Causal Discovery and Logic

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen,
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op vrijdag 14 juni 2013
om 10:00 uur precies

door

Thomas Claassen

geboren op 8 september 1969
te Sittard

Promotor:

Prof. dr. Tom Heskes

Manuscriptcommissie:

Prof. dr. Peter Lucas
Prof. dr. Thomas Richardson (University of Washington)
Dr. Ricardo Silva (University College London)

SIKS Dissertation Series No. 2013-22

The research in this thesis has been carried out under the auspices of the Dutch Research School for Information and Knowledge Systems (SIKS), and the Institute for Computing and Information Sciences (iCIS) of the Radboud University Nijmegen.

This research was supported by NWO Vici grant nr.639.023.604.

ISBN 978-90-820674-0-8
Cover design: TMQ

Contents

Chapter 1

Introduction

1.1 'Scientists have established a link between ...'

On an almost daily basis one can hear about the latest scientific breakthroughs and discoveries. News bulletins report on findings that can help us to better understand the world, decide on new policies, or improve our quality of life. Popular media are often interested in links related to physical and emotional well-being, such as behavioural patterns associated with heart diseases, food supplements related to a reduced risk on dementia, and steps to improve the chance on a successful career or relationship; but also in other areas, such as the link between crime and socioeconomical status, or the rise in unemployment figures since the latest austerity measures. Scientific journals tend to report on more complex and detailed relations, such as genes linked to cancer signalling pathways, the connection between nitrogen levels and bio-diversity, and the link between solar activity and climate change.

Nearly always the link is understood to imply a **causal** connection, where the first element somehow triggers, contributes to, or alters the chance on the second. The importance of finding causal relations is universally understood: knowing what influences what provides a level of *control* over one's life or environment. But strictly speaking the reported link only claims the discovery of a statistical correlation. And by the famous adage 'correlation does not imply causation' this in itself is not enough to warrant a causal interpretation. So how *can* we go from one to te other? That is the subject of the field of causal discovery

Statistical vs. causal risk factors

In medicine, the gold standard in establishing the effect of a treatment is the double-blind randomized controlled trial, where subjects are allocated to different treatment groups at random, and both subjects and analysts do not know who belongs to which

group: any statistically significant difference can then (only) be attributed to the *causal effect* of that treatment.

However, the majority of research is not of this form: in many cases it would be too expensive, unethical, or simply impossible to realize. Other types of experiments range from single-blind to open or uncontrolled (no control group) trials, all the way to purely observational studies. In such cases, finding variables that positively correlate with the disease/condition under scrutiny are known as **risk factors**. But no matter how strong the correlation is or how plausible the explanation may seem, simply interpreting a risk factor as causal may have unexpected, possibly even harmful consequences.

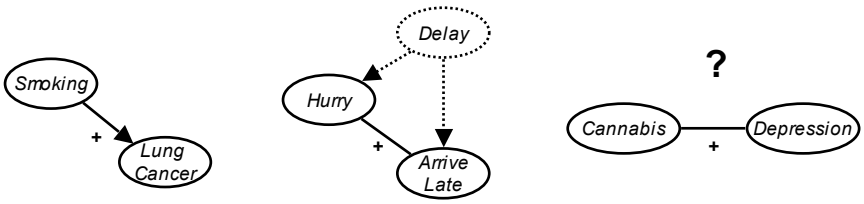

Figure 1.1: Statistical vs. causal risk/prediction factors. (a) Lung cancer and smoking (b) Hurrying and Arriving Late, (c) Cannabis and Depression/Schizophrenia

For example, in Figure 1.1 *Smoking* is a well-established risk factor for *Lung cancer*, as a higher proportion of smokers will develop lung cancer than of non-smokers. Indeed, we know *Smoking* $\Rightarrow$ *Lung cancer*, and so quitting smoking is an effective way to reduce the risk on lung cancer. But in a classic example attributed to Judea Pearl, by the same token *Hurrying* is a strong risk factor for *Arriving Late*, as people that are late still try to arrive on time. But clearly, taking your time will not help you arrive earlier. In fact, not hurrying is likely to *increase* the number of times you arrive late - the exact opposite of what the link suggested - as both factors result from, e.g. being delayed.

In this case the difference is obvious because we know from experience what the causal relations are. But the point of causal discovery is to identify relations we do not already know. So what to do if we find a link between *Cannabis* and mental disorders such as *Depression* or psychosis? It is possible that cannabis has a negative influence on the brain, and so banning it would be beneficial to the public health. But it is also possible that people start using cannabis to cope with depression, or that people who are susceptible to depression are also susceptible to drug use. It could even be that cannabis relief offers some form of protection against depression.

So, only if the risk factor is also a **causal risk factor** does a ban have a positive effect on reducing depression: in all other cases the effect is either zero or negative.

What makes a relation *causal*?

In order to decide which risk factors are causal, we need to know what exactly a causal relation *is*. In everyday speak the term 'cause' (or absence thereof) is used for many subtly different concepts. For example:

- a sedentary lifestyle can cause a heart attack (increase likelihood),
- the new fertilizer caused the grass to grow beautifully (significant factor),
- he fell into crime because he never finished school (most important),
- she won Olympic gold because she kept believing (necessary, not sufficient),
- no, you were not late because you had to wash-up your cup (insignificant),
- alcohol causes breast cancer through increased oestrogen levels (mechanism),
- he hit the jackpot because he kept betting on black (specific instance),
- what caused the system to break down? (explanation),
- the driver in the blue car caused the accident (blame).

Usually it is clear from the context what is meant, but that is not good enough for implementation in an algorithm. Given this inherent ambiguity, we want to capture the essence of what people mean when talking about causality, without getting bogged down in the philosophical quagmire of the 'correct' definition, be it through counterfactuals [Lewis, 1973], hypothetical interventions, structural equations models [Pearl, 2000], probability raising preceding factors, INUS conditions[1], or epistemological constructs [Williamson, 2005].

A key aspect behind a causal relation $X \Rightarrow Y$ is the idea of **effective manipulation**. The *manipulation* part captures the notion that a cause can or must influence its effect, which in turn suggests a degree of *deliberate control* of X, either in size, number or likelihood, over Y. The *effective* part captures the notion of relevance: if you cannot really notice an effect, it may as well not be there at all. Together they emphasize a causal model as a 'summary of influence': concluding that 2,500 out of 10,000 genes contribute to some degree to a disease is near useless (even if true); but finding five major markers constitutes an important medical breakthrough. In other words: causal discovery should distinguish between the relations that matter and the ones that don't.

We can identify a few other properties of causal relations: they relate to the real, physical world (mathematical equality and logical implication are not causal), they are constant/persistent (relations do not stop/start being causal), time-dependent (causation takes time), stable (in similar circumstances on average similar things happen), **transitive** (if $X \Rightarrow Y$ and $Y \Rightarrow Z$ then $X \Rightarrow Z$), and irreflexive ($X \not\Rightarrow X$, unless time is explicitly ignored as in cyclic models).

Finally an example of the distinction between singular causation ('what happened') and generic causal relations In the famous *Butterfly effect* from chaos theory [Lorenz, 1963], the flap of a hypothetical butterfly in Brazil causes a tornado

[1] insufficient but non-redundant part of a condition which is itself unneccessary but sufficient for the occurrence of the effect, [Mackie, 1988]

in Texas, in the sense that in the same conditions without the flap the tornado does not appear. But this just illustrates sensitive dependence on initial conditions in nonlinear systems in a specific instance: almost any change is likely to lead to radical differences given sufficient time. In practice, training butterflies is not an effective means to create hurricanes. In contrast, an expected consequence of global warming is an increase in extreme weather, including the incidence of hurricanes [IPCC, 2008]. Given the consensus that a large part of this trend can be attributed to human activity, it follows that in terms of effective manipulation: human induced climate change causes hurricanes ... butterflies do not.

Causal explanations for observed correlations

So far we found that a *statistical* risk factor is good for predicting likely cases, but that you need a *causal* risk factor to find a good treatment. They look the same, so how to distinguish. To do so it helps to know the basic causal configurations that can produce a correlation. We will illustrate these by an example from an area of research in micro-biology that has gained much attention in recent years: the human gut flora.

After it was discovered that the human digestive tract, and in particular the gut, contained a great many different types of bacteria (> 1000 species), researchers quickly realized that the impact of presence, absence, or abundance of these organisms on their environment, and so indirectly on the general state of health was largely unknown. Inspired by the famous example of stomach ulcers that were tought to be related to stress, but turned out to be result from an infection with the Helicobacter Pilori bacterium, people started to look for associations between such organisms and all kinds of intestinal conditions, in an attempt to identify unknown risk factors, ultimately leading to new treatments.

Example 1.1. *Figure 1.2 displays four basic causal configurations that can produce an observed association between abundance of a micro-organism (red) and the condition of the gut (blue), resulting in certain abdominal symptoms.*

Case (c) depicts that Helicobacter is a (direct) ***cause*** *of ulcers, as mentioned above. In (b) the dangerous hospital bacterium Clostridium difficile establishes itself when competing organisms are killed off by antibiotics: C.diff flourishes as an* ***effect*** *of changes in its environment. Fig.(d) displays how coeliac disease is caused by an overreaction of the immune system to the presence of transglutaminase, which results in inflammation of the villi lining the small intestine. It is not unlikely that one or more of the hundreds of species of bacteria in the gut also suffers adversely from this immune reaction as a* ***common cause*** *without having any direct impact on/from the tissue inflammation.*

Finally, in (e) the abundance presence of a certain bacteria is the result of a high-fat diet that induces symptoms such as acid reflux and abdominal cramps that are similar to certain inflammatory bowel diseases (IBDs). A study under people with such symptoms will show a negative correlation, suggesting some sort of

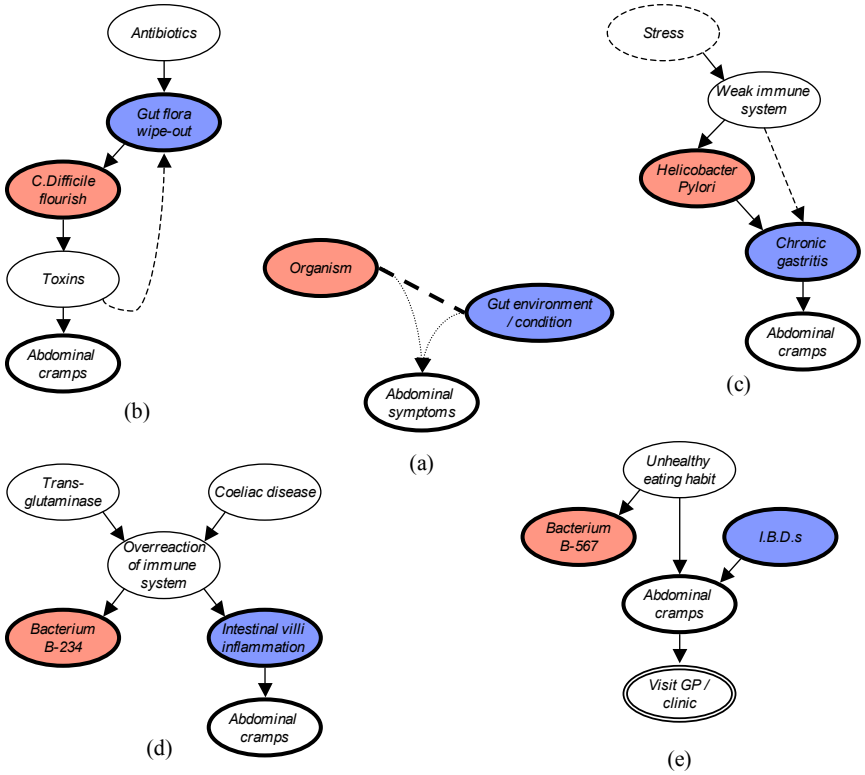

Figure 1.2: Four different causal configurations that produce an observed link between abundance of a micro-organism and the state of health in the gut. (a) Gut flora (red) as established risk factor for intestinal condition (blue) and resulting abdominal symptoms (white), together with four possible causal explanations: (b) change in environment causes *C.diff* to flourish, (c) *Helicobacter* infection causes ulcer, (d) immune system affects both stomach lining and gut flora, (e) selection on symptoms suggests bacteriological link with inflammatory bowel disease (IBD)

protective/preventative effect that could be commercially very attractive, even though actively adding the bacterium to your diet would not make you any less susceptible to IBDs. This spurious relation, resulting from focussing on cases with certain symptoms, is known as ***selection bias****.*

Four possible associations, but only in one case was the organism the actual cause of the condition. To establish which one we have we need to find additional relations that can help to eliminate the others. How to do that efficiently is detailed in chapter 3.

Causal discovery

The basic challenge of causal discovery can now be depicted as in Figure 1.3: given a data base of results from one or more experimental and/or observational studies, and armed with the combined knowledge of previous research and literature, the causal discovery algorithm should produce a precise and informative model of all the causal relations (or absence thereof) it could or could not find. To make this connection it needs information on how to distinguish between causal relations and 'mere' associations. Here we can do a little better than the famous 'no causes in, no causes out' from [Cartwright, 1989] in the form of 'no causal assumptions in, no causal relations out'. The output can take any form, but in this thesis we primarily focus on graphical causal models.

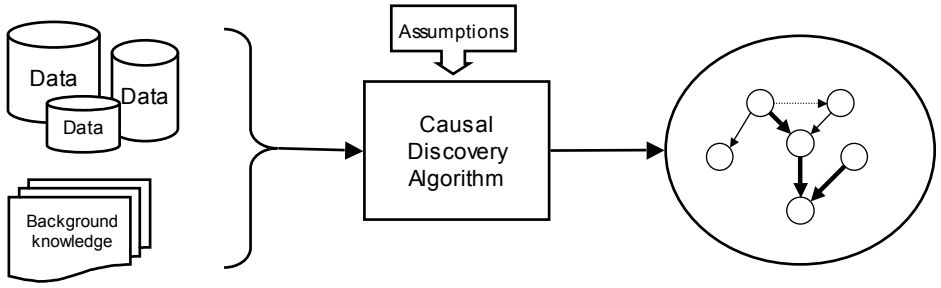

Figure 1.3: Causal discovery overview

Causal discovery is not about truth: we can never prove something about the real world; all we can (and should) aim for is that it is *valid* in the sense that if the conclusions turn out to be false, then the input (data, knowledge, or assumptions) must be wrong.

Our ultimate goal is to make the entire process completely transparant, where all relevant information is available to the algorithm prior to analysis, and every subsequent causal conclusion can be traced back to the exact pieces of information it was based on, irrespective of the nature of the experiments or the specific area of research the causal model applies to. In practice, for now we still have to cut a number of corners, but the results in this thesis should represent a decent step towards that final goal.

In the remainder of this thesis we are no longer concerned with meaning and interpretations of causality: for that the reader is referred to, e.g. Lewis [1973]; Dawid [2000]; Pearl [2000]; Cartwright [2004]; Williamson [2005]. However, before we focus entirely on causal relations in abstract models we take a closer look at how 'effective manipulation' translates to results from real-world observational and experimental studies.

1.2 From dynamical systems to causal discovery

In this section we take a look at a toy system to see if we can capture our intuitive 'effective manipulation' in a useful, objective definition, and what possible complications to expect when trying to discover them in practice. After all: if we want to claim to understand causality and find 'real' causal relations from probabilistic data we should at least be able to understand/indicate how and where they reside in the representation of a simple real-world experiment.

After this section we will not go back to this level of detail, but instead work directly from the high-level abstract model representation of causal relations between observed variables. But it is helpful to realize what actually underpins all the neat and tidy graphical models in subsequent sections, and how certain assumptions and simplifications relate to properties of underlying real-world systems.

The behaviour of many (if not all) physical systems can be described, at least in principle, in terms of a dynamical system[2]:

Definition 1.2. *A **dynamical system** is a tuple $(T, \mathcal{M}, \Phi)$, where*

- *T represents a set of **time** parameters,*
- *$\mathcal{M}$ is a manifold called the **state space** or **phase space** of the system,*
- *$\Phi^t(x) : T \times \mathcal{M} \to \mathcal{M}$ is an **evolution rule**.*

The evolution rule maps states $x \in \mathcal{M}$ at t_0 to states x' at $t_0 + t$. Dynamical systems have the **memorylessness** property, which means that a state $\mathbf{x}(t > t')$ is *independent* of all states $\mathbf{x}(t < t')$ given the (full) state $\mathbf{x}(t')$. It includes discrete systems, such as the *logistic map*:

$$x_{n+1} = r x_n (1 - x_n), \quad r \in [0, 4]$$

which maps the interval $[0, 1]$ onto itself and shows chaotic behaviour for $r \gtrsim 3.57$, but also cellular automata like Conway's Game of Life. However, in this thesis we focus on the subclass known as **real-time dynamical systems** or **flows**, where time is continuous: $T = \mathbb{R}$, a state $\mathbf{x} \in \mathcal{M}$ is a d-dimensional vector of variables, and the evolution rule Φ^t is the integrated solution to a set of d (coupled) differential equations $\dot{\mathbf{x}} = v(\mathbf{x})$ such as, for example, Newton's equations of motion. A **trajectory** $\mathbf{x}(t) = \Phi^t(\mathbf{x}_0)$ is the path through phase-space traced out by the evolution rule for a given initial point $\mathbf{x}(0) = \mathbf{x}_0$. Note that constants are simply variables with time-derivative $\dot{\mathbf{c}} = 0$, and that reversible systems can be run backwards in time with negative t.

[2] See, e.g. [Cvitanović *et al.*, 2010] for a comprehensive introduction to this topic with ample focus on describing and deriving (causal) properties of dynamical systems.

Example: throwing a drawing pin

To see how the properties of real-world dynamical systems, observed probabilistic in/dependencies, and inferred causal relations are related we take a look at one of the simplest dynamical systems that displays all this interdependence: throwing a drawing pin or thumbtack.

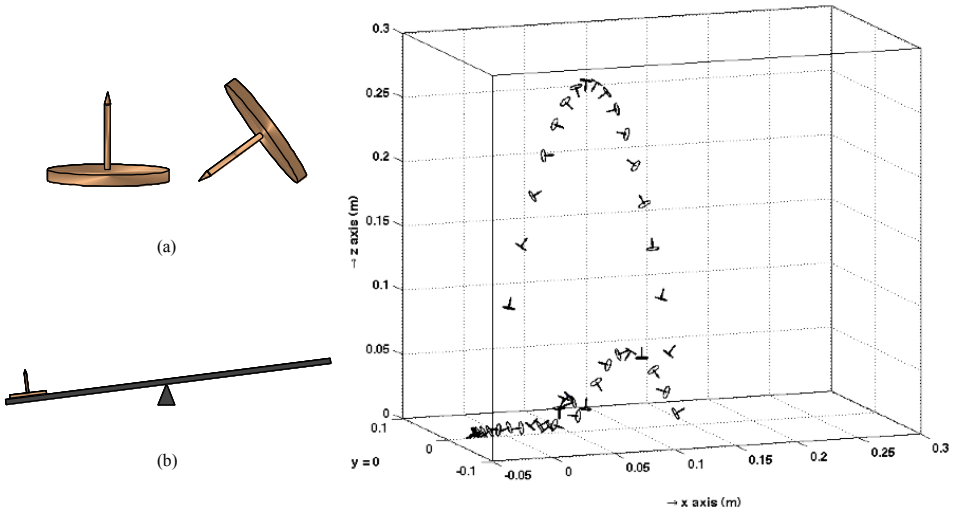

Figure 1.4: (a) Stable outcomes of throwing a drawing pin: 'pin up' and 'pin down', (b) Experimental setup, and time-lapse visualization of a sample throw, see text for details.

Figure 1.4 shows the general setup: a drawing pin is launched in a certain direction from a certain height with a certain speed, bounces over a flat surface and finally comes to rest with either pin up or pin down (a). Despite the simplicity of the system it exhibits a rich and chaotic bouncing behaviour, including transitions high speed/low rotational motion → low speed/high rotation (and vice versa), bouncing backwards, sliding with friction → stick → flip over, static → dynamic contact (slipping), balancing/bouncing near unstable equilibrium positions, etc.

We implemented a simulation algorithm in Matlab using a $4/5^{th}$ order Runge-Kutta ode-solver with discontinuity detection and interpolation.[3] The physics modeling for the pin was based on [Baraff and Witkin, 1997], where we implemented static/dynamic friction at contact points according to [Bender and Schmitt, 2006], and used the Guendelman *et al.* [2003] approach to handle bounces. To reduce the computational complexity we limited the simulation to symmetric motion in the $x - z$ plane. Parameters for the simulation include:

[3] Available physics engines for use in computer games do not accurately simulate this bouncing behaviour, as they need to handle potentially very many objects in real-time.

- **constants** pin dimensions h_{pin} etc., density ρ, inertial tensor $\mathbf{I}_{pin}$, gravity g, (rotational) drag ν, static/dynamic friction $\mu_{s/d}$, bounce-restitution κ;
- **variables** position vector $\mathbf{x}$, orientation matrix $\mathbf{R}$, speed $\mathbf{v}$, rotation ω, angular momentum $\mathbf{L}$, force F, torque τ;
- **initial values** at $t = 0$ for speed $\mathbf{v}_0$, height z_0, orientation $\mathbf{R}_0$, rotation ω_0;
- **thresholds** transition dynamic $\leftrightarrow$ static, bounce $\leftrightarrow$ contact, ode accuracy ϵ

while the equations of motions are described as:

$$\begin{array}{lcccl} \text{position} & : & \dot{\mathbf{x}} & = & \mathbf{v} \\ \text{orientation} & : & \dot{\mathbf{R}} & = & \omega^{\star}\mathbf{R} \\ \text{linear momentum} & : & \dot{\mathbf{p}} & = & F \\ \text{angular momentum} & : & \dot{\mathbf{L}} & = & \tau \end{array} \tag{1.1}$$

with linear velocity $\mathbf{v} = \mathbf{p}/M$, and angular velocity given by $\omega = \mathbf{R} \cdot \mathbf{I}_{pin}^{-1} \cdot \mathbf{R}^T \cdot \mathbf{L}$. The total external force F and torque τ are summed contributions from gravity, drag, friction, normal contact forces, etc., according to whether the pin is in free flight, bounce, or contact mode.

Apart from the system parameters we define a number of additional **observables** (random variables), including:

- R : final result pin-up/down-left/-right, see Figure 1.4(a),
- X : final horizontal distance from starting point along x-axis,
- B : horizontal distance at first bounce,
- H : max. height along z-axis *after* first bounce, etc.

The r.h.s. of Figure 1.4(b) shows a sample bounce for an 8mm long copper pin (diameter head = 1.0cm) on a flat vinyl surface (friction coefficient $\mu = 0.5$, restitution $\kappa = 0.6$), with $v_0 = 2.1$m/s, $\omega_0 = 18$rad/sec. Launched softly from $z_0 = 0.1$ at an angle of 0.2rad from the vertical, the pin completes just over one full rotation before it bounces backwards (due to friction) with much higher rotational velocity, from distance $B = 13.9$cm to reach height $H = 5.8$cm. After a few hops and bounces the pin slides to rest at $X = -3.2$cm with result $R =$ 'pin down (left)'. Trajectories for slightly different initial values quickly diverge, for example $v_0 = 2.0999$m/s already results in a different outcome R. But some random variables show coherence over larger intervals, e.g. final horizontal distance X correlates strongly with the direction of the first bounce (determined by speed and orientation of the pin) which in this case remains backwards for values $\omega \in [16.6 - 18.4]$.

Probability and proportions

As stated, in a real-world dynamical system each initial condition $\mathbf{x}_0$ is mapped unambiguously by the evolution rule Φ^t to future states $\mathbf{x}_0(t)$, including the outcome of any additionally defined random variable V. Just as we can consider the evolution of a single point, we can also consider the evolution of an entire *region* $\mathcal{M}_i$ through phase-space, including the proportion of trajectories from that region that result in

event $V = v$ at time t_V. The initial regions represent a **state of knowledge** $\mathcal{I}$ about (the parameters of) the system. Some parameters are known/defined exactly, others are assumed to be in a some interval, perhaps with certain values preferred over others, possibly in complex combinations. We can represent this knowledge through a normalized **initial density** $\rho(\mathbf{x})$ over the region $\mathcal{M}_i$, such that $\int_{\mathcal{M}_i} \mathrm{d}x\, \rho(\mathbf{x}) = 1$.

With this the **probability** (density) of outcome $V = v$ given a state of knowledge $\mathcal{I} = \{\rho(\mathbf{x}), \mathcal{M}_i\}$ becomes:

$$p(V = v \,|\, \mathcal{I}) = \int_{\mathcal{M}_i} \mathrm{d}\mathbf{x}\, \rho(\mathbf{x}) \delta\big(v - V(\Phi^{t_V}(\mathbf{x})\big) \tag{1.2}$$

In words: in a dynamical system setting probabilities take the form of relative proportions of (densities over) regions of initial values in phase-space that are mapped to the given outcome.

Given a state of knowledge, the probability $p(V = v \,|\, \mathcal{I})$ is unambiguously defined. If *all* parameters are known exactly, then the region $\mathcal{M}_i$ reduces to a single initial value $\mathbf{x}_0$ and probability collapses to a trivial yes/no property.

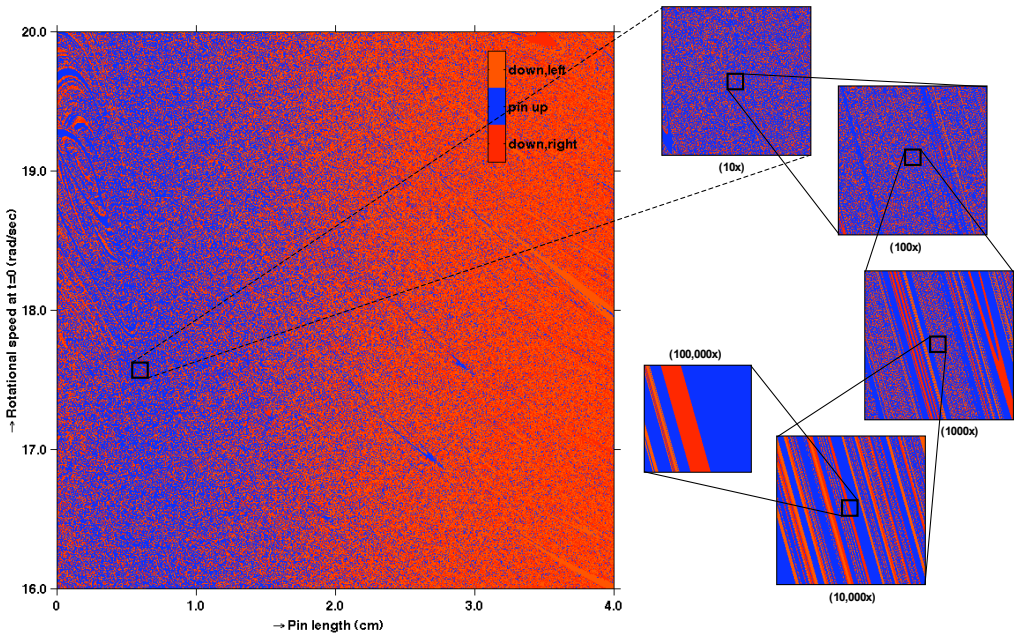

Figure 1.5: Outcomes pin-up/down in the drawing pin experiment for pins with length h_{pin} varying from 0 (effectively a coin) to 4cm (nail), and initial rotation ω_0 between $[16-20]$rad/sec; each time for a copper pin with $d_{head} = 1$cm, from height $z_0 = 0.1$m with initial speed $v_0 = 8$m/s under an angle of 0.25rad with the vertical, and constant friction, drag, etc. The zoom plots illustrate how the proportion $p(R = \text{`}pin\ up\text{'} \,|\, \mathcal{I})$ changes with different regions of initial values.

The probability concept from eq.(1.2) is illustrated in Figure 1.5. As can be

seen, the outcomes are highly dependent on tiny variations in initial conditions due to the chaotic nature of trajectories in the system, resulting in seemingly random distribution of possible outcomes. Some large-scale structure is recognizable, e.g. the elliptical spiral pattern in top left, the diagonal streaks for longer pins in the r.h.s., and the higher blue (pin up) component in the l.h.s. of the main figure.

However, the entire system is completely deterministic, and indeed if we zoom in on any particular part we discover detailed structure, corresponding to similar types of motion. Note that for the final diagram (at 100,000x zoom) the width of the red diagonal band corresponds to a difference in length of about 40nm (well within the manufacturing specifications of a real drawing pin), while all other parameters remain exactly equal.

The successive zoomplots can be thought of as representing increasingly precise (specific) states of knowledge. Note that:

- different states of knowledge correspond to different values for the probability,
- no convergence to a 'true' probability for more detailed information

In short: probabilities are a relative measure that quantify uncertainty about an outcome w.r.t. a given state of knowledge. There is no objective or preferred state of knowledge, but *given* a state of knowledge the probability is completely and unambiguously defined. In real-world systems a small amount of uncertainty is already sufficient to produce and explain all probabilistic properties and behaviour, without the need to invoke 'magical' random or noise factors that do not satisfy the equations of motion.

Perhaps surprisingly, we find that the a priori probability $p(R = \text{`}pin\text{-}up\text{'})$ does not exist: only $p(R = \text{`}pin\text{-}up\text{'} \,|\, \mathcal{I})$ is meaningful in the sense that it corresponds to an exact, unambiguous value. For systems with intrinsic symmetries, such as fair coins and dice, the intuitive probabilities $p(\text{`}heads\text{'}) = 50\%$ and $p(\text{`}six\text{'}) = \frac{1}{6}$ correspond to an implicit state of knowledge $\mathcal{I}$ that is invariant w.r.t. the symmetry. But for most real-world systems we cannot rely on such degeneracies, and we should aim to quantify (make explicit) our assumptions / knowledge about the system.

For a more in-depth philosophical treatment of probabilities as relative measures the reader is referred to Jaynes [2003]. As a final remark we simply note that most, if not all real-world systems can be described (at least in principle) in terms of a dynamical system, and that our interpretation of 'random' as arising from uncertain, incomplete knowledge about the state of a system applies even down to the quantum level.

Causal relation as effective manipulation

In a real-world dynamical system we can compute exactly what happens to the outcome of a random variable V as a function of a parameter x for a given initial state $\mathbf{x}_0$. If it changes then parameter x allows a level of control over the outcome of V. Similar for (densities over) regions $\mathcal{M}_i$ of phase-space, corresponding to states

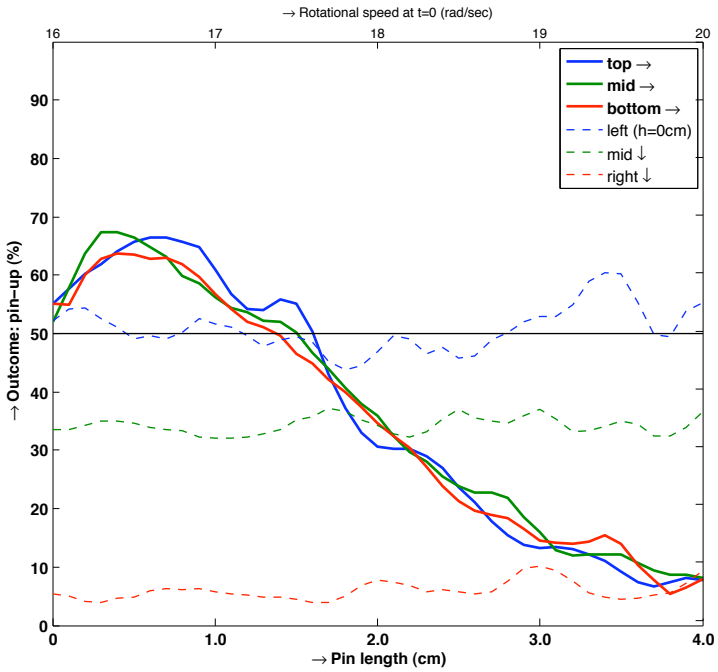

Figure 1.6: Causal relation as effective control (solid lines) vs. little to no control (dashed). Depicts gliding averages of outcome '*pin up*' as a function of **pin length** $h_{pin} = [0, 4]$cm. (solid lines) and initial **rotation** $\omega_0 = [16, 20]$rad/sec. (dashed). Solid {blue, green, red} correspond to horizontal cross sections along resp. {top, middle, bottom} in Figure 1.5, whereas the dashed lines correspond to vertical cuts across {left, mid, right}.

of knowledge $\mathcal{I}$ about the system: if the probability $p(V = v \,|\, \mathcal{I})$ in eq.(1.2) changes for states $\mathcal{I}'$ that *only* differ in $x \in \mathbf{x}$, then parameter x provides some level of deliberate control over V given $\mathcal{I}_{\backslash x}$.

For two initial states of knowledge, $\mathcal{I}$ and $\mathcal{I}' = \mathcal{I} + \Delta x$, that differ only in parameter x such that $\int \mathrm{d}x\, \rho'(\mathbf{x}) = \int \mathrm{d}x\, \rho(\mathbf{x})$, we can define the **causal effect** of x on outcome $V = v$ as:

$$c(x \Rightarrow v \,|\, \mathcal{I}, \Delta x) = \int_{\mathcal{M}_i} \mathrm{d}\mathbf{x}\, (\rho'(\mathbf{x}) - \rho(\mathbf{x}))\delta\big(v - V(\Phi^{t_V}(\mathbf{x})\big) \tag{1.3}$$

with $t_V > t_0$ and $\mathcal{M}_i$ the union of non-zero regions of $\rho(\mathbf{x}), \rho'(\mathbf{x})$. In words: the causal effect corresponds to the change in proportion (density) of outcomes $V = v$ given change Δx in state $\mathcal{I}$.[4] Substituting random variable X as the value of parameter x at t_0, eq.(1.3) describes deliberate control between random variables.

[4]Note that, similar to standard probabilistic interpretations, causality is defined in terms of co-changing probabilities, but the restriction $\int \mathrm{d}x\, \rho'(\mathbf{x}) = \int \mathrm{d}x\, \rho(\mathbf{x})$ automatically isolates causal influence from 'mere' correlation.

We could then say there exists a **causal relation** $X \Rightarrow V$ if there is *some* triple $\{\mathcal{I}, \Delta x, v\}$ such that $c(x \Rightarrow v \,|\, \mathcal{I}, \Delta x) \neq 0$. We can generalize to relations between arbitrary random variables, with careful interpretation of ΔX in case X depends on multiple parameters $\{x_i, x_j, ...\}$, and measures for the *strength* of the causal effect/relation by quantifying the difference in distribution over all possible outcomes V relative to the difference in distribution over X, etc., but we do not pursue this further here.

Instead we note that the definition via eq.(1.3) is still too broad to capture the desired 'effective manipulation' in general: the required change in x may not be feasible (too large or too specific), and/or the resulting change in $p(V = v|\mathcal{I})$ may be deemed insignificant (absolute or relative). In both cases, the level of control ceases to be *effective* in practice, and for all intents and purposes x is not different from parameters that provide no control.

Example 1.3. *Figure 1.6 illustrates* ***effective manipulation****: the solid lines depict gliding averages for outcome 'pin up' as a function of pin length h_{pin} around three initial rotation speeds ω_0. The proportion corresponds to probability $p(R =$ 'pin up'$|\mathcal{I})$ in eq.(1.2), where the state of information $\mathcal{I}$ is a uniform rectangular region $\mathcal{M}_i$ around (h_{pin}, ω_0) but all other parameters known exactly.*

The zoomplots in Figure 1.5 showed that outcome R is sensitive to tiny changes in both h_{pin} and ω_0. At the macroscopic scale in Figure 1.6, the proportion 'pin up' first increases with length to a maximum at $\Delta h_{pin} \approx 0.5cm.$, and subsequently decreases towards zero as the pin gets longer and longer, ***irrespective*** *of the value for ω_0 (or most other parameters). However, as $\Delta\omega_0$ is increased (dashed lines) the proportion fluctuates a bit, but remains essentially the same: even the 15% increase for $\omega_0 : 18.5 \rightarrow 19.5$ in very short pins (dashed blue) quickly cancels out with some uncertainty in other initial values, and no effective control remains. Only with extremely detailed knowledge and fine-grained control over the initial rotation does it become an effective means to influence the outcome.*

In short: in practice adjusting the pin-length h_{pin} is an effective means of influencing the probability on $R =$ 'pin up', whereas adjusting the initial rotation ω_0 is not. In terms of ***relevant*** *causal relations: $h_{pin} \Rightarrow R$, but not $\omega_0 \not\Rightarrow R$.*

Undesired, irrelevant causal relations can be excluded by limiting $\mathcal{I}$ and Δx to realistic scenarios and putting a threshold on the causal effect $|c(x \Rightarrow v \,|\, \mathcal{I}, \Delta x)| > \epsilon$. But this introduces a level of ambiguity in the sense that the 'right' network now depends on the (discretionary) chosen thresholds, and may also lead to a breakdown of transitivity. But in practice, with limited data weak causal relations are already very hard to distinguish from absent causal relations. Therefore, we interpret 'relevant/effective' as 'identifiable from the available information'. It means that we purposely **do not distinguish between very weak and absent causal relations**. This should allow for sufficiently sparse (= meaningful) causal models, without the need to introduce questionable a priori thresholds on effect size. It en-

sures that transitivity etc. remain satisified[5], while the discretionary choice changes to 'how reliable do you want the inferred relations to be'.

In terms of random variables X, Y in a dynamical system: Let $\mathcal{I}_X, \mathcal{I}'_X$ represent subsets of the available information $\mathcal{I}$ on the system that correspond to regions in state-space that *only* differ w.r.t. variable X. Then, in terms of the state-space probability from eq.(1.2), there is an **identifiable causal relation** $X \Rightarrow Y \,|\, \mathcal{I}$ if we can establish that:

$$X \Rightarrow Y \,|\, \mathcal{I} \;:\quad p(Y = y \,|\, \mathcal{I}_X) \neq p(Y = y \,|\, \mathcal{I}'_X) \tag{1.4}$$

In words: there is an identifiable causal relation $X \Rightarrow Y \,|\, \mathcal{I}$ if we can derive from $\mathcal{I}$ that the proportion of outcomes $Y = y$ varies with changes on/to X. Possibly also indirect via transitivity, if we already established that: $\exists Z : (X \Rightarrow Z \,|\, \mathcal{I}) \wedge (Z \Rightarrow Y \,|\, \mathcal{I})$. It is similar to the inferred effect of a $do(X)$-operator in Pearl [2000], where the difference between $\mathcal{I}_X$ and $\mathcal{I}'_X$ corresponds to a surgical intervention on X. *How* to establish the condition in eq.(1.4) from available information and data does not immediately follow from the definition: that is part of the subject of this thesis. Also, for finite data the state-space probabilities given $\mathcal{I}$ are approximated by, but not equivalent to the observed sample proportions, which introduces another level of ambiguity, see 5.1.

Identifiable **absence of a causal relation** $X \not\Rightarrow Y \,|\, \mathcal{I}$ follows from establishing that eq.(1.4) does not hold for any ΔX, which in general requires stronger assumptions (e.g. faithfulness, see 2.2). We also obtain the class of **not-identifiable causal relations**, $X \overset{?}{\Rightarrow} Y \,|\, \mathcal{I}$, where the available information is not (yet) sufficient to decide whether there is a causal relation or not. In section 2.3 we find a similar class of undecidable relations in the form of circle marks in the PAG.

In short, we incorporate the *effective* part of manipulation by assuming that **all identifiable causal relations are relevant**. Whether or not a relation $X \Rightarrow Y$ is identifiable depends on the available information $\mathcal{I}$, including data and assumptions. More information may lead to more causal relations, but can never alter relations already found: if a causal relation turns out not to be true, then (part of) the input information or assumptions must be false. When working with real, finite data sets only probabilities remain, and the best we can hope for is something like $p(X \Rightarrow Y \,|\, \mathcal{I})$; chapter 5 makes a few tentative steps towards that goal.

Conditional independence in state space

Having defined identifiable causal relations, sofar all we found is that correlation does not imply causation, but that we also cannot infer absence of such a relation from a lack of dependence. So the question arises: how *do* we find causal relations?

[5] Weak direct causal paths from large data sets can always be filtered out *after* the entire inference process has completed

What is missing is something that captures the structural flow of manipulation: do this and then that happens. Our intuition from the drawing pin example is that the answer lies in the aforementioned memorylessness property of dynamical systems, which says that states $\mathbf{x}(t > t')$ are independent of states $\mathbf{x}(t < t')$ given $\mathbf{x}(t')$. This would manifest itself in the form of two variables that become independent *conditional* on one or more others: these (actively) separating variables then all play a role in intercepting/mediating the causal effects that produce the original dependence. It is virtually impossible that such a conditional independence arises by accident, as it would require variables that vary together without any coordinating mechanism. Note that the conclusion only holds if *all* separating variables are needed for the independence, which will lead to our focus on *minimal* conditional independencies, see section 2.2.

We do not necessarily need to find/cover the value of all parameters (dimensions) of the separating state $\mathbf{x}(t')$: often only a few parameters appear in the equations of motion for another, e.g. for the drawing pin the time development of position only depends on the speed $\dot{\mathbf{x}} = \mathbf{v}$, which in turn depends on gravity, drag, etc. Which ones are directly or indirectly in effect at a given stage can change with different parts of state space, e.g. for the pin the translational and rotational components only interact at bounce/contact, but not during free flight. In such transitions from one active mechanism to another, just one or two variables can already screen off the effect of another parameter on a third.

Figure 1.7 illustrates how screening off/conditional independence shows up in state space mappings.[6] It depicts results from two different 'double pin throw' experiments, where in each case both a standard pin (1) and a long pin (2) are thrown the same way under an angle with the vertical, except that in one experiment ('sequential') the initial speed of pin 2 depends on the final horizontal distance X_1 achieved by pin 1, whereas in the other ('simultaneous') both initial speeds are the same. Plots depict the final horizontal distances $\Delta X_{1,2}$ as a function of initial speed v_0 and rotation ω_0 of pin 1 (and angle θ_0 in the r.h.s.), where all other parameters are kept constant.

Plot (b), top-left-middle, shows not surprisingly that the horizontal distance ΔX_1 of pin 1 tends to be larger (more reddish) for higher launch speeds v_0. For slightly different initial angles θ_0, drag, friction, etc. the plot would be similar but differ in detail (position of the bands). For 'realistic' experiments this would average out towards a roughly homogenous orange-cyan gradient from top to bottom, corresponding to an identifiable *dependence* between distance and speed, denoted $X \not\perp\!\!\!\perp v_0$, but not (hardly) between distance and rotation: $X \perp\!\!\!\perp \omega_0$. The dependence between distance and speed also holds for the longer pin 2 in the 'simultaneous' experiment in plot (a). To a somewhat lesser degree it also holds for X_2 in the

[6] Note that in real data from experiments the conditional independences are found through appropriate statistical tests, e.g. a Chi-squared test for discrete variables or the kernel-based KCI-test [Zhang *et al.*, 2011] for continuous variables.

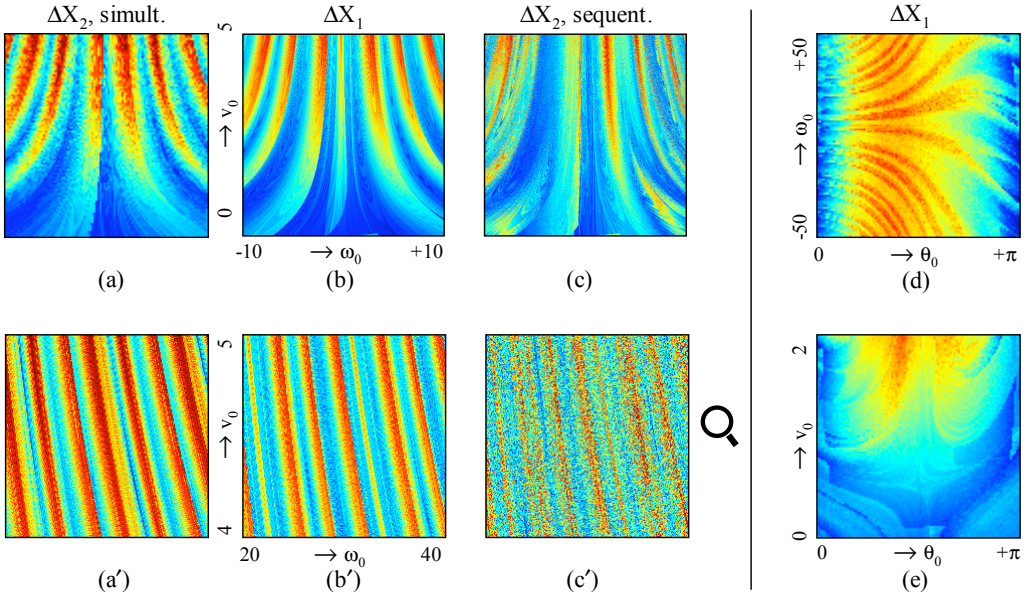

Figure 1.7: Conditional independence in results for two 'double pin throw' experiments, each with two pins with lengths $h_1 = 8$mm., resp. $h_2 = 24$mm., both launched under a slight angle $\theta_0 = 0.25$ with the vertical, and either otherwise identical initial conditions ('simultaneous'), or with initial speed of pin 2 depending on ΔX_1: the final horizontal distance of pin 1 ('sequential'), ranging from about -20cm (deep-dark blue) to $+80$cm (dark red). Plots (a)-(c): horizontal distances $\Delta X_{1,2}$ as function of initial speed v_0 and rotation ω_0 of pin 1; (a′)-(c′): idem, close-ups at higher rotation; (d): ΔX_1 as a function of initial angle θ_0 and rotation ω_0; (e): idem, vs. θ_0 and v_0; see main text for details

sequential version in plot (c): more red/less blue in the top part. Furthermore, if X_1 is large (red) then in both versions of the 'double pin' experiment X_2 on average tends to be larger as well, and so v_0, X_1, X_2 are all dependent.

However, in the close-up plot (a′) for the simultaneous experiment we can see that knowing the outcome X_1, corresponding to the color of the equivalent pixel in plot (b′), gives little to no extra information on the value of X_2 once v_0 is known: the bands in plots (a′) and (b′) look qualitatively the same but appear in unrelated locations, and so *conditional* on v_0 variables X_1 and X_2 become independent, denoted $X_2 \perp\!\!\!\perp X_1 \,|\, v_0$. In contrast, for the sequential version in plot (c′) we see that even though there is some additional 'noise' the bands with high-value X_2 coincide exactly with the high-value bands for X_1 in plot (b′). Therefore in the sequential experiment: $X_2 \perp\!\!\!\perp v_0 \,|\, X_1$.

