Discovering Reliable Causal Rules

Kailash Budhathoki[*] Mario Boley[○] Jilles Vreeken[●]

Abstract

We study the problem of deriving policies, or *rules*, that when enacted on a complex system, *cause* a desired outcome. Absent the ability to perform controlled experiments, such rules have to be inferred from past observations of the system's behaviour. This is a challenging problem for two reasons: First, observational effects are often unrepresentative of the underlying causal effect because they are skewed by the presence of confounding factors. Second, naive empirical estimations of a rule's effect have a high variance, and, hence, their maximisation typically leads to spurious results.

To address these issues, we first identify conditions on the underlying causal system that—by correcting for the effect of potential confounders—allow estimating the causal effect from observational data. Importantly, we provide a criterion under which causal rule discovery is possible. Moreover, to discover reliable causal rules from a sample, we propose a conservative and consistent estimator of the causal effect, and derive an efficient and exact algorithm that maximises the estimator. Extensive experiments on a variety of real-world and synthetic datasets show that the proposed estimator converges faster to the ground truth than the naive estimator, recovers causal rules even at small sample sizes, and the proposed algorithm efficiently discovers meaningful rules.

1 Introduction

The ultimate goal of data analysis is to understand the underlying data-generating process in terms of cause and effect. Towards this goal, rule mining [1, 8, 24] has been studied extensively over the years. Most rule miners measure the effect of a rule in terms of correlation or dependence. Correlation, however, does not imply causation. Thus, maximizing those measures does typically not reflect the causal structure of the data-generating process.

The gold standard for establishing the causal relationship between variables is through a controlled experiment, such as a randomized controlled trial (RCT) [10]. In many cases, however, it is infeasible to perform an RCT. We hence most often have to infer causal dependencies from observational data, which is data that was collected without full control. In this work, we study discovering causal rules from observational data that maximise causal effect. Though simple to state, this is a very hard task. Not only do we have to cope with an intricate combination of two semantic problems—one statistical and one structural—but in addition the task is also computationally difficult.

The structural problem is exemplified in Simpson's paradox. Even strong and confidently measured effects of a rule might not actually reflect true domain mechanisms, but can be mere artefacts of the effect of other variables. Notably, such confounding effects can not only attenuate or amplify the marginal effect of a rule on the target variable, in the most misleading cases they can even result in sign reversal, i.e. when interpreted naively, the data might indicate a negative effect even though in reality there is a positive effect [16, Chap. 6]. For example, a drug might appear to be effective for the treatment of a disease for the overall population. However, if the treatment assignment was affected by sex that also affects the recovery (say males, who recover—regardless of the drug—more often than females, are also more likely to use the drug than females), we *falsely* find that the treatment is *not* effective at all to male and female subpopulations.

The statistical problem is the well-known phenomenon of overfitting. In practice, we often estimate the causal effect of a rule from a sample. The overfitting phenomenon then results from the high variance of the naive empirical (or "plug-in") estimator of causal effect for rules with too small sample sizes for the instances either covered, or excluded by the rule. Combined with the maximization task over a usually very large rule language, this variance turns into a strong positive bias that dominates the search and causes essentially random results of either extremely specific or extremely general rules.

Finally, the rule space over which we maximise causal effect is exponential in size and does not exhibit structure (e.g. monotonicity) that is trivially exploited. We therefore need an efficient optimization algorithm. In this paper, we present a theoretically sound approach to discovering causal rules that remedies each of these problems.

1. To address the structural problem, instead of measuring the marginal observational effect of a rule, we propose to measure its conditional observational effect given a set of potential confounder variables. Moreover, we give a criterion under which causal rule discovery is possible.

2. To address overfitting, we measure and optimise the *reliable* effect of a rule. By down-weighing rules with a small effect evidence (subsample covered or excluded by the rule), we propose a conservative empirical estimate of the population effect, that is not prone to

[*] kaibud@amazon.de, Amazon Research Tübingen (work done prior to joining Amazon when the author was working at MPI-INF)

[○] mario.boley@monash.edu, Monash University

[●] jv@cispa.de, CISPA Helmholtz Center for Information Security

overfitting. Additionally, and in contrast to other known rule optimisation criteria, it is also *consistent*, i.e., with increasing amounts of evidence, the measure converges to the actual population effect of a rule.

3. We develop a practical algorithm for efficiently discovering the top-k strongest reliable causal rules. In particular, we show how the optimisation function can be cast into a branch-and-bound approach based on a computationally efficient and tight optimistic estimator.

We support our claims by experiments on both synthetic and real-world datasets as well as by reporting the required computation times on a large set of benchmark datasets.

2 Related Work

2.1 Association rules In rule-based classification, the goal is to find a set of rules that optimally predict the target label. Classic approaches include CN2 [13], and FOIL [17]. In more recent work, the attention shifted from accuracy to optimising more reliable scores, such as area under the curve (AUC) [7].

In association rule mining [1], we can impose hard constraints on the relative occurrence frequency to get reliable rules. In emerging and contrast pattern mining [3, 6], we can get reliable patterns whose *supports* differ significantly between datasets by performing a statistical hypothesis test.

Most subgroup discovery [11, 24] methods optimise a surrogate function based on some null hypothesis test. The resulting objective functions are usually a multiplicative combination of coverage and effect.

All these methods optimise associational effect measures that are based on the observed joint distribution. Thus they capture correlation or dependence between variables. They do not reflect the effect if we were to intervene in the system.

2.2 Causal rules Although much of literature is devoted in mining reliable association rules, a few proposals have been made towards mining causal rules. Silverstein et al. [20] test for pairwise dependence and conditional independence relationships to discover causal associations rules that consist of a univariate antecedent given a univariate control variable. Li et al. [14] discover causal rules from observational data given a target by first mining association rules with the target as a consequent, and performing cohort studies per rule.

Atzmueller & Puppe [2] propose a semi-automatic approach to discovering causal interactions by mining subgroups using a chosen quality function, inferring a causal network over these, and visually presenting this to the user. Causal falling rule lists [22] are sequences of “if-then” rules over the covariates such that the effect of a specific intervention decreases monotonically down the list from *experimental* data. Shamsinejadbabaki et al. [19] discover actions from a partial directed acyclic graph for which the post-intervention probability of Y differs from the observational probability. Papaxanthos et al. [15] discover statistically significant feature combinations by conditioning on a categorical covariate.

While all these methods have opened the research direction, we still lack a theoretical understanding. These methods have introduced the idea of conditioning observational effect measures on certain covariates. We continue this line of work by providing a framework that describe which covariates principally need to be considered for conditioning and describing how the resulting effect measure can be estimated reliably and efficiently from data.

