VARIATION-BASED CAUSE EFFECT IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mining genuine mechanisms underlying the complex data generation process in real-world systems is a fundamental step in promoting interpretability of (and thus trust in) data-driven models. Therefore, we propose a variation-based cause effect identification (VCEI) framework for causal discovery in bivariate systems from a single observational setting. Our framework relies on the principle of independence of cause and mechanism (ICM) under the assumption of an existing acyclic causal link, and offers a practical realization of this principle. Principally, we artificially construct two settings in which the marginal distributions of one covariate, claimed to be the cause, are guaranteed to have non-negligible variations. This is achieved by re-weighting samples of the marginal so that the resultant distribution is notably distinct from this marginal according to some discrepancy measure. In the causal direction, such variations are expected to have no impact on the effect generation mechanism. Therefore, quantifying the impact of these variations on the conditionals reveals the genuine causal direction. Moreover, we formulate our approach in the kernel-based maximum mean discrepancy, lifting all constraints on the data types of cause and effect covariates, and rendering such artificial interventions a convex optimization problem. We provide a series of experiments on real and synthetic data showing that VCEI is, in principle, competitive to other cause effect identification frameworks.

1 INTRODUCTION

Building trust in our machine learning models requires that they extend beyond their current limits of learning associational patterns and correlations. We need to be able to use them in interacting with our surroundings, in taking action to change or improve our environment, or in querying them for hypothetical scenarios that requires transparency. Yet, their black-box characteristics constitute significant barriers to their wide-scale adoption in, e.g., safety-critical domain. Causal inference relies on genuine cause-effect relationships rather purely statistical associations, thus promoting our understanding of the underlying data generation process.

While inferring genuine causal relations (oftentimes termed *causal discovery*) is, in general, a challenging task, it is even more challenging in bivariate systems where many of the early methods (based on conditional independence tests Spirtes et al. (2000); Sun et al. (2007); Pearl (2009)) fall short. Moreover, bivariate causal discovery is a fundamental step in mining implicit asymmetries in larger structures. In bivariate systems, asymmetries in the functional relationship (e.g. causal relationships tend to be functionally simpler, more elementary, and easier to learn with limited-capacity models than purely associational ones) is an example of a characteristic permitting identifiability of causal structure from observational data.

Another example of such an asymmetry is the postulate of independent mechanisms, on which our framework relies. In this principle, it is assumed that causal relationships tend to decompose

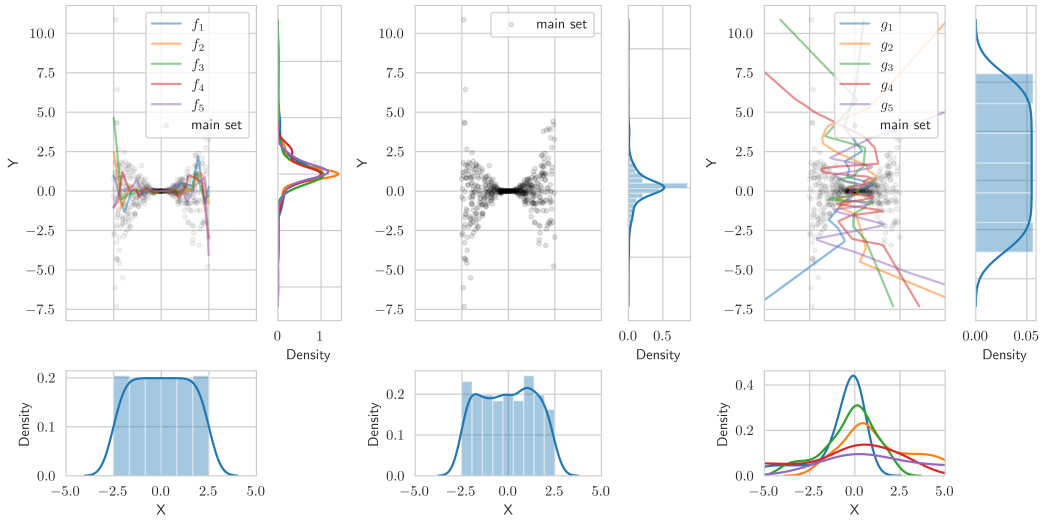

Figure 1: a toy example illustrating the asymmetry induced by the principle of independent mechanisms, and the effect of variations. The genuine data generation process is $y = -\frac{1}{2}x^2 * \epsilon$ with $x \sim \mathcal{U}[-2.5, 2.5]$ and the multiplicative noise ϵ from a standard normal distribution. A sample set from such a process are depicted in the scatter plot (middle). To naïvly introduce variations, we randomly draw a new set from the data generation process and train a model in the causal direction $f_i : x \mapsto f_i(x)$ (left), and similarly in the acausal direction $g_i : y : g_i(y)$ (right). The figure illustrates stability of the causal predictive models compared to acausal ones.

into invariant, stable sub-mechanisms. Such a principle has been the core asymmetry exploited in numerous bivariate causal discovery frameworks as shall be discussed in section 6. In this work, we exploit a barely explored interpretation of this principle, namely that these sub-mechanisms do not influence each other. To this end, we introduce variations to cause generation mechanism and quantify the influence on the effect generation mechanism.

Introducing variations to an empirical distribution can be as naïve as drawing random subsets. While this is not guaranteed to introduce non-negligible variations, fig. 1 shows a crafted toy setup that illustrates the effect of these variations on the effect generation mechanism, and the asymmetry revealed as a result.

