Distinguishing Cause from Effect on Categorical Data: The Uniform Channel Model

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

Distinguishing cause and effect from observations of a pair of random variables is a core problem in causal discovery. Most approaches that have been proposed for this task (*i.e.*, to break the symmetry of the probabilistic formulation), namely *additive noise models* (ANM), are only adequate for quantitative data. We propose a criterion to address the cause-effect problem with categorical variables (living in sets with no meaningful order), inspired by seeing a conditional *probability mass function* (pmf) as a discrete memoryless channel. We select as the most likely causal direction that in which the conditional pmf is closer to a *uniform channel* (UC). The underlying rationale is the following: in a UC, the conditional entropy (of the effect given the cause) is independent/decoupled from the distribution of the cause, in agreement (as ANMs) with the principle of *independence of cause and mechanism*. Our approach thus extends the ANM rationale to categorical variables, and we call it *uniform channel model* (UCM). To assess how *close* a conditional pmf (estimated from data) is to being a UCM, we use classical statistical testing, supported by a closed-form estimate of a UC channel. On the theoretical front, we prove identifiability of the UC and show its equivalence with a structural causal model. Finally, the proposed method compares favorably with recent state-of-the-art alternatives in experiments on synthetic, benchmark, and real data.

1. Introduction

Causal inference is a key problem in many areas of science and data analysis (Pearl, 2009). In principle, distinguishing statistical dependencies from causal relationships requires interventions (Pearl, 2009; Peters et al., 2017). However, intervening is often impossible (*e.g.*, analyzing past data), impractical, or unethical (*e.g.*, forcing people to smoke), which has stimulated much research aimed at inferring causal relationships (*causal discovery*) from purely observational data (Janzing, 2019; Mooij et al., 2016; Peters et al., 2017), or mixed observational-interventional data (Faria et al., 2022)

There is a vast literature on methods to learn *directed acyclic graphs* (DAGs) from data (Vowels et al., 2023), by inferring *conditional independences* (CIs) among variables (Koller and Friedman, 2010). However, as is well known, those methods cannot distinguish among causal DAGs entailing the same CIs. The simplest instance of this problem involves a pair of variables (X, Y): purely statistical methods cannot recover the causal graph because $X \to Y$ and $Y \to X$ constitute a so-called *Markov equivalence class* (MEC), corresponding to the two possible factorizations of the joint distribution: $p_{X,Y}(x, y) = p_{Y|X}(y|x)\, p_X(x) = p_{X|Y}(x|y)\, p_Y(y)$.

Without interventions, choosing one of the members of a MEC requires additional assumptions about the underlying data-generating mechanism. For instance, *additive noise models* (ANM) (Shimizu et al., 2006; Hoyer et al., 2009; Peters et al., 2011, 2014) assume the effect is a function of the cause plus a noise term independent of the cause ($Y = f_Y(X) + N_Y, X \perp\!\!\!\perp N_Y$); if the same doesn't hold in the reverse direction, the model is said to be *identifiable*. ANMs are generically identifiable (Peters et al., 2011, 2014) in the following sense: if the joint density $p_{X,Y}$ corresponds to

© 2023 .

an ANM, the conditional density $p_{Y|X}(\cdot|x)$ has identical *shape* for any x, simply being *shifted* by $f_Y(x)$, but $p_{X|Y}(\cdot|y)$ typically depends on y in a more complicated way. Under the ANM criterion, if such a model exists in one direction but not the other, the former is selected as the causal direction.

The ANM can be seen as an instance of the principle of *independence of cause and mechanism* (ICM) (Janzing et al., 2010, 2012), according to which the cause-effect mechanism (a deterministic function followed by addition of noise, in an ANM) is *independent*[1] of the cause, thus of its distribution. The ICM principle has been exploited using different tools to define and assess the notion of *independence*: information geometry (Janzing et al., 2012); Kolmogorov complexity (not computable, but approximable (Li and Vitányi, 2009)) by Janzing et al. (2012) and Mian et al. (2021); stochastic complexity, via the *minimum description length* (MDL (Rissanen, 1998)) principle (Budhathoki and Vreeken, 2017a; Marx and Vreeken, 2019; Tagasovska et al., 2020) or the *minimum message length* (Wallace and Boulton, 1968) criterion, by Stegle et al. (2010).

Very few methods have been proposed to address the cause-effect problem with categorical variables. Peters et al. (2011) have extended ANMs to the discrete case, but only for variables taking values in a set equipped with a meaningful order, in which an operation similar to addition (a shift) is defined. They consider $\mathbb{Z}$ (or subsets of $\mathbb{Z}$) for variables without cyclic structure, and $\mathbb{Z}/n\mathbb{Z}$ with modulo-n addition, for variables with cyclic structure (*e.g.*, seasons or months). However, purely categorical variables live in sets with no order, where a meaningful notion of addition/shift cannot be defined, precluding the direct use of ANMs. Peters et al. (2011) consider what they call "structureless" sets, but only with a particular form of conditional distribution, not generally applicable. Some of the MDL-based methods mentioned above have been used for categorical variables: Budhathoki and Vreeken (2017a) proposed CISC (*causal inference by stochastic complexity*); Cai et al. (2018) proposed HCR (*hidden compact representation*), based on the *Bayesian information criterion* (BIC). Finally, Liu and Chan (2016) assess mechanism independence via a *distance correlation* (DC) between the cause distribution and the conditional distribution of the effect.

We propose a new approach to the cause-effect problem for categorical variables, inspired by viewing the causal mechanism as a communication channel. This view allows extending to the categorical case a key feature of ANMs: the conditional (differential) entropy of the effect given the cause is independent[1] of the distribution of the cause (Janzing et al., 2012). For categorical variables (*i.e.*, symbols, in channel terminology), a memoryless channel corresponds to the conditional *probability mass function* (pmf) of the output given the input, the *channel matrix* $\boldsymbol{\theta}^{X\to Y}$, where $\boldsymbol{\theta}^{X\to Y}_{x,y} = p_{Y|X}(y|x) = \mathbb{P}[Y = y|X = x]$. In a so-called *uniform channel* (UC (Hamming, 1986)), the rows of this matrix are permutations of each other[2], implying (as shown below) that the conditional entropy $H(Y|X)$ is independent of the distribution of X. Paralleling the ANM rationale, given a pair of categorical variables (X, Y), if the conditional pmf in one direction, say of Y given X, corresponds to a UC and the same is not true in the other direction, then the causal structure is declared to be $X \to Y$. This criterion, which we refer to as the UCM (*uniform channel model*) is supported by an *identifiability* result proved in this paper: if a joint distribution corresponds to a UCM in one direction, in general (*i.e.*, with probability 1 under any continuous density on the model parameters), it does not correspond to a UCM in the reverse direction.

1. The term "independent" here does not have a probabilistic sense, but a functional sense: changing the distribution of the cause does not affect the causal mechanim, and vice-versa.
2. This is not necessarily what is called a *symmetric channel*, which requires additionally that all the columns are also permutations of each other (Cover and Thomas, 2006).

The proposed UCM approach is further supported by the fact (proved below) that if, and only if, $\boldsymbol{\theta}^{X\to Y}$ corresponds to a UCM, is it possible to write a *structural causal model* (SCM) (Pearl, 2009) of the form $Y = f_Y(X, U_Y)$, where f_Y is a deterministic function and U_Y is an exogenous random variable, taking values in in the same set as Y and independent of X. The importance of this independence was recently highlighted by Papineau (2022), "the probabilistic independence of exogenous terms in (...) structural equations holds the key to causal direction".

The final question that needs to be addressed is how to instantiate the UCM principle with a finite amount of data; this question parallels the question of how to estimate the underlying function and noise distribution in the ANM approach for quantitative variables. Naturally, with a finite amount of data, we only have access to an estimate of the underlying distribution, and the probability that this estimate corresponds exactly to a UCM in one of the two possible directions is vanishingly small. Although there are other conceivable ways to address this issue, here we resort to statistical hypothesis testing to decide in which direction, if any, the conditional pmf can be considered an UCM. An essential building block of this approach is the ability to estimate a channel under the constraint that it is uniform; this is a problem that, to the best of our knowledge, had not be studied before and for which we derive a closed-form solution. We also extend the approach to the case where the rows of the channel matrix are cyclic permutations of each other (a *cyclic UCM* – CUCM), applicable when the effect variable has cyclic nature, but in this case the channel estimate has to be obtained iteratively.

