Identifying Causal Mechanism Shifts Under Additive Models with Arbitrary Noise

Yewei Xia[1,*], **Xueliang Cui**[2,3,*], **Hao Zhang**[2,†], **Yixin Ren**[1], **Feng Xie**[4], **Jihong Guan**[5], **Ruxin Wang**[2,†], **Shuigeng Zhou**[1,†]

[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai, China

[2]SIAT, Chinese Academy of Sciences, Shenzhen, China

[3]Southern University of Science and Technology, Shenzhen, China

[4]Department of Applied Statistics, Beijing Technology and Business University, Beijing, China

[5]Department of Computer Science and Technology, Tongji University, Shanghai, China

{ywxia23, yxren21}@m.fudan.edu.cn, {xl.cui2, h.zhang10, rx.wang}@siat.ac.cn, fengxie@btbu.edu.cn, jhguan@tongji.edu.cn, sgzhou@fudan.edu.cn

Abstract

In many real-world scenarios, the goal is to identify variables whose causal mechanisms change across related datasets. For example, detecting abnormal root nodes in manufacturing, and identifying key genes that influence cancer by analyzing differences in gene regulatory mechanisms between healthy individuals and cancer patients. This can be done by recovering the causal structure for each dataset independently and then comparing them to identify differences, but the performance is often suboptimal. Typically, existing methods directly identify causal mechanism shifts based on linear additive noise models (ANMs) or by imposing restrictive assumptions on the noise distribution. In this paper, we introduce CMSI, a novel and more general algorithm based on nonlinear ANMs that identifies variables with shifting causal mechanisms under arbitrary noise distributions. Evaluated on various synthetic datasets, CMSI consistently outperforms existing baselines in terms of F1 score. Additionally, we demonstrate CMSI's applicability on gene expression datasets of ovarian cancer patients at different disease stages.

1 Introduction

Causal discovery aims to uncover underlying causal relationships between variables of interest in observational data [Chen *et al.*, 2024a; Zhang *et al.*, 2024], which play a significant role in multiple disciplines such as genomic analysis [Xu *et al.*, 2024] and epidemiology [Robins *et al.*, 2000]. Numerous causal discovery algorithms have been proposed, including prominent methods such as PC [Spirtes *et al.*, 2000] and GES [Meek, 1997]. More recently, methods based on properly defined Structural Causal Models (SCMs) have been proposed to distinguish the correct graph underlying the observational data, by adding additional assumptions on the functional class of the SCM [Hoyer *et al.*, 2008; Zhang *et al.*, 2021; Zhang *et al.*, 2022].

In practice, the ultimate goal in many applications is not to recover the entire underlying DAG, but rather to detect the changes in the causal mechanisms among multiple related domains. For example, one application involves identifying the root causes of faults in large-scale manufacturing and industrial control systems [Bogatinovski *et al.*, 2021]. Another important application lies in identifying differences in gene regulatory networks between healthy individuals and cancer patients. The key genes that explain these differences may provide valuable information on potential therapeutic targets for specific types of cancer [Hudson *et al.*, 2009]. In these applications, compared with the size of the whole causal graph, the number of variables whose causal mechanisms change among domains tends to be relatively sparse. As a result, estimating causal graphs for each individual domain and subsequently detecting changes is inefficient and redundant. Therefore, it is of significant importance to develop a practical method for directly detecting changes in causal mechanisms across different environments.

Most existing methods for identifying causal mechanism shifts rely on restrictive assumptions within the functional class of the SCM, which can limit their applicability in real-world scenarios. For instance, both DCI [Wang *et al.*, 2018] and the method proposed by [Varici *et al.*, 2021] assume that the underlying SCM adheres to a linear ANM with Gaussian noise. To relax the linearity constraint, UT-IGSP [Squires *et al.*, 2020] introduces a nonparametric approach for identifying intervention targets. However, this method depends on nonparametric conditional independence (CI) tests, which are computationally expensive and inefficient. More recently, iSCAN [Chen *et al.*, 2024b] leverages the Jacobian of the score function of the mixture distribution to detect mechanism shifts. Despite its advancements, this approach imposes strict assumptions on the noise distribution in the ANM, which further restricts its applicability in practical scenarios.

Building on recent advancements in identifying causal relationships within multivariate ANMs with arbitrary noise distributions [Montagna *et al.*, 2023], we propose a novel method for detecting shifts in causal mechanisms that is applicable to arbitrary noise distributions. Specifically, we es-

tablish a connection between the score function of the mixture distribution across environments and the identification of variables whose causal mechanisms shift under arbitrary noise distributions. Based on this theoretical insight, we develop a practical algorithm that iteratively detects common leaf variables across all environments and evaluates whether their causal mechanisms have shifted by leveraging the score function. In summary, our main contributions are as follows:

- We prove that the score function of the mixture distribution unveils information to detect mechanism shifts for the common leaf nodes, which does not rely on any structural assumptions on the individual DAGs or parametric assumptions on noise distributions.
- We propose an algorithm that iteratively selects a common leaf variable among the DAGs in different environments and evaluates whether such variable is shifted.
- We evaluate the performance and applicability of our method by extensive experiments on both synthetic and ovarian cancer datasets. The results demonstrate the superiority of our method.

The remainder of this paper is organized as follows: In Section 2, we list the previous works in related topics. In Section 3, we introduce the fundamental notions and background. In Section 4, we present our method for detecting shifted variables. In Section 5, we evaluate the performance of our algorithm on both synthetic and ovarian cancer datasets. In Section 6, we summarize the content of this paper.

2 Related Work

Mechanism Shifts Identification. In the past decade, various approaches have been proposed to identify mechanism changes across environments. [Zhao *et al.*, 2014; Yuan *et al.*, 2017; Liu *et al.*, 2017] propose algorithms for directly detecting the changes by estimating the difference between two precision matrices. However, these methods are only suitable for undirected graphs. For directed graphs that provide more essential information, [Li *et al.*, 2023] offers a potential approach to detect heterogeneous functional relationships between a variable X_j and its parents. However, the applicability of their method is limited, as it relies on the assumption of a common distribution for covariates across different environments. This assumption is potentially violated if the mechanism shifts have affected the ancestors of X_j. Additionally, DCI [Wang *et al.*, 2018] estimates changes by testing the invariance of regression coefficients and noise variances, while [Varici *et al.*, 2021] repeatedly identifies intervention sites in subsets of variables by comparing precision matrices, both of which are restricted to linear additive noise models (ANMs) with Gaussian noise. To address the linearity constraint, UT-IGSP [Squires *et al.*, 2020] introduces a nonparametric method for identifying intervention targets. However, this approach relies on nonparametric conditional independence (CI) tests, which can be computationally expensive and inefficient. iSCAN [Chen *et al.*, 2024b] detects shifts by leveraging the Jacobian of the score function of the mixture distribution. Nonetheless, this method assumes that the noise distribution in the ANM satisfies specific conditions.

