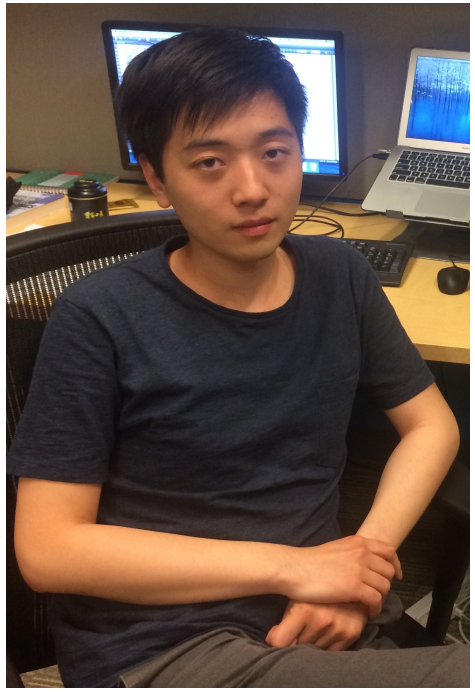
Causal Structure Learning with Unknown Mechanism Shifts

Chandler Squires, Yuhao Wang, Caroline Uhler

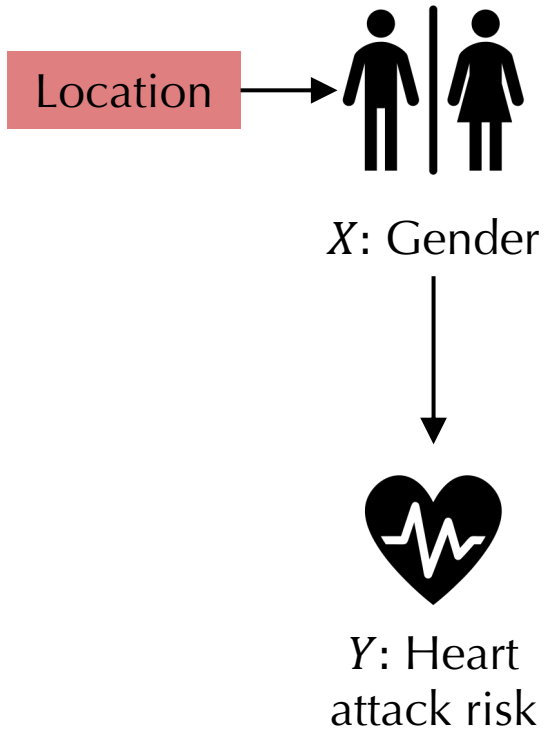


Yuhao
Wang

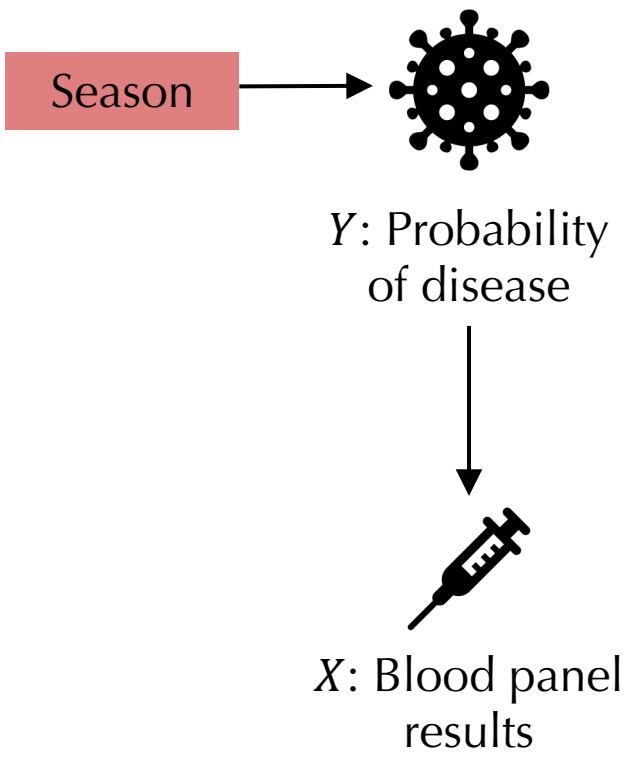
Caroline
Uhler

Structural causal models (SCMs) provide a natural language for distribution shifts

