An Efficient Search-and-Score Algorithm for Ancestral Graphs using Multivariate Information Scores for Complex Non-linear and Categorical Data

Nikita Lagrange [1] **Hervé Isambert** [1]

Abstract

We propose a greedy search-and-score algorithm for ancestral graphs, which include directed as well as bidirected edges, originating from unobserved latent variables. The normalized likelihood score of ancestral graphs is estimated in terms of multivariate information over relevant "ac-connected subsets" of vertices, $\boldsymbol{C}$, that are connected through collider paths confined to the ancestor set of $\boldsymbol{C}$. For computational efficiency, the proposed two-step algorithm relies on local information scores limited to the close surrounding vertices of each node (step 1) and edge (step 2). This computational strategy, although restricted to information contributions from ac-connected subsets containing up to two-collider paths, is shown to outperform state-of-the-art causal discovery methods on challenging benchmark datasets.

1. Introduction

The likelihood function plays a central role in the selection of a graphical model $\mathcal{G}$ based on observational data $\mathcal{D}$. Given N independent samples from $\mathcal{D}$, the likelihood $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$ that they might have been generated by the graphical model $\mathcal{G}$ is given by (Koller & Friedman, 2009),

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = \frac{1}{Z_{\mathcal{D},\mathcal{G}}} \exp\Big(-NH(p,q)\Big) \tag{1}$$

where $H(p,q) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x})$ is the cross-entropy between the empirical probability distribution $p(\boldsymbol{x})$ of the observed data $\mathcal{D}$ and the theoretical probability distribution $q(\boldsymbol{x})$ of the model $\mathcal{G}$ and $Z_{\mathcal{D},\mathcal{G}}$ a data- and model-dependent factor ensuring proper normalization condition for finite dataset. In short, Eq.1 results from the asymptotic probability that the N independent samples, $\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}$, are drawn from the model distribution, $q(\boldsymbol{x})$, *i.e.* $\mathcal{L}_{\mathcal{D}|\mathcal{G}} \equiv q(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}) = \prod_i q(\boldsymbol{x}^{(i)})$, rather than the empirical distribution, $p(\boldsymbol{x})$. This leads to, $\log \mathcal{L}_{\mathcal{D}|\mathcal{G}} = \sum_i \log q(\boldsymbol{x}^{(i)})$, which converges towards $N \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x}) = -N\, H(p,q)$ in the large sample size limit, $N \to \infty$, with $\log Z_{\mathcal{D},\mathcal{G}} = \mathcal{O}(\log N)$.

The structural constraints of the model $\mathcal{G}$ translate into the factorization form of the theoretical probability distribution, $q(\boldsymbol{x})$ (Pearl & Paz, 1985; Pearl, 1988; 2009; Spirtes et al., 2000; Richardson, 2009). In particular, the probability distribution of Bayesian networks (BN) factorizes in terms of conditional probabilities of each variable given its parents, as $q_{\text{BN}}(\boldsymbol{x}) = \prod_i q(x_i|\mathbf{pa}_{X_i})$, where $\mathbf{pa}_{X_i}$ denote the values of the parents of node X_i in $\mathcal{G}$, $\mathbf{Pa}_{X_i}$. For Bayesian networks, the factors of the model distribution, $q(x_i|\mathbf{pa}_{X_i})$, can be directly estimated with the empirical conditional probabilities of each node given its parents as, $q(x_i|\mathbf{pa}_{X_i}) \equiv p(x_i|\mathbf{pa}_{X_i})$, leading to the well known estimation of the likelihood function in terms of conditional entropies $H(X_i|\mathbf{Pa}_{X_i}) = -\sum_{\boldsymbol{x}} p(x_i, \mathbf{pa}_{X_i}) \log p(x_i|\mathbf{pa}_{X_i})$,

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\text{BN}}} = \frac{1}{Z_{\mathcal{D},\mathcal{G}_{\text{BN}}}} \exp\Big(-N \sum_{X_i \in \boldsymbol{V}}^{\text{vertices}} H(X_i|\mathbf{Pa}_{X_i})\Big) \tag{2}$$

This paper concerns the experimental setting for which some variables of the underlying Bayesian model are not observed. This frequently occurs in practice for many applications. We derive an explicit likelihood function for the class of ancestral graphs, which include directed as well as bidirected edges, arising from the presence of unobserved latent variables. Tian and Pearl 2002 showed that the probability distribution of such graphs factorizes into c-components including subsets of variables connected through bidirected paths (*i.e.* containing only bidirected edges). Richardson 2009 later proposed a refined factorization of the model distribution of the broader class of acyclic directed mixed graphs in terms of conditional probabilities over "head" and "tail" subsets of variables within each ancestrally closed subsets of vertices. However, unlike with Bayesian networks, the contributions of c-components or head-and-tail factors to the likelihood function cannot simply be estimated in terms of empirical distribution $p(\boldsymbol{x})$, as shown below. This

[1]CNRS, UMR168, Institut Curie, PSL University, Sorbonne University, Paris, France. Correspondence to: Hervé Isambert <herve.isambert@curie.fr>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

leaves the likelihood function of ancestral graphs difficult to estimate from empirical data, in general, although iterative methods have been developped when the data is normally distributed (Richardson & Spirtes, 2002; Drton et al., 2009; Evans & Richardson, 2010; Triantafillou & Tsamardinos, 2016; Rantanen et al., 2021; Claassen & Bucur, 2022).

The present paper provides an explicit decomposition of the likelihood function of ancestral graphs in terms of multivariate cross-information over relevant 'ac-connected' subsets of variables, Fig. 1., which do not rely on the head-and-tail factorization but coincide with the parametrizing sets (Hu & Evans, 2020) derived from the head-and-tail factorization. It suggests a natural estimation of these revelant contributions to the likelihood function in terms of empirical distribution $p(\boldsymbol{x})$. This result extends the likelihood expression of Bayesian Networks (Eq. 2) to include the effect of unobserved latent variables and enables the implementation of a greedy search-and-score algorithm for ancestral graphs. For computational efficiency, the proposed two-step algorithm relies on local information scores limited to the surrounding vertices of each node (step 1) and edge (step 2). This computational strategy is shown to outperform state-of-the-art causal discovery methods on challenging benchmarks. This paper extends a manuscript opted-in at NeurIPS 2024 (Lagrange & Isambert, 2024) to include benchmarks on linear as well as non-linear continuous models, Figs. 2 & E.2.

2. Theoretical Results

2.1. Multivariate Cross-entropy and Cross-information

The theoretical result of the paper (Theorem 1) is expressed in terms of multivariate cross-information derived from multivariate cross-entropies through the Inclusion-Exclusion Principle. The same expressions can be written between multivariate information and multivariate entropies by simply substituting $q(\{x_i\})$ with $p(\{x_i\})$ in the equations below and will be used to estimate the likelihood function of ancestral graphs (Proposition 3).

As recalled above, the cross-entropy between m variables, $\boldsymbol{V} = \{X_1, \cdots, X_m\}$, is defined as,

$$H(\boldsymbol{V}) \;=\; -\sum_{\{x_i\}} p(x_1, \cdots, x_m) \log q(x_1, \cdots, x_m) \quad (3)$$

where $p(\{x_i\})$ is the empirical joint probability distribution of the variables $\{X_i\}$ and $q(\{x_i\})$ the joint probability distribution of the model. Bayes formula, $q(\{x_i\}, \{y_j\}) = q(\{x_i\}|\{y_j\})\; q(\{y_j\})$, directly translates into the definition of conditional cross-entropy through the decomposition,

$$H(\{X_i\}, \{Y_j\}) = H(\{X_i\}|\{Y_j\}) + H(\{Y_j\}) \quad (4)$$

Multivariate (cross) information, $I(\boldsymbol{V}) \equiv I(X_1; \cdots; X_m)$, are defined from multivariate (cross) entropies through Inclusion-Exclusion formulas over all subsets of variables (McGill, 1954; Ting, 1962; Han, 1980; Yeung, 1991) as,

$$\begin{aligned} I(X) &= H(X) \\ I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ I(\boldsymbol{V}) &= -\sum_{\boldsymbol{S}\subseteq\boldsymbol{V}} (-1)^{|\boldsymbol{S}|} H(\boldsymbol{S}) \end{aligned} \quad (5)$$

where the semicolon separators are needed to distinguish multipoint (cross) information from joint variables as in $I(\{X,Z\};Y) = I(X;Y) + I(Z;Y) - I(X;Y;Z)$. Below, implicit separators between non-conditioning variables in multivariate (cross) information will always correspond to semicolons, *e.g.* as in $I(\boldsymbol{V})$ in Eq. 5. Unlike multivariate (cross) entropies, which are always positive, $H(X_1, \cdots, X_k) \geqslant 0$, multivariate (cross) information, $I(X_1; \cdots; X_k)$, can be positive or negative for $k \geqslant 3$, while they remain always positive for $k < 3$, *e.g.* $I(X;Y) \geqslant 0$.

In turn, multivariate (cross) entropies can be expressed through the Principle of Inclusion-Exclusion into the same expression but in terms of multivariate (cross) information,

$$H(\boldsymbol{V}) = -\sum_{\boldsymbol{S}\subseteq\boldsymbol{V}} (-1)^{|\boldsymbol{S}|} I(\boldsymbol{S}), \quad (6)$$

Conditional multivariate (cross) information $I(\boldsymbol{V}|Z)$ are defined similarly as multivariate (cross) information $I(\boldsymbol{V})$ but in terms of conditional (cross) entropies as,

$$I(\boldsymbol{V}|Z) \;=\; -\sum_{\boldsymbol{S}\subseteq\boldsymbol{V}} (-1)^{|\boldsymbol{S}|} H(\boldsymbol{S}|Z) \quad (7)$$

Eqs. 5 & 7 lead to a decomposition rule relative to a variable Z, Eq. 8, which can be conditioned on a set of joint variables, $\boldsymbol{A} = \{A_1, \cdots, A_m\}$, with implicit comma separators,

$$I(\boldsymbol{V}) \;=\; I(\boldsymbol{V}|Z) + I(\boldsymbol{V};Z) \quad (8)$$

$$I(\boldsymbol{V}|\boldsymbol{A}) \;=\; I(\boldsymbol{V}|Z,\boldsymbol{A}) + I(\boldsymbol{V};Z|\boldsymbol{A}) \quad (9)$$

Alternatively, conditional (cross) information, such as $I(X;Y|\boldsymbol{A})$, can be expressed in terms of non-conditional (cross) entropies using Eq. 4,

$$\begin{aligned} I(X;Y|\boldsymbol{A}) &= H(X|\boldsymbol{A}) + H(Y|\boldsymbol{A}) - H(X,Y|\boldsymbol{A}) \\ &= H(X,\boldsymbol{A}) + H(Y,\boldsymbol{A}) - H(X,Y,\boldsymbol{A}) - H(\boldsymbol{A}) \end{aligned} \quad (10)$$

which can in turn be expressed in terms of non-conditional (cross) information as,

$$\begin{aligned} I(X;Y|\boldsymbol{A}) = I(X;Y) &- \cdots (-1)^k \sum_{i_1<\cdots<i_k} I(X;Y;A_{i_1};\cdots;A_{i_k}) \\ &+ \cdots (-1)^m I(X;Y;A_1;\cdots;A_m) \\ &= \sum_{\boldsymbol{S}'\subseteq\boldsymbol{S}}^{X,Y\in\boldsymbol{S}'} (-1)^{|\boldsymbol{S}'|} I(\boldsymbol{S}'), \end{aligned} \quad (11)$$

where $\boldsymbol{S} = \{X,Y\} \cup \boldsymbol{A}$. This corresponds to *all (cross) information terms including both X and Y* in the expression of the multivariate (cross) entropy, $H(X,Y,\boldsymbol{A})$, Eq. 6.

2.2. Graphs and Connection Criteria

2.2.1. Directed mixed graphs, ancestral graphs

Two vertices are said to be **adjacent** if there is an edge (of any type) between them, $X*\!\!-\!\!*Y$, where $*$ stands for any (head or tail) end mark. X and Y are said to be **neighbors** if $X - Y$, **parent** and **child** if $X \to Y$ and **spouses** if $X \longleftrightarrow Y$ in $\mathcal{G}$.

A **path** in $\mathcal{G}$ is a sequence of distinct vertices $V_1, \ldots, V_n$ consecutively adjacent in $\mathcal{G}$, as, $V_1 *\!\!-\!\!* V_2 *\!\!-\!\!* \cdots *\!\!-\!\!* V_{n-1} *\!\!-\!\!* V_n$. In particular, a **collider path** between V_1 and V_n has the form $V_1 *\!\!\to V_2 \longleftrightarrow \cdots \longleftrightarrow V_{n-1} \leftarrow\!\!* V_n$ and a **directed path** corresponds to $V_1 \to V_2 \to \cdots \to V_n$.

X is called an **ancestor** of Y and Y a **descendant** of X if $X = Y$ or there is a **directed path** from X to Y, $X \to \cdots \to Y$. $\mathbf{An}_{\mathcal{G}}(Y)$ denotes the **set of ancestors** of Y in $\mathcal{G}$. By extension, for any subset of vertices, $\boldsymbol{C} \subseteq \boldsymbol{V}$, $\mathbf{An}_{\mathcal{G}}(\boldsymbol{C})$ denotes the set of ancestors for all $Y \in \boldsymbol{C}$ in $\mathcal{G}$.

A **directed mixed graph** (**DMG**) is a vertex-edge graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ that can contain two types of edges: directed ($\to$) and bidirected ($\longleftrightarrow$) edges.

A **directed cycle** occurs if $X \in \mathbf{An}_{\mathcal{G}}(Y)$ and $X \leftarrow Y$. An **almost directed cycle** occurs if $X \in \mathbf{An}_{\mathcal{G}}(Y)$ and $X \leftrightarrow Y$.

Definition 1. An **ancestral graph** is a **DMG**:
- *i)* without directed cycles;
- *ii)* without almost directed cycles.

An **ancestral graph** is said to be **maximal** if every missing edge corresponds to a structural independence. If an ancestral graph $\mathcal{G}$ is not maximal, there exists a unique maximal ancestral graph $\bar{\mathcal{G}}$ by adding bidirected edges to $\mathcal{G}$ (Richardson & Spirtes, 2002).

2.2.2. *ac*-connecting paths, *ac*-connected subsets

Let us now define **ancestor collider connecting paths** or ***ac*-connecting paths**, which entail simpler path connecting criterion than the traditional **m-connecting criterion**, discussed in the Appendix A. Yet, ***ac*-connecting paths** and ***ac*-connected subsets** will turn out to be directly relevant to characterize the likelihood decomposition and Markov equivalent classes of ancestral graphs.

Definition 2. [ac-connecting path] An ac-connecting path between X and Y given a subset of variables $\boldsymbol{C}$ (possibly including X and Y) is a collider path, $X * \!\!\to Z_1 \longleftrightarrow \cdots \longleftrightarrow Z_K \leftarrow\!\!* Y$, with all $Z_i \in \mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{C})$, that is, with Z_i in $\boldsymbol{C}$ or connected to $\{X, Y\} \cup \boldsymbol{C}$ by an ancestor path, *i.e.* $Z_i \to \cdots \to T$ with $T \in \{X, Y\} \cup \boldsymbol{C}$.