3 Reliable Causal Rules

We consider a system of discrete random variables with a designated **target variable** Y and a number of covariates, which we differentiate into **actionable variables**[1] $\mathbb{X} := (X_1, \ldots, X_\ell)$ and **control variables** $\mathbb{Z} := (Z_1, \ldots, Z_m)$. For example, Y might indicate recovery from a disease, $\mathbb{X}$ different medications that can be administered to a patient, and $\mathbb{Z}$ might be attributes of patients, such as blood group. Let $\mathscr{X}_j$ denote the domain of X_j, and $\mathscr{Z}_j$ be that of Z_j. As such, the domain of $\mathbb{X}$ is the Cartesian product $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_\ell$, and that of $\mathbb{Z}$ is $\mathscr{Z} = \mathscr{Z}_1 \times \cdots \times \mathscr{Z}_m$.

We use Pearl's do-notation [16, Chap. 3] $do(X := x)$, or $do(x)$ in short, to represent the **atomic intervention** on variable X which changes the system by assigning X to a value x, keeping everything else in the system fixed. The distribution of Y after the intervention $do(x)$ is represented by the **post-intervention** distribution $P(Y \mid do(X := x))$. This may not be the same as the **observed** conditional distribution $P(Y \mid X = x)$. As we observe $P(Y \mid X = x)$ without controlling the system, other variables might have influenced Y, unlike in case of $P(Y \mid do(X := x))$. Therefore, to capture the underlying data-generating mechanism, we have to use the post-intervention distribution $P(Y \mid do(X := x))$.

Let $\mathscr{S}$ be the set of all possible vector values of all possible subsets of actionable variables. More formally,

$$\mathscr{S} = \bigcup_{\mathbf{x} \in \mathscr{P}(\{\mathscr{X}_1, \ldots, \mathscr{X}_\ell\})} \mathbf{x} \, ,$$

where $\mathscr{P}(\bullet)$ is the powerset function. In this work, we are concerned with **rules** $\sigma : \mathscr{S} \to \{\top, \bot\}$ that for a given value $\mathbf{x} \in \mathscr{S}$ evaluate to either true ($\top$) or false ($\bot$). Specifically, we investigate the **rule language** $\mathscr{L}$ of conjunctions of **propositions** $\sigma \equiv \pi_1 \wedge \cdots \wedge \pi_l$ that can be formed from inequality and equality conditions on actionable variables X_js (e.g. $\pi \equiv \text{dosage} \geq 450$).

Let $\mathbf{X} \subseteq \mathbb{X}$ denote the subset of actionable variables, with their joint domain $\mathscr{X}$, on which propositions of a rule σ are

[1] Although an actionable variable (e.g. blood group) may not be directly physically manipulable, a causal model such as a structural equation model [16] permits us to compute the effect of intervention on such variables.

defined. Most rule miners measure the effect of a rule using the observed conditional distribution,

$$P(Y \mid \sigma = \top) = \sum_{\sigma(\mathbf{x})=\top} P(Y \mid \mathbf{X} = \mathbf{x}) \,,$$

which captures the correlation or more generally dependence between the rule and the target. To understand the underlying data-generating mechanism, however, we need post-intervention distributions.

One caveat with rules is that, in general, there are many values $\mathbf{x}$ that can satisfy a rule σ (e.g., $\sigma \equiv X_j \leq 3$ is satisfied by $X_j = 3, 2, \ldots$). As a result, we have a multitude of atomic interventions to consider (e.g. for $\sigma \equiv X_j \leq 3$, we have $P(Y \mid do(X_j := 3)), P(Y \mid do(X_j := 2)), \ldots$). Depending on the atomic intervention we choose, we may get different answers. This ambiguity can be avoided by considering the average of all post-intervention distributions where the probability of each atomic intervention is defined by some **stochastic policy** Q_σ [16, Chap. 4]. In reinforcement learning, for instance, a stochastic policy is the conditional probability of an action given some state. Formally, the post-intervention distribution of Y under the stochastic policy Q_σ is given by

$$P(Y \mid do(Q_\sigma)) = \sum_{\sigma(\mathbf{x})=\top} P(Y \mid do(\mathbf{X} := \mathbf{x})) Q_\sigma(do(\mathbf{X} := \mathbf{x})) \,.$$

Let $\bar{\sigma}$ denote the logical negation of σ. Our goal is to identify rules σ that have a high **causal effect** on a specific **outcome** y for the target variable Y, which we define as the difference in the post-intervention probabilities of y under the stochastic policies corresponding to σ and $\bar{\sigma}$, i.e.,

$$e(\sigma) = p(y \mid do(Q_\sigma)) - p(y \mid do(Q_{\bar{\sigma}})) \,,$$

where p represents the probability mass function. Next we show how to compute the above from observational data, and state the stochastic policy to this end.

3.1 Causal Effect from Observational Data

In observational data, we have observed conditional distributions $P(Y \mid \mathbf{X} = \mathbf{x})$ which may not be the same as post-intervention distributions $P(Y \mid do(\mathbf{X} := \mathbf{x}))$. A well-known reason for this discrepancy is the potential presence of **confounders**, i.e., variables that influence both, our desired intervention variable(s) and the target. More generally, to measure the causal effect, we have to eliminate the influence of all *spurious path* in the **causal graph**, i.e., the directed graph that describes the conditional independences of our random variables (with respect to all post-intervention distributions).

In more detail, when estimating the causal effect of X on Y, any undirected path connecting Y and X that has an incoming edge towards X is a **spurious path**. A node (variable) is a **collider** on a path if its in-degree is 2, e.g., Z is a collider on the path $X \to Z \leftarrow Y$. A spurious path is **blocked** by a set of nodes $\boldsymbol{Z}$, if the path contains a collider that is not in $\boldsymbol{Z}$, or a non-collider on the path is in $\boldsymbol{Z}$ [16, Def. 1.2.3]. A set of nodes $\boldsymbol{Z}$ satisfies the **back-door criterion** for a set of nodes $\mathbf{X}$ and a node Y if it blocks all spurious paths from any X in $\mathbf{X}$ to Y, and there is no direct path from any X in $\mathbf{X}$ to any Z in $\boldsymbol{Z}$ [16, Def. 3.3.1]. For $\mathbf{X}$ and Y, if a set $\boldsymbol{Z}$ satisfies the back-door criterion, then observational and post-intervention probabilities are equal within each $\mathbf{z}$ stratum of $\boldsymbol{Z}$:

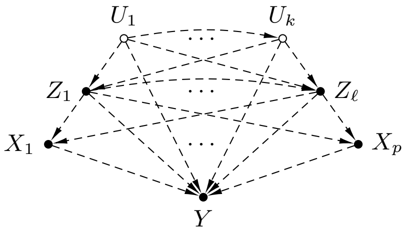


Figure 1: A skeleton causal graph of an admissible input to causal rule discovery (see Def. 1). A dashed edge from a node u to v indicates that u potentially affects v.

$$p(y \mid do(\mathbf{X} := \mathbf{x}), \mathbf{z}) = p(y \mid \mathbf{x}, \mathbf{z}) \,,$$

and averaging the observational probabilities over $\boldsymbol{Z}$ gives $p(y \mid do(\mathbf{X} := \mathbf{x}))$ [16, Thm. 3.3.2].

