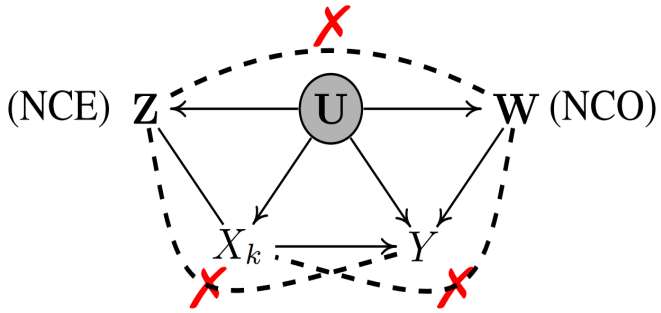


Automating the Selection of Proxy Variables of Unmeasured Confounders

Feng Xie

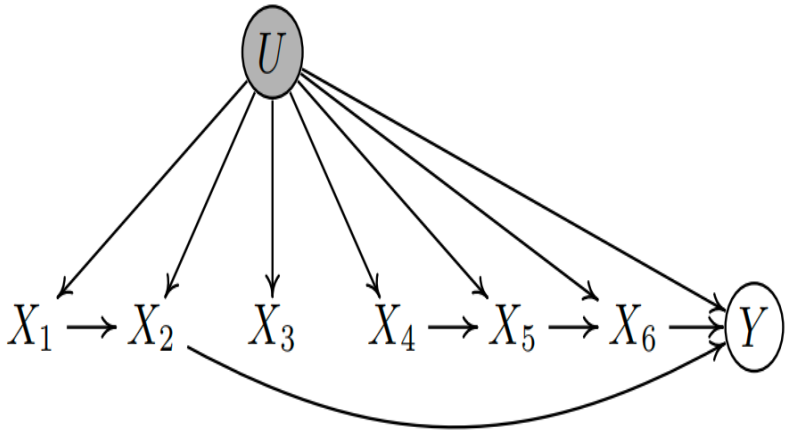
Joint Work with Zhengming Chen, Shanshan Luo*, Wang Miao, Ruichu Cai, and Zhi Geng

Problem Definition

- $\mathbf{X} = \{X_1, \ldots, X_p\}$ denotes a vector of treatments;
- Y denotes an outcome;
- $\mathbf{U} = \{U_1, \ldots, U_p\}$ denotes a vector of unmeasured confounders.

	X_1	X_2	…	X_6	Y
1	$X_{1,1}$	$X_{2,1}$	…	$X_{6,1}$	Y_1
2	$X_{1,2}$	$X_{2,2}$	⋮	$X_{6,2}$	Y_2
3	$X_{1,3}$	$X_{2,3}$	…	$X_{6,3}$	Y_3
⋮	⋮	⋮	⋮	⋮	⋮
n	$X_{1,n}$	$X_{2,n}$	…	$X_{6,n}$	Y_n

Observational dataset



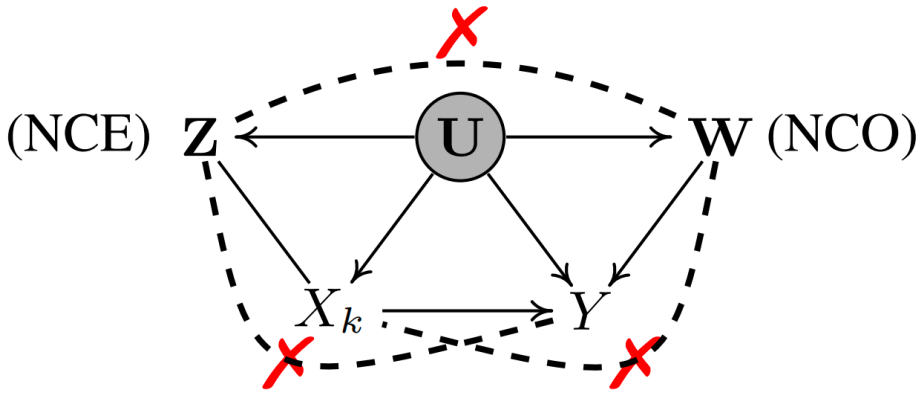
A simple causal graph

Can we estimate the causal effect of X_k *on* Y in the presence of an unmeasured confounder U?