Definition 3. [ac-connected subset] A subset $\boldsymbol{C}$ is said to be ac-connected if $\forall X, Y \in \boldsymbol{C}$, X and Y are connected (through any type of edge) or there is an ac-connecting path between X and Y given $\boldsymbol{C}$.

2.3. Likelihood Decomposition of Ancestral Graphs

Theorem 1. [**likelihood of ancestral graphs**] *The cross-entropy $H(p, q)$ and likelihood $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$ of an ancestral graph $\mathcal{G}$ is decomposable in terms of multivariate cross-information, $I(\boldsymbol{C})$, summed over all ac-connected subsets of variables, $\boldsymbol{C}$ (Definition 3),*

$$
\begin{aligned}
H(p, q) &= -\sum_{\boldsymbol{C} \subseteq \boldsymbol{V}}^{\text{ac-connected}} (-1)^{|\boldsymbol{C}|} I(\boldsymbol{C}) \\
\mathcal{L}_{\mathcal{D}|\mathcal{G}} &= \frac{1}{Z_{\mathcal{D},\mathcal{G}}} \exp\Big(N \sum_{\boldsymbol{C} \subseteq \boldsymbol{V}}^{\text{ac-connected}} (-1)^{|\boldsymbol{C}|} I(\boldsymbol{C})\Big) \quad (12)
\end{aligned}
$$

where N is the number of iid samples in the dataset $\mathcal{D}$ and $Z_{\mathcal{D},\mathcal{G}}$ a data- and model-dependent normalization constant.

The proof of Theorem 1 is left to Appendix B. It is based on a partition of the cross-entropy (Eq. 6) into cross-information contributions from ac-connected and non-ac-connected subsets of variables, which does not rely on head-and-tail factorizations nor on imset formalism[1] Hu and Evans proposed an equivalent result (Proposition 3.3 in (Hu & Evans, 2020)) with a proof using head-and-tail decomposition to define parametrizing sets, which happen to coincide with the ac-connected sets defined here (Definition 3). Theorem 1 characterizes in particular the Markov equivalence class of ancestral graphs (Richardson & Spirtes, 2002; Richardson, 2003; Ali & Richardson, 2002; Ali et al., 2005; Tian, 2005; Ali et al., 2009) as,

Corollary 2. *Two ancestral graphs are Markov equivalent iff they have the same ac-connected subsets of vertices.*

Note, in particular, that Eq. 12 holds for *maximal ancestral graphs* (MAG), for which all pairs of ac-connected variables are connected by an edge, and their Markov equivalent representatives, the *partial ancestral graphs* (PAG) (Richardson & Spirtes, 2002; 1999; Zhang, 2007; 2008).

Proposition 3. The likelihood decomposition of ancestral graphs (Eq. 12, Theorem 1) can be estimated by replacing the model distribution q by the empirical distribution p in the retained multivariate cross-information terms $I(\boldsymbol{C})$ corresponding to all *ac*-connected subsets of variables, $\boldsymbol{C}$.

Hence, Proposition 3 amounts to estimating all relevant cross-information terms in the likelihood function with the corresponding multivariate information terms computed from the available data, while assuming by construction that the model distribution obeys all local and global condi-

[1]The genesis of Theorem 1 and Proposition 3 is further discussed in `https://openreview.net/forum?id=Z2f4Laqi8U&noteId=8GLWeaAKc9`.

tional independences entailed by the ancestral graph. The corresponding factorization of the model distribution can be expressed in terms of empirical distribution, assuming positive distributions, see Appendix C.

Fig. 1 illustrates the cross-entropy decomposition for a few graphical models in terms of cross-information contributions from their ac-connected subsets of vertices. In particular, an unshielded non-collider (*e.g.* $X \to Z \to W$, Fig. 1A), is less likely (*i.e.* higher cross-entropy) than an unshielded collider or 'v-structure' (*e.g.* $X \to Z \leftarrow W$, Fig. 1B), if the corresponding three-point information term is negative, $I(X;Z;W) < 0$, in agreement with earlier results (Affeldt & Isambert, 2015; Verny et al., 2017). However, this early approach, exploiting the sign and magnitude of three-point information to orient v-structures, does not include higher order terms involving multiple v-structures, which can lead to orientation conflicts between unshielded triples, in practice. Resolving such orientation conflicts requires to include information contributions from higher-order ac-connected subgraphs, such as star-like ac-connected subsets including three or more parents, Fig. 1C. Similarly, the cross-entropies of collider paths involving several colliders also include higher-order terms, as with the simple example of a two-collider path, Fig. 1D. By contrast, the cross-entropy based on the head-and-tail factorization of the same two-collider path, *i.e.* $q(x,z,y,w) = q(z,y|x,w)q(x)q(w)$ (Richardson, 2009), is found to be equivalent to the cross-entropy of a Bayesian graph without bidirected edge, Fig. 1E, when estimated with the empirical distribution $p(.)$, see Appendix C. This observation illustrates the difficulty to estimate the likelihood functions of ancestral graphs using head-and-tail factorization.

Further examples of graphical models, Figs. 1F-I, show the relative simplicity of the decomposition with only few (non-trivial) ac-connected contributing subsets $\boldsymbol{C}$ with $|\boldsymbol{C}| \geqslant 3$, as compared to the much larger number of non-ac-connected non-contributing subsets, that cancel each other by construction due to conditional independence constraints of the underlying model. Note, in particular, that most contributing multivariate information $I(\boldsymbol{C})$ only concern direct connections or collider paths within a single component subgraph induced by $\boldsymbol{C}$ (solid line edges in Fig. 1). However, occasionally, collider paths extending beyond $\boldsymbol{C}$ into $\mathbf{An}_{\mathcal{G}}(\boldsymbol{C}) \setminus \boldsymbol{C}$ (marked with wiggly edges) with corresponding ancestor path(s) (marked with dashed edges) do occur, as shown in Fig. 1G.

In addition, the present information-theoretic decomposition of the likelihood of ancestral graphs can readily distinguish their Markov equivalence classes according to Corollary 2. For instance, the ancestral graphs of Fig. 1F and Fig. 1G, despite sharing the same edges and the same unshielded collider ($X \to Z \leftarrow T$), turn out not to be Markov equivalent, as discussed in (Ali et al., 2009). Indeed, their cross-entropy decompositions differ by two ac-connected contributing terms: a three-point cross information $I(X;Y;T)$ with a collider path not confined in $\boldsymbol{C}$ (*i.e.* $X \rightsquigarrow Z \nleftrightsquigarrow T \longleftrightarrow Y$ and corresponding ancestor path $Z \dashrightarrow Y$) and a four-point information term $I(X;Y;Z;T)$ due to the two-collider path ($X \to Z \longleftrightarrow T \longleftrightarrow Y$). More quantitatively, it shows that the graph of Fig. 1G with a two-collider path is more likely than the graph of Fig. 1F whenever $I(X;Y;T) - I(X;Y;Z;T) = I(X;Y;T|Z) = I(X;Y|Z) - I(X;Y|Z,T) < 0$. Finally, the Markov equivalent graphs of Fig. 1H and Fig. 1I, also due to (Ali et al., 2009), illustrate the fact that the actual ancestor collider path between unconnected pairs does not need to be unique nor conserved between Markov equivalent graphs (as long as their cross-entropies share the same multivariate cross-information decomposition).

3. Efficient Search-and-Score Causal Discovery using Local Information Scores

The likelihood estimation of ancestral graphs (Theorem 1 and Proposition 3) enables the implementation of a search-and-score algorithm for this broad class of graphs, which has attracted a number of contributions recently (Triantafillou & Tsamardinos, 2016; Rantanen et al., 2021; Claassen & Bucur, 2022; Andrews et al., 2022; Hu & Evans, 2024b;a). Our specific objective is not to develop an exact method limited to simple graphical models with a few nodes and small datasets but to implement an efficient and reliable heuristic method applicable to more challenging graphical models and large datasets.

Indeed, search-and-score structure learning methods need to rely on heuristic rather than exhaustive search, in general, given that the number of ancestral graphs grows super-exponentially as the number of vertices increases. This can be implemented for instance with a Monte Carlo algorithmic scheme with random restarts, which efficiently probes relevant graphical models. Here, we opt, instead, to use the prediction of an efficient hybrid causal discovery method, MIIC (Verny et al., 2017; Cabeli et al., 2021; Ribeiro-Dantas et al., 2024), as starting point for a subsequent search-and-score approach based on the proposed likelihood estimation of ancestral graphs (Eq. 12 and Proposition 3).

Moreover, while the likelihood decomposition of ancestral graphs may involve extended ac-connected subsets of variables, as illustrated in Fig. 1, we aim to implement a computationally efficient search-and-score causal discovery method based on approximate local scores limited to the close surrounding vertices of each node and edge. Yet, while MIIC only relies on unshielded triple scores, the novel search-and-score extension, MIIC_search&score, uses also higher-order local information scores to compare alternative

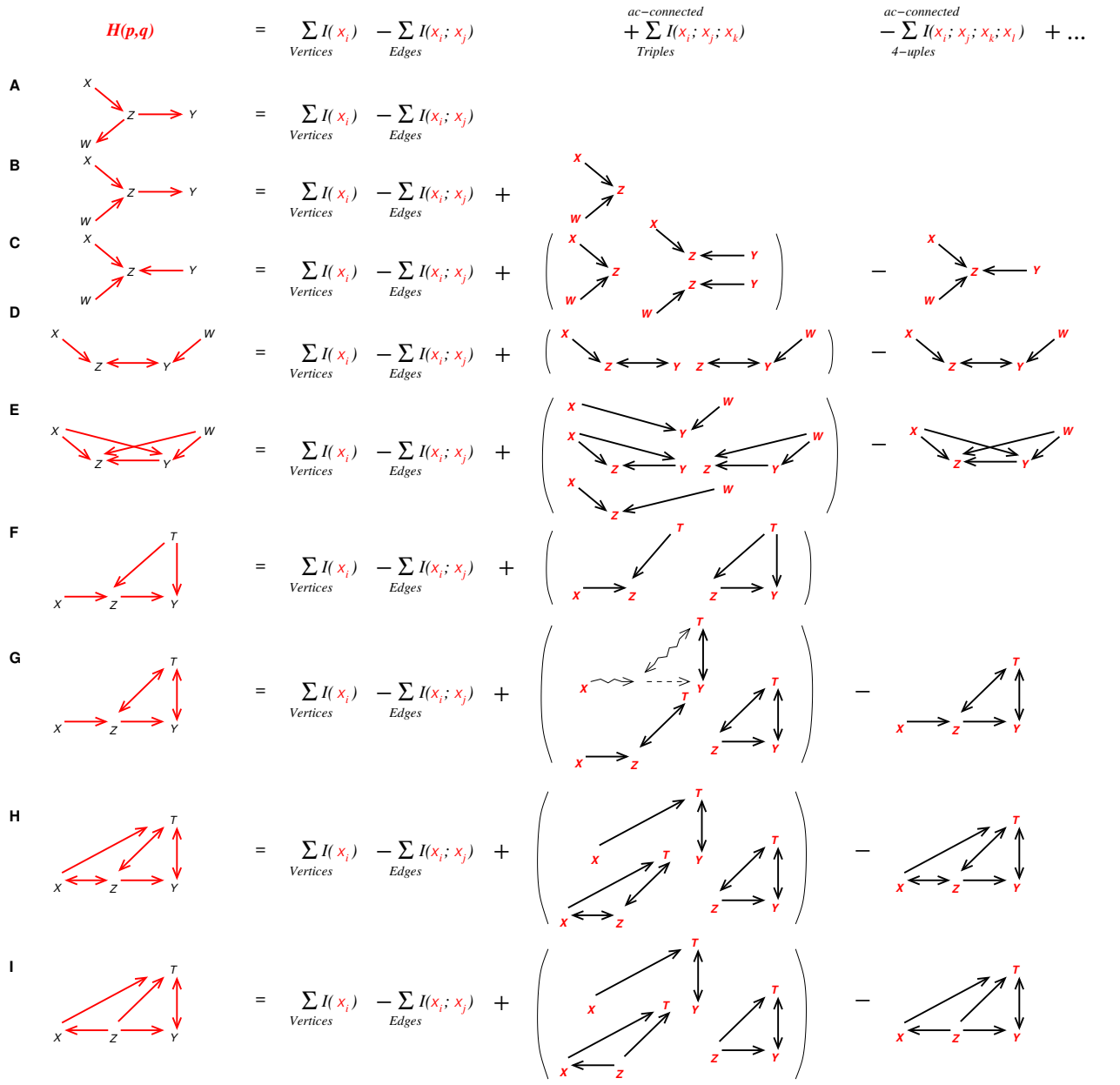

Figure 1. **Cross-entropy decomposition of ancestral graphs.** Examples of cross-entropy decomposition of ancestral graphs (red edges, lhs) in terms of relevant multivariate cross-information contributions $I(\boldsymbol{C})$ with $\boldsymbol{C} \subseteq \boldsymbol{V}$ (red nodes, rhs). Simple graphs: (**A**) without unshielded colliders, (**B**) with a single or non-overlapping unshielded colliders, (**C**) with overlapping unshielded colliders through three or more (conditionally) independent parents or (**D**) through a two-(or more)-collider path. (**E**) Bayesian graph corresponding to the head-and-tail factorization of the two-collider path in (**D**) estimated using the empirical distribution $p(.)$, see Appendix C. (**F**) Simple Bayesian graph not Markov equivalent to an ancestral graph (**G**) sharing the same edges and unshielded collider (Ali et al., 2009). Solid black edges correspond to direct connections or collider paths confined to the corresponding ac-connected subset $\boldsymbol{C}$, while wiggly edges indicate collider paths extending beyond $\boldsymbol{C}$ yet indirectly connected to $\boldsymbol{C}$ by an ancestor path, marked with dashed edges, see Definition 2. By contrast, graphs **H** and **I** illustrate the fact that collider paths may not be unique nor conserved between two Markov equivalent graphs (*i.e.* sharing the same cross-information terms) (Ali et al., 2009).

subgraphs, as detailed below.

The proposed method is shown to outperform MIIC and other state-of-the-art causal discovery methods on challenging datasets including latent variables.

