Scalable Causal Structure Learning via Amortized Condtional Independence Testing

James Leiner[1], Brian Manzo[2], Aaditya Ramdas[1,3], Wesley Tansey[4]

[1]Department of Statistics and Data Science, Carnegie Mellon University

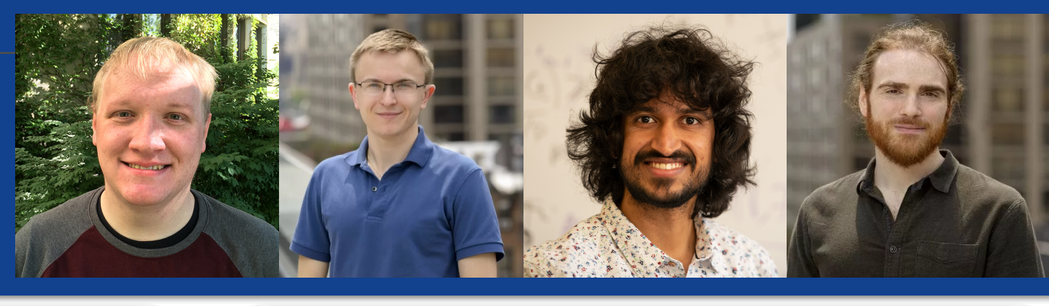
[2]Machine Learning Department, Carnegie Mellon University

[3]Department of Statistics, University of Michigan

[4]Computational Oncology, Memorial Sloan Kettering Cancer Center

Link to paper

(arXiv: 2310.16626)

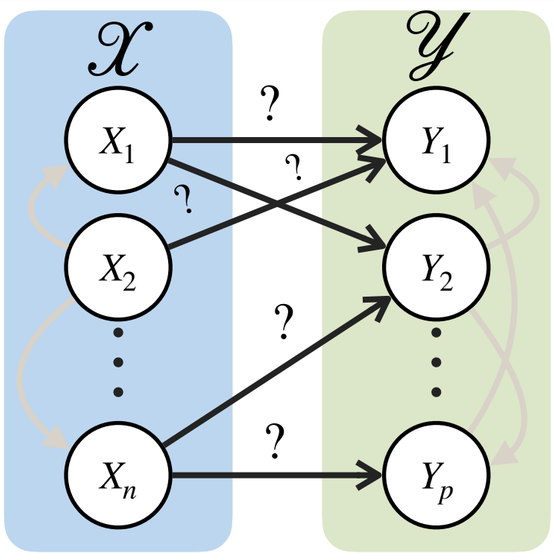


Problem Setup

Suppose we observe graph data containing two sets of nodes, $\mathcal{X}$ and $\mathcal{Y}$. Assume that:

- No edge can be directed from $\mathcal{Y}$ to $\mathcal{X}$
- Edges between nodes in the same set can be oriented in any direction

Key Question: Which edges exist between $\mathcal{X}$ and $\mathcal{Y}$?



Motivating Application

Any dataset where groups of variables are known to be ordered in time will have this structure.

We consider a cancer dataset as a running example:

- $\mathcal{X}$ contains binary variables indicating whether certain mutations are contained in the primary tumor site
- $\mathcal{Y}$ contains binary variables indicating whether metastases have developed in secondary locations

Discovering connections of the form $X_j \to Y_k$ allow us to proactively screen at-risk patients and better understand the progression of the disease.

Causal p-values

Under certain assumptions, a hypothesis that an edge is present between nodes X_j and Y_k. is reducible to testing for conditional independence between X_j and Y_k given other sets of nodes on the graph.

Proposition 1

Assume a graph $\mathcal{G} := (\mathcal{X}, \mathcal{Y})$ satisfies the global directed Markov property and the probability distribution is d-separation faithful.

Assume no element can be directed from any element in $\mathcal{Y}$ to any element in $\mathcal{X}$.