While several previous works relied on this principle for causal discovery in bivariate systems, they either impose strict constraints on the data types (e.g. continuous data in regression-based approaches or identical data spaces for cause and effect), tend to show high sensitivity to the capacity of the chosen model class, or suffer from prohibitive computational complexities that renders them practically applicable only to certain (e.g., binary) data types. In this current work, we address these limitation, propose a new cause-effect identification framework based on artificially generated variations. The choice of the discrepancy measure along with the kernel embedding of the marginal distributions renders our framework applicable for a variety of data types[1] (e.g., timeseries data) and offers a practical realization leveraging convex optimization tools.

The problem we address in this paper is identifying the causal structure of a bivariate system from a single observational setting. To that end, we proposed a two-step variation-based causal discovery approach relying on convex optimization to introduce non-negligible variations and the kernel-based MMD metric to quantify the impact of these variations. Our contribution can thus be summarized as:

1. We introduce a new frame work bivariate causal discovery from observational settings.
2. We propose a kernel-based method that is independent of the data types used.
3. The framework entails an optimization problem that has been cast as a convex optimization problem.

[1]That is, within the identifiability limitations of the ICM postulate as shall be discussed in section 3.4.

2 PRELIMINARIES

Assumptions: We will consider a bivariate system (x, y) for cause-effect inference from an observational setting. In such a system, we assume acyclicity and the existence of a causal link (i.e. either $x \to y$ or $y \to x$). We additionally assume *causal sufficiency* in the sense that all relevant covariates are observed.

Independence of Causal Mechanisms (ICM): Our identification framework relies principally on the ICM concept Sgouritsa et al. (2015); Peters et al. (2017) which postulates that the genuine data generation process decomposes into *independent* modules that neither inform nor influence each other. Such independence will not necessarily (and is in practice less likely to) hold in acausal decompositions. In a bivariate causal graph $x \to y$ with a joint distribution p_{xy}, ICM implies *independence* between the marginal p_x and the conditional $p_{y|x}$, and shall be henceforth denoted by $p_{y|x} \perp p_x$. ICM induces an asymmetry in bivariate systems that has been leveraged in several causal inference approaches Mooij et al. (2009); Janzing & Schölkopf (2010); Stegle et al. (2010); Janzing et al. (2012); Daniusis et al. (2012); Schölkopf et al. (2012); Sgouritsa et al. (2015); Kocaoglu et al. (2017); Marx & Vreeken (2017); Tagasovska et al. (2018); Blöbaum et al. (2018); Budhathoki & Vreeken (2018); Marx & Vreeken (2021). Janzing & Schölkopf (2010) formulated this notion of *independence* in terms of Kolmogorov complexities Kolmogorov (1968) of the constituent distributions. Many works thereafter relied on the minimum description length (MDL) Rissanen (1978) as a proxy for the intractable Kolmogorov complexity Budhathoki & Vreeken (2017; 2018); Marx & Vreeken (2018); Mitrovic et al. (2018); Tagasovska et al. (2018); Kalainathan (2019); Marx & Vreeken (2019).

Maximum Mean Discrepancy (MMD): For analytical tractability, we will mainly consider kernel-based MMD as a metric of disparity between distributions Gretton et al. (2008; 2012). Given a kernel k, the MMD can be expressed as norm in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ between the kernel embeddings of the distributions p and q:

$$\text{MMD}_k^2(p, q) = \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \tag{1}$$

where μ_p and μ_q are the mean embeddings of p and q, respectively, in the Hilbert space $\mathcal{H}$ through the feature mapping $k(x, \cdot)$. From a practical perspective, squared MMD has an analytically tractable empirical estimator of a quadratic form given by:

$$\text{MMD}_k^2(p, q) \simeq \frac{1}{N^2} \sum_{i,j=1}^{N} k(x_i, x_j) - \frac{2}{NM} \sum_{i,j=1}^{N,M} k(x_i, y_j) + \frac{1}{M^2} \sum_{i,j=1}^{M} k(y_i, y_j) \tag{2}$$

with $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^M$ being finite sample sets drawn from p and q, respectively Sriperumbudur et al. (2009); Gretton et al. (2012). This efficient estimator renders MMD practically appealing for various applications amongst which is causal discovery Goudet et al. (2017); Baumann et al. (2020); Gao et al. (2021).

3 VARIATION-BASED CAUSE EFFECT IDENTIFICATION

In this section, we introduce our variation-based cause effect identification (VCEI) framework, a two-step procedure performed at least once in each direction of a bivariate system to infer the genuine causal structure from a single observational setting. Hypothesizing that the underlying causal structure is $x \to y$, the first step of VCEI is to introduce artificial variations to the marginal distribution p_x (see section 3.1). In the second step, we quantify the impact of these variations on the conditional $p_{y|x}$ (see section 3.2). According to the ICM postulate, variations on p_x are expected to have minimal impact on the conditional $p_{y|x}$ in the genuine causal direction.

Notation: let $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ denote a set of N i.i.d samples passively obtained, i.e. in an observational setting p_{xy}, from a bivariate system, where $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ are two random variables following the marginals p_x and p_y, respectively. Let further $\mathcal{D}_x = \{x_n \,|\, (x_n, y_n) \in \mathcal{D}\}$ denote the x-covariate view of the dataset, and likewise for $\mathcal{D}_y$.

