Causal Discovery via Monotone Triangular Transport Maps

Sina Akbari
EPFL, Switzerland
sina.akbari@epfl.ch

Luca Ganassali
EPFL, Switzerland
luca.ganassali@epfl.ch

Negar Kiyavash
EPFL, Switzerland
negar.kiyavash@epfl.ch

Abstract

We study the problem of causal structure learning from data using transport maps. Specifically, we first provide a constraint-based method which builds upon lower-triangular monotone parametric transport maps to design conditional independence tests which are agnostic to the noise distribution. We provide an algorithm for causal discovery up to Markov Equivalence for general structural equations and noise distributions, which allows for settings with latent variables. Our approach also extends to score-based causal discovery by providing a novel means for defining scores. This allows us to uniquely recover the causal graph under additional identifiability and structural assumptions, such as additive noise or post-nonlinear models. We provide experimental results to compare the proposed approach with the state of the art on both synthetic and real-world datasets.

1 Introduction

Recovering the causal structure between the variables of a system from observational data is a coveted goal in several disciplines of science. The importance of this task has become increasingly evident in the realm of artificial intelligence over the past few decades. This is mainly because a clear understanding of the causal structure in data can greatly enhance predictions of variables under external manipulations, and eliminate systematic biases in inference.

The existing approaches for recovering the causal mechanisms can be largely categorized into *score-based* and *constraint-based* methods. Most of the existing score-based methods impose constraints on either the functional assignment model, or the data distribution. For instance, they may limit the problem to linear models [33, 32, 45], or models with additive noise [12, 20, 25, 18], or restrict the data distribution to a limited class, e.g. Gaussian or discrete [14, 16, 27, 18]. These methods can be sensitive to the choice of model assumptions, and may fail to recover the correct causal model if the relationships between variables are complex, or latent variables exist. However, there is abundant evidence that information about the sparsity of the underlying causal graph improves the estimation efficiency of these methods [42, 30, 34, 9, 11]. An established approach for uncovering the sparsity pattern of the graph involves conducting conditional independence tests, as commonly employed in constraint-based methods, such as PC [36]. Unfortunately, conditional independence (CI) testing is only well understood – theoretically speaking – for either Gaussian or discrete data distributions, and proves to be inefficient in practice outside of this scope. Despite recent progress in the study of kernel-based CI tests [10, 6, 43, 8], conducting CI tests for more general data generating processes remains a daunting task. This presents a significant challenge, since non-Gaussian continuous data is prevalent in various natural phenomena.

In this study, we employ the optimal transport (OT) framework to characterize arbitrary (possibly non-Gaussian) continuous distributions. Through this approach, we offer several advantages compared to existing approaches in the literature. To begin with, this method provides a means to conduct conditional independence tests on continuous data with non-Gaussian distribution. This can hence be

Accepted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023.

used as a building block for constraint-based causal discovery algorithms. Further, it allows for a straightforward way to define and determine scores in the context of score-based causal discovery by characterizing the joint distribution of variables. We will elaborate on this point further in Section 4.

Related work To the best of our knowledge, the work presented in [38] is the only work to date that has explored the application of optimal transport (OT) in the context of learning causal structure from data. In this work the authors considered a two-dimensional additive noise model of the form

$$(X_1, X_2) \coloneqq (U_1, f(X_1) + U_2), \tag{1}$$

where U_1, U_2 are independent exogenous noise variables, and sought to distinguish cause from effect among the two variables X_1 and X_2. The main idea in their approach comes from viewing the distribution ν of $\mathbf{X} = (X_1, X_2)$ as the pushforward measure of the distribution μ of noises $\mathbf{U} = (U_1, U_2)$. The authors made the crucial assumption that the solution T_{ot} to the following standard optimal transport problem with L^2 cost coincides with the structural model of Eq. (1):

$$T_{\mathrm{ot}} := \underset{T:\mathbb{R}^2\to\mathbb{R}^2, T_1=\mathrm{id}, T_\sharp\mu=\nu}{\arg\min} \mathbb{E}_\mu[\|\mathbf{U} - T(\mathbf{U})\|^2], \tag{2}$$

Note that in (2), the authors impose the first coordinate of the transport map to be identity. A simple criterion for T_{ot} to correspond to a cause-effect additive noise model is given by $\mathrm{div}(T_{\mathrm{ot}} - \mathrm{id}) = 0$, where div is the divergence. In practice, the authors rely on a conditional variance test of the form $\mathrm{Var}((T_{\mathrm{ot}})_2(\mathbf{U}) - U_2 \,|\, X_1) = 0$, where $(T_{\mathrm{ot}})_2$ is the second coordinate of the map T_{ot}, to test causal direction. While [38] paves the way for applications of OT in causal discovery, it remains unclear how their method can be generalized to multivariate models; for instance, in higher dimensions (≥ 3), the proposed divergence-based criterion turns out to be only necessary, but far from sufficient. Moreover, the OT problem in Eq. (2) requires knowledge of the distribution μ which is often unavailable. It is unclear how robust the approach of [38] is to noise distribution misspecification in higher dimensions. It is not clear either how it extends to models that violate the additive noise assumption.

Other works have explored the use of optimal transport framework to extract certain relations among variables from the joint distribution. Most relevant to our study is the work [35] which highlighted the presence of a specific type of pairwise conditional independence relations, i.e., the independence of two variables given all the rest of variables, in the Hessian information of the log density of the joint distribution. In line with this observation, [19] emphasized that these independence relations can be leveraged to recover the independence map of a Markov random field.

Contributions Our contributions can be summarized as follows:

(i) We propose a novel causal discovery method based on optimal transport (OT), designed to be agnostic to the noise distribution. The foundation of this method draws inspiration from the work by Morrison et. al. [19], which originally focused on structure learning in Markov random fields. They utilized a parametric OT framework to infer the structure by constructing a lower triangular monotone map, denoted as S, between an unknown data distribution and a reference distribution, typically a standard isotropic Gaussian. We extend this method to causal discovery domain by incorporating additional conditional independence tests. It is noteworthy that our method does not rely on any assumptions regarding the structural or noise properties, and it produces the underlying causal graph up to Markov equivalence. Moreover, our approach is applicable in the presence of latent variables.

(ii) Under additional structural assumptions, e.g. additive noise or post-nonlinear models, we introduce methods to uniquely recover the causal graph based on the same construction as in (i) and a notion of score based on structural assumptions and the shape of the transport map. We demonstrate the dual purpose of the OT-based framework: first, it facilitates the recovery of the sparsity map, and thereby enhances efficiency. Second, it offers a unified framework for evaluating scores.

(iii) We provide novel characterizations of additive noise and post-nonlinear models, generalizing the divergence criterion of [38] to higher dimensions. We show that our criteria are necessary and sufficient for assessing whether data is generated from a model belonging to these two classes of SEMs. Our OT framework offers an effective way of determining the validity of these criteria.

Paper organization We define the problem in Section 2.1 and review the definitions of structural equation models and identifiability. We give some background on transport maps in Section 2.2, introducing the lower-triangular monotone maps, also called Knothe-Rosenblatt maps, and argue that

these maps are particularly well-adapted for causal discovery, as illustrated by Theorem 1 in Section 2.3. Section 3 is dedicated to the description of our OT-based causal discovery method: we first discuss the parameterization of the learned maps, and how to extract conditional independencies from these maps. We then describe our algorithm, which we present as a variation of the PC algorithm [36]. Section 4 describes our score-based approaches under further additive noise model or post non-linear model assumptions. Numerical experiments are presented in Section 5.

2 Preliminaries

2.1 Problem setup

A directed acyclic graph (DAG) is defined as $\mathcal{G} = (\mathbf{X}, E)$, where $\mathbf{X} = \{X_1, \dots, X_d\}$, and $E \subseteq \mathbf{X} \times \mathbf{X}$ denote the set of vertices and edges of this graph, respectively, such that $\mathcal{G}$ cointains no directed cycle. Each vertex $X_k \in \mathbf{X}$ represents a random variable. For each vertex X_k, $\mathrm{Pa}(X_k)$ denotes the set of it parents in $\mathcal{G}$. We say two DAGs are Markov equivalent if they share the same d-separation relations [22]. Throughout this work, we assume that the random variables $\{X_1, \dots, X_d\}$ are governed by a *structural equation model* (SEM) [23],

$$\forall\, 1 \leq k \leq d, \quad X_k \coloneqq f_k((X_\ell)_{X_\ell \in \mathrm{Pa}(X_k)}, U_k), \tag{3}$$

where $(U_k)_{1 \leq k \leq d}$ are mutually independent noise variables. Let $\pi_{\mathbf{X}}$ denote the probability distribution over $\mathbf{X}$ induced by the SEM defined in Eq. (3). $\pi_{\mathbf{X}}$ is commonly referred to as the *observational distribution* in the literature. We drop subscript $\mathbf{X}$ whenever it is clear from context.

Causal discovery refers to the task of learning the causal graph $\mathcal{G}$ from i.i.d. samples drawn from observational distribution π. Two causal DAGs within the same Markov equivalence class are not distinguishable from merely the observational data. In other words, the causal graph is identifiable up to Markov equivalence class using the observational data [23]. However, the causal graph may become uniquely identifiable under further assumptions.

Let $\mathcal{C}$ be a class of SEMs, that is, a subclass of structural models of the form (3).

Definition 1 (Identifiability within a class of SEMs)**.** *Given a distribution π, we say causal graph $\mathcal{G}$ is* identifiable *within the class $\mathcal{C}$ if every SEM $\mathcal{S} \in \mathcal{C}$ that induces distribution π yields the causal structure $\mathcal{G}$.*

In other words, the causal DAG is identifiable in class $\mathcal{C}$ if there is a *unique* DAG $\mathcal{G}$ compatible with (i) the data distribution and (ii) a SEM in class $\mathcal{C}$. Examples of such classes $\mathcal{C}$ and some of these assumptions required for identifiability will be given in Section 4.

