Linear Causal Disentanglement via Interventions

Chandler Squires

My co-authors

Anna Seigal*

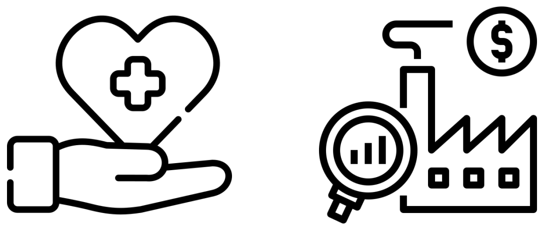
Salil Bhate

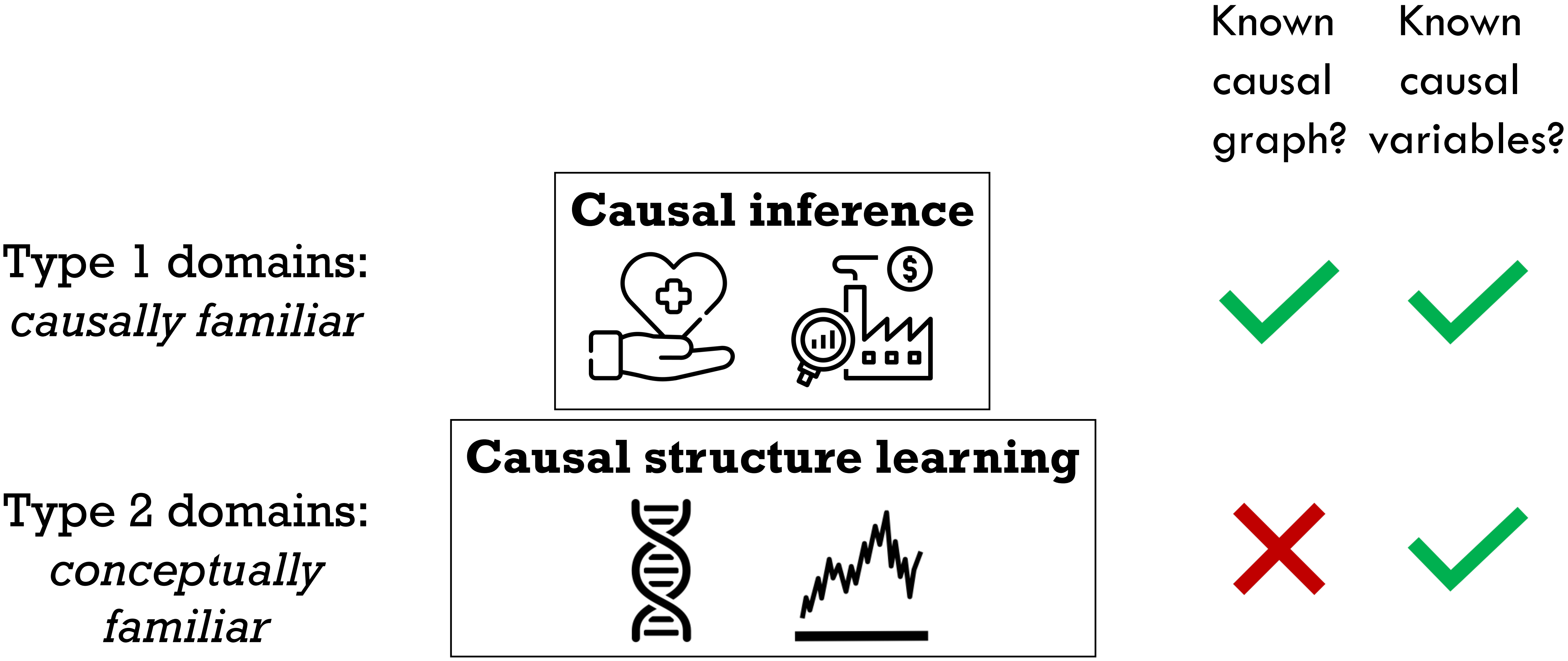
Caroline Uhler

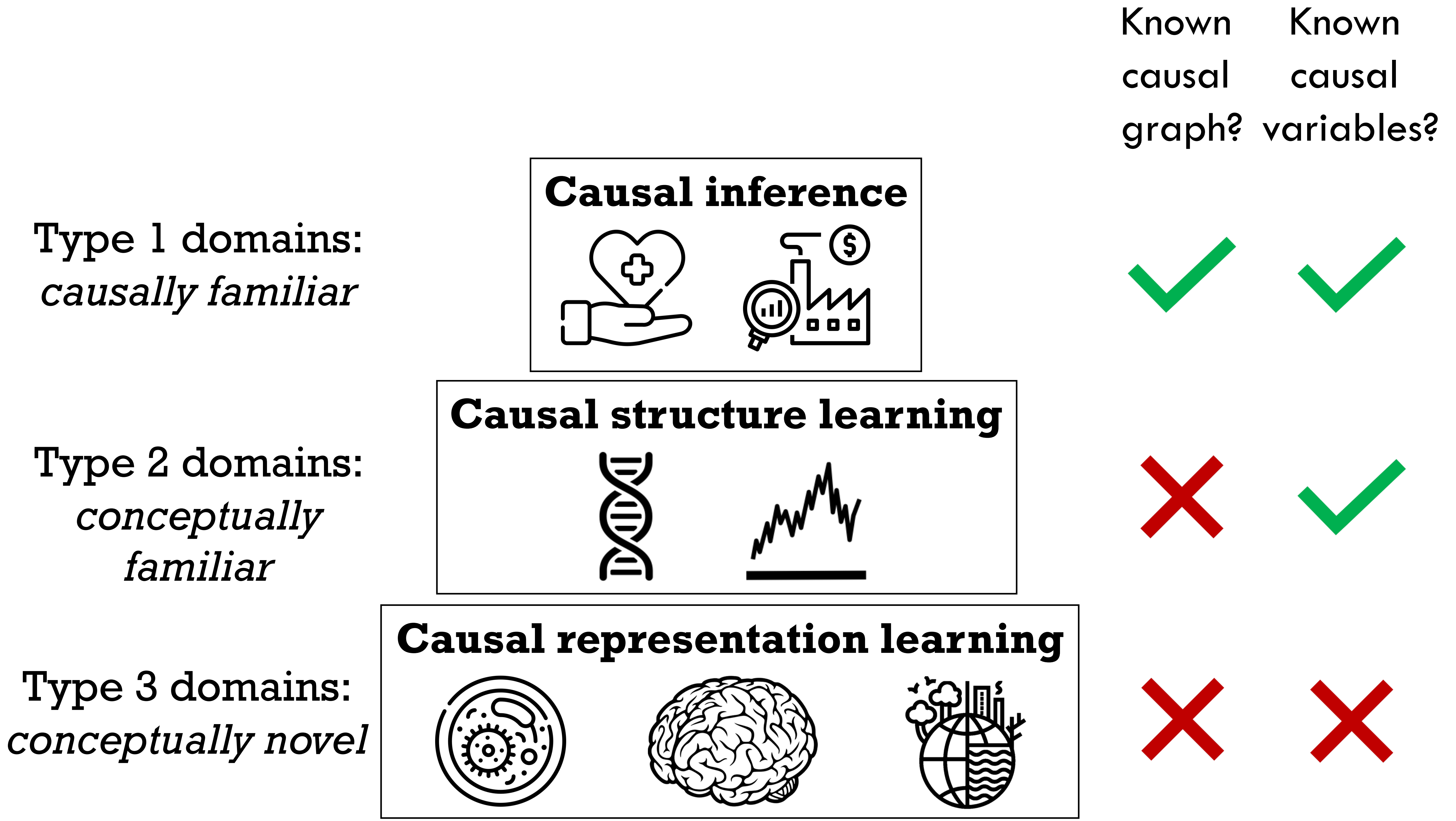
*Equal contribution

	Known causal graph?	Known causal variables?
Type 1 domains: *causally familiar*	✓	✓

Causal inference




Known causal graph?