The main contributions of this paper are the following:

- A new instantiation, for categorical variables, of the principle of *independence of cause and mechanism*: the *uniform channel model* (UCM) principle.
- A proof of identifiability of the UCM.
- A proof that the joint distribution of a pair of categorical random variables is entailed by a structural causal model in which the exogenous noise has the same cardinality as the effect variable if and only if it corresponds to a UCM.
- An instantiation of the UCM principle using classical statistical hypothesis testing, supported on a closed-form estimate of a UCM (which, to the best of our knowledge, is also a new result and may be of independent interest).

The paper is organized as follows. Section 2 describes the UC model for causal inference and presents the corresponding identifiability theorem. Section 3 addresses the problem of estimating uniform and cyclic uniform channels from data. Section 4 describes how the criterion is applied with observed data. Experimental results are reported in Section 5, and Section 6 concludes the paper.

2. Uniform Channel Models – UCM – for Categorical Variables

This section describes the proposed causal inference principle for categorical variables, after introducing notation and reviewing the notion of uniform channel. Finally, we prove an identifiability theorem for the proposed model and show its equivalence with an SCM.

2.1. Categorical Variables and Uniform Channels

Let $X \in \mathcal{X} = \{1, ..., |\mathcal{X}|\}$ and $Y \in \mathcal{Y} = \{1, ..., |\mathcal{Y}|\}$ be two *categorical random variables* (although their outcomes are shown as integers, no role is played by their order). The joint pmf $p_{X,Y}$ can be factored in two different ways, $p_{X,Y}(x, y) = p_{Y|X}(y|x)\, p_X(x) = p_{X|Y}(x|y)\, p_Y(y)$, corresponding to a *Markov equivalence class*. If $X \perp\!\!\!\perp Y$, the joint pmf factors trivially $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$. Let the vector of parameters of the first factorization be denoted as $\boldsymbol{\theta} = (\boldsymbol{\theta}^X, \boldsymbol{\theta}^{X\to Y})$, *i.e.*,

$$\theta_x^X = p_X(x) = \mathbb{P}[X = x] \quad \text{and} \quad \theta_{x,y}^{X\to Y} = p_{Y|X}(y|x) = \mathbb{P}[Y = y|X = x],$$

where $\boldsymbol{\theta}^X \in \Delta_{|\mathcal{X}|-1}$, with Δ_{m-1} being the probability simplex in $\mathbb{R}^m$. The conditional probabilities are arranged in a $|\mathcal{X}| \times |\mathcal{Y}|$ row-stochastic matrix $\boldsymbol{\theta}^{X\to Y}$, with the x-th row denoted as $\boldsymbol{\theta}_x^{X\to Y}$.

Definition 1 (Discrete Memoryless Channel) *(Cover and Thomas, 2006; Hamming, 1986) A discrete memoryless channel (DMC) is a probabilistic system with a discrete input alphabet $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ and a discrete output alphabet $\mathcal{Y} = \{1, \ldots, |\mathcal{Y}|\}$, specified by the conditional probabilities $p_{Y|X}(y|x)$, for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The adjective "memoryless" means that, given a sequence of random inputs, the corresponding outputs are conditionally independent.*

Definition 2 (Uniform channel (UC)) *(Hamming, 1986) A UC is a DMC in which each row of the conditional probability (channel) matrix $\boldsymbol{\theta}^{X\to Y}$ is a permutation of every other row.*

Definition 3 (Cyclic uniform channel (CUC)) *A CUC is a UC where each row of the channel matrix $\boldsymbol{\theta}^{X\to Y}$ is a cyclic permutation of every other row.*

Let $\mathbb{S}_{|\mathcal{Y}|}$ denote the set of all $|\mathcal{Y}|!$ permutations of $(1, ..., |\mathcal{Y}|)$. In a UC, each row $\boldsymbol{\theta}_x^{X\to Y}$ is a row-specific permutation $\sigma_x \in \mathbb{S}_{|\mathcal{Y}|}$ of a common vector $\boldsymbol{\gamma} \in \Delta_{|\mathcal{Y}|-1}$, *i.e.*,

$$\boldsymbol{\theta}_x^{X\to Y} = (\gamma_{\sigma_x(1)}, \ldots, \gamma_{\sigma_x(|\mathcal{Y}|)}) \quad \Leftrightarrow \quad p_{Y|X}(y|x) = \gamma_{\sigma_x(y)}. \tag{1}$$

In the case of a CUC, $\sigma_x \in \mathbb{C}_{|\mathcal{Y}|}$, the set of all $|\mathcal{Y}|$ cyclic permutations of $(1, ..., |\mathcal{Y}|)$.

2.2. Uniform Channel Model – UCM – for Categorical Variables

We propose a new principle to infer the most likely causal direction between two categorical variables by following the rationale behind ANMs. Recall that the ANM principle for real variables is as follows: if Y satisfies an ANM $Y = f_Y(X) + N_Y$, where $N_Y \perp\!\!\!\perp X$ (*i.e.*, the *exogenous* noise is independent of X), but the same is not true in the reverse direction, then the most likely causal direction is $X \to Y$. In the ANM for real variables, the conditional *probability density function* (pdf) has identical *shape* for all values of x, simply being *shifted* by $f_Y(x)$, *i.e.*, $p_{Y|X}(y|x) = p_{N_Y}(y - f_Y(x))$, where p_{N_Y} is the pdf of the noise variable N_Y. Consequently, the *conditional differential entropy* $h(Y|X)$ does not depend on the pfd of X, as shown next.

Proposition 4 *If real-valued variables X and Y admit an ANM from X to Y, then the conditional differential entropy $h(Y|X) = h(N_Y)$, independently of the distribution of X.*

Proof: Using the shift-invariance (a) of differential entropy (Cover and Thomas, 2006),

$$\begin{aligned} h(Y|X) &= \mathbb{E}_{X,Y}[-\log p_{Y|X}(Y|X)] \;=\; \mathbb{E}_X\big[\mathbb{E}_{Y|X}[-\log p_{Y|X}(Y|X)]\big] \\ &= \mathbb{E}_X\big[\mathbb{E}_{Y|X}[-\log p_{N_Y}(Y - f_Y(X))]\big] \overset{(a)}{=} \mathbb{E}_X\big[h(N_Y)\big] \;=\; h(N_Y). \end{aligned}$$ ■

For categorical variables, the sets $\mathcal{X}$ and $\mathcal{Y}$ lack any meaningful order, thus there is no notion of *addition*, and an ANM is not directly applicable. However, the conditional entropy invariance property of ANMs can be preserved by considering the transformation group under which discrete entropy is invariant: *permutations*. Consequently, our proposed causal inference principle is:

UCM causal inference principle for categorical variables: given two categorical variables X and Y, if the conditional pmf $\boldsymbol{\theta}^{X\to Y}$ corresponds to a UCM, but the conditional pmf $\boldsymbol{\theta}^{Y\to X}$ does not, then we infer the causal direction to be $X \to Y$.

Paralleling Proposition 4, the following result is a simple consequence of the invariance of (discrete) entropy to symbol permutations.

Proposition 5 *If $\boldsymbol{\theta}^{X\to Y}$ corresponds to a UC (each row of $\boldsymbol{\theta}^{X\to Y}$ is a permutation of a vector $\boldsymbol{\gamma} \in \Delta_{|\mathcal{Y}|-1}$), then the conditional entropy $H(Y|X) = H(\boldsymbol{\gamma})$, independently of p_X.*

Proof: Due to the permutation-invariance property (a) of entropy (Cover and Thomas, 2006),

$$\begin{aligned} H(Y|X) = \mathbb{E}_{X,Y}[-\log p_{Y|X}(Y|X)] \;&=\; \mathbb{E}_X\big[\mathbb{E}_{Y|X}[-\log p_{Y|X}(Y|X)]\big] \\ &= \mathbb{E}_X\big[\mathbb{E}_{Y|X}[-\log \gamma_{\sigma_X(Y)}]\big] \overset{(a)}{=} \mathbb{E}_X\big[H(\boldsymbol{\gamma})\big] \;=\; H(\boldsymbol{\gamma}). \end{aligned}$$ ■

The proposed causal inference problem can be seen as an embodiment of the *independence of cause and mechanism* principle, in which *independence* corresponds to the following property: the conditional pmf of the effect given the cause has the same collection of probability values, only their positions depend on the cause. Thus, the conditional uncertainty (entropy) of the effect, given the cause, is independent of the distribution of the cause. As in the ANM approach, we assume *causal sufficiency*, *i.e.*, that there are no unobserved/hidden common causes of the two observed variables.

A relevant fact that provides further support to the proposed principle is that the UCM can be written as a *structural causal model* (SCM) (Pearl, 2009), as shown in the following proposition.