Application. Identification of causal mechanism shifts has wide-ranging applications across numerous domains. In biology, understanding differences in gene regulatory networks across different populations in biological systems can provide crucial insights into disease mechanisms and potential therapeutic targets [Hudson *et al.*, 2009; Pimanda *et al.*, 2007]. Specially, in the analysis of EEG signals it is of interest to detect neurons or different brain regions that interact differently when the subject is performing different activities [Sanei and Chambers, 2013]. In economics, [Shi *et al.*, 2020] apply a new method to detect shifts in Granger causality, uncovering structural breaks in the US money-income relationship. Another application area lies in fault detection in large-scale Internet of things and cloud applications [Bogatinovski *et al.*, 2021]. Other examples include epidemiology [Robins *et al.*, 2000], medicine [Plis *et al.*, 2010], etc. These diverse applications highlight the importance of developing robust and efficient methods for causal mechanism shift detection.

Score Matching. Score matching, a recently developed parameter learning method, is particularly effective for high-dimensional density models with intractable partition functions [Lyu, 2012; Ren *et al.*, 2025]. Recent advancements have extended score matching to generative modeling, resulting in score-based generative models that achieve remarkable performance in image generation and other domains [Song and Ermon, 2019; Song *et al.*, 2020]. This method has also been explored in causal discovery. For instance, SCORE [Rolland *et al.*, 2022] leverages the Jacobian of the score function to identify the topological order in additive noise models (ANMs) with Gaussian noise. NoGAM [Montagna *et al.*, 2023] builds on this by learning causal relations through regression on the score function, relaxing the noise distribution constraints of iSCAN. Additionally, score matching's ability to capture subtle distribution changes makes it valuable for detecting distribution shifts. For example, iSCAN [Chen *et al.*, 2024b] uses the diagonal of the Jacobian of the score function to identify shifts in causal mechanisms.

3 Preliminaries and Background

In this section, we first introduce the notation and assumptions that will be used throughout the paper. We then conclude by formalizing the problem setting.

Definition 1 (Structural Causal Model (SCM))**.** *Let $V = \{1, ..., d\}$ be the vertices of a directed acyclic graph G, and $X = \{X_1, ..., X_d\}$ be a d-dim vector of random variables, in one-to-one correspondence with V. An SCM $\mathcal{M} = (X, f, \mathbb{P}_N)$ over d variables is a set of d structural equations with the same form:*

$$\forall i \in V, X_i = f_i(X_{\mathbf{PA}_i}, N_i), \tag{1}$$

where $\mathbf{PA}_i \subseteq V \setminus \{i\}$ are the direct parents of vertex i in the underlying DAG of the SCM. $\mathbb{P}_N$ is the joint distribution of noise terms $N = \{N_1, ..., N_d\}$ which are assumed to be jointly independent[1].

[1]Note that we assume there is no latent confounders.

Definition 2 (Causal Mechanism)**.** *Every SCM induces a joint distribution $\mathbb{P}(X)$ of X and admits the following factorization:*

$$\mathbb{P}(X) = \prod_{i=1}^{d} \mathbb{P}(X_i | X_{\mathbf{PA}_i}), \tag{2}$$

where $\mathbb{P}(X_i|X_{\mathbf{PA}_i})$ is defined as the causal mechanism of X_i.

Definition 3 (Independent Causal Mechanisms (ICM) Principle)**.** *A change in $\mathbb{P}(X_j|\mathbf{PA}_j)$ has no effect on and provides no information on $\mathbb{P}(X_k|\mathbf{PA}_k)$ for any $k \neq j$.*

Denote an underlying SCM $\mathcal{M}^*$ with DAG structure G^* and joint distribution $\mathbb{P}^*(X) = \prod_{j=1}^{d} \mathbb{P}^*(X_j|X_{\mathrm{PA}_j^*})$. An environment arises from intervening causal mechanisms of a subset of X. Formally, we have the following definition.

Definition 4 (Environment)**.** *An environment $\mathcal{E}_h$ is derived from an underlying ANM $\mathcal{M}^* = (X, f, \mathbb{P}_N)$ by applying interventions on an unknown subset I of variables X. According to the ICM Principle (Def. 3), then the new joint distribution $\mathbb{P}^h(X)$ can be factorized as follows:*

$$\mathbb{P}^h(X) = \prod_{i \in I} \mathbb{P}^h(X_i | X_{\mathbf{PA}_i^h}) \prod_{j \in V \setminus I} \mathbb{P}^*(X_j | X_{\mathbf{PA}_j^*}), \tag{3}$$

where $\mathbb{P}^h(X_i|X_{\mathbf{PA}_i^h})$ is the causal mechanism after intervention. Here, we allow an intervention to modify the causal mechanism of a variable while optionally removing part of its direct parents, i.e., $\mathbf{PA}_i^h \subseteq \mathbf{PA}_i^*$.

With Def. 4, we have the definition of the shifted variables.

Definition 5 (Shifted Variable)**.** *A shifted variable is a variable whose causal mechanism differs across different environments. Formally, variable X_i is a shifted variable, if there exists two environments $\mathcal{E}_a$ and $\mathcal{E}_b$, such that*

$$\mathbb{P}^a(X_i|\mathbf{PA}_i^a) \neq \mathbb{P}^b(X_i|\mathbf{PA}_i^b), \tag{4}$$

where $\mathbb{P}^a(X_i|\mathbf{PA}_i^a)$ and $\mathbb{P}^b(X_i|\mathbf{PA}_i^b)$ denotes the causal mechanisms of X_i in environment $\mathcal{E}_a$ and $\mathcal{E}_b$, respectively.

Definition 6 (Additive Noise Model (ANM))**.** *An additive noise model $\mathcal{M}^*$ is an SCM whose each structural equation has the following form:*

$$\forall i \in V, \qquad X_i = f_i(X_{\mathbf{PA}_i}) + N_i, \tag{5}$$

where f_i is referred to as the data generation function.

In this paper, we assume that all SCMs in different environments satisfy the ANM model. That is, intervention can modify function f_i, parent $\mathbf{PA}_i$ or noise distribution N_i. However, the new SCM after intervention still follows the ANM model. Next, we introduce a lemma from [Montagna *et al.*, 2023] which serves as the basis of our approach.

