
Supplementary Material

Causal Mixture Models: Characterization and Discovery

Anonymous Author(s)

Affiliation

Address

email

1 A Preliminaries

2 In this section, we provide a brief overview of helpful preliminary concepts that, although relevant to
3 our analysis, we assume to be widely familiar in the main target audience of our work; for brevity and
4 clarity, we therefore chose to postpone them away from the main manuscript and into this appendix.

5 A.1 Structural Causal Models

6 Given random variables $\mathbf{X} = \{X_1, \dots, X_n\}$ we can encode the underlying data generating process
7 (DGP) as a structural equation model (SEM) [Koller and Friedman, 2009], which encodes a set of
8 hypotheses on this process in the form of one functional dependency f_j for each random variable
9 $X_j \in \mathbf{X}$, so that

$$X_j = f_j(\mathbf{X}_j) \quad \text{with } \mathbf{X}_j \subseteq \mathbf{X} \setminus \{X_j\}. \quad (1)$$

10 Of special interest is a structural causal model (SCM) [Bollen, 1989], which is a particular kind
11 of an SEM with additional assumptions that allow it to also model the causal mechanisms of the
12 DGP. Here, the set of random variables \mathbf{X} is extended to also include random unobserved variables
13 $\mathbf{U} = \{U_1, \dots, U_n\}$, which play the role of noise. Hence, each functional dependency takes the form

$$X_j = f_j(\mathbf{X}_j, U_j) \quad \text{with } \mathbf{X}_j \subseteq \mathbf{X} \setminus \{X_j\} \text{ and } U_j \in \mathbf{U}. \quad (2)$$

14 To further study SCMs, we need to establish their correspondence with causal graphs [Pearl, 2009].

15 A.2 Causal Graphs

16 Consider a set of random variables $\mathbf{X} = \{X_1, \dots, X_n\}$ that follow a distribution $\mathcal{L}_{\mathbf{X}}$ that has a joint
17 probability density $p_{\mathbf{X}}$ with respect to some appropriate measure, and an (arbitrary) total ordering
18 $X_1 < X_2 < \dots < X_n$. Then the joint probability density factorises as

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{X_j \in \mathbf{X}} p_{X_j|\mathbf{Y}_j}(x, \mathbf{y}_j), \quad \text{where } \mathbf{Y}_j \subseteq \{X_1, \dots, X_j\} \quad (3)$$

19 and \mathbf{y}_j are those values out of \mathbf{x} corresponding to the same indices as \mathbf{Y}_j . This comes as a direct result
20 of the chain rule and the conditional independence rules. Any such factorisation can be represented
21 as a (fully) directed acyclic graph (DAG) $\mathcal{G} = (\mathbf{X}, E)$ with nodes the random variables \mathbf{X} and edges
22 the set $E = \cup_{j=1}^n \{i \rightarrow j | X_i \in \mathbf{Y}_j\}$. In other words, in this graph we add an edge to the dependent
23 variable X_j from each variable in the corresponding conditioning set \mathbf{Y}_j that appears in each factor
24 $p_{X_j|\mathbf{Y}_j}$ of Eq. (3). We further make this relation explicit, by instead writing $\mathbf{Pa}_j = \mathbf{Y}_j$ to indicate
25 that the conditioning set \mathbf{Y}_j serves as the set of direct parents of node X_j in \mathcal{G} . Such a graph is called
26 a *Bayesian network* [Koller and Friedman, 2009] and allows for a visual representation of all those

27 independencies that are implied solely from this factorisation of the joint density and irrespective of
 28 the form of each factor.

29 These independencies can be read from the graph in terms of the d-separation [Pearl, 2009].

30 **Definition A.1** (d-separation). For any pairwise disjoint subsets $\mathbf{U}, \mathbf{V}, \mathbf{W} \subseteq \mathbf{X}$, and $\mathbf{U}, \mathbf{V} \neq \emptyset$, it is

$$\mathbf{U} \perp\!\!\!\perp \mathbf{V} | \mathbf{W} \iff \text{all paths from any variable in } \mathbf{U} \text{ to any of } \mathbf{V} \text{ are blocked by } \mathbf{W}. \quad (4)$$

31 We call a path blocked if it either

- 32 • traverses a section $\rightarrow V \rightarrow, \leftarrow V \leftarrow$ or $\leftarrow V \rightarrow$ for some variable $V \in \mathbf{W}$, or
- 33 • traverses a section $\rightarrow V \leftarrow$ where neither V nor any of its descendants are contained in \mathbf{W} .

34 Hence, a Bayesian network \mathcal{G} can be seen as a description of an entire family of distributions that
 35 fulfill a given set of conditional independencies. When a distribution $\mathcal{L}_{\mathbf{X}}$ exhibits all the conditional
 36 independencies that one can read from the graph \mathcal{G} , we call \mathcal{G} an I-map of $\mathcal{L}_{\mathbf{X}}$ and write $\mathcal{L}_{\mathbf{X}} \in I_{\mathcal{G}}$.

37 In other words, the I-map defines an equivalence relation among all DAGs via the relation $\mathcal{G} \equiv_M$
 38 $\mathcal{G}' \iff I_{\mathcal{G}} = I_{\mathcal{G}'}$, of which each equivalence class is called the *Markov equivalence class* (MEC).
 39 When, in addition, each of the factors in the factorisation of Eq. (3) correspond to a functional
 40 dependency of an SCM, we call the resulting DAG causal.