Finally, from Figure 1.7(e) we see that X_1 varies with both the speed v_0 and the launch angle θ_0 (with lowest values for straight up/down, and highest value for slightly up and to the right, see also (d)). But we also see that *given* the distance

X_1, variables θ_0 and v_0 become dependent: if X_1 is small (blueish) and v_0 is high, then θ_0 is likely close to vertical. In other words, in both experiments: $v_0 \not\perp\!\!\!\perp \theta_0 \,|\, X_1$.

Causal information from conditional independence

So what does finding a variable that separates two others tell us about causal effects? For that we need to know how screening off is accomplished in the flow through high-dimensional state space: rather hard to visualize, but ultimately leading to two simple conclusions. Remember that system parameters (incl. time) become dimensions, and that random variables become functions on surfaces that intercept trajectories from initial states, ranging from flat hyperplanes across the time dimension to complex, even discontinuous manifolds. Data from an experiment corresponds to sampling from a distribution over a region of initial states mapped by the evolution rule to values for random variables at the corresponding hypersurfaces.

Consider a dynamical system described by state $\mathbf{q} = \{x, y, z, ..\}$, with three defined random variables $\{X, Y, Z\}$ that correspond to resp. the value of parameters $\{x, y, z\}$ at times $\{t_X, t_Y, t_Z\}$ in state space. Suppose we find that X and Y are (probabilistically) dependent, $X \not\perp\!\!\!\perp Y$, but independent given Z, so $X \perp\!\!\!\perp Y \,|\, Z$. The dependence means that, whatever the distribution over initial region $\mathcal{M}_0$, there must exist at least two initial states $\mathbf{q}_0$ and $\mathbf{q}'_0$ that differ in both parameter x at t_X and in parameter y at t_Y along their trajectories through state space (and so 'vary together in X and Y'). These two initial states then *must* also differ in parameter z at t_Z, because the independence $X \perp\!\!\!\perp Y \,|\, Z$ implies that two trajectories with the *same* value for Z cannot differ in both X and Y. We now look at the possible configurations in which two initial states can be mapped through state space to produce the required combination of differences in parameters.

Assume that initial states $\mathbf{q}'_0 = \mathbf{q}_0 + \Delta\mathbf{q}$ at t_0 are mapped by the evolution rule to states $\mathbf{q}'_X = \mathbf{q}_X + \{\Delta x, ..\}$ at t_X, to $\mathbf{q}'_Y = \mathbf{q}_Y + \{., \Delta y, ..\}$ at t_Y, and to $\mathbf{q}'_Z = \mathbf{q}_Z + \{., ., \Delta z, ..\}$ at t_Z, where $\{\Delta x, ..\}$ represents a change in parameter x, and possibly changes in one or more other parameters as well. Without loss of generality we assume (for now) that $t_0 < t_X \leq t_Y$, i.e. variable X obtains its value at or before Y is reached. From the dependence, if $t_X < t_Y$ then $\mathbf{q}_X$ is mapped to $\mathbf{q}_Y$, and so the difference $\{\Delta x, ..\}$ is mapped to $\{., \Delta y, ..\}$. By eq.(1.3), that implies there is either a nonzero causal effect $x \Rightarrow y$ in the system, or a nonzero causal effect from one of the other differing parameters at t_X to y, or both. In case $t_X = t_Y$ then there are no causal effects between X and Y.

For the separating variable Z there are now three possible cases:

(1) $t_X < t_Z < t_Y$: the difference $\{\Delta x, ..\}$ at t_X is mapped to $\{., ., \Delta z, ..\}$ at t_Z, which in turn is mapped to $\{., \Delta y, ..\}$ at t_Y. If there are changes in other parameters at t_Z that have a causal effect on y at t_Y, then in general there can be states in the system that differ in both X and Y for a given value of Z.

Therefore, the only remaining option is that the difference Δz is responsible for (mapped to) Δy, which by definiton implies a **causal effect** $z \Rightarrow y$.

(2) $t_0 < t_Z \leq t_X$: if there were any causal effect $x \Rightarrow y$ then in general for two initial states mapped to the same value for Z there can still be a difference Δx (through variation in other parameters at t_Z) which gets mapped to a difference Δy, contrary the conditional independence. But if there is no causal effect $x \nRightarrow y$, then a difference in some other parameter at t_X must be responsible for (mapped to) the variation Δy. Then like before, if there are changes in parameters other than Δz at t_Z that together have a causal effect on both x at t_X *and* on y at t_Y, then X and Y can vary together for fixed Z, contrary the given. The only remaining option is that at least one of the differences Δx or Δy must originate from Δz instead. By definition this implies a **causal effect** $z \Rightarrow x$ **or** $z \Rightarrow y$ **(or both)**.

(3) $t_Y \leq t_Z$: if different changes $\{., \Delta y, ..\}$ can lead to the same value for Δz, then (small) changes Δy can be compensated for by changes in other parameters at t_Y to give $\Delta z = 0$, leaving X and Y varying together for fixed Z, *contrary* the given $X \perp\!\!\!\perp Y \mid Z$, so this case cannot occur.

In short: if z intercepts all simultaneous variation in x and y, then whatever the exact configuration, there is a causal effect $z \Rightarrow x$ or $z \Rightarrow y$ or both. We assumed $t_X \leq t_Y$, but this conclusion also holds true if $t_Y \leq t_X$. The argument can be extended to random variables that are functions of multiple parameters on complex, possibly intersecting manifolds. This suggests the conclusion that **conditional independence implies at least one causal relation** applies generally:

$$(X \not\perp\!\!\!\perp Y) \wedge (X \perp\!\!\!\perp Y \mid Z) \quad \text{implies} \quad (Z \Rightarrow X) \vee (Z \Rightarrow Y) \tag{1.5}$$

The caveat mentioned in case (3) applies to the other cases as well: essentially it assumes that there is no deterministic, one-to-one functional relationship $X = f(Z)$ between random variables, see also [Zhang and Spirtes, 2008].

We can do a similar analysis for the case where X and Y are independent but have a shared dependence with a third variable Z, such that $X \perp\!\!\!\perp Y$, $X \not\perp\!\!\!\perp Z$, and $Y \not\perp\!\!\!\perp Z$. Analogous to the previous argument: if any difference Δz at t_Z would be mapped by the evolution rule to a difference in Δx at t_X or Δy at t_Y, then X and Y would vary together, *contrary* $X \perp\!\!\!\perp Y$, unless there was some other contribution that just happened to cancel out the effect of Δz exactly. This suggests case (3) (above) as the only feasible configuration, which implies **no causal effect** from z on either x or y. This configuration is described succinctly as:

$$(X \perp\!\!\!\perp Y) \wedge (X \not\perp\!\!\!\perp Y \mid Z) \quad \text{implies} \quad (Z \nRightarrow X) \wedge (Z \nRightarrow Y). \tag{1.6}$$

The 'no accidental cancellation' clause returns as the faithfulness assumption in section 2.3. The two rules can be combined to infer new causal information, e.g. by

using (1.6) to eliminate one of the options in (1.5) to obtain a definite causal relation.

▶ This is the main idea behind the approach developed in this thesis: in real dynamical systems, certain local independence patterns signify presence or absence of causal relations, *irrespective* of what we know or do not know about the rest of the system.

This idea is fundamentally different from many existing approaches to discovering causal relations: instead of working from a **global criterium** (minimality) for a network of relations over variables and using that to infer that certain relations are causal, we look for **local patterns** that signify causal information and combine these to build up an equivalent global picture. In chapter 3 we will provide a more formal proof, based on graphical model theory. The end result is a method that works well in practice: it competes with current state-of-the-art techniques, but is also remarkably straightforward and flexible, as demonstrated in the subsequent applications in chapters 4 and 5. It suggests that this bottom-up approach to causality, inspired by starting from an underlying dynamical system, is somehow more 'natural' than the more conventional top-down approach.

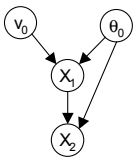

Figure 1.8: Graphical causal model for sequential two-pin system

With a combination of conditional in/dependencies we can start to build up a causal model. Suppose that we are presented with a data set from the sequential 'double pin' experiment from Figure 1.7 in which random variables $\{v_0, \theta_0, X_1, X_2\}$ are measured for a reasonable number of throws, with initial parameters $\{v_0, \omega_0, \theta_0\}$ chosen independent at random. Clearly, from this we cannot infer anything on the detailed bounce behaviour of the pins. But we are able to find that $v_0 \perp\!\!\!\perp \theta_0$, $v_0 \not\perp\!\!\!\perp \theta_0 \mid X_1$, and $X_2 \perp\!\!\!\perp v_0 \mid X_1, \theta_0$, which is sufficient to construct the intuitive causal model in Figure 1.8. From this we can read there is a causal relation $X_1 \Rightarrow X_2$ (which is indeed true in the sequential experiment), but definitely no causal relations from X_1 or X_2 to v_0 or θ_0; see chapter 2 for details.

▶ From this point on we leave dynamical systems, differential equations, and state-space diagrams behind and work directly from the graphical causal model representation inferred from data/probabilistic independencies.

However, we keep the idea inspired by memorylessness that local independence

patterns imply presence or absence of causal relations. We assume that all identifiable causal relations are relevant for a model, and that we can ignore pathological cases, e.g. no ambiguities through deliberately ill-defined random variables etc.

1.3 Outline of the thesis

After the previous 'zooming in' on causal relations, where even a toy dynamical system already leads to all kinds of technical intricacies and ambiguities, reasonably inferring such relations for large real-world systems may seem like a daunting if not impossible task. Fortunately, once cast in graphical model form 'effective manipulation' is an intuitively clear and straightforward concept, and most complications only arise from highly artificial cases that have no bearing on the large majority of causal relations we are interested in.

Therefore, in the rest of this thesis we work directly from observed probabilities generated from a clear-cut causal model over well-defined random variables. We focus on probabilistic independencies and background knowledge as the source for causal information, ignoring other (distributional) properties, such as non-linear relations and non-Gaussianity. Only when apparent contradictions arise, or when trying to answer fundamental questions like 'What actually *is* the best causal model/correct prior/true probability?', or 'How realistic is an acyclic model?' should we need to go back to the ground-truth level of relevant relations in a real-world dynamical system.

Chapter 2 starts with an overview of basic concept and definitions from standard graphical model theory, and may be skipped by readers familiar with this topic. It focusses on the class of *mixed graphical models* that are most suited to represent the relation between causal models and probabilistic in/dependencies. It defines the causal DAG as the basic underlying model and introduces a number of key assumptions such as the causal Markov condition and faithfulness. For reference purposes the last part of this chapter details the Fast Causal Inference (FCI) algorithm as the gold-standard in constraint-based causal discovery.

In **Chapter 3** the main idea from our excursion into dynamical system territory in section 1.2 is developed: it shows that causal discovery can take the form of straightforward deduction on a set of logical statements about causal relations that are directly derived from in/dependence patterns observed in the data. It belongs to the category of constraint-based approaches to causal discovery, but differs from existing methods in that it splits up causal inference in a series of modular steps that can be executed in arbitrary order. As a result, causal discovery becomes a very flexible and intuitive process, with promising applications and extensions, illustrated in the next two chapters. We first show that the 7 FCI rules to find all identifiable absent causal relations $X \nRightarrow Y$, are instances of just two underlying patterns. We extend this result to find three rules that translate inferred probabilistic in/dependencies directly into logical statements about presence or absence

of causal relations. We then show that straightforward logical deduction on these statements is sufficient to find all invariant features in the underlying causal model, corresponding to all identifable presence and/or absence of direct causal relations. The proof for completeness relies on a mapping from all graphical orientation rules in FCI to instances of our three logical rules. The method is implemented in the anytime LoCI-algorithm and evaluated on behaviour and performance.

Chapter 4 deals with the problem of how to infer more causal information from the *combination* of different experiments than just the sum of causal relations from each separately. The two main problems faced in this task are that of partially overlapping variables, i.e. not all variables are always measured, and that due to different experimental circumstances or interventions some dependencies are present in one but not in another experiment. Pooling the data would be similar to learning from incomplete data with values not missing at random, but existing techniques are ill-equipped to handle data from different experiments. Here we show how the logical framework from chapter 3 can be used to create a larger, more informative model provided that no selection bias is present. We implement the approach into the Multiple model Causal Inference (MCI) algorithm, and test against two reference methods to assess its efficacy.

Finally, **Chapter 5** tackles the problem of robustness and reliability of causal models inferred from data. Sometimes causal discovery algorithms confidently assert relations that researchers know to be wrong, or produce very different conclusions for slightly different data sets, which tends to make the entire output suspect. We introduce a Bayesian approach to constraint-based causal discovery that obtains reliability estimates for the individual logical causal statements employed in the LoCI algorithm. We score the Bayesian likelihood for possible DAG patterns over (small) subsets of variables which, together with an appropriate prior, gives a posterior probability for each pattern. Adding these probabilities for all patterns that imply a logical causal statement then provides the reliability estimate. The statements are then sorted and processed in decreasing order of reliability. One complicating aspect is that the underlying causal relations between variables in these subsets do not necessarily match a DAG: we derive some rules to ensure that even then only valid causal statements are inferred. The resulting Bayesian Constraint-based Causal Discovery (BCCD) algorithm outperforms all current state-of-the-art algorithms in terms of the accuracy of inferred causal relations. On top of that it also provides a reasonable estimate for the reliability of each causal conclusion in the output, which should help to improve confidence in the overall model as well.

Chapter 2

Graphical Models and Causal Discovery

This chapter introduces basic concepts and definitions from graphical model theory, with focus on the relation between causal models and probabilistic independence. For more details the reader is referred to Koller and Friedman [2009]; Neapolitan [2004]; Pearl [2000]; Spirtes et al. [2000].

Throughout this thesis we use upper-case letters X, V_i, etc. to denote single nodes or random variables, and matching lower-case letters x, v_i for states or values of those variables. Similarly we use boldface capitals $\mathbf{X}$, $\mathbf{V}$, etc. to denote *sets* of variables, and corresponding blod lower-case letters $\mathbf{x}$ to denote values for each variable in the set. Calligraphic $\mathcal{G}$, $\mathcal{B}$, $\mathcal{M}$ are used to denote structures or models.

2.1 Mixed graphical models

Definition 2.1. *A **graph** $\mathcal{G}$ is an ordered pair $(\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is a non-empty, finite set of vertices or nodes, and $\mathbf{E}$ is a set of edges between pairs of nodes $(X, Y) : X, Y \in \mathbf{V}$.*

A **graphical model** is a graph where nodes represent variables, and edges represent direct interactions or relations between variables. A **directed graph** $\mathcal{G}$ contains only edges (arcs) of the form $X \to Y$, where the end marks at the nodes are known as tails '$-$' and arrowheads '$>$'. We use '$*$' to indicate an arbitrary edge mark. A **mixed graph** $\mathcal{M}$ is a graph that can contain more than one type of edge between different pairs of nodes, e.g. directed $\to$ and bi-directed $\leftrightarrow$ edges.

In a graph $\mathcal{G}$, a **path** $\pi = \langle X_1, \ldots, X_n \rangle$ is a sequence of distinct nodes such that for $1 \leq i \leq n-1$, X_i and X_{i+1} are **adjacent** (connected by an edge) in $\mathcal{G}$. We

sometimes use $X \not\sim Y$ to indicate **non-adjacency** (absence of an edge between X and Y). An edge or path is **into** (or **out of**) a node X if it has an arrowhead (or tail) at X. If $X \to Y$ in $\mathcal{G}$, then X is a **parent** of Y, and Y is a **child** of X; if $X \leftrightarrow Y$ is in $\mathcal{G}$, then X and Y are called a **spouse** of the other; if $X - Y$, then they are **neighbours**. The **skeleton** of a graph $\mathcal{G}$ is the set of adjacencies in $\mathcal{G}$ (all edges between nodes without distinguishing end marks).

A **directed path** from X_1 to X_n is a path along which each node X_i is a parent of its successor X_{i+1} in $\mathcal{G}$. A node X is **ancestor** of Y (and Y is a **descendant** of X), if there is a directed path from X to Y in $\mathcal{G}$, or if $X = Y$. A **(directed) cycle** is a (directed) path from a node back to itself. In a mixed graph $\mathcal{M}$ there is an **almost directed cycle** if there is a directed path from some node X to some node Y, and $X \leftrightarrow Y$ is also in $\mathcal{M}$. A **directed acyclic graph (DAG)** $\mathcal{G}$ is a graph containing only arcs $\to$ as edges, but without any directed cycles.

A vertex Z is a **collider** on a path $\pi = \langle \ldots, X, Z, Y, \ldots \rangle$ if π contains the subpath $X *\!\!\to Z \leftarrow\!\!* Y$, i.e. if both edges from X and Y are into Z; otherwise Z is a **non-collider**. A **trek** is a path that does not contain any collider. A (sub)path over a triple $\langle X, Z, Y \rangle$ is **unshielded** if X and Y are not adjacent in $\mathcal{G}$. An unshielded collider $X \to Z \leftarrow Y$ is also known as a ***v*-structure**. A path $\pi = \langle X_1, \ldots, X_n \rangle$ is **uncovered** if each successive triple along π is unshielded.

For disjoint (sets of) nodes X, Y, and $\mathbf{Z}$ in a DAG $\mathcal{G}$, X is ***d*-connected** to Y given $\mathbf{Z}$, denoted $X \not\perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z}$, iff there exists an **unblocked path** $\pi = \langle X, \ldots, Y \rangle$ in $\mathcal{G}$ on which every collider is ancestor of some $Z \in \mathbf{Z}$ and every non-collider is not in $\mathbf{Z}$. If not, then all such paths are **blocked**, and X is said to be ***d*-separated** from Y given $\mathbf{Z}$ in the graph $\mathcal{G}$, denoted $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z}$. When applied to a *mixed graph*, the notion of d-separation is known as ***m*-separation**, sometimes explicitly indicated as $X \perp\!\!\!\perp_M Y \,|\, \mathbf{Z}$.

Definition 2.2. *An **ancestral graph** is a mixed graph containing directed ($\to$), bi-directed ($\leftrightarrow$), and/or undirected ($-$) edges, and in which:*

- *there is no (almost) directed cycle,*
- *there are no arrowheads at nodes on undirected edges.*

The graph is called *ancestral* because arrowheads on an edge signify 'non-ancestorship'. A ***maximal* ancestral graph (MAG)** $\mathcal{M}$ is an ancestral graph in which any two non-adjacent vertices (X, Y) in $\mathcal{M}$ can be m-separated by some set $\mathbf{Z}$. Note that DAGs are just a special subclass of MAGs.

Two MAGs (or DAGSs) are **Markov equivalent** if they imply the same set of m-separations. The **(Markov) equivalence class** of a graph $\mathcal{M}$, denoted $[\mathcal{M}]$, is the set of all graphs that are Markov equivalent to $\mathcal{M}$. In a MAG $\mathcal{M}$, a path $\pi = \langle X, \ldots, W, Z, Y \rangle$ is a **discriminating path** for a disciminated node Z, if X is not adjacent to Y, and every node between X and Y is both a collider along π and a parent of Y.

With this the equivalence class can be characterized as:

Proposition 2.3. *Two MAGs are Markov equivalent if and only if:*

- *they have the same skeleton,*
- *they have the same v-structures, and*
- *they have the same discriminated nodes with the same non/collider property along corresponding discriminating paths.*

Proof. See Richardson and Spirtes [2002]. □

For DAGs equivalence reduces to the first two items, i.e. the same skeleton and v-structures. For MAGs/DAGs the corresponding equivalence class can be represented intuitively by a mixed graph with circles '$\circ$' – signifying 'undecided' – as third type of edge mark, resulting in six possible edges.

Definition 2.4. *A **partial ancestral graph (PAG)** $\mathcal{P}$ is a mixed graph that represents the equivalence class $[\mathcal{M}]$ of a MAG $\mathcal{M}$, such that:*

- *it has the same skeleton as $\mathcal{M}$,*
- *all non-circle edge marks (tails and arrowheads) in $\mathcal{P}$ correspond to invariant edge marks in $[\mathcal{M}]$.*

The PAG is **complete** (cPAG) if all circle marks in $\mathcal{P}$ correspond to variant (non-invariant) edge marks in $[\mathcal{M}]$. A cPAG is **maximally informative** in the sense that it explicitly captures *all* invariant features of the equivalence class of $\mathcal{M}$ (shared by all members of $[\mathcal{M}]$).
For representation purposes we also introduce a generalized version:

Definition 2.5. *A **causal PAG** $\mathcal{P}$ is an ancestral graph, possibly with circle marks '$\circ$' on edges to denote 'unknown tail or arrowhead', that represents all known causal information on a MAG $\mathcal{M}$, such that:*

- *all edges in $\mathcal{M}$ are also in $\mathcal{P}$,*
- *all implied ancestral relations in $\mathcal{P}$ correspond to ancestral relations in $\mathcal{M}$.*

If the skeletons of $\mathcal{M}$ and $\mathcal{P}$ are the same then all tails and arrowheads on edges in $\mathcal{P}$ are also in $\mathcal{M}$. As a result, any (complete) PAG is also a causal PAG, but not the other way around as a causal PAG may contain more or less information than necessary to specify the equivalence class, e.g. from background information or external interventions. A fully connected $\circ\!\!-\!\!\circ$ graph is a causal PAG $\mathcal{P}$ representing no known causal information at all.

Finally, in a PAG $\mathcal{P}$ a path $\pi = \langle X_1, \ldots, X_n \rangle$ is said to be a **possibly directed (p.d.) path**, if no edge $X_i * \!\!-\!\!* X_{i+1}$ has an arrowhead at X_i or a tail at X_{i+1}. If each successive triple along π is also unshielded it is an **uncovered possibly directed (u.p.d.) path**. If all edges on a u.p.d. path are of the form $\circ\!\!-\!\!\circ$, then the path is called an **uncovered circle (u.c.) path**.

2.2 Graphical models and probabilistic independence

In a graphical model $\mathcal{G}$, we assume that the set of random variables $\mathbf{V}$ can be partitioned into three subsets: $\mathbf{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$, in which $\mathbf{O}$ represents the subset of *observed* variables, $\mathbf{L}$ the set of hidden or *latent* variables, and $\mathbf{S}$ the set of *selection* variables that determine inclusion (selection) into a data set $\mathbf{D}$. If the distinction is not made explicitly we simply assume $\mathbf{V} = \mathbf{O}$.

Definition 2.6. *A **Bayesian network** (BN) is a pair $\mathcal{B} = (\mathcal{G}, \mathbf{\Theta})$, where $\mathcal{G} = (\mathbf{V}, \mathbf{A})$ is DAG over $\mathbf{V}$, and the parameters $\boldsymbol{\theta}_V \subset \mathbf{\Theta}$ represent the conditional probability of variable $V \in \mathbf{V}$ given its parents $\mathbf{Pa}(V)$ in the graph $\mathcal{G}$.*

A joint probability distribution $p(\mathbf{V})$ over random variables $\mathbf{V}$ factors according to a Bayesian network:

$$p(X_1 = x1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i \mid \mathbf{Pa}(X_i) = \mathbf{pa}_i, \boldsymbol{\theta}_i) \tag{2.1}$$

We use $X \perp\!\!\!\perp_p Y \mid \mathbf{Z}$ to denote that two variables X, Y are **probabilistically independent** given (or conditional on) a (possibly empty) subset $\mathbf{Z}$. Idem, we use $X \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}$ to denote that X and Y are probabilistically **dependent** given $\mathbf{Z}$.

An important concept is that of a ***minimal* conditional in/dependence**, capturing the notion that all variables in the set indicated by square brackets are needed to make or break an independence:

- $X \perp\!\!\!\perp_p Y \mid \mathbf{W} \cup [\mathbf{Z}] \equiv \forall \mathbf{Z}' \subsetneq \mathbf{Z} \colon X \not\perp\!\!\!\perp_p Y \mid \mathbf{W} \cup \mathbf{Z}'$,
- $X \not\perp\!\!\!\perp_p Y \mid \mathbf{W} \cup [\mathbf{Z}] \equiv \forall \mathbf{Z}' \subsetneq \mathbf{Z} \colon X \perp\!\!\!\perp_p Y \mid \mathbf{W} \cup \mathbf{Z}'$.

The following property links the structure of a Bayesian network to a set of probabilistic independence relations:

Definition 2.7. *A probability distribution $p(\mathbf{V})$ satisfies the **Markov condition** w.r.t. a DAG $\mathcal{G}$, iff each variable $V \in \mathbf{V}$ is (probabilistically) independent of its non-descendants, given its parents in $\mathcal{G}$.*

A probability distribution that factorizes according to (2.1), e.g. a Bayesian network, satisfies the Markov condition, cf. Pearl [2000]. To complete the connection between the structure of a Bayesian network and implied probabilistic independencies we define the following property

Definition 2.8. *A probability distribution $p(\mathbf{V})$ is **faithful** w.r.t. a DAG $\mathcal{G}$, iff there are no (probabilistic) independencies between variables in $\mathbf{V}$ that are not entailed by the Markov condition applied to $\mathcal{G}$.*

We say that a distribution $p(\mathbf{V})$ is *faithful* if there exists a DAG to which it is faithful. Similarly, we use shorthand 'a faithful DAG $\mathcal{G}$' to indicate that it is faithful w.r.t. some implied distribution.

Assuming the Markov condition and faithfulness hold, then graphical d-separation and probabilistic independence become equivalent:

Theorem 2.9. *Let X, Y, and $\mathbf{Z}$ be disjoint (subsets of) variables in a Bayesian network with faithful DAG $\mathcal{G}$, then:*

$$(X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Z}) \Leftrightarrow (X \perp\!\!\!\perp_{p} Y \mid \mathbf{Z})$$

Proof. See, e.g. Pearl [1988]. □

If we only consider the *marginal* distribution over a subset of variables $\mathbf{V}' \subset \mathbf{V}$ from a faithful Bayesian network $\mathcal{B}$, then the resulting distribution $p(\mathbf{V}')$ may no longer correspond to a faithful DAG $\mathcal{G}'$ over $\mathbf{V}'$. In that case the independence relations between the observed variables can be represented in the form of a *maximal ancestral graph* (MAG), see Definition 2.2. MAGs form an extension of the class of DAGs that is closed under marginalization and conditioning.

We say that a MAG $\mathcal{M}$ is *faithful* to a distribution over $p(\mathbf{V}')$ if they can both be obtained from a faithful DAG $\mathcal{G}$, resp. distribution $p(\mathbf{V})$, through marginalization and/or conditioning. With this the ancestral analog of Theorem 2.9 becomes:

Proposition 2.10. *Let X, Y, and $\mathbf{Z}$ be disjoint (subsets of) variables in a faithful MAG $\mathcal{M}$, then:*

$$(X \perp\!\!\!\perp_{M} Y \mid \mathbf{Z}) \Leftrightarrow (X \perp\!\!\!\perp_{p} Y \mid \mathbf{Z})$$

Proof. See Richardson and Spirtes [2002]. □

2.3 Causal models and ancestral graphs

This section introduces the model for the causal relations we try to find, and specifies the assumptions we are willing to/have to make in order to 'get it to work'.

A popular and intuitive way of representing a causal system is in the form of a **causal DAG** $\mathcal{G}_C$, a graphical model where the arrows represent direct causal interactions between the variables in the system [Zhang, 2008; Pearl, 2000].

Definition 2.11. *In a causal DAG $\mathcal{G}_C$ there is a **causal relation** $X \Rightarrow Y$, iff there is a directed path from X to Y in $\mathcal{G}_C$. Absence of such a path is denoted $X \not\Rightarrow Y$.*

With this definition causal relations are **transitive** $(X \Rightarrow Y) \wedge (Y \Rightarrow Z) \vdash (X \Rightarrow Z)$, **irreflexive** $(X \not\Rightarrow X)$, and **acyclic** $(X \Rightarrow Y) \vdash (Y \not\Rightarrow X)$. Every edge $X \Rightarrow Y$ in a causal DAG $\mathcal{G}_C$ represents a **direct** causal relation. A relation $X \Rightarrow Y$ that is mediated by other nodes in the network is an **indirect** causal relation.

In order to link observed probabilistic in/dependencies to causal relation in the underlying system, we make the following assumptions:

- the **Causal DAG** assumption implies that the systems we consider in this thesis can be described by *some* underlying causal DAG $\mathcal{G}_C$,

- the **Causal Markov** assumption implies that each variable in a causal DAG $\mathcal{G}_C$ is (probabilistically) independent of its non-descendants, given its parents,
- the **Causal Faithfulness** assumption states that there are no independencies between variables that are not entailed by the Causal Markov assumption.

Note: we do *not* rely on **causal sufficiency** in this thesis, that is we do not want to make the simplifying assumption that there are no unobserved common causes between variables in the system.

Under these assumptions the probabilistic independence relations between observed variables are described by a (faithful) maximal ancestral graph (MAG) $\mathcal{M}_C$. All MAGs that can encode the same set of independence relations through Proposition 2.10 are indistinguishable in terms of (Markov) implied independencies.[1]. Invariant features in the corresponding equivalence class $[\mathcal{M}_C]$ can then be represented in the form of a complete *partial ancestral graph* (PAG) $\mathcal{P}$, see Definition 2.4. These invariant features contain all valid causal information that is identifiable from observed independence constraints, [Zhang, 2008].

Reading the PAG

A few pointers on how to read causal/independence relations from a PAG $\mathcal{P}$:

- probabilistic in/dependence follows from standard m-separation,
- a definite *causal* relation $X \Rightarrow Y$ if there is a directed edge/path from X to Y *and* $U \ast\!\!\rightarrow X$ in $\mathcal{P}$ for some node U (to exclude selection bias),
- identifiable *absence* $X \not\Rightarrow Y$ iff there is no p.d. path from X to Y in $\mathcal{P}$.

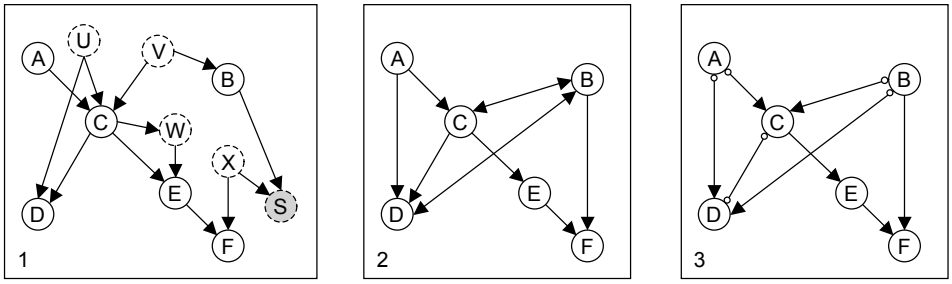

Figure 2.1: 1) Causal DAG $\mathcal{G}_C$ with hidden variables; 2) MAG over observed nodes; 3) complete PAG

Example 2. Figure 2.1 illustrates the relation between: (1) the (unknown) underlying causal DAG, (2) the MAG of ancestral relations between observed variables, and (3) the corresponding PAG of all causal information from in/dependencies.

[1] We follow the standard assumption that Markov equivalence implies statistical equivalence [Spirtes, 2010].

Figure (1) shows a causal DAG $\mathcal{G}_C$ over six observed nodes $\{A, B, C, D, E, F\}$, with a number of hidden variables including selection node S. There is a direct causal link $A \Rightarrow C$ and indirect causal link $A \Rightarrow F$, while there is no causal relation $B \not\Rightarrow C$. In (2) direct causal dependencies have become arcs in the MAG, but there are also some new edges: the bi-directed link $C \leftrightarrow B$ signals the presence of confounder V, and the arc $A \rightarrow D$ appears because conditioning on C unblocks the path $\langle A, C, U, D \rangle$. The PAG equivalence class in (3) shows all invariant marks: it tells us that D is definitely not an ancestor of B, so $D \not\Rightarrow B$, but that we cannot be sure about the reverse, nor about the causal relation between D and C. We *can* infer $C \Rightarrow E \Rightarrow F$, but *not* $B \Rightarrow F$, as there is no invariant arrowhead at B in (3).

2.4 Constraint-based Causal Discovery

From the previous section, our goal in causal discovery is to reconstruct as much as possible of the complete PAG from observed independence relations. A large class of causal discovery algorithms, known as **constraint-based** methods, is based directly on the Markov and faithfulness assumptions: iff an independence $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ can be found for *any* set of variables $\mathbf{Z}$ then there is no direct causal relation between X and Y in the underlying causal graph $\mathcal{G}_C$, and hence no edge between X and Y in the equivalence class $\mathcal{P}$.

Members of this group include the IC-algorithm [Pearl and Verma, 1991], PC [Spirtes *et al.*, 2000], Grow-Shrink [Margaritis and Thrun, 1999], TC [Pellet and Elisseef, 2008], and many others. All involve repeated independence tests in the adjacency-search phase to uncover the skeleton of $\mathcal{P}$, followed by an orientation-phase using rules such as in [Meek, 1995] to find invariant tails and arrowheads. The differences lie mainly in the search strategy employed, size of the conditioning sets, and additional assumptions imposed.

Of these, the **Fast Causal Inference (FCI)** algorithm [Spirtes *et al.*, 2000] in conjunction with the additional orientation rules in [Zhang, 2008] was the first to be shown sound and complete in the large-sample limit, even when hidden common causes and/or selection bias may be present. As such, it has become the de facto gold standard for constraint-based causal inference. For future reference we provide a brief description of the elements that are most important to our work below.[2]

Loosely speaking, the augmented FCI algorithm consists of an ingenious adjacency search based on conditional independencies (details of which will not concern us here) to find the skeleton of the PAG $\mathcal{P}$, followed by an orientation phase based on a the graphical rules detailed in Table 1, where we follow the numbering from [Zhang, 2008]. A graphical depiction of the rules is shown in Figures 2.2 and 2.3. Inspection reveals a certain hierarchy in which rules can trigger which others, reflected in the basic structure of FCI in Algorithm 2.1.

[2] An implementation of FCI is available from `www.phil.cmu.edu/projects/tetrad/current`.

$\mathcal{R}$0a	If $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, then $X \not\!\!\perp\!\!\!\perp Y$, $Sep(X,Y) \leftarrow \mathbf{Z}$.
$\mathcal{R}$0b	If $X *\!\!-\!\!* Z \circ\!\!-\!\!* Y$ and $X \not\!\!\perp\!\!\!\perp Y$, then if $Z \notin Sep(X,Y)$, then $X *\!\!\rightarrow Z \leftarrow\!\!* Y$.
$\mathcal{R}$1	If $X *\!\!\rightarrow Z \circ\!\!-\!\!* Y$, and $X \not\!\!\perp\!\!\!\perp Y$, then $Z \longrightarrow Y$.
$\mathcal{R}$2a	If $Z \longrightarrow X *\!\!\rightarrow Y$ and $Z *\!\!-\!\!\circ Y$, then $Z *\!\!\rightarrow Y$.
$\mathcal{R}$2b	If $Z *\!\!\rightarrow X \longrightarrow Y$ and $Z *\!\!-\!\!\circ Y$, then $Z *\!\!\rightarrow Y$.
$\mathcal{R}$3	If $X *\!\!\rightarrow Z \leftarrow\!\!* Y$, $X *\!\!-\!\!\circ W \circ\!\!-\!\!* Y$, $X \not\!\!\perp\!\!\!\perp Y$, and $W *\!\!-\!\!\circ Z$, then $W *\!\!\rightarrow Z$.
$\mathcal{R}$4a	If $u = \langle X, .., Z_k, Z, Y\rangle$ is a discr. path between X and Y for Z, and $Z \circ\!\!-\!\!* Y$, then if $Z \in Sep(X,Y)$, then $Z \longrightarrow Y$.
$\mathcal{R}$4b	Idem, if $Z \notin Sep(X,Y)$ then $Z_k \longleftrightarrow Z \longleftrightarrow Y$.
$\mathcal{R}$5	If $u = \langle Z, X, .., W, Y, Z, X\rangle$ is an u.c. path, then $Z \!-\!\!- Y$ (= all edges on u).
$\mathcal{R}$6	If $X \!-\!\!- Z \circ\!\!-\!\!* Y$, then orient as $Z \!-\!\!* Y$.
$\mathcal{R}$7	If $X \!-\!\!\circ Z \circ\!\!-\!\!* Y$, and $X \not\!\!\perp\!\!\!\perp Y$, then $Z \!-\!\!* Y$.
$\mathcal{R}$8a	If $Z \longrightarrow X \longrightarrow Y$ and $Z \circ\!\!\rightarrow Y$, then $Z \longrightarrow Y$.
$\mathcal{R}$8b	If $Z \!-\!\!\circ X \longrightarrow Y$ and $Z \circ\!\!\rightarrow Y$, then $Z \longrightarrow Y$.
$\mathcal{R}$9	If $Z \circ\!\!\rightarrow Y$, $u = \langle Z, X, W, .., Y\rangle$ is an u.p.d. path, and $X \not\!\!\perp\!\!\!\perp Y$, then $Z \longrightarrow Y$.
$\mathcal{R}$10	If $Z \circ\!\!\rightarrow Y$, $X \longrightarrow Y \longleftarrow W$, and $u_1 = \langle Z, S, .., X\rangle$, $u_2 = \langle Z, V, .., W\rangle$ are u.p.d. paths (possibly $S = X$ and/or $V = W$), then if $S \not\!\!\perp\!\!\!\perp V$, then $Z \longrightarrow Y$.

Table 2.1: Graphical edge orientation rules in augmented FCI.

Algorithm 2.1 Augmented FCI algorithm

Input : independence oracle for $\mathbf{V}$
Output : complete PAG $\mathcal{P}$ over $\mathbf{V}$

1: $\mathcal{P} \leftarrow$ fully $\circ\!\!-\!\!\circ$ connected graph over $\mathbf{V}$
2: **for all** $\{X, Y\} \in \mathbf{V}$ **do**
3: search *in some clever way* for a $\mathbf{Z} : X \perp\!\!\!\perp_p Y \mid \mathbf{Z}$
4: $\mathcal{P} \leftarrow \mathcal{R}$0a (eliminate $X \not\!\!\perp\!\!\!\perp Y$)
5: record $Sep(X,Y) \leftarrow \mathbf{Z}$
6: **end for**
7: $\mathcal{P} \leftarrow \mathcal{R}$0b (unshielded colliders)
8: **repeat** $\mathcal{P} \leftarrow \mathcal{R}1 - \mathcal{R}$4b **until** finished
9: $\mathcal{P} \leftarrow \mathcal{R}$5 (uncovered circle paths)
10: **repeat** $\mathcal{P} \leftarrow \mathcal{R}6 - \mathcal{R}7$ **until** finished
11: **repeat** $\mathcal{P} \leftarrow \mathcal{R}$8a$-\mathcal{R}$10 **until** finished

Starting from the fully $\circ\!\!-\!\!\circ$ connected graph in line 1, $\mathcal{R}$0a eliminates all edges between conditionally independent nodes to obtain the skeleton of $\mathcal{P}$ with only $\circ\!\!-\!\!\circ$ edges (line 4). Then rules $\mathcal{R}$0b-$\mathcal{R}$4b (lines 7-8) obtain all invariant arrowheads (as well as some tails). Rules $\mathcal{R}5 - \mathcal{R}10$ (lines 9-11) then suffice to identify all and only the remaining invariant tails. For example, in Figure 2.1, the arrowheads at C from A and B are identified by $\mathcal{R}$0b, and the tailmark at $B \longrightarrow F$ follows from $\mathcal{R}$9.

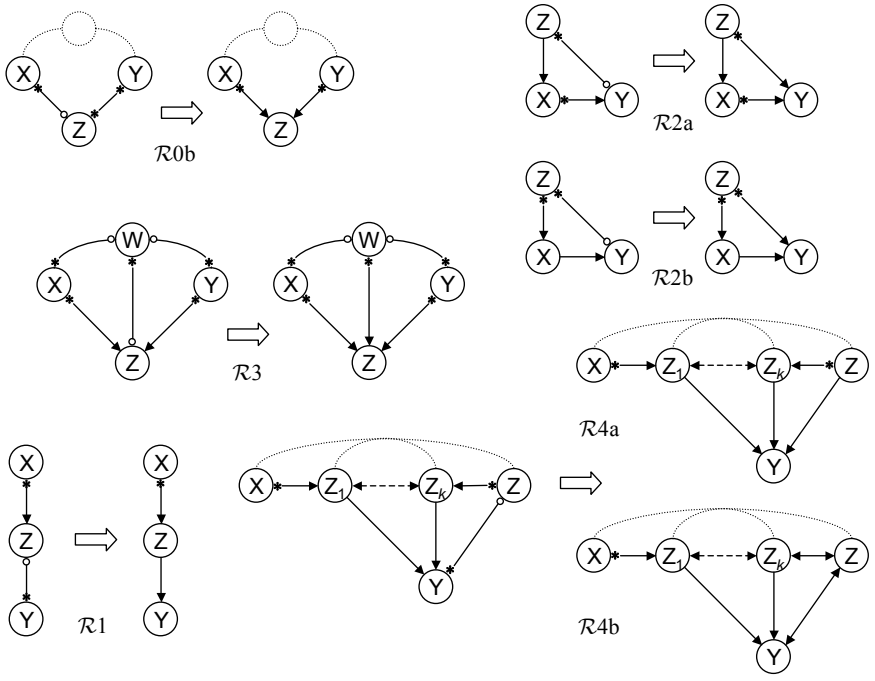

Figure 2.2: FCI - Arrowhead orientation rules

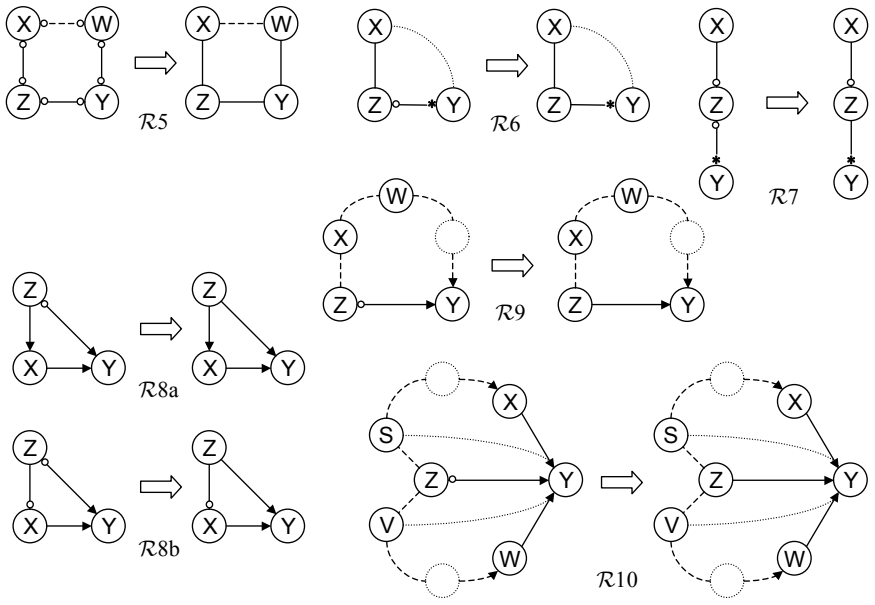

Figure 2.3: FCI - Tail orientation rules

Chapter 3

A Logical Characterization of Causal Discovery

We present a novel approach to constraint-based causal discovery, that takes the form of straightforward logical inference, applied to a list of simple, logical statements about causal relations that are derived directly from observed (in)dependencies. We show that two logical inference rules can find all invariant arrowheads in a causal model – signifying absence of causal relations – where before a set of seven graphical orientation rules were needed. An additional third rule then suffices to make the method both sound and complete, in the sense that all invariant features of the corresponding partial ancestral graph (PAG) are identified, even in the presence of latent variables and selection bias. The approach shows that identifiable causal relations can be reduced to one of just two fundamental forms. More importantly, as the basic building blocks of the method do not rely on the detailed (graphical) structure of the PAG, it opens up a range of new opportunities, including breaking up the inference process in modular, bite-size steps, and application to multiple models (see sections 4 and 5).

3.1 Introduction

Discovering causal relationships from data has long been an active area of research. From earlier, heated discussions in the 80's and before on whether it was possible at all to infer causality from observations alone, the debate has now shifted to

This chapter is based on: [Claassen and Heskes, 2011a] "A Logical Characterization of Constraint-Based Causal Discovery", published at the 27th Conference on Uncertainty in Artificial Intelligence, and [Claassen and Heskes, 2011b] "A Structure Independent Algorithm for Causal Discovery", published at the 19th European Symposium on Artificial Neural Networks.

means and methods to infer as much causal information from any source of data as possible, using the fewest, least restrictive assumptions.

The famous *Fast Causal Inference* (FCI) algorithm [Spirtes *et al.*, 2000], see section 2.4, was one of the first algorithms that was able to validly infer causal relations from conditional independence statements in the large sample limit, even in the presence of latent and selection variables. It consists of an efficient search for a conditional independence between each pair of variables to identify the skeleton of the underlying causal MAG, followed by an orientation stage to identify invariant tail and arrowhead marks.

It was shown to be sound in the large sample limit [Spirtes *et al.*, 1999], although not yet complete. Ali *et al.* [2005] proved that the seven graphical orientation rules employed by FCI, see Figure 2.2, were sufficient to identify all invariant arrowheads in the equivalence class $[\mathcal{M}]$, given a single MAG $\mathcal{M}$. Later, Zhang [2008] introduced another set of seven rules (Figure 2.3) to orient all remaining invariant tails. Augmented with this set of rules the FCI algorithm is also provably complete. The resulting maximally informative PAG $\mathcal{P}$ contains all causal relations identifiable from the set of observed independencies (see section 2.3 on how to read causal information from a PAG).

In this chapter we introduce three rules to convert minimal in/dependencies into logical statements about causal relations. We show that straightforward inference on these logical statements, using standard properties of causality, is sufficient to obtain all identifiable causal information. The result is the first provably sound and complete alternative to the augmented FCI algorithm. The logical approach produces a much simpler, arguably more natural method for causal inference, that is used as the basis in subsequent chapters to handle results from different experiments, and to develop a new method that outcompetes existing algorithms for causal discovery from real-world data sets.