Proximal Causal Learning

Def. 1 [NCE and NCO (Miao et al., 2018a; Shi et al., 2020b)] Given a target causal effect of X_k on Y in the case where **U** are the set of unmeasured confounding between X_k and Y, sets **Z** and **W** are the valid NCE and NCO respectively if the following conditions hold:

- **Z** is independent of Y conditional on (**U**, X_k), and
- **W** is independent of (X_k, **Z**) conditional on **U**.



"✗" indicates that the current active paths should not exist here.

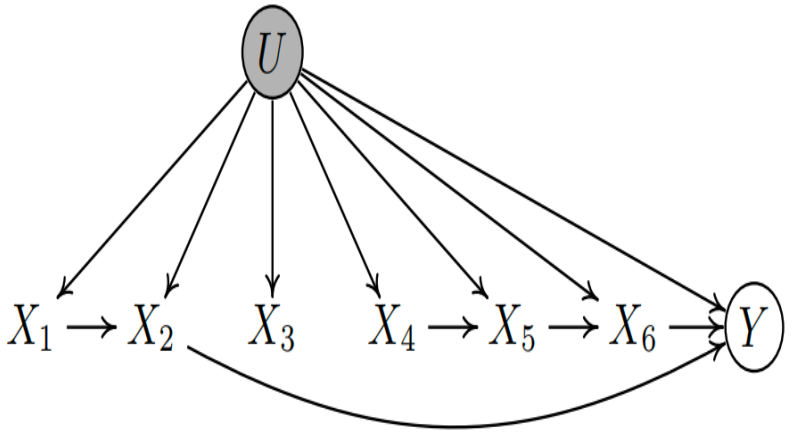
One can estimate the causal effect of X_k *on* Y using Z (NCE) and W (MCO)!

Problem Definition

- $\mathbf{X} = \{X_1, \ldots, X_p\}$ denotes a vector of treatments;
- Y denotes an outcome;
- $\mathbf{U} = \{U_1, \ldots, U_p\}$ denotes a vector of unmeasured confounders.

	X_1	X_2	…	X_6	Y
1	$X_{1,1}$	$X_{2,1}$	…	$X_{6,1}$	Y_1
2	$X_{1,2}$	$X_{2,2}$	⋮	$X_{6,2}$	Y_2
3	$X_{1,3}$	$X_{2,3}$	…	$X_{6,3}$	Y_3
⋮	⋮	⋮	⋮	⋮	⋮
n	$X_{1,n}$	$X_{2,n}$	…	$X_{6,n}$	Y_n

Observational dataset



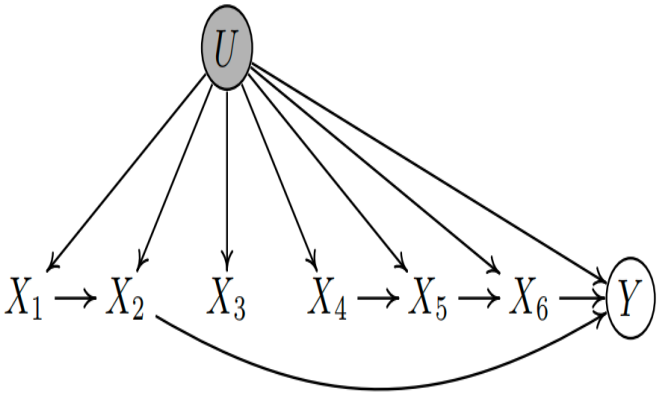
A simple causal graph

Is it possible to find valid proxy variables (NCE and NCO) of unmeasured confounders *U* relative to a causal relation $X_k \rightarrow Y$ only from measured variables X?