41 A graphical representation of MECs can be given through the common notion of *completed partially*
 42 *directed acyclic graphs* (CPDAGs) [Chickering, 2002] also known under other names such as
 43 maximally oriented graphs [Meek, 1997]. A partially directed graph (PDAG) \mathcal{P} contains both
 44 undirected and directed edges, and can be associated to an equivalence class $\mathcal{M}(\mathcal{P})$ with $\mathcal{G} \in \mathcal{M}(\mathcal{P})$
 45 if and only if \mathcal{G}, \mathcal{P} have the same skeleton and v-structures. The notion of completion of PDAGs
 46 allows for representing equivalence classes uniquely. To this end, for a given equivalence class \mathcal{M}
 47 one distinguishes between *compelled* edges with the same directionality in every member of \mathcal{M} , and
 48 *reversible* edges otherwise. The completed PDAG \mathcal{P} for \mathcal{M} is then the one having a directed edge
 49 for every compelled edge in \mathcal{M} , and an undirected edge for every reversible edge in \mathcal{M} .

50 A.3 SCMs and causal graphs

51 We now return to the assumptions implicit in an SCM.

52 **Assumption A.2** (Causal Interpretation). Each functional dependency f_j corresponds to a true causal
 53 mechanism in the data, with \mathbf{X}_j being the direct causes of the direct effect X_j .

54 As a corollary, the corresponding causal graph can have no recurrence.

55 **Assumption A.3** (Orientation and Acyclicity). The causal graph of an SCM is a DAG.

56 Hence, we can once again identify the direct causes of each effect X_j with its parents in the
 57 corresponding DAG, $\mathbf{X}_j = \mathbf{Pa}_j$.

58 **Assumption A.4** (Exogeneity of Noise). The random variables \mathbf{U} are exogenous; that is, there are
 59 no edges $X_j \rightarrow U_i$, for any X_j and any U_i .

60 **Assumption A.5** (Independence of Noise). The random variables \mathbf{U} are mutually independent.

61 In our work, we consider that part of the noise is the latent variable \mathbf{L}_j , in addition to the typical U_j .
 62 Hence, even though for the \mathbf{U} we do assume Assumption A.5 to hold, we note that this fails to hold
 63 for the entire set of exogenous noise sources, which, in our case, is further extended to include all \mathbf{Z} .
 64 As a result, to be more rigorous, we make claims to find the CPDAG corresponding to the conditional
 65 distribution $\mathcal{L}_{\mathbf{X}|\mathbf{Z}}$, rather than the marginal $\mathcal{L}_{\mathbf{X}}$.

66 We remark that, if one is interested in the entire marginal $\mathcal{L}_{\mathbf{X}}$, this can be achieved as a two-level
 67 algorithm, in which we first use our presented method to infer the Markov equivalence class of
 68 $\mathcal{L}_{\mathbf{X}|\mathbf{Z}}$, then using the computed values of each \mathbf{L}_j to identify which variables correspond to the
 69 same underlying latent Z_i , and finally performing causal structure inference over the values of \mathbf{Z} to
 70 complete the Markov equivalence class of $\mathcal{L}_{\mathbf{X}}$.

71 In the scope of our work, we focus on the conditional $\mathcal{L}_{\mathbf{X}|\mathbf{Z}}$. To formally show that we can recover
 72 the corresponding underlying Markov equivalence class, as represented by a CPDAG \mathcal{G} , will be the
 73 focus in the following section. Specifically, we next move to formally showing our main claims in
 74 Theorem 3.4 on using the proposed latent-aware BIC as a consistent scoring criterion for this purpose.

75 B Proof of Theorem B.2

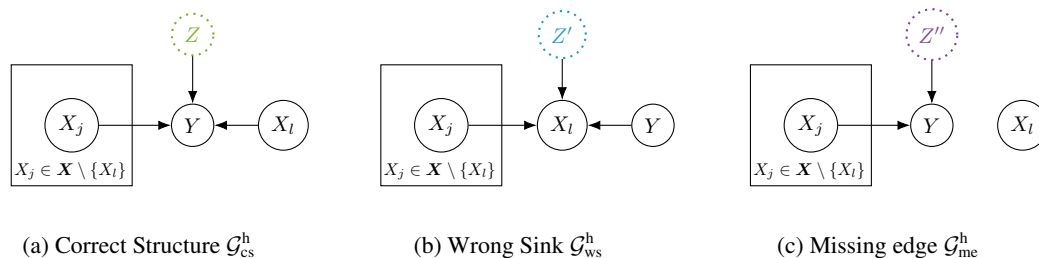


Figure B.1: *Identifiable Cases*: Under mild assumptions, when Z has at least two mixing components, all of the shown DAGs are identifiable.

