Configurational causal modeling and logic regression

Michael Baumgartner and Christoph Falk*

Abstract

Configurational comparative methods (CCMs) and logic regression methods (LRMs) are two families of methods that employ very different techniques to analyze data generated by causal structures featuring conjunctural causation and equifinality. Aiming for the same by different means carries a substantive synergy potential, which, however, remains untapped so far because representatives of the two frameworks know little of each other. The purpose of this article is to change that. We first level the field for readers from both backgrounds by providing cursory introductions to the basic ideas behind CCMs and LRMs. Then, we carve out the strengths and weaknesses of both methods by benchmarking their performance under a variety of different data scenarios. It turns out that CCMs and LRMs have complementary strengths and weaknesses. This creates various promising avenues for cross-validation.

Keywords: INUS causation, conjunctural causation, component causation, equifinality, Coincidence Analysis, multi-method research, cross-validation

Introduction

Many disciplines investigate causal structures with one or both of the following features: (i) causes are arranged in complex bundles that only become operative when all of their components are properly co-instantiated, each of which in isolation is ineffective or leads to different outcomes, and (ii) outcomes can be brought about along alternative causal routes such that, when one route is suppressed, the outcome may still be produced via another

*University of Bergen, Norway, michael.baumgartner@uib.no, christoph.falk@uib.no. The authors gratefully acknowledges the support by the Toppforsk program of the University of Bergen, co-financed by the Trond Mohn Foundation (grant nr. 811886).

one. For example, of a given set of implementation strategies available to hospitals some strategies yield a desired outcome (e.g. high vaccination uptake or shorter hospitalization times) in combination with certain other strategies, whereas in other combinations the same strategies may have opposite effects; and the same outcome can be obtained via different bundles of strategies (e.g. Yakovchenko et al. 2020). Or, a variation in a phenotype only occurs if many single-nucleotide polymorphisms interact, and various such interactions can independently induce the same phenotype (e.g. Culverhouse et al. 2002). Different labels are used for features (i) and (ii): "component causation", "conjunctural causation", "alternative causation", "equifinality", etc. For uniformity's sake, we will subsequently refer to (i) as *conjunctivity* and to (ii) as *disjunctivity* of causation, reflecting the fact that causes form conjunctions and disjunctions, that is, Boolean AND- and OR-connections.

Causal structures featuring conjunctivity and disjunctivity pose severe challenges for methods of causal data analysis. Because many theories of causation entail that it is necessary (though not sufficient) for X to be a cause of Y that there be some kind of dependence (e.g. probabilistic or counterfactual) between X and Y, standard methods—most notably Bayesian network methods (Spirtes et al. 2000)—infer that X is *not* a cause of Y if X and Y are *not* pairwise dependent (i.e. correlated).[1] However, structures displaying conjunctivity and disjunctivity often do not exhibit such pairwise dependencies. As an illustration, consider the interplay between a person's skills to perform an activity, the challenges posed by that activity, and the actor's autotelic experience of complete involvement with the activity called *flow* (Csikszentmihalyi 1975). A simplified (binary) model of this interplay involves the factors S, with values 0/1 representing low/high skills, C, with 0/1 standing for low/high challenges, and F, with 0/1 representing the absence/presence of flow. According to Csikszentmihalyi's (1975, ch. 4) flow theory, flow is triggered by skills and challenges being either both high or both low, meaning that F=1 has two alternative causes S=1 & C=1 and S=0 & C=0. If the flow theory is true, ideal (i.e. unbiased, non-confounded, noise-free) data on this structure feature the four configurations c_1 to c_4 in Table 1a, and no others. As can

[1]Methods of causal inference must be distinguished from methods of causal reasoning (e.g. Peters et al. 2017, 5-6). The former are methods discovering or learning causal model from data, and the latter are methods testing given models by, for example, inferring predictions from them. This paper is only concerned with the former type of methods.

#	S	C	F
c_1	1	1	1
c_2	0	0	1
c_3	1	0	0
c_4	0	1	0

(a)

	S	C	F
S	1.00	0.00	0.00
C	0.00	1.00	0.00
F	0.00	0.00	1.00

(b)

Table 1: Table (a) contains ideal data with F being the outcome, table (b) the corresponding correlation matrix.

easily be seen from the corresponding correlation matrix in Table 1b, there are no pairwise dependencies. In consequence, Bayesian network methods and standard regression methods will struggle to find the flow model, even when processing ideal data on it. Although there exist various protocols for tracing interaction effects involving two or three exogenous factors, these interaction calculations face tight computational complexity restrictions when more exogenous factors are involved and quickly run into multicollinearity issues (Brambor et al. 2006). Standard methods of causal data analysis are simply not designed to group causes conjunctively and disjunctively—rather, their main aim is to quantify effect sizes.

Discovering causal structures exhibiting conjunctivity and disjunctivity calls for methods that track causation as defined by a theory not treating pairwise dependencies as necessary for causation and that embed individual factors in complex Boolean AND- and OR-functions, fitting those functions as a whole to the data. The problem, however, is that the space of possible Boolean functions over even a handful of factors is vast. For n binary factors there exist 2^{2^n} possible Boolean functions, and if we also include factors with more than two values that number explodes beyond controllability. That means methods capable of discovering causal structures with conjunctivity and disjunctivity must, in addition to relying on a suitable theory of causation, find ways to efficiently navigate in that vast space of possibilities.

The methods explicitly built for this purpose are the so-called *configurational comparative methods* (CCMs; Ragin 1987; Rihoux and Ragin 2009; Baumgartner and Ambühl 2020). They rely on tools from Boolean algebra, take data on binary, multi-value or continuous (fuzzy-set) factors as input, and infer causal structures as defined by the so-called *INUS* or *MINUS theory* (Mackie 1974; Baumgartner and Falk 2019),[2] which spells out causation in

[2]Mackie (1974, 62) introduced the label "INUS" as an acronym standing for *I*nsufficient but *N*on-redundant parts of *U*nnecessary but *S*ufficient conditions. As there are more elegant ways to capture the idea expressed by that expansion, "INUS" is often used as a mere name for a theoretical framework today—void of its original

terms of redundancy-free Boolean dependency structures. One of the distinctive features of this account is that it does not imply that causes and their outcomes are pairwise dependent.

CCMs, whose main base is in the social sciences, are not the only methods designed for the analysis of structures with conjunctivity and disjunctivity. In biostatistics, the problems posed by such structures have led to the development of *logic regression methods* (LRMs; Ruczinski et al. 2003; Schwender and Ickstadt 2007). LRMs are primarily used to model higher-order interactions in genetic association studies, to which end they express binary outcomes as Boolean functions, which they fit to data by embedding them in a generalized regression framework. Contrary to CCMs, the primary target of LRMs is not causation but prediction. Accordingly, LRMs are not expressly linked to the (M)INUS theory and their models have a non-standarized syntax that may contain redundant elements prohibiting a causal interpretation. Still, as this paper will show, the redundancies in LRM models can be eliminated and their syntax transformed into the form of (M)INUS models by suitable post-processing. That is, although employing very different techniques, LRMs can be tweaked to search for the same types of structures as CCMs. Targeting the same by different means creates a substantial potential for synergies. It is the main goal of this paper to bring that potential to light, as it remains entirely untapped so far because representatives of the two frameworks know little of each other.[3]

To level the field for readers from different backgrounds, the first part of the paper provides cursory introductions to the (M)INUS theory of causation and to the basic ideas behind CCMs and LRMs. In the second part, we then carve out the strengths and weaknesses of both methods by benchmarking their performance under a variety of different data scenarios simulated from causal structures analyzable by both methods, that is, structures over binary factors featuring one outcome and up to 9 interacting causes. It turns out that CCMs and LRMs have complementary strengths and weaknesses, which yields a considerable potential for cross-validation.

meaning. Accordingly, "MINUS" is a name, without an expansion, locating the corresponding theory in the INUS tradition.

[3]Of the 1381 CCM articles listed in the bibliography data base of the COMPASSS network (compasss.org), which is dedicated to CCM research, only one article mentions logic regression (in passing), *viz.* Clarke (2020). Google Scholar, which has 11500 records for CCMs and 4070 for LRMs (in March 2021), additionally finds Rohwer (2011), who is concerned with CCMs and refers to logic regression in a footnote.

(M)INUS causation

Even though causation is everywhere in human interaction with the world, it is not pre-theoretically clear what causation is. Is it an objective feature of our world or is it something we, as observers, project onto the world? Does it govern what occurs around us or is it a concept that merely facilitates theorizing about those occurrences? Is it a matter of the instantiation of regularities or laws, or of counterfactual dependence, or of probability raising, or of manipulability, mechanisms, or powers? Theories of causation answer these questions by providing explicit definitions of causation. But as there are good arguments for conflicting answers to these questions, there exist many conflicting theories. They resist being meaningfully pitted against each other, ultimately because they are embedded in irreconcilable background metaphysics.[4] As the purpose of this paper is not to contribute to the theoretical literature on causation but to compare two methods discovering structures featuring conjunctivity and disjunctivity, we can confine ourselves to reviewing the core tenets of the theory custom-built to account for these features, the (M)INUS theory, without thereby claiming to be presenting the only or ultimate truth about causation.

The (M)INUS theory belongs to the family of so-called *regularity theories*,[5] according to which causal relations are nothing over and above specific forms of regular or lawlike behavior patterns. The (M)INUS theory stipulates that *general causation*, that is, causal relations between types of events or properties, as in "High skills combined with high challenges cause flow", are conceptually prior to relations of *singular causation* among token events, as in "Peter's high skills combined with the high challenges of his tasks cause Peter's flow on day x". In other words, there is nothing in a sequence of token events that would make it causal; rather, a causal relation between two token events is a matter of them properly instantiating causally related event types. Event types or properties are modeled using factors (or variables or predictors)[6] taking specific values. Hence, the (M)INUS theory provides a

[4] In that light, the position of *causal pluralism*, which accepts the existence of multiple concepts or multiple variants of causation, has seen a rise in popularity in recent years (e.g. Godfrey-Smith 2010).

[5] Through much of the 20$^{\text{th}}$ century, regularity theories, which have roots going back to Hume (1748) and Mill (1843), were widely criticized (cf. e.g. Armstrong 1983; Hausman 1998), but the groundbreaking work of Mackie (1974) has revived that theoretical framework and has led to the development of modern regularity theories that can deal with classical objections (cf. Graßhoff and May 2001; Baumgartner and Falk 2019).

[6] As the methods discussed in this paper have emerged from different disciplines with different terminologies, the corresponding literatures use different terms with identical (or easily translatable) meanings. Through-

definition of what it means for a factor A taking some value α (i.e. $A{=}\alpha$) to be causally relevant for another factor B taking a value β (i.e. $B{=}\beta$), where "causal relevance" designates the relation of "… is a type-level cause of …".

