Intervals of Causal Effects for Learning Causal Graphical Models

Samuel Montero-Hernandez SAMUEL@INAOEP.MX
Felipe Orihuela-Espina F.ORIHUELA-ESPINA@INAOEP.MX
Luis Enrique Sucar ESUCAR@INAOEP.MX
Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico.

Abstract

Structure learning algorithms aim to retrieve the true causal structure from a set of observations. Most times only an equivalence class can be recovered and a unique model cannot be singled out. We hypothesized that casual directions could be inferred from the assessment of the strength of potential causal effects and such assessment can be computed by intervals comparison strategies. We introduce SLICE (Structural Learning with Intervals of Causal Effects), a new algorithm to decide on unresolved relations, which taps on the computation of causal effects and an acceptability index; a strategy for intervals comparison. For validation purposes, synthetic datasets were generated varying the graph size and density with samples drawn from Gaussian and non-Gaussian distributions. Comparison against LiNGAM is made to establish the performance of SLICE over 1440 scenarios using the normalised structural Hamming distance (SHD). The retrieved structures with SLICE showed smaller SHD values in the Gaussian case, improving the structure of the retrieved causal model in terms of correctly found directions. The acceptability index is a good predictor of the true causal effects ($R^2 = 0.62$). The proposed strategy represents a new tool for discovering unravelled causal relations in the presence of observational data only.

Keywords: causal discovery, causal effects, structure learning.

1. INTRODUCTION

A set of causal relations can be expressed through a *causal graphical model* (CGM) (Pearl, 2009). The structure of a CGM can be specified by an expert or retrieved from data by a *structure learning algorithm* (SLA). Several SLAs split the learning problem into the search and the orientation phases. The search stage decides on the existence of a link, and the orientation stage decides on the direction of the information flow. Search strategies can produce an equivalence class – a collection of statistically equivalent solutions – (where the true causal model is expected to be a member) or a single structure of the causal model based on different assumptions. Under certain circumstances, SLA can achieve success in retrieving relations but fail when disentangling the cause from the effect of such retrieved relation. When departing merely from observational data (samples of the system of interest without manipulating any variables), under Gaussianity, traditional SLA can only retrieve an equivalence class. They are unable to pick the true causal model among the statistically equivalent members of such equivalence class. A specific assumption of non-Gaussian disturbances could yield a single structure (Shimizu et al., 2006). However, deviations from a multivariate Gaussian distribution are hard to estimate in high dimensions (Maathuis et al., 2009). Then assuming Gaussian interventional (experimental) information could be helpful. The incorporation of interventional information becomes a necessary aid to isolate the solely causal model.

Causal inferences can be formulated without manipulation (Pearl, 2009, Ch. 11.4.5), manipulation permits detecting useful properties of a system, and hence, contributing to the identification of cause-effect pairs (Eberhardt et al., 2005). Given that an essential feature of causation is responsiveness, that is a change in the *effect* given a change in the *cause*, it makes sense to devise strategies to orient undefined causal relations based on observed effects.

In this paper, we address the problem of orienting unresolved directions of causal graphical models. We describe our proposal as founded on the relationship between the size and directionality of the effects. We hypothesise that an orientation strategy based on comparing the size and direction of potential causal effects will unravel undefined causal relations.

1.1 Related Work

Typical classification of SLA groups them into score-based (Lam and Bacchus, 1994; Campos, 2006) and constraint-based (Cooper and Herskovits, 1992; Claassen and Heskes, 2012) approaches. Score-based methods retrieve a structure which obtains a maximum score fitting the data. Constraint-based strategies start with a complete connected/disconnected network and remove/add links following conditional independence tests. Further, strategies for learning CGM can be broadly organised: (i) depending on the input data which can be observational, interventional or hybrid, (ii) regarding the learning strategy, these can either maximise the number of edges or attempt to obtain full identifiability, and (iii) depending on the criterion to decide the directionality of the edges they could be based on interventions or causal effects.

Works starting from observational data and orienting edges based on interventions include (Meganck et al., 2006; Borchani et al., 2007; Masegosa and Moral, 2013; Shimizu et al., 2006). In these works, different utility functions based on entropy, connectivity and costs are explored. Both single and multiple interventions are evaluated, and the causal structures are determined based on the obtained score. In (He and Geng, 2008) randomised experiments and quasi-experiments are performed for orientation. Eberhardt (2008) proved the conjecture of the sufficient and necessary number of experiments to uniquely identify the causal structure from a Markov equivalence class. Hyttinen et al. (2013) adapted existing concepts in combinatorics for experimental selection to address the causal discovery problem. From an interventional dataset, Eaton and Murphy (2007) retrieved the causal structure while the manipulated variables are discovered. Hauser and Peter (2012) explored strategies for orienting the maximum number of edges per intervention and the discovery of a full identifiable model. Using a hybrid solution, Hauser and Bühlmann (2015) obtained an interventional equivalence class smaller than the one obtained by only observations by jointly combining observational and interventional data. Regarding full identifiability, Shimizu et al. proposed the Linear Non-Gaussian Acyclic Models (LiNGAM) algorithm for learning a unique graph structure from observational data (Shimizu et al., 2006). LiNGAM relies on the assumptions of linear generating functions, non-Gaussian noise in structural equation models and no unobserved confounders. LiNGAM yields good estimations when the non-Gaussian disturbances assumption hold and large sample sizes (e.g., > 10000) are available.

