R.S. was supported by the Wellcome Trust [222836/Z/21/Z]. B.G. acknowledges the support of the UKRI AI programme, and the Engineering and Physical Sciences Research Council, for CHAI - EPSRC Causality in Healthcare AI Hub (grant no. EP/Y028856/1). A.K. is supported by UKRI (grant number EP/S023356/1), as part of the UKRI Centre for Doctoral Training in Safe and Trusted AI. A.D. acknowledges the support of G-Research's March 2025 grant.

Imperial College London | UNIVERSITY OF OXFORD | ICML International Conference On Machine Learning

Bayesian model selection allows for less restrictive functional assumptions when doing multivariate causal discovery on observational data

Bayesian Model Selection for Multivariate Causal Discovery

Anish Dhir*[1], Ruby Sedgwick*[1,2], Avinash Kori[1], Ben Glocker[1], Mark van der Wilk[3]

1 Department of Computing, Imperial College London, UK *equal contribution
2 Xyme, Oxford, UK
3 University of Oxford, UK

paper

Introduction

- Current causal discovery approaches require **restrictive model assumptions** to ensure identifiability, e.g. additive noise (ANM). In real-world data these restrictive assumptions are commonly violated, **losing identifiability guarantees**.
- Bayesian model selection has been proven to improve performance in the bivariate case by allowing **more flexible functional assumptions**.
- However, the naive discrete Bayesian model selection approach isn't feasible for the multivariate case where the number of possible graphs scales super-exponentially.
- We propose a **continuous Bayesian model selection** approach that **scales well to large numbers of variables** while still allowing more flexible functional assumptions.

Bayesian model selection solves this problem

$$P(\mathcal{M}_\mathcal{G}|X) \propto P(X|\mathcal{M}_\mathcal{G})P(\mathcal{M}_\mathcal{G}),$$

(posterior; marginal likelihood; prior)

$$P(X|\mathcal{M}_\mathcal{G}) = \int \prod_{i=1}^{D} \left(P_i(X_i|X_{\mathrm{PA}_\mathcal{G}(i)})\pi_i(\mathrm{d}P_i)\right)$$

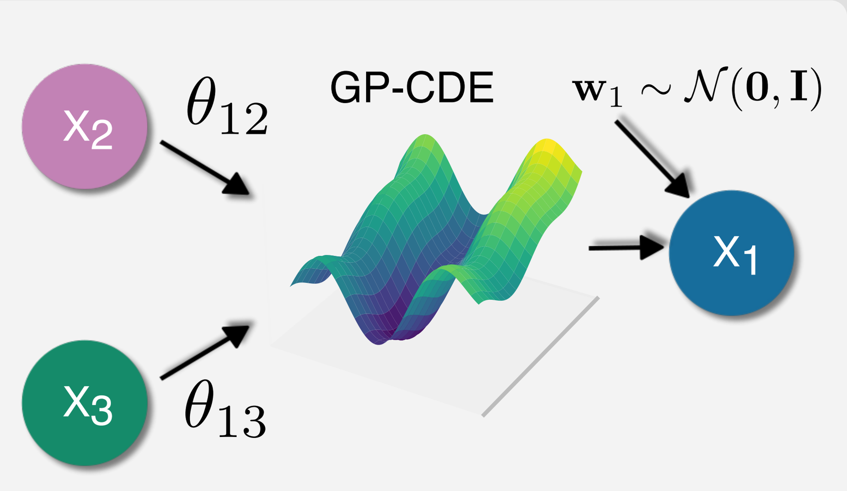
(marginal likelihood; parent set; prior on distribution)

Causal model: $\mathcal{M}_\mathcal{G}$ Data: $X \in \mathbb{R}^D$

- Factorised prior on distributions encodes the **independent mechanism (ICM) assumption**.
- The **marginal likelihood** will prefer models whose ICM assumption aligns with the properties of the data generating process.

Causal Gaussian Process Conditional Estimator (CGP-CDE)

Aim: given a set of variables, find the causal graph.



Fit a Gaussian process conditional estimator (GP-CDE) for each variable. Latent variable $\mathbf{w}_1$ allows non-Gaussian and heteroskedastic densities.