Therefore, to compute the post-intervention probability of y under the stochastic policy Q_σ for a rule σ, i.e. $p(y \mid do(Q_\sigma))$, we need a set of variables $\boldsymbol{Z}$ that satisfy the back-door criterion for actionable variables $\mathbf{X} \subseteq \mathbb{X}$ and Y. As we consider the rule language $\mathscr{L}$ over all actionable variables $\mathbb{X}$, we require a set of control variables $\mathbb{Z}$ that satisfy the back-door criterion for *all* the actionable variables $\mathbb{X}$. This also implies that there are no other spurious paths via potentially unobserved variables $\boldsymbol{U}$. In the special case when $\mathbb{Z}$ is empty, Y must not cause any actionable variable $X_j \in \mathbb{X}$. We formalise these conditions in the definition below.

Definition 1 (ADMISSIBLE INPUT TO CAUSAL RULE DISCOVERY) *The causal system $(\mathbb{X}, Y, \mathbb{Z})$ of actionable variables, target variable, and control variables is an admissible input to causal rule discovery if the underlying causal graph of the variables satisfy the following:*

(a) there are no outgoing edges from Y to any X in $\mathbb{X}$,
(b) no outgoing edges from any X in $\mathbb{X}$ to any Z in $\mathbb{Z}$,
(c) no edges between actionable variables $\mathbb{X}$, and
(d) no edges between any unobserved U and X in $\mathbb{X}$.

In Fig. 1, we show a skeleton causal graph of an admissible input to causal discovery. The proposition below shows that the control variables $\mathbb{Z}$ block all spurious paths between any subset of actionable variables $\mathbf{X} \in \mathbb{X}$ and Y if the input is admissible.

PROPOSITION 3.1. *Let ($\mathbb{X}$, Y, $\mathbb{Z}$) be an admissible input to causal rule discovery. Then the control variables $\mathbb{Z}$ block all spurious paths between any subset of actionable variables $\mathbf{X} \subseteq \mathbb{X}$ and Y.*

Proof. We provide the proof in the arXiv submission [5]. □

Using admissible control variables $\mathbb{Z}$, we can compute $p(y \mid do(Q_\sigma))$ for any rule σ from the rule language $\mathscr{L}$ as

$$\begin{aligned} p(y \mid do(Q_\sigma)) &= \sum_{\sigma(\mathbf{x})=\top} \sum_{\mathbf{z}\in\mathscr{Z}} p(y \mid \mathbf{x},\mathbf{z}) p(\mathbf{z}) Q_\sigma(do(\mathbf{x})) \\ &= \sum_{\mathbf{z}\in\mathscr{Z}} p(\mathbf{z}) \sum_{\sigma(\mathbf{x})=\top} p(y \mid \mathbf{x},\mathbf{z}) Q_\sigma(do(\mathbf{x})) \ , \end{aligned}$$

where the first expression is obtained by applying the backdoor adjustment formula [16, Thm. 3.3.2] and the second expression is obtained from the first by exchanging the inner summation with the outer one. What is left now is to define the stochastic policy Q_σ which in some sense we treated as an oracle so far. The following theorem shows that, with a specific choice of Q_σ, we can compute the causal effect of any rule on the target, from observational data, in terms of simple conditional expectations (akin to conditional average treatment effect [10]).

THEOREM 3.1. *Given an admissible input to causal rule discovery, $(\mathbb{X}, Y, \mathbb{Z})$, and a stochastic policy $Q_\sigma(do(\mathbf{x})) = p(\mathbf{X} = \mathbf{x} \mid \sigma = \top, \mathbb{Z} = \mathbf{z})$, the causal effect of any rule σ, from the rule language $\mathscr{L}$, on Y in observational data is given by*

$$e(\sigma) = \mathbb{E}[p(y \mid \sigma, \mathbb{Z})] - \mathbb{E}[p(y \mid \bar{\sigma}, \mathbb{Z})] \ . \tag{3.1}$$

Proof. We provide the proof in the arXiv submission [5]. □

That is, for admissible input $(\mathbb{X}, Y, \mathbb{Z})$, the expression above on the r.h.s. gives us the causal effect of any rule σ from the rule language $\mathscr{L}$ on Y from observational data. Importantly, we have shown that causal rule discovery is a difficult problem in practice—any violation of Def. 1 would render Eq. (3.1) non-causal. Having said that, criterion (a) is an implicit assumption in rule discovery, and criterion (b) and (d) are a form of *causal sufficiency* [18], which is a fairly standard assumption in causal inference literature. Criterion (c) is also common in prediction setting in machine learning/statistics literature where the assumption is that independent input features co-cause a target.

Exceptional cases aside, in practice, we often do not know the complete causal graph. While with some assumptions, we can discover a partially directed graph from observational data [21], a rather promising approach is to leverage domain knowledge to eliminate certain variables following the guidelines in Def. 1. For instance, *smoking* causes *tar deposits* in a person's lungs, therefore both *smoking* and *tar deposits* cannot be in $\mathbb{X}$; this ensures that criterion (c) of Def 1 is not violated. Moreover, *smoking* may affect a person's *blood pressure*. Thus it is unsafe to include *blood pressure* in $\mathbb{Z}$—criterion (b) would be violated otherwise. This way, we can get a practical solution that is closer to the truth.

3.2 Statistical Considerations In practice, we want to estimate $e(\sigma)$ (Eq. (3.1)) from a sample drawn from the population. Suppose that we have a sample of N instances **stratified** by $\mathbb{Z}$ from the population (or in practice, the sample size is large enough to give relatively accurate estimates of the marginal distribution of $\mathbb{Z}$). The naive estimator of the causal effect $e(\sigma)$ is the estimator based on the empirical distribution $\hat{P}$ (resp. $\hat{p}$ for pmf), i.e. the **plug-in** estimator:

$$\begin{aligned} \widehat{e}(\sigma) &= \mathbb{E}[\hat{p}(y \mid \sigma, \mathbf{z}) - \hat{p}(y \mid \bar{\sigma}, \mathbf{z})] \\ &= \sum_{\mathbf{z}\in\mathscr{Z}} \Big(\hat{p}(y \mid \sigma, \mathbf{z}) - \hat{p}(y \mid \bar{\sigma}, \mathbf{z})\Big) \hat{p}(\mathbf{z}) \\ &= \sum_{\mathbf{z}\in\mathscr{Z}} (\hat{p}_{\sigma,\mathbf{z}} - \hat{p}_{\bar{\sigma},\mathbf{z}}) \hat{p}(\mathbf{z}) \ , \end{aligned}$$

where $\hat{p}_{\sigma,\mathbf{z}} = \hat{P}(y \mid \sigma, \mathbf{z})$, and $\hat{p}_{\bar{\sigma},\mathbf{z}} = \hat{P}(y \mid \bar{\sigma}, \mathbf{z})$. In a stratified sample, $\hat{p}(\mathbf{z})$ is the same as $p(\mathbf{z})$. As the empirical distribution is a consistent estimator of the population distribution, $\widehat{e}(\sigma)$ is a consistent estimator of $e(\sigma)$.

