
Meta-Dependence in Conditional Independence Testing

Bijan H. S. Mazaheri^{1,2}

Jiaqi Zhang^{2,3}

Caroline Uhler^{2,3}

¹Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire, USA

²Schmidt Center, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

³Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Abstract

Conditional independence testing is a critical component of feature screening, invariant statistical models, and causal discovery. Many of these algorithms rely on the sequential application of conditional independence tests, and their stability hinges on how their outcomes interact. We study this “meta-dependence” between conditional independence properties using the following geometric intuition: satisfying each conditional independence property constrains the space of possible joint distributions to a manifold. The “meta-dependence” of multiple conditional independences in a probability distribution is informed by its position relative to these manifolds. We provide a simple-to-compute measure of this meta-dependence using moment projections, with a closed-form expression for multivariate Gaussian distributions, and consolidate our findings empirically using both synthetic and real-world data. Our measure of meta-dependence does not rely on graphical properties of the distribution and can be computed directly from summary statistics such as a covariance matrix, allowing for various applications. We demonstrate one use case of meta-dependence, using a simple redundancy metric to tune significance thresholds and improve causal discovery.

1 INTRODUCTION

Structural Causal Models (SCMs), popularized by Pearl [1998], are networks of causal dependencies that drive data-generating processes. Knowing the graphical structure of these networks enables us to identify causal effects [Pearl, 2009], efficiently search for root causes [Ikram et al., 2022], and design more informative experiments [Zhang et al., 2023]. In practice, however, the SCM is rarely available

a priori and must be inferred from data. This is known as *causal discovery* [Spirtes et al., 2000].

Data generated from SCMs exhibit predictable conditional independence (CI) structures from the causal Markov condition, and these structures can be read off the causal graph. For example, a chain $A \rightarrow B \rightarrow C$ exhibits a Markov property whereby the first and last random variables are independent given the middle one. *D-separation* conditions formalize exactly how graphical structure constrains the joint distribution with respect to conditional independence, and the resulting properties are the machinery underlying most graphical methods for causal identification [Pearl, 2009]. The same reasoning underlies methods well beyond effect identification, from invariant feature selection [Rojas-Carulla et al., 2018] to the transportability of causal and statistical findings across settings [Pearl and Bareinboim, 2014].

Motivation Conditional independence properties are not separate degrees of freedom within a causal system. The “graphoid axioms” [Pearl and Paz, 1986] and the “Verma constraints” [Verma and Pearl, 1990] (the latter arising with latent variables) capture how CIs constrain one another, but these are *qualitative, graph-level* rules. Such dependence matters for any algorithm that uses multiple CI tests: if an algorithm fails when *at least one test* is incorrect, independent failures are costlier than co-occurring ones. This is most stark in hypothesis testing, where the standard Bonferroni correction treats test errors as independent, overcorrecting when they are redundant. Yet no metric quantifies the information shared between CI tests at the level of the distribution, let alone one computable without the causal graph. We provide a geometric insight into this dependence and a simple-to-compute metric for it.

1.1 PROBLEM AND SUMMARY

Suppose we repeatedly subsample a dataset, keeping track of the outcomes of two CI tests. Under this stochastic process,

the two tests become random variables that can correlate. We call this phenomenon “meta-dependence,” and we call its sub-sampling estimate “Finite Sample CI Dependence” (FS-CID). While FS-CID can clearly be computed by the bootstrapped approach just described, it is time-consuming and provides no mathematical insight into the reason for these dependencies.

In this paper, we quantify the meta-dependence between different CI tests geometrically. We do so without access to a causal graph or the raw data, requiring only summary statistics of the distribution (e.g., a covariance matrix). The geometric picture comes from intuition built in Section 3, which shows that a slight change to a distribution may bring it closer to some CI properties and farther from others.

Such distribution “changes” are easy to interpret as perturbations on the parameters of the structural equations, but this framework disappears when the parametric form of the underlying data-generating distribution is unknown. As such, Section 4 extends the intuition to a “moment projection” [Nielsen, 2018] from an empirical distribution onto a manifold of distributions that obey some CI property. We use these moment projections to formally define Conditional Independence Meta-Dependence (CIMD), which measures the change in Kullback–Leibler (KL) divergence associated with one CI property when the distribution is projected onto another. To make these notions practical, Section 4.3 shows that both moment projections and CIMD have closed forms for multivariate Gaussian distributions.

We find that CIMD can be positive or negative and, somewhat surprisingly, that such variations can arise within models that obey the *same causal structure* — only differing in their structural equations. These observations suggest that the co-occurrence or anti-correlation of CI-testing errors is not a purely graphical phenomenon (as is the case for graphoid axioms and Verma constraints), but also driven by the underlying ground-truth probability distribution. This dependence manifests in finite-sample dependence as follows: finite-sample uncertainty perturbs the ground-truth distribution into an empirical one that is likely “close” in distributional distance,¹ so properties that correspond to similar perturbations are likely to co-occur. We establish this link by showing that variations in CI-test outcomes under FS-CID simulations closely match the purely distributional CIMD for both synthetic and real-world data (Section 5).

While the implications of this insight are fundamental, our results also have immediate consequences for the tuning of algorithms that rely on CI tests. To support this claim, we provide a simple adjustment to the widely used PC algorithm from causal discovery that assigns different significance thresholds across tests (Section 5.3), improving accuracy in data-sparse settings.

¹We will use the KL divergence, which is not a distance, but is close in concept.

1.2 RELATED WORKS

Causal Discovery. 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 (see Vowels et al. [2022], Squires and Uhler [2023] for reviews). One approach to causal discovery involves a *constraint-based* search guided by many CI tests, e.g., the PC-algorithm from Spirtes et al. [2000]. As such, meta-dependence between CI properties is of great interest to researchers in this area, both for tuning of hypothesis significance thresholds and in understanding the problem’s inherent complexity.

Causal Strength and Stability. Related work quantifies the strength of *causal* relationships [Janzing et al., 2013] and the stability of causal models [Schulman and Srivastava, 2016]. These notions analyze causal quantities and assume a *known* structure; we instead target causal discovery, where the structure is *unknown*.

Complexity of DAGs. Relatedly, some DAGs are “harder” to learn than others — a “DAG complexity” tied to the data needed to verify a structure. One such quantity is the number of CI queries required to confirm a structure [Zhang et al., 2024], but this purely graphical count ignores the distribution-driven *dependencies* between queries that we study here. A related notion is “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.

Feature Selection. Many feature-selection algorithms take advantage of the observation that conditioning on a sufficiently informative set of features makes the rest of the features independent from the label that we seek to predict [Tsamardinos and Aliferis, 2003]. Such a set has been characterized as a Markov blanket or Markov boundary [Richardson, 2003, Pearl, 2009] (see Yu et al. [2020] for a review of causality-informed feature selection procedures). Causal structure has also been applied to understand relationship invariances in the presence of shifting domains and distributions [Magliacane et al., 2018, Rojas-Carulla et al., 2018, Mazaheri et al., 2023], and these domains may benefit from a better understanding of CI meta-dependence.

