Causal Structure Discovery from Distributions Arising from Mixtures of DAGs

Basil Saeed [1] **Snigdha Panigrahi** [2] **Caroline Uhler** [1 3]

Abstract

We consider distributions arising from a mixture of causal models, where each model is represented by a directed acyclic graph (DAG). We provide a graphical representation of such mixture distributions and prove that this representation encodes the conditional independence relations of the mixture distribution. We then consider the problem of structure learning based on samples from such distributions. Since the mixing variable is latent, we consider causal structure discovery algorithms such as FCI that can deal with latent variables. We show that such algorithms recover a "union" of the component DAGs and can identify variables whose conditional distribution across the component DAGs vary. We demonstrate our results on synthetic and real data showing that the inferred graph identifies nodes that vary between the different mixture components. As an immediate application, we demonstrate how retrieval of this causal information can be used to cluster samples according to each mixture component.

1. INTRODUCTION

Determining causal structure from data is a central task in many applications. (Friedman et al., 2000; Heckerman et al., 1995) Causal structure is often modeled using a *directed acyclic graph* (DAG), where the nodes represent the variables of interest, and the directed edges represent the direct causal effects between these variables (Pearl, 2009). Assuming that the generating distribution of the data factors according to the DAG provides a way to relate the conditional independence relations in the distribution to separation statements in the DAG (known as *d-separation*) through the *Markov property* (Lauritzen, 1996). When not all variables of interest can be measured, DAGs are not sufficient to represent the observed distribution, since latent variables may introduce confounding effects between the observed variables. Instead, a family of mixed graphs known as *maximal ancestral graphs* (MAGs) can be used to model the observed variables by depicting the presence of latent confounders between pairs of variables through bidirected edges (Richardson and Spirtes, 2002).

With respect to learning the causal graph from data, the most ubiquitous methods infer d-separation relations by estimating conditional independence relations from the data; examples are the PC and GSP algorithms in the fully observed setting, and the FCI algorithm in the presence of latent variables (Spirtes et al., 2000; Solus et al., 2017; Zhang, 2008). These algorithms are consistent under the *faithfulness assumption*, which asserts that every conditional independence relation in the distribution corresponds to a d-separation relation in the graph. Note that even under faithfulness, the causal graph is in general not fully identifiable from observational data; it can in general only be identified up to its *Markov equivalence class* (Spirtes et al., 2000).

In various applications, data used for causal structure discovery is *heterogeneous* in that it stems from different causal models on the same set of variables (Gates and Molenaar, 2012; Chu et al., 2003; Ramsey et al., 2011). This is relevant for example in biomedical applications, where the goal is to learn a gene regulatory network based on gene expression data from a disease that consists of multiple not well characterized subtypes (as is the case for many neurological diseases). In such scenarios, the samples stem from a mixture of different causal models on the same set of variables, and the causal effects of the mixture distribution can in general not be faithfully represented by a single DAG.

Furthermore, a single DAG inferred from such samples cannot identify differences between the component DAGs in the mixture, which may be critical for personalized biomedical interventions, and may lead to flawed conclusions downstream.

In this work, we consider distributions arising as mixtures

[1]Laboratory for Information and Decision Systems and Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, MA, USA [2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA [3]Department of Biosystems Science and Engineering, ETH Zurich, Switzerland. Correspondence to: Caroline Uhler <cuhler@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

of causal DAGs. Our main contributions are as follows:

- We introduce the *mixture graph* to represent such mixture distributions. We prove that this graph encodes the conditional independence relations in the mixture distribution through separation statements (Theorem 3.2) and show that the separation statements in every such graph can be realized by independence relations in some mixture distribution (Proposition 3.5).
- We introduce the *union graph*, a graph defined from the mixture graph. We prove that, under a faithfulness and ordering assumption on the DAGs in the mixture, the FCI algorithm applied to data from a mixture of DAGs outputs the union graph (Theorem 4.4).
- We prove that the union graph can be used to identify variables whose conditional distribution across the component DAGs changes (Proposition 4.6). We demonstrate the implication of this result for identifying critical nodes and for clustering samples according to their mixture component on synthetic data and data from genomics.

2. PRELIMINARIES & RELATED WORK

2.1. Graphical representations: DAGs and MAGs

In this paper, we consider two types of graphs: *directed acyclic graphs* (*DAGs*) and *mixed graphs* with directed ($\to$) and bidirected ($\leftrightarrow$) edges. We denote the former by $\mathcal{D} = (V, E)$ and the latter by $\mathcal{M} = (V, D, B)$, where V denotes the set of vertices, E and D denote the set of directed edges and B denotes the set of bidirected edges. A mixed graph is said to be *ancestral* if it has no directed cycles, and whenever there is a bidirected edge $u \leftrightarrow v$, then there is no directed path from u to v (Richardson and Spirtes, 2002). While ancestral graphs have been defined more generally to allow also for undirected edges, in this work we will only make use of graphs with directed and bidirected edges.

Throughout, we will use the notation $\mathrm{ch}_{\mathcal{M}}(v)$, $\mathrm{pa}_{\mathcal{M}}(v)$ and $\mathrm{an}_{\mathcal{M}}(v)$ to denote the children, parents and ancestors, respectively, of a node v in the graph $\mathcal{M}$. Furthermore, we use the standard definitions of *path* and *directed path* in a graph; for these definitions, see e.g. Lauritzen (1996). We will use the notation $v \leftrightarrow_{\mathcal{M}} u$ as a shorthand to denote "the edge $v \leftrightarrow u$ between nodes u, v in $\mathcal{M}$", and use similar notations for other types of edges.

The notion of d-separation from DAGs can be generalized to ancestral graphs by accounting for the new possible ways to obtain a collider from bidirected edges (Richardson and Spirtes, 2002). In ancestral graphs, unlike in DAGs, it is possible to have a pair of nodes that are not adjacent, but cannot be d-separated given any subset of nodes. An ancestral graph where any non-adjacent pair of nodes is d-separated given some subset of nodes is called *maximal*, and a non-maximal ancestral graph can be made maximal by adding a bidirected edge between all such pairs. An ancestral graph that is maximal is called a *Maximal Ancestral Graph* (MAG) (Richardson and Spirtes, 2002).

Ancestral graphs are a useful representation of DAGs with unobserved nodes.

Specifically, Richardson and Spirtes (2002) showed that given a DAG $\mathcal{D} = (V \cup L, E)$, with observed nodes V and unobserved nodes L, satisfying a set of d-separation statements of the form "A d-separated from B given C" for disjoint $A, B, C \subseteq V$, there exists an ancestral graph $\mathcal{M} = (V, D, B)$ with the same d-separation statements, called the *marginal ancestral graph of $\mathcal{D}$ with respect to L*. Sadeghi et al. (2013) gave a local criterion to construct this graph from $\mathcal{D}$. Throughout our paper, we will make use of this in the special case where L consists of a single node of in-degree 0. The specialization of Sadeghi's algorithm to this case is provided in Algorithm 1.