Section 3.2 in this chapter shows that all invariant arrowheads are instances of just two fundamental cases. Section 3.3 extends this into a theory for inference from causal logic. Section 3.4 shows that the method is sound and complete, after which section 3.5 shows an implementation into the so-called LoCI algorithm. Proofs are relegated to the appendices (except for two results in 3.2 that can be found in [Claassen and Heskes, 2010a]). For a quick overview of standard concepts and terminology in causal modeling and ancestral graphs see chapter 2.

3.2 Invariant arrowheads and minimal independence

This section derives the fairly remarkable property that the seven graphical orientation rules in Figure 2.2 to find all invariant arrowheads are, in fact, different manifestations of just *two* fundamental rules. Invariant arrowheads correspond to identifiable absence of causal relations $X \not\Rightarrow Y$. Ruling out possible causes or contributing factors can in many cases be almost as important as finding factors that *do* contribute. Note again that below $\mathcal{G}_C$ always refers to the unknown but faithful

underlying causal DAG, and that $\mathbf{S}$ indicates the (possibly empty) set of selection variables in $\mathcal{G}_C$; see also Example 1.1 and section 2.2.

We start from the following observation that brings out the fundamental connection between a node that *makes* or *breaks* a conditional independence, and the *presence* or *absence* of certain causal relations:

Lemma 3.1. *Let X, Y, Z, and $\mathbf{W}$ be four disjoint (sets of) observed nodes in a causal DAG $\mathcal{G}_C$, and $\mathbf{S}$ be a set of (unobserved) selection nodes. If a node Z makes or breaks an independence relation between X and Y given $\mathbf{W}$, then:*

1. $X \perp\!\!\!\perp_p Y \,|\, \mathbf{W} \cup [Z] \quad \vdash \quad Z \Rightarrow (X \cup Y \cup \mathbf{W} \cup \mathbf{S})$,
2. $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{W} \cup [Z] \quad \vdash \quad Z \not\Rightarrow (X \cup Y \cup \mathbf{W} \cup \mathbf{S})$.

with special case

1′. $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{W} \cup Z] \quad \vdash \quad Z \Rightarrow (X \cup Y \cup \mathbf{S})$[1].

Together, the rules allow to infer causal relations, even in the presence of latent variables and selection bias: find a minimal conditional independence $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{Z}]$, and for some $Z \in \mathbf{Z}$ eliminate $Z \not\Rightarrow X$ and $Z \not\Rightarrow \mathbf{S}$ by a conditional dependence $X \not\perp\!\!\!\perp_p U \,|\, \mathbf{W} \cup [Z]$ to infer $Z \Rightarrow Y$. With this we can derive the following result (where the notation is chosen to match the graphical rules in Figure 2.2):

Proposition 3.2. *In a faithful PAG $\mathcal{P}$, all invariant arrowheads are instances of:*

rule (1): $Y \ast\!\!\rightarrow Z$, obtained from a dependence $U \not\perp\!\!\!\perp_p V \,|\, \mathbf{W} \cup [Z]$ created by Z from a minimal independence $U \perp\!\!\!\perp_p V \,|\, [\mathbf{W}]$, with $Y \in \{U, V, \mathbf{W}\}$, or

rule (2): $Z \longrightarrow Y$, from a minimal $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{W} \cup Z]$, with an arrowhead $X \ast\!\!\rightarrow Z$ from either rule (1) or rule (2).

Proof sketch. Both rules are sound, as they are direct applications of Lemma 3.1. The proof that they are also complete follows by induction on the graphical orientation rules $\mathcal{R}0\text{b}-\mathcal{R}4\text{b}$ (see Figure 2.2), showing that each time one of these is triggered, they always satisfy rules (1) and (2) above. As $\mathcal{R}0\text{b}-\mathcal{R}4\text{b}$ are sufficient for arrowhead completeness, rules (1) and (2) hold for *all* invariant arrowheads. For the full proof see lemmas 5-10 for Theorem 1 in [Claassen and Heskes, 2010a]. □

So all invariant arrowheads start from a minimal independence. Fortunately, we do not need to find *all* minimal independencies:

Lemma 3.3. *In a faithful PAG $\mathcal{P}$, checking rules (1) and (2) for a single, arbitrary minimal independence for each pair of nodes $\{X, Y\}$ (if it exists) is sufficient to orient all invariant arrowheads.*

Proof. See lemmas 11-13 in [Claassen and Heskes, 2010a]. □

[1] Many thanks to Peter Spirtes for pointing out that this case was already mentioned in Spirtes *et al.* [1999] (corollary to lemma 14) to prove correctness of the FCI-algorithm, although never used as an orientation rule.

The standard implementation of FCI, see Algorithm 2.1, already finds such single minimal sets, as it looks for separating sets $\mathbf{Z}$ of increasing size (lines 2-3). That means that for a (new) algorithm to find all invariant arrowheads we can take lines 1-8 of FCI, and replace the seven orientation rules from Figure 2.2 with just the two(!) rules from Proposition 3.2, where rule (1) executes once on each minimal dependence, and then rule (2) executes until no more arrowheads are found. But the next sections show we can do much better. Note that, when starting from the full set of independence statements in lemma 3.1, it is *not* necessary to consider discriminating paths in order to guarantee arrowhead completeness, contrary to when starting from a MAG as in [Ali *et al.*, 2005].[2]

3.3 Inference from Causal Logic

Instead of trying to match invariant features in the graph (PAG), we can also use Lemma 3.1 to translate observed minimal in/dependencies *directly* into logical statements about the presence or absence of certain causal relations. We can then reason with these by logical deduction on standard causal properties to obtain new information. This turns out to be a very natural and efficacious method for causal discovery.

Note: we again use X, Y, $\mathbf{Z}$, etc. to denote disjoint (sets of) observed variables, and $\mathbf{S}$ to denote the (possibly empty) set of selection nodes in a causal DAG $\mathcal{G}_C$.

3.3.1 Logical rules from minimal independence

First we state the standard properties of causal relations from section 2.3 into causal logic form:

Proposition 3.4. *Causal relations in a DAG $\mathcal{G}_C$ are:*

irreflexive	$: (X \Rightarrow X)$	$\vdash \mathit{false}$
acyclic	$: (X \Rightarrow Y)$	$\vdash Y \not\Rightarrow X$
transitive	$: (X \Rightarrow Y) \wedge (Y \Rightarrow Z)$	$\vdash X \Rightarrow Z$

Next we cast Lemma 3.1 in a form that translates observed minimal (in)dependencies *directly* into logical statements about causal relations:

Lemma 3.5. *For minimal in/dependencies between nodes in a causal DAG $\mathcal{G}_C$:*

1. $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{W} \cup Z] \;\vdash\; (Z \Rightarrow X) \vee (Z \Rightarrow Y) \vee (Z \Rightarrow \mathbf{S})$
2. $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{W} \cup [Z] \;\vdash\; (Z \not\Rightarrow X) \wedge (Z \not\Rightarrow Y) \wedge (Z \not\Rightarrow \mathbf{W}) \wedge (Z \not\Rightarrow \mathbf{S})$

[2] This may seem contradictory, as a MAG is just an encoding of an independence model, but it is not possible to read *which* set separates X and Y in $\mathcal{R}$4a/b from the MAG, without actually checking for the discriminating path.

By establishing which minimal in/dependencies hold in a distribution, a **list of logical statements Ł** can be compiled, of the form:

$$\begin{array}{ll} 1: & (Z \Rightarrow X) \ \vee \ (Z \Rightarrow Y) \ \vee \ (Z \Rightarrow \mathbf{S}) \\ 2: & (X \not\Rightarrow Y) \\ 3: & (Y \Rightarrow X) \ \vee \ (Y \Rightarrow W) \ \vee \ (Y \Rightarrow \mathbf{S}), \text{ etc.} \end{array}$$

Each line states a truth, for one specific node, about the causal relations it has with one or more others. New statements can be inferred by substituting the subject of one line in another, and then reducing via $X \not\Rightarrow Y \stackrel{\text{def}}{=} \neg(X \Rightarrow Y)$ and the standard causal properties in Proposition 3.4. The following two examples illustrate the inference process in deriving (new) causal information:

Example 3.6. *Suppose in a causal system $\mathcal{G}_C$ both $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$ and $X \not\perp\!\!\!\perp_p U \mid \mathbf{W} \cup [Z]$ have been observed, for some $Z \in \mathbf{Z}$, then*

$$\begin{array}{ll} 1: & (Z \Rightarrow X) \ \vee \ (Z \Rightarrow Y) \ \vee \ (Z \Rightarrow \mathbf{S}) \\ 2: & (Z \not\Rightarrow X) \ \wedge \ (Z \not\Rightarrow U) \ \wedge \ (Z \not\Rightarrow \mathbf{S}) \ \wedge \ (Z \not\Rightarrow \mathbf{W}) \end{array}$$

Using (2:) to eliminate $Z \Rightarrow X$ and $Z \Rightarrow \mathbf{S}$ from (1:) then gives (3:)

$$\begin{array}{lllll} \vdash & (\bot) & \vee \ (Z \Rightarrow Y) & \vee & (\bot) \\ 3: & & (Z \Rightarrow Y) & & \end{array}$$

This case corresponds to the embedded Y-structure from Mani *et al.* [2006], and matches the conditions for orientation rule $\mathcal{R}1$ in Table 2.1.

Example 3.7. *Now suppose both $Z \perp\!\!\!\perp_p W \mid [\mathbf{U}_{ZW} \cup X]$ and $X \perp\!\!\!\perp_p Y \mid [\mathbf{U}_{XY} \cup Z \cup W]$ have been observed, with $\mathbf{U}_{XY}$ and $\mathbf{U}_{ZW}$ two possibly empty/overlapping sets of nodes. Then for the inference list this gives (amongst others) statement (1:) from the first independence, and (2:) and (3:) from the second:*

$$\begin{array}{lllllll} 1: & (X \Rightarrow Z) & \vee \ (X \Rightarrow W) & \vee & & \vee \ (X \Rightarrow \mathbf{S}) \\ 2: & (Z \Rightarrow X) & \vee & & \vee \ (Z \Rightarrow Y) & \vee \ (Z \Rightarrow \mathbf{S}) \\ 3: & & (W \Rightarrow X) & \vee \ (W \Rightarrow Y) & \vee \ (W \Rightarrow \mathbf{S}) \end{array}$$

Using transitivity to substitute (1:) in (2:) this reduces by acyclicity to (4:)

$$\begin{array}{lllllll} \vdash & (Z \Rightarrow Z) & \vee \ (Z \Rightarrow W) & \vee \ (Z \Rightarrow Y) & \vee \ (Z \Rightarrow \mathbf{S}) \\ 4: & & (Z \Rightarrow W) & \vee \ (Z \Rightarrow Y) & \vee \ (Z \Rightarrow \mathbf{S}) \end{array}$$

Idem, (1:) in (3:) gives (5:), which after substitution in (4:) reduces to (6:)

$$\begin{array}{lllllll} 5: & (W \Rightarrow Z) & \vee & & \vee \ (W \Rightarrow Y) & \vee \ (W \Rightarrow \mathbf{S}) \\ \vdash & (Z \Rightarrow Z) & \vee \ (Z \Rightarrow Y) & \vee \ (Z \Rightarrow Y) & \vee \ (Z \Rightarrow \mathbf{S}) \\ 6: & & & (Z \Rightarrow Y) & \vee \ (Z \Rightarrow \mathbf{S}) \end{array}$$

This case matches $\mathcal{R}9$ in Figure 2.3, where all alternatives for Z from the second minimal independence necessarily lead to a causal relation to node Y (or $\mathbf{S}$). Note that now selection bias cannot be eliminated. The latter is illustrated by Fig. 2.1(3), with nodes (B, F, C, E) in the role of resp. (Z, Y, X, W) in Example 3.7, leading to conclusion $(B \Rightarrow F) \vee (B \Rightarrow \mathbf{S})$, but it is possible that only $B \Rightarrow \mathbf{S}$ holds true in the (unknown) underlying causal DAG $\mathcal{G}_C$, as seen in (1).

3.3.2 Inferred statements

Remarkably enough, Lemma 3.1 is already sufficient to identify almost all causal information that can be discovered from probabilistic independencies.There is just one more piece of information needed to complete the puzzle.

Lemma 3.8 (**Inferred blocking node**). *In a causal system* $\mathcal{G}_C$*, if* $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{Z}]$ *and there is a subset* $\{Z_1, \ldots, Z_k, Z\} \subseteq \mathbf{Z}$*, such that in the sequence* $[U_0, \ldots, U_{k+2}] = [X, Z_1, \ldots, Z_k, Z, Y]$ *it holds that:*

- $U_i \not\twoheadleftarrow \{U_{i-1}, U_{i+1}\}$,
- $U_j \not\perp\!\!\!\perp_p U_{j+1} \,|\, \mathbf{Z}'$,

with $i = 1..k$*,* $j = 0..(k+1)$*, and* $\forall \mathbf{Z}' \subseteq \mathbf{Z} \setminus \{U_j, U_{j+1}\}$*, then* $Z \Rightarrow (Z_k \cup Y \cup \mathbf{S})$.

In other words, if we find an 'inferred blocking node', then we can add the following statement to the list Ł:

n: $\quad (Z \Rightarrow Z_k) \;\vee\; (Z \Rightarrow Y) \;\vee\; (Z \Rightarrow \mathbf{S})$

Lemma 3.8 is clearly a generalization of rule $\mathcal{R}$4a: if the nodes in the sequence $[\mathbf{U}]$ are adjacent in $\mathcal{P}$, then it corresponds to a discriminating path for Z, and the non-independence tests in the second item can be omitted. Note that resulting statement (n:) reveals that the discriminating path for Z in $\mathcal{R}$4a behaves identical to a node Z observed in a minimal independence between Z_k and Y. As a result, whether or not we observe $Z_k \perp\!\!\!\perp Y \,|\, [.. \cup Z]$, the fact that in the given conditions Z does *not* create a dependency between X and Y, allows us to infer that Z blocks *some* path between Z_k and Y; hence 'inferred blocking node'.

One remark: the set of independence relations possibly involved in lemma 3.8 may seem quite daunting. However, in section 3.5 we will see that ultimately only a handful need to be checked.

3.3.3 Direct and indirect causal relations

Reasoning with presence or absence of causal relations implies that we are not limited to direct causal influences only, but can draw on other, indirect sources of causal information as well: both can be used to derive new information in exactly the same way. But often it is very informative to know what direct and what indirect causes are, and so we would like to be able to distinguish between them:

Lemma 3.9. *In a causal system* $\mathcal{G}_C$*, a conditional independence* $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$ *implies that all causal paths between* X *and* Y *in* $\mathcal{G}_C$ *are mediated by nodes in* $\mathbf{Z}$.

All causal paths mediated by other (observed) nodes implies no direct causal link (or confounder) between X and Y, so then $X \Rightarrow Y$ is an *indirect* causal relation. For independent nodes $X \perp\!\!\!\perp Y \,|\, \varnothing$ it also implies that neither is a cause of the other: $(X \not\twoheadleftarrow Y) \wedge (Y \not\twoheadleftarrow X)$.

Lemma 3.9 gives the skeleton (structure) of the PAG, where a missing edge represents absence of a direct cause or confounding link. If we want to distinguish

between direct and indirect causal relations in our list Ł, we can simply 'project' the causal information onto the PAG skeleton:

Lemma 3.10. *The causal information in the list Ł can be transferred to invariant edge marks between adjacent nodes in the corresponding PAG $\mathcal{P}$:*

- *if* $(X \not\Rightarrow Y) \in Ł$, *then* $X \leftarrow\!\!* Y$,
- *if* $(X \Rightarrow Y)$ *and/or* $(X \Rightarrow \mathbf{S}) \in Ł$, *then* $X —\!\!* Y$,

Proof. Follows from MAG definition for tail/arrowhead marks in [Richardson and Spirtes, 2002, §4.2] □

In the next section we find that reasoning on logical statements obtained from (inferred) minimal in/dependencies is not only sound but also complete. It means that after the logical causal inference (LoCI) process has completed, we can choose to reproduce the complete PAG by projecting all ancestral information (tails and arrowheads) onto the global skeleton. If only a subset of independence relations is available, for example because the size of the conditioning sets is restricted as in Anytime FCI [Spirtes, 2001], or because unreliable decisions are rejected, then the skeleton may contains superfluous edges. But even then all inferred causal relations in the list Ł remain valid, and so does the resulting causal PAG, even when the orientation rules in Table 2.1 can no longer be applied.

3.4 A Logical Characterization of Causal Information

From the two previous sections we know that all invariant arrowheads in a PAG are instances of two cases that are covered by the two rules in Lemma 3.1 in combination with logical deduction on Proposition 3.4. In this section we show that, together with a third rule from Lemma 3.8, they are also sufficient to find all invariant tails in the PAG. We do this by matching each graphical orientation rule in Table 2.1 to specific instances of the lemmas and already inferred information, where we rely on the known completeness of the FCI algorithm. As this covers all invariant features in the PAG, lemmas 3.1 and 3.8 also cover *all* identifiable causal information from independence relations.

3.4.1 Invariant tails

The rules in section 3.2 will not only find all arrowheads, but also a number of invariant tails, as rule (2) in Proposition 3.2 already covers all instances of rule $\mathcal{R}1$, including the tail $Z \longrightarrow Y$. In this section we show that all remaining invariant tails from rules $\mathcal{R}5 - \mathcal{R}10$ in Figure 2.3 correspond to *three* cases that can be found from minimal in/dependencies in combination with the standard causal rules in Proposition 3.4.

We want to emphasize that there is no need to search for any of the specific cases that match individual orientation rules discussed here or in the proofs of this section or that of section 3.2: they automatically 'pop up' when running the causal logic rules in Proposition 3.4 on the list of statements Ł. We only need/use them here to characterize the causal information that can be identified in this way, and thus, since the augmented FCI algorithm is complete, by any algorithm for causal discovery.

To do that, we first introduce the following concept:

Definition 3.11. *A sequence* $[X, Z_1, \ldots, Z_k, V_1, \ldots, V_m, Y]$ *is a* ***transitive relation*** *from* X *to* Y *if it holds that:*

- $\forall Z_i, \exists \mathbf{U}_i : Z_{i-1} \perp\!\!\!\perp Z_{i+1} \,|\, [\mathbf{U}_i \cup Z_i]$,
- $\forall V_j : V_j \Rightarrow (V_{j+1} \cup \mathbf{S})$,

with $Z_0 = X$, $Z_{k+1} = V_1$, $V_{m+1} = Y$, *for* $k, m \geq 0$.

In words: a series of overlapping minimal conditional independencies, followed by a causal relation. A transitive relation can be as short as a single independence $X \perp\!\!\!\perp Y \,|\, [Z_1]$, or a relation $X \Rightarrow Y$. As such, it is a generalization of the u.p.d. path in section 2.1. The reason for this introduction is the property:

Corollary 3.12. *In a causal system* $\mathcal{G}_C$, *if there is a transitive relation* $[X, Z_1, \ldots, Y]$, *then:*

- $X \Rightarrow (Z_1 \cup \mathbf{S}) \;\vdash\; X \Rightarrow (Y \cup \mathbf{S})$.

With this we can state:

Proposition 3.13. *In a PAG* $\mathcal{P}$, *all invariant tails* $Z \multimap\!\!* Y$ *from graphical orientation rules* $\mathcal{R}4a$, $\mathcal{R}5$, $\mathcal{R}7$, $\mathcal{R}9$, *and* $\mathcal{R}10$ *are instances of:*

(2b): $X \perp\!\!\!\perp Y \,|\, [\mathbf{W} \cup Z]$, *with* $X \Rightarrow (Z \cup \mathbf{S})$ *from either case (3) or another (2b),*

(3): $U \perp\!\!\!\perp V \,|\, [\mathbf{W} \cup W]$, *with two transitive relations* $[W, U, .., Y] + [W, V, .., Y]$, *and* $Z \in \{U, V, W\}$,

(4): $X \perp\!\!\!\perp Y \,|\, [\mathbf{Z}]$, *with inferred blocking node* $Z \in \mathbf{Z}$, *together with* $Z_k \Rightarrow (Y \cup \mathbf{S})$ *from either case (2) or case (4).*

Case (2b) covers rule $\mathcal{R}7$, and is so named because of its similarity/overlap with case (2) for $\mathcal{R}1$. Case (3) covers all instances of rules $\mathcal{R}5$, $\mathcal{R}9$, and $\mathcal{R}10$, and case (4) accounts for tails from orientation rule $\mathcal{R}4a$. In most instances of case (3) the transitive relation requires only a single minimal conditional independence, even for long paths. Often, both transitive relations can be captured together in a single independence, as in Example 3.7.

As a bonus all identifiable selection nodes $X \Rightarrow \mathbf{S}$ also pop out 'automatically' by applying the rules in Proposition 3.4 on instances of case (3):

Corollary 3.14. *In a PAG* $\mathcal{P}$, *all identifiable selection nodes* $X \Rightarrow \mathbf{S}$ *are covered by case (3), in the form of a minimal independence with two transitive relations back to itself.*

That leaves just tails from three more orientation rules to handle. However, these too follow implicitly from the existing cases:

Corollary 3.15. *In a PAG $\mathcal{P}$, all invariant tails from orientation rules $\mathcal{R}6$, $\mathcal{R}8a$, and $\mathcal{R}8b$, are covered by the causal logic rules applied to cases (1)-(4).*

3.5 Logical Causal Discovery

In this section we turn the logical causal inference rules into an efficient anytime algorithm for deriving both all explicit causal relations and the complete PAG.

3.5.1 Inference process

A nice property is that the logical substitute/reduce steps take on a particularly simple form: it only involves statements that are either a logical disjunction of at most two causal relations and possible selection bias, or a single term for the absence of a causal relation. In other words, the list of logical causal statements Ł always keeps the simple form shown in section 3.3.1. Each step consists of a substitution of one statement in another followed by a reduction to this standard form, see Corollary 3.30 in Appendix 3.D. Furthermore, as more information becomes available, statements in the list can simplify from three to two or even one term. Cf. Example 3.6, where inferred statement (3:) replaces (1:), as there is no point in keeping the original.

The next result limits the independence search:

Lemma 3.16. *In the logical causal inference (LoCI) approach, finding a single, arbitrary $X \perp\!\!\!\perp Y \mid [\mathbf{Z}]$, for each pair of nodes (X, Y) in the graph (if it exists) is sufficient to find all invariant features of the PAG.*

Fortunately, the current implementation of the FCI algorithm already finds only *minimal* conditional independencies for each pair of nodes (if it exists), as it looks for sets of increasing size until it finds one that separates the two. This is also the dominant factor in the time-complexity of the algorithm. For each pair found, we still need to check for other nodes that can destroy this independence (Lemma 3.1, item 2), however, this is negligible compared to the independece search itself.

Furthermore, the inferred blocking node from lemma 3.8, can be tackled efficiently *after* all invariant arrowheads have been found from cases (1) and (2): it makes it possible to establish all 'non-ancestor' conditions in the sequence in one go. Together with a restriction to a sequence of non-separated nodes (avoiding the additional dependence tests), this greatly reduces the number of candidates to check. The rest of the inference process is straightforward logical deduction to ultimately obtain a matrix $\mathbf{M}_C$ of in/direct causal relations, stating for each ordered pair (X, Y) of variables whether: $X \Rightarrow Y$, $X \not\Rightarrow Y$, or "don't know".

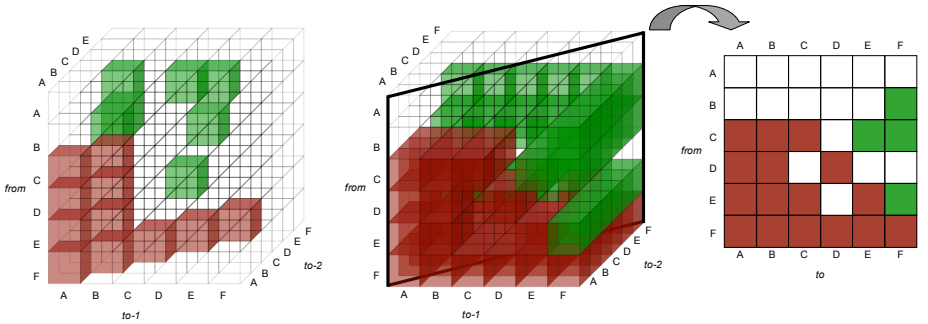

Figure 3.1: a) 3D map of list Ł of logical causal statements; b) idem, after exhaustive application of causal rules 3; c) matrix $\mathcal{M}_C$ of inferred causal relations (diagonal plane)

Example 3.17. *The remaining inference process on the logical list Ł can be visualized as filling in a 3D-cube, in Figure 3.1. We only need logical statements of at most three terms, that either indicate (possible) presence, or absence of relations. Using coordinates* $(Z, X, Y) \equiv (top \rightarrow down, left \rightarrow right, front \rightarrow back)$ *in Figure 3.1(a), this becomes:*

- $(Z, X, Y) = +1$ *encodes* $(Z \Rightarrow X) \vee (Z \Rightarrow Y) \vee (Z \Rightarrow \mathbf{S})$,
- $(Z, X, X) = +1$ *encodes* $(Z \Rightarrow X) \vee (Z \Rightarrow \mathbf{S})$,
- $(Z, X, X) = -1$ *encodes* $(Z \not\Rightarrow X)$,
- $(X, X, X) = +1$ *encodes* $(X \Rightarrow \mathbf{S})$.
- $(X, X, X) = -1$ *encodes* $(X \not\Rightarrow \mathbf{S})$.

where green equals $+1$*, red equals* -1*, and white represents empty cells with default value* 0*, that could not be inferred (yet).*

Proposition 3.4 now fills in new cells, until no more can be found, Figure 3.1(b). At that point the diagonal plane $(from, to-1 = to-2)$ *represents the causal matrix* $\mathbf{M}_C$ *in Figure 3.1(c). This matrix can be 'projected' onto the skeleton (green=tail, red=arrowhead, white=circle) to obtain the corresponding PAG in Figure 2.1(c).*

Note that the causal matrix $\mathbf{M}_C$ *in (c) clearly shows that* $C \Rightarrow E \Rightarrow F$ *represent definite causal relations, as* $(C, C) = (E, E) = -1$ *(red), whereas the link* $B \rightarrow F$ *cannot exclude selection bias, given that cell* $(B, B) = 0$ *is still white.*

In practice, the inference process illustrated in Figure 3.1 is so fast compared to the independence search that it makes sense to execute it each time a new minimal in/dependence is found, to turn it into an efficient anytime algorithm. The final step uses lemmas 3.9 and 3.10 to transfer the logical information in the list Ł to invariant edge marks in the skeleton of $\mathcal{P}$.

3.5.2 The LoCI algorithm

We can now give the outline of the so-called LoCI algorithm that is able to infer the complete PAG, using the logical causal inference approach described in the previous sections.

Algorithm 3.1 Logical Causal Inference (LoCI) algorithm

Input : independence oracle for $\mathbf{V}$
Output : complete PAG $\mathcal{P}$ over $\mathbf{V}$
1: **for all** $\{X, Y\} \in \mathbf{V}$ **do**
2: search *in some clever way* for a $X \perp\!\!\!\perp Y \mid [\mathbf{Z}]$
3: $\forall Z \in \mathbf{Z} : Ł \leftarrow Z \Rightarrow (X \cup Y \cup \mathbf{S})$
4: $\forall W, X \not\perp\!\!\!\perp Y \mid \mathbf{Z} \cup W :$
5: $Ł \leftarrow W \not\Rightarrow (X \cup Y \cup \mathbf{Z} \cup \mathbf{S})$
6: **repeat** Ł $\leftarrow$ substitute/reduce **until** finished
7: **end for**
8: Ł $\leftarrow Z \Rightarrow (Z_k \cup Y \cup \mathbf{S}), \forall Z$: inferred block. node
9: **repeat** Ł $\leftarrow$ substitute/reduce **until** finished
10: $\mathcal{P} \leftarrow$ fully $\circ\!\!-\!\!\circ$ connected graph over $\mathbf{V}$
11: eliminate $X \not\!\!-\!\!- Y$, iff $X \perp\!\!\!\perp Y \mid [*]$
12: orient $X -\!\!- * Y$, iff $X \Rightarrow (Y \cup \mathbf{S}) \in Ł$
13: orient $X \leftarrow\!\!- * Y$, iff $X \not\Rightarrow Y \in Ł$

Algorithm 3.1 borrows the initial search for (minimal) conditional independencies from the standard FCI algorithm. If it finds one it is recorded in the list Ł, line 3, and checked for nodes that destroy this independence (also recorded in Ł). Each time a minimal independence has been found, line 6 runs the inference rules to update the identifiable causal information. This step could be run just once, after the independence search has completed, but in practice the impact on performance is negligble and far outweighed by the fact that most causal information is already identifiable (available) in the early stages of the process. At line 8, all non-ancestor relations ($X \not\Rightarrow Y$) have been found (see lemma 7), which makes it relatively easy to find the remaining 'inferred blocking nodes' from lemma 3.8 in line 8. If any are found that contain new information, then line 9 infers the remaining relations. Finally, lines $10 - 13$ construct the equivalent PAG representation from the list Ł.

The LoCI algorithm above finds a minimal separating set for each pair of nodes (if it exists). By Lemma 3.16 that means that it can reconstruct the PAG, which from [Zhang, 2008] is known to be a sound and complete representation of all identifiable (absence of) causal relations from independencies. With this the main result of this section can now be summarized as:

Theorem 3.18. *The Logical Causal Inference (LoCI) algorithm is sound and complete for causal discovery from probabilistic independence relations.*

3.6 Discussion and Conclusion

In this chapter we developed a new approach to constraint-based causal discovery: observed minimal (in)dependencies are converted into logical statements about causal relations, and these statements are subsequently combined using basic properties of causality.

It leads to a remarkably simple characterization, in which all identifiable causal relations take the form of an (inferred) minimal conditional independence with either elimination of one alternative, or both alternatives leading to the same conclusion.

The resulting logical causal inference (LoCI) method was put to work in an efficient anytime algorithm, the first alternative to the augmented FCI-algorithm shown to be both sound and complete. The LoCI algorithm is strikingly simpler than its counterpart in section 2.4. Even though it is not necessarily faster, as for both the overall complexity is dominated by the independence search, the fact that the implementation takes on this very simple and elegant form suggests it is somehow more 'natural' to causal discovery.

The way in which the LoCI algorithm builds up this causal information is markedly different from many other constraint-based methods: instead of focussing on combinations of node pairs that may or may not be separable (the essence of graphical orientation rules), the LoCI algorithm focusses on the nodes that separate them. In particular, as it does not need to search for pairs of nodes that cannot be separated by any set (the edges forming the skeleton of the PAG), the approach taken by the algorithm could be dubbed 'structure independent'. As a result, it can be adapted to search for target causal relations in large models, updating each time as new independence information becomes available; of course, if we want to ensure completeness, we still have to find all of them.

The simplicity of the LoCI algorithm raises the question if a similar approach is viable in other applications as well. For example, incorporation of causal information from background knowledge or derived from other properties of the distribution [Shimizu *et al.*, 2006; Mooij *et al.*, 2010], is straightforward. The same holds for additional assumptions, such as 'no selection bias'. Including interventional information also fits nicely in this framework, and requires only minor modifications of the minimal independence lemma 3.1. An extension to multiple models, similar to [Triantafillou *et al.*, 2010; Claassen and Heskes, 2010b], seems feasible as well. To prove completeness, we had to rely on the known completeness of the augmented FCI algorithm, but we suspect that a more direct proof should be possible.

Perhaps the most promising aspect of the LoCI approach lies in the flexibility it offers in deriving causal information. For example, we are free to ignore any suspect, 'borderline' (in)dependence decisions, by not including them in the list Ł in lines 3 and 5: all inferred causal relations remain valid. This should definitely increase the reliability of the output, even though it is no longer guaranteed to be complete. Finally, the 'structure independent' aspect implies there are many different ways to arrive at the same conclusion. This makes it possible to choose

the most reliable combination(s) of independencies for a more robust conclusion and to detect inconsistencies. Tracking which logical statements in Ł are combined to identify new relations could also improve accountability for the output, indicating exactly *why* a causal relation was found.

3.A Proofs: causal relations from in/dependence

The first lemma is derived from familiar results, see [Spirtes *et al.*, 1999; Claassen and Heskes, 2010b], to bring out the symmetry between a node that makes and a node that breaks an independence relation.

Lemma 3.1 *Let X, Y, Z, and $\mathbf{W}$ be four disjoint (sets of) observed nodes in a causal DAG $\mathcal{G}_C$, and $\mathbf{S}$ be a set of (unobserved) selection nodes. If a node Z makes or breaks an independence relation between X and Y given $\mathbf{W}$, then:*

1. $X \perp\!\!\!\perp_p Y \,|\, \mathbf{W} \cup [Z] \quad \vdash \quad Z \Rightarrow (X \cup Y \cup \mathbf{W} \cup \mathbf{S})$,
2. $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{W} \cup [Z] \quad \vdash \quad Z \not\Rightarrow (X \cup Y \cup \mathbf{W} \cup \mathbf{S})$.

with special case

1′. $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{W} \cup Z] \quad \vdash \quad Z \Rightarrow (X \cup Y \cup \mathbf{S})$

Proof. (1.) To block, node Z must be a noncollider on a path $\pi = \langle X, .., Z, .., Y \rangle$ in $\mathcal{G}_C$ that is unblocked given $\mathbf{W} \cup \mathbf{S}$. As Z is a noncollider it has at least one outgoing arc along π. Follow π in this direction until either a collider is encountered or the end of π is reached. Every collider along π has to be an ancestor of $(\mathbf{W} \cup \mathbf{S})$, which implies that in either case Z has a directed path in $\mathcal{G}_C$ (=causal relation) to at least one node from $(X \cup Y \cup \mathbf{W} \cup \mathbf{S})$.
(2.) To unblock, node Z must be (a descendant of) at least one collider on a path $\pi = \langle X, .., Y \rangle$ in $\mathcal{G}_C$ that is blocked given $\mathbf{W} \cup \mathbf{S}$. Any directed path in $\mathcal{G}_C$ from Z to a node in $\mathbf{W} \cup \mathbf{S}$ implies that the collider(s) would already be unblocked when conditioning on just $\mathbf{W} \cup \mathbf{S}$. No directed paths from Z to $(\mathbf{W} \cup \mathbf{S})$ implies that if there existed a directed path from Z to X or Y, then it could not be blocked by any node $(\mathbf{W} \cup \mathbf{S})$. But then such a path would make Z a noncollider on an unblocked path between X and Y given $(\mathbf{W} \cup \mathbf{S})$: starting from X, let θ_X be the first collider encountered along π that is unblocked by conditioning on Z, and similarly θ_Y the first collider along π starting from Y, (possibly $\theta_X = \theta_Y$, but $\{\theta_X, \theta_Y\} \notin \mathbf{W}$ (otherwise Z not needed)); then the paths $\langle X, \theta_X, Z \rangle$ and $\langle Z, \theta_Y, Y \rangle$ are into Z and unblocked given $\mathbf{W} \cup Z$, so a directed path $Z \Rightarrow X$ would make Z a noncollider on unblocked path $\langle X, \theta_X, Z, Y \rangle$ given $\mathbf{W}$, contradicting $X \perp\!\!\!\perp_p Y \,|\, \mathbf{W}$; idem for $Z \Rightarrow Y$.
The special case (1′.) for minimal $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{W} \cup Z]$ follows from (1.) and acyclicity. By contradiction: suppose $\exists U_1 \in (\mathbf{W} \cup Z) : U_1 \not\Rightarrow (X \cup Y \cup \mathbf{S})$, then, as (1.) applies to all nodes $(\mathbf{W} \cup Z)$, there must be a node $U_2 \in (\mathbf{W} \cup Z) \setminus U_1$ (acyclicity) such that $U_1 \Rightarrow U_2$. But (transitivity) U_2 also cannot have a directed path to $(X \cup Y \cup \mathbf{S})$, and

so there must be a node $U_3 \in (\mathbf{W} \cup Z) \setminus \{U_1, U_2\}$ (acyclicity) such that $U_2 \Rightarrow U_3$. This can continue until all nodes in $(\mathbf{W} \cup Z)$ have been allocated at which stage the last node cannot have a directed path to any $(X \cup Y \cup \mathbf{W} \cup Z \cup \mathbf{S})$, in contradiction with (1.). □

3.B Proofs: causal logic rules

Observed probabilistic minimal conditional in/dependences can be converted into **logical statements about causal relations**:

Lemma 3.5 *For minimal in/dependencies between nodes in a causal DAG $\mathcal{G}_C$:*

1. $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{W} \cup Z] \;\vdash\; (Z \Rightarrow X) \vee (Z \Rightarrow Y) \vee (Z \Rightarrow \mathbf{S})$
2. $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{W} \cup [Z] \;\vdash\; (Z \not\Rightarrow X) \wedge (Z \not\Rightarrow Y) \wedge (Z \not\Rightarrow \mathbf{W}) \wedge (Z \not\Rightarrow \mathbf{S})$

Proof. Just rewriting Lemma 3.1 with the definition of (absence of) causal relations to *sets* of nodes, where:

- rule 1. corresponds to $\exists U \in (X \cup Y \cup \mathbf{S}) : Z \Rightarrow U$,
- rule 2. corresponds to $\forall U \in (X \cup Y \cup \mathbf{W} \cup \mathbf{S}) : Z \not\Rightarrow U$.

□

Next two lemmas used in the proof of Lemma 3.8.

Lemma 3.19. *For two observed nodes, X and Y, in a causal DAG $\mathcal{G}_C$: $X \not\perp\!\!\!\perp Y$, iff they are connected by a trek in $\mathcal{G}_C$ or they both have treks into* $\mathbf{S}$

Proof. Almost by definition. Assuming causal Markov and faithfulness, two observed nodes X and Y are dependent given a set $\mathbf{Z}$, iff they are connected by a path π in $\mathcal{G}_C$ on which all noncolliders are not in $\mathbf{Z}$ and all colliders are (ancestor of) nodes in $(\mathbf{Z} \cup \mathbf{S})$. For $\mathbf{Z} = \varnothing$ this reduces to a path π on which all colliders are in $An(\mathbf{S})$. Starting from X, follow π until the first collider. Then X has a colliderless path to a node with a directed path to $\mathbf{S}$, which implies a trek from X to $\mathbf{S}$. If Y is reached, then π is by definition a colliderless path, or trek, to Y. Idem for Y. □

Lemma 3.20. *In a causal DAG $\mathcal{G}_C$, if $X \not\perp\!\!\!\perp Y$, then identifiable absence of a causal relation $X \not\Rightarrow Y$ implies absence of selection bias $X \not\Rightarrow \mathbf{S}$.*

Proof. For adjacent nodes in a PAG $\mathcal{P}$, the proof is trivial: identifiable absence of a causal relation means identifiable non-ancestorship, and so an invariant arrowhead $X \leftarrow\!\!* Y$ in $\mathcal{P}$. By definition of the MAG, see §4.2 in [Richardson and Spirtes, 2002], this means that $X \notin An(Y \cup \mathbf{S})$.
For nonadjacent nodes we can use Theorem 2 from [Claassen and Heskes, 2010b], which states that there is identifiable absence of a causal relation $X \not\Rightarrow Y$, iff it is impossible to go from X to Y in the graph $\mathcal{P}$, without going against an arrowhead. By contradiction: suppose that $X \Rightarrow \mathbf{S}$. This implies $An(X) \Rightarrow \mathbf{S}$, so nodes that are

ancestor of X have no (invariant) arrowheads (only tails). By lemma 4.1, dependent nodes either have a trek between them, or both have treks to $\mathbf{S}$. But if there is a trek between X and Y, then no node between X and the source of that trek can have an arrowhead, and all nodes between the source and Y are going 'with' the arrowhead, so then not all paths go against an arrowhead. Similarly for treks to $\mathbf{S}$. Therefore also for nonadjacent nodes $X \not\Rightarrow \mathbf{S}$. □

Lemma 3.8 (Inferred blocking node). *In a causal system $\mathcal{G}_C$, if $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$ and there is a subset $\{Z_1, \ldots, Z_k, Z\} \subseteq \mathbf{Z}$, such that in the sequence $[U_0, \ldots, U_{k+2}] = [X, Z_1, \ldots, Z_k, Z, Y]$ we know that:*

- *$U_i \not\Rightarrow \{U_{i-1}, U_{i+1}\}$,*
- *$U_j \not\perp\!\!\!\perp_p U_{j+1} \mid \mathbf{Z}'$,*

with $i = 1..k$, $j = 0..(k+1)$, and $\forall \mathbf{Z}' \subseteq \mathbf{Z} \setminus \{U_j, U_{j+1}\}$, then $Z \Rightarrow (Z_k \cup Y \cup \mathbf{S})$.

Proof. In words: if no node Z_i in the minimal independence $X \perp\!\!\!\perp Y \mid [\mathbf{Z}]$ has a causal relation (directed path in $\mathcal{G}_C$) to either of its neighbors in the sequence $[X, Z_1, \ldots, Z_k, Z, Y]$, and all neighboring nodes in the sequence are dependent given any subset of $\mathbf{Z}$, then Z has a causal relation to Z_k, Y, and/or $\mathbf{S}$.

First we show that there is an unblocked path from X to Z in $\mathcal{G}_C$ relative to $\mathbf{Z}_{\setminus Z}$. The first item, in combination with Lemma 3.20 implies that there is no selection bias on any of the nodes Z_i. By Lemma 3.1, this, together with the given $Z_1 \not\Rightarrow X$, implies $Z_1 \Rightarrow Y$, and so it also follows that there is no selection bias on Y (otherwise $Z_1 \Rightarrow (Y) \Rightarrow \mathbf{S}$).
By the second item, all neighbors in the sequence $[\mathbf{U}]$ are dependent (given empty set), and so by the previous observation in combination with Lemma 3.19 this implies that each successive pair is connected by a trek (but not a directed path, by item 1) in $\mathcal{G}_C$, with the possible exception of the edges to X and Z, that can still correspond to directed paths and/or treks to $\mathbf{S}$.
As each successive pair in the sequence is connected by an unblocked (sub)path given $\mathbf{Z}_{\setminus Z}$ that is *into* both Z_i and Z_{i+1}, it follows (by concatenating them) that there is also an unblocked path from X to Z in $\mathcal{G}_C$ relative to $\mathbf{Z}_{\setminus Z}$. Nodes Z and Y are also not separated by any subset from $\mathbf{Z}$, and so are connected by an unblocked subpath relative to $\mathbf{Z}_{\setminus Z}$.

In conclusion, by construction there are unblocked paths from X (via Z_k) and Y to Z in $\mathcal{G}_C$, given $\mathbf{Z}_{\setminus Z}$. If both paths from Z_k and Y are into Z, then the sequence $[\mathbf{U}]$ would represent an unblocked path between X and Y given $\mathbf{Z}$ in $\mathcal{G}_C$, which would make X and Y dependent, contrary to the given. Therefore Z must be an ancestor of Z_k and/or Y, and/or have a directed path to $\mathbf{S}$ in $\mathcal{G}_C$. In other words, then: $Z \Rightarrow Z_k \vee Z \Rightarrow Y \vee Z \Rightarrow \mathbf{S}$. □

Lemma 3.9. *In a causal system $\mathcal{G}_C$, a conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ implies that all causal paths between X and Y in $\mathcal{G}_C$ are mediated by nodes in $\mathbf{Z}$.*

Proof. See [Spirtes *et al.*, 2000]. Assumes that no causal paths are blocked by selection nodes, which is implicitly covered by the faithfulness assumption. Implies that (unconditionally) independent nodes have no in/direct causal relation or confounder between them. □

3.C Proofs: logical characterization

Proposition 3.2. *In a faithful PAG $\mathcal{P}$, all invariant arrowheads are instances of:*

rule (1): $Y \ast\!\!\rightarrow Z$, obtained from a dependence $U \not\perp\!\!\!\perp_p V \mid \mathbf{W} \cup [Z]$ created by Z from a minimal independence $U \perp\!\!\!\perp_p V \mid [\mathbf{W}]$, with $Y \in \{U, V, \mathbf{W}\}$, or

rule (2): $Z \longrightarrow Y$, from a minimal $X \perp\!\!\!\perp_p Y \mid [\mathbf{W} \cup Z]$, with an arrowhead $X \ast\!\!\rightarrow Z$ from either rule (1) or rule (2).

Proof sketch. Both cases are sound:
(1.) By Lemma 3.1, the first gives $(Z \not\Leftarrow Y) \wedge (Z \not\Leftarrow \mathbf{S})$, which, by definition, implies that if Y has an edge to Z in $\mathcal{P}$, then the mark at Z is an (invariant) arrowhead.
(2.) The second likewise through elimination, giving $(Z \Rightarrow X) \vee (Z \Rightarrow Y) \vee (Z \Rightarrow \mathbf{S})$, where the first and third are eliminated by the arrowhead at $X \ast\!\!\rightarrow Z$ (by definition). Therefore $Z \Rightarrow Y$, and so (acyclicity) also $Y \not\Leftarrow Z$, but also $Y \not\Leftarrow \mathbf{S}$, otherwise (transitivity) $Z \Rightarrow \mathbf{S}$. Therefore, if Y has an edge to Z in $\mathcal{P}$, then it has an arrowhead mark at Y.

The proof that they are also complete follows from the lemmas below, by induction on the graphical orientation rules $\mathcal{R}0\text{b}-\mathcal{R}4\text{b}$, showing that none of them introduces a violation of Lemma 3.2. As these rules are sufficient for arrowhead completeness [Ali *et al.*, 2005; Zhang, 2008], it follows that the theorem holds for all invariant arrowheads. □

The next lemmas show that none of the arrowhead orientation rules in Table 2.1 introduce a violation of Proposition 3.2.

Lemma 3.21. *The arrowheads at Z from rules $\mathcal{R}0b$, $\mathcal{R}3$, and $\mathcal{R}4b$ are covered by case* (1) *and the arrowhead at Y from rule $\mathcal{R}1$ is covered by case* (2)*.*

Proof. Implied directly by the corresponding patterns, see also Figure 2.2:

$\mathcal{R}0$b: If this rule fires, then it implies $X \perp\!\!\!\perp Y \mid [\mathbf{W}]$ for some set $\mathbf{W}$ (possibly empty), with $X \not\perp\!\!\!\perp Y \mid \mathbf{W} \cup Z$. Therefore case (1) applies and Z gets arrowheads on the edges from X and Y in $\mathcal{G}$, just as in the consequent of $\mathcal{R}0$b in fig.2.2.

