Discovering Invariant and Changing Mechanisms from Data

Sarah Mameche
sarah.mameche@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

David Kaltenpoth
david.kaltenpoth@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Jilles Vreeken
jv@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

ABSTRACT

While *invariance* of causal mechanisms has inspired recent work in both robust machine learning and causal inference, causal mechanisms often *vary* over domains due to, for example, population-specific differences, the context of data collection, or intervention. To discover invariant and changing mechanisms from data, we propose extending the algorithmic model for causation to mechanism changes and instantiating it via Minimum Description Length. In essence, for a continuous variable Y in multiple contexts C, we identify variables X as causal if the regression functions $g : X \to Y$ have succinct descriptions in all contexts. In empirical evaluations we show that our method, Vario, reveals mechanism changes, discovers causal variables by invariance, and finds causal networks, such as on real-world data that gives insight into the signaling pathways in human immune cells.

CCS CONCEPTS

• **Mathematics of computing** → **Causal networks**; **Exploratory data analysis**; **Information theory**; *Regression analysis.*

KEYWORDS

invariance, mechanism changes, causal inference, information theory, regression

ACM Reference Format:
Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. 2022. Discovering Invariant and Changing Mechanisms from Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539479

1 INTRODUCTION

We consider the task of finding structural causal mechanisms f for a variable Y when data comes from heterogeneous sources – such as different hospitals, experimental conditions, or moments in time. The distribution of covariates X changes in these environments, violating the commonly taken i.i.d. assumption. As an added difficulty, the variable Y may *itself* be subject to changes – whether due to confounding factors, sampling bias, or external intervention.

Here, we propose a causal model that permits each context c a separate structural mechanism f^c. That is, in contrast to the literature [23, 28], we do not assume there exists a *single* invariant mechanism f which causes Y in the same way in *all* contexts but acknowledge that external factors unknown to us may influence its distribution in practice. For example, in medical treatments, we may see local variations and group-specific trends; in fMRI measurements, data collection conditions affect the causal strengths; and lastly, experts often study biological systems under different conditions, manipulating parts of the system through controlled interventions [15, 35, 40].

We illustrate this situation in Fig. 1 for three contexts and variables under distribution shift (lightning bolts). The same graphical causal model (DAG) applies in each condition. As the structural causal model (SCM) shows, however, the generating process f of Y is different in the third context. We are interested in discovering the partitioning Π of contexts into those groups that share an invariant mechanism for Y. By discovering Π, we find out whether there is a single or whether there are multiple mechanisms at play, what these are, how they differ, and where they apply. Aside from informing us in which contexts the generating process is the same and hence what datasets one can safely pool for learning, we show that we can use the discovered partitions to establish causality.

To achieve this, we build upon the algorithmic model of causation [18] and extend it to multiple contexts and mechanisms. That is, we adopt an information-theoretic perspective and identify the *true causal mechanisms* as those that allow the *simplest*, as measured by Kolmogorov complexity, description of the conditional distribution of Y given X across all contexts.

We show that we can achieve a computable score via the Minimum Description Length (MDL) principle [6]. As a proof of concept, we instantiate our framework for linear functional models. We introduce the Vario algorithm, which divides contexts into groups by invariance and, if possible, uses these invariances to tell apart causal and non-causal variables.

We give an overview of our approach in Fig. 1 (dashed parts). Our algorithmic causal model (ACM, top) represents each cause-effect relationship by a set of mechanisms. Our approach (bottom) is two-pronged. Using linear models g to approximate the generating process f, we partition the contexts by the similarity of the models. Based on the partitions found for different variable sets and their scores, we discover the causal variables, here X_1.

On synthetic and real-world data, we confirm that our method finds stable mechanisms, intervention targets, and causal variables even in the presence of nonlinearities. On the protein interaction dataset by Sachs et al. [35], Vario outperforms existing approaches for causal discovery from multiple environments.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539479

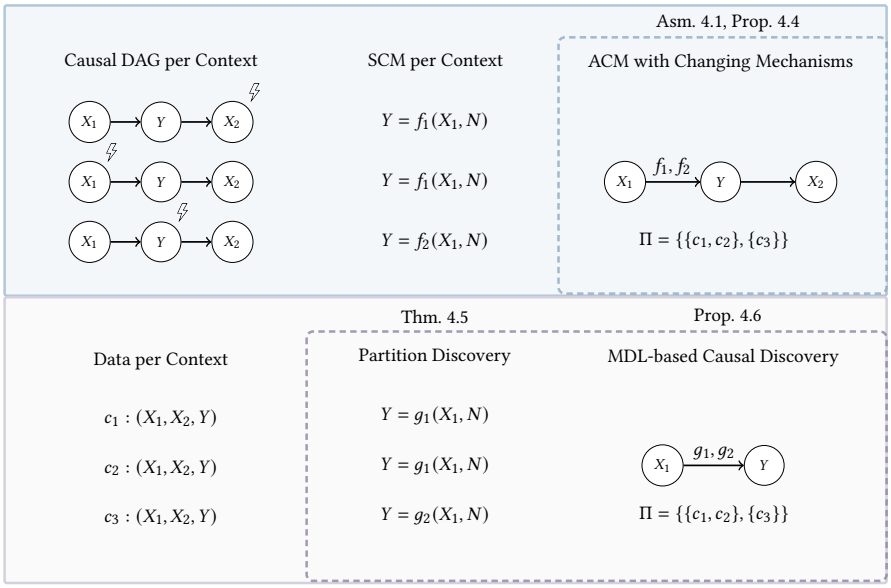


Figure 1: *Causal Model (top):* We consider multiple contexts that share the causal graph, but which may be observed under distribution shift or intervention (lightning bolts). For a target variable Y we are interested in discovering the *invariant* causal mechanisms (f_1) as well as where the generating process *changes* (f_2). *Approach (bottom):* From data, VARIO partitions contexts into groups (Π) based on function similarity, and discovers causal variables (X_1) based on invariance in groups.

We organize the remainder of this paper as follows. Next, we cover related work in Section 2. We introduce notation and preliminaries in Section 3. In Section 4, we develop the theory of our approach by defining our causal model, Π-invariance, and a practical score, and we discuss how to use those for causal discovery. In Section 5 we present the VARIO algorithm for discovering partitions and Π-invariant sets of covariates. We empirically evaluate VARIO and its competitors on synthetic and real-world data in Section 6, and wrap up with a discussion and conclusions in Section 7.

2 RELATED WORK

Causal discovery from observational data is an actively studied problem, and almost all approaches to it are either constraint-based, such as the FCI [39] and PC algorithms [38], or score-based, such as GES [3]. Traditional methods consider identically distributed data. There is growing literature on causal inference in the *non-i.i.d.* case. This includes nonstationary [14] as well as *heterogenous* data as we assume it, collected from different contexts or environments. We assume that the assignment of observations to contexts is known, and further that we observe an identical set of variables; see Huang et al. [13, 15] and Hu et al. [12] for relaxations of this.

Approaches for causal inference from multiple contexts generally fall into one of two categories: combining locally learned models or learning a model jointly from all data. The former include graph merging approachs [9, 44], model averaging [37], or direct estimation of differences in causal graphs [45]. Approaches that discover a shared causal model do so by considering in one form or another the principle of *independent mechanisms* [16, 29, 36, 41, 47].

One variant of this principle is *invariance*, the idea that causal mechanisms remain the same when the distribution of the cause changes. Peters et al. [2, 28] are the main proponents of invariance as an asymmetry between causal and non-causal associations. Their approach to Invariant Causal Prediction (ICP) comprises a subset search over X, and an intersection of those subsets that satisfy invariance w.r.t. a target Y, which they prove to be a subset of the causal variables for Y. Extensions of ICP cover the nonlinear case [11], and accommodate an influence of contexts on Y in the form of hidden variables [34].

Several authors [2, 4, 10, 23, 32] propose to leverage invariance for out-of-distribution learning, suggesting to prefer invariant predictors for robustness. As it is generally easier to identify invariant dependencies than strictly causal ones, several approaches to risk minimization make use of invariance [1, 20, 33].

More recent approaches go beyond invariant causal relations and consider *mechanism changes*. As such, the Joint Causal Inference (JCI) framework [27] introduces a context variable C to the causal graph, making traditional causal discovery algorithms applicable to multiple contexts; the method CD-NOD [47] adapts the PC algorithm specifically. Other authors discuss structure discovery under uncertain interventions [5, 7, 17, 40, 43] with the aim to discover causal DAGs in the interventional Markov Equivalence Class [8]. Key methods include GIES [7], which assumes known intervention targets, and UT-IGSP [40] for uncertain interventions.

This work addresses a general setting where mechanism changes or interventions with unknown types and targets may exist. Instead of conditional independence testing [27, 40, 47], we use functional modeling to be able to use the invariance asymmetry explicitly. In particular, we rely on the *algorithmic* causal modeling framework and the associated model inference criteria [18, 21, 36] which allow to identify causal relationships beyond Markov Equivalence [24, 26].

