Appendices

Table of Contents

A Notations and Abbreviations **15**

B NSCSL for Nonlinear Models: The POC-based Learning Algorithm **15**
B.1 Nonlinear Structural Equation Model and Estimation of POC 15
B.1.1 Nonlinear Structural Equation Model 15
B.1.2 Estimation of POC 16
B.2 Learning Algorithm based on POCs 16
B.2.1 Step 1: Nonlinear Causal Structural Learning 16
B.2.2 Step 2: Causal Relevance Measurement Using POCs 16
B.2.3 Step 3: Necessary and Sufficient Causal Structural Learning 17
B.3 The Computational Complexity of the NSCSL Algorithm 17
B.4 Discussion on Scale Invariance 17

C Extension to Markov Equivalence Class **17**
C.1 Additional Graph Terminology 17
C.2 Model Identifiabilities 17
C.3 Extended Algorithm for Markov Equivalence Class 18

D Technical Proofs **18**
D.1 Proofs of Thm. 4.4 19
D.1.1 Part 1: Lower Bound for M-POC 19
D.1.2 Part 2: Lower Bound for C-POC 20
D.1.3 Part 3: Conditions to Achieve Lower Bounds 21
D.2 Proofs of Thm. 4.6 21
D.3 Proofs of Thm. 5.1 22

E Additional Simulation Results **25**
E.1 Simulation Configurations 25
E.2 More Real Data Analyses on Yeast Data 25
E.3 Additional Simulation Results: True and Estimated Matrix 27
E.4 Additional Simulation Results: True and Estimated Graphs 32

A Notations and Abbreviations

Table A.1: The table of notations and abbreviation.

Concept	Description
Graph $\mathcal{G}$	A graph with a node set $\boldsymbol{X}$ and edge set $\boldsymbol{D_X}$
Node X_i	A member of the node set $\boldsymbol{X}$
Parent of X_j	A node X_i such that there is a directed edge from X_i to X_j (i.e., X_i is a direct cause of X_j)
Ancestor of X_j	A node X_k such that there's a directed path from X_k to X_j regulated by at least one additional node X_i for $i \neq k$ and $i \neq j$ (i.e., X_k is an indirect cause of X_j)
$\text{PA}_{X_j}(\mathcal{G})$	The set of all parents/ancestors of node X_j in $\mathcal{G}$
DAG	A directed acyclic graph; a directed graph $\mathcal{G}$ that does not contain directed cycles
SCM	The structural causal model characterizes the causal relationship among $\lvert\boldsymbol{X}\rvert = d$ nodes via a DAG $\mathcal{G}$ and noises $\boldsymbol{e_X} = [e_{X_1}, \cdots, e_{X_d}]^\top$ such that $X_i := h_i\{\text{PA}_{X_i}(\mathcal{G}), e_{X_i}\}$ for some unknown h_i and $i = 1, \cdots, d$
$Y(Z_i = z_i)$	Potential outcome after setting individual variable Z_i to z_i
$\mathcal{G}_{\boldsymbol{O}}$	DAG characterizing the causal relationship among $\boldsymbol{O}$
$\mathbb{P}_{\mathcal{G}}$	The mass/density function for an SCM with its DAG $\mathcal{G}$
$\boldsymbol{Z}_{-i} \equiv \boldsymbol{Z} \setminus Z_i$	The complementary variable set of Z_i

B NSCSL for Nonlinear Models: The POC-based Learning Algorithm

In this section, we outline a method for learning the NSCG for a nonlinear model. Given the lack of an explicit form for both causal effects and POCs in such models, we employ an iterative learning method based on Thm. 4.4 to capture complex nonlinear relationships among variables. The method begins with a pre-screening process, identifying necessary and sufficient features from $\boldsymbol{Z}$ with high causation scores. Following this, the causal graph is estimated among the selected nodes and Y, approximating $\mathcal{G}_{\boldsymbol{V}}$ in the nonlinear structural equation model. This iterative approach is applicable for general SCMs and the process is repeated until convergence is achieved. App. B.1 introduces the nonlinear structural equation model and describes the estimation of POC in such a model. The main algorithm based on POCs is presented in App. B.2.

B.1 Nonlinear Structural Equation Model and Estimation of POC

B.1.1 Nonlinear Structural Equation Model

While the linear structural equation model has good properties such as easy implementation and nice interpretation, it cannot capture complex nonlinear causal relationships. To address this, we consider the non-linear additive form following [25; 47; 28]. Specifically, for variables $\boldsymbol{D} = \{g(\boldsymbol{Z}), Y\}$ as a $d+1$-dimensional vector, we consider replacing the model in (1) as follows,

$$D_i := \psi_i\{\text{PA}_{D_i}(\mathcal{G})\} + e_{D_i}, \tag{B.1}$$

for the i-th element/node D_i in $\boldsymbol{D}$ with some unknown nonlinear function ψ_i and independent noise e_{D_i}, for $i = 1, \cdots, d+1$. As mentioned in [47; 28], this model is identifiable from observational data. We further define a new functional matrix $\boldsymbol{B}(\psi) = \boldsymbol{B}(\psi_1, \cdots, \psi_{d+1})$ that encodes the unknown causal relationship among variables. The element in the i-th row and j-th column is defined as:

$$[\boldsymbol{B}(\psi)]_{ij} := ||\partial_j \psi_i||_2, \tag{B.2}$$

where $||\cdot||_2$ presents the L_2 norm. This matrix thus describes the dependency among $\boldsymbol{D}$; if D_i does not depend on D_j, we have $||\partial_j \psi_i||_2 = 0$. We can also incorporate background knowledge as done in § 5.2 to reflect the causal roles in different types of variables. We do this by specifying the last column of $\boldsymbol{B}(\psi)$ to be all zeros, so that $h_c \boldsymbol{B}(\psi) = \sum_{i=1}^{d+1} |[\boldsymbol{B}(\psi)]_{i,d+1}| = 0$.

B.1.2 Estimation of POC

We detail the estimation of POC based on Thm. 4.4 and the model (B.1) as follows. Denote the estimator of the conditional probability of the outcome given the i-th selected feature $g_i(\boldsymbol{Z})$ as $\widehat{\mathbb{P}}(Y = y|g_i(\boldsymbol{Z}))$. This can be achieved by either parametric models (such as logistic regression for the binary outcome) or non-parametric models (such as random forest or neural network). Then, the estimated marginal POC across n data points is given by

$$\widehat{\text{M-POC}}(g_i|\{\boldsymbol{o}^{(j)}\}) = \prod_{j=1}^{n} \left| \widehat{\mathbb{P}}\{Y = y^{(j)}|g_i(\boldsymbol{Z}) = g_i(\boldsymbol{z}^{(j)})\} - \widehat{\mathbb{P}}\{Y = y^{(j)}|g_i(\boldsymbol{Z}) \neq g_i(\boldsymbol{z}^{(j)})\} \right|,$$

where $g_i(\boldsymbol{Z})$ is the i-th dimension of $g(\boldsymbol{Z})$. Similarly, we can estimate the conditional probability of the outcome given all selected feature $g(\boldsymbol{Z})$ as $\widehat{\mathbb{P}}(Y = y|g(\boldsymbol{Z}))$. Likewise, we have the estimated conditional POC as

$$\begin{aligned}\widehat{\text{C-POC}}(g_i|\{\boldsymbol{o}^{(j)}\}) = \prod_{j=1}^{n} \Big| &\widehat{\mathbb{P}}\{Y = y^{(j)}|g_i(\boldsymbol{Z}) = g_i(\boldsymbol{z}^{(j)}), g_{-i}(\boldsymbol{Z}) = g_{-i}(\boldsymbol{z}_{-i}^{(j)})\} \\ &- \widehat{\mathbb{P}}\{Y = y^{(j)}|g_i(\boldsymbol{Z}) \neq g_i(\boldsymbol{z}^{(j)}), g_{-i}(\boldsymbol{Z}) = g_{-i}(\boldsymbol{z}_{-i}^{(j)})\} \Big|,\end{aligned}$$

where $g_{-i}(\cdot) \equiv g(\cdot) \setminus g_i(\cdot)$ is the complement of $g_i(\cdot)$.

B.2 Learning Algorithm based on POCs

Given the lack of an explicit form for causal quantities in the nonlinear models, we employ an iterative learning method based on Thm. 4.4 to capture complex nonlinear relationships among variables. The method begins with an initialized causal graph and then identifies necessary and sufficient features from $\boldsymbol{Z}$ with high POCs by setting g as a subset selection function. Following this, the causal graph can be updated among the selected nodes and Y, approximating $\mathcal{G}_{\boldsymbol{V}}$ in the nonlinear structural equation model based on data $\{\boldsymbol{o}^{(j)} = (\boldsymbol{z}^{(j)}, y^{(j)})\}_{1 \leq j \leq n}$. This process is repeated until convergence is achieved. The main algorithm based on POCs is presented as follows.

B.2.1 Step 1: Nonlinear Causal Structural Learning

In the first step, we employ a causal structural learning algorithm for nonlinear models to estimate the functional matrix $\boldsymbol{B}(\psi)$ as presented in (B.2). This estimation is performed considering the selector g^k at the k-th iteration, where $g^1(\boldsymbol{Z}) := \boldsymbol{Z}$. Aligning with the main text, we adopt the nonparametric acyclicity constraint on $\boldsymbol{B}(\psi)$ proposed in Zheng et al. [47], represented as $h_n(\boldsymbol{B}(\psi)) = 0$. The loss function, defined by the augmented Lagrangian for the k-th iteration, is then formulated as:

$$\tilde{L}(\boldsymbol{B}(\psi), \theta, \lambda|g^k, \{\boldsymbol{o}^{(j)}\}) = \tilde{f}(\boldsymbol{B}(\psi), \theta|g^k, \{\boldsymbol{o}^{(j)}\}) + \lambda\{h_c(\boldsymbol{B}(\psi)) + h_n(\boldsymbol{B}(\psi))\}, \quad \text{(B.3)}$$

where $\tilde{f}(\boldsymbol{B}(\psi), \theta|g^k, \{\boldsymbol{o}^{(j)}\})$ denotes a nonlinear loss function with parameters θ, and λ represents the Lagrange multiplier. The objective in (B.3) can be solved using existing nonlinear causal structural learning methods (refer to Yu et al. [44], Zhu et al. [48], Zheng et al. [47]) given the selector g^k and data $\{\boldsymbol{o}^{(j)} = (\boldsymbol{z}^{(j)}, y^{(j)})\}_{1 \leq j \leq n}$. The estimated functional matrix resulting from this process is represented as $\widehat{\boldsymbol{B}}^k(\widehat{\psi})$.

B.2.2 Step 2: Causal Relevance Measurement Using POCs

Following the estimation of the functional matrix $\widehat{\boldsymbol{B}}^k(\widehat{\psi})$, we assess the causal relevance of nodes through the probabilities of causation (POCs) to update the selection function g. The loss function relating to the selector is defined as follows:

$$\tilde{L}^P(g, \gamma|\widehat{\boldsymbol{B}}^k(\widehat{\psi}), \boldsymbol{o}^{(j)}) = -\sum_{i=1}^{d} \widehat{\text{P}}(g_i|\boldsymbol{o}^{(j)}) + \gamma|g|, \quad \text{(B.4)}$$

where $|g|$ signifies the number of selected nodes in g with a penalty γ for controlling the complexity of the selector. The term g_i refers to the i-th dimension of g for $i = 1, \cdots, d$. Here, $\widehat{\mathrm{P}}(\cdot)$ can represent either the estimated M-POC or C-POC as outlined in App. B.1. By optimizing different POCs, we can learn the corresponding best selector g according to the loss function detailed in (B.4) using the CAUSAL-REP algorithm as proposed in Wang & Jordan [42] by considering the selector g as the subset of $\boldsymbol{Z}$. The resulting selector is denoted as g^{k+1}.

