Ordering-based Causal Discovery from Discrete Data

Anonymous Authors[1]

Abstract

Learning DAG structures from purely observational data is a long-standing challenge, though significant progress has been made in recent years. An emerging line of research leverages the score of the data function to identify a topological order of the underlying DAG and ultimately perform causal discovery as combined with edge pruning algorithms. This paper extends the original score matching framework for causal discovery, which is originally designated for continuous data, and introduces a novel leaf discriminant criterion based on the discrete score function. Through simulated and real-world experiments, we demonstrate that our theory enables accurate inference of true causal orders from observed discrete data and that our identified ordering can significantly boost the accuracy of existing causal discovery baselines.

1. Introduction

Discovering the causal structure, often a *directed acyclic graph* (DAG), within a system of variables has long been an active pursuit across diverse scientific fields (Sachs et al., 2005; Richens et al., 2020; Wang et al., 2020). This paper focuses on causal discovery from observational data, a central problem in causality that presents two key challenges.

First, identifiability remains a major issue: multiple causal models can generate the same observational data distribution. To this end, certain assumptions on the data generative process are required to ensure the causal model is identifiable from purely observed data (Peters et al., 2010; 2014).

Second, structure learning is computationally intractable in the general case, as searching over the combinatorial space of DAGs is known to be NP-hard (Chickering, 1996; Chickering et al., 2004). An important fact one can possibly exploits is that any DAG has at least one topological order and the ordering exists if and only if it is a DAG. The prior knowledge of partial orderings is typically available in some real-world scenarios, such as genetics (Olson, 2006), healthcare (Denton et al., 2007) or meteorology (Bruffaerts et al., 2018). Incorporating such prior information can significantly reduces the complexity of DAG search since acyclicity constraint is naturally enforced given a causal order (Ban et al., 2024).

Ordering-based causal discovery is a line of research that addresses the case where partial orderings are not given (Teyssier & Koller, 2012; Bühlmann et al., 2014). The algorithm entails into two stages: (1) determining a topological ordering and (2) subsequent post-processing to remove spurious edges. Research in ordering-based causal discovery recently takes off with the use of score matching (Rolland et al., 2022; Sanchez et al., 2022; Montagna et al., 2023b;a; Xu et al., 2024), wherein a valid causal order can be estimated by sequentially identifying the leaf nodes based on the score of data distribution. The approach has proven practically effective and offers some robustness to noise misspecifications or to assumptions violations such as faithfulness and measurement errors (Montagna et al., 2024).

Despite their successes, ordering-based causal discovery frameworks with score matching are currently limited to continuous data. Extending the methods to discrete data remains a largely unexplored area. The core difficulty lies in the fact that the concept of a "score function" (i.e., Jacobian of the data log-likelihood) is not well-defined for discrete random variables. Our work is motivated by a fundamental question: **can the score matching paradigm be applied for recovering a causal order from discrete data?** Given the growing literature on surrogate "scores" for discrete data (Hyvärinen, 2007; Lyu, 2012; Meng et al., 2022; Sun et al., 2022), we investigate whether any of the proposed discrete score functions can effectively serve as a leaf node discriminant criterion. It turns out the answer is affirmative, and we further develop an identifiability result that guarantees the recovery of causal orders from observational discrete data.

Contributions In summary, our work presents the following contributions:

- We characterize the identifiability of a topological or-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

der underlying a discrete structural causal model and demonstrate how it can be estimated based on the *discrete score* of data distribution (see §3.3).

- Our theoretical results shed light on a novel connection of causal discovery with majorization theory (Marshall et al., 1979; Hickey, 1982; 1983) for quantifying the randomness of a system. While making no assumptions about the additive structure, our theory gives rise to a condition that generalizes existing sufficient conditions for identification of leaf node in non-linear and linear Gaussian additive models.
- We enrich the ordering-based literature with an extension to discrete data, while provide a fresh view to learning structures from discrete data, which is currently dominated by classical independence test-based and score-and-search approaches.
- Lastly, we validate our theory through synthetic and real-world experiments, demonstrating empirical effectiveness in recovering true causal orders, ultimately yielding a significant boost in the accuracy of the inferred causal structures.

2. Related Work

Causal discovery algorithms broadly fall into two categories: constraint-based methods, such as PC (Spirtes & Glymour, 1991) and FCI (Spirtes et al., 2000), which detect edge existence and direction by conditional independence tests; and score-based[1] methods that search for DAGs that optimizes a given objective/loss function function (Ott & Miyano, 2003; Chickering, 2002; Teyssier & Koller, 2012; Cussens et al., 2017). Research on continuous data particularly enjoys remarkable progress over the years, driven by the development of non-convex characterization of the acyclicity constraints. This gives rise to a family of scalable DAG learning frameworks via continuous optimization programs, notably Lachapelle et al. (2019); Zheng et al. (2020); Yu et al. (2019); Bello et al. (2022). We refer readers to Glymour et al. (2019); Vowels et al. (2022); Kitson et al. (2023) for excellent reviews of the related methods. In the following, we focus on ordering-based algorithms and structure learning approaches for discrete observational data.

Ordering-based Causal Discovery This family of methods often assume the (continuous) observational data is generated from an additive noise model. They first estimate a topological ordering of the causal variables, and prune the resulting fully connected DAG by some variable selection procedure. CAM (Bühlmann et al., 2014) is an early order-based approach; CAM uses a greedy search to determine the topological ordering and rely on significance tests to prune the DAG. Ghoshal & Honorio (2018) and Chen et al. (2019) infer the causal graph of linear additive modes, by sequentially identifying leaf nodes based on an estimation of the precision matrix. In the same spirit, Rolland et al. (2022) tackles non-linear Gaussian models and proposes to identify the leaf node by the Hessian matrix of the data log-likelihood. This method offers several advantages such as robustness to assumption violations (Montagna et al., 2024) or scalability in high-dimensional graphs (Montagna et al., 2023c). Additionally, it provides guarantees on finite sample complexity bounds (Zhu et al., 2024), further enhancing its appeal for practical applications. Several extensions to handle arbitrary continuous noise settings have recently been developed, including Sanchez et al. (2022); Montagna et al. (2023a;b;c); Xu et al. (2024).

Causal Discovery from Discrete Data Constraint-based causal discovery can be extended to discrete data with G-tests (Quine & Robinson, 1985) or $chi\text{-}squared$ tests (Cochran, 1952). Score-based methods, such as GES (Chickering, 2002; Teyssier & Koller, 2012), can be applied on multinomial Bayesian networks with BIC (Schwarz, 1978) or BDeu (Heckerman et al., 1995) scoring functions. However, it is well-known that graphs are generally identifiable up to the Markov equivalence class. To this end, several identifiability results have been proposed, under specific assumptions, for nominal/categorical data (Peters et al., 2010; Liu & Chan, 2016; Cai et al., 2018; Compton et al., 2020; Qiao et al., 2021), ordinal data (Luo et al., 2021; Ni & Mallick, 2022), or mixed data (Tsagris et al., 2018; Sedgewick et al., 2019; Wenjuan et al., 2018). The vast majority of these existing methods are designed for bivariate settings. Algorithmically, these approaches typically resort to constraint-based or score-based algorithms to search for the true graphs.

Score Matching Score matching is a family of parameter learning methods alternative to the maximum likelihood principle. The objective entails matching two log probability density functions by their first-order derivatives using the Fisher divergence metric. First introduced in (Hyvärinen & Dayan, 2005), score matching obviates the intractability of the normalizing partition functions as well as the ground-truth data score. and leads to a consistent estimate. Further developments in score estimation include kernel-based estimators (Li & Turner, 2017), denoising score matching (Vincent, 2011), slice score matching (Song et al., 2020), denoising likelihood score matching (Chao et al., 2022), and score-based generative modelling (Song & Ermon, 2019). In these line, the score function is learned by fitting a neural network minimizing the empirical Fisher divergence.

[1]The term *score* in traditional causality literature refers to an objective of a DAG optimization problem. This is to distinguish with the *score* of data distribution $\nabla \log p(x)$ in score matching literature.