3.1 ARTIFICIALLY GENERATED EXPERIMENTAL SETUPS

In this step, we propose an approach to introduce variations to the marginal distributions. For simplicity though, we will describe our approach for the first random variable x, but it should be clear that this step takes place once for each covariate. It should also be noted that such variations are intended to reveal potential dependencies between the marginal and the corresponding conditional, and do not necessarily retain similar dynamics to an *intervention*.

Given $\mathcal{D}_x$ with their unknown marginal p_x, we define the *empirical distribution* on these samples to be the uniform mixture of the Dirac delta distributions δ_{x_n} defined on each sample individually:

$$p_{x,N}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta(x - x_n) = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}(x) \tag{3}$$

which is a probability density function with the corresponding empirical cumulative distribution function $F_{x,N}(x)$ (eCDF) defined on the sample set as $F_{x,N}(x) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{x_n \leq x}$ where $\mathbb{1}_{(\cdot)}$ is the indicator function and the inequality is to be understood entry-wise (Scott, 1992). A generalization of the empirical distribution is a weighed mixture of the constituent Dirac distributions δ_{x_n} which we will denote by $p^{\boldsymbol{\alpha}}_{x,N}$ and define as (see appendix A for a brief discussion on this modelling choice):

$$p^{\boldsymbol{\alpha}}_{x,N}(x) = \sum_{n=1}^{N}\alpha_n\delta_{x_n} \tag{4}$$

where $\boldsymbol{\alpha} = [\alpha_n]_{n=1}^{N} \in [0,1]^{N\times 1}$ is a non-negative weight vector satisfying $\mathbf{1}^\top\boldsymbol{\alpha} = 1$ where $\mathbf{1}$ is the all-ones vector. From eq. (2), the MMD between the empirical distribution $p_{x,N}$ and the weighted version thereof $p^{\boldsymbol{\alpha}}_{x,N}$ becomes:

$$\mathrm{MMD}^2_k(p^{\boldsymbol{\alpha}}_{x,N}, p_{x,N}) \simeq \boldsymbol{\alpha}^\top\mathbf{K}_{xx}\boldsymbol{\alpha} - \frac{2}{N}\boldsymbol{\alpha}^\top\mathbf{K}_{xx}\mathbf{1} + \frac{1}{N^2}\mathbf{1}^\top\mathbf{K}_{xx}\mathbf{1} \tag{5}$$

where $\mathbf{K}_{xx} = [k(x_i, x_j)]_{i,j=1}^{N}$ is the Gram matrix of the kernel k on the sample set $\mathcal{D}_x$.

With this defined, and with the objective of introducing a non-negligible variation to the marginal of x, we are interested in solving the following problem:

Problem 1 *Given a set of samples $\{x_n\}_{n=1}^{N}$, find the weight vector $\boldsymbol{\alpha}$ that renders the mixture distribution $p^{\boldsymbol{\alpha}}_{x,N}$ maximally distinct from $p_{x,N}$ in some discrepancy measure $D(\cdot,\cdot)$.*

For analytical tractability, we will mainly consider the (MMD) metric[2] w.r.t a positive definite kernel function $k_{\mathbb{X}} : \mathbb{X}^2 \to \mathbb{R}$. By adopting a kernel-based approach, we mask the data space (in the sense that data space, along with its type and dimensionality, is subsumed in the kernel design/function) with an appropriately chosen kernel $k_{\mathbb{X}}$ function rendering our VCEI framework widely applicable to various data types[3] as opposed to e.g., regression-based identification frameworks.

Based on the squared MMD as a discrepancy measure, Problem 1 can be formally stated as:

$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad \mathrm{MMD}^2_{k_{\mathbb{X}}}(p^{\boldsymbol{\alpha}}_{x,N},\, p_{x,N}) \tag{6}$$

$$\text{subject to} \quad \mathbf{1}^\top\boldsymbol{\alpha} = 1 \tag{7}$$

$$\boldsymbol{\alpha} \geqslant 0 \;\; \text{(entry-wise)} \tag{8}$$

In spite of convexity of the objective (since MMD is jointly convex in both arguments as can be deduced from eq. (1)) and linearity of both constraints, the optimization problem remains non-convex. This is due the fact that the convex objective is being maximized rather than minimized which renders the objective a concave function in the standard form of a convex optimization problem.

[2]While we introduce our framework based on the MMD metric, similar relaxations or heuristics Park & Boyd (2017) can be applied to render Problem 1 a convex optimization problem for other discrepancy measures. This is, however, outside the scope of this contribution.

[3]For instance, in inferring summary graphs of temporal data using a timeseries kernel, or an embedding+kernel design for e.g. natural languages.

Noting that the closed-form estimator of the squared MMD is also quadratic in the optimization variable $\boldsymbol{\alpha}$ (see eq. (5)), Park & Boyd (2017) address this problem in a two-step procedure referred to as semidefinite relaxation (SDR). They first *lift* the problem to a higher dimensional space by defining $\mathbf{A} = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$ in which the objective function becomes linear, then apply a convex *relaxation* to the intractable constraints. As a result, the following formulation is a relaxation of 6–8 (see appendix Bfor a derivation) which is a quadratically constraint quadratic program (QCQP) that can make use of off-the-shelf convex optimization tools[4]:

$$\underset{\mathbf{A}}{\text{maximize}} \quad \mathbf{A} \bullet \left(\mathbf{K}_{xx} - \frac{2}{N}\mathbf{K}_{xx}\mathbf{1}\mathbf{1}^\top \right) + \frac{1}{N^2}\mathbf{1}^\top \mathbf{K}_{xx}\mathbf{1} \tag{9}$$