B.2.3 Step 3: Necessary and Sufficient Causal Structural Learning

Repeat the optimization process for loss functions in (B.3) and (B.4) for $k = 1, \cdots, K$ iterations until either the maximum iteration number K is reached, or the change in loss functions in (B.3) and (B.4) falls below a predefined tolerance level τ. The resultant estimated matrix is denoted as $\widehat{\boldsymbol{B}}(\widehat{\psi})$, from which the estimated causal graph, $\widehat{\mathcal{G}}_{\boldsymbol{V}}$, can be derived.

B.3 The Computational Complexity of the NSCSL Algorithm

The computational complexity of NSCSL comprises two parts: the cost from causal discovery as $G(n, p)$, and the estimation of causal effects/scores $F(n, p)$, where n is the data sample size and p is the number of nodes. In the linear case, our method learns the features and causal graph through single-step optimization in (5), with complexity cubic in the number of nodes, $G(n, p) = \mathcal{O}(p^3)$ following Zheng et al. [46]. Here, the causal effect computation is linear-time and thus is dominated. In the nonlinear case, according to App. B.2, the time complexity depends on the base causal discovery method and the number of max iterations K, yielding $\mathcal{O}[K(G(n, p) + F(n, p))]$. Supporting runtime details are provided in § 6.

B.4 Discussion on Scale Invariance

NSCSL is scale-invariant when we appropriately choose the causal discovery base learner and model the treatment effects/POCs. Though NOTEARS lacks scale invariance, our method's flexibility allows for the integration of scale-invariant causal discovery methods like NSCGL with FCI. Additionally, under LSEM, the rescaling will not affect the relative rank of the features based on absolute causal effects. In the nonlinear case, we proposed to use POCs which by their definitions are scale-invariant.

C Extension to Markov Equivalence Class

Our proposed algorithm can also be extended to manage the Markov equivalence class of partial directed acyclic graphs when the causal graph cannot be uniquely identified from the observational studies.

C.1 Additional Graph Terminology

A graph $\mathcal{G}$ that contains directed and/or undirected edges is termed as a partially directed graph. If this graph doesn't contain a directed cycle, it's referred to as a partially directed acyclic graph or PDAG. The DAG $\mathcal{G}$ is generally not identifiable from the distribution of $\boldsymbol{X}$ based on conditional independence relationships, as per observational data [22]. This is because multiple DAGs can represent the same conditional independence relationships, and these DAGs form a Markov equivalence class (MEC), denoted as $MEC(\mathcal{G})$. Two DAGs belong to the same MEC if and only if they have the same skeleton and the same v-structures [22]. A MEC of DAGs can be uniquely symbolized by a completed partially directed acyclic graph (CPDAG) [35; 7], a graph that may contain both directed and undirected edges. A CPDAG adheres to the following: if $X_i \to X_j$ exists in the CPDAG, then $X_i \to X_j$ is present in every DAG in the MEC; and if $X_i - X_j$ exists in the CPDAG, then the MEC contains a DAG for which $X_i \to X_j$ as well as a DAG for which $X_j \to X_i$.

C.2 Model Identifiabilities

In the absence of further assumptions regarding the form of functions and/or noises, the model in (1) can only be identified up to MEC following the Markov and faithful assumptions [35; 25]. Below, we

explore the conditions for the unique identifiability of the DAG and potential strategies for addressing scenarios involving the MEC.

Initially, we consolidate cases where the DAG is uniquely identifiable. In the context of the LSEM, when the noises $\boldsymbol{\epsilon}$ follow a Gaussian distribution, the resulting model corresponds to the standard linear-Gaussian model class, as investigated in Spirtes et al. [35] and Peters et al. [26]. In instances where the noises $\boldsymbol{\epsilon}$ maintain equal variances, according to Peters & Bühlmann [24], the DAG $\mathcal{G}$ can be uniquely identified from observational data. Further, when the functions are linear but the noises are non-Gaussian, one can derive the LiNGAM as described in Shimizu et al. [34], where the true DAG can be uniquely identified under certain favorable conditions. In addition, as cited in Zheng et al. [47]; Rolland et al. [28], the nonlinear additive model can be identified from observational data. Another scenario of note arises when the corresponding MEC encompasses only one DAG; here, the DAG can be inherently identified from observational data. Recent score-based causal discovery algorithms [46; 44; 48; 5] typically take into account synthetic datasets generated from fully identifiable models, which provides practical relevance in evaluating the estimated graph in relation to the true DAG.

In instances where the true DAG is not identifiable, we reference the discussion in App. C.1. In such cases, a CPDAG uniquely symbolizes a MEC of DAGs that yield the same joint distribution of variables. This CPDAG can be inferred from observational data via a variety of causal discovery algorithms [see e.g., 35; 7; 34; 14; 4; 27]. One feasible approach to dealing with MEC involves enumerating all DAGs in the MEC derived from a given CPDAG [6]. It is conventional to encapsulate a range of potential effects or probabilities by their average or the minimum absolute value [6; 33]. However, such an approach typically proves computationally prohibitive for large graphs, necessitating computational shortcuts to acquire the causal effects or probabilities of causation without enumerating all DAGs in the MEC of the estimated CPDAG.

C.3 Extended Algorithm for Markov Equivalence Class

In contrast to existing causal discovery algorithms, we consider a causal graph that is necessary and sufficient to portray causal relationships influencing the outcome variable Y. This aim is expressed as a causal identification constraint designed to restrict the causal structural learning to a smaller class of DAGs, as detailed in the ensuing section. Additionally, we employ causal effects or probabilities of causation as another loss function within the objective. This assists in identifying the DAG with the highest score of conditional scores of causation or causal effects, wherein the v-structures of interest are also constrained to have an endpoint at Y. Based on the estimated CPDAG, which involves a significantly smaller number of nodes, we can generate all DAGs within the MEC and prune superfluous nodes following aggregation.

Specifically, the proposed NSCSL algorithm can be adapted to handle MEC of PDAGs when the causal graph is not uniquely identifiable from observational data. Let's recall either the estimated matrix $\widehat{\boldsymbol{B}}$ obtained by NSCSL based on causal effects, as presented in § 5.2 for the *linear* model, or the estimated functional matrix $\widehat{\boldsymbol{B}}(\widehat{\psi})$ by the POC-based NSCSL in App. B for the *nonlinear* model. Based on these estimated (functional) matrices, we can derive the estimated causal graph $\widehat{\mathcal{G}}_{\boldsymbol{V}}$. This can further lead to estimation under its MEC by averaging over possible DAGs, as follows:

$$\widehat{\mathcal{G}}^{*}_{\boldsymbol{V}} = \frac{1}{|MEC(\widehat{\mathcal{G}}_{\boldsymbol{V}})|} \sum_{\mathcal{G}_i \in MEC(\widehat{\mathcal{G}}_{\boldsymbol{V}})} \mathcal{G}_i, \tag{C.1}$$

where $|MEC(\widehat{\mathcal{G}}_{\boldsymbol{V}})|$ is the size of MEC for $\widehat{\mathcal{G}}_{\boldsymbol{V}}$. As we have previously noted, the number of nodes in $\widehat{\mathcal{G}}_{\boldsymbol{V}}$ is much smaller than p, making it feasible to generate all DAGs in the MEC and prune extraneous nodes following aggregation.

D Technical Proofs

In this section, we provide proofs for Thms. 4.4 and 4.6. Let the symbol $\wedge$ denote the logical connective and, and the symbol $\vee$ denote the logical connective or. For two events A and B, $A \wedge B =$ True if $A = B =$ True, and $A \wedge B =$ False otherwise. Furthermore, $A \vee B =$ False if $A = B =$ False, and $A \vee B =$ True otherwise.

D.1 Proofs of Thm. 4.4

Consider the marginal probability of causation (M-POC) for Z_i in Def. 4.2 such that

$$\text{M-POC}_i(y) \equiv \mathbb{P}\{Y(Z_i \neq z_i) \neq y, Y(Z_i = z_i) = y\},$$

and the conditional probability of causation (C-POC) for Z_i in Def. 4.3 such that

$$\text{C-POC}_i(y) \equiv \mathbb{P}\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y, Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}.$$

Our goal is to establish their lower bounds. In the following, we derive the lower bound for M-POC in Part 1 and the lower bound for C-POC in Part 2. Finally, we discuss conditions for the lower bound equality to be held in Part 3.

D.1.1 Part 1: Lower Bound for M-POC

We focus on M-POC first. Based on the consistency assumption (A1), we have either $\{Y(Z_i \neq z_i) = y\}$ or $\{Y(Z_i \neq z_i) \neq y\}$ holds. Since the events $\{Y(Z_i \neq z_i) = y\}$ and $\{Y(Z_i \neq z_i) \neq y\}$ are disjoint, we have

$$\{Y(Z_i \neq z_i) = y\} \vee \{Y(Z_i \neq z_i) \neq y\} = \text{True}, \tag{D.1}$$

where the symbol $\vee$ denote the logical connective *or*, meaning one of the above event holds. Based on this fact, we focus on the second event $\{Y(Z_i = z_i) = y\}$ in M-POC which yields

$$\begin{aligned} & \{Y(Z_i = z_i) = y\} \\ = \ & \{Y(Z_i = z_i) = y\} \wedge \text{True} \\ = \ & \{Y(Z_i = z_i) = y\} \wedge [\{Y(Z_i \neq z_i) = y\} \vee \{Y(Z_i \neq z_i) \neq y\}] \\ = \ & [\{Y(Z_i = z_i) = y\} \wedge \{Y(Z_i \neq z_i) = y\}] \vee [\{Y(Z_i = z_i) = y\} \wedge \{Y(Z_i \neq z_i) \neq y\}], \end{aligned} \tag{D.2}$$

where the first equality is owing to the definition of the logical connective $\wedge$, the second equality comes from (D.1), and the last equality follows the rule of interchange in the logical connectives, i.e, $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$ for events A, B, and C. By noticing the last line is the logical connective *or* of two events, taking the probability on both sides of (D.2) gives

$$\begin{aligned} & \mathbb{P}\{Y(Z_i = z_i) = y\} \\ \leq \ & \mathbb{P}[\{Y(Z_i = z_i) = y\} \wedge \{Y(Z_i \neq z_i) = y\}] + \mathbb{P}[\{Y(Z_i = z_i) = y\} \wedge \{Y(Z_i \neq z_i) \neq y\}] \\ = \ & \underbrace{\mathbb{P}[Y(Z_i = z_i) = y, Y(Z_i \neq z_i) = y]}_{\eta_1} + \mathbb{P}[Y(Z_i = z_i) = y, Y(Z_i \neq z_i) \neq y] \\ = \ & \eta_1 + \text{M-POC}_i(y), \end{aligned} \tag{D.3}$$

where the first inequality is owing to $\mathbb{P}(A \vee B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, and the equalities are due to the definitions of probabilities.

Similarly, by (A1), since the events $\{Y(Z_i = z_i) = y\}$ and $\{Y(Z_i = z_i) \neq y\}$ are disjoint, we have

$$\{Y(Z_i = z_i) = y\} \vee \{Y(Z_i = z_i) \neq y\} = \text{True}.$$

Based on this fact, the first event $\{Y(Z_i \neq z_i) = y\}$ in M-POC is

$$\begin{aligned} & \{Y(Z_i \neq z_i) = y\} \\ = \ & \{Y(Z_i \neq z_i) = y\} \wedge \text{True} \\ = \ & \{Y(Z_i \neq z_i) = y\} \wedge [\{Y(Z_i = z_i) = y\} \vee \{Y(Z_i = z_i) \neq y\}] \\ = \ & [\{Y(Z_i \neq z_i) = y\} \wedge \{Y(Z_i = z_i) = y\}\}] \vee [\{Y(Z_i \neq z_i) = y\} \wedge \{Y(Z_i = z_i) \neq y\}]. \end{aligned} \tag{D.4}$$

By noticing the last line is the logical connective *or* of two events, taking the probability on both sides of (D.4) gives

$$\begin{aligned} \mathbb{P}\{Y(Z_i \neq z_i) = y\} &\geq \mathbb{P}[\{Y(Z_i \neq z_i) = y\} \wedge \{Y(Z_i = z_i) = y\}] \\ &= \mathbb{P}[Y(Z_i \neq z_i) = y, Y(Z_i = z_i) = y] = \eta_1, \end{aligned} \tag{D.5}$$

where the first inequality is owing to $\mathbb{P}(A \vee B) \geq \mathbb{P}(A)$ and the last equality comes from the definition of η_1.

