Automating the Selection of Proxy Variables of Unmeasured Confounders

Feng Xie [1] **Zhengming Chen** [2 3] **Shanshan Luo** [1] **Wang Miao** [4] **Ruichu Cai** [2] **Zhi Geng** [1]

Abstract

Recently, interest has grown in the use of proxy variables of unobserved confounding for inferring the causal effect in the presence of unmeasured confounders from observational data. One difficulty inhibiting the practical use is finding valid proxy variables of unobserved confounding to a target causal effect of interest. These proxy variables are typically justified by background knowledge. In this paper, we investigate the estimation of causal effects among multiple treatments and a single outcome, all of which are affected by unmeasured confounders, within a linear causal model, without prior knowledge of the validity of proxy variables. To be more specific, we first extend the existing proxy variable estimator, originally addressing a single unmeasured confounder, to accommodate scenarios where multiple unmeasured confounders exist between the treatments and the outcome. Subsequently, we present two different sets of precise identifiability conditions for selecting valid proxy variables of unmeasured confounders, based on the second-order statistics and higher-order statistics of the data, respectively. Moreover, we propose two data-driven methods for the selection of proxy variables and for the unbiased estimation of causal effects. Theoretical analysis demonstrates the correctness of our proposed algorithms. Experimental results on both synthetic and real-world data show the effectiveness of the proposed approach.

[1]Department of Applied Statistics, Beijing Technology and Business University, Beijing, China [2]School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China [3]Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE [4]Department of Probability and Statistics, Peking University, Beijing, China. Correspondence to: Shanshan Luo <shanshanluo@btbu.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Estimating the causal effect from observational data is a fundamental problem in various fields of scientific research, including social sciences (Pearl, 2009; Spirtes et al., 2000), economics (Imbens & Rubin, 2015), public health (Hernán & Robins, 2006a), and machine learning (Spirtes, 2010; Peters et al., 2017; Fernández-Loría & Provost, 2022). Within the framework of causal graphical models, covariate adjustment, such as the use of the back-door criterion, emerges as a powerful and primary tool for estimating causal effects from observational data (Pearl, 2009; Van Der Zander et al., 2019). However, although this method has been used in a range of fields, it should be noted that biased causal effects can arise when unmeasured confounders are present and the covariate adjustment set does not exist in the system (Pearl, 2009; Rotnitzky & Smucler, 2020; Cheng et al., 2022).

The method of instrumental variables is a general approach used to estimate the causal effect of interest in the presence of unobserved confounders (Pearl, 2009; Wright, 1928; Goldberger, 1972; Bowden & Turkington, 1990). This method has been extensively studied in practical sciences, including economics (Imbens & Rubin, 2015; Imbens, 2014), sociology (Pearl, 2009; Spirtes et al., 2000) and epidemiology (Hernán & Robins, 2006b; Baiocchi et al., 2014). In practice, it can be quite challenging to identify a valid instrumental variable (Pearl, 1995; Kuroki & Cai, 2005; Kang et al., 2016; Silva & Shimizu, 2017; Gunsilius, 2021; Xie et al., 2022a; Cheng et al., 2023). Sometimes, in the system of interest, an instrumental variable may not even exist.

Recently, the proximal causal learning method, also referred to as negative control, has emerged as an alternative strategy to address unmeasured confounders and estimate the unbiased causal effects of interest (Kuroki & Pearl, 2014; Miao et al., 2016; de Luna et al., 2017; Miao et al., 2018a;b; Wang & Blei, 2019; Shi et al., 2020a; Tchetgen et al., 2020; Singh, 2020; Wang & Blei, 2021; Mastouri et al., 2021; Xu et al., 2021; Shpitser et al., 2023). This method allows us to infer the causal effect of interest by observing suitable proxy variables for unmeasured confounding, with these proxy variables often being termed Negative Controls (NCs). NCs are readily applicable in various domains (Lipsitch et al., 2010; Sofer et al., 2016). For instance, one study of the causal effect of the flu shot (X_k) on influenza-related

hospitalization (Y), where there exists unmeasured health-seeking behavior (U) (Shi et al., 2020b). The proximal causal learning method operates on the following principles: (i) find a variable e.g., a person's annual wellness visit history (Z), that is influenced by confounder U and has no direct effect on the outcome Y, referred to as the Negative Control Exposure (NCE); (ii) find another variable, e.g., a person's injury/trauma hospitalization (W), that is influenced by confounder U and is not causally affected by the treatment X_k, referred to as the Negative Control Outcome (NCO); and (iii) use these two proxy variables to estimate the causal effect of flu shot on influenza-related hospitalization. Figure 1 illustrates the corresponding causal graph that satisfies the above conditions respectively, with further details in Section 2.2. However, although these methods have been used in a range of fields, the valid proxy variables are typically justified by background knowledge in those works. Thus, it is vital to develop statistical methods for selecting proxy variables of unmeasured confounding from observational data.

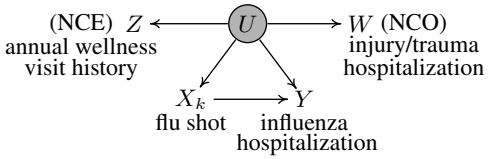


Figure 1. A typical confounder proxy causal diagram. Z and W are NCE and NCO of unmeasured confounder U for the causal relationship $X_k \to Y$.

Recently, Kummerfeld et al. (2022) established sufficient conditions for selecting valid NCE and NCO of one unmeasured confounder in a linear causal model, making valuable contributions to the field. However, their work is limited to single-treatment settings, where potential proxy variables cannot directly affect both the treatment and the outcome of interest. In reality, multiple-treatment scenarios exist, where unmeasured confounders influence both treatments and the outcome. For instance, in gene expression studies, there exist multiple gene expressions may affect the trait of a human of interest (e.g., body weight) (Miao et al., 2022). Besides, their work only considers a particular class of proxy variables of an unmeasured confounder, where those proxy variables are independent of both the treatment and outcome conditional on the unmeasured confounder while our work does not restrict this condition. In this paper, we tackle the challenge of proxy variable identification in a more complex scenario, where the proxy variables can have effects on the outcome, and multiple unmeasured confounders may exist in the system. Specifically, we make the following contributions:

1. We extend the existing proxy variable estimator that deals with a single unmeasured confounder, as previously discussed by Kuroki & Pearl (2014), to accommodate scenarios where multiple unmeasured confounders exist between treatments and the outcome.
2. We present two different sets of precise identifiability conditions for selecting proxy variables of unmeasured confounders, based on the second-order statistics [1] and higher-order statistics [2], respectively.
3. We propose two efficient algorithms for selecting proxy variables of unmeasured confounders. The first algorithm leverages the rank-deficiency properties of covariance matrices, while the second algorithm takes advantage of the non-Gaussianity of the data. Both algorithms consistently estimate the desired causal effect and come with theoretical proofs that establish their correctness.
4. We demonstrate the efficacy of the proposed algorithms on both synthetic and real-world data.

2. Preliminaries

2.1. Notations

Our work is in the framework of causal graphical models (Pearl, 2009; Spirtes et al., 2000). In a directed acyclic graph (DAG) $\mathcal{G}$, a **path** is a sequence of nodes $\{X_1, \ldots, X_r\}$ such that X_i and X_{i+1} are adjacent in $\mathcal{G}$, where $1 \leq i < r$. A **collider** on a path $\{X_1, ...X_p\}$ is a node X_i , $1 < i < p$, such that X_{i-1} and X_{i+1} are parents of X_i. A **trek** between X_i and X_j is a path that does not contain any colliders in $\mathcal{G}$. A **source** in a trek is a unique node such that no arrows point to it. We use the ordered pair of directed paths (P_1, P_2) denotes a trek in $\mathcal{G}$ from X_i to X_j, where P_1 has sink X_i, P_2 has sink X_j, and both P_1 and P_2 have the same source.[3] Other commonly used concepts in graphical models, such as d-separation, can be found in standard sources (Pearl, 1988; 2009; Spirtes et al., 2000).

We denote vectors and matrices by boldface letters. The (i, j) entry of matrix $\mathbf{M}$ is denoted by $\mathbf{M}_{i,j}$. The notation $|\mathbf{A}|$ denotes the cardinality of set $\mathbf{A}$. The notation $\mathbf{\Sigma}_{\mathbf{A},\mathbf{B}}$ denotes the cross-covariance matrix of set $\mathbf{A}$ (rows) and $\mathbf{B}$ (columns). The notation $\mathrm{rk}(\mathbf{C})$ denotes the rank of matrix $\mathbf{C}$, e.g., $\mathrm{rk}(\mathbf{\Sigma}_{\mathbf{A},\mathbf{B}})$ denotes the rank of cross-covariance matrix of set $\mathbf{A}$ and $\mathbf{B}$. The determinant of a matrix $\mathbf{A}$ is denoted $\det(\mathbf{A})$. We use the notation $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}$ for "$\mathbf{A}$ is independent of $\mathbf{B}$ given $\mathbf{C}$", and $\mathbf{A} \not\perp\!\!\!\perp \mathbf{B}|\mathbf{C}$ for the negation of the same sentence (Dawid, 1979).

2.2. Proximal Causal Learning

The proximal causal learning approach offers a new strategy for inferring the causal effect of interest in the presence of unmeasured confounders (Kuroki & Pearl, 2014; de Luna

[1] Second-order statistics means the second-order moments (like covariances or correlations).

[2] Higher-order statistics means beyond the second-order moments in statistics, e.g., skewness, kurtosis, etc. of the data.

[3] A sink of a graph $\mathcal{G}$ is any node that is not a parent of any other node.

et al., 2017; Miao et al., 2018a; Wang & Blei, 2019; Shi et al., 2020a; Tchetgen et al., 2020). Specifically, suppose that X_k is the treatment, Y is the outcome, and $\mathbf{U}$ represents the set of unmeasured confounders between X_k and Y. The theory around the proximal causal learning approach says that the target causal effect of X_k on Y can be identified when two sets of proxy variables, $\mathbf{Z}$ and $\mathbf{W}$, are available for the unmeasured confounder $\mathbf{U}$. In such cases, the proxy set $\mathbf{Z}$, referred to as the Negative Control Exposure (NCE), does not causally affect the primary outcome Y, and another proper proxy set $\mathbf{W}$, called the Negative Control Outcome (NCO), is not causally affected by the treatment X_k. The graphical condition for NCE and NCO relative to a target causal effect of X_k on Y is described in Definition 1, and an illustrative example is provided accordingly.

Definition 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 $\mathbf{U}$ are the set of unmeasured confounding between X_k and Y, sets $\mathbf{Z}$ and $\mathbf{W}$ are the valid NCE and NCO respectively if the following conditions hold:*

1. *$\mathbf{Z}$ is independent of Y conditional on $(\mathbf{U}, X_k)$, i.e., $\mathbf{Z} \perp\!\!\!\perp Y|(\mathbf{U}, X_k)$, and*
2. *$\mathbf{W}$ is independent of $(X_k, \mathbf{Z})$ conditional on $\mathbf{U}$, i.e., $\mathbf{W} \perp\!\!\!\perp (X_k, \mathbf{Z})|\mathbf{U}$.*

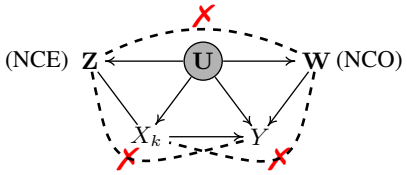


Figure 2. Diagram of one possible violation of NCE and NCO assumptions. Dashed lines represent active paths. The symbol "✗" indicates that the current active paths should not exist here.

For the rest of the paper, we will call the above two conditions the graphical criteria for proxy variables validity, or simply ***proximal criteria***. Figure 2 is an illustration of the NCE and NCO conditions and one potential violation of the NCE and NCO conditions. Notice that the validity of both proxy variables NCE and NCO are mutually dependent on each other. Therefore, when we say that proxy variables are valid for a causal relationship, we mean that both NCE and NCO are valid simultaneously.

Definition 2 (**Connected (Disconnected) NCE and NCO**)**.** *Assume $\mathbf{Z}$ and $\mathbf{W}$ are NCE and NCO of unmeasured confounders $\mathbf{U}$ for the causal relationship $X_k \to Y$. We refer to the set $\mathbf{Z}$ as Connected (Disconnected) NCE if $\mathbf{Z} \not\perp\!\!\!\perp X_k|\mathbf{U}$ ($\mathbf{Z} \perp\!\!\!\perp X_k|\mathbf{U}$). Similarly, we refer to the set $\mathbf{W}$ as Connected (Disconnected) NCO if $\mathbf{W} \not\perp\!\!\!\perp Y|\mathbf{U}$ ($\mathbf{W} \perp\!\!\!\perp Y|\mathbf{U}$).*

Definition 3 (**Quadruple-disconnected NC**)**.** *Assume $\mathbf{Z}$ and $\mathbf{W}$ are NCE and NCO of unmeasured confounder $\mathbf{U}$ for the causal relationship $X_k \to Y$. We say a variable Q is a Quadruple-disconnected NC if $Q \perp\!\!\!\perp X_k|\mathbf{U}$, $Q \perp\!\!\!\perp Y|\mathbf{U}$, $Q \perp\!\!\!\perp \mathbf{Z}|\mathbf{U}$, and $Q \perp\!\!\!\perp \mathbf{W}|\mathbf{U}$.*

Example 1. *Consider the causal relationship $X_2 \to Y$ in Figure 3. X_1 and X_6 are valid disconnected NCE and disconnected NCO relative to $X_2 \to Y$, respectively. Because $X_3 \perp\!\!\!\perp X_2|\mathbf{U}$, $X_3 \perp\!\!\!\perp Y|\mathbf{U}$, $X_3 \perp\!\!\!\perp X_1|\mathbf{U}$, and $X_3 \perp\!\!\!\perp X_6|\mathbf{U}$, X_3 can serve as a Quadruple-disconnected NC.*

Proposition 1 (**Proxy Variables Estimator** (Kuroki & Pearl, 2014))**.** *Assume the system is a linear causal model, i.e., all variables are continuous and the causal relationships among variables are linear. Further, assume that there exist 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, e.g., the causal graph in Figure 1, the unbiased estimator for the causal effect $\beta_{X_k \to Y}$ of X_k on Y is as follows,*

$$\beta_{X_k \to Y} = \frac{\sigma_{X_k Y}\sigma_{WZ} - \sigma_{X_k W}\sigma_{YZ}}{\sigma_{X_k X_k}\sigma_{WZ} - \sigma_{X_k W}\sigma_{X_k Z}} \tag{1}$$

where $\sigma_{X_k Y}$ is the covariance between X_k and Y, etc.

It is worth noting that this standard estimator is only applicable in the case of a single U. For scenarios involving multiple confounders $\mathbf{U}$, please refer to the extended estimator introduced in Section 3.

2.3. Model Definition

In this paper, let $\mathbf{X} = \{X_1, \ldots, X_p\}^\intercal$ denote a vector of p-dimensional treatments, Y denote an outcome, and $\mathbf{U} = \{U_1, \ldots, U_q\}^\intercal$ denote a vector of q-dimensional unmeasured confounders. Analogous to Wang & Blei (2019); Ogburn et al. (2019); D'Amour (2019b), we consider the case that $\mathbf{U}$ affects both treatments $\mathbf{X}$ and outcome Y. Without loss of generality, we assume that all variables have a zero mean. We here restrict our attention to a linear acyclic causal model,

$$\begin{aligned} \mathbf{X} &= \mathbf{B}\mathbf{X} + \mathbf{C}\mathbf{U} + \varepsilon_{\mathbf{X}}, \quad c_{ij} \neq 0, \\ Y &= \boldsymbol{\beta}^\intercal \mathbf{X} + \boldsymbol{\delta}^\intercal \mathbf{U} + \varepsilon_Y, \quad \boldsymbol{\delta}_i \neq 0, \end{aligned} \tag{2}$$

where $\boldsymbol{\beta}$ is the column vector that signifies the causal effects of interest. The noise terms in $\varepsilon_{\mathbf{X}}$ and $\mathbf{U}$ are independent of each other, ε_Y is independent of $\mathbf{X}$ and $\mathbf{U}$. We assume that the generating process is recursive. That is to say, the causal relationships among variables can be represented by a DAG (Pearl, 2009; Spirtes et al., 2000).

In contrast to the single-treatment model studied in Kummerfeld et al. (2022), where potential variables cannot directly influence the treatment and outcome of interest, our work explores a more general scenario. In our study, the system of interest accommodates multiple unmeasured confounders and multiple treatments, where latent variables can have a direct effect on both the treatment and the outcome of interest. Figure 3 provides a simple graph that satisfies our model while violating the model in Kummerfeld et al. (2022), where the variables $X_i, i = 1, \ldots, 6$ are potential

treatments, Y is the outcome, and U is an unmeasured confounder.

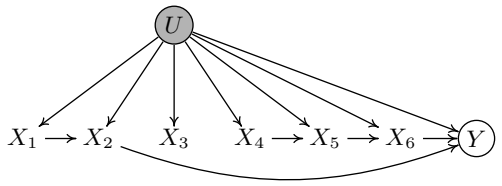


Figure 3. A simple causal graph involving 6 potential treatments and one outcome.

For convenience, we further assume that the number of unmeasured confounders is known. When the number of confounders is unknown, consistent estimation of the number of confounders has been well-established by Bai & Ng (2002) under factor models. In practical applications, one may directly use open-source software, to perform a significance test to determine whether the number of confounders in a factor model is sufficient to capture the full dimensionality of the dataset, as stated in Miao et al. (2022). Notice that the model of Equation 2 assumes that $\mathbf{U}$ affects both treatments $\mathbf{X}$ and outcome Y, i.e., all entries of $\mathbf{C}$ are non-zero, as the studied in Kummerfeld et al. (2022). In Appendix H, we explored a scenario in which, after applying certain necessary preprocessing steps, our theory remains applicable, even when certain entries of $\mathbf{C}$ are zero.

Goal: The goal of this paper is to identify sets NCE and NCO of unmeasured confounders that satisfy proximal criteria for a given casual relationship $X_k \to Y, X_i \in \mathbf{X}$ and estimate the total causal effect of treatment X_i on Y simultaneously.