Algorithm 1: Construction of the marginal ancestral graph

Input: DAG $\mathcal{D} = (V \cup \{y\}, E)$, where y has in-degree 0.
Output: the marginal ancestral graph of $\mathcal{D}$ w.r.t. y.

(0) Initialize $D = \emptyset$, $B = \emptyset$
(1) For $u, v \in \mathrm{ch}_{\mathcal{D}}(y)$: add $u \leftrightarrow v$ to B.
(2) For t, u, v such that $(t \to u) \in E$ and $(u \leftrightarrow v) \in B$: if $u \in \mathrm{an}_{\mathcal{D}}(v)$, then add $t \to v$ to D.
(3) For u, v such that $u \leftrightarrow v \in B$: if $u \in \mathrm{an}_{\mathcal{D}}(v)$, then remove $u \leftrightarrow v$ from B and add $u \to v$ to D.
(4) Return the ancestral graph $\mathcal{M} = (V, D, B)$.

Although, in general, the ancestral graph constructed using Sadeghi's criterion is not maximal, the relevant restriction considered here, i.e., when L consists of a single node with in-degree 0, is always a MAG. The following proposition states this; a proof is provided in section A of the Appendix

Proposition 2.1. *The output of Algorithm 1 is a MAG.*

2.2. Markov Properties

Given a graph $\mathcal{M}$ with nodes V, we associate to each node $v \in V$ a random variable X_v and denote the joint distribution of $X_V := (x_v : v \in V)$ by p_{X_V}. The *Markov property* associates missing edges in $\mathcal{M}$ with conditional independence statements in p_{X_V}: a distribution p_{X_V} is said to satisfy the Markov property with respect to $\mathcal{M}$ if for any disjoint $A, B, C \subseteq V$ such that A and B are d-separated given C in $\mathcal{M}$, it holds that $X_A \perp\!\!\!\perp X_B \mid X_C$ in p_{X_V} (Lauritzen, 1996). For DAGs, an equivalent condition to the Markov property is for p_{X_V} to factorize as $p_{X_V}(x_V) = \prod_{v \in V} p(x_v | x_{\mathrm{pa}_{\mathcal{G}}(v)})$; see Lauritzen (1996).

Considering latent variables X_L, Richardson and Spirtes (2002) showed that given a distribution p_{X_V,X_L} that is Markov with respect to a DAG $\mathcal{D}$ over $V \cup L$, the marginal $p_{X_V}(x_V) = \sum_{x_L} p_{X_V,X_L}(x_V, x_L)$ is Markov with respect to the marginal ancestral graph of $\mathcal{D}$ with respect to L.

It is possible for two different DAGs $\mathcal{D}_1, \mathcal{D}_2$ over the same set of nodes to satisfy the same set of d-separation statements. In this case, $\mathcal{D}_1$ and $\mathcal{D}_2$ are said to be *Markov equivalent*, and the set of all DAGs that are Markov equivalent to a DAG $\mathcal{D}$ is called the *Markov equivalence Class of* $\mathcal{D}$. These definitions trivially extend to MAGs. The Markov equivalence class of a MAG can be represented by a *partial ancestral graph* (*PAG*): the edges in such a graph have three types of tips: arrowheads ($\leftarrow$), tails ($-$) and circles $\circ\!\!-$, where arrowhead (tail) signifies that this arrowhead exists in all graphs in the Markov equivalence class (Zhang, 2008).

2.3. Causal Structure Discovery

The goal of structure learning is to recover the graph $\mathcal{D}$ or $\mathcal{M}$ from data generated from the distribution p_{X_V}. This task often requires assumptions beyond the Markov property. One common such assumption is the so-called *faithfulness assumption* which states that for any disjoint $A, B, C \subseteq V$, it holds that A and B are d-separated given C whenever $X_A \perp\!\!\!\perp X_B \mid X_C$ in p_{X_V} (Spirtes et al., 2000). The faithfulness assumption allows making inference about the structure of $\mathcal{D}$ or $\mathcal{M}$ from conditional independence tests on the data. Various algorithms have been proposed for this task that are provably consistent, such as the PC, GES or GSP algorithms for learning DAGs (Spirtes et al., 2000; Chickering, 2002; Solus et al., 2017), and the FCI algorithm for learning MAGs (Spirtes et al., 2000). Note that even under the faithfulness assumption, it is in general only possible to retrieve the Markov equivalence class of a graph $\mathcal{D}$ or $\mathcal{M}$ from data; this is the output of the above algorithms. For example, FCI in general does not return a specific MAG, but a PAG representing a Markov equivalence class of MAGs.

2.4. Causal Inference from Mixtures of DAGs

While the problem of learning appropriate representations from data of DAG mixtures arises in various applications, little work has been done on theory and methodology in this direction. Spirtes (1994) investigated the conditional independence properties of such mixture distributions; he defined a cyclic graphical model derivable from the component DAGs and proved that the mixture distribution is Markov with respect to it. However, this graph does not capture the full set of conditional independence relations for any reasonable mixture. In fact, as we discuss later, this graph is similar to the representation we define in Section 4, which also only provides partial information about the structure of the component DAGs. To capture the full set of independences in the mixture distribution, a representation sparser than that of Spirtes (1994) is necessary. Strobl (2019a;b) built on this work to define a sparser graph. However, we provide examples in Section B of the Appendix showing that the Markov condition in general does not hold for this graph, i.e., there can be d-separation statements in the graph that do not correspond to conditional independence relations in the mixture distribution. Finally, Ramsey et al. (2011) provided conditions for the mixture distribution to be representable by a graph that is a union of the component DAGs.

To learn the component DAGs from mixture data, a simple approach is to cluster the data using, for example, the Expectation-Maximization (EM) algorithm and then learn a DAG from each cluster. This, however, uses a reduced sample size to learn each DAG (corresponding to the size of the associated cluster). In the case where the cluster labels are known and the DAGs are related, Wang et al. (2020) showed that learning each DAG separately can lead to loss in accuracy compared to when the full sample size is used to learn the DAGs jointly. When the expectation in the EM algorithm can be computed, as e.g. for Gaussians, Thiesson et al. (1997) proposed a heuristic approach based on the EM algorithm to directly learn the component DAGs from the mixture data. In this work, we consider a different problem. Instead of learning the component DAGs we provide a graphical representation of the mixture distribution and identify critical aspects of the component DAGs that are captured by this graph and can be identified by algorithms such as FCI when applied directly to the mixture distribution.

3. MIXTURE DAG AND MARKOV PROPERTY

In this section, we provide our first main result: after formally introducing distributions that arise as mixtures of DAGs, we define the *mixture DAG* and prove in Theorem 3.2 and Proposition 3.5 that it is a valid representation of the model, i.e., the DAG encodes the conditional independence relations of the mixture distributions. More precisely, not only is the Markov condition satisfied (i.e., all separation statements in the mixture DAG correspond to conditional independence relations in the mixture distribution), but in addition, every mixture DAG is also realizable by a mixture distribution (meaning that the mixture DAG cannot be made sparser without losing the Markov property).