Definition 2 ($\mathcal{G}$-compatible orderings)**.** *Given a causal DAG $\mathcal{G}$, we say that a permutation σ of $\{1, \dots, d\}$ is a* $\mathcal{G}$-compatible (causal) ordering *if for all k, ℓ,*

$$X_\ell \in \mathrm{Pa}(X_k) \implies \sigma(\ell) < \sigma(k)\,.$$

2.2 Background on transport maps

We assume all probability distributions are absolutely continuous with respect to Lebesgue measure. With a slight abuse of notation, we use the same symbol to denote a distribution and its density.

A *transport map* S between two distributions μ and ν in $\mathbb{R}^d$ is a map $S : \mathbb{R}^d \to \mathbb{R}^d$ such that the pushforward of μ by S is ν. In other terms $S(X) \sim \nu$ when $X \sim \mu$. Assuming S is invertible, we denote by $S_\# \mu$ the *pushforward* of μ by the map S, and by $S^\# \nu$ the *pullback* of ν by S. These are easily obtained by a multi-dimensional change of variables as follows:

$$S_\# \mu(\mathbf{y}) = \mu \circ S^{-1}(\mathbf{y})\, |\det(\nabla S^{-1}(\mathbf{y}))|, \quad S^\# \nu(\mathbf{x}) = \nu \circ S(\mathbf{x})\, |\det(\nabla S(\mathbf{x}))|. \tag{4}$$

In general, there are many such transport maps S. A special class of transport maps, well suited to the problem of recovering a causal graph (see Section 2.3), or a causal ordering, is the lower-triangular maps. In particular, when μ and ν have positive densities with respect to Lebesgue measure, there exists a unique lower-triangular map S of the following form

$$S(\mathbf{x}) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ \vdots \\ S_d(x_1, \dots, x_d) \end{bmatrix}, \tag{5}$$

which satisfies the measure transformations of Eq. (4) and for each component k, S_k is strictly increasing in the last variable [15, 28, 2, 4]. This map is sometimes referred to as the Knothe-Rosenblatt (KR) map [15, 28]. Note that each component S_k only depends on $x_1, \dots, x_k$ and the strict monotonocity of S_k in x_k implies that this map is invertible[1]. We refer the interested reader to [4], in which Knothe-Rosenblatt maps have been studied thoroughly. In particular, KR maps are shown to be characterized as the limit of solutions of the standard optimal transport problem in the regime where the quadratic cost becomes degenerate (see Carlier et al. [4], Theorem 2.1).

2.3 Knothe-Rosenblatt maps for causal discovery

As we will see in Section 3.1, KR maps are used to provide good estimates of the joint distribution π. There are several non-parametric methods to estimate joint densities, the most popular of which being multivariate Kernel density estimation, and KNN density estimation [31]. Rather than relying on these, we propose a method based on KR maps because these maps precisely reveal the causal structure in a SEM. The Theorem below, proof of which is given in Appendix C, brings some mathematical evidence of the previous statement and shows that KR maps are particularly well-suited for causal discovery tasks.

Theorem 1. *Suppose random variables $\{X_1, \dots, X_d\}$ are governed by the SEM given in* (3)*, Moreover, assume that for all $1 \leq k \leq d$, map f_k is strictly increasing in the last variable, and that the cumulative distribution function (c.d.f) of U_k, denoted by F_{U_k}, is strictly increasing.*

(i) For any $\mathcal{G}$-compatible ordering σ, KR map $S(\sigma)$ between the distribution of $(U_{\sigma(1)}, \dots, U_{\sigma(d)})$ and that of $(X_{\sigma(1)}, \dots, X_{\sigma(d)})$ coincides with the SEM equations (3)*, that is, for all $1 \leq k \leq d$,*

$$S(\sigma)_k(u_{\sigma(1)}, \dots, u_{\sigma(k-1)}, u_{\sigma(k)}) = f_{\sigma(k)}\Big((S(\sigma)_\ell(u_{\sigma(1)}, \dots, u_{\sigma(\ell)}))_{\ell : X_{\sigma(\ell)} \in \mathrm{Pa}(X_{\sigma(k)})}, u_{\sigma(k)}\Big). \quad (6)$$

Further, KR maps corresponding to any $\mathcal{G}$-compatible ordering are the same up to a permutation[2]*.*

(ii) If the causal mechanism is identifiable within a class $\mathcal{C}$ of SEMs (Def. 1), then σ is a $\mathcal{G}$-compatible ordering if and only if the KR map $S(\sigma)$ provides a SEM in class $\mathcal{C}$.

In other words, given a $\mathcal{G}$-compatible ordering, KR map recovers the true underlying SEM. Moreover, if the structure is uniquely identifiable (for instance in the case of additive noise models), computing KR maps allows us to identify a $\mathcal{G}$-compatible causal ordering, and hence $\mathcal{G}$ itself. Theorem 1 gives an identifiablity result through the use of KR maps, highlighting the fact that they are exceptionally suitable for revealing the causal relations among the variables. However, it is noteworthy that our transport-based causal discovery approach, which will be introduced in the sequel, does not rely on this result. Indeed, while Theorem 1 accomplishes a more extensive objective, namely, recovering the entire SEM, it necessitates stronger assumptions than those needed for our primary focus in this paper: recovering the mere causal structure (the essential graph).

Our approach for OT-based causal discovery comprises two steps. The first step, discussed in Section 3, recovers the causal graph up to Markov equivalence, without requiring any model or structural assumptions. The second step, detailed in Section 4, aims at learning a $\mathcal{G}$-compatible ordering, which identifies the causal graph, under additional assumptions, e.g. restricting the model to the ANM class.

3 Recovering the essential graph via monotone triangular transport maps

3.1 A parametrization of the transport maps

Henceforth, we consider the KR map from π (or some of its marginals) to a known, smooth, log-concave source distribution η with the same dimension. Throughout this work, η will be taken to be the multivariate isotropic normal distribution. The idea behind this approach is that if the KR map S from the data distribution π to η can be estimated efficiently with finitely many samples, then π can be simply represented as the pullback of η by S, namely $\pi = S^{\#}\eta$. In practice, we parameterize the KR map S with a vector of parameters $\boldsymbol{\alpha}$. The parameterized transport map $S_{\boldsymbol{\alpha}}$ is estimated by

[1] Note that by this definition, the transport map defined in [38] as the solution of problem (2) is exactly the KR map between the distributions of $\mathbf{U} = (U_1, U_2)$ and $\mathbf{X} = (X_1, X_2)$.

[2] Note that these permutations necessarily preserve the lower-triangular structure.

optimizing over the set of parameters $\boldsymbol{\alpha}$ such that the Kullback-Leibler divergence between π and the pullback of the source distribution η by the map $S_{\boldsymbol{\alpha}}$ is minimized:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} D_{\mathsf{KL}}(\pi \| S_{\boldsymbol{\alpha}}^{\#}\eta) = \arg\min_{\boldsymbol{\alpha}} \mathbb{E}_{\pi}[\log \pi - \log S_{\boldsymbol{\alpha}}^{\#}\eta] \approx \arg\max_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^{n} \log\left(S_{\boldsymbol{\alpha}}^{\#}\eta(\mathbf{x}^i)\right), \quad (7)$$

where $\{\mathbf{x}^i\}_{1\leq i\leq n}$ are the i.i.d. samples of data. Following [19], an efficient way to parameterize $S_{\boldsymbol{\alpha}}$ in order to enforce both the lower-triangular shape and the monotonicity assumption is the following:

$$(S_{\boldsymbol{\alpha}})_k(x_1,\ldots,x_k) = c_{k,\boldsymbol{\alpha}}(x_1,\ldots,x_{k-1}) + \int_0^{x_k} g \circ h_{k,\boldsymbol{\alpha}}(x_1,\ldots,x_{k-1},t)\mathrm{d}t, \quad (8)$$

where $g : \mathbb{R} \to \mathbb{R}$ is a positive map, and $c_{k,\boldsymbol{\alpha}} : \mathbb{R}^{k-1} \to \mathbb{R}$ (resp. $h_{k,\boldsymbol{\alpha}} : \mathbb{R}^k \to \mathbb{R}$) is a linear combination of multivariate Hermite polynomials $\{\phi_s\}_{s\geq 0}$ (resp. Hermite functions $\{\psi_s\}_{s\geq 0}$).

Note that map $S_{\boldsymbol{\alpha}}$ defined in (8) has the desired lower-triangular shape and each $(S_{\boldsymbol{\alpha}})_k$ is strictly increasing in the last variable. The well-known fact that Hermite functions form a Hilbert basis of $L^2(\mathbb{R})$ justifies this parametrization [19, 35]. The expressiveness of the model depends on the maximal degree of the Hermite polynomials/functions in $(c_{k,\boldsymbol{\alpha}}, h_{k,\boldsymbol{\alpha}})_{1\leq k\leq d}$. Moreover, since η is log-concave and the parametrization of $S_{\boldsymbol{\alpha}}$ is linear in α, the optimization problem (7) is convex. In some recent work was proposed to learn maps $c_{k,\boldsymbol{\alpha}}$ and $h_{k,\boldsymbol{\alpha}}$ with neural networks [21].

In this work, the positive map g in (8) is always fixed as the square function. This allows us to compute all the integrals in maps $S_{\boldsymbol{\alpha}}$ easily and in closed form (as they correspond to joint moments of Gaussian variables), as well as all the partial derivatives of $S_{\boldsymbol{\alpha}}$ – which we will need in the sequel.