Whereas representing a probability distribution by the score of its density has proven effective for continuous data, the notion of gradient is not defined for discrete modalities, rendering score matching inapplicable. To this end, a popular surrogacy of the typical score function is what is known as the *concrete* score (Meng et al., 2022), that is the ratio of two marginal probabilities for different state-value pairs $\frac{p(y)}{p(x)}$. Analogous to the score function $\nabla \log p(x)$, this ratio arises in the reverse process for discrete diffusion models, where the evolution of discrete variables is described through a continuous-time Markov chain (Anderson, 2012; Campbell et al., 2022; Sun et al., 2022; Lou et al., 2024), leading to a natural realization of score matching in discrete domains.

In this paper, we focus on categorical score estimation by matching marginal probabilities for each dimension. This approach, called **ratio matching**, is initially proposed by Hyvärinen (2007) for binary data, where it also preserves consistency (under some regularity conditions) and bypasses the computation of normalizing constant. Extensions to general discrete data are developed in Lyu (2012); Sun et al. (2022). We summarize the technicalities in §3.3.

3. Preliminaries

Notation We use upper case letters (e.g., X) for random variables and lower case letters (e.g., x) for values. We reserve bold capital letters (e.g., $\mathbf{G}$) for notations related to graphs and calligraphic letters (e.g., $\mathcal{X}$) for spaces. Finally, we use $[d]$ to denote a set of integers $\{1, 2, \cdots, d\}$.

This work deals with discrete random variables X of finite domain where each variable X_i has n_i states ($n_i \geq 2$) and its domain is $[n_i]$. Let $\mathcal{X} := \prod_{i=1}^{d} [n_i]$ denote the domain of X and $p(x)$ be the joint probability density function.

3.1. Structural Causal Model

A directed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ consists of a set of nodes $\mathbf{V}$ and an edge set $\mathbf{E} \subseteq \mathbf{V}^2$ of ordered pairs of nodes with $(v, v) \notin \mathbf{E}$ for any $v \in \mathbf{V}$ (one without self-loops). For a pair of nodes i, j with $(i, j) \in \mathrm{E}$, there is an arrow pointing from i to j and we write $i \to j$. Two nodes i and j are adjacent if either $(i, j) \in \mathbf{E}$ or $(j, i) \in \mathbf{E}$. If there is an arrow from i to j then i is a parent of j and j is a child of i. Let pa_i and ch_i denote the set of variables respectively associated with parents and children of node i in $\mathbf{G}$.

The data generative process for a set of random variables $X = \{X_i\}_{i \in [d]}$ is characterized via a *structural causal model* (SCM, Pearl, 2009) over the tuple $\langle U, X, f \rangle$ that generally consists of a sets of assignments

$$X_i := f_i\big(X_{\mathrm{pa}_i}, \epsilon_i\big), \quad i \in [d], \tag{1}$$

where $\{\epsilon_1, \cdots, \epsilon_d\}$ are mutually independent exogenous variables with strictly positive density. Given a joint distribution over the exogenous variables ϵ, the (deterministic) functions $f = [f_i]_{i \in [d]}$ define a joint distribution P_X over the endogenous variables X. An SCM induces a causal graph $\mathbf{G}$, which is often assumed to be a DAG. An important property of DAGs is that there exists a non-unique *topological ordering* $\pi = (\pi_1, \cdots, \pi_d)$ that represents directions of edges such that i comes before j in the ordering for every directed edge $(i, j) \in \mathbf{E}$, written as $\pi_i < \pi_j$ if $(i, j) \in \mathbf{E}$ where π_i, π_j denote the positions of nodes i and j in the ordering.

In this work, we make standard causal discovery assumptions: (1) the distribution P_X and the induced graph $\mathbf{G}$ satisfies Markov properties (Pearl, 2009) and (2) there are no latent confounders of the observed variables. This model allows the probability density of X to be factorized as:

$$p(x) = \prod_{i=1}^{d} p(x_i | x_{\mathrm{pa}_i}). \tag{2}$$

3.2. Score Matching for Causal Discovery

An important class of causal models for continuous data is additive noise model (ANM, Peters et al., 2014; Hoyer et al., 2008) where the graph $\mathbf{G}$ can be uniquely identifiable. In ANMs, (1) takes the form $X_i := f_i\big(X_{\mathrm{pa}_i}\big) + \epsilon_i, \quad i \in [d]$.

SCORE (Rolland et al., 2022) is the pioneering work that sheds light on the connection between score function and causal discovery. Assuming the model is a nonlinear ANM with Gaussian noise (i.e., the noise variables $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$), the authors show that the causal ordering of the DAG $\mathbf{G}$ can be recovered from the score function $\nabla \log p(x)$.

Given the Markovian factorization in (2), the joint log density under this model can be written as

$$\begin{aligned} \log p(x) &= \sum_{i=1}^{d} \log p(x_i | x_{\mathrm{pa}_i}) \\ &= -\frac{1}{2} \sum_{i=1}^{d} \left(\frac{x_i - f_i(x_{\mathrm{pa}_i})}{\sigma_i} \right)^2 - \frac{1}{2} \sum_{i=1}^{d} \log(2\pi\sigma_i^2). \end{aligned}$$

Thus, the score function $\mathbf{s}(x) := \nabla \log p(x)$ reads

$$\mathbf{s}_j(x) = -\frac{x_j - f_j(x_{\mathrm{pa}_j})}{\sigma_j^2} + \sum_{i \in \mathrm{ch}_j} \frac{\partial f_i}{\partial x_j}(\mathrm{pa}_i) \frac{x_i - f_i(x_{\mathrm{pa}_i})}{\sigma_i^2}.$$

It is observed that if j is a leaf node, then the second summand vanishes due to having no children. This gives rise to $\partial_j \mathbf{s}_j(x) := \frac{\partial \mathbf{s}_j(x)}{\partial x_j} = -1/\sigma_j^2$, thus $\mathrm{Var}_X[\partial_{x_j} \mathbf{s}_j(x)] = 0$. Rolland et al. (2022) shows that this condition is sufficient to identify a leaf of the graph.

The algorithm initially estimates the Jacobian of the score for all data points and then selects as a leaf node the diagonal element that yields the smallest variance over the data. The column corresponding to the selected leaf node is then removed from the data matrix, and the process is repeated until the entire ordering is determined. If the Hessian matrix is accurately estimated, a true causal order can be identified. Thereafter, the authors propose to apply CAM (Bühlmann et al., 2014) for eliminating spurious edges from the super DAG to recover the true graph.

SCORE nonetheless cannot distinguish leaf nodes in linear causal models, where the diagonal values of the score's Jacobian is constant for any nodes. Utilizing this fact, Xu et al. (2024) proposes an alternative leaf discriminant criterion applicable to both linear and non-linear relations, where the outer variance is replaced with expectation. A sufficient condition for identifiability is that the noise variances are non-decreasing, restated as follows:

Assumption 3.1. *(Non-decreasing variance of noises) (Xu et al., 2024) For any two noises ϵ_i and ϵ_j, $\sigma_i \leq \sigma_j$ if $\pi_i < \pi_j$.*

The non-decreasing variance condition extends the standard equal variance assumption in previous literature e.g., Peters & Bühlmann (2014); Ghoshal & Honorio (2018) and can be regarded as a representation of prior knowledge about the uncertainty inherent in the system.

3.3. Generalized Score Matching

In this section, we briefly describe the generalized score matching principle proposed in Lyu (2012). We note that despite the conceptual similarities, this generalized "version" of score function departs from the diffusion setup of concrete score matching. We later show how the following generalization facilitates the identification of a topological order of $\mathbf{G}$ from discrete data.