76 In this section we elaborate on the theoretical justification of our method. For this, we make
 77 use of the relaxed space constraints to break down our analysis in finer finer steps. We first
 78 show that under mild conditions, the MLR distribution does not degenerate to a Gaussian.
 79

80 **Lemma B.1** (Non Gaussianity of Direct Effect). *Let $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_r\}$ be*
 81 *observations of random variables \mathbf{X}, Y , such that $\mathbf{X}|Y \sim \text{MLR}(\mathbf{B}, \gamma, \sigma^2)$,*
 82 *with the parameters $\theta = (\mathbf{B}, \gamma, \sigma^2) \in \cup_{K \geq 2} \Theta_K$, where Θ_K is the parameter*
 83 *space for K mixtures. Also assume that the $\gamma_k \neq 0$ for any configuration*
 84 *above, that the prior distribution of \mathbf{B} has a positive density over the entire*
 85 *euclidean space, with the full dimensionality for each K . Then the marginal*
 86 *distribution of $Y|\mathbf{X}$ is not a Gaussian with probability 1.*

87 *Proof.* We need to show that the MLR distribution with density

$$p_{Y|\mathbf{X}}^{\text{MLR}}(y, \mathbf{x}; \mathbf{B}, \gamma, \sigma^2) = \sum_{k=1}^K \gamma_k p_X^{\mathcal{N}}(x; \beta_k^\top \mathbf{y}, \sigma^2) \quad (5)$$

88 cannot be formulated into a single Gaussian. For this to happen either of two
 89 conditions would have to hold:

- 90 1. there is only one component in the mixture, $\mathbb{P}(Z = 1) = 1$, or
- 91 2. each component has the same linear coefficients.

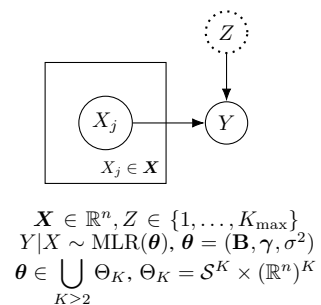
92 The former condition can only happen with probability zero, since we required that $\gamma_k \neq 0$ for all
 93 $1 \leq k \leq K$ and all $\gamma \in \mathcal{S}^K$ and that $K \geq 2$; hence, only finite points out of the domain of \mathcal{S}^K
 94 would satisfy this condition, which can only happen with probability 0.

95 To study the latter condition, we treat the domain of the linear coefficients as $\mathbf{B} \in \mathbb{R}^{n \times K}$, which is
 96 homeomorphic to $(\mathbb{R}^n)^K$, for each K . For all (or in fact, any two) of the linear coefficients to be the
 97 same, the matrix \mathbf{B} would need to have a rank less than K , where we assume that $K \leq K_{\max} \leq n$.
 98 For this to happen, it would also mean that \mathbf{B} should lie in a subspace of $\mathbb{R}^{n \times K}$ with dimension
 99 strictly lower than K ; in this case, we know that the Lebesgue measure of any subset of this space
 100 has zero measure. Hence, the measure of this set under an (absolutely continuous) prior on the
 101 parameters \mathbf{B} that is positive in the entire $\mathbb{R}^{n \times K}$ would have a zero measure under the corresponding
 102 push-forward measure.

103 Overall, the probability of these conditions happening has probability equal to that of the union of
 104 zero probability events, which itself happens with probability 0. \square

105 To show identifiability, we need to be able to distinguish, under the true hypothesis $Y|\mathbf{X} \sim \text{MLR}$,
 106 between all competing hypotheses of Fig. B.1.

107 **Theorem B.2** (Local Consistency of $\text{BIC}_{\hat{Z}}^{\text{ML}}$). *Let $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_r\}$ be observations of random*
 108 *variables \mathbf{X}, Y , such that $\mathbf{X}|Y \sim \text{MLR}(\mathbf{B}, \gamma, \sigma^2)$, with the parameters θ satisfying the conditions*
 109 *of Lemma B.1. Then, out of the structural hypotheses depicted in Fig. B.1 the $\text{BIC}_{\hat{Z}}^{\text{ML}}$ score of the*



110 ground truth hypothesis \mathcal{G}_{cs}^h is asymptotically larger than any of the alternative ones, \mathcal{G}_{ws}^h and \mathcal{G}_{me}^h ,
 111 with probability 1.

112 *Proof.* This claim builds on the established properties of the vanilla BIC score. Here, we treat
 113 $\mathbf{X} \setminus \{X_l\}$ as nuisance parameters, and we focus on single edge modifications, between the sing of
 114 each structural hypothesis and the node on the right of each depiction. We first consider the pair
 115 \mathcal{G}_{cs}^h and \mathcal{G}_{ws}^h . In this case, since the $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}$ is based on the Maximum Likelihood Estimate (MLE)
 116 estimates $\hat{\boldsymbol{\theta}}$ of the true parameter values $\boldsymbol{\theta}$, and the MLE estimate is asymptotically biased, then the
 117 the correct model \mathcal{G}_{cs}^h and the one arising from the alternate hypothesis \mathcal{G}_{ws}^h have the same number of
 118 parameters, while at the same time the likelihood under \mathcal{G}_{cs}^h is larger than that of \mathcal{G}_{ws}^h . Hence, in this
 119 case the $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}$ value is an increasing function of only the likelihood, and hence it mustbe that also
 120 $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}(\mathcal{G}_{cs}^h) > \text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}(\mathcal{G}_{ws}^h)$.

121 For the pair \mathcal{G}_{cs}^h and \mathcal{G}_{me}^h the number of parameters in the latter hypothesis is smaller than that of the
 122 \mathcal{G}_{cs}^h . Under similar reasoning as in Lemma B.1, we can claim that $Y \not\perp\!\!\!\perp X_l$ with probability 1. The
 123 rest follows from established asymptotic behaviour of $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}$ as a special case of BIC, due to the
 124 decomposability property that $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}$ inherits from BIC. \square

125 Using this result, we can extend the local consistency of $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}$ to its global consistency.

