

# Discovering Reliable Causal Rules

Kailash Budhathoki\*

Mario Boley<sup>°</sup>

Jilles Vreeken<sup>•</sup>

## 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 can lead to random results.

To address these issues, first we measure the causal effect of a rule from observational data—adjusting for the effect of potential confounders. Importantly, we provide a graphical criteria 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. On synthetic data, the proposed estimator converges faster to the ground truth than the naive estimator and recovers relevant causal rules even at small sample sizes. Extensive experiments on a variety of real-world datasets show that the proposed algorithm is efficient and discovers meaningful rules.

## 1 Introduction

The ultimate goal of meaningful data analysis is to understand how the data was generated, by reasoning in terms of cause and effect. Towards this goal, rule mining [1, 6, 8, 23] 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. As a result, rules that maximise such effect measures are in no way guaranteed to reflect the underlying 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 impossible or at the very least impractical 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 often referred to as 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 [15, 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 may find that the treatment is *not* effective at all to male and female subpopulations.

The statistical problem is the well-known phenomenon of overfitting. This phenomenon 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.

Third, the rule space over which we maximise causal effect is exponential in size and does not exhibit structure 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, we propose to measure the causal effect of a rule from observational data. To this end, we control for the effect of a given set of potential confounder variables. In particular, we give a graphical criteria under which it is possible to discover causal rules. While in practice the set of control variables will rarely be complete, i.e., not contain all potential confounders, this approach can rule out specific alternative explanations of findings as well as eliminate misleading observations caused by selected observables that are

\*Amazon Research Tübingen (work done prior to joining Amazon)

<sup>°</sup>Monash University

<sup>•</sup>CISPA Helmholtz Center for Information Security

known to be strong confounders. In fact, this pragmatic approach is usually a necessity caused by sparsity.

2. To address the overfitting problem, we propose to measure and optimise the *reliable* effect of a rule. In contrast to the plug-in estimator, we propose a conservative empirical estimate of the population effect, that is not prone to overfitting. Additionally, and in contrast to other known rule optimisation criteria, it is also *consistent*, i.e., with increasing amounts of evidence (data), 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

**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 [16]. 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, 5], we can get reliable patterns whose *supports* differ significantly between datasets by performing a statistical hypothesis test. Most subgroup discovery [23] 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.

**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. [19] 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 sub-

groups using a chosen quality function, inferring a causal network over these, and visually presenting this to the user. Causal falling rule lists [21] 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. [18] discover actions from a partial directed acyclic graph for which the post-intervention probability of  $Y$  differs from the observational probability.

While all these methods have opened the research direction, we still lack a theoretical understanding. Roughly speaking, all these methods propose to condition “some” effect measure on “some” covariates. In this work, we present a theoretical result showing which covariates to condition upon, under what conditions causal rule discovery is possible, and how an effect measure must be constructed to capture causal effect. Overall, despite the importance of the problem, to the best of our knowledge there does not exist a theoretically well-founded, efficient approach to discovering reliable causal rules from observational 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**<sup>1</sup>  $\mathbf{X} := (X_1, \dots, X_\ell)$  and **control variables**  $\mathbf{Z} := (Z_1, \dots, Z_m)$ . For example,  $Y$  might indicate recovery from a disease,  $\mathbf{X}$  different medications that can be administered to a patient, and  $\mathbf{Z}$  might be attributes of patients, such as blood group. Let  $\mathcal{X}_j$  denote the domain of  $X_j$ , and  $\mathcal{Z}_j$  be that of  $Z_j$ . As such, the domain of  $\mathbf{X}$  is the Cartesian product  $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_\ell$ , and that of  $\mathbf{Z}$  is  $\mathcal{Z} = \mathcal{Z}_1 \times \dots \times \mathcal{Z}_m$ .

We use Pearl’s do-notation [15, 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  $\mathcal{S}$  be the set of all possible vector values of all possible subsets of actionable variables. More formally, we have the following definition:

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

<sup>1</sup>Although an actionable variable (e.g. blood group) may not be directly physically manipulable, a causal model such as a structural equation model [15] permits us to compute the effect of intervention on such variables.