2 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., $\mathbf{V} = \{V_1, V_2, \dots\}$, and $\mathbf{v}$ is an assignment to $\mathbf{V}$. Parents and children in graphs will also follow these conventions, e.g., $\mathbf{PA}(\mathbf{V}) = \mathbf{pa}^{\mathbf{v}}(\mathbf{V})$, 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., $\mathbf{PA}_{\mathcal{G}}(\mathbf{V})$, but this may be omitted without ambiguity. 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 $\mathcal{G}$) [Pearl, 2009]. The distribution factorizes according to $\mathcal{G}$:

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

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 $A \perp_d B \mid \mathbf{C}$. The causal Markov condition gives $A \perp_d B \mid \mathbf{C} \Rightarrow A \perp B \mid \mathbf{C}$.

Faithfulness. The formulation of structural causal models yields the causal Markov condition, but the converse is not necessarily true [Ramsey et al., 2012]. That is, an “unblocked” causal pathway between two variables does not necessarily require that they be statistically dependent. 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. The success of constraint-based causal discovery hinges on this correspondence between graphical properties (d-separation) and statistical independence.

Strong Faithfulness. Perfect statistical independence is extremely unlikely within empirical probability distributions. To address this, Zhang and Spirtes [2012] introduced a stronger notion of faithfulness that requires a minimal threshold of dependence between all causally linked variables; a closely related assumption underlies the high-dimensional analysis of the PC-algorithm by Kalisch and Bühlmann [2007]. For mathematical simplicity, the definition is only given for linear structural equations with additive Gaussian noise, where conditional dependence can be quantified using partial correlation.

Definition 2.1 (λ -strong faithfulness). For constant $\lambda \in (0, 1)$, a multivariate Gaussian distribution P is said to be λ -strong-faithful to DAG $\mathcal{G}$ with respect to two variables

A, B and set $\mathbf{C}$,² if and only if the absolute value of the partial correlation between A, B conditioned on $\mathbf{C}$ (denoted $\rho_{A,B|\mathbf{C}}$) is bounded below by λ . That is, λ -strong faithfulness requires

$$A \not\perp_d B \mid \mathbf{C} \iff |\rho_{A,B|\mathbf{C}}| > \lambda.$$

The introduction of λ -strong faithfulness enabled a more critical examination into the faithfulness assumption. While a set of unfaithful distributions has Lebesgue measure 0, the study of Uhler et al. [2013] showed that the geometry of unfaithful distributions is a sufficiently “bendy” manifold such that the majority of distributions are “close” to an unfaithful distribution. As such, most models will be in danger of (empirically) deviating towards a faithfulness violation, leading to unstable results. This paper utilizes this geometric perspective to study how these instabilities are related, which serves as a fundamental motivation for the perspective we pose on CI meta-dependencies.

Information-theoretic Measures. While strong faithfulness was originally defined with respect to Gaussian vectors with linear dependencies, we will generalize this notion using information theory. Kullback–Leibler (KL) divergence gives an asymmetric notion of difference between distributions [Kullback, 1997]. For two distributions P and Q over $x \in \mathbb{X}$ with densities (or, for discrete variables, probability mass functions) p and q , the KL divergence is the expected log-likelihood ratio under P :

$$D(P||Q) := \mathbb{E}_{x \sim P} \left[\log \frac{p(x)}{q(x)} \right]. \quad (2)$$

The CIMD quantities we define below apply to both discrete and continuous formulations of KL divergence, while our closed-form results (Section 4.3) instantiate the continuous, Gaussian setting. One interpretation of $D(P||Q)$ is the (average) “surprise” when assuming Q and seeing P .

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

$$I(X : Y) := D(P||P_{X \perp Y}). \quad (3)$$

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

3 ILLUSTRATING META-DEPENDENCE

In this section, we illustrate the geometric intuition between CI dependencies by perturbing the relationships between different variables. In particular, we consider a joint distribution P that is generated from a linear additive Gaussian

²Here $\mathbf{C} \cap \{A, B\} = \emptyset$. In cases where $\mathbf{C} = \emptyset$, we omit it for simplicity.

noise model. By describing the joint Gaussian distributions using an arbitrary set of structural equations, local perturbations on equation parameters can mimic the estimation errors of empirical distributions on finite samples. These perturbations alter the partial correlation between variables, which can result in weaker or stronger dependencies.

Intuitively, for each tuple A, B, C , the set of joint distributions $\mathcal{P}_{A \perp\!\!\!\perp B \mid C}$ that satisfy $\rho_{A,B \mid C} = 0$ constitutes a manifold in the space of possible joint distributions. Weaker conditional dependence then corresponds, informally, to the distribution being “closer” to this manifold.³ A perturbation that alters the joint distribution might move it towards two such manifolds simultaneously, constituting a positive dependence, or towards one while moving away from the other, a negative dependence. By observing how proximities to different CI manifolds change after these parameter perturbations, we obtain insights into the dependencies between different CI tests. In particular, we will demonstrate how identical causal structures with different parameters can exhibit different dependencies between their CI properties.

3.1 PARAMETER PERTURBATIONS

In linear additive Gaussian models, each variable V is associated with an additive, independent Gaussian noise term N_V . To simplify our examples, we will assume that N_V is chosen so that all random variables are standardized, meaning that $\rho_{A,B \mid C} = \text{Cov}(A, B \mid C)$.

Consider the following system with $\alpha \neq 0$:

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

According to this characterization, $\rho_{A,B} = \alpha$. Perturbing α allows us to directly change the covariance between A, B .

We now consider two CI properties jointly and give examples of both “positive dependence” and “negative dependence”. A *positive* dependence involves a perturbation *towards* one CI manifold that also moves the joint distribution *towards* another CI manifold. In contrast, a *negative* dependence involves a perturbation *towards* one CI manifold that moves the joint distribution *away from* another.

3.2 POSITIVE DEPENDENCE

Consider the following structural equations that capture a fully connected DAG: $A \rightarrow B \rightarrow C, A \rightarrow C$,

$$A = N_A, \quad B = \alpha_1 A + N_B, \quad C = \alpha_2 A + \beta B + N_C. \quad (4)$$

³This proximity is not necessarily Euclidean distance in a chosen parameterization. Section 4 will sharpen this using an information-geometric notion of closeness, which is captured by the covariance matrix for Gaussians.

The variances of the noise terms are again chosen so that the random variables are standardized. Writing each variable in terms of N_A, N_B, N_C ,

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

Using $\text{Var}(N_B) = 1 - \alpha_1^2$, the three correlations for this model are,

$$\rho_{A,B} = \alpha_1, \quad \rho_{A,C} = \alpha_2 + \alpha_1 \beta, \quad \rho_{B,C} = \alpha_1 \alpha_2 + \beta.$$

For the first example, let us assume $\alpha_2 = 0$, which corresponds to a simplified DAG $\mathcal{G} : A \rightarrow B \rightarrow C$. Notice that a perturbation lowering α_1 consequently lowers the dependencies between A, B and A, C simultaneously by reducing the magnitudes of both $|\rho_{A,B}|$ and $|\rho_{A,C}|$. This constitutes a *positive dependence*.

