Imperial College
London

23rd April 2024

Causal Discovery for Contestable and Explainable AI

From Contestable Neural Networks to Transparent Causal Discovery

Fabrizio Russo

Imperial College London

Outline

Introduction

Imperial College London

Introduction

- Measuring *Causal Effects* is necessary when one wants to evaluate the impact of actions e.g. the effect of a treatment on a patient's health

Imperial College London

Introduction

- Measuring *Causal Effects* is necessary when one wants to evaluate the impact of actions e.g. the effect of a treatment on a patient's health
- Randomized Control Trials (RCTs) provide *experimental data* which are the go-to for scientists to evaluate such effects

Imperial College London

Introduction

- Measuring *Causal Effects* is necessary when one wants to evaluate the impact of actions e.g. the effect of a treatment on a patient's health
- Randomized Control Trials (RCTs) provide *experimental data* which are the go-to for scientists to evaluate such effects
- RCTs are not always possible, ethical or economically feasible

Imperial College London

Introduction

- Measuring *Causal Effects* is necessary when one wants to evaluate the impact of actions e.g. the effect of a treatment on a patient's health
- Randomized Control Trials (RCTs) provide *experimental data* which are the go-to for scientists to evaluate such effects
- RCTs are not always possible, ethical or economically feasible

The need to extract as much causal information as possible from observational data

Imperial College London

Causal Effects Estimation

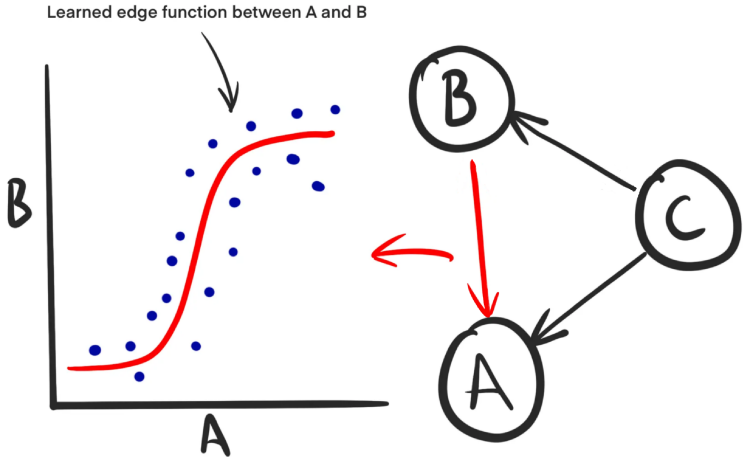


Image from https://towardsdatascience.com/how-to-understand-the-world-of-causality-c698cdc9f27c

Image from https://towardsdatascience.com/how-to-understand-the-world-of-causality-c698cdc9f27c

Causal Graphs for Contestable Neural Networks

Russo, Fabrizio & Francesca Toni. Causal Discovery and Knowledge Injection for Contestable Neural Networks. ECAI 2023: 2025-2032

Motivation

- Neural Networks are black boxes

Motivation

- Neural Networks are black boxes
- Contestability has been advocated by majour AI ethics frameworks and regulations (e.g. OECD, ACM and GDPR)

Motivation

- Neural Networks are black boxes
- Contestability has been advocated by majour AI ethics frameworks and regulations (e.g. OECD, ACM and GDPR)

Can we make neural networks both more transparent and human-aligned?

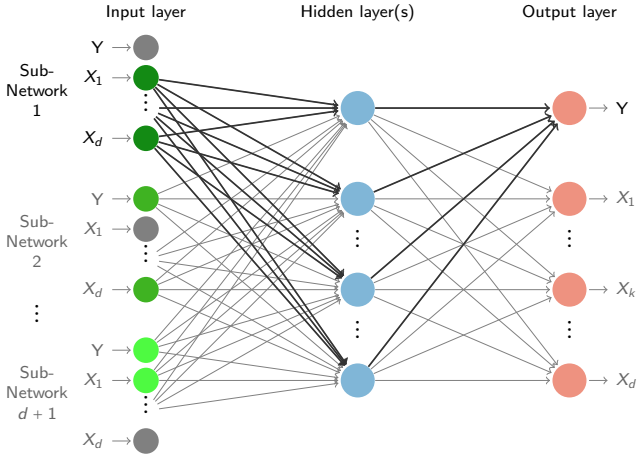
Contribution

Contestable Neural Networks

We propose *Knowledge Injection* empowered by *Causal Discovery* as a mean to make neural networks *Contestable* and improve them through experts' feedback.
This enables human-in-the-loop debugging and increased transparency.

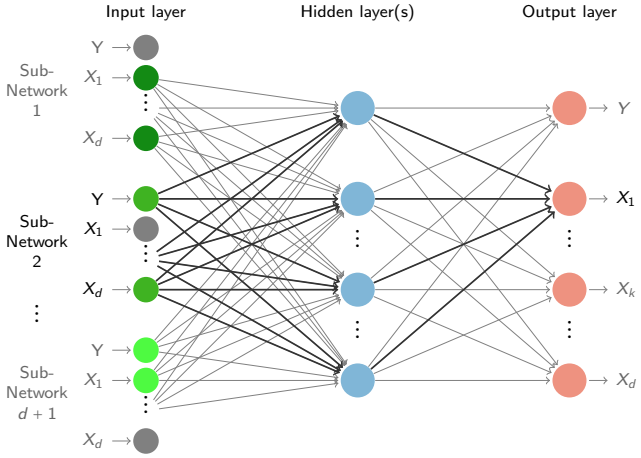
Imperial College London

Joint Network Structure (Kyono, Y. Zhang and Schaar 2020)



Imperial College London

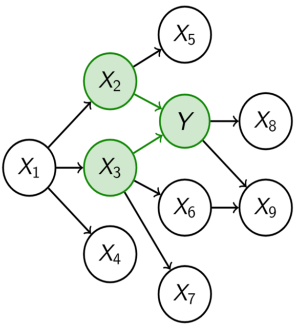
Joint Network Structure (Kyono, Y. Zhang and Schaar 2020)



Imperial College London

The Intuition

- **Objective:** have the network use only the relationships contained in the DAG i.e. predict each of the features using only its parents.