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

Let  $\mathbf{X} \subseteq \mathbf{X}$  denote the subset of actionable variables, with their joint domain  $\mathcal{X}$ , on which propositions of a rule  $\sigma$  are 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  $\sigma$  (e.g.,  $\sigma \equiv X_j \leq 3$  is satisfied by  $X_j = 3, 2, \dots$ ). As a result, we have a multitude of atomic interventions to consider (e.g. for  $\sigma \equiv X_j \leq 3$ , we have  $P(Y \mid do(X_j := 3)), P(Y \mid do(X_j := 2)), \dots$ ). 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**  $Q_\sigma$  [15, 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  $Q_\sigma$  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  $\sigma$ . Our goal is to identify rules  $\sigma$  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  $\sigma$  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  $X \rightarrow Z \leftarrow Y$ . A spurious path is **blocked** by a set of nodes  $\mathbf{Z}$ , if the path contains a collider that is not in  $\mathbf{Z}$ , or a non-collider on the path is in  $\mathbf{Z}$  [15, Def. 1.2.3]. A set of nodes  $\mathbf{Z}$  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  $\mathbf{Z}$  [15, Def. 3.3.1]. For  $\mathbf{X}$  and  $Y$ , if a set  $\mathbf{Z}$  satisfies the back-door criterion, then observational and post-intervention probabilities are equal within each  $\mathbf{z}$  stratum of  $\mathbf{Z}$ :

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

and averaging the observational probabilities over  $\mathbf{Z}$  gives  $p(y \mid do(\mathbf{X} := \mathbf{x}))$  [15, Thm. 3.3.2].

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

**Definition 1** (ADMISSIBLE INPUT TO CAUSAL RULE DISCOVERY) *The causal system  $(\mathbf{X}, Y, \mathbf{Z})$  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  $\mathbf{X}$ ,*
- (b) *no outgoing edges from any  $X$  in  $\mathbf{X}$  to any  $Z$  in  $\mathbf{Z}$ ,*
- (c) *no edges between actionable variables  $\mathbf{X}$ , and*
- (d) *no edges between any unobserved  $\mathbf{U}$  and  $X$  in  $\mathbf{X}$ .*

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

**PROPOSITION 3.1.** *Let  $(\mathbf{X}, Y, \mathbf{Z})$  be an admissible input to causal rule discovery. Then the control variables  $\mathbf{Z}$  block all*

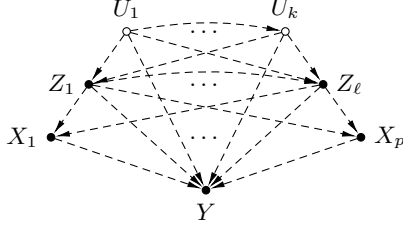


Figure 1: A skeleton causal graph of an admissible input to causal rule discovery (see Def. 1). A dashed edge from a node  $u$  to  $v$  indicates that  $u$  potentially affects  $v$ .

spurious paths between any subset of actionable variables  $\mathbf{X} \subseteq \mathbf{X}$  and  $Y$ .

*Proof.* We postpone the proof to the Appendix.  $\square$

Using admissible control variables  $\mathbf{Z}$ , we can then compute  $p(y | do(Q_\sigma))$  for any rule  $\sigma$  from the rule language  $\mathcal{L}$  as

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

where the first expression is obtained by applying the back-door adjustment formula [15, 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  $Q_\sigma$  which in some sense we treated as an oracle so far. The following theorem shows that, with a specific choice of  $Q_\sigma$ , 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 [11]).

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

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

*Proof.* We postpone the proof to the Appendix.  $\square$

That is, for admissible input  $(\mathbf{X}, Y, \mathbf{Z})$ , the expression above on the r.h.s. gives us the causal effect of any rule  $\sigma$  from the rule language  $\mathcal{L}$  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* [17], which is a fairly standard assumption in causal inference literature.

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 [20], a rather pragmatic 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  $\mathbf{X}$ ; 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  $\mathbf{Z}$ —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  $\mathbf{Z}$  from the population (or in practice, the sample size is large enough to give relatively accurate estimates of the marginal distribution of  $\mathbf{Z}$ ). 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:

$$\begin{aligned} \hat{e}(\sigma) &= \mathbb{E}[\hat{p}(y | \sigma, \mathbf{z}) - \hat{p}(y | \bar{\sigma}, \mathbf{z})] \\ &= \sum_{\mathbf{z} \in \mathcal{Z}} \left( \hat{p}(y | \sigma, \mathbf{z}) - \hat{p}(y | \bar{\sigma}, \mathbf{z}) \right) \hat{p}(\mathbf{z}) \\ &= \sum_{\mathbf{z} \in \mathcal{Z}} (\hat{p}_{\sigma, \mathbf{z}} - \hat{p}_{\bar{\sigma}, \mathbf{z}}) \hat{p}(\mathbf{z}), \end{aligned}$$