3.1. Mixture of Causal DAGs

To introduce the mixture model, we consider K DAGs $\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}\}$ with $\mathcal{D}^{(j)} = (V, E^{(j)})$ for $1 \leq j \leq K$, i.e., these K DAGs are defined on the *same* set of nodes.

Associated with each component DAG $\mathcal{D}^{(j)}$ is a random vec-

tor X_V with distribution $p^{(j)}(x_V)$. Let V_{INV} denote the set of nodes that are *invariant* across the K component DAGs, i.e., nodes whose conditional distribution in the factorization does not vary across $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$; that is

$$V_{\text{INV}} = \Big\{ v \in V \colon p^{(j)}(x_v | x_{\text{pa}_{\mathcal{D}^{(j)}}(v)}) = p^{(k)}(x_v | x_{\text{pa}_{\mathcal{D}^{(k)}}(v)}) \quad \text{for all } j, k \in \{1, 2, \cdots, K\} \Big\}. \tag{1}$$

Assuming that each distribution $p^{(j)}(x_V)$ admits a factorization according to DAG $\mathcal{D}^{(j)}$, we then obtain:

$$\begin{aligned} p^{(j)}(x_V) &= \prod_{v \in V \setminus V_{\text{INV}}} p^{(j)}(x_v | x_{\text{pa}_{\mathcal{D}^{(j)}}(v)}) \prod_{v \in V_{\text{INV}}} p^{(j)}(x_v | x_{\text{pa}_{\mathcal{D}^{(j)}}(v)}) \\ &= \prod_{v \in V \setminus V_{\text{INV}}} p^{(j)}(x_v | x_{\text{pa}_{\mathcal{D}^{(j)}}(v)}) \prod_{v \in V_{\text{INV}}} p^{(1)}(x_v | x_{\text{pa}_{\mathcal{D}^{(1)}}(v)}) \end{aligned}$$

for all $1 \leq j \leq K$, i.e., each distribution decouples into two components: one over the variables associated with V_{INV} that remains constant across all K distributions, and another over the remaining variables which may differ with j.

Let J be a discrete variable taking values in $\{1, \ldots, K\}$ with probabilities $p_J(j)$ for each $j \in \{1, \ldots, K\}$. Defining a joint distribution p_μ over $X_V \cup J$ by

$$p_\mu(x_V, j) := p_J(j) \cdot p^{(j)}(x_V), \tag{2}$$

this joint distribution satisfies $p^{(j)}(x_V) = p_\mu(x_V | J = j)$ and the observed mixture distribution is obtained by marginalizing p_μ over the unobserved index variable J. With a slight abuse of notation, we denote the resulting mixture distribution also by p_μ. Given samples from this distribution, i.e., without knowledge of the membership of each sample to its generating DAG, we analyze what can still be inferred regarding the structure of $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$.

3.2. Mixture DAG and Markov Property

We now present the *mixture DAG*, a DAG that is representative of the independence relations induced amongst the observed variables after marginalizing over the index variable J in (2). Denoting the number of vertices in V by $|V|$, the mixture DAG is a graph on $K \cdot |V| + 1$ nodes constructed by placing the K component DAGs next to each other, giving rise to a DAG on $K \cdot |V|$ nodes, and using an additional node to represent J. We now provide the precise definition.

Definition 3.1 (Mixture DAG)**.** *Let $v^{(j)}$ denote vertex v in DAG j and let $[V] := \cup_{1 \leq j \leq K} V^{(j)}$ denote the vertices of the K component DAGs. The* mixture DAG*, denoted by $\mathcal{D}_\mu$, has nodes $[V] \cup \{y\}$ and edges E_μ consisting of edges in each component DAG, namely*

$$\bigcup_{j=1}^{K} \Big\{ v^{(j)} \to \tilde{v}^{(j)} : v, \tilde{v} \in V,\ v \to \tilde{v} \in E^{(j)} \Big\},$$

and additional edges from node y to some nodes in $[V]$, namely those corresponding to variables that have conditionals that are not the same for all j, i.e.,

$$\bigcup_{j=1}^{K} \Big\{ y \to v^{(j)} : v \in V \setminus V_{\text{INV}} \Big\}.$$

Figure 1 provides an example of the mixture DAG arising from a mixture with $K = 2$ and $|V| = 4$. Note that, while the results of this section hold even when

the DAGs $\mathcal{D}^{(j)}$ have no common topological ordering (meaning that there exists no ordering π such that $v < u$ in π only if $u \notin \text{an}_{\mathcal{D}^{(j)}}(v)$ for all $1 \leq j \leq K$), the mixture DAG is sparsest, and hence provides information about the component DAGs through separation statements, when a common topological ordering exists (as in Figure 1). When there is no common ordering, the set V_{INV} is generally smaller, since $\text{pa}_{\mathcal{D}^{(j)}} \neq \text{pa}_{\mathcal{D}^{(k)}}$ implies $p^{(j)}(x_v | x_{\text{pa}_{\mathcal{D}^{(j)}}(v)}) \neq p^{(k)}(x_v | x_{\text{pa}_{\mathcal{D}^{(k)}}(v)})$, which implies a denser mixture DAG.

We emphasize here that the DAG in Definition 3.1 is not a graphical model representation of the mixture distribution in the standard sense. This is already clear from the fact that the mixture DAG has $K \cdot |V| + 1$ nodes, whereas the mixture distribution is only $|V|$-dimensional. Yet, in the following theorem we show that it is possible to read off conditional independence relations that hold in the mixture distribution p_μ from the mixture graph in an intuitive manner.

For $A \subset V$, we use the notation $[A]$ to denote all K copies of the nodes in A, i.e., $A = \cup_{1 \leq j \leq K} A^{(j)}$.

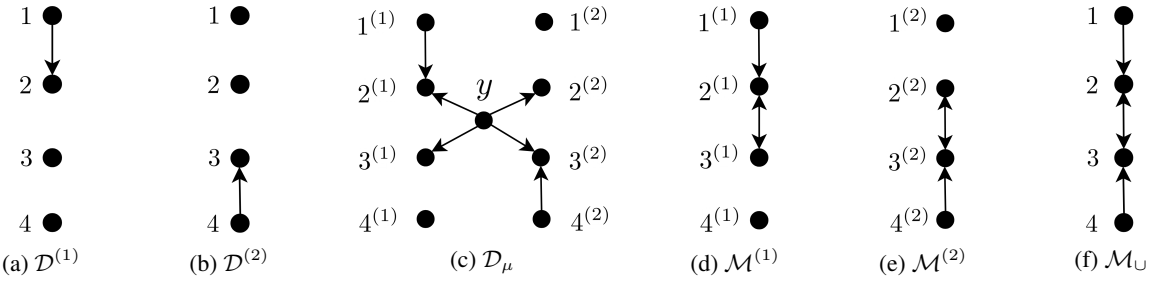
Theorem 3.2 (Markov Property)**.** *Let $A, B, C \subseteq V$ be disjoint. If $[A]$ and $[B]$ are d-separated given $[C]$ in the mixture DAG $\mathcal{D}_\mu$, then $X_A \perp\!\!\!\perp X_B | X_C$ in the mixture distribution p_μ.*

To illustrate this result, consider the example in Figure 2. Since $[1] = \{1^{(1)}, 1^{(2)}\}$ and $[4] = \{4^{(2)}, 4^{(2)}\}$ are d-separated given $\emptyset$ in the mixture DAG, then the mixture distribution $p_\mu(x_1, x_2, x_3, x_4)$ satisfies $X_1 \perp\!\!\!\perp X_4$.