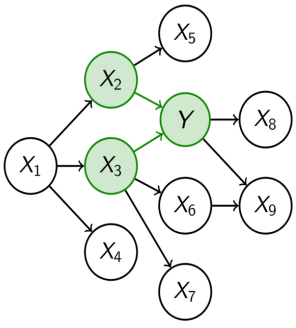


	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.005	0.017	0.008	0.002	0.042	0.02	0.005	0.059	0.05
X_1	0.006	0.0	0.063	0.054	0.068	0.009	0.006	0.013	0.006	0.008
X_2	0.088	0.036	0.0	0.022	0.019	0.124	0.008	0.011	0.006	0.008
X_3	0.087	0.034	0.021	0.0	0.024	0.005	0.107	0.104	0.006	0.009
X_4	0.009	0.032	0.02	0.023	0.0	0.01	0.013	0.01	0.005	0.005
X_5	0.026	0.006	0.017	0.004	0.004	0.0	0.012	0.002	0.005	0.018
X_6	0.025	0.006	0.008	0.011	0.005	0.017	0.0	0.014	0.002	0.114
X_7	0.029	0.003	0.007	0.011	0.002	0.024	0.029	0.0	0.011	0.01
X_8	0.036	0.002	0.004	0.003	0.004	0.006	0.009	0.006	0.0	0.006
X_9	0.024	0.003	0.003	0.004	0.003	0.005	0.079	0.01	0.004	0.0

Imperial College London

The Intuition

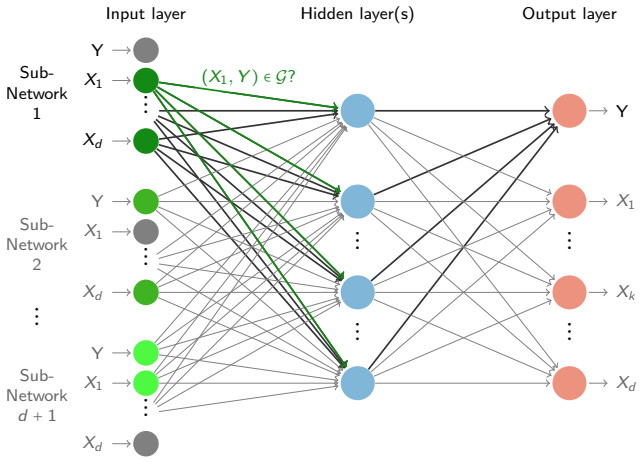
- **Objective:** have the network use only the relationships contained in the DAG i.e. predict each of the features using only its parents.



	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.059	0.05
X_1	0.0	0.0	0.063	0.054	0.068	0.0	0.0	0.0	0.0	0.0
X_2	0.088	0.0	0.0	0.0	0.0	0.124	0.0	0.0	0.0	0.0
X_3	0.087	0.0	0.0	0.0	0.0	0.0	0.107	0.104	0.0	0.0
X_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.114
X_7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X_9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

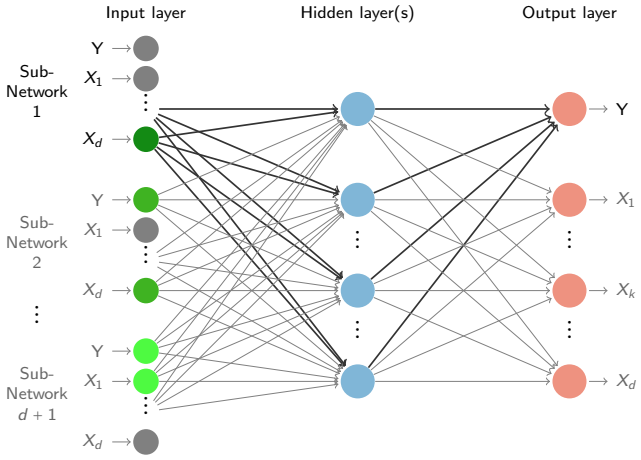
Imperial College London

Inject Causal Knowledge



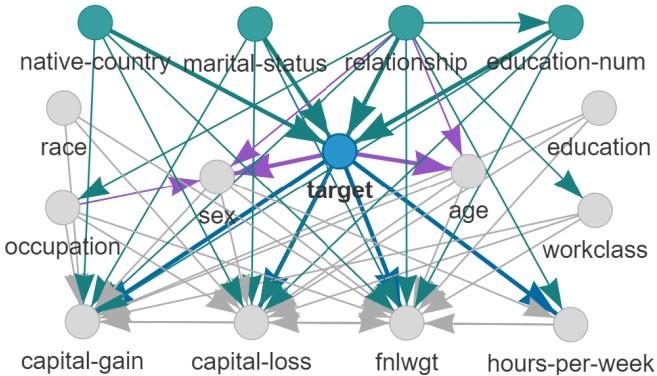
Imperial College London

Inject Causal Knowledge



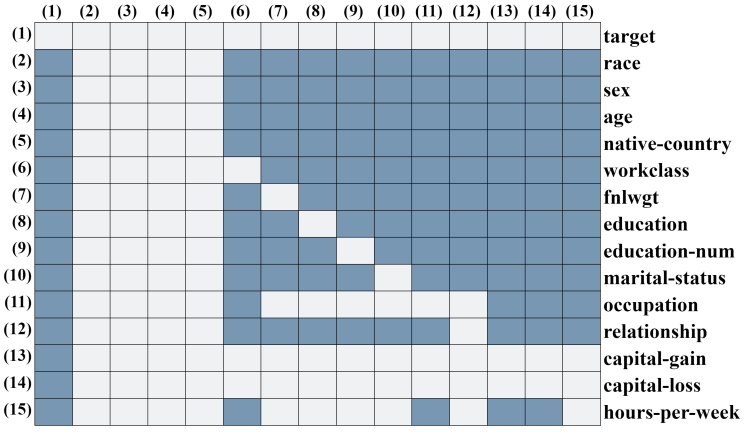
Imperial College London

Case Study - Predict Income



Imperial College London

Common Sense Constraints



Imperial College London

Contested Network - Results

Table 1: Average AUC over 25 runs for the Adult dataset in four experimental scenarios.