Remark 1. *In the problem described above, conventional constraint-based causal discovery methods that account for unmeasured confounders, such as the FCI (Fast Causal Inference) algorithm (Spirtes et al., 1995; Zhang, 2008) or its variants, like the RFCI algorithm (Really Fast Causal Inference) (Colombo et al., 2012), result in a fully connected causal graph. This occurs because unmeasured confounders $\mathbf{U}$ affect both the treatments X and Y. As a result, it becomes challenging to identify valid Negative Control Exposure (NCE) and Negative Control Outcome (NCO) for unmeasured confounders from the resulting (fully connected) graph.*

3. Extended Proxy Variables Estimator with Multiple Unmeasured Confounders

In this section, we will extend the existing proxy variable estimator with a single unmeasured confounder (Proposition 1) to handle the case when there exist multiple unmeasured confounders between treatment and outcome. Specifically, we build upon the work of Kuroki & Pearl (2014) on the proxy variable estimator with an unmeasured confounder and extend it to include multiple unmeasured confounders in the case of a linear causal model. To improve readability, we defer all proofs to Appendix K.

Proposition 2 (Extended Proxy Variables Estimator). *Assume the system is a linear causal model, i.e., all variables are continuous and the causal relationships among variables are linear, and assume there exist q unmeasured confounders, denoted by $\mathbf{U}$, that affect both treatment X_k and outcome Y. Let $\mathbf{Z}$ with $|\mathbf{Z}| = q$ and $\mathbf{W}$ with $|\mathbf{W}| = q$ be two valid NCE and NCO of $\mathbf{U}$ respectively. Thus, the unbiased estimator for the total causal effect $\beta_{X_k \to Y}$ of X_k on Y is as follows,*

$$\beta_{X_k \to Y} = \frac{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{Z}\},\{Y \cup \mathbf{W}\}})}{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{Z}\},\{X_k \cup \mathbf{W}\}})}. \tag{3}$$

Proposition 2 asserts that, given two valid q-dimensional NCE and q-dimensional NCO for the causal relationship $X_k \to Y$ when there exist q-dimensional unmeasured confounders, then the total causal effect of X_k on Y can be consistently estimated using Eq. 3. Note that if the dimension of NCE and NCO is less than q, the estimated $\beta_{X_k \to Y}$ will be biased (see an example described in Appendix. A).

Remark 2. *If $q = 1$, i.e., there exists only one unmeasured confounder between X_k and Y, the estimator in Proposition 2 is equal to the estimator in Proposition 1.*

Remark 3. *According to Proposition 2, for a given causal relationship $X_k \to Y$, the necessary conditions for the NCE $\mathbf{Z}$ and NCO $\mathbf{W}$ to be valid are that $\mathbf{\Sigma}_{\{X_k,\mathbf{Z}\},\{Y,\mathbf{W}\}}$ and $\mathbf{\Sigma}_{\{X_k,\mathbf{Z}\},\{X_k,\mathbf{W}\}}$ both are full rank, i.e., $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{Z}\},\{Y,\mathbf{W}\}}) = q + 1$, $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{Z}\},\{X_k,\mathbf{W}\}}) = q + 1$.*

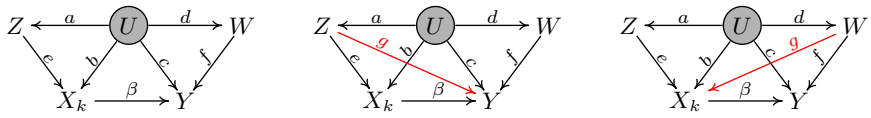
4. Identifiability with Second-Order Statistics

In this section, we first investigate the identifiability of the proxy variables in the model described in Eq. 2 with second-order statistics. Then, we provide a data-driven method for selecting valid proxy variables (i.e., NCE and NCO) of each treatment X_k on outcome Y and obtaining its corresponding unbiased causal effect of X_k on Y simultaneously. All proofs are included in Appendix K.

4.1. Identification of Proxy Variables with Second-Order Statistics

In this section, we investigate the identifiability of proxy variables using second-order statistics. Before giving our main results, we first introduce the concept of rank constraints (which is an extension of the famous Tetrad constraints presented in Spearman (1928)), which is an essential constraint that leverages the second-order statistics derived from the data (Sullivant et al., 2010; Spirtes, 2013).

Definition 4 (Rank Constraint). *Suppose all variables follow the linear acyclic causal model. Let $\mathbf{A}$ and $\mathbf{B}$ be two*



(a) (b) (c)

Figure 4. A linear causal model with any of the graphical structures above entails all possible rank constraints in the marginal covariance matrix of $\{X_k, Y, Z, W\}$.

sets of random variables. A rank constraint in the submatrix of the covariance matrix $\mathrm{rk}(\mathbf{\Sigma_{A,B}})$ *is any constraint of the type* $\mathrm{rk}(\mathbf{\Sigma_{A,B}}) \leq r$*, where* r *is some constant.*

It is noteworthy that if one uses rank constraints to structural constraints with unobserved variables, the rank-faithful assumption is necessary (Spirtes, 2013).

Definition 5 (rank-faithfulness)**.** *Let a probability distribution* P *be rank-faithful to a DAG* $\mathcal{G}$ *if every rank constraint on a sub-covariance matrix that holds in* P *is entailed by every linear structural equation model with respect to* $\mathcal{G}$*.*

The rank-faithfulness assumption allows us to use the rank-deficiency constraints to impose structural constraints with unobserved variables. Intuitively speaking, the set of values of free parameters for which $\mathrm{rk}(\mathbf{\Sigma_{A,B}}) \leq r$ has a Lebesgue measure of 0. Note that this assumption does not restrict the data distribution, making it a distribution-free assumption. Furthermore, the practicality of rank-faithfulness has been demonstrated through simulation results and applications in Kummerfeld & Ramsey (2016); Xie et al. (2022b); Huang et al. (2022), as well as in our paper. For a further discussion of rank-faithfulness assumptions, please refer to Section 4 in Spirtes (2013) for more details.

A Motivating Example: Before showing the theoretical results, we give a simple example to illustrate the basic idea. Consider the causal diagram in Figure 3. We observe a causal relationship $X_2 \to Y$. In this case, X_1 and X_6 serve as valid NCE and NCO, respectively, for this causal relationship. The cross-covariance matrix $\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}$ is singular, that is,

$$\det(\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}) = 0. \tag{4}$$

By Eq.4, we quickly know that $\mathrm{rk}(\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}) \leq 2$. We introduce an edge $X_1 \to Y$ in the graph of Figure 3, causing X_1 and X_6 to become invalid NCE and NCO, respectively, concerning the causal relationship $X_2 \to Y$. The cross-covariance matrix $\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}$ will no longer have a vanishing determinant, and instead,

$$\det(\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}) \neq 0. \tag{5}$$

That is to say, $\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}$ is full rank. Assuming the distribution is rank-faithful to the graph, the above facts show that lack of edge $X_1 \to Y$, i.e., the variable of NCE does not causally affect the primary outcome, has a testable implication.

We now investigate the conditions under which the valid NCE and NCO of unmeasured confounder $\mathbf{U}$ relative to a causal relationship $X_k \to Y$ can be identified in terms of rank constraints. To estimate the causal effect of $X_k \in \mathbf{X}$ on Y in the system, according to Proposition 2, the minimal condition is as follows,

Assumption 1. *For a given causal relationship* $X_k \to Y$ *in the system, there exist at least* q *variables in* $\mathbf{X}$ *that qualify as NCE and* q *variables in* $\mathbf{X}$ *that qualify as NCO.*

Assumption 1 is a very natural condition that one expects to hold. This is because if Assumption 1 fails, i.e., there are no valid sets of NCE and NCO for the causal relationship in the system, then we can not estimate the unbiased causal effect of interest using the extended proxy variables estimator. Unfortunately, Assumption 1 is an *insufficient* condition for identifying the sets of NCE and NCO in terms of rank constraints. An illustrative example is given below.

Example 2 (Counterexample)**.** *Consider the causal diagrams shown in Figure 4. Assume that the data are generated from a linear causal model and rank-faithfulness holds. We find that all possible rank constraints are full-rank (no rank-deficiency) in the marginal matrix of* $\{X_k, Y, Z, W\}$*, e.g.,* $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k,Y\},\{Z,W\}}) = 2$ *in three subgraphs. However, according to proximal criteria, we know that only in subgraph (a),* Z *and* W *are NCE and NCO of unmeasured confounder* U *for the causal relationship* $X_k \to Y$*, while they are not in other subgraphs (b) and (c). The above facts imply that one can not identify valid NCE and NCO using rank constraints under Assumption 1.*

We next give two sufficient conditions that render the sets of NCE and NCO of the unmeasured confounders $\mathbf{U}$ relative to a causal relationship $X_k \to Y$ identifiable, respectively.

Assumption 2. *For a given causal relationship* $X_k \to Y$ *in the system, the following conditions hold: i) there exist at least* q *variables in* $\mathbf{X}$ *that qualify as NCE and* q *variables in* $\mathbf{X}$ *that qualify as NCO, and ii) there exist at least one Quadruple-disconnected NC relative to* $X_k \to Y$*.*

Assumption 2 says that apart from satisfying the minimum number of NCE and NCO, the system also requires at least one additional Quadruple-disconnected NC. Assumption 2 is much milder than the assumptions considered in Kummerfeld et al. (2022), since for a causal relationship $X_k \to Y$ in the presence of one unmeasured confounder, we only need one Quadruple-disconnected NC, one unrestricted NCE, and one unrestricted NCO, while they need three Quadruple-

disconnected NCs. Roughly speaking, the existence of a quadruple disconnected negative control, termed as Q, simplifies the condition testing in the proximal criteria. That is to say, when verifying the two criteria of proximal criteria, that is, when finding two Rank-Deficiency Constraints, it is only necessary to add this variable to the rows and columns of the cross-covariance matrix. For example, the variable X_3 in Example 3 is a quadruple disconnected negative control.

Lemma 1. *Assume that the input data $\mathbf{X}$ and Y strictly follow the Equation 2 and the rank-faithfulness holds. If Assumption 2 holds, then the underlying NCE and NCO relative to the causal relationship $X_k \to Y$ can be identified by using the following rule.*

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

Example 3 ($\mathcal{R}1$)**.** *Let's consider the causal diagram shown in Figure 3. We consider the causal relationship $X_2 \to Y$ (Assumption 2 holds for $X_2 \to Y$). Let $\mathbf{A} = \{X_1\}$, $\mathbf{B} = \{X_6\}$, and $Q = X_3$. We check $\mathcal{R}1$ and obtain that 1) $\mathrm{rk}(\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}) \leq 2$, and 2) $\mathrm{rk}(\mathbf{\Sigma}_{\{X_2,X_1\},\{X_3,X_6\}}) \leq 1$. These facts imply that X_1 and X_6 are valid NCE and NCO relative to $X_2 \to Y$, respectively.*

We next introduce another sufficient condition when there is no proper Quadruple-disconnected NC mentioned in Assumption 2 in the system.

Assumption 3. *For a given causal relationship $X_k \to Y$ in the system, there exist at least $q+1$ variables in $\mathbf{X}$ that qualify as NCE and $q+1$ variables in $\mathbf{X}$ that qualify as NCO.*

Assumption 3 states that apart from satisfying the minimum number of NCE and NCO, i.e., q NCE and q NCO, the system also requires at least one additional NCE and one additional NCO.

Lemma 2. *Assume that the input data $\mathbf{X}$ and Y strictly follow the Equation 2 and the rank-faithfulness holds. If Assumption 3 holds, then the underlying NCE and NCO relative to the causal relationship $X_k \to Y$ can be identified by using the following rule.*

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

Example 4 ($\mathcal{R}2$)**.** *Continue to consider the causal diagram shown in Figure 3. We now consider the causal relationship $X_6 \to Y$. Let $\mathbf{A} = \{X_4, X_5\}$, and $\mathbf{B} = \{X_1, X_2\}$. We check $\mathcal{R}2$ and obtain that 1) $\mathrm{rk}(\mathbf{\Sigma}_{\{X_6,X_4,X_5\},\{X_6,Y,X_1,X_2\}}) \leq 2$, and 2) $\mathrm{rk}(\mathbf{\Sigma}_{\{X_6,X_4,X_5\},\{X_1,X_2\}}) \leq 1$. These facts imply that $\{X_4, X_5\}$ and $\{X_1, X_2\}$ are valid NCE and NCO relative to $X_6 \to Y$ respectively.*

Building upon Lemmas 1 and 2, we provide graphical conditions that are sufficient for the identifiability of NCE and NCO in terms of rank constraints.

Theorem 1 (Identifiability of NCE and NCO with Rank Constraints)**.** *Assume that the input data $\mathbf{X}$ and Y strictly follow the Equation 2 and the rank-faithfulness holds. Then the underlying NCE and NCO relative to the causal relationship $X_k \to Y$ can be identified if Assumption 2 or 3 is satisfied.*

4.2. Algorithm

In this section, we will leverage the above theoretical results and propose a data-driven method called Proxy-Rank to estimate the total causal effects of treatment $X_k \in \mathbf{X}$ on outcome Y:

Proxy-Rank algorithm

1. Given a p-dimensional treatments $\mathbf{X}$, outcome Y, the number of unmeasured confounders q. Initialize the sets of NCE, and NCO, causal effect, as $\mathcal{NCE}$, $\mathcal{NCO}$, and $\mathcal{C}$, respectively, with an empty set, i.e., $\mathcal{NCE} := \emptyset$, $\mathcal{NCO} := \emptyset$, and $\mathcal{C} := \emptyset$.
2. Find valid NCE and NCO of unmeasured confounder $\mathbf{U}$ relative to per causal relationship $X_k \to Y$ according to Lemmas 1 and 2.
3. Estimate the corresponding unbiased causal effect by Proposition 2 given NCE and NCO for per causal relationship $X_k \to Y$. Otherwise, output a value (NA) indicating the lack of knowledge to obtain the unbiased causal effect.

The specific details of algorithm execution are provided in the Appendix E.

We now show the correctness of the proposed algorithm. That is to say, for the causal relationships of interest, our algorithm outputs the true proxy variables and the unbiased estimation of causal effects.

Theorem 2 (Correctness)**.** *Assume that the data Y and $\mathbf{X}$ strictly follow the Equation 2 and the rank-faithfulness holds. Given infinite samples, the Proxy-Rank algorithm outputs the true causal effect $\mathcal{C}$ correctly.*

For more discussion on the consistency result and convergence rate of the above theorem, please refer to Appendix G.

We finally analyze the complexity of the Proxy-Rank algorithm. Let q be the number of latent confounders, and p be

the number of treatments. There are two dominant parts. One dominant part is to check $\mathcal{R}_1$ of Lemma 1 with worst-case complexity is $\mathcal{O}\left(\frac{p!}{q!\cdot q!\cdot(p-2q-1)!}\right)$. The other dominant part is to check $\mathcal{R}_2$ of Lemma 2 with worst-case complexity is also $\mathcal{O}\left(\frac{p!}{q!\cdot q!\cdot(p-2q-1)!}\right)$. Hence, the worst-case complexity of the Proxy-Rank algorithm is $\mathcal{O}\left(\frac{p!}{q!\cdot q!\cdot(p-2q-1)!}\right)$.

5. Identifiability with Higher-Order Statistics

In the above section, we have shown that the proxy variables can be identified with the help of rank constraints of the covariance matrix (second-order statistics) under some mild assumptions. However, if Assumption 2 or Assumption 3 is violated, e.g., the dimension of the sets of NCE and NCO are exactly q, the underlying NCE and NCO are not guaranteed to be identified with second-order statistics. To tackle this issue, below we show that we can benefit from higher-order statistics of the noise terms. Then, we provide another data-driven method with higher-order statistics for selecting valid proxy variables for a given causal relationship $X_k \to Y$ and obtaining its unbiased causal effect of X_k on Y.

5.1. Identification of Proxy Variables with Higher-Order Statistics

We assume the model of interest is a linear causal model with non-Gaussian error terms (also known as Linear, Non-Gaussian, Acyclic Model, shortly LiNGAM) (Shimizu et al., 2006). Specifically, the assumption is as follows,

Assumption 4. *The noise terms of variables follow non-Gaussian distributions.*

Assumption 4 states the non-Gaussianity of data, which is expected to be ubiquitous, due to Cramér Decomposition Theorem (Cramér, 1962), as stated in Spirtes & Zhang (2016). Within the framework of this assumption, a significant body of research has already been initiated (Hyvärinen et al., 2010; Wang & Drton, 2020; Salehkaleybar et al., 2020; Zhao et al., 2022). For further reference, we recommend consulting the work of Shimizu (2022).

We next introduce an important constraint, Generalized Independent Noise (GIN) (which is an extension of the familiar Independent Noise (IN) constraint presented in Shimizu et al. (2011)), which is an essential constraint that exploits the non-Gaussianity (higher-order statistics) from data (Xie et al., 2020; Cai et al., 2019; Xie et al., 2023).

Definition 6 (GIN Condition)**.** *Suppose all variables follow the linear non-Gaussian acyclic causal model. Let $\mathcal{Y}$, $\mathcal{Z}$ be two sets of random variables. We say that $(\mathcal{Z}, \mathcal{Y})$ follows the GIN condition if and only if $\omega^\intercal \mathcal{Y} \perp\!\!\!\perp \mathcal{Z}$, where ω satisfies $\omega^\intercal \mathbb{E}(\mathcal{Y}\mathcal{Z}^\intercal) = \mathbf{0}$ and $\omega \neq \mathbf{0}$.*

By Darmois–Skitovich theorem (Darmois, 1953; Skitovitch, 1953)[4], GIN (that is linear transformation $\omega^\intercal \mathcal{Y} \perp\!\!\!\perp \mathcal{Z}$) implies that $\omega^\intercal \mathcal{Y}$ shares no common non-Gaussian exogenous noise components with $\mathcal{Z}$.