3.2 Capturing conditional independencies in marginal densities

We denote the independence of X_ℓ and X_k conditioned on $\mathbf{Z}$ by $X_\ell \perp\!\!\!\perp X_k | \mathbf{Z}$. Recall that π denotes the distribution (or, density) of the data $\mathbf{X} = \{X_1, \ldots, X_d\}$. In order to explain how the conditional independencies can be read directly off the marginals of π, we need the following assumption on π.

Assumption 1. *Density π is positive and its second-order partial derivatives are defined everywhere.*

Lemma 2 of [35] establishes a characterization of conditional independence in terms of Hessian information of the density π. Herein, we adapt their lemma for our purpose.

Lemma 3.1. *[Adapted from [35]] Suppose π satisfies Assumption 1. Let $\pi_{\mathbf{Z}}$ denote the marginal density over $\mathbf{Z} \subseteq \mathbf{X}$. For any two variables $X_k, X_\ell \in \mathbf{Z}$, the following equivalence holds:*

$$X_k \perp\!\!\!\perp X_\ell \mid \mathbf{Z} \setminus \{X_k, X_\ell\} \iff \frac{\partial^2 \log \pi_{\mathbf{Z}}}{\partial x_k \partial x_\ell} = 0 \text{ on } \mathbb{R}^{|\mathbf{Z}|}.$$

Proof of Lemma 3.1 appears in Appendix C. For a fixed subset $\mathbf{Z} \subseteq \mathbf{X}$, let $S_{\mathbf{Z}}$ be a transport map that pushes the marginal density $\pi_{\mathbf{Z}}$ forward to a multivariate isotropic Gaussian η. In light of Lemma 3.1, the conditional independence relations can be determined by assessing the partial derivatives of $S_{\mathbf{Z}}^{\#}\eta$ at all points and observing if they are all zero. When the variables are continuous, determining whether these derivatives are zero everywhere is impractical. Instead, we propose the following *conditional independence score* to test the conditional independence $X_k \perp\!\!\!\perp X_\ell \mid \mathbf{Z} \setminus \{X_k, X_\ell\}$:

$$\Omega_{k\ell}^{\mathbf{Z}} := \mathbb{E}_{\pi_{\mathbf{Z}}}\left[\left(\tfrac{\partial^2}{\partial x_k \partial x_\ell} \log \pi_{\mathbf{Z}}(\mathbf{z})\right)^2\right] = \mathbb{E}_{\pi_{\mathbf{Z}}}\left[\left(\tfrac{\partial^2}{\partial x_k \partial x_\ell} \log S_{\mathbf{Z}}^{\#}\eta(\mathbf{z})\right)^2\right] \approx \frac{1}{n}\sum_{i=1}^{n}\left(\tfrac{\partial^2}{\partial x_k \partial x_\ell} S_{\mathbf{Z}}^{\#}\eta(\mathbf{z}^i)\right)^2, \quad (9)$$

where $\{\mathbf{z}^i\}_{i=1}^n$ are observed samples of $\mathbf{Z}$. In practice, finite sample approximations of $\boldsymbol{\alpha}$ and $\Omega_{k\ell}^{\mathbf{Z}}$ could yield small but non-zero entries when the corresponding independence holds. To deal with this issue, we compare the conditional independence score to a properly chosen threshold $\tau_{k\ell}^{\mathbf{Z}}$. The threshold is chosen in proportion to the standard deviation of $\Omega_{k\ell}^{\mathbf{Z}}$, driven by the objective of isolating those entries whose standard deviation renders them indistinguishable from zero [1]. Morrison et al. [19] take a similar thresholding approach, albeit employing the absolute value of the partial derivatives of log density as the independence score rather than the squared form of Eq. (9). The standard deviation of $\Omega_{k\ell}^{\mathbf{Z}}$ is approximated as

$$\varsigma(\Omega_{k\ell}^{\mathbf{Z}}) \approx \frac{1}{n}(\nabla_{\boldsymbol{\alpha}}\Omega_{k\ell}^{\mathbf{Z}})^T \Gamma(\boldsymbol{\alpha})^{-1} (\nabla_{\boldsymbol{\alpha}}\Omega_{k\ell}^{\mathbf{Z}})\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}, \quad (10)$$

where $\Gamma(\boldsymbol{\alpha})$ is the Fisher information matrix [5], and $\nabla_{\boldsymbol{\alpha}}\Omega^{\mathbf{Z}}_{k\ell}$ is the gradient of $\Omega^{\mathbf{Z}}_{k\ell}$ w.r.t. $\boldsymbol{\alpha}$. See [19] for further details of the rationale behind this choice of thresholding and its consistency analysis.

It is crucial to note that this criterion does not require any assumption on the distribution class. SING algorithm [19] employs the criterion above only for the set $\mathbf{Z} \coloneqq \mathbf{X}$ to recover the Markov random field structure. In the context of DAGs, their approach is termed as *total conditioning* (TC) by [24], and is shown to recover the moralized graph of $\mathcal{G}$, where each vertex is adjacent to its Markov boundary [24]. In our method, the maps $S_{\mathbf{Z}}$ for every $\mathbf{Z} \subseteq \mathbf{X}$ will be parameterized in the same fashion as S in (8), except that dimension d will be replaced with a smaller $d' \coloneqq |\mathbf{Z}|$.

3.3 Description of the algorithm

Assuming faithfulness [36], conditional independence relations encoded in the data distribution are equivalent to d-separations in the causal DAG. The conditional independence test developed in the previous section can therefore be employed as a module in constraint-based causal discovery methods. To illustrate this, we present our method PC-OT as a variation of the PC algorithm [36], summarized as Algorithm 1 in Appendix A. Note that even in the presence of latent variables, the joint distribution of the observable variables can still be represented as a pullback measure of a standard Gaussian distribution of the same dimension. Although this map may have nothing to do with the underlying SEM in this case, once a representation $\pi = S^{\#}\eta$ (or[3] $\pi_{\mathbf{Z}} = S^{\#}_{\mathbf{Z}}\eta$) is obtained, Lemma 3.1 applies, i.e. the conditional independence relationships the observable data are revealed, enabling the recovery of the maximally oriented partial ancestral graph (PAG). In the presence of latent variables, OT-based variants of algorithms such as FCI and RFCI [36, 7] can be developed analogous to Alg. 1.

4 Refined OT-based structure learning under structural assumptions

In this section, we shall undertake an analysis of our OT-based causal discovery approach under the consideration of class assumptions that enable the unique identification of the causal DAG. The results stated in this section are applicable to any class within which the causal graph is identifiable (refer to Def. 1) and membership in that class is testable. As two illustrative examples of such classes, we discuss additive noise models (ANMs) and post-nonlinear models (PNLs). While ANMs are discussed below, PNLs are postponed to Appendix D due to space limitations.

4.1 Additive noise models

Additive noise models (ANMs) are defined as follows.

Definition 3 (ANM)**.** *We say that the SEM of Eq.* (3) *forms an additive noise model if,*

$$\forall\, 1 \leq k \leq d: \quad f_k((X_\ell)_{X_\ell \in \mathrm{Pa}(X_k)}, U_k) \coloneqq g_k((X_\ell)_{X_\ell \in \mathrm{Pa}(X_k)}) + U_k, \tag{11}$$

that is, the structural equation pertaining to any variable is additive in the corresponding noise.

Note that as long as the noise variables U_k have a strictly positive density, the ANMs satisfy the conditions of Theorem 1. For any ordering σ, we denote by π_σ the joint distribution of $(X_{\sigma(1)}, \ldots, X_{\sigma(d)})$. If σ is $\mathcal{G}-$compatible, then in view of Theorem 1, the KR map $S(\sigma)$ from π_σ to η is of the form

$$S(\sigma)_k(x_{\sigma(1)}, \ldots, x_{\sigma(k)}) = M_k(\sigma)\left(x_{\sigma(k)} - g_k((x_\ell)_{X_\ell \in \mathrm{Pa}(X_{\sigma(k)})})\right), \tag{12}$$

where $M_k(\sigma)$ is the strictly increasing transport map from the distribution of $U_{\sigma(k)}$ to a standard Gaussian $\mathcal{N}(0,1)$. Note that, up to a one-dimensional monotonous map, the partial derivative of $S(\sigma)_k$ with respect to its last variable is a constant, and specifically equal to 1. This observation constitutes a characterization of ANMs, formalized below.

Lemma 4.1. *Suppose π satisfies Assumption 1. Let σ be an ordering. The following are equivalent.*

- *π is induced by an ANM, and σ is $\mathcal{G}-$compatible.*
- *for all $1 \leq k \leq d$, there exists a strictly increasing map $B_k(\sigma) : \mathbb{R} \to \mathbb{R}$ such that*

$$\tfrac{\partial}{\partial x_{\sigma(k)}} B_k(\sigma) \circ S(\sigma)_k(x_{\sigma(1)}, \ldots, x_{\sigma(k)}) - 1 = 0\,. \tag{13}$$

[3] In general, the marginal distribution $\pi_{\mathbf{Z}}$ for $\mathbf{Z} \subseteq \mathbf{X}$ is not modeled by a SEM with independent noises, and hence the marginal model for nodes in $\mathbf{Z}$ is equivalent to a SEM with latent variables.

(a) for ANM–OT (b) for PC–OT

Figure 1: Underlying causal graphs in the numerical experiments.