$$\text{subject to} \quad \begin{bmatrix} \mathbf{A} & \mathbf{A}\mathbf{1} \\ \mathbf{1}^\top\mathbf{A} & 1 \end{bmatrix} \succeq 0 \quad \text{(positive semidefiniteness)} \tag{10}$$

$$\mathbf{A} \geqslant 0 \qquad \text{(entry-wise)} \tag{11}$$

$$\mathbf{1}^\top\mathbf{A}\mathbf{1} = 1 \tag{12}$$

$$\mathbf{A} = \mathbf{A}^\top \tag{13}$$

where $\mathbf{K}_{xx} = [k_{\mathbb{X}}(x, \tilde{x})]_{x,\tilde{x}\in\mathcal{D}_x}$ is the Gram matrix, and $\bullet$ denotes the dot-product in matrix space defined as $\mathbf{A} \bullet \mathbf{K}_{xx} = \mathbf{trace}(\mathbf{A}\mathbf{K}_{xx})$.

The solution $\mathbf{A}^{\text{SDR}}$ to 9–13 is an optimal solution to the original formulation $\mathbf{A}^\star$ 6-8 (i.e. $\mathbf{A}^{\text{SDR}} \equiv \mathbf{A}^\star$) if the condition $\mathbf{A}^\star = \boldsymbol{\alpha}^\star\boldsymbol{\alpha}^{\star\top}$ is satisfied (i.e. if $\mathbf{A}^{\text{SDR}}$ is rank one which will be the case if $\mathbf{A}^{\text{SDR}}$ is a feasible solution to 6-8 Park & Boyd (2017))[5]. In this case, the distribution weights can be recovered as $\boldsymbol{\alpha}^\star = \mathbf{A}^\star\mathbf{1}$.

With the solution to Problem 1, we would have obtained a new marginal $p^{\boldsymbol{\alpha}^\star}_{x,N}$ that is constructed from the passively obtained observational data $\mathcal{D}_x$ and is maximally distinct from the original marginal p_x. Finally, this optimization is performed on the second covariate y to obtain a weighted marginal $p^{\boldsymbol{\beta}}_{y,N}$ with weight vector $\boldsymbol{\beta} \in [0, 1]^{N\times 1}$ that is maximally distinct from $p_{y,N}$.

3.2 Quantifying the Impact of Distributional Variations

In the second step, we quantify the impact of the artificially generated variations (i.e. within the marginals $p_{x,N}$ and $p^{\boldsymbol{\alpha}}_{x,N}$ and similarly from $p_{y,N}$ to $p^{\boldsymbol{\beta}}_{y,N}$) on the conditionals $p_{x|y}$ and $p_{y|x}$, respectively. This can be achieved by fitting predictive models to each of these settings leading to the two models $\hat{f}_{y|x}$ and $\hat{f}^{\boldsymbol{\alpha}}_{y|x}$ in the $x \to y$ direction, and $\hat{g}_{x|y}$ and $\hat{g}^{\boldsymbol{\beta}}_{x|y}$ in the opposite direction. Each model is attainable from a model class $\mathcal{M}_{x\to y}$ or $\mathcal{M}_{y\to y}$ with their corresponding training paradigms $\text{Train}_{\mathcal{M}_{x\to y}}[\cdot]$ and $\text{Train}_{\mathcal{M}_{y\to x}}[\cdot]$.

In order to fit a predictive model on a weighted empirical distribution e.g. $p^{\boldsymbol{\alpha}}_{x,N}$, the corresponding weights can be considered sample weights and the training paradigms $\text{Train}_{\mathcal{M}_\cdot}[\cdot]$ supports sample importance[6] (see, for example, Wen et al. (2018) for a weighted Gaussian Process (GP) model or Steininger et al. (2021) for neural networks).

ICM postulates that, if $x \to y$ is the true causal direction of the data generation process, then the impact of the introduced variations on the $\hat{g}$ models are likely to be more apparent. We quantify this impact via model disagreement on a (potentially unlabeld) set Nakkiran & Bansal (2020), which is in turn quantified as the MMD discrepancy between each model's prediction on a common set:

$$S_{x\to y} = \text{MMD}^2_{k_{\mathbb{Y}}}\left(\hat{f}_{y|x}(x), \hat{f}^{\boldsymbol{\alpha}}_{y|x}(x) \right) \tag{14}$$

where $x \sim p_x(x)$ (which empirically could simply be all samples in $\mathcal{D}_x$ or a random subset thereof) and similarly for $S_{y\to x}$. Finally, the lower of either scores[7] $S_{x\to y}$ and $S_{y\to x}$ is an indicator of a lesser impact on the conditionals, and in turn the genuine causal direction. An overview of the VCEI framework for identical data spaces is presented in algorithm 1.

[4]For instance, we used the open-source library `cvxpy` Diamond & Boyd (2016) for all experiments.

[5]In section 3.3, we discuss situations in which $\mathbf{A}^{\text{SDR}}$ is not a rank one matrix.

[6]Alternatively, model fitting can be preceded by a re-sampling step.

[7]Similarly, see section 3.3 for a discussion the implicit assumptions this decision criterion entails.