3 PRELIMINARIES

In the following, we define the notation and basic concepts we use throughout the paper.

3.1 Invariance in Causal Models

We consider continuous random variables $X = \{X_1, \dots X_n\}$, writing $Y \in X$ to refer to a designated variable of interest. We measure X in m different contexts $C = \{c_1, \dots, c_m\}$, e.g. corresponding to data collected in different hospitals. Writing C for the random variable denoting the context, we hence collect data from a joint distribution $P(X, C)$. We assume that the directed acyclic graph (DAG) $\mathcal{G}$ describing the causal relationships is the same in all contexts but that the structural equation can differ between them,

$$X_i = f_i^c(\mathrm{pa}_{\mathcal{G}}(X_i), N_i) \text{ for all } i \ ,$$

where $N_i \perp\!\!\!\perp (X, C)$ is a noise variable independent of X and C. In other words, $P(X \mid C = c)$ can differ for different contexts c.

We are interested in finding out in which contexts the mechanisms f_i^c differ and where they remain the same. The case where $f_i^c = f_i$ for all contexts c is commonly known as *invariance* [2, 28]. Invariance is based on the principle that the mechanism f_i causing X_i should be independent of $P(\mathrm{pa}_{\mathcal{G}}(X_i))$, and hence that it remains in place when the distribution of the covariates changes [28, 41]. While in principle, physical mechanisms stay the same, we want to model interventional and observational data alongside, or account for variance over different contexts due to limited samples. We would therefore like to tell between *which* contexts invariance holds and between which there are differences in treatment effects. In the latter case, f_i changes independently of $P(\mathrm{pa}_{\mathcal{G}}(X_i))$, which is commonly known as the *independent change* principle [41, 47].

To find sets of contexts where invariance holds as well as those where it does not, we take an information-theoretic approach based on Kolmogorov complexity and the algorithmic model of causality.

3.2 Kolmogorov Complexity

The Kolmogorov complexity of an object x is the length of the shortest program p for a universal Turing machine $\mathcal{U}$ that computes x and halts [22]. That is, p is the shortest possible description, or, optimal lossless compression of x. Formally, we have

$$K(x) = \min_{p \in \{0,1\}^*} \{|p| : \mathcal{U}(p) = x\} \ .$$

For a distribution P, it is defined as the length of the shortest program that approximates P to within any precision $1/q$,

$$K(P) = \min_{p \in \{0,1\}^*} \{|p| : |\mathcal{U}(p, x, q) - P(x)| \leq \frac{1}{q}\} \ .$$

Kolmogorov complexity is an integral concept in algorithmic causal modelling, which we move to next.

3.3 Algorithmic Model of Causality

In the algorithmic model of causation [18], causal mechanisms are considered *programs* that operate on objects. For variables X, Y it states that if X causes Y, denoted $X \to Y$, then

$$K(P(X)) + K((P(Y \mid X)) \leq K(P(Y)) + K((P(X \mid Y))) \ .$$

In general, this implies the algorithmic variant of the well-known Markov condition, introduced by Janzing and Schölkopf [18, 21].

Postulate 3.1. *[18] (Algorithmic Markov Condition). A DAG $\mathcal{G}$ formalizing the causal structure of X is only acceptable as the true causal structure if*

$$K(P(X_1, \dots X_n)) \stackrel{+}{=} \sum_i K(P(X_i \mid pa_{\mathcal{G}}(X_i)) \ .$$

where $\stackrel{+}{=}$ denotes equality up to an additive constant.

In particular, if one can state the true causal model as a DAG over X, the above principle permits inferring all causal directions. It states that the true causal model corresponds to the simplest, in terms of Kolmogorov complexity, factorization of the joint distribution. Consequently, it allows deciding between DAGs in the same Markov Equivalence Class [18]. It has inspired approaches to various causal inference problems [19, 24, 26].

Unlike our setting where data comes from multiple environments, all of these approaches rely on the idea that the data X are i.i.d., i.e., from the same environment. Therefore, we next develop an extension of the algorithmic framework suitable for our task.

4 THEORY

We now extend the algorithmic model of causality to multiple contexts and state its properties. We then develop a practical approach based on the MDL principle and finally give results on identifiability.

4.1 Causal Models with Changing Mechanisms

We assume that the causal DAG $\mathcal{G}$ is the same for all environments to capture invariant structure. For each variable Y, however, multiple causal mechanisms $f^c : \mathrm{pa}_{\mathcal{G}}(Y) \to Y$ may exist that govern it in different contexts.

The mechanisms f^c can be shared within a group π_k of multiple contexts, which we specify using a partition Π of C into groups. That is, $\Pi = \{\pi_k\}_k$, for which $\bigcup_k \pi_k = C$ and $\pi_i \cap \pi_j = \emptyset$ for $i \neq j$. We write $\Pi(c)$ to denote the group π_k containing c, $c \in \pi_k$. Overall, we state our causal model as follows.

Assumption 4.1 (Causal Model with Changing Mechanisms)**.** *Given a DAG $\mathcal{G}$ for X as well as partitions Π_i of C for each variable X_i, we consider the structural equation model*

$$X_i = f_i^{\Pi_i(C)}(pa_{\mathcal{G}}(X_i), N_i) \ ,$$

where $N_i \perp\!\!\!\perp (pa_{\mathcal{G}}(X_i), C)$ is independent of $pa_{\mathcal{G}}(X_i)$ and C.

An example is given in Fig. 1, where the variable of interest Y is governed by a different structural equation in the intervened context c_3, so that it has the partition $\Pi = \{\{c_1, c_2\}, \{c_3\}\}$.

In this example, the same function $Y = f_1(X_1, N)$ is applied to X_1 for both sets c_1, c_2, i.e. X_1 respects the structure of the partition Π of Y. We will call this Π-invariance of X_1 w.r.t. Y according to the following definition.

Definition 4.2 (Π-invariance)**.** *Given a DAG $\mathcal{G}$ and a partition Π_i of C for X_i, we call a set $S \subset X$ Π-invariant w.r.t. X_i, denoted as $S \in \mathcal{I}_\Pi(X_i)$, if for all $c_1, c_2 \in C$ with $\Pi_i(c_1) = \Pi_i(c_2)$ we have*

$$P(X_i \mid S, C = c_2) = P(X_i \mid S, C = c_1) \ .$$

In particular, the causal parents of each variable are Π-invariant.

Proposition 4.3 (Π-invariance for causal variables)**.** *If the generating model is as in Assumption 4.1 with DAG $\mathcal{G}$ where a variable Y has partition Π, then Π-invariance holds for the causal parents of Y,*

$$pa_{\mathcal{G}}(Y) \in \mathcal{I}_{\Pi}(Y) \ .$$

We now revisit the independent change (IC) principle. To do so we extend the algorithmic Markov condition from Postulate 3.1 to the case of multiple contexts. We arrive at the following criterion as to which partition Π and causal parents $X = \mathrm{pa}_{\mathcal{G}}(Y)$ are plausible for a given variable Y.

Proposition 4.4 (Algorithmic Markov Condition for Changing Mechanisms)**.** *Let $\mathcal{G}$ be a causal DAG over X with causal model with partitions Π_i of C for each X_i as in Assumption 4.1. Then the pair $(\mathcal{G}, \{\Pi_i\})$ is acceptable only if for all i*

$$\Pi_i, pa_{\mathcal{G}}(X_i) = \arg\min_{\Pi, S} \sum_{\pi \in \Pi} K(P(X_i \mid S, C \in \pi)). \tag{1}$$

That is, the true causal model leads to a factorization of P into distributions $P(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i), C \in \pi)$ that can be described independently of non-causal variables for a given target, as well as independently of the context in a given group.

Guided by the above properties, we now develop our approach for causal discovery in multiple contexts.

4.2 Discovering Invariances using MDL

We next explain how to use the algorithmic Causal Model over multiple contexts to discover changing mechanisms from data.

While Kolmogorov complexity is not computable [22], it can be approximated from above using the Minimum Description Length (MDL) principle [6]. Rather than taking the minimum over all programs p, we instead consider the best compression of the data using a restricted model class $\mathcal{M}$ of programs. Writing $L(x, M)$ for the compression of X using model $M \in \mathcal{M}$, we have $K(x) \leq \min_{M \in \mathcal{M}} L(x, M)$ with equality if $\mathcal{M}$ contains all programs.

We here use a two-part score $L(x, M) = L(M) + L(x \mid M)$, i.e. we separately encode the model and the data given the model [6].