Different from previous works, ours gives the possibility to go a step further when an SLA can only find a set of potential causal structures. SLICE offers the plausibility to take a partially directed causal graph and generates a complete casual structure with stability in high-dimensional scenarios and able to cope with Gaussian and non-Gaussian variables.

2. PRELIMINARY DEFINITIONS

We shall denote variables by capital letters (e.g., X, Y, Z) and the values assigned to them with lower-case (e.g., x, y, z). Sets will be denoted by bold letters e.g., $\mathbf{X} = \{X_1, \ldots, X_n\}$, $\mathbf{x} = \{x_1, \ldots, x_n\}$. Formally, a causal system can be represented by means of a structural causal model (Definition 1).

Definition 1 ***((Pearl, 2009, Sect. 1.5) Structural Causal Model)*** *A structural causal model (SCM) is a triplet $\{\mathbf{U}, \mathbf{V}, \mathcal{F}\}$ (Pearl et al., 2016) where $\mathbf{U}$ is a set of exogenous variables, $\mathbf{V}$ is a set endogenous variables, and $\mathcal{F}$ is a set of functions, such that:*

i Every variable $V \in \mathbf{V}$ is caused (takes its value) by $f : (\mathbf{U} \cup V') \to \mathbf{V}$, such that $V' \subset \mathbf{V}$ and $f \in \mathcal{F}$.

ii Every $V \in \mathbf{V}$ is takes its value from at least one exogenous variable.

iii Variables in $\mathbf{U}$ are caused by some mechanism that we decide not include in the system.

An SCM can be represented by means of a directed acyclic graph (DAG). A DAG G is an ordered pair $(\mathbf{X}, \mathbf{E})$ comprising a set of vertices $\mathbf{X} = \{\mathbf{U} \cup \mathbf{V}\}$ and a set of ordered pairs $\mathbf{E} = (X_i, X_j)$. In a causal context, edges of the form $X_i \to X_j$ means that X_i is a cause of X_j and the set of functions $\mathcal{F}$ has the form $X_i = f_i(pa_i)$ where pa_i is the set of its parents (direct causes).

Definition 2 (Equivalence class (of graphs)) *A class of equivalence (of graphs) is a subset $\mathcal{G} = \{G : G \sim \mathcal{G}\ ,\ G, \mathcal{G} \in G^*\}$, where $G \sim \mathcal{G}$ indicates that there is an equivalence relation between G and $\mathcal{G}$, with G^* being the set of all possible equivalence classes.*

Definition 3 (Intervention (on a causal probabilistic model)) *The atomic intervention denoted by $do(X_i = x_i)$ or $do(\hat{X}_i)$, amounts to removing the equation $X_i = f_i(pa_i, U_i)$ from the model and substituting $X_i = x_i$ in the remaining equations. The new model represents the behaviour of the system under the intervention $do(X_i = x_i)$ and, when solved for the probability distribution (P) of X_j, yields the causal effect of X_i on X_j denoted as $P(X_j|\hat{x}_i)$.*

Interventions in a graphical model are performed by applying the *do()* operator (Pearl, 2009). An *intervention* on a variable in the graphical model generates a manipulated version of P. The probability of a variable Y given the intervention of a variable X, is expressed as $P(Y|do(X = x))$, or $P(Y|\hat{x})$ for short. The post-intervention joint distribution P' over $X_1, X_2, \ldots, X_n$ when $do(\hat{x_i})$ is given by Equation 1:

$$P(v_1, v_2, \ldots, v_n|\hat{v_i}) = P'(v_i) \prod_{j|V_j \notin V_i} P(v_j|pa_j), \tag{1}$$

Definition 4 (Causal Effect) *Given two disjoint sets of variables, X and Y, the causal effect of X on Y, denoted as $\theta_{y|\hat{x}}$, is a function from X to the space of probability distributions on Y. For each realisation x of X, $\theta_{y|\hat{x}}$ gives the probability of $Y = y$ induced by deleting from $X_i = f_i(pa_i)$ all equations corresponding to variables in X and substituting $X = x$ in the remaining equations.*

3. CAUSAL EFFECTS FOR EDGE ORIENTATION

In CGMs, interventions quantify the extent of influence over a variable given the manipulation of another. To compute the causal effect of one variable over another, a complete defined CGM is needed. The rules of the *do* calculus can be applied when a single structure (DAG) is present. There are algorithms for computing causal effects when a DAG is given (Lauritzen, 2001; Tikka and Karvanen, 2017). However, in many scenarios, the CGM has to be estimated from observational data only. In this case, it is usually infeasible to isolate the true causal model, but an equivalence class of the true causal model can be probably obtained. Either a completed partially directed acyclic graph (CPDAG) or a partially ancestral graph (PAG) can be obtained depending on whether the assumption of causal sufficiency (allowing for hidden common causes) is made or not. Here, we rely on the causal sufficiency assumption; consequently, we deal with CPDAG structures. Given a data set, a CPDAG can be retrieved by different algorithms (Chickering, 2002; Spirtes et al., 2000).