126 **Lemma 3.3.** *The latent-aware score $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}$ is a consistent scoring criterion.*

127 *Proof.* By considering any sequence of appropriate single edge modifications between adjacent struc-
 128 tural hypotheses as in Fig. B.1, we can extend the global consistency of BIC to that of $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{ML}}$ Chick-
 129 ering [2002]. \square

130 C Implementation Considerations

131 We note that our main theory covers the Greedy Equivalent Search (GES) algorithm. In our imple-
 132 mentation, however, we have used TOPIC Xu et al. [2025], a more recent greedy score-based search
 133 that has similar guarantees to GES, and when similar requirements are met by the used score. Hence,
 134 we replace within TOPIC our proposed $\text{BIC}_{\hat{\mathbf{Z}}}^{\text{EM}}$ score, to thus derive $\text{TOPIC}_{\text{BIC}}$, and he here analyse
 135 the two algorithms.

136 We first assume access to the MLE oracle. Then, although the output of both algorithms lies on the
 137 domain of all CPDAGs over \mathbf{X} , $\text{TOPIC}_{\text{BIC}}$ builds on TOPIC, which is both asymptotically and
 138 practically more efficient than GES. Indeed, in each iteration, it first limits the candidate hypotheses
 139 from the set of all representatives of the Markov equivalence classes which perform a single-edge
 140 modification from the current best, to only those which differ with respect to the most likely modified
 141 source. Formally, the combination of these two steps are two subsequent maximisations over exactly
 142 the same domain (of all hypotheses with single edge modifications), only performed first over the
 143 possible sources, and then over the rest of the hypotheses, conditioned on the chosen source. Hence,
 144 at each iteration, the asymptotic greedy optimum is the same in both algorithms.

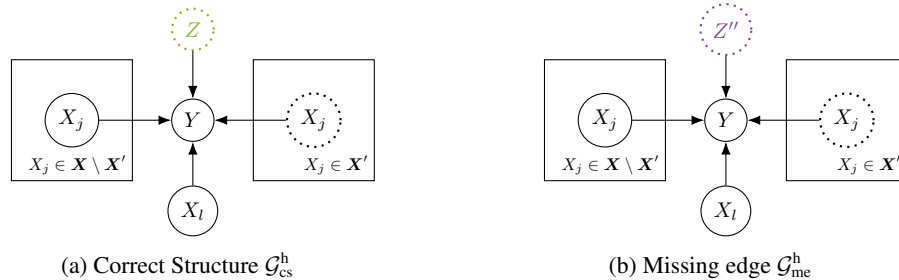


Figure C.1: Decomposability of the MLR model in the intermediate stages of $\text{TOPIC}_{\text{BIC}}$. The so-far-undiscovered edges in $\mathbf{X}' \subseteq \mathbf{X}$ are akin to noise that affects both cases equally.

145 To see the asymptotical consistency, in the particular assumptions of our causal model, we treat two
 146 different cases. First, when no latent variable affects the result, $\text{TOPIC}_{\text{BIC}}$ would have similar ease to
 147 detect edge additions as in the case of TOPIC/GES. In the case that a the true structure is an Mixture
 148 of Linear Regressions (MLR) model, we can consider the sequence of edge modifications that the
 149 $\text{TOPIC}_{\text{BIC}}$ algorithm would produce. Within this sequence, we can revisit the cases of Fig. B.1,
 150 and notice that when a subset $\mathbf{X}' \subset \mathbf{X}$ is not yet discovered in the structure of the algorithm, the
 151 effects of the so-far-undiscovered variables \mathbf{X}' can be seen as added noise, which equally affects an
 152 appropriate hypothesis $\mathcal{G}_{\text{cs}}^h$ and $\mathcal{G}_{\text{me}}^h$, as shown in Fig. C1.

153 Hence, intuitively, we expect a point at which one of the edges of the type $X_l \rightarrow Y$ would be added
 154 to the model, until all parents \mathbf{X} will be discovered. Finally, we posit that the practical use of $\text{BIC}_{\hat{Z}}^{\text{EM}}$
 155 in lieu of $\text{BIC}_{\hat{Z}}^{\text{ML}}$ is equally affecting both frameworks, as long as Conjecture 1 holds.

D Detailed Evaluation

Experimental Setup We give a more detailed description of our synthetic data generation here. In iteration $i \in \{1, \dots, N_I\}$ of each experiment, we randomly sample a DAG \mathcal{G} over $N_X := |\mathbf{X}|$ observed variables under an Erdős R nyi model with edge density $p_G \in [0, 1]$. In addition, we draw $N_Z := |\mathbf{Z}|$ latent mixing variables $Z_i \sim \text{Categorical}(\gamma^i)$ with $Z_i \in \{1, \dots, K_i\}$, where we fix all $K_i =: K$ to the same hyperparameter K . We then sample a set of so-called mixing targets $\mathbf{T} = \{X_j \mid \exists Z_i : \text{La}_j = Z_i\} \subseteq \mathbf{X}$ where $X_j \in \mathbf{T}$ with probability $p_Z \in [0, 1]$. We distribute the effect of the N_Z mixing variables equally across these targets, resulting in $0 \leq i \leq N_Z$ many disjoint sets $\mathbf{T}_i = \{X_j \mid \text{La}_j = Z_i\}$. For example, we have $\mathbf{T}_1 = \{X_1, X_2\}$ and $\mathbf{T}_2 = \{X_4\}$ in Fig. 2.