Given two probability densities $p(x)$ and $q(x)$ and a linear operator (functional) $\mathcal{L}$, the *generalized Fisher divergence* is defined as

$$D_{\mathcal{L}}(p\|q) = \sum_{\mathcal{X}} p(x) \left| \frac{\mathcal{L}p(x)}{p(x)} - \frac{\mathcal{L}q(x)}{q(x)} \right|^2, \tag{3}$$

where $\frac{\mathcal{L}p(x)}{p(x)}$ is termed as *generalized score function*. A valid linear operator $\mathcal{L}$ should be complete, meaning that two densities $p(x) = q(x)$ (a.e) if $p(x)$ and $q(x)$ satisfies $\frac{\mathcal{L}p(x)}{p(x)} = \frac{\mathcal{L}q(x)}{q(x)}$ (a.e). It is easy to see that the gradient operator ∇ is complete, under which $D_{\mathcal{L}}$ reduces to the original Fisher divergence, since $\nabla \log p(x) = \nabla \frac{p(x)}{p(x)}$.

For discrete data, Lyu (2012) proposes to choose $\mathcal{L}$ to be the marginalization operator $\mathcal{M}$. Let $\mathcal{M}_i p(x) := p(x_{-i}) = \sum_{x_i} p(x)$ be the marginal density induced from $p(x)$, where x_{-i} denote the vector formed by dropping x_i from x. This gives rise to

$$\frac{\mathcal{M}_i p(x)}{p(x)} = \frac{p(x_{-i})}{p(x)} = \frac{1}{p(x_i|x_{-i})}. \tag{4}$$

The **discrete score function** is thus defined as $\mathcal{M}p(x) := [\mathcal{M}_i p(x)]_{i=1}^d$ where each $\mathcal{M}_i p(x)$ is a reciprocal of the singleton conditional density $p(x_i|x_{-i})$.

The operator $\mathcal{M}$ is complete due to a well-known result in statistics (Brook, 1964; Lyu, 2012) that the joint density $p(x)$ is completely determined by the ensemble of the singleton conditionals $p(x_i|x_{-i}), \forall i \in [d]$.

It can be seen that the normalizing constant does not affect the computation as it gets cancelled out in the generalized score function. The generalized Fisher divergence can also be re-expressed into a form as an expectation of functions of the unnormalized model, which enables Monte Carlo sampling for estimation. It is worth noting that the above construction is also applicable to continuous data where the summation is replaced with integration.

4. Ordering-based Causal Discovery via Discrete Score Matching

From this point we will mainly deal with the singleton conditional densities $p(x_i|x_{-i})$, which are referred to as the **reciprocal discrete score functions**.

Our task is to identify a criterion to discriminate leaf nodes of a causal graph from i.i.d observational samples. Motivated by the **non-decreasing variance** condition (Park, 2020; Xu et al., 2024), we investigate whether the knowledge of the system's uncertainty can facilitate the identification of the leaf variables in a discrete SCM.

Translation of the non-decreasing variance condition to discrete variables, particular categorical ones, is however not straightforward, as they lack inherent quantitative values that can directly reflect the system's uncertainty. As revealed shortly, there fortunately exists a broad class of randomness measures of discrete probability distributions that only deals with the probabilities rather than the values on the associated sample space. Building on this construction, we develop a generalized framework for characterizing the system's randomness. This framework plays a crucial role in identifying the leaf variable with the reciprocal discrete score function.

Let $\mathcal{P}$ denote a class of all discrete probability vectors. With no loss of generality we assume that all the probabilities distributions we deal with have been ordered in non-increasing order. We also assume the vectors have an equal length of $n = \max(n_1, \cdots, n_d)$ by properly padding the shorter one with the appropriate number of 0's at the end.

Definition 4.1. *Given two probability distributions* $\mathbf{a} = (a_1, \cdots, a_n)$ *and* $\mathbf{b} = (b_1, \cdots, b_n)$ *with* $a_1 \geq \cdots a_n \geq 0$ *and* $b_1 \geq \cdots b_n \geq 0$*, we say that* $\mathbf{a}$ *majorizes* $\mathbf{b}$*, written as* $\mathbf{a} \succeq \mathbf{b}$*, if and only if*

$$\sum_{i=1}^{k} a_i \geq \sum_{i=1}^{k} b_i, \quad \textit{for all } k = 1, \cdots, n.$$

Hickey (1982; 1983) formalizes the randomness or spreadness of a discrete probability distribution via majorization theory (Marshall et al., 1979). For two discrete distributions $\mathbf{a}$ and $\mathbf{b}$, we say $\mathbf{a}$ is more spread or more uniform/random than $\mathbf{b}$ if $\mathbf{a} \preceq \mathbf{b}$. A function $\phi : \mathbb{R}^n \mapsto \mathbb{R}$ is a *Schur-concave* function if $\phi(\mathbf{a}) \leq \phi(\mathbf{a})$ for all vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ such that $\mathbf{a} \succeq \mathbf{b}$. An interesting fact is the uniform vector i.e., $(\frac{1}{n}, \cdots, \frac{1}{n})$ is majorized by all probability vectors (thus being most random) and a degenerate vector e.g., $(1, 0, \cdots 0)$ and its permutations majorize all other vectors (thus being least random). These properties motivate the construction a measure of randomness as follows:

Definition 4.2. *(Hickey, 1982) A real-valued continuous function* ϕ*, taking finite values in* $\mathcal{P}$ *is a measure of randomness if it is symmetric and concave, and the concavity being strict on the sub-class of distribution finite number of positive probabilities.*

A popular class of measures of randomness has the form:

$$\phi(\mathbf{p}) = \sum_{k=1}^{n} g\left(p_k\right), \tag{5}$$

where $g : \mathbb{R} \mapsto \mathbb{R}$ is continuous, strictly concave with $g(0) = 0$. The entropy function is one of the best-known measures of the above form, where $g(p_k) = -p_k \log p_k$.

We are now ready to state our identifiability results. First, we need the following assumption to ensure the singleton conditional densities are defined.

Assumption 4.3. *Let* $x \in \mathcal{X}$ *be a discrete random vector defined by an SCM (1). For any node* $i \in [d]$*, the conditional densities* $p(x_i|x_{\mathrm{pa}_i})$ *are non-zero* $\forall x \in \mathcal{X}$*.*

The assumption also implies the conditional densities are non-degenerate for all variables. With a slight abuse of notation, let $\phi(X)$ denote the randomness, under ϕ, in the probability vector of the distribution of X. Let X and Y be joint distributed discrete random variables. The conditional information about X given Y is defined as $\phi(X|Y) = \mathbb{E}_Y\left[\phi(X|y)\right]$, accordingly in the probability vector $p(X|y)$ for a given value y.

Definition 4.4. *(Non-decreasing randomness) Given a valid topological ordering* π *of the true graph* $\mathbf{G}$ *and a measure of randomness* ϕ *as defined in (4.2),* ϕ *is said to satisfy non-decreasing randomness if for any two nodes* $i, j \in [d]$ *such that* $\pi_i < \pi_j$*, one has* $\phi(X_i|X_{\mathrm{pa}_i}) \leq \phi(X_j|X_{\mathrm{pa}_j})$*.*

The non-decreasing randomness property characterizes along the causal order the relative uncertainty among local densities (representing independent local generative systems). It indicates that, intuitively, the root nodes should carry the least randomness as it only depends on the noise variables. Meanwhile, other variables inherit the uncertainty from both the noises and their parent variables. For a leaf node particularly, it also indirectly accumulates uncertainty from the entire system, thus likely to be more random. One can see that for any pair of nodes i, j such that $\pi_i < \pi_j$, if the probability vector $p(X_i|x_{\mathrm{pa}_i})$ majorizes $p(X_j|x_{\mathrm{pa}_j}), \forall x \in \mathcal{X}$, the non-decreasing randomness property holds for any function ϕ defined in (4.2).

Theorem 4.5. *Let* $x \in \mathcal{X}$ *be a discrete random vector defined via an SCM (1), and let* $\mathbf{r}_i(x_{-i}) := p(X_i|x_{-i})$ *be the reciprocal discrete score function for every node* $i \in [d]$*. If there exists a randomness measure* ϕ *satisfying the non-decreasing randomness property w.r.t the true graph* $\mathbf{G}$*, then* X_j *is a leaf node* $\Leftrightarrow$ $j = \arg\max_{i \in [d]} \mathbb{E}_{X_{-i}}\Big[\phi\big(\mathbf{r}_i(x_{-i})\big)\Big]$.