Combining (D.3) and (D.5), we have

$$\begin{aligned} & \text{M-POC}_i(y) \qquad\qquad (D.6)\\ \text{(by (D.3))} \quad &\geq \mathbb{P}\{Y(Z_i = z_i) = y\} - \eta_1 \\ \text{(by (D.5))} \quad &\geq \mathbb{P}\{Y(Z_i = z_i) = y\} - \mathbb{P}\{Y(Z_i \neq z_i) = y\} \\ &= \mathbb{P}\{Y = y | Z_i = z_i\} - \mathbb{P}\{Y = y | Z_i \neq z_i\}, \end{aligned}$$

where the last equation follows the results that $\mathbb{P}\{Y = y|do(X = x)\} = \mathbb{P}\{Y = y|X = x\}$ under the ignorability assumption (A2) following Rosenbaum & Rubin [29] and Pearl et al. [22, 23]. The proof of the first part thus is completed.

D.1.2 Part 2: Lower Bound for C-POC

We next show the lower bound of conditional POC in Thm. 4.4 following the same logic as in Part 1. Based on the consistency assumption (A1), we have either $\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}$ or $\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\}$ holds, i.e., these two events are disjoint, thus, we have

$$\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \vee \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\} = \text{True}, \qquad (D.7)$$

where the symbol $\vee$ denote the logical connective *or*, meaning one of the above event holds. Based on this fact, we focus on the second event $\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}$ in C-POC which yields

$$\begin{aligned} & \{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \qquad (D.8)\\ = \; & \{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \text{True} \\ = \; & \{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \\ & \qquad \wedge [\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \vee \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\}] \\ = \; & [\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}] \\ & \qquad \vee [\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\}], \end{aligned}$$

where the first equality is owing to the definition of the logical connective $\wedge$, the second equality comes from (D.7), and the last equality follows the rule of interchange in the logical connectives, i.e, $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$ for events A, B, and C. By noticing the last line is the logical connective *or* of two events, taking the probability on both sides of (D.8) gives

$$\begin{aligned} & \mathbb{P}\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \qquad (D.9)\\ \leq \; & \mathbb{P}[\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}] \\ & \qquad + \mathbb{P}[\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\}] \\ = \; & \underbrace{\mathbb{P}[Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y, Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y]}_{\eta_2} \\ & \qquad + \mathbb{P}[Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y, Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y] \\ = \; & \eta_2 + \text{C-POC}_i(y), \end{aligned}$$

where the first inequality is owing to $\mathbb{P}(A \vee B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, and the equalities are due to the definitions of probabilities. Similarly, by (A1), since the events $\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}$ and $\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\}$ are disjoint, we have

$$\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \vee \{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\} = \text{True}.$$

Based on this fact, the first event $\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}$ in C-POC is

$$\begin{aligned} & \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \qquad (D.10)\\ = \; & \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \text{True} \\ = \; & \{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \\ & \qquad \wedge [\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \vee \{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\}] \\ = \; & [\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}\}] \\ & \qquad \vee [\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) \neq y\}]. \end{aligned}$$

By noticing the last line is the logical connective *or* of two events, taking the probability on both sides of (D.10) gives

$$\begin{aligned} & \mathbb{P}\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \qquad (D.11)\\ \geq \; & \mathbb{P}[\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \wedge \{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\}] \\ = \; & \mathbb{P}[Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y, Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y] = \eta_2, \end{aligned}$$

where the first inequality is owing to $\mathbb{P}(A \vee B) \geq \mathbb{P}(A)$ and the last equality comes from the definition of η_2. Combining (D.9) and (D.11), we have

$$\begin{aligned} & \text{C-POC}_i(y) \qquad (D.12) \\ \text{(by (D.9))} \quad & \geq \mathbb{P}\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} - \eta_2 \\ \text{(by (D.11))} \quad & \geq \mathbb{P}\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} - \mathbb{P}\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}) = y\} \\ & = \mathbb{P}\{Y = y | Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}\} - \mathbb{P}\{Y = y | Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}\}, \end{aligned}$$

where the last equation follows the results that $\mathbb{P}\{Y = y | do(X = x)\} = \mathbb{P}\{Y = y | X = x\}$ under the ignorability assumption (A2) following Rosenbaum & Rubin [29] and Pearl et al. [22, 23]. The proof of the second part thus is completed.

D.1.3 Part 3: Conditions to Achieve Lower Bounds

In this part, we discuss the conditions under which the POCs are equal to their corresponding lower bounds. To this end, we introduce the following monotonicity condition.

(C1*). **Monotonicity**:
(i) $\{Y(\boldsymbol{Z} \neq \boldsymbol{z}) = y\} \wedge \{Y(\boldsymbol{Z} = \boldsymbol{z}) \neq y\} = \text{False}$;
(ii) $\{Y(Z_i \neq z_i) = y\} \wedge \{Y(Z_i = z_i) \neq y\} = \text{False}$.

Here, (C1*.i) is proposed in Section 9.2.3 in Pearl et al. [22] and also Tian & Pearl [39] to establish the identifiability of the probability of causation. We generalize the condition (C1*.i) to the condition (C1*.ii) so that we can extend the results in Theorem 9.2.14 in Pearl et al. [22] for POCs in Defs. 4.2 and 4.3.

We detail the case of M-POC first. By noticing the monotonicity condition in (C1*.ii) such that $[\{Y(Z_i \neq z_i) = y\} \wedge \{Y(Z_i = z_i) \neq y\}] = \text{False}$, we can simplify (D.4) as

$$\{Y(Z_i \neq z_i) = y\} = [\{Y(Z_i \neq z_i) = y\} \wedge \{Y(Z_i = z_i) = y\}]. \qquad (D.13)$$

Substituting (D.13) into (D.2) yields

$$\{Y(Z_i = z_i) = y\} = \{Y(Z_i \neq z_i) = y\} \vee [\{Y(Z_i = z_i) = y\} \wedge \{Y(Z_i \neq z_i) \neq y\}]. \qquad (D.14)$$

Based on the consistency assumption (A1), we have either $\{Y(Z_i \neq z_i) = y\}$ or $\{Y(Z_i \neq z_i) \neq y\}$ holds, and thus the events $\{Y(Z_i \neq z_i) = y\}$ and $[\{Y(Z_i = z_i) = y\} \wedge \{Y(Z_i \neq z_i) \neq y\}]$ are disjoint. Therefore, taking the probability on both sides of (D.14) gives

$$\mathbb{P}\{Y(Z_i = z_i) = y\} = \mathbb{P}\{Y(Z_i \neq z_i) = y\} + \mathbb{P}\{Y(Z_i = z_i) = y, Y(Z_i \neq z_i) \neq y\}. \qquad (D.15)$$

Recall Def. 4.2. Based on (D.15), we have

$$\begin{aligned} \text{M-POC}_i(y) & = \mathbb{P}\{Y(Z_i \neq z_i) \neq y, Y(Z_i = z_i) = y\} \\ & = \mathbb{P}\{Y(Z_i = z_i) = y\} - \mathbb{P}\{Y(Z_i \neq z_i) = y\} \\ & = \mathbb{P}\{Y = y | Z_i = z_i\} - \mathbb{P}\{Y = y | Z_i \neq z_i\}, \end{aligned}$$

where the last equation follows the results that $\mathbb{P}\{Y = y | do(X = x)\} = \mathbb{P}\{Y = y | X = x\}$ under the Ignorability assumption by Rosenbaum & Rubin [29] and Pearl et al. [22, 23]. Thus, the lower bound equality for M-POC holds when an additional monotonicity condition is imposed. Following the same logic, we can show the lower bound equality for C-POC holds given the monotonicity condition. We omit the details for brevity.

D.2 Proofs of Thm. 4.6

As a direct result of Thm. 4.4, below we can establish the relationship between POC and the corresponding expected mean outcome given different combinations of the confounders involving Z_i. Specifically, we take expectations over Y on both sides of (D.6) and (D.12). When Y is nonnegative, this yields

$$\begin{aligned} \sum_{y \in \mathcal{L}} y\text{M-POC}_i(y) \geq & \sum_{y \in \mathcal{L}} y \left[\mathbb{P}\{Y(Z_i = z_i) = y\} - \mathbb{P}\{Y(Z_i \neq z_i) = y\}\right] \\ \geq & \mathbb{E}\{Y(Z_i = z_i)\} - \mathbb{E}\{Y(Z_i \neq z_i)\}, \end{aligned}$$

and

$$\begin{aligned}\sum_{y\in\mathcal{L}} y\text{C-POC}_i(y) \geq& \sum_{y\in\mathcal{L}} y\left[\mathbb{P}\{Y(Z_i=z_i,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i})=y\}-\mathbb{P}\{Y(Z_i\neq z_i,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i})=y\}\right]\\ \geq& \mathbb{E}\{Y(Z_i=z_i,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i})\}-\mathbb{E}\{Y(Z_i\neq z_i,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i})\}.\end{aligned}$$

Under the ignorability assumption (A2) following Rosenbaum & Rubin [29] and Pearl et al. [22, 23], we have

$$\begin{aligned}&\sum_{y\in\mathcal{L}} y\text{M-POC}_i(y)\geq \delta_M(z_i)\equiv \mathbb{E}\{Y|Z_i=z_i\}-\mathbb{E}\{Y|Z_i\neq z_i\},\\ &\sum_{y\in\mathcal{L}} y\text{C-POC}_i(y)\geq \delta_C(z_i)\equiv \mathbb{E}\{Y|Z_i=z_i,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i}\}-\mathbb{E}\{Y|Z_i\neq z_i,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i}\},\end{aligned}\tag{D.16}$$

where $\delta_M(z_i)$ and $\delta_C(z_i)$ are defined as the marginal and conditional causal effects using the differences of expectations based on the corresponding POC. Recall the definitions of natural causal effects for Z_i as

$$\begin{aligned}TE_i&=\mathbb{E}\{Y(Z_i=z_i+1)\}-\mathbb{E}\{Y(Z_i=z_i)\},\\ DE_i&=\mathbb{E}\{Y(Z_i=z_i+1,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i}^{(z_i)})\}-\mathbb{E}\{Y(Z_i=z_i)\},\end{aligned}$$

where $\boldsymbol{z}_{-i}^{(z_i)}$ is the value of $\boldsymbol{Z}_{-i}$ if setting $do(Z_i=z_i)$. When Z_i is binary, by comparing the definitions with $Z_i\in\{0,1\}$, we have

$$|TE_i|\geq\delta_M(z_i),\quad |DE_i|\geq\delta_C(z_i).\tag{D.17}$$

Therefore, combining (D.16) and (D.17) yields the second conclusion in Thm. 4.6 that

$$\begin{aligned}&\min\{\sum_{y\in\mathcal{L}} y\text{M-POC}_i(y),|TE_i|\}\geq \mathbb{E}\{Y|Z_i=z_i\}-\mathbb{E}\{Y|Z_i\neq z_i\},\\ &\min\{\sum_{y\in\mathcal{L}} y\text{C-POC}_i(y),|DE_i|\}\geq \mathbb{E}\{Y|Z_i=z_i,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i}\}-\mathbb{E}\{Y|Z_i\neq z_i,\boldsymbol{Z}_{-i}=\boldsymbol{z}_{-i}\}.\end{aligned}$$

Further, when Z_i is binary, the absolute values of $\delta_M(z_i)$ and $\delta_C(z_i)$ is equal to the absolute values of TE_i and DE_i, respectively. Combining this with (D.16) yields the second conclusion in Thm. 4.6. Finally, following the same logic in App. D.1.3, we can show the lower bound equality for Thm. 4.6 holds when the monotonicity condition is imposed. We omit the details for brevity. The proof is thus completed.