To generate samples, we traverse \mathbf{X} in topological order of the induced \mathcal{G} . For each X_j , we sample $\mathbf{B}_j = \{\beta_{j1}, \dots, \beta_{jK_i}\}$ coefficient vectors, where $\beta_{jk} \in \mathbb{R}^{|\mathbf{Pa}_j|}$ for all $k \in \{1, \dots, K_i\}$ with $K_i = K$ if $X_j \in \mathbf{T}$ and $K_i = 1$ otherwise. We draw $\beta_{jk} \in [-1, -\epsilon] \cup [\epsilon, 1]$ to avoid causal effects close to zero, and if possible also ensure that $|\beta_{jk} - \beta_{jk'}| > \epsilon$ for all pairs k, k' to create sufficient class separation, where $\epsilon = 0.25$ by default. We then draw S samples from a (mixture of) linear regression model(s) $(X_j \mid \mathbf{Pa}_j = \mathbf{y}) \sim \text{MLR}(\mathbf{B}_j, \gamma^j, \sigma^2)$, and standardize the resulting samples to generate an internally-standardized structural causal model (iSCM) [Ormaniec et al., 2024].

In the case studies, we consider a mixture of interventional datasets, as well as the flow cytometry dataset by Sachs et al. [2005] under the experimental setup in Wang et al. [2017]. For both cases, we use the same scripts as in Kumar et al. [2024]¹. For the mixture of interventions, we have $N_Z = 1$ with $K = N_X + 1$ classes which defines a split into one observational and K interventional datasets. Under a so-called diagonal or atomic setting, one node at a time undergoes intervention, resulting in disjoint sets $\mathbf{I}_k \subseteq \mathbf{X}$, here with hard interventions that fix $\beta_{jk} = 0$ if $X_j \in \mathbf{I}_k$. A similar structure applies to the real-world dataset with 5846 samples and known manipulations on 5 of the 11 variables, namely $\mathbf{I}_1 = \{\text{Akt}\}$, $\mathbf{I}_2 = \{\text{PKC}\}$, $\mathbf{I}_3 = \{\text{PIP2}\}$, $\mathbf{I}_4 = \{\text{Mek}\}$, $\mathbf{I}_5 = \{\text{PIP3}\}$.

Evaluation Metrics To evaluate the discovered number of mixing components and assignments, standard metrics in clustering evaluation are appropriate, where we show the v-measure and the Adjusted Mutual Information (AMI) as two examples (e.g., Vinh et al. [2010]). We average each score over \mathbf{X} since we can associate each variable X_j to a fixed assignment with K_i components if $\text{La}_j = Z_i$, else $K_i = 1$. To validate statements on *whether* observed variables are mixing targets, we compute F1 scores (called F1-target) over the statement $X_j \in \mathbf{T}$ averaged over \mathbf{X} . To validate results on *which* mixing variables affect which mixing targets, we compute the Jaccard index (called Jacc) comparing the true sets $\{\mathbf{T}_1, \dots, \mathbf{T}_{Z_m}\}$ to those returned by our algorithm.

We also compare the induced DAG \mathcal{G} against the discovered DAG or PDAG \mathcal{G}' . As a simplistic score that gives intuitive insight into correctly oriented vs. spurious edge counts, we show F1 scores over directed edges E in \mathcal{G}' (called F1-dir), where we note that in the case of PDAGs we only count edges that are directed with certainty. A classical distance score for two DAGs $\mathcal{G}, \mathcal{G}'$ is the Structural Hamming Distance (SHD). More suitable to causal DAGs $\mathcal{G}, \mathcal{G}'$ are the distance scores proposed in Wahl and Runge [2024]. The *s/c-metrics* (S/C) are based on counting separation statements and comparing their validity in $\mathcal{G}, \mathcal{G}'$. Scalable variants thereof are the *separation distances* (SD) that associate each pair of separable nodes in \mathcal{G}' with a separation set S under a given separation strategy, and validate whether S remains separating in \mathcal{G} . We report the SC (without randomization) and the SD (with the 'parent' resp. 'pparent' type) which are defined both when \mathcal{G}' is a DAG or PDAG².

Baselines As our method is the only one to discover the full \mathcal{G}^Z , we show (i) AMI and F1-target scores over \mathbf{X} for all clustering baselines and wherever applicable for Mixture-UTIGSP, (ii) metrics on \mathcal{G} for all causal discovery baselines, and (iii) Jaccard scores over $\{\mathbf{T}_i\}_i$ only for our method. We note that we apply all baselines without optimization of their hyperparameters using the standard implementations in the causal-learn, causal discovery toolbox (cdt), causalDisco and dodiscover Python libraries. For all conditional-independence tests, we use the Fisher-Z test given the linearity of our functional model. We ran the evaluations on an 11th Gen Intel Core i9 CPU.