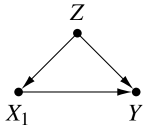
The plug-in estimator, however, shows high variance for rules with overly small sample sizes for either of the two events, σ or $\bar{\sigma}$. In Fig. 3 (left), we show the estimated mean and variance of the plug-in estimator for a very specific rule of five conditions, and see that while it is close to the true causal effect, it shows very high variance in small samples. This high variance is problematic, as it leads to overfitting: if we use this estimator for the optimisation task over a very large space of rules, the variance will turn into a strong positive bias—we will overestimate the effects of rules from the sample—that dominates the search. We, hence, end up with random results of either extremely specific or general rules.

We address this problem of high variance by biasing the plug-in estimator. In particular, we introduce bias in terms of our confidence in the point estimates using confidence intervals. Note that we need not quantify the confidence of the point estimate $\hat{p}(\mathbf{z})$ as $\hat{p}(\mathbf{z}) = p(\mathbf{z})$; the point estimates of concern are the conditional probabilities $\hat{p}_{\sigma,\mathbf{z}}$ and $\hat{p}_{\bar{\sigma},\mathbf{z}}$.

In repeated random samples of instances with $\sigma = \top$ and $\mathbb{Z} = \mathbf{z}$ from the population, the number of instances with *successful* outcome y is a binomial random variable with the success probability $p(y \mid \sigma, \mathbf{z})$. In a stratum $\mathbf{z}$ of $\mathbb{Z}$, let $n_{\sigma,\mathbf{z}}$ and $n_{\bar{\sigma},\mathbf{z}}$ be the number of instances that satisfy σ and $\bar{\sigma}$, respectively. Then the one-sided binomial confidence interval of $\hat{p}_{\sigma,\mathbf{z}}$, using a normal approximation of the error distribution, is given by $\beta\sqrt{\hat{p}_{\sigma,\mathbf{z}}(1-\hat{p}_{\sigma,\mathbf{z}})/n_{\sigma,\mathbf{z}}}$, where β is the $1 - \alpha/2$ quantile of a standard normal distribution for an error rate α, or simply the **z-score** corresponding to the confidence level. For a 95% confidence level, for instance, the error rate is $\alpha = 0.05$, thereby $\beta = 1.96$. We can

$P(Z=1)$	$P(Z=0)$
0.9	0.1

Z	$P(X_1=1\,|\,Z)$	$P(X_1=0\,|\,Z)$
1	0.8	0.2
0	0.5	0.5



		$P(Y\,|\,X_1,Z)$	
Z	X_1	$Y=1$	$Y=0$
1	1	0.7	0.3
1	0	0.5	0.5
0	1	0.5	0.5
0	0	0.4	0.6

Figure 2: A causal graph with corresponding conditional probability table for each node.

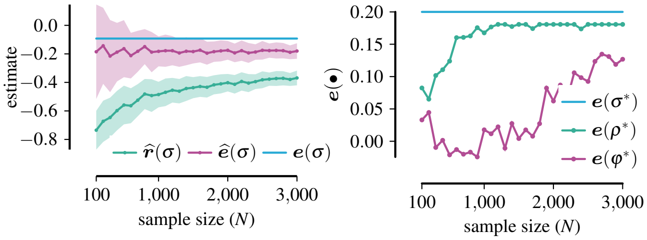


Figure 3: From the population generated using the causal graph of Fig. 2 together with 5 additional independent actionable variables $X_2,\ldots,X_6$, we show (left) variance of the plug-in and the reliable estimator of the causal effect for a specific rule that contains variables that are independent of the target, and (right) overfitting phenomenon in rule learning, i.e. the sample-optimal rule φ^* chosen by the plug-in estimator leads to a causal effect farther from the reference $e(\sigma^*)$ than the sample-optimal rule ρ^* chosen by the reliable estimator.

easily verify that the maximum value of $\hat{p}_{\sigma,\mathbf{z}}(1-\hat{p}_{\sigma,\mathbf{z}})$ is $1/4$, and hence the maximum value of the one-sided confidence interval is $\beta/(2\sqrt{n_{\sigma,\mathbf{z}}})$. Taking a conservative approach, we bias the difference $\hat{p}_{\sigma,\mathbf{z}}-\hat{p}_{\bar{\sigma},\mathbf{z}}$ by subtracting the sum of the maximum values of the one-sided confidence intervals of the point estimates, this results in

$$\tau(\mathbf{z}) = (\hat{p}_{\sigma,\mathbf{z}}-\hat{p}_{\bar{\sigma},\mathbf{z}}) - \left(\beta/(2\sqrt{n_{\sigma,\mathbf{z}}}) + \beta/(2\sqrt{n_{\bar{\sigma},\mathbf{z}}})\right) .$$

Note that $\tau(\mathbf{z})$ lower bounds the true probability mass difference in the population with confidence $1-\alpha$. That is, there is a $1-\alpha$ chance that the true difference is larger than $\tau(\mathbf{z})$. For a fixed β, the lower bound gets tighter with increasing sample size. In fact, it is easy to see that $\tau(\mathbf{z})$ is a **consistent** estimator of the true probability mass difference in the population; the introduced bias term vanishes asymptotically. More formally, for a fixed finite β, we have

$$\lim_{\min(n_{\sigma,\mathbf{z}},n_{\bar{\sigma},\mathbf{z}})\to\infty} \beta/(2\sqrt{n_{\sigma,\mathbf{z}}}) + \beta/(2\sqrt{n_{\bar{\sigma},\mathbf{z}}}) = 0 .$$

As we deal with empirical probabilities, we can express $\tau(\mathbf{z})$ in terms of cell counts in a contingency table. Suppose that we have a contingency table as shown in Tab. 1 (left) for a $\mathbf{z}$ stratum. Then we can express $\tau(\mathbf{z})$ as

$$\tau(\mathbf{z}) = \frac{a}{n_{\sigma,\mathbf{z}}} - \frac{c}{n_{\bar{\sigma},\mathbf{z}}} - \frac{\beta}{2\sqrt{n_{\sigma,\mathbf{z}}}} - \frac{\beta}{2\sqrt{n_{\bar{\sigma},\mathbf{z}}}} .$$

In the extreme case, however, a rule may select all or none of the instances in a stratum, resulting in $n_{\sigma,\mathbf{z}}=0$ or $n_{\bar{\sigma},\mathbf{z}}=0$, and hence the empirical conditional probability mass functions can be undefined. In practice, we encounter this problem often, both due to specificity of a rule as well as small sample sizes to begin with.