$B_k(\sigma)$ in Eq. (13) corresponds to $M_k(\sigma)^{-1}$ in Eq. (12). In practice, we parameterize each map $B_k(\sigma)$ with vector $\boldsymbol{\beta}_k$, and $\boldsymbol{\beta}_k^*$ is estimated by optimizing the natural loss given by Lemma 4.1:

$$\begin{aligned}\boldsymbol{\beta}_k^* &:= \arg\min_{\boldsymbol{\beta}_k} \mathbb{E}_\pi \Big[\Big| \tfrac{\partial}{\partial x_{\sigma(k)}} [B_k(\sigma)]_{\boldsymbol{\beta}_k} \circ (S(\sigma)_{\boldsymbol{\alpha}^*})_k(X_{\sigma(1)}, \ldots, X_{\sigma(k)}) - 1 \Big| \Big] \\ &\approx \arg\min_{\boldsymbol{\beta}_k} [\mathrm{ANMloss}_k(\sigma, \mathbf{x})](\boldsymbol{\beta}_k), \quad \text{where} \end{aligned} \tag{14}$$

$$[\mathrm{ANMloss}_k(\sigma, \mathbf{x})](\boldsymbol{\beta}_k) := \sum_{i=1}^{n} \Big| \tfrac{\partial}{\partial x_{\sigma(k)}} [B_k(\sigma)]_{\boldsymbol{\beta}_k} \circ (S(\sigma)_{\boldsymbol{\alpha}^*})_k(x^i_{\sigma(1)}, \ldots, x^i_{\sigma(k)}) - 1 \Big|, \tag{15}$$

and (we recall) $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are observed samples, and we used $S(\sigma)_{\boldsymbol{\alpha}^*}$ with $\boldsymbol{\alpha}^*$ being the solution of the optimization problem (7), as in Section 3. An efficient way to parameterize $[B_k(\sigma)]_{\boldsymbol{\beta}_k}$ is:

$$[B_k(\sigma)]_{\boldsymbol{\beta}_k}(u) := \int_0^u g \circ b_{k,\boldsymbol{\beta}_k}(t) \mathrm{d}t, \tag{16}$$

where $g : \mathbb{R} \to \mathbb{R}$ is positive (the quadratic function in our case), and $b_{k,\boldsymbol{\beta}_k} : \mathbb{R} \to \mathbb{R}$ is a linear combinations of Hermite functions $\{\psi_s\}_{s \geq 0}$. Note that the parameterization in (16) enforces strict monotonicity of $u \mapsto [B_k(\sigma)]_{\boldsymbol{\beta}_k}(u)$.

Note that given the Markov equivalence class, and a $\mathcal{G}$-compatible ordering, the causal graph is uniquely identified. Under identifiability assumptions for ANMs [12], every $\mathcal{G}$-compatible ordering is consistent with the true underlying causal order. As such, identifying one $\mathcal{G}$-compatible ordering suffices to recover the causal DAG. On account of Lemma 4.1, we devise a method to decide $\mathcal{G}$-compatibility of an ordering under ANM assumption. To this end, we compute the *ANM loss* corresponding to an ordering, introduced subsequently.

ANM loss. The *ANM loss* of an ordering σ, parameterized by $\boldsymbol{\gamma} \in (\mathbb{R}_{>0})^d$ is defined as

$$\mathrm{ANMloss}_{\boldsymbol{\gamma}}(\sigma, \mathbf{x}) := \sum_{k=1}^{d} \gamma_k [\mathrm{ANMloss}_k(\sigma, \mathbf{x})](\boldsymbol{\beta}_k^*), \tag{17}$$

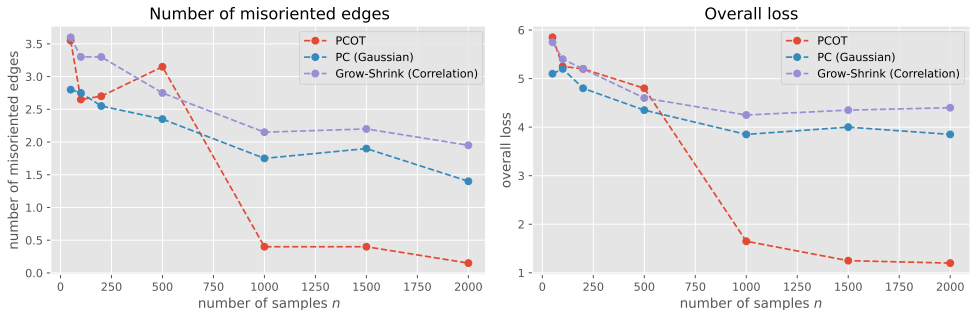
where $\boldsymbol{\alpha}^*$, $\boldsymbol{\beta}_k^*$ and $\mathrm{ANMloss}_k(\sigma, \mathbf{x})$ are defined in (7), (14) and (15). Evidently, an ordering σ is $\mathcal{G}$-compatible iff its ANM loss is zero. In practice, we choose the ordering with the lowest ANM loss.

Possible orderings Given an essential graph $\hat{\mathcal{G}}$, we need to test every possible causal graph $\mathcal{H}$ in the Markov equivalence class $\mathcal{M}(\hat{\mathcal{G}})$ of $\hat{\mathcal{G}}$. For any such graph $\mathcal{H}$, choose an arbitrary $\mathcal{H}-$compatible ordering $\sigma_{\mathcal{H}}$. We define the set of *possible orderings* as $\Sigma(\hat{\mathcal{G}}) := \{\sigma_{\mathcal{H}}, \mathcal{H} \in \mathcal{M}(\hat{\mathcal{G}})\}$. Note that depending on $\hat{\mathcal{G}}$, $\Sigma(\hat{\mathcal{G}})$ is often much smaller than the set of permutations of $\{1, \ldots, d\}$; a trivial upper bound on $\Sigma(\hat{\mathcal{G}})$ is $2^{e(\hat{\mathcal{G}})}$, with $e(\hat{\mathcal{G}})$ denoting the edge count in $\hat{\mathcal{G}}$.

5 Numerical experiments

This section presents the performance of our methods through illustrative examples. Further numerical experiments, plots and details can be found in Appendix E. We utilized TransportMaps package [37] for recovering parametric maps.

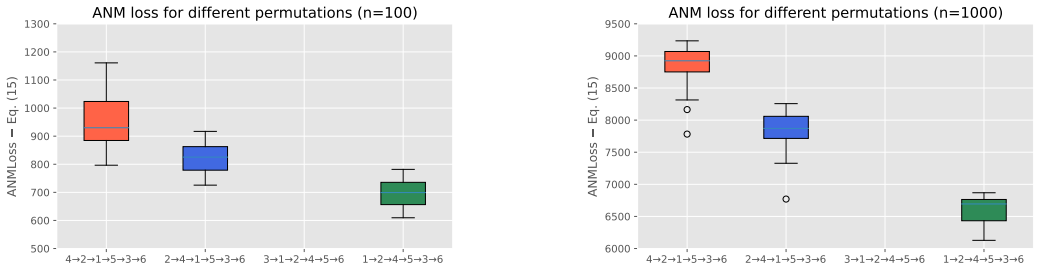
Experiments for PC–OT. We compared the performance of our PC–OT method with both PC and Grow-Shrink (GS) algorithms [36], provided with conditional independence tests which are designed



(a) average number of misoriented edges

(b) overall average loss

Figure 2: comparison of PC–OT with PC and GS algorithms.



(a) ANMloss with $n = 100$ samples available

(b) ANMloss with $n = 1000$ samples available

Figure 3: ANMloss for the four different DAGs within the Markov equivalence class. For both plots, $\gamma_k = 1$ was chosen for every $1 \leq k \leq d$ (see Eq. 17.) The plots are clipped for better visualization: the loss corresponding to ordering $3 \to 1 \to 2 \to 4 \to 5 \to 6$ is not included due to a large gap.

for Gaussian distributions (using CDT package [13]). On the contrary, our method is equipped with the OT-based conditional independence criterion, which is agnostic to the noise distribution.

We worked with synthetic data from a SEM where the exogenous noises are non-Gaussian, details of which can be found in Appendix E. The underlying causal graph is represented in Figure 1b. We see on Figure 2 that PC–OT outperforms these two methods as soon as the number of samples is large enough. This superior performance is illustrated for the number of misoriented edges of the output as well as the overall loss, which is defined as the total number of missing, extra and misoriented edges. Results are averaged over 20 tests. See Appendix E for further details and experiments.

Experiments for ANM-OT. To illustrate the ANM-OT method, we worked with the DAG of Figure 1a. The Markov equivalence class of this graph contains four DAGs. For each of these DAGs, we chose a compatible ordering and compared the ANMloss (17) of each of these orderings. Results shown in Figure 3 show that the ANMloss of the true ordering (the rightmost one) is significantly lower than the other three.

6 Concluding remarks

In this work, we proposed a novel causal discovery method based on transport, designed to be agnostic to the noise distribution. This framework both helps recovering the causal graph up to Markov equivalence, and offers a coherent framework for evaluating scores assessing the validity of structural assumptions such as additive noise or post-nonlinear models.

Bringing the optimal transport framework to the fore as a valuable toolkit for causal discovery, we believe that this study may pave the way to future works of interest to the causality community.

References

[1] Ricardo Baptista, Youssef Marzouk, Rebecca E. Morrison, and Olivier Zahm. Learning non-gaussian graphical models via hessian scores and triangular transport, 2023.

[2] Vladimir Igorevich Bogachev, Aleksandr Viktorovich Kolesnikov, and Kirill Vladimirovich Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, 2005.

[3] Nicolas Bonnotte. From knothe's rearrangement to brenier's optimal transport map. *SIAM Journal on Mathematical Analysis*, 45(1):64–87, 2013. doi: 10.1137/120874850. URL `https://doi.org/10.1137/120874850`.

[4] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From knothe's transport to brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

[5] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2002.

[6] Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *International Conference on Machine Learning*, pages 1422–1430. PMLR, 2014.

[7] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[8] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014.

[9] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[10] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

[11] Stefan Haufe, Klaus-Robert Müller, Guido Nolte, and Nicole Krämer. Sparse causal discovery in multivariate time series. In *causality: objectives and assessment*, pages 97–106. PMLR, 2010.

[12] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

[13] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019.

[14] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Structural agnostic modeling: Adversarial learning of causal graphs. *arXiv preprint arXiv:1803.04929*, 2018.

[15] Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52, 1957.