Loss:

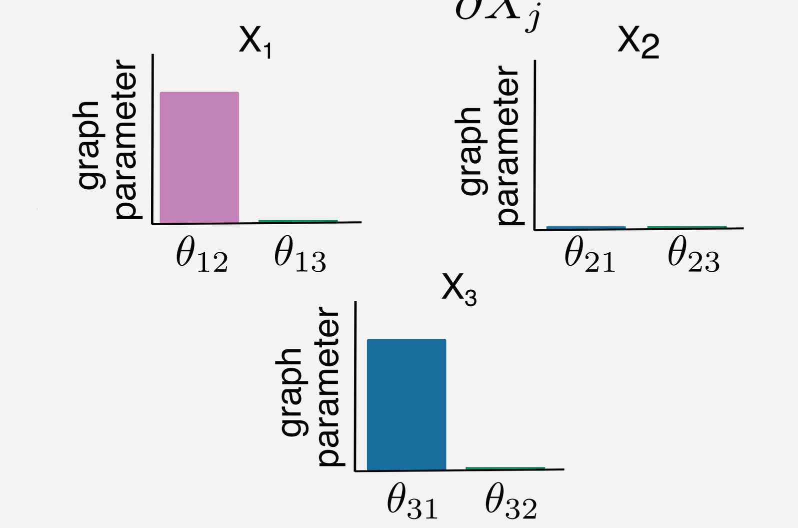
$$\mathcal{L}_{\mathrm{ELBO}}(q, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\phi}) + \log p(\boldsymbol{\theta}) - \gamma_t h(\mathbf{A}_{\boldsymbol{\theta}})$$

(marginal likelihood; acyclicity constraint)

$$A_{ij} = \begin{cases} \theta_{ij} & \text{if } j \neq i, \\ 0 & \text{otherwise.} \end{cases}$$

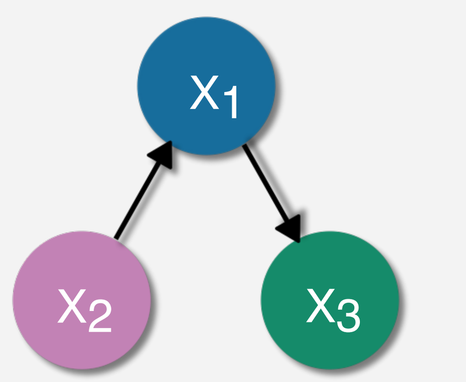
An adjacency matrix is constructed from the graph parameters (inverse lengthscales).

$$\theta_{ij} = 0 \implies \frac{\partial f_i}{\partial X_j} = 0.$$



The ARD properties of the kernel mean edges that aren't evidenced by the data are "switched off" by making the graph parameter small.

$$\mathbf{A}_{\boldsymbol{\theta}} = \begin{bmatrix} 0 & 0.6 & 10^{-12} \\ 10^{-21} & 0 & 10^{-14} \\ 1.2 & 10^{-13} & 0 \end{bmatrix}$$



The causal graph is extracted from the graph parameters via thresholding.

Probability of error

$$P(E|\mathcal{M}_\mathcal{G}) = \int_\mathcal{R} p(X|\mathcal{M}_\mathcal{G})dX,$$

(probability of error)

$$\mathcal{R} = \{X : \mathcal{S}(X|\mathcal{M}_\mathcal{H}) > \mathcal{S}(X|\mathcal{M}_\mathcal{G}) \text{ for } \mathcal{H} \neq \mathcal{G}\}$$

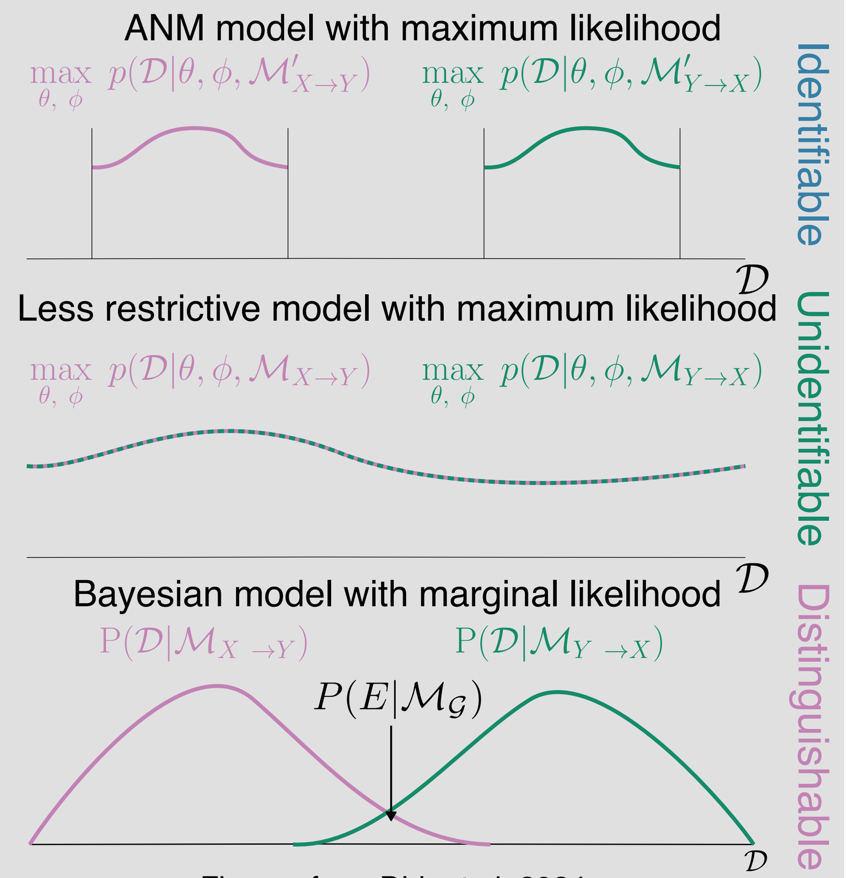
(score)