As a remedy, we apply the Laplace correction to the score. That is, we increment count of each cell in the contingency table by one. This way we start with a uniform distribution within each stratum of $\mathbb{Z}$. Hence a stratum of size n increases to $n+4$, and the total effective sample size increases from N to $N+4|\mathscr{Z}|$. After applying Laplace correction, we have $\hat{P}(\mathbf{z}) = (n+4)/(N+4|\mathscr{Z}|)$, and $\tau(\mathbf{z})$ is given by

$$\tau(\mathbf{z}) = \frac{a+1}{n_{\sigma,\mathbf{z}}+2} - \frac{c+1}{n_{\bar{\sigma},\mathbf{z}}+2} - \frac{\beta}{2\sqrt{n_{\sigma,\mathbf{z}}+2}} - \frac{\beta}{2\sqrt{n_{\bar{\sigma},\mathbf{z}}+2}} .$$

Combining these, we obtain the **reliable** estimator of $e(\sigma)$ as

$$\widehat{r}(\sigma) = \sum_{\mathbf{z}\in\mathscr{Z}} \tau(\mathbf{z})\hat{p}(\mathbf{z}) . \tag{3.2}$$

Note that $\widehat{r}(\sigma)$ is still a **consistent** estimator of the causal effect. To demonstrate this, consider the following example.

Suppose that we generate the population using the causal graph in Fig. 2. In addition, we generate five uniformly distributed binary actionable variables $X_2,X_3,\ldots,X_6$ that are independent of each other as well as the rest of the variables. We can now numerically estimate the variance of the two estimators for a specific rule, e.g. $\sigma \equiv X_1=1 \wedge X_2=0 \wedge X_3=1 \wedge X_4=1 \wedge X_5=0 \wedge X_6=0$, which does not only contain causal variable X_1 but also five actionable variables that are independent of the target Y.

To this end, we draw stratified samples of increasing sizes from the population, and report $\widehat{e}(\sigma)$ and $\widehat{r}(\sigma)$ scores averaged over 25 simulations along with one sample standard deviation in Fig. 3 (left). We observe that variances of both estimators decrease with increasing sample size. Although the reliable estimator is biased, its variance is relatively low compared to the plug-in estimator. As a result of this low variance, unlike the plug-in estimator, the reliable estimator is able to avoid overfittinga. Let σ^* denote the top-1 rule in the population, i.e. $\sigma^* = \text{argmax}_{\sigma\in\mathscr{L}}\, e(\sigma)$. Let φ^* denote the top-1 rule using the plug-in estimator, i.e. $\varphi^* = \text{argmax}_{\sigma\in\mathscr{L}}\, \widehat{e}(\sigma)$, and ρ^* denote the top-1 rule using the reliable estimator, i.e. $\rho^* = \text{argmax}_{\sigma\in\mathscr{L}}\, \widehat{r}(\sigma)$. In Fig. 3 (right), we plot $e(\varphi^*)$ against $e(\rho^*)$. We observe that with increasing sample sizes $e(\rho^*)$ is both relatively closer, as well as converges much faster to the reference $e(\sigma^*)$.

4 Discovering Rules

Now that we have a reliable and consistent estimator of the causal effect, we turn to discovering rules that maximize this estimator. Below, we provide the formal problem definition.

Definition 2 (TOP-k CAUSAL RULE DISCOVERY) *Given a sample and a positive integer k, find a set $\mathscr{F}_k \subseteq \mathscr{L}$, $|\mathscr{F}_k| = k$, such that for all $\sigma \in \mathscr{F}_k$ and $\varphi \in \mathscr{L} \setminus \mathscr{F}_k$, $\widehat{\boldsymbol{r}}(\sigma) \geq \widehat{\boldsymbol{r}}(\varphi)$.*

Given the hardness of empirical effect maximisation problems [23], it is unlikely that the optimisation of the reliable causal effect allows a worst-case polynomial algorithm. While the exact computational complexity of the causal rule discovery problem is open, here we proceed to develop a practically efficient algorithm using the branch-and-bound.

4.1 Branch-and-Bound Search The branch-and-bound search scheme finds a solution that optimises the objective function $f : \Omega \to \mathbb{R}$, among a set of admissible solutions Ω, also called the search space. Let $\text{ext}(\sigma)$, also called the extension of σ, denote the subset of instances in the sample that satisfy σ. The generic search scheme for a branch-and-bound algorithm requires the following two ingredients:

- A **refinement operator** $\boldsymbol{b} : \mathscr{L} \to \mathscr{P}(\mathscr{L})$ that is monotone, i.e. for $\sigma, \varphi \in \mathscr{L}$ with $\varphi = \boldsymbol{b}(\sigma)$ it holds that $\text{ext}(\varphi) \subseteq \text{ext}(\sigma)$, and that non-redundantly generates the search space $\mathscr{L}$. That is, for every rule $\sigma \in \mathscr{L}$, there is a unique sequence of rules $\sigma_0, \sigma_1, \ldots, \sigma_\ell = \sigma$ with $\sigma_i = \boldsymbol{b}(\sigma_{i-1})$.
- An **optimistic estimator** $\tilde{f} : \Omega \to \mathbb{R}$ that provides an upper bound on the objective function attainable by extending the current rule to more specific rules. That is, it holds that $\tilde{f}(\sigma) \geq f(\varphi)$ for all $\varphi \in \mathscr{L}$ with $\text{ext}(\varphi) \subseteq \text{ext}(\sigma)$.

A branch-and-bound algorithm simply enumerates the search space $\mathscr{L}$ starting from the root ϕ using the refinement operator $\boldsymbol{b}$ (branch), but based on the optimistic estimator $\tilde{f}$ prunes those branches that cannot yield improvement over the best rules found so far (bound).

The optimistic estimator depends on the objective function, and there are many optimistic estimators for an objective function. Not all of these are equally well-suited in practice, as the tightness of the optimistic estimator determines its pruning potential. We consider the **tight optimistic estimator** [9]

$$\begin{aligned}\tilde{f}(\sigma) &= \max\{f(Q) \mid Q \subseteq \text{ext}(\sigma)\} \\ &\geq \max\{f(\varphi) \mid \text{ext}(\varphi) \subseteq \text{ext}(\sigma) \text{ for all } \varphi \in \mathscr{L}\} .\end{aligned}$$

The branch-and-bound search scheme also provides an option to trade-off the optimality of the result for the speed. Instead of asking for the f-optimal result, we can ask for the γ-approximation result for some approximation factor $\gamma \in (0, 1]$. This is done by relaxing the optimistic estimator, i.e. $\tilde{f}(\sigma) \geq \gamma f(\varphi)$ for all $\varphi \in \mathscr{L}$ with $\text{ext}(\varphi) \subseteq \text{ext}(\sigma)$. Lower γ generally yields better pruning, at the expense of guarantees on the quality of the solution.