where  $\hat{p}_{\sigma, \mathbf{z}} = \hat{P}(y | \sigma, \mathbf{z})$ , and  $\hat{p}_{\bar{\sigma}, \mathbf{z}} = \hat{P}(y | \bar{\sigma}, \mathbf{z})$ . 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,  $\hat{e}(\sigma)$  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,  $\sigma$  or  $\bar{\sigma}$ . To illustrate this, in Fig. 3 (left), we show the estimated distribution for 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, and we end up with random results of either extremely specific or extremely 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  $\mathbf{Z} = \mathbf{z}$  from the population, the number of instances with *successful* outcome  $y$  is a binomial random variable with

the success probability  $p(y|\sigma, \mathbf{z})$ . In a stratum  $\mathbf{z}$  of  $\mathbf{Z}$ , let  $n_{\sigma, \mathbf{z}}$  and  $n_{\bar{\sigma}, \mathbf{z}}$  be the number of instances that satisfy  $\sigma$  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  $\beta$  is the  $1 - \alpha/2$  quantile of a standard normal distribution for an error rate  $\alpha$ , 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 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  $\beta$ , 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  $\beta$ , we have

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

As we deal with empirical probabilities, we can express  $\tau(\mathbf{z})$  in terms of 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  $\tau(\mathbf{z})$  in terms of the cell counts in the contingency table 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  $\mathbf{Z}$ . Hence a stratum of size  $n$  increases to  $n + 4$ , and the total effective sample size increases from  $N$  to  $N + 4|\mathcal{Z}|$ . After applying Laplace correction, we have  $\hat{P}(\mathbf{z}) = (n + 4)/(N + 4|\mathcal{Z}|)$ , and  $\tau(\mathbf{z})$  is given by

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

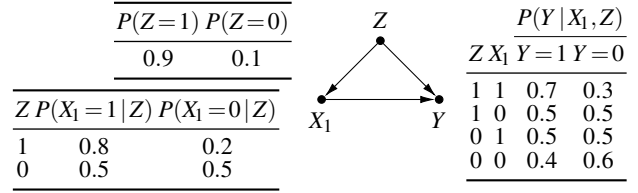


Figure 2: A causal graph of three variables  $X_1$ ,  $Y$  and  $Z$ , and alongside their corresponding conditional probabilities as used in the running example.

After introducing the bias and applying the Laplace correction to the plug-in estimator, we obtain the **reliable** estimator of the causal effect as

$$(3.2) \quad \hat{r}(\sigma) = \sum_{\mathbf{z} \in \mathcal{Z}} \tau(\mathbf{z}) \hat{p}(\mathbf{z}).$$

Note that  $\hat{r}(\sigma)$  is still a **consistent** estimator of the causal effect. In contrast to the plug-in estimator, the reliable estimator is much better at generalisation as it avoids overfitting.

To demonstrate this, let us 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  $X_2, X_3, \dots, X_6$  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 do so, we draw stratified samples of increasing sizes from the population, and report  $\hat{e}(\sigma)$  and  $\hat{r}(\sigma)$  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 indeed able to avoid overfitting, and hence, better at generalisation. Let  $\sigma^*$  denote the top-1 rule in the population, i.e.  $\sigma^* = \arg\max_{\sigma \in \mathcal{L}} e(\sigma)$ . Let  $\varphi^*$  denote the top-1 rule using the plug-in estimator, i.e.  $\varphi^* = \arg\max_{\sigma \in \mathcal{L}} \hat{e}(\sigma)$ , and  $\rho^*$  denote the top-1 rule using the reliable estimator, i.e.  $\rho^* = \arg\max_{\sigma \in \mathcal{L}} \hat{r}(\sigma)$ . 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  $e(\sigma^*)$ , which is in agreement with both theory and intuition.

#### 4 Discovering Rules

Now that we have a reliable and consistent estimator of the causal effect, we turn to discovering rules that maximize this

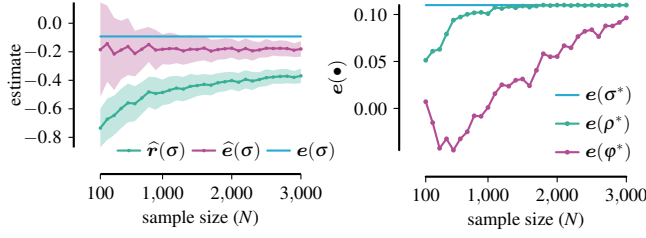


Figure 3: From the population generated using the causal graph of Fig. 2 together with 5 additional independent actionable variables  $X_2, \dots, X_6$ , we show (left) variance of the plug-in and reliable estimator of the causal effect for a specific rule that contains variables that are independent of the target, and (right) generalisation error of the effect estimators.