3.1. MIIC, an Hybrid Causal Discovery Method based on Unshielded Triple Scores

MIIC is an hybrid causal discovery method combining constraint-based and information-theoretic frameworks

(Verny et al., 2017; Cabeli et al., 2020). Unlike traditional constraint-based methods (Pearl, 2009; Spirtes et al., 2000), MIIC does not directly attempt to uncover conditional independences but, instead, iteratively substracts the most significant three-point (conditional) information contributions of successive contributors, A_1, A_2, ..., A_n, from the mutual information between each pair of variables, $I(X;Y)$,

$$I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1) - \cdots \\ \cdots - I(X;Y;A_n|\{A_i\}_{n-1}) = I(X;Y|\{A_i\}_n) \quad (13)$$

where $I(X;Y;A_k|\{A_i\}_{k-1}) > 0$ is the *positive* information contribution from A_k to $I(X;Y)$ (Affeldt & Isambert, 2015; Affeldt et al., 2016). Conditional independence is eventually established when the residual conditional mutual information on the right hand side of Eq. 13, $I(X;Y|\{A_i\}_n)$, becomes smaller than a complexity term, *i.e.* $k_{X;Y|\{A_i\}}(N) \geqslant I(X;Y|\{A_i\}_n) \geqslant 0$, which dependents on the considered variables and sample size N.

This leads to an undirected skeleton, which MIIC then (partially) orients based on the sign and amplitude of the regularized conditional 3-point information terms (Affeldt & Isambert, 2015; Verny et al., 2017). In particular, negative conditional 3-point information terms, $I(X;Y;Z|\{A_i\}) < 0$, correspond to the signature of causality in observational data (Affeldt & Isambert, 2015) and lead to the prediction of a v-structure, $X \to Z \leftarrow Y$, if X and Y are not connected in the skeleton. By contrast, a positive conditional 3-point information term, $I(X;Y;Z|\{A_i\}) > 0$, implies the absence of a v-structure and suggests to propagate the orientation of a previously directed edge $X \to Z - Y$ as $X \to Z \to Y$.

In practice, MIIC's strategy to circumvent spurious conditional independences significantly improves recall, that is, the fraction of correctly recovered edges, compared to traditional constraint-based methods. Yet, MIIC only relies on unshielded triple scores to reliably uncover significant contributors and orient v-structures, as outlined above. MIIC has been recently improved to ensure the consistency of the separating set in terms of indirect paths in the final skeleton or (partially) oriented graphs (Li et al., 2019; Ribeiro-Dantas et al., 2024) and to improve the reliably of predicted orientations (Cabeli et al., 2021; Ribeiro-Dantas et al., 2024).

The predictions of this recent version of MIIC, which include 3 types of edges (directed, bidirected and undirected), have been used as starting point for the subsequent local search-and-score method implemented in the present paper.

3.2. New Search-and-Score Method based on Higher-Order Local Information Scores

Starting from the structure predicted by MIIC, as detailed above, MIIC_search&score proceeds in two steps, first to remove likely false positive edges (Step 1) and then to orient the remaining edges based on their estimated contributions to the global likelihood decomposition, Eq. 12 (Step 2). The two steps are illustrated on a running example, Fig. D.1.

3.2.1. Step 1: Node scores for edge orientation priming and edge removal

The first step consists in minimizing a node score corresponding to the local normalized log likelihood of each node w.r.t. its possible parents or spouses amongst the connected nodes predicted by MIIC. To this end, the node score assesses the conditional entropy of each node w.r.t. a selection of parents, spouses or neighbors, $\mathbf{Pa}'_{X_i} \subseteq \mathbf{Pa}_{X_i} \cup \mathbf{Sp}_{X_i} \cup \mathbf{Ne}_{X_i}$, and a factorized Normalized Maximum Likelihood (fNML) regularization (Affeldt & Isambert, 2015) (conditional entropies can be estimated as differences of mutual information for continuous variables, Appendix D),

$$\mathrm{Score_n}(X_i) = H(X_i|\mathbf{Pa}'_{X_i}) + \frac{1}{N}\sum_{j}^{q_{x_i}} \log \mathcal{C}_{n_j}^{r_{x_i}} \quad (14)$$

where q_{x_i} corresponds to the combination of levels of $\mathbf{Pa}'_{X_i}$, while r_{x_i} is the number of levels of X_i, and n_j the number of samples corresponding to a particular combination of levels j in each summand, with $\sum_j n_j = N$, the total number of samples. $\log \mathcal{C}_{n_j}^{r_{x_i}}$ is the fNML regulatization cost summed over all combinations of levels, q_{x_i}, (Kontkanen & Myllymäki, 2007; Roos et al., 2008), see Appendix D.

This first algorithm is looped over each node, priming the orientations of their surrounding edges (as directed, bidirected or undirected), until convergence. Edges without orientation priming at either extremity are assumed to be false positive edges and removed at the end of Step 1.

3.2.2. Step 2: Edge orientation scores, as local contributions to the global likelihood

The second step consists in orienting the edges retained after Step 1, based on the optimization of their local contributions to the global likelihood score, Eq. 12, restricted to ac-connected subsets containing up to two-collider paths. This amounts to minimizing each edge orientation score w.r.t. its nodes' parents and spouses, corresponding to minus the conditional information plus a fNML complexity cost, Table 1, given three sets of parents and spouses of X and Y, *i.e.* $\mathbf{Pa}'_{X\setminus Y} = \mathbf{Pa}_X \cup \mathbf{Sp}_X \setminus Y$, $\mathbf{Pa}'_{Y\setminus X} = \mathbf{Pa}_Y \cup \mathbf{Sp}_Y \setminus X$ and $\mathbf{Pa}'_{XY} = \mathbf{Pa}'_{X\setminus Y} \cup \mathbf{Pa}'_{Y\setminus X}$ with their corresponding combinations of levels, $q_{y|x}$, $q_{x\setminus y}$ and q_{xy}. These orientation scores, Table 1, include symmetrized fNML complexity terms to enforce Markov equivalence, if X and Y share the same parents or spouses (excluding X and Y), see Appendix D. Indeed, all three scores become equals if $\mathbf{Pa}'_{Y\setminus X} = \mathbf{Pa}'_{X\setminus Y} = \mathbf{Pa}'_{XY}$ implying also the same combinations of parent and spouse levels, $q_{y|x} = q_{x\setminus y} = q_{xy}$.

While orientation scores cannot be summed over individual edges due to multiple countings of ac-connected contri-

Table 1. Local scores for the orientation of a single directed or bidirected edge, see Appendix D.

Edge	Information	Symmetrized fNML complexity (Markov equivalent)
$X \to Y$	$-I(X;Y\|\mathbf{Pa}'_{Y\setminus X})$	$\frac{1}{2N}\left(\sum_j^{q_{x\setminus y}{r_y}} \log \mathcal{C}_{n_j}^{r_x} - \sum_j^{q_{x\setminus y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\setminus x}{r_x}} \log \mathcal{C}_{n_j}^{r_y} - \sum_j^{q_{y\setminus x}} \log \mathcal{C}_{n_j}^{r_y}\right)$
$X \leftarrow Y$	$-I(X;Y\|\mathbf{Pa}'_{X\setminus Y})$	$\frac{1}{2N}\left(\sum_j^{q_{x\setminus y}{r_y}} \log \mathcal{C}_{n_j}^{r_x} - \sum_j^{q_{x\setminus y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\setminus x}{r_x}} \log \mathcal{C}_{n_j}^{r_y} - \sum_j^{q_{y\setminus x}} \log \mathcal{C}_{n_j}^{r_y}\right)$
$X \leftrightarrow Y$	$-I(X;Y\|\mathbf{Pa}'_{XY})$	$\frac{1}{2N}\left(\sum_j^{q_{xy}{r_y}} \log \mathcal{C}_{n_j}^{r_x} - \sum_j^{q_{xy}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{yx}{r_x}} \log \mathcal{C}_{n_j}^{r_y} - \sum_j^{q_{yx}} \log \mathcal{C}_{n_j}^{r_y}\right)$

butions, score differences between alternative orientations provide an estimate of the global score change. Hence, step 2 algorithm is looped over each edge to compute an orientation score decrement, given its current orientation and the orientations of surrounding edges. The orientation change corresponding to the largest global score decrement, without forming new directed or almost directed cycles, is then chosen at each iteration until convergence or until a limit cycle is reached. Limit cycles may originate from the local two-collider approximation of the global score.

4. Experimental Results

We first tested whether MIIC_search&score orientation scores (Table 1) effectively predicts bidirected orientations on three simple ancestral models, Fig. E.1 in Appendix E, when the end nodes do not share the same parents (Fig. E.1, Model 1), share some parents (Fig. E.1, Model 2) or when the bidirected edge is part of a longer than two-collider paths (Fig. E.1, Model 3). The prediction of the edge orientation scores are summarized in Table E.1 and show good predictions for large enough datasets.

We then tested MIIC_search&score performance on ancestral graphs obtained by hiding up to 20% of variables in either linear or non-linear continuous Bayesian networks of 50-150 nodes and average degree 3-5. Figs. 2 & E.2 compare MIIC_search&score performance to MIIC results used as starting point for MIIC_search&score and to M3HC (Triantafillou & Tsamardinos, 2016), GFCI (Ogarrio et al., 2016) and DAG-GNN (Yu et al., 2019), a DL method based on continuous constrained optimization. The parameter settings for each method are detailed in Appendix E.

While MIIC_search&score is outperformed on linear Gaussian models by causal discovery methods assuming linear combinations of the variables (with possible pre- or post-nonlinear transformations for DAG-GNN), MIIC_search&score clearly outperforms these methods on more complex models including non-linear couplings between variables, Figs. 2 & E.2. MIIC_search&score also outperforms MIIC in terms of edge precision with little to no decrease in edge recall, demonstrating the benefit of MIIC_search&score's rationale to improve MIIC predictions by extending MIIC information scores from unshielded triples to higher-order information contributions.

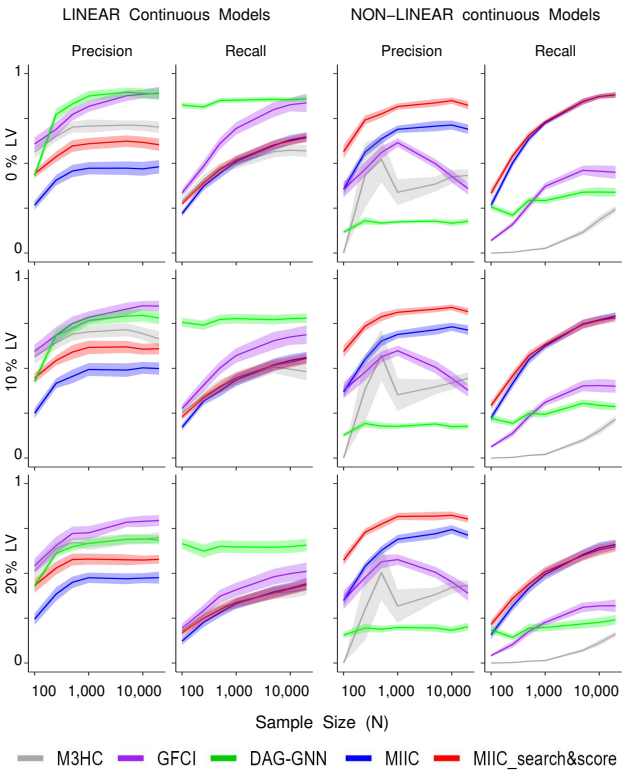

Figure 2. **Benchmark results on continuous datasets.** Benchmark results are averaged over 30 independent ancestral graph models obtained by hiding 0%, 10% or 20% of variables in linear Gaussian models (left) or in more complex models with non-linear couplings between variables (right), and including 50 nodes of average degree 5. MIIC_search&score results are compared to MIIC results used as starting point for MIIC_search&score, M3HC (Triantafillou & Tsamardinos, 2016), GFCI (Ogarrio et al., 2016) and DAG-GNN (Yu et al., 2019). See also Fig. E.2. Causal discovery performance is assessed in terms of *Precision* and *Recall* relative to the theoretical PAGs, while counting as false positive all correctly predicted edges but with a different orientation as the directed or bidirected edges of the PAG. Error bars: 95% confidence interval.

We also analyzed challenging benchmarks for categorical datasets from the bnlearn repository (Scutari, 2010). They concern ancestral graphs obtained by hiding up to 20% of variables in Discrete Bayesian Networks of increasing complexity (number of nodes and parameters), such as Alarm (37 nodes, 509 parameters), Insurance (27 nodes, 1,008 parameters), Barley (48 nodes, 84 links, 114,005 parameters), and Mildew (35 nodes, 540,150 parameters), Figs. 3 & E.3.

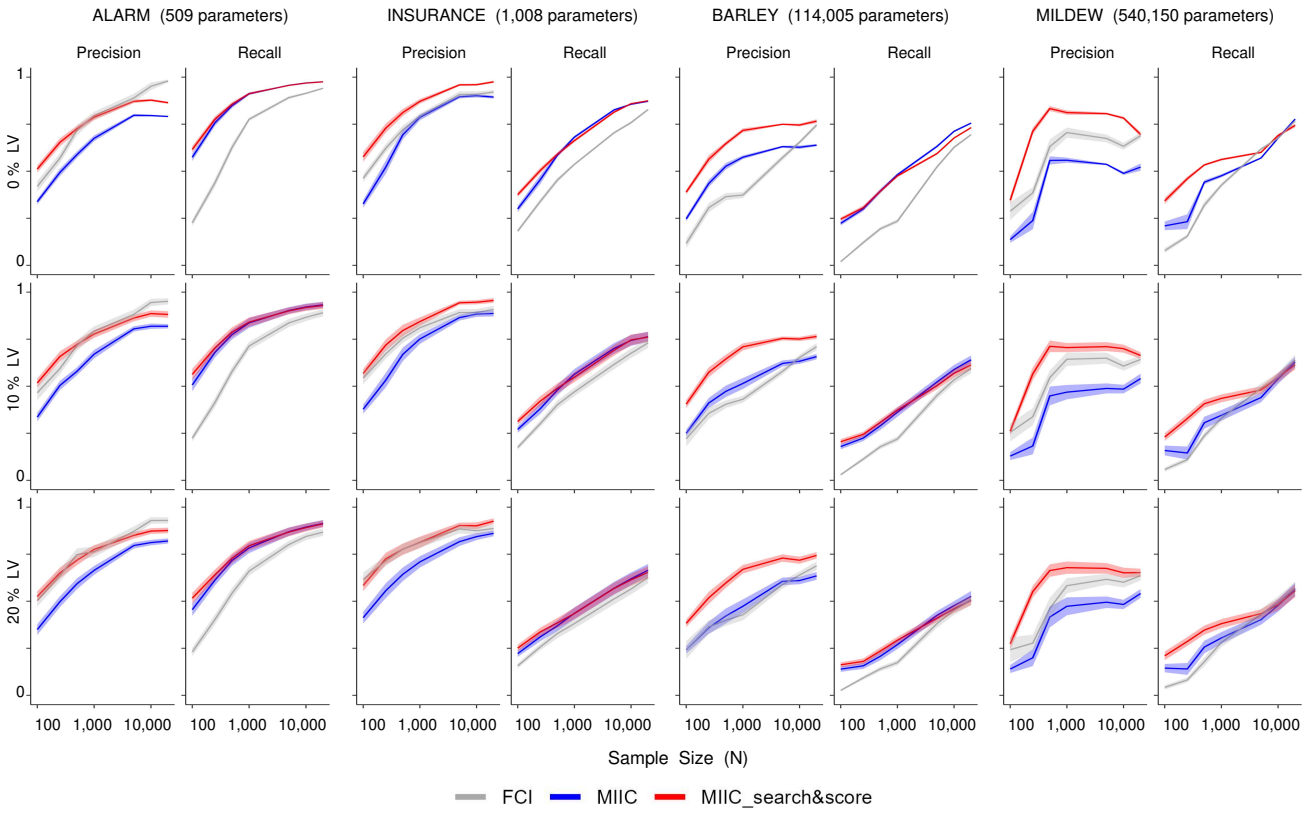

Figure 3. **Benchmark results on 'real-world' categorical datasets from the bnlearn repository.** Benchmark results are averaged over 50 independent categorical datasets from ancestral graphs obtained by hiding 0%, 10% or 20% of variables in Discrete Bayesian Networks of increasing complexity (see main text): Alarm, Insurance, Barley and Mildew. MIIC_search&score results are compared, as in Fig. 2, to MIIC results used as starting point for MIIC_search&score and to FCI (Zheng et al., 2024). Error bars: 95% confidence interval.