[16] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

[17] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410, 1995.

[18] Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. *arXiv preprint arXiv:2304.03382*, 2023.

[19] Rebecca Morrison, Ricardo Baptista, and Youssef Marzouk. Beyond normality: Learning sparse probabilistic graphical models in the non-gaussian setting. *Advances in neural information processing systems*, 30, 2017.

[20] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

[21] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

[22] Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

[23] Judea Pearl. *Causality*. Cambridge university press, 2009.

[24] Jean-Philippe Pellet and André Elisseeff. Using markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9(7), 2008.

[25] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models, 2014.

[26] Marc Rieger. Monotonicity of transport plans, 03 2012.

[27] Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

[28] Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

[29] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

[30] Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, Madelyn RK Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.

[31] David Scott and Stephan Sain. Multi-dimensional density estimation. *Data Mining and Data Visualization*, 24, 01 2005.

[32] Anna Seigal, Chandler Squires, and Caroline Uhler. Linear causal disentanglement via interventions. *arXiv preprint arXiv:2211.16467*, 2022.

[33] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[34] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

[35] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

[36] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[37] T.M.Team. Transportmaps package. `http://transportmaps.mit.edu/`, 2015.

[38] Ruibo Tu, Kun Zhang, Hedvig Kjellström, and Cheng Zhang. Optimal transport for causal discovery. In *International Conference on Learning Representations*, 2022.

[39] Kento Uemura and Shohei Shimizu. Estimation of post-nonlinear causal models using autoencoding structure. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3312–3316. IEEE, 2020.

[40] Kento Uemura, Takuya Takagi, Kambayashi Takayuki, Hiroyuki Yoshida, and Shohei Shimizu. A multivariate causal discovery based on post-nonlinear model. In *Conference on Causal Learning and Reasoning*, pages 826–839. PMLR, 2022.

[41] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 647–655, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

[42] Kun Zhang, Heng Peng, Laiwan Chan, and Aapo Hyvärinen. Ica with sparse connections: Revisited. In *Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings 8*, pages 195–202. Springer, 2009.

[43] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[44] Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

[45] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Appendix

This appendix is organized as follows. We present our OT-based version of PC algorithm in Section A. In Section B, we briefly review concepts from KR maps and their relation to Brenier maps for the sake of comprehensiveness and their subsequent utilization in our proofs. We present the proofs of our main results in Section C. Section D includes the derivation of PNLloss based on Lemma C.1 in the text. Section E is devoted to further experimental results, along with details of the experiments included in the main text.

A OT-based PC Algorithm

As discussed in the main text, any constraint-based causal discovery algorithm can be modified to utilize the proposed OT-based conditional independence test. For illustration purposes, we present an OT-based version of PC algorithm [36] in this section. The pseudo-code is provided as Algorithm 1.

High-level description. Like classic PC, the algorithm begins with a complete undirected graph. It keeps track of a counter Δ, increasing it by one at each iteration. At each iteration, subsets $\mathbf{Z}$ of size $\Delta + 2$ are chosen, and SING [19] is called as a subroutine using only the samples corresponding to variables in $\mathbf{Z}$. The output of this subroutine is the matrix $\Omega^{\mathbf{Z}}$, with entries as defined in Eq. (9). As long as the entry $\Omega^{\mathbf{Z}}_{k\ell}$ of the matrix does not exceed the threshold $\tau^{\mathbf{Z}}_{k\ell}$, we conclude that X_k and X_ℓ are conditionally independent with respect to π, which results in removing the corresponding edge from $\hat{\mathcal{G}}$. The corresponding separating set $\mathbf{Z} \setminus \{X_k, X_\ell\}$ is stored for orienting the v-structures at the end of the algorithm. The algorithm iterates as long as the maximum degree of the remaining graph $\hat{\mathcal{G}}$ is at least Δ. Finally, Meek rules [17] are applied to output the essential graph.

Algorithm 1 PC–OT

input: n i.i.d. samples from the observational distribution $\{\mathbf{x}^i\}_{i=1}^n \sim \pi$
output: essential graph corresponding to the causal DAG $\mathcal{G}$

function PCOT($\{\mathbf{x}^i\}_{i=1}^n$)
  $\hat{\mathcal{G}} \leftarrow$ complete undirected graph on $\mathbf{X}$, $\quad \Delta \leftarrow 0$
  for every $k \neq \ell \in \{1, \dots, d\}$ **do** SepSet $(X_k, X_\ell) \leftarrow$ null
  while True **do**
    for every subset $\mathbf{Z} \subseteq \mathbf{X}$ of size $\Delta + 2$ **do**
      $\Omega^{\mathbf{Z}} \leftarrow$ SING($\{\mathbf{x}^i_{\mathbf{Z}}\}_{i=1}^n \sim \pi_{\mathbf{Z}}$)
      for each pair $\{X_k, X_\ell\} \subseteq \mathbf{Z}$ **do**
        if $\Omega^{\mathbf{Z}}_{k\ell} < \tau^{\mathbf{Z}}_{k\ell}$ **then**
          delete the edge between X_k and X_ℓ in $\hat{\mathcal{G}}$
          SepSet $(X_k, X_\ell) \leftarrow \mathbf{Z} \setminus \{X_k, X_\ell\}$
    $\Delta \leftarrow \Delta + 1$
    if $\Delta >$ maxdegree$(\hat{\mathcal{G}})$ **then break**
  for every triplet $k, \ell, m \in \{1, \dots, d\}$ **do**
    if $\exists$ an edge between X_k and X_ℓ, and X_ℓ and X_m, no edge between X_k and X_m **then**
      Orient $X_k \to X_\ell$ and $X_\ell \leftarrow X_m$ in $\hat{\mathcal{G}}$ if and only if $X_\ell \notin$ SepSet (X_k, X_m)
  Apply Meek rules on $\hat{\mathcal{G}}$ [17]
  return $\hat{\mathcal{G}}$

B On Knothe-Rosenblatt transport maps

In this short part we give details on construction of Knothe-Rosenblatt transport maps between two distributions μ and ν in $\mathbb{R}^d$. For a general recap on transport maps, we refer to Section 2.2. In the following we assume that μ and ν have positive densities on $\mathbb{R}^d$, with respect to Lebesgue measure. These assumptions are made for the sake of simplicity, but can be further relaxed [26, 4, 15, 28, 3].

A fundamental building block for KR maps is given in this Lemma, which characterizes the one-dimensional monotone transport maps:

Lemma B.1 (See Carlier et al. [4] and Proposition 2.5 of Rieger [26]). *When μ and ν are one-dimensional ($d = 1$), there exists a unique strictly increasing transport map T from μ to ν, given by $T := F_\nu^{-1} \circ F_\mu$, where F_μ (resp. F_ν) is the cumulative distribution function (c.d.f.) of distribution μ (resp. of ν). This map will be referred to as the* (one-dimensional) Brenier map *between distributions μ and ν.*

We are now ready to describe the construction of KR maps. Recall that these maps are of the following form

$$S(\mathbf{x}) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ \vdots \\ S_d(x_1, \dots, x_d) \end{bmatrix},$$

with S_k is strictly increasing in the last variable for all k.

Let $(X_1, \dots, X_d) \sim \mu$ and $(Y_1, \dots, Y_d) \sim \nu$. The KR map S is built recursively. First, let S_1 be the (unique) Brenier map from the distribution of X_1 to that of Y_1. Then, when $S_1, \dots, S_{k-1}$ are already constructed, we define S_k as follows. For every fixed $x_1, \dots, x_{k-1}$, the map $S_k(x_1, \dots, x_{k-1}, \cdot)$ is defined as the (unique) Brenier map from the distribution of

$$(X_k \,|\, X_{k-1} = x_{k-1}, \dots, X_1 = x_1)$$

to that of

$$(Y_k \,|\, Y_{k-1} = S_{k-1}(x_1, \dots, x_{k-1}), \dots, Y_1 = S_1(x_1)) \,.$$

It can be easily checked that this map S previously defined is (i) transporting μ onto ν and (ii) satisfying the lower-triangular and monotonicity properties.

C Proofs

Theorem 1. *Suppose random variables $\{X_1, \dots, X_d\}$ are governed by the SEM given in* (3)*, Moreover, assume that for all $1 \le k \le d$, map f_k is strictly increasing in the last variable, and that the cumulative distribution function (c.d.f) of U_k, denoted by F_{U_k}, is strictly increasing.*

(i) For any $\mathcal{G}$-compatible ordering σ, KR map $S(\sigma)$ between the distribution of $(U_{\sigma(1)}, \dots, U_{\sigma(d)})$ and that of $(X_{\sigma(1)}, \dots, X_{\sigma(d)})$ coincides with the SEM equations (3)*, that is, for all $1 \le k \le d$,*

$$S(\sigma)_k(u_{\sigma(1)}, \dots, u_{\sigma(k-1)}, u_{\sigma(k)}) = f_{\sigma(k)}\Big((S(\sigma)_\ell(u_{\sigma(1)}, \dots, u_{\sigma(\ell)}))_{\ell : X_{\sigma(\ell)} \in \mathrm{Pa}(X_{\sigma(k)})}, u_{\sigma(k)}\Big). \quad (6)$$

Further, KR maps corresponding to any $\mathcal{G}$-compatible ordering are the same up to a permutation[4]*.*

(ii) If the causal mechanism is identifiable within a class $\mathcal{C}$ of SEMs (Def. 1), then σ is a $\mathcal{G}$-compatible ordering if and only if the KR map $S(\sigma)$ provides a SEM in class $\mathcal{C}$.

Proof. First note that statement (ii) follows from the identifiablity assumption.