For $A{=}\alpha$ to be causally relevant for $B{=}\beta$, $A{=}\alpha$ must be a difference-maker of $B{=}\beta$, meaning—roughly—that there exists a context in which other causes take constant values and a change from $A{\neq}\alpha$ to $A{=}\alpha$ is associated with a change from $B{\neq}\beta$ to $B{=}\beta$. Factors in (M)INUS structures can either be *crisp-set* (binary), taking two possible values 0 and 1, *fuzzy-set*, taking continuous values from the unit interval $[0, 1]$, or *multi-value*, taking an open (but finite) number of non-negative integers as possible values. For simplicity of exposition, we subsequently focus on crisp-set factors, which allows for conveniently abbreviating the 'Factor=value' notation. As is conventional in Boolean algebra, we will use 'A' as shorthand for $A{=}1$ and 'a' for $A{=}0$. The (M)INUS theory borrows much of the formal machinery from Boolean algebra, in particular, the operations of *negation*, $\neg A$ (expressing 'NOT $A{=}1$'), *conjunction*, $A{*}B$ ('$A{=}1$ AND $B{=}1$'), *disjunction*, $A + B$ ('$A{=}1$ OR $B{=}1$'), *implication*, $A \rightarrow B$ ('IF $A{=}1$, THEN $B{=}1$'), and *equivalence* $A \leftrightarrow B$ ('$A{=}1$ IF, AND ONLY IF, $B{=}1$').[7] In case of crisp-set factors, Boolean operations are given a rendering in classical logic, which we do not reiterate here (see e.g. Lemmon 1965, ch. 1).

Based on the implication operator[8] the notions of *sufficiency* and *necessity* are defined, which are the two core Boolean dependence relations exploited by the (M)INUS theory: $A{*}C{*}E$ is sufficient for B if, and only if (iff), $A{*}C{*}E \rightarrow B$ (i.e. whenever A AND C AND E are true, B is true); $A + C + E$ is necessary for B iff $B \rightarrow A + C + E$ (i.e. whenever B is true, A OR C OR E are true). Many of these relations, however, have nothing to do with causation. To use a standard (oversimplified) example, the sinking of a barometer in combination with high temperatures and blue skies is sufficient for weather changes, but it does not cause the weather; or whenever there is an election, votes are cast or public speeches are made, so casting votes or making public speeches is necessary for an election, but it does not cause

out the paper, we will hence indicate terminological variations in brackets.

[7]Note that "$*$" and "$+$" are used as in Boolean algebra here, which means, in particular, that they do not represent the linear algebraic (arithmetic) operations of multiplication and addition (notational variants of Boolean "$*$" and "$+$" are "$\wedge$" and "$\vee$"). For a standard introduction to Boolean algebra see Bowran (1965).

[8]By "implication" we always mean Boolean implication in this paper, which is also known as *material implication* (or *material conditional*).

it. Still, some Boolean dependencies are in fact due to underlying causal dependencies: long-term exposure to an active virus combined with lacking immunity is both sufficient and causally relevant for infection; striking a match or exposing it to heat or to inflammable chemicals is both necessary and causally relevant for the match to catch fire.

That means in order to define causal relevance in terms of Boolean dependencies, those relations of sufficiency and necessity that are due to underlying causal dependencies must be filtered out. The main reason why most sufficiency and necessity relations do not reflect causation is that they either contain redundancies or are themselves redundant to account for the behavior of the outcome, whereas causal conditions do not feature redundant elements and are themselves indispensable to account for the outcome in at least one context. Accordingly, to filter out the causally interpretable Boolean dependencies, they need to be freed of redundancies. In Mackie's (1974, 62) words, causes are *I*nsufficient but *N*on-redundant parts of *U*nnecessary but *S*ufficient conditions (thus the acronym INUS).

While Mackie's INUS theory only requires that sufficient conditions be freed of redundancies, he himself formulates a problem for that theory, *viz.* the *Manchester Factory Hooters* problem (Mackie 1974, 81-87), which Graßhoff and May (2001) solve by eliminating redundancies also from necessary conditions. Accordingly, modern versions of the INUS theory stipulate that whatever can be removed from sufficient or necessary conditions without affecting their sufficiency and necessity is not a difference-maker and, hence, not a cause. The causally interesting sufficient and necessary conditions are *minimal* in the sense that they do not contain sufficient and necessary proper parts. Minimally sufficient and minimally necessary conditions can be combined in *MINUS-formulas* (Beirlaen et al. 2018): a MINUS-formula of an outcome B is a minimally necessary disjunction of minimally sufficient conditions of B, in disjunctive normal form.[9] The following is a simple example:

$$A{*}e \;+\; C{*}d \;\leftrightarrow\; B \tag{1}$$

(1) being a MINUS-formula of B entails that $A{*}e$ and $C{*}d$, but neither A, e, C, nor d

[9] An expression is in disjunctive normal form iff it is a disjunction of one or more conjunctions of one or more factor values (Bowran 1965, 13). Note moreover that to do justice to the different types of redundancies that Boolean dependency structures may be affected by, the complete definition of the notion of a MINUS-formula is intricate and beyond the scope of this article (for the latest definition and a discussion of its remaining limitations see Baumgartner and Falk 2019).

alone, are sufficient for B and that $A{*}e + C{*}d$, but neither $A{*}e$ nor $C{*}d$ alone, are necessary for B. If this holds, it follows that for each factor value in (1) there exists a *difference-making pair*, meaning a pair of cases (or units of observation) such that a change in that factor value alone accounts for a change in the outcome (Baumgartner and Falk 2019, 9). For example, A being part of the MINUS-formula (1) entails that there are two cases σ_i and σ_j such that e is given and $C{*}d$ is not given in both σ_i and σ_j while A and B are present in σ_i and absent in σ_j. Only if such a difference-making pair $\langle\sigma_i, \sigma_j\rangle$ exists is A indispensable to account for B.

For an adequate definition of causation an additional constraint is needed because not all MINUS-formulas faithfully represent causation. Complete redundancy elimination is relative to the set of analyzed factors **F**, meaning that factor values contained in MINUS-formulas relative to some **F** may fail to be part of a MINUS-formulas relative to supersets of **F** (Baumgartner and Falk 2019). In other words, by adding factors to the analysis, factor values that originally appeared to be non-redundant to account for an outcome can turn out to be redundant after all. Hence, a *permanence* constraint needs to be imposed: only factor values that are permanently non-redundant, meaning that cannot be rendered redundant by expanding factor sets, are causally relevant.

These considerations yield the following definition of causation:

Causal Relevance (MINUS). A is causally relevant to B iff (I) A is part of a MINUS-formula of B relative to a factor set **F** and (II) A remains part of a MINUS-formula of B across all expansions of **F**.

Two features of the (MINUS) definition make it particularly well suited for the analysis of structures affected by conjunctivity and disjunctivity. First, (MINUS) does not require that causes and effects are pairwise dependent. The following is a well-formed MINUS-formula expressing the flow model from the introduction: $S{*}C + s{*}c \leftrightarrow F$. As shown in Table 1, ideal data generated from that model feature no pairwise dependencies. Nonetheless, if, say, high skills are permanently non-redundant for flow in combination with high challenges, they are causally relevant for flow subject to (MINUS), despite being uncorrelated with flow. Second, MINUS-formulas whose elements satisfy the permanence constraint not only identify causally relevant factor values but also place a Boolean ordering over these causes, such

that conjunctivity and disjunctivity can be directly read off their syntax.

Still, discovering causation as defined in (MINUS) faces various challenges. First, as it is possible that data δ produced by a MINUS structure only feature dependencies between complex Boolean functions of exogenous factors and a corresponding outcome, δ cannot be analyzed by searching for cause-effect pairs and then combining them to complex structures. Rather, analyzing δ calls for fitting complex Boolean functions as a whole to δ. But, as we have seen in the introduction, the space of Boolean functions over more than five factors is so vast that it cannot be exhaustively scanned. Hence, algorithmic strategies are needed to purposefully narrow down the search.

Second, condition (MINUS.II) is not comprehensively testable. Once a MINUS-formula of an outcome B containing a factor value A has been inferred from data δ, the question arises whether the non-redundancy of A in accounting for B is an artefact of δ, due, for example, to the uncontrolled variation of confounders, or whether it is genuine and persists when further factors are taken into consideration. But in practice, expanding the set of factors is only feasible within narrow confines. To make up for the impossibility to test (MINUS.II), data δ should be collected in such a way that Boolean dependencies in δ are not induced by an uncontrolled variation of latent causes but by actual causal dependencies among the measured factors. If the dependencies in δ are not artefacts of latent causes, they cannot be neutralized by factor set expansions, meaning they are permanent and, hence, causal. It follows that in order for it to be guaranteed that causal inferences drawn from δ are error-free, δ must meet very high quality standards. In particular, the uncontrolled causal background of δ must be *homogenous*, meaning that latent causes not connected to the outcome on causal paths via the measured exogenous factors (so-called *off-path* causes) take constant values (i.e. do not vary) in the cases recorded in δ (Baumgartner and Ambühl 2020, online appendix).

However, third, real-life data often do not meet very high quality standards. To make this concrete, consider Table 2 featuring a simple small-n data set over the set of factors $\mathbf{F}_1 = \{A, C, D, E, B\}$ simulated from the MINUS structure in expression (1)—to which we will henceforth refer as the *ground truth*. Each row in that table represents a configuration of the factors in $\mathbf{F}_1$, and the column "n.obs" indicates how many cases instantiate a particular configuration. Those data have been simulated such that they feature various imperfections

conf.	A	C	D	E	B	n.obs
σ_1	1	1	1	1	0	6
σ_2	0	1	1	1	0	1
σ_3	1	0	1	1	0	2
σ_4	0	0	1	1	0	2
σ_5	1	1	0	1	1	3
σ_6	1	0	0	1	0	2
σ_7	0	0	0	1	0	1
σ_8	1	1	1	0	1	4
σ_9	0	1	1	0	0	5
σ_{10}	1	0	1	0	1	6
σ_{11}	0	0	1	0	0	2
σ_{12}	1	1	0	0	1	3
σ_{13}	1	0	0	0	1	1
σ_{14}	0	0	0	0	0	3
σ_{15}	1	1	1	0	0	2
σ_{16}	0	0	1	0	1	2

Table 2: Example data with a total of 45 units of observation (cases) instantiating 16 configurations σ_1 to σ_{16} with "n.obs" indicating how many cases instantiate a particular configuration. B is endogenous, the other factors are exogenous.

typical for real-life data. They are *fragmented*, meaning they do not comprise all configurations that can be generated by (1). For example, if the behavior of the factors in $\mathbf{F}_1$ is regulated by (1), we should be able to observe the configuration $a{*}C{*}d{*}e{*}B$, which, however, is not contained in Table 2. Furthermore, the data contain *noise* due to measurement error or confounding, meaning that not all configurations in Table 2 are compatible with (1). In σ_{15}, for instance, $A{*}e$ is combined with b, even though $A{*}e$ is sufficient for B according to (1); or, in σ_{16} B is given without any of its causes in (1)—hence, its occurrence must be due to latent causes.[10] As a result, Table 2 does not feature relations of strict Boolean sufficiency or necessity. In such cases, methods for the discovery of MINUS causation can only approximate strict MINUS structures by fitting their models more or less closely to the data using suitable parameters of model fit. Moreover, the fact that the instances of B in σ_{16} must be due to latent causes indicates that the unmeasured background of Table 2 is not (entirely) homogeneous, which, in turn, entails that causal inferences drawn from that table are not guaranteed to be

[10]Table 2 was more specifically simulated from (1) by, first, assembling all configurations compatible with (1), each instantiated by one case, second, introducing noise by randomly adding 10% of cases incompatible with (1), and third, randomly multiplying some cases and deleting others (see the replication script in the supplementary material for a stepwise generation of Table 2).

error-free. In order to nonetheless distill some causal information from such data, strategies for estimating the error risk and the reliability of issued models are needed.