	Adult Dataset ($\|V\| = 14$)			
Data (N)	**CASTLE** $\|E\| = 210$	**Injected** $\|E\| = 46$	**Partial** $\|E\| = 116$	**Contested** $\|E\| = 30$
100	0.67 (0.03)	**0.69 .**	0.66	**0.69 .**
500	0.72 (0.04)	**0.74 ***	0.71	**0.74 ***
1000	0.75 (0.03)	**0.76**	0.74	**0.76**
2000	0.74 (0.03)	**0.77 *****	**0.76 ***	**0.77 *****
5000	0.75 (0.03)	**0.79 *****	**0.76**	**0.79 *****
10000	0.75 (0.02)	**0.85 *****	**0.76 .**	**0.85 *****
20000	0.76 (0.02)	**0.86 *****	**0.77 .**	**0.86 *****

Significance levels against CASTLE (Kyono, Y. Zhang and Schaar 2020): 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Summary

- Contestability has been advocated by majour AI ethics frameworks and regulations (e.g. OECD, ACM and GDPR)

Summary

- Contestability has been advocated by majour AI ethics frameworks and regulations (e.g. OECD, ACM and GDPR)
- We propose a framework to empower experts to contest neural networks predictions

Summary

- Contestability has been advocated by majour AI ethics frameworks and regulations (e.g. OECD, ACM and GDPR)
- We propose a framework to empower experts to contest neural networks predictions
- We demonstrate how it can help both accuracy and transparency

Limitations

- Contestable Neural Networks are not a tool for Causal Inference

Limitations

- Contestable Neural Networks are not a tool for Causal Inference
- Causal Graphs make Networks more transparent, what about transparency and assurance around the Causal Graphs themselves?

Causal Discovery with Shapley Values

Russo, Fabrizio & Francesca Toni. Shapley-PC: Constraint-based Causal Structure Learning with Shapley Values. CoRR abs/2312.11582 (2023) Preprint Under Review

Imperial College London

Causal Discovery

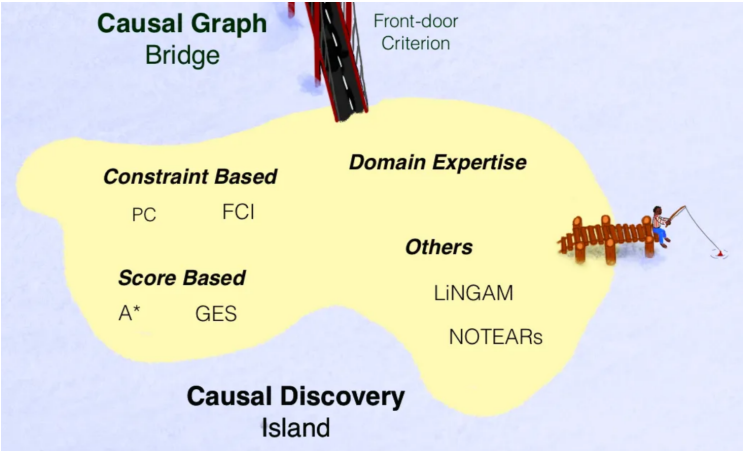
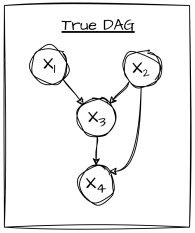


Image from https://towardsdatascience.com/how-to-understand-the-world-of-causality-c698cdc9f27c

PC-Algorithm (Spirtes, C. N. Glymour and Scheines 2000)

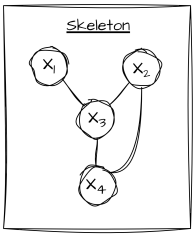


Input: Conditional Independence Information for all variables

1: Build skeleton $\mathcal{C}$ via adjacency search (save separating sets)
2: Orient v-structures in $\mathcal{C}$ (using separating sets from step 1)
3: Propagate *d-separation* via Meek rules

Return: CPDAG

PC-Algorithm (Spirtes, C. N. Glymour and Scheines 2000)



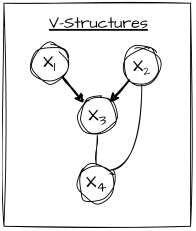
Input: Conditional Independence Information for all variables

1: **Build skeleton $\mathcal{C}$ via adjacency search (save separating sets)**
2: Orient v-structures in $\mathcal{C}$ (using separating sets from step 1)
3: Propagate *d-separation* via Meek rules

Return: CPDAG

Imperial College London

PC-Algorithm (Spirtes, C. N. Glymour and Scheines 2000)

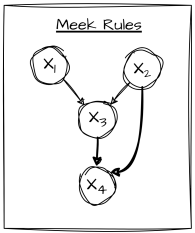


Input: Conditional Independence Information for all variables

1: Build skeleton $\mathcal{C}$ via adjacency search (save separating sets)
2: **Orient v-structures in $\mathcal{C}$ (using separating sets from step 1)**
3: Propagate *d-separation* via Meek rules

Return: CPDAG

PC-Algorithm (Spirtes, C. N. Glymour and Scheines 2000)

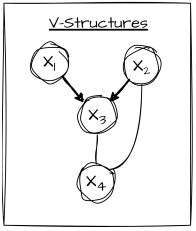


Input: Conditional Independence Information for all variables

1: Build skeleton $\mathcal{C}$ via adjacency search (save separating sets)
2: Orient v-structures in $\mathcal{C}$ (using separating sets from step 1)
3: **Propagate d-separation via Meek rules**

