Meta-Dependence in Conditional Independence Testing

Abstract

Constraint-based causal discovery algorithms utilize many statistical tests for conditional independence to uncover networks of causal dependencies. These approaches to causal discovery rely on a correspondence between the graphical properties of a causal structure and the conditional independence properties of observed variables, formally defined as faithfulness assumptions. Finite data yields an empirical distribution that is "close" to the actual distribution. Across these many possible empirical distributions, the statistical dependencies between variables in a given model have a "meta-dependence". An alternative perspective is that tests for conditional independence have shared information. This paper links the two ideas by observing that the geometry of conditional independencies informs their co-occurrence with respect to local perturbations from the true distribution to the empirical one. We use information projections to measure this meta-dependence and show that they can be computed efficiently in practice. We consolidate our findings empirically using both synthetic and real-world data.

1 INTRODUCTION

Structural Casual Models (SCMs), popularized by Pearl [1998, 2009], describe data-generating processes tied to causal relationships. These models graphically describe causal networks with arrows denoting the flow of causality. SCM structures give rise to covariate adjustment formulae, which allow the computation of causal effects without access to "randomized control trials" — the causality gold standard formalized by Fisher et al. [1931].

Causal discovery is the task of recovering a causal structure from data, often in the form of a *directed acyclic graph* (DAG) or a representation of an equivalence class of DAGs, such as a "CP-DAG" (see Squires and Uhler [2023] for a review). One approach to causal discovery involves a *constraint-based* search guided by conditional independence ("CI-tests"), e.g., the PC-algorithm from Spirtes et al. [2000]. Constraint-based approaches make formal use of the observation that variables without a direct causal link can be made independent by conditioning on intermediary causal paths. This is known as the causal Markov condition. For example, a famous causal pathway in biology is the transcription of DNA to RNA, followed by RNA translation to proteins. This pathway leads to an association between DNA and proteins, *but only through* RNA. DNA information is independent from protein expression when conditioned on RNA, as RNA contains all the relevant information that links DNA and proteins.

The formulation of structural causal models implies the causal Markov condition, but the converse is not necessarily true. A causal pathway between two variables does not necessarily require that they be statistically dependent under all conditionings. Statistical dependence from causal linkage is called "faithfulness," which is often assumed in order to give a two-way correspondence between CI-test outcomes and graphical properties.

Assuming faithfulness is sometimes considered "mild" — the set of unfaithful distributions has Lebesgue measure 0. However, the study of Uhler et al. [2013] on linear structural equations and additive Gaussian noise showed that the geometry of unfaithful distributions is a sufficiently "bendy" manifold such that the majority of distributions are "close"[1] to an unfaithful distribution. This finding suggests that unfaithful distributions (with respect to a given DAG) are difficult to distinguish from faithful ones — uncertainty in the true distribution will likely always overlap with violations of faithfulness. As a result, verifying faithfulness

[1] A distribution is "close" to a violation of faithfulness if two causally connected variables are almost independent, e.g. they have small variance.

Submitted to the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025). **To be used for reviewing only**.

is data-intensive, requiring a sufficiently refined empirical distribution to ensure that the true distribution does not fall on a faithfulness violation. This could lead to instability in the statistics and an erroneous output graph.

1.1 MOTIVATION

Significance Thresholds. Causal discovery algorithms have many hyperparameters that must be tuned [Biza et al., 2020]. Constraint-based causal discovery exhaustively employs many CI tests, many of which correspond to the same action, i.e., removing an edge between the conditionally independent variables. If the tests have no shared information, we must lower the significance thresholds of these tests. For example, a p-value of $.05$ will reject the null 5% of the time under the null hypothesis. If these tests have shared information and co-occur, then we do not need to lower the p-value to avoid "p-value hacking". The "graphoid axioms" given in Pearl and Paz [2022] are one known example of this dependence, from which the "Verma constraints" [Verma and Pearl, 1990] are derived. Both of these constraints are graphical and therefore difficult to apply while *learning* an unknown causal DAG. There is currently no metric for quantifying the shared information between CI tests without knowledge of the DAG structure.

Validating Causal Discovery. To validate the success of causal discovery, we require a notion of distance between the true ($\mathcal{G}$) and recovered ($\hat{\mathcal{G}}$) graphical models. A popular choice is the *Structural Hamming Distance* (SHD) [de Jongh and Druzdzel, 2009], i.e., the number of edge removals and additions to transform $\hat{\mathcal{G}}$ into $\mathcal{G}$. Another choice introduced by Peters and Bühlmann [2015] uses interventions to measure causally relevant changes. This is a special case of "adjustment identification distance" [Henckel et al., 2024].

These graphical perspectives fail to consider the *co-occurrence* of edge errors driven by the statistical properties of the empirical distribution. For example, many errors may be driven by a small perturbation in the data, whereas a significant perturbation may drive only a single error. In this case, the single mistake may be as (or even more) severe than many co-occurring errors. Liu et al. [2010] suggested regularizing causal discovery by focusing on properties that are stable to subsampling. The dependencies between these instabilities have not yet been studied.

Uncertainty in Causal Discovery While we can quantify uncertainty in conditional independence testing [de Morais Andrade et al., 2014], it remains unclear how to propagate this uncertainty into a causal structure without subsampling as in Liu et al. [2010]. A "confidence ensemble" of causal structures relative to empirical uncertainty may not necessarily correspond to a neighborhood of structures that are close in SHD. In order to capture model uncertainty, we need to study the dependence between errors in CI tests.

1.2 CONTRIBUTIONS

This paper studies the shared information between different CI tests and links it to the geometry of conditional independence properties under local perturbations. Section 4 motivates the paper by observing that a perturbation on the parameters of a distribution may bring it closer to some faithfulness violations and farther from others.

To formally study this geometry with *unknown* models, Section 5 defines an "information projection" [Nielsen, 2018] from an empirical distribution onto a conditional independence property. We use KL divergence to study the distance between distributions and their projections, which simplifies to mutual information.

We use information projections to define Conditional Independence Meta-Dependence (CIMD), which measures the change in proximity to one conditional independence that results from projecting onto another conditional independence. CIMD represents shared information between two tests for conditional independence – projecting out components of a distribution that go against both conditional independencies moves those properties together into feasibility. Other components of the distribution can give rise to negative CIMD if they are in conflict with information that was in favor of the other property. To make these notions practical, Section 6 shows that our information projection can be computed using maximum likelihood estimators, and both information projections and CIMD have closed forms for multivariate Gaussian distributions.

Finite sample uncertainty perturbs the ground-truth distribution to an empirical one that is likely close in KL divergence. Other than this intuition, we provide no theoretical results directly linking CIMD to the dependence between CI-tests under finite sample uncertainty. Instead, we bootstrap perturbations by randomly subseting a dataset and keeping track of the co-occurrence of significant conditional independencies. We call this FS-CID and provide an empirical study linking CIMD to FS-CID in both synthetic and real-world data in Section 7.