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  $\mathcal{F}_k \subseteq \mathcal{L}$ ,  $|\mathcal{F}_k| = k$ , such that for all  $\sigma \in \mathcal{F}_k$  and  $\varphi \in \mathcal{L} \setminus \mathcal{F}_k$ ,  $\hat{r}(\sigma) \geq \hat{r}(\varphi)$ .*

Given the hardness of empirical effect maximisation problems [22], 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 paradigm.

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

- A **refinement operator**  $b : \mathcal{L} \rightarrow \mathcal{P}(\mathcal{L})$  that is monotone, i.e. for  $\sigma, \varphi \in \mathcal{L}$  with  $\varphi = b(\sigma)$  it holds that  $\text{ext}(\varphi) \subseteq \text{ext}(\sigma)$ , and that non-redundantly generates the search space  $\mathcal{L}$ . That is, for every rule  $\sigma \in \mathcal{L}$ , there is a unique sequence of rules  $\sigma_0, \sigma_1, \dots, \sigma_\ell = \sigma$  with  $\sigma_i = b(\sigma_{i-1})$ .
- An **optimistic estimator**  $\tilde{f} : \Omega \rightarrow \mathbb{R}$  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  $\varphi \in \mathcal{L}$  with  $\text{ext}(\varphi) \subseteq \text{ext}(\sigma)$ .

A branch-and-bound algorithm simply enumerates the search space  $\mathcal{L}$  starting from the root  $\phi$  using the refinement operator  $b$  (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  $f$ . 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] given by

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

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  $\gamma$ -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  $\varphi \in \mathcal{L}$  with  $\text{ext}(\varphi) \subseteq \text{ext}(\sigma)$ . Lower  $\gamma$  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:

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

where  $\pi$  is the set of propositions and  $\pi^{(\sigma)}$  is the subset of  $\pi$  used in  $\sigma$ . 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  $\hat{r}$ .

**4.2 Efficient optimistic estimator** If we look at the definition of  $\hat{r}(\sigma)$  in Eq. (3.2), we see that, regardless of  $\sigma$ ,  $\hat{p}(\mathbf{z})$  remains the same for a  $\mathbf{z}$  stratum. Thus, we can obtain an optimistic estimator of  $\hat{r}(\sigma)$  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  $\hat{r}(\sigma)$  is given by

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

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  $\sigma$ . The refinement of  $\sigma$ ,  $\sigma' = b(\sigma)$ , results in a contingency 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  $\sigma' = b(\sigma)$ , we have the following relations:  $a' \leq a$  and  $b' \leq b$ .

This implies that the subsets of the extensions of  $\sigma$  will have contingency table counts  $a'$  in the range  $\{0, 1, \dots, a\}$ ,

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

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

and  $b'$  in the range  $\{0, 1, \dots, b\}$ . Let  $\mathcal{C} = \{0, 1, \dots, a\} \times \{0, 1, \dots, b\}$ . Then the optimistic estimator of  $\tau(\mathbf{z})$  can be defined in terms of  $\mathcal{C}$  as

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

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

**PROPOSITION 4.1.** *Let  $\mathcal{C} = \{0, 1, \dots, a\} \times \{0, 1, \dots, b\}$  be the set of all possible configurations of  $(a', b')$  in Tab. 1 (right) that can result from refinements of a rule  $\sigma$  from the contingency table of Tab. 1 (left). Then the **tight optimistic 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 postpone the proof to the Appendix.  $\square$

## 5 Experiments

We implemented the branch-and-bound search with priority-queue in the free and open source realKD<sup>2</sup> Java library, and provide the source code online.<sup>3</sup> 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  $e(\sigma)$ . 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

$e(\sigma^*)$ , where  $\sigma^*$  is the maximiser in the population, i.e.  $\sigma^* = \operatorname{argmax}_{\sigma \in \mathcal{L}} e(\sigma)$ . Using the reliable estimator  $\hat{r}$ , for instance, we get the reliable effect maximiser  $\rho^*$  in the sample, i.e.  $\rho^* = \operatorname{argmax}_{\sigma \in \mathcal{L}} \hat{r}(\sigma)$ . As such,  $e(\rho^*)$  is our estimate of the estimand  $e(\sigma^*)$ , using  $\hat{r}$ . Note that  $e(\rho^*)$  is a function of the sample, and thus a random variable. Therefore the MSE of  $e(\rho^*)$  is given by