Return: CPDAG

PC-Algorithm (Spirtes, C. N. Glymour and Scheines 2000)



Input: Conditional Independence Information for all variables

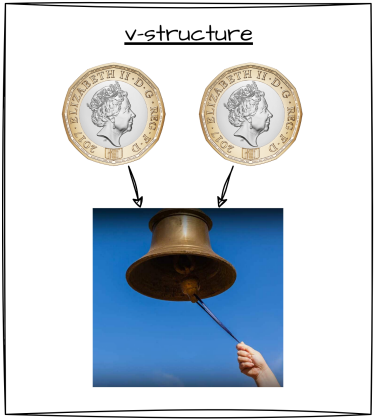
1: Build skeleton $\mathcal{C}$ via adjacency search (save separating sets)
2: **Orient v-structures in $\mathcal{C}$ (using separating sets from step 1)**
3: Propagate *d-separation* via Meek rules

Return: CPDAG

Imperial College London

V-structures or Immoralities

- Two variables are marginally independent $Coin1 \perp\!\!\!\perp Coin2$
- But they are dependent when conditioned on a common descendant $Coin1 \not\perp\!\!\!\perp Coin2 \mid Bell$



PC-based Algorithms

i. Conservative-PC (Ramsey, Spirtes and J. Zhang 2006)
   → Orient v-structure only if the candidate collider renders dependence in **all** tests when in the conditioning set
ii. Majority-PC (Colombo and Maathuis 2014)
   → Orient v-structure only if the candidate collider renders dependence in the **majority** of the tests when in the conditioning set
iii. Max-PC (Ramsey 2016)
   → Orient v-structure only if the test with the **maximum p-value** does not contain the candidate collider
iv. ML4C (Dai et al. 2023)
   → Build a machine learning model to **predict** v-structures
v. Shapley-PC (Russo and Toni 2023)
   → Orient v-structure only if the candidate collider has the **lowest Shapley Independence Value**

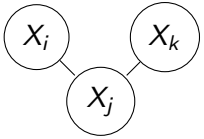
Shapley Values

- Shapley value in cooperative game theory (Shapley 1953)
- Suppose a **team** $N = 1, \ldots, n$ of players cooperates to earn **value** $v(N)$:

$$\phi_v(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \tag{1}$$

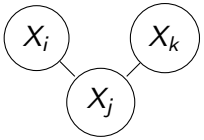
- Represents a **player** i's marginal value-added upon joining a team

Shapley Independence Values (SIV)



- Suppose we are assessing whether $X_i - X_j - X_k$ in $\mathcal{C}$ is a v-structure

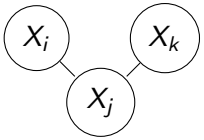
Shapley Independence Values (SIV)



- Suppose we are assessing whether $X_i - X_j - X_k$ in $\mathcal{C}$ is a v-structure
- For a given skeleton $\mathcal{C}$ our **team** of variablesis:

$$\mathbf{N} = \{\mathbf{S} : \mathbf{S} \subseteq \text{adj}(\mathcal{C}, X_i) \setminus \{X_j\} \vee \mathbf{S} \subseteq \text{adj}(\mathcal{C}, X_k) \setminus \{X_j\}\} \tag{2}$$

Shapley Independence Values (SIV)



- Suppose we are assessing whether $X_i - X_j - X_k$ in $\mathcal{C}$ is a v-structure
- For a given skeleton $\mathcal{C}$ our **team** of variablesis:

$$\mathbf{N} = \{\mathbf{S} : \mathbf{S} \subseteq \text{adj}(\mathcal{C}, X_i) \setminus \{X_j\} \vee \mathbf{S} \subseteq \text{adj}(\mathcal{C}, X_k) \setminus \{X_j\}\} \tag{2}$$

- Our **value** function is the p-value returned by the conditional independence test $I(X_i, X_k \mid \mathbf{S})$

Shapley Independence Values (SIV)

Shapley Independence Value (SIV)

$$\phi_I(X_j, \{X_i, X_k\}) = \sum_{\mathbf{S} \in \mathbf{N}} \frac{|S|!(n - |S| - 1)!}{n!} [I(X_i, X_k \mid \mathbf{S} \cup \{X_j\}) - I(X_i, X_k \mid \mathbf{S})] \quad (3)$$

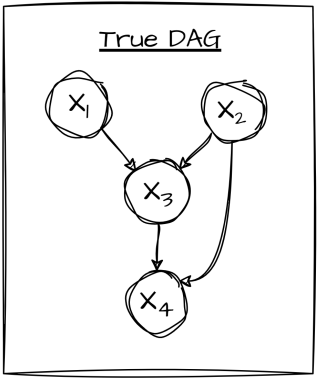
Shapley Independence Values (SIV)

Shapley Independence Value (SIV)

$$\phi_I(X_j, \{X_i, X_k\}) = \sum_{\mathbf{S} \in \mathbf{N}} \frac{|S|!(n - |S| - 1)!}{n!} [I(X_i, X_k \mid \mathbf{S} \cup \{X_j\}) - I(X_i, X_k \mid \mathbf{S})] \quad (3)$$

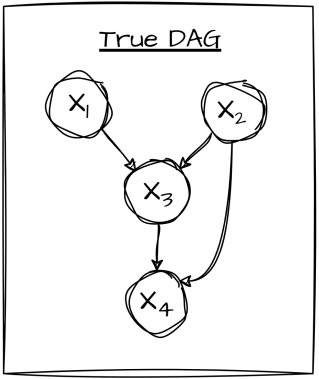
The higher $\phi_I(X_j, \{X_i, X_k\})$ the higher is the X_j's contribution the independence between X_i and X_k.