Proposition 6 *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be a pair of dependent random variables such that the conditional pmf $\boldsymbol{\theta}^{X\to Y}$ corresponds to a UCM specified by $\boldsymbol{\gamma}$ and $\sigma_1, ..., \sigma_{|\mathcal{X}|}$, as in (1), and $\boldsymbol{\theta}^X$ be arbitrary. Then, the joint pmf of X and Y is entailed by the following SCM:*

$$X := U_X, \qquad Y := f_Y(X, U_Y), \tag{2}$$

with $U_X \perp\!\!\!\perp U_Y$ (independent exogenous variables), $U_X \in \mathcal{X}$ is categorical with pmf $\boldsymbol{\theta}^X$, $U_Y \in \mathcal{Y}$ is categorical with pmf $\boldsymbol{\gamma}$, and $f_Y : \mathcal{X}\times\mathcal{Y} \to \mathcal{Y}$ is a function given by $f_Y(x, u) = \tau_x(u)$, with $\tau_x = \sigma_x^{-1}$ (inverse[3] permutation of σ_x). Conversely, if the conditional pmf $\boldsymbol{\theta}^{X\to Y}$ does not correspond to a UCM, it is impossible to write an SCM of the form (2), with $U_Y \in \mathcal{Y}$, entailing the joint pmf of X and Y.

3. Given a permutation $\sigma \in \mathbb{S}_{|\mathcal{Y}|}$, its inverse $\sigma^{-1} \in \mathbb{S}_{|\mathcal{Y}|}$ is such that $\sigma^{-1}(\sigma(i)) = i$, for any $i = 1, ..., |\mathcal{Y}|$.

Proof: Given $x \in \mathcal{X}$, $Y = f_Y(x, U)$ is a categorical random variable with conditional pmf

$$p_{Y|X}(y|x) = \mathbb{P}[Y = y|X = x] = \mathbb{P}[U_Y = \tau_x^{-1}(y)] = \gamma_{\sigma_x(y)}$$

(which coincides with (1)), where the second equality stems from the definition of f_Y and permutations being bijections. Conversely, if the conditional pmf $\boldsymbol{\theta}^{X\to Y}$ does not correspond to a UCM (neither all its rows are equal to each other, by assumption), depending on what value/category X takes, the conditional pmf of Y takes different probability values, not just a permutation of a common pmf, making it impossible to write an SCM of the form (2) with $U_Y \in \mathcal{Y}$ independent of U_X. ■

Remark 7 *If none, or both, of the conditional pmfs, $\boldsymbol{\theta}^{X\to Y}$ and $\boldsymbol{\theta}^{Y\to X}$, correspond to a UCM, the proposed criterion does not select a causal direction. Of course, in practice, the conditional pmfs are estimated from a finite data set, thus there is a very small chance that one of these estimates corresponds exactly to a UCM. In Section 4, we come back to this issue, proposing statistical tests do decide if a conditional pmf estimate can be considered to correspond to a UCM. In the following subsection, we assume that we have an infinite amount of data (equivalently, the true underlying pmf) and address the identifiability issue in this ideal condition. If this model was not identifiable in this ideal setting, it would be hard to argue that it could be useful with a finite amount of data.*

Remark 8 *Our UCM contains as a particular case the model for "structureless" sets proposed by Peters et al. (2011). Their model assumes a function $\phi : \mathcal{X} \to \mathcal{Y}$ and $p_{Y|X}(y|x) = p$, if $y = \phi(x)$, and $p_{Y|X}(y|x) = (1-p)/(|\mathcal{Y}|-1)$, if $y \neq \phi(x)$. This corresponds to a UC (in fact, a CUC), with*

$$\boldsymbol{\gamma} = \big(p, (1-p)/(|\mathcal{Y}|-1), ..., (1-p)/(|\mathcal{Y}|-1)\big)$$

and any set of permutations such that $\sigma_x(y) = 1$, for $y = f(x)$. Our UC and CUC models are much more general, as they do not constraint the conditional pmf to have only two different values.

2.3. Identifiability

For the proposed criterion to be useful, it should be supported by an indentifiability guarantee, *i.e.*, that the set of joint probability mass functions $p_{X,Y}$ such that both $p_{Y|X}$ and $p_{X|Y}$ correspond to uniform channels should be as small as possible, ideally have zero Lebesgue measure in the space of valid parameters, thus zero probability under any continuous density (Peters et al., 2011). Before stating and proving the general identifiability result, we illustrate it for the case where both variables are binary: $\mathcal{X} = \mathcal{Y} = \{1, 2\}$. Let $p_X(1) = \theta_1^X = \beta$ and let $\boldsymbol{\theta}^{X\to Y}$ correspond to a UCM (in this case, simply a *binary symmetric channel*) with error probability α (Cover and Thomas, 2006):

$$\boldsymbol{\theta}^{X\to Y} = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix}.$$

Of course, a channel matrix where the two rows are equal to $(1-\alpha, \alpha)$ is also a UC, but in that case X and Y are independent, which is an uninteresting case. The channel in the reverse direction, *i.e.*, $\boldsymbol{\theta}^{Y\to X}$, can be easily derived using Bayes law, yielding

$$\boldsymbol{\theta}^{Y\to X} = \begin{bmatrix} \dfrac{(1-\alpha)\beta}{(1-\alpha)\beta + \alpha(1-\beta)} & \dfrac{\alpha(1-\beta)}{(1-\alpha)\beta + \alpha(1-\beta)} \\ \dfrac{\alpha\beta}{\alpha\beta + (1-\alpha)(1-\beta)} & \dfrac{(1-\alpha)(1-\beta)}{\alpha\beta + (1-\alpha)(1-\beta)} \end{bmatrix}.$$

Notice that in matrix $\boldsymbol{\theta}^{Y\to X}$, the variable Y indexes rows and X indexes columns, so that it is row-stochastic as is standard for channel matrices. Matrix $\boldsymbol{\theta}^{Y\to X}$ represents a UC if and only if one (or both) of two conditions are satisfied: the diagonal elements are equal to each other; the elements in the first column are equal to each other (in which case, $X \perp\!\!\!\perp Y$). Simple algebraic manipulation allows showing that this is equivalent to having $(\alpha,\beta) \in \{(\alpha,\beta) \in [0,1]^2 : \alpha = 0 \vee \alpha = 1/2 \vee \alpha = 1 \vee \beta = 0 \vee \beta = 1/2 \vee \beta = 1\}$, which has zero Lebesgue measure. The following theorem generalizes this result for arbitrary $|\mathcal{X}|$ and $|\mathcal{Y}|$.

Theorem 9 *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two categorical random variables with a joint pmf such that the conditional $\boldsymbol{\theta}^{X\to Y}$ corresponds to a UC. Assume also that the marginals have full support[4]: $p_Y(y) \neq 0$, for any $y \in \mathcal{Y}$, and $p_X(x) \neq 0$, for any $x \in \mathcal{X}$. Further assume that the rows of the channel matrix $\boldsymbol{\theta}^{X\to Y}$ are not all equal to each other (i.e., X and Y are not independent[5]). Then, the set of parameters such that the reverse channel $\boldsymbol{\theta}^{Y\to X}$ is also a UCM has zero Lebesgue measure.*

The proof, presented in Appendix A, essentially boils down to showing that the UC condition on the reverse channel $\boldsymbol{\theta}^{Y\to X}$ corresponds to the zero set of a polynomial that is not identically zero, thus has zero Lebesgue measure, a classical result from polynomial theory (Federer, 1969).

3. Channel Estimation from Data

This section addresses the problem of estimating channel parameters $\boldsymbol{\theta}^{X\to Y}$ from N independent and identically distributed samples of $(X,Y) : (x_1,y_1),...,(x_N,y_N)$. This will play a key role in translating the causal inference criterion proposed in Section 2 to the realistic scenario where there is only access to a finite amount of data, rather than perfect knowledge of the joint pmf $p_{X,Y}$.

Before estimating the channel parameters (conditional pmf), notice that estimating the marginal pmf $\boldsymbol{\theta}^X$ is trivial: with N_x denoting the number of samples (x_i,y_i) with $x_i = x$, the *maximum likelihood* (ML) estimate of $\boldsymbol{\theta}^X$ is given by

$$\hat{\boldsymbol{\theta}}^X = \arg\max_{\boldsymbol{\theta}\in\Delta_{|\mathcal{X}|-1}} \sum_{x\in\mathcal{X}} N_x \log \theta_x^X = \Big(\frac{N_1}{N},...,\frac{N_{|\mathcal{X}|}}{N}\Big). \tag{3}$$

For estimating $\boldsymbol{\theta}^{X\to Y}$, we consider the following 4 scenarios: **(1)** arbitrary channel; **(2)** UCM, with known permutations; **(3)** UCM, with *unknown* permutations; **(4)** CUCM, with *unknown* cyclic permutations. Scenarios 1 and 2 are trivial and considered only as they provide the building blocks to address scenarios 3 and 4.