D.3 Proofs of Thm. 5.1

We investigate the theoretical consistency of the proposed causal structural learning methods under Model (1) with independent Gaussian error and equal variance, using the score-based method such as NOTEARS that minimizes the loss in (5).

Notations and Conditions: We first detail some notations and the required conditions in Thm. 5.1 below. Denote $\boldsymbol{W}_j\in\mathbb{R}^{n\times 1}$ as the row vector of a matrix $\boldsymbol{W}\in\mathbb{R}^{(d+1)\times n}$ for $j=1,\cdots d+1$. Let $\text{supp}(v)=\{i:v_i\neq 0\}$ denote the support of a vector v, which is the set of indices of nonzero terms of v. The first condition requires a bounded true matrix $\boldsymbol{B}$ as follows.

Condition D.1. The true matrix $\boldsymbol{B}\in\mathbb{R}^{(d+1)\times(d+1)}$ is bounded such that $\|\boldsymbol{B}\|_2=\mathcal{O}(1)$, and the maximum degree across different rows is less than the number of nodes in the graph, i.e., $s_0=\max_j \text{supp}(\boldsymbol{B}_j)\leq d+1$.

Next, we recall the linear structural equation model condition on $\boldsymbol{W}\equiv[g(\boldsymbol{Z})^\top,Y]^\top$ as follows.

Condition D.2. Suppose Model (1) such that $\boldsymbol{W}=\boldsymbol{B}^\top\boldsymbol{W}+\boldsymbol{\epsilon}$, with independent Gaussian error $\boldsymbol{\epsilon}$ and the error variance is a constant σ^2.

Furthermore, define the order as $\nu=[\nu_1,\nu_2,\cdots,\nu_{d+1}]^\top$ is a permutation of indices $\{1,2,...,d,d+1\}$ of nodes in the causal graph, such that $\nu_j\in\{1,2,...,d+1\}$. If we set the last node to be the target

variable, then the last index is fixed to be $d+1$. Let $\boldsymbol{B}(\nu)$ as $\boldsymbol{B}(\nu) = [\boldsymbol{B}_1(\nu), \cdots, \boldsymbol{B}_{d+1}(\nu)]^\top$, such that

$$\boldsymbol{B}_j(\nu) = \underset{\boldsymbol{\beta}:\text{supp}(\boldsymbol{\beta})\in\{\nu_1,\cdots,\nu_{j-1}\}}{\arg\min} \mathbb{E}(\boldsymbol{W}_{\nu_j} - \boldsymbol{\beta}^\top \boldsymbol{W})^2, \quad j = 1, 2, \cdots, d+1.$$

Note that the true causal graph $\boldsymbol{B}$ that generates $\boldsymbol{W}$ also has a topological order, ν^*, which is called the true topological order (this true order may not be unique). Then the true causal graph $\boldsymbol{B}$ can also be denoted as $\boldsymbol{B}(\nu^*)$. Denote the set of all permuations of $\{1, 2, \cdots, d+1\}$ as Υ and the set of true order as Υ^* such that $\Upsilon^* \subset \Upsilon$. Denote the order of $\widehat{\boldsymbol{B}}$ as $\nu \in \Upsilon$, so $\widehat{\boldsymbol{B}}$ can also be notated as $\widehat{\boldsymbol{B}}(\nu)$. Then let $s_j(\nu) = \text{supp}(\boldsymbol{B}_j(\nu)), \widehat{s}_j(\nu) = \text{supp}(\widehat{\boldsymbol{B}}_j(\nu))$. The last condition assumes the consistency of the topological ordering.

Condition D.3. The true topological ordering of $\boldsymbol{B}$, i.e., ν^*, is consistently estimated.

Condition D.3 is commonly imposed when proving the error bound of causal structural learning results [see Condition (A6) in 33].

Overview of Proof: With the aforementioned three conditions, our proof follows similar strategies of the causal structural learning literature [e.g., 33] but accounts for the extra penalty term from causal effects. Notice that the explicit forms of causal effects under LSEM are linear combinations of elements of $\boldsymbol{B}$. This implies our new regulation can similarly vanish away as n goes to infinity. For simplicity, in the rest of the proof, we show the consistency given a selection function g for brevity. To start with, proving the consistency of $\widehat{\boldsymbol{B}}$ returned by NSCSL with NOTEARS as baseline is equivalent to showing that the $\widehat{\boldsymbol{B}}$ solving

$$\widehat{\boldsymbol{B}} = \arg\min \|\boldsymbol{W} - \boldsymbol{B}^\top \boldsymbol{W}\|_2^2 + \lambda h_2(\boldsymbol{B}), \quad \text{subject to } h_1(\boldsymbol{B}) = 0. \tag{D.18}$$

is consistent. Then the proof of Thm. 5.1 follows the same strategy as the proof of Proposition 1 in [33]. The key difference between Thm. 5.1 and Proposition 1 in [33] is that the loss function in (D.18) contains extra penalty term $h_2(\boldsymbol{B}) = \delta^* - \sum_{i=1}^{d} |\widehat{CE}_i(\boldsymbol{B})| + \sum_{i=1}^{d+1} |b_{i,d+1}|$, that involves causal effect information compared to the original loss function

$$\widetilde{\boldsymbol{B}} = \arg\min \|\boldsymbol{W} - \boldsymbol{B}^\top \boldsymbol{W}\|_2^2, \quad \text{subject to } h_1(\boldsymbol{B}) = 0,$$

in [46]. In this proof, we need to show a main statement and the rest of the proof will follow the same procedure of Steps 2-3 in Appendix Section 9 of [33]. The **Main Statement** is:

$$\left\|\widehat{\boldsymbol{B}}(\nu) - \boldsymbol{B}(\nu)\right\|_2 = \mathcal{O}\Big(\sqrt{\sum_{j=1}^{d+1} \frac{\log n}{n}(s_j(\nu) + \widehat{s}_j(\nu))} + \frac{\lambda}{n}\sqrt{s_0(d+1)}\Big),$$

with λ satisfy the bound of $\mathcal{O}((n \log n)^{1/2})$ for arbitrary $\nu \in \Upsilon$.

Proof of Main Statement: Since $\widehat{\boldsymbol{B}}(\nu)$ solves (D.18), we have

$$\|\boldsymbol{W} - \widehat{\boldsymbol{B}}(\nu)^\top \boldsymbol{W}\|_2^2 + \lambda h_2(\widehat{\boldsymbol{B}}(\nu)) \leq \|\boldsymbol{W} - \boldsymbol{B}(\nu)^\top \boldsymbol{W}\|_2^2 + \lambda h_2(\boldsymbol{B}(\nu)). \tag{D.19}$$

Following Theorem 7.1 in Van de Geer & Bühlmann [40] and Shi & Li [33], we can transform (D.19) to

$$\frac{1}{2}\|\boldsymbol{B}(\nu)^\top \boldsymbol{W} - \widehat{\boldsymbol{B}}(\nu)^\top \boldsymbol{W}\|_2^2 + \lambda h_2(\widehat{\boldsymbol{B}}(\nu)) \leq 2\sum_{j=1}^{d+1} \kappa(s_j(\nu) + \widehat{s}_j(\nu)) \log n + \lambda h_2(\boldsymbol{B}(\nu)), \tag{D.20}$$

for some $\kappa > 0$. Recall Condition D.2 such that $\boldsymbol{W} = \boldsymbol{B}^\top \boldsymbol{W} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}^{n\times 1}, \boldsymbol{I}_n \times \sigma^2)$. From Equation (65) in [33], we have $0 < \kappa^* < \sigma^2$ such that

$$\kappa^* n \sum_{j=1}^{d+1} \|\boldsymbol{B}_j(\nu) - \widehat{\boldsymbol{B}}_j(\nu)\|_2^2 \leq \frac{1}{2}\|\boldsymbol{B}(\nu)^\top \boldsymbol{W} - \widehat{\boldsymbol{B}}(\nu)^\top \boldsymbol{W}\|_2^2 \leq \frac{1}{\kappa^*} n \sum_{j=1}^{d+1} \|\boldsymbol{B}_j(\nu) - \widehat{\boldsymbol{B}}_j(\nu)\|_2^2, \forall \nu. \tag{D.21}$$

Combining (D.21) with (D.20), we have

$$\kappa^* n \sum_{j=1}^{d+1} \|\boldsymbol{B}_j(\nu) - \widehat{\boldsymbol{B}}_j(\nu)\|_2^2 + \lambda h_2(\widehat{\boldsymbol{B}}(\nu)) \leq 2\sum_{j=1}^{d+1} \kappa(s_j(\nu) + \widehat{s}_j(\nu)) \log n + \lambda h_2(\boldsymbol{B}(\nu)). \tag{D.22}$$

If $\lambda h_2(\widehat{\boldsymbol{B}}(\nu)) > \lambda h_2(\boldsymbol{B}(\nu))$, the main statement directly holds.

In the other case, we focus on showing the main statement when $\lambda h_2(\widehat{\boldsymbol{B}}(\nu)) \leq \lambda h_2(\boldsymbol{B}(\nu))$. Specifically, we first define a vector

$$\boldsymbol{\mu}(\nu) = [\widehat{\boldsymbol{B}}_1(\nu) - \boldsymbol{B}_1(\nu), \widehat{\boldsymbol{B}}_2(\nu) - \boldsymbol{B}_2(\nu), \cdots, \widehat{\boldsymbol{B}}_p(\nu) - \boldsymbol{B}_{d+1}(\nu)]^\top \in \mathbb{R}^{(d+1)\times 1}.$$

Our goal is to bound $\|\boldsymbol{\mu}(\nu)\|_2 = \sum_{j=1}^{d+1} \|\widehat{\boldsymbol{B}}_j(\nu) - \boldsymbol{B}_j(\nu)\|_2^2$. Define $\mathcal{M}(\nu) = \operatorname{supp}(\boldsymbol{\mu}(\nu))$ and $\mathcal{M}^c(\nu)$ is the complementary set. Then denote $\boldsymbol{\mu}(\nu)_{\mathcal{M}(\nu)}$ as the vector formed by elements of $\boldsymbol{\mu}(\nu)$ in $\mathcal{M}(\nu)$. Denote $\boldsymbol{\mu}(\nu)_{\mathcal{M}^c(\nu)}$ as the vector formed by elements of $\boldsymbol{\mu}(\nu)$ in $\mathcal{M}^c(\nu)$. Hence, we can show

$$\begin{aligned}
&\|\lambda h_2(\boldsymbol{B}(\nu))\|_1 - \|\lambda h_2(\widehat{\boldsymbol{B}}(\nu))\|_1 \\
\leq &\left\| \lambda \sum_{i=1}^{d} \left\{ CE_i(\boldsymbol{B}(\nu)) - \widehat{CE}_i(\widehat{\boldsymbol{B}}(\nu)) \right\} + \lambda \left\{ \boldsymbol{B}_{d+1}(\nu) - \widehat{\boldsymbol{B}}_{d+1}(\nu) \right\} \right\|_1 \\
\leq &\lambda \sum_{i=1}^{d} \left\| CE_i(\boldsymbol{B}(\nu)) - \widehat{CE}_i(\widehat{\boldsymbol{B}}(\nu)) \right\|_1 + \lambda \left\| \boldsymbol{B}_{d+1}(\nu) - \widehat{\boldsymbol{B}}_{d+1}(\nu) \right\|_1.
\end{aligned} \tag{D.23}$$