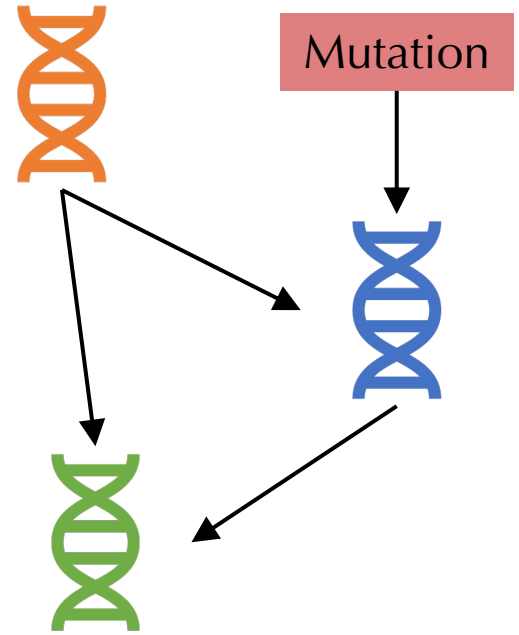
Covariate shift



Label shift



Disease genomics



Mechanism shifts are changes to the data-generating process defined by an SCM.

A (Markovian) **structural causal model** over random variables $X_1, \ldots, X_p$ is a set of equations of the form

$$X_i = f_i(X_{\mathrm{pa}(i)}, U_i)$$

Where $\mathrm{pa}(i) \subseteq [p] \setminus \{i\}$ are called the **parents** of i and

U_i are independent exogenous noise variables.

The associated **causal graph** $\mathcal{G}^*$ consists of nodes $1, \ldots, p$ and edges $j \to i$ for all $i = 1, \ldots, p$ and $j \in \mathrm{pa}(i)$. We assume $\mathcal{G}^*$ is acyclic.

Mechanism shifts are changes to the data-generating process defined by an SCM.

Given a structural causal model, a **mechanism change** on a set of variables C consists of replacing the functions f_i for $i \in C$.

For example, a gene editing technique called a *knockdown* can reduce the activity of a gene.

$$X_2 = 300 \cdot \sigma(X_1^{0.42} + X_1^{0.61} + U_1)$$

$$\Downarrow$$

$$X_2 = 50 \cdot \sigma\left(X_1^{0.28} + X_1^{0.45} + U_1\right)$$

Mechanism shifts are changes to the data-generating process defined by an SCM.

Note that a **do-intervention** (setting a variable deterministically to some value) is a special case of a mechanism change.

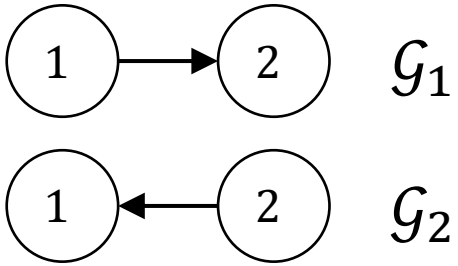
$$X_2 = 300 \cdot \sigma(X_1^{0.42} + X_1^{0.61} + U_1)$$

$$\Downarrow$$

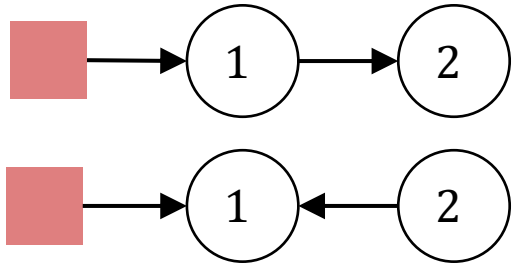
$$X_2 = 0$$

Mechanism changes are also called **soft** interventions, and the set C are called the intervention **targets**.

Shifts help us to learn causal structure from data.



$\mathcal{G}_1$ and $\mathcal{G}_2$ are **Markov equivalent** (indistinguishable from observational data).



However, they are not **interventionally Markov equivalent** for an intervention on $C = \{1\}$.

Shifts help us to learn causal structure from data.

Several algorithms have been proposed for using interventional data to learn causal structure. To name a few:

- Greedy Interventional Equivalence Search (GIES)
  - Hauser and Bühlmann 2012
- Interventional Greedy Sparsest Permutation (IGSP)
  - Wang et al. 2017, Yang et al. 2018

These approaches assume that the intervention targets are *known*.