We say that the leaf variable X_l is ϕ-identifiable if Theorem 4.5 holds for a certain measure ϕ. In connection with the previous literature, it is natural to ask whether the (expected conditional) variance function is applicable. Let us consider the function $\mathrm{Var}(\mathbf{p}) = \sum_{k=1}^{n} p_k(\log p_k - \mu)^2$ with $\mu = \sum_{k=1}^{n} p_k \log p_k$. It is well-known that the variance function is convex in the variables. In the presence of symmetry, convexity implies Schur-convexity. Hence, $\mathrm{Var}(\mathbf{p})$ is Schur-convex, thus its negative, defined as $\phi_{Var}(\mathbf{p}) := -\mathrm{Var}(\mathbf{p})$ is Schur-concave and qualifies as a randomness measure. The variance function can therefore be used for causal order search. This result is formalized in the following corollary.

Corollary 4.6. *Let* $x \in \mathcal{X}$ *be a discrete random vector defined via an SCM (1), and let* $\mathbf{r}_i(x_{-i}) := p(X_i|x_{-i})$ *be the reciprocal discrete score function for every node* $i \in [d]$*. If* ϕ_{Var} *satisfying the non-decreasing randomness property w.r.t the true graph* $\mathbf{G}$*, then* X_j *is a leaf node* $\Leftrightarrow$ $j = \arg\min_{i \in [d]} \mathbb{E}_{X_{-i}}\Big[\mathrm{Var}\big(\mathbf{r}_i(x_{-i})\big)\Big]$.

Corollary 4.6 simply follows from Theorem 4.5 and the proof is direct from Schur-convexity of the $\mathrm{Var}(\mathbf{p})$. Furthermore, one may notice that in additive noise models, the uncertainty of the system is entirely captured in the noise variables. In this case, non-decreasing randomness of the local densities is reduced to non-decreasing randomness of the corresponding noise variables. If the variance function is considered as the measure ϕ, our condition (4.4) can be viewed as a generalization of the equal/non-decreasing variance of noises introduced in the previous literature.

Let $\phi_H(\mathbf{p}) = \sum_{n}^{k=1} -p_k \log p_k$ be the entropy function and $\phi_U(\mathbf{p}) = \sum_{n}^{k=1} \log p_k$ be the sum of logarithmic probabilities. Let us denote $\phi_{\mathrm{KL}}(\mathbf{p}) = -\mathrm{KL}(\mathbf{p}\|\mathbf{u}) - \mathrm{KL}(\mathbf{u}\|\mathbf{p})$,

where $\mathbf{u}$ denotes the uniform distribution of appropriate dimension.

Proposition 4.7. *For ϕ_H and ϕ_U defined above, if both measures satisfy the non-decreasing randomness property w.r.t the true graph $\mathbf{G}$, then the leaf variable X_l is identifiable from $\phi_{\mathrm{KL}}(\cdot)$ as defined above.*

If multiple functions ϕ satisfy the non-decreasing randomness condition, the leaf variable may also be identifiable from their linear combination. Proposition 4.7 introduces an interesting instance. It is clear that ϕ_{KL} is non-positive and achieves the maximum at zero when $\mathbf{p}$ is uniform. Higher ϕ_{KL} thus indicates more randomness in a distribution. We provide the proofs for the above results in Appendix A. We now empirically verify the proposals in 4.6 and 4.7 through numerical experiments.

Estimation of Discrete Score Function We employ the continuous-time discrete diffusion framework proposed in Sun et al. (2022) to estimate the singleton conditionals. The framework generalizes the ratio matching objective for binary variables from Hyvärinen (2007). The objective elegantly circumvents the calculation of the data score function and its minimizer is shown to be consistent.

As for the parameterization of the score function, Sun et al. (2022) introduces an efficient Transformer architecture that only requires $O(1)$ forward evaluations, which is adopted in our implementation. The model is designed in an amortized fashion where an entire ensemble of singletons is returned per input. In our implementation, models are trained with Adam optimizer (Kingma, 2014) at fixed 300 epochs, 3000 time steps and learning rates of 0.0001. The size of hidden units is set as $2d$ where d is the number of variables in the data (i.e., sequence length). For details on architecture design, we refer readers to §4.1 and §5.3 in Sun et al. (2022). We recap the fundamentals of score matching in Appendix B and the categorical ratio matching objective in Eq. (15).

Remark In terms of the order search alone, the time complexity is linear in the number of nodes. The dominant factor is the training time of the continuous-time diffusion model for estimating the discrete score functions. The algorithm involves recursively estimating the score function from the data where the identified leaf variable is removed. Consequently, a new model must be trained at every iteration, which unfortunately increases the training time. As our current work focuses on establishing the identifiability theory, we leave the exploration of more efficient training regimes to future research. We summarize our causal order search procedure in Algorithm 1.

Algorithm 1 Causal Order Search with Discrete SCORE

Input: Data matrix $X \in [n]^{N\times d}$ and base measure ϕ.
Output: Topological ordering π.
Initialize $\pi = [\quad]$, nodes = $\{1, \cdots, d\}$
for $i = 1, \cdots, d$ **do**
  Estimate the conditionals set $\left\{p(X_j|x_{-j})\right\}_{j\in\text{nodes}}$ with a continuous-time diffusion model by Eq. (15).
  Estimate $V_j = \mathbb{E}\Big[\phi\big(p(X_j|x_{-j})\big)\Big], \forall j \in$ nodes.
  Find leaf $l \leftarrow$ nodes $[\arg\max_j V_j]$.
  Update $\pi \leftarrow [l, \pi]$, nodes $\leftarrow$ nodes $- \{l\}$.
  Remove l-the column of X.
end for

5. Experimental Setup

Datasets We evaluate the effectiveness of our proposed framework on both simulated and real-world datasets. We generate random DAGs from Erdos-Rényi (ER) or Scale-Free (SF) with number of nodes d up to 60 nodes and expected node degrees at $2d$ (ER2) and $4d$ (ER4). We use `pgmpy`[2] library (Ankan & Textor, 2024) to construct a Bayesian network based on the generated structures and populate the conditional probability distributions with normalized uniformly random weights. This strategy aims to create a system of approximately constant randomness, thus enabling the verification of our identifiability results. The cardinality of variables runs from 3 to 6, and samples of $10,000$ observations are simulated from the given models.

We additionally experiment with Sachs dataset (Sachs et al., 2005), a popular benchmark of causal discovery with the ground-true causal network of protein signalling pathways. We analyze the preprocessed interventional dataset[3] with 11 categorical features and total of 5400 samples across 6 experimental conditions. For each synthetic setting, we generate 10 random datasets. For every experiment, we run our models at 10 different initializations and report the average results. Our codes are anonymously published at `anonymous.4open.science/r/discrete-SCORE-C1F3/`.