2 RELATED WORKS

Causal Strength and Stability. Our work is related to the quantification of causal models and their stability. Previous works have attempted to quantify the strength of *causal* relationships, most notably the seminal work of Janzing et al. [2013] and more recent stability results in Schulman and Srivastava [2016]. Causal strength does not necessitate statistical dependence, such as when faithfulness is violated. These notions are developed to analyze the computation of causal quantities and rely heavily on *known causal structures*. This paper is motivated by causal discovery, which is applied in settings with *unknown causal structure*.

Complexity of DAGs. Embedded in the idea that CI-tests are dependent is the notion that some DAGs (or CP-DAGs) are "more difficult" to learn than others. This notion of "DAG complexity" is related to the amount of data needed to verify a structure with statistical significance. A relevant quantity is the number of CI queries required to confirm a structure, studied in Zhang et al. [2024]. This purely graphical quantity lacks consideration for the *dependencies* between these queries due to distribution specifics. Our work seeks to understand these dependencies, pushing toward an entropy-based DAG-complexity.

Separating Set Size. Independence queries that condition on many variables are highly data-intensive. For example, a separating set of k Bernoulli random variables requires 2^k independence queries, i.e., one for each possible assignment to the separating set. Kocaoglu [2023] explores the notion of Markov equivalence classes with restrictions on the size of conditioning. This is an important step towards a more coarse-grained grouping of causal structures that are difficult to statistically distinguish. This approach does not take into account potential dependencies arising from specific distributions.

Falsifying Causal Models. Falsification of SCMs was recently studied using "self-compatibility" with respect to structures on subsets of variables [Faller et al., 2024]. Self-compatibility posits that changes in the recovered structure on different subsets of the observed variables contain information about the stability of the resulting graphical properties. Self-compatibility is likely related to the information geometry of the empirical distribution, since the exclusion of some observables removes potentially redundant CI tests from constraint-based algorithms.

3 PRELIMINARIES

Notation. We will generally use the capital Roman alphabet to denote random variables (e.g., A, B, C, V) and the lowercase Roman alphabet to denote assignments to those random variables (e.g., $A = a$ or just a). We will use blackboard bold font to denote the set of possible values a variable can take, e.g., $a \in \mathbb{A}$. Bold will indicate a set of random variables, e.g., $\mathbf{V} = (V_1, V_2, \ldots)^\top$, and $\mathbf{v}$ is an assignment to $\mathbf{V}$. Parents and children in graphs will also follow these conventions, e.g., $\mathbf{PA}(V) = \mathbf{pa}^{\mathbf{v}}(v)$, where the assignments to those parents come from values specified in $\mathbf{v}$. We sometimes use subscripts to indicate the relevant graph structure for the parents, e.g., $\mathbf{PA}_{\mathcal{G}}(V)$, but this may be omitted without ambiguity. In order to distinguish between vectors of random variables and lists of realizations of those variables *from data*, we will also use $\vec{a}$ to indicate a list of entries for A. We will generally use the Greek alphabet (e.g., α, β) to represent parameters for structural equations.

Structural Causal Models. A structural causal model (SCM) is a tuple $\langle \mathcal{G}, P \rangle$ consisting of a directed acyclic graph (DAG) $\mathcal{G}$ and a probability measure on the variables (which are vertices of $\mathcal{G}$). The distribution factorizes according to $\mathcal{G}$:

$$\Pr(\mathbf{v}) = \prod_{v \in \mathbf{v}} P(v \mid \mathbf{pa}^{\mathbf{v}}_{\mathcal{G}}). \tag{1}$$

This gives rise to other conditional independencies from the causal Markov condition that can be described using the d-separation rules. If two variables A, B are d-separated by $\mathbf{C}$, we write $A \perp\!\!\!\perp_d B \mid \mathbf{C}$. $A \perp\!\!\!\perp_d B \mid \mathbf{C} \Rightarrow A \perp\!\!\!\perp B \mid \mathbf{C}$ is implied by the causal Markov condition. See Pearl [2009] for more information on SCMs.

Faithfulness. Constraint-based causal discovery requires two variables that are not d-separated to be statistically dependent. There are two levels to this requirement: (1) Dependence in the true distribution and (2) dependence in the empirical distribution. The first prohibits causal pathways from "canceling out." The second type of faithfulness violation can occur even if the true distribution is faithful with no cancellations. Faithfulness and the causal Markov condition combine to give conditional independence $A \perp\!\!\!\perp B \mid \mathbf{C}$ if *and only if* A, B are d-separated by $\mathbf{C}$.

Information-theoretic Measures. Kullback-Liebler (KL) divergence gives an asymmetric notion of "distance" between distributions [Kullback, 1997]. For two probability mass functions $P(x)$ and $Q(x)$ on $x \in \mathbb{X}$, the KL divergence is given by the expectation of the log-likelihood ratio under $P(x)$:

$$D(P||Q) \coloneqq \sum_{x \in \mathbb{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right). \tag{2}$$

One interpretation of $D(P||Q)$ is the (average) "surprise" when assuming Q and seeing P. Surprising/falsifying observations will correspond to large log-likelihood terms, weighted according to their probability in P.

The KL divergence between a joint probability distribution on two random variables, $P(x, y)$, and their product, $P_{X \perp\!\!\!\perp Y} = P(x)P(y)$, gives "mutual information":

$$I(X : Y) \coloneqq D(P||P_{X \perp\!\!\!\perp Y}). \tag{3}$$

Conditional mutual information uses conditional probabilities in Eq. (3). For additional background, see Cover [1999].

4 DEPENDENCE OF FAITHFULNESS

In this section, we examine the dependence between different faithfulness violations in $\langle P, \mathcal{G} \rangle$ by perturbing relationships (covariances) between different variables. In particular, we consider a joint distribution P that is generated from a linear additive Gaussian noise model. By describing the

joint probability distributions using an arbitrary set of structural equations, faithfulness is exhibited from the parameters of those equations, and violations can be created by locally perturbing these parameters.

Local perturbations on equation parameters can mimic the estimation errors of empirical distributions on finite samples. By observing the co-occurrence of faithfulness violations after these parameter perturbations, we obtain insights into the dependencies between different faithfulness violations.

4.1 FAITHFULNESS VIOLATIONS

In linear additive Gaussian models, each variable V is associated with additive Gaussian noise N_V that is mutually independent between different variables. Assume without loss of generality that N_V is standardized. Consider the following system:

$$A = N_A, \quad B = \alpha A + N_B.$$

According to this characterization, $\mathrm{Cov}(A, B) = \alpha$. If $A \to B$, then P is faithful to $\mathcal{G}$ if and only if $A \not\perp\!\!\!\perp B$, which occurs when $\mathrm{Cov}(A, B) = \alpha \neq 0$. Therefore, perturbing α to α' with $|\alpha'| < |\alpha|$ moves toward a violation of faithfulness, by lowering the covariance.

It is important to understand that this section perturbs relationships between different variables, which are not necessarily the true structural equations. As a result, $\alpha = 0$ does not mean $A \not\to B$, but rather that the covariance is zero (which may be due to the cancellation of other causal pathways in the system).

We now consider two faithfulness violations jointly and give examples of both "positive dependence" and "negative dependence". A *positive* dependence involves a perturbation *towards* one faithfulness violation that also moves the joint distribution *towards* another faithfulness violation. In contrast, a *negative* dependence involves a perturbation *towards* one faithfulness violation that moves the joint distribution *away from* another faithfulness violation.