We prove (i) recursively. Without loss of generality we assume that σ is the identity permutation id and denote $S = S(\sigma) = S(\mathrm{id})$. For any random variable Y, F_Y will denote its cumulative distribution function (c.d.f). By definition, transport map S has the following form

$$S(u_1, u_2, \dots, u_d) = \begin{bmatrix} S_1(u_1) \\ S_2(u_1, u_2) \\ \vdots \\ S_d(u_1, \dots, u_d) \end{bmatrix}.$$

By definition of a compatible ordering, X_1 has no parent in $\mathcal{G}$, hence $X_1 := f_1(U_1)$. The map S_1 is by definition (see Appendix B) the monotone Brenier map between the distribution of U_1 and that of $X_1 = f_1(U_1)$. Since by assumption f_1 and F_{U_1} are strictly increasing, then $F_{f_1(U_1)}$ is invertible and this Brenier 1D map is given by

$$S_1(u_1) = F^{-1}_{f_1(U_1)} \circ F_{U_1}(u_1) = (F_{U_1} \circ f_1^{-1})^{-1} \circ F_{U_1}(u_1) = f_1(u_1) \,.$$

[4]Note that these permutations necessarily preserve the lower-triangular structure.

Then, assume that (6) holds for $1 \leq k < \ell$. Fix $u_1, \dots, u_{\ell-1} \in \mathbb{R}$. By definition again, $u_\ell \mapsto S_\ell(u_1, \dots, u_{\ell-1}, u_\ell)$ is the Brenier map between the first marginal distribution of

$$(U_\ell \,|\, U_{\ell-1} = u_{\ell-1}, \dots, U_1 = u_1) \overset{(d)}{=} U_\ell$$

and that of

$$(X_\ell \,|\, X_{\ell-1} = S_{\ell-1}(u_1, \dots, u_{\ell-1}), \dots, X_1 = S_1(u_1)) \overset{(d)}{=} f_\ell\left((S_k(u_1, \dots, u_k))_{k:X_k \in \mathrm{Pa}(X_\ell)}, u_\ell\right),$$

The second equality in distribution being justified by the fact that $\sigma = \mathrm{id}$ is a compatible ordering. Since by assumption f_ℓ is strictly increasing in the last variable and F_{U_ℓ} is strictly increasing, then $F_{f_\ell\left((S_k(u_1,\dots,u_k))_{k:X_k\in\mathrm{Pa}(X_\ell)}, U_\ell\right)} = F_{U_\ell} \circ f_\ell^{-1}\left((S_k(u_1, \dots, u_k))_{k:X_k \in \mathrm{Pa}(X_\ell)}, \cdot\right)$ is invertible and this Brenier 1D map is given by

$$\begin{aligned}
S_\ell(u_1, \dots, u_{\ell-1}, u_\ell) &= F^{-1}_{f_\ell\left((S_k(u_1,\dots,u_k))_{k:X_k\in\mathrm{Pa}(X_\ell)}, U_\ell\right)} \circ F_{U_\ell}(u_\ell) \\
&= \left[F_{U_\ell} \circ f_\ell^{-1}\left((S_k(u_1, \dots, u_k))_{k:X_k \in \mathrm{Pa}(X_\ell)}, \cdot\right)\right]^{-1} \circ F_{U_\ell}(u_\ell) \\
&= f_\ell^{-1}\left((S_k(u_1, \dots, u_k))_{k:X_k \in \mathrm{Pa}(X_\ell)}, u_\ell\right) .
\end{aligned}$$

□

Lemma 3.1. *[Adapted from [35]] Suppose π satisfies Assumption 1. Let $\pi_{\mathbf{Z}}$ denote the marginal density over $\mathbf{Z} \subseteq \mathbf{X}$. For any two variables $X_k, X_\ell \in \mathbf{Z}$, the following equivalence holds:*

$$X_k \perp\!\!\!\perp X_\ell \mid \mathbf{Z} \setminus \{X_k, X_\ell\} \iff \frac{\partial^2 \log \pi_{\mathbf{Z}}}{\partial x_k \partial x_\ell} = 0 \text{ on } \mathbb{R}^{|\mathbf{Z}|} .$$

Proof. Suppose the independence holds. Then the marginal density factorizes as follows.

$$\pi_{\mathbf{Z}} = \pi_{\mathbf{Z}\setminus\{X_k,X_\ell\}} \cdot \pi_{X_k|\mathbf{Z}\setminus\{X_k,X_\ell\}} \cdot \pi_{X_\ell|\mathbf{Z}\setminus\{X_k,X_\ell\}},$$

and therefore,

$$\log(\pi_{\mathbf{Z}}) = \log(\pi_{\mathbf{Z}\setminus\{X_k,X_\ell\}}) + \log(\pi_{X_k|\mathbf{Z}\setminus\{X_k,X_\ell\}}) + \log(\pi_{X_\ell|\mathbf{Z}\setminus\{X_k,X_\ell\}}). \tag{18}$$

It is clear from Eq. 18 that $\dfrac{\partial^2 \log \pi_{\mathbf{Z}}}{\partial x_k \partial x_\ell} = 0$ on $\mathbb{R}^{|\mathbf{Z}|}$.

For the opposite direction, note that the general solution to the PDE $\dfrac{\partial^2 \log \pi_{\mathbf{Z}}}{\partial x_k \partial x_\ell} = 0$ on $\mathbb{R}^{|\mathbf{Z}|}$ is given by $\log(\pi_{\mathbf{Z}})(\mathbf{z}) = f(\mathbf{z} \setminus \{x_k\}) + g(\mathbf{z} \setminus \{x_\ell\})$, for some functions f and g. The marginal density $\pi_{\mathbf{Z}}$ is then of the form

$$\pi_{\mathbf{Z}}(\mathbf{z}) = \exp(f(\mathbf{z} \setminus \{x_k\})) \exp(g(\mathbf{z} \setminus \{x_\ell\})).$$

Relying on the positivity of the density, we can compute the conditional density as follows:

$$\begin{aligned}
\pi_{X_k,X_\ell|\mathbf{Z}\setminus\{X_k,X_\ell\}}(\mathbf{z}) &= \frac{\pi_{\mathbf{Z}}(\mathbf{z})}{\pi_{\mathbf{Z}\setminus\{X_k,X_\ell\}}(\mathbf{z})} = \frac{\exp(f(\mathbf{z} \setminus \{x_k\})) \exp(g(\mathbf{z} \setminus \{x_\ell\}))}{\iint e^{f(\mathbf{z}\setminus\{x_k\})} e^{g(\mathbf{z}\setminus\{x_\ell\})} \mathrm{d}x_k \mathrm{d}x_\ell} \\
&= \frac{\exp(f(\mathbf{z} \setminus \{x_k\})) \exp(g(\mathbf{z} \setminus \{x_\ell\}))}{\int e^{f(\mathbf{z}\setminus\{x_k\})} \mathrm{d}x_k \int e^{g(\mathbf{z}\setminus\{x_\ell\})} \mathrm{d}x_\ell} \\
&= \frac{\exp(f(\mathbf{z} \setminus \{x_k\})) \int e^{g(\mathbf{z}\setminus\{x_\ell\})} \mathrm{d}x_\ell}{\int e^{f(\mathbf{z}\setminus\{x_k\})} \mathrm{d}x_k \int e^{g(\mathbf{z}\setminus\{x_\ell\})} \mathrm{d}x_\ell} \cdot \frac{\exp(g(\mathbf{z} \setminus \{x_\ell\})) \int e^{f(\mathbf{z}\setminus\{x_k\})} \mathrm{d}x_k}{\int e^{f(\mathbf{z}\setminus\{x_k\})} \mathrm{d}x_k \int e^{g(\mathbf{z}\setminus\{x_\ell\})} \mathrm{d}x_\ell} \\
&= \frac{\pi_{\mathbf{Z}\setminus\{X_\ell\}}(\mathbf{z})}{\pi_{\mathbf{Z}\setminus\{X_k,X_\ell\}}(\mathbf{z})} \cdot \frac{\pi_{\mathbf{Z}\setminus\{X_k\}}(\mathbf{z})}{\pi_{\mathbf{Z}\setminus\{X_k,X_\ell\}}(\mathbf{z})} = \pi_{X_k|\mathbf{Z}\setminus\{X_k,X_\ell\}}(\mathbf{z}) \cdot \pi_{X_\ell|\mathbf{Z}\setminus\{X_k,X_\ell\}}(\mathbf{z}),
\end{aligned}$$

which implies the desired conditional independence relation. □

Lemma 4.1. *Suppose π satisfies Assumption 1. Let σ be an ordering. The following are equivalent.*

- *π is induced by an ANM, and σ is $\mathcal{G}-$compatible.*
- *for all $1 \leq k \leq d$, there exists a strictly increasing map $B_k(\sigma) : \mathbb{R} \to \mathbb{R}$ such that*

$$\tfrac{\partial}{\partial x_{\sigma(k)}} B_k(\sigma) \circ S(\sigma)_k(x_{\sigma(1)}, \dots, x_{\sigma(k)}) - 1 = 0 . \tag{13}$$

Proof. The first direction is proved in the main text, applying Theorem 1 and considering the form of the KR map in equation (12). For the other direction, without loss of generality we assume that σ is the identity permutation id, the proof being identical for any other permutation. The general solution to the PDE

$$\frac{\partial}{\partial x_k} B_k \circ S_k(x_1, \dots, x_k) - 1 = 0$$

is

$$B_k \circ S_k(x_1, \dots, x_k) = x_k - h_k(x_1, \dots, x_{k-1}), \tag{19}$$

for some function $h_k : \mathbb{R}^{k-1} \to \mathbb{R}$. By definition of transport map S_k, $S_k(X_1, \dots, X_k)$ is a standard Gaussian variable. By (19), denoting $U_k := -B_k \circ S_k(X_1, \dots, X_k)$, we have for all k, $X_k = h_k(X_1, \dots, X_{k-1}) + U_k$. By independence of the Gaussian marginals, the U variables are independent. □

Lemma C.1. *Suppose π satisfies Assumption 1. Let σ be an ordering. The following are equivalent.*

- *π is induced by a PNL model, and σ is $\mathcal{G}-$compatible.*
- *For all $1 \leq k \leq d$, there exists a strictly increasing map $B_k(\sigma) : \mathbb{R} \to \mathbb{R}$ such that for all $1 \leq \ell \leq d$ where $\ell \neq k$,*

$$\frac{\partial^2}{\partial x_{\sigma(l)} \partial x_{\sigma(k)}} B_k(\sigma) \circ S(\sigma)_k(x_{\sigma(1)}, \dots, x_{\sigma(k)}) = 0. \tag{20}$$

Proof. Here again, without loss of generality we assume that σ is the identity permutation id, the proof being identical for any other permutation.