3.3 NEGATIVE DEPENDENCE

Now, consider the joint distribution P generated by

$$\alpha_1 = 0.5, \quad \alpha_2 = 0.3, \quad \beta = -0.3. \quad (5)$$

The value $\alpha_2 = 0.3$ fixes a specific base distribution P ; treating α_2 as an indeterminate in a neighborhood of this value lets us analyze how small perturbations move P relative to nearby CI manifolds. Concretely, we track the change in $\rho_{B,C}$ and $\rho_{A,C}$ as α_2 is perturbed around 0.3:

$$\begin{aligned} \rho_{A,C} &= \alpha_2 - 0.15 = 0.15, \\ \rho_{B,C} &= 0.5\alpha_2 - 0.3 = -0.15. \end{aligned} \quad (6)$$

Notice that $\rho_{A,C} > 0$ while $\rho_{B,C} < 0$. As a result, changing α_2 moves the absolute values of the two correlations in opposite directions. Therefore, in the neighborhood of P as given by the parameters in Equation (5), perturbations of α_2 that shift P towards $A \perp\!\!\!\perp C$ are unlikely to move the distribution “closer” to $B \perp\!\!\!\perp C$.

Remarks. Our example of positive dependence in Section 3.2 holds for *all* parameter specifications with $\alpha_2 = 0$ — i.e., distributions that follow the DAG structure $A \rightarrow B \rightarrow C$. That is, $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$ co-occur when perturbing α_1 regardless of the specification of $\beta > 0$. In contrast, the example of negative dependence is specific to perturbations of the distribution P specified by Eq. (5). For example, consider an alternative joint distribution Q generated using

$$\alpha_1 = 0.5, \quad \alpha_2 = 0.3, \quad \beta = 0.3. \quad (7)$$

Now, both correlations are positive, meaning that perturbations in α_2 move their absolute values in the same direction. Hence, although the same graphical model underlies P and

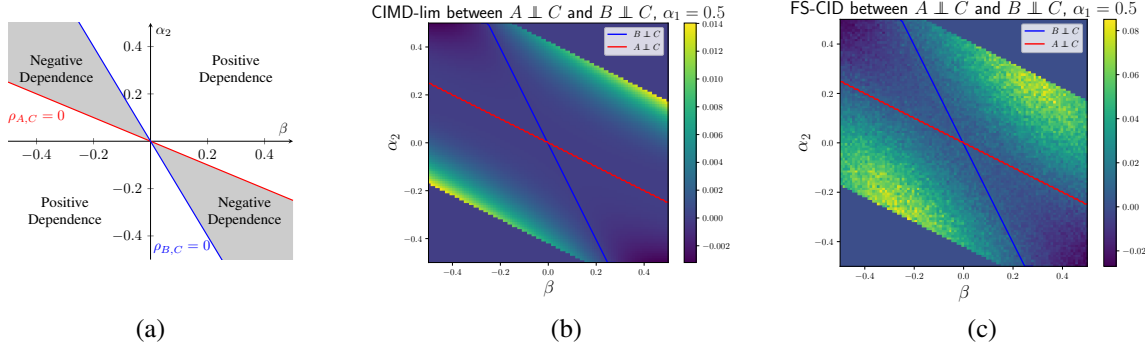


Figure 1: (a) Dependence of CI properties for structural equations given in Eq. (4) when $\alpha_1 = .5$. The shaded region between the red and blue lines corresponds to models where moving towards blue moves *away* from red. Hence, the shaded region has a “negative” dependence between $A \perp C$ and $B \perp C$, while the unshaded region has a “positive” dependence. (b) The introduced metric, CIMD (Section 4), measuring the meta-dependence between CI-tests, computed from the covariance matrix. (c) Estimation of CIMD with finite samples (Sections 4.3 and 5).

Q , these distributions differ in their dependencies between (local) conditional independencies. 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. We visualize the manifolds of independencies using a red line for $A \perp C$ and a blue line for $B \perp C$.

It is clear that we cannot rely solely on the graph structure when studying the dependence between CI tests under a random perturbation. Instead, we need to quantify 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. This quantity will not require access to the parameters of the structural equations.

4 QUANTIFYING META-DEPENDENCE

We now make meta-dependence precise. CIMD quantifies the shared information between two conditional independence tests via moment projections and requires no access to the causal graph. Section 5 will then CIMD to the co-occurrence of test outcomes under finite-sample uncertainty.

4.1 PROJECTING INTO INDEPENDENCE

To define CIMD, we first relate a joint distribution to a single conditional independence through moment projections.

Consider a graphical model on the two vertices $\{X, Y\}$. Restricting attention to directed acyclic graphs without latent confounders, a joint distribution P with $X \not\perp Y$ is modeled as $X \rightarrow Y$ or $X \leftarrow Y$, while a distribution with $X \perp Y$ is modeled by the *empty graph* on $\{X, Y\}$ (no

edges). The M-projection of P onto this empty graph is the product of its marginals, $\Pi_{X \perp Y}(P)$ (defined below), and the mutual information measures how far P sits from it in KL divergence:

$$I(X : Y) = D(P || \Pi_{X \perp Y}(P)).$$

This product-of-marginals construction is special to *marginal* independence; for a conditional independence $A \perp B | C$ we instead obtain the projection from a general characterization of projections onto graphs. The M-projection⁴ of P onto a set of distributions $\mathcal{P}$ is its KL-closest member,

$$\Pi_{\mathcal{P}}(P) := \operatorname{argmin}_{Q \in \mathcal{P}} D(P || Q), \quad (8)$$

and when $\mathcal{P} = \mathcal{P}_{\mathcal{G}}$ is the set of distributions Markov to a graph $\mathcal{G}$, it takes a factorized form.

Theorem 4.1 (Thm 8.7 in Koller and Friedman [2009]). *A moment projection of P onto $\mathcal{G}$ is given by the factorization implied by $\mathcal{G}$ using the conditional probabilities from P :*

$$[\Pi_{\mathcal{G}}(P)](\mathbf{v}) = \prod_{V_i \in \mathbf{V}} P(v_i | \mathbf{pa}(v_i)).$$

The left-hand side $\Pi_{\mathcal{G}}(P)$ is the projected distribution (at assignment $\mathbf{v}$) and the right-hand side is its factorization using the conditionals of the *original* P .

Theorem 4.1 projects onto a *graph*, not a single conditional independence. To reuse it for a CI, we introduce an *auxiliary graph*: a placeholder DAG chosen only to induce the target independence. For $A \perp B | C$, the fork $A \leftarrow C \rightarrow B$ suffices.⁵

⁴See Koller and Friedman [2009] Ch 8.5.