In our problem setting, we can define the refinement operator based on the lexicographical ordering of propositions:

$$\boldsymbol{b}(\sigma) = \{\sigma \wedge \pi_i \mid \pi_i \in \boldsymbol{\pi}, i > \max\{j : \pi_j \in \boldsymbol{\pi}^{(\sigma)}\}\} ,$$

where $\boldsymbol{\pi}$ is the set of propositions and $\boldsymbol{\pi}^{(\sigma)}$ is the subset of $\boldsymbol{\pi}$ used in σ. In practice, we need more sophisticated refinement operators in order to avoid the inefficiency resulting from a combinatorial explosion of equivalent rules. This, we can do by defining a closure operator on the rule language (see, e.g. Boley & Grosskreutz [4]), which we also employ in our experimental evaluation. Next we derive an optimistic estimator for the objective function $\widehat{\boldsymbol{r}}$.

4.2 Efficient optimistic estimator If we look at the definition of $\widehat{\boldsymbol{r}}(\sigma)$ in Eq. (3.2), we see that, regardless of σ, $\hat{p}(\mathbf{z})$ remains the same for a $\mathbf{z}$ stratum. Thus, we can obtain an optimistic estimator of $\widehat{\boldsymbol{r}}(\sigma)$ by simply bounding $\tau(\mathbf{z})$ for each $\mathbf{z}$ stratum. Let $\tilde{\tau}(\mathbf{z})$ denote the optimistic estimator of $\tau(\mathbf{z})$. Then the optimistic estimator of $\widehat{\boldsymbol{r}}(\sigma)$ is given by

$$\tilde{\boldsymbol{r}}(\sigma) = \sum_{\mathbf{z} \in \mathscr{Z}} \tilde{\tau}(\mathbf{z}) \hat{p}(\mathbf{z}) .$$

To derive the optimistic estimator $\tilde{\tau}(\mathbf{z})$, for clarity of exposition we first project $\tau(\mathbf{z})$ in terms of free variables a and b, such that we can write

$$\tau(a,b) = \frac{a+1}{a+b+2} - \frac{n_1 - a + 1}{n - a - b + 2} - \frac{0.5\beta}{\sqrt{a+b+2}} - \frac{0.5\beta}{\sqrt{n-a-b+2}} .$$

Suppose that we have a contingency table as shown in Tab. 1 (left) for a $\mathbf{z}$ stratum with the rule σ. The refinement of σ, $\sigma' = \boldsymbol{b}(\sigma)$, results in a table as shown in Tab. 1 (right). Note that n_1, n_0, and n do not change within a $\mathbf{z}$ stratum regardless of the rule. Since $\text{ext}(\sigma') \subseteq \text{ext}(\sigma)$ holds for any $\sigma' = \boldsymbol{b}(\sigma)$, we have the following relations: $a' \leq a$ and $b' \leq b$.

This implies that the subsets of the extensions of σ will have contingency table counts a' in the range $\{0, 1, \ldots, a\}$, and b' in the range $\{0, 1, \ldots, b\}$. Let $\mathscr{C} = \{0, 1, \ldots, a\} \times \{0, 1, \ldots, b\}$. Then the optimistic estimator of $\tau(\mathbf{z})$ can be defined in terms of $\mathscr{C}$ as

$$\tilde{\tau}(\mathbf{z}) \geq \max_{(a',b') \in \mathscr{C}} \tau(a', b') .$$

The following proposition shows that we can obtain the **tight optimistic estimate** of $\tau(\mathbf{z})$ in linear time.

PROPOSITION 4.1. *Let $\mathscr{C} = \{0, 1, \ldots, a\} \times \{0, 1, \ldots, b\}$ be the set of all possible configurations of (a', b') in Tab. 1 (right) that can result from refinements of a rule σ from the contingency table of Tab. 1 (left). Then the* ***tight optimistic***

Table 1: Contingency tables for (left) a rule σ, and (right) its refinement $\sigma' = \boldsymbol{b}(\sigma)$ for a $\mathbf{z}$ stratum of $\mathbb{Z}$.

	$Y=y$	$Y\neq y$	
$\sigma=\top$	a	b	
$\sigma=\bot$	c	d	
Σ	n_1	n_0	n

	$Y=y$	$Y\neq y$	
$\sigma'=\top$	a'	b'	
$\sigma'=\bot$	c'	d'	
Σ	n_1	n_0	n

***estimator** of $\tau(\mathbf{z})$ is given by*

$$\tilde{\tau}_t(\sigma,\mathbf{z}) = \max_{a'\in\{0,1,\dots,a\}} \frac{a'+1}{a'+2} - \frac{n_1-a'+1}{n-a'+2} - \frac{\beta}{2\sqrt{a'+2}} - \frac{\beta}{2\sqrt{n-a'+2}}.$$

Proof. We provide the proof in the arXiv submission [5]. □

5 Experiments

We implemented the branch-and-bound search with priority-queue in realKD[2] Java library, and provide the source code online.[3] All experiments were executed single threaded on Intel Xeon E5-2643 v3 machine with 256 GB memory running Linux. We report the results at $\beta = 2.0$, which corresponds to a 95.45% confidence level, and search for optimal top-k rules, i.e. $\gamma = 1.0$, unless stated otherwise.

5.1 Performance of the Estimators First we evaluate the performance of the proposed estimators of causal effect $\boldsymbol{e}(\sigma)$. To this end, we measure the *statistical efficiency* of an estimator by its **mean squared error (MSE)** as it captures the two most important properties of an estimator: bias and variance. As the optimistic bias is strongest for the best rule and decreases monotonically, we consider the top-1 search here. Thus the parameter of interest in the population is the maximum value of the causal effect $\boldsymbol{e}(\sigma^*)$, where σ^* is the maximiser in the population, i.e. $\sigma^* = \text{argmax}_{\sigma\in\mathscr{L}}\, \boldsymbol{e}(\sigma)$. Using the reliable estimator $\widehat{\boldsymbol{r}}$, for instance, we get the reliable effect maximiser ρ^* in the sample, i.e. $\rho^* = \text{argmax}_{\sigma\in\mathscr{L}}\, \widehat{\boldsymbol{r}}(\sigma)$. As such, $\boldsymbol{e}(\rho^*)$ is our estimate of the estimand $\boldsymbol{e}(\sigma^*)$, using $\widehat{\boldsymbol{r}}$. Note that $\boldsymbol{e}(\rho^*)$ is a function of the sample, and thus a random variable. Therefore the MSE of $\boldsymbol{e}(\rho^*)$ is given by

$$\text{MSE}\left(\boldsymbol{e}(\rho^*)\right) = \mathbb{E}_{\boldsymbol{e}(\rho^*)}\left[\left(\boldsymbol{e}(\rho^*) - \boldsymbol{e}(\sigma^*)\right)^2\right] .$$

Likewise, we can obtain the MSE of $\boldsymbol{e}(\varphi^*)$ using the plug-in estimator $\widehat{\boldsymbol{e}}$, where $\varphi^* = \text{argmax}_{\sigma\in\mathscr{L}}\, \widehat{\boldsymbol{e}}(\sigma)$.