4.2 POSITIVE DEPENDENCE

Consider the following structural equations for a fully connected DAG $\mathcal{G} : A \to B \to C, A \to C$,

$$A = N_A, \quad B = \alpha_1 A + N_B, \quad C = \alpha_2 A + \beta B + N_C. \tag{4}$$

Writing each variable in terms of N_A, N_B, N_C,

$$\begin{aligned} A &= N_A, \\ B &= \alpha_1 N_A + N_B, \\ C &= \alpha_2 N_A + \beta\alpha_1 N_A + \beta N_B + N_C. \end{aligned}$$

We will study two covariances for this model,

$$\mathrm{Cov}(A, B) = \alpha_1, \quad \mathrm{Cov}(A, C) = \alpha_2 + \alpha_1\beta.$$

For the first example, let's assume $\alpha_2 = 0$, which corresponds to a simplified model $A \to B \to C$. Faithfulness of P to $\mathcal{G}$ requires $A \not\perp\!\!\!\perp B$ and $A \not\perp\!\!\!\perp C$, i.e., $\mathrm{Cov}(A, B) \neq 0$ and $\mathrm{Cov}(A, C) \neq 0$.

Notice that a perturbation from α_1 to α_1' with $|\alpha_1'| < |\alpha_1|$ also lowers the dependency between A, C as $\mathrm{Cov}(A, C) = |\alpha_1'\beta| < |\alpha_1\beta|$. Therefore, faithfulness violations $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$ co-occur under these types of perturbations, so long as $|\beta|$ is not too large relative to $|\alpha_1|$.

4.3 NEGATIVE DEPENDENCE

Now, consider the joint distribution P generated by

$$\alpha_1 = 1, \quad \alpha_2 = 1, \quad \beta = -1. \tag{5}$$

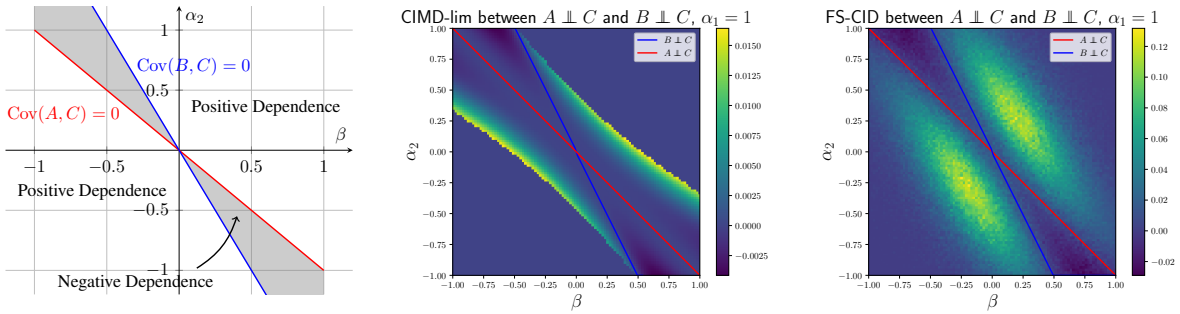
This specification of parameters gives $\mathrm{Cov}(A, C) = 0$, a faithfulness violation ($A \perp\!\!\!\perp C$). On the other hand, $\mathrm{Cov}(A, B) = \alpha_1 = 1$, which is faithful with respect to $A \not\perp\!\!\!\perp {}_d B$. A perturbation reducing $\alpha_1 = 0$ will induce $A \not\perp\!\!\!\perp B$ — a faithfulness violation with respect to $A \not\perp\!\!\!\perp {}_d B$). This perturbation simultaneously makes $\mathrm{Cov}(A, C) = 1$, meaning we are now faithful with respect to $A \not\perp\!\!\!\perp {}_d C$. Therefore, in the neighborhood of P as given by the parameters in Equation (5), deviations towards $A \perp\!\!\!\perp C$ are unlikely to also be close to $A \perp\!\!\!\perp B$.

Remarks. Our example of positive dependence in Section 4.2 holds for *all* parameter specifications with $\alpha_2 = 0$ — i.e., distributions that follow the DAG structure $A \to B \to C$. That is, $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$ co-occur when perturbing α_1 regardless of the specification of β. In contrast, the example of negative dependence is only local to perturbations of the distribution P specified by Eq. (5). For example, consider an alternative joint distribution Q generated using

$$\alpha_1 = 1, \quad \alpha_2 = 1, \quad \beta = 1. \tag{6}$$

Perturbing α_1 moves the faithfulness condition $A \not\perp\!\!\!\perp B$ and $A \not\perp\!\!\!\perp C$ in the same direction, giving another example of *positive* dependence. Hence, although the same graphical model underlies P and Q, these distributions differ in their dependencies between (local) faithfulness violations. More generally, Figure 1 (a) illustrates the space of possible models generated by different specifications of Eq. (4), with regions of positive and negative dependence annotated.

It is clear that we cannot rely solely on the graph structure (such as the graphoid axioms in Pearl [2014]) when studying the dependence between faithfulness violations under a random perturbation. Instead, we need to measure how measures of dependence change relative to local perturbations of the joint-distribution. In the next section, we will define an information-projection-based quantity to formally measure shared information between two CI tests.



(a) (b) (c)

Figure 1: (a) Dependence of faithfulness for structural equations given in Eq. (4) when $\alpha_1 = 1$. The shaded region between the red and blue lines corresponds to models where moving towards blue moves *away* from red. Hence, the shaded region has a "negative" dependence between faithfulness violations $A \perp\!\!\!\perp C$ and $B \perp\!\!\!\perp C$, while the unshaded region has a "positive" dependence. (b) The defined metric, CIMD (section 5, measuring the meta-dependence between CI-tests, computed based on access to parameters of the structural equation. (c) Estimation of CIMD with finite samples (Sections 6 and 7).

5 CIMD VIA PROJECTIONS

In this section, we formally define the CI Meta-Dependence (CIMD) between two conditional independence tests. This test uses the geometric intuition built in Section 4 to quantify the shared information between the tests. In Section 7, we will link CIMD to the co-occurrence of significance between two tests under perturbations due to finite sample uncertainty.

5.1 PROJECTING INTO INDEPENDENCE

To define CIMD, we first consider the relationship between a joint distribution and a particular conditional independence via information projections.

Consider a motivating example where we need to describe X, Y using a graphical model on $\{X, Y\}$. A joint distribution P with $X \not\perp\!\!\!\perp Y$ would be modeled as $X \to Y$, $X \leftarrow Y$ or $X \leftrightarrow Y$. In contrast, a joint distribution with $X \perp\!\!\!\perp Y$, e.g., obtained by independently drawing X, Y from their respective marginal distributions and denoted by $P_{X \perp\!\!\!\perp Y}$, would be modeled using an empty graph. The mutual information between X, Y,

$$I(X : Y) = D(P || P_{X \perp\!\!\!\perp Y}),$$

can be interpreted as the KL divergence between a joint distribution P and its *information projection* onto a class of distributions that follow an alternative DAG, namely the empty graph in this case.