When a causal structure is known, for example $X_0 \rightarrow X_1 \rightarrow X_2$, and supposing that $X_0 = \epsilon_0$, $X_1 = 2X_0 + \epsilon_1$ and $X_2 = 3X_1 + \epsilon_2$, then the operation $P(X_2|do(X_1))$ can be computed by removing all the terms of X_0 in the equation of X1 and graphically by deleting all incoming arcs to X_1. Further, under linearity, the manipulation of X_1 and observation of its effect in X_2, that is $P(X_2|do(X_1 = x_1)) - P(X_2|do(X_1 = x_1 + 1))$, can be established by the regression coefficient of X_1 in $X_2 \sim \beta \cdot X_1 + \epsilon_2$. This coefficient is $\beta \neq 0$ when the observed variable (X_2) is not in the parents set of the intervened variable (X_1) and $\beta = 0$ zero when the observed variable is in the parents set of the intervened one. For the reverse effect, when X_2 is intervened and X_1 is observed, the causal effect is zero. It follows that the causal effect computed in the true causal direction (X_1 manipulated and X_2 observed) should result in a greater causal effect than the one computed in the reverse of the true causal direction (X_2 manipulated and X_1 observed) regardless of the value of β, since $|\beta| > 0$ is expected. This example can be extended to the case when the true causal structure is not known, but we have a set of potential causal structures e.g., for equivalence classes of DAGs. In such cases, it is possible to compute causal effects when the true causal structure is not known but, instead of obtaining a scalar, a multi-set of potential causal effects is computed (Maathuis et al., 2009). Multi-sets can be characterised in terms of their lower and upper bounds. We propose to consider these bounds as *intervals of causal effects (ICE)* (see formal definition below) for their later comparison. Let $X1 \leftrightarrow X2$ be a CPDAG representing an equivalence class of an unknown causal structure and let ICE_1 and ICE_2 be two intervals of causal effects computed by manipulating X_1 and observing X_2 and vice versa obtained from the potential structures $X_1 \rightarrow X_2$ and $X_1 \leftarrow X_2$ respectively. The true causal direction between X_1 and X_2 can be found by identifying whether $ICE_1 > ICE_2$ or $ICE_1 < ICE_2$ (as the base case where $|\beta| > 0$). In this sense, a suitable procedure for intervals comparison is requiered for the computation of the estimated causal effects.

Intervention calculus when the DAG is absent (IDA) is an algorithm to compute the bounds of causal effects when the true DAG is not known (Maathuis et al., 2009). The details of IDA can be found in (Maathuis et al., 2009, 2010), but briefly, IDA aims to compute the causal effect on a variable Y given the manipulation of a variable X departing from a set of observational data. An equivalence class is estimated from the observations using the PC algorithm (Spirtes et al., 2000), and then, by applying the *do* calculus the multiset of causal effects observed in Y by intervening X, denoted by $\mathbf{\Theta}_{Y|\hat{X}}$, is computed for the m members in the equivalence class, that is $\mathbf{\Theta}_{Y|\hat{X}} = \{\theta^j_{Y|\hat{X}}\}$ for $j = 1 \ldots m$. If all $\theta^j_{Y|\hat{X}}$ are the same, it is understood that the causal effect has been uniquely determined. Otherwise, different potential causal effects would be obtained. In such case, the lower and upper bounds of $\mathbf{\Theta}_{Y|\hat{X}}$ can be obtained as the minimum and maximum absolute values as $[\min|\theta^j_{Y|\hat{X}}|, \max|\theta^j_{Y|\hat{X}}|]$.

If given a set of observations by an unknown CGM entailed in the equivalence class $\mathcal{G}$, it is possible to estimate a multiset of potential causal effects $\mathbf{\Theta}$, then, we hypothesise it is possible to go backwards, whereby $\mathbf{\Theta}$ encodes clues about the undefined links in $\mathcal{G}$. Consequently, the equivalence class can be shrunk (i.e. diminish the number of undefined causal links) perhaps to a single causal structure. We argue that it is possible to estimate the causal direction in an undefined edge $X - Y$ based on the assessment of the potential causal effects of X on Y in contrast to the potential causal effects of Y on X, that is $\mathbf{\Theta}_{Y|\hat{X}}$ *vs.* $\mathbf{\Theta}_{X|\hat{Y}}$. The problem here is elaborating a suitable strategy for the assessment of both multisets. We propose the use of the lower and upper bounds of causal effects as *intervals of causal effects* established in Definition 5.

Definition 5 (Interval of causal effects) *The interval of causal effects (ICE) of X on Y is defined as $ICE_{Y|\hat{X}} = [min|\theta^j_{Y|\hat{X}}|, max|\theta^j_{Y|\hat{X}}|]$, where $|\cdot|$ is the absolute value and $\theta^j_{Y|\hat{X}}$ is the potential causal effect on Y given the manipulation of X for every member of an equivalence class $\mathcal{G}$.*

3.1 Acceptability index for intervals comparison

Multisets in $\mathbb{R}$ are proxied here by closed intervals in $\mathbb{R}$. (Sengupta and Pal, 2000) proposed a strategy for comparing closed intervals in $\mathbb{R}$. Let I be the set of all closed intervals in $\mathbb{R}$, and let $A, B \in I$ be the interval proxies of the multisets with $[a_L, a_R] \subseteq \mathbb{R}$ and $[b_L, b_R] \subseteq \mathbb{R}$ as lower and upper bounds of A and B respectively. A and B can be represented by their midpoint $m(\cdot)$ and half-width $w(\cdot)$. For example, $A = \langle m(A), w(A) \rangle$, with $m(A) = (a_L + a_R)/2$ and $w(A) = (a_R - a_L)/2$.

