings.mlr.press/v235/dhir24a.html**

Garnelo, M., Rosenbaum, D., Maddison, C. J., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D. J., & Eslami, S. M. A. (2018b). Conditional Neural Processes. *Proceedings of the 35th International Conference on Machine Learning*, 1704–1713.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S. M. A., & Teh, Y. W. (2018a). Neural Processes. *Arxiv Preprint Arxiv:1807.01622*.

Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., & Turner, R. (2019,). Meta-Learning Probabilistic Inference for Prediction. *International Conference on Learning Representations*. **https://openreview.net/forum?id=HkxStoC5F7**

Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023,). TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. *The Eleventh International Conference on Learning Representations*. **https://openreview.net/forum?id=cp5PvcI6w8_**

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, S. M. A., Rosenbaum, D., Vinyals, O., & Teh, Y. W. (2019,). Attentive Neural Processes. *International Conference on Learning Representations*.

Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., & Hutter, F. (2022,). Transformers Can Do Bayesian Inference. *International Conference on Learning Representations*. **https://openreview.net/forum?id=KSugKcbNf9**

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.