Formally, we define the information projection of $P(\mathbf{V})$ onto a class of distributions that are Markovian with respect to $\mathcal{G}$, denoted by $\mathbb{P}_{\mathcal{G}}$, as

$$P_{\mathcal{G}}(\mathbf{V}) \coloneqq \mathrm{argmin}_{Q \in \mathbb{P}_{\mathcal{G}}} D(P||Q). \tag{7}$$

Similarly, we define the information projections onto a class of distributions that satisfy the conditional independence of $A \perp\!\!\!\perp B \mid \mathbf{C}$, denoted by $\mathbb{P}_{A \perp\!\!\!\perp B \mid \mathbf{C}}$, as

$$P_{A \perp\!\!\!\perp B \mid \mathbf{C}} \coloneqq \mathrm{argmin}_{Q \in \mathbb{P}_{A \perp\!\!\!\perp B \mid \mathbf{C}}} D(P||Q). \tag{8}$$

5.2 CI META-DEPENDENCE

To get the shared information between two CI-tests, we examine whether the (conditional) mutual information corresponding to one test changes after we apply an information projection with respect to the other test.

Definition 5.1 (CIMD)**.** Consider two conditional independencies:

$$(T_1) \;\; A \perp\!\!\!\perp B \mid C \qquad (T_2) \;\; A' \perp\!\!\!\perp B' \mid C'$$

The *CI Meta-Dependence (CIMD)* between these two tests on an empirical distribution P is given by the difference

$$\mathrm{CIMD}(T_1, T_2, P) \coloneqq I_P(A : B \mid \mathbf{C}) - I_{P_{T_2}}(A : B \mid \mathbf{C}).$$

Here $I_P(\cdot)$ is computed using the joint distribution P while $I_{P_{T_2}}(\cdot)$ is computed using the projection of P onto T_2, as defined in Eq. (8).

A standard result from information geometry is

$$I(A : B \mid \mathbf{C}) = D(P || P_{A \perp\!\!\!\perp B \mid \mathbf{C}}). \tag{9}$$

That is, conditional mutual information is the distance to the class of conditional independence that the projection achieves. To see why this does not depend on any other variables $\mathbf{X}$, write out the expression for KL divergence and cancel out $P(\mathbf{X} \mid a, b, \mathbf{c})$ from the numerator and denominator of the log-likelihood ratio.

CIMD measures whether first projecting onto the closest distribution that satisfies T_2 brings the joint distribution closer to the class of distributions that satisfy T_1. By interpreting an information projection as the minimum perturbation of the joint distribution into a CI property, this quantity measures whether this perturbation also moves towards another CI property. If the CIMD between two tests is 0, it means the two tests are "orthogonal" and share no information.

Remarks. We note three important properties of CIMD. First, there is an asymmetry of how CIMD depends on T_1 and T_2, i.e., $\text{CIMD}(T_1, T_2, P) \neq \text{CIMD}(T_2, T_1, P)$. The rationale behind this can be understood using an example: in the DAG $A \to B \to C \to D$, a perturbation towards $B \perp\!\!\!\perp C$ necessarily implies $A \perp\!\!\!\perp D$ because $A \perp\!\!\!\perp D$ relies on the connection between B and C. However, a perturbation towards $A \perp\!\!\!\perp D$ need not imply $B \perp\!\!\!\perp C$, as it could be a result of a $C \perp\!\!\!\perp D$ faithfulness violation.

Second, CIMD is *different* from Euclidean distance in the parameter space of the structural equation model. To see this, revisit the examples in Section 4: information projections do not move to the nearest point on the red and blue faithfulness-violation lines in Figure 1 — instead, they correspond to zeroing out coefficients of the structural equations.

Third, CIMD is dependent on the joint distribution P. The CIMD between the same two tests may be positive for some P and negative for other Q, as in Section 4.3.

6 PRACTICAL IMPLEMENTATION

In this section, we show how CIMD can be approximated and computed efficiently in practice. In particular, we show that a maximum likelihood procedure recovers the information projection for generic joint distributions in Section 6.1. We then show how to compute information projections and CIMD in closed-form for Gaussian variables in Section 6.2.

6.1 REFACTORING EMPIRICAL DISTRIBUTIONS

In practice, we need to compute CIMD using finite samples drawn from some unknown joint distribution P. Denote the empirical distribution estimated via finite samples by $\hat{P}$. We first consider how to project $\hat{P}$ onto a class of distributions that are Markovian with respect to $\mathcal{G}$, as defined in Eq. (7). We show that the projection of $\hat{P}$ is the composition of maximum likelihood estimators (MLEs) for each vertex, given their parents in $\mathcal{G}$ as inputs.

Our approach closely follows the argument in Goodfellow [2016] (page 128) that maximum likelihood classifiers minimize KL divergence. The additional element comes from observing the separate parametrization for each structural equation describing the distribution of a variable given its parents, which allows us to consider each structural equation as a separate optimization.

Lemma 6.1. *Fix some function-class for the structural equations and let $\mathbb{P}_\mathcal{G}$ be the set of distributions generated according to the DAG $\mathcal{G}$. Projecting $\hat{P}$ into $\mathbb{P}_\mathcal{G}$ corresponds to composing maximum likelihood estimators in the factorization of $\mathcal{G}$. That is,*

$$\hat{P}_\mathcal{G}(\mathbf{v}) = \prod_{v \in \vec{v}} \hat{P}(v \mid \mathbf{pa}^\mathbf{v}(v)), \tag{10}$$

where $\hat{P}(v \mid \mathbf{pa}^\mathbf{v}(v))$ are the MLEs for predicting V using $\mathbf{PA}(V)$.

Proof. Let $P^\theta_\mathcal{G}(\mathbf{v})$ be a probability distribution, parameterized by θ, that factorizes according to $\mathcal{G}$. The maximum likelihood parameterization for dataset $\vec{\mathbf{v}}$ is given by

$$\hat{\theta} \coloneqq \text{argmax}_\theta P^\theta_\mathcal{G}(\vec{\mathbf{v}}).$$

We will expand $P^\theta_\mathcal{G}(\vec{\mathbf{v}})$ in two ways — one across *data entries* of $\mathbf{v}^{(i)} \in \vec{\mathbf{v}}$, assuming they have been gathered independently from P, and another across *variables* $v \in \mathbf{v}$ according to the factorization in $\mathcal{G}$. We begin with an expansion across m data entries that converts our optimization to one on the empirical expectation of log-likelihood:

$$\begin{aligned} \hat{\theta} &= \text{argmax}_\theta \prod_{i=1}^{m} P^\theta_\mathcal{G}(\mathbf{v}^{(i)}) \\ &= \text{argmax}_\theta \sum_{i=1}^{m} \log(P^\theta_\mathcal{G}(\mathbf{v}^{(i)})) \\ &= \text{argmax}_\theta \mathbb{E}_{\mathbf{V} \sim \hat{P}(\mathbf{v})}[\log(P^\theta_\mathcal{G}(\mathbf{V})] \end{aligned} \tag{11}$$

We now show that Eq. (11) is both the minimization of KL divergence and the composition of maximum likelihood models for each variable using their parents.