The following two sections review how these problems are addressed by configurational comparative methods (CCMs) and logic regression methods (LRMs), respectively.

Configurational comparative methods

The best known CCM is *Qualitative Comparative Analysis* (QCA; Ragin 1987, 2008); a more recent addition to the family of CCMs is *Coincidence Analysis* (CNA; Baumgartner 2009; Baumgartner and Ambühl 2020). Both QCA and CNA have been developed with a focus on the analysis of data with low noise levels and no more than 15 to 20 exogenous factors. They aim to build all data-fitting models within user-defined complexity constraints.

Their core parameters of model fit are *consistency* and *coverage* (Ragin 2008, ch. 3). In crisp-set and multi-value data δ, consistency (*con*) and coverage (*cov*) of a Boolean dependence $\phi \rightarrow \psi$ are defined as follows:

$$con(\phi \rightarrow \psi) := \frac{|\, \phi * \psi \,|_\delta}{|\, \phi \,|_\delta} \qquad\qquad cov(\phi \rightarrow \psi) := \frac{|\, \phi * \psi \,|_\delta}{|\, \psi \,|_\delta} \tag{2}$$

where ϕ and ψ stand for Boolean functions of the factors in δ and $|\ldots|_\delta$ for the cardinality of the set of cases in δ instantiating the enclosed expression. What counts as acceptable scores on these fit parameters is defined in thresholds set by the analyst prior to the application of QCA or CNA. These thresholds determine how close a dependence in the data must approximate a strict Boolean dependence in order to pass as one of sufficiency or necessity. By convention, thresholds are often set between 0.75 and 1, the latter of which corresponds to strict Boolean dependence. For instance, if the consistency threshold is set to 0.8, A does not count as sufficient condition for B in Table 2 because $con(A \rightarrow B) = {}^{17}/_{29} = 0.59$, whereas $A{*}e$ does count as sufficient for B—despite the two cases instantiating configuration σ_{15}, which features $A{*}e$ without B—because $con(A{*}e \rightarrow B) = {}^{14}/_{16} = 0.88$.

QCA infers MINUS-formulas from data as in Table 2 by means of Quine-McCluskey optimization (McCluskey 1965) from switching circuit theory. It conducts a *top-down* search

that first assembles maximal conjunctions of exogenous factor values that meet the chosen consistency threshold, and thus count as sufficient, in a so-called *truth-table*; then it successively eliminates redundant conjuncts, and finally it combines minimally sufficient conjunctions to minimally necessary disjunctions. While this approach works fine for ideal data, it faces two problems when applied to non-ideal data. First, when data are fragmented it tends to require the introduction of unobserved configurations as simplifying assumptions. If these assumptions are unwarranted, complete redundancy elimination is blocked (Schneider and Wagemann 2012, sect. 8.2). Second, a top-down search may abort the minimization prematurely because finding redundancy-free Boolean predictors is not always possible via successive factor elimination but may require eliminating multiple factor values at the same time (Baumgartner and Ambühl 2020, sect. 3.1).

CNA, by contrast, infers MINUS-formulas by means of an algorithm custom-built for causal modeling that adopts a *bottom-up* search strategy bypassing truth-tables and is not affected by either of QCA's problems when processing fragmented and noisy data. As a result, CNA is more successful than QCA at avoiding redundancies and, because redundancy-freeness is crucial for MINUS causation, at inferring correct MINUS-formulas from non-ideal data. We will therefore use CNA as our CCM of choice in the remainder of this paper.

The CNA algorithm, which is implemented in the `cna()` function of the **cna** R package (Ambühl and Baumgartner 2020a), takes as inputs a data set δ with crisp-set, fuzzy-set or multi-value factors, consistency and coverage thresholds `con` and `cov`, an upper bound `maxstep` for the complexity of the models to be built, and an optional `ordering` parameter specifying candidate outcomes in δ. The algorithm then starts by searching for all *atomic* MINUS-formulas—single-outcome models—that meet `con` and `cov` in δ within the confines of `maxstep` for all candidate outcomes in the `ordering`. To this end, it tests, for all candidate outcomes, whether the consistency scores of the values of single exogenous factors in δ meet `con`; if that is not the case, conjunctions of two factor values are tested, then conjunctions of three, and so on, until `maxstep` is reached. Whenever a conjunction meets `con`, it is a minimally sufficient condition, and supersets of it are not tested any more. Next, single minimally sufficient conditions are checked for compliance with `cov`; if that check is negative, disjunctions of two are tested, then disjunctions of three, and so on, until `maxstep` is

reached. Whenever a disjunction meets `con` and `cov`, it is a minimally necessary disjunction of minimally sufficient conditions, that is, an atomic MINUS-formula. Finally, CNA combines atomic models to *complex* MINUS-formulas—multi-outcome models—representing the entire causal structure underlying δ. But as LRMs analyze structures with single outcomes only, we will not further discuss CNA's generation of multi-outcome models here. To maintain comparability with LRMs, the remainder of the paper will illustrate and benchmark CNA's performance by focusing on single-outcome structures only.

If CNA is run on Table 2 with a conventional threshold setting of `con` = `cov` = 0.8, a `maxstep` restricting model complexity to no more than 9 exogenous factor values (or *leaves* in LRM jargon), and an `ordering` specifying B as only candidate outcome, it returns the following two models with corresponding consistency and coverage scores:

$$A{*}e \;+\; C{*}d \;\leftrightarrow\; B \qquad con = 0.895\;;\;\; cov = 0.895 \tag{3}$$

$$A{*}e \;+\; c{*}D{*}e \;\leftrightarrow\; B \qquad con = 0.800\;;\;\; cov = 0.840 \tag{4}$$

These are all the MINUS-formulas inferable from Table 2 satisfying the chosen tuning parameters. An output consisting of multiple models is common for CNA—just as for many other methods. It means that the data underdetermine their own causal modeling at the chosen tuning parameters. Accordingly, the above CNA output is to be interpreted disjunctively, entailing that the ground truth is either (3) or (4).

Model (3), for example, identifies two alternative causal paths to B, one featuring A and e as parts of a complex cause and another one with C and d as parts of a complex cause. CNA models are to be interpreted relative to the data from which they have been inferred and to the threshold settings chosen for that inference. That means, in particular, that they do not purport to be complete representations of underlying causal structures. Rather, they only detail those causally relevant factor values along with those conjunctive and disjunctive groupings for which the data contain evidence at the chosen threshold settings. Thus, even though (3) and (4) do not ascribe causal relevance to E, they must not be interpreted to exclude that E is causally relevant for B. Given the frequent fragmentation of data processed by CCMs, their models only entail claims about causal relevance, not about causal irrelevance.

Another feature of CCM models that deserves emphasis is that they are sensitive to

changes in tuning parameters. CCMs track difference-making relations on the level of individual cases in the data (and not marginal effect sizes on the population level), and what counts as difference-making evidence changes with changes in tuning parameters, meaning that resulting models change as well. For instance, if we increase the coverage threshold to cov = 0.95, CNA returns model (5) for Table 2, or if we lower the consistency and coverage thresholds to con = cov = 0.7, models (6), (7), and (8) are issued.

$$A{*}e \;+\; C{*}d \;+\; c{*}D{*}e \;\leftrightarrow\; B \qquad con = 0.826\;;\;\; cov = 1.00 \tag{5}$$

$$A{*}e \;\leftrightarrow\; B \qquad con = 0.875\;;\;\; cov = 0.737 \tag{6}$$

$$C{*}d \;+\; c{*}D{*}e \;\leftrightarrow\; B \qquad con = 0.875\;;\;\; cov = 0.737 \tag{7}$$

$$A{*}d \;+\; c{*}D{*}e \;\leftrightarrow\; B \qquad con = 0.789\;;\;\; cov = 0.789 \tag{8}$$

While sensitivity to tuning settings is problematic from the perspective of methods quantifying effect sizes on the population level because varying effect sizes cannot be given a consistent causal interpretation, a lot of variance in CCM models merely reflects varying amounts of inferentially exploited difference-making evidence without implying any inconsistent causal conclusions. Two different models inferred with different tuning parameters do not contradict one another if the causal claims entailed by them stand in a subset relation, that is, if one of them is a *submodel* of the other. A model $\mathbf{m}_i$ is a submodel of another model $\mathbf{m}_j$ iff all causal relevance ascriptions as well as conjunctive and disjunctive groupings entailed by $\mathbf{m}_i$ are also entailed by $\mathbf{m}_j$. For example, (6) ascribes causal relevance to A and e and it places the two causes on the same path. This (and more) also follows from (3), meaning that (6) is a submodel of (3)—which makes (3) a *supermodel* of (6). A submodel does not conflict with its supermodel but merely makes less (or the same) causal claims.

But not all models inferable from Table 2 are mutually compatible. Model (8), for example, places A and d on the same path, whereas (3) places them on different ones. And even if two models are compatible, it does not follow that they both correctly reflect an underlying data-generating structure. Hence, criteria are needed to select among all the models inferable from data. A straightforward selection criterion is overall model fit, which can be defined as the product of a model's consistency and coverage scores. Based on that criterion,

(3), which is the ground truth behind the data in Table 2, is preferable over (4), (6), (7), and (8). However, model (5) has an even higher fit and it entails that not only $A{*}e$ and $C{*}d$ are causes of B but also $c{*}D{*}e$, which is false. (5) thus increases the fit at the cost of entailing false positives, meaning it is *overfitted*. This illustrates a common problem of CCMs: in noisy discovery contexts, the best fitting models often overfit the data (Arel-Bundock 2019). Hence, additional selection criteria are needed to counterbalance overall fit.

Parkkinen and Baumgartner (2021) propose a robustness criterion, tailor-made for CCMs, to reduce the overfitting risk. According to that proposal, the robustness of a model $\mathbf{m}_i$ is measured in terms of the degree to which $\mathbf{m}_i$'s causal attributions overlap with the causal attributions of all other models obtained from a series of data re-analyses under systematically varied `con` and `cov` thresholds. More specifically, the robustness of $\mathbf{m}_i$ corresponds to the number of sub- and supermodels $\mathbf{m}_i$ has among all the models inferred in such a re-analysis series. For example, if we re-analyze Table 2 at all `con` and `cov` settings in the interval $[0.65, 0.95]$, varied at increments of 0.1, it turns out that model (6) has 21 sub- and supermodels among all the resulting models, which is the highest number of all models and, thus, yields a normalized robustness score of 1. (3) has 16 sub- and supermodels, while (5) has 15, resulting in robustness scores of 0.76 and 0.71, respectively (cf. the replication script for details). That means the most robust model identifies $A{*}e$ as a conjunctive cause of B, which is true according to the ground truth and, hence, does not overfit the data. But (6) avoids overfitting at the cost of not completely recovering the ground truth, as it misses the causal relevance of $C{*}d$. Model (3), which not only correctly but also completely represents the ground truth, has significantly better fit than (6), yet significantly lower robustness. Moreover, (3) is slightly more robust than (5), yet fits the data slightly worse. In practice, the final model choice is a matter of weighing up these scores. In this particular case, the marginal gain in fit coupled with a loss in robustness and an increase in model complexity disqualifies model (5), whereas the choice between (6) and (3) is undetermined—but whichever of these two models ends up selected, only correct causal inferences will be drawn.