Recall the close form of causal effects in Def. 4.5 derived in § 5.1 under the LSEM model. We have

$$DE_i(\boldsymbol{B}; g) = \theta_i,$$

where θ_i presents the weight of the direct edge $g(\boldsymbol{Z})_i \to Y$ according to (1) and Def. 4.5. In addition, the total causal effect can be quantified by the path method [see e.g., 43; 20] as

$$TE_i(\boldsymbol{B}; g) = \sum_{k=1}^{m_i} PE\{\pi_i^{(k)}\},$$

where $PE\{\pi_i^{(k)}\} = b_{i,l_1} \cdots b_{l_{\tau_k},(d+1)}$ is the causal effect of $g_i(\boldsymbol{Z})$ on Y through the directed path $\pi_i^{(k)} = \{i, l_1, \cdots, l_{\tau_k}, d+1\} \in \pi_i$ with length $\tau_k + 1$, and $b_{i,j}$ is the weight of the edge $g_i(\boldsymbol{Z}) \to g_j(\boldsymbol{Z})$ if it exists, and $b_{i,j} = 0$ otherwise, for $i, j \in \{1, \cdots, d\}$, and $b_{l_{\tau_k},(d+1)} = \theta_{l_{\tau_k}}$ as the direct edge from $g_{l_{\tau_k}}(\boldsymbol{Z})$ to Y. Both TE_i and DE_i can be explicitly calculated given a matrix $\boldsymbol{B}$ under a selector g. We denote their estimates as $\widehat{TE}_i$ and $\widehat{DE}_i$ given the estimated matrix $\widehat{\boldsymbol{B}}$ and g. Using the direct causal effects as an example, we have (D.23) be further bounded by

$$\begin{aligned}
&\lambda \sum_{i=1}^{d} \left\| CE_i(\boldsymbol{B}(\nu)) - \widehat{CE}_i(\widehat{\boldsymbol{B}}(\nu)) \right\|_1 + \lambda \left\| \boldsymbol{B}_{d+1}(\nu) - \widehat{\boldsymbol{B}}_{d+1}(\nu) \right\|_1 \\
\leq &C_1 \lambda \sum_{i=1}^{d} \left\| \boldsymbol{B}_i(\nu) - \widehat{\boldsymbol{B}}_i(\nu) \right\|_1 + \lambda \left\| \boldsymbol{B}_{d+1}(\nu) - \widehat{\boldsymbol{B}}_{d+1}(\nu) \right\|_1 \\
\leq &\max\{C_1, 1\} \lambda \sum_{i=1}^{d+1} \left\| \boldsymbol{B}_i(\nu) - \widehat{\boldsymbol{B}}_i(\nu) \right\|_1,
\end{aligned} \tag{D.24}$$

for some constant $C_1 > 0$. Recall $\|\boldsymbol{\mu}(\nu)\|_2 = \sum_{j=1}^{d+1} \|\widehat{\boldsymbol{B}}_j(\nu) - \boldsymbol{B}_j(\nu)\|_2^2$. Combining (D.23) and (D.24), we have

$$\begin{aligned}
&\|\lambda h_2(\boldsymbol{B}(\nu))\|_1 - \|\lambda h_2(\widehat{\boldsymbol{B}}(\nu))\|_1 \leq \max\{C_1, 1\} \lambda \sum_{i=1}^{d+1} \left\| \boldsymbol{B}_i(\nu) - \widehat{\boldsymbol{B}}_i(\nu) \right\|_1 \\
\leq &\max\{C_1, 1\} \lambda \|\boldsymbol{\mu}(\nu)\|_1 \leq \max\{C_1, 1\} \lambda (\|\boldsymbol{\mu}(\nu)_{\mathcal{M}(\nu)}\|_1 + \|\boldsymbol{\mu}(\nu)_{\mathcal{M}^c(\nu)}\|_1) \\
\leq &2 \max\{C_1, 1\} \lambda \sqrt{s_0(d+1)} \|\boldsymbol{\mu}(\nu)_{\mathcal{M}(\nu)}\|_2,
\end{aligned} \tag{D.25}$$

where the last inequality follows Equation (67) in Shi & Li [33]. Together with (D.22), we have

$$\begin{aligned}
&\kappa^* n \sum_{j=1}^{d+1} \|\boldsymbol{B}_j(\nu) - \widehat{\boldsymbol{B}}_j(\nu)\|_2^2 - 2 \sum_{j=1}^{d+1} \kappa(s_j(\nu) + \widehat{s}_j(\nu)) \log n \\
= &\kappa^* n \|\boldsymbol{\mu}(\nu)\|_2 - 2 \sum_{j=1}^{d+1} \kappa(s_j(\nu) + \widehat{s}_j(\nu)) \log n \leq 2 \max\{C_1, 1\} \lambda \sqrt{s_0(d+1)} \|\boldsymbol{\mu}(\nu)_{\mathcal{M}(\nu)}\|_2,
\end{aligned}$$

and thus

$$\kappa^*\|\boldsymbol{\mu}(\nu)\|_2 \leq \quad 2\sum_{j=1}^{d+1}\kappa(s_j(\nu)+\widehat{s}_j(\nu))\frac{\log n}{n}+2\max\{C_1,1\}\frac{\lambda\sqrt{s_0(d+1)}}{n}\|\boldsymbol{\mu}(\nu)_{\mathcal{M}(\nu)}\|_2. \tag{D.26}$$

Rearranging (D.26) leads to

$$\|\boldsymbol{\mu}(\nu)\|_2 \leq \mathcal{O}\big(\sqrt{\sum_{j=1}^{d+1}\frac{\log n}{n}(s_j(\nu)+\widehat{s}_j(\nu))}+\frac{\lambda\sqrt{s_0(d+1)}}{n}\big),$$

for arbitrary $\nu \in \Upsilon$. Hence the proof of the **Main Statement** is completed. The rest of the proofs follow the similar arguments in Proposition 1 of [33]. For λ satisfying the bound of $\mathcal{O}((n\log n)^{1/2})$, the consistency of the estimated matrix holds for arbitrary $\nu \in \Upsilon$. The proof of Thm. 5.1 is hence completed.

E Additional Simulation Results

In this section, we provide additional simulation configurations and results.

E.1 Simulation Configurations

Table E.1: Hyper-parameters information.

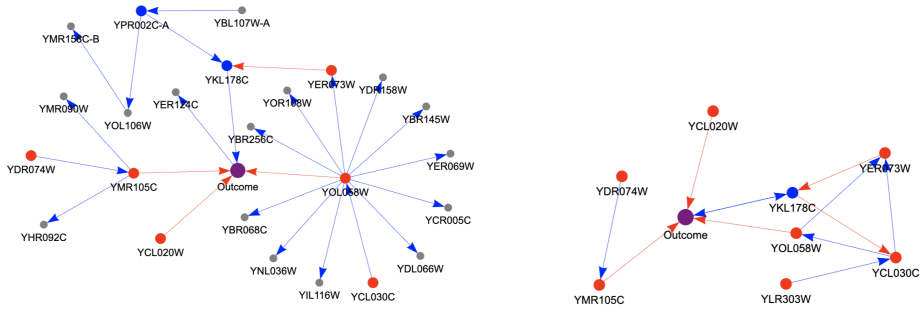
Hyper-parameters	Values
Maximum number of dual ascent steps in NOTEARS/NSCSL	100
Tolerance τ of acyclic constraint h_1 to be violated in NOTEARS/NSCSL	1e-8
Maximum of parameters for the hard constraints in NOTEARS/NSCSL	1e+16
L_1 penalty term l in NOTEARS/NSCSL	0
Conditional independent testing in PC	"Fisher-Z"
Pruning threshold for all methods	0.3

E.2 More Real Data Analyses on Yeast Data

The estimated causal graph among candidate QTLs and the outcome is shown in Fig. E.1 under the proposed method and NOTEARS [46] for illustration. The purple node represents the outcome, the blue nodes indicate QTLs with a positive causal impact, and the red nodes denote QTLs with a negative impact. Grey nodes are noisy QTLs without causal impact. Blue and red arrows represent positive and negative causal links, respectively. Causal effects from candidate genes on the genetic variant YER124C in yeast gene data are summarized using NSCSL in Table E.2. Fig. E.1 demonstrates that the proposed algorithm can discover necessary and sufficient causal relationships with better performance compared to the current causal discovery benchmark. Specifically, all nodes with causal effects (either blue or red nodes in the causal graph) on the outcome are identified under NSCSL. Furthermore, the proposed algorithm identifies an additional gene, 'YLR303W', which is not found in NOTEARS. Here, 'YLR303W' is essential for sulfur amino acid synthesis [3], with an estimated total causal effect of -0.06 on the target gene expression, as shown in Table E.2. And 'YER124C' of interest is a daughter cell-specific protein involved in cell wall metabolism [8]. It has been shown that sulfur amino acid synthesis can influent cell wall metabolism [37; 9]. These indicate that NSCGL which additionally identified 'YLR303W' performs better than NOTEARS. These observations align with findings from our simulation studies, further supporting NSCSL's superiority in revealing important causal features.

Table E.2: Summary of candidate genes that affect the variant YER124C in yeast gene data by NSCSL.

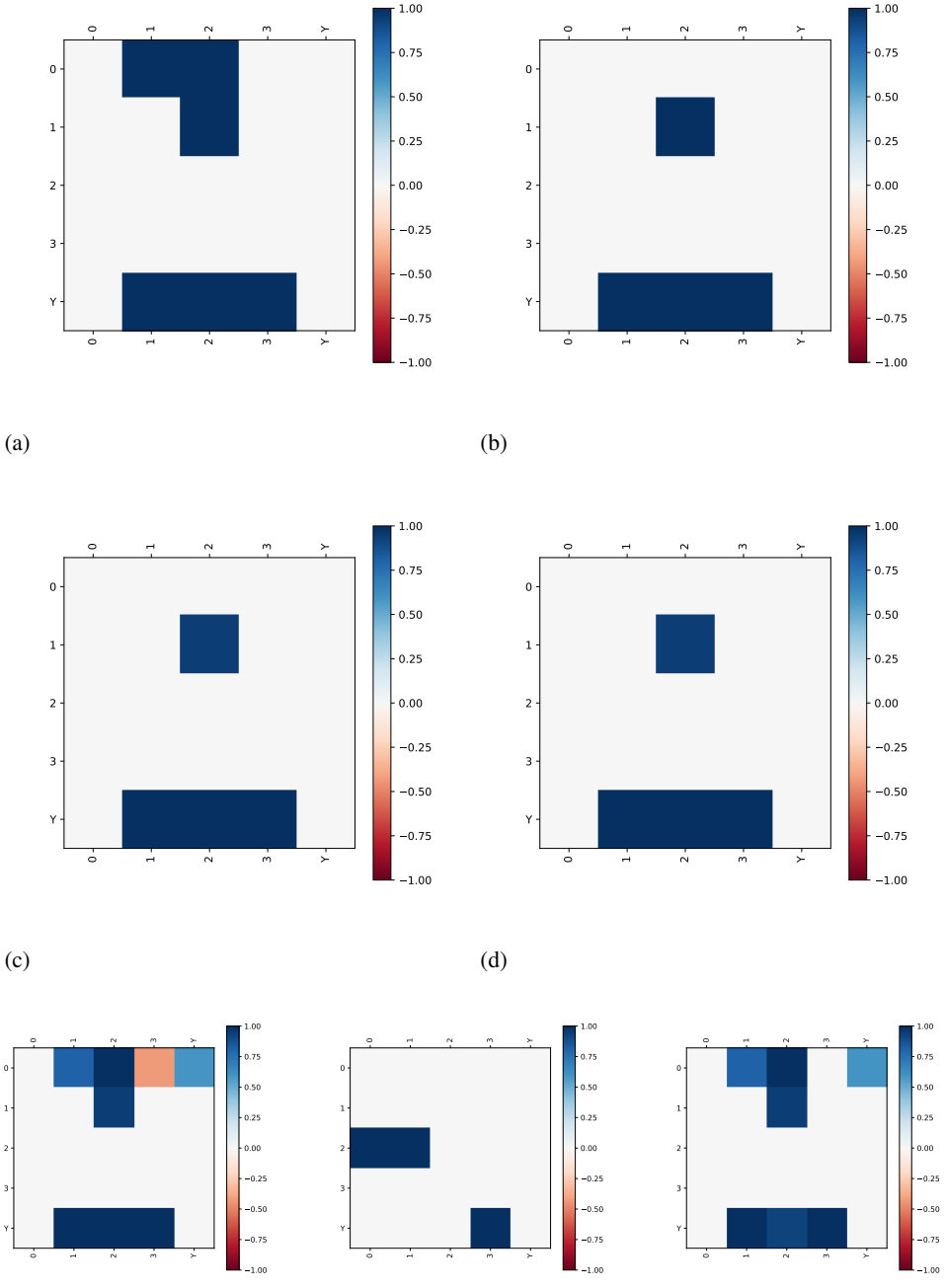
Gene Code	Gene Function	Direct Effect	Total Effect
YOL058W	Arginosuccinate synthetase	-0.20	-0.22
YCL020W	Genotype regulators	-0.26	-0.26
YDR074W	Trehalose-6-phosphate phosphatase	0.0	-0.15
YMR105C	Phosphoglucomutase	-0.28	-0.28
YKL178C	Cell surface a factor receptor	0.06	0.07
YLR303W	Required for sulfur amino acid synthesis	0.0	-0.06
YCL030C	Multifunctional enzyme	0.0	-0.22
YER073W	Mitochondrial aldehyde dehydrogenase	0.0	-0.06



(a) (b)

Figure E.1: Causal graphs for candidate genes that affect variant YER124C in yeast: (a). the estimated graph $\widehat{\mathcal{G}}$ by NOTEARS (benchmark); (b). the estimated graph $\widehat{\mathcal{G}}$ by NSCSL using TE.