A Motivating Example: To illustrate the intuitions behind it, we will begin by providing a straightforward example before presenting the theoretical results. Let's consider the causal relationship $X_k \to Y$ shown in Figure 4. Assume that the data are generated from a linear causal model with non-Gaussian error terms. In the subgraph (a), Z and W are the valid NCE and NCO for the causal relationship $X_k \to Y$. We have that $(\{X_k, Z\}, \{X_k, Y, W\})$ follows the GIN constraint, as explained below. The causal models of latent variables is $U = \varepsilon_U$, and $\{X_k, Y, W\}$ and $\{X_k, Z\}$ can then be represented as

$$\underbrace{\begin{bmatrix} X_k \\ Y \\ W \end{bmatrix}}_{\mathcal{Y}} = \begin{bmatrix} 1 & 0 \\ \beta & c+fd \\ 0 & d \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ f\varepsilon_W + \varepsilon_Y \\ \varepsilon_W \end{bmatrix}}_{\mathcal{E}_\mathcal{Y}},$$

$$\underbrace{\begin{bmatrix} X_k \\ Z \end{bmatrix}}_{\mathcal{Z}} = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \varepsilon_Z \end{bmatrix}}_{\mathcal{E}_\mathcal{Z}}.$$

According to the above equations, $\omega^\intercal \mathbb{E}(\mathcal{Y}\mathcal{Z}^\intercal) = \mathbf{0} \Rightarrow \omega = (d\beta, -d, c+df)^\intercal$. Then we can see $\omega^\intercal \mathcal{Y} = \omega^\intercal \mathcal{E}_\mathcal{Y} = c\varepsilon_W - d\varepsilon_Y$. By Darmois–Skitovich theorem, $\omega^\intercal \mathcal{Y}$ is independent of $\mathcal{Z}$ because there is no common non-Gaussian noise terms between $c\varepsilon_W - d\varepsilon_Y$ and $\mathcal{Z}$ (including noise terms ε_U and ε_Z). That is to say, $(\{X_k, Z\}, \{X_k, Y, W\})$ follows the GIN constraint.

Next, we discuss the subgraph (b), where Z and W are the invalid NCE and NCO, respectively, concerning the causal relationship $X_k \to Y$. $\{X_k, Y, W\}$ and $\{X_k, Z\}$ can then be represented as

$$\underbrace{\begin{bmatrix} X_k \\ Y \\ W \end{bmatrix}}_{\mathcal{Y}} = \begin{bmatrix} 1 & 0 \\ \beta & c+fd+ag \\ 0 & d \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ f\varepsilon_W + a\varepsilon_Z + \varepsilon_Y \\ \varepsilon_W \end{bmatrix}}_{\mathcal{E}_\mathcal{Y}},$$

$$\underbrace{\begin{bmatrix} X_k \\ Z \end{bmatrix}}_{\mathcal{Z}} = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \varepsilon_Z \end{bmatrix}}_{\mathcal{E}_\mathcal{Z}}.$$

We have $\omega^\intercal \mathcal{Y}$ is *dependent* of $\mathcal{Z}$ because there exists common non-Gaussian noise terms ε_Z between $\omega^\intercal \mathcal{Y}$ and $\mathcal{Z}$, no matter $\omega^\intercal \mathbb{E}(\mathcal{Y}\mathcal{Z}^\intercal) = \mathbf{0}$ or not. That is to say, $(\{X_k, Z\}, \{X_k, Y, W\})$ violates the GIN constraint. Assuming the distribution is rank-faithful to the graph, the above facts show that the lack of edge $Z \to Y$, i.e., the variable of NCE does not causally affect the primary outcome,

[4] Assume that V_1 and V_2 are linear combinations of independent noise terms $\varepsilon_i (i = 1, ..., n)$. If V_1 and V_2 are statistically independent, there are no common non-Gaussian noise terms between V_1 and V_2. See Theorem 6 of Appendix. K.

has a testable implication with the help of non-Gaussianity. For further details regarding the above example, please refer to the Appendix. C.

We now demonstrate that if valid NCE and NCO relative to a causal relationship $X_k \to Y$ exist (Assumption 1 holds), we can identify them using GIN constraints.

Lemma 3. *Assume that the input data* $\mathbf{X}$ *and* Y *strictly follow the Equation 2 and the rank-faithfulness holds. If Assumptions 1 and 4 hold, then the underlying NCE and NCO relative to the causal relationship* $X_k \to Y$ *can be identified by using the following rule.*

$\mathcal{R}3$. Let $\mathbf{A}$ *and* $\mathbf{B}$ *be two disjoint subsets of* $\mathbf{X}$*, where* $|\mathbf{A}| = q$ *and* $|\mathbf{B}| = q$*. 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* $\mathbf{A}$ *and* $\mathbf{B}$ *are valid NCE and NCO relative to* $X_k \to Y$ *respectively.*

Note that the result of Lemma 3 does not strictly require adherence to Assumption 3-that is, not all noise variables need to follow a non-Gaussian distribution. For further discussion, please see Appendix D.

Example 5 ($\mathcal{R}3$)**.** *Consider the causal diagram shown in Figure 3. We consider the causal relationship* $X_2 \to Y$*. Assume that the data are generated from a linear non-Gaussian acyclic causal model. Let* $\mathbf{A} = \{X_1\}$*, and* $\mathbf{B} = \{X_6\}$*. We check $\mathcal{R}3$ and obtain that* $(\{X_2, X_1\}, \{X_2, Y, X_6\})$ *follows the GIN constraint, and 2) that* $(X_6, \{X_2, X_1\})$ *follows the GIN constraint. These facts imply that* X_1 *and* X_6 *are valid NCE and NCO relative to* $X_2 \to Y$ *respectively.*

Based on Lemma 3, we present the identifiability of NCE and NCO in terms of GIN constraints.

Theorem 3 (Identifiability of NCE and NCO with GIN Constraints)**.** *Assume that the input data* $\mathbf{X}$ *and* Y *strictly follow the Equation 2 and the rank-faithfulness holds. Furthermore, assume Assumption 4 holds. Then the underlying NCE and NCO relative to the causal relationship* $X_k \to Y$ *can be identified if Assumption 1 holds.*

5.2. Algorithm

In this section, we will leverage the above theoretical results and propose another data-driven method called Proxy-GIN to estimate the total causal effects of treatment $X_k \in \mathbf{X}$ on outcome Y:

Proxy-GIN algorithm

1. Given a p-dimensional treatments $\mathbf{X}$, outcome Y, the number of unmeasured confounders q. Initialize the sets of NCE, and NCO, causal effect, as $\mathcal{NCE}$, $\mathcal{NCO}$, and $\mathcal{C}$, respectively, with an empty set, i.e., $\mathcal{NCE} := \emptyset$, $\mathcal{NCO} := \emptyset$, and $\mathcal{C} := \emptyset$.
2. Find valid NCE and NCO of unmeasured confounder $\mathbf{U}$ relative to per causal relationship $X_k \to Y$ according to Lemmas 3.
3. Estimate the corresponding unbiased causal effect by Proposition 2 given NCE and NCO for per causal relationship $X_k \to Y$. Otherwise, output a value (NA) indicating the lack of valid NCE and NCO for this causal relationship $X_k \to Y$ to obtain the unbiased causal effect.

The specific details of algorithm execution are provided in the Appendix F.

We now show that, in the large sample limit, for the causal relationships of interest, our algorithm outputs the true proxy variables and the unbiased estimation of causal effect.

Theorem 4 (Correctness)**.** *Assume that the input data* $\mathbf{X}$ *and* Y *strictly follow the Equation 2 and the rank-faithfulness holds. Furthermore, assume that Assumption 4 holds. Given infinite samples, the Proxy-GIN algorithm outputs the true causal effect* $\mathcal{C}$ *correctly.*

We finally analyze the complexity of the Proxy-GIN algorithm. Let q be the number of latent confounders, and p be the number of treatments. The dominant part is to check $\mathcal{R}_3$ of Lemma 3 with worst-case complexity is $\mathcal{O}\left(\frac{p!}{(q+1)!\cdot(q+1)!\cdot(p-2q-3)!}\right)$. Hence, the worst case complexity of the Proxy-GIN algorithm is $\mathcal{O}\left(\frac{p!}{(q+1)!\cdot(q+1)!\cdot(p-2q-3)!}\right)$.

6. Experimental Results on Synthetic Data

In this section, we evaluate the performance of the proposed methods in estimating causal effects from synthetic data. We here consider the following two typical settings: **Gaussian case:** The data are generated according to the causal graph in Figure 3, with the noise terms being generated from standard normal distributions; **Non-Gaussian case**: The data are generated according to the graph obtained by removing variable X_3 from Figure 3, with the noise terms being generated from standard exponential distributions. In three cases, the connected coefficient β_k is sampled from a uniform distribution between $[-1, 1]$. Note that in the *Gaussian case*, either Assumption 2 or 3 holds for the causal relationships $X_2 \to Y$, $X_5 \to Y$, and $X_6 \to Y$. However, in the *non-Gaussian case*, both Assumption 2 and 3 are violated for the causal relationships $X_2 \to Y$ and $X_5 \to Y$, but Assumption 1 holds. As a result, we will focus on these three causal relationships in this context.

The methods we compare against are: 1) NAIVE, the least-squares regression coefficient of Y on $X_k \in \mathbf{X}$; 2) FindNC, the algorithm 1 of in (Kummerfeld et al., 2022) + standard confounder proxy estimator; 3) Proxy-Rank, our method, using T. W. Anderson's canonical correlation-based rank

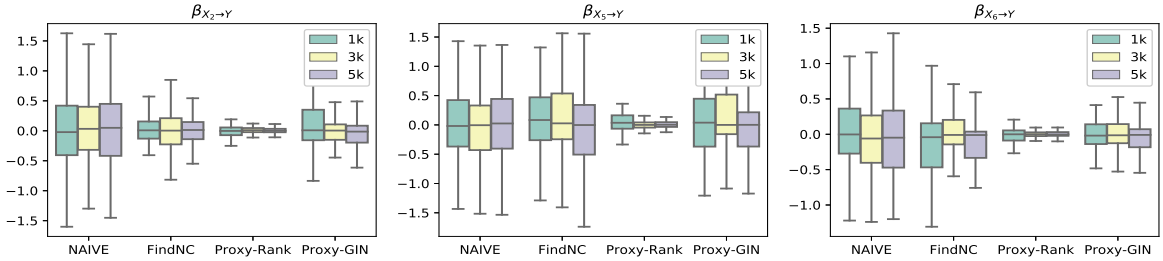


Figure 5. Performance of NAIVE, FindNC, Proxy-Rank, and Proxy-GIN on the Gaussian case.

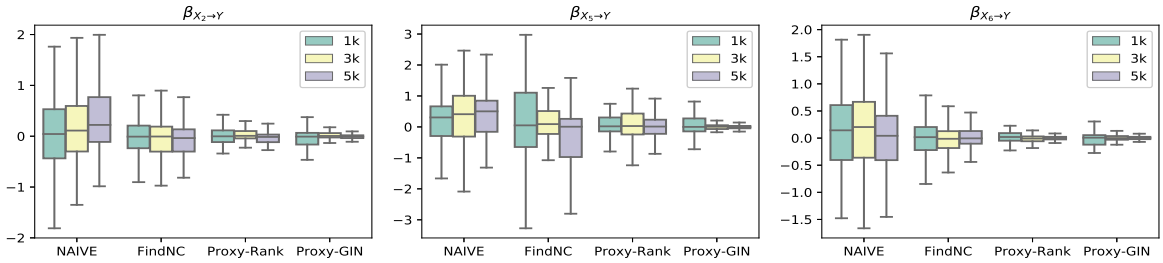


Figure 6. Performance of NAIVE, FindNC, Proxy-Rank, and Proxy-GIN on the Non-Gaussian case.

test (Anderson, 1984) to evaluate the Rank constraint, and 4) Proxy-GIN, our method, using HSIC-based independence test (Zhang et al., 2018) to evaluate the GIN constraint, due to the non-Gaussianity of the data. For the sake of comparison, if the algorithm fails to find the valid NCE and NCO for a causal relationship, we will randomly select variables as NCE and NCO to estimate the causal effect of interest. Each experiment was repeated 100 times with randomly generated data and the results were averaged. The sample size is selected from $\{1,000(1k), 3,000(3k), 5,000(5k)\}$. The source code is in the Supplementary file.

Figures 5 $\sim$ 6 summarize the bias of the estimators of each parameter. As expected, our proposed Proxy-Rank algorithm almost outperforms other methods (with little bias for all causal effects) in all two cases (except for $\beta_{X_2 \to Y}$ and $\beta_{X_5 \to Y}$ in the Non-Gaussian case), with all sample sizes. The reason that $\beta_{X_2 \to Y}$ and $\beta_{X_5 \to Y}$ cannot be consistently estimated by the Proxy-Rank algorithm is that, in the Non-Gaussian case, there are no valid NCE and NCO for the causal relationship $X_2 \to Y$ and $X_5 \to Y$ in the ground-truth graph. The Proxy-GIN algorithm outperforms the NAIVE algorithm and the FindNC algorithm in non-Gaussian cases, with all sample sizes, which verifies the correctness of the Proxy-GIN algorithm. The the NAIVE algorithm and FindNC algorithm are expected to perform poorly, since there exists unmeasured confounder U and the FindNC algorithm needs three Quadruple-disconnected NCs for per causal relationships. More experimental results are provided in Appendix I.

7. Experimental Results on Real-world Data

In this section, we apply the proposed methods 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). Additionally, it includes body weight as the outcome variable and data collected from 227 mice. As discussed in Miao et al. (2022), gene expression studies like this one may encounter unmeasured confounding issues stemming from batch effects or unobserved phenotypes.

Following the analysis conducted by Miao et al., we assume that there is only one latent variable underlying the common influence, and the data generation mechanism adheres to a linear causal model. We observed that the majority of our findings align with those presented by Miao et al. (2022). For instance, 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. Detailed results and analysis are included in Appendix J.

8. Conclusion

This paper focuses on the identifiability conditions for selecting proxy variables for unmeasured confounders in observational data. Initially, we introduce an extended proxy variable estimator to handle multiple unmeasured confounders between treatments and outcomes. Subsequently, we provide two specific identifiability conditions based on second-order and higher-order statistics. Additionally, the paper proposes two efficient algorithms for selecting proxy variables, utilizing Rank-deficiency and GIN properties, with their effectiveness substantiated by experimental results.

Acknowledgements

FX would like to acknowledge the support by the Natural Science Foundation of China (62306019) and the Disciplinary funding of Beijing Technology and Business University (STKY202302). FX would like to thank Kun Zhang for his valuable insights and discussions during the initial phase of the research. ZM's research was supported by the China Scholarship Council (CSC). We appreciate the comments from anonymous reviewers, which greatly helped to improve the paper.

Impact Statement

Learning causal effects is essential throughout the data-driven sciences and has attracted much attention. Our research focuses on estimating causal effects where proxy variables exist to help adjust for unmeasured confounders in fields like social sciences, economics, public health, and neuroscience. We assess the impact of our work in the context of these fields. However, the applicability of existing methods is often limited in practice, as the validity of proxy variables typically relies on background knowledge. Notable merits of our work include providing a practical method for selecting proxy variables and estimating unbiased causal effects from purely observational data.

References

Anderson, T. W. *An Introduction to Multivariate Statistical Analysis. 2nd ed.* John Wiley & Sons, 1984.

Bai, J. and Ng, S. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Baiocchi, M., Cheng, J., and Small, D. S. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33:2297–2340, 2014.

Bollen, K. A. Structural equations with latent variables wiley. *New York*, 1989.

Bowden, R. J. and Turkington, D. A. *Instrumental variables*. Number 8. Cambridge university press, 1990.

Cai, R., Xie, F., Glymour, C., Hao, Z., and Zhang, K. Triad constraints for learning causal structure of latent variables. In *Advances in Neural Information Processing Systems*, pp. 12863–12872, 2019.

Cheng, D., Li, J., Liu, L., Yu, K., Le, T. D., and Liu, J. Toward unique and unbiased causal effect estimation from data with hidden variables. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Cheng, D., Li, J., Liu, L., Yu, K., Le, T. D., and Liu, J. Discovering ancestral instrumental variables for causal inference from observational data. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pp. 294–321, 2012.

Cramér, H. *Random variables and probability distributions*. Cambridge University Press, Cambridge, 2nd edition, 1962.

D'Amour, A. Comment: Reflections on the deconfounder. *Journal of the American Statistical Association*, 114: 1597–1601, 2019b.

Darmois, G. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pp. 2–8, 1953.

Dawid, A. P. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

de Luna, X., Fowler, P., and Johansson, P. Proxy variables and nonparametric identification of causal effects. *Economics Letters*, 150:152–154, 2017.

Draisma, J., Sullivant, S., and Talaska, K. Positivity for gaussian graphical models. *Advances in Applied Mathematics*, 50(5):661–674, 2013.

Drton, M., Robeva, E., and Weihs, L. Nested covariance determinants and restricted trek separation in gaussian graphical models. *Bernoulli*, 26(4), 2020.

Fernández-Loría, C. and Provost, F. Causal classification: Treatment effect estimation vs. outcome prediction. *The Journal of Machine Learning Research*, 23(1):2573–2607, 2022.

Goldberger, A. S. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pp. 979–1001, 1972.

Gunsilius, F. F. Nontestability of instrument validity under continuous treatments. *Biometrika*, 108(4):989–995, 2021.

Hernán, M. A. and Robins, J. M. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006a.

Hernán, M. A. and Robins, J. M. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17 (4):360–372, 2006b.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pp. 689–696, 2009.

Huang, B., Low, C. J. H., Xie, F., Glymour, C., and Zhang, K. Latent hierarchical causal structure discovery with rank constraints. *Advances in Neural Information Processing Systems*, 35:5549–5561, 2022.

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Imbens, G. W. Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3):323–358, 2014.

Imbens, G. W. and Rubin, D. B. Causal inference for statistics, social, and biomedical sciences: An introduction. *Cambridge University Press*, 2015.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American statistical Association*, 111:132–144, 2016.