Logic regression

The first and best known LRM is *Logic Regression* (LR; Ruczinski et al. 2003). There exist various extensions of LR, for example, *Monte Carlo Logic Regression* (Kooperberg and Ruczinski 2005), *Logic Feature Selection* (Schwender and Ickstadt 2007), or, very recently, *Bayesian Logic Regression* (Hubin et al. 2020).[11] Apart from differences in the underlying algorithms, especially in the fitting and model selection protocols, and in the processed data types, the main difference between these methods concerns the logical form of their Boolean outputs. While LR outputs complete Boolean models furnishing sufficient and necessary conditions for the outcome, though in no standardized syntax (i.e. in no normal form) and without systematic minimization, the other LRMs output lists of best fitting sufficient conditions of the outcome, each of which syntactically standardized to a conjunction of factor values, but without combining them to a complete and minimized model that also furnishes a necessary condition. In light of our previous discussion of the MINUS theory, it is clear that neither of these outputs lends itself to a causal interpretation—which would require minimized necessary disjunctions of sufficient conditions in disjunctive normal form.

LRMs are not designed for causal discovery—in fact, no reference to the INUS or MINUS theory (or to any other theory of causation) appears anywhere in the LRM literature. Instead of tracing causation, LRMs search for association patterns that allow for *prediction*, or, as Ruczinski et al. (2003, 476) put it:

> "we attempt to find decision rules such as 'if X_1, X_2, X_3, *and* X_4 are true,' or 'X_5 or X_6 *but not* X_7 are true,' then the response is more likely to be in class 0. In other words, we try to find Boolean statements involving the binary predictors that enhance the prediction for the response. In the near future, one such example could arise from SNP microarray data (...), where one is interested in finding an association between variations in DNA sequences and a disease outcome such as cancer."

Correspondingly, the main field of application of LRMs are genetic association studies where higher order interactions among single nucleotide polymorphisms (SNPs) are investigated

[11] An overview over LR and its main extensions is provided in (Schwender and Ruczinski 2010).

for their associations with variations in phenotype, for example, in disease risk.

Still, as we shall see below, the Boolean models of LR can be minimized and brought into the standardized syntax of MINUS-formulas by means of suitable post-processing, without thereby changing the truth conditions, the predictive content, or the fit of these models. Such post-processing is possible because LR models not only provide sufficient but also necessary conditions for the outcome. As the other LRMs abstain from issuing necessary conditions, their outputs cannot be analogously post-processed without changing the content or fit of these outputs, meaning without diverting the methods from their intended use. For that reason, we will subsequently focus on LR and its models only.

LR can efficiently analyze high-dimensional, large-n data with high noise levels, but it is devised for binary (crisp-set) outcomes only. The **LogicReg** R package (Kooperberg and Ruczinski 2019), which implements LR, can process data with up to 1000 factors and build models with up to 128 exogenous factor values. No CCM can process such data and construct models of that complexity. But while CNA returns all data-fitting models within user-defined complexity bounds, including models with multiple outcomes, LR issues one best fitting model with one outcome only. To this end, it embeds Boolean expressions in a generalized linear regression model of the following form:

$$g(\mathsf{E}[Y]) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j, \tag{9}$$

where $\mathsf{E}[Y]$ is the expected value of the outcome Y, g is a link function, and L_j are Boolean expressions, for example, $L_j = X_2{}^* x_4{}^* X_7$ (Ruczinski et al. 2003, 479).[12] Models are fit to the data using standard scoring functions from regression analysis. In the linear case, the most common score is the residual sum of squares, meaning that LR searches for models of form (9) such that L_j and the estimated parameters minimize the residual sum of squares.

LR represents Boolean expressions as logic trees where factor values appear as leaves connected via branches through the operators AND and OR (Ruczinski et al. 2003, 478).

[12] In order to avoid confusion with arithmetic operations, disjunction is commonly symbolized by '$\vee$', conjunction by '$\wedge$', and negation by a superscripted '^{c}' in the LRM literature. For reasons of compatibility with the CCM notation, we cannot follow this convention here. Hence, the sign "$*$" in L_j is to be interpreted in terms of conjunction, not multiplication, yet "+" in (9) stands for addition, not disjunction. Everywhere else in this paper (and in the replication script), "+" represents disjunction, not addition.

The search for best fitting models (standardly) implements a *simulated annealing algorithm* (Otten and van Ginneken 1989) that starts from the empty or null model and proceeds by iteratively performing tree transformations. These transformations are, in each iteration, randomly selected from a set of six *possible moves* consisting in addition, deletion, and alternation of operators or of leaves (Ruczinski et al. 2003, 481). After each move, the fit of the new tree is contrasted with the fit of the old tree. If the fit is equal or better, the move is always accepted and the next transformation is performed; if the fit is worse, the move is accepted with the following probability:

$$a(\mathbf{m}_o, \mathbf{m}_n, T) = \min\{1, \exp((f(\mathbf{m}_o) - f(\mathbf{m}_n))/T)\}, \tag{10}$$

where $f(\mathbf{m}_o)$ is the fit of the old tree, $f(\mathbf{m}_n)$ is the fit of the new tree and T is a parameter called *temperature* that decreases with the progression of the algorithm as specified in a simulated annealing cooling scheme. (10) entails that, in the early stages of the search, new trees with worse fit than old trees are accepted with high probability, whereas that acceptance probability tends towards zero in later stages. The rationale behind initially accepting trees with decreased fit is to allow the algorithm to scan large portions of the search space without getting stuck in mere local optima.

As anticipated above, the models resulting from this procedure often are not causally interpretable—for three main reasons. First, no syntactic constraints are imposed on the tree transformations, meaning that the ultimately selected tree may have any non-standardized syntactic form. However, to interpret a Boolean model in terms of causal conjunctivity and disjunctivity, it must have disjunctive normal form. Second, LR's set of possible moves allows for the introduction of logically redundant elements into the trees. For instance, if $a + B$ is the old tree, one possible move is to add a conjunct as follows: $a + A{*}B$. This transformation, however, adds nothing whatsoever to the content of the model because the new tree is logically equivalent to the old one and, hence, induces the same residual sum of squares. As tree transformations with equal fit are always accepted, that new tree is accepted despite its redundant leave. But, of course, the logically redundant A does not make a difference to the outcome and, hence, is not a cause. Third, LR can embed multiple trees in one and the same regression model. Yet, multiple tress cannot be interpreted in terms

of one causal structure; causal structures with conjunctivity and disjunctivity are represented by single Boolean models, that is, by single trees.

This third obstacle to a causal interpretation of LR models is easily overcome. The `logreg()` function, which implements LR in the package **LogicReg**, provides an argument `ntrees` by means of which LR can be constrained to build models with no more than one embedded tree. Removing the first two obstacles is more intricate. It requires a post-processing of the Boolean expressions (trees) in LR models that standardizes their syntax and minimizes them by eliminating redundancies. Transforming a non-standardized Boolean expression into a minimized disjunctive normal form is known as *Boolean minimization* and there exist various algorithms for this task. But as the task is NP-complete, the running time of all these algorithms grows exponentially with the number of factors in the Boolean expressions, such that applying them to LR models caps the maximal complexity of these models somewhere between 20 and 30 factors. Moreover, most algorithms for Boolean minimization are not tailored towards causal data analysis but towards simplicity and cost-reduction (e.g. in electrical engineering). As a consequence, they only output one minimized expression, even though there often exist multiple equivalent ones, each of which might correspond to the data-generating causal structure (Baumgartner and Thiem 2017). An exception is the *ereduce* algorithm as implemented in the `ereduce()` function of the **cnaOpt** package (Ambühl and Baumgartner 2020b), which tackles Boolean minimization against the background of causal data analysis and returns all logically equivalent minimized disjunctive normal forms.[13] In what follows, we therefore render the Boolean expressions output by LR causally interpretable by post-processing them with `ereduce()`.

Such post-processing yields that, despite its original focus on prediction, LR can be used to search for the same causal target as CNA and its models can be interpreted in the same way as CNA models: they entail causal relevancies as defined by the MINUS theory but no irrelevancies, submodels do not conflict with supermodels, and if post-processing by `ereduce()` yields multiple causal models, they are to be interpreted disjunctively. The main remaining difference between LR and CNA then is that the latter purposefully builds all data-fitting

[13] In a nutshell, `ereduce(x)` searches for minimal hitting sets in the Boolean expression `x` that prevent `x` from being false in the data. Another more well-known approach to find all minimalized disjunctive normal forms is *Petrick's method* (Roth and Kinney 2010, section 6.3), but there does not exist a ready-made implementation of that approach in R.

models within given complexity bounds, whereas the former randomly moves through the search space, honing in on one, or (after post-processing) a small number of equivalent best fitting models. One upshot of this difference is that the output of CNA does not vary between re-analyses of the data using the same tuning parameters, whereas the model(s) issued by LR may vary from re-analysis to re-analysis. Two central tuning parameters controlling the LR output are the number (`iter`) of iterated tree transformations and the maximum number (`nleaves`) of leaves (exogenous factor values) in the fitted tree. Repeatedly re-running LR on the data in Table 2 at `iter` = 25000, setting the same complexity upper bound as in our previous CNA application, *viz.* `nleaves` = 9, and post-processing the resulting trees by `ereduce()`, yields a wide array of Boolean models. Here are the models with frequencies n and fit scores e resulting from one particular series of 100 re-runs:[14]

$$A{*}C{*}d + A{*}e + c{*}D{*}e \leftrightarrow B \qquad n = 70 \quad e = 0.277 \tag{11}$$

$$A{*}C{*}d + A{*}e \leftrightarrow B \qquad n = 8 \quad e = 0.291 \tag{12}$$

$$A{*}e + c{*}D{*}e + C{*}d{*}E \leftrightarrow B \qquad n = 5 \quad e = 0.277 \tag{13}$$

$$A{*}e + C{*}d{*}E \leftrightarrow B \qquad n = 2 \quad e = 0.291 \tag{14}$$

$$A{*}e + C{*}d + c{*}D{*}e \leftrightarrow B \qquad n = 9 \quad e = 0.277 \tag{15}$$

$$A{*}e + C{*}d \leftrightarrow B \qquad n = 6 \quad e = 0.291 \tag{16}$$

To recall, the data in Table 2 are simulated from the MINUS structure in expression (1), which corresponds to (16) in the above list and, hence, is only returned in 6 of the 100 re-runs. In the vast majority of re-runs, LR outputs an overly complex model entailing false causal relevancies and, hence, overfitting the data.