E.3 Additional Simulation Results: True and Estimated Matrix

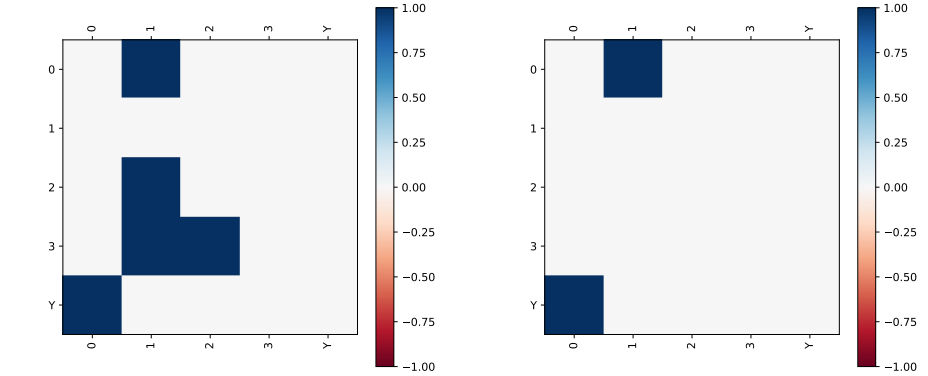


(a) (b)

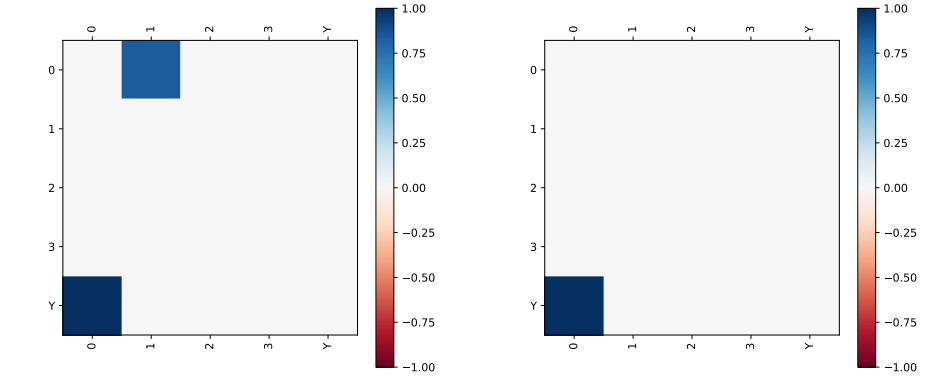
(c) (d)

(e) (f) (g)

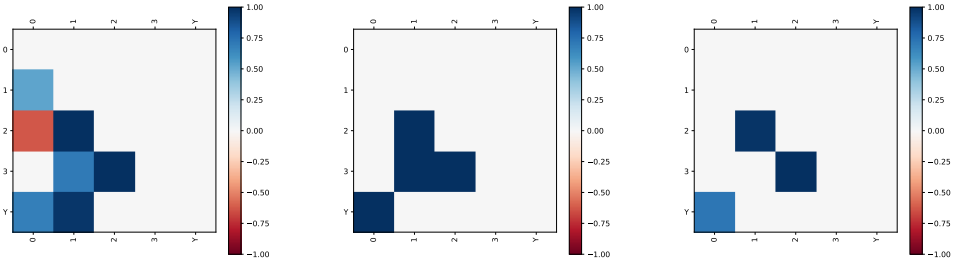
Figure E.2: Estimated matrix under S1 ($n = 20$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.



(a) (b)

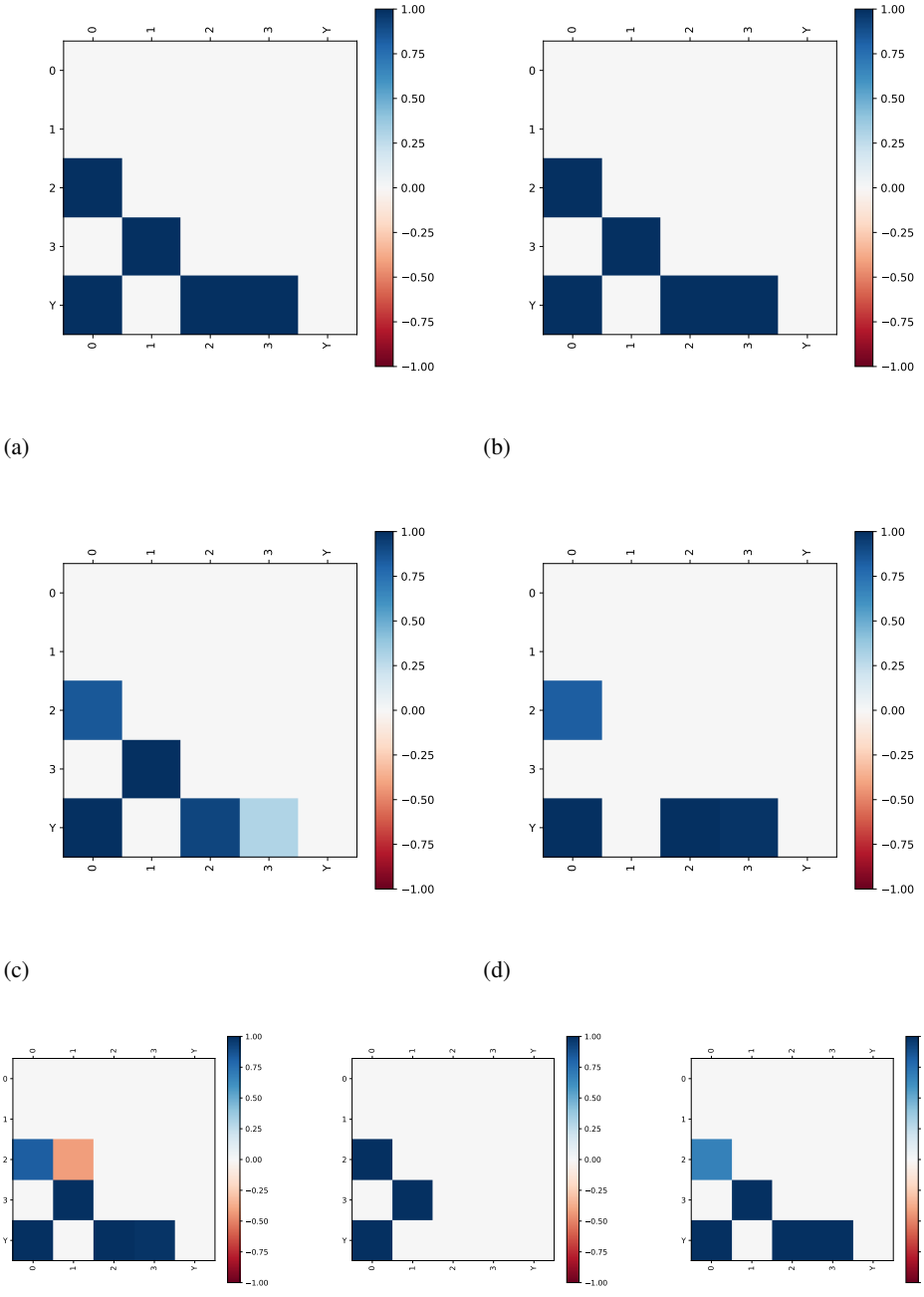


(c) (d)



(e) (f) (g)

Figure E.3: Estimated matrix under S2 ($n = 20$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.



(a) (b)

(c) (d)

(e) (f) (g)

Figure E.4: Estimated matrix under S3 ($n = 20$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.

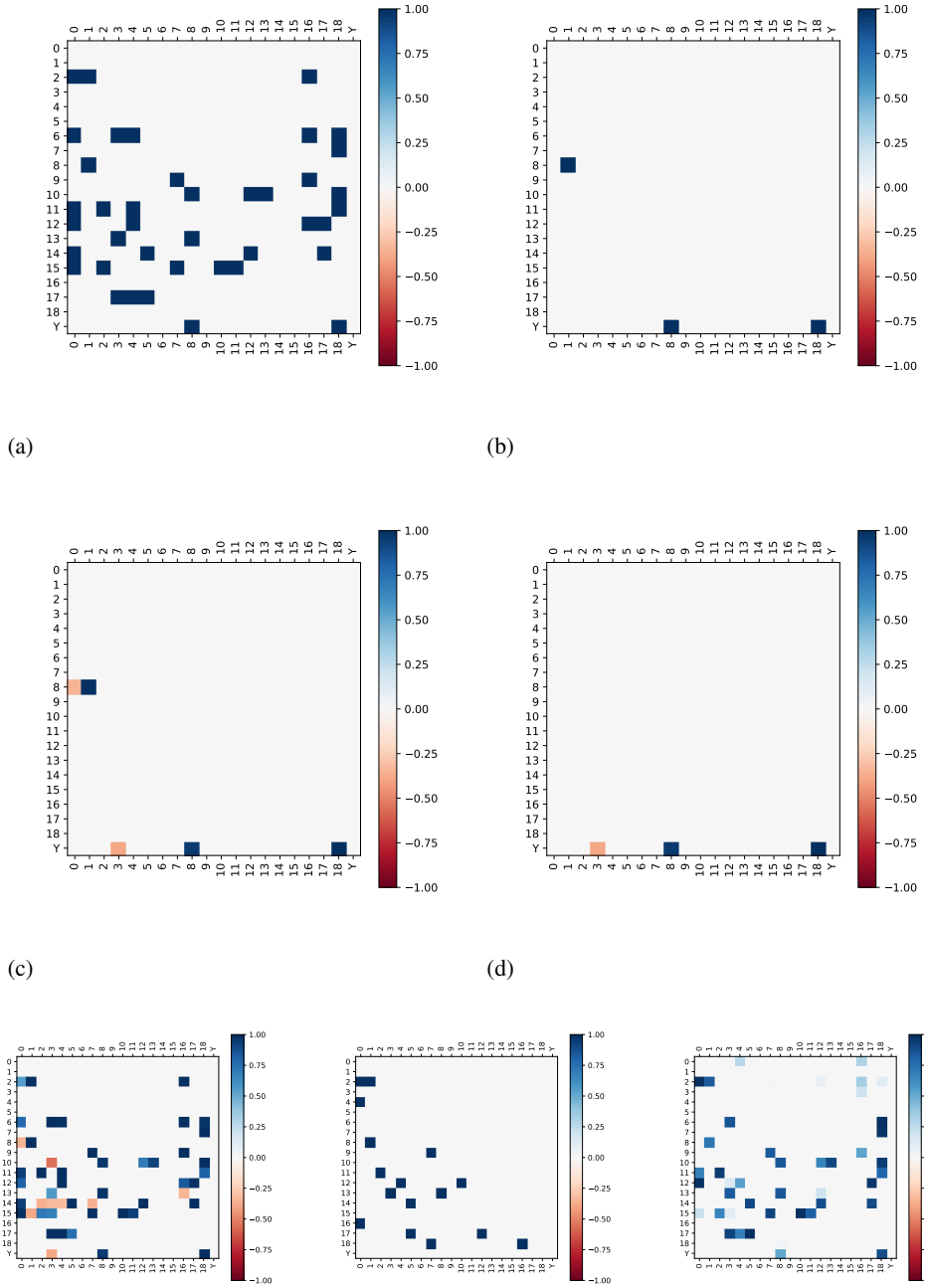
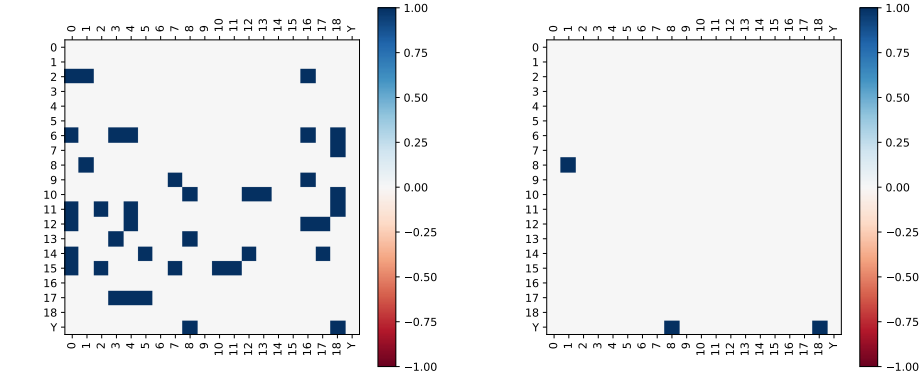
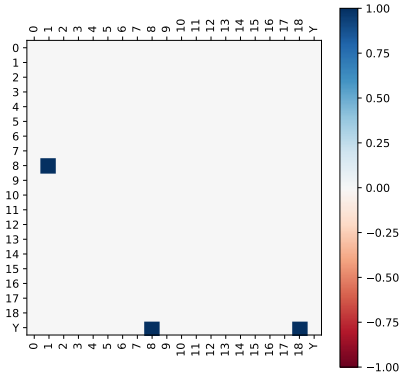