3.1. Scenario 1: Arbitrary Channel

Let $N_{x,y}$ be the number of samples (x_i,y_i) such that $x_i = x$ and $y_i = y$. In the absence of constraints other than each row of $\boldsymbol{\theta}^{X\to Y}$ must be a valid pmf, the ML estimate is

$$\hat{\boldsymbol{\theta}}^{X\to Y} = \arg\max_{\boldsymbol{\theta}\in(\Delta_{|\mathcal{Y}|-1})^{|\mathcal{X}|}} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} N_{x,y} \log \theta_{x,y}. \tag{4}$$

4. There is no loss of generality in this assumption; if there are zeros in the marginals, we simply redefine $\mathcal{X}$ or/and $\mathcal{Y}$ by removing the zero-probability elements.
5. If all the rows are equal to each other, then $Y \perp\!\!\!\perp X$; since independence is a symmetrical relationship, the reverse channel $\boldsymbol{\theta}^{Y\to X}$ will also have all its rows equal to each other, thus being a special case of a UC channel.

Since both the objective function and the constraints in (4) are separable across $x = 1, ..., |\mathcal{X}|$, the problem is also separable into a collection of problems, each yielding the classical ML estimates

$$\hat{\theta}_{x,y}^{X\to Y} = N_{x,y}/N_x, \quad \text{for } (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{5}$$

3.2. Scenario 2: UCM with Known Permutations

In a UC, each row $\boldsymbol{\theta}_x^{X\to Y}$ is as given in (1). If the permutations $\sigma_1, \ldots, \sigma_{|\mathcal{X}|}$ are known, the ML estimate of $\boldsymbol{\gamma}$ is given by

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\Delta_{|\mathcal{Y}|-1}}{\arg\max} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} N_{x,y} \log \gamma_{\sigma_x(y)}. \tag{6}$$

This problem is not separable, as all the rows of the channel matrix share the same probability values, although with different permutations. Swapping the summation order and using the inverse permutations $\tau_x = \sigma_x^{-1}$ to do a change of variable in the sum over $\mathcal{Y}$, problem (6) can be rewritten as

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\Delta_{|\mathcal{Y}|-1}}{\arg\max} \sum_{z\in\mathcal{Y}} \log \gamma_z \sum_{x\in\mathcal{X}} N_{x,\tau_x(z)}. \tag{7}$$

Problem (7) has the same form as (3), the solution being simply

$$\hat{\gamma}_y = \frac{1}{N} \sum_{x\in\mathcal{X}} N_{x,\tau_x(y)}, \quad \text{for } y \in \mathcal{Y}. \tag{8}$$

Notice that $N_{x,\tau_x(y)}$ is the number of samples (x_i, y_i) such that $x_i = x$ and $y_i = \tau_x(y)$.

3.3. Scenario 3: UC with Unknown Permutations

In this case, the log-likelihood is maximized, not only w.r.t. $\boldsymbol{\gamma}$, but also the permutations. Although at first sight this may look like a very hard problem, as there are $(|\mathcal{Y}|!)^{|\mathcal{X}|}$ combinations of permutations, we show next that it can be solved very efficiently. The optimization problem in hand (formulated w.r.t. the inverse permutations, denoted as $\tau_1, \ldots, \tau_{|\mathcal{X}|} \in \mathbb{S}_{|\mathcal{Y}|}$) is

$$\hat{\boldsymbol{\gamma}}, \hat{\tau}_1, \ldots, \hat{\tau}_{|\mathcal{X}|} = \underset{\substack{\boldsymbol{\gamma}\in\Delta_{|\mathcal{Y}|-1} \\ \tau_1, ..., \tau_{|\mathcal{X}|} \in \mathbb{S}_{|\mathcal{Y}|}}}{\arg\max} \mathcal{L}(\boldsymbol{\gamma}, \tau_1, ..., \tau_{|\mathcal{X}|}), \tag{9}$$

where

$$\mathcal{L}(\boldsymbol{\gamma}, \tau_1, ..., \tau_{|\mathcal{X}|}) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} N_{x,\tau_x(y)} \log \gamma_y. \tag{10}$$

The following proposition (proved in Appendix B) provides the solution to this problem.

Proposition 10 *A globally optimal solution to the problem specified in* (9)–(10) *is given as follows. For $x \in \mathcal{X}$, $\hat{\tau}_x$ is any permutation that sorts $\{N_{x,1}, ..., N_{x,|\mathcal{Y}|}\}$ into non-increasing order,*

$$\hat{\tau}_x \ \textit{ is such that } \ N_{x,\hat{\tau}_x(1)} \geq \cdots \geq N_{x,\hat{\tau}_x(|\mathcal{Y}|)}, \tag{11}$$

and, for $y \in \mathcal{Y}$,

$$\hat{\gamma}_y = \frac{1}{N} \sum_{x\in\mathcal{X}} N_{x,\hat{\tau}_x(y)}. \tag{12}$$

The solution in (11)–(12) is a global optimum, but is not unique; for any other pmf $\hat{\boldsymbol{\xi}}$ that is a permutation of $\hat{\boldsymbol{\gamma}}$, *i.e.*, $\hat{\gamma}_y = \hat{\xi}_{\rho(y)}$, where $\rho \in \mathbb{S}_{|\mathcal{Y}|}$, we have

$$\hat{\theta}^{X\to Y}_{x,y} = \hat{\gamma}_{\hat{\sigma}_x(y)} = \hat{\xi}_{\rho(\hat{\sigma}_x(y))}. \tag{13}$$

That is, $\boldsymbol{\gamma}$ is identifiable only up to a permutation, since any permutation of $\hat{\boldsymbol{\gamma}}$, combined with the inverse of that permutation composed with each σ_x, yields the same conditional pmf estimate $\hat{\boldsymbol{\theta}}^{X\to Y}$. Finally, notice that the cost of computing this solution scales as $O(|\mathcal{X}|\,|\mathcal{Y}|\,\log|\mathcal{Y}|)$, due to the $|\mathcal{X}|$ sorting operations of size $|\mathcal{Y}|$.

3.4. Scenario 4: CUC with Unknown Permutations

The difference between this and the previous case is that the permutations are now cyclic. Thus, the corresponding optimization problem is identical to (9)–(10), but with the constraint $\tau_1, ..., \tau_{|\mathcal{X}|} \in \mathbb{S}_{|\mathcal{Y}|}$ replaced with $\tau_1, ..., \tau_{|\mathcal{X}|} \in \mathbb{C}_{|\mathcal{Y}|}$, where $\mathbb{C}_{|\mathcal{Y}|}$ is the set of cyclic permutations of $\{1, ..., |\mathcal{Y}|\}$.

Although the cardinality of $\mathbb{C}_{|\mathcal{Y}|}$ is $|\mathcal{Y}|$, much smaller than that of $\mathbb{S}_{|\mathcal{Y}|}$, which is $|\mathcal{Y}|!$, this problem is harder than (9)–(10). Whereas the cost of the exact solution of (9)–(10) scales with $O(|\mathcal{X}|\,|\mathcal{Y}|\,\log|\mathcal{Y}|)$, exactly solving this problem by exhaustive search costs $O(|\mathcal{Y}|^{|\mathcal{X}|})$. It happens that this problem is a variant of a class of problems known as *multireference alignment*, which is known to be NP-hard (Bandeira et al., 2014). Exact solutions are thus out of the question for large problems. Here, we propose an alternating maximization approach with two steps:

- Given the current permutation estimates $\hat{\tau}_1, ..., \hat{\tau}_{|\mathcal{X}|}$, update $\hat{\boldsymbol{\gamma}}$ according to (8), with $\tau_x = \hat{\tau}_x$.
- Given the current $\hat{\boldsymbol{\gamma}}$, maximize w.r.t. the permutations, which is separable across $\tau_1, ..., \tau_{|\mathcal{X}|}$:

$$\hat{\tau}_x = \arg\max_{\tau\in\mathbb{C}_{|\mathcal{Y}|}} \sum_{y=1}^{|\mathcal{Y}|} N_{x,\tau(y)} \log\hat{\gamma}_y, \quad \text{for } x = 1, ..., |\mathcal{X}|. \tag{14}$$

This maximization is carried out exactly by considering all the $|\mathcal{Y}|$ cyclic permutations.