Known causal variables?
Type 1 domains:
causally familiar
Causal inference
Type 2 domains:
conceptually familiar
Causal structure learning



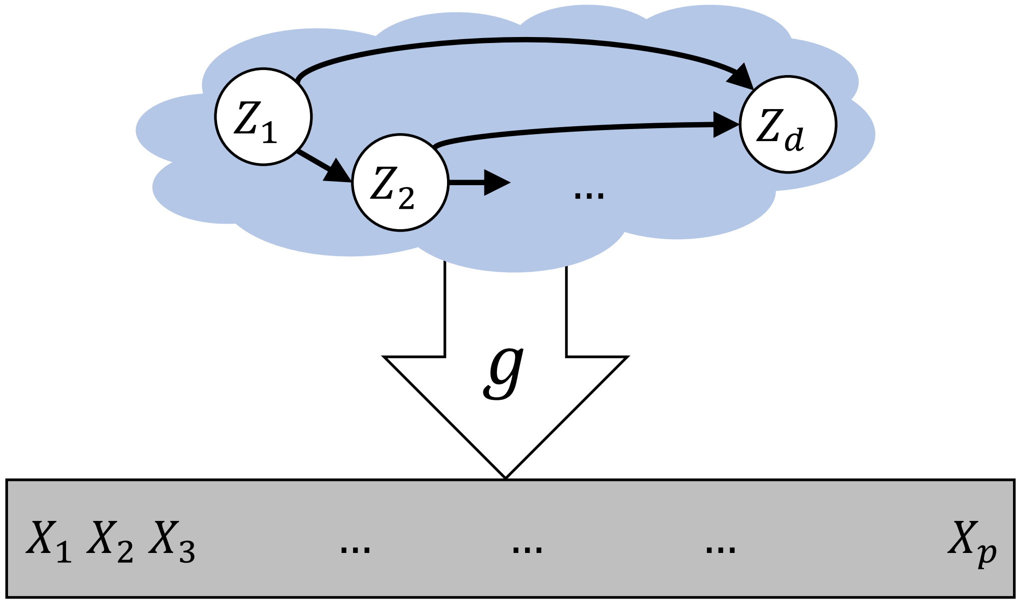
Known causal graph?
Known causal variables?
Causal inference
Causal structure learning
Causal representation learning
Type 1 domains:
causally familiar
Type 2 domains:
conceptually familiar
Type 3 domains:
conceptually novel