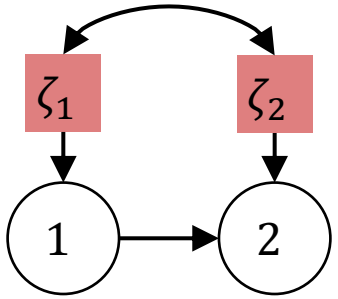
In practice, the targets may be (partially) unknown – e.g., when gene knockdowns have *off-target* effects.

We can learn the target of interventions alongside learning the causal structure.

- Eaton and Murphy 2007: A dynamic programming approach
- Mooij et al. 2020: Joint Causal Inference (JCI)
- Squires et al. 2020: Unknown-Target IGSP (UT-IGSP)
- More recently:
  - Jaber et al. 2020: Ψ-FCI
  - Brouillard et al. 2020: Differentiable Causal Discovery with Interventions (DCDI)

We can learn the target of interventions alongside learning the causal structure.

We introduce a binary indicator variable ζ_k for each intervention (environment) $k = 1, \ldots, K$, with the intervention targets as children.



$\zeta_1, \zeta_2 = (0, 0)$: observational
$\zeta_1, \zeta_2 = (1, 0)$: data from intervention 1
$\zeta_1, \zeta_2 = (0, 1)$: data from intervention 2

$$X_1 = f_1(U_1)\,\mathbb{1}_{\zeta_1=0} + f_1'(U_1)\mathbb{1}_{\zeta_1=1}$$

$$X_2 = f_2(X_1, U_2)\,\mathbb{1}_{\zeta_2=0} + f_2'(X_1, U_2)\mathbb{1}_{\zeta_2=1}$$

Ordering-based approaches to causal structure learning

Causal structure learning is simple when the variable order is known.

Let $\mathcal{G}$ be a DAG over nodes $1, \ldots, p$.
Let $\mathcal{I}(\mathcal{G}) = \{ (i, j, C) : i \text{ and } j \text{ are d-separated by } C \text{ in } \mathcal{G}\}$.

Let $\mathbb{P}_X$ be a distribution on variables $X_1, \ldots, X_p$.
Let $\mathcal{I}(\mathbb{P}_X) = \left\{ (i, j, C) : X_i \text{ and } X_j \text{ are independent given } C \text{ in } \mathbb{P}_X \right\}$.

$\mathcal{G}$ is an **independence map (I-MAP)** of $\mathbb{P}_X$ if $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(\mathbb{P}_X)$.

For example, the complete graph $\mathcal{G}^c$ is an I-MAP of any distribution since $\mathcal{I}(\mathcal{G}^c) = \emptyset$.

Causal structure learning is simple when the variable order is known.

$\mathcal{G}$ is a **minimal I-MAP** it is an I-MAP and no subgraph of $\mathcal{G}$ is an I-MAP.

Let π be a permutation of $[p]$. A DAG is **compatible** with π if $i \to j$ in $\mathcal{G}$ implies that $i <_\pi j$.

Then (generically) there is a unique minimal I-MAP $\mathcal{G}_\pi$ that is compatible with π.

$\mathcal{G}_\pi$ has the edge $i \to j$ if $i <_\pi j$ and

X_i is not independent of X_j given $X_{pre_\pi(j)\setminus\{i\}}$.

The true variable ordering generates the sparsest minimal I-MAP.

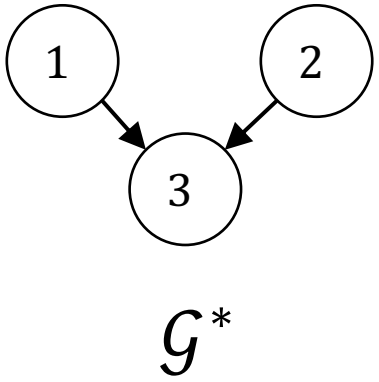
Raskutti and Uhler 2018: Let $S(\pi)$ denote the number of edges in $\mathcal{G}_\pi$. Then (generically) for any $\pi^* \in \operatorname{argmin} S(\pi)$, we have that $\mathcal{G}_{\pi^*}$ is Markov equivalent to $\mathcal{G}^*$.

The **sparsest permutation** algorithm: solve the above combinatorial optimization problem by enumerating over permutations.