Algorithm 1 Variation-based cause-effect identification (VCEI) on identical data spaces $\mathbb{X} \equiv \mathbb{Y}$

Require: $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, a kernel function k, model classes $\mathcal{M}_{x\to y}$, $\mathcal{M}_{y\to x}$, corresponding training paradigms $\text{Train}_{\mathcal{M}_{x\to y}}[\cdot]$ and $\text{Train}_{\mathcal{M}_{y\to x}}[\cdot]$, and a regularization parameter b_α.
Ensure: $\mathbb{X} \equiv \mathbb{Y}$ (where $x \in \mathbb{X}$ and $y \in \mathbb{Y}$)
Estimate $S_{x\to y}$**:** Solve SDR of Problem 1 (Equation (9)–13 and 15) in $\mathcal{D}_x$ to estimate $\boldsymbol{\alpha}$
$\hat{f}_{y|x} \leftarrow \text{Train}_{\mathcal{M}_{x\to y}}\left[p_{xy,N}\right]$
$\hat{f}^{\boldsymbol{\alpha}}_{y|x} \leftarrow \text{Train}_{\mathcal{M}_{x\to y}}\left[p^{\boldsymbol{\alpha}}_{xy,N}\right]$
$S_{x\to y} \leftarrow \text{MMD}^2_k\left(\hat{f}_{y|x}(p_{x,N}), \hat{f}^{\boldsymbol{\alpha}}_{y|x}(p_{x,N})\right)$
Estimate $S_{y\to x}$**:** Solve SDR of Problem 1 (Equation (9)–13 and 15) in $\mathcal{D}_y$ to estimate $\boldsymbol{\beta}$
$\hat{g}_{x|y} \leftarrow \text{Train}_{\mathcal{M}_{y\to x}}\left[p_{xy,N}\right]$
$\hat{g}^{\boldsymbol{\beta}}_{x|y} \leftarrow \text{Train}_{\mathcal{M}_{y\to x}}\left[p^{\boldsymbol{\beta}}_{xy,N}\right]$
$S_{y\to x} \leftarrow \text{MMD}^2_k\left(\hat{g}_{x|y}(p_{y,N}), \hat{g}^{\boldsymbol{\beta}}_{x|y}(p_{y,N})\right)$
Return: "$x \to y$" **if** $S_{x\to y} < S_{y\to y}$ **otherwise** "$y \to x$"

3.3 Practical Considerations

While Problem 1 tends to construct setups with maximal disparity from the given empirical distribution $p_{x,N}$, we are not necessarily interested in such extreme scenarios as long as these variations are non-negligible so that they reveal dependencies between the marginal and the conditional distributions in the acausal direction. Therefore, we would oftentimes prefer a sub-optimal, yet more appealing, solution to the optimal solution of Problem 1 for practical considerations. Such practical aspects are discussed in the sequel.

Scalability: one directly notes that the SDR formulation 9–13 hardly scales to larger datasets since the dimensionality of the optimization space is quadratic in the number of data points N (as a result of the lifting step). Therefore, we rather restrict the weighted distribution $p^{\boldsymbol{\alpha}}_{\cdot,N}$ to a reasonable number of samples $M < N$ drawn randomly from the original dataset. This is denoted henceforth by $p_{\cdot,M}$ for the M-sample subset and $p^{\tilde{\boldsymbol{\alpha}}}_{\cdot,M}$ for the weighted version thereof. The size of the reference empirical distribution $p_{\cdot,N}$ (2nd argument of eq. (5)) does not affect the dimensionality of the optimization problem and, thus, can grow as needed within the Gram matrix computational limits.

Dirac Distributions: an artifact of the choice of the discrepancy measure (and the formulation of problem 1) is that attainable solutions to 9–13 are in practice Dirac-like probability measures in the sense that $\|\boldsymbol{\alpha}\|_\infty \sim 1$ where $\|\cdot\|_\infty$ is the supremum norm. One can avoid such extreme scenarios by augmenting the optimization problem with regularizing constraints such as

$$\|\mathbf{A}\|_\infty \leqslant b_\alpha \tag{15}$$

with the supremum norm of a matrix given by $\|\mathbf{A}\|_\infty \coloneqq \max_i \|\mathbf{a}_{i\cdot}\|_1$ which directly constraints the maximum probability mass that is allowed on a single data point and $b_\alpha \in [1/M,\, 1.0]$ becomes a hyper-parameter in our framework. Figure 2 illustrates the effect of this regularization constraint on a 2D sample set drawn from a standard Normal distribution. Likewise, one can constrain maximum deviation from the uniform mixture as in

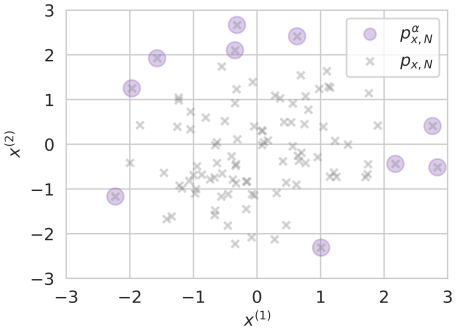

Figure 2: An illustrative example of solving problem 1 on a 2D Gaussian dataset. The true distribution is $p_x = \mathcal{N}(\mathbf{0}, \mathbf{1})$ from which $N = 100$ samples are depicted in grey. Purple markers represent the weights $\boldsymbol{\alpha}$ of the weighted distribution $p^{\boldsymbol{\alpha}}_{x,100}$.

$$\text{MMD}^2_k\left(p^{\tilde{\boldsymbol{\alpha}}}_{\cdot,M}, p_{\cdot,M}\right) \;\leqslant\; \text{MMD}^2\left(p_{\cdot,M}, p_{\cdot,N}\right) + b_D \tag{16}$$

where b_D is a slack variable, and the l.h.s is a linear function of the optimization variable $\mathbf{A}$ similar to Eq. 9 with a different Gram matrix. Given the convexity of both regularization constraints above,

eqs. (15) and (16), the SDR formulation 9–13 remains a convex optimization problem if augmented with either of these constraints.