This illustrates that LR also faces a severe overfitting risk, just as do CCMs (or statistical methods). To reduce that risk, LR provides various instruments, among which there is a `penalty` parameter punishing model complexity, analogous to the Akaike Information Criterion (AIC), and a permutation test randomly permuting the outcome and checking whether the best model fit obtainable from the permuted data is equal or even better than the best fit obtained from the original data. The model inferred from the original data should only be in-

[14]The lower e, the better the fit. Moreover, e does not express the fit of the Boolean expression alone but of its embedding in a linear regression model of form (9), which we do not reproduce here.

terpreted to reflect an actual signal in the data if that check is negative in most permutations. If we repeat the above re-analysis series of Table 2 (with the same replication seed as before) setting `penalty` = 2 (which, according to Kooperberg and Ruczinski 2019, corresponds to AIC) and performing the permutation test, overfitting disappears entirely, as all 100 re-runs yield the same MINUS-formula:

$$A{*}e \leftrightarrow B \qquad n = 100 \qquad e = 0.459 \tag{17}$$

(17) is a submodel of the ground truth (1) and, as such, only makes true causal claims. Of course, it has worse fit than models (11) to (16) and it does not completely reflect the ground truth. But without re-introducing an excessive overfitting risk the complete ground truth cannot be recovered by LR from Table 2. In particular, increasing `iter` has no effect on the output and reducing `penalty` to 1, again, yields an overfitted model in the majority of re-runs. This is essentially due to the data's small sample size of only 45 cases, which is a size way below LR's ordinary domain of application. Nevertheless, when its models are suitably post-processed and complexity sufficiently penalized, LR consistently infers the very same model from Table 2 that is also the most robust CNA model.

In sum, both CNA and LR can be tuned to correctly analyze the MINUS structure used to simulate the data in Table 2. Of course, showing that these methods successfully detect MINUS causation in one specific example serves mere illustration purposes. The next section therefore performs systematic benchmark tests on a broad array of examples.

Benchmarking

To benchmark the performance of CNA and LR under a variety of discovery contexts, we set up a series of inverse search trials, first, randomly generating data-generating structures (or ground truths), second, simulating different types of data from those structures, and third, processing that data with CNA and LR to measure the degree to which their outputs comply with various benchmark criteria. This section first explains the details of the test setups and benchmark criteria and then discusses the test results and the ensuing synergy potential.

Test setups and benchmarks

To ensure the comparability of CNA and LR, the trials must be confined to data dimensions and data-generating structures analyzable by both methods, meaning that important features of both methods cannot be tested in the following. As CNA is more restricted in regard to the dimensionality of the data, we confine the trials to a set $\mathbf{F} = \{A, B, C, D, E, F, G, H\}$ of 8 factors and to sample sizes of no more than 1000 cases, and because LR can only treat one factor in $\mathbf{F}$ as outcome, which moreover must be binary, we restrict $\mathbf{F}$ to binary factors and randomly generate ground truths Δ from $\mathbf{F}$ with a single outcome and between 1 and 9 causes (leaves) each.[15] To get a statistically significant performance assessment, we generate a total of 1000 ground truths Δ.[16]

The performance of CNA and LR is influenced by the sample size of the data, by the level of noise, and by the data's *fragmentation* (or *limited diversity*). As that latter concept is specific to methods tracing MINUS causation, it must be explicitly introduced here. A causal structure Δ with one outcome and n mutually independent exogenous factors, each of which can take y values, is compatible with y^n possible configurations of those exogenous factors. If all of these possible configurations are observed at least once, the resulting data are non-fragmented (or saturated). But real-life data are hardly ever non-fragmented because some possible configurations are rare and thus unobserved. That is, fragmentation is the ratio of unobserved configurations to all possible configurations. The higher the sample size and the lower the noise level and fragmentation, the better the performance of CNA and LR. To ensure the replicability of our test series, however, we cannot vary the sample size, the noise level, and the fragmentation in a controlled manner all at once. We therefore decided to randomize fragmentation and systematically vary only the sample size and noise level.

To this end, we first produce ideal data δ^{id} for every Δ comprising one case per configuration of the factors in $\mathbf{F}$ compatible with Δ. Then, a randomly drawn percentage of

[15] Note that a cause is a factor taking a value (not a factor). As multiple values of the same factor can be causes, it is possible to draw structures with more causes than there are factors in $\mathbf{F}$. Examples are (11) or (13).

[16] The number of ground truths determines the number of trials in each test type. 1000 trials were chosen because the means of the resulting benchmark scores calculated from different samples of that size were found to stabilize with standard errors of the means between 0.0004 and 0.015 (see Figure 5 in the Online Appendix). In other words, we can have high confidence that trials on a sample of 1000 ground truths drawn from $\mathbf{F}$ are representative of the population of all ground truths that can be built from $\mathbf{F}$.

configurations, between 0% and 50%, is removed from every δ^{id} to yield fragmented data sets δ^{fr} with fragmentations anywhere between 0% and 50%. Next, samples of 60, 200, and 1000 cases are drawn from each δ^{fr}, with replacement and equal selection probability for each case in δ^{fr}. This results in small-sized data sets δ^{fr}_{60}, intermediate-sized data δ^{fr}_{200}, and large-sized data δ^{fr}_{1000}. Finally, from each δ^{fr}_{60}, δ^{fr}_{200}, and δ^{fr}_{1000}, four noisy data sets are created by replacing, respectively, 5%, 15%, 25%, and 35% of the cases compatible with Δ by randomly drawn cases incompatible with Δ—which incompatibilities can be thought of as resulting from measurement error or confounding. Each case compatible with Δ has equal probability of being replaced by an incompatible case and each incompatible case has equal probability of being drawn, meaning that noise is unbiased. The result of this procedure are 12 data types comprising 1000 data sets each, $\delta^{5\%}_{60}, \delta^{15\%}_{60}, \ldots, \delta^{15\%}_{200}, \delta^{25\%}_{200}, \ldots, \delta^{25\%}_{1000}, \delta^{35\%}_{1000}$, where subscripts indicate the sample sizes and superscripts the noise levels.

Next, each of these 12000 data sets is analyzed by CNA and LR. CNA is run with a robustness check systematically re-analyzing each data set at all `con` and `cov` settings in the interval [0.6, 1], varied at increments of 0.1, and retaining the models in the 95$^{\text{th}}$ percentile of robustness scores. The complexity of models to be built is limited to 9 factor values. LR is induced to fit exactly one logic tree to the data with the same upper complexity bound of 9 leaves; it is run with `penalty` = 2, `iter` = 25000, and a permutation test checking for signal in the data. The logic tree output by LR is then post-processed by `ereduce()` in order to generate all MINUS-formulas corresponding to that tree.

The sets of MINUS-formulas **S** output by CNA and LR are tested against three increasingly stringent benchmark criteria, measuring first, whether they are *error-free*, second, whether they contain a *correct* model, and third, to what degree correct models in **S** *completely* reflect the ground truth. A set **S** is error-free iff it does not entail a causal claim that is false of the ground truth Δ (i.e. no false positive). That can be satisfied in two ways: either (i) **S** is empty, meaning no causal inferences are drawn (e.g. because CNA's fit thresholds cannot be met or because LR's permutation test is negative), or (ii) **S** contains at least one[17] model $\mathbf{m}_i$ that is correct of the ground truth Δ, which is the case iff $\mathbf{m}_i$ is a submodel of Δ.

[17]Recall that an output containing multiple models is to be interpreted disjunctively; and a disjunction of models is true iff at least one model is true.

So, **S** satisfies the first benchmark criterion iff it satisfies conditions (i) or (ii).[18]

The second benchmark focuses on non-empty sets **S** only and checks whether condition (ii) is satisfied, meaning whether **S** actually contains at least one model $\mathbf{m}_i$ that is a submodel of Δ, and thus correct. That is, an empty set **S** does not pass the second benchmark. Finally, the third criterion assesses the informativeness of correct models. Of two different correct models one can be more complex than the other and, hence, reveal Δ more completely. The completeness benchmark, therefore, measures the degree to which the correct models in **S** exhaustively reveal Δ. More specifically, completeness amounts to the ratio of the complexity of the most complex correct model in **S** to the complexity of Δ, where complexity of a model $\mathbf{m}_i$ is understood (as is standard for both CNA and LR) as the number of factor values (leaves) in $\mathbf{m}_i$. That is, contrary to the first and second benchmarks, which can only be passed or not, the third benchmark can be passed by degree, but when **S** is empty or does not contain a correct model, completeness is 0 by default.[19]

Results

The results are presented in the bar-charts of Figure 1. The *x*-axes break them down by the noise levels, the columns by the sample sizes, and the rows by the benchmark criteria. The error-freeness and correctness scores represent the ratios of CNA and LR analyses that produce outputs complying with error-freeness and correctness, respectively. For example, CNA's score of 0.99 in the first bar of the top panel in the left column means that CNA's output is error-free in 99% of the trials run on the 1000 data sets of type $\delta_{60}^{5\%}$. By contrast, the completeness scores represent the degrees to which correct models completely exhibit the ground truths averaged over all 1000 analyses of a corresponding data type. For example, LR's score of 0.39 in the last bar of the bottom panel in the right column means that LR recovers 39% of the ground truth, on average, in 1000 trials performed on data of type $\delta_{1000}^{35\%}$.

[18]The reader may wonder why we test a benchmark that can be passed by a trivial method producing empty outputs by default. The reason is that such a method would be entirely uninformative, which would be visible in its failing the second and third benchmarks, correctness and completeness; but an empty output produced by a method that does not fail on the other benchmarks is a valuable piece of information entailing that the data do not warrant any causal conclusions. The capacity to abstain from drawing causal inferences when no such inferences are warranted is a crucial methodological asset that deserves to be benchmarked.

