# Causal Discovery in the Presence of Missing Data

**Ruibo Tu**[*1], **Cheng Zhang**[*2], **Paul Ackermann**[3], **Karthika Mohan**[4], **Hedvig Kjellström**[1], **and Kun Zhang**[*5]

[1]KTH Royal Institute of Technology, [2]Microsoft Research, Cambridge, [3]Karolinska Institute, [4]University of California, Berkeley, [5]Carnegie Mellon University

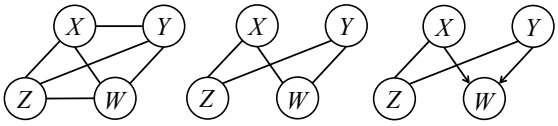
## INTRODUCTION

- **Missing Data Problem:** Analyze the influence of the missing data on the constraint-based causal discovery algorithms —— *Propositions*
- **Solution:** Propose a framework *MVPC* and *two correction methods* for correcting the results of the deletion-based PC algorithm.
- What is the relationship between IQ and gender?

| *IQ* | *Gender* | *College* |
|---|---|---|
| high | female | in |
| medium | female | not in |
| low | female | not in |
| high | male | in |
| medium | male | in |
| low | male | in |

## DELETION-BASED PC

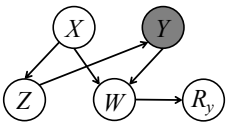
- List-wise deletion and **Test-wise deletion**
- Deletion might produce errors in the **skeleton search** step. **Observed** data distribution vs **Full** data distribution
- PC Recap



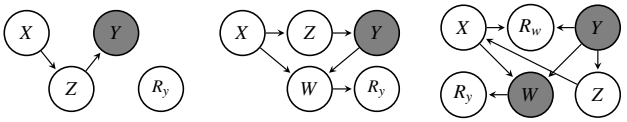

## MISSINGNESS GRAPH

- Missingness Graphs: missingness indicators and proxy variables

| $X$ | $Y$ | $Z$ | $W$ | $R_y$ |
|---|---|---|---|---|
| 1 | 2.2 | 2.1 | 3.1 | 0 |
| 3 | 6.1 | 6.3 | 9.3 | 0 |
| 2 | 4.1 | 4.2 | 6.1 | 0 |
| 7 | *M* | 14.1 | 21.4 | 1 |
| 4 | *M* | 8.1 | 12.3 | 1 |




- MCAR, MAR, MNAR:




(a) A MCAR graph $\{\mathbf{V}_m, \mathbf{V}_o\} \perp\!\!\!\perp_d \mathbf{R}$

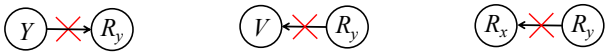
(b) A MAR graph $\mathbf{V}_m \perp\!\!\!\perp_d \mathbf{R} \mid \mathbf{V}_o$

(c) A MNAR graph

- Self-Masking Missingness


## ASSUMPTIONS

- Recoverability of joint distributions




- Faithful observability

$$X \perp\!\!\!\perp Y \mid \{\mathbf{Z}, \mathbf{R_K} = \mathbf{0}\} \iff X \perp\!\!\!\perp Y \mid \{\mathbf{Z}, \mathbf{R_K} = \mathbf{1}\}$$

## PROPOSITIONS

- **Proposition 1:** The erroneous edges are extraneous edges

$$X \perp\!\!\!\perp Y \mid \{\mathbf{Z}, R_x = 0, R_y = 0, \mathbf{R_z} = \mathbf{0}\} \Longrightarrow X \perp\!\!\!\perp Y \mid \mathbf{Z}$$
$$X \not\perp\!\!\!\perp Y \mid \{\mathbf{Z}, R_x = 0, R_y = 0, \mathbf{R_z} = \mathbf{0}\} \not\Longrightarrow X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$$

- **Proposition 2:** Determine circumstances where the extraneous edges happen. *If there is an extraneous edge between X and Y, under the four assumptions there is at least a missingness indicator of X, Y, and Z is either the direct common effect or a descendant of the direct common effect of X and Y. ( Z: any variable set which satisfies that given Z, X is independent of Y )*


## MVPC

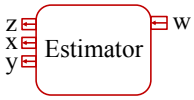
- Skeleton Search (PC)
  1. Detecting direct causes of *missingness indicators*
  2. Detecting potential extraneous edges
  3. Recovering the true causal graph skeleton
- Determining the orientation (PC)

### Recovering the true causal graph skeleton

- Get access to the data from the joint distribution *P(X, Y, Z)*

**(A)** Permutation-based correction:

$$P(X,Y,Z) = \int_W P(X,Y,Z \mid W)P(W)dW$$
$$= \int_W P(X,Y^*,Z \mid W, R_y = 0)P(W)dW$$

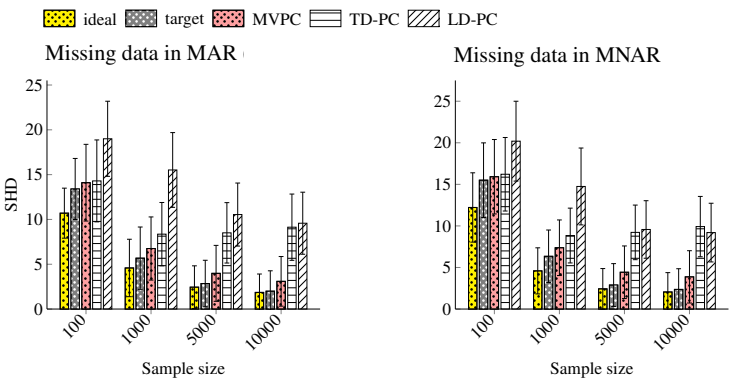



**(B)** Density ratio weighted corection

$$P(X,Y,Z) = \frac{P(X,Y,Z \mid R_z = 0)P(R_z = 0)}{P(R_z = 0 \mid X,Y)}$$
$$= P(X,Y,Z \mid R_z = 0)\frac{P(X,Y)}{P(X,Y \mid R_z = 0)}$$

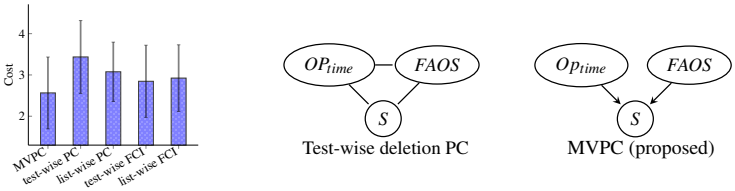
### Experiments

- Synthetic data




- Real-world datasets

Cognition and Aging USA ATR rehabilitation Study




## REFERENCES

- Mohan, Karthika, Judea Pearl, and Jin Tian. "Graphical models for inference with missing data." Advances in neural information processing systems. 2013.
- Strobl, Eric V., Shyam Visweswaran, and Peter L. Spirtes. "Fast causal inference with non-random missingness by test-wise deletion." International Journal of Data Science and Analytics (2017): 1-16.