For a given partition Π and potential covariates S for Y, we write the score as $L(Y, M_Y(\Pi, S))$, but we will drop the dependency of M on Y, S, Π when clear from the context. We have

$$L(Y \mid M) + L(M) = \sum_c L(Y \mid M_c) + \left(\sum_c L(M_c \mid M_\Pi) + L(M_\Pi) \right)$$

where M_c is the best-fitting model for Y in a single context. Note that $L(Y \mid M)$ does not depend on Π itself, but only on the local models M_c. Equivalently, Π depends on Y only through the models M_c. We can therefore ignore the data cost $L(Y \mid M)$ and find the partition Π^* satisfying

$$\begin{aligned} \Pi^* &= \arg\min_{\Pi} L(M) + L(Y \mid M) \\ &= \arg\min_{\Pi} L(\Pi) + L(M_\Pi \mid \Pi) + \sum_c L(M_c \mid M_\Pi) \ . \end{aligned}$$

To encode the model M, we must now make assumptions on the model class used. To do so, we use linear functions $g^c(X) = \alpha_c^t X + \alpha_0$. While simple, we show next that under some conditions using linear functions is not detrimental to the identifiability of the correct partitions as well as Π-invariant sets.

Theorem 4.5 (Linear approximation preserves partitioning and invariance)**.** *Given a target Y, let Π be its partition of the contexts C. Let $U \subset \mathbb{R}^{n-1}$ be a bounded set and $\mathcal{F} \subset L^2(U)$ be a subset of square-integrable functions containing the set $\mathcal{L}$ of linear functions $x \mapsto \alpha^t x + \alpha_0$. Further assume that each function f^c is sampled from $\mathcal{F}$ according to an absolutely continuous probability measure Q on $\mathcal{F}$. Let $\alpha_f = \arg\min_{\alpha, \alpha_0} E\left(\left(f(X) - \alpha^t X - \alpha_0 \right)^2 \right)$ be the projection of f onto its best linear approximation and write $\alpha_c = \alpha_{f^c}$. Then Q-almost surely*

$$f^{c_i} = f^{c_j} \text{ if and only if } \alpha_{c_i} = \alpha_{c_j} \ .$$

In particular, the partition Π as well as the Π-invariant sets $\mathcal{I}_\Pi(Y)$ according to $\{\alpha_c\}$ are Q-almost surely the same as for the $\{f^c\}$.

Now, to encode M_Π, we must encode both the partition Π as well as the aggregate model M_Π given Π. To encode the partition Π, we first encode the number of contexts using the MDL-optimal code for integers $L_{\mathbb{N}}$ [31]. The assignment of each context to one of at most $|C|$ groups can be encoded under a uniform distribution using $\log |C|$ bits. Overall

$$L(\Pi) = L_{\mathbb{N}}(|C|) + |C| * \log |C| \ .$$

It remains to transmit the model M_Π given Π. For each group $\pi \in \Pi$ we encode the group mean parameter $\alpha_\pi = \mathrm{mean}(\alpha_c : \Pi(c) = \pi)$ using the asymptotically optimal precision of $|C|^{1/2}$ per dimension [6], so that we obtain

$$L(M_\Pi \mid \Pi) = \sum_{\pi \in \Pi} L(M_\pi) = \frac{(|S| + 1)|\Pi|}{2} \log |C| \ . \tag{2}$$

Last, we need to encode the individual parameters α_c for each mechanism in M_Π. We do so by modeling each α_c as Normally distributed with, $\alpha_c \sim N(\alpha_{\Pi(c)}, \sigma^2)$ with unit variance $\sigma^2 = 1$,

$$\begin{aligned} L(M_c \mid M_\Pi) &= L(M_c \mid M_{\Pi(c)}) \\ &= \frac{|S| + 1}{2} \log(2\pi) + \frac{1}{2} \left\| \alpha_c - \alpha_{\Pi(c)} \right\|_2^2 \ . \end{aligned} \tag{3}$$

With the above, we have a principled way of encoding the coefficients. While guaranteed to work well with large sample sizes, there is a risk of underfitting when $|C|$ takes small values, as we expect in our use case. Hence, an adequate adjustment of the model cost in Eq. (2) to the error in Eq. (3) may be needed. Rather than pre-specifying a model cost that only depends on $|C|$ as above, we can choose the appropriate value *adaptively* to the data at hand.

To give intuition, we are interested in that number of groups $k = |\Pi|$ which contains the most information about the coefficients yet does not overfit. To find such a k we can use the so-called elbow property of the log mean squared error over $k = 1, \ldots |\Pi|$ [42]. We find that k for which the error decreases most steeply, and so that for $k + 1$ onward, the error flattens out, for which we use the following score

$$L'(M_\Pi, M_C) = \frac{1}{k - 1} \log \sum_{c=1}^{|C|} (\alpha_c - \alpha_\pi)^2 \ . \tag{4}$$

Moving forward, we use the MDL score L unless stated otherwise.

With a score that for each variable or set X can find partition Π in linear models, we turn to causal discovery next.

4.3 MDL-based Causal Discovery

We now explain how we use Π−invariance for causal discovery. As in Section 4.2, we will assume linear Gaussian functional models. This means that causal variables have invariant regression coefficients in each group of Π (Proposition 4.3). More importantly, under appropriate conditions the converse holds, i.e. the invariance property is obtained *only* for causal variables. To see this, we can extend the identifiability result of Peters et al. [2, 28].

Proposition 4.6 (Π−invariance for causal discovery)**.** *Let Y be generated according to the causal model of Assumption* (4.1)*. Further assume that the mechanism f^{Π} is linear, $Y = \alpha^T_{\Pi(C)}X + N$ and $N \sim \mathcal{N}(0, \sigma^2), N \perp\!\!\!\perp (C, X)$. Also assume that Π can be written as $\Pi = \{\pi, C \setminus \pi\}$ where $|\pi| \geq 2$, and for each variable X_i, there is at least one pair $c_1, c_2 \in \pi$ s.t. $\Pi_i(c_1) \neq \Pi_i(c_2)$. Then, a subset of causal variables $\mathcal{S}_\Pi \subseteq pa_{\mathcal{G}}(Y)$ are identifiable by invariance of the regression coefficients α_k over the contexts $c_k \in \pi$.*

To illustrate, in Fig. 1 the causal variable X_1 admits the *same* linear function g_1 in two contexts, irrespective of the distribution shift of X_1 in one of these contexts. For X_2 in contrast, only a spurious dependency to Y exists, so that the linear function *changes* when X_2 undergoes distribution shift in one of these contexts.

Note that this asymmetry holds due to the heterogeneity of the covariates. We also remark that we cannot use invariance when Y has a different generating process in every context.

To find sets of invariant variables in practice, we can proceed similarly to invariant prediction without mechanism changes [2, 28] and consider different subset regressions. That is, we find $\mathcal{S}_{\Pi^*}$ as

$$\begin{aligned}\mathcal{S}_{\Pi^*} &= \{X_i \mid X_i \text{ admits } \Pi^*\} \\ &= \bigcap \{X_S \mid X_S \in \mathcal{I}_{\Pi^*}(Y)\} \\ &= \bigcap \{X_S \mid \Pi^* = \arg\min_{\Pi} L(M_Y(X_S, \Pi))\} \ . \end{aligned} \tag{5}$$

We have $X_S \in \mathcal{I}_{\Pi^*}$ if the function $f : X_S \to Y$ is Π^*−invariant, in which case the partition Π^* leads to the best scoring model. We keep all variables X_i in the intersection of such sets X_S. Most importantly, by stating Eq. (5) in terms of our MDL-score, we can, with Π^* unknown, choose the partition with the lowest score.

We address the methodological details in the following.

5 THE VARIO ALGORITHM

We now introduce the Vario algorithm for discovering stable and changing mechanisms from data based on the theory outlined above. We present the pseudo-code of the principal method as Algorithm 1. Unless otherwise stated, we use the MDL-score L .

Given data X, Y over contexts C, Vario returns a partition Π^* and the set $\mathcal{S}^*_\Pi$ of variables admitting this partition. We also return other sets of admissible variables $\mathcal{S}_\Pi$ to be used in postprocessing.

In the first step (lines 2-4), we find potential partitions for each subset X_S. We linearly regress Y onto X_S in each context and cluster the resulting functions by similarity. We do so using subprocedure Vario-Π, which given coefficients $\{\alpha_1, \ldots, \alpha_{|C|}\}$ for all contexts, returns candidate partitions ordered by their scores. We postpone a detailed description of Vario-Π to Appendix B.

In a second step (lines 5-10) we address the intersection in Eq. (5).