⁵Any Markov-equivalent choice, such as the chain $A \rightarrow C \rightarrow B$, gives the same projection.

Corollary 4.2. Let $\mathcal{P}_{A \perp B \mid C}$ be the set of distributions satisfying $A \perp B \mid C$, and realize the projection onto it through the auxiliary graph $A \leftarrow C \rightarrow B$. With $\mathbf{X} = \mathbf{V} \setminus (\{A, B\} \cup C)$ the remaining variables, the marginal of $\Pi_{A \perp B \mid C}(P)$ over $A, B, \mathbf{X}$ (with C integrated out) is

$$[\Pi_{A \perp B \mid C}(P)](a, b, \mathbf{x}) = \mathbb{E}_C[P(a \mid C)P(b \mid C)P(\mathbf{x} \mid a, b, C)].$$

Proof. $A \perp B \mid C$ implies the factorization $P(\mathbf{V}) = P(C)P(A \mid C)P(B \mid C)P(\mathbf{X} \mid A, B, C)$, i.e., the auxiliary graph $A \leftarrow C \rightarrow B$, to which we apply Theorem 4.1. $\square$

Crucially, this projection requires no knowledge of the underlying graphical structure.

4.2 CI META-DEPENDENCE

To measure the shared information between two CI tests, we ask whether the conditional mutual information for one test changes after projecting onto the other.

Definition 4.3 (CIMD). Consider two conditional independencies:

$$(T_1) \ A \perp B \mid C, \quad (T_2) \ A' \perp B' \mid C'.$$

The *CI Meta-Dependence (CIMD)* between them on an empirical distribution P is

$$\text{CIMD}(T_1, T_2, P) := I_P(A : B \mid C) - I_{\Pi_{T_2}(P)}(A : B \mid C).$$

Here I_P is the conditional mutual information under P , and $I_{\Pi_{T_2}(P)}$ is the same functional under $\Pi_{T_2}(P)$, the projection of P onto the null of T_2 (Eq. (8)). The two tests play different roles: T_1 fixes *which* mutual information is measured, $I(A : B \mid C)$, while T_2 fixes the projection applied beforehand. The variables need not be disjoint — A, B, C and A', B', C' may overlap, and the primed variables are not separate copies of the unprimed ones.

Reading a moment projection as the minimal perturbation of P onto a CI property, CIMD asks whether that perturbation also moves toward a second CI property. Zero CIMD means the two tests are “orthogonal”: the CI manifolds are perpendicular, so moving onto one leaves the distance to the other unchanged. Positive CIMD means projecting onto T_2 also moves P toward T_1 ’s null, so the manifolds point the same way; negative CIMD, which is rarer (Section 5), means they lie on opposite sides of P . Figure 1 illustrates both cases.

Remarks. We note three properties of CIMD. First, it is asymmetric: $\text{CIMD}(T_1, T_2, P) \neq \text{CIMD}(T_2, T_1, P)$ in general. In the DAG $A \rightarrow B \rightarrow C \rightarrow D$, perturbing toward $B \perp C$ forces $A \perp D$, since severing B – C disconnects

A from D ; but perturbing toward $A \perp D$ need not force $B \perp C$, as it may instead arise from a $C \perp D$ faithfulness violation. This fixes the order of computation: to ask whether T_1 is *redundant* once T_2 holds — e.g., when T_2 ’s outcome was already spent earlier in a sequence of tests — one computes $\text{CIMD}(T_1, T_2, P)$, where a large positive value means T_1 adds little beyond T_2 .

Second, CIMD is related to, but distinct from, Euclidean distance in the parameter space of the structural equation model: in Figure 1, the moment projections do not land on the nearest point of the red and blue lines but reach them by zeroing out correlations.

Third, CIMD depends on P : the same pair of tests can yield positive CIMD for one distribution and negative for another (Section 3.3).

We use KL divergence because its M-projection has a closed form in exponential families (Section 4.3) and it carries the directional “surprise” reading of Section 2, matching our intended meaning — how far a distribution must move to satisfy a test. A symmetric divergence could be substituted, at the cost of this tractability.

4.3 A GAUSSIAN CLOSED FORM

For multivariate Gaussians, both the moment projection and CIMD can be computed in closed form directly from the covariance matrix, with no access to the raw data. Let Σ be the covariance of P over $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{X}$ (with $\mathbf{X}$ the remaining variables), and let $\Sigma^\perp$ be the covariance of the projection $\Pi_{A \perp B \mid C}(P)$. By Corollary 4.2, $\Sigma^\perp$ agrees with Σ except in the two blocks involving $\mathbf{A} \cup \mathbf{B}$ and $\mathbf{X}$: one zeroes the $\mathbf{A}, \mathbf{B}$ conditional covariance given $\mathbf{C}$, the other propagates that change through $P(\mathbf{X} \mid \mathbf{A}, \mathbf{B}, \mathbf{C})$ (full construction in Appendix A). Letting Φ and $\Phi^\perp$ denote the conditional covariance of $\mathbf{Z} = \mathbf{A} \cup \mathbf{B}$ given $\mathbf{C}$ under P and under $\Pi_{A \perp B \mid C}(P)$ respectively, the Gaussian mutual-information formula gives

$$\begin{aligned} \text{CIMD}(T_1, T_2) = & \frac{1}{2} \log \left(\frac{\det(\Phi_{\mathbf{A}, \mathbf{A}}) \det(\Phi_{\mathbf{B}, \mathbf{B}})}{\det(\Phi)} \right) \\ & - \frac{1}{2} \log \left(\frac{\det(\Phi_{\mathbf{A}, \mathbf{A}}^\perp) \det(\Phi_{\mathbf{B}, \mathbf{B}}^\perp)}{\det(\Phi^\perp)} \right), \end{aligned} \quad (9)$$

where $\Phi_{\mathbf{A}, \mathbf{A}}$ and $\Phi_{\mathbf{B}, \mathbf{B}}$ reduce to scalars when A and B are single variables. Both terms depend on P only through Σ , so CIMD is a function of the covariance matrix alone.

5 EMPIRICAL TESTS

CIMD does not directly address finite sample deviations. However, KL divergence still measures statistical difference: a small $D(P \parallel Q)$ indicates that samples drawn from P are, on average, not very surprising under Q — equivalently, Q

assigns high likelihood to data typical of P . In keeping with the asymmetry of KL, this is a directional statement and does not imply the reverse. 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.⁶

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 variable. 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.⁷ 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

$$\text{FS-CID} := \hat{P}(t_1, t_2) - \hat{P}(t_1)\hat{P}(t_2), \quad (10)$$

where $\hat{P}$ are empirical estimates from 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; the interesting meta-dependence instead occurs near the CI-test violations, where tests actually flip between rejection and non-rejection under resampling. To focus on this regime, we zero CIMD out whenever either mutual information term is large (> 0.1 nats), and call the result “CIMD-lim.” The cutoff of 0.1 nats is a deliberately small amount of (conditional) mutual information: a term exceeding it corresponds to a dependence strong enough that the associated CI-test almost never fails to reject in FS-CID, so its meta-dependence is not resolved by the finite-sample simulation and would only add noise to the comparison.