To equate this with minimal KL divergence, we observe that the empirical log-likelihood $\log(\hat{P}(\mathbf{V}))$ is invariant to the parametrization θ of $P^\theta_\mathcal{G}$. Therefore, we can add empirical log-likelihood to the optimization. Negating the objective and adding in this empirical likelihood term shows the following connection to KL divergence minimization:

$$\hat{\theta} = \text{argmin}_\theta D(\hat{P} || P^\theta_\mathcal{G}).$$

To break Equation (11) into the composition of MLE models for variables given their parents, we utilize the factorization of $P^\theta_\mathcal{G}(\mathbf{v})$ and linearity of expectation to break the expectation into the conditional log-likelihoods of n variables $V_j \in \mathbf{V}$.

$$\hat{\theta} = \text{argmax}_\theta \sum_{j=1}^{n} \mathbb{E}_{\mathbf{V} \sim \hat{P}(\mathbf{v})}[\log(P^\theta_\mathcal{G}(V_j \mid \mathbf{pa}^\mathbf{V}(V_j))].$$

The factorization according to $\mathcal{G}$ means that we can parameterize each conditional distribution (conditioned on parents)

separately. That is, $P_{\mathcal{G}}^{\theta}$ has separate parameters θ_j for the conditional distribution of each variable V_j given its parents. This means we can partition $\theta = \cup_j \theta_j$ and distribute the argmax:

$$\hat{\theta} = \sum_{j=1}^{n} \text{argmax}_{\theta_j} \mathbb{E}_{\mathbf{V}\sim\hat{P}(\mathbf{v})}[\log(P_{\mathcal{G}}^{\theta_j}(V_j \mid \mathbf{pa}^{\mathbf{v}}(V_j))].$$

The optimization has now been separated into a sum of subproblems, each solved with its corresponding MLE. □

Lemma 6.1 shows how to find an information projection onto general $\mathcal{G}$. By applying this for a $\mathcal{G}$ that induces $A \perp\!\!\!\perp B \mid \mathbf{C}$, we get the information projection for a specific conditional independence.

Theorem 6.2. *Consider the joint probability distribution on $\mathbf{V} = \{A, B\} \cup \mathbf{C} \cup \mathbf{X}$. Fix some function-class for structural equations on variables $\mathbf{V}$ and let $\mathbb{P}_{A \perp\!\!\!\perp B \mid \mathbf{C}}$ be the set of distributions that can be generated according to these functions with the property that $A \perp\!\!\!\perp B \mid \mathbf{C}$. Using $\mathbb{C}$ to denote the set of possible assignments to $\mathbf{C}$, projecting $\hat{P}$ into $\mathbb{P}_{A \perp\!\!\!\perp B \mid \mathbf{C}}$, corresponds to*

$$\hat{P}_{A \perp\!\!\!\perp B \mid \mathbf{C}}(\vec{v}) = \sum_{\mathbf{c}\in\mathbb{C}} \hat{P}(\mathbf{c})\hat{P}(a \mid \mathbf{c})\hat{P}(b \mid \mathbf{c})\hat{P}(\mathbf{x} \mid a, b, \mathbf{c}),$$

where $\hat{P}(a \mid \mathbf{c})$ and $\hat{P}(b \mid \mathbf{c})$ are MLEs trained to predict a, b using $\mathbf{c}$ and $\hat{P}(\mathbf{x} \mid a, b, \mathbf{c})$ is an MLE trained to predict $\mathbf{x}$ using $a, b, \mathbf{c}$.

Theorem 6.2 can be proved by factorizing according to $P(\mathbf{V}) = P(\mathbf{C})P(A, B \mid \mathbf{C})P(\mathbf{X} \mid A, B, \mathbf{C})$ and then applying Lemma 6.1 to the resulting graphical structure, i.e. $A \leftarrow \mathbf{C} \rightarrow B$ and all of $\{A, B\} \cup \mathbf{C}$ pointing to all of $\mathbf{X}$.

Theorem 6.2 shows how to sample from $\hat{P}_{A \perp\!\!\!\perp B|\mathbf{C}}$ by training two models for A, B using input $\mathbf{C}$. We then sample A, B from these models and use a third model trained for $\mathbf{X}$ based on $A, B, \mathbf{C}$ to sample everything else. Crucially, this computation does not rely on the underlying P, nor does it require any knowledge of the underlying graphical structure.

6.2 GAUSSIAN VECTORS

Lemma 6.1 and Theorem 6.2 show that we can project an empirical probability distribution into a class of distributions by substituting empirical conditional probabilities into a factorization consistent with that class. For general function classes, this corresponds to composing models for each variable given its parents. For conditional independence of $A \perp\!\!\!\perp B \mid \mathbf{C}$, this corresponds to sampling independently from models for A and B (with input $\mathbf{C}$), then sampling the rest of the variables from a model built on input $A, B, \mathbf{C}$.

We now consider a special case of linear models with additive Gaussian noise, for which we are able to derive closed-form solutions. In this section, we will compute information projections by directly modifying the covariance matrix for P on $A, B, \mathbf{C}, \mathbf{X}$:

$$\Sigma = \begin{bmatrix} \Sigma_{\{A,B\},\{A,B\}} & \Sigma_{\{A,B\},\mathbf{C}} & \Sigma_{\{A,B\},\mathbf{X}} \\ \Sigma_{\mathbf{C},\{A,B\}} & \Sigma_{\mathbf{C},\mathbf{C}} & \Sigma_{\mathbf{C},\mathbf{X}} \\ \Sigma_{\mathbf{X},\{A,B\}} & \Sigma_{\mathbf{X},\mathbf{C}} & \Sigma_{\mathbf{X},\mathbf{X}} \end{bmatrix}. \quad (12)$$

According to Theorem 6.2,

$$\begin{aligned} P_{A \perp\!\!\!\perp B \mid \mathbf{C}}(\mathbf{C}) &= P(\mathbf{C}), \\ P_{A \perp\!\!\!\perp B \mid \mathbf{C}}(A \mid \mathbf{C}) &= P(A \mid \mathbf{C}), \\ P_{A \perp\!\!\!\perp B \mid \mathbf{C}}(B \mid \mathbf{C}) &= P(B \mid \mathbf{C}), \\ P_{A \perp\!\!\!\perp B \mid \mathbf{C}}(\mathbf{X} \mid A, B, \mathbf{C}) &= P(\mathbf{X} \mid A, B, \mathbf{C}). \end{aligned} \quad (13)$$

Denote the covariance matrix for $P_{A \perp\!\!\!\perp B \mid \mathbf{C}}$ using $\Sigma^{\perp\!\!\!\perp}$. Eq. (13) tells us that $\Sigma^{\perp\!\!\!\perp}$ is the same as Σ except for two sub-matrices: $\Sigma^{\perp\!\!\!\perp}_{\{A,B\},\{A,B\}}$ and $\Sigma^{\perp\!\!\!\perp}_{\vec{X},\vec{X}}$.