Algorithm 1: Vario(X, Y)

input : target variable Y, covariates X in contexts C
output: partition Π^* and set $\mathcal{S}_{\Pi^*}$, as in Eq. (5)

1 admissible $\leftarrow \{\}$; $\Pi^* \leftarrow \{\}$; $\mathcal{S}_{\Pi^*} \leftarrow \{\}$;
2 **foreach** set X_S **do**
3 regression $Y \sim \alpha^T X_S$ in each $C \in \mathcal{C}$;
4 admissible(X_S) = Vario-Π $(\{\alpha_1, \ldots \alpha_{|C|}\})$;
5 **foreach** variable X_i **do**
6 $\mathcal{X} = \{X_S \mid X_i \in X_S\}$;
7 admissible(X_i) = $\bigcap_{X_S \in \mathcal{X}}$ admissible(X_S) ;
8 **foreach** partition Π in admissible(X_i) **do**
9 add X_i to $\mathcal{S}_\Pi$;
10 $\bar{L}(X_i, \Pi) = \frac{1}{|\mathcal{X}|} \sum_{X_S \in \mathcal{X}} L(M_Y(\Pi, X_S))$;
11 $\Pi^*, \mathcal{S}_{\Pi^*} = \arg\min_{\Pi, \mathcal{S}_\Pi} \sum_{X_i \in S_\Pi} \bar{L}(X_i, \Pi)$;
12 **return** $\Pi^*, \mathcal{S}_{\Pi^*}$, each non-empty set $\mathcal{S}_\Pi$;

We consider each X_i and add it to the set S_Π if it is a member of all subsets that admit Π (line 7). To find the best S_Π without going through all partitions, we associate a cost to X_i and Π (line 10). The cost corresponds to $L(M_Y(\Pi, X_S)$ in Eq. (5) averaged over all supersets X_S. That is, we account for how well Π describes the linear regressions with X_i. Finally, Vario returns Π^* and $\mathcal{S}_{\Pi^*}$ with the lowest cost. We explain their interpretation in the following.

Vario for Intervention Discovery. First, we can use Vario to locate interventions. By Assumption 4.1 and its instantiation through MDL, we take Π^* to be the most likely partition for Y, and can thus detect interventions on Y. If a context c_0 with observational data exists, we estimate the interventional contexts as

$$I(Y) = \{c_i \mid \Pi^*(c_i) \neq \Pi^*(c_0)\} \ .$$

If c_0 is unknown, we assume that the largest π in Π is the observational group and find $I(Y)$ accordingly.

Vario for Local Causal Discovery. We primarily propose Vario to find causal variables for a target Y based on Π-invariance. These correspond to the set $\mathcal{S}_{\Pi^*}$ that we find along with Π^*. We have

$$\Pi^* = \arg\min_{\Pi, \mathcal{S}_\Pi} L(M_Y(\mathcal{S}_\Pi, \Pi))$$

where the minimization ranges over all sets $\mathcal{S}_\Pi$ of Π-invariant variables, which is consistent with our modeling goal in Sec. 4.2. Under the conditions of Sec. 4.3, we have $S_{\Pi^*} \subseteq \text{pa}_{\mathcal{G}}(Y)$. Vario can therefore discover the causal parents of the target.

Vario for Global Causal Discovery. Although not our primary goal, we now describe a natural extension of our approach for global causal discovery, which we call Vario-$\mathcal{G}$. It not only discovers the causally relevant variables for Y, but a complete causal network.

The core idea of Vario-$\mathcal{G}$ is that instead of naively aggregating all edges found with Vario, we can refine the result by utilizing partitions and scores along *paths* in the network. To this end, in the first step (lines 1-5) of Algorithm 2, we include *each* set $\mathcal{S}_\Pi$ found with Vario, not just the best one. Specifically, for each X in one of the sets $\mathcal{S}_\Pi$, we add an edge to our graph, writing $X \to_{\Pi_{xy}} Y$ in $\mathcal{G}$ to say that Π_{xy} is the partition for Y which X admits.

Algorithm 2: VARIO-$\mathcal{G}(X)$

input : variables X in contexts C
output: $\mathcal{G}$ with edges $X \to_{\Pi} Y$

1 **foreach** target variable Y **do**
2 $\mathcal{S}_\Pi$ = VARIO $(X \setminus \{Y\}, Z)$;
3 **foreach** pair (X, Π) s.t. $X \in \mathcal{S}_\Pi$ **do**
4 add edge $X \to_{\Pi_{xy}} Y$ to $\mathcal{G}$;
5 label with $S_{xy} = L(M_Y(\Pi, X))$;
6 **foreach** path $X \to_{\Pi_{xy}} Y \to_{\Pi_{yz}} Z$ exists in $\mathcal{G}$ **do**
7 **if** $X \to_{\Pi_{xz}} Z$ exists in $\mathcal{G}$ **then**
8 **if** consistent$(\Pi_{xz}, \Pi_{xy} \cup \Pi_{yz})$ **then**
9 **if** $S_{xz} > S_{xy}$ **and** $S_{xz} > S_{yz}$ **then** prune $X \to Z$;
10 **return** $\mathcal{G}$;

We take the union of admissible edges for all targets (lines 1-5). The edges found in this way can include indirect, *ancestral* causal relations $X \to Z$ if there is a path $X \to_{\Pi_{xy}} Y \to_{\Pi_{yz}} Z$ in $\mathcal{G}$. However, in such a case, the mechanism changes for Z on the direct path must match those along the path via the intermediate node: for example, if Y was intervened upon in context c_1 and Z was intervened upon in c_2, the partition Π_{xz} for the direct path shows interventions in *both* c_1 and c_2, since it shows the coefficient changes of the composed linear function $g_{X\to Z} = g_{Y\to Z} \circ g_{X\to Y}$. Moreover, if $g_{X\to Z}$ emerges from composition rather than a direct causal edge, its MDL model score does not improve upon that of either $g_{Y\to Z}$ or $g_{X\to Y}$ (see Appendix A). We can make this property about ancestral paths precise by defining *consistency* of partitions.

Definition 5.1 (Consistency)**.** *A partition Π_1 is consistent w.r.t. Π_2 if for each pair $c, c' \in C$, if $\Pi_2(c) = \Pi_2(c')$ then $\Pi_1(c) = \Pi_1(c')$.*

Intuitively, Π_1 is consistent w.r.t. Π_2 if Π_2 has same or more fine-grained groups than Π_1. We also define the *union* of partitions as $\Pi_{12} := \Pi_1 \cup \Pi_2$ containing the combined interventions of both partitions, $\Pi_{12}(c) = \Pi_{12}(c')$ iff $\Pi_1(c) = \Pi_1(c')$ and $\Pi_2(c) = \Pi_2(c')$.

We then obtain for each chain of three nodes an ancestral path with the following properties.

Proposition 5.2. *For an ACM (Assumption 4.1), with DAG $\mathcal{G}$ and linear models, i.e. $X_i = \alpha^T_{\Pi_i(C)} pa_{\mathcal{G}}(X_i) + N$ and $N \sim \mathcal{N}(0, \sigma^2)$, let*

$$X \to_{\Pi_{xy}} Y \to_{\Pi_{yz}} Z$$

be a path in $\mathcal{G}$. Let $S_{xy} = L(M_Y(\Pi_{xy}, X)), S_{yz} = L(M_Z(\Pi_{yz}, Y))$ be the MDL-scores. Then there is an edge

$$X \to_{\Pi_{xz}} Z\,,$$

that is, X admits Π_{xz} for Z with score $S_{xz} = L(M_Z(\Pi_{xz}, X))$, and

- *Π_{xz} is consistent w.r.t. $\Pi_{xy} \cup \Pi_{yz}$,*
- *$S_{xz} \geq S_{xy}$ and $S_{xz} \geq S_{yz}$.*

We use these properties to prune ancestral variables in the second step (lines 5-8) of Algorithm 2. While we can make an analogous statement for longer paths, we limit consideration to paths over one intermediate node. For simplicity, we prune edges in one pass and examine edges by increasing score gain, which we define next.

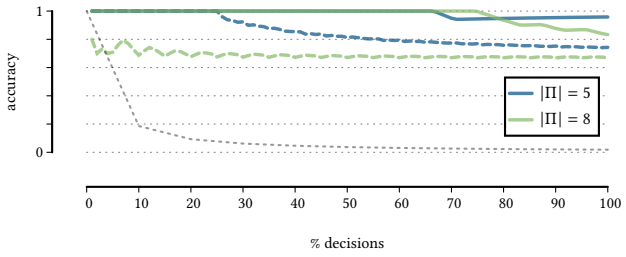


Figure 2: *VARIO recovers partitions exactly.* Given is the accuracy of VARIO-Π at finding Π for Y given $X = \mathbf{pa}_{\mathcal{G}}(Y)$. We show results per the x% decisions ordered by confidence, where solid lines show the empirical score L', dashed ones the the MDL score L, for $|C| = 5, 8$, $|\Pi| < 5$. Baseline accuracy of randomly choosing a partition is given by the gray line.