Then, there is an edge between $X_j \in \mathcal{X}$ and $Y_k \in \mathcal{Y}$ if and only if X_j and Y_k are conditionally dependent given $S \cup X_{-j}$ for all $S \subseteq Y_{-k}$.

H_0: $X_j \to Y_k$ is absent → H_0: There exists $S \subseteq Y_{-k}$ such that $X_j \perp Y_k \mid S, X_{-j}$

H_0: $X_j \to Y_k$ is absent → H_1: $X_j \not\perp Y_k \mid S, X_{-j}$ for all $S \subseteq Y_{-k}$

This implies that… $p_{X_j \to Y_k} \leq \max_{S \subseteq Y_{-k}} p_{X_j \perp Y_k \mid S, X_{-j}}$

So $p_{X_j \to Y_k}$ can be bounded by an exhaustive computation of conditional independence tests over all possible conditioning subsets, but this is not always feasible.

Causal Search

In lieu of brute force computation, our strategy consists of two steps:

1. Find a function $T_{X_j, Y_k}(\cdot)$ that takes in S as an input and outputs a statistic for the hypothesis $X_j \perp Y_k \mid S, X_{-j}$
2. Use discrete optimization to find $\hat{S} := \arg\min_{S \subseteq Y_{-k}} T_{X_j, Y_k}(S)$

Generalized Covariance Measure (GCM)

We focus on the GCM [Shah and Peters, 2018]. This tests whether the expected conditional covariance,

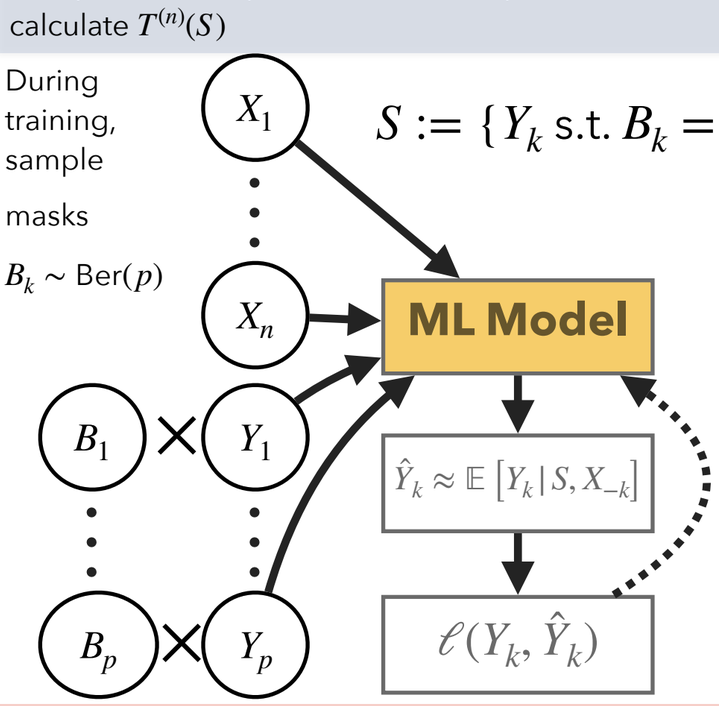
$$\mathbb{E}\left[\mathbb{E}[X_j Y_k \mid S, X_{-j}] - \mathbb{E}[X_j \mid S, X_{-j}]\mathbb{E}[Y_k \mid S, X_{-j}]\right]$$

is non-zero. The method's statistic $T^{(n)}$ is computed from well-trained model-based estimates $\hat{X}_j$ and $\hat{Y}_k$ targeting $\mathbb{E}\left[Y_k \mid S, X_{-j}\right]$ and $\mathbb{E}\left[X_j \mid S, X_{-j}\right]$.[1]

Amortized Predictive Models

Desiderata: train models $\hat{Y}_k(\cdot)$ and $\hat{X}_j(\cdot)$ that takes S as an inputs and outputs conditional expectations to calculate $T^{(n)}(S)$

During training, sample masks $B_k \sim \text{Ber}(p)$



When using model, manually let $B_k = 0$ for all $Y_k \notin S$ (given arbitrary choice of S).