For this evaluation, first we generate the population using the causal graph in Fig. 2, and add five *independent* uniformly distributed binary actionable variables $X_2, X_3, \dots, X_6$. Then, for a given sample size N, we sample N observations from that population, and compute the MSE of the two estimators over 100 samples. In Fig. 4 (left), we show the MSE of the estimators for increasing sample sizes $N = 100, 200, \dots, 3000$. As expected, we observe that the MSE decreases for both estimators as the sample size increases. The reliable estimator, however, has a consistently lower MSE than the plug-in estimator. These results show that the proposed reliable estimator is a better choice for optimisation (search) than the naive plug-in estimator.

5.2 Comparison with the state-of-the-art Next we investigate the quality of rules inferred using the reliable estimator, and compare against other state-of-the-art measures. Although there exists a number of algorithms to infer interesting rules from data, most of them do not provide us *optimal* causal rules. Therefore, in this evaluation, we focus mainly on effect measures they employ, as we can always exhaustively search for optimal rules using those effect measures as long as we keep the rule language small.

From the exhaustive list of effect measures, we consider the weighted relative accuracy [12] for comparison, primarily because it is widely used in inductive rule learners. In addition, we also consider the plug-in estimator without control variables, i.e. $\widehat{\boldsymbol{e}}(\sigma)$ with $\mathbb{Z} := \emptyset$. In our case, the weighted relative accuracy of the event σ for an outcome y at the population level is given by

$$\boldsymbol{w}(\sigma) = p(\sigma)\Big(p(y \mid \sigma) - p(y)\Big) .$$

In particular, we apply Laplace correction to the plug-in estimators of both the weighted relative accuracy, $\widehat{\boldsymbol{w}}(\sigma)$, and $\widehat{\boldsymbol{e}}(\sigma) \mid \mathbb{Z} := \emptyset$.

To obtain synthetic data with the known ground truth, we sample observations from the population in our previous evaluation (Sec. 5.1). In the causal graph (Fig. 2), only one actionable variable (X_1) affects the target Y; other actionable variables $X_2, \dots, X_6$ are independent. As such, only one rule $\sigma \equiv X_1 = 1$ is *relevant*.[4] We assess an effect measure by evaluating the probability of recovering that single "true" rule.

To this end, we take 100 samples, find optimal top-1 rule from each sample, and then calculate the proportion of "true" rule among those 100 optimal top-1 rules. In Fig. 4 (right), we report the probability of recovering the "true" rule at various sample sizes. We observe that the reliable causal effect is consistently better than other effect measures, and its probability of recovering the core rule exactly approaches 1.0 rapidly with increasing sample size. Other effect measures perform worse particularly when the sample size is small. These results demonstrate that by conditioning on the control

[2] `https://bitbucket.org/realKD/`

[3] https://www.dropbox.com/s/frlv8os29gobcby/rce.zip?dl=0

[4] The complementary rule $\bar{\sigma} \equiv X_1 = 0$ has a *negative* effect.

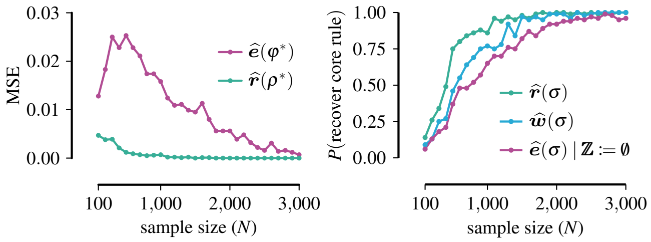


Figure 4: (left) Mean squared error (MSE) of the plug-in estimator $\widehat{e}$ and the reliable estimator $\widehat{r}$ of the population optimal causal effect $e(\sigma^*)$. (right) The probability of recovering the core rule from 100 samples, for the Laplace-corrected plug-in estimator of causal effect, $\widehat{e}(\sigma)$, with an empty $\mathbb{Z}$, the Laplace corrected plug-in estimator of weighted relative accuracy, $\widehat{w}(\sigma)$, and the reliable estimator of causal effect, $\widehat{r}(\sigma)$.

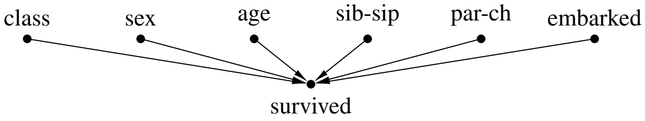


Figure 5: Assumed causal graph of the `titanic` dataset. *par-ch* stands for the number of parents/children aboard, and *sib-sp* for the number of siblings/spouses aboard.

Table 2: Top-3 causal rules discovered on the `titanic` dataset with "survival" as a target variable.

Top-3 rules (σ)	cvg(σ)	$\widehat{r}(\sigma)$
class $\leq 2 \land$ sex = female	0.1907	0.576
class $\leq 2 \land$ sex = female $\land$ par-ch ≤ 2	0.1885	0.573
class $\leq 2 \land$ sex = female $\land$ sib-sp ≤ 2	0.1874	0.572

variables, reliable causal effect infers relevant causal rules, even on small sample sizes.

5.3 Qualitative Study on Real-World Data Next we investigate whether rules discovered by reliable causal effect are meaningful. To this end, we consider the `titanic` training set from the Kaggle repository.[5] The sinking of RMS Titanic is one of the most notorious shipwrecks in history. One of the reasons behind such tragic loss of lives was the lack of lifeboats. During the evacuation, some passengers were treated differently than the others; some groups of people were, hence, more likely to survive than the others. Thus, it is of interest to find the conditions that have causal effect on the *survival* (Y). The dataset contains demographics and travel attributes of the passengers.

Existing causal discovery methods are not applicable as we have a mixed data set. We also do not know the complete causal graph. Therefore we take a pragmatic approach using domain knowledge. If we could perform a hypothetical intervention of changing the sex of a person, this will also change their title, but not the other way around. Thus it is reasonable to assume that *sex* causes *title*. As putting them together in $\mathbb{X}$ would violate criterion (c) of Def. 1, we only keep one of them, i.e. *sex*, in $\mathbb{X}$. Similarly we can argue that fare causes passenger class. Therefore we only keep *class* in $\mathbb{X}$. Overall, none of the variables seem to confound (co-cause) Y and other variables. Thus, we have the causal graph as shown in Fig. 5, where $\mathbb{Z} := \emptyset$, $Y :=$ *survived*, and $\mathbb{X} := \{class, sex, age, sib\text{-}sip, par\text{-}ch, embarked\}$.