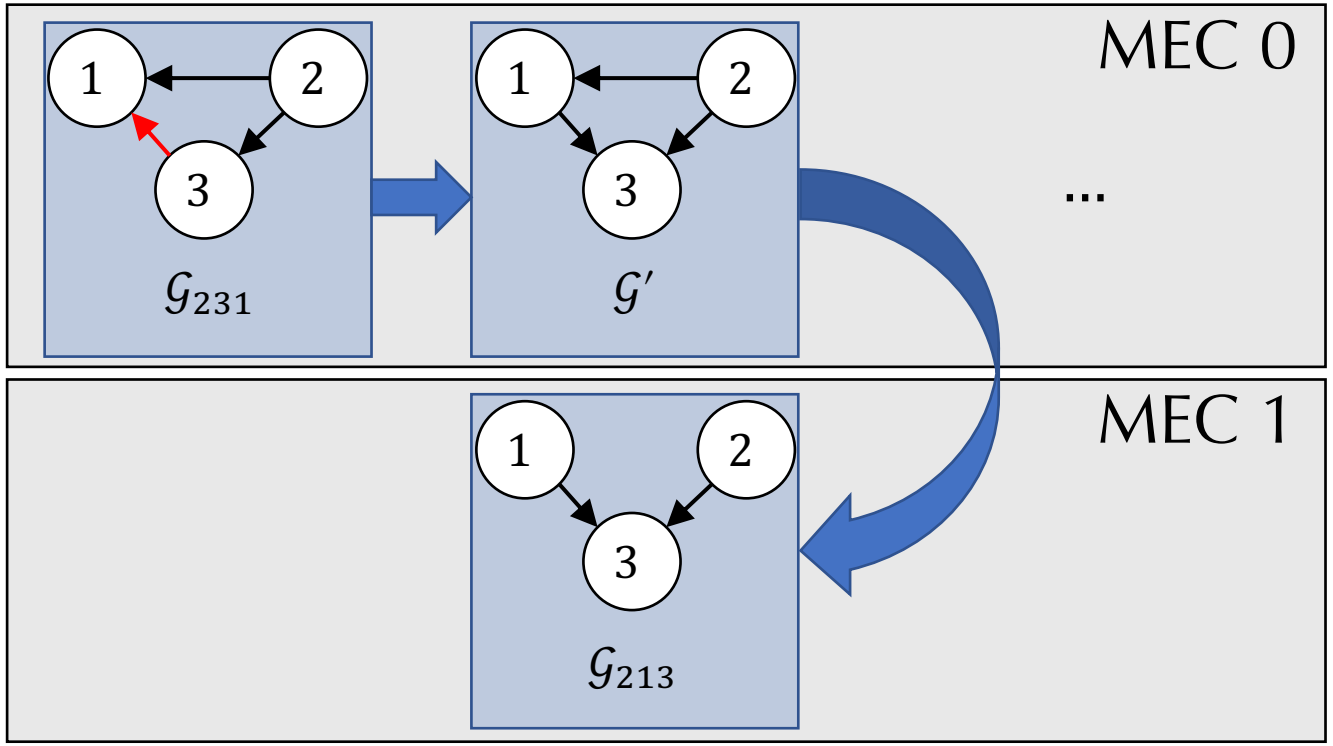
This approach isn't scalable, since the number of permutations of p items is $p!$

The true variable ordering generates the sparsest minimal I-MAP.

The **greedy sparsest permutation (GSP)** algorithm instead starts from some initial permutation π_0 and iteratively swaps the order of variables to find sparser permutations.