Confidence. For partition discovery with VARIO-Π, we can readily give a confidence statement. We want to measure the score gain that results from using the discovered groups of contexts rather than no groups, so we compare the encoded lengths of Π and the singleton partitioning $\Pi_0 = \{\{c_1\}, ..\{c_k\}\}$ as follows

$$L_{\text{conf}}(\Pi) = L(M_Y(\Pi, X)) - L(M_Y(\Pi_0, X))\ .$$

For causal discovery with VARIO-$\mathcal{G}$, we define the confidence in an edge $X \to_{\Pi} Y$ as the gain in MDL-score of using X in addition to the remaining causal parents, $X_S = S_{\Pi^*} \setminus \{X\}$. In detail, we estimate this set as $X_S = \arg\min_{X'_S} L(M_Y(\Pi, X'_S))$, and define

$$L_{\text{gain}}(X \to_{\Pi} Y) = L(M_Y(\Pi, X_S)) - L(M_Y(\Pi, X \cup X_S))\ .$$

Complexity. The time and space complexity of VARIO-Π are in $O(b)$ depending on the search strategy. For exhaustive search, b is a bell number of C. For a heuristic that we use in practice, $b = |C|^3$, and for a greedy version we have $b = |C|^2$ (Appendix B).

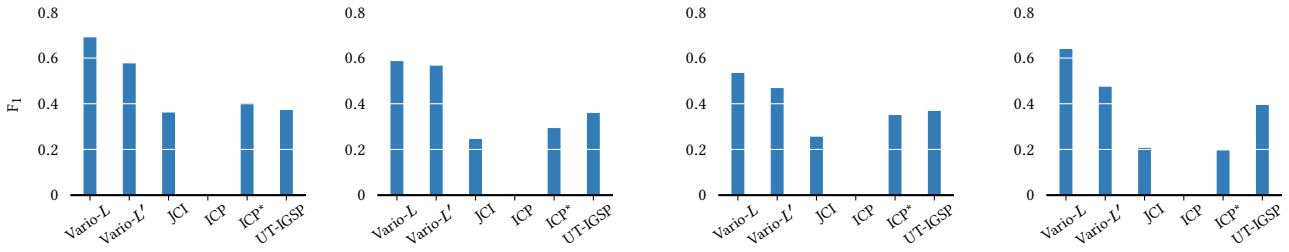
For VARIO, we have worst-case time complexity $O(2^{|X|}b)$ and space complexity $O(|X|^2 b)$. For this to be the case, we designed VARIO to traverse only X and not all sets S_Π (lines 5-10 of Algorithm 1). Note that this requires estimating scores on a per-variable basis (line 10) rather than directly computing them.

For VARIO-$\mathcal{G}$, we have complexity $O(2^{|X|}b|X|)$ for the first phase, $O(|X|^3)$ for considering ancestral paths. We deem the search over partitions feasible since we do not expect a large number of different domains in practice and as we found our approach to work well when few contexts are given (see Section 6). The main bottleneck is searching over all subsets in VARIO, similar to existing invariance-based approaches [2]. To circumvent this, one can preprocess X using a standard feature selection for Y [23] as our interest is in distinguishing *causal* features from correlated ones.

6 EVALUATION

In this section, we evaluate VARIO on synthetic and real data. Since we propose different use cases of our method, we structure our analysis by the following questions.

(1) Can VARIO-Π find good partitions?
(2) Can VARIO use partitions to discover causal variables?
(3) Can VARIO-$\mathcal{G}$ use partitions on paths to find causal graphs?



(a) **Mean shift, one at a time.** $\Pi = \{\{c_1\}, C \setminus c_1\}$.
(b) **Mean shift, context groups.** $\Pi = \{\pi_1, \pi_2, \pi_3\}$.
(c) **Hard interventions.** $\Pi = \{\pi_1, \pi_2\}$.
(d) **Mechanism changes.** $\Pi = \{\pi_1, \pi_2, \pi_3\}$.

Figure 3: *Vario discovers causal variables in different settings.* Shown is the F_1-score on discovering causal variables $X_i \in \mathbf{pa}_{\mathcal{G}}(Y)$ for each target Y in a random causal graph $\mathcal{G}$, with $|C| = 5$,$|\mathcal{G}| = 5$. Example partitions for Y are shown in each setting.

6.1 Experimental Setup

We implemented Vario in R. We compare to ICP [2] as state of the art in invariance-based causal discovery; the JCI [27] framework for constraint-based causal discovery from multiple contexts under mechanism changes; and UT-IGSP [40] for score-based causal discovery from multiple contexts under uncertain interventions. We use the original implementations recommended by the authors, e.g., instantiating JCI with the FCI algorithm [27]. All experiments finished within one day on a standard commodity laptop.

For the experiments with synthetic data, we first generate random acyclic networks $\mathcal{G}$ over X with density p of connections. Based on $\mathcal{G}$ we generate data for multiple contexts, with $|X| = |C| = 5, p = 1$ unless otherwise specified. We consider a linear Gaussian model that is augmented by context variables C, i.e.,

$$X_i = \sum_j \alpha_{ij}^c X_j + \beta_i^c C + N, \quad N \sim \mathcal{N}(0,1), c \in C\,, \tag{6}$$

where coefficients α_i^c are the causal strengths for X_i in context c, and β_i^c serve to simulate shift interventions. We choose the causal coefficients uniformly from $[-1, -0.25] \cup [0.25, 1]$, where the entries are $\alpha_{ij} \neq 0$ when $X_j \in \mathrm{pa}_{\mathcal{G}}(X_i)$ is a causal parent.

In our setting of multiple contexts, α, β may change across contexts depending on a partition chosen randomly for each variable. For instance, we can use the same α and have in each context c where $\beta_i^c = 1$ a shift intervention on X_i. Alternatively, we can modify α to model a soft intervention or mechanism change for Y.

6.2 Vario-Π for Partition Discovery

First, we check whether Vario can discover good partitions. For this purpose we consider mechanism changes, meaning that between groups π, the causal coefficients α take different values.

We run Vario-Π on input $(Y, \mathrm{pa}_{\mathcal{G}}(Y))$ for each target in random graphs $\mathcal{G}$ and match the partition it discovers to the ground truth.

In Fig. 2 we report, over the top-k% most confident decisions, the accuracy of exactly discovering the correct partition. We use L_{conf}, respectively L', to order decisions by. In the figure, we give the results for $|C| = 5, 8$. As a baseline, we give the favorably ordered accuracy of a randomly guessed partition, for illustration with $|C| = 5$. We find that Vario recovers the ground truth Π^* with much larger accuracy than is possible by chance. The empirical score L' (solid lines) slightly outperforms the MDL score L (dashed lines). This is what we expect because the MDL score sets a constant coefficient penalty for a given value of $|C|$, as shown in Eq. (2). In contrast, the empirical score uses the elbow criterion, which depends on the given coefficient values α_c as shown in Eq. (4), making the latter option more adaptive.

6.3 Vario for Local Causal Discovery

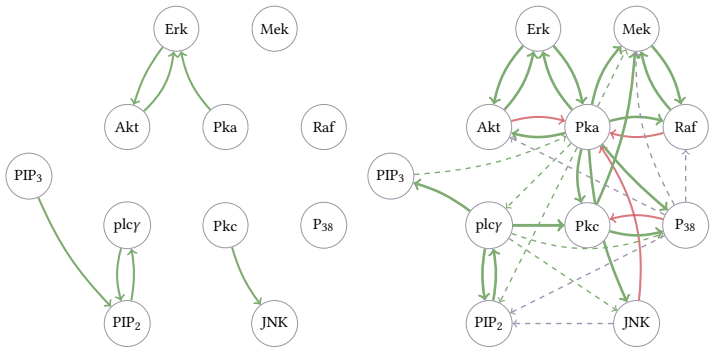
Since Vario-Π reliably finds good partitions, we now examine whether our weaker form of invariance still allows causal discovery.

We hereby evaluate Vario's main algorithm using L and L' respectively, and compare the set $\mathcal{S}_\Pi$ of invariant variables that Vario discovers with the ground truth $\mathrm{pa}_{\mathcal{G}}(Y)$, using F_1 scores.

Since ICP works well for two contexts, we merge all observational contexts into one and all interventional contexts $c \in I(Y)$ into one. In addition we are interested in whether our *score-based* approach offers any advantages over ICP's *hypothesis testing*. Thus, we also include a version named ICP* that applies ICP to two contexts at a time. It considers an edge causal if found in any pair of contexts, which we consider reasonable given that ICP's results are sparse, and that this design was used to apply ICP in practice [2, 25]. We here omit causal discovery approaches disregarding contexts, as preliminary results showed them to perform worse.