$$\Sigma^{\perp\!\!\!\perp} = \begin{bmatrix} \Sigma^{\perp\!\!\!\perp}_{\{A,B\},\{A,B\}} & \Sigma_{\{A,B\},\mathbf{C}} & \Sigma_{\{A,B\},\mathbf{X}} \\ \Sigma_{\mathbf{C},\{A,B\}} & \Sigma_{\mathbf{C},\mathbf{C}} & \Sigma_{\mathbf{C},\mathbf{X}} \\ \Sigma_{\mathbf{X},\{A,B\}} & \Sigma_{\mathbf{X},\mathbf{C}} & \Sigma^{\perp\!\!\!\perp}_{\mathbf{X},\mathbf{X}} \end{bmatrix}. \quad (14)$$

First, $\Sigma_{\{A,B\},\{A,B\}}$ must be changed so that $A \perp\!\!\!\perp B | \mathbf{C}$. This is done by setting the conditional covariance equal to zero and rearranging the terms,

$$\Sigma^{\perp\!\!\!\perp}_{\{A,B\},\{A,B\}} = \Sigma_{\{A,B\},\mathbf{C}}(\Sigma_{\mathbf{C},\mathbf{C}})^{-1}\Sigma_{\mathbf{C},\{A,B\}}. \quad (15)$$

From here, we need to generate a new covariance matrix $\Sigma_{\mathbf{X},\mathbf{X}}$ using the new distribution on A, B and the same conditional $P(\mathbf{X} \mid A, B, \mathbf{C})$. Using $\overline{\mathbf{X}} \coloneqq \{A, B\} \cup \mathbf{C}$,

$$\Sigma^{\perp\!\!\!\perp}_{\overline{\mathbf{X}},\overline{\mathbf{X}}} = \begin{bmatrix} \Sigma^{\perp\!\!\!\perp}_{\{A,B\},\{A,B\}} & \Sigma_{\{A,B\},\mathbf{C}} \\ \Sigma_{\mathbf{C},\{A,B\}} & \Sigma_{\mathbf{C},\mathbf{C}} \end{bmatrix}. \quad (16)$$

If this is invertible, we can compute

$$\begin{aligned} \Omega_{\overline{\mathbf{X}},\overline{\mathbf{X}}} &\coloneqq (\Sigma^{\perp\!\!\!\perp}_{\overline{\mathbf{X}},\overline{\mathbf{X}}})^{-1} - \Sigma^{-1}_{\overline{\mathbf{X}},\overline{\mathbf{X}}} \\ \Sigma^{\perp\!\!\!\perp}_{\mathbf{X},\mathbf{X}} &= \Sigma_{\mathbf{X},\mathbf{X}} + \Sigma_{\mathbf{X},\overline{\mathbf{X}}}\Omega_{\overline{\mathbf{X}},\overline{\mathbf{X}}}\Sigma_{\overline{\mathbf{X}},\mathbf{X}}. \end{aligned} \quad (17)$$

To get a closed-form expression for CIMD from this projection, we first calculate Φ, the covariance of $\mathbf{Z} = \{A, B\}$ conditional on $\mathbf{C}$,

$$\Phi = \Sigma^{\perp\!\!\!\perp}_{\mathbf{Z},\mathbf{Z}} - \Sigma^{\perp\!\!\!\perp}_{\mathbf{Z},\mathbf{C}}(\Sigma^{\perp\!\!\!\perp}_{\mathbf{C},\mathbf{C}})^{-1}\Sigma^{\perp\!\!\!\perp}_{\mathbf{C},\mathbf{Z}} \quad (18)$$

Then, we apply the closed form for mutual information between multivariate Gaussians,

$$\text{CIMD}(T_1, T_2) = \frac{1}{2}\log\left(\frac{\det(\Phi^{\perp\!\!\!\perp}_{A,A})\det(\Phi^{\perp\!\!\!\perp}_{B,B})}{\det(\Phi^{\perp\!\!\!\perp})}\right). \quad (19)$$

7 EMPIRICAL TESTS

CIMD does not directly address finite sample deviations. However, KL divergence measures statistical distance — small $D(P||Q)$ indicates that data drawn from P is likely to look like Q. Motivated by this intuition, we give experiments on both synthetic and real-world data that show CIMD aligns with the dependence between outcomes of CI-tests under finite-data uncertainty.[2]

FS-CID. We will use "Finite-Sample CI Dependence" (FS-CID) to measure the dependence between two CI-tests due to finite-sample uncertainty. FS-CID bootstraps perturbations of the true distribution to empirical ones and keeps track of the co-occurrence of CI-test significance. To do this, we randomly generate 1000 smaller datasets. For synthetic data, these smaller datasets come from sampling a multivariate Gaussian. For real-world data, these datasets are random sub-samples from a larger dataset.

For each of these datasets, we compute the result of a CI-test using partial correlation with a Fisher-Z transformation.[3] If t_1 corresponds to a failure to reject test 1 and t_2 failure to reject test 2, then we report the FS-CID as

$$\text{FS-CID} \coloneqq \hat{P}(t_1, t_2) - \hat{P}(t_1)\hat{P}(t_2), \tag{20}$$

where $\hat{P}$ are empirical probability estimates from the 1000 datasets.

Limited CIMD. When either null hypothesis is very likely to be rejected (i.e., strong dependence), we will likely have $\hat{P}(t) = 0$ for an FS-CID of 0. In contrast, CIMD still projects the distribution onto that (very likely to reject) CI-test, likely giving a non-zero value. In such cases, CIMD is uninformative because the distribution is far from either CI property. To compare CIMD to FS-CID, we zero CIMD out when either mutual information term is large ($> .1$). We call this "CIMD-lim" in our experiments.

7.1 SYNTHETIC DATA

For synthetic data, we calculate FS-CID by generating 1000 datasets of size 20 using the generative model specified by the structural equations in Eq. (4) with $\alpha_1 = 1$ and $\alpha_2, \beta \in [-1, 1]$.[4] These parameters correspond exactly to those used in Figure 1 (a), with results shown in (b) and (c).

Unlike the single-parameter perturbations studied in Section 4, finite-sample deviations extend in any direction—potentialy affecting more than one parameter. Nonetheless, the regions of positive and negative dependence match up for all three plots. We provide additional synthetic experiments on different graphical structures in Appendix B.

7.2 REAL DATA

To demonstrate the relevance of CIMD in practical studies, we run experiments on the "California Housing" [Pace and Barry, 1997], "Apple Watch and Fitbit" [Fuller, 2020], and "Auto MPG" [Quinlan, 1993] datasets. These real-world datasets were chosen arbitrarily and have few continuous covariates, keeping our experiment size within reason. FS-CID is computed in these settings by sub-sampling 50 data points from these datasets 1000 times.[5]

We give the results for the "Apple Watch and Fitbit" dataset in Figure 2. Notice that CIMD-lim and FS-CID recover very similar meta-dependencies across the many conditional-independence tests. This dataset has relatively sparse dependence, giving an easy-to-recognize pattern. The results for the other two datasets are given in Appendix A. These two datasets have more complicated meta-dependencies but still share signal across the two metrics.

8 CONCLUSION

This paper pioneers the study of CI Meta-Dependence (CIMD) by defining an information-theoretic measure for the shared information between tests for conditional independence. CIMD is empirically linked to the co-occurrence of conditional independencies under finite-sample uncertainty.

While the empirical and theoretical results of this paper focus on projecting covariance matrices, the results of Section 6 generalize to projections on any distribution by composing generative models.