5.1 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 $\alpha_2, \beta \in [-1, 1]$.⁸ These parameters correspond exactly to those used in Figure 1 (a), with results shown in (b) and (c).

⁶Code can be found on [GitHub](#).

⁷This hypothesis test is implemented in conditional-independence [Chandler Squires, 2018]. All experiments are conducted on 1 CPU.

⁸We use relatively small (20 samples) dataset sizes in order to induce a sufficient perturbation from the true distribution to the

Unlike the single-parameter perturbations studied in Section 3, finite-sample deviations extend in any direction, possibly 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.

5.2 REAL-WORLD 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.⁹

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. Much of the recovered meta-dependence concentrates on pairs of CI-tests that share variables; this is expected, since shared variables are precisely where a moment projection most strongly perturbs the dependencies that both tests probe. We therefore read Figure 2 as evidence that CIMD tracks FS-CID wherever meta-dependence is present, rather than as a claim about how prevalent such dependence is across arbitrary pairs of tests. The results for the other two datasets are given in Appendix C. These two datasets have more complicated meta-dependencies but still share signal across the two metrics.

5.3 TUNING CAUSAL DISCOVERY

As a proof of concept for using CIMD inside an algorithm, we apply it to the multiple-testing correction in constraint-based causal discovery. The PC algorithm removes an edge $\{A, B\}$ whenever *some* conditioning set $\mathbf{S}$ yields $A \perp\!\!\!\perp B | \mathbf{S}$, so each candidate edge is associated with an entire family of CI tests. A Bonferroni correction controls the family-wise error rate by dividing the significance level α by the total number of such tests, implicitly treating them as independent. Many of these tests are highly redundant, however: for a fixed edge, $A \perp\!\!\!\perp B | \mathbf{S}$ and $A \perp\!\!\!\perp B | \mathbf{S}'$ probe overlapping dependencies. We use CIMD to estimate this redundancy and relax the correction accordingly.

For each edge, let n_{edge} be the number of conditioning sets it empirical one to resolve a dependence.

⁹We use a relatively small sub-sample size (50 points) in order to induce a sufficient perturbation from the true distribution to the empirical one to resolve a dependence.

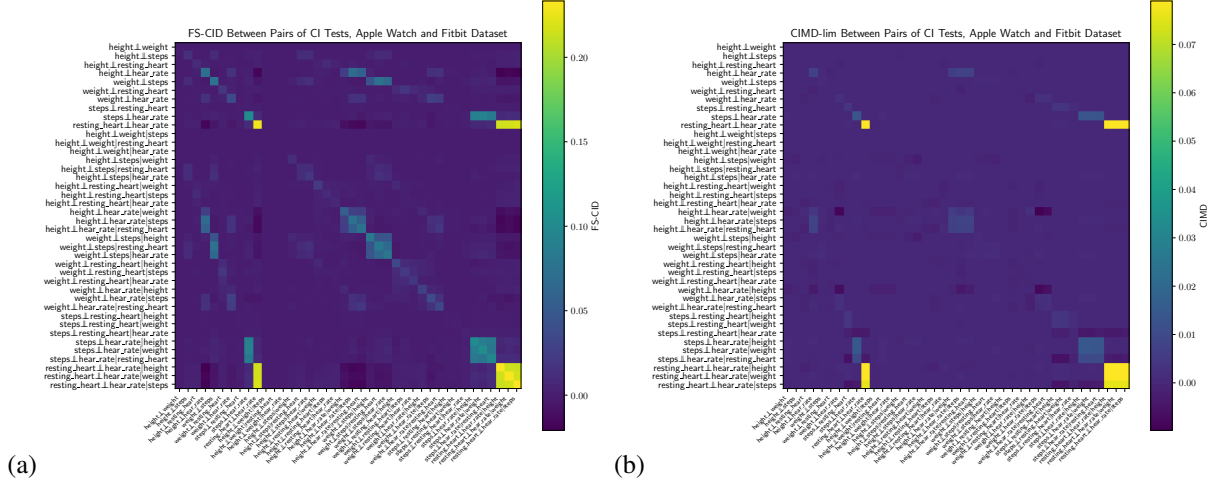


Figure 2: Meta-dependencies of CI-tests as measured by FS-CID in (a) and CIMD-lim in (b).

is tested against. We estimate a mean redundancy $\bar{r} \in [0, 1]$ by averaging, over sampled pairs of conditioning sets, the CIMD of Section 4 normalized by $I_P(a : b \mid S)$ and gated to the small-dependence regime exactly as the CIMD-lim of Section 5. We then replace n_{edge} with an *effective* test count $n_{\text{eff}} = n_{\text{edge}} (1 - \bar{r})$ in the Bonferroni denominator, so that edges with redundant conditioning-set tests receive a less stringent threshold; when $\bar{r} = 0$ the rule reduces exactly to Bonferroni. Because the normalized CIMD is a function of the covariance matrix (Section 4.3), the entire correction is computed from summary statistics of the data.

We evaluate on sparse linear-Gaussian DAGs ($p = 18$ variables, ≈ 1 expected parent per node, conditioning sets up to size 3), estimating the skeleton with a partial-correlation (Fisher-Z) test at $\alpha = 0.05$ and averaging over 100 random graphs at each sample size. We compare three edge-thresholding rules applied to the *same* skeleton search: uncorrected (“vanilla”), Bonferroni, and the CIMD-relaxed correction. Figure 3 reports the skeleton F_1 . The redundancy CIMD detects is substantial ($\bar{r} \approx 0.6$ – 0.7), so it treats only 30–40% of each edge’s conditioning-set tests as effective and relaxes the per-edge threshold accordingly. Two patterns emerge. First, uncorrected PC trades precision for recall: a fixed α leaves a roughly constant false-positive rate that more data does not remove, so its precision stays near 0.93 at every sample size while its recall climbs; its F_1 therefore rises only modestly (from 0.93 to 0.95) and plateaus below both corrections, which reach precision ≈ 0.99 and $F_1 \approx 0.97$ by $N = 5000$. The corrections overtake uncorrected PC once $N \geq 2000$; at the smallest samples their conservatism costs enough recall that uncorrected PC retains the highest F_1 . Second, and consistent with the modest claim of Section 6, CIMD-PC never underperforms Bonferroni-PC and improves on it exactly where the two differ most — the small-sample regime: at $N = 500$ it recovers more true edges at matched precision, giving $\Delta F_1 = +0.015 \pm 0.002$

and 0.47 ± 0.06 fewer skeleton errors, with the gap closing to zero by $N \geq 2000$. This setting is deliberately favorable to correction — a sparse graph with a large, redundant family of conditioning-set tests — so we present it as evidence that CIMD’s redundancy estimate can be turned into a usable, distribution-aware multiplicity correction, not as a claim of universal improvement.

