Some Research Projects and Ideas

Qing Zhou

zhou@stat.ucla.edu

UCLA Department of Statistics & Data Science

March, 2026

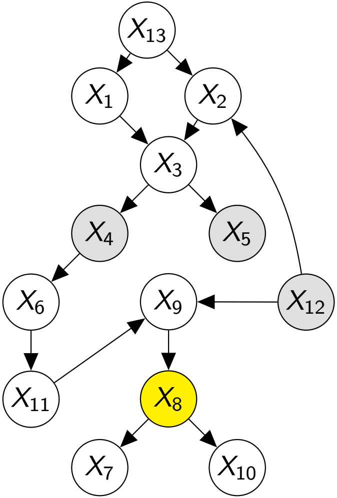
Outline

1. Sequential intervention
2. Causal discovery on dependent data
3. Causal discovery under latent confounding
4. Causal representation learning
5. Applications in genomics data

Acknowledgement: Some of these research projects are supported by NIH and NSF.

Sequential intervention

Causal bandit problem



- reward variable $Y = X_8$
- may intervene on $\{X_4, X_5, X_{12}\}$
  $\mathrm{do}(X_4 = 1)$
  $\mathrm{do}(X_5 = -1)$, etc.
- rewards:
  $\mathrm{E}[X_8 \mid \mathrm{do}(X_4 = 1)]$
  $\mathrm{E}[X_8 \mid \mathrm{do}(X_5 = -1)]$, etc.
- Design a sequential decision process to to find the optimal intervention that maximizes the reward (causal effect).

Sequential intervention

Design sequential intervention for structure learning of causal DAGs (active learning)

Iterative between the following steps:

- Given the current estimated graph structure (partial DAG), choose node(s) to perform intervention
- Generate experimental data under the chosen intervention
- Combine all data collected so far to update the graph structure

Papers: Huang and Zhou (2023) TMLR; Zhao and Zhou (2025) arXiv.

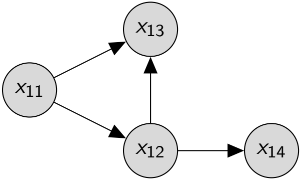
Causal discovery on dependent data

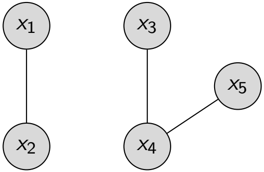
Model sample dependence under causal DAGs.

- Data matrix $X = (x_{ij})_{n \times p}$, $n = 5$ units and $p = 4$ variables.

$$\begin{array}{ll} \text{unit 1} & x_1 = (x_{11}, x_{12}, x_{13}, x_{14}) \\ & \dots \\ \text{unit 5} & x_5 = (x_{51}, x_{52}, x_{53}, x_{54}) \end{array}$$

- DAG (left) modeling causality among $(x_{i1}, x_{i2}, x_{i3}, x_{i4})$, $i \in [n]$; undirected graph (right) over all units $x_1, x_2, \dots, x_5$ for sample dependence.





Causal discovery on dependent data

Main ideas:

- SEM for x_i: $x_{ij} = f_j(pa_{ij}, \varepsilon_{ij})$, $j = 1, \ldots, p$.
  If $\varepsilon_j = (\varepsilon_{1j}, \ldots, \varepsilon_{nj})$ are i.i.d., then $x_1, \ldots, x_n$ are i.i.d.
- To model sample dependence, assume

$$\varepsilon_j \sim \mathcal{N}_n(0, \omega_j \Sigma) \quad \forall j \in [p].$$

  Σ ($n \times n$) models the correlation among $x_1, \ldots, x_n$.
- De-correlation: (1) Estimate $\widehat{\Sigma}$; (2) Use $\widehat{\Sigma}$ to transform data X so that the sample dependence is removed $\Rightarrow \widehat{X}$.
- Apply standard algorithms to $\widehat{X}$ for structure learning.

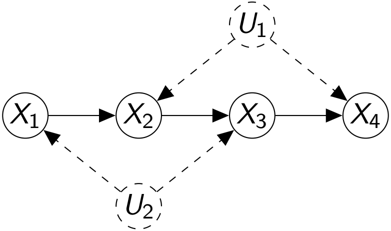
Papers: Li et al. (2024) JMLR; Chen and Zhou (2025) AISTATS.

Future work

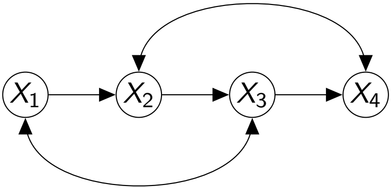
- Generalization to nonlinear SEM.
- Relax the Gaussian error assumption.
- Theoretical analysis of $\widehat{\Sigma}$ or $\widehat{\Theta}$ in more general settings (e.g. any data matrix with row dependence).

Causal discovery under latent confounding

DAG with latent variables



Acyclic directed mixed graph:



Structural equation models associated with ADMG:

- SEM $X_j = f_j(PA_j, \varepsilon_j)$, $\forall\, j$
- If $X_a \leftrightarrow X_b$, then ε_a and ε_b may be dependent (share latent common cause). [Semi-Markov causal model]
- Causal discovery from observational data of (X_1, X_2, X_3, X_4) $\Longrightarrow$ Structure learning of ADMGs

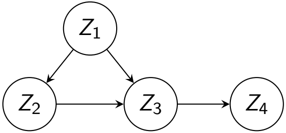
Causal discovery under latent confounding

Some ideas about structure learning of ADMGs:

- Two-step approach:
  1. Learn a maximal ancestral graph (MAG) or a DAG (or their equivalence class) to represent the CI among observed variables.
  2. Modify edge orientations or delete edges using score-based update.
- Incorporate experimental data
- Sequential intervention

Some relevant papers on DAG learning:
Huang and Zhou (2025) arXiv; Ye et al (2021) IEEE PAMI; Aragam and Zhou (2015) JMLR; Fu and Zhou (2013) JASA

Causal representation learning



$\xrightarrow{f}$ $(X_1, \ldots, X_p)$

latent causality ($\mathcal{G}_Z$) observed data (unstructured)

- $X = f(Z)$, where f is nonlinear and unknown.
- Given observed data X, wish to recover $Z = (Z_1, \ldots, Z_d)$
- Without intervention, at best we may recover

$$\widetilde{Z} = AZ,$$

A is $d \times d$ matrix and $\widetilde{Z}$ is entangled (mixing of $Z_1, \ldots, Z_d$).

Causal representation learning

Consider interventions

- Assume we can intervene on subsets Z_S.
- What parts of Z and $\mathcal{G}_Z$ can be recovered?
- How to recover them?
- How to design intervention targets S?

If f is linear, related to latent factor model: Kim and Zhou (2023) ICML.