Identifiable: $P(E|\mathcal{M}_\mathcal{G}) = 0$

Unidentifiable: $P(E|\mathcal{M}_\mathcal{G}) = P(E|U(\mathcal{G}))$

Distinguishable: $0 < P(E|\mathcal{M}_\mathcal{G}) < P(E|U(\mathcal{G}))$

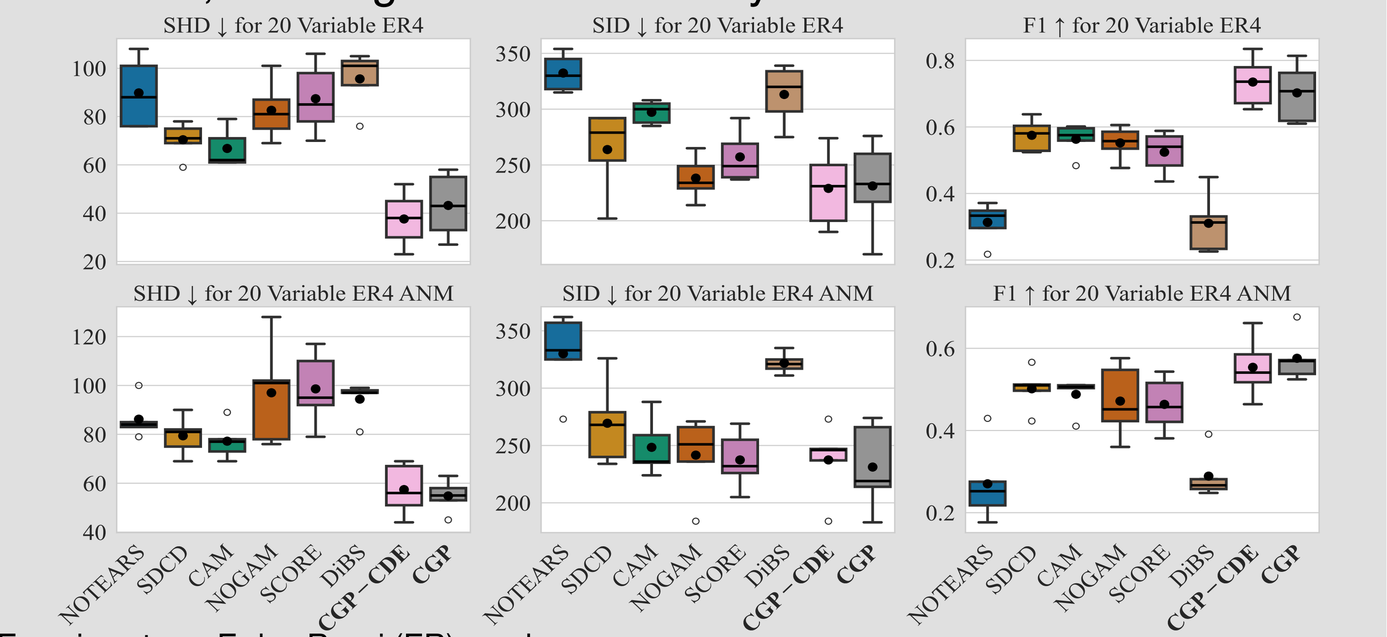
(uniformly random selected graph)



Figures from Dhir et al. 2024

Is performance on identifiable data worse?