SDR Relaxation: a solution $d_{\mathbb{X}}^{\text{sdr}}$ obtained from the SDR formulation is a lower bound on the optimal value of the original formulation 6–8 that is tight only if the rank one condition $\mathbf{A} = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$ is satisfied Park & Boyd (2017). Yet, the rank-one condition is not guaranteed, and is even unlikely to be satisfied as additional constraints (e.g., eqs. (15) and (16)) are included in the optimization problem. Practically, however, estimating the weight vector as $\boldsymbol{\alpha} \simeq \mathbf{A}^{\text{SDR}}\mathbf{1}$ remains a reasonable estimate for the weighted empirical that notably outperforms naive baselines (e.g. drawing random subsets).

Disagreement Bias: in the second step of our identification framework, we quantify disparity between two models (e.g., $\hat{f}_{y|x}$ and $\hat{f}^{\boldsymbol{\alpha}}_{y|x}$) via their MMD-based disagreement on a common input distribution. However, for some model classes (e.g., neural networks) such an approach is likely to be biased. In fact, it was observed recently that two identical neural network classifiers would disagree even when trained on identical data as long as a randomization factor plays a roll (i.e. different initial weights, batching, data shuffling, or different random seeds in general) Nakkiran & Bansal (2020); Jiang et al. (2021). In fact, it was conjectured that this sort of disagreement correlates with the generalization performance of the classifier.

Our empirical observations extend the claims of Nakkiran & Bansal (2020); Jiang et al. (2021) to regression problems with MMD as a disagreement metric. Since all our models are trained on limited data, they are likely to disagree (i.e. generalize poorly) even if the training distributions were identical. This *disagreement bias* is not accounted for in our work, and is left as an open question for future contribution. Figure 3 depicts an example of such a bias in the non-zero disagreement score $S_{y\to x}$ even though the genuine causal direction is indeed $y \to x$.

Trend as a Score: the final decision criterion, that is comparing the MMD-based disagreements scores, implicitly imposes a strong assumption of the data spaces $\mathbb{X} \equiv \mathbb{Y}$ and similarly on the kernels $k_{\mathbb{X}} \equiv k_{\mathbb{Y}}$ (admittedly, this has been an implicit assumption in numerous previous works e.g., roughly all approaches relying on regression performance). At the expense of additional computational demands, one can circumvent this limitation with the following observation. It is observed, and also intuitive, that the attainable solution to 9–13 augmented with 15 is monotonic in the hyper-parameter b_α (refer to appendix C for an illustrative example). According to ICM, repeating the optimization problem with increasing values for b_α is likely to be reflected in an increasing trend of the disagreement score of the acausal direction. In the causal direction, however, the disagreement score is expected to remain roughly constant.

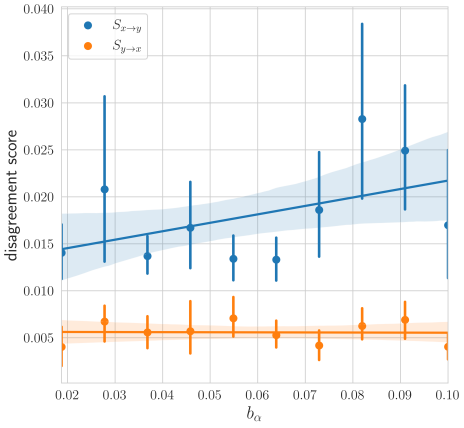

Figure 3: An illustration of the behaviour of the disagreement scores $S_{x\to y}$ (upper) and $S_{y\to x}$ (lower) for different values of the hyper-parameter b_α where the true causal structure is $y \to x$. Example from the 1st pair of the `SIM` dataset Mooij et al. (2016a).

Treating these disagreement score as functions of the regularization hyper-parameter (e.g., linearly regressing $S_\cdot$ on b_α for different solutions of the optimization problem) gives an alternative decision criterion (e.g., trend of these regression lines) that is independent of the data spaces, kernels, and kernel hyper-parameters. This is briefly illustrated in fig. 3 (and a similar effect can be observed w.r.t the number of samples M), but is not thoroughly investigated in this work, and is rather left as another open point for future contribution. Interestingly, and also left open for future work, using this decision mechanism may also mitigate the causal sufficiency assumption leading to broader identifiability.

3.4 Identifiability

The proposed VCEI framework is viewed as a practical realization of the ICM principle, and thus, inherits all identifiability limitations of that postulate. When viewed from e.g., Kolmogorov complexities $K(p_x) + K(p_{y|x}) \leq K(p_y) + K(p_{x|y})$ if $x \to y$ as formulated by Janzing & Schölkopf

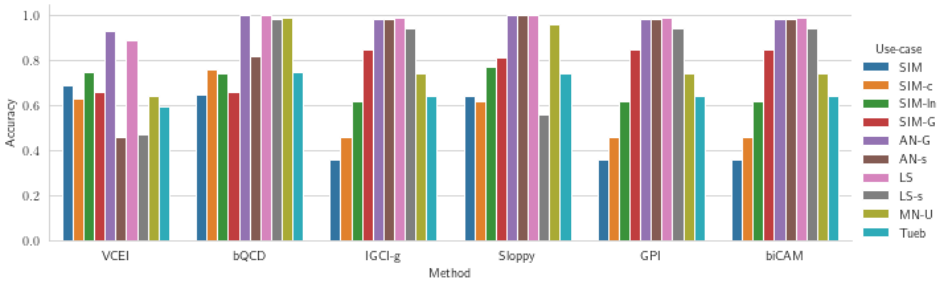

Figure 4: Accuracy of VCEI on benchmark datasets compared to baseline methods. Identification accuracies for baseline methods were taken from Tagasovska et al. (2018). For `SIM-`$\star$ the sample size was $M = 200$, while for the remaining datasets the sample was limited to $M = 100$.