Figs. 3 & E.3 compare MIIC_search&score performance to MIIC results and to FCI from the python causal-learn package (Zheng et al., 2024). The parameter settings for each method are detailed in Appendix E. Fig. 3 results are obtained from independent datasets for each ancestral graph and sample size, while Fig. E.3 results provide a bootstrap sensitivity analysis to sampling noise for each method based on independent resamplings with replacement of single datasets of increasing size.

As with continuous data, MIIC_search&score outperforms MIIC in terms of edge precision with little to no decrease in edge recall, Figs. 3 & E.3. MIIC_search&score is also found to outperform FCI on both precision and recall on small datasets (*e.g.* $N \leqslant 10,000$ samples) of complex graphical models (*i.e.* Barley and Mildew), while reaching similar performance at larger sample sizes or for simpler graphical models (*i.e.* similar precision on Alarm and Insurance), as expected from the asymptotic consistency of FCI for very large datasets, Fig. 3. We also observed that FCI had a hard time to converge on bootstrapped datasets, explaining the lack of FCI comparison with MIIC and MIIC_search&score in Fig. E.3 for complex models at large sample sizes.

Hence, MIIC_search&score is shown to outperform all other tested causal discovery methods on complex non-Gaussian datasets including non-linear couplings between variables, while GFCI and DAG-GNN are the best performers on linear Gaussian datasets. These results demonstrate GFCI, M3HC and DAG-GNN's clear advantage in assuming intrinsically linear combinations of the variables, when analyzing multivariate Gaussian datasets. However, these results also highlight the clear limitation of this assumption, when analyzing more complex datasets including non-linear couplings between variables, which goes beyond the variable transformation of pre- or post-nonlinear causal models. By contrast, MIIC_search&score and MIIC make no particular assumption on the data distributions and achieve similar good performance across a broad range of Gaussian or non-Gaussian multimodal distributions Figs. 2 & E.2 as well as on complex real-world categorical datasets Figs. 3 & E.3.

Importantly, the benchmark PAGs used to score the causal discovery methods with increasing proportions of latent variables (Figs. 2 & 3 and Figs. E.2 & E.3) include not only bidirected edges originating from hidden common causes but also additional directed or undirected edges arising, in

particular, from indirect effects of hidden variables with observed parents. Irrespective of their orientations, all these additional edges originating from indirect effects of hidden variables generally correspond to weaker effects (*i.e.* lower mutual information of indirect effects due to the Data Processing Inequality) and are more difficult to uncover than the edges of the original DAG model without hidden variables. This explains the steady decrease in recall for complex ancestral models with higher proportions of hidden variables, while precision remains essentially unaffected. This is clearly apparent for complex random models with average degree 5 (Figs. 2 & E.2) and for complex real-world models, Insurance, Barley and Mildew (Figs. 3 & E.3).

MIIC_search&score scalability is primarily limited by the quadratic complexity of MIIC w.r.t. the number of nodes, with only a small time increase when including latent variables thanks to MIIC's greedy approach, see Fig. S5 in (Verny et al., 2017). By contrast, the two-step search-and-score scheme of MIIC_search&score is essentially linear in the numbers of nodes (step 1) and edges (step 2) for a fixed degree. In practice, Steps 1 and 2 take similar running time as for MIIC to provide a starting graph for MIIC_search&score. Hence, MIIC_search&score's scalability remains competitive against recent differential causal discovery methods using continuous optimization techniques, whose complexity is generally cubic in the number of variables, although more scalable methods have recently been developped, e.g. (Lopez et al., 2022; Montagna et al., 2023; Amin & Wilson, 2024).

All in all, these results highlight MIIC_search&score capacity to efficiently and robustly learn complex graphical models from limited available data, which is a frequent setting for many real-world applications. In addition, MIIC_search&score, which has been implemented to analyze challenging categorical or complex continuous datasets including non-linear couplings between variables, is quite unique in this regard, as all other search-and-score methods for ancestral graphs (Triantafillou & Tsamardinos, 2016; Rantanen et al., 2021; Claassen & Bucur, 2022; Andrews et al., 2022; Hu & Evans, 2024a) have only been demonstrated on continuous datasets from linear Gaussian models.

5. Limitations

The main limitation of the paper concerns the local scores used in the search-and-score algorithm, which are limited to ac-connected subsets of vertices with a maximum of two-collider paths.

While this approach could be extended to higher-order information contributions including three-or-more-collider paths, it allows for a simple two-step search-and-score scheme at the level of individual nodes (step 1) and edges (step 2), as detailed in section 3. This already shows a significant improvement in causal discovery performance (*i.e.* combining good precision and good recall on challenging benchmarks) as compared to existing state-of-the-art methods.

In addition, for ancestral graphs including only ac-connected subsets with a maximum of two-collider paths, the local consistency of the likelihood score (Eq. 12) guarantees that the two-step algorithm gives a correct estimate of the global likelihood, based on Proposition 3, just like with the likelihood decomposition of Bayesian networks (Eq. 2). Besides, for more complex ancestral graphs including longer collider paths, an oracle version of the approach could also be proposed, as the general likelihood decomposition (Eq. 12) could in principle be used in conjunction with an exhaustive search-and-score algorithm over MAGs or PAGs, which can be generated rather efficiently (Hu & Evans, 2020). Alternatively, and in practice for graphs with more than about 10-15 nodes, an MCMC algorithm based on the likelihood decomposition (Eq. 12) could be used to efficiently search for high-scoring MAGs or PAGs.

Finally, another limitation of the proposed likelihood decomposition (Eq. 12) is its restriction to MAGs without undirected edge, as shown in the proof of Theorem 1, Appendix B section (iii). However, it would be interesting to see how this result could be extended to MAGs including undirected edges. In addition, there are other existing scores under latent variables for specific parametric assumptions like discrete variables, exponential or stratified Gaussian families. In particular, discrete chain graph models (Drton, 2009), fully bidirected graph models (Drton & Richardson, 2008), and discrete nested Markov models (Richardson et al., 2023) have been shown to be curved exponential models, which can be scored consistently using BIC scores. One extension of these BIC scores has also been recently proposed (Bellot et al., 2024) to allow for the distinction between different acyclic directed mixed graphs (ADMG) from (Verma & Pearl, 1990), see Fig. 1c & d in (Bellot et al., 2024). These ADMGs imply the same set of conditional independence constraints and yet are distinguishable because they imply an equality between different functionals of the probability distribution (see Eq. 1 in (Bellot et al., 2024)). It would be interesting to further explore the relation of these other existing scores under latent variables with the present information theoretical score (Eq. 12).

Acknowledgements

We would like to thank MIIC Team members for their contributions to the MIIC projects and the reviewers for their constructive remarks and suggestions. NL acknowledges funding from the CNRS-Imperial College London joint PhD programme. HI acknowledges funding from CNRS, Institut Curie, ANR PEPR AI4scMed (ANR-22-PESN-0002) and

affiliation to Prairie Institute.

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.

References

Affeldt, S. and Isambert, H. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*, pp. 42–51, 2015.

Affeldt, S., Verny, L., and Isambert, H. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics*, 17(S2), jan 2016. doi: 10.1186/s12859-015-0856-x.

Ali, R. A. and Richardson, T. S. Markov equivalence classes for maximal ancestral graphs. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pp. 1–9, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1-55860-897-4.

Ali, R. A., Richardson, T. S., Spirtes, P., and Zhang, J. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'05, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

Ali, R. A., Richardson, T. S., and Spirtes, P. Markov equivalence for ancestral graphs. *Ann. Statist.*, 37(5B):2808–2837, oct 2009. doi: 10.1214/08-aos626.

Amin, A. and Wilson, A. Scalable and flexible causal discovery with an efficient test for adjacency. In *Proceedings of the 41 st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. doi: 10.48550/arXiv.2406.09177.

Andrews, B., Cooper, G. F., Richardson, T. S., and Spirtes, P. The m-connecting imset and factorization for ADMG models. Preprint, 2022. arxiv 2207.08963.

Bellot, A., Zhang, J., and Bareinboim, E. Scores for learning discrete causal graphs with unobserved confounders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10): 11043–11051, 2024. doi: 10.1609/aaai.v38i10.28980.

Cabeli, V., Verny, L., Sella, N., Uguzzoni, G., Verny, M., and Isambert, H. Learning clinical networks from medical records based on information estimates in mixed-type data. *PLoS Comput. Biol.*, 16(5):e1007866, 2020. doi: 10.1371/journal.pcbi.1007866.

Cabeli, V., Li, H., da Câmara Ribeiro-Dantas, M., Simon, F., and Isambert, H. Reliable causal discovery based on mutual information supremum principle for finite datasets. In *WHY21, 35rd Conference on Neural Information Processing Systems*. NeurIPS, 2021.

Chickering, D. M. A Transformational Characterization of Equivalent Bayesian Network Structures. In *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pp. 87–98. Morgan Kaufmann, 1995.

Claassen, T. and Bucur, I. G. Greedy equivalence search in the presence of latent confounders. In Cussens, J. and Zhang, K. (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 443–452. PMLR, 01–05 Aug 2022.

Drton, M. Discrete chain graph models. *Bernoulli*, 15(3):736–753, aug 2009. doi: 10.3150/08-bej172.

Drton, M. and Richardson, T. S. Graphical methods for efficient likelihood inference in gaussian covariance models. *J. Mach. Learn. Res.*, 9:893–914, 2008.

Drton, M., Eichler, M., and Richardson, T. S. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(81): 2329–2348, 2009.

Evans, R. J. and Richardson, T. S. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, UAI'10, Corvallis, OR, USA, 2010. AUAI Press.

Han, T. S. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46(1):26–45, 1980.

Hu, Z. and Evans, R. Faster algorithms for Markov equivalence. In Peters, J. and Sontag, D. (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 739–748. PMLR, 03–06 Aug 2020.

Hu, Z. and Evans, R. A fast score-based search algorithm for maximal ancestral graphs using entropy. Preprint, 2024a. arxiv 2402.04777.

Hu, Z. and Evans, R. J. Towards standard imsets for maximal ancestral graphs. *Bernoulli*, 30(3), 2024b. doi: 10.3150/23-bej1663.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. Causal inference using graphical models with the r package pcalg. *J. Stat. Softw.*, 47(11):1–26, 2012.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Kontkanen, P. *Computationally Efficient Methods for MDL-Optimal Density Estimation and Data Clustering*. PhD thesis, 2009.

Kontkanen, P. and Myllymäki, P. A linear-time algorithm for computing the multinomial stochastic complexity. *Inf. Process. Lett.*, 103(6):227–233, 2007.

Kontkanen, P., Buntine, W., Myllymäki, P., Rissanen, J., and Tirri, H. Efficient computation of stochastic complexity. *in: C. Bishop, B. Frey (Eds.) Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics*, 103:233–238, 2003.

Lagrange, N. and Isambert, H. An efficient search-and-score algorithm for ancestral graphs using multivariate information scores. Preprint, 2024. arxiv 2412.17508.

Li, H., Cabeli, V., Sella, N., and Isambert, H. Constraint-based causal structure learning with consistent separating sets. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Lopez, R., Huetter, J.-C., Pritchard, J., and Regev, A. Large-scale differentiable causal discovery of factor graphs. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 19290–19303. Curran Associates, Inc., 2022.

McGill, W. J. Multivariate information transmission. *Trans. of the IRE Professional Group on Information Theory (TIT)*, 4: 93–111, 1954.

Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. Scalable causal discovery with score matching. In van der Schaar, M., Zhang, C., and Janzing, D. (eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 752–771. PMLR, 2023.

Ogarrio, J. M., Spirtes, P., and Ramsey, J. A hybrid causal search algorithm for latent variable models. *JMLR Workshop Conf. Proc.*, 52:368–379, 2016.

Pearl, J. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo, CA, 1988.

Pearl, J. *Causality: models, reasoning and inference*. Cambridge University Press, 2nd edition, 2009.

Pearl, J. and Paz, A. Graphoids: A graph-based logic for reasoning about relevance relations, or when would x tell you more about y if you already know z. Technical report, UCLA Computer Science Department, December 1985.

Rantanen, K., Hyttinen, A., and Järvisalo, M. Maximal ancestral graph structure learning via exact search. In de Campos, C. and Maathuis, M. H. (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1237–1247. PMLR, 27–30 Jul 2021.

Ribeiro-Dantas, M. d. C., Li, H., Cabeli, V., Dupuis, L., Simon, F., Hettal, L., Hamy, A.-S., and Isambert, H. Learning interpretable causal networks from very large datasets, application to 400, 000 medical records of breast cancer patients. *iScience*, 27(5): 109736, 2024. doi: 10.1016/j.isci.2024.109736.

Richardson, T. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

Richardson, T. and Spirtes, P. Scoring ancestral graph models. Technical report, 1999. Available as Technical Report CMU-PHIL 98.

Richardson, T. and Spirtes, P. Ancestral graph Markov models. *Ann. Statist.*, 30(4):962–1030, aug 2002. doi: 10.1214/aos/1031689015.

Richardson, T. S. A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pp. 462–470, Arlington, VA, USA, 2009. AUAI Press.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1), 2023. doi: 10.1214/22-aos2253.

Rissanen, J. and Tabus, I. Kolmogorov's structure function in mdl theory and lossy data compression. In *Adv. Min. Descrip. Length Theory Appl.*, pp. 245–264. MIT Press, 2005.

Roos, T., Silander, T., Kontkanen, P., and Myllymäki, P. Bayesian network structure learning using factorized nml universal models. In *Proc. 2008 Information Theory and Applications Workshop (ITA-2008)*. IEEE Press, 2008.

Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Softw.*, 35(3):1–22, 2010.

Shtarkov, Y. M. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, 1987.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT press, , 2nd edition, 2000.

Szpankowski, W. *Average case analysis of algorithms on sequences*. John Wiley & Sons, , 2001.

Tian, J. Generating Markov equivalent maximal ancestral graphs by single edge replacement. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'05, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 567–573. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002.