Mean Shift, one at a time. We first consider a simple setup where each variable may be subject to a mean shift intervention in at most one context, and each variable is affected in a different context. For example, this design might correspond to a typical diagonal experiment with interventions Mooij et al. [27]. Fig. 3a shows the methods' performance on identifying the causal parents $\mathrm{pa}_{\mathcal{G}}(Y)$. While JCI, ICP*, and UT-IGSP have high precision in finding causal edges, they are conservative and find fewer edges than Vario, with ICP finding no edges as expected under interventions. We obtain the best overall results with Vario, with the MDL score benefitting slightly over the empirical one (Fig. 3a).

Mean Shift, context groups. Next, we consider more general mean shifts. Each variable has a partition with groups, and the affected contexts may overlap for different variables. We observe in Fig. 3b that competitors' performance degrades, while Vario's performance is not notably impaired. In practice, we cannot rule out that more than one variable changes per context, especially as interventions may have off-target effects that are not known a priori.



(a) Estimated by ICP. (b) Estimated by VARIO.

Figure 4: *VARIO discovers causal interactions in the Sachs et al. [35] dataset.* Green edges are correct, dashed green are ancestral (but not anticausal), red edges are anticausal, and dashed gray are spurious edges with regard to the consensus network. ICP obtains an F_1 score of 0.38, VARIO 0.56.

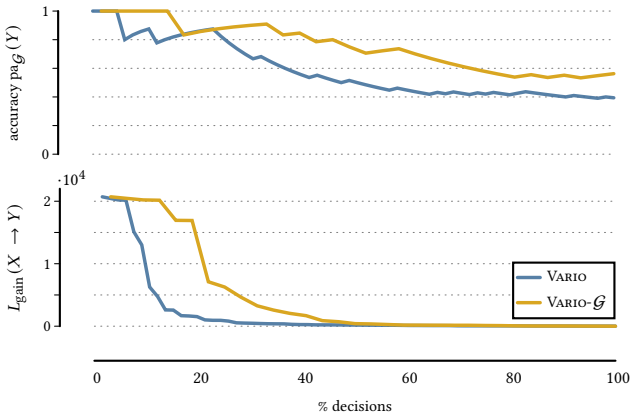


Figure 5: *VARIO is accurate when it is confident.* Given are the accuracies (top) over the top-k% discovered edges ordered by gain (bottom) for the Sachs et al. [35] dataset.

Hard Intervention. We also simulate surgical interventions that remove the influence of the causal parents altogether. That is, α attains value zero in one group, $\alpha_{i(1)} = 0, \alpha_{i(2)} = 1$. Although VARIO assumes fixed causal parents to Y over contexts, it does not suffer from this in practice (Fig 3c).

Mechanism Changes. Last, we consider changes in causal strength. In this setting, $\beta = 0$, and context groups arise from differences in α. This setting corresponds to causal mechanism changes or soft intervention on Y, and we observe the most considerable difference between VARIO and its competitors (Fig 3d).

We conclude that VARIO can find not only partitions but also causal variables based on them, whereby it can handle different types of mechanism changes. For experiments evaluating intervention discovery in the same settings, see Appendix C. In the remainder of this chapter, we focus on the performance of VARIO in the non-linear case.

6.4 VARIO-$\mathcal{G}$ for Global Causal Discovery in Real Data

We round up by evaluating VARIO on the protein signaling dataset by Sachs et al. [35] with measurements of proteins in human immune cells. The aim is to gain insight into the cell signaling pathways underlying such tissues. The data include eight experimental settings with molecular interventions on eleven proteins.

As there is controversy as to which pathways constitute the ground truth, we here use the network that Meinshausen et al. [25] propose, included in Appendix C, which serves as a summary of causal pathways reported in the literature and should be sufficient to allow a broad comparison to related methods. As a complication, path signaling may include feedback loops. Causal inference methods, including ours, assume acyclicity, however. Perfect identifiability may thus not be achievable on this dataset [25, 27].

We show the networks found by ICP [2] and VARIO-$\mathcal{G}$ in Fig. 4. We find remarkably many edges that agree with the literature, including 18 causal and only four reverse arrows. We remark that edges are reported for ICP whenever it finds invariance between any pair of contexts, given that its invariance test is conservative [25]. As JCI finds similarly few edges [27] we postpone the result to Appendix C. UT-IGSP [40] here has a high type II error (e.g., the maximum retrieved number of true positives is 11, for which it finds 26 false positives) so that we refrain from showing the network. Concerning other existing methods that do not consider multiple contexts or invariance, we note that most rely on explicit background knowledge about interventions [27, 30] whereas VARIO is applicable without known intervention targets.

On intervention detection, we achieve an F_1-score of 0.36 when using both the direct and indirect interventions reported in Sachs et al. [35] as ground truth, as compared to 0.3 for JCI [27].

We last investigate how confidence in edges relates to accuracy. To do this, we evaluate edges $X \to Y$ that VARIO and VARIO-$\mathcal{G}$ find, and consider the gain L_{gain} of using X in the causal parent set for Y. In Figure 5 we show the accuracies over edges (top) over the top-k% decisions as ordered by gain (bottom). We see that high gain values strongly correlate with high accuracy (top). For example, we found that all five spurious edges and two of the anticausal edges in Fig. 4 have the smallest gains among all edges. We also confirm Fig. 5 that the associated gain values are informative (bottom).

We also point out the difference between all admissible edges (blue) and those in the final result (yellow). It is due to VARIO finding many variables that admit a partitioning, some of which are removed in the pruning step. Hence, many false-positive decisions of VARIO are *indirectly* causal relationships and, since VARIO-$\mathcal{G}$ removes edges based on partitions and MDL-scores, both of these offer *qualitative information* about causal effects in the network.

Overall, this experiment shows that VARIO serves to gain insight into real-world networks. In particular, we found that besides constraint- and score-based methods, an invariance-based approach is effective for discovering causes and effects.

7 DISCUSSION AND CONCLUSION

In this work, we address the problem of finding a set of invariant causal mechanisms for a variable Y that we observe in different contexts C. We base our approach on the algorithmic model of causation and the principle of independent change, where we search for the simplest, in terms of Kolmogorov complexity, factorization of the joint distribution into independent mechanisms. We propose a practical instantiation using Minimum Description Length that allows discovering such mechanisms in the linear case, and discuss how doing so connects to causal discovery.

Our algorithm Vario can serve different goals, including intervention discovery, causal discovery, and gaining insight into where the generating process for Y changes, showing which contexts can be pooled together to base predictions on. Our evaluations have confirmed that Vario can handle different data generating processes, outperforms state of the art in invariance-based causal discovery, and can compete with non-linear approaches.

For the future, we would like to extend Vario from finding groups for one target variable at a time to discovering causal networks in a score-based fashion. To do this efficiently, we need heuristics to bypass the exact search over partitions and variable subsets that we currently use. Another promising direction is to investigate the theory and practice of using Vario to determine causal orientations; empirically, we have found that it works remarkably well in this regard, rarely returning any causal children rather than parents. We further plan to extend Vario such that it can discover changes due to noise interventions or selection bias.

REFERENCES

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. 2019. Invariant Risk Minimization. *CoRR* abs/1907.02893 (2019).

[2] P. Bühlmann. 2018. Invariance, Causality and Robustness. *CoRR* abs/1812.08233 (2018).

[3] D. M. Chickering. 2002. Optimal structure identification with greedy search. *JMLR* 3 (2002), 507–554.

[4] R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters. 2021. A causal framework for distribution generalization. PP (2021).

[5] D. Eaton and K. Murphy. 2007. Exact Bayesian structure learning from uncertain interventions. In *AISTATS*, Vol. 2. PMLR, San Juan, Puerto Rico, 107–114.

[6] P. Grünwald. 2007. *The Minimum Description Length Principle.* MIT Press.

[7] A. Hauser and P. Bühlmann. 2012. Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *JMLR* 13, 1 (aug 2012), 2409–2464.

[8] A. Hauser and P. Bühlmann. 2013. Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Statist. Soc. B* 77 (03 2013). DOI: http://dx.doi.org/10.1111/rssb.12071

[9] Y. He and Z. Geng. 2016. Causal Network Learning from Multiple Interventions of Unknown Manipulated Targets. *arXiv* (Oct. 2016), arXiv:1610.08611.

[10] C. Heinze-Deml and N. Meinshausen. 2021. Conditional Variance Penalties and Domain Shift Robustness. *Mach. Learn.* (2021).

[11] C. Heinze-Deml, J. Peters, and N. Meinshausen. 2018. Invariant causal prediction for nonlinear models. *J. Causal Inf.* 6, 2 (2018).

[12] S. Hu, Z. Chen, V. P. Nia, L. Chan, and Y. Geng. 2018. Causal Inference and Mechanism Clustering of a Mixture of Additive Noise Models. In *NeurIPS (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 5212–5222.

[13] B. Huang, K. Zhang, M. Gong, and C. Glymour. 2020. Causal Discovery from Multiple Data Sets with Non-Identical Variable Sets. In *AAAI*. 10153–10161.

