Causal Discovery from Shifted Multiple Environments

Dezhi Yang
School of Software, Shandong University
Jinan, Shandong, China
dzyang@mail.sdu.edu.cn

Guoxian Yu
School of Software, Shandong University
Jinan, Shandong, China
gxyu@sdu.edu.cn

Jun Wang*
School of Software, Shandong University
Jinan, Shandong, China
kingjun@sdu.edu.cn

Jinglin Zhang
School of Control Science and Engineering, Shandong Universiry
Jinan, Shandong, China
jinglin.zhang@sdu.edu.cn

Carlotta Domeniconi
Department of Computer Science, George Mason University
Fairfax, VA, USA
carlotta@cs.gmu.edu

Abstract

A fundamental problem in many science domains is learning the causal structure of a system from observed data. The observed data canonically come from multiple environments (i.e. different times, locations, and measurements), and causal models may have unobserved shifts. Although the causal graphs can be identified by modeling the distribution changes among different environments, existing solutions can only learn causal structures when given environmental information. In contrast, we propose a causal discovery approach (`CausalSME`) which automatically identifies pseudo environments and unobserved distribution shifts. Specifically, `CausalSME` learns a causal model containing unobserved variables, which can correct the distribution shifts with mixed environments. The heart of `CausalSME` is a variational autoencoder that infers shifted causal effects of unobserved variables and guides the identification of environment information. It further divides the shifted samples by the identified environments to jointly learn an invariant causal model. We prove the structure identifiability of `CausalSME` with the causal additive model. In our extensive experiments we show that `CausalSME` achieves state-of-the-art performance.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Bayesian network models**; *Learning latent representations*; *Latent variable models*; *Causal reasoning and diagnostics.*

Keywords

Causal Discovery; Multiple Environments; Distribution Shifts; Bilevel Optimization; Adversarial Learning

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '25, August 3–7, 2025, Toronto, ON, Canada
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1245-6/25/08
https://doi.org/10.1145/3690624.3709247

ACM Reference Format:
Dezhi Yang, Guoxian Yu, Jun Wang, Jinglin Zhang, and Carlotta Domeniconi. 2025. Causal Discovery from Shifted Multiple Environments. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3690624.3709247

1 Introduction

Causal discovery is the process of learning cause-and-effect relationships in the form of a directed acyclic graph (DAG) from a given system. It is interpretable as we can see the generative pattern of the observed data and estimated effects. Furthermore, it is generalizable across different and novel environments, and thus has an essential role in many fields, such as biology [6], healthcare [27], and economics [9]. Nonetheless, it is quite challenging for scientists to collect enough data from multiple environments and learn a causal structure with distribution shifts. In most situations, environmental changes cause unobserved variables to have different distributions, as well as exogenous noise, which greatly corrupt the learned causal structures. Methods based on the single-environment assumption [46, 47] cannot remove correlations caused by distribution shifts, and hence give a causal graph with many spurious edges.

Throughout this work, we focus on a general and practical scenario that the data come from *multiple environments* with unobserved *distribution shifts*. Under this condition, identifying the underlying causal structure is generally not possible for methods that test optimal conditional independence relations [33, 34, 43] and fit functions [5, 31, 36], which only find the Markov equivalence class. Methods that rely on identifiable models, such as LiNGAM [32], CAM [2], and PNL [45], additionally require the arguably strong assumption of an independent and identical distribution. Indeed, gradient-based methods (DAG-GNN [41], and Gran-DAG [16]) similarly fail to consistently estimate the underlying causal graph with unobserved shifts.

It is inevitable to measure data from multiple environments in most applications, and estimating multi-environment distributions and detecting cross-environment shifts have become an important research topic in causal discovery. DICD [38] and Meta-DAG [19] combine multi-environment data to learn an invariant causal structure. MSS [24] and MC/IB [8] identify a causal structure by testing changes in conditional distributions across environments. One of the cornerstone conditions of these methods is that the

environmental information should be known, namely the number of environments and samples contained in each environment. However, this information is generally unknown in practice.

In this paper, we propose a novel causal discovery method for learning a causal directed acyclic graph (DAG) from mixed multi-environment data with unobserved distribution shifts, by automatically identifying unobserved causal effects and environments. Given data over observed variables, our goal is to determine the effects of unobserved variables, divide data into different environments, and jointly learn causal networks over the observed variables.

To achieve this goal, we define a causal additive model with unobserved variables, and prove our model is identifiable when unobserved causal effects are mutually independent. For environment recognition, we develop a framework to automatically partition environments based on variational inference. Specifically, we leverage a variational autoencoder to infer the causal effects of unobserved variables, combine it with the reconstruction process of observed data, and formulate this process as a bilevel optimization problem. To avoid over-reconstruction causality of unobserved shifts, we introduce independence constraints of reconstruction residuals and unobserved causal effects. With a designed clustering strategy, the unobserved causal effects approximated by the additional variational autoencoder efficiently guide the identification of environmental information. After the environments' identification, we set up independent variational autoencoders for the samples of the different environments, learn the unobserved causal effects of each environment in parallel, and combine them with the same reconstruction process, where the reconstruction model restores an original causal graph. Figure 1 shows the conceptual framework of `CausalSME`. Our main contributions are outlined as follows:
(i) We propose a novel causal discovery algorithm (`CausalSME`) for mixed multi-environment observed data by accounting for causal effects of unobserved variables, and uncover causal graphs from various distributions with unobserved shifts, which is a challenging, practical but largely unexplored topic.
(ii) `CausalSME` develops a bilevel optimization for multi-environment data, which is divided by a residual-effect clustering strategy, to infer environment-specific unobserved causal effects and environment-invariant causal graphs, and guarantees the structural identifiability by pursuing mutually independent unobserved causal effects and reconstructed residuals in an adversarial way.
(iii) Experiments on synthetic and real datasets demonstrate that `CausalSME` significantly improves the performance of causal discovery without knowing the environment information a-priori, in both single-environment and multiple-environment scenarios.

2 Related Work

Causal discovery is one of the most critical problems in statistics and has gained increasing attention among researchers [23]. However, the complex causal relationships and unknown noises make it difficult to infer an accurate causal structure without sufficient observed data. When the observed data are sufficient, both constraint-based [33, 43, 44] and score-based [4, 11, 28] methods can identify the causal structure up to a Markov equivalence class. These methods suffer from a huge combinatorial optimization overhead and do not benefit from continuous optimization methods.

To estimate causal graphs using continuous optimization, NOTEARS [46] generalizes causal discovery as a continuous optimization problem by introducing a differentiable equality to constrain the number of cycles in the graph. To extend the framework of NOTEARS, DAG-GNN [41], Gran-DAG [16], and NOTEARS-MLP [47] find nonlinear (or generalized linear) causal relationships by implementing a similar acyclic constraint on neural networks. RCL-OG[40] uses reinforcement learning to search for accurate causal graphs. By introducing more complex neural networks, GraN-LCS [17] and HetDAG [18] enable gradient-based methods to learn causal graphs from large-scale complex data. The aforementioned methods can only distinguish Markov equivalence classes, unless an identifiable structure function model is assumed. Based on the asymmetry of distribution in the causal and the anti-causal directions, [7, 12, 21] can determine causal directions with asymmetry. [2, 25, 32, 45] show that causal direction can be identified by assuming some specific structural causal models.

In the presence of multiple environments or distribution shifts, several algorithms, such as DARING [10] and ReScore [42], can find causal graphs with heterogeneous noise, where the causal model avoids over-reconstruction due to residual independence or sample reweighting. DICD [38] and Meta-DAG [19] combine data from different environments to learn a uniform causal network that conforms to the distribution of each environment. Based on the federated learning, FedCausal [39] combines multi-environment heterogeneous data to learn the global causal structure. However, none of these methods can explain the effect of unknown interventions and unobserved variables across multiple environments.

To relieve this issue, [22, 35] attempt to identify unknown intervention targets. [13, 37] focus on estimating shifted causal mechanisms and confounders. However, these methods *cannot* infer causal graphs with unobserved distribution shifts. MSS [24] identifies causal structures by the sparsity of the causal mechanism shifts. MC/IB [8] applies ideas of invariant causal prediction [26] to identify the causal DAG. Unfortunately, the above approaches need to *know* the environmental information beforehand. In contrast, our `CausalSME` can achieve robust causal discovery by automatically identifying environmental information and explicitly modeling the causal effects of unobserved variables.

3 Preliminaries

3.1 Structural Causal Model

A Structural Causal Model (SCM) describes the data generation process. It encodes a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{V_1, \cdots, V_d\}$ is the set of variables and $\mathcal{E}$ is the set of edges between variables, which represent the direct causal effect of one variable on another. Let $X \in \mathbb{R}^{n \times d}$ denote the observed matrix of n samples and d variables, and $X_i = X[:, i]$ represent the observed values of V_i. For ease of understanding, we use X_i instead of V_i in the following. We assume that $\mathcal{G}$ satisfies faithfulness, sufficiency, and causal Markov conditions [15], and corresponds to a joint distribution $P(\mathbf{X}) = \prod_{i=1}^{d} P(X_i|Pa(X_i))$, where $Pa(X_i)$ is the set of parents of X_i in $\mathcal{G}$. If all causal relationships are linear, we can define a standard linear structural causal model as follows:

$$X = XA + \epsilon \tag{1}$$

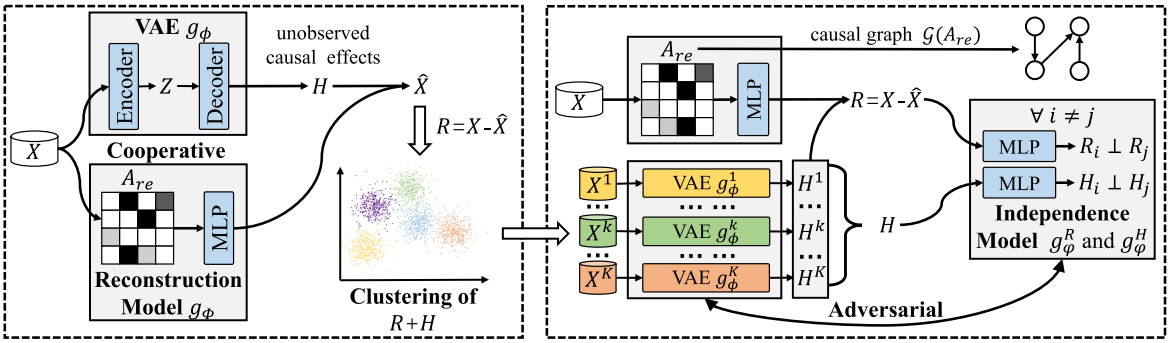


Figure 1: The overview of CausalSME. VAE g_ϕ and reconstruction model g_Φ cooperate to reconstruct the observed data. CausalSME identifies unknown environmental information by clustering samples based on the reconstruction loss $R = X - \hat{X}$ and unobserved causal effects H. Then it leverages independent VAEs ($\{g_\phi^k\}_{k=1}^K$) to infer unobserved causal effects of K environments in parallel, and trains these VAEs and the reconstruction model to learn the causal graph $\mathcal{G}(A_{re})$. The independence models g_φ^R and g_φ^H are trained via an adversarial process to pursue mutual independence of reconstruction losses and unobserved causal effects.

where the non-zero entry of the matrix $A \in \mathbb{R}^{d\times d}$ encodes a causal DAG, if $A_{ji} = 0$, there is no directed edge from X_j to X_i. $\epsilon \in \mathbb{R}^{n\times d}$ is a matrix composed by mutually independent exogenous noise. We further include nonlinear functions in the model:

$$X_i = \sum f_j(X_j) + \epsilon_i, X_j \in Pa(X_i) \tag{2}$$

It is straightforward to show this model conforms to the causal additive model (CAM) [2], and X_i is equal to the sum of all its parents $X_j \in Pa(X_i)$ after independent nonlinear mapping $f_j(\cdot)$. This model is known to permit identifiability of causal structures. Based on the above models, we study the identifiability of causal graphs under the shifted effects of unobserved variables.

3.2 Unobserved Variables in SCMs

The key variables in a system are not always fully observable. To consider unobserved causal effects, we introduce unobserved variables into a linear model:

$$X = XA + ZB + \epsilon \tag{3}$$

where $Z \in \mathbb{R}^{n\times d}$ represents the numerical matrix of unobserved variables, and the matrix $B \in \mathbb{R}^{d\times d}$ encodes causal relationships from the unobserved variables to the observed ones. As we assume Causal Sufficiency [33] (there are no confounders), it is still possible for the causal graph to be identified. Without loss of generality, we assume that there is no causality from X to Z and no causality among the elements of Z, and that each observed variable has one and only one unobserved parent variable, without hindering the effects of the distribution $P(X|Z)$. Let $H \in \mathbb{R}^{n\times d}$ denote the causal effects of unobserved variables, we then write the model as follows:

$$X = XA + H + \epsilon \tag{4}$$

where $H = ZB$. We can therefore focus on the causal effects of Z on X without distinguishing each unobserved variable.

3.3 Problem Setting

Let the data be generated by the following model:

$$X = XA + ZB + \epsilon, \quad Z \sim N(\mu_Z, \delta_Z), \quad \epsilon \sim N(\mu_\epsilon, \delta_\epsilon) \tag{5}$$

Our primary goal is to recover the matrix A, which directly encodes the causal relationships between the observed variables. Our second goal is to infer the causal effects of the unobserved variables, otherwise we would learn a biased A. In addition, mixed data from multiple environments may have different distributions of unobserved variables (i.e. their means and variances may be shifted), which further increases the difficulty of the causal effect inference. Therefore, our additional goal is to distinguish data from different environments, which is necessary to infer unobserved causal effects specific to each environment.

4 Methodology

In this section, we show that, with several additional independence constrains, the causal model containing unobserved variables is identifiable. We then elaborate on CausalSME for inferring unobserved causal effects and identifying environmental information. We learn invariant causal structures by combining data and unobserved causal effects in all environments.

4.1 Identifiability

Consider the model from Eq. (5) with $X = (X_1, \cdots, X_d)$, $Z = (Z_1, \cdots, Z_d)$, and B diagonal matrix with diagonal elements $\{B_{ii}\}_{i=1}^d$. We write the generation function of X_i as follows:

$$X_i = X_{-i}A[:, i] + Z_i B_{ii} + \epsilon_i \tag{6}$$

If we try to fit A over only the observed variables, we will get a biased result $A' = \arg\min_{A'} \mathbf{E}(|X - XA'|), A' \neq A$. To accurately discover A, we infer the causal effects of the unobserved variables $H' = (H'_1, \cdots, H'_d)$ and include them in the fitting of the observed data $\hat{X}_i = X_{-i}A'[:, i] + H'_i$. With the goal of minimizing the reconstruction loss $R_i = \mathbf{E}(|X_i - \hat{X}_i|)$, there exist multiple A' and H'

which make $\hat{X}_i$ close to X_i. We state that, with some additional constraints, the above model can identify a unique matrix $A' = A$.

Theorem 4.1. ***Identifiability.*** *When the reconstruction loss is minimized, if the reconstruction residuals and the unobserved causal effects of all variables are mutually independent, then A is identifiable.*

Proof. Taking a single variable X_i, it is easy to see if $A'[:, i] = A[:, i]$ and $H'_i = Z_i B_{ii}$, then the reconstruction residuals $R_i = \mathbf{E}(|\epsilon_i|) = 0, R_i \perp X_{-i}$ and $H'_i \perp X_{-i}$. We reason by contradiction, and assume $A'[:, i] = A[:, i]$ or $H'_i = Z_i B_{ii}$ does not hold when $R_i = \mathbf{E}(|\epsilon_i|)$, $R_i \perp X_{-i}$ and $H'_i \perp X_{-i}$:

• If $A'[:, i] \neq A[:, i]$ and $H'_i = Z_i B_{ii}$, then $R_i = \mathbf{E}(|X_{-i}A[:, i] - X_{-i}A'[:, i] + \epsilon_i|)$. Obviously, $R_i \not\perp X_{-i}$, which is not consistent with the assumption.

• If $A'[:, i] = A[:, i]$ and $H'_i \neq Z_i B_{ii}$, then $R_i = \mathbf{E}(|Z_i B_{ii} - H'_i + \epsilon_i|)$. Obviously, R_i is not minimized and $R_i \neq \mathbf{E}(|\epsilon_i|)$, which is not consistent with the assumption.

• If $A'[:, i] \neq A[:, i]$ and $H'_i \neq Z_i B_{ii}$, then $R_i = \mathbf{E}(|X_{-i}A[:, i] - X_{-i}A'[:, i] + Z_i B_{ii} - H'_i + \epsilon_i|)$. If $X_{-i}A[:, i] - X_{-i}A'[:, i] + Z_i B_{ii} - H'_i \neq 0$, then R_i is not minimized, which is not consistent with the assumption. Otherwise, if $X_{-i}A[:, i] - X_{-i}A'[:, i] + Z_i B_{ii} - H'_i = 0$, then $H'_i = X_{-i}A[: i] - X_{-i}A'[:, i] + Z_i B_{ii}$ and $H'_i \not\perp X_{-i}$, which is also not consistent with the assumption.

□ We have therefore proved that if $R_i = \mathbf{E}(|\epsilon_i|)$, $R_i \perp X_{-i}$ and $E_i \perp X_{-i}$, then $A'[:, i] = A[:, i]$ and $H'_i = Z_i B_{ii}$. Generalizing the above proof to all observed variables, we show that if $\forall i, R_i = \mathbf{E}(|\epsilon_i|)$, $\forall i \neq j, R_i \perp R_j$ and $E_i \perp E_j$, then $A' = A$ and $H' = ZB$.

With multiple environments, the distribution of both unobserved variables and exogenous noise may shift. When the mean of the noise is shifted, i.e. $\epsilon \sim N(\mu_\epsilon, \delta_\epsilon), \mu_\epsilon \neq 0$, we assume that the mean shift is included in the unobserved causal effect, i.e. $H = ZB + \mu_\epsilon$, and that the exogenous noise still follows the Gaussian distribution with $\mu_\epsilon = 0$ and $\mathbf{E}(|\epsilon|) = 0$. With data from multiple environments, if we infer unobserved causal effects in each environment and ensure they are independent of each other, we can minimize the reconstruction loss and recover the matrix A.

4.2 Model Setting

To infer the causal effects of unobserved variables so that the causal graph is correctly identified, we make use of variational autoencoders (VAE) [14]. Different from traditional VAEs, our VAE infers the causal effects of unobserved variables rather than reconstructing the observed data. The encoder inferences the mean and the variance of potential variables:

$$\begin{aligned} \mu &= \mathrm{MLP}_\mu(\sigma(XA_{en})) \\ \delta &= \mathrm{MLP}_\delta(\sigma(XA_{en})) \end{aligned} \tag{7}$$

where σ and $A_{en} \in \mathbb{R}^{d \times dm}$ are the shared activation function and the encoding matrix, MLP_μ and MLP_δ are parallel encoding networks, and the observed data X are encoded as the potential variable mean μ and variance δ. Let A_{en} be reshaped as $W_{en} \in \mathbb{R}^{d \times m \times d}$, which encodes the reverse inference process from observed variables to unobserved ones. MLP_μ and MLP_δ are actual combinations of d multilayer perceptrons with mapping $f|\mathbb{R}^m \to \mathbb{R}^m$. By randomly sampling ξ from the Gaussian distribution $N(0, 1)$, we obtain the potential variables matrix $Z = \mu + \delta \times \xi$. We then decode the potential variables as unobserved causal effects:

$$H = \mathrm{MLP}_{de}(\sigma(ZB_{de})) \tag{8}$$

where σ is the activation function, $B_{de} \in \mathbb{R}^{md \times md}$ is the decoding matrix and MLP_{de} is the decoding network. m is a model type adjustment parameter; by setting $m = 1$ and skipping the activation function, the VAE is simplified to fit linear data, otherwise the model is nonlinear. In particular, we reshape B_{de} as $W_{de} \in \mathbb{R}^{m \times d \times m \times d}$ and set $\forall i \neq j, W[:, i, :, j] = 0$ to simulate the one-to-one causality from unobserved variables to observed ones. MLP_{de} is a combination of d multilayer perceptrons with mapping $f|\mathbb{R}^m \Rightarrow \mathbb{R}$. We denote the VAE as $g_\phi(\cdot)$ with parameters ϕ.

Based on the model from Eq. (4), we introduce the output H of the VAE into the reconstruction process of the observed data, and define the reconstruction model as follows:

$$\hat{X} = \mathrm{MLP}_{re}(\sigma(XA_{re})) + H \tag{9}$$

where $A_{re} \in \mathbb{R}^{d \times dq}$ is reshaped as $W_{re} \in \mathbb{R}^{d \times q \times d}$ to represent the causality between the observed variables, σ is the activation function, MLP_{re} is a combination of d multilayer perceptrons with mapping $f|\mathbb{R}^q \to \mathbb{R}$, and q is a model type adjustment parameter similar to m. Setting $q = 1$, the reconstruction model is rewritten as $\hat{X} = XA_{re} + H$. To interpret A_{re} as a causal graph, we constrain A_{re} to be acyclic and use the equality constraint proposed by [41, 46]:

$$h(\mathcal{G}) = tr((I + \mathcal{G} \odot \mathcal{G}/d)^d) - d = 0 \tag{10}$$

where I is an identify matrix, $\mathcal{G}_{ij} = \sum W_{re}[i, :, j]$, $\mathcal{G} \odot \mathcal{G}$ represents the Hadamard product $(\mathcal{G} \odot \mathcal{G})_{ij} = \mathcal{G}^2_{ij}$, and $tr(\cdot)$ is the trace of the matrix. We define the reconstruction model as $g_\Phi(\cdot)$ with parameters Φ.

We aim at reconstructing the observed data, but also at achieving the mutual independence between the reconstructed residuals and the unobserved causal effects. Let $R_i = X_i - \hat{X}_i$ and H_i be the reconstructed residual and the unobserved causal effect of X_i, while R_{-i} and H_{-i} are the reconstructed residuals and the unobserved causal effects of the other variables. We approximate the independence of each pair (R_i, R_{-i}) and (H_i, H_{-i}):

$$\begin{aligned} M(R) &= \sum_{i=1}^{d} || \frac{\mathrm{Cov}[f_i(R_i), f_{-i}(R_{-i})]}{\sqrt{\mathrm{Var}[f_i(H_i)]} \cdot \sqrt{\mathrm{Var}[f_{-i}(H_{-i})]}} || \\ M(H) &= \sum_{i=1}^{d} || \frac{\mathrm{Cov}[f_i(H_i), f_{-i}(H_{-i})]}{\sqrt{\mathrm{Var}[f_i(H_i)]} \cdot \sqrt{\mathrm{Var}[f_{-i}(H_{-i})]}} || \end{aligned} \tag{11}$$

Referring to the Theorem proposed by [10], if and only if $\forall f_i$, $\forall f_{-i}$, $M(R) = 0$, $M(H) = 0$, then (R_i, R_{-i}) and (H_i, H_{-i}) are both mutually independent. In practice, we set $f_i(R_i) = R_i$ and approximate f_{-i} with a multilayer perceptron to simplify the independence calculation. We define the approximated independence models as g^R_φ and g^H_φ with parameters φ_R and φ_H.

4.3 Model Optimization

As described in Section 4.2, we build models to infer the causal effects of unobserved variables, reconstruct the observed data and approximate the mutual independence of reconstructed residuals and unobserved causal effects. To train the models, we propose a bilevel optimization framework, and alternate the updating of the

VAE and the reconstruction model. The goal of a general VAE is to optimize the evidence lower bound defined as follows:

$$L_{elbo} = \mathbf{E}_{q(Z|X)}(\log p(X|Z)) - D(q(Z|X)|p(Z)) \tag{12}$$

Differently, our VAE infers the unobserved causal effect H instead of the observed data X, and guarantees that the elements of H are mutually independent. We therefore define a new objective function for our VAE:

$$\min_{\phi} L_{vae}(X,\phi) = -\mathbf{E}_{q(Z|X)}(\log p(H|Z)) + D(q(Z|X)|p(Z)) + \lambda_H M(H)(\phi) \tag{13}$$

where D is the Kullback-Leibler divergence, $D(q(Z|X)|p(Z)) = (\mu)^2 + (\delta)^2 - 2\log(\delta) - 1$, and λ_H is a penalty coefficient. We set $\mathbf{E}_{q(Z|X)}(\log p(H|Z)) = ||X - \hat{X}||_2^2/2n$. $M(H)(\phi)$ is the mutually independent penalty of the unobserved causal effects from Eq. (11).

For the reconstruction model, we fit the observed data under the acyclic and sparsity constraint, while ensuring mutually independent reconstructed residuals. We define the objective function of the reconstruction model as follows:

$$\min_{\Phi} L_{re}(X,\Phi) = ||X - \hat{X}||_2^2/2n + L_{DAG}(\Phi) + \beta L_{sparse}(\Phi) + \lambda_R M(R)(\Phi) \tag{14}$$

where β and λ_R are penalty coefficients, and $L_{sparse}(\Phi) = |\Phi|_1$ is a sparsity penalty, which also ensures the sparsity of causal graphs. $L_{DAG}(\Phi) = \rho/2|h(\mathcal{G})|^2 + \alpha h(\mathcal{G})$ is an acyclic penalty derived by the standard machinery of augmented Lagrange, where $h(\mathcal{G})$ is from Eq. (10), ρ is a penalty parameter, α is a dual variable. $M(R)(\Phi)$ is the mutually independent penalty of residuals from Eq. (11).

In continuous optimization, the VAE and the reconstruction model recover real unobserved causal effects and causal structures. We further formulate the independent constraint as a min-max problem and solve it with adversarial learning. We optimize the independence approximation model to maximize $M(R)$ and $M(H)$, while the VAE and the reconstruction model optimize R and H to minimize $M(R)$ and $M(H)$. The maximized $M(R)$ and $M(H)$ gradually get closer to, or equal to, zero in ideal settings, and we guarantee that R and H are mutually independent.

Algorithm 1 outlines the main process of `CausalSME`. Line 1 initializes the parameters of all models; Line 5 updates the independence approximation model g_φ^H to maximize $M(H)$; Lines 6-7 infer unobserved causal effects and update the VAE g_ϕ; Line 11 updates the independence approximation model g_φ^R to maximize $M(R)$; Lines 12-13 reconstruct observed data and update the reconstruction model g_Φ. The overall goal of `CausalSME` is as follows:

$$\min_{\phi,\Phi}\max_{\varphi_R,\varphi_H} L(X,\phi,\Phi,\varphi_R,\varphi_H) = L_{vae}(X,\phi) + L_{re}(X,\Phi) + M(R) + M(H) \tag{15}$$

where $M(R)$ and $M(H)$ are adversarial optimization targets of independence approximation models, rather than penalty terms in $L_{vae}(X,\phi)$ and $L_{re}(X,\Phi)$; as such there is no penalty coefficient. Finally, the reconstruction model outputs a causal graph $\mathcal{G}(W_{re})$, where $\mathcal{G}(W_{re})_{ij} = 1$ if and only if $\sum W_{re}[i,:,j] \geq \tau$, otherwise $\mathcal{G}(W_{re})_{ij} = 0$. We obtain a binary adjacency matrix through a threshold τ, which disregards sufficiently small values in $\mathcal{G}(W_{re})$

Algorithm 1 `CausalSME`: Causal Discovery from Shifted Multiple Environments

Input: Observed data $\mathbf{X}$ and hyper-parameters $\{\beta, \lambda_H, \lambda_R, t_1, t_2, \tau\}$
Output: Causal graph $\mathcal{G}$

1: Initialize the parameters of all models ϕ, Φ, φ_R and φ_H
2: **while** not reach max iterations or trigger termination conditions **do**
3: ▷ Optimize variational inference model
4: **for** *epoch* = 1 to t_1 **do**
5: Update φ_H to maximize $M(H)$
6: Fix Φ and infer H in Eq. (7) and (8)
7: Update ϕ to minimize L_{vae} from Eq. (13)
8: **end for**
9: ▷ Optimize reconstruction model
10: **for** *epoch* = 1 to t_2 **do**
11: Update φ_R to maximize $M(R)$
12: Fix ϕ and reconstruct X in Eq. (9)
13: Update Φ to minimize L_{re} from Eq. (14)
14: **end for**
15: **end while**
16: **return** $\mathcal{G}(W_{re})$

and improves the accuracy of the resulting graph. We set $\tau = 0.3$, a value which is widely used in differentiable causal discovery.

The computational complexity of CausalSME is $O(nd^2(m+q) + ndm^2 + d^2q + d^3)$, where m and q are constants that regulate the size of models. Compared to the computational complexity of NOTEARS-MLP $O(nd^2q + d^2q + d^3)$, CausalSME has an extra $O(nd^2m + ndm^2)$ of VAEs per iteration. However, the VAE is used to infer unobserved causal effects, which is necessary and enables `CausalSME` to identify the unbiased causal model and the causal structure in multiple environments with shifted distributions. In practice, the computation of acyclic penalty $O(d^3)$ is still the highest cost.

4.4 Environment Identification

Although our VAE can infer unobserved causal effects, it is generally not effective in multiple environments. This may be especially problematic as the VAE can approximate only one distribution, and the number of environments is typically unknown. For this reason, we divide samples into pseudo environments.

As proved in Theorem 4.1, we cannot minimize the reconstruction loss because in a multiple environment setting $H \neq ZB$. For loss minimization, we partition samples to learn an environment-specific H. Assuming that samples are divided into K groups (each group has n^k observed data X^k, with exogenous noise ϵ^k and unobserved variables Z^k), we denote the inferred unobserved causal effects as $\{H^k\}_{k=1}^K$ and the reconstruction loss as follows:

$$R = \sum_{k=1}^{K} ||X^k - \hat{X}^k|| = \sum_{k=1}^{K} ||X^k(A - A') + Z^kB - H^k + \epsilon^k|| \tag{16}$$

Consider an ideal case where $H^k = \mathbf{E}(X^k(A-A') + Z^kB + \epsilon^k)$ to minimize R, i.e., H^k is the center point of $X^k(A-A') + Z^kB + \epsilon^k$. We treat loss minimization as a clustering problem, namely to find optimal clusters $\{X^k\}_{k=1}^K$ such that $X^k - \hat{X}^k + H^k = X^k(A-A') + Z^kB + \epsilon^k$

has the smallest distance to its center point. We therefore use *K-Means* [1, 20] to divide samples and determine the optimal number of clusters with the Silhouette Coefficient [29]. The initial cluster centers are randomly chosen with a probability proportional to the shortest distance from the sample to the existing cluster centers. We then assign one VAE $g_\phi^k \in \{g_\phi^k\}_{k=1}^K$ with parameters $\phi_k \in \{\phi_k\}_{k=1}^K$ to each cluster to infer environment-specific $H^k \in \{H^k\}_{k=1}^K$. We retrain the models according to the framework of Algorithm 1, with the difference that Lines 6-7 infer $\{H_k\}_{k=1}^K$ of all environments and update VAEs $\{g_\phi^k\}_{k=1}^K$. When the distribution shifts are large enough to ensure that samples from different environments do not overlap, *K-Means* can accurately divide samples into their original environments; otherwise, *K-Means* can identify the samples which are not shared across environments.

5 Experiments

In this section, we conduct extensive experiments to evaluate our `CausalSME` when operating in environments with unobserved distribution shifts. All experiments were run on the same server (Ubuntu 18.04.5, Intel Xeon Gold 6248R and Nvidia RTX 3090). The code of CausalSME is shared at https://www.sdu-idea.cn/codes.php?name=CausalSME.

Baselines: We compare our approach against state-of-the-art differentiable causal discovery methods, including DICD [38] for multiple environment settings, DARING [10] and ReScore [42] for experiments with heterogeneous noise, and NOTEARS [46] (NOTEARS-MLP [47]) and DAG-GNN [41] for experiments with a single environment. In addition, we set up a baseline called CausalSME/MI which operates under the ideal setting of known environmental information. This shows how well `CausalSME` would perform when the environment to which each sample belongs to is known.

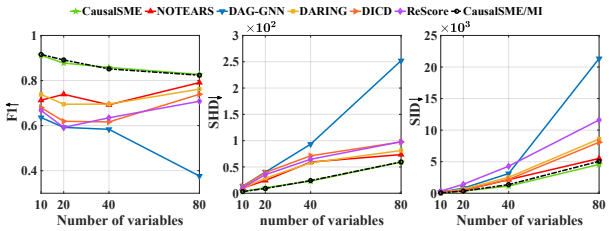
Synthetic data: We adopt the widely used Erdos-Renyi (ER) model to generate random DAGs with d observed variables. Specifically, we construct d unobserved variables, each of which directly affects one observed variable. The observed data is then generated according to a linear model as described in Eq. (3), where each value of the matrices A and B is uniformly sampled in $(-2, -0.5) \cup (0.5, 2)$. The nonlinear data is generated by the causal additive model as in Eq. (2), where f_j is a stochastic nonlinear function. To simulate different environments for the unobserved variables, we use a Gaussian distribution $N(\mu, \delta)$ with randomly selected means $\mu \in [-2, 2]$ and standard deviations $\delta \in [0.1, 2]$. We only take the values of the observed variables to estimate the causal structure.

Metrics: We use three canonical metrics to evaluate the discovered graphs: F1 is the harmonic average of precision and recall; SHD is the structural Hamming distance that measures the number of error edges; and SID is the structural intervention distance that measures the number of erroneous interventional distributions.

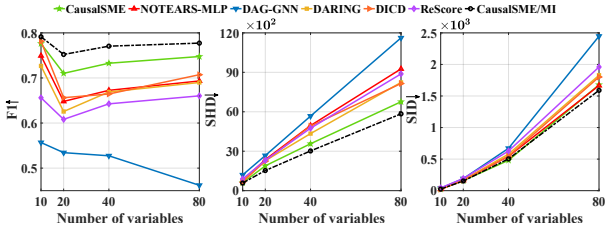
Hyper-parameters: We specify hyper-parameter settings for all baselines and for `CausalSME` in Appendix A. In addition, we present experiments on hyper-parameter analysis for `CausalSME` in Appendix B, to select appropriate hyper-parameter values and explain their meaning. In subsequent experiments, we use the hyper-parameter settings stated in the Appendix.

5.1 Results on Unobserved Distribution Shifts

To see how `CausalSME` performs in mixed environments with distribution shifts, we test all methods for true graphs with {10, 20, 40, 80} nodes. We set 5 environments, each with 30% of the variables affected by shifted unobserved variables, and 400 samples. We show the results in Figure 2 and make the following observations.



(a) Linear setting.



(b) Nonlinear setting.

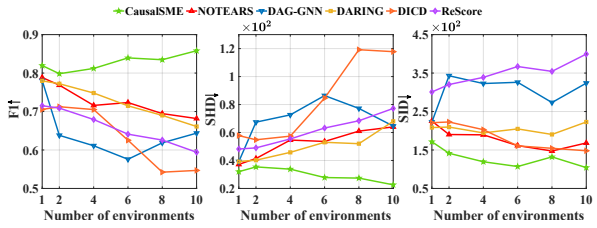
Figure 2: Results on unobserved distribution shifts.

(i) For all metrics, `CausalSME` outperforms baseline methods in finding linear causality. DAG-GNN performs poorly and occasionally fails to train, because it is unable to effectively learn a causal model with a matrix that cannot cover the causal effects of unobserved variables. DARING and ReScore (which deal with heterogeneous noise distribution) and DICD (which deals with real multi-environment information) do not achieve significantly good results. This is because our experiments explore a problem which is more complex than unobserved shifts, and heterogeneous noise can be seen as a special case of this problem. Our `CausalSME` explicitly models unobserved causal effects, it can directly learn unbiased causal models separated from distribution shifts, and uncovers true causal graphs. (ii) `CausalSME` brings significant improvements in all metrics with nonlinear data. Discovering causal structures from nonlinear data with unobserved variable shifts is more difficult, because of the inherent complexity of nonlinear models. As a result, all methods show a sharp drop in performance. NOTEARS-MLP is unable to identify or constrain distribution shifts and blend them into the learned causal model, thus retaining more edges in order to include the correct causal structure. DARING, DICD, and ReScore all tend to remove more error edges to ensure the accuracy of structure learning. This is because they introduce various constraints to avoid over-reconstructed models, which results in sparse graphs. Limited to a single model, these methods cannot balance different distributions of multiple environments and abandon possible causal structures. `CausalSME` achieves significantly better performance than DICD without knowing a-priori the environment information.

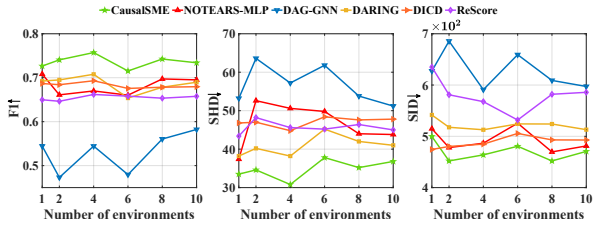
The advantage of CausalSME is that it can comprehensively mitigate missing, redundant, and reverse edges caused by unobserved variable shifts in multiple environments. This is because the tailored VAEs of CausalSME enhance the learning ability of the whole model for multi-environment distributions while ensuring a unique causal model.
(iii) As the graph size increases, the F1 of most methods increases. This is primarily due to the fact that the causal effect of unobserved variables is diluted by the large number of causal relationships as the graph size increases. Even in this case, CausalSME outperforms other methods by a large margin.
(iv) Our pseudo-environment partition strategy is effective. We provide CausalSME with accurate environment information (i.e. the true sample partition), and define this baseline as CausalSME/MI. CausalSME slightly loses to CausalSME/MI, which has accurate environmental information, and outperforms other baselines. We tested the statistical significance of the results of all methods with a p-value of 0.05. The test results showed that CausalSME and CausalSME/MI are significantly better than other baselines, and there is no significant difference between their results. These results prove that our pseudo-environment partition effectively distinguishes samples from different environments. At the same time, CausalSME can find even more accurate causal structures when accurate environmental information is available.

5.2 The Impact of Mixed Environments

To evaluate the robustness of CausalSME with respect to the number of environments, we fix the total sample size to 2000 and adjust the number of environments within $\{1, 2, 4, 6, 8, 10\}$. We show the experimental results on graphs with 40 nodes in Figure 3.



(a) Linear setting.



(b) Nonlinear setting.

Figure 3: Results on the number of environments.

(i) Most methods suffer from a decreased performance as the number of environments increases with linear data, while CausalSME gains in performance. This is because the baselines do not recognize the effects of unobserved variables, and it becomes increasingly difficult for them to learn a single causal model that fits the heterogeneous distribution across multiple environments as the number of environments increases. CausalSME effectively learns unobserved variable shifts in multiple environments via VAEs. As such it benefits from the additional environments to balance distribution differences across environments, and learn an unbiased causal model. We emphasize that we do *not* provide any environment information to CausalSME. The performance margin proves that our pseudo-environment partition strategy works well with a varying number of environments.
(ii) CausalSME achieves the best F1 and SHD in the nonlinear setting. Except for DAG-GNN and ReScore, the other methods are insensitive to the number of environments. This might be because the nonlinearity of the unobserved distribution shifts may pose a harder challenge to the causal discovery than the changes in the number of environments, and the impact of environmental variation is weakened. Thanks to the pseudo-environment partition strategy, CausalSME again outperforms the baselines in the nonlinear setting, and maintains a stable performance against the different numbers of environments.
(iii) CausalSME works well also in a single environment setting. We find that homogeneous shifts also undermine the causal discovery. We apply the same shift to all samples when working with one environment. The baselines' performance is clearly inferior to that of CausalSME. This shows that CausalSME works well not only with multi-environment data. It is able to learn unobserved causal effects in a variety of settings and uncovers the causal graphs.
(iv) CausalSME achieves a good balance between finding correct edges and filtering out incorrect ones, hence its F1 and SHD results are better than its competitors. The pseudo-environment partition strategy enables CausalSME to recognize distribution differences among environments and to retain more real edges. Furthermore, CausalSME captures multi-environment unobserved causal effects via tailored VAEs, and thus learns unbiased causal models and avoids spurious edges. CausalSME's SID only shows the improvement due to the additional correct edges, as missing correct edges have a greater influence on SID than redundant incorrect ones, and its gap with other methods is not as prominent as for F1 and SHD.

5.3 Robustness to Imbalance

We investigate the robustness of CausalSME with respect to multiple environments with imbalanced sample size. We set up five environments, each containing a random sample size, and a total sample size of 2000. Specifically, we divide all 2000 samples into 40 packages, each containing 50 samples, and randomly distribute these 40 packages to 5 environments (each environment holds at least one package) to simulate environments with significantly different sample sizes. We also test all methods on graphs with 40 nodes and show the results in Table 1.

CausalSME maintains a good performance with an imbalanced sample size. Imbalanced samples have little effect on NOTEARS (NOTEARS-MLP), DARING, and ReScore, because these algorithms do not consider the environment a sample belongs to. Unlike DICD, whose performance is largely degraded compared to the case of

Table 1: Results on imbalanced sample sizes. The boldfaced result highlights the method that performed the best among all compared methods.

linear	F1↑	SHD↓	SID↓
CausalSME	**82.97±5.0**	**28.50±11.4**	**148.20±77.1**
NOTEARS	72.16±3.7	53.30±10.6	186.00±55.0
DARING	71.12±3.8	53.50±13.8	262.60±135.8
ReScore	63.99±4.4	62.70±14.8	438.2±151.1
DAG-GNN	58.52±7.3	83.56±23.1	303.11±94.7
DICD	57.16±9.2	104.10±35.2	209.90±65.8
nonlinear	F1↑	SHD↓	SID↓
CausalSME	**74.37±5.5**	**34.89±6.6**	**442.44±115.7**
NOTEARS-MLP	68.62±7.0	47.00±11.3	476.30±142.2
DARING	67.57±2.8	43.40±5.8	541.70±124.6
ReScore	66.74±6.1	45.90±7.7	593.40±163.2
DAG-GNN	49.01±8.6	59.00±11.5	678.00±154.6
DICD	64.59±7.0	54.20±10.9	521.00±146.7

uniform sample allocation, CausalSME is robust against imbalanced sample sizes and maintains a stable causal discovery accuracy. DICD directly combines multi-environment data to learn a unified causal graph. Therefore, it is prone to ignore the distribution of an environment with a small sample size, and thus obtains an inaccurate causal model. Although CausalSME also focuses on multi-environment data, it reconstructs unobserved causal effects to eliminate shifts in data, which reduces the distribution differences among environments and enables the model to better combine multi-environment data. These results validate that reconstructing unobserved causal effects can more accurately and reliably recover the causal model than directly combining multi-environment data, which makes CausalSME more robust than DICD.

5.4 Robustness to Shift Sizes

We further verify the robustness of CausalSME to the shift size. We gradually increase the shift of the mean and variance of the unobserved variables from one to five (the maximum shift is more than twice the shift in previous experiments), and evaluate CausalSME in different shift settings with five environments and 2000 samples. We conduct experiments on graphs with 40 nodes and show the results in Figure 4. In addition, we show the performance of CausalSME in partitioning pseudo environments and inferring unobserved causal effects in Appendix C.

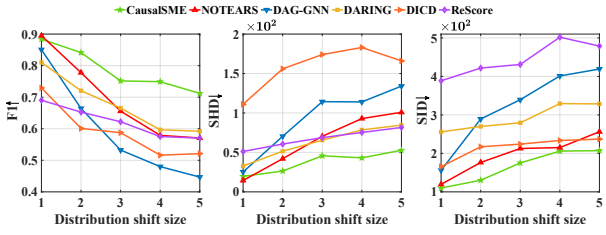


Figure 4: Results on the shift sizes.

CausalSME is the most robust method against the gradually increasing shift. All methods achieve the best performance on data with small shifts, especially NOTEARS and DAG-GNN, which are close in performance to CausalSME. This also validates the causal discovery capability of NOTEARS and DAG-GNN on data with no or small shifts. However, as the shift increases, the performance of the baselines rapidly decreases. The larger shift distorts the observed distribution and poses greater difficulties for causal discovery. In particular, ReScore achieves good F1 and SHD, but the worst SID. We observe that ReScore estimates bidirectional edges or cycle structures, which illustrates that ReScore only builds structures similar to the true graph, but does not learn the correct causality. The performance of CausalSME degrades slowly compared to the baselines. By considering unobserved causal effects, CausalSME infers the shifted effects and mitigates the impact of shifts to learn a more accurate causal graph.

We have checked the correspondence between the real environments and the pseudo environments partitioned by CausalSME, and we have compared the real and inferred unobserved causal effects. The analysis is presented in Appendix C. The results demonstrate that the pseudo environments identified by CausalSME closely match the real environments, and thus CausalSME identifies reliable unobserved causal effects.

5.5 Real Data

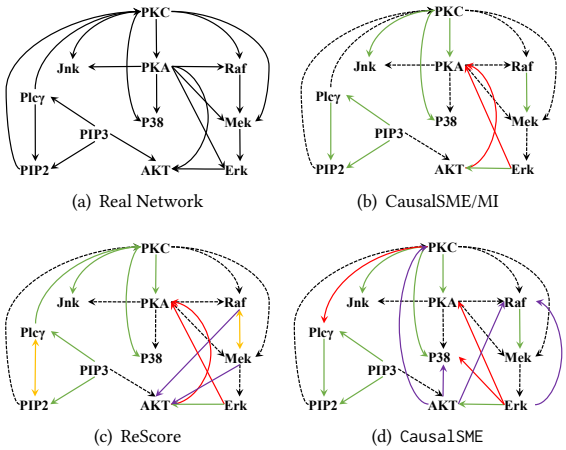
To see how well CausalSME performs on real-world tasks, we evaluate a protein and phospholipid signaling network on the widely used Sachs dataset [30]. It contains 7466 samples and 11 variables and provides a real graph with 20 edges. This dataset actually consists of 9 observed datasets with different stimulations, which can be thought of as coming from different environments and being influenced by shifted unobserved variables. We conduct experiments with all samples and show results in Table 2.

Table 2: Results on the Sachs dataset. The boldfaced result highlights the method that performed the best among all compared methods.

	F1↑	SHD↓	SID↓	Total edges
CausalSME	45.71	16	71	15
CausalSME/MI	**53.33**	**12**	**61**	10
ReScore	45.71	15	77	13
DARING	40.00	18	69	20
DICD	38.71	18	73	11
NOTEARS	35.56	24	70	25
DAG-GNN	28.59	30	75	29

CausalSME achieves competitive results without environmental information and finds the most accurate causal structure once environmental information is available. NOTEARS and DAG-GNN have the lowest results, because they are not good at identifying causal graphs from data with distribution shifts. CausalSME has a performance similar to ReScore and to DARING, which shows that CausalSME can identify heterogeneous data well in the absence of environmental information. By introducing environmental information, DICD greatly reduces the total number of edges, but discards

some correct edges with lower F1 and higher SHD. In contrast, CausalSME/MI estimates the causal graph with fewer edges under real environmental information while maintaining the highest F1. These findings confirm the ability of our CausalSME to capture unobserved shifts and restore the original causal structure.



(a) Real Network (b) CausalSME/MI

(c) ReScore (d) CausalSME

Figure 5: Results on the Sachs dataset. (a) is the true causal graph. (b), (c), and (d) are the estimated graphs of CausalSME/MI, ReScore, and CausalSME, respectively, where the reverse edges are red, the redundant edges are purple, the bidirectional edges are yellow, and the missing edges are dashed lines.

We further show the estimated graphs of CausalSME, CausalSME/MI, and ReScore, and the real graph in Figure 5. We see that the result of ReScore contains bidirectional edges, which means that ReScore's acyclic constraint is violated. Although CausalSME has one more incorrect edge with respect to ReScore, it finds a more reliable acyclic graph. CausalSME/MI further removes most of the reverse and redundant edges. Overall, our method (CausalSME or CausalSME/MI) produces more accurate and reliable results than the baselines.

6 Conclusion

In this paper, we focus on reconstructing biased causal models due to distribution shifts of unobserved variables, and propose CausalSME to tackle this unexplored, challenging but practical problem. CausalSME can effectively capture unobserved variable shifts from mixed multi-environment data with a combination of variational inference and observational reconstruction, and naturally identify environmental information based on inferred unobserved causal effects. It further enforces mutually independent unobserved causal effects and reconstructed residuals, and thus ensures the identifiability of causal graphs. Extensive experiments on synthetic and real data show that CausalSME has outstanding accuracy and robustness for causal discovery in shifted multiple environments with unobserved causal effects.

Acknowledgments

This work is supported by National Key Research and Development Program of China (No. 2023YFF0725500), NSFC (62031003, 62272276 and 62432006), Shandong Provincial Natural Science Foundation (No. ZR2024JQ001) and Taishan Scholars Program (No. tsqn202306007).

References

[1] David Arthur and Sergei Vassilvitskii. 2006. *k-means++: The advantages of careful seeding*. Technical Report. Stanford.
[2] Peter Bühlmann, Jonas Peters, and Jan Ernest. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* 42, 6 (2014), 2526–2556.
[3] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16, 5 (1995), 1190–1208.
[4] David Maxwell Chickering. 2002. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research* 2 (2002), 445–498.
[5] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3 (2002), 507–554.
[6] Jukka Corander, William P Hanage, and Johan Pensar. 2022. Causal discovery for the microbiome. *The Lancet Microbe* 3, 11 (2022), e881–e887.
[7] P Daniusis, D Janzing, J Mooij, J Zscheischler, B Steudel, K Zhang, and B Schölkopf. 2010. Inferring deterministic causal relations. In *26th Conference on Uncertainty in Artificial Intelligence*. 143–150.
[8] A Ghassami, N Kiyavash, B Huang, and K Zhang. 2018. Multi-domain causal structure learning in linear systems. In *Advances in Neural Information Processing Systems*. 6266–6276.
[9] Shawkat Hammoudeh, Ahdi Noomen Ajmi, and Khaled Mokni. 2020. Relationship between green bonds and financial and environmental variables: a novel time-varying causality. *Energy Economics* 92 (2020), 104941.
[10] Yue He, Peng Cui, Zheyan Shen, Renzhe Xu, Furui Liu, and Yong Jiang. 2021. DARING: Differentiable causal discovery with residual independence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 596–605.
[11] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. 2018. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1551–1560.
[12] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* 182, 183 (2012), 1–31.
[13] Dominik Janzing and Bernhard Schölkopf. 2018. Detecting non-causal artifacts in multivariate linear regression models. In *International Conference on Machine Learning*. 2245–2253.
[14] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[15] Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.
[16] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. 2019. Gradient-based neural dag learning. In *International Conference on Learning Representations*.
[17] Jiaxuan Liang, Jun Wang, Guoxian Yu, Carlotta Domeniconi, Xiangliang Zhang, and Maozu Guo. 2023. Gradient-based local causal structure learning. *IEEE Transactions on Cybernetics* 54, 1 (2023), 486–495.
[18] Jiaxuan Liang, Jun Wang, Guoxian Yu, Wei Guo, Carlotta Domeniconi, and Maozu Guo. 2023. Directed acyclic graph learning on attributed heterogeneous network. *IEEE Transactions on Knowledge and Data Engineering* 35, 10 (2023), 10845–10856.
[19] Songtao Lu and Tian Gao. 2023. Meta-DAG: Meta causal discovery via bilevel optimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 1–5.
[20] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. 1, 14 (1967), 281–297.
[21] Osman A Mian, Alexander Marx, and Jilles Vreeken. 2021. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8975–8982.
[22] Joris M Mooij, Sara Magliacane, and Tom Claassen. 2020. Joint causal inference from multiple contexts. *Journal of Machine Learning Research* 21 (2020), 1–108.
[23] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress* 19, 2 (2000), 3.
[24] Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. 2022. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems*. 10904–10917.
[25] Jonas Peters and Peter Bühlmann. 2013. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* 101, 1 (2013), 219–228.

[26] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78, 5 (2016), 947–1012.
[27] Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2, 7 (2020), 369–375.
[28] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 2017. 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 (2017), 121–129.
[29] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.
[30] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 5721 (2005), 523–529.
[31] Mauro Scanagatta, Cassio P de Campos, Giorgio Corani, and Marco Zaffalon. 2015. Learning Bayesian networks with thousands of variables. In *Advances in Neural Information Processing Systems*. 1864–1872.
[32] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7 (2006), 2003–2030.
[33] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT Press.
[34] Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal inference in the presence of latent variables and selection bias. In *11th Conference on Uncertainty in Artificial Intelligence*. 499–506.
[35] Chandler Squires, Yuhao Wang, and Caroline Uhler. 2020. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*. 1039–1048.
[36] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65 (2006), 31–78.
[37] Yixin Wang and David M Blei. 2019. The blessings of multiple causes. *J. Amer. Statist. Assoc.* 114, 528 (2019), 1574–1596.
[38] Yu Wang, An Zhang, Xiang Wang, Yancheng Yuan, Xiangnan He, and Tat-Seng Chua. 2022. Differentiable invariant causal discovery. *arXiv preprint arXiv:2205.15638* (2022).
[39] Dezhi Yang, Xintong He, Jun Wang, Guoxian Yu, Carlotta Domeniconi, and Jinglin Zhang. 2024. Federated causality learning with explainable adaptive optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 16308–16315.
[40] Dezhi Yang, Guoxian Yu, Jun Wang, Zhengtian Wu, and Maozu Guo. 2023. Reinforcement causal structure learning on order graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 10737–10744.
[41] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. DAG-GNN: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*. 7154–7163.
[42] An Zhang, Fangfu Liu, Wenchang Ma, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. 2023. Boosting causal discovery via adaptive sample reweighting. In *International Conference on Learning Representations*.
[43] Hao Zhang, Shuigeng Zhou, and Jihong Guan. 2018. Measuring conditional independence by independent residuals: theoretical results and application in causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2029–2036.
[44] Jiji Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* 172, 16-17 (2008), 1873–1896.
[45] Kun Zhang and Aapo Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence*. 647–655.
[46] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*. 9492–9503.
[47] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. 2020. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*. 3414–3425.

A Hyper-parameter Settings

We start by listing the hyper-parameters shared by all methods as follows:

The MLP structure of reconstruction model has input dimension d and output dimension 1 without hidden layer on the linear setting, and has an additional hidden layer with 10 units on the nonlinear setting, d is the number of observed variables. Both l_1 and l_2 regularization coefficients of the model are 0.01. Since the reconstruction models of DARING, ReScore, NOTEARS, NOTEARS-MLP and `CausalSME` are optimized by the L-BFGS-B algorithm [3], there is no need to set their learning rates. The convergence tolerance of equality constraints is 10^{-8}, the pruning threshold is $\tau = 0.3$. In particular, DARING and `CausalSME` both introduce the independence approximate model. Their independent approximate models are composed of d MLPs with input dimension d and output dimension 1. The penalty coefficient of the independence approximation is 0.01, the optimizer learning rate of the independence approximate model is 0.01 and the momentum is 0.9.

In addition to the shared hyper-parameters, we list the special hyper-parameters of each algorithm as follows:

• DAG-GNN: The number of hidden layers of the encoder is 1, the number of hidden units of the encoder is 64, the number of hidden layers of the decoder is 1, the number of hidden units of the decoder is 64, and the learning rate is 0.003.

• DICD: the penalty term coefficient is 0.1 and the learning rate is 0.001.

• ReScore: The Gumbel softmax temperature is 20, the number of iterations for adaptive reweight is 20, the epoch begin to reweighting is 0, the learning rate for reweight model is 0.001, and the l_1 regularization coefficient for reweighting model is 0.001. The MLP structure of reweighting model has input dimension d and output dimension 1 without hidden layer on the linear setting, and has an additional hidden layer with $2d$ units on the nonlinear setting, d is the number of observed variables.

• `CausalSME`: The MLP structure of variational autoencoder has input dimension d and output dimension 1 without hidden layer on the linear setting, and has an additional hidden layer with $m = 10$ units on the nonlinear setting, d is the number of observed variables. For each alternate training, the reconstruction model is optimized for $t_1 = 10$ epochs and the variational autoencoder is optimized for $t_2 = 10$ epochs.

For all baseline methods, our hyper-parameter settings are consistent with the default parameters in their exposed codes. We set the hyper-parameters of `CausalSME` such as model structure and learning rate to be the same as other methods to ensure fair comparisons. In particular, for the hyper-parameters t_1 and t_2 specific to `CausalSME`, we select appropriate values for them in hyper-parameter analysis experiments (Appendix B) and explain why.

B Hyper-parameter Selection

To search the best number of epochs (t_1 and t_2) for the reconstruction model and the variational autoencoder during each alternate training, we test `CausalSME` for different t_1 and t_2 on graphs with 40 nodes. We experiment with t_1 and t_2 selected from $\{1, 10, 20, 30\}$ and report the metrics and runtime of results in Figure S1.

We observe that `CausalSME` improves as t_2 increases. This indicates that the fully trained variational autoencoder can improve the performance of `CausalSME`. However, the runtime of `CausalSME` also increases rapidly with the increased t_2. In addition, we find that `CausalSME` has significantly better accuracy in the setting $t_1 = 10$ than in other settings. If t_1 is too small, the reconstruction model is not fully trained. Conversely, if the reconstruction model is already stable with large t_1, the variational autoencoder does not interact

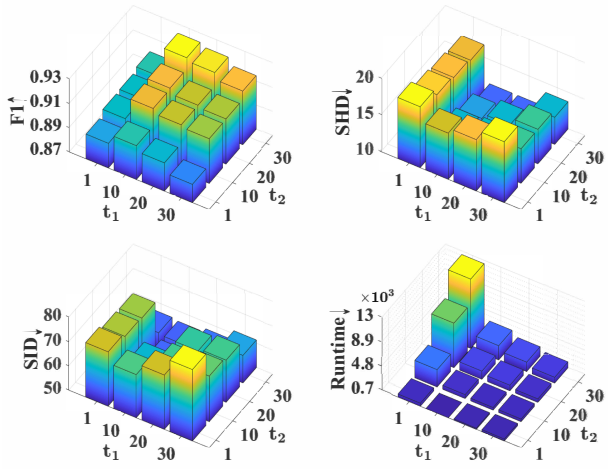


Figure S1: Results on the different hyper-parameters.

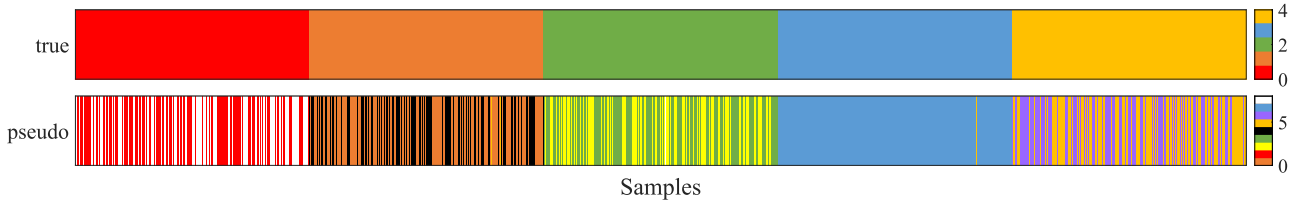
with the reconstruction model and loses the ability to infer unobserved causal effects. In summary, we set $t_1 = 10$ and $t_2 = 10$ for both the accuracy and efficiency of CausalSME.

C Robustness to Shift Size

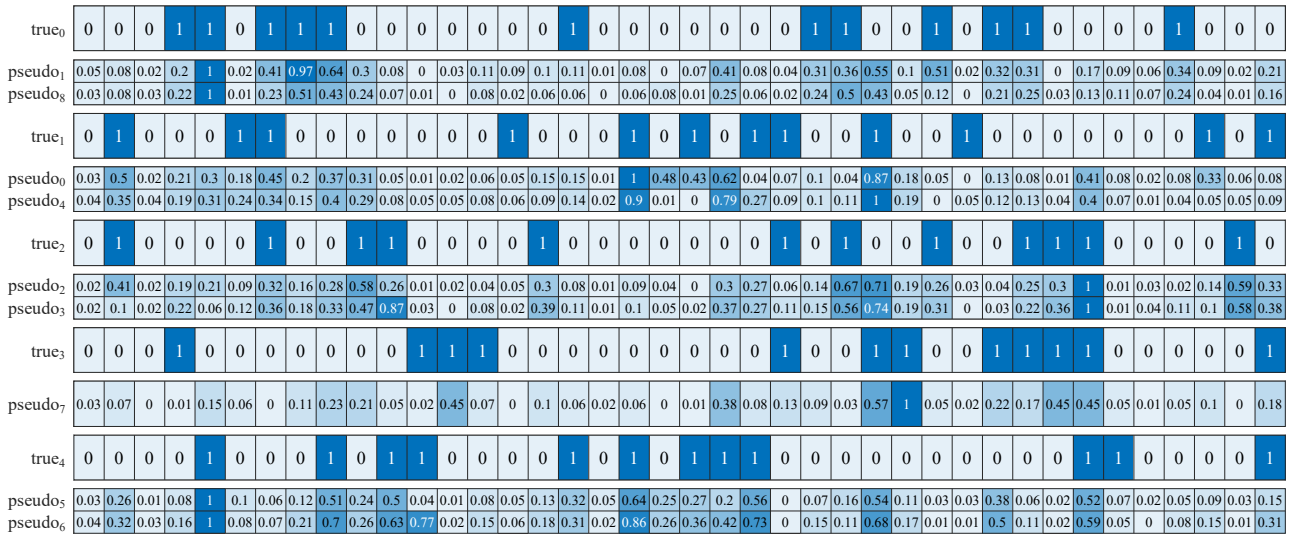
We study the correspondence between the real environments and the pseudo environments identified by CausalSME, and we compare the real and inferred unobserved causal effects. Specifically, we use a heat map to represent all samples, where samples belonging to different real and pseudo environments are set to different colors. For unobserved causal effects, we normalize the inferred unobserved causal effects into the interval $[0, 1]$ and present them with a vector, where each element of the vector represents the inferred magnitude of the unobserved causal effect on the observed variable. Then, we represent the true unobserved causal effects with a binary vector, where the element with a value of 1 indicates that there is an unobserved variable with shifted causal effect on the corresponding observed variable, otherwise there is no shifted causal effect. We reveal the results in Figure S2.

We observed that CausalSME recognizes more pseudo environments than real ones. However, this does not mean that the pseudo-environment partition strategy of CausalSME fails. In fact, samples belonging to the same real environment are basically assigned to one or a group of pseudo environments, which indicates that there is a clear correspondence between the pseudo environment and the real one. In other words, our pseudo-environment partition strategy successfully groups samples from different real environments, which supports CausalSME to identify unobserved causal effects in mixed multiple environments.

The results in Figure 2(b) show that the unobserved causal effects inferred by CausalSME mostly coincide with the real causal effects. This further demonstrates the effectiveness of CausalSME in identifying unobserved causal effects across multiple environments.



(a) Real environments ($K = 5$) vs. Pseudo environments ($K' = 9$). CausalSME infers pseudo environments that have a clear correspondence with real environments. We define the pseudo environments inferred by CausalSME as $\{pseudo_{k'}\}_{k'=0}^{8}$ and the true environments as $\{true_k\}_{k=0}^{4}$, where the pseudo environments $pseudo_1$ and $pseudo_8$ correspond to the real environment $true_0$, the pseudo environments $pseudo_0$ and $pseudo_4$ correspond to the real environment $true_1$, the pseudo environments $pseudo_2$ and $pseudo_3$ correspond to the real environment $true_2$, the pseudo environment $pseudo_7$ corresponds to the real environment $true_3$, and the pseudo environments $pseudo_5$ and $pseudo_6$ correspond to the real environment $true_4$. CausalSME recognize reliable pseudo environments.



(b) Real unobserved effects vs. inferred unobserved effects. We show the true unobserved causal effects on the observed variables in the real environments $\{true_k\}_{e=0}^{4}$ (0 in the heat map indicates that the observed variable is not affected by the unobserved causal effects and 1 is the opposite), and the unobserved causal effect size inferred by CausalSME in pseudo environments $\{pseudo_{k'}\}_{e'=0}^{8}$ (the darker the blue in the heat map, the larger the inferred unobserved causal effect). The results show that the dark blue areas in the pseudo-environment heat map basically match the areas with the value of 1 in the real-environment heat map. CausalSME effectively infer unobserved causal effects on observed variables.

Figure S2: Results on pseudo environments partition and unobserved effects inference.