6 CONCLUSION

This paper introduces the study of CI Meta-Dependence (CIMD), an information-theoretic measure of the shared information between two tests for conditional independence. Our concrete contributions are twofold. First, we show empirically — on both synthetic and real-world data (Section 5) — that CIMD aligns with FS-CID, the co-occurrence of CI-test outcomes under finite-sample uncertainty. Second, we derive a closed form for CIMD in the multivariate Gaussian case (Section 4.3) that is computable directly from a covariance matrix, making the measure tractable in practice. The definition itself (Section 4) is general and applies to any distribution that factorizes over a graph; only the closed-form computation is specific to the Gaussian setting.

Beyond these results, CIMD suggests several directions for algorithms that rely on sequences of CI tests. We describe them below, distinguishing what the present paper supports from what remains open.

Significance Thresholds. Constraint-based causal discovery employs many CI tests, many of which correspond to removing an edge between two variables. Multiple-testing corrections (e.g., Bonferroni) typically treat these tests as if they were independent, tightening each significance threshold in proportion to the total number of tests. When two tests share information — and therefore tend to co-occur — treating them as independent is overly conservative. CIMD

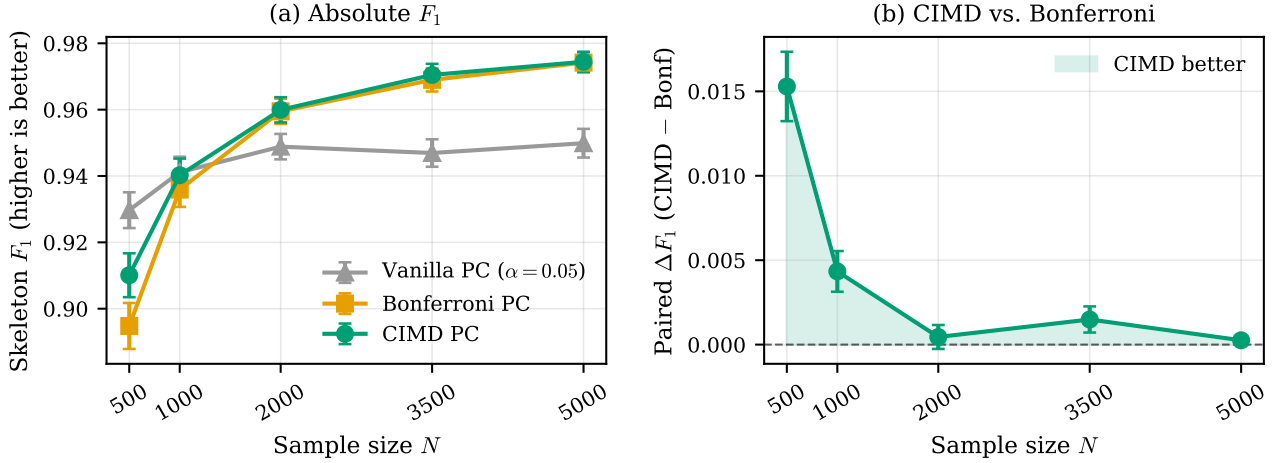


Figure 3: Using CIMD to relax the multiple-testing correction in PC-based causal discovery, on sparse linear-Gaussian DAGs ($p = 18$, averaged over 100 random graphs; error bars are ± 1 SEM). (a) Skeleton F_1 for uncorrected PC, Bonferroni-corrected PC, and CIMD-corrected PC; both corrections overtake uncorrected PC as N grows, while uncorrected PC’s precision plateaus. (b) Paired difference $\Delta F_1 = F_1^{\text{CIMD}} - F_1^{\text{Bonf}}$ per graph: CIMD-PC improves on Bonferroni-PC at small N and coincides with it for $N \geq 2000$.

quantifies this redundancy directly from the distribution, which suggests replacing the uniform correction with a redundancy-aware one. We make a deliberately modest claim: *when a multiple-testing correction is going to be applied anyway*, accounting for the redundancy measured by CIMD is a strict improvement over assuming independence between tests. We demonstrate this concretely in Section 5.3: replacing Bonferroni’s test count with a CIMD-based effective count matches or slightly improves the skeleton recovered by the PC algorithm, at no cost in precision. We do not claim it delivers exact finite-sample error control, which is infeasible even for a single CI test [Shah and Peters, 2020].

Validating Causal Discovery. Validating causal discovery requires a distance between the true ($\mathcal{G}$) and recovered ($\hat{\mathcal{G}}$) graphs, such as the Structural Hamming Distance [de Jongh and Druzdzel, 2009] or intervention-based metrics [Peters and Bühlmann, 2015, Henckel et al., 2024, Wahl and Runge, 2025]. Because CI meta-dependence is distribution-specific, a fixed structural difference can be more or less detectable across different distributions, so such graphical error measures should be read with the distribution in mind. CIMD instead quantifies *co-occurring* edge errors: a single distribution-driven mistake may matter as much as many co-occurring ones. Projecting onto a manifold for a *full* graphical structure (Section 4) could then yield distribution-aware DAG losses that focus validation on statistically significant errors.

Uncertainty in Causal Discovery. CIMD is a step toward quantifying higher-order uncertainty in causal structure learning. While methods do exist to quantify uncertainty in CI 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]. CIMD provides a moment-based (or covariance-based) metric that aligns with sub-sampling dependence (FS-CID). Since CIMD does not require access to the raw data, CIMD may be useful in large data and security-critical settings.

Future Work. Several questions remain open. (i) *Non-Gaussian computation.* Our closed form is specific to multivariate Gaussians; computing CIMD more generally requires estimating moment projections and KL divergences non-parametrically, for instance via kernel-based mutual-information estimators or normalizing-flow projections onto CI manifolds. Characterizing the difficulty and accuracy of these estimators is an important next step. (ii) *A formal CIMD-FS-CID link.* The agreement between CIMD and FS-CID is, at present, established only empirically. Making it rigorous appears to require controlling the local curvature of the CI manifolds around the true distribution, whose geometry has so far been characterized mainly through its singularities [Uhler et al., 2013]. Our empirical agreement is consistent with this curvature being mild in typical neighborhoods, but we do not prove it, and we regard closing this gap as the central open question raised by this work. (iii) *Gaussian finite-sample guarantees.* The Gaussian setting is the most promising place to make progress: because the partial-correlation (Fisher-Z) test is level and has known power, one could aim for a finite-sample statement expressing the joint rejection probability of two CI tests as a function of CIMD plus an error term controlled by the distance to the true distribution.