Regardless, it should not be assumed that the original multisets A or B are continuous, nor that the mid-point belongs to the original multisets. The acceptability function $\mathcal{A}_< : I \times I \rightarrow [0, \infty)$ is defined as per Eq. 2:

$$\mathcal{A}_<(A, B) = \frac{m(B) - m(A)}{w(B) + w(A)}\,, \tag{2}$$

with A and B pre-ordered by their midpoints i.e. $m(A) \leq m(B)$ and $w(A) + w(B) \neq 0$. $\mathcal{A}_<(A, B)$ can be interpreted as a rate of satisfaction of premise $A < B$, where $<$ implies an order relation between A and B. Since the image of $\mathcal{A}_<$ is $[0, \inf)$, the acceptability of $A < B$ can be classified according to expression 3:

$$\mathcal{A}_<(A < B) = \begin{cases} = 0 & \text{if } m(A) = m(B), \\ > 0 \text{ and } < 1 & \text{if } m(A) < m(B) \text{ and } a_R > b_L \\ > 1 & \text{if } m(A) < m(B) \text{ and } a_R \leq b_L \end{cases} \tag{3}$$

If $\mathcal{A}_<(A, B)$ = 0 we reject the idea that A is inferior to B. If $0 < \mathcal{A}_<(A, B) < 1$, then A is inferior to B to some extent with uncertainty. And for $\mathcal{A}_<(A, B) \geq 1$ we have "absolute" certainty that B is superior to A. Other interval comparison strategies exist in the literature, but we choose (Sengupta and Pal, 2000) because the acceptability index in Equation 2, can be easily calculated in terms of the upper and lower bounds of a multiset of causal effects. Moreover, a threshold could be set to control an allowed uncertainty level for the orientation of undefined edges.

4. STRUCTURE LEARNING WITH INTERVALS OF CAUSAL EFFECTS

Our algorithm attempts for resolving the causal direction from a set undefined causal edges. It iteratively selects the pairs of variables with the smallest interval-overlapping regarding their intervals of causal effect and the chosen pair indicates the cause and the effect in that relation.

Given a departing equivalence class $\mathcal{G}$ represented by a CPDAG, we extract the set $\mathbf{T}$ of undefined edges from $\mathcal{G}$; $(X, Y) \in \mathbf{T} | X - Y \in \mathcal{G}$. For every pair $(X, Y) \in \mathbf{T}$, the causal effects of intervening X over Y, $\mathbf{\Theta}_{Y|\hat{X}}$ and vice versa $\mathbf{\Theta}_{X|\hat{Y}}$ are computed. The lower and upper bounds of multi-sets $\mathbf{\Theta}_{Y|\hat{X}}$ and $\mathbf{\Theta}_{X|\hat{Y}}$ are used to establish ICE$_{Y|\hat{X}}$ and ICE$_{X|\hat{Y}}$, The midpoints of both ICEs are calculated in order to conform the premise $A < B$ where A and B are the ICE with the minimum and maximum midpoint respectively. Consistent with the previous ordering step, one of the pairs, (X, Y) or (Y, X), is stored in the set $\mathbf{T}_o$ as the ordered pair meaning the potential causal direction. Then, the acceptability index ($\mathcal{A}_<$ in Eq. 2) is computed across all pairs of intervals and is stored in $\mathcal{A}_{set}$. The argument of the maxima $\mathcal{A}_<$ in $\mathcal{A}_{set}$ is chosen as the index pointing out the pair in $\mathbf{T}_o$, which is estimated as the causal orientation. The application of the set of Meek rules[1] (Meek, 1995) is attempted. Finally, the set $\mathbf{T}$ is updated and the whole process is iteratively carried out until no more undefined edges remain. Our algorithm has been developed in R language and can be downloaded from `https://github.com/smonterohdz/slice`.