$\mathcal{G}^*$

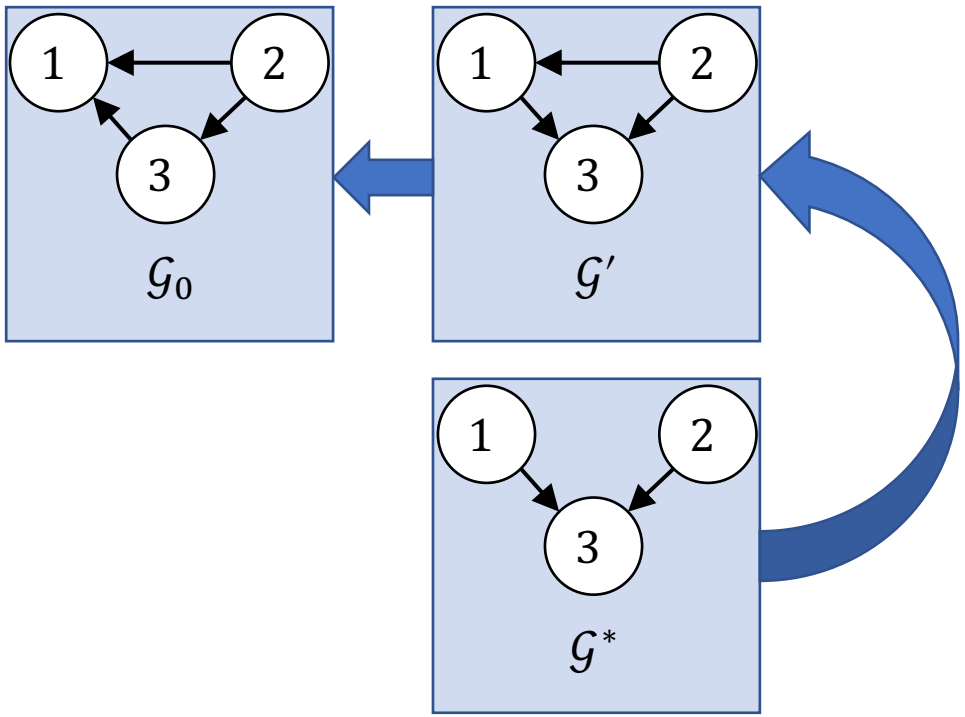


"Chickering sequences" guarantee the consistency of greedy search.

Chickering 2002: Let $\mathcal{G}_0$ be an I-MAP of $\mathcal{G}^*$ that is not Markov equivalent to $\mathcal{G}^*$, i.e., $\mathcal{I}(\mathcal{G}_0) \subsetneq \mathcal{I}(\mathcal{G}^*)$.

Then, some graph in the MEC of $\mathcal{G}_0$ is not a minimal I-MAP of $\mathcal{G}^*$.

The proof is constructive: given $\mathcal{G}^*$ and $\mathcal{G}_0$, it gives a sequence of edge additions and edge flips from $\mathcal{G}^*$ to $\mathcal{G}_0$.



UT-IGSP modifies GSP to search over DAGs which include the intervention variables.

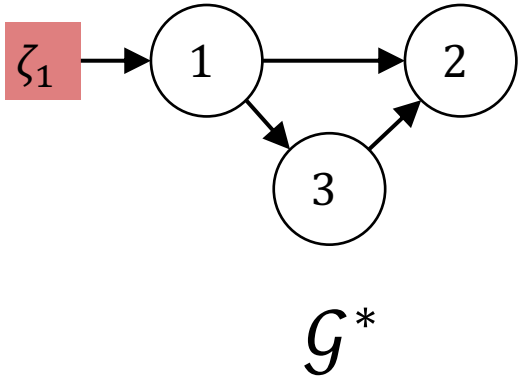
By construction, the intervention variables and the original "system" variables satisfy two forms of background knowledge:

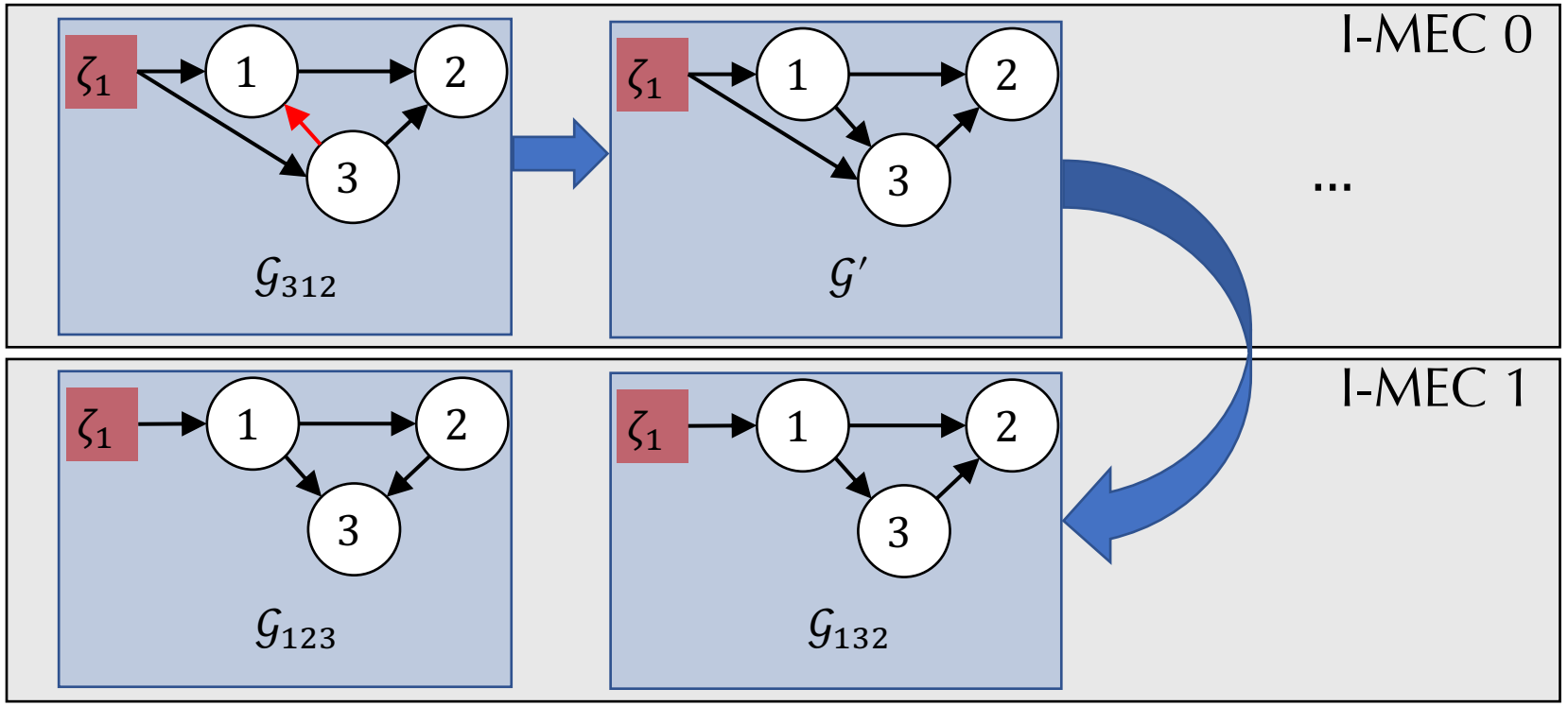
- **Exogeneity:** Intervention variables are upstream of system variables.
- **Known adjacencies**: There is a complete graph over intervention variables, since they are deterministically related.

UT-IGSP only considers permutations which obey exogeneity, and uses the known adjacencies in constructing each $\mathcal{G}_\pi$.

The Chickering sequence never reverses an edge which is oriented the same way in $\mathcal{G}_0$ and $\mathcal{G}^*$, so UT-IGSP is still consistent.

UT-IGSP modifies GSP to search over DAGs which include the intervention variables.





Our review article

Causal Structure Learning: a Combinatorial Perspective

- Permutation-based causal structure learning
- Causal structure learning under unobserved confounding
- Bayesian approaches for uncertainty quantification in causal structure learning
- Open problems in causal structure learning and related areas

Zooming out

- Structural causal models provide a natural language for distribution shift.
- As a field, we have a number of algorithms for learning causal models from observation and/or interventional data.
- Two major limitations:
  - These algorithms assume that the variables we observe correspond to those in a SCM. In many domains (e.g., image processing), we may need to *learn* causal representations.
  - These algorithms are designed with the goal of learning causal structures, not with downstream tasks in mind. *Targeted* approaches may perform better on downstream tasks.

Thanks!

Slides will be made available at http://chandlersquires.com

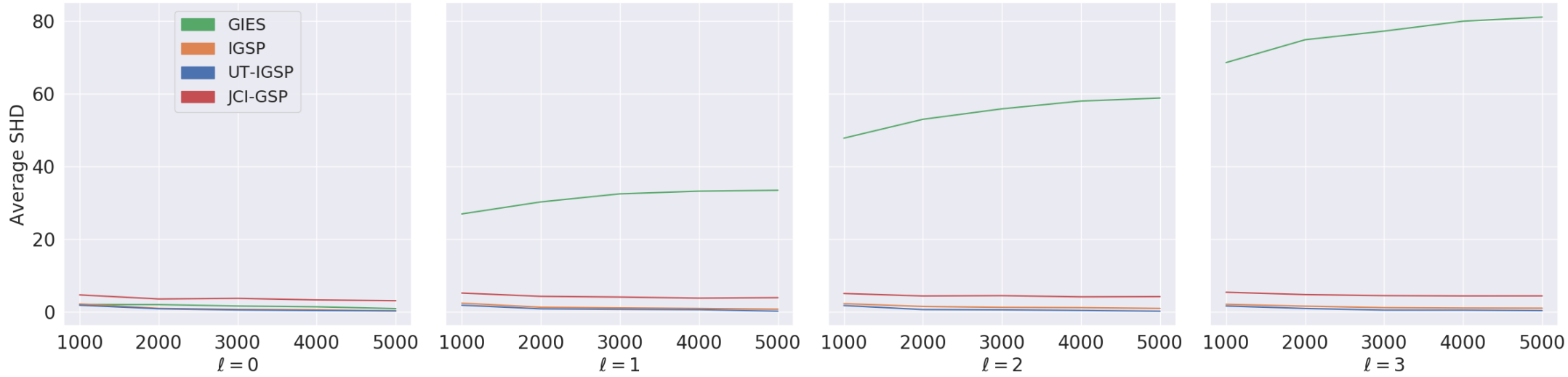
References (alphabetical)

- Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020). Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, *33*, 21865-21877.
- Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, *3*(Nov), 507-554.
- Eaton, D., & Murphy, K. (2007, March). Exact Bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics* (pp. 107-114). PMLR.
- Hauser, A., & Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, *13*(1), 2409-2464.
- Jaber, A., Kocaoglu, M., Shanmugam, K., & Bareinboim, E. (2020). Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, *33*, 9551-9561.
- Mooij, J. M., Magliacane, S., & Claassen, T. (2020). Joint causal inference from multiple contexts.
- Raskutti, G., & Uhler, C. (2018). Learning directed acyclic graph models based on sparsest permutations. *Stat*, *7*(1), e183.
- Squires, C., Wang, Y., & Uhler, C. (2020). Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence* (pp. 1039-1048). PMLR.
- Wang, Y., Solus, L., Yang, K., & Uhler, C. (2017). Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, *30*.
- Yang, K., Katcoff, A., & Uhler, C. (2018). Characterizing and learning equivalence classes of causal DAGs under interventions. In *International Conference on Machine Learning* (pp. 5541-5550). PMLR.

Appendix

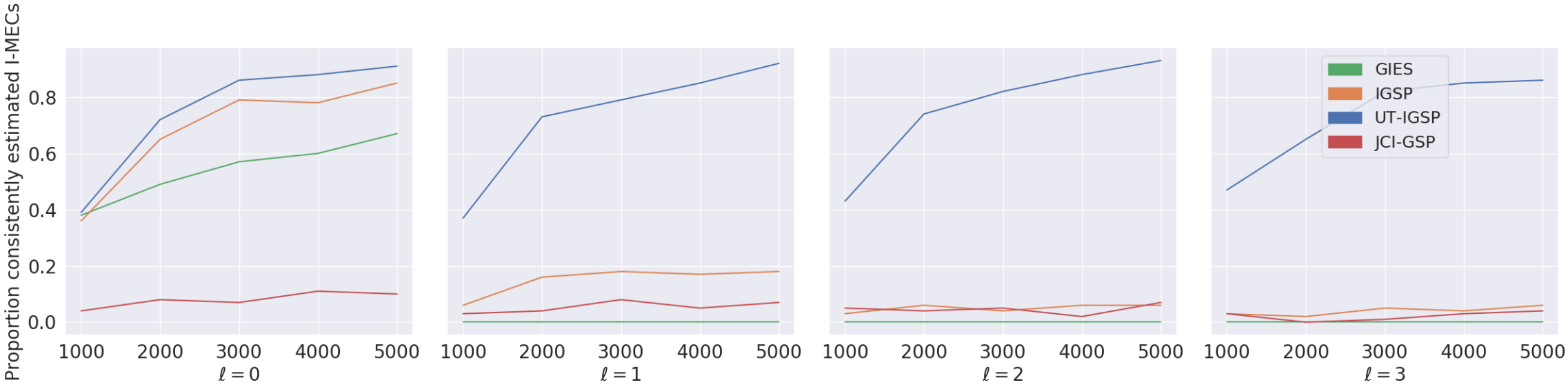
Empirical evaluation of UT-IGSP

- Data from *linear* structural causal models with Gaussian noise, with parent coefficients randomly sampled.
- $p = 20$ node DAGs sampled from an Erdös-Rényi distribution with expected neighborhood size 1.5.
- 5 *known* intervention targets and a varying number ℓ of *unknown* intervention targets.
- Interventions shift the mean of the intervened variable by 1.
- Hypothesis testing based on partial correlation for conditional independence and Chow tests for conditional invariances.
- Results averaged over 100 DAGs for each number of samples $n = 1000, 2000, \ldots, 5000$.