Metrics Rolland et al. (2022) proposes D_{top}, a topological divergence metric quantifying the number of edges that cannot be recovered due to the errors in the topological order. For an ordering π and a target adjacency matrix A, the metric is defined as

$$D_{top}(\pi, A) = \sum_{j=1}^{d} \sum_{i:\pi_j>\pi_i} A_{ji}.$$

[2] `pgmpy.org/index.html`
[3] available at `bnlearn.com/book-crc/code/sachs.interventional.txt.gz`

Table 1. Synthetic experiment for ER graphs of $2d$ degree.

d	5			10			15			20		
ER2	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}
SCORE + PC (Ours)	2.40	**0.78**	1.50	5.20	**0.82**	2.00	8.00	**0.70**	1.50	9.00	**0.77**	3.00
PC	1.80	0.56	-	6.40	0.50	-	9.40	0.48	-	16.00	0.45	-
SCORE + GAM (Ours)	1.40	**0.82**	1.50	5.00	**0.71**	2.00	5.60	**0.75**	1.50	6.60	**0.80**	3.00
GAM	0.80	0.64	-	3.80	0.58	-	3.20	0.62	-	4.20	0.62	-
OCD	2.60	0.58	-	8.40	0.42	-	11.60	0.45	-	14.40	0.49	-

Table 2. Synthetic experiment for ER graphs of $2d$ degree.

d	30			40			50			60		
ER2	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}
SCORE + PC (Ours)	12.80	**0.72**	5.00	16.60	**0.75**	3.50	20.20	**0.72**	12.00	19.20	**0.76**	6.00
PC	20.40	0.45	-	26.60	0.47	-	31.60	0.48	-	37.60	0.47	-
SCORE + GAM (Ours)	9.00	**0.76**	5.00	12.80	**0.79**	3.50	18.00	**0.73**	12.00	21.40	**0.74**	6.00
GAM	4.60	0.62	-	8.40	0.61	-	19.20	0.50	-	12.80	0.60	-
OCD	16.00	0.50	-	29.00	0.41	-	23.00	0.52	-	38.00	0.50	

Table 3. Synthetic experiment for SF graphs of $2d$ degree.

d	5			10			15			20		
SF2	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}
SCORE + PC (Ours)	3.00	**0.71**	0.20	5.00	**0.72**	0.80	10.60	**0.50**	3.80	15.60	**0.44**	4.80
PC	3.20	0.44	-	6.40	0.46	-	13.00	0.40	-	19.80	0.35	-
SCORE + GAM (Ours)	0.60	**0.89**	0.20	4.80	**0.60**	0.80	9.20	**0.48**	3.80	12.80	0.46	4.80
GAM	0.40	0.63	-	4.20	0.51	-	7.40	0.47	-	10.00	**0.48**	-
OCD	1.80	0.56	-	6.00	0.44	-	10.00	0.40	-	14.00	0.35	-

Table 4. Experiment on Sachs dataset.

d	11		
SACHS	SHD	F1	D_{top}
SCORE + PC (Ours)	38.20	**0.39**	3.80
PC	35.00	0.28	-
SCORE + GAM (Ours)	30.20	**0.29**	3.80
GAM	29.00	0.25	-
OCD	38.00	0.24	-

If a node j appears after node i in the true ordering, i.e., $\pi_j > \pi_i$, the edge $j \to i$ must not exist. $D_{top}(\pi, A)$ returns zero if π is a correct order. As a result, the inferred partial order introduces the forbidden links from a node to its preceding nodes, which can be used to impose constraints on DAG search algorithms. Hence, the quality of the estimated order can be further assessed by how well the provided knowledge from the ordering can improve causal discovery baselines. For comparing the estimated DAG with the ground-truth one, we report the commonly used metrics: F1 score and Structural Hamming Distance (SHD). F1 score measures the balances between precision and recall, while SHD counts the smallest number of edge additions, deletions, and reversals required to transform the recovered DAG into the true one. D_{top} is thus a lower bound on the SHD of the final algorithm. Lower D_{top}, SHD (↓) and higher F1 are desirable (↑).

Baselines We investigate the representative methods of 3 families of causal discovery approaches for discrete data: (1) constraint-based methods with the PC algorithm (Spirtes & Glymour, 1991), (2) OCD (Ni & Mallick, 2022), a recent score-based algorithm, and (3) generalized additive models (GAM, Wood, 2017). We report the performance of the PC algorithm with G-tests (Quine & Robinson, 1985) at p-value of 0.5, which gives the best results in most of our experiments. Though designed for ordinal data, OCD (Ni & Mallick, 2022) is empirically shown to achieve better accuracy and scalability than traditional score-based methods in various real-world settings. We adopt the authors' proposed configuration that uses BIC score for greedy search. Lastly, we explore additive models for variable selection by fitting a multinomial logistic regression model with factor covariates. We begin with a fully connected graph and prune redundant edges based on a cut-off value of 0.0001.

6. Results & Discussion

Tables 1-4 report the results of our experiments in recovering the topological ordering and the true causal graph. `SCORE+X` refers to the application of our inferred causal order on a structure learning baseline `X`. As shown, our method is capable of recovering a true causal order with relatively low errors.

Once a causal order is found, the next step is pruning to recover the DAG. Methods for continuous data often rely on regression to identify parent variables, which requires knowledge of the appropriate model forms. However, our approach does not assume an additive structure, making it more general but also more challenging in terms of selecting suitable pruning methods. As a result, the performance of DAG recovery inherently depends on the underlying causal discovery algorithms. Here, our main goal is to assess the extent to which the inferred causal order can enhance performance, rather than to achieve state-of-the-art results. Nevertheless, ordering-based causal discovery offers flexibility, as it can be applied on top of any existing algorithm.

In our experiments, we consider PC algorithm and GAM for post-processing as OCD gives sub-optimal performance. It is crucial to note that because SHD quantifies the number of errors in absolute value, given a sparse graph, a method could achieve low SHD by predicting few edges, which obviously would compromise the accuracy score. Therefore, one need to examine both SHD and F1 metrics to assess the causal discovery effectively thoroughly.

Concretely, using the knowledge of an topological ordering to prune forbidden edges, given a correct one, both lower SHD and higher F1 scores are expected. However, if the inferred causal order is imperfect, this would result in an increased number of false deletions, increasing SHD. Therefore, we expect that an effective algorithm should yield a boost in accuracy (i.e., higher F1 score) with minimal increase in SHD. It can be observed that our algorithm achieves this goal in nearly all settings, with a significant improvement in accuracy in the baselines on both simulated and real-world settings.

It is worth noting that our simulated data process is in fact quite general, which assumes no precise knowledge of the base measure. We find that two proposed choices of ours in Corollary 4.6 and Proposition 4.7 yield the best performance in the synthetic and real-world experiments respectively. We have also experimented with other variants, such as ϕ_U or entropy function ϕ_H alone. Though these alternatives give slightly higher D_{top} on average, they exhibit the same behaviour with a consistent boost in F1 score. Additional evidence on $4d$ degree settings are provided in Appendix C.

7. Conclusion

In this work, we have explored the application of discrete score matching to causal discovery and contributed to the current score matching literature a new identifiability result to infer causal orders from observational discrete data. One limitation of the proposed algorithm is that it requires iteratively training diffusion models to estimate the discrete score function, leading to increased computational time. Future improvements in model design, through potentially adaptive masking strategies, may enable single-model approximation of the discrete scores with various data patterns of missing columns. This could effectively reduce the order search complexity strictly to $O(d)$.

While the partial knowledge of a topological ordering yields improvement in accuracy, the DAG learning performance is still constrained by the quality of estimation of the score functions as well as the post-processing methods, a persisting challenge of this line of research. Furthermore, as in any identifiability theory, the non-decreasing randomness condition necessitates prior knowledge about the data-generating process, which may be untestable in some cases. Future research could explore methods to verify these assumptions and identify appropriate randomness measures in practical applications.

Impact Statement

This paper presents the use of machine learning to efficiently solve a class of statistical estimation problems in a scalable way. Although we are not aware of any immediate negative societal effects of our approach, machine learning often leads to unforeseen consequences across different fields. Therefore, it is important to carefully assess both the potential benefits and risks to society when applying the proposed method in practical settings.

References

Anderson, W. J. *Continuous-time Markov chains: An applications-oriented approach*. Springer Science & Business Media, 2012.

Ankan, A. and Textor, J. pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research*, 25(265):1–8, 2024. URL http://jmlr.org/papers/v25/23-0487.html.

Ban, T., Chen, L., Wang, X., Wang, X., Lyu, D., and Chen, H. Differentiable structure learning with partial orders. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Bello, K., Aragam, B., and Ravikumar, P. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Brook, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51 (3/4):481–483, 1964.

Bruffaerts, N., De Smedt, T., Delcloo, A., Simons, K., Hoebeke, L., Verstraeten, C., Van Nieuwenhuyse, A., Packeu, A., and Hendrickx, M. Comparative long-term trend analysis of daily weather conditions with daily pollen concentrations in brussels, belgium. *International journal of biometeorology*, 62:483–491, 2018.