$\mathcal{R}1$: Implies $X \perp\!\!\!\perp Y \mid [\mathbf{Z}]$ with $Z \in \mathbf{Z}$ and an arrowhead at Z from a rule that fired before. If no violations before $\mathcal{R}1$ fires, then case (2) applies, and there is an arrowhead at $Z \longrightarrow Y$ in $\mathcal{G}$, just as in the consequent of $\mathcal{R}1$.

$\mathcal{R}3$: Implies $X \perp\!\!\!\perp Y \mid [\mathbf{W}]$ with $W \in \mathbf{W}$, and $X \not\perp\!\!\!\perp Y \mid \mathbf{W} \cup Z$. Therefore case (1) applies, to give $W \ast\!\!\rightarrow Z$ in $\mathcal{G}$, just as in $\mathcal{R}3$.

$\mathcal{R}$4b: By construction of the discriminating path, $\mathcal{R}$4b implies $X \perp\!\!\!\perp Y \,|\, [\mathbf{Z}]$, with $\{Z_1, \ldots, Z_k\} \in \mathbf{Z}$, but $Z \notin \mathbf{Z}$ as $X \not\perp\!\!\!\perp Y \,|\, \mathbf{Z} \cup Z$. Therefore case (1) applies, resulting in the addition of $Z_k \ast\!\!\rightarrow Z \leftarrow\!\!\ast\, Y$ to $\mathcal{G}$, just as in $\mathcal{R}$4b.

□

Lemma 3.22. *The arrowheads at Y from rules $\mathcal{R}$2b, $\mathcal{R}$4a, and $\mathcal{R}$4b are covered by cases* (1) *and* (2)*.*

Proof. First $\mathcal{R}$2b. The arrowhead at $Z \ast\!\!\rightarrow X$ either appeared by case (1) as a node X that creates the dependency $U \not\perp\!\!\!\perp V \,|\, \mathbf{W} \cup X$ from $U \perp\!\!\!\perp V \,|\, [\mathbf{W}]$, with $Z \in \{U, V, \mathbf{W}\}$ (case 1a), or by case (2), as a minimal conditional independence $X \perp\!\!\!\perp U \,|\, [\mathbf{W} \cup Z]$, with a (somehow) established $Z \not\twoheadrightarrow (U \cup \mathbf{S})$, for which either U and Y are also independent given $\mathbf{W} \cup Z$ (case 2a), or not (case 2b). (Note: $Y \notin \mathbf{W}$ in case (2), otherwise (from $Y \not\twoheadrightarrow X$, lemma 2) $Y \Rightarrow U$, which, together with $Z \Rightarrow X$ and $X \Rightarrow Y$, would imply $Z \Rightarrow U$). For these three instances:

1a) If conditioning on X creates $U \not\perp\!\!\!\perp V \,|\, \mathbf{W} \cup X$, then conditioning on Y as a descendant of X implies $U \not\perp\!\!\!\perp V \,|\, \mathbf{W} \cup Y$, and so case (1) also applies to $Z \ast\!\!\rightarrow Y$.

2a) If $Y \perp\!\!\!\perp U \,|\, \mathbf{W} \cup Z$, then also $Y \perp\!\!\!\perp U \,|\, [\mathbf{W} \cup Z]$, as no subset can block the path between Y and U via X, and so case (2) applies to $Z \longrightarrow Y$.

2b) If $Y \not\perp\!\!\!\perp U \,|\, \mathbf{W} \cup Z$, then there is an unblocked path π between U and Y given $\mathbf{W} \cup Z$. The path π is *into* Y, since otherwise the path $\langle X, Y \rangle + \pi$ would be an unblocked path between X and U given $\mathbf{W} \cup Z$, contrary to $X \perp\!\!\!\perp U \,|\, [\mathbf{W} \cup Z]$. Therefore, conditioning on collider Y on the path creates the dependency $X \not\perp\!\!\!\perp U \,|\, \mathbf{W} \cup Z \cup Y$, and so case (1) applies.

In $\mathcal{R}$4a and $\mathcal{R}$4b, the arrowhead at Y is simply an instance of $\mathcal{R}$2b with $Z_k = X$. □

This leaves rule $\mathcal{R}$2a as the only remaining case to prove. For that we use the observation:

Lemma 3.23. *If two nodes X and Y are conditionally independent given a set of nodes $\mathbf{Z}$, $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$, then an arbitrary node V is either:*

(a) part of the conditional independence, i.e. $V \in (X \cup Y \cup \mathbf{Z})$,

(b) conditionally independent of X and/or Y given $\mathbf{Z}$, i.e. logical statement $(V \perp\!\!\!\perp X \,|\, \mathbf{Z}) \vee (V \perp\!\!\!\perp Y \,|\, \mathbf{Z})$, or

(c) (descendant of) a collider between U and V such that $X \not\perp\!\!\!\perp Y \,|\, \{\mathbf{Z} \cup V\}$.

Proof. If neither (a) nor (b), i.e. $V \notin (X \cup Y \cup \mathbf{Z})$ *and* $V \not\perp\!\!\!\perp \{X, Y\} \,|\, \mathbf{Z}$, then there are paths $\pi_X = \langle X, \ldots, V \rangle$ and $\pi_Y = \langle Y, \ldots, V \rangle$ in the corresponding graph that are unblocked given $\mathbf{Z}$. Node V has to be a collider on the path $\pi = \pi_X + \pi_Y$, otherwise π would be unblocked given $\mathbf{Z}$ (as $V \notin \mathbf{Z}$), contrary to $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$. But then conditioning on $\mathbf{Z} \cup V$ will make them dependent, i.e. then (c). □

Note that if $\mathbf{Z}$ is a *minimal* set that makes X and Y independent, then case (b) does not imply that it is also minimal for $V \perp\!\!\!\perp X/Y \,|\, \mathbf{Z}$, as shown by the example in Figure 3.2: from $X \perp\!\!\!\perp Y \,|\, [\{Z_1, Z_2\}]$, for node V we find $V \perp\!\!\!\perp X \,|\, \{Z_1, Z_2\}$ (as none of the other options in Lemma 3.23 applies), but this is only *minimal* for subset $V \perp\!\!\!\perp X \,|\, [Z_2]$.

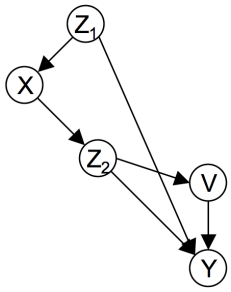

Figure 3.2: Example of case (b) in Lemma 3.23 with 'minimal' only for subset

For the proof of $\mathcal{R}$2a we also use:

Lemma 3.24. *In an ancestral graph $\mathcal{G}$, if a node Z unblocks a blocked path $\pi = \langle U, \ldots, V \rangle$ between two nodes U and V given some set $\mathbf{W}$, then there are unblocked paths from both U and V into Z relative to $\mathbf{W}$, and so $Z \not\perp\!\!\!\perp \{U, V\} \,|\, \mathbf{W}$.*

Proof. By definition, a path π is unblocked relative to $\mathbf{W}$ if all noncolliders on the path are not in $\mathbf{W}$ and all colliders are in $An(\mathbf{W})$. Adding a node Z to the conditioning set can never remove a noncollider, so it can only unblock on a collider that is in $An(\mathbf{W} \cup Z)$, but not in $An(\mathbf{W})$. So the node Z must be (a descendant of) a collider C on the path (possibly $C = Z$). No node $W \in \mathbf{W}_{\setminus C}$ blocks the path $X \Rightarrow Z$ (otherwise conditioning on Z would not be needed), therefore if π is unblocked relative to $\mathbf{W}$, then so are the two paths $\pi_U = \langle U, ..(, X, ..), Z \rangle$ and $\pi_V = \langle Z, (.., X,)..., V \rangle$, which implies $Z \not\perp\!\!\!\perp \{U, V\} \,|\, \mathbf{W}$. □

Finally we need the following result (see also Figure 3.3, below):

Lemma 3.25. *In an ancestral graph $\mathcal{G}$, if there are (sets of) nodes U, Y, Z and $\mathbf{W}$, such that $U \perp\!\!\!\perp Z \,|\, \mathbf{W}$ and $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W} \cup Y$, with $Z *\!\!\rightarrow Y$ in $\mathcal{G}$, then there is a node $W \in (U \cup \mathbf{W})$, and a set $\mathbf{Q} \subseteq \mathbf{W}$, such that $W \perp\!\!\!\perp Z \,|\, [\mathbf{Q}]$ and $W \not\perp\!\!\!\perp Z \,|\, \mathbf{Q} \cup Y$.*

Proof. In words: *if conditioning on a node Y destroys some conditional independence for a neighbouring node Z (unblocks a path), then the same holds for at least some minimal conditional independence between Z and one of the other nodes involved.*
By definition, there is a $\mathbf{W}' \subseteq \mathbf{W}$ such that $U \perp\!\!\!\perp Z \,|\, [\mathbf{W}']$. If then also $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W}' \cup Y$, then the lemma applies with $W = U$ and $\mathbf{Q} = \mathbf{W}'$. If not, i.e. if $U \perp\!\!\!\perp Z \,|\, \mathbf{W}' \cup Y$,

then we can show that there is a node $W \in \mathbf{W}$ for which the lemma holds.
Let $\mathcal{G}'$ be the MAG obtained from $\mathcal{G}$ by marginalizing out all nodes in $\mathcal{G}$ that are not in $\{U, Y, Z\} \cup \mathbf{W}$. From the original $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W} \cup Y$, by Lemma 3.24, there is an unblocked path $\pi = \langle U, \ldots, Y \rangle$ in $\mathcal{G}'$ that is into Y given $\mathbf{W}$. The path π contains one or more (say k) colliders in $\mathcal{G}'$, some of which are (ancestors of) nodes from $\mathbf{W}$, but not from $\mathbf{W}'$ (otherwise the path to Y would also be unblocked given $\mathbf{W}'$, which, together with edge $Z \ast\!\!\rightarrow Y$, would imply $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W}' \cup Y$, contrary to the assumed). Number the colliders as $W_1, \ldots, W_k$, as they are encountered along π when starting from Y, such that $\pi = U \ast\!\!\rightarrow W_k \longleftrightarrow \ldots \longleftrightarrow W_2 \longleftrightarrow W_1 \longleftrightarrow Y$. By induction: if there is no edge between W_1 and Z in $\mathcal{G}'$, then they are (minimally) conditionally independent given some set $\mathbf{Q}_1 \subset \mathbf{W}$ (possibly empty), but dependent given Y, as the paths from both W_1 and Z into Y are not blocked by any node from $\mathbf{W}$ (as a bi-directed edge in $\mathcal{G}'$, resp. (direct) edge into Y), and so the lemma is satisfied. If not, i.e. if there is an edge in $\mathcal{G}'$, then this edge is *out of* W_1, otherwise the path $\langle U, W_k, .., W_1, Z \rangle$ would be unblocked relative to $\mathbf{W}$, making U and Z dependent given $\mathbf{W}$, contrary the given. But then for W_2, if there is no edge between W_2 and Z in $\mathcal{G}'$, then $W_2 \perp\!\!\!\perp Z \,|\, [\mathbf{Q}_2]$, with $W_1 \in \mathbf{Q}_2$, because it is the only node from $\mathbf{W}$ that blocks the trek $W_2 \longleftrightarrow W_1 \longrightarrow Z$. But that also means that the path from W_2 to Y is unblocked given $\mathbf{Q}_2$, and so $W_2 \not\perp\!\!\!\perp Z_2 \,|\, \mathbf{Q}_2 \cup Y$. If not, then the edge to Z is (again) *out of* W_2, otherwise $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W}$, contrary the given. This applies to all successive colliders W_i on the path π. But if all, up to and including W_k, have an edge in $\mathcal{G}'$ into Z, then no unblocked path between U and Z implies that W_k is needed to block $U \ast\!\!\rightarrow W_k \longrightarrow Z$, and so *all* W_i on π are in $\mathbf{W}'$, implying an unblocked path to Y, and so $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W}' \cup Y$. □

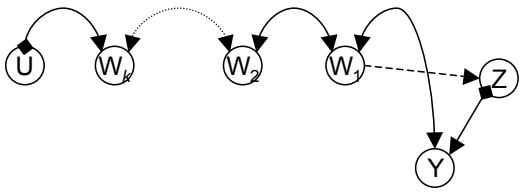

Figure 3.3: Configuration for Lemma 3.25

Now at last we can show:

Lemma 3.26. *The arrowhead at Y from $\mathcal{R}2a$ is covered by cases* (1) *and* (2).

Proof. If the arrowhead at $X \ast\!\!\rightarrow Y$ originates from case (2), then the edge appears as $X \longrightarrow Y$, and is therefore also an instance of $\mathcal{R}2$b, which we already found to be valid. If $X \ast\!\!\rightarrow Y$ originates from case (1), then there is a minimal $U \perp\!\!\!\perp V \,|\, [\mathbf{W}]$, with $X \in (U \cup V \cup \mathbf{W})$, and the node Y creates $U \not\perp\!\!\!\perp V \,|\, \mathbf{W} \cup Y$. By Lemma 3.23 there are now three cases for node Z:

(a) $Z \in (U \cup V \cup \mathbf{W})$,

(b) $Z \perp\!\!\!\perp U \,|\, \mathbf{W}$, (and/or $Z \perp\!\!\!\perp V \,|\, \mathbf{W}$)

(c) $U \not\perp\!\!\!\perp V \,|\, \mathbf{W} \cup Z$.

For case (a), both X and Z are in $(U \cup V \cup \mathbf{W})$, and so if rule (1) applies to $X * \!\!\rightarrow Y$ it also applies to $Z * \!\!\rightarrow Y$. Case (c) cannot occur, as that would imply $Z \not\twoheadrightarrow (U \cup V \cup \mathbf{W} \cup \mathbf{S})$ by Lemma 3.1, with $X \in (U \cup V \cup \mathbf{W})$, while $\mathcal{R}$2a has $Z \rightarrow X$.
For the remaining case (b), w.l.o.g. we assume $U \perp\!\!\!\perp Z \,|\, \mathbf{W}$. Lemma 3.24 implies $U \not\perp\!\!\!\perp Y \,|\, \mathbf{W}$ which, together with $Z * \!\!\rightarrow Y$, implies $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W} \cup Y$, because the unblocked path from U to Y given $\mathbf{W}$ cannot contain Z, as that would create an unblocked path from U via Z to X given $\mathbf{W}$, contrary $U \perp\!\!\!\perp Z \,|\, \mathbf{W}$. Then from Lemma 3.25 it follows that there is at least one minimal conditional independence between Z and some node from $(U \cup \mathbf{W})$ that is destroyed by conditioning on Y. Therefore, the arrowhead $Z * \!\!\rightarrow Y$ is then covered by case (1). □

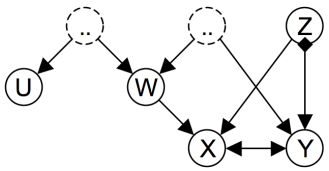

Figure 3.4: Example of non-minimal case (b) in Lemma 3.26.

Example 3.27. *Figure 3.4 shows an instance of case (b) for $\mathcal{R}$2a where the initial separating set is not minimal. Here, $\mathcal{R}$2a applies to $Z * \!\!\rightarrow Y$, after $X * \!\!\rightarrow Y$ is derived via case (1) from $U \perp\!\!\!\perp X \,|\, [W]$ with $U \not\perp\!\!\!\perp X \,|\, W \cup Y$ (the origin of the edge $Z \longrightarrow X$ is not depicted). By Lemma 3.23, for node Z indeed $U \perp\!\!\!\perp Z \,|\, W$ holds (case b), but not as a minimal independence, as $U \perp\!\!\!\perp Z \,|\, [\varnothing]$. As a result, edge $Z * \!\!\rightarrow Y$ does not follow from case (1) applied to this combination of nodes as conditioning on Y does not make U and Z dependent, i.e. $U \perp\!\!\!\perp Z \,|\, Y$. However, as in the proof of Lemma 3.26, Z is minimally conditionally independent of 'eliminated' node W, but dependent when conditioning on Y. Therefore, case (1) applies to $W \perp\!\!\!\perp Z \,|\, [\varnothing]$ and $W \not\perp\!\!\!\perp Z \,|\, Y$, from which follows that $Z * \!\!\rightarrow Y$.*

We can now complete the proof of the main invariant arrowhead result:

Proof of Proposition 3.2. Follows from the arrowhead completeness of rules $\mathcal{R}$0b-$\mathcal{R}$4b, the fact that after $\mathcal{R}$0a the theorem holds (no arrowheads), in combination with the proof in lemmas 3.21-3.26 that none of the rules $\mathcal{R}$0b-$\mathcal{R}$4b introduces a violation of Proposition 3.2. □

Note that rule (1) covers all $X \circ\!\!\rightarrow Y$ and $X \leftrightarrow Y$ edges (see section on reading PAGs). Next we continue with the invariant tails.

Proposition 3.13 (Invariant tails). *In a PAG $\mathcal{P}$, all invariant tails $Z - * Y$ from graphical orientation rules $\mathcal{R}$4a, $\mathcal{R}$5, $\mathcal{R}$7, $\mathcal{R}$9, and $\mathcal{R}$10 are instances of:*

(2b): $X \perp\!\!\!\perp Y \mid [\mathbf{W} \cup Z]$, with $X \Rightarrow (Z \cup \mathbf{S})$ from either case (3) or another (2b),

(3): $U \perp\!\!\!\perp V \mid [\mathbf{W} \cup W]$, with two transitive relations $[W, U, .., Y] + [W, V, .., Y]$, and $Z \in \{U, V, W\}$,

(4): $X \perp\!\!\!\perp Y \mid [\mathbf{Z}]$, with inferred blocking node $Z \in \mathbf{Z}$, together with $Z_k \Rightarrow (Y \cup \mathbf{S})$ from either case (2) or case (4).

Proof. Case (2b) covers rule $\mathcal{R}$7, and is so named because of its similarity/overlap with case(2) for $\mathcal{R}$1. Case (3) covers all instances of rules $\mathcal{R}$5, $\mathcal{R}$9, and $\mathcal{R}$10, and case (4) accounts for tails from orientation rule $\mathcal{R}$4a.
All three cases are sound:

(2b): By Lemma 3.1, $X \perp\!\!\!\perp Y \mid [\mathbf{W} \cup Z]$ gives $(Z \Rightarrow X) \vee (Z \Rightarrow Y) \vee (Z \Rightarrow \mathbf{S})$. Combined with $(X \Rightarrow Z) \vee (X \Rightarrow \mathbf{S})$ this gives $(Z \Rightarrow Y) \vee (Z \Rightarrow \mathbf{S})$, and so a tail at Z if it has an edge to Y in $\mathcal{P}$.

(3): Idem, $U \perp\!\!\!\perp V \mid [\mathbf{W} \cup W]$ gives $(W \Rightarrow U) \vee (W \Rightarrow V) \vee (W \Rightarrow \mathbf{S})$. From Corollary 3.12, the transitive relations give $(W \Rightarrow \{U \cup \mathbf{S}\}) \vdash (W \Rightarrow \{Y \cup \mathbf{S}\})$, and $(W \Rightarrow \{V \cup \mathbf{S}\}) \vdash (W \Rightarrow \{Y \cup \mathbf{S}\})$. Substituting these two in the first then gives $(W \Rightarrow Y) \vee (W \Rightarrow \mathbf{S})$. This holds for all nodes on the two transitive chains, hence if $Z \in \{U, V, W\}$, then $(Z \Rightarrow Y) \vee (Z \Rightarrow \mathbf{S})$, and therefore a tail $Z - * Y$, if they are connected in $\mathcal{P}$.

(4): By Lemma 3.8, as Z is an inferred blocking node between X and Y given $\mathbf{Z}$, there is a $Z_k \in \mathbf{Z}$ such that $Z \Rightarrow Z_k \vee Z \Rightarrow Y \vee Z \Rightarrow \mathbf{S}$. Together with the given $Z_k \Rightarrow Y \vee Z_k \Rightarrow \mathbf{S}$, this reduces to $Z \Rightarrow Y \vee Z \Rightarrow \mathbf{S}$, and hence an invariant tail $Z \longrightarrow Y$.
In rule $\mathcal{R}$7, X and Y are nonadjacent, so conditionally independent given some set, and Z as a noncollider between the two is needed in all such sets, hence $X \perp\!\!\!\perp Y \mid [\mathbf{W} \cup Z]$. Only rules $\mathcal{R}$6 and $\mathcal{R}$7 can produce the required $X - \circ Z$ edge to trigger $\mathcal{R}$7, however, every (chain of) $\mathcal{R}$7 orientations needs to start from an instance of $\mathcal{R}$6. Rule $\mathcal{R}$6 implies identifiable selection bias on X (undirected edges to other nodes), and so, if it triggers $\mathcal{R}$7 then this satisfies case (2b), and therefore any subsequent tail oriented by $\mathcal{R}$7 as well.

Rule $\mathcal{R}$5 triggers on an uncovered circle path. In Figure 2.3, let U be the node next to X on the circle path (U could be W, in which case dashed line $X - W$ becomes edge), so that we have $Z \perp\!\!\!\perp U \mid [X \cup ..]$, as Z and U by definition not adjacent. Furthermore, there are two transitive relations $[X, U, .., W, Y]$ and $[X, Z, Y]$ that are both from X to Y, and so $\mathcal{R}$5 satisfies case (3) and gives (among others) $Z - Y$.

Rule $\mathcal{R}9$ similarly: now $Z \perp\!\!\!\perp W \mid [X \cup ..]$, with two transitive relations $[X, W, .., Y]$ and $[X, Z, Y]$ that are both from X into Y, and so $\mathcal{R}9$ satisfies case (3) and will orient $Z \!-\!\!- Y$.

Rule $\mathcal{R}10$ idem: now $V \perp\!\!\!\perp S \mid [Z \cup ..]$, with two transitive relations $[Z, S, .., X, Y]$ and $[Z, V, .., W, Y]$ that are both from Z into Y, and so $\mathcal{R}10$ satisfies case (3) and will orient $Z \!-\!\!- Y$.

In rule $\mathcal{R}4$a, from the description of the graphical orientation, it follows that X and Y are non-adjacent in $\mathcal{P}$, and that all nodes $Z_1, .., Z_k, Z$, see also Figure 2.2, are needed to make them independent, and hence appear in the set $X \perp\!\!\!\perp Y \mid [\mathbf{Z} \cup Z]$. Furthermore, each neighboring node in the sequence is adjacent in the graph, so not separated by any set, let alone a subset from $\mathbf{Z}$. All nodes Z_i have arrowheads at edges to their neighbors in the sequence, implying non-ancestorship, so no causal relation to either. Therefore Z is an inferred blocking node. The tail at $Z_k \longrightarrow Y$ implies $Z_k \Rightarrow (Y \cup \mathbf{S})$, and so $\mathcal{R}4$a satisfies case (4) (in fact, even stronger, as identifiable $Z_k \Rightarrow Y$).

By construction of the discriminating path, all nodes Z_i in the sequence, except perhaps Z_1, also satisfy the conditions in case (4). For Z_2, the arc $Z_1 \longrightarrow Y$ follows from case (2). For Z_3, the invariant arc $Z_2 \longrightarrow Y$ therefore satisfies case (4) (although it may also be derived in other ways as well). Similar for all subsequent nodes up to Z_k. □

Corollary 3.14 (Identifiable selection). *In a PAG $\mathcal{P}$, all identifiable selection nodes $X \Rightarrow \mathbf{S}$ are covered by case (3), in the form of a minimal independence with two transitive relations back to itself.*

Proof. Identifiable selection bias $X \Rightarrow \mathbf{S}$ corresponds to a node with an undirected edge. Only $\mathcal{R}5$, $\mathcal{R}6$, and $\mathcal{R}7$ can produce undirected edges. When the transitive relations reach all the way back to the node Z from the initial minimal conditional independence, then the conclusion becomes $(Z \Rightarrow Z) \vee (Z \Rightarrow \mathbf{S})$, which (irreflexivity) reduces to $(Z \Rightarrow \mathbf{S})$. In other words, then there is *identifiable selection bias* on Z, and therefore also on all other nodes involved in the transitive relation (including the $\mathbf{U}_i$s).

This is what happens in $\mathcal{R}5$. Afterwards, if $\mathcal{R}6$ is always executed before $\mathcal{R}7$ when a new undirected edge is found, then $\mathcal{R}6$ will never (need to) identify a new selection node, as it produces only tails on nodes that are already established selection nodes.

That leaves just $\mathcal{R}7$. It is possible that part of the transitive relation in case (3) is traversed in both ways (that is where the 'not necessarily disjoint' part in the definition of transitive relation comes in). This occurs for single nodes that separate two nonchordal undirected subgraphs in $\mathcal{P}$. Then $\mathcal{R}7$ will orient $Z \!-\!\!* Y$ in the direction away from one undirected subgraph, and $Y \!-\!\!* Z$ when orienting in the direction away from the other subgraph, resulting in $Y \!-\!\!- Z$, and so identifiable selection bias on both Y and Z. These are also the *only* ways in which undirected edges can be created by the orientation rules. □

3.D Proofs: LoCI and the complete PAG

Lemma 3.16 (Single minimal independence). *In the logical causal inference (LoCI) approach, finding a single, arbitrary $X \perp\!\!\!\perp Y \,|\, [\mathbf{Z}]$, for each pair of nodes (X, Y) in the graph (if it exists) is sufficient to find all invariant features of the PAG.*

Proof sketch. This stems from the fact that the graphical orientation rules are defined on sets of adjacent nodes, which ensures that most nodes are almost always needed to separate two nonadjacent nodes in the same rule, and so will be found as part of the separating set, no matter how large/variable the set of nodes to block all paths between the two can be. Note: once a minimal set is found for a pair of nodes, then all remaining nodes are checked to see if including them destroys the independence (so lemma 3, item 1 applies). □

In the proof of rule $\mathcal{R}$2a we use that if conditioning on a node Y destroys (unblocks) some minimal conditional independence for a neighbouring node Z, then it does so in *all* minimal independencies between Z and at least one node in $\mathcal{G}$:

Lemma 3.28. *In an ancestral graph $\mathcal{G}$, if there are (sets of) nodes U, Y, Z and $\mathbf{W}$, such that $U \perp\!\!\!\perp Z \,|\, [\mathbf{W}]$ and $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W} \cup Y$, with $Z \ast\!\!\rightarrow Y$ in $\mathcal{G}$, then there is a node V (possibly $V = U$), such that for all sets $\mathbf{Q}$ for which $V \perp\!\!\!\perp Z \,|\, [\mathbf{Q}]$ it holds that $V \not\perp\!\!\!\perp Z \,|\, \mathbf{Q} \cup Y$.*

Proof. Follows along the lines of Lemma 3.25. From that proof, there is an unblocked path π in $\mathcal{G}'$ (the graph $\mathcal{G}$, marginalized over $\{U, Y, Z\} \cup \mathbf{W}$, of the form $\pi = U \ast\!\!\rightarrow W_k \longleftrightarrow \ldots \longleftrightarrow W_2 \longleftrightarrow W_1 \longleftrightarrow Y$ that is into Y given $\mathbf{W}$, for some $k \geq 0$. Here we consider the corresponding path(s) θ in $\mathcal{G}$, where the colliders on the path are now indicated by $\{U_1, .., U_m\}$. (Note that edges in $\mathcal{G}'$ may correspond to multiple unblocked paths relative to $\mathbf{W}$ in $\mathcal{G}$, and that θ may contain different nodes than π, including other (ancestors of) colliders from $\mathbf{W}$).

So, let $\theta = U \ast\!\!\rightarrow U_m \longleftrightarrow \ldots \longleftrightarrow U_2 \longleftrightarrow U_1 \longleftrightarrow Y$ be an unblocked path in $\mathcal{G}$ that is into Y given $\mathbf{W}$. We look at nodes V along θ, starting from Y, and try to find one that does not have a link to Z. Suppose $V = V_1$ is encountered on the first leg (trek) $U_1 \longleftarrow\!\ast \,.. \ast\!\!\rightarrow Y$. If V_1 does not have an edge to Z in $\mathcal{G}$, then there is some $\mathbf{Q}$ such that $V_1 \perp\!\!\!\perp Z \,|\, [\mathbf{Q}]$, while also $V_1 \not\perp\!\!\!\perp Z \,|\, \mathbf{Q} \cup Y$ (both edges to Y), and so the lemma is satisfied. If there is a link, then it can only be of the form $U_1 \longleftarrow\!\ast \,.. \ast\!\!\rightarrow V_1 \longleftarrow\!\ast Z$, otherwise there would be an unblocked path between U and Z given $\mathbf{W}$.

For a second node, V_2, a similar story holds: if there is no edge $V_2 - Z$ in $\mathcal{G}$, then there is some $\mathbf{Q}$ such that $V_2 \perp\!\!\!\perp Z \,|\, [\mathbf{Q}]$; but note that now $V_1 \notin \mathbf{Q}$, as the edges from both V_2 and Z are into V_1, and therefore also/again $V_2 \not\perp\!\!\!\perp Z \,|\, \mathbf{Q} \cup Y$ (unblocked path resp. edge into Y given $\mathbf{Q}$), and so the lemma is satisfied. If there is an edge, then again it must be of the form $U_1 \longleftarrow\!\ast \,.. \ast\!\!\rightarrow V_2 \longleftarrow\!\ast Z$, otherwise there would be an unblocked path to Z, and we can continue until we reach U_1. At that point, again, if U_1 has no edge to Z, then there is some $U_1 \perp\!\!\!\perp Z \,|\, [\mathbf{Q}]$, with none of

the V_i encountered on the first leg in $\mathbf{Q}$ (as V_j is not ancestor of either U_1 or Z, from which, by $V_j \longrightarrow .. \longrightarrow V1$, follows that neither are any of the other V_i), and so there are unblocked paths from U_1 and Z into Y given $\mathbf{Q}$. If it has an edge, then, contrary the edges from Z into V_i (if any), it must be an edge *out of* $U_1 \longrightarrow Z$, otherwise there would be an unblocked path given $\mathbf{W}$.

Continuing with the second leg, $U_2 \longleftarrow * .. * \longrightarrow U_1$, we now find that a node V encountered in going from U_1 to U_2 along θ (if any) *cannot* have a direct edge to Z without creating an unblocked path from U to Z relative to $\mathbf{W}$ (either as noncollider between U_2 and Z, or as collider that is ancestor of U_1). So then $V \perp\!\!\!\perp Z \,|\, [\mathbf{Q}]$, with $U_1 \in \mathbf{Q}$ (as it is the only node blocking the path $V * \!\!\rightarrow U_1 \longrightarrow Z$). But like before, none of the previous V_i (if any) are part of $\mathbf{Q}$: by contradiction, if $V_i \in \mathbf{Q}$, then $V_i \longrightarrow V$ (because $V_i \notin An(Z)$), so if $U_2 \longleftrightarrow V \longrightarrow U_1$ along θ, then the 'top' V_j would be ancestor of U_1, contrary the arrowhead $U_1 \longleftrightarrow V_j$, and if $U_2 \longleftarrow V * \!\!\rightarrow U_1$ along θ, then the path via $U_2 \longleftarrow V \longleftarrow V_i \longleftarrow * Z$ would be unblocked (end-of-by-contradiction). So, if none of the V_i are part of $\mathbf{Q}$, then the paths from both V and Z to Y are unblocked given $\mathbf{Q}$, and the lemma is satisfied.

If no node V on $U_2 \longleftrightarrow U_1$, then again, for U_2 there is either a minimal independence that satisfies the lemma or an edge $U_2 \longrightarrow Z$. This can be repeated along θ until a node is found or the final node U is reached. At that point, if no other node has been found before, it can be applied to any set for which $U \perp\!\!\!\perp Z \,|\, [\mathbf{Q}]$, as all other colliders $U_i \in \mathbf{Q}$, are needed to separate U and Z, but none of the V_i, and so the path to Y is always unblocked. □

To prove Lemma 3.16 it is actually easier to use a more restricted variant of Proposition 3.2, where case (1) only needs to be applied to instances where Z is **always** part of the minimal conditional independence, and that will always be unblocked when conditioning on Y; idem for case (2):

Lemma 3.29. *In a PAG $\mathcal{P}$, all invariant arrowheads $Z * \!\!\rightarrow Y$ are instances of:*

case (1′): $U \not\perp\!\!\!\perp V \,|\, \mathbf{W} \cup [Y]$, $Z \in \{U, V, \mathbf{W}\}$, *and for all sets* $\mathbf{W}' : U \perp\!\!\!\perp V \,|\, [\mathbf{W}']$ *the paths from U and V to Y are unblocked relative to $\mathbf{W}'$, and either $Z \in \{U, V\}$ or (necessarily) $Z \in \mathbf{W}'$,*

case (2′): $X \perp\!\!\!\perp Y \,|\, [\mathbf{W}]$ *with* $Z \in \mathbf{W}$, *and* $Z \not\twoheadrightarrow (X \cup \mathbf{S})$ *from either case (1′) or case (2′), and Z in all sets* $\mathbf{W}' : X \perp\!\!\!\perp Y \,|\, [\mathbf{W}']$.

Proof. We now show this holds for each arrowhead rule:

$\mathcal{R}$0b fires on any (minimal) conditional independence $X \perp\!\!\!\perp Y \,|\, \mathbf{W}$ between X and Y, and for any such $\mathbf{W}$, including Z will unblock the path $\langle X, Z, Y \rangle$, so case (1′) applies,

$\mathcal{R}$1 node Z is part of *any* set (minimal or not) that separates X and Y, and so case (2′) applies,

$\mathcal{R}$3 similar to $\mathcal{R}$0b, fires on a node W that is part of all sets separating X and Y, and including Z will unblock the path $\langle X, Z, Y \rangle$, and so case (1′) applies,

$\mathcal{R}$4b (arrowheads at Z) all nodes $Z_1, \ldots, Z_k$ are part of all sets separating X and Y, and including Z then makes them dependent, so case (1′) applies,

$\mathcal{R}$2b for instance (1a) in Lemma 3.22, if case (1′) applies to $Z \ast\!\!\rightarrow X$, then it also applies to $Z \ast\!\!\rightarrow Y$, as X is never part of the minimal conditional independence involving Z, and so unblocked paths to X imply unblocked paths to Y; for instance (2a), Z is present in all sets that make X and U independent, and so also in all sets that make Y and U independent (as it implies $Z \longrightarrow Y$), and so case (2′) applies; for instance (2b), if $Y \not\perp\!\!\!\perp U \,|\, [\mathbf{W} \cup Z]$ holds for all sets for which $X \perp\!\!\!\perp U \,|\, [\mathbf{W} \cup Z]$, then it is an instance of case (1′) with $V \equiv X$ and $\mathbf{W} \equiv (\mathbf{W} \cup Z)$, and so for all $\mathbf{W} \cup Z$ there are unblocked paths from X and U into Y, resulting in $X \not\perp\!\!\!\perp U \,|\, \mathbf{W} \cup Z \cup Y$. If not, then there is *some* $\mathbf{W}'$ for which $X \perp\!\!\!\perp U \,|\, [\mathbf{W}' \cup Z]$ and *not* $Y \not\perp\!\!\!\perp U \,|\, [\mathbf{W}' \cup Z]$. But as Z is needed in all sets that block a path $\pi = U .. \ast\!\!\rightarrow Z \longrightarrow X$ between U and X, it means that Z is also needed in all sets that separate U and Y, because if there is any remaining unblocked path π from U to either X or Z, then $\pi+$ either $X \longrightarrow Y$ or $Z \longrightarrow Y$ is an unblocked path from U into Y. Therefore Z is also needed in all sets that separate U and Y, which, together with $Z \not\twoheadrightarrow (U \cup \mathbf{S})$, implies that it is an instance of case (2′).

$\mathcal{R}$4a/b (arrowhead at Y) instances of $\mathcal{R}$2b with $Z_k = Y$,

$\mathcal{R}$2a if the arrowhead between X and Y originates from case (2′) then $X \longrightarrow Y$, and so is an instance of $\mathcal{R}$2b. If not, then the arrowhead $X \ast\!\!\rightarrow Y$ originates from case (1′) with node X in $\mathcal{R}$2a in the role of Z in (1′).

Now, if Z, like X, is also a *necessary* member of the minimal independence $U \perp\!\!\!\perp V \,|\, [\mathbf{W}]$, i.e. $Z \in \{U, V\}$ or $\forall \mathbf{W}', U \perp\!\!\!\perp V \,|\, [\mathbf{W}'] : Z \in \mathbf{W}'$, then case (1′) also applies immediately to Z.

If Z is *not* necessary, then there is some $U \perp\!\!\!\perp V \,|\, [\mathbf{W}']$, with $Z \notin (U \cup V \cup \mathbf{W}')$, for which $U \not\perp\!\!\!\perp V \,|\, \mathbf{W}' \cup Y$. For node Z then instance (b) in Lemma 3.23 applies, say as $U \perp\!\!\!\perp Z \,|\, \mathbf{W}'$, as instance (a) is excluded by the assumed $Z \notin (U \cup V \cup \mathbf{W}')$, and instance (c) still cannot occur, as $Z \longrightarrow X$. But then also $U \not\perp\!\!\!\perp Z \,|\, \mathbf{W}' \cup Y$, as both U and Z have unblocked paths into Y relative to $\mathbf{W}'$ (by Lemma 3.24, applied on the $U \not\perp\!\!\!\perp V \,|\, \mathbf{W}' \cup Y$ from case (1′), together with edge $Z \ast\!\!\rightarrow Y$). But then by Lemma 3.25 there is also a $W \perp\!\!\!\perp Z \,|\, [\mathbf{Q}]$ with $W \not\perp\!\!\!\perp Z \,|\, \mathbf{Q} \cup Y$, and therefore, by Lemma 3.28 there is also some node Q for which for every set $Q \perp\!\!\!\perp Z \,|\, [\mathbf{Q}']$ also $Q \not\perp\!\!\!\perp Z \,|\, \mathbf{Q}' \cup Y$, i.e. then case (1′) is also satisfied for the arrowhead $Z \ast\!\!\rightarrow Y$. □

(Note that $\mathcal{R}$2b is not simply an instance of Lemma 3.28, as that only says that X has a conditional independence with some node $V \in \mathbf{W}$ that will always be destroyed by conditioning on Y, but not that Z is necessarily part of this set.)

So all arrowhead rules are covered by Lemma 3.29. As the two cases in Lemma 3.29 are just a restricted form of the cases in Proposition 3.2, it follows that all rules are also covered by Proposition 3.2 if just a (any) single minimal independence is

found between each pair (if it exists). We now complete the proof that a single minimal independence suffices to find all invariant marks:

Proof of Lemma 3.16. By Lemma 3.29, all invariant arrowheads are instances of cases (1′) and (2′), and so will always be found if at least one minimal conditional independence is found between each pair of nodes (if it exists). For the remaining invariant tails: *For each rule:*

- $\mathcal{R}5$ The circle path corresponds to a transitive chain of minimal conditional independencies where each node is necessarily part of any set that separates its two neighbors.
- $\mathcal{R}7$ Similar to $\mathcal{R}1$.
- $\mathcal{R}9$ Node X is part of every conditional independence between Z and W; the same holds for the successive nodes in both transitive relations from X via Z to Y and from X via W to Y.
- $\mathcal{R}10$ Similar to $\mathcal{R}5/\mathcal{R}9$.

Remaining rules $\mathcal{R}6$, $\mathcal{R}8$a, and $\mathcal{R}8$b do not require any separate independence statement. Therefore all orientation rules that trigger on an instance of case (1) do so for at least one that is part of a set that is present in all minimal conditional independencies for a given pair of nodes. □

Theorem 3.18 *The Logical Causal Inference (LoCI) algorithm is sound and complete for causal discovery from probabilistic independence relations.*

Proof. Soundness follows from the validity of the lemmas 3.5 and 3.8, that produce the logical statements in the list Ł, in combination with the causal logic rules in Proposition 3.4. Completeness follows from the fact that all rules are instances of cases (1)-(4) (Propositions 3.2 and 3.13), for a single, arbitrary minimal independence between nodes and in combination with subsequent dependencies (Lemma 3.16), the fact that all logical inference in each of the cases (1)-(4) is covered by Proposition 3.4, the fact that case (1) and (2) will find all required non-ancestor relations / invariant arrowheads (Proposition 3.2, see also [Zhang, 2008]), needed to obtain the only remaining piece of information (inferred blocking node for case (4) from Lemma 3.8). After running the logical rules on this set of statements to completion, all invariant edge marks have been found and can be transferred to the PAG. □

As a final step we demonstrate that for each case (1)-(4), the logical inference steps in the LoCI algorithm do indeed keep the simple form of the list of statements in section 3.3.1, i.e. either a statement on the absence of a specific causal causal relation, or a disjunction of possible causal relations from one variable to at most two other variables, and/or the selection set **S**.

Corollary 3.30. *Keeping only statements of at most two disjunctive positive (presence) causal relations or a single negative (absence) causal relations in the list Ł is sufficient.*

Proof. We do the proof by demonstration for each of the cases (1)-(4) for invariant arrowheads and tails in Propositions 3.2 and 3.13:

Case (1): Follows directly from Lemma 3.5, item 2.

1:	$Y \not\Rightarrow Z$	$\wedge$	$Y \not\Rightarrow ..$	$\wedge$	$Y \not\Rightarrow \mathbf{S}$

Case (2): A minimal independence in combination with already inferred information on the absence of a causal relation.

1:	$Z \Rightarrow X$	$\vee$	$Z \Rightarrow Y$	$\vee$	$Z \Rightarrow \mathbf{S}$
2:	$Z \not\Rightarrow X$	$\wedge$		$\wedge$	$Z \not\Rightarrow \mathbf{S}$
$\vdash$			$Z \Rightarrow Y$		
3:			$Y \not\Rightarrow Z$		

Note that (1:) becomes obsolete with the consequence $Z \Rightarrow Y$ in the third line, so that the latter effectively *replaces* the first in the list Ł.

Case (2b): Idem, but now in combination with an already inferred causal relation and/or selection bias on a specific node.

1:	$Z \Rightarrow X$	$\vee$	$Z \Rightarrow Y$	$\vee$	$Z \Rightarrow \mathbf{S}$
2:	$X \Rightarrow Z$	$\wedge$		$\wedge$	$X \Rightarrow \mathbf{S}$
$\vdash$	$Z \Rightarrow Z$	$\vee$	$Z \Rightarrow Y$	$\vee$	$Z \Rightarrow \mathbf{S}$
3:			$Z \Rightarrow Y$	$\vee$	$Z \Rightarrow \mathbf{S}$

Again, (3:) replaces (1:) in Ł.

Case (3): A minimal independence for Z, with both alternatives leading to Y.

1:	$W \Rightarrow U$	$\vee$	$W \Rightarrow V$	$\vee$	$W \Rightarrow \mathbf{S}$
2:	$U \Rightarrow W$	$\vee$	$U \Rightarrow Y$	$\vee$	$U \Rightarrow \mathbf{S}$
3:	$V \Rightarrow W$	$\vee$	$V \Rightarrow Y$	$\vee$	$V \Rightarrow \mathbf{S}$
$\vdash$	$W \Rightarrow Y$	$\vee$	$W \Rightarrow V$	$\vee$	$W \Rightarrow \mathbf{S}$
$\vdash$	$W \Rightarrow Y$	$\vee$	$W \Rightarrow Y$	$\vee$	$W \Rightarrow \mathbf{S}$
4:			$W \Rightarrow Y$	$\vee$	$W \Rightarrow \mathbf{S}$

This demonstrates the case for $Z = W$; cases $Z = U/V$ go the same.

Case (4): An inferred blocking node Z, with an earlier/afterwards established causal relation to Y or selection bias.

1:	$Z \Rightarrow Z_k$	$\vee$	$Z \Rightarrow Y$	$\vee$	$Z \Rightarrow \mathbf{S}$
2:			$Z_k \Rightarrow Y$	$\vee$	$Z_k \Rightarrow \mathbf{S}$
3:			$Z \Rightarrow Y$	$\vee$	$Z \Rightarrow \mathbf{S}$

With (3:) replacing (1:) in Ł.

This shows that all cases can be found from straightforward deduction on the simple list structure Ł. □

Chapter 4

Causal discovery from different experiments

A long-standing open research problem is how to use information from ***multiple****, overlapping observational and experimental studies, including background knowledge, to infer new causal relations. It is well-known that simply pooling such data sets together may lead to spurious relations and erroneous conclusions that have no bearing on the actual underlying causal system at the heart of these experiments. Recent developments have shown ways to use multiple, partially overlapping data sets, but they have to rely on the assumption that all data sets originate from essentially* ***identical*** *experiments (with different observed variables).*

In this chapter we present a new approach, embodied by the Multiple Causal Inference (MCI) algorithm, that is the first method that can infer provably valid causal relations in the large sample limit from ***different*** *experiments.*

It extends results from the previous chapter that show that constraint-based causal discovery is decomposable into a candidate pair identification and subsequent elimination step. Introducing the framework of an invariant causal system in different external contexts, a key observation is that these steps can be applied separately from different models. We test the algorithm on a variety of synthetic input model sets to assess its behavior and the quality of the output. It produces fast, reliable, and easily interpretable output. The method shows promising signs that it can be adapted to suit causal discovery in real-world application areas as well.

This chapter is based on: [Claassen and Heskes, 2010b] "Causal discovery in multiple models from different experiments" in "Adv. in Neural Information Processing Systems 23", and [Claassen and Heskes, 2010c] "Learning causal network structure from multiple (in)dependence models", published at the Fifth European Workshop on Probabilistic Graphical Models.

4.1 Introduction

Discovering causal relations from observational data is an important, ubiquitous problem in science. In many application areas there is a multitude of data from different but related experiments. Often the set of measured variables is not the same between trials, or the circumstances under which they were conducted differed, making it difficult to compare and evaluate results, especially when they seem to contradict each other, e.g. when a certain dependency is observed in one experiment, but not in another.