Ting, H. K. On the amount of information. *Theory Probab. Appl.*, 7(4):439–447, 1962. doi: 10.1137/1107041.

Triantafillou, S. and Tsamardinos, I. Score-based vs constraint-based causal learning in the presence of confounders. In *CFA@UAI*, 2016.

Verma, T. and Pearl, J. Causal networks: Semantics and expressiveness. In Shachter, R. D., Levitt, T. S., Kanal, L. N., and Lemmer, J. F. (eds.), *Uncertainty in Artificial Intelligence*, volume 9 of *Machine Intelligence and Pattern Recognition*, pp. 69–76. North-Holland, 1990. doi: 10.1016/B978-0-444-88650-7.50011-1.

Verny, L., Sella, N., Affeldt, S., Singh, P. P., and Isambert, H. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.*, 13(10): e1005662, 2017.

Yeung, R. W. A new outlook on shannon's information measures. *IEEE transactions on information theory*, 37(3):466–474, 1991. doi: 10.1109/18.79902.

Yu, Y., Chen, J., Gao, T., and Yu, M. DAG-GNN: DAG structure learning with graph neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7154–7163. PMLR, 2019.

Zhang, J. A characterization of Markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'07, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172(16-17):1873–1896, 2008.

Zheng, Y., Huang, B., Chen, W., Ramsey, J., Gong, M., Cai, R., Shimizu, S., Spirtes, P., and Zhang, K. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

Appendices

A. Preliminaries: connection and separation criteria

A.1. *m*-connection *vs* *m'*-connection criteria

An ancestral graph can be interpreted as encoding a set of conditional indepencence relations by a graphical criterion, called m-separation, based on the concept of m-connecting paths, which generalizes the separation criteria of Markov and Bayesian networks to ancestral graphs.

Definition 4. [m-connecting path] A path $\boldsymbol{\pi}$ between X and Y is m-connecting given a (possibly empty) subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ (with $X, Y \notin \boldsymbol{C}$) if:
i) its non-collider(s) are not in $\boldsymbol{C}$, and
ii) its collider(s) are in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{C})$.

Definition 5. [m-separation criterion] The subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be m-separated by $\boldsymbol{C}$, noted $\boldsymbol{A} \perp_m \boldsymbol{B} | \boldsymbol{C}$, if there is no m-connecting path between any vertex in $\boldsymbol{A}$ and any vertex in $\boldsymbol{B}$ given $\boldsymbol{C}$.

The probabilistic interpretation of ancestral graph is given by its global and pairwise Markov properties (which are equivalent (Richardson & Spirtes, 2002)): if $\boldsymbol{A}$ and $\boldsymbol{B}$ are m-separated by $\boldsymbol{C}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are conditionally independent given $\boldsymbol{C}$ and $\forall X \in \boldsymbol{A}$ and $\forall Y \in \boldsymbol{B}$, there is a probability distribution P faithful to $\mathcal{G}$ such that their conditional mutual information vanishes, *i.e.* $I_P(X; Y | \boldsymbol{C}) = 0$, also noted $X \perp\!\!\!\perp_P Y | \boldsymbol{C}$.

However, as discussed above, the proof of Theorem 1 will require to introduce a weaker m'-connection criterion defined below.

Definition 6. [m'-connecting path] A path $\boldsymbol{\pi}$ between X and Y is m'-connecting given a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ (with X, Y possibly in $\boldsymbol{C}$) if:
i) its non-collider(s) are not in $\boldsymbol{C}$, and
ii) its collider(s) are in $\mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{C})$.

Note, in particular, that an m-connecting path is necessary an m'-connecting path but that the converse is not always true. For example, the path $X \rightarrow Z \longleftrightarrow T \longleftrightarrow Y$ in Fig. 1G (with $Z \rightarrow Y$) is an m'-connecting path given T (as $Z \in \mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup T)$) but not an m-connecting path given T (as $Z \notin \mathbf{An}_{\mathcal{G}}(T)$).

However, Richardson and Spirtes 2002 have shown the following lemma,

Lemma 4. [Corollary 3.15 in (Richardson & Spirtes, 2002)] *In an ancestral graph $\mathcal{G}$, there is a m'-connecting path $\boldsymbol{\mu}$ between X and Y given $\boldsymbol{C}$ if and only if there is a (possibly different) m-connecting path $\boldsymbol{\pi}$ between X and Y given $\boldsymbol{C}$.*

Hence, Lemma 4 implies that m'-separation and m-separation criteria are in fact equivalent, as an absence of m'-connecting paths implies an absence of m-connecting paths and vice versa. This enables to reformulate the m-separation criterion above as,

Definition 7. [m'-separation (and m-separation) criteria] The subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be m'-separated (or m-separated) by $\boldsymbol{C}$, if all paths from any $X \in \boldsymbol{A}$ to any $Y \in \boldsymbol{B}$ have either
i) a non-collider in $\boldsymbol{C}$, or
ii) a collider *not* in $\mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{C})$.

The probabilistic interpretation of an ancestral graph is given by its (global) Markov property: if $\boldsymbol{A}$ and $\boldsymbol{B}$ are m-separated (or m'-separated) by $\boldsymbol{C}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are conditionally independent given $\boldsymbol{C}$, noted as, $\boldsymbol{A} \perp_m \boldsymbol{B} | \boldsymbol{C}$.

A.2. *ac*-connecting paths and *ac*-connected subsets

Let us now recall the definition of **ancestor collider connecting paths** or ***ac*-connecting paths**, which is directly relevant to characterize the likelihood decomposition and Markov equivalent classes of ancestral graphs (Theorem 1). We give here a different yet equivalent definition of *ac*-connecting paths as defined in the main text (Definition 2) in order to underline the similarities and differencies with the notion of m'-connecting path (Definition 6).

Definition 8. [ac-connecting path] A path $\boldsymbol{\pi}$ between X and Y is an ac-connecting path given a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ (with X and Y possibly in $\boldsymbol{C}$) if:

i) $\boldsymbol{\pi}$ does not have any noncollider, and
ii) its collider(s) are in $\mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{C})$.

Hence, more simply (following Definition 2 in the main text), an ac-connecting path given $\boldsymbol{C}$ is a collider path, $X *\!\!\rightarrow Z_1 \leftrightarrow \cdots \leftrightarrow Z_K \leftarrow\!\!* Y$, with all $Z_i \in \mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{C})$, *i.e.* with Z_i in $\boldsymbol{C}$ or connected to $\{X, Y\} \cup \boldsymbol{C}$ by an ancestor path, $Z_i \rightarrow \cdots \rightarrow T$ with $T \in \{X, Y\} \cup \boldsymbol{C}$.

Definition 9. [ac-separation criterion] The subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be ac-separated by $\boldsymbol{C}$ if there is no ac-connecting path between any vertex in $\boldsymbol{A}$ and any vertex in $\boldsymbol{B}$ given $\boldsymbol{C}$.

Previous definitions and Lemma 4 readily lead to the following corollary between the different connection and separation criteria:

Corollary 5.
i) m*-connecting path* $\pi \implies m'$*-connecting path* π
ii) ac*-connecting path* $\pi \implies m'$*-connecting path* π
iii) m*-separation* $\iff m'$*-separation*
iv) m/m'*-separation* $\implies ac$*-separation*

Finally, we recall the notion of **ac-connected subset** (Definition 3 in the main text), which is central for the decomposition of the likelihood of ancestral graphs (Theorem 1): A subset $\boldsymbol{C}$ is said to be ac-connected if $\forall X, Y \in \boldsymbol{C}$, there is an ac-connecting path between X and Y w.r.t. $\boldsymbol{C}$.

B. Proof of Theorem 1.

In order to prove that the likelihood function of an ancestral graph, Eq. 12, contains all and only the ac-connected subsets of vertices in $\mathcal{G}$ (Definition 3), we will first show (***i***) that all non-ac-connected subsets $\boldsymbol{S'}$ are included in a cancelling combination of multivariate cross-information terms, $I(X; Y|\boldsymbol{A}) = 0$, with $X, Y \in \boldsymbol{S'}$ and $\boldsymbol{S'} \subseteq \boldsymbol{S} = \{X, Y\} \cup \boldsymbol{A}$, Eq. 11. Conversely, we will then show (***ii***) that cancelling combinations of multivariate cross-information terms associated to pairwise conditional independence, $I(X; Y|\boldsymbol{A}) = \sum_{\boldsymbol{S'} \subseteq \boldsymbol{S}}^{X,Y \in \boldsymbol{S'}} (-1)^{|\boldsymbol{S'}|} I(\boldsymbol{S'}) = 0$, do not contain any ac-connected subset $\boldsymbol{S'}$. Finally, we will prove (***iii***) that the information terms which appear in multiple cancelling combinations from different pairwise independence constraints do not modify the multivariate information decomposition of the likelihood function of ancestral graphs, Eq. 12, as these shared/overlapping terms in fact all cancel through more global Markov independence relationships involving higher order (three or more points) vanishing multivariate information terms, such as $I(X; Y; Z|\boldsymbol{A}) = 0$.

i) Let's first prove that all non-ac-connected subsets $\boldsymbol{S'}$ are included in at least one cancelling combination of multivariate cross-information, $I(X; Y|\boldsymbol{A}) = 0$, with $X, Y \in \boldsymbol{S'}$ and $\boldsymbol{S'} \subseteq \{X, Y\} \cup \boldsymbol{A}$.

If $\boldsymbol{S'}$ is a non-ac-connected subset, there is at least one disconnected pair X and Y for which each path π_j between X and Y contains either some collider(s) not in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ or, if all colliders along π_j are in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$, there must be some non-collider(s) at node(s) $\boldsymbol{Z}_j$ but not necessarily in $\boldsymbol{S'}$. Let's define $\boldsymbol{S} = \boldsymbol{S'} \cup_j \boldsymbol{Z}_j$. X and Y can be shown to be m-separated given $\boldsymbol{S} \setminus \{X, Y\}$, as for each path π_j between X and Y, its non-collider(s) are in $\boldsymbol{S}$ at node(s) $\boldsymbol{Z}_j$ (when all collider(s) along π_j are in $\boldsymbol{S'}$) or there is some collider(s) not in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$, which are not in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S})$ either. The latter statement is proven by contradiction assuming that there is a collider at $Z \notin \mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ such that $Z \in \mathbf{An}_{\mathcal{G}}(\boldsymbol{S})$. There is therefore a directed path $Z \rightarrow \cdots \rightarrow W$ with $W \in \boldsymbol{S}$. Hence, $W \in \boldsymbol{S'}$ or there is a noncollider at $W \in \boldsymbol{Z}_j$ which is on a path π_j between X and Y along which all colliders are in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ by construction of $\boldsymbol{S}$. This leads by induction to $Z \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow T$ where $T \in \boldsymbol{S'}$ and thus $Z \in \mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$, which is a contradiction. Hence, all non-ac-connected subsets $\boldsymbol{S'}$ are included in a cancelling combination of multivariate cross-information terms, $I(X; Y|\boldsymbol{A}) = 0$, with $X, Y \in \boldsymbol{S'}$ and $\boldsymbol{S'} \subseteq \boldsymbol{S} = \{X, Y\} \cup \boldsymbol{A}$.

ii) Conversely, we will now show that cancelling combinations of multivariate cross-information terms associated to pairwise conditional independence, $I(X; Y|\boldsymbol{A}) = \sum_{\boldsymbol{S'} \subseteq \boldsymbol{S}}^{X,Y \in \boldsymbol{S'}} (-1)^{|\boldsymbol{S'}|} I(\boldsymbol{S'}) = 0$, do not contain any ac-connected subset $\boldsymbol{S'}$, where $\boldsymbol{S} = \{X, Y\} \cup \boldsymbol{A}$.

We will prove it by contradiction assuming that there exists a subset $\boldsymbol{W} \subseteq \boldsymbol{A}$, such that $\boldsymbol{S'} = \{X, Y\} \cup \boldsymbol{W}$ is ac-connected. In particular, there should be an ac-connecting path between X and Y confined to $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ and thus to $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S}) \supseteq \mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$, which is an m'-connecting path between X and Y given $\boldsymbol{A}$, contradicting the above hypothesis of m'-separation given $\boldsymbol{A}$, *i.e.* $I(X; Y|\boldsymbol{A}) = 0$. The use of m'-separation, *i.e.* the absence of m'-connecting paths with

colliders in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S})$ rather than m-connecting paths with colliders in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{A})$, is necessary here, see Definitions 4 and 6. Hence, no ac-connected subset $\boldsymbol{S'}$ is included in cancelling combinations of multivariate cross-information terms associated to pairwise conditional independence, $I(X;Y|\boldsymbol{A}) = \sum_{\boldsymbol{S'}\subseteq\boldsymbol{S}}^{X,Y\in\boldsymbol{S'}} (-1)^{|\boldsymbol{S'}|} I(\boldsymbol{S'}) = 0$.

iii) Finally, we will show that the information terms which appear in multiple cancelling combinations from different pairwise independence constraints do not modify the multivariate cross-information decomposition of the likelihood function of ancestral graphs, Eq. 12, as these shared/overlapping terms in fact all cancel through more global Markov independence relationships involving higher order (three or more points) vanishing multivariate cross-information terms, such as $I(X;Y;Z|\boldsymbol{A}) = 0$.

This result requires to use an ordering of the nodes, $X_k \succ X_j \succ X_i$, that is compatible with the directed edges of the ancestral graph assumed to have no undirected edges, *i.e.* $X_j \notin \mathbf{An}(X_i)$ if $X_j \succ X_i$. Under this ordering, higher order nodes $X_k \succ X_j \succ X_i$ can be a priori excluded from all separating sets $\boldsymbol{A}_{ij}$ of pairs of lower order nodes, *i.e.* if $I(X_i;X_j|\boldsymbol{A}_{ij}) = 0$ then $X_k \notin \boldsymbol{A}_{ij}$.

In particular, the two pairwise conditional independence relations $I(X_k;X_\ell|\boldsymbol{A}_{k\ell}) = 0$, with $X_\ell \succ X_k$, and $I(X_i;X_j|\boldsymbol{A}_{ij}) = 0$, with $X_j \succ X_i$, do not share any multivariate cross-information terms, if $X_\ell \neq X_j$. Indeed, as $I(X_i;X_j|\boldsymbol{A}_{ij})$ contains all cross-information terms including both X_i and X_j as well as every subset (possibly empty) of $\boldsymbol{A}_{ij}$, none of them includes X_ℓ if $X_\ell \succ X_j$. Therefore $I(X_i;X_j|\boldsymbol{A}_{ij})$ does not contain any cross-information term of $I(X_k;X_\ell|\boldsymbol{A}_{k\ell})$ which contains both X_k and X_ℓ as well as every subset (possibly empty) of $\boldsymbol{A}_{k\ell}$. This property eliminates all multiple countings of multivariate cross-information terms if $X_\ell \neq X_j$. Note that this result does not hold in general for ancestral graphs including undirected edges.