The costs of both steps of this algorithm scale as $O(|\mathcal{X}|\,|\mathcal{Y}|)$. Convergence can be proved via the approach used to prove convergence of the K-means algorithm (Selim and Ismail, 1984), as both algorithms share a common structure: alternate between exact maximization w.r.t. real quantities (cluster centers, in K-means, $\boldsymbol{\gamma}$ in our algorithms) and an exact combinatorial optimization (the point-to-cluster assignments in K-means, the cyclic permutations in the proposed algorithms).

4. Applying the UCM Principle from Data

Applying the proposed causal inference principle amounts to performing hypothesis testing concerning the UC nature of the conditional pmf estimates $\hat{\boldsymbol{\theta}}^{X\to Y}$ and $\hat{\boldsymbol{\theta}}^{Y\to X}$. This is closely related to classical tests for two-way contingency tables (Agresti, 2013; Read and Cressie, 1988). Given a table of counts $N_{x,y}$, let the *null hypothesis* H_0 be that these counts can be explained by a UCM in the $X \to Y$ direction. To test this hypothesis, consider the corresponding maximum log-likelihood (noting that $p_{X,Y}(x,y) = \mathbb{P}[X = x, Y = y] = \theta^X_x\,\gamma_{\sigma_x(y)}$, for a UCM),

$$\mathcal{L}_{H_0} = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} N_{x,y}\log\big(\hat{\theta}^X_x\,\hat{\gamma}_{\hat{\sigma}_x(y)}\big) = \sum_{x\in\mathcal{X}} N_x \log\hat{\theta}^X_x + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} N_{x,y}\log\big(\hat{\gamma}_{\hat{\sigma}_x(y)}\big), \tag{15}$$

where $\hat{\boldsymbol{\theta}}^X$ is the ML estimate of $\boldsymbol{\theta}^X$, and $\hat{\sigma_1}, ..., \hat{\sigma}_{|\mathcal{X}|}$ and $\hat{\boldsymbol{\gamma}}$ are the ML estimates of the UCM, obtained as shown in Section 3.

The alternative hypothesis is that the channel is arbitrary, with maximum log-likelihood

$$\mathcal{L}_{\bar{H}_0} = \sum_{x \in \mathcal{X}} N_x \log \hat{\theta}_x^X + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} N_{x,y} \log\big(N_{x,y}/N_x\big), \tag{16}$$

since the ML estimates of the conditional pmf parameters are as given in (5), and the ML estimate of the marginal $\boldsymbol{\theta}^X$ is the same, regardless of the channel being uniform or not. These models are nested: a UCM is a particular case of the set of all valid channels, thus it is always true that $\mathcal{L}_{H_0} \leq \mathcal{L}_{\bar{H}_0}$.

The so-called *likelihood-ratio statistic* (LRS), denoted G^2, is then given by

$$G^2_{X \to Y} = 2(\mathcal{L}_{\bar{H}_0} - \mathcal{L}_{H_0}) = 2 \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} N_{x,y} \log\Big(\frac{N_{x,y}}{\hat{\gamma}_{\hat{\sigma}_x(y)} N_x}\Big); \tag{17}$$

notice that $\hat{\gamma}_{\hat{\sigma}_x(y)} N_x$ is the expected value of $N_{x,y}$ under the null hypothesis. This is the LRS in the $X \to Y$ direction, which we indicate with the subscript $X \to Y$. The LRS in the reverse direction, denoted $G^2_{Y \to X}$ is computed in the same way, after swapping the roles of X and Y.

It is well known that G^2 is asymptotically χ^2-distributed with df $= (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)$ *degrees of freedom*, yielding the p-value $\mathtt{p} = \mathbb{P}[\chi^2_{\text{df}} \geq G^2] = 1 - \mathbb{P}[\chi^2_{\text{df}} < G^2]$, where $\mathbb{P}[\chi^2_{\text{df}} < G^2]$ is the cumulative distribution function of a χ^2_{df} distribution. If $\mathtt{p}$ is less than some significance level (*i.e.*, the test statistic G^2 is too large), the null hypothesis is rejected.

Let $\mathtt{p}^{X \to Y}$ and $\mathtt{p}^{Y \to X}$ be the $\mathtt{p}-$values of the LRS for testing the uniformity of the channels in both directions, and let α be a significance level for the test (Agresti, 2013; Read and Cressie, 1988), *i.e.*, the null hypothesis is rejected if the $\mathtt{p}-$value is less than α. Having a statistical test of whether an estimated conditional pmf corresponds to a UCM, we adopt a procedure similar to the one proposed by Peters et al. (2011).

- If $\mathtt{p}^{X \to Y} \geq \alpha$ and $\mathtt{p}^{Y \to X} < \alpha$, declare $X \to Y$.
- If $\mathtt{p}^{X \to Y} < \alpha$ and $\mathtt{p}^{Y \to X} \geq \alpha$, declare $Y \to X$.
- If $\mathtt{p}^{X \to Y} < \alpha$ and $\mathtt{p}^{Y \to X} < \alpha$, declare "undecided: wrong model".
- If $\mathtt{p}^{X \to Y} \geq \alpha$ and $\mathtt{p}^{Y \to X} \geq \alpha$, declare "undecided: both directions possible".

The fourth case (*i.e.*, the hypotheses that the channel is uniform in both directions cannot be rejected) is asymptotically improbable, unless X and Y are independent, due to the identifiability guarantee. Alternatively, to force the method to make a decision between the two causal directions, one may simply decide for $X \to Y$, if $\mathtt{p}^{X \to Y} > \mathtt{p}^{Y \to X}$, and for $Y \to X$, otherwise.

5. Results

We evaluate the proposed UCM approach on synthetic, benchmark, and real data, comparing it to two state-of-the-art causal direction selection methods for categorical variables: DC (Liu and Chan, 2016) (`eda.mmci.uni-saarland.de/prj/cisc/`, with $\epsilon = 0$) and HCR (Cai et al., 2018) (a Python version of the R code available at `cran.r-project.org/web/packages/HCR/index.html`). The code for all the experiments will be made available upon acceptance of the manuscript.

In the ML estimates underlying our approach (namely (5) and (8)), to avoid the problem of zero or vanishing probabilities, we use a small amount (10^{-3}) of additive (a.k.a. Dirichlet) smoothing.

5.1. Identifying the UCM Direction

The first set of experiments is essentially a sanity check, assessing the ability of the proposed log-likelihood-based criterion to identify the UCM direction, using synthetic data, with different sample sizes N and different sizes of the support sets, $|\mathcal{X}|$, and $|\mathcal{Y}|$. For each pair $(|\mathcal{X}|, |\mathcal{Y}|)$ and each N, we generate 100 independent datasets using randomly generated UCMs in the $X \to Y$ direction and the results reported for each N are the corresponding averages. The decision rule is simply to choose $X \to Y$, if $\mathrm{p}^{X\to Y} \geq \mathrm{p}^{Y\to X}$ (equivalently, $G^2_{X\to Y} \leq G^2_{Y\to X}$), and $Y \to X$ (which is wrong), otherwise. The results in Fig. 1 show that the accuracy achieves high values, close to 100%, for $N > 500 \sim 1000$, and there is no clear effect of the sizes of the support sets or difference between the non-cyclic and the cyclic cases.

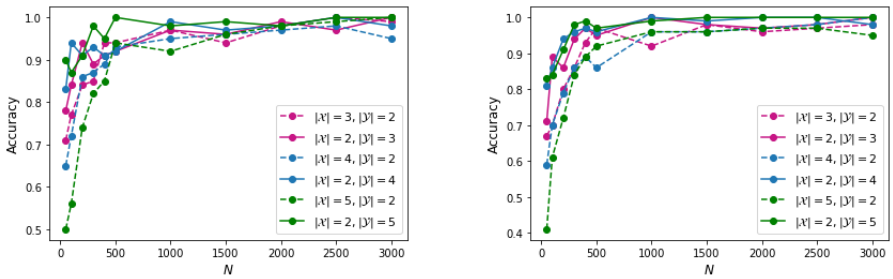

Figure 1: Accuracy in selecting the UCM direction, using the proposed criterion, for different sample and support sizes. The left plot corresponds to general UCMs and the one on the right to cyclic UCMs (CUCM).

5.2. Benchmark Data

We now consider the *cause-effect pairs* benchmark set by Guyon et al. (2013). We use only the 112 pairs in which both variables are categorical or binary and that have as ground truth that $X \to Y$ or $Y \to X$. We compare the UCM approach (using $\alpha = 0.05$ in the decision procedure described in Section 4) with the two methods mentioned above: DC (Liu and Chan, 2016) and HCR (Cai et al., 2018). In any of the methods testes, a decision of "undecided" is counted as wrong. The results in Table 1 show that UCM outperforms both HCR and DC on this dataset.