Linear, Acyclic Causal Model

- Assume variables were generated by the Linear, Acyclic Causal Model. We assume that **U** affects both treatments **X** and outcome *Y*.

$$\mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{C}\mathbf{U} + \varepsilon_{\mathbf{X}}, \quad c_{ij} \neq 0,$$
$$Y = \boldsymbol{\beta}^{\intercal}\mathbf{X} + \boldsymbol{\delta}^{\intercal}\mathbf{U} + \varepsilon_{Y}, \quad \boldsymbol{\delta}_i \neq 0,$$



Find the sufficient testable conditions that render the proxy variables relative to a causal relation $X_k \rightarrow Y$ identifiable?

Proxy Variables Estimator

Prop. 1 [Proxy Variables Estimator (Kuroki & Pearl, 2014)] Assume the system is a linear causal model. Further, assume that there exists one unmeasured confounder U that affects both treatment X_k and outcome Y, and that Z and W are NCE and NCO of confounder U. the unbiased estimator for the causal effect $\beta_{X_k \to Y}$ of X_k on Y is as follows,

$$\beta_{X_k \to Y} = \frac{\sigma_{X_k Y}\sigma_{wz} - \sigma_{X_k w}\sigma_{Yz}}{\sigma_{X_k X_k}\sigma_{wz} - \sigma_{X_k w}\sigma_{X_k z}}.$$

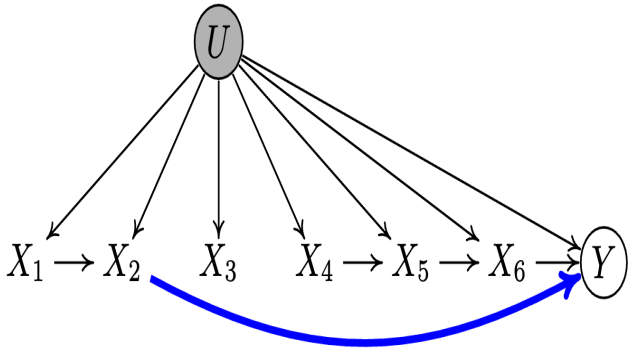
Prop. 2 [Extended Proxy Variables Estimator (Ours)] Assume the system is a linear causal model. Further, assume that there exists q unmeasured confounders, denoted by **U**, that affect both treatment X_k and outcome Y. Let **Z** with $|\mathbf{Z}| = q$ and **W** with $|\mathbf{W}| = q$ be two valid NCE and NCO of confounders **U** respectively. Thus, the unbiased estimator for the causal effect $\beta_{X_k \to Y}$ of X_k on Y is as follows,

$$\beta_{X_k \to Y} = \frac{\det(\Sigma_{\{X_k \cup \mathbf{Z}\},\{Y \cup \mathbf{W}\}})}{\det(\Sigma_{\{X_k \cup \mathbf{Z}\},\{X_k \cup \mathbf{W}\}})}.$$

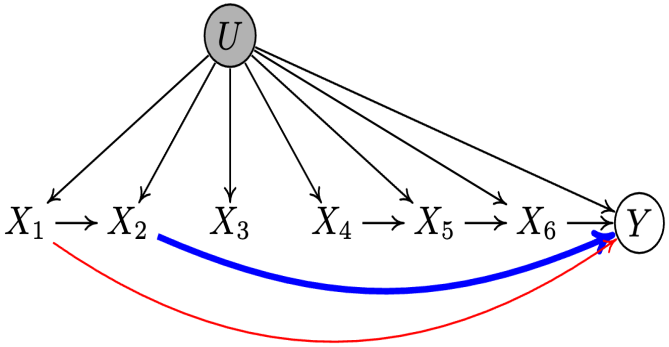
The extended proxy variables estimator allows us to obtain an unbiased causal effect, given valid NCE and NCO of confounders.

Identification of Proxy Variables with Second-Order Statistics

A Motivating Example: Consider the causal relationship $X_2 \rightarrow Y$ in the following causal diagrams. (a) X_1 and X_6 serve as valid NCE and NCO, respectively. (b) X_1 and X_6 serve as invalid NCE and NCO, respectively.