Kummerfeld, E. and Ramsey, J. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1655–1664, 2016.

Kummerfeld, E., Lim, J., and Shi, X. Data-driven automated negative control estimation (dance): Search for, validation of, and causal inference with negative controls. *arXiv preprint arXiv:2210.00528*, 2022.

Kuroki, M. and Cai, Z. Instrumental variable tests for directed acyclic graph models. In *International Workshop on Artificial Intelligence and Statistics*, pp. 190–197. PMLR, 2005.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101:423–437, 2014.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lin, W., Feng, R., and Li, H. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288, 2015.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383, 2010.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, pp. 7512–7523. PMLR, 2021.

Miao, W., Ding, P., and Geng, Z. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111:1673–1683, 2016.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105:987–993., 2018a.

Miao, W., Shi, X., and Tchetgen, E. T. A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*, 2018b.

Miao, W., Hu, W., Ogburn, E. L., and Zhou, X.-H. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, pp. 1–15, 2022.

Ogburn, E. L., Shpitser, I., and Tchetgen, E. J. T. Comment on "Blessings of multiple causes". *Journal of the American Statistical Association*, 114:1611–1615, 2019.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

Pearl, J. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 435–443, 1995.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

Peters, J., Mooij, J. M., D., J., and Schölkopf, B. Minimal nonlinear distortion principle for nonlinear independent component analysis. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference*. MIT Press, 2017.

Rotnitzky, A. and Smucler, E. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *The Journal of Machine Learning Research*, 21(1):7642–7727, 2020.

Salehkaleybar, S., Ghassami, A., Kiyavash, N., and Zhang, K. Learning linear non-gaussian causal models in the presence of latent variables. *The Journal of Machine Learning Research*, 21(1):1436–1459, 2020.

Schneeberger, M. Irx3, a new leader on obesity genetics. *EBioMedicine*, 39:19–20, 2019.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540, 2020a.

Shi, X., Miao, W., and Tchetgen, E. T. A selective review of negative control methods in epidemiology. *Current epidemiology reports*, 7:190–202, 2020b.

Shimizu, S. *Statistical Causal Discovery: LiNGAM Approach*. Springer, 2022.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct): 2003–2030, 2006.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.

Shpitser, I., Wood-Doughty, Z., and Tchetgen, E. J. T. The proximal id algorithm. *Journal of Machine Learning Research*, 23:1–46, 2023.

Silva, R. and Shimizu, S. Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49, 2017.

Singh, R. Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. *arXiv preprint arXiv:2012.10315*, 2020.

Skitovitch, V. P. On a property of the normal distribution. *DAN SSSR*, 89:217–219, 1953.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen, E. J. T. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(3):348, 2016.

Spearman, C. Pearson's contribution to the theory of two factors. *British Journal of Psychology. General Section*, 19(1):95–101, 1928.

Spirtes, P. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5), 2010.

Spirtes, P. Calculation of entailed rank constraints in partially non-linear and cyclic models. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 606–615. AUAI Press, 2013.

Spirtes, P. and Zhang, K. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):1–28, 2016.

Spirtes, P., Meek, C., and Richardson, T. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (UAI)*, pp. 499–506. Morgan Kaufmann Publishers Inc., 1995.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2000.

Sullivant, S., Talaska, K., and Draisma, J. Trek separation for gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Van Der Zander, B., Liśkiewicz, M., and Textor, J. Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence*, 270:1–40, 2019.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. Confounder adjustment in multiple hypothesis testing. *Annals of statistics*, 45(5):1863, 2017.

Wang, S., Yehya, N., Schadt, E. E., Wang, H., Drake, T. A., and Lusis, A. J. Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genetics*, 2:e15, 2006.

Wang, Y. and Blei, D. A proxy variable view of shared confounding. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2021.

Wang, Y. and Blei, D. M. The Blessings of Multiple Causes. *Journal of the American Statistical Association*, 114: 1574–1596, 2019.

Wang, Y. S. and Drton, M. High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1):41–59, 2020.

Wheatcroft, S. B., Kearney, M. T., Shah, A. M., Ezzat, V. A., Miell, J. R., Modo, M., Williams, S. C., Cawthorn, W. P., Medina-Gomez, G., Vidal-Puig, A., Sethi, J. K., and Crossey, P. A. Igf-binding protein-2 protects against the

development of obesity and insulin resistance. *Diabetes*, 56:285–294, 2007.

Wright, P. G. *Tariff on Animal and Vegetable Oils*. Macmillan, New York, 1928.

Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., and Zhang, K. Generalized independent noise conditionfor estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems*, pp. 14891–14902, 2020.

Xie, F., He, Y., Geng, Z., Chen, Z., Hou, R., and Zhang, K. Testability of instrumental variables in linear non-gaussian acyclic causal models. *Entropy*, 24(4):512, 2022a.

Xie, F., Huang, B., Chen, Z., He, Y., Geng, Z., and Zhang, K. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pp. 24370–24387. PMLR, 2022b.

Xie, F., Huang, B., Chen, Z., Cai, R., Glymour, C., Geng, Z., and Zhang, K. Generalized independent noise condition for estimating causal structure with latent variables. *arXiv preprint arXiv:2308.06718*, 2023.

Xu, L., Kanagawa, H., and Gretton, A. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems*, 34:26264–26275, 2021.

Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896, 2008.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *UAI*, pp. 647–655. AUAI Press, 2009.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

Zhao, R., He, X., and Wang, J. Learning linear non-gaussian directed acyclic graph with diverging number of nodes. *The Journal of Machine Learning Research*, 23(1):12314–12347, 2022.

A. More Details on Extended Proxy Variables Estimator

We provide simulation results to demonstrate that when the dimension of NCE and NCO is lower than the number of unmeasured confounders, the estimated causal effect becomes biased when using a standard proxy variable estimator (Proposition 1). In particular, we examine the causal relationship $X_k \to Y$ depicted in Figure 7. We consider different quantities of unmeasured confounders, denoted as $q = 1, 2, 3, 4$. We employ the following estimators:

- The *Traditional Proxy Variables Estimator* described in Proposition 1, which is given by:

$$\beta_{X_k \to Y} = \frac{\sigma_{X_k Y}\sigma_{W_1 Z_1} - \sigma_{X_k W_1}\sigma_{Y Z_1}}{\sigma_{X_k X_k}\sigma_{W_1 Z_1} - \sigma_{X_k W_1}\sigma_{X_k Z_1}}. \tag{6}$$

- The *Extended Proxy Variables Estimator* described in Proposition 2, which is expressed as:

$$\beta_{X_k \to Y} = \frac{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{Z}\},\{Y \cup \mathbf{W}\}})}{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{Z}\},\{X_k \cup \mathbf{W}\}})}, \tag{7}$$

  where $\mathbf{Z} = \{Z_1, \ldots, Z_q\}$ and $\mathbf{W} = \{W_1, \ldots, W_q\}$.

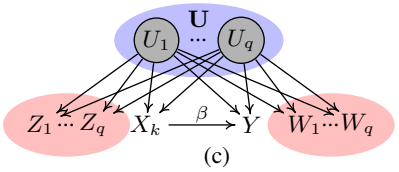


Figure 7. Causal Diagram used in our simulation studies, where $q = 1, 2, 3, 4$.

Figures 8 summarizes the bias of the estimators of each parameter. As expected, our proposed *Extended Proxy Variables Estimator* outperforms *Traditional Proxy Variables Estimator* (with little bias for the causal effect of X_k on Y) when the number of unmeasured confounders is greater or equal to 2.

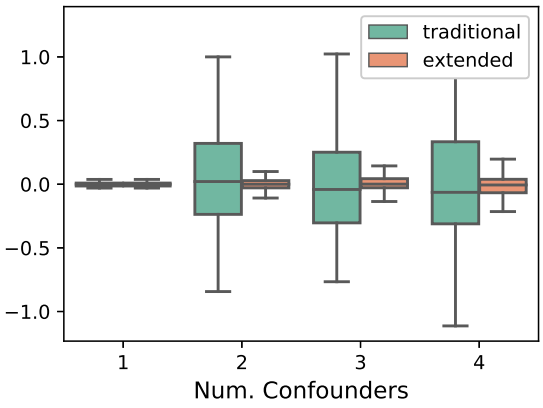


Figure 8. Performance of extended estimator and traditional estimator with varying numbers of unmeasured confounders.

B. More Details on the Motivating Example for Second-Order Statistics (in Section 4.1)

In this section, we will give the details of the motivating example described in Section 4.1.

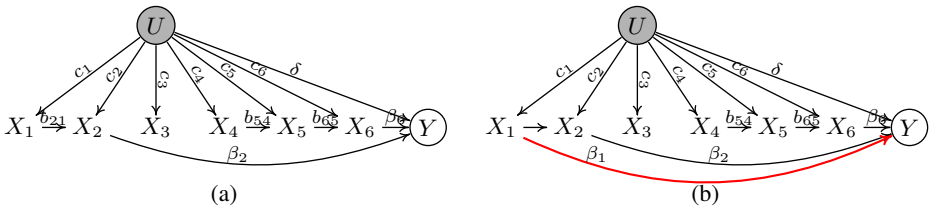


Figure 9. A simple causal graph involving 6 potential treatments and one outcome.

We assume that all random variables have mean zero for simplicity.

Structure (a): Since the variables strictly follow the linear causal model, we obtain

$$
\begin{aligned}
U &= \varepsilon_U, \\
X_1 &= c_1 U + \varepsilon_{X_1} = c_1 \varepsilon_U + \varepsilon_{X_1}, \\
X_2 &= c_2 U + b_{21} X_1 + \varepsilon_{X_2} = (c_1 b_{21} + c_2)\varepsilon_U + b_{21}\varepsilon_{X_1} + \varepsilon_{X_2}, \\
X_3 &= c_3 U + \varepsilon_{X_3} = c_3 \varepsilon_U + \varepsilon_{X_3}, \\
X_4 &= c_4 U + \varepsilon_{X_4} = c_4 \varepsilon_U + \varepsilon_{X_4}, \\
X_5 &= c_5 U + b_{54} X_4 + \varepsilon_{X_5} = (c_5 + c_4 b_{54})\varepsilon_U + b_{54}\varepsilon_{X_4} + \varepsilon_{X_5}, \\
X_6 &= c_6 U + b_{65} X_5 + \varepsilon_{X_6} \\
&= (c_6 + b_{65}c_5 + b_{65}c_4 b_{54})\varepsilon_U + b_{65}b_{54}\varepsilon_{X_4} + b_{65}\varepsilon_{X_5} + \varepsilon_{X_6}, \qquad (8) \\
Y &= \delta U + \beta_6 X_6 + \beta_2 X_2 + \varepsilon_Y \\
&= \left[\delta + \beta_6(c_6 + b_{65}c_5 + b_{65}c_4 b_{54}) + \beta_2 c_1 b_{21}\right]\varepsilon_U + \beta_2 b_{21}\varepsilon_{X_1} + \beta_2\varepsilon_{X_2} \\
&\quad + \beta_6 b_{65} b_{54}\varepsilon_{X_4} + \beta_6 b_{65}\varepsilon_{X_5} + \beta_6\varepsilon_{X_6} + \varepsilon_Y. \qquad (9)
\end{aligned}
$$

We now consider the causal relationship $X_2 \to Y$ in Figure 9(a). X_1 and X_6 are the valid NCE and NCO for the causal relationship $X_2 \to Y$, respectively. We have the vanishing determinants on the cross-covariance matrix $\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}$, i.e.,

$$
\det(\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}) = \det\left(\begin{bmatrix} \sigma_{X_2X_2} & \sigma_{X_2Y} & \sigma_{X_2X_6} \\ \sigma_{X_3X_2} & \sigma_{X_3Y} & \sigma_{X_3X_6} \\ \sigma_{X_1X_2} & \sigma_{X_1Y} & \sigma_{X_1X_6} \end{bmatrix}\right) = 0 \qquad (10)
$$

By Eq.10, we quickly know that $\mathrm{rk}(\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}) \leq 2$. We next add an edge $X_1 \to Y$ to Figure 9(a) such that X_1 and X_6 are the invalid NCE and NCO relative to the causal relationship $X_2 \to Y$ (as shown in Figure 9(b)). Now, the vanishing determinant on the cross-covariance matrix $\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}$ will fail, i.e.,

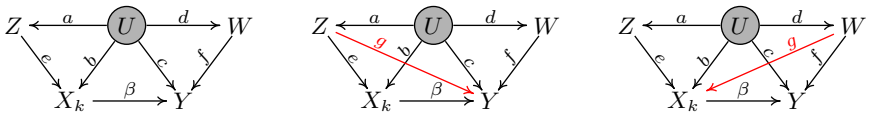
$$
\begin{aligned}
\det(\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}) &= \det\left(\begin{bmatrix} \sigma_{X_2X_2} & \sigma_{X_2Y} & \sigma_{X_2X_6} \\ \sigma_{X_3X_2} & \sigma_{X_3Y} & \sigma_{X_3X_6} \\ \sigma_{X_1X_2} & \sigma_{X_1Y} & \sigma_{X_1X_6} \end{bmatrix}\right) \\
&= \sigma_U^2 \sigma_{X_1}^2 \sigma_{X_2}^2 c_3 \beta_1 (c_4 b_{54} b_{65} + c_5 b_{65} + c_6) \neq 0 \qquad (11)
\end{aligned}
$$

That is to say, $\mathbf{\Sigma}_{\{X_2,X_3,X_1\},\{X_2,Y,X_6\}}$ is full rank. Assuming the distribution is rank-faithful to the graph, the above facts show that lack of edge $X_1 \to Y$, i.e., the variable of NCE does not causally affect the primary outcome, has a testable implication.

C. More Details on the Motivating Example for Higher-Order Statistics (in Section 5.1)

In this section, we will give the details of another motivating example described in Section 5.1.

We assume that all random variables have mean zero for simplicity. We consider the linear causal model with non-Gaussian noise terms.



(a) (b) (c)

Figure 10. A linear causal model with any of the graphical structures above entails the same rank constraints in the marginal covariance matrix of $\{X_k, Y, Z, W\}$, but it entails different GIN constraints.

Structure (a): Since the variables strictly follow the linear causal model, we obtain

$$\begin{aligned} U &= \varepsilon_U, \\ Z &= aU + \varepsilon_Z = a\varepsilon_U + \varepsilon_Z, \\ W &= dU + \varepsilon_W = d\varepsilon_U + \varepsilon_W, \\ X_k &= bU + eZ + \varepsilon_{X_k} = (ae+b)\varepsilon_U + e\varepsilon_Z + \varepsilon_{X_k}, \\ Y &= cU + \beta X_k + fW + \varepsilon_Y \\ &= (ae+b+c+df)\varepsilon_U + e\beta\varepsilon_Z + \beta\varepsilon_{X_k} + f\varepsilon_W + \varepsilon_{X_k}. \end{aligned} \tag{12}$$

1). By the above equations, $\{X_k, Y, W\}$ and $\{X_k, Z\}$ can then be represented as

$$\underbrace{\begin{bmatrix} X_k \\ Y \\ W \end{bmatrix}}_{\mathcal{Y}} = \begin{bmatrix} 1 & 0 \\ \beta & c+fd \\ 0 & d \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ f\varepsilon_W + \varepsilon_Y \\ \varepsilon_W \end{bmatrix}}_{\mathcal{E}_\mathcal{Y}}, \tag{13}$$

$$\underbrace{\begin{bmatrix} X_k \\ Z \end{bmatrix}}_{\mathcal{Z}} = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \varepsilon_Z \end{bmatrix}}_{\mathcal{E}_\mathcal{Z}}, \tag{14}$$

According to the above equations, $\omega^\intercal \mathbb{E}[\mathcal{Y}\mathcal{Z}^\intercal] = \mathbf{0} \Rightarrow \omega = [d\beta, -d, c+df]^\intercal$. Then we can see $\omega^\intercal \mathcal{Y} = \omega^\intercal \mathcal{E}_\mathcal{Y} = c\varepsilon_W - d\varepsilon_Y$. By Darmois–Skitovich theorem, $\omega^\intercal \mathcal{Y}$ is independent of $\mathcal{Z}$ because there is no common non-Gaussian noise terms between $c\varepsilon_W - d\varepsilon_Y$ and $\mathcal{Z}$ (including ε_U and ε_Z). That is to say, $(\{X_k, Z\}, \{X_k, Y, W\})$ *follows the GIN constraint*.

2). $\{X_k, Z\}$ and W can then be represented as

$$\underbrace{\begin{bmatrix} X_k \\ Z \end{bmatrix}}_{\mathcal{Y}} = \begin{bmatrix} ae+b \\ a \end{bmatrix} [U] + \underbrace{\begin{bmatrix} e\varepsilon_Z + \varepsilon_{X_k} \\ \varepsilon_Z \end{bmatrix}}_{\mathcal{E}_\mathcal{Y}}, \tag{15}$$

$$\underbrace{[W]}_{\mathcal{Z}} = [d]\,[U] + \underbrace{[\varepsilon_W]}_{\mathcal{E}_\mathcal{Z}}, \tag{16}$$

According to the above equations, $\omega^\intercal \mathbb{E}[\mathcal{Y}\mathcal{Z}^\intercal] = \mathbf{0} \Rightarrow \omega = [-a, e+b]^\intercal$. Then we can see $\omega^\intercal \mathcal{Y} = \omega^\intercal \mathcal{E}_\mathcal{Y} = b\varepsilon_Z - a\varepsilon_{X_k}$. By Darmois–Skitovich theorem, $\omega^\intercal \mathcal{Y}$ is independent of $[W]$ because there is no common non-Gaussian noise terms between $b\varepsilon_Z - a\varepsilon_{X_k}$ and $[W]$. That is to say, $(\{W\}, \{X_k, Z\})$ *follows the GIN constraint*.