The most immediate implication for CIMD and its relationship to FS-CID is the ability to more finely tune significance thresholds in constraint-based causal discovery. Tests for the removal of a single edge with shared information (non-zero CIMD) do not need to lower their p-value thresholds. Tests for the same edge with low CIMD must adjust their p-values to avoid p-value hacking, which may otherwise result in an overly sparse structure.

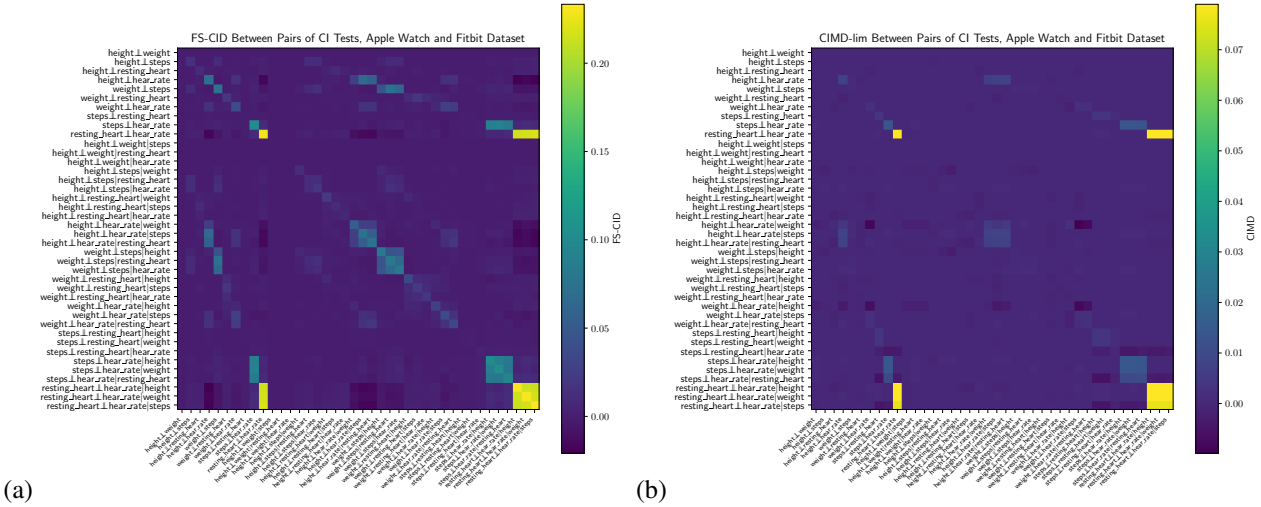
One important observation is that the dependencies between edges and conditional independence properties are functions of the observed empirical distribution. Consequently, two

[2]Code can be found at: `https://anonymous.4open.science/r/CIMD_experiments-1124/`.

[3]This hypothesis test is implemented in the `conditional-independence` python package [Chandler Squires, 2018].

[4]We use relatively small (20 samples) dataset sizes in order to induce a sufficient perturbation from the true distribution to the empirical one to resolve a dependence.

[5]We use relatively small (20 samples) dataset sizes in order to induce a sufficient perturbation from the true distribution to the empirical one to resolve a dependence.



(a) (b)

Figure 2: Meta-dependencies of CI-tests as measured by FS-CID in (a) and CMID-lim in (b).

DAG structures $\mathcal{G}$ and $\mathcal{G}'$ may be farther apart in some empirical distributions than others. The projections defined in Section 5 use a generative structure implied by the conditional independence, and can therefore be easily generalized to projections *onto a full graphical structure*. Advancements in distribution-specific graphical distances will improve the validation of causal discovery algorithms, allowing the field to focus on minimizing statistically significant errors in a causal DAG.

CIMD marks a crucial step towards quantifying higher-order uncertainty in causal structure learning. This will help us work towards learning "confidence ensembles" of causal structure. A deeper understanding of the dependencies between CI-tests will also have fundamental scientific implications for the learnability of different causal processes relative to their context within a larger causal network.

Finally, we emphasize that CIMD can be computed using an empirical estimate of a covariance matrix (or generative models), and does not require access to the raw data. The (empirical) correspondence between CIMD and FS-CID also points towards using CIMD to detect *heterogeneity in relationships* that would otherwise require $\mathcal{O}(n^2)$ clustering approaches. As such, CIMD may be useful in very large data settings or in systems where privacy or security concerns limit access to the raw data.

References

Konstantina Biza, Ioannis Tsamardinos, and Sofia Triantafillou. Tuning causal discovery algorithms. In *International Conference on Probabilistic Graphical Models*, pages 17–28. PMLR, 2020.

Chandler Squires. `causaldag`*: creation, manipulation, and learning of causal models*, 2018. URL `https://github.com/uhlerlab/causaldag`.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Martijn de Jongh and Marek J Druzdzel. A comparison of structural distance measures for causal bayesian network models. *Recent advances in intelligent information systems, challenging problems of science, computer science series*, pages 443–456, 2009.

Pablo de Morais Andrade, Julio Michael Stern, and Carlos Alberto de Bragança Pereira. Bayesian test of significance for conditional independence: the multinomial model. *Entropy*, 16(3):1376–1395, 2014.

Philipp M Faller, Leena C Vankadara, Atalanti A Mastakouri, Francesco Locatello, and Dominik Janzing. Self-compatibility: Evaluating causal discovery without ground truth. In *International Conference on Artificial Intelligence and Statistics*, pages 4132–4140. PMLR, 2024.

RA Fisher et al. The technique of field experiments. *Harpenden: Rothamsted Experimental Station*, pages 11–13, 1931.

Daniel Fuller. Replication Data for: Using machine learning methods to predict physical activity types with Apple Watch and Fitbit data using indirect calorimetry as the criterion., 2020. URL `https://doi.org/10.7910/DVN/ZS2Z2J`.

Ian Goodfellow. *Deep learning*, volume 196. MIT press, 2016.

Leonard Henckel, Theo Würtzen, and Sebastian Weichwald. Adjustment identification distance: A gadjid for causal structure learning. In *Uncertainty in Artificial Intelligence*, pages 1569–1598. PMLR, 2024.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. 2013.

Murat Kocaoglu. Characterization and learning of causal graphs with small conditioning sets. *Advances in Neural Information Processing Systems*, 36, 2023.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in neural information processing systems*, 23, 2010.

Frank Nielsen. What is an information projection. *Notices of the AMS*, 65(3):321–324, 2018.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

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

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Judea Pearl and Azaria Paz. Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 189–200. 2022.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

R. Quinlan. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

Leonard J Schulman and Piyush Srivastava. Stability of causal inference. In *UAI*, 2016.

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

Chandler Squires and Caroline Uhler. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815, 2023.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

Jiaqi Zhang, Kirankumar Shiragur, and Caroline Uhler. Membership testing in markov equivalence classes via independence queries. In *International Conference on Artificial Intelligence and Statistics*, pages 3925–3933. PMLR, 2024.

Meta-Dependence of Conditional Independence Testing (Supplementary Material)

A REAL-WORLD DATASET CI META-DEPENDENCE

Figure 3 shows six plots on three real-world datasets. These experiments demonstrate the occurrence of conditional independence meta-dependence in real-world datasets. A shared structure is expressed between FS-CID (calculated via sub-sampling the data and computing correlation in null hypothesis rejection) and limited CIMD.