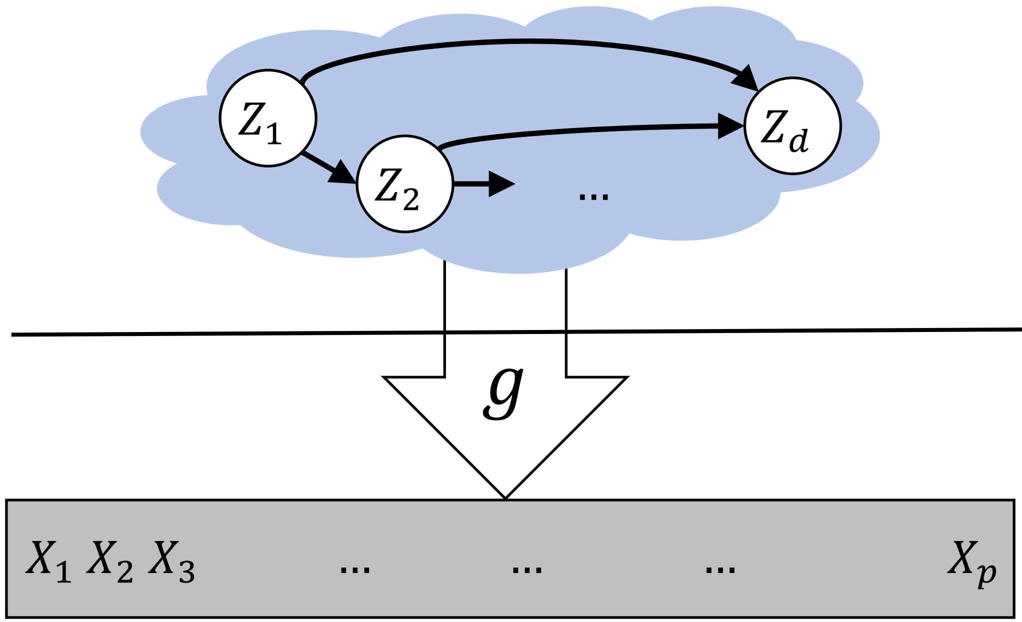
Causal Disentanglement

Macro-variables

Mixing function

Micro-variables



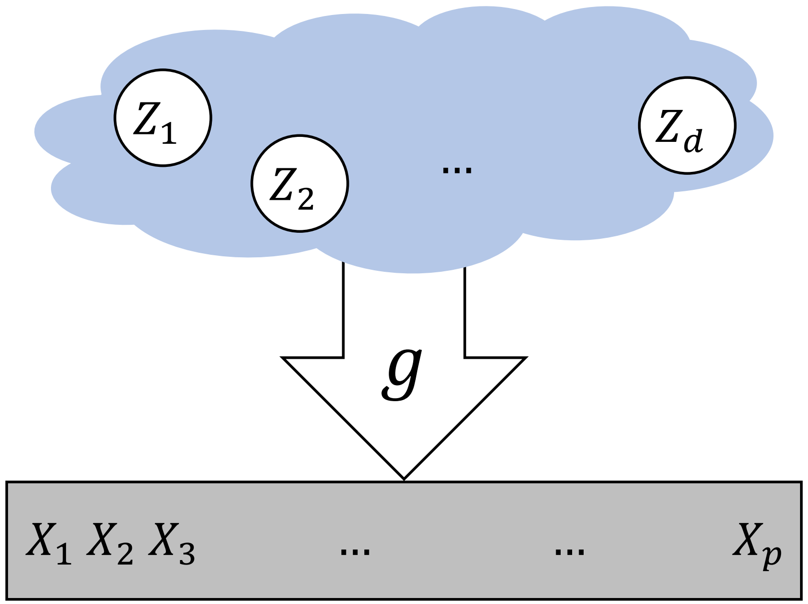


	Cellular Biology	Neuroscience
Macro-variables	• Protein concentrations • Cellular morphology (e.g. nucleus shape)	• Neurotransmitter concentrations • Reuptake rate
Micro-variables	• Fluorescent microscopy images • Gene expression (RNAseq)	• Neuroimaging data (fMRI) • Electrical activity (LFP)