We note that while the graphical representation provided by Strobl (2019b) (the *mother graph*) is similar to the mixture DAG, it critically differs in how the component DAGs are connected via the node y. Importantly, we show in Section B in the Appendix that the mixture distribution p_μ is *not* Markov with respect to the mother graph[1].

In the following, we provide a proof for Theorem 3.2. For each $1 \leq j \leq K$, let $\widetilde{\mathcal{D}}^{(j)}$ be the sub-DAG induced by $\mathcal{D}_\mu$ on the vertices $V^{(j)} \cup \{y\}$.The main ingredient of the proof is the following lemma, which connects d-separation statements in the mixture DAG to conditional independence

[1] Strobl (2019a;b) provides two different constructions; we show that the Markov property does not hold in either.



(a) $\mathcal{D}^{(1)}$ (b) $\mathcal{D}^{(2)}$ (c) $\mathcal{D}_\mu$ (d) $\mathcal{M}^{(1)}$ (e) $\mathcal{M}^{(2)}$ (f) $\mathcal{M}_\cup$

Figure 1: (a)-(b): component DAGs for a mixture model with $K = 2$; (c): corresponding mixture DAG (see Definition 3.1); (d)-(e): associated component MAGs (see Section 4); (f): associated union graph (see Definition 4.2).

relations in the mixture distribution via d-separation in $\widetilde{D}^{(j)}$.

Lemma 3.3. *Let $A, B, C \subseteq V$ be disjoint. If for all $1 \leq j \leq K$ it holds that*

(a) $A^{(j)}$ and $B^{(j)}$ are d-separated given $C^{(j)}$, and;

(b) $A^{(j)}$ and y are d-separated given $C^{(j)}$ in $\widetilde{\mathcal{D}}^{(j)}$,

then $X_A \perp\!\!\!\perp J \mid X_C$ in p_μ, implying the factorization

$$p^{(j)}(x_A, x_B|x_C) = p^{(1)}(x_A|x_C)p^{(j)}(x_B|x_C)$$

for all $1 \leq j \leq K$.

We now provide the proof for Theorem 3.2.

Proof of Theorem 3.2. We start by showing that the conditions of Lemma 3.3 are satisfied. First, note that $[A]$ and $[B]$ are d-separated given $[C]$ in $\mathcal{D}_\mu$ implies that $A^{(j)}$ and $B^{(j)}$ are d-separated given $C^{(j)}$ in $\mathcal{D}^{(j)}$ for all $1 \leq j \leq K$. Second, note that since y has in-degree 0, we cannot have both a d-connecting path given $[C]$ between $[A]$ and y and one between $[B]$ and y in $\mathcal{D}_\mu$. Hence, we may assume without loss of generality that $[A]$ and y are d-separated given $[C]$ (otherwise, $[B]$ and y are d-separated given $[C]$).

We now use Lemma 3.3 to show that $p_\mu(x_A, x_B|x_C)$ factorizes as $f_A(x_A, x_C)f_B(x_B, x_C)$, which would prove that $X_A \perp\!\!\!\perp X_B|X_C$ in p_μ. By definition of p_μ in (2),

$$p_\mu(x_A, x_B|x_C) = \sum_{j=1}^{K} p^{(j)}(x_A, x_B|x_C)p_J(j),$$

and hence as a consequence of Lemma 3.3 we obtain

$$\begin{aligned} p_\mu(x_A, x_B|x_C) &= \sum_{j=1}^{K} p^{(1)}(x_A|x_C)p^{(j)}(x_B|x_C)p_J(j) \\ &= p^{(1)}(x_A|x_C) \sum_{j=1}^{K} p^{(j)}(x_B|x_C)p_J(j), \end{aligned}$$

providing a factorization of the desired form.

□

In Theorem 3.2, we established that every separation statement in the mixture DAG $\mathcal{D}_\mu$ corresponds to a conditional independence relation in the mixture distribution p_μ. Next, we show that every mixture DAG is *realizable*, i.e., that for any mixture DAG $\mathcal{D}_\mu$, there exists a p_μ whose conditional independence relations are faithfully represented by the separation statements of $\mathcal{D}_\mu$. This implies that $\mathcal{D}_\mu$ is the "correct" graphical representation of a mixture of DAGs and cannot be made sparser without losing the Markov property.

3.3. Faithfulness

We define faithfulness of a mixture distribution p_μ with respect to a mixture DAG $\mathcal{D}_\mu$ analogously to how faithfulness is defined for a distribution with respect to a DAG model.

Definition 3.4 (Mixture Faithfulness)**.** *The mixture distribution p_μ is faithful with respect to a mixture DAG $\mathcal{D}_\mu$ if for any disjoint $A, B, C \subseteq V$ with $X_A \perp\!\!\!\perp X_B|X_C$ in p_μ it holds that $[A]$ and $[B]$ are d-separated given $[C]$.*

We next provide an example showing that mixture faithfulness is not implied by faithfulness of each component distribution $p^{(j)}$ with respect to the corresponding DAG $\mathcal{D}^{(j)}$. Hence, to establish realizability of the mixture graph, it is not sufficient to rely on the fact that for every DAG $\mathcal{D}^{(j)}$, there exists a distribution $p^{(j)}$ that is faithful to it.

Example 1. *Consider the distributions $p^{(1)}(x_V), p^{(2)}(x_V)$ on $V = \{1, 2, 3, 4\}$ that factor according to the DAGs $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$, respectively, shown in Figure 1. Namely*

$$\begin{aligned} p^{(1)}(x_V) &= p^{(1)}(x_1)p^{(1)}(x_2|x_1)p^{(1)}(x_3)p^{(1)}(x_4), \\ p^{(2)}(x_V) &= p^{(2)}(x_1)p^{(2)}(x_2)p^{(2)}(x_3|x_4)p^{(2)}(x_4), \end{aligned}$$

where

$$\begin{aligned} &p^{(1)}(x_1) = \mathcal{N}(x_1; 0, 1), && p^{(2)}(x_1) = \mathcal{N}(x_1; 0, 1), \\ &p^{(1)}(x_2|x_1) = \mathcal{N}(x_2; x_1, 1), && p^{(2)}(x_2) = \mathcal{N}(x_2; 0, 2), \\ &p^{(1)}(x_3) = \mathcal{N}(x_3; 0, 1), && p^{(2)}(x_3|x_4) = \mathcal{N}(x_3; x_4, 1), \\ &p^{(1)}(x_4) = \mathcal{N}(x_4; 0, 1), && p^{(2)}(x_4) = \mathcal{N}(x_4; 0, 1). \end{aligned}$$