[14] B. Huang, K. Zhang, and B. Schölkopf. 2015. Identification of Time-Dependent Causal Model: A Gaussian Process Treatment. In *IJCAI*. AAAI, 3561–3568.

[15] B. Huang, K. Zhang, P. Xie, M. Gong, E. P. Xing, and C. Glymour. 2019. Specific and Shared Causal Relation Modeling and Mechanism-Based Clustering. In *NeurIPS*, Vol. 32. Curran.

[16] B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. 2017. Behind Distribution Shift: Mining Driving Forces of Changes and Causal Arrows. In *ICDM*. 913–918. DOI: http://dx.doi.org/10.1109/ICDM.2017.114

[17] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. 2020. Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning. In *NeurIPS*. Curran, 9551–9561.

[18] D. Janzing and B. Schölkopf. 2010. Causal inference using the Algorithmic Markov Condition. *IEEE TPAMI* 56 (2010), 5168–5194.

[19] D. Kaltenpoth and J. Vreeken. 2019. We Are Not Your Real Parents: Telling Causal from Confounded by MDL. In *SDM*. SIAM.

[20] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*. PMLR, 5815–5826.

[21] J. Lemeire and D. Janzing. 2013. Replacing Causal Faithfulness with Algorithmic Independence of Conditionals. *Minds and Machines* 23 (2013).

[22] M. Li and P. Vitányi. 1993. *An Introduction to Kolmogorov Complexity and its Applications.* Springer.

[23] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. 2018. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. In *NeurIPS*.

[24] A. Marx and J. Vreeken. 2019. Identifiability of Cause and Effect using Regularized Regression. In *KDD*. ACM.

[25] N. Meinshausen, A. Hauser, J. M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. 2016. Methods for causal inference from gene perturbation experiments and validation. *PNAS* 113, 27 (2016), 7361–7368.

[26] O. Mian, A. Marx, and J. Vreeken. 2021. Discovering Fully Oriented Causal Networks. In *AAAI*.

[27] J. Mooij, S. Magliacane, and T. Claassen. 2020. Joint Causal Inference from Multiple Contexts. *JMLR* 21 (2020), 99:1–99:108.

[28] J. Peters, P. Bühlmann, and N. Meinshausen. 2016. Causal inference using invariant prediction: identification and confidence intervals. *J. R. Statist. Soc. B* 78, 5 (2016), 947–1012. DOI: http://dx.doi.org/10.1111/rssb.12167 (with discussion).

[29] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. 2014. Causal Discovery with Continuous Additive Noise Models. *JMLR* 15, 58 (2014), 2009–2053.

[30] J. Ramsey and B. Andrews. 2018. FASK with Interventional Knowledge Recovers Edges from the Sachs Model. *arXiv* abs/1805.03108 (2018).

[31] J. Rissanen. 1983. A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals Stat.* 11, 2 (1983), 416–431.

[32] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. 2018. Invariant Models for Causal Transfer Learning. *JMLR* 19, 36 (2018), 1–34.

[33] E. Rosenfeld, P. K. Ravikumar, and A. Risteski. 2021. The Risks of Invariant Risk Minimization. In *ICLR*.

[34] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. 2021. Anchor regression: Heterogeneous data meet causality. *J. R. Statist. Soc. B* 83, 2 (2021), 215–246.

[35] K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. 2005. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science* (2005), 523–9.

[36] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. 2012. On Causal and Anticausal Learning. In *ICML (ICML'12)*. Omnipress, Madison, WI, USA, 459–466.

[37] M. Sinha, P. Tadepalli, and S. A. Ramsey. 2021. Voting-based integration algorithm improves causal network learning from interventional and observational data: An application to cell signaling network inference. *PLoS One* 16, 2 (2021), e0245776.

[38] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. 2000. *Causation, prediction, and search.* MIT Press.

[39] P. Spirtes, C. Meek, and T. Richardson. 1999. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery* 21 (1999), 1–252.

[40] C. Squires, Y. Wang, and C. Uhler. 2020. Permutation-Based Causal Structure Learning with Unknown Intervention Targets. In *UAI*. PMLR, 1039–1048.

[41] J. Tian and J. Pearl. 2013. Causal Discovery from Changes. (2013).

[42] R. Tibshirani, G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. B* 63, 2 (2001), 411–423.

[43] R. Tillman and P. Spirtes. 2011. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *AISTATS*. PMLR, 3–15.

[44] S. Triantafillou and I. Tsamardinos. 2015. Constraint-based Causal Discovery from Multiple Interventions over Overlapping Variable Sets. *JMLR* 16, 66 (2015), 2147–2205.

[45] Y. Wang, C. Squires, A. Belyaeva, and C. Uhler. 2018. Direct Estimation of Differences in Causal Graphs. In *NeurIPS*, Vol. 31. Curran.

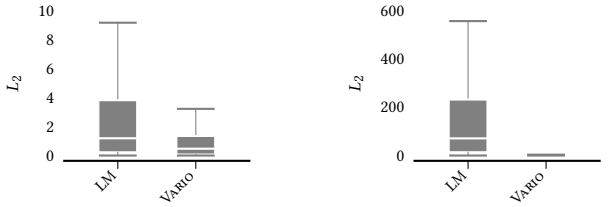
[46] K. Zhang, M. Gong, and B. Schölkopf. 2015. Multi-Source Domain Adaptation: A Causal View. In *AAAI*. AAAI Press, 3150–3157.

[47] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. 2017. Causal Discovery from Nonstationary/Heterogeneous Data: Skeleton Estimation and Orientation Determination. *IJCAI* 2017 (2017), 1347–1353.

Algorithm 3: VARIO-Π(X_S, Y, α)

input : X_S, Y, coefficients $\alpha = (\alpha_1, ..\alpha_{|C|})$ of $g : X_S \rightarrow Y$
output: partitions Π ranked by score L

1 Ranking = {};
2 **foreach** number of groups $k = 1, ...|C|$ **do**
3 **if** greedy **then** candidates = *topdown_partitions* (α, Π^*_{k-1});
4 **if** heuristic **then** candidates = *ordered_partitions* (α, k);
5 **foreach** Π_k in candidates **do**
6 Ranking(Π) = $L(M_Y(X_S, \Pi))$;
7 $\Pi^*_k = \arg\min_{\Pi_k}$ Ranking(Π_k) ;
8 sort(Ranking);
9 **return** Ranking



(a) No distribution shift of X in c'. **(b) Distribution shift of X in c'.**

Figure 6: *VARIO serves to find out which contexts can be pooled together for learning.* We compare L_2-loss of a linear model, trained on all contexts (LM) or observational ones (VARIO) for an unseen context c', where $|C| = 5$ and $|\Pi| = 2, 3, 4$.

A APPENDIX

We here state the proofs accompanying the propositions in the main paper.

PROOF OF PROPOSITION 4.3. Π−invariance follows by definition from the algorithmic Causal Model in Assumption 4.1 for each variable X_i and partition Π_i. □

For Proposition 2, we first formally define algorithmic Causal Models on *string* representations. An ACM over the strings $x_1, ...x_n$ associates with each string x_i a program q_i computing x_i from its causes $\text{pa}_{\mathcal{G}}(x_i)$ and an independent noise term, based on the true causal structure between the strings given as a DAG $\mathcal{G}$.

Postulate A.1. *(Algorithmic Model of Causality). Let $x_1, \ldots, x_n$ be strings and $\mathcal{G}$ the DAG capturing their causal structure, where no latent confounders exist. Then each x_i is computable by a program q_i of length $O(1)$ from its parents $pa_{\mathcal{G}}(x_i)$ and n_i as input,*

$$x_i = q_i(pa_{\mathcal{G}}(x_i, n_i))$$

where all n_i are independent.

To establish the connection between causality and algorithmic descriptions, we need the algorithmic Markov Property on string representations, of which we consider the following recursive formulation.

Postulate A.2. *(Algorithmic Markov Condition). For any ACM, the complexity $K(x_1, ...x_n)$ factorizes according to the DAG $\mathcal{G}$,*

$$K(x_1, ...x_n) \stackrel{+}{=} \sum_{x_i} K(x_i \mid pa_{\mathcal{G}}(x_i))$$

where $\stackrel{+}{=}$ denotes equality up to an additive constant.

Conceptually, with string representations instead of random variables to describe properties of interest, we need not assume that we can collect i.i.d. observations on each variable. We can thus, for the same random variable Y, use different nodes y^i that describe individual realizations of Y [18], or groups thereof with nodes y^π. Consequently, we can apply the algorithmic Markov Condition to graphs $\mathcal{G}$ where the target Y has any partition Π.