Hence, the issue of redundant multivariate cross-information terms in the likelihood decomposition, Eq. 12, is related to the conditional independences of two or more pairs, $\{X_i,X_r\}$, $\{X_j,X_r\}$, ..., $\{X_\ell,X_r\}$, sharing the same higher order node, X_r, *i.e.*, $I(X_k;X_r|\boldsymbol{A}_{kr}) = 0$ for $k = i,j,\cdots,\ell$. However, this situation also entails a more global Markov independence constraint between X_r and $\{X_i,X_j,\cdots,X_\ell\}$, given a separating set $\boldsymbol{A}$, with $\boldsymbol{A}_{kr} \subseteq \boldsymbol{A} \cup \{X_i,\cdots,X_\ell\}$ for $k = i,j,\cdots,\ell$. Such a global Markov independence constraint can be decomposed into more local independence constraints using the chain rule (in *any* order of the variables $X_i,X_j,\cdots,X_\ell$) and the decomposition rules of multivariate (cross) information (Eq. 9),

$$
\begin{aligned}
0 &= I(\{X_i,X_j,\cdots,X_\ell\};X_r|\boldsymbol{A}) \\
&= \big(I(X_i;X_r|\boldsymbol{A}) + I(X_j;X_r|\boldsymbol{A},X_i)\big) + \big[I(X_k;X_r|\boldsymbol{A},X_i,X_j)\big] + \cdots + I(X_\ell;X_r|\boldsymbol{A},\cdots) \\
&= \big(I(X_i;X_r|\boldsymbol{A}) + I(X_j;X_r|\boldsymbol{A}) - I(X_i;X_j;X_r|\boldsymbol{A})\big) \\
&\quad + \big[I(X_k;X_r|\boldsymbol{A},X_i) - I(X_j;X_k;X_r|\boldsymbol{A},X_i)\big] + \cdots + I(X_\ell;X_r|\boldsymbol{A},\cdots) \\
&= \big(I(X_i;X_r|\boldsymbol{A}) + I(X_j;X_r|\boldsymbol{A}) - I(X_i;X_j;X_r|\boldsymbol{A})\big) \\
&\quad + \big[I(X_k;X_r|\boldsymbol{A}) - I(X_j;X_k;X_r|\boldsymbol{A}) - I(X_i;X_k;X_r|\boldsymbol{A}) + I(X_i;X_j;X_k;X_r|\boldsymbol{A})\big] + \cdots
\end{aligned}
$$

where all the conditional multivariate cross-information terms vanish by induction due to the non-negativity of (conditional) mutual (cross) information. In particular, the conditional multivariate cross-information terms in the last expression, *i.e.* between X_r and each subset of $\{X_i,X_j,\cdots,X_\ell\}$ given the separating set $\boldsymbol{A}$, all vanish. This result can be readily extended to any subsets $\{X_r,X_s,\cdots,X_z\}$ (conditionally) independent of $\{X_i,X_j,\cdots,X_\ell\}$ given a separating set $\boldsymbol{A}$, *i.e.* $I(\{X_i,X_j,\cdots,X_\ell\};\{X_r,X_s,\cdots,X_z\}|\boldsymbol{A}) = 0$. Hence, as the final conditional multivariate cross-information terms of the decomposition all vanish while not sharing any subsets of variables, it proves the absence of redundancy and a global cancellation of non-ac-connected subsets (from pairwise and higher order conditional independence relations) in the likelihood function of ancestral graphs without undirected edges, Eq. 12.

Hence, only ac-connected subsets effectively contribute to the cross-entropy of an ancestral graph with only directed and bidirected edges, Eq. 12. □

C. Factorization of the probability distribution of ancestral graphs

C.1. Factorization resulting from Theorem 1 and Proposition 3

Before presenting the factorization of the model distribution of ancestral graphs resulting from Theorem 1 and Proposition 3, it is instructive to obtain an equivalent factorization for Bayesian graphs, assuming a positive empirical distributions, $p(x_1, \cdots, x_m) = \prod_{i=1}^{m} p(x_i|x_{i-1}, \cdots, x_1) > 0$,

$$\begin{aligned} q(x_1, \cdots, x_m) &= \prod_{i=1}^{m} q(x_i|\mathbf{pa}_{x_i}) = \prod_{i=1}^{m} p(x_i|\mathbf{pa}_{x_i}) \\ &= p(x_1, \cdots, x_m) \prod_{i=1}^{m} \frac{p(x_i|\mathbf{pa}_{x_i})}{p(x_i|x_{i-1}, \cdots, x_1)} \\ &= p(x_1, \cdots, x_m) \prod_{i=1}^{m} \frac{p(x_i|\mathbf{pa}_{x_i}) p(\boldsymbol{x}_{i-1} \backslash \mathbf{pa}_{x_i}|\mathbf{pa}_{x_i})}{p(x_i, \boldsymbol{x}_{i-1} \backslash \mathbf{pa}_{x_i}|\mathbf{pa}_{x_i})} \end{aligned} \tag{15}$$

This leads to the following alternative expressions for the cross-entropy $H(p, q) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x})$ in terms of multivariate entropy and information, which only depend on the empirical joint distribution $p(\boldsymbol{x})$,

$$\begin{aligned} H(p, q) &= \sum_{i=1}^{m} H(x_i|\mathbf{Pa}_{X_i}) \\ &= H(X_1, \cdots, X_m) + \sum_{i=1}^{m} I(X_i; \boldsymbol{X}_{i-1} \backslash \mathbf{Pa}_{X_i}|\mathbf{Pa}_{X_i}) \end{aligned} \tag{16}$$

where $\sum_{i=1}^{m} I(X_i; \boldsymbol{X}_{i-1} \backslash \mathbf{Pa}_{X_i}|\mathbf{Pa}_{X_i})$ can be decomposed, using the chain rule and Eq. 11, into unconditional multivariate information terms, which exactly cancel all the multivariate information of the non-ac-connected subsets of variables in the multivariate entropy decomposition, Eq. 6.

Note, however, that this result obtained for Bayesian networks requires an explicit factorization of the global model distribution, $q(\boldsymbol{x})$, in terms of the empirical distribution, $p(\boldsymbol{x})$, which is not known and presumably does not exist, in general, for ancestral graphs.

Alternatively, assuming that the empirical and model distributions are positive ($\forall \boldsymbol{x}, p(\boldsymbol{x}) > 0$, $q(\boldsymbol{x}) > 0$), it is always possible to factorize them into factors associated to each (cross) information term in the (cross) entropy decomposition, Eq. 6, as,

$$q(\boldsymbol{x}) = \prod_{i=1}^{m} q(x_i) \times \prod_{i<j}^{m} \frac{q(x_i, x_j)}{q(x_i) q(x_j)} \times \prod_{i<j<k}^{m} \frac{q(x_i, x_j, x_k) q(x_i) q(x_j) q(x_k)}{q(x_i, x_j) q(x_i, x_k) q(x_j, x_k)} \times \cdots \tag{17}$$

where all the marginal distributions over a subset of variables, *e.g.* $q(x_i, x_j, x_k) = \sum_{\ell \neq i,j,k} q(\boldsymbol{x})$ or $p(x_i, x_j, x_k) = \sum_{\ell \neq i,j,k} p(\boldsymbol{x})$, cancel two-by-two by construction.

This can be illustrated on a simple example of a two-collider path including one bidirected edge, $X \to Z \longleftrightarrow Y \leftarrow W$ (Fig. 1D), valid for $q(.)$ and $p(.)$ alike,

$$\begin{aligned} q(x, z, y, w) &= q(x)\, q(z)\, q(y)\, q(w) \\ &\quad \times \frac{q(x, z)}{q(x)\, q(z)} \frac{q(z, y)}{q(z)\, q(y)} \frac{q(y, w)}{q(y)\, q(w)} \frac{q(x, y)}{q(x)\, q(y)} \frac{q(x, w)}{q(x)\, q(w)} \frac{q(z, w)}{q(z)\, q(w)} \\ &\quad \times \frac{q(x)\, q(z)\, q(y)\, q(x, z, y)}{q(x, z)\, q(x, y)\, q(z, y)} \frac{q(z)\, q(y)\, q(w)\, q(z, y, w)}{q(z, y)\, q(z, w)\, q(y, w)} \\ &\quad \times \frac{q(x)\, q(z)\, q(w)\, q(x, z, w)}{q(x, z)\, q(x, w)\, q(z, w)} \frac{q(x)\, q(y)\, q(w)\, q(x, y, w)}{q(x, y)\, q(x, w)\, q(y, w)} \\ &\quad \times \frac{q(x, z)\, q(z, y)\, q(y, w)\, q(x, y)\, q(x, w)\, q(z, w)\, q(x, z, y, w)}{q(x, z, y)\, q(x, z, w)\, q(x, y, w)\, q(z, y, w)\, q(x)\, q(y)\, q(z)\, q(w)} \end{aligned} \tag{18}$$

where all individual distribution marginals on subsets of variables, *e.g.* $q(x)$, $q(x,z)$, $q(x,z,y)$ (or $p(x)$, $p(x,z)$, $p(x,z,y)$), cancel two-by-two by construction, except $q(x,z,y,w)$ (or $p(x,z,y,w)$).

In addition and *only for the model distribution* $q(.)$, all ratios in gray in Eq. 18 also cancel due to Markov independence relations across non-ac-connected subsets (see proof of Theorem 1). This leaves a truncated factorization retaining all and only the ac-connected subsets of variables in the graph, which we propose to estimate on empirical data by substituting the remaining $q(.)$ terms by their empirical counterparts $p(.)$, see Proposition 3.

This leads to the following global factorization for $q(.)$ in terms of $p(.)$,

$$\begin{aligned} q(x,z,y,w) \;\equiv\; & p(x)\,p(z)\,p(y)\,p(w)\,\frac{p(x,z)}{p(x)\,p(z)}\,\frac{p(z,y)}{p(z)\,p(y)}\,\frac{p(y,w)}{p(y)\,p(w)} \\ & \times \frac{p(x)\,p(z)\,p(y)\,p(x,z,y)}{p(x,z)\,p(x,y)\,p(z,y)}\,\frac{p(z)\,p(y)\,p(w)\,p(z,y,w)}{p(z,y)\,p(z,w)\,p(y,w)} \\ & \times \frac{p(x,z)\,p(z,y)\,p(y,w)\,p(x,y)\,p(x,w)\,p(z,w)\,p(x,z,y,w)}{p(x,z,y)\,p(x,z,w)\,p(x,y,w)\,p(z,y,w)\,p(x)\,p(y)\,p(z)\,p(w)} \\ =\; & p(x,z,y,w)\,\frac{p(x)\,p(y)}{p(x,y)}\,\frac{p(x)\,p(w)}{p(x,w)}\,\frac{p(z)\,p(w)}{p(z,w)} \\ & \times \frac{p(x,z)\,p(x,w)\,p(z,w)}{p(x)\,p(z)\,p(w)\,p(x,z,w)}\,\frac{p(x,y)\,p(x,w)\,p(y,w)}{p(x)\,p(y)\,p(w)\,p(x,y,w)} \end{aligned} \tag{19}$$

where the terms in gray have been passed to the lhs of Eq. 18 applied to $p(.)$. This ultimately leads to the analog of the Bayesian Network factorization in Eq. 15 but for the two-collider path, $X \to Z \longleftrightarrow Y \leftarrow W$ (Fig. 1D),

$$q(x,z,y,w) \;\equiv\; p(x,z,y,w)\,\frac{p(x)\,p(w)}{p(x,w)}\,\frac{p(z|x)\,p(w|x)}{p(z,w|x)}\,\frac{p(x|w)\,p(y|w)}{p(x,y|w)} \tag{20}$$

where the last three factors "correct" the expression of $p(x,z,y,w)$ for the three (conditional) independences entailed by the underlying graph, that is, $X \perp W$, $Z \perp W|X$, and $X \perp Y|W$.

C.2. Relation to the head-and-tail factorizations

The head-and-tail factorizations of the model distribution of an acyclic directed mixed graph, introduced by Richardson 2009, do not correspond to a single factorized equation (as with Bayesian graphs, Eq. 15) but to multiple factorized equations, which enable the parametrization of the joint probability distribution with independent parameters for ancestrally closed subsets of vertices.

For instance, the head-and-tail factorizations of the simple two-collider path including one bidirected edge, $X \to Z \longleftrightarrow Y \leftarrow W$, introduced above, Fig. 1D, correspond to the following equations (Richardson, 2009),

$$\begin{aligned} q(x,w) &= q(x)\,q(w) \\ q(x,z) &= q(z|x)\;q(x) \\ q(y,w) &= q(y|w)\;q(w) \\ q(x,z,w) &= q(z|x)\;q(x)\,q(w) \\ q(x,y,w) &= q(y|w)\;q(w)\,q(x) \\ q(x,z,y,w) &= q(z,y|x,w)\;q(x)\,q(w) \end{aligned} \tag{21}$$

Importantly, these head-and-tail factorizations imply additional relations such as $q(y|w) = q(y|x,w)$ (*i.e.* $X \perp Y|W$) obtained by comparing the last two relations in Eqs. 21 after marginalizing $q(x,z,y,w)$ over z. However, such implicit conditional independence relations are *not verified by the empirical distribution* $p(.)$ *in general* and prevent the estimation of the head-and-tail factorizations by substituting the rhs $q(.)$ terms in Eqs. 21 with their empirical counterparts $p(.)$, as in the case of Bayesian networks, Eq. 15.

Indeed, while the head-and-tail factorization relations, Eqs. 21, obey the local and global Markov independence relations entailed by the graphical model, Fig. 1D, leading to the cancellation of all factors associated to non-ac-connected subsets in gray in Eq. 18, the remaining head-and-tail factors cannot be readily estimated with the empirical distribution $p(.)$.

In particular, the cross-entropy of the two-collider path of interest, Fig. 1D, obtained with the head-and-tail factorizations corresponds to[2] $H(p,q) = -\sum p(x,z,y,w) \log q(z,y|x,w)\, q(x)\, q(w)$. Then, estimating the $q(.)$ terms with their $p(.)$ counterparts leads to the cross-entropy of a Bayesian graph, Fig. 1E, with a different Markov equivalent class than the ancestral graph of interest, Fig. 1D. A similar discrepancy is obtained with a c-component factorization which leads to the cross-entropy of the Bayesian graph of Fig. 1E without edge $X \to Y$, corresponding to a different Markov equivalence class than the previous two graphs, Figs. 1D & E.

These examples illustrate the difficulty to exploit the c-component or head-and-tail factorizations to estimate the likelihood of ancestral graphs including bidirected edge(s).

D. Node and edge scores based on Normalized Maximum Likelihood criteria

MIIC_search&score's two-step implementation is illustrated with a running example on Fig. D.1. It relies on Node scores (Step 1), for edge orientation priming and edge removal, and on Edge orientation scores (Step 2). Both scores are based on Normalized Maximum Likelihood criteria, derived in this Appendix.