[19]Unlike completeness, we do not quantify correctness because there currently does not exist an adequate quantitative correctness measure for MINUS models. It is not trivial to meaningfully quantify the seriousness

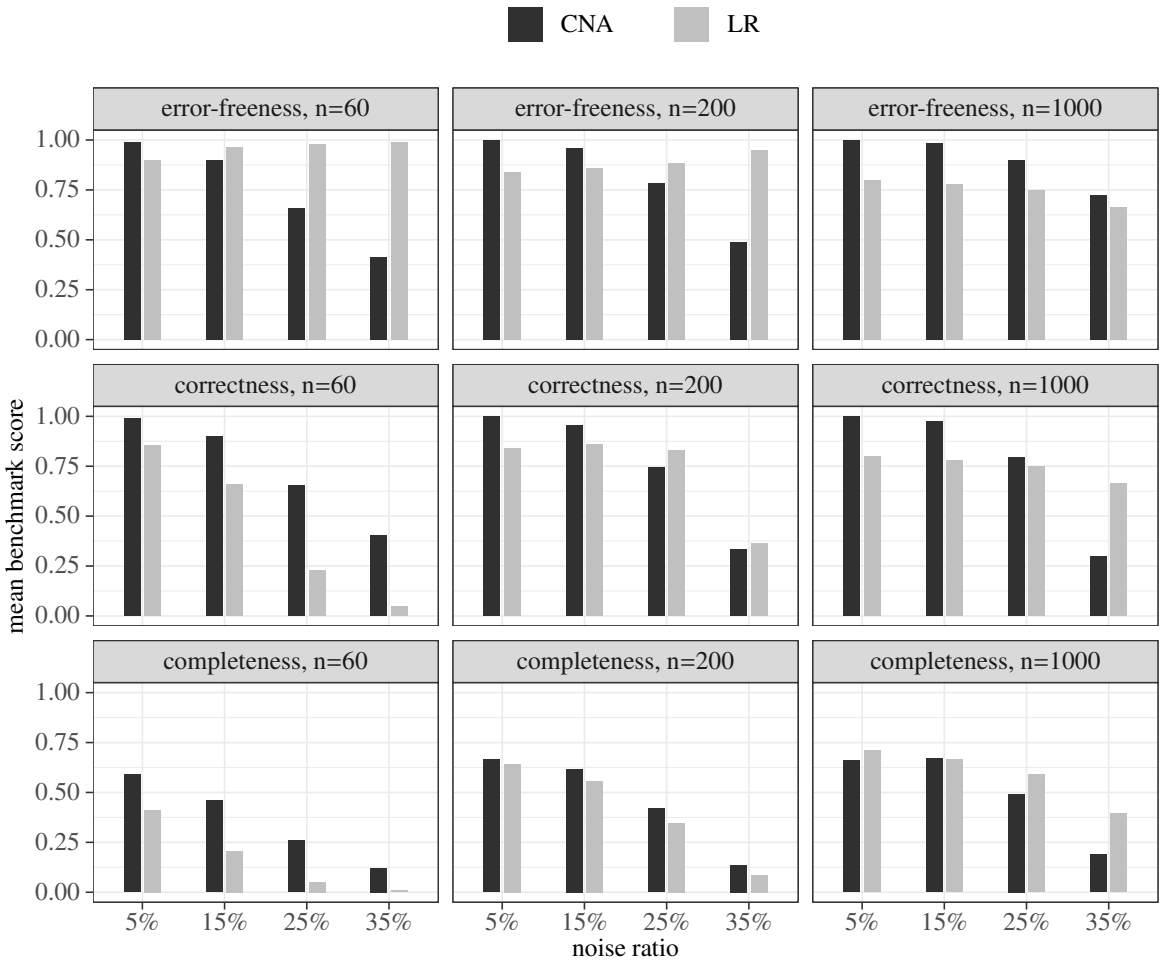

Figure 1: Bar-plots displaying mean scores of CNA (in black) and LR (in grey) in error-freeness (i.e. ratio of trials without false positive), correctness (i.e. ratio of trials in which a correct model is recovered), and completeness (i.e. ratio of the ground truth complexity recovered), averaged over 1000 trials in each test type, broken down by noise levels (on x-axis) and sample sizes (from left to right). These means have standard errors between 0.0004 and 0.015.

The overall finding is that CNA and LR have strengths and weaknesses in different benchmarks. LR has advantages in error-freeness, CNA in correctness, and the completeness scores are in favor of CNA in small-sized data and in favor of LR in large-sized data. More specifically, when averaged over all data types, LR avoids erroneous inferences in 86% of the trials, finds a correct model in 64%, and recovers 39% of the ground truth complexity, whereas CNA's output is error-free in 82%, contains a correct model in 75% of the trials, and recovers 44% of the ground truths. Further differentiation shows that CNA outperforms LR in both error-freeness and correctness in low noise data, committing almost no false pos-

of errors. For some more details on this problem see Parkkinen and Baumgartner (2021).

itives and finding a correct model in at least 90% of the trials. But those scores plummet when CNA is applied to high noise data, independently of the sample size. At 35% noise, CNA issues a correct model in only 34% of the trials, drawing an erroneous inference in more than half of the trials when the data have small or intermediate size. Contrary to its correctness scores, CNA's error-freeness scores in high noise data increase with increasing sample sizes, but only reach an acceptable value (i.e. 0.73) if $n = 1000$.

LR, by contrast, maintains a constant score on error-freeness independently of the noise. Remarkably though, scores in large-n trials are roughly 20 percentage points lower than the corresponding scores in small-n trials. This finding requires explanation. LR successfully avoids false positives in trials on high noise data with small sample sizes because it mostly outputs no models at all—which can be read off the huge differences between LR's error-freeness and correctness scores in those trials (the only way to avoid false positives without actually recovering correct models is by issuing no model). Whereas LR is right to mostly abstain from drawing inferences from data of types $\delta_{60}^{35\%}$ and $\delta_{200}^{35\%}$, data of type $\delta_{60}^{25\%}$ would allow for finding a correct model in 65% of the trials, which is CNA's correctness score in those trials. Hence, LR is overly cautious in drawing inferences from small-n data with no more than 25% noise. At the same time, LR avoids erroneous inferences equally frequently as CNA in the high noise trials on large-n data and outputs significantly more correct models.

That is, LR's permutation test effectively induces LR to abstain from drawing an inference when none is warranted (e.g. in $\delta_{60}^{35\%}$ and $\delta_{200}^{35\%}$). But when LR actually draws an inference that inference tends to be less frequently correct than CNA's, except for the high noise trials on large-n data. By contrast, running CNA with a robustness check in the threshold interval $[0.6, 1]$ very reliably leads to the recovery of a correct model, but it does not prevent CNA from too frequently committing false positives when analyzing small- and intermediate-sized data with high noise levels.

The cautiousness of LR when processing small-sized data yields that it recovers the ground truths only half as completely when $n = 60$ as does CNA; but both methods find only fractions of the complete ground truths, *viz.* 17% (LR) and 36% (CNA), on average. These low completeness scores are due to the fact that small-n data have high fragmentation. It follows that a lot of information about the possible behavior patterns of the analyzed

factors and, thus, about the underlying causal structure is missing. As is to be expected, the completeness scores of both methods increase with increasing sample sizes—despite all data types in our test series being fragmented to some (randomized) degree. Averaged over all noise levels, CNA recovers 50% of the ground truths from large-n data and LR 59%.

These findings must be further contextualized by relativizing them to the number of models output.[20] Averaged over all trials of the series, CNA's exhaustive model search over a large interval of threshold settings results in an output comprising 5.4 models, whereas LR's search heuristic coupled with its frequent abstinence to draw a causal inference yields an average of only 0.78 models. That means, in return, that in the vast majority of trials in which LR draws an inference that inference (after post-processing) is unambiguous, featuring exactly one model, whereas when CNA draws an inference it mostly issues multiple models scoring equally on fit and robustness. The difference in model numbers is particularly large in the trials on high noise data of small to intermediate sizes. On the one hand, an analyst is unlikely to draw determinate causal conclusions when she is presented with more than a handful of models, which reduces the risk of drawing false conclusions from high noise data, but on the other hand, it means that the edge CNA has over LR as regards the correctness of its output comes at the price of a significantly higher ambiguity ratio. In other words, although there mostly exists a multitude of equally well fitting MINUS models, LR's average correctness score of 0.64 shows that its simulated annealing algorithm coupled with a complexity penalty, a permutation test, and suitable post-processing is remarkably successful at honing in on one model that is actually true of the ground truth.

This finding, again, needs contextualization. First, in those trials in which LR issues a (non-empty) model, that model has a mean complexity of 3.3 leaves, whereas CNA's models have a mean complexity of 4.1. The less complex a model, the fewer causal ascriptions it makes, and the more likely it is that no false ascriptions are made. It follows that it is less difficult to find at least one correct model at LR's mean output complexity than at CNA's. Second, LR's permutation test and the need to post-process its models come at a significant cost to computing times.[21] When our test series is run, in parallel, on a computer with 16

[20]The number of models output in a particular test type are collected in the columns `LR.ambig` and `CNA.ambig` of the score object `results` in the replication script.

[21]The times a method takes to analyze a data set are collected in the lists `speedCNA` and `speedLR` in the

EPYC cores, 3 GHz and 32 GB RAM, LR needs 0.15 seconds, on average, to analyze a data set without a permutation test and post-processing, but adding such a test and post-processing increases the average execution time to 4.8 seconds—in which time, to repeat, LR does not produce all MINUS models complying with the tuning settings. By contrast, it takes 0.88 seconds, on average, for CNA to find all models meeting the tuning settings. Additionally checking the robustness of these models increases execution times to 2.7 seconds, on average.

Synergy potential

We have seen that CNA and LR can both be used to search for MINUS causation. The fact that the two methods conduct this search by means of very different techniques creates an ideal setting for synergies. On the one hand, exhibiting a performance difference in a particular discovery context of a benchmark test gives the underperforming method a clear indication of possible improvements. For example, our results demonstrate that CNA needs to be complemented by a procedure for assessing signal strength in the data, analogous to LR's permutation test. Such a procedure will prevent CNA from misfiring so frequently in high noise data; at the same time, judging by LR's template, it can be expected to increase CNA's computation times considerably. Our results also show that LR's permutation test is too cautious and the complexity penalty approximating AIC too restrictive when analyzing low noise data. There might be a way to relax those constraints and, thereby, increase LR's correctness and completeness scores without falling back into the overfitting pitfall.

On the other hand, the complementarity of the two methods' strengths and weaknesses opens the way for cross-validation studies applying both methods to the same data. To make this concrete, we identified those trials in our benchmark experiment in which CNA and LR output at least one identical model (which also obtains if both return the empty model). The ratios of those trials are plotted in Figure 2, along with the error-freeness, correctness, and completeness scores reached by those identical models. CNA and LR issue an identical model in one fourth of the trials, on average. Independently of the noise or sample sizes, these identical models very rarely induce a false positive, as they reach a mean error-freeness score of 0.95, which is significantly higher than either of the methods' individual scores. Av-

replication script.

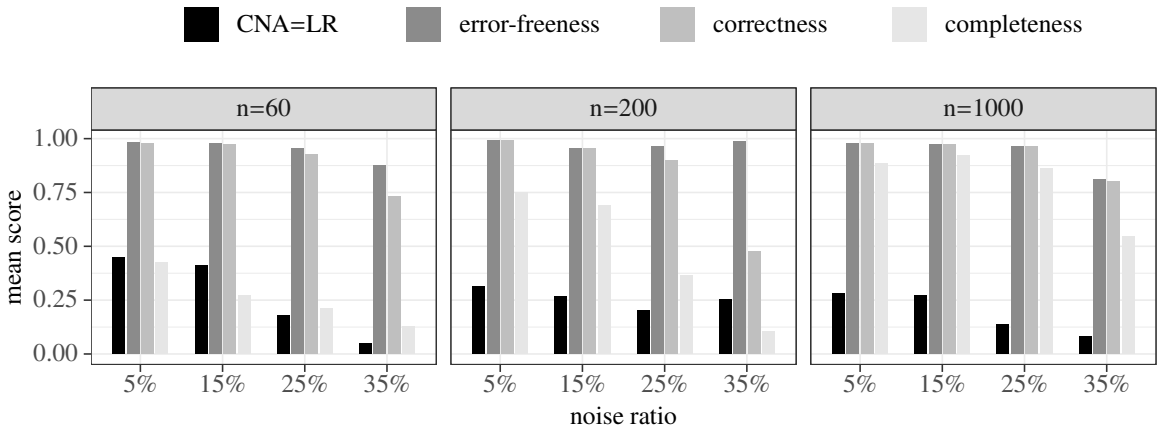

Figure 2: Black bars represent the ratios of trials in which CNA and LR produce an identical model. Increasingly lighter shades of grey represent the error-freeness, correctness, and completeness scores reached by the subpopulation of identical models.

eraged over all trials, there is a correct model among 89% of identical models. When only looking at the data types for which both CNA and LR typically output a non-empty model, *viz.* all data types except for $\delta_{60}^{25\%}$, $\delta_{60}^{35\%}$, and $\delta_{200}^{35\%}$, there is a correct model among 95% of identical models. Comparing these scores to the mean correctness scores in the population of CNA models (0.75) and LR models (0.64) renders the potential of cross-validation palpable. An identical output of both methods is significantly more likely to be true of the data-generating structure than one that is not cross-validated.