¹using the implementation of Mixture-UTIGSP at https://github.com/BigBang0072/mixture_mec

²using the implementation of the metrics at https://github.com/JonasChoice/sep_distances

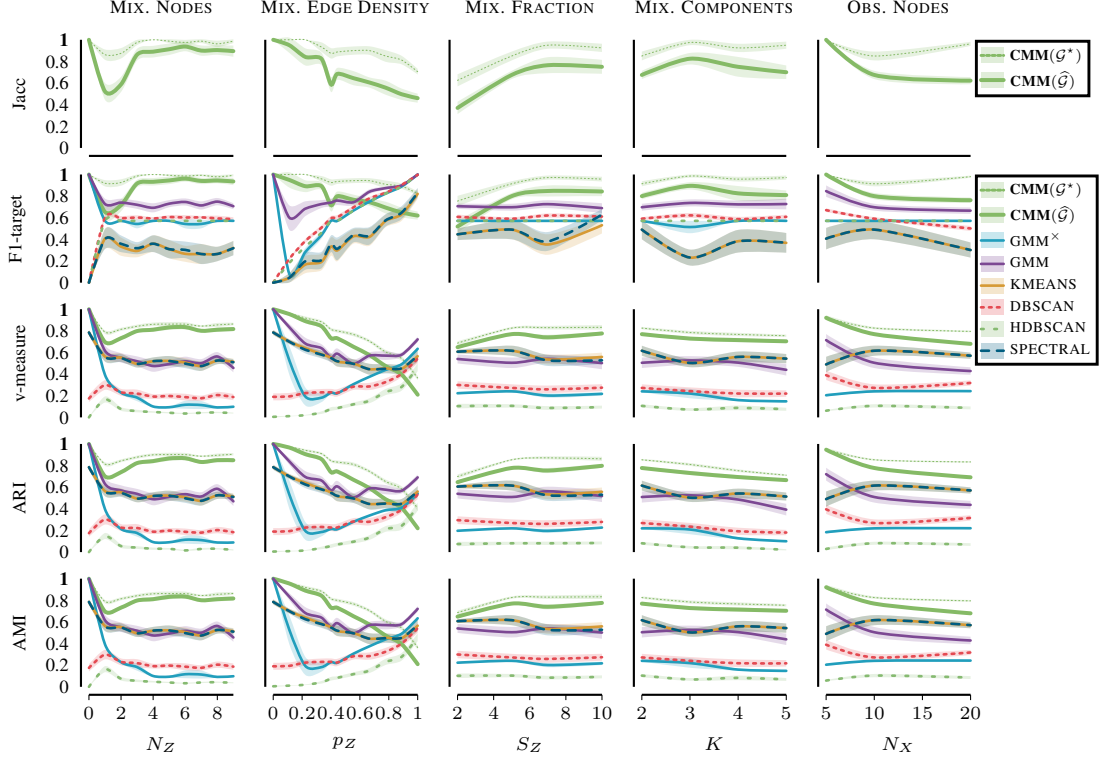


Figure D.1: *Discovering Mixing Structure.*

Discovering Mixing Structure Fig. D.1 shows an extended variant of Fig. 4 in the main manuscript. The parameters are $N_X = 10$, $N_Z = 2$, $K = 2$, $p_Z = 0.4$, $p_G = 0.4$, $S = 1000$, $S_Z = 5$, which are held fixed while changing one parameter of interest (columns in Fig. D.1), where we run $N_I = 10$ iterations for each parameter configuration. Other choices of clustering algorithm besides the GMM, here shown in color for better readability, perform either worse on recovering targeted nodes (KMEANS, SPECTRAL) or mixing assignments (DBSCAN, HDBSCAN). We observe no noticeable differences between the clustering metric choices, so we report the AMI in the main manuscript.

Discovering Causal Structure Fig. D.2 shows an extended variant of Fig. 5 in the main manuscript. The parameters are $N_X = 10$, $N_Z = 4$, $K = 2$, $p_Z = 0.5$, $p_G = 0.4$, $S = 500$, $S_Z = 5$, $N_I = 10$. All structural metrics (SD, S/C, and SHD) show stable performance of the CMM across the settings. The intuitive score TPR-dir furthermore suggests that our method performs well in distinguishing causal edge directions, improving as sample size S increases, and as the likelihood of mixing p_Z increases. In particular, the results suggest that "sparse" mixing $0 < p_Z < 1$ is most beneficial. We connect this to previous findings in the multi-environment setting [Perry et al., 2022] showing that identification of edge orientations is possible under the sparse mechanism shift hypothesis. This could inspire future work directions to explore this property also in the more general mixed setting.

Discovering Interventional Mixtures Finally, Fig. D.2 extends Fig. 6 to show both mixing structure (Jacc, F1-iv) and causal structure (SD, S/C, SHD) discovery for the mixture of interventions. As a well-defined split into observational and interventional datasets exists for this setting, we also include Mixture-UTIGSP in the presentation (dark blue). The parameters $N_X \in \{4, 6, 8\}$ are as in Kumar et al. [2024], while we restrict the setting to up to $S = 1000$ samples, given that the true positive rates over \mathcal{G} for the CMM (and VARSORT) already approach 1 for $S = 1000$; we refer to Kumar et al. [2024] to see the performance of Mixture-UTIGSP with more samples. Compared to Fig. D.1, the CMM performs much better on discovering whether $(X_j \in \mathcal{T})$ and which nodes $(X_j \in \mathcal{T}_i)$ are mixed, which we explain by the fact that hard interventions $\beta = 0$ create a more distinct separation than re-sampling of β with $|\beta_{jk} - \beta_{jk'}| > \epsilon$.