Structure (b): The data generation process is as follows:

$$\begin{aligned} U &= \varepsilon_U, \\ Z &= aU + \varepsilon_Z = a\varepsilon_U + \varepsilon_Z, \\ W &= dU + \varepsilon_W = d\varepsilon_U + \varepsilon_W, \\ X_k &= bU + eZ + \varepsilon_{X_k} = (ae+b)\varepsilon_U + e\varepsilon_Z + \varepsilon_{X_k}, \\ Y &= cU + \beta X_k + fW + gZ + \varepsilon_Y \\ &= (ae+b+c+df+ag)\varepsilon_U + (e\beta+g)\varepsilon_Z + \beta\varepsilon_{X_k} + f\varepsilon_W + \varepsilon_{X_k}. \end{aligned} \tag{17}$$

1). By the above equations, $\{X_k, Y, W\}$ and $\{X_k, Z\}$ can then be represented as

$$\underbrace{\begin{bmatrix} X_k \\ Y \\ W \end{bmatrix}}_{\mathcal{Y}} = \begin{bmatrix} 1 & 0 \\ \beta & c+fd+ag \\ 0 & d \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ f\varepsilon_W + a\varepsilon_Z + \varepsilon_Y \\ \varepsilon_W \end{bmatrix}}_{\mathcal{E}_\mathcal{Y}}, \tag{18}$$

$$\underbrace{\begin{bmatrix} X_k \\ Z \end{bmatrix}}_{\mathcal{Z}} = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \varepsilon_Z \end{bmatrix}}_{\mathcal{E}_\mathcal{Z}}, \tag{19}$$

We have $\omega^\intercal \mathcal{Y}$ is dependent of $[X_k, Z]$ because there exists common non-Gaussian noise terms ε_Z between $\mathcal{Y}$ and $\mathcal{Z}$, no matter $\omega^\intercal \mathbb{E}[\mathcal{Y}\mathcal{Z}^\intercal] = \mathbf{0}$ or not. That is to say, $(\{X_k, Z\}, \{X_k, Y, W\})$ *violates the GIN constraint*.

2). $\{X_k, Z\}$ and W can then be represented as

$$\underbrace{\begin{bmatrix} X_k \\ Z \end{bmatrix}}_{\mathcal{Y}} = \begin{bmatrix} ae+b \\ a \end{bmatrix} \begin{bmatrix} U \end{bmatrix} + \underbrace{\begin{bmatrix} e\varepsilon_Z + \varepsilon_{X_k} \\ \varepsilon_Z \end{bmatrix}}_{\mathcal{E}_\mathcal{Y}}, \tag{20}$$

$$\underbrace{\begin{bmatrix} W \end{bmatrix}}_{\mathcal{Z}} = \begin{bmatrix} d \end{bmatrix} \begin{bmatrix} U \end{bmatrix} + \underbrace{\begin{bmatrix} \varepsilon_W \end{bmatrix}}_{\mathcal{E}_\mathcal{Z}}, \tag{21}$$

According to the above equations, $\omega^\intercal \mathbb{E}[\mathcal{Y}\mathcal{Z}^\intercal] = \mathbf{0} \Rightarrow \omega = [-a, e+b]^\intercal$. Then we can see $\omega^\intercal \mathcal{Y} = \omega^\intercal \mathcal{E}_\mathcal{Y} = b\varepsilon_Z - a\varepsilon_{X_k}$. By Darmois–Skitovich theorem, $\omega^\intercal \mathcal{Y}$ is independent of $[W]$ because there is no common non-Gaussian noise terms between $b\varepsilon_Z - a\varepsilon_{X_k}$ and $[W]$. That is to say, $(\{W\}, \{X_k, Z\})$ *follows the GIN constraint*.

Structure (c): The data generation process is as follows:

$$\begin{aligned} U &= \varepsilon_U, \\ Z &= aU + \varepsilon_Z = a\varepsilon_U + \varepsilon_Z, \\ W &= dU + \varepsilon_W = d\varepsilon_U + \varepsilon_W, \\ X_k &= bU + eZ + gW + \varepsilon_{X_k} = (ae+b+gd)\varepsilon_U + e\varepsilon_Z + g\varepsilon_W + \varepsilon_{X_k}, \\ Y &= cU + \beta X_k + fW + \varepsilon_Y \\ &= (ae+b+gd+c+df)\varepsilon_U + e\beta\varepsilon_Z + \beta\varepsilon_{X_k} + (\beta g + f)\varepsilon_W + \varepsilon_{X_k}. \end{aligned} \tag{22}$$

1). By the above equations, $\{X_k, Y, W\}$ and $\{X_k, Z\}$ can then be represented as

$$\underbrace{\begin{bmatrix} X_k \\ Y \\ W \end{bmatrix}}_{\mathcal{Y}} = \begin{bmatrix} 1 & 0 \\ \beta & c+fd \\ 0 & d \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ f\varepsilon_W + \varepsilon_Y \\ \varepsilon_W \end{bmatrix}}_{\mathcal{E}_\mathcal{Y}}, \tag{23}$$

$$\underbrace{\begin{bmatrix} X_k \\ Z \end{bmatrix}}_{\mathcal{Z}} = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} X_k \\ U \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \varepsilon_Z \end{bmatrix}}_{\mathcal{E}_\mathcal{Z}}, \tag{24}$$

According to the above equations, $\omega^\intercal \mathbb{E}[\mathcal{Y}\mathcal{Z}^\intercal] = \mathbf{0} \Rightarrow \omega = [d\beta, -d, c+df]^\intercal$. Then we can see $\omega^\intercal \mathcal{Y} = \omega^\intercal \mathcal{E}_\mathcal{Z} = c\varepsilon_W - d\varepsilon_Y$. By Darmois–Skitovich theorem, $\omega^\intercal \mathcal{Y}$ is independent of $\mathcal{Z}$ because there is no common non-Gaussian noise terms between $c\varepsilon_W - d\varepsilon_Y$ and $\mathcal{Z}$. That is to say, $(\{X_k, Z\}, \{X_k, Y, W\})$ *follows the GIN constraint* (This result is the same as the result in Structure (a)).

2). However, we have that $(\{W\}, \{X_k, Z\})$ *violates the GIN constraint*, as explained below. $\{X_k, Z\}$ and W can then be represented as

$$\underbrace{\begin{bmatrix} X_k \\ Z \end{bmatrix}}_{\mathcal{Y}} = \begin{bmatrix} ae+b+gd \\ a \end{bmatrix} [U] + \underbrace{\begin{bmatrix} e\varepsilon_Z + g\varepsilon_W + \varepsilon_{X_k} \\ \varepsilon_Z \end{bmatrix}}_{\mathcal{E}_\mathcal{Y}}, \tag{25}$$

$$\underbrace{[W]}_{\mathcal{Z}} = [d]\,[U] + \underbrace{[\varepsilon_W]}_{\mathcal{E}_\mathcal{Z}}, \tag{26}$$

We have $\omega^\intercal \mathcal{Y}$ is dependent of $\mathcal{Z}$ because there exists common non-Gaussian noise terms ε_W between $\mathcal{Y}$ and $\mathcal{Z}$, no matter $\omega^\intercal \mathbb{E}[\mathcal{Y}\mathcal{Z}^\intercal] = \mathbf{0}$ or not.

Conclusion: Assuming the distribution is rank-faithful to the graph, the above facts show that lack of edges $Z \to Y$ (i.e., the variable of NCE does not causally affect the primary outcome) or $W \to X_k$ (i.e., the variable of NCO does not causally affect the primary treatment) has a testable implication.

D. More Details about Depending on Assumption 4 in Lemma 3

We here would like to mention that the result of Lemma 3 does not strictly require adherence to Assumption 3—that is, not all noise variables need to follow a non-Gaussian distribution. For instance, consider the causal graphs shown in Figure 4, the identification of valid NCO and NCE in subfigure (a) solely depends on the non-Gaussian distribution of the noise components associated with variables Z and W, making them valid, whereas the other subfigures demonstrate invalid cases. Specifically, as elaborated in Appendix C, for subfigure (b), we find that $\omega^\intercal \mathcal{Y}$ is dependent on $[X_k, Z]$ due to the presence of common non-Gaussian noise terms ε_Z between $\mathcal{Y}$ and $\mathcal{Z}$. This dependence aligns with the Darmois-Skitovitch theorem, which necessitates that ε_Z must be non-Gaussian. Similarly, for subfigure (c), we observe that $\omega^\intercal \mathcal{Y}$ is dependent on $\mathcal{Z}$ due to the presence of common non-Gaussian noise terms ε_W between $\mathcal{Y}$ and $\mathcal{Z}$, again requiring ε_W to be non-Gaussian as per the Darmois-Skitovitch theorem.

E. More Details on Proxy-Rank Algorithm (in Section 4.2)

The specific details of the Proxy-Rank algorithm are provided in the following,

F. More Details on Proxy-GIN Algorithm (in Section 5.2)

The specific details of the Proxy-GIN algorithm are provided in the following,

G. Discussion on the Consistency Result and Convergence Rate of Theorem 2

G.1. Discussion on the Consistency Result

General description. Theorem 2 shows that the true causal effect obtained by the Proxy-Rank is correct in the sense that the proxy variables relative to the causal relationship of interest are valid and the causal effect in the output $\mathcal{C}$ is the true value. In fact, the consistency results of our estimation depend on two processes: first, appropriately selecting proximal variables; next, based on the selected proxy variables, the nonparametric estimator $\hat{\beta}_{X_k \to Y}$ is obtained using Eq. (3). For Theorem 2, the formal statement can be expressed as follows: let $\text{cond1} = (\text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k, Q, \mathbf{A}_1\}, \{X_k, Y, \mathbf{B}_1\}}) \leq q+1, \text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k, \mathbf{A}_1\}, \{Q, \mathbf{B}_1\}} \leq q)$, where $|\mathbf{A}_1| = q$ and $|\mathbf{B}_1| = q$ (Lemma 1), and $\text{cond2} = (\text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k, \mathbf{A}_2\}, \{X_k, Y, \mathbf{B}_2\}}) \leq q+1, \text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k, \mathbf{A}_2\}, \{\mathbf{B}_2\}})) \leq q)$, where $|\mathbf{A}_2| = q+1$ and $|\mathbf{B}_2| = q+1$ (Lemma 2).

$$\lim_{n \to \infty} \text{pr}(|\hat{\beta}_{X_k \to Y} - \beta_{X_k \to Y}| > \epsilon \mid \text{cond1 } \textit{or } \text{cond2}) = 0 \tag{27}$$

for all $\epsilon > 0$, which implies that the estimated causal effect $\hat{\beta}_{X_k \to Y}$ obtained by the Proxy-Rank algorithm is consistent.

Proof details. Before presenting the detailed proof, we first introduce two lemmas that will aid in our demonstration. Let

$$\Delta_1 = \text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k, \mathbf{A}\}, \mathbf{B}}) - \text{rk}(\mathbf{\Sigma}_{\{X_k, \mathbf{A}\}, \mathbf{B}}),\ \Delta_2 = \text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k, \mathbf{A}\}, \{X_k, Y, \mathbf{B}\}}) - \text{rk}(\mathbf{\Sigma}_{\{X_k, \mathbf{A}\}, \{X_k, Y, \mathbf{B}\}}),$$

Algorithm 1 Proxy-Rank

Input: A dataset of treatments $\mathbf{X} = \{X_1, ..., X_p\}$, outcome Y, and the number of unmeasured confounders q.

1: Initialize sets $\mathcal{C} = \emptyset$, $\mathcal{NCE} = \emptyset$, and $\mathcal{NCO} = \emptyset$
2: **for** every variable X_k in $\mathbf{X}$ **do**
3: **repeat**
4: Select two subsets $\mathbf{A}$, $\mathbf{B}$ and one variable Q from $\mathbf{X} \setminus X_k$ such that $\mathbf{A} \cap \mathbf{B} \cap \{Q\} = \emptyset$, and that $|\mathbf{A}| = q$, $|\mathbf{B}| = q$.
5: **if** $\mathbf{A}$, $\mathbf{B}$, and Q satisfy $\mathcal{R}_1$ of Lemma 1 **then**
6: $\mathcal{NCE}_k \leftarrow \mathbf{A}$, $\mathcal{NCO}_k \leftarrow \mathbf{B}$;
7: $\mathcal{C}_k = \frac{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{A}\}, \{Y \cup \mathbf{B}\}})}{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{A}\}, \{X_k \cup \mathbf{B}\}})}$;
8: Break the loop of line 3;
9: **end if**
10: **until** all possible disjoint subsets $\mathbf{A}$ with $|\mathbf{A}| = q$, $\mathbf{B}$ with $|\mathbf{B}| = q$, and variable Q in $\mathbf{X} \setminus X_k$ are selected.
11: **if** $\mathcal{NCO}_k \neq \emptyset$ **then**
12: Continue;
13: **end if**
14: **repeat**
15: Select two subsets $\mathbf{A}$ and $\mathbf{B}$ from $\mathbf{X} \setminus X_k$ such that $\mathbf{A} \cap \mathbf{B} = \emptyset$, and that $|\mathbf{A}| = q + 1$, $|\mathbf{B}| = q + 1$.
16: **if** $\mathbf{A}$ and $\mathbf{B}$ satisfy $\mathcal{R}_2$ of Lemma 2 **then**
17: $\mathcal{NCE}_k \leftarrow \mathbf{A}$, $\mathcal{NCO}_k \leftarrow \mathbf{B}$;
18: $\mathcal{C}_k = \frac{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{A}\}, \{Y \cup \mathbf{B}\}})}{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{A}\}, \{X_k \cup \mathbf{B}\}})}$;
19: Break the for loop of line 14
20: **end if**
21: **until** all possible disjoint subsets $\mathbf{A}$ with $|\mathbf{A}| = q + 1$, and $\mathbf{B}$ with $|\mathbf{B}| = q + 1$ in $\mathbf{X} \setminus X_k$ are selected.
22: **end for**
23: **if** $\mathcal{NCO}_k \neq \emptyset$ **then**
24: Continue;
25: **else**
26: $\mathcal{C}_k = \text{NA}$. // indicating the lack of knowledge to obtain the unbiased causal effect.
27: **end if**

Output: $\mathcal{C}$, a set that collects the total causal effects of $X_k \in \mathbf{X}$ on Y.

Algorithm 2 Proxy-GIN

Input: A dataset of treatments $\mathbf{X} = \{X_1, ..., X_p\}$, outcome Y, and the number of unmeasured confounders q.

1: Initialize sets $\mathcal{C} = \emptyset$, $\mathcal{NCE} = \emptyset$, and $\mathcal{NCO} = \emptyset$
2: **for** every variable X_k in $\mathbf{X}$ **do**
3: **repeat**
4: Select two disjoint subsets $\mathbf{A}$ and $\mathbf{B}$ from $\mathbf{X} \setminus X_k$ such that $|\mathbf{A}| = q$, $|\mathbf{B}| = q$.
5: **if** $\mathbf{A}$ and $\mathbf{B}$ satisfy $\mathcal{R}3$ of Lemma 3 **then**
6: $\mathcal{NCE}_k \leftarrow \mathbf{A}, \mathcal{NCO}_k \leftarrow \mathbf{B}$;
7: $\mathcal{C}_k = \frac{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{A}\}, \{Y \cup \mathbf{B}\}})}{\det(\mathbf{\Sigma}_{\{X_k \cup \mathbf{A}\}, \{X_k \cup \mathbf{B}\}})}$;
8: Break the loop of line 3;
9: **end if**
10: **until** all possible disjoint subsets $\mathbf{A}$ with $|\mathbf{A}| = q$ and $\mathbf{B}$ with $|\mathbf{B}| = q$ in $\mathbf{X} \setminus X_k$ are selected.
11: **if** $\mathcal{NCO}_k \neq \emptyset$ **then**
12: Continue;
13: **end if**
14: **end for**
15: **if** $\mathcal{NCO}_k \neq \emptyset$ **then**
16: Continue;
17: **else**
18: $\mathcal{C}_k = \text{NA}$. // indicating the lack of knowledge to obtain the unbiased causal effect.
19: **end if**

Output: $\mathcal{C}$, a set that collects the total causal effects of $X_k \in \mathbf{X}$ on Y.

Lemma 4. $\Delta_1 = o_p(1)$ *and* $\Delta_2 = o_p(1)$*, or equivalently,* $\lim_{n\to\infty} \text{pr}(|\text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) - \text{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}})| > \epsilon) = 0$ *and* $\lim_{n\to\infty} \text{pr}(|\text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) - \text{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}})| > \epsilon) = 0$ *for all* $\epsilon > 0$.

Proof. The rank of $\text{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}})$ is equal to r if and only if there exists an invertible $m \times m$ matrix C and an invertible $n \times n$ matrix D such that

$$C\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}D = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}.$$

By the law of large numbers, we have that $\hat{\mathbf{\Sigma}}_{\{X_k,\mathbf{C}\},\mathbf{B}} = \mathbf{\Sigma}_{\{X_k,\mathbf{C}\},\mathbf{B}} + o_p(1)$ as well as

$$C\hat{\mathbf{\Sigma}}_{\{X_k,\mathbf{C}\},\mathbf{B}}D = C\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}D + o_p(1) = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} + o_p(1).$$

As a result, we have $\text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) = r + o_p(1) = \text{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) + o_p(1)$. The proof for $\text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}})$ is similar, we thus omit for simplicity. □

Lemma 5. $\lim_{n\to\infty} \text{pr}(|\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon \mid \text{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \leq q, \text{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \leq q+1) = 0$ *for all* $\epsilon > 0$.

Proof. According to Lemma 2 in that main text, we know that the event $\text{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \leq q, \text{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \leq q+1$ implies $\mathbf{A}$ and $\mathbf{B}$ are valid NCE and NCO relative to $X_k \to Y$ respectively.

Once we have selected suitable proxies $\mathbf{A}$ and $\mathbf{B}$, we can utilize Eq.(3) to establish nonparametric estimator $\hat{\beta}_{X_k\to Y}$, and according to standard M-estimation theory, specifically Theorem 5.41 in Van der Vaart (2000), we can demonstrate the consistency of $\hat{\beta}_{X_k\to Y}$. □

Proposition 3. $\lim_{n\to\infty} \text{pr}(|\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon \mid \text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \leq q, \text{rk}(\hat{\mathbf{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \leq q+1) = 0$ *for all* $\epsilon > 0$.