Approaches to identifiability the causal disentanglement problem

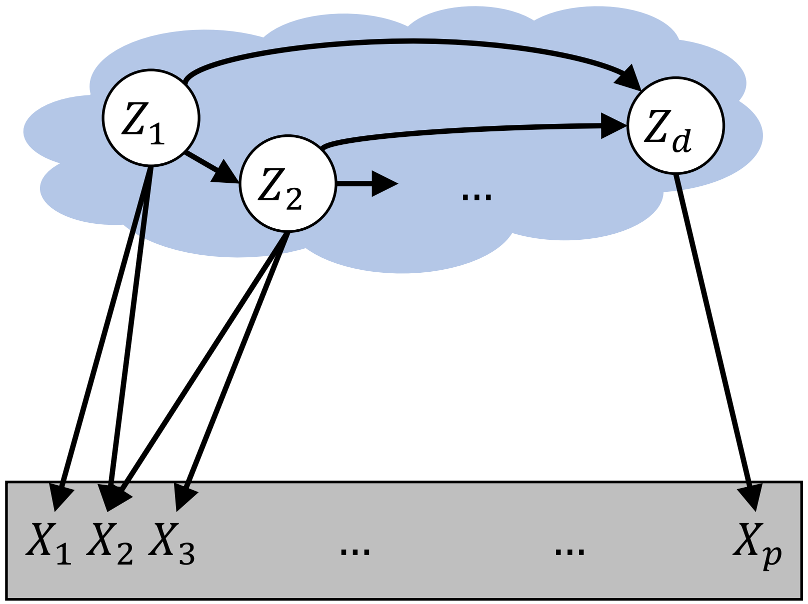
Identifiability = A unique model explains
the data we observe.

Restrict latent DAG $\mathcal{G}$



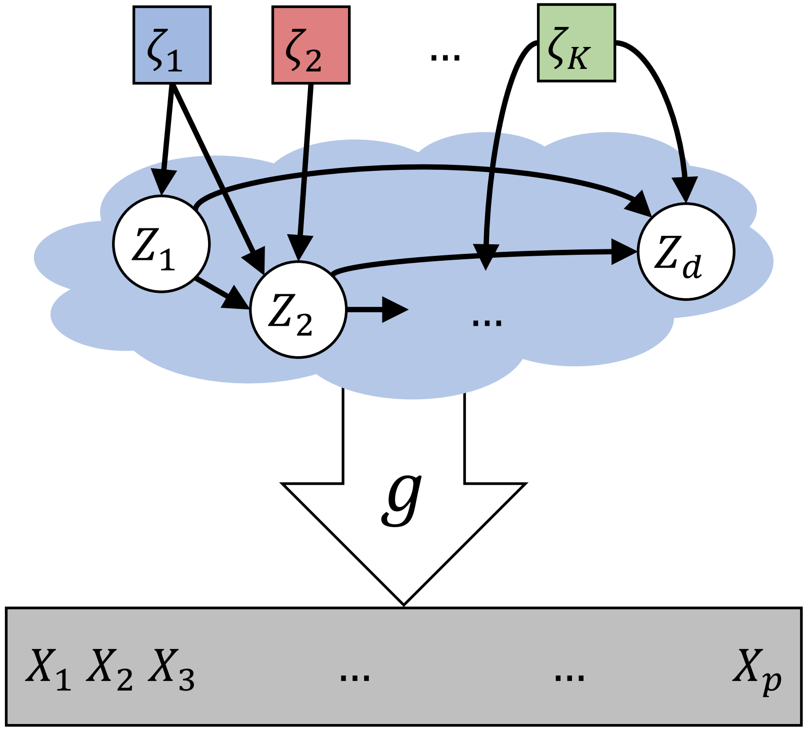
Linear ICA (Comon 1994)
Nonlinear ICA (Hyvärinen '19)

Restrict mixing function g

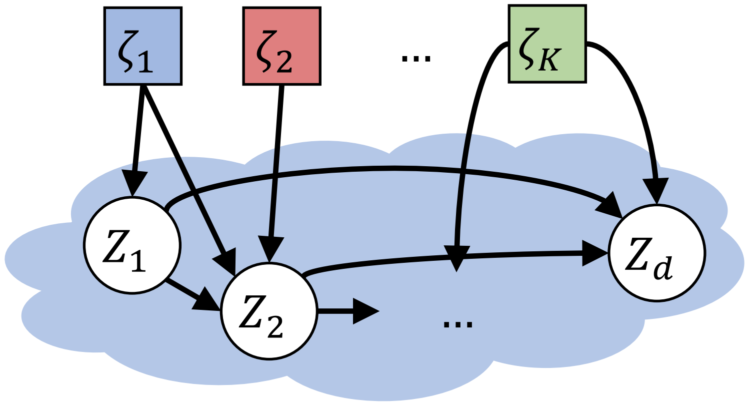


Most work on latent DAG recovery
(Silva '06, Halpern '15, Cai '19,
Kivva '21, Xie '20, Xie '22)