For the first direction, since h_k is strictly increasing, Theorem 1 applies, and the KR map S from π to η is of the form

$$S_k(x_1, \dots, x_k) = M_k \left(h_k^{-1}(x_k) - g_k((x_\ell)_{X_\ell \in \mathrm{Pa}(X_k)}) \right), \tag{21}$$

where M_k is the strictly increasing transport map from the distribution of U_k to a standard Gaussian $\mathcal{N}(0, 1)$. With $B_k := M_k^{-1}$, S_k is thus a solution to PDE (20).

For the other direction, let us assume that for all $1 \leq \ell < k$,

$$\frac{\partial^2}{\partial x_k \partial x_\ell} B_k \circ S_k = 0\,. \tag{22}$$

Applying (22) for $\ell = 1$, this implies that for all $B_k \circ S_k$ is of the form $B_k \circ S_k(x_1, \dots, x_k) = a_1(x_2, \dots, x_k) - b_1(x_1, \dots, x_{k-1})$, where a_1 again satisfies (22) for $\ell = 2$, which again implies that a_1 is of the form $a_1(x_2, \dots, x_k) = a_2(x_3, \dots, x_k) - b_2(x_2, \dots, x_{k-1})$. Iterating over $1 \leq \ell \leq k-1$, we obtain that $B_k \circ S_k$ is of the form

$$B_k \circ S_k(x_1, \dots, x_k) = g(x_k) - h(x_1, \dots, x_{k-1})\,. \tag{23}$$

By definition, S_k is strictly increasing in x_k, and B_k is a strictly increasing transport map. Therefore, Eq. (23) implies that g is strictly increasing in x_k, and g^{-1} is well-defined. By definition of transport map S_k, $S_k(X_1, \dots, X_k)$ is a standard Gaussian variable. By (19), denoting $U_k := -B_k \circ S_k(X_1, \dots, X_k)$, we have for all k, $g(X_k) = h(X_1, \dots, X_{k-1}) + U_k$. By independence of the Gaussian marginals, the U variables are independent. Finally, applying the function g^{-1} to both sides, we get $X_k = g^{-1}(h(X_1, \dots, X_{k-1}) + U_k)$, which is a PNL considering the independence of U variables.

□

D PNLs

Due to space limitations, the discussion on post-nonlinear (PNL) models and the derivations of PNLloss were postponed to this appendix.

D.1 Post non-linear models

Post non-linear (PNL) models [40, 39, 44], known to be a general identifiable class of models, are defined as follows.

Definition 4 (PNL). *We say that the SEM of Eq.* (3) *forms a post-nonlinear model if for all* $1 \leq k \leq d$,

$$f_k((X_\ell)_{X_\ell \in \mathrm{Pa}(X_k)}, U_k) \coloneqq h_k \left(g_k((X_\ell)_{X_\ell \in \mathrm{Pa}(X_k)}) + U_k \right), \tag{24}$$

where the functions h_k *are strictly increasing.*

Note that the PNL model reduces to an ANM when h_k is the identity for all k. That is, ANMs are a special case of PNLs. The PNL class is identifiable if we prevent some singular functions and noise distributions [41]. We show the following characterization of PNLs.

Lemma C.1. *Suppose* π *satisfies Assumption 1. Let* σ *be an ordering. The following are equivalent.*

- π *is induced by a PNL model, and* σ *is* $\mathcal{G}-$*compatible.*
- *For all* $1 \leq k \leq d$, *there exists a strictly increasing map* $B_k(\sigma) : \mathbb{R} \to \mathbb{R}$ *such that for all* $1 \leq \ell \leq d$ *where* $\ell \neq k$,

$$\tfrac{\partial^2}{\partial x_{\sigma(l)} \partial x_{\sigma(k)}} B_k(\sigma) \circ S(\sigma)_k(x_{\sigma(1)}, \ldots, x_{\sigma(k)}) = 0. \tag{20}$$

In view of Lemma C.1, for a given ordering σ, we can parameterize each map $B_k(\sigma)$ with vector $\boldsymbol{\beta}_k$ as in (16), and $\boldsymbol{\beta}_k^*$ is now estimated by optimizing the loss given by Lemma C.1:

$$\begin{aligned} \boldsymbol{\beta}_k^* &:= \arg\min_{\boldsymbol{\beta}_k} \sum_{\substack{1 \leq \ell \leq k \\ \ell \neq k}} \mathbb{E}_\pi \left[\left| \frac{\partial^2}{\partial x_{\sigma(l)} \partial x_{\sigma(k)}} [B_k(\sigma)]_{\boldsymbol{\beta}_k} \circ (S(\sigma)_{\boldsymbol{\alpha}^*})_k(X_{\sigma(1)}, \ldots, X_{\sigma(k)}) \right| \right] \\ &\approx \arg\min_{\boldsymbol{\beta}_k} [\mathrm{PNLloss}_k(\sigma, \mathbf{x})](\boldsymbol{\beta}_k), \end{aligned} \tag{25}$$

where

$$[\mathrm{PNLloss}_k(\sigma, \mathbf{x})](\boldsymbol{\beta}_k) := \sum_{\substack{1 \leq \ell \leq k \\ \ell \neq k}} \sum_{i=1}^{n} \left| \frac{\partial^2}{\partial x_{\sigma(l)} \partial x_{\sigma(k)}} [B_k(\sigma)]_{\boldsymbol{\beta}_k} \circ (S(\sigma)_{\boldsymbol{\alpha}^*})_k(x^i_{\sigma(1)}, \ldots, x^i_{\sigma(k)}) \right|. \tag{26}$$

PNL loss. The *PNL loss* of an ordering σ, parameterized by $\boldsymbol{\gamma} \in (\mathbb{R}_{>0})^d$ is defined as

$$\mathrm{PNLloss}_{\boldsymbol{\gamma}}(\sigma, \mathbf{x}) := \sum_{k=1}^{d} \gamma_k [\mathrm{PNLloss}_k(\sigma, \mathbf{x})](\boldsymbol{\beta}_k), \tag{27}$$

where $\boldsymbol{\alpha}^*$, $\boldsymbol{\beta}_k^*$ and $\mathrm{ANMloss}_k(\sigma, \mathbf{x})$ are defined in (7), (25) and (26).

We note in particular that

$$\mathrm{ANMloss}_{\boldsymbol{\gamma}}(\sigma) = 0 \implies \mathrm{PNLloss}_{\boldsymbol{\gamma}}(\sigma) = 0,$$

which agrees with the fact that $\{\text{ANMs}\} \subsetneq \{\text{PNLs}\}$.

E Further on numerical experiments

In this section, we first provide comprehensive details of the numerical experiments included in the main text. Subsequently, we unveil novel numerical experiments, including a numerical experiment with the real-world dataset 'Sachs' [29].

Parameters α **and** τ**.** In all our experiments, the parameters α are chosen so that all Hermite polynomials/functions involved are of degree 2. The thresholds $\tau_{k\ell}^{\mathbf{Z}}$ in Alg. 1 are defined as follows

$$\tau_{k\ell}^{\mathbf{Z}} := \delta_{|\mathbf{Z}|} \times \varsigma(\Omega_{k\ell}^{\mathbf{Z}}), \tag{28}$$

where $\varsigma(\Omega_{k\ell}^{\mathbf{Z}})$ is defined in (10), and the $\delta_{|\mathbf{Z}|}$s are constants depending on $|\mathbf{Z}|$, i.e. the size of the subspace of variables. They are tuned as follows:

$\lvert\mathbf{Z}\rvert$	2	3	4	5	6
Value of $\delta_{\lvert\mathbf{Z}\rvert}$	0.17	0.3	0.4	0.5	0.6

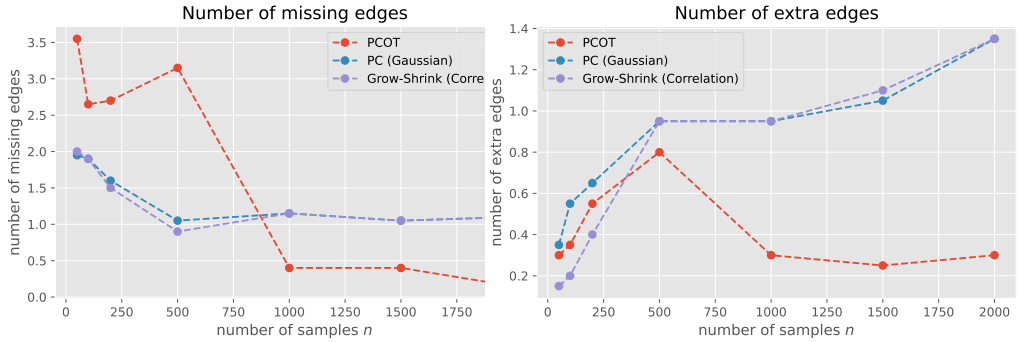
Table 1: Values of $\delta_{|\mathbf{Z}|}$ for the thresholds $\tau_{k\ell}^{\mathbf{Z}}$ in Alg. 1, defined by (28).