(i) Output from MIIC (ii) **Step 1**: Edge orientation priming and Edge removal (iii) **Step 2**: Edge orientation

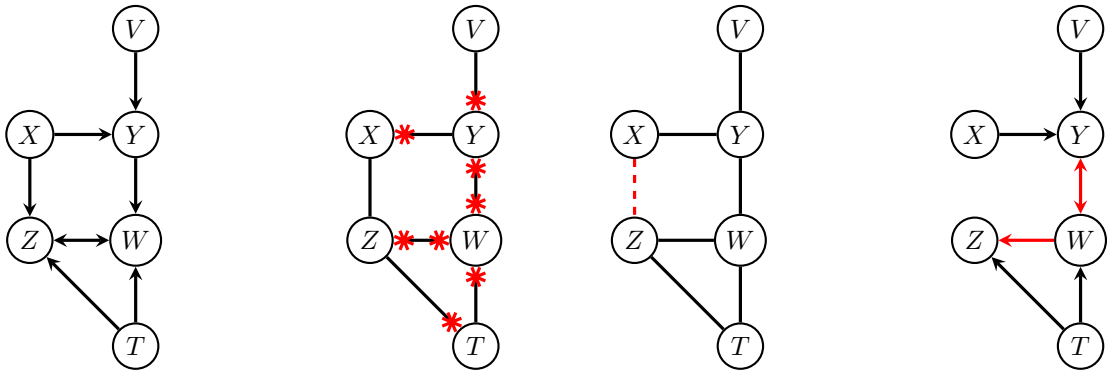

Figure D.1. **A simple running example of MIIC_search&score's two-step implementation.** (i) The graph obtained by MIIC is used as starting point for MIIC_search&score two-step algorithm. (ii) Step 1 (Node scores): Edge orientation primings (*) are obtained from the minimization of the node scores Eqs. 14 & 34. Edges without priming at either extremity are removed at the end of Step 1 (e.g. X−−Z). (iii) Step 2 (Edge orientation scores) : Edge orientations for directed or bidirected edges are obtained by optimizing their local contributions (Table 1) to the global likelihood score (Eq. 12) restricted to ac-connected subsets containing up to two-collider paths. This amounts to minimizing each edge orientation score w.r.t. its nodes' parents and spouses (Table 1). The orientations of the edges highlighted in red have been modified with respect to their orientations in the MIIC network (i) used as starting point for MIIC_search&score.

Search-and-score methods based on likelihood estimates need to properly account for finite sample size, as cross-entropy minimization leads to ever more complex models, resulting in model overfitting for finite datasets. BIC regularization is valid in the asymptotic limit of very large datasets and leads to the following finite size corrections of the cross-information terms in the likelihood decomposition Eq. 12,

$$I(\boldsymbol{C}) \to I'(\boldsymbol{C}) = I(\boldsymbol{C}) - \frac{1}{2} \prod_{k=1}^{|\boldsymbol{C}|} (1 - r_k) \frac{\log N}{N} \tag{22}$$

for categorical datasets, where r_k is the number of categories or levels of the kth variable of $\boldsymbol{C}$.

However, BIC regularization tends to overestimate finite size corrections, leading to lower recall, in general. In order to better take into account finite sample size, we used instead the (universal) Normalized Maximum Likelihood (NML) criterion (Shtarkov, 1987; Rissanen & Tabus, 2005; Kontkanen & Myllymäki, 2007; Roos et al., 2008), which amounts to normalizing the likelihood function over all possible datasets with the same number N of samples.

[2] Indeed, all terms in Eq. 18 actually cancel two-by-two by construction, *whatever their factorized expression*, except for the remaining joint-distribution over all variables, $q(x,z,y,w) = q(z,y|x,w)\, q(x)\, q(w)$.

Moreover, as search-and-score structure learning methods need to rely on heuristic rather than exhaustive search, we have implemented a computationally efficient search-and-score method based on the likelihood decomposition of ancestral graphs (Eq. 12) limited to the close surrounding vertices of each node and edge. These node and edge scores, detailed below, extend MIIC's unshielded triple scores to higher-order local information scores including ac-connected subsets of vertices with a maximum of two-collider paths.

Node score. We first used the factorized Normalized Maximum Likelihood (fNML) complexity (Kontkanen & Myllymäki, 2007; Roos et al., 2008) to define a local score for each node X_i, which extends the decomposable likelihood of Bayesian graphs given each node's parents, Eq. 2, to all non-descendant neighbors, $\mathbf{Pa}'_{X_i}$,

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\mathrm{X_i}}} = e^{-N.\,\mathrm{Score_n}(X_i)} = \frac{e^{-NH(X_i|\mathbf{Pa}'_{X_i})}}{\sum_{|\mathcal{D}'|=N} e^{-NH(X_i|\mathbf{Pa}'_{X_i})}} \tag{23}$$

$$= e^{-NH(X_i|\mathbf{Pa}'_{X_i}) - \sum_j^{q_i} \log \mathcal{C}_{n_j}^{r_i}} \tag{24}$$

$$= e^{N \sum_j^{q_i} \sum_k^{r_i} \frac{n_{jk}}{N} \log\left(\frac{n_{jk}}{n_j}\right) - \sum_j^{q_i} \log \mathcal{C}_{n_j}^{r_i}} \tag{25}$$

$$= \prod_j^{q_i} \frac{\prod_k^{r_i} \left(\frac{n_{jk}}{n_j}\right)^{n_{jk}}}{\mathcal{C}_{n_j}^{r_i}} \tag{26}$$

where n_{jk} corresponds to the number of data points for which X_i is in its kth state and its non-descendant neighbors in their jth state, with $n_j = \sum_k^{r_i} n_{jk}$. The universal normalization constant $\mathcal{C}_n^r$ is then computed by summing the numerator over all possible partitions of the n data points into a maximum of r subsets, $\ell_1 + \ell_2 + \cdots + \ell_r = n$ with $\ell_k \geqslant 0$,

$$\mathcal{C}_n^r = \sum_{\ell_1+\ell_2+\cdots+\ell_r=n} \frac{n!}{\ell_1!\ell_2!\cdots\ell_r!} \prod_{k=1}^{r} \left(\frac{\ell_k}{n}\right)^{\ell_k} \tag{27}$$

which can in fact be computed in linear-time using the following recursion (Kontkanen & Myllymäki, 2007),

$$\mathcal{C}_n^r = \mathcal{C}_n^{r-1} + \frac{n}{r-2}\mathcal{C}_n^{r-2} \tag{28}$$

with $\mathcal{C}_n^1 = 1$ for all n and applying Eq. 31 below for $r = 2$. However, for large n and r, $\mathcal{C}_n^r$ computation tends to be numerically unstable, which can be circumvented by implementing the recursion on parametric complexity ratios $\mathcal{D}_n^r = \mathcal{C}_n^r / \mathcal{C}_n^{r-1}$ rather than parametric complexities themselves (Cabeli et al., 2020) as,

$$\mathcal{D}_n^r = 1 + \frac{n}{(r-2)\mathcal{D}_n^{r-1}} \tag{29}$$

$$\log \mathcal{C}_n^r = \sum_{k=2}^{r} \log \mathcal{D}_n^k \tag{30}$$

for $r \geqslant 3$, with $\mathcal{C}_n^1 = 1$ and $\mathcal{C}_n^2 = \mathcal{D}_n^2$, which can be computed directly with the general formula, Eq. 27, for $r = 2$,

$$\mathcal{C}_n^2 = \sum_{h=0}^{n} \binom{n}{h} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{n}\right)^{n-h} \tag{31}$$

or its Szpankowski approximation for large n (needed for $n > 1000$ in practice) (Szpankowski, 2001; Kontkanen et al., 2003; Kontkanen, 2009),

$$\mathcal{C}_n^2 = \sqrt{\frac{n\pi}{2}} \left(1 + \frac{2}{3}\sqrt{\frac{2}{n\pi}} + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right) \tag{32}$$

$$\simeq \sqrt{\frac{n\pi}{2}} \exp\left(\sqrt{\frac{8}{9n\pi}} + \frac{3\pi - 16}{36n\pi}\right) \tag{33}$$

This leads to the following local score for each node X_i, which is minimized over alternative combinations of non-descendant neighbors, $\mathbf{Pa}'_{X_i} \subseteq \mathbf{Pa}_{X_i} \cup \mathbf{Sp}_{X_i} \cup \mathbf{Ne}_{X_i}$, in the first step of the local search-and-score algorithm (step 1) detailed in the main text,

$$\mathrm{Score_n}(X_i) = H(X_i|\mathbf{Pa}'_{X_i}) + \frac{1}{N}\sum_{j}^{q_{x_i}} \log \mathcal{C}_{n_j}^{r_{x_i}} \tag{34}$$

Note that Eq. 34 is defined for categorical variables for which the number of categories r_{x_i} of variable X_i, the number of combinatorial categories q_i of its non-descendant neighbors, and the corresponding number of samples n_j (for $j = 1, \cdots, q_i$) are readily obtained from the dataset. However, this node score can also be extended for continuous (or mixed-type) variables, by subtracting the unconditional entropy $H(X_i)$ to the conditional entropy $H(X_i|\mathbf{Pa}'_{X_i})$ in Eq. 34, to obtain an equivalent node score (up to a $H(X_i)$ constant) in terms of mutual information, $H(X_i|\mathbf{Pa}'_{X_i}) - H(X_i) = -I(X_i; \mathbf{Pa}'_{X_i})$. This extended node score can then be estimated following the optimization procedure proposed by Cabeli *et al* 2020 to estimate (conditional) mutual information between continuous (or mixed-type) variables. Besides, this extended node score leads to the same definition of edge scores, listed in Table 1 and presented below for categorical variables, which can similarly be estimated for continuous (or mixed-type) variables using the same optimization procedure proposed by Cabeli *et al* 2020.

Edge scores. We now define several edge scores to optimize the orientation of each edge, $X - Y$, given its close surrounding vertices.

To this end, we first introduce a local score for node pairs which simply sums the node scores, Eq. 34, for each node. The resulting pair scores are listed in Table D.1 for unconnected node pairs and for pairs of nodes connected by a directed edge, where $\mathbf{Pa}'_{X\setminus Y} = \mathbf{Pa}_X \cup \mathbf{Sp}_X \setminus Y$ and $\mathbf{Pa}'_{Y\setminus X} = \mathbf{Pa}_Y \cup \mathbf{Sp}_Y \setminus X$ with their corresponding combinations of levels, $q_{y\setminus x}$ and $q_{x\setminus y}$.

Table D.1. Local scores for node pairs

Pair score	Information	fNML Complexity
$X \not\to Y$	$H(X\|\mathbf{Pa}'_{X\setminus Y}) + H(Y\|\mathbf{Pa}'_{Y\setminus X})$	$\frac{1}{N}\Big(\sum_j^{q_{x\setminus y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\setminus x}} \log \mathcal{C}_{n_j}^{r_y} \Big)$
$X \to Y$	$H(X\|\mathbf{Pa}'_{X\setminus Y}) + H(Y\|\mathbf{Pa}'_{Y\setminus X}, X)$	$\frac{1}{N}\Big(\sum_j^{q_{x\setminus y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\setminus x}{}^{r_x}} \log \mathcal{C}_{n_j}^{r_y} \Big)$
$X \leftarrow Y$	$H(X\|\mathbf{Pa}'_{X\setminus Y}, Y) + H(Y\|\mathbf{Pa}'_{Y\setminus X})$	$\frac{1}{N}\Big(\sum_j^{q_{x\setminus y}{}^{r_y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\setminus x}} \log \mathcal{C}_{n_j}^{r_y} \Big)$

Then, edge scores for directed edges, $X \to Y$ and $Y \to X$, are defined w.r.t. to the edge removal score, $X \not\to Y$, by substracting the pair scores of unconnected pairs to the pair scores of directed edges, leading to the following edge orientation scores,

$$\mathrm{Score}(X \to Y) \;=\; -I(X;Y|\mathbf{Pa}'_{Y\setminus X}) + \frac{1}{N}\Big(\sum_j^{q_{y\setminus x}{}^{r_x}} \log \mathcal{C}_{n_j}^{r_y} - \sum_j^{q_{y\setminus x}} \log \mathcal{C}_{n_j}^{r_y} \Big) \tag{35}$$

$$\mathrm{Score}(Y \to X) \;=\; -I(X;Y|\mathbf{Pa}'_{X\setminus Y}) + \frac{1}{N}\Big(\sum_j^{q_{x\setminus y}{}^{r_y}} \log \mathcal{C}_{n_j}^{r_x} - \sum_j^{q_{x\setminus y}} \log \mathcal{C}_{n_j}^{r_x} \Big) \tag{36}$$

However, if $r_x \neq r_y$, the fNML complexities of these orientation scores are not identical for Markov equivalent edge orientations between nodes sharing the same parents (or spouses) (Chickering, 1995), $\mathbf{Pa}'_{Y\setminus X} = \mathbf{Pa}'_{X\setminus Y} = \mathbf{Pa}'$ and $q_{y\setminus x} = q_{x\setminus y}$, despite sharing the same conditional mutual information,

$$I(X;Y|\mathbf{Pa}') \;=\; \frac{1}{2}\Big(H(X|\mathbf{Pa}') + H(Y|\mathbf{Pa}', X)\Big) + \frac{1}{2}\Big(H(X|\mathbf{Pa}', Y) + H(Y|\mathbf{Pa}')\Big) \tag{37}$$

This suggests to symmetrize the fNML complexities for edge orientation scores by averaging them over each directed orientation, as for the conditional information in Eq. 37, leading to the proposed fNML complexity for directed edges given in Table 1 in the main text.

For bidirected edges, the proposed local orientation score accounts for all ac-connected subsets in close vicinity of the bidirected edge, which concerns all subsets including either X and any combination (possibly void) of parents or spouses different from Y (*i.e.* corresponding to the information contributions $H(X|\mathbf{Pa}'_{X\setminus Y})$) or Y and any combination of parents or spouses different from X (*i.e.* corresponding to the information contributions $H(Y|\mathbf{Pa}'_{Y\setminus X})$) or, else, including both nodes X and Y plus any combination of their parents or spouses, corresponding to the following information contribution, $-I(X;Y|\mathbf{Pa}'_{XY})$, where $\mathbf{Pa}'_{XY} = \mathbf{Pa}'_{X\setminus Y} \cup \mathbf{Pa}'_{Y\setminus X}$. This last term, $-I(X;Y|\mathbf{Pa}'_{XY})$, contains all the remaining information contributions once the bidirected orientation score is given relative to the edge removal score (Table D.1) as for the two directed orientation scores, above. Finally, the symmetrized fNML complexity associated with a bidirected edge should be computed with the whole set of conditioning parents or spouses, $\mathbf{Pa}'_{XY}$, as indicated in Table 1. Note that this bidirected orientation score becomes also Markov equivalent to the two directed orientation scores, as required, when the nodes share the same parents and spouses, *i.e.* $\mathbf{Pa}'_{XY} = \mathbf{Pa}'_{Y\setminus X} = \mathbf{Pa}'_{X\setminus Y}$ and $q_{xy} = q_{y\setminus x} = q_{x\setminus y}$ in Table 1.