Example I (Perfect Independence Information)



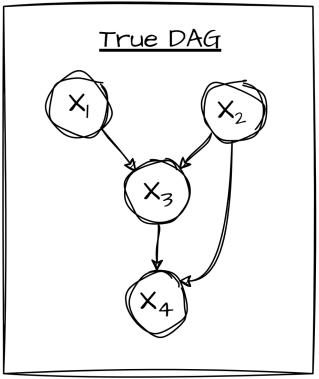
- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Perfect Conditional Independence Information
  1. $I(X_1, X_2) = 1 \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$

Example I (Perfect Independence Information)



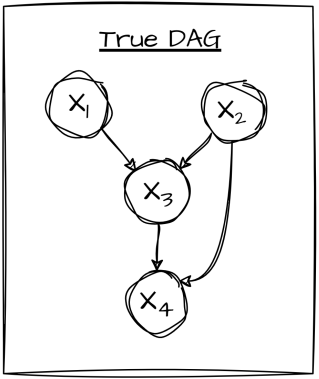
- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Perfect Conditional Independence Information
  1. $I(X_1, X_2) = 1 \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$
- $\phi_I(X_3, \{X_1, X_2\}) = \phi_I(X_4, \{X_1, X_2\}) = -0.5$

Example I (Perfect Independence Information)



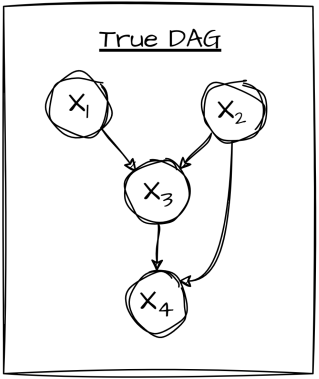
- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Perfect Conditional Independence Information
  1. $I(X_1, X_2) = 1 \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$
- $\phi_I(X_3, \{X_1, X_2\}) = \phi_I(X_4, \{X_1, X_2\}) = -0.5$
- All PC-based algorithms can recover the true DAG

Example II (Imperfect Independence Information)



- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Finite Sample Conditional Independence Tests
  1. $I(X_1, X_2) = 0.7 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0.01 < \alpha \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0.1 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0.75 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$

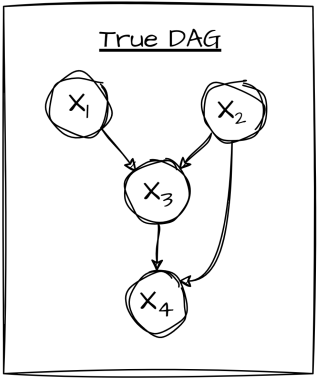
Example II (Imperfect Independence Information)



- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Finite Sample Conditional Independence Tests
  1. $I(X_1, X_2) = 0.7 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0.01 < \alpha \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0.1 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0.75 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$
- $\phi_I(X_3, \{X_1, X_2\}) = -0.03, \phi_I(X_4, \{X_1, X_2\}) = 0.08$

Imperial College London

Example II (Imperfect Independence Information)



- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Finite Sample Conditional Independence Tests
  1. $I(X_1, X_2) = 0.7 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0.01 < \alpha \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0.1 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0.75 \geq \alpha \rightarrow X_1 \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$
- $\phi_I(X_3, \{X_1, X_2\}) = -0.03, \phi_I(X_4, \{X_1, X_2\}) = 0.08$
- Only Shapley-PC correctly orients the v-structure

Advantages of using SIV

- More robust to wrong independence tests by looking at relations among tests with different conditioning sets
- Remove the dependency on the usual significance threshold α
- Keep the theoretical soundness and asymptotic consistency of PC
- No computational overhead (compared to modifications of PC)

Advantages of using SIV

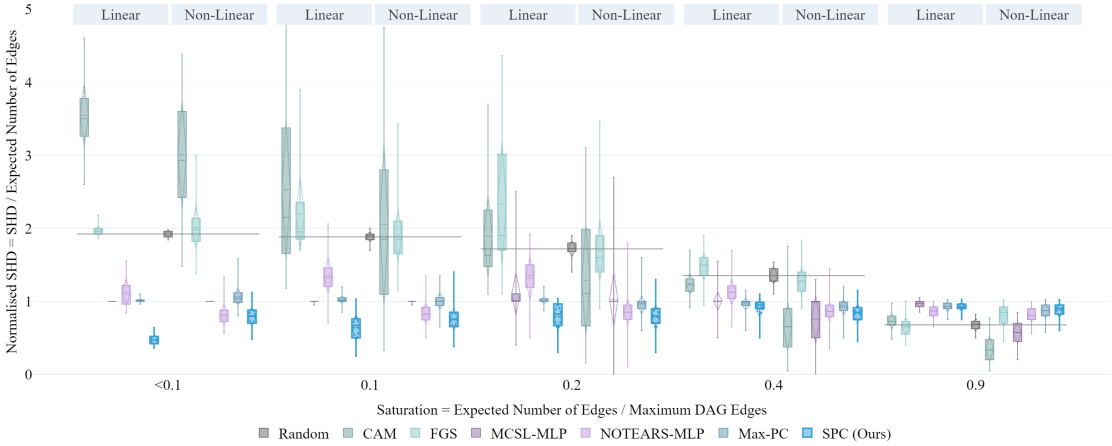
- More robust to wrong independence tests by looking at relations among tests with different conditioning sets
- Remove the dependency on the usual significance threshold α
- Keep the theoretical soundness and asymptotic consistency of PC
- No computational overhead (compared to modifications of PC)
- Better empirical results

Synthetic Data - DGPs

- We analyse all the different DGPs in (Zheng, Aragam et al. 2018; Zheng, Dan et al. 2020) for Erdös-Rényi (ER) graphs.
  - 10 random graphs $\mathcal{G}_i = (\mathbf{V}, E)$
  - Number of nodes $|\mathbf{V}| \in \{10, 20, 50\}$
  - Number of edges $|E| = |\mathbf{V}| \times d$ with $d = \{1, 2, 4\}$
- Given the ground truth DAGs $\mathcal{G}_i$, we simulate Structural Equation Models (SEMs) belonging to the Additive Noise Model ($X_j = f_j(\text{pa}(\mathcal{G}, X_j)) + u_j$)
  - 4 linear SEMs (Gaussian, Exponential, Gumbel, Uniform noise) and 4 non-linear SEMs (GPs, Additive GPs, Mixed Models and MLPs) with Gaussian noise
  - Number of samples $N = s \times |\mathbf{V}|$, $s \in \{10, 50, 100, 500\}$