Bühlmann, P., Peters, J., and Ernest, J. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Cai, R., Qiao, J., Zhang, K., Zhang, Z., and Hao, Z. Causal discovery from discrete data using hidden compact representation. *Advances in neural information processing systems*, 31, 2018.

Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Chao, C.-H., Sun, W.-F., Cheng, B.-W., Lo, Y.-C., Chang, C.-C., Liu, Y.-L., Chang, Y.-L., Chen, C.-P., and Lee, C.-Y. Denoising likelihood score matching for conditional score-based data generation. *arXiv preprint arXiv:2203.14206*, 2022.

Chen, W., Drton, M., and Wang, Y. S. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4): 973–980, 2019.

Chickering, D. M. Learning bayesian networks is np-complete. *Learning from data: Artificial intelligence and statistics V*, pp. 121–130, 1996.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3 (Nov):507–554, 2002.

Chickering, M., Heckerman, D., and Meek, C. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

Cochran, W. G. The $\chi 2$ test of goodness of fit. *The Annals of mathematical statistics*, pp. 315–345, 1952.

Compton, S., Kocaoglu, M., Greenewald, K., and Katz, D. Entropic causal inference: Identifiability and finite sample results. *Advances in Neural Information Processing Systems*, 33:14772–14782, 2020.

Cussens, J., Haws, D., and Studenỳ, M. Polyhedral aspects of score equivalence in bayesian network structure learning. *Mathematical Programming*, 164:285–324, 2017.

Denton, B., Viapiano, J., and Vogl, A. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health care management science*, 10:13–24, 2007.

Ghoshal, A. and Honorio, J. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1466–1475. PMLR, 2018.

Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Heckerman, D., Geiger, D., and Chickering, D. M. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243, 1995.

Hickey, R. J. A note on the measurement of randomness. *Journal of Applied Probability*, 19(1):229–232, 1982.

Hickey, R. J. Majorisation, randomness and some discrete distributions. *Journal of applied probability*, 20(4):897–902, 1983.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Hyvärinen, A. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., and Chobtham, K. A survey of bayesian network structure learning. *Artificial Intelligence Review*, 56(8):8721–8814, 2023.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2019.

Li, Y. and Turner, R. E. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

Liu, F. and Chan, L. Causal inference on discrete data via estimating distance correlations. *Neural computation*, 28 (5):801–814, 2016.

Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024.

Luo, X. G., Moffa, G., and Kuipers, J. Learning bayesian networks from ordinal data. *Journal of Machine Learning Research*, 22(266):1–44, 2021.

Lyu, S. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.

Marshall, A. W., Olkin, I., and Arnold, B. C. Inequalities: theory of majorization and its applications. 1979.

Meng, C., Choi, K., Song, J., and Ermon, S. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35: 34532–34545, 2022.

Montagna, F., Noceti, N., Rosasco, L., and Locatello, F. Shortcuts for causal discovery of nonlinear models by score matching. *arXiv preprint arXiv:2310.14246*, 2023a.

Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pp. 726–751. PMLR, 2023b.

Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. Scalable causal discovery with score matching. In *Conference on Causal Learning and Reasoning*, pp. 752–771. PMLR, 2023c.

Montagna, F., Mastakouri, A., Eulig, E., Noceti, N., Rosasco, L., Janzing, D., Aragam, B., and Locatello, F. Assumption violations in causal discovery and the robustness of score matching. *Advances in Neural Information Processing Systems*, 36, 2024.

Murphy, K. P. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

Ni, Y. and Mallick, B. Ordinal causal discovery. In *Uncertainty in Artificial Intelligence*, pp. 1530–1540. PMLR, 2022.

Olson, E. N. Gene regulatory networks in the evolution and development of the heart. *Science*, 313(5795):1922–1927, 2006.

Ott, S. and Miyano, S. Finding optimal gene networks using biological constraints. *Genome Informatics*, 14:124–133, 2003.

Park, G. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.

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

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

Peters, J., Janzing, D., and Schölkopf, B. Identifying cause and effect on discrete data using additive noise models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 597–604. JMLR Workshop and Conference Proceedings, 2010.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. 2014.

Qiao, J., Bai, Y., Cai, R., and Hao, Z. Learning causal structures using hidden compact representation. *Neurocomputing*, 463:328–333, 2021.

Quine, M. P. and Robinson, J. Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. *The Annals of Statistics*, pp. 727–742, 1985.

Richens, J. G., Lee, C. M., and Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020.

Rolland, P., Cevher, V., Kleindessner, M., Russell, C., Janzing, D., Schölkopf, B., and Locatello, F. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Sanchez, P., Liu, X., O'Neil, A. Q., and Tsaftaris, S. A. Diffusion models for causal discovery via topological ordering. *arXiv preprint arXiv:2210.06201*, 2022.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, pp. 461–464, 1978.

Sedgewick, A. J., Buschur, K., Shi, I., Ramsey, J. D., Raghu, V. K., Manatakis, D. V., Zhang, Y., Bon, J., Chandra, D., Karoleski, C., et al. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics*, 35 (7):1204–1212, 2019.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020.

Spirtes, P. and Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

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

Sun, H., Yu, L., Dai, B., Schuurmans, D., and Dai, H. Score-based continuous-time discrete diffusion models. *arXiv preprint arXiv:2211.16750*, 2022.

Teyssier, M. and Koller, D. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *arXiv preprint arXiv:1207.1429*, 2012.

Tsagris, M., Borboudakis, G., Lagani, V., and Tsamardinos, I. Constraint-based causal discovery with mixed data. *International journal of data science and analytics*, 6: 19–30, 2018.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Vowels, M. J., Camgoz, N. C., and Bowden, R. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Wang, Y., Liang, D., Charlin, L., and Blei, D. M. Causal inference for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pp. 426–431, 2020.

Wenjuan, W., Lu, F., and Chunchen, L. Mixed causal structure discovery with application to prescriptive pricing. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 5126–5134, 2018.

Wood, S. N. *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2017.

Xu, Z., Li, Y., Liu, C., and Gui, N. Ordering-based causal discovery for linear and nonlinear relations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. Pmlr, 2020.

Zhu, Z., Locatello, F., and Cevher, V. Sample complexity bounds for score-matching: causal discovery and generative modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

A. Proof

Theorem 4.5. *Let $x \in \mathcal{X}$ be a discrete random vector defined via an SCM (1), and let $\mathbf{r}_i(x_{-i}) := p(X_i|x_{-i})$ be the reciprocal discrete score function for every node $i \in [d]$. If there exists a randomness measure ϕ satisfying the non-decreasing randomness property w.r.t the true graph $\mathbf{G}$, then X_j is a leaf node $\Leftrightarrow j = \arg\max_{i\in[d]} \mathbb{E}_{X_{-i}}\Big[\phi\big(\mathbf{r}_i(x_{-i})\big)\Big]$.*

Proof. A measure ϕ as defined in (4.2) is a valid measure of randomness in the sense that it satisfies the conditional information property: if X and Y are discrete random variables defined on a finite sample space, the expected randomness remaining in X after Y has been observed is less than or equal to the marginal randomness of X with equality if and only X and Y are independent (Hickey, 1982). Formally, it reads

$$\phi(X|Y) \leq \phi(X). \tag{6}$$

It is easy to see that this appealing property applies to all measures of the form (5), which is a simple result of the concavity of ϕ by Jensen's inequality. We now apply this result to prove our leaf discriminant criterion.

For ease of notation, let $\mathbf{p}_i(x_{\mathrm{pa}_i}) = p(X_i|x_{\mathrm{pa}_i})$ and $\mathbf{r}_i(x_{-i}) = p(X_i|x_{-i})$ for any node i.

We first prove the " $\Rightarrow$ " direction.

With no loss of generality, we assume that there is only one leaf node.