Acknowledgements

We thank Murat Kocaoglu, Abhinav Kumar, Beshr Islam Bouli, and the anonymous reviewers for their helpful discussions. B.M. is partially supported by the Advanced Research Concepts (ARC) COMPASS program, sponsored by the Defense Advanced Research Projects Agency (DARPA) under agreement number HR001-25-3-0212. J.Z. was partially supported by the Eric and Wendy Schmidt Center. C.U. was partially supported by NCCIH/NIH (1DP2AT012345), ONR (N00014-24-1-2687), DOE-ASCR (DE-SC0023187), and the Eric and Wendy Schmidt Center at the Broad Institute.

References

- Chandler Squires. *causal dag: creation, manipulation, and learning of causal models*, 2018. URL <https://github.com/uhlerlab/causal dag>.
- 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.
- 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>.
- 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.
- Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Saini, Saurabh Bagchi, and Murat Kocaoglu. Root Cause Analysis of Failures in Microservices through Causal Discovery. *Advances in Neural Information Processing Systems*, 35:31158–31170, December 2022.
- Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. 2013.
- Markus Kalisch and Peter Bühlmann. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *The Journal of Machine Learning Research*, 8:613–636, May 2007. ISSN 1532-4435. URL <https://dl.acm.org/doi/10.5555/1248659.1248681>.
- Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- 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.
- Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.
- Bijan Mazaheri, Atalanti Mastakouri, Dominik Janzing, and Michaela Hardt. Causal information splitting: engineering proxy features for robustness to distribution shifts. In *Uncertainty in Artificial Intelligence*, pages 1401–1411. PMLR, 2023.
- 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 and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579–595, November 2014. ISSN 0883-4237, 2168-8745. doi: 10.1214/14-STS486.
- 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? *Probabilistic and Causal Inference*, 1986. URL <https://api.semanticscholar.org/CorpusID:262089256>.
- 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>.

- Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.
- Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1): 145–157, 2003.
- Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36): 1–34, 2018.
- Leonard J Schulman and Piyush Srivastava. Stability of causal inference. In *UAI*, 2016.
- Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), June 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1857.
- 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.
- Ioannis Tsamardinos and Constantin F Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *International Workshop on Artificial Intelligence and Statistics*, pages 300–307. PMLR, 2003.
- 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.
- Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D’ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.
- Jonas Wahl and Jakob Runge. Separation-based distance measures for causal graphs. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL <https://openreview.net/forum?id=K07fATqh2W>.
- Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhao-long Ling, and Xindong Wu. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)*, 53(5):1–36, 2020.
- Jiaqi Zhang, Louis Cammarata, Chandler Squires, Themistoklis P Sapsis, and Caroline Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, pages 1–10, 2023.
- 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.
- Jiji Zhang and Peter L Spirtes. Strong faithfulness and uniform consistency in causal inference. *arXiv preprint arXiv:1212.2506*, 2012.

Meta-Dependence in Conditional Independence Testing (Supplementary Material)

Bijan H. S. Mazaheri^{1,2}

Jiaqi Zhang^{2,3}

Caroline Uhler^{2,3}

¹Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire, USA

²Schmidt Center, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

³Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

A DERIVATION OF THE GAUSSIAN CLOSED FORM

Here we derive the closed form of Eq. (9). For Gaussian vectors, the moment projection and CIMD are obtained by directly modifying the covariance of P over $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{X}$,

$$\Sigma = \begin{bmatrix} \Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{A} \cup \mathbf{B}} & \Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{C}} & \Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{X}} \\ \Sigma_{\mathbf{C}, \mathbf{A} \cup \mathbf{B}} & \Sigma_{\mathbf{C}, \mathbf{C}} & \Sigma_{\mathbf{C}, \mathbf{X}} \\ \Sigma_{\mathbf{X}, \mathbf{A} \cup \mathbf{B}} & \Sigma_{\mathbf{X}, \mathbf{C}} & \Sigma_{\mathbf{X}, \mathbf{X}} \end{bmatrix}.$$

Writing $P^\perp := \Pi_{\mathbf{A} \perp \mathbf{B} | \mathbf{C}}(P)$, Corollary 4.2 gives $P^\perp(\mathbf{C}) = P(\mathbf{C})$, $P^\perp(\mathbf{A} | \mathbf{C}) = P(\mathbf{A} | \mathbf{C})$, $P^\perp(\mathbf{B} | \mathbf{C}) = P(\mathbf{B} | \mathbf{C})$, and $P^\perp(\mathbf{X} | \mathbf{A}, \mathbf{B}, \mathbf{C}) = P(\mathbf{X} | \mathbf{A}, \mathbf{B}, \mathbf{C})$. Hence the covariance $\Sigma^\perp$ of $P^\perp$ matches Σ except in the sub-matrices $\Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{A} \cup \mathbf{B}}^\perp$ and $\Sigma_{\mathbf{X}, \mathbf{X}}^\perp$.

First, $\Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{A} \cup \mathbf{B}}$ is changed so that $\mathbf{A} \perp \mathbf{B} | \mathbf{C}$, by setting the conditional covariance (the Schur complement of block $\mathbf{C}$) to zero and rearranging:

$$\Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{A} \cup \mathbf{B}}^\perp = \Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{C}} (\Sigma_{\mathbf{C}, \mathbf{C}})^{-1} \Sigma_{\mathbf{C}, \mathbf{A} \cup \mathbf{B}}.$$

Next we regenerate $\Sigma_{\mathbf{X}, \mathbf{X}}$ using the new distribution on $\mathbf{A}, \mathbf{B}$ and the same conditional $P(\mathbf{X} | \mathbf{A}, \mathbf{B}, \mathbf{C})$. With $\bar{\mathbf{X}} := \mathbf{A} \cup \mathbf{B} \cup \mathbf{C}$,

$$\Sigma_{\bar{\mathbf{X}}, \bar{\mathbf{X}}}^\perp = \begin{bmatrix} \Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{A} \cup \mathbf{B}}^\perp & \Sigma_{\mathbf{A} \cup \mathbf{B}, \mathbf{C}} \\ \Sigma_{\mathbf{C}, \mathbf{A} \cup \mathbf{B}} & \Sigma_{\mathbf{C}, \mathbf{C}} \end{bmatrix},$$

and, when invertible,

$$\begin{aligned} \Omega_{\bar{\mathbf{X}}, \bar{\mathbf{X}}} &:= (\Sigma_{\bar{\mathbf{X}}, \bar{\mathbf{X}}}^\perp)^{-1} - \Sigma_{\bar{\mathbf{X}}, \bar{\mathbf{X}}}^{-1}, \\ \Sigma_{\mathbf{X}, \mathbf{X}}^\perp &= \Sigma_{\mathbf{X}, \mathbf{X}} + \Sigma_{\mathbf{X}, \bar{\mathbf{X}}} \Omega_{\bar{\mathbf{X}}, \bar{\mathbf{X}}} \Sigma_{\bar{\mathbf{X}}, \mathbf{X}}. \end{aligned}$$

Finally, we form the conditional covariance of $\mathbf{Z} = \mathbf{A} \cup \mathbf{B}$ given $\mathbf{C}$ for the original distribution (Φ) and the projected distribution ($\Phi^\perp$):

$$\begin{aligned} \Phi &= \Sigma_{\mathbf{Z}, \mathbf{Z}} - \Sigma_{\mathbf{Z}, \mathbf{C}} (\Sigma_{\mathbf{C}, \mathbf{C}})^{-1} \Sigma_{\mathbf{C}, \mathbf{Z}}, \\ \Phi^\perp &= \Sigma_{\mathbf{Z}, \mathbf{Z}}^\perp - \Sigma_{\mathbf{Z}, \mathbf{C}}^\perp (\Sigma_{\mathbf{C}, \mathbf{C}}^\perp)^{-1} \Sigma_{\mathbf{C}, \mathbf{Z}}^\perp. \end{aligned}$$

Substituting Φ and $\Phi^\perp$ into the closed form for the mutual information between multivariate Gaussians yields Eq. (9).