What is more, the increase in correctness scores is not due to the fact that identical models would be less informative (and thus more likely to be correct). Though models returned by both methods are somewhat less complete than CNA's models in small-n data, they reach higher completeness scores than CNA and LR individually in the trials with intermediate-n and large-n data. In particular when analyzing data with $n = 1000$, a cross-validation study finding identical models reveals 80% of the ground truths, on average—independently of the noise. In sum, although it only happens in one fourth of the trials that CNA and LR produce the exact same models, when it happens, the corresponding models receive a validation boost in error-freeness and correctness without compromising on completeness.

While it is not very frequent that CNA's and LR's outputs contain identical models, it happens in 70% of the trials that a CNA model and an LR model are related by the submodel relation. In trials in which LR does not mostly abstain from returning a model (i.e. in all tri-

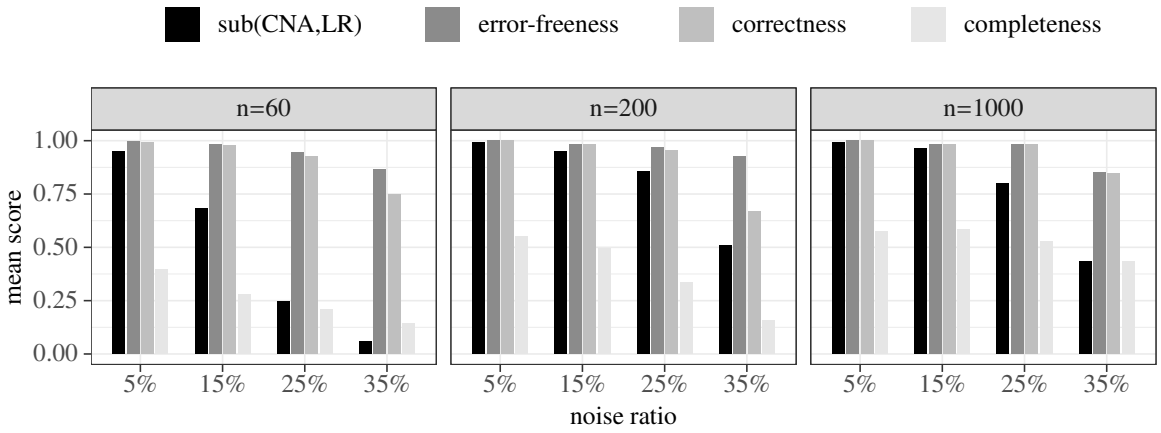

Figure 3: Black bars represent the ratios of trials in which some of CNA's and LR's models are related by the submodel relation. Increasingly lighter shades of grey represent the error-freeness, correctness, and completeness scores reached by the subpopulation of models related by the submodel relation.

als except for the ones on $\delta_{60}^{25\%}$, $\delta_{60}^{35\%}$, and $\delta_{200}^{35\%}$), it holds in 85% of the analyses, on average, that at least one of the CNA models is a submodel of one of the LR models, or vice versa. That is, CNA and LR often output closely related models that differ only in the degree of detail with which they represent the underlying causal structures. The bar-charts in Figure 3 depict the ratios of trials in which the two methods output models related by the submodel relation, along with the error-freeness, correctness, and completeness scores in the population of those related models. As in case of identical models, submodels reach a very solid mean error-freeness score of 0.96. Because being identical is a special case of being related by the submodel relation, most submodels are less complex than their supermodels; and because making fewer causal claims correlates with making fewer mistaken claims, the mean correctness score over all trials among the submodels is even higher than among the identical models, *viz.* 0.92. For the same reason, however, the mean completeness score among submodels is only 0.39 and, thus, lower than among identical models and CNA models taken separately. That means if CNA models are submodels of LR models, or vice versa, those submodels receive a validation boost in correctness at the price of reduced completeness.

Limitations and Conclusions

Before we conclude, we want to highlight the relevant limitations of our analysis. Our test series only simulates a proper subset of possible discovery scenarios. For instance, it does not analyze the effects of varying latent causes on the performance of CNA and LR. We simulate data from the complete data-generating structures, meaning there are no unmeasured causes in our tests. As long as latent causes are homogenized in the unmeasured causal background of the data, they do confound the data and, hence, do not constitute a problem for CNA or LR (see Baumgartner and Ambühl 2020). But if unmeasured causes, in particular *common* causes of two (or more) measured factors, vary in an uncontrolled manner, they negatively affect the performance of any causal discovery method. It would be an important topic for a follow-up study to determine whether or not the performances of CNA and LR are differently affected by uncontrolled variation in latent (common) causes.

Furthermore, our analysis introduces fragmentation and noise at random. Yet of course, such data deficiencies may be non-random in real-life data. Certain configurations may be more likely than others to be unobserved, certain factors may be more easily affected by measurement error, or certain types of measurement error may be more frequent than others. If fragmentation or noise are biased, they tend to induce stronger spurious dependencies than if they are unbiased; and stronger spurious dependencies are more likely to be mistaken for causal dependencies. An important question we cannot answer with our analysis, hence, is whether there are relevant differences in how CNA and LR handle non-random fragmentation and noise and whether the cross-validation potential is similar under these circumstances.

Finally, it should be reiterated that our benchmark experiments only scrutinize parts of the inferential power of CNA and LR. To ensure comparability of the results, we restrict our tests to data processable by both of them. CNA can also process multi-value and fuzzy-set data as well as data generated by multi-outcome structures. But the dimensionality of the data CNA can handle in reasonable time is limited to about 20 exogenous factors. Without permutation test and model post-processing, LR, by contrast, quickly finds models for data featuring a couple hundred exogenous factors. But the applicability of LR is restricted to binary outcomes and it cannot process data generated by structures with multiple outcomes.

Still, within the scope limitations of our analysis, we find a substantive cross-validation potential. Any study analyzing data processable by both methods can expect to profit, in one way or another, from such cross-validation. In case of low noise data, CNA should be the primary tool of analysis. But CNA often outputs more than one equally fitting model, leaving the analyst with the task of model selection. Our results suggest that also finding one of those model candidates or a submodel of it by LR provides a strong incentive to select that model, because a cross-validated (sub)model has a roughly 90% chance of being true of the data-generating structure. That is, in low noise settings, cross-validation is an instrument for ambiguity reduction. When analyzing data with intermediate noise levels, cross-validation serves the purpose of enhancing the correctness of the inference. A non-empty model returned by both methods is about 20% more likely to be correct than a model that is only returned by one of them. In high noise settings, LR—with permutation test and post-processing—should be the primary tool of analysis. If LR abstains from issuing a model, any model CNA might infer from the same data should be met with skepticism. But if LR draws an inference from high noise data that can be cross-validated with CNA, our results suggest that that model may be causally interpreted despite the noisy discovery context. As the mean correctness of cross-validated models that are related by the submodel relation is roughly 75% in high noise contexts, it is even justified to give preference to the CNA model in that case, if it is a supermodel of the LR model.

Moreover, our results provide no reason to conclude that this synergy potential is restricted to data processable by both methods. Rather, we expect CNA and LR to be implementable sequentially as data pre-processing tools for one another as well. If an analyzed process is hypothesized to involve multiple outcomes, a preliminary CNA analysis might partition the data into batches pertaining to the separate outcomes, each of which could then also be processed by LR. Or, if data dimensionality is beyond CNA's limitations, LR might be employed to build a preliminary model identifying a limited group of relevant factors, based on which the data could then be subsetted to a dimensionality manageable by CNA, which, in turn, would render the data amenable to a cross-validation study as described in the previous paragraph. Of course, these suggestions are tentative. Properly fleshing them out will require separate studies. But in light of the promising results obtained in the restricted

scope of our analysis, we conclude that such follow-up studies are worthwhile at any rate.

This paper aimed to bridge the gap between methodological communities with very little mutual exchange by showcasing two methods that can be used to uncover causal structures featuring conjunctivity and disjunctivity. Configurational comparative methods as exemplified by CNA and logic regression method as LR apply very different techniques for that purpose and have complementary strengths and weaknesses. While the former have correctness and completeness advantages, the latter have advantages in fallacy- and ambiguity-freeness. In light of that complementarity, the potential for cross-validation is considerable. So far, however, it remains entirely untapped. We have made a first attempt at paving the way for the future exploitation of that potential.

References

Ambühl, M. and M. Baumgartner (2020a). ***cna****: Causal modeling with Coincidence Analysis.* R Package Version 3.0.1. https://cran.r-project.org/package=cna.

Ambühl, M. and M. Baumgartner (2020b). ***cnaOpt****: Optimizing Consistency and Coverage in Configurational Causal Modeling.* R Package Version 0.2.0. https://cran.r-project.org/package=cnaOpt.

Arel-Bundock, V. (2019). The double bind of Qualitative Comparative Analysis. *Sociological Methods & Research*, 1–20. doi: 10.1177/0049124119882460.

Armstrong, D. M. (1983). *What is a Law of Nature?* Cambridge: Cambridge University Press.

Baumgartner, M. (2009). Uncovering deterministic causal structures: A Boolean approach. *Synthese 170*, 71–96. doi: 10.1007/s11229-008-9348-0.

Baumgartner, M. and M. Ambühl (2020). Causal modeling with multi-value and fuzzy-set Coincidence Analysis. *Political Science Research and Methods 8*, 526–542. doi: 10.1017/psrm.2018.45.

Baumgartner, M. and C. Falk (2019). Boolean difference-making: A modern regularity theory of causation. *The British Journal for the Philosophy of Science*. doi:

10.1093/bjps/axz047.

Baumgartner, M. and A. Thiem (2017). Model ambiguities in configurational comparative research. *Sociological Methods & Research 46*(4), 954–987. doi: 10.1177/0049124115610351.

Beirlaen, M., B. Leuridan, and F. Van De Putte (2018). A logic for the discovery of deterministic causal regularities. *Synthese 195*(1), 367–399. doi: 10.1007/s11229-016-1222-x.

Bowran, A. P. (1965). *A Boolean Algebra. Abstract and Concrete*. London: Macmillan.

Brambor, T., W. R. Clark, and M. Golder (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis 14*(1), 63–82. doi: 10.1093/pan/mpi014.

Clarke, K. A. (2020). Logical constraints: The limitations of QCA in social science research. *Political Analysis*, 1–17. doi: 10.1017/pan.2020.7.

Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety*. San Francisco: Jossey-Bass Publishers.

Culverhouse, R., B. K. Suarez, J. Lin, and T. Reich (2002). A perspective on epistasis: Limits of models displaying no main effect. *The American Journal of Human Genetics 70*(2), 461–471. doi: 10.1086/338759.