(a)



(b)

$\mathrm{rk}\left(\Sigma_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}\right) \leq 2$

$\Sigma_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}$ is full rank

The above facts show that lack of edge $X_1 \rightarrow Y$, i.e., the variable of NCE does not causally affect the primary outcome, has a testable implication.

Identification of Proxy Variables with Second-Order Statistics

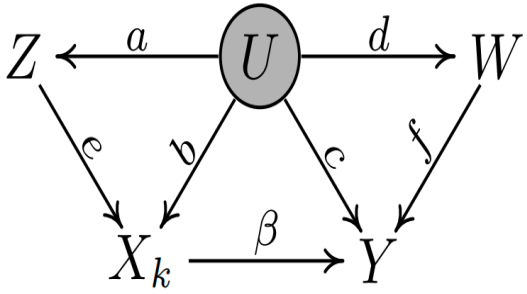
Rule 1. Let **A** and **B** be two disjoint subsets of **X**, where $|\mathbf{A}| = q$ and $|\mathbf{B}| = q$. Furthermore, let Q be a variable in $\{\mathbf{X} \backslash \{\mathbf{A} \cup \mathbf{B} \cup X_k\}\}$.
If 1) $\mathrm{rk}\left(\Sigma_{\{X_k, Q, \mathbf{A}\}, \{X_k, Y, \mathbf{B}\}}\right) \leq q + 1$, and 2) $\mathrm{rk}\left(\Sigma_{\{X_k, \mathbf{A}\}, \{Q, \mathbf{B}\}}\right) \leq q$, then **A** and **B** are valid NCE and NCO relative to $X_k \rightarrow Y$ respectively.

Rule 2. Let **A** and **B** be two disjoint subsets of **X**, where $|\mathbf{A}| = q + 1$ and $|\mathbf{B}| = q + 1$.
If 1) $\mathrm{rk}\left(\Sigma_{\{X_k, \mathbf{A}\}, \{X_k, Y, \mathbf{B}\}}\right) \leq q + 1$, and 2) $\mathrm{rk}\left(\Sigma_{\{X_k, \mathbf{A}\}, \{\mathbf{B}\}}\right) \leq q$, then **A** and **B** are valid NCE and NCO relative to $X_k \rightarrow Y$ respectively.

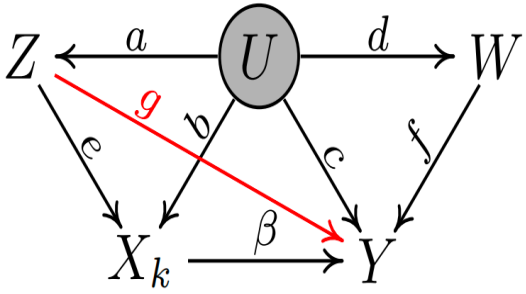
Roughly speaking, condition (1) of Rule 1(or 2) ensures that A is independent of Y conditional on (U, X_k), and condition (2) of Rule 1(or 2) ensures that A is independent of (X_k, B) conditional on U.

Identification of Proxy Variables with Second-Order Statistics

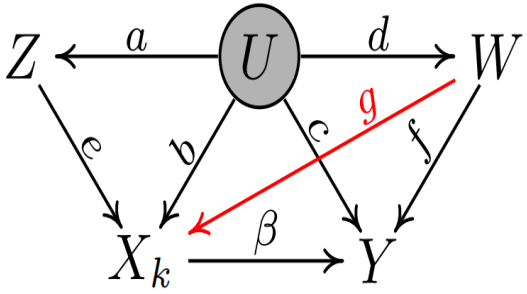
A Counter-example: However, the above three causal graphs entails all possible rank constraints in the marginal covariance matrix of $\{X_k, Y, Z, W\}$. 😞



(a)



(b)