B SYNTHETIC DATA EXPERIMENTS ON OTHER GRAPH STRUCTURES

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

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

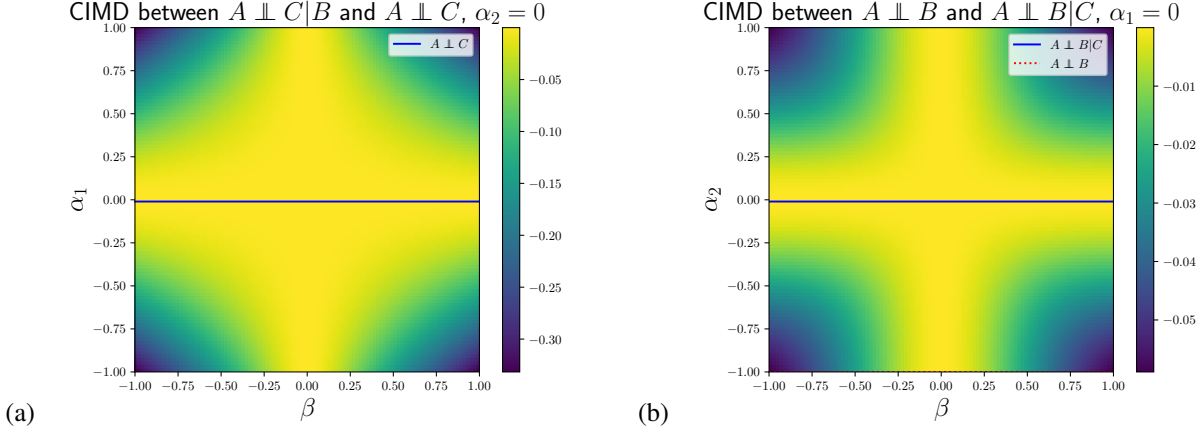


Figure 4: Both (a) and (b) plot the CIMD for various parameters from the structural equations given in Eq. (4). The effective structure in (a) is a Markov chain, and the effective structure in (b) is a collider.

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

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

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

C REAL-WORLD DATASET CI META-DEPENDENCE

Figure 5 shows four plots on two 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.

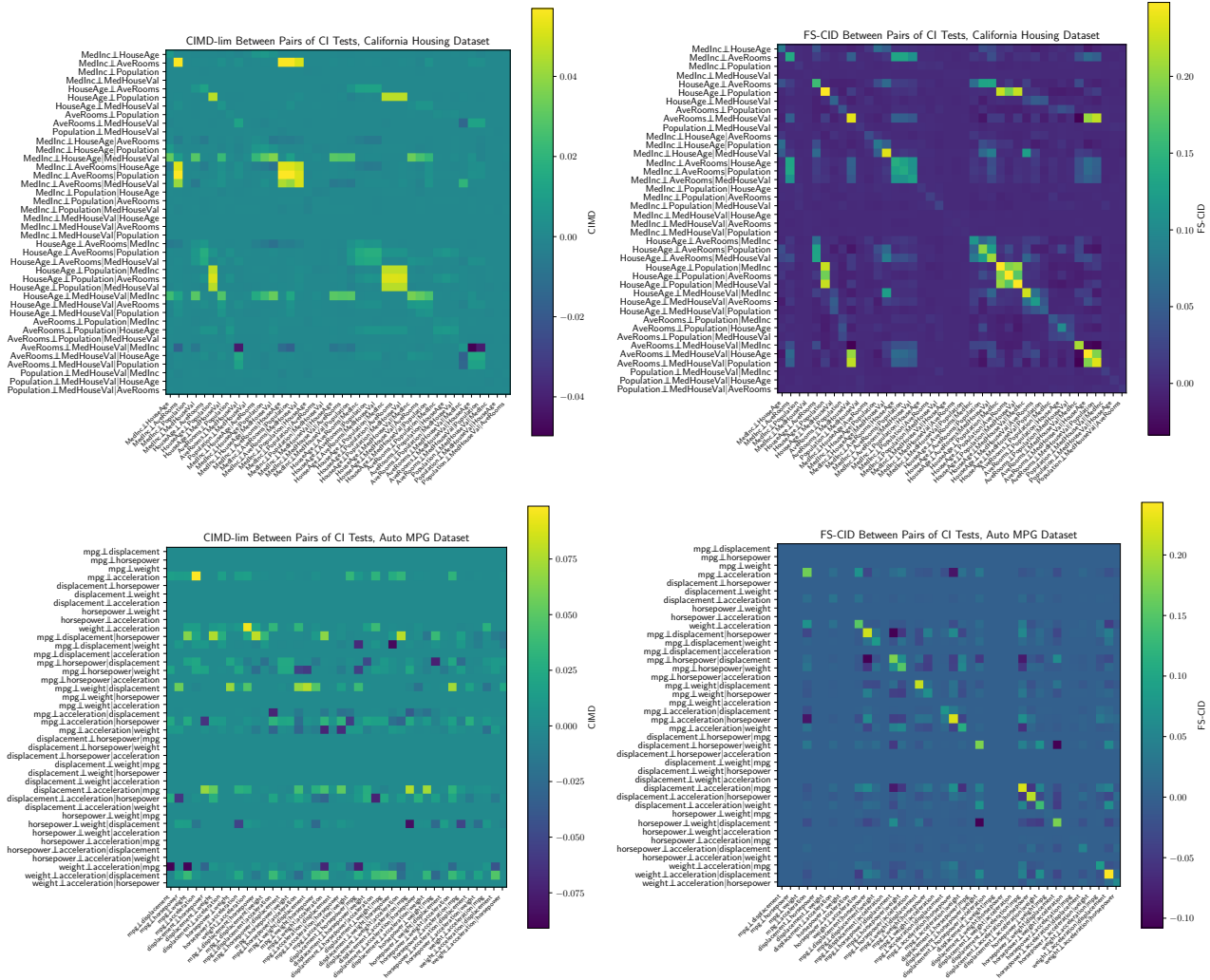


Figure 5: Demonstrating the dependence between CI-tests in real-world data using both FS-CID and limited CIMD.