(2010), one directly notes a limitation of ICM-based frameworks, that is when equality occurs and thus the ICM-based asymmetry vanishes. Asymmetry vanishes if the underlying system can be described with the same functional form and distributional families in either direction (Mitrovic et al., 2018). A very common example thereof are linear models with additive Gaussian noise(Hoyer et al., 2008). Loosely speaking, identifiability of ICM-based frameworks increase with increasing non-linearity of the functional form, smaller noise effects (Mooij et al., 2016b), and less (or no) confounding bias .

In addition, and as stated earlier, we assume existence of a causal link and causal sufficiency. The former, however, can be mitigated with either an independence test. The latter can also be mitigated with the use of disagreement trends rather than single scores (as discussed in the preceding subsection) where confounding may lead to a positive trend in either direction, but it is expected be more observable (i.e. steeper) in the acausal direction.

4 Experimental Validation

In the sequel, we report empirical validation of our proposed method. For a benchmark, we tested VCEI on the same use-cases presented in the work of Tagasovska et al. (2018).

Simulated data: simulation data[8] were originally generated in the work of Mooij et al. (2016b). Four different scenarios were considered: `SIM` which is the default use-case without confounder-bias, `SIM-c` which includes a single latent confounder, `SIM-ln` a use-case with low noise levels, and finally `SIM-G` which has a Gaussian-like distribution for both the cause X and the additive noise. We additionally, included the 5 additional synthetic datasets published by Tagasovska et al. (2018) (namely `AN(-s)`, `LS(-s)`, and `MN-U`) (see appendix C.1 for a more detailed description).

Real-world data: the Tübingen Cause-Effect (CE) benchmark was considered for real-data validation, which consists of 108 pairs from 37 different domains. We only used 103 pairs, which have univariate (continuous or discrete) cause and effect variables.

Baselines: we included a selected set of the methods from the baselines reported in the work of Tagasovska et al. (2018). We namely compare our VCEI framework to: biCAM Bühlmann et al. (2014), which are additive noise model (ANM)-based, IGCI Janzing & Schölkopf (2010), bQCD Tagasovska et al. (2018)), Sloppy Marx & Vreeken (2019), and finally GPI Stegle et al. (2010).

Sample Size: due to the limited scalability of the proposed framework (and the limited computational budget), the number of samples used in the optimization step to construct the different setting 3.1, and latter for training of the predictive models, was chosen to be relatively low.

Figure 4 depicts the identification accuracies of our method on the selected benchmark datasets, and compared to other causal discovery baseline algorithms. We use the same metric as in Mooij et al. (2016b) namely, *accuracy for forced decisions*. In principle, each algorithm is forced to take a decision about the causal direction from which the identification accuracy corresponds to the how frequent the algorithm was able to reach correct decisions over the number of dataset files.

[8]All synthetic dataset have been obtained from: https://github.com/tagas/bQCD

While our framework does not show an unprecedented performance on the benchmark dataset, it is certainly competitive to many previous methods, in addition to being generic w.r.t data types, and robust w.r.t choice of model class and the learning capacity thereof. We refer to appendix C.3 for further detailled analysis of our framework.

5 CONCLUSION

We introduce Variation-based Cause Effect Identification (VCEI), a kernel-based framework for causal discovery in a bivariate systems. Our method combines the principle of independent causal mechanism (ICM) with convex-optimization under semi-definite relaxation (SDR) and the learning power of data-driven models to identify the genuine causal structure of a bivariate system. With the kernel-based scores, we impose only mild assumptions on the the data types, thus giving the advantage of its implementation for a wide range of applications. Additionally, our framework is robust to the model capacity as long as it is capacitive enough to learn variations of conditionals.

6 RELATED WORK

In this section, we briefly review relevant work on causal discovery in bivariate systems. The intent is not to provide an extensive review (for which the interested reader is referred to e.g. Mooij et al. (2016b) specifically for cause-effect identification or Vowels et al. (2021) for a more recent review on causal discovery). Rather, we review works that notably share similarities and analogies to our proposed framework in order to highlight and emphasize our contributions.

Works on causal discovery started with conditional independence tests Spirtes et al. (2000); Sun et al. (2007); Pearl (2009) which fell short in bivariate cause-effect identification scenarios due to lack of conditioning covariates. Lines of work that addressed this problem postulated a sort of an inherent asymmetry in the cause-effect relationship. An example of such is the functional and distributional asymmetries proposed by the early works in this direction Shimizu et al. (2006); Hoyer et al. (2008); Mooij et al. (2009); Zhang & Hyvarinen (2012). Contrary to these frameworks, our proposed approach does not impose functional or distributional constraints on the causal relationship.

