Optimal experimental design via Bayesian optimization: active causal structure learning for Gaussian process networks

Max Planck Institute for Intelligent Systems

Julius von Kügelgen[1,2], Paul K. Rubenstein[1,2], Bernhard Schölkopf[1], Adrian Weller[2,3]

[1]Max Planck Institute for Intelligent Systems, [2]University of Cambridge, [3]The Alan Turing Institute

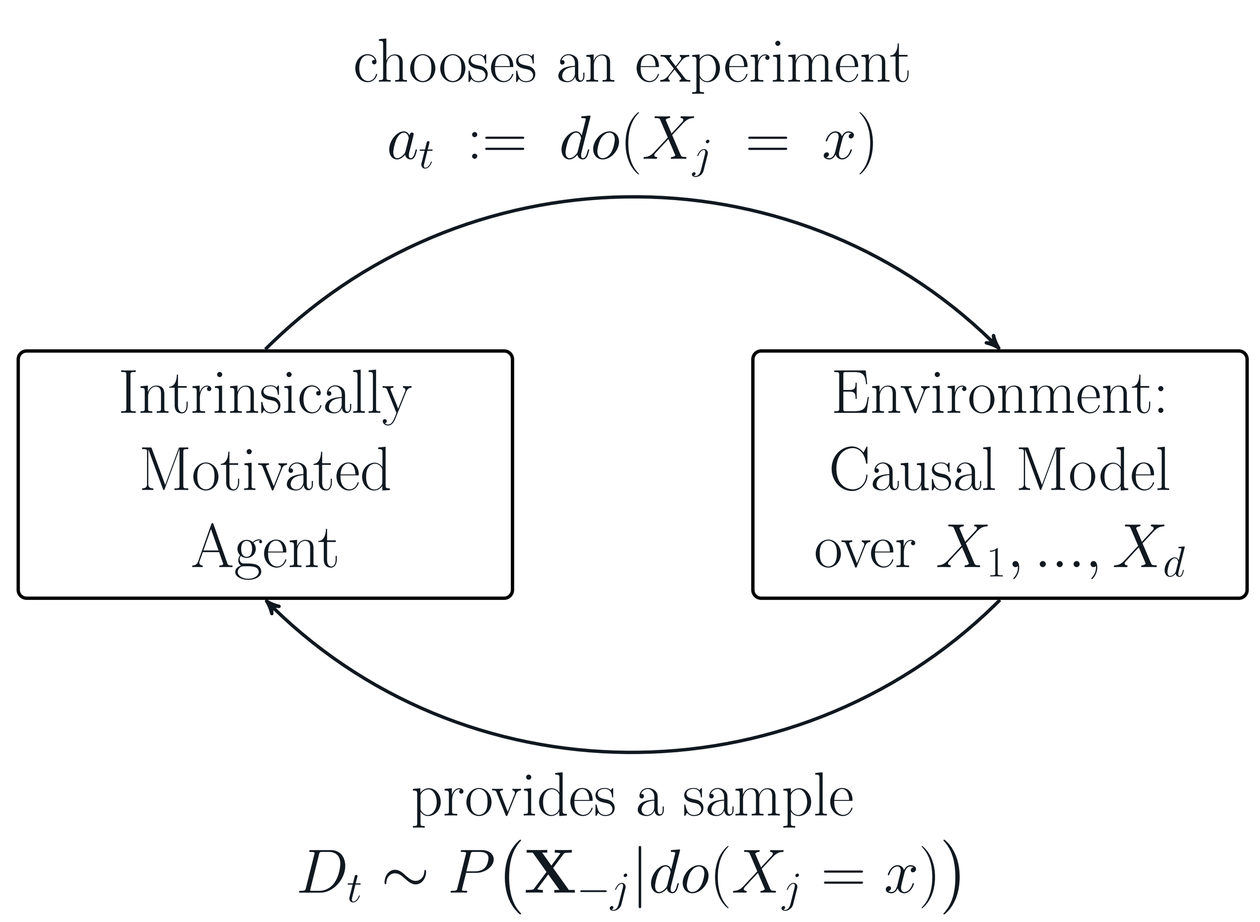
UNIVERSITY OF CAMBRIDGE

Summary

- optimal experimental design for continuous random variables with unknown causal structure
- allow for non-linear functional relationships modelled with Gaussian process priors
- Bayesian active learning approach to perform maximally informative experiments for causal structure
- Bayesian optimisation to efficiently maximise a Monte Carlo estimate of expected information gain

Motivation

Humans learn causal models not just from large amounts of observational or interventional data. Rather, we constantly interact with our environment: conducting experiments to test hypotheses and updating our beliefs based on the outcomes → *"child as a scientist"*.



Problem setting

Assume a structural causal model (SCM) over $\mathbf{X} = \{X_1, ..., X_d\}$,

$$X_i := f_i(\mathbf{Pa}_i^{G^*}) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \qquad (i = 1, ..., d),$$

where **both** G^* and the (possibly) **nonlinear** f_i are **unknown**.

Q: Which intervention $do(X_j = x)$ should we try next to learn G^*?