Table 1: Results on the benchmark dataset: accuracy on the 112 pairs considered.

UCM	DC	HCR
0.61	0.41	0. 47

5.3. Real Data

Finally, we evaluate the UCM method (again with $\alpha = 0.05$) on real data from the *UCI Machine Learning Repository* (Bache and Lichman, 2013). We use pairs of variables from the following datasets: Adult, Pittsburgh Bridges, Accute Inflamation, Temperature, and Horse Colic. The datasets

and selected pairs, as well as the criteria used to decide what is the ground truth causal direction, are described in Appendix C. We include only pairs for which a test of independence, at significance level 0.05 (Agresti, 2013), rejects the null hypothesis of independence. Furthermore, we include only pairs for which at least one of the three tested methods chooses one of the causal directions.

Table 2: Results on real data. Wrong decisions are shown in red; UWM stands for "undecided: wrong model". Month is a cyclic variable, thus a CUC was used in the $Y \to X$ direction.

Dataset	X	Y	UCM	DC	HCR
Adult	Occupation	Income	UWM	$X \to Y$	$X \to Y$
Adult	Work Class	Income	UWM	$X \to Y$	$X \to Y$
Acute Inflammation	Inflam. of urinary bladder	Lumbar pain	$Y \to X$	Inconcl.	Inconcl.
Acute Inflammation	Inflam. of urinary bladder	Nausea	$Y \to X$	Inconcl.	Inconcl.
Acute Inflammation	Inflam. of urinary bladder	Burning urethra	$Y \to X$	Inconcl.	Inconcl.
Pittsburgh Bridges	Material	Lanes	$X \to Y$	$Y \to X$	$X \to Y$
Pittsburgh Bridges	Purpose	Type	UWM	$Y \to X$	$X \to Y$
Temperature	Month	Temperature	$X \to Y$	$X \to Y$	$Y \to X$
Horse Colic	Abdomen Status	Surgical Lesion	UWM	$X \to Y$	$X \to Y$

Table 2 shows that the UCM and HCR approaches found the “correct" causal direction in 5 out of 9 pairs, and DC in 4 pairs. UCM returned only correct decisions or abstained from deciding. This small number of experiments do now allow reaching any strong conclusions, but suggest that UCM performs on par, arguably somewhat better, with DC and HCR.

6. Conclusions

We presented a new approach, termed UCM (*uniform channel model*), for the cause-effect problem with categorical variables. The proposed method extends the rationale of *additive noise models* (ANM) to the case of categorical data, by exploiting the interpretation of conditional distribution as communication channels. In fact, UCM can be seen as an ANM-type instantiation of the principle of independence of cause and mechanism, which preserves a key feature of ANM for quantitative data: the conditional entropy of the effect given the cause is independent of distribution of the cause. The rationale of the approach is in the right causal direction, the conditional probability mass function is closer to a uniform channel than in the reverse direction.

The core results of this paper are a proof of identifiability of the UCM and a proof of its equivalence to a structural causal model.

To instantiate the approach on finite data, we resorted to statistical tests to decide in which of the two directions (if any) is the conditional probability mass function sufficiently close to corresponding to a UCM. The experimental results confirmed the adequacy of the proposed method. By comparing our method with other recent methods (DC, by Liu and Chan (2016), and HCR, by Cai et al. (2018)), we found that UCM outperforms those other methods on benchmark and performs on par with them on real data.

As future work, we will extend the proposed method to handle more than two variables and to cases where one variable is categorical and the other is continuous.

References

A. Agresti. *Categorical Data Analysis*. Wiley, 2013. 3rd Edition.

K. Bache and M. Lichman. UCI Machine Learning Repository, 2013. http://archive.ics.uci.edu/ml/index.php.

A. Bandeira, M. Charikar, A. Singer, and A. Zhu. Multireference alignment using semidefinite programming. In *5th Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 459–470, 2014.

M. Best and N. Chakravarti. Active set algorithms for isotonic regression: An unifying framework. *Mathematical Programming*, 47:425–439, 1990.

K. Budhathoki and J. Vreeken. MDL for causal inference on discrete data. In *IEEE International Conference on Data Mining*, pages 751–756, 2017a.

K. Budhathoki and J. Vreeken. Causal inference by stochastic complexity, 2017b. arXiv:1702.06776.

K. Budhathoki and J. Vreeken. Accurate causal inference on discrete data. In *IEEE International Conference on Data Mining*, pages 881–886, 2018.

R. Cai, J. Qiao, K. Zhang, Z. Zhang, and Z. Hao. Causal discovery from discrete data using hidden compact representation. In *Neural Information Processing Systems*, 2018.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

G. Faria, A. Martins, and M. Figueiredo. Differentiable causal discovery under latent interventions. In B. Schölkopf, C. Uhler, and K. Zhang, editors, *First Conference on Causal Learning and Reasoning (CLeaR)*, volume 177 of *Proc. of Machine Learning Research*, pages 253–274, 2022.

H. Federer. *Geometric Measure Theory*. Springer–Verlag, 1969.

I. Guyon, A. Statnikov, and M. Henaff. Causality challenge #3: Cause-effect pairs, 2013. http://www.causality.inf.ethz.ch/cause-effect.php.

R. Hamming. *Coding and Information Theory*. Prentice-Hall, 1986.

G. Hardy, J. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 1952.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Processing Information Systems*, page 689–696, 2009.

D. Janzing. The cause-effect problem: Motivation, ideas, and popular misconceptions. In I. Guyon, A. Statnikov, and B. Batu, editors, *Cause Effect Pairs in Machine Learning*, pages 3–26. Springer, 2019.

D. Janzing, P. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *International Conference on Machine Learning*, pages 479—486, 2010.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183: 1–31, 2012.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2010.

S. Lang. *Algebra*. Springer–Verlag, 2002. 3rd Edition.

M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2009.

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

A. Marx and J. Vreeken. Indentifiability of cause and effect using regularized regression. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 852–861, 2019.

O. Mian, A. Marx, and J. Vreeken. Discovering fully oriented causal networks. *AAAI Conference on Artificial Intelligence*, 35(10):8975–8982, 2021.

J. Mooij, J.Peters, D. Janzing, J. Zscheischelr, and B.Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17: 1–102, 2016.

D. Papineau. The statistical nature of causation. *The Monist*, 105(2):247–275, 2022.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 2011.

J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. In *Journal of Machine Learning Research*, volume 15, pages 2009–2053, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

T. Read and N. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, 1988.

J. Rissanen. Stochastic complexity in statistical inquiry theory. *World Scientific*, 1998.

S. Selim and M. Ismail. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (1):81–87, 1984.

S. Shimizu, P. Hoyer, A. Hyvarinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. In *Journal of Machine Learning Research*, volume 7, pages 2003–2030, 2006.

O. Stegle, D. Janzing, K. Zhang, J. Mooij, and B. Schöolkopf. Probabilistic latent variable models for distinguishing cause and effect. In *Neural Information Processing Systems*, 2010.

N. Tagasovska, V. Chavez-Demoulin, and T. Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*, volume PMLR 119, pages 9311–9323, 2020.

M. Vowels, N. Camgoz, and R.Bowden. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys*, 2023. Accepted, available at https://dl.acm.org/doi/pdf/10.1145/3527154.

C. Wallace and D. Boulton. An information measure for classification. *Computer Journal*, 11: 185–194, 1968.

Appendix A. Proof of Theorem 9

Proof: Let us denote $\boldsymbol{\theta}^X = \boldsymbol{\beta} \in \Delta_{|\mathcal{X}|-1}$ and recall that under the UC assumption, matrix $\boldsymbol{\theta}^{X\to Y}$ has the form

$$\boldsymbol{\theta}^{X\to Y} = \begin{bmatrix} \gamma_{\sigma_1(1)} & \gamma_{\sigma_1(2)} & \cdots & \gamma_{\sigma_1(|\mathcal{Y}|)} \\ \gamma_{\sigma_2(1)} & \gamma_{\sigma_2(2)} & \cdots & \gamma_{\sigma_2(|\mathcal{Y}|)} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{\sigma_{|\mathcal{X}|}(1)} & \gamma_{\sigma_{|\mathcal{X}|}(2)} & \cdots & \gamma_{\sigma_{|\mathcal{X}|}(|\mathcal{Y}|)} \end{bmatrix},$$

where $\sigma_1, ..., \sigma_{|\mathcal{X}|} \in \mathbb{S}_{|\mathcal{Y}|}$ are permutations and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{|\mathcal{Y}|}) \in \Delta_{|\mathcal{Y}|-1}$. The assumption that the rows of this matrix are not all equal to each other precludes the two following condition from holding: $\sigma_1 = \sigma_2 = \cdots = \sigma_{|\mathcal{X}|}$ and $\boldsymbol{\gamma} = (1, 1, ..., 1)/|\mathcal{Y}|$.