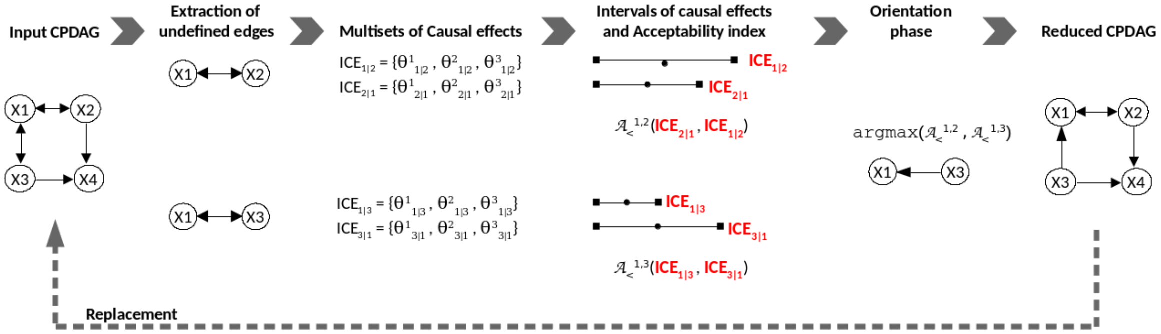

Input CPDAG
Extraction of undefined edges
Multisets of Causal effects
Intervals of causal effects and Acceptability index
Orientation phase
Reduced CPDAG
X1
X2
X3
X4
$ICE_{1|2} = \{\theta^1_{1|2}, \theta^2_{1|2}, \theta^3_{1|2}\}$
$ICE_{2|1} = \{\theta^1_{2|1}, \theta^2_{2|1}, \theta^3_{2|1}\}$
$ICE_{1|2}$
$ICE_{2|1}$
$\mathcal{A}_<^{1,2}(ICE_{2|1}, ICE_{1|2})$
$\text{argmax}(\mathcal{A}_<^{1,2}, \mathcal{A}_<^{1,3})$
$ICE_{1|3} = \{\theta^1_{1|3}, \theta^2_{1|3}, \theta^3_{1|3}\}$
$ICE_{3|1} = \{\theta^1_{3|1}, \theta^2_{3|1}, \theta^3_{3|1}\}$
$ICE_{1|3}$
$ICE_{3|1}$
$\mathcal{A}_<^{1,3}(ICE_{1|3}, ICE_{3|1})$
Replacement

Figure 1: The general idea for orientation via causal effects. Starting from a CPDAG from which a multiset of causal effects is computed, a set of intervals of causal effects (ICEs) is generated from the lower and upper bounds of the multisets. Then, the acceptability index $\mathcal{A}$ is computed from both ICEs. Orientation is performed regarding the maximum $\mathcal{A}$ and the CPDAG is then reduced.

The proposed strategy for causal structure learning is illustrated in Figure 1 and formalised as the structure learning with intervals of causal effects (SLICE) presented in Algorithm 1. SLICE aims at the identification of a single causal structure. The computational complexity of SLICE is $O(g^u)$ with $g = \#(\{G_i \in \mathcal{G}\})$ and $u = \#(\mathbf{T})$ where $\#(\cdot)$ is the cardinality. The current implementation of SLICE assumes linear functions since we employ IDA for causal effects estimation. Extension to other probability distributions can be achieved by applying the inference rules of *do-calculus* (Pearl, 2009, Sec. 3.4.2).

SLICE assumes it starts from a *correct* CPDAG, that is, a correct skeleton with the all the v-structures correctly identified. When SLICE finds an undefined edge $X - Y$, it attempts orienting $X \rightarrow Y$ or $X \leftarrow Y$ according to the procedure of assessing intervals. This assumes that $X - Y$ was a correctly identified dependency. However, in cases of false positives (FP), SLICE will fail as

[1] A set of four directionality rules to partially orient a causal graph given some edges patterns.

it will be attempting to orient an edge that should not exist. In other words, SLICE has not have integrated a mechanism to detect FP in the input CPDAG.

```
Algorithm: SLICE: Structure learning with intervals of causal effects.
Data: Observations from the set of variables $\mathbf{V}$, an initial CPDAG $\mathcal{G}$
Result: An estimated directed acyclic graph $G_e$
Extract the set of undefined edges $\mathbf{T}$ from $\mathcal{G}$;
repeat
    $\mathcal{A}_{set} := \emptyset$ ;
    $\mathbf{T}_o := \emptyset$;
    foreach $(X, Y) \in \mathbf{T}$ do
        Compute the multisets of causal effects $\Theta_{Y|\hat{X}}$ and $\Theta_{X|\hat{Y}}$ with IDA ;
        Determine ICE$_{X|\hat{Y}}$ and ICE$_{Y|\hat{X}}$ (Def. 5);
        if $m(ICE_{X|\hat{Y}}) \leq m(ICE_{Y|\hat{X}})$ then
            $A :=$ICE$_{X|\hat{Y}}$ ;
            $B :=$ICE$_{Y|\hat{X}}$ ;
            $\mathbf{T}_o := \mathbf{T}_o \cup \{(X, Y)\}$;
        else
            $A :=$ICE$_{Y|\hat{X}}$ ;
            $B :=$ICE$_{X|\hat{Y}}$ ;
            $\mathbf{T}_o := \mathbf{T}_o \cup \{(Y, X)\}$;
        end if
        $\mathcal{A}_{set} := \mathcal{A}_{set} \cup \{\mathcal{A}_{<}(A, B)\}$;
    end foreach
    $(V_1, V_2) := \mathbf{T}_o[argmax(\mathcal{A}_{set})]$;
    Orient as $V_1 \rightarrow V_2$ in $\mathcal{G}$ ;
    Apply Meek rules to $\mathcal{G}$;
    Update the set of undefined edges $\mathbf{T}$ from $\mathcal{G}$;
    $G_e := \mathcal{G}$;
until $\mathbf{T} = \emptyset$;
return $G_e$
```