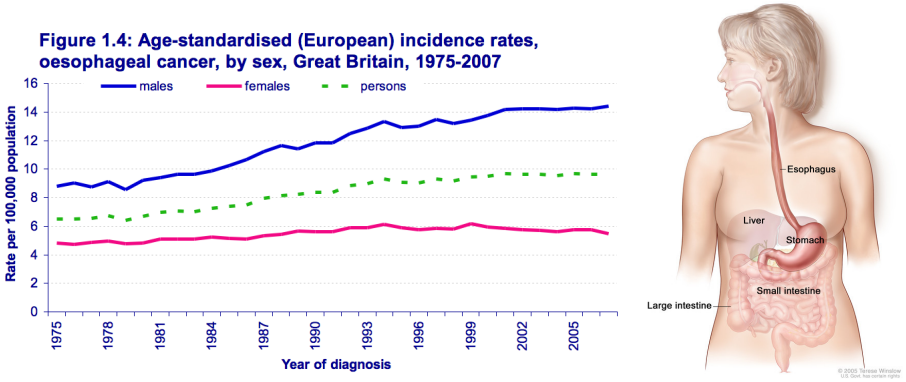

Figure 4.1: Incidence rates of oesophageal cancer (*Source: Cancer Research UK, press release 25 August, 2010*)

Example 4.1. *Oesophageal cancer rates in men have risen by 50 per cent over the last 25 years, see Figure 4.1. It is thought that the rise in obesity is to blame, in combination with people eating less fruit and vegatables. Many related studies are available, but how or why body fat would raise risk is not clear. Key question is: if people exercise more to reduce obestiy, would that reduce the incidence of oesophageal cancer? A complicating factor is that different studies find contradictory information. For example, some show a clear link between drinking hot beverages and squamous cell carcinoma (a type of oesophageal cancer), whereas others do not.*

Results obtained from one data set are often used to either corroborate or challenge results from another, but we would like to learn more from the *combination* of experiments. In general, simply pooling results from different tests into a single large data set is not a good idea: this is long known, see [Yule, 1903], to lead to complications through possible spurious associations, i.e. variables that are independent in two data sets can exhibit an apparent dependency in the combined data set. One option is to present several alternative causal explanations to the end-user, but this is can be confusing and hard to interpret when many models are presented. A pragmatical approach could be to combine the different results through some kind of majority voting, e.g. where arcs in the output are kept if they

appear in all or most models from the separate experiments. But this still cannot adequately explain why sometimes dependencies are observed in one experiment, but not in another. Being able to use and combine results from different, but related experiments, studies, and existing background knowledge would be very desirable. However, there is currently no principled framework that allows us to do so, which means that even if lots of important information (data) is available, the answers to key causal questions could still be missed.

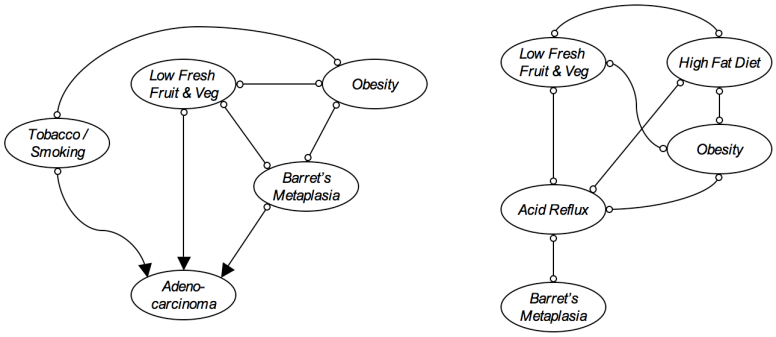

Figure 4.2: Independence models, e.g. as learned by the FCI-algorithm in section 2.4, from two different (hypothetical) studies into oesophageal cancer risk factors.

Example 4.2. *Suppose that, in relation to example 4.1, there are two separate (hypothetical) studies into riks factors. Results of these two studies are depicted as the models in Figure 4.2. Neither of these models really explains the relation between obesity and adenocarcinoma (a type of oesophageal cancer in the region close to the stomach). Only a proper combination of both models can reveal that it is not so much body fat, as people's eating behaviour (via acid reflux) that is likely behind much of this increase.*

In short, methods to relate results from different studies must be able to:

- handle data from partly overlapping sets of measured variables,
- handle different *types* of input information, e.g. causal (from background knowledge) and probabilistic,
- explain apparent contradictions: risk factors identified by one study, but not by another,
- produce coherent, clear, and concise output

Constraint-based methods like the PC-algorithm [Spirtes *et al.*, 2000] are provably correct in the large sample limit, as are Bayesian methods like the greedy search algorithm GES [Chickering, 2002] (with additional post-processing steps to handle hidden confounders). Both are defined in terms of modeling a single data set and have no principled means to relate to results from other sources in the process.

Recent developments show how, under certain strict assumptions, multiple partially overlapping data sets can be combined by searching through possible solutions

over all variables in the combined set. For example the ION-algorithm by Tillman *et al.* [2008], looks through all possible PAGs that are consistent with the PAG models that describe each data set. It uses local information from each model to restrict the search space as much as possible, but is still limited to a few nodes in practice. A more efficient version is available that can handle up to 20 nodes, but the number of possible PAG classes in the output quickly becomes prohibitive and difficult to interpret. An interesting approach, focussing on pairwise causal relations instead of the PAG, is taken by [Triantafillou *et al.*, 2010] who translate this problem into a series of SAT problems to which powerful standard SAT solvers can be applied. With preprocessing the resulting cSAT+ algorithm can be scaled up to about 50 nodes.

However all these algorithms are still essentially single model learners in the sense that they assume there is one, single encapsulating structure that accounts for *all* observed dependencies in the different models. In practice, observed dependencies often differ between data sets, precisely because the experimental circumstances were *not* identical in different experiments, even when the causal system at the heart of it was the same. The method we develop in this article shows how to distinguish between causal dependencies internal to the system under investigation and merely contextual dependencies.

In chapter 3 we recognized the modular aspect of causal discovery: causal relations can be derived from combinations of minimal in/dependencies between certain variables, without having to know or uncover anything else about the entire graph. In section 4.3 we take this one step further by showing that such causal discovery can be decomposed into two separate steps: a conditional independency to identify a pair of possible causal relations (one of which is true), and then a conditional dependency that eliminates one option, that can be taken from *different* models, as the corresponding logical causal statements do not depend on the context. This forms the basis underpinning the MCI-algorithm in section 4.4.

Our goal in this chapter is to find a model that can explain and use in/dependence information from multiple experiments in order to derive new, valid causal relations (sections 4.2 and 4.3). It is implemented in section 4.4 as the MCI algorithm, and evaluated in section 4.5.

4.2 Modeling the system

Random variation in a system corresponds to the impact of unknown external variables, see [Pearl, 2000]. Some of these external factors may be actively controlled, e.g. in clinical trials, or passively observed as the natural embedding of a system in its environment. We refer to both observational and controlled studies as *experiments*. External factors that affect two or more variables in a system simultaneously, can lead to dependencies that are not part of the system. Different external factors may bring about observed dependencies that differ between models, seemingly

contradicting each other. By modeling this external environment explicitly as a set of unobserved (hypothetical) context nodes that causally affect the system under scrutiny we can account for this effect.

Definition 4.3. *The* ***external context*** $\mathcal{G}_E$ *of a causal DAG* $\mathcal{G}_C$ *is a set of independent nodes* $\mathbf{U}$ *in combination with links from every* $U \in \mathbf{U}$ *to one or more nodes in* $\mathcal{G}_C$*. The total causal structure of an experiment then becomes* $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$.

Figure 4.3 depicts a causal system in three different experiments (double lined arrows indicate direct causal relations; dashed circles represent unobserved variables). The second and third experiment will result in an observed dependency between variables A and B, whereas the first one will not. The context only introduces arrows from nodes in $\mathcal{G}_E$ to $\mathcal{G}_C$ which can never result in a cycle, therefore the structure of an experiment $\mathcal{G}_T$ is also a causal DAG. Note that differences in dependencies can only arise from different *structures* of the external context.

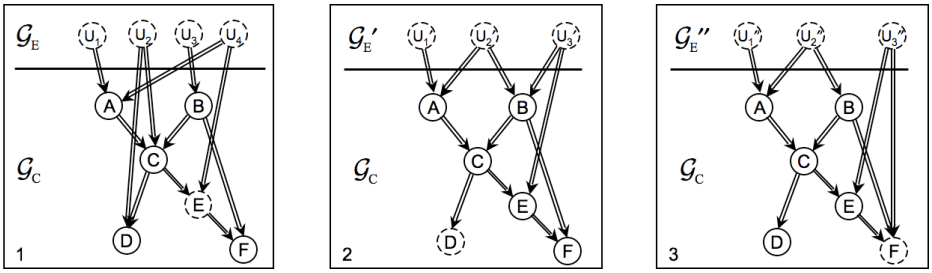

Figure 4.3: A causal system $\mathcal{G}_C$ over six variables $\{A, B, C, D, E, F\}$, in three different experiments corresponding to contexts $\mathcal{G}_E$, $\mathcal{G}'_E$, and $\mathcal{G}''_E$

In this paradigm different experiments become variations in the context of a constant causal system. The goal of causal discovery from multiple models can then be stated as: "Given experiments with unknown total causal structures $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, $\mathcal{G}'_T = \{\mathcal{G}'_E + \mathcal{G}_C\}$, etc., and known joint probability distributions $p(\mathbf{V} \subset \mathcal{G}_T)$, $p'(\mathbf{V}' \subset \mathcal{G}'_T)$, etc., which variables are connected by a directed path in $\mathcal{G}_C$?".

We assume that the large sample limit distributions $p(\mathbf{V})$ are known and can be used to obtain categorical statements about probabilistic (in)dependencies between sets of nodes in each experiment. In this chapter we largely ignore selection bias, which should follow straightforward from these and previous results, see also [Spirtes *et al.*, 1999]. Finally, section 4.3.2 discusses how to handle interventions, i.e. experimental context nodes that externally force a variable to a particular value. These effectively alter the structure of the causal system $\mathcal{G}_C$ by eliminating (blocking) all incoming causal paths into the intervention node, and can be a great source of causal information. However, as our benchmark reference methods in section 4.5 cannot handle these properly, we do not consider interventions in the evaluation, and assume there are no such 'unknown' blocking interventions.

4.3 Causal relations in multiple models

From chapter 3 we know there is a close connection between minimal in/dependencies and presence/absence of a causal relation between variables. Restating the key results from section 3.2 in terms of a causal system $\mathcal{G}_C$ in an external context $\mathcal{G}_E$ (ignoring selection bias) gives the following three rules:

Lemma 4.4. *Let X, Y, $\mathbf{Z}$, and W be four disjoint (sets of) observed nodes in an experiment with causal structure $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, then for arbitrary context $\mathcal{G}_E$:*

(1) $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$ implies causal paths $Z \Rightarrow X$ and/or $Z \Rightarrow Y$ from every $Z \in \mathbf{Z}$ to X and/or Y in $\mathcal{G}_C$,

(2) $X \not\perp\!\!\!\perp_p Y \mid \mathbf{Z} \cup [W]$ implies no causal paths $W \not\Rightarrow X$, $W \not\Rightarrow Y$, or $W \not\Rightarrow Z$ for any $Z \in \mathbf{Z}$ in $\mathcal{G}_C$,

(3) $X \perp\!\!\!\perp_p Y$ implies absence of causal paths $X \not\Rightarrow Y$ and $Y \not\Rightarrow X$ in $\mathcal{G}_C$.

The in/dependence patterns above have a *causal* origin, independent of the (unobserved) external background: rule (1) identifies a candidate pair of causal relations, rule (2) identifies the absence of causal paths, and rule (3) eliminates direct causal links between variables. Causal discovery from multiple models now takes the obvious form:

Corollary 4.5. *Let X, Y and $Z \in \mathbf{Z}$ be disjoint (sets of) variables in two experiments with causal structures resp. $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, and $\mathcal{G}'_T = \{\mathcal{G}'_E + \mathcal{G}_C\}$, then if there exists both:*

- *a minimal conditional independence $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$ in $\mathcal{G}_T$, and*
- *established absence of a causal path $Z \not\Rightarrow X$ from $\mathcal{G}'_T$,*

then there is a causal link (directed path) $Z \Rightarrow Y$ in $\mathcal{G}_C$.

In fact, the origin of the information $Z \not\Rightarrow X$ is irrelevant: be it from in/dependencies via rule (2), other properties of the distribution, e.g. non-Gaussianity [Shimizu *et al.*, 2006] or nonlinear features [Hoyer *et al.*, 2009], or existing background knowledge. The only prerequisite for bringing results from various sources together is that the causal system at the centre is *invariant*, i.e. that the causal structure $\mathcal{G}_C$ remains the same across the different experiments $\mathcal{G}_T$, $\mathcal{G}'_T$ etc.

4.3.1 Combining information from multiple models

In this section we focus on efficiently combining multiple conditional independence models represented by (complete) PAGs. We want to use these models to convey as much about the underlying causal structure $\mathcal{G}_C$ as possible. For that we choose a **causal PAG** as the target output model: similar in form and interpretation to a PAG, where tails and arrowheads now represent *all known* (non)causal relations, see Definition 2.5. Note this is not necessarily an equivalence class in accordance with the rules in [Zhang, 2008], as it may contain additional explicit information.

Ingredients for extracting this information are the rules in Lemma 4.4, in combination with the standard properties of causal relations: acyclic (if $X \Rightarrow Y$ then $Y \not\Rightarrow X$) and transitivity (if $X \Rightarrow Y$ and $Y \Rightarrow Z$ then $X \Rightarrow Z$), see Proposition 3.4. As the causal system is assumed invariant, the established (absence of) causal relations in one model are valid in all models.

Algorithm 4.1 Brute force implementation of rules (1)-(3)

Input : set of complete PAGs $\mathcal{P}_i$, fully $\circ\!\!-\!\!\circ$ connected graph $\mathcal{G}$
Output : causal PAG $\mathcal{G}$

1: **for all** $\mathcal{P}_i$ **do**
2: $\quad\mathcal{G} \leftarrow$ eliminate all edges not appearing between nodes in $\mathcal{P}_i$
3: $\quad\mathcal{G} \leftarrow$ (non)causal connections between nodes in $\mathcal{P}_i$
4: **end for**
5: **repeat**
6: $\quad$**for all** $\mathcal{P}_i$ **do**
7: $\quad\quad$**for all** $\{X, Y, Z, \mathbf{W}\} \in \mathcal{P}_i$ **do**
8: $\quad\quad\quad\mathcal{G} \leftarrow (Z \not\Rightarrow \{X, Y, \mathbf{W}\})$, if $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{W} \cup [Z]$
9: $\quad\quad\quad\mathcal{G} \leftarrow (Z \Rightarrow Y)$, if $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{W} \cup Z]$ and $(Z \not\Rightarrow X) \in \mathcal{G}$
10: $\quad\quad$**end for**
11: $\quad$**end for**
12: **until** no more new non/causal information found

A straightforward brute-force implementation is given by Algorithm 4.1. The input is a set of PAG models $\mathcal{P}_i$, representing the conditional (in)dependence information between a set of observed variables, e.g. as learned by the augmented FCI-algorithm from section 2.4, from a number of different experiments $\mathcal{G}_T^{(i)}$ on an invariant causal system $\mathcal{G}_C$. The output is the single causal PAG $\mathcal{G}$ over the *union* of all nodes in the input models $\mathcal{P}_i$.

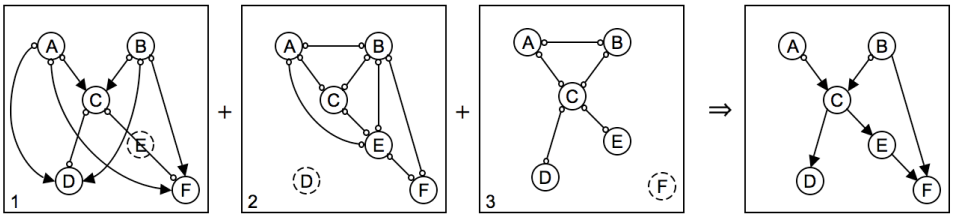

Figure 4.4: Three different experiments, one causal model

Example 4.6. *Consider the three PAG models in the l.h.s. of figure 4.4. None of these identifies a causal relation, yet despite the different (in)dependence relations, it is easily verified that Algorithm 4.1 terminates after two loops with the nearly complete causal PAG on the r.h.s. as the final output. Figure 4.3 shows corresponding experiments that explain the observed dependencies above.*

To the best of our knowledge, Algorithm 4.1 is the first algorithm ever to perform such a derivation. Nevertheless, this brute-force approach exhibits a number of serious shortcomings. In the first place, the computational complexity of the repeated loop over all subsets in line 7 makes it not scalable: for small models like the ones in figure 4.4 the derivation is almost immediate, but for larger models it quickly becomes unfeasible. Secondly, for sparsely overlapping models, i.e. when the observed variables differ substantially between the models, the algorithm can miss certain relations: when a causal relation is found to be absent between two non-adjacent nodes, then this information cannot be recorded in $\mathcal{G}$, and subsequent causal information identifiable by rule (1) may be lost. These problems are addressed in the section 4.4, resulting in the MCI-algorithm.

4.3.2 Including interventions

In many experiments certain factors are actively influenced instead of just observed. When one or more variables are externally *forced* to specific values, irrespective of the value of their parents in the underlying causal graph $\mathcal{G}_C$, this is known as a **(hard) intervention**. Examples are setting the amount of artificial light in greenhouse trials, administering medicine 'X', or enforcing a work-out regime. *Soft* interventions are when the relevant variables are not set but only stimulated, e.g. reducing carbon emissions through taxation, or providing only an incentive to exercise: these change probabilities but not the (internal) structure of $\mathcal{G}_C$, and are covered by the standard context model in section 4.2.

If the hard intervention is on a variable X with parents in $\mathcal{G}_C$, the result is a distribution that is faithful to a **modified** causal system $\mathcal{G}'_C$ where all causal links to X in the original $\mathcal{G}_C$ have been severed and replaced by an additional control node U_M (M for manipulation) in the context, or at least 'external to the system $\mathcal{G}_C$', with X as its child; see also [Pearl, 2000] for an extensive discussion. If we are certain that the intervention on X does not directly affect any other variables that are (parents of) nodes in $\mathcal{G}_C$, then it is known as a **surgical intervention**. If not, i.e. if we do a hard intervention on X but cannot rule out that we accidentally affected/influenced other variables in the system, then we have to account for possible **side effects** of the intervention. Assuming that any possible side effects (if present) at most correspond to soft interventions (no 'accidental' hard interventions), then these take the form of (unknown) additional links from U_M to one or more other nodes in $\mathcal{G}'_C$. Despite the fact that in such cases the (internal) causal system has changed from $\mathcal{G}_C$ to $\mathcal{G}'_C$, knowing which variable(s) were the target of intervention(s) can provide valuable information about the original causal structure $\mathcal{G}_C$.

We can use experimental knowledge through modified inference rules, depending on the role of the intervention node in the minimal independence, and whether or not side effects can be excluded. The key observation here is that interventions do not introduce causal relations between variables within the system $\mathcal{G}_C$, but only block (overrule) certain ones from having an effect. Therefore, any positive causal

relation (or disjunctive pair of possible causal relations) identified in the modified system $\mathcal{G}'_C$ is also valid in the original one. Note this remains true for multiple interventions, leading to the following modified inference rule:

Lemma 4.7. *Let X, Y and $\mathbf{Z}$ be three disjoint non-empty (sets of) variables in an experiment with causal structure $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, then an observed minimal independence $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{Z}]$, in combination with (hard) intervention on a node ...*

(1) X, without side effects, implies causal links $X \Rightarrow Y$, $\mathbf{Z} \Rightarrow Y$ and $X \Rightarrow Z$ for some $Z \in \mathbf{Z}$,

(2) X, with possible side effects, implies causal links $\mathbf{Z} \Rightarrow Y$ for all $Z \in \mathbf{Z}$ in $\mathcal{G}_C$,

(3) Z, without side effects, implies the familiar $(Z \Rightarrow X) \vee (Z \Rightarrow Y)$, and that Z is the source of some trek between two nodes from $\{X, Y, \mathbf{Z}_{\setminus Z}\}$ in $\mathcal{G}_C$,

(4) Z, with possible side effects, implies only the familiar $(Z \Rightarrow X) \vee (Z \Rightarrow Y)$.

Proof. Follows from rule 1 in Lemma 3.5, in combination with the observation that all causal relations in the intervened system $\mathcal{G}'_C$ are present in the underlying $\mathcal{G}_C$. □

If $\mathbf{Z}$ is just a single node Z, then this reduces to:

Corollary 4.8. *Let X, Y and Z be three distinct variables in an experiment with causal structure $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, then an observed minimal conditional independence $X \perp\!\!\!\perp_p Y \,|\, [Z]$, in combination with intervention on node ...*

(1) X, without side effects, implies causal links $X \Rightarrow Z \Rightarrow Y$ in $\mathcal{G}_C$,

(2) X, with possible side effects, implies a causal link $Z \Rightarrow Y$ in $\mathcal{G}_C$,

(3) Z, without side effects, implies causal links $X \Leftarrow Z \Rightarrow Y$ (and $X \not\Rightarrow Y$) in $\mathcal{G}_C$,

(4) Z, with possible side effects: no change.

Proof. Follows immediately from Lemma 4.7. □

This does not apply to *absence* of certain causal relations from unconditional independence or conditional dependence. For example, $X \Rightarrow Y$ with intervention on Y makes them independent in the interventional distribution, and eliminates option $Y \Rightarrow X$, but does not preclude the possibility $X \Rightarrow Y$. Similarly, for a causal system $\mathcal{G}_C$ with $X \Rightarrow Z \Rightarrow W \Rightarrow Y$, $Z \Leftarrow U \Rightarrow Y$, and intervention on W, we can find $X \not\perp\!\!\!\perp_p Y \,|\, [Z]$, from which we *can* conclude by Lemma 3.5 that $Z \not\Rightarrow Y$ in $\mathcal{G}'_C$, but not automatically that the same also holds in the underlying $\mathcal{G}_C$, *unless* we can somehow rule out $Z \Rightarrow W$.

We will not pursue inference from interventional distributions further in this chapter: for a full treatment the promising new 'nested Markov model' theory from

Shpitser *et al.* [2012] (based on latent projections instead of ancestral graphs, see [Verma and Pearl, 1991]), is more appropriate.

Case (2) in Corollary 4.8, intervention on X with possible side effects, forms the basis of the Local Causal Discovery (LCD) algorithm in [Cooper, 1997], who already employed a notion of a context of possible causal graphs (although in slightly different form) for the observed independency given intervention on X. There it was based on an exhaustive enumeration of all possible configurations over three nodes including possible confounders, but we see that it also follows immediately from theorem 1, rule (1) via the elimination of strict conditional independencies involving causal links to X. The first case (intervention on X without side effects) is similar to the situation exploited by the Trigger algorithm in Chen *et al.* [2007], where X corresponds to the known, uncaused locus L_i (location on a gene) for a transcription factor T_i, that separates variable L_i from another transcription factor T_j in a randomized trial, from which $L_i \Rightarrow T_i \Rightarrow T_j$ is inferred.

Note that interventions can detect/create minimal independencies that do not exist in the underlying causal graph $\mathcal{G}_C$. Example: for a model with $X \Rightarrow Z_1 \Rightarrow Y$ plus $X \Leftarrow Z_2 \Rightarrow Y$, we have $X \perp\!\!\!\perp_p Y \,|\, [Z_1, Z_2]$, but with intervention on X we find $X \perp\!\!\!\perp_p Y \,|\, [Z_1]$, and so $Z_1 \Rightarrow Y$.

4.4 The MCI algorithm

To tackle the computational complexity noted at the end of section 4.3.1 we have the following result, in which we use the notion of a possibly directed (p.d.) path in a PAG[1] to indicate a path that can be converted into a directed path by changing circle marks into appropriate tails and arrowheads, see section 2.1.

Proposition 4.9. *Let X and Y be two variables in an experiment with causal structure $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, and let $\mathcal{P}_{[G]}$ be the corresponding PAG over a subset of observed nodes from $\mathcal{G}_T$. Then the absence of a causal link $X \not\Rightarrow Y$ is detectable from the conditional in/dependence structure in this experiment iff there exists no p.d. path from X to Y in $\mathcal{P}_{[G]}$.*

In other words: X cannot be a cause (ancestor) of Y if all paths from X to Y in the graph $\mathcal{P}_{[G]}$ go against an invariant arrowhead (signifying non-ancestorship) and vice versa. We refer to this as inference rule (4). Calculating which variables are connected by a p.d. path from a given PAG is straightforward: turn the graph into a $\{0, 1\}$ adjacency matrix by setting all arrowheads to zero and all tails and circle marks to one, and compute the resulting reachability matrix. As this will uncover *all* detectable 'non-causal' relations in a PAG in one go, it needs to be done only once for each model, and can be aggregated into a matrix $\mathbf{M}_C$ to make all tests for rule (2) in line 8 of Algorithm 4.1 superfluous. If we also record all other established

[1] From here on we ignore the distinction between a PAG and a complete PAG, as any PAG can be easily checked for/converted into completeness by executing the orientation rules in Table 2.1.

(non)causal relations in the matrix $\mathbf{M}_C$ as the algorithm progresses, then indirect causal relations are no longer lost when they cannot be transferred to the output graph $\mathcal{G}$. The next lemma propagates indirect (non)causal information from $\mathbf{M}_C$ to edge marks in the graph:

Lemma 4.10. *Let X, Y and $\mathbf{Z}$ be disjoint sets of variables in an experiment with causal structure $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, then for every $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$:*

- *every (indirect) causal relation $X \Rightarrow Y$ implies causal links $\mathbf{Z} \Rightarrow Y$,*
- *every absence of (indirect) causal relation $X \not\Rightarrow Y$ implies no links $X \not\Rightarrow \mathbf{Z}$.*

The first makes it possible to orient indirect causal chains, the second shortens indirect non-causal links. We refer to these as rules (5) and (6), respectively. As a final improvement it is worth noting that for rules (1), (5) and (6) it is only relevant to know that a node Z occurs in *some* $\mathbf{Z}$ in a minimal conditional independence relation $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$ separating X and Y, but not what the other nodes in $\mathbf{Z}$ are or in what model(s) it occurred. We can introduce a structure $\mathbf{S}_{CI}$ to record all nodes Z that occur in some minimal conditional independency in one of the models $\mathcal{P}_i$ for each combination of nodes (X, Y), *before* any of the rules (1), (5) or (6) is processed. As a result, in the repeated causal inference loop no conditional independence / m-separation tests need to be performed at all.

Algorithm 4.2 MCI algorithm

Input : set of PAGs $\mathcal{P}_i$, fully $\circ\!\!-\!\!\circ$ connected graph $\mathcal{G}$
Output : causal graph $\mathcal{G}$, causal relations matrix $\mathbf{M}_C$

1: $\mathbf{M}_C \leftarrow \mathbf{0}$
2: **for all** $\mathcal{P}_i$ **do**
3: $\quad\mathcal{G} \leftarrow$ eliminate all edges not appearing between nodes in $\mathcal{P}_i$
4: $\quad\mathbf{M}_C \leftarrow (X \not\Rightarrow Y)$, if no p.d. path $\langle X, \ldots, Y\rangle \in \mathcal{P}_i$
5: $\quad\mathbf{M}_C \leftarrow (X \Rightarrow Y)$, if causal path $\langle X \Rightarrow \ldots \Rightarrow Y\rangle \in \mathcal{P}_i$ ▷ *transitivity*
6: $\quad$**for all** $(X, Y, Z) \in \mathcal{P}_i$ **do**
7: $\quad\quad\mathbf{S}_{CI} \leftarrow$ triple (X, Y, Z), if $Z \in \mathbf{Z}$ for which $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$
8: $\quad$**end for**
9: **end for**
10: **repeat**
11: $\quad$**for all** $(X, Y, Z) \in \mathcal{G}$ **do**
12: $\quad\quad\mathbf{M}_C \leftarrow (Z \Rightarrow Y)$, for unused $(X, Y, Z) \in \mathbf{S}_{CI}$ with $(Z \not\Rightarrow X) \in \mathbf{M}_C$
13: $\quad\quad\mathbf{M}_C \leftarrow (X \not\Rightarrow Z)$, for unused $(X \not\Rightarrow Y) \in \mathbf{M}_C$ with $(X, Y, Z) \in \mathbf{S}_{CI}$
14: $\quad\quad\mathbf{M}_C \leftarrow (Z \Rightarrow Y)$, for unused $(X \Rightarrow Y) \in \mathbf{M}_C$ with $(X, Y, Z) \in \mathbf{S}_{CI}$
15: $\quad$**end for**
16: **until** no more new causal information found
17: $\mathcal{G} \leftarrow$ non/causal info in $\mathbf{M}_C$ ▷ *tails/arrowheads*

With these results we can now give an improved version of the brute-force approach: the *Multiple model Causal Inference* (MCI) algorithm, above. The input is

still a set of PAG models from different experiments, but the output is now twofold: the graph $\mathcal{G}$, containing the causal structure uncovered for the underlying system $\mathcal{G}_C$, as well as the matrix $\mathbf{M}_C$ with an explicit representation of all (non)causal relations between observed variables, including remaining indirect information that cannot be read from the graph $\mathcal{G}$.

The first stage (lines 2-9) is a pre-processing step to extract all necessary information for the second stage from each of the models separately. Building the $\mathbf{S}_{CI}$ matrix is the most expensive step as it involves testing for conditional independencies (m-separation) for increasing sets of variables. This can be efficiently implemented by noting that nodes connected by an edge will not be separated and that many other combinations will not have to be tested as they contain a subset for which a (minimal) conditional independency has already been established. If a (non)causal relation is found between adjacent variables in $\mathcal{G}$, or one that can be used to infer other intermediate relations (lines 13-14), then it can be marked as 'processed' to avoid unnecessary checks. Similar for the entries recorded in the minimal conditional independence structure $\mathbf{S}_{CI}$.

The MCI algorithm is provably sound in the sense that if all input PAG models $\mathcal{P}_i$ are valid, then all (absence of) causal relations identified by the algorithm in the output graph $\mathcal{G}$ and (non)causal relations matrix $\mathbf{M}_C$ are also valid, provided that the causal system $\mathcal{G}_C$ is an invariant causal DAG and the causal faithfulness assumption is satisfied.

4.5 Experimental results

We tested the MCI-algorithm on a variety of synthetic data sets to verify its validity and assess its behaviour and performance in uncovering causal information from multiple models. For the generation of random causal DAGs we used a variant of [Ide and Cozman, 2002] to control the distribution of edges over nodes in the network. The random experiments in each run were generated from this causal DAG by including a random context and hidden nodes. For each network the corresponding (complete) PAG was computed, and together used as the set of input models for the MCI-algorithm. The generated output $\mathcal{G}$ and $\mathbf{M}_C$ was verified against the true causal DAG and expressed as a percentage of the true number of (absent) causal relations.

To assess the performance we introduced two reference methods to act as a benchmark for the MCI-algorithm (in the absence of other algorithms that can validly handle different contexts). The first is a common sense method, indicated as 'sum-FCI', that utilizes the transitive closure of all causal relations in the input PAGs, that could have been identified by FCI in the large sample limit. As the second benchmark we take all causal information contained in the PAG over the union of observed variables, independent of the context, hence 'nc-CPAG' for 'no context'. Note that this is not really a method as it uses information directly derived from the true causal graph.

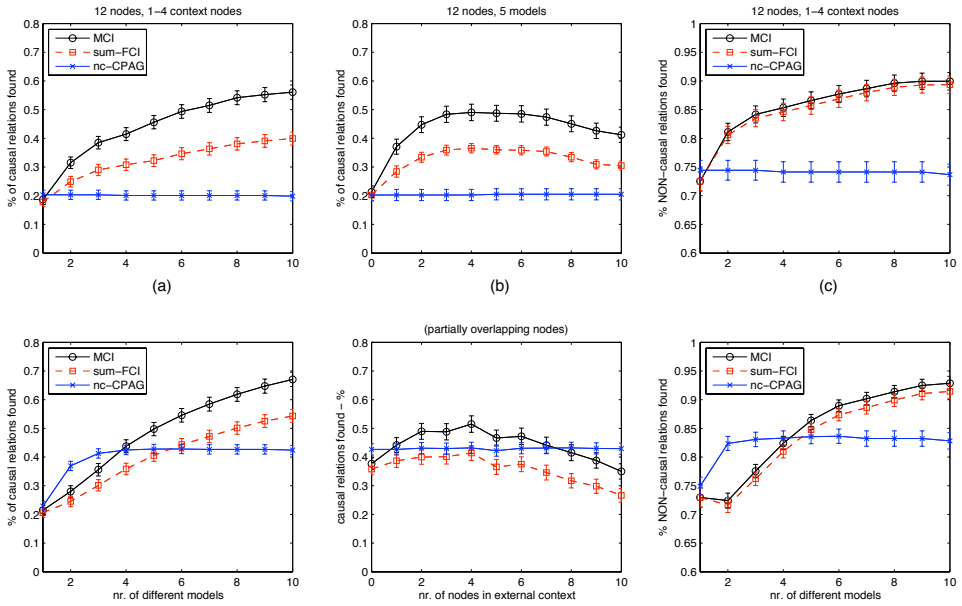

Figure 4.5: Proportion of causal relations discovered by the MCI-algorithm vs. sum-FCI and nc-CPAG in different settings; (a) causal relations vs. nr. of models, (b) causal relations vs. nr. of context nodes, (c) non-causal relations $\not\rightarrow$ vs. nr. of models; (top) identical observed nodes in input models, (bottom) only partially overlapping observed nodes; see main text for details.

In Figure 4.5, each graph depicts the percentage of causal (a&b) or non-causal (c) relations uncovered by each of the three methods: MCI, sum-FCI and nc-CPAG, as a function of the number of input models (a&c) or the number of nodes in the context (b), averaged over 200 runs, for both identical (top) or only partly overlapping (bottom) observed nodes in the input models. Performance is calculated as the proportion of uncovered relations as compared to the actual number of non/causal relations in the true causal graph over the union of observed nodes in each model set. In these runs the underlying causal graphs contained 12 nodes with edge degree $\leq$ 5. Tests for other, much sparser/denser graphs up to 20 nodes, showed comparable results.

Some typical behaviour is easily recognized: MCI always outperforms sum-FCI, and more input models always improve performance. Also non-causal information (c) is much easier to find than definite causal relations (a&b). For single models / no context the performance of all three methods is very similar, although not necessarily identical. The perceived initial drop in performance in Figure 4.5(c,bottom) is only because in going to two models the number of non-causal relations in the union rises more quickly than the number of new relations that is actually found (due to lack of overlap). A striking result that is clearly brought out is that adding random

context actually *improves* detection rate of causal relations. The rationale behind this effect is that externally induced links can introduce conditional dependencies, allowing the deduction of non-causal links that are not otherwise detectable; these in turn may lead to other causal relations that can be inferred, and so on. If the context is expanded further, at some point the detection rate will start to deteriorate as the causal structure will be swamped by the externally induced links (b).

As to be expected: all of the (non)causal relations identified by a valid MCI algorithm were found to be present in the true causal graph as well. For $\gtrsim 8$ nodes the algorithm spends the majority of its time building the $\mathbf{S}_{CI}$ matrix in lines 6-8. The actual number of minimal conditional independencies found, however, is quite low, typically in the order of a few dozen for graphs of up to 12 nodes.

4.6 Conclusion

We have shown the first principled algorithm that can use results from *different* experiments to uncover new (non)causal information. It is provably sound in the large sample limit, provided the input models are learned by a valid procedure like the augmented FCI algorithm from section 2.4. In its current implementation the MCI-algorithm is a fast and practical method that can easily be applied to sets of models of up to 20 nodes (presuming the graph is sufficiently sparse). Compared to related algorithms like ION, it produces very concise and easily interpretable output, and does not suffer from the inability to handle any differences in observed dependencies between data sets [Tillman *et al.*, 2008]. For larger models it can be converted into an anytime algorithm by running over minimal conditional independencies from subsets of increasing size: at each level all uncovered causal information is valid, and, for reasonably sparse models, most will already be found at low levels. For very large models an exciting possibility is to target only specific causal relations: finding the right combination of (in)dependencies is sufficient to decide if it is causal, even when there is no hope of deriving a global PAG model.

From the construction of the MCI-algorithm it is sound, but not necessarily complete. From chapter 3 we know that Lemma 4.4 already covers all invariant arrowheads in the single model case, and only needs one additional rule (the 'inferred blocking node') to cover all tails as well. We aim to extend this result to the multiple model domain. Integrating our approach with recent developments in causal discovery that are not based on independence constraints [Shimizu *et al.*, 2006; Hoyer *et al.*, 2009] can provide for even more detectable causal information. When applied to real data sets the large sample limit no longer applies and inconsistent causal relations may result. It should be possible to exclude the contribution of such links (when detected) on the final output. Alternatively, output might be generalized to quantities like 'the probability of a causal relation' based on the strength of appropriate conditional (in)dependencies in the available data. The next chapter contains an important step in that direction.

4.A Proofs

Proposition 4.9

Proof. '$\Leftarrow$' follows from the fact that a directed path $\pi = \langle X, \dots, Y \rangle$ in the underlying causal DAG $\mathcal{G}_C$ implies existence of a directed path in the true MAG over the observed nodes and therefore at least the existence of a p.d. path in the PAG $\mathcal{P}_{[G]}$. '$\Rightarrow$' follows from the completeness of the PAG in combination with theorem 2 in [Zhang, 2008] about orientability of PAGs into MAGs. This, together with Meek's algorithm [Meek, 1995] for orienting chordal graphs into DAGs with no unshielded colliders, shows that it is always possible to turn a p.d. path into a directed path in a MAG that is a member of the equivalence class $\mathcal{P}_{[G]}$. Therefore, a p.d. path from X to Y in $\mathcal{P}_{[G]}$ implies there is at least some underlying causal DAG in which it is a causal path, and so cannot correspond to a valid, detectable absence of a causal link. □

Lemma 4.10

Proof. From rule (1) in Lemma 4.4, $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$ implies causal links $\mathbf{Z} \Rightarrow X$ and/or $\mathbf{Z} \Rightarrow Y$. If $X \Rightarrow Y$ then by transitivity $\mathbf{Z} \Rightarrow X$ also implies $\mathbf{Z} \Rightarrow Y$. If $X \not\Rightarrow Y$ then for any $Z \in \mathbf{Z}$, $X \Rightarrow Z$ implies $Z \Rightarrow Y$ and so (transitivity) also $X \Rightarrow Y$, in contradiction of the given; therefore $X \not\Rightarrow \mathbf{Z}$. □

Chapter 5

Bayesian Constraint-based Causal Discovery

This chapter targets the problem of accuracy and robustness in causal inference from finite data sets. Some state-of-the-art algorithms produce clear output complete with solid theoretical guarantees but are susceptible to propagating erroneous decisions, while others are very adept at handling and representing uncertainty, but need to rely on undesirable assumptions. Our aim is to combine the inherent robustness of the Bayesian approach with the theoretical strength and clarity of constraint-based methods. We use a Bayesian score to obtain probability estimates on the input statements used in a constraint-based procedure. These are subsequently processed in decreasing order of reliability, letting more reliable decisions take precedence in case of conflicts, until a single output model is obtained. Tests show that a basic implementation of the resulting Bayesian Constraint-based Causal Discovery (BCCD) algorithm already outperforms established procedures such as FCI and Conservative PC. It can also indicate which causal decisions in the output have high reliability and which do not.

5.1 Introduction: Robust Causal Discovery

Causal discovery lies at the heart of most scientific research today. Perhaps surprisingly then, 'proper' causal discovery algorithms are still not as routinely applied in practice as one might expect. Researchers can find them too difficult, slow, or cumbersome to use for everyday analysis. But arguably one of the main obstacles is the

This chapter is based on: [Claassen and Heskes, 2012a] "A Bayesian Approach to Constraint Based Causal Inference", which received the Best Paper Award at the 28th Conference on Uncertainty in Artificial Intelligence.

perceived *lack of robustness* of such methods. Small changes in the input can lead to substantial changes in the output, as (erroneous) borderline decisions become decisive statements that are propagated through the network. But this ambiguity is usually not apparent in the resulting causal model. In particular, if causal relations are identified that are known (e.g. from background knowledge) to be clearly wrong, then this tends to have a very negative impact on the confidence assigned by a researcher to *any* causal relation found by the algorithm. Outputting a single model without further qualifications may seem clear and informative, but suggests a level of reliability that cannot be justified in practice, and makes researchers reluctant to go through the trouble of applying them in the first place.

Other methods can provide some measure of confidence by outputting multiple models with the implied assumption that arcs present in many are more likely to be true. The obvious downside is that it becomes much harder to interpret succinctly what causal relations are actually implied by the output. On top of that they often have to rely on undesirable, unrealistic assumptions like causal sufficiency, which makes them less suited for causal analysis in general. This is somewhat compensated for by the fact that other sources of information, e.g. background literature, are incorporated naturally through a prior on structures and/or parameters.

For a more robust solution, we want a method that is less susceptible to the impact of single, categorical decisions, but can handle unobserved common causes. Ideally we would like a robust and efficient method that requires few assumptions, and outputs a single clear model that indicates explicitly how reliable all the individual causal relations are. Sections 5.2-5.4 in this chapter implement the first part, while 5.5 and 5.6 make a good step towards the second.

Background

We briefly state a few terms and concepts that play an important role in this chapter; for details the reader is referred to sections 2.1-2.3.

In many real-world systems the relations and interactions between variables can be modeled in the form of a causal DAG $\mathcal{G}_C$ over a set of variables $\mathbf{V}$. A directed path from A to B in such a graph indicates a *causal relation* $A \Rightarrow B$ in the system, where cause A *influences* the value of its effect B, but not the other way around. An edge $A \rightarrow B$ in $\mathcal{G}_C$ indicates a *direct* causal link. In data from a system with a causal relation $A \Rightarrow B$ (direct or indirect), the values of A and B have a tendency to vary together, i.e. they become probabilistically dependent.

The *Causal Markov Condition* and *Causal Faithfulness Condition* link the underlying, asymmetric causal relations to observable, symmetric probabilistic dependencies. Together, they imply that the causal DAG $\mathcal{G}_C$ is also *minimal*, in the sense that no proper subgraph can satisfy both assumptions *and* produce the same probability distribution [Zhang and Spirtes, 2008]. The joint probability distribution induced by a causal DAG $\mathcal{G}_C$ factors according to a *Bayesian network* (BN). If some of the variables in the causal DAG are hidden then the independence rela-

tions between the observed variables may be represented in the form of a *maximal ancestral graph* (MAG) $\mathcal{M}$. Its *equivalence class* $[\mathcal{M}]$ (indistinguishable via independencies), is represented as a *partial ancestral graph* (PAG) $\mathcal{P}$, which keeps the skeleton and invariant tail ($-$) and arrowhead ($>$) edge marks in $[\mathcal{M}]$, and turns the rest into circles ($\circ$); see Definitions 2.2-2.6.

Causal discovery procedures

With this in mind, the task of a causal discovery algorithm is to find as many invariant features of the equivalence class corresponding to a given data set as possible. From this, all identifiable, present or absent causal relations can be read.

A large class of **constraint-based** causal discovery algorithms is based directly on the faithfulness assumption: if a conditional independence $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$ can be found for *any* set of variables $\mathbf{Z}$, then there is no direct causal relation between X and Y in the underlying causal graph $\mathcal{G}_C$, and hence no edge between X and Y in the equivalence class $\mathcal{P}$. In this way, an exhaustive search over all pairs of variables can uncover the entire skeleton of $\mathcal{P}$. In the subsequent stage a number of orientation rules are executed that find the invariant tails and arrowheads.

Members of this group include the IC-algorithm [Pearl and Verma, 1991], PC/FCI [Spirtes *et al.*, 2000], Grow-Shrink [Margaritis and Thrun, 1999], TC [Pellet and Elisseef, 2008], and many others. All involve repeated independence tests in the adjacency search phase, and employ orientation rules as described in Meek [1995]. The differences lie mainly in the search strategy employed, size of the conditioning sets, and additional assumptions imposed. Of these, only the FCI algorithm in conjunction with the additional orientation rules in [Zhang, 2008] is sound and complete in the large-sample limit when hidden common causes and/or selection bias may be present.

Constraint-based procedures tend to output a single, reasonably clear graph, representing the class of all possible causal DAGs. The downside is that for finite data they give little indication of which parts of the network are stable (reliable), and which are not: if unchecked, even one erroneous, borderline independence decision may be propagated through the network, leading to multiple incorrect orientations [Spirtes, 2010].

To tackle the perceived lack of robustness of PC, Ramsey *et al.* [2006] proposed a **conservative approach** for the orientation phase. The standard rules draw on the implicit assumption that, after the initial adjacency search, a single $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$ should suffice to orient an unshielded triple $\langle X, Z, Y \rangle$, as Z should be either part of *all* or part of *no* sets that separate X and Y. The Conservative PC (CPC) algorithm tests explicitly whether this assumption holds, and only orients the triple into a noncollider resp. v-structure $X \rightarrow Z \leftarrow Y$ if found true. If not, then it is marked as *unfaithful*. Tests show that CPC significantly outperforms standard PC

in terms of overall accuracy, albeit often with less informative output, for only a marginal increase in run-time.

This idea can be extended to FCI: the set of potential separating nodes is now conform FCI's adjacency search, and any of Zhang's orientation rules that relies on a particular unshielded (non-)collider does not fire on an unfaithful triple. See [Glymour *et al.*, 2004; Kalisch *et al.*, 2011] for an implementation of Conservative FCI (CFCI) and many related algorithms.

The **score-based** approach is an alternative paradigm that builds on the implied *minimality* of the causal graph: define a scoring criterion $S(\mathcal{G}, \mathbf{D})$ that measures how well a Bayesian network with structure $\mathcal{G}$ fits the observed data $\mathbf{D}$, while preferring simpler networks, with fewer free parameters, over more complex ones. If the causal relations between the variables in $\mathbf{D}$ form a causal DAG $\mathcal{G}_C$, then in the large sample limit the highest scoring structure $\mathcal{G}$ must be part of the equivalence class of $[\mathcal{G}_C]$.