Using Bayes law, it is trivial to obtain the reverse channel, the elements of which are given by

$$\theta^{Y\to X}_{y,x} = p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)\, p_X(x)}{p_Y(y)} = \frac{\gamma_{\sigma_x(y)}\, \beta_x}{\sum_{x'\in\mathcal{X}} \gamma_{\sigma_{x'}(y)}\, \beta_{x'}} = \frac{a_{y,x}}{A_y}, \tag{18}$$

where $a_{y,x} = \beta_x \gamma_{\sigma_x(y)}$ and $A_y = p_Y(y) \neq 0$ (by assumption). As in the binary example, using variables Y and X to index rows and columns, respectively, $\boldsymbol{\theta}^{Y\to X}$ is a row-stochastic matrix:

$$\boldsymbol{\theta}^{Y\to X} = \begin{bmatrix} a_{1,1}/A_1 & \cdots & a_{1,|\mathcal{X}|}/A_1 \\ \vdots & \ddots & \vdots \\ a_{|\mathcal{Y}|,1}/A_{|\mathcal{Y}|} & \cdots & a_{|\mathcal{Y}|,|\mathcal{X}|}/A_{|\mathcal{Y}|} \end{bmatrix}.$$

For $\boldsymbol{\theta}^{Y\to X}$ to correspond to a UC, its rows must be permutations of each other, which is equivalent to all being permutations of one of them, say the first, without loss of generality. We exclude the case where these permutations are all equal to identity, since that would correspond to all rows of $\boldsymbol{\theta}^{Y\to X}$ being equal to each other, *i.e.*, $X \perp\!\!\!\perp Y$, which is excluded in the conditions of the theorem. The condition that all the rows are permutations of the first one can be written formally as

$$\exists(\rho_2, ..., \rho_{|\mathcal{Y}|}) \in \mathbb{L} : \forall y \in \mathcal{Y} \setminus \{1\},\, \forall x \in \mathcal{X},\ a_{1,x}/A_1 = a_{y,\rho_y(x)}/A_y, \tag{19}$$

where $\mathbb{L} = (\mathbb{S}_{|\mathcal{X}|})^{|\mathcal{Y}|-1} \setminus \mathbb{I}$, with $\mathbb{I} = \big\{\rho_2, ..., \rho_{|\mathcal{Y}|} : \rho_2 = ... = \rho_{|\mathcal{Y}|} = \iota\big\}$, and ι is the identity permutation. In words, $\mathbb{L}$ is the set of all $(|\mathcal{Y}| - 1)$-tuples of permutations of $|\mathcal{X}|$ elements, except for the one in which all permutations are identity.

The equality $a_{1,x}/A_1 = a_{y,\rho_y(x)}/A_y$ is equivalent to $(a_{1,x}\, A_y - a_{y,\rho_y(x)}\, A_1)^2 = 0$, thus the following equivalence holds:

$$\big(\forall y \in \mathcal{Y} \setminus \{1\},\, \forall x \in \mathcal{X},\ a_{1,x}/A_1 = a_{y,\rho_y(x)}/A_y\big) \quad \Leftrightarrow \quad Q_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = 0,$$

with

$$Q_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \sum_{y\in\mathcal{Y}\setminus\{1\}} \sum_{x\in\mathcal{X}} (a_{1,x}\, A_y - a_{y,\rho_y(x)}\, A_1)^2, \tag{20}$$

where we have written the model parameters compactly as $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Delta_{|\mathcal{X}|-1} \times \Delta_{|\mathcal{Y}|-1} \subset \mathbb{R}^{|\mathcal{X}|+|\mathcal{Y}|}$, and denoted $\boldsymbol{\rho} = (\rho_2, ..., \rho_{|\mathcal{Y}|}) \in \mathbb{L}$. A key observation is that $Q_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ is a polynomial in the elements of $\boldsymbol{\theta}$, since the $a_{y,x}$ and the A_y are themselves polynomials (either products of two elements or sums of products of pairs of elements) as is clear in (18).

Finally, the existential quantifier in (19) can be re-written using a product, *i.e.*,

$$\big(\exists \boldsymbol{\rho} \in \mathbb{L}: \; Q_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = 0\big) \quad \Leftrightarrow \quad R(\boldsymbol{\theta}) = 0, \qquad \text{where } \; R(\boldsymbol{\theta}) \; = \prod_{\boldsymbol{\rho} \in \mathbb{L}} Q_{\boldsymbol{\rho}}(\boldsymbol{\theta}).$$

Since $R(\boldsymbol{\theta})$ a product of polynomials, it is itself a polynomial. Consequently, we have shown that the UC condition in (19) corresponds to having $\boldsymbol{\theta}$ as a root of a polynomial.

The rest of the proof relies on a classical result about polynomials (Federer, 1969): let $S : \mathbb{R}^n \to \mathbb{R}$ be a polynomial that is not identically zero; then, the set $S^{-1}(0) = \{\mathbf{u} \in \mathbb{R}^n : \; S(\mathbf{u}) = 0\}$ has zero Labesgue measure in $\mathbb{R}^n$. All that is left to show then is that $R(\boldsymbol{\theta})$ is not identically zero. For this purpose, we can ignore the valid parameter space $\Delta_{|\mathcal{X}|-1} \times \Delta_{|\mathcal{Y}|-1}$, because if $R^{-1}(0)$ has zero Lebesgue measure in $\mathbb{R}^{|\mathcal{X}|+|\mathcal{Y}|}$, so does the intersection $R^{-1}(0) \cap (\Delta_{|\mathcal{X}|-1} \times \Delta_{|\mathcal{Y}|-1})$. We can also ignore the condition $\boldsymbol{\gamma} \neq (1, ..., 1)/|\mathcal{Y}|$, since this is a single point, thus a set of zero measure.

A sufficient and necessary condition for $R(\boldsymbol{\theta})$ not to be identically zero is that none of its factors $Q_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ is identically zero[6]. To show that no $Q_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ is identically zero, let us write it explicitly, using the definitions of $a_{y,x}$ and A_y in (18):

$$Q_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \sum_{y \in \mathcal{Y} \setminus \{1\}} \sum_{x \in \mathcal{X}} \Big(\beta_x \gamma_{\sigma_x(1)} \sum_{x' \in \mathcal{X}} \beta_{x'} \gamma_{\sigma_{x'}(y)} - \beta_{\rho_y(x)} \gamma_{\sigma_{\rho_y(x)}(y)} \sum_{x' \in \mathcal{X}} \beta_{x'} \gamma_{\sigma_{x'}(1)} \Big)^2. \tag{21}$$

Since $Q_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ is a sum of non-negative terms, to show that it is not identically zero, it suffices to show that one of the terms in the sum is strictly positive for some choice of $\boldsymbol{\theta}$. The condition $\boldsymbol{\rho} = (\rho_2, ..., \rho_{|\mathcal{Y}|}) \in \mathbb{L}$ means that at least one of the permutations $\rho_2, ..., \rho_{|\mathcal{Y}|}$ is not the identity, which implies that there is at least one pair (x, y) such that $\rho_y(x) \neq x$. Let y and x be one such pair. Choosing $\boldsymbol{\gamma} = (1, ..., 1)$ and $\boldsymbol{\beta} \in \Delta_{|\mathcal{X}|-1}$ such that all components are different from each other ($i \neq j \; \Rightarrow \beta_i \neq \beta_j$), we have (noticing that $\sum_{x' \in \mathcal{X}} \beta_{x'} = 1$)

$$Q_{\boldsymbol{\rho}}(\boldsymbol{\theta}) = \Big(\beta_x \sum_{x' \in \mathcal{X}} \beta_{x'} - \beta_{\rho_y(x)} \sum_{x' \in \mathcal{X}} \beta_{x'} \Big)^2 + \sum_{y' \neq y} \sum_{x' \neq x} (\cdots)^2 \tag{22}$$

$$= \big(\beta_x - \beta_{\rho_y(x)} \big)^2 + \text{ non-negative terms} > 0. \tag{23}$$

In conclusion, since none of the $Q_{\boldsymbol{\rho}}(\boldsymbol{\theta})$ polynomials is identically zero, $R(\boldsymbol{\theta})$ is also not identically zero, consequently its zero set has zero Lebesgue measure. ■