Proof. Without loss of generality, we assume that $0 < \epsilon < 1$. We thus have,

$$
\begin{aligned}
&\mathrm{pr}(\mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1, |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon) \\
&= \mathrm{pr}\begin{pmatrix} \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) - \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \\ \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) - \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1, \\ |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon \end{pmatrix} \\
&= \mathrm{pr}(\Delta_1 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \Delta_2 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1, |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon, |\Delta_1| > \epsilon, |\Delta_2| > \epsilon) \\
&\quad + \mathrm{pr}(\Delta_1 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \Delta_2 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1, |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon, |\Delta_1| > \epsilon, |\Delta_2| \le \epsilon) \\
&\quad + \mathrm{pr}(\Delta_1 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \Delta_2 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1, |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon, |\Delta_1| \le \epsilon, |\Delta_2| > \epsilon) \\
&\quad + \mathrm{pr}(\Delta_1 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \Delta_2 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1, |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon, |\Delta_1| \le \epsilon, |\Delta_2| \le \epsilon) \\
&\le \mathrm{pr}(|\Delta_1| > \epsilon) + \mathrm{pr}(|\Delta_1| > \epsilon) + \mathrm{pr}(|\Delta_2| > \epsilon) \\
&\quad + \mathrm{pr}\begin{pmatrix} \Delta_1 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \Delta_2 + \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1, |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon, \\ -\epsilon \le -\Delta_1 \le \epsilon, -\epsilon \le -\Delta_2 \le \epsilon \end{pmatrix} \\
&\le \mathrm{pr}(|\Delta_1| > \epsilon) + \mathrm{pr}(|\Delta_1| > \epsilon) + \mathrm{pr}(|\Delta_2| > \epsilon) \\
&\quad + \mathrm{pr}(\mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q+\epsilon, \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1+\epsilon, |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon) \\
&\le o_p(1) + \mathrm{pr}(\mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1, |\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon) \\
&= o_p(1) + \mathrm{pr}(|\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon \mid \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1) \\
&\quad \times \mathrm{pr}(\mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1) \\
&\le o_p(1) + o_p(1)O_p(1) \\
&= o_p(1),
\end{aligned}
$$

where the two-to-last inequality holds because of (i) Lemma 4 and (ii) the event $\{\mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q+\epsilon\}$ is exactly equivalent to $\{\mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q\}$ for $0 < \epsilon < 1$, and the last inequality holds because of Lemma 5.

Therefore, we have

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{pr}(|\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon \mid \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1) \\
&\quad = \lim_{n\to\infty} \frac{\mathrm{pr}(|\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon, \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1)}{\mathrm{pr}(\mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1)} \\
&\quad = \lim_{n\to\infty} \frac{\mathrm{pr}(|\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y}| > \epsilon, \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\hat{\boldsymbol{\Sigma}}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1)}{\mathrm{pr}(\mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \le q, \mathrm{rk}(\boldsymbol{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \le q+1)} + o_p(1) \\
&\quad = 0.
\end{aligned}
$$

□

G.2. Discussion on the Convergence Rate

Proxy-Rank algorithm involves two states: first, correctly selecting appropriate proxy variables; next, based on the selected proxies, obtaining the nonparametric estimator $\hat{\beta}_{X_k\to Y}$ using Eq. (3). Once we have chosen suitable proxy variables, the nonparametric estimator $\hat{\beta}_{X_k\to Y}$ of the causal effects based on Eq. (3) will exhibit a $\sqrt{n}$ rate of convergence, meaning that the random variables $\sqrt{n}(\hat{\beta}_{X_k\to Y} - \beta_{X_k\to Y})$ will converge to a normal distribution. These asymptotic results can still be obtained using standard M-estimation theory, particularly Theorem 5.41 of Van der Vaart (2000).

H. Discussion and Further Work

The preceding sections have provided two different sets of precise identifiability conditions for the selection of proxy variables of unmeasured confounders, along with their corresponding search algorithms. It's worth noting that these methods theoretically assume that the unobserved confounding variable $\mathbf{U}$ affects both the treatments $\mathbf{X}$ and Y, meaning that all

entries of matrix $\mathbf{C}$ are non-zero. This assumption inherently leads to the following two conditions: 1) all variables are mutually dependent on each other, and 2) the two necessary conditions of the extended proxy variables estimator, that is, $\mathbf{\Sigma}_{\{X_k,\mathbf{Z}\},\{Y,\mathbf{W}\}}$ and $\mathbf{\Sigma}_{\{X_k,\mathbf{Z}\},\{X_k,\mathbf{W}\}}$ both are full rank. In some real-world applications, unobserved confounding might not affect all potential treatment variables. Thus, given a causal relationship $X_k \to Y$, before applying the proposed rules $\mathcal{R}1 \sim \mathcal{R}3$, one can perform an initial screening as follows:

1. identify the maximal clique set containing both X_k and Y, ensuring the removal of variables that are statistically independent of T and Y given the subset in $\mathbf{X}$, and

2. identify the candidate sets $\mathbf{Z}$ and $\mathbf{W}$ that satisfy the following two conditions: $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{Z}\},\{Y,\mathbf{W}\}}) = q + 1$ and $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{Z}\},\{X_k,\mathbf{W}\}}) = q + 1$.

The above operation ensures that even if not all entries of matrix $\mathbf{C}$ are non-zero, the identification methods we propose can still be applied.

In this paper, we restrict our attention to linear causal models, which are common in the social sciences and ought to be more common in economics and elsewhere (Bollen, 1989; Spirtes et al., 2000). One of the future research directions is to address the discrete model or the non-linear causal model, existing techniques, e.g., extended trek separation in Spirtes (2013) or additive noise model in Hoyer et al. (2009), Zhang & Hyvärinen (2009) and Peters et al. (2014) may help to address this issue. Another direction of future work is to extend our results to multiple outcomes setting (Wang et al., 2017).

I. More Results on Experimental Results

We here evaluate the performance of the proposed method in an additional **Mixture case** setting. The data are generated according to the causal graph in Figure 3, with the noise terms being randomly selected from standard normal distributions and standard exponential distributions.

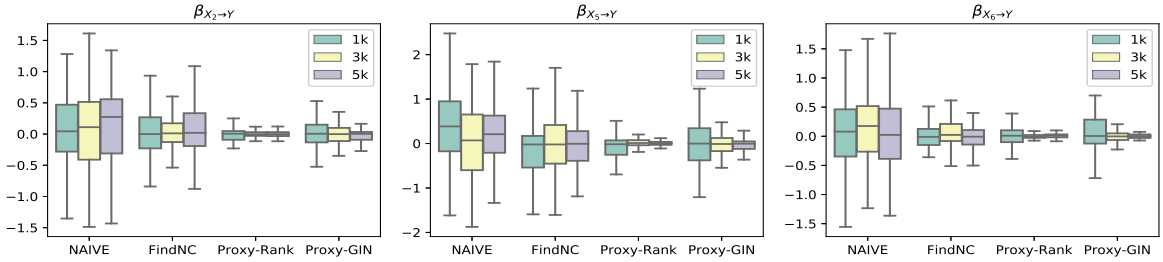


Figure 11. Performance of NAIVE, FindNC, Proxy-Rank, and Proxy-GIN on the Mixture case.

Figure 11 summarizes the bias of the estimators of each parameter. As expected, our proposed Proxy-Rank algorithm almost outperforms other methods (with little bias for all causal effects) with all sample sizes, which indicates that the Proxy-Rank algorithm is a distribution-free method. An interesting conclusion is that in the Mixture case, the Proxy-GIN algorithm performs equally well, even though Assumption 4 is not fully satisfied. We further noticed that the Proxy-GIN algorithm does not have the same level of stability in the small sample size as the Proxy-Rank algorithm, e.g., 1k for the causal effect of X_5 on Y (See Remark 4 for more details). One possible reason is that reliable estimation of higher-order statistics requires much more samples than that of second-order statistics (Hyvärinen et al., 2004).

Remark 4. *Figure 12 presents a comparative graph of the results obtained from the Proxy-Rank algorithm and the Proxy-GIN algorithm, considering two different distributions: the normal distribution and the exponential distribution, respectively. Our findings reveal that the Proxy-GIN algorithm does not exhibit the same level of stability in cases of small sample sizes, for instance, when the sample size is equal to 1k. One possible explanation for this behavior is that reliable estimation of higher-order statistics typically requires a substantially larger number of samples compared to second-order statistics (Hyvärinen et al., 2004).*

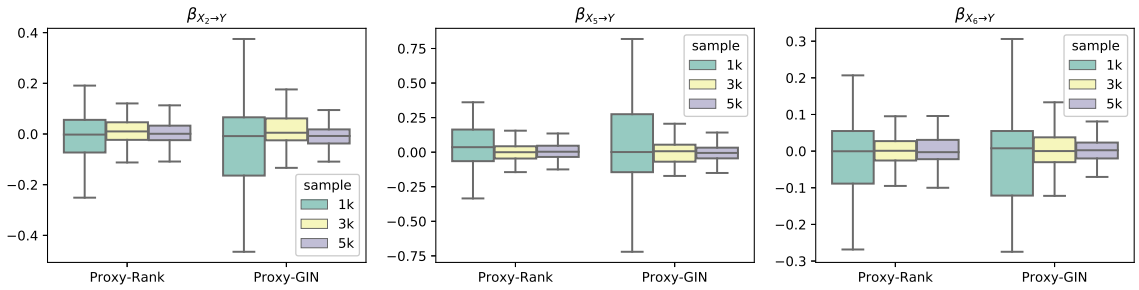


Figure 12. Performance of Proxy-Rank in Gaussian data and Proxy-GIN in Non-Gaussian.

J. More Details of Real-World Application

In this section, we apply the proposed methods 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). Additionally, it includes body weight as the outcome variable and data collected from 227 mice. As discussed in Miao et al. (2022), gene expression studies like this one may encounter unmeasured confounding issues stemming from batch effects or unobserved phenotypes. Diverging from the approach taken by Miao et al. (2022), we intentionally refrained from incorporating five additional potential instrumental variables from the raw data as prior knowledge in our analysis. This choice was made to underscore the superiority of the proposed algorithm.

Following the analysis conducted by Miao et al., we assume that there is only one latent variable underlying the common influence, and the data generation mechanism adheres to a linear causal model. Figure 13 presents the causal effects of the 17 genes on mouse weight with our proposed methods and the NAIVE method. We observed that the majority of our findings align with those presented by Miao et al. (2022). For instance, 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. Furthermore, some of our conclusions coincide with prior research results. In particular, *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). Similarly,*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).

K. Proofs

Before we proceed with presenting the proofs of our results, we require a few additional theorems and definitions.

Definition 7 (Trek)**.** *A trek in $\mathcal{G}$ from i to j is an ordered pair of directed paths (P_1, P_2) where P_1 has sink i, P_2 has sink j, and both P_1 and P_2 have the same source k. The common source k is called the top of the trek, denoted* $\mathrm{top}(P_1, P_2)$*. Note that one or both of P_1 and P_2 may consist of a single vertex, that is, a path with no edges.*

Definition 8 (trek-separation (t-separation))**.** *Let $\mathbf{A}, \mathbf{B}, \mathbf{C_A}$, and $\mathbf{C_B}$ be four variable subsets. We say the order pair $(\mathbf{C_A}, \mathbf{C_B})$ t-separates $\mathbf{A}$ from $\mathbf{B}$ if for every trek $(P_1; P_2)$ from a vertex in $\mathbf{A}$ to a vertex in $\mathbf{B}$, either P_1 contains a vertex in $\mathbf{C_A}$ or P_2 contains a vertex in $\mathbf{C_B}$.*

Note that the notion of t-separation is a more general separation criterion than d-separation in a graph (See Theorem 2.11 in (Sullivant et al., 2010)). Sullivant et al. (2010) characterized the vanishing determinants of a cross-covariance matrix by using the notion of t-separation.

Theorem 5 (Graphical Representation of Rank Constraints)**.** *Let $\mathcal{G}$ be a linear directed graphical model and $\mathbf{A}$ and $\mathbf{B}$ be two subsets of the variables in $\mathcal{G}$. The* $\mathrm{rk}(\mathbf{\Sigma_{A,B}})$ *less than or equal to r for all covariance matrices consistent with the graph $\mathcal{G}$ if and only if there exist subsets $\mathbf{C_A}, \mathbf{C_B}$ with $|\mathbf{C_A}| + |\mathbf{C_B}| \leq r$ such that $(\mathbf{C_A}, \mathbf{C_B})$ t-separates $\mathbf{A}$ from $\mathbf{B}$.*

We now quote Darmois-Skitovitch theorem (Darmois, 1953; Skitovitch, 1953).

Theorem 6 (Darmois-Skitovitch Theorem)**.** *Define two random variables V_1 and V_2 as linear combinations of independent*

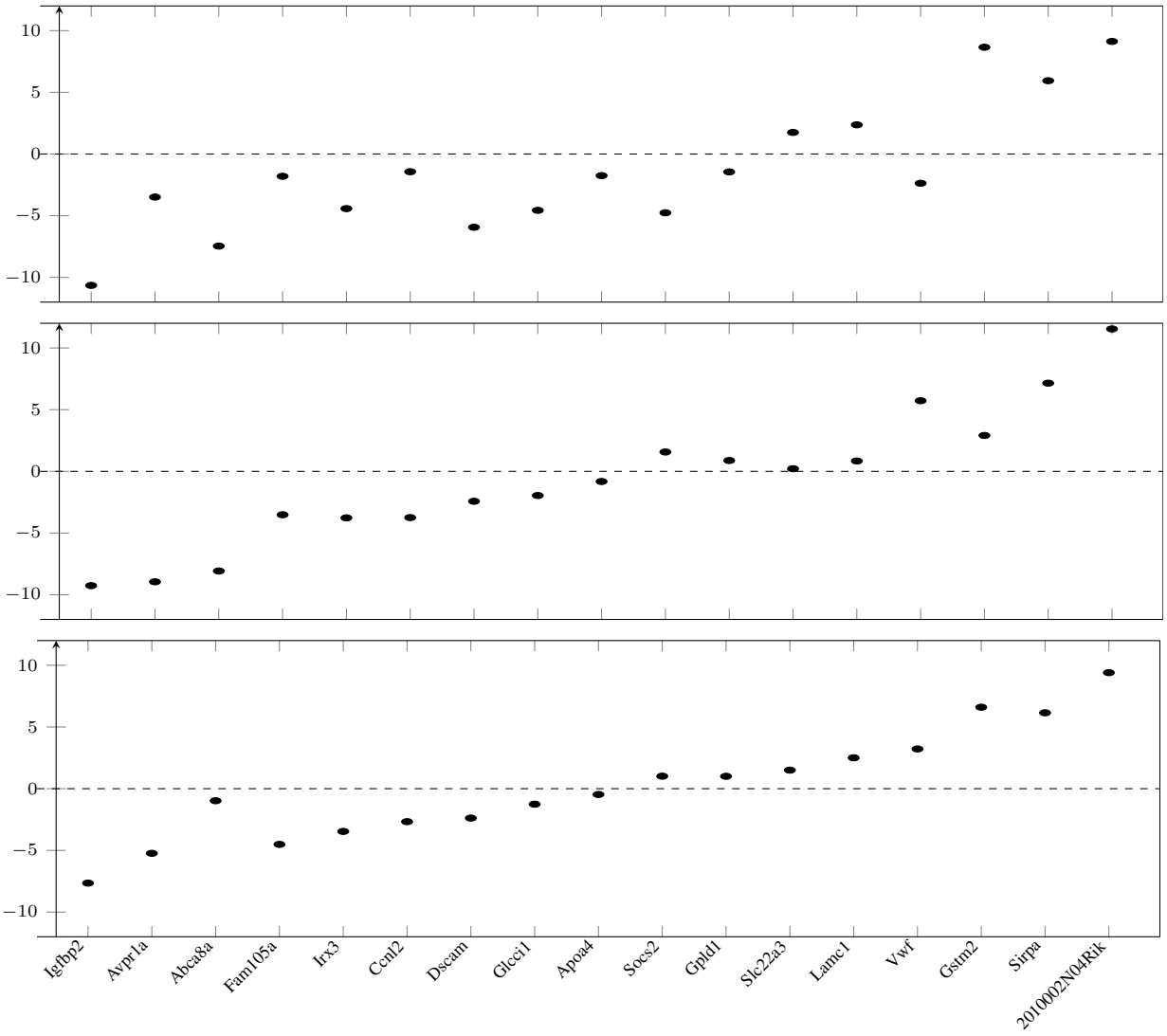


Figure 13. Causal Effect estimates for 17 gene expressions on body weight. The first panel depicts Proxy-Rank estimation, the second displays Proxy-GIN estimation, and the last one showcases NAIVE estimation.

random variables $\varepsilon_i (i = 1, ..., n)$*:*

$$V_1 = \sum_{i=1}^{n} \alpha_i \varepsilon_i, \quad V_2 = \sum_{i=1}^{n} \beta_i \varepsilon_i. \tag{28}$$

Then, if V_1 *and* V_2 *are statistically independent, all variables* ε_j *for which* $\alpha_j \beta_j \neq 0$ *are Gaussian. In other words, if there exists a non-Gaussian* ε_j *for which* $\alpha_j \beta_j \neq 0$*,* V_1 *and* V_2 *are dependent.*

We next introduce the graphical representation of GIN constraints as presented in (Xie et al., 2023).