Figure E.5: Estimated matrix under S4 ($n = 100$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.

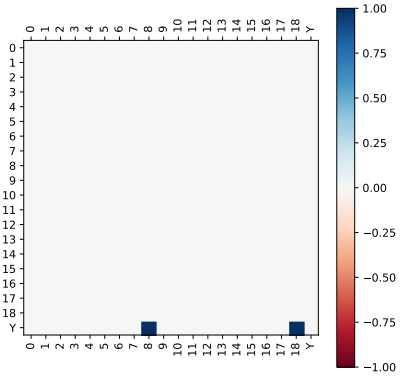


(a)

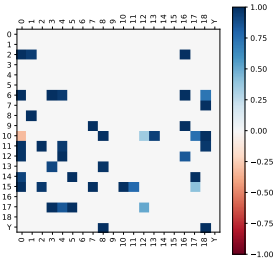
(b)



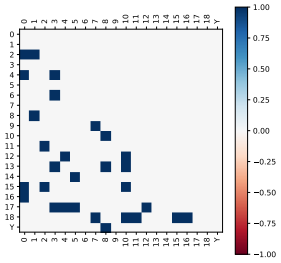
(c)



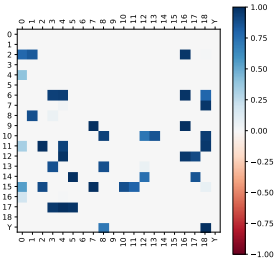
(d)



(e)



(f)



(g)

Figure E.6: Estimated matrix under S4 ($n = 300$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.

E.4 Additional Simulation Results: True and Estimated Graphs

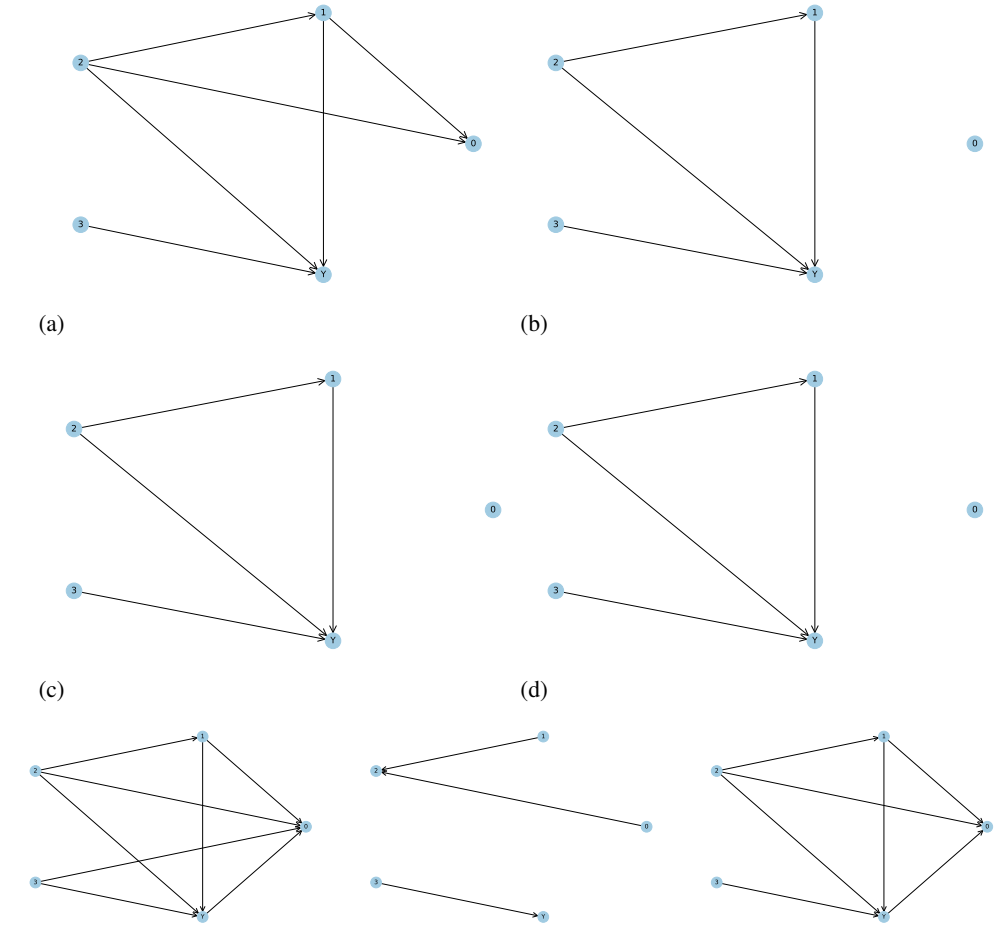


Figure E.7: Graphs under S1 ($n = 20$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.

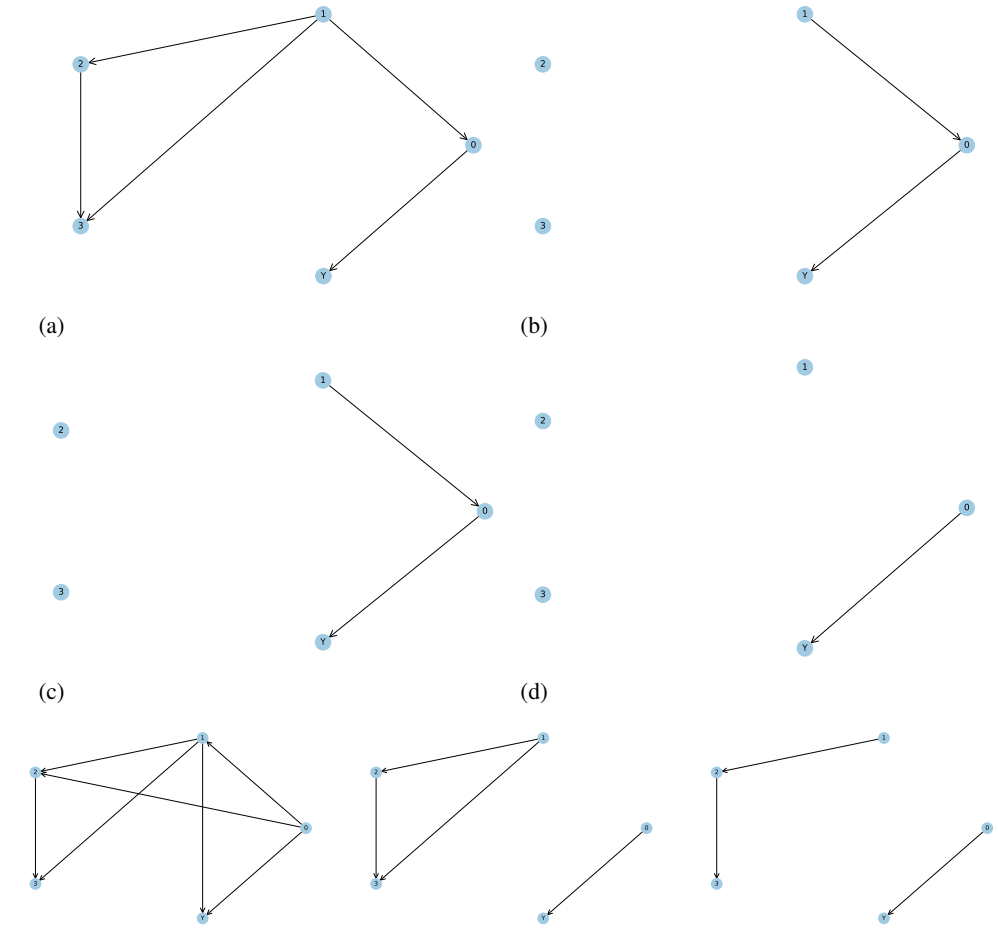


Figure E.8: Graphs under S2 ($n = 20$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.

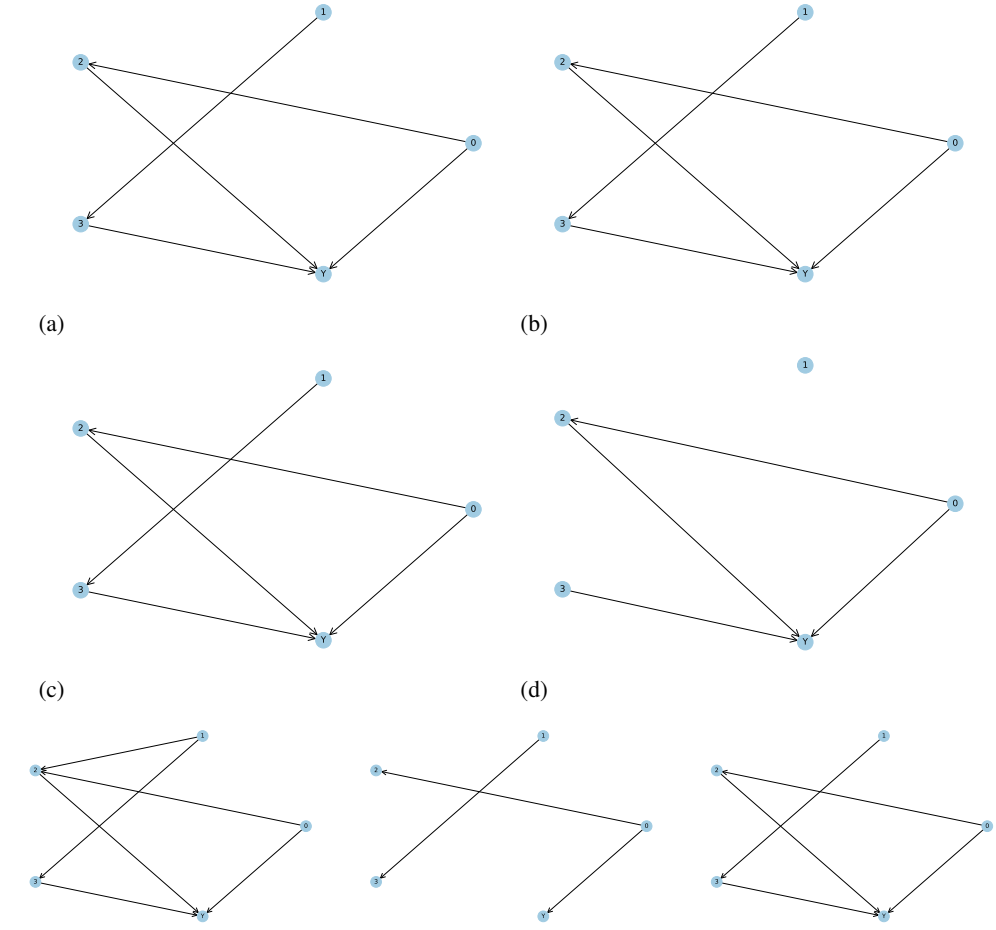


Figure E.9: Graphs under S3 ($n = 20$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.