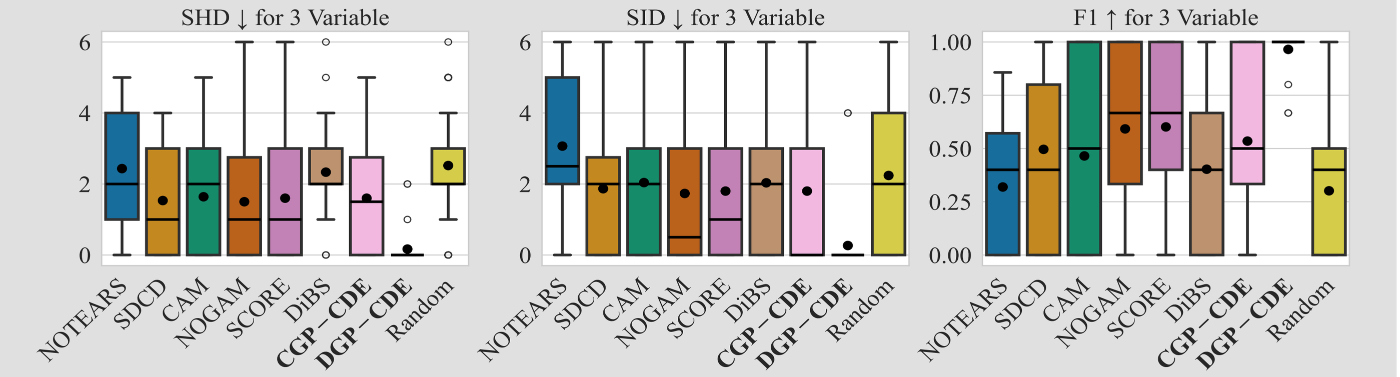
- CGP-CDE outperforms ANM methods on ANM and non-ANM data.
- Causal Gaussian process (CGP), which has ANM assumption, does well, showing the benefit of Bayesian model selection.



Experiments on Erdos-Renyi (ER) graphs.

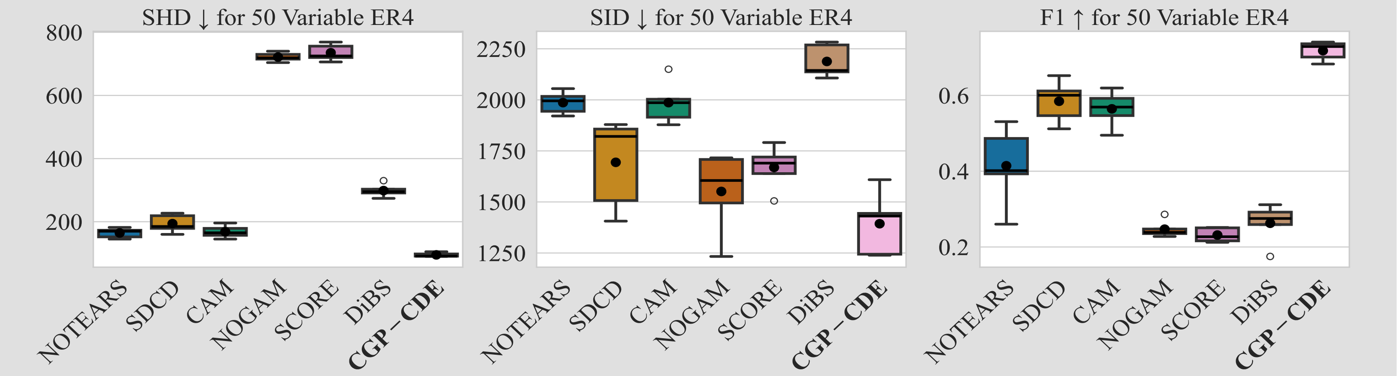
How bad is the continuous approximation?

- Excellent performance of discrete Bayesian model selection (DGP-CDE) shows the **probability of error is small**.
- The difference in performance between CGP-CDE and DGP-CDE is due to the continuous relaxation in CGP-CDE.



Does it scale to large numbers of variables?

- Experiments on ER graphs with 4 expected edges per variable and 50 variables. The CGP-CDE outperforms the other models.



How well does it perform on semi-synthetic data?

- Experiments on Syntren dataset, derived from a gene regulatory network simulator which has 10 datasets of 20 nodes. The CGP-CDE again outperforms the other models.