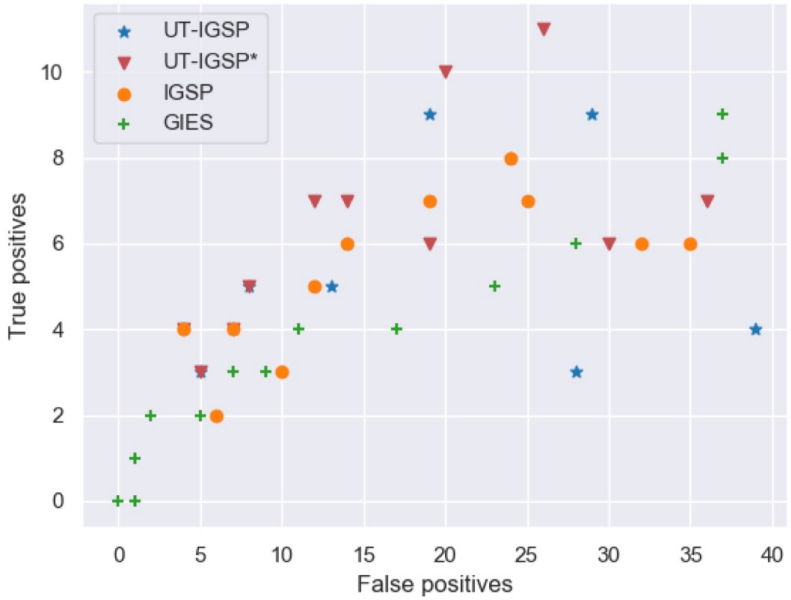
Empirical evaluation of UT-IGSP

- Data from *linear* structural causal models with Gaussian noise, with parent coefficients randomly sampled.
- $p = 20$ node DAGs sampled from an Erdös-Rényi distribution with expected neighborhood size 1.5.
- 5 *known* intervention targets and a varying number ℓ of *unknown* intervention targets.
- Interventions shift the mean of the intervened variable by 1.
- Hypothesis testing based on partial correlation for conditional independence and Chow tests for conditional invariances.
- Results averaged over 100 DAGs for each number of samples $n = 1000, 2000, \ldots, 5000$.

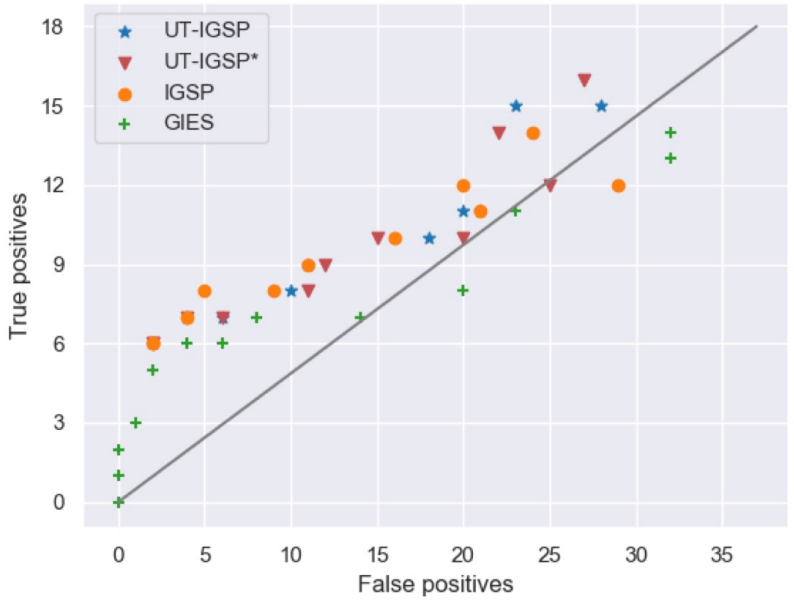


Empirical evaluation of UT-IGSP

- Sachs 2005 protein mass spectrometry data.
- 1,755 observational samples and 4,091 interventional samples.
- Compared to the conventionally accepted ground truth network.
- UT-IGSP*: all knowledge of intervention targets is removed.



(a) Directed edge recovery



(b) Skeleton recovery