Then, defining $p_\mu(x_V) := \sum_{j=1}^{2} p^{(j)}(x_V)p_J(j)$, *for some* $J \sim p_J(j)$, *we obtain that*

$$\begin{aligned}
p_\mu(x_2, x_3) &= \int p_\mu(x_V)dx_1dx_4 \\
&= \int p_J(1)p^{(1)}(x_1)p^{(1)}(x_2|x_1)p^{(1)}(x_3)p^{(1)}(x_4)dx_1dx_4 \\
&\quad + \int p_J(2)p^{(2)}(x_1)p^{(2)}(x_2)p^{(2)}(x_3|x_4)p^{(2)}(x_4)dx_1dx_4 \\
&= p_J(1)\,\mathcal{N}(x_2; 0, 2)\,\mathcal{N}(x_3; 0, 1) \\
&\quad + p_J(2)\,\mathcal{N}(x_2; 0, 2)\,\mathcal{N}(x_3; 0, 2) \\
&= \mathcal{N}(x_2; 0, 2)\Big(p_J(1)\mathcal{N}(x_3; 0, 1) + p_J(2)\mathcal{N}(x_3; 0, 2)\Big) \\
&= f(x_2)g(x_3),
\end{aligned}$$

which implies that $X_2 \perp\!\!\!\perp X_3$ *in* p_μ, *although in the mixture DAG corresponding to* p_μ *shown in Figure 1 the nodes* 2 *and* 3 *are d-connected via the path through* y. □

This example was carefully crafted; even a slight perturbation such as choosing $p^{(2)}(x_2) = \mathcal{N}(x_2; 0, 2.001)$ would have meant that $p_\mu(x_2, x_3)$ does not factor, indicating that mixture-faithfulness violations are rare. More precisely, consider the family of Gaussian mixture models where each $p^{(j)}$ is a Gaussian distribution that is faithful with respect to $\mathcal{D}^{(j)}$. A violation of mixture-faithfulness occurs if and only if $\sum_j p^{(j)}(x_A, x_B|x_C)$ factors as $p_\mu(x_A|x_C)p_\mu(x_B|x_C)$, i.e.,

$$\sum_j p^{(j)}(x_A, x_B|x_C) = \sum_i p^{(i)}(x_A|x_C) \sum_j p^{(j)}(x_B|x_C),$$

when $[A]$ and $[B]$ are d-connected given $[C]$ in $\mathcal{D}_\mu$. This represents an equality constraint on the parameters of the Gaussians $p^{(j)}$ for $1 \le j \le K$. As a consequence, mixture-faithfulness holds almost surely and any $\mathcal{D}_\mu$ is realizable by a mixture of Gaussians, thereby proving the following.

Proposition 3.5 (Realizability of $\mathcal{D}_\mu$)**.** *For any mixture DAG* $\mathcal{D}_\mu$, *there exists a mixture distribution* p_μ *that is faithful with respect to* $\mathcal{D}_\mu$.

4. LEARNING FROM MIXTURE DATA

Without knowing the membership of each sample to a component DAG, we cannot generally learn the structure of $\mathcal{D}^{(j)}$ for each j from the data. Since the mixing variable is latent, an intuitive approach is to apply FCI to learn a MAG representation of p_μ. In this section, we will characterize the output of FCI. In particular, we will show that FCI identifies critical nodes in the component DAGs: those whose conditionals across the component DAGs vary.

A difficulty for structure discovery using MAG-based learning algorithms such as FCI, is that even under the mixture-faithfulness assumption the conditional independence relations in a mixture distribution p_μ may not be representable by any MAG. We illustrate this in the following example and then provide conditions to avoid this phenomenon.

Example 2. *Consider* $\mathcal{D}_\mu$ *shown in Figure 2a. We show that there does not exist any MAG* $\widetilde{\mathcal{M}}$ *over the variables* $V = \{1, \dots, 5\}$ *that satisfies:* A *d-sep from* B *given* C *in* $\widetilde{\mathcal{M}}$ *if and only if* $[A]$ *d-sep from* $[B]$ *given* $[C]$ *in* $\mathcal{D}_\mu$. *First, note that such a MAG would need to have the same skeleton as the graph in Figure 2b to respect the adjacencies in* $\mathcal{M}_\mu$. *Otherwise it would have an extra or missing d-separation with no analog in* $\mathcal{M}_\mu$. *In addition,* $\widetilde{\mathcal{M}}$ *would also need to contain the colliders* $4 \to 5 \leftarrow 2$ *and* $1 \to 2 \leftarrow 5$ *to respect the d-separation relations resulting from* $4^{(2)} \to 5^{(2)} \leftarrow y \to 2^{(2)}$ *and* $1^{(1)} \to 2^{(1)} \leftarrow y \to 5^{(1)}$ *respectively. This implies the existence of* $2 \leftrightarrow_{\widetilde{\mathcal{M}}} 5$. *Further note that conditioning on either* $[2], [3]$ *or* $[4]$ *(or any subset of these) connects* $[5]$ *and* $[1]$ *in* $\mathcal{D}_\mu$ *which are d-separated given* $\emptyset$. *The only orientation of arrowheads compatible with both the skeleton and these separation/connection relations is* $2 \to 3 \to 4$. *Hence,* $4 \in \mathrm{de}_{\widetilde{\mathcal{M}}}(2)$. *Finally, the existence of an arrowhead* $4 \leftarrow\!* 5$ *would violate the separation:* $[5]$ *d-separated from* $[1]$ *given* $\emptyset$. *Hence,* $2 \leftrightarrow_{\widetilde{\mathcal{M}}} 5$ *and* $2 \in \mathrm{an}_{\widetilde{\mathcal{M}}}(5)$, *violating the ancestral property.* □

We now identify a class of mixture models for which the d-separations in the mixture DAG are equivalent to d-separation statements in a MAG.

Definition 4.1. *Let* $\mathcal{M}^{(j)}$ *be the MAG constructed via Algorithm 1 from the induced sub-DAG* $\widetilde{\mathcal{D}}^{(j)}$ *defined in Section 3.2. The MAGs* $\mathcal{M}^{(1)}, \dots, \mathcal{M}^{(K)}$ *are said to be* compatible with the same poset *if there exists a partial order* π *on* V *such that for all* $1 \le j \le K$ *it holds that* (a) $u \in \mathrm{an}_{\mathcal{M}^{(j)}}(v) \Rightarrow u <_\pi v$; *and* (b) $u \leftrightarrow_{\mathcal{M}^{(j)}} v \Rightarrow u \not\lessgtr_\pi v$.