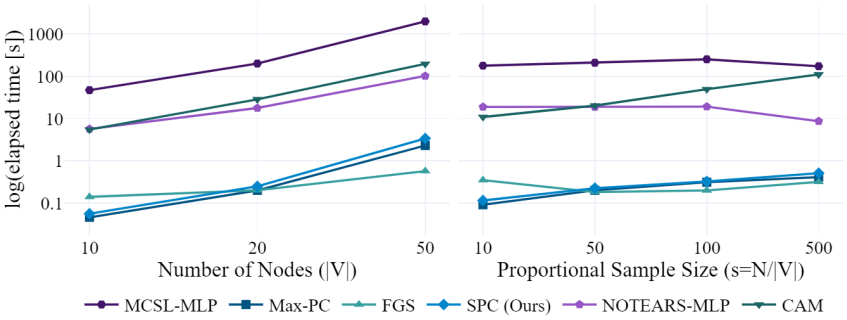
Imperial College London

Synthetic Data - Structural Hamming Distance (↓)



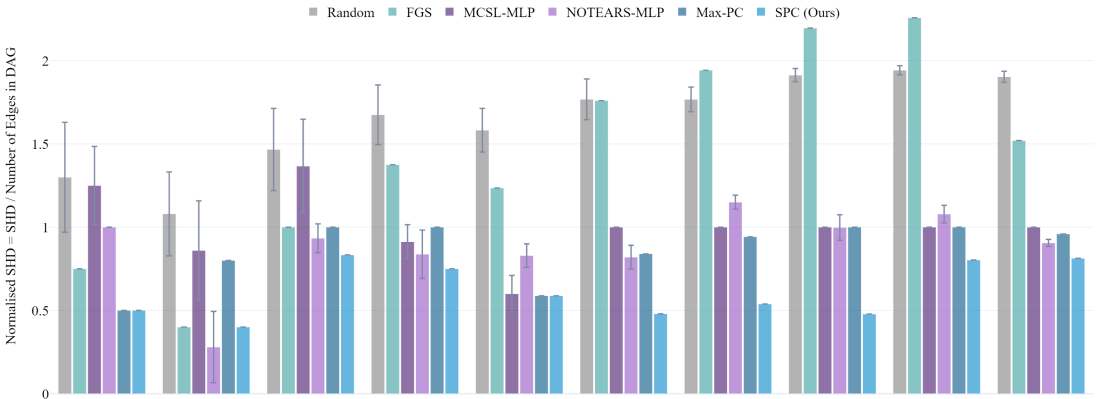
Imperial College London

Synthetic Data - Runtime



Imperial College London

bnlearn data



Summary

- We propose Shapley-PC, a novel constraint-based causal discovery algorithm
- Improves upon the robustness against errors in independence tests in a small sample setting improving on SOTA baselines
- Maintains the soundness and large sample consistency of its predecessors

Conclusion & Future Work

Conclusions & Future Work

- We discussed two novel methods to
  - Employ Causal Graphs as means to contest Neural Networks
  - Discovery Causal Graphs more robustly

Conclusions & Future Work

- We discussed two novel methods to
  - Employ Causal Graphs as means to contest Neural Networks
  - Discovery Causal Graphs more robustly
- We are working on integrating these ideas withing and Argumentation Framework that allows both discovery and explainability of Causal Graphs

Thanks for your attention!
Questions?

Imperial College London

References I

Bühlmann, Peter, Jonas Peters and Jan Ernest (2014). 'CAM: Causal additive models, high-dimensional order search and penalized regression'. In: *The Annals of Statistics* 42.6, pp. 2526–2556. DOI: 10.1214/14-AOS1260. URL: https://doi.org/10.1214/14-AOS1260.

Chickering, David Maxwell (Mar. 2002). 'Learning equivalence classes of bayesian-network structures'. In: *Journal of Machine Learning Research* 2, pp. 445–498. ISSN: 1532-4435. DOI: 10.1162/153244302760200696. URL: https://doi.org/10.1162/153244302760200696.

Colombo, Diego and Marloes H. Maathuis (Jan. 2014). 'Order-Independent Constraint-Based Causal Structure Learning'. In: *Journal of Machine Learning Research* 15.1, pp. 3741–3782. ISSN: 1532–4435.

Dai, Haoyue et al. (2023). 'Ml4c: Seeing causality through latent vicinity'. In: *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, pp. 226–234.

Harris, Naftali and Mathias Drton (Jan. 2013). 'PC algorithm for nonparanormal graphical models'. In: *Journal of Machine Learning Research* 14.1, pp. 3365–3383. ISSN: 1532-4435.

Hung, H. M. James et al. (1997). 'The Behavior of the P-Value When the Alternative Hypothesis is True'. In: *Biometrics* 53.1, pp. 11–22. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2533093 (visited on 09/10/2023).

Kalisch, Markus and Peter Bühlman (2007). 'Estimating high-dimensional directed acyclic graphs with the PC-algorithm.'. In: *Journal of Machine Learning Research* 8.22, pp. 613–636.

Imperial College London

References II

Kyono, Trent, Yao Zhang and Mihaela van der Schaar (2020). 'CASTLE: Regularization via Auxiliary Causal Graph Discovery'. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. URL: https://proceedings.neurips.cc/paper/2020/hash/1068bceb19323fe72b2b344ccf85c254-Abstract.html.

Meek, Christopher (1995). 'Causal Inference and Causal Explanation with Background Knowledge'. In: *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. UAI 1995. Montréal, Qué, Canada: Morgan Kaufmann Publishers Inc., pp. 403–410. ISBN: 1558603859.