PROOF OF PROPOSITION 4.4. The statement follows by applying the algorithmic Markov Condition A.2 to a class of algorithmic Causal Models $\mathcal{G}_\Pi$ that separate the string y into nodes $\{y^{\pi_1}, ...y^{\pi_m}\}$ for each $\pi_i \in \Pi$ and that otherwise use the DAG-structure of $\mathcal{G}$. □

We further point out that the results on invariant causal prediction in the linear case apply *within* each group of contexts $\pi \in \Pi$ wrt. Y. We refer to Peters et al. [28] for detailed proofs.

PROOF OF PROPOSITION 4.6. We can apply Proposition 1 by Peters et al. [28] to $X, Y \mid \Pi(c) = \pi$, as within groups contexts do not affect the generating process $f^\pi : \text{pa}_{\mathcal{G}}(Y) \rightarrow Y$, thus the assumption that Y is not directly affected by contexts holds in π. □

PROOF OF PROPOSITION 5.2. Let $X_1 \rightarrow X_2 \rightarrow X_3$ in $\mathcal{G}$, with Π_{12} for X_2 and Π_{23} for X_3. We want to show that X_1 admits a consistent partition Π_{13} for X_3, i.e. VARIO finds a direct path $X_1 \rightarrow X_3$ with Π_{13}, and that there is an upper bound on the model score of Π_{13}.

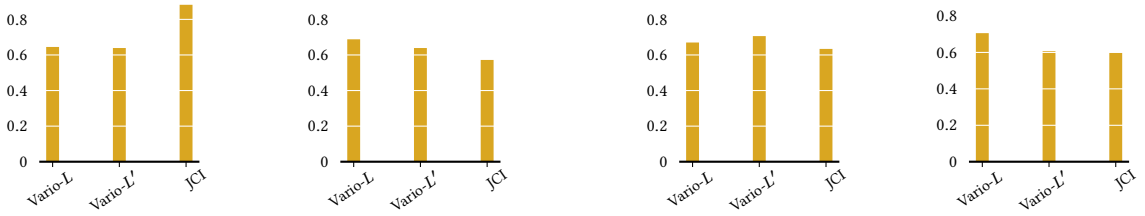
Let $g_{12} : X_1 \rightarrow X_2, g_{23} : X_2 \rightarrow X_3$, then there is a transitive function $g_{13} : X_1 \rightarrow X_3$ with coefficients $\alpha^c_{13} = \alpha^c_{12}\alpha^2_{23}$ for each c, thus $\alpha^c_{12} \neq \alpha^{c'}_{12} \vee \alpha^c_{23} \neq \alpha^{c'}_{23} \Rightarrow \alpha^c_{13} \neq \alpha^{c'}_{13}$, that is, $\Pi_{12}(c) \neq \Pi_{12}(c') \vee \Pi_{23}(c) \neq \Pi_{23}(c')$ and Π_{13} is consistent w.r.t. $\Pi_{12} \cap \Pi_{23}$.

Let $S_{12} = L(M_{X_2}(\Pi_{12}, X_1))$ and $S_{23} = L(M_{X_3}(\Pi_{23}, X_2))$ be the MDL-scores for encoding the linear functions g_{12}, g_{23}, respectively. For all partitions, we obtain zero cost for encoding the coefficient errors in the data limit, $L(M_C \mid M_\Pi) = 0$. Since $|\Pi_{13}| \geq |\Pi_{12}|, |\Pi_{13}| \geq |\Pi_{23}|$ by consistency, we have $L(M_{\Pi_{13}}) \geq L(M_{\Pi_{13}})$ and $L(M_{\Pi_{13}}) \geq L(M_{\Pi_{23}})$, so the same holds for the model scores. □

PROOF OF THEOREM 4.5. Since the functional $\alpha : f \mapsto \alpha_f$ is linear, we can write $\mathcal{F} = \ker(\alpha) \oplus \mathcal{F}'$. Thus, if two functions f, g map to the same $\alpha_f = \alpha_g$ we know that $f - g \in \ker(\alpha)$. Thus if $Q(\alpha_f = \alpha_g) > 0$ we have $Q(\ker(\alpha)) > 0$. Equivalently, due to the disintegration theorem the restriction of $Q_{|\mathcal{F}'}$ to $\mathcal{F}'$ has an atom at the origin, $Q_{|\mathcal{F}'}(\{0\}) > 0$. However, since Q is absolutely continuous, so is $Q_{|\mathcal{F}'}$, which is a contradiction. □

B PSEUDOCODE

VARIO for Partition Discovery. We here show the missing piece of our approach, VARIO-Π. As shown in Algorithm 3, we evaluate partitions Π_k for each number of groups k (lines 2-5). For each partition we consider the model $M_Y(X_S, \Pi))$ and score it (line 4) as



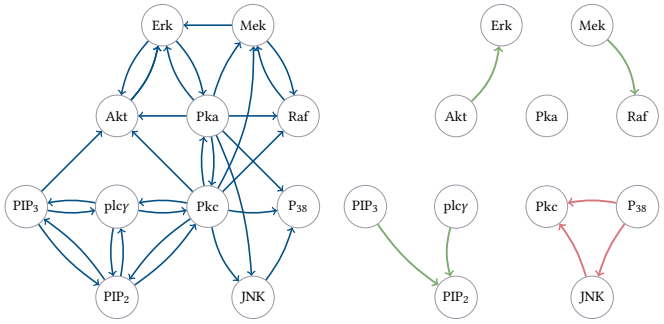
(a) Mean shift, one at a time. (b) Mean shift, context groups. (c) Hard interventions. (d) Mechanism changes.

Figure 7: *Vario detects interventions in different settings.* Shown is the F_1-score on discovering interventional contexts $c \in I(Y)$ for each target Y in a random causal graph $\mathcal{G}$, with $|C| = 5$,$|\mathcal{G}| = 5$. Settings are the same as in Fig. 3.

described in Section 4.2. We finally return a ranked list of partitions ordered by L (respectively L') to use in the main Algorithm 1.

To avoid enumerating all partitions, we propose a heuristic. We order each dimension of α_i and on each set $\{\alpha_i^c \mid c \in C\}$, only allow partitions without gaps. We denote these as *ordered_partitions* (k). To justify, given infinite data the partitions for all causal variables X_i coincide (Assumption 4.1). Thus, the orderings of each dimension α_i will agree, and the true partition will be among those we consider. At each k we hereby have $(k-1)$ options for d dimensions of X, leading to runtime in $O(d|C|^3)$. We can also opt for a greedy version that does not consider all ordered partitions at k, but only those that result from splitting one group of the best partition at $k-1$, called *topdown_partitions* (k). The greedy version is in $O(d|C|^2)$.

Vario for Out-of-Distribution Learning. We briefly describe an additional use case of our method: generalization to *unseen* contexts. It has two aspects. First, the discovered *partitions* can, by revealing in which contexts interventions or latent variables are present, inform us which contexts we can safely combine for learning. We do so by only using observational contexts, $C \setminus I(Y)$, for prediction. Secondly, the use of *causal* variables rather than *all* covariates may also lead to more reliable predictions, particularly in case that the unseen context is subject to distribution shift [4, 23, 46].



(a) Ground truth [25]. (b) Estimated by JCI.

Figure 8: *Causal interactions in the Sachs et al. [35] dataset.* Shown is the consensus network [25] and the causal graph found with JCI [27], with green edges if present in (a).

C ADDITIONAL RESULTS

Vario for Out-of-distribution Learning. As a concluding experiment, we consider predicting Y by the Π–invariant mechanism that governs the majority of contexts. In detail, we assume that the underlying generating process $f : X \to Y$ applies to at least two observed contexts and will continue to generate Y in future contexts; all the while, different mechanisms may apply in interventional contexts. In Figure 6 we confirm that pooling only those contexts where f is in place and learning a linear model leads to more reliable predictions in a future context c' (Fig. 6a). In particular, the effect improves with covariate shift: if X is distributed differently from C in c', the gap is more prominent (Fig. 6b). We refer to Magliacane et al. [23] to illustrate similar effects regarding the use of causal variables rather than all correlated variables.

Vario for Intervention Discovery. Complementary to our results on causal discovery in Fig. 3, in Fig 7 we present the results on intervention discovery in the same settings. Although we found the empirical score L' to perform better on partition discovery when the causal parents are *known* (Fig. 2), on a causal graph with unknown parents we do not observe a notable difference between using L' and L anymore (Fig.7), so that we suggest relying on the principled MDL-approach with Vario-L in practice. While JCI [27] here shows better performance in the first setting, as soon as there are multiple groups, results are on par with those of Vario.

Competitors for Global Causal Discovery in Real-World Data. In continuation of the results on the Sachs et al. [35] dataset, we show the consensus network [25] and the causal network that JCI finds in Figure 8. For JCI, we favorably combined the results of different variants that they consider [25, 27]. For results of UT-IGSP on this dataset, we refer to Squires et al. [40] as we obtain many false-positive edges with this approach.