5. EXPERIMENTS AND RESULTS

5.1 Generation of synthetic models

Algorithm 1 was tested on synthetic models by varying the number of variables, the density of connection between nodes and number of data samples. Each synthetic model consists of a DAG G_t, a set of linear functions $\mathcal{F}$, and a set of observations $\mathbf{D}$. The synthetic DAGs and their set of observations $\mathbf{D}$ were generated using the `randomDAG` and `rmvDAG` functions of the R package pcAlg (Kalisch et al., 2012) (v2.5-0, `https://CRAN.R-project.org/package=pcalg`). The procedure for generating the synthetic models and data sets is described in the pcAlg package documentation. Briefly, every X_i node ($i = 1, \ldots, n$) in G_t is forward connected ($X_i \rightarrow X_{\{i+1,\ldots,n\}}$) to k neighbours, with $k < (n - i)$ chosen with probability p_c. The edges in G_t are randomly weighted between 0.1 and 1.0. To generate $\mathbf{D}$, every X_i node in G_t is visited in topological order and its value is given by function $X_i = f_i(pa_i, e_i) \in \mathcal{F}$ where pa_i is the set of parents of the X_i node and e_i is an error term with a specific probability distribution. Every $f_i(\cdot)$ function is defined as $f_i = w_1 \cdot X_1 + \ldots + w_h \cdot X_h + e_i$ where $1, \ldots, h$ indicates the h parents of the X_i node. Observations in $\mathbf{D}$ are sampled by evaluating each $f_i(\cdot)$ as many times as needed.

5.2 Model evaluation

For evaluating models, we use the *normalised structural Hamming distance* (SHD). SHD has been used before in the assessment of algorithms for causal structure learning (Peters and Bühlmann, 2015; Hauser and Peter, 2012; Tsamardinos et al., 2006; Colombo and Maathuis, 2014; Kalisch and Bühlmann, 2014). The SHD is defined as the minimum number of edge insertions, deletions, and changes needed to transform one model into another (Tsamardinos et al., 2006, Algorithm 4). We computed the SHD between the sub-graphs of the true and the estimated causal model induced by the set of undefined edges in the initial equivalence class. Let graphs $G_t = (V, E_t)$, $\mathcal{G} = (V, E_c)$ and $G_e = (V, E_e)$ be the true model, the initial equivalence class and the estimated model, respectively, where V, E_t, E_c and E_e are the sets of nodes, edges in the true model, edges in the initial equivalence class and edges in the estimated model, respectively. We defined the subset of undirected edges as $\mathbf{T} \subseteq E_c | (V_i, V_j) \in \mathbf{T}$ iff $(V_i, V_j) \wedge (V_j, V_i) \in E_c$, we computed $SHD(G_t[\mathbf{T}], G_e[\mathbf{T}])$. An SHD=0 indicates a perfect match, and a value of SHD=1 means that no estimated edge matches any edge in the true causal model. Values of SHD > 1 indicate that, although there are no coincident edges, the estimated model resulted with more (spurious) edges. Figure 2 illustrate the induced sub-graph on the true model, the equivalence class and the estimated model.

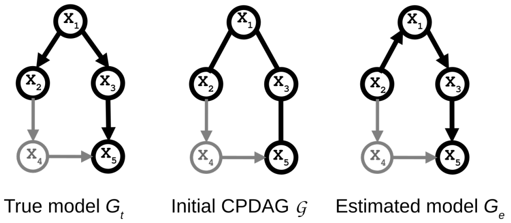


Figure 2: Example of model evaluation in induced sub-graphs. The subset S is determined by the undefined edges (bold edges and nodes) in the initial equivalence class (middle), then is extracted from the true (left) and estimated model (right) model. The SHD is computed between the induced sub-graphs $G_t[S]$ and $G_e[S]$, the sub-graphs in grey are not considered in SHD computation.

5.3 Validation

Concurrent validity is shown by comparing SLICE output against true synthetic data. We generated several synthetic causal models by varying the sample size ($\mathbf{D}$), number of variables or nodes (n), and the connection density (p_c). For every model two sets $\mathbf{D}$ sized $1,000$ and $10,000$ were generated. We evaluate the performance of the algorithm for $n = \{4, 8, 16, 32, 64, 100\}$ and a connection density $p_c = \{0.25, 0.85\}$. Besides, we explored two scenarios regarding the initial CPDAG; when the initial CPDAG is estimated and may contain errors in the skeleton, and when the initial CPDAG is correct. Thirty replications (model instances) were made for every combination of parameters (number of variables and density of connections) for a total of 1440 cases. Initial CPDAG were estimated with the PC algorithm (Spirtes et al., 2000) with alpha=0.01 by using the version implemented in the R package pcAlg (Kalisch et al., 2012) (v2.5-0). The SHD between the estimated and the true models is reported (Fig. 3a).