An example is the (Bayesian) likelihood score: given a Bayesian network $\mathcal{B} = (\mathcal{G}, \mathbf{\Theta})$, the likelihood of observing a particular data set $\mathbf{D}$ can be computed recursively from the network. Integrating out the parameters $\mathbf{\Theta}$ in the conditional probability tables (CPTs) then results in:

$$p(\mathbf{D}|\mathcal{G}) = \int_{\mathbf{\Theta}} p(\mathbf{D}|\mathcal{G}, \mathbf{\Theta}) f(\mathbf{\Theta}|\mathcal{G}) \, \mathrm{d}\mathbf{\Theta}, \tag{5.1}$$

where f is a conditional probability density function over the parameters $\mathbf{\Theta}$ given structure $\mathcal{G}$.

A closed form solution to eq.(5.1) is used in algorithms such as K2 [Cooper and Herskovits, 1992] and the Greedy Equivalence Search (GES) [Chickering, 2002] to find an optimal structure by repeatedly comparing scores for slightly modified alternatives until no more improvement can be found. See also Bouckaert [1995] for an evaluation of different strategies using these and other measures such as the BIC-score and minimum description length.

Score-based procedures can output a set of high-scoring alternatives. This ambiguity makes the result arguably less straightforward to read, but does allow for a measured interpretation of the reliability of inferred causal relations, and is not susceptible to incorrect categorical decisions [Heckerman *et al.*, 1999]. The main drawback is the need to rely on the causal sufficiency assumption.

5.2 The Best of Both Worlds

The strength of a constraint-based algorithm like FCI is its ability to handle data from arbitrary faithful underlying causal DAGs and turn it into sound and clear, unambiguous causal output. The strength of the Bayesian score-based approach lies in the robustness and implicit confidence measure that a likelihood weighted combination of multiple models can bring.

- Our idea is to improve on conservative FCI by using a Bayesian approach to estimate the reliability of different constraints, and use this to decide if, when, and how that information should be used.

Instead of classifying pieces of information as reliable or not, we want to rank and process constraints according to a confidence measure. This should allow to avoid propagating unreliable decisions while retaining more confident ones. It also provides a principled means for conflict resolution. The end-result is hopefully a more informative output model than CFCI, while obtaining a higher accuracy than standard FCI can deliver.

To obtain a confidence measure that can be compared across different estimates we want to compute the *probability* that a given independence statement holds from a given data set $\mathbf{D}$. In an ideal Bayesian approach we could compute a likelihood $p(\mathbf{D}|\mathcal{M})$ for each $\mathcal{M} \in \mathbf{M}$ (see section 5.3 on how to approximate this). If we know that the set $\mathbf{M}$ contains the 'true' structure, then the probability of an independence hypothesis I follows from normalized summation as:

$$p(I|\mathbf{D}) \propto \sum_{\mathcal{M} \in \mathbf{M}(I)} p(\mathbf{D}|\mathcal{M}) p(\mathcal{M}), \tag{5.2}$$

[Heckerman *et al.*, 1999], where $\mathbf{M}(I)$ denotes the subset of structures that entail independence statement I, and $p(\mathcal{M})$ represents a prior distribution over the structures (see §5.3.4).

Two remarks. Firstly, it is well known that the number of possible graphs grows very quickly with the number of nodes $\mathbf{V}$. But eq.(5.2) equally applies when we limit data and structures to *subsets* of variables $\mathbf{X} \subset \mathbf{V}$. For sparse graphs we can choose to consider only subsets of size $K \ll |\mathbf{V}|$. We opt to go one step further and follow a search strategy similar to PC/FCI, using structures of increasing size. Secondly, it would be very inefficient to compute eq.(5.2) for each independence statement we want to evaluate. From a single likelihood distribution over structures over $\mathbf{X}$ we can immediately compute the probability of *all* possible independence statements between variables in $\mathbf{X}$, including complex combinations such as those implied by v-structures, just by summing the appropriate contributions for each statement.

Having obtained probability estimates for a list of in/dependence statements $\mathcal{I}$, we can rank these in decreasing order of reliability, and keep the ones based on a decision threshold $p(I|\mathbf{D}) > \theta$, with $\theta = 0.5$ as intuitive default. In case of remaining conflicting statements, the ones with higher confidence take precedence. The resulting procedure is outlined in Algorithm 5.1.

In this form, the Bayesian estimates are only used to guide the adjacency search (update skeleton $\mathcal{G}$, l.7), and to filter the list of independencies $\mathcal{I}$ (l.12). Ideally, we would like the probabilities to guide the orientation phase as well. This implies processing the independence statements sequentially, in decreasing order of reliability. For that we can use the Logical Causal Inference (LoCI) algorithm from

Algorithm 5.1 Outline of Bayesian FCI

Start : database $\mathbf{D}$ over variables $\mathbf{V}$
Stage 1 - Adjacency search
fully connected graph $\mathcal{P}$, empty list $\mathcal{I}$, $K = 0$
repeat
  for all $X - Y$ still connected in $\mathcal{P}$ **do**
    for all adjacent sets $\mathbf{Z}$ of K nodes in $\mathcal{P}$ **do**
      estimate $p(\mathcal{M}|\mathbf{D})$ over $\{X, Y, \mathbf{Z}\}$
      sum to $p(I|\mathbf{D})$ for independencies I
      update $\mathcal{I}$ and $\mathcal{P}$ for each $p(I|\mathbf{D}) > \theta$
    end for
  end for
  $K = K + 1$
until all relevant found
Stage 2 - Orientation rules
rank and filter $\mathcal{I}$ in decreasing order of reliability
orient unshielded triples in $\mathcal{P}$
run remaining orientation rules
return causal model $\mathcal{P}$

section 3.5.2: it breaks up the overall inference process into a series of modular steps that can be executed in arbitrary order by translating observed minimal independence constraints into **logical statements** L about presence or absence of causal relations:

1. $X \perp\!\!\!\perp Y \,|\, [\mathbf{W} \cup Z] \quad \vdash \quad (Z \Rightarrow X) \;\vee\; (Z \Rightarrow Y)$,
2. $X \not\perp\!\!\!\perp Y \,|\, \mathbf{W} \cup [Z] \quad \vdash \quad Z \not\Rightarrow (\{X, Y\} \cup \mathbf{W})$.

New information is found by deduction on the standard causal properties *transitivity* and *irreflexivity*, see section 3.3.

5.3 Sequential Causal Inference

This section discusses the steps needed to turn the previous idea into a working algorithm in the next section. Main issues are: probability estimates for logical causal statements from substructures, Bayesian likelihood computation, and inference from unfaithful DAGs. Proofs are detailed in Appendices 5.A, 5.B.

A word on notation: we use Ł denotes the set of logical causal statements L over two or three variables in $\mathbf{V}$, of the form given in the r.h.s. of rules 1 and 2, above. We use $\mathbf{M_X}$ to represent the set of MAGs over $\mathbf{X}$, and $\mathbf{M_X}(L)$ to denote the subset that entails logical statement L. $\mathbf{D}$ denotes a data set over variables $\mathbf{V}$ from a distribution that is faithful to some (larger) causal DAG $\mathcal{G}_C$. We also use $\mathcal{G}$ to explicitly indicate a DAG, $\mathcal{M}$ for a MAG, and $\mathcal{P}$ for a PAG.

5.3.1 A Modular Approach

In order to process available information in (decreasing) order of reliability we need to obtain probability estimates for logical statements on causal relations from data. Using the notational conventions introduced above:

Lemma 5.1. *The probability of a logical causal statement L given a data set $\mathbf{D}$ is given by*

$$p(L|\mathbf{D}) = \frac{\sum_{\mathcal{M}\in\mathbf{M}(L)} p(\mathbf{D}|\mathcal{M})p(\mathcal{M})}{\sum_{\mathcal{M}\in\mathbf{M}} p(\mathbf{D}|\mathcal{M})p(\mathcal{M})}. \tag{5.3}$$

Proof. Follows from summing the normalized posterior likelihoods of all MAGs that entail that statement through m-separation, similar to eq.(5.2). □

As stated, in many cases considering only a small subset of the variables in $\mathbf{V}$ is already sufficient to infer L. But that also implies that there are multiple subsets that imply L, each with different probability estimates. As these relate to different sets of variables, they should not be combined as in standard multiple hypothesis tests, but instead we want to look for the *maximum* value that can be found.

Lemma 5.2. *Let $\mathbf{D}$ be a data set over variables $\mathbf{V}$. Then $\forall\mathbf{X}\subseteq\mathbf{V} : p(L|\mathbf{D}) \geq \sum_{\mathcal{M}\in\mathbf{M}_\mathbf{X}(L)} p(\mathcal{M}|\mathbf{D})$.*

Proof. Let $p(\mathcal{M}_\mathbf{V}|\mathbf{D}_\mathbf{V})$ be the posterior probability of MAG $\mathcal{M}_\mathbf{V}$ given data $\mathbf{D}_\mathbf{V}$ over variables $\mathbf{V}$. Let $\mathcal{M}(\mathbf{X})$ denote the MAG $\mathcal{M}$ marginalized to variables $\mathbf{X}$, then:

$$\begin{aligned} p(L|\mathbf{D}_\mathbf{V}) &= \sum_{\mathcal{M}_\mathbf{V}\in\mathbf{M}_\mathbf{V}(L)} p(\mathcal{M}_\mathbf{V}|\mathbf{D}_\mathbf{V}) \\ &\geq \sum_{\mathcal{M}_\mathbf{V}\in\mathbf{M}_\mathbf{V}(L):\mathcal{M}_\mathbf{V}(\mathbf{X})\in\mathbf{M}_\mathbf{X}(L)} p(\mathcal{M}_\mathbf{V}|\mathbf{D}_\mathbf{V}) \\ &= \sum_{\mathcal{M}_\mathbf{X}\in\mathbf{M}_\mathbf{X}(L)} p(\mathcal{M}_\mathbf{X}|\mathbf{D}_\mathbf{V}) \end{aligned}$$

Where by definition $p(\mathcal{M}_\mathbf{X}|\mathbf{D}_\mathbf{V}) = \sum_{\mathcal{M}_\mathbf{V}\in\mathbf{M}_\mathbf{V}:\mathcal{M}_\mathbf{V}(\mathbf{X})=\mathcal{M}_\mathbf{X}} p(\mathcal{M}_\mathbf{V}|\mathbf{D}_\mathbf{V})$. The inequality follows from the fact that, by definition, no marginal MAG $\mathcal{M}(\mathbf{X})$ entails a statement not entailed by $\mathcal{M}$, whereas the converse can (and does) occur. □

It means that while searching for logical causal statements Ł, it makes sense to keep track of the maximum probabilities obtained so far. However, computing $p(\mathcal{M}|\mathbf{D})$ for $\mathcal{M}\in\mathbf{M}_\mathbf{X}$ still involves computing likelihoods over all structures over $\mathbf{V}$, which is precisely what we want to avoid. A reasonable approximation is provided by $p(\mathcal{M}|\mathbf{D}_\mathbf{X})$, i.e. the estimates obtained by only including data in $\mathbf{D}$ from the variables $\mathbf{X}$. It means that the lower bound is no longer guaranteed to hold universally, but should still be adequate in practice. As a result, a conservative estimate is given by:

$$p(L|\mathbf{D}) \gtrapprox \max_{\mathbf{X}\subset\mathbf{V}} \sum_{\mathcal{M}\in\mathbf{M}_\mathbf{X}(L)} p(\mathcal{M}|\mathbf{D}_\mathbf{X}) \tag{5.4}$$

5.3.2 Obtaining likelihood estimates

If we know that the 'true' structure over a subset $\mathbf{X} \subseteq \mathbf{V}$ takes the form of a DAG, then computing the required likelihood estimates $p(\mathbf{D}_\mathbf{X}|\mathcal{G})$ is relatively straightforward. Cooper and Herskovits [1992] showed that, under some reasonable assumptions, for discrete random variables the integral (5.1) has a closed-form solution. In the form presented in [Heckerman *et al.*, 1995] this score is known as the *Bayesian Dirichlet* (BD) metric:

$$p(\mathbf{D}|\mathcal{G}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N_{ij} + N'_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + N'_{ijk})}{\Gamma(N'_{ijk})}, \tag{5.5}$$

with n the number of variables, r_i the multiplicity of variable X_i, q_i the number of possible instantiations of the parents of X_i in $\mathcal{G}$, N_{ijk} the number of cases in data set $\mathbf{D}$ in which variable X_i has the value $r_{i(k)}$ while its parents are instantiated as $q_{i(j)}$, and with $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. The $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$ represent the pseudocounts for a Dirichlet prior over the parameters in the corresponding CPTs.

Different strategies for choosing the prior exist: for example, choosing $N'_{ijk} = 1$ (uniform prior) leads to the original *K2-metric*, see [Cooper and Herskovits, 1992]. Setting $N'_{ijk} = N'/(r_i q_i)$ gives the popular *BDeu*-metric, which is *score equivalent* in the sense that structures from the same equivalence class $[\mathcal{G}]$ receive the same likelihood score, cf. [Buntine, 1991]. In this article, we opt for the $K2$-metric, as it seems more appropriate in causal settings [Heckerman *et al.*, 1995]. But having to consider only one instance of every equivalence class may prove a decisive advantage of the BDe(u)-metric in future extensions.

However, eq.(5.5) only applies to DAGs. We know that, even when assuming causal sufficiency applies for the variables $\mathbf{V}$, considering arbitrary subsets size $|\mathbf{X}| \geq 4$ in general will require MAG representations to account for common causes that are not in $\mathbf{X}$. Extending the derivation of eq.(5.5) requires additional assumptions on multiplicity and number of the hidden variables, and turns the nice closed-form solution into an intractable problem that requires approximation, e.g. through sampling [Heckerman *et al.*, 1999]. This would make each step in our approach much more expensive. Recently, Evans and Richardson [2010] showed a maximum likelihood approach to fit acyclic directed mixed graphs (a superset of MAGs) directly on binary data. Unfortunately, this method cannot provide the likelihood estimates per model we need for our purposes. Silva and Ghahramani [2009] do present a Bayesian approach, but need to put additional constraints on the distribution in the form of (cumulative) Gaussian models.

In short, even though we would like to use MAGs to compute $p(\mathbf{D}_\mathbf{X}|\mathcal{M})$ directly in eq.(5.3), at the moment we have to rely on DAGs to obtain approximations to the 'true' value. This will result in less accurate reliability estimates for $p(L|\mathbf{D})$, but also means that we may miss certain pieces of information, or, even worse, that the inference may become invalid.

5.3.3 Inference from unfaithful DAGs

Fortunately we can show that, even when the true independence structure over a subset $\mathbf{X} \subset \mathbf{V}$ is a MAG, we can still do valid inference via $p(L|\mathbf{D}_\mathbf{X})$ from likelihood scores over an exhaustive set of DAGs over $\mathbf{X}$, provided we account for unfaithful DAG representations. This part discusses unfaithful inference, and how the mapping from structures to logical causal statements can be modified. Much of what we show builds on [Bouckaert, 1995].

In the large-sample limit, the Bayesian likelihood score picks the smallest DAG structure(s) that can capture the observed probability distribution exactly.

Definition 5.3 (**Optimal uDAG**). *A DAG $\mathcal{G}$ is an (unfaithful) **uDAG** approximation to a MAG $\mathcal{M}$ over a set of nodes $\mathbf{X}$, iff for any probability distribution $p(\mathbf{X})$, generated by an underlying causal graph faithful to $\mathcal{M}$, there is a set of parameters $\mathbf{\Theta}$ such that the Bayesian network $\mathcal{B} = (\mathcal{G}, \mathbf{\Theta})$ encodes the same distribution $p(\mathbf{X})$. The uDAG is **optimal** if there exists no uDAG to $\mathcal{M}$ with fewer free parameters.*

In other words: a uDAG is just a DAG for which we do not know if it is faithful or not. Reading in/dependence relations from a uDAG goes as follows:

Lemma 5.4. *Let $\mathcal{B} = (\mathcal{G}, \mathbf{\Theta})$ be a Bayesian network over a set of nodes $\mathbf{X}$, with $\mathcal{G}$ a uDAG for some MAG that is faithful to a distribution $p(\mathbf{X})$. Let $\mathcal{G}_{X\|Y}$ be the graph obtained by eliminating the edge $X - Y$ from $\mathcal{G}$ (if present). Then, if $X \perp\!\!\!\perp_{\mathcal{G}_{X\|Y}} Y \,|\, \mathbf{Z}$ then:*

$$(X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z}) \Leftrightarrow (X \perp\!\!\!\perp_{p} Y \,|\, \mathbf{Z}).$$

Proof sketch. The independence rule $(X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z}) \Rightarrow (X \perp\!\!\!\perp_{p} Y \,|\, \mathbf{Z})$ follows from the standard rule for d-separation, see e.g. [Pearl, 1988].
The dependence rule $(X \perp\!\!\!\perp_{\mathcal{G}_{X\|Y}} Y \,|\, \mathbf{Z}) \wedge (X \not\perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z}) \Rightarrow (X \not\perp\!\!\!\perp_{p} Y \,|\, \mathbf{Z})$ is similar to the 'coupling' theorem (3.11) in [Bouckaert, 1995], but stronger. As we assume a faithful MAG, a dependence $X \not\perp\!\!\!\perp_{p} Y \,|\, \mathbf{Z}$ cannot be destroyed by in/excluding a node U that has no unblocked path in the underlying MAG to X and/or Y given $\mathbf{Z}$. This eliminates one of the preconditions in the coupling theorem. See Appendix 5.A for details. □

So, in a uDAG all independencies from d-separation are still valid, but the identifiable dependencies are restricted.

Example 5.5. *Treating the uDAG in Figure 5.1(b) as a faithful DAG would suggest $X \perp\!\!\!\perp_{p} T \,|\, [Z]$, and hence $(Z \Rightarrow X) \vee (Z \Rightarrow T)$. This is wrong: Figure 5.1(a) shows that Z is ancestor of neither X nor T. Lemma 5.4 would not make this mistake, as it allows to deduce $X \perp\!\!\!\perp_{p} T \,|\, Z$, but not the erroneous $X \not\perp\!\!\!\perp_{p} T$.*

We can generalize Lemma 5.4 to *indirect* dependencies.

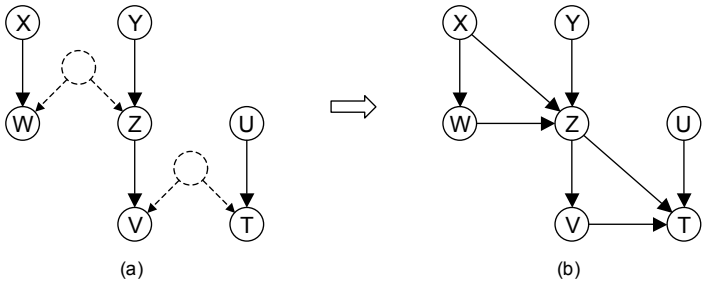

Figure 5.1: (a) causal DAG with hidden variables, (b) uDAG with unfaithful $X \perp\!\!\!\perp_{\mathcal{G}} T \,|\, [Z]$

Lemma 5.6. *Let $\mathcal{G}$ be a uDAG for a faithful MAG $\mathcal{M}$. Let X, Y, and $\mathbf{Z}$ be disjoint (sets of) nodes. If $\pi = \langle X, .., Y \rangle$ is the only unblocked path from X to Y given $\mathbf{Z}$ in $\mathcal{G}$, then $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$.*

Example 5.7. *With Lemma 5.6 we infer from Figure 5.1(b) that $X \not\perp\!\!\!\perp_p T \,|\, \{V, W\}$. We also find that $Y \not\perp\!\!\!\perp_p V$, from which, in combination with $Y \perp\!\!\!\perp_p V \,|\, [Z]$, we (rightly) conclude that $(Z \Rightarrow Y) \vee (Z \Rightarrow V)$, see (a).*

In general, lemmas 5.4 and 5.6 assert different dependencies for different uDAG members of the same equivalence class. If the uDAG $\mathcal{G}$ is also *optimal*, then all in/dependence statements from any uDAG member of the corresponding equivalence class $[\mathcal{G}]$ are valid. In that case we can do the inference based on the PAG representation $\mathcal{P}$ of $[\mathcal{G}]$. This provides additional information, but also simplifies some inference steps. Again, see Appendix 5.A for details.

Identifying an absent causal relation (arrowhead) $X \not\Rightarrow Y$ from an optimal uDAG becomes identical to the inference from a faithful MAG. Let a *potentially directed path* (p.d.p.) be a path in a PAG that could be oriented into a directed path by changing circle marks into appropriate tails/arrowheads, then

Lemma 5.8. *Let $\mathcal{G}$ be an optimal uDAG to a faithful MAG $\mathcal{M}$, then the absence of a causal relation $X \not\Rightarrow Y$ can be identified, iff there is no potentially directed path from X to Y in the PAG $\mathcal{P}$ of $[\mathcal{G}]$.*

Proof sketch. The optimal uDAG $\mathcal{G}$ is obtained by (only) adding edges between variables in the MAG $\mathcal{M}$ to eliminate invariant bi-directed edges, until no more are left. At that point the uDAG is a representative of the corresponding equivalence class $\mathcal{P}$ (Theorem 2 in Zhang [2008]). For any faithful MAG all and only the nodes not connected by a p.d.p. in the corresponding PAG have a definite non-ancestor relation in the underlying causal graph. At least one uDAG instance in the equivalence class of an optimal uDAG over a given skeleton leaves the ancestral relations of the original MAG intact. Therefore, any remaining invariant arrowhead in the PAG $\mathcal{P}$ matches a non-ancestor relation in the original MAG. □

For the presence of causal relations (tails) a similar, but more complicated criterion can be found, see Appendix 5.B.2. Ultimately, the impact of having to use uDAGs boils down to a modified mapping of structures to logical causal statements, based on the inference rules above.

Finally, it is worth mentioning that in the large-sample limit, matching uDAGs over increasing sets of nodes we are guaranteed to find all independencies needed to obtain the skeleton, as well as all invariant arrowheads and many invariant tails. However, as the primary goal remains to improve accuracy/robustness when the large-sample limit does *not* apply, we do not pursue this matter further here.

5.3.4 Consistent prior over structures

The computation of $p(L|\mathbf{D_X})$ requires a prior distribution $p(\mathcal{M})$ over the set of MAGs over $\mathbf{X}$. A straightforward solution is to use a uniform prior, assigning equal probability to each $\mathcal{M} \in \mathbf{M}$. Alternatively, we can use a predefined function that penalizes complexity or deviation w.r.t. some reference structure [Chickering, 2002; Heckerman *et al.*, 1995]. If we want to exploit score-equivalence with the BDe(u) metric in eq.(5.5), we can weight DAG representatives according to the size of their equivalence class.

If we have background information on expected (or desired) properties of the structure, such as max. node degree, average connectivity, or small-world/scale-free networks, we can use this to construct a prior $p(\mathcal{M})$ through *sampling*: generate random graphs over all variables in accordance with the specified characteristics, sample a random subset of K variables from that graph, and compute the marginal uDAG/MAG structure over that subset. Repeat until the empirical distribution of structures over K variables obtained through this sampling procedure converges to an adequate approximation of the prior $p(\mathcal{M})$. Averaging over structures that are PAG-isomorphs (equivalence classes identical under relabeling) improves both consistency and convergence.

Irrespective of the method to obtain a prior, it is essential to ensure it is also *consistent* over structures of different size. Perhaps surprisingly, this is *not* obtained by applying the same strategy at different levels: a uniform distribution over DAGs over $\{X, Y, Z\}$ implies $p(\text{“}X \perp\!\!\!\perp Y\text{”}) = 6/25$, whereas a uniform distribution over two-node DAGs implies $p(\text{“}X \perp\!\!\!\perp Y\text{”}) = 1/3$. We obtain a consistent multi-level prior by starting from a preselected level K, and then extend to different sized structures through marginalization.

5.4 The BCCD algorithm

We can now turn the results from the previous section into a working algorithm. The implementation largely follows the outline in Algorithm 5.1, except that now uDAGs (instead of MAGs) are used to obtain a list of logical causal statements (instead of

independencies), and that logical inference takes the place of the orientation rules, resulting in the Bayesian Constraint-based Causal Discovery (BCCD) algorithm.

Algorithm 5.2 Bayesian Constrained-based Causal Discovery (BCCD)

In : database $\mathbf{D}$ over variables $\mathbf{V}$, background info $\mathcal{I}$
Out: causal relations matrix $\mathbf{M}_C$, causal PAG $\mathcal{P}$
Stage 0 - Mapping
1: $\mathcal{G} \times \text{Ł} \leftarrow Get\_uDAG\_Mapping(\mathbf{V}, K_{max} = 5)$
2: $p(\mathcal{G}) \leftarrow Get\_Prior(\mathcal{I})$
Stage 1 - Search
3: fully connected $\mathcal{P}$, initialize list $\forall L \in \text{Ł} : p(L) = 0$, $K = 0$, $\theta = 0.5$
4: **while** $K \leq K_{max}$ **do**
5: **for all** $X \in \mathbf{V}, Y \in Adj(X)$ in $\mathcal{P}$ **do**
6: **for all** $\mathbf{Z} \subseteq Adj(X)_{\setminus Y}$, $|\mathbf{Z}| = K$ **do**
7: $\mathbf{W} \leftarrow Check\_Unprocessed(X, Y, \mathbf{Z})$
8: $\forall \mathcal{G} \in \mathcal{G}_{\mathbf{W}}$: compute $p(\mathcal{G}|\mathbf{D}_{\mathbf{W}})$
9: $\forall L : p(L_{\mathbf{W}}|\mathbf{D}_{\mathbf{W}}) \leftarrow \sum_{\mathcal{G} \to L_{\mathbf{W}}} p(\mathcal{G}|\mathbf{D}_{\mathbf{W}})$
10: $\forall L : p(L) \leftarrow \max(p(L), p(L_{\mathbf{W}}|\mathbf{D}_{\mathbf{W}}))$
11: **for all** $\{W_i, W_j\} \subset \mathbf{W}$ **do**
12: **if** $p(\text{“}W_i \not\leftrightarrow W_j\text{”}|\mathbf{D}_{\mathbf{W}}) > \theta$ **then**
13: remove edge $W_i - W_j$ from $\mathcal{P}$ (if present)
14: **end if**
15: **end for**
16: **end for**
17: **end for**
18: $K = K + 1$
19: **end while**
Stage 2 - Inference
20: $\mathbf{L}_C =$ empty 3D-matrix size $|\mathbf{V}|^3$, $i = 1$
21: Ł $\leftarrow Sort\_Descending$ (Ł, $p(L)$)
22: **while** $p(L_i) > \theta$ **do**
23: $\mathbf{L}_C \leftarrow Run\_Causal\_Logic(\mathbf{L}_C, L_i)$
24: $i \leftarrow i + 1$
25: **end while**
26: $\mathbf{M}_C \leftarrow Get\_Causal\_Matrix(\mathbf{L}_C)$
27: $\mathcal{P} \leftarrow Map\_To\_PAG(\mathcal{P}, \mathbf{M}_C)$

A crucial step in the algorithm is the mapping $\mathcal{G} \times$ Ł from optimal uDAG structures to causal statements in line 9. This mapping is the same for each run, so it can be precomputed from the rules in section 5.3.3, and stored for use afterwards (line 1). The uDAGs $\mathcal{G}$ are represented as adjacency matrices. For speed and efficiency purposes, we choose to limit the structures to size $K \leq 5$, which gives a list of 29,281 DAGs over 5 variables [Robinson, 1973], and so also 29,281 uDAGs at this

highest level. For details about representation and rules, see the Appendix.

The adjacency search (lines 3-15), loops over subsets from neigbouring nodes for identifiable causal information, while keeping track of adjacencies that can be eliminated (line 11). For structures over five or more nodes we need to consider nodes from FCI's *Possible-D-Sep* set [Spirtes *et al.*, 1999]. In practice, it rarely finds any additional (reliable) independencies, and we opt to skip this step for speed and simplicity (line 6), similar to the RFCI approach in [Colombo *et al.*, 2011]. As the set $\mathbf{W} = \{X, Y\} \cup \mathbf{Z}$ can be encountered in different ways, line 7 checks if the test on that set has been performed already. A list of probability estimates $p(L|\mathbf{D})$ for each logical causal statement is built up (line 10), until no more information is found.

The inference stage (lines 16-21) then processes the list Ł in decreasing order of reliability, until the threshold is reached. Statements in Ł are added one-by-one to the matrix of logical causal statements $\mathbf{L}_C$ (encoding identical to Ł), with additional information inferred from the causal logic rules. Basic conflict resolution is achieved through not overriding existing information (from more reliable statements). The final step (lines 22,23) retrieves all explicit causal relations in the form of a causal matrix $\mathbf{M}_C$, and maps this onto the skeleton $\mathcal{P}$ obtained from Stage 1 to return a graphical PAG representation.

We verified the resulting mapping from uDAG rules to logical statements (line 1) against a *brute-force approach* that checks the intersection of all logical causal statements implied by all MAGs that can have a particular uDAG as an optimal approximation. We found that at least for uDAG structures up to five nodes our rules are still sound and complete. Interestingly enough, for uDAGs up to four nodes all implied statements are *identical* to those that would be obtained if we just treated uDAGs as faithful DAGs. Only at five+ nodes do we have to take into account that the DAGs we score with eq.(5.5) may be unfaithful.

5.5 Experimental Evaluation

We tested various aspects of the BCCD algorithm in many different circumstances, and against various other methods. The principal aim at this stage is to verify the viability of the Bayesian approach. We compare our results and that of other methods from data against known ground-truth causal models. For that, we generate random causal graphs with certain predefined properties (adapted from Melancon *et al.* [2000]; Chung and Lu [2002]), generate random data from this model, and marginalize out one or more hidden confounders. We looked at the impact of the number of data points, size of the models, sparseness, choices for parameter settings etc. on the performance to get a good feel for expected strengths and weaknesses in real-world situations.

It is well-known that the relative performance of different causal discovery methods can depend strongly on the performance metric and/or specific test problems

used in the evaluation. Therefore, we will not claim that our method is inherently better than others based on the experimental results below, but simply note that the fact that in nearly all test cases the BCCD algorithm performed as good or better than other methods is a clear indication of the viability and potential of this approach.

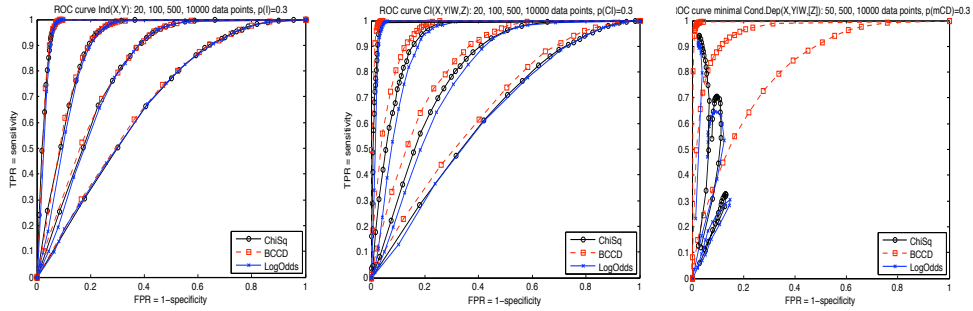

Figure 5.2: BCCD approach to (complex) independence test in eq.(5.2) for different sized data sets; (a) independence test $X \perp\!\!\!\perp_p Y$ (for, left to right: 10000, 500, 100, and 20 data points per set), (b) conditional independence $X \perp\!\!\!\perp_p Y \,|\, W, Z$ (again: 10000, 500, 100, and 20 data points), (c) *minimal* conditional dependence $X \not\perp\!\!\!\perp_p Y \,|\, W \cup [Z]$ (now for: 10000, 500, resp. 50 data points).

First we implemented the BCCD approach as a (minimal) independence test. Figure 5.2 shows a typical example in the form of ROC-curves for different sized data sets, compared against a chi-squared test and a Bayesian log-odds test from [Margaritis and Bromberg, 2009], with the α decision threshold for the p-value as tuning parameter for the chi-squared test along the curve, idem for the log-odds decision threshold, and with the prior on independence as the tuning parameter for BCCD. For 'regular' independence tests, Figure 5.2(a), there was no significant difference (BCCD marginally ahead). For the *conditional* independencies in Figure 5.2(b) the BCCD version started to outperform the other tests, an effect that becomes stronger as the size of the conditioning set increases. The reason for this behaviour is that 'traditional' tests look for dependencies in any of the separate subsets of data (slices) for each possible value of the conditioning set, effectively performing multiple tests on smaller data sets, whereas the BCCD test always uses all data on the variables in one single test. A drawback is that the BCCD test becomes relatively more expensive. But the real difference in performance only becomes apparent when testing for complex independence statements, involving combination(s) of in- and dependencies, such as in the minimal dependence in Figure 5.2(c). Other methods always reject for both high and low decision thresholds (on resp. the independence and the dependence), resulting in the awkward looped curves in (c), but the BCCD method has no such problem.

Obviously, in such cases we can/should introduce multiple thresholds for the chi-squared test and the likes, but then the problem becomes how to choose the optimal

combination; and even then it would still lag behind BCCD for the same reason as in (b). For BCCD the optimal setting is always given by the prior probability on that statement, with similar performance in a reasonable interval around that setting, as illustrated by Figure 5.3.

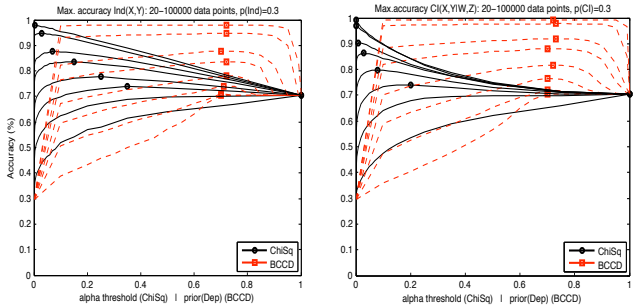

Figure 5.3: Accuracy for chi-squared and BCCD independence test as function of threshold setting for different sized data sets (lines from bottom to top: 20, 50, 100, 200, 500, 1000, 10000, 100000 data points); (a) independence $X \perp\!\!\!\perp_p Y$, (b) conditional independence $X \perp\!\!\!\perp_p Y \,|\, W, Z$; the bold mark on each line indicates the optimal setting.

Note that for the chi-squared test the optimal decision threshold is not only highly dependent on the *size* of the data set (the larger $\mathbf{D}$, the smaller the optimal decision threshold α), but also on the size of the conditioning set.

Next we look at the performance of the BCCD algorithm itself, against two other state-of-the-art methods that can handle hidden confounders: FCI as the de facto benchmark, and its equivalent adapted from conservative PC. For the evaluation we use two complementary metrics: (1) the *PAG accuracy* looks at the graphical causal model output and counts the number of edge marks that matches the PAG of the true equivalence class (excluding self-references), and (2) the *causal accuracy* looks at the proportion of all causal decisions, either explicit as BCCD does or implicit from the PAG for FCI, that are correct compared to the generating causal graph.

Figure 5.4 shows typical results for the PAG accuracy test: for a data set of 1000 records the *PAG accuracy* for both FCI and conservative FCI peaks around a threshold $\alpha \approx 0.05$ - lower for more records, higher for less - with conservative FCI consistently outperforming standard FCI. The BCCD algorithm peaks at a cut-off value $\theta \in [0.45, 0.7]$ with an accuracy that is slightly higher than the maximum for conservative FCI. The PAG accuracy tends not to vary much over this interval, making the default choice $\theta = 0.5$ fairly safe, even though the number of invariant edge marks does increase significantly when lowering θ (more correct decisions, but also more mistakes).

In summary: we found that in most circumstances conservative FCI outperforms vanilla FCI by about $3 - 4\%$ in terms of PAG accuracy and slightly more in terms of causal accuracy. In its standard form, with a uniform prior over structures

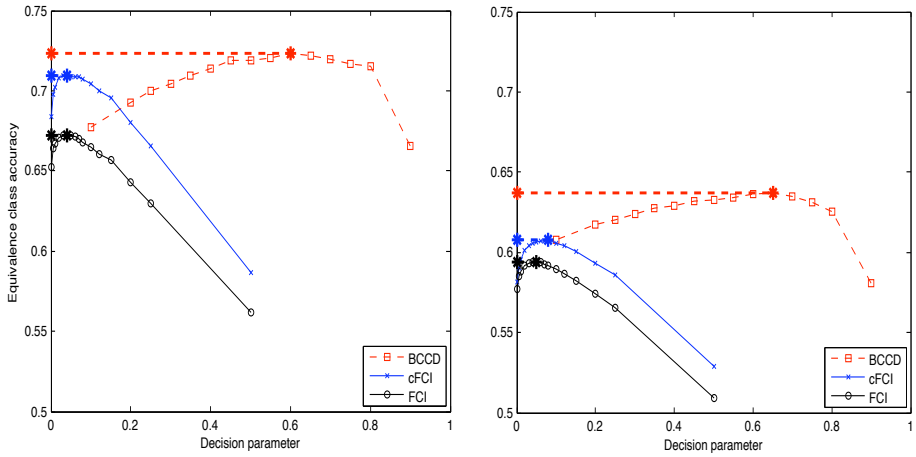

Figure 5.4: Equivalence class accuracy (% of edge marks in PAG) vs. decision parameter; for BCCD and (conservative) FCI, from 1000 random models; (a) 6 observed nodes, 1-2 hidden, 1000 points, (b) idem, 12 observed nodes.

of 5 nodes, the BCCD algorithm consistently outperforms conservative FCI by a small margin of about $1-2\%$ at default decision thresholds ($\theta = 0.5$ for BCCD, $\alpha = 0.05$ for FCI). Including additional tests / nodes per test and using an extended mapping often increases this difference to about $2-4\%$ at optimal settings for both approaches (cf. Figure 5.4). This gain does come at a cost: BCCD has an increase in run-time of about a factor 2.5 compared to conservative FCI, which in turn is marginally more expensive than standard FCI. Evaluating many large structures can increases this cost even further, unless we switch to evaluating equivalence classes via the BDe metric in §5.3.2.

To obtain additional insight in (the quality of) the output of the different algorithms, we also looked at their orientation behavior in the form of the so called *confusion matrices* for the corresponding PAGs. Table 5.1 shows the average orientation behavior for edge marks in the PAG model at standard threshold settings for each of the three methods. Each cell in the confusion matrix indicates the percentage of (row,column) = (true,output) oriented edge marks in the PAG output. For example, cell $(\rightarrow, -\circ) = 2.1\%$ in the FCI-table on the l.h.s. indicates that on average 2.1% of the edge marks in the PAG output of standard FCI were circle marks that should have been invariant arrowheads. The main (red) diagonal indicates the percentage of correctly oriented edge marks of that type in the output, with the sum-total in the top-left in bold the total average PAG-accuracy. The cells in the bottom row of each matrix sum all contributions in that column, and so indicate the average percentage of the corresponding edge mark in the output PAG. For example the bold number in the bottom-right represents the number of circle marks in the output.

68.2	⨯	→	—	−○
⨯	42.5	1.7	0.0	0.6
→	3.0	11.6	0.1	2.1
—	0.9	3.1	1.2	1.5
−○	4.8	12.3	1.6	12.9
total	51.2	28.7	2.9	**17.1**

FCI

74.5	⨯	→	—	−○
⨯	42.5	0.9	0.0	1.5
→	3.0	10.7	0.1	3.1
—	0.9	1.4	1.5	2.9
−○	4.8	5.9	1.1	19.8
total	51.2	18.9	2.7	**27.3**

cFCI

75.5	⨯	→	—	−○
⨯	44.0	0.2	0.1	0.5
→	3.2	10.9	0.5	2.3
—	0.9	1.7	2.4	1.7
−○	5.2	5.5	2.7	18.2
total	53.3	18.3	5.7	**22.7**

BCCD

Table 5.1: Confusion matrices showing average orientation behaviour in PAG output: (a) Standard FCI algorithm, (b) Conservative FCI, (c) BCCD ; rows = true value, columns = output edge mark (1000 random models over 6 nodes, 10,000 data points).

We can recognize how FCI is very informative, in the sense that it makes the most explicit decisions (only 17.1% circle marks), but also includes many mistakes. Conservative FCI starts from the same skeleton (first column is identical to FCI), but is then more reluctant to orient uncertain v-structures. It manages to increase the overall accuracy as a result (sum of diagonal entries: 74.5 vs. 68.2), but does so by simply leaving a lot of default "don't know" circle marks (27.3%), which are much more likely than the invariant arrowhead/tail options to be correct (= increase accuracy), but do not provide any real information. The BCCD algorithm increases this accuracy even further, but also manages to be much more informative than cFCI: it avoids mostly erroneous (overly conservative) circle marks, resulting in a slight increase in overall performance.

However, the main benefit of the BCCD approach lies not in a slight improvement in accuracy, but in the added insight it provides into the generated causal model: even in this simple form, the algorithm gives a useful indication of which causal decisions are reliable and which are not, which seems very useful to have in practice.

Figure 5.5 depicts the *causal accuracy* as a function of the tuning parameter for the three methods. The BCCD dependency is set against $(1-\theta)$ so that going from $0 \rightarrow 1$ matches processing the list of statements in decreasing order of reliability. As hoped/expected: changing the decision parameter θ allows to access a range of accuracies, from a few very reliable causal relations to more but less certain indications. In contrast, the accuracy of the two FCI algorithms cannot be tuned effectively through the decision parameter α. The reason behind this is apparent from Figure 5.2(b): changing the decision threshold in an independence test shifts the balance between dependence and independence decisions, but it cannot identify

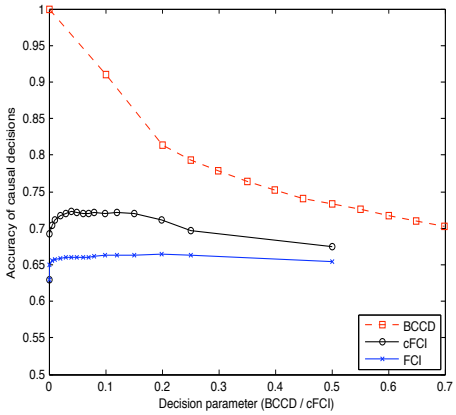

Figure 5.5: Accuracy of causal decisions as a function of the decision parameter; averages from 1000 data sets of 1000 points each from models over 6 observed nodes.

or alter the balance in favor of more reliable decisions. We consider the fact that the BCCD algorithm *can* do exactly that as one of the most promising aspects of the Bayesian approach.

5.6 Discussion and future work

The experimental results confirm that the Bayesian approach is both viable and promising: even in a basic implementation the BCCD algorithm already outperforms other state-of-the-art causal discovery algorithms. It yields slightly better accuracy, both for optimal and standard settings of the decision parameters. Furthermore, BCCD comes with a decision threshold that is easy to interpret and can be used to vary from making just a few but very reliable causal statements to many possibly less certain decisions. Perhaps counterintuitively, changing the confidence level in (conservative) FCI does not lead to similar behavior as it only affects the balance between dependence and independence decisions, which in itself does not increase the reliability of either.

We do not claim that the current optimal uDAG mapping is complete in the sense that it is guaranteed to extract the maximum amount of causal information. The brute-force check at the end of section 5.4 showed that the combined inference rules cover all optimal uDAGs up to five nodes: they could be extended to infer more dependencies and/or causal information from larger graphs. If possible this should take the form of an easy-to-use graphical criterion in the vein of d-separation, but based on our impressions so far this seems rather ambitious.

But there are many other opportunities left for improvement, both in speed and accuracy. An easy option is to try to squeeze out as much as possible from the current framework: scoring equivalence classes with the BDe metric should

bring a performance gain for large structures, without any obvious drawback, as the likelihood contributions of all members are aggregated anyway. Furthermore, we now sometimes miss likely causal information in the mapping from uDAGs to causal statements L, because they are not implied in a small percentage of possible matching MAGs. To use this information we can choose to weigh the different causal statements L by the proportion of underlying MAGs in which they hold instead of a logical yes/no value. This results in a more informative mapping $\mathcal{G} \rightarrow \text{Ł}$, for more accurate estimates $p(L|\mathbf{D})$. It seems not possible to infer this type of mapping from graphical rules alone, and so it would have to rely on brute-force computation.

For larger (sub)graphs some form of Monte Carlo sampling can be applied to obtain the (weighted) mapping to causal statements. Reasonable reliability estimates can be obained from a limited number of high scoring alternatives, see, e.g. [Bouckaert, 1995]. It would be very expensive to compute, but could provide valuable information to decide on borderline cases, or simply as an independent confirmation for parts of the inferred structure.

Further improvements

An interesting question is how far off from the theoretical optimum we are: what is the maximum attainable accuracy of all causal information that can be extracted from a given data set? At the moment it is not clear whether we are fighting for the last few tenths of a percent or if substantial gains can still be made.

A hopeful but ambitious path is to tackle some of the fundamental problems: scoring MAGs (or even PAGs) directly would eliminate the need for 'unfaithful inference' altogether [Evans and Richardson, 2010]. This would improve both the mapping and the probability estimates of the inferred logical causal statements, although it is likely to impact the overall running time. Sampling MAGs could also be employed to obtain or confirm reliability estimates for causal information derived from *combinations* of separately obtained logical statements, as described in [Claassen and Heskes, 2011a]. We expect such combinations to be highly dependent, but especially for less certain ones, e.g. two statements with $p(L_{1,2}|\mathbf{D}) = 0.6$, the difference between a fully independent estimate: $p(L_1|\mathbf{D}) \cdot p(L_2|\mathbf{D}) = 0.36$, and a fully dependent estimate: $min\left(p(L_1|\mathbf{D}), p(L_2|\mathbf{D})\right) = 0.6$, is considerable. Having a means to obtain a more principled reliability estimate for the combination can improve the overall accuracy of the BCCD algorithm.

Finally, we would like to turn the current reliability estimates into principled probabilities for initial logical causal statements and new statements derived during the inference process: this would improve conflict resolution, and would ultimately allow to give a meaningful estimate for the probability of all causal relations inferred from a given data set. In fact, for high values the relatively crude BCCD reliability estimates in Figure 5.5 already form a decent approximation to $p(X \Rightarrow Y \,|\, \mathbf{D})$: for example at $(1 - \theta) = 0.2$, corresponding to using all statements with reliability ≥ 0.8, the inferred causal relations are also correct in about 80% of the cases. Still,

much is needed before we can claim to have a proper probability estimate.