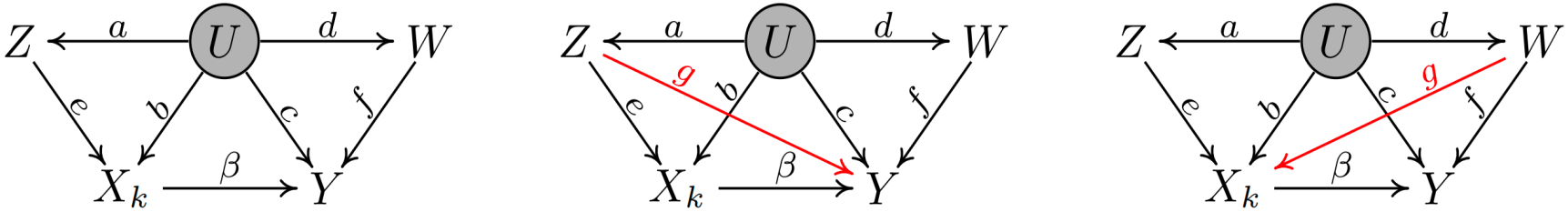
(c)

The above facts indicate that the conditions involving second-order statistics (marginal covariance matrix) are not necessary.

Identification of Proxy Variables with Higher-Order Statistics

Def. [GIN condition] Suppose all variables follow the linear non-Gaussian acyclic causal model. Let $\mathbf{Y}$, $\mathbf{Z}$ be two sets of random variables. We say that $(\mathbf{Z}, \mathbf{Y})$ follows the GIN condition if and only if $\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Y} \perp \mathbf{Z}$, where $\boldsymbol{\omega}$ satisfies $\boldsymbol{\omega}^{\mathrm{T}}\mathbb{E}(\boldsymbol{Y}\boldsymbol{Z}^{\mathrm{T}}) = \mathbf{0}$ and $\boldsymbol{\omega} \neq \mathbf{0}$.

A Motivating Example: consider the causal relationship $X_k \rightarrow Y$ in the above causal diagram, where Z and W serve as valid NCE and NCO for the $X_k \rightarrow Y$ in the subgraph (a). We analyze the observed variables and note the following result:



(a) (b) (c)

(a). $(\{X_k, Z\}, \{X_k, Y, W\})$ follows the GIN, and $(\{W\}, \{X_k, Z\})$ follows the GIN.
(b). $(\{X_k, Z\}, \{X_k, Y, W\})$ violates the GIN, and $(\{W\}, \{X_k, Z\})$ follows the GIN.
(c). $(\{X_k, Z\}, \{X_k, Y, W\})$ follows the GIN, and $(\{W\}, \{X_k, Z\})$ violates the GIN.

The facts show that lack of edges $\boldsymbol{Z} \rightarrow Y$ (i.e., the variable of NCE does not affect the outcome), or $\boldsymbol{W} \rightarrow X_k$ (i.e., the variable of NCO does not affect the treatment), has a testable implication.

Identification of Proxy Variables with Higher-Order Statistics

Rule 3. Let **A** and **B** be two disjoint subsets of **X**, where $|\mathbf{A}| = q$ and $|\mathbf{B}| = q$. Assume that all noise variables follow the non-Gaussian distributions.
If 1) $(\{X_k, \mathbf{A}\}, \{X_k, Y, \mathbf{B}\})$ follows the GIN constraint, and 2)$(\mathbf{B}, \{X_k, \mathbf{A}\})$ follows the GIN constraint, then **A** and **B** are valid NCE and NCO relative to $X_k \rightarrow Y$ respectively.

Roughly speaking, condition (1) of Rule 3 ensures that A is independent of Y conditional on (U, X_k), and condition (2) of Rule 3 ensures that A is independent of (X_k, B) conditional on U.

Simulation

We here consider the following two typical settings:

- **Gaussian case**: The noise terms are generated from $\mathbb{N}(0,1)$;
- **Non-Gaussian case**: The noise terms are generated from $\mathrm{Exp}(\lambda)$.