In Tab. 2, we present optimal top-3 causal rules discovered from the input above using the proposal method. The coverage of a rule is a fraction of instances that belong to its extension, i.e. $\text{cvg}(\sigma) = |\text{ext}(\sigma)|/N$. We observe that being a female passenger from the first, or the second class has the highest effect on survival with a reliable causal effect estimate of $\widehat{r}(\sigma_1) = 0.576$. It is well-known that passengers from different classes were treated differently during evacuation. What is interesting is that although females were more likely to survive, this only applied to the females from the first and the second class; this is also corroborated by the fact that roughly half of the females from the third class did not survive the mishap compared to the only one-tenth from the other two classes combined. The other two rules corroborate the adage of "women and children" first. Those rules appearing on top with mere $< 20\%$ coverage shows that reliable causal effect can discover rare rules.

6 Discussion

The main focus of this discussion are the assumptions (in Def. 1) required for causal rule discovery and their practical implications. First we note that it is *impossible* to do causal inference from observational data without making assumptions, as the joint distribution alone cannot tell us what happens when we intervene on the system [16]. Through causal diagrams, we make such assumptions more explicit and transparent. Often, the more explicit the assumption, the more criticism it invites. Explicit assumptions, however, can be good as they provide a way to verify our models, and improve them. Def. 1, for instance, provides guidelines for variable selection process for causal rule discovery.

For the inferred rules to be causal, the input must be admissible. Although criterion (a), (b) and (d) are fairly standard in causality literature, criterion (c) is specific to causal rule discovery. However. in machine learning, we also implicitly assume that independent input features co-cause the target. Over a large group of actionable variables, this can be a strong assumption. The naive way to remove

[5] `https://www.kaggle.com/c/titanic`

this assumption would be to include rest of the actionable variables $\mathbb{X} \setminus \mathbf{X}$ in the set of control variables to block any spurious path between actionable variables $\mathbf{X}$ in a rule σ and the target Y via $\mathbb{X} \setminus \mathbf{X}$. By doing so, however, we may not only violate other criteria, but the search also gets complicated.

A direct correction for controlling the familywise error rate of confidence intervals would not lead to an effective approach to discover causal effects. Therefore, we followed the statistical learning approach and designed an estimator that avoids overfitting. This use of confidence intervals is reminiscent of upper confidence bound strategies in multi-armed bandit problems (yielding an optimal policy despite not controlling the familywise error rate of reward estimates).

7 Conclusion

Traditional descriptive rule discovery techniques do not suffice for discovering reliable causal rules from observational data. Among the sources of inconsistency we have that observational effect sizes are often skewed by the presence of confounding factors. Second, naive empirical effect estimators have a high variance, and, hence, their maximisation is highly optimistically biased unless the search is artificially restricted to high frequency events. In this work, we presented a causal rule discovery approach that addresses both these issues. We measured the causal effect of a rule from observational data by adjusting for the effect of potential confounders. In particular, we gave the graphical criteria under which causal rule discovery is possible. To discover reliable causal rules from a sample, we proposed a conservative and consistent estimator of the causal effect, and derived an efficient and exact algorithm based on branch-and-bound search that maximises the estimator. The proposed algorithm is efficient and finds meaningful rules. It would make an interesting future work to develop an effect measure, and the algorithm that works for a continuous real-valued target and gracefully handles a large set of control variables.

Acknowledgements

The authors thank the anonymous reviewers for detailed comments. KB was supported by IMPRS-CS. MB is supported by Monash University. JV is supported by CISPA Helmholtz Center for Information Security.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, 1993.

[2] M. Atzmueller and F. Puppe. A knowledge-intensive approach for semi-automatic causal subgroup discovery. In *Knowledge Discovery Enhanced with Semantic and Social Information*, pages 19–36. Springer, 2009.

[3] S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *DAMI*, 5(3):213–246, 2001.

[4] M. Boley and H. Grosskreutz. Non-redundant subgroup discovery using a closure system. In *MLKDD*, pages 179–194. Springer, 2009.

[5] K. Budhathoki, M. Boley, and J. Vreeken. Discovering reliable causal rules, 2020, arXiv:2009.02728.

[6] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *KDD*, pages 43–52, New York, NY, USA, 1999. ACM.

[7] J. Fürnkranz and P. A. Flach. ROC 'n' rule learning - towards a better understanding of covering algorithms. *Machine Learning*, 58(1):39–77, 2005.

[8] J. Fürnkranz, D. Gamberger, and N. Lavrač. *Foundations of Rule Learning*. Cognitive Technologies. Springer, 2012.

[9] H. Grosskreutz, S. Rüping, and S. Wrobel. Tight optimistic estimates for fast subgroup discovery. In *MLKDD*, pages 440–456. Springer, 2008.

[10] M. A. Hernán and J. M. Robins. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, 2020.

[11] J. Kalofolias, M. Boley, and J. Vreeken. Efficiently discovering locally exceptional yet globally representative subgroups. In *IEEE ICDM*, pages 197–206, 2017.

[12] N. Lavrač, P. Flach, and B. Zupan. Rule evaluation measures: A unifying view. In *Inductive Logic Programming*, pages 174–185. Springer, 1999.

[13] N. Lavrač, B. Kavsek, P. A. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *JMLR*, 5:153–188, 2004.

[14] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma. From observational studies to causal rule mining. *ACM Trans. Intell. Syst. Technol.*, 7(2):14:1—14:27, 2015.

[15] L. Papaxanthos, F. Llinares-López, D. Bodenham, and K. Borgwardt. Finding significant combinations of features in the presence of categorical covariates. In *Advances in Neural Information Processing Systems 29*, pages 2279–2287. Curran Associates, Inc., 2016.

[16] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.

[17] J. R. Quinlan and R. M. Cameron-Jones. Induction of logic programs: FOIL and related systems. *New Generation Comput.*, 13(3&4):287–312, 1995.

[18] R. Scheines. An introduction to causal inference. In *Causality in Crisis? University of Notre Dame*, pages 185–200, 1997.

[19] P. Shamsinejadbabaki, M. Saraee, and H. Blockeel. Causality-based cost-effective action mining. *Intelligent Data Analysis*, 17(6):1075–1091, 2013.

[20] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. *DAMI*, 4(2):163–192, 2000.

[21] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT press, 2000.

[22] F. Wang and C. Rudin. Causal falling rule lists. In *FATML Workshop*, 2017.

[23] L. Wang, H. Zhao, G. Dong, and J. Li. On the complexity of finding emerging patterns. *Theoretical Computer Science*, 335(1):15–27, 2005.

[24] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.