Learning from contexts



Squires '23
Liu '22, Ahuja '22, Varici '23

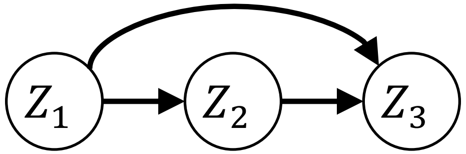


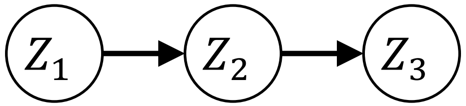
Control			…
$Z_1 = f_1(\varepsilon_1)$	$Z_1 = f_1'(\varepsilon_1)$	$Z_1 = f_1(\varepsilon_1)$	
$Z_2 = f_2(Z_1, \varepsilon_2)$	$Z_2 = f_2'(Z_1, \varepsilon_2)$	$Z_2 = f_2''(Z_1, \varepsilon_2)$	
$\vdots$	$\vdots$	$\vdots$	
$Z_d = f_d(Z_1, Z_2, \dots, \varepsilon_d)$	$Z_d = f_d(Z_1, Z_2, \dots, \varepsilon_d)$	$Z_d = f_d(Z_1, Z_2, \dots, \varepsilon_d)$	

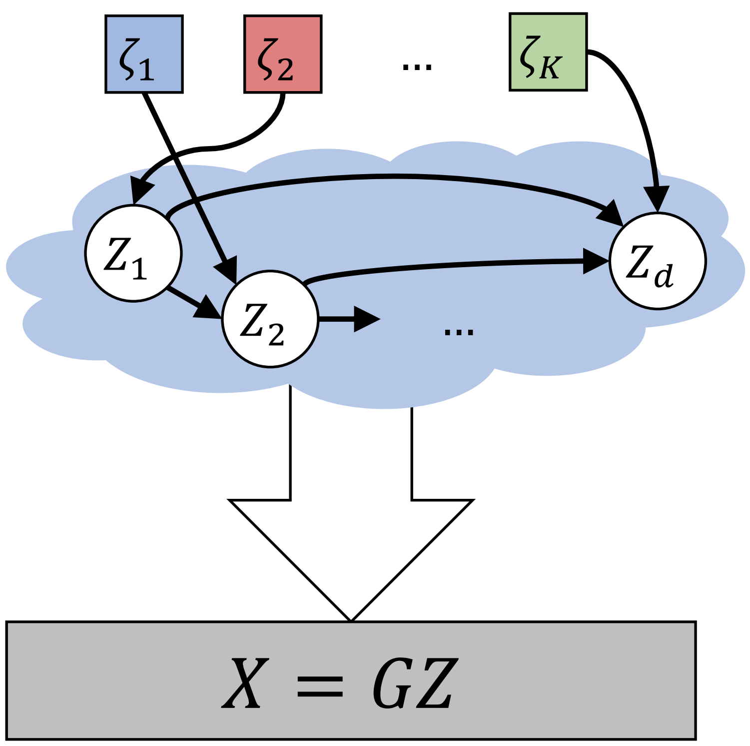
Do-intervention	Replaces mechanism with a constant	$Z_2 = \hat{z}_2$	More general
Perfect intervention	Removes dependence of parents	$Z_2 = f_2'(\varepsilon_2)$	
Soft intervention (mechanism shift)	Changes mechanism to any function	$Z_2 = f_2'(Z_1, \varepsilon_2)$	

ICML 2023 paper

- We show that causal disentanglement problem is identifiable from single-node perfect interventions, assuming:
  - A linear mixing function, and
  - A linear latent structural causal model.
- One intervention on each node is sufficient for identification.
- One intervention on each node is also, in the worst case, necessary for identification.
- For single-node soft interventions, the latent graph is only identifiable up to its transitive closure.







$G \in \mathbb{R}^{p \times d}$ with full column rank

Control

$$Z_1 = \sigma_1 \varepsilon_1$$
$$Z_2 = A_{12} Z_1 + \sigma_2 \varepsilon_2$$
$$\vdots$$
$$Z_d = A_{1d} Z_1 + A_{2d} Z_2 + \cdots + \sigma_d \varepsilon_d$$

$$Z_1 = \sigma_1' \varepsilon_1$$
$$Z_2 = A_{12} Z_1 + \sigma_2 \varepsilon_2$$
$$\vdots$$
$$Z_d = A_{1d} Z_1 + A_{2d} Z_2 + \cdots + \sigma_d \varepsilon_d$$

$$Z_1 = \sigma_1 \varepsilon_1$$
$$Z_2 = A_{12}' Z_1 + \sigma_2' \varepsilon_2$$
$$\vdots$$
$$Z_d = A_{1d} Z_1 + A_{2d} Z_2 + \cdots + \sigma_d \varepsilon_d$$