B SYNTHETIC DATA EXPERIMENTS ON OTHER GRAPH STRUCTURES

We also explore different graphical structures on three nodes by setting different parameters to 0.

Markov Chain When setting $\alpha_2 = 0$, we suppress the $A \to C$, giving rise to a Markov chain $A \to B \to C$, in which $A \perp\!\!\!\perp C \mid B$.

Collider When setting $\alpha_1 = 0$, we suppress the $A \to B$, giving rise to a collider $A \to C \leftarrow B$, in which $A \perp\!\!\!\perp B$ but $A \not\perp\!\!\!\perp B \mid C$.

Both the Markov chain and the collider experiments traverse surfaces in which the second test is satisfied *before* any projection. For $A \to B \to C$, we start with $A \perp\!\!\!\perp C \mid B$ already being satisfied — projecting onto $A \perp\!\!\!\perp C$ does not take us any closer, giving a large yellow region representing a CIMD of 0. However, once the edges between $A \to B$ and $B \to C$ get stronger (in the edges of the plot), projecting onto $A \perp\!\!\!\perp C$ seems to move us away from $A \perp\!\!\!\perp C \mid B$, represented by the darker blue (with a negative CIMD).

Similarly, the collider begins with $A \perp\!\!\!\perp B$ being satisfied, and projecting onto $A \perp\!\!\!\perp B \mid C$ does not affect its satisfaction (shown again by a large yellow region of 0 CIMD). Again, stronger edges between $A \to C$ and $B \to C$ in the corners create a negative CIMD.

From these experiments, we observe that negative CIMD appears to occur between tests on the same two variables with different conditioning sets, particularly when one test holds in the ground truth distribution, and the other does not.

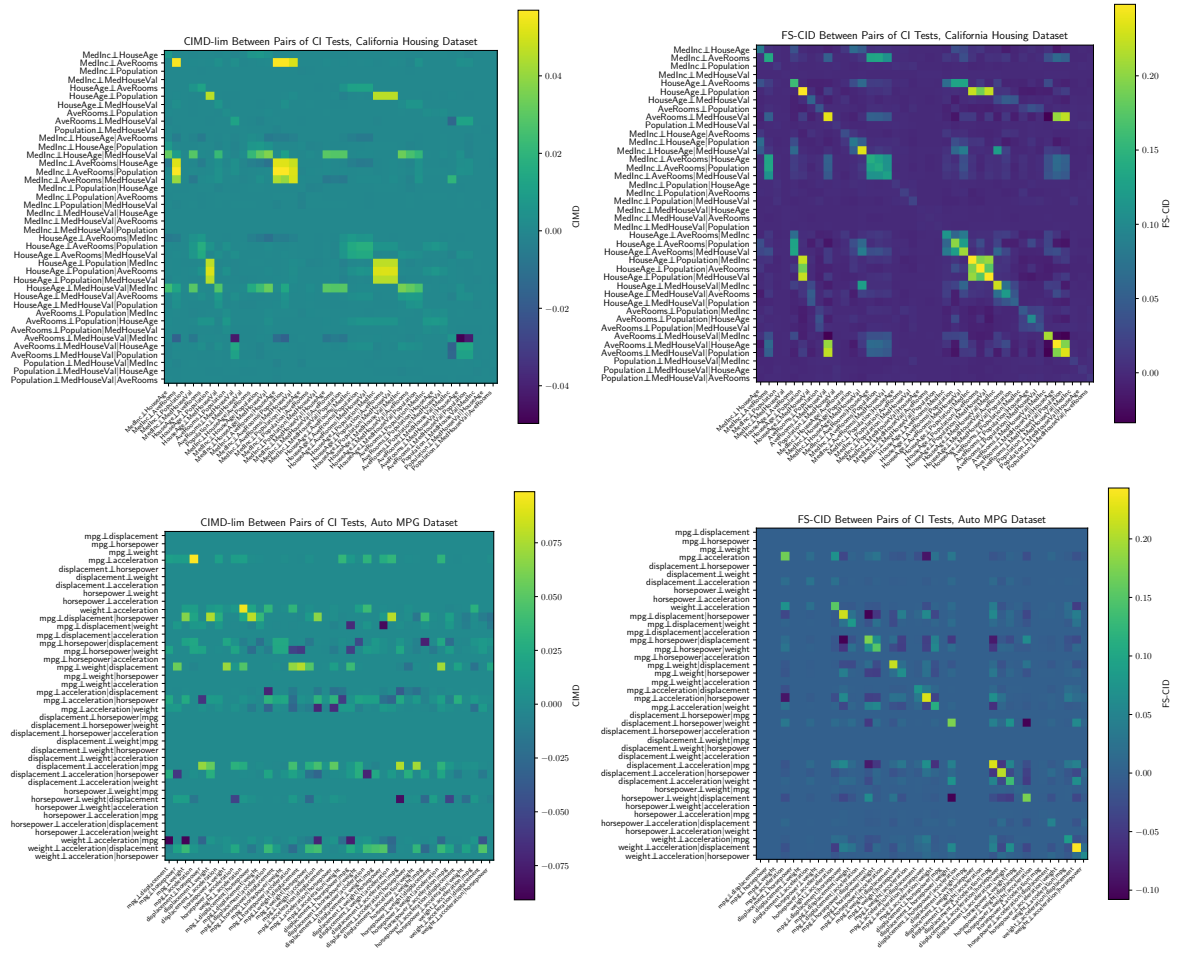
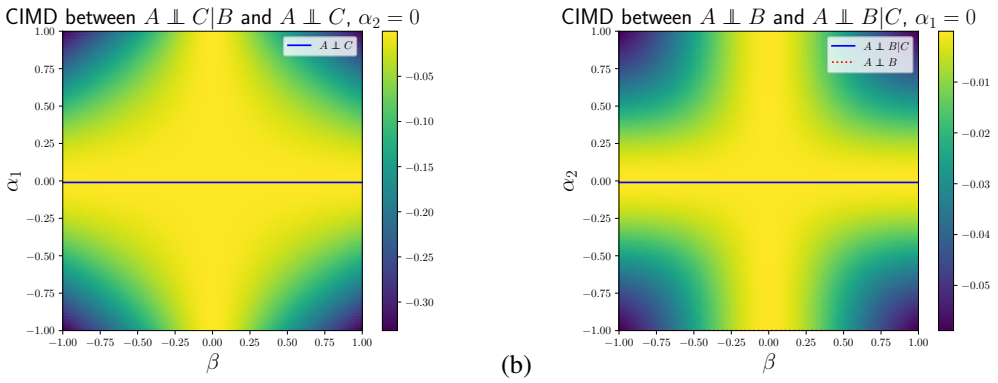


Figure 3: Demonstrating the dependence between CI-tests in real-world data using both FS-CID and limited CIMD.



(a) (b)

Figure 4: Both (a) and (b) plot the CIMD for various parameters from the structural equations given in Equation 4. The effective structure in (a) is a Markov chain, and the effective structure in (b) is a collider.