Lemma 1. *Given X be a d-dim random variables generated according to Def. 5 with underlying DAG G. Let $p(x)$ be the pdf of X and $s(x) = \nabla \log p(x)$ be the corresponding score function. Denote R_i be the residual of X_i by regressing on $X_{\setminus\{i\}}$, and then obtain $g(R_i)$ as the estimator for $s_i(X)$ by regressing on R_i. $\forall\, X_i \in X$, we have:*

$$\mathbb{E}\left[(s_i(X) - g(R_i))^2\right] = 0 \Longleftrightarrow \textit{node } i \textit{ is a leaf in } G. \tag{6}$$

Lemma 1 provides a sufficient and necessary condition for detecting leaf nodes for those models generated according to ANM. Notably, it requires no specific parametric assumption on the noise distribution. Finally, to conclude this section, we formally define the whole problem.

Problem definition. Given H datasets $\mathbf{X}^1, ..., \mathbf{X}^H$ sampled from K different environments where $\mathbf{X}^k \in \mathbb{R}^{m_h \times d}$ consists of m_h independent and identically distributed samples from environment $\mathcal{E}_h$, our task is to identify the shifted variables relative to H environments without any assumptions on the specific noise distribution in the ANM.

4 Method

In this section, we propose CMSI to identify mechanism shifts across environments, without any assumptions about the noise distribution in the ANM.

Let $\mathbf{X}$ be the row concatenation of all the datasets $\mathbf{X}^h$, i.e., $\mathbf{X} = [(\mathbf{X}^1)^\top | \cdots | (\mathbf{X}^H)^\top]^\top \in \mathbb{R}^{m \times d}$, where $m = \sum_{h=1}^{H} m_h$. The pooled data $\mathbf{X}$ represents an aggregation of data sampled from H heterogeneous environments. To account for this aggregation, we introduce the probability mass $w_h \triangleq \frac{m_h}{m}$, which represents the probability that an observation belongs to the environment $\mathcal{E}_h$, i.e., $\sum_{h=1}^{H} w_h = 1$. Let $\mathbb{Q}(X)$ denote the distribution of the mixture data with density function $q(x)$, i.e., $q(x) = \sum_{h=1}^{H} w_h p^h(x)$. We use $s^h(x) \triangleq \nabla \log p^h(x)$ to denote the score function of the joint distribution of environment h with density $p^h(x)$. Also, we use $s(x) \triangleq \nabla \log q(x)$ to denote the score function of the mixture distribution with density $q(x)$.

4.1 Main Results

Inspired by Lemma 1 that identifies the leaf nodes in a single environment via the score function. In the sequel, we will show that the score function of the mixture distribution $s(x) \triangleq \nabla \log q(x)$ can help reveal causal mechanism shifts among different environments. Firstly, we make the following assumptions on the underlying SCM.

Condition 1 ([Peters *et al.*, 2014])**.** *Given a bivariate model $X_i := N_i$ and $X_j := f_j(X_i) + N_j$, we call the SCM an identifiable bivariate ANM if the triple (f_j, p_{N_i}, p_{N_j}) does not solve the following differential equation for all pairs x_i, x_j with $f_j'(x_i)g''(x_j - f_j(x_i)) \neq 0$:*

$$\begin{aligned} k''' = k'' \left(-\frac{g''' f'}{g''} + \frac{f''}{f'} \right) - 2g'' f'' f' \\ + g' f''' + \frac{g' g''' f'' f'}{g''} - \frac{g'(f'')^2}{f'}. \end{aligned} \tag{7}$$

Here, $f := f_j$, $k := \log p_{N_i}$, $g := \log p_{N_j}$. To improve readability, the arguments $x_j - f_j(x_i), x_i$ and x_i of g, k and f, respectively, have been removed.

Assumption 1. *For all $j \in \mathbf{V}, i \in \mathbf{PA}_j$ and all sets $\mathbf{S} \subseteq \mathbf{V}$ with $\mathbf{PA}_j \setminus \{i\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_j \setminus \{i, j\}$, there is an $x_{\mathbf{S}}$ with $p_{\mathbf{S}}(x_{\mathbf{S}}) > 0$, s.t.*

$$(f_j(x_{\mathbf{PA}_j \setminus \{i\}}, X_i), \mathcal{L}\left(X_i \mid X_{\mathbf{S}} = x_{\mathbf{S}}\right), \mathcal{L}(N_j)) \tag{8}$$

Algorithm 1 Regression the Score on Residual (ReSR)

Input: Dataset $\mathbf{X}$.
Output: Estimator $\hat{g}(\mathbf{R})$ of the score $s(\mathbf{X})$.

1: $s(\mathbf{X}) \leftarrow$ estimate the score function of $\mathbf{X}$.
2: $\mathbf{R}_i \leftarrow$ residual by regressing $\mathbf{X}_i$ on $\mathbf{X}_{\setminus\{i\}},\ \forall i \in \mathbf{V}$.
3: $\hat{g}(\mathbf{R}) \leftarrow$ obtained by regressing $s(\mathbf{X})$ on $\mathbf{R}_i,\ \forall i \in \mathbf{V}$.
4: **return** $\hat{g}(\mathbf{R})$.

satisfies Condition 1. In particular, we require the noise variables to have non-vanishing densities and the functions f_j to be continuous and three times continuously differentiable.

Theorem 1. *For all $h \in [H]$, let G^h and $p^h(x)$ denote the underlying DAG structure and pdf of environment $\mathcal{E}_h$, respectively, and let $q(x)$ be the pdf of the mixture distribution of the H environments such that $q(x) = \sum_{h=1}^{H} w_h p^h(x)$. Also, let $s(x) = \nabla \log q(x)$ be the associated score function, $R_i := X_i - \mathbb{E}_{X_i \sim \mathbb{Q}(X_i)}[X_i | X_{\setminus\{i\}}]$, $g^*(R_i) := \mathbb{E}[s_i(X)|R_i]$. Then, under Assumption 1, we have: if l is a leaf in all DAGs G^h, then l is a shifted node if and only if:*

$$\mathbb{E}\left[\left(g^*(R_l) - s_l(X)\right)^2\right] > 0, \tag{9}$$

except for the case that the causal functions are identical across environments, i.e. $\forall h, h^* \in [H], f_l^h = f_l^{h^*}$.