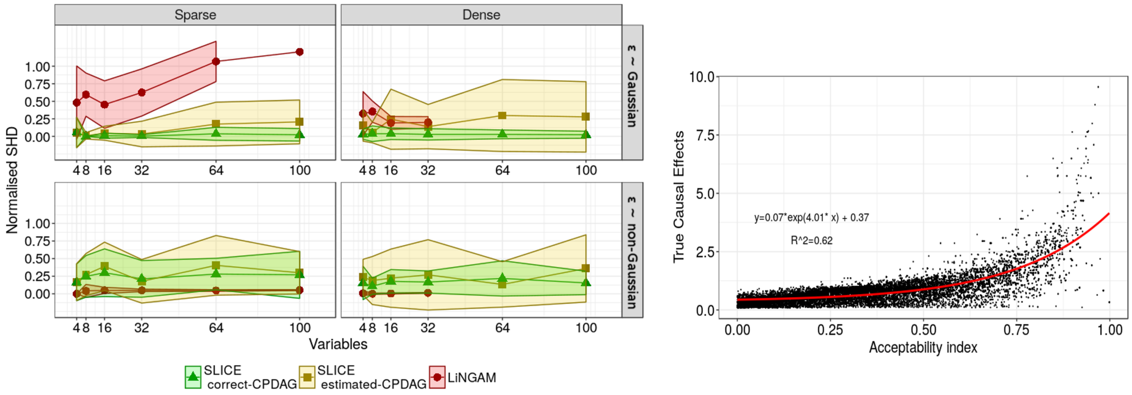
We explored the performance when errors are drawn from a Gaussian distribution $\mathcal{N}(0, 1)$ and from a non-Gaussian distribution. The non-Gaussian distribution was generated as described in Shimizu et al. (2006); briefly a Gaussian distribution was raised to an exponent in the range $[0.5, 0.8] \cup [1.2, 2.0]$. For reference, SLICE results are shown against LiNGAM (Shimizu et al., 2006) over the same set of synthetic models. LiNGAM was chosen for comparison because it assumes linear functions and no unobserved confounders (assumptions shared with SLICE). Results are summarised in Fig. 3a. In both scenarios, under Gaussian and non-Gaussian errors disturbances, when the initial CPDAG does not contain all the conditional (in)dependence relations, SLICE fails to correctly orient about 25% of the links, and the dispersion increases when the true causal model has a 85% of connectivity density.

Note that this is not an error of SLICE but of the algorithm used for equivalence class estimation, and SLICE simply propagates that error. If the structure search algorithm is affected by the number of variables and samples, then it is probable that it generates false positive or Type I errors (spurious undefined edges) or false negatives or Type II errors (missing links) in the estimated CPDAG. In cases of false positives, SLICE attempts to orient spurious links and regardless on the decided orientation, the directed edge will make the final structure more distant from the true causal model. Analogously, in the case of missing edges, SLICE will be limited since the set of undefined edges will be incomplete. For the non-Gaussian case, LiNGAM obtained approximately the complete causal structures, as expected. However, SLICE was able to identify the 75% in most cases. Expectedly, the best results are obtained when SLICE departs from a correct CPDAG from Gaussian distribution. SLICE results are shown to be more consistent when the number of variables increases where LiNGAM seems to make random guesses or fails to finish the procedure. SLICE seems to be less affected by connectivity density, although the dispersion increases as the number of variables does. Regarding the execution time for learning a single model, in average, SLICE scales better than LiNGAM on a 3.6 GHz Intel Core i7 processor with 15 GB RAM using R 3.4.1 in Debian 9.4 OS. SLICE takes $\mu = 0.26 \pm 1.02$ seconds when starts with a correct CPDAG and $\mu = 0.05 \pm 0.07$ seconds when the CPDAG is estimated and LiNGAM delays $\mu = 4.82 \pm 9.78$ seconds.

5.4 Relation between the true causal effects and the acceptability index

The idea of considering multisets of causal effects as intervals (ICE) combined with the assessment of them by computing the acceptability value ($\mathcal{A}$) form a heuristic able to orient undefined causal relations in most cases. To investigate whether the $\mathcal{A}$ index encodes any information regarding causal effects we perform a regression analysis between the acceptability values and the true causal effects. For every pair of variables forming an undefined edge in an initial equivalence class we computed the $\mathcal{A}$ index of ICEs in both directions, also we computed the true causal effects considering the same pair of variables but now in the true causal model. If $X - Y$ is in $\mathcal{G}$ and $X \rightarrow Y$ is in G_t we contrast $\mathcal{A}_{<}(\text{ICE}_{X|\hat{Y}}, \text{ICE}_{Y|\hat{X}})$ versus $\theta^{G_t}_{Y|\hat{X}}$.

Figure 3b shows the behaviour of causal effects associated with the $\mathcal{A}_{<}$ index. Regression analysis suggests that the acceptability index (without outliers $\mathcal{A}_{<} = 1.0$) among potential causal effects is a good predictor of the true causal direction ($R^2 = 0.62$). Our heuristic of orienting undefined edges on the base of greater acceptability value appears appropriate for choosing the



(a) (b)

Figure 3: a) Normalised SHD (mean – line – and standard deviation – shaded area –) for SLICE given a correct CPDAG, SLICE given an estimated CPDAG and LiNGAM algorithms. Results for the 0.2 and 0.85 connectivity density are presented in columns and error distribution (Gaussian and non-Gaussian) in rows. Smaller values (close to zero) indicate better match with the true causal model (best seen in color). b) Relation between acceptability index and true causal effects. Regression analysis (red line) indicates that acceptability values significantly predicted true causal effects.

most probable correct orientation. A threshold on the acceptability value may provide some error control at the expense of not isolating a single model.