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

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

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.  $\hat{e}(\sigma)$  with  $\mathbf{Z} := \emptyset$ . In our case, the weighted relative accuracy of the event  $\sigma$  for an outcome  $y$  at the population level is given by

$$w(\sigma) = p(\sigma) \left( p(y | \sigma) - p(y) \right).$$

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

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

<sup>2</sup><https://bitbucket.org/realKD/>

<sup>3</sup><https://www.dropbox.com/s/fr1v8os29gobcby/rce.zip?dl=0>



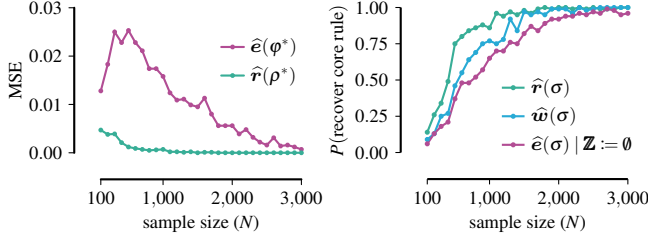


Figure 4: (left) Mean squared error (MSE) of the plug-in estimator  $\hat{e}$  and the reliable estimator  $\hat{r}$  of the population optimal causal effect  $e(\sigma^*)$ . (right) The probability of recovering the core rule from 100 samples, for the Laplace-corrected plug-in estimator of causal effect,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

$\sigma \equiv X_1 = 1$  is *relevant*.<sup>4</sup> 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 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.<sup>5</sup> 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

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

Top-3 rules ( $\sigma$ )	cv $g(\sigma)$	$\hat{r}(\sigma)$
$\text{class} \leq 2 \wedge \text{sex} = \text{female}$	0.1907	0.576
$\text{class} \leq 2 \wedge \text{sex} = \text{female} \wedge \text{par-ch} \leq 2$	0.1885	0.573
$\text{class} \leq 2 \wedge \text{sex} = \text{female} \wedge \text{sib-sp} \leq 2$	0.1874	0.572

change their title, but not the other way around. Thus it is reasonable to assume that *sex* causes *title*. As putting them together in  $\mathbf{X}$  would violate criterion (c) of Def. 1, we only keep one of them, i.e. *sex*, in  $\mathbf{X}$ . Similarly we can argue that fare causes passenger class. Therefore we only keep *class* in  $\mathbf{X}$ . Overall, none of the variables seem to confound (co-cause)  $Y$  and other variables. Altogether, we have

$$\mathbb{Z} := \emptyset$$

$$Y := \text{survived}$$

$$\mathbf{X} := \{\text{class}, \text{pname}, \text{sex}, \text{age}, \text{sib\_sip}, \text{par\_ch}, \text{embarked}\}$$

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.  $\text{cv}g(\sigma) = |\text{ext}(\sigma)|/N$ .

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  $\hat{r}(\sigma_1) = 0.576$ . 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. The fact that all those rules came out 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 the system undergoes changes through interventions [15]. 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,

<sup>4</sup>The complementary rule  $\bar{\sigma} \equiv X_1 = 0$  has a *negative* effect.

<sup>5</sup><https://www.kaggle.com/c/titanic>



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 the literature, criterion (c) is new and specific to rule discovery. It requires that there are no edges between actionable variables. Over a large group of actionable variables, this can be a strong assumption. The naive way to remove this assumption would be to include rest of the actionable variables  $\mathbf{X} \setminus \mathbf{X}$  in the set of control variables to block any spurious path between actionable variables  $\mathbf{X}$  in a rule  $\sigma$  and the target  $Y$  via  $\mathbf{X} \setminus \mathbf{X}$ . By doing so, however, we may not only violate other criteria, but the search also gets complicated.

On the statistical side, 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 with small generalisation error. This use of confidence intervals is reminiscent of, e.g., 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.