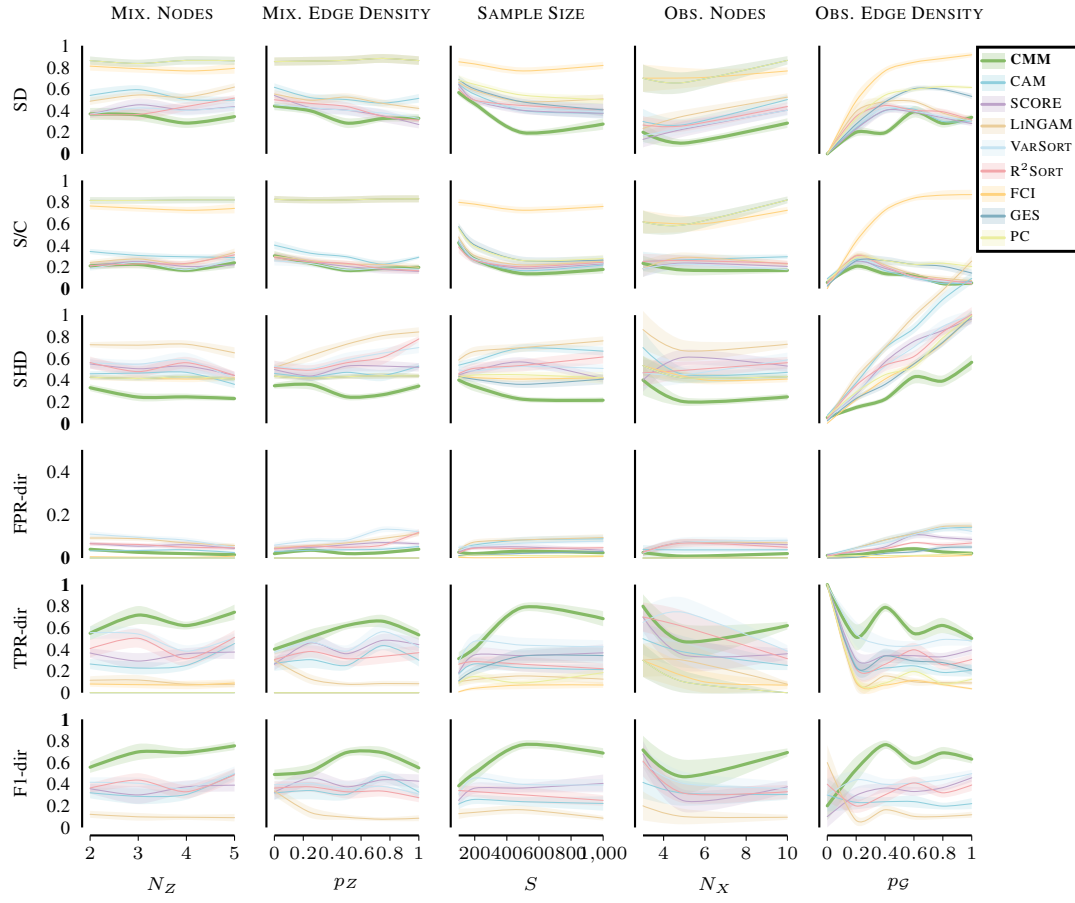


Figure D.2: *Discovering Causal Graphs.*

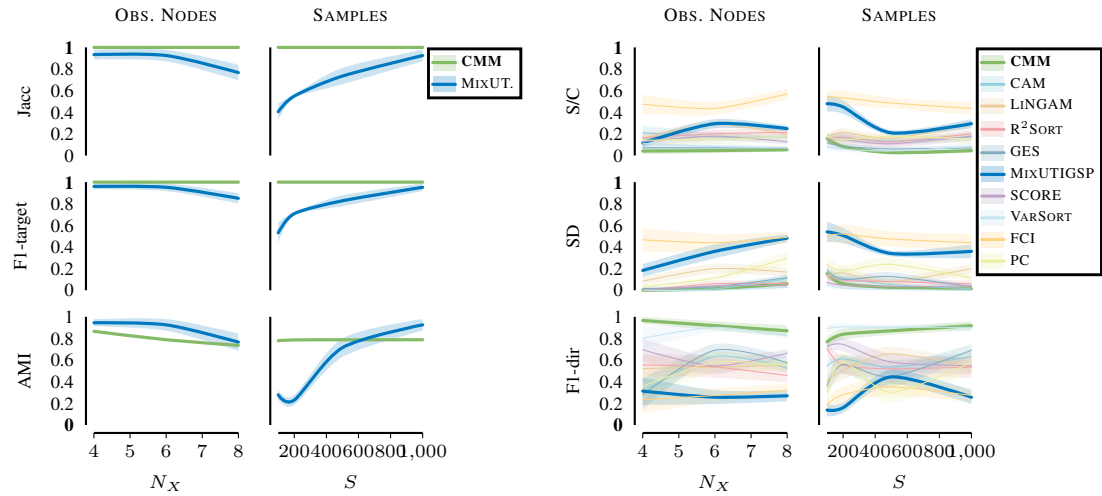


Figure D.3: *Discovering mixing (left) and graphs (right) in interventional mixtures.*

E Ablation Studies

As additions to our main experimental evaluation, we perform ablation studies on two questions.

- (i) *Choice of Score-Based Algorithm.* We address the choice of the score-based causal discovery algorithm used together with the latent-aware BIC, in particular the GES variants.
- (ii) *Effect of Latent Mixing Variables.* We take a closer look at how latent mixing affects causal discovery algorithms in practice. We hypothesize that that spurious edges will appear between mixing targets, and study to which extent the latent-aware BIC can prune these.