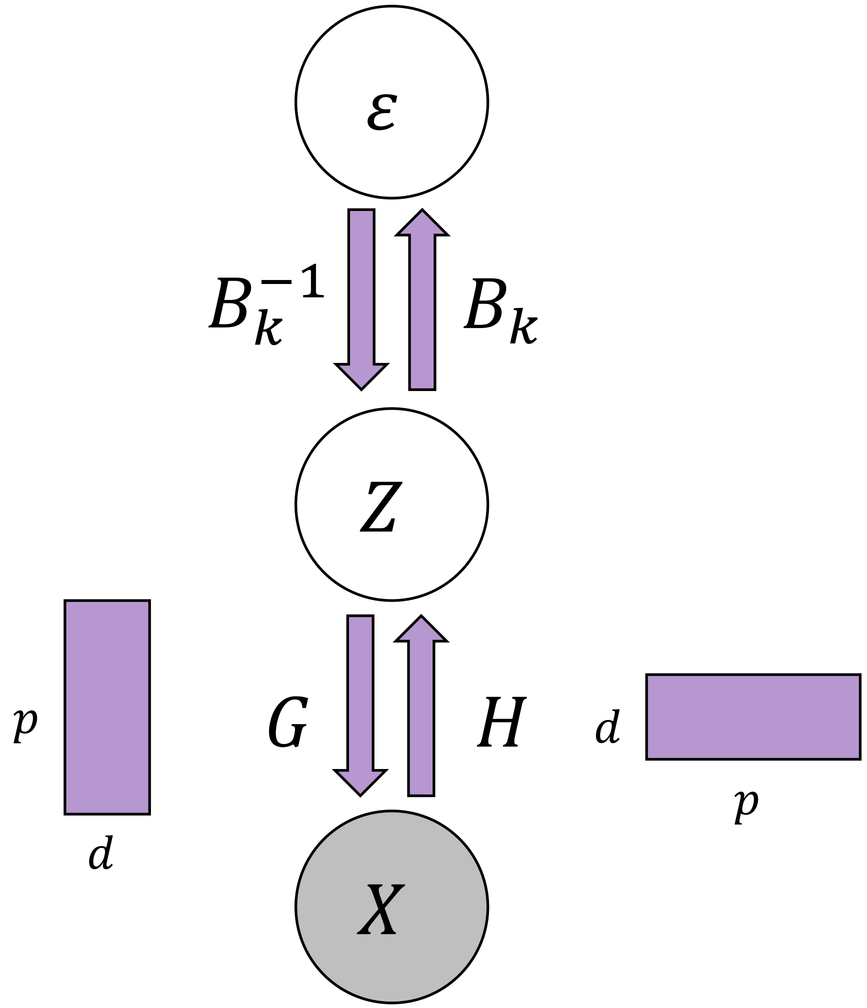
…

Compact version:

In context k, $Z = A_k Z + \Omega_k^{1/2} \varepsilon$.

Equivalently,

$Z = B_k^{-1} \varepsilon$ for $\boxed{B_k = \Omega_k^{-1/2} (I - A_k).}$ ← Upper triangular