Suppose l is the leaf node, we have $p(x_l|x_{-l}) = p(x_l|x_{\mathrm{pa}_l})$. Hence

$$\mathbb{E}_{X_{-l}}\Big[\phi(\mathbf{r}_l)\Big] = \mathbb{E}_{X_{-l}}\Big[\phi(\mathbf{p}_l)\Big] = \phi(X_l|X_{\mathrm{pa}_l}). \tag{7}$$

For any non-leaf i, we have

$$\mathbb{E}_{X_{-i}}\Big[\phi(\mathbf{r}_i)\Big] = \phi(X_i|X_{-i}) = \phi(X_i|X_{\mathrm{mb}_i}), \tag{8}$$

where mb_i denotes the Markov blanket of X_i.

Since i is a non-leaf node, $\pi_i < \pi_l$ holds in a causal order π. Applying the properties of conditional information (6) and non-decreasing randomness (4.4) respectively, we have

$$\phi(X_i|X_{\mathrm{mb}_i}) \leq \phi(X_i|X_{\mathrm{pa}_i}) \leq \phi(X_l|X_{\mathrm{pa}_l}). \tag{9}$$

Since mb_i is the minimal set of nodes that renders X_i independent from the other variables and, by Assumption 4.3, no local density is degenerate, the first inequality is therefore strict.

We conclude that $\mathbb{E}_{X_{-l}}\Big[\phi(\mathbf{r}_l)\Big] > \mathbb{E}_{X_{-i}}\Big[\phi(\mathbf{r}_i)\Big], \forall i \neq l$.

We now prove the " $\Leftarrow$ " direction.

Suppose there exists a non-leaf node i such that $i = \arg\max_{i\in[d]} \mathbb{E}_{X_{-i}}\phi(\mathbf{r}_i)$.

It follows that for any leaf node l, we have

$$\mathbb{E}_{X_{-i}}\phi(\mathbf{r}_i) = \phi(X_i|X_{\mathrm{mb}_i}) > \mathbb{E}_{X_{-l}}\phi(\mathbf{r}_l) = \phi(X_l|X_{\mathrm{pa}_l}).$$

Since ϕ satisfies the non-decreasing randomness property, we have $\phi(X_l|X_{\mathrm{pa}_l}) \geq \phi(X_i|X_{\mathrm{pa}_i})$ since $\pi_i < \pi_l$.

This leads to $\phi(X_i|X_{\mathrm{mb}_i}) > \phi(X_i|X_{\mathrm{pa}_i})$, which contradicts the conditional information inequality (6).

Therefore, we must have that $\phi(X_i|X_{\mathrm{mb}_i}) \leq \phi(X_l|X_{\mathrm{pa}_l})$ and the equality occurs when $\phi(X_i|X_{\mathrm{mb}_i}) = \phi(X_i|X_{\mathrm{pa}_i})$. This happens if and only if X_i is independent from other nodes given its parents. Then, i must be a leaf node. □

Let $\phi_H(\mathbf{p}) := H(\mathbf{p}) = \sum_n^{k=1} -p_k \log p_k$ be the entropy function and $\phi_U(\mathbf{p}) = \sum_n^{k=1} \log p_k$ be the sum of log probabilities. Let us denote $\phi_{\mathrm{KL}}(\mathbf{p}) = -\mathrm{KL}(\mathbf{p}\|\mathbf{u}) - \mathrm{KL}(\mathbf{u}\|\mathbf{p})$, where $\mathbf{u}$ denotes the uniform distribution of appropriate dimension.

Proposition 4.7 *For ϕ_H and ϕ_U defined above, if both measures satisfy the non-decreasing randomness property w.r.t the true graph $\mathbf{G}$, then the leaf variable X_l is identifiable from $\phi_{\mathrm{KL}}(\cdot)$ as defined above.*

Proof. We have the following derivations:

$$\begin{aligned}
\mathrm{KL}(\mathbf{p}\|\mathbf{u}) &= \log n - H(\mathbf{p}) \\
\mathrm{KL}(\mathbf{u}\|\mathbf{p}) &= -\log n - \frac{1}{n}\sum_{k=1}^{n}\log p_k \\
\Rightarrow \phi_{\mathrm{KL}}(\mathbf{p}) &= -\mathrm{KL}(\mathbf{p}\|\mathbf{u}) - \mathrm{KL}(\mathbf{u}\|\mathbf{p}) \\
&= H(\mathbf{p}) + \frac{1}{n}\sum_{k=1}^{n}\log p_k = \phi_H(\mathbf{p}) + \frac{1}{n}\phi_U(\mathbf{p}).
\end{aligned}$$

It suffices to show that ϕ_{KL} is a valid randomness measure as defined in (4.2). This is true since ϕ_{KL} is the sum of two Schur-concave functions and a constant, thereby preserving Schur-concavity and satisfying the conditional information property. That ϕ_H and ϕ_U satisfy non-decreasing randomness implies ϕ_{KL} also fulfils this property.

Applying Theorem 4.5, one can easily show that if X_l is a leaf variable , then $l = \arg\max_{i\in[d]} \mathbb{E}_{X_{-i}}\left[\phi_{\mathrm{KL}}\left(\mathbf{r}_i(x_{-i})\right)\right]$, rendering X_l being identifiable from ϕ_{KL}.

□

B. Score Matching

We begin by presenting a background of score matching. Consider an energy-based model over random vector $x \in \mathbb{R}^d$ written as a Gibbs distribution as follows:

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta},$$

where $E_\theta(x) \geq 0$ is the energy function with parameters θ and $Z_\theta = \int \exp(-E_\theta(x))\mathrm{d}x$ is the partition function

If two continuously differentiable real-valued functions $f(x)$ and $g(x)$ have equal first derivatives everywhere i.e., $\nabla_x f(x) = \nabla_x g(x)$, and they are log probability density functions with normalization requirement $\int \exp(f_\theta(x))\mathrm{d}x = \exp(f_\theta(x))\mathrm{d}x = 1$, then $f(x) \equiv g(x)$.

The first-order gradient function of the log-density function is called the **score function** of that distribution. The above property suggests we can learn the model θ by matching its score function with the score of the data distribution. The **score matching** objective minimizes the **Fisher divergence** between two distributions

$$D_F\left[p_{\mathrm{data}}(x)\|p_\theta(x)\right] = \mathbb{E}_{p_{\mathrm{data}}(x)}\left[\frac{1}{2}\|\nabla_x \log p_{\mathrm{data}}(x) - \nabla_x \log p_\theta(x)\|_2^2\right]. \tag{10}$$

For the second term, we can parametrize a neural network $\mathbf{s}_\theta(x) \triangleq \nabla_x \log p_\theta(x)$ to approximate the score function. This can helps us ignore the intractable normalizing constant Z_θ. However, the first term $\nabla_x \log p_{\mathrm{data}}(x)$ is intractable since it requires the knowledge of the data density.

B.1. Basic Score Matching

Under certain regularity conditions, Hyvärinen & Dayan (2005) establishes an objective that avoids computing $\nabla_x \log p_{\mathrm{data}}(x)$. With integration by parts, the Fisher divergence can be rewritten as

$$\begin{aligned}
D_F\left[p_{\mathrm{data}}(x)\|p_\theta(x)\right] &= \mathbb{E}_{p_{\mathrm{data}}(x)}\left[\frac{1}{2}\sum_{i=1}^{d}\left(\frac{\partial E_\theta(x)}{\partial x_i}\right)^2 + \frac{\partial^2 E_\theta(x)}{(\partial x_i)^2}\right] + \mathrm{const}, \\
&= \mathbb{E}_{p_{\mathrm{data}}(x)}\left[\frac{1}{2}\|\mathbf{s}_\theta(x)\|^2 + \mathrm{Tr}\left(\mathbf{J}_x \mathbf{s}_\theta(x)\right)\right] + \mathrm{const},
\end{aligned} \tag{11}$$

where $\mathbf{J}_x \mathbf{s}_\theta(x)$ is the Jacobian of the score function. The estimator under objective (11) is consistent. However, it takes $O(d^2)$ time to compute the trace of the Jacobian.