Theorem 7 (Graphical Representation of GIN Constraints Theorem)**.** *Let* $\mathcal{G}$ *be a linear directed graphical model. Let* $\mathcal{Y}$*,* $\mathcal{Z}$ *be two sets of observed variables in* $\mathcal{G}$*.* $(\mathcal{Y}, \mathcal{Z})$ *satisfies the GIN condition (while with the same* $\mathcal{Z}$*, no proper subset of* $\mathcal{Y}$ *does) if and only if there exists a* $\mathcal{S}$ *with* $0 \leq |\mathcal{S}| \leq min(|\mathcal{Y}| - 1, |\mathcal{Z}|)$ *such that 1) the order pair* $(\emptyset, \mathcal{S})$ *t-separates* $\mathcal{Z}$ *and* $\mathcal{Y}$*, and that 2) the covariance matrix of* $\mathcal{S}$ *and* $\mathcal{Z}$ *has rank s, and so does that of* $\mathcal{S}$ *and* $\mathcal{Y}$*.*

K.1. Proof of Proposition 1

Proof. The proof can be found in Kuroki & Pearl (2014) or Miao et al. (2018a) when an unmeasured confounder exists in a linear causal model. □

K.2. Proof of Proposition 2

Here, we offer two methods of proof. The first utilizes the back door criterion and the conditional instrumental variable approach. The second utilizes the properties of the Trek rules (Sullivant et al., 2010). The details are as follows.

Proof. We initially define $\mathbf{\Sigma}_{(\mathbf{A},\mathbf{B})\cdot\mathbf{C}} = \mathbf{\Sigma}_{\mathbf{A},\mathbf{B}} - \mathbf{\Sigma}_{\mathbf{A},\mathbf{C}} \mathbf{\Sigma}_{\mathbf{C},\mathbf{C}}^{-1} \mathbf{\Sigma}_{\mathbf{C},\mathbf{B}}$. In this context, $\mathbf{\Sigma}_{(\mathbf{A},\mathbf{B})\cdot\mathbf{C}}$ can be interpreted as the conditional covariance matrices of $\mathbf{A}$ and $\mathbf{B}$ given $\mathbf{C}$, and $\mathbf{\Sigma}_{\mathbf{C},\mathbf{C}}^{-1}$ represents the inverse of $\mathbf{\Sigma}_{\mathbf{C},\mathbf{C}}$. We then apply the *back door criterion* and obtain:

$$\beta_{X_k \to Y} = \frac{\mathbf{\Sigma}_{(X_k,Y)\cdot\mathbf{U}}}{\mathbf{\Sigma}_{(X_k,X_k)\cdot\mathbf{U}}} \tag{29}$$

Hence, we have

$$\left(\mathbf{\Sigma}_{X_k,X_k} - \mathbf{\Sigma}_{X_k,\mathbf{U}} \mathbf{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1} \mathbf{\Sigma}_{\mathbf{U},X_k}\right) \beta_{X_k \to Y} = \mathbf{\Sigma}_{X_k,Y} - \mathbf{\Sigma}_{X_k,\mathbf{U}} \mathbf{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1} \mathbf{\Sigma}_{\mathbf{U},Y}. \tag{30}$$

According to the proximal criteria, i.e., $\mathbf{W} \perp\!\!\!\perp (X_k, \mathbf{Z})|\mathbf{U}$, the following two conditions hold: $\mathbf{Z} \perp\!\!\!\perp \mathbf{W}|\mathbf{U}$ and $\mathbf{W} \perp\!\!\!\perp X_k|\mathbf{U}$. Consequently, this will imply that $\mathbf{\Sigma}_{(\mathbf{Z},\mathbf{W})\cdot\mathbf{U}} = \mathbf{0}$ and $\mathbf{\Sigma}_{(\mathbf{W},X_k)\cdot\mathbf{U}} = \mathbf{0}$. Let's expand the above three equations to obtain:

$$\mathbf{\Sigma}_{\mathbf{Z},\mathbf{W}} = \mathbf{\Sigma}_{\mathbf{Z},\mathbf{U}} \mathbf{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1} \mathbf{\Sigma}_{\mathbf{U},\mathbf{W}}, \tag{31}$$

$$\mathbf{\Sigma}_{\mathbf{W},X_k} = \mathbf{\Sigma}_{\mathbf{W},\mathbf{U}} \mathbf{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1} \mathbf{\Sigma}_{\mathbf{U},X_k} \tag{32}$$

By solving the above Equations 31∼32, we obtain

$$\mathbf{\Sigma}_{X_k,\mathbf{W}} \mathbf{\Sigma}_{\mathbf{Z},\mathbf{W}}^{-1} = \mathbf{\Sigma}_{X_k,\mathbf{U}} \mathbf{\Sigma}_{\mathbf{Z},\mathbf{U}}^{-1} \tag{33}$$

To verify the conclusion, we next consider the following two scenarios. In Scenario 1, we assume independence between $\mathbf{Z}$ and X_k given $\mathbf{U}$, while in Scenario 2, we assume dependence of $\mathbf{Z}$ on X_k given $\mathbf{U}$, as illustrated in Figure 14 below. It is worth noting that, although from the graphical representation, Scenario 2 encompasses the situation in Scenario 1, they are not inclusive from a proof perspective. Specifically, Scenario 1 relies on the condition $\mathbf{Z} \perp\!\!\!\perp X_k|\mathbf{U}$, while the proof in Scenario 2 capitalizes on the property of conditional instrumental variables. In other words, in Scenario 2, given the condition $\mathbf{U}$, $\mathbf{Z}$ can serve as the instrumental variable set for the causal relationship $X_k \to Y$, whereas $\mathbf{Z}$ is not the valid instrumental variable set in Scenario 1. In summary, although the proof strategies for both scenarios are independent, it is intriguing that they share the same expression for the causal effect $X_k \to Y$.

Scenario 1: Because $\mathbf{Z} \perp\!\!\!\perp X_k|\mathbf{U}$, and according to the proximal criteria, i.e., $\mathbf{Z} \perp\!\!\!\perp Y|(\mathbf{U}, X_k)$, we further have $\mathbf{Z} \perp\!\!\!\perp Y|\mathbf{U}$. Consequently, this will imply that $\mathbf{\Sigma}_{(\mathbf{Z},X_k)\cdot\mathbf{U}} = \mathbf{0}$ and $\mathbf{\Sigma}_{(\mathbf{Z},Y)\cdot\mathbf{U}} = \mathbf{0}$. Thus, we obtain

$$\mathbf{\Sigma}_{\mathbf{Z},X_k} = \mathbf{\Sigma}_{\mathbf{Z},\mathbf{U}} \mathbf{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1} \mathbf{\Sigma}_{\mathbf{U},X_k} \tag{34}$$

$$\mathbf{\Sigma}_{\mathbf{Z},Y} = \mathbf{\Sigma}_{\mathbf{Z},\mathbf{U}} \mathbf{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1} \mathbf{\Sigma}_{\mathbf{U},Y}, \tag{35}$$

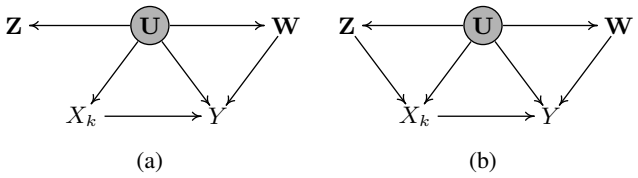


Figure 14. (a) Scenario 1: $\mathbf{Z} \perp\!\!\!\perp X_k|\mathbf{U}$, and (b) Scenario 2: $\mathbf{Z} \not\perp\!\!\!\perp X_k|\mathbf{U}$

By solving the above Equations 33∼35, we obtain

$$\boldsymbol{\Sigma}_{X_k,\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{Z},\mathbf{W}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},X_k} = \boldsymbol{\Sigma}_{X_k,\mathbf{U}}\boldsymbol{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1}\boldsymbol{\Sigma}_{\mathbf{U},X_k}, \tag{36}$$

$$\boldsymbol{\Sigma}_{X_k,\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{Z},\mathbf{W}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},Y} = \boldsymbol{\Sigma}_{X_k,\mathbf{U}}\boldsymbol{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1}\boldsymbol{\Sigma}_{\mathbf{U},Y} \tag{37}$$

By combing Equations 30, 36, and 37, we have

$$\left(\boldsymbol{\Sigma}_{X_k,X_k} - \boldsymbol{\Sigma}_{X_k,\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{Z},\mathbf{W}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},X_k}\right)\beta_{X_k\to Y} = \boldsymbol{\Sigma}_{X_k,Y} - \boldsymbol{\Sigma}_{X_k,\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{Z},\mathbf{W}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},Y}. \tag{38}$$

Hence, we finally have

$$\left(\det(\boldsymbol{\Sigma}_{\{X_k\cup\mathbf{Z}\},\{X_k\cup\mathbf{W}\}})\right)\beta_{X_k\to Y} = \det(\boldsymbol{\Sigma}_{\{X_k\cup\mathbf{Z}\},\{Y\cup\mathbf{W}\}}), \tag{39}$$

which is consistent with the Equation 3.

Scenario 2: We here apply the *conditional instrumental variable approach* and obtain:

$$\beta_{X_k\to Y} = \frac{\boldsymbol{\Sigma}_{Z_iY\cdot\mathbf{U}}}{\boldsymbol{\Sigma}_{Z_iX_k\cdot\mathbf{U}}}, \quad Z_i \in \mathbf{Z}. \tag{40}$$

That is,

$$\left(\boldsymbol{\Sigma}_{Z_i,X_k} - \boldsymbol{\Sigma}_{Z_i,\mathbf{U}}\boldsymbol{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1}\boldsymbol{\Sigma}_{\mathbf{U},X_k}\right)\beta_{X_k\to Y} = \boldsymbol{\Sigma}_{Z_k,Y} - \boldsymbol{\Sigma}_{Z_k,\mathbf{U}}\boldsymbol{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1}\boldsymbol{\Sigma}_{\mathbf{U},Y}. \tag{41}$$

Based on the proximal criteria, i.e., $\mathbf{Z} \perp\!\!\!\perp Y|(\mathbf{U}, X_k)$, we conclude that, for $Z_i \in \mathbf{Z}$, Z_i serves as a valid instrumental variable for the causal relationship $X_k \to Y$ given $\mathbf{U}$. Hence, the equation above can be expressed in vector form as:

$$\left(\boldsymbol{\Sigma}_{\mathbf{Z},X_k} - \boldsymbol{\Sigma}_{\mathbf{Z},\mathbf{U}}\boldsymbol{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1}\boldsymbol{\Sigma}_{\mathbf{U},X_k}\right)\beta_{X_k\to Y} = \boldsymbol{\Sigma}_{\mathbf{Z},Y} - \boldsymbol{\Sigma}_{\mathbf{Z},\mathbf{U}}\boldsymbol{\Sigma}_{\mathbf{U},\mathbf{U}}^{-1}\boldsymbol{\Sigma}_{\mathbf{U},Y}. \tag{42}$$

By solving Equations 30, 33, and 42 for $\beta_{X_k\to Y}$, we obtain

$$\left(\boldsymbol{\Sigma}_{X_k,X_k} - \boldsymbol{\Sigma}_{X_k,\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{Z},\mathbf{W}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},X_k}\right)\beta_{X_k\to Y} = \boldsymbol{\Sigma}_{X_k,Y} - \boldsymbol{\Sigma}_{X_k,\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{Z},\mathbf{W}}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},Y}. \tag{43}$$

This will imply that

$$\left(\det(\boldsymbol{\Sigma}_{\{X_k\cup\mathbf{Z}\},\{X_k\cup\mathbf{W}\}})\right)\beta_{X_k\to Y} = \det(\boldsymbol{\Sigma}_{\{X_k\cup\mathbf{Z}\},\{Y\cup\mathbf{W}\}}), \tag{44}$$

which is consistent with the Equation 3. □

We hereby present an alternative proof strategy employing Trek rules. Before delving into the proof of this result, we initially introduce two definitions and a theorem that play a crucial role in our argument.

Definition 9 (Trek System & Intersection)**.** *Let A and B be two subsets of the vertex set of a DAG $\mathcal{G}$, with $|A| = |B|$. A* system of treks *from A to B is a set of treks that each are between a vertex in A and a vertex in B. Let $\mathcal{T}$ be such a system. Then $\mathcal{T}$ has* no sided intersection *if any two distinct treks in $\mathcal{T}$ have disjoint left sides and disjoint right sides.*

Definition 10 (Trek Rule (Sullivant et al., 2010))**.** *Let $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{\mathcal{D}}$ be the cofficients matrix and $\Omega = (\sigma_{ij})$ be the variance of noise term. To any trek τ, associate a* trek monomial

$$\sigma(\tau) \;=\; \sigma_{i_1j_1}\prod_{k=1}^{\ell-1}\lambda_{i_ki_{k+1}}\prod_{k=1}^{r-1}\lambda_{j_kj_{k+1}}. \tag{45}$$

Theorem 8 ((Drton et al., 2020; Draisma et al., 2013)). *Suppose the underlying graph $\mathcal{G}$ is acyclic. Then the determinant of $\mathbf{\Sigma}_{A,B}$ equals $\mathcal{P}_{A,B}$, and $\mathcal{P}_{A,B}$ is defined as*

$$\mathcal{P}_{A,B} = \sum (-1)^{\mathcal{T}} \prod_{\tau \in \mathcal{T}} \sigma(\tau) \tag{46}$$

with the summation being over all systems of treks $\mathcal{T}$ from A to B with no sided intersection.

Now, we prove the Proposition 2 based on the above theorem.

Proof. Let us first consider the trek system between $\mathbf{Z}$ and $\mathbf{W}$. Based on the model definition, we know that $\mathbf{Z}$ and $\mathbf{W}$ are the child set of confounders $\mathbf{U}$ with $|\mathbf{Z}| = |\mathbf{W}| = |\mathbf{U}| = q$. Since $\mathbf{Z} \perp\!\!\!\perp \mathbf{W}|\mathbf{U}$, there are no direct path from $\mathbf{Z}$ to $\mathbf{W}$ but $Z_i \leftarrow U_i \rightarrow W_i$ is only trek between Z_i and W_i. Thus, the trek system without sided intersection between $\mathbf{Z}$ and $\mathbf{W}$ must have the source $\mathbf{U}$ according to *Pigeonhole principle*. That is any two treks τ_i and τ_j has source U_i and U_j respectively, $U_i \neq U_j$ in all trek system (note that if $U_i = U_j$ in a trek system, this trek system has sided intersection in source U_i).

Now, consider the numerator term of Eq(3), i.e., the determinant of covariance matrice between $A = \{X_k, \mathbf{Z}\}$ and $B = \{Y, \mathbf{W}\}$. To do so, by Theorem 8, it must discuss the trek system between A and B which is no side intersection. There are two cases for A and B: Case I: $\mathbf{Z} \perp\!\!\!\perp X_k|\mathbf{U}$; and Case II: $\mathbf{Z} \not\perp\!\!\!\perp X_k|\mathbf{U}$.

Case I: $\mathbf{Z} \perp\!\!\!\perp X_k|\mathbf{U}$, i.e., there are no edges between $\mathbf{Z}$ and X_k. For the trek system $\mathcal{T}$ between A and B with $|A| = |B| = q + 1$, there are $q + 1$ trek in $\mathcal{T}$. if $\mathcal{T}$ is no side intersection, the trek with the sink node X on the left side and sink node Y on the right side, denote as τ_i, $Top(\tau_i) \notin \mathbf{U}$ (i.e., the source of this trek can no be U_i). Otherwise, there exist two treks in the trek system that have a common source that violates the no-sided intersection (according to the above analysis in $\mathbf{Z}$ and $\mathbf{W}$). For instance, the trek between X_k and Y, $(U_i \rightarrow X_k, U_i \rightarrow Y)$, is intersecting with one of the trek between $\mathbf{Z}$ and $\mathbf{W}$. *An illustrative example is given in Example 1*. Therefore, the trek between $\{X_k, Y\}$ must be $(X_k; X_k \rightarrow Y)$ in the no side-intersection trek system between A and B, and meanwhile, other q trek between $\mathbf{Z}$ and $\mathbf{W}$ has source $\mathbf{U}$. According to Theorem 8, denote the trek between X_k and Y as $\tau_{\{X_k,Y\}}$, the determinant equals

$$\mathcal{P}_{A,B} = \sum (-1)^{\mathcal{T}} \prod_{\tau \in \mathcal{T} \setminus \tau_{\{X_k,Y\}}} \sigma(\tau) \cdot \sigma(\tau_{\{X_k,Y\}}). \tag{47}$$

Since $\sigma(\tau_{\{X_k,Y\}}) = \sigma(X_i)\beta_{X_k \rightarrow Y}$ (by *Trek Rule*), the above equation can be rewritten as

$$\mathcal{P}_{A,B} = \beta_{X_k \rightarrow Y} \sigma(X_i) \sum (-1)^{\mathcal{T}} \prod_{\tau \in \mathcal{T} \setminus \tau_{\{X_k,Y\}}} \sigma(\tau). \tag{48}$$

To show $\beta_{X_k \rightarrow Y}$ can be unbiasedly estimated, now we consider the denominator term of Eq. (3). Similarly, for two vectors $C = \{X_k, \mathbf{Z}\}$ and $D = \{X_k, \mathbf{W}\}$, the determinant can be formalized as

$$\mathcal{P}_{C,D} = \sigma(X_i) \sum (-1)^{\mathcal{T}'} \prod_{\tau \in \mathcal{T} \setminus \tau_{\{X_k,Y\}}} \sigma(\tau), \tag{49}$$

where the covariance of the trek between X_k and X_k equals $\sigma(X_i)$. Based on the above analysis, we can get the unbiased estimation of $\beta_{X_k \rightarrow Y}$ by the ratio of two determinants, i.e., $\mathcal{P}_{A,B}/\mathcal{P}_{C,D}$.

Case II: $\mathbf{Z} \not\perp\!\!\!\perp X_k|\mathbf{U}$, i.e., there exits a edges between Z_i and X_k for some $Z_i \in \mathbf{Z}$. According to the proximal criteria, e.g., $\{X_k, \mathbf{Z}\} \perp\!\!\!\perp \mathbf{W}|\mathbf{U}$, there are no other trek between Z_i and W_i except for the $(U_i \rightarrow Z_i, U_i \rightarrow W_i)$. For this case, the key difference to Case I is that there may exist a trek between Z_i and Y in which the source $Top(Z_i, Y) \notin \mathbf{U}$.