## 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] 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.
- [6] J. H. Friedman and N. I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143, 1999.
- [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] K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.
- [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] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.
- [16] 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.
- [17] R. Scheines. An introduction to causal inference. In *Causality in Crisis? University of Notre Dame*, pages 185–200. Press, 1997.
- [18] P. Shamsinejadbabaki, M. Saraee, and H. Blockeel. Causality-based cost-effective action mining. *Intelligent Data Analysis*, 17(6):1075–1091, 2013.
- [19] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. *DAMI*, 4(2):163–192, 2000.
- [20] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.
- [21] F. Wang and C. Rudin. Causal falling rule lists. In *FATML Workshop*, 2017.
- [22] L. Wang, H. Zhao, G. Dong, and J. Li. On the complexity of finding emerging patterns. *Theoretical Computer Science*, 335(1):15–27, 2005.
- [23] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.

### Proof of Proposition 3.1

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

*Proof.* We prove this proposition graphically. Recall the criteria for the input  $(\mathbf{X}, Y, \mathbf{Z})$  to be admissible.

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

Observe that any spurious path between any  $X$  in  $\mathbf{X}$  and  $Y$  can be formed through one of the following ways:

- from  $Y$  to  $\mathbf{X}$  directly,
- via control variables  $\mathbf{Z}$ ,
- via other actionable variables  $\mathbf{X} \setminus \mathbf{X}$ ,
- via latent variables  $U$ .

Criterion (a) rules out trivial spurious paths from  $Y$  to  $\mathbf{X} \supseteq \mathbf{X}$  that cannot be blocked by any  $\mathbf{Z}$ . Criterion (b) ensures that any spurious path unblocked by one control variable is blocked by another. Criterion (c) ensures that there are no spurious paths between any subset of actionable variables  $\mathbf{X}$  in the rule  $\sigma$  and  $Y$  via other actionable variables  $\mathbf{X} \setminus \mathbf{X}$ . To see this, suppose that we have two actionable variables  $X_1$  and  $X_2$ , and a rule  $\sigma \equiv X_1 = 1$ . If the causal graph contains the path  $X_1 \leftarrow X_2 \rightarrow Y$ , we will have a biased estimate of the causal effect. Criterion (d) is really just a form of standard causal sufficiency [17]. As there are no edges between any latent variable  $U$  in  $U$  and any  $X$  in  $\mathbf{X}$ , by conditioning on  $\mathbf{Z}$ , we block any spurious path between any  $\mathbf{X}$  and  $Y$  via  $U$ . Thus, if the input is admissible, the control variables  $\mathbf{Z}$  block all spurious paths between any subset  $\mathbf{X}$  of  $\mathbf{X}$  and  $Y$ .  $\square$

### Proof of Theorem 3.1

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

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

*Proof.* Recall that  $p(y \mid \text{do}(Q_\sigma))$  is given by

$$p(y \mid \text{do}(Q_\sigma)) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{z}) \sum_{\sigma(\mathbf{x}) = \top} p(y \mid \mathbf{x}, \mathbf{z}) Q_\sigma(\text{do}(\mathbf{x}))$$

Using the stochastic policy,  $p(y \mid \text{do}(Q_\sigma))$  reduces to

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

for  $\sigma(\mathbf{x}) = \top$ , it holds that  $p(\mathbf{x}, \sigma \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})$ ; thus

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

since  $\sum_{\sigma(\mathbf{x}) = \top} P(y, \mathbf{x} \mid \mathbf{z}) = P(y, \sigma \mid \mathbf{z})$ , this results in

$$\begin{aligned} p(y \mid \text{do}(Q_\sigma)) &= \sum_{\mathbf{z} \in \mathcal{Z}} \frac{p(\mathbf{z})}{p(\sigma \mid \mathbf{z})} p(y, \sigma \mid \mathbf{z}) \\ &= \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{z}) p(y \mid \sigma, \mathbf{z}) \\ &= \mathbb{E}[p(y \mid \sigma, \mathbf{Z})]. \end{aligned}$$

Substituting the above in the definition of  $e(\sigma)$ , we get

$$\begin{aligned} e(\sigma) &= p(y \mid \text{do}(Q_\sigma)) - p(y \mid \text{do}(Q_{\bar{\sigma}})) \\ &= \mathbb{E}[p(y \mid \sigma, \mathbf{Z})] - \mathbb{E}[p(y \mid \bar{\sigma}, \mathbf{Z})]. \end{aligned}$$

$\square$

### Proof of Proposition 4.1

PROPOSITION 4.1. *Let  $\mathcal{C} = \{0, 1, \dots, a\} \times \{0, 1, \dots, b\}$  be the set of all possible configurations of  $(a', b')$  in Tab. 1 (right) that can result from refinements of a rule  $\sigma$  from the contingency table of Tab. 1 (left). Then the **tight optimistic estimator** of  $\tau(\mathbf{z})$  is given by*