5.A Appendix: Probabilistic inference from uDAGs

This section describes how to read in/dependencies from uDAGs; it extends results in [Bouckaert, 1995] through the added assumption of an underlying *faithful MAG*; see also [Claassen and Heskes, 2012b].

Note: a DAG $\mathcal{G}$ is an (unfaithful) **uDAG** approximation to a MAG $\mathcal{M}$ over a set of nodes $\mathbf{X}$, iff for any probability distribution $p(\mathbf{X})$, generated by an underlying causal graph faithful to $\mathcal{M}$, there is a set of parameters $\mathbf{\Theta}$ such that the Bayesian network $\mathcal{B} = (\mathcal{G}, \mathbf{\Theta})$ encodes the same distribution $p(\mathbf{X})$.

We use $\mathcal{G}$ to explicitly indicate a DAG, $\mathcal{M}$ for a MAG, and $\mathcal{P}$ for a PAG. For details regarding other graph theoretical / causal concepts and terminology used in this Appendix the reader is referred to sections 2.1-2.3.

A uDAG is a DAG for which we do not know if it is faithful or not. We can apply standard d-separation to read presence or absence of an independence '$X \perp\!\!\!\perp_p Y \mid \mathbf{Z}$?' from a uDAG $\mathcal{G}$, *provided* the set $\mathbf{Z}$ d-separates nodes X and Y when edge $X - Y$ (if present) is removed:

Lemma 5.4. Let $\mathcal{B} = (\mathcal{G}, \mathbf{\Theta})$ be a Bayesian network over a set of nodes $\mathbf{X}$, with $\mathcal{G}$ a uDAG for a MAG $\mathcal{M}$ that is faithful to a distribution $p(\mathbf{X})$. Let $\mathcal{G}_{X\|Y}$ be the graph obtained by eliminating the edge $X - Y$ from $\mathcal{G}$ (if present), then:

$$(X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Z}) \Rightarrow (X \perp\!\!\!\perp_p Y \mid \mathbf{Z}),$$

$$(X \perp\!\!\!\perp_{\mathcal{G}_{X\|Y}} Y \mid \mathbf{Z}) \wedge (X \not\perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Z}) \Rightarrow (X \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}).$$

or, alternatively: if $\mathbf{Z}$ is a set that d-separates X and Y in $\mathcal{G}_{X\|Y}$, then

$$(X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Z}) \Leftrightarrow (X \perp\!\!\!\perp_p Y \mid \mathbf{Z}),$$

Proof sketch. The independence rule ($\Rightarrow$) follows [Pearl, 1988]. The dependence rule (through negation) is similar to the 'coupling' theorem (3.11) in [Bouckaert, 1995], but stronger. As we assume a faithful MAG, a dependence $X \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}$ cannot be destroyed by in/excluding a node U that has no unblocked path in the underlying MAG to X and/or Y given $\mathbf{Z}$. This eliminates one of the preconditions in the coupling theorem (see below). □

So, all independencies from d-separation remain valid, but identifiable dependencies put restricitions on the set $\mathbf{Z}$. For a more formal proof, we first introduce the notion of *coupling*:

Definition 3.10. [Bouckaert, 1995] *In a DAG $\mathcal{G}$, two variables X and Y are **coupled** given $\mathbf{Z}$, denoted $\rangle X, \mathbf{Z}, Y \langle_{\mathcal{G}}$, if: $(X \to Y) \in \mathcal{G}$, $Pa(Y)_{\mathcal{G}} \subseteq (X \cup \mathbf{Z})$, and $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z}$ in $\mathcal{G}_{X\|Y}$ (or vice versa for X).*

The relevance of this notion comes courtesy of the following result, based on the graphoid axioms for independence:

Theorem 3.11. [Bouckaert, 1995] *In a DAG $\mathcal{G}$, if two variables X and Y are coupled given $\mathbf{Z}$ in $\mathcal{G}$ then $X \not\perp\!\!\!\perp Y \,|\, \mathbf{Z}$.*

In contrast with this coupling Theorem, Lemma 5.4 does not require the explicit inclusion of $Pa(Y)_{\mathcal{G}} \subseteq (X \cup \mathbf{Z})$ in the separating set. This is a direct consequence of the (stronger) assumption of an underlying faithful MAG $\mathcal{M}$, as we now show in the detailed proof of Lemma 5.4:

Proof of Lemma 5.4. Let $X \to Y$ in $\mathcal{G}$, and let X and Y be coupled given disjoint sets $\mathbf{Z} \cup \mathbf{W}$, with $\mathbf{W} \subset Pa(Y)$, so that $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z} \cup \mathbf{W}$ by Theorem 3.11. Now assume that also $X \perp\!\!\!\perp_{\mathcal{G}\|} Y \,|\, \mathbf{Z}$, then there is no unblocked path given $\mathbf{Z}$ from X to any $W \in \mathbf{W}$ in $\mathcal{G}$, otherwise the $W \longrightarrow Y$ would imply an unblocked path $\langle X, .., W, Y \rangle$ given $\mathbf{Z}$, in contradiction with $X \perp\!\!\!\perp_{\mathcal{G}\|} Y \,|\, \mathbf{Z}$. As a result, $\forall W \in \mathbf{W}$: $X \perp\!\!\!\perp_{\mathcal{G}} W \,|\, \mathbf{Z}$ in $\mathcal{G}$, so also [Pearl, 1988] $X \perp\!\!\!\perp_p W \,|\, \mathbf{Z}$, and so by definition also $X \perp\!\!\!\perp_M W \,|\, \mathbf{Z}$ in the underlying faithful MAG $\mathcal{M}$.

We now show by contradiction that this implies that X and Y are dependent given $\mathbf{Z}$, i.e. $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$:

Suppose that $X \perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$ while given $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z} \cup \mathbf{W}$ and $X \perp\!\!\!\perp_p \mathbf{W} \,|\, \mathbf{Z}$. For a faithful MAG $\mathcal{M}$ this implies that X and Y are m-separated given only $\mathbf{Z}$, but *not* given $(\mathbf{Z} \cup \mathbf{W})$. An unblocked path π given a set $(\mathbf{Z} \cup \mathbf{W})$ means that all noncolliders on π are not in $\mathbf{Z} \cup \mathbf{W}$ and all colliders on π are in $An(\mathbf{Z} \cup \mathbf{W} \cup \mathbf{S})$. A path π in $\mathcal{M}$ blocked by the removal of $\mathbf{W}$ therefore contains one or more colliders in $An(\mathbf{W})$, as any collider in $\mathbf{Z}$ that blocks the path π would also block the path given $(\mathbf{Z} \cup \mathbf{W})$ before. Let W be the first collider in $An(\mathbf{W})$ encountered along the unblocked path π given $(\mathbf{Z} \cup \mathbf{W})$, then faithfulness implies $X \not\perp\!\!\!\perp_p W \,|\, \mathbf{Z}$, contrary the given. Therefore the assumption $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$ cannot hold, and so we can infer $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$. □

Note that this does *not* follow from the (in)dependence axioms - otherwise it would hold always. As a consequence unblocked paths are always preserved.

Corollary 5.9. *Let $\mathcal{G}$ be a uDAG for a faithful MAG $\mathcal{M}$, then all unblocked paths in $\mathcal{M}$ are preserved in $\mathcal{G}$, i.e. then $X \not\perp\!\!\!\perp_M Y \,|\, \mathbf{Z}$ implies $X \not\perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z}$.*

Proof. Follows immediately from Lemma 5.4. □

We can apply Lemma 5.4 twice to get a two-step, *indirect* dependence, as in Example 5.5:

Lemma 5.10. *Let $\mathcal{G}$ be a uDAG for a faithful MAG $\mathcal{M}$. Let X, Y, $\mathbf{Z}$, W be disjoint (sets of) nodes with $\langle X, W, Y\rangle \in \mathcal{G}$, for which $X \perp\!\!\!\perp_{\mathcal{G}\|} W \,|\, \mathbf{Z}$ and $W \perp\!\!\!\perp_{\mathcal{G}\|} Y \,|\, \mathbf{Z}$. If $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z} \cup W$, then $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$.*

Proof. By Lemma 5.4, $X \not\perp\!\!\!\perp_p W \,|\, \mathbf{Z}$ and $W \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$, and so there are unblocked paths from X and Y to W given $\mathbf{Z}$ in $\mathcal{M}$. These paths in $\mathcal{M}$ cannot be both *into* W, because that would imply that X and Y are m-connected given $\mathbf{Z} \cup W$, so $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z} \cup W$, contrary the given $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z} \cup W$. But that means that W is a noncollider in $\mathcal{M}$ on an unblocked path given $\mathbf{Z}$ from X to Y, and so $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$ without W. □

It is easy to see that, similar to the direct edge in Lemma 5.4, the path $\langle X, W, Y\rangle$ is also the *only* unblocked path between X and Y given $\mathbf{Z}$. This turns out to hold generally for identifiable dependence in a uDAG $\mathcal{G}$ to a faithful MAG $\mathcal{M}$: if there is *one and only one* unblocked path in $\mathcal{G}$ from X to Y given $\mathbf{Z}$, then $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$. Though perhaps intuitively obvious ('one unblocked path implies there is no other path/mechanism that can cancel out the dependence due to this one'), to prove it we need to relate observations in $\mathcal{G}$ to configurations in $\mathcal{M}$, and deduce the dependency holds from there.

We first show this for an extended version of Lemma 5.10, in which there is now a single, unblocked directed path from X to Y given $\mathbf{Z} \subset An(Y)$ (so no unblocked paths 'created' by $\mathbf{Z}$). After that we do the general version.

Lemma 5.11. *Let $\mathcal{G}$ be a uDAG for a faithful MAG $\mathcal{M}$. Let X, Y, $\mathbf{Z}$, $\mathbf{W} = \{W_1, .., W_k\}$ be disjoint (sets of) nodes. If $\pi = \langle X \to W_1 \to ..W_k \to Y\rangle$ is the only unblocked path from X to Y given $\mathbf{Z}$ in $\mathcal{G}$, then $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$.*

Proof. By induction. We show that at each step, if there was an unblocked path in $\mathcal{M}$ between X and W_i given $\mathbf{Z}$, then there is also one between X and W_{i+1}. We do this by deriving a contradiction from the assumption that a necessary collider in $\mathcal{M}$ is not in the set $\mathbf{Z}$.

Let $\mathbf{V} = Pa(\mathbf{W})_{\mathcal{G}} \setminus (\mathbf{W} \cup \mathbf{Z})$ represent the set all parents in $\mathcal{G}$ from nodes in $\mathbf{W}$ that are not on π or in $\mathbf{Z}$. Let $\mathbf{V}_i \subseteq \mathbf{V}$ denote a minimal subset of nodes from $\mathbf{V}$ needed to block all alternative paths between $W_i \to W_{i+1}$ in $\mathcal{G}$, such that $W_i \perp\!\!\!\perp_{\mathcal{G}\|} W_{i+1} \,|\, \mathbf{Z} \cup \mathbf{V}_i$. By Lemma 5.4 this implies:

$$W_i \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \cup \mathbf{V}_i \tag{5.6}$$

The invariant after induction step i is:

$$X \not\perp\!\!\!\perp_p W_i \,|\, \mathbf{Z} \tag{5.7}$$

By definition, W_i separates all its predecessors on the unblocked path π in $\mathcal{G}$ from all its successors given $\mathbf{Z} \cup \mathbf{V}$, so that in particular:

$$X \perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \cup \mathbf{V}_i \cup W_i \tag{5.8}$$

Below (in B) we show by contradiction that this implies that:

$$W_i \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \tag{5.9}$$

For the proof in (B) we first show in (A) that:

$$\forall \mathbf{V}'_i \subseteq (\mathbf{V}_i \setminus V) : X \perp\!\!\!\perp_p V \,|\, \mathbf{Z} \cup \mathbf{V}'_i \tag{5.10}$$

From this it follows that the unblocked path in $\mathcal{M}$ given $\mathbf{Z}$ corresponding to equation (5.7) cannot be blocked by any node from $\mathbf{V}_i$, and so also

$$X \not\perp\!\!\!\perp_p W_i \,|\, \mathbf{Z} \cup \mathbf{V}_i \tag{5.11}$$

Equations (5.11) and (5.6) correspond to unblocked paths in $\mathcal{M}$ from to W_i to X and W_{i+1} given $\mathbf{Z} \cup \mathbf{V}_i$. Node W_i cannot be a collider between these two paths in $\mathcal{M}$, otherwise $X \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \cup \mathbf{V}_i \cup W_i$, contrary to (5.8), and so it follows that:

$$X \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \cup \mathbf{V}_i \tag{5.12}$$

But if $X \perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \cup (\mathbf{V}_i \setminus V)$, then V must be a collider in $\mathcal{M}$ with unblocked paths to X and W_{i+1} given $\mathbf{Z} \cup (\mathbf{V}_i \setminus V)$. But that implies $X \not\perp\!\!\!\perp_p V \,|\, \mathbf{Z} \cup (\mathbf{V}'_i \setminus V)$, in contradiction with equation (5.10), and so $X \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \cup (\mathbf{V}_i \setminus V)$. This argument can be repeated to eliminate all nodes $\mathbf{V}_i$, so that we find:

$$X \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \tag{5.13}$$

which corresponds to the invariant for the next step. By induction over the entire path this ultimately proves that $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$.

The remaining elements in this process are detailed below:

The invariant in equation (5.7) holds at the induction base for edge $X \to W_1$. In $\mathcal{G}$ we find that $\mathbf{Z}$ blocks all other paths between X and W_1, otherwise an alternative unblocked path from X to Y given $\mathbf{Z}$ would exist, contrary the given. Therefore $X \perp\!\!\!\perp_{\mathcal{G}\|} W_1 \,|\, \mathbf{Z}$, from which we conclude $X \not\perp\!\!\!\perp_p W_1 \,|\, \mathbf{Z}$ by Lemma 5.4, and so equation (5.7) is satisfied.

Part (A): X is independent of any node in $\mathbf{V}_i$ given $\mathbf{Z}$, eq.(5.10).
Let $V_j \in \mathbf{V}_i \cap Pa(W_j)$ be a node that is a parent from W_j (with $j \in \{0, .., i\}$) needed to block some secondary path (apart from π) in $\mathcal{G}$ between W_i and W_{i+1}. This path consists of one half as a directed path $\pi_1 = V_j \to W_j .. \to W_i$, and the second half as an unblocked path $\pi_2 = V_j .. \to W_{i+1}$ given $\mathbf{Z} \cup (\mathbf{V}_i \setminus V_j)$ in $\mathcal{G}$, connected by noncollider V_j. The path π_2 may contain zero, one or more collider nodes (that are ancestors of/) from $(\mathbf{V}_i \setminus V_j)$. Denote these as $V_{k(1)}, .., V_{k(m)}$ respectively, as encountered along π_2 when going from V_j to W_{i+1}. By contradiction. Suppose that $X \not\perp\!\!\!\perp_p V_j \,|\, \mathbf{Z}$, then there is an unblocked path $X \not\perp\!\!\!\perp_M V_j \,|\, \mathbf{Z}$ in $\mathcal{M}$, and so

(Corollary 5.9) there is an unblocked path π' in $\mathcal{G}$. If this unblocked path π' in $\mathcal{G}$ does *not* go via W_j (the child of V_j, on π), then the path $\pi' + \pi_1 + W_i \rightarrow ..Y$ is an alternative unblocked path between X and Y via V_j given $\mathbf{Z}$, and so is not allowed. If it *does* go via W_j and there are no collider nodes along π_2 (i.e. $m = 0$ in the sequence $V_{k(1)}, .., V_{k(m)}$), then the path $\pi' + \pi_2 + W_{i+1} \rightarrow ..Y$ is an alternative unblocked path between X and Y via V_j given $\mathbf{Z}$, and so is also not allowed. Alternatively, if there are one or more collider nodes along π_2 (i.e. with $m \geq 1$ in the sequence $V_{k(1)}, .., V_{k(m)}$), then the fact that π' can only go via W_j if this node is (ancestor of) a collider in $\mathcal{G}$ that is (ancestor of) a node in $\mathbf{Z}$, implies that there is at least one leg $\{W_{k(l)}, W_{k(l+1)}\}$ in the corresponding sequence $W_{k(0)}(= W_j), W_{k(1)}, .., W_{k(m)}, W_{k(m+1)}(= W_{i+1})$ for which $k(l) \leq j$ and $W_{k(l)} \prec W_{k(l+1)}$ along π. As node $W_{k(l)}$ is an ancestor of (or equal to) node W_j, which in turn was (ancestor of) a collider in $\mathcal{G}$ that is (ancestor of) a node in $\mathbf{Z}$, it means that the path $X.. \rightarrow W_{k(l)} \leftarrow .. \rightarrow W_{k(l+1)}.. \rightarrow Y$ is an alternative unblocked path between X and Y given $\mathbf{Z}$, and so again not allowed.

In short: assuming $X \not\perp\!\!\!\perp_p V_j \,|\, \mathbf{Z}$ results in a contradiction with the original assumption of a single unblocked path π connecting X and Y in $\mathcal{G}$, and therefore $X \perp\!\!\!\perp_p V_j \,|\, \mathbf{Z}$ must apply for any node from $\mathbf{V}_i$ that is needed to create the $W_i \not\perp\!\!\!\perp_{\mathcal{G}\|} W_{i+1} \,|\, \mathbf{Z} \cup \mathbf{V}_i$. But if $X \perp\!\!\!\perp_p V \,|\, \mathbf{Z}$ holds for all $V \in \mathbf{V}_i$, then (by faithfulness) for an arbitrary node $V' \in (\mathbf{V}_i \setminus V)$ it must hold that $X \perp\!\!\!\perp_p V \,|\, \mathbf{Z} \cup V'$, otherwise, to create the dependence through conditioning on V', there have to be unblocked paths from X (and V) to V' given $\mathbf{Z}$ in $\mathcal{M}$, contradicting $X \perp\!\!\!\perp_p V' \,|\, \mathbf{Z}$. This argument can be extended to arbitrary subsets of $\mathbf{V}_i$, ergo: $\forall \mathbf{V}' \subseteq (\mathbf{V}_i \setminus V) : X \perp\!\!\!\perp_p V \,|\, \mathbf{Z} \cup \mathbf{V}'$. (end-of-proof part A)

Part (B): two successive nodes along π are dependent given $\mathbf{Z}$, equation (5.9). By contradiction. Assume the invariant $X \not\perp\!\!\!\perp_p W_i \,|\, \mathbf{Z}$ holds up to node W_i along the path, and suppose that $W_i \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z} \cup \mathbf{V}_i$, but $W_i \perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z}$. Then there is at least one $V \in \mathbf{V}_i$ needed to block all paths between W_i and W_{i+1} in $\mathcal{G}$ that is a collider between unblocked paths from these two nodes given $\mathbf{Z} \cup (\mathbf{V}_i \setminus V)$ in $\mathcal{M}$, necessary to create the dependence. By (A) we know that no subset of nodes in $\mathbf{V}_i$ can block the unblocked path in $\mathcal{M}$ between X and W_i given $\mathbf{Z}$, and so we also have (in particular): $X \not\perp\!\!\!\perp_p W_i \,|\, \mathbf{Z} \cup (\mathbf{V}_i \setminus V)$, corresponding to an unblocked path in $\mathcal{M}$ between X and W_i given $\mathbf{Z} \cup (\mathbf{V}_i \setminus V)$. Node W_i cannot be a collider between these paths in $\mathcal{M}$, because that would imply $X \not\perp\!\!\!\perp_M W_{i+1} \,|\, \mathbf{Z} \cup \mathbf{V}_i \cup W_i$, contrary equation (5.8). But if W_i is a noncollider between these paths in $\mathcal{M}$, then without W_i the path from X to V is unblocked given $\mathbf{Z} \cup (\mathbf{V}_i \setminus V)$, contrary to (A).

As W_i has to be either a collider or a noncollider between these unblocked paths in $\mathcal{M}$, it follows that the assumption of a node $V \in \mathbf{V}_i$ needed in the conditioning set to create the dependence between W_i and W_{i+1} is false. Hence they must also be dependent without conditioning on $\mathbf{V}_i$, or in other words: $W_i \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z}$. (end-of-proof part B); this completes the proof of Lemma 5.11. □

We can now formulate the general version to infer dependence from arbitrary, single unblocked paths.

Lemma 5.6. Let $\mathcal{G}$ be a uDAG for a faithful MAG $\mathcal{M}$. Let X, Y, and $\mathbf{Z}$ be disjoint (sets of) nodes. If $\pi = \langle X, .., Y \rangle$ is the *only* unblocked path from X to Y given $\mathbf{Z}$ in $\mathcal{G}$, then $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$.

Proof. The path π can be split into three parts: $\pi = \pi_1 + \pi_2 + \pi_3$, with $\pi_1 = X \leftarrow .. \leftarrow U$, the part of π that is a directed path into X, $\pi_2 = U \rightarrow .. \rightarrow C_1 \leftarrow .. \rightarrow C_k \leftarrow .. \leftarrow V$, the part with directed paths into colliders C_i along π, and $\pi_3 = V \rightarrow .. \rightarrow Y$, a directed path into Y. Note that any π can be written as a combination of one, two or all three subpaths from $\{\pi_1, \pi_2, \pi_3\}$, possibly with X and/or Y taking the role of U and/or V. For example, the case in Lemma 5.11 corresponds to $\pi = \pi_3$ with $X = V$.

For the proof, we first show in part (A) that each of the three subpaths π_1, π_2, π_3 represents a dependency $U \not\perp\!\!\!\perp_p X \,|\, \mathbf{Z}$, $U \not\perp\!\!\!\perp_p V \,|\, \mathbf{Z}$, and $V \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$, corresponding to unblocked paths in $\mathcal{M}$ given $\mathbf{Z}$. Then we show in part (B) that these can be stitched together in any combination to obtain $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$.

Part (A): Subpaths π_1 and π_3 already satisfy the antecedent of Lemma 5.11, and so represent identifiable dependencies $U \not\perp\!\!\!\perp_p X \,|\, \mathbf{Z}$, and $V \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$.

For each pair of nodes W_i, W_j on the path π_2 it holds that $\mathbf{Z}$ blocks all alternative paths π'_{ij} between them in $\mathcal{G}$ (except along π_2), otherwise the path $\pi' = \langle X, .., W_i \rangle + \pi'_{ij} + \langle W_j, .., Y \rangle$ is an alternative unblocked path in $\mathcal{G}$ between X and Y given $\mathbf{Z}$: whether $W_{i/j} \in An(\mathbf{Z})$ is a collider or noncollider along π', it does not block the path. As we assumed that π was the *only* unblocked path between X and Y, it follows there is no unblocked path π'_{ij} in $\mathcal{G}$ given $\mathbf{Z}$.
This implies in particular that for each successive pair of nodes W_i, W_{i+1} along π_2 it holds that $W_i \perp\!\!\!\perp_{\mathcal{G}\|} W_{i+1} \,|\, \mathbf{Z}$, and so (by Lemma 5.11) that $W_i \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z}$.

Furthermore, each node W_i that is not a collider along π in $\mathcal{G}$ is also not a collider between its neighbouring legs $W_{i-1} - W_i - W_{i+1}$ along the corresponding unblocked path in $\mathcal{M}$, otherwise conditioning on $\mathbf{Z} \cup W_i$ would unblock a path in $\mathcal{M}$, whereas in $\mathcal{G}$ it implies $W_{i-1} \perp\!\!\!\perp_{\mathcal{G}} W_{i+1} \,|\, \mathbf{Z} \cup W_i$, and so there is an unblocked path in $\mathcal{M}$ without W_i, corresponding to $W_{i-1} \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z}$.

But if W_i *is* a collider along π in $\mathcal{G}$ then it is also a collider between its neighbouring legs $W_{i-1} - W_i - W_{i+1}$ along the corresponding unblocked path in $\mathcal{M}$. The single unblocked path implies that there is a subset $\mathbf{Z}' \subset (\mathbf{Z} \backslash W_i)$ (in case collider W_i in $\mathcal{G}$ is itself part of $\mathbf{Z}$) such that both $W_{i-1} \perp\!\!\!\perp_{\mathcal{G}\|} W_i \,|\, \mathbf{Z}'$ and $W_i \perp\!\!\!\perp_{\mathcal{G}\|} W_{i+1} \,|\, \mathbf{Z}'$. This subset also separates W_{i-1} and W_{i+1} in $\mathcal{G}$ (otherwise it would not block all alternative paths to W_i) so that $W_{i-1} \perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z}'$. That implies that W_i is a collider in $\mathcal{M}$ between unblocked paths from W_{i-1} and W_{i+1} given $\mathbf{Z}'$, i.e. $W_{i-1} \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z}' \cup W_i$. We can expand $\mathbf{Z}'$ to include all nodes in $\mathbf{Z}$ that are not descendant of W_i in $\mathcal{G}$. The remaining subset $\mathbf{Z}^* = \mathbf{Z} \setminus \mathbf{Z}'$ contains only descendants of W_i in $\mathcal{G}$ and can

only destroy this dependence if it blocks at least one leg, say $W_{i-1} - W_i$, of this unblocked path in $\mathcal{M}$ given $\mathbf{Z}'$, so that $W_{i-1} \not\perp\!\!\!\perp_p W_i \,|\, \mathbf{Z}$. This also implies an unblocked path in $\mathcal{G}$ from W_{i-1} to a node $Z^* \in \mathbf{Z}^*$ given $\mathbf{Z} \setminus Z^*$ that does *not* go via W_i. Node W_{i+1} cannot have a similar alternative path to Z^* in $\mathcal{G}$ that does not go via W_i, because that would imply an alternative unblocked path between X and Y, bypassing W_i. Therefore, similar to the situation in Lemma 5.11, W_{i+1} and Z^* can be separated (in $\mathcal{G}$) by some set including W_i, whereas in $\mathcal{M}$ they are dependent given W_i (blocking the path $W_{i-1} - \mathbf{Z}^* .. \rightarrow W_i \leftarrow .. W_{i+1}$). The contradiction implies that the assumption the nodes in $\mathbf{Z}^*$ can block the dependence via W_i given $\mathbf{Z}'$ is false, and hence that again $W_{i-1} \not\perp\!\!\!\perp_p W_{i+1} \,|\, \mathbf{Z}$.

As this applies to each overlapping triple we can extend the dependence (again, similar to Lemma 5.11) along the entire path π_2 to obtain $U \not\perp\!\!\!\perp_p V \,|\, \mathbf{Z}$.

Part (B): If π consists of just a single subpath π_i, then the dependence is already shown above. For combinations we can connect the subpaths on root nodes U and V along π in $\mathcal{G}$ in the same fashion: Node U cannot be a collider between π_1 and π_2 in $\mathcal{M}$, because in $\mathcal{G}$ conditioning on U cannot unblock any new paths (as U was already in $An(\mathbf{Z})$, and so $X \perp\!\!\!\perp_{\mathcal{G}} V \,|\, \mathbf{Z} \cup U$, and so without U there is an unblocked path in $\mathcal{M}$ corresponding to $X \not\perp\!\!\!\perp_p V \,|\, \mathbf{Z}$. Similarly V cannot be a collider between π_2 and π_3, and therefore also for a single unblocked path $\pi = \pi_1 + \pi_2 + \pi_3$ in $\mathcal{G}$ it holds that $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$.

For empty π_2, we also find that $U (= V)$ cannot be a collider between π_1 and π_3 in $\mathcal{M}$, as conditioning on U blocks the last unblocked path in $\mathcal{G}$ between X and Y. Otherwise, any path in $\mathcal{G}$ unblocked by adding U to the conditioning set $\mathbf{Z}$ goes via a collider $C \in An(U)$. But C already has an unblocked path to $X (/Y)$ given $\mathbf{Z}$, which means that the path $\pi' = X \leftarrow .. \rightarrow C \rightarrow .. \rightarrow U \rightarrow .. \rightarrow Y$ (or vice versa for Y) is an alternative unblocked path in $\mathcal{G}$ given $\mathbf{Z}$. This is contrary the orignal assumption, and therefore conditioning on U blocks the path π but cannot open up any new path in $\mathcal{G}$, and therefore $X \perp\!\!\!\perp_p Y \,|\, \mathbf{Z} \cup U$. This implies U must be a noncollider connecting π_1 and π_3 in $\mathcal{M}$, and therefore again $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$. □

A powerful way to obtain more dependence statements is to eliminate nodes from the conditioning set $\mathbf{Z}$ that can be shown not to be needed to ensure the dependence.

Lemma 5.12. *Let $\mathcal{G}$ be a uDAG for a faithful MAG $\mathcal{M}$. Let X, Y, and $\mathbf{Z}$ be disjoint (sets of) nodes such that $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$. Let $\mathbf{Z}' \subseteq \mathbf{Z}$ be a subset such that for each $Z \in \mathbf{Z}'$ there are no (disjoint) unblocked paths $\pi_X = \langle X, .., Z \rangle$ and $\pi_Y = \langle Z, .., Y \rangle$ between X and Y in $\mathcal{G}$ given $\mathbf{Z} \setminus Z$, then $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z} \setminus \mathbf{Z}'$.*

Proof. The given $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$ establishes the existence of an unblocked path π in $\mathcal{M}$ given $\mathbf{Z}$. All nodes $Z \in \mathbf{Z}^* \subseteq \mathbf{Z}$ that are (descendants of) colliders along this unblocked path π have unblocked paths to both X and Y given $\mathbf{Z} \setminus Z$ (or given the union of $(\mathbf{Z}^* \setminus Z)$ and any subset $(\mathbf{Z} \setminus Z)$), and are therefore (by Corollary

5.9) not in $\mathbf{Z}'$. So, removing any (subset of) node(s) $\mathbf{Z}'$ from $\mathbf{Z}$ cannot introduce a noncollider on π, nor remove a necessary collider from π. Hence the unblocked path π remains unblocked in $\mathcal{M}$ given $\mathbf{Z} \setminus \mathbf{Z}'$, and so (by faithfulness) the dependence $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z} \setminus \mathbf{Z}'$ also holds. □

This approach can be extended to read even more dependencies. For example, the single unblocked path requirement in Lemma 5.6 can be relaxed, ultimately leading to a graphical criterion to read dependencies from uDAGs. However, a full analysis of inference from uDAGs would go far beyond the scope of the current chapter. Instead we focus on the mapping to the logical causal statements in the BCCD algorithm.

5.B Causal statements from uDAGs

This part of the Appendix focuses on the mapping from optimal uDAGs to logical causal statements as used in the BCCD algorithm.

Note: a *logical causal statement* L is a statement about the presence or absence of causal relations between two or three variables of the form $(X \Rightarrow Y)$, $(X \Rightarrow Y) \vee (X \Rightarrow Z)$, or $(X \not\Rightarrow Y) \equiv \neg(X \Rightarrow Y)$. We use Ł to denote the set of possible causal statements L over variables in $\mathbf{V}$. A uDAG approximation to a faithful MAG $\mathcal{M}$ is *optimal* if there exists no uDAG to $\mathcal{M}$ with fewer free parameters.

5.B.1 Minimal in/dependencies

From section 3.4 we know that all causal information can be found by identifying variables Z that *make* or *break* an independence relation between $\{X, Y\}$:

1. $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{W} \cup Z] \quad \vdash \quad (Z \Rightarrow X) \;\vee\; (Z \Rightarrow Y)$,
2. $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{W} \cup [Z] \quad \vdash \quad Z \not\Rightarrow (\{X, Y\} \cup \mathbf{W})$.

In words: a minimal independence identifies the presence of at least one from two causal relations, whereas a dependence identifies the absence of causal relations.

We can infer a minimal independence $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{Z}]$ from a uDAG if we can establish that in a given independence $X \perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$ all nodes $Z \in \mathbf{Z}$ are noncollider on some unblocked path between X and Y given the other nodes $\mathbf{Z}_{\setminus Z}$.

Lemma 5.13. *Let $\mathcal{G}$ be a uDAG to a faithful MAG $\mathcal{M}$. Then $X \perp\!\!\!\perp_p Y \,|\, [\mathbf{Z}]$ can be read from $\mathcal{G}$, iff we can infer that $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, [\mathbf{Z}]$ and that $\forall Z \in \mathbf{Z} : X, Y \not\perp\!\!\!\perp_p Z \,|\, \mathbf{Z}_{\setminus Z}$.*

Proof. In words: it suffices to establish that given a separating set $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, \mathbf{Z}$ in $\mathcal{G}$, each node Z in $\mathbf{Z}$ is dependent on both X and Y given the others.
Clearly, if the independence is not minimal in $\mathcal{G}$ then we cannot infer it is minimal in $\mathcal{M}$ (otherwise $\mathcal{M} = \mathcal{G}$ is a trivial counter), so we can start from $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, [\mathbf{Z}]$. If there is a node Z for which it is *not* possible to establish a dependence to X and Y

given the rest, then there exists a corresponding MAG in which Z is independent from X/Y given $\mathbf{Z}_{\setminus Z}$, and so by Lemma 4a not always needed in the minimal independence.
If it *does* hold, then each node $Z \in \mathbf{Z}$ is noncollider on some unblocked path between X and Y given all the others: for each node there are unblocked paths π_{XZ} and π_{ZY} from X and Y to Z given $\mathbf{Z}_{\setminus Z}$, connected by noncollider Z (otherwise not $X \perp\!\!\!\perp_p Y \mid \mathbf{Z}$), which makes $\pi_{XY} = \pi_{XZ} + \pi_{ZY}$ the required unblocked path between X and Y. Therefore $\mathbf{Z}$ is needed to block all paths in $\mathcal{M}$, and so it also represents a minimal independence in $\mathcal{M}$, i.e. $X \perp\!\!\!\perp_p Y \mid [\mathbf{Z}]$. □

Note that to establish in Lemma 5.13 that a node Z has unblocked paths to both X and Y given the others we can either show that $X \not\perp\!\!\!\perp_p Z \mid \mathbf{Z}_{\setminus Z}$ and $Z \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}_{\setminus Z}$ hold, or directly show that $X \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}_{\setminus Z}$ can be inferred from uDAG $\mathcal{G}$.

Identifying a node that breaks an independence from a uDAG follows straightforward from the definition:

Corollary 5.14. *Let $\mathcal{G}$ be a uDAG to a faithful MAG $\mathcal{M}$. Then $X \not\perp\!\!\!\perp_p Y \mid \mathbf{W} \cup [Z]$ can be read from $\mathcal{G}$, iff $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{W}$, and both $X \not\perp\!\!\!\perp_p Z \mid \mathbf{W}$ and $Z \not\perp\!\!\!\perp_p Y \mid \mathbf{W}$ can be inferred.*

Proof. If Z has unblocked paths to X and Y given $\mathbf{W}$, then Z is a collider between these paths in $\mathcal{M}$ (otherwise not $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{W}$), and so including Z makes them dependent, i.e. $X \not\perp\!\!\!\perp_p Y \mid [\mathbf{W}]Z$. □

It means that for inferring (both types of) causal information from uDAGs, reading dependencies remains the crucial bottleneck. From Lemma 5.6 we know that the existence of a single unblocked path is sufficient to infer a dependence, but that would miss out on many others. One way to increase the number of readable dependencies is to identify patterns of nodes that can invalidate a given unblocked path π in uDAG $\mathcal{G}$: if we find these patterns are not present for said path, then we can also infer the dependence. For that we introduce the following notion:

Definition 5.15. *In a uDAG $\mathcal{G}$, a node Z lies on an **(indirect) triangle detour** for an edge $X \to Y$, iff Z is a non-collider on a triangle with X and Y in $\mathcal{G}$, or $X \to (Z' \to Y) \leftarrow Z$ in $\mathcal{G}$. A node Z lies on an **(indirect) collider detour** for $X \to Y$, iff $X \to Z \leftarrow Y$ in $\mathcal{G}$, or if it has disjoint incoming directed paths from X and Y via (only) other (indirect) collider detour nodes for X and Y.*

The relevance lies in the following property:

Lemma 5.16. *In a uDAG $\mathcal{G}$ to a faithful MAG $\mathcal{M}$, an edge $X \to Y$ is guaranteed to imply $X \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}$ if $\mathbf{Z}$ contains all (indirect) triangle detour nodes for $X \to Y$, but no (indirect) collider detour nodes.*

Proof sketch. If $X \to Y$ in $\mathcal{G}$, then this tells us that $X \not\perp\!\!\!\perp_p Y \mid An(Y)_{\mathcal{G}}$. Suppose X and Y are not adjacent in $\mathcal{M}$, then this implies that either:

(1) a set of nodes $\mathbf{U}$, necessary for separating X and Y in $\mathcal{M}$, is not in $An(Y)_{\mathcal{G}}$, and/or
(2) a set of nodes $\mathbf{W}$ that unblock a path between m-separated X and Y in $\mathcal{M}$ are in $An(Y)_{\mathcal{G}}$.

In case of (1): let U be the first node from $\mathbf{U}$ in (some global ordering that satisfies) the partial order induced by uDAG $\mathcal{G}$, then $X \to U \leftarrow Y$ in $\mathcal{G}$, and so U is part of a collider detour for $X \to Y$. This follows from the fact that all nodes in $\mathbf{U}$ are part of some minimal separating subset (though not necessarily all together), and so U has a directed path to at least X or Y in $\mathcal{M}$, and is a noncollider on some unblocked path given the other nodes in a minimal separating set $\mathbf{Z}_i$ containing U. Therefore, conditional on any subset $An(Y)_{\setminus X}$ that still includes Y, U has an unblocked path to X in $\mathcal{M}$, so $X \to U$ in $\mathcal{G}$; similar for any subset $An(Y)_{\setminus Y}$ that still includes X, U has an unblocked path to Y in $\mathcal{M}$, so $Y \to U$ in $\mathcal{G}$. Similar for subsequent alternative blocking nodes from $\mathbf{U}$, except that now these may (or may not) be separated from X and/or Y in $\mathcal{M}$ by preceding nodes from $\mathbf{U}$ in $\mathcal{G}$, in which case they have an unblocked path to one or more of those nodes from $\mathbf{U}$, and so are part of an in direct collider detour.

In case of (2): let $\mathbf{Z}$ be the subset of predecessors of Y that are in $An(X, Y)_{\mathcal{M}}$, and let $\mathbf{W}$ be its complement $\mathbf{W} = An(Y)_{\mathcal{G}} \setminus \mathbf{Z}$. Then $X \perp\!\!\!\perp_p Y \mid \mathbf{Z}$, but there is also at least one unblocked path $\pi = \langle X, .., (Y) \rangle$ in $\mathcal{M}$ (partly) via a subset of collider nodes $\{W_1, .., W_k\} \subseteq \mathbf{W}$. All nodes along π (including X) have an unblocked path to Y in $\mathcal{M}$ given $\mathbf{Z} \cup \mathbf{W}$, and so arcs $\pi \to Y$ in $\mathcal{G}$. Using $\pi_{\setminus X}$ as shorthand for all nodes along π except X (and Y), then if $\pi_{\setminus X} \prec X$ in $\mathcal{G}$, then the same holds for arcs $\pi_{\setminus X} \to X$, and so all non-endpoint nodes along π form triangle detours for the edge $X \to Y$. If $\pi_{\setminus W} \prec W$, then W has unblocked paths to all other nodes along π in $\mathcal{M}$, and so $\pi_{\setminus W} \to W \to Y$ in $\mathcal{G}$, which again means they all form an (indirect) triangle detour for $X \to Y$.

For an arc $X \to Y$ in $\mathcal{G}$, if X and Y are also adjacent in $\mathcal{M}$ then they are dependent given any set. If not, then in case of (1) including nodes from $\mathbf{U}$ may separate them, but these nodes are all part of an (indirect) collider detour for $X \to Y$ in $\mathcal{G}$, and so excluding these from $\mathbf{Z}$ avoids destroying the dependence. Similarly, in case of (2) the dependence can be the result of an unblocked path due to conditioning on non-ancestors of X and Y in $\mathcal{M}$, but these are then all part of (indirect) triangle detours in $\mathcal{G}$, and so as long as all of these are included in $\mathbf{Z}$ the dependence $X \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}$ is ensured. □

We can string these dependencies together to form longer paths.

Corollary 5.17. *In a uDAG $\mathcal{G}$ to a faithful MAG $\mathcal{M}$, if W is a noncollider between non-adjacent X and Y, and we can infer that $X \not\perp\!\!\!\perp_p W \mid \mathbf{Z}$ and $W \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}$, then it also follows that $X \not\perp\!\!\!\perp_p Y \mid \mathbf{Z}$.*

Proof. By contradiction: assume $X \perp\!\!\!\perp_p Y \mid \mathbf{Z}$. From the given there are unblocked paths π_{XW} and π_{WY} in $\mathcal{M}$ given $\mathbf{Z}$, and the assumption implies W would need to

be a collider between these paths in $\mathcal{M}$. There cannot exist alternative directed paths out of W to X and/or Y in $\mathcal{M}$: these would need to be blocked by $\mathbf{Z}$ to ensure the independence, but that would unblock the collider path, resulting in $X \perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$, contrary the assumed.

As X and Y are not adjacent in $\mathcal{G}$ we can choose $\mathbf{U}$ from $An(X, Y, W)_{\mathcal{G}}$ such that there is only one unblocked path (edge) between X and W in $\mathcal{G}$, corresponding to an unblocked path π'_{XW} in $\mathcal{M}$, and likewise π'_{WY} in $\mathcal{M}$ for edge $W - Y$ in $\mathcal{G}$. By Lemma 4 this also implies identifiable dependency $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{U}$. But by construction, adding W to the conditioning set will separate X and Y in $\mathcal{G}$, giving $X \perp\!\!\!\perp_p Y \,|\, \mathbf{U} \cup W$ by Lemma 3. This implies that W cannot be a collider between π'_{XW} and π'_{WY} in $\mathcal{M}$, whereas the original assumption implies it must be. This means that the assumption must be false, and so indeed $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$. □

We can use the uDAG rules above to test observed (minimal) independencies in $\mathcal{G}$ for the required dependencies in Lemma 5.13 and Corollary 5.14: if there are multiple unblocked paths for a given dependence, then validating any one of them via Lemma 5.16 or Corollary 5.17 corresponds to identifying an unblocked path in faithful MAG $\mathcal{M}$, which is sufficient to infer the dependence.

We can try to find additional uDAG rules to read even more dependencies, but that would neglect another important piece of information, namely that the uDAG is also *optimal*.

5.B.2 Causal inference from optimal uDAGs

In general, the previous results assert different dependencies for different uDAG members of the same equivalence class. For *optimal* uDAGs (oDAGs for short) additional information can be inferred.

Lemma 5.18. *If $\mathcal{G}$ is an optimal uDAG to a faithful MAG $\mathcal{M}$, then all in/dependence statements that can be inferred for any uDAG instance of the corresponding equivalence class $[\mathcal{G}]$ are valid.*

Proof. All (DAG) instances in an equivalence class $[\mathcal{G}]$ can describe the same distribution with the same in/dependencies, and have the same number of free parameters. Therefore, if one is a valid (optimal) uDAG to the faithful MAG $\mathcal{M}$, then they all are. That means that all in/dependence statements derived for any of these via proper uDAG inference rules, e.g. Lemma 5.4, are valid in $\mathcal{M}$. □

Even though there can be different oDAGs (optimal uDAGs) for a MAG $\mathcal{M}$, it does mean that no edge in a given oDAG $\mathcal{G}$ can be removed without either requiring an invariant bi-directed edge in the corresponding equivalence class, or implying in/dependence statements not present in $\mathcal{M}$.

For example, knowing that Figure 5.6(c) is an *optimal* uDAG implies $X \not\perp\!\!\!\perp Y$, whereas this does not follow for 'ordinary' uDAGs (only implies $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, ZW$).

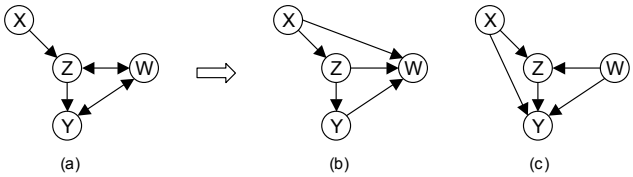

Figure 5.6: (a) MAG with invariant bi-directed edges ($\mathcal{R}$4b), (b) optimal uDAG if $r_W \leq (r_Y - 1)r_Z + 1$, (c) idem, if $r_W \geq (r_Y - 1)r_Z + 1$; with r_X the multiplicity of random variable X, etc.

It also follows that inference is most naturally done on the PAG representation $\mathcal{P}$ of the graph. In this Appendix we focus on deriving causal statements. For that we need to establish a connection between the underlying faithful MAG $\mathcal{M}$ and an optimal uDAG representation $\mathcal{G}$ (where we ignore selection bias).

Lemma 5.19. *For a faithful MAG $\mathcal{M}$, an optimal uDAG $\mathcal{G}$ is a member of an equivalence class $[\mathcal{M}']$ obtained by (only) adding arcs to $\mathcal{M}$, necessary to eliminate an arrowhead from a bi-directed edge in the PAG $\mathcal{P}(\mathcal{M}')$, until no more invariant bi-directed edges are left.*

Proof sketch. If the PAG $\mathcal{P}(\mathcal{M})$ does not contain a bi-directed edge, then there exists a DAG representative of the corresponding equivalence class $[\mathcal{M}]$, see Theorem 2 in [Zhang, 2008]. As fewer edges and fewer invariant edge marks require fewer free parameters, any such DAG is also optimal.
If the PAG $\mathcal{P}(\mathcal{M})$ *does* contain edges with two invariant arrowheads then a uDAG approximation is needed. From Section 3.2 we know that all invariant arrowheads at a node Z on a bi-directed edge $Z \leftrightarrow Y$ in a PAG are inferred from (minimal) conditional independencies $U \perp\!\!\!\perp_p V \,|\, [\mathbf{W}]$ with $Y \in (\{U, V\} \cup \mathbf{W})$ that are destroyed by conditioning on the arrowhead node Z. In this minimal dependence $U \not\perp\!\!\!\perp_p V \,|\, [\mathbf{W}]Z$ node Z has distinct unblocked incoming paths in $\mathcal{P}$ from U and V given $\mathbf{W}$ (so Z also has another invariant arrowhead to some other node, apart from Y). As a uDAG leaves every unblocked path in $\mathcal{M}$ intact, the only way to eliminate the invariant arrowhead is to 'hide' the conditional independence, by either adding an edge $U - V$, or adding edges to extend the required separating set $\mathbf{W}$. But additional nodes in the separating set can only hide the independence if at least one of these, say W', helps to block all paths between Z and, say, U. But then W' and V would also need to be separated in $\mathcal{M}$ (otherwise $W' \in \mathbf{W}$), and so the invariant arrowhead at Z still follows from a conditional independence destroyed by Z, i.e. $W' \not\perp\!\!\!\perp_p V \,|\, [..]Z$, *unless* an edge is added between the two separated nodes.
In short: to eliminate an invariant arrowhead $Z \leftarrow\!\!* Y$ edges need to be added in $\mathcal{M}$ between the two separated nodes in a non-empty subset of (minimal) conditional independencies destroyed by Z to obtain an unfaithful MAG $\mathcal{M}'$. Each added edge is in the form of an arc with the arrowhead at Y (or arbitrary orientation if

$Y \in \mathbf{W}$), *unless* this necessarily results in an almost directed cycle (not permitted in a MAG)), in which case the added edge itself becomes a bi-directed edge, which then has to be eliminated in a subsequent step. How to find which (minimal set of) edges need to be added in each step is not important to us here.