Figures 1d and 1e show examples of MAGs $\mathcal{M}^{(j)}$ that satisfy this poset compatibility condition. One can further check that the MAGs $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ associated with the mixture DAG in Figure 2a do not satisfy this condition. This example shows that there exist DAGs $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(K)}$ with a common topological ordering whose corresponding MAGs $\mathcal{M}^{(1)}, \dots, \mathcal{M}^{(K)}$ do not satisfy the poset compatibility condition 4.1. On the other hand, it can be readily verified that the compatibility assumption on $\mathcal{M}^{(1)}, \dots, \mathcal{M}^{(K)}$ implies that $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(K)}$ have a common topological ordering.

In the following, we show that poset compatibility ensures that d-separation relations in $\mathcal{D}_\mu$ are representable by a MAG, which we call the *union graph* since it is obtained as a union of the edges of $\mathcal{M}^{(1)} \dots, \mathcal{M}^{(K)}$.

Definition 4.2 (Union Graph)**.** *The* union graph $\mathcal{M}_\cup := (V, D_\cup, B_\cup)$ *has vertices* V, *directed edges*

$$D_\cup = \{v \to u : u, v \in V,\ \exists_j v^j \to_{\mathcal{M}^{(j)}} u^j\},$$

and bidirected edges

$$B_\cup = \{v \leftrightarrow u : v, u \in V,\ \exists_j v^j \leftrightarrow_{\mathcal{M}^{(j)}} u^j\}.$$

We remark that Spirtes (1994) studied a similar graph and proved the Markov property for a DAG with vertices $V \cup \{y\}$ and directed edges given by the union of $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$.

An example of a union graph $\mathcal{M}_\cup$ is given in Figure 1f. In general, $\mathcal{M}_\cup$ may neither be maximal nor ancestral (see Figure 2b for an example). However, the following lemma states that under poset compatibility it is guaranteed to be both. The proof is given in Section D of the Appendix,

Lemma 4.3. *Under the assumption that $\mathcal{M}^{(1)} \ldots, \mathcal{M}^{(K)}$ are compatible with the same poset, $\mathcal{M}_\cup$ is a MAG.*

We now state the main results of this section, characterizing the output of FCI when run on mixtures of DAGs.

Theorem 4.4. *Let $A, B, C \subseteq V$ be disjoint. If the component MAGs satisfy the poset compatibility assumption, then A and B are d-separated given C in $\mathcal{M}_\cup$ if and only if $[A]$ and $[B]$ are d-separated given $[C]$ in $\mathcal{D}_\mu$.*

The proof is provided in Section E in the Appendix. The following corollary follows directly from the asymptotic consistency of FCI (Spirtes et al., 2000).

Corollary 4.5. *If the distribution p_μ is faithful with respect to a mixture DAG whose component MAGs satisfy the poset compatibility assumption 4.1, then FCI outputs the Markov equivalence class of the corresponding union MAG $\mathcal{M}_\cup$.*

We end this section by pointing out an important structural property of $\mathcal{M}_\cup$, which can be used to recover key information about the component distributions in the mixture. We leave the proof to Section F of the Appendix.

Proposition 4.6. *A bidirected edge $u \leftrightarrow v$ in the union graph $\mathcal{M}_\cup$ implies that $u \in V \setminus V_{INV}$. Additionally, this implies that $p^{(j)}(x_u | x_{\mathrm{pa}_{\mathcal{D}^{(j)}}(u)}) \neq p^{(i)}(x_u | x_{\mathrm{pa}_{\mathcal{D}^{(i)}}(u)})$.*

Hence bidirected edges identify nodes in the component DAGs whose conditional distribution varies across mixture components. As we show in the following section, these nodes are natural candidates for features when clustering.

5. EXPERIMENTS

5.1. Synthetic Data

In the following, we demonstrate the effectiveness of learning the union graph from mixture data, analyze the performance when estimating $V \setminus V_{INV}$ using Proposition 4.6, and investigate the performance of clustering using mixture data when $V \setminus V_{INV}$ are used as features.

We generated K component DAGs each with $|V| = 10$ nodes and the same topological ordering from an Erdös-Rényi model with expected degree $d = 1.5/K$ so that the nodes in the $\mathcal{M}_\cup$ have expected degree less than 1.5. From these DAGs, the corresponding MAGs $\mathcal{M}^{(j)}$ were computed using Algorithm 1. If the MAGs were not compatible with the same poset, the DAGs were discarded to ensure poset-compatibility (2 out of 270 graphs were discarded).

Data was sampled from each DAG based on a linear structural equation model with additive Gaussian noise, where each edge weight (u, v) was sampled uniformly in $[-2, -0.25] \cup [0.25, 2]$ (to ensure that it was bounded away from zero) and set to be equal for the edges $(u^{(j)}, v^{(j)})$ for all $1 \leq j \leq K$ if this edge existed in DAG $\mathcal{D}^{(j)}$. In this case, $v \in V_{INV}$ if and only if the parents of X_v are the same across all K DAGs. The mean for the Gaussian noise was sampled uniformly in $[-2, 2]$ with standard deviation 1. From each DAG $\mathcal{D}^{(j)}$, we generated $n\, p_j$ observations where $\sum_{j=1}^{K} p_j = 1$ yielding a total of n samples. For the plots in the main paper, we chose $p_j = 1/K$. We present additional plots in Appendix G for when $p = (p_k : 1 \leq k \leq K)$ is sampled from a Dirichlet distribution.

Learning the Union MAG. To evaluate Corollary 4.5, we ran the R implementation of FCI from the `pcalg` library on this synthetic data using Gaussian conditional independence tests (despite the true distribution being a mixture of Gaussians) with threshold α. The output is a PAG $\widehat{\mathcal{P}}_\cup$ representing the Markov equivalence class of the union graph. As comparison, we computed the true union graph $\mathcal{M}_\cup$ based on the MAGs $\mathcal{M}^{(j)}$, generated n samples from this graph (using a structural equation model with the same parameters as in the mixture) and ran FCI on these samples to obtain an estimate $\widetilde{\mathcal{P}}_\cup$ for the PAG of the union graph. This offsets the estimation errors that are intrinsic to FCI. The difference between the PAGs $\widehat{\mathcal{P}}_\cup$ and $\widetilde{\mathcal{P}}_\cup$ was measured via a *normalized structural Hamming distance*; the structural Hamming distance (SHD) between PAGs counts the occurrences of $*\!\!\rightarrow$ in one of the PAGs versus $*\!\!-$ in the other, plus the number of adjacencies present in one graph but not the other. The normalization is done by dividing over the possible number of errors for the realization at hand to keep the value in $[0, 1]$ and make the numbers comparable. Figure 2e shows the normalized SHD averaged over 30 realizations of synthetic datasets. We used $K = 4$ and $n = 5000$ in this plot; in Section G in the Appendix, we provide plots for $K \in \{2, 6\}$ and $n \in \{1000, 10000\}$.

Identifying Nodes in $V \setminus V_{\mathrm{INV}}$. To evaluate Proposition 4.6, we estimated $V \setminus V_{\mathrm{INV}}$ by determining all nodes incident to bidirected edges in the PAG $\widehat{\mathcal{P}}_\cup$ estimated using FCI. This set was compared to the ground truth; Figure 2f shows true positive and false positive rates for varying significance levels[2], averaged over 30 realizations. We used $K = 4$ and

[2]We do not use ROC plots since while increasing the threshold monotonically increases the true positive rate of the estimated adjacencies, it generally does not monotonically increase the number of correctly inferred edge orientations.