A different aspect of asymmetry is the ICM postulate on which numerous cause-effect identification frameworks have relied Sgouritsa et al. (2015); Mooij et al. (2009); Janzing & Schölkopf (2010); Stegle et al. (2010); Janzing et al. (2012); Daniusis et al. (2012); Schölkopf et al. (2012); Kocaoglu et al. (2017); Marx & Vreeken (2017); Tagasovska et al. (2018); Blöbaum et al. (2018); Budhathoki & Vreeken (2018); Marx & Vreeken (2021); Budhathoki & Vreeken (2017); Kalainathan (2019); Marx & Vreeken (2018; 2019); Mitrovic et al. (2018), mainly utilizing the MDL as a proxy in place of the intractable Kolmogorov complexities. Yet, most of the works are limited specific data spaces, e.g. numeric data for regression-based frameworks Sgouritsa et al. (2015); Mooij et al. (2009); Tagasovska et al. (2018); Marx & Vreeken (2019). Notable exceptions are works relying on kernel-embeddings Mitrovic et al. (2018); Lopez-Paz et al. (2015) Likewise, our contribution lifts all constraints on the data spaces via the adopted kernel-based MMD metric (except for a mild assumption discussed in section 3.3) to the choice of a characteristic kernel.

Kernel-based MMD was utilized a loss function in Goudet et al. (2017) for learning bivariate causal structures. Their approach relies on the simplicity of the functional relationship in the causal direction, and thus can be identified with a model class of limited-capacity. The higher the model capacity, the less identifiable a causal structure would be to their model. In contrast, our framework is more robust to the model choice in the sense that it only requires a model class of a capacitive power to learn the functional relationship in either direction equally well.

Finally, and aside from bivariate systems, Peters et al. (2016) proposed a causal discovery framework in scenarios of multiple experimental setups (including an observational one) with random, unknown interventions. Yet, in the case of a single observational setup, they introduce conditional splitting of the dataset (under predefined conditions) to emulate an artificial scenario of multiple experimental setups. In spite of the distinction, their contribution was an inspiration for our proposed framework.

REFERENCES

Dominik Baumann, Friedrich Solowjow, Karl H Johansson, and Sebastian Trimpe. Identifying causal structure in dynamical systems. *arXiv preprint arXiv:2006.03906*, 2020.

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pp. 900–909. PMLR, 2018.

Kailash Budhathoki and Jilles Vreeken. Mdl for causal inference on discrete data. In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 751–756. IEEE, 2017.

Kailash Budhathoki and Jilles Vreeken. Origo: causal inference by compression. *Knowledge and Information Systems*, 56(2):285–307, 2018.

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

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*, 2012.

Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

Yinghua Gao, Li Shen, and Shu-Tao Xia. DAG-GAN: Causal Structure Learning with Generative Adversarial Nets. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3320–3324. IEEE, 2021.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Learning Functional Causal Models with Generative Neural Networks. *arXiv.org*, September 2017. doi: 10.1007/978-3-319-98131-4,10.48550/arXiv.1709.05321. URL `https://arxiv.org/abs/1709.05321v3`.

Arthur Gretton, Karsten Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368*, 2008.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

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.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*, 2021.

Diviyan Kalainathan. *Generative Neural Networks to infer Causal Mechanisms: algorithms and applications*. PhD thesis, Université Paris Saclay (COmUE), 2019.

Murat Kocaoglu, Alexandros G Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Andrei Nikolaevic Kolmogorov. Three approaches to the quantitative definition of information. *International journal of computer mathematics*, 2(1-4):157–168, 1968.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pp. 1452–1461. PMLR, 2015.

Alexander Marx and Jilles Vreeken. Telling cause from effect using mdl-based local and global regression. In *2017 IEEE international conference on data mining (ICDM)*, pp. 307–316. IEEE, 2017.

Alexander Marx and Jilles Vreeken. Causal inference on multivariate and mixed-type data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 655–671. Springer, 2018.

Alexander Marx and Jilles Vreeken. Identifiability of cause and effect using regularized regression. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 852–861, 2019.

Alexander Marx and Jilles Vreeken. Formally justifying mdl-based inference of cause and effect. *arXiv preprint arXiv:2105.01902*, 2021.

Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. *Advances in neural information processing systems*, 31, 2018.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, pp. 745–752, 2009.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *J. Mach. Learn. Res.*, 17(1):1103–1204, jan 2016a. ISSN 1532-4435.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016b.

Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020.

Jaehyun Park and Stephen Boyd. General heuristics for nonconvex quadratically constrained quadratic programming. *arXiv preprint arXiv:1703.07870*, 2017.

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

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. ISSN 1533-7928. URL `http://jmlr.org/papers/v15/peters14a.html`.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. A Wiley-interscience publication. Wiley, 1992. ISBN 9780471547709.

Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. Inference of Cause and Effect with Unsupervised Inverse Regression. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 847–855, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/sgouritsa15.html.

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.

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

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. *Advances in neural information processing systems*, 23, 2010.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Distinguishing between cause and effect via kernel-based complexity measures for conditional distributions. In *15th European Symposium on Artificial Neural Networks (ESANN 2007)*, pp. 441–446. D-Side Publications, 2007.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing Cause from Effect Using Quantiles: Bivariate Quantile Causal Discovery. *arXiv.org*, January 2018. doi: 10.48550/arXiv.1801.10579. URL https://arxiv.org/abs/1801.10579v4.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582*, 2021.

Junfeng Wen, Negar Hassanpour, and Russell Greiner. Weighted gaussian process for estimating treatment effect. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, 2018.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.