There are two cases for the trek between Z_i and Y: (i). $\tau_1 = (Z_i, Z_i \rightarrow X_k \rightarrow Y)$ or (ii). $\tau_2 = (X_k \rightarrow Z_i, X_k \rightarrow Y)$. By *Trek Rule*, we have $\sigma(\tau_1) = \sigma(Z_i)\beta_{Z_i \rightarrow X_k}\beta_{X_k \rightarrow Y}$ for case (i) while $\sigma(\tau_2) = \sigma(X_k)\beta_{X_k \rightarrow Z_i}\beta_{X_k \rightarrow Y}$ for case (ii). Furthermore, if there are more than one Z_i, for example, Z_i and Z_j have the trek of the above cases, then this trek system has a side intersection in the node X_k (as $\sigma(\tau_1)$ and $\sigma(\tau_2)$ has side intersection in the node X_k). Thus, a trek system without sided intersection between $\{X_k, \mathbf{Z}\}$ and $\{Y, \mathbf{W}\}$ can only be one trek following the above cases, i.e., only a Z_i follows the above cases. Thus, the set of no side-intersection trek system is q treks with source $\mathbf{U}$ between $\{\mathbf{Z} \setminus Z_i, X_k\}$ and $\{\mathbf{W}\}$ and plus a trek between Z_i and Y. Denote the trek between Z_i and Y as $\tau_{\{Z_i,Y\}}$, $\tau_{\{Z_i,Y\}}$ follows one of the above cases.

Therefore, by Theorem 8, the determinant can be formalized as

$$\beta_{X_k \to Y} \sum (-1)^{\mathcal{T}'} \prod_{\tau' \in \mathcal{T}'} \sigma(\tau') + \sum_{1}^{q} \tau_{Z_i - X_k} \beta_{X_k \to Y} \sum (-1)^{\mathcal{T}''} \prod_{\tau \in \mathcal{T}''} \sigma(\tau), \tag{50}$$

where the first term represents the summation of no sided intersection system including the trek with source $\mathbf{U}$ between $\mathbf{Z}$ and $\mathbf{W}$ and a trek from X_k to Y; the second term presents the summation of the no sided intersection system inlcuding the trek with source $\mathbf{U}$ between $\{\mathbf{Z} \setminus Z_i, X_k\}$ and $\mathbf{W}$ and a trek between Z_i and Y. Similarly, for two vector $\{X_k, \mathbf{Z}\}$ and $\{X_k, \mathbf{W}\}$, the determinant can be formalized as

$$\sum (-1)^{\mathcal{T}'} \prod_{\tau' \in \mathcal{T}'} \sigma(\tau') + \sum_{1}^{q} \tau_{Z_i - X_k} \sum (-1)^{\mathcal{T}''} \prod_{\tau \in \mathcal{T}''} \sigma(\tau). \tag{51}$$

In the end, we can obtain the unbiased estimation of $\beta_{X_k \to Y}$ by the ratio of two determinants. □

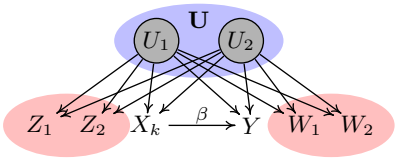


Figure 15. Illustration of Proof of Proposition 2.

Example 6 (Illustration of Proof of Proposition 2)**.** *Consider a graph in Fig. 15. There are two confounders U_1 and U_2 that affect the X_k and Y. Let $\mathbf{Z} = \{Z_1, Z_2\}$ and $\mathbf{W} = \{W_1, W_2\}$. For two vector $\{X_k, Z_1, Z_2\}$ and $\{Y, W_1, W_2\}$, all no intersection trek system between $\mathbf{Z}$ and $\mathbf{W}$ must has the source $\{U_1, U_2\}$ due to $\mathbf{Z} \perp\!\!\!\perp X_k | \{U_1, U_2\}$, that is,*

- *($U_1 \to Z_1$, $U_1 \to W_1$) and ($U_2 \to Z_2$, $U_2 \to W_2$)*
- *($U_2 \to Z_1$, $U_2 \to W_1$) and ($U_1 \to Z_2$, $U_1 \to W_2$)*
- *($U_1 \to Z_2$, $U_1 \to W_1$) and ($U_2 \to Z_1$, $U_2 \to W_2$)*
- *($U_2 \to Z_2$, $U_2 \to W_1$) and ($U_1 \to Z_1$, $U_1 \to W_2$)*

For the trek between X_k and Y, it must be $(X_k, X_k \to Y)$, otherwise, it will be an intersection with the trek between $\mathbf{Z}$ and $\mathbf{W}$ in the source node $\mathbf{U}$. For example, a trek between X_k and Y with source U_1, $(U_1 \to X_k, U_1 \to Y)$, is intersecting with ($U_1 \to Z_1$, $U_1 \to W_1$) in the trek system: ($U_1 \to Z_1$, $U_1 \to W_1$), ($U_2 \to Z_2$, $U_2 \to W_2$) and $(U_1 \to X_k, U_1 \to Y)$. Therefore, according to Theorem 1, the determinant equals $\beta_{X_k \to Y}[\sigma^2_{U_1}\sigma^2_{U_2}\sigma^2_{X_k}(a_1a_3b_2b_4 - a_1a_4b_2b_3 - a_2a_3b_1b_4 + a_2a_4b_1b_3)]$, where a_i represent the effect from U_1 to $\mathbf{Z} \cup \mathbf{W}$ while b_i represent the effect from U_2 to $\mathbf{Z} \cup \mathbf{W}$.

For two vectors $\{X_k, Z_1, Z_2\}$ and $\{X_k, W_1, W_2\}$, based on Theorem 1, we also obtain the determinant as $\sigma^2_{U_1}\sigma^2_{U_2}\sigma^2_{X_k}(a_1a_3b_2b_4 - a_1a_4b_2b_3 - a_2a_3b_1b_4 + a_2a_4b_1b_3)$. Therefore, one may obtain an unbiased estimation of $\beta_{X_k \to Y}$ by the ratio of two determinants.

K.3. Proof of Lemma 1

Proof. Firstly, in accordance with Equation 2, where all entries of the matrix $\mathbf{C}$ are non-zero, and under the faithfulness assumption, we can directly infer that both$\mathbf{\Sigma}_{\{X_k, \mathbf{A}\}, \{Y, \mathbf{B}\}}$ and $\mathbf{\Sigma}_{\{X_k, \mathbf{A}\}, \{X_k, \mathbf{B}\}}$ are always full rank. Therefore, these necessary conditions are satisfied for any sets $\mathbf{Z}$ and $\mathbf{W}$ (otherwise, it would be required to test these conditions).

Secondly, due to condition (1), namely $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k, Q, \mathbf{A}\}, \{X_k, Y, \mathbf{B}\}}) \leq q + 1$, and the faithfulness assumption, and based on the "Graphical Representation of Rank Constraints" Theorem, we can assert that there exist subsets $\mathbf{C_A}$ and $\mathbf{C_B}$ with $|\mathbf{C_A}| + |\mathbf{C_B}| \leq q + 1$ such that $(\mathbf{C_A}, \mathbf{C_B})$ t-separates $\{X_k, Q, \mathbf{A}\}$ from $\{X_k, Y, \mathbf{B}\}$. According to the data generation process, the treks between $\{Q, A\}$ and $\{Y, B\}$ must go through unmeasured confounder $\mathbf{U}$ (except for X_k). Therefore,

$\mathbf{C_A} \cup \mathbf{C_B} = \{X_k, \mathbf{U}\}$. Since $|X_k, \mathbf{U}| = q + 1$, we can conclude that all treks between $\{X_k, Q, \mathbf{A}\}$ and $\{X_k, Y, \mathbf{B}\}$ must go through a node in $\{X_k, \mathbf{U}\}$. This will imply that $\mathbf{A} \perp\!\!\!\perp Y|(\mathbf{U}, X_k)$, i.e., condition 1 of proximal criteria holds, and $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|(\mathbf{U}, X_k)$.

Furthermore, because of condition 2), i.e., $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{Q,\mathbf{B}\}}) \leq q$, and because according to the "Graphical Representation of Rank Constraints" Theorem, we know that there exist subsets $\mathbf{C_A}, \mathbf{C_B}$ with $|\mathbf{C_A}| + |\mathbf{C_B}| \leq q$ such that $(\mathbf{C_A}, \mathbf{C_B})$ t-separates $\{X_k, \mathbf{A}\}$ from $\{Q, \mathbf{B}\}$. According to the generation of data (Equation 2), all treks between $\{X_k, \mathbf{A}\}$ and $\{Q, \mathbf{B}\}$ must go through unmeasured confounders $\mathbf{U}$. Hence, $\mathbf{C_A} \cup \mathbf{C_B} = \mathbf{U}$. This will imply that $\mathbf{B} \perp\!\!\!\perp X_k|\mathbf{U}$. Because $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|(\mathbf{U}, X_k)$, we have $\mathbf{B} \perp\!\!\!\perp (X_k, \mathbf{A})|\mathbf{U}$, i.e., condition 2 of proximal criteria holds.

Based on the above analysis, $\mathbf{A}$ and $\mathbf{B}$ are valid NCE and NCO with respect to $X_k \to Y$, respectively. Furthermore, due to Assumption 2, we know that such sets $\mathbf{A}$ and $\mathbf{B}$ must exist in the system. □

K.4. Proof of Lemma 2

Proof. The proof strategy for this theorem is similar to the proof strategy for Lemma 1.

Firstly, according to Equation 2 (where all entries of matrix $\mathbf{C}$ are non-zero) and the faithfulness assumption, we can directly infer that $\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{Y,\mathbf{B}\}}$ and $\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,\mathbf{B}\}}$ are both full rank. Therefore, these necessary conditions always hold for any sets $\mathbf{Z}$ and $\mathbf{W}$ (otherwise, it would be required to test these conditions).

Secondly, due to condition (1), namely $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,Y,\mathbf{B}\}}) \leq q + 1$, and the faithfulness assumption, and based on the "Graphical Representation of Rank Constraints" Theorem, we can assert that there exist subsets $\mathbf{C_A}$ and $\mathbf{C_B}$ with $|\mathbf{C_A}| + |\mathbf{C_B}| \leq q + 1$ such that $(\mathbf{C_A}, \mathbf{C_B})$ t-separates $\{X_k, \mathbf{A}\}$ from $\{X_k, Y, \mathbf{B}\}$. According to the data generation process, the treks between $\{Q, A\}$ and $\{Y, B\}$ must go through unmeasured confounder $\mathbf{U}$ (except for X_k). Therefore, $\mathbf{C_A} \cup \mathbf{C_B} = \{X_k, \mathbf{U}\}$. Since $|\{X_k, \mathbf{U}\}| = q + 1$, we can conclude that all treks between $\{X_k, \mathbf{A}\}$ and $\{X_k, Y, \mathbf{B}\}$ must go through a node in $\{X_k, \mathbf{U}\}$. This will imply that $\mathbf{A} \perp\!\!\!\perp Y|(\mathbf{U}, X_k)$, i.e., condition 1 of proximal criteria holds, and $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|(\mathbf{U}, X_k)$.

Furthermore, because of condition 2), i.e., $\mathrm{rk}(\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\mathbf{B}}) \leq q$, and because according to the "Graphical Representation of Rank Constraints" Theorem, we know that there exist subsets $\mathbf{C_A}, \mathbf{C_B}$ with $|\mathbf{C_A}| + |\mathbf{C_B}| \leq q$ such that $(\mathbf{C_A}, \mathbf{C_B})$ t-separates $\{X_k, \mathbf{A}\}$ from $\{\mathbf{B}\}$. According to the generation of data (Equation 2), all treks between $\{X_k, \mathbf{A}\}$ and $\{\mathbf{B}\}$ must go through unmeasured confounders $\mathbf{U}$. Hence, $\mathbf{C_A} \cup \mathbf{C_B} = \mathbf{U}$. This will imply that $\mathbf{B} \perp\!\!\!\perp X_k|\mathbf{U}$. Because $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|(\mathbf{U}, X_k)$, we have $\mathbf{B} \perp\!\!\!\perp (X_k, \mathbf{A})|\mathbf{U}$, i.e., condition 2 of proximal criteria holds.

Based on the above analysis, $\mathbf{A}$ and $\mathbf{B}$ are valid NCE and NCO with respect to $X_k \to Y$, respectively. Furthermore, due to Assumption 3, we know that such sets $\mathbf{A}$ and $\mathbf{B}$ must exist in the system. □

K.5. Proof of Theorem 1

Proof. Assuming Assumption 2 holds, then according to Lemma 1, for a given causal relationship $X_k \to Y$ in the system, the underlying NCE and NCO relative to the causal relationship $X_k \to Y$ can be identified using $\mathcal{R}1$.

Similarly, assuming Assumption 3 holds, then according to Lemma 2, for a given causal relationship $X_k \to Y$ in the system, the underlying NCE and NCO relative to the causal relationship $X_k \to Y$ can be identified using $\mathcal{R}2$. □

K.6. Proof of Theorem 2

Proof. The correctness of Proxy-Rank originates from the following observations:

- Firstly, for a given causal relationship $X_k \to Y$ in the system, by Lemma 1 and Proposition 2, valid set of NCE and NCO in $\mathbf{X} \setminus X_k$ have been exactly discovered, and the unbiased causal effect $\mathcal{C}_k$ is obtained if Assumption 2 satisfies (Lines 3∼13 of Algorithm 1).
- Secondly, for a given causal relationship $X_k \to Y$ in the system, by Lemma 2 and Proposition 2, valid set of NCE and NCO in $\mathbf{X} \setminus X_k$ have been exactly discovered, and the unbiased causal effect $\mathcal{C}_k$ is obtained if Assumption 2 violates but Assumption 3 satisfies (Lines 14∼21 Algorithm 1).
- Lastly, value (NA) is obtained, which indicates the lack of knowledge to obtain the unbiased causal effect (Lines 23∼27 Algorithm 1).

□

K.7. Proof of Lemma 3

Proof. Firstly, according to Equation 2 (where all entries of matrix $\mathbf{C}$ are non-zero) and the faithfulness assumption, we can directly infer that $\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{Y,\mathbf{B}\}}$ and $\mathbf{\Sigma}_{\{X_k,\mathbf{A}\},\{X_k,\mathbf{B}\}}$ are both full rank. Therefore, these necessary conditions always hold for any sets $\mathbf{Z}$ and $\mathbf{W}$ (otherwise, it would be required to test these conditions).

Secondly, due to condition (1), namely $(\{X_k, \mathbf{A}\}, \{X_k, Y, \mathbf{B}\})$ follows the GIN constraint, the faithfulness assumption, Assumption 4 (Non-Gaussianity), and based on the "Graphical Representation of GIN Constraints" Theorem, we can assert that there exist $\mathcal{S}$ with $0 \leq |\mathcal{S}| \leq \min(|\{X_k, Y, \mathbf{B}\}| - 1, |\{X_k, \mathbf{A}\}|) = q + 1$ such that 1) the order pair $(\emptyset, \mathcal{S})$ t-separates $\mathcal{Z}$ and $\mathcal{Y}$, and that 2) the covariance matrix of $\mathcal{S}$ and $\mathcal{Z}$ has rank s, and so does that of $\mathcal{S}$ and $\mathcal{Y}$. According to the data generation process, the treks between $\{\mathbf{A}\}$ and $\{Y, \mathbf{B}\}$ must go through unmeasured confounder $\mathbf{U}$ (except for X_k). Therefore, $\mathcal{S} = \{X_k, \mathbf{U}\}$. Since $|\{X_k, \mathbf{U}\}| = q + 1$, we can conclude that all treks between $\{X_k, Q, \mathbf{A}\}$ and $\{X_k, Y, \mathbf{B}\}$ must go through a node in $\{X_k, \mathbf{U}\}$. This will imply that $\mathbf{A} \perp\!\!\!\perp Y | (\mathbf{U}, X_k)$, i.e., condition 1 of proximal criteria holds.

Furthermore, because of condition 2), i.e., $(\mathbf{B}, \{X_k, \mathbf{A}\})$ follows the GIN constraint, because of Assumption 4, and because according to the "Graphical Representation of Rank Constraints" Theorem, we know that there exist $\mathcal{S}$ with $0 \leq |\mathcal{S}| \leq \min(|\{X_k, \mathbf{A}\}| - 1, |\mathbf{B}|) = q$ such that the order pair $(\emptyset, \mathcal{S})$ t-separates $\mathbf{B}$ and $\{X_k, \mathbf{A}\}$. According to the generation of data (Equation 2), all treks between $\mathbf{B}$ and $\{X_k, \mathbf{A}\}$ must go through unmeasured confounders $\mathbf{U}$. Hence, $\mathcal{S} = \mathbf{U}$. This will imply that $|\mathbf{U}| = q$. Thus, we have $\mathbf{B} \perp\!\!\!\perp (X_k, \mathbf{A}) | \mathbf{U}$, i.e., condition 2 of proximal criteria holds.

Based on the above analysis, $\mathbf{A}$ and $\mathbf{B}$ are valid NCE and NCO with respect to $X_k \to Y$, respectively. Furthermore, due to Assumptions 1 and 4, we know that such sets $\mathbf{A}$ and $\mathbf{B}$ must exist in the system. □

K.8. Proof of Theorem 3

Proof. Assuming Assumptions 1 and 4 hold, then according to Lemma 3, for a given causal relationship $X_k \to Y$ in the system, the underlying NCE and NCO relative to the causal relationship $X_k \to Y$ can be identified using $\mathcal{R}3$. □

K.9. Proof of Theorem 4

Proof. The correctness of Proxy-Rank originates from the following observations:

- Firstly, for a given causal relationship $X_k \to Y$, by Lemma 3 and Proposition 2, valid set of NCE and NCO in $\mathbf{X} \setminus X_k$ have been exactly discovered, and the unbiased causal effect $\mathcal{C}_k$ is obtained if Assumption 1 satisfies (Lines 2∼10 of Algorithm 2).
- Then, value (NA) is obtained, which the lack of valid NCE and NCO for this causal relationship $X_k \to Y$ to obtain the unbiased causal effect (Lines 15∼19 Algorithm 2).

□