Active Bayesian causal discovery

Given: prior beliefs about DAGs and their associated parameters, $P(G)P(\theta_G|G)$; likelihood function, $P(\mathbf{D}|\theta_G, G)$.

Goal: maximise the information gain in G w.r.t. the expected outcome of the experiment, subject to current beliefs.

$$\sum_{G \in \mathcal{G}} P(G) \int P\big(\mathbf{x}_{-j}|G, do(X_j = x)\big) \log P\big(G|\mathbf{x}_{-j}, do(X_j = x)\big) d\mathbf{x}_{-j}$$

Problem: intractable due to integration over θ_G and $\mathbf{x}_{-j}$.

Approach: 1. Gaussian process priors over $f_i \sim \mathcal{GP}(0, k_i)$

$\Longrightarrow$ predictive posterior $P\big(\mathbf{x}_{-j}|G, do(X_j = x)\big)$ and marginal likelihood $P(\mathbf{D}|G)$ available in closed form

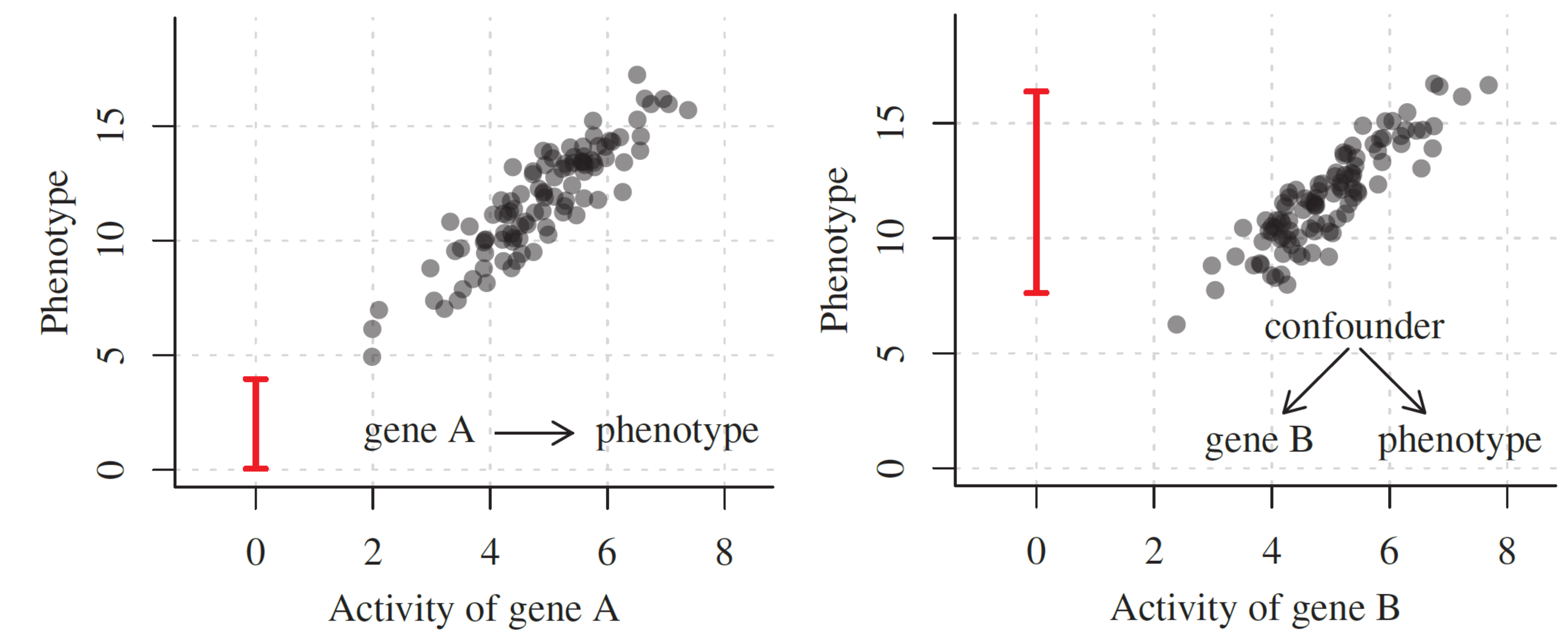
2. Monte Carlo estimate using M sampled outcomes $\mathbf{x}_{-j}^{(m)}$

$$\underset{j \in \{1,...,d\}, x \in \mathcal{X}_j}{\arg\max} \sum_{G \in \mathcal{G}} P(G) \frac{1}{M} \sum_{m=1}^{M} \log P\big(G|\mathbf{x}_{-j}^{(m)}, do(X_j = x)\big) \qquad (1)$$

3. Approximate arg max over continuous $x \in \mathcal{X}_j$ in (1) using Bayesian optimisation

Contrast to linear setting

Nonlinear functional relationships introduce additional uncertainty and complicate causal discovery–even from experimental data!



[Figure from Peters et al. (2017)]

Bivariate example

Intervention	$G_1 : X \to Y$	$G_2 : Y \to X$
$p\big(y|do(x)\big)$	$p(y|x)$	$p(y)$
$p\big(x|do(y)\big)$	$p(x)$	$p(x|y)$