B.2. Denoising Score Matching

Vincent (2011) proposes a denoising score matching objective that can completely avoid both the unknown term $p_{\text{data}}(x)$ and computationally expensive second-order derivatives. This is done by adding a bit of noise to each data point: $\tilde{x} = x + \epsilon$ where the noise distribution $p(\epsilon)$ is smooth. Let $p(\tilde{x}) = \int p(\tilde{x}|x) p_{\text{data}}(x) \mathrm{d}x$ denote the noisy data distribution.

$$
\begin{aligned}
D_F\Big[p(\tilde{x}) \| p_\theta(\tilde{x})\Big] &= \mathbb{E}_{p(\tilde{x})}\left[\frac{1}{2}\Big\|\nabla_x \log p(\tilde{x}) - \nabla_x \log p_\theta(\tilde{x})\Big\|_2^2\right] \\
&= \mathbb{E}_{p(x,\tilde{x})}\left[\frac{1}{2}\Big\|\nabla_x \log p(\tilde{x}|x) - \nabla_x \log p_\theta(\tilde{x})\Big\|_2^2\right] + \text{const.}
\end{aligned} \tag{12}
$$

Vincent (2011) proves that minimizing (12) is equivalent to minimizing the explicit score matching objective (10). Denoising score matching however is not a consistent objective. The inconsistency becomes non-negligible when $q(\tilde{x})$ significantly differs from $p_{\text{data}}(x)$. Furthermore, if we use small noise perturbation, this often significantly increase the variance of objective (Murphy, 2023).

B.3. Multi-scale Denoising Score Matching

Another issue is that score matching can have difficulty in recovering the true distribution when there are regions of low data density that are highly disconnected. Song & Ermon (2019) proposes to overcome the difficulties by perturbing the data with different scales of noise. Consider a sequence of positive noise scales $\alpha_{\text{min}} = \alpha^1 < \alpha^2 < \cdots < \alpha^T = \alpha_{\text{max}}$, for each data point $x \sim p_{\text{data}}(x)$, a discrete Markov chain $\{x^0 = x, x^1, \cdots, x^t\}$ is constructed such that $p_{\alpha^t}(x^t|x) = \mathcal{N}(x^t|x, \alpha^t \mathbf{I})$ and the marginal distribution is given by $p_{\alpha^t}(x^t) = \int p_{\alpha^t}(x^t|x) p_{\text{data}}(x) dx$.

The noise scales are prescribed such that α_{min} is small enough for $p_{\alpha_{\text{min}}} \approx p_{\text{data}}(x)$ and α_{max} is large enough for x^T to be approximately distributed according to $\mathcal{N}(0, \mathbf{I})$. Then we seek to minimize the expected of Fisher divergences between $p_{\alpha^t}(x^t)$ and $p_\theta(x^t)$ as follows:

$$
\int_0^T \alpha^t \, \mathbb{E}_{p_{\alpha^t}(x,x^t)}\left[\frac{1}{2}\Big\|\nabla_{x^t} \log p_{\alpha^t}(x^t|x) - \nabla_{x^t} \log p_\theta(x^t)\Big\|_2^2\right] dt. \tag{13}
$$

As in basic score matching, we can model $\nabla_{x^t} \log p_\theta(x^t)$ with a time-dependent neural network $\mathbf{s}_\theta(x^t, t)$.

B.4. Continuous-Time Discrete Score Matching

Consider a finite discrete state space $\mathcal{X}$, Sun et al. (2022) model a continuous-time Markov chain forward process $\{X^t\}_{t\in[0,T]}$ with the transition probability characterized by rate matrices $Q^t \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$. If the forward process starts at the target distribution $q^0 = p_{\text{data}}(x)$, the marginal time t is given by $q^t(x^t) = \int q^t(x^t|x) p_{\text{data}}(x) dx$.

To estimate the discrete score function (4) is essentially to learn, from observed i.i.d samples, the set of the singleton conditional distributions $\big\{p(X_i|x_{-i})\big\}_{i\in[d]} := \big\{q^0(X_i|x_{-i})\big\}_{i\in[d]}$.

In a similar setup with the multi-scale denoising framework, we match the $q^t(X_i|x_{-i})$ with a time-dependent neural network $p_\theta^t(X_i|x_{-i})$ by minimizing the weighted sum cross entropy along the forward process as

$$
\int_0^T \mathbb{E}_{q^t(x_t)}\left[\sum_{i=1}^d \left(-\sum_{x_i} q_t(X_i^t = x_i|x_{-i}^t) \log p_\theta^t(X_i^t = x_i|x_{-i}^t)\right)\right] dt. \tag{14}
$$

One can notice the bottleneck lies in the intractable term $q_t(X_i^t = x_i|x_{-i}^t)$. Fortunately, using the factorization property of conditional distribution, we can simplify the above objective as

$$\int_0^T \mathbb{E}_{q^t(x_t)} \left[\sum_{i=1}^{d} \left(- \sum_{x_i} \log p_\theta^t(X_i^t = x_i | x_{-i}^t) \right) \right] dt. \tag{15}$$

See Appendix B.4 (Sun et al., 2022) for the full derivation. We minimize objective (15) to train the discrete score models of interest. We follow the authors' suggested sub-rate transition matrix $Q_i^t = Q\beta(t)$ where $Q = \mathbf{1}\mathbf{1}^T - n_i I$ is the uniform base rate and $\beta(t)$ is the time schedule function.

C. Additional Experiments

We present additional experiments of graphs of $4d$ degree. We encounter a memory explosion issue in the data generation process as SF graphs tend to be concentrated on high-degree nodes. The experiments on SF graphs are conducted up to 20 nodes due to our memory constraints.

Table 5. Synthetic experiment for ER graphs of $4d$ degree.

d	5			10			15			20		
ER4	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}
SCORE + PC (Ours)	0.20	**0.98**	0.50	9.80	**0.71**	3.00	13.60	**0.72**	1.00	21.20	**0.67**	5.50
PC	0.00	0.67	-	8.60	0.52	-	15.00	0.49	-	20.20	0.50	-
SCORE + GAM (Ours)	2.80	**0.82**	0.50	8.60	**0.68**	3.00	13.75	**0.69**	1.00	19.20	**0.65**	5.50
GAM	2.60	0.60	-	6.60	0.57	-	15.60	0.43	-	15.80	0.54	-
OCD	5.80	0.50	-	15.60	0.28	-	20.40	0.43	-	29.00	0.38	-

Table 6. Synthetic experiment for ER graphs of $4d$ degree.

d	30			40			50			60		
ER4	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}
SCORE + PC (Ours)	35.80	**0.58**	14.20	42.60	**0.63**	10.50	53.00	**0.59**	13.50	69.40	**0.60**	18.00
PC	28.00	0.52	-	43.00	0.49	-	50.20	0.49	-	66.80	0.49	-
SCORE + GAM (Ours)	35.20	**0.56**	14.20	44.20	**0.60**	10.50	53.40	**0.57**	13.50	69.40	**0.59**	18.00
GAM	27.40	0.53	-	38.00	0.52	-	42.00	0.53	-	55.60	0.53	-
OCD	47.00	0.34	-	61.20	0.36	-	69.00	0.35	-	70.00	0.40	-

Table 7. Synthetic experiment for SF graphs of $4d$ degrees.

d	5			10			15			20		
SF4	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}	SHD	F1	D_{top}
SCORE + PC (Ours)	2.60	**0.81**	0.40	12.20	**0.51**	1.60	21.60	**0.46**	2.00	30.80	**0.44**	5.40
PC	2.40	0.54	-	10.80	0.45	-	24.00	0.34	-	27.20	0.41	-
SCORE + GAM (Ours)	1.80	**0.83**	0.40	12.40	**0.40**	1.60	19.00	**0.44**	2.00	29.00	**0.35**	5.40
GAM	1.40	0.61	-	11.40	0.39	-	19.00	0.35	-	27.00	0.34	-
OCD	4.80	0.40	-	13.80	0.29	-	23.40	0.22	-	32.20	0.22	-