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

*Proof.* The expression for  $\tau(a', b')$  from the contingency table in Tab. 1 (right) is given by

$$\begin{aligned} \tau(a', b') &= \frac{a' + 1}{a' + b' + 2} - \frac{n_1 - a' + 1}{n - a' - b' + 2} - \frac{\beta}{2\sqrt{a' + b' + 2}} - \\ &\quad \frac{\beta}{2\sqrt{n - a' - b' + 2}}. \end{aligned}$$

Combining the first and the third term above, we get

$$\begin{aligned} \lambda_z(a', b') &= \frac{2a' + 2 - \beta\sqrt{a' + b' + 2}}{2(a' + b' + 2)} - \frac{n_1 - a' + 1}{n - a' - b' + 2} - \\ &\quad \frac{\beta}{2\sqrt{n - a' - b' + 2}}. \end{aligned}$$

Note that if we fix the value of  $a'$ , then the value of  $b'$  that maximises  $\tau(a', b')$  has to maximise the first term above, but minimise the other two terms. Observe that  $b' = 0$ , out of  $b' \in \{0, 1, \dots, b\}$ , does both simultaneously. Thus we have the following relation:  $\tau(a', 0) > \tau(a', b')$  for all  $b' > 0$ .

The tight optimistic estimator of  $\tau(\mathbf{z})$  is then the maximum value over all possible configurations  $\mathcal{C}$ , i.e.

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

□

### Effect of $\beta$

The reliable estimator  $\hat{r}$  of causal effect has a user-defined parameter  $\beta$  that represents our confidence in the point estimate. It is easy to see that a 0% confidence level corresponds to the plug-in estimator  $\hat{e}$ , as then we would have a z-score of  $\beta = 0$ . As we increase the confidence level,  $\hat{r}(\sigma)$  gets more conservative. How does this affect the performance of the reliable estimator?

To answer this question, we sample observations from the population in our previous evaluation (Sec. 5.1). In Fig. 5, we plot the mean squared error (MSE) of  $e(p^*)$  that uses the reliable estimator  $\hat{r}$  at various confidence levels: 50%, 60%, 70%, 80%, 90%, 99%. We observe that  $\text{MSE}(e(p^*))$  decreases with increasing sample sizes at all confidence levels. The MSE is much better at higher confidence levels particularly in the beginning when sample sizes are small. These results suggest that higher confidence levels lead to more reliable rules, in terms of their closeness to the maximum true effect. It might be tempting then to go for a 100% confidence level. However, 100% confidence is only achievable with an infinite sample size. Moreover, at a 100% confidence level, we have  $\beta = \infty$ , and in turn  $\hat{r}(\sigma) = \infty$ . This suggests that we can calibrate the optimal  $\beta$  for a given sample size somewhere below the 100% confidence level.

### Efficiency of the Branch-and-Bound Search

Next we assess efficiency of the branch-and-bound search. To this end, first we search for top-1 rule in all the standard classification datasets from the KEEL repository.<sup>6</sup> The diversity of these datasets in terms of their sample size and number of actionable variables provides a reasonable picture on the efficiency of the proposed search algorithm in a real-world scenario. For each dataset, we select the classification target as the target, and randomly select one of the attributes

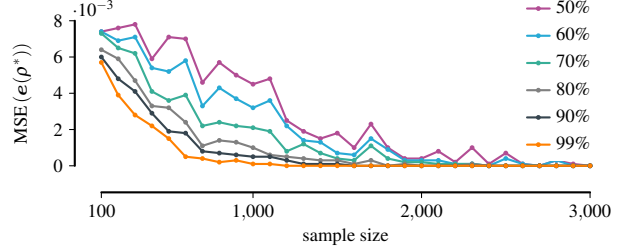


Figure 5: Mean squared error of the reliable estimator  $\hat{r}$  of the population optimal causal effect  $e(\sigma^*)$  at various  $\beta$ .