Appendix B. Proof of Proposition 10

Proof: Noticing that the permutations that map $\boldsymbol{\gamma}$ to each row of $\boldsymbol{\theta}^{Y|X}$ are arbitrary, there is no loss of generality in assuming $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_{|\mathcal{Y}|}$, *i.e.*, $\boldsymbol{\gamma} \in \mathcal{K}_{|\mathcal{Y}|}$, the so-called *monotone cone* (Best and Chakravarti, 1990). Furthermore, it is more convenient to formulate the problem w.r.t. the inverse permutations, denoted as $\tau_1, \ldots, \tau_{|\mathcal{X}|} \in \mathbb{S}_{|\mathcal{Y}|}$. The problem can thus be written as

$$\hat{\boldsymbol{\gamma}}, \hat{\tau}_1, \ldots, \hat{\tau}_{|\mathcal{X}|} \; = \underset{\substack{\boldsymbol{\gamma} \in (\Delta_{|\mathcal{Y}|-1} \cap \mathcal{K}_{|\mathcal{Y}|}) \\ \tau_1, ..., \tau_{|\mathcal{X}|} \in \mathbb{S}_{|\mathcal{Y}|}}}{\text{arg max}} \quad \mathcal{L}(\boldsymbol{\gamma}, \tau_1, ..., \tau_{|\mathcal{X}|}), \tag{24}$$

6. Recall that a product of two polynomials with real coefficients is identically zero only if at least one of the factors is identically zero. This is a classical result from abstract algebra, which in the language thereof is stated as follows: the ring of all polynomials in n variables with real coefficients is an *integral domain* or *entire ring*, that is, it does not have divisors of zero (Lang, 2002). The result generalizes trivially, by induction, to products of more than two polynomials.

where

$$\mathcal{L}(\boldsymbol{\gamma}, \tau_1, ..., \tau_{|\mathcal{X}|}) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} N_{x,\tau_x(y)} \log \gamma_y. \tag{25}$$

The assumption $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_{|\mathcal{Y}|}$ makes the maximization w.r.t. $\tau_1, ..., \tau_{|\mathcal{X}|}$ independent of the particular values of $\boldsymbol{\gamma}$ as well as separable into a collection of $|\mathcal{X}|$ independent maximizations,

$$\hat{\tau}_x = \underset{\tau_x \in \mathbb{S}_{|\mathcal{Y}|}}{\arg\max} \sum_{y \in \mathcal{Y}} N_{x,\tau_x(y)} \log \gamma_y, \tag{26}$$

for $x \in \mathcal{X}$. Solving (26) is a simple application of the *rearrangement inequality*[7] (Hardy et al., 1952). Since $\log \gamma_1 \geq \log \gamma_2 \geq \cdots \geq \log \gamma_{|\mathcal{Y}|}$, the solution $\hat{\tau}_x$ is any permutation that also sorts $\{N_{x,1}, ..., N_{x,|\mathcal{Y}|}\}$ into non-increasing order:

$$\hat{\tau}_x \text{ is such that } N_{x,\hat{\tau}_x(1)} \geq \cdots \geq N_{x,\hat{\tau}_x(|\mathcal{Y}|)}. \tag{27}$$

If all the elements of $\{N_{x,1}, ..., N_{x,|\mathcal{Y}|}\}$ are different, the optimal permutation is unique; otherwise, there are several optimal permutations, all achieving the same maximum. The cost of finding $\hat{\tau}_1, \ldots, \hat{\tau}_{|\mathcal{X}|}$ is $O(|\mathcal{X}|\,|\mathcal{Y}|\,\log|\mathcal{Y}|)$, since it requires $|\mathcal{X}|$ sorting operations, each with $|\mathcal{Y}|$ elements.

Plugging $\hat{\tau}_1, \ldots, \hat{\tau}_{|\mathcal{X}|}$ back into (9)–(10) and swapping the summation order, yields

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in (\Delta_{|\mathcal{Y}|-1} \cap \mathcal{K}_{|\mathcal{Y}|})}{\arg\max} \sum_{y \in \mathcal{Y}} \log \gamma_y \sum_{x \in \mathcal{X}} N_{x,\hat{\tau}_x(y)}. \tag{28}$$

This problem is the same as (7), with $\hat{\tau}_x$ in the place of τ_x and with the additional constraint $\boldsymbol{\gamma} \in \mathcal{K}_{|\mathcal{Y}|}$. Temporarily ignoring this constraint leads to (see (8)),

$$\hat{\gamma}_y = \frac{1}{N} \sum_{x \in \mathcal{X}} N_{x,\hat{\tau}_x(y)}, \quad \text{for } y \in \mathcal{Y}. \tag{29}$$

The fact that $N_{x,\hat{\tau}_x(1)} \geq N_{x,\hat{\tau}_x(2)} \geq \cdots \geq N_{x,\hat{\tau}_x(|\mathcal{Y}|)}$, for any $x \in \mathcal{X}$, implies that $\hat{\boldsymbol{\gamma}} \in \mathcal{K}_{|\mathcal{Y}|}$, without having to include this constraint. Consequently, problem (9)–(10) has a global solution given by (29) and (27). ■

Appendix C. Detailed Description of the Datasets Used in Section 5.3

Adult - This dataset consists of 48832 records from the census database of the US in 1994. We consider the following pairs: *(occupation, income)* and *(work class,income)*. The variable *occupation* takes values in the set {*admin, armed-force, blue-collar, white-collar, service, sales, professional, other-occupation*}. The variable *work class* takes categories in {*private, self-employed, public servant, unemployed*}. Finally, *income* is a binary variable taking value in $\{> 50, \leq 50\}$. Following

7. Given any two non-decreasing sequences of n real numbers, $x_1 \leq \ldots \leq x_n$ and $y_1 \leq \ldots \leq y_n$,

$$\forall \sigma \in \mathbb{S}_n, \quad \sum_{i=1}^{n} x_{\sigma(i)} y_i \leq \sum_{i=1}^{n} x_i y_i.$$

Budhathoki and Vreeken (2017b), we assume that the ground truth is *occupation* $\rightarrow$ *income* and *work class* $\rightarrow$ *income*.

Pittsburgh Bridges - This dataset contains records about 108 bridges and some of their characteristics. We consider the pairs (*purpose*, *type*) and (*material*,*lanes*). The variable *purpose* takes values in $\{$*Walk, Aqueduct, RR, Highway*$\}$; variable *type* takes values in $\{$*Wood, Suspen, Simple-T, Arch, Cantilev, CONT-T*$\}$; variable *material* takes values in $\{$*Steel, Iron, Wood*$\}$; variable *lanes* takes values in $\{1, 2, 4, 6\}$. Following Cai et al. (2018), the ground truths assumed are *material* $\rightarrow$ *lanes* and *purpose* $\rightarrow$ *type*.

Acute Inflammations - This dataset consists of 120 patients and whether each patient is experiencing a specific symptom, the temperature, and whether he/she suffers from acute inflammations of the urinary bladder and/or acute nephritis. We consider the binary variables *ocurrence of nausea* (Y_1), *lumbar pain* (Y_2), and *burning of urethra* (Y_3) (naturally, taking value 1 if the patient has that symptom, and 0 otherwise). Moreover, X represent the diagnosis *inflammation of urinary bladder*. Following Peters et al. (2011), the goal is to model the diagnosis process, thus we expect $Y_j \rightarrow X$, for $j = 1, ..., 3$. Notice that the variables X_i only correspond to the diagnosis, not necessarily the truth, otherwise they would be considered the cause, rather than the effect.

The χ^2 test did not reject the null hypothesis of independence (at a significance level of 5%) for the following pairs of variables in this dataset: (*Inflammation of urinary bladder*, *Occurrence of nausea*); (*Inflammation of urinary bladder*, *Burning of urethra*); (*Nephritis*, *Micturition pains*). Although, intuitively, we would expected a causal relation between those, the statistical evidence is not strong enough in favour of their mutual dependency (according to the χ^2 test), therefore, these pairs were discarded from the experiments.

Temperature - This dataset consists of 9162 daily values of temperature measured in Furtwangen (Germany), with the variable *day of the year* taking integer values from 1 to 365 (or 366 for leap years) and *temperature* in $^\circ C$. Here, we aggregate days associated with each month and take *month* $\rightarrow$ *temperature* as the ground truth, as Peters et al. (2011). Notice that *month* assumes a cyclic structure.

Horse Colic - This dataset contains 368 medical records of horses. We study the causal relationship between the variable *abdomen status*, which takes the values in $\{$*Normal, Other, Firm feces in the large intestine, Distended small intestine, Distended large intestine*$\}$, and the binary variable *surgical lesion*, indicating whether the lesion was surgical or not. As Budhathoki and Vreeken (2018), we regard *abdomen status* $\rightarrow$ *surgical lesion* as ground truth.