The proof of Theorem 1 is presented in the appendix. Theorem 1 provides an effective approach for detecting mechanism shifts for common leaf nodes. As the induced graph by $\mathcal{M}^*$ is acyclic and each intervention will only optionally remove edges without adding some new edges, we can always find a common leaf node. Furthermore, after removing the common leaf node found, we can always find a new one, until the whole graph is empty. Consequently, all the shift nodes can be found in this procedure by Theorem 1.

The idea above is outlined in Alg. 2. Concretely, we first integrate the primary computational procedure of Lemma 1 into Alg. 1 as a component. In Alg. 1, to obtain the estimator $\hat{g}(\mathbf{R})$ of the score $s(\mathbf{X})$, we first regress each variable $\mathbf{X}_i$ against the remaining variables $\mathbf{X}_{\setminus\{i\}}$ to get the residual $\mathbf{R}_i$, and finally regress the score function $s(\mathbf{X})$ against the residuals $\mathbf{R}$. In Alg. 2, we first identify the leaf nodes of the underlying DAG for each environment based on Lemma 1 and then obtain the common leaf nodes L of all DAGs (Line 4-5). Subsequently, based on Theorem 1, we detect whether the common leaf nodes are shift nodes (Line 6-8). Repeat the above procedure until all variables have been detected (Line 9). Note that for each iteration, the common leaf nodes will not be involved in the subsequent identification of leaf nodes, regardless of whether they are shifted.

Computational Complexity. The computational complexity of both score function estimation and regression within a single environment $\mathcal{E}_h$ is $\mathcal{O}(dm_h^3)$. In Alg. 2, the primary computational cost stems from the score function estimation on the pooled data $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and regression on $\boldsymbol{X}$, each with a complexity of $\mathcal{O}(dm^3)$. In the worst-case scenario, where only one node is removed per iteration, the algorithm requires at most d iterations. Therefore, the total computational complexity of Alg. 2 is $\mathcal{O}(d^2m^3)$.

Algorithm 2 Causal Mechanism Shifts Identification (CMSI)

Input: Dataset $\mathbf{X}^1, ..., \mathbf{X}^H$.
Output: Estimated shifted variables set $\hat{\mathbf{I}}$.

1: **Initialization**: $\hat{\mathbf{I}} = \{\}, N = \{1, ..., d\}$.
2: $\mathbf{X} = \left[(\mathbf{X}^1)^\top | \cdots | (\mathbf{X}^H)^\top\right]^\top \in \mathbb{R}^{m \times d}$.
3: **while** $N = \emptyset$ **do**
4: $\quad \hat{g}(\mathbf{R}^h) \leftarrow \text{ReSR}(\mathbf{X}^h),\ \forall h \in [H]$.
5: $\quad L \leftarrow \cap_{h \in [H]} \{l \mid \mathbb{E}[(s_l(\mathbf{X}^h) - \hat{g}(R_l^h))^2] = 0, l \in \mathbf{V}\}$.
6: $\quad \hat{g}(\mathbf{R}) \leftarrow \text{ReSR}(\mathbf{X})$.
7: $\quad \hat{\mathbf{I}} \leftarrow \hat{\mathbf{I}} \cup \{X_i \mid \mathbb{E}[s_i(\mathbf{X}) - \hat{g}(R_i))^2] \neq 0,\ i \in L\}$.
8: $\quad$ remove $\mathbf{X}^h[:, L]$ and $\mathbf{X}[:, L], \forall h \in [H]$.
9: $\quad N \leftarrow N - \{L\}$.
10: **end while**
11: **return** $\hat{\mathbf{I}}$.

4.2 Practical Implementation

In this section, we present some practical implementation details of Alg. 2 used in the experiments.

Score Function Estimation. Given an environment $\mathcal{E}$ and its dataset $\mathbf{X} = \{x^1, ..., x^m\} \in \mathbb{R}^{m \times d}$. To estimate the corresponding score function $s(x) = \nabla \log p(x)$, Stein's identity [Stein, 1972] provides an estimator using:

$$\mathbb{E}_{p(x)}\left[h(x)\nabla \log p(x)^\top + \nabla h(x)\right] = 0, \tag{10}$$

where $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ s.t. $\lim_{x \to \infty} h(x)p(x) = 0$. In practice, we adopt a similar approach to the method in [Li and Turner, 2017] and present the estimator for the point-wise first-order partial derivative, corresponding to Eq. (10):

$$\hat{\mathbf{G}} = -(\mathbf{K} + \eta \mathbf{I})^{-1}\langle \nabla, \mathbf{K} \rangle, \tag{11}$$

where $\mathbf{H} = (h(x^1), \ldots, h(x^m)) \in \mathbb{R}^{d' \times m}$, $\mathbf{K} = \mathbf{H}^\top \mathbf{H}$, $\mathbf{K}_{ij} = \kappa(x^i, x^j) = h(x^i)^\top h(x^j)$, $\overline{\nabla h} = \frac{1}{m}\sum_{k=1}^{m} \nabla h(x^k)$, $\langle \nabla, \mathbf{K} \rangle = m\mathbf{H}^\top \overline{\nabla h}$, and $\eta \geq 0$ is a regularization parameter. $\mathbf{G} \equiv (\nabla \log p(x^1), \ldots, \nabla \log p(x^m))^\top \in \mathbb{R}^{m \times d}$ and $\hat{\mathbf{G}}$ is used to approximate $\mathbf{G}$.

Regression. In both stages of regression in Alg. 1, we utilize kernel ridge regression due to its capability to model nonlinear relationships. To enhance computational efficiency, we implement kernel ridge regression with NVIDIA's cuML library from the RAPIDS AI open-source software suite. The regularization coefficient in the kernel ridge regression is set to $\alpha = 0.1$, and we use the radial basis function (RBF) kernel with a width parameter of $\gamma = 0.1$.

Selection of Common Leaf. Before employing Theorem 1 to examine the shifted nodes, it is necessary to leverage Lemma 1 first to obtain the leaf nodes in each individual environment $\mathcal{E}_h$ and then deduce the common leaf nodes among all environments. Estimating $\mathbb{E}[(s_i^h(X) - \hat{g}(R_i^h))^2]$ is performed by computing the Mean Squared Error (MSE) between the prediction $\hat{s}(R_i^h)$ and the ground truth $s_i(\mathbf{X}^h)$, denoted as MSE_i^h. In practice, due to the limited number of samples, it is not feasible to select leaf nodes solely by $\text{MSE}_i^h = 0$ because MSE_i^h is susceptible to errors. Therefore, we first rank the nodes based on their individual MSE_i^h

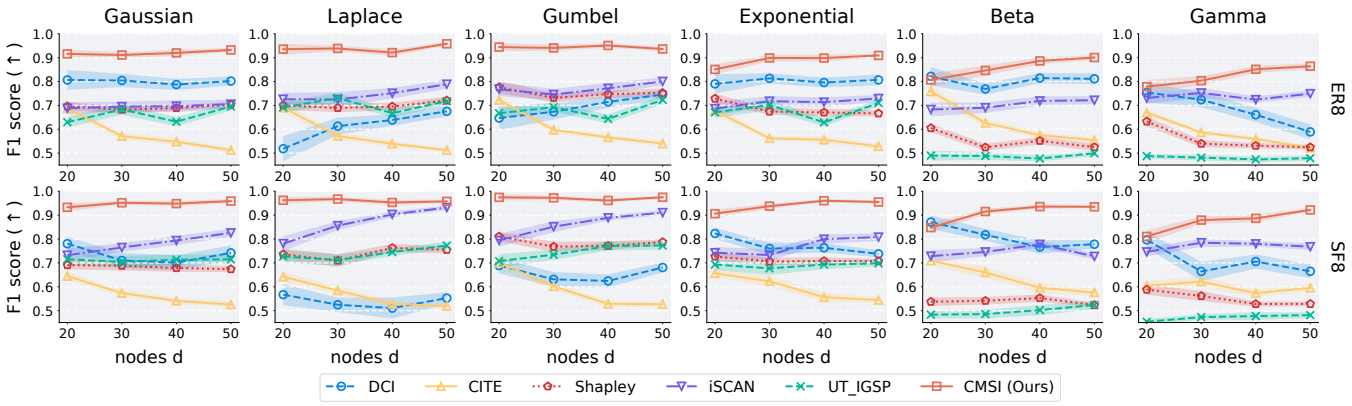

Figure 1: Simulation results on synthetic datasets with d nodes, $d \in \{20, 30, 40, 50\}$. The average degree of each node is set to 8. The x-axis represents the number of nodes, the top axis indicates the noise distribution, and the right-side axis labels the graph model type (Erdös-Rényi (ER)/Scale-Free (SF)). Each subplot shows the performance of CMSI and other baselines for a specific setting. The points indicate the average values obtained from 30 simulations. The shaded areas around the lines represent the standard error.

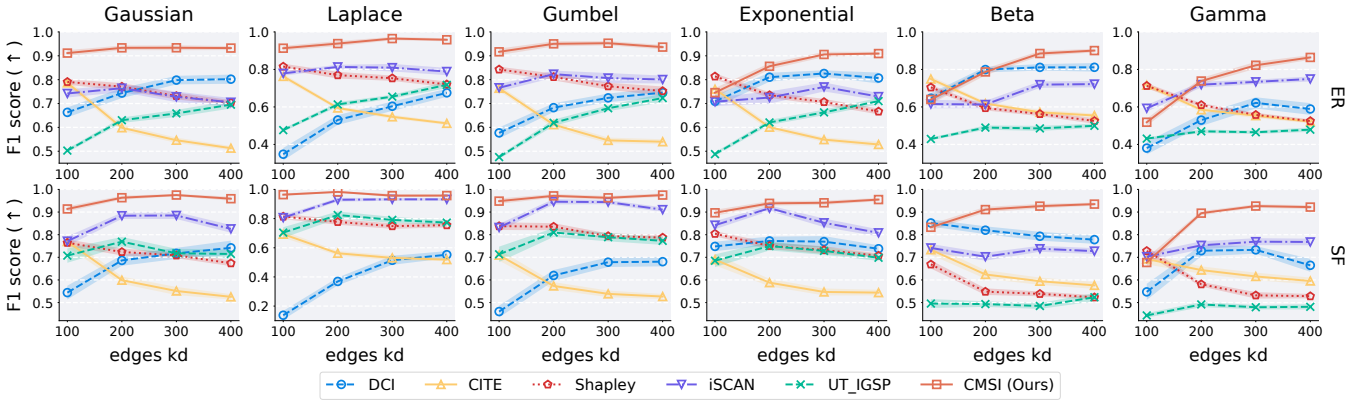

Figure 2: Simulation results on synthetic datasets with 50 nodes and kd edges, $kd \in \{100, 200, 300, 400\}$. The x-axis represents the number of edges, the top axis indicates the noise distribution, and the right-side axis labels the graph model type (Erdös-Rényi (ER)/Scale-Free (SF)). Each subplot shows the performance of CMSI and other baselines for a specific setting. The points indicate the average values obtained from 30 simulations. The shaded areas around the lines represent the standard error.

values in each environment $\mathcal{E}_h$. Subsequently, we select the node with the minimum sum of ranks across all environments as the common leaf.

Selection of Shifted Variables. Based on $\mathbf{X}$, the row concatenation of all datasets $\mathbf{X}^h$, we compute the score function $s(\mathbf{X})$ and its estimator $\hat{s}(\mathbf{R})$ based on the residuals $\mathbf{R}$. Similarly to the selection of common leaf, the estimation of $\mathbb{E}\left[(s_i(\mathbf{X}) - \hat{s}(R_i))^2\right]$ is performed by computing MSE_i and limited samples can introduce errors in MSE_i. Therefore, similar to [Chen *et al.*, 2024b], we introduce the following statistic for each common leaf l:

$$\text{Stats} = \frac{\text{MSE}_l}{\min_h \text{MSE}_l^h + \varepsilon}. \tag{12}$$

The ratio $\frac{\text{MSE}_l}{\min_h \text{MSE}_l^h}$ distinguishes between shifted and non-shifted common leaf nodes l. For shifted nodes, the ratio is large. For non-shifted nodes, the ratio is small, as MSE_l converges to zero faster than MSE_l^h, supported by the larger dataset used for MSE_l computation (Theorem 1). The small ϵ in the denominator ensures numerical stability, e.g., 10^{-9}. We iteratively identify common leafs and calculate the value of this ratio for each common leaf. This process ultimately generates a dictionary where each key-value pair comprises a node index and its corresponding stat value. After sorting this dictionary in non-increasing order, we utilize the Python library *kneed* to find the *knee* point which is the point of maximum curvature in a curve. Finally, we selected the *knee* point as the boundary between shifted and non-shifted nodes.

5 Experiments

In this section, we illustrate the capability of our algorithm CMSI through extensive experiments on synthetic and real-world datasets. In Section 5.1, we evaluate the algorithm's ability to detect shifted variables on synthetic datasets, where the data generation function is characterized as a composite trigonometric function. In Section 5.2, we demonstrate the application of CMSI to real-world scenarios by experiments on an ovarian cancer dataset. Additionally, we provide addi-

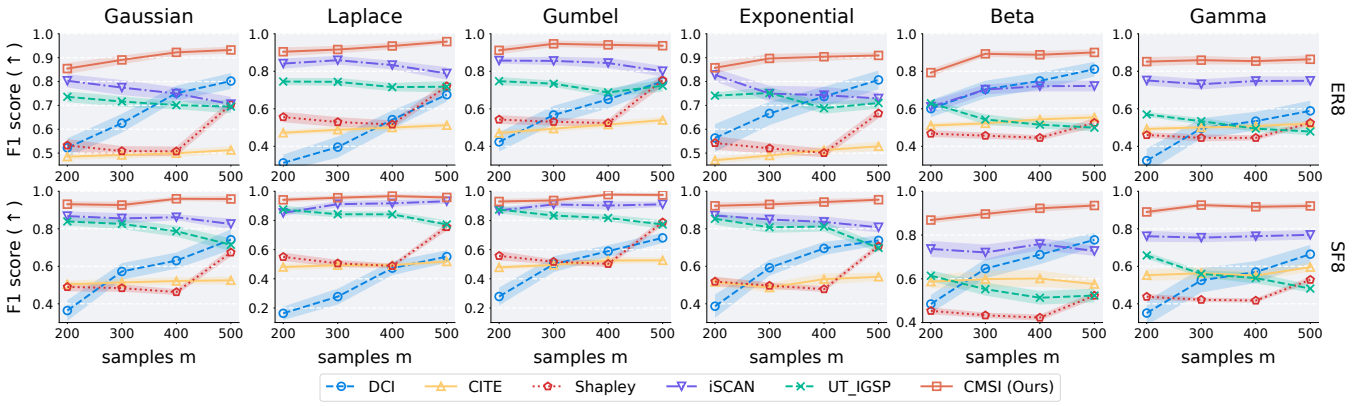

Figure 3: Simulation results on synthetic datasets with 50 nodes and 400 edges under various sample counts $m \in \{200, 300, 400, 500\}$. The x-axis represents the number of samples m, the top axis indicates the noise distribution, and the right-side axis labels the graph model type (Erdös-Rényi (ER)/Scale-Free (SF)). Each subplot shows the performance of CMSI and other baselines for a specific setting. The points indicate the average values obtained from 30 simulations. The shaded areas around the lines represent the standard error.

tional experiments in the Appendix including: (1) Extended evaluations on shifted variable identification using the data-generating functions of Section 5.1. (2) Identifying shifted variables when the data generation is generated by sampling Gaussian processes. (3) Identifying shifted variables when the underlying causal graph structures are different.

5.1 Synthetic Experiments

Basic Settings. We use the Erdös-Rényi (ER) and Scale-Free (SF) graph models to generate random DAGs with d nodes ($d \in \{20, 30, 40, 50\}$) and kd edges ($k \in \{2, 4, 6, 8\}$). The default sample size in each environment equals 500.

Synthetic Dataset. Based on the generated causal graph and the additive noise model (noise $\in$ {Gaussian(0, 1), Laplace(0, 1), Gumbel(0, 1), Exponential(1), Beta(1, 1), Gamma(0.5, 0.5)}), we construct the synthetic datasets. The non-shifted variables are generated by following additive noise model:

$$X_i = \sum_{j \in \mathbf{PA}_i} \sin(X_j^2) + N_i. \tag{13}$$

To generate the shifted variables, we first select a subset of non-root nodes as shifted nodes, based on a ratio $r \in \{0.15, 0.20, 0.25, 0.30\}$. For each shifted node, a single environment is randomly selected from all available environments to represent its shifted environment. The node does not undergo a shift in the remaining environments. (For example, consider a scenario with $H = 3$ environments, each containing $d = 20$ nodes. Using a ratio $r = 0.2$, we randomly select 4 non-root nodes as shifted nodes. Then, we randomly assign shifted node 1 to environment 2 for its shift, shifted node 2 to environment 1, and so on for the remaining shifted nodes. Each shifted node only exhibits a shift in its assigned environment.) For a shifted variable $X_i \in I$, we randomly select t direct parents $\widetilde{\mathbf{PA}}_i \subseteq \mathbf{PA}_i$ and then intervene in the relationships between $\widetilde{\mathbf{PA}}_i$ and X_i. Similar as [Chen *et al.*, 2024b], the intervention is implemented by changing the generation function f_i from $\sin(X_j^2)$ to $4\cos(2X_j^2 - 3X_j)$. So $\forall X_i \in \mathbf{I}$:

$$X_i = \sum_{j \in \mathbf{PA}_i \setminus \widetilde{\mathbf{PA}}_i} \sin(X_j^2) + \sum_{j \in \widetilde{\mathbf{PA}}_i} 4\cos(2X_j^2 - 3X_j) + N_i.$$

Basic Experimental Setup. We use F1 Score metric to evaluate the algorithm's ability to detect intervened variables. Experiments were conducted on a system equipped with an Intel Xeon(R) Platinum 8255C CPU and two NVIDIA GeForce RTX 2080 Ti GPUs.

Baselines. We compared the performance of CMSI against several baselines, which include: iSCAN [Chen *et al.*, 2024b], CITE [Varici *et al.*, 2021], Shapley [Budhathoki *et al.*, 2021], DCI [Wang *et al.*, 2018], UT-IGSP [Squires *et al.*, 2020].

Overall Results Analysis. From Fig. 1∼4 we can see that CMSI consistently achieves state-of-the-art performance in terms of F1 score, compared with other well-performing baselines. In contrast, CITE and DCI exhibit the worst performance and the limitations are likely attributable to their inherent reliance on linearity restrictions on SCMs. In addition, CSMI exhibits the lowest standard deviation in general, indicating greater stability and reliability regardless of data variations. More significantly, considering different noise types, CMSI exhibits the best performance in most scenarios. This indicates that CMSI exhibits stronger robustness and generalization capabilities across different noise types.

Results w.r.t. number of nodes. Fig. 1 illustrates the results with $d \in \{20, 30, 40, 50\}$ nodes and $8d$ edges. As the number of nodes increases, the performance of CMSI steadily improves under any noise conditions. In contrast, the performance of CITE decreases as the number of nodes increases. Furthermore, the trends exhibited by certain methods differ according to the type of noise present. For instance, DCI's performance improves with increasing node numbers under Laplace noise, but declines under Gamma noise.

Results w.r.t. number of edges. Fig. 2 displays the results with 50 nodes and $50k$ edges ($k \in \{2, 4, 6, 8\}$). The performance of CMSI consistently improves as the number of

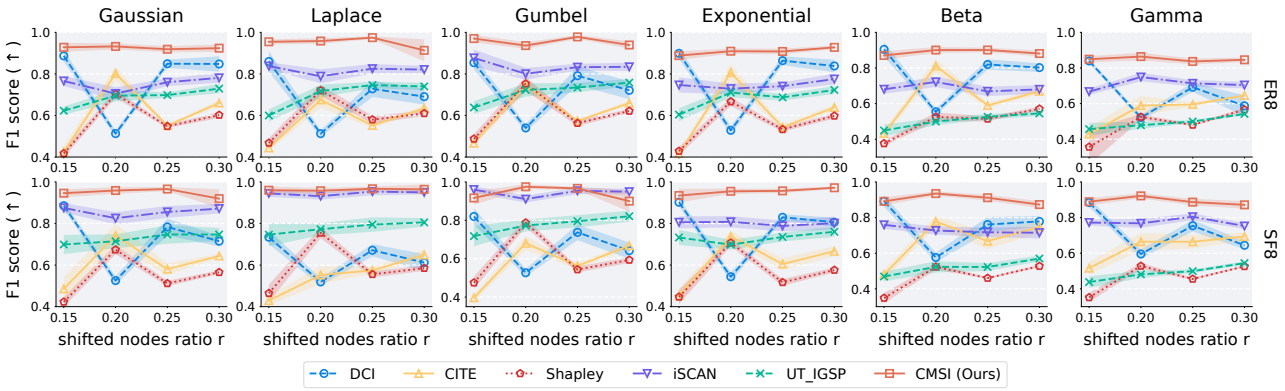

Figure 4: Simulation results on synthetic datasets with 50 nodes and 400 edges under various shifted nodes ratio $r \in \{0.15, 0.20, 0.25, 0.30\}$. The shifted nodes ratio r means the proportion of shifted nodes across all environments relative to the total number of nodes in a single environment. The x-axis represents the shifted nodes ratio r, the top axis indicates the noise distribution, and the right-side axis labels the graph model type (Erdös-Rényi (ER)/Scale-Free (SF)). Each subplot shows the performance of CMSI and other baselines for a specific setting. The points indicate the average values obtained from 30 simulations. The shaded areas around the lines represent the standard error.

edges grows. This suggests that CMSI has stronger generalization abilities and excels at identifying shifted variables in denser causal graphs. A notable exception is the scenario presented in Fig. 2 where the edge count equal to 100 and the node count equal to 50. This represents the sparsest graph structure among all those tested with the largest number of nodes. Therefore, the lack of sufficient edge information (i.e., information about causal relationships) in this scenario hinders CMSI's ability to make accurate judgments, leading to slightly poorer performance. Nevertheless, numerous real-world applications (such as gene regulatory networks) involve dense causal graphs, ensuring the wide applicability of CMSI.

Results w.r.t. number of samples. Fig. 3 displays the results with 50 nodes, 400 edges and m samples ($m \in \{200, 300, 400, 500\}$). In general, an increase in the number of samples results in enhanced performance for algorithms, such as DCI. However, CMSI demonstrates stable performance across different sample sizes, indicating its ability to effectively acquire knowledge from the samples. This makes it suitable for shift detection in small-sample scenarios. In contrast, algorithms like DCI, Shapley, and CITE perform poorly with small sample sizes, while iSCAN and UT_IGSP exhibit performance fluctuations as the sample size increases.

Results w.r.t. shifted nodes ratio. Fig. 4 displays the results with 50 nodes, 400 edges and shifted nodes ratio $r \in \{0.15, 0.20, 0.25, 0.30\}$). When the ratio r varies, the performance of DCI, and Shapley fluctuates significantly, but CMSI's performance remains stable and consistently the best. This suggests that CMSI can detect shifts in causal mechanisms under varying degrees of shift.

Additional Experiments. Due to space limitations, we leave more experiments in the Appendix. Additional experiments include 1) increasing the number of environments $H \in [2, 3, 4, 5]$, 2) more flexible data generation processes using Gaussian process, 3) shifted variable identification with diverse causal structures. All results consistently imply that CMSI possesses greater general applicability than baseline methods. See the Appendix for more details.

5.2 Experiments on Ovarian Cancer Dataset

We evaluated CMSI on an ovarian cancer dataset [Tothill *et al.*, 2008] that was previously analyzed by iSCAN [Chen *et al.*, 2024b] and DCI [Wang *et al.*, 2018]. This dataset was collected from patients with stage III or stage IV ovarian cancer and divided into two subsets based on survival duration. Based on these two subsets, CMSI identified the two most heavily intervened genes in the apoptosis pathway: BIRC3 and FAS. BIRC3 was also identified by iSCAN and DCI, but FAS was exclusively identified by CMSI as a highly intervened gene. In fact, BIRC3 is categorized as an inhibitor of apoptosis proteins (IAPs) , and the changes in BIRC3 expression levels induced by intervention affect the survival time of ovarian cancer patients [Hu *et al.*, 2019]. In addition, FAS is highly expressed in ovarian tumors [Mondal *et al.*, 2023] and the FAS receptor (CD95) expressed by FAS contributes to tumor growth [Ceppi *et al.*, 2014].

6 Conclusion

In this work, we propose CMSI, an applicable method for detecting causal mechanism shifts under ANM with arbitrary noise distributions. We prove that the score function of the mixture distribution unveils information to detect distribution shifts for the common leaf nodes for any noise distribution. Then an algorithm is proposed that iteratively selects a common leaf variable among the DAGs in different environments and evaluates whether such variable is shifted. The effectiveness of CMSI was assessed using various synthetic datasets and real-world ovarian cancer dataset. In the future, we are eager to explore how to extend CMSI to scenarios in which several latent confounders exist and how to identify mechanism shifts for these latent confounders.

Acknowledgements

This work was supported by National Natural Science Foundation (62372116, 62306019, 12471308 and 62472415) and Guangdong Basic and Applied Basic Research Foundation (2025A1515010103).

Contribution Statement

Yewei Xia and Xueliang Cui contributed equally to this work. Corresponding authors: Hao Zhang, Ruxin Wang, Shuigeng Zhou.

References

[Bogatinovski *et al.*, 2021] Jasmin Bogatinovski, Sasho Nedelkoski, Alexander Acker, Florian Schmidt, Thorsten Wittkopp, Soeren Becker, Jorge Cardoso, and Odej Kao. Artificial intelligence for it operations (aiops) workshop white paper. *arXiv preprint arXiv:2101.06054*, 2021.

[Budhathoki *et al.*, 2021] Kailash Budhathoki, Dominik Janzing, Patrick Bloebaum, and Hoiyi Ng. Why did the distribution change? In *International Conference on Artificial Intelligence and Statistics*, pages 1666–1674. PMLR, 2021.

[Ceppi *et al.*, 2014] Paolo Ceppi, Abbas Hadji, Frederick J Kohlhapp, Abhinandan Pattanayak, Annika Hau, Xia Liu, Huiping Liu, Andrea E Murmann, and Marcus E Peter. Cd95 and cd95l promote and protect cancer stem cells. *Nature communications*, 5(1):5238, 2014.

[Chen *et al.*, 2024a] Mingjie Chen, Hongcheng Wang, Ruxin Wang, Yuzhong Peng, and Hao Zhang. Cdrm: Causal disentangled representation learning for missing data. *Knowledge-Based Systems*, 299:112079, 2024.

[Chen *et al.*, 2024b] Tianyu Chen, Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. iscan: identifying causal mechanism shifts among nonlinear additive noise models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Hoyer *et al.*, 2008] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

[Hu *et al.*, 2019] Xiaoyan Hu, Yang Meng, Lian Xu, Lei Qiu, Mingtian Wei, Dan Su, Xu Qi, Ziqiang Wang, Shengyong Yang, Cong Liu, et al. Cul4 e3 ubiquitin ligase regulates ovarian cancer drug resistance by targeting the antiapoptotic protein birc3. *Cell death & disease*, 10(2):104, 2019.

[Hudson *et al.*, 2009] Nicholas J Hudson, Antonio Reverter, and Brian P Dalrymple. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS computational biology*, 5(5):e1000382, 2009.

[Li and Turner, 2017] Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

[Li *et al.*, 2023] Xinran Li, Bo Jiang, and Jun S Liu. Kernel-based partial permutation test for detecting heterogeneous functional relationship. *Journal of the American Statistical Association*, 118(542):1429–1447, 2023.

[Liu *et al.*, 2017] Song Liu, Kenji Fukumizu, and Taiji Suzuki. Learning sparse structural changes in high-dimensional markov networks: A review on methodologies and theories. *Behaviormetrika*, 44:265–286, 2017.

[Lyu, 2012] Siwei Lyu. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.

[Meek, 1997] Christopher Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University, 1997.

[Mondal *et al.*, 2023] Tanmoy Mondal, Himanshu Gaur, Brice EN Wamba, Abby Grace Michalak, Camryn Stout, Matthew R Watson, Sophia L Aleixo, Arjun Singh, Salvatore Condello, Roland Faller, et al. Characterizing the regulatory fas (cd95) epitope critical for agonist antibody targeting and car-t bystander function in ovarian cancer. *Cell Death & Differentiation*, 30(11):2408–2431, 2023.

[Montagna *et al.*, 2023] Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pages 726–751. PMLR, 2023.

[Peters *et al.*, 2014] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

[Pimanda *et al.*, 2007] John E Pimanda, Katrin Ottersbach, Kathy Knezevic, Sarah Kinston, Wan YI Chan, Nicola K Wilson, Josette-Renée Landry, Andrew D Wood, Anja Kolb-Kokocinski, Anthony R Green, et al. Gata2, fli1, and scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proceedings of the National Academy of Sciences*, 104(45):17692–17697, 2007.

[Plis *et al.*, 2010] Sergey M Plis, Vince D Calhoun, Michael P Weisend, Tom Eichele, and Terran Lane. Meg and fmri fusion for non-linear estimation of neural and bold signal changes. *Frontiers in neuroinformatics*, 4:114, 2010.

[Ren *et al.*, 2025] Yixin Ren, Haocheng Zhang, Yewei Xia, Hao Zhang, Jihong Guan, and Shuigeng Zhou. Fast causal discovery by approximate kernel-based generalized score functions with linear computational complexity. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, page 1197–1208, 2025.

[Robins *et al.*, 2000] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[Rolland *et al.*, 2022] Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

[Sanei and Chambers, 2013] Saeid Sanei and Jonathon A Chambers. *EEG signal processing*. John Wiley & Sons, 2013.

[Shi *et al.*, 2020] Shuping Shi, Stan Hurn, and Peter CB Phillips. Causal change detection in possibly integrated

systems: Revisiting the money–income relationship. *Journal of Financial Econometrics*, 18(1):158–180, 2020.

[Song and Ermon, 2019] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[Song *et al.*, 2020] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[Squires *et al.*, 2020] Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

[Stein, 1972] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pages 583–603. University of California Press, 1972.

[Tothill *et al.*, 2008] Richard W Tothill, Anna V Tinker, Joshy George, Robert Brown, Stephen B Fox, Stephen Lade, Daryl S Johnson, Melanie K Trivett, Dariush Etemadmoghadam, Bianca Locandro, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical cancer research*, 14(16):5198–5208, 2008.

[Varici *et al.*, 2021] Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Scalable intervention target estimation in linear models. *Advances in Neural Information Processing Systems*, 34:1494–1505, 2021.

[Wang *et al.*, 2018] Yuhao Wang, Chandler Squires, Anastasiya Belyaeva, and Caroline Uhler. Direct estimation of differences in causal graphs. *Advances in neural information processing systems*, 31, 2018.

[Xu *et al.*, 2024] Wenwei Xu, Hao Zhang, Yewei Xia, Yixin Ren, Jihong Guan, and Shuigeng Zhou. Hybrid causal feature selection for cancer biomarker identification from rna-seq data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2024.

[Yuan *et al.*, 2017] Huili Yuan, Ruibin Xi, Chong Chen, and Minghua Deng. Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4):755–770, 2017.

[Zhang *et al.*, 2021] Hao Zhang, Kun Zhang, Shuigeng Zhou, Jihong Guan, and Ji Zhang. Testing independence between linear combinations for causal discovery. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6538–6546, 2021.

[Zhang *et al.*, 2022] Hao Zhang, Shuigeng Zhou, Kun Zhang, and Jihong Guan. Residual similarity based conditional independence test and its application in causal discovery. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 5942–5949, 2022.

[Zhang *et al.*, 2024] Hao Zhang, Yixin Ren, Yewei Xia, Shuigeng Zhou, and Jihong Guan. Towards effective causal partitioning by edge cutting of adjoint graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10259–10271, 2024.

[Zhao *et al.*, 2014] Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.