E. Toy models, benchmark data generation and method settings

E.1. Toy models

Fig. E.1 shows three simple ancestral models used to test MIIC_search&score orientation scores (Table 1) to effectively predict bidirected orientations when the end nodes do not share the same parents (Model 1), share some parents (Model 2) or when the bidirected edge is part of a longer than two-collider paths (Model 3).

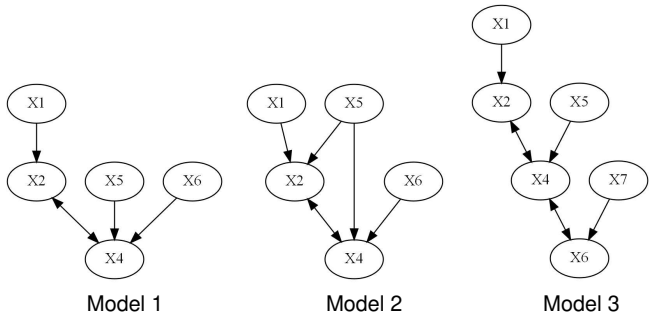

Figure E.1. **Simple ancestral graphs.**

The data is generated from the theoretical DAG using the rmvDAG function in the pcalg package (Kalisch et al., 2012). Each node follows a normal distribution, and the data is discretized using bnlearn's discretize function using Hartemink's pairwise mutual information method (Scutari, 2010). For these toy models, the edge orientation scores are computed assuming the correct parents of each node.

The prediction of the edge orientation scores are summarized in Table E.1 in % of replicates displaying directed edges (wrong) or bidirected edge (correct) as a function of increasing dataset size N.

Table E.1.

	Model 1, $X_2 - X_4$			Model 2, $X_2 - X_4$			Model 3, $X_2 - X_4$			Model 3, $X_4 - X_6$		
N	$\leftarrow$	$\rightarrow$	$\leftrightarrow$	$\leftarrow$	$\rightarrow$	$\leftrightarrow$	$\leftarrow$	$\rightarrow$	$\leftrightarrow$	$\leftarrow$	$\rightarrow$	$\leftrightarrow$
1000	92	6	2	42	54	4	98	2	0	0	100	0
5000	14	20	66	14	12	74	100	0	0	0	100	0
10000	0	2	98	0	0	100	88	0	12	0	76	24
20000	0	0	100	0	0	100	0	0	100	0	0	100
35000	0	0	100	0	0	100	0	0	100	0	0	100
50000	0	0	100	0	0	100	0	0	100	0	0	100

E.2. Generation of linear and non-linear continuous datasets

The benchmark ancestral graphs for continuous datasets were obtained by hiding 0%, 10% or 20% of variables in either linear or non-linear continuous Bayesian networks of 50 or 150 nodes of connectivity degree between 1 and 5 and average degree of 3 or 5, corresponding to 360 ancestral graphs in total.

Undirected graphs were generated using igraph's sample_degseq function with the vl option https://igraph.org/r/html/1.2.5/sample_degseq.html. DAGs were then obtained using a random ordering of their nodes.

Linear Gaussian datasets were generated from these DAGs using rmvDAG with uniformly distributed weights in $[-2, -0.1]$ and $[0.1, 2]$ (https://www.rdocumentation.org/packages/pcalg/versions/2.6-8/topics/rmvDAG).

Non-linear multimodal datasets were generated from these DAGs using Gaussian mixtures and non-linear structural equation models following (Cabeli et al., 2020):

- Nodes without parents: Gaussian mixtures with 2 to 4 modes randomly distributed in $[-20, 20]$ with Gaussian noise $\sigma = 1$, and subsequently rescaled betwenen 0 and 1.
- Nodes with parents: non-linear structural equation models including pairwise coupling between the parents and non-linear transforms (x^β, $\beta \in [1, 3]$, $\exp(x)$, $\sin(x)$, $\cos(x)$) (Cabeli et al., 2020).

E.3. Causal discovery methods and parameter settings

MIIC_search&score's source code is accessible at https://github.com/miicTeam/MIICsearchscore.

MIIC R package is accessible at https://github.com/miicTeam/miic_R_package.

MIIC and MIIC_search&score settings were set as described in section 3 of the main text.

M3HC was obtained at https://github.com/mensxmachina/M3HC and run with the following parameter settings: MaxCondSetM3HC = 10, Threshold = 0.05, Tol = 0.001, TABUsize = 1, skeleton = MMPC.

GFCI was obtained at https://github.com/cmu-phil/py-tetrad and run with the following parameter settings for linear data: depth = 5, max_disc_path_length = 5, SEM_BIC (penalty_discount = 5), Fisher Z test (α = 0.05) and for non-linear data: depth = 3, max_disc_path_length = 2, basis function BIC (truncation_limit = 3, penalty_discount = 2), degenerate Gaussian test (α = 0.05).

DAG-GNN was obtained from https://github.com/ronikobrosly/DAG_from_GNN and run using the following key parameters: encoder = MLP, decoder = MLP, optimizer = Adam, graph_threshold = 0.3, h_tol = 1e-8.
The training hyperparameters were adapted to the number of variables (N) as follows:

- N = 100, epochs = 300, batch_size = 20, lr = 0.002, k_max_iter = 100
- N = 250, epochs = 300, batch_size = 25, lr = 0.002, k_max_iter = 100
- N = 500, epochs = 200, batch_size = 50, lr = 0.001, k_max_iter = 80
- N = 1000, epochs = 200, batch_size = 100, lr = 0.001, k_max_iter = 80
- N = 5000, epochs = 100, batch_size = 250, lr = 0.0001, k_max_iter = 50
- N = 10000, epochs = 100, batch_size = 500, lr = 0.0001, k_max_iter = 50
- N = 20000, epochs = 75, batch_size = 1000, lr = 5e-05, k_max_iter = 20

FCI was obtained at https://github.com/py-why/causal-learn and run with G^2-conditional independence test for categorical data and default parameter $\alpha = 0.05$. For bootstrapped data, FCI was run with the following parameters: Alarm — α = 0.05; Insurance — α = 0.005, depth = 2, max_path_length = 1; Barley — α = 0.0005, depth = 1, max_path_length = 0; Mildew — α = 0.0005, depth = 1, max_path_length = 0.

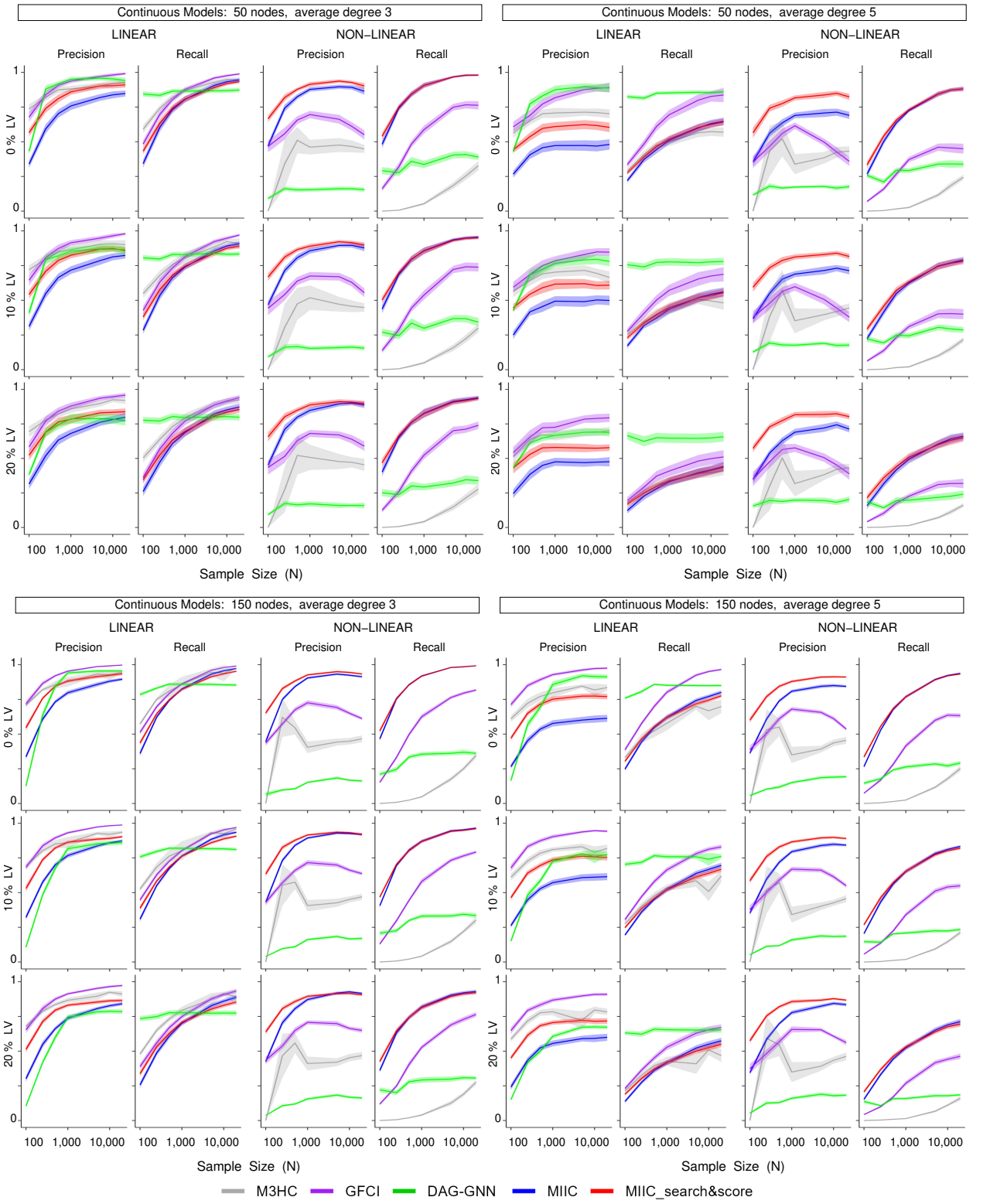

Figure E.2. **Benchmark results on continuous datasets.** Benchmark results are averaged over 30 independent ancestral graph models obtained by hiding 0%, 10% or 20% of variables in linear Gaussian models or more complex non-linear multimodal models including 50 or 150 nodes of average degree 3 or 5. MIIC_search&score results are compared to MIIC results used as starting point for MIIC_search&score, M3HC (Triantafillou & Tsamardinos, 2016), GFCI (Ogarrio et al., 2016), and DAG-GNN (Yu et al., 2019). Causal discovery performance is assessed in terms of *Precision* and *Recall* relative to the theoretical PAGs, while counting as false positive all correctly predicted edges but with a different orientation as the directed or bidirected edges of the PAG. Error bars: 95% confidence interval.

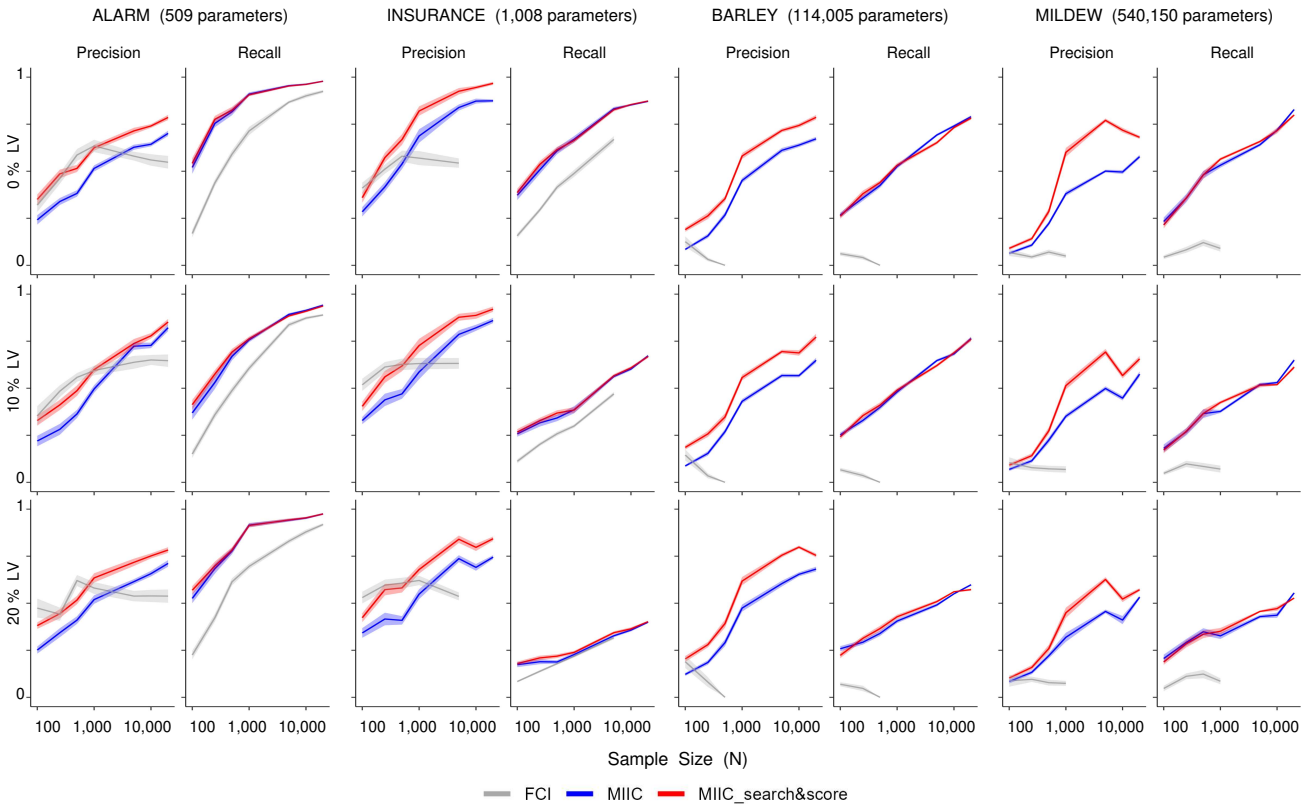

Figure E.3. **Benchmark results on bootstrapped 'real-world' categorical datasets from the bnlearn repository.** Benchmark results on bootstrap sensitivity analysis to sampling noise based on 30 independent resamplings with replacement of single categorical datasets of increasing sizes. Ancestral graphs are obtained by hiding 0%, 10% or 20% of variables in Discrete Bayesian Networks of increasing complexity (see main text): Alarm, Insurance, Barley and Mildew. Causal discovery performance is assessed in terms of *Precision* and *Recall* relative to the theoretical PAGs, while counting as false positive all correctly predicted edges but with a different orientation as the directed or bidirected edges of the PAG. MIIC_search&score results are compared to MIIC results used as starting point for MIIC_search&score and to FCI (Zheng et al., 2024). The lack of FCI results for complex models at large sample sizes stems from FCI difficulty to converge on bootstrapped datasets. Error bars: 95% confidence interval.