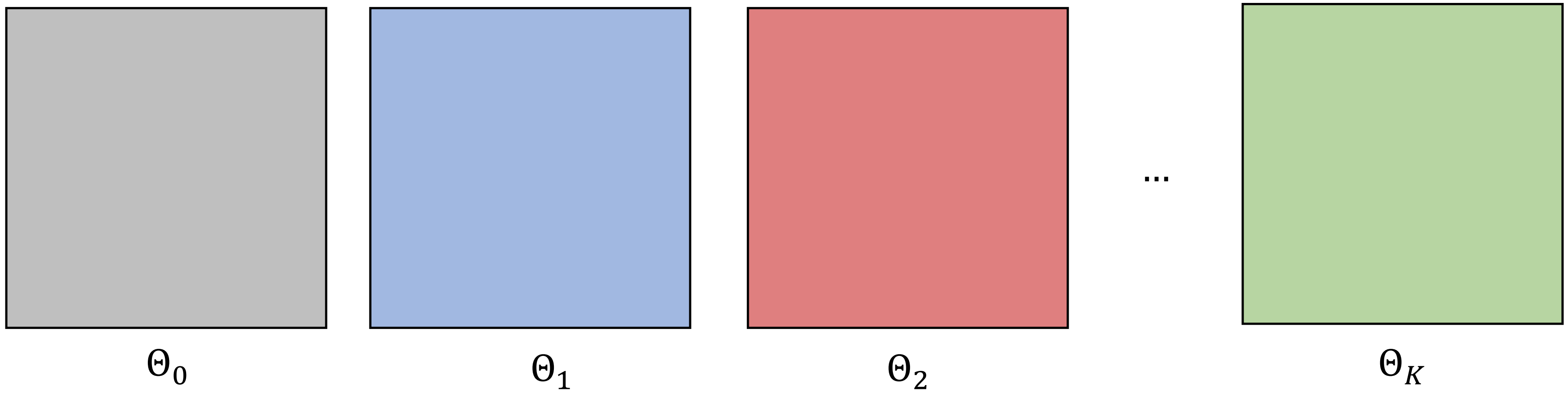
$H = G^{\dagger}$

$$\mathrm{Cov}(\varepsilon)^{-1} = I_d$$

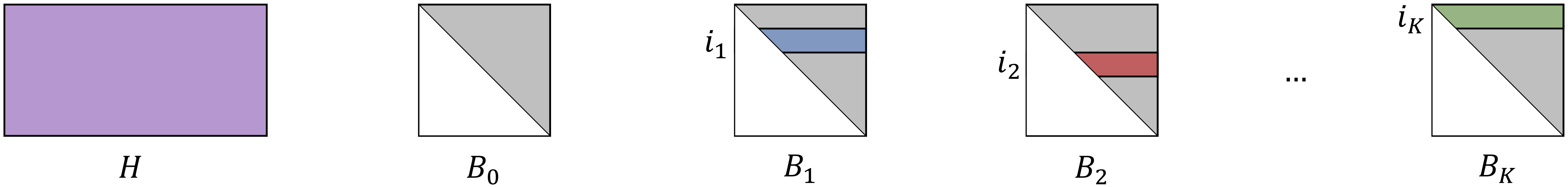
$$\mathrm{Cov}_k(Z)^{-1} = B_k^{\top} B_k$$

$$\Theta_k := \mathrm{Cov}_k(X)^{\dagger} = H^{\top} B_k^{\top} B_k H$$

Input:



Output:



such that $\Theta_k = H^\top B_k^\top B_k H$ **for all** k.

Without loss of generality, we assume:

1. The latent dimension d is known.
2. We know which context is observational.
3. For each node, there is only one intervention targeting that node.

Key identity

$$\boxed{\Theta_k - \Theta_0 = \left(H^\top B_k^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2} - \left(H^\top B_0^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2}}$$

$$\boldsymbol{v}^{\otimes 2} := \boldsymbol{v}\boldsymbol{v}^\top$$

First takeaway:

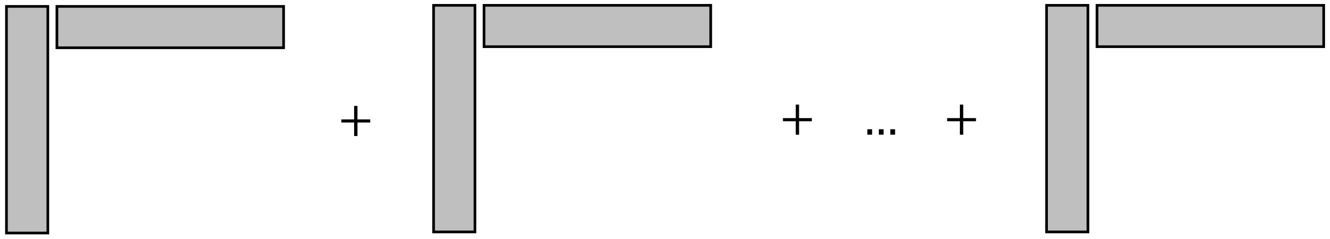
Let $r_{k,0} = rank(\Theta_k - \Theta_0)$.

Then $r_{k,0} = 1$ if i_k is a source node, otherwise $r_{k,0} = 2$.

Basic proof sketch of the key identity

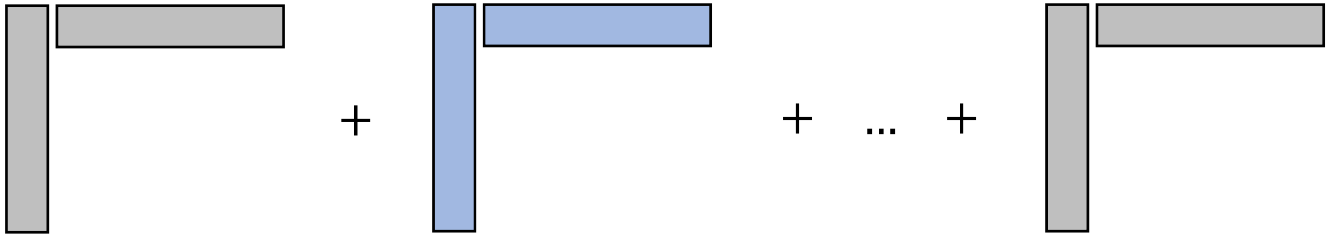
It's easy to show that we can decompose a product into a sum of rank-one terms.

$B_0^\top B_0 =$



$(B_0^\top \boldsymbol{e}_1)^{\otimes 2} \quad + \quad (B_0^\top \boldsymbol{e}_2)^{\otimes 2} \quad + \; \dots \; + \quad (B_0^\top \boldsymbol{e}_d)^{\otimes 2}$

$B_k^\top B_k =$



$\left(B_k^\top \boldsymbol{e}_1\right)^{\otimes 2} \quad + \quad (B_0^\top \boldsymbol{e}_2)^{\otimes 2} \quad + \; \dots \; + \quad (B_0^\top \boldsymbol{e}_d)^{\otimes 2}$

$$\Rightarrow B_k^\top B_k - B_0^\top B_0 = \left(B_k^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2} - \left(B_0^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2}$$

$$\Rightarrow \Theta_k - \Theta_0 = \left(H^\top B_k^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2} - \left(H^\top B_0^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2}$$

Sketch for a constructive proof of identifiability

$$\boxed{\Theta_k - \Theta_0 = \left(H^\top B_k^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2} - \left(H^\top B_0^\top \boldsymbol{e}_{i_k}\right)^{\otimes 2}}$$

Algorithm sketch:

1. Test ranks to find a source node i_k.
2. Recover the i_k-th row of H up to scale.
3. Remove* i_k and repeat.

*Involves projecting all matrices onto the orthogonal complement of the i_k-th row of H.