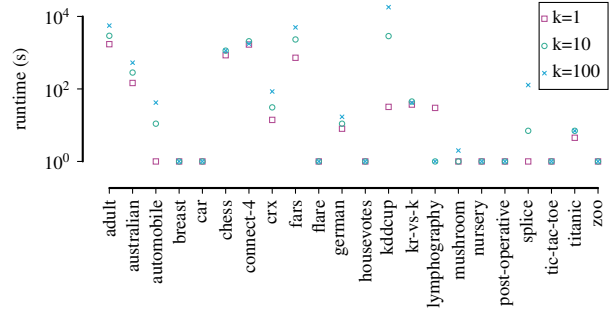


Figure 6: Runtime, in seconds, of the top- $k$  branch-and-bound search algorithm for different values of  $k$  for the real-world datasets in Tab. 3.

as the control variable. As outcome  $y$ , we select one of the outcomes of the target  $Y$ . We discretise a real-valued actionable variable into maximum 8 equi-frequent bins.

In the Tab. 3, we provide the summary of the datasets along with the efficiency results. For each dataset, we report the target ( $Y$ ), the set of control variables ( $\mathbf{Z}$ ),<sup>7</sup> the sample size ( $N$ ), the number of actionable variables ( $|\mathbf{X}|$ ), the approximation factor ( $\gamma$ ) such that the branch-and-bound search finishes within an hour, the runtime in seconds, and the number of nodes expanded during the search. We observe that the branch-and-bound search with the tight optimistic estimator retrieves the *optimal* top-1 result within seconds for most datasets, taking up to roughly an hour for a few datasets.

In practice, it may be of interest to look for multiple rules for various reasons. Therefore, next we evaluate the scalability of the branch-and-bound algorithm with respect to  $k$  in top- $k$  search. In Fig. 6, we show the runtime of the branch-and-bound algorithm for  $k = 1, 10, 100$  in all the standard

<sup>7</sup>Although we use only one variable in  $\mathbf{Z}$ , it does not really work to our advantage here. Assuming that observations are roughly uniformly distributed across values of  $\mathbf{Z}$ , we only have  $N/|\mathcal{Z}|$  observations within each group of  $\mathbf{Z}$ . Thus group sizes gets smaller as we add more variables in  $\mathbf{Z}$ . This in turn increases the bias term in the reliable estimator, thereby decreasing the number of rules with *positive* reliable causal effect. That fact combined with the tightness of the optimistic estimator eventually speeds up the branch-and-bound search as we add more variables to  $\mathbf{Z}$ .

<sup>6</sup><https://sci2s.ugr.es/keel/datasets.php>

Table 3: Summary of the datasets used for the empirical evaluation along with the efficiency results. For each dataset, we report the chosen target variable ( $Y$ ), the chosen control variables ( $\mathbf{Z}$ ), the sample size ( $N$ ), the number of actionable variables ( $|\mathbf{X}|$ ), the approximation factor ( $\gamma$ ), the runtime in seconds, and the number of nodes expanded during search.

Dataset	Target ( $Y$ )	Control ( $\mathbf{Z}$ )	$N$	$ \mathbf{X} $	$\gamma$	time (s)	#nodes
adult	class	sex	48,842	13	0.8	1,717	258,575
australian	class	a4	690	13	1.0	146	952,175
automobile	output	engine-type	205	24	1.0	1	15,167
breast	class	age	286	8	1.0	78	420
car	acceptability	safety	1,728	5	1.0	0.02	33
chess	class	bkblk	3,196	35	1.0	851	1,613,398
connect-4	class	a1	67,557	61	0.3	1,679	140,707
crx	class	a1	690	14	1.0	14	101,621
fars	injury-severity	case-state	100,968	28	0.8	724	22,328
flare	class	prev24hour	1,066	10	1.0	0.014	32
german	customer	statusAndSex	1,000	19	1.0	8	43,007
housevotes	class	el-salvador-aid	435	15	1.0	0.007	57
kddcup	class	atr-6	494,020	40	0.99	37	219
kr-vs-k	game	white-king-col	28,056	5	1.0	30	7,304
lymphography	classes	changes-in-lym	148	17	1.0	0.14	1,666
mushroom	class	gill-size	8,124	21	1.0	0.307	215
nursery	class	social	12,690	7	1.0	0.66	279
post-operative	decision	l-core	90	7	1.0	0.016	258
splice	class	pos1	3,190	59	1.0	1.03	1,855
tic-tac-toe	class	toleft	958	8	1.0	0.11	488
titanic	survived	sex	891	9	1.0	4.5	26,700
zoo	type	aquatic	101	15	1.0	0.009	96

classification datasets from the KEEL repository, using same approximation factors. For most datasets, we observe that the branch-and-bound search finishes within seconds at all values of  $k$ . For a few datasets, even though runtime of the algorithm increases with increasing value of  $k$ , it finishes within a couple of hours.