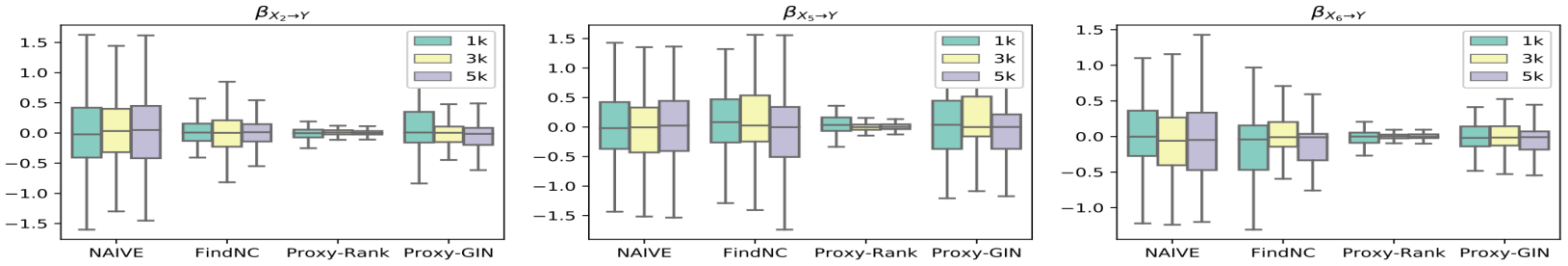


Figure 5. Performance of NAIVE, FindNC, Proxy-Rank, and Proxy-GIN on the Gaussian case.

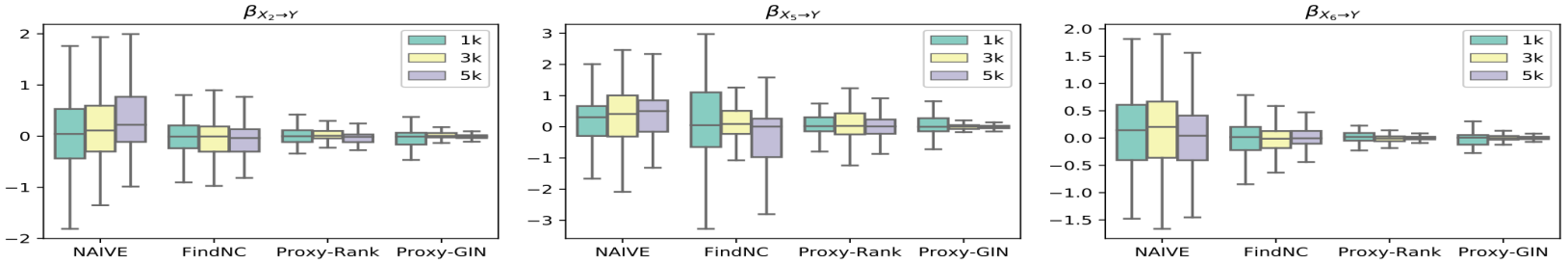


Figure 6. Performance of NAIVE, FindNC, Proxy-Rank, and Proxy-GIN on the Non-Gaussian case.

Our algorithms outperform other methods (with little bias for all causal effects) in all two settings, with all sample sizes.

Application to Real-World Data

We apply our method to analyze the causal effects of gene expressions on the body weight of F2 mice using the mouse obesity dataset as described by Wang et al. (2006). The dataset we used comprises 17 gene expressions that are known to potentially influence mouse weight, as reported by Lin et al. (2015).

- The gene expressions Gstm2, Sirpa, and 2010002N04Rik exhibit positive and significant effects on body weight, whereas the gene expression Dscam demonstrates a negative impact on body weight.

- Igfbp2 (Insulin-like growth factor binding protein 2) displays negative and significant effects on body weight, attributable to its role in mitigating the development of obesity, as supported by Wheatcroft et al. (2007).

- Irx3 (Iroquois homebox gene 3) exhibits negative and significant effects on body weight, which can be attributed to its association with lifestyle changes and its pivotal role in weight regulation through energy balance, as elucidated in Schneeberger (2019).

Conclusions and Future work

Conclusions

- Introduce an extended proxy variable estimator to handle multiple unmeasured confounders between treatments and outcomes;
- Provide two specific identifiability conditions for selecting proxy variables, based on the second-order and higher-order statistics;
- Proposes two efficient algorithms for selecting proxy variables, with their effectiveness substantiated by experimental results.

Future work

- Investigate the identifiability conditions for selecting proxy variables under nonlinear causal models.