Once all required edges have been added, the collider(s) at Z is/are no longer invariant, and the newly implied possible dependence in the MAG $\mathcal{M}'$ via Z is compensated for by the implied dependencies via the added edges, in combination with parameter constraints that ensure these separate paths cancel out each other exactly. After this step the PAG $\mathcal{P}(\mathcal{M}')$ is recomputed. This process is repeated until all invariant bi-directed edges have been eliminated. At that point a there is a DAG instance $\mathcal{G}$ in the equivalence class $[\mathcal{M}']$, which is a uDAG to the faithful MAG $\mathcal{M}$, for which the number of free parameters can be calculated.
Choosing different arrowheads to eliminate in each step can lead to different uDAGs with different numbers of free parameters: the smallest one(s) correspond to the optimal uDAG(s) $\mathcal{G}$ to the faithful MAG $\mathcal{M}$. □

Note that a given MAG can have different optimal uDAG representations, possibly depending on the multiplicity of the variables as well.

Having established a connection between an optimal uDAG $\mathcal{G}$ and the underlying faithful MAG $\mathcal{M}$, we can translate this information into causal inference from the PAG representation $\mathcal{P}(\mathcal{G})$ of the observed uDAG. Fortunately the inference rule for absent causal relations takes a particularly simple form, identical to that for regular, faithful PAGs. It uses the notion of a potentially directed path (p.d.p.), introduced in Section 2.1.

Lemma 5.20. *Let $\mathcal{G}$ be an optimal uDAG to a faithful MAG $\mathcal{M}$, then the absence of a causal relation $X \not\Rightarrow Y$ can be identified, iff there is no potentially directed path from X to Y in the PAG $\mathcal{P}$ of $[\mathcal{G}]$.*

Proof. From Lemma 5a we know that the optimal uDAG $\mathcal{G}$ is obtained by (only) adding arcs between variables in the MAG $\mathcal{M}$ to eliminate invariant bi-directed edges, until no more are left. By construction, all arrowheads on arcs added in each step to obtain the next $\mathcal{M}'$ satisfy the non-ancestor relations in $\mathcal{M}$. Therefore, any remaining invariant arrowhead in the corresponding PAG $\mathcal{P}(\mathcal{M}')$ matches a non-ancestor relation in the original MAG $\mathcal{M}$. For a MAG all nodes not connected by a potentially directed path (p.d.p.) in the corresponding PAG have a definite non-ancestor relation in the underlying causal graph, see Proposition 4.9. As all unblocked paths in $\mathcal{M}$ are left intact in $\mathcal{M}'$ and $\mathcal{G}$, and a p.d.p. is by definition an unblocked path given the empty set, adding edges at each step can only *hide* non-ancestor relations still identifiable in the previous step, but never introduce new ones.
For an optimal uDAG $\mathcal{G}$ it holds that $\mathcal{P}(\mathcal{G}) = \mathcal{P}(\mathcal{M}')$ (at least for one of the possible MAG solutions), and so if there is no p.d.p from X to Y in the PAG $\mathcal{P}(\mathcal{G})$, then there is no p.d.p. from X to Y in $\mathcal{M}'$, and so also none in $\mathcal{M}$, which implies the

absence of a causal relation $X \not\to Y$. But no more then these can be inferred, as the uDAG also matches itself as faithful MAG, and for that MAG the nodes not connected by a p.d.p. in $\mathcal{P}$ are all that can be identified. □

For causal alternatives a result can also be found. Like for 'regular' uDAGs, it is based on identifying minimal conditional independencies via Lemma 4b: start from a minimal separating set $X \perp\!\!\!\perp_{\mathcal{G}} Y \,|\, [\mathbf{Z}]$ in the graph $\mathcal{P}(\mathcal{G})$, and establish dependencies from each node $Z \in \mathbf{Z}$ to X and Y given the other $\mathbf{Z}_{\setminus Z}$.

Naturally, to identify independencies in *optimal* uDAGs we can use all uDAG lemmas from Section 5.A, including in particular the 'only one unblocked path' result (Lemma 5.6). But we now also utilize a variant of the triangle/collider detours in Lemma 5.16, based on the fact that if an edge in $\mathcal{P}(\mathcal{G})$ cannot 'hide' an invariant bi-directed edge, then it cannot invalidate that edge as a dependence.

Lemma 5.21. *Let $\mathcal{G}$ be an optimal uDAG to a faithful MAG $\mathcal{M}$, and $\mathcal{P}$ the corresponding PAG of $\mathcal{G}$. Then, an edge $X - Y$ in $\mathcal{P}$ corresponds to an identifiable dependence if all nodes Z in a triangle with X and Y either satisfy:*
(1) $Z \ast\!\!\to X$ and/or $Z \ast\!\!\to Y$ are not in $\mathcal{P}$, or
(2) $Z \,—\!\ast\, X$ and/or $Z \,—\!\ast\, Y$ are in $\mathcal{P}$.

Proof. The proof of Lemma 5.16 for regular uDAGs showed that if an edge $X - Y$ was present in $\mathcal{G}$ but not in the underlying MAG $\mathcal{M}$, then it showed in the presence of either a triangle or collider detour. For optimal uDAGs collider detours do not apply, as they only introduce additional arrowheads in $\mathcal{G}$ using more parameters, and so do not appear in the construction of a optimal uDAG in Lemma 5.19. Remains to show that a node in a triangle in $\mathcal{P}(\mathcal{G})$ cannot correspond to a triangle detour:
If (1) applies, then if Z was oriented as a collider between X and Y (removing edge $X - Y$) then it would represent the same equivalence class but with fewer parameters, and so the fact that this did not occur implies this is not the case for Z.
If (2) applies, then Z is definitely a noncollider between X and Y (though not necessarily ancestor of), and so Z cannot have an invariant bi-directed edge to either that is 'hidden' in $\mathcal{G}$ by edge $X - Y$.
If this holds for all nodes in a triangle with edge $X - Y$, then there is no node that can hide an implicit collider that invalidates the edge, and so the edge represents a direct dependence. □

We can extend this to identifiable dependencies by finding a path that can be validated through Lemma 6a. In this we use the term *base path* to indicate a path/edge that is not itself a triangle detour of another path/edge in $\mathcal{P}$.

Corollary 5.22. *Let $\mathcal{G}$ be an optimal uDAG to a faithful MAG $\mathcal{M}$, and $\mathcal{P}$ the corresponding PAG of $\mathcal{G}$. Then a dependence $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$ can be inferred if there is an unblocked base path in $\mathcal{P}$ between X and Y given $\mathbf{Z}$ along which all edges can be verified to represent a direct dependence.*

Proof. If we can validate all edges along the path, e.g. by Lemma 5.6 or Lemma 5.21, then we can string these together similar to Corollary 5.17, to establish the existence of an unblocked path in the underlying faithful MAG $\mathcal{M}$ between X and Y given $\mathbf{Z}$, which ensures the dependence $X \not\perp\!\!\!\perp_p Y \,|\, \mathbf{Z}$. □

And one special alternative to validate edges as dependencies from already identified in/dependencies:

Lemma 5.23. *Let $\mathcal{G}$ be an optimal uDAG to a faithful MAG $\mathcal{M}$, and $\mathcal{P}$ the corresponding PAG of $\mathcal{G}$. Then for adjacent $X - Y$ in $\mathcal{P}$ a dependence $X \not\perp\!\!\!\perp_p Y \,|\, Z$ can be inferred, if there exist identifiable $X \not\perp\!\!\!\perp_p Y \,|\, W$ and $X \perp\!\!\!\perp_p Z \,|\, [W]$.*

Proof. If $X - Y$ in $\mathcal{P}(\mathcal{G})$, then if $X \not\perp\!\!\!\perp_p Y \,|\, W$ and $X \perp\!\!\!\perp_p Z \,|\, [W]$, then $X \not\perp\!\!\!\perp_p Y \,|\, Z$ Proof: if also $X - Y$ in $\mathcal{M}$ then X and Y are dependent given any set, so also given Z. If $X \not\perp\!\!\!\perp_p Y$ but not adjacent in $\mathcal{M}$ then they remain dependent given Z, as if Z blocks all paths between X and Y, and W blocks all paths between X and Z, then W would also block all paths between X and Y, contrary $X \not\perp\!\!\!\perp_p Y \,|\, W$. Finally, if $X \perp\!\!\!\perp_p Y$ in $\mathcal{M}$, but conditioning on W unblocks a path between them, then W cannot have a directed path to X (or Y). But that means that $X \perp\!\!\!\perp_p Z \,|\, [W]$ implies that W does have a directed path to Z in $\mathcal{M}$, and so if W unblocks the path, then so does descendant Z, and so $X \not\perp\!\!\!\perp_p Y \,|\, Z$. □

Note that lemmas 5.21 and 5.23 do not identify the minimal independencies themselves, but only verify the dependencies (in Lemma 5.13) required to find them. Also note that Lemma 5.23 builds on minimal independencies already found, which implies a recursive approach is needed to find the full mapping.

For every oDAG we may infer additional causal information by applying the standard causal inference rules on the statements obtained via the lemmas in this Appendix. Together this results in the mapping from each oDAG $\mathcal{G}$ to the set of logical causal statements Ł as used in the BCCD algorithm. Interestingly enough, for optimal uDAGs up to four nodes the mapping is *identical* to that for regular, faithful DAGs. Only at five or more nodes the distinction becomes relevant.

Bibliography

R. Ali, T. Richardson, P. Spirtes and J. Zhang (**2005**). "Towards characterizing markov equivalence classes for directed acyclic graphs with latent variables". In "Proc. of the 21st Conference on Uncertainty in Artificial Intelligence", pages 10–17.

D. Baraff and A. Witkin (**1997**). "Physically based modeling: Principles and practice, SIGGRAPH 1997". http://www.cs.cmu.edu/~baraff/sigcourse/index.html.

J. Bender and A. Schmitt (**2006**). "Constraint-based collision and contact handling using impulses". In "Proceedings of the 19th international conference on computer animation & social agents", .

R. Bouckaert (**1995**). *Bayesian Belief Networks: From Construction to Inference*. Ph.D. thesis, University of Utrecht.

W. Buntine (**1991**). "Theory refinement on Bayesian networks." In "Proc. of the 7th Conference on Uncertainty in Artificial Intelligence", pages 52–60. Morgan Kaufmann, Los Angeles, CA.

N. Cartwright (**1989**). *Natures Capacities and Their Measurement*. Clarendon Press, Oxford.

N. Cartwright (**2004**). "Causation: one word, many things". *Philosophy of Science*, (71):805–819.

L. Chen, F. Emmert-Streib and J. Storey (**2007**). "Harnessing naturally randomized transcription to infer regulatory relationships among genes". *Genome Biology*, 8(10):R219.1–13.

D. Chickering (**2002**). "Optimal structure identification with greedy search". *Journal of Machine Learning Research*, 3(3):507–554.

F. Chung and L. Lu (**2002**). "Connected components in random graphs with given expected degree sequences". *Annals of Combinatorics*, 6(2):125–145.

T. Claassen and T. Heskes (**2010**a). "Arrowhead completeness from minimal conditional independencies". Technical report, Faculty of Science, Radboud University Nijmegen.

T. Claassen and T. Heskes (**2010**b). "Causal discovery in multiple models from different experiments". In J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor and

A. Culotta, editors, "Adv. in Neural Information Processing Systems 23 (NIPS)", pages 415–423.

T. Claassen and T. Heskes (**2010**c). "Learning causal network structure from multiple (in)dependence models". In "Proc. of the Fifth European Workshop on Probabilistic Graphical Models (PGM)", pages 81–88.

T. Claassen and T. Heskes (**2011**a). "A logical characterization of constraint-based causal discovery". In "Proc. of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)", pages 135–144.

T. Claassen and T. Heskes (**2011**b). "A structure independent algorithm for causal discovery". In "Proc. of the 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)", pages 309–314.

T. Claassen and T. Heskes (**2012**a). "A Bayesian approach to constraint based causal inference". In "Proc. of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)", pages 207 – 216.

T. Claassen and T. Heskes (**2012**b). "Supplement to: A Bayesian approach to constraint based causal inference". Technical report. http://www.cs.ru.nl/~tomc/docs/BCCD_Supp.pdf.

D. Colombo, M. Maathuis, M. Kalisch and T. Richardson (**2011**). "Learning high-dimensional dags with latent and selection variables (uai2011)". Technical report, ArXiv, Zurich.

G. Cooper (**1997**). "A simple constraint-based algorithm for efficiently mining observational databases for causal relationships". *Data Min. Knowl. Discov*, 1(2):203–224.

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

P. Cvitanović, R. Artuso, R. Mainieri, G. Tanner and G. Vattay (**2010**). *Chaos: Classical and Quantum*. Niels Bohr Institute, Copenhagen. http://ChaosBook.org.

A. Dawid (**2000**). "Causal inference without counterfactuals". *Journal of the American Statistical Association*, 95(450):407–424.

R. Evans and T. Richardson (**2010**). "Maximum likelihood fitting of acyclic directed mixed graphs to binary data". In "Proc. of the 26th Conference on Uncertainty in Artificial Intelligence", pages 177–184.

C. Glymour, R. Scheines, P. Spirtes and J. Ramsey (**2004**). "The TETRAD project: Causal models and statistical data". www.phil.cmu.edu/projects/tetrad/current.

E. Guendelman, R. Bridson and R. Fedkiw (**2003**). "Nonconvex rigid bodies with stacking". *ACM Transactions on Graphics (TOG)*, 22(3):871–878.

D. Heckerman, D. Geiger and D. Chickering (**1995**). "Learning Bayesian networks: The combination of knowledge and statistical data". *Machine Learning*, 20:197–243.

D. Heckerman, C. Meek and G. Cooper (**1999**). "A Bayesian approach to causal discovery". In "Computation, Causation, and Discovery", pages 141–166.

P. Hoyer, D. Janzing, J. Mooij, J. Peters and B. Schölkopf (**2009**). "Nonlinear causal discovery with additive noise models". In "Advances in Neural Information Processing Systems 21 (NIPS*2008)", pages 689–696.

J. Ide and F. Cozman (**2002**). "Random generation of Bayesian networks". In "Advances in Artificial Intelligence", pages 366–376. Springer Berlin.

IPCC (**2008**). *Climate Change 2007, Synthesis Report*. Intergovernmental Panel on Climate Change, Geneva, Switzerland. eds. R. Pachauri and A. Reisinger.

E. Jaynes (**2003**). *Probability Theory : The Logic of Science*. Cambridge Univ. Press.

M. Kalisch, M. Mächler, D. Colombo, M. Maathuis and P. Bühlmann (**2011**). "Causal inference using graphical models with the R package pcalg." http://cran.r-project.org/web/packages/pcalg/vignettes/pcalgDoc.pdf.

D. Koller and N. Friedman (**2009**). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

D. Lewis (**1973**). *Counterfactuals*. Blackwell Publishers, England.

E. Lorenz (**1963**). "Deterministic nonperiodic flow". *J. Atmos. Sci.*, 20:130–141.

J. Mackie (**1988**). *The Cement of the Universe: A study in Causation*. Clarendon Press, Oxford, England.

S. Mani, G. Cooper and P. Spirtes (**2006**). "A theoretical study of Y structures for causal discovery". In "Proc. of the 22nd Conference in Uncertainty in Artificial Intelligence", pages 314–323.

D. Margaritis and F. Bromberg (**2009**). "Efficient Markov network discovery using particle filters". *Computational Intelligence*, 25(4):367–394.

D. Margaritis and S. Thrun (**1999**). "Bayesian network induction via local neighborhoods". In "Advances in Neural Information Processing Systems 12", pages 505–511.

C. Meek (**1995**). "Causal inference and causal explanation with background knowledge". In "UAI", pages 403–410. Morgan Kaufmann.

G. Melancon, I. Dutour and M. Bousquet-M'elou (**2000**). "Random generation of DAGs for graph drawing". Technical Report INS-R0005, Centre for Mathematics and Computer Sciences.

J. M. Mooij, O. Stegle, D. Janzing, K. Zhang and B. Schölkopf (**2010**). "Probabilistic latent variable models for distinguishing between cause and effect". In "Advances in Neural Information Processing Systems 23", pages 1687–1695.

R. Neapolitan (**2004**). *Learning Bayesian Networks*. Prentice Hall, 1st edition.

J. Pearl (**1988**). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman Publishers, San Mateo, CA.

J. Pearl (**2000**). *Causality: models, reasoning and inference*. Cambridge University Press.

J. Pearl and T. Verma (**1991**). "A theory of inferred causation". In "Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.", pages 441–452.

J. Pellet and A. Elisseef (**2008**). "Using Markov blankets for causal structure learning". *Journal of Machine Learning Research*, 9:1295–1342.

J. Ramsey, J. Zhang and P. Spirtes (**2006**). "Adjacency-faithfulness and conservative causal inference". In "Proc. of the 22nd Conference on Uncertainty in Artificial Intelligence", pages 401–408.

T. Richardson and P. Spirtes (**2002**). "Ancestral graph Markov models". *Ann. Stat.*, 30(4):962–1030.

R. Robinson (**1973**). "Counting labeled acyclic digraphs". In F. Harary, editor, "New Directions in the Theory of Graphs", pages 239–273. Academic Press.

S. Shimizu, P. Hoyer, A. Hyvärinen and A. Kerminen (**2006**). "A linear non-Gaussian acyclic model for causal discovery". *Journal of Machine Learning Research*, 7:2003–2030.

I. Shpitser, T. Richardson, J. Robins and R. Evans (**2012**). "Parameter and structure learning in nested Markov models". In "UAI2012 Workshop on Causal Structure Learning", .
URL http://www.stat.washington.edu/tsr/uai-causal-structure-learning-workshop/papers/shpitser.pdf

R. Silva and Z. Ghahramani (**2009**). "The hidden life of latent variables: Bayesian learning with mixed graph models". *Journal of Machine Learning Research*, 10:1187–1238.

P. Spirtes (**2001**). "An anytime algorithm for causal inference". In "Proc. of the Eighth International Workshop on Artificial Intelligence and Statistics (AISTATS)", pages 213–221.

P. Spirtes (**2010**). "Introduction to causal inference". *Journal of Machine Learning Research*, 11:1643–1662.

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

P. Spirtes, C. Meek and T. Richardson (**1999**). "An algorithm for causal inference in the presence of latent variables and selection bias". In "Computation, Causation, and Discovery", pages 211–252.

R. Tillman, D. Danks and C. Glymour (**2008**). "Integrating locally learned causal structures with overlapping variables". In "Advances in Neural Information Processing Systems, 21", .

S. Triantafillou, I. Tsamardinos and I. Tollis (**2010**). "Learning causal structure from overlapping variable sets". In "Proc. of the 13th Int. Conference on Artificial Intelligence and Statistics", pages 860–867.

T. Verma and J. Pearl (**1991**). "Equivalence and synthesis of causal models". In "Proc. of the 6th Annual Conference on Uncertainty in Artificial Intelligence", pages 255–270.

J. Williamson (**2005**). *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press, Oxford, UK.

G. Yule (**1903**). "Notes on the theory of association of attributes in statistics". *Biometrika*, 2:121–134.

J. Zhang (**2008**). "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias". *Artificial Intelligence*, 172(16-17):1873 – 1896.

J. Zhang and P. Spirtes (**2008**). "Detection of unfaithfulness and robust causal inference". *Minds and Machines*, 2(18):239–271.

K. Zhang, J. Peters, D. Janzing and B. Schölkopf (**2011**). "Kernel-based conditional independence test and application in causal discovery". In "Proc. of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)", pages 804–813.

Samenvatting

Een van de belangrijkste problemen waar wetenschappelijk onderzoekers telkens mee geconfronteerd worden is de vraag 'waardoor wordt dit verschijnsel veroorzaakt'? Weten welke factoren daadwerkelijk van invloed zijn geeft de mogelijkheid om doelgericht een bepaald gewenst effect te bewerkstelligen. Andersom is het soms even belangrijk te weten dat een bepaalde factor juist geen (nadelige) gevolgen heeft.

Een moeilijkheid die zich hierbij voordoet is dat een oorzaak niet altijd hoeft te leiden tot een bepaald gevolg en omgekeerd; bijv. we weten inmiddels dat roken longkanker veroorzaakt, maar niet iedereen die rookt krijgt longkanker, en longkanker komt ook voor bij mensen die niet roken. Ook is vaak niet direct duidelijk of en hoe een bepaald verband causaal is: bijv. zelfs als er een relatie is aangetoond tussen geweld op straat en gewelddadige computerspelletjes, leidt dan het spelen van de spelletjes tot geweld of zijn agressieve jongeren eerder geneigd stoere spelletjes leuk te vinden? Of zijn beide wellicht enkel een gevolg van opvoeding of sociale omstandigheden? Vaak is het niet mogelijk (te duur of onethisch) om gecontroleerde experimenten uit te voeren die uitsluitsel kunnen geven, en moeten onderzoekers proberen de causale relaties uit enkel de beschikbare data af te leiden. Dit is het terrein van de zgn. *causal discovery* methoden, en het hoofdonderwerp van dit proefschrift. Om precies te zijn: het probeert antwoord te vinden op de vraag 'waar zit nou precies de causale informatie in een data set, en hoe kun je deze er het beste uit halen?'.

In een notedop: voor zogenaamde *dynamische systemen*, d.w.z. systemen die voldoen aan (Newtoniaanse) dynamica, oftewel bijna alle voorkomende systemen, bestaat er een direct, logisch verband tussen (probabilistische) afhankelijkheden en onafhankelijkheden tussen variabelen en de aan- of afwezigheid van bepaalde causale relaties. De kern van dit proefschrift is dat het laat zien dat wanneer geobserveerde on/afhankelijkheden tussen variabelen direct vertaald worden naar deze logische uitspraken over causale relaties, en daar vervolgens mee verder te redeneren, dat het achterhalen van identificeerbare aspecten van het onderliggende causale model een bijzonder elegante en eenvoudige vorm aanneemt. De daaruit volgende mogelijkheden wordt gedemonstreerd aan de hand van een aantal concrete toepassingen.

In hoofdstuk 1 wordt gekeken naar wat we eigenlijk precies bedoelen met het doorgaans intuïtief duidelijke, maar toch verrassend ambigue begrip 'causale relatie'.

Aan de hand van het voorbeeld van een gesimuleerde worp met een punaise wordt een verband gelegd met 'effectief beïnvloedbare kansen'. Uiteindelijk leidt dit tot de conclusie dat juist *conditionele* onafhankelijkheid, d.w.z. waarbij twee variabelen onafhankelijk worden gegeven een of meer andere variabele, de cruciale link vormt tussen waarneembare patronen in data en identificeerbare causale relaties in de vorm van een grafische model, mits aan een aantal zeer redelijke aannames is voldaan.

Hoofdstuk 2 begint met een korte introductie in de theorie van grafische modellen, gevolgd door een beschrijving van een aantal bestaande methoden voor causal discovery. Hoofdstuk 3 laat vervolgens zien dat ditzelfde ook bereikt kan worden met behulp van veel minder regels (3 i.p.v. 14) via de tussentijdse transformatie van conditionele onafhankelijkheden naar logische uitspraken over causale relaties.

Vervolgens laten we een tweetal krachtige toepassingen van dit principe zien. Allereerst geeft hoofdstuk 4 aan hoe de methode gebruikt kan worden om extra causale relaties af te leiden uit de *combinatie* van meerdere data sets, zelfs als deze afkomstig zijn van *verschillende* experimenten/onderzoeken. Dit was voorheen een fundamenteel probleem waarvoor geen correcte oplossing beschikbaar was, maar volgt eenvoudig uit de in het voorgaande hoofdstuk ontwikkelde 'logische machinerie'. Daarna wordt in hoofdstuk 5 gedemonstreerd hoe deze aanpak ook gebruikt kan worden om te komen tot schattingen voor de *kans* dat een bepaalde causale relatie bestaat gegeven de beschikbare data. Door het opbreken van het causal discovery process in kleine hapklare brokjes wordt het mogelijk aan te geven welke causale relaties vrijwel zeker en welke minder waarschijnlijk zijn en waarom. Hiermee kan de betrouwbaarheid van de uiteindelijke conclusies sterk vergroot worden, hetgeen van cruciaal belang is voor het vergroten van het vertrouwen van onderzoekers in causal discovery methoden.

Tot slot biedt de in dit proefschrift getoonde aanpak tal van handvaten voor andere, veelbelovende verbeteringen: dit vormt dan ook onderwerp van ons huidige vervolgonderzoek.

Acknowledgments

I would like to take this opportunity to thank all the people who in one way or another contributed to this thesis and my time as a Ph.D. student.

It was Monique's master thesis on language and meter in Old English poetry (Beowulf) that indirectly triggered this dissertation. I realized that if I wanted to turn my interest in artificial intelligence into something more serious, I should go for an external Ph.D. Many thanks to Theo van der Weide for his support in this ambition, and in particular for putting me into contact with Tom Heskes. Tom was, understandably, rather sceptical on the chances of success for what must have seemed like 'some silly bloke wanting to do a bit of machine learning on the side', but he decided to give it a go. My interest gradually started to focus on causality and learning network structures, and I came up with an idea to infer causal relations from multiple data sets. Ultimately, we worked this into a successful proposal, which gave me this now-or-never opportunity.

Tom: I hope you feel your gamble paid off in the end. You were an excellent promotor with far more patience than I would ever have. I also really appreciate your constructive criticisms, which turned my waffling into an almost coherent read, as well as your refreshingly sane perspective on other matters.

I want to thank my roommates over the years, Tjeerd, Adriana, Frank, and Perry for putting up with random discourse, and in particular Botond and Evgeni for also being excellent travel companions abroad. Special thanks to Joris for having the stamina to work through some of my proofs and for introducing me to top-level research in many other aspects of causality. I also want to say thanks to all my other current and former colleagues: Elena, Janos, Marcel, Rasa, Ali, Saiden, Fabio, Pavol, Dimitris, Max, Daniel, Twan, Wout, Elena S., Jonce, Christiaan, Stefan, Wessel, Maya, Saskia, and Suzan, as well as Nicole and Simone for making up an inspirational machine learning group.

Finally, I want to thank friends and family for their interest. But 'First and Last and Always', my main source of support and inspiration is and always has been MQ. But if it was hard to get one of my 'brilliant ideas' past Tom, it was even harder to convince Monique. Yet despite heated debates on topics ranging from selection bias to the correct stress pattern for 'ingear dagum', she still laughs at my terrible jokes and I love her to bits. Now it is your turn!

SIKS Dissertatiereeks

====
2009
====

2009-01 Rasa Jurgelenaite (RUN)
Symmetric Causal Independence Models
2009-02 Willem Robert van Hage (VU)
Evaluating Ontology-Alignment Techniques
2009-03 Hans Stol (UvT)
A Framework for Evidence-based Policy Making Using IT
2009-04 Josephine Nabukenya (RUN)
Improving the Quality of Organisational Policy Making using Collaboration Engineering
2009-05 Sietse Overbeek (RUN)
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality
2009-06 Muhammad Subianto (UU)
Understanding Classification
2009-07 Ronald Poppe (UT)
Discriminative Vision-Based Recovery and Recognition of Human Motion
2009-08 Volker Nannen (VU)
Evolutionary Agent-Based Policy Analysis in Dynamic Environments
2009-09 Benjamin Kanagwa (RUN)
Design, Discovery and Construction of Service-oriented Systems
2009-10 Jan Wielemaker (UVA)
Logic programming for knowledge-intensive interactive applications
2009-11 Alexander Boer (UVA)
Legal Theory, Sources of Law and the Semantic Web
2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)
Operating Guidelines for Services
2009-13 Steven de Jong (UM)
Fairness in Multi-Agent Systems
2009-14 Maksym Korotkiy (VU)
From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)
2009-15 Rinke Hoekstra (UVA)
Ontology Representation - Design Patterns and Ontologies that Make Sense
2009-16 Fritz Reul (UvT)
New Architectures in Computer Chess
2009-17 Laurens van der Maaten (UvT)
Feature Extraction from Visual Data
2009-18 Fabian Groffen (CWI)
Armada, An Evolving Database System
2009-19 Valentin Robu (CWI)
Modeling Preferences, Reasoning and Collaboration in Agent-Mediated Electronic Markets
2009-20 Bob van der Vecht (UU)
Adjustable Autonomy: Controling Influences on Decision Making
2009-21 Stijn Vanderlooy (UM)
Ranking and Reliable Classification
2009-22 Pavel Serdyukov (UT)
Search For Expertise: Going beyond direct evidence
2009-23 Peter Hofgesang (VU)
Modelling Web Usage in a Changing Environment
2009-24 Annerieke Heuvelink (VUA)
Cognitive Models for Training Simulations
2009-25 Alex van Ballegooij (CWI)
RAM: Array Database Management through Relational Mapping
2009-26 Fernando Koch (UU)
An Agent-Based Model for the Development of Intelligent Mobile Services
2009-27 Christian Glahn (OU)
Contextual Support of social Engagement and Reflection on the Web
2009-28 Sander Evers (UT)
Sensor Data Management with Probabilistic Models

2009-29 Stanislav Pokraev (UT)
Model-Driven Semantic Integration of Service-Oriented Applications
2009-30 Marcin Zukowski (CWI)
Balancing vectorized query execution with bandwidth-optimized storage
2009-31 Sofiya Katrenko (UVA)
A Closer Look at Learning Relations from Text
2009-32 Rik Farenhorst (VU) and Remco de Boer (VU)
Architectural Knowledge Management: Supporting Architects and Auditors
2009-33 Khiet Truong (UT)
How Does Real Affect Affect Affect Recognition In Speech?
2009-34 Inge van de Weerd (UU)
Advancing in Software Product Management: An Incremental Method Engineering Approach
2009-35 Wouter Koelewijn (UL)
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling
2009-36 Marco Kalz (OUN)
Placement Support for Learners in Learning Networks
2009-37 Hendrik Drachsler (OUN)
Navigation Support for Learners in Informal Learning Networks
2009-38 Riina Vuorikari (OU)
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context
2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)
Service Substitution – A Behavioral Approach Based on Petri Nets
2009-40 Stephan Raaijmakers (UvT)
Multinomial Language Learning: Investigations into the Geometry of Language
2009-41 Igor Berezhnyy (UvT)
Digital Analysis of Paintings
2009-42 Toine Bogers
Recommender Systems for Social Bookmarking
2009-43 Virginia Nunes Leal Franqueira (UT)
Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients
2009-44 Roberto Santana Tapia (UT)
Assessing Business-IT Alignment in Networked Organizations
2009-45 Jilles Vreeken (UU)
Making Pattern Mining Useful
2009-46 Loredana Afanasiev (UvA)
Querying XML: Benchmarks and Recursion

====
2010
====

2010-01 Matthijs van Leeuwen (UU)
Patterns that Matter
2010-02 Ingo Wassink (UT)
Work flows in Life Science
2010-03 Joost Geurts (CWI)
A Document Engineering Model and Processing Framework for Multimedia documents
2010-04 Olga Kulyk (UT)
Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments
2010-05 Claudia Hauff (UT)
Predicting the Effectiveness of Queries and Retrieval Systems
2010-06 Sander Bakkes (UvT)
Rapid Adaptation of Video Game AI
2010-07 Wim Fikkert (UT)
Gesture interaction at a Distance
2010-08 Krzysztof Siewicz (UL)
Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments
2010-09 Hugo Kielman (UL)
A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging
2010-10 Rebecca Ong (UL)
Mobile Communication and Protection of Children

2010-11 Adriaan Ter Mors (TUD)
The world according to MARP: Multi-Agent Route Planning
2010-12 Susan van den Braak (UU)
Sensemaking software for crime analysis
2010-13 Gianluigi Folino (RUN)
High Performance Data Mining using Bio-inspired techniques
2010-14 Sander van Splunter (VU)
Automated Web Service Reconfiguration
2010-15 Lianne Bodenstaff (UT)
Managing Dependency Relations in Inter-Organizational Models
2010-16 Sicco Verwer (TUD)
Efficient Identification of Timed Automata, theory and practice
2010-17 Spyros Kotoulas (VU)
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications
2010-18 Charlotte Gerritsen (VU)
Caught in the Act: Investigating Crime by Agent-Based Simulation
2010-19 Henriette Cramer (UvA)
People's Responses to Autonomous and Adaptive Systems
2010-20 Ivo Swartjes (UT)
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative
2010-21 Harold van Heerde (UT)
Privacy-aware data management by means of data degradation
2010-22 Michiel Hildebrand (CWI)
End-user Support for Access to Heterogeneous Linked Data
2010-23 Bas Steunebrink (UU)
The Logical Structure of Emotions
2010-24 Dmytro Tykhonov
Designing Generic and Efficient Negotiation Strategies
2010-25 Zulfiqar Ali Memon (VU)
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective
2010-26 Ying Zhang (CWI)
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines
2010-27 Marten Voulon (UL)
Automatisch contracteren
2010-28 Arne Koopman (UU)
Characteristic Relational Patterns
2010-29 Stratos Idreos(CWI)
Database Cracking: Towards Auto-tuning Database Kernels
2010-30 Marieke van Erp (UvT)
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval
2010-31 Victor de Boer (UVA)
Ontology Enrichment from Heterogeneous Sources on the Web
2010-32 Marcel Hiel (UvT)
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems
2010-33 Robin Aly (UT)
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval
2010-34 Teduh Dirgahayu (UT)
Interaction Design in Service Compositions
2010-35 Dolf Trieschnigg (UT)
Proof of Concept: Concept-based Biomedical Information Retrieval
2010-36 Jose Janssen (OU)
Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification
2010-37 Niels Lohmann (TUE)
Correctness of services and their composition
2010-38 Dirk Fahland (TUE)
From Scenarios to components
2010-39 Ghazanfar Farooq Siddiqui (VU)
Integrative modeling of emotions in virtual agents
2010-40 Mark van Assem (VU)
Converting and Integrating Vocabularies for the Semantic Web
2010-41 Guillaume Chaslot (UM)
Monte-Carlo Tree Search

2010-42 Sybren de Kinderen (VU)
Needs-driven service bundling in a multi-supplier setting - computational e3-service approach
2010-43 Peter van Kranenburg (UU)
A Computational Approach to Content-Based Retrieval of Folk Song Melodies
2010-44 Pieter Bellekens (TUE)
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain
2010-45 Vasilios Andrikopoulos (UvT)
A theory and model for the evolution of software services
2010-46 Vincent Pijpers (VU)
e3alignment: Exploring Inter-Organizational Business-ICT Alignment
2010-47 Chen Li (UT)
Mining Process Model Variants: Challenges, Techniques, Examples
2010-48 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets
2010-49 Jahn-Takeshi Saito (UM)
Solving difficult game positions
2010-50 Bouke Huurnink (UVA)
Search in Audiovisual Broadcast Archives
2010-51 Alia Khairia Amin (CWI)
Understanding and supporting information seeking tasks in multiple sources
2010-52 Peter-Paul van Maanen (VU)
Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention
2010-53 Edgar Meij (UVA)
Combining Concepts and Language Models for Information Access

====
2011
====

2011-01 Botond Cseke (RUN)
Variational Algorithms for Bayesian Inference in Latent Gaussian Models
2011-02 Nick Tinnemeier(UU)
Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language
2011-03 Jan Martijn van der Werf (TUE)
Compositional Design and Verification of Component-Based Information Systems
2011-04 Hado van Hasselt (UU)
Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms
2011-05 Base van der Raadt (VU)
Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.
2011-06 Yiwen Wang (TUE)
Semantically-Enhanced Recommendations in Cultural Heritage
2011-07 Yujia Cao (UT)
Multimodal Information Presentation for High Load Human Computer Interaction
2011-08 Nieske Vergunst (UU)
BDI-based Generation of Robust Task-Oriented Dialogues
2011-09 Tim de Jong (OU)
Contextualised Mobile Media for Learning
2011-10 Bart Bogaert (UvT)
Cloud Content Contention
2011-11 Dhaval Vyas (UT)
Designing for Awareness: An Experience-focused HCI Perspective
2011-12 Carmen Bratosin (TUE)
Grid Architecture for Distributed Process Mining
2011-13 Xiaoyu Mao (UvT)
Airport under Control. Multiagent Scheduling for Airport Ground Handling
2011-14 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets
2011-15 Marijn Koolen (UvA)
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16 Maarten Schadd (UM)
Selective Search in Games of Different Complexity
2011-17 Jiyin He (UVA)
Exploring Topic Structure: Coherence, Diversity and Relatedness
2011-18 Mark Ponsen (UM)
Strategic Decision-Making in complex games
2011-19 Ellen Rusman (OU)
The Mind ' s Eye on Personal Profiles
2011-20 Qing Gu (VU)
Guiding service-oriented software engineering - A view-based approach
2011-21 Linda Terlouw (TUD)
Modularization and Specification of Service-Oriented Systems
2011-22 Junte Zhang (UVA)
System Evaluation of Archival Description and Access
2011-23 Wouter Weerkamp (UVA)
Finding People and their Utterances in Social Media
2011-24 Herwin van Welbergen (UT)
Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
2011-25 Syed Waqar ul Qounain Jaffry (VU)
Analysis and Validation of Models for Trust Dynamics
2011-26 Matthijs Aart Pontier (VU)
Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
2011-27 Aniel Bhulai (VU)
Dynamic website optimization through autonomous management of design patterns
2011-28 Rianne Kaptein(UVA)
Effective Focused Retrieval by Exploiting Query Context and Document Structure
2011-29 Faisal Kamiran (TUE)
Discrimination-aware Classification
2011-30 Egon van den Broek (UT)
Affective Signal Processing (ASP): Unraveling the mystery of emotions
2011-31 Ludo Waltman (EUR)
Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
2011-32 Nees-Jan van Eck (EUR)
Methodological Advances in Bibliometric Mapping of Science
2011-33 Tom van der Weide (UU)
Arguing to Motivate Decisions
2011-34 Paolo Turrini (UU)
Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
2011-35 Maaike Harbers (UU)
Explaining Agent Behavior in Virtual Training
2011-36 Erik van der Spek (UU)
Experiments in serious game design: a cognitive approach
2011-37 Adriana Burlutiu (RUN)
Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
2011-38 Nyree Lemmens (UM)
Bee-inspired Distributed Optimization
2011-39 Joost Westra (UU)
Organizing Adaptation using Agents in Serious Games
2011-40 Viktor Clerc (VU)
Architectural Knowledge Management in Global Software Development
2011-41 Luan Ibraimi (UT)
Cryptographically Enforced Distributed Data Access Control
2011-42 Michal Sindlar (UU)
Explaining Behavior through Mental State Attribution
2011-43 Henk van der Schuur (UU)
Process Improvement through Software Operation Knowledge
2011-44 Boris Reuderink (UT)
Robust Brain-Computer Interfaces
2011-45 Herman Stehouwer (UvT)
Statistical Language Models for Alternative Sequence Selection

2011-46 Beibei Hu (TUD)
Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
2011-47 Azizi Bin Ab Aziz(VU)
Exploring Computational Models for Intelligent Support of Persons with Depression
2011-48 Mark Ter Maat (UT)
Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
2011-49 Andreea Niculescu (UT)
Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

====
2012
====

2012-01 Terry Kakeeto (UvT)
Relationship Marketing for SMEs in Uganda
2012-02 Muhammad Umair(VU)
Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
2012-03 Adam Vanya (VU)
Supporting Architecture Evolution by Mining Software Repositories
2012-04 Jurriaan Souer (UU)
Development of Content Management System-based Web Applications
2012-05 Marijn Plomp (UU)
Maturing Interorganisational Information Systems
2012-06 Wolfgang Reinhardt (OU)
Awareness Support for Knowledge Workers in Research Networks
2012-07 Rianne van Lambalgen (VU)
When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
2012-08 Gerben de Vries (UVA)
Kernel Methods for Vessel Trajectories
2012-09 Ricardo Neisse (UT)
Trust and Privacy Management Support for Context-Aware Service Platforms
2012-10 David Smits (TUE)
Towards a Generic Distributed Adaptive Hypermedia Environment
2012-11 J.C.B. Rantham Prabhakara (TUE)
Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
2012-12 Kees van der Sluijs (TUE)
Model Driven Design and Data Integration in Semantic Web Information Systems
2012-13 Suleman Shahid (UvT)
Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
2012-14 Evgeny Knutov(TUE)
Generic Adaptation Framework for Unifying Adaptive Web-based Systems
2012-15 Natalie van der Wal (VU)
Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes
2012-16 Fiemke Both (VU)
Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
2012-17 Amal Elgammal (UvT)
Towards a Comprehensive Framework for Business Process Compliance
2012-18 Eltjo Poort (VU)
Improving Solution Architecting Practices
2012-19 Helen Schonenberg (TUE)
What's Next? Operational Support for Business Process Execution
2012-20 Ali Bahramisharif (RUN)
Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
2012-21 Roberto Cornacchia (TUD)
Querying Sparse Matrices for Information Retrieval
2012-22 Thijs Vis (UvT)
Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

2012-23 Christian Muehl (UT)
Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
2012-24 Laurens van der Werff (UT)
Evaluation of Noisy Transcripts for Spoken Document Retrieval
2012-25 Silja Eckartz (UT)
Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
2012-26 Emile de Maat (UVA)
Making Sense of Legal Text
2012-27 Hayrettin Gurkok (UT)
Mind the Sheep! User Experience Evaluation and Brain-Computer Interface Games
2012-28 Nancy Pascall (UvT)
Engendering Technology Empowering Women
2012-29 Almer Tigelaar (UT)
Peer-to-Peer Information Retrieval
2012-30 Alina Pommeranz (TUD)
Designing Human-Centered Systems for Reflective Decision Making
2012-31 Emily Bagarukayo (RUN)
A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
2012-32 Wietske Visser (TUD)
Qualitative multi-criteria preference representation and reasoning
2012-33 Rory Sie (OUN)
Coalitions in Cooperation Networks (COCOON)
2012-34 Pavol Jancura (RUN)
Evolutionary analysis in PPI networks and applications
2012-35 Evert Haasdijk (VU)
Never Too Old To Learn – Online Evolution of Controllers in Swarm- and Modular Robotics
2012-36 Denis Ssebugwawo (RUN)
Analysis and Evaluation of Collaborative Modeling Processes
2012-37 Agnes Nakakawa (RUN)
A Collaboration Process for Enterprise Architecture Creation
2012-38 Selmar Smit (VU)
Parameter Tuning and Scientific Testing in Evolutionary Algorithms
2012-39 Hassan Fatemi (UT)
Risk-aware design of value and coordination networks
2012-40 Agus Gunawan (UvT)
Information Access for SMEs in Indonesia
2012-41 Sebastian Kelle (OU)
Game Design Patterns for Learning
2012-42 Dominique Verpoorten (OU)
Reflection Amplifiers in self-regulated Learning
2012-43 Withdrawn
2012-44 Anna Tordai (VU)
On Combining Alignment Techniques
2012-45 Benedikt Kratz (UvT)
A Model and Language for Business-aware Transactions
2012-46 Simon Carter (UVA)
Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
2012-47 Manos Tsagkias (UVA)
Mining Social Media: Tracking Content and Predicting Behavior
2012-48 Jorn Bakker (TUE)
Handling Abrupt Changes in Evolving Time-series Data
2012-49 Michael Kaisers (UM)
Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
2012-50 Steven van Kervel (TUD)
Ontologogy driven Enterprise Information Systems Engineering
2012-51 Jeroen de Jong (TUD)
Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

====
2013
====

2013-01 Viorel Milea (EUR)
News Analytics for Financial Decision Support
2013-02 Erietta Liarou (CWI)
MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
2013-03 Szymon Klarman (VU)
Reasoning with Contexts in Description Logics
2013-04 Chetan Yadati(TUD)
Coordinating autonomous planning and scheduling
2013-05 Dulce Pumareja (UT)
Groupware Requirements Evolutions Patterns
2013-06 Romulo Goncalves(CWI)
The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
2013-07 Giel van Lankveld (UT)
Quantifying Individual Player Differences
2013-08 Robbert-Jan Merk(VU)
Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
2013-09 Fabio Gori (RUN)
Metagenomic Data Analysis: Computational Methods and Applications
2013-10 Jeewanie Jayasinghe Arachchige(UvT)
A Unified Modeling Framework for Service Design.
2013-11 Evangelos Pournaras (TUD)
Multi-level Reconfigurable Self-organization in Overlay Services
2013-12 Maryam Razavian (VU)
Knowledge-driven Migration to Services
2013-13 Mohammad Zafiri (UT)
Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
2013-14 Jafar Tanha (UVA)
Ensemble Approaches to Semi-Supervised Learning Learning
2013-15 Daniel Hennes (UM)
Multiagent Learning - Dynamic Games and Applications
2013-16 Eric Kok (UU)
Exploring the practical benefits of argumentation in multi-agent deliberation
2013-17 Koen Kok (VU)
The PowerMatcher: Smart Coordination for the Smart Electricity Grid
2013-18 Jeroen Janssens (UvT)
Outlier Selection and One-Class Classification
2013-19 Renze Steenhuisen (TUD)
Coordinated Multi-Agent Planning and Scheduling
2013-20 Katja Hofmann (UVA)
Fast and Reliable Online Learning to Rank for Information Retrieval
2013-21 Sander Wubben (UvT)
Text-to-text generation by monolingual machine translation