E.1 Details of the experiments in the text

PC-OT experiments. These experiments were conducted based on the following SEM:

$$\begin{aligned} X_1 &:= U_1 \\ X_2 &:= U_2 \\ X_3 &:= X_1^2 + X_2 + U_3 \\ X_4 &:= U_4 \\ X_5 &:= 0.5X_1^2 - 0.5X_4^2 + X_1X_4 + U_5 \\ X_6 &:= X_4^3 - X_5 + U_6 \end{aligned} \quad \text{with} \quad \begin{aligned} U_1 &\sim 0.2\mathcal{N}(0,1) \times \mathcal{N}(0,1) \\ U_2 &\sim (\text{Gumbel}(0,0.7) - 2.5)/2.5 \\ U_3 &\sim \mathcal{N}(0,1) \times \text{Exp}(1)/8 \\ U_4 &\sim (\text{Ber}(1/2) \times \text{Exp}(1) - 3)/2 \\ U_5 &\sim (\text{Ber}(1/2) \times \Gamma(2,3) - 24)/12 \\ U_6 &\sim \text{Gumbel}(0,0.5) - 1.5 \end{aligned} \tag{29}$$

The underlying causal graph $\mathcal{G}$ is given by Figure 1b. Figure 4 illustrates the decomposition of these error terms. Note that the comparison between the number of misoriented edges was included in Figure 2a.



(a) average number of missing edges (b) average number of extra edges

Figure 4: Decomposition of the errors made by PC-OT, PC (Gaussian) and Grow-Shrink.

ANM experiments. Within these experiments, we worked with the following SEM:

$$\begin{aligned} X_1 &:= U_1 \\ X_2 &:= 0.5X_1^2 + U_2 \\ X_3 &:= \log(X_1^2) + U_3 \\ X_4 &:= 2X_2(X_2+1) + U_4 \\ X_5 &:= 0.5X_1^2 - 0.5X_4^2 + X_1X_4 + U_5 \\ X_6 &:= 0.25X_3^2 - X_5 + U_6 \end{aligned} \quad \text{with} \quad \begin{aligned} U_1 &\sim 0.2\mathcal{N}(0,1)^2 \\ U_2 &\sim 0.5\mathcal{N}(-2.5,1) \\ U_3 &\sim \log(\mathcal{N}(0,1)^2 + 1) \\ U_4 &\sim 0.3\mathcal{N}(0,1)^2 \\ U_5 &\sim \log(\mathcal{N}(0,1)^2 + 1) \\ U_6 &\sim \mathcal{N}(0,1) \end{aligned} \tag{30}$$

The underlying causal graph $\mathcal{G}$ is given by Figure 1a. For each sample size, we repeated the experiment 20 times, and the box plots of the ANMlosses corresponding to each permutation was depicted in Figure 3. The permutation corresponding to the true causal order was $1 \to 2 \to 4 \to 5 \to 3 \to 6$, which had the lowest ANMloss among all compatible permutations. Further, the gap between the ANMlosses increased as the number of samples grew larger.

E.2 Further experiments

Real-world data. In this section, we consider a dataset corresponding to the causal relations among components of a cellular signaling network based on single-cell data, namely 'Sachs' dataset [29]. This dataset comprises samples of 7446 primary human immune system cells. We consider a

subnetwork of this dataset corresponding to the proteins Pcl_γ, $PIP3$, $PIP2$, PKC and Akt. The causal mechanisms between these proteins are depicted in Figure 5.

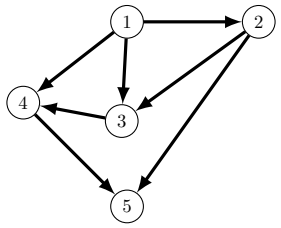


Figure 5: Causal mechanisms pertaining to the proteins $1 \coloneqq Pcl_\gamma$, $2 \coloneqq PIP3$, $3 \coloneqq PIP2$, $4 \coloneqq PKC$ and $5 \coloneqq Akt$.

Since the dataset comprises values between 1.0 and 9058, we applied a logarithm function so that the support spans the real numbers. We then provided the Markov equivalence class of Figure 5 (which consists of 10 different DAGs) to our ANM-OT algorithm. Table 2 below demonstrates the ANMlosses corresponding to each compatible permutation. As can be seen in Table 2, the ground truth permutation $1 \to 2 \to 3 \to 4 \to 5$ has the second lowest ANMloss, following permutation $2 \to 1 \to 3 \to 4 \to 5$, which is a transposition of the true permutation.

Permutation	ANMloss (ANM-OT)
$2 \to 1 \to 3 \to 4 \to 5$	28,517.74
$1 \to 2 \to 3 \to 4 \to 5$	31,732.97
$1 \to 4 \to 3 \to 2 \to 5$	33,776.88
$1 \to 3 \to 4 \to 2 \to 5$	35,402.31
$2 \to 3 \to 1 \to 4 \to 5$	36,563.22
$4 \to 1 \to 3 \to 2 \to 5$	38,816.92
$3 \to 2 \to 1 \to 4 \to 5$	46,310.82
$4 \to 3 \to 1 \to 2 \to 5$	48,449.60
$3 \to 1 \to 4 \to 2 \to 5$	49,448.33
$3 \to 4 \to 1 \to 2 \to 5$	49,686.34

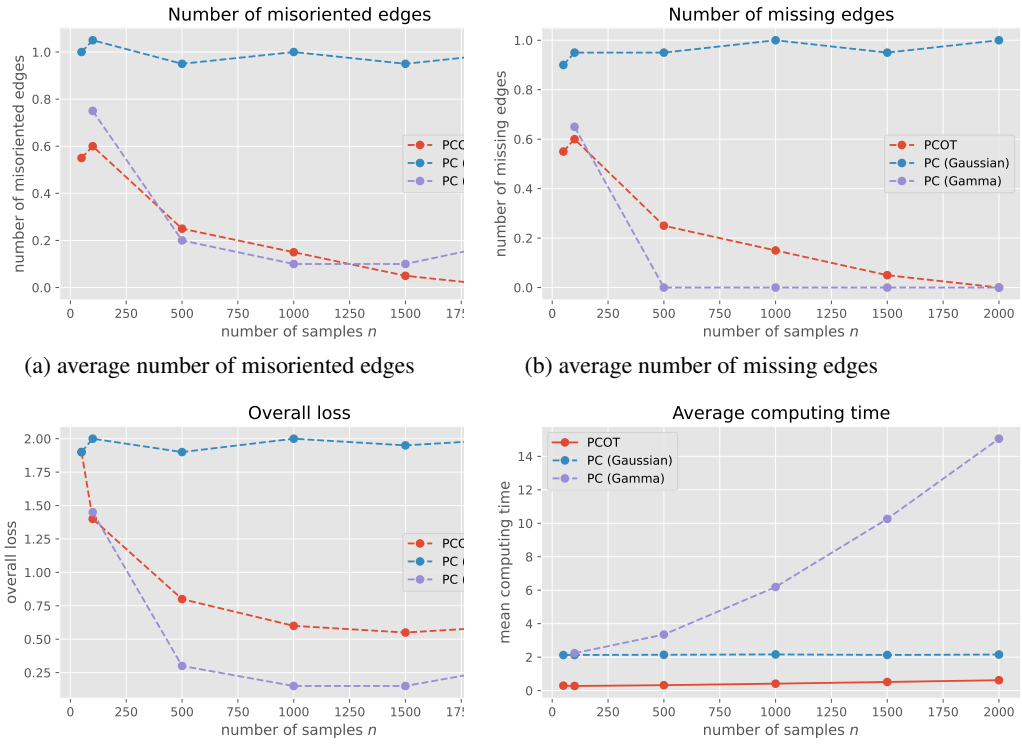
Table 2: ANMlosses pertaining to the permutations compatible with the Markov equivalence class of the DAG in Figure 5.

Another illustrative example for PC-OT. To illustrate the effectiveness of PC-OT on data with non-Gaussian noise, we provide a numerical experiment on a small model. The SEM we consider is as follows.

$$\begin{aligned} X_1 &:= U_1/450 \\ X_2 &:= U_2 \\ X_3 &:= (X_2^3 + \log(|X_2|X_1^2 + U_3))/15 \end{aligned} \quad \text{with} \quad \begin{aligned} U_1 &\overset{(d)}{=} \sqrt{4/3}(\mathrm{Pow}(4) - 3/2) \text{ conditioned to be } \leq 1000 \\ U_2 &\overset{(d)}{=} (\mathrm{Gumbel}(0, 0.7) - 2.5)/2.5 \\ U_3 &\overset{(d)}{=} \mathrm{Ber}(1/2) \times \mathrm{Exp}(1/2) \end{aligned} \tag{31}$$

Note that the DAG corresponding to the SEM of Eq. (31) is a v-structure, namely $X_1 \to X_3 \leftarrow X_2$. We repeated the experiments of Section 5 using the SEM of Eq. (31). For comprehensiveness, we also included a version of PC algorithm provided with a kernel-based CI test, namely HSIC-Gamma [10] provided in the CDT package [14]. The results are depicted in Figure 6. As witnessed in Figure 6, PC-OT performs significantly better than PC with Gaussian CI tests.

Time complexity. Although the performance of PC-OT is comparable to PC with the kernel-based CI tests, the computing time of the kernel-based algorithm appears to be drastically growing with sample size. In contrast, PC-OT does not suffer from a growing runtime. It is noteworthy that with 2000 samples, the kernel-based method necessitates a runtime that is 14 times greater compared to that of PC-OT.



(a) average number of misoriented edges

(b) average number of missing edges

(c) overall average loss

(d) average computing time of the algorithms

Figure 6: Performance of PC-OT, PC (Gaussian CI test) and PC (HSIC-Gamma CI test) on the illustrative example with SEM of Eq. (31).