6. CONCLUSIONS

We have addressed the learning of the structure of causal graphical models and contributed with a new algorithm capable of deciding a single DAG by using causal effects during the orientation stage. Our proposal's rational is founded on the idea of the size and directionality of the causal effects. The new orientation strategy unravels undefined causal relations by computing causal effects of a pair of variables conforming undefined edges. When the initial CPDAG is correct, our results show a significant improvement in terms of SHD compared to LiNGAM algorithm. The performance of SLICE subtly decreases in cases where the initial CPDAG is not correctly estimated, e.g., as it may be the case when using PC for the structure search stage, yet it still outperforms LiNGAM in many cases. Our proposal can be extended by incorporating more elaborated methods for causal effects estimation as presented in (Hyttinen et al., 2015) and exploring alternative methods for multisets comparison. The performance of SLICE when error disturbances are generated from non-Gaussian distributions (assumed by LiNGAM) suggested SLICE outperforms LiNGAM. Finally, the current approach relies on the causal sufficiency assumption and dropping such assumption is critical for applicability in many real-world domains.

ACKNOWLEDGEMENTS

This research has been funded by the Mexican research council through project CB-2014-01-237251 and SMH is supported by a PhD scholarship from the CONACYT.

References

H. Borchani, M. Chaouachi, and N. Ben Amor. Learning Causal Bayesian Networks from Incomplete Observational Data and Interventions. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 17–29, 2007.

L. M. D. Campos. A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests. *Journal of Machine Learning Research*, 7:2149–2187, 2006.

D. M. Chickering. Learning Equivalence Classes of Bayesian-Network Structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

T. Claassen and T. Heskes. A Bayesian Approach to Constraint Based Causal Inference. *UAI 2012, Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 207–216, 2012.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 10 1992.

D. Eaton and K. P. Murphy. Exact Bayesian structure learning from uncertain interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 107–114, 2007.

F. Eberhardt. Almost Optimal Intervention Sets for Causal Discovery. *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 161–168, 2008.

F. Eberhardt, C. Glymour, and R. Scheines. On the Number of Experiments Sufficient and in the Worst Case Necessary to Identify All Causal Relations Among N Variables. *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 178–184, 2005.

A. Hauser and P. Bühlmann. Jointly Interventional and Observational Data: Estimation of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 77(1):291–318, 2015.

A. Hauser and B. Peter. Two Optimal Strategies for Active Learning of Causal Models from Interventions. *Workshop on Probabilistic Graphical Models*, 55(2008):123–130, 6 2012.

Y.-B. He and Z. Geng. Active Learning of Causal Networks with Intervention Experiments and Optimal Designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Experiment Selection for Causal Discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013.

A. Hyttinen, F. Eberhardt, and J. Matti. Do-calculus when the True Graph Is Unknown. *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 395–404, 2015.

M. Kalisch and P. Bühlmann. Causal Structure Learning and Inference: A Selective Review. *Quality Technology & Quantitative Management*, 11(1):3–21, 2014.

M. Kalisch, M. Machler, D. Colombo, M. H. Maathuis, P. Buhlmann, M. Mächler, D. Colombo, and M. H. Maathuis. Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software*, 47(11):26, 2012.

W. Lam and F. Bacchus. Learning Bayesian Belief Networks: An approach Based on the MDL Principle. *Computational Intelligence*, 10(4):269–293, 1994.

S. Lauritzen. Causal inference from graphical models. *Complex Stochastic Systems*, pages 63–107, 2001.

M. H. Maathuis, M. Kalisch, and P. Buhlmann. Estimating High-Dimentional Intervention Effects From Observational Data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

M. H. Maathuis, D. Colombo, M. Kalisch, P. Buhlmann, H. M. Maathuis, D. Colombo, M. Kalisch, P. Buhlmann, M. H. Maathuis, D. Colombo, M. Kalisch, P. Buhlmann, and P. Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 4 2010.

A. R. Masegosa and S. Moral. An interactive approach for Bayesian network learning using domain/expert knowledge. *International Journal of Approximate Reasoning*, 54(8):1168–1181, 2013.

C. Meek. Causal Inference and Causal Explanation with Background Knowledge. In *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*, pages 403–418, 1995.

S. Meganck, P. Leray, and B. Manderick. Learning Causal Bayesian Networks from Observations and Experiments: A decision Theoretic Approach. In *LNCS*, volume 3885, pages 58–69, 2006.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge Univ Press, second edition, 2009.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Pirmer*. Wiley, 2016.

J. Peters and P. Bühlmann. Structural Intervention Distance for Evaluating Causal Graphs, 2015.

A. Sengupta and T. K. Pal. On Comparing Interval Numbers. *European Journal of Operational Research*, 127(1):28–43, 11 2000.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, Massachusetts, USA, 2nd edition, 2000.

S. Tikka and J. Karvanen. Identifying Causal Effects with the R Package causaleffect. *Journal of Statistical Software*, 76(12), 2017.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, 65(1):31–78, 10 2006.