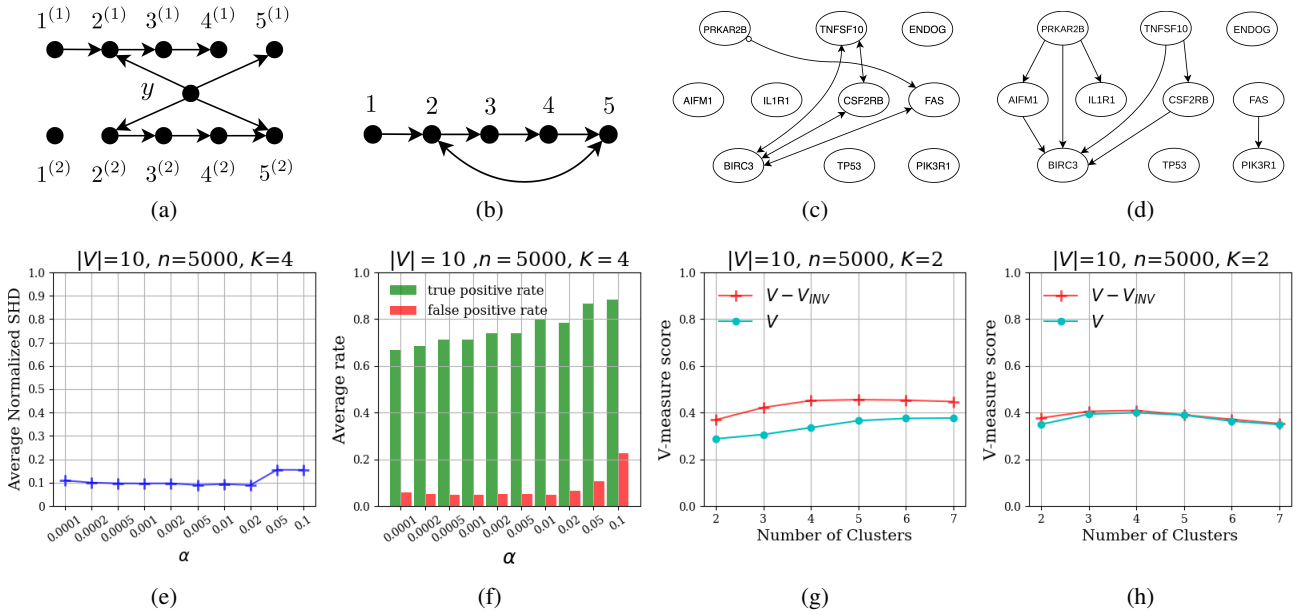


Figure 2: (a) shows a mixture DAG $\mathcal{D}_\mu$ and (b) shows the associated union graph $\mathcal{M}_\cup$. In this model, each DAG $\mathcal{D}^{(j)}$ has a common topological ordering, however, the union MAG is not ancestral; (c) shows the output of FCI on genes in the apoptosis pathway using mixture data without knowledge of the cluster membership for each sample, while (d) shows the difference graph of Wang et al. (2018) on the same genes learned when cluster membership of each sample is known; (e) shows the average normalized SHD between the PAG $\widehat{\mathcal{P}}_\cup$ estimated using the mixture data, and $\widetilde{\mathcal{P}}_\cup$ estimated using data sampled from $\mathcal{M}_\cup$; (f) shows the true and false positive rate in estimating $V \setminus V_{\text{INV}}$; (g) shows the performance of clustering when the set $[V \setminus V_{\text{INV}}]$ has no descendants in $\mathcal{D}_\mu$, while (h) shows the same plot when $[V \setminus V_{\text{INV}}]$ has descendants in $\mathcal{D}_\mu$.

$n = 5000$ in this plot. In Section G in the Appendix, we show plots for $K \in \{2, 6\}$ and $n \in \{1000, 10000\}$.

Clustering. Under mixture-faithfulness, $X_{V \setminus V_{\text{INV}}}$ represents the set of nodes whose conditionals vary across the component DAGs. This motivates using the nodes $X_{V \setminus V_{\text{INV}}}$ and their descendents as features for clustering since these are the only nodes with different marginals across the mixture components. Since FCI generally cannot identify all the descendents of $X_{V \setminus V_{\text{INV}}}$, we used only $X_{V \setminus V_{\text{INV}}}$ for clustering. As a proof-of-concept demonstrating that these features can be useful, we considered two settings, one in which $[V \setminus V_{\text{INV}}]$ has no descendants in $\mathcal{D}_\mu$ (see Figure 2g), and another one in which this set has descendants (Figure 2h).

In both settings, we used $\widetilde{K}$-*means* clustering for various values of $\widetilde{K}$. To compare the quality of clustering using $[V \setminus V_{\text{INV}}]$ versus all nodes as features, we used the V-measure score from Rosenberg and Hirschberg (2007) which is based on ground truth cluster assignments; a higher score represents better performance. As per what is expected from our theoretical results, Figure 2g shows that clustering based on the reduced number of features $[V \setminus V_{\text{INV}}]$ results in higher quality clusters as compared to using all features for clustering in the setting where $[V \setminus V_{\text{INV}}]$ has no descendants in $\mathcal{D}_\mu$, while otherwise both feature sets perform equally.

5.2. Real Data

Ovarian Cancer. We applied this framework to gene expression data from ovarian cancer in $K = 2$ patient groups (with 93 and 168 observations, respectively) with different survival rates (Tothill et al., 2008). We followed the analysis of Wang et al. (2018), where the difference-DAG was estimated for the two groups based on the apoptosis pathway consisting of $|V| = 10$ genes. The resulting difference-DAG is shown in Figure 2d. While the difference-DAG can identify edges that are different between the two DAGs $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$ and hence provides more information than the union graph, computing the difference-DAG requires knowledge of the membership of each observation to the two disease subgroups, which is not available for many diseases. The estimated PAG $\widehat{\mathcal{P}}_\cup$ based on the combined samples from the two patient groups is shown in Figure 2c. It was estimated using FCI with stability selection. FCI identified BIRC3 as the node with the highest number of incident bidirected edges; BIRC3 is known to be one of the major disregulated genes in ovarian cancer and an inhibitor of apoptosis (Johnstone et al., 2008; Jönsson et al., 2014).