Training process mimics process of an end user arbitrarily evaluating different conditioning subsets.

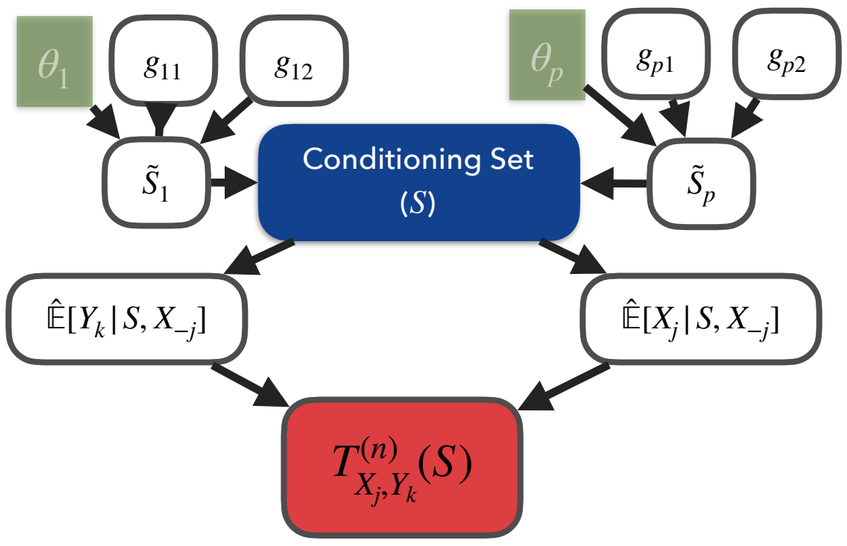
Gumbel-Softmax Optimization

Desiderata: Learn $\arg\min_{\theta_1, \ldots, \theta_p} \mathbb{E}\left[T_n(S)\right]$ where $1_{Y_k \in S} \sim \text{Ber}(\theta_k)$

Replace $\frac{\partial T_n}{\partial S} \approx \frac{\partial T_n}{\partial \tilde{S}}$ to enable back propagation. $\tilde{S}$ is a continuous relaxation of S using the Gumbel-Softmax trick [Jang et al., 2017].

$$\tilde{S}_i = \frac{\exp\left((\log\theta_i + g_{i1})/\tau\right)}{\exp\left((\log\theta_i + g_{i1})/\tau\right) + \exp\left((\log(1-\theta_i) + g_{i2})/\tau\right)}$$

$g_{i1}, g_{i2} \sim \text{Gumbel}(0,1)$ $\tau \to 0$ approximates discrete distribution



Results

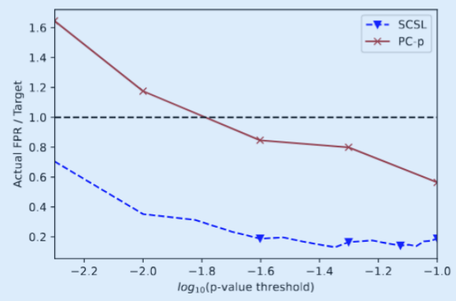
Dataset: $n = 22{,}352$ combining metastatic events with pre-metastatic tumor mutation info [Nguyen et al., 2022]

Semi-Synthetic Simulations

We posit a logistic model $\mathcal{P}$ relating $\mathcal{X}$ and $\mathcal{Y}$. For reach patient, we calculate $\pi_i := \mathcal{P}(\mathcal{Y}_i \mid \mathcal{X}_i)$ as the likelihood of this row under the assumed model.

Construct new dataset by sampling $\text{Cat}(\pi_1, \ldots \pi_n)$. This preserves marginal distributions of $\mathcal{X}$ and $\mathcal{Y}$ while providing ground truth knowledge of causal relationships

The only other causal discovery method that produces p -values has inflated type I error, while SCSL is conservative.