Ng, Ignavier et al. (2022). 'Masked gradient-based causal structure learning'. In: *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, pp. 424–432.

Peters, Jonas and Peter Bühlmann (2015). 'Structural intervention distance for evaluating causal graphs'. In: *Neural computation* 27.3, pp. 771–799.

Ramsey, Joseph (2016). *Improving accuracy and scalability of the pc algorithm by maximizing p-value*.

Ramsey, Joseph, Madelyn Glymour et al. (2017). 'A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images'. In: *International journal of data science and analytics* 3, pp. 121–129.

Imperial College London

References III

Ramsey, Joseph, Peter Spirtes and Jiji Zhang (2006). 'Adjacency-Faithfulness and Conservative Causal Inference'. In: *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. UAI 2006. Cambridge, MA, USA: AUAI Press, pp. 401–408. ISBN: 0974903922.

Russo, Fabrizio and Francesca Toni (2023). 'Shapley-PC: Constraint-based Causal Structure Learning with Shapley Values'. In: *CoRR* abs/2312.11582. DOI: 10.48550/ARXIV.2312.11582. arXiv: 2312.11582. URL: https://doi.org/10.48550/arXiv.2312.11582.

Shapley, Lloyd S (1953). 'A value for n-person games (1953)'. In: *Contribution to the Theory of Games*.

Spirtes, Peter, Clark N Glymour and Richard Scheines (2000). *Causation, prediction, and search*. MIT press.

Zheng, Xun, Bryon Aragam et al. (2018). 'DAGs with NO TEARS: Continuous Optimization for Structure Learning'. In: *Prooceedings of the 31st Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9492–9503. URL: https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html.

Zheng, Xun, Chen Dan et al. (Nov. 2020). 'Learning Sparse Nonparametric DAGs'. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 3414–3425. URL: https://proceedings.mlr.press/v108/zheng20a.html.

Appendices

Imperial College London

Shapley-PC (Step 2)

Input: Skeleton $\mathcal{C}$

$SV \leftarrow \varnothing$
for $X_i - X_j - X_k \in \mathcal{C}$ **do**
  for $\mathbf{S} \in \mathbf{N}$ **do** ▷All subsets of adjacent nodes (Eq. 2)
    for $X_c \in \mathbf{S}$ **do**
      $SV \leftarrow SV \cup \{\phi_I(X_c, \{X_i, X_k\})\}$
  $\phi_I^* = min(SV)$ ▷Least contribution to $I(X_i, X_j \mid \mathbf{N})$
  if $\phi_I^* = \phi_I(X_j, \{X_i, X_k\})$ **then**
    if $X_i - X_j - X_k$ not fully directed **then**
      if do not add a cycle **then**
        orient: $X_i \rightarrow X_j \leftarrow X_k$

Return: CPDAG

Imperial College
London

PC-Algorithm (Propagation, Meek 1995)

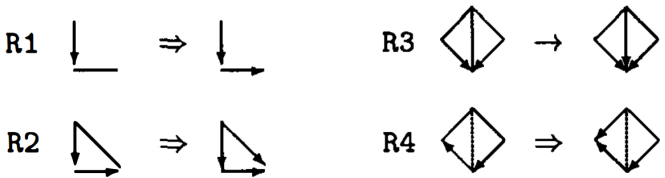
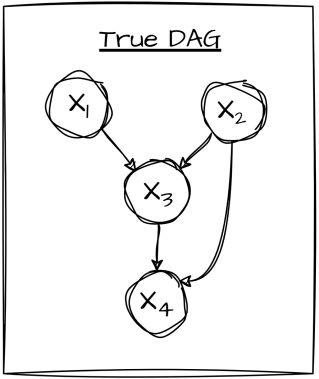


Figure 1: Orientation rules for patterns

Imperial College London

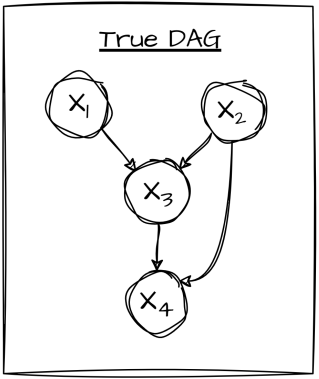
Example III (Little imperfect Independence Information)



- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Finite Sample Conditional Independence Tests
  1. $I(X_1, X_2) = 1 \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0.1 \rightarrow X_1 \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$

Imperial College London

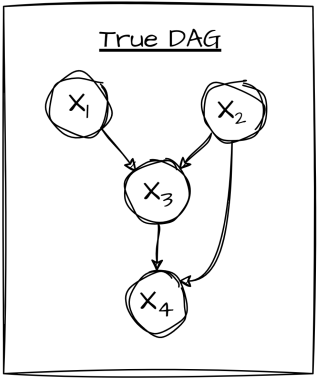
Example III (Little imperfect Independence Information)



- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Finite Sample Conditional Independence Tests
  1. $I(X_1, X_2) = 1 \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0.1 \rightarrow X_1 \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$
- $\phi_I(X_3, \{X_1, X_2\}) = \phi_I(X_4, \{X_1, X_2\}) = -0.45$

Imperial College London

Example III (Little imperfect Independence Information)



- Assess whether $X_1 - X_3 - X_2$ is a v-structure
- Finite Sample Conditional Independence Tests
  1. $I(X_1, X_2) = 1 \rightarrow X_1 \perp\!\!\!\perp X_2$
  2. $I(X_1, X_2 \mid X_3) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_3$
  3. $I(X_1, X_2 \mid X_4) = 0 \rightarrow X_1 \not\perp\!\!\!\perp X_2 \mid X_4$
  4. $I(X_1, X_2 \mid \{X_3, X_4\}) = 0.1 \rightarrow X_1 \perp\!\!\!\perp X_2 \mid \{X_3, X_4\}$
- $\phi_I(X_3, \{X_1, X_2\}) = \phi_I(X_4, \{X_1, X_2\}) = -0.45$
- Only Shapley-PC and PC-max correctly orient the v-structure

Imperial College London

Soundness

Definition (Perfect Conditional Independence Test (CIT))

$$I(X_i, X_j \mid \mathbf{S}) = \begin{cases} 1 \text{ iff } X_i \perp\!\!\!\perp X_j \mid \mathbf{S} \\ 0 \text{ otherwise} \end{cases}$$

Lemma (Negative SIV)

Given a skeleton $\mathcal{C}$ and an unshielded triple $(X_i, X_j, X_k) \in \mathcal{C}$, With a perfect CIT I, $\phi_I(X_j, \{X_i, X_k\}) < 0$ if X_j is a collider or a descendant thereof and $\phi_I(X_j, \{X_i, X_k\}) > 0$ otherwise.