T cell activation. We also applied our framework to single-cell gene expression data of naive and activated T cells (i.e. $K = 2$, with 298 and 377 samples, respectively)

from Singer et al. (2016). Following the analysis in Wang et al. (2018), we performed the analysis on 60 genes that had a fold expression change above 10. The FCI output on these 60 nodes is shown in Section G.2 in the Appendix. The following nodes have the highest number of incident bidirected edges, indicating that they may play important roles in T cell activation: CDC6, CDC20, SHCBP1, NKG2A, GZMB4 and KIF2C. All these genes have been discribed before as critical: CDC6 and CDC20 are essential regulators of the cell division cycle. Shorter cell cycle time for increased proliferation is a hallmark of T cell activation (Qiao et al., 2016; Borlado and Méndez, 2008). SHCBP1 has been shown to be tightly linked to cell proliferation and strongly correlates with proliferative stages of T cell development (Schmandt et al., 1999; Buckley et al., 2014). NKG2A functions to limit excessive activation, prevent apoptosis, and preserve the specific T cell response (Rapaport et al., 2015). GZMB4 has been shown to regulate antiviral T cell response (Salti et al., 2011). Finally, the gene KIF2C encodes a Kinesin-like protein that functions as a microtubule-dependent molecular motor. It is over-expressed in a variety of solid tumors and induces frequent T cell responses (Gnjatic et al., 2010).

6. DISCUSSION

In this paper, we provided a graphical representation (via the mixture DAG) of distributions that arise as mixtures of causal DAGs. We showed that the mixture DAG not only satisfies the Markov property with respect to such mixture distributions, but is also always realizable by a mixture distribution, meaning that it cannot be made sparser without losing the Markov property. In addition, we characterized the output of the prominent FCI algorithm when applied to data from such mixture distributions. FCI is a natural candidate in this setting due to the presence of the latent mixing variable. We proved that FCI can identify variables whose conditionals vary across the different components and showed how this property can be used to infer cluster membership of samples. This is relevant for many applications, as for example when studying diseases consisting of multiple not well characterized subtypes. In such studies, genomic perturbation experiments can now be performed relatively routinely, leading to high-throughput interventional data. In future work it would be interesting to study how interventional data could be used to enhance causal inference based on mixtures of DAGs or which interventions to perform in order to enhance identifiability of pathways that are shared among the different subtypes as well as those that are different across the subtypes for personalized interventions.

ACKNOWLEDGEMENTS

Basil Saeed was partially supported by the Abdul Latif Jameel Clinic for Machine Learning in Health at MIT. Caroline Uhler was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-2765), IBM, and a Simons Investigator Award.

REFERENCES

Borlado, L. R. and Méndez, J. (2008). CDC6: from DNA replication to cell cycle checkpoints and oncogenesis. *Carcinogenesis*, 29(2):237–243.

Buckley, M. W., Arandjelovic, S., Trampont, P. C., Kim, T. S., Braciale, T. J., and Ravichandran, K. S. (2014). Unexpected phenotype of mice lacking SHCBP1, a protein induced during T cell proliferation. *PloS ONE*, 9(8).

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554.

Chu, T., Glymour, C., Scheines, R., and Spirtes, P. (2003). A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, 19(9):1147–1152.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620.

Gates, K. M. and Molenaar, P. C. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, 63(1):310–319.

Gnjatic, S., Cao, Y., Reichelt, U., Yekebas, E. F., Nölker, C., Marx, A. H., Erbersdobler, A., Nishikawa, H., Hildebrandt, Y., Bartels, K., et al. (2010). NY-CO-58/KIF2C is overexpressed in a variety of solid tumors and induces frequent T cell responses in patients with colorectal cancer. *International Journal of Cancer*, 127(2):381–393.

Heckerman, D., Mamdani, A., and Wellman, M. P. (1995). Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3):24–26.

Johnstone, R. W., Frew, A. J., and Smyth, M. J. (2008). The TRAIL apoptotic pathway in cancer onset, progression and therapy. *Nature Reviews Cancer*, 8(10):782–798.

Jönsson, J.-M., Bartuma, K., Dominguez-Valentin, M., Harbst, K., Ketabi, Z., Malander, S., Jönsson, M., Carneiro, A., Måsbäck, A., Jönsson, G., et al. (2014). Distinct gene expression profiles in ovarian cancer linked to Lynch syndrome. *Familial Cancer*, 13(4):537–545.

Lauritzen, S. L. (1996). *Graphical Models*, volume 17. Clarendon Press.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Qiao, R., Weissmann, F., Yamaguchi, M., Brown, N. G., VanderLinden, R., Imre, R., Jarvis, M. A., Brunner, M. R., Davidson, I. F., Litos, G., et al. (2016). Mechanism of APC/CCDC20 activation by mitotic phosphorylation. *Proceedings of the National Academy of Sciences*, 113(19):E2570–E2578.

Ramsey, J., Spirtes, P., and Glymour, C. (2011). On meta-analyses of imaging data and the mixture of records. *NeuroImage*, 57(2):323–330.

Rapaport, A. S., Schriewer, J., Gilfillan, S., Hembrador, E., Crump, R., Plougastel, B. F., Wang, Y., Le Friec, G., Gao, J., Cella, M., et al. (2015). The inhibitory receptor NKG2A sustains virus-specific CD8+ T cells in response to a lethal poxvirus infection. *Immunity*, 43(6):1112–1124.

Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Sadeghi, K. et al. (2013). Stable mixed graphs. *Bernoulli*, 19(5B):2330–2358.

Salti, S. M., Hammelev, E. M., Grewal, J. L., Reddy, S. T., Zemple, S. J., Grossman, W. J., Grayson, M. H., and Verbsky, J. W. (2011). Granzyme B regulates antiviral CD8+ T cell responses. *The Journal of Immunology*, 187(12):6301–6309.

Schmandt, R., Liu, S. K., and McGlade, C. J. (1999). Cloning and characterization of mPAL, a novel Shc SH2 domain-binding protein expressed in proliferating cells. *Oncogene*, 18(10):1867–1879.

Singer, M., Wang, C., Cong, L., Marjanovic, N. D., Kowalczyk, M. S., Zhang, H., Nyman, J., Sakuishi, K., Kurtulus, S., Gennert, D., et al. (2016). A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating T cells. *Cell*, 166(6):1500–1511.

Solus, L., Wang, Y., Matejovicova, L., and Uhler, C. (2017). Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*.

Spirtes, P. (1994). Conditional independence properties in directed cyclic graphical models for feedback. Technical report, Carnegie Mellon University.

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

Strobl, E. V. (2019a). The global Markov property for a mixture of DAGs. *arXiv preprint arXiv:1909.05418*.

Strobl, E. V. (2019b). Improved causal discovery from longitudinal data using a mixture of DAGs. In *Proceedings of Machine Learning Research*, volume 104, pages 100–133.

Thiesson, B., Meek, C., Chickering, D. M., and Heckerman, D. (1997). Learning mixtures of DAG models. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 504–513.

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., Johnson, D. S., Trivett, M. K., Etemadmoghadam, D., Locandro, B., et al. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14(16):5198–5208.

Wang, Y., Segarra, S., and Uhler, C. (2020). High-dimensional joint estimation of multiple directed Gaussian graphical models. *Electronic Journal of Statistics*.

Wang, Y., Squires, C., Belyaeva, A., and Uhler, C. (2018). Direct estimation of differences in causal graphs. In *Advances in Neural Information Processing Systems*, pages 3770–3781.

Zhang, J. (2008). Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(Jul):1437–1474.