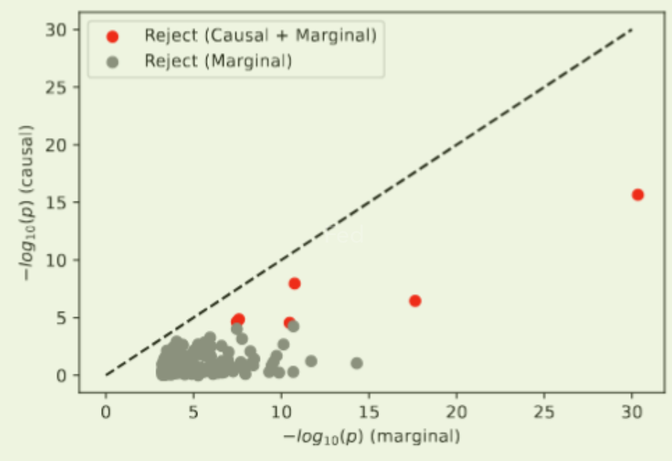


SCSL also often has improved performance even when compared to methods not designed for frequentist error control…

n	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{Y}\rvert$	F1 Score: SCSL	PC-p	PC	BOSS	CCD	FCI	FGES	GFCI	GRASP	GRaSP-FCI
200	5	5	**0.26**	0.24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	10	10	0.07	**0.10**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	15	15	**0.09**	0.07	0.0	0.0	0.0	0.0	0.03	0.03	0.03	0.06
	20	20	0.04	0.04	0.02	**0.11**	0.02	0.02	0.04	0.04	0.06	0.06
2000	5	5	**0.71**	0.38	0.0	0.18	0.0	0.0	0.17	0.17	0.0	0.17
	10	10	**0.30**	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	15	15	**0.12**	0.10	0.0	0.03	0.0	0.0	0.0		0.0	0.03
	20	20	**0.08**	0.06	0.0	0.04	0.0	0.0	0.02		0.04	0.04
20,000	5	5	0.87	0.57	**0.95**	0.84	0.95	0.82	0.95	0.89	0.84	0.89
	10	10	**0.78**	0.37	0.29	0.46	0.29	0.06	0.46	0.24	0.38	0.24
	15	15	**0.49**	0.16		0.15			0.13	0	0.15	0.06
	20	20	**0.33**	0.06		0.06			0.04	0.02	0.08	

Real Data Results

The original study identified 161 discoveries rejected using associative p-values with a Benjamini-Hochberg (BH) adjustment. Only 6 discoveries remain when substituting causal p-values with the same BH adjustment.



Primary	Gene	Secondary	p-value: Causal	Marginal
Breast	CDH1	Lung	3.5×10^{-7}	2.3×10^{-18}
Colon	KRAS	Lung	1.4×10^{-5}	2.6×10^{-8}
Liver	TERT	Liver	2.3×10^{-5}	3.4×10^{-8}
Lung	EGFR	CNS (Brain)	2.8×10^{-5}	3.3×10^{-11}
Pancreas	KRAS	Lymph	2.2×10^{-16}	4.5×10^{-31}
Pancreas	TP53	Lymph	1.1×10^{-8}	1.7×10^{-11}

Footnotes

1: Letting $R_i = \left(X_j^i - \widehat{X}_j^i\right)\left(Y_k^i - \hat{Y}_k^i\right)$, then under the null (and given appropriate regularity conditions ensuring fast convergence of the estimated conditional means),

$$T^{(n)}_{X_j, Y_k} := \frac{\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^n R_i}{\left(\frac{1}{n}\sum_{i=1}^n R_i^2 - \left(\frac{1}{n}\sum_{r=1}^n R_r\right)^2\right)^{1/2}} \approx N(0,1)$$