E.1 Choice of Score-Based Algorithm

While we used the latent-aware score $\text{BIC}_{\mathcal{Z}}^{\text{EM}}$ within the topological-ordering-based framework in the main evaluation, we can also use it within the GES algorithm, compared to standard GES with BIC. We compare these three variants in Table E.1 for the basic experimental parameters (as in Fig. D.2). Regarding the choice of the algorithm, the topological-ordering-based variant (**CMM (TOP.)**) appears to have better practical performance in our experiments. The experiment also confirms that regarding the score, $\text{BIC}_{\mathcal{Z}}^{\text{EM}}$ with MLR fitting (**CMM (GES)**) provides a benefit over plain BIC (GES).

METRIC	CAUSAL GRAPH		
	CMM (TOP.)	CMM (GES)	GES
SHD	0.17 ± 0.01	0.31 ± 0.02	0.36 ± 0.04
S/C	0.13 ± 0.03	0.19 ± 0.03	0.26 ± 0.03
SD	0.17 ± 0.02	0.33 ± 0.03	0.48 ± 0.04

Table E.1: *Choice of Score-Based Algorithm.*

E.2 Effect of Latent Mixing Variables

We are also interested in how exactly the presence of latent mixing variables influences the graphs \mathcal{G}' returned by causal discovery methods unaware of mixing. As the latents \mathcal{Z} introduce dependencies between the mixing targets \mathcal{T} , we presume that the reported \mathcal{G}' will contain additional spurious (FP) edges.

In Fig. E.1 (light gray) we observe that the FPR in \mathcal{G}' indeed increases as the probability $p_{\mathcal{Z}}$ of $X_j \in \mathcal{T}$ increases (shown for SCORE and CAM). Given this trend, we investigate whether we can correct the result by pruning any FPs that arise from mixing. Thus, we apply the CMM to each node X_j given its causes in \mathcal{G}' , fit an MLR, and use the $\text{BIC}_{\mathcal{Z}}^{\text{EM}}$ to remove any redundant parents of X_j under this model. This results in a graph \mathcal{G}'' , also shown in Fig. E.1 (dark gray).

The shaded regions show to which extent FP edges are removed correctly (green) resp. TP edges removed incorrectly (red). The $\text{BIC}_{\mathcal{Z}}^{\text{EM}}$ prunes some of the spurious and almost no causal edges. However, there still remain additional FPs in \mathcal{G}'' when $p_{\mathcal{Z}}$ increases. This is perhaps due to practical limitations of EM in estimating the correct mixing, leaving room for future improvements.

We reach a similar conclusion from our questions in Sections E.1 and E.2 that we can expect not a substantial, but at least some improvement in discovering causal graphs \mathcal{G} using the latent-aware BIC, be it via correcting the outputs of causal discovery algorithms (E.2), as an improved scoring criterion in GES (E.1), or with our main algorithm (Fig. D.2), while in addition being able to discover the latent structure $\mathcal{G}^{\mathcal{Z}}$ which can point us to subsamples with distinct causal generating process.

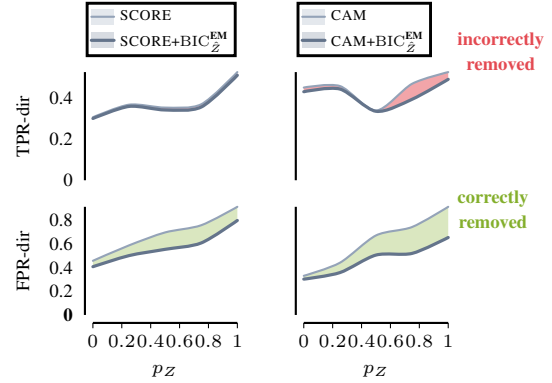


Figure E.1: *Effect of Latent Mixing.*

References

- Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, July 2009. ISBN 978-0-262-01319-2.
- Kenneth A. Bollen. Structural Equation Models with Observed Variables. In *Structural Equations with Latent Variables*, chapter Four, pages 80–150. John Wiley & Sons, Ltd, 1989. ISBN 978-1-118-61917-9.
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.
- David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.
- Christopher Meek. Graphical Models: Selecting causal and statistical models. 1 1997. doi: 10.1184/R1/22696393.v1. URL https://kilthub.cmu.edu/articles/thesis/Graphical_Models_Selecting_causal_and_statistical_models/22696393.
- Sascha Xu, Sarah Mameche, and Jilles Vreeken. Information-theoretic causal discovery in topological order. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL <https://openreview.net/forum?id=9pjJXQWYXc>.
- Weronika Ormaniec, Scott Sussex, Lars Lorch, Bernhard Schölkopf, and Andreas Krause. Standardizing Structural Causal Models. *arXiv e-prints*, art. arXiv:2406.11601, June 2024. doi: 10.48550/arXiv.2406.11601.
- Karen Sachs, Omar Perez, Dana Pe’er, Douglas Lauffenburger, and Garry Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, pages 523–9, 2005.
- Yuhao Wang, Liam Solus, Karren Dai Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 58245833, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- Abhinav Kumar, Kirankumar Shiragur, and Caroline Uhler. Learning mixtures of unknown causal interventions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=aC9mB1PqYJ>.
- Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010. URL <http://jmlr.org/papers/v11/vinh10a.html>.
- Jonas Wahl and Jakob Runge. Separation-based distance measures for causal graphs. *arXiv e-prints*, art. arXiv:2402.04952, February 2024. doi: 10.48550/arXiv.2402.04952.
- Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. 2022.