Theorem (Soundness)

Let P be faithful to a DAG $\mathcal{G} = (\mathbf{V}, E)$, and assume that we have a perfect CIT. Then the output of the SPC-algorithm is the CPDAG that represents $\mathcal{G}$.

Asymptotic Consistency

- In the sample limit, the original PC-algorithm has been shown to be consistent for sparse graphs and multivariate Gaussian distributions (Kalisch and Bühlman 2007) or Gaussian copulas (Harris and Drton 2013)
- The results are contingent on PC only performing CIT between pairs of variables, with the size of the conditioning sets **S** less or equal to the degree of the graph
- Our proposed method has the same characteristic, hence the consistency results are equally applicable

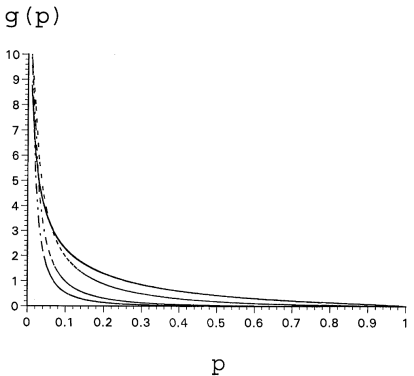
Complexity

- As in PC, the complexity of our algorithm depends on the number of vertices and their maximal degree (Spirtes, C. N. Glymour and Scheines 2000, p.85)
- As in CPC, MPC and PC-Max, we perform additional tests compared to the original PC that used the separating sets derived from the adjacency test in Step 1
- The number of tests though still depends on the degree as we only add the tests about the adjacency set to calculate $\phi_I(X_j)$
- The majority of the testing is still done in the adjacency search of Step 1 of PC (Ramsey, Spirtes and J. Zhang 2006)

Imperial College London

Why the minimum? (Hung et al. 1997)

- Under the alternative hypothesis of dependence
  - $g(p)$ depends on the sample size and the value of the parameter in the alternative hypothesis
  - $g(p)$ decrease monotonically and concentrate around 0, the further towards 1 the lower is the probability of dependence



Evaluation Metrics

- Structural Hamming Distance (SHD) is a purely graphical metric summing up the number of changes to be made to the estimated graph to match the true one (SHD = Extra + Missing + Reversed)

- Structural Intervention Distance (SID) (Peters and Bühlmann 2015) quantifies the closeness between two DAGs in terms of their corresponding causal inference statements

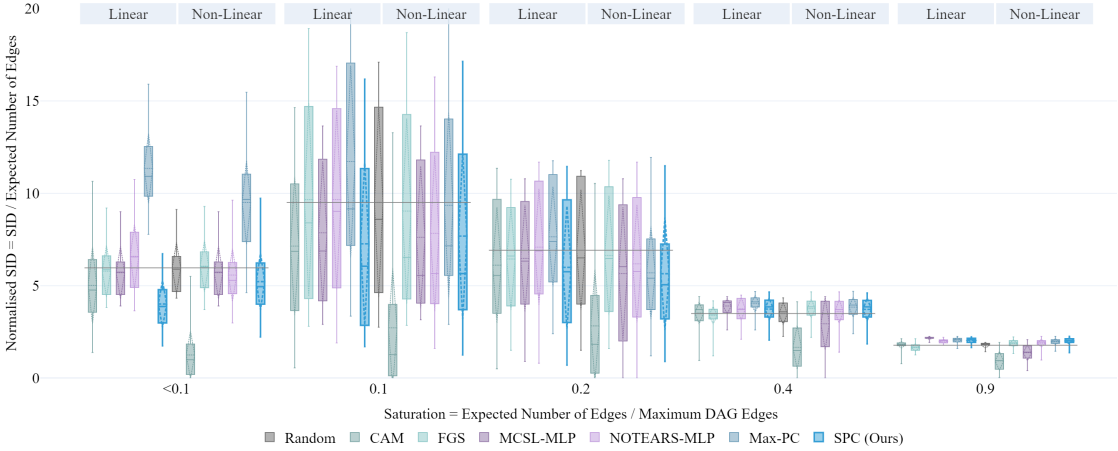
- Saturation is the number of edges in the DAG compared to the maximum number of edges that a DAG with $\mathbf{V}$ nodes can have to remain acyclic

Baselines

1. PC-Max (Ramsey 2016) is a modification of the PC-algorithm
2. Fast Greedy Equivalence Search (FGS) (Ramsey, M. Glymour et al. 2017) is a fast implementation of GES (Chickering 2002) where graphs are evaluated using the Bayesian Information Criterion
3. Causal Additive Model (CAM) (Bühlmann, Peters and Ernest 2014) learns an additive SEM by decoupling feature selection and causal order estimation
4. NOTEARS-MLP (Zheng, Dan et al. 2020) learns a non-linear SEM via continuous optimisation
5. Masked-CSL-MLP(Ng et al. 2022) learns a non-linear SEM via continuous optimisation in the same fashion as Zheng, Dan et al. 2020 but applying the gumbel-softmax trick

Imperial College London

Synthetic Data - SID



Imperial College London

bnlearn data - SID