Ongoing work on identifiability

Non-linear latent model

Under submission, preprint coming soon.

Jiaqi Zhang

Multi-node interventions

Álvaro Ribot

Cathy Cai

Algorithmic Approaches

- Our constructive identifiability proof can be easily adapted to a finite-sample algorithm by replacing rank checks with hypothesis tests for rank constraints.
- However, this algorithm suffers from error propagation and empirically has poor finite-sample performance.

Causal Structure Learning Approaches

- **Constraint-based**
  - PC algorithm (Spirtes ‘00)
- **Hybrid**
  - MMHC (Tsamardinos ‘06)
  - GSP (Raskutti ‘18)
- **Score-based**
  - **Exact**
    - GOBNILP (Cussens ‘20)
  - **Greedy**
    - GES (Chickering ‘02)
  - **Gradient-based**
    - NO-TEARS (Zheng ‘18)

Causal Structure Learning: A Combinatorial Perspective (Squires and Uhler, 2022)

Penalized maximum-likelihood score

- Assume X is jointly Gaussian, and let $\ell(H, B; \mathcal{D})$ denote the log-likelihood of a dataset $\mathcal{D}$ with precision matrix $\Theta = H^\top B^\top B H$.
- Let $\mathcal{B}(i_1, \dots, i_K) \subseteq \left(\mathbb{R}^{d\times d}\right)^{\otimes(K+1)}$ denote all tuples $(B_0, B_1, \dots, B_K)$ of upper triangular matrices such that B_k can be derived from B_0 using an intervention on i_k.
- Given datasets $\mathcal{D}_0, \mathcal{D}_1, \dots, \mathcal{D}_K$, we want to solve

$$\max_{\substack{H \in \mathbb{R}^{d\times p} \\ i_1,\dots,i_K \\ (B_0,B_1,\dots,B_K) \in \mathcal{B}(i_1,\dots,i_K)}} \sum_{k=0}^{K} \ell(H, B_k; \mathcal{D}_k) + \rho(B_0)$$

Sparsity penalty, e.g. BIC

Simplified problem

- Fix $i_1, \dots, i_K$ and a sparsity pattern for B_0 (given by binary matrix A).
- Let $\mathcal{B}_A(i_1, \dots, i_K)$ denote tuples of matrices which are consistent with the sparsity pattern.
- Then the simplified optimization problem becomes:

$$\max_{\substack{H \in \mathbb{R}^{d \times p} \\ (B_0, B_1, \dots, B_K) \in \mathcal{B}_A(i_1, \dots, i_K)}} \sum_{k=0}^{K} \ell(H, B_k; \mathcal{D}_k)$$

Let’s Chat!

Selected References

Ahuja, K., Wang, Y., Mahajan, D., & Bengio, Y. (2022) Interventional Causal Representation Learning.

Cai, R., Xie, F., Glymour, C., Hao, Z., & Zhang, K. (2019). Triad constraints for learning causal structure of latent variables.

Chickering, D. M. (2002). Optimal structure identification with greedy search.

Comon, P. (1994). Independent component analysis, a new concept?

Cussens, J. (2020). GOBNILP: Learning Bayesian network structure with integer programming.

Halpern, Y., Horng, S., & Sontag, D. (2015). Anchored discrete factor analysis.

Hyvarinen, A., Sasaki, H., & Turner, R. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning.

Kivva, B., Rajendran, G., Ravikumar, P., & Aragam, B. (2021). Learning latent causal graphs via mixture oracles.

Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., Hengel, A. V. D., ... & Shi, J. Q. (2022). Identifying Weight-Variant Latent Causal Models.

Raskutti, G., & Uhler, C. (2018). Learning directed acyclic graph models based on sparsest permutations.

Silva, R., Scheines, R., Glymour, C., Spirtes, P., & Chickering, D. M. (2006). Learning the Structure of Linear Latent Variable Models.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm.

Varici, B., Acarturk, E., Shanmugam, K., Kumar, A., & Tajer, A. (2023). Score-based Causal Representation Learning with Interventions.

Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., & Zhang, K. (2020). Generalized independent noise condition for estimating latent variable causal graphs.

Xie, F., Huang, B., Chen, Z., He, Y., Geng, Z., & Zhang, K. (2022). Identification of linear non-Gaussian latent hierarchical structure.

Zheng, X., Aragam, B., Ravikumar, P. K., & Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning.