Discovering Mixtures of Structural Causal Models from Time Series Data

Sumanth Varambally [1] **Yi-An Ma** [1] **Rose Yu** [2 1]

Abstract

Discovering causal relationships from time series data is significant in fields such as finance, climate science, and neuroscience. However, contemporary techniques rely on the simplifying assumption that data originates from the same causal model, while in practice, data is heterogeneous and can stem from different causal models. In this work, we relax this assumption and perform causal discovery from time series data originating from *a mixture of causal models*. We propose a general variational inference-based framework called MCD to infer the underlying causal models as well as the mixing probability of each sample. Our approach employs an end-to-end training process that maximizes an evidence-lower bound for the data likelihood. We present two variants: MCD-Linear for linear relationships and independent noise, and MCD-Nonlinear for nonlinear causal relationships and history-dependent noise. We demonstrate that our method surpasses state-of-the-art benchmarks in causal discovery tasks through extensive experimentation on synthetic and real-world datasets, particularly when the data emanates from diverse underlying causal graphs. Theoretically, we prove the identifiability of such a model under some mild assumptions. Implementation is available at `https://github.com/Rose-STL-Lab/MCD`.

1. Introduction

Causal discovery extends and complements the scope of prediction-focused machine learning with the notions of controllability and counterfactual reasoning. It aims to infer the underlying causal structure among observed variables in the data (Spirtes et al., 2000). Many methods have been developed for causal discovery from time-series data based on structural causal models (SCMs) (Hyvärinen et al., 2010; Pamfil et al., 2020; Yao et al., 2022; Gong et al., 2022), conditional independence tests (Malinsky & Spirtes, 2018; Runge et al., 2019; Runge, 2020), as well as the weaker notion of Granger causality (Granger, 1969; Khanna & Tan, 2020; Tank et al., 2021).

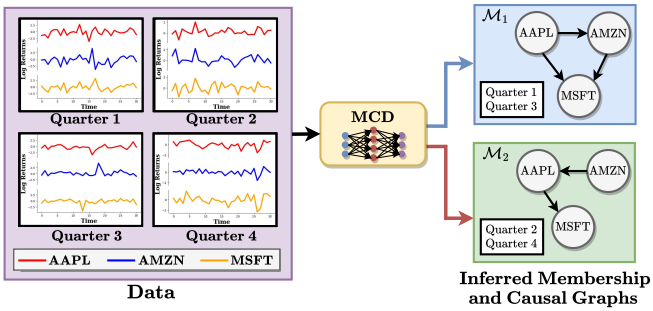

Figure 1. MCD discovers multiple causal graphs from time-series data by determining the mixture component membership for each sample and inferring one graph per mixture component.

Unfortunately, the existing methods predominantly assume that a single causal model applies to the entire data set. In machine learning tasks, data are often multi-modal and highly heterogeneous. For example, gene regulatory networks are particular to different cells at different developmental stages. But during experiments for cell lineage, one can only track the RNA expression levels of different cells with related but distinct gene regulatory networks, since every measurement destroys the cell (Qiu et al., 2022). Similarly, stock market interactions can vary over different time periods. Using a single causal model to explain the data can result in oversimplification and an inability to capture diverse causal mechanisms.

The task of discovering mixtures of causal graphs from data has received limited attention in the literature. Recent work, such as Thiesson et al. (1998); Saeed et al. (2020); Markham et al. (2022), have tackled the challenge of inferring causal models from mixture distributions. However, these approaches primarily focus on independent data and do not specifically address time series data. Löwe et al. (2022) touched upon this problem by inferring a per-sample summary graph in an amortized framework, but their ap-

[1]Halıcıoğlu Data Science Institute, University of California, San Diego, La Jolla, USA [2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, USA. Correspondence to: Sumanth Varambally <svarambally@ucsd.edu>.

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

proach is limited to inferring Granger causal relationships and does not account for instantaneous effects.

In this paper, we investigate a more realistic setting in which time series data is generated from a mixture of unknown structural causal models (SCMs). We assume there are K mixture components. The membership of which time series comes from which SCM component is also unknown. Our goal is to perform causal discovery by learning the complete SCMs as well as the corresponding membership for each time series sample. A complete SCM includes both the causal graph and its associated functional equations. Figure 1 summarizes the problem setting that MCD tackles.

We propose a variational inference-based framework, Mixture Causal Discovery (MCD), for causal discovery from heterogeneous time series data. Our approach learns the complete SCM and the mixture membership of each sample. To compute the intractable posterior, we derive and optimize a novel Evidence Lower Bound (ELBO) of the data likelihood. We present two variants: (1) MCD-Linear, which models linear relationships and independent noise, and (2) MCD-Nonlinear, which uses neural networks to model functional relationships and history-dependent noise. Theoretically, we characterize a necessary and sufficient condition for the identifiability of a mixture of linear Gaussian SCMs and derive a sufficient condition for the identifiability of general SCMs under some mild assumptions. In summary, our contributions are as follows:

- We tackle the realistic and challenging setting of discovering mixtures of SCMs for time series data with additive noise. We propose a novel variational inference approach, MCD, to simultaneously infer the complete SCM and the mixture membership of each sample.
- Theoretically, we show that under mild assumptions, mixtures of identifiable causal models are identifiable for both linear Gaussian and general SCMs. We also derive the relationship between our proposed ELBO objective and true data likelihood.
- We derive two instances of MCD: (1) MCD-Linear, which models linear relationships with independent noise, and (2) MCD-Nonlinear, which models nonlinear relationships with history-dependent noise.
- Experimentally, we demonstrate the strong performance of our method on both synthetic and real-world datasets. Notably, MCD can accurately infer the mixture causal graphs and mixture membership information, even when the number of SCMs is misspecified.

2. Related work

In this section, we provide a focused literature survey on causal discovery for time series and multiple causal models.

Causal Discovery for time series data. Many works on time series causal discovery use the notion of Granger causality (Granger, 1969). Tank et al. (2021) use component-wise Multi-Layer Perceptrons (cMLP) or Long Short Term Memory Networks (cLSTMs) with sparsity constraints on weight matrices to infer non-linear Granger causal links. Khanna & Tan (2020) use component-wise Statistical Recurrent Units (SRU), which incorporate single and multi-scale summary statistics from multi-variate time series for Granger causal detection. Amortized Causal Discovery. (Löwe et al., 2022) aims to infer Granger causality from time series data using a variational auto-encoder framework in conjunction with Graph Neural Networks. However, Granger causality is not true causality; it only indicates the presence of a predictive relationship. Granger causality also cannot account for instantaneous effects, latent confounders, or history-dependent noise (Peters et al., 2017).

In contrast to Granger Causality, the framework of SCMs can theoretically model instantaneous effects, latent confounders, and history-dependent noise. Hyvärinen et al. (2010) incorporate vector autoregressive models to the LiNGAM (Shimizu et al., 2006) algorithm to propose the VARLiNGAM algorithm for time series data. DYNOTEARS, proposed in Pamfil et al. (2020), uses the NOTEARS DAG constraint (Zheng et al., 2018) to learn a dynamic Bayesian network. However, VARLiNGAM and DYNOTEARS only account for *linear* causal relationships and do not account for history-dependent noise. Runge et al. (2019) extend the PC algorithm to time series data with the PCMCI method. PCMCI$^+$ (Runge, 2020) can handle instantaneous edges. Rhino (Gong et al., 2022) learns the temporal adjacency matrix given observational data while modeling the exogenous history-dependent noise distribution. However, these methods assume a single causal graph for the whole data distribution.

Learning mixtures of causal models. Several works focus on the problem of causal discovery from heterogeneous independent data, but not time series. Thiesson et al. (1998) use a heuristic search-and-score method to learn mixture of directed acyclic graph (DAG) models. However, this method only models linear causal relationships and Gaussian noise. Saeed et al. (2020) use the FCI algorithm (Spirtes, 2001) to recover the maximal ancestral graph (MAG) and use it to detect variables with varying conditional distributions across the mixture components. Strobl (2022) propose using longitudinal data, i.e., data about the same variables measured across different, potentially irregularly-spaced points of time, to infer a mixture of DAGs. However, they do not infer causal relationships *across* time. Markham et al. (2022) define a distance covariance-based kernel used to cluster sample points based on the underlying causal model. Any causal discovery algorithm can be used to infer a DAG for each inferred cluster. Huang et al. (2019) pre-

sume 'individuals' have multiple associated samples and cluster them into groups. They learn individual-specific and shared causal structures across groups using a linear non-Gaussian mixture model.

Recent work (Huang et al., 2020; Zhou et al., 2022) tackled causal discovery from data governed by heterogeneous and non-stationary causal mechanisms over time. Unlike our approach, they infer a single graph for all samples. In contrast, we model the heterogeneity of causal models across samples. Our method learns one SCM per inferred mixture component and the mixture membership of each sample.

Another line of work deals with causal discovery from non-stationary time-series. Regime-PCMCI (Saggioro et al., 2020) assumes that a time series can be divided into different regimes (albeit with linear causal relationships) with distinct DAGs, and aims to infer the appropriate regime for each time index. PCMCI$_\Omega$ (Gao et al., 2023) uses conditional independence tests for semistationary time series, in which a finite number of causal models occur sequentially and periodically over time. This differs from our setting, in which we assume different causal graphs govern different samples. In certain scenarios, for example, analyzing climate patterns over different locations, our approach can pool information from different samples. On the other hand, Regime PCMCI and PCMCI$_\Omega$ would have to infer causal graphs from different locations separately.

Preliminaries. A Structural Causal Model (Pearl, 2009) (SCM) explicitly encodes the causal relationships between variables. Formally, an SCM over D variables consists of a 5-tuple $\langle \mathcal{X}, \varepsilon, \mathcal{F}, \mathcal{G}, P(\epsilon) \rangle$:

1. A set of endogenous (observed) variables $\mathcal{X} = \{X^1, X^2, \ldots, X^D\}$;
2. A set of exogenous (noise) variables $\varepsilon = \{\epsilon^1, \epsilon^2, \ldots, \epsilon^m\}$ which influence the endogenous variables. In general, $m \geq D$ due to latent confounders; but we assume causal sufficiency, i.e., $m = D$;
3. A Directed Acyclic Graph (DAG) $\mathcal{G}$, denoting the causal links amongst the members of $\mathcal{X}$;
4. A set of D functions $\mathcal{F} = \{f^1, f^2, \ldots, f^D\}$ determining $\mathcal{X}$ through the structural equations $X^i = f^i(\text{Pa}^i_\mathcal{G}, \epsilon^i)$, where $\text{Pa}^i_\mathcal{G} \subset \mathcal{X}$ denotes the parents of node i in graph $\mathcal{G}$ and $\epsilon^i \subset \varepsilon$;
5. $P(\epsilon)$, which describes a distribution over noise ϵ.

Given time series data $X \in \mathbb{R}^{D \times T}$, where T is the number of time steps, we can describe the causal relationships as:

$$X^d_t = f^d_t(\text{Pa}^d_\mathcal{G}(< t), \text{Pa}^d_\mathcal{G}(t), \epsilon^d_t), \tag{1}$$

where X^d_t denotes the value of the d^{th} variable of the time-series at time step t, $\text{Pa}^d_\mathcal{G}(< t)$ denote the parents of node d from the preceding time-steps (lagged parents) and $\text{Pa}^d_\mathcal{G}(t)$ are the parents at the current time-step (instantaneous parents). We assume that X_t is influenced by at most time-lag L preceding time steps, i.e. $\text{Pa}_\mathcal{G}(< t) \subseteq \{X_{t-1}, ..., X_{t-L}\}$. This is a common assumption, shared with Rhino (Gong et al., 2022), VARLiNGaM (Hyvärinen et al., 2010) and PCMCI (Runge et al., 2019) amongst others. The causal relationships can be modeled as a temporal adjacency matrix $\mathcal{G}_{0:L}$, where $\mathcal{G}_{1:L}$ represents the lagged relationships, and $\mathcal{G}_0$ represents the instantaneous edges. We set $\mathcal{G}^{i,j}_\tau = 1$ if $X^i_{t-\tau} \to X^j_t$, and 0 otherwise. In practice, we input L as a hyperparameter. We use the additive noise model due to its identifiability (Gong et al., 2022):

$$X_t = f_t(\text{Pa}_\mathcal{G}(< t), \text{Pa}_\mathcal{G}(t)) + \epsilon_t \tag{2}$$

We mute the variable index d for simplicity. $X_t \in \mathbb{R}^D$ represents the values of all variables at time t.

Our model shares similar assumptions to Gong et al. (2022), including causal stationarity, minimality and sufficiency, and some mild conditions on the likelihood function. These assumptions are restated in Appendix A.2.

3. Mixture Causal Discovery (MCD)

In this section, we detail our approach to learning mixtures of structural causal models from time-series data.

3.1. Problem setting.

We are given N samples of multi-variate time series with D variables, each of length T, denoted by $\left\{X^{1:D,(n)}_{1:T}\right\}_{n=1}^{N}$. We assume that each sample is generated by one of the K unknown SCMs, each consisting of a DAG $\mathcal{G}_k$, represented as a temporal adjacency matrix of size $(L+1) \times D \times D$, and its structural equations.

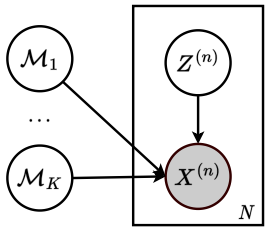

Figure 2. Probabilistic graphical model diagram of a mixture of SCMs. Shaded circles are observed and hollow circles are latent.

Our goal is to infer the DAG, the structural equations for all K SCMs, as well as the mixture membership of each sample in an unsupervised fashion. As shown in the graphical model of Figure 2, we represent the K SCMs as random variables $\mathcal{M}_{1:K}$. For each data sample indexed by n, we assign a categorical variable $Z^{(n)} \in \{1, \ldots, K\}$ to represent the membership of each sample to an SCM component.

We model each SCM $\mathcal{M}_k$ as a pair $(\mathcal{G}_k, \Theta_k)$, where $\mathcal{G}_k$ is the adjacency matrix, and Θ_k represents the trainable parameters of the structural equations and noise models. We model the causal relationships of $X^{(n)}_t$ under SCM k as:

$$X^{(n)}_t\Big|_k = f_k(\text{Pa}_{\mathcal{G}_k}(\leq t)) + g_k(\text{Pa}_{\mathcal{G}_k}(< t), \epsilon_t), \tag{3}$$

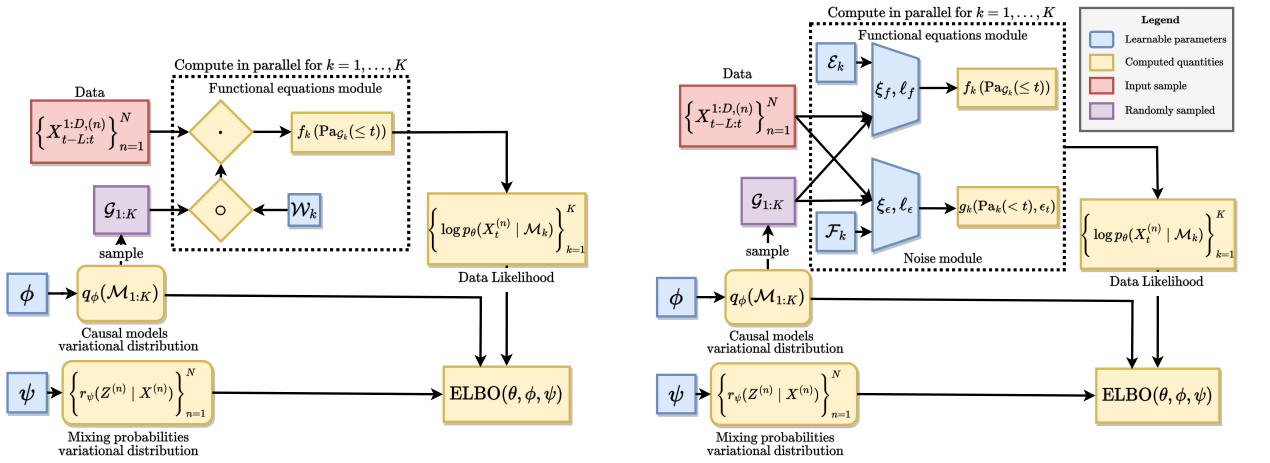

Figure 3. Overview of how the ELBO from Eq. (4) is calculated for (left) MCD-Linear, and (right) MCD-Nonlinear. Given time-series data $\left\{X_{t-L}^{1:D,(n)}\right\}_{n=1}^{N}$, and a DAG sample $\mathcal{G}_{1:K}$ from the variational distribution $q_\phi(\mathcal{M}_{1:K})$, we calculate the likelihood of the data under all the K causal models. The likelihood is weighted by the mixing probabilities $\left\{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)\right\}_{n=1}^{N}$ to calculate the ELBO.

where the function f_k denotes the structural equation model and g_k denotes the exogenous noise model.

3.2. Variational inference

Our goal is to infer the true posterior distribution $p\left(\mathcal{M}_{1:K} \mid X^{(1:N)}\right)$. However, it is intractable due to the presence of latent variables $\mathcal{M}_{1:K}$ and $\{Z^{(n)}\}$. We propose a variational inference framework to infer the parameters of the data generation process.

Proposition 1. *Under the data generation process described in Figure 2, the data likelihood admits the following evidence lower bound (ELBO):*

$$
\begin{aligned}
&\log p_\theta\left(X_{1:T}^{(1:N)}\right) \\
&\geq \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})}\left[\mathbb{E}_{r_\psi\left(Z^{(n)} \mid X_{1:T}^{(n)}\right)}\left[\log p_\theta\left(X_{1:T}^{(n)} \mid \mathcal{M}_{Z^{(n)}}\right)\right.\right. \\
&\left.\left. + \log p\left(Z^{(n)}\right) + H\left(r_\psi\left(Z^{(n)} \mid X_{1:T}^{(n)}\right)\right)\right]\right] \\
&+ \sum_{k=1}^{K} \mathbb{E}_{q_\phi(\mathcal{M}_k)}\left[\log p(\mathcal{M}_k)\right] + H\left(q_\phi(\mathcal{M}_k)\right) \qquad (4) \\
&\equiv \mathrm{ELBO}(\theta, \phi, \psi).
\end{aligned}
$$

For derivation details, we refer the reader to Section A.1. Here, $\log p_\theta\left(X_{1:T}^{(n)} \mid \mathcal{M}_{Z^{(n)}}\right)$ represents the marginal likelihood of $X^{(n)}$ under model $\mathcal{M}_{Z^{(n)}}$, $q_\phi\left(\mathcal{M}_k\right)$ represents the variational distribution of the causal model $\mathcal{M}_k$, and $r_\psi(Z^{(n)} \mid X^{(n)})$ represents the variational posterior distribution of the mixing rate for sample $X^{(n)}$. The number of causal models K is a hyperparameter. $p(Z)$ represents our prior belief about the membership of samples to the causal models, typically considered to be a uniform distribution.

3.3. Model implementation

We describe how to parameterize the different loss terms in Eq. (4). Figure 3 shows the ELBO calculation for MCD.

Causal model. We parameterize the variational distribution of the causal models as $q_\phi\left(\mathcal{M}_{1:K}\right) = \prod_{k=1}^{K} q_{\phi_k}\left(\mathcal{G}_k\right) \delta(\Theta_k)$, where δ represents the Dirac-δ function, and Θ_k represents the (learned) parameters of the structural equations and noise models. The distribution of the DAG adjacency matrix $q_{\phi_k}\left(\mathcal{G}_k\right)$ is represented as a product of independent Bernoulli distributions. The expectation over $q_\phi\left(\mathcal{M}_{1:K}\right)$ is computed by sampling once through Monte-Carlo sampling, using the Gumbel-Softmax trick (Jang et al., 2017).

Mixing probabilities. We specify the mixing rates variational distribution $r_\psi\left(Z^{(n)} \mid X^{(n)}\right)$ as a K-way categorical distribution, and learn it for each sample. We set $r_\psi\left(Z^{(n)} \mid X^{(n)}\right) = \mathrm{softmax}\left(\frac{w^{(n)}}{\tau_r}\right)$, where $w^{(n)} = \left[w_1^{(n)}, \ldots, w_K^{(n)}\right] \in \mathbb{R}^K$ are learnable weight parameters for each sample, and τ_r is a temperature hyperparameter. The number of parameters in the learned membership matrix grows linearly with the number of samples. This dependence on the sample size can be eliminated using a classifier that inputs a $D \times T$ multivariate time series and outputs a categorical distribution over the mixture components. However, in practice, we observe that the number of parameters

for the learned membership matrix W is quite small.

We also need an expectation of the likelihood term over $r_\psi\left(Z^{(n)} \mid X^{(n)}\right)$. This involves computing the marginal likelihood of each sample under all K causal models, theoretically requiring K times more operations than using a single model. In practice, we can calculate the marginal likelihoods under all causal models in parallel using PyTorch vectorization. Empirically, the computational complexity increase over using a single model leads to a modest increase in run-time, much less than a factor of K (Appendix B.8).

In theory, any likelihood-based causal structure learning algorithm can be used to implement the marginal likelihood loss term $\log p_\theta\left(X_{1:T}^{(n)} \mid \mathcal{M}_{Z^{(n)}}\right)$ in Eq. (4). Appendix C details the computation of $\log p_\theta\left(X_{1:T}^{(n)} \mid \mathcal{M}_{Z^{(n)}}\right)$ based on the model for f_k in Eq (3). We implement two variants of MCD to show the flexibility of our framework: (1) MCD-Linear, which handles linear causal relationships and independent noise, and (2) MCD-Nonlinear, which handles nonlinear structural equations and history-dependent noise.

MCD-Linear. We implement each of the K models using a linear model:

$$f_k^d\left(\mathrm{Pa}_{\mathcal{G}_k}(\leq t)\right) = \sum_{\tau=0}^{L}\sum_{j=1}^{D}\left(\mathcal{G}_k \circ \mathcal{W}_k\right)_\tau^{j,d} \times X_{t-\tau}^{j,(n)}, \quad (5)$$

where $\circ$ denotes the Hadamard product, and $\mathcal{W}_k \in \mathbb{R}^{(L+1)\times D\times D}$ is a learned weight tensor. The parameters of each SCM are given by $\Theta_k = \{\mathcal{W}_k\}$. Since we do not model the history-dependence of the noise, we set $g_k(\mathrm{Pa}_{\mathcal{G}_k}(<t), \epsilon_t) = \epsilon_t$, i.e., the identity function.

MCD-Nonlinear. We implement each of the K causal models based on Rhino (Gong et al., 2022) as it can handle instantaneous effects and history-dependent noise. We parameterize the structural equations f_k in (2) with embeddings $\mathcal{E}_k$, which are used in conjunction with neural networks, denoted by Ξ_f and ℓ_f:

$$f_k^d\left(\mathrm{Pa}_{\mathcal{G}_k(\leq t)}\right) = \Xi_f\left(\left[\sum_{\tau=0}^{L}\sum_{j=1}^{D}\left(\mathcal{G}_k\right)_\tau^{j,d} \times \ell_f\left(\left[X_{t-\tau}^{j,(n)}, \left(\mathcal{E}_k\right)_\tau^j\right]\right), \left(\mathcal{E}_k\right)_0^d\right]\right). \quad (6)$$

$\mathcal{E}_k \in \mathbb{R}^{(L+1)\times D\times e}$ are trainable embeddings (with embedding dimension e) corresponding to model $\mathcal{M}_k$, and Ξ_f and ℓ_f are multi-layer perceptron networks that are shared across all nodes and causal models $\mathcal{M}_{1:K}$. The noise model $g_k(\mathrm{Pa}_{\mathcal{G}_k}(<t), \epsilon_t)$ is described using conditional spline flow. The network that predicts parameters for the conditional spline flow model uses a similar architecture, utilizing embeddings $\mathcal{F}_k$ with neural networks Ξ_ϵ and ℓ_ϵ. Thus, the SCM parameters are $\Theta_k = \{\mathcal{E}_k, \mathcal{F}_k, \Xi_f, \ell_f, \Xi_\epsilon, \ell_\epsilon\}$.

4. Theoretical analysis

In this section, we examine (1) conditions under which the mixture model is identifiable (2) the relationship between the derived ELBO objective and the true data likelihood.

Structural identifiability. We examine when the mixtures of SCM models are identifiable. Structural identifiability dictates that two distinct mixtures of SCMs cannot result in the same observational distribution. Identifiability is an important statistical property to ensure that the causal discovery problem is meaningful (Peters et al., 2011).

We establish a necessary and sufficient condition for the identifiability of mixtures of linear structural vector autoregressive models (SVARs) with Gaussian noise.

Theorem 2 (Identifiability of linear SVARs with equal–variance additive Gaussian noise)**.** *Let $\mathcal{F}$ be a family of distributions of K structural vector autoregressive (SVAR) models of lag $L \geq 1$ with zero-mean Gaussian noise of equal variance, i.e.*

$$\mathcal{F} = \Big\{\mathcal{L}_{\mathcal{M}^{(k)}} : \mathcal{M}^{(k)} \text{ is specified by the equations } \mathbf{X}_t = \mathbf{W}^{(k)}\mathbf{X}_t + \sum_{\tau=1}^{L}\mathbf{A}_\tau^{(k)}\mathbf{X}_{t-\tau} + \varepsilon^{(k)}, \; \varepsilon^{(k)} \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right), 1 \leq k \leq K\Big\}$$

and let $\mathcal{H}_K$ be the family of all $K-$finite mixtures of elements from $\mathcal{F}$. Then the family $\mathcal{H}_K$ is identifiable if and only if the following condition is met: The ordered pairs

$$\left(\left[\mathbf{B}^{(k)}\right]^{-1}\mathbf{A}_1^{(k)}, ..., \left[\mathbf{B}^{(k)}\right]^{-1}\mathbf{A}_L^{(k)}, \left[\mathbf{B}^{(k)}\right]\left[\mathbf{B}^{(k)}\right]^T\right), \quad (7)$$

are distinct over all k, where $\mathbf{B}^{(k)} = \mathbf{I} - \mathbf{W}^{(k)}$.

To illustrate this condition, we consider the 2D SCM case when $L = 0$, i.e. there are no lagged effects. The SCM is specified by $W^{(k)} = \begin{bmatrix} 0 & w_1^{(k)} \\ w_2^{(k)} & 0 \end{bmatrix}$, and the matrix $B^{(k)}$ takes the form $B^{(k)} = \begin{bmatrix} 1 & -w_1^{(k)} \\ -w_2^{(k)} & 1 \end{bmatrix}$, where $w_1^{(k)}w_2^{(k)} = 0$ due to acyclicity. Then $\left[B^{(k)}\right]\left[B^{(k)}\right]^T = \begin{bmatrix} 1+\left(w_1^{(k)}\right)^2 & -(w_1^{(k)}+w_2^{(k)}) \\ -(w_1^{(k)}+w_2^{(k)}) & 1+\left(w_2^{(k)}\right)^2 \end{bmatrix}$, and the condition is violated iff $W^{(i)} = W^{(j)}$, i.e., they have the same SCM equations. The mixture of linear Gaussian SCMs is identifiable when the structural equation matrices are distinct.

We now examine the identifiability of mixtures of general SCMs. We derive an intuitive sufficient condition for mixture model identifiability in terms of the existence of K representative points from the sample space $\mathbb{X}$. These points

exhibit a key characteristic: their association with a particular causal model is unequivocal, as determined by their marginal likelihood functions of the mixture components.

Theorem 3 (Identifiability of finite mixture of causal models)**.** *Let $\mathcal{F}$ be a family of K identifiable causal models, $\mathcal{F} = \left\{\mathcal{L}_{\mathcal{M}}^{(k)} : \mathcal{M} \text{ is an identifiable causal model }, 1 \leq k \leq K\right\}$ and let $\mathcal{H}_K$ be the family of all $K-$finite mixtures of elements from $\mathcal{F}$, i.e.*

$$\mathcal{H}_K = \Bigg\{ h : h = \sum_{k=1}^{K} \pi_k \mathcal{L}_{\mathcal{M}_k}, \mathcal{L}_{\mathcal{M}_k} \in \mathcal{F}, \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1 \Bigg\}$$

where $\mathcal{L}_{\mathcal{M}_k}(x) = \sum_{\mathcal{M}} p(x \mid \mathcal{M}) p(\mathcal{M}_k = \mathcal{M})$ denotes the likelihood of x evaluated with causal model $\mathcal{M}_k$. Further, assume that the following condition is met:

For every $k = 1, \ldots, K$, $\exists a_k \in \mathbb{X}$ such that

$$\frac{\mathcal{L}_{\mathcal{M}_k}(a_k)}{\sum_{j=1}^{K} \mathcal{L}_{\mathcal{M}_j}(a_k)} > \frac{1}{2}. \qquad (*)$$

Then the family $\mathcal{H}_K$ is identifiable.

Appendix A.3 contains the relevant definitions and proofs. To draw a parallel with clustering, this implies that each cluster has at least one point whose membership can be established with a high level of certainty to that specific cluster. Directly verifying the condition (*) is generally difficult because we rarely know the exact form of the likelihood function. However, this condition can be verified approximately using the estimated likelihood functions for each mixture component, as with our approach, MCD. The validity of this verification critically depends on how closely the estimated likelihood function approximates the true likelihood function.

Furthermore, as a direct consequence of the structural identifiability of the Rhino model (Gong et al., 2022), a mixture of Rhino models is also identifiable, provided that the assumptions in Section A.2 and condition $(*)$ are satisfied.

Relationship between ELBO and log-likelihood. We verify the soundness of our derived ELBO objective in Eq. (4). By maximizing the ELBO, we can simultaneously learn the K underlying causal graphs, their associated functional equations, and the membership of each sample to its respective causal model. We show that (Appendix A.4):

$$\begin{aligned} &\log p_\theta(X) = \text{ELBO}(\theta, \phi, \psi) \\ &+ \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \Big[\\ &\text{KL}\left(r_\psi\left(Z^{(n)} | X^{(n)}\right) \| p(Z^{(n)} | X^{(n)}, \mathcal{M}_{1:K}) \right) \Big] \\ &+ \text{KL}\left(q_\phi\left(\mathcal{M}_{1:K}\right) \| p\left(\mathcal{M}_{1:K} | X\right) \right). \end{aligned}$$

Maximizing $\text{ELBO}(\theta, \phi, \psi)$ with respect to (θ, ϕ, ψ) is equivalent to jointly (1) maximizing the log-likelihood $\log p_\theta(X)$ (2) minimizing the KL divergence between the variational distribution $q_\phi\left(\mathcal{M}_{1:K}\right)$ and the true posterior $p\left(\mathcal{M}_{1:K} \mid X\right)$, and (3) minimizing the expectation, under the variational distribution $q_\phi(\mathcal{M}_{1:K})$, of the KL divergence between the variational posterior $r_\psi\left(Z^{(n)} \mid X^{(n)}\right)$ for each sample $X^{(n)}$ and the true posterior for mixture component selection $p\left(Z^{(n)} \mid X^{(n)}, \mathcal{M}_{1:K}\right)$.

5. Experiments

5.1. Experimental setup

We train the model on 80% of the data and validate on the remaining 20%. We pick the model with the lowest validation likelihood and evaluate the corresponding causal graphs. Details about model validation are in Appendix D.2.

Baselines. We benchmark against several state-of-the-art temporal causal discovery methods, including Rhino (Gong et al., 2022), PCMCI$^+$ (Runge, 2020), DYNOTEARS (Pamfil et al., 2020), and VARLiNGaM (Hyvärinen et al., 2010). PCMCI$^+$ and DYNOTEARS can be used with two options - one where the algorithm predicts one causal graph per sample and one where the algorithm predicts one graph to explain the whole dataset. We denote these options with suffixes -s and -o, respectively. Since these baseline methods cannot discover mixtures of causal graphs, we also report results by grouping samples by their true causal graph. We then predict one causal graph per group. This option is reported for PCMCI$^+$, DYNOTEARS, and Rhino and is denoted by the suffix -g in the results. Appendix D.4 details the steps for post-processing PCMCI$^+$'s output.

In practice, the number of mixture components, which we treat as a hyperparameter, is often unknown. We use K^* to denote the true number of SCMs, and K to represent the input to MCD. We report the clustering accuracy for MCD in addition to traditional causal discovery metrics like orientation F1 score and AUROC (Area Under the Receiver Operator Curve). We define clustering accuracy as:

$$\text{Cluster Acc.}\left(\tilde{Z}, Z\right) = \max_{\pi \in S_K} \frac{1}{N} \sum_{n=1}^{N} 1\left(\pi(\tilde{Z}_n) = Z_n\right),$$

where $\tilde{Z}$ are the assigned mixture labels and Z are the true

labels. We refer the reader to Appendix C.1 for details about the calculation of clustering accuracy.

5.2. Datasets

Synthetic datasets. We generate a pool of K^* random graphs (specifically, Erdős-Rényi graphs) and treat them as ground-truth causal graphs. To generate a sample $X^{(n)}$, we first randomly sample a graph $\mathcal{G}_k$ from this pool and use it to model relationships between the variables. We experiment with $D = 5, 10, 20$ nodes. For each value of D, we generate datasets with $K^* = 1, 5, 10, 20$ graphs having $N = 1000$ samples each. The time series length T is 100, and the time lag L is set to 2 for all the methods, which is the lag value used to simulate the data. The number of causal graphs K is set to $2K^*$. We experiment with two sets of synthetic datasets: (1) Linear datasets, in which linear causal relationships are modeled with Gaussian noise; (2) Nonlinear datasets, in which the functional relationships are modeled as randomly generated multi-layer perceptions with history-dependent noise. We refer the reader to Appendix D.1 for more details about the setup.

Netsim Brain Connectivity. The Netsim benchmark dataset (Smith et al., 2011) consists of simulated blood oxygenation level-dependent (BOLD) imaging data. Each variable represents a region of the brain, with the goal being to infer the interactions between the different regions. The dataset has 28 different simulations, which differ in the number of variables and time length over which the measurements are recorded. In our experiments, we consider the samples from simulation 3 comprising $N = 50$ time series, each with $D = 15$ nodes and $T = 200$ timepoints. These samples share the same ground-truth causal graph. We introduce heterogeneity by considering a pool of $K^* = 3$ random permutations and applying a randomly chosen one to the nodes of each sample and its corresponding ground truth causal graph. This setup is denoted as **Netsim-mixture**. We use a uniform prior for $p(Z)$ and set $L = 2$ and $K = 5$.

DREAM3 Gene Network. The DREAM3 dataset (Prill et al., 2010) is a real-world biology dataset consisting of measurements of gene expression levels obtained from yeast and E.coli cells. There are 5 distinct ground-truth networks, comprising 2 for E.coli and 3 for Yeast, each with $D = 100$ nodes. Each time series consists of $T = 21$ timesteps, with 46 trajectories recorded per graph. Thus, a total of $N = 230$ samples are combined across all the networks. We mix samples from all 5 networks to simulate the scenario in which the identity of the cell from which the data is obtained is unknown. This is a challenging dataset due to the high dimensionality of the data and the small number of available samples. We set the time lag $L = 2$ and $K = 10$. We discuss how we post-process the model outputs on the Netsim and DREAM3 datasets in Appendix D.5.

	Netsim-mixture		DREAM3	
Method	AUROC(↑)	**F1**(↑)	AUROC(↑)	**F1** (↑)
PCMCI$^+$-s	0.82	0.67	0.50	0.01
PCMCI$^+$-o	0.71	0.49	0.51	0.04
PCMCI$^+$-g	0.72	0.52	0.51	0.05
VARLiNGAM	0.78	0.60	NA	NA
DYNOTEARS-s	0.85	0.28	0.50	0.03
DYNOTEARS-o	0.83	0.45	0.50	0.03
DYNOTEARS-g	0.85	0.46	0.50	0.03
Rhino	0.84 ± 0.01	0.62 ± 0.01	0.57 ± 0.01	0.08 ± 0.01
MCD-Nonlinear (this paper)	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.69 \pm 0.08}$	$\mathbf{0.58 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$
MCD-Linear (this paper)	0.73 ± 0.02	0.62 ± 0.02	0.51 ± 0.01	0.00 ± 0.00

Table 1. Results on Netsim-mixture and DREAM3. -s indicates that the baseline predicts one graph per sample. -o indicates that the baseline predicts one graph for the whole dataset. -g signifies that the baseline is run on samples grouped according to the ground truth causal graph. VARLiNGAM does not run on the DREAM3 dataset. MCD-Nonlinear achieves a clustering accuracy of $86.8 \pm 26.3\%$ on Netsim-mixture and $95.6 \pm 4.8\%$ on DREAM3.

S&P 100. We also run MCD on daily stock returns of companies from the S&P 100 index. We use the `yahoofinancials` package to retrieve the daily closing prices of $D = 100$ stocks from January 1, 2016 to July 1, 2023. Similar to the setup in Pamfil et al. (2020), we use log-returns, i.e., differences of the logarithm of the closing prices of successive days. In addition, we normalize the log-returns to have zero mean and unit variance. We chunk the data into segments of length $T = 31$ each, resulting in $N = 60$ samples. We train our model on the first 48 samples and validate with the last 12 samples. Following Pamfil et al. (2020), we set $L = 1$. We set $K = 5$ and threshold the edge probabilities at 0.4. We qualitatively analyze the results on this dataset since it lacks ground truth causal graphs.

5.3. Results on synthetic and real-world datasets

Synthetic datasets results. Results are presented in Figure 4. On the nonlinear dataset, MCD-Nonlinear handily outperforms all the baselines except Rhino-g. Notably, it performs better than PCMCI$^+$-g, even though PCMCI$^+$ has additional information (i.e., ground truth membership information) that MCD-Nonlinear does not. MCD-Nonlinear achieves comparable, and sometimes better, performance than Rhino-g, especially on the $D = 10$ and $D = 20$ datasets. The baseline variants that predict one graph per sample perform poorly as expected since one sample does not provide adequate information to infer all the causal relationships. DYNOTEARS and VARLiNGAM, which assume that the causal relationships are linear, perform poorly on these datasets. We also omit MCD-Linear for this reason.

On the linear dataset, MCD-Linear achieves a similar or better F1 score than the grouped baseline methods when the number of graphs K^* is 5, 10, 20. It achieves a comparable level of performance to the baselines for $K^* = 1$ despite the misspecification of the number of graphs. MCD-Linear expectedly outperforms MCD-Nonlinear across all settings

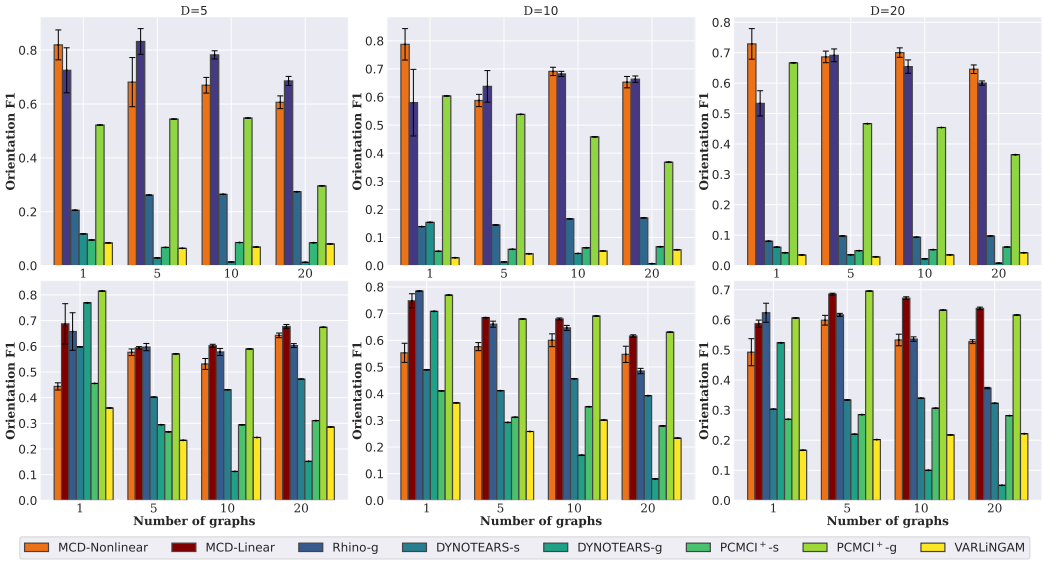

Figure 4. Results on the **nonlinear** (top) and **linear** (bottom) synthetic datasets for dimension $D = 5, 10, 20$. We report the orientation F1 scores. (-s) indicates that the baseline predicts one graph per sample. (-g) signifies that the baseline was executed on samples grouped according to the ground truth causal graph. These methods use additional information that MCD does not. Average of 5 runs reported.

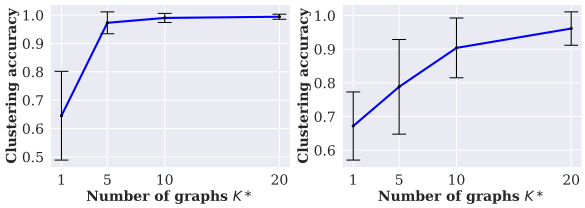

Figure 5. Clustering accuracy of (left) MCD-Nonlinear on the nonlinear synthetic datasets (right) MCD-Linear on the linear synthetic datasets, as a function of the true number of causal graphs K^*. The accuracy is averaged across 5 runs and data dimensionality $D = 5, 10, 20$. Hyperparameter K is set to $2K^*$ for all settings.

since its model matches the parametric form of the data generation process. Nevertheless, MCD-Nonlinear performs better than all the baselines that do not use ground truth membership information. Although VARLiNGAM is a linear model, it performs poorly since it cannot handle linear SCMs with Gaussian noise.

We report the clustering accuracy for MCD in Figure 5 for different values of K^*, averaged over the data dimensionalities $D = 5, 10, 20$. MCD-Nonlinear achieves near-perfect clustering for scenarios with multiple underlying graphs. MCD-Linear also achieves strong clustering performance, although it shows high variability across D. The low clustering accuracy and relatively low F1 and AUROC scores for $K^* = 1$ are explained by the observation that MCD learns two similar mixture components to explain the single underlying mode in the distribution.

Netsim-mixture results. The results on the Netsim dataset are presented in Table 1. MCD-Nonlinear outperforms all baselines as measured by AUROC and F1 scores. This setting illustrates the benefits of modeling heterogeneity, even when it comes from a simple permutation of nodes. In this setting, MCD-Nonlinear achieves a clustering accuracy of $86.8 \pm 26.3\%$, highlighting its ability to accurately group samples when the underlying causal models are sufficiently diverse. MCD-Linear learns a single mode for the dataset and achieves similar results to Rhino.

DREAM3 results. The results on the DREAM3 dataset are presented in Table 1. All methods fare poorly at inferring the causal relationships. However, out of all the considered baselines, MCD-Nonlinear achieves relatively better performance in terms of AUROC and F1 score. It is especially encouraging that MCD-Nonlinear can accurately cluster samples by their causal models, with a remarkable clustering accuracy of $95.6 \pm 4.8\%$. MCD-Linear infers a single mode from the dataset.

S&P100 results. MCD-Nonlinear infers two distinct causal graphs for the S&P100 dataset. We aggregate the adjacency matrices across time as described in Appendix D.5. Figure 6 shows subgraphs of the two discovered causal graphs for stocks in the energy, financials, industrials, and real estate sectors. The model identifies that companies from the same sector interact more than those across sectors, which is evident from the block-diagonal structure of the inferred graphs. Further, the two inferred causal graphs have important differences; e.g., Graph 1 shows more interactions between the financial and industrial sectors than Graph 2, and the direction of the causal influences for XOM in the

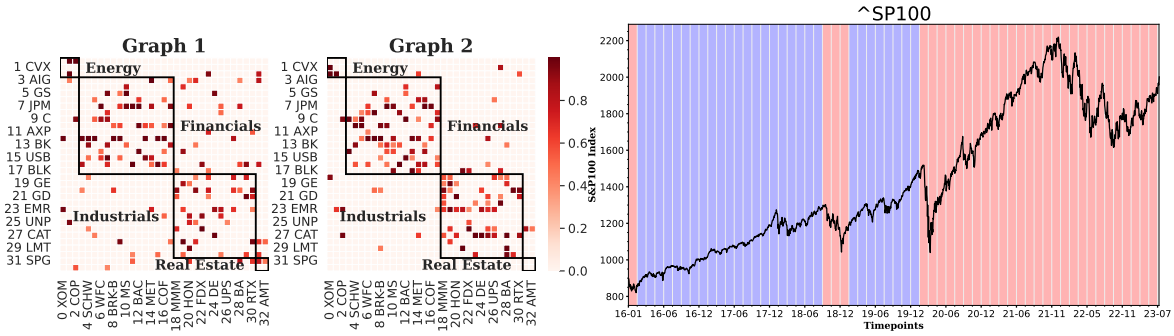

Figure 6. (Left) Sub-graphs of the two inferred graphs from MCD-Nonlinear (Right) S&P 100 index overlaid with the mixture membership from MCD-Nonlinear. Red indicates that Graph 1 from Figure 11 is used, while blue indicates that Graph 2 is active.

energy sector is reversed between the two graphs.

We also overlay the inferred mixture membership for each sample onto the S&P100 index in Figure 6 (right). We note that the model automatically identifies that several consecutive time windows are governed by the same causal graph. In addition, the changes in the governing causal graph successfully capture several important events. For example, the stock market crashes in December 2018 and March 2020 (due to COVID-19) are captured by the red (first) causal graph. The 'blue' periods, in which Graph 2 is active, exhibit relatively less pronounced trends. Additionally, Graph 2 is much sparser than Graph 1. We show the full causal graphs and interesting patterns that MCD-Nonlinear captures in selected stocks in Appendix B.2.

5.4. Ablation Studies

Robustness of MCD to the misspecification of number of components. We examine the performance of MCD when the number of mixture components K is misspecified, and does not equal the true number of underlying components K^*. Figure 7 shows the performance of MCD-Nonlinear as a function of K on the nonlinear synthetic dataset with dimensionality $D = 10$ and ground truth number of graphs $K^* = 10$. We note that when the number of models is underspecified, our model performs poorly as expected since it cannot fully explain all the modes in the data. Surprisingly, the performance increases with increasing K. The clustering accuracy and performance metrics show high standard deviation when K is set to the true number of mixture components $K^* = 10$. While some random seeds achieve high clustering accuracy, others tend to saturate at a suboptimal grouping when $K = K^*$. On the other hand, when $K > K^*$, the additional SCMs are used as 'buffers,' and the correct grouping is learned during the later epochs as the SCMs are inferred more accurately. This phenomenon is further explored in Appendix B.6

We perform additional ablation studies on synthetic datasets to investigate the behavior of MCD as K^* increases, compare causal discovery performance using ground-truth membership assignments, and examine the impact of the similarity of the causal graphs. Details of these studies can be found in Appendix B.4.

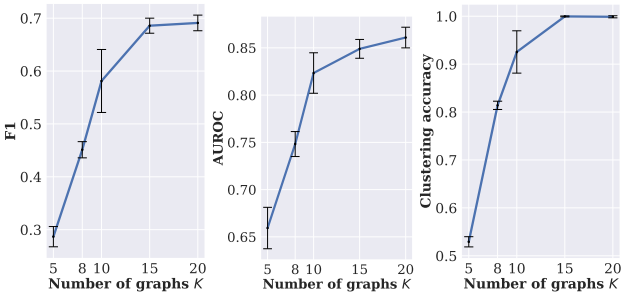

Figure 7. Performance of MCD-Nonlinear as a function of hyperparameter input K on synthetic data with $D = 10, K^* = 10$. Surprisingly, MCD-Nonlinear performs better when the number of graphs is overspecified.

6. Conclusion and Discussion

In this work, we examine the problem of discovering mixtures of structural causal models from time series data. This problem has far-reaching applications in climate, finance, and healthcare, among other fields, since multimodal and heterogeneous data is ubiquitous in practice. We propose MCD, an end-to-end variational inference method, to learn both the underlying SCMs and the mixture component membership of each sample. We demonstrate the empirical efficacy of our method on both synthetic and real-world heterogeneous datasets. We conduct ablation studies on synthetic datasets to investigate MCD's behavior with varying numbers of causal graphs, its robustness to misspecification of the number of graphs, and the impact of similarity among the causal graphs associated with each mixture component. In addition, we discuss the structural identifiability of mixtures of causal models. Future work could tackle data with latent confounders and non-stationarity in time.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

This work was supported in part by the U.S. Army Research Office under Army-ECASE award W911NF-07-R-0003-03, the U.S. Department Of Energy, Office of Science, IARPA HAYSTAC Program, NSF Grants SCALE MoDL-2134209, CCF-2112665 (TILOS), #2205093, #2146343, #2134274, CDC-RFA-FT-23-0069 and DARPA AIE FoundSci.

References

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Gao, S., Addanki, R., Yu, T., Rossi, R. A., and Kocaoglu, M. Causal discovery in semi-stationary time series. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., Sharma, A., Lamb, A., Kukla, M., Pawlowski, N., et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Gong, W., Jennings, J., Zhang, C., and Pawlowski, N. Rhino: Deep causal temporal relationship learning with history-dependent noise. *The Eleventh International Conference on Learning Representations*, 2022.

Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Huang, B., Zhang, K., Xie, P., Gong, M., Xing, E. P., and Glymour, C. Specific and shared causal relation modeling and mechanism-based clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

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

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Khanna, S. and Tan, V. Y. F. Economy statistical recurrent units for inferring nonlinear granger causality. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SyxV9ANFDH.

Löwe, S., Madras, D., Zemel, R., and Welling, M. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pp. 509–525. PMLR, 2022.

Malinsky, D. and Spirtes, P. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*, pp. 23–47. PMLR, 2018.

Markham, A., Das, R., and Grosse-Wentrup, M. A distance covariance-based kernel for nonlinear causal clustering in heterogeneous populations. In *Conference on Causal Learning and Reasoning*, pp. 542–558. PMLR, 2022.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1595–1605. PMLR, 2020.

Pearl, J. *Causality*. Cambridge university press, 2009.

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

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 589–598, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

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

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, 5(2):e9202, 2010.

Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Hoi (Joseph) Min, K., Wang, L., Grody, E. I., Shurtleff, M. J., Yuan, R., Xu, S., Ma, Y., Replogle, J. M., Lander, E. S., Darmanis, S., Bahar, I., Sankaran, V. G., Xing, J., and Weissman, J. S. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711.e45, 2022.

Runge, J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7), 2018.

Runge, J. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1388–1397. PMLR, 2020.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

Saeed, B., Panigrahi, S., and Uhler, C. Causal structure discovery from distributions arising from mixtures of dags. In *International Conference on Machine Learning*, pp. 8336–8345. PMLR, 2020.

Saggioro, E., de Wiljes, J., Kretschmer, M., and Runge, J. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11), 2020.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Spirtes, P. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pp. 278–285. PMLR, 2001.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Strobl, E. V. Causal discovery with a mixture of dags. *Machine Learning*, pp. 1–25, 2022.

Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. B. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Thiesson, B., Meek, C., Chickering, D. M., and Heckerman, D. Learning mixtures of dag models. In *Proc. of the 14th Conference on Uncertainty in Attificial Intelligence*, pp. 504–513, 1998.

Yakowitz, S. J. and Spragins, J. D. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RDlLMjLJXdq.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Zhou, F., He, K., and Ni, Y. Causal discovery with heterogeneous observational data. In *Uncertainty in Artificial Intelligence*, pp. 2383–2393. PMLR, 2022.

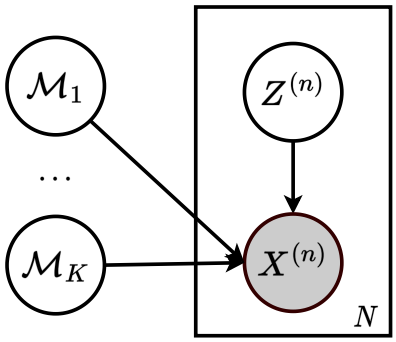

Figure 8. The assumed data generation model. First, the mixture index $Z^{(n)}$ is drawn from a K-way categorical distribution ($Z^{(n)} \sim \text{Cat}(K), Z^{(n)} \in \{1, \dots, K\}$), and a causal model is drawn from the corresponding mixture component distribution $\mathcal{M} \sim p\left(\mathcal{M}_{Z^{(n)}}\right)$. A sample $X^{(n)}$ is then drawn in according to the chosen causal model $\mathcal{M}$.

A. Theory

A.1. ELBO derivation

Proposition 1. Under the data generation process described in Figure 2, the data likelihood admits the following evidence lower bound (ELBO):

$$\begin{aligned}
&\log p_\theta \left(X_{1:T}^{(1:N)}\right) \\
&\geq \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\mathbb{E}_{r_\psi\left(Z^{(n)} | X_{1:T}^{(n)}\right)} \left[\log p_\theta \left(X_{1:T}^{(n)} \mid \mathcal{M}_{Z^{(n)}}\right) + \log p\left(Z^{(n)}\right) \right] + H\left(r_\psi\left(Z^{(n)} \mid X_{1:T}^{(n)}\right)\right) \right] \\
&+ \sum_{i=1}^{K} \mathbb{E}_{q_\phi(\mathcal{M}_i)} \left[\log p(\mathcal{M}_i)\right] + H\left(q_\phi(\mathcal{M}_i)\right)
\end{aligned}$$

Proof. Denote the causal models as $\mathcal{M}_{1:K} = (\mathcal{M}_1, \dots, \mathcal{M}_K)$ and the sample $X = \left\{X^{(n)}\right\}_{n=1}^{N}$. Then, we can write the log-likelihood under the assumed model as follows:

$$\begin{aligned}
\log p_\theta(X) &= \log \left[\sum_{\mathcal{M}_{1:K}} p_\theta\left(X \mid \mathcal{M}_{1:K}\right) p(\mathcal{M}_{1:K}) \times \frac{q_\phi(\mathcal{M}_{1:K})}{q_\phi(\mathcal{M}_{1:K})} \right] \\
&= \log \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\frac{p_\theta\left(X \mid \mathcal{M}_{1:K}\right) p(\mathcal{M}_{1:K})}{q_\phi(\mathcal{M}_{1:K})} \right] \\
&\geq \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\log \frac{p_\theta\left(X \mid \mathcal{M}_{1:K}\right) p(\mathcal{M}_{1:K})}{q_\phi(\mathcal{M}_{1:K})} \right] \qquad \text{(using Jensen's inequality)} \\
&= \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\log p_\theta\left(X \mid \mathcal{M}_{1:K}\right) + \log p(\mathcal{M}_{1:K}) - \log q_\phi(\mathcal{M}_{1:K})\right]
\end{aligned}$$

Since the sample points are conditionally independent given the causal models, we can write:

$$\begin{aligned}
\log p_\theta(X) \geq \sum_{n=1}^{N} &\mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\log p_\theta \left(X^{(n)} \mid \mathcal{M}_{1:K}\right) \right] \\
&+ \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\log p(\mathcal{M}_{1:K}) - \log q_\phi(\mathcal{M}_{1:K})\right]
\end{aligned}$$

Further, note that:

$$\begin{aligned}
\log p_\theta(X^{(n)} \mid \mathcal{M}_{1:K}) &= \log \left[\sum_{Z^{(n)}} p_\theta(X^{(n)} \mid Z^{(n)}, \mathcal{M}_{1:K}) p(Z^{(n)} \mid \mathcal{M}_{1:K}) \right] \\
&= \log \left[\sum_{Z^{(n)}} p_\theta(X^{(n)} \mid Z^{(n)}, \mathcal{M}_{1:K}) p(Z^{(n)}) \times \frac{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)}{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)} \right] \\
&= \log \mathbb{E}_{r_\psi(Z^{(n)} \mid X^{(n)})} \left[\frac{p_\theta(X^{(n)} \mid Z^{(n)}, \mathcal{M}_{1:K}) p(Z^{(n)})}{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)} \right] \\
&\geq \mathbb{E}_{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)} \left[\log \frac{p_\theta(X^{(n)} \mid Z^{(n)}, \mathcal{M}_{1:K}) p(Z^{(n)})}{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)} \right] \qquad \text{(using Jensen's inequality)} \\
&= \mathbb{E}_{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)} \left[\log p_\theta(X^{(n)} \mid Z^{(n)}, \mathcal{M}_{1:K}) + \log p(Z^{(n)}) - \log r_\psi\left(Z^{(n)} \mid X^{(n)}\right) \right].
\end{aligned}$$

We use the fact that $p_\theta(X^{(n)} \mid Z^{(n)}, \mathcal{M}_{1:K}) = p_\theta\left(X^{(n)} \mid \mathcal{M}_{Z^{(n)}}\right)$. Putting it all together, and using the independence of the causal models, we obtain:

$$\begin{aligned}
\log p_\theta(X) \geq & \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\mathbb{E}_{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)} \left[\log p_\theta(X^{(n)} \mid \mathcal{M}_{Z^{(n)}}) + \log p(Z^{(n)}) - \log r_\psi\left(Z^{(n)} \mid X^{(n)}\right) \right] \right] \\
& + \sum_{i=1}^{K} \mathbb{E}_{q_\phi(\mathcal{M}_i)} \left[\log p(\mathcal{M}_i) - \log q_\phi(\mathcal{M}_i) \right] \\
& \equiv \text{ELBO}(\theta, \phi, \psi)
\end{aligned}$$

□

A.2. Theoretical assumptions

In this section, we list out the theoretical assumptions used in Rhino (Gong et al., 2022). Our model also operates under similar assumptions, since we implement the component SCMs as Rhino models.

Assumption 1 (Causal Stationarity). (Runge, 2018) The time series X with a graph G is called causally stationary over a time index set $\mathcal{T}$ if and only if for all links $X^i_{t-\tau} \to X^j_t$ in the graph

$$X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid X_t \backslash \left\{X^i_{t-\tau}\right\} \qquad \text{holds for all } t \in \mathcal{T}.$$

Informally, this assumption states that the causal graph does not change over time, i.e., the resulting time series is stationary.

Assumption 2 (Causal Markov Property). (Peters et al., 2017) Given a DAG G and a probability distribution p, p is said to satisfy the causal Markov property, if it factorizes according to G, i.e. $p(x) = \prod_{i=1}^{D} p\left(x_i \mid \text{Pa}^i_G(x_i)\right)$. In other words, each variable is independent of its non-descendent given its parents.

Assumption 3 (Causal Minimality). Given a DAG G and a probability distribution p, p is said to satisfy the causal minimality with respect to G, if p is Markovian with respect to G but not to any proper subgraph of G.

Assumption 4 (Causal Sufficiency). A set of observed variables V is said to be causally sufficient for a process X_t if, in the process, every common cause of two or more variables in V is also in V, or is constant for all units in the population. In other words, causal sufficiency implies the absence of latent confounders in the data.

Assumption 5 (Well-defined Density). The likelihood of each mixture component (i.e. the likelihood function of each Rhino model) is absolutely continuous with respect to a Lebesgue or counting measure and $|\log p\left(X_{0:T}; G\right)| < \infty$ for all possible G.

A.3. Identifiability of the mixture of causal models

Definition 1 (Identifiability)**.** Let $P = \{p_\theta : \theta \in \mathcal{T}\}$ be a family of distributions, each member of which is parameterized by the parameter θ from a parameter space $\mathcal{T}$. Then P is said to be identifiable if

$$p_{\theta_1} = p_{\theta_2} \implies \theta_1 = \theta_2 \quad \forall \theta_1, \theta_2 \in \mathcal{T}.$$

Definition 2 (Identifiability of finite mixtures)**.** Let $\mathcal{F}$ be a family of distributions. The family of $K-$mixture distributions on $\mathcal{F}$, defined as $\mathcal{H}_K = \left\{ h : h = \sum_{k=1}^{K} \pi_k f_k, f_k \in \mathcal{F}, \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1 \right\}$, is said to be identifiable if

$$\sum_{k=1}^{K} \pi_k f_k = \sum_{j=1}^{K} \pi'_j f'_j \implies \forall k \; \exists j \text{ such that } \pi_k = \pi'_j \text{ and } f_k = f'_j.$$

Here, we quote a result from Yakowitz & Spragins (1968) that established a necessary and sufficient condition for the identifiability of finite mixtures of multivariate distributions.

Theorem A (Identifiability of finite mixtures of distributions (Yakowitz & Spragins, 1968))**.** Let $\mathcal{F} = \{F(x;\alpha), \alpha \in \mathbb{R}^m, x \in \mathbb{R}^n\}$ be a finite mixture of distributions. Then $\mathcal{F}$ is identifiable if and only if $\mathcal{F}$ is a linearly independent set over the field of real numbers.

In other words, this theorem states that a mixture of distributions is identifiable if and only if none of the individual mixture components can be expressed as a mixture of distributions from the same family. In general, it can be difficult to comment on the identifiability of a mixture of arbitrary random distributions. However, it is known that the mixture of multivariate Gaussian distributions is identifiable. We use this result to prove the identifiability of a mixture of linear SCMs with Gaussian noise.

Proposition A (Identifiability of mixture of multivariate Gaussian distributions (Yakowitz & Spragins, 1968))**.** The family of n-dimensional Gaussian distributions generates identifiable finite mixtures.

Theorem B (Identifiability of linear SCMs with equal-variance additive Gaussian noise)**.** Let $\mathcal{F}$ be a family of distributions of K linear causal models with Gaussian noise of equal variance, i.e.

$$\mathcal{F} = \left\{ \mathcal{L}_{\mathcal{M}^{(k)}} : \mathcal{M}^{(k)} \text{ is specified by the equations } \mathbf{X} = \mathbf{W}^{(k)}\mathbf{X} + \varepsilon^{(k)}, \varepsilon^{(k)} \sim \mathcal{N}\left(\mu^{(k)}, \sigma^2\mathbf{I}\right), 1 \le k \le K \right\}$$

and let $\mathcal{H}_K$ be the family of all $K-$finite mixtures of elements from $\mathcal{F}$, i.e.

$$\mathcal{H}_K = \left\{ h : h = \sum_{k=1}^{K} \pi_k \mathcal{L}_{\mathcal{M}^{(k)}}, \mathcal{L}_{\mathcal{M}^{(k)}} \in \mathcal{F}, \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1 \right\}$$

where $\mathcal{L}_{\mathcal{M}^{(k)}}(x) = p\left(x \mid \mathcal{M}^{(k)}\right)$ denotes the likelihood of x evaluated with causal model $\mathcal{M}^{(k)}$.

Then the family $\mathcal{H}_K$ is identifiable if and only if the following condition is met:

$$\text{The ordered pairs } \left(\left[\mathbf{B}^{(k)}\right]^{-1} \mu^{(k)}, \left[\mathbf{B}^{(k)}\right] \left[\mathbf{B}^{(k)}\right]^T \right) \text{ are distinct over all } k,\ 1 \le k \le K, \tag{8}$$

where $\mathbf{B}^{(k)} = \mathbf{I} - \mathbf{W}^{(k)}$.

Proof. Note that the equations for a linear SCM can equivalently be written as:

$$\mathbf{X}^{(k)} = \left[\mathbf{B}^{(k)}\right]^{-1} \varepsilon^{(k)} \sim \mathcal{N}\left(\left[\mathbf{B}^{(k)}\right]^{-1} \mu^{(k)}, \sigma^2 \left[\mathbf{B}^{(k)}\right]^{-1} \left[\mathbf{B}^{(k)}\right]^{-T} \right)$$

where $\mathbf{B}^{(k)} = \left(\mathbf{I} - \mathbf{W}^{(k)}\right)$.

A linear SCM with equal variance Gaussian additive noise is known to be identifiable (Peters & Bühlmann, 2014). Thus, from Proposition A, we have that the finite mixture is identifiable, as long as the parameters of the resultant Gaussian distributions are distinct, as required by the condition in (8).

Conversely, if condition (8) does not hold, then $\exists k \neq j$ such that

$$p(x \mid \mathcal{M}^{(k)}) = p(x \mid \mathcal{M}^{(j)}),$$

i.e. there are two mixture components with identical distributions. This family cannot be identifiable, since any mixture of the form $h = \alpha p\left(x \mid \mathcal{M}^{(k)}\right) + (1-\alpha)p\left(x \mid \mathcal{M}^{(j)}\right) = p(x \mid \mathcal{M}^{(k)})$ for any $\alpha \in [0,1]$ □

Theorem 2 (Identifiability of linear SVARs with equal-variance additive Gaussian noise)**.** Let $\mathcal{F}$ be a family of distributions of K structural vector autoregressive (SVAR) models of lag $L \geq 1$ with zero-mean Gaussian noise of equal variance, i.e.

$$\mathcal{F} = \Bigg\{ \mathcal{L}_{\mathcal{M}^{(k)}} : \mathcal{M}^{(k)} \text{ is specified by the equations } \mathbf{X}_t = \mathbf{W}^{(k)}\mathbf{X}_t + \sum_{\tau=1}^{L} \mathbf{A}_\tau^{(k)}\mathbf{X}_{t-\tau} + \varepsilon^{(k)}, \\ \varepsilon^{(k)} \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right), 1 \leq k \leq K \Bigg\}$$

and let $\mathcal{H}_K$ be the family of all $K-$finite mixtures of elements from $\mathcal{F}$, i.e.

$$\mathcal{H}_K = \left\{ h : h = \sum_{k=1}^{K} \pi_k \mathcal{L}_{\mathcal{M}^{(k)}}, \mathcal{L}_{\mathcal{M}^{(k)}} \in \mathcal{F}, \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1 \right\}$$

where $\mathcal{L}_{\mathcal{M}^{(k)}}(x) = p\left(x \mid \mathcal{M}^{(k)}\right)$ denotes the likelihood of x evaluated with causal model $\mathcal{M}^{(k)}$.

Then the family $\mathcal{H}_K$ is identifiable if and only if the following condition is met:

$$\text{The ordered pairs } \left(\left[\mathbf{B}^{(k)}\right]^{-1} \mathbf{A}_1^{(k)}, ..., \left[\mathbf{B}^{(k)}\right]^{-1} \mathbf{A}_L^{(k)}, \left[\mathbf{B}^{(k)}\right] \left[\mathbf{B}^{(k)}\right]^T \right) \text{ are distinct over all } k, \tag{9}$$

where $\mathbf{B}^{(k)} = \mathbf{I} - \mathbf{W}^{(k)}$.

Proof. Note that the SVAR equations can equivalently be written as:

$$\mathbf{X}_t = \left[\mathbf{B}^{(k)}\right]^{-1} \sum_{\tau=1}^{L} \mathbf{A}_\tau^{(k)}\mathbf{X}_{t-\tau} + \left[\mathbf{B}^{(k)}\right]^{-1} \varepsilon^{(k)}$$

where $\mathbf{B}^{(k)} = \left(\mathbf{I} - \mathbf{W}^{(k)}\right)$.

This implies that

$$p\left(\mathbf{X}_t \mid \mathbf{X}_{t-1}, ..., \mathbf{X}_{t-L}, \mathcal{M}^{(k)}\right) \sim \mathcal{N}\left(\left[\mathbf{B}^{(k)}\right]^{-1} \sum_{\tau=1}^{L} \mathbf{A}_\tau^{(k)}\mathbf{X}_{t-\tau}, \sigma^2 \left[\mathbf{B}^{(k)}\right]^{-1} \left[\mathbf{B}^{(k)}\right]^{-T} \right).$$

Following a similar argument as in the proof of Theorem B, the finite mixture is identifiable if and only if the parameters of the resultant Gaussian distributions are distinct as a function of $\{\mathbf{X}_{t-\tau}\}_{\tau=1}^{L}$. Hence, the condition. □

It can be difficult to reason about the identifiability of a mixture of SCMs whose structural equations come from a general class of functions, or whose noise distribution is non-Gaussian. However, the likelihood can be evaluated quite easily on a finite number of points, at least approximately if not exactly.

Here, we describe a sufficient condition for the identifiability of finite mixtures of *identifiable* causal models.

Theorem 3 (Identifiability of finite mixture of causal models)**.** Let $\mathcal{F}$ be a family of K identifiable causal models, i.e. $\mathcal{F} = \left\{\mathcal{L}_{\mathcal{M}}^{(k)} : \mathcal{M} \text{ is an identifiable causal model }, 1 \leq k \leq K\right\}$ and let $\mathcal{H}_K$ be the family of all $K-$finite mixtures of elements from $\mathcal{F}$, i.e.

$$\mathcal{H}_K = \left\{ h : h = \sum_{k=1}^{K} \pi_k \mathcal{L}_{\mathcal{M}_k}, \mathcal{L}_{\mathcal{M}_k} \in \mathcal{F}, \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1 \right\}$$

where $\mathcal{L}_{\mathcal{M}_k}(x) = \sum_{\mathcal{M}} p(x \mid \mathcal{M}) p(\mathcal{M}_k = \mathcal{M})$ denotes the likelihood of x evaluated with causal model $\mathcal{M}_k$. Further, assume that the following condition is met:

$$\text{For every } k, 1 \leq k \leq K, \exists a_k \in \mathbb{X} \text{ such that } \frac{\mathcal{L}_{\mathcal{M}_k}(a_k)}{\sum_{j=1}^{K} \mathcal{L}_{\mathcal{M}_j}(a_k)} > \frac{1}{2}. \tag{*}$$

Then the family $\mathcal{H}_K$ is identifiable, i.e., if $h_1 = \sum_{k=1}^{K} \pi_k \mathcal{L}_{\mathcal{M}_k}$ and $h_2 = \sum_{j=1}^{K} \pi'_j \mathcal{L}_{\mathcal{M}'_j} \in \mathcal{H}_K$ then:

$$h_1 = h_2 \implies \forall k \in \{1, \ldots, K\} \ \exists j \in \{1, \ldots, K\} \text{ such that } \pi_k = \pi'_j \text{ and } \mathcal{M}_k = \mathcal{M}'_j.$$

Proof. From Theorem A, we have that $\mathcal{H}_K$ is identifiable if and only if for any $\alpha_1, \ldots, \alpha_K \in \mathbb{R}$,

$$\sum_{j=1}^{K} \alpha_j \mathcal{L}_{\mathcal{M}_j} = 0 \implies \alpha_j = 0 \ \ \forall j \in \{1, \ldots, K\}$$

Note that $\sum_{j=1}^{K} \alpha_j \mathcal{L}_{\mathcal{M}_j} = 0 \implies \sum_{j=1}^{K} \alpha_j \mathcal{L}_{\mathcal{M}_j}(x) = 0 \quad \forall x \in \mathbb{X}$. In particular,

$$\sum_{j=1}^{K} \alpha_j \mathcal{L}_{\mathcal{M}_j}(a_k) = 0 \quad \forall k \in \{1, \ldots, K\}, \tag{10}$$

where a_k is as defined in Condition ((*)). Denote $\mathcal{L}_{\mathcal{M}_j}(a_k) = \beta_{kj}$. Then Equation (10) can be written as:

$$\begin{bmatrix} \beta_{11} & \ldots & \beta_{1K} \\ \vdots & & \vdots \\ \beta_{K1} & \ldots & \beta_{KK} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix} = \mathbf{0}. \tag{11}$$

Or equivalently

$$\boldsymbol{\beta}\boldsymbol{\alpha} = \mathbf{0}. \tag{12}$$

Note that $\boldsymbol{\alpha} = 0$ if and only if $\boldsymbol{\beta}$ is full rank. We now show that Condition ((*)) implies that $\boldsymbol{\beta}$ is strictly diagonally dominant and hence full rank. Note that Condition ((*)) can be equivalently written as:

$$\begin{aligned} \frac{\beta_{kk}}{\sum_{j=1}^{K} \beta_{kj}} > \frac{1}{2} &\implies 2\beta_{kk} > \sum_{j=1}^{K} \beta_{kj} \\ &\implies \beta_{kk} > \sum_{j=1, j \neq k}^{K} \beta_{kj} \end{aligned}$$

which implies strict diagonal dominance since $\beta_{kj} \geq 0 \quad \forall k, j$. Hence $\boldsymbol{\alpha} = 0$ thus implying linear independence. □

Note that a_k refers to any point in the support of the mixture distribution such that the condition (*) is satisfied. It does not constitute a 'sample' from the k^{th} SCM in the conventional sense of being randomly drawn from the SCM. Instead, it can be intentionally chosen to meet the specified condition.

A.4. Relationship between ELBO and log-likelihood

In this section, we derive an exact relationship between the derived evidence lower bound $\text{ELBO}(\theta, \phi, \psi)$ and the log-likelihood $\log p_\theta(X)$.

First, note that:

$$p_\theta(X) p\left(\mathcal{M}_{1:K} \mid X\right) = p_\theta\left(X \mid \mathcal{M}_{1:K}\right) p\left(\mathcal{M}_{1:K}\right)$$

and hence:

$$p_\theta(X) = \frac{p_\theta\left(X \mid \mathcal{M}_{1:K}\right) p\left(\mathcal{M}_{1:K}\right)}{p\left(\mathcal{M}_{1:K} \mid X\right)}.$$

The log-likelihood can be written as:

$$\begin{aligned}
\log p_\theta(X) &= \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})}\left[\log p_\theta(X)\right] \\
&= \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})}\left[\log \frac{p_\theta\left(X \mid \mathcal{M}_{1:K}\right) p\left(\mathcal{M}_{1:K}\right)}{p\left(\mathcal{M}_{1:K} \mid X\right)} \times \frac{q_\phi(\mathcal{M}_{1:K})}{q_\phi(\mathcal{M}_{1:K})}\right] \\
&= \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})}\left[\log p_\theta\left(X \mid \mathcal{M}_{1:K}\right) + \log p\left(\mathcal{M}_{1:K}\right)\right] \\
&+ \sum_{i=1}^{K} \mathrm{H}\left(q_\phi(\mathcal{M}_i)\right) + \mathrm{KL}\left(q_\phi\left(\mathcal{M}_{1:K}\right) \,||\, p\left(\mathcal{M}_{1:K} \mid X\right)\right) \\
&= \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})}\left[\sum_{n=1}^{N} \log p_\theta\left(X^{(n)} \mid \mathcal{M}_{1:K}\right) + \sum_{i=1}^{K} \log p\left(\mathcal{M}_i\right)\right] \\
&+ \sum_{i=1}^{K} \mathrm{H}\left(q_\phi(\mathcal{M}_i)\right) + \mathrm{KL}\left(q_\phi\left(\mathcal{M}_{1:K}\right) \,||\, p\left(\mathcal{M}_{1:K} \mid X\right)\right)
\end{aligned}$$

Also note that, using the rules of conditional probability:

$$\begin{aligned}
\frac{p_\theta(X^{(n)} \mid \mathcal{M}_{1:K})}{p_\theta(X^{(n)} \mid \mathcal{M}_{1:K}, Z^{(n)})} &= \frac{p_\theta(X^{(n)}, \mathcal{M}_{1:K})}{p(\mathcal{M}_{1:K})} \times \frac{p(Z^{(n)}, \mathcal{M}_{1:K})}{p_\theta(X^{(n)}, Z^{(n)}, \mathcal{M}_{1:K})} \\
&= \frac{p(Z^{(n)} \mid \mathcal{M}_{1:K})}{p(Z^{(n)} \mid X^{(n)}, \mathcal{M}_{1:K})} \\
&= \frac{p(Z^{(n)})}{p(Z^{(n)} \mid X^{(n)}, \mathcal{M}_{1:K})}
\end{aligned}$$

where the last step follows from the fact that $Z^{(n)}$ and $\mathcal{M}_i$ are independent.

Thus, we can write:

$$\begin{aligned}
p_\theta(X^{(n)} \mid \mathcal{M}_{1:K}) &= \mathbb{E}_{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)}\left[p_\theta(X^{(n)} \mid \mathcal{M}_{1:K})\right] \\
&= \mathbb{E}_{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)}\left[\frac{p_\theta(X^{(n)} \mid \mathcal{M}_{1:K}, Z^{(n)}) p(Z^{(n)})}{p(Z^{(n)} \mid X^{(n)}, \mathcal{M}_{1:K})}\right] \\
&= \mathbb{E}_{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)}\left[\frac{p_\theta(X^{(n)} \mid \mathcal{M}_{Z^{(n)}}) p(Z^{(n)})}{p(Z^{(n)} \mid X^{(n)}, \mathcal{M}_{1:K})} \times \frac{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)}{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)}\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
\log p_\theta(X) = \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})}\Bigg[\sum_{n=1}^{N} \mathbb{E}_{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)}\left[\log p_\theta\left(X^{(n)} \mid \mathcal{M}_{Z^{(n)}}\right) + \log p(Z^{(n)})\right] + \mathrm{H}\left(r_\psi\left(Z^{(n)} \mid X^{(n)}\right)\right) \\
+ \mathrm{KL}\left(r_\psi(Z^{(n)} \mid X^{(n)}) \,||\, p(Z^{(n)} \mid X^{(n)}, \mathcal{M}_{1:K})\right) + \sum_{i=1}^{K} \log p\left(\mathcal{M}_i\right)\Bigg] + \sum_{i=1}^{K} \mathrm{H}\left(q_\phi(\mathcal{M}_i)\right) \\
+ \mathrm{KL}\left(q_\phi\left(\mathcal{M}_{1:K}\right) \,||\, p\left(\mathcal{M}_{1:K} \mid X\right)\right).
\end{aligned}$$

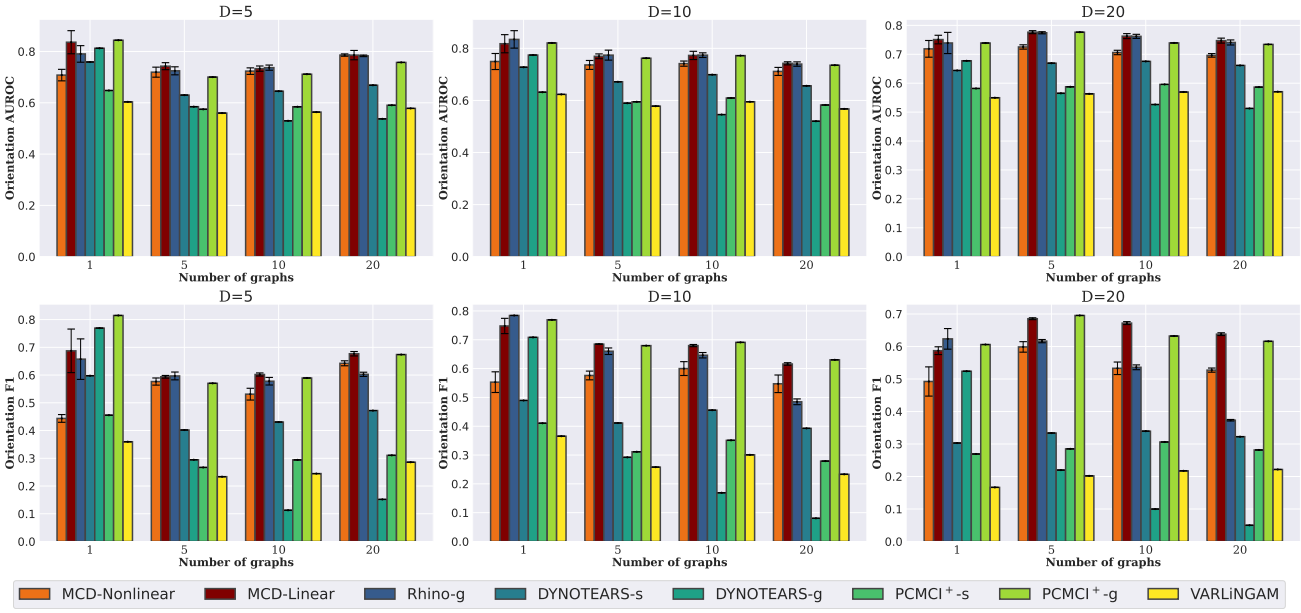

Figure 9. Results on the linear synthetic dataset for $D = 5, 10, 20$. We report both the orientation F1 and AUROC scores. Average of 5 runs reported.

Noting that

$$\begin{aligned}\text{ELBO}(\theta, \phi, \psi) \equiv \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\mathbb{E}_{r_\psi\left(Z^{(n)} \mid X^{(n)}\right)} \left[\log p_\theta(X^{(n)} \mid \mathcal{M}_{Z^{(n)}}) + \log p(Z^{(n)}) - \log r_\psi \left(Z^{(n)} \mid X^{(n)} \right) \right] \right] \\ + \sum_{i=1}^{K} \mathbb{E}_{q_\phi(\mathcal{M}_i)} \left[\log p(\mathcal{M}_i) - \log q_\phi(\mathcal{M}_i) \right]\end{aligned}$$

we obtain that:

$$\begin{aligned}\log p_\theta(X) = \text{ELBO}(\theta, \phi, \psi) + \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathcal{M}_{1:K})} \left[\text{KL} \left(r_\psi \left(Z^{(n)} \mid X^{(n)} \right) \mid\mid p(Z^{(n)} \mid X^{(n)}, \mathcal{M}_{1:K}) \right) \right] \\ + \text{KL} \left(q_\phi \left(\mathcal{M}_{1:K} \right) \mid\mid p \left(\mathcal{M}_{1:K} \mid X \right) \right).\end{aligned}$$

B. Additional experiments

B.1. More results on the synthetic datasets

Figure 9 shows the results of all methods on the linear synthetic datasets, and Figure 10 shows the results on the nonlinear synthetic datasets. We observe that the difference in performance between MCD and Rhino-g is much lower in terms of AUROC compared to orientation F1. MCD is able to achieve similar performance to the 'gold-standard' baseline Rhino-g despite not having ground-truth membership information.

B.2. S&P100

Setup. We provide more details on the setup for the experiment with the S&P 100 dataset. We used grid search to iterate over multiple values for the sparsity term λ in equation (18), number of graphs K, and 5 random seeds. We picked the setting that yielded the lowest validation loss and reported results using the inferred causal graphs. We aggregate the temporal adjacency matrix following the procedure described in Appendix D.5.

Additional results. Figure 11 shows the heatmap of the two aggregated causal graphs inferred by MCD-Nonlinear. As noted in the main paper, we observe several interesting differences between the two graphs. Many sectors such as 'Industrials',

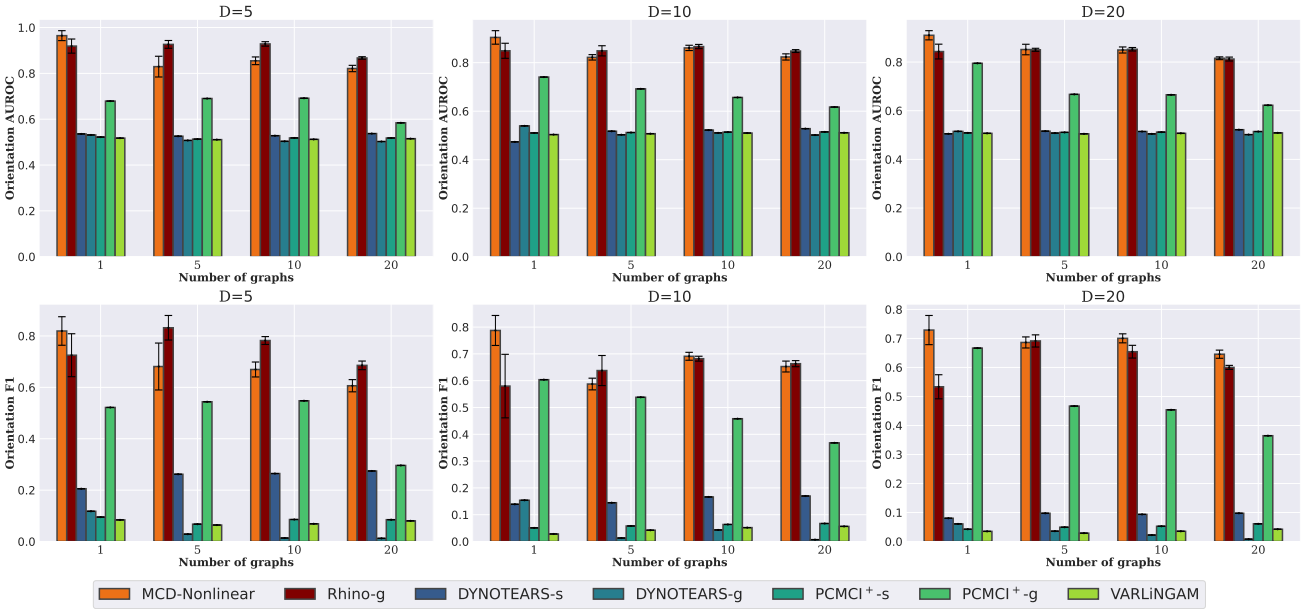

Figure 10. Results on the nonlinear synthetic dataset for $D = 5, 10, 20$. We report both the orientation F1 and AUROC scores. Average of 5 runs reported.

'Utilities' and 'Technology' seem to have different patterns of intra-sector interactions in the two graphs. Further, Graph 1 shows more marked interactions for stocks in the 'Real Estate' sector.

We also visualize the stock prices of the companies whose interactions changed between the two graphs and overlaid the membership information over their indices, as we did with Figure 6. Figure 12 provides examples of 6 such stocks. We observe that in most cases, the 'red' periods, i.e., periods in which Graph 1 is active, show more pronounced trends and marked movements of the stock prices compared to the 'blue' periods.

Additionally, we run MCD-Linear on this dataset and observe that it only discovers a single mode in the dataset. Figure 13 shows the discovered causal graph from MCD-Linear.

B.3. Netsim

Setup. We additionally experiment with a different setup on the Netsim Brain connectivity dataset. We combine the time series with length $T = 200$ and number of nodes $D = 5$ from simulations 1, 8, 10, 13, 14, 15, 16, 18, 21, 22, 23, and 24. This dataset comprises $N = 600$ samples, with $K^* = 14$ distinct underlying causal graphs. We refer to this setup as **Netsim**. This dataset exhibits significant graph membership imbalance, with the top 3 causal graphs accounting for 500 out of the 600 samples. Hence, we consider an exponentially weighted prior for the membership indicators, i.e. $p(Z = k) \propto \exp(-\lambda_p k)\ \forall k \in \{1, \dots, K\}$. We set $\lambda_p = 5$ and $K = 20$.

Results. In this setup, we observe that MCD-Nonlinear and MCD-Linear are outperformed by the baselines PCMCI$^+$ and Rhino, even though they only predict one graph for the entire dataset. This is attributed to the similarity among the various underlying graphs in the Netsim dataset and the strong imbalance in the data. Our model faces sample complexity issues because it learns multiple causal graphs, whereas other methods perform reasonably well by predicting only one. This highlights the idea that learning a mixture model is only beneficial when the underlying SCMs differ from one another significantly. In such a scenario, the benefits of learning multiple graphs outweigh the drawbacks of limited samples per model. This explanation is also supported by the observation that PCMCI$^+$ (grouped) achieves lower performance than its single graph counterpart. Further, MCD-Nonlinear and MCD-Linear achieve relatively low clustering accuracy of $35.2 \pm 6.6\%$ and $35.4 \pm 5.2\%$, due to the inherent similarities in the underlying SCMs.

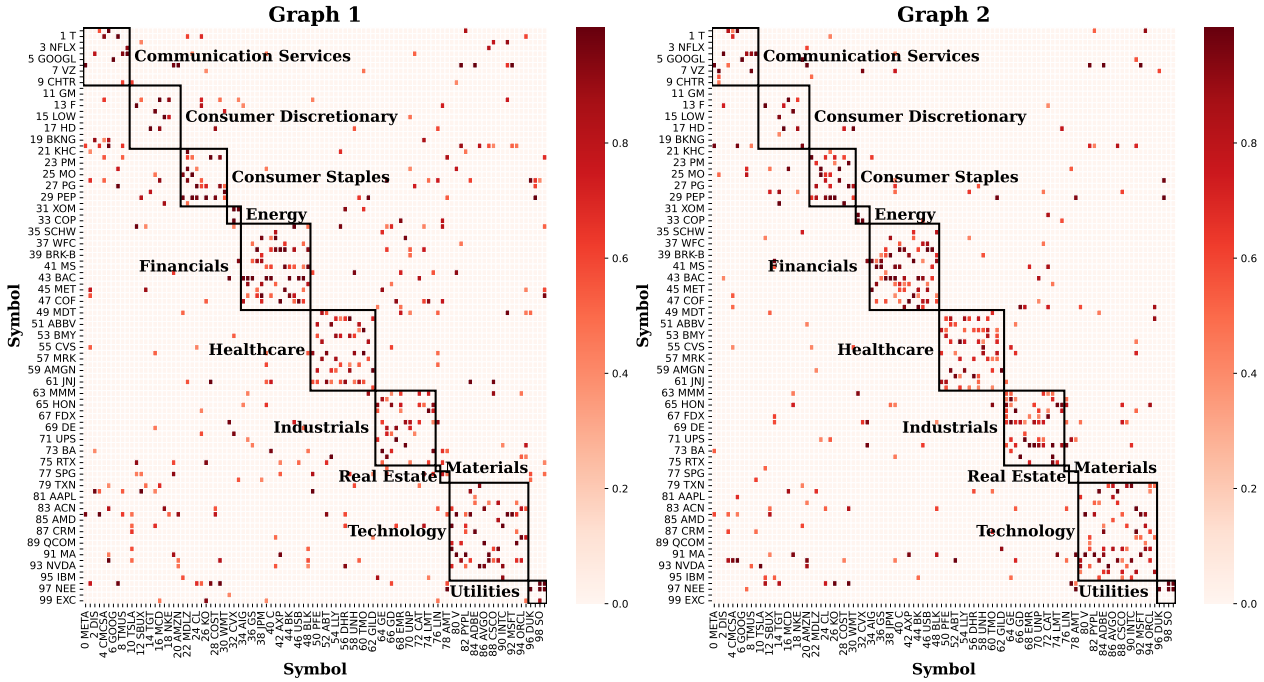

Figure 11. Discovered causal graphs from S&P 100. MCD-Nonlinear discovers two distinct graphs from the dataset.

	Netsim	
Method	**AUROC** (↑)	**F1** (↑)
PCMCI$^+$-s	0.702	0.648
PCMCI$^+$-o	0.827	**0.803**
PCMCI$^+$-g	0.810	0.785
VARLiNGAM	0.638	0.598
DYNOTEARS-s	0.706	0.588
DYNOTEARS-o	0.674	0.626
DYNOTEARS-g	0.629	0.584
Rhino	$\mathbf{0.873 \pm 0.026}$	0.707 ± 0.033
MCD-Nonlinear (this paper)	0.733 ± 0.060	0.607 ± 0.052
MCD-Linear (this paper)	0.728 ± 0.018	0.623 ± 0.022

Table 2. Results on the Netsim dataset. -s indicates that the baseline predicts one graph per sample. -o indicates that the baseline predicts one graph for the whole dataset. -g signifies that the baseline is run on samples grouped according to the ground truth causal graph. VARLiNGAM does not run on the DREAM3 dataset. MCD-Nonlinear achieves a clustering accuracy of $35.2 \pm 6.6\%$ on Netsim, while MCD-Linear has a clustering accuracy of $35.4 \pm 5.2\%$.

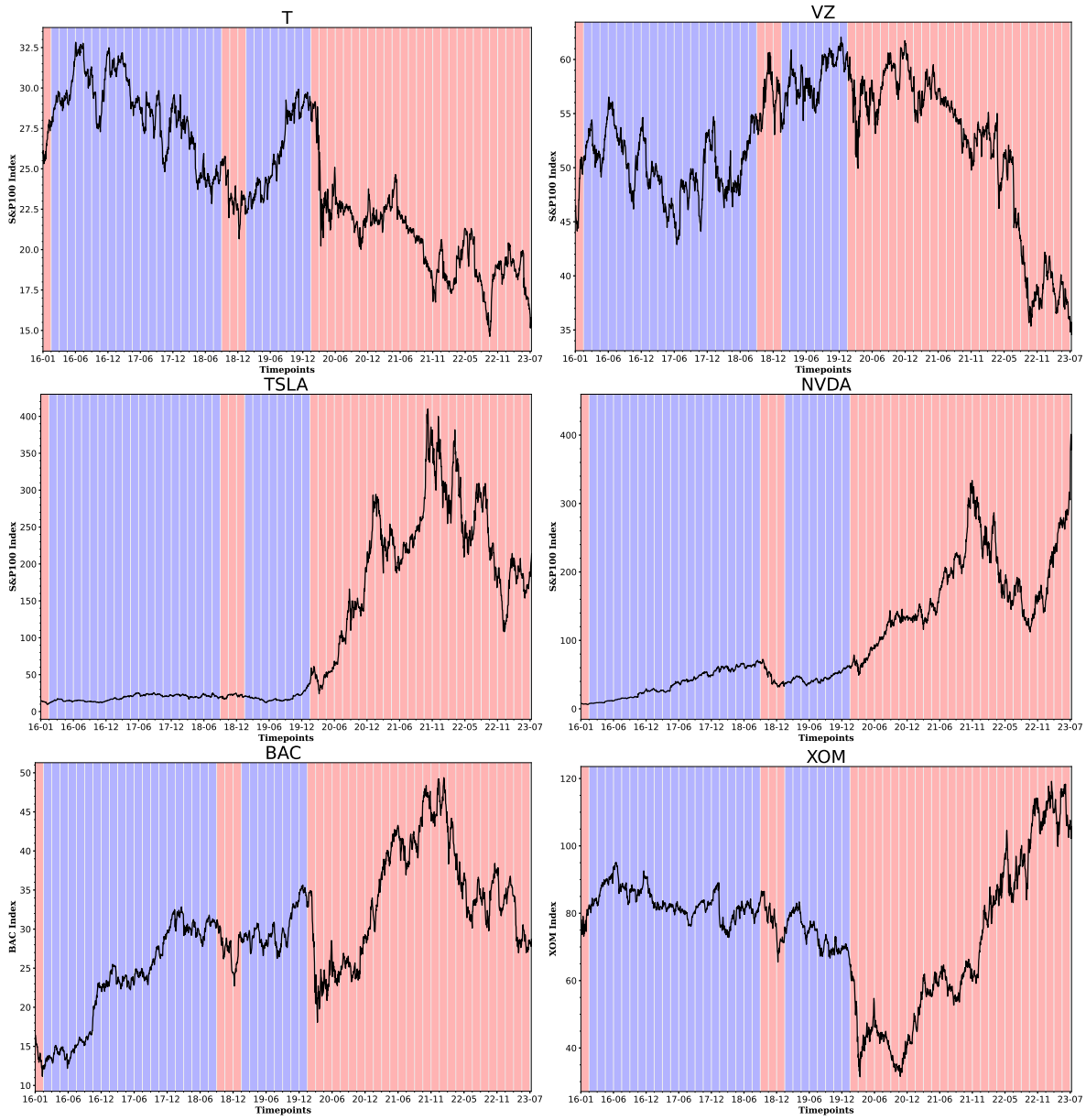

Figure 12. Segmentation of the stock prices of AT&T, Verizon, Tesla, NVIDIA, Bank of America, and Exxon Mobil stock prices with respect to the inferred causal graphs from MCD-Nonlinear.

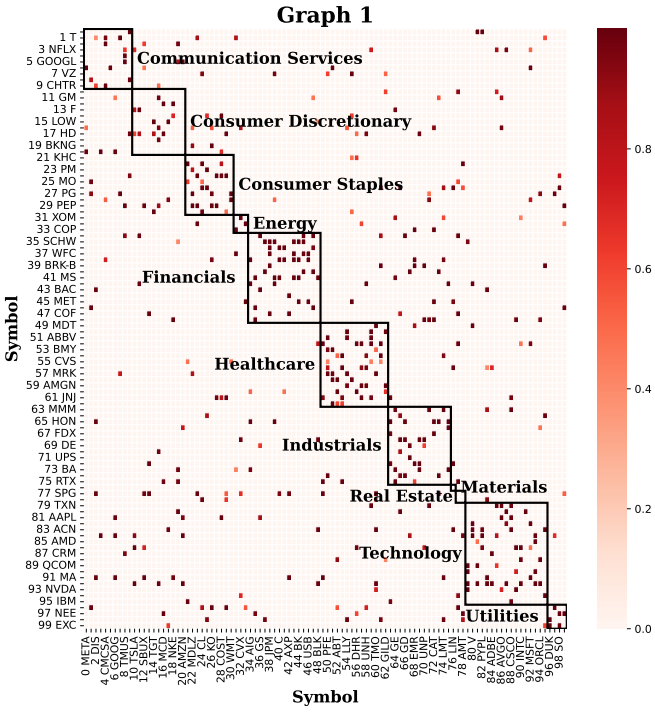

Figure 13. Discovered causal graph from S&P100 using MCD-Linear. MCD-Linear only discovers a single mode in the dataset

B.4. Ablation studies

Effect of number of samples per component. We investigate the effect of a decreasing number of samples per mixture component on the performance of MCD, as the number of ground truth SCMs K^* increases. We consider synthetic nonlinear data of dimension $D = 10$ with $K^* = 40, 60, 80, 100$ ground truth graphs in addition to the settings discussed in the main paper. We run MCD-Nonlinear, Rhino and PCMCI$^+$ on $N = 1000$ samples generated from K^* SCMs for increasing values of K^*, with $K = 2K^*$. The results are presented in Figure 14. MCD-Nonlinear suffers a gradual decrease in model performance, with roughly a 40% decrease in F1 and 22% decrease in AUROC from $K^* = 1$ to $K^* = 100$. Meanwhile, the performance of Rhino falls off more drastically and becomes equivalent to random guessing for large K^*. The performance of PCMCI$^+$ (grouped) also decreases quite rapidly with the increase in K^*.

Using ground truth membership assignments. We assess MCD performance with learned versus ground-truth membership associations on synthetic data with $D = 10$. As before, we set $K = 2K^*$. Figure 15 shows the results of this ablative experiment run with MCD-Nonlinear on the nonlinear synthetic dataset. The performance of MCD-Nonlinear with ground truth labels and Rhino-g is theoretically an upper bound on its performance. Encouragingly, we observe that our model performs close to this upper bound.

Effect of similarity of the graphs on performance We examine the performance of MCD when the causal graphs are similar to each other. The dataset setup is as follows: we first generate a random ER graph. We then perturb each edge with a probability p, i.e. flip the entry in the adjacency matrix from 0 to 1, and vice versa with probability p. We check if the resulting graph is a DAG. If yes, we add it to the pool of generating graphs. We repeat this procedure until we obtain K^* DAGs. We then generate the synthetic dataset ($N = 1000$ samples) with these K^* DAGs using randomly generated MLPs and spline functions. We run MCD-Nonlinear with $K = 2K^*$ on the resulting datasets with $D = 10$ and $K = 5, 10$. We set $p = 0.005, 0.008, 0.01, 0.05, 0.1$, corresponding to varying levels of similarity of the underlying graphs. We also attach the pair-wise statistics for the resulting graphs in Table 4. The results are reported in Figure 16

The results indicate that MCD can achieve good clustering accuracy and causal discovery performance even when the constituent causal graphs are similar. In the $K^* = 5$ case, we notice that the clustering accuracy remains high for all considered settings, and the F1 score decreases slightly for the higher values of $-\log p$ when the causal graphs are similar. For $K^* = 10$, the clustering accuracy is considerably lower for $p = 0.005$, but nevertheless the F1 score remains high.

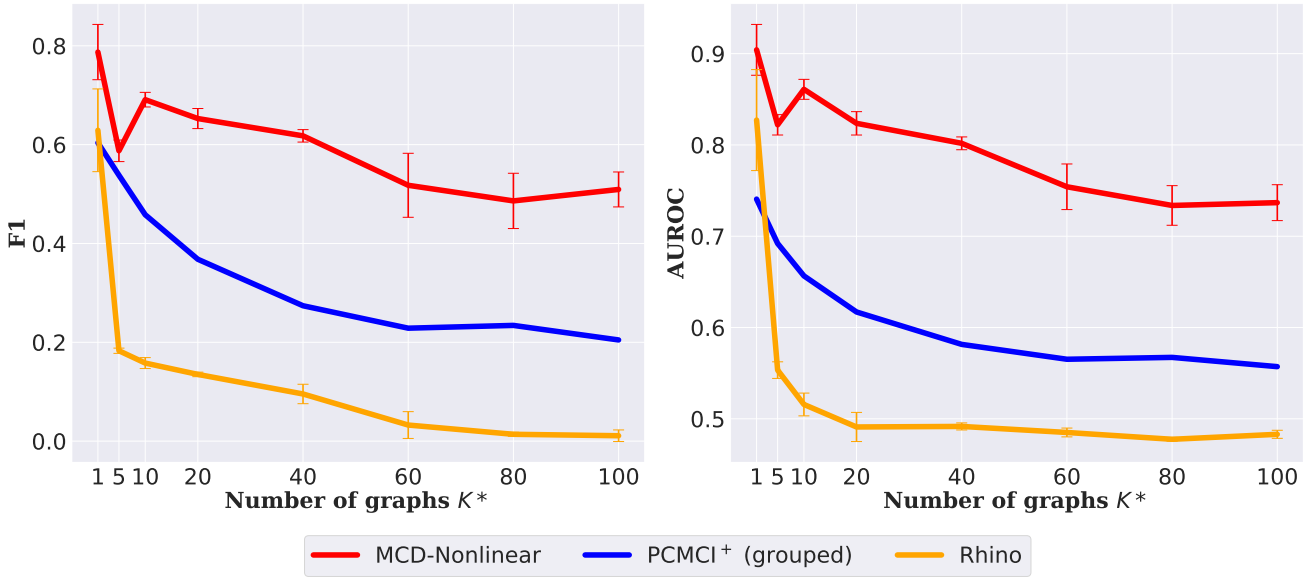

Figure 14. Effect of increasing number K^* of underlying SCMs on F1 and AUROC on synthetic data with $D = 10$. MCD's performance declines gradually with decreased number of samples per mixture component, while Rhino's performance decays drastically. The performance of Rhino (grouped) decays slightly faster than MCD.

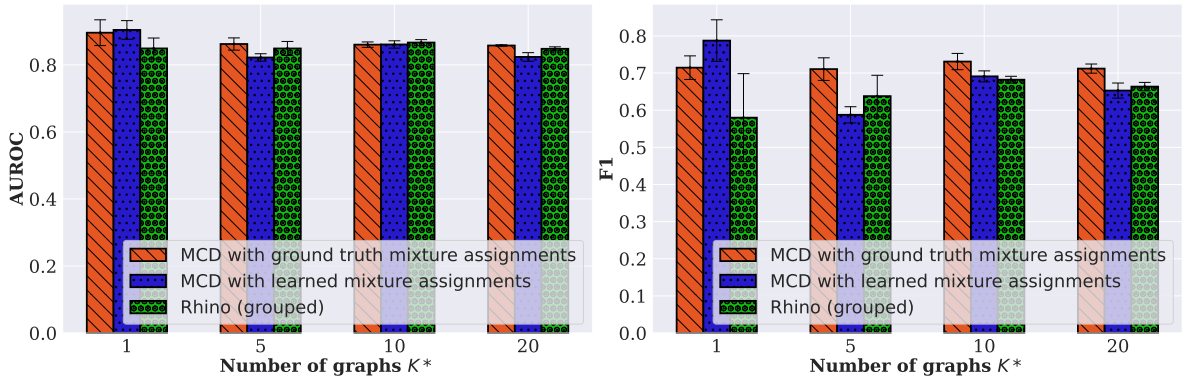

Figure 15. Comparison of model performance of MCD with ground-truth versus learned mixture assignments and Rhino (grouped) on synthetic data with $D = 10$. Expectedly, MCD performs better with explicit information about the cluster assignments, but it achieves comparable performance even with learned membership information.

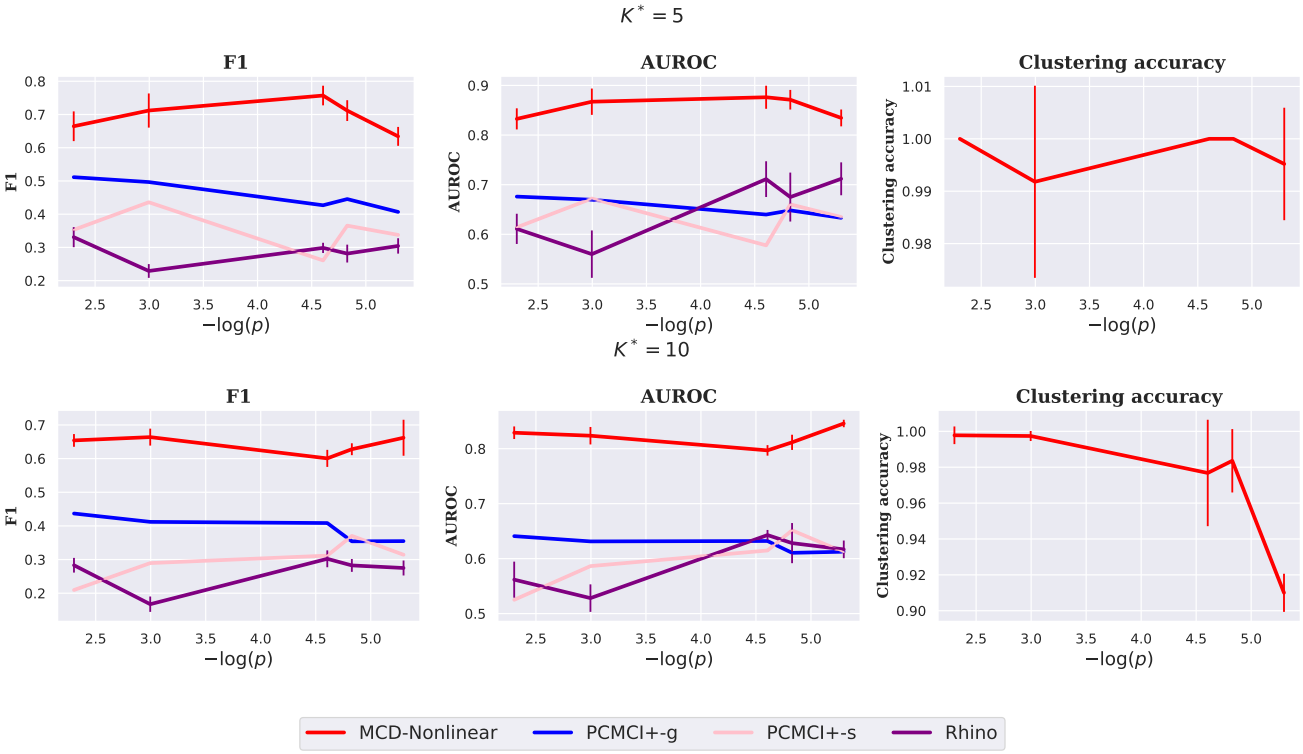

Figure 16. Causal discovery performance and clustering accuracy of MCD on synthetic datasets with similar graphs. MCD achieves good causal discovery performance and clustering accuracy even when $-\log p$ is high, i.e., the graphs are similar.

On the other hand, PCMCI$^+$-g suffers a gradual performance drop as the graphs become similar, while the single graph baselines Rhino and PCMCI$^+$-s have slightly better performance when the constituent causal graphs are more similar.

B.5. Clustering accuracy for $D = 5, 10, 20$ on synthetic datasets

Figure 17 shows the clustering accuracy of MCD-Linear and MCD-Nonlinear for different values of D on the linear and nonlinear synthetic datasets, respectively. For all settings, we set the hyperparameter $K = 2K^*$. We observe that for most values of $K^* > 1$, the clustering accuracy is quite high, while it remains low for $K^* = 1$. Both MCD-Linear and MCD-Nonlinear are particularly good at clustering for higher number of nodes D. As noted earlier, the low clustering accuracy for $K^* = 1$ is expected since the single mode in the data distribution is 'split' across two learned causal graphs.

B.6. Clustering progression with training

We analyze the progression of clustering accuracy and the number of unique graphs learned with the number of training steps for MCD-Nonlinear on the nonlinear synthetic dataset with $D = 10, K^* = 10$. As training progresses, not all K graphs are utilized. We count only those graphs for which at least one associated sample exists. Figure 18 shows the plots. We observe that when $K = 20$, as training progresses, the algorithm groups together points from different causal graphs until they converge to the "true" number of causal graphs $K^* = 10$ and clustering accuracy converges to (approximately) 100%; however, when $K = 10$, we observe that the number of unique graphs can sometimes fall below $K^* = 10$, resulting in suboptimal clustering accuracy.

B.7. Netsim visualization

Figure 19 shows a visualization of a heatmap of the predictions for the Netsim-mixture dataset. The 3 ground truth adjacency matrices and the top-3 discovered adjacency matrices, ranked by the prediction frequency, are shown. All 3 matrices achieve a high AUROC score, even though the poor calibration of scores results in the prediction of many spurious edges.

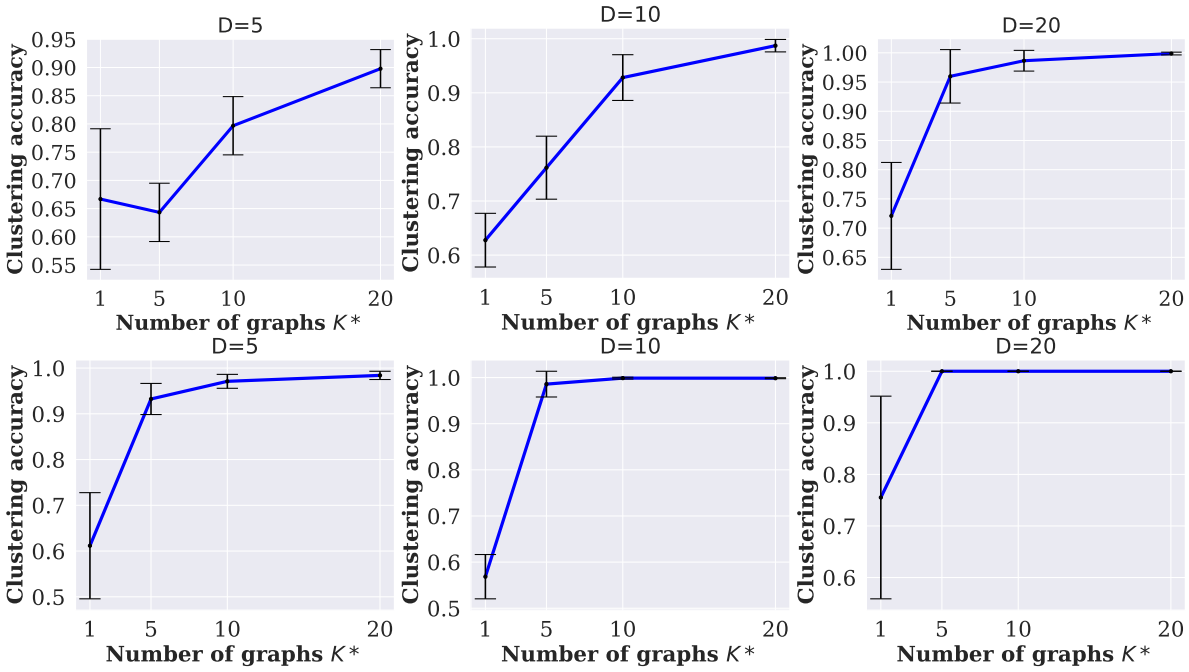

Figure 17. Plots showing the clustering accuracy of (top) MCD-Linear and (bottom) MCD-Nonlinear vs K^* on the linear and nonlinear synthetic datasets for $D = 5, 10, 20$.

B.8. Timing analysis

In this section, we analyze the run-time of MCD-Nonlinear as a function of the hyperparameter K. As noted in Section 3.3, MCD-Nonlinear, in theory, needs roughly K times more operations than Rhino in each epoch due to the evaluation of the expectation over the variational distribution $r_\psi\left(Z^{(n)} \mid X^{(n)}\right)$ while calculating the ELBO. However, this does not translate to a K times increase in model runtime. We measure and plot the total runtime for training our model for the synthetic dataset with $D = 10$ nodes as a function of K. Figure 20 shows the plot.

We observe that although the plot shows an approximately linear trend, the slope is much lesser than 1. In fact, a $100\times$ increase in K from 2 to 200 results in a less than $4\times$ increase in run-time. Thus, MCD scales reasonably well with the number of mixture components K.

C. Implementation details

In this section, we elaborate more on how we model the terms in equation (4) for MCD-Linear and MCD-Nonlinear.

We model the K SCMs as additive noise models. For the k^{th} causal model, we have:

$$X_t^{(n)}\Big|_k = f_k(\text{Pa}_{\mathcal{G}_k}(< t), \text{Pa}_{\mathcal{G}_k}(t)) + g_k(\text{Pa}_{\mathcal{G}_k}(< t), \epsilon_t),$$

where the function f_k models the structural equation between the nodes and g_k models the exogenous noise under causal model $\mathcal{M}_k$.

MCD-Linear. We implement the each of the K models using a linear model:

$$f_k^d\left(\text{Pa}_{\mathcal{G}_k}(\leq t)\right) = \sum_{\tau=0}^{L}\sum_{j=1}^{D}\left(\mathcal{G}_k \circ \mathcal{W}_k\right)_\tau^{j,d} \times X_{t-\tau}^{j,(n)}, \tag{13}$$

where $\circ$ denotes the Hadamard product, and $\mathcal{W}_k \in \mathbb{R}^{(L+1)\times D\times D}$ is a learned weight tensor. We only model independent noise with this model, and set $g_k(\text{Pa}_{\mathcal{G}_k}(< t), \epsilon_t) = \epsilon_t$, i.e., the identity function.

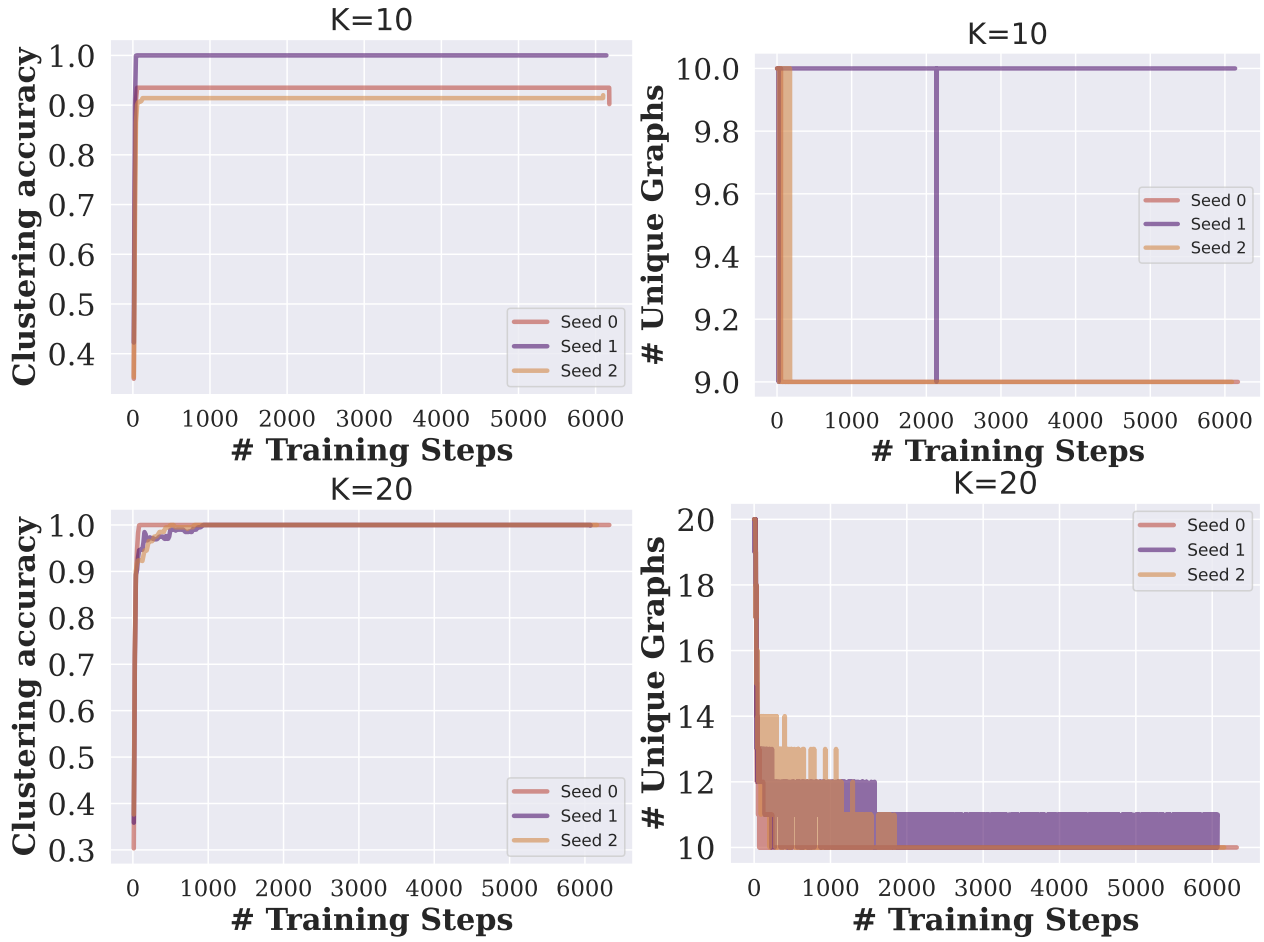

Figure 18. Plots showing the progression of (left) clustering accuracy (right) number of unique learned graphs for MCD-Nonlinear with the number of training steps on the nonlinear synthetic dataset with $D = 10, K^* = 10$. We observe that as training progresses, clustering accuracy increases for both the $K = 10$ and $K = 20$ runs; however, when $K = 10$, some runs tend to learn a lower number of graphs, thus resulting in suboptimal clustering accuracy.

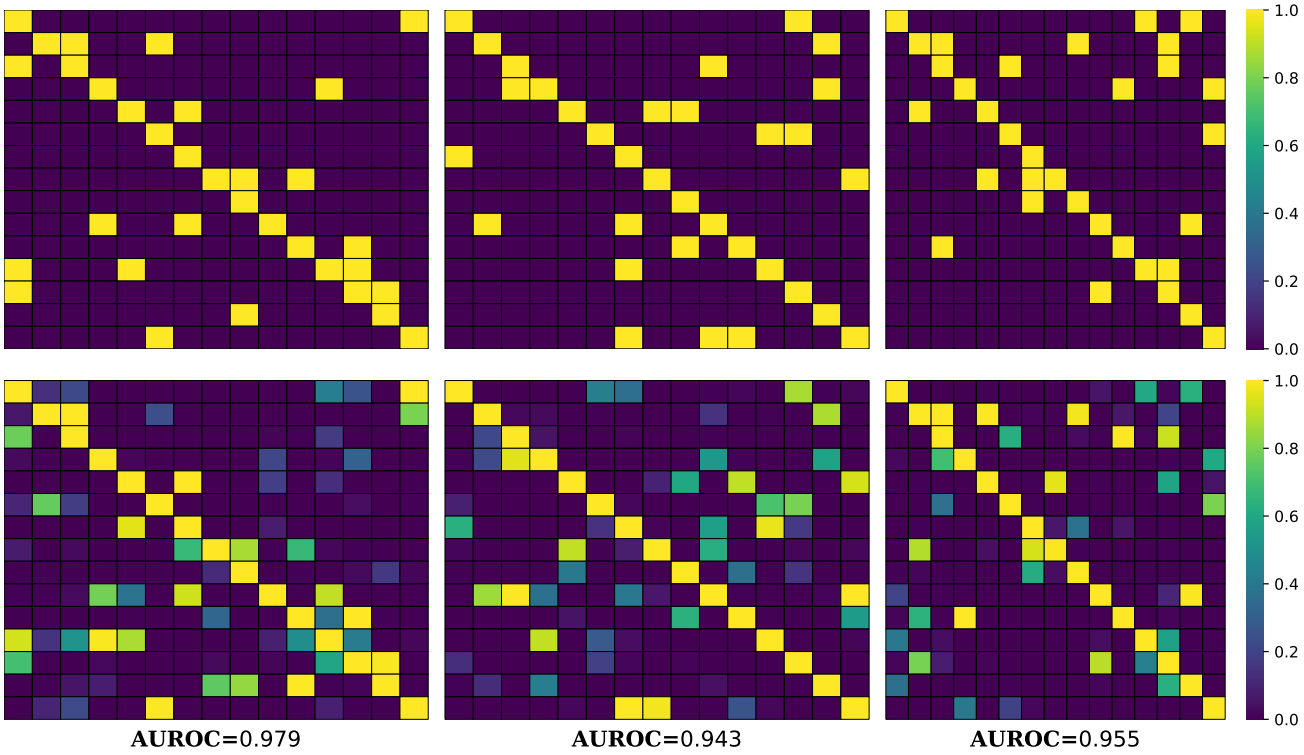

Figure 19. Heatmap for the Netsim-mixture dataset showing the (top) adjacency matrices of the ground-truth causal graphs, and the (bottom) edge probabilities for the top-3 discovered adjacency matrices (ranked by frequency of occurrence). We also report the graph-wise AUROC metrics.

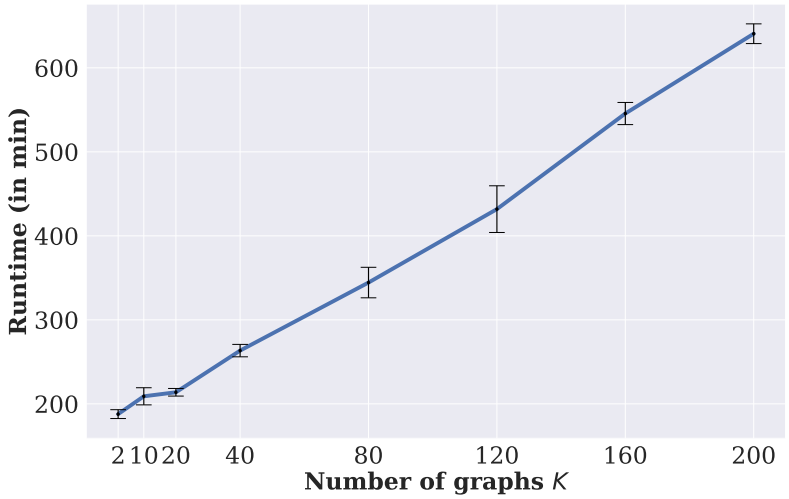

Figure 20. Run time plot of MCD as a function of K. A $100\times$ increase in K from 2 to 200 results in a less than $4\times$ increase in run-time.

Assuming that $\epsilon_t \sim \mathcal{N}(0, I)$, the marginal likelihood under each model $\mathcal{M}_k$ can be further simplified as follows, using the causal Markov assumption:

$$\begin{aligned}\log p_\theta\left(X_{1:T}^{(n)}\middle|\mathcal{M}_k\right) &= \sum_{t=L}^{T}\sum_{d=1}^{D}\log p_\theta\left(X_t^{d,(n)}\middle|\mathrm{Pa}_{\mathcal{G}_k^d}(<t), \mathrm{Pa}^d_{\mathcal{G}_k}(t)\right)\\ &= \sum_{t=L}^{T}\sum_{d=1}^{D}\left[X_t^{d,(n)} - f_k^d\left(\mathrm{Pa}^d_{\mathcal{G}_k}(<t), \mathrm{Pa}_{\mathcal{G}_k}\right)\right]^2\\ &= \sum_{t=L}^{T}\sum_{d=1}^{D}\left[X_t^{d,(n)} - \sum_{\tau=0}^{L}\sum_{j=1}^{D}\left(\mathcal{G}_k \circ \mathcal{W}_k\right)_\tau^{j,d} \times X_{t-\tau}^{j,(n)}\right]^2. \end{aligned} \tag{14}$$

MCD-Nonlinear. We use the Rhino model (Gong et al., 2022) to model each of the K SCMs. The functions f_k are modeled as follows:

$$f_k^d\left(\mathrm{Pa}_{\mathcal{G}_k}(\leq t)\right) = \Xi_f\left(\left[\sum_{\tau=0}^{L}\sum_{j=1}^{D}\left(\mathcal{G}_k\right)_\tau^{j,d} \times \ell_f\left(\left[X_{t-\tau}^{j,(n)}, \left(\mathcal{E}_k\right)_\tau^{j,(n)}\right]\right), \left(\mathcal{E}_k\right)_0^{d,(n)}\right]\right), \tag{15}$$

where Ξ_f and ℓ_f are multi-layer perceptron networks that are shared across all K causal models $\mathcal{M}_{1:K}$ and all D nodes, and $\mathcal{E}_k \in \mathbb{R}^{(L+1)\times D\times e}$ are embeddings (with embedding dimension e) corresponding to model k. A similar architecture is used for the hypernetwork that predicts parameters for the conditional spline flow model, with embeddings $\mathcal{F}_k$, and hypernetworks Ξ_ϵ and ℓ_ϵ. The only difference is that the output dimension of Ξ_ϵ is different, being equal to the number of spline parameters. The noise variables ϵ_t^d are described using a conditional spline flow model,

$$p_{g_k^d}(g_k^d(\epsilon_t^d) \mid \mathrm{Pa}^d_{\mathcal{G}_k}(<t)) = p_\epsilon(\epsilon_t^d)\left|\frac{\partial (g_k^d)^{-1}}{\partial \epsilon_t^d}\right|, \tag{16}$$

with ϵ_t^d modeled as independent Gaussian noise.

Using the causal Markov assumption:

$$\begin{aligned}\log p_\theta\left(X_{1:T}^{(n)}\middle|\mathcal{M}_k\right) &= \sum_{t=L}^{T}\sum_{d=1}^{D}\log p_\theta\left(X_t^{d,(n)}\middle|\mathrm{Pa}^d_{\mathcal{G}_k}(<t), \mathrm{Pa}^d_{\mathcal{G}_k}(t)\right)\\ &= \sum_{t=L}^{T}\sum_{d=1}^{D}\log p_{g_k^d}\left(u_t^{d,(n)}\middle|\mathrm{Pa}^d_{\mathcal{G}_k}(<t)\right)\end{aligned} \tag{17}$$

where $u_t^{d,(n)} = X_t^{d,(n)} - f_k^d\left(\mathrm{Pa}^d_{\mathcal{G}_k}(<t), \mathrm{Pa}^d_{\mathcal{G}_k}(t)\right)$.

The prior distribution $p(\mathcal{M}_{1:K})$ is modeled as follows:

$$p_\theta(\mathcal{M}_{1:K}) \propto \prod_{k=1}^{K}\exp\left(-\lambda\left\|\left(\mathcal{G}_k\right)_{1:T}\right\|^2 - \sigma h\left(\left(\mathcal{G}_k\right)_0\right)\right). \tag{18}$$

The first term is a sparsity prior and $h\left(\left(\mathcal{G}_k\right)_0\right)$ is the acyclicity constraint from (Zheng et al., 2018).

C.1. Calculation of clustering accuracy

We would like to evaluate the accuracy of our method in grouping samples based on the underlying SCMs. However, the assigned cluster indices by the model and the 'ground-truth' cluster indices might not match nominally, even though they refer to the same grouping assignment. For example, the cluster assignment of $(1, 1, 1, 2, 2)$ for $N = 5$ points is equivalent to the assignment $(2, 2, 2, 1, 1)$. In other words, we want a permutation invariant accuracy metric between the inferred cluster assignments $\tilde{Z}$ and true cluster assignments Z with $\tilde{Z}, Z \in \mathbb{N}^N$. We define

$$\text{Cluster Acc.}\left(\tilde{Z}, Z\right) = \max_{\pi\in S_K}\frac{1}{N}\sum_{n=1}^{N}1\left(\pi(\tilde{Z}_n) = Z_n\right)$$

with S_K denoting the permutation group over K elements. Evaluating the cluster accuracy naively would require $K!$ operations. However, we use the Hungarian algorithm to find the correct permutation in $O(K^3)$ time[1].

D. Experimental details

D.1. Synthetic datasets setup

This section provides more details about how we set up and run experiments using MCD on synthetic datasets. We set the number of mixture components K to twice that of true graphs (i.e., $K = 2K^*$) to showcase its robustness against over-specification of the number of components. We set a uniform prior for the mixing probabilities $p\left(Z^{(n)}\right)$, i.e. $p(Z^{(n)} = k) = \frac{1}{K}\, \forall k \in \{1, \dots, K\}$. Our implementation of the likelihood function for Rhino-g on the synthetic datasets matches the type of causal relationships modeled, i.e., we use the linear model described in Equation (5) on the linear dataset and the nonlinear variant described in Equation (6) for the nonlinear datasets.

Dataset generation. We generate two separate sets of synthetic datasets: a linear dataset with independent Gaussian noise and a nonlinear dataset with history-dependent noise modeled using conditional splines (Durkan et al., 2019). We generate a pool of K^* random graphs (specifically, Erdős-Rényi graphs) and treat them as ground-truth causal graphs. To generate a sample $X^{(n)}$, we assign it to a graph by drawing $Z^{(n)} \sim \text{Categorical}(K^*)$, and use the corresponding graph $\mathcal{G}_{Z^{(n)}}$ from this pool to model relationships between the variables.

Linear dataset. We model the data as:

$$X_t^{d,(n)} = \sum_{\tau=0}^{L} \sum_{j=1}^{D} (\mathcal{G}_{Z^{(n)}} \circ \mathcal{W}_{Z^{(n)}})_\tau^{j,d} \times X_{t-\tau}^{j,(n)} + \epsilon_t^d,$$

with $\epsilon_t^d \sim \mathcal{N}(0, 0.25)$. Each entry of the matrices $\mathcal{W}_k, \quad k = 1, \dots, K$ is drawn independently from $\mathcal{U}[0.1, 0.5] \cup \mathcal{U}[-0.5, -0.1]$.

Nonlinear dataset. We model the data as:

$$X_t^{d,(n)} = f_k^d \left(\text{Pa}^d_{\mathcal{G}_k}(< t), \text{Pa}^i_{\mathcal{G}_k}(t)\right) + \epsilon_t^d,$$

where f_k^d are randomly initialized multi-layer perceptrons (MLPs), and the random noise ϵ_t^d is generated using history-conditioned quadratic spline flow functions (Durkan et al., 2019).

D.2. Implementation of validation step for MCD

MCD learns a sample-wise membership variable $Z^{(n)}$ for every sample in the dataset by optimizing the ELBO. In order to evaluate the log-likelihood of the samples in the validation set, we still need to infer their corresponding membership information $Z^{(n)}$. Hence, during each validation step, we fix the weights of the other parameters and perform one step of gradient descent for the membership weights $Z^{(n)}$ with respect to the ELBO for the samples in the validation set. This way, we ensure that all samples in the input dataset are assigned to a mixture component, and the validation likelihood can be evaluated.

D.3. Hyperparameter details

Since MCD uses the acyclicity constraint from Zheng et al. (2018), we use an augmented Lagrangian training procedure to ensure that our model produces DAGs. We closely follow the implementation of the procedure from Geffner et al. (2022); Gong et al. (2022) with one exception: we modify the convergence criteria by increasing the number of outer steps for which the DAG penalty needs to be lower than a threshold (set to 10^{-8}). This modification enables the model to train for longer and prevents premature stopping. In the interest of fairness, we make this change to both MCD and Rhino.

We used the rational spline flow model described in (Durkan et al., 2019). We use the quadratic or linear rational spline flow model in all our experiments, both with 8 bins. The MLPs ℓ and Ξ have 2 hidden layers each, with hidden dimensions

[1]This approach and implementation are adapted from https://smorbieu.gitlab.io/accuracy-from-classification-to-clustering-evaluation/

Dataset	Synthetic ($D = 5, 10, 20$)	Netsim-mixture	DREAM3	S&P100	Netsim
Hyperparameter					
Matrix LR	10^{-2}	10^{-2}	10^{-3}	10^{-2}	10^{-2}
Likelihood LR	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
Batch Size	128	64	64	64	64
# Outer auglag steps	100	60	60	60	60
# Max inner auglag steps	6000	2000	6000	2000	2000
Embedding dim e	$= D$	15	32	100	15
Sparsity factor λ	5	25	10	20	25
Spline type	Quadratic	Linear	Linear	Linear	Linear

Table 3. Table showing the hyperparameters used with MCD.

set to $\max(4D, e, 64)$ with LeakyReLU activation functions, where e is the embedding dimension. We also use layer normalization and skip connections. The temperature for sampling the adjacency matrix from $q_\phi(\mathcal{M}_{1:K})$ using the Gumbel Softmax distribution was set to 0.25, and the temperature τ_r for sampling from the mixing rates variational distribution was set to 1. We used the same hyperparameters for both MCD-Nonlinear and MCD-Linear. Table 3 summarizes the hyperparameters used for training.

Baselines. Rhino was trained with similar hyperparameters as MCD on all datasets. For all other baselines, the default hyperparameter values are used. For Rhino and MCD, which parameterize the causal graphs as Bernoulli distributions over each edge, we use the inferred edge probability matrix as the "score", and evaluate the AUROC metric between the score matrix and the true adjacency matrix. For DYNOTEARS, we use the absolute value of the output scores and evaluate the AUROC. Since PCMCI+ and VARLiNGAM only output adjacency matrices, we directly evaluate the AUROC between the predicted and true adjacency matrices.

D.4. Post-processing the output of PCMCI$^+$

PCMCI$^+$ produces Markov equivalence classes rather than fully oriented causal graphs for the instantaneous adjacency matrix. To make its outputs comparable, we post-process the resultant edges. We follow the setup in Gong et al. (2022) and enumerate up to 3000 DAGs for the instantaneous matrix. We ignore the edges (i.e., set the corresponding entries in the adjacency matrix to 0) whose orientations are undecided. We compare all outputs against the ground truth during the evaluation and return the average metric across all enumerations.

D.5. Aggregating the temporal adjacency matrix across time

The Netsim and DREAM3 datasets used in the evaluation provide ground-truth time-aggregated causal graphs. In order to make our model output comparable, we follow the procedure outlined in (Gong et al., 2022) to convert the time-lag adjacency matrix to an aggregated matrix. The $(i, j)^{\text{th}}$ entry of the aggregated matrix $\mathcal{G}_{\text{agg}}$ is 1 iff $\mathcal{G}_\ell^{ij} = 1$ for some lag value ℓ in the time-lag matrix $\mathcal{G}$. Both Rhino and MCD represent the edges as Bernoulli random variables, hence output a probability score for each edge. For evaluating the F1 score, we threshold the probability values at 0.5, i.e., edges with a probability ≥ 0.5 are considered as predicted edges.

D.6. Pair-wise graph distance in the mixture distributions

Table 4 shows the pairwise graph distances between the ground-truth graphs of the mixture distributions used in the paper. We calculate the Structural Hamming Distance (SHD) between every pair of graphs in the mixture, and report the mean, standard deviation, minimum and maximum values.

E. Toy example

We provide a toy example to illustrate the importance of modeling the heterogeneity of a multi-modal dataset. Consider a dataset where each sample $X^{(n)}$ from the dataset $\left\{X_{1:T}^{1:D,(n)}\right\}_{n=1}^{N}$ is generated from one out of the two following SCMs,

Dataset	D	K^*	Avg. SHD	Std. dev. SHD	Min. SHD	Max. SHD
Synthetic (nonlinear)	5	5	21.00	2.57	15	24
Synthetic (nonlinear)	5	10	22.09	2.60	16	26
Synthetic (nonlinear)	5	20	22.19	2.92	14	32
Synthetic (nonlinear)	10	5	49.80	2.64	43	54
Synthetic (nonlinear)	10	10	53.42	3.03	47	59
Synthetic (nonlinear)	10	20	52.73	3.90	43	64
Synthetic (nonlinear)	20	5	120.20	2.14	117	124
Synthetic (nonlinear)	20	10	114.89	4.42	104	123
Synthetic (nonlinear)	20	20	114.28	5.13	101	127
Synthetic (linear)	5	5	21.40	3.14	16	27
Synthetic (linear)	5	10	22.18	2.81	16	28
Synthetic (linear)	5	20	22.07	3.16	12	30
Synthetic (linear)	10	5	48.60	3.75	41	54
Synthetic (linear)	10	10	54.16	3.53	45	60
Synthetic (linear)	10	20	54.52	4.14	42	67
Synthetic (linear)	20	5	114.20	2.23	111	118
Synthetic (linear)	20	10	111.58	5.17	103	124
Synthetic (linear)	20	20	113.42	4.82	101	126
Synthetic-Perturbed ($p = 0.005$)	10	5	2.40	0.92	1	4
Synthetic-Perturbed ($p = 0.008$)	10	5	5.20	0.98	4	7
Synthetic-Perturbed ($p = 0.01$)	10	5	7.00	0.89	5	8
Synthetic-Perturbed ($p = 0.05$)	10	5	21.40	5.37	15	30
Synthetic-Perturbed ($p = 0.1$)	10	5	48.20	2.64	45	53
Synthetic-Perturbed ($p = 0.005$)	10	10	2.44	1.04	1	5
Synthetic-Perturbed ($p = 0.008$)	10	10	3.56	1.36	1	6
Synthetic-Perturbed ($p = 0.01$)	10	10	6.11	2.51	1	12
Synthetic-Perturbed ($p = 0.05$)	10	10	23.04	4.17	15	32
Synthetic-Perturbed ($p = 0.1$)	10	10	47.56	8.56	29	65
DREAM3	100	5	517.60	202.13	234	896
Netsim-mixture	15	3	34.00	1.63	32	36
Netsim	5	14	2.59	1.17	1	5

Table 4. Pair-wise graph statistics for experimental datasets used in the paper.

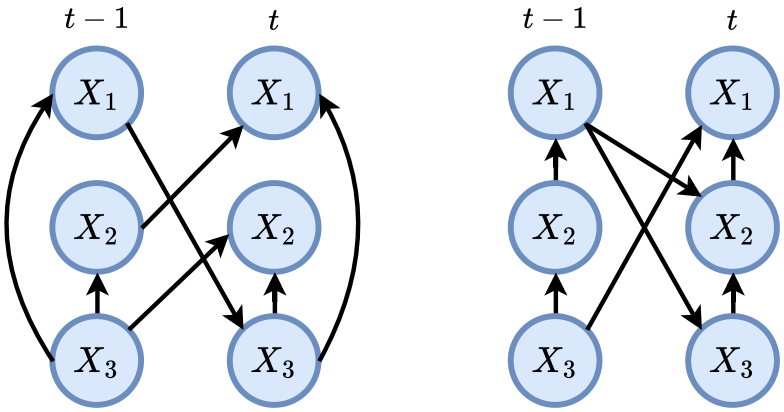

Figure 21. Temporal causal graphs which represent the causal relationships encoded by the SCMs.

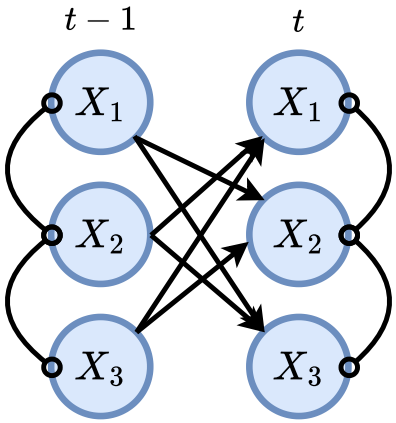

Figure 22. PCMCI$^+$ output on the toy-example. The algorithm infers a dense graph with many spurious causal relationships.

chosen with equal probability:

$$
\begin{aligned}
X_t^{1,(n)} &= 0.4X_{t-1}^{2,(n)} + 0.6X_t^{3,(n)} + \epsilon_1^{(n)} \\
X_t^{2,(n)} &= 0.3X_{t-1}^{3,(n)} + 0.3X_t^{3,(n)} + \epsilon_2^{(n)} \\
X_t^{3,(n)} &= 0.5X_{t-1}^{1,(n)} + \epsilon_3^{(n)}
\end{aligned}
$$

(or)

$$
\begin{aligned}
X_t^{1,(n)} &= 0.7X_{t-1}^{3,(n)} - 0.2X_t^{2,(n)} + \epsilon_1^{(n)} \\
X_t^{2,(n)} &= 0.2X_{t-1}^{1,(n)} + 0.4X_t^{3,(n)} + \epsilon_2^{(n)} \\
X_t^{3,(n)} &= -0.3X_{t-1}^{1,(n)} + \epsilon_3^{(n)}.
\end{aligned}
$$

These SCMs can be represented through the temporal causal graphs given in Figure 21.

However, if the graph membership of the samples is unknown, inferring a single causal graph to explain the causal relationships from the dataset would result in spurious causal relationships. For example, going by conditional independence tests, note that none of the nodes would be conditionally independent of each other for any conditioning set. This is also shown in the output of the PCMCI$^+$ algorithm, where a dense graph is inferred as shown in Figure 22. Thus, it is crucial to use a mixture distribution to model observational data from such heterogeneous distributions.