Godfrey-Smith, P. (2010). Causal pluralism. In H. Beebee, C. Hitchcock, and P. Menzies (Eds.), *Oxford Handbook of Causation*, pp. 326–337. Oxford: Oxford University Press.

Graßhoff, G. and M. May (2001). Causal regularities. In W. Spohn, M. Ledwig, and M. Esfeld (Eds.), *Current Issues in Causation*, pp. 85–114. Paderborn: Mentis.

Hausman, D. (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.

Hubin, A., G. Storvik, and F. Frommlet (2020). A novel algorithmic approach to Bayesian Logic Regression (with discussion). *Bayesian Analysis 15*(1), 263 – 333. doi: 10.1214/18-BA1141.

Hume, D. (1999 (1748)). *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press.

Kooperberg, C. and I. Ruczinski (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genetic Epidemiology 28*(2), 157–170. doi: 10.1002/gepi.20042.

Kooperberg, C. and I. Ruczinski (2019). ***LogicReg**: Logic Regression*. R package version 1.6.2. https://CRAN.R-project.org/package=LogicReg.

Lemmon, E. J. (1965). *Beginning Logic*. London: Chapman & Hall.

Mackie, J. L. (1974). *The Cement of the Universe. A Study of Causation*. Oxford: Clarendon Press.

McCluskey, E. J. (1965). *Introduction to the Theory of Switching Circuits*. Princeton: Princeton University Press.

Mill, J. S. (1843). *A System of Logic*. London: John W. Parker.

Otten, R. H. J. M. and L. P. P. P. van Ginneken (1989). *The Annealing Algorithm*. Boston: Kluwer Academic Publishers.

Parkkinen, V.-P. and M. Baumgartner (2021). Robustness and model selection in configurational causal modeling. *Sociological Methods & Research*. doi: 10.1177/0049124120986200.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Ragin, C. C. (1987). *The Comparative Method*. Berkeley: University of California Press.

Ragin, C. C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Rihoux, B. and C. C. Ragin (Eds.) (2009). *Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related Techniques*. Thousand Oaks: Sage.

Rohwer, G. (2011). Qualitative Comparative Analysis: A Discussion of Interpretations. *European Sociological Review 27*(6), 728–740. doi: 10.1093/esr/jcq034.

Roth, C. H. and L. L. Kinney (2010). *Fundamentals of Logic Design* (6 ed.). Stamford: Cengage Learning.

Ruczinski, I., C. Kooperberg, and M. LeBlanc (2003). Logic regression. *Journal of Computational and Graphical Statistics 12*(3), 475–511. doi: 10.1198/1061860032238.

Schneider, C. Q. and C. Wagemann (2012). *Set-Theoretic Methods: A User's Guide for Qualitative Comparative Analysis (QCA) and Fuzzy-Sets in the Social Sciences*. Cambridge: Cambridge University Press.

Schwender, H. and K. Ickstadt (2007). Identification of SNP interactions using logic regression. *Biostatistics 9*(1), 187–198. doi: 10.1093/biostatistics/kxm024.

Schwender, H. and I. Ruczinski (2010). Logic regression and its extensions. *Advances in Genetics 72*, 25–45. doi: 10.1016/B978-0-12-380862-2.00002-3.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search* (2 ed.). Cambridge: MIT Press.

Yakovchenko, V., E. J. Miech, et al. S. S. Rogal (2020). Strategy configurations directly linked to higher Hepatitis C virus treatment starts: An applied use of configurational comparative methods. *Medical Care 58*(5). doi: 10.1097/MLR.0000000000001319.

Online Appendix

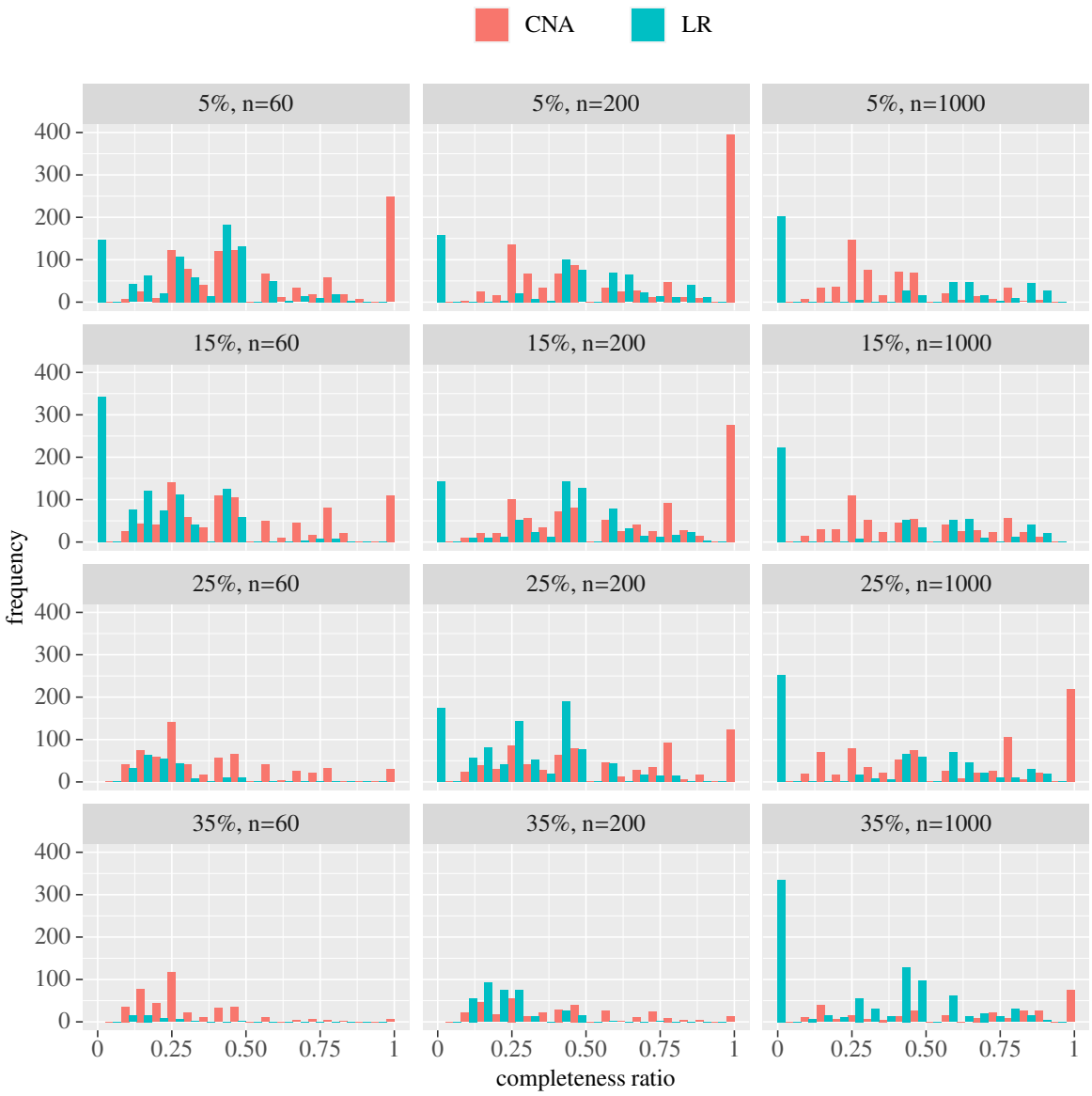

Figure 4: Histograms displaying the frequency (on the *y*-axis) of correct CNA models (in red) and LR models (in turquoise) reaching the completeness ratios on the *x*-axis. The rows break the trials down by the noise levels, the columns by sample size.

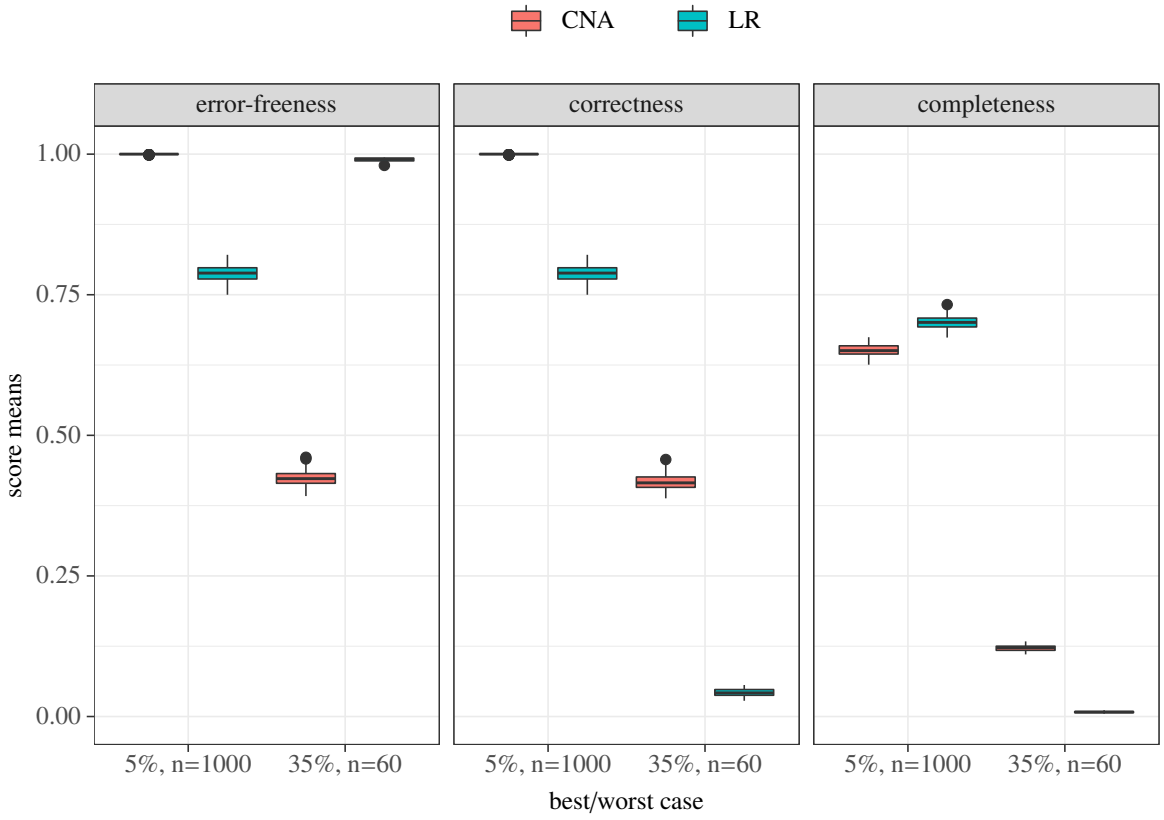

Figure 5: Box-and-whisker plots showing the distribution of mean benchmark scores obtained from 60 replications of the trials on data of types $\delta_{1000}^{5\%}$ and $\delta_{60}^{35\%}$ (i.e. the best and worst case trials) simulated from 1000 different ground truths each. Boxes represent the IQR (i.e. inter-quartile range between the 25^{th} and 75^{th} percentile). Whiskers extend to $1.5 \times$ IQR. Dots represent outliers.