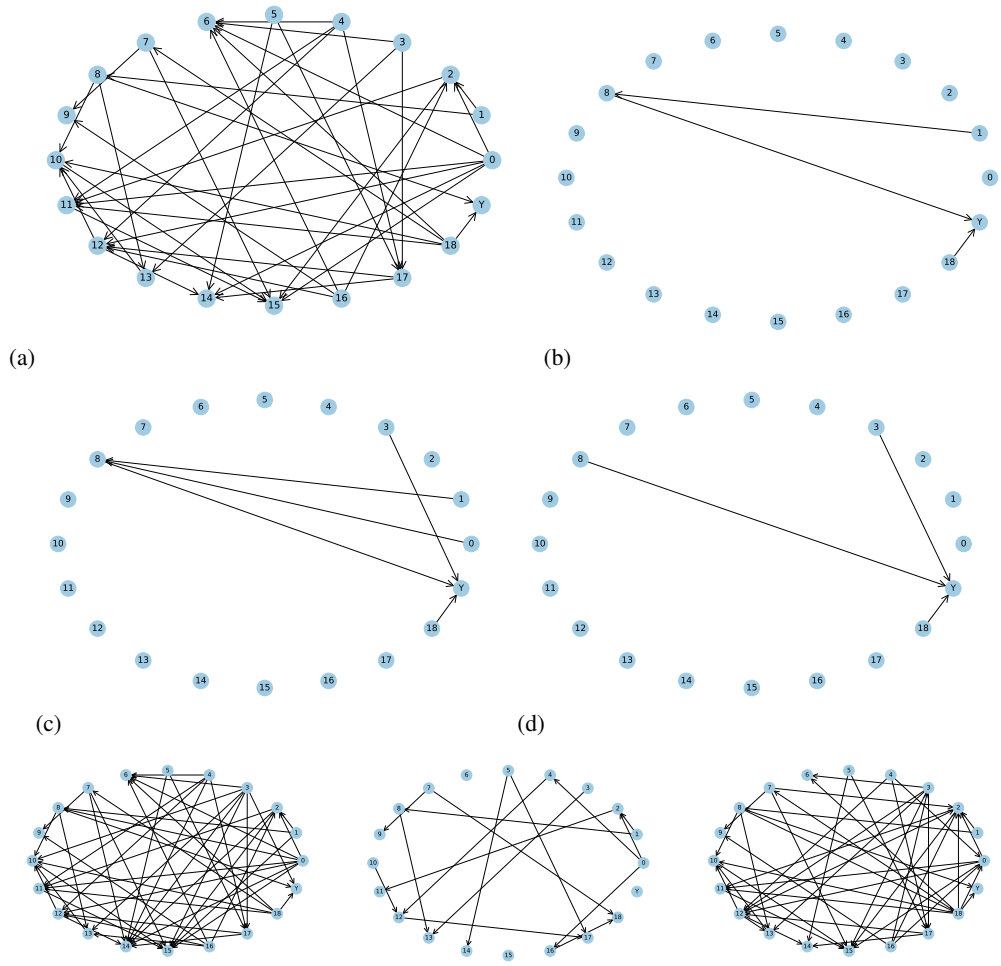


Figure E.10: Graphs under S4 ($n = 100$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.

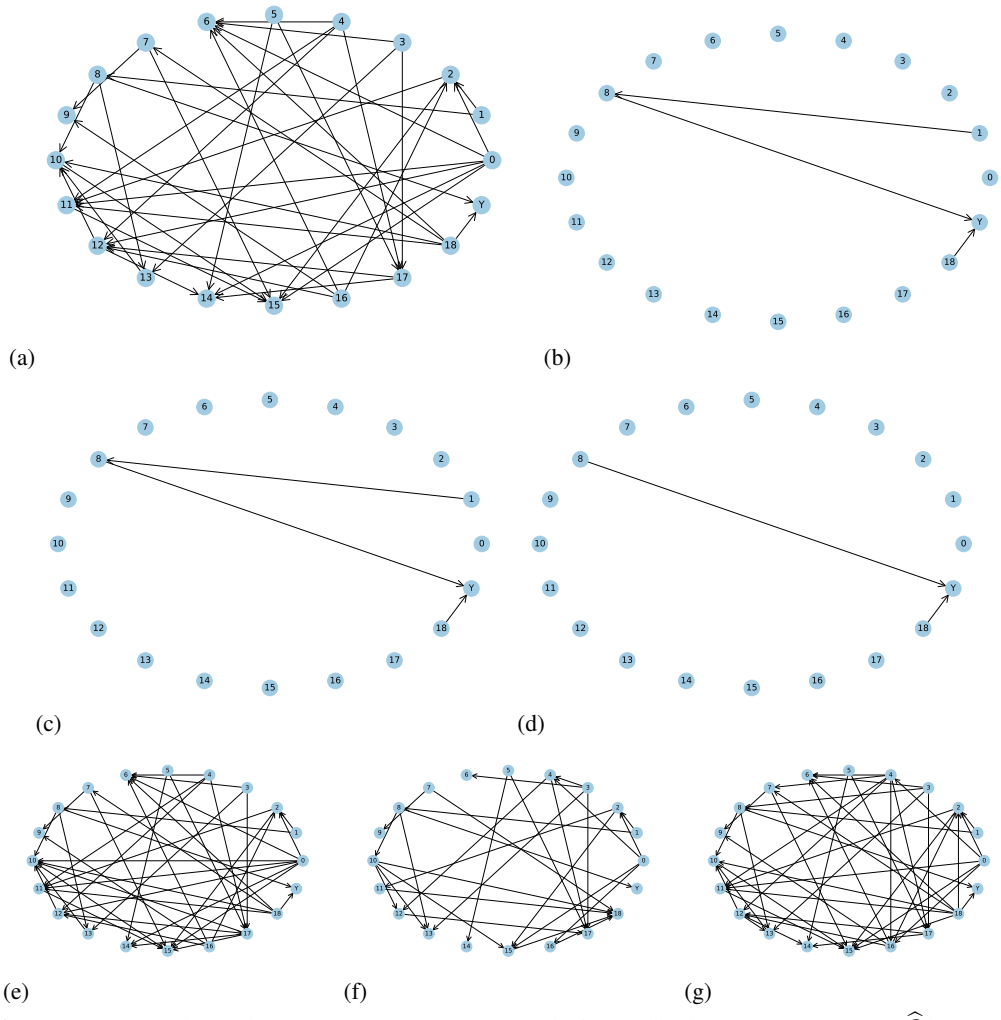


Figure E.11: Graphs under S4 ($n = 300$): (a). true whole graph; (b). true NSCG; (c). $\widehat{\mathcal{G}}$ by NSCSL with TE; (d). $\widehat{\mathcal{G}}$ by NSCSL with DE; (e). $\widehat{\mathcal{G}}$ by NOTEARS; (f). $\widehat{\mathcal{G}}$ by PC; (g). $\widehat{\mathcal{G}}$ by LiNGAM.

References

[1] Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010.

[2] Brem, R. B. and Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102(5):1572–1577, 2005.

[3] Brzywczy, J. and Paszewski, A. Role of o-acetylhomoserine sulfhydrylase in sulfur amino acid synthesis in various yeasts. *Yeast*, 9(12):1335–1342, 1993.

[4] Bühlmann, P., Peters, J., Ernest, J., et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

[5] Cai, H., Song, R., and Lu, W. Anoce: Analysis of causal effects with multiple mediators via constrained structural learning. In *International Conference on Learning Representations*, 2020.

[6] Chakrabortty, A., Nandy, P., and Li, H. Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*, 2018.

[7] Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[8] Colman-Lerner, A., Chin, T. E., and Brent, R. Yeast cbk1 and mob2 activate daughter-specific genetic programs to induce asymmetric cell fates. *Cell*, 107(6):739–750, 2001.

[9] de Melo, A. T., Martho, K. F., Roberto, T. N., Nishiduka, E. S., Machado Jr, J., Brustolini, O. J., Tashima, A. K., Vasconcelos, A. T., Vallim, M. A., and Pascon, R. C. The regulation of the sulfur amino acid biosynthetic pathway in cryptococcus neoformans: the relationship of cys3, calcineurin, and gpp2 phosphatases. *Scientific Reports*, 9(1):11923, 2019.

[10] Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*, 2021.

[11] Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1551–1560, 2018.

[12] Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schölkopf, B. Quantifying causal influences. 2013.

[13] Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pp. 2907–2916. PMLR, 2020.

[14] Kalisch, M. and Bühlmann, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[15] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Kumar, V. and Minz, S. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

[17] Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

[18] Lee, S. and Bareinboim, E. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.

[19] Lee, S. and Bareinboim, E. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in neural information processing systems*, 33:8565–8576, 2020.

[20] Nandy, P., Maathuis, M. H., Richardson, T. S., et al. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2): 647–674, 2017.

[21] Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.-S., and Wen, J.-R. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12700–12710, 2021.

[22] Pearl, J. et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

[23] Pearl, J. et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[24] Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[25] Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. 2014.

[26] Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[27] Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3(2):121–129, 2017.

[28] Rolland, P., Cevher, V., Kleindessner, M., Russell, C., Janzing, D., Schölkopf, B., and Locatello, F. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.

[29] Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[30] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[31] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[32] Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

[33] Shi, C. and Li, L. Testing mediation effects using logic of boolean matrices. *Journal of the American Statistical Association*, pp. 1–14, 2021.

[34] 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.

[35] Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V., and Wimberly, F. Constructing bayesian network models of gene expression networks from microarray data. 2000.

[36] Spirtes, P. L., Meek, C., and Richardson, T. S. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

[37] Takahashi, H., Braby, C. E., and Grossman, A. R. Sulfur economy and cell wall biosynthesis during sulfur limitation of chlamydomonas reinhardtii. *Plant physiology*, 127(2):665–673, 2001.

[38] Tang, K., Huang, J., and Zhang, H. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*, 2020.

[39] Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

[40] Van de Geer, S. and Bühlmann, P. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. 2013.

[41] Vowels, M. J., Camgoz, N. C., and Bowden, R. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys (CSUR)*, 2021.

[42] Wang, Y. and Jordan, M. I. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021.

[43] Wright, S. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

[44] Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019.

[45] Zhang, W., Wu, T., Wang, Y., Cai, Y., and Cai, H. Towards trustworthy explanation: On causal rationalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 41715–41736. PMLR, 2023.

[46] Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pp. 9472–9483, 2018.

[47] Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

